Experiments in Word Domain Disambiguation for Parallel Texts

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

This paper describes some preliminary results about Word Domain Disambiguation, a variant of Word Sense Disambiguation where words in a text are tagged with a *domain* label in place of a *sense* label. The English WordNet and its aligned Italian version, MultiWordNet, both augmented with domain labels, are used as the main information repositories. A baseline algorithm for Word Domain Disambiguation is presented and then compared with a *mutual help* disambiguation strategy, which takes advantages of the shared senses of parallel bilingual texts.

1 Introduction

This work describes some preliminary results about Word Domain Disambiguation (WDD), a variant of Word Sense Disambiguation (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. Domain labels, such as Medicine and Architecture, provide a natural way to establish semantic relations among word senses, grouping them into homogeneous clusters. A relevant consequence of the application of domain clustering over the WordNet senses is the reduction of the word polysemy (i.e. the number of domains for a word is generally lower than the number of senses for that word).

We wanted to investigate the hypothesis that the polysemy reduction caused by domain clustering can profitably help the word domain disambiguation process. A preliminary experiment has been set up with two goals: first, providing experimental evidences that a frequency based WDD algorithm can outperform a WSD baseline algorithm; second, exploring WDD in the context of parallel, not aligned, text disambiguation.

The English WordNet and the Italian aligned version MultiWordNet, both augmented with domain labels, are used as the main information repositories. A baseline algorithm for Word Domain Disambiguation is presented and then compared with a *mutual help* disambiguation strategy, which makes use of the shared senses of parallel bilingual texts.

Several works in the literature have remarked that for many practical purposes the fine-grained sense distinctions provided by WordNet are not necessary (see for example [Wilks and Stevenson, 98], [Gonzalo *et al.*, 1998], [Kilgarriff and Yallop, 2000] and the SENSEVAL initiative) and make it hard word sense disambiguation. Two related works are also [Buitelaar, 1998] and [Buitelaar, 2000], where the reduction of the WordNet polysemy is obtained on the basis of regular polysemy relations. Our approach is based on sense clusters derived by domain proximity, which in some case may overlap with regular polysemy derived clusters (e.g. both "book" as composition and "book" as physical object belong to Publishing), but in many cases may not (e.g. "lamb" as animal belongs to Zoology, while "lamb" as meat belongs to Food). Following this line we propose Word Domain Disambiguation as a practical alternative for applications that do not require fine grained sense distinctions.

The paper is organized as follows. Section 2 introduces domain labels, their organization and the extensions to WordNet. Section 3 discusses Word Domain Disambiguation and presents the algorithms used in the experiment. Section 4 gives the experimental setting. Results are discussed in Section 5.

2 WordNet and Subject Field Codes

In this work we will make use of an augmented WordNet, whose synsets have been annotated with one or more subject field codes. This resource, discussed in [Magnini and Cavaglià, 2000],

currently covers all the noun synsets of WordNet 1.6 [Miller, 1990], and it is under development for the remaining lexical categories.

Subject Field Codes (SFC) group together words relevant for a specific domain. The best approximation of SFCs are the field labels used in dictionaries (e.g. Medicine, Architecture), even if their use is restricted to word usages belonging to specific terminological domains. In WordNet, too, SFCs seem to be used occasionally and without a consistent design.

Information brought by SFCs is complementary to what is already in WordNet. First of all a SFC may include synsets of different syntactic categories: for instance Medicine[1] groups together senses from Nouns, such as `doctor#1` and `hospital#1`, and from Verbs such as `operate#7`. Second, a SFC may also contain senses from different WordNet sub-hierarchies (i.e. deriving from different "unique beginners" or from different "lexicographer files"). For example, the Sport SFC 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`.

We have organized about 250 SFCs in a hierarchy, where each level is made up of codes of the same degree of specificity: for example, the second level includes SFCs such as Botany, Linguistics, History, Sport and Religion, while at the third level we can find specializations such as American_history, Grammar, Phonetics and Tennis.

A problem arises for synsets that do not belong to a specific SFC, but rather can appear in almost all of them. For this reason, a Factotum SFC has been created which basically includes two types of synsets:

- *Generic* synsets, which are hard to classify in a particular SFC, are generally placed high in the WordNet hierarchy and are related senses of highly polysemous words. For example:

  `man#1` an adult male person (as opposed to a woman)
  `man#3` the generic use of the word to refer to any human being
  `date#1` day of the month
  `date#3` appointment, engagement

- *Stop Senses* synsets which appear frequently in different contexts, such as numbers, week days, colors, etc. These synsets usually belong to non polysemous words and they behave much as *stop words*, because they do not significantly contribute to the overall meaning of a text.

A single domain label may group together more than one word sense, resulting in a reduction of the polysemy. Figure 1 shows an example. The word "book" has seven different senses in WordNet 1.6: three of them are grouped under the Publishing domain, causing the reduction of the polysemy from 7 to 5 senses.

3 Word Domain Disambiguation

In this section we present two baseline algorithms for word domain disambiguation and we propose some variants of them to deal with WDD in the context of parallel texts.

3.1 Baseline algorithms

To decide a proper baseline for Word Domain Disambiguation we wanted to be sure that it was applicable to both the languages (i.e. English and Italian) used in the experiment. This caused the exclusion of a selection based on the domain frequency computed as a function of the frequency of the WordNet senses, because we did not have a frequency estimation for Italian senses. We adopted two alternative frequency measures, based respectively on the intra text frequency and the intra word frequency of a domain label. Both of them are computed with a two-stage disambiguation process, structurally similar to the algorithm used in [Voorhees, 1998].

Baseline 1: Intra text domain frequency. The baseline algorithm follows two steps. First, all the words in the text are considered and for each domain label allowed by the word the label score is incremented by one. At the second step each word is reconsidered, and the domain label (or labels, depending on how many best solutions are requested) with the highest score is selected as the result of the disambiguation.

Baseline 2: Intra word domain frequency. In this version of the baseline algorithm, step 1 is modified in that each domain label allowed by the word is incremented by the frequency of the label among the senses of that word. For instance,

[1] Throughout the paper subject field codes are indicated with this Typeface while word senses are reported with this `typeface#1`, with their corresponding numbering in WordNet 1.6. Moreover, we use *subject field code*, *domain label* and *semantic field* with the same meaning.

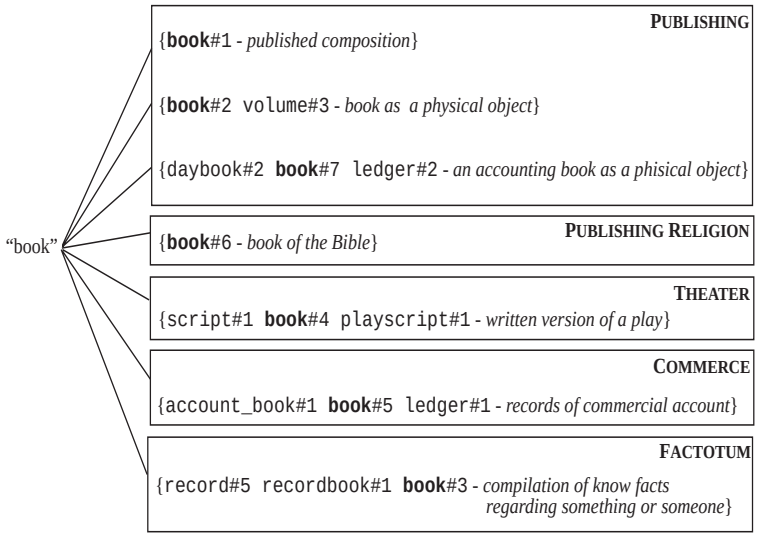

Figure 1: An example of polysemy reduction

if "book" is the word (see Figure 1), PUBLISHING will receive .42 (i.e. three senses out of seven belong to PUBLISHING), while the others domain labels will receive .14 each.

3.1.1 The "factotum" effect

As we mentioned in Section 2, a FACTOTUM label is used to mark WORDNET senses that do not belong to a specific domain, but rather are highly widespread across texts of different domains. A consequence is that very often, at the end of step 1 of the disambiguation algorithm, FACTOTUM outperforms the other domains, this way affecting the selection carried out at step 2 (i.e. in case of ambiguity FACTOTUM is often preferred).

For the purposes of the experiment described in the next sections the FACTOTUM problem has been resolved with a slight modification at step 2 of the baseline algorithm: when FACTOTUM is the best selection for a word, also the second available choice is considered as a result of the disambiguation process.

3.2 Extensions for parallel texts

We started with the following working hypothesis. Using aligned wordnets to disambiguate parallel texts allows us to calculate the intersection among the synsets accessible from an English text through the English WORDNET and the synsets accessible from the parallel Italian text through the Italian WORDNET. It would seem reasonable that the synset intersection maximizes the number of significant synsets for the two texts, and at the same time tends to exclude synsets whose meaning is not pertinent to the content of the text.

Let us try to make the point clearer with an example. Suppose we find in an English text the word "bank" and in the Italian parallel text the word "banca", which we do not know being the translation of "bank", because we do not have word alignments. For "bank" we get ten senses from WORDNET 1.6 (reported in Figure 2), while for "banca" we get two senses from MULTIWORDNET (reported in Figure 2). As the two wordnets are aligned (i.e. they share synset offsets), the intersection can be straightforwardly determined. In this case it includes 06227059, corresponding to `bank#1` and `banca#1`, and 02247680, corresponding to `bank#4` and `banca#2`, which both pertain to the BANKING domain, and excludes, among the others, `bank#2`, which happens to be an homonym sense in English but not in Italian.

Incidentally, if "istituto di credito" were not in the synset `06227059` (e.g. because of the incompleteness of the Italian WORDNET) and it were the only word present in the Italian news to denotate the `bank#1` sense, the synset intersection would have been empty.

As far as disambiguation is concerned it seems a reasonable hypothesis that the synset intersection could bring constraints on the sense selection for a word (i.e. it is highly probable that the correct choice belongs to the intersection). Following this line we have elaborated a *mutual help* disambiguation strategy where the synset intersection can be accessed to help the disambiguation process of both English and Italian texts.

In addition to the synset intersection, we wanted to consider the intersection of domain labels, that is domains that are shared among the

Bank (from WordNet 1.6)

1. {06227059} depository financial institution, bank, banking concern, banking company -- (a financial institution that accepts deposits and channels the money into lending activities;)
2. {06800223} bank -- (sloping land (especially the slope beside a body of water))
3. {09626760} bank -- (a supply or stock held in reserve especially for future use (especially in emergencies))
4. {02247680} bank, bank building -- (a building in which commercial banking is transacted;)
5. {06250735} bank -- (an arrangement of similar objects in a row or in tiers;)
6. {03277560} savings bank, coin bank, money box, bank -- (a container (usually with a slot in the top) for keeping money at home;)
7. {06739355} bank -- (a long ridge or pile; "a huge bank of earth")
8. {09616845} bank -- (the funds held by a gambling house or the dealer in some gambling games;)
9. {06800468} bank, cant, camber -- (a slope in the turn of a road or track;)
10. {00109955} bank -- (a flight maneuver; aircraft tips laterally about its longitudinal axis (especially in turning))

Banca (from MultiWordnet)

1. {06227059} istituto_di_credito cassa banco banca
2. {02247680} banca

Figure 2: An example of sysnet intersection in MULTIWORDNET

senses of the parallel texts. In the example above the domain intersection would include just one label (i.e. BANKING), in place of the two synsets of the synset intersection. The hypothesis is that domain intersection could reduce problems due to possible misalignments among the synsets of the two wordnets.

Two mutual help algorithms have been implemented, *weak mutual help* and *strong mutual help*, which are described in the following.

***Weak* Mutual help.** In this version of the mutual help algorithm, step 1 of the baseline is modified in that, if the domain label is found in the synset or domain intersection, a bonus is assigned to that label, doubling its score. In case of empty intersection (i.e. either no synset or no domain is shared by the two texts) this algorithm guarantees the same performances of the baseline.

***Strong* Mutual help.** In the strong version of the mutual help strategy, step 1 of the baseline is modified in that the domain label is scored if and only if it is found in the synset or domain intersection. While this algorithm does not guarantee the baseline performance (because the intersection may not contain all the correct synsets or domains), the precision score will give us indications about the quality of the synset intersection.

4 Experimental Setting

The goal of the experiment is to establish some reference figures for Word Domain Disambiguation. Only nouns have been considered, mostly because the coverage of both MULTIWORDNET and of the domain mapping for verbs is far from being complete.

	Lemmas	*Senses*	*Mean Polysemy*
WN 1.6	94474	116317	1.23
Ital WN	19104	25226	1.32
DISC	56134	118029	2.10

Table 1: Overview of the used resources (Noun part of speech)

4.1 Lexical resources

Besides the English WORDNET 1.6 we used MULTIWORDNET [Artale *et al.*, 1997; Magnini and Strapparava, 1997], an Italian version of the English WORDNET. It is based on the assumption that a large part of the conceptual relations defined for the English language can be shared with Italian. From an architectural point of view, MULTIWORDNET implements an extension of the WORDNET lexical matrix to a "multilingual lexi-

Mean Values for Nouns		*Italian News*	*English News*
Lexical Coverage	WN 1.6	-	98%
	ItalWN	93%	-
	Disc	100%	-
# Synsets	English	-	155.21
	Italian	111.38	-
	Intersection	35.48	

Table 2: Mean lexical coverage and synset amount for AdnKronos news

Mean Values for Nouns		*Italian News*	*English News*
Sense Polysemy	WN 1.6	-	4.37
	ItalWN	3.22	-
	Disc	6.82	-
Domain Polysemy	English	-	3.58
	Italian	2.68	-

Table 3: Mean sense and domain polysemy for AdnKronos news

cal matrix" through the addition of a third dimension relative to the language. MultiWordNet currently includes about 30,000 lemmas.

As a matter of comparison, in particular to estimate the lack of coverage of MultiWordNet, we consider some data from the Italian dictionary "DISC" [Sabatini and Coletti, 1997], a large size monolingual dictionary, available both as printed version and as CD-ROM.

Table 1 shows some general figures (only for nouns) about the number of lemmas, the number of senses and the average polysemy for the three lexical resources considered.

4.2 Parallel Texts

Experiments have been carried out on a news corpus kindly placed at our disposal by AdnKronos, an important Italian news provider. The corpus consists of 168 parallel news (i.e. each news has both an Italian and an English version) concerning various topics (e.g. politics, economy, medicine, food, motors, fashion, culture, holidays). The average length of the news is about 265 words.

Table 2 reports the average lexical coverage (i.e. percent of lemmas found in the news corpus) for WordNet 1.6, MultiWordNet and the Disc dictionary. A practically zero variance among the various news is exhibited. We observe a full coverage for the Disc dictionary; in addition, the incompleteness of MultiWordNet is limited to 5% with respect to WordNet 1.6. The table also reports the average amount of unique synsets for each news. In this case the incompleteness of Italian WordNet with respect to WordNet 1.6 raises to 30%, showing that a significant amount of word senses is missing.

Table 3 shows the average polysemy of the news corpus considering both word senses and word domain labels. The figures reveal a polysemy reduction of 17-18% when we deal with domain polysemy.

Manual Annotation. A subset of forty news pairs (about half of the initial corpus) have been manually annotated with the correct domain label. Annotators were instructed about the domain hierarchy and then asked to select one domain label for each lemma among those allowed by that lemma.

Uncertain cases have been reviewed by a second annotator and, in case of persisting conflict, a third annotator was consulted to take a decision. Lemmatization errors as well as cases of incomplete coverage of domain labels have been detected and excluded. The whole manual set consists of about 2500 annotated nouns.

Although we do not have empirical evidences, our practical experience confirms the intuition that annotating texts with domain labels is an easier task than sense annotation.

Forty-two domain labels, representing the more informative level of the domain hierarchy mentioned in Section 1, have been used for the experiment. Table 4 reports the complete list.

5 Results and Discussion

WSD and WDD on the Semcor Brown Corpus. In the first experiment we wanted to verify that, because of the polysemy reduction induced by domain clustering, WDD is a simpler task than

administration	agriculture	alimentation	anthropology	archaeology	architecture
art	artisanship	astrology	astronomy	biology	chemistry
commerce	computer_science	earth	economy	engineering	factotum
fashion	history	industry	law	linguistics	literature
mathematics	medicine	military	pedagogy	philosophy	physics
play	politics	psychology	publishing	religion	sexuality
sociology	sport	telecommunication	tourism	transport	veterinary

Table 4: Domain labels used in the experiment.

	Baseline 1	*Baseline 2*	*Weak Mutual (baseline 2)*		*Strong Mutual (baseline 2)*	
			Synset Inter.	Domain Inter.	Synset Inter.	Domain Inter.
Italian	.83	.86	.87	.88	.74 / .68	.77 / .91
English	.85	.86	.87	.87	.70 / .57	.80 / .91

Table 5: Precision and Recall (English and Italian) for different WDD algorithms

WSD. For the experiment we used a subset of the Semcor corpus. As for WSD we obtained .66 of correct disambiguation with a sense frequency algorithm on polysemous noun words and .80 on all nouns (this last is also reported in the literature, for example in [Mihalcea and Moldovan, 1999]). As for WDD, precision has been computed considering the intersection between the word senses belonging to the domain label with the higher score and the sense tag for that word reported in Semcor. Baseline 1 and baseline 2, described in section 3.1, respectively gave .81 and .82 in precision, with a significant improvement over the WSD baseline, which confirms the initial hypothesis.

WDD in parallel texts. In this experiment we wanted to test WDD in the context of parallel texts. Table 5 reports the precision and recall (just in case it is not 1) scores for six different WDD algorithms applied to parallel English/Italian texts. Numbers refer to polysemous words only.

Both the baseline algorithms perform quite well: 83% for Italian and 85% for English in case of baseline 1, and 86% for both languages in case of baseline 2 are similar to the results obtained on the SemCor corpus.

The algorithm which includes word domain frequency (i.e. baseline 2) reaches the highest score in both languages, indicating that the combination of domain word frequency (considered at step 1 of the algorithm) and domain text frequency (considered at step 2) is a good one. In addition, the fact that results are the same for both languages indicates that the method can smooth the coverage differences among the wordnets.

We expected a better result for the bilingual extensions. The weak mutual strategy, either considering the synset intersection or the domain labels intersection, brings just minor improvements with respect to the baselines; the strong mutual strategy lowers both the precision and the recall. There are several explanations for these results. The difference in sense coverage between the two wordnets, about 30%, may affect the quality of the synset intersection: this would also explain the low degree of recall (68% for Italian and 57% for English). This is particularly evident for the strong mutual strategy, where the relative lexical poorness of the Italian synsets can strongly reduce the number of synsets in the intersection. Note also that the length of the synset intersection is about 30-40% of the mean synset number for Italian and English news respectively. This means less material which the disambiguation algorithms can take advantage of: relevant sysnsets can be left out of the intersection. For these reasons it is crucial having wordnet resources at the same level of completion to exploit the *mutual help* hypothesis.

Furthermore, there may be a significant amount of senses which are "quasi" aligned. This may happen when two parallel senses map into close synsets, but not in the same one (e.g. one is the direct hypernym of the other). This problem could be overcome considering the IS-A relations during the computation of the intersection. In this situation it is also probable that the senses maintain the same domain label. This would explain why the domain intersection behaves better than the synset intersection (from 74%-68% to 77%-91% for the Italian and from 70%-57% to 80%-91% for the English).

6 Conclusions

We have introduced Word Domain Disambiguation, a variant of Word Sense Disambiguation where words in a text are tagged with a *domain* label in place of a *sense* label. Two baseline algorithms has been presented as well as some extensions to deal with domain disambiguation in the context of parallel translation texts.

Two aligned wordnets, the English WordNet 1.6 and the Italian MultiWordNet, both augmented with domain labels, have been used as the main information repositories.

The experimental results encourage to further investigate the potentiality of word domain disambiguation. There are two interesting perspectives for the future work: first, we want to exploit the relations among different lexical categories (mainly nouns and verbs) when they share the same domain label; second, it seems reasonable that the disambiguation process may take advantage of both WDD and WSD, where the initial word ambiguity is first reduced with WDD and then resolved with more fine grained information. Finally, an in-depth investigation is necessary for what we called *factotum effect*, which is peculiar of WDD.

As for the applicative scenarios, we want to apply WDD to the problem of content based user modelling. In particular we are developing a personal agent for a news web site that learns user's interests from the requested pages that are analyzed to generate or to update a model of the user [Strapparava *et al.*, 2000]. Exploiting this model, the system anticipates which documents in the web site could be interesting for the user. Using MultiWordNet and domain disambiguation algorithms, a content-based user model can be built as a semantic network whose nodes, independent from the language, represent the word sense frequency rather then word frequency. Furthermore, the resulting user model is independent from the language of the documents browsed. This is particular valuable with multilingual web sites, that are becoming very common especially in news sites or in electronic commerce domains.

References

A. Artale, B. Magnini, and C. Strapparava. WordNet for italian and its use for lexical discrimination. In *AI*IA97: Advances in Artificial Intelligence*. Springer Verlag, 1997.

P. Buitelaar. Corelex: An ontology of systematic polysemous classes. In *Proceedings of FOIS98, International Conference on Formal Ontology in Information Systems, Trento, Italy, June 6-8 1998*. IOS Press, 1998.

P. Buitelaar. Reducing lexical semantic complexity with systematic polysemous classes and underspecification. In *Proceedings of ANLP2000 Workshop on Syntactic and Semantic Complexity in Natural Language Processing Systems, Seattle, USA, April 30 2000*, 2000.

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.

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.

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.

B. Magnini and C. Strapparava. Costruzione di una base di conoscenza lessicale per l'italiano basata su WordNet. In M. Carapezza, D. Gambarara, and F. Lo Piparo, editors, *Linguaggio e Cognizione*. Bulzoni, Palermo, Italy, 1997.

R. Mihalcea and D. Moldovan. A method for word sense disambiguation of unrestricted text. In *Proc. of ACL-99*, College Park Maryland, June 1999. held in conjunction with UM'96.

G. Miller. An on-line lexical database. *International Journal of Lexicography*, 13(4):235–312, 1990.

F. Sabatini and V. Coletti. *Dizionario Italiano Sabatini Coletti*. Giunti, 1997.

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.

E. Voorhees. Using wordnet for text retrieval. In C. Fellbaum, editor, *WordNet - an Electronic Lexical Database*. MIT Press, 1998.

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