Using Domain Information for Word Sense Disambiguation

Bernardo Magnini, Carlo Strapparava, Giovanni Pezzulo and **Alfio Gliozzo**
ITC-irst, Istituto per la Ricerca Scientifica e Tecnologica, I-38050 Trento, ITALY
email: {magnini, strappa, pezzulo, gliozzo}@itc.it

Abstract

The major goal in ITC-irst's participation at SENSEVAL-2 was to test the role of domain information in word sense disambiguation. The underlying working hypothesis is that domain labels, such as MEDICINE, ARCHITECTURE and SPORT provide a natural way to establish semantic relations among word senses, which can be profitably used during the disambiguation process. For each task in which we participated (i.e. English all words, English 'lexical sample' and Italian 'lexical sample') a different mix of knowledge based and statistical techniques were implemented.

1 Introduction

Current investigation in Word Sense Disambiguation (WSD) at ITC-irst focuses on the role of *domain information*. The hypothesis is that domain labels (such as MEDICINE, ARCHITECTURE and SPORT) provide a natural and powerful way to establish semantic relations among word senses, which can be profitably used during the disambiguation process. In particular, domains constitute a fundamental feature of text coherence, such that word senses occurring in a coherent portion of text tend to maximize domain similarity. The importance of domain information in WSD has been remarked in several works, including (Gonzalo et al., 1998) and (Buitelaar and Sacaleanu, 2001). In (Magnini and Strapparava, 2000) we introduced "Word Domain Disambiguation" (WDD) as a variant of WSD 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. We also argued that WDD can be applied to disambiguation tasks that do not require fine grained sense distinctions, such as information retrieval and content-based user modeling. For SENSEVAL-2 the goal was to evaluate the role of domain information in WSD: no other syntactic or semantic information has been used (e.g. semantic relations in WORDNET) except domain labels. Three systems have been implemented, integrating knowledge-based and statistical techniques, for the three tasks we participated in, i.e. English 'all words', English 'lexical sample' and Italian 'lexical sample'. The main lexical resource for domains is "WordNet Domains", an extension of English Wordnet 1.6 (Fellbaum, 1998) developed at ITC-irst, where synsets have been annotated with domain information.

2 WordNet Domains

The basic lexical resource we used in SENSEVAL-2 is "WordNet Domains", an extension of WORDNET 1.6 where each synset has been annotated with at least one domain label, selected from a set of about two hundred labels hierarchically organized (see (Magnini and Cavaglià, 2000) for the annotation methodology and for the evaluation of the resource). The information from the domains that we added is complementary to what is already in WORDNET. First of all a domain may include synsets of different syntactic categories: for instance MEDICINE groups together senses from Nouns, such as `doctor#1` and `hospital#1`, and from Verbs such as `operate#7`. Second, a domain may include senses from different WORDNET sub-hierarchies (i.e. deriving from different "unique beginners" or from different "lexicographer files"). For example, SPORT contains senses such as `athlete#1`, deriving from `life_form#1`, `game_equipment#1` from `physical_object#1`, `sport#1` from `act#2`, and `playing_field#1` from `location#1`. Finally, domains may group senses of the same word into homogeneous clusters, with the side

effect of reducing word polysemy in WORDNET. Table 1 shows an example. The word "bank" has ten different senses in WORDNET 1.6: three of them (i.e. sense 1, 3 and 6) can be grouped under the ECONOMY domain, while sense 2 and 7 both belong to GEOGRAPHY and GEOLOGY, causing the reduction of the polysemy from 10 to 7 senses. For the purposes of SENSEVAL-2 we have considered 41 disjoint labels which allow a good level of abstraction without loosing relevant information (i.e. in the experiments we have used SPORT in place of VOLLEY or BASKETBALL, which are subsumed by SPORT).

Sense	*Synset & Gloss*	*Domains*	*Semcor occurr.*
#1	depository financial institution, bank, banking concern, banking company (a financial institution...)	ECONOMY	20
#2	bank (sloping land...)	GEOGRAPHY, GEOLOGY	14
#3	bank (a supply or stock held in reserve...)	ECONOMY	-
#4	bank, bank building (a building...)	ARCHITECTURE, ECONOMY	-
#5	bank (an arrangement of similar objects...)	FACTOTUM	1
#6	savings bank, coin bank, money box, bank (a container...)	ECONOMY	-
#7	bank (a long ridge or pile...)	GEOGRAPHY, GEOLOGY	2
#8	bank (the funds held by a gambling house...)	ECONOMY, PLAY	
#9	bank, cant, camber (a slope in the turn of a road...)	ARCHITECTURE	-
#10	bank (a flight maneuver...)	TRANSPORT	-

Table 1: WORDNET senses, domains and occurrences in Semcor for the word "bank"

Two mapping procedures have been implemented for SENSEVAL-2 in order to use domain information. For the English tasks a mapping from WORDNET 1.6 to the WORDNET 1.7 pre-release made available to participants; for the Italian task a mapping from WORDNET 1.6 to WORDNET 1.5, because the interlingual index of EuroWordNet (Vossen, 1998) is in that version. The mapping to WORDNET 1.7 is based on a set of heuristics (e.g. correspondences between synonyms, glosses and hypernyms) which discover corresponding synset pairs. Then, an inheritance algorithm is applied to WORDNET 1.7 in order to fill unassigned synsets with domain labels. As far as the Italian wordnet is concerned the same procedure used for the WORDNET 1.7 mapping has been applied to WORDNET 1.5, resulting in the annotation of the Interlingual Index. Then the equivalence links (we excluded eq_hyperonym and eq_hyphonym) from the ILI to the Italian synsets were used to bring the domain information to Italian words.

There was no time for a complete evaluation of the quality of the mapping procedures.

3 Algorithms

The starting point in the algorithm design was the previous work in word domain disambiguation reported in (Magnini and Strapparava, 2000). One drawback of that approach is that, for rather long texts, it does not consider domain variations. To overcome this problem we have introduced *contexts* within which domains are calculated. A second direction of work has been the acquisition of domain information from annotated texts (i.e. Semcor and the training data). The following sections presents details of the disambiguation procedures implemented for SENSEVAL-2.

3.1 Linguistic Processing

XML files made available by the task organizers have been processed with an XML parser. As for lemmatization and part-of-speech tagging the Tree Tagger, developed at the University of Stuttgart (Schmid, 1994) has been used, both for English and Italian. The WordNet morphological analyser has also been used in order to resolve ambiguities and lemmatization mistakes. After this process texts are represented as vectors of triples: word lemma, WORDNET part of speech and position in the text.

3.2 Scoring Domains for a Lemma

The basic procedure in domain driven disambiguation is a function that, given a lemma L,

associates a score to each domain defined for that lemma in Wordnet Domains. Such a score is the relative frequency of the domain in L, computed on the basis of the occurrences of the synsets of L in Semcor. Semcor occurrences for synsets with multiple domain annotations are repeated for each domain (e.g. if a synset has 2 occurrences and 2 labels it is counted as having 4 occurrences), while synsets with 0 occurrences are counted as 0.5. As an example, consider the lemma "bank" in Table 1. According to our scoring method, it has 57 total occurrences in Semcor. The GEOLOGY domain collects contributions from senses 2 and 7, for a total of 16 occurrences in Semcor, which corresponds to a frequency .28 (i.e. $fq[D_{Geology}](bank) = 0.28$).

3.3 Domain Vectors

The data structure that collects domain information is called a *Domain Vector* (DV). Intuitively a DV represents the domains that are relevant for a certain lemma (or word sense) in a certain context. We have considered three kinds of DV's: a DV for a lemma L within a context C (DV_L^C), for the case of test data; a DV for a synset S of a lemma L within a context C (DV_S^C), for the case of training data; and a DV for a synset S of a lemma L in WORDNET (DV_S), which is used when no training data are available.

DV for a lemma in context (DV_L^C). Given a set of domains $D_1 \ldots D_n$, a DV for a lemma L in a position K within a text represents the relevance of those domains for that lemma, i.e. each component $DV_L[i]$ gives the degree of relevance of the domain D_i for the lemma L. Given a context of $\pm C$ words before and after the lemma L in the position K, each component of the domain vector is defined with the following formula:

$$DV_L^C[i] = \sum_{k=-C}^{+C} Fq[D_i](L_k) * gauss$$

where *gauss* is the normal distribution centered on the position K. In the current algorithms C is set to 50 because our experiments with Semcor showed that the precision decreases below that thresold.

Intuitively, the above formula takes into account the contribution of the lemmas in the context C to the sense of the target lemma L. In addition a DV actually selects a set of relevant domains rather than just one domain.

DV for a synset in context (DV_S^C) In case a training corpus is available where lemmas are annotated with the correct sense, Domain Vectors are computed with the formula above. Instead of considering a lemma in a position K within a text, we have a sense for that lemma (i.e. a synset). DV_S^C represents a "typical" vector for a sense S of a lemma L.

DV for a synset without context (DV_S) When a training corpus is not available (as for the 'all words' task), a simpler way to build a DV for a certain synset is to compute it with respect to WordNet Domains. Given a synset S in WordNet Domains, the domain vector DV_S is a vector that has 1's in the position of its domain(s) and 0's otherwise. A more accurate DV could be obtained by considering contextual information such as the synset gloss.

3.4 Comparing Domain Vectors

To disambiguate a lemma L (i.e. the target lemma) in a text, first its DV_L^C is computed. The next step consists of comparing the DV of the target lemma L with the domain vectors for each sense of L derived either from the training set, when available, or from WordNet Domains, when training data are not available. The sense vector DV_S which maximizes the similarity is selected as the appropriate sense of L in that text. The similarity between two DV's is calculated with the standard scalar product: $DV_1 \cdot DV_2 = \sum_i DV_1[i] * DV_1[i]$.

4 Results and Discussion

Table 2 presents the results, in terms of precision and recall, obtained at the SENSEVAL-2 initiative for the three tasks in which we participated.

Task	*Precision*	*Recall*
English All Words (fine g.)	.748	.357
English All Words (coarse g.)	.748	.357
English Lexical Sample (fine g.)	.665	.249
English Lexical Sample (coarse g.)	.720	.269
Italian Lexical Sample (fine g.)	.375	.371

Table 2: Final results of ITC-irst systems at SENSEVAL-2

4.1 English 'All Words'

The 'all words' task seems to benefit from the domain approach. One reason for this is that texts are enough long to provide an accurate context (as mentioned in section 3.3, we used a window of 100 content words around the target word) within which domains are coherent. The rather low degree of recall reflects the fact that few words in a text carry relevant domain information. Most of the words actually behave such as a "factotum" (see (Magnini and Cavaglià, 2000) for a preliminary discussion on this problem) that can equally occur in almost every domain. Some words lie outside the domain approach and their senses could be captured with the integration of local (e.g. syntactic) information.

4.2 English 'Lexical Sample'

From the point of view of domain driven disambiguation, the 'lexical sample' task was inherently more difficult than the 'all words' task for two reasons. First the context provided for disambiguation was generally shorter than the 100 words we used to build a semantic vector. Second, the high number of "factotum" words to be disambiguated resulted in a recall even lower (i.e. about 0.24) than for the 'all words' task. The improvement of performance from the fine grained to the coarse grained evaluation seems to confirm that, at least to some degree, domain clustering corresponds to the sense grouping created by the task organizers.

4.3 Italian 'Lexical Sample'

The low results obtained for the Italian 'lexical sample' task may have several causes. First of all, the absence of a training set and the absence of any tagged text for Italian forced us to use a similarity function (see 3.4) trained to an English corpus. This was possible because we maintained the mappings between the English and the Italian wordnets. However, these multiple mappings (i.e. from WORDNET1.6 to WORDNET1.5 and then to the Italian synsets through the equivalence links) are another source of possible errors, especially concerning the domain information associated with Italian synsets.

5 Conclusions

We have described an approach to word sense disambiguation based on domain information. The underlying assumption is that domains constitute a fundamental feature of text coherence. As a consequence, word senses occurring in a coherent portion of text tend to maximize domain similarity. Three systems have been implemented, integrating knowledge-based and statistical techniques, for the three task we participated in. As for lexical resources, the systems make use of WordNet Domains, an extension of English Wordnet 1.6, where synsets have been annotated with domain information. The disambiguation algorithm is based on domain vectors that collect contextual information with respect to the target word. At this moment only domain information is used in our system. A promising research direction is the use of local information (e.g. syntax) to capture word behaviors that lie outside the domain approach.

References

P. Buitelaar and B. Sacaleanu. 2001. Ranking and selecting synsets by domain relevance. In *Proc. of NAACL Workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations*, Pittsburgh, PA, June.

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

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

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

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

H. Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing.*

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