A statistical clustering method to provide a semantic indexing of texts

Thierry Poibeau
Laboratoire d'Informatique de Paris-Nord
Institut Galilée
Av. J.-B. Clément
93430 Villetaneuse
Thomson-CSF/LCR
Domaine de Corbeville
F-91404 Orsay
Thierry.Poibeau@lcr.thomson-csf.com

Abstract

Information Extraction is intended to extract only a few pieces of information from a text. The main topics of a document, disseminated through the text, can't be automatically discovered by classical techniques based on local parsing. To solve this problem, we present the integration of a semantic indexing method in an Information Extraction system. We show the interest of such an analysis based on stochastic methods in order to extract information disseminated through the text and propose a thematic indexing. The algorithm is presented and evaluated on a concrete example.

1 Introduction

Information Extraction (IE) is a technology dedicated to the extraction of structured information from texts to fill pre-defined templates. The American Message Understanding Conferences (MUC) provided a relevant framework for the development of research in this area. IE is known to have established a now established linguistic architecture based on cascading automata (Appelt *et al.*, 1993). The analysis is exclusively local and it has been proved that general systems have inferior results than specific systems performing only local analyses.

Even if IE seems to be now a relatively mature technology, it suffers a number of yet unsolved limitations, which cause problems for its dissemination through industrial applications. IE operates on a very homogeneous amount of texts and performances decrease rapidly if one gives to the system a wide variety of texts as input. Besides, some information disseminated through texts cannot be extracted by a strictly local syntactic analysis (Allan *et al.*, 1998). Both these limitations call for the integration of statistical techniques traditionally used for Information Retrieval into IE systems. We present an experiment which automatically detects the main topics of a film from semantic regularities appearing in its reviews. The list of the main topics is displayed in an interface that allows the end-user to choose what he wants to see from thematic criteria.

Firstly, the IE system developed in the framework of the European ECRAN project is presented. We then introduce the LexiconTuner, a tool that analyses texts to extract semantic clusters. Finally, the integration of a statistical analysis in IE systems is considered, with examples taken from the ECRAN project.

2 An Information Extraction System involving a Statistical Analysis

In the last few years, a new kind of Information Extraction system emerged which was more portable and adaptive. Such an Information Extraction system has been developed in the framework of ECRAN (a European research project that aims to develop a new technology for Information Extraction systems). Specific developments have been made for French, English and Italian. All the tools are integrated through the platform for linguistic engineering Gate from the University of Sheffield (Cunningham *et al.*, 1996).

The IE system developed for French can be divided into three parts: modules for shallow parsing, modules for named entity recognition and then modules to complete a result template. The shallow parsing stage produces a tagged text, with morphological and POS information. The second stage allows recognition of proper names, organizations, locations, dates, etc. The last stage links all this information together to produce a result template which presents a synthetic view of the text.

The corpus is made of movie reviews from the French newspaper *Le Monde*. The idea is to complete a

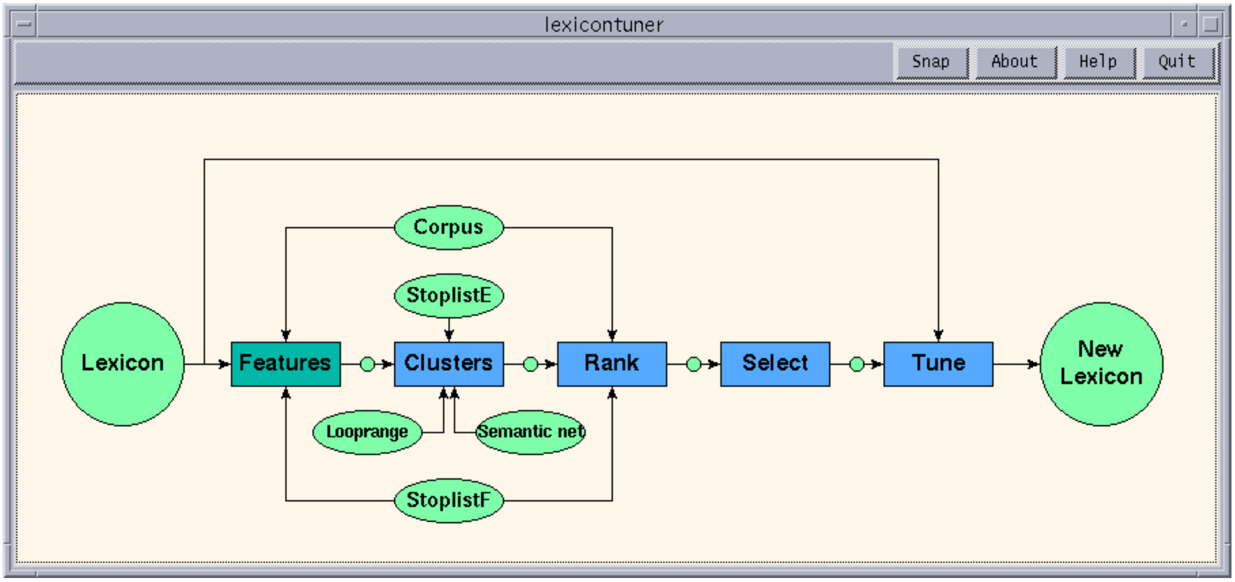

Figure 1: The

database about films starting from their reviews: the extracted information concerns the title of the film, the director, the actors and its main topics (`love`, `war`, `music`...). The end-user can then choose a film according to these different criteria through the interface shown above. A "cluster extractor" module provides a semantic indexing of the texts to enable the end-user to choose a film from its main topics. This thematic indexing and the traditional extraction task complement each other, the former providing disseminated information, which is not strictly expressed by any syntactic phrase. The cluster extractor module is part of a more general system for semantic analysis of corpora.

This system, the "LexiconTuner", dedicated to the automatic analysis of the semantics of texts, was developed in the framework of the ECRAN project (Ecran, 1996b). It provides a semi-automatic extraction of semantic clusters, which can be manually controlled and corrected. It is actually used to enhance the quality of extraction systems in producing:

- a semantic indexing of texts reflecting their main topics,
- homogeneous semantic classes to help the user fill extraction patterns during the elaboration of the system.

These issues are considered in the section describing the experiment on *Le Monde* corpus, in section 4.

3 A Method for Semantic Indexing

We developed a mainly automatic method that combines knowledge contained in on-line resources and statistical data obtained from the training corpus of the chosen domain. This method is implemented through the LexiconTuner, which was initially developed for French but can be used for any language if correct resources are provided.

The system can be divided into two main parts.

- The cluster generator produces a list of semantic clusters that occur in the corpus, by clustering together related lexical items found in the corpus.
- The cluster ranker then measures the salience of each semantic cluster with respect to the task of Information Extraction and ranks them accordingly.

3.1 Extraction of Semantic Clusters from Lexical Resources

Semantic clusters–that is to say clusters of semantically related words–can be acquired from on-line resources including dictionaries, thesauri and taxonomies (Wilks *et al.*, 1995). Dictionaries model the meanings of lexical items using textual definitions (Rastier *et al.*, 1994). Textual definitions are written in natural language and the full information encoded in these definitions can't be extracted easily. However, partial information, which can be used as semantic clusters, can be extracted from the definitions relatively easily.

The idea of building networks from definitions was first proposed by Véronis and Ide (1990), along with a propagation technique for computing clusters of related words. Many authors have proposed techniques for deriving some kind of clusters or associations of concepts from graphs of concepts, among others transitive closures and computation of graph "cliques", simulated annealing, etc. Some of these experiments derived information from a dictionary; see for example the early work at New Mexico State University (Slator, 1988a and 1988b) (Fowler and Dearholt, 1989). In our approach, the content words in the definitions of a word sense are used as component concepts of the word sense (Luk, 1995), that is to say that a concept is a lexical item considered as a member of a semantic cluster. For example, the word *opera*, which is defined as "dramatic performance or composition of which music is an essential part..." in the *Concise Oxford Dictionary*, can be assigned the semantic concepts of `dramatic`, `performance`, `composition` and `music`.

In the ECRAN prototype, the generation of the clusters is carried out in two steps. Firstly, the list of all the concepts occurring in the corpus is generated. For each word in the corpus, the program looks for its textual definition(s) in the lexicon. The words in the definitions are treated as concepts and are recorded. This step allows us to go from a French word (e.g. *opéra*) to some concepts labelled with English words (`dramatic`, `performance`, `composition`, `music`), by means of a bilingual dictionary (*opera* being defined as a "dramatic performance or composition of which music is an essential part").

Secondly, semantic clusters are generated by clustering related concepts in the list. The program uses a semantic net as its semantic knowledge source for determining the semantic relatedness between the concepts. The net is derived from the *Oxford Advanced Learner's Dictionary* (OALD). Each node represents a word, and for any two words *w1* and *w2*, *w1* is linked to *w2* if *w2* appears in the definitions of *w1* in OALD (`music` will be linked to `composition` if `music` appears in the definition of `composition`). The link from *w1* to *w2* is weighted inversely proportionally to the product of the number of senses of *w1* as defined in OALD and the number of words in the definition of *w1* that contains *w2*. The clusters are formed in 3 steps:

- Every circle in the semantic net which length is equal to a user-defined number is identified. For example, if the number is 3, the system will generate all the sets consisting of three related words according to the dictionary definition. For `opera`, the following associations will be considered: (`opera`, `performance`, `dramatic`) and (`dramatic`, `composition`, `music`)..

- Circles that share one or more nodes are merged. Thus, (`opera`, `performance`, `dramatic`) and (`dramatic`, `composition`, `music`) are merged into a single set (`opera`, `performance`, `dramatic`, `composition`, `music`). This is called a "core cluster".

- Lastly, peripheral words (words which are part of a circle whose length is inferior to the user-defined number) are related to the core cluster. If there is two words related like (`dramatic`, `comic`), then `comic` will be added to the cluster, being related to `dramatic`.

We can ask ourselves if this algorithm really capture all the complexity of dictionary definitions. For example, although cycles are common among dictionary definition terms, there may also be chains of words which are highly related but which do not form a cycle. To solve this problem, (Lesk, 1986) suggested a method to cluster word pairs based on "degree of overlap" between their head words or definition words, in the context of word sense disambiguation. In our system, the merging step and the addition of peripheral words partially allow taking into account chains of words.

Lastly, the membership of a cluster is "weighed". The weight of a core member is taken as the inverse of the number of senses of the word as defined in OALD. The weight of a non-core member is taken as the mean of the weights of the core members of the cluster to which it is related. Then, semantic clusters can capture notions appearing in texts independently from the lexical items expressing these notions.

3.2 Identification of Useful Semantic Clusters from a Text

Not all semantic clusters are equally useful for Information Extraction in any particular domain. The usefulness of a semantic cluster can be determined statistically from the corpus itself. (Salton, 1988) suggested measures that reveal the usefulness of a term: applied to the members of a cluster, these measures reveal the usefulness of a semantic cluster in the corpus (Resnik, 1995) (Aguirre & Rigau, 1996). A comparison of 3 measures is suggested in (Sta, 1995): frequency, discrimination value and variance. An experiment is made on a technical corpus and the evaluation reveals that the variance offers the best discriminatory power, even if a combination of the different measures can give even better results.

The ECRAN prototype for French implemented a method to measure the usefulness of semantic clusters using a single function which captures the variance of the occurrence frequency (Luk, 1995) (Ecran, 1996a).

The variance of the occurrence frequency of a semantic cluster s is given as:

$$Var(f(s,A)) = E\left[(f(s,A) - E(f(s,A)))^2\right]$$

where $f(s,A)$ is the occurrence frequency of s in a collection of articles A, and $E(X)$ is the mean of X for any X.

The measurement captures both the discriminatory power and the relevancy of a semantic cluster: if the occurrence frequency of a semantic cluster has a high variance, then it must be both discriminative and relevant. If a semantic cluster has low relevancy (i.e. the cluster has low occurrence frequency in most of the articles), then the variance of occurrence frequency will be low relatively to a cluster with a higher frequency. At the same time, if a semantic cluster has low discriminatory power (i.e. the semantic cluster has similar occurrence frequency in most articles), then the variance of occurrence frequency will also be low. Lastly, it also takes into account the dependency of the discriminatory power on the occurrence frequencies, normalized according to the size of the document. The weight of a cluster is then normalized according to the size of the document using the following function:

$$Wn(f,d) = W(f,d) * \frac{1}{\log(N(d)/200)+1} \quad \text{if } N(d) \geq 200$$

$$= W(f,d) * \log(200/N(d)) + 1 \quad \text{otherwise}$$

where $Wn(f,d)$ and $W(f,d)$ are the normalized and non-normalized weights of the feature f in document d, and $N(d)$ is the number of content words in document d. The use of a logarithmic expression allows to smooth the result in spite of the differences in the size of the documents. The constant 200 is the average size of a document in the corpus, without taking into account empty words.

After a rough discrimination between the more useful semantic clusters and the less useful ones obtained by these statistical methods, domain experts can manually refine the selections of the semantic clusters which are to be used in the Information Extraction system. Lastly a set of representative clusters is associated to each text. These clusters are those which have a higher weight in the text than in the corpus.

4 Semantic indexing to enhance the quality of an IE system

The LexiconTuner is used to enhance the ECRAN prototype by providing some information about the main topics of a film in addition to the information concerning the director, the actors, etc. This information is disseminated through the text and may be extracted by stochastic methods.

4.1 Extracting Concepts reflecting the main Topics of Texts

If some explicit information can be linguistically extracted, some other is disseminated through the articles and semantic regularities must be registered in order to get the main topics of the text. The hypothesis was made that there is a close link between the main topics of the film and those of its reviews. We automatically extract about 1000 critics from the French newspaper *Le Monde* for this experiment.

The algorithm used for semantic indexing is an implementation of the method we saw before. Firstly, the Cluster Generator is used to generate a list of clusters of the concepts that occur in the corpus. We chose a circle size of 2, which seems the best value to produce pertinent clusters regarding the semantic relatedness of the members of the clusters. A total of 676 different clusters are generated. The figure below presents the cluster for `affection`.

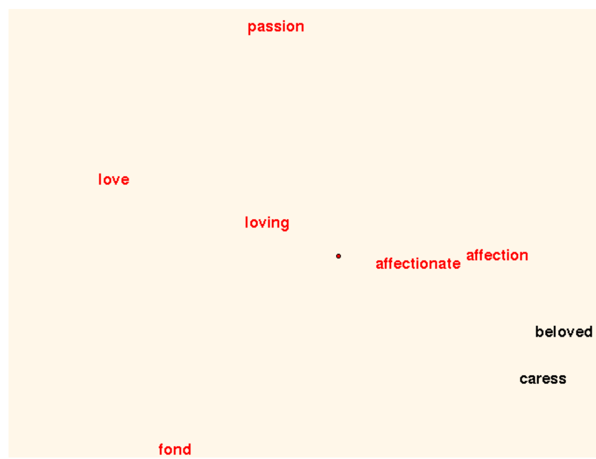

Figure 2: a cluster generated by the system reflecting the notion of `affection`

Clusters can be divided into three main types according to the nature of the core members of the clusters (which are mainly adjectives, verbs or nouns).

A manual step is then necessary to choose among clusters those which are the most descriptive according to the corpus. This work is guided by the results produced by the LexiconTuner. The clusters are sorted into the three different classes manually because the distinction between clusters of different types is not always obvious and classification very often requires intelligent judgement. In some cases, the core members are of different parts of speech and in others, the core members are polysemous words with senses of different parts of speech (e.g. *worry*). After the sorting process, there are 129 adjective clusters, 273 noun clusters and 131 verb clusters. The rest of the clusters (129) are semantically non-coherent and are discarded in the sorting process.

We finally retain only the noun clusters that appear to be representative of the content of the corpus. They express objects and notions whereas adjectives express qualities and verbs express relations between objects, which subjectively seemed to be less representative of the texts. However, in an IE process we must keep in mind that verbs are as important as nouns to structure the extracted information. Verbs allow to assign thematic roles to entities and to bind them to extraction patterns.

4.2 Some results from the Movie Reviews corpus

The set of noun clusters is ranked according to the corpus using the cluster ranker. The top 40 ranked clusters are extracted and reflect the main topics of the corpus. According to the analyst, two clusters are found to be very similar to other clusters in the list and are removed. A name is given to each of the remaining clusters.

```
Lover               (1.02e+00)
Picture             (6.89e-01)
Actor               (6.53e-01)
Finance             (5.73e-01)
Movie               (5.02e-01)
Marriage            (4.20e-01)
Event               (3.77e-01)
Result              (3.76e-01)
Mail                (3.73e-01)
Family              (3.39e-01)
Castle              (2.92e-01)
Illustration        (2.45e-01)
Sex                 (2.35e-01)
View                (2.33e-01)
Experience          (2.28e-01)
Motor-vehicle       (2.15e-01)
World               (2.04e-01)
Image               (2.03e-01)
Affection           (1.88e-01)
Impression          (1.87e-01)
Travel              (1.78e-01)
Space               (1.76e-01)
Plant               (1.51e-01)
Youth               (1.47e-01)
Standard            (1.47e-01)
Drawing             (1.35e-01)
War                 (1.24e-01)
Music               (1.18e-01)
Health              (1.17e-01)
Birth               (1.12e-01)
Evidence            (1.12e-01)
Difficulty          (1.04e-01)
Interest            (9.98e-02)
Benefit             (9.75e-02)
Home                (9.39e-02)
Desire              (9.19e-02)
Sorrow              (9.13e-02)
Style               (9.01e-02)
```

The labels of the 38 most salient clusters and their salience values are shown above. Each text is tagged by a set of about ten of the most salient clusters according to the comparison of their weight in the text and in the corpus. These clusters are displayed in an interface and allow to select a film from its topics and characteristics (see section 2). Most of the time, people do not know exactly what they want to see, but select a film by its topics. The ECRAN prototype aims at fulfilling these needs.

4.3 Evaluating the Quality of the Semantic Indexing Task

An evaluation of the semantic indexing has been performed on 50 manually tagged texts. The texts were tagged twice, by two different readers, who are referred to as H1 and H2 below.

	S:H1	H2:H1
Recall	56.2 (27.4)	83.2 (39.3)
Precision	49.2 (23.9)	50.6 (23.9)

	S:H2	H1:H2
Recall	47.1 (27.4)	50.6 (23.9)
Precision	67.6 (39.3)	83.2 (39.3)

Accuracy of the tuned system in processing the corpus is shown in the table above. Accuracy of the system is measured in terms of recall and precision concerning the assignments of each of the human judges. For example, the column designated as S:H1 contains the recall and precision achieved by the system (S) compared to the assignments of the human judge H1. The recall and precision measured from the comparison of the results of each of the human judges are also shown (H2:H1 and H1:H2). Baseline recall and precision values are also shown for comparison (in brackets following the actual value achieved). These baseline values are taken as the expected recall and precision obtained by randomly assigning the same total number of semantic clusters to the set of articles. Note that the baseline values are different for each set of results because each set of results has a different total of assignments.

The accuracy of the system with respect to the individual human judges is lower than, but still comparable to those of the human judges with respect to each other, taking into account the varying baseline values (e.g. with respect to H1, the system achieves the recall of 56.2% compared to the 83.2% achieved by H2). These results are also significantly higher than other results obtained by randomly assigning semantic clusters to the set of articles.

5 Conclusion

We presented a method to extract salient semantic clusters from a corpus. Our approach is semi-automatic and provides good results. It is also inexpensive and not time-consuming. The results can be significantly improved if the automatic process is aided by assistance from human experts. The experiments we made concerned the field of Information Extraction. We proved that a part of the information contained in texts could not be extracted by a strictly local linguistic analysis. We showed that semantic indexing then offers a good framework to enhance the quality of current IE systems and offers a new kind of access to information to end-users.

More generally, this paper raises some questions about the integration of stochastic methods in IE systems. These methods allow us to explore texts and to extract their main topics. They could provide a homogeneous input to IE systems in offering high quality pre-filtering. A refinement of the analysis has to be considered in order to deal with small linguistic variations (Riloff, 1995) and textual genres (Biber *et al.*, 1998), which may have important consequences on the quality of the results. The integration of IE systems, Topic Detection methods and intelligent access to text via hypertexts offer new approaches to electronic documentation.

Acknowledgements

This paper reflects some results from the European ECRAN project, partially fund by the EC (LE-2110). I would like to thank Alpha Luk for his comments on a previous version of this paper. This paper reuses some of the material he produced in the framework of the ECRAN project. I would like to thank Adeline Nazarenko and two anonymous reviewers for their valuable comments that slightly improve the quality of this paper. Lastly, Evi Commins and François Toussenel carefully reread this paper to correct by faulty English. Of course, all the remaining errors are mine.

References

(Aguirre & Rigau, 1996) E. Aguirre & G. Rigau, "Word Sense Disambiguation using conceptual density", *Proceedings of the 16th International Conference on Computational Linguistics (COLING)*, Copenhagen, 1996.

(Allan *et al.*, 1998) J. Allan, J. Carbonnel, G. Doddington, J. Yamron and Y. Yang, *Topic Detection and Pilot Study (Final report)*, Proceedings of the DARPA broadcast news transcription and understanding workshop, 1998.

(Appelt *et al.*, 1993) D.E. Appelt, J. Hobbs, J. Bear, D. Israel, M. Kameyana & M. Tyson, "FASTUS: a finite-state processor for information extraction from real-world text", *Proceedings of the 13th International Joint Conference on Artificial Intelligence*, Chambéry, France, 1993.

(Biber *et al.*, 1998) D. Biber, S. Conrad & R. Reppen, *Corpus linguistics: investigating language structure and use*, Cambridge University Press, Cambridge, MA, 1998.

(Cowie *et al.*, 1992) J. Cowie, J. Guthrie. & L. Guthrie, "Lexical disambiguation using simulated annealing", *Proceedings of the 14th International Conference on Computational Linguistics (COLING'92),* Nantes, 1992.

(Cunningham *et al.*, 1996) H. Cunningham, K. Humphreys, R. Gaizauskas & Y. Wilks, "GATE: a general architecture for text engineering", *Proceedings of the 14th International Conference on Computational Linguistics (COLING)*, Copenhagen, 1996.

(Ecran, 96a) *D-2.5.1 - Methods for Lexical Items Modification/Creation*, ECRAN Project Deliverable, 1996 ("ftp://ftp.dcs.shef.ac.uk/share/ilash/Ecran/DEL.2-5-1.ps.gz".

(Ecran, 96b) *D-2.6.1 - Heuristics for Automatic Tuning*, ECRAN Project Deliverable, 1996 ("ftp://ftp.dcs.shef.ac.uk/share/ilash/Ecran/DEL.2-6-1.ps.gz").

(Fowler and Dearholt, 1989) R.H. Fowler and D.W. Dearholt, D.W., *Pathfinder networks in information retrieval*, Las Cruces, New Mexico State University. (Memorandums in Computer and Cognitive Science, MCCS-89-147, Computer Science Laboratory).

C (Lesk, 1986) M. Lesk, "Automatic Sense Disambiguation: How to Tell a Pine Cone from an Ice Cream Cone", *Proceedings of 1986 ACM SIGDOC Conference, ACM*, NY, 1986.

(Luk, 1995) A. K. Luk, "Statistical Sense Disambiguation with Relatively Small Corpora using

Dictionary Definitions", *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, 1995.

(Rastier *et al.*, 1994) F. Rastier, M. Cavazza & A. Abeillé, *Sémantique pour l'analyse*, Masson, Paris, 1994.

(Resnik, 1995) Resnik, "Using information content to evaluate semantic similarity in a taxonomy", *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, Montreal, 1995.

(Riloff, 95) E. Riloff, "Little Words Can Make a Big Difference for Text Classification", *Proceedings of the 18th Annual International Conference on research and Development in Information Retrieval (SIGIR¹95*), 1995.

(Salton, 1988) G. Salton, *Automatic Text Processing*, Addison-Wesley, Reading, MA, 1988.

(Slator, 1988a) B.M. Slator, "Constructing contextually organized semantic knowledge bases", *Proceedings of the Third Annual Rocky Mountain Conference on Artificial Intelligence*, Denver, Colorado, 1988.

(Slator 1988b) B.M. Slator, *Lexical semantics and a preference semantics analysis*, Doctoral dissertation. Las Cruces, New Mexico State University. (Memorandums in Computer and Cognitive Science, MCCS-88-143, Computer Science Laboratory), 1988.

(Sta, 1995) J.-D. Sta, "Comportement statistique des termes et acquisition terminologique à partir de corpus", *TAL*, 36(1-2), 1995.

(Véronis and Ide, 1990) J. Véronis & N. M. Ide, "Word Sense Disambiguation with Very Large Neural Networks Extracted from Machine Readable Dictionaries", *Proceedings of the 13th International Conference on Computational Linguistics (COLING'90)*, Helsinki, Finland, 1990.

(Wilks *et al.*, 1995) Y. Wilks, B. Slator & L. Guthrie, *Electric words: dictionaries, computers and meanings*, MIT Press, Cambridge, MA, 1995.

(Wilks & Stevenson, 1997) Y. Wilks & M. Stevenson, "Sense tagging: Semantic tagging with a lexicon", *Proceedings of the SIGLEX Workshop "Tagging Text with Lexical Semantics: What, why and how?"*, Washington D.C., 1997.