ARABIC GOAL-ORIENTED CONVERSATIONAL AGENTS USING SEMANTIC SIMILARITY TECHNIQUES

ZAID IZZADIN MOHAMMED NOORI

A thesis submitted in partial fulfilment of the requirements of the Manchester Metropolitan University for the degree of Doctor of Philosophy

April 2015

Supervisors

Dr Keeley Crockett and Dr Zuhair Bandar

Declaration

I declare that no portion of the work referred to in this thesis has been previously submitted for a degree or qualification at any other university or other institute of learning

Zaid I M Noori

Dedication

To my beloved parents who raised me on

Hard work and devotion

بسم الله الرحمن الرحيم

﴿ وَأَنْ لَيْسَ لِلْإِنْسَانِ إِلَّا مَا سَعَى (39) وَأَنَّ سَعْيَهُ سَوْفَ يُرَى (40) ثُمَّ يُجْزَاهُ الْجَزَاءَ الْأَوْفَى (41)﴾

سورة النجم

Acknowledgement

I would like to express my sincere gratitude to my supervisors Dr Keeley Crockett and Dr Zuhair Bandar for their continuous support, true guidance, patience and encouragement during my study.

Besides, I would like to thank Dr Jim O'Shea for his help and support during my research.

Special thanks also to my mother, brothers and family for their continuous help, patience and support during my studies, and also to Mr Mohammed Wed who helped me with his skills in web programming.

Finally, my thanks to my colleagues at the Iraqi Consulate in Manchester who helped in testing and evaluating the programmes, and also my colleagues at the Ministry of Foreign Affairs in Iraq for their help and support.

Abstract

Conversational agents (CAs) are computer programs used to interact with humans in conversation. Goal-Oriented Conversational agents (GO-CAs) are programs that interact with humans to serve a specific domain of interest; its' importance has increased recently and covered fields of technology, sciences and marketing. There are several types of CAs used in the industry, some of them are simple with limited usage, others are sophisticated. Generally, most CAs were to serve the English language speakers, a few were built for the Arabic language, this is due to the complexity of the Arabic language, lack of researchers in both linguistic and computing. This thesis covered two types of GO-CAs. The first is the traditional pattern matching goal oriented CA (PMGO-CA), and the other is the semantic goal oriented CA (SGO-CA).

Pattern matching conversational agents (PMGO-CA) techniques are widely used in industry due to their flexibility and high performance. However, they are labour intensive, difficult to maintain or update, and need continuous housekeeping to manage users' utterances (especially when instructions or knowledge changes). In addition to that they lack for any machine intelligence.

Semantic conversational agents (SGO-CA) techniques utilises humanly constructed knowledge bases such as WordNet to measure word and sentence similarity. Such measurement witnessed many researches for the English language, and very little for the Arabic language.

In this thesis, the researcher developed a novelty of a new methodology for the Arabic conversational agents (using both Pattern Matching and Semantic CAs), starting from scripting, knowledge engineering, architecture, implementation and evaluation. New tools to measure the word and sentence similarity were also constructed. To test performance of those CAs, a domain representing the Iraqi passport services was built. Both CAs were evaluated and tested by domain experts using special evaluation metrics. The evaluation showed very promising results, and the viability of the system for real life.

List of publications

The following paper has reported some of work related to this thesis:

Noori, Z., Bandar, Z. , and Crockett, K., 2014. Arabic Goal-oriented Conversational Agent, Based on Pattern Matching and Knowledge Trees. Proceedings of the World Congress on Engineering 2014 Vol I, ISBN: 978-988-19252-7-5, ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online) 2014.

List of abbreviations	
AI	Artificial intelligence
AIML	Artificial Intelligence Markup Language
ALICE	The Artificial Linguistic Internet Computer Entity
API	Application Programming Interface
ASU	Applied Science University
AWN	Arabic WordNet
AWSD	Arabic Word Sense Disambiguation
AWSS	Arabic Word Semantic Similarity
BNF	Backus–Naur Form
CA	Conversation Agent
DARPA	Defence Advanced Research Projects Agency
DF	Difference between sentences
ECA	Embodied Conversational Agents
FAQs	Frequently Asked Questions
FW	Function Words
GO-CA	Goal-Oriented Conversational Agent
GPS	Ground positioning System
HR	Human Rating
IEEE	Institute of Electrical and Electronics Engineers
IPS	Iraqi Passport Services
IR	Information Retrieval
KE	Knowledge Engineering
KIF	Knowledge Interchange Format
KMP	Knuth–Morris–Pratt string searching algorithm
LAW	List of Arabic Words
LCA	Linguistic Conversational Agents
LCS	Lowest Common Subsumer
MIT	Massachusetts Institution of Technology
MMU	Manchester Metropolitan University
NLP	Natural Language Processing
PM	Pattern Matching
PMGO-CA	Pattern Matching Goal-Oriented Conversational Agent.
SCA	Semantic Conversational Agent
SCAF	Semantic Conversational Agents Framework
SDS	Spoken Dialogue System
SGO-CA	Semantic Goal-Oriented Conversational Agent
SM	Semantic Matrix
SQL	Structured Query Language
STASIS	Sentence Similarity Based on Semantic Nets and Corpus Statistics
SUMO	Suggested Upper Merged Ontology
SUO-KIF	Suggested Upper Merged Ontology Knowledge Interchange Format
SV	Similarity Vector
TDS	Text-based Dialogue Systems
WPPMI	weighted positive point-wise mutual information

List of abbreviations	
WSD	Word Sense Disambiguation
WST	Word Similarity Threshold
XML	Extensible Markup Language

Table of Contents

List of figures

List of tables

Chapter 1
Introduction

Machine intelligence has focused researchers' interest since 1950, it was first inspired by Alan Turing (Turing, 1950) through his famous question "Can machines think?" This question was the motivation for researchers to seek an answer.

To answer this question, researchers developed what were known as chatbots (Chatbots.org, 2005), these Chatbots were designed to converse with human just for the sake of keeping up the conversation for as long as needed to pass the test. Most of these chatbots relied on rephrasing users' utterances to generate what looked like a viable and reasonable response, however those chatbots lacked any form of intelligence.

At a later stage, Chatbots were developed further into what is known now as Conversational Agents (CAs) (Crockett, et al., 2011) to help solving real life problems by simulating human knowledge not just to strive aimlessly to prolong the conversations. Since then machine intelligence has been an ultimate goal in the history of computer science.

The work in this thesis shall answer the following questions:

1. Can pattern matching CAs be used effectively for the Arabic language in a domain of interest?
2. Is it possible to develop an Arabic semantic conversational agent?
3. Does the semantic CA introduce a significant improvement over pattern matching CAs?
4. Is it possible to simulate human short and long term memory?
5. Can pattern matching or semantic CAs effectively cover the domain of interest?

6. Are existing methods for sentence similarity suitable to be embedded within an Arabic semantic CA?

1.1. Research Aims and Objectives:

To answer the research questions, the following objectives were set by the author to be achieved

1. Review the existing Arabic and English Conversational Agents, with an emphasis on the Goal Oriented CAs, and also emphasis on adaptable Conversation Agents.
2. Research into semantic word and sentence similarity measures in both English and Arabic language.
3. Investigate the use of short and long term memory within CAs through human semantic memory system and examine if memory mechanism can be developed within CAs
4. Using appropriate Knowledge Engineering methods to obtain user goals which are used to model the domain knowledge.
5. Design and develop a novel semantic based adaptable Arabic Goal-Oriented Conversational Agent (AGO-CA) which incorporates word and sentence similarity measures.
6. Development of a human semantic similarity memory system to capture and recall short term memory from conversation.
7. Conduct usability evaluation of the AGO-CA for the selected domain.

1.2. Research Contributions

- A novel Architecture for Arabic CAs using knowledge trees.

- Long-term memory management in CAs.
- An evaluation methodology for Conversational Agents.
- Development and evaluation of an Arabic pattern matching goal-oriented CA.
- Development and evaluation of an Arabic semantic goal-oriented CA.
- Development and evaluation of new measure for word semantic similarity in Arabic.
- Including sentence difference in sentence similarity measurement.
- Conversational Agent Development Tools.

1.3. Background

Researchers have succeeded in developing many types of CAs; most of them revolved around the idea of using pattern matching techniques, where the scripter writes many different patterns of users' utterances in order to script a conversation.

Although pattern matching CAs can offer good performance, they still lack any form of intelligence, it is up to the scripter to write enough patterns to handle different users' utterances. As time passes and information changes, the CA would need an effort by the scripter to update the scripts. This makes the conversational agent cumbersome to manage and these patterns might eventually conflict with one another.

To overcome the weaknesses of pattern matching, another approach to develop conversational agents has emerged recently, relying on semantic relations between texts to estimate similarity instead of the pattern matching approach.

An attempt has been made in English to incorporate similarity measures into conversational agents as a replacement to pattern matching (O'Shea, et al., 2010). Semantic CAs focus on estimating the relatedness of user utterance and the canonical sentences stored within the CA to generate responses.

Unlike pattern matching CAs, semantic CAs are expected to achieve more machine intelligence by eliminating the need for patterns and replace them with similarity measurement between users utterances and prototype sentences stored within the CA.

The use of semantic similarity measures also reduces the effort required to update CA's patterns and rules. Yet, the research on these types of agents is still in its early days, as the work was focused on developing similarity measurement methods and their related tools such as semantic networks and WordNet.

The Arabic language received little attention in the field of CAs development, the only work in this field was attempted by (Hijjawi, 2011) to develop a pattern-matching Arabic CA.. To the best of the researcher knowledge no attempt has been made to develop Arabic semantic conversational agents, this is mainly due to the linguistic complexities of the Arabic language and the absence of a well-developed semantic similarity measures for the Arabic language.

The research presented in this thesis proposes a new architecture for the Arabic conversational agents, which is used to construct an Arabic pattern matching goal-oriented CA to overcome the weaknesses associated with previous Arabic CA constructed by (Hijjawi, 2011). This research also studied the need for semantic CAs and proposed a new one for the Arabic language. Both CAs developed in this work were evaluated by human participants. This thesis ends with a comparative study between pattern matching CAs and semantic CAs and a conclusion.

1.4. Thesis Outline

Chapter 2: Conversational Agents.

This chapter gives an overview of previous work and techniques used within conversational agents, their evaluation methodology and a general outline of the knowledge engineering process. The chapter also describes some of the challenges associated with developing conversational agents, and the linguistic complexities of the Arabic language such as Arabic grammar and morphology.

Chapter 3: Sentence Similarity Measurement.

Chapter three gives an overview on sentence similarity measurement methods, and the resources used to measure them such as WordNet and SUMO ontology. Then the chapter delves into the details of existing word and sentence similarity measures, discussing their strengths and weaknesses.

This chapter also covers some of the problems associated with using these methods in the Arabic language, and the limitation of linguistic tools used to perform word and sentence similarity in Arabic.

Chapter 4: Arabic Conversational Agents: Architecture and Scripting Language

This chapter begins with the methodology of developing conversational agents, starting with knowledge engineering, architecture design, implementation and evaluation.

The chapter describes the knowledge engineering process starting by gathering information about the domain of study and how this information is modelled and transformed into a knowledge trees to serve as a knowledge base for CAs.

Then this chapter introduces a new architecture for Arabic conversational agents to overcome the weaknesses of previous Arabic CA, such as poor dialogue flow control and slow performance. Each part of the new architecture is explained in details and the role of each in the overall performance of CAs.

A full description of all the features of the new architecture and the new pattern matching goal-oriented CA (PMGO-CA) is also covered, these include: dialogue flow control, Accuracy, user-agent Interaction, Flexibility, Adaptability, and Memory management.

The new scripting language PMGO-CA is also covered with full explanation of the pattern matching process between user's utterances and patterns stored within the knowledge tree of the CA.

This chapter also provides full description of software tools used to construct the Arabic PMGO-CA; their features and advantages.

Chapter 5: Pattern Matching Goal Oriented Conversational Agent Evaluation

The chapter introduces an evaluation methodology for the conversational agents, including subjective and objective evaluation metrics, human participants, and the questionnaire used to evaluate the agent.

This chapter also covers the results of PMGO-CA evaluation with elaboration and analysis.

Chapter 6: Semantic Goal-Oriented Conversational Agent

This chapter introduces a novel semantic goal orientated Conversational Agent (SGO-CA). The new semantic CA incorporates a similarity measure instead of pattern matching techniques.

The chapter also covers the information sources used to estimate the similarity between words and sentences, and the similarity measures used to calculate them. Finally, the chapter proposed modifications and adaptations for the existing measures and introduces a new measure for computing Arabic word similarity. The chapter ends with the implementation of SGO-CA for the domain of study.

Chapter 7: Experiments and Evaluation of SGO-CA

This chapter is split into two parts. The first part covers a series of empirical experiments to test the proposed word and sentence similarity measures and make changes and adaptations for this measure in the context of SGO-CA. While the second part of the chapter covers human evaluation for the semantic goal-oriented conversational agent according to the same evaluation methodology developed in chapter (5).

Chapter 8: Conclusion and Future work

This chapter summarises all the work and novelties that have been achieved during the course of this research, and highlights the research results. The chapter ends with a set of recommendations for further research in the field of conversational agents and semantic similarity measurement.

Chapter 2
Conversational Agents

2.1. Introduction

Communicating with computers using natural language has been a goal in artificial intelligence for many decades. It was stimulated by the British code breaker Alan Turing, who designed what is known as the Turing Test 'TT' to test whether computers can replace humans in communicating with other humans (Turing, 1950).

Turing proposed an imitation game which is played with a man (A), a woman (B) and an interrogator (C) whose gender is unimportant. The interrogator stays in a room apart from A and B. The objective of the interrogator is to determine which of the other two is the woman is while the objective of both the man and the woman is to convince the interrogator that he/she is the woman and the other is not. This situation is depicted in Figure (2-1).

Figure 2-1 Turing Test

What would happen when a machine takes the part of A in this game? Would the results differ when the game is played with a machine instead of a woman? These questions replace the original, "Can machines think?" (Turing, 1950)

Turing's ideas have been widely discussed, attacked, and defended. (Saygin, et al., 2000). The Turing test was criticised for the fact that it has a woman and a machine each trying to convince the judge that they are a woman and the judge's task is still to decide which the woman is, and which is not. But this judge is not thinking about the differences between women and machines, but between women and men. (Hayes, et al., 1995).

Others believe that the game has been misunderstood and judged according to the performance of systems in the Loebner Prize. (Shah, 2011)

In 1990 Hugh Loebner (An American inventor) agreed with The Cambridge Centre for Behavioural Studies to underwrite a contest designed to implement the Turing Test. Dr. Loebner pledged a Grand Prize of $100,000 and a Gold Medal for the first computer whose responses were indistinguishable from a human's. Such a computer can be said "to think.". Each year an annual cash prize and a bronze medal are awarded to the most human-like computer chatbot. This encouraged researchers and experts to develop more CAs to win this prize. Some good examples of the CAs developed especially for the Loebner Prize was TIPS, CONVERSE (Wiks, 2000), ALICE (Wallace, 2003), Ella, Jabberwacky (Carpenter, 2006), Mitsuku (Worswick, 2013) and other CAs.

Computer programmers that interact with users using natural languages are called Chatbots, they usually try to keep the conversation going with users aimlessly in variety of topics. According to (Shawar, 2007) the aim of chatbots was to see if they could fool users that they were real humans.

The first chatbots was known as ELIZA, which was a simple computer program written at the Massachusetts Institute of Technology (M.I.T.) by Professor Joseph Weizenbaum between the years 1964-1966 (Weizenbaum, 1966). ELIZA used few tricks in answering questions by other questions giving the impression that the program is listening and responding to questions by answers.

ELIZA was primitive Chatbot and incapable of developing any real-world knowledge or considered application of self-awareness. However, it was the first step towards more developed and sophisticated chatbots. PARRY (Colby, 1975) was a development of ELIZA with some modifications. It was developed in 1972 by a psychiatrist called Kenneth Colby at Stanford University and was modelled on the paranoid mind. It tried to add more personality through beliefs and emotional classification (accept, reject, neutral) instead of matching trigger words (Kao, 2007) . PARRY also suffered from drawbacks, it was unable to generate responses, except for a limited number of unrepeated questions. It is worth mentioning that PARRY did not pass the Turing test.

Unlike chatbots which try to keep the conversation going aimlessly, conversational agents are designed to help users in a specific domain of interest through consistent dialogue.

(O'Shea, et al., 2011) defined Conversational agents (CAs) as "a computer program which interacts with a user through natural language dialogue and provides some form of service by processing user's input and providing a suitable response".

Conversational agents exploit natural language technologies to engage users in text-based information-seeking and task-oriented dialogs for a broad range of applications (Lester, 2004), like web-based guidance (Latham, 2010), database interfaces (Owda, et al., 2011) and tutoring (Graesser, 2005) (Latham, et al.), customer service (Noori, et al., 2014), help desk (Harbusch, et al.), guided selling (Anna3) and technical support. (Acomb, et al.).

The on-going development of internet technologies, web applications, computational linguistics, and the increasing business needs for customer service have contributed into the development of commercial conversational agents, a good sample of these CAs is Anna (Anna3), and Spleak (Chatbots.org, 2005).

The CA Anna engaged with users on a text-based dialogue to help them exploring and buying products. Anna can also respond to other non-related utterance with smooth answers trying to change the conversation to the products domain. The CA answers questions about products, prices, sizes, delivery, spare-parts, and opening hours. Anna has an animated cartoon figure which displays emotions related to her responses, like smiling while she welcomes users, etc.

Anna can respond to non-related utterance by trying to direct the conversation towards the products and services; the most remarkable thing about Anna that it picks up a clue about what the customer wants in abstracts and then offers a menu in which a user can click and select from, once a selection is made, Anna navigates the user to the desired product page where all information is available, thus Anna is not purely based on conversation, it provides services based on both conversation and web navigation.

Spleak is a spoken chatbot, it talks to people in a variety of subjects. It has an access to a number of sites like weather forecast, horoscopes, dictionaries, news, etc. and use information from such sites to keep conversation running with users. Unlike Anna who helps customers with products and services using a meaningful dialogue and web navigation, conversations carried out with Spleak were often random with the sole aim of making a conversation going for the longest period of time.

It is worth to mention that both Anna and Spleak won the Loebner prize in the years 2003 and 2006 consequently.

Conversational agents take too long and cost too much to develop (Razmerita, et al., 2004). They require expertise in the scripting of conversations and a good understanding of the written form of the language (i.e. English or Arabic). Researchers must design their own system architecture, develop knowledge representation and reasoning mechanisms, gather the required domain knowledge, and implement all system modules.

There are many challenges associated with the development of conversational agents, starting with capturing and interpreting users' utterance, disambiguating the utterance according to a given domain or context, knowledge representation and reasoning about the world or a particular domain, in addition to other challenges related to agents responsiveness, adaptability and usability.

Many English-based CAs have been developed, some of which were text-based such as ELIZA (Weizenbaum, 1966), ALICE (ALICE, 1995), PARRY (Colby, 1975) among many others. However, the Arabic language Conversational Agents has witnessed less attention, this is mainly due to the complexity of the language itself and the limited researches in Arabic linguistics, in addition to the lack of social acceptance for such applications.

This chapter covers:

- A background and review about conversational agents.
- The approaches used to develop CAs and the associated challenges
- A background and review about Arabic language and its challenges.
- The evaluation of conversational agents.
- Knowledge organisation and representation.

2.2. Natural Language Processing

Chowdhury (Chowdhury, 2005) defined Natural Language Processing (NLP) as "The area of research and application that explores how computers can be used to understand and manipulate natural language text or speech to do useful things". According to (Madnani, 2007), the term "Natural Language Processing encompasses a broad set of techniques for automated generation, manipulation and analysis of natural or human languages". (Kao, 2007), also defined NLP as "the attempt to extract a fuller meaning representation from free text". NLP aims to convert human language into a formal representation that is easy for computers to manipulate, and determine who did what to whom, when, where, how and why.

NLP utilises variety of tools and techniques including grammar rules, lexical and morphological analysis (Altabbaa, et al., 2010) (Mohtasseb, et al.) (Mazroui, 2014), noun phrase generation, word segmentation (Monroe W., 2014), semantic and discourse analysis, word meaning and knowledge representation, lexicons, thesaurus, corpus such as WordNet (Miller, 1994), VerbNet, FrameNet (Ruppenhofer, et al., 2010), the Brown corpus (Francis, et al., 1979) and the Canadian Hansard. (Roukos, et al., 1997)

According to (Nadkarni, 2011), NLP tasks are classified into low-level and a high-level tasks. Low-level NLP tasks include:

- Sentence boundary detection (READ, 2012), to determine the beginning and end of sentence.
- Tokenization (Stanford tokenizer), which divide texts into a sequence of tokens, which roughly correspond to "words".
- Part-of-speech tagging (Brill, 1992), also called grammatical tagging or word-category disambiguation, is the process of marking up a word in a text (corpus) as corresponding to a particular part of speech, based on both its definition, as well as its context
- Morphological Analysis (Altabbaa, et al., 2010), to decompose words and extract their stems and affixes.
- Shallow parsing (Abney, 1994), identifying phrases from constituent part-of-speech tagged tokens. For example, a noun phrase may comprise an adjective sequence followed by a noun. (Nadkarni, 2011)

  Higher-level tasks build on low-level tasks and are usually problem-specific (Nadkarni, 2011). Including:

- Spelling error detection (Gupta, 2012).
- Grammatical error identification (Andersen, 2011), to identify poorly formed sentences.
- Named entity recognition (Manning), identifying specific words or phrases ('entities') and categorizing them.

- Word sense disambiguation (WSD) (Rindflesch, 1994) (Weeber, 2001), determine the exact meaning of a word in a given context or sentence.
- Negation and uncertainty identification (Chapman, 2001) (Weeber, 2001), inferring whether a named entity is present or absent.
- Relation extraction (Bach) to determine relation between words, entities and concepts

NLP have been used widely in applications, including Machine Learning, information extraction ((Gupta, 2014), InQuery (Callan, 1992)), Document Retrieval (Liddy, 2001) (Richardson, 1998), machine translation, text-summarisation, web-search, human computer interfaces, education, parsing (Green, et al., 2010), customer service (Rosenfeld, et al., 2000), weather forecast (Hazen T., 1998), text mining (NetOwl, 2014) (TextWise, 2014) and conversational Agent (Rozinaj, 2012).

(Chowdhury, 2005) Stated that "at the core of any NLP task there is the important issue of natural language understanding. Building an NLP system imposes several challenges related to the interpretation and analysis of linguistic input, and knowledge representation".

Thus, a layered approach must be followed to construct an NLP system, starting at the word layer to determine the morphological structure, then the sentence layer to check the syntax according to a defined grammar in order to understand the meaning of the sentence, (who did what to whom, when, why and how) and then to the context layer to determine what this sentence means in this specific context, and what is the required action to be taken.

Accurate and efficient natural language processing is essential for an effective conversational agent to respond appropriately to users' utterances.

According to (Lester, 2004) "A conversational agent must interpret the utterance, determine and perform the actions that should be taken to respond to the utterance"; therefore A language understanding system must have a considerable knowledge about the structure of the language including the meaning of words, the grammar, and how words are combined into phrases and sentences.

However, language grammar is not always applicable since people are always changing the rules to meet their needs; therefore it is not always possible to determine the exact and complete characterisation of utterances.

An example of an NLP based CA is GALAXY (Seneff, 1998), which is a natural language system for spoken language developed at MIT. GALAXY supports English spoken and textual dialogues to help users to access online information. GALAXY interprets the utterance and frames it into defined attributes. This framing helps GALAXY to understand the utterance's topic and the information that is requested. Then, GALAXY uses a template-based response generator in order to reply with a relevant response.

However, there are strong arguments why NLP approaches are not suitable in the development of CAs. According to (Sammut, 2001), "traditional methods for Natural Language Processing (Allen J., 1995) have failed to deliver the expected performance required in a Conversational Agent" because exact grammar is rarely used in conversations; therefore the CA must have a mechanism to deal with poorly formed utterances. In addition to that, people in their daily life often use some colloquial language and expressions which might look ambiguous to the CA. For example someone might use the phrase "I've never been into baseball" to state that He/ She does not find baseball interesting.

Thus, pragmatic knowledge about the context of the conversation turns out to be a much more important factor in understanding an utterance than traditional linguistic analysis. (Sammut, 2001).

2.3. Types of Conversational Agents

There are two main types of CAs depending on their interfaces. They are Embodied Conversational Agents (ECA), and Linguistic Conversational Agents (LCA).

2.3.1. Embodied Conversational Agents

(Cassell, 2000) Defined embodied conversational agents (ECA) as "computer-generated cartoon-like characters that demonstrate many of the same properties as humans in face-to-face conversation, including the ability to produce and respond to verbal and nonverbal communication". ECAs stimulate human appearance and behaviour to communicate with people to answer questions and perform tasks for the user through natural language dialogues. (Valle, 2010) described the structure of ECAs consisting of the following main components shown in figure (2-2), they are:

- An interface to capture language or gesture input into the ECA, such as audio and gesture analysis.
- An engine or a dialogue manager to determine the ECA's behaviour.
- A visual component to perform gestures and movement, such as audio and gesture synthesis

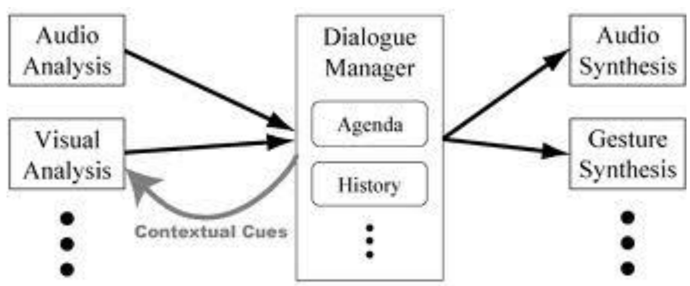


Figure 2-2 ECA's structure (Morency, et al., 2005)

ECAs are beneficial in human-computer interactions for a number of reasons. Agents could act as smart assistants, much like the ones used in travel agents or investment advisors (Catrambone, et al., 2002). A conversational interface appears to be a more natural dialogue style because the user does not have to learn complex command structure and functionality. Furthermore, an embodied agent could use intonation, gaze patterns, facial expressions and gestures.

One common trend discovered in studies is that embodied conversational agents appear to attract people's attention, both in positive and negative senses. Studies have shown that the attention gained by an embodied conversational agent had a more positive, desired effect. According to (Walker, 1994), people who interacted with a talking face spend more time on an on-line questionnaire, made fewer mistakes, and wrote more comments than those who answered a text questionnaire.

The development of an ECA requires advanced techniques for gesture and speech recognition, ECAs have challenges related to understanding human psychology. For example the ECA must capture and analyse facial expressions and gaze patterns which are different from one person to another and from one situation to another.

ECAs supporting speech commands and utterance encompasses the same challenges the Speech dialogue systems do, recognition systems must support variety of pronunciations and dialects, furthermore there are many differently spelled words but yet they are pronounced the same way such as “sea” and “see”, therefore a disambiguation mechanism is required for such utterances, the matter is even more complicated in Arabic where there are many dialects, with different words and pronunciations.

In addition, an ECA has the same other challenges associated with LCAs which are related to word sense disambiguation, morphological analysis, knowledge representation, reasoning, responsiveness, adaptability, usability, memory, etc.

ECAs have been developed for variety of applications such as tutoring (Massaro, et al.) and customer service (Kopp, et al., 2005).

Due to the complexity of the Arabic language and the variety of Arabic dialects used in the Arab countries, and the limitation of Arabic speech analysis systems, this research is focusing on the development of textual Arabic conversational agent to tackle the problems associated with conversational agents in general, and overcome the challenges associated with the Arabic language in specific. This text-based CA can serve as a base for future development of an Arabic ECA, by adding speech analysis and synthesis modules.

2.3.2. Linguistic Conversational Agents

Linguistic Conversational Agents are categorised according to their interfaces as Spoken Dialogue Systems (SDS) and Textual Dialogue Systems (TDS).

- SDS: Spoken Dialogue Systems (SDSs) are concerned with the conversion of speech into text. The average user might expect to interact with a CA by speaking to it directly and having the speech interpreted by SDS algorithms (O'Shea, et al., 2011). The goal of spoken dialogue systems (SDS) is to offer efficient and natural access to applications and services, such as email and calendars, travel and entertainment booking, and product recommendation. (Demberg, 2006).

  During the last few years several SDS were developed in many applications including:
    - Voiced-based control of home appliances, such as light and air conditioning. (Baig, et al., 2012)
    - GPS systems. (Trovato, et al., 1998)
    - E-mail services, to help users write, listen and navigate through their emails. (Walker, et al., 1997)
    - Other applications such as cinema schedules and bus trip information (P. Madeira, 2003).
- TDS: A textual Conversational Agent enables communication through a “User Interface” that has input and output textual boxes in order to receive/send an utterance/response respectively (Hijjawi, 2011).

The problem with SDS is the challenge related to capturing user’s voice, isolating it from other noise in environment, and converting voice utterance into text utterance, considering the fact that users pronounce words differently, in addition to the disambiguation part, where the agent would have to identify the intended word among many similarly pronounced words.

Furthermore an SDS would face the same complexities associated with the TDS after converting the voice utterance into text; all of these issues are magnified with the development of an Arabic conversational agent, due to the diversity of the Arabic dialects, and the lack of neat Arabic speech analysis systems.

TDS also encompasses many challenges in sentence structuring, language grammar, and morphological analysis, and word sense disambiguation. These challenges are fully covered in section 4.2.

Many LCAs have been developed since the last century such as ArabChat (Hijjawi, 2011), InfoChat (Allen J., 1995), ELIZA (Weizenbaum, 1966), (ALICE, 1995) and many others.

2.4. Approaches to Developing Linguistic Conversational Agents

2.4.1. Pattern Matching

Pattern recognition is usually considered as an engineering area which focuses on the development and evaluation of systems that imitate or assist humans in their ability of recognizing patterns (Duin, 2007). Text-based pattern matching systems can be classified into three categories,

- Question and Answering systems
- Natural Language interfaces to databases. (Susie M. Stephens).
- Conversational agents (ArabChat (Hijjawi, 2011), Student debt advisor (Crockett, et al., 2009), Bullying and harassment advisor (Latham, 2010), Intelligent tutoring system (Latham, et al.).

From a CA perspective Text-based Pattern Matching (PM) is the process of searching for a string or sequence of strings in a piece of text to find all occurrences of these strings inside that text. (Hijjawi, 2011).

Pattern matching is a technique that uses an algorithm to handle user conversations by matching CA's patterns against a user's utterance. AIML (Wallace, 2003) is the widely used pattern matching technique in conversational agents; a typical pattern consists of words, spaces, and wildcards. A wildcard is a symbol used to match a portion of the user's utterance.

Several other pattern matching algorithms have been developed by Knuth (Knuth, et al.), Boyer-Moore (Robert, et al.) Karp-Rabin (Karp, et al.).

The Knuth–Morris–Pratt string searching algorithm (or KMP algorithm) searches for occurrences of a "word" **W** within a main "text string" **S** by employing the observation that when a mismatch occurs, the word itself embodies sufficient information to determine where the next match could begin, thus bypassing re-examination of previously matched characters.

The Boyer-Moore algorithm uses information gathered during the pre-process step to skip sections of the text, resulting in a lower constant factor than many other string algorithms. In general, the algorithm runs faster as the pattern length increases. The key feature of the algorithm is to match on the tail of the pattern rather than the head, and to skip along the text in jumps of multiple characters rather than searching every single character in the text.

The Rabin–Karp algorithm or Karp–Rabin algorithm is a string searching algorithm created by Richard M. Karp and Michael O. Rabin (1987) that uses hashing to find any one of a set of pattern strings in a text.

The scripting language is the language in which patterns are defined; the most widely known scripting languages are AIML used by ALICE (Wallace, 2003) and Pattern Script used by Info Chat (Sammut, 2001).

2.4.1.1. ALICE

The Artificial Linguistic Internet Computer Entity (ALICE) is a chatbot that converse with users in variety of topics. ALICE uses as a scripting language known as AIML (Artificial Intelligence Mark-up Language) which was originally adapted from a non-XML grammar developed by Prof. Richard S. Wallace (Wallace, 2003), AIML is a scripting language which enables people to input knowledge into chatbots based on the A.L.I.C.E free software technology.

ALICE is designed to keep the client talking as long as possible, without necessarily providing any useful information along the way. The longer average conversation lengths measured over the years have in fact been a measure of A. L. I. C. E.'s progress.

According to (Wallace, 2003), AIML describes a class of data objects called AIML objects and partially describes the behaviour of computer programs that process them. AIML objects are made up of units called topics and categories. Each category consists of a pattern, a template and an optional context, pattern represents a question, while template represents an answer. The AIML pattern language is simple, consisting only of words, spaces, and the wildcard symbols as demonstrated in figure (2-3).

```
<category >
        <pattern>Hi</pattern>
         <template>Hi there!</template>
</category >
<category >
         <pattern>Hello *</pattern>
        <template><srai>Hello</srai></template>
</category >
<category>
        <pattern> What do you know about Isaac Newton</pattern>
        <template>
                <srai>Who is Isaac Newton</srai>
        </template>
</category>
```

Figure 2-3 A sample of AIML script

AIML elements begin and end with opening and closing tags, rules are organized into categories, each category contains pattern and template; the pattern is compared against user’s utterances and the template is the response which is fired once the pattern is matched.

There are three types of categories:

- Atomic Categories: are those with patterns that does not have wildcards "*" or "_"
- Default Categories: are those with patterns has a wildcards "*" or "_" which are used to replace a part of user's utterance.
- Recursive categories: It is a property of template not pattern. The template calls the pattern matcher recursively using <srai> and <sr> tags which refers to simply recursive artificial intelligence and symbolic reduction.

2.1.1.1 InfoChat and Pattern Script

InfoChat is a pattern matching conversational agent developed by ConvAgent in collaboration with the Human Computer Learning Foundation (Sammut, 2001). InfoChat has its own scripting language that structure any applied domain into a collection of text files, each text file represents a context, and each context has many rules. A rule has many patterns and associated responses.

```
museum ::

#new topic(museum, museum topics, eliza)

        init ==>

            [

                Welcome to the Powerhouse Museum and our exhibition on
                the Universal Machine. We can talk about lots of things,
                including Alan Turing and his ideas on Artificial Intelligence.

                |

                We have a great exhibit on Charles Babbage and computers in
                general.
```

```
                                        |

                                        We can talk about other things, like Robotics and Machine
                                        Learning.

                              ]

museum topics ::

{* comput~ * | * universal *} ==>

                    [

                              #goto(universal, [init])

                    ]

{* control * | * information * | * processing *} ==>

                    [

                              #goto(control, [init])

                    ]

{* communications * | * media *} ==>

                    [

                              #goto(media, [init])
```

Figure 2-4 sample of InfoChat scripts (Sammut, 2001)

As shown in figure (2-4) InfoChat scripting rules are of the form pattern ==> response. Pattern expressions may contain wildcards such as ’*’, indicating that zero or more words may match and ’~’ to indicate that zero or more characters may be matched.

Patterns can also contain non-terminal symbols, i.e., references to other pattern expressions. This enables the script writer to create abbreviations for common expressions such as lists of alternatives for the various ways in which the user can enter affirmative and negative answers. Since the definitions of non-terminal symbols may be recursive, pattern expressions are equivalent in expressive power to BNF notation.

Response expressions contain two different types of alternative constructs. Alternatives surrounded by braces ("{", "}") indicate that any element may be chosen; at random for output to the user. Alternatives surrounded by brackets ("[","]") are chosen in sequence. Thus, if the same rule fires more than once, the first alternative is chosen on the first firing, the second element on the second firing, and so on.

2.4.1.2. Issues Related to Pattern Matching

In general pattern matching based CAs encompasses the following issues:

- It is a process of searching for an occurrence in a string of text, it does not include any semantics about the words and sentence in general.
- It requires careful definitions for patterns, as some of those patterns may overlap (apply to different user utterances with different meaning) causing misfired responses.
- The scripting of pattern matching is time consuming, for each utterance the scripter must define countless number of patterns.

However Pattern matching has some advantages, in terms of responsiveness, the pattern matching process is fairly fast providing almost real-time response.

2.4.2. Sentence Semantic Similarity

(Oxford dictionary, 2015) defined a sentence as "a set of words that is complete in itself, containing a subject and predicate, conveying a statement, question, exclamation, or command, and consisting of a main clause, and sometimes one or more subordinate clauses".

Sentence semantic similarity is a measurement of the extent in which two or more sentences are similar to each other from logical perspective. An effective similarity measure should be able to determine whether sentences are semantically equivalent or not, taking into account the variability of natural language expression (Achananuparp, et al., 2008).

Measurement of text similarity have been used for a long time in wide range of applications in natural language processing and related areas, including information retrieval, automatic evaluation of machine translation (Papineni, 2002), relevance feedback and text classification, word-sense disambiguation (Resnik., 1999), language modelling (Rosenfeld, 1996), synonym extraction (Lin, 1998), and automatic thesauri extraction (Curran., 2002).

In general there are two approaches to measure semantic similarity, a statistical approach which is purely based on mathematical formulae; the other uses humanly constructed sources such as knowledge bases and thesaurus to measure semantic similarity.

Latent Semantic Analysis (LSA) is a statistical method of measuring sentence similarity. According to (Landauer, 1998) LSA is “A fully automatic mathematical/statistical technique for extracting and inferring relations of expected contextual usage of words in passages of discourse”, this method is fully explained in section (3.1)

The other approach for measuring semantic similarity utilises knowledge bases, Corpus statistics, lexicons, grammar and part-of-speech, etc. to determine sentence similarity, this method has been researched by scholars and some algorithms were proposed. The semantic similarity of two sentences is often calculated using information from a structured lexical database and corpus statistics and the semantic distance between sysnets in WordNet. Details about this method and WordNet can be found in sections (3.5.1) and (3.3) respectively

2.5. Arabic Conversational Agents

Little work has been achieved in the development of Arabic Conversational Agents. Mohammed Hijjawi (Hijjawi, 2011) developed the first known Arabic agent known as

ArabChat. ArabChat used pattern matching algorithms and classified users' utterances to be either question or non-question in order to speed up matching. The prototype agent was developed for the Applied Science University (ASU) in Jordan to work as an information point advisor for their visitor students who are Arabic native speakers. Some good trials were made to test ArabChat and showed some degree of success. However, amending the scripts in the domain in any way resulted in complex reformulation of rules within contexts and was very time consuming similar to English CAs.

Despite being successful in terms of robustness as the first Arabic conversational agent, ArabChat had some drawbacks including slow responsiveness to users' utterances and a complexity to modify rules and patterns; the scripting of ArabChat requires expertise and careful consideration in rules definition.

This section examines the challenges related to the development of an Arabic conversational agent.

2.5.1. Arabic Dialects

There are three main categories of Arabic language, they are Classic, Modern, and Colloquial Arabic (Ryding, 2005). Arab speakers usually use these different types of Arabic depending on the nature of conversation.

Classical Arabic is the original Arabic language which is used in the Quran. Classical Arabic is very rich in terms of grammar and vocabulary and encompasses a number of diacritics which are used to distinguish Arabic words and determine their pronunciation and grammatical meaning that facilitates and detect their grammatical cases (for instance, noun or verb).

However these diacritics are no longer used in Modern Arabic language, the grammatical meaning is being understood by the context of the sentence or paragraph. Modern Arabic is used as the official language in Arab countries.

The third type, known as Colloquial Arabic, is the dialect language being used in different Arab countries. A dialect spoken in one Arabic country might not be understood by people living in another country. The Arab dialects may have different vocabulary and may even contain words from other languages.

There is no standard grammar for any of the Arabic dialects, this will increase the challenge associated with the development of an Arabic CA to understand or recognize user utterances from various Arab countries.

2.5.2. Arabic Morphology

In linguistics, morphology is the study of the internal structure of words (El Kholy, 2010). It is the identification, analysis and description of the structure of morphemes and other units of meaning in a language (Altabbaa, et al., 2010).

One of the main distinguishing features of the Arabic language is the root-and-pattern morphology. The root is the semantic abstraction consisting of two, three or (less commonly) four constants from which words are derived through the superimpositions of templatic pattern. In Arabic, the word ” KTB” has the broad lexical sense of ‘writing’ from which the words for ‘book’ (KiTaab), ‘writing’ (Maktub), ‘writer’ (KaaTiB), ‘office’ (maKTab) and ‘document’ (KaTi-iBa) are derived, nouns have feminine and masculine gender and singular and plural number, and also dual in some Semitic languages. Adjectives are morphologically like nouns.

Arabic is a morphologically rich and complex language, characterised by a combination of template and affixation morphemes, complex morphological, phonological and orthographic rules, and a rich feature system. (Altantawy, et al., 2011)

Morphology usually focuses on two fundamental issues: derivational morphology, which concerns how words are formed, and secondly, inflectional morphology, which concerns how words interact with the syntax (Ryding, 2005). However, derivational morphology governs the principles of a word’s transformation

Morphological analysis and generation are important to many NLP applications such as machine translation (Diab, 2007); (El Kholy, 2010) and information retrieval (Aljlayl, et al., 2002), and conversational agents (O'Shea, et al., 2010). Much work has been done on Arabic morphological analysis and generation in a variety of approaches and at different degrees of linguistic depths. Habash (Altantawy, et al., 2011) Morphology analyser (XEROX) (Khoja, 1999), ISRI (Taghva, 2005) and Light Stemming (Larkey, 2002). (AraMorph), (Mazroui, 2014), (Mohtasseb, et al.), Qutuf (Altabbaa, et al., 2010)

2.5.3. Language Ambiguity

In Arabic language, multiple words may have different meanings. There are, two types of ambiguity in Arabic: morphological ambiguity and word sense ambiguity. Morphological ambiguity is often a result of not using the Arabic diacritics. For example the word "ذَهبْ" means gold while the same word with slightly different diacritics "ذَهبَ" means "went" .

Morphological ambiguity increases the challenge associated with the development of the Arabic conversational agents, because diacritics is usually omitted in modern Arabic language, therefore CA users are not expected to include Arabic diacritics in their utterances, which makes it hard to determine the intended word.

Word sense ambiguity occurs when two words with an exact syntactic form (including diacritics) have different meanings for example the word"يسير" "walks", means "يسير" "easy" as well, the word "يُسَلمْ" means "salute " and it also means "deliver" at the same time.

In addition, word sense ambiguity is a challenge in the development of a semantic conversational agent. A method is needed to distinguish the intended meaning of word. Many word sense disambiguation techniques (Agirre, et al., 2009) (Zouaghi, et al., 2012) (Liu, et al., 2007) (Ide, et al., 2002) (Li, 1995) have been developed but they usually require additional computational and time complexity which might not be desired in a CA environment.

WSD was the central topic of research in Natural Language Processing (NLP) for years, and more recently it was found that it is so important in many NLP tasks such as parsing, machine translation, information retrieval, question answering, conversational agent, information extraction and text mining. WSD is considered as the key step to approach language understanding (Agirre, et al., 2009)

2.5.4. Non Arabic Words Used in Arabic Dialect

The Arabic language contains countless number of non-Arabic words, for instance the word "موبايل" "Mobile" is widely used to express the cell phone devices, the word "كومبيوتر" "computer" is used for computer devices, "باسبورت" is used to define the travel document or passport. These words and other dialect words do not follow the same rules of morphological analysis and grammar. This, of course, is another challenge when developing Arabic Conversational Agent.

2.5.5. Arabic Grammar

The Arabic language has the flexibility of sentence structuring in terms of word order. The sentence structure in Arabic has three forms (El Kholy, 2010), which are: (from right to left): [object][subject][verb] (أكل محمد التفاحة), [object][verb][subject] (محمد أكل التفاحة), and [verb][subject][object] (التفاحة أكلها محمد).In contrast, the sentence structure in English might be [subject][verb][object] ('Mohammad ate an apple). Consequently, this flexibility of sentence structuring in Arabic will increase the complexity of building an Arabic CA in terms of actual sentence understanding. In Arabic language, the research into computational semantics is much smaller than other areas in NLP, due to high complexity (El Kholy, 2010).

The challenges of Arabic language can be summarised as:

- The rich morphology and the many inflectional morphological categories for Arabic language

- Arabic diacritics, there are many Arabic words with a similar spelling but they differ in diacritics (which are not being used in the modern Arabic language) which causes a morphological ambiguity.
- Word sense ambiguity, there are some Arabic words with the exact spelling and diacritics, yet they might have different meaning based on the context.
- The diverse Arabic dialects used around the world, those dialects may have foreign words, and they usually do not follow the standard Arabic grammar.

Building a pattern matching CA does not suffer from any of the challenges stated above, since pattern matching is just a process of searching for a pattern in a string of texts regardless of the spelling and grammatical structure of the sentence.

However these issues impose a real challenge in the development of an Arabic conversational agent based on semantic similarity or natural language processing in which semantic analysis and sentence understanding is required.

2.6. Evaluation of Conversational Agents:

Conversational agents like other programs, must pass rigours testing and evaluation before releasing them for public use. The evaluation of CAs is the process of performing tests on various aspects of a conversational agent by a selected group of qualified participants from different backgrounds to decide whether the agent is suitable to interact with users in real environment, and uncover any weaknesses associated within the agent based on evaluators feedback.

Chatbot evaluations have been conducted using a variety of criteria (usability, user satisfaction, Agent credibility, ease of understanding, efficiency, effectiveness, speed, and error rates etc.). Some evaluation methods tend not to assess all criteria and as there is no benchmark metrics and consistency across evaluation methods. Instead they conclude that evaluations should be adapted to user needs and the application at hand (Shawar, 2007).

Traditional evaluation methods often focus on usability criteria in a narrow sense, which correspond roughly to the concepts of usability goals (Preece, 2002). More recent approaches focused on both subjective and objective reactions in the evaluation. In addition to that, emotional aspects and user satisfaction are also included in the evaluation. These are usually referred to as user experience goals.

Although there is no standard methodology adopted by researchers to evaluate the agents, the evaluation can be classified into two major categories, subjective and objective evaluation.

Subjective evaluation usually focuses on user's satisfaction criteria, including:

- Task Ease: to measure how easy it is for a user to reach out the required information.
- Performance: which measures the level on which conversations were easy to understand
- User Expertise: to evaluate the level on which the evaluator knew what He/ She could say or do at each point of the dialogue
- Expected Behaviour: To evaluate the degree of the agent ability to meet user's expectations.
- Future use: the degree in which the user is willing to use the system instead of human experts.

Objective evaluation focuses on the actual gain of using the agent; according to (O'Shea, et al., 2010) objective evaluation metrics include:

- Dialogue / Conversation length.
- Count of dialogue turns.
- Various measures of success at utterance or task completion level.
- Various counts of errors, corrections or percentage error rates.
- Various counts of correct actions by the agent (e.g. answering questions).

- Various speech recognition accuracy measures.

Evaluation of a CA is mainly done either by distributing a questionnaire to the users trying to reveal their assessment of using the agent or by studying the resulting dialogue (Silvervarg A., 2011).

Generally, questionnaires are a particularly efficient method to apply and analyse since they enable many users with different backgrounds to evaluate several items on variety of aspects, including usability, functionality and responsiveness in addition to several other criteria which varies from one system to another. They also allow an efficient quantitative measurement of product features. Some questionnaires can under certain circumstances be used as a stand-alone evaluation method.

(Walker, et al., 1998) Identified three major limitations in subjective and objective evaluation methods:

- The use of reference answers makes it impossible to compare systems that use different dialogue strategies for carrying out the same task; such comparison requires a standard answer to be defined for every user utterance.
- Various evaluation metrics may be highly correlated with one another and thus provide redundant feedback on performance.
- The inability to trade-off or combine various metrics to make generalisations.

To overcome these limitations (Walker, et al., 1998) introduced a general framework for evaluating and comparing the performance of spoken dialogue agents called PARADISE. This was used to evaluate the DARPA communicator SDS (Walker, 2001). PARADISE uses range of methods from decision theory to combine a disparate set of performance measures such as user satisfaction task success and dialogue cost into a single performance evaluation function.

(O'Shea, et al., 2011) introduced what is known as “Wizard of Oz” to evaluate rule-based systems separately from the rest of the CA’s components. This wizard simulates the CA interface and operates the rule-based system, allowing the user to test and evaluate system rules independently but this is a very time consuming approach and less commonly used in commercial application development.

(Walker, 2001) conducted an exploratory experiment with nine participating communicator systems. All systems supported travel planning and utilised some form of mixed-initiative interaction, the evaluation of these systems included both subjective and objective evaluation, objective metrics were extracted from the logs, while subjective metrics were collected via a survey.

(O'Shea, et al., 2009) introduced an evaluation methodology for the semantic conversational agents (SCA) . Evaluation process is divided into two phases:

- Phase one: evaluates the SCA's interaction capabilities from the users' perspective: this phase is divided into two parts:
    - Part A involves an experiment which evaluates the SCA interaction using a set of participants. The evaluation included the following metrics:
        1. Usability – is the SCA easy to use?
        2. Accuracy– is the interaction with the SCA correct/ without misunderstanding?
        3. Satisfaction – is the interaction with the SCA pleasing/ trouble-free?
        4. Naturalness/Believability – is the SCA human-like?
        5. Task success – is the goal of the interaction with the SCA achieved?
        6. Repeated use – would the user consider using the SCA in the future?

    - Part B involved a comparative assessment of two CAs. The first CA was the SCA developed using the SCAF and the second was a text-based CA InfoBot. The aim of the comparative evaluation was to assess any differences between the interactions of the CAs by measuring satisfaction from the user's perspective. This was gauged by examining the different aspects of the interaction, such as usability and naturalness of the dialogue.

- Phase two: assesses natural language scripting, which is used to script the SCA. The aim of the evaluation was to determine whether or not natural language scripting enables the construction of scripts with ease, efficiency and without flaws from the script writer's perspective, the evaluation included the following metrics:
  - Intuitiveness: denotes ease-of-use.
  - Usefulness: denotes whether the features are beneficial and contribute to the ease of functionality of the SCA.
  - Flawlessness: denotes errors or deficiencies affecting the SCA's capabilities and, thus, interaction.

2.7. Knowledge Organisation in Goal-Oriented Conversational Agents

Goal-orientated Conversational Agents (GO-CAs) are a special family of conversational agents that are designed to converse with humans through the use of natural language dialogue to achieve a specific task (Crockett, et al., 2011). GO-CAs help users to satisfy their goals in a specific domain of interest, Unlike Chatbots, which strive to keeps the conversation going randomly as long as possible. The GO-CAs emulate the decision-making ability of a human expert.

One of the components of a GO-CA is a knowledge base of the domain and a set of rules similar to those found in an expert system.

2.7.1. Knowledge Acquisition

Knowledge acquisition is the accumulation, transfer, and transformation of problem-solving expertise from experts or documented knowledge sources to a computer program for constructing or expanding the knowledge base. (Trappey, 2006).

Shadbolt (Shadbolt, et al., 1999) classified knowledge according to three perspectives, they are:

- The first considers the distinction between declarative knowledge which refers to the knowledge of facts and procedural knowledge which refers to the knowledge of how to do things. These two types of knowledge are also referred to as static knowledge and dynamic knowledge.

- The second is well-known classification of knowledge is that of tacit knowledge which is difficult to articulate and explicit knowledge which is easier to articulate.
- The third perspective is related to what extent the knowledge is abstract across many situations; or specific which applies only to one or a few situations. Methods of making knowledge more abstract or specific has been a major effort in knowledge engineering.

The steps below summarises the knowledge acquisition process:

- Conduct initial interviews with the expert(s) to establish a basic understanding of the domain, key terminology and determine what knowledge to gather.
- Analyse the resulting documents, and produce a set of questions about any misunderstanding, ambiguities and issues related to the domain
- Conduct a second interview with the expert(s), using the prepared questions to reach a better understanding of the domain, also ask experts for any guides and documentations related to the domain, and also ask for a sample of procedural documents used within the domain.
- Analyse the results of the interview and the acquired documents to identify higher level information about the domain such as entities, attributes, rules, concepts and relationships between concepts
- Translate this higher level knowledge to a better understood format such as trees, organisation diagrams, work flow diagrams or flow charts
- Discuss the resulting representation with the expert(s) to expand the knowledge.
- Refine the resulting knowledge by gathering higher level information and repeating the analysis and representation process
- Validate the knowledge acquired with other experts if possible, and make modifications where necessary.

2.7.2. Knowledge Base

The knowledge base contains the relevant knowledge necessary for understanding, formulating, and solving problems. It includes two basic elements, facts such as the problem situation and the theory of the problem area; and special heuristics or rules that direct the use of the knowledge to solve specific problems in a particular domain (Trappey, 2006).

There are different ways to represent knowledge depending on the type of problem. Solving a problem almost completely determined by the way the problem is conceptualised and represented.

According to (Ramirez, et al., 2012), the types of representation models used for knowledge systems include distributed, symbolic, non-symbolic, declarative, probabilistic, ruled based, among others, each of them suited for a particular type of reasoning.

Symbolic systems are human understandable representations which use symbols as the basic representation unit; each symbol represents something like an entity, a concept, an attribute or a word. Symbolic systems were in fact the original and predominant approach in AI until the late 80's (Haugeland, 1989). Symbolic systems include structures such as semantic networks, rule based systems and frames, whereas distributed systems include different types neural or probabilistic networks.

According to (Ramirez, et al., 2012) "Non-symbolic systems use machine understandable representations based on the configuration of items, such as numbers, or nodes to represent an idea, a concept, a skill, a word. These systems are also known as distributed systems".

In Semantic networks, concepts are graphically represented as nodes, while relations between concepts are represented as arcs, nodes appear as circles or ellipses or rectangles to represent objects such as physical objects, concepts or situations while links appear as arrows to express the relationships between objects, and link labels specify particular relations. Relationships provide the basic structure for organizing knowledge. Associations have a grade which represents knowledge or strength of the association (Ramirez, et al., 2012). Semantic networks are mainly used to model declarative knowledge. However, they are flexible enough to be used with procedural knowledge. Figure (2-5) demonstrates semantic network, the IS-A link is seen by (Brachman, 1983) as a relation between the representational objects, which forms a taxonomic hierarchy, a tree or a lattice-like structures for categorising classes of things in the world being represented.

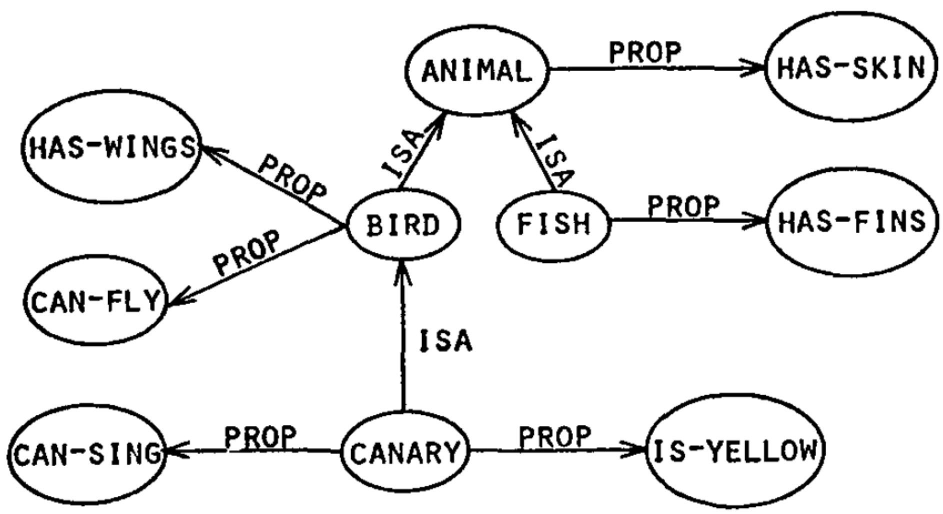


Figure 2-5 Semantic network (Shapiro, 1978)

Gruber (Gruber, 1993) defines ontology as an explicit specification of a conceptualisation. Ontologies represent knowledge as a hierarchy of concepts within a domain, using a shared vocabulary to denote the types, properties and interrelationships of those concepts.

Rule based systems are symbolic representation models which are commonly used in procedural knowledge, they contain a set of organised rules each rule is structured as a conditions and actions. Actions are fired when the associated condition is satisfied.

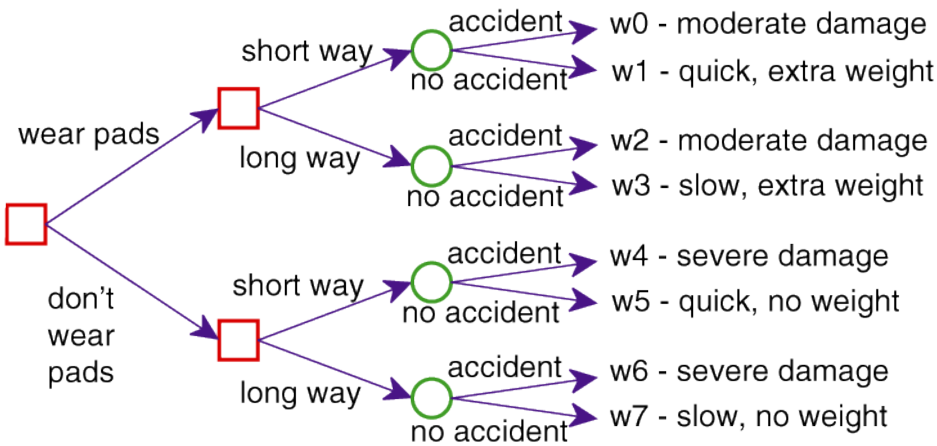


Figure 2-6 decision trees for the delivery robot. Square boxes represent decisions that the robot can make. Circles represent random variables that the robot cannot observe before making its decision (Poole, et al., 2010)

A frame is a type of semantic network which mixes declarative knowledge and structured procedural knowledge (Ramirez, et al., 2012). Frames are constructed in a way to simulate human memory in situations that mix procedural and declarative knowledge. Each symbol in a frame has associated procedures, and a group of attributes of the situation.

2.7.2.1. Knowledge Trees

(Owda, et al., 2011) defined knowledge tree as " a tree where the knowledge is organised in a hierarchical structure based on the expert knowledge which has been extracted and developed by a knowledge engineer".

Knowledge trees are used for knowledge representation in many systems (Crockett, et al., 2009) (Owda, et al., 2011). It is used to simulate the structure in which humans represent knowledge. Knowledge trees offer an easy method to revise and update knowledge bases; and serve as a map for conversational flow in a specific domain. Figure (2-7) below shows an example of knowledge trees in which the information and the knowledge is modelled in the shape connected nodes to represent domain rules and regulations.

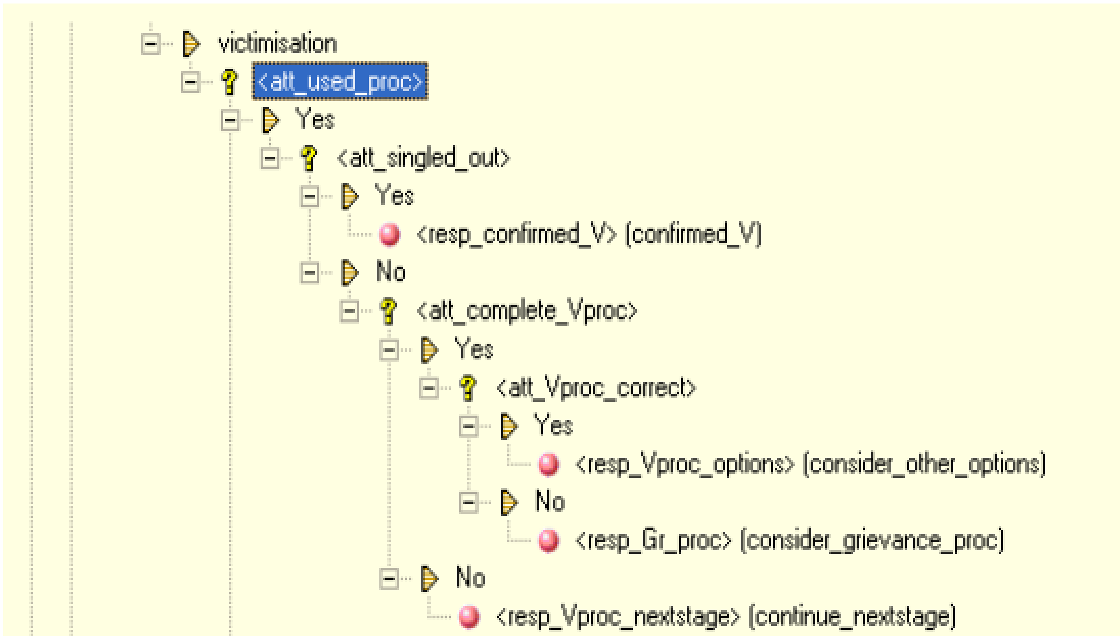


Figure 2-7 snapshot of knowledge tree used in HR Bullying and Harassment Advisor (Crockett, et al., 2009)

2.7.3. Inference engine

The inference engine provides a methodology for reasoning about information in the knowledge base to formulate a conclusion (Trappey, 2006).

According to (Wojna, 2005) decision making as a human activity is often performed on different levels of abstraction. It includes both simple everyday decisions, such as selection of products while shopping, choice of itinerary to a workplace, and more compound decisions, e.g. in marking a student's work or in investments. Decisions are always made in the context of a current situation on the basis of the knowledge and experience acquired in the past. Several research directions have been developed to support computer-aided decision making, among them are decision and game theory (Luce, 1957), planning (Nilsson., 1971), control theory (Rosenblueth, et al., 1943), and machine learning (Mitchell, 1997). The development of these directions has led to different methods of knowledge representation (introduced in section 3.4.2) and reasoning about the real world for solving decision problems.

There are different formal reasoning systems used by computers, such as:

- Deductive reasoning which is based on the assumption that knowledge is represented and extended within a deductive system. This approach is very general and it encompasses a wide range of problems. However, real-life problems are usually very complex, and depend on many factors, some of them quite unpredictable; deductive reasoning does not allow for such uncertainty.
- Inductive reasoning (Friedman, et al., 2001) (Maimon, et al., 2002) (Michalski, et al., 1986) is more suitable for real-life problems; it is based on the assumption that knowledge about a decision problem is given in the form of a set of exemplary objects with known decisions. This set is called a training set. In the learning phase the system constructs a data model on the basis of the training set and then uses the constructed model to reason about the decisions for new objects called test objects.

  The most popular Computational models used in inductive reasoning are neural networks (Bishop, 1996), decision trees (Quinlan, 1993), rule based systems (Michalski, et al., 1986), rough sets (Pawlak, 1991), Bayesian networks (Jensen., 1996), and analogy-based systems (Quinlan, 1993) (Russell., 1989)..

2.7.4. Interfaces to Expert Systems

Experts systems contain a language processor for friendly, problem-oriented, communication between the user and the computer. This communication can best be carried out in a natural language. (Trappey, 2006), Expert systems vary according to their communication interfaces, such as:

- Menu based systems, where a choice is made by selecting a choice from available menu Such as Frequently Asked Question systems (FAQs), those are used by many companies and organizations to satisfy users' questions; for example the FAQ's system of Microsoft download centre

- Speech recognition systems, in which, the system analyses user's voice to determine the input; such as the speech recognition system used in smart phones to execute specific commands based on user's voice utterance.
- Facial recognition systems, which analyses human facial expression to gain more insights about the person's attitude
- Text based expert systems, which interacts with users by analysing their textual utterances; such as ALICE (ALICE, 1995) and ArabChat (Hijjawi, 2011).

2.7.5. Problems Associated With Knowledge Engineering

Problems associated with knowledge engineering can be classified into two types. The first is knowledge acquisition process, this includes issues associated with information sources and how to obtain information from them; the other is related to the representation and modelling of this acquired information.

Knowledge acquisition process include challenges related to both language and communication as experts often use different languages, acronyms and shortcuts within their domain, they usually find it difficult to break out of this when they talk to people who are not experts in their domain, assuming that their audience has a lot more knowledge and understanding than it really does.

Language is also rather imprecise which adds another challenge. People use the same word to mean different things and use different words to mean the same thing. These characteristics of language can lead to major problems for knowledge acquisition such as lack of knowledge dissemination, and misunderstandings.

As discussed earlier in "knowledge acquisition" section, knowledge is majorly classified into tacit knowledge which is difficult to articulate and explicit knowledge which is relatively easy to articulate, both contain such a vast amount of knowledge that mapping all of it would be both impossible and a waste of time. (Shadbolt, et al., 1999)

As a result of having different types of knowledge, there are different types of experts with variety levels of experience, ranging from those whose knowledge of a domain is almost completely tacit to those whose knowledge is almost completely explicit. In addition, experts may not be able to remember the same things during interviews as they can when they are performing a task; the ability to recall the same information in different tasks can vary between individuals. For instance, those with experience of teaching others in a classroom setting are usually better at explaining their knowledge than those without such experience.

2.8. Summary

This chapter gave an overview on Conversational Agents (CAs), their definition, origin, types, and usage, with an elaboration on some CAs used and tested, showing their facilities and shortcomings. It also gave some definition and history about the Natural Language Processing (NLP) and Latent Semantic Analysis (LSA) and their role to build an understanding between the language texts and computers.

Special concentration was given to the Arabic Conversational Agents, their usage and challenges of the Arabic language; in addition to a short overview about sentence semantic similarity methods.

In general the challenges associated with the development of Arabic Conversational Agents can be summarised as:

- The complexity of the Arabic language
- The variety of spoken Arabic dialects in different Arab countries
- Word sense ambiguity
- Knowledge acquisition and modelling.
- Dialogue flow control.
- CA’s Responsiveness, Usability and Adaptability.

Chapter 3
Sentence Similarity Measurement

3.1. Introduction

Semantic similarity can be defined as the measurement of extent in which two words or sentences are similar to one another from logical perspective. Semantic similarity has important applications in many Artificial intelligence and natural language processing (NLP) fields, such as automatic question answering system (Harabagiuo, et al., 2004), Information Extraction (Hliaoutakis, et al., 2006), Machine Translation (Jeong, 2005), Conversational Agents (O'shea, 2012), Text Analysis (Malandrakis, et al., 2013), and Automatic Text Summarization (Ramiz, 2009).

This chapter gives an overview about word and sentence similarity measurement and the different methods used to compute them, along with the advantages and disadvantages of each method. It also focuses on the Arabic word and sentence similarity, and the challenges associated with these methods, the tools used to measure semantic similarity are also discussed in details such as WordNet (Princeton University, 2005), AraMorph (Buckwalter, 2002) and Suggested Upper Merged Ontology (SUMO). This chapter also covers the evaluation methods for both word and sentence similarity.

The methods and techniques described in this chapter shall be used to match user’s utterance against standard sentences stored in agent rules.

Sentence similarity for English language has been deeply researched by many scholars. Generally, there are two main approaches to measure sentence similarity. The first is based on semantic networks such as WordNet (Princeton University, 2005) by calculating similarity between each word in both sentences, then calculating sentence semantic similarity; which can be a function of the similarity between each pair of words. An example of this approach is the STASIS method developed by (Li, et al., 2006) which was covered in section (3.5.1)

The second method is called Latent Semantic Analysis (LSA) (Landauer, 1998) LSA is a fully automatic mathematical/statistical technique for extracting and inferring relations of expected contextual usage of words in passages of discourse. LSA takes only raw text as input such as sentences or paragraphs, it does not utilise any humanly constructed dictionaries, knowledge bases, semantic networks, grammars, syntactic parsers, or morphologies.

LSA takes raw text as input parsed into words defined as unique character strings and separated into meaningful passages or samples such as sentences or paragraphs. Then LSA constructs a matrix, which has rows representing unique words, and columns representing passages. Each cell contains the frequency of occurrence the word of its row in the passage denoted by its column, and then each cell frequency is weighted by a function that expresses both the word’s importance in the particular passage and the degree to which the word type carries information in the domain of discourse in general.

In LSA, a sentence is represented in a very high-dimensional space with hundreds or thousands of dimensions (Landauer, 1998). This results in a very sparse sentence vector which is consequently computationally inefficient. High dimensionality and high sparsity can also lead to unacceptable performance in similarity computation (Li, et al., 2006)

(O’Shea, et al., 2008) compared between LSA and STASIS by using a dataset of 65 sentence pairs, a questionnaire was distributed among number of participants who were asked to rate “how similar the sentences are in meaning.” The rating scale ran from 0 (minimum similarity) to 4.0 (maximum similarity). Then the same dataset were calculated through LSA and STASIS

Both LSA and STASIS have performed well using the same dataset (O’Shea, et al., 2008), and the experiment showed that similarity judgements made using these algorithms are reasonable and consistent with human rating. LSA scored (0.838) correlation with human rating while STASIS scored (0.816).

Although LSA is able to capture and represent significant components of the lexical and passage meaning evinced in judgement and behaviour by humans, it does, however, lack important cognitive abilities that humans use to construct and apply knowledge from experience. (Landauer, 1998)

Unlike LSA, the STASIS method (covered in section 3.5.1) is based entirely on semantic networks (WordNet) to measure sentence similarity, where relations between words and synsets are identified based on human perspective. The researcher has found that the STASIS method is more suitable to develop semantic conversational agent, because it measures sentence similarity based on a knowledge base constructed from human's experience instead of depending on statistical approach to compute semantic similarity.

This chapter is focusing on semantic sentence similarity of text exchanged through dialogue between a human and a conversational agent based on word similarity and corpus statistics. In general the measurement is performed on the following stages:

- Word similarity: by measuring semantic similarity between all words within the short texts being compared.
- Sentence similarity: by measuring total sentence similarity based on the similarity scores between each pair of words in both short texts.

These stages are tightly coupled and it is hard to separate them, since word similarity is part of sentence similarity both will be referred as "semantic similarity".

Little attention was given to the Arabic language regarding word and sentence similarity, the only trial observed in Arabic was conducted by (Almarsoomi, et al., 2013), in which algorithm for measuring Arabic word semantic similarity using Arabic WordNet was developed.

3.2. Challenges of Sentence Semantic Similarity for the Arabic language

The challenges of using Arabic word and sentence semantic similarity in the application of conversational agents can be divided into three main categories: technical challenges related to speed and performance, challenges related to the Arabic language itself, and conceptual challenges related to the philosophy behind using semantic similarity methods in CAs.

3.2.1. Technical challenges

These include the challenges associated with integrating the CA with existing systems such as Arabic WordNet, these are described in details in section (3.3.7)

3.2.2. Linguistic Challenges

This type of challenges are related to the Arabic language and was already covered in section (2.5), these can be summarised as:

- The variations of Arabic dialect.
- The complexity of Arabic grammar.
- Arabic diacritics and morphological ambiguity.
- Word sense ambiguity.

3.2.3. Challenges Associated with Sentence Similarity Measurement

The third type of challenges is related to the similarity concept itself, this include:

- The variant meaning of similarity: words or sentences are not always similar in the same way. They might be highly similar in some domains and contexts and counter wise in other contexts or domain, in some contexts some details may not be critical as some other contexts. For example, if someone lost a passport and he is talking to a friend, the phrase “I have lost my passport” is highly similar to the phrase “I do not have a passport” since they both lead to the same fact that he does not have a passport now. But, if this person is talking to a police officer those two sentences are not at the same level of similarity.
- Function words: Arabic language like other languages contains function words; (like في (in) على(on) من (from) which often contain rich semantic information about the

sentence, yet they cannot be classified in the ontology or knowledge base as something that truly exists in real world. For example, the word "في" (in) is used to relate between an entity and a place, but the word itself cannot be classified as an object that truly exist. Good sentence similarity measurements in the author's opinion must consider function words as well.

- Negation: sentence similarity measurement does not deal with negated phrases properly, so a good similarity measurement must include a method that gives more consideration for negated phrases. For example, the two sentences "I want a new passport""اريد جواز جديد"and "I do not want a passport" " لا اريد جواز" contain highly similar words but one of the sentences is totally negates the other.
- Type of sentences: in general, sentences can be classified into informative, negative, Affirmative, and questionable sentences, each of which must be recognised before measuring similarity. For example, the sentence "Do I have to apply for a new passport?" "هل يجب ان اقدم للحصول على جواز جديد ؟" must not be similar to the sentence "I want to apply for a new passport" " اريد التقديم للحصول على جواز جديد". Sentence similarity measurement is unable to conclude facts from sentences, it wouldn't detect similarity between "I lost my passport" "فقدت جوازي" and "I have no passport" "لا املك جواز". Although they are not similar but they still share the same fact that the person does not have a passport now.
- The compound nature of Arabic words: Arabic words are usually rich of semantic information due to the affixes added to Arabic words. These affixes contain rich information about tense, plural, dual, and singular forms, and other information about the sentence. For example the Arabic word "يكتبون"which means (they are writing), the word indicates a plural masculine, and a tense in which the act of writing occurs is present in this case.

3.3. WordNet

WordNet is a large lexical database of English nouns, verbs, adjectives, and adverbs grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual relations (Princeton University, 2005).

WordNet has been widely used as a rich linguistic resource in fields of semantic similarity. (Datamuse, 2003)(Fong, 2003) (Pedersen, 2007)(Alcock, 2004). The relation between words in the WordNet lexical hierarchy provides a valuable source of information for calculating semantic similarity.

The WordNet project started in the Princeton University Department of Psychology, by George A. Miller in the mid-1980s, to provide a tool to organise lexical information in terms of word meanings, rather than word forms, providing an alternative to classic dictionaries that group words according to their meaning regardless of their semantic. Therefore, WordNet resembles a thesaurus more than a dictionary (Miller, et al., 1990).

Most of the methods used to measure similarity described in this thesis use WordNet as information source to evaluate word and sentence similarity, therefore an overview of WordNet structure and semantic relations is covered in the following sections.

According to (Elkateb, et al., 2006), Arabic WordNet AWN (BLACK, et al., 2006) was constructed according to the same methods developed for Euro WordNet (Vossen, 1997). Euro WordNet is a multilingual database with WordNets for several European languages, it is structured in the same way as the English WordNet except that the synsets of supported languages are linked to an Inter-Lingual-Index based on English WordNet, the languages are interconnected so that it is possible to go from the words in one language to similar words in any other. The Euro WordNet approach maximises compatibility across WordNets

Since all WordNets including the Arabic, English and Euro WordNet have the same hierarchical structure; specific concepts can be linked and translated with great accuracy by following a top-down procedure. Base abstract concepts are defined and extended via

Hyponymy relations to derive a core WordNet and the set of more specific concepts are encoded as synsets, the concepts are ontology terms which represent classes such as “Human” and “Animal”, other language-specific concepts are translated manually to the closest synset in Arabic. The same step is performed for all English synsets that currently have an equivalence relation in SUMO ontology(Vanderhulst, 2005) which is the knowledge base used by WordNet. The SUMO ontology is discussed in section (3.3.5).

Arabic WordNet uses the same ontology base concepts as the English WordNet. However, AWN needs more effort to add more Arabic words and structure. At the time of writing this thesis the number of Arabic words did not exceed 24,000 words (The Global WordNet Association, 2014) compared to more than 150,000 words in English according to English WordNet statistics (Princeton University, 2014). In addition to that, Arabic WordNet does not have rich API’s (Application Programming Interfaces). Therefore, it has limited applications compared to the English WordNet. These applications include Question Answering (Abouenour, et al.), NLP (Rodríguez, et al., 2008), semantic web annotation (Bin Saleh, et al., 2009) and search engines (Al Ameed, et al., 2006).

3.3.1. Semantic Relations

The main relation among words in WordNet is synonymy (like the relation between the words shut and close). Words that have the same concept and are interchangeable in many contexts are grouped into unordered sets (synsets).

In WordNet, a synset is linked to another synsets by a number of “conceptual relations”. Additionally, each synset contains a brief definition (“gloss”) and one or more short sentences demonstrating the use of the synset members.

WorldNet’s conceptual relations between synsets can be summarised as:

- Hyponymy or (is-a) relation which is the most frequently encoded relation among synsets. It links more general synsets like “حيوان” “Animal” to increasingly specific ones like “ثديات” “Mammal” and “طيور” “Birds”. Thus, WordNet states that the category “حيوان”“Animal” includes “ثدييات”“Mammal” which in turn includes “ ثديات مائية”“Aquatic Mammal”. Conversely, concepts like“الثديات المائية”“Aquatic Mammal” and “ثدييات” “Mammal” make up the category “حيوان”“Animal”.

- Meronymy: the part-whole relation holds between synsets like “chair” and “مسند”“backrest”, “مقعد” “seat” and “ساق” “leg”. Parts are inherited from their super ordinates. If a chair has legs, then an armchair has legs as well.
  Parts are not inherited “upward” as they may be characteristic only of specific kinds of things rather than the class as a whole: chairs and kinds of chairs have legs, but not all kinds of furniture have legs.
- Antonym: is an opposite relation between two synsets like “سريع” “fast” and “بطيء” “slow”, “طويل” “tall” and “قصير” “short”

3.3.2. Part of Speech

The words covered in WordNet can be classified into three categories:

- Nouns
- Verbs
- Adjectives and adverbs

3.3.2.1. Nouns

The most obvious relations between nouns in WordNet is “Synonymy” and “Hyponymy”, nouns such as people’s names, cities, countries, species and other entities are organized into a tree hierarchy. For example, the term “قطة” “Cat” is a “ثدييات” “Mammal”; and“ثدييات” “Mammal” is a subordinate of“حيوان” “Animal”. All noun hierarchies eventually go up the root node “entity”. There are some nouns that might be synonyms as well such as “لهب” “flame” and “نار” “fire”.

WordNet distinguishes among types (common nouns) and instances (specific persons, countries and geographic entities). Thus, armchair is a type of chair, but the cat’s name

“Garfield” is an instance of a “Cat”. Instances are always leaf (terminal) nodes in their hierarchies

3.3.2.2. Verbs

Verbs are the most important lexical and syntactic category of a language. All English sentences must contain at least one verb (Fellbaum, 1990).

Verb synsets are arranged into hierarchies, verbs towards the bottom of the trees express increasingly specific manner, as in “يتواصل” “communicate”, “يتكلم” ”talk” and “يهمس” “whisper”. The specific manner expressed depends on the semantic criteria, such as volume in the above example that is just one dimension along which verbs can be elaborated.

(Fellbaum, 1990) also stated that the sentence frame used to test hyponymy between nouns, is not suitable for verbs. For example people might be familiar with the sentence “الكلب هو حيوان” “a dog is an animal” but they are likely to reject such statements as“ الركض هو حركة” “jogging is moving” or “ الهمس هو كلام”“whispering is talking”. The semantic distinction between two verbs is different from the features that distinguish two nouns in a “hyponymy” relation.

3.3.2.3. Adjectives and adverbs

According to (Fellbaum, et al., 1993), WordNet divides adjectives into two major classes: descriptive and relational. Descriptive adjectives are often bipolar attributes and consequently are organised in terms of binary; opposite in meaning (antonym), and similar in meaning (synonym).

Adjectives are organised in terms of antonyms: pairs of “direct” antonyms like “رطب-جاف” “wet-dry” and “شاب-مسن” young-old reflect the strong semantic contrast of their members. Each of these adjectives in turn is linked to a number of “semantically similar” adjectives. For example, dry is linked to parch. (Princeton University, 2005).

Relational adjectives are assumed to be variants of modifying nouns and so are cross-referenced to the nouns for such as colour adjectives.

There are only few adverbs in WordNet (hardly, mostly, really, etc.) as the majority of the English adverbs are straightforwardly derived from adjectives via morphological affixation (like surprisingly, strangely, etc.)

3.3.3. Database Structure

According to (BLACK, et al., 2006) the database structure of the Arabic WordNet comprises of four categories, they are:

- Items; which are conceptual entities, including synsets, ontology classes, and instances. Each item has a unique identifier, and descriptive information.
- Word entity, or word sense: each word is associated with an item via an identifier
- A form: it is a special form that is considered dictionary information (not an inflectional variant) such as the broken plural form.
- A link; which represents conceptual relation relates two items, and has a type such as "Synonym" or "Hyponym". Links connect synset items to other synset items

3.3.4. Morphological Analysis

Morphology is concerned with lexical relations between word forms. Morphological analysis is crucial in WordNet. For example, if someone looks up the word "books" in WordNet, WordNet won't be able to find the word with some type of morphological analysis since it has only the word "book" stored in its database. Therefore a program is needed to strip off the plural suffix to and then to look up the root of word in lexical database.

The following sections cover the details of Arabic morphology and AraMorph. In Arabic WordNet only the root of words is stored in the lexical database. Therefore, it Is important to run or implement morphological analysis of words to derive their roots and isolate their affixes before performing semantic similarity measurement between them, compound words will not be found in the lexical database and therefor semantic similarity would fail to give any result.

Morphological analysis is also important to detect the part-of-speech categorisation of words (noun, verb, adverb etc.) which has an important role in semantic similarity measurement.

In this research AraMorph (Brihaye, 2003) is used as a tool for morphological analysis, AraMorph is explained in details in the next section.

3.3.4.1. Arabic Morphology (AraMorph)

According to (Brihaye, 2003) AraMorph is a tool written in java used to perform Arabic morphology analysis and part of speech tagging. It is based on Backwater’s transliteration system (Habash, et al., 2007), which is a method of transforming Arabic letters into Latin letters and vice versa. Table (3-1) demonstrates how Arabic letters are translated to Latin letters.

Symbol	Arabic letter	Symbol	Arabic letter
'	HAMZA (ء)	_	TATWEEL (~(
|	ALEF WITH MADDA ABOVE(آ)	F	FEH (ف)
>	ALEF WITH HAMZA ABOVE (أ)	Q	QAF (ق)
&	WAW WITH HAMZA ABOVE (ؤ)	K	KAF(ك)
<	ALEF WITH HAMZA BELOW (إ)	L	LAM (ل)
}	YEH WITH HAMZA ABOVE (ئ)	M	MEEM (م)
A	ALEF (ا)	N	NOON (ن)
B	BEH (ب)	H	HEH(ه)
P	TEH MARBUTA (ت)	W	WAW(و)
T	THE(ذ)	Y	ALEF MAKSURA (ى)
V	THEH (ث)	Y	YEH (Y(
J	JEEM (ج)	F	FATHATAN (ً)
H	HAH (ح)	N	DAMMATAN (ٌ)
X	KHAH (خ)	K	KASRATAN(ٍ)
D	DAL (د)	A	FATHA (َ)
*	THAL (ذ)	U	DAMMA(ُ)
R	REH (ر)	I	KASRA (ِ)
Z	ZAIN(ز)	~	SHADDA (ّ)
S	SEEN (س)	O	SUKUN (ْ)

$	SHEEN (ش)	`	SUPERSCRIPT ALEF
S	SAD (ص)	{	ALEF WASLA
D	DAD(ض)	P	PEH
T	TAH (ط)	J	TCHEH
Z	ZAH (ظ)	V	VEH
E	AIN (ع)	G	GAF
G	GHAIN (غ)		

Table 3-1 Buckwalter transliteration (Buckwalter, 2002)

AraMorph performs Morphological analysis for Arabic words in the steps below:

1. Arabic words are converted to Latin characters based on the transliteration table (3-1)
2. AraMorph uses an algorithm developed by (Buckwalter, 2002) to decompose the word in a sequence of possible prefix, stem, and suffix.
3. AraMorph checks the presence of each element in three dictionaries:
    - The prefix dictionary
    - The stem dictionary
    - The suffix dictionary
4. AraMorph grabs the morphological information for each element. If applicable, AraMorph then checks if the morphologies of each element are compatible between each other by looking-up three tables containing valid combinations:
    - Between the prefix and the stem.
    - Between the prefix and the suffix.
    - Between the stem and the suffix.

For example, using AraMorph to process of the Arabic verb (yEmlwn, يعملون), AraMorph extracts the root (عمل,Eml) and prefix (y) (ي) which refers to a third person, and the suffix (wn) (ون) which indicates a plural masculine suffix.

Morphological analysis is essential in processing word similarity, because WordNet keeps only the root of each word in the lexical database. The semantic similarity is measured between the roots of words regardless of their morphological affixes.

The types of Arabic morphological categories are discussed in the next section.

Arabic Morphological Categories

According to (AraMorph, 2003), each Arabic stem is assigned a morphological category using a form of mnemonic notation (N, Ndu, NduAt, Nprop, PV, IV, FW, FW-Wa, FW-WaBi, etc.). These notations denote both the basic part of speech classification (Noun, Verb, or Function Word) and the set of prefixes and suffixes that can be attached to that stem; Morphological categories can be highlighted as:

- Function Word stems
- Noun stems
- Verb stems

More details about the morphological categories can be found on (AraMorph, 2003)

3.3.5. Suggested Upper Merged Ontology (SUMO)

According to (Gruber, 2008) ontology defines a set of representational primitives used to model a domain of knowledge or discourse. Ontologies are typically written in declarative languages to define levels of abstraction rather than data structures and implementation strategies, these languages are powerful to express concepts unlike the languages used for procedural programming.

In the context of semantic similarity, the presence of ontology is essential to serve as a knowledge base for measuring semantic similarity based on the relations defined between ontology subclasses and concepts.

The Suggested Upper Merged Ontology (SUMO) (Vanderhulst, 2005) and its domain ontologies form the largest formal public ontology in existence today. They are used for research and applications in search, linguistics and reasoning. Figure (3-1) shows a portion of SUMO ontology taxonomies. Detailed explanation about SUMO can be found in section (6.2.2)

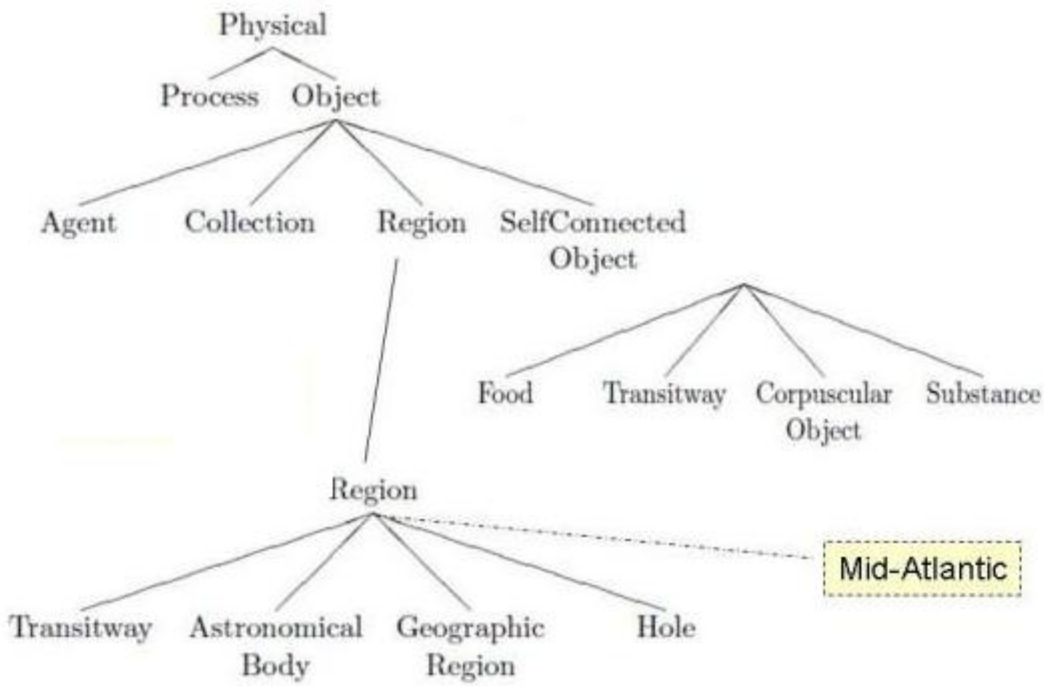


Figure 3-1 the Suggested Upper Merged Ontology (Pease, 2011)

SUMO is the only formal ontology that has been mapped to the entire lexicon of all WordNets. It is written in the SUO-KIF language (Standard Upper Ontology Knowledge Interchange Format) and it is free and owned by the IEEE. The ontologies that extend SUMO are available under the General Public License (Free Software Foundation, 2007).

SUO-KIF language

Standard Upper Ontology Knowledge Interchange Format (SUO-KIF) (Pease, 2009) is a language designed for use in the authoring and interchange of knowledge. SUO-KIF is also logically comprehensive at its most general, it provides for the expression of arbitrary logical sentences. In this way, it differs from relational database languages (like SQL) and logic programming languages (like Prolog).

SUO-KIF combines terms into statements of facts, for example: "The 82nd Airborne is a military unit", which would be stated in SUO-KIF as (instance The82ndAirborne MilitaryUnit) and "The class of all Person(s) is a subclass of the class of all animals" is expressed in SUO-KIF as (subclass Person Animal)

SUO-KIF also support logical relations between statements such as "And", "or"; and also supports implications and other logical functions, more details about SUO-KIF can be found at (Nolt, et al., 2011)

It's important to have a tool to edit the WordNet ontology to add new terms or modify the existing terms and relations between WordNet synsets. Today there are standard languages and a variety of commercial and open source tools for creating and working with ontologies such as (protégé, 2014) and SIGMA (Pease, et al., 2013).

The SIGMA knowledge engineering environment (Pease, et al., 2013)is a system for developing, viewing and debugging theories in first order logic. It works with Knowledge Interchange Format (KIF) and is optimised for the Suggested Upper Merged Ontology (SUMO).

SIGMA includes a number of useful features for knowledge engineering, including term and hierarchy browsing, the ability to load different files of logical theories, a full first order inference capability with structured proof results, a natural language paraphrase capability for logical axioms, support for displaying mappings to the WordNet lexicon and a number of knowledge base diagnostics.

The only one disadvantage of SIGMA that it was not designed for editing the ontology, the ontology has to be modified directly in a text file which requires expertise in SUO-KIF language.

(protégé, 2014) protégé is a free, open-source ontology editor and framework for building intelligent systems, it has a simple customisable user interface, and provides a graphic representation for ontology. Until this research is conducted, protégé does not support the "KIF" format which is used by the Arabic WordNet ontology.

3.3.6. AWN browser

The Arab WordNet (AWN) browser (The Global WordNet Association, 2014) is a combination of tools written in Java to browse the Arabic WordNet. AWN browser uses AraMorph (Brihaye, 2003) as morphological analysis to decompose Arabic words and isolate their stems and affixes. The AWN browser also has modules used to lookup the lexical database and SUMO ontology where users can either lookup an Arabic word or they can look up an ontology term, AWN browser also provides an instant translation between Arabic words and English words.

As illustrated in figure (3-2), for example when a user looks up the word (يعملون) (They are working), AWN first performs morphological analysis using AraMorph to decompose the word, then the AWN browser looks up the word in the Arabic lexical hierarchy and provides a graphical view for its position. In addition the AWN browser finds the equivalent English word which is (work) in this case; based on the semantic position in the English lexical hierarchy.

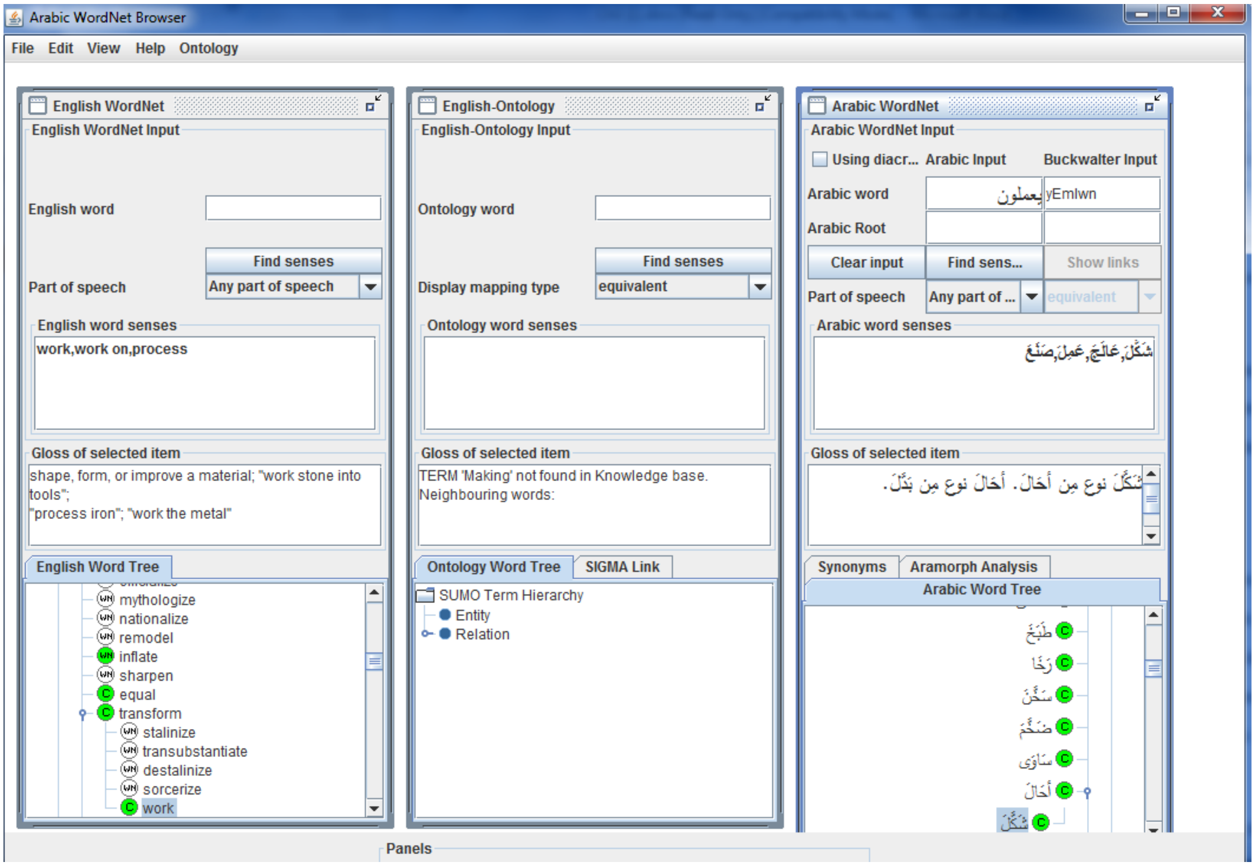


Figure 3-2AWN browser

3.3.7. Problems and Challenges Associated with Arabic WordNet

- Incompleteness: as discussed earlier in section (3.3) Arabic words added to the AWN do not exceed 24,000 words (The Global WordNet Association, 2014), this represent less than 10% of the total Arabic stems. Therefore, when developing Arabic semantic conversational agents, AWN must be expanded to include all Arabic words.
- Lack of tools: The AWN browser is designed for browsing purposes only; it does not have any functionality to modify the lexical database. Therefore, it's not possible to add new words through the AWN browser. Although the database of Arabic words is available in XML format, it is up to the researchers to adapt or reformat it according to their needs.

  In addition, some domains may require modification to the ontology to add new entities or relations. The AWN browser does not have this functionality, other tools such as SIGMA discussed in section (3.3.5) is also designed for browsing purposes, on the other hand protégé has a very simple interface to create and modify ontologies with graphical representation but it does not support the “KIF” format used by Arabic WordNet.

- Similarity measurement: AWN browser was not designed to be used in measuring word or sentence similarity. Although the AWN browser source code is publically available, there is not sufficient software documentation to enable researchers to reuse AWN software. This increases the effort needed by researchers to reuse or modify the source code.

To overcome these challenges, the research presented in this thesis developed a new tool to manage the lexical hierarchy and ontology concepts. The ontology of WordNet was copied to the new tree, and Arabic words were inserted in their appropriate places. This tool is described and discussed with further detail in chapter (6).

3.4. Word Semantic Similarity

According to Liu (Liu, et al., 2007) the similarity between two concepts is identified by humans through comparing their common and different attributes. These attributes are used to derive equations used to measure word and sentence semantic similarity.

Word Similarity can be defined as the measurement of semantic relatedness between two words based on the attributes they share, these attributes may include lexical attributes such as part-of-speech, tense, and numeral; or semantic attributes such as “part of” and “instance of” which are defined by the ontology.

As explained in section (3.3), in WordNet, words are organised into synonym sets (synsets) these synsets are linked logically through (IS-A) relation creating a hierarchical structure.

One method for measuring word similarity is the edge-counting based method introduced by (Rada, et al., 1989) which finds the minimum path length between two words (Rada, et al., 1989) applied this method to a medical domain, and found that the path length function simulated well human assessments of conceptual similarity.

However, (Richardson, et al., 1998) had concerns that this measure was less accurate than expected when applied to a comparatively broad domain (e.g. WordNet taxonomy). They found that irregular densities of links between concepts resulted in an unexpected conceptual distance outcomes.

Resnik's measure (Resnik, 1995) introduced an information content method for semantic similarity measurement; it was the first to combine the use of ontology and a corpus for ontology concept similarity measurement. The concept can be a node in ontology such as an entity or relation using the below equations:

$$sim(c1, c2) \underset{c \ \in \ (c1, c2)}{Max} [-\log P(c)] \tag{3-1}$$

$$P(c) = \frac{freq(c)}{N} \tag{3-2}$$

Where sim(c1, c2) is the set of concepts that subsume both concepts c1 and c2; and P(c) is the probability of encountering an instance of concept (c). N is the total number of nouns in corpus.

(Jiang, et al., 1997) Conducted a comparative study between the edge-based method and the information content method, according to (Jiang, et al., 1997), the distance measure is highly dependent upon the subjectively pre-defined network hierarchy.

Since the original purpose of the design of the WordNet was not for similarity computation purpose, some local network layer constructions may not be suitable for the direct distance manipulation.

(Jiang, et al., 1997) also stated that the information content method requires less information on the detailed structure of taxonomy, but it is still dependent on the skeleton structure of the taxonomy.

Therefore (Jiang, et al., 1997) presented a hybrid method on the basis of the edge-based notion through adding the information content as a decision factor. (Jiang, et al., 1997) Included link strength and link weight factor which is calculated based on local density, node depth, and link type.

According to (Lin, 1998) previous similarity measures such as edge-count based method (Rada, et al., 1989) are tied to a particular application or assumes a particular domain model. For example, the method introduced by (Rada, et al., 1989) assume that the domain is represented in a network. If a collection of words is not present in the network, the edge-based measures do not apply.

(Lin, 1998) Proposed a new formula derived from information theory which combines information content of the compared words based on the argument that the similarity between two words is a ratio between the information need to express their commonality and the information needed to fully describe both of them:

$$sim(A,B) = \frac{\log P(common(A,B))}{\log P(description(A,B))} \tag{3-3}$$

For example, if A is an orange and B is an apple, the proposition that states the commonality between A and B is “fruit (A) and fruit (B)”. In information theory the information contained in a statement is measured by the negative logarithm of the probability of the statement. Therefore:

$$commonality(A,B) = -\log P(\text{fruit (A) and fruit (B)} \tag{3-4}$$

According to (Lin, 1998) description *(A,B)* is a proposition that describes what A and B are.

As an improvement to edge-based similarity methods (Leacock, et al., 1998) proposed a method for measuring the similarity between two concepts, taking into consideration the maximum depth of the noun taxonomy.

$$sim_{ab} = \max[-\log(\frac{Np}{2D})] \tag{3-5}$$

Where (Np) is the number of nodes in path (p) from concept (a) to concept (b) and D is the maximum depth of taxonomy.

(Li, et al., 2003) Included the attributes of path length (different attributes) and depth (common attributes) as a function to measure the semantic similarity between two words:

$$S(w1, w2) = f(f_1(l).f_2(h)) \qquad (3\text{-}6)$$

$$f_1(l) = e^{-\alpha l} \qquad (3\text{-}7)$$

$$f_2(h) = \frac{e^{\beta h} - e^{-\beta h}}{e^{\beta h} + e^{-\beta h}} \qquad (3\text{-}8)$$

Where (*l*) is the shortest path between two words in the lexical hierarchy of WordNet; (*h*) is the depth of the concept that subsumes the two words, (α) is a constant and (β) is a smoothing factor.

More recently (Liu, et al., 2013) Introduced Word similarity measurement using WordNet as improvement to the edge-based similarity method, the measurements included density, depth, and path length between concepts in WordNet lexical hierarchy. (Liu, et al., 2013) Stated that the greater the density of the lexical tree, the closer the distance between the nodes. Density can be measured by the number of subordinate nodes in a branch of lexical hierarchy. (Liu, et al., 2013) also stated that “The deeper the depth of the nodes located, the higher the similarity of them”, based on the fact that deeper concepts in the WordNet hierarchy contain more semantic information than higher concepts. This method of word similarity also considered the path length as an important factor in measuring word similarity. According to (Liu, et al., 2013) “The shorter path is contained within the longer path in a ‘is-a’ taxonomy, the concept nodes pair with shorter path between them has greater concept similarity than those with longer path between them”.

(Batet, et al., 2013) stated that “ similarity measurements based on path-based function provide absolute similarity values with non-comparable scales when they are obtained from different ontologies”, therefor they introduced a concept similarity measurement across multiple ontology, because path length would depend on the ontology size, depth and granularity.

$$sim(c1, c2) = -\log_2 \frac{|T(c1) \cup T(c2)| - |T(c1) \cap T(c2)|}{|T(c1) \cup T(c2)|} \tag{3-9}$$

Where c is a concept and T(C*i*) is defined as the set of super concepts of the concept (c).

(Tian, et al., 2014) also introduced a domain specific word similarity measurement, they developed a new metric for the software domain called (WordSimSE) to compute the similarity of two words by representing them as vectors and then compute the similarity between these two vectors. Each word is represented as a feature vector where each element in the vector is the co-occurrence weight of that word with other (contextual) word in the corpus. These contextual words serve as semantic anchors forming a basis to compare the semantic distance of two words. The co-occurrence weight is measured using a weighted positive point-wise mutual information (WPPMI).

As for the Arabic language, the researcher found that it has received relatively less effort in the field of word similarity measurement. (Almarsoomi, et al., 2013) Developed an Algorithm for Measuring Arabic Word Semantic Similarity (AWSS) based on Li's original work (Li, et al., 2006).

According to (Almarsoomi, et al., 2013) the depth of the concepts should also be taken into account when measuring semantic similarity between two words, because the concepts at upper levels of the lexical hierarchy have more general semantics and less similarity between them. This is done by measuring the depth of the concept that subsume the concepts containing the two words, this concept is known as Lowest Common Subsumer (LCS) as illustrated in the example below.

Figure (3-3) demonstrates a portion of AWN noun hierarchy. The shortest path length between (أب) father and (أم) mother is 2 and the concept (شخص) parent is called Lowest-Common Subsumer (LCS) for the words (أب) father and (أم) mother; while the shortest path between (جد) grandparent and (أب) father is 6. In this case, the word (أم) mother is more similar to (أب) father than (جد) grandparent is to (أب) father. Also in this figure, the shortest path length between (جد) grandparent and (تاجر عملة)“money handler” is 5, less than from (جد) grandparent to (أب)father, but it’s not possible to say that (جد)grandparent is more similar to (تاجر عملة)“money handler” than to father. This case illustrates the importance of the depth of LCS where the similarity of compared words grows higher if the depth of LCS increases as the lexical hierarchy goes deeper.

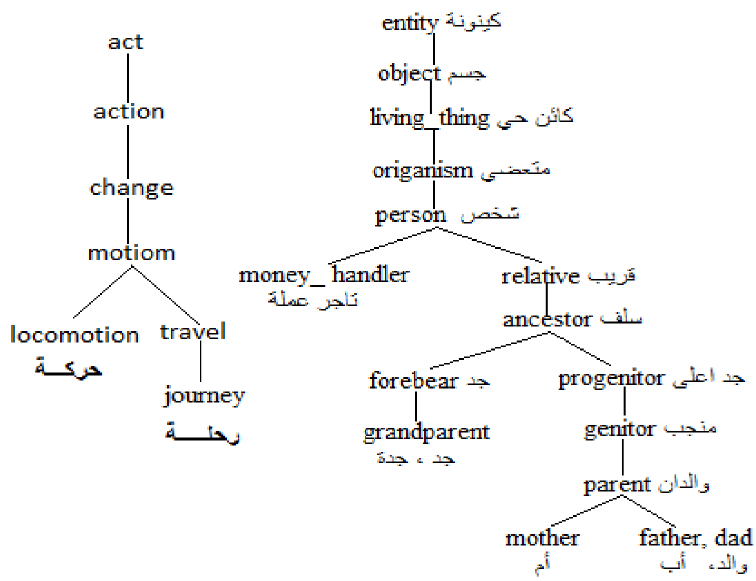


Figure 3-3 A portion of Arabic WordNet (Almarsoomi, et al., 2013)

(Almarsoomi, et al., 2013) also defined the semantic similarity between two words W1 and W2 as a function of the attributes path length and depth as follows:

$$\mathrm{S(W1,W2)} = F(f1(l), f2(d)) \qquad (3\text{-}10)$$

Where, (l) is the length of the shortest path between w1 and w2. (d) is the depth of the LCS of w1 and w2 in a lexical hierarchy. f1 and f2 are transfer functions of path and depth respectively.

For example, in Figure (3-3), أب father and والدdad are in the same concept, and length between them is 0. This case implies that the two words have the same meaning. So, f1 is set to be a monotonically decreasing function of l and is selected in exponential form to meet l constraints.

When d=0, there is no common attributes between the compared words and the similarity of s (w1, w2) = 0. As shown in Figure (3-3) , رحلة journey and أبfather are classified under separate substructure and no LCS subsumes the compared words, hence the similarity between them is 0. Furthermore and as shown in the example of جدgrandparent and تاجر عملةmoney-handler, the similarity grows higher if the depth of LCS of compared words increases in a lexical hierarchy. To meet this constraints, f2 is set to be increasing function of d.

The overall similarity is calculated using the following nonlinear formula:

$$sim(W1, W2) = e^{(-\alpha * l)} * tanh(\beta \, * \, d) \tag{3-11}$$

Where, α and β are the length and depth factors respectively which signify the contribution of the path length (*l)* which can be calculated using:

$$l \, = \, \mathrm{d1} + \mathrm{d2} - \, (2 * \mathrm{d}) \tag{3-12}$$

Where d1 and d2 are the depth of w1 and w2 respectively.

3.4.1. Challenges Associated with Word Similarity Measurements

There are several challenges associated with the development of word similarity measurements for the Arabic language, they can be highlighted as:

- Arabic grammar and morphology: Arabic words have much more affixes than English words, those affixes usually contain rich semantic information about the word. For example, the English sentence “they are writing” can be expressed in one Arabic word “يكتبون” this word is derived from the base verb “كتب” “write” with additional affixes to indicate plural, and present tense, this would increase the challenge of measuring word semantic similarity because Arabic words contain many affixes which are directly attached to the word and must be separated to obtain more accurate measurement.

- Flexibility of Arabic expressions: for example in the Arabic language a noun can be substituted with a verb without any change of meaning for example the sentence "أريد الذهاب الى السوق" and "أريد ان اذهب الى السوق"; both sentences mean "I want to go to the market" but one of them is phrased with the verb "اذهب" "go", while the other is rephrased with the noun "ذهاب". Some similar nouns and verbs might be located at different parts of the lexical hierarchy, which might change similarity measurement scores.
- Arabic diacritics: as explained in section (2.5.1), Arabic words include diacritics which are often used to disambiguate words and part-of-speech category. But, in modern Arabic these diacritics are usually omitted and it is up to the human reader to disambiguate the word according to the context. This is an important issue when measuring word similarity.

3.4.2. Evaluation of word similarity measures

The purpose of evaluating word similarity measurement is to calculate how close the machine rating (sentence similarity scores) is to human rating (sentence similarity according to human perspective).

In general, the evaluation process can be summarised in the steps below:

- Identifying a dataset of word pairs.
- Distribute the dataset among a number of qualified participants (e.g. native language speakers with reasonable age and different educational background).
- The same dataset is processed by the machine to compute semantic similarity.
- Measuring the correlation between human rating and machine rating.

(Li, et al., 2003) Evaluated variety of word similarity strategies to achieve a good similarity measure, for each of the proposed strategies, experiments were carried out with two steps:

- First, strategy parameters are tuned on the training data set (D_1). Given the value of a parameter, semantic similarity values of the word pairs are calculated. Then, the correlation coefficient between the computed semantic similarity values and the human ratings of Rubenstein-Goodenough is calculated. Thus, a set of correlation coefficients is obtained by changing the value of the strategy parameters. The parameters resulting in the greatest correlation coefficient are considered as the optimal parameters for that particular strategy.
- Second, the identified optimal parameters are used to calculate semantic similarity for word pairs in test data set (D_0). Again, the correlation coefficient between computed similarity values and human ratings of Rubenstein-Goodenough's is calculated for words pairs in (D_0). This correlation coefficient is used to judge the suitability of the particular strategy compared with other strategies and previously published results.

(Almarsoomi, et al., 2013) developed a new semantic measure and an Arabic data set to evaluate the new algorithm. To achieve that she conducted an experiment on Arabic word similarity measurements by comparing the results of word similarity measurements with human ratings. A benchmark of Arabic words created by (Almarsoomi, et al., 2012) is used in the evaluation process. The production of this dataset is divided into three major stages:

- Creating a List of Arabic Words (LAW). 27 Arabic categories were produced to cover different semantic themes and contain ordinary Arabic words. These categories were employed to generate a set of 56 stimulus Arabic words by selecting the first two words from each category.
- Constructing the set of Arabic word pairs, LAW was presented to 22 Arabic Native speakers from 5 Arabic countries to construct a set of word pairs covering the range of similarity of meaning (high to low). The participants were asked to create two lists

of word pairs which include high and medium similarity of meaning. The final set of Arabic word pairs contains 70 pairs of words which were selected using high and medium similarity word pair lists generated by participants plus the low similarity word pairs list selected randomly.

- Collecting the human ratings for the set of 70 word pairs: This experiment used a sample of 60 Arabic Native speakers from 7 Arabic countries who had not taken part in the first experiment. Each of 70 word pairs was printed on a separate card and those cards were presented to participants for rating how similar the word pair on each card was in meaning. The order of 70 cards was randomised before presentation. Each of 60 participants was requested to sort the 70 cards based on the similarity of meaning and rate them using scales which ranged from 0.0 (low similarity) to 4.0 (high similarity). Finally, each of the 70 Arabic word pairs was assigned a semantic similarity score calculated as the mean of the ratings provided by 60 Arabic native speakers.

The AWSS measure obtained a good value of Pearson correlation coefficient (r = 0.894) with the human judgments. The AWSS measure is performing well at (r= 0.894) with the average value of the correlations of human participants (r = 0.893). Furthermore, the performance of the Arabic word measure is substantially better than the worst human (lower bound) performance at (r=0.716).

The AWSS measure parameters (α and β) have been tuned using the training dataset to find the optimal values within the interval [0, 1]. In this experiment, the strongest correlation coefficient was obtained at α= 0.162 and β= 0.234.

One of the main disadvantages of AWSS evaluation is that it is limited to Arabic nouns, no attention or evaluation was given to verbs despite to their importance.

As mentioned earlier in section (3.1), semantic similarity measurement is performed in two stages, first is word similarity, and second is sentence similarity. The AWSS measure described in this section is used in this work for word similarity measurement, for both verbs and noun words.

3.5. Sentence Semantic Similarity

Sentence similarity can be defined as the level on which two sentences are related to each other.

There are several criteria that can be considered as attributes for sentence similarity including:

- Type of sentence: informative, negative, affirmative, and questionable.
- The tense in which the action in the sentence occurred (if applicable), and the involved participants.
- The part of speech categorisation of words in the sentence.
- The grammatical structure of the sentence.
- The semantic relatedness between words in the sentences based on lexical resources.
- The frequency in which the words of the sentences occurs in corpus.
- Facts that can be extracted from sentences

For example, consider these sentences

"I do not have a job" ***and*** *"I have a job interview tomorrow"*

If those two sentences are considered based on their sentence type, they are totally different because the first sentence is negated and the second is informative. But, when considering the tense of the sentences, the first one indicates a fact about the present; while the other indicates something about the future. Both sentences also contain the same entity that performed the act (human) in this case, which gives some similarity. If words part of speech categorisation for both sentences is also considered, some level of similarity will be found, the grammatical structure of both sentences is also close. Considering the semantic similarity between individual words also leads to different levels of similarity, the logical significance of each word in the sentence may also give insights about similarity; since not all words have the same amount of information.

Finally, if facts extracted from the sentences are considered, different levels of similarity will be found. Therefore, the real challenge is to find a similarity measure that best fit with the Arabic conversational agent.

3.5.1. Sentence Similarity Based on Semantic Nets and Corpus Statistics (STASIS)

(Li, et al., 2006) Introduced sentence similarity based on semantic nets and corpus statistics to measure sentence similarity, this method combined path length and depth in lexical hierarchy of WordNet, it also includes other factors such as word frequency in corpus and word order similarity.

In general the STASIS method (Li, et al., 2006) can be summarised in the following steps:

1- Identify the joint word set of two short texts; which includes all unique words from the two sentences.
2- Each sentence is evaluated separately with the word set.
3- A matrix is formed by measuring the similarity of word pairs of the sentence and the word set.
4- The corpus frequency of the similar pairs is also included in the calculation.
5- The result of the matrices is evaluated in a function to calculate the overall similarity
6- The word order similarity of both sentences is calculated separately, and then it is combined in a function with the overall similarity to calculate the total similarity

The upcoming sections will discuss the details of the STASIS except the word order similarity which is not included in this thesis due the flexible structuring of Arabic language.

3.5.2. Challenges Associated with STASIS When Using Arabic Language

Although the STASIS method achieved outstanding evaluation results (Li, et al., 2006), there are several challenges associated with using (STASIS) and other sentence similarity measurements in general, these can be summarised as:

- Similarity is not the same thing as meaning, sometimes there is similarity but the meaning is very different. STASIS and other sentence similarity measures in general focus on sentence similarity instead of sentence meaning. For example the sentences "I'm looking for a house" and the sentence "look at that beautiful house". Both sentences are similar but they mean two different things.
- In the application of conversational agent, there is no standard semantic similarity threshold that can be applied to all utterances, some utterances have much information and require strong similarity, while others contain less information. An example the sentences "I've lost my passport and I need to go to Baghdad soon", and the sentence "I've lost my passport, and I need another one". Both sentences have the same meaning, but the first one contains more information, therefore it is difficult to set a standard threshold for utterances exchanged between users and CA.
- Similarity measurements do not deal with different types of sentences (informative, negative and questionable), therefore it does not deal with facts extracted from utterance, it only measures how the words in utterance are close to the stored utterance in the agent. For example STASIS does not include a method to distinguish between questionable and informative utterance.
- There are many linguistic problems associated with sentence similarity, such as word sense ambiguity, and part-of speech tagging. For example a sentence can be rephrased to other sentences with the exact meaning but with nouns instead of verbs. Nouns and verbs might be located at different places in the lexical hierarchy and the connecting path might change when replacing a noun with a verb or vice versa (as explained in section 3.5.1) which leads to different similarity score. In such cases sentence similarity would fail to give accurate results.

- A sentence similarity measure does not provide any reasoning of the problem; instead it only measures how close the sentences are based on the words of each sentence. Therefore this method is not expected to extract facts from utterance.
- Sentence similarity does not consider grammar: therefore a non-logical sentence would be treated the same way as a logical sentence with the same words. Although there is a word order similarity measures (Li, et al., 2006) which considers word order in similarity measurements, it is not applicable to the Arabic language due to the flexible structuring of the sentence.

3.6. Evaluation of Semantic Sentence similarity

Jim O'Shea (O'shea, et al., 2013) described three methods of evaluating sentence similarity:

- Systems-Level Evaluation in dialog systems: in which the similarity measure could be evaluated through the performance of a system in which it is used.
- Indirect Measurement Using IR (information retrieval) Techniques: these measures require a corpus; Pairs of texts from the corpus are already rated as paraphrase and non-paraphrase by human judges. The same texts are classified by the semantic similarity algorithm. A high similarity rating is interpreted as a paraphrase whereas low similarity means non-paraphrase.
- Specifically Designed Methodology: by using a benchmark dataset of sentence pairs with similarity values derived from human judgment. The performance of the similarity measurement algorithm is evaluated using its correlation (usually Pearson's product-moment correlation coefficient) with the human ratings.

Li (Li, et al., 2006) evaluated the STASIS similarity measure by collecting human ratings for pairs of sentences. The participants consisted of (32) volunteers, all native speakers of English educated to graduate level or above. The participants were asked to complete a questionnaire, rating the similarity of meaning of the sentence pairs on the scale from 0.0 (minimum similarity) to 4.0 (maximum similarity). This measure achieved a reasonably good Pearson correlation coefficient off 0.816 with the human ratings.

3.7. Summary

This chapter defined Semantic Similarity (words and sentences). WordNet and Arabic WordNet (AWN) were introduced and explained briefly. Morphological analysis, ontology used in WordNet and the semantic relations was also discussed in some details. Some concentration was given to the Arabic Morphology (AraMorph), SUMO and SUO-KIF. The methods for measuring similarity between words and sentences were also discussed with some examples.

Challenges of sentence similarity for the Arabic language and weaknesses associated with using semantic similarity method in conversational agents were also covered and discussed with some further comments on how to overcome these weaknesses which can be highlighted as below:

- Incompleteness of Arabic WordNet and lack of tools: to overcome this challenge a new tool was developed to manage the lexical hierarchy and ontology concepts, the ontology of WordNet was copied into the new scripted tree and Arabic words are inserted in their appropriate places. Further elaboration on this tool is given in chapter (6).
- Word sense disambiguation: although there are many methods developed for WSD (Zouaghi, et al., 2011) (Agirreand, et al., 2009) (M., et al., 2012) (Liu, et al., 2007) but using them would cause more time complexity. In addition to that, adding one of these methods to the system would make it hard for the researcher to evaluate sentence similarity measurement because the result of evaluation would reflect the performance of sentence similarity and WSD method. During the experiment in this thesis, only Arabic words related to the domain were added to the lexical hierarchy, this would eliminate the need for WSD during this experiment.
- Lack of research on Arabic word and sentence similarity: to the best of the researcher knowledge, the only effort in this field was made by (Almarsoomi, et al., 2013) and it only covered Arabic nouns. There is a lack of research on the field of Arabic verb similarity and the semantic information contained within verbs. AWSS

measurement (Almarsoomi, et al., 2013) will be used during the course of this work to measure the similarity between nouns and verbs as well.

Chapter 4
Arabic Conversational Agents: Architecture and Scripting Language

4.1. Introduction

The main goal of this research is to develop a novel Arabic semantic conversational agent to overcome difficulties found when applying other types of CAs. But, it is not possible to construct a semantic CA without having a reliable design and architecture to insure its smoothness and viability. Also, once this CA is completed, it should be tested, evaluated and compared to a well-known and successful type of CAs.

ArabChat (Hajjawi 2011) was the only true trial of the Arabic conversational agents found. Although this CA was successful with its pilot application domain, and had rich scripting features, the researcher found that it suffers some drawbacks like irresponsiveness, and complexity associated with managing the conversational agents, and dialogue flow. It also lacks any information structure to the domain.

Therefore, and in an attempt to improve features of the ArabChat regarding dialogue flow, speed, and usability, an architecture was designed and tested using pattern matching conversational agent (PMGO-CA), this architecture was later used to construct the Arabic semantic conversational agent.

Following to the development of PMGO-CA in this chapter and its evaluation (Chapter 5), a modified version of the architecture is used to develop a semantic goal oriented conversational agent (SGO-CA) which is covered in chapter (6). Using the same methodology to construct a pattern matching CA and semantic CA makes it easier for the researcher to conduct a fair comparative study between the performance of pattern matching approach and semantic similarity approach in CAs

The main features of the novel unified architecture for Arabic goal-oriented conversational agents introduced in this chapter and used for both pattern matching goal-oriented conversational agent (PMGO-CA); and the semantic goal oriented conversational agent (SGO-CA) covered in chapter six can be highlighted as:

1- Dialogue flow control: the new architecture provides control over dialogue flow and consistency through the use of knowledge trees to control conversations and track contexts, this makes dialogue questions and answers more organised, details about dialogue flow are covered in section (4.3.3.1)
2- Increased speed: structuring domain rules as tree nodes reduces the number of patterns to be evaluated against user utterances. In this case, patterns of the current context are evaluated first, if no match is found, PMGO-CA searches other contexts for a match. this makes the agent more efficient, details about context-switching can be found in section (4.3.3.2)
3- Usability: the new architecture and software tools were developed and optimised for usability, all software tools contain friendly interfaces with self-explanatory options, making the agent easier to script, implement, and maintain.
4- Adaptability: the use of knowledge trees has significantly contributed to make the agent adaptable for other domains, simply by replacing the knowledge tree file with another knowledge tree of other domain.
5- Memory: PMGO-CA asks users a set of questions at the start of each conversation, these questions are related to users such as name, age and current location, this information are used to identify users when they converse with PMGO-CA again. The questions are customisable by the PMGO-CA scripter. More details about memory are covered in sections (4.3.4) and (4.3.5).

   In addition to user's information, PMGO-CA keeps a record of the fired rules (fired rules are rules used throughout the dialogue to generate a response to user's utterance) and store them with users information in a database to be used in future conversations with the same users.

PMGO-CA built in this research offers the following improvements over the ArabChat reviewed in section (2.5):

- The use of knowledge trees to script the domain.
- More speed in processing users' utterance
- Easier housekeeping for the CA, in terms of usability and user-friendly interfaces
- PMGO-CA tackles long term memory issues in CAs.

This chapter describes the following novel contributions:

- The methodology of developing PMGO-CA.
- Domain knowledge engineering and transformation.
- The architecture of PMGO-CA.
- The knowledge tree of the knowledge domain.
- Pattern matching algorithm used to match users utterances
- Mechanisms used to traverse the knowledge tree, in order to respond to users utterances
- Memory management in PMGO-CA.
- Software tools used to construct PMGO-CA.

4.1.1. The Methodology for Developing New Arabic Goal-Oriented Conversational Agent (PMGO-CA):

The development of an Arabic goal-oriented conversational agent and associated scripting language comprised of the following stages:

1- Knowledge engineering: this is a process of gathering all information about the domain, modelling them to create a knowledge representation.
2- PMGO-CA Architecture design and implementation to support the modelled information.

3- Implementation: the development of software tools and a new scripting language which takes into consideration challenges of the Arabic language.

4- Evaluation of the new PMGO-CA and scripting language (covered in chapter 5)

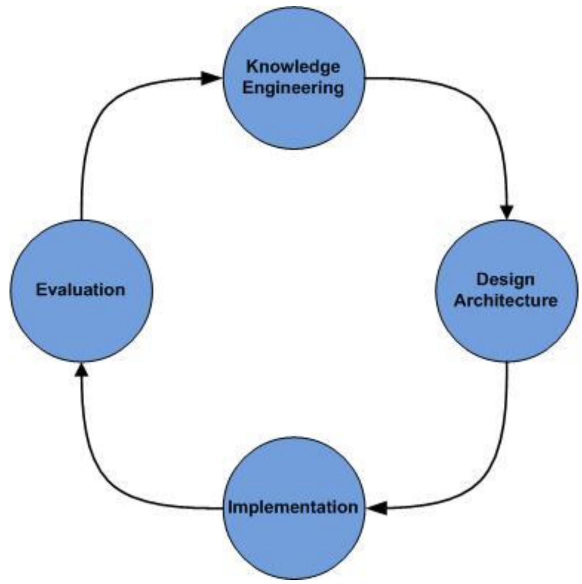


Figure 4-1 GO-CA phases of development

As shown in figure (4-1), these stages were iterative because there are many issues to be discovered in the PMGO-CA's performance during the evaluation phase which leads to more changes to the agent. Some of these issues required modification for search algorithms and context-switching (covered in sections 4.3.3.1 and 4.3.3.2 respectively), these changes lead to additional modification to the architecture and software code modification of the PMGO-CA

Other issues were discovered during testing, this required modifications to the knowledge representation and resulted to knowledge tree modification and patterns re-scripting.

4.2. Knowledge Engineering of the Domain

After selecting a domain of interest, the knowledge engineering process (Trappey, 2006) begins by gathering information about the domain from knowledge sources, these sources include stakeholders, domain experts, books, manuals, regulations, guides, and any other formal documents or work procedures.

After gathering the raw information, the analysis process begins by analysing each piece of information and formalise them in a consistent manner; then revise this refined information with domain experts to clarify any missing or ambiguous information.

Then a high level representation for this knowledge is established, typically a flow chart, a knowledge tree, or a graph; the representational model differs based on the domain type and target application and users. The higher level representation must also be revised and checked by domain experts and stakeholders.

The process of knowledge engineering used can be highlighted in these steps:

1- Gathering information about the domain, including all laws, work procedures, regulations and list of FAQs.
2- Identifying the processes of the domain and formalizing them into process charts.
3- Reviewing these process charts with domain experts.
4- Transforming the process charts into a flow charts.
5- Converting the flowcharts to knowledge trees.

4.2.1. Iraqi Passport Domain and Knowledge Sources

The Iraqi Passport Services (IPS) was chosen as the domain of knowledge to develop an Arabic conversational agent for. It is well known that the passport is one of the documents used to prove the identity of an individual. It becomes the only important document to prove the citizenship and identity when used outside the borders and territory of the native country.

Iraqi citizens, especially immigrants, experienced a large number of problems due to frequent changes in Iraqi passports after 2003. The different types of passport forms and the procedure to apply for new ones were very confusing. This coincided with the changes in the citizenship and passport laws. This resulted in long delays and queues at the Iraqi missions abroad when applying for passports or inquiring about passport issues.

To make life easier for Iraqi immigrants and those living abroad, and in an attempt to answer their queries and questions in an efficient way, an Arabic Pattern Matching Goal-Oriented Conversational Agent PMGO-CA was constructed to offer online service.

PMGO-CA can access, interpret and discuss the correct and updated information about the Iraqi Passports, and reply to user enquiries in a natural language in real time for Iraqis seeking advice about passport services.

Information gathering started by first studying the crisis which took place due to suspension of all passport services in the year 2003 and the following years. Then frequently asked questions by people about IPS were gathered from The Passports Directorate in Baghdad, The Consulate department at the Ministry of Foreign Affairs, and the Iraqi missions abroad. These questions were analysed and organised to cover all questions and inquiries about passports raised by Iraqis living outside Iraq.

Rules and regulations about the passports were gathered from The Iraqi Passport law (The Iraqi Passport Law, 2006), Iraqi Citizenship law (The Iraqi Citizenship Law, 2006), and the Consular Works Reference Guide (AbdulRazak, 2012).

The researcher found that those references cannot answer all the questions and queries raised by people, and there still some questions without an adequate answers. Therefore He interviewed some passport and citizenship officers at the Ministry of Interior, experts in passports at the Consular Department at the Ministry of Foreign Affairs in Iraq, in addition to some consuls at the Iraqi missions in London, Paris, Cairo, and Manchester.

A special concentration was given to the frequently asked questions raised by Iraqi's living abroad, and the work procedures at the missions to answer these questions and sort out their problems.

The gathered information was engineered to take the form of a general process chart with five main processes about the passports (Issuing new passports, extending passports

Figure (4-2) shows a sample of the process charts produced during the knowledge engineering process, all process charts are attached in the appendix (2) of this thesis.

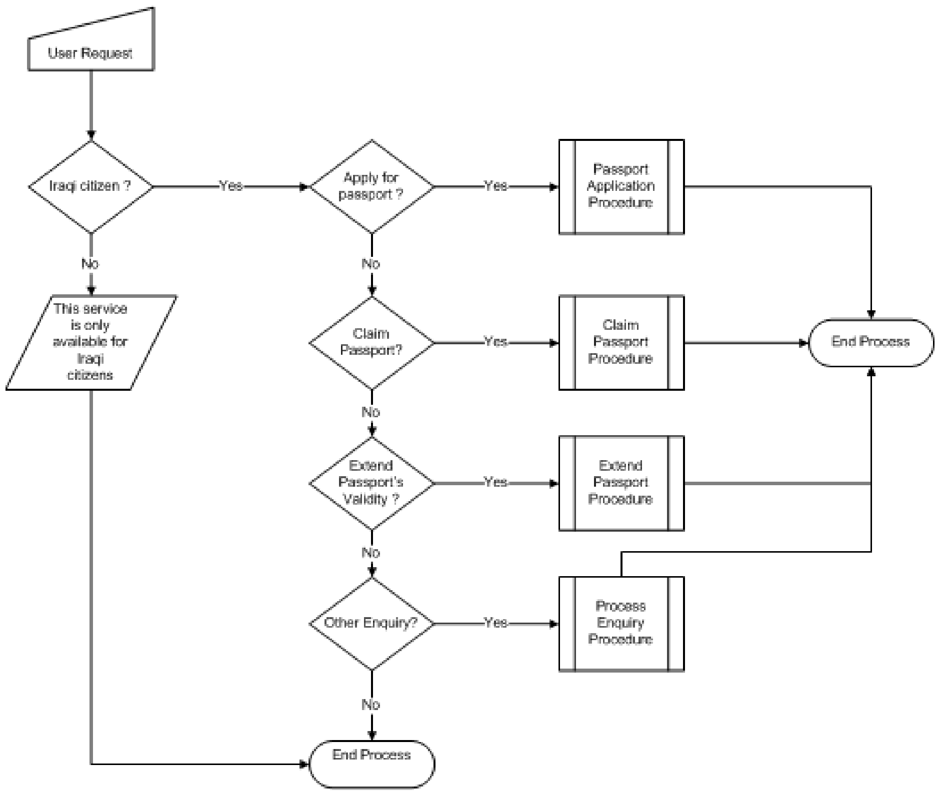


Figure 4-2 A sample of process chart of IPS domain with 4 sub-processes

This process chart was also clarified and discussed with some of the domain experts (consuls) before converting it to a knowledge tree.

4.2.2. Knowledge Transformation

The process chart of the Iraqi passport domain services was converted into a flow chart. Each branch of the flow chart represented one of the main categories for passport services, terms of services were modelled as (if statements), where each condition leads to different results This flow chart representation was the most suitable for the IPS domain, because it is considered to be a procedural domain where each procedure has a set of requirements to be satisfied.

Figure (4-3) shows a flow chart for a new passport procedure that was generated as part of the knowledge engineering process. All flowcharts produced during the knowledge engineering process can be found in Appendix 2 of this thesis.

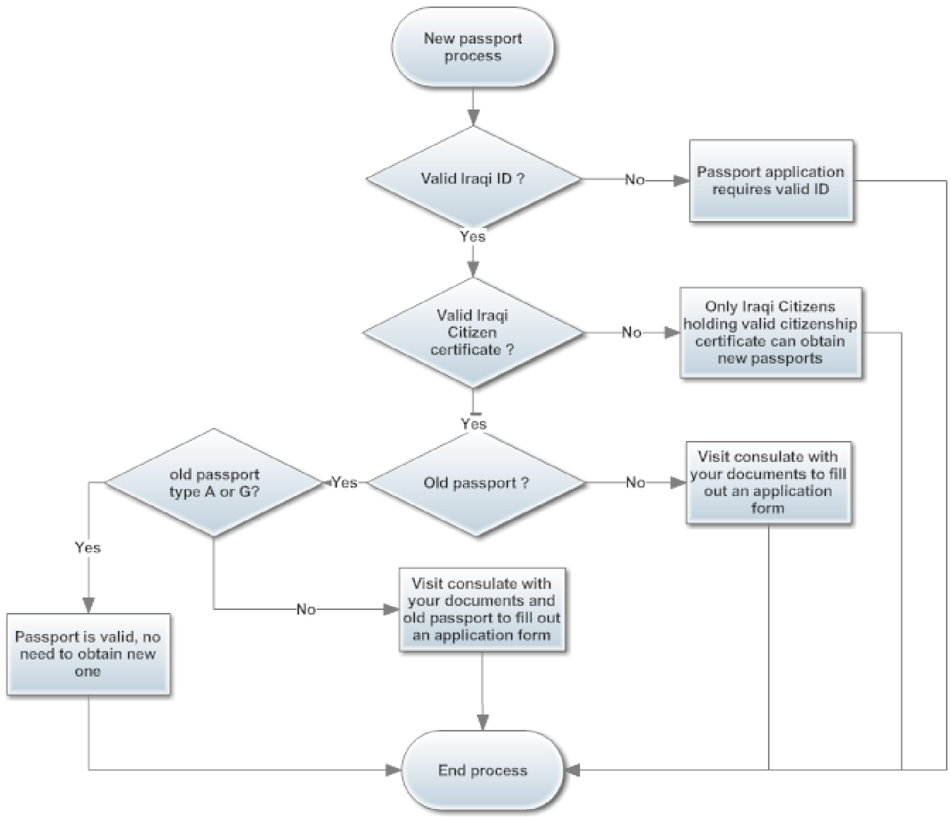


Figure 4-3 Sample flow chart for new passport procedure

In addition to the five main processes of passport domain, there are several FAQs in the IPS domain that could not be considered as a part of the procedure, these questions and topics were categorised as general questions.

For example a user may ask about the validity period of certain type of passport, such a question cannot be classified as a part of the main IPS procedures.

These FAQs were organised under a new node in the knowledge tree called general questions nodes.

4.2.3. Creation of Knowledge Trees

The knowledge tree methodology used in this research was inspired by the conversational bullying and harassment system developed by Latham et al. (Latham, 2010), and was adapted for the purpose of structuring knowledge within an Arabic Conversational Agent with some minor modifications.

The flowcharts produced during knowledge engineering process were converted to knowledge trees by converting each step in the flowchart to a node in the knowledge tree. Figure (4-4) shows a portion of the knowledge tree which is used as basis for domain scripting.

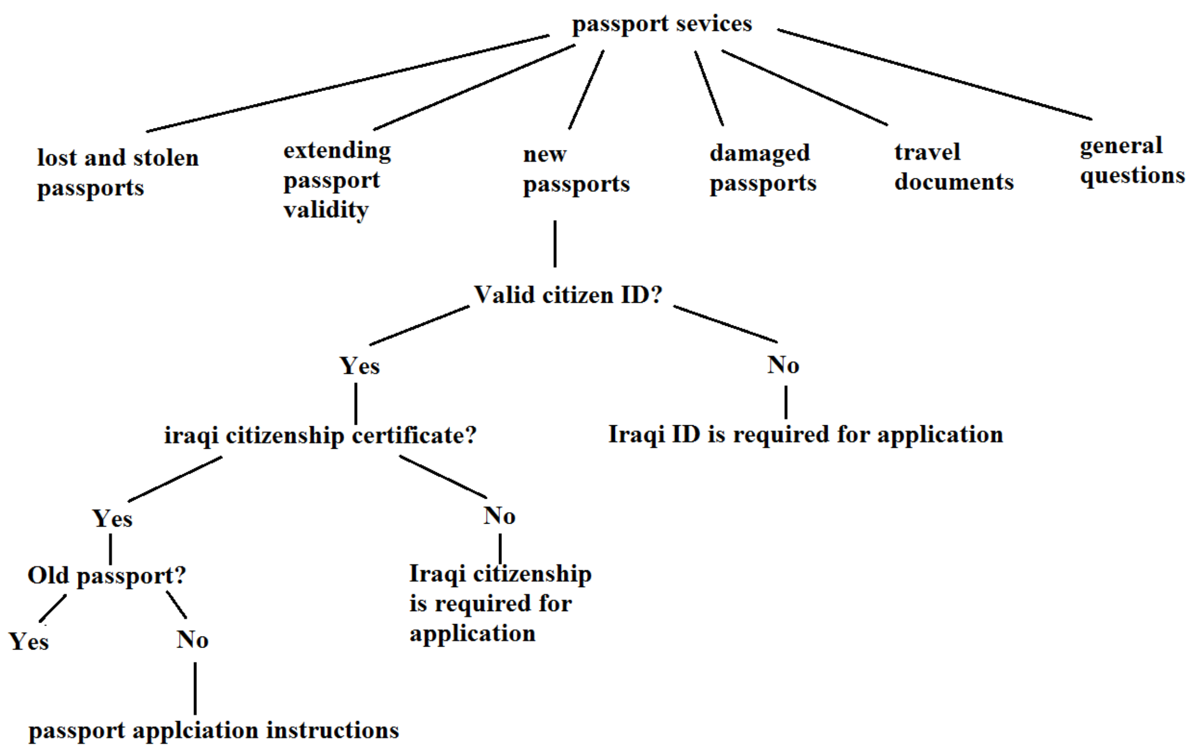


Figure 4-4 a portion of the knowledge tree produced during the KE process

4.2.4. Mapping Conversations to Goals

The knowledge trees described in Section (4.2.3) form the basis for scripting conversation around the user's goals.

The scripting of the knowledge tree was based on identifying the major goals (abbreviated as G throughout this section) and problems for users in general sense first;

G1: Issuing new passports

G2: Extending passports validity

G3: Lost and stolen passports

G4: Passport damage

G5: Travel documents

Normally when users begin a conversation they would more likely give a headline about the subject rather than getting into the details. For example if someone needs a new passport they would more likely say "I want to apply for a new passport" without getting into the details of their case.

4.3. The Proposed Architecture of PMGO-CA

The new architecture was built on the concept of modularity. PMGO-CA functionality was distributed among several modules to facilitate maintenance and future development. Figure (4-5) shows a high level architecture for PMGO-CA and described below:

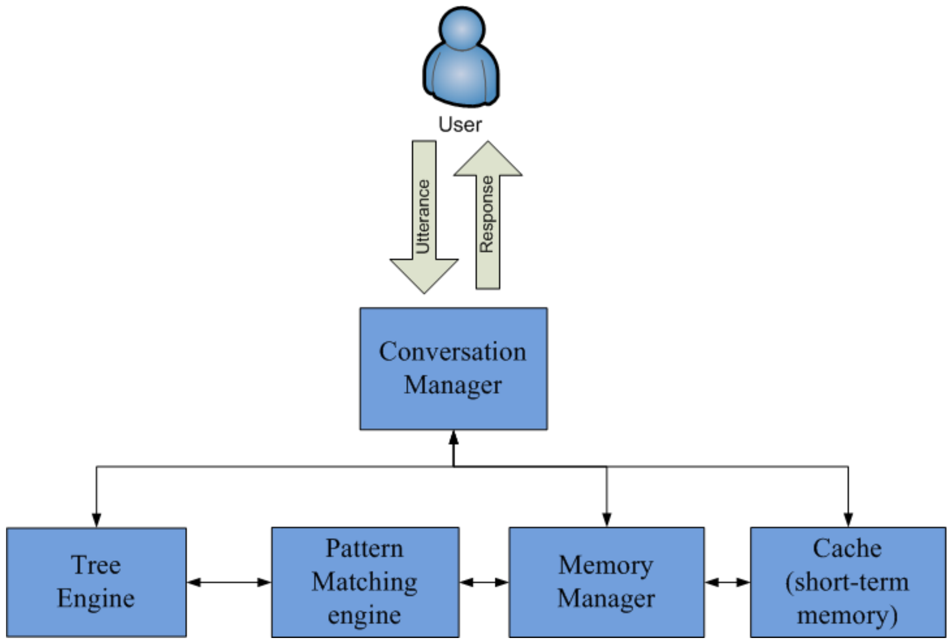


Figure 4-5 PMGO-CA high level architecture

In this section the user-agent interaction is described at a high level to give an idea about the CA's operation before getting into the details of its architecture.

1- Upon the start of a conversation session, PMGO-CA request from the user some personal information which is stored in memory variables about specific information encoded in memory variables (explained in section 4.3.4.1), these variables contain information about users such as name, age and location.

2- Once users have answered these initial questions about their information; the conversation begins by the agent asking the users about the type of service they need, for example

   Agent: how can I help you?

3- When users answer with their purpose of conversation, PMGO-CA performs a search in the knowledge tree for a rule that matches user's utterance. Details about the search algorithm is also covered in section (4.3.3.1), the agent searches for a proper match for user's utterance based on pattern matching algorithm described in section (4.3.2.1).

4- Once a match is found the matched rule is triggered, and the agent responds according to the triggered node type, (node types are described in detail in section 4.3.1.2). If no match is found PMGO-CA asks users to rephrase their utterance for number of times defined by the scripter.
5- Throughout the conversation, PMGO-CA keeps the user's information and triggered nodes in short-term memory (section 4.3.4). Triggered nodes are these which were fired during the conversation. When the conversation ends, PMGO-CA stores this information in long-term memory described in section (4.3.5).
6- The conversation continues until the goal of the user is met or the user ends the conversation.

In order to explain the functionality of the components in the architecture, first the new proposed scripting language needs to be introduced in the next section.

4.3.1. Arabic Pattern Matching Scripting Language

As described earlier in section (4.2.2), the details of domain processes where gathered and represented in a flow chart which was used to shape the form of a knowledge tree. This representation was found to be the most suitable for the IPS domain, due to its procedural nature. Other frequently asked questions (FAQs) which were not accommodated in the procedures were organised in a separate "general context" on the same knowledge tree.

A Goal-oriented approach is used to script the knowledge tree, this approach can be defined as identifying the user's goal first and then gathering other relevant information to achieve this goal. User's goals are also referred to as "Context" in the rest of this chapter.

The scripting of the knowledge tree was performed by converting the flow chart into a conversation and each part of the resulted conversation tree is scripted as a suitable tree node. Figures (4-6) and (4-7) show the knowledge tree in Arabic and its English translation

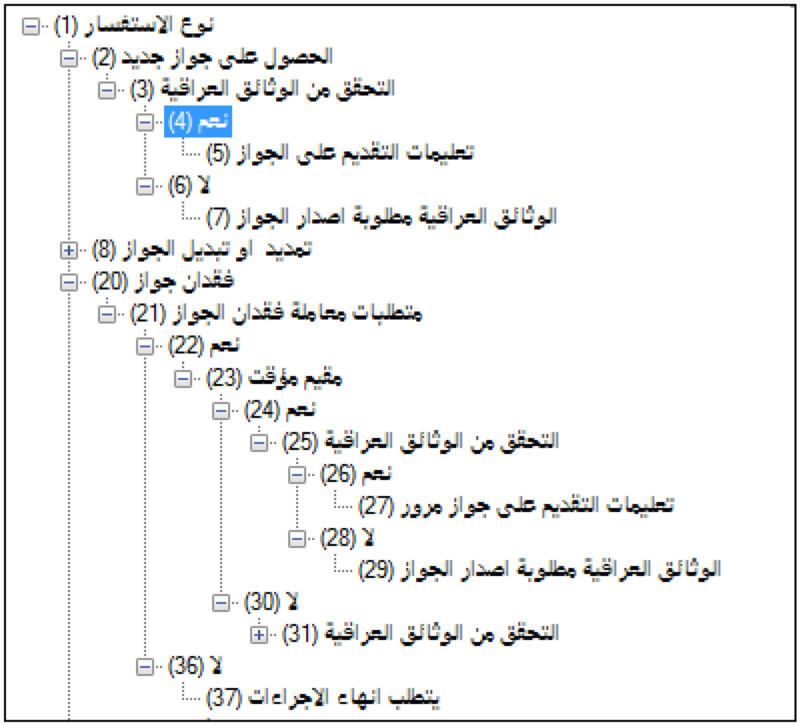


Figure 4-6 IPS knowledge tree (Arabic)

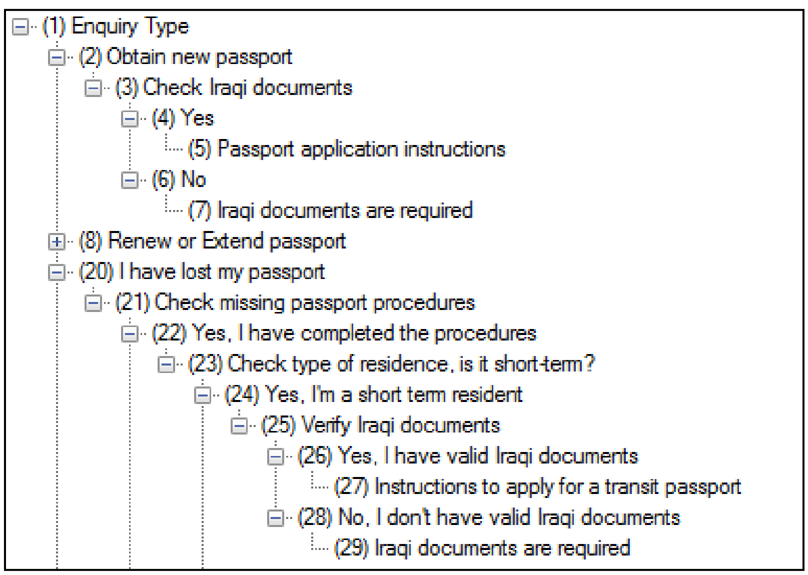


Figure 4-7 IPS knowledge tree (translated to English)

The upcoming sections discuss the details of the scripting language of the knowledge tree, and the tree nodes with their attributes, with some elaboration on how these attributes impact the CA's behaviour.

4.3.1.1. The Tree Script Editor

After gathering knowledge information and modelling them as a flowchart it was necessary to convert the flowchart into a machine-readable data structure. Therefore, a tree script editor was developed to model the domain flowchart as knowledge tree.

The Tree Script Editor is a client-side application used by PMGO-CA administrator to create and maintain the knowledge tree, add and modify rules of the current domain, and create trees for other domains.

Figure (4-8) displays the Tree Script Editor for PMGO-CA. Rules were structured as nodes and organised into a tree structure (described in section 4.3.1.2). After PMGO-CA scripter completes the tree, the scripted tree is saved to a text file and then uploaded to the conversational agent, as described in section (4.3.6.2.4)

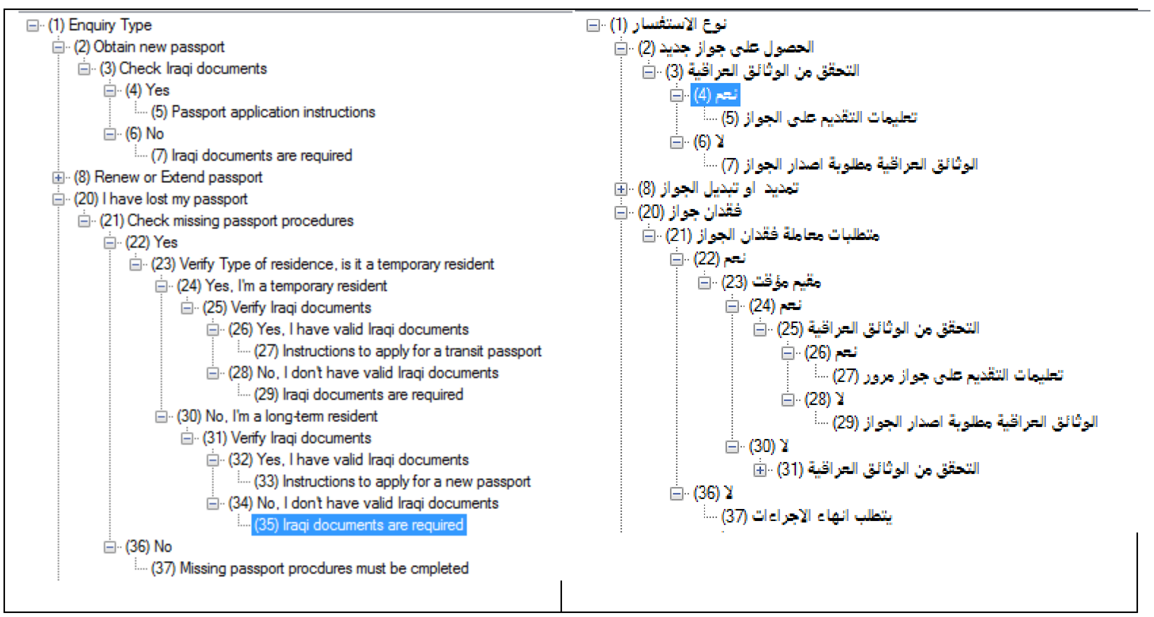


Figure 4-8 Tree Script Editor

4.3.1.2. The Tree Structure

The knowledge tree and its nodes are scripted into a file using the tree scripter editor tool (described in section 4.3.1.1). Knowledge tree files are scripted with the JavaScript Object Notation (Ihrig, 2013). Figure (4-9) shows types of nodes within the knowledge tree:

1. Question nodes.
2. Value nodes.
3. Report nodes.

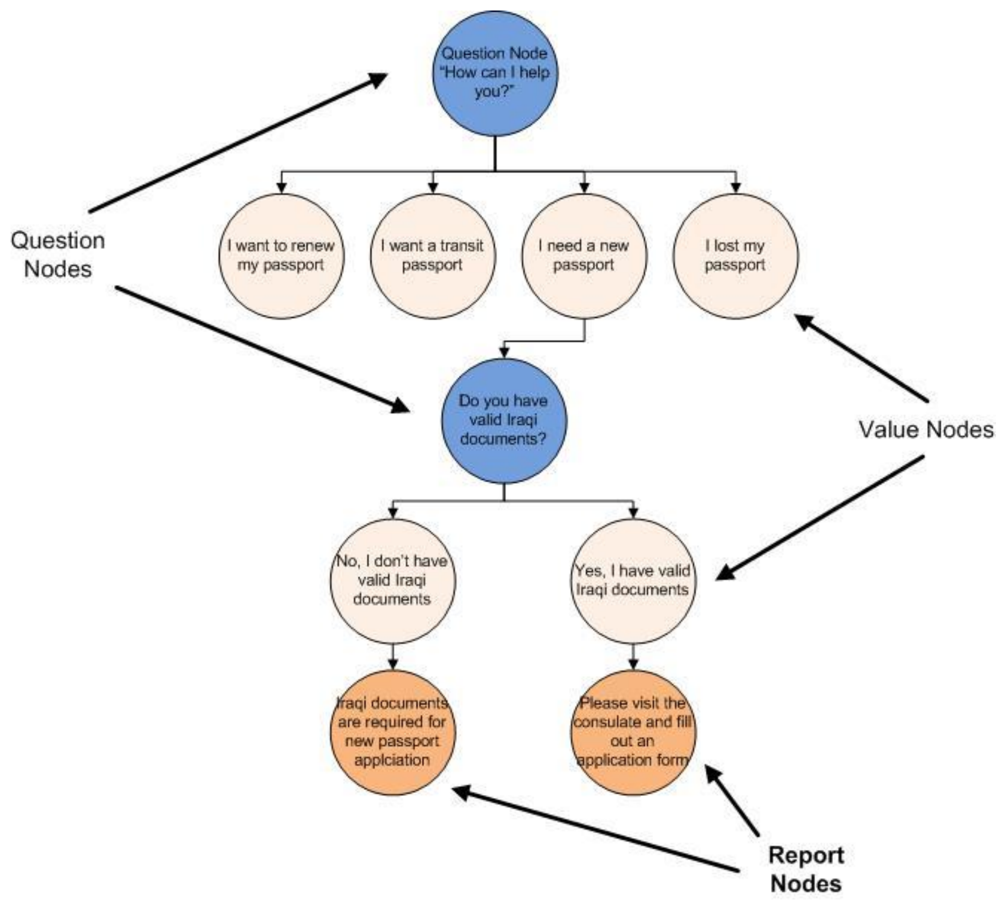


Figure 4-9 types of tree nodes

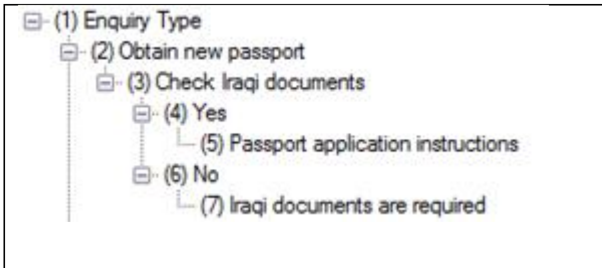
Tree nodes are scripted in a hierarchical format where each node contain the nodes underneath, therefore this section will explain the scripting features of each node type, however there are some common attributes shared among all types of nodes; these include:

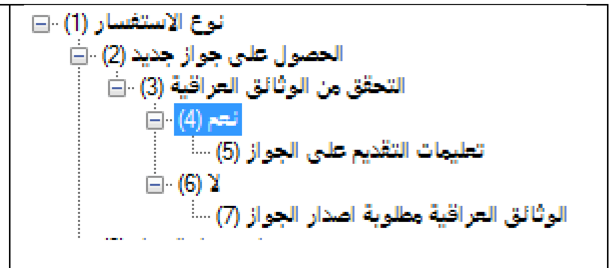
- Unique node identified "ID", used to distinguish nodes from each other.
- Node description, which appears as a node title in the graphical view of the knowledge tree.
- Node type "NType" which is an integer that denotes the type of nodes, as nodes can be any of the followings:
  - Question nodes: they are encoded as type (3).
  - Value nodes: they are encoded as type (4).
  - Report nodes: they are encoded as type (5).
- An array of nodes that contain all the nodes underneath.

Question node

The question node represents a question which the agent asks the user to obtain specific information. When this node is triggered, the question contained within this node will be fired and displayed to the user. Figure (4-9) shows a portion of the knowledge tree which demonstrate node types.

As illustrated in conversation sample (4-1) Node number (1) titled "Enquiry type" contains a question to be asked to the user about the type of help they need; this node is the root node of the tree and is triggered at the beginning of each conversation, as shown in the conversation snippet line number (1).





Agent: How Can I help you?**(1)** User: I need a new passport.**(2)** Agent: Do you have valid Iraqi documents?**(3)**	النظام: كيف يمكننا المساعدة؟ **(1)** المستخدم: أريد الحصول على جواز جديد **(2)** النظام: هل لديك وثائق عراقية نافذة؟ **(3)**

Conversation Sample 4-1 example of question and value nodes

Figure (4-10) demonstrates the scripting features of question nodes; these nodes include a "Question field", which contains a question to be asked to the user by the agent

```
"Root":
{
    "ID":1,
    "NType":3,
    "Description":"(1) نوع الاستفسار",
    "Question":"كيف يمكننا المساعدة ؟",
    "Nodes":[],
}
```

Figure 4-10 attributes of question nodes

Figure (4-11) shows the interface used to add question nodes which have two simple fields. The first is a short descriptive text for this rule which will appear on the tree as the node title, and the second is the question that the agent shall ask the user.

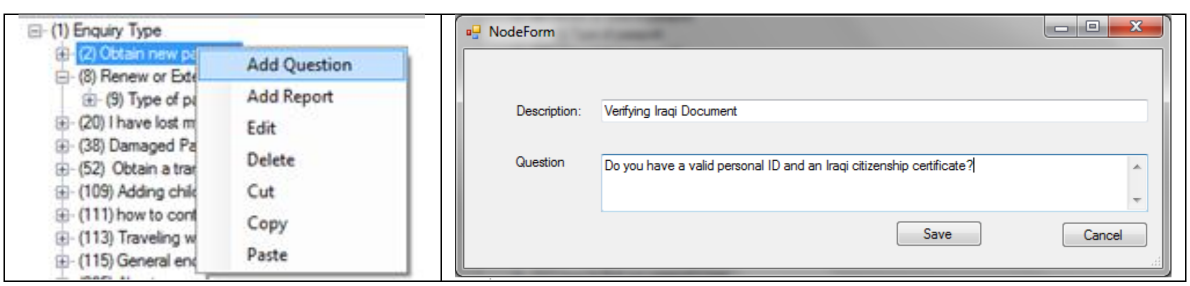


Figure 4-11 adding question nodes

Value node

When the scripter defines a question node they must also define potential alternative answers that the user may respond to. For example when the agent asks the user if they have a valid Iraqi documents, the user often responds with Yes or No. However, sometimes they may answer with “Yes, I have an Iraqi Civil ID, but I don’t have a citizenship certificate”, whatever the case is, the scripter must define all the possible case scenarios in which a user may respond, each of these possible responses is represented with a value node, as shown in figure (4-9).

Value nodes can only be added as sub-nodes to question nodes, which is logical since value nodes represent potential answers for a question asked by the agent. Value nodes contain the patterns associated with that answer to be matched with user utterance in order to activate the node.

Referring to conversation sample (4-1), in the second line of conversation the user responds to the agent by asking for help regarding “new passports” with an utterance that activates the value node number (2) as illustrated in the portion of the knowledge tree.


```
{
    "ID":2,
    "NType":4,
    "Description":"(2) الحصول على جواز جديد",
    "NodeValue":"اريد جواز جديد",
    "patterns":"اريد*اصدار*جواز\r\n*اريد*جواز\r\nاريد*جواز
    "Abuse":false,
    "DisableSearch":false,
    "Nodes":[]
}
```


Figure 4-12attributes of value nodes

Figure (4-12) illustrates the attributes of value nodes, these include:

- “Node Value”, the canonical form of potential answer in natural language.

- “patterns”, field contains all the patterns associated with this canonical form
- “Abuse”: a Boolean field that determines if this node contains abusive patterns.
- “Disable Search” a Boolean field, used to mark this node as a context sensitive node described in the section below.

Figure (4-13) shows the interfaces used to add value nodes to the knowledge tree

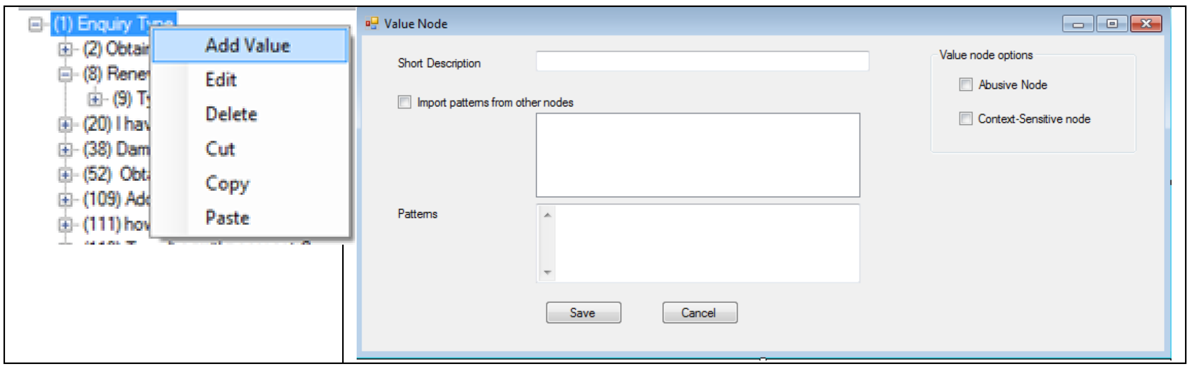
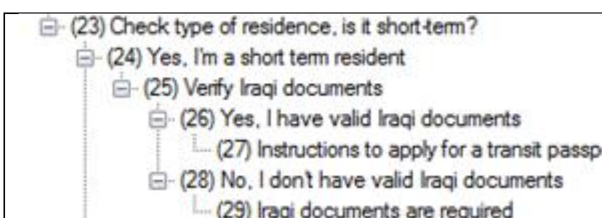
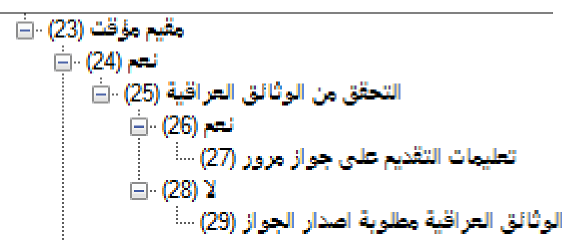


Figure 4-13Adding Value Nodes

Context-Sensitive Node

Context-sensitive nodes are special type of value nodes, they are only active in a specific context (domain goal). For example when the agent asks whether a user has valid Iraqi documents, the user may respond with “Yes, I have valid documents”, this answer is a context related answer, it is only valid when the dialogue flows into that context, as illustrated in conversation sample (4-2).

	
*Agent: Are you a short-term resident?***(23)**	النظام: هل انت مقيم بصورة مؤقتة؟**(23)**

*User: Yes, I'm a tourist.***(24)**	المستخدم: نعم، انا سائح**(24)**

Conversation Sample 4-2 example of context-sensitive nodes

The node which contains the patterns for the answer "Yes, I'm a tourist" can only be activated when the agent asks the related question, but in different context or at the beginning of each conversation the utterance "Yes, I'm a tourist" will not trigger this node, instead PMGO-CA shall ask the user to rephrase their utterance.

Abusive Nodes

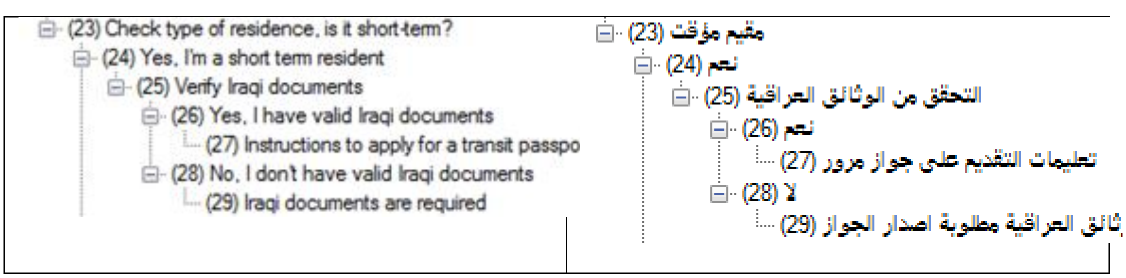
Abusive nodes are value nodes that contain patterns for swearing or other abusive words. If the user utterance contains words which are in the patterns of abusive nodes, the agent shall terminate the conversation.

Report Node

When users answer the agent's question, the agent matches the answer with the patterns of the value nodes of the current question node; when a match is found, the value node is activated.

Value nodes can either contain a question node (if there's information needs to be acquired from the user) or a report node (which contain a respond to user's utterance based on the provided information), this response is encapsulated with a report node.

As a result report nodes are always leaf nodes, in other words report nodes do not contain any descendant nodes, triggering a report node means that a user has completed the goal of their conversation.



Agent: Do you have valid Iraqi documents? **(25)** User: Yes, I do. **(26)** Agent: please fill in the application form an attend the embassy with your documents and (4) personal photos with white background**(27)**	النظام: هل لديك وثائق عراقية نافذة؟**(25)** المستخدم: نعم **(26)** النظام: يرجى ملء الاستمارة و الحضور الى البعثة مع الوثائق و (4) صور شخصيةبخلفية بيضاء**(27)**

Conversation Sample 4-3

The conversation sample (4-3) illustrates the concept of report nodes, in the first line of the conversation the agents asks the user if they have valid documents, the user responds with an utterance that activates the value node number (26), as shown in the tree snippet, once this value node is matched, PMGO-CA expands it and examines the nodes underneath it, PMGO-CA finds a report node (number 27) and fires a response with instructions on how to apply for new passport.

Figure (4-14) shows an example for the attributes of report nodes, these attributes include:

- The “Answer” field contains a final response given to user once all necessary information is gathered.
- “Activation Times” the number of times this node has been triggered in the current conversation for particular user, this option is auto calculated by PMGO-CA and kept in short-term memory. It is always set to “0” at the beginning of conversation, and cannot be altered by the scripter or the user.
- “Activation Limit” the maximum number of activation times for this node, the value of these parameters usually ranges between (1) and (3) this value is defined by the scripter.
- “Activation Limit Message”: a message displayed to users when they reach the activation limit
- “Terminate Conversation”: On Limit Violation”: if this option is checked and the node’s “activation times” becomes equal to its activation limit, PMGO-CA closes the conversation.
- “Mentioned Before”: this Boolean variable is used to check if the current node has been triggered in past conversations with the same user, this option is

controlled by PMGO-CA and maintained in short-term and long-term memory; the scripter cannot alter this option.

- “Memorise”: a Boolean variable determines whether this node (if triggered) will be stored in user’s record in memory database or not.

```
{
    "ID":5,
    "NType":5,
    "Description":"تعليمات التقديم على الجواز (5)",
    "Answer":"املأ الاستمارة الخاصة باصدار الجواز
              واتصل بالبعثة لاخذ موعد للمراجعة
              وجلب هوية الاحوال المدنية وشهادة
              الجنسية العراقية واربعه صور شخصية",
    "ActivationTimes":0,
    "ActivationLimit":3,
    "ActivationLimitMessage":"لقد قمنا بالاجابة اكثر من مرة
                         لذلك سنقوم باغلاق المحادثة",
    "TerminateConvOnLimitViolation":true,
    "MentionedBefore":false,
    "Memorize":true,
}
```

Figure 4-14 attributes of report nodes

Figure (4-15) shows the interfaces used to add report nodes to the knowledge tree

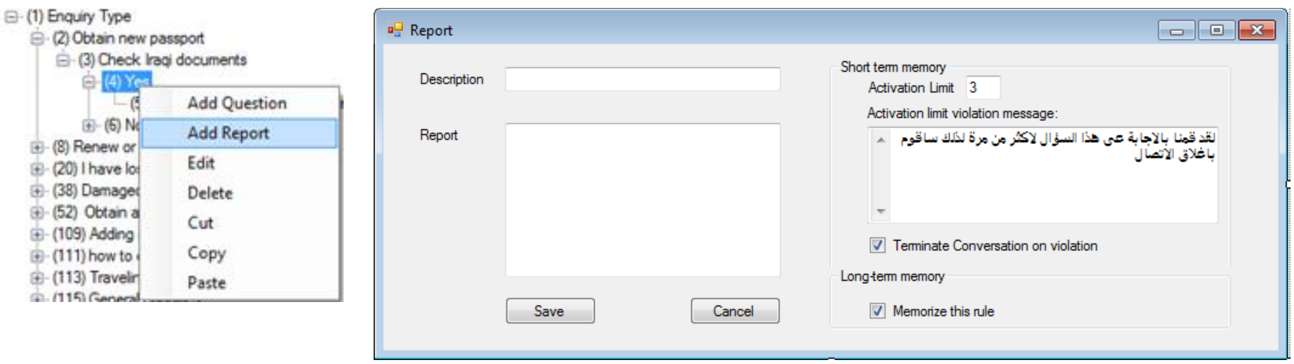


Figure 4-15 Adding report nodes

4.3.2. Pattern Matching Engine

The PMGO-CA utilises a pattern matching algorithm to match a user's utterance against the patterns of domain rules (tree nodes), these patterns are defined by PMGO-CA scripter. Wild cards are used to replace part of text within an utterance (a wildcard is a symbol that may be substituted for any of a defined subset of all possible characters). Wildcards might represent a letter, a number, a word, or series of words, these symbols are the same wildcards used by ArabChat but with simple modification.

wildcard	Meaning
%	An alphabet letter
#	A number
$	One word
*	Null, any character, word, or words

Table 4-1 Pattern wildcards

Table (4-1) lists the symbols used in PMGO-CA; the Original work of (Hijjawi, 2011) used the wildcard "*" to replace many words, but in PMGO-CA the same wild card is used to express anything, ranging from null characters to alphanumeric characters, a word, or a series of words. This eliminates the need to write extra patterns and facilitate the scripting of these rules and make the process less complicated; for example let's consider the following statements:

I want a new passport	اريد اصدار جواز جديد
I want new passports for my kids	اريد اصدار جوازات جديدة لاولادي
I want a new passport for my wife	اريد اصدار جواز جديد لزوجتي

These three utterance indicate the same goal which is “passport issuing”, but with different level of details, the scripter can write a single generic pattern which includes all of these utterances instead of writing tens of patterns for each utterance, this pattern would be “I*new*passport*” “اريد*جواز*جديد*”, which means that any sentence begins with (I) and ends with (passport) followed by any character, word or series of words will match this rule.

4.3.2.1. Pattern Matching Algorithm

This section covers the pattern matching algorithm that PMGO-CA used to evaluate user’s utterance against patterns defined within knowledge tree nodes. First, the following terms must be defined:

- Users utterances (*U*): a unit of dialogue containing a communicative action (Keizer, 2001)
- Keywords: these are words included within patterns, which are separated by pattern symbols

Pattern matching between the user utterance and each pattern within a value node proceed as follows:

1- Identify keywords in user utterance.
2- The pattern is divided into parts according to the keywords (with retaining these keywords), to form a pattern vector (A)
3- The utterance is also divided into parts according to the keywords (with retaining these keywords) to form a sentence vector (B)
4- if the two vectors A,B differ in length then the utterance does not match the pattern
5- For each element in vectors A,B a token is formed such that <A[*i*],B[*i*]> where (i) is the index, forming a vector of tokens (T)

6- Each element (token) of the vector (T) is examined, if A[*i*] is a pattern symbol, evaluate A[*i*] with B[*i*] according the symbol table (4-1) described in section (4.3.2.1). If A[*i*], B[*i*] are compatible then the token is valid.
7- If A[*i*] is a character or word, and B[*i*] is identical to A[*i*] then the token is valid, otherwise the token is not valid.
8- If all tokens are valid, then the utterance matches the pattern, otherwise the sentence does not match the pattern.

Table (4-2) below demonstrates some pattern matching examples

Utterance		**Pattern**		**Result**
اريد الحصول على جوازجديد	I want to obtain a new passport	اريد*جواز	I want*passport	Match
كيف يمكن الاتصال بالبعثة	How can I contact the embassy	*اتصال %بعثة	*Contact the % embassy	Match
لا اريد جواز	I do not want a new passport	اريد*جواز	Want*passport	Not match

Table 4-2 patterns examples

The first example in table (4-2) demonstrates pattern matching process between the utterance "اريد الحصول على جواز جديد" " I want to obtain a new passport", and the pattern "اريد*جواز" "I want*passport", contained within one of the CA's nodes, pattern matching proceeds as follows:

1- Keywords are identified {أريد ، جواز}
2- The pattern is divided according to the stop words, as shown in table (4-3)

A= [اريد، *، جواز]

3- The sentence is divided according to the stop words, as shown in table (4-3)

B= [اريد، الحصول على، جواز]

4- The tokens vector T is formed from the elements of the vector A and vector B

T = [< اريد ، اريد> ,<*, الحصول على> , <جواز ، جواز>]

5- All tokens are evaluated as shown in table (4-3), the table also shows comments on why tokens match.

A	B	T	result	Comment
اريد	اريد	< اريد, اريد>	Valid	Both token elements are identical
*	الحصول على	<*, الحصول على>,	Valid	Symbol (*) replaces many words
جواز	جواز	<جواز, جواز>	Valid	Both token elements are identical
Overall result			match	All tokens are valid

Table 4-3 example of pattern match

Table (4-4) shows an example of the matched pattern “اريد*جواز” and sentence “لا اريد جواز”, with explanatory comments.

A	B	T	Result	Comments
اريد	لا	<اريد, لا>	Invalid	Words in token are not identical
*	اريد	<*, اريد>,	Valid	Symbol (*) can replace any word
جواز	جواز	<جواز ,جواز >	Valid	Words in token are identical
Overall result			Mismatch	The first token is invalid therefor the sentence and pattern do not match

Table 4-4 example of pattern mismatch

4.3.2.2. Conflict Resolution Strategy

As the knowledge domain grows larger, the number of domain rules will also increase. This increased number of rules may cause rules to conflict with each other; as two or more different rules might have patterns that match the same user's utterance, as shown in table (4-5):

Utterance		Pattern		Result
اريد الحصول على جواز نوع أ	I want to obtain a passport type A	اريد*جواز*	I want*passport*	Match
اريد الحصول على جواز نوع أ	I want to obtain a passport type A	اريد*جواز نوع أ	I want*passport type A	Match

Table 4-5patterns conflict

The same utterance " اريد الحصول على جواز نوع أ" matches patterns for two different rules. In such cases there is a need for a mechanism to decide which rule to be triggered. In PMGO-CA, the pattern length is used as a factor to determine pattern weight:

$$W(p) = length(P) \qquad (4\text{-}1)$$

Where W(P) is the weight of the pattern P, and length(P) is a function that can be described as the number of characters contained within the pattern P, these include alphanumeric characters, spaces , and other symbols.

Longer patterns tend to have more information than shorter patterns, thus in PMGO-CA, longer patterns have greater weight (w (p)) than shorter patterns. In case of a conflict between patterns of two rules, the node with the highest weight (pattern length) will be activated.

4.3.3. Tree Engine

The tree Engine controls the dialogue flow according to domain rules which are scripted as tree nodes (explained in section (4.3.1.2), the process of matching an utterance is performed by the pattern matching engine described in section (4.3.2.1).

The tree engine processes the knowledge trees to request information to lead a user towards their goal and then uses the appropriate scripts attached to each node to respond to the user.

This section explains the following:

- The tree search algorithm: this algorithm defines how the knowledge tree is processed to guide users for their goals. This algorithm is the essence of the tree engine
- Management of Context Switching: the context switching determines when and how to switch the conversation between user's goals
- Promotion / Demotion and Activation of Rules: these rules determines in which situation a specific types of nodes are activated/deactivated or given higher priority over other nodes.

4.3.3.1. Tree Search Algorithm

The tree search algorithm is used to control the dialogue flow and to decide which nodes are evaluated, before getting into the details of the search algorithm, the following terms must be identified:

- ***R***: the root node of the knowledge tree.
- ***C***: current node, this node represents the location of current conversation in the knowledge tree.
- **D**: sub-nodes of the current node ***C*** , also called "Candidate nodes"
- ***M:*** matched node, the node that have the best patterns matching to user's utterance.
- ***U:*** user's utterance.
- ***A:*** Agent's response.
- ***T:*** number of times a particular node is activated.
- ***L:*** maximum number of activation times for a particular node, also called "Activation Limit".
- ***V:*** invalid answer violation message, a message that appears when users exceed the (L) of a particular node

The tree search algorithm can be summarized in the steps below:

1- The current node (context) ***C*** is set to the root node of the tree ***R*** at the beginning of the conversation. ***R*** is always a question node, step (1) in figure (4-16)

C = R, where R<>null

2- The agent asks the user the question contained within the node ***C***

3- The user replies with an utterance ***U***

4- A search is performed in the candidate nodes ***D*** of the current node ***C*** to evaluate the user utterance ***U*** against the patterns of these candidate nodes, to find a match node ***M,*** step (2) in figure (4-16)

5- If no match is found PMGO-CA performs a recursive search on all tree nodes, for the rule with the longest pattern (the node with the highest priority pattern), except context-sensitive nodes, to find a match node ***M.*** step (5) in figure (4-16)

6- If a match node ***M*** is found in any of previous steps PMGO-CA examines the descendant node of the matched node ***M***, if it was a question node, PMGO-CA replies with a question and sets the current node ***C*** to that descendent node. step (4) in figure (4-16)

If that descendent node was a report node PMGO-CA checks the node activation time ***T***, if T is equal to 0 PMGO-CA fires the response contained with the node D, increases ***T*** by 1 and resets the current node C to the tree root node R.

A = answer of node D, where T = 0

IF ***T***> 0 and less than the activation limit ***L***, PMGO-CA fires the response plus a notification that this topic has been discussed earlier in the same conversation.

After the report node is activated PMGO-CA increases the number of activation times ***T*** by 1 and resets the current node C to the tree root node R

T=T+1, where T<L

C=R

If the number of activation times (T) is equal to the activation limit L, PMGO-CA replies with an invalid answer violation message, as illustrated in steps (11-15) in figure (4-16)

$$\boldsymbol{A = V}$$

7- If PMGO-CA fails to find a matched node M, the checks if the utterance is related to the domain by comparing the words of the utterance with domain defined keywords, if the utterance contains domain keywords PMGO-CA replies to the user asking them to rephrase their utterance; otherwise it replies with an answer asking the user to be subjective and stick to the domain of conversation, as shown in steps (8,9,10) in figure (4-16).

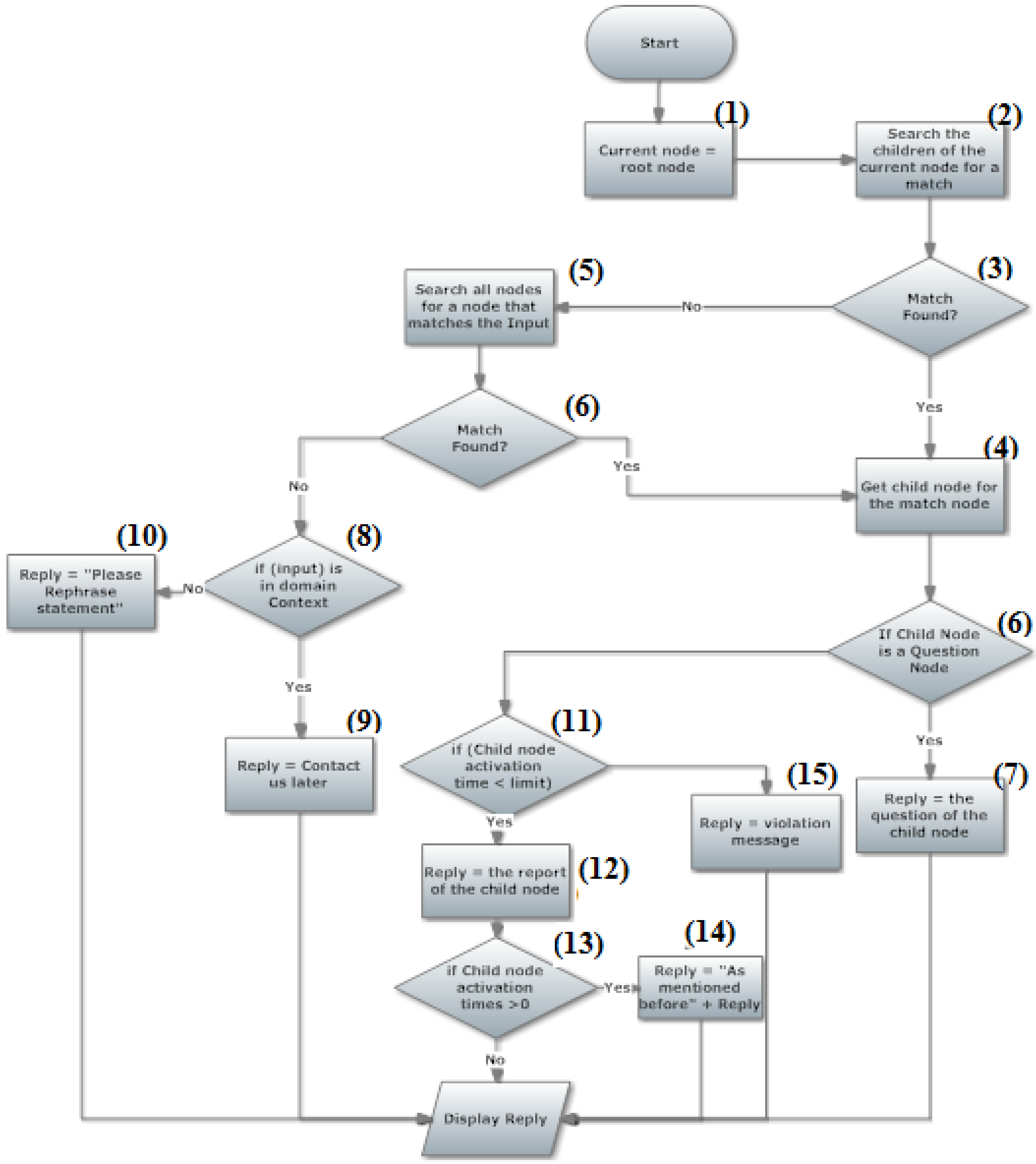


Figure 4-16 Tree search flowchart

4.3.3.2. Context Switching

Context switching in CAs can be defined as a process of switching from one topic of conversation to another.

As discussed earlier in section (4.2), the knowledge of the Iraqi passport domain has been classified in two types; a context or topic, and general question or frequently asked questions (FAQs).

The nodes contained within topics nodes are marked as context-sensitive nodes; because they are associated and activated with a dialogue flow about one topic, these nodes are never activated or considered as candidate nodes when the conversation flows into another context.

Other general questions (non-context) nodes are not considered as topics, they are not restrictive to any dialogue flow or context, and they might be activated even when the conversation is flowing into different context, but only when PMGO-CA does not find a match in the current context.

PMGO-CA was designed in this way for two reasons. First, there are some questions that cannot be classified as a part of context, because they are either related to different topics in the passport domain, or they do not belong to the main five topics of the IPS domain as discussed earlier in section (4.2),

The second is the nature of human conversation, during their dialogues humans do not commit to a single topic, and even if they decide to, it is hard to anticipate what the conversation flow would be like between two people. Furthermore, humans may say something irrelevant during the conversation then they return to the same original context.

Context switching is demonstrated in conversation sample (4-4), where a user has a problem with their lost passport, in line (2), PMGO-CA asks the user whether they have completed passport loss procedures, such as reporting the police and making a journal announcement; the context expects the user to answer either with a patterns of “Yes” or “No”, but the user responds with a question which is out of the current context.

When the user asks about the number of journals to be announced in, line number (5), which is categorised as a general question.

1. You: I lost my passport, what shall I do? 2. System: When a passport is lost, you must inform the police and publish an announcement in a journal about losing your passport, then you can visit the consulate to apply for a new passport, Have you completed these procedures? 3. You: Yes, I have completed these procedures. 4. System: Do you live there for short-term and you want to go back to Iraq? 5. You: In how many journals the announcement must be published? 6. System: the announcement should be made in one news appear and for a period of one day at least. 7. You: Yes, I'm a tourist 8. System: do you have valid Iraqi ID, and citizenship certificate or copies of them?	1. انت: فقدت جوازي ماذاافعل؟ 2. النظام:عند فقدان الجواز يتطلب تقديم بلاغ للشرطة عن حالة الفقدان و نشر الفقدان في الصحيفة و بعد ذلك يتم مراجعة البعثة لاصدار جواز جديد، هل قمت بهذه الاجراءات؟ 3. انت: نعم، قمت بهذه الاجراءات 4. النظام: هل انت مقيم بصورة مؤقته و تريد العودة الى العراق ؟ 5. انت: كم صحيفة يتطلب ان انشر فيها الاعلان؟ 6. النظام: يتم نشر الاعلان في صحيفة محلية واحدة و لمدة يوم واحد على الاقل 7. انت: نعم، انا سائح 8. النظام: هل تملك هوية الاحوال المدنية و شهادة الجنسية العراقية النافذه او صور ملونه منهما

Conversation Sample 4-4

Context switching rules in PMGO-CA can be summarised as:

1- Context switching occurs when the conversation's current context is changed from a context-sensitive node to another context-sensitive node; because it is not possible for both the agent and the user to carry on two conversations on two different topics and maintain their information separately.
Context nodes usually includes topics that needs more information to generate a response to the user, hence it is not possible to have PMGO-CA asking two different questions at the same time and have these questions answered by the user at once.

You: Hello System: Hello You: I want to obtain a new passport System: do you have valid Iraqi ID and citizenship certificate? You: May I extend my current passport instead of applying for a new one? System: this is a different situation; however I can help you with that, what is your passport's type?	انت: مرحبا النظام: اهلاً و سهلاً انت: اريد الحصول على جواز جديد النظام: هل تملك النسخ الاصلية لكل من هوية الاحوال المدنية و شهادة الجنسية العراقية و باصدارات حديثة؟ انت: هل يمكن تمديد جوازي بدل الحصول على جواز جديد ؟ النظام: هذا موضوع مختلف لكن على اي حال يمكنني مساعدتك في ذلك ، ما هو نوع جوازك ؟

Conversation Sample 4-5 Switching from context node to other context node

This is demonstrated in conversation sample (4-5), a conversation occurs between a user and PMGO-CA about obtaining new passport, in line number (4) PMGO-CA asks the user if they have valid Iraqi documents. This topic is a context and is recognised as one of the main processes of the IPS domain, however the user decides not to go with the context of new passports and requests passport renewal service, in line number (5); the agent switches the context to passport extending service and notifies the user that this is a different topic, then PMGO-CA asks the user a question about the type of the passport to be extended, in line number (6)

2- Switching the conversation from a context-sensitive node to non-context node (general question node) is not considered context switching. Because when conversing with the agent in a particular topic, users may ask some unrelated questions, but they do not intend to choose another topic or discuss another type of passport services, therefore when a non-context (general) node is triggered it's neither suitable nor useful to switch the current context to the a non-context node, it would be more useful to retain the current context and remind the user to stick to it.

You: I want a travel document. System: travel documents are issued to citizens with lost passport and wish to go back to Iraq urgently, they are also issued for Iraqi student's children to travel back to Iraq, why do you need a travel document? You: what is the validity period of travel documents? System: travel documents are valid for six months starting from the issue date, please reply to the previous questions so I can help you better? You: my passport was burned System: are you living there for short-term bases and want to return to Iraq?	1 –انت: اريد جواز مرور 2 –النظام: يصدر جواز المرور لمن فقد جواز سفره ، او في الحالات التي توجب المواطن على العودة للعراق بشكل طوعي ، ويصدر لاطفال الطلبة العراقيين للسفر الى العراق، لماذا تريد الحصول على جواز مرور؟ 3 –انت: ما هي مدة صلاحية جواز المرور 4 –النظام: مدة نفاد الجواز هي ستة اشهر تبدأ من يوم الاصدار ، الرجاء الرد على السؤال السابق لاتمكن من مساعدتك بشكل افضل 5 –انت: جوازي احترق 6 –النظام: هل انت مقيم بصورة مؤقته و تريد العودة الى العراق؟

Conversation Sample 4-6

Conversation sample (4-6) illustrates this case. A user is requesting a travel document in line number (1) and the agent asks the user about the purpose of requesting it in line number (2).

Instead of responding to PMGO-CA's question, the user asks about the validity period of travel documents in line number (3). This question is categorised as a general question (non-context).

PMGO-CA responds to the question about passport validity in line number (4) but it retains the previous context and reminds the user about it in line number (4), so the user still has the option to go on with the "travel document" context.

In line number (5), the user says that his passport was burned to justify his need to get a travel document, and the conversation flows in the same context (travel document).

3- Switching the conversation from non-context node to context-sensitive nodes also does not count as context-switching; non-context nodes not only cover IPS domain's FAQs, but they also include nodes to answer some of the frequent statements exchanged among people, such as greeting and courtesy statements. Therefore, it's not suitable to call it context-switching when a user asks a question after greeting or saying "Hi" to the CA, this can be clarified in the conversation sample (4-7)

You: Hello System: Hello You: do I have to take appointment to visit the consulate? System: no need for appointments, you can visit the embassy during the working hours You: fine, I want to obtain a new passport. System: Do you have valid Iraqi documents and citizenship certificate?	انت: مرحبا النظام: اهلاً و سهلاً انت: هل يجب اخذ موعد لمراجعة القنصلية النظام: لا داعي لاخذ موعد بالامكان الحضور خلال ساعات الدوام الرسمي انت: جيد، اريد الحصول على جواز جديد النظام: هل تملك النسخ الاصلية لكل من هوية الاحوال المدنية و شهادة الجنسية العراقية و باصدارات حديثة؟

Conversation Sample 4-7

4.3.3.3. Nodes Activation and Promotion

In general, CAs need mechanisms to decide when to trigger certain rules, and where not to, and how many times they should be triggered during a conversation. These mechanisms are summarised as:

- Rules activation/deactivation: to determine when a rule is activated and when it is not.
- Rule strength: to determine which rule of conflicting rules is triggered, also called conflict resolution strategies as explained in section (4.3.2.2)
- Rules promotion/demotion: to decide in which context certain rules can have more priority than other rules.

These mechanisms were described implicitly in the tree search algorithm covered in section (4.3.3.1), this section focuses on these mechanisms and how they are encoded through the search algorithm.

Previous conversational agents like ArabChat and InfoChat used numerical attributes associated with rules to control these factors. These numeric attributes may cause confusion to the scripter and make the scripting process slower and more cumbersome. In PMGO-CA a simpler approach is followed by replacing some of these numerical attributes with other methods, which are encoded inside the tree search algorithm. This approach can be summarised as:

1- When a conversation flows into a particular context, only context-sensitive nodes and non-context sensitive nodes are activated and considered as candidate nodes for matching, but context-sensitive nodes are promoted over non-context nodes. Because once the user gets into a context they expected to proceed with the conversation flow, so the antecedent nodes of the current node are evaluated first, then if no match is found, the agent performs a search in other non-context nodes. In other words context sensitive nodes are only activated in their current context. Conversation sample (4-8) shows a conversation sample between a user and PMGO-CA about obtaining new passport, PMGO-CA asks the user in line number (2) about their documents, the user responds in line number (3) stating that they have valid Iraqi documents, this response triggers a report nodes with instruction on obtaining new passport.

   Later in the conversation, line number (5) the user asks PMGO-CA if he/she can extend their current passport instead of issuing a new one in line number (6). PMGO-CA asks about the type of user's passport, but in line number (7) the user responds with the same utterance that they used in the "new passport" context, line number (3), PMGO-CA does not trigger the same node trigged in the previous context, and asks the user to rephrase their utterance because it has no meaning neither in the current context nor in the general questions.

1- You: I want a new passport. 2- Agent: Do you have the original copies of valid Iraqi ID and citizenship certificate?	1 –انت: اريد الحصول على جواز جديد 2 –هل تملك النسخ الاصلية لكل من هوية الاحوال المدنية و شهادة الجنسية العراقية و باصدارات

3- User: Yes, I have valid Iraqi documents. 4- Agent: please fill in the application form an attend the embassy with your documents and (4) personal photos with white background 5- You: May I extend my current passport instead of applying for a new one? 6- What is you passport's type? **7-** User: Yes, I have valid Iraqi documents. 8- Please rephrase statement	حديثة؟ 3 –انت: نعم املك وثائق عراقية نافذة 4 –املأ الاستمارة الخاصة باصدار الجواز واتصل بالبعثة لاخذ موعد للمراجعة وجلب هوية الاحوال المدنية وشهادة الجنسية العراقية واربعه صور شخصية 5 –انت: هل يمكن تمديد جوازي بدل التقديم للحصول على جواز جديد؟ 6 –ما هو نوع جوازك ؟ 7 –انت: نعم املك وثائق عراقية نافذة 8 –يرجى اعادة صياغة الجملة من فضلك

Conversation Sample 4-8

2- If there is no particular context in the current conversation, all context-sensitive nodes are deactivated as described earlier in section (4.3.3.1), conversation sample (4-9) shows a conversation snippet between a user and the agent, in line number (3) the user states that he has valid Iraqi documents with an utterance related to the "new passport' context, since the conversation is not going through that context, PMGO-CA does not consider it and asks the user to rephrase his/her statement.

You: Hello System: Hello User: Yes, I have valid Iraqi documents. System: Please rephrase statement	انت: مرحبا النظام: اهلاً و سهلاً انت: نعم املك وثائق عراقية تافذة النظام: يرجى اعادة صياغة الجملة من فضلك

Conversation Sample 4-9

3- There is a numerical attribute associated with each report node, called "activation limit", which is covered in tree search algorithm section (4.3.3.1). The report node can only be triggered for the given number of times, after that the node becomes deactivated, once the activation times reaches the activation limit PMGO-CA responds with the "activation limit violation" message and closes the conversation as discussed in section (4.3.4).

4- Conflicting nodes (nodes with conflicted patterns): when this case occurs the node with the highest weight is triggered, pattern weight is covered in section (4.3.2.2)

4.3.4. Short-Term Memory (Cache)

Short-term memory or (cache) is the memory used by PMGO-CA during the conversation session with users, each conversation with each user has its own version of cache.

Upon the beginning of the conversation session, a cache is created and associated with that particular conversation only, this cache contains:

1- Memory variables queue: their attributes and parameters memory variables are covered in section (4.3.5.1)
2- A copy of knowledge tree, its nodes and their attributes associated with each specific user.

Short-term memory or cache is used to keep user's information captured during the memory conversation (section 4.3.5) along with domain node which were activated during the conversation,

This leads to significant performance improvement, since all nodes and attributes are kept in the server's RAM, there's no additional time cost to look up in a database. In addition, this helps to separate the current context of each user and the activated nodes and maintains separate cache for each conversation of each user.

The cache also keeps track of the current node (context), the knowledge tree and the activated nodes with their activation times and other attributes. These attributes are summarized as:

- ***R***: the root node of the knowledge tree.
- ***C***: current node, this node represents the location of current conversation in the knowledge tree.

- **D**: sub-nodes of the current node ***C*** , also called “candidate nodes”
- ***M:*** matched node, the node that have the best patterns matching to user’s utterance.
- ***T:*** number of times a particular node is activated.
- ***L:*** maximum number of activation times for each node, also called “Activation Limit”.
- ***V:*** invalid answer violation message, a message that appears when users exceed the (L) of each node

The activation time attribute is used to control the number of times a report node is activated; when the node is activated PMGO-CA checks the number of activation times, if this node has been activated more than 0 times and less than the activation limit defined by the scripter for that node, PMGO-CA responds normally to the user but notifies them that this topic has already been mentioned before. This is shown in the conversation sample (4-10).

You: I lost my passport, what shall I do? System: When a passport is lost, you must inform the police and put an advertisement in a journal about losing your passport, then you can visit the consulate to apply for a new passport, Have you completed all these procedures? You: Yes, I have System: Do you live there for short-term and you want to go back to Iraq? You: Yes, I’m a tourist. System: do you have valid personal ID and Citizenship certificate, or copies of them? You: Yes System: As mentioned before, you can send a written request to the consulate to obtain a transit passport, which requires copies of your Iraqi documents and 4 personal pictures with white	انت: فقدت جوازي ماذا افعل النظام: عند فقدان الجواز يتطلب تقديم بلاغ للشرطة عن حالة الفقدان ، نشر الفقدان في الصحيفة و بعد ذلك يتم مراجعة البعثة لاصدار جواز جديد ، هل قمت بهذه الاجراءات ؟ انت: نعم قمت بذلك النظام: هل انت مقيم بصورة مؤقته و تريد العودة الى العراق ؟ انت: نعم انا سائح النظام: هل تملك هوية الاحوال المدنية و شهادة الجنسية العراقية النافذه او صور ملونه منهما ؟ انت: نعم النظام: كما ذكرنا سابقا يمكن تقديم طلب خطي الى البعثة للحصول على جواز مرور و يتطلب ذلك وجود نسخ من الوثائق العراقية و 4 صور شخصية بخلفية بيضاء

background.	

Conversation Sample 4-10

However, when the number of activation times reaches the activation limit PMGO-CA responds with an "Activation violation" message and closes the conversation session if the scripter chooses this option for that node, as shown in the conversation sample (4-10).

If a report node is activated several times during the same conversation, the agent stops the conversation based on an activation limit defined by the scripter which is set to (3) by default, however the scripter has the option to tweak this number according to their needs. A sample this occurrence is shown in conversation sample (4-11) line number (8) where PMGO-CA responds to the user stating that this topic has already been activated several times during the conversation, and then the agent terminates the conversation.

1- You: I lost my passport, what shall I do? 2- System: When a passport is lost, you must inform the police and make a journal advertisement about losing your passport, then you can visit the consulate to apply for a new passport, Have you done all of these procedures? 3- You: Yes, I have 4- System: Do you live there for short-term and you want to go back to Iraq? 5- You: Yes, I'm a tourist. 6- System: do you have valid personal ID and Citizenship certificate, or copies of them? 7- You: Yes 8- System: I have answered this enquiry multiple times, thus this conversation will be closed, thank you.	1 –انت: فقدت جوازي ماذا افعل 2 –النظام: عند فقدان الجواز يتطلب تقديم بلاغ للشرطة عن حالة الفقدان ، نشر الفقدان في الصحيفة و بعد ذلك يتم مراجعة البعثة لاصدار جواز جديد ، هل قمت بهذه الاجراءات ؟ 3 –انت: نعم قمت بذلك 4 –النظام: هل انت مقيم بصورة مؤقته و تريد العودة الى العراق؟ 5 –انت: نعم انا سائح 6 –النظام: هل تملك هوية الاحوال المدنية و شهادة الجنسية العراقية النافذه او صور ملونه منهما ؟ 7 –انت: نعم 8 –النظام: لقد قمت بالاجابة اكثر من مرة عن هذا الاستفسار، لذلك سأقوم بأغلاق المحادثة، شكراً لكم

Conversation Sample 4-11

Another memory feature has been already described implicitly in the context switching section (4.3.3.2). Considering the conversation sample (4-12) a user is conversing with

PMGO-CA about extending his/her passport, when the agent ask about the user's passport type, the user does not respond to the question and states that he needs to call the embassy instead of using the CA; Although PMGO-CA responds with instructions on how to contact the embassy, but it retains the current context at the same time so the user can proceed to converse with the about extending their passport.

You: Can I extend my passport instead of applying for a new one? System: what's your passport's type? You: I want to call the embassy instead of chatting here System: you can reach the Iraqi consulate or embassy on the following numbers 00000 0044and 004400002, respectively, please answer my previous answer so I can help you better	انت: هل يمكن تمديد جوازي بدل التقديم للحصول على جواز جديد؟ النظام: ما هو نوع جوازك ؟ انت: اريد الاتصال بالسفارة و لا اريد التحدث هنا النظام: عنوان السفارة هو 22 كوينزكيت ورقم تليفون السفارة هو 0044 00000 . وعنوان القنصلية في مانجستر هو 24 اوكسفورد كورت ورقم الهاتف هو 004400002، الرجاء الرد على السؤال السابق لاتمكن من مساعدتك بشكل افضل

Conversation Sample 4-12

4.3.5. Long-Term Memory

Long-term memory has been a real challenge in CAs development. This research introduces a modest attempt to simulate long-term memory in CAs to store users' information and the activated nodes during their conversation for long-term bases.

Questions, answers, and patterns related to user's information were encapsulated in structure called "memory variable"; these long-term memory variables are kept in a database and it's the scripter's task to organize and manage them.

Memory variables are stored in a database separately from the knowledge tree to provide more control over their options and attributes. In addition, each user has different parameters for these variables, table (4-6) summarizes the memory variables used within PMGO-CA with their parameters range

Memory variable	Parameter range
Name	List of (1000) frequently used Arabic names
Age	Ranges between (18-90) years
Nationality	Iraqi / Non-Iraqi
Current country of residence	a list of (200) countries
Province of origin	A list of (18) Iraqi provinces

Table 4-6 memory variables and parameters

The functionality of long-term memory in PMGO-CA is to gather user's information according to the defined variables and store them in a database along with the activated nodes, when the same user initiates the conversation and gives the same information PMGO-CA assumes that it is the same user and retrieves the activated nodes from the database which contains the users parameters and the description of the activated nodes of previous conversation. The description of nodes is used instead of the nodes ID number because when the tree is modified some nodes may obtain different ID numbers. In addition, processing the log file to extract contexts was not viable either. Therefore, only the description of activated nodes is stored in the database. Conversation logs (discussed in section 4.3.6.2.1) are also captured during the conversations, but they do not play any role in long term memory of PMGO-CA; they are used for the sole purpose of monitoring and observing the performance of the CA.

The problem with the memory approach used within PMGO-CA is that many users may share the same information like age and location; this makes it impossible to accurately identifying users. Another problem is that user's utterances are based on patterns; and the user when initiates other conversation may give the same information but with different pattern, in this case CA's memory will fail to identify them. Conversation sample (4-13) demonstrates an example of two conversations with the same user but with different times.

You: Hello System: Hello, May I have your name please?	انت: مرحباً

<table>
<tr><td>You: Mohammed.
System: from which Iraqi governorate are you?
Mohammed: Baghdad.
System: How old are you?
Mohammed: 24 years old.
System: Do you live outside Iraq now?
Mohammed: Yes.
System: in which country do you live now?
Mohammed: India.
System: How can I help you?
Mohammed: I want to obtain a new passport.
“output omitted for brevity”
System: Please fill out the application form and contact the embassy to schedule an appointment to apply for the new passport; you shall need your personal ID and your Iraqi Citizenship certificate, in addition to 4 personal photos.</td><td dir="rtl">النظام: اهلاً و سهلاً ممكن اعرف اسم المتحدث؟
انت: محمد
النظام: من اي محافظة انت ؟
محمد: من بغداد
النظام: كم يبلغ عمرك؟
محمد: 24 سنة
النظام: هل انت مقيم خارج العراق حاليا؟
محمد: نعم
النظام: في اي دولة تقيم الان؟
محمد: الهند
النظام: كيف يمكننا المساعدة؟
محمد: اريد اصدار جواز سفر
"تم اخفاء جزء من النص للاختصار "
النظام: يرجى ملء استماة التقديم و الاتصال بالبعثة لتحديد موعد للتقديم على جواز، و يجب احظار هوية الاحوال المدنية و شهادة الجنسية العراقية بالاضافة الى 4 صور شخصية</td></tr>
<tr><td colspan="2">When the same user initiates a conversation later</td></tr>
<tr><td>You: Hello
System: Hello, May I have your name please?
You: Mohammed.
System: Mohammed, from which Iraqi governorate are you?
Mohammed: Baghdad.
System: How old are you?
Mohammed: 24 years old.
System: Do you live outside Iraq now?
Mohammed: Yes.
System: in which country do you live now?
Mohammed: India.
System: I remember we’ve talked previously about (obtaining a new passport), how can I help</td><td dir="rtl">انت: مرحبا
النظام: اهلاً و سهلاً ممكن اعرف اسم المتحدث؟
انت: محمد
النظام: من اي محافظة انت ؟
محمد: من بغداد
النظام: كم يبلغ عمرك ؟
محمد: 24 سنة
النظام: هل انت مقيم خارج العراق حالياً؟
محمد: نعم
النظام: في اي دول تتقيم الان؟
محمد: الهند
النظام: اذكر اننا تحدثنا سابقاعن (اصدارجواز)
كيف يمكنني مساعدتك</td></tr>
</table>

you?	

Conversation Sample 4-13Memory variables

4.3.5.1. Memory Variables

As explained earlier, memory variables are data structures holding all memory related information for a particular topic, the attributes of a memory variable are described in table (4-7)

Attribute name	Explanation	Examples	
		Arabic	English
Name:	A descriptive name for the memory variable.	السكن	Location
Logical Name:	A unique identifier for each memory variable, this name is used as an identifier to keep user's information in database.	This attribute must be defined in English only	Location
Query Question:	A question to be asked to users to gain information.	"اين تسكن ؟"	"Where are you located?"
Answers:	Patterns for potential answers for a questions asked by the agent	"العراق" "مصر"	"Jordan" "Egypt"
Prefix:	This field contains characters or words that may proceed the answers	"انا اعيش في", "انا ساكن في"	"I live in" "I'm located in"
Suffix:	This field contains characters or words that may proceed the answers	"منذ سنة" "منذ شهر"	"for a year" "for a month"
Invalid Answer Message:	a message displayed to user when his answer does not match any pattern	"يرجى كتابة اسم البلد بوضوح"	("Please type country name properly")
Invalid Answer Limit:	The maximum number of times a user can answer with invalid utterance to memory variable	"1".	"1".
Invalid Answer Violation Message:	a message displayed to users when the exceed the Error limit.	"لم تقم بادخال المعلومات بشكل صحيح لعدة مرات، لذلك سنضطر لاغلاق الاتصال"	"you've typed invalid information for several times, therefore the conversation will be closed"
Terminate Conversation On Violation:	This is a Boolean attributes that causes the conversation to be closed if the invalid	True	True

Attribute name	Explanation	Examples	
		Arabic	English
	answer limit is reached		

Table 4-7 A attributes of memory variables

Conversation sample (4-14) illustrates the behaviour of memory variables; a user is conversing with the CA. In line number (11) PMGO-CA asks the user about their location, which is a query question. Then the user responds with a random utterance which has no meaning in line (12). This utterance is evaluated against a combination of answers defined in the answers field and their affixes (defined in prefix and suffix fields); PMGO-CA does not find any match with these combination; so the agent tries to perform an approximate string matching with all answers to check if there's an approximate match (approximate string matching is covered shortly). After the CA fails to find an approximate answer; PMGO-CA increases the Invalid answer counter for this variable by (1) and fires an "invalid answer message" in line number (13).

1- System: Hello, May I have your name please?	1. النظام: اهلاً و سهلاً ممكن اعرف اسم المتحدث؟
2- You: Mohammed.	2. انت: محمد
3- System: Mohammed, from which Iraqi governorate are you?	3. النظام: انت من اي محافظة يامحمد؟
4- Mohammed: Baghad.	4. محمد: من بغاد
5- System: do you mean (Baghdad)?	5. النظام: هل تقصد (بغداد) ؟
6- Mohammed: Yes.	6. محمد: نعم.
7- System: How old are you?	7. النظام: كم يبلغ عمرك؟
8- Mohammed: I'm 24 years old.	8. محمد: عمري 24 سنة
9- System: Do you live outside Iraq now?	9. النظام: هل انت مقيم خارج العراق حاليا؟
10- Mohammed: Yes.	10. محمد: نعم
11- System: in which country do you live now?	11. النظام: في اي دولة تقيم الان؟
12- Mohammed: "some random answer".	12. محمد: "كلام مبعثر"
13- System: Please type your city name properly	13. النظام: يرجى اعادة كتابة اسم البلد بطريقة صحيحة.
14- Mohammed: "any random answer".	14. محمد: "كلام مبعثر"
15- System: you've typed invalid information for several times, therefore the conversation will be closed	15. النظام: لقد قمت بادخال معلومات غير صحيحة لعدة مرات لذلك سنقوم باغلاق المحادثة.

Conversation Sample 4-14 Memory procedure

When the user answers with a random utterance again in line number (14), PMGO-CA checks the Invalid answer counter which is (1) and compare it with the "invalid answer limit" which is (1) in this case PMGO-CA responds with an "invalid message violation" message in line number (15) and closes the conversation because the option "Terminate conversation on violation" is active.

Affixes are extra information that users usually include in their utterance but they cannot be considered as part of the answer. For example, when users are asked about their location; they may answer with "I've been living in Jordan for five years", in such case the phrases "I've been living in" and "for five years" are not a part of the answer which is (Jordan); the answer only will be stored in the memory database.

Let us consider the conversation example (4-14), in line number (7) PMGO-CA asks about user's age; the user responds with "I'm 24 years old" PMGO-CA looks up the prefix, answers, and prefix fields (explained in table (4-6)) of the "age" variable and forms a combination of patterns among them; if the prefix field has the pattern "I'm" and the answer field has a pattern of "24" and the suffix field has the pattern "years old"; then one of the combinations would be "I'm 24 years old" and it would match the user's utterance in this case, but the actual answer is retained as "24" and this answer only will be considered in the database of long-term memory.

There are two types of affixes, the first is prefixes which proceed the answers such as the phrase "I've been living in" in the above example; and suffixes which follows the answer, such us the phrase "for five years" in the example above, these potential affixes are listed in the "prefix" and the "suffix" fields of memory variables.

Approximate string matching is performed using Levenshtein distance (Gonzalo, 2001), if there is a good similarity score between the user's answer and the patterns defined within the "answers" field in memory variable, PMGO-CA asks the user if they meant this approximate answer. For example, in line number (3) PMGO-CA asks the user about their governorate of origin, the user mistakenly responds with “Baghad” instead of “Baghdad” in line number (4), when PMGO-CA checks the combination of answers and affixes it does not find a match, so PMGO-CA performs approximate string matching with the defined answers and find an approximate answer “Baghdad” to user’s utterance “Baghad”, then the agent asks the user if they meant the approximate answer “Baghdad” in line number (5), and the user responds with “Yes” in line number (6) to confirm that they meant “Baghdad”.

4.3.5.2. Memory Algorithm

Figure (4-7) shows a flowchart explaining how memory variables are processed at the start of the conversation session between the user and the agent.

When users initiates a session with PMGO-CA, the memory algorithm starts by retrieving memory variables from database and organising them in a queue according to their priority, then PMGO-CA starts processing them and enquiring user’s about their information, when users respond to all of these variables PMGO-CA compares the user’s parameters with all parameters of previous conversations with all users, which are stored in user’s database, if the same parameters are found, PMGO-CA retrieves a list of the activated nodes in previous conversation, otherwise PMGO-CA keeps the parameters of the new user’s in short-term memory.

When the memory algorithm ends, the tree search algorithm (explained in section 4.3.3.1) starts to converse with the user to help them achieve their conversation goal.

At the end of the conversation session, PMGO-CA creates a record for the user with their information and the activated nodes. If the user already had a record then their record is updated with the new activated nodes.

Before getting into the memory algorithm the following terms are defined:

- Q: memory variables queue
- X: Current memory variable.
- I: query question associated with the current memory variable Q

- A: answer patterns defined within memory variable
- P: prefix
- S: suffix
- M: invalid answer message
- L: invalid answer limit
- T: invalid answer counter
- V: invalid answer limit violation message

A pointer in memory called “Current memory variable” indicates the memory variable being processed; the steps below summarizes the memory procedure, as shown in figure (4-8)

1- Memory variables are retrieved from the database and organized in a memory queue ***Q*** based on their priority.
2- Retrieve memory variable form memory queue and assign the current memory variable to it, as shown in figure (4-17), steps (1), (2) and (3)
3- Display a query question associated with the current memory variable ***X***, as shown in step (4) and (5) in figure (4-17)
4- When the user responds to the question, PMGO-CA checks user’s answer against the patterns of answers associated with the memory variable, step (6) and (7) in figure (4-17),the pattern is computed by combining the values of the answer (***A***), prefix (***P***), suffix (***S***) fields. Each item in these fields is cross joined with the items of the other fields to create a list of combinations; user utterance is evaluated against these combinations to find a match
5- If PMGO-CA does not find a match in the list of the combinations of (answer (***A***), prefix (***P***), suffix (***S***) fields); PMGO-CA searches for an approximate answer, as shown in step (8) and (9) in figure (4-17), with the use of approximate string matching using Levenshtein distance (Gonzalo, 2001) between each item of the answers field and the answer of the user. If an approximate match is found, PMGO-CA displays a message to the user to ask if they meant this answer. If the user answers with “yes” "نعم", their answer is kept in the cache and the next variable is processed.

6- If no match is found in the previous steps PMGO-CA checks the number of invalid answers ***T***, as shown in step (10) if it is less than the invalid answer limit ***L***; PMGO-CA replies to user with "invalid answer message" (***M)***defined by the scripter(step 11) and increases the number of invalid answer by (1) (step 12)

If the number of invalid answers ***T*** is equal to the limit ***L***, PMGO-CA responds with the "invalid answer violation" message (***V***) defined by the administrator (step 13), then PMGO-CA checks the option "terminate conversation on violation", if this option is activated by the scripter PMGO-CA closes the conversation step (14) and (15)

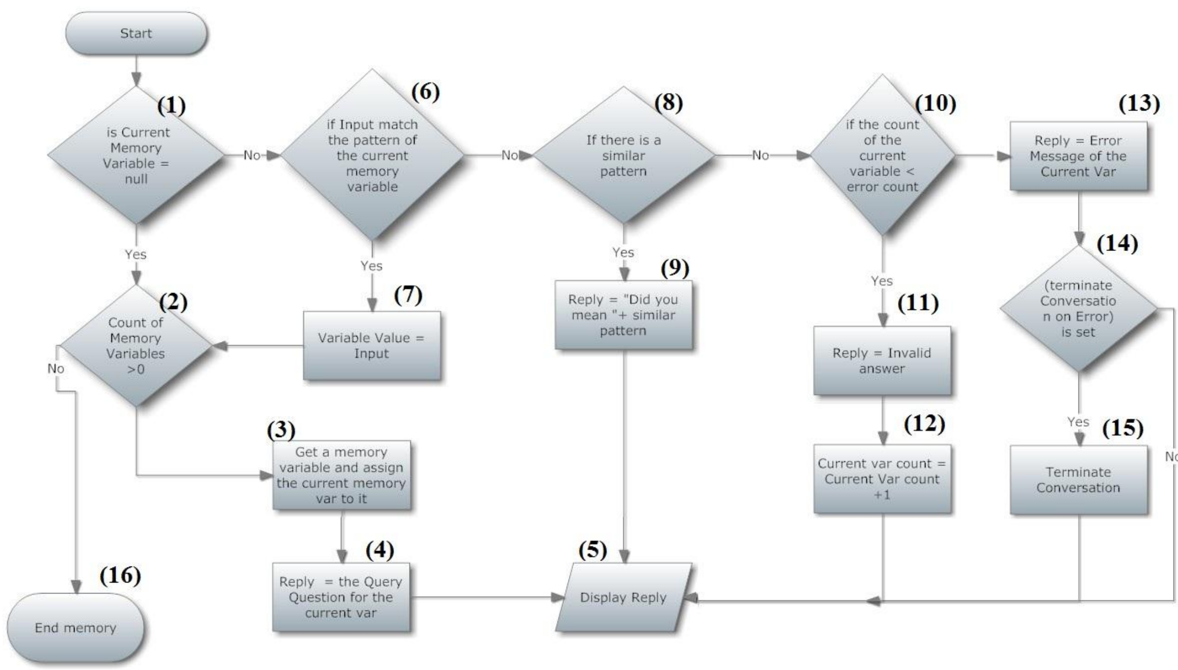


Figure 4-17 Memory flowchart

4.3.6. Conversation Manager

This module serves as an interface between the user and the agent; it receives user's input and displays CA's response. In addition, this module performs the coordination among other CA's modules.

PMGO-CA has been implemented as web site hosted online at (http://www.iraq-pass-ca.net), this web site has a user interface to interact with users in addition to administrative interfaces to manage all agent's aspects including interfaces to manage and view all PMGO-CA options such as memory variables (discussed in section 4.3.5.1), conversation logs, unrecognized user utterances and an interface to upload the scripted knowledge tree discussed in section (4.3.6.2.4)

4.3.6.1. User interface

As shown in figure (4-18), the user interface has an area to type in the utterance and another area to show the utterance and the system response; this interface also has a button to download the current conversation as a text file and another button to close the conversation.

Normal users have access to the conversation boxes only, if they attempt other features such as logs or tree management, they are prompted to provide the administrative credentials for PMGO-CA

Figure 4-18 PMGO-CA user interface

4.3.6.2. Administrative Interfaces

The PMGO-CA administrator is authorised to browse and manage all aspects of PMGO-CA including viewing conversation logs, managing knowledge tree, and long-term memory management. The upcoming sections explain these interfaces in depth.

Conversation Logs

Conversations history between users and PMGO-CA are stored in a database and can be browsed by the system administrator, these logs contain all conversations with the time and date of each one and do not include any other information or statistics. Figure (4-19) shows the interface of browsing and viewing the conversation logs.

Figure 4-19 conversation logs interface

Unrecognised Utterances

In conversations, sometimes there are some utterances that are not recognised by PMGO-CA. These utterances are stored in a separate database of unrecognised utterances to be viewed by PMGO-CA administrator directly. This enables the administrator to find out which patterns can be used to update the knowledge tree in order to respond to this unrecognised question. Figure (4-20) shows the interface of unrecognised utterances captured during conversations with users.

Figure 4-20 Unrecognized utterances interface

Memory variables

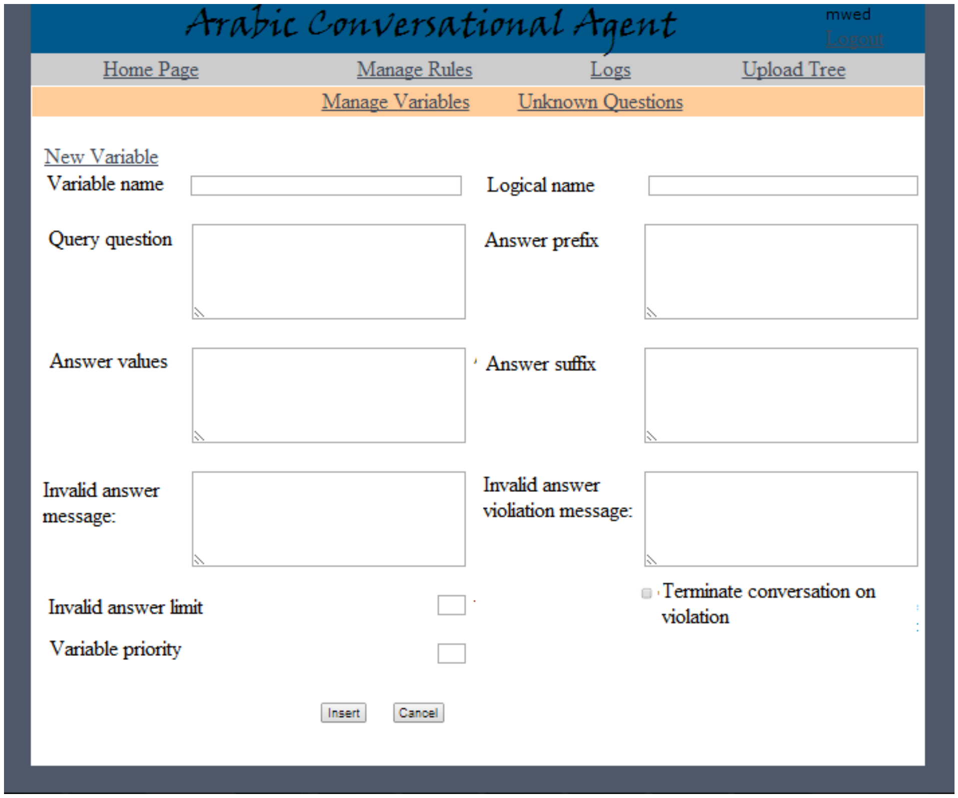


Figure 4-21 Memory variables interface

Figure (4-21) shows the friendly user interface used to add a new memory variable to PMGO-CA.

Upload Knowledge Tree

When knowledge is changed or modified whether due to change in the domain itself or to modify the scripts to handle more users utterance, the scripter can modify the domain knowledge tree using the tree script editor (described in section 4.3.1.1), then upload the new tree to PMGO-CA. Figure (4-22) shows the interface of uploading a new knowledge tree file to help the scripter modify the knowledge of PMGO-CA.

Figure 4-22 Upload tree file interface

4.4. Contributions of PMGO-CA

The contributions of the proposed architecture can be highlighted as:

- Accuracy: PMGO-CA engages users through a consistent dialogue, before replying to users, the agent asks users about any related information that might help to satisfy their enquiries, after gathering all the required information, the agent replies based on that conversation scenario. PMGO-CA gives answers based on the user's feedback and to provide accurate answers based on user's situation
- Interaction: the new Goal-Oriented approach makes the agent highly interactive with users and takes the conversation into another level of intelligence, unlike ArabChat (Hijjawi, 2011), PMGO-CA engages users with a consistent sequential dialogue, exchanging questions and answers with users to help them achieve their goal; while ArabChat is more similar to Question & Answering system than a conversational agent.
- Responsiveness: since PMGO-CA and the user are going through a dialogue which is based on the current context, the agent matches the new user's utterance within that context only, which makes the agent much faster because it will only examine those rules within that context.
- Flexibility: PMGO-CA offers high level of flexibility in switching from one context to another, if the user input does not match any of the current context's nodes; the agent searches the whole tree to find the proper context and moves to it.

- Adaptability: PMGO-CA architecture is totally adaptable and manageable by the scripter throughout the interfaces of PMGO-CA and the tree script editor without the need to high programming skills.
- Memory management: PMGO-CA tackled long-term memory issues in conversational agents, PMGO-CA can identify users based on the information they provided in previous conversation and keeps records of the contexts discussed within these conversations
- Scripting language: PMGO-CA introduced an enhancement for the pattern matching algorithm, by using the (*) as a replacement for any character, word or null characters as described in section (4.3.2.1); this helps to reduce the number of patterns; in addition to the pattern weighting mechanism described in section (4.3.2.2), to resolve conflicting patterns issue.

4.5. Summary

In this chapter, an overview was given about Arabic conversational agents, and the issues associated with them, a new architecture for Arabic conversational agent was introduced, based on knowledge trees. Full discussion about the features of the architecture was covered showing specifications of the new PMGO-CA and facilities offered by the CA. This chapter also defined the domain used in this CA, and the justification behind using this domain. Pattern matching algorithm used in PMGO-CA was also discussed in details.

Advantages of the new architecture and the pattern matching algorithm were expressed in detail. Some examples and experiments about the new architecture were given showing the response of PMGO-CA on them. Chapter (5) will contain details of an evaluation methodology of PMGO-CA and discuss its results.

Chapter 5
PMGO-CA Evaluation

5.1. Introduction

A new architecture and scripting language for the development of an Arabic conversational Agent based on pattern matching was proposed in chapter (4) which is expected to offer high level of robustness and user-agent interaction.

Conversational Agents (like other programs) are evaluated and tested before they are used in real environments. Evaluation is typically conducted by, either a questionnaire distributed among several participants, or through monitoring the performance of the agent itself and check its response to users' utterances (Silvervarg A., 2011).

Because of their diversity, there is no standard methodology adopted by researchers to evaluate conversational agents. Furthermore, there is no particular methodology that can be applied to all types of conversational agents. However, (O'Shea et al., 2011) classified the evaluation of conversational agents in two distinctive forms; they are:

- Subjective Evaluation: this is usually focuses on the user's satisfaction criteria, such as (task ease, efficiency, user expertise, expected behaviour and future use etc.).
- Objective Evaluation: this is usually focuses on the performance of the CA in a real environment (dialogue coverage, conversation length, count of dialogue turns, task completion level, counts of errors, and speech recognition accuracy, etc.).

This chapter introduces a new methodology based on (Oshea' et al 2011) for evaluating and testing Arabic PMGO-CA proposed in Chapter 4. This evaluation shall cover the architecture, the domain information sufficiency, and the scripting language and their capability to deal with the Arabic language through the subjective and objective metrics.

The contents of this chapter can be outlined as follows:

1- Evaluation methodology of PMGO-CA conversational agent
2- Subjective and objective evaluation metrics
3- Evaluation questionnaires, conversation logs and statistics
4- Evaluation results and discussion

5.2. Evaluation Methodology

The purpose of the evaluation is to appraise the performance of the PMGO-CA according to subjective and objective metrics.

The following hypotheses are related to PMGO-CA's capability to handle the user's requests and satisfy them in real time:

H0a: The PMGO-CA can be used to successfully satisfy users' queries and allow them to achieve their goals.

H1a: The PMGO-CA's cannot be used to successfully satisfy users' queries and allow them to achieve their goals.

In order to test these two hypotheses, PMGO-CA shall meet several criteria including responsiveness, conversation length, information accessibility, ability to correct user utterance, etc. A set of metrics were chosen to evaluate these criteria and determine PMGO-CA behaviour and performance; these metrics are:

M1: Responsiveness.

Responsiveness refers to the specific ability of a system or functional unit to complete assigned tasks within a given time (Weik, 2000). It has significant impact on the overall performance of conversational agents and other software systems in general, and plays an important role in user-agent interaction, encouraging users to communicate with CAs.

This subjective metric is evaluated by domain experts through a questionnaire (explained in section 5.2.1.1). The measurement is based on their observations during the conversations they carried out with PMGO-CA; these participants are required to evaluate the speed of PMGO-CA interaction based on their expectations of a CA's performance

M2: Conversation length (Kopp, et al., 2005)

This metric is based on the number of utterances (dialogue turns) exchanged between the participants and the PMGO-CA to reach the goal. This metric is evaluated by each participant through a questionnaire and is measured by the number of utterances exchanged between the participant and PMGO-CA to achieve dialogue goals.

Normally there is a fixed number of dialogue turns for each conversational goal based on the knowledge tree, but sometimes a user may not find all of the information about their goals in one path (goal) of the tree. Consequently, they may have to switch to frequently asked questions or another goal during a conversation related to a particular topic. As a result some users may find their information in one question while others may have to go through a long dialogue before they reach their desired goal. Therefore, this metric is evaluated by questionnaire participants to gain more insight about users opinions regarding conversation length.

M3: Information Accessibility

There are several ways to seek information regarding the passport services. For example some users may choose to call the consular section, while others prefer to browse a website, other users may prefer to find more details in the official laws and regulations of passport domains.

The information accessibility metric evaluates how easy for the user to reach certain required information. Did they find PMGO-CA more suitable instead of calling the consular section and wait for staff availability? Is PMGO-CA helpful to give the right information? Would they prefer to seek other alternatives to acquire information (Like browsing websites for passport regulation guides)?

This metric is evaluated by participants based on their willingness to use PMGO-CA instead of other methods.

M4: Correcting User Utterance (Semeraro, et al., 2003)

When conversing with CAs, users often commit mistakes such as misspelling or switching to many different topics, PMGO-CA must have the flexibility to handle user's mistakes, such as correcting misspelled words and maintaining a record of the previous context to enable users to return to them directly.

Although the PMGO-CA was not designed to perform spelling corrections, this metric reflects whether enough patterns were scripted to handle and tolerate the majority of users' utterances.

In addition, this metric also reflects PMGO-CA's flexibility to maintain a record of the previous context before the user engaged in conversation that digressed away from the goal. Therefore, the PMGO-CA must tolerate users' mistakes when they go off-topic.

This metric is measured both subjectively (by questionnaire participants after conversing with PMGO-CA), and objectively (by computing the ratio between number of mistyped utterances recognized by PMGO-CA, to the total number of mistyped utterances).

M5: CA Understanding of Users' Utterances (Forbes-Rileya, et al., 2009)

This metric measures the percentage of recognised utterances (whether misfired or correctly fired) to the total number of utterances. It is evaluated both subjectively, through by questionnaires during the conversations carried out with PMGO-CA, and objectively by examining conversations logs and calculating the misfired percentage.

M6: Accuracy (Bickmore, et al., 2006)

Accuracy of scripting (keeping misfiring to minimum) evaluates the rate of correct responses. Unlike M5, accuracy measures the percentage of the responses triggered correctly as expected to the total number of the recognized utterances.

This metric reflects the accuracy of the efforts carried out to script patterns within PMGO-CA, by writing the correct patterns to handle expected utterances. This metric also reflects the ability of the strategies used to distinguish among conflicting patterns.

This metric is measured both subjectively and objectively. First, the questionnaire participants are requested to evaluate these criteria by observing PMGO-CA's responses during conversation, and second, conversation logs are examined to compute the percentage of accurate answers given to the total number of recognized utterances.

M7: Conversation Consistency

Conversation consistency is a measurement of dialogue flow and consistency. This subjective metric is evaluated by questionnaires based on participant observation during conversations.

Conversation consistency reflects the smoothness and naturalness of conversation flow. This metric also reflects the performance of tree search algorithms and context-switching mechanisms.

M8: Memory

This metric measures the performance of both short-term and long-term memory of PMGO-CA. It reflects PMGO-CA's ability to remember activated nodes during conversations, and to recognize users when they initiate new conversations in future.

This subjective metric is evaluated by questionnaire participants based on the conversations they carried out with PMGO-CA.

M9: Validity of CAs Responses

This metric evaluates whether the instructions given by PMGO-CA are identical to official domain rules and regulations, this metric is evaluated both subjectively and objectively. Subjective evaluation is performed through a questionnaire filled by domain experts based on their observations during conversations.

The objective evaluation is performed by examining the log files and calculating the percentage of valid responses given to the number of correctly-fired utterances.

Valid responses contain information identical to the official laws and regulations of the Iraqi passport domain, whilst an invalid response contains wrong or old information about these laws and regulation. Invalid responses are result of errors in knowledge engineering process, while misfired responses are resulted from patterns scripting errors.

Unlike accuracy, which is an indicator of scripting skills, validity measures on which level PMGO-CA's instructions are identical to the official laws and regulations of the domain. Therefore, validity is an indicator of good knowledge engineering.

Another hypotheses related to knowledge engineering was also studied within the evaluation of PMGO-CA.

H0b: CAs can be knowledge engineered to cover the topics and rules of a particular domain of interest.

H1b: CAs cannot be knowledge engineered to cover the topics and rules of a particular domain of interest.

To prove these hypotheses a metric is chosen to evaluate PMGO-CA coverage for the knowledge domain which is the Iraqi passport domain in this experiment:

M10: Domain Coverage

To use PMGO-CA as a tool to help citizens, it must be inclusive to all topics, regulations, laws and services related to the passport domain. This subjective metric is evaluated by domain experts based on their observations during conversations they carried out with PMGO-CA.

There are also other metrics related to the future expansion and usage of PMGO-CA which are not related to any hypotheses:

M11: Use of CA to replace human experts

This metric measures the possibility of replacing a human passport expert with PMGO-CA; this subjective metric is evaluated by domain experts based on the their observations during the conversations they carried out with PMGO-CA

M12: Use of CA as a training tool

This metric measures the possibility of using PMGO-CA as a training tool for junior consuls and provide a good knowledge base about the domain. This subjective metric is evaluated by domain experts based on their observations during the conversations they carried out with PMGO-CA.

5.2.1. Subjective Evaluation

Subjective evaluation metrics are rated by participants based on their observations during their conversations with the PMGO-CA. These metrics are evaluated using a questionnaire with questions related to these metrics. Participants are required to answer these questions with a rating between (1) and (5); as follows:

(1) Weak

(2) Acceptable

(3) Good

(4) Very good

(5) Excellent

The rating system within this questionnaire was inspired by the questionnaire designed by (O'shea, 2012) to evaluate the SCAF framework.

Table (5-1) lists the subjective metrics with explanations on how they are evaluated through the questionnaire.

Metric		Evaluator	Evaluation method
M1	Responsiveness	Domain experts	1 – 5 Rated based on observation during conversations. (1) indicates low responsiveness (5) indicates high responsiveness
M2	Conversation length	Domain experts	1 – 5 Rated based on observation during conversations. (1) indicates a long conversation (5) indicates very short conversation
M3	Information accessibility	Domain experts	1 – 5 Rated based on observation during conversations. (1) indicates difficult accessibility (5) indicates very easy accessibility
M4	Correcting user utterance	Domain experts	1 – 5 Rated based on observation during conversations. (1) Indicates inability to tolerate users mistakes (5) Indicates great ability to tolerate users mistakes
M5	CA's understanding of users' utterances	Domain experts	1 – 5 Rated based on observation during conversations. (1) Indicates poor understanding of users utterances (5) Indicates excellent understanding of users utterances
M6	Accuracy	Domain experts	1 – 5 Rated based on observation during conversations. (1) Indicates low percentage of accurate answers (5) Indicates high percentage of accurate answers
M7	Conversation consistency	Domain experts	1 – 5 Rated based on observation during conversations.. (1) indicates inconsistent dialogue (5) indicates highly consistent dialogue
M8	Memory	Domain experts	1 – 5 Rated based on observation during conversations. (1) indicates strong memory (5) indicates weak memory

Metric		Evaluator	Evaluation method
M9	Validity	Domain experts	1 – 5 Rated based on observation during conversations. (1) indicates low percentage of accurate responses (5) indicates high percentage of accurate responses
M10	Domain coverage	Domain experts	1-5 Rated based on observation during conversations. (1) indicates low coverage of domain topics (5) indicates full coverage of domain topics
M11	Use of CA to replace human experts	Domain experts	1 – 5 Rated based on observation during conversations. (1) indicates low possibility to replace experts with PMGO-CA (5) indicates high possibility to replace experts with PMGO-CA
M12	Use of CA as a training tool	Domain experts	1 – 5 Rated based on observation during conversations. (1) low possibility of using PMGO-CA to train junior consuls (5) high possibility of using PMGO-CA to train junior consuls

Table 5-1 subjective evaluation metrics

5.2.1.1. Questionnaire

The evaluation was conducted through a questionnaire designed especially for PMGO-CA, the questionnaire starts with some explanation and instructions about the test and domain, and how to test and evaluate the agent. It also requests some information about the age, gender, status, and experience of the participants themselves (the aim of this personal information is to give the researcher the chance to evaluate the participants experience themselves. However no personal identification data is requested or stored). The questionnaire included questions concentrated on the subjective metrics discussed in section (5.2.1). Questionnaire participants were required to read the instructions thoroughly and rate the questionnaire items form (1-5), where (1) shows poor feedback and (5) shows excellent feedback. Table (5-3) lists the evaluation metrics and the related question.

Evaluation metric		Related question
M1	Responsiveness	Rates the speed of PMGO-CA when responding to questions?
M2	Conversation length	Rates the length of the conversation carried out with PMGO-CA in terms of the number of utterances exchanged between users and PMGO-CA
M3	Information accessibility	Rates how PMGO-CA would be more useful to get information than other methods such as phone calls or browsing the website of the Iraqi ministry of foreign affairs
M4	Correcting user utterance	Rates how PMGO-CA tolerated the spelling mistakes encountered during the conversation
M5	CA's understanding to user's utterance	Rates the level of PMGO-CA understanding to utterance during the conversations, in terms of the percentage of the number of utterances not understood, to the total number of utterances typed during the conversation.
M6	Accuracy	Rates the accuracy of CA's answers during the conversation, in terms of the number of PMGO-CA's responses to the expected topics, to the total number of PMGO-CA's responses.
M7	Conversation consistency	Rates the dialogue flow of PMGO-CA and the flexibility in switching the conversation from one topic to another
M8	Memory	Rates the long-term and short-term memory of PMGO-CA
M9	Validity	Rates the validity of PMGO-CA instructions according to the Iraqi passport domain laws and regulations, in terms of percentage of number of correct instructions, to the total number of instructions provided by PMGO-CA.
M10	Domain coverage	Rates whether PMGO-CA completely covers all domain topics with the exact laws and regulations
M11	Use of CA to replace human experts	Rates the possibility to replace a human expert with PMGO-CA
M12	Use of CA as a training tool	Rates the possibility of using PMGO-CA as a training tool for junior diplomats

Table 5-2 questions related to evaluation metrics

A copy of the questionnaire and instructions can be found in Appendix (3) of this thesis.

5.2.1.2. Evaluation Participants and Experimental Methodology

It was not easy to find experts within the passport domain to test and evaluate the PMGO-CA. The researcher managed to select only (10) qualified participants who are experts in the Iraqi passport domain to fill out an electronic version of the questionnaire sent to their emails.

Participants were asked to do the following:

1- Log on to the online system. Using the web site www.iraq–pass–ca.net/
2- Converse with the system with questions regarding passport issues and topics, (Passport issue, Extending passport validity, Lost and stolen passports, Passport damage, Travel documents)
3- Use the modern Arabic language, and avoid colloquial Arabic words as much as possible.
4- Use the dialogue as if they were Iraqi citizens living abroad
5- Initiate several conversations with the system to be familiar with it before evaluating and making any observations or judgements.
6- Fill out this questionnaire with their information, which are used to evaluate the process without disclosure of these information,
7- Submit the questionnaire by email once they are completed.

The reason for selecting domain experts as participants is that normal users cannot evaluate PMGO-CA precisely. Domain experts are totally familiar with the Iraqi passport domain, so they can evaluate PMGO-CA performance and validity of answers better than non-expert l participants. The author believes that selecting experts to test the PMGO-CA was successful, as their professional and continuous testing helped to improve and develop the scripting and added more rules, questions and reports to the CA.

5.2.2. Objective Evaluation

Objective evaluation is used to evaluate the expected performance of PMGO-CA to achieve its' design objectives to offer online help to users covering all topic related to the Iraqi passport domain. Conversation logs were stored in PMGO-CA's database, those logs contain all conversations carried out by users with the CA, they were used to measure the metrics listed in table (5-3)

	Metric	Evaluation method
M5	CA understanding of user utterances	This metric is measured objectively by examining conversation logs and calculating the percentage of recognised utterances given to the total number of utterances.
M6	Accuracy	This metric is measured objectively by examining conversation logs and calculating the percentage of correctly answered utterances to the number of recognised utterances
M9	Validity	This metric is measured objectively by examining conversation logs and calculating the percentage of valid responses to the number of correctly answered utterances.

Table 5-3 Objective evaluation metrics

5.3. Evaluation Results and Discussion

This section presents the experimental results and their discussion. Subjective evaluation results are based on the level of agreement with each metric (M) by means of a five-point rating scale, as described in section (5.2.1).

Table (5-4) displays the number of participants when rating each metric. For example, with respect to responsiveness, eight participant gave a rating of 5 (Excellent), two participants gave a rating of 4 (very good).

Metric	**Rating frequency**					**average**
	5	**4**	**3**	**2**	**1**	
M1: Responsiveness	8	2	0	0	0	4.8
M2: Conversation length	4	5	0	1	0	4.2
M3: Information accessibility	4	3	3	0	0	4.1
M4: Correcting user utterance	0	6	4	0	0	3.6
M5: CA's understanding of user's utterance	3	4	2	1	0	3.9
M6: Accuracy	6	4	0	0	0	4.6
M7: Conversation consistency	4	5	1	0	0	4.3
M8: Memory	0	3	4	3	0	3.0
M9: Validity	6	2	1	1	0	4.3
M10: Domain coverage	3	4	2	1	0	3.9
M11: Use of CAs to replace human experts	1	5	3	1	0	3.6
M12: Use of CA to train junior consuls	0	5	2	2	1	3.1

Table 5-4 subjective evaluation frequency

This work uses the same significance test used by (O'shea, 2012). Results are measured for significance using the Wilcoxon Signed-Ranks test. The assumption made for the Wilcoxon test is that the variable being tested is symmetrically distributed about the median, and that the responses are symmetrically distributed about (Good), a hypothesis that users assess each metric as agreeable can be tested. Users that assess a metric as agreeable will give a rating more than 3. The null and alternative hypotheses are stated as follows:

H0: the median response is 3.

H1: the median response is more than 3.

A (1 tail) test set at a significance level of 5% was proposed. Example analysis explaining statistical significance can be found in appendix (4) of this thesis. Table 5-5 summarises the opinion of each metric from the perspective of the ten participants in PMGO-CA evaluation questionnaire.

Metric	User opinion
M1: Responsiveness	PMGO-CA was responsive and interactive with users
M2: Conversation length	It doesn't take long conversations to reach users goals
M3: Information accessibility	It's easy to use PMGO-CA to obtain information about the Iraqi passport domain
M4: Correcting user utterance	PMGO-CA can handle users' mistakes during conversations
M5: CA's understanding to user's utterance	PMGO-CA can understand and process users' utterances
M6: Accuracy	PMGO-CA responses are accurate
M7: Conversation consistency	The conversations flow is consistent and organized
M8: Memory	PMGO-CA memory can remember previous contexts and user information
M9: Validity	PMGO-CA responses are valid according to official laws and regulations of Iraqi passport domain
M10: Domain coverage	PMGO-CA covers the topics of Iraqi passport domain
M11: Use of CAs to replace human experts	PMGO-CA cannot be used as replacement of human experts in Iraqi passport domain
M12: Use of CA to train junior consuls	PMGO-CA cannot be used as a training tool for junior specialists in Iraqi passport domain

Table 5-5 User's opinion about PMGO-CA

Table (5-6) shows the results of objective evaluation by examining conversation logs and gathering related statistics.

Metric		Statistics	percentage
M5	CA understanding of user utterances	Total number of utterances: 1120 recognized utterances: 870	77 %
M6	Accuracy	recognized utterances: 870 Correctly answered utterances: 620	71 %
M9	Validity	Correctly answered utterances: 620 Valid responses: 520	84 %

Table 5-6 Results of objective metrics

Considering the results of objective evaluation in table (5-6) and the results of subjective evaluation in table (5-4) it is noticeable that M5 and M9 scored similar results to their subjective evaluation and objective evaluation. While the results of the metric M6 in subjective evaluation by domain experts (4.6) which equals (92%) differs from its objective evaluation results gathered from log files (71%). The reason for this difference is that users often type the same utterance repeatedly when PMGO-CA did not fire the correct response, and all these trials are shown in the log files. Once they became familiar with the CA they can have a better judgement of what the CA can understand. For this reason, the researcher decided to rely more on the results of evaluation questionnaire instead of log files statistics.

The average results of the subjective evaluation shown in table (5-4) were converted to a percentage scale for consistency, Tables (5-5) and (5-6) shows these results with elaboration on their outcomes.

Evaluation results and outcomes			
Metric		Score	Outcome
M1	Responsiveness	96%	The score indicates high level of agent performance, in other words user utterances are processed and answered in milliseconds
M2	Conversation length	84%	The score reflects very good feedback on the time consumed by PMGO-CA to converse with users and answer their questions

Evaluation results and outcomes			
M3	Information accessibility	82%	The score reflects a good level of user satisfaction about using PMGO-CA as a method to access information regarding the passport domain
M4	Correcting user's utterance	72%	The score reflects a good level of PMGO-CA's ability to handle user's mistakes, to improve this rate many more patterns have to be add to PMGO-CA's rules
M5	CA's understanding to user's utterance	78%	The score indicates that PMGO-CA can understand most users utterances', the score can be further improved by adding more patterns to the agent's rules
M6	Accuracy	92%	The score reflects very low level of misfired replies
M7	conversation consistency	86%	The score indicates that PMGO-CA is able to maintain consistent dialogue flow through the conversation, this also reflects the effectiveness of context switching mechanisms
M8	Memory	60%	The score shows good level of memory management however more work needs to be achieved to improve the memory of PMGO-CA
M9	Validity	86%	The score shows very good level of valid responses in PMGO-CA which reflects very good effort to knowledge engineering
M10	Domain coverage	78%	The score shows high level of covering all topics of Iraqi passport domain
M11	Use of CA to replace human experts	--	The results of this metric was discarded in the overall results of PMGO-CA evaluation due to the lack of sufficient information and methods to estimate it
M12	Use of CA as a training tool	--	The results of this metric was discarded in the overall results of PMGO-CA evaluation due to the lack of sufficient information and methods to estimate it

Table 5-7 results of PMGO-CA evaluation

In relation to evaluation hypotheses, the results have shown that:

H0a: The PMGO-CA can be used successfully to satisfy users' queries and allow them to achieve their goal.

This hypothesis was proven through this experiment, due to the outstanding results of the (9) metrics: responsiveness, conversation length, information accessibility, correcting user’s utterance, conversation consistency, memory, CA’s understanding of user’s utterance, accuracy and validity.

H0b: CA’s can be knowledge engineered to cover the topics and rules of a particular domain of interest

The results of the metric “**M10:** Domain coverage” showed high level (78%) of domain coverage including most domain topics and FAQs related to the Iraqi passport services.

The results of the metrics (M11) and (M12) were disregarded in the overall results of PMGO-CA evaluation due to the lack of sufficient information and proofs to evaluate them.

Using PMGO-CA as a replacement to human experts requires more than a questionnaire and a pilot study, PMGO-CA must be used first in real environment under the supervision of domain experts, before it can be proved useful to replace human experts or not.

The same thing applies for using PMGO-CA to train junior consuls, although it can offer quick help to junior consuls when needed, it’s too early to rely on it a training tool.

5.4. Conversation Samples

This section provides some samples of Arabic conversations carried out in the testing of PMGO-CA and their English translation, these samples were taken from logs of conversations carried out with (10) participants.

Conversation sample (5-1) demonstrates a typical conversation for a lost passport. As shown in the dialogue the user tells the system about his case; and the system replies with a set of procedures that should be followed when a passport is lost and then asks the user whether he/ she has completed these procedures or not. In this dialogue the user replies that he had completed these procedures, then the system asks another question about the type of residence; because in the IPS domain, there are different procedures for each type of residency which are, temporary residency and long-term residency; the user replies that he/ she is a tourist on a short-term residency; then the system asks the user whether he/ she has valid Iraqi ID and Citizenship certificates or at least copies of them. As short-term residents are usually given a travel document or transit passport to allow them to get back to Iraq, this type of travel document does not require the presence of the original documents, while a standard passport application requires their presence. Finally the agent replies to the user with instruction on how to obtain the travel document.

You: I lost my passport, what shall I do? System: When a passport is lost, you must inform the police and publish an advertisement in a journal about losing your passport, then you can visit the consulate to apply for a new passport, Have you completed all these procedures? You: Yes, I have System: Do you live there for short-term and you want to go back to Iraq? You: Yes, I'm a tourist. System: do you have valid personal ID and Citizenship certificate, or copies of them? You: Yes System: As mentioned before, you can send a written request to the consulate to obtain a transit passport, which requires copies of your Iraqi documents and 4 personal pictures with white background.	انت: فقدت جوازي ماذا افعل النظام: عند فقدان الجواز يتطلب تقديم بلاغ للشرطة عن حالة الفقدان ، نشر الفقدان في الصحيفة و بعد ذلك يتم مراجعة البعثة لاصدار جواز جديد ، هل قمت بهذه الاجراءات ؟ انت: نعم قمت بذلك النظام: هل انت مقيم بصورة مؤقته و تريد العودة الى العراق ؟ انت: نعم انا سائح النظام: هل تملك هوية الاحوال المدنية و شهادة الجنسية العراقية النافذه او صور ملونه منهما ؟ انت: نعم النظام: كما` ذكرنا سابقاً يمكنك تقديم طلب خطي الى البعثة بشأن اصدار جواز مرور، يتطلب ذلك وجود نسخ من الوثائق العراقية و صور شخصية ملونة بخلفية بيضاء عدد (4)

Conversation sample 5-1

Conversation sample (5-2) demonstrates another conversation about the same topics of the previous example, but this time the user replies that he/ she did not complete the legal procedures of lost passports. This time the system gives a different reply and instructs the user to complete these procedures and apply for an alternative passport or a travel document to get back to Iraq.

You: I lost my passport, what shall I do?	انت: فقدت جواز سفري ماذا افعل؟
System: When a passport is lost, you must inform	النظام: عند فقدان الجواز يتطلب تقديم بلاغ للشرطة عن حالة

the police and publish an advertisement in a journal about losing your passport, then you can visit the consulate to apply for a new passport, Have you completed these procedures? You: No, I have not. System: You must complete these procedures, then you may apply for a new passport or obtain a transit passport to go back to Iraq.	الفقدان، نشر الفقدا ن في الصحيفة وبعد ذلك يتم مراجعة البعثةلاصدارجواز جديد، هل قمت بهذه الاجراءات؟ انت: لا لم اقم بذلك النظام: يجب اكمال اجراءات فقدان الجواز قبل الشروع باصدارجواز بديل، بعد ذلك يمكنك اصدار جواز جديد او اصدار جواز مرور سريع لغرض العودة الى العراق

Conversation sample 5-2

Conversation sample (5-3) is also related to the above examples, but it demonstrates the ability of the agent to switch from a context to another. In this sample when the system asks the user about lost passport procedures, the user replies with an utterance related to another goal (Context): “I want a transit passport”, in this case the agent performs search to find the context related to this utterance and moves the conversation to that context.

You: I lost my passport, what shall I do? System: When a passport is lost, you must inform the police and publish an advertisement in a journal about losing your passport, then you can visit the consulate to apply for a new passport, Have you completed these procedures? You: I want a travel document. System: this is a different situation however I can help you with that, travel documents are issued for citizens who have lost their passport and wish to go back to Iraq urgently, they are also issued for Iraqi student’s children to travel back to Iraq,	انت: فقدت جوازي ماذا افعل النظام: عند فقدان الجواز يتطلب تقديم بلاغ للشرطةعن حالة الفقدان، نشر الفقدان في الصحيفة وبعد ذلك يتم مراجعة البعثة لاصدار جواز جديد، هل قمت بهذه الاجراءات؟ انت: اريد جواز مرور النظام: هذا موضوع مختلف لكن على اي حال يمكنني مساعدتك في ذلك، يصدر جواز المرور لمن فقد جواز سفره، او في الحالات التي توجب المواطن على العودة للعراق بشكل طوعي، ويصدر لاطفال الطلبة العراقيين للسفر الى العراق

Conversation sample 5-3

5.5. Problems Associated with Pattern Matching

Despite having good advantages in helping users and covering all Iraqi passport topics, the process of scripting patterns remains cumbersome due to the nature of patterns themselves. There are many topics that a scripter must consider before writing patterns within a specific domain, such as:

- The use of generic, patterns instead of non-generic, more specific patterns:
  Although adding generic patterns (explained in section 4.3.2) may save time and effort, this may cause accuracy drawbacks, a generic pattern that matches many utterances may save the scripter a lot of time and effort and gives an extraordinary results and responsiveness, but in the meantime it might cause high percentage of misfired answers. Writing generic patterns might be desirable for some knowledge domains. But, generally speaking most domains require patterns to be specific and restrictive. On the other hand, if a scripter tends to be extremely specific in writing patterns, they might lose the advantage of flexibility and responsiveness, but the agent response would likely be highly accurate. This decision of using generic patterns is to be made by the scripter according to the nature of the knowledge domain itself.

- Patterns Conflict
  Generally speaking, CAs contain set of rules to be evaluated against a users' utterance in order to fire a response, the number of these rules varies from one domain to another. When adding too many rules to the agent, some of these rules will eventually contain patterns that conflict with patterns of other rules which lead to misfired responses by the agent. Although pattern matching CAs contain conflict resolution strategies but these do not guaranty optimal results. The researcher was so keen to avoid this drawback when scripting the rules and patterns; however he faced this problem during testing the agent. This led him to do some fundamental changes to sort it out. Still, this conflict might take place during housekeeping and updating the agent, Therefore, domain scripters need to be careful when updating the agent to avoid the conflict. More details about conflict resolution strategies can be found in chapter (4) section (4.3.2.2)

- Exact definition for spacing and other characters in patterns

  Scripting patterns is complicated in Arabic language, due to the use of colloquial language in conversations with Arabic CAs, the scripter must include many patterns to deal with the variation of words; or write generic patterns to apply to many utterances and have accuracy drawbacks. More details about this can be found in chapter (4) section (4.3.2.1). PMGO-CA with its current state holds more than (800) patterns distributed over more than (50) nodes.

5.6. Summary

In this chapter, the researcher evaluated the newly constructed PMGO-CA using a set of hypotheses and associated metrics. The evaluation methodology and metrics (both subjective and objective) were explained thoroughly. To achieve the evaluation, the researcher designed a questionnaire list to asses various aspects of PMGO-CA, and selected the participants to implement the evaluation.

The evaluation results in general showed good feedback on using PMGO-CA to satisfy users' enquiries with very good coverage of the Iraqi passport domain topics and procedures.

From the results, it was clear that PMGO-CA was responding positively to users' utterances with high accuracy (92%). This means that misfiring was kept to the minimum. Users also considered the conversations carried out with PMGO-CA to be simple, consistent and short by the results of conversation length (84%) and conversation consistency (86%).

PMGO-CA also proved to be a good method to access information regarding the Iraqi passport domain by the results of the information accessibility metric (82%). The flow of conversation was also smooth and the agent managed to reach the goal of the user within a very reasonable time by the results of the responsiveness metric (96%).

Results have also shown a high level (86%) in the validity of responses given by PMGO-CA and covering very good percentage of the topics related to the domain.

However, PMGO-CA showed less ability to understand user’s utterances, it was obvious through the results of the metric (M5) which scored (78%). This is mainly due to the use of colloquial Arabic language while conversing with PMGO-CA; which has no standard spelling or grammatical structure. This requires the scripting of many patterns to handle users’ utterances, making the housekeeping of PMGO-CA very labour intensive.

Chapter 6
Semantic Goal-Oriented Conversational Agent (SGO-CA)

6.1. Introduction

Architecture for the development of an Arabic goal-oriented conversational agent was introduced in chapter 4. This architecture was used to construct a Pattern Matching Goal-Oriented Conversational Agent (PMGO-CA). It was then tested and evaluated for its' viability and performance in chapter 5.

Although evaluation of the PMGO-CA showed good results, the researcher observed that it is difficult to maintain and script (i.e. the current knowledge tree holds (5) main contexts and more than (70) frequently asked questions, and contains more than (800) patterns). It is time consuming and complex sometimes to write enough patterns to handle all potential users' utterances. Furthermore, when domain rules or regulations are changed it would be cumbersome to re-script all these patterns, especially for large domain CAs, not to mention the conflicts that might occur between rules sometimes during the maintenance.

A new approach to developing English CAs was attempted recently by (O.Shea, 2014), using semantic relations between texts to compute similarity between user's utterances and the sentences defined within CA's rules. This approach is believed to offer a high level of intelligence and minimises the effort required to manage the scripting of conversational agents. However, this attempt was conducted using English language for a limited prototype domain. No trials were conducted for the Arabic language.

This chapter offers a novel architecture to construct a goal-oriented semantic conversational agent for the Arabic language (SGO-CA) using the semantic structure of Arabic WordNet and SUMO ontology as an information source to measure the similarity.

SGO-CA was constructed using the same methodology used within PMGO-CA explained in section (4.1.1). However, a complete new semantic similarity engine (described in section 6.2.2) to measure the similarity between user utterances and prototype sentences was used to replace the pattern-matching engine used in PMGO-CA.

To construct the new SGO-CA, the researcher adapted and modified some well selected word and sentence similarity measures and strategies covered in the literature (The word similarity measure (Li, et al., 2003), the AWSS word similarity measure (Almarsoomi, et al., 2013) and the STASIS sentence similarity measure (Li, et al., 2006)). A novel measure and measuring tools was also applied to implement the CA. This novel SGO-CA was tested and evaluated for its viability using those tools through series of experiments in chapter (7) of this thesis.

The novel contributions of this chapter can be summarised as follows:

- A new methodology of developing an Arabic Semantic Goal- Orientated Conversational Agent. (SGO-CA)
- A new word similarity measure to be used in SGO-CA.
- Adapting previous measures in sentence similarity for use in SGO-CA for the Arabic language.
- Utilising the mapping between WordNet and SUMO ontology to develop an information source for similarity measurement.
- Introducing a new equation for sentence difference to be incorporated in with the overall similarity between two sentences
- Inclusion of Arabic function words in similarity measurement.
- The construction and implementation of SGO-CA for the Iraqi passport domain.
- A set of software tools used to construct SGO-CA to allow for future generalisation.

6.2. SGO-CA Overview

As mentioned before, the new SGO-CA was constructed based on the same architecture introduced in chapter (4). The only amendment to the architecture was in replacing the pattern matching engine with the semantic similarity engine as shown in figure (6-1). SGO-CA uses an approach derived from the STASIS method (Li, et al., 2006) (covered in section 3.5) to calculate word and sentence similarity between user's utterances, and the sentences kept within CA's nodes in the knowledge tree introduced in chapter (4).

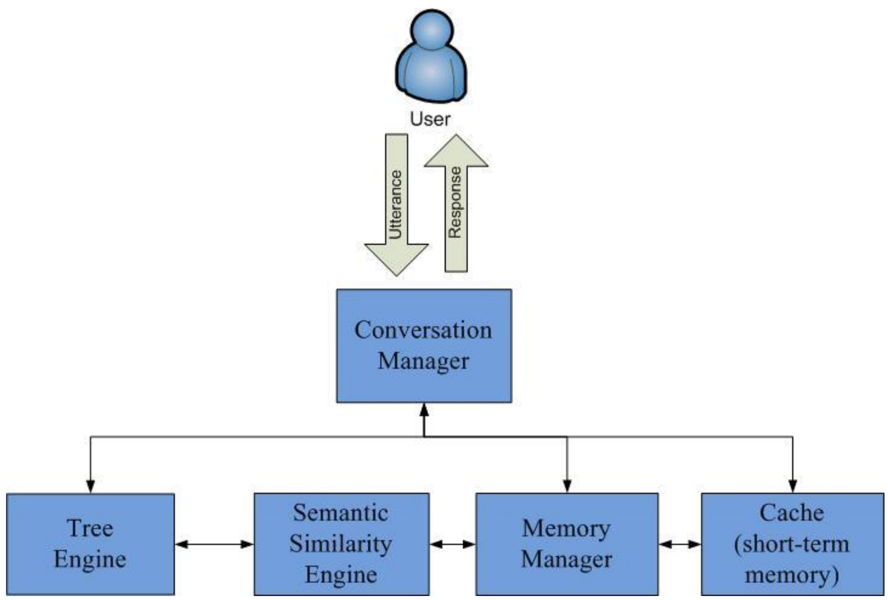


Figure 6-1 SGO-CA architecture

SGO-CA follows the same design methodology used in PMGO-CA, including the interface for the user-agent interaction, knowledge-tree for structuring goal orientated knowledge, the tree search algorithm and the memory management algorithm. These are all identical to the PMGO-CA. In short, this chapter is only focusing on the novel components introduced within SGO-CA.

6.2.1. Semantic Similarity Engine

The semantic similarity engine is the heart of SGO-CA. This engine takes a user utterance and a regular answer and calculates the similarity between them. It has access to the information source (described in section 6.2.2) and the corpus (described in section 6.3.2.3) and uses them to calculate the similarity between the user's utterance and regular answers by using the similarity measures described in section (6.3).

6.2.2. Information Sources

Previous research on semantic similarity between texts (Li, et al., 2003) (Almarsoomi, et al., 2013) proposed the use of WordNet as an information source to evaluate the similarity between two words. In both Arabic and English WordNets (Black, et al., 2006) (Miller, et al., 1993) words are classified according to their part of speech into four categories: nouns, verbs, adjectives and adverbs.

According to (Miller, et al., 1993), there are five relations between words in the WordNet database; these are:

- Hyponymy or (is-a) relation: this relation connects nouns or verbs to other nouns or verbs they are related to for example "man" is "person"
- Troponymy (manner-name): This relation relates two verbs together. Troponymy for verbs is the same as hyponymy for nouns, although the resulting hierarchies are much shallower: The troponymy relation between two verbs can be expressed such that the first verb is related to the second in some particular manner. Troponyms of communication verbs often encode the speaker's INTENTION or motivation for communicating, as in examine, confess, or preach, or the MEDIUM of communication: fax, e-mail, phone, telex.

- Synonymy: This relation connects two identical nouns or verbs in meaning such as “close” and “shut”.
- Meronymy: The part-whole relation holds between synsets like “chair” and “مسند”“backrest”, “مقعد” “seat” and “ساق” “leg”. Parts are inherited from their super ordinates: if a chair has legs, then an armchair has legs as well.
- Antonym: It is an opposite relation between two words like “fast” and “slow”

In WordNet, nouns and verbs are organised in a hierarchical form thus forming a tree of IS-A relations. This hierarchical structure was used by researchers (Li, et al., 2003) (Almarsoomi, et al., 2013) to evaluate the similarity between two words.

In the English language, verbal nouns (nouns derived from verbs) have similar grammatical structure of the verb itself for example “singing” as a verb has the same grammar as the verbal noun “singing”; but this is not the case in Arabic language where verbal nouns have different grammatical structure than Arabic verbs for example “gained” “حصول” and "حصل", therefore in Morphological analysis tools such as AraMorpth (AraMorph, 2003) they are given a different part of speech, which cause a verbal noun and a noun to exist in different parts of the WordNet tree. This has a negative impact on similarity measurement.

In addition, the hierarchical structure of both English and Arabic WordNet covers only nouns and verbs. Other parts of speech such as adjectives and adverbs are not linked to super ordinate words. While adjectives are related to other adjectives using the “Synonymy” and “Antonymy” relations.

Another problem with the Arabic language is that the same word (regardless of its part of speech) might have different meanings, and therefore might appear in different locations in the WordNet tree, this requires the application of a word sense disambiguation method. This problem is common in almost all languages, and makes measuring the similarity difficult.

Previous research in word similarity (Li, et al., 2003) (Almarsoomi, et al., 2013) used datasets of nouns to test and evaluate similarity measures. But to run a conversational agent, other parts of speech such as verbs and adjectives must also be considered in similarity measurement.

In addition to the relations between words in WordNet, words are also mapped to a particular concept in SUMO ontology (Pease, 2011) as illustrated in figure (6-2).

SUMO (Suggested Upper Merged Ontology) is a collection of well-defined and well-documented concepts, interconnected into semantic network and accompanied by a number of axioms. The concepts range from very general ones, such as Quantity, to very specific, such as Bird. The axioms mostly reflect common-sense notions that are generally recognized among the concepts.

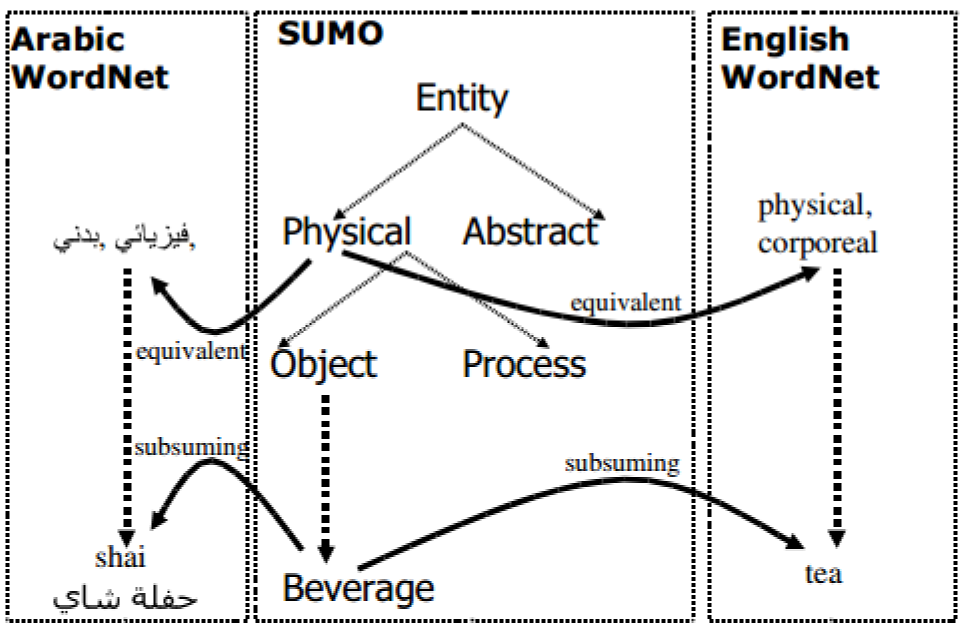


Figure 6-2 SUMO mapping to WordNet (Black, et al., 2006)

Concepts in SUMO are organised into a single hierarchy with a root of “Entity”, representing the most general concept. The first two levels of the hierarchy are depicted in figure (6-3).

Entities are divided into physically existent (Physical), and conceptual (Abstract). Physical things are further distinguished as objects and processes, etc.

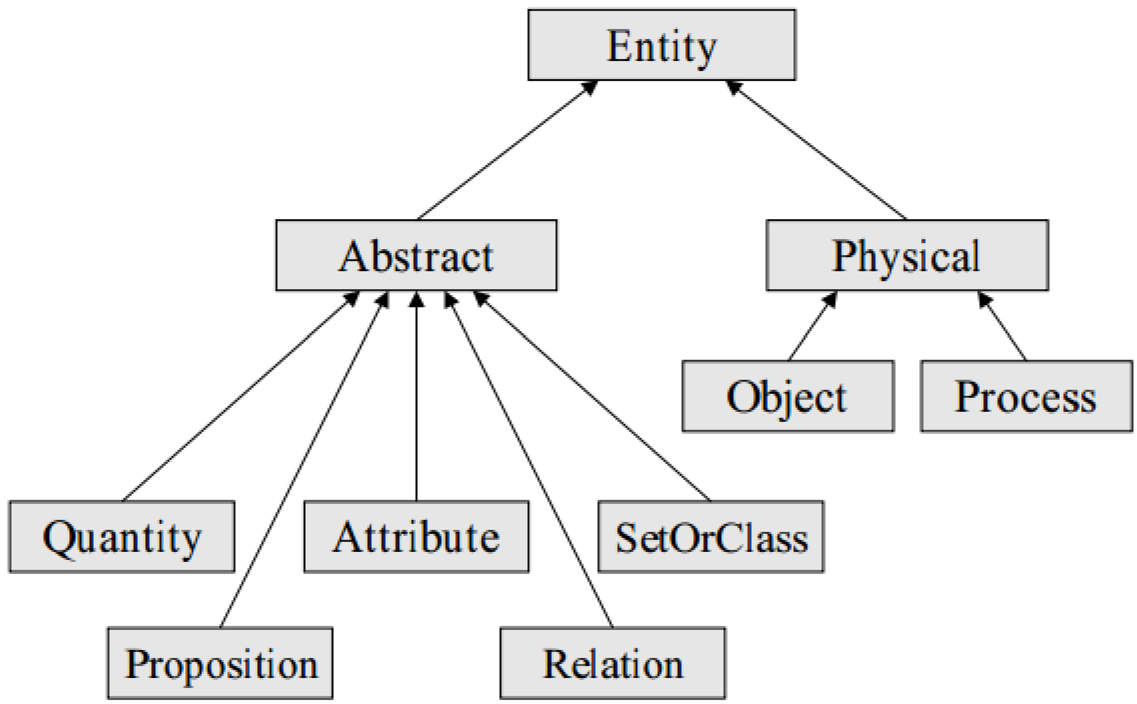


Figure 6-3 Portion of SUMO ontology (Sevcenko, 2003)

Subclasses of a class are usually mutually exclusive, i.e. they do not share common instances. For example, nothing can be both an abstract and a physical, neither both an object and a process. This property is explicitly specified in SUMO.

One of the drawbacks of SUMO is its relatively low coverage that does not allow its deployment for open-domain applications. It also lacks a connection between its concepts and natural language words. These limitations have been partially overcome by connecting SUMO to the WordNet lexicon. (Sevcenko, 2003)

Given the above mentioned limitations and issues associated with WordNet, this research makes advantage of words mapping to the SUMO ontology, this significantly helps to enhance the similarity measurement between sentences because Arabic words are mapped to their equivalent or subsuming SUMO concepts regardless of their part-of-speech. Unlike WordNet which classifies words according to their part-of-speech.

Despite being a valuable source of linguistic information when mapping to the SUMO ontology and English WordNet, the researcher believed that Arabic WordNet has some limitations when applied for conversational agents, these are:

- Slow performance: Sometimes it takes several seconds to look up a word in WordNet browser, for example looking up the word "ذهب" "Gold" takes about a second to view the available word senses and another 4 seconds to expand the selected word sense.
- Morphological ambiguity: Due to the diacritics used in Arabic language, Arabic WordNet browser cannot distinguish between some words when they are typed without diacritics.
- Word sense ambiguity: Same words have different meaning in different contexts for example AWN browser displays (8) word senses for the word (ذهب), one of these word senses means (gold) while other word senses are variety of senses for the verb (leave).
- Classification of Arabic words according to their part of speech. (Nouns, Verbs, Adjectives and adverbs) (Sevcenko, 2003)
- Lack of function words classification in WordNet databases (Sevcenko, 2003): Despite their importance and direct effect on the meaning, function words are not defined and incurred in the AWN.

- Limited number of words incurred in the Arabic WordNet: As the number of words does not exceed 24,000 words. (The Global WordNet Association, 2014)

Due to the incompleteness and slow performance of Arabic WordNet browser, the lexical tree created by the researcher in this work (described in section (6.2.2)) combined both the hierarchical structure of the SUMO ontology and the mapped Arabic words.

More than (2000) Arabic words were added to the new lexical tree; most of these words were related to the Iraqi passport domain, while others were frequently used words in daily life and are not strictly related to the domain. These words were added according to their mapping to the SUMO ontology.

Therefore in SGO-CA, the lexical tree is the main information source used in calculating semantic similarity between words, based on the path length between the words and the depth of words within the lexical tree. More details about the lexical tree can be found in section (6.2.2)

6.3. Methodology for the Application of Semantic Similarity within SGO-CA

In SGO-CA, semantic similarity is calculated between user's utterances and sentences stored within the rules of SGO-CA called "regular answers" using semantic similarity measures discussed in the upcoming sections.

To understand and explain the adaptation of semantic similarity measures in semantic conversational agents, this section is focusing on how two sentences are semantically measured within SGO-CA, assuming that one of the sentences is the user's utterance and the other is the "regular answer". Section (6.5) explains in detail where the" regular answers" are encoded and stored within the SGO-CA.

6.3.1. Word Similarity

As explained in section (3.4), word similarity measures were studied by many researchers. (Li, et al., 2003) presented different strategies to calculate the semantic similarity using multiple information sources, (i.e. the shortest path length, depth and local density). The strategy obtained the best result was the one that implemented non-linear functions containing both the shortest path and depth. This strategy also obtained the best performance among the reported word similarity measures by using the following equation

$$\mathrm{sim}(\mathrm{W1}, \mathrm{W2}) = e^{-\alpha l} . \frac{e^{\beta h} - e^{-\beta h}}{e^{\beta h} + e^{-\beta h}} \qquad (6\text{-}1)$$

Where W1 and W2 are two words to be compared, (*l*) is the shortest path between two words in the lexical hierarchy of WordNet; (*h*) is the depth of the concept that subsumes the two words, (α) is a constant and (β) is a smoothing factor. The same equation was used by (O'Shea, et al., 2010) to develop a Semantic Conversational Agents Framework (SCAF) for the English language.

This equation was originally developed and evaluated for English language by (Li, et al., 2003) using the English WordNet (1.6) as an information source (not the Arabic WordNet). Their experiments covered different strategies and tested several hypotheses; therefore it was selected as basis in this work.

More recently (Almarsoomi, et al., 2013) performed a study on word similarity measurement for the Arabic language and developed the Arabic Word Semantic Similarity (AWSS) also linking the path length and depth of Arabic words for Arabic WordNet (3.0).

The AWSS algorithm measures the similarity between two words using the equation:

$$sim(W1, W2) = e^{(-\alpha * l)} * tanh(\beta * d) \qquad (6\text{-}2)$$

Where:

- W1 and W2 are two words to be compared
- α and β are the length and depth factors respectively which signify the contribution of path length between two words, and the depth of the Least Common Subsumer (LCS). The values of Alpha and Beta where set by (Almarsoomi, et al., 2013) to $\alpha = 0.162$ and $\beta = 0.234$.
- (d) is the depth of LCS
- (*l*) Is the length of the shortest path connecting W1 and w2, (*l*) can be calculated as:

$$l = d1 + d2 - (2 * d) \quad (6\text{-}3)$$

Where (d1) is the path length between W1 and the root of the lexical tree and (d2) is the path length between w2 and the root of lexical tree, and (d) is the path distance between the LCS and the tree root.

The Least Common Subsumer is the concept which subsumes two words, in other words LCS is the first common concept between W1 and W2. Figure (6-4) demonstrates the concepts of depth and length. Taking the two words “father” "اب"and “grandparent” "جد" for example the path length (*l*) between these two words is the count of the links connecting both words which is (6) and the depth of the LCS “ancestor” "سلف" which subsumes both words is the count of the links between this LCS and the root of the tree “entity” which is (6) in this example. Appendix (6) of this thesis illustrates an example for the AWSS word similarity calculation.

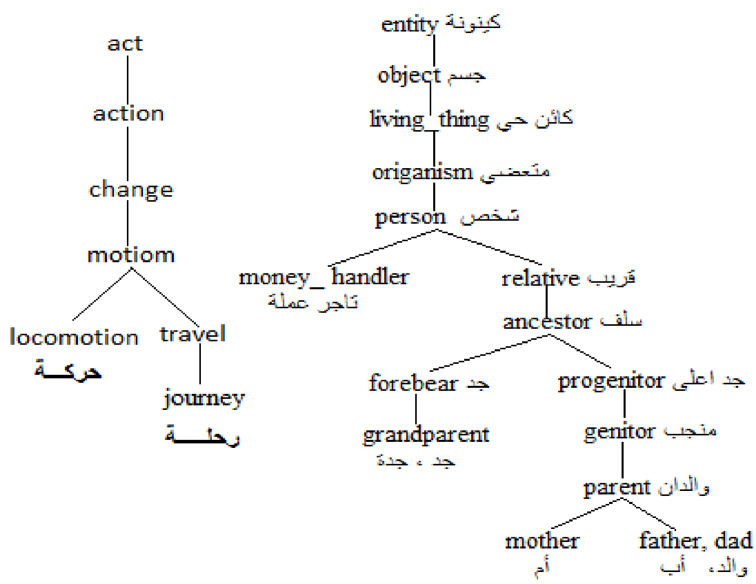


Figure 6-4 A portion of Arabic WordNet (Almarsoomi, et al., 2013)

6.3.1.1. The Proposed Word Similarity Measure

As mentioned in section (6.3.1) above, the AWSS similarity measure proposed by (Almarsoomi, et al., 2013) linked between the length and depth of words in the Arabic WordNet. This measure showed promising results in evaluating nouns dataset using Arabic WordNet (3.0) in terms of correlation between similarity scores and human ratings.

The researcher found that this measure can be improved using the same dataset for the same version of the Arabic WordNet (3.0). Therefore he proposed the following new non-linear equation also linking between both length, and depth of the words in Arabic WordNet (3.0). This alternative equation was simpler and showed stronger correlation with human rating throughout empirical experiments covered in chapter (7)

$$sim(W1, W2) = \alpha^{l}.tanh(\beta * d) \qquad (6\text{-}4)$$

Where:

- W1 and W2 are two words to be compared
- (d) Is the depth of LCS subsuming two words W1 and W2.

- *(l)* is the path length between W1 and W2

α and β are factors equal to (α = 0.801 and β = 0.218) for the Arabic WordNet(3.0), and (α = 0.881 and β = 1) for the lexical tree developed in this work, These factors represent the length and depth factors, and signify the contribution of path length between two words, and the depth of the Least Common Subsumer (LCS). It is worth mentioning that the values for the parameters (α and β) varies depending on the used information source (Arabic WordNet (3.0) or lexical tree). The best values for these parameters were obtained throughout a series of empirical experiments covered in chapter (7)

This equation was tested several times and compared to the AWSS measure using the same Arabic datasets used in AWSS (Almarsoomi, et al., 2013) referred to as (WS) in this work. The correlation coefficient with the human ratings was found to be equal to (r=0.9) using the proposed measure compared to (r=0.894) using the AWSS measure (Almarsoomi, et al., 2013).

Further elaboration on the proposed word similarity measure is covered in chapter (7). The experiments on both word and sentence similarity shall provide more insights about the performance of these measures and their test results which shall give an indication on the best procedure to be used in word similarity measurements in SGO-CA. Appendix (6) of this thesis illustrates a calculation example for the proposed word similarity measure.

6.3.2. Sentence Similarity

Chapter (3) discussed several methods used to measure sentence similarity, however the STASIS method developed by (Li, et al., 2006) showed the most outstanding results in evaluation STASIS is the most heavily cited measure and believed to be the most appropriate method for comparing a pair of sentences by the time of writing this thesis. The researcher used an approach derived from the STASIS method with some modification and adaptation for sentence similarity measurement within SGO-CA.

In the STASIS method (Li, et al., 2006), the sentence similarity measurement is performed in two stages, word similarity and sentence similarity measurement. The STASIS method proposed the use of the word similarity measure developed and evaluated by (Li, et al., 2003) covered in section (6.2.1). Then, sentence similarity measurement is performed as a function between word similarity results. Details about the STASIS method is outlined in chapter (3) and explained in the following sections.

Both word similarity measures (Almarsoomi, et al., 2013), and the proposed measure discussed in section (6.3.1.1) were experimented as a part of the STASIS method and evaluated in chapter (7) to decide which of them is the most appropriate for the Arabic language domain used in this research. The word similarity method with the best results shall be admitted and used within sentence similarity measurements in SGO-CA.

In addition to the adaptation of the word similarity measure, another adaptation was made to STASIS by removing the word order similarity. This was due to the flexible structuring of the Arabic language where word order may not indicate high significance, for example consider the following two sentences.

The two user utterances (الجواز لا يستبدل) and (لا يستبدل الجواز) both mean that “passport is not to be replaced” with the same words but with different order. More details about the structuring of Arabic language were discussed in section (2.5.5).

The following modifications and adaptations were made to STASIS method:

- Using either AWSS word similarity measure or the proposed measure instead of the measure proposed by (Li, et al., 2003)
- Using the mapping between Arabic WordNet and SUMO ontology to calculate the similarity between words regardless of part of speech.
- Removing word order from similarity calculation.

- Introducing the difference between two sentences as an added factor measuring the similarity.
- Using an Arabic corpus to calculate information content values.
- Considering function words in sentence similarity measurement.

6.3.2.1. The Proposed Sentence Similarity Measurement

The following steps describe how the sentence similarity measurement was incorporated within SGO-CA. These steps are derived from STASIS measure with some modifications:

1- Identify the pair of sentences to be compared, let (***U***) be the user utterances and (***R***) is a regular answer stored within one of SGO-CA's knowledge tree nodes.

2- Identify the joint word set (***T***) of two sentences ($\boldsymbol{U}$) and ($\boldsymbol{R}$); which includes all unique words (uncommon) from the both U and ***R***.

3- ($\boldsymbol{U}$) is evaluated against the word set $\boldsymbol{T}$ using these steps:

   a. A similarity matrix ($\boldsymbol{SM1}$) is formed by measuring the similarity of word pairs of each sentence ($\boldsymbol{U}$) and ($\boldsymbol{T}$), using one of word similarity measures (Almarsoomi, et al., 2013) or the new proposed measuring equation described in section (6.3.1.1). The selection of the best equation is covered in the empirical experiments on SGO-CA in chapter (7).

   b. Word similarity scores below the word similarity threshold ($\boldsymbol{WST}$) (covered in the following sections) are set to (0) to eliminate any noise to the semantic matrix.

c. A semantic vector (***SV1***) is formed by taking the maximum similarity score of each column in the matrix and multiplying it with the information content value (***I***(***w***)) of both of the corresponding words in the similarity matrix. Information content value is explained in section (6.3.2.4.1).

4- the regular answer (***R***) is also evaluated against the word set (***T***) using the same stages described in step (3) above, forming another similarity matrix (***SM2***) and similarity vector (***SV2***).

5- the similarity (***S***(***U***,***R***)) between (***U***) and (***R***) (covered in section (6.3.2.4)) is calculated as a cosine similarity between two similarity vectors (***SV1***) and (***SV2***)

6- To signify the contribution of cells containing the value of (0) in similarity vectors (***SV1***) and (***SV2***), the researcher introduced a sentence difference measure ***DF*** (***U***,**R**) (covered in section 6.3.3) and included it in the overall all similarity (***Sim*** (***U***,***R***)).

7- If the overall similarity score (***Sim*** (***U***,***R***)) is greater than or equal to the sentence similarity threshold (***SST***) (explained in section 6.3.6), ***U*** and ***R*** are considered similar, and therefore the user utterances (***U***) is said to match the SGO-CA node containing the regular answer (***R***).

The following sections describe each step of similarity measurement in detail, using this example:

Regular answer (***R***): جوازي مفقود و اريد جواز بديل (My passport is lost and I want another one)

User utterance (***U***): فقدت جوازي يوم امس في المطار (I lost my passport yesterday at the airport)

6.3.2.2. The Joint Word Set

As explained in section (3.5.1.1), the joint word set is defined as a set that contains all the roots of distinct words from user utterances (***U***) and regular answer (***R***), for example:

Regular answer (***R***): جوازي مفقود و اريد جواز بديل (My passport is lost and I want another one)

User utterance (***U***): فقدت جوازي يوم امس في المطار (I lost my passport yesterday at the airport)

Joint word set (***T***): { مطار ، في ، امس ، يوم ، بديل ، اريد ، مفقود ، جواز }{passport, at, yesterday, another, lost, airport, want, my, I}.

The root of each word is extracted using morphological analysis described in section (3.3.4) and the roots of words from both sentences are used to formulate the joint word set.

6.3.2.3. Similarity Matrices

A similarity matrix (***SM1***) between regular answers (**R**) and joint word set (**T**), where the root words of the joint word sets as the first row of the matrix and the root words of the regular answer (**R**) as the first column in the matrix. The value of each cell of the similarity matrix is populated by calculating the similarity between the word pairs corresponding to that cell. Table (6-1) shows the similarity matrix formed between a regular answer (**R**) and the joint word set (**T**) using the same example used in section (6.3.2.1).

Another similarity matrix (***SM2***) is formed between user utterance (***U***) and the joint word set (***T***); this similarity matrix is shown in table (6-2).

No		0	1	2	3	4	5	6	7	8
		جواز	مفقود	اريد	بديل	فقد	يوم	امس	في	مطار
0	جواز	1	0	0	0	0	0	0	0.22	0
1	مفقود	0	1	0	0	0.67	0	0	0	0
2	اريد	0	0	1	0	0	0	0	0	0
3	جواز	1	0	0	0	0	0	0	0.22	0
4	بديل	0	0	0	1	0	0	0	0.22	0

Table 6-1 Similarity matrix between regular answer and the joint word set

No		0	1	2	3	4	5	6	7	8
		جواز	مفقود	اريد	بديل	فقد	يوم	امس	في	مطار
0	فقد	0	0.67	0	0	1	0	1	0	0
1	جواز	1	0	0	0	0	0	0	0.22	0
2	يوم	0	0	0	0	0	1	0.68	0	0
3	امس	0	0	0	0	0	0.68	1	0	0
4	في	0.22	0	0	0	0.27	0	0	1	0
5	المطار	0	0	0	0	0	0	0	0	1

Table 6-2 Similarity matrix between user's utterance and the joint word set

The following steps highlight the population of the similarity matrix:

- If any of the compared words does not exists in the lexical tree then the similarity is (0)
- If both words are identical then the similarity is (1)

- If both words are synonyms then the similarity is also (1)
- Otherwise word similarity is calculated using either (Almarsoomi, et al., 2013) word similarity measure(AWSS) (6-2):

$$sim(W1, W2) = e^{(-\alpha * l)} * \tanh(\beta * d) \quad (6\text{-}2)$$

Or the newly proposed measure (6-4)

$$sim(W1, W2) = \alpha^{l} . \tanh(\beta * d) \quad (6\text{-}4)$$

If the result of word similarity measure fails to pass the word similarity threshold (***WST***) explained in the next section, then the similarity is set to (0).

These two measures are explained in section (6.3.1) of this chapter; and the selection of the best method for word similarity measurement is covered in the experiments in chapter (7).

Word Similarity Threshold

According to (Li, et al., 2006) the word similarity score should pass a predefined threshold referred to as (***WST***) in this work, if it fails to do so, the similarity is set to (0) in the similarity matrix to avoid adding such noise to the matrix, this threshold was set to (0.2) by (Li, et al., 2006) for the English language.

This threshold will be empirically determined in the experiments described in chapter (7) for the Arabic language.

6.3.2.4. Similarity Vectors

The similarity vector is a result of taking the highest value of each column in the similarity matrix described in section (6.3.2.3) and multiplying it by the information content value ($\boldsymbol{I}(\boldsymbol{w})$) of the two corresponding words in the similarity matrix. Consider the similarity matrix shown in table (6-1). The similarity vector (***SV1***) between regular answer (***R***) and the joint word set (***T***) can be calculated as shown in the example below:

$\boldsymbol{SV1}[0]$ = Max ($\boldsymbol{SM1}[0,0]$, $\boldsymbol{SM1}[0,1]$, $\boldsymbol{SM1}[0,2]$, $\boldsymbol{SM1}[0,3]$, $\boldsymbol{SM1}[0,4]$) * $\boldsymbol{I}$(W1) * $\boldsymbol{I}$(W2)

Max ($\boldsymbol{SM2}[0,0]$, $\boldsymbol{SM2}[0,1]$, $\boldsymbol{SM2}[0,2]$, $\boldsymbol{SM2}[0,3]$, $\boldsymbol{SM2}[0,4]$) = $\boldsymbol{SM2}[0,0]$ = 1

W1 = جواز (passport)

W2 = جواز (passport)

I(passport) = 0.58

Regardless of the word similarity measure used within STASIS, the STASIS method assigns a similarity of (1) for any identical words or synonyms, in the case of the above example both words W1 and w2 are identical therefor their similarity is set to (1)

The calculation of word information content value ($\boldsymbol{I}(\boldsymbol{W})$) is covered in the following section. Semantic similarity calculated based on the similarity vectors ***SV1*** and ***SV2*** between the user utterances (***R***) and the joint word set (***T***). Tables (6-3) and (6-4) demonstrate the process of calculating the values of semantic vectors ***SV1*** and ***SV2.***

Similarity vector 1	1* I(جواز) *I(جواز)	1* I(مفقود)* I(مفقود)	1* I(يريد)* I(يريد)	1* I(بديل)* I(بديل)	0.67* I(فقد)* I(مفقود)	0	0	0.22* I(في)* I(جواز)	0

Table 6-3 Similarity vector (1)

Similarity vector 2	1* I(جواز)* I(جواز)	0.67* I(فقد)* I(مفقود)	0	0	1* I(فقد)* I(فقد)	1* I(يوم)* I(يوم)	1* I(امس)* I(امس)	1* I(في)* I(في)	1* I(مطار)*I(مطار)

Table 6-4 Similarity vector (2)

Information Content Value

As discussed in section (3.5.1.3), words that occur more frequently within texts contain less semantic information than words that occur less frequently. In this research a corpus of Arabic words has been collected from Al-Watan newspaper which was collected by (Abbas, et al., 2011). This corpus contains more than (9,000,000) words and was used to estimate word significance based on the frequency of occurrence to calculate information content values.

According to (Li, et al., 2006) word information content value can be calculated from a corpus using the following equation:

$$I(\mathrm{w}) = 1 - \frac{log(n+1)}{log(N+1)} \tag{6-5}$$

Where (w) is the word, (n) is the frequency of occurrence in corpus and (N) is the total number of words in corpus.

Applying corpus statistics to the previous examples leads to the following similarity vectors shown in tables (6-5) and (6-6). Appendix (6) of this thesis illustrates a sample of calculating the information content values.

Similarity Vector 1	0.33	0.51	0.42	0.55	0.25	0	0	0.029	0

Table 6-5 similarity vector (1)

Similarity Vector 2	0.33	0.25	0	0	0.27	0.22	0.37	0.05	0.34

Table 6-6 Similarity vector (2)

6.3.2.5. Sentence Similarity Calculation

According to (Li, et al., 2006) the semantic similarity between the user utterance and regular answer $\boldsymbol{S}(\boldsymbol{U},\boldsymbol{R})$ is defined as the cosine similarity between the two similarity vectors using the following equation:

$$\mathrm{S}_s = \frac{S_1 . S_2}{||S_1||.||S_2||} \tag{6-6}$$

This equation can be elaborated as follows:

$$\boldsymbol{S}(\boldsymbol{U},\boldsymbol{R}) = \frac{\sum_{i=1}^{n}(SV1_i * SV2_i)}{\sqrt{\sum_{i=1}^{n}(SV1_i)^2} * \sqrt{\sum_{i=1}^{n}(SV2_i)^2}} \tag{6-6}$$

Where $\boldsymbol{S}(\boldsymbol{U},\boldsymbol{R})$ is the similarity between user utterance ($\boldsymbol{U}$) and regular answer ($\boldsymbol{R}$), $\boldsymbol{SV1}$ and $\boldsymbol{SV2}$ are similarity vectors and (n) is the length of similarity vectors. Sentence similarity score ranges from 0.0 to 1.0. Where (1) indicates identical similarity and (0) indicates no similarity.

Applying equation (6-6) to the similarity vectors in tables (6-5) and (6-6), gives a result of 0.43 for the sentence similarity

$$S(U,R) = \frac{\begin{bmatrix} 0.33 * 0.33 + 0.51 * 0.25 + 0 + 0 + 0.25 * 0.27 \\ +0 + 0 + 0.029 * 0.05 + 0 \end{bmatrix}}{\begin{bmatrix} \sqrt{(0.33)^2 + (0.51)^2 + (0.42)^2 + (0.55)^2 + (0.25)^2 + (0)^2 + (0)^2 + (0.029)^2 + (0)^2} * \\ \sqrt{(0.33)^2 + (0.25)^2 + (0)^2 + (0)^2 + (0.27)^2 + (0.22)^2 + (037)^2 + (0.05)^2 + (0.34)^2} \end{bmatrix}}$$

$$S(U,R) = 0.43$$

Another example for sentence similarity calculation can be found in Appendix (6) of this thesis.

6.3.3. Sentence Difference Calculation

According to (Lin, 1998) the similarity between two concepts is related to the differences between them. The more differences they have, the less similar they are. This section proposes a novel contribution in similarity measurement, by including the difference between two sentences as a factor in the sentence similarity calculation. This novelty will be fully experimented in chapter (7) to study the impact of including this factor in similarity measurement.

Chapter (3) discussed many methods for sentence similarity. But those methods were not developed specifically for the use of CAs; they focused on sentence similarity but not sentence difference (i.e. sentence length).

Long sentences tend to score higher in similarity than short ones, because in short sentences only few number of words are compared while in long ones there is a better chance of scoring higher similarity ratings among several words.

When comparing sentences with different lengths the comparison does not always lead to fair results, as longer sentences have considerably more rich semantic features than shorter ones. It is also not possible to decide whether short sentences are similar or not due to the lack of these semantic features. In other words, the only thing that can be said about these sentences is that they are different at some level.

For example consider the two sentences "I lost my passport" "فقدت جوازي" and "I lost my passport last month" ."فقدت جوازي الشهر الماضي". The second sentence contains more details about the time in which the action took place, while the first sentence does not include such details. Therefore it is hard to determine the exact similarity of these two sentences.

Furthermore sentence difference cannot be only judged by the length of sentences, because the words of the shorter sentence might all be similar to the words of the longer one.

As discussed earlier in section (6.3.2.4.1) STASIS use information content values to signify the contribution of words that occur less frequently than other words. But STASIS only deals with these information content values for words scoring above than the word similarity threshold ($\boldsymbol{WST}$).

When a word in the joint word set has a similarity score higher than the word similarity threshold ($\boldsymbol{WST}$) with other word in similarity matrices, STASIS use the information content values of the two words to signify the contribution of their importance; but when a word in the word set fails to pass the ($\boldsymbol{WST}$) with any of the other words in one of the similarity matrices, the similarity is set to (0) regardless of its information content value.

Therefore, the researcher proposes calculating the information content values of the words that scores (0) similarity in any of the similarity vectors, and including these content values in sentence difference measurement, and considering them later in the overall similarity measurement between two sentences and see its effect on the final result.

Reconsidering similarity vectors shown in table (6-5) and (6-6) respectively, these similarity vectors contain cells with (0) as a similarity score.

Sentence difference is computed by calculating the average of the information content values for words scoring (0) in the similarity vectors and dividing it by the average of the content values for all words in the word set. Sentence difference is calculated as follows

- If $COUNT(X_k)$ is 0 then the difference is set to (1)
- Otherwise sentence difference is calculated using this equation:

$$DF(U,R) = \frac{\sum_{k=0}^{n} I(X_k)/(COUNT(X_k) + \alpha)}{\sum_{i=0}^{n} I(Y_i)/(COUNT(Y_i) + \alpha)} \qquad (6\text{-}7)$$

- If $(\mathrm{DF}(\mathrm{U}, \mathrm{R}))$ >1 then $(\mathrm{DF}(\mathrm{U}, \mathrm{R}))$ is set to (1).

Where X_k are words having a similarity of zero in the similarity vectors $\boldsymbol{SV1}$ and $\boldsymbol{SV2}$ and $I(X_k)$ is the information content of words having a similarity of (0) in the similarity vectors $\boldsymbol{SV1}$ and $\boldsymbol{SV2}$, and $I(Y_i)$ is the information content values words of the joint wordset, $\boldsymbol{T}$ is the joint word set,(α) is a constant to avoid division by (0). ($\boldsymbol{DF}$) is the level of sentence difference which ranges between (0) and a maximum of (1), and $\boldsymbol{U}$ and $\boldsymbol{R}$ are the user utterances and regular answer respectively.

After calculating sentence similarity and difference, total sentence semantic similarity can be calculated by using the following proposed equation:

$$Sim(\boldsymbol{U},\boldsymbol{R}) = S(U,R) * DF(U,R) \tag{6-8}$$

Using the same similarity vectors in tables (6-5) and (6-6) for the same example used within this chapter, sentence difference can be calculated as follows:

$$DF(U,R) = \frac{I(\text{يوم}) + I(\text{امس}) + I(\text{مطار}) + I(\text{اريد}) + I(\text{بديل})/5}{\begin{bmatrix} I(\text{جواز}) + I(\text{مفقود}) + I(\text{اريد}) + I(\text{بديل}) \\ +I(\text{فقد}) + I(\text{يوم}) + I(\text{امس}) + I(\text{في}) + I(\text{مطار}) \end{bmatrix} /9}$$

$$DF(U,R) = \frac{[0.22 + 0.37 + 0.34 + 0.42 + 0.55]/5}{\begin{bmatrix} 0.33 + 0.51 + 0.42 + 0.55 \\ +0.25 + 0.22 + 0.37 + 0.029 + 0.34 \end{bmatrix} /9}$$

$$DF(U,R) = 0.38/0.35 = 1.08$$

The maximum value for sentence difference is (1), therefore, any results higher than (1) will be set to (1), and total similarity between two sentences can be calculated as:

$$Sim(\boldsymbol{U},\boldsymbol{R}) = 0.43 * 1$$

$$Sim(\boldsymbol{U},\boldsymbol{R}) = 0.43$$

Appendix (6) of this thesis illustrates an example for sentence difference calculation.

6.3.4. Function Words

Function words are words that have little lexical meaning yet they serve to express grammatical relationships with other words within a sentence, such as articles, prepositions, determiners etc.

According to (Miller, et al., 1993) the most obvious difference between WordNet and a standard dictionary is that WordNet divides the lexicon into five categories: nouns, verbs, adjectives, adverbs, and function words.

But WordNet contains only nouns, verbs, adjectives, and adverbs. The relatively small set of English function words is omitted on the assumption that they are probably stored separately as part of the syntactic component of language (Miller, et al., 1993).

According to (Li, et al., 2006) function words contribute less to the meaning of a sentence than other words, while (O'Shea, et al., 2010) stated that function words alone can discriminate between one major class of speech act (questions) and others (affirmative, informative etc.).

Originally the STASIS method (Li, et al., 2006) did not remove function words (such as in, what. etc.) from the joint word set. These function words were retained but they only scored similarity if the two words are identical function words, because function words are not classified somewhere in the WordNet tree.

The researcher believes that function words contain rich semantic and have a significant impact on sentence similarity measurement. Therefore, it was decided to conduct an experiment by including these function words in sentence similarity measurements. But before that, they need to be defined and included in the information source developed in this work which is the lexical tree.

One problem related to adding function words to lexical tree is that they are not classified as a part of something or as a type of an entity, one possible solution for this is to add function words where they are related. For example, prepositions related to time should be added somewhere near time terms in lexical tree, other related to location with the location, and so on.

There is an issue associated with the approach of classifying function words in the lexical tree as some of these function words can refer to variety of things in different contexts, they may refer to time in a context and to place in another, for example the function word ("في" "at") can refer to place in the example "انا في الجامعة" "I'm at the university" or to time as " ساراك في الساعة الثانية" "I'll see you at 2 o'clock". For this reason it cannot be added in a place where they are related.

Function words often serve as a relation between concepts for example "I'm at home". In this example the function words "at" was used to relate between the person and their location. The researcher proposed to classify and place function words in the lexical tree under the term "Relation".

Figure (6-5) shows the classification of function words. An experiment is conducted in chapter (7) to study the impact of including the function words in sentence similarity measurement in relation to the SGO-CA performance in the Iraqi passport domain.

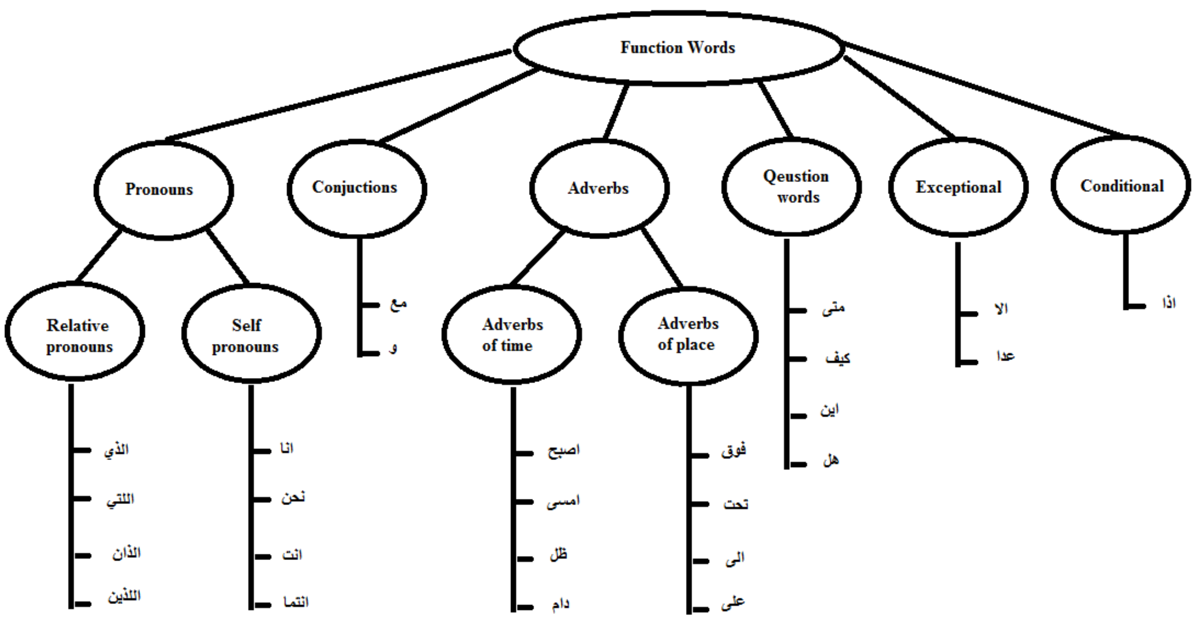


Figure 6-5 Classification of function words in lexical tree

6.3.5. Dialogue Act Classification

According to (O'Shea, et al., 2010), dialogue act classification is a crucial first step in measuring the semantic similarity between a pair of sentences. For example, dialogue acts can distinguish between instructive utterances such as "close the door" and question utterances such as "is the door closed?"

As discussed in chapter (3), sentences in Arabic language can be classified into four recognised types, these are: informative, negative, affirmative, and questionable sentences. The four types can sometimes lead to the same meaning. Table (6-7) shows examples of the four sentence types in Arabic.

It is true that the sentence type or "dialogue act" can have significant impact on sentence meaning as demonstrated in the examples shown in Table 6-7.

However, classifying the user utterance according to the type of Arabic sentences would add more computational complexity to SGO-CA, because the knowledge tree would be searched several times for several types of sentences to find an appropriate match, and this will have a direct effect on the response time of the agent.

Sentence type	Example	translation
Informative sentence	لدي جواز قديم	I have an old passport
Negative sentence	لا املك جواز سفر	I do not have a passport
Questionable sentence	هل تملك جواز سفر ؟	Do you have a passport?
Affirmative sentence	نعم، لدي جواز	Yes, I have a passport

Table 6-7 Types of Arabic sentences

In addition, this classification will not improve the performance of SGO-CA, because same ideas can be expressed by users in different types of sentences. For example the informative sentence "اريد جواز جديد" "I want a new passport" and the questionable sentence "كيف احصل على جواز جديد" "How do I get a new passport" are not similar in type, but in a goal-oriented CA like SGO-CA, these two sentences have exactly the same meaning, that the user wants help with issuing a new passport.

Another example is the instructive sentence "ساعدني في الحصول على جواز" "Help me to get a new passport" and the informative sentence "اريد جواز جديد" "I want a new passport"; they both indicate the same goal in SGO-CA although they are different in types. For these reasons dialogue act classification was not included in SGO-CA.

6.3.6. Semantic Matching

Sentence similarity measurement used in SGO-CA calculates a numeric value between (0) and (1). This value reflects the similarity between the users' utterance and one of the regular answers stored with one of SGO-CA's nodes. However, this value does not indicate whether the user utterance matches exactly the compared sentence or not.

In order to make a decision whether there is a match or not between users' utterances and regular answers stored within SGO-CA, a similarity threshold must be defined. This threshold is a numerical value with a range between 0 and 1. If the similarity is greater than or equal to the threshold then the user utterance match the SGO-CA containing the compared regular answer ($\boldsymbol{R}$). If it fails to pass the threshold then the user utterance does not match the compared sentences and SGO-CA shall continue to evaluate other regular answers of the same node or other nodes.

In SGO-CA a base threshold (0.5) is used for context sensitive nodes, while a lower threshold of (0.2) is used for other nodes. This difference in thresholds is because we need to ensure higher similarity to trigger a context sensitive node, while in non-sensitive nodes lower similarity is quite enough to trigger them.

6.4. Modification to Scripting Language

In chapter 4, a knowledge tree of the Iraqi passport services domain was introduced. This tree was constructed based on domain laws, regulations, and experience collected together to form the knowledge engineering process described in section (4.2).

Value nodes of this knowledge tree contained patterns of potential users' utterances, the conversation between users and the agent flow based on which node matched each user's utterance. This matching was performed with the pattern matching engine described in section (4.3.2).

SGO-CA uses the same knowledge tree used in the PMGO-CA But instead; SGO-CA evaluates users' utterances against regular answers using semantic similarity measures instead of evaluating users' utterances against patterns.

Therefore a modification has been made to the scripting language by converting the value nodes to semantic value nodes. This was achieved by changing the pattern field within the value nodes to become the "Answers" field in the semantic value nodes. Each Answer field contains the regular answers which are evaluated against user utterances to find a semantic match. This evaluation is performed by the semantic engine described in section (6.2.1).

The use of a semantic similarity engine has contributed significantly to reduce the number of patterns/answers stored within SGO-CA's knowledge tree nodes. For example one of the nodes concerned with "new passport" procedures had more than (50) patterns, but in SGO-CA the same node has less than (5) answers; making SGO-CA a lot easier to script and update.

Semantic matching is performed based on a predefined threshold, if similarity score passes this threshold then there is a match, otherwise there isn't. This threshold is called a sentence similarity threshold ($\boldsymbol{SST}$).

If the utterance was inconvenient or inconsistent with the domain, like " what is the weather like", the agent shall respond by saying this issue is outside the conversation please stick to the same subject as shown in conversation sample (6-1).

System: Welcome to the online guide for Iraqi consulate service. User: is it going to rain tomorrow System: this system is for Iraqi consulate issues, please stick to the subject	النظام: اهلا و سهلا بكم في نظام دليل الجوازات الخاص بالقنصلية العراقية انت: هل ستمطر صباح يوم غد النظام: هذا البرنامج مخصص لنظام الجوازات العراقية، يرجى الالتزام بالموضوع

Conversation Sample 6-1 conversation sample

6.4.1. Conflict Resolution Strategy

In pattern matching, a user utterance may match more than one pattern within different nodes of PMGO-CA and a mechanism to resolve this conflict has already been introduced in section (4.3.2.2).

In SGO-CA, the same problem might take place, but in a different way. User utterances might pass the sentence similarity threshold (***SST***), when evaluated with regular answers of different SGO-CA nodes. To overcome this, the researcher decided to use the highest similarity score as a method to eliminate conflicts among SGO-CA nodes, therefore the SGO-CA containing a regular answer with the highest similarity score with user utterance is triggered. If similarity scores are the same between a user utterance and two regular sentences which belong to different SGO-CA nodes, in this case the node with the regular answer that scored high similarity first will be triggered. for example:

- Regular answer 1 : (**اريد الحصول على جواز**) (I want to obtain a passport)
- Regular answer 2 : (**اريد جواز مرور**) (I want a travel document)
- User utterance 2 : (**كيف يمكن ان احصل على جواز**)

The regular answer (1) scored a similarity of (0.77) with user utterances; while the second regular answer which belongs to other node scored a similarity of (0.48) therefor the SGO-CA node containing the regular answer (1) will be triggered

6.4.2. Software Tools Used to Construct SGO-CA

Chapter (4) introduced the software tools used to construct PMGO-CA. Some of these tools were modified and adapted to be used to construct SGO-CA. Other tools were added and developed specifically to manage the information sources used by SGO-CA in similarity measurements. The following sections describe these software tools in detail.

As discussed in section (3.3.6), the Arabic WordNet browser was not designed for sentence similarity measurement. It also lacked sufficient interfaces to modify both the ontology and the lexical database. Therefore, developing SGO-CA using the existing AWN browser (The Global WordNet Association, 2014) was not possible at the time of this work. Instead, a new tool called "SGO-CA Manager" has been developed by the researcher using parts of WordNet software. The SGO-CA manager contains an editor tool to manage the lexical tree described in section (6.2.2) and also tools to calculate word and sentence similarity. This software tool contains the following features:

1- Facility to add, remove and modify ontology concepts and Arabic words directly.
2- Functionality to perform word similarity and sentence similarity using variety of word similarity measures
3- Word frequency calculation according to corpus: Word frequency calculation is a part of the sentence similarity method described in section (6.3.2)
4- Full integration within the CA manager to perform semantic similarity between user's utterance and the answers stored within the CA.

The “SGO-CA manager” tool makes SGO-CA easier to script and implement with all options in one place, but this tool maintain its information separately from Arabic WordNet database making it suitable to be used to script multiple domains.

Figure (6-6) shows the main interface for the SGO-CA manager tool. This interface contains three options which will now be explained in detail.

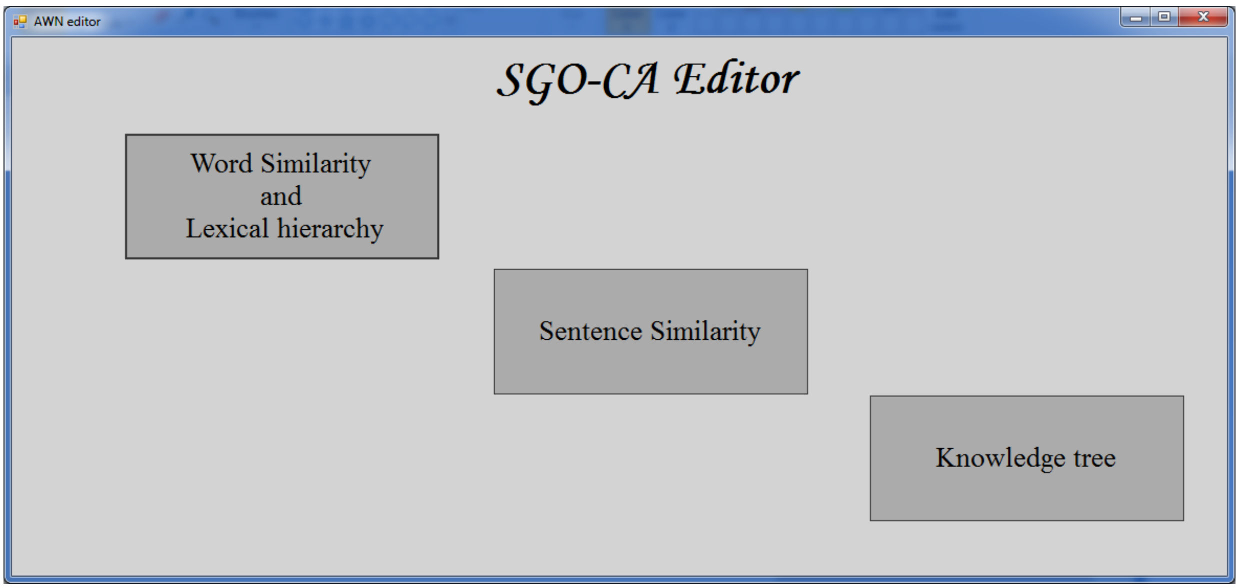


Figure 6-6 Main interface of SGO-CA script editor

6.4.2.1. Lexical Hierarchy Editor

The lexical tree editor enabled the SGO-CA scripter to manage the lexical hierarchy whose structure is identical to the ontology structure of Arabic WordNet in order to evaluate the proposed and existing word and sentence similarity measures. More than (2000) Arabic words were added to the lexical tree and organised according to their mapping to SUMO ontology concepts. Some of those Arabic words covered most of the words used within the passport domain, the remaining were some common words used in conversations and not strictly related to a specific domain. Figure (6-7) shows the interface of the lexical tree editor.

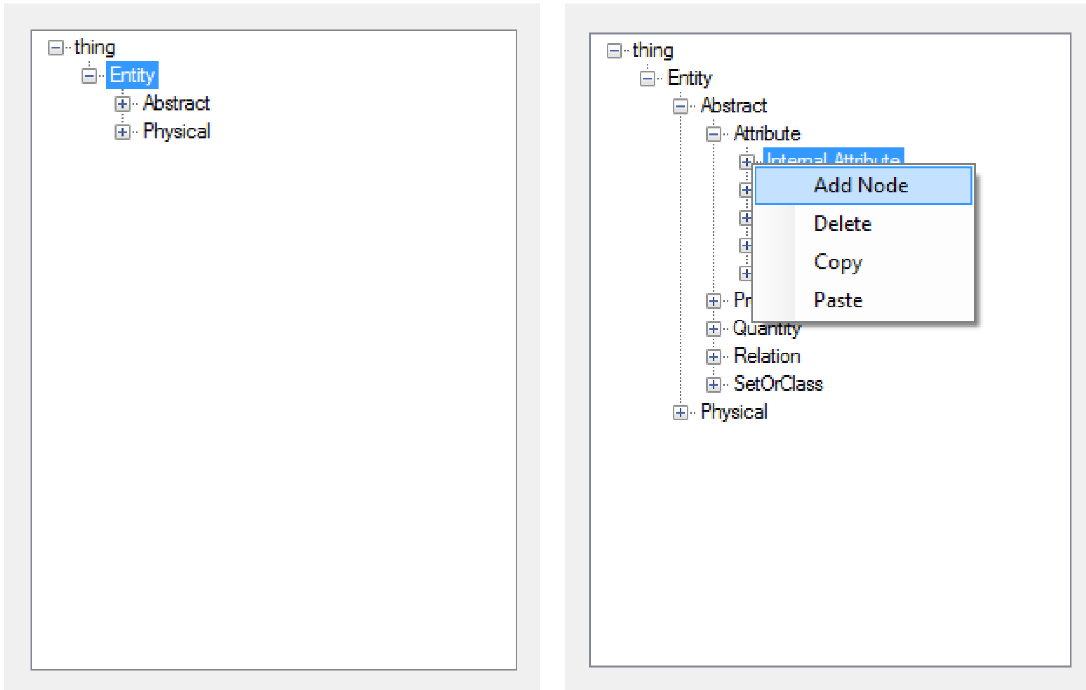


Figure 6-7 Lexical tree editor

The lexical tree contains two types of nodes:

- Term node: this node represents a class of an entity or relationship within the ontology, details about ontology and classes can be found in section (3.3.5). Figure (6-9) shows a part of the lexical tree, with the term "License" which descends from the Term "Certificate" which in turn descends from the term "Text" and so forth.
- Arabic word node: This node contains an Arabic word which is an instance of the Ontology node containing it. Figure (6-8) shows the Arabic word (جواز) (passport) as a type of the term (License).

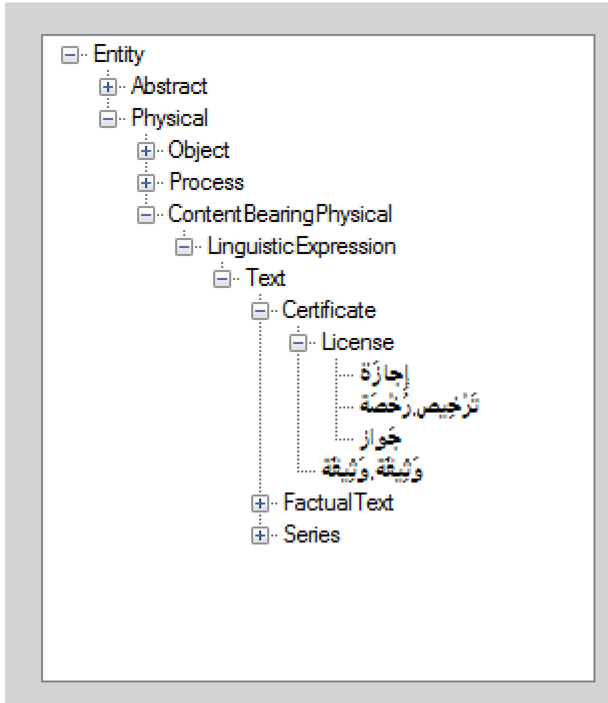


Figure 6-8 portion of the lexical tree

Ontology terms are written and maintained in English language, because the structure of the lexical tree was taken from the universal structure of the SUMO ontology which also maintained ontology terms in English language. Arabic words were attached to their corresponding English words mapped to the SUMO ontology.

The Scripter can modify the lexical tree to add/ delete new ontology terms. Figure (6-9) demonstrates adding ontology term 'Legal document' to the lexical tree by typing the term in the designated box, and then clicking the "Add term" button.

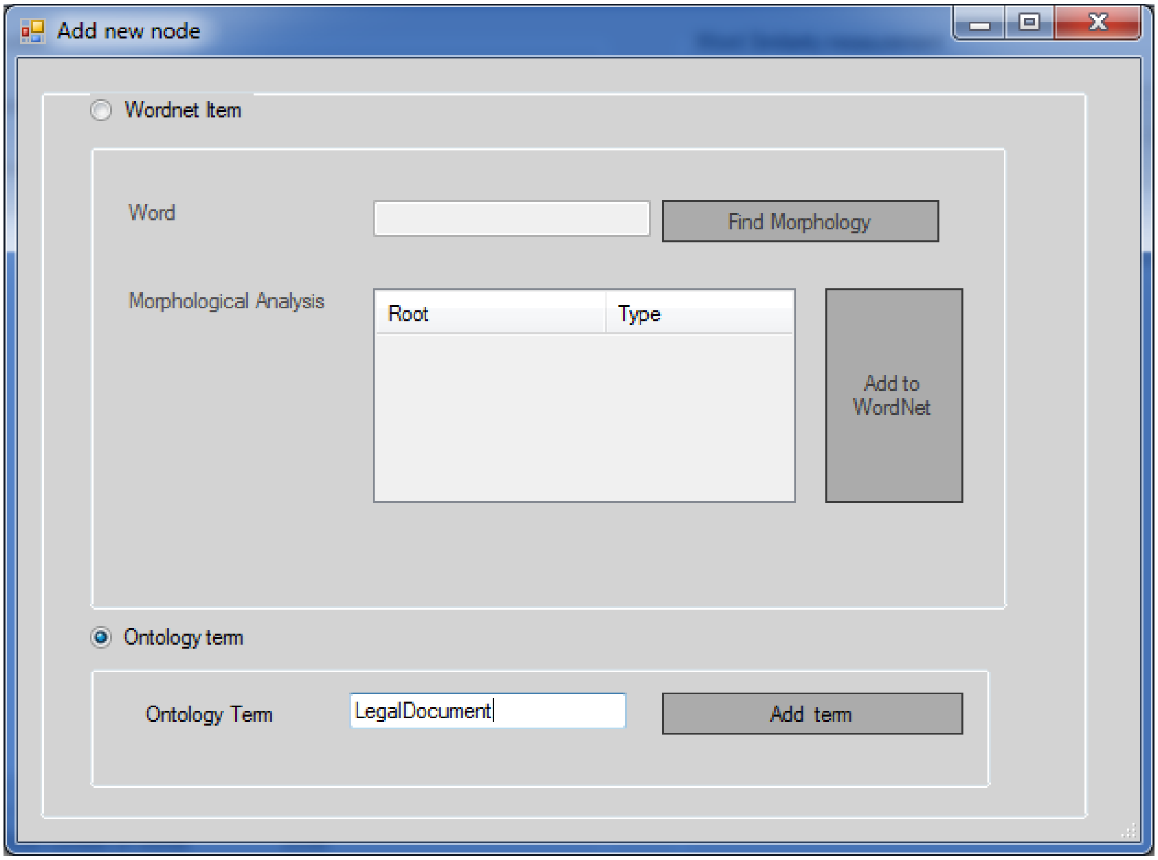


Figure 6-9 Interface of adding ontology term to the lexical tee

The scripter can also add new Arabic words, but before doing that, the roots of these words is identified using morphological analysis (AraMorph, 2003), because only the root of words can be added to the lexical tree. The Morphological analysis tool (AraMorph, 2003) is integrated within the "SGO-CA" manager tool.

The reason for adding only the root of words is to eliminate the need to add several morphological forms for each word to the tree. This helps to minimize the size of the lexical tree and makes words look up much faster.

Figure (6-10) shows the interface of adding Arabic words to the lexical tree. The scripter types the word (i.e. passport (جواز)) in any morphological form in the designated box and clicks the "find morphology" button to extract the word root in order to be added to the lexical hierarchy. The list box shows a list of available morphological roots and categories for this word; the scripter then selects the required root and clicks the “Add to WordNet” button to include the selected word in its desired location of the lexical tree.

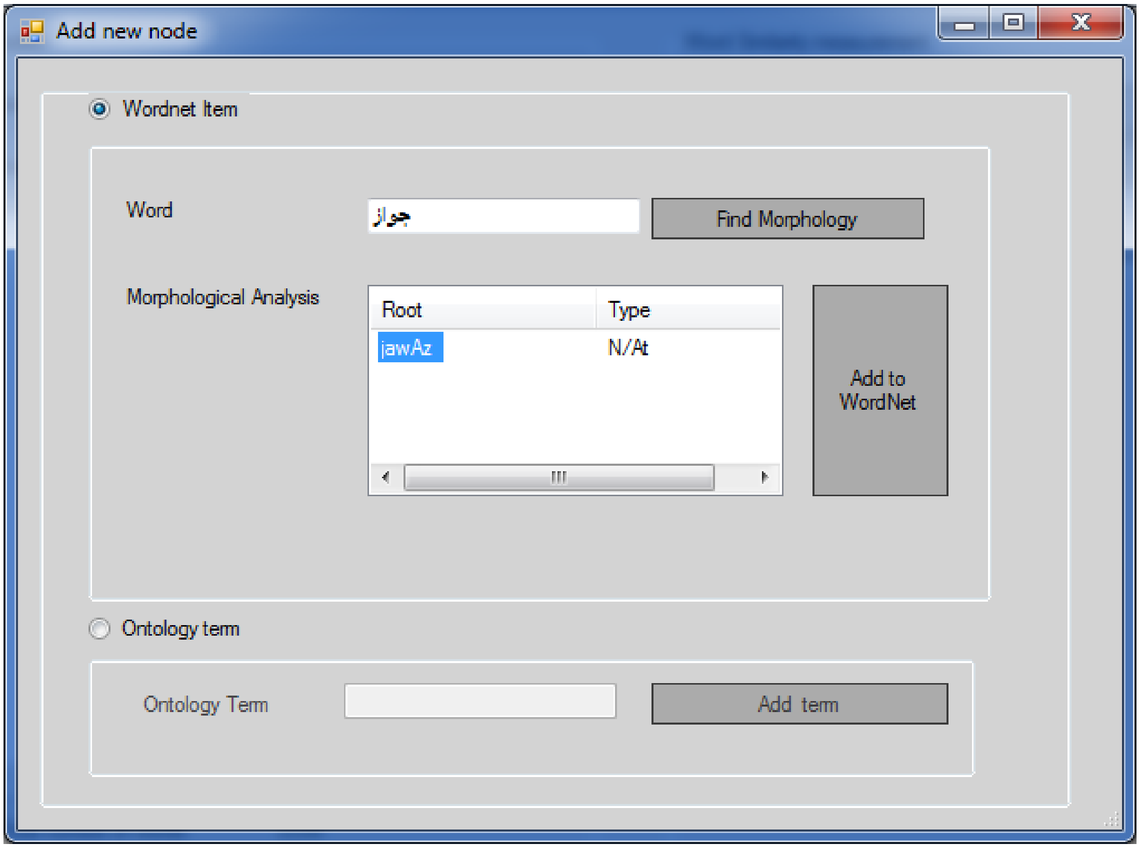


Figure 6-10 Interface of adding Arabic word to the lexical tree

The scripter can also use the search facility to look up a word in the lexical tree to find its location. For example figure (6-11) shows the interface of searching for the word (جواز) (passport) in the lexical tree. The scripter may type a word in the designated box and click the search button to show the location of the searched word and its information content value.

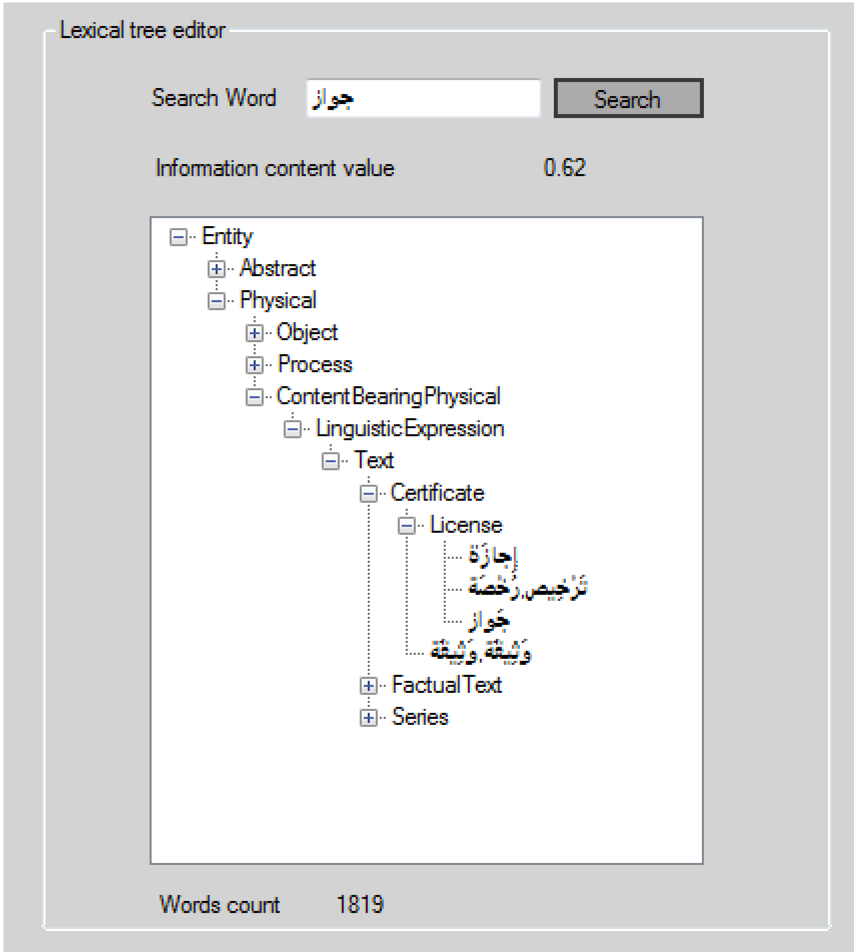


Figure 6-11 Interface of word search

6.4.2.2. Similarity Calculator

The similarity calculator is a part of SGO-CA manager tool that enables the scripter to gain access to the semantic similarity engine and perform similarity measurement between words and sentences.

Figure (6-12) shows the interface of the word similarity measurement using the AWSS measure (Almarsoomi, et al., 2013) . By typing each word in its designated box and hitting the “Measure Similarity” button. The program will extract the root of each word and perform a quick look up in the lexical hierarchy to obtain their location and calculate word similarity according to the methods described in section (6.3.1).

The interface shows the similarity between the compared pairs of words and the information related to it, such as the path distance between two words, and the term subsuming this pair of words (LCS). The depth of LCS is also shown in the interface. More details about this information can be found in section (6.3.1) of this chapter.

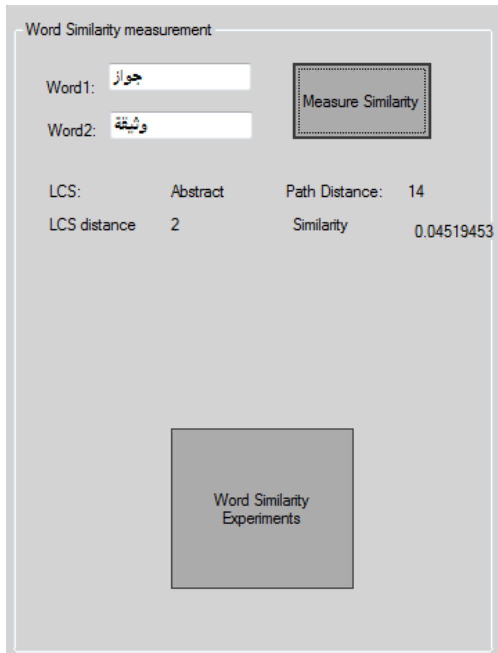


Figure 6-12 interface of word similarity measurement

Figure (6-13) shows the interface of the sentence similarity measurement, this interface displays the similarity matrices and similarity vectors comparing the two sentences (اريد جواز جديد) (I want to obtain a new passport) and (لا اريد جواز)) I do not want a passport). The interface also has a designated area for “unknown words”: those which do not exist in the lexical tree enabling the scripter to add them. The total sentence similarity score is also displayed by the interface.

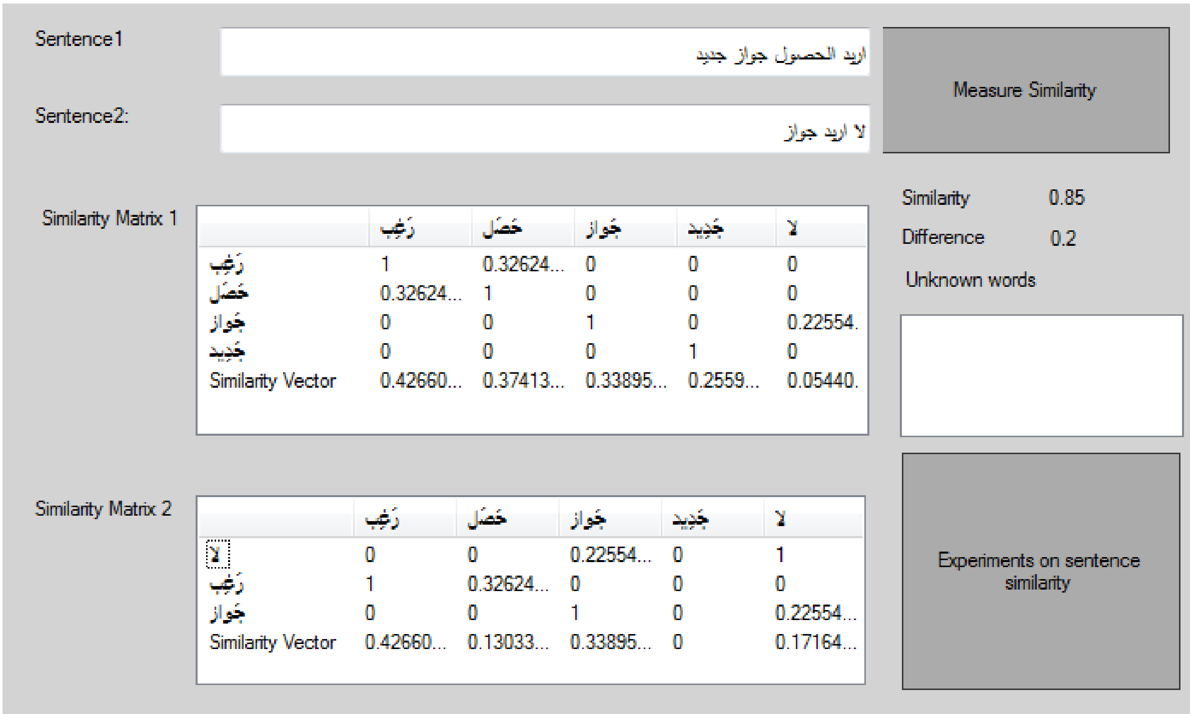


Figure 6-13 interface of sentence similarity measurement

The purpose of the similarity calculator is to provide an interface on which the scripter can conduct experiments and observe results more closely; the semantic calculator is used as a part of the semantic similarity engine used to evaluate users utterances (*U*) against regular answers (*R*) stored within SGO-CA nodes.

6.5. Implementation of SGO-CA in the Iraqi Passport Domain

Information about the Iraqi passport domain was gathered and modelled into process charts and flowcharts as mentioned in chapter (4), then it was converted to a knowledge tree having three types of nodes (question nodes, value nodes and report nodes).

Value nodes contained patterns of expected user's utterances which PMGO-CA used to evaluate users utterances in order to decide whether to trigger that particular node or not.

The same knowledge tree with the same methodology was used to construct SGO-CA, the value node was transformed to semantic value nodes, these semantic nodes no longer contain patterns, instead they contain a list of regular answers which are Arabic sentences used to evaluate users utterances semantically by measuring their similarity using the measures proposed in this chapter.

Therefore the same knowledge tree was used, by removing the patterns and replacing them with 3 or 4 sentences to be used for semantic evaluation and decide whether to trigger the semantic value node containing the matched answer.

The mechanism used for context switching, promotion/ demotions, and activation/ deactivation remained the same as used in PMGO-CA.

6.6. Summary

This chapter proposed a novel Arabic goal-oriented semantic conversational agent to overcome the scripting complexity and maintenance associated with pattern matching conversational agents. Unlike pattern matching CAs (which is domain dependent), semantic conversational agents use information sources such as WordNet and SUMO ontology to calculate similarity between sentences.

The chapter began with an introduction to SGO-CA, the theory used in calculating sentence similarity measures, and how these methods were adopted to be used in SGO-CA. In addition to that, the chapter proposed improvements on the existing methods, these improvements are evaluated using empirical experiments described in chapter (7).

The architecture of SGO-CA comprises of a semantic similarity engine which is used to perform the matching between users' utterances and regular answers stored within the knowledge tree of SGO-CA.

Semantic similarity uses information sources such as WordNet to calculate the similarity; however the slow performance of Arabic WordNet browser made it almost impossible to be used in this research, especially when the researcher needed fast tools to test results and make observations, which was not possible using the Arabic WordNet browser.

To overcome the insufficiency of the Arabic WordNet, the researcher developed a new information source called the “lexical tree” which utilises the SUMO mapping to the Arabic WordNet.

A New software tools were created in this work to manage SGO-CA, these tools were used to edit the lexical tree, calculate word and sentence similarity, and to manage the knowledge tree of SGO-CA, these tools helped the researcher to manage SGO-CA and make modifications to the lexical tree to test the results directly.

This Arabic SGO-CA developed is expected to offer significant improvements over PMGO-CA developed in chapter (4). The experiments and evaluation carried out in chapter (7) shall examine the validity of this assumption.

The key contributions of this chapter can be highlighted as:

- Introducing a novel new word similarity measure to provide stronger results than the measures used in literature
- Creating of a new Arabic lexical tree based on the SUMO mapping WordNet in word similarity measurement.
- Adapting sentence similarity measures from literature to be used to construct an Arabic semantic Goal-Oriented Conversational Agent (SGO-CA).
- A novel contribution of using sentence difference as a factor in overall sentence similarity.
- Including function words in similarity measurements.

Chapter 7
Experiments and Evaluation of SGO-CA

7.1. Introduction

Chapter (6) proposed a novel Arabic word similarity measure and software tools used to measure word and sentence semantic similarity. This chapter is concerned with testing these tools and evaluating the proposed the SGO-CA.

One of the main tests in this chapter is the adaptation of the sentence semantic similarity measure (STASIS) introduced earlier for the English language by (Li, et al., 2006), by incorporating the new proposed Arabic word measure (6-4) discussed in section (6.3.1.1).

This chapter is divided into two parts: the first part describes a series of empirical experiments to examine the proposed similarity measures; the following list highlights the experiments:

- Developing a suitable word similarity measure to be used in SGO-CA.
- Define best values for α and β in word similarity measures that correlate best with human rating.
- Selection of the word similarity threshold ($\boldsymbol{WST}$) for the given measure.
- Use of function words in similarity measure calculations
- Inclusion of sentence difference in overall similarity measurement.

The second part of the chapter covers an evaluation of SGO-CA carried out by human participants. The aim of this evaluation is to test the viability of the new proposed architecture, and then the results will be compared to the (PMGO-CA) developed in chapter (4) and evaluated in chapter (5).

7.2. Experimental Methodology

Experiments were designed to test and tune the proposed word similarity measure (6-4). This is conducted by selecting datasets of Arabic words (AWSS) developed by (Almarsoomi, et al., 2013) to obtain the best correlation between human rating and machine calculation. This dataset is shown in table (7-1) and will be referred to as (***WS***) throughout this chapter.

Word Pairs		ازواج الكلمات		Human Ratings	AWSS measure
Coast	Endorsement	تصديق	ساحل	0.01	0.0
Noon	String	خيط	ظهر	0.01	0.27
Stove	Walk	مشي	موقد	0.02	-
Slave	Vegetable	خضار	عبد	0.04	0.06
Smile	Village	قرية	ابتسامة/بسمة	0.05	0.0
Wizard	Infirmary	مشفى	ساحر	0.06	-
Hill	Pigeon	حمامة	تل	0.08	0.06
Glass	Diamond	الماس	كأس	0.09	0.05
Cord	Mountain	جبل	حبل	0.13	0.17
Forest	Shore	شاطئ	غابة	0.21	0.17
sepulcher	Sheikh	شيخ	ضريح	0.22	0.06
Tool	Pillow	مخدة	أداة	0.25	0.32
Coast	Mountain	جبل	ساحل	0.27	0.45
Tool	Tumbler	قدح	أداة	0.33	0.54
Journey	Shore	شاطئ	رحلة	0.37	0.0
Coach	Travel	سفر	حافلة	0.40	0.0
Food	Oven	فرن	طعام	0.44	-
Feast	Fasting	صيام	عيد	0.49	0.17
Coach	Means	وسيلة	حافلة	0.52	0.38
Girl	Sister	اخت	فتاة	0.60	0.37
Hill	Mountain	جبل	تل	0.65	-
Master	Sheikh	شيخ	سيد	0.67	0.67
Food	Vegetable	خضار	طعام	0.69	0.53
Slave	Odalisque	جارية	عبد	0.71	0.93
Run	Walk	مشي	جري	0.75	0.60
Cord	String	خيط	حبل	0.77	0.70
Forest	Woodland	أحراش	غابة	0.79	0.82
Cushion	Pillow	مخدة	مسند	0.85	0.82
Countryside	Village	قرية	ريف	0.85	0.82
Coast	Shore	شاطئ	ساحل	0.89	0.89
Tool	Means	وسيلة	أداة	0.92	0.93
Boy	Lad	فتى	صبي	0.93	0.95
Sepulcher	Grave	قبر	ضريح	0.94	0.82
Wizard	Magician	مشعوذ	ساحر	0.94	-
Glass	Tumbler	قدح	كأس	0.95	0.89

Table 7-1 AWSS evaluation dataset (WS)

To test the adaptation of the semantic sentence similarity (STASIS), a set of (30) sentence pairs were selected from an English dataset developed originally by (O'Shea et al 2013). These (30) sentence pairs were translated from English to Arabic language by an Arabic linguistic expert. Human ratings scaled between (0) to (4) was converted to read (0) to (1) for consistency with human rating as shown in table (7-2). This dataset shall be referred to (SD) in this chapter.

SP	Sentences	الجمل	HR
1	You're not a good friend if you're not prepared to be present when I need you.	أنت لست صديقا جيدا إذا كنت غير مستعد لتكون حاضرا عندما أحتاجك.	0.785
	A good friend always seems to be present when you need them.	الصديق الجيد يكون دائما حاضرا عند الحاجة إليه.	
2	If you continuously use these products, I guarantee you will look very young.	إذا كنت تستخدم هذه المنتجات بشكل مستمر، أنا أضمن لك سوف تظهرصغير السن جدا.	0.895
	I assure you that, by using these products consistently over a long period of time, you will appear really young.	أؤكد لك أنه باستخدام هذه المنتجات بشكل ملائم لفترة طويلة من الزمن سوف تبدو صغيرحقا.	
3	Water freezes at a certain temperature, which is zero degrees Celsius.	يتجمد الماء عند حرارة معينة، وهي صفر مئوي.	0.77
	The temperature of boiling water is 100 C and the temperature of ice is 0 C	درجة حرارة الماء المغلي هي مئة مئوية ودرجة حرارة الجليد هي صفر مئوية.	
4	We got home safely in the end, although it was a long journey.	وصلنا البيت بسلام في النهاية، على الرغم أنها كانت رحلة طويلة.	0.765
	Though it took many hours travel, we finally reached our house safely.	رغم ان ساعات السفر كانت عديدة ، اخيرا وصلنا منزلنا بسلام.	
5	A man called Dave gave his fiancée a large diamond ring for their engagement.	رجل يدعى سامر قدم لخطيبته خاتم كبيرمن الماس في الخطوبة.	0.805

SP	Sentences	الجمل	HR
	The man presented a diamond to the woman and asked her to marry him.	قدم رجل الماس للمرأة وطلب منها أن تتزوجه.	
6	Midday is 12 o'clock in the middle of the day.	منتصف اليوم هو الساعة الثانية عشر في منتصف النهار	0.99
	Noon is 12 o'clock in the middle of the day.	الظهر هو الساعة الثانية عشر في منتصف النهار	
7	The first thing I do in a morning is make myself a cup of coffee.	اول شئ أفعله في الصباح هو اصنع لنفسي فنجان من القهوة.	0.962
	The first thing I do in the morning is have a cup of coffee.	اول شئ أفعله في الصباح هو تناول فنجان من القهوة.	
8	Meet me on the hill behind the church in half an hour.	قابلني على التل وراء الكنيسة خلال نصف ساعة.	0.982
	Join me on the hill at the back of the church in thirty minutes time	التحق بي على التلة خلف الكنيسة خلال ثلاثين دقيقة من الوقت.	
9	Get that wet dog off my brand new white sofa.	ابعد هذا الكلب الرطب من أريكتي البيضاء الجديدة.	0.898
	Make that wet hound get off my white couch I only just bought it.	اجعل هذا الكلب الرطب ينزل من اريكتي البيضاء لقد اشتريتها للتو.	
10	Could you climb up the tree and save my cat from jumping please?	هل يمكنك تسلق الشجرة وانقاذ قطتي من القفز رجاءأ؟	0.958
	Can you get up that tree and rescue my cat otherwise it might jump?	هل يمكنك صعود تلك الشجرة وانقاذ قطتي وإلا فإنها قد تقفز؟	
11	I have invited a variety of people to my party so it should be interesting.	لقد دعوت مجموعة متنوعة من الناس لحفلتي لذ ستكون ممتعة	0.545
	A number of invitations were given out to a variety of people inviting them down the pub.	قدمت عددا من الدعوات إلى لمجموعة متنوعة من الناس الى تدعوهم الى الحانه.	
12	Do you want to come with us to the pub behind the hill?	هل تريد أن تأتي معنا إلى الحانة وراء التل؟	0.455

SP	Sentences	الجمل	HR
	We are going out for drinks tonight in Salford Quays if you would like to come	سوف نخرج هذه الليلة لتناول المشروبات في بغداد إذا رغبت أن تأتي.	
13	You shouldn't be covering what you really feel	أنت لا ينبغي أن تخفي ما تشعر به حقا.	0.552
	There is no point in covering up what you said, we all know	لا يوجد أي نقطة في اخفاء ما قلته، نحن نعلم جميعا	
14	You must realize that you will definitely be punished if you play with the alarm	يجب أن تدرك أنك بالتأكيد ستعاقب اذا كنت تلعب بالمنبه.	0.71
	He will be harshly punished for setting the fire alarm off.	ستعاقب بقسوة لاطفائك منبه الحريق .	
15	It seems like I've got eczema on my ear doctor, can you recommend something for me?	يبدو ان عندي الأكزيما في أذني ايها الطبيب، هل تفضل لي شيئا؟	0.512
	I had to go to a chemist for a special rash cream for my ear.	علي أن أذهب إلى الصيدلية لكريم طفح خاص لأذني.	
16	Roses can be different colors, it has to be said red is the best though.	الورود تكون بألوان مختلفة ، لكن لا بد القول ان الأحمر هو الأفضل .	0.708
	Roses come in many varieties and colors, but yellow is my favorite	الورود تأتي بأصناف وألوان متنوعة، لكن الأصفر هو الافضل لدي.	
17	Would you like to go out to drink with me tonight?	هل ترغب في الخروج للشرب معي الليلة؟	0.252
	I really don't know what to eat tonight so I might go out somewhere	أنا حقا لا اعلم ماذا ساكل الليلة لذا قد أذهب الى مكان ما	
18	I am so hungry I could eat a whole horse plus dessert	أنا جائع جدا لدرجة يمكنني أكل حصان بأكمله بالإضافة إلى حلوى	0.765
	I could have eaten another meal, I'm still starving.	كنت استطيع اكل وجبة اخرى، انا لازلت متضورا.	
19	We ran farther than the other children that day	ركضنا أبعد من الأطفال الآخرين ذلك اليوم	0.608

SP	Sentences	الجمل	HR
	You ran farther than anyone today	ركضت أبعد من الآخرين اليوم	
20	I am proud of our nation, well, most of it.	أنا فخور بأمتنا، حسنا، اغلبها.	0.428
	I think of myself as being part of a nation	أفكر في نفسي بأني جزء من أمة	
21	Does music help you to relax, or does it distract you too much?	هل تساعدك الموسيقى على الاسترخاء، أم أنها تلهيك كثيراً؟	0.025
	Does this sponge look wet or dry to you?	هل تبدو هذه الاسفنجة رطبة ام جافة بالنسبة لك	
22	The children crossed the road very safely thanks to the help of the lollipop lady	الأطفال عبروا الطريق بسلام جدا شكرا لمساعدة بائعة المصاصات.	0.032
	It was feared that the child might not recover, because he was seriously ill.	كان يخشى من أن الطفل قد لا يتعافى، لأنه كان مريضا بجد.	
23	Boats come in all shapes and sizes but they all do the same thing.	القوارب تأتي بجميع الأشكال والأحجام ولكنها جميعا تفعل الشيء نفسه	0.125
	Chairs can be comfy and not comfy, depending on the chair	الكراسي تكون مريحة او غيرمريحة، اعتمادا على الكرسي	
24	There was a heap of rubble left by the builders outside my house this morning	كان هناك كومة من الأنقاض من قبل البنائين تركت خارج داري هذا الصباح	0.022
	Sometimes in a large crowd accidents may happen, which can cause deadly injuries.	أحيانا تقع حوادث بوجود حشد كبير ، وقد يمكن أن تسبب اصابات قاتلة	
25	I love to laugh as it makes me happy as well as those around me.	أنا أحب أن أضحك لأنه يجعلني سعيدا وكذلك الاخرين من حولي.	0.02
	I thought we bargained that it would only cost me a pound.	اعتقدت اننا تفاوضنا بانه سيكلفني باوند فقط.	
26	He was harshly punished for setting the fire alarms off.	هو عوقب بقسوة لانه اطفأ جهازتنبيه الحريق .	0.055

SP	Sentences	الجمل	HR
	He delayed his response, in order to create a tense atmosphere.	تأخر رده ليخلق جو من التوتر.	
27	Someone spilt a drink accidentally on my shirt, so I changed it.	شخص ما اسقط شراب بطريق الخطأ على قميصي، لذا غيرته.	0.12
	It appears to have shrunk; it wasn't that size before I washed it	يبدو أنها تقلصت، لم تكن بهذا الحجم قبل غسلها.	
28	The damp was mostly in the very corner of the room	الرطوبة في الغالب في الزاوية البعيدة من الغرفة	0.028
	The young lady was somewhat partially burnt from the sun.	احترقت الشابة جزئيا من الشمس.	
29	Flies can also carry a lot of disease and cause maggots.	يمكن أن يحمل الذباب الكثير من المرض ويسبب اليرقات.	0.03
	I dry my hair after I wash it or I will get ill.	انا اجف شعري بعد غسله والا سوف امرض.	
30	They said they were hoping to go to America on holiday.	قالوا انهم كانوا يأملون ان يذهبوا إلى أمريكا في اجازة.	0.04
	I like to cover myself up in lots of layers, I don't like the cold.	أحب تغطية نفسي بالكثير من الطبقات، أنا لا أحب البرد.	

Table 7-2 dataset of English sentence pairs with Arabic translation (SD)

7.2.1. Experiment (1): Investigation of Word Similarity Measures

In this experiment the dataset (***WS***) of the Arabic nouns is used to compare the word similarity measure (6-2) developed by (Almarsoomi, et al., 2013) (discussed in section 60301) and the proposed word similarity measure (6-4).

This experiment aims to test the following hypotheses:

- H0: the proposed word similarity measure (6.4) can be used as an alternative to the AWSS measure (Almarsoomi, et al., 2013) as it provides higher correlation with human ratings.

$$sim(W1, W2) = e^{(-\alpha * l)} * tanh(\beta * d) \quad (6\text{-}2)$$

$$sim(W1, W2) = \alpha^{l}.tanh(\beta * d) \quad (6\text{-}4)$$

- H1: the proposed word similarity measure cannot be used as an alternative to the AWSS measure (Almarsoomi, et al., 2013) as it does not provide stronger correlation with human rating.

The comparison is performed by measuring the Pearson correlation coefficient between both similarity measures and the rating of human participants for the dataset of Arabic nouns (***WS***). The similarity measure with the strongest correlation will be chosen as a word similarity measure within SGO-CA.

Table (7-1) in appendix (7) shows the word pairs of dataset (***WS***) with information about their path length and the depth of their lowest common subsumer (LCS) along with human rating and machine rating (shown in the AWSS column in table (7-1) appendix (7)) for both AWSS similarity measures and the proposed measure (6-4). For this experiment, the same Arabic WordNet version 3 used by (Almarsoomi, et al., 2013)was used. This is to ensure that the comparison used the same testing features.

The experiment shows excellent results for the new proposed similarity measure with a correlation coefficient of ($r = 0.9$) when $\alpha = 0.801$ and $\beta = 0.218$ compared to the correlation coefficient obtained from the AWSS measure of ($r = 0.894$) with the optimised values of $\alpha = 0.162$ and $\beta = 0.234$ (Almarsoomi, et al., 2013). Therefore the null hypothesis can be accepted and the proposed word similarity measure (6-4) can be used as an alternative for the AWSS measure (Almarsoomi, et al., 2013).

7.2.2. Experiment (2) Tuning the Proposed Word Measure

Section (7.2.1) discussed an experiment to test out the new proposed word similarity measure (6-4) on the dataset ($\boldsymbol{WS}$) used to evaluate the AWSS measure (6-2) (Almarsoomi, et al., 2013). The new measure showed stronger correlation with human judgment as explained in section (7.2.1).

In this section the word similarity measure (6-4) is tuned to fit the new information source developed in this work, which utilises the mapping between WordNet and SUMO ontology (the lexical tree) instead of the original hierarchy of WordNet used in the previous experiment. Using this new mapping changed the path and depth of Arabic words, therefore the new measure must be tuned to fit the new information source and obtain the best correlation results with human rating of the dataset (*SW0*).

The aim of this experiment is to obtain the optimum results of the parameters (α) and (β) for the new word similarity measure (6-4) using the lexical tree developed in this work as the information source to evaluate semantic similarity between two words, the optimized parameters shall be used in measure(6-4) to estimate word similarity within SGO-CA.

This experiment is performed using the same word dataset ($\boldsymbol{WS}$); but the similarity is evaluated based on their path length and LCS depth of words in the lexical tree instead of their path length and LCS depth in Arabic WordNet.

Table (7-2) in appendix (7) show the word pairs of ($\boldsymbol{WS}$) with the path length and LCS depth parameters, the table also shows the human rating of the dataset ($\boldsymbol{WS}$), and the similarity results of word pairs using the new measure (6-4).

The tuning of the parameters (α) and (β) is performed throughout scanning the combinations of (α) and (β) starting from (0) to (1) with an increment of (0.001), and measure the Pearson correlation against human rating for each combination and select the combination with the strongest correlations.

The optimised values for (α) and (β) in this experiment were (α = 0.881) and (β=1) which obtained a correlation of (r=0.868) with human rating. Therefore, these values shall be used to calculate word similarity in SGO-CA.

7.2.3. Experiment (3) Incorporating the New Word Similarity Measure in Sentence Similarity Calculation

The aim of this experiment is to test out the new word similarity measure (introduced in chapter 6) and optimised in section (7.2.2) within the proposed sentence similarity measures (introduced in chapter 6).

This experiment is performed using the sentences dataset (***SD***), and the lexical tree developed in this work, and the Arabic corpus introduced in chapter (6) to calculate the semantic similarity of sentence pairs in the dataset ($\boldsymbol{SD}$)

Table (7-3) in appendix (7) list the sentence pairs from ($\boldsymbol{SD}$) and their similarity scores using the sentence similarity measure (introduced in chapter 6) incorporating and the new proposed measure (6-4) to calculate word similarity.

The experiment results showed that the proposed sentence similarity measure achieved a correlation of (r=0.886) with human rating, this is considered as an outstanding result, compared to a correlation of (r=0.816) obtained during the experiments conducted by (Li, et al., 2006) to evaluate the STASIS method for sentence similarity in Arabic language; therefore, the proposed adaptations for sentence similarity measure will be used in SGO-CA.

7.2.4. Experiment (4) Selection of Word Similarity Threshold (WST)

As a part of the STASIS method for sentence measurement, (Li, et al., 2006) identified a threshold of (0.2) for word similarity scores; the similarity score between each word pair should be greater than or equal to this threshold in order for their similarity score to be kept in the similarity matrix, otherwise the similarity of word pair is set to 0 in similarity matrix to eliminate the noise in similarity matrix.

Therefore, similarity scores of words which are greater than or equal to this threshold are retained in the similarity matrix and dealt with, otherwise scores are set to (0). The main purpose of this is to reduce the noise in similarity matrix. Details about this threshold and the similarity matrix were discussed earlier in section (6.3.2.3).

The aim of this experiment is to find the optimal results for the word similarity threshold WST which leads to the strongest correlation with human rating. This experiment is conducted by scanning proposed values of WST in the range between (0.0) and (0.5) with an increment of (0.1), the value of (WTS) which leads to the strongest correlation with human judgment will be used for sentence similarity measurement within SGO-CA.

Table (7-4) in appendix (7) show the results of this experiment. The best correlation with human rating (referred to as HR) of (0.886) was obtained when (WST = 0.2). This complies with the value of WST set by (Li, et al., 2003). Therefore the value of word similarity threshold is set to (0.2) in SGO-CA.

7.2.5. Experiment (5): Using Function Words in Similarity Measurement

In section (6.3.4) of this thesis, the researcher proposed adding Arabic function words to the lexical hierarchy and including these words in sentence similarity measurement. More than (60) Arabic function words were added to the lexical tree. Appendix (5) of this thesis contains a complete comprehensive list of Arabic function words and their variations.

The aim of this experiment is to test the following hypotheses:

- H0: Including function words can improve similarity measurement through better correlation with human ratings.
- H1: Including function words cannot improve similarity measurement through better correlation with human ratings.

This experiment is conducted in the sentences dataset (***DS0***), by measuring the semantic similarity of sentence pairs twice, the first when function words are removed and the second is when function words are retained, the approach that leads to higher correlation with human judgment will be followed in sentence similarity calculation within SGO-CA.

The experiment in this section compares the results of incorporating these function words or removing them from the similarity measurement. The best correlation with human ratings obtained in this experiment was (r=0.886) when function words are removed from the sentences. While a correlation of (r=0.7) was obtained when these function words are retained. Therefore the (H0) hypothesis is rejected and function words will be removed from similarity measurement in SGO-CA. Results of this experiment are listed in table (7-5) in appendix (7).

7.2.6. Experiment (6): Including Sentence Difference in Similarity Measurement

In section (6.3.3) of this thesis, the researcher proposed an enhancement to the STASIS method (Li, et al., 2006) by including the difference between two sentences as a factor in the overall sentence similarity calculation.

The aim of this experiment is to test the following hypotheses:

- H0: Sentence difference can improve similarity measurement to offer stronger correlation with human ratings, using these equations:

$$DF(\boldsymbol{U},\boldsymbol{R}) = \frac{\sum_{k=0}^{n} I(X_k)/(COUNT(X_k)+\alpha)}{\sum_{i=0}^{n} I(Y_i)/(COUNT(Y_i)+\alpha)} \qquad (6\text{-}7)$$

$$Sim(\boldsymbol{U},\boldsymbol{R}) = S(U,R) * DF(U,R) \qquad (6\text{-}8)$$

- H1: Sentence difference cannot improve similarity measurement and cannot offer stronger correlation with human rating.

The experiment is performed by comparing human ratings to the results of the two approaches (similarity measurement with and without sentence difference). The first is similarity results without including difference between sentences obtained a correlation of (r=0.886), while the second approach was conducted by including the difference between two sentences in overall similarity calculation, which obtained a stronger correlation of (r= 0.89). Therefore, the null (H0) can be accepted and sentence difference will be included in similarity measurement in SGO-CA. Table (7-6) in appendix (7) shows the results of this experiment.

In this experiment, the results showed the importance of including the sentence difference in measuring similarity due to the high contribution content value of the words with low similarity scores in the similarity vector, and their effect on the measurement.

7.2.7. Experiments Conclusion and Discussion:

Experiments in this chapter gave the researcher an excellent insight on the performance of word and sentence similarity measures, and the proposed tuning and adaptation of these measures. This section summarises the observations made throughout these experiments:

- **Word similarity measure**

  The proposed word similarity measure (6-4) provided stronger correlation of (r=0.9) compared to a correlation of (0.894) using AWSS measure on the evaluation datasets ($\boldsymbol{WS}$).

  The new word similarity measure also showed good correlation of (r = 0.868) on the same dataset ($\boldsymbol{WS}$) by using the lexical tree developed in section (6.2.2) as an information source to evaluate the similarity between words.

  It is notable also from the experiment that the correlation coefficient decreases when using the lexical tree due to the fact that AWSS dataset contains nouns only and was designed to apply Arabic WordNet. Therefore, the role of the lexical tree is not effective. It is expected that the use of lexical tree as information source can enhance word similarity measurement and achieve stronger results with human rating when a dataset of both nouns and verbs are used in the experiments, but unfortunately such dataset has not been published or tested until the time this thesis has written.

- **Sentence similarity measure**

  Adapting the STASIS method for sentence similarity for the Arabic language was done by:

  - Using the lexical tree as an information source
  - Incorporating the new word similarity measure
  - Using an Arabic corpus to calculate word significance
  - Removing function words entirely from compared sentences

  These adaptations from the researcher point of view form the optimum application to evaluate sentence similarity. Testing results have shown an outstanding performance in terms of correlation with human rating (r= 0.886) using the evaluation dataset developed by (O'Shea, et al, 2013). This result comes higher than the correlation coefficient measured by (Li, et al., 2006) which showed (r=0.816). However, more testing is needed to be performed on larger datasets to optimise the sentence similarity measurement.

- **Function words**

  From the experiments, it was found that including function words in STASIS failed to enhance the similarity measurement. The experiments also showed that removing function words from the sentence before performing similarity measurement can give enhanced performance and stronger correlation with human judgment (r=0.886) compared to a weaker correlation of (r=0.7) when function words are included.

The main reason for the poor contribution of function words is due to their frequent repetition in the corpus; this gave them a very low information content values. Consequently, their contribution in similarity scores is low. In addition to that, having many function words in sentences makes the joint word set longer, and lowers the similarity between the two sentences due to the low similarity scores for an increased number of words.

Therefore, a better definition and placement of function words in the lexical tree and altering their frequency in corpus might improve their contribution in sentence similarity.

- **Word Similarity Threshold (WST)**

As explained in section (6.3.2.3) the STASIS method introduced a threshold of (0.2) for word similarity to be stored in the similarity matrix. Word pairs that score less than this threshold are set to a similarity of (0) in the similarity matrix. This threshold is referred to as Word Similarity Threshold (WST) in this thesis. The experiments in this chapter showed the strongest correlation can be obtained(r= 0.886) when (WST = 0.2), this confirms the hypothesis introduced by (Li, et al., 2006).

- **Sentence Difference**

Another limitation of the STASIS method was found when conducting the experiments. STASIS used information content values to signify the contribution of words that occur less frequently than other words in the corpus. However, the contribution of information content values is considered only when these high value words have similar words in the other sentence. Otherwise, their information content values shall be neglected and their score in similarity becomes (0).

In this thesis, the researcher introduced a method for sentence difference to signify the contribution of information content values scoring a similarity of (0) in similarity vectors. The experiment showed that including sentence difference in measuring similarity between sentences resulted in a significant improvement on the test results. Including the difference between sentences obtained stronger correlation with human rating (r= 0.89), compared to a weaker correlation of (r=0.886) when sentence difference is not included.

7.3. Evaluation of SGO-CA

Chapter (5) introduced and discussed an evaluation methodology for pattern matching conversational agents, with evaluation hypotheses, and metrics. Therefore this section shall focus only on evaluation results of SGO-CA.

The evaluation was conducted using a questionnaire similar to the one designed for PMGO-CA (discussed in section 5.2.1.1) which starts with some explanation about SGO-CA and instructions about the evaluation and the domain, and how to test and evaluate the agent. The researcher decided to use the same participants who evaluated the PMGO-CA, as they were experienced in both the domain of the Arabic language, and to ensure a fair comparison between the evaluation results of SGO-CA and PMGO-CA**.**

The participants were requested to read the instructions thoroughly and rate the questionnaire items from (1-5), where (1) shows poor feedback and (5) shows excellent feedback; the questionnaire and instructions can be found in appendix (3) of this thesis

Table (7-3) list the evaluation results of SGO-CA and shows the average human rating for each evaluation metric for the 10 participants.

Metric	**Rating frequency**					**average**	**percent**
	5	**4**	**3**	**2**	**1**		
M1: Responsiveness	1	1	3	4	1	2.7	54%
M2: Conversation length	1	7	2	0	0	3.9	78%
M3: Information accessibility	4	4	2	0	0	4.2	84 %

Metric	Rating frequency					average	percent
M4: Correcting user utterance	0	1	9	0	0	3.1	62 %
M5: CA's understanding of user's utterance	2	5	3	0	0	3.9	78%
M6: Accuracy	2	6	1	1	0	3.9	78%
M7: Conversation consistency	3	2	3	2	0	3.7	74%
M8: Memory	0	6	1	3	0	3.3	66%
M9: Validity	4	5	1	0	0	4.3	86%
M10: Domain coverage	0	6	3	1	0	3.5	70 %

Table 7-3 SGO-CA evaluation questionnaire results

7.3.1. Evaluating Results and Discussion

The purpose of the evaluation of SGO-CA is to measure its performance compared to the PMGO-CA developed in chapter (4) and evaluated in chapter (5). This comparison study is used to establish the base for the development of semantic conversational agents. Table (7-4) lists the evaluation metrics and the average rating for both SGO-CA and PMGO-CA.

Metric	PMGO-CA average	SGO-CA Average
M1: Responsiveness	4.8	2.7
M2: Conversation length	4.2	3.9
M3: Information accessibility	4.1	4.2
M4: Correcting user utterance	3.6	3.1
M5: CA's understanding of user's utterance	3.9	3.9
M6: Accuracy	4.6	3.9
M7: Conversation consistency	4.3	3.7
M8: Memory	3.0	3.3

M9: Validity	4.3	4.3
M10: Domain coverage	3.9	3.5

Table 7-4 Evaluation results for SGO-CA and PMGO-CA

M1: Responsiveness

SGO-CA scored an average of (2.7) in responsiveness metric. This indicates that SGO-CA is more time consuming than PMGO-CA, it is mainly due to the computational complexity associated with mathematical calculations of word and sentence similarity measures, unlike the pattern matching techniques which requires much less computational overhead.

M2: Conversation length

Since both SGO-CA and PMGO-CA use the same knowledge tree to control the dialogue flow, theoretically speaking, they should both score same results. The difference noted between the two results (3.9) for SGO-CA; and (4.2) for PMGO-CA is caused by the higher percentage of misfiring in SGO-CA.

This misfiring is sometimes attributed to the misspelling of Words committed by the user himself which leads to rephrasing or correcting the utterance, and consequently gives an impression of longer conversation compared to PMGO-CA.

M3: Information accessibility

Both CAs were built according to the same architecture, therefore they scored similar average for this evaluation metric, (4.1) for SGO-CA and (4.2) for PMGO-CA, which means that there is no users' preference to use any of the agents.

M4: Correcting user utterance

The structuring for patterns in PMGO-CA is flexible to handle misspelling in user's utterance by using wildcards to replace a letter or a part of the words. While this is not the case in

SGO-CA, where misspell in any of the words of the utterance might cause failure in morphological analysis, and un-recognition of the word giving it a similarity of (0). This is why PMGO-CA scored higher average of (3.6) for this evaluation metric compared to an average of (3.1) for SGO-CA. This can be overcome by spell checking user utterances before sentence similarity calculation, which was not included in SGO-CA to avoid additional computational overhead.

M5: CA understanding of user's utterance

Both SGO-CA and PMGO-CA scored identical results for this evaluation metric, as both CAs can process user's utterances and respond to them but with different levels of accuracy as explained shortly in the discussion of the accuracy metric (M6)

M6: Accuracy

PMGO-CA scored relatively higher average of (4.6) for this evaluation metric compared to (3.9) for SGO-CA. This indicates that SGO-CA has higher rates of misfired responses than PMGO-CA. It is mainly due to the flexibility of patterns scripting, where the scripter has the freedom to add more patterns with wildcards to handle various utterances, unlike semantic similarity which is automatically calculated by the machine.

M7: Conversation Consistency

Due to higher rates of misfired answers, conversation flow in SGO-CA is often interrupted by switching to other contexts or to frequently asked questions giving the impression of inconsistent conversation. Therefore SGO-CA scored lower average of (3.7) for this evaluation metric compared to an average of (4.3) to PMGO-CA.

M8: Memory

Both CAs used the same architecture with the same memory features, therefore they scored similar results in memory management; with (3.0) for PMGO-CA and (3.3) for SGO-CA.

M9: Validity

Regardless of misfired responses, both CAs use the same knowledge tree, therefore the responses given by both CAs scored identical average of (4.3) for this evaluation metric.

M10: Domain Coverage

PMGO-CA scored higher average of (3.9) for this evaluation metric compared to an average of (3.5) for SGO-CA. This is mainly due to the higher rates of non-understandable utterances by SGO-CA which gives an impression of weak coverage for the domain.

7.3.2. Scripts Comparison

Although PMGO-CA seems to have stronger evaluation results, but SGO-CA has bigger advantage over PMGO-CA. In SGO-CA, the scripter might define regular answers only once for each knowledge tree node, which means less or almost no housekeeping except when the knowledge of the domain is changed. While PMGO-CA needs continuous effort to maintain, monitor and modify patterns to accommodate the changes in users' utterances, in addition to that, patterns may sometimes conflict with one another especially when the knowledge tree is big and diverse.

For example, an average of (5) regular answers per node is usually defined in SGO-CA while the average number of patterns in value nodes of PMGO-CA exceeds (30)

Figure (7-1) shows typical regular answers field in SGO-CA which contains three regular answers separated by (|), these regular answers constitute almost all forms of utterances and are evaluated against users' utterances by SGO-CA to decide whether to trigger this node or not. While the same node in PMGO-CA has more than (40) patterns as shown in figure (7-2). This makes PMGO-CA very hard to script and maintain compared to SGO-CA.

```
{
    "ID":2,
    "NType":4,
    "Description":"اريد الحصول على جواز (2)",
    "NodeValue":"اريد جواز جديد",
    "ErrorCount":4,
    "ErrorCountMessage":"لقد قمنا بالاجابة اكثر من مرة لذلك سنضطر لاغلاق المحادثة",
    "TerminateConvOnErrorCount": true,
    "Abuse": false,
    "DisableSearch": false,
    "Answers":"اريد وثيقة سفر جديدة | اريد اصدار جواز جديد | اريد الحصول على جواز جديد",
}
```


Figure 7-1 SGO-CA sample script

```
{
    "ID":2,
    "NType":4,
    "Description":"(2) اريد الحصول على جواز",
    "NodeValue":"اريد جواز جديد",
    "ErrorCount":4,
    "ErrorCountMessage":"لقد قمنا بالاجابة اكثر من مرة لذلك سنضطر لاغلاق المحادثة",
    "TerminateConvOnErrorCount": true,
    "Abuse": false,
    "DisableSearch": false,
    "patterns":"| *جواز*اصدار*اريد | *جواز $ اريد | جواز*احصل* | *جواز*احصل* | جواز اريد
    | *جواز*حصول*اريد | جواز*استخراج*اريد | *جواز*استخراج*اريد | جواز*اصدار*اريد
    | *جواز*استخراج*ستطيع | جواز*اصدار*ستطيع | *جواز*اصدار*ستطيع | جواز*حصول*اريد
    | جديد*جواز*اريد | جواز*حصول*ستطيع | *جواز*حصول*ستطيع | جواز*استخراج*ستطيع
    | جواز*احصل*اريد | *جواز*احصل*اريد | جواز*حتاج* | *جواز*حتاج* | *جديد*جواز*اريد
    | جواز*حصل*تسلون | جواز*حصل*تسلون* | *جواز*حصل*تسلون* | *جواز*طلع* | جواز*طلع*
    | *جواز*استخراج*اريد | جواز*اصدار*اريد | *جواز*اصدار*اريد | *جواز*حصل*تسلون
    | *جواز*اصدار*ستطيع | جواز*حصول*اريد | *جواز*حصول*اريد | جواز*استخراج*اريد
    | جواز*استخراج*ستطيع | *جواز*استخراج*ستطيع | جواز*اصدار*ستطيع
    | *جديد*جواز*اريد | جديد*جواز*اريد | جواز*حصول*ستطيع | *جواز*حصول*ستطيع
    | *جديد*جواز*محتاج | جديد*جواز*محتاج | جديد*جواز*اريد$ | *جديد*جواز*اريد$
    | *جواز*اصدار* | *جواز*اصدار | جواز*اصدار | جديد*جواز*محتاج$ | *جديد*جواز*تاجهح$
    *جواز*حصول* | جواز*حصول* | *جواز*حصول | جواز*حصول | جواز*اصدار*",

}
```

Figure 7-2 PMGO-CA sample scripts

7.3.3. Semantic CAs vs. Pattern Matching CAs

Table (7-5) highlights the comparison between PMGO-CA and SGO-CA. The table is a result of testing and evaluating both CAs using pre-set objective and subjective metrics:

No	Item	PMGO-CA	SGO-CA
1.	Responsiveness	Very short response time (less than 1 second)	High response time (more than 15 seconds)
2.	Correcting user utterance	The flexible nature of pattern matching scripts enables the scripter to have more control and flexibility to write patterns handling variety of users utterances	Semantic CAs are restrictive to spelling, any spelling error may cause the engine not to recognize words, or interpret it as different word, therefor SGO-CA has limitation in this area
3.	CA understanding of users' utterances	PMGO-CA showed very good understanding to users utterances than SGO-CA; PMGO-CA scored (78%) in responding to users utterance regardless whether these response were correctly fired or misfired	SGO-CA also showed the same ability to handle users utterances scoring (78%) in responding to users utterance regardless whether these response were correctly fired or misfired
4.	Accuracy	PMGO-CA showed higher levels of accuracy (92%) in firing the correct response for user utterance than SGO-CA	SGO-CA showed poor accuracy compared to PMGO-CA; SGO-CA scored (78%) of correctly fired responses to users' utterances
5.	Conversation consistency	Conversations carried out with PMGO-CA seemed more consistent than the ones carried out with SGO-CA due to the high accuracy of PMGO-CA; therefor PMGO-CA scored (86%) in conversation consistency	Due to the higher rates of misfiring SGO-CA leads the conversation to incorrect contexts therefore it showed lower level of conversation consistency; therefor SGO-CA scored (0.74%) in conversation consistency
6.	Scripting Complexity	PMGO-CA scripts are complicated and require intellectual challenge to write and maintain, in addition it's very time consuming	SGO-CA scripts are very easy to write and maintain therefore less time consuming

Table 7-5 comparison between PMGO-CA and SGO-CA

7.4. Summary

This chapter covered experiments on words and sentence similarity measures from literature, and experimented with new measures and adaptations over these measures that might be used in Arabic semantic conversational agents.

Evaluation of the new Arabic semantic conversational agent (SGO-CA) was also conducted according to the same methodology used to evaluate (PMGO-CA); although PMGO-CA showed better evaluation results than SGO-CA due to the linguistic complexities of Arabic language and other challenges related to information sources and semantic similarity measures; but being the first semantic CA, (SGO-CA) evaluation results were very encouraging; and it's believed once these challenges are resolved semantic CAs can offer significant improvements over pattern matching in the field of conversational agents. Chapter (8) of this thesis covers some of these challenges with recommendations for other researchers in the field of semantic CAs.

The main contributions of this chapter can be summarized as follows:

- Evaluating the new word similarity measure proposed in chapter (6); and tuning it to obtain strong correlation with human judgment.
- Experimenting adaptation of an existing sentence similarity measure to suit the Arabic language by incorporating the new word similarity measures and using Arabic corpus to estimate the significance of Arabic words.
- Evaluating the proposed measures and their adaptations to develop an Arabic Semantic CA
- Evaluation of SGO-CA on the Iraqi passports domain and compare its metrics with the same metrics used in evaluating the (PMGO-CA).

The work in this chapter showed some promising results in the field of using semantic similarity measures to develop conversational agents. However, these measures can be further improved and adapted to optimise the performance of semantic conversational agents.

The evaluation of SGO-CA covered in this chapter have also shown good results, although pattern matching is still faster than SGO-CA and has better accuracy results, the efforts to maintain SGO-CA is less and easier than PMGO-CA.

From studying the results of evaluation of SGO-CA. the researcher believes that better results can be achieved when completing information sources such as SUMO ontology and WordNet (used to evaluate the semantic similarity), and linguistic tools such as morphological analysis tools. Once all these issues are resolved, semantic conversational agent are expected to outperform pattern matching CAs. The researcher believes that the contributions of this thesis have opened the door wide for other researchers to develop conversational agents in general, and specifically Arabic CAs, and work to resolve their related issues.

Chapter 8
Conclusions and Further Work

The objectives of this research were outlined in section (1.1) of the thesis. The research presented in this thesis began by reviewing existing conversational agents (CAs) with a special focus on Arabic goal oriented conversational agents.

At first the research covered a general methodology for CAs' development, starting with knowledge engineering process, architecture design, implementation and evaluation of CAs. Each process in this methodology was discussed and elaborated in chapter (4) of this thesis.

The knowledge engineering process for the Arabic goal oriented CA is concerned with gathering and modelling information and procedures of the proposed domain (The Iraqi passport laws and regulations used in this research) and transforming this information to shape the knowledge tree. This is a new approach of structuring knowledge for an Arabic domain for the purpose of conversational agent domain design.

A new architecture, with four main components (tree engine, short-term memory, long-term memory, and matching engine) was designed to develop both Arabic (semantic and pattern matching) goal-oriented CAs and their tools, these CAs were tested and evaluated for their viability, adaptability, flexibility, accessibility, and other criteria.

Tree engine is used to process the domain knowledge tree and control the dialogue flow between users and the CA. This engine has also an access to the matching engine, which evaluates users' utterances against defined nodes within the knowledge tree.

Short-term and long-term memory are components used to keep the activated nodes for both short and long term to be retrieved either during the same conversation, or later when the same user initiates a new conversation.

The implementation of CAs was also studied in depth; in the course of this work two types of conversational agents were developed. The first is a pattern matching goal-oriented CA

(PMGO-CA) developed in chapter (4). While the second is a semantic goal-oriented CA (SGO-CA) developed in chapter (6) which utilises semantic similarity measures instead of pattern matching techniques to respond to users' utterances.

Pattern matching techniques used to construct CA were discussed in this thesis, including new scripting language, algorithms for the Arabic pattern matching CA (PMGO-CA) and conflict resolution strategies. Although PMGO-CA showed an outstanding performance in terms of speed and accuracy, the process of scripting patterns and maintaining the CA is still labour intensive, as it is difficult for the scripter to predict all users' utterances.

Semantic similarity between words and sentences were examined, this research also highlighted the problems and complexities of developing semantic CAs for the Arabic language. An adapted sentence similarity measure was introduced incorporating a new measure for word similarity; these measures were used to construct the Arabic semantic CA (SGO-CA)

An information source called the "lexical tree" was also developed in this research, this tree utilised the mapping between Arabic WordNet and SUMO ontology concepts. It was used to calculate semantic similarity between words.

The similarity measures proposed were tested and tuned throughout empirical experiments to suit the Arabic language and the new information source. Experiments on words similarity were conducted using a dataset of Arabic nouns, while experiments on sentences used another dataset for sentence pairs.

Human evaluation for both pattern matching CA (PMGO-CA) and semantic CA (SGO-CA) based on a new evaluation methodology was also conducted in this work and a comparative study between the two types of CAs was performed based on the results of this evaluation.

8.1 Research Contributions

The research conducted in this thesis offers the following contributions to the knowledge in the field of Arabic CAs:

8.1.1 Knowledge Trees

Adapting knowledge trees and using them for the construction of goal orientated Arabic conversational agents. Although knowledge trees have been used in previous researches on conversational agents, this research is the first of its' kind to use knowledge trees for the Arabic CAs. In addition, this thesis provided modifications for these knowledge trees by introducing algorithms for short term memory, dialogue flow control, and mechanisms for context switching and nodes activation and deactivation. The new knowledge tree constructed in this research was simple, user friendly and adaptable for any type of domains.

8.1.2 Evaluation Methodology for Conversational Agents

Chapter (5) of this thesis introduced a new methodology to evaluate PMGO-CA. The new methodology focused on measuring CAs' performance through subjective and objective metrics. Those metrics were selected to reflect the usability, flexibility, accessibility, adaptability of the CAs that enables them to interact with users and offer good service.

8.1.3 Construction of Arabic Semantic CA

This thesis introduced the first Arabic Semantic Goal-Oriented Conversational Agent (SGO-CA) to overcome weaknesses of the pattern matching technique. The new CA which is the first of its type has significantly reduced the scripting complexities and the continuous maintenance of the PMGO- CA. It was evaluated in chapter (7) by experts in the domain to test its viability, response and compliance with the official laws and regulations of the Iraqi passport domain. Results of the testing were positive and clearly illustrated the effectiveness of the technique.

The evaluation was conducted using the same metrics developed for PMGO-CA evaluation; these evaluation metrics focused on user satisfaction criteria such as responsiveness, accuracy, accessibility and domain coverage.

8.1.4 Long-term Memory Management in CA

The application of long-term memory in CAs was introduced also for the first time. This memory was used to recognise users based on their personal information (such as name, date and place of birth and location). The memory proved to be effective when retrieving information about users' utterances queries and questions.

8.1.5 Utilising SUMO Mapping with Arabic WordNet

Previous research on semantic similarity measurement focused on the WordNet database to measure the similarity between nouns. In this thesis an alternative approach was created by including the mapping between WordNet and concepts encoded within the SUMO ontology. This mapping resulted in developing a new information source called the "lexical tree" that includes verbs and adjective, in addition to nouns. The new information source was used successfully in this work to evaluate the similarity between words.

8.1.6 New Measure for Word Semantic Similarity

A novel word similarity measure (6-4) was developed in this thesis, it obtained stronger correlation to human ratings than other measures covered in the literature. The correlation coefficient obtained by this measure was (r=0.9), compared to (r=0.894) obtained by the AWSS similarity measure (6-2) developed by (Almarsoomi, et al., 2013).

The new word similarity measure was also applied to the new information source (lexical tree) developed in this work and scored a very good correlation of (r=0.868).

8.1.7 Including Sentence Difference in Sentence Similarity Measurement

This research was the first to test sentence difference as part of sentence similarity measurement in CAs. An experiment in chapter (7) showed that including sentence difference can lead to stronger correlation with human rating (r=0.89) compared to a correlation of (r=0.884) when sentence difference is not in use.

8.1.8 Conversational Agent Development Tools

A new set of software tools were developed to make CAs scripting easier and less labour intensive; these tools were developed to keep all the options available for the scripter in one place.

The tools include a tree script editor tool to maintain the knowledge tree files, short and long-term memory management tools, PMGO-CA and SGO-CA management tools integrated to manage both PMGO-CA and the SGO-CA and the lexical tree.

The researcher believes that these tools are the first of their kind for the Arabic language, and they provide the facilities to observe the behaviour of the knowledge tree, pattern and semantic engines, in addition to the ability to observe all calculations performed by the system.

8.1.9 Adaptability

Although the concept of knowledge trees has been introduced earlier in literature, but this research was the first to utilise it for the Arabic CAs. The use of knowledge tree files has significantly contributed into making Arabic CAs more adaptable for multiple domains.

Adaptability can be achieved by collecting and engineering the new domain information, and transforming it to create a new knowledge tree to operate the CA.

The use of semantic information source (lexical tree based on mapping between WordNet and SUMO ontology) shapes another form of adaptability. Once this tree is complete and validated, it can be used for any semantic CA for the Arabic language regardless of the domain.

8.2 Research Questions

This section answers the research questions and aims and objectives raised in sections (1-1) and (1-2) respectively, they are:

1. **Can pattern matching CAs be used effectively for Arabic language in a domain of interest?**

   Related objective: (1)

   From the general review of the goal oriented CAs, especially the Arabic pattern matching CAs, and the development of the new architecture, the researcher found that encouraging results were obtained through the evaluation of PMGO-CA; PMGO-CA showed very good levels of performance, responsiveness, accuracy, adaptability, and domain coverage.
2. **Is it possible to develop an Arabic semantic conversational agent?**

   Related objectives: (2) and (5)

   From the research into semantic word and sentence similarity in both English and Arabic languages, An Arabic semantic CA (SGO-CA) has been successfully developed. Evaluation of this SGO-CA showed encouraging results.
3. **Does the semantic CA introduce significant improvements over pattern matching CAs?**

   Related objective: (5) and (7)

   The semantic CA (SGO-CA) developed in this work showed significant improvements in terms of reducing scripting complexity and CA maintenance.

4. **Is it possible to simulate human memory throughout the CA?**

   Related objectives: (3) and (6)

   The attempt made in this research to build a memory to identify users based on their personal information (such as their names, age and location) showed promising results.

5. **Can these pattern matching or semantic CAs cover an entire domain of interest and help user gain information about it?**

   Related objective: (4)

   Evaluation results showed that both PMGO-CA and SGO-CA covered almost all related items related to the domain.

6. **Are existing methods for sentence similarity suitable to be embedded within an Arabic semantic CA?**

   Related objective: (2)

   Embedding semantic sentence similarity within Arabic CAs showed promising results in this thesis. However, some further research is needed improve the techniques of semantic sentence similarity and performance of the CA.

8.3 Future Work

Being the first research to study the Arabic semantic goal-oriented CA, the researcher has encountered some challenges and issues. Some of these challenges were related to the nature of the Arabic language, others related to the available information sources of the Arabic language, the philosophy of semantic similarity measures, and in addition to some other technical challenges.

The researcher believes that this thesis can be considered as a good reference for those interested in the field of Arabic semantic similarity and Arabic conversational agents. This section highlights some recommendations for other future researchers in this field to focus on and study:

- Completion of information sources:

  It is recommended to spend more efforts to complete the information sources, especially for the Arabic language such as Arabic WordNet (AWN) and SUMO ontology. This will encourage more researchers to investigate and develop Arabic CAs.

- Arabic function words:

  Some focus is needed to include function words in the AWN as they have rich semantic information. The attempts by the researcher to include them in measuring semantic similarity showed negative results due to their high frequency of occurrence in the Arabic corpus.

- Using a spell checking technique to correct users' utterances before processing them by the semantic similarity engine.
- Incorporating a method of Word Sense Disambiguation (WSD) in semantic CAs to determine the intended word sense within the given context among many different other word senses with the same spelling. This shall reduce the number of misfiring and resulted in more correct regular answers.

Table of References

Abney, S. P. 1994. *Parsing By Chunks.* Bell Communications Research. 1994.

Abouenour, L., Bouzoubaa, K. and Rosso, P. *IMPROVING Q/A USING ARABIC WORDNET.*

Achananuparp, P., Achananuparp, P. and S., Xiajiong. 2008. *The Evaluation of Sentence Similarity Measures.* College of Information Science and Technology. 2008.

Acomb, K., et al. *Technical Support Dialog Systems: Issues, Problems, and Solutions.* SpeechCycle . New York

Agirre, E., Lopez, O. and A., Soroa. 2009. *Knowledge-Based WSD on Specific Domains: Performing Better than Generic Supervised WSD.* University of the Basque Country. 2009.

Agirreand, E., De Lacalleand, O. and Soroa, A. 2009. *Knowledge-Based WSD on Specific Domains: Performing Better than Generic Supervised WSD.* 2009.

Al Ameed, H.K., et al. 2006. *Arabic Search Engines Improvement: A New Approach using Search Key Expansion Derived from Arabic Synonyms Structure.* 2006.

Alcock, Keith. 2004. WordNet Relationship Browser. [Online] 2004. http://www.keithalcock.com.

ALICE. 1995. *ALICE, Artificial Intelligence Foundation.* [Online] 1995.

Aljlayl, M. and Frieder, O. 2002. *improving the retrieval effectiveness via a light stemming.* 2002.

Allen J. 1995. *Natural Language Understanding.* s.l. : Benjamin/Cummings, 1995.

Almarsoomi, F., et al. 2014. *Arabic Word and Semantic Similarity Measures.* School of Computing, Maths and Digital Technology, Manchester Metropolitan University. 2014.

Almarsoomi, F., et al. 2012. *AWSS: An Algorithm for Measuring Arabic Word Semantic Similarity.* 2012.

—. 2013. *AWSS: An Algorithm for Measuring Arabic Word Semantic Similarity.* Department of Computing and Mathematics, Manchester Metropolitan University. 2013.

Altabbaa, M., Al-Zaraee, A. and M., Shukairy. 2010. *An Arabic Morphological Analyzer and Part-Of-Speech Tagger.* department of Informatics Engineering , Arab International University. Damascus, Syria : s.n., 2010.

Altantawy, M., Habash, N. and Rambow, O. 2011. *Fast Yet Rich Morphological Analysis.* 2011.

Andersen, E., Ø. 2011. *Grammatical error prediction.* Cambridge, Uk : s.n., 2011.

Anna3. *http://www.chatbots.org/virtual_assistant/anna3/.*

AraMorph. *http://www.nongnu.org/aramorph/english/index.html.*

—. 2003. *http://www.nongnu.org/aramorph/english/index.html.* 2003.

Bach, N., Badaskar, S. *A Review of Relation Extraction.* Canergie Mellon University, Canergie Mellon University, Language Technologies Institute.

Baig, F., Baig, S. and Khan, M. 2012. *Controlling Home Appliances Remotely through Voice Command.* 2012.

Batet, M., et al. 2013. *Semantic Similarity Estimation from Multiple Ontologies.* 2013.

Bickmore, T and Giorgino, T. 2006. *Health dialog systems for patients and consumers.* 2006.

Bin Saleh, L. and Al-Khalifa, H. 2009. *AraTation: an Arabic semantic annotation tool.* 2009.

Bishop, M. 1996. *Neural Networks for Pattern Recognition.* Oxford, England : Oxford University Press, 1996.

BLACK, W., ELKATEB, S. and VOSSEN, P. 2006. *Introducing the Arabic WordNet Project.* 2006.

Brachman, J. 1983. *What IS-A is and isn't. An Analysis of Taxonomic Links in Semantic Networks.* 1983.

Brihaye, P. 2003. *AraMorph.* [Online] 2003. http://www.nongnu.org/aramorph/english/index.html.

Brill, E. 1992. *A Simple Rule-Based Part of Speech Tagger.* Department of Computer Science, University of Pennsylvania. Pennsylvania, U.S.A. : s.n., 1992.

Buckwalter, T. 2002. Arabic Morphology Analysis. [Online] 2002. http://www.qamus.org/morphology.htm.

Callan, P., J. , Croft W., Harding, M., S. 1992. *Database and Expert Systems Applications, Chapter 3: The INQUERY Retrieval System.* Massachusetts, USA : s.n., 1992. ISBN: 978-3-211-82400-9 (Print) 978-3-7091-7557-6 (Online).

Carpenter, R. 2006. *Jabberwacky. http://jabberwacky.com.* 2006.

Cassell, J. 2000. *Embodied conversational interface agents.* 2000.

Catrambone, R., Stasko, J. and Xiao, J. 2002. *Anthropomorphic Agents as a User Interface Paradigm.* 2002.

Chapman, W., Bridewell, W., Hanbury, P., Cooper F. 2001. *A simple algorithm for identifying negated ?ndings and diseases in discharge summaries.* 2001.

Chatbots.org. 2005. Spleak. *Chatbots.org.* [Online] 2005. http://www.chatbots.org/chatbot/spleak/.

Chowdhury, G. 2005. *Natural Language Processing.* s.l. : Dept. of Computer and Information Sciences University of Strathclyde, 2005.

Colby, K. 1975. *Artificial Paranoia: A Computer Simulation of Paranoid Process.* New York : Pregamon Press, 1975.

Crockett, K., et al. 2009. *Bullying and Debt: Developing Novel Applications of Dialogue Systems.* 2009.

Crockett, K., O'Shea, J. and Bandar, Z. 2011. *Goal Orientated Conversational Agents: Applications to benefit society.* The Intelligent Systems Group, School of Computing, Mathematics & Digital Technology, Manchester Metropolitan University. 2011.

Curran., J. 2002. *Ensemble menthods for automatic thesaurus extraction.* 2002. In Proc. of EMNLP.

Datamuse. 2003. *FreeNet.* [Online] 2003. http://www.lexfn.com.

Demberg, V., Moore D.J. 2006. *Information Presentation in Spoken Dialogue Systems.* University of Edinburgh, University of Stuttgart. 2006.

Diab, M., Hacioglu, K., Jurafsky, D. 2007. *Automated methods for processing Arabic text: From tokenization to base phrase chunking.* s.l. : Kluwer/Springer, 2007.

Duin, P. R., Pekalska E. 2007. *The Science of Pattern Recognition Achievements and Perspectives.* 2007.

El Kholy, A., N Habash. 2010. *Techniques for Arabic morphological detokenization and orthographic denormalization.* 2010. In proceeding of: Workshop on Semitic Languages. .

Elkateb, S., Black, W. and Farwell, D. 2006. *Arabic WordNet and the Challenges of Arabic.* 2006.

Fellbaum, C. 1990. *English Verbs as a Semantic Net.* Princeton University. 1990.

Fellbaum, C., Gross, D. and Miller, K. 1993. *Adjectives in WordNet.* 1993.

Fong, S. 2003 . WordNet Connect. [Online] 2003 . http://dingo.sbs.arizona.edu/~sandiway/wnconnect/index.html.

Forbes-Rileya, K. and Litmanb, D. 2009. *Designing and evaluating a wizarded uncertainty-adaptive spoken dialogue tutoring system.* 2009.

Francis, W. and Kucera, H. 1979. *Brown corpus manual.* Department of Linguistics, Brown University . 1979.

Frederick, H.R. 1985. *Rule-based systems, Communications of the ACM.* 1985. pp. 921-924. Vol. 28.

Free Software Foundation, Inc. 2007. GNU.org. [Online] 2007. http://www.gnu.org/copyleft/gpl.html.

Friedman, J., Hastie, T. and R., Tibshirani. 2001. *The Elements of Statistical Learning.* New York : Springer, 2001.

Graesser, A., Chipman, P., Haynes, B.C., Olney, A. 2005. *An Intelligent Tutoring System With Mixed-Initiative Dialogue.* 2005. pp. 612--618.

Green, S. and Manning, C. 2010. *Better Arabic Parsing: Baselines, Evaluations, and Analysis.* Computer Science Department, Stanford University. 2010.

Gruber, R. 1993. *A Translation Approach to Portable Ontology Specifications.* Stanford University. 1993.

Gruber, T. 2008. Ontology. [Online] 2008. http://web.dfc.unibo.it/buzzetti/IUcorso2007-08/mdidattici/ontology-definition-2007.htm.

Gupta, N., Mathur P. 2012. Spell Checking Techniques in NLP: A Survey. December 2012, Vol. 2, 12.

Gupta, S., Manning, D., C. 2014. *SPIED: Stanford Pattern-based Information Extraction and Diagnostics.* Department of Computer Science, Stanford University. 2014.

Habash, N., Soudi, A. and Buckwalter, T. 2007. *On Arabic Transliteration.* 2007.

Harabagiuo, S. and Narayanan, S. 2004. *Question Answering Based on Semantic Structures.* University of Texas; International Computer Science Institute;. 2004.

Harbusch, k., Knapp, M. and Laumann, C. *Modelling user-initiative in an automatic help desk system.* Computer Science Department, University Koblenz-Landau . Koblenz, GERMANY : s.n.

Haugeland, John C. 1989. *Artificial Intelligence : The Very Idea.* 1989.

Hayes, P. and Ford, K. 1995. *Turing Test Considered Harmful.* 1995.

Hazen T., Glass J. 1998. *TELEPHONE-BASED CONVERSATIONAL SPEECH RECOGNITION IN THE JUPITER DOMAIN.* Spoken Language Systems Group Laboratory for Computer Science , Massachusetts Institute of Technology Cambridge. Massachusetts, USA : s.n., 1998.

Hijjawi, M. 2011. *ArabChat: an Arabic Conversational Agent November.* Manchester Metropolitan University. 2011.

Hliaoutakis, A., et al. 2006. *Information Retrieval by Semantic Similarity.* Technical University of Crete; Dalhousie University;. 2006.

Ide, N. and Erjavec, T. 2002. *Sense Discrimination with Parallel Corpora.* 2002.

Jensen., F. V. 1996. *An Introduction to Bayesian Networks.* New York : Springer Verlag, 1996.

Jeong, B. 2005. *Machine-Learning based Semantic Similarity Measures to Assist Discovery and Reuse of Data Exchange XML Schema.* Department of Industrial and Management Engineering, Pohang University of Science & Technology . 2005.

Jiang, J. and Conrath, D. 1997. *Semantic similarity based on corpus statistics and lexical taxonomy.* 1997.

Kao, A., Poteet, S. 2007. *Natural Language Processing and Text Mining.* s.l. : pringer-Verlag London Limited, 2007.

Khoja, Shereen. 1999. *Stemming Arabic Text http://zeus.cs.pacificu.edu/shereen/research.htm.* 1999.

Kopp, S., et al. 2005. *A Conversational Agent as Museum Guide- Design and Evaluation of a Real-World Application.* A.I Group, University of Bielefeld. Germany : s.n., 2005.

Landauer, T., Foltz, P., Laham D. 1998. *An Introduction to Latent Semantic Analysis.* Department of Psychology, University of Colorado at Boulde. 1998.

Larkey, L., Connell, M., Ballesteros L. 2002. *Light Stemming for Arabic Information Retrieval.* 2002.

Latham, A., Crockett, K., Bandar, Z. 2010. *A Conversational Expert System Supporting Bullying and Harassment Policies.* 2010. pp. 163 --168.

Latham, A., et al. *Adaptive Tutoring in an Intelligent Conversational Agent System.*

Leacock, C. and Chodorow, M. 1998. *Combining local context and WordNet similarity for word sense identification.* 1998.

Lester, J., Branting, k., Mott, B. 2004. *Conversational Agents.* 2004.

Li, X., Szpakowicz S., Matwin S. 1995. *A WordNet-based Algorithm for Word Sense Disambiguation.* 1995.

Li, Y., Bandar, Z. and McLean, D. 2003. *An approach for measuring semantic similarity between words using multiple information sources.* 2003.

Li, Y., et al. 2006. *Sentence Similarity Based on Semantic Nets and Corpus Statistics.* 2006.

Li, Y., Mclean, D. and Bandar, Z. 2006. *A.Sentence similarity based on semantic nets and corpus statistics.* 2006.

Liddy, Elizabeth D. Liddy & Jennifer H. 2001. *An NLP Approach for Improving Access to Statistical Information for the Masses.* School of Information Studies, Center for Natural Language Processing, Syracuse University. 2001.

Lin, D. 1998. *An information-theoretic definition of similarity.* 1998.

—. 1998. *An information-theoretic definition of similarity.* 1998. pp. 296-304, In Proc. of the 15th ICML.

Liu, H. and Wang, P. 2013. *Assessing Sentence Similarity Using WordNet based Word Similarity.* 2013.

Liu, X., Zhou, Y. and Zheng, R. 2007. *Measuring Semantic Similarity in Wordnet.* 2007.

Liu, Y., et al. 2007. *Using WordNet to Disambiguate Word Senses for Text Classification.* 2007.

Liu, Y., et al. 2007 . *Using WordNet to Disambiguate Word Senses for Text Classification.* 2007 .

Luce, R., Raiffa, H. 1957. *Games and Decisions.* New York : Wiley, 1957.

M., El Bachir and Alsaeedan, W. 2012. *Genetic Algorithm for Arabic Word Sense Disambiguation.* 2012.

Madnani, N. 2007. Getting Started on Natural Language Processing with Python. 2007, Vol. Volume 13, Issue 4.

Maimon, O. and Rokach, L. 2002. *Handbook of Data Mining and Knowledge Discovery.* New York, USA : Oxford University Press, 2002.

Malandrakis, N., et al. 2013. *Distributional Semantic Models for Affective Text Analysis.* 2013.

Manning, Jenny Rose Finkel and Christopher D. *Nested Named Entity Recognition.* Computer Science Department, Stanford University.

Massaro, D., et al. *A Multilingual Embodied Conversational Agent for Tutoring Speech and Language Learing.* Department of Psychology, Perceptual Science Laboratory , University of California, Santa Cruz.

Mazroui, Azzeddine. 2014. Alkhalil Morphology System. [Online] 8 June 2014. Last accessed on May, 2014. http://sourceforge.net/projects/alkhalil/.

Michalski, I., Mozetic, J. and Lavrac, H. 1986. *The multi-purpose incremental learning system AQ15 and its testing application to three medical domains.* 1986. In Proceedings of the fifth National Conference on Articial Intelligence .

Miller, G. 1994. *WordNet: A Lexical Database for English.* s.l. : Communications of the ACM , 1994. Vol. Vol. 38.

Miller, G., et al. 1990. *Introduction to WordNet: An On-line Lexical Database.* 1990.

Mitchell, T. M. 1997. *Machine Learning.* McGraw-Hill, Portland : s.n., 1997.

Mohtasseb, H. and Issa, K. Sarf - Arabic Morphology System. [Online] [Cited: 8 June 2014.] Last accessed on May, 2014. http://sourceforge.net/projects/sarf/.

Monroe W., Green S.,Manning .D .C. 2014. *Word Segmentation of Informal Arabic with Domain Adaptation.* Computer Science Department, Stanford University. 2014.

Morency, L. and Darrell, T. 2005. Dialog Context for Head Gesture Recognition. *CASIL Research Abstracts.* [Online] 2005. http://publications.csail.mit.edu/abstracts/abstracts05/lmorency/lmorency.html.

Nadkarni, P., Ohno-Machado L., Chapman W. 2011. *Natural language processing: an introduction.* 2011.

NetOwl. 2014. NetOwl. [Online] 2014. Last accessed on May, 2014. http://www.netowl.com/.

Nilsson., R. 1971. *STRIPS: A new approach to the application of theorem proving to problem solving.* 1971. pp. 189-208.

Nolt, J., Rohatyn, D. and Varzi, A. 2011. *Schaum's Outline of Logic.* second edition. s.l. : McGraw-Hill, 2011.

Noori, Z., Bandar, Z. and Crockett, K. 2014. *Arabic Goal-oriented Conversational Agent,Based on Pattern Matching and Knowledge Trees.* 2014.

O'Shea, J., et al. 2008. *A Comparative Study of Two Short Text Semantic Similarity Measures .* 2008.

O'Shea, K., Bandar, Z. and K., Crockett. 2010. *A Conversational Agent Framework using Semantic Analysis.* Manchester Metropolitan University. United Kingdom : s.n., 2010.

O'Shea, K. Crockett, k. Bandar, Z, O'Shea, *Natural language scripting within conversational agent design*. J. Applied Intelligence, January 2014, Volume 40, Issue 1, pp 189-197.

The online version of the original article can be found under doi:10.1007/s10489-012-0408-2

O'Shea, J., Bandar, Z. and Crockett, K. 2011. *System Engineering and Conversational Agents.* s.l. : Springer, 2011.

O'shea, J., Bander, Z. and Crockett, K. 2013. *A New Benchmark Dataset with Production Methodology for Short Text Semantic Similarity Algorithm.* 2013.

O'shea, K. 2012. *An approach to conversational agent design using semantic sentence similarity.* Manchester Metropolitan University. 2012.

O'Shea, K., Bandar, Z. and Crockett, K. 2009. *A semantic-based conversational agent framework.* Manchester Metropolitan Univ. Manchester, UK : s.n., 2009.

Owda, M., Bandar, Z. and Crockett, K. 2011. *Information Extraction for SQL Query Generation in the Conversation-Based Interfaces to Relational Databases.* s.l. : Springer, 2011. pp. 44--53.

P. Madeira, M. Mourão, N. Mamede. 2003. *STAR Frames - A step ahead in the design of conversational systems capable of handling multiple domains , in Proc. ICEIS.* Angers, France : s.n., 2003.

Papineni, K., Roukos, S., Ward, R. T, Zhu, W.-Z. 2002. *BLEU: A Method for Automatic Evaluation of Machine Translation.* Philadelphia, PA. : s.n., 2002. Proceedings of the 40th Annual Meeting of the ACL.

Pawlak, Z. 1991. *Rough Sets - Theoretical Aspects of Reasoning about Data.* Dordrecht : Kluwer Academic Publishers, 1991.

Pease, A. and Benzmüller, C. 2013. *Sigma: An Integrated Development Environment for Formal Ontology.* 2013.

Pease, A. 2011. OntologyPortal . *Ontology Portal.* [Online] 2011. http://www.ontologyportal.org/.

—. 2009. Standard Upper Ontology Knowledge Interchange Format. [Online] 2009. http://sigmakee.sourceforge.net.

Pedersen, T. 2007. WordNet Similarity. [Online] 2007. http://wn-similarity.sourceforge.net.

Poole, D. and Mackworth, A. 2010. *ARTIFICIAL INTELLIGENCE - FOUNDATIONS OF COMPUTATIONAL AGENTS.* [Online] 2010. http://artint.info/html/ArtInt_216.html.

Preece, J., Rogers, Y., Sharpe, H. 2002. *Interaction design: Beyond human-computer interaction.* New York : Wiley, 2002.

Princeton University. 2014. *http://wordnet.princeton.edu/wordnet/man/wnstats.7WN.html.* 2014.

—. 2005. WordNet - A lexical database for English. *Princeton University.* [Online] 2005. http://wordnet.princeton.edu.

Princeton University, P. 2005. WordNet - A lexical database for English. *Princeton University.* [Online] 2005. http://wordnet.princeton.edu.

protégé. 2014. [Online] 2014. http://protege.stanford.edu/.

Quinlan, J. R. 1993. *C4.5: Programs for Machine Learning.* San Mateo, CA : Morgan Kaufmann, 1993.

Rada, R., et al. 1989. *Development and application of a metric on semantic nets.* 1989.

Ramirez, C. and Valdes, V. 2012. *Advances in Knowledge Representation, chapter 3: A General Knowledge Representation Model of Concepts.* s.l. : InTech, 2012. p. 49. ISBN: 978-953-51-0597-8.

Ramiz, M. 2009. *Expert Systems with Applications, Chapter three: A new sentence similarity measure and sentence based extractive technique for automatic text summarization.* 2009. pp. 7764–7772.

Razmerita, L., et al. 2004. *INCA: AN INTELLIGENT COGNITIVE AGENT-BASED FRAMEWORK FOR ADAPTIVE AND INTERACTIVE LEARNING.* Centre for Advanced Learning Technologies, Fontainebleau France,INRIA, The National Institute for Research in Computer Science, American University of Paris. 2004. Proc. the IADIS International Conference Cognition and Exploratory Learning in Digital Age.

READ, J., DRIDAN, R., OEPEN, S., SOLBERG, L. 2012. *Sentence Boundary Detection: A Long Solved Problem.* , Department of Informatics, University of Oslo. 2012.

Resnik, P. 1995. *Using information content to evaluate semantic similarity.* 1995.

Resnik., P. 1999. Semantic similarity in a taxonomy: An information based measure and its application to problems of ambiguity in natural language. 1999.

Richardson, D., Dolan, B. and Vanderwende, L. 1998. *MindNet: acquiring and structuring semantic information from text.* WA, U.S.A. : Microsoft Research, 1998.

Richardson, D., S., Dolan, B., W., Vanderwende L. 1998. *MindNet: acquiring and structuring semantic information from text.* WA, U.S.A. : Microsoft Research, 1998.

Rindflesch, C., T., Aronson, R., A. 1994. *Ambiguity resolution while mapping free text to the UMLS Metathesaurus.* 1994.

Rodríguez, H., et al. 2008. *Arabic WordNet: Semi-automatic Extensions using Bayesian Inference.* 2008.

Rosenblueth, A., Wiener, N. and J., Bigelow. 1943. *Behavior, purpose, and teleology. Philosophy of Science.* 1943. pp. 18-24.

Rosenfeld, R. 1996. *A maximum entropy approach to adaptive statistical modelling. Computer Speech and Language.* 1996.

Rosenfeld, R., et al. 2000. *TOWARDS A UNIVERSAL SPEECH INTERFACE .* School of Computer Science , Carnegie Mellon University. 2000.

Roukos, S. and Graff, D., Melamed, D. 1997. *Hansard French/English.* 1997. 1-58563-048-9.

Rozinaj, G., Páleník, A., Cepko, J., Juhár, J., Cižmár, A. , Rusko, M. , Jarina R. 2012. *The First Spoken Slovak Dialogue System Using Corpus Based TTS.* Department of Telecommunications . Bratislava : s.n., 2012.

Ruppenhofer, J., et al. 2010. *FrameNet II: Extended Theory and Practice.* 2010.

Russell., S. J. 1989. *Use of Knowledge in Analogy and Induction.* s.l. : Morgan Kaufmann, 1989.

Ryding. 2005. *A Reference Grammar of Modern Standard Arabic.* 2005.

Sammut, C. 2001. *Managing Context in a Conversational Agent.* School of Computer Science and Engineering, University of New South Wales. Sydney, Australia : s.n., 2001.

Saygin, A., Cickeli, I. and AKMAN, V. 2000. *Turing Test: 50 Years Later.* 2000.

Semeraro, G., et al. 2003. *Evaluation and Validation of a Conversational Agent Embodied in a Bookstore.* 2003.

Seneff, S., Hurley E., Lau, R., Pao C., Schmid P., Zue V. 1998. *GALAXY-II: A REFERENCE ARCHITECTURE FOR CONVERSATIONAL SYSTEM DEVELOPMENT.* Spoken Language Systems Group Laboratory for Computer Science, Massachusetts Institute of Technology Cambridge. Massachusetts, USA. : s.n., 1998.

Shadbolt, N. and Milton, N. 1999. *From Knowledge Engineering to Knowledge Management.* Artificial Intelligence Group, University of Nottingham. 1999.

Shah, H. 2011. *Turing's misunderstood imitation game.* 2011.

Shapiro, S. 1978. *PATH-BASED AND NODE-BASED INFERENCE IN SEMANTIC NETWORKS.* Department of computer science, State University of New York. 1978.

Shawar, B., Atwell E. 2007. *Chatbots: are they really useful.* 2007.

Silvervarg A., Jonsson, A. 2011. *Subjective and Objective Evaluation of Conversational Agents in Learning Environments for Young Teenagers.* Department of Computer and Information Science., Link¨oping University. Link¨oping, Sweden : s.n., 2011.

Souza, M. and Ferreira, M. 2002. *Designing reusable rule-based architectures with design patterns,Expert Systems with Applications.* 2002. pp. 395-403. Vol. Vol. 23.

Susie M. Stephens, Jake Y. Chen, Marcel G. Davidson, Shiby Thomas and Barry M. Trute. *a platform for BLAST search and Regular Expression pattern matching in life sciences.*

Taghva, K., Elkoury, R., and Coombs, J. 2005. *Arabic Stemming without a root dictionary, http://www.isri.unlv.edu/publications/isripub/Taghva2005b.pdf.* 2005.

TextWise. 2014. *TextWise.* [Online] 2014. Last accessed on May, 2014. http://www.textwise.com/.

The Global WordNet Association. 2014. [Online] 2014. http://globalwordnet.org/arabic-wordnet/awn-browser/.

Tian, Y., Lo, D. and Lawall, J. 2014. *Automated Construction of a Software-Specific Word Similarity Database.* 2014.

Trappey, A. 2006. *Knowledge Engineering Applying Knowledge-based Expert System to Develop an Intelligent Collaborative Design Platform.* 2006.

Trovato, A. and Pelletier, D. 1998. *Portable system for providing voice driving directions.* 1998.

Turing, A. M. 1950. *Computing machinery and intelligence.* 1950.

Valle, K. 2010. *The Credibility of Embodied Conversational Agents (ECAs) .* 2010.

Vanderhulst, G. 2005. *Mapping Real Domain Ontologies to SUMO: a Case Study.* 2005.

Vossen, P. 1997. *EuroWordNet: a multilingual database for information retrieval.* 1997.

Walker, J.H., Sproull, L., & Subramani, R. 1994. *Using a human face in an interface.* Boston, MA : s.n., 1994. pp. 85-91.

Walker, M, et al. 1998. *Evaluating Spoken Dialogue Agents with PARADISE.* s.l. : AT&T labs, 1998.

Walker, M. A. et al. 2001. *DARPA communicators dialog travel planning system.* Denmark : s.n., 2001. pp. 1371-1374.

Walker, M., et al. 1997. *Evaluating Competing Agent Strategies for a Voice Email Agent.* 1997.

Wallace, R. 2003. *The Elements of AIML Style.* ALICE A.I Foundation. 2003.

Weeber, M., Mork J., Aronson AR. 2001. *Developing a Test Collection for Biomedical Word Sense Disambiguation.* 2001.

Weik, Martin H. 2000. *Computer science and communications dictionary.* 2000.

Weizenbaum, J. 1966. *Eliza – a Computer Program for Study of Natural Communication Between Man and Machine.* 1966. pp. 36-45. Vol. 9.

Wiks, Y., Catizone, R. 2000. *Human- Computer Conversation.* 2000.

Wojna, A. 2005 . *Analogy-based reasoning in classifier construction.* Faculty of Mathematics, Informatics and mechanics , Warsaw University. 2005 .

Worswick, S. 2013. *http://www.mitsuku.com/.* 2013.

XEROX. *http://open.xerox.com/Services/arabic-morphology.*

Zouaghi, A., L., Merhbene and Zrigu, M. 2012. *word Sense disambiguation for Arabic language using the variants of the Lesk algorithm.* 2012.

Zouaghi, A., Merheben, L and Zrigui, M. 2011. *A Hybrid Approach for Arabic Word Sense Disambiguation.* 2011.

APPENDICES

APPENDIX ONE

ARABIC GOAL-ORIENTED CONVERSATIONAL AGENT BASED ON PATTERN MATCHING AND KNOWLEDGE TREES

Arabic Goal-oriented Conversational Agent Based on Pattern Matching and Knowledge Trees

Zaid Noori, Zuhair Bandar, Keeley Crockett

Abstract- **Conversational Agents (CA's) are computer agents used in applications to converse with humans using natural language dialogues. They are widely used in different fields like industry, education, marketing, health, and other services. Goal Oriented Conversational Agents (GO-CAs) are agents having a deep strategic purpose which enables them to direct conversations to achieve a certain goal using a specific domain. Typically (CA's) are programmed to have a set of rules that guide the conversation with the user. One technique used to script CA's is through pattern matching algorithms. Such algorithms are used to match the user's dialogue and instigate the conversation through writing a series of scripts that contains the rules and patterns relevant to the domain. Throughout the conversation, values can be extracted from the user's dialogue which allows the CA to respond with the correct answer. CA's have been mainly developed for the English language and very limited work has been carried out in Arabic. This is mainly due to the complexity of the language and the lack of resources supporting the Arabic language. This paper proposes a new CA architecture based on a pattern matching algorithm for the development of a goal orientated Arabic Conversational Agents (ACA). The ACA incorporates a new scripting language and knowledge engineering is used to construct the domain. A prototype ACA was developed and the Iraqi passport system was used as a domain to evaluate the new ACA. The ACA was tested and evaluated by experts within the Iraq Consulate with encouraging results and received positive feedback.**

Index Terms- **Conversational Agent, Goal Oriented, Goal Oriented Arabic Conversational Agent, Pattern matching.**

I. INTRODUCTION

The idea of engaging machines to communicate with humans was inspired by the Turning Test in 1950 [1]. Since then a lot of researchers have worked to change this idea into reality. A Conversational Agent (CA) is an agent which uses natural language dialogue to communicate with humans [2]. It has also the ability to reason and pursue a course of action based on its interaction with humans and other agents [3]. The first CA's were known as Chatbots and were designed with the sole aim of holding and maintaining a conversation with users which was often aimless [4].

Zaid Noori
PhD student, School of Computing, Mathematics and Digital Technology, Manchester Metropolitan University
E-mail: zaid.noori@stu.mmu.ac.uk
Zuhair Bandar
Retired Prof, School of Computing, Mathematics and Digital Technology, Manchester Metropolitan University
E-mail: zbandar@hotmail.co.uk
Keely Crockett
Senior Lecturer, School of Computing, Mathematics and Digital Technology, Manchester Metropolitan University
E-mail: K.Crockett@mmu.ac.uk

More recently, Goal Oriented Conversational Agents (GO-CA) were developed to focus the conversation on a particular business [5]. GO-CAs, like other CA's, offer the ability to provide 24/7 consistent support and advice to the user regardless of their computer skills and ability.

They can also provide individual interactions with a different number of users simultaneously. Some good examples of CA's are those used in sales services, education, student debit advisor, and bullying and harassment polices [6, 16].

Traditionally, CA's are scripted using traditional Pattern Matching (PM) algorithms [5]. These algorithms operate on a set of rules organized into contexts that represent the domain; The CA matches each user utterance to patterns within the rules where the highest scoring rule causes a response to the user to fire. Conflict resolution strategies exist within most CA engines to deal with rules that score the same. The main issue with pattern matching is that each domain can take substantial time to script and must be done by domain experts with excellent linguistic skills.

Although the Arabic language is spoken by more than 350 Million all over the world, being the language of the Holy Quran, and also one of the six languages accredited in The United Nations Agencies, it lacks real and active researches in both language resource and CA development. Arabic Conversational Agents (ACA) using Pattern Matching or any other techniques are also rare. little work has been done in developing Arabic CAs [7]. ArabChat [7] was designed at first to act as an Arabic Conversational agent using the same principles as the traditional CA. However, when tested, it was found that it has some weaknesses like irresponsiveness, domain limitation, and inconsistent dialogue flow, in addition to the complexity associated with scripting, maintaining and managing the CA. The new Arabic Goal-orientated Conversational Agent architecture proposed in this paper is designed to overcome these weaknesses. A new CA architecture is introduced to provide a better dialogue flow, usability, adaptability and responsiveness.

In 2003, the Iraq passport system crashed which caused suffering to Iraqi citizens inside and outside Iraq. It was necessary to establish a new system completely. To overcome this problem, temporary solutions were used, by issuing travel documents, and passports with limited validity period until the system is put back into order. This temporary solution caused other problems in itself. The number of valid official travelling documents and passports were confusing for both Iraqis and International Authorities. The burden of these problems was put on Iraqi missions around the globe. Daily phone calls and visits to consular sections by Iraqi immigrants and citizens to inquire about the passport services (issuing, renewal, replacing, etc.) was necessary. The new passport service was taken as the case study to build an Arabic conversational agent for Iraqi's living abroad and will be taken as the experimental domain in this paper.

This paper is organized as follows: section II provides an overview of related work in CA's. Section III describes the architecture of the proposed ACA. Section IV describes the passport service domain and Section V provides details of the knowledge engineering phase. Section VI describes the evaluation of the agent using a pilot study. Finally, section VII concludes by looking at the future use of the agent in a real live environment.

II. Conversational Agents

7.1 Related works

Conversational Agent can be divided into two main types, Embodied Conversational Agent (ECA), and Linguistic Conversational Agent (LCA). ECA's are usually characterized by a multimedia interface which includes facial display, hand gestures, posture, etc. interaction with a human (or representative of a human in a computer environment). ECAs are generally used in applications where risks and impact are not significant if the CA does operates improperly [8,9]. ECAs are complex with a relatively limited number of dialogue tasks. Linguistic Conversational Agents (LCA) are usually categorized into the following: Spoken Dialogue Systems (SDS): In which a speech conversation with the agent is converted to a text through speech recognition algorithms. This type of CA's is insufficiently developed and not commonly used due to the relatively high error rates when converting audio input to text [10]. Chatterbots: In which pattern matching algorithms are used to script conversations with humans, where the aim of this type of CA's is to pass the Turning test (converse with humans successfully for 5 minutes) [11]. There is limited usage of this type of CA's in practical life as they are usually used only to generate conversation with no specific goal.

Goal-Oriented Conversational Agents (GO-CA's) are a type of CA's that have a deep strategic purpose which enable them to direct the conversation to achieve a goal [12]. In this type of CA's, Pattern Matching (PM) is used to search for a string in a piece of text to find all occurrences of these strings inside that text [7]. It is considered as one of the most successful methods for developing CA's that demonstrates or at least gives the impression of some kind of intelligence. To achieve this, knowledge engineering must take place on the domain. From this process knowledge trees are generated and scripted to form the rules used in the CA (patterns and responses). Rules are usually divided into contexts to simplify the management of the CA. During the conversation, rules are scanned to compare their patterns with the user sentences, matched patterns shall be captured and responses shall be fired as a reply to the user. The usage of this type of agent is expanding, especially in marketing and medicine as it offers good services. Short Text Semantic Similarity algorithms (STSS): are also used to develop (GO-CA's) [20]. Essentially, pattern matching algorithms are replaced with sophisticated algorithms for the measurement of Short Text Semantic Similarity [13]. A semantic similarity measure would interpret the semantic content of the sentence as opposed to its structural form. This means fewer patterns are needed in each rule. Throughout the applications of semantics the quantity of scripting can be reduced (patterns) and the user inputs are then matched against the natural language sentences of each rule [20]. The use of such measures is in its infancy and only been trialed on English CA's.

7.2 Arabic Conversational Agents

As mentioned previously, little work has been achieved in the development of Arabic Conversational Agents. Hijjawi et al, [7] developed the first known Arabic agent known as Arabchat. Arabchat used pattern matching algorithms and classified users' utterances as either question or non-question in order to improve matching. The prototype agent was developed for the Applied Science University (ASU) in Jordan to work as an information point advisor for their visitor students who are Arabic native speakers. Some good trials were made to test ArabChat and showed some degrees of success. However, amending the scripts in the domain in any way resulted in complex reformulation of rules within contexts and was very time consuming– similar to English CA's [7]. ArabChat represented the first attempt in ACA development. It was simple in design, with very limited information and knowledge. The contexts were poorly organized which led to slow responsiveness of the agent. However, for a first trial it was successful in terms of robustness and usability [7].

7.3 EVALUATION OF CONVERSATIONAL AGENTS

Evaluation of CA's takes place before releasing them for commercial usage. Both subjective and objective evaluations are usually conducted; however there is no standard methodology adopted by researchers. Evaluation of CA's is mainly done either by distributing a questionnaire to the users trying to reveal their subjective assessment of using the agent or by studying the resulting dialogue [15]. The PARADISE framework [16] was one of the earliest works in creating an evaluation system; it was used to evaluate the DARPA communicator SDS. Chatbot evaluations [21] have also been conducted using a variety of criteria (usability, user satisfaction, Agent credibility, ease of understanding, efficiency, effectiveness, speed, and error rates etc.) that tries to combine subjective and objective measures. Some evaluations tend not to assess all criteria and as there is no benchmark metrics there is no consistency across evaluations. Instead they conclude that evaluations should be adapted to user needs and the application at hand [16].

In this paper, the proposed CA was tested by experts in consular works for both subjective and objective goals. This included its reliability, consistency, speed and its ability to replace the experts or to work as a training tool. Details of the evaluation can be found in section VI.

III. Arabic Goal-Oriented Conversational Agent

The AGO-CA proposed in this paper used the pattern matching approach. Knowledge Engineering (described in section V) was undertaken to structure knowledge in a goal orientated manner. Each node in the knowledge tree was mapped onto a context that contains a series of rules consisting of patterns. Details of the scripting language can be found in sub-section A. The main focus of AGO-CA was to build a modular architecture to provide a robust Conversational Agent with features such as:

- Conversational flow control to ensure the user stays on target to achieve their goal. This is achieved through the creation of knowledge trees (see section V).
- Domain adaptability for ease of maintenance,
- Usability for all audiences regardless of their expertise.

Figure 1 shows the high level architecture for the new AGO-CA.

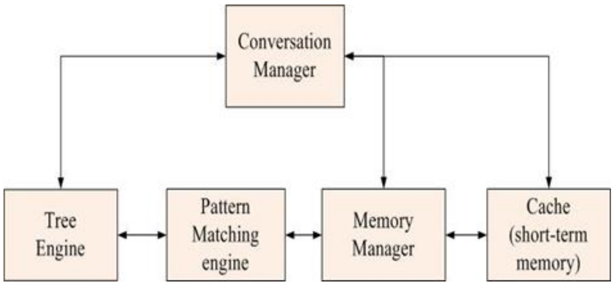


Figure III-1 Agent's Architecture

Each component will now be described:

- The Tree Engine is a module responsible for the flow of dialogue towards the goals of the system. This tree engine contains the scripted knowledge tree and also all the required operations and interfaces to search, modify and maintain the tree. The tree engine uses a scripted knowledge tree defined and maintained by the AGO-CA administrator to inspect and interact with users' utterances; all rules of the domain are organized in a hierarchical tree structure. The tree engine also interacts with the Cache; which keeps all the information related to users and fired rules.

- The Pattern Matching Engine is responsible for managing patterns and patterns operations. Pattern matching engine compares the user's utterance against the predefined rules; it's also used to select the best pattern from a group of matched patterns. Higher priority is given to the most appropriate pattern.
- The Conversation manager performs the coordination between other system modules; it also acts as the main interface between the user and other system modules.
- Memory Manager & Cache are modules related to both long-term and short-term memory; The Memory Manager is used to collect user's information and stores them after achieving the user goal for later use.

When a user initiates a conversation, the agent shall act as follows:

1. The User enters natural language text known as an utterance.
2. The conversation manager requests a reply from the tree engine, if no utterance is being processed the tree engine replies with a query about what the user is requesting help for.
3. When the user answers, his/her utterance is sent to the tree engine for inspection. The tree engine will inspect the utterance by consulting the pattern matching engine to determine which context the user is requesting. Once the tree engine defines the context it inspects the nodes within that context.
4. If there's a direct answer for this query (determined by high scoring patterns) the tree engine fires the associated answer and it will be sent to the conversation manager to be displayed to the user.
5. If that context has more than one option, the tree engine expands the current context and begins a dialogue with a user to gain all required information to be able to provide a appropriate response accordingly.
6. During this dialogue if the user gives an utterance which does not belong to the current context, the tree engine performs a recursive search on all rules defined in the scripted tree to find the appropriate context.
7. If no match is found, the agent shall notify the user and encourage him/her to rephrase the question because the CA did not find the appropriate match.

7.4 Scripting language

Unlike the mechanism used in Arab Chat which evaluates the user utterance against a set of rules and fires the rule based on a numeric activation level value, the proposed AGO-CA introduces a new technique by organizing the rules in a tree structure where each node represents a context, and each context contains rules related to that context only. This structure provides a consistent method to organize the domain topics. The creation of this tree structure can be found in Section V on Knowledge Engineering. This structure enables the AGO-CA to follow the conversation appropriately and helps the AGO-CA to be fully interactive with the user. For example, when a user has an enquiry and the AGO-CA needs additional information to formulate a response, it fires a query about the possible case (the query usually gives two options, Yes or No). Based on the feedback from the user, the tree can follow the proper tree path and fires the accurate answer.

This new tree structure also reduces complexity associated with assigning a numerical strength and activation level values for each rule, as the AGO-CA automatically controls flow of conversation based on the scripted tree. It also provides optimized usability for the AGO-CA administrator by having a consistent interface and appropriate structure to organize the rules and topics within the domain. This is simply done through enabling the administrator to add, delete, or amend nodes and values on the tree without any effect on the main structure.

Conversational Expert System (CES) have been used in the past [14]. Hence, knowledge trees were adopted for the new AGO-CA. This required development of a knowledge tree r tool by which allowed creation of three types of rules, Question Rules, Value Rules, and Report rules**.**

The question rule is a query question which the agent asks the user for input. The value rule contains a potential answer along with patterns in which the user might respond; for example consider the following conversation in figure 2.

1. *Agent: welcome to the Consulate Adviser System, how can I help you?*
2. *User: I need a new passport.*
3. *Agent: Do you have valid Iraqi documents?*
4. *User: Yes, I do.*
5. *Agent: Please fill the online form, print it and submit it with four recent photos of you to the passport department at the consulate.*

Figure III-2 conversation example

The first and third lines represent Question rules, in which the agent is asking the user. The second line is a value rule, in which the user is giving an answer to the agent.

The fourth line also a value node, but it is context-sensitive it means that the answer "Yes, I do" comes in the current conversation flow. If this reply comes at the beginning of the conversation or in another context, it would be considered invalid.

The fifth line is report node which is the final reply fired by the agent when all the queries are completed.

In addition to the facilities mentioned above, the new architecture provides an almost a real-time responsiveness since the AGO-CA is expecting the potential answers from the user, therefore the AGO-CA does not have to scan and search through all the rules in all the contexts to match an answer. The AGO-CA is also flexible in that, it allows switching from one context to another if the user initiates such a dialogue. Conversation example in figure 3 demonstrates this:

1. *User: I lost my passport, what should I do?*
2. *Agent: When a passport is lost, you must inform the police and make a journal advertisement about losing your passport, then you can visit the consulate to apply for a new passport, Have you done all of these procedures?*
3. *User: I want a transit passport.*
4. *Agent: Transit passports are issued for citizens who have lost their passport and wish to go back to Iraq urgently, they are also issued for Iraqi student's children to travel outside Iraq.*

Figure III-3 conversation example

Let's consider the 3rd line, when the agent is asking the user whether he had completed the legal procedures for losing a passport, the user moves into other context asking about transit passport. As demonstrated the agent is flexible in that it switchers to another context that discusses Transit passports and thus provides the user with the correct response.

IV. IRAQ Passport – domain description

A passport is one of the documents that prove the identity of an individual. It becomes the only important document to prove the citizenship when used outside the borders and territory of the native country. Iraqi citizens, especially immigrants, experienced a large number of problems due to frequent changes in Iraqi passports after 2003. The different types of passport forms and the releases of new passports were very confusing. This coincided with the changes in the passport laws. As a result, there were long delays and queues at the Iraqi missions abroad when applying or investigating about passport issues. To make life easier for citizens, and in an attempt to answer their queries and questions in a better and quicker way, an Arabic Goal-Oriented Conversational Agent (AGO-CA) was constructed using the proposed architecture to offer an online service. The CA can access, interpret and discuss the correct and updated information about the Iraqi Passports, and reply in a natural language on frequently asked questions in natural language and queries of the Iraqis seeking advice about passport services.

V. Knowledge Engineering Passport Services

Knowledge Engineering is the extraction of information about the domain from different sources like regulations, legislation, experts in the domain and work procedures. In this paper, information about passports was gathered from the Iraqi Passport law [17], Iraqi Citizenship law [18], Consoler Works Reference Guide [19]. In addition to that, information was also collected on work procedures and advice from experts in this field.

The information gathered was engineered to take the form of an organization diagram with six main contexts about the passports (issuing new passports, renewal, extension, correction, sorting lost passports and travel documents). These contexts were sub divided into about 45 sub contexts. The organization diagram was then converted to take the shape of a knowledge tree having Question Nodes, Value Nodes, and Report Nodes. When conversing with the agent, the matched node shall be expanded and considered as a context, and the user is lead through a dialogue flow to the right response by matching the utterance with node patterns saved in the tree. If the user decides to switch from one context to another (ask questions about a different subject), the agent shall search for the nearest context that matches the subject in the user's utterance. Figures (4, 5) show multi-level knowledge trees in both English and Arabic (for purpose of translation).

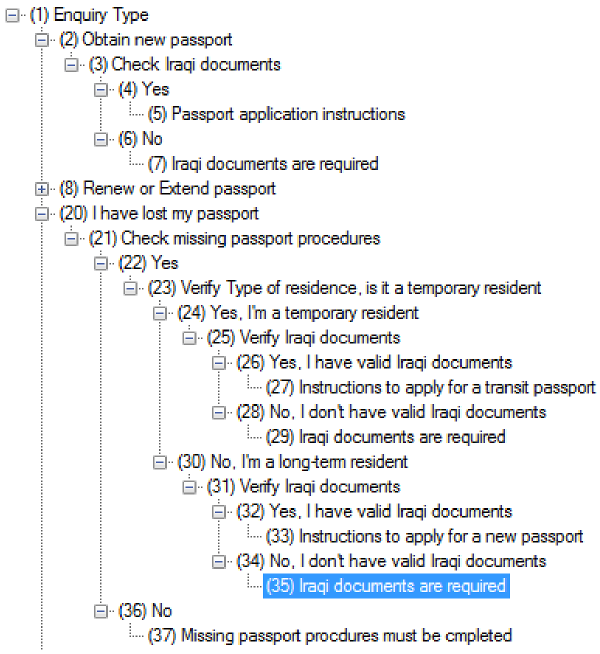


Figure V-1. English knowledge tree

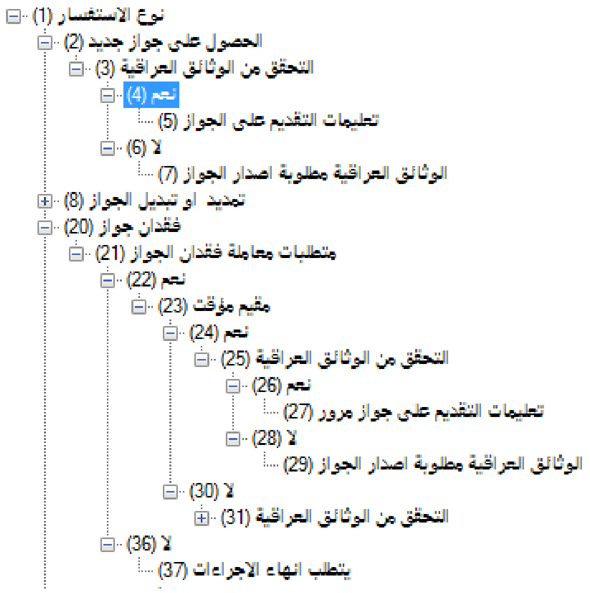


Figure V-2. Arabic knowledge tree

7.5 Conversation Samples

Figures (5 and 6) below show a conversation samples in Arabic and it's translation in English taken from the agent.

Let's consider Figure 5. A user is having a problem since he lost his passport, and asking the agent what he should do about that.

The agent then responds asking the user about the completion of legal procedures regarding the loss of his passport.

It's clear from the conversation sample that the agent is guiding the user through a dialogue to give a correct response and ultimately solving the user's problem – hence reaching the conversational goal.

.

You: I lost my passport, what should I do? System: When a passport is lost, you must inform the police and make a journal advertisement about losing your passport, then you can visit the consulate to apply for a new passport, Have you done all of these procedures? You: Yes, I have System: Do you live there for short-term and you want to go back to Iraq? You: Yes, I'm a tourist. System: do you have valid personal ID and Citizenship certificate, or copies of them? You: Yes System: You can send a written request to the consulate to obtain a transit passport, which requires copies of your Iraqi documents and 4 colored personal pictures with white background.	انت: فقدت جوازي ماذا افعل النظام: عند فقدان الجواز يتطلب تقديم بلاغ للشرطة عن حالة الفقدان ، نشر الفقدان في الصحيفة و بعد ذلك يتم مراجعة البعثة لاصدار جواز جديد ، هل قمت بهذه الاجراءات ؟ انت: نعم قمت بذلك النظام: هل انت مقيم بصورة مؤقته و تريد العودة الى العراق ؟ انت: نعم انا سائح النظام: هل تملك هوية الاحوال المدنية و شهادة الجنسية العراقية النافذه او صور ملونه منهما ؟ انت: نعم النظام: يمكنك تقديم طلب خطي الى البعثة بشأن اصدار جواز مرور و يتطلب وجود نسخ من الوثائق العراقية و صور ملونة بخفية بيضاء عدد 4

Fig. 5

The conversation sample in figure 6 is another example. When the user responded in a different way to the agents question (negative response), the Agent fires a different response instructing the user on the right procedure before a travel document or a passport can be issued to him.

You: I lost my passport, what should I do? System: When a passport is lost, you must inform the police and make a journal advertisement about losing your passport, then you can visit the consulate to apply for a new passport, Have you done all of these procedures? You: No, I have not. System: You must complete these procedures, then you may apply for a new passport or obtain a transit passport to go back to Iraq.	انت: فقدت جواز سفري ماذا افعل النظام: عند فقدان الجواز يتطلب تقديم بلاغ للشرطة عن حالة الفقدان ، نشر الفقدان في الصحيفة و بعد ذلك يتم مراجعة البعثة لاصدار جواز جديد ، هل قمت بهذه الاجراءات ؟ انت: لا لم اقم بذلك النظام: يجب اكمال اجراءات فقدان الجواز قبل الشروع باصدار جواز بديل، بعذ ذلك يمكنك اصدار جواز جديد او اصدار جواز مرور سريع لغرض العودة الى العراق

Fig. 6

VI. Evaluation

7.6 Methodology

The evaluation was conducted through a questionnaire designed especially for this case. It contains some explanation and instructions on the domain, and how to test and evaluate the agent. It also requests some information about the age, gender, status, and experience of the participants themselves. 13 questions were put in the questionnaire, these questions concentrated on subjective issues (agent speed, conversation flow, time to reach the correct answer etc.), and objective ones (like the domain, possibility of using CA's to replace humans in consular activities). The questions were rated between (1-5), where (1) shows poor feedback and (5) shows excellent feedback.

It was not easy to find experts in passport issues to evaluate this work. We managed to finally to select only 10 qualified participants. In addition to the instructions mentioned in the questionnaire, participants were given 6 scenarios designed to test the Agent, those scenarios covered the domain contexts. After reading them, they were engaged in a conversation with the AGO-CA. The conversations were captured in a log file for further analysis and computation of the evaluation metrics.

7.7 Results and Discussion

Table I shows the results of the subjective evaluation. It was clear that the AGO-CA was responding positively with good understanding of the questions with 92.5% accuracy, this mean that misfiring is kept to minimum. The flow of conversation was smooth and the agent managed to reach the goal of the user within a very reasonable time (as indicated the percentage 85%). As for the objective evaluation, it was clear that the possibility of using the agent to replace humans is a little early (only 72.5%); this is mainly due to the culture of people when conversing with passport professionals. The overall evaluation indicated that AGO-CA is impressive. However some further work is needed to make it more acceptable to converse with humans.

Table I: Evaluation Results

Subjective Evaluation	
Item	Rate
Information accessibility	82.5%
Time to reach required information	85%
How well the CA understands user utterance	77.5%
The accuracy of CA answers	92.5%
CA's ability to correct user utterance	72.5%
The validity of answers given by the CA	87.5%
CA responsiveness	95%
CA ability to control dialogue flow during conversation	85%
Overall rate	84.68%
Objective Evaluation	
Item	Rate
How well the CA covers domain topics and issues	77.5%
The possibility of replacing a real passport expert with the CA	72.5%
The possibility of using the CA to provide services to citizens	82.5%
The possibility to use the CA to train consuls	62.5%
Overall rate	73.75%

VII. Conclusions and Further work

The overall ratings achieved of the objective and subjective tests showed that AGO-CA can be used successfully as a real time tool offering services to different users. An expanding market can be expected if such CA's are constructed to serve other fields of life. The knowledge tree architecture proposed simplified and facilitated the work of scripters and enabled them to manage changes and variations in an easier way. In addition to that, these AGO-CA's can be used in training junior diplomats on consular passports activities and becomes a good tool to capture expert knowledge and updated information on the domain.

Although the pattern matching technique is a good tool to run conversational agents, we believe that further work for the Arabic conversational agents is needed using semantic similarity to compare between the two techniques.

References

[1] Turning A.M., 1950 Computing Machinery and Intelligence, Mind, New Series, V59, issue 236, pp433-460 (1950).

[2] Massaro, D.W., Cohen M. M., Besko, J., & Cole, R. A., “Development and Evaluating conversational agents”., In Casell J., Sullivan S., Pervost, &E. Churchill (Eds.) Embodied Conversational Agents, MIT Press, Cambridge, MA, (2000).

[3] Crockett, K., O'Shea, J., Bandar, Z. Goal orientated conversational agents: Applications to benefit society, Agent and Multi-Agent Systems: Technologies and Applications pp. 16-25, Springer Berlin Heidelberg, 2011

[4] Carpenter, R 2007, jabberwacky-live chatbot website viewed 24 Jan 2014 http://www.jabberwacky.com

[5] O'Shea, J. Bandar, Z., Crockett, K. “System Engineering and Conversational Agents” pp 8.2 Chapter 8 pp-204.

[6] O'Shea, J., Bandar, Z., Crockett, K., “System Engineering and Conversational Agents” pp8.3.1, pp8.3.2 p207-2011.

[7] Hijjawi, M.H. “ArabChat: An Arabic Conversational Agent” Ph.D. thesis presented to School of Computing, mathematics and Digital Technology at Manchester Metropolitan University 2011.

[8] Robinson, S., et al.: “What would you ask a CA?”. Observations of Human-Agent Dialogues in a Museum Setting. In: Language Resources and Evaluation Conference 2008, Marrakech, Morocco. Pp1125-1131 (2008).

[9] Babu, S., et al,: “What Would You Like to Talk About?”. An Evaluation of Social Conversation with a Virtual Receptionist. In: Gratch, J., Young M., Aylett, R.S., Ballin, D., Oliver P. (eds.) IVA 2006. LNCS (LANI), vol.4133, pp. 169-180. Springer, Heideberg (2006).

[10] Hunt, M. J.: Figures of Merits for Assessing Connected Word Recognizers, Speech Communication 9, pp239-336 (1990).

[11] Zendek, S., Passing Leobner's Turning Test: A Case of Conflicting Discourse Functions. Minds and Machines 11 pp53-76 (2001). Publishing Company Amsterdam (2007).

[12] O'Shea, J., Bandar Z., Crockett, K.,: “System Engineering and Conversational Agents” In intelligent Based System Engineering ISRL 10, pp 210-230, Springer-Verlag (2011).

[13] Li, Y., et al., Semantic Similarity Based on Semantic Nets and Corpus Statistics. IEEE Transaction on Knowledge and Data Engineering, (2006). 18(8): pp1138-1150.

[14] Annabel Latham, Keely Crockett, Zuhair Bandar; A Conversational Expert System Supporting Bullying and Harassment Polices.

[15] Annika Silvervarg, Arne J¨onsson; “Subjective and Objective Evaluation of Conversational Agents in Learning Environments for Young Teenagers”. Department of Computer and Information Science. Link¨oping University, Link¨oping, Sweden

[16] Shawar, B. A. A., and Atwell, E. S. 2007. Chatbots: are they really useful? LDV-Forum 22:31–50.

[17] The Iraqi Passport Law No. 32 – 1999; Iraqi Gazette Newspaper 2006.

[18] The Iraqi Citizenship Law No. 26 - 2006; Iraqi Gazette Newspaper 2006.

[19] Hamoudi, AbdulRazak; Consolers Manual Guide; Ministry of Foreign Affairs, Republic of Iraq 1st Ed. 2012

[20] O'Shea, K. S.; “A Semantic Based Conversational Agent Framework”. Ph.D. Thesis submitted to MMU-2011

[21] Walker, M. A. et al; DARPA communicators dialog travel planning system. The June 2000 data collection. In: EUROSPEECH 2000. 7th European Conference on Speech Communication and Technology 2nd INTERSPEECH Event, Denmark. Pp1371-1374 (2001)

APPENDIX TWO

KNOWLEDGE ENGINEERING

Introduction

This report outlines the outcomes which have been completed in phase one of development of Pattern Matching Goal-Oriented conversational agent (PMGO-CA); Knowledge engineering covered all procedures of the Iraqi Passport Services domain (IPS).

These procedures were used to produce a series of flowcharts, which were then verified by domain experts. In addition, a set of frequently asked questions and answers were also acquired in IPS domain

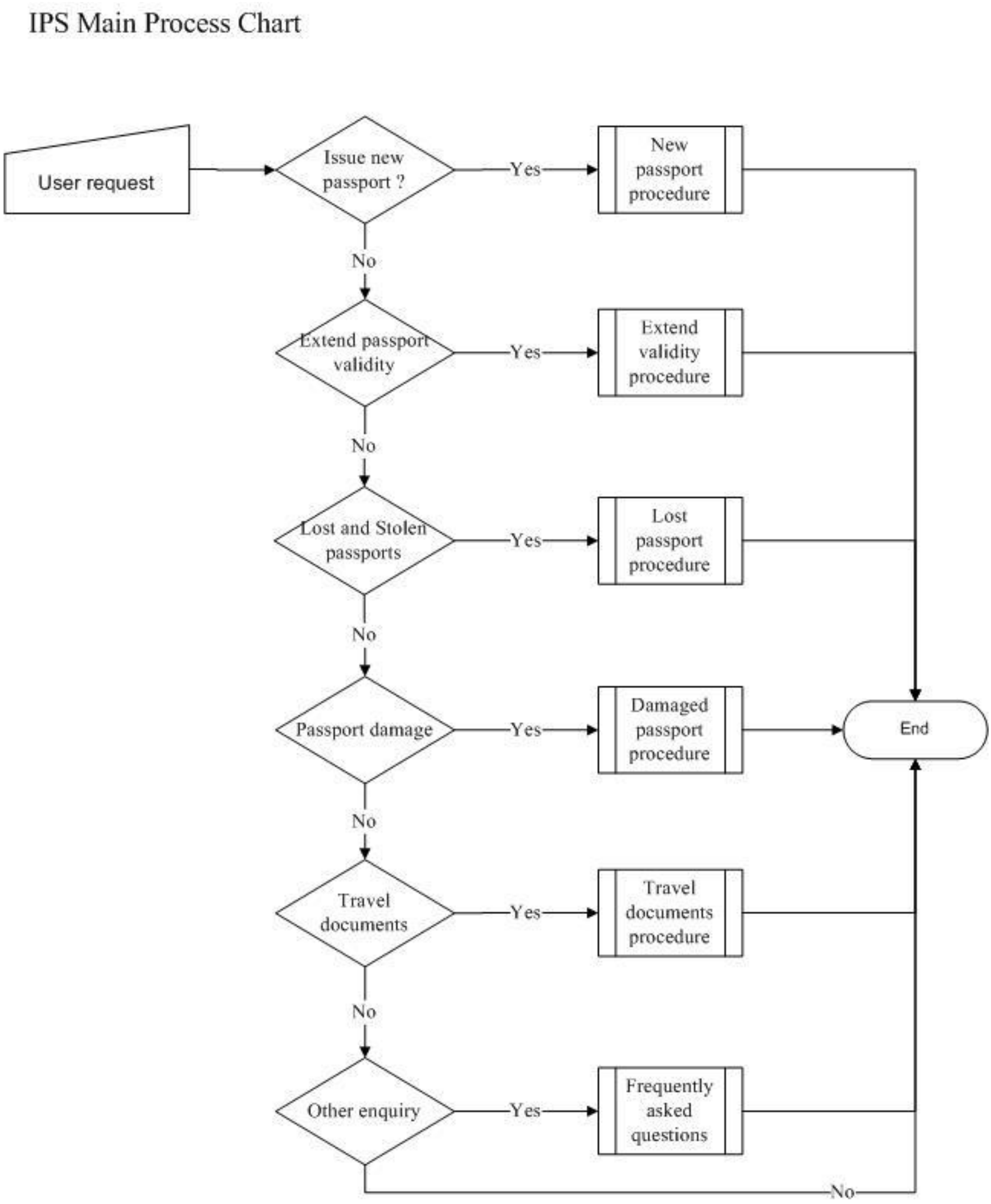
Finally the flowcharts were transformed into a knowledge tree which provides the backbone of PMGO-CA

VIII. CONTENTS

- Flow charts of IPS procedures

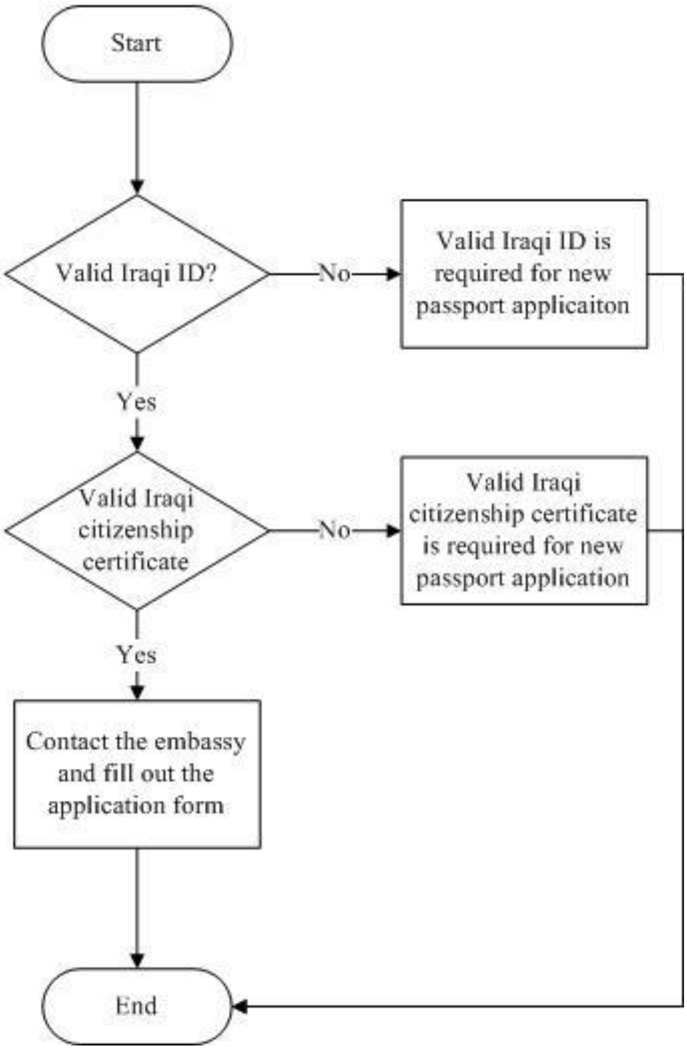
- Frequently asked questions about the IPS domain

7.8 **IPS Main process Chart**

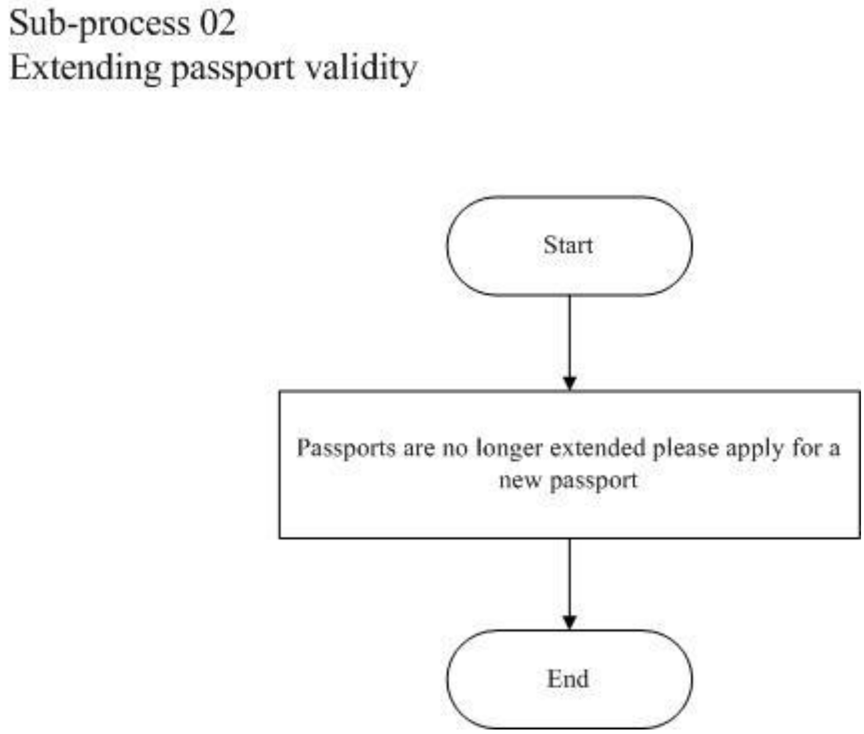


7.9 New passport procedure

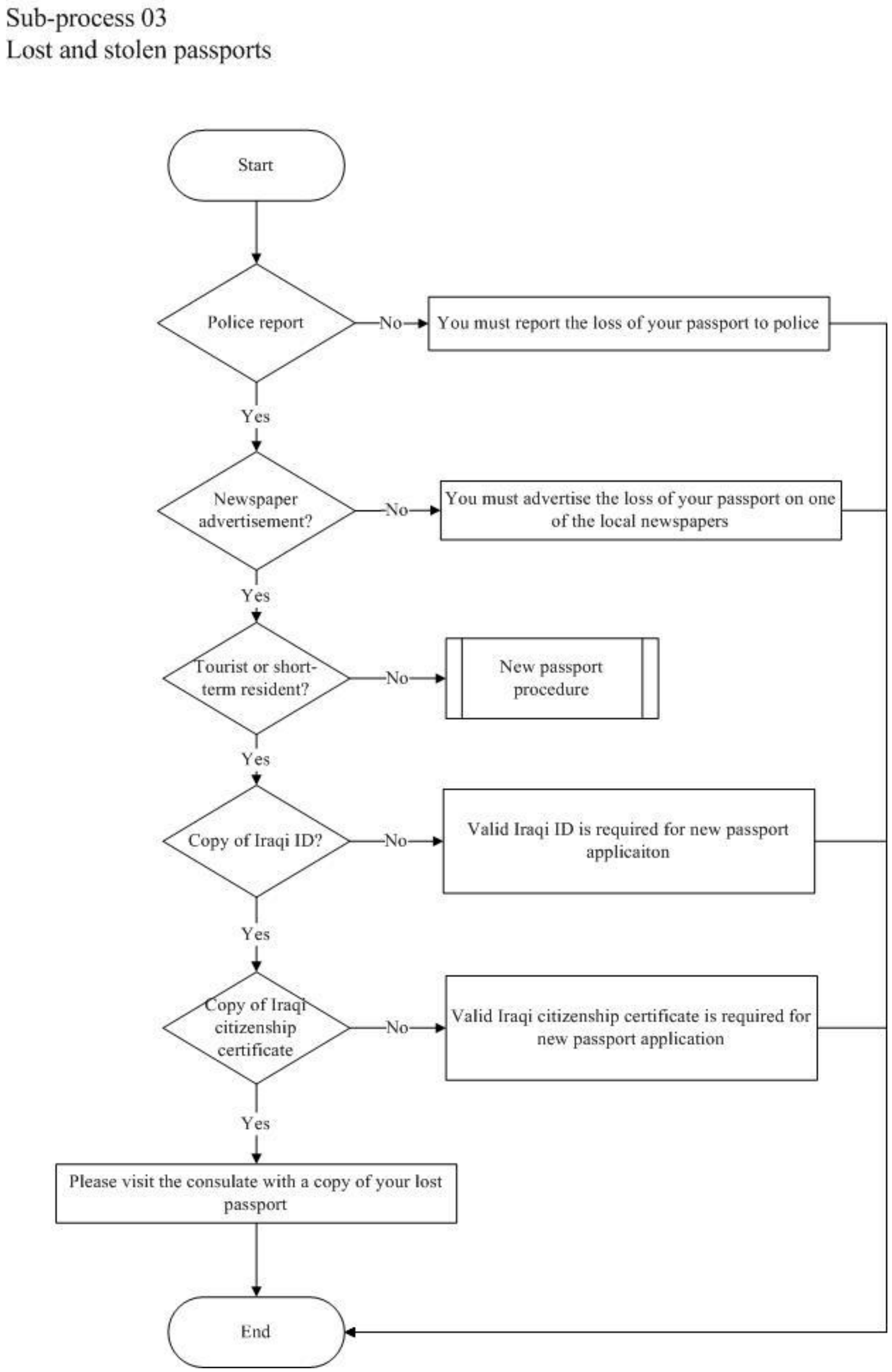
Sub-process 01
Issue new passport



7.10 **Extending passport validity procedure**

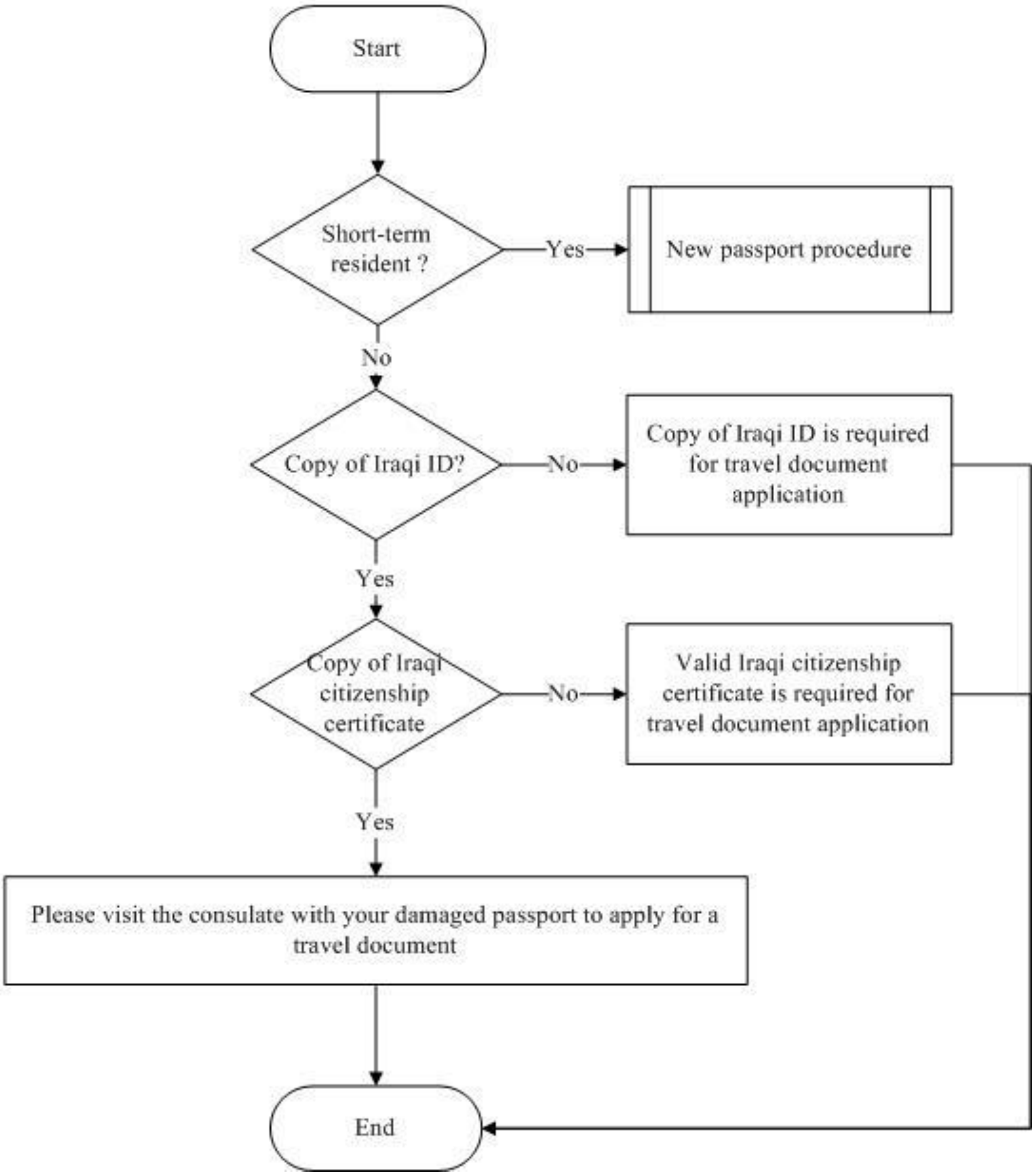


7.11 Lost and stolen passport procedure

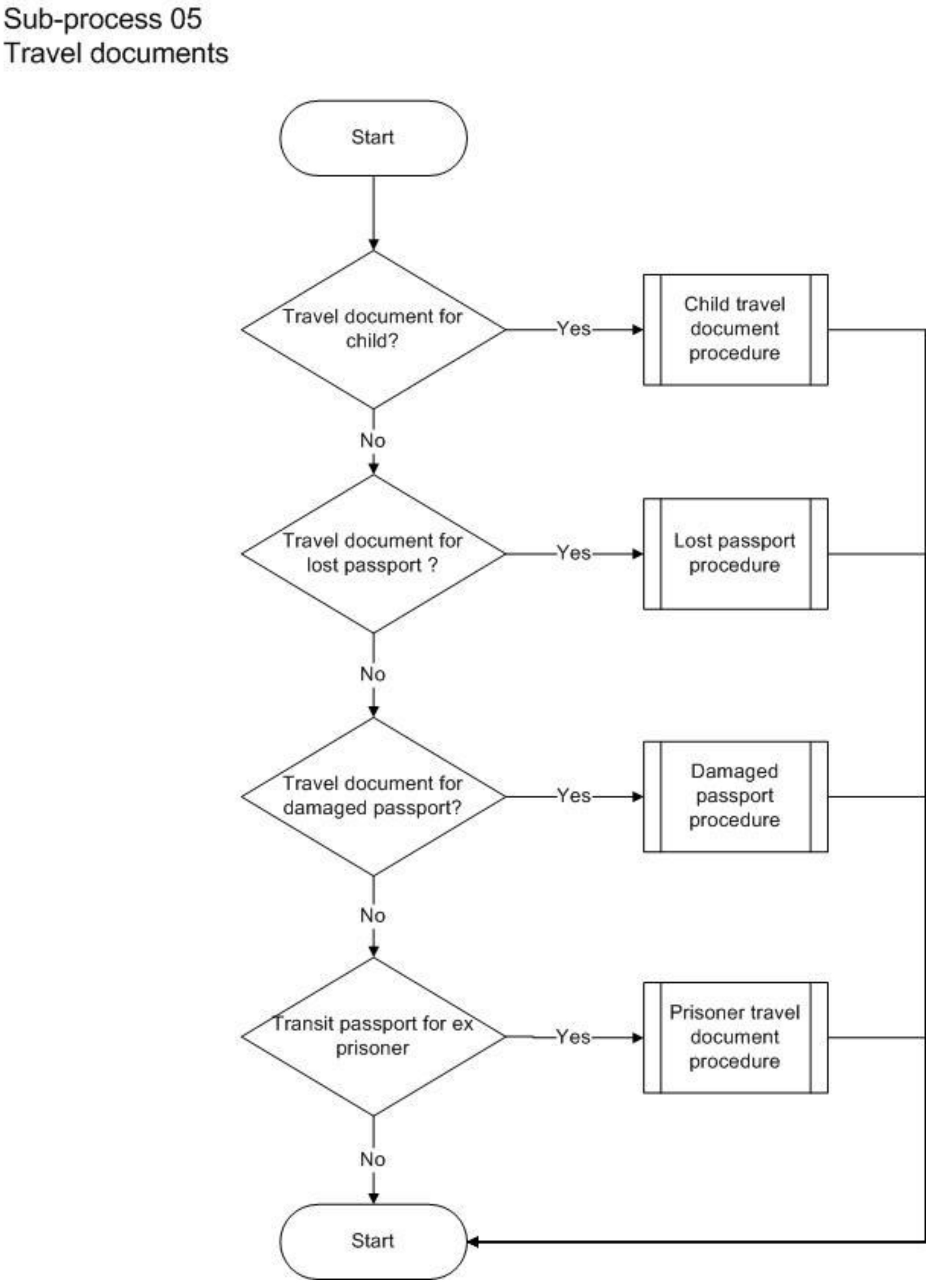


7.12 **Damaged passport procedure**

Sub-process 04
Damaged passports

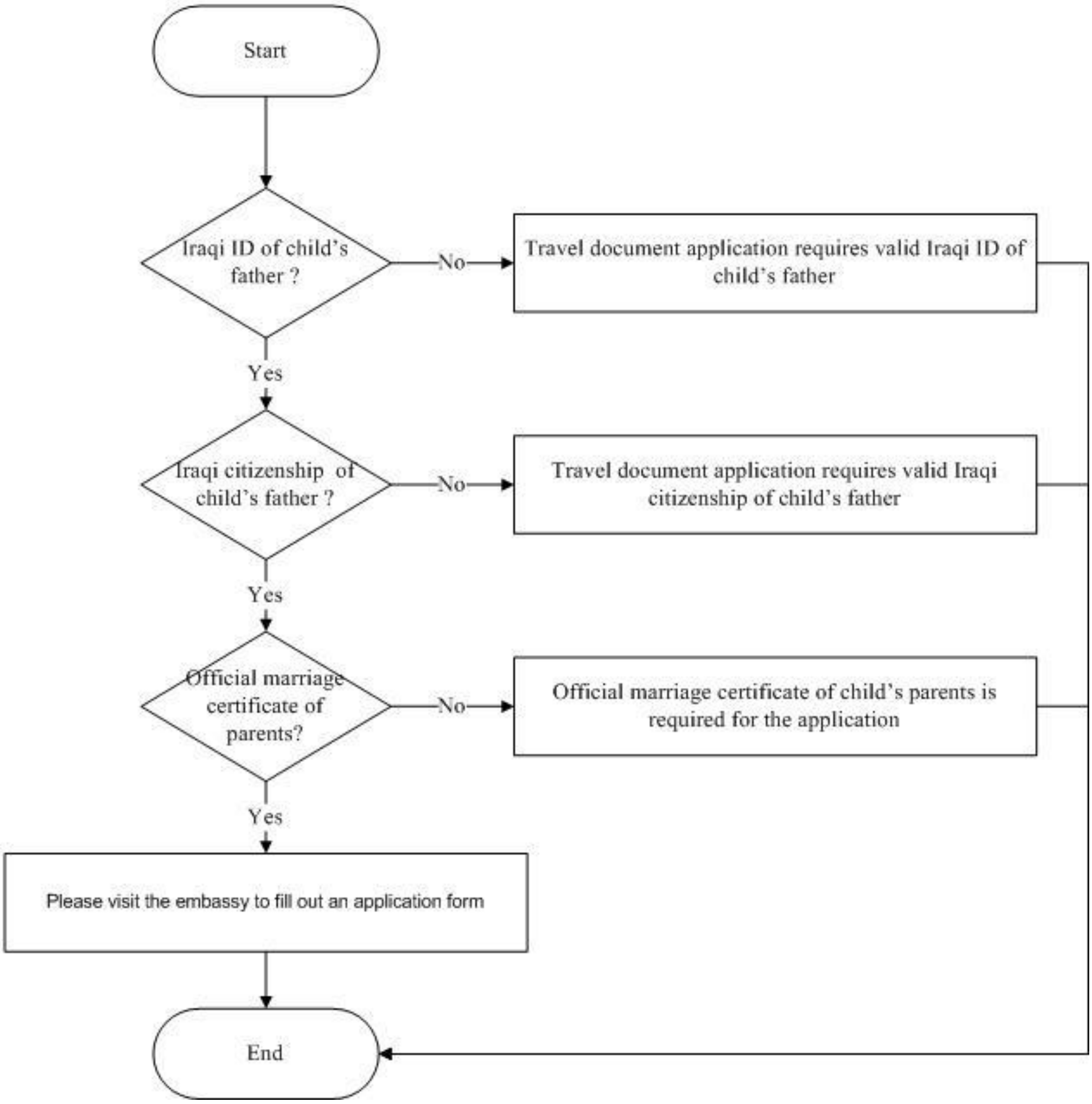


7.13 **Travel document procedure**



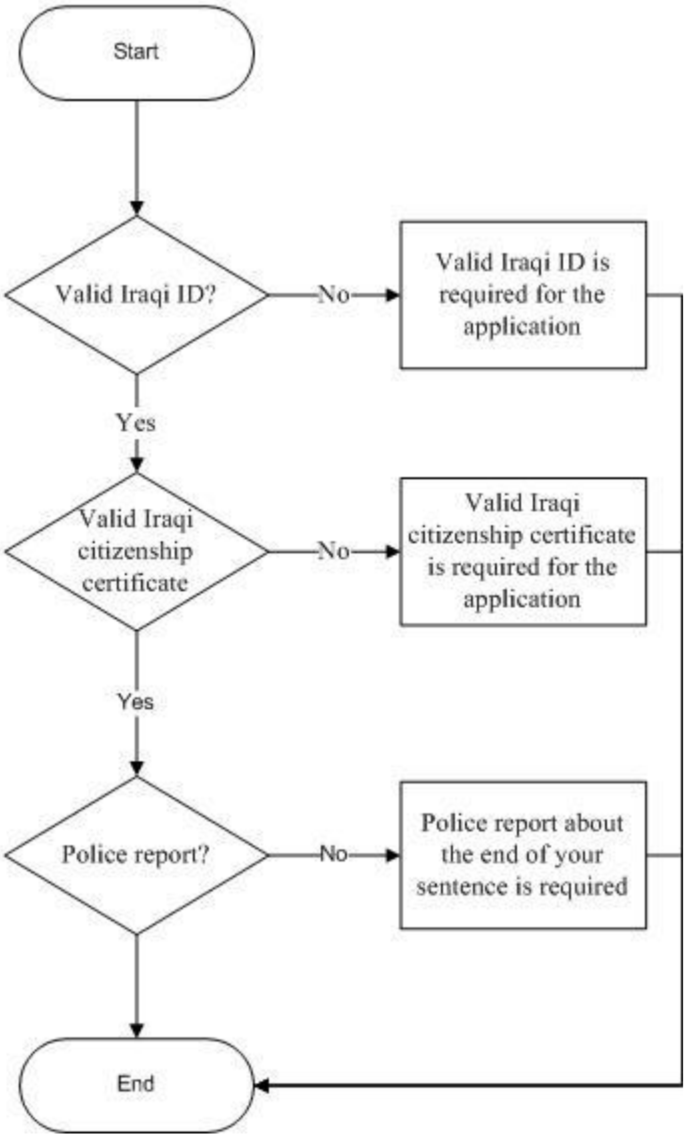
7.14 **Children travel document procedure**

Sub-process 5.1
Children travel documents



7.15 **Ex prisoner travel document**

Sub-process 5.2
Travel Document for ex prisoner



7.16 **Frequently asked questions**

What are the requirements for obtaining a passport?	ماهي متطلبات اصدار الجواز؟	1
Passport application requires an application form with Iraqi ID and citizenship certificate, (4) personal photos with white background and an appointment to capture fingerprint	يتطلب اصدار الجواز بطاقة الاحوال المدنية الجديدة ، شهادة الجنسية العراقية ، ملء استمارة التقديم ، صور ملونة عدد 4 بخلفية بيضاء ، صور عن الوثائق العراقية و يجب اخذ موعد من البعثة لاخذ البصمة .	
How to get a passport application form?	كيف استطيع ملء الاستمارة الخاصة بالجواز (A) ؟	2
You can fill out the application form throughout this website http://www.iraqinationality.gov.iq	بالامكان ملء الاستمارة عن طريق موقع مديرية السفر والجنسية http://www.iraqinationality.gov.iq	
Do I need an appointment to apply for a new passport?	هل احتاج الى موعد للتقديم على الجواز A ؟	3
Yes, you should make an appointment with the consulate	نعم يجب اخذ موعد مسبق من البعثة .	
Is fingerprint mandatory to get a new passport?	هل البصمة شرط اساسي للحصول على جواز ؟	4
Yes, it is; but it's possible to get the finger print of father or son instead, but with the presence of the passport own after failing to obtain their fingerprint	البصمة الحية شرط اساسي ويمكن ان تؤخذ بصمة الابن او الاب بدلا عن ذلك شرط حضور مقدم الطلب شخصيا وتجربة اخذ بصمته الحية	
Can I claim my passport form a consulate different than the one that I have applied to ?	قدمت على جواز سفر في بعثة ما فهل يمكن استلامه من بعثة اخرى ؟	5
Passports must be claimed from the same consulate; but it's possible to call or write to the consulate in case of emergency	يجب استلام الجواز في نفس البعثة ، ويمكن الاتصال بالبعثة او الكتابة اليها في الحالات الطارئة .	
Is it possible to claim the passport by an authorized person?	هل يمكن استلام الجواز عن طريق الوكالة او التخويل؟	6
No, it is not, personal presence is required to acquire the owner's fingerprint	كلا لا يمكن ذلك ويشترط الحضور الشخصي لاخذ بصمة الاتسلام لنفس الشخص .	
How much does it cost to obtain a passport and is it possible to pay them online?	ما هي تكاليف اصدار الجواز وهل يمكن دفعها Online؟	7
New passport fee is currently (25) us dollars, and it's paid directly at the consulate	كاليف اصدار الجواز هي (25) وفي الوقت الحاضر يتم الدفع مباشرة نقدا للبعثة .	
Is it possible to apply for a passport using one document, iraqi ID or Iraqi citizenship certificate?	هل يمكن اصدار جواز سفر استنادا الى وثيقة واحدة فقط بطاقة الاحوال او شهادة الجنسية العراقية؟	8
No, both documents are required	كلا يجب توفر الوثيقتين .	
Is it mandatory to have recent documents to apply for a new passport?	هل يتشرط باصدار الجواز الالكتروني ان تكون الوثائق اصدار جديد؟	9
Yes it is, but it's possible to contact the consulate in some cases	نعم يشترط ذلك وفي الحالات الاخرى يمكن الاتصال بالبعثة او الكتابة اليها لمعرفة القرار .	
Is it possible to apply for a new passport	هل يمكن التقديم على الجواز المقروء اليا بالاستناد الى	10

based on copies of Iraqi documents	نسخ ملونة عن الوثائق؟	
No, original documents should be provided	كلا يشترط تقديم الوثائق الاصلية .	
I have a different name and information in documents abroad, than the names stated in Iraqi documents, may I obtain a passport based on these information?	احمل اسما اخر وبيانات مختلفة في الوثائق الاجنبية يختلف عن الوثائق العراقية هل استطيع الحصول على جواز سفر بالبيانات المذكورة في الوثايق الاجنبية ؟	11
No, Iraqi documents are only considered as a source of information to obtain Iraqi passport	كلا يعتمد باصدار الجواز على بيانات الوثائق العراقية فقط .	
I have changed my name and information in my documents abroad; is it possible to apply for a new passport based on the information stated in my Iraqi documents	غيرت اسمي وبياناتي في الوثائق الاجنبية فهل يمكن الحصول على جواز سفر استنادا الى الوثائق العراقية؟	12
You can send a written request to the embassy to look into it.	يمكن تقديم طلب للبعثة وشرح القضية للبت في الامر .	
My husband/wife is not Iraqi can they obtain an Iraqi passport?	زوجي – زوجتي ليست عراقية هل تستطيع – يستطيع الحصول على جواز سفر عراقي ؟	13
No, they are not allowed to obtain an Iraqi passport until they acquire Iraqi nationality	كلا لا يحق له – لها الحصول على جواز سفر عراقي قبل اكتساب الجنسية العراقية .	
I'm a woman married to a foreigner; can my children apply for an Iraqi passport?	انا متزوجة من اجنبي هل يحق لاطفالي التقديم على جواز السفر العراقي ؟	14
No, they can't, until they acquire Iraqi nationality	كلا لا يمكن ذلك الا بعد اكتساب الجنسية العراقية .	
My wife is foreigner, can my children apply for an Iraqi passport	زوجتي اجنبية هل يحق لابنائي الحصول على جواز السفر العراقي ؟	15
Children of Iraqi citizens from a foreign mother are Iraqi citizens, they can apply for passport after they obtain Iraqi ID	ابناء العراقي من ام اجنبية عراقيون ويمكن الحصول لهم على جواز سفر بعد الحصول على بطاقة الاحوال المدنية.	
Is it possible to extend the validity of (G) and (A) passports	هل يمدد الجواز من الفئة ج او أ ؟	16
No, they cannot be extended	الجوازات من الفئتين ج و أ لا تمدد .	
What is the validity period of (G) and (A) passports?	ما هو فترة استخدام الجواز المقروء اليا من الفئة ج و أ ؟	17
These passports are valid for (8) years	فترة استخدام الجواز من الفئة ج و أ هي ثمان سنوات .	
Is it possible to add occupation to (G) or (A) passports?	هل يمكن اضافة المهنة في الجوازات المقروءة اليا ج و أ ؟	18
There's no occupation field in these passports and it cannot be added in other page	لا يوجد حقل المهنة في الجوازات المقروءة اليا ولا يمكن اضافتها في ورقة اضافية .	
Is it possible to add children to (G) or (A) passports?	هل يمكن اضافة الاولاد في الجوازات المقروءة اليا من الفئتين ج و أ ؟	19
No, they cannot be added, new passports are given to children regardless of their age	كلا لا يمكن اضافة الاطفال على الجوازات بل يصدر جواز للطفل ايا كان عمره .	
Is it possible to add children to (S) passports?	هل يمكن اضافة الاطفال على الجواز من الفئة س ؟	20
Yes it is	نعم يمكن ذلك .	
Is it possible to replace (G) passport due to	هل يمكن تغيير الجواز المقروء الكترونيا بسبب تغغير	21

changes in the owners looks	في شكل صاحبه ؟	
Yes, the consulate must be informed in case of any obvious change to obtain a new passport	نعم في حالة وجود اي تغغير خلقي ظاهر يجب مخاطبة البعثة للتقديم على جواز جديد .	
In which cases a travel document is issued?	ماهي الحالات التي يصدر فيها جواز المرور ؟	22
Travel documents are given to whom they lost their passports or in cases in which a citizen is willing to return to iraq voluntarily, it is also issued to children of Iraqi students abroad.	يصدر جواز المرور لمن فقد جواز سفره ، او في الحالات التي توجب المواطن على العودة للعراق بشكل طوعي ، ويصدر لاطفال الطلبة العراقيين للسفر الى العراق لاصدار جواز للطفل .	23
What are the requirements to obtain a travel document?	ماهي متطلبات اصدار جواز المرور ؟	24
The original Iraqi documents (Iraqi ID and Iraqi citizenship certificate and copies of them), (4) personal photographs in white background, a written request to the consulate, an application form and a travel ticket	تقديم الوثائق العراقية الاصلية بطاقة الاحوال المدنية وشهادة الجنسية العراقية ونسخ ملونة عنها ، صور ملونة بخلفية بيضاء عدد 4 ، طلب خطي للبعثة ، ملء استمارة التقديم في البعثة ،تذكرة سفر .	
What is the validity period of travel documents?	ما هي المدة المحددة لجواز المرور ؟	25
Travel documents are valid for (6) months starting form the issue date	مدة نفاد الجواز هي ستة اشهر تبدأ من يوم الاصدار .	
Is it possible to obtain transit passports for children	هل يمكن اصدار جواز مرور للاطفال ؟	26
It is possible for the children of Iraqi students born abroad during study period, or for families wishing to return voluntarily to Iraq due to refusal of residence permit, and other emergency situations that are looked into by the chief of the consulate staff	يمكن ذلك لابناء الطلبة العراقيين المبتعثين الذين رزقوا بابناء في الخارج اثناء الدراسة ، او للعوائل التي ترغب بالعودة الطوعية للعراق بسبب رفض طلبات اقامتهم ، الحالات الطارئة الاخرى التي يقرر بشأنها رئيس البعثة .	
Is there a temporary passport other than the travel document?	هل هناك جواز سفر مؤقت غير جواز المرور ؟	27
No, there isn‘t	كلا لا يوجد .	
What are the fees for the transit documents	ما هي رسوم جواز المرور ؟	28
Fees are (25) us dollar	الرسوم هي (25) دولار	

APPENDIX THREE

EVALUATION QUESTIONNAIRE

Dialogue system for Iraqi passport domain

Questionnaire

1) Introduction

This questionnaire aims to evaluate the dialogue systems used to converse with citizens, which is my subject of study to find an appropriate method to communicate with Iraqi citizens living abroad automatically without the need for specialized employees to solve some of the consul issues concerning the Iraqi passport and offer fast responses to the citizens.

2) Instructions

- Conversation with the dialogue system can be initiated online by visiting the CA web site www.iraq-pass-ca.net/ for PM or (www.sem.iraq-pass-ca.net/ for semantic CA)
- After greetings, the conversation can be started; it is very similar to conversations carried out in instant messenger (i.e. yahoo messenger)
- The dialogue system shall converse with you regarding passport issues and topics, and will not answer any other questions regarding anything else; these topics are:
  - Passport issue
  - Extending passport validity
  - Lost an d stolen passports
  - Passport damage
  - Travel documents
- The language of conversation is the modern Arabic language, the system can also understand some of colloquial Arabic words which can be used when needed

- The system starts with asking for some personal information right before the conversation starts
- Please use the dialogue as if you were an Iraqi citizen living abroad
- Please click the (close conversation) button once you’re done with the system, in order to record your conversation by the system so I can review them and consider any comments that you may have; you can also download a text file with the conversation to your computer by clicking (download conversation)
- Please consider initiating several conversations with the system to be familiar with it before evaluating and making any observations
- After conversing with the system; please fill out this questionnaire with your information, which are used to evaluate the process without disclosure of these information, your name might not be familiar to the system which stores hundreds of names, in such case please use any common Iraqi name
- Please submit the questionnaire to this email zaidnori@yahoo.co.uk once your done with it

3) Basic information

Gender	○ Male ○ Female
Years of professional experience	○ Less than 5 ○ 5 to 10 years ○ to 15 years ○ 15 to 20 years ○ More than 20 years
Years of experience in the passport domain	
Current position	
Professional degree	
Years of experience in consular work in general	

4) Evaluation

Please rate each of the aspects of the dialogue system list in table (1), the rate should be an integer ranging between (1) and (5), as follows:

(6) Weak
(7) Acceptable
(8) Good
(9) Very good
(10) Excellent

You can also write any comment in the designated field of each row in table (1)

	Question	**Rating (5-1)**	**Comments**
1	You have found the dialogue system very responsive, in terms of speed		
2	You could reach your desired information in short and direct conversation		
3	You find the dialogue system helpful to access information regarding the Iraqi passport domain, and you would prefer to use it instead of other methods such as browsing a website or calling the consulate staff		
4	The system was able to correct mistyped words		
5	The system was able to understand your questions, (regardless of incorrect responses)		
6	You found the answers of the system accurate		

7	You found the conversations with the system to be consistent and organized		
8	The system was able to memorize personal information and previous topics in the conversation		
9	The responses of the system were identical to Iraqi passport laws and regulations		
10	The system covered all domain topics and regulations		
11	You agree that the system can be used instead of an expert		
12	You agree that this system can be used as a good tool for training of junior diplomats.		

Table 1 Questionnaire form

APPENDIX FOUR

STATISTICAL ANALYSIS OF USER QUESTIONNAIRE USING THE WILCOXON RANKED TEST

1. Introduction

This appendix contains a statistical analysis for the results of PMGO-CA evaluation questionnaire (which is covered in chapter (5)) the statistical analysis is performed using the Wilcoxon Ranked Test with a significance level of 5%. The assumption made for the Wilcoxon test is that the variable being tested is symmetrically distributed about the median, and that the responses are symmetrically distributed about (Good), a hypothesis that Participants assess each metric as agreeable can be tested. Participants that assess a metric as agreeable will give a rating more than 3. The null and alternative hypotheses are, stated as follows:

H0: the median response is 3

H1: the median response is more than 3

The null hypothesis indicate that there's no difference between the median (3) and the rating of questionnaire participants, while the alternative hypothesis assumes that the participant rating is above the median, so if the null hypothesis is rejected, it means that users think of each aspect they evaluated to be above (good).

2. The Wilcoxon signed-rank test

The Wilcoxon signed-rank test is a non-parametric statistical hypothesis test used when comparing two related samples, matched samples, or repeated measurements on a single sample to assess whether their population mean ranks differ.

3. Procedure

Let N be the sample size, the number of pairs. Thus, there are a total of *2N* data points. For $i = 1, \ldots, N$, let $x_{1,i}$ and $x_{2,i}$ denote the measurements.

H_0: median difference between the pairs is zero

H_1: median difference is not zero.

- For $i = 1, ..., N$, calculate $|x_{2,i} - x_{1,i}|$ and $\text{sgn}(x_{2,i} - x_{1,i})$, where sgn is the sign function.
- Exclude pairs with $|x_{2,i} - x_{1,i}| = 0$. Let N_r be the reduced sample size.
- Order the remaining N_r pairs from smallest absolute difference to largest absolute difference, $|x_{2,i} - x_{1,i}|$.
- Rank the pairs, starting with the smallest as 1. Ties receive a rank equal to the average of the ranks they span. Let R_i denote the rank.
- Calculate the test statistic W

$$W = |\sum_{i=1}^{N_r}[\text{sgn}(x_{2,i} - x_{1,i}) \cdot R_i]|$$

Decision rules:

If the alternative is H_1: $\mu \neq \mu_0$, reject H_0: $\mu = \mu_0$ if $w \leq w_\alpha^*$

If the alternative is H_1: $\mu > \mu_0$, reject H_0: $\mu = \mu_0$ if $w^- \leq w_\alpha^*$

If the alternative is H_1: $\mu < \mu_0$, reject H_0: $\mu = \mu_0$ if $w^+ \leq w_\alpha^*$

n	Two-Tailed Test		One-Tailed Test	
	α = .05	α = .01	α = .05	α = .01
5	--	--	0	--
6	0	--	2	--
7	2	--	3	0
8	3	0	5	1
9	5	1	8	3
10	8	3	10	5
11	10	5	13	7
12	13	7	17	9
13	17	9	21	12
14	21	12	25	15
15	25	15	30	19
16	29	19	35	23
17	34	23	41	27
18	40	27	47	32
19	46	32	53	37
20	52	37	60	43
21	58	42	67	49
22	65	48	75	55
23	73	54	83	62
24	81	61	91	69
25	89	68	100	76
26	98	75	110	84
27	107	83	119	92
28	116	91	130	101
29	126	100	140	110
30	137	109	151	120

Table 3-1 Critical values

4. Test results

Table (2) shows the results of Participant ratings for Metric 1 (Responsiveness).

Responsiveness						
Observation (i)	Participant rating	Rating-Median rating(3)	Absolute value	Ranking	+	-
4	4	1	1	1.5	1.5	0
7	4	1	1	1.5	1.5	0
1	5	2	2	6.5	6.5	0
2	5	2	2	6.5	6.5	0
3	5	2	2	6.5	6.5	0
5	5	2	2	6.5	6.5	0
6	5	2	2	6.5	6.5	0
8	5	2	2	6.5	6.5	0
9	5	2	2	6.5	6.5	0
10	5	2	2	6.5	6.5	0
Total					55	0

Table 4-1 Ratings for Metric 1 (responsiveness)

Positive difference W+ = 55

Negative difference W-= 0

Taking the value of W- of 0 for the calculated value of T, that being T = 0, at the 5% significance level, that being p = 0.05, for a sample size of 10, that being n = 10, the Wilcoxon T statistic table value of T is 10 for a one-tailed test. As the calculated value of T is smaller than the critical value in table (1), the null hypothesis may be rejected.

In short, for the metric of “Responsiveness” Participants agree that PMGO-CA is responsive

Table (3) shows the results of Participant ratings for Metric 2 (Conversation length).

Conversation length						
Observation (i)	Participant rating	Rating–Median rating(3)	Absolute value	Ranking	+	-
1	4	1	1	3.5	3.5	0
2	4	1	1	3.5	3.5	0
4	2	-1	1	3.5	0	3.5
6	4	1	1	3.5	3.5	0
7	4	1	1	3.5	3.5	0
8	4	1	1	3.5	3.5	0
3	5	2	2	8.5	8.5	0
5	5	2	2	8.5	8.5	0
9	5	2	2	8.5	8.5	0
10	5	2	2	8.5	8.5	0
Total					51.5	3.5

Table 4-2 Ratings for Metric 2 (conversation length)

Positive difference W+ = 51.5

Negative difference W-= 3.5

Taking the value of W- of 3.5 for the calculated value of T, that being T = 3.5, at the 5% significance level, that being p = 0.05, for a sample size of 10, that being n = 10, the Wilcoxon T statistic table value of T is 10 for a one-tailed test. As the calculated value of T is smaller than the critical value in table (1), the null hypothesis may be rejected.

In short, for the metric of “conversation length” Participants agree that the length of conversation was acceptable

Table (4) shows the results of Participant ratings for Metric 3 (Information accessibility).

Information accessibility						
Observation (i)	Participant rating	Rating – Median rating(3)	Absolute value	Ranking	+	-
4	3	0	0	0	0	0
8	3	0	0	0	0	0
1	4	1	1	2.5	2.5	0
3	4	1	1	2.5	2.5	0
7	4	1	1	2.5	2.5	0
10	4	1	1	2.5	2.5	0
2	5	2	2	6.5	6.5	0
5	5	2	2	6.5	6.5	0
6	5	2	2	6.5	6.5	0
9	5	2	2	6.5	6.5	0
Total					36	0

Table 4-3 Ratings for Metric 3 (Information accessibility)

Positive difference W+ = 36

Negative difference W-= 0

Taking the value of W- of 0 for the calculated value of T, that being T = 0, at the 5% significance level, that being p = 0.05, for a sample size of 8, that being n = 8, the Wilcoxon T statistic table value of T is 5 for a one-tailed test. As the calculated value of T is smaller than the critical value in table (1), the null hypothesis may be rejected.

In short, for the metric of “Information accessibility”, participants agreed to prefer PMGO-CA over other methods to acquire information about the Iraqi Passport Domain

Table (5) shows the results of Participant ratings for Metric 4 (Correcting users' utterances).

Correcting users' utterances						
Observation (i)	Participant rating	Rating–Median rating(3)	Absolute value	Ranking	+	-
9	3	0	0	0	0	0
7	3	0	0	0	0	0
1	3	0	0	0	0	0
5	3	0	0	0	0	0
2	4	1	1	3.5	3.5	0
3	4	1	1	3.5	3.5	0
6	4	1	1	3.5	3.5	0
8	4	1	1	3.5	3.5	0
10	4	1	1	3.5	3.5	0
4	4	1	1	3.5	3.5	0
Total					21	0

Table 4-4 Ratings for Metric 4 (Correcting Participants' utterances)

Positive difference W+ = 21

Negative difference W-= 0

Taking the value of W- of 0 for the calculated value of T, that being T = 0, at the 5% significance level, that being p = 0.05, for a sample size of 6, that being n = 6, the Wilcoxon T statistic table value of T is 2 for a one-tailed test. As the calculated value of T is smaller than the critical value in table (1), the null hypothesis may be rejected.

In short, for the metric of "Correcting users' utterances", participants agreed that PMGO-CA was able to handle misspelled utterances

Table (6) shows the results of Participant ratings for Metric 5 (CA understanding of Participants' utterances).

CA understanding of Participants' utterances						
Observation (i)	Participant rating	Rating– Median rating(3)	Absolute value	Ranking	+	-
3	3	0	0	0	0	0
4	4	1	1	4	4	0
5	4	1	1	4	4	0
6	2	-1	1	4	0	4
7	4	1	1	4	4	0
8	4	1	1	4	4	0
9	4	1	1	4	4	0
10	4	1	1	4	4	0
1	5	2	2	8.5	8.5	0
2	5	2	2	8.5	8.5	0
Total					41	4

Table 4-5 Ratings for Metric 5 (CA understanding of Participants' utterances)

Positive difference W+ = 41

Negative difference W-= 4

Taking the value of W- of 0 for the calculated value of T, that being T = 6, at the 5% significance level, that being p = 0.05, for a sample size of 9, that being n = 9, the Wilcoxon T statistic table value of T is 8 for a one-tailed test. As the calculated value of T is smaller than the critical value in table (1), the null hypothesis may be rejected.

In short, for the metric of "Information accessibility", participants agreed to prefer PMGO-CA over other methods to acquire information about the Iraqi Passport Domain

Table (7) shows the results of Participant ratings for Metric 6 (Accuracy).

Accuracy						
Observation (i)	Participant rating	Rating–Median rating(3)	Absolute value	Ranking	+	-
1	4	1	1	2.5	2.5	0
2	4	1	1	2.5	2.5	0
4	4	1	1	2.5	2.5	0
10	4	1	1	2.5	2.5	0
3	5	2	2	7.5	7.5	0
5	5	2	2	7.5	7.5	0
6	5	2	2	7.5	7.5	0
7	5	2	2	7.5	7.5	0
8	5	2	2	7.5	7.5	0
9	5	2	2	7.5	7.5	0
Total					55	0

Table 4-6 Ratings for Metric 7 (Accuracy)

Positive difference W+ = 55

Negative difference W-= 0

Taking the value of W- of 0 for the calculated value of T, that being T = 0, at the 5% significance level, that being p = 0.05, for a sample size of 10, that being n = 10, the Wilcoxon T statistic table value of T is 10 for a one-tailed test. As the calculated value of T is smaller than the critical value in table (1), the null hypothesis may be rejected.

In short, for the metric of “Accuracy”, participants agreed that PMGO-CA answers are accurate

Table (8) shows the results of Participant ratings for Metric 7 (Conversation consistency).

Conversation consistency						
Observation (i)	Participant rating	Rating– Median rating(3)	Absolute value	Ranking	+	-
7	3	0	0	0	0	0
3	4	1	1	3	3	0
5	4	1	1	3	3	0
8	4	1	1	3	3	0
9	4	1	1	3	3	0
10	4	1	1	3	3	0
1	5	2	2	7.5	7.5	0
2	5	2	2	7.5	7.5	0
4	5	2	2	7.5	7.5	0
6	5	2	2	7.5	7.5	0
Total					45	0

Table 4-7 Ratings for Metric 8 (conversation consistency)

Positive difference W+ = 45

Negative difference W-= 0

Taking the value of W- of 0 for the calculated value of T, that being T = 0, at the 5% significance level, that being p = 0.05, for a sample size of 9, that being n = 9, the Wilcoxon T statistic table value of T is 8 for a one-tailed test. As the calculated value of T is smaller than the critical value in table (1), the null hypothesis may be rejected.

In short, for the metric of “Conversation consistency”, participants agreed that their conversations with PMGO-CA were organized and consistent

Table (9) shows the results of Participant ratings for Metric 8 (Memory).

Memory						
Observation (i)	Participant rating	Rating–Median rating(3)	Absolute value	Ranking	+	-
3	3	0	0	0	0	0
10	3	0	0	0	0	0
2	4	1	1	4	4	0
4	4	1	1	4	4	0
5	4	1	1	4	4	0
6	4	1	1	4	4	0
7	2	-1	1	4	0	4
8	4	1	1	4	4	0
9	4	1	1	4	4	0
1	5	2	2	8	8	0
Total					32	4

Table 4-8 Ratings for Metric 8 (Memory)

Positive difference W+ = 32

Negative difference W-= 4

Taking the value of W- of 4, for the calculated value of T, that being T = 4, at the 5% significance level, that being p = 0.05, for a sample size of 8, that being n = 8, the Wilcoxon T statistic table value of T is 5 for a one-tailed test. As the calculated value of T is equal to the critical value, the null hypothesis may be rejected

In short, for the metric of “Memory”, participants agreed that PMGO-CA memory was good.

Table (10) shows the results of Participant ratings for Metric 9 (Validity).

Validity						
Observation (i)	Participant rating	Rating–Median rating(3)	Absolute value	Ranking	+	-
8	3	0	0	0	0	0
3	4	1	1	2	2	0
4	2	-1	1	2	0	2
9	4	1	1	2	2	0
1	5	2	2	6.5	6.5	0
2	5	2	2	6.5	6.5	0
5	5	2	2	6.5	6.5	0
6	5	2	2	6.5	6.5	0
7	5	2	2	6.5	6.5	0
10	5	2	2	6.5	6.5	0
Total					43	2

Table 4-9 Ratings for Metric 9 (Validity)

Positive difference W+ = 43

Negative difference W-= 2

Taking the value of W- of 2 for the calculated value of T, that being T = 2, at the 5% significance level, that being p = 0.05, for a sample size of 9, that being n = 9, the Wilcoxon T statistic table value of T is 8 for a one-tailed test. As the calculated value of T is equal to the critical value in table (1), the null hypothesis may be rejected.

In short, for the metric of “Validity”, participants agreed that PMGO-CA responses were accurate

Table (11) shows the results of Participant ratings for Metric 10 (Domain coverage).

Domain coverage						
Observation (i)	Participant rating	Rating–Median rating(3)	Absolute value	Ranking	+	-
3	3	0	0	0	0	0
6	3	0	0	0	0	0
9	3	0	0	0	0	0
10	3	0	0	0	0	0
2	4	1	1	2	2	0
4	4	1	1	2	2	0
8	4	1	1	2	2	0
1	5	2	2	5	5	0
5	5	2	2	5	5	0
7	5	2	2	5	5	0
Total					21	0

Table 4-10 Ratings for Metric 10 (Domain coverage)

Positive difference W+ = 21

Negative difference W-= 0

Taking the value of W- of 0 for the calculated value of T, that being T = 0, at the 5% significance level, that being p = 0.05, for a sample size of 6, that being n = 6, the Wilcoxon T statistic table value of T is 2 for a one-tailed test. As the calculated value of T is smaller than the table value, the null hypothesis may be rejected.

In short, for the metric of “Domain coverage”, participants agreed that PMGO-CA well covered domain topics, laws and regulations

APPENDIX FIVE

LIST OF ARABIC FUNCTION WORDS

Num.	Word	Num.	Word	Num.	Word	Num.	Word
1	الان	86	اى	171	على	256	لمتى
2	التي	87	اي	172	عليك	257	لمن
3	الذي	88	ايا	173	عن	258	لن
4	الذين	89	ايان	174	عند	259	لنا
5	اللاتي	90	اية	175	عوض	260	له
6	اللتان	91	ايضا	176	عيانا	261	لها
7	اللتيا	92	اين	177	عين	262	لهم
8	اللتين	93	ايها	178	غادر	263	لهما
9	اللذان	94	ايهم	179	غالبا	264	لهن
10	اللذين	95	ايهم	180	غدا	265	لهو
11	اللواتي	96	ايهما	181	غداة	266	لهي
12	اثر	97	ايهن	182	غير	267	لو
13	احدى	98	اها	183	فاالى	268	لولا
14	اذ	99	اي	184	فاالى	269	لوما
15	اذا	100	بات	185	فانى	270	لي
16	اذما	101	باي	186	فاي	271	ليت
17	اذن	102	بس	187	فايان	272	ليس
18	ازاء	103	بضع	188	فاين	273	لئلا
19	الا	104	بعد	189	فبكم	274	ما
20	الام	105	بعدا	190	فضلا	275	مادام
21	الى	106	بعض	191	فعلى	276	ماذا
22	اليك	107	بك	192	فعن	277	مازال
23	اليكم	108	بكم	193	ففي	278	مافتئ
24	اليكما	109	بكما	194	فقط	279	متى
25	اليكن	110	بكن	195	فكم	280	مثل
26	اما	111	بل	196	فكما	281	مذ
27	ان	112	بله	197	فكيف	282	مرة
28	انا	113	بلى	198	فلاين	283	مرحبا
29	اننا	114	بم	199	فلم	284	مساء
30	اى	115	بما	200	فلماذا	285	مع
31	ايا	116	بماذا	201	فلمتى	286	مقابل
32	اياك	117	بمن	202	فم	287	مكانك
33	اياكم	118	بنا	203	فما	288	مكانكم
34	اياكما	119	به	204	فماذا	289	فبكم
35	اياكن	120	بها	205	فمتى	290	فضلا
36	ايانا	121	بهم	206	فمن	291	فعلى
37	اياه	122	بهما	207	فهل	292	فعن
38	اياها	123	بهن	208	فهو	293	ففي
39	اياهم	124	بوسا	209	فهي	294	فقط
40	اياهما	125	بي	210	فو	295	فكم
41	اياهن	126	بيد	211	فور	296	فكما
42	اياي	127	بين	212	فوق	297	فكيف

43	ايه	128	تجاه	213	في	298	فلاين
44	االى	129	تحت	214	فيم	299	فلم
45	ااي	130	تلك	215	فيما	300	فلماذا
46	اثناء	131	تلو	216	قبل	301	نحو
47	اجل	132	ثم	217	قد	302	نعم
48	اسفل	133	ثمة	218	قلما	303	نعما
49	اصبح	134	جانب	219	قليلا	304	نفس
50	اصلا	135	جعل	220	كاد	305	ها
51	اضحى	136	جلل	221	كان	306	هاتان
52	اعلى	137	جمعاء	222	كان	307	هاته
53	اعن	138	حوالي	223	كانما	308	هاتي
54	اف	139	حول	224	كاي	309	هاتين
55	افي	140	حيث	225	كاين	310	هاك
56	اقبل	141	حيثما	226	كثيرا	311	هب
57	الا	142	حين	227	كخ	312	هذا
58	البتة	143	حينما	228	كذا	313	هذان
59	اليس	144	خارج	229	كذلك	314	هذه
60	ام	145	خاصة	230	كفى	315	هذي
61	اما	146	داخل	231	كل	316	هذين
62	امام	147	درى	232	كلا	317	هكذا
63	امامك	148	ذلك	233	كلتا	318	هل
64	امد	149	ذه	234	كلما	319	هلا
65	امس	150	ذوات	235	كم	320	هلم
66	امسى	151	ذي	236	كما	321	هم
67	امن	152	رغم	237	كمن	322	هما
68	ان	153	رويدك	238	كي	323	هن
69	انا	154	سال	239	كيت	324	هنا
70	انت	155	سوف	240	كيف	325	هناك
71	انتم	156	سوى	241	كيفما	326	هنالك
72	انتما	157	شبه	242	لا	327	هو
73	انتن	158	شتان	243	لأي	328	هولاء
74	أنشأ	159	شرع	244	لأيان	329	هي
75	أنى	160	شمال	245	لأين	330	هيا
76	أنّى	161	صدقا	246	لدى	331	هيهات
77	أها	162	صراحة	247	لديك	332	و
78	أهلا	163	صوب	248	لعل	333	وراء
79	أو	164	ضد	249	لك	334	ورائك
80	أوشك	165	ضمن	250	لكم	335	ورد
81	أول	166	طاق	251	لكما	336	وسط
82	أولاء	167	طالما	252	لكن	337	وهب
83	أولالك	168	عبر	253	لم	338	ويل
84	أولى	169	عسى	254	لما	339	يومئذ

85	أولئك	170	عقب	255	لماذا		

Table 4-11 List of Arabic function words (Hijjawi, 2011)

APPENDIX SIX

SIMILARITY MEASURES CALCULATION SAMPLES

A. Word similarity measures

A.1 Using AWSS measure:

- First word (***W1***): Cushion (مسند)
- Second word (***W2***): Pillow (مخده)
- Information source: Arabic WordNet
- Path length (***l***): 0
- Depth of lowest common subsumer (***d***): 5

$$sim(W1, W2) = e^{(-\alpha * l)} * tanh(\beta * d)$$

$$sim(W1, W2) = e^{(-0.162*0)} * tanh(0.234 * 5)$$

$$sim(W1, W2) = e^{(0)} * tanh(1.17)$$

$$sim(W1, W2) = 1 * 0.82$$

$$sim(W1, W2) = 0.82$$

A.2 Using proposed measure (equation 6.6)

- First word (***W1***): Cushion (مسند)
- Second word (***W2***): Pillow (مخده)
- Information source: lexical tree
- Path length (***l***): 0
- Depth of lowest common subsumer (***d***): 5

$$sim(W1, W2) = \alpha^{(l)} * tanh(\beta * d)$$

$$sim(W1, W2) = 0.881^{(2)} * tanh(1 * 4)$$

$$sim(W1, W2) = \alpha^{(l)} * tanh(4)$$

$$sim(W1, W2) = 0.776 * 0.99$$

$$sim(W1, W2) = 0.775$$

B. Information content values:

Example word (w) : مستعد

Number of occurrence in corpus (n): 299

Total number of words in corpus (N): 9071655

$$I(w) = 1 - \frac{log(n+1)}{log(N+1)}$$

$$I(\text{مستعد}) = 1 - \frac{log(300)}{log(9071656)}$$

$$I(\text{مستعد}) = 1 - \frac{2.47}{6.95}$$

$$I(\text{مستعد}) = 1 - 0.355$$

$$I(\text{مستعد}) = 0.645$$

C. Sentence similarity measure:

First sentence: انت لست صديقا اذا كنت غير مستعد لتكون حاضرا عندما احتاجك

Second sentence: الصديق الجيد يكون دائما حاضرا عند الحاجة اليه

Joint word set: { صاحب، جيد، مستعد، حضور، احتاج، دائما }

C.1 Find similarity matrices

First Similarity matrix (***SM1***):

	صاحب	جيد	مستعد	حضور	احتاج	دائما
صاحب	1	0.53	0	0	0	0.47
جيد	0.53	1	0	0	0	0.68
مستعد	0	0	1	0.45	0	0
حضور	0	0	0.45	1	0	0
احتاج	0	0	0	0	1	0

Second similarity matrix (***SM2***):

	صاحب	جيد	مستعد	حضور	احتاج	دائما
صاحب	1	0.53	0	0	0	0.47
جيد	0.53	1	0	0	0	0.68
دائماً	0.47	0.68	0	0	0	1
حضور	0	0	0.45	1	0	0
احتاج	0	0	0	0	1	0

C.2 calculating similarity Vectors

First similarity vector (***SV1***):

1*I(صاحب) *I(صاحب)	1*I(جيد) *I(جيد)	1*I(مستعد) *I(مستعد)	1*I(حضور) *I(حضور)	1*I(احتاج) *I(احتاج)	0.68*I(دائم) *I(جيد)
0.2	0.22	0.41	0.22	0.26	0.17

Second similarity vector (***SV2***):

1*I(صاحب) *I(صاحب)	1*I(جيد) *I(جيد)	1*I(مستعد) *I(حضور)	1*I(حضور) *I(حضور)	1*I(احتاج) *I(احتاج)	1*I(دائم) *I(دائم)
0.2	0.22	0.13	0.22	0.26	0.28

C.3 Sentence similarity calculation

Sentence similarity calculation (***S(S1,S2)***):

$$\boldsymbol{S}(\boldsymbol{S1},\boldsymbol{S2}) = \frac{\sum_{i=1}^{n}(SV1_i * SV2_i)}{\sqrt{\sum_{i=1}^{n}(SV1_i)^2} * \sqrt{\sum_{i=1}^{n}(SV2_i)^2}}$$

$$\boldsymbol{S}(\boldsymbol{S1},\boldsymbol{S2}) = \frac{0.2 * 0.2 + 0.22 * 0.22 + 0 + 0.22 * 0.22 + 0.26 * 0.26 + 0.17 * 0.28}{\sqrt{0.2^2 + 0.22^2 + 0.41^2 + 0.22^2 + 0.26^2 + 0.17^2} * \sqrt{0.2^2 + 0.22^2 + 0.13^2 + 0.22^2 + 0.26^2 + 0.28^2}}$$

$$\boldsymbol{S}(\boldsymbol{S1},\boldsymbol{S2}) = \frac{0.252}{0.63 * 0.457}$$

$$\boldsymbol{S}(\boldsymbol{S1}, \boldsymbol{S2}) = 0.875$$

C.4 Sentence difference calculation:

Since there is no cell in any of the similarity vectors holding a value of zero, sentence difference is set to (1)

$$DF(S1, S2) = 1$$

C.5 Overall similarity:

$$Sim(\boldsymbol{S1}, \boldsymbol{S2}) = S(S1, S2) * DF(S1, S2)$$

$$Sim(\boldsymbol{S1}, \boldsymbol{S2}) = 0.875 * 1$$

$$Sim(\boldsymbol{S1}, \boldsymbol{S2}) = 0.875$$

APPENDIX SEVEN

RESULTS OF EMPERICAL EXPERIMENTS ON SEMANTIC SIMILARITY MEASURES

A. Experiment (1): Investigation of Word Similarity Measures

Word pairs				Path length	Depth of LCS	Human Rating	AWSS	Measure 6.4
Coast	**Endorsement**	ساحل	تصديق	0	0	0.01	0	0
Noon	**String**	ظهر	خيط	5	3	0.01	0.27	0.19
Slave	**Vegetable**	عبد	خضار	16	5	0.04	0.06	0.02
Smile	**Village**	ابتسامة	قرية	0	0	0.05	0	0
Hill	**Pigeon**	تل	حمامة	16	5	0.08	0.06	0.02
Glass	**Diamond**	كاس	الماس	17	5	0.09	0.05	0.02
Cord	**Mountain**	حبل	جبل	9	4	0.13	0.17	0.1
Forest	**Shore**	غابة	شاطئ	9	4	0.21	0.17	0.1
sepulcher	**Sheikh**	ضريح	شيخ	16	5	0.22	0.06	0.02
Tool	**Pillow**	اداة	مخدة	5	4	0.25	0.32	0.23
Coast	**Mountain**	ساحل	جبل	3	4	0.27	0.45	0.36
Tool	**Tumbler**	اداة	قدح	3	6	0.33	0.55	0.44
Journey	**Shore**	رحلة	شاطئ	0	0	0.37	0	0
Coach	**Travel**	حافلة	سفر	0	0	0.40	0	0
Feast	**Fasting**	عيد	صيام	9	4	0.49	0.17	0.1
Coach	**Means**	حافلة	وسيلة	4	4	0.52	0.38	0.29
Girl	**Sister**	فتاة	اخت	3	3	0.60	0.37	0.3
Master	**Sheikh**	سيد	شيخ	2	7	0.67	0.67	0.58
Food	**Vegetable**	طعام	خضار	2	4	0.69	0.53	0.45
Slave	**Odalisque**	عبد	جارية	0	7	0.71	0.93	0.91
Run	**Walk**	جري	مشي	3	10	0.75	0.6	0.5
Cord	**String**	حبل	خيط	2	10	0.77	0.7	0.63
Forest	**Woodland**	غابة	احراش	0	5	0.79	0.82	0.8
Cushion	**Pillow**	مسند	مخدة	0	5	0.85	0.82	0.8
Countryside	**Village**	ريف	قرية	0	5	0.85	0.82	0.8
Coast	**Shore**	ساحل	شاطئ	0	6	0.89	0.89	0.86
Tool	**Means**	اداة	وسيلة	0	7	0.92	0.93	0.91
Boy	**Lad**	صبي	فتى	0	8	0.93	0.95	0.94
Sepulcher	**Grave**	ضريح	قبر	0	5	0.94	0.82	0.8
Glass	**Tumbler**	كاس	قدح	0	6	0.95	0.89	0.86

Table 4-1 similarity results of the proposed measure using AWSS evaluation dataset

B. Experiment (2) Tuning the Proposed Word Measure

Word pairs				Path length	Depth of LCS	Human rating	Measure(6.4)
Coast	Endorsement	ساحل	تصديق	17	0	0.01	0
Noon	String	ظهر	خيط	12	0	0.01	0
Slave	Vegetable	عبد	خضار	13	0	0.04	0
Smile	Village	ابتسامة	قرية	14	1	0.05	0.13
Hill	Pigeon	تل	حمامة	15	2	0.08	0.14
Glass	Diamond	كاس	الماس	9	3	0.09	0.32
Cord	Mountain	حبل	جبل	9	2	0.13	0.31
Forest	Shore	غابة	شاطئ	4	5	0.21	0.6
sepulcher	Sheikh	ضريح	شيخ	13	2	0.22	0.19
Tool	Pillow	اداة	مخدة	5	3	0.25	0.53
Coast	Mountain	ساحل	جبل	5	5	0.27	0.53
Tool	Tumbler	اداة	قدح	5	4	0.33	0.53
Journey	Shore	رحلة	شاطئ	10	1	0.37	0.21
Coach	Travel	حافلة	سفر	13	1	0.4	0.15
Feast	Fasting	عيد	صيام	9	2	0.49	0.31
Coach	Means	حافلة	وسيلة	9	3	0.52	0.32
Girl	Sister	فتاة	اخت	5	5	0.6	0.53
Master	Sheikh	سيد	شيخ	5	5	0.67	0.53
Food	Vegetable	طعام	خضار	8	2	0.69	0.35
Slave	Odalisque	عبد	جارية	2	5	0.71	0.78
Run	Walk	جري	مشي	4	5	0.75	0.6
Cord	String	حبل	خيط	2	4	0.77	0.78
Forest	Woodland	غابة	احراش	2	6	0.79	0.78
Cushion	Pillow	مسند	مخدة	2	4	0.85	0.78
Countryside	Village	ريف	قرية	0	6	0.85	1
Coast	Shore	ساحل	شاطئ	0	7	0.89	1
Tool	Means	اداة	وسيلة	2	5	0.92	0.78
Boy	Lad	صبي	فتى	0	7	0.93	1
Sepulcher	Grave	ضريح	قبر	0	9	0.94	1
Glass	Tumbler	كاس	قدح	0	7	0.95	1

Table 4-2 tuning of the proposed word similarity measure

C. Experiment (3) Incorporating the New Word Similarity Measure in Sentence Similarity Calculation

SP	Sentences	الجمل	HR	Proposed SM
1	You're not a good friend if you're not prepared to be present when I need you.	أنت لست صديقا جيدا إذا كنت غير مستعد لتكون حاضرا عندما أحتاجك.	0.785	0.88
	A good friend always seems to be present when you need them.	الصديق الجيد يكون دائما حاضرا عند الحاجة إليه.		
2	If you continuously use these products, I guarantee you will look very young.	إذا كنت تستخدم هذه المنتجات بشكل مستمر، أنا أضمن لك سوف تظهرصغير السن جدا.	0.895	0.81
	I assure you that, by using these products consistently over a long period of time, you will appear really young.	أؤكد لك أنه باستخدام هذه المنتجات بشكل ملائم لفترة طويلة من الزمن سوف تبدو صغيرحقا.		
3	Water freezes at a certain temperature, which is zero degrees Celsius.	يتجمد الماء عند حرارة معينة، وهي صفر مئوي.	0.77	0.75
	The temperature of boiling water is 100 C and the temperature of ice is 0 C	درجة حرارة الماء المغلي هي مئة مئوية ودرجة حرارة الجليد هي صفر مئوية.		
4	We got home safely in the end, although it was a long journey.	وصلنا البيت بسلام في النهاية، على الرغم أنها كانت رحلة طويلة.	0.765	0.89
	Though it took many hours travel, we finally reached our house safely.	رغم ان ساعات السفر كانت عديدة ، اخيرا وصلنا منزلنا بسلام.		
5	A man called Dave gave his fiancee a large diamond ring for their engagement.	رجل يدعى سامر قدم لخطيبته خاتم كبيرمن الماس في الخطوبة.	0.805	0.86
	The man presented a diamond to the woman and asked her to marry him.	قدم رجل الماس للمرأة وطلب منها أن تتزوجه.		
6	Midday is 12 o'clock in the middle of the day.	منتصف اليوم هو الساعة الثانية عشر في منتصف النهار	0.99	0.85

SP	Sentences	الجمل	HR	Proposed SM
	Noon is 12 o'clock in the middle of the day.	الظهر هو الساعة الثانية عشر في منتصف النهار		
7	The first thing I do in a morning is make myself a cup of coffee. The first thing I do in the morning is have a cup of coffee.	اول شئ أفعله في الصباح هو اصنع لنفسي فنجان من القهوة. اول شئ أفعله في الصباح هو تناول فنجان من القهوة.	0.962	0.96
8	Meet me on the hill behind the church in half an hour. Join me on the hill at the back of the church in thirty minutes time	قابلني على التل وراء الكنيسة خلال نصف ساعة. التحق بي على التلة خلف الكنيسة خلال ثلاثين دقيقة من الوقت.	0.982	0.72
9	Get that wet dog off my brand new white sofa. Make that wet hound get off my white couch I only just bought it.	ابعد هذا الكلب الرطب من أريكتي البيضاء الجديدة. اجعل هذا الكلب الرطب ينزل من اريكتي البيضاء لقد اشتريتها للتو.	0.898	0.92
10	Could you climb up the tree and save my cat from jumping please? Can you get up that tree and rescue my cat otherwise it might jump?	هل يمكنك تسلق الشجرة وانقاذ قطتي من القفز رجاءأ؟ هل يمكنك صعود تلك الشجرة وانقاذ قطتي وإلا فإنها قد تقفز؟	0.958	0.89
11	I have invited a variety of people to my party so it should be interesting. A number of invitations were given out to a variety of people inviting them down the pub.	لقد دعوت مجموعة متنوعة من الناس لحفلتي لذ ستكون ممتعة قدمت عددا من الدعوات إلى لمجموعة متنوعة من الناس الى تدعوهم الى الحانه.	0.545	0.62
12	Do you want to come with us to the pub behind the hill? We are going out for drinks tonight in Salford Quays if you would like to come	هل تريد أن تأتي معنا إلى الحانة وراء التل؟ سوف نخرج هذه الليلة لتناول المشروبات في بغداد إذا رغبت أن تأتي.	0.455	0.6
13	You shouldn't be covering what you really feel	أنت لا ينبغي أن تخفي ما تشعر به حقا.	0.552	0.68

SP	Sentences	الجمل	HR	Proposed SM
	There is no point in covering up what you said, we all know	لا يوجد أي نقطة في اخفاء ما قلته، نحن نعلم جميعا		
14	You must realize that you will definitely be punished if you play with the alarm	يجب أن تدرك أنك بالتأكيد ستعاقب اذا كنت تلعب بالمنبه.	0.71	0.75
	He will be harshly punished for setting the fire alarm off.	ستعاقب بقسوة لاطفائك منبه الحريق .		
15	It seems like I've got eczema on my ear doctor, can you recommend something for me?	يبدو ان عندي الأكزيما في أذني ايها الطبيب، هل تفضل لي شيئا؟	0.512	0.47
	I had to go to a chemist for a special rash cream for my ear.	علي أن أذهب إلى الصيدلية لكريم طفح خاص لأذني.		
16	Roses can be different colours, it has to be said red is the best though.	الورود تكون بألوان مختلفة ، لكن لا بد القول ان الأحمر هو الأفضل .	0.708	0.89
	Roses come in many varieties and colours, but yellow is my favourite	الورود تأتي بأصناف وألوان متنوعة، لكن الأصفر هو الافضل لدي.		
17	Would you like to go out to drink with me tonight?	هل ترغب في الخروج للشرب معي الليلة؟	0.252	0.63
	I really don't know what to eat tonight so I might go out somewhere	أنا حقا لا اعلم ماذا ساكل الليلة لذا قد أذهب الى مكان ما		
18	I am so hungry I could eat a whole horse plus dessert	أنا جائع جدا لدرجة يمكنني أكل حصان بأكمله بالإضافة إلى حلوى	0.765	0.85
	I could have eaten another meal, I'm still starving.	كنت استطيع اكل وجبة اخرى، انا لازلت متضورا.		
19	We ran farther than the other children that day	ركضنا أبعد من الأطفال الآخرين ذلك اليوم	0.608	0.94
	You ran farther than anyone today	ركضت أبعد من الآخرين اليوم		
20	I am proud of our nation, well, most of it.	أنا فخور بأمتنا، حسنا، اغلبها.	0.428	0.48

<table>
<tr><th>SP</th><th>Sentences</th><th>الجمل</th><th>HR</th><th>Proposed SM</th></tr>
<tr><td></td><td>I think of myself as being part of a nation</td><td>أفكر في نفسي بأني جزء من أمة</td><td></td><td></td></tr>
<tr><td rowspan="2">21</td><td>Does music help you to relax, or does it distract you too much?</td><td>هل تساعدك الموسيقى على الاسترخاء، أم أنها تلهيك كثيرا؟</td><td rowspan="2">0.025</td><td rowspan="2">0.29</td></tr>
<tr><td>Does this sponge look wet or dry to you?</td><td>هل تبدو هذه الاسفنجة رطبة ام جافة بالنسبة لك</td></tr>
<tr><td rowspan="2">22</td><td>The children crossed the road very safely thanks to the help of the lollipop lady</td><td>الأطفال عبروا الطريق بسلام جدا شكرا لمساعدة بائعة المصاصات.</td><td rowspan="2">0.032</td><td rowspan="2">0.32</td></tr>
<tr><td>It was feared that the child might not recover, because he was seriously ill.</td><td>كان يخشى من أن الطفل قد لا يتعافى، لأنه كان مريضا بجد.</td></tr>
<tr><td rowspan="2">23</td><td>Boats come in all shapes and sizes but they all do the same thing.</td><td>القوارب تأتي بجميع الأشكال والأحجام ولكنها جميعا تفعل الشيء نفسه</td><td rowspan="2">0.125</td><td rowspan="2">0.54</td></tr>
<tr><td>Chairs can be comfy and not comfy, depending on the chair</td><td>الكراسي تكون مريحة او غيرمريحة، اعتمادا على الكرسي</td></tr>
<tr><td rowspan="2">24</td><td>There was a heap of rubble left by the builders outside my house this morning</td><td>كان هناك كومة من الأنقاض من قبل البنائين تركت خارج داري هذا الصباح</td><td rowspan="2">0.022</td><td rowspan="2">0.44</td></tr>
<tr><td>Sometimes in a large crowd accidents may happen, which can cause deadly injuries.</td><td>أحيانا تقع حوادث بوجود حشد كبير ، وقد يمكن أن تسبب اصابات قاتلة</td></tr>
<tr><td rowspan="2">25</td><td>I love to laugh as it makes me happy as well as those around me.</td><td>أنا أحب أن أضحك لأنه يجعلني سعيدا وكذلك الاخرين من حولي.</td><td rowspan="2">0.02</td><td rowspan="2">0.33</td></tr>
<tr><td>I thought we bargained that it would only cost me a pound.</td><td>اعتقدت اننا تفاوضنا بانه سيكلفني باوند فقط.</td></tr>
<tr><td rowspan="2">26</td><td>He was harshly punished for setting the fire alarms off.</td><td>هو عوقب بقسوة لانه اطفأ جهازتنبيه الحريق .</td><td rowspan="2">0.055</td><td rowspan="2">0.57</td></tr>
<tr><td>He delayed his response, in order to create a tense atmosphere.</td><td>تأخر رده ليخلق جو من التوتر.</td></tr>
</table>

SP	Sentences	الجمل	HR	Proposed SM
27	Someone spilt a drink accidentally on my shirt, so I changed it.	شخص ما اسقط شراب بطريق الخطأ على قميصي، لذا غيرته.	0.12	0.5
	It appears to have shrunk; it wasn't that size before I washed it	يبدو أنها تقلصت، لم تكن بهذا الحجم قبل غسلها.		
28	The damp was mostly in the very corner of the room	الرطوبة في الغالب في الزاوية البعيدة من الغرفة	0.028	0.42
	The young lady was somewhat partially burnt from the sun.	احترقت الشابة جزئيا من الشمس.		
29	Flies can also carry a lot of disease and cause maggots.	يمكن أن يحمل الذباب الكثير من المرض ويسبب اليرقات.	0.03	0.49
	I dry my hair after I wash it or I will get ill.	انا اجف شعري بعد غسله والا سوف امرض.		
30	They said they were hoping to go to America on holiday.	قالوا انهم كانوا يأملون ان يذهبوا إلى أمريكا في اجازة.	0.04	0.29
	I like to cover myself up in lots of layers, I don't like the cold.	أحب تغطية نفسي بالكثير من الطبقات، أنا لا أحب البرد.		

Table 4-3 similarity scores for dataset (SD) using word similarity measures

D. Experiment (4) Selection of Word Similarity Threshold (WST)

SP	Sentences	الجمل	scores similarity Sentence						
			HR	WST (0.0)	WST (0.1)	WST (0.2)	WST (0.3)	WST (0.4)	WST (0.5)
1	You're not a good friend if you're not prepared to be present when I need you.	أنت لست صديقا جيدا إذا كنت غير مستعد لتكون حاضرا عندما أحتاجك.	0.785	0.88	0.88	0.88	0.88	0.88	0.74
	A good friend always seems to be present when you need them.	الصديق الجيد يكون دائما حاضرا عند الحاجة إليه.							
2	If you continuously use these products, I guarantee you will look very young.	إذا كنت تستخدم هذه المنتجات بشكل مستمر، أنا أضمن لك سوف تظهرصغير السن جدا.	0.895	0.81	0.81	0.81	0.8	0.78	0.71
	I assure you that, by using these products consistently over a long period of time, you will appear really young.	أؤكد لك أنه باستخدام هذه المنتجات بشكل ملائم لفترة طويلة من الزمن سوف تبدو صغيرحقا.							
3	Water freezes at a certain temperature, which is zero degrees Celsius.	يتجمد الماء عند حرارة معينة، وهي صفر مئوي.	0.77	0.75	0.75	0.75	0.74	0.37	0.35
	The temperature of boiling water is 100 C and the temperature of ice is 0 C	درجة حرارة الماء المغلي هي مئة مئوية ودرجة حرارة الجليد هي صفر مئوية.							
4	We got home safely in the end, although it was a long journey.	وصلنا البيت بسلام في النهاية، على الرغم أنها كانت رحلة طويلة.	0.765	0.9	0.9	0.89	0.89	0.89	0.89

SP	Sentences	الجمل	scores similarity Sentence						
			HR	WST (0.0)	WST (0.1)	WST (0.2)	WST (0.3)	WST (0.4)	WST (0.5)
	Though it took many hours travel, we finally reached our house safely.	رغم ان ساعات السفر كانت عديدة ، اخيرا وصلنا منزلنا بسلام.							
5	A man called Dave gave his fiancee a large diamond ring for their engagement.	رجل يدعى سامر قدم لخطيبته خاتم كبيرمن الماس في الخطوبة.	0.805	0.86	0.86	0.86	0.85	0.76	0.76
	The man presented a diamond to the woman and asked her to marry him.	قدم رجل الماس للمرأة وطلب منها أن تتزوجه.							
6	Midday is 12 o'clock in the middle of the day.	منتصف اليوم هو الساعة الثانية عشر في منتصف النهار	0.99	0.85	0.85	0.85	0.84	0.84	0.84
	Noon is 12 o'clock in the middle of the day.	الظهر هو الساعة الثانية عشر في منتصف النهار							
7	The first thing I do in a morning is make myself a cup of coffee.	اول شئ أفعله في الصباح هو اصنع لنفسي فنجان من القهوة.	0.962	0.96	0.96	0.96	0.96	0.95	0.95
	The first thing I do in the morning is have a cup of coffee.	اول شئ أفعله في الصباح هو تناول فنجان من القهوة.							
8	Meet me on the hill behind the church in half an hour.	قابلني على التل وراء الكنيسة خلال نصف ساعة.	0.982	0.75	0.75	0.72	0.7	0.68	0.6
	Join me on the hill at the back of the church in thirty minutes time	التحق بي على التلة خلف الكنيسة خلال ثلاثين دقيقة من الوقت.							

SP	Sentences	الجمل	HR	scores similarity Sentence WST (0.0)	WST (0.1)	WST (0.2)	WST (0.3)	WST (0.4)	WST (0.5)
9	Get that wet dog off my brand new white sofa.	ابعد هذا الكلب الرطب من أريكتي البيضاء الجديدة.	0.898	0.92	0.92	0.92	0.9	0.9	0.89
	Make that wet hound get off my white couch I only just bought it.	اجعل هذا الكلب الرطب ينزل من اريكتي البيضاء لقد اشتريتها للتو.							
10	Could you climb up the tree and save my cat from jumping please?	هل يمكنك تسلق الشجرة وانقاذ قطتي من القفز رجاءأ؟	0.958	0.89	0.89	0.89	0.86	0.83	0.83
	Can you get up that tree and rescue my cat otherwise it might jump?	هل يمكنك صعود تلك الشجرة وانقاذ قطتي وإلا فإنها قد تقفز؟							
11	I have invited a variety of people to my party so it should be interesting.	لقد دعوت مجموعة متنوعة من الناس لحفلتي لذ ستكون ممتعة	0.545	0.62	0.62	0.62	0.51	0.39	0.36
	A number of invitations were given out to a variety of people inviting them down the pub.	قدمت عددا من الدعوات إلى لمجموعة متنوعة من الناس الى تدعوهم الى الحانه.							
12	Do you want to come with us to the pub behind the hill?	هل تريد أن تأتي معنا إلى الحانة وراء التل؟	0.455	0.62	0.62	0.6	0.58	0.21	0.21
	We are going out for drinks tonight in Salford Quays if you would like to come	سوف نخرج هذه الليلة لتناول المشروبات في بغداد إذا رغبت أن تأتي.							

SP	Sentences	الجمل	HR	Sentence similarity scores WST (0.0)	WST (0.1)	WST (0.2)	WST (0.3)	WST (0.4)	WST (0.5)
13	You shouldn't be covering what you really feel	أنت لا ينبغي أن تخفي ما تشعر به حقا.	0.552	0.73	0.73	0.68	0.6	0.6	0.6
	There is no point in covering up what you said, we all know	لا يوجد أي نقطة في اخفاء ما قلته، نحن نعلم جميعا							
14	You must realize that you will definitely be punished if you play with the alarm	يجب أن تدرك أنك بالتأكيد ستعاقب اذا كنت تلعب بالمنبه.	0.71	0.75	0.75	0.75	0.64	0.6	0.56
	He will be harshly punished for setting the fire alarm off.	ستعاقب بقسوة لاطفائك منبه الحريق .							
15	It seems like I've got eczema on my ear doctor, can you recommend something for me?	يبدو ان عندي الأكزيما في أذني ايها الطبيب، هل تفضل لي شيئا؟	0.512	0.47	0.47	0.47	0.47	0.12	0.1
	I had to go to a chemist for a special rash cream for my ear.	علي أن أذهب إلى الصيدلية لكريم طفح خاص لأذني.							
16	Roses can be different colors, it has to be said red is the best though.	الورود تكون بألوان مختلفة ، لكن لا بد القول ان الأحمر هو الأفضل .	0.708	0.89	0.89	0.89	0.85	0.85	0.85
	Roses come in many varieties and colors, but yellow is my favorite	الورود تأتي بأصناف وألوان متنوعة، لكن الأصفر هو الافضل لدي.							
17	Would you like to go out to drink with me tonight?	هل ترغب في الخروج للشرب معي الليلة؟	0.252	0.66	0.66	0.63	0.44	0.44	0.44

SP	Sentences	الجمل	HR	Sentence similarity scores					
				WST (0.0)	WST (0.1)	WST (0.2)	WST (0.3)	WST (0.4)	WST (0.5)
	I really don't know what to eat tonight so I might go out somewhere	أنا حقا لا اعلم ماذا ساكل الليلة لذا قد أذهب الى مكان ما							
18	I am so hungry I could eat a whole horse plus dessert	أنا جائع جدا لدرجة يمكنني أكل حصان بأكمله بالإضافة إلى حلوى	0.765	0.85	0.85	0.85	0.83	0.79	0.77
	I could have eaten another meal, I'm still starving.	كنت استطيع اكل وجبة اخرى، انا لازلت متضورا.							
19	We ran farther than the other children that day	ركضنا أبعد من الأطفال الآخرين ذلك اليوم	0.608	0.97	0.97	0.94	0.94	0.94	0.94
	You ran farther than anyone today	ركضت أبعد من الآخرين اليوم							
20	I am proud of our nation, well, most of it.	أنا فخور بأمتنا، حسنا، اغلبها.	0.428	0.55	0.55	0.48	0.48	0.45	0.45
	I think of myself as being part of a nation	أفكر في نفسي بأني جزء من أمة							
21	Does music help you to relax, or does it distract you too much?	هل تساعدك الموسيقى على الاسترخاء، أم أنها تلهيك كثيراً؟	0.025	0.29	0.29	0.29	0.23	0.12	0.12
	Does this sponge look wet or dry to you?	هل تبدو هذه الاسفنجة رطبة ام جافة بالنسبة لك							
22	The children crossed the road very safely thanks to the help of the lollipop lady	الأطفال عبروا الطريق بسلام جدا شكرا لمساعدة بائعة المصاصات.	0.032	0.37	0.37	0.32	0.04	0.03	0.03

SP	Sentences	الجمل	HR	scores similarity Sentence					
				WST (0.0)	WST (0.1)	WST (0.2)	WST (0.3)	WST (0.4)	WST (0.5)
	It was feared that the child might not recover, because he was seriously ill.	كان يخشى من أن الطفل قد لا يتعافى، لأنه كان مريضا بجد.							
23	Boats come in all shapes and sizes but they all do the same thing.	القوارب تأتي بجميع الأشكال والأحجام ولكنها جميعا تفعل الشيء نفسه	0.125	0.54	0.54	0.54	0.44	0	0
	Chairs can be comfy and not comfy, depending on the chair	الكراسي تكون مريحة او غيرمريحة، اعتمادا على الكرسي							
24	There was a heap of rubble left by the builders outside my house this morning	كان هناك كومة من الأنقاض من قبل البنائين تركت خارج داري هذا الصباح	0.022	0.47	0.47	0.44	0.39	0.29	0
	Sometimes in a large crowd accidents may happen, which can cause deadly injuries.	أحيانا تقع حوادث بوجود حشد كبير، وقد يمكن أن تسبب اصابات قاتلة							
25	I love to laugh as it makes me happy as well as those around me.	أنا أحب أن أضحك لأنه يجعلني سعيدا وكذلك الاخرين من حولي.	0.02	0.33	0.33	0.33	0.16	0.14	0.14
	I thought we bargained that it would only cost me a pound.	اعتقدت اننا تفاوضنا بانه سيكلفني باوند فقط.							
26	He was harshly punished for setting the fire alarms off.	هو عوقب بقسوة لانه اطفأ جهازتنبيه الحريق .	0.055	0.6	0.6	0.57	0.57	0.23	0.23

SP	Sentences	الجمل	HR	scores similarity Sentence					
				WST (0.0)	WST (0.1)	WST (0.2)	WST (0.3)	WST (0.4)	WST (0.5)
	He delayed his response, in order to create a tense atmosphere.	تأخر رده ليخلق جو من التوتر.							
27	Someone spilt a drink accidentally on my shirt, so I changed it.	شخص ما اسقط شراب بطريق الخطأ على قميصي، لذا غيرته.	0.12	0.5	0.5	0.5	0.39	0.39	0.33
	It appears to have shrunk; it wasn't that size before I washed it	يبدو أنها تقلصت، لم تكن بهذا الحجم قبل غسلها.							
28	The damp was mostly in the very corner of the room	الرطوبة في الغالب في الزاوية البعيدة من الغرفة	0.028	0.44	0.44	0.42	0.24	0	0
	The young lady was somewhat partially burnt from the sun.	احترقت الشابة جزئيا من الشمس.							
29	Flies can also carry a lot of disease and cause maggots.	يمكن أن يحمل الذباب الكثير من المرض ويسبب اليرقات.	0.03	0.57	0.57	0.49	0.49	0.4	0.17
	I dry my hair after I wash it or I will get ill.	انا اجف شعري بعد غسله والا سوف امرض.							
30	They said they were hoping to go to America on holiday.	قالوا انهم كانوا يأملون ان يذهبوا إلى أمريكا في اجازة.	0.04	0.4	0.4	0.29	0.13	0.13	0
	I like to cover myself up in lots of layers, I don't like the cold.	أحب تغطية نفسي بالكثير من الطبقات، أنا لا أحب البرد.							

Table 4-4 Experiment (5) Word Similarity Threshold

E. Experiment (5): Using Function Words in Similarity Measurement

SP	Sentences	الجمل	HR	With-out FW	With FW
1	You're not a good friend if you're not prepared to be present when I need you.	أنت لست صديقا جيدا إذا كنت غير مستعد لتكون حاضرا عندما أحتاجك.	0.785	0.88	0.7
	A good friend always seems to be present when you need them.	الصديق الجيد يكون دائما حاضرا عند الحاجة إليه.			
2	If you continuously use these products, I guarantee you will look very young.	إذا كنت تستخدم هذه المنتجات بشكل مستمر، أنا أضمن لك سوف تظهرصغير السن جدا.	0.895	0.81	0.92
	I assure you that, by using these products consistently over a long period of time, you will appear really young.	أؤكد لك أنه باستخدام هذه المنتجات بشكل ملائم لفترة طويلة من الزمن سوف تبدو صغيرحقا.			
3	Water freezes at a certain temperature, which is zero degrees Celsius.	يتجمد الماء عند حرارة معينة، و هي صفر مئوي.	0.77	0.75	0.75
	The temperature of boiling water is 100 C and the temperature of ice is 0 C	درجة حرارة الماء المغلي هي مئة مئوية ودرجة حرارة الجليد هي صفر مئوية.			
4	We got home safely in the end, although it was a long journey.	وصلنا البيت بسلام في النهاية، على الرغم أنها كانت رحلة طويلة.	0.765	0.89	0.83
	Though it took many hours travel, we finally reached our house safely.	رغم ان ساعات السفر كانت عديدة ، اخيرا وصلنا منزلنا بسلام.			
5	A man called Dave gave his fiance a large diamond ring for their engagement.	رجل يدعى سامر قدم لخطيبته خاتم كبيرمن الماس في الخطوبة.	0.805	0.86	0.72
	The man presented a diamond to the woman and asked her to marry him.	قدم رجل الماس للمرأة وطلب منها أن تتزوجه.			

SP	Sentences	الجمل	HR	With-out FW	With FW
6	Midday is 12 o'clock in the middle of the day.	منتصف اليوم هو الساعة الثانية عشر في منتصف النهار	0.99	0.85	0.86
	Noon is 12 o'clock in the middle of the day.	الظهر هو الساعة الثانية عشر في منتصف النهار			
7	The first thing I do in a morning is make myself a cup of coffee.	اول شئ أفعله في الصباح هو اصنع لنفسي فنجان من القهوة.	0.962	0.96	0.96
	The first thing I do in the morning is have a cup of coffee.	اول شئ أفعله في الصباح هو تناول فنجان من القهوة.			
8	Meet me on the hill behind the church in half an hour.	قابلني على التل وراء الكنيسة خلال نصف ساعة.	0.982	0.72	0.87
	Join me on the hill at the back of the church in thirty minutes time	التحق بي على التلة خلف الكنيسة خلال ثلاثين دقيقة من الوقت.			
9	Get that wet dog off my brand new white sofa.	ابعد هذا الكلب الرطب من أريكتي البيضاء الجديدة.	0.898	0.92	0.82
	Make that wet hound get off my white couch I only just bought it.	اجعل هذا الكلب الرطب ينزل من اريكتي البيضاء لقد اشتريتها للتو.			
10	Could you climb up the tree and save my cat from jumping please?	هل يمكنك تسلق الشجرة وانقاذ قطتي من القفز رجاءأ؟	0.958	0.89	0.93
	Can you get up that tree and rescue my cat otherwise it might jump?	هل يمكنك صعود تلك الشجرة وانقاذ قطتي وإلا فإنها قد تقفز؟			
11	I have invited a variety of people to my party so it should be interesting.	لقد دعوت مجموعة متنوعة من الناس لحفلتي لذ ستكون ممتعة	0.545	0.62	0.53
	A number of invitations were given out to a variety of people inviting them down the pub.	قدمت عددا من الدعوات إلى لمجموعة متنوعة من الناس الى تدعوهم الى الحانه.			

SP	Sentences	الجمل	HR	With-out FW	With FW
12	Do you want to come with us to the pub behind the hill?	هل تريد أن تأتي معنا إلى الحانة وراء التل؟	0.455	0.6	0.75
	We are going out for drinks tonight in Salford Quays if you would like to come	سوف نخرج هذه الليلة لتناول المشروبات في بغداد إذا رغبت أن تأتي.			
13	You shouldn't be covering what you really feel	أنت لا ينبغي أن تخفي ما تشعر به حقا.	0.552	0.68	0.65
	There is no point in covering up what you said, we all know	لا يوجد أي نقطة في اخفاء ما قلته، نحن نعلم جميعا			
14	You must realize that you will definitely be punished if you play with the alarm	يجب أن تدرك أنك بالتأكيد ستعاقب اذا كنت تلعب بالمنبه.	0.71	0.75	0.49
	He will be harshly punished for setting the fire alarm off.	ستعاقب بقسوة لاطفائك منبه الحريق .			
15	It seems like I've got eczema on my ear doctor, can you recommend something for me?	يبدو ان عندي الأكزيما في أذني ايها الطبيب، هل تفضل لي شيئا؟	0.512	0.47	0.6
	I had to go to a chemist for a special rash cream for my ear.	علي أن أذهب إلى الصيدلية لكريم طفح خاص لأذني.			
16	Roses can be different colors, it has to be said red is the best though.	الورود تكون بألوان مختلفة ، لكن لا بد القول ان الأحمر هو الأفضل .	0.708	0.89	0.56
	Roses come in many varieties and colors, but yellow is my favorite	الورود تأتي بأصناف وألوان متنوعة، لكن الأصفر هو الافضل لدي.			
17	Would you like to go out to drink with me tonight?	هل ترغب في الخروج للشرب معي الليلة؟	0.252	0.63	0.63
	I really don't know what to eat tonight so I might go out somewhere	أنا حقا لا اعلم ماذا ساكل الليلة لذا قد أذهب الى مكان ما			

SP	Sentences	الجمل	HR	With-out FW	With FW
18	I am so hungry I could eat a whole horse plus dessert	أنا جائع جدا لدرجة يمكنني أكل حصان بأكمله بالإضافة إلي حلوى	0.765	0.85	0.83
	I could have eaten another meal, I'm still starving.	كنت استطيع اكل وجبة اخرى، انا لازلت متضورا.			
19	We ran farther than the other children that day	ركضنا أبعد من الأطفال الآخرين ذلك اليوم	0.608	0.94	0.54
	You ran farther than anyone today	ركضت أبعد من الآخرين اليوم			
20	I am proud of our nation, well, most of it.	أنا فخور بأمتنا، حسنا، اغلبها.	0.428	0.48	0.71
	I think of myself as being part of a nation	أفكر في نفسي بأني جزء من أمة			
21	Does music help you to relax, or does it distract you too much?	هل تساعدك الموسيقى على الاسترخاء، أم أنها تلهيك كثيرا؟	0.025	0.29	0.56
	Does this sponge look wet or dry to you?	هل تبدو هذه الاسفنجة رطبة ام جافة بالنسبة لك			
22	The children crossed the road very safely thanks to the help of the lollipop lady	الأطفال عبروا الطريق بسلام جدا شكرا لمساعدة بائعة المصاصات.	0.032	0.32	0.44
	It was feared that the child might not recover, because he was seriously ill.	كان يخشى من أن الطفل قد لا يتعافى، لأنه كان مريضا بجد.			
23	Boats come in all shapes and sizes but they all do the same thing.	القوارب تأتي بجميع الأشكال والأحجام ولكنها جميعا تفعل الشيء نفسه	0.125	0.54	0.39
	Chairs can be comfy and not comfy, depending on the chair	الكراسي تكون مريحة او غيرمريحة، اعتمادا على الكرسي			
24	There was a heap of rubble left by the builders outside my house this morning	كان هناك كومة من الأنقاض من قبل البنائين تركت خارج داري هذا الصباح	0.022	0.44	0.48

SP	Sentences	الجمل	HR	With-out FW	With FW
	Sometimes in a large crowd accidents may happen, which can cause deadly injuries.	أحيانا تقع حوادث بوجود حشد كبير، وقد يمكن أن تسبب اصابات قاتلة			
25	I love to laugh as it makes me happy as well as those around me.	أنا أحب أن أضحك لأنه يجعلني سعيدا وكذلك الاخرين من حولي.	0.02	0.33	0.6
	I thought we bargained that it would only cost me a pound.	اعتقدت اننا تفاوضنا بانه سيكلفني باوند فقط.			
26	He was harshly punished for setting the fire alarms off.	هو عوقب بقسوة لانه اطفأ جهاز تنبيه الحريق .	0.055	0.57	0.59
	He delayed his response, in order to create a tense atmosphere.	تأخر رده ليخلق جو من التوتر.			
27	Someone spilt a drink accidentally on my shirt, so I changed it.	شخص ما اسقط شراب بطريق الخطأ على قميصي، لذا غيرته.	0.12	0.5	0.45
	It appears to have shrunk; it wasn’t that size before I washed it	يبدو أنها تقلصت، لم تكن بهذا الحجم قبل غسلها.			
28	The damp was mostly in the very corner of the room	الرطوبة في الغالب في الزاوية البعيدة من الغرفة	0.028	0.42	0.39
	The young lady was somewhat partially burnt from the sun.	احترقت الشابة جزئيا من الشمس.			
29	Flies can also carry a lot of disease and cause maggots.	يمكن أن يحمل الذباب الكثير من المرض ويسبب اليرقات.	0.03	0.49	0.76
	I dry my hair after I wash it or I will get ill.	انا اجف شعري بعد غسله والا سوف امرض.			
30	They said they were hoping to go to America on holiday.	قالوا انهم كانوا يأملون ان يذهبوا إلى أمريكا في اجازة.	0.04	0.29	0.74

SP	Sentences	الجمل	HR	With-out FW	With FW
	I like to cover myself up in lots of layers, I don't like the cold.	أحب تغطية نفسي بالكثير من الطبقات، أنا لا أحب البرد.			

Table 4-5 using function words in similarity measurement

F. Experiment (6): Including Sentence Difference in Similarity Measurement

SP	Sentences	الجمل	HR	Without Diff	With Diff
1	You're not a good friend if you're not prepared to be present when I need you.	أنت لست صديقا جيدا إذا كنت غير مستعد لتكون حاضرا عندما أحتاجك.	0.785	0.88	0.88
	A good friend always seems to be present when you need them.	الصديق الجيد يكون دائما حاضرا عند الحاجة إليه.			
2	If you continuously use these products, I guarantee you will look very young.	إذا كنت تستخدم هذه المنتجات بشكل مستمر، أنا أضمن لك سوف تظهرصغير السن جدا.	0.895	0.81	0.8
	I assure you that, by using these products consistently over a long period of time, you will appear really young.	أؤكد لك أنه باستخدام هذه المنتجات بشكل ملائم لفترة طويلة من الزمن سوف تبدو صغيرحقا.			
3	Water freezes at a certain temperature, which is zero degrees Celsius.	يتجمد الماء عند حرارة معينة، و هي صفر مئوي.	0.77	0.75	0.75
	The temperature of boiling water is 100 C and the temperature of ice is 0 C	درجة حرارة الماء المغلي هي مئة مئوية ودرجة حرارة الجليد هي صفر مئوية.			
4	We got home safely in the end, although it was a long journey.	وصلنا البيت بسلام في النهاية، على الرغم أنها كانت رحلة طويلة.	0.765	0.89	0.8

SP	Sentences	الجمل	HR	Without Diff	With Diff
	Though it took many hours travel, we finally reached our house safely.	رغم ان ساعات السفر كانت عديدة ، اخيرا وصلنا منزلنا بسلام.			
5	A man called Dave gave his fiancée a large diamond ring for their engagement.	رجل يدعى سامر قدم لخطيبته خاتم كبيرمن الماس في الخطوبة.	0.805	0.86	0.69
	The man presented a diamond to the woman and asked her to marry him.	قدم رجل الماس للمرأة وطلب منها أن تتزوجه.			
6	Midday is 12 o'clock in the middle of the day.	منتصف اليوم هو الساعة الثانية عشر في منتصف النهار	0.99	0.85	0.85
	Noon is 12 o'clock in the middle of the day.	الظهر هو الساعة الثانية عشر في منتصف النهار			
7	The first thing I do in a morning is make myself a cup of coffee.	اول شئ أفعله في الصباح هو اصنع لنفسي فنجان من القهوة.	0.962	0.96	0.96
	The first thing I do in the morning is have a cup of coffee.	اول شئ أفعله في الصباح هو تناول فنجان من القهوة.			
8	Meet me on the hill behind the church in half an hour.	قابلني على التل وراء الكنيسة خلال نصف ساعة.	0.982	0.72	0.72
	Join me on the hill at the back of the church in thirty minutes time	التحق بي على التلة خلف الكنيسة خلال ثلاثين دقيقة من الوقت.			
9	Get that wet dog off my brand new white sofa.	ابعد هذا الكلب الرطب من أريكتي البيضاء الجديدة.	0.898	0.92	0.82
	Make that wet hound get off my white couch I only just bought it.	اجعل هذا الكلب الرطب ينزل من اريكتي البيضاء لقد اشتريتها للتو.			
10	Could you climb up the tree and save my cat from jumping please?	هل يمكنك تسلق الشجرة وانقاذ قطتي من القفز رجاءأ؟	0.958	0.89	0.89
	Can you get up that tree and rescue my cat otherwise it might jump?	هل يمكنك صعود تلك الشجرة وانقاذ قطتي وإلا فإنها قد تقفز؟			

SP	Sentences	الجمل	HR	Without Diff	With Diff
11	I have invited a variety of people to my party so it should be interesting.	لقد دعوت مجموعة متنوعة من الناس لحفلتي لذ ستكون ممتعة	0.545	0.62	0.62
	A number of invitations were given out to a variety of people inviting them down the pub.	قدمت عددا من الدعوات إلى لمجموعة متنوعة من الناس الى تدعوهم الى الحانه.			
12	Do you want to come with us to the pub behind the hill?	هل تريد أن تأتي معنا إلى الحانة وراء التل؟	0.455	0.6	0.48
	We are going out for drinks tonight in Salford Quays if you would like to come	سوف نخرج هذه الليلة لتناول المشروبات في بغداد إذا رغبت أن تأتي.			
13	You shouldn't be covering what you really feel	أنت لا ينبغي أن تخفي ما تشعر به حقا.	0.552	0.68	0.63
	There is no point in covering up what you said, we all know	لا يوجد أي نقطة في اخفاء ما قلته، نحن نعلم جميعا			
14	You must realize that you will definitely be punished if you play with the alarm	يجب أن تدرك أنك بالتأكيد ستعاقب اذا كنت تلعب بالمنبه.	0.71	0.75	0.75
	He will be harshly punished for setting the fire alarm off.	ستعاقب بقسوة لاطفائك منبه الحريق .			
15	It seems like I've got eczema on my ear doctor, can you recommend something for me?	يبدو ان عندي الأكزيما في أذني ايها الطبيب، هل تفضل لي شيئا؟	0.512	0.47	0.47
	I had to go to a chemist for a special rash for my ear.	علي أن أذهب إلى الصيدلية لكريم طفح خاص لأذني.			
16	Roses can be different colors, it has to be said red is the best though.	الورود تكون بألوان مختلفة ، لكن لا بد القول ان الأحمر هو الأفضل .	0.708	0.89	0.89
	Roses come in many varieties and colors, but yellow is my favorite	الورود تأتي بأصناف وألوان متنوعة، لكن الأصفر هو الافضل لدي.			
17	Would you like to go out to drink with me tonight?	هل ترغب في الخروج للشرب معي الليلة؟	0.252	0.63	0.55

SP	Sentences	الجمل	HR	Without Diff	With Diff
	I really don't know what to eat tonight so I might go out somewhere	أنا حقا لا اعلم ماذا ساكل الليلة لذا قد أذهب الى مكان ما			
18	I am so hungry I could eat a whole horse plus dessert	أنا جائع جدا لدرجة يمكنني أكل حصان بأكمله بالإضافة إلى حلوى	0.765	0.85	0.85
	I could have eaten another meal, I'm still starving.	كنت استطيع اكل وجبة اخرى، انا لازلت متضورا.			
19	We ran farther than the other children that day	ركضنا أبعد من الأطفال الآخرين ذلك اليوم	0.608	0.94	0.85
	You ran farther than anyone today	ركضت أبعد من الآخرين اليوم			
20	I am proud of our nation, well, most of it.	أنا فخور بأمتنا، حسنا، اغلبها.	0.428	0.48	0.47
	I think of myself as being part of a nation	أفكر في نفسي بأني جزء من أمة			
21	Does music help you to relax, or does it distract you too much?	هل تساعدك الموسيقى على الاسترخاء، أم أنها تلهيك كثيراً؟	0.025	0.29	0.29
	Does this sponge look wet or dry to you?	هل تبدو هذه الاسفنجة رطبة ام جافة بالنسبة لك			
22	The children crossed the road very safely thanks to the help of the lollipop lady	الأطفال عبروا الطريق بسلام جدا شكرا لمساعدة بائعة المصاصات.	0.032	0.32	0.3
	It was feared that the child might not recover, because he was seriously ill.	كان يخشى من أن الطفل قد لا يتعافى، لأنه كان مريضا بجد.			
23	Boats come in all shapes and sizes but they all do the same thing.	القوارب تأتي بجميع الأشكال والأحجام ولكنها جميعا تفعل الشيء نفسه	0.125	0.54	0.54
	Chairs can be comfy and not comfy, depending on the chair	الكراسي تكون مريحة او غيرمريحة، اعتمادا على الكرسي			

SP	Sentences	الجمل	HR	Without Diff	With Diff
24	There was a heap of rubble left by the builders outside my house this morning	كان هناك كومة من الأنقاض من قبل البنائين تركت خارج داري اذا الصباح	0.022	0.44	0.44
	Sometimes in a large crowd accidents may happen, which can cause deadly injuries.	أحيانا تقع حوادث بوجود حشد كبير، وقد يمكن أن تسبب اصابات قاتلة			
25	I love to laugh as it makes me happy as well as those around me.	انا أحب أن أضحك لأنه يجعلني سعيدا وكذلك الاخرين من حولي.	0.02	0.33	0.32
	I thought we bargained that it would only cost me a pound.	اعتقدت اننا تفاوضنا بانه سيكلفني باوند فقط.			
26	He was harshly punished for setting the fire alarms off.	هو عوقب بقسوة لانه اطفا جهازتنبيه الحريق .	0.055	0.57	0.57
	He delayed his response, in order to create a tense atmosphere.	تأخر رده ليخلق جو من التوتر.			
27	Someone spilt a drink accidentally on my shirt, so I changed it.	شخص ما اسقط شراب بطريق الخطأ على قميصي، لذا غيرته.	0.12	0.5	0.43
	It appears to have shrunk; it wasn't that size before I washed it	يبدو أنها تقلصت، لم تكن بهذا الحجم قبل غسلها.			
28	The damp was mostly in the very corner of the room	الرطوبة في الغالب في الزاوية البعيدة من الغرفة	0.028	0.42	0.4
	The young lady was somewhat partially burnt from the sun.	احترقت الشابة جزئيا من الشمس.			
29	Flies can also carry a lot of disease and cause maggots.	يمكن أن يحمل الذباب الكثير من المرض ويسبب اليرقات.	0.03	0.49	0.49
	I dry my hair after I wash it or I will get ill.	انا اجف شعري بعد غسله والا سوف امرض.			
30	They said they were hoping to go to America on holiday.	قالوا انهم كانوا يأملون ان يذهبوا إلى أمريكا في اجازة.	0.04	0.29	0.29

SP	Sentences	الجمل	HR	Without Diff	With Diff
	I like to cover myself up in lots of layers, I don’t like the cold.	أحب تغطية نفسي بالكثير من الطبقات، أنا لا أحب البرد.			

Table 4-6 Sentence difference experiment results