Comparing Ontology-Based and Corpus-Based Domain Annotations in WordNet

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}@fififi.eu

Abstract

Domain information has been regarded as an emerging topic of interest in relation to WordNet. A lexical resource, WordNet Domains, is presented, where WordNet synsets have been annotated with domain labels such as Medicine, Architecture and Sport. This annotation reflects the lexico-semantic criteria adopted by humans involved in the annotation. However, from a corpus-based perspective, domains reflect term distribution in a given text collection. The paper proposes a preliminary investigation aiming at comparing and integrating ontology-based and corpus-based domain information.

1 Introduction

The starting point of this paper is WordNet Domains (Magnini and Cavaglià, 2000), an extension of WordNet 1.6 (Fellbaum, 1998) in which synsets have been annotated with one or more domain labels. The annotation methodology was mainly manual and based on lexico-semantic criteria which take advantage of the already existing conceptual relations in WordNet. However, a question is how well this annotation reflects the way synsets occur in a certain text collection. This issue is particularly relevant when we want to use the manual annotation for text processing tasks (e.g. word sense disambiguation). More than this, analyzing texts in term of domain distribution is a potential source of additional information with respect to the manual annotation. For the purposes of this paper we have limited our investigation to a corpus of news (i.e. the Reuters corpus), mainly because this collection is freely available for research purposes and because news are categorized by means of topic codes, which makes the comparison with WordNet Domains easier.

To make the problem more concrete, let us consider the WordNet synset {`heroin, diacetyl morphine, H, horse, junk, scag, shit, smack`} which is annotated with the Medicine domain because heroin is a drug (i.e. something that is used as a medicine or narcotic) and drug behavior is best described as part of medical knowledge. In the Reuters corpus the word "heroin" likely occurs in the context either of crime news or administrative news, without any strong relation with the medical field. This is a typical case of inconsistency, where the manual annotation considers a technical use of a word, while the corpus of news records a wider context of use. The negative consequence is that the influence of technical (i.e. scientific) domains in generic texts, such as news, is often overestimated.

However, for a quite large number of non technical domains (e.g. Sport, Politics), the above problem is less dramatic and a comparison between the lexico-semantic annotation in WordNet Domains and the data automatically acquired from the Reuters corpus can be carried out. We report the results of a preliminary experiment performed with two main goals: first, to provide a methodology for a corpus-based evaluation of WordNet Domains; second, to integrate the manual, ontology-based annotations with distributional information over the Reuters collection.

Our interest in domain information is motivated by its utility in many concrete scenarios, including word sense disambiguation (WSD) and text categorization (TC). In WSD the underlying hypothesis is that information provided by domain labels offers a natural 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. Results recently obtained at the SENSEVAL-2 initiative with a system based on domain information (Magnini et al., 2001) confirm the working hypothesis.

In TC, categories are symbolic labels, and no additional knowledge, neither of procedural nor of declarative nature, related to their meaning is available. A resource such as WordNet Domains could improve the integration of linguistic knowledge into traditional statistical approaches.

The paper is structured as follows. Section 2 presents WordNet Domains, describing the annotation methodology and some problematic aspects. Section 3 describes the procedure to automatically acquire domain information from the Reuters collection. Section 4 shows the results of an experiment performed on a subset of the corpus. Finally, Section 5 reports some relevant related works.

2 WordNet Domains

Domains have been used both in Linguistics (i.e. Semantic Fields) and in Lexicography (i.e. Subject Field Codes) to mark technical usages of words. Although this is a useful information for sense discrimination, in dictionaries it is typically used only for a small portion of the lexicon. WordNet Domains is an attempt to extend the coverage of domain labels within an already existing lexical database, WordNet (version 1.6). Synsets have 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 details about the domain taxonomy).

Information brought by domains 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 the experiment reported in this paper we have considered a set of 41 disjoint labels which allows 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).

The procedure for synsets annotation with domain labels is based on lexico-semantic criteria which exploit the WordNet taxonomy. First, a small number of high level synsets are manually annotated with their pertinent domain. Then, an automatic procedure exploits some of the WordNet relations

Sense	*Synset & Gloss*	*Domains*
#1	depository financial institution, bank, banking concern, banking company (a financial institution...)	Economy
#2	bank (sloping land...)	Geography, Geology
#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
#6	savings bank, coin bank, money box, bank (a container...)	Economy
#7	bank (a long ridge or pile...)	Geography, Geology
#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 and domains for the word "bank"

(i.e. hyponymy, troponymy, meronymy, antonymy and pertain-to) to extend the manual assignments to all the reachable synsets. As an example, this inheritance-based procedure allows to automatically mark the synset {`beak, bill, neb, nib`} with the code Zoology, starting from the synset {`bird`} and following a "part-of" relation.

There are cases in which the inheritance procedure has to be blocked, inserting an "exception", to prevent a wrong propagation. For instance, `barber_chair#1`, being a "part-of" `barbershop#1`, which in turn is annotated with Commerce, would wrongly inherit the same domain. To deal with these cases, the inheritance procedure allows the declaration of exceptions, such as:

```
assign shop#1 to Commerce
   with exception [part, isa, shop#1]
```

which assigns the synset `shop#1` to Commerce, but excludes all the parts of the daughters of `shop#1`, such as `barbershop#1`.

2.1 Factotum

There are a number of WordNet synsets that do not belong to a specific domain, but rather they can appear in almost all of them. For this reason, a

FACTOTUM label has been created which basically includes two types of synsets:

- *Generic* synsets, which are hard to classify in a particular domain, such as:

  `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

  They are generally placed high in the WORDNET hierarchy and are related senses of highly polysemous words. Many verb synsets fall in this category.

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

2.2 Specialistic VS Generic Usages

Domain labels (about 250) in WORDNET DOMAINS have been selected from dictionaries and then structured in a taxonomy which follows the Dewey Decimal Classification (DCC (Diekema, 1998)). The annotation task consists of interpreting a WORDNET synset with respect to the DCC classification. A relevant problem arises for those synsets that occur in a well defined context (i.e. specialistic) in the WORDNET hierarchy, but having a wider (i.e. generic) textual usage.

An example of this situation is the synset {`feeling -- (the psychological feature of experiencing affective and emotional states...)`}, which, given its definition, could be annotated under the PSYCHOLOGY domain. However, at least intuitively, the use of this synset in documents is broader then the psychological discipline, and a FACTOTUM annotation would be coherent with the constant distribution of this term in all the domains of the Reuters corpus.

What we expect from the corpus-based acquisition procedure described in the next Section is domain information which we can use either to modify or to integrate the domain annotation based on DCC.

3 Corpus-based Acquisition

This section reports the methodology used to automatically acquire domain information from the Reuters corpus and to compare it with domain annotations already present in WORDNET DOMAINS.

The following steps have been carried out: (i) linguistic processing of the corpus, which includes POS tagging, multiwords identification and filtering on WORDNET; (ii) acquisition of domain information for WORDNET synsets, based on probability distribution in the corpus; (iii) matching of the acquired information with domain manual annotation.

3.1 Experimental Setting

The Reuters Corpus (Reuters, 1997) is a collection of about 390,000 English news freely available for research purposes. Each news is annotated with one or more *Topic Code*, selected from a set of 127 labels, covering a variety of topics, even if economy and politics are prevalent. The mapping between Reuters Topic Codes and WORDNET DOMAINS domain labels is not trivial: domains are different both in their extension and in their structure. For the acquisition experiment described in this paper, a limited subset of the Reuters Topic Codes has been considered, which can be easily mapped to the domain labels used in WORDNET DOMAINS.

The selected Topic Codes (i.e. the domain set) are GREL (Religion), GENT (Arts, Culture, Entertainment), GVIO (War, Civil War), GCRIM (Crime, Law Enforcement) and GSPO (Sports), for a total of about 18 million tokens.

Domain	*Topic codes*	*# Reuters tokens*
RELIGION	GREL	307219
ART	GENT	400637
MILITARY	GVIO	3798848
LAW	GCRIM	2864378
SPORT	GSPO	2230613

Table 2: Domain Set for the acquisition experiment.

3.2 Linguistic Processing

The subset of the Reuters corpus was first lemmatized and annotated with part-of-speech tags. The Tree Tagger, developed at the University of Stuttgart (Schmid, 1994) has been used in this phase, as well as the WORDNET morphological analyzer, to resolve ambiguities and lemmatization mistakes. Then a filter was applied to identify the words actually contained in WORDNET 1.6, including multiwords. This process resulted in the individuation of 36,503 lemmas which include 6,137 multiwords.

Table 2 shows the mapping between Reuters Topic Codes and WORDNET domains, as well as the number of tokens in the Reuters Corpus which are in WORDNET 1.6 for each domain.

3.3 Acquisition Procedure

Given a synset in WORDNET DOMAINS, the acquisition procedure aims at identifying which domain is relevant in the Reuters corpus for that synset.

As a first step a *Relevant Lemma List* was built as the union of the synonyms and of the gloss for that synset. This list represents the context of the synset in WORDNET, and will be used to estimate the probability of a domain given a synset in the corpus.

This information is collected in a vector, called *Reuters Vector*, with a dimension in each domain. The value of each dimension is the probability of the domain, given the synset and it is calculated with the following formula:

$$P(D \mid s) = \frac{P(s \mid D)P(D)}{\sum_{i=1}^{n} P(s \mid D_i)P(D_i)}$$

where D is the given WordNet domain and s is a given synset.

The probability of the synset for a domain is assumed to be conditioned by the probability of its most related lemmas (those in the Relevant Lemma List) and is calculated with the following formula:

$$P(s \mid D) = P(l_1 ... l_m \mid D) = \prod_{i=1}^{m} P(l_i \mid D)$$

where l is a lemma in the Relevant Lemma List.

The probability of a lemma given a domain is its relative frequency in that domain; it is calculated with the following formula:

$$P(l_i \mid D) = \frac{c(l_i, D) + \epsilon}{c(D) + \epsilon |L|}$$

where $c(l, D)$ is the number of occurrences of lemma l in a domain, $c(D)$ is the total number of occurrences of the domain and $|L|$ is the total number of the lemmas in the considered corpus. For each domain, $P(D)$ was assumed to be 1, because we have not special domain requirements to fit.

3.4 Matching with manual annotation

In addition to the Reuters Vector, for each synset we have built a vector, called *Wordnet Vector*, with five dimensions, one for each selected domain; this was simply done by scoring 1 the selected domain and 0 all the other four domains.

At this point we have, for each synset, a Wordnet Vector and a Reuters Vector. These vectors are first normalized, then the scalar product between them has been calculated. We adopt this measure as an index of the *Proximity Score* between the two sources of domain information. This measure ranges from 0 to 1 and it is indicative of the similarity between the two annotations.

The scoring for each dimension of the Reuters domain vector (i.e. the probability to be in a given domain if a synset occurs), is interpreted as the relevance of a synset in a domain. Note that just five domains were selected for the experiment and a closed world is assumed.

4 Results and Discussion

In WordNet Domains there are 11,993 synsets with at least one manual annotation belonging to the selected domain set: 2094 for Religion, 2878 for Art, 4250 for Sport, 1376 for Law and 1395 for Military.

In the rest of this Section we discuss the application of the acquisition and comparison methodology to some relevant cases.

4.1 Synsets with unique manual annotation

In this experiment two restrictions have been applied to the initial set of 11,993 synsets: (i) a synset must have at least one word among its synonyms occurring at least one time in the Reuters Corpus and (ii) it must have just one domain annotation in WordNet Domains. This selection produced an experimental set of 867 synsets. As expected, the average value of Proximity Score obtained after the comparison of the Reuters and WordNet vectors for this subset of synsets was very high (i.e. 0.96), indicating that this subset of synsets is very relevant for the selected domains.

As an example, consider the synset, {`baseball, baseball game, ball game -- (a game played with a bat and ball between two teams of 9 players; teams take turns at bat trying to score run)`}, which was manually annotated with the Sport domain. Its WordNet Vector shows 1 in Sport and 0 elsewhere. Applying the acquisition procedure described in Section 3 over the corpus with the Relevant Lemma List extracted from the synset, the resulting Reuters Vector for the synset is reported in Table 3. The high score of Sport with respect to the other domains indicates a strong collocation of the synset in the Reuters, which is also confirmed by the Proximity Score (i.e. 1) resulting by the comparison of the synset manual annotation.

Law	Art	Religion	Sport	Military
$1.82e^{-60}$	$2.44e^{-56}$	$1.71e^{-152}$	1	$2.45e^{-63}$

Table 3: Reuters Vector for {`baseball, baseball game, ball game`}

Similar results are obtained for words whose senses belong to different domains among the five we have selected. This is the case, as instance, of the lemma "icon", that has three senses in Wordnet 1.6. `Icon#1 (Computer Science) a graphic symbol (...) that denotes a program...` was marked by WordNet Domains annotators as `Computer Science` and thus was not used for the experiment. `Icon#2` {`picture, image, icon, ikon -- (a visual representation of an object or scene...)`} was annotated with Art and `Icon#3` {icon, ikon -- (a conventional religious picture...)} was annotated with RELIGION.

The Reuters Vectors obtained for Icon#2 and Icon#3 are represented in Tables 4 and 5. While the annotated domains are different, in both cases the Proximity Score is 0.99.

LAW	ART	RELIGION	SPORT	MILITARY
$1.52e^{-05}$	0.99	$7.28e^{-05}$	$4.090e^{-08}$	$3.87e^{-08}$

Table 4: Reuters Vector for {picture, image, icon, ikon}

LAW	ART	RELIGION	SPORT	MILITARY
0.0005	$4.20e^{-45}$	0.99	$1.92e^{-51}$	0.0006

Table 5: Reuters Vector for {icon, ikon}

4.2 Synsets with multiple manual annotation

A number of synsets of the initial set have been annotated with multiple domain labels in WORDNET DOMAINS. This is the case of the adjective canonic#2: {canonic, canonical -- (of or relating to or required by canon law)}, which is annotated with two labels, RELIGION and LAW. We used this Relevant Lemma List: canonic#a canonical#a relate#v require#v canon#n law#n.

The Reuters Vector for canonic#2 is shown in Table 6. We can see that the most relevant domains are RELIGION and LAW, with a Proximity Score of 0.99.

LAW	ART	RELIGION	SPORT	MILITARY
0.41	$9.48e^{-47}$	0.56	0.004	0.02

Table 6: Reuters Vector for {canonic, canonical}

4.3 Factotum annotations

FACTOTUM synsets do not belong to any specific domain, and should have high frequency in all the Reuters texts. For instance, the synset containing the verb "to be" in its first sense, {be -- (have the quality of being)}, corresponds to the Reuters Vector represented in Table 7, which manifests a very high Proximity Score (i.e. 0.99) with respect to the WORDNET vector.

LAW	ART	RELIGION	SPORT	MILITARY
0.21	0.29	0.20	0.16	0.20

Table 7: Reuters Vector for {be}

4.4 Mismatching annotations

For some synsets the two vectors produce different classifications. As instance, the synset {wrath, anger, ire, ira -- (belligerence aroused by a real or supposed wrong (personified as one of the deadly sins))}, is annotated with RELIGION, inherited from its hypernym {mortal sin, deadly sin}. However, the preferred domain acquired from the Reuters Vector, shown in Table 8, is MILITARY. This is mainly due to the fact that many lemmas in this synset pertain to the MILITARY domain. The only lemma that forces our choice in the RELIGION domain is "deadly_sin", which is very rare in the Reuters corpus.

LAW	ART	RELIGION	SPORT	MILITARY
$1.4e^{-45}$	3.5^{-44}	5.2^{-13}	3.7^{-48}	1

Table 8: Reuters Vector for {wrath, anger, ire, ira}

4.5 Covering problems

For some synsets the Relevant Lemma List is not covered enough in the Reuters corpus to produce a significant domain classification. This is the case, for instance, of the synset {Loki -- (trickster; god of discord and mischief; contrived death of Balder and was overcome by Thor)}, manually annotated with RELIGION, due also to its hypernym {deity, divinity, god, immortal}. Its Reuters Vector is shown in Table 9.

LAW	ART	RELIGION	SPORT	MILITARY
$2.10e^{-44}$	1.45^{-131}	2.63^{-13}	6.78^{-46}	1

Table 9: Reuters Vector for {Loki}

The preferred domain, MILITARY, depends both on the absence in the corpus of some lemmas (i.e. Loki, Balder, Thor) and on the presence of terms strongly related to the MILITARY domain (i.e. discord, death, overcome).

5 Related Works

The importance of domain information in relation to WORDNET has been remarked by several works in the last years.

(Gonzalo et al., 1998) emphasizes the role of domains for WSD. Following this line, (Magnini and Strapparava, 2001b) 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 example, (Magnini and Strapparava, 2001a) describes the use of a content-based document representation as a starting point to build a model of the user's interests.

A closely related work is that of (Buitelaar and Sacaleanu, 2001), which describes a method for determining the relevance of GermaNet synsets with respect to a specific domain. A case study on three domains (i.e. BUSINESS, SOCCER and MEDICAL) is reported, where three different corpora have been used. Term Relevance of a synset with respect to a domain is calculated summing up term relevances for words in the synset and in its hyponyms (with a penalty for missing hyponyms).

Finally, a methodology for the integration of domain specific information into generic synsets is suggested in (Vossen, 2001).

6 Conclusions

In this paper we have presented a lexical resource, WORDNET DOMAINS, where synsets have been annotated with domain labels following lexico-semantic criteria. A preliminary investigation has been presented aiming at comparing the lexico-semantic annotation with a corpus-based annotation. The goal was twofold: on the one hand to provide an evaluation of the manual annotation based on distribution of domain information in a large corpus; on the other hand to integrate probabilistic information into WORDNET DOMAINS.

A procedure for the automatic acquisition of domain information from domain-annotated corpora, as well as the results of an experiment over a subset of the Reuters corpus, have been described. Results show that an high degree of matching between ontology-based and corpus-based annotations can be reached for a limited number of domain relevant synsets. The lower degree of correspondence reached for other synsets is due either to limitations of coverage (i.e. words in WORDNET not present in the corpus) or to different interpretations of the synset (e.g. specialistic versus generic interpretation).

We consider the work presented here as a first step in the direction of a full and automatic procedure for the acquisition of domain information from corpora. For the future we plan to collect and use large and diverse domain-annotated corpora, the long term goal of this research being the integration of corpus-based domain information within the WORDNET taxonomy. At the same time we will continue exploring the role of domain information in relation to the use of WORDNET, mainly for word sense disambiguation tasks.

References

P. Buitelaar and B. Sacaleanu. 2001. Ranking and selecting synsets by domain relevance. In *Proc. of NAACL Workshop WordNet and Other Lexical Resources: Applications, Extensions and Customizations*, Pittsburgh, June. held in conjunction with NAACL2001.

A. Diekema. 1998. *Dewey decimal classification.* DDC.

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. 2001a. Improving user modelling with content-based techniques. In *UM2001 User Modeling: Proc. of 8^{th} International Conference on User Modeling (UM2001)*, Sonthofen (Germany), July. Springer Verlag.

B. Magnini and C. Strapparava. 2001b. Using wordnet to improve user modelling in a web document recommender system. In *Proc. of NAACL Workshop WordNet and Other Lexical Resources: Applications, Extensions and Customizations*, Pittsburgh, June. held in conjunction with NAACL2001.

B. Magnini, C. Strapparava, G. Pezzulo, and A. Gliozzo. 2001. Using domain information for word sense disambiguation. In *Proc. of SENSEVAL-2.* to appear.

Reuters. 1997. http://about.reuters.com/researchandstandards/corpus/.

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

P. Vossen. 2001. Extending, trimming and fusing wordnet for technical documents. In *Proc. of NAACL Workshop WordNet and Other Lexical Resources: Applications, Extensions and Customizations*, Pittsburgh, June. held in conjunction with NAACL2001.