GPLSI: Word Coarse-grained Disambiguation aided by Basic Level Concepts[*]

Rubén Izquierdo **Armando Suárez**
GPLSI Group, DLSI
University of Alicante
Spain
{ruben, armando}@dlsi.ua.es

German Rigau
IXA NLP Group
EHU/UPV
Donostia, Basque Country
german.rigau@ehu.es

Abstract

We present a corpus-based supervised learning system for coarse-grained sense disambiguation. In addition to usual features for training in word sense disambiguation, our system also uses Base Level Concepts automatically obtained from WordNet. Base Level Concepts are some synsets that generalize a hyponymy sub–hierarchy, and provides an extra level of abstraction as well as relevant information about the context of a word to be disambiguated. Our experiments proved that using this type of features results on a significant improvement of precision. Our system has achieved almost 0.8 F1 (fifth place) in the coarse–grained English all-words task using a very simple set of features plus Base Level Concepts annotation.

1 Introduction

The GPLSI system in SemEval's task 7, *coarse–grained English all-words*, consists of a corpus-based supervised-learning method which uses local context information. The system uses Base Level Concepts (BLC) (Rosch, 1977) as features. In short, BLC are synsets of WordNet (WN) (Fellbaum, 1998) that are representative of a certain hyponymy sub–hierarchy. The synsets that are selected to be BLC must accomplish certain conditions that will be explained in next section. BLC are slightly different from Base Concepts of EuroWordNet[1] (EWN) (Vossen et al., 1998), Balkanet[2] or Meaning Project[3] because of the selection criteria but also because our method is capable to define them automatically. This type of features helps our system to achieve 0.79550 F1 (over the First–Sense baseline, 0.78889) while only four systems outperformed ours being the F1 of the best one 0.83208.

WordNet has been widely criticised for being a sense repository that often offers too fine–grained sense distinctions for higher level applications like Machine Translation or Question & Answering. In fact, WSD at this level of granularity, has resisted all attempts of inferring robust broad-coverage models. It seems that many word–sense distinctions are too subtle to be captured by automatic systems with the current small volumes of word–sense annotated examples. Possibly, building class-based classifiers would allow to avoid the data sparseness problem of the word-based approach.

Thus, some research has been focused on deriving different sense groupings to overcome the fine–grained distinctions of WN (Hearst and Schütze, 1993) (Peters et al., 1998) (Mihalcea and Moldovan, 2001) (Agirre et al., 2003) and on using predefined sets of sense-groupings for learning class-based classifiers for WSD (Segond et al., 1997) (Ciaramita and Johnson, 2003) (Villarejo et al., 2005) (Curran, 2005) (Ciaramita and Altun, 2006). However, most of the later approaches used the original Lexicographical Files of WN (more recently called Super-

[*] This paper has been supported by the European Union under the project QALL-ME (FP6 IST-033860) and the Spanish Government under the project Text-Mess (TIN2006-15265-C06-01) and KNOW (TIN2006-15049-C03-01)

[1] http://www.illc.uva.nl/EuroWordNet/

[2] http://www.ceid.upatras.gr/Balkanet

[3] http://www.lsi.upc.es/ nlp/meaning

Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), pages 157–160,
Prague, June 2007. ©2007 Association for Computational Linguistics

senses) as very coarse–grained sense distinctions. However, not so much attention has been paid on learning class-based classifiers from other available sense–groupings such as WordNet Domains (Magnini and Cavaglia, 2000), SUMO labels (Niles and Pease, 2001), EuroWordNet Base Concepts or Top Concept Ontology labels (Atserias et al., 2004). Obviously, these resources relate senses at some level of abstraction using different semantic criteria and properties that could be of interest for WSD. Possibly, their combination could improve the overall results since they offer different semantic perspectives of the data. Furthermore, to our knowledge, to date no comparative evaluation have been performed exploring different sense–groupings.

This paper is organized as follows. In section 2, we present a method for deriving fully automatically a number of Base Level Concepts from any WN version. Section 3 shows the details of the whole system and finally, in section 4 some concluding remarks are provided.

2 Automatic Selection of Base Level Concepts

The notion of Base Concepts (hereinafter BC) was introduced in EWN. The BC are supposed to be the concepts that play the most important role in the various wordnets[4] (Fellbaum, 1998) of different languages. This role was measured in terms of two main criteria:

- A high position in the semantic hierarchy;
- Having many relations to other concepts;

Thus, the BC are the fundamental building blocks for establishing the relations in a wordnet and give information about the dominant lexicalization patterns in languages. BC are generalizations of features or semantic components and thus apply to a maximum number of concepts. Thus, the Lexicografic Files (or Supersenses) of WN could be considered the most basic set of BC.

#rel.	synset
18	group_1,grouping_1
19	social_group_1
37	organisation_2,organization_1
10	establishment_2,institution_1
12	faith_3,religion_2
5	Christianity_2,**church_1**,Christian_church_1

#rel.	synset
14	entity_1,something_1
29	object_1,physical_object_1
39	artifact_1,artefact_1
63	construction_3,structure_1
79	building_1,edifice_1
11	place_of_worship_1, ...
19	**church_2**,church_building_1

#rel.	synset
20	act_2,human_action_1,human_activity_1
69	activity_1
5	ceremony_3
11	religious_ceremony_1,religious_ritual_1
7	service_3,religious_service_1,divine_service_1
1	**church_3**,church_service_1

Table 1: Possible Base Level Concepts for the noun *Church*

Basic Level Concepts (Rosch, 1977) should not be confused with Base Concepts. BLC are the result of a compromise between two conflicting principles of characterization:

- Represent as many concepts as possible;
- Represent as many features as possible;

As a result of this, Basic Level Concepts typically occur in the middle of hierarchies and less than the maximum number of relations. BC mostly involve the first principle of the Basic Level Concepts only.

Our work focuses on devising simple methods for selecting automatically an accurate set of Basic Level Concepts from WN. In particular, our method selects the appropriate BLC of a particular synset considering the relative number of relations encoded in WN of their hypernyms.

The process follows a bottom-up approach using the chain of hypernym relations. For each synset in WN, the process selects as its Base Level Concept the first local maximum according to the relative number of relations. For synsets having multiple hypernyms, the path having the local maximum with higher number of relations is selected. Usually, this process finishes having a number of “fake” Base Level Concepts. That is, synsets having no descendants (or with a very small number) but being the first local maximum according to the number of relations considered. Thus, the process finishes checking if the number of concepts subsumed by the

[4] http://wordnet.princeton.edu

	Senses	BLC	SuperSenses
Nouns	4.92	4.10	3.01
Verbs	11.00	8.67	1.03
Nouns + Verbs	7.66	6.16	3.47

Table 2: Polysemy degree over SensEval–3

preliminary list of BLC is higher than a certain threshold. For those BLC not representing enough concepts according to a certain threshold, the process selects the next local maximum following the hypernym hierarchy.

An example is provided in table 1. This table shows the possible BLC for the noun "church" using WN1.6. The table presents the hypernym chain for each synset together with the number of relations encoded in WN for the synset. The local maxima along the hypernym chain of each synset appears in bold.

Table 2 presents the polysemy degree for nouns and verbs of the different words when grouping its senses with respect the different semantic classes on SensEval–3. Senses stand for the WN senses, BLC for the Automatic BLC derived using a threshold of 20 and SuperSenses for the Lexicographic Files of WN.

3 The GPLSI system

The GPLSI system uses a publicly available implementation of Support Vector Machines, SVMLight[5] (Joachims, 2002), and Semcor as learning corpus. Semcor has been properly mapped and labelled with both BLC[6] and sense-clusters.

Actually, the process of training-classification has two phases: first, one classifier is trained for each possible BLC class and then the SemEval test data is classified and enriched with them, and second, a classifier for each target word is built using as additional features the BLC tags in Semcor and SemEval's test.

Then, the features used for training the classifiers are: lemmas, word forms, PoS tags[7], BLC tags, and first sense class of target word (S1TW). All features were extracted from a window $[-3..+3]$ except for the last type (S1TW). The reason of using S1TW features is to assure the learning of the baseline. It is well known that Semcor presents a higher frequency on first senses (and it is also the baseline of the task finally provided by the organizers).

Besides, these are the same features for both first and second phases (obviously except for S1TW because of the different target set of classes). Nevertheless, the training in both cases are quite different: the first phase is class-based while the second is word-based. By word-based we mean that the learning is performed using just the examples in Semcor that contains the target word. We obtain one classifier per polysemous word are in the SemEval test corpus. The output of these classifiers is a sense-cluster. In class-based learning all the examples in Semcor are used, tagging those ones belonging to a specific class (BLC in our case) as positive examples while the rest are tagged as negatives. We obtain so many binary classifiers as BLC are in SemEval test corpus. The output of these classifiers is $true$ or $false$, "the example belongs to a class" or not. When dealing with a concrete target word, only those BLC classifiers that are related to it are "activated" (i.e, "animal" classifier will be not used to classify "church"), ensuring that the word will be tagged with coherent labels. In order to avoid statistical bias because of very large set of negative examples, the features are defined from positive examples only (although they are obviously used to characterize all the examples).

4 Conclusions and further work

The WSD task seems to have reached its maximum accuracy figures with the usual framework. Some of its limitations could come from the sense–granularity of WN. In particular, SemEval's coarse-grained English all-words task represents a solution in this direction.

Nevertheless, the task still remains oriented to words rather than classes. Then, other problems arise like data sparseness just because the lack of adequate and enough examples. Changing the set of classes could be a solution to enrich training corpora with many more examples Another option seems to be incorporating more semantic information.

[5] http://svmlight.joachims.org/

[6] Because BLC are automatically defined from WN, some tuning must be performed due to the nature of the task 7. We have not enough room to present the complete study but threshold 20 has been chosen, using SENSEVAL-3 English all-words as test data. Moreover, our tests showed roughly 5% of improvement against not using these features.

[7] TreeTagger (Schmid, 1994) was used

Base Level Concepts (BLC) are concepts that are representative for a set of other concepts. A simple method for automatically selecting BLC from WN based on the hypernym hierarchy and the number of stored relationships between synsets have been used to define features for training a supervised system.

Although in our system BLC play a simple role aiding to the disambiguation just as additional features, the good results achieved with such simple features confirm us that an appropriate set of BLC will be a better semantic discriminator than senses or even sense-clusters.

References

E. Agirre, I. Aldezabal, and E. Pociello. 2003. A pilot study of english selectional preferences and their cross-lingual compatibility with basque. In *Proceedings of the International Conference on Text Speech and Dialogue (TSD'2003)*, CeskBudojovice, Czech Republic.

J. Atserias, L. Villarejo, G. Rigau, E. Agirre, J. Carroll, B. Magnini, and P. Vossen. 2004. The meaning multilingual central repository. In *Proceedings of Global WordNet Conference (GWC'04)*, Brno, Czech Republic.

M. Ciaramita and Y. Altun. 2006. Broad-coverage sense disambiguation and information extraction with a supersense sequence tagger. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'06)*, pages 594–602, Sydney, Australia. ACL.

M. Ciaramita and M. Johnson. 2003. Supersense tagging of unknown nouns in wordnet. In *Proceedings of the Conference on Empirical methods in natural language processing (EMNLP'03)*, pages 168–175. ACL.

J. Curran. 2005. Supersense tagging of unknown nouns using semantic similarity. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL'05)*, pages 26–33. ACL.

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

M. Hearst and H. Schütze. 1993. Customizing a lexicon to better suit a computational task. In *Proceedingns of the ACL SIGLEX Workshop on Lexical Acquisition*, Stuttgart, Germany.

Thorsten Joachims. 2002. *Learning to Classify Text Using Support Vector Machines*. Kluwer Academic Publishers.

B. Magnini and G. Cavaglia. 2000. Integrating subject fields codes into wordnet. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*.

R. Mihalcea and D. Moldovan. 2001. Automatic generation of coarse grained wordnet. In *Proceding of the NAACL workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations*, Pittsburg, USA.

I. Niles and A. Pease. 2001. Towards a standard upper ontology. In *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001)*, pages 17–19. Chris Welty and Barry Smith, eds.

W. Peters, I. Peters, and P. Vossen. 1998. Automatic sense clustering in eurowordnet. In *First International Conference on Language Resources and Evaluation (LREC'98)*, Granada, Spain.

E. Rosch. 1977. Human categorisation. *Studies in Cross-Cultural Psychology*, I(1):1–49.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of NemLap-94*, pages 44–49, Manchester, England.

F. Segond, A. Schiller, G. Greffenstette, and J. Chanod. 1997. An experiment in semantic tagging using hidden markov model tagging. In *ACL Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, pages 78–81. ACL, New Brunswick, New Jersey.

L. Villarejo, L. Màrquez, and G. Rigau. 2005. Exploring the construction of semantic class classifiers for wsd. In *Proceedings of the 21th Annual Meeting of Sociedad Espaola para el Procesamiento del Lenguaje Natural SEPLN'05*, pages 195–202, Granada, Spain, September. ISSN 1136-5948.

P. Vossen, L. Bloksma, H. Rodriguez, S. Climent, N. Calzolari, A. Roventini, F. Bertagna, A. Alonge, and W. Peters. 1998. The eurowordnet base concepts and top ontology. Technical report, Paris, France, France.