Improving User Modelling with Content-Based Techniques

Bernardo Magnini and Carlo Strapparava

ITC-irst, Istituto per la Ricerca Scientifica e Tecnologica, I-38050 Trento, ITALY
email: {magnini,strappa}@irst.itc.it

Abstract. SiteIF is a personal agent for a bilingual news web site that learns user's interests from the requested pages.
In this paper we propose to use a content-based document representation as a starting point to build a model of the user's interests. Documents passed over are processed and relevant senses (disambiguated over WordNet) are extracted and then combined to form a semantic network. A filtering procedure dynamically predicts new documents on the basis of the semantic network.
There are two main advantages of a content-based approach: first, the model predictions, being based on senses rather then words, are more accurate; second, the model is language independent, allowing navigation in multilingual sites. We report the results of a comparative experiment that has been carried out to give a quantitative estimation of these improvements.

Keywords: Content-Based User Modelling, Natural Language Processing, WordNet.

1 Introduction

SiteIF [Stefani and Strapparava, 1998; Strapparava *et al.*, 2000] is a personal agent for a multilingual news web site, that takes into account the user's browsing by "watching over the user's shoulder". It learns user's interests from the requested pages that are analyzed to generate or to update a model of the user. Exploiting this model, the system tries to anticipate which documents in the web site could be interesting for the user.

Many systems (e.g. [Lieberman *et al.*, 1999; Armstrong *et al.*, 1995; Minio and Tasso, 1996] that exploit a user model to propose relevant documents, build a representation of the user's interest which takes into account some properties of words in the document, such as their frequency and their co-occurrence. However, assuming that interest is strictly related to the semantic content of the already seen documents, a purely word based user model is often not accurate enough. The issue is even more important in the Web world, where documents have to do with many different topics and the chance to misinterpret word senses is a real problem.

In this paper we propose to use a content-based document representation as a starting point to build a model of the user's interests. As the user browses the

documents, the system builds the user model as a semantic network whose nodes represent senses (not just words) of the documents requested by the user. Then, the filtering phase takes advantage of the word senses to retrieve new documents with high semantic relevance with respect to the user model.

The use of senses rather than words implies that the resulting user model is not only more accurate but also independent from the language of the documents browsed. This is particularly important for multilingual web sites, that are becoming very common especially in news sites or in electronic commerce domains.

The sense-based approach adopted for the user model component of the SiteIF system makes use of MultiWordNet [Artale *et al.*, 1997], a multilingual lexical database where English and Italian senses are aligned. A technique, recently proposed in [Magnini and Strapparava, 2000], called Word Domain Disambiguation, has been adopted to disambiguate the word senses that define the user interest model.

As for the filtering phase, our approach is supported by experimental evidences (e.g. [Gonzalo *et al.*, 1998a]) that have shown that a content based match can significantly improve the accuracy of the retrieval.

The paper also describes an empirical evaluation of a content-based versus a traditional word-based user modelling. This experiment shows a substantial improvement in performance with respect to the word based approach.

The paper is organized as follows. Section 2 gives a sketch of the kind of documents the system deals with and describes how MultiWordNet and the disambiguation algorithms can be exploited to represent the documents in terms of lexical concepts. Section 3 describes how the user model is built, maintained and used to propose new relevant documents to the user. Section 4 gives an account of the experiment that evaluates and compares a synset-based user model versus a word-based user model. Some final comments about future developments conclude the paper.

2 Content Based Document Representation

The SiteIF web site has been built using a news corpus kindly put at our disposal by AdnKronos, an important Italian news provider. The corpus consists of about 5000 parallel news (i.e. each news has both an Italian and an English version) partitioned by AdnKronos in a number of fixed categories: `culture`, `food`, `holidays`, `medicine`, `fashion`, `motors` and `news`. The average length of the news is about 265 words. Figure 1 shows an example of parallel (English-Italian) news.

The main working hypothesis underlying our approach to user modelling is that a content based analysis of the document can improve the accuracy of the model. There are two crucial questions to address: first, a repository for word senses has to be identified; second, the problem of word sense disambiguation, with respect to the sense repository, has to be solved.

CULTURE: GIOTTO- PAID BY MONKS TO WRITE ANTI-FRANCISCAN POETRY

Rome,10 Jan. -(Adnkronos)- Giotto was 'paid' to attack a faction of the Franciscans, the Spiritual ones, who opposed church decoration in honour of Poverello di Assisi. This has been revealed in the research of an Italian scholar who is a professor at Yale University, Stefano Ugo Baldassarri, who thinks he has solved the mystery of the only known poetry by the famous Tuscan painter: the Giotto verses have in fact always provoked wonder because they seem to be a criticism of the ideals of St. Francis and all the more so since their author was also the man who painted the famous frescoes of the Basilica at Assisi. ...

CULTURA: GIOTTO- PAGATO DA FRATI PER SCRIVERE POESIA ANTI-FRANCESCANA

Roma, 10 gen. -(Adnkronos)- Giotto fu 'pagato' per attaccare una fazione dei Francescani, quella degli Spirituali, che si opponevano alla decorazione delle chiese in onore del Poverello di Assisi. Lo rivela una ricerca di uno studioso italiano docente alla Yale University, Stefano Ugo Baldassarri, che ritiene di aver svelato il mistero dell'unica poesia conosciuta del celebre pittore toscano: i versi giotteschi, infatti, avevano sempre destato meraviglia perché apparivano come una critica agli ideali di San Francesco, tanto piu' mossa proprio dall'autore dei celebri affreschi della Basilica di Assisi.
...

Fig. 1. Sample of parallel news texts.

As for sense repository we have adopted WordNet (version 1.6) [Fellbaum, 1998], a large lexical database for English, freely available, which has received a lot of attention within the computational linguistics community. Nouns, verbs, adjectives and adverbs are organized into synonym sets (i.e. *synsets*), each representing one underlying lexical concept. Synsets are linked by different semantic relations (is-a, part-of, etc...) and organized in hierarchies. The main advantage in using WordNet is that versions in languages other than English are now available (even if none is still complete). In particular in SiteIF we use MultiWordNet, a multilingual extension of the English WordNet. The Italian part of MultiWordNet currently covers about 35,000 lemmas, completely aligned with the English WordNet (i.e. with correspondences to English senses).

The advantages of a synset-based document representation are that: (*i*) each ambiguous term in the document is disambiguated, therefore allowing its correct interpretation and consequently a better precision in the user model construction (e.g. if a user is interested in financial news, a document containing the word "bank" in the context of geography will not be relevant); (*ii*) synonym words belonging to the same synset can contribute to the user model definition. For example both "bank" and "bank building" bring evidences for financial documents, improving the coverage of the document retrieval.

As far as word disambiguation is concerned, we have addressed the problem starting with the hypothesis that many sense distinctions are not relevant for a document representation useful in user modelling. This line is also supported

by several works (see for example [Wilks and Stevenson, 98], [Gonzalo *et al.*, 1998b], [Kilgarriff and Yallop, 2000] and the SENSEVAL initiative) which remark that for many practical purposes (e.g. cross lingual information retrieval) the fine-grained sense distinctions provided by WordNet are not necessary. To reduce the WordNet polysemy, and, as a consequence, the complexity of word sense disambiguation, we have used Word Domain Disambiguation (WDD), a technique proposed in [Magnini and Strapparava, 2000] based on sense clustering through the annotation of the MultiWordNet synsets with domain labels. Section 2.1 gives some details about WDD, while Section 2.2 shows how WDD is applied to represent documents in our context.

2.1 Word Domain Disambiguation

Word Domain Disambiguation is a variant of Word Sense Disambiguation where for each word in a text a *domain* label (among those allowed by the word) has to be chosen instead of a *sense* label. Domain labels, such as Medicine and Architecture, provide a natural way to establish semantic relations among word senses, grouping them into homogeneous clusters. Figure 2 shows an example. The word "book" has seven different senses in WordNet 1.6: three of them can be grouped under the Publishing domain, causing the reduction of the polysemy from 7 to 5 senses.

In MultiWordNet the synsets have been annotated with one or more domain labels ([Magnini and Cavaglià, 2000]). This resource currently covers all the noun synsets, and it is under development for the remaining lexical categories.

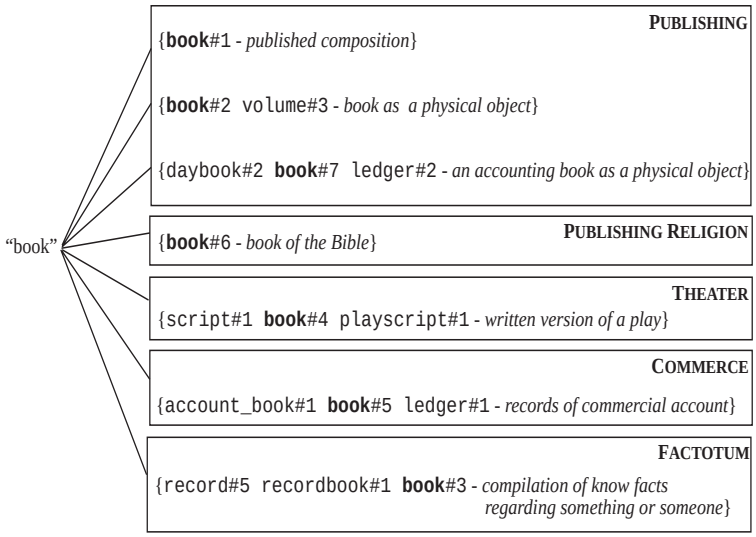

Fig. 2. An example of polysemy reduction

The domain disambiguation algorithm follows two steps. First, each word in the text is considered and for each domain label allowed by that word a score is given. This score is determined by the frequency of the label among the senses of the word. At the second step each word is reconsidered, and the domain label with the highest score is selected as the result of the disambiguation. In [Magnini and Strapparava, 2000] it is reported that this algorithm reaches .83 and .85 accuracy in word domain disambiguation, respectively for Italian and English, on a corpus of parallel news. This result makes WDD appealing for applications where fine-grained sense distinctions are not required, such as document user modelling.

2.2 Document Representations

Each document maintained in the SiteIf site is processed to extract its semantic content. Given that we relay on MultiWordNet, the final representation consists in a list of synsets relevant for a certain document. The text processing is carried out whenever a new document is inserted in the web site, and includes two basic phases: (i) lemmatization and part-of-speech tagging; (ii) synset identification with WDD.

As for lemmatization and part-of-speech tagging we use the LinguistX tools produced by InXight™, which allow to process texts in a number of languages including English and Italian. During this phase the text is first tokenized (i.e. lexical units are identified), then for each word the possible lemmas as well as their morpho-syntactic features are collected. Finally part of speech ambiguities are solved. This is the input for the synset identification phase, which is mainly based on the word domain disambiguation procedure described in Section 2.1. The WDD algorithm, for each word (currently just nouns are considered, due to the limited coverage of the domain annotation), proposes the domain label appropriate for the word context. Then, the word synsets associated to the proposed domain are selected and added to the document representation. As an example, Figure 3 shows a fragment of the Synset Document Representation (SDR) for the document presented in Figure 1. Words are presented with the preferred domain label as well as with the selected synsets. For readability reasons we show the synonyms belonging to each synsets in place of the synset unique identifier used in the actual implementation. In addition, only the English part of the synset is displayed.

3 Sense-Based User Modelling

In SiteIF the user model is implemented as a semantic net whose goal is to represent the contextual information derived from the documents. Previous versions of SiteIF were purely word-based, that is the nodes in the net represented the words and the arcs the word co-occurrences. However the resulting user models were fixed to the precise words of the browsed news. One key issue in automating the retrieval of potentially interesting news was to find document

Word lemma	*Domain label*	*Synsets*
faction	Factotum	{faction-2, sect-2} {cabal-1, faction-1, junta-1, junto-1, camarilla-1}
franciscan	Religion	{Gray_Friar-1, Franciscan-1}
church	Religion	{church-1, Christian_church-1, Christianity-2}
		{church-2, church_building-1} {church_service-1, church-3}
research	Factotum	{research-1} {inquiry-1, enquiry-2, research-2}
scholar	Pedagogy	{scholar-1, scholarly_person-1, student-2}
		{learner-1, scholar-2} {scholar-3}
professor	Pedagogy	{professor-1}
mystery	Literature	{mystery-2, mystery_story-1, whodunit-1}
poetry	Literature	{poetry-1, poesy-1, verse-1} {poetry-2}
painter	Art	{painter-1}
verse	Literature	{poetry-1, poesy-1, verse-1} {verse-2, rhyme-2} {verse-3, verse_line-1}
criticism	Factotum	{criticism-1, unfavorable_judgment-1}
ideal	Factotum	{ideal-1} {ideal-2}
man	Factotum	{man-1, adult_male-1} {man-3} {man-7} {man-8}
author	Literature	{writer-1, author-1}
fresco	Art	{fresco-1} {fresco-2}
basilica	Religion	{basilica-1}

Fig. 3. Synset Document Representation for a fragment of text

representations that are semantically rich and accurate, keeping to a minimal level the participation of the user.

Our hypothesis is that maintaining the same user model network structure but with nodes representing synsets and arcs the co-occurrence of synsets helps to define semantic chains through which the filtering has a better chance to catch documents semantically closer to the topics already touched by the user.

Possibly modelling with synsets or with words will bring to different choices and optimizations in the semantic network representation. However in this paper one purpose is to compare the results of word-based and of synset-based user model, and then we keep uniform the machinery of the user model data structures and algorithms.

3.1 Modelling Phase

In the modelling phase SiteIF considers the browsed documents during a user navigation session. The system uses the document representation of the browsed news. Every synset has a score that is inversely proportional to its frequency over all the news corpus. The score is higher for less frequent synsets, avoiding that very common meanings become too prevailing in the user model. Likewise, in the word-based case we considered a word list document representation, where every word has a score inversely proportional to the word frequency in the news corpus.

The system builds or augments the user model as a semantic net whose nodes are synsets and arcs between nodes are the co-occurrence relation (cooccuring presence in a document) of two synsets. Weights on nodes are incremented by the score of the synsets, while weights on arcs are the mean of the connected

nodes weights[1]. For each browsed news, the weights of the net are periodically reconsidered and possibly lowered, depending on the time passed from the last update. Also no longer useful nodes and arcs may be removed from the net. In this way it is possible to consider changes of the user's interests and to avoid that uninteresting concepts remain in the user model.

Figure 4 sketches the modelling process showing an example of user model augmentation.

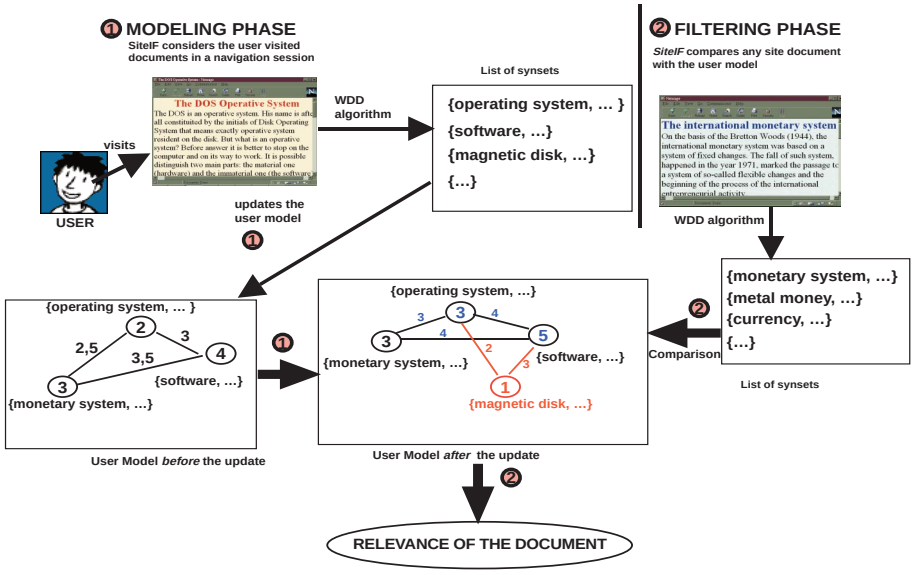

Fig. 4. Modelling and Filtering Processes

3.2 Filtering Phase

During the filtering phase, the system compares any document (i.e. the representation of any documents in terms of synsets) in the site with the user model. A matching module receives as input the internal representation of a document and the current user model and it produces as output a classification of the document (i.e. whether it is worth or not the user's attention). The relevance of any single document is estimated using the Semantic Network Value Technique (see for

[1] As far as the arcs are concerned, an indication of the semantic similarity between the synsets using word sense disambiguation techniques is also present. This is useful to build cohesive chains of synsets in the user model network. (See [Resnik, 1995] for an introduction about word sense disambiguation and semantic similarity issue.) However, we do not take advantage of this information in the evaluation experiment in section 4.

details [Stefani and Strapparava, 1998]). The idea behind the SiteIF algorithm consists of checking, for every concept in the representation of the document, whether the context in which it occurs has been already found in previously visited documents (i.e. already stored in the semantic net). This context is represented by a co-occurrence relationship, i.e. by the couples of terms included in the document which have already co-occurred before in other documents. This information is represented by arcs of the semantic net.

Here below there is the formula used to calculate the relevance of a document using the Semantic Network Value Technique:

$$Relevance(doc) = \sum_{i \in \{syns(doc)\}} w(i) * freq_{doc}(i) + \sum_{i,j \in \{syns(doc)\}} w(i,j) * w(j) * freq_{doc}(j)$$

where $w(i)$ is the weight of synset-node i in the UM network, $w(i,j)$ is the weight of the arc between i and j.

See figure 4 for a summary sketch of the filtering process.

4 Evaluation

We wanted to estimate how much the new version of SiteIF (synset based) actually improves the performances with respect to the previous version of the system (word based). However, setting a comparative test among user models, going beyond a generic user satisfaction is not straightforward. To evaluate whether and how the exploitation of the synset representation improves the accuracy of the semantic network modelling and filtering, we arranged an experiment whose goal was to compare the output of the two systems against the judgements of a human advisor.

We proceeded in the following way. First, a test set of about one hundred English news from the AdnKronos corpus were selected homogeneously with respect to the overall distribution in categories (i.e. `culture`, `motors`, etc...). The test set has been made available as a Web site, and then 12 ITC-irst researchers were asked to browse the site, simulating a user visiting the news site. Users were instructed to select a news, according to their personal interests, to completely read it, and then to select another news, again according to their interests. This process was repeated until ten news were picked out.

After this phase, a human advisor, who was acquainted with the test corpus, was asked to analyze the documents chosen by the users, and to propose new potential interesting documents from the corpus. The advisor was requested to follow the same procedure for each document set: documents were first grouped according to their AdnKronos category, and a new document was searched in the test corpus within that category. If a relevant document was found, it was added to the advisor proposals, otherwise none document for that category is proposed. Eventually, an additional document, outside the categories browsed by the user could be added by the advisor. On average, the advisor proposed 3 documents for a user document set.

At this point we compared the advisor proposals with the results of the two systems. To simulate the advisor behavior (i.e. it is allowed that for a given category no proposal is selected), all the system documents whose relevance was minor of a fixed difference (20%) from the best document, were eliminated. After this selection, on average, the system proposed 10 documents for a user document set.

Standard figures for precision and recall have been calculated considering the matches among the advisor and the systems documents. Precision is the ratio of recommended documents that are relevant, while the recall is the ratio of relevant documents that are recommended. In terms of our experiment we have $precision = \frac{|H \cap S|}{|H|}$ and $recall = \frac{|H \cap S|}{|S|}$, where H is the set of the human advisor proposals and S is the set of the system proposals.

Table 1 shows the result of the evaluation. The first column takes into account the document news, the second only the AdnKronos categories. We can note that precision considerably increases (34%) with the synset-based user model. This confirms the working hypothesis that substituting words with senses both in the modelling and in the filtering phase produces a more accurate output. The main reason, as expected, is that a synset-based retrieval allows to prefer documents with high degree of semantic coherence, which is not guaranteed in case of a word-based retrieval.

	News		Categories	
	Precision	*Recall*	*Precision*	*Recall*
Word-Based UM	0.51	0.21	0.89	0.40
Synset-Based UM	0.85	0.36	0.97	0.43

Table 1. Comparison between word-based UM and synset-based UM

As for recall, it also gains some points (15%), even if it remains quite low. However, this does not a seems a serious drawback for a pure recommender system, where there is no the need to answer an explicit query (as it happens, for instance, in information retrieval systems), but rather the need is for an high quality (i.e. the precision) of the proposals.

5 Conclusions

We have presented a new version of SiteIF, a recommender system for a Web site of multilingual news. Exploiting a content-based document representation, we have described a model of the user's interests based on word senses rather that on simply words. The main advantages of this approach are that semantic accuracy increases and that the model is independent from the language of the news.

To give a quantitative estimation of the improvements induced by a content-based approach, a comparative experiment - sense-based vs. word-based user model - has been carried out, which has showed a significant higher precision in the system recommendations.

There are several areas for future developments. One point is to improve the disambiguation algorithms which are at the basis of the document representation. A promising direction (proposed in [Magnini and Strapparava, 2000]) is to design specific algorithms which consider the synset intersection of parallel news.

A second working direction concerns the possibility to develop clustering algorithms over the senses of the semantic network. For example, once the user model network is built, it could be useful to have the capability to dynamically infer some homogeneous user interest areas. This would allow to arrange in uniform dynamic groups the recommended documents.

References

[Armstrong *et al.*, 1995] R. Armstrong, D. Freitag, T. Joachim, and T. Mitchell. Webwatcher: A learning apprentice for the world wide web. In *Proc. of AAAI Spring Symposium on Information Gathering from Heterogeneous and Distributed Environments*, Stanford, March 1995.

[Artale *et al.*, 1997] A. Artale, B. Magnini, and C. Strapparava. WordNet for italian and its use for lexical discrimination. In *AI*IA 97: Advances in Artificial Intelligence*. Springer Verlag, 1997.

[Fellbaum, 1998] C. Fellbaum. *WordNet. An Electronic Lexical Database*. The MIT Press, 1998.

[Foltz *et al.*, 1998] T. Foltz, W. Kintsch, and T. Landauer. The measurement of textual coherence with latent semantic analysis. *Discourse Processes*, 25(2-3), 1998. Special Issue: Quantitative Approaches to Semantic Knowledge Representations.

[Gonzalo *et al.*, 1998a] J. Gonzalo, F. Verdejio, Chugur, and J. Cigarran. Indexing with wordnet synsets can improve text retrieval. In S. Harabagiu, editor, *Proceeding of the Workshop "Usage of WordNet in Natural Language Processing Systems"*, Montreal, Quebec, Canada, August 1998.

[Gonzalo *et al.*, 1998b] J. Gonzalo, F. Verdejio, C. Peters, and N. Calzolari. Applying eurowordnet to cross-language text retrieval. *Computers and Humanities*, 32(2-3):185–207, 1998.

[Kilgarriff and Yallop, 2000] A. Kilgarriff and C. Yallop. What's in a thesaurus? In *Proceedings of LREC-2000, Second International Conference on Language Resources and Evaluation*, Athens, Greece, June 2000.

[Lieberman *et al.*, 1999] Henry Lieberman, Neil W. Van Dyke, and Adrian S. Vivacqua. Let's browse: A collaborative web browsing agent. In *Proceedings of the 1999 International Conference on Intelligent User Interfaces*, Collaborative Filtering and Collaborative Interfaces, pages 65–68, 1999.

[Magnini and Cavaglià, 2000] B. Magnini and G. Cavaglià. Integrating subject field codes into WordNet. In *Proceedings of LREC-2000, Second International Conference on Language Resources and Evaluation*, Athens, Greece, June 2000.

[Magnini and Strapparava, 2000] B. Magnini and C. Strapparava. Experiments in word domain disambiguation for parallel texts. In *Proc. of SIGLEX Workshop on Word Senses and Multi-linguality*, Hong-Kong, October 2000. held in conjunction with ACL2000.

[Minio and Tasso, 1996] M. Minio and C. Tasso. User modeling for information filtering on internet services: Exploiting an extended version of the UMT shell. In *Proc. of Workshop on User Modeling for Information Filtering on the World Wide Web*, Kailia-Kuna Hawaii, January 1996. held in conjunction with UM'96.

[Resnik, 1995] P. Resnik. Disambiguating noun groupings with respect to WordNet senses. In *Proc. of third workshop on very large corpora*, MIT, Boston, June 1995.

[Stefani and Strapparava, 1998] A. Stefani and C. Strapparava. Personaliziong access to web sites: The siteif project. In *Proc. of second Workshop on Adaptive Hypertext and Hypermedia*, Pittsburgh, June 1998. held in conjunction with HYPERTEXT 98.

[Strapparava *et al.*, 2000] C. Strapparava, B. Magnini, and A. Stefani. Sense-based user modelling for web sites. In *Adaptive Hypermedia and Adaptive Web-Based Systems - Lecture Notes in Computer Science 1892*. Springer Verlag, 2000.

[Vossen, 1998] P. Vossen. Special issue on eurowordnet. *Computers and Humanities*, 32, 1998.

[Wilks and Stevenson, 98] Y. Wilks and M. Stevenson. Word sense disambiguation using optimised combination of knowledge sources. In *Proc. of COLING-ACL'98*, 98.