The UCREL Semantic Analysis System

Paul Rayson[a], Dawn Archer[b], Scott Piao[b] and Tony McEnery[b]

UCREL, Lancaster University
[a]Computing Department and [b]Department of Linguistics and Modern English Language,
Lancaster University, Lancaster, LA1 4YR
{p.rayson, d.archer, s.piao, t.mcenery}@lancaster.ac.uk

Abstract

The UCREL semantic analysis system (USAS) is a software tool for undertaking the automatic semantic analysis of English spoken and written data. This paper describes the software system, and the hierarchical semantic tag set containing 21 major discourse fields and 232 fine-grained semantic field tags. We discuss the manually constructed lexical resources on which the system relies, and the seven disambiguation methods including part-of-speech tagging, general likelihood ranking, multi-word-expression extraction, domain of discourse identification, and contextual rules. We report an evaluation of the accuracy of the system compared to a manually tagged test corpus on which the USAS software obtained a precision value of 91%. Finally, we make reference to the applications of the system in corpus linguistics, content analysis, software engineering, and electronic dictionaries.

Introduction

Understanding the meaning of words seems to present little difficulty to human beings. Indeed, children as young as seven years old seem to be able to disambiguate the various meanings of polysemous words in context. Yet, this seemingly trivial task has presented a serious challenge to the NLP research community.

Researchers in machine translation (MT) have been aware of the difficulty posed by multiple meanings of words since the 1950s and 1960s (Gale *et al*, 1993). However, whilst some researchers have allegedly left the field in frustration (Bar Hillel, for example, left when he could see no way of automatically resolving the meaning of the word *pen* in the sentence "The box was in the pen"), some others have devoted remarkable efforts to word sense disambiguation (WSD).

The WSD algorithms and systems that have been suggested and developed since the 1950s tend to draw on AI-based methods, knowledge-based methods and corpus-based methods (Ide and Véronis, 1998). However, more recently, researchers have started to combine various approaches together, as a means of obtaining better results (see, for example, Stevenson and Wilks, 2001).

A WSD system generally selects a sense from a pool of possible senses of a word that matches a given context. For example, it would tag the word "bank" as *a financial institution*[1] if it finds that the surrounding words talk about financial issues, and as *river bank* if its context talks about a river. Some WSD systems can even distinguish between "bank" as *a financial institution* and "bank" as the *building containing that institution* (or one branch of it), even though such fine-grained sense disambiguation is not always necessary within NLP (many NLP problems can be solved without access to the full set of dictionary definitions).

Let's imagine a scenario in which we only want to know the domain of a journalistic report. In order to understand that the report talks about a crime case, it should be enough to know that many words in the news are about crime, law and the court[s]. For this type of task, what we need is a system that can determine the semantic category (or categories) of each word rather than a system that finds actual word sense definitions.

In this paper, we describe a semantic analysis system (USAS) developed at UCREL, Lancaster, which assigns semantic categories to English words. This system is different from most WSD systems in that it does not provide word meaning definitions. Rather, it assigns a semantic category to each word employing a comprehensive semantic category scheme that was originally based on the *Longman Lexicon of Contemporary English* (LLOCE) (McArthur, 1981). It is also different from the named entity identification systems, such as LaSIE in the GATE of Sheffield (Humphreys *et al*, 1999), in that it does not focus on one or two specific classes of words but, rather, assigns a tag or tags to every word in a running text. USAS combines several resources and approaches including the CLAWS POS tagger, semantic lexicons, a template list, contextual rules etc. And, as shown in our evaluation, the system performs to a high standard. Indeed, USAS obtained a precision of 91% on our evaluation corpus.

Our system has various applications in corpus linguistics and NLP. For example, it has been used to carry out content analysis of spoken and written discourse since 1990 (see Wilson and Rayson, 1993; Wilson and Leech, 1993; Wilson and Moudraia, forthcoming; Archer and Rayson, forthcoming). We have also used it to extract multiword expressions (MWE).[2] Currently, the UCREL team are incorporating USAS into an intelligent multilingual electronic dictionary, as part of the Benedict Project.[3] We believe that past experience points to wider possible applications of our system in practical NLP tasks.

[1] Definition can vary depending on the dictionary it uses.

[2] The results were extremely encouraging, particularly when extracting low-frequency MWEs (see Piao et al, 2003).

3 This is an EU project IST-2001-34237. Website: http://mot.kielikone.fi/benedict/.

Related work

The research areas closely related to our work include automatic word sense disambiguation (WSD) and semantic tagging. Research on the issue of word sense disambiguation has a long history, and a large body of literature in this area has been published. As mentioned in the previous section, approaches to WSD can generally be divided into AI-based, knowledge-based, and corpus-based ones.

The Al-based approaches were especially popular in the 1970s, but declined after the 1980s, when they were found to be impractical for large-scale language understanding (Ide and Veronis, 1998: 6-8). As large-scale lexical resources such as machine-readable dictionaries and WordNet (Fellbaum, 1998) have become increasingly available, the focus of WSD research has shifted towards WSD approaches using lexical resources (McRoy, 1992; Cowie et al, 1992; Harley and Glennon, 1997; Stevenson and Wilks, 2001).

Stevenson and Wilks (2001) provide an impressive example of a knowledge-based WSD approach. They combined several knowledge sources, tools and approaches, including LDOCE (*Longman Dictionary of Contemporary English*), a lemmatiser, a name entity identifier, Brill POS tagger, the simulated annealing optimisation algorithm (Cowie et al, 1992), selectional preferences, word subject codes and a feature extractor based on collocations and, as such, developed an "all-words" WSD system, which tags *all* content words in the input text. Stevenson and Wilks (2001) evaluated their system on the SEMCOR Corpus containing 200,000 words, and reported an accuracy of 94%.

Researchers who adopt a corpus-based approach to WSD research attempt to disambiguate word sense based on word usage information extracted from corpora (Brown et al, 1991; Yarowsky, 1995; Ng and Lee, 1996; Ng 1997). Often, statistical and machine learning algorithms are applied to distinguish different senses of a word based on pragmatic information extracted from the training corpora. Such approaches alone are unlikely to solve large-scale WSD problems. Consequently, corpus-based researchers often focus on small number of words (for example, Yarowsky (1995) conducted experiment on 12 words).

Other WSD work seeks to assign each content word with a semantic category using a pre-defined semantic taxonomy, e.g. tagging the word "father" as [HUMAN, MALE, ADULT] and "cucumber" as [NON-HUMAN, VEGETABLE], etc. A number of projects in this paradigm have been reported in the past decade, including Basili *et al*. (1997), Lowe *et al*. (1997), Lua (1997), Humphreys *et al* (1999), Demetriou and Atwell (2001).

Recently, SENSEVAL[4] has been developed to provide a framework for evaluating and comparing different WSD algorithms and systems. In spite of all these efforts, however, a generic WSD system efficient enough for practical application is yet to be developed.

[4] http://www.senseval.org/

The USAS system we present in this paper points to another generic semantic disambiguation system. Using this system, we attempt to attack the WSD problem by employing a broad semantic taxonomy rather than fine-grained word sense definitions. While such a system may fall short of orthodox WSD systems, our past experience has shown that it provides a practical means of coping with large-scale semantic disambiguation tasks. Furthermore, if we can design the same or similar semantic taxonomies for multiple languages, such a system can potentially provide a bridge for cross-language WSD and MT (cf. KAIST Multilingual WordNet (Oh et al, 2002)).

The USAS System

Architecture

Currently, the USAS system consists of the CLAWS POS tagger (Garside and Smith, 1997), a lemmatiser, a semantic tagger and some auxiliary format manipulating components. For POS tagging, we employ the C7 tagset[5]. Subsequent semantic disambiguation, to a large extent, depends on POS information encoded in this tagset. Evaluated over the large number of domains in the British National Corpus, CLAWS performs with success rates of between 96%-98%[6].

The core part of the USAS system is a semantic annotation component, which consists of semantic lexical resources, a set of context rules and programs implementing algorithms of disambiguation and assigning semantic tags to each word in a running text. The semantic lexicon resource is composed of two main parts: a single word lexicon and a collection of multi-word semantic templates. The former is used for providing candidate semantic categories for single words, while the latter is used for identifying multi-word expressions (MWE), including discontinuous MWEs, which depict single semantic concepts. Another knowledge source is a set of context rules, which provides context cues for some highly ambiguous words. Such words include "have" and "do", which can be used either as semantically significant content words or semantically "empty" function words.

USAS semantic taxonomy and tagset

The Lancaster USAS semantic tagset[7] was initially based on the *LLOCE* taxonomy, which also adopts a general ontological approach to semantic field analysis. However, it has been modified and revised in the light of practical tagging problems met in the course of applied research. This has included the splitting of several top level categories in LLOCE. For example the LLOCE top-level category "Arts and crafts, science and technology, industry and education" became three USAS top-level categories "Arts and crafts", "Science and technology" and "Education".

We have compared the scheme to other semantic category systems in detail and described the criteria underlying USAS in Archer et al (forthcoming). As USAS

[5] See http://www.comp.lancs.ac.uk/ucrel/claws7tags.html
[6] See http://www.comp.lancs.ac.uk/ucrel/bnc2/bnc2error.htm
[7] For the full tagset see http://www.comp.lancs.ac.uk/ucrel/usas/

automatically tags every word in a text, we have also added a category "Names and grammatical words" that captures words traditionally considered to be 'empty' of content (i.e. closed class words) and proper nouns. The revisions reflect our responses to problems met in light of tagging English texts from a variety of domains across different historical periods (Piao et al, 2004), and for a variety of purposes (e.g. market research, content analysis, information extraction, keyword extraction, etc.).

Currently the scheme includes 21 major discourse fields (shown in Table 1), which, in turn, expand into 232 categories. Letters are used to denote the major semantic fields while numbers are used to indicate subdivisions of the fields.

A	General & Abstract Terms
B	The Body & the Individual
C	Arts & Crafts
E	Emotional Actions, States & Processes
F	Food & Farming
G	Government & the Public Domain
H	Architecture, Building, Houses & the Home
I	Money & Commerce
K	Entertainment, Sports & Games
L	Life & Living Things
M	Movement, Location, Travel & Transport
N	Numbers & Measurement
O	Substances, Materials, Objects & Equipment
P	Education
Q	Linguistic Actions, States & Processes
S	Social Actions, States & Processes
T	Time
W	The World & Our Environment
X	Psychological Actions, States & Processes
Y	Science & Technology
Z	Names & Grammatical Words

Table 1 USAS tagset top level domains

USAS semantic lexical resources

As mentioned above, the USAS lexical resource consists of two main parts: a single word lexicon and a multi-word expression (MWE) lexicon. Currently, the former contains over 42,000 entries while the latter contains over 18,400 entries. Additionally, there is a small 'auto-tagging' single word lexicon where the entries are words containing wildcard characters. This lexicon contains around 50 entries such as '*kg' and '*km' to match weights and measures for example.

The single-word lexicon provides possible semantic categories for each word. Direct mapping between lemmas and semantic categories was not found to be viable in all cases. Stubbs (1996: 40) observed that "meaning is not constant across the inflected forms of a lemma" and Tognini-Bonelli (2001: 92) noted that lemma variants have different senses. Each word is combined with a POS tag, and they are mapped (together) to semantic categories. Since a word can have multiple POS tags in different contexts, a word may combine with each of the possible POS tags to form several entries. Fig. 2 shows some sample lexicon entries.

The MWE list aims to identify expressions such as phrasal verbs (*stubbed out*), noun phrases (*riding boots*), proper names (*United States of America*), true idioms (*living the life of Riley*) and their semantic categories. The semantic tags in template entries are arranged in the same way as in the single-word lexicon (see Fig. 3 for sample MWE lexicon entries).

```
occasion      NN1     T1.2 S1.1.1
occasion      VV0     A2.2
occasional    JJ      N6-
occasionally  RR      N6-
occult        NN1     S9
occupancy     NN1     H4
occupants     NN2     H4/S2mf M3/S2mf
occupation    NN1     I3.1 S7.1+
```

Fig 2: Sample of USAS word lexicon

```
stub*_* {Np/P*/R*} out_RP     O4.6-
ski_NN1 boot*_NN*             B5/K5.1
United_* States_N*            Z2
life_NN1 of_IO Riley_NP1      K1
```

Fig 3: Sample of USAS multiword templates

Notice that some entries are templates. These templates use simplified pattern matching codes, such as wildcards, as a means of capturing MWEs that have similar structures. For example, "*_* Ocean_N*1" will capture "Pacific Ocean", "Atlantic Ocean", etc. The templates not only match continuous MWEs, but also match discontinuous ones. In fact, numerous MWEs allow other words to be embedded within them. For example, the set phrase "turn on" may occur as "turn it on", "turn the light on", "turn the TV on" etc. Using the template " turn*_* {N*/P*/R*} on_RP " we can identify this set phrase in various contexts.

Semantic field disambiguation

As in the case of grammatical tagging, the task of semantic tagging subdivides broadly into two phases: Phase I (Tag assignment): attaching a set of potential semantic tags to each lexical unit and Phase II (Tag disambiguation): selecting the contextually appropriate semantic tag from the set provided by Phase I. USAS makes use of seven major techniques or sources of information in phase II. Below, we briefly describe the techniques (for further details, see Garside and Rayson 1997).

1. *POS tag.* Some candidate semantic tags can be eliminated by POS tagging. For example, consider the word "spring". There is a lexicon entry for spring that specifies (i) the possibility of a common noun tag, temporal noun tag or a verb tag, and (ii) the possibility that the common noun may have the 'coil' sense or the 'water source' sense. By choosing the common noun tag, the POS tagger can filter out the senses of 'jump' and 'season'. Hence the semantic tagger's task is simplified to choosing between the 'water source' and the 'coil':

word	*POS tag*	*semantic tag*
spring	temporal noun	[season]
spring	common noun	[coil] [water source]
spring	verb	[jump]

2. *General likelihood ranking for single-word and MWE tags*. The candidate senses in lexicon entries are ranked in terms of frequency, even though at present such ranking is derived from limited or unverified sources such as frequency-based dictionaries, past tagging experience and intuition. For example, "green" referring to colour is generally more frequent than "green" meaning inexperienced.

3. *Overlapping template resolution*. Normally, semantic multi-word expressions take priority over single word tagging, but in some cases a set of MWEs will produce overlapping candidate taggings for the same set of words. A set of heuristics is applied to determine the most likely MWE for tag assignment. The heuristics take account of length and span of the MWEs and how much of a template is matched in each case.

4. *Domain of discourse*. Knowledge of the current domain or topic of discourse is used to alter rank ordering of semantic tags in the lexicon and MWE list for a particular domain. Consider the adjective "battered" which has three candidate tags: 'Violence' (e.g. battered wife), 'Judgement of Appearance' (e.g. battered car), and 'Food' (e.g. battered cod). If the topic of conversation was known to be food, then we automatically raise the likelihood of the 'Food' semantic tag, at the expense of the other two tags.

5. *Text-based disambiguation*. Gale et al (1992) have used corpus analysis techniques to show that a given word largely keeps the same meaning within a text. For example, if a text uses "bank" in the sense of 'side of a river', all other occurrences of bank are likely to have that sense. In USAS, this method works together with step 4.

6. *Contextual rules*. The template mechanism is also used in identifying regular contexts in which a word is constrained to occur in a particular sense. Consider the meaning of the noun *account*: if it occurs in a sequence such as *NP's account of NP* it almost certainly means 'narrative explanation', whereas if it occurs in a financial context, in such collocations as *savings account* or the *balance of … account* it almost certainly has the meaning of a 'bank account'.

7. *Local probabilistic disambiguation*. It is generally supposed that the correct semantic tag for a given word is substantially determined by the local surrounding context. To return to the example of account: if this noun occurs in the company of words such as *financial*, *bank*, *overdrawn*, *money*, there is little doubt that the financial meaning is the correct one. However, we could identify the surrounding context not only in terms of (a) the words themselves, but also in terms of (b) their grammatical tags, (c) their semantic tags, or (d) some combination of (a) - (c). This method is still under development and future work includes experimentation, using a training corpus and a test corpus, to determine what weight to give each of these contextual factors for selecting the correct semantic tag of given word or word class. These and other factors are discussed in more detail in Garside and Rayson (1997).

Evaluation

Elsewhere, we have reported on the precision and recall of the MWE component (Piao et al, 2003), and the coverage of the lexicon across a variety of corpora (Piao et al, 2004). Here we report the breakdown of the errors for each word class and show the relative activation of the tagging methods when used in running text.

To evaluate the performance of the USAS system, we tested it on a corpus containing about 124,900 words. This corpus consists of transcriptions of 36 informal conversations, usually between two people in each case. After running the corpus though the semantic tagger, the output was manually corrected by a team of four post-editors. A team leader cross-checked post-editing decisions semi-automatically to ensure consistency within the team. Finally the machine-tagged version was compared against the hand-corrected one. Although we acknowledge that some human errors were inevitable, we assumed that human judgement is correct, and any machine outputs different from the hand-corrected version were counted as errors.

POS tag first letter	**Word class**	**Error relative to test-bed**	**Error relative to tag frequency**
A	Article	0.21	2.47
B	before clause marker	0.00	0.00
C	conjunction	0.05	0.60
D	determiner	0.21	4.69
E	existential there	0.01	1.22
F	formulae and foreign words	0.00	0.31
G	Genitive	0.01	6.62
I	preposition	0.36	4.16
J	Adjective	0.87	17.65
M	Number	0.29	23.93
N	Noun	2.62	16.29
P	Pronoun	0.06	0.51
R	Adverb	1.08	13.47
T	infinitive marker - to	0.11	7.52
U	interjection	0.02	0.94
V	Verb	3.03	13.21
X	negative	0.01	1.25
Z	Letter	0.00	2.67
Total		**8.95**	

Table 2 Breakdown of errors by POS

The rule-based methods produced a success rate of 91.05% on the post-edited test-bed. After applying the

various disambiguation methods, the initial ambiguity ratio[8] of 47.73% was reduced to 17.06%. Finally, the tagger selects the first choice (most likely) tag for each word and this produces the reported error rate (8.95%). Table 2 shows the breakdown by word-class of the automatic semantic tagging errors. Such an error analysis allows us to identify where the errors occur and thus helps us to improve the accuracy of the semantic tagger.

As Table 2 illustrates, most of the errors (7.60% out of 8.95%) occurred within those word classes that relate to *content* as opposed to *function*: verb (3.03%), noun (2.62%), adverb (1.08%) and adjective (0.87%). Such a result can be expected, as the sense disambiguation of content words is generally more difficult than that of function words. The number category has the largest error rate relative to tag frequency (23.93%). This is mainly due to weights and measures being mistagged. However, because numbers occurred infrequently in our running text, they account for a mere 0.29% of the overall errors in the corpus. The tagger achieved high accuracies in respect of other word classes.

In order to examine the efficacy of the different components of the tagger, we also analysed the number of times when each component was triggered for disambiguation in running text. Table 3 shows the relative hitting rates of the 14 methods we used when tagging words and MWEs in the test corpus.

Tagging method	**Relative frequency**
Lexicon	63.68
Lexicon with stemming	3.41
Lexicon with lemmatisation	0.03
Auto-tag rule	0.39
Domain of discourse	7.67
Auxiliary verb	6.76
Context rules	0.83
Lexicon ignoring POS	0.92
Lexicon with stemming ignoring POS	0.07
WordNet unknown word look-up	0.05
Wildcard multi-word-expression	0.54
Multi-word-expression	11.60
Multi-word-expression and domain of discourse	4.06
Total	**100.00**

Table 3 Breakdown of tagging methods

Notice that, for almost 70% of the time, the semantic field was disambiguated through lexicon look-up, i.e. a combination of lexicon look-up of the surface forms and that of the stemmed or lemmatised forms. The MWE component was applied to just over 15% of words in the test corpus while the semi-automatic algorithm of assigning a domain of discourse covered almost 8%. Auxiliary verb identification appears to be particularly important since the CLAWS POS tagger does not distinguish between auxiliary and lexical verbs at the POS level. Note that, as the statistical disambiguation component is still under development, it was not included in our experiment, and hence this table does not reflect the performance of the statistical disambiguation algorithm.

[8] We define *initial ambiguity ratio* as the percentage of words in a text with more than one possible semantic tag assigned from the semantic lexicon and MWE list before the application of disambiguation techniques.

Conclusion and future work

In this paper, we described the USAS semantic tagging system. Employing a hierarchical semantic taxonomy, semantic lexical resources and a number of disambiguation algorithms such as templates, context rules etc., USAS assigns semantic categories to words and MWEs in a running text. Although different from many existing WSD systems, we believe that our system provides a practical tool for large-sale semantic annotation tasks, and that it can also support/enhance WSD systems. We also contend that such an approach would be useful for cross-language WSD and machine translation, if parallel systems were developed for other languages.

In Lancaster, further research work is under way, aiming to improve and apply the USAS system for linguistic study and language engineering tasks. For example, USAS has been used in the software engineering domain for the analysis of large volumes of technical documentation (Sawyer et al, 2002), and in decision management (Rayson et al, 2003). We are also modifying it to make it capable of tagging historical text semantically (Archer et al, 2003). Other current work includes mapping its tagset to WordNet synsets, investigating techniques to automatically detect new MWEs, and developing a mirror semantic tagger for Finnish (Lofberg et al, 2003) as part of the effort to enhance electronic dictionaries. We envisage that the USAS system will find wider applications and provide useful tool for both corpus linguistics and NLP communities.

Acknowledgements

This work is continuing to be supported by the Benedict project, EU funded IST-2001-34237. Much of the early development of the system was funded under two EPSRC projects (running between 1990 and 1996) involving Andrew Wilson and Paul Rayson and supervised by Geoffrey Leech, Roger Garside and Jenny Thomas.

References

Archer, D., Rayson, P., Piao, S., McEnery, T. (forthcoming). Comparing the UCREL Semantic Annotation Scheme with Lexicographical Taxonomies. To be presented at European Association for Lexicography 11th International Congress (Euralex 2004), Lorient, France.

Archer, D., McEnery, T., Rayson, P., Hardie, A. (2003). Developing an automated semantic analysis system for Early Modern English. In Archer, D., Rayson, P., Wilson, A. and McEnery, T. (eds.) Proceedings of the Corpus Linguistics 2003 conference. UCREL technical paper number 16. (pp. 22--31) UCREL, Lancaster University.

Archer, D. and Rayson, P. (forthcoming). Using the UCREL automated semantic analysis system to investigate differing concerns in refugee literature. In proceedings of the *Keywords workshop, February 5-6,*

2004. Office for Humanities Communication, Centre for Computing in the Humanities, King's College London.

Basili, R., M. Della Rocca, and M.T. Pazienza. (1997). Towards a bootstrapping framework for corpus semantic tagging. In *Proceedings of the SIGLEX Workshop "Tagging Text with Lexical Semantics: What, why and how?"*, Washington, D.C., April. ANLP.

Brown, P., Pietra S., Pietra, V. and Mercer, R. (1991) Word sense disambiguation using statistical methods. In Proceedings of the 29th Annual Meeting of the ACL, (pp 264-270) Berkeley, California.

Cowie, J., Guthrie, J. and Guthrie, L. (1992) Lexical Disambiguation using Simulated Annealing. *Proceedings of the 16th International Conference on Computational Linguistics (COLING-92)* (pp. 359-365) Nantes, France, July.

Demetriou, G. and Atwell, E. (2001) A domain-independent semantic tagger for the study of meaning associations in English text. In *Proceedings of the 4th International Workshop on Computational Semantics (IWCS 4)* (pp. 67-80). Tilburg, Netherlands.

Fellbaum, C. (1998) A Semantic Network of English Verbs. In Fellbaum, C. (ed.).*WordNet: An Electronic Lexical Database*. (pp. 69-104) Cambridge, Mass.: MIT Press.

Gale, W., Church, K., and Yarowsky, D. (1992), One Sense Per Discourse. *Proceedings of the 4th DARPA Speech and Natural Language Workshop*. (pp.233-237).

Gale, W., Church, K. and Yarowsky, D. (1993) A method for disambiguating word senses in a large corpus. Computers and the Humanities (26), pp. 415 - 439.

Garside, R., and Smith, N. (1997) A hybrid grammatical tagger: CLAWS4, in Garside, R., Leech, G., and McEnery, A. (eds.) *Corpus Annotation: Linguistic Information from Computer Text Corpora*. (pp. 102-121) Longman, London.

Garside, R. and Rayson, P. (1997) Higher-level annotation tools, in Garside, R., Leech, G., and McEnery, A. (eds.) *Corpus Annotation: Linguistic Information from Computer Text Corpora*. (pp. 179-193) Longman, London.

Harley, A. and D. Glennon. (1997). Sense tagging in action: Combining different tests with additive weights. In *Proceedings of the SIGLEX Workshop "Tagging Text with Lexical Semantics"*. Association for Computational Linguistics, (pp. 74-78) Washington, D.C.

Humphreys, K., Gaizauskas, R., Huyck, S., Mitchell, B., Cunningham, H., and Wilks Y. (1999) Description of the University of Sheffield LaSIE-II System as used for MUC-7**.** In Proceedings of MUC-7. Morgan Kaufmann.

Ide, N. and Veronis, J. (1998) Introduction to the special issue on word sense disambiguation: The state of art. Computational Linguistics, 24(1): 1—40.

Lofberg, L., Archer, D., Piao, S. L., Rayson, P., McEnery, T., Varantola, K., Juntunen, J-P. (2003). Porting an English semantic tagger to the Finnish language. In D. Archer, P. Rayson, A. Wilson and T. McEnery (eds.) Proceedings of the CL2003 conference. UCREL technical paper number 16. (pp. 457 – 464) UCREL, Lancaster University.

Lowe, J. B. Baker, C. and Fillmore, C. (1997) 'A frame-semantic approach to semantic annotation'. In Proceedings of the SIGLEX workshop "Tagging Text with Lexical Semantics'. (pp. 18-24) Washington. D.C.

Lua. K. T. (1997) 'An efficient inductive unsupervised semantic tagger'. *Computer Processing of Oriental Languages*, 1(1), pp. 35-47.

McRoy, S. (1992). Using multiple knowledge sources for word sense disambiguation. *Computational Linguistics*, 18(1):1-30.

McArthur, T. (1981) *Longman Lexicon of Contemporary English*. Longman, London.

Ng, H. T. and Lee, H. B. (1996) Integrating multiple knowledge sources to disambiguate word sense: an exemplar-based approach. In *Proceedings of AC'L96*, (pp. 40-47) Santa Cruz, CA.

Ng, H. T. (1997) Exemplar-based word sense disambiguation: some recent improvements. In *Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing (EMNLP).* (pp. 208-213) Somerset, New Jersey.

Oh J-H., Saim S., Yong-Seok C.i, Key-Sun C. (2002) Word Sense Disambiguation with Information Retrieval Technique. In proceedings of LREC 2002, Las Palmas, Spain, May 2002.

Piao, S. L., Rayson, P., Archer, D., Wilson, A. and McEnery, T. (2003) Extracting Multiword Expressions with a Semantic Tagger. In *proceedings of the Workshop on Multiword Expressions: Analysis, Acquisition and Treatment, at ACL 2003*, (pp. 49-56) Sapporo, Japan, July 12, 2003.

Piao, S. L., Rayson, P. Archer, D., McEnery, T. (2004). Evaluating Lexical Resources for A Semantic Tagger. *LREC 2004*, May 2004, Lisbon, Portugal.

Rayson P., Sharp B., Alderson A., et al (2003). Tracker: a framework to support reducing rework through decision management. In *Proceedings of ICEIS2003*. (pp. 344 – 351) Angers - France, April 23-26, 2003. Volume 2.

Sawyer, P., Rayson, P., and Garside, R. (2002) REVERE: support for requirements synthesis from documents. *Information Systems Frontiers Journal*. Volume 4, Issue 3, Kluwer, Netherlands, pp. 343 - 353.

Stevenson, M. and Wilks, Y. (2001) The interaction of knowledge sources in word sense disambiguation. *Computational Linguistics* 27(3).

Stubbs, M. (1996). *Text and corpus analysis: computer-assisted studies of language and culture*. Blackwell, Oxford.

Tognini-Bonelli, E. (2001). *Corpus linguistics at work*. Benjamins, The Netherlands.

Wilson, A. and Rayson, P. (1993). Automatic content analysis of spoken discourse. In C. Souter and E. Atwell (eds.), *Corpus Based Computational Linguistics*. (pp. 215-226) Amsterdam: Rodopi.

Wilson, A. and Leech, G.N. (1993). Automatic Content Analysis and the Stylistic Analysis of Prose Literature. Revue: *Informatique et Statistique dans les Sciences Humaines 29*: 219-234.

Wilson, A. and Moudraia, O. (forthcoming) Quantitative or Qualitative Content Analysis? Experiences from a cross-cultural comparison of female students' attitudes to shoe fashions in Germany, Poland and Russia. To appear in Wilson, A., Rayson, P. and Archer, D. (eds.) *Corpus Linguistics around the world*. Rodopi, Amsterdam.

Yarowsky, D. (1995) Unsupervised word sense disambiguation rivalling supervised methods. In *Proceedings of ACL-95*. (pp. 189-196) Cambridge. Massachusetts.