Using Relevant Domains Resource for Word Sense Disambiguation

Sonia Vázquez, Andrés Montoyo
Department of Software and Computing Systems
University of Alicante
Alicante, Spain
{svazquez,montoyo}@dlsi.ua.es

German Rigau
Department of Computer Languages and Systems
Euskal Herriko Unibertsitatea
Donostia, Spain
jipriclg@si.ehu.es

Abstract

This paper presents a new method for Word Sense Disambiguation based on the WordNet Domains lexical resource [4]. The underlaying working hypothesis is that domain labels, such as ARCHITECTURE, SPORT and MEDICINE provide a natural way to establish semantic relations between word senses, that can be used during the disambiguation process. This resource has already been used on Word Sense Disambiguation [5], but it has not made use of glosses information. Thus, we present in first place, a new lexical resource based on WordNet Domains glosses information, named "Relevant Domains". In second place, we describe a new method for WSD based on this new lexical resource ("Relevant Domains"). And finally, we evaluate the new method with English all-words task of SENSEVAL-2, obtaining promising results.

Keywords: Word Sense Disambiguation, Computational Lexicography.

1. Introduction and motivation

The development and convergence of computing, telecommunications and information systems has already led to a revolution in the way that we work, communicate with other people, buy news and use services, and even in the way that we entertain and educate ourselves. The revolution continues and one of its results is that large volumes of information will be shown in a format that is more natural for users than the typical data presentation formats of past computer systems. Natural Language Processing (NLP) is crucial in solving these problems and language technologies will make an indispensable contribution to the success of information systems.

Designing a system for NLP requires a large knowledge on language structure, morphology, syntax, semantics and pragmatic nuances. All of these different linguistic knowledge forms, however, have a common associated problem, their many ambiguities, which are difficult to resolve.

In this paper we concentrate on the resolution of the lexical ambiguity that appears when a given word has several different meanings. This specific task is commonly referred as Word Sense Disambiguation (WSD). The disambiguation of a word sense is an "intermediate task" [8] and it is necessary to resolve such problems in certain NLP applications, as Machine Translation (MT), Information Retrieval (IR), Text Processing, Grammatical Analysis, Information Extraction (IE), hypertext navigation and so on. In general terms, WSD intents to assign a definition to a selected word, in a text or a discourse, that endows it with a meaning that distinguishes it from all of the other possible meanings that the word might have in other contexts. This association of a word to one specific sense is achieved by acceding to two different

information sources, known as context[1] and external knowledge sources[2].

The method we propose in this paper is based on strategic knowledge (knowledge-driven WSD), that is, the disambiguating of nouns by matching the context in which they appear with the information from WordNet lexical resource.

WordNet is not a perfect resource for word-sense disambiguation, because it has the problem of the fined-grainedness of WordNet´s sense distinctions [2]. This problem causes difficulties in the performance of automatic word-sense disambiguation with free-running texts. Several authors [8, 3] have stated that the divisions of a proposed sense in the dictionary are too fine for Natural Language Processing. To solve this problem, we propose a WSD method for applications that do not require a fine granularity for senses distinctions. This method consists of labelling texts words with a domain label instead of a sense label. We named domains to a set of words with a strong semantic relation. Therefore, applying domains to WSD contributes with a relevant information to establish semantic relations between word senses. For example, "bank" has ten senses in WordNet 1.6 but three of them "bank#1", "bank #3" and "bank #6" are grouped into the same domain label "Economy", whereas "bank#2" and "bank#7" are grouped into domains labels "Geography" and "Geology".

A lexical resource with domain labels associated to word senses is necessary for the WSD proposed method. Thus, a new lexical resource has been developed, named Relevant Domains obtained from "WordNet Domains" [4].

A proposal in WSD using domains has been developed in [5]; they use WordNet Domains as lexical resource, but from our point of view they don't make good use of glosses information. Thus, in this paper we present a new lexical resource obtained from glosses information of WordNet Domains and a new WSD method that use this new lexical resource. This new method is evaluated with English all-words task of SENSEVAL-2, obtaining promising results.

The organisation of this paper is: after this introduction, in section 2 we describe the new lexical resource, named Relevant Domains. In section 3, the new WSD method is presented using the Relevant Domains resource. In section 4, an evaluation of WSD method is realized, and finally conclusions and an outline of further works are shown.

2. New resource: Relevant Domains

WordNet Domains [4] is an extension of WordNet 1.6 where each synset has one or more domain labels. Synsets associated to different syntactic categories can have the same domain labels. These domain labels are selected from a set of about 250 hundred labels, hierarchically organized in different specialization levels. This new information added to WordNet 1.6., allows to connect words that belong to different subhierarchies and to include into the same domain label several senses of the same word. Thus, a single domain label may group together more than one word sense, obtaining a reduction of the polysemy. Table 1 shows an example. The word "music" has six different senses in WordNet 1.6.: four of them are grouped under the MUSIC domain, causing the reduction of the polysemy from six to three senses.

Table 1. Domains associated to word "music"

Synset	Domain	Noun	Gloss
05266809	Music	music#1	an artistic form of auditory …
04417946	Acoustics	music#2	any agreeable (pleasing…
00351993	Music, and Free_time	music#3	a musical diversion; his music…
05105195	Music	music#4	a musical composition in…
04418122	Music	music#5	the sounds produced by singers..
00755322	Law	music#6	punishment for one's actions;…

1 Context is a set of words which are around the word to disambiguate along with syntactical relations, semantic categories and so on.

2 External knowledge resources are lexical resources, as WordNet, manually developed to give valuable information for associating senses to words.

In this work, WordNet Domains will be used to collect examples of domains associations to the different meanings of the words. To realize this task, WordNet Domains glosses will be used to collect the more relevant and representative domain labels for each English word. In this way, the new resource named Relevant Domains, contains all words of WordNet Domains glosses, with all their domains and they are organised in an ascendant way because of their relevance in domains.

To collect the most representative words of a domain, we use the "Mutual Information" formula (1) as follows:

$$MI(w, D) = \log_2 \frac{\Pr(w \mid D)}{\Pr(w)} \quad (1)$$

W: word.
D: domain.

Intuitively, a representative word is that appears in a domain context most frequently. But we are interested about the importance of words in a domain, that is, the most representative and common words in a domain. We can appreciate this importance with the "Association Ratio" formula:

$$AR(w, D) = \Pr(w \mid D) \log_2 \frac{\Pr(w \mid D)}{\Pr(w)} \quad (2)$$

W: word.
D: domain.

Formula (2) shows "Association Ratio" that is applied to all words with noun grammatical category obtained from WordNet Domains glosses. Later, the same process is applied to verbs, adjectives and adverbs grammatical categories. A proposal in this sense has been made in [6], but using Lexicography Codes of WordNet Files.

In order to obtain Association Ratio for nouns of WordNet Domains glosses, it is necessary to use a parser which obtains all nouns appeared in each gloss. For this task, we use "Tree Tagger" parser [7].

For example, the gloss associated to sense "music#1" is the following: *"An artistic form of auditory communication incorporating instrumental or vocal tones in a structured and continuous manner"*. Then, Table 2 shows the domains associated with gloss nouns of "music#1".

Table 2. Domains association with gloss nouns of "music#1"

Domain	Noun
Music	form
Music	communication
Music	tone
Music	manner

This process is realized with all the WordNet Domains glosses to obtain all the domains associated to each noun for begining with the Association Ratio calculus. Finally, we obtain a list of nouns with their associated domains sorted by Association Ratio. With this format, the domains that appear in first positions of a noun are the most representatives. The results of the Association Ratio for noun "music" are showed in Table 3. Thus, the most representative domains for noun "music" are: MUSIC, FREE-TIME and ACOUSTICS.

After the Association Ratio for nouns, the same process is done to obtain Association Ratio for verbs, adjectives and adverbs.

Table 3. Association Ratio of "music"

Noun	Domain	A.R.
music	Music	0.240062
music	Free_time	0.093726
music	Acoustics	0.072362
music	Dance	0.065254
music	University	0.046024
music	Radio	0.042735
music	Art	0.020298
music	Telecommunication	0.006069
…	…	…

3. WSD method

The method presented here is basically about the automatic sense-disambiguation of words that appear

into the context of a sentence, with their different possible senses quite related. The context is taken from the words that co-occur with the proposed word into a sentence and from their relations to the word to be disambiguated. The WSD method that we propose in this paper, is connected with the strategic knowledge, because it uses the new resource "Relevant Domains" as an information source to disambiguate word senses into a text.

So that our WSD method needs a new structure that contains the most representative domains sorted by the Association Ratio formula in the context of a sentence. This structure is named context vector. Furthermore, each polysemic word in the context has different senses and for each sense we need a structure which contains the most represenative domains sorted equally by the Association Ratio formula. This structure is named sense vector.

In order to obtain the correct word senses into the context, we must measure the proximity between context vector and sense vectors. This proximity is measured with cosinus between both vectors, that is, the more cosinus the more proximity between both vectors.

Next subsections describe each one of the structures and their integration in the WSD method.

3.1. Context vector

Context vector combines in only one structure the most relevant and representative domains related to the words from the text to be disambiguated, that is, the information of all the words (nouns, verbs, adjectives and adverbs) of the text to be disambiguated. With this information we try to know which domains are the most relevant and representative into the text. In order to obtain this vector we use information from the Relevant Domains lexical resource. Thus, we will obtain domains sorted by Association Ratio values for nouns, verbs, adjectives and adverbs taken from the text to be disambiguated. Then each word is measured according to a list of relevant domain labels. Finally, we obtain a sorted vector where the most relevant and representative domain labels are in the first positions.

A formal representation of context vector is showed in formula (3).

$$CV = \sum_{w \in context} AR(W, D) \quad (3)$$

Figure 1 shows the context vector obtained from the following text: "There are a number of ways in which the chromosome structure can change, which will detrimentally change the genotype and phenotype of the organism".

$$CV = \begin{pmatrix} \textbf{Domain} & \textbf{A.R.} \\ \text{Biology} & 0.03102837 \\ \text{Ecology} & 0.00402855 \\ \text{Botany} & 3.20408e-05 \\ \text{Zoology} & 1.77959e-05 \\ \text{Anatomy} & 1.29592e-05 \\ \text{Physiology} & 1.00022e-06 \\ \text{Chemistry} & 1.00017e-06 \\ \text{Geology} & 1.66327e-07 \\ \text{Meteorology} & 1.00371e-07 \\ \ldots & \ldots \end{pmatrix}$$

Figure 1: Context Vector

3.2. Sense vector

Sense vector groups the most relevant and representative domains of the gloss that is associated with each one of the word senses into the same structure. That is, we take advantage of the information of the glosses of WordNet. In this way, the glosses are analyzed syntactically and their words are pos-tagged (nouns, verbs, adverbs and adjectives). Then the same calculus done with the context vector will be done with the sense vector, in order to obtain one vector for each sense of all words in the text. For example, we obtain the sense vector showed in Figure 2 for sense "genotype#1".

Domain	**A.R.**

$$VS = \begin{pmatrix} \text{Ecology} & 0.084778 \\ \text{Biology} & 0.047627 \\ \text{Bowling} & 0.019687 \\ \text{Archaeology} & 0.016451 \\ \text{Sociology} & 0.014251 \\ \text{Alimentation} & 0.006510 \\ \text{Linguistics} & 0.005297 \\ \ldots & \ldots \end{pmatrix}$$

Figure 2: Sense vector associated to "genotype#1"

3.3. Vectors comparison

The new WSD proposed method begins with the syntactic analysis of the text, using "Tree tagger". We calculate the context and sense vectors from these words that are tagged with their pos. From these vectors it is necessary to estimate, with the cosinus measure, which of them are more approximated to the context vector. We will select the senses with the cosinus more approximated to 1.

To calculate the cosinus we use the normalized correlation coefficient in formula (4):

$$\cos(CV, SV) = \sum_{i=1..n} \frac{CV * SV}{\sqrt{\sum_{i=1..n} CV^2} * \sqrt{\sum_{i=1..n} SV^2}} \quad (4)$$

CV: Context vector
SV: Sense vector

In order to select the appropriate sense, we made a comparison between all the sense vectors and the context vector, and we select the senses more approximated to the context vector.

For example, the cosinus between the context vector and the sense vectors of "genotype" has the next values:

genotype#1 = 0.00804111
genotype#2 = 0.00340548

Therefore, we select the genotype#1 sense, because its cosinus is nearest to 1.

4. Evaluation and discussion

In this section we evaluated the new method WSD from texts of English all-words task from SENSEVAL-2. In these texts, nouns, verbs, adjectives and adverbs are tagged with their senses. These words will be disambiguated using the new method WSD, and later, the results obtained will be compared with the senses obtained in SENSEVAL-2 by other WSD methods. In order to measure the evaluation, we use precision and coverage values. To obtain the precision measure we divide the number of senses correctly disambiguated by the number of senses answered. And to obtain the recall measure we divide the number of senses correctly disambiguated by total number of senses.

The evaluation has been carried out taking different windows sizes. Thus, the first evaluation takes one sentence as window size. In this way, the WSD method disambiguate all the words that appear in the sentence. Therefore, the context of the words to be disambiguated is not too large, because the number of words is very limited. The results obtained in the first evaluation are showed in the row 1 of the Table 4.

In the second evaluation, we select a window of 100 words that contains the ambiguous word. In this evaluation the ambiguous word is related to a large group of words, that perform the context and give more information about domain relations. The results obtained in the second evaluation are showed in the row 2 of the Table 4.

In the third evaluation, we reduce the domain specialization levels, that is, the domains are grouped in a more general domain level. This reduction is realized over the WordNet Domains hierarchy structure. Therefore, 43 domains are obtained from the 165 previous ones. Really, the domains are grouped from the top levels. For example, domain level "Medicine" contains the following domains: Dentistry, Pharmacy, Psychiatry, Radiology and Surgery. These domains are included into "Medicine", so the specialization and the search space are reduced. The results of the third evaluation are showed in the row 3 of the Table 4.

The last evaluation, is realized considering the WordNet granularity. As WordNet has a subtle granularity it is very difficult to establish distinctions between different senses. So, in this evaluation we use 165 domains, but when we obtain the words senses, all

senses labeled with the same domain are returned. For example, if the WSD process returns the domain "Economy" for the word "bank", the results showed will be: "bank#1, bank#3 and bank#6. These senses have been labeled with the domain Economy. The results of the fourth evaluation are showed in the row 4 of the Table 4.

Table 4: Results obtained in WSD evaluations

Decision	Precision	Recall
Sentence	0.44	0.32
Window 100 words	0.47	0.38
43 domains	0.48	0.41
Domain level WSD	0.54	0.43

First line in table 4 shows a precision of 44%, that is obtained when we evaluate from a sentence that contains the ambiguos word. This result is due to the reduced number of words in the sentence context. Then, the WSD method can not obtain a context vector with a correct information.

In the second evaluation, we use a window of 100 words containing the ambiguous word. In this case a 47% precision is obtained. This result confirms that context vector is better in a 100 words window.

In the third evaluation, where specification level is reduced with a 100 words window, the results obtained in relation to the second evaluation have not a signifier difference. Nevertheless, when we try to disambiguate with domain levels, the results are better. This improvement is related with the WordNet granularity, because the senses obtained have the same associated domain and it is very difficult to select the correct sense.

The results obtained with our WSD method in English all-words task of SENSEVAL-2 are showed in the Table 4. In comparison with the results obtained by other systems, we are in a middle position, just as we can see in the Table 5.

Table 5: Classification attending to the results of English all-words task in SENSEVAL-2.

System	Precision	Recall
SMWaw-	0.690	0.690
Ave-Antwerp	0.636	0.636
LIA-Sinequa-AllWords	0.618	0.618
David-fa-UNED-AW-T	0.575	0.569
David-fa-UNED-AW-U	0.556	0.550
Gchao2-	0.475	0.454
Gchao3-	0.474	0.453
Ken-Litkowski-clr-aw	0.451	0.451
Gchao-	0.500	0.449
WSD-UA	**0.540**	**0.430**
cm.guo-usm-english-tagger2	0.360	0.360
Magnini2-irst-eng-all	0.748	0.357
Cmguo-usm-english-tagger	0.345	0.338
c.guo-usm-english-tagger3	0.336	0.336
Agirre2-ehu-dlist-all	0.572	0.291
Judita-	0.440	0.200
Dianam-system3ospdana	0.545	0.169
Dianam-system2ospd	0.566	0.169
Dianam-system1	0.598	0.140
Woody-IIT2	0.328	0.038
Woody-IIT3	0.294	0.034
Woody-IIT1	0.287	0.033

5. Conclusions and further works

In this paper we present a new lexical resource named Relevant Domains from glosses of WordNet Domains, and a new WSD method based in this new Lexical Resource. This new WSD method with Relevant Domains improves Magnini and Strapparava's work, because they didn't take advantage of WordNet Domains glosses information. Nevertheless, the Relevant Domains resource and the new WSD method get information about the glosses of the WordNet Domains.

The results obtained in the evaluation process confirm that the new WSD method obtains a promising precision and recall measures, for the word sense disambiguation task.

We extract an important conclusion about domains because they establish semantic relations between the word senses, grouping them into the same semantic

category (sports, medicine…). With our WSD method also we can resolve WordNet Granularity for senses.

Also, the new lexical resource Relevant Domains, is a new information source that can complete other WSD methods like Information Retrieval Systems, Question Answering...

In further works we will try to attach new information about the Relevant Domains, using Semcor or other tagged corpus. Therefore the WSD method will be evaluated again.

Finally we will try building a multilingual process addapting the WSD method and the Relevant Domains to each possible language.

References

[1] Church K. and Hanks P., *Word association norms, mutual information, and leogicography*. Computational Linguistics, vol. 16, ns. 1, 22-29. 1990. Also in proceedings of the 27th Annual Meeting of the Association for Computational Linguistics (ACL'89). Pittsburg, Pennsylvannia, 1989.

[2] Ide N. and Véronis J. (1998) *Introduction to the Special Issue on Word Sense Disambiguation: The State of the Art.* Computational Linguistics. 24 (1), 1-40.

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

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

[5] Magnini B. and Strapparava C., *Experiments in Word Domain Disambiguation for Parallel Texts*.In Proc. Of SIGLEX Workshop on Word Senses and Multi-linguaty, Hong-Kong, October 2000.

[6] Rigau G., Atserias J. and Agirre E., *Combining Unsupervised Lexical Knowledge Methods for Word Sense Disambiguation*. Proceedings of joint 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics ACL/EACL'97. Madrid, Spain, 1997.

[7] Schmid Helmut (1994) *Probabilistic part-of-speech tagging using decision tre*, Proceedings International Conference on New Methods in Language Processing.Manchester, pp 44-49. UK

[8] Wilks Y. And Stevenson M. (1996) *The grammar of sense: Is word sense tagging much more than part-of-speech tagging?* Technical Report CS-96-05, University of Sheffield, UK.