Building a network of topical relations from a corpus

Olivier Ferret

CEA–LIST/LIC2M
18, route du Panorama
92265 Fontenay-aux-Roses, France
ferreto@zoe.cea.fr

Abstract

Lexical networks such as WordNet are known to have a lack of topical relations although these relations are very useful for tasks such as text summarization or information extraction. In this article, we present a method for automatically building from a large corpus a lexical network whose relations are preferably topical ones. As it does not rely on resources such as dictionaries, this method is based on self-bootstrapping: a network of lexical cooccurrences is first built from a corpus and then, is filtered by using the words of the corpus that are selected by the initial network. We report an evaluation about topic segmentation showing that the results got with the filtered network are the same as the results got with the initial network although the first one is significantly smaller than the second one.

1. Introduction

Since the first versions of WordNet (Miller, 1995) were developed, lexico-semantic networks have given rise to a strong interest in Computational Linguistics. The principles of WordNet were applied to other languages than English, as in the EuroWordNet project (Vossen, 1998), and similar resources were developed, as the MindNet network (Richardson et al., 1998) for instance. However, some researchers (Harabagiu et al., 1999) have pointed out the insufficiencies of the WordNet-like networks. One of these insufficiencies is their lack of topical relations, defined as the "tennis problem" by Roger Chaffin (Fellbaum, 1998). These relations, which are considered by Halliday and Hasan (1976) as non-systematic semantic relations and classified as collocation relations, account for the fact that two words refers to the same topic or to the same situation. Associations like *doctor–hospital*, *burglar–policeman* or *plane–airport* are examples of this kind of relations. Such relations have been introduced in WordNet 2.0 by associating domains to synsets but the number of these links remains still very low: for instance, WordNet 2.1 does not contain any of the three commonsense associations mentioned here above.

The topical knowledge carried by these relations is very useful in information extraction, question-answering (Moldovan and Novischi, 2002) or for automatic summarizing (Harabagiu and Maiorano, 2002) to determine what the characteristic elements of a situation or a topic are and to define what must be extracted from a text according to the current task. Its broad extent makes its manual building very difficult on a large scale and motivated some work for its automatic learning. The topical knowledge got in such a way has two main forms: a set of topical relations that appears as a network of lexical associations or sets of words related to the same topic.

The first approach is represented by work such as the one of Harabagiu and Moldovan (1998) concerning the extraction of topical relations from the glosses associated with the synsets of WordNet. In the context of query expansion, Mandala et al. (1999) proposed another way for performing such kind of extension by integrating co-occurrences and a thesaurus to WordNet. The work of Lin and Hovy (Lin and Hovy, 2000) about the acquisition of topic signatures is a typical example of the second approach. A topic signature is the representation of a topic got from the selection and the weighting of the vocabulary of a set of texts related to this topic. Their acquisition is performed by a supervised learning process. Ferret and Grau (1998) showed that the same kind of representations can be learnt from texts in an unsupervised way. Half way between these two approaches, (Agirre et al., 2001; Agirre and Lopez de Lacalle Lekuona, 2004) built topic signatures that were focused on WordNet's synsets. The knowledge associated to some synsets was used for selecting texts in relation to them and the vocabulary of these texts was then used for building topic signatures. Finally, Magnini and Cavaglia (2000) added topical knowledge to WordNet by annotating its synsets with Subject Field Codes. Their work was extended in (Avancini et al., 2003) by expanding the domains built from this annotation.

Although work based on WordNet for the acquisition of topical knowledge is interesting, its is intrinsically limited by the fact of using WordNet, that is the only resource of this kind at the moment, especially concerning its more elaborated features (such as the glosses associated with the synsets). These features are often exploited by acquisition methods, but they are generally not present in similar lexico-semantic networks. For its part, work that aims at building topic representations must face the fact that topics are expressed in texts in many different ways, which makes difficult to recognize that two textual units refer to the same topic. This last point has led us to choose an approach based on a lexical network, with the potentiality to build later on topic signatures with the advantage of exploiting selected relations from the network. We have also chosen not to rely on significant resources in order to make possible applying our acquisition method to a wide set of languages. Building a co-occurrence network[1] seems to be an interesting solution in this respect. But as the relations underlying

[1] A network of lexical co-occurrences in the present case is a set of co-occurrences collected from a large corpus. These co-occurrences are linked to each other by their words and globally form a network.

co-occurrences are rather heterogeneous, it is necessary to add to this first solution a filtering process for preferably selecting topical co-occurrences.

2. Overview

The starting point of our work is the recording of co-occurrences between words on a large corpus. The space in which this recording is performed, generally a fixed-size window, and the type of the considered words depend on the type of the relations to catch. For building a network of topical relations, it would be necessary to record co-occurrences in text segments that are topically homogeneous and between words that are representative of the topic of these segments. Initially, this principle seems to be circular: segmenting texts into topically homogeneous passages and finding words that are representative of the topics of these segments are two subtasks of topic analysis of texts that have to rely on the kind of knowledge underlying topical relations.

For getting out of this circle, we resort, as it is frequently done in such a situation, to a method based on bootstrapping: first, we build a network of lexical co-occurrences by recording co-occurrences from a corpus according to criteria that favor topical relations (see Section 3) but that are not very selective with regard to other kinds of relations. Then, this initial network is used by a topic analyzer, called TOPICOLL, to delimit in the same corpus text segments that are topically homogeneous and in these segments, to select the words that are representative of their topic (see Sections 4.1 and 4.2). Despite the heterogeneousness of the knowledge it relies on, such a system gets high enough results (Ferret, 2002) for bootstrapping our process. In the third step, a first network of topical co-occurrences is built by recording lexical co-occurrences in the set of segments produced by the previous step (see Section 4.3). Finally, this network is used for selecting the more significant co-occurrences of the initial network from the topical viewpoint (see Section 4.4).

3. Building of the initial co-occurrence network

The corpus we used for building the initial co-occurrence network was made up of 24 months of the French newspaper *Le Monde* taken from 1990 to 1994 in a balanced way. Its size was around 39 million words. The corpus was first pre-processed in order to characterize texts by their significant words from the topical viewpoint. Thus, we retained only the canonical form of plain words, that is, nouns, verbs and adjectives. Co-occurrences were extracted according to the method described in (Church and Hanks, 1990) by moving a window on texts. The parameters of this extraction were set in order to catch more probably topical relations: the window was quite large (20-words wide), took into account the boundaries of texts and ignored the order of co-occurrences. As in (Church and Hanks, 1990), we adopted an evaluation of mutual information as a measure of the cohesion between two words. The finite size of the corpus permits us to normalize this measure according to the maximal mutual information relative to the corpus[2]. After filtering the less significant co-occurrences (co-occurrences whose cohesion < 0.1 and frequency < 10), we obtained a network with 22,749 words and 2,572,589 collocations. As an example, Table 1 gives the most cohesive co-occurrences the word *livre* (book) is part of in this network.

4. Topical filtering

As mentioned in the overview section, the first step of the topical filtering of a co-occurrence network consists in defining, from the corpus used for building this network, a set of textual units having the two following characteristics: first, each unit corresponds to a text segment that is topically homogeneous; second, its words are the words of the segment that are representative of the topic of the segment. These units, that were introduced in (Ferret and Grau, 2000), are called Topical Units.

4.1. Building of Topical Units

The building of Topical Units relies on the use of both a text segmenter and the co-occurrence network to filter. The segmenter delimits text segments that are topically homogeneous while the selection of those of their words that are representative of their topic is based on the network. These two tasks are achieved in our case by a system called TOPICOLL (Ferret, 2002) that exploits a co-occurrence network for performing topic analysis. We will not explain in this paper the way text segmentation works in TOPICOLL as this aspect is not specifically related to our problem. The topic segmentation of texts could also be done by systems that only rely on lexical recurrence such as TextTiling (Hearst, 1994) or C99 (Choi, 2000).

On the other hand, we have taken up the way TOPICOLL selects words from a co-occurrence network. A lot of the words of the network presented in Section 3, especially verbs and adjectives, are found in several different topical contexts because of their generality. As a consequence, this network only contains a restrictive set of co-occurrences that are topically specific. The selection mechanism of TOPICOLL is based on the hypothesis that in a text segment that refers to a particular topical context, the number of relations in the co-occurrence network between the words that are representative of the topic of the segment is higher than the number of relations between the words of the segment that are general ones or related to another topic. More precisely: let w_1 and w_2 be two words of a text segment that are representative of its topic; let w_3 be another word of the same segment but with none specific relation with its topic; let $cooc(w_i)$ be the set of words to which the word w_i is linked in the co-occurrence network. According to the hypothesis underlying the selection of words in TOPICOLL, the two following conditions should be generally valid:

$$card(cooc(w_1) \cap cooc(w_2)) > card(cooc(w_1) \cap cooc(w_3))$$

$$card(cooc(w_1) \cap cooc(w_2)) > card(cooc(w_2) \cap cooc(w_3))$$

The results of TOPICOLL concerning topical segmentation of texts (Ferret, 2002) tend to confirm this hypothesis. In this case, the rule is more restrictive as it is applied with

[2] In theory, cohesion values are between 0 and 1 but actually, the maximal value is equal to 0.4

word	freq.	coh.	word	freq.	coh.	word	freq.	coh.
tyre *(tyre)*	14	0.21	relieur *(bookbinder)*	15	0.19	libraire *(bookseller)*	237	0.19
intranquillité *(non quietness)*	23	0.20	names *(names)*	54	0.19	reliure *(bookbinding)*	62	0.19
tractatus *(tractatus*	13	0.20	novélisation *(novelisation)*	10	0.19	bibliophile *(booklover)*	26	0.18
sterling *(sterling)*	145	0.20	vélin *(vellum)*	10	0.19	presse_parisien *(parisian_newspapers)*	92	0.18
rotativiste *(rotary_printer)*	11	0.19	prix_littéraire *(literary_price)*	17	0.19	reprographie *(reprography)*	12	0.18

Table 1: The most cohesive co-occurrences for the word *livre* (book)

trigrams of topically representative words and not only with bigrams.

In concrete terms, this hypothesis supports the following process in TOPICOLL: a window is moved over the text to be analyzed in order to limit the focus space of the analysis. This latter contains the lemmatized plain words of the text coming from its pre-processing[3]. For each position of this window, we select those words of the co-occurrence network that are linked to at least three words of the window. This process leads to select both words that are in the window and words only coming from the network. The second ones are called inferred words. In order to limit the impact of the "noise" in the initial co-occurrence network, co-occurrences from that network were filtered according to their cohesion value. Co-occurrences whose cohesion was under 0.12 were discarded. This threshold was experimentally set from the results of TOPICOLL concerning text segmentation.

From a more general point of view, we suppose that the focus window is moved in a text segment that was already delimited by a topic segmenter. In TOPICOLL, segmentation and word selection are done simultaneously and the window is moved until the next topic shift is detected. The words that were selected for each position of the window are accumulated and finally, only those that were selected for 75% of the positions of the segment are kept for building the Topical Unit associated to the segment. This condition is taken from (Ferret and Grau, 2000) and aims once again at reducing the number of words selected from non-topical co-occurrences.

4.2. Filtering of Topical Units

Topical Units built in this way are then filtered twice. The first filtering aims at discarding the less significant Topical Units from a topical point of view. In the previous section, we have supposed that a text segment only refers to one topic. But the topical structure of texts is often more complex and topics are sometimes so intermingled that it is impossible to get a linear segmentation of a text. In such a case, the results of a "classical" topical segmenter are not reliable and the Topical Units built from them are not as homogeneous as they should be. We consider that such a situation occurs when no word from the text segment can be selected as a word of the associated Topical Unit, *i.e.* as a representative of the topic of the segment. A lack of selection of words from the text segment can also happen when a passage is weakly marked from the topical viewpoint or when its topic is expressed through a general vocabulary. Moreover, we only keep the Topical Units that contain at least two words from their original segment. As Rastier (Rastier, 1995), we consider that a topic is a pattern of semantic units. Hence, the most reliable way to identify a topic is to identify at least two of its components.

The second filtering is applied to the inferred words of each Topical Unit. Keeping only the words coming from the text segments would be too restrictive as their number is generally small (around three words). But our aim is to filter the initial co-occurrence network and the inferred words added from this network must be as topically close as possible to the words selected from texts. The principle of the filtering of the inferred words is the same as the principle of their selection described in Section 4.1: an inferred word is kept if it is linked, in the co-occurrence network, to at least three text words of the Topical Unit. Moreover, a selective threshold is applied both on the frequency and the cohesion of the co-occurrences supporting these links: only co-occurrences whose frequency $\leq$ 15 and cohesion $\leq$ 0.15 are used.

4.3. Building of a network of topical co-occurrences

After the filtering step, the remaining Topical Units contain the subset of the words from their original text segment that were selected by the means of the initial co-occurrence network and the words of this network that are the most strongly linked to this subset. Thus, each Topical Unit gathers a set of words that are supposed to be strongly coherent from the topical point of view. Next, we record the co-occurrences between these words for all the Topical Units kept after filtering. Hence, we get a large set of co-occurrences likely to be topical in nature, even though a significant number of non-topical co-occurrences remain as the filtering of Topical Units is an unsupervised process. The frequency of a co-occurrence in this case is given by the number of Topical Units in which its two words are simultaneously found. No distinction concerning the origin of the words of the Topical Units is made.

[3] The pre-processing of texts for TOPICOLL is exactly the same as the one applied to texts for building the co-occurrence network of Section 3.

word	freq.	coh.	word	freq.	coh.	word	freq.	coh.
libraire	237	0.19	bouquin	70	0.18	grasset	142	0.17
(bookseller)			*(book)*			*(grasset)*		
reliure	62	0.19	bibliographie	66	0.17	éditeur	919	0.17
(bookbinding)			*(bibliography)*			*(editor*		
presse_parisien	92	0.18	imprimerie	262	0.17	imprimeur	51	0.17
(parisian_newspaper			*(printing_house)*			*(printer)*		
best-seller	74	0.18	imposable	104	0.17	maison_d'édition	109	0.17
(best-seller)			*(taxable)*			*(publisher)*		
librairie	415	0.18	éd	54	0.17	préfacer	53	0.17
(bookshop)			*(eds)*			*(to_preface)*		

Table 2: The most cohesive co-occurrences for the word *livre* (book) after filtering

4.4. Filtering of the initial network

The network of topical co-occurrences resulting from the previous step could be considered as a solution to the problem of building a network of topical co-occurrences. However, the filtering of the initial network appears to be a more reliable method. The building of Topical Units showed that using a co-occurrence network, even with selective criteria, for enriching the topical representation of a text segment brings "noise", even though it also brings interesting words. The network of topical co-occurrences built from Topical Units is a subset of the initial network but it also contains co-occurrences that are not part of it, *i.e.* co-occurrences that were not extracted from the corpus used for setting the initial network or co-occurrences whose frequency in this corpus was too low. Some of these "new" co-occurrences are topical ones but not all of them. As it is difficult to globally estimate what part of them are interesting, we have chosen to let them aside and to focus our attention on the co-occurrences of the topical network that are also present in the initial network.

Thus, we only use the network of topical co-occurrences as a filter for the initial co-occurrence network. Beforehand, the topical network itself is filtered in order to discard co-occurrences whose frequency is too low, that is, co-occurrences that are not stable and therefore, not representative. More precisely, only the co-occurrences whose frequency is higher than 5 are taken. This threshold was experimentally set from the use of the final network (see Section 6). It should be noted that it is lower than the threshold set for the initial network, probably because of the selection of words from texts. Finally, the initial network is filtered by keeping only the subset of its co-occurrences that are also present in the topical network. Their frequency and their cohesion are taken from the initial network, which follows our general viewpoint. Frequencies given by the topical network are potentially interesting, because they are topically more significant, while the difficulty in evaluating the results of the filtering of Topical Units justifies our choice of not using them.

5. Results

We applied the method presented to the co-occurrence network of Section 3. The first step produced 382,208 Topical Units. 59% of them were kept after filtering. The network built from these Topical Units was made of 11,674 words and 2,864,473 co-occurrences. 70% of these co-occurrences were new with regard to the initial network and were discarded. Finally, the filtered network contains 7,223 words and 400,963 co-occurrences. The significant drop in the number of co-occurrences goes together with a more significant drop in the size of the vocabulary of the network. This point could be considered as negative as it could have an influence on the topical covering of the network. However, the evaluation presented in Section 6 does not show it. Although the *Le Monde* newspaper discusses a large set of topics, the co-occurrences of the initial network seem to be reliable only for a subset of these topics. As a consequence, the loss of vocabulary due to the filtering of the network does not seem to have an significant influence from the topical point of view.

Table 2 gives an idea of the filtering of the most cohesive co-occurrences for the word *livre* (book) presented in Table 1. Some non-topical co-occurrences such as those with *intranquillité* (non quietness), *sterling* (*livre sterling* is a compound noun that means sterling pound) or *tyre* (the name of a town) are actually discarded. But the filtering can still be improved: some words without a specific topical link with *livre* (book), such as *imposable* (taxable), are kept whereas others, like *relieur* (bookbinder), are removed although they are part of the same domain as *livre*.

6. Evaluation

Evaluating the topical quality of a co-occurrence network, as any evaluation of a linguistic resource, is a difficult task. Apart from the direct human judgment, two solutions are generally considered: comparing the new resource with a similar resource that was built manually or using the new resource in a system that can be evaluated. As a lexical topical network for French does not exist as far as we know, we adopted the second solution. Moreover, the use of TOPICOLL, which relies on a topical network for performing a topic analysis of texts, makes it clearly an interesting candidate for this kind of evaluation. More precisely, if the filtering process we have previously described is selective enough, it should keep the topical co-occurrences that TOPICOLL is supposed to exploit and the use by TOPICOLL of a co-occurrence network that was topically filtered should not have a negative impact on its results.

Systems	Recall	Precision	F1-measure	Miss	False alarm	P_k
BASE	0.51	0.28	0.36	0.46	0.55	0.50
(Bigi et al., 1998)	0.80	0.75	0.77	unknown	unknown	unknown
TOPICOLL$_1$ (initial network)	0.85	0.79	0.82	0.19	0.20	0.20
TOPICOLL$_2$ (topical filtering)	0.85	0.79	0.82	0.20	0.21	0.21
TOPICOLL$_3$ (frequency filtering)	0.83	0.71	0.77	0.26	0.24	0.25

Table 3: Precision/recall and P_k for the *Le Monde* corpus

The evaluation we present in this article is more specifically dedicated to the task of topic segmentation, which consists in finding the boundaries of a set of concatenated documents. Its conditions were taken from (Ferret, 2002). The evaluation corpus was made up of 49 texts from the *Le Monde* newspaper about 11 topics. These texts were 133 words long on average, which is equivalent to the size of a paragraph. Results in Table 3 are average values computed from 10 different sequences of them. We classically used the recall/precision[4] measure and the probabilistic error metric P_k (Beeferman et al., 1999) for measuring segmentation accuracy [5]. We give for information the results of a baseline procedure, called base in Table 3, that consisted in randomly choosing a fixed number of sentence ends as boundaries. We also give the results of the system described in (Bigi et al., 1998) which was evaluated on the same kind of data as TOPICOLL. These results illustrate the fact that TOPICOLL is comparable to the other systems in the field of text segmentation. TOPICOLL$_1$ is a version of TOPICOLL that relies on the initial co-occurrence network, TOPICOLL$_2$ relies on the network resulting from the topical filtering and TOPICOLL$_3$ relies on a network whose size is close to the size of the filtered network but that results from the application of a threshold on both the frequency and the cohesion of co-occurrences. The network of TOPICOLL$_3$ contains 17,639 words and 196,374 co-occurrences with a threshold set to 0.14 for cohesion and to 14 for frequency. As a threshold concerning co-occurrences' frequency and cohesion is also used for the network of TOPICOLL$_1$ (0.13 for cohesion and 13 for frequency) in order to discard the less significant co-occurrences, the same thresholds were applied to the network of TOPICOLL$_2$ in order to get comparable results. Finally, the network of TOPICOLL$_2$ is made of 7,160 words and 183,074 co-occurrences while the network of TOPICOLL$_1$ contains 18,958 words and 341,549 co-occurrences.

Table 3 clearly shows that the results of TOPICOLL do not decrease when it makes use of the network that was topically filtering although the size of this network is 46% lower than the size of the initial network. Moreover, it also shows that using a network that was filtered according to the frequency and the cohesion of its co-occurrences only has a significant negative impact on TOPICOLL's results even if the size of the network is comparable to the size of the topically filtered network. These results tend to show that the filtering method we have proposed is an effective way of preferably selecting topical co-occurrences.

7. Conclusion and future work

The work we have presented in this article aims at automatically building a lexical network that mainly relies on topical relations. We have chosen to solve this problem by filtering a network of lexical co-occurrences and proposed a method based on bootstrapping for doing it. Its indirect evaluation through the use of its result by a topic segmentation system has shown the interest of this approach. However, this evaluation must be carried on further, especially by comparing the lexical networks that are produced in such a way with similar networks that were built or at least controlled manually. The topical relations extracted from the definitions associated to the synsets of WordNet (Harabagiu et al., 1999) should be an interesting resource in this respect when it will be available. Finally, another solution we think about for evaluating our lexical network is to facilitate human judgment by structuring this network into representations such as topic signatures, which makes easier a global judgment.

8. References

Eneko Agirre and Oier Lopez de Lacalle Lekuona. 2004. Publicly available topic signatures for all wordnet nominal senses. In *4rd International Conference on Languages Resources and Evaluation (LREC 2004)*, Lisbon, Portugal.

Eneko Agirre, Olatz Ansa, David Martinez, and Eduard Hovy. 2001. Enriching wordnet concepts with topic signatures. In *SIGLEX workshop on "WordNet and Other Lexical Resources: Applications, Extensions and Customizations" in conjunction with NAACL'01.*

Henri Avancini, Alberto Lavelli, Bernardo Magnini, Fabrizio Sebastiani, and Roberto Zanoli. 2003. Expanding domain-specific lexicons by term categorization. In *18th ACM Symposium on Applied Computing (SAC-03)*, pages 793–797, Melbourne, US. ACM Press, New York, US.

Doug Beeferman, Adam Berger, and John Lafferty. 1999. Statistical models for text segmentation. *Machine Learning*, 34(1):177–210.

[4] Precision is given by $\frac{N_t}{N_b}$ and recall by $\frac{N_t}{D}$, with D the number of document breaks, N_b the number of boundaries found by TOPICOLL and N_t the number of boundaries that are document breaks (the boundary should not farther than 9 plain words from the document break).

[5] P_k evaluates the probability that a randomly chosen pair of words, separated by k words, is wrongly classified, *i.e.* they are found in the same segment by TOPICOLL while they are actually in different ones (miss of a document break) or they are found in different segments by TOPICOLL while they are actually in the same one (false alarm).

Brigitte Bigi, Renato de Mori, Marc El-Bèze, and Thierry Spriet. 1998. Detecting topic shifts using a cache memory. In 5^{th} *International Conference on Spoken Language Processing*, pages 2331–2334, Sydney, Australia.

Freddy Y. Y. Choi. 2000. Advances in domain independent linear text segmentation. In *NAACL'00*, pages 26–33.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

Christiane Fellbaum, editor. 1998. *WordNet:An Electronic Lexical Database*. The MIT Press.

Olivier Ferret and Brigitte Grau. 1998. A thematic segmentation procedure for extracting semantic domains from texts. In *ECAI'98*, pages 155–159, Brighton, UK.

Olivier Ferret and Brigitte Grau. 2000. A topic segmentation of texts based on semantic domains. In *ECAI 2000*, pages 426–430, Berlin, Germany.

Olivier Ferret. 2002. Using collocations for topic segmentation and link detection. In *COLING 2002*, pages 260–266, Tapei, Taiwan.

M. A. K. Halliday and R. Hasan. 1976. *Cohesion in English*. Longman, London.

Sanda M. Harabagiu and Steven J. Maiorano. 2002. Multi-document summarization with GISTexter. In *Third International Conference on Language Ressources and Evaluation (LREC 2002)*, Las Palmas, Canary Islands, Spain.

Sanda M. Harabagiu and Dan I. Moldovan. 1998. Knowledge processing on an extended wordnet. In Christiane Fellbaum, editor, *WordNet – An Electronic Lexical Database*, pages 379–405. MIT Press, Cambridge, Massachusetts and London, England.

Sanda M. Harabagiu, George A. Miller, and Dan I. Moldovan. 1999. Wordnet 2 - a morphologically and semantically enhanced resource. In *ACL-SIGLEX99: Standardizing Lexical Resources*, pages 1–8, Maryland.

Marti A. Hearst. 1994. Multi-paragraph segmentation of expository text. In *32th Annual Meeting of the Association for Computational Linguistics*, pages 9–16, Las Cruces, New Mexico, USA.

Chin-Yew Lin and Eduard Hovy. 2000. The automated acquisition of topic signatures for text summarization. In *COLING 2000*.

Bernardo Magnini and Gabriela Cavaglia. 2000. Integrating subject field codes into wordnet. In M. Gavrilidou, G. Crayannis, S. Markantonatu, S. Piperidis, and G. Stainhaouer, editors, *Second International Conference on Language Resources and Evaluation (LREC 2000)*, pages 1413–1418, Athens, Greece.

Rila Mandala, Takenobu Tokunaga, and Hozumi Tanaka. 1999. Complementing wordnet with roget's and corpus-based thesauri for information retrieval. In *EACL 99*, Bergen.

George A. Miller. 1995. Wordnet: A lexical database. *Communications of the ACM*, 38(11):39–41.

Dan I. Moldovan and Adrian Novischi. 2002. Lexical chains for question answering. In *COLING 2002*, Tapei, Taiwan.

François Rastier, 1995. *L'analyse thématique des données textuelles*, chapter La sémantique des thèmes ou le voyage sentimental. Didier, Paris.

Stephen D. Richardson, William B. Dolan, and Lucy Vanderwende. 1998. Mindnet: acquiring and structuring semantic information from text. In *ACL/COLING'98*.

Piek Vossen, editor. 1998. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publisher, Dordrecht.