The TALP Systems for Disambiguating WordNet Glosses

Mauro Castillo, Francis Real, Jordi Atserias and German Rigau[†]
TALP Research Center. UPC. {castillo,fjreal,batalla}@lsi.upc.es
[†] IXA Group. UPV/EHU. rigau@si.ehu.es

1 Introduction

This paper describes the TALP systems presented at Senseval-3 task 12 "Word-Sense Disambiguation of WordNet Glosses". Our method combines a set of knowledge-based heuristics integrating several information sources and techniques.

Using large scale lexico–semantic knowledge bases, such as WN, has become a usual, often necessary, practice for most current Natural Language Processing systems. Building appropriate resources of this nature for broad–coverage semantic processing is a hard and expensive task, involving large research groups during long periods of development. For example, dozens of person–years are been invested world–wide into the development of wordnets for various languages (Fellbaum, 1998), (Atserias et al., 1997), (Agirre et al., 2002), (Pianta et al., 2002).

Dictionaries are special texts describing the meaning of a language. They provide a wide range of information of words by giving definitions of the word senses and as, a side effect, they supply knowledge about the world itself.

WordNet (WN) (Fellbaum, 1998) can be also seen as an structured dictionary with thouthands of semantic relations, defining the most common concepts of the English language. Although the importance of (WN) has widely exceeded the purpose of its creation (Miller et al., 1990), and it has become an essential semantic resource for many applications, at the moment is not rich enough to directly support advanced semantic processing (Harabagiu et al., 1999).

Sense disambiguation of definitions in any lexical resource is an important objective in the language engineering community because this process can increase the semantic conectivity among concepts. The first significant disambiguation of dictionary definitions took place 20 years ago (see (Rigau, 1998) for an extended survey on acquiring lexical knowledge from Machine Readable Dictionaries). Recently, several research groups have presented different approaches to perform this process on WN.

In the eXtended WordNet[1] (Mihalcea and Moldovan, 2001) the WN glosses have been syntactically parsed, transformed into logic forms and the content words are also semantically disambiguated. Being derived from an automatic process, disambiguated words included into the glosses have assigned a confidence label indicating the quality of the annotation (gold, silver or normal).

The OntoWordNet project aims to achieve a formal specification of WN. As an intermediate step, they also apply an automatic WSD system to the wordnet glosses (Gangemi et al., 2003). In this case, they use also a set of heuristics but in an iterative and incremental process.

The Senseval-3 task 12 "Word-Sense Disambiguation of WordNet Glosses"[2] has been designed as an "all-words" task using as a gold standard the hand–tagged words provided by the XWN.

[1] http://xwn.hlt.utdallas.edu/
[2] http://www.clres.com/SensWNDisamb.html

2 The TALP Systems

Our main goal is to build a robust WSD system based initially on the main heuristics of (Mihalcea and Moldovan, 2001; Novischi, 2002; Gangemi et al., 2003), but considering the current content of the MEANING[3] Multilingual Central Repository (MCR) (Atserias et al., 2004) (i.e. SUMO, MultiWordNet Domains, etc.).

Given a word from a gloss, each heuristic votes for different synsets.

The program simply adds up the votes for each word sense, selecting the most voted sense.

We have presented two different systems using two different preprocess.

- PRE-XWN (XWN preprocess) uses the gloss segmentation and PoS tagging provided by the XWN.

- In PRE-TALP (TALP preprocess) first, the sentences are tokenized and then passed on to a multiword identification module. Then, the output containing the multiwords is POS tagged using Eric Brill's tagger (Brill, 1995). Tagged words and multiword expressions are lemmatized using WN.

PRE-XWN 00256298n the restoration#n(s) of run-down#a(g) urban#a(n) areas#n(n) by the middle#a class#n(n) (resulting#v(n) in the displacement#n of lower#a(n) - income#n(s) people#n(n))
PRE-TALP 00256298n the restoration#n of run-down#v urban-areas#n by the middle-class#n (resulting#v in the displacement#n of lower#a - income#n people#n)

Table 1: PRE-XWN and PRE-TALP Example

Table 1 shows an example of the different behaviours of both preprocessing systems. PRE-TALP recognizes the Multi Word Expression *urban-area* and *middle-class*, while XWN splited then in several words. On the other hand, the tagger did not recognize *run-down* as an adjective. Obviously, different segmentation and tagging preprocessing causes different word counts and scoring.

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

2.1 Heuristics

The main heuristics used in the disambiguation process are:

1. **Monosemous**: Applying a closed–world assumption, this heuristic identifies monosemous words.

2. **Most Frequent**: Based on WN2.0 sense frequencies, this heuristic only selects those synsets having frequencies higher than the 85% of the most frequent senses.

3. **Hypernym**: This method follows the hypernym chain looking for words appearing in the gloss (e.g. the genus term).

4. **WordNet Relations**: This heuristic follows any synset relation looking for words appearing in its gloss. The method does not only use direct relations, but also performs a chaining search following all relations and stopping at distance five.

5. **MultiWordNet Domains** (Magnini and Cavagli, 2000): Having a synset with a particular WN Domain label, this method selects those synsets from the words of the gloss having the same Domain label.

6. **Patterns**: This method uses the "One sense per collocation" heuristic (Yarowsky, 1993), implementing those patterns appearing in (Novischi, 2002).

7. **Lexical Parallelism**: This heuristic identifies the words with the same part of speech separated by comas or conjunctions and marks them, when possible, with senses that belong to the same hierarchy.

8. **SUMO** (Niles and Pease, 2001): Having a synset with a particular SUMO label, this method selects those synsets from the words of the gloss having the same SUMO label.

9. **Category**: Having a synset being connected to a particular WN CATEGORY, this method selects those synsets from the words of the gloss connected the same CATEGORY.

10. **Bigram**: This heuristic uses high frequency word sense pairs occurring in SemCor.

11. **Sense One**: Finally, this heuristic always assigns the first WN sense.

3 Test Data

The test set consist of 15,179 gold assigments form 9,257 glosses taked directly from XWN2.0-1.1 (see table 2).

POS	words	gold
Noun	35539	10985
Verb	2863	2105
Adj	370	263
Adv	3719	1826
Total	42491	15179

Table 2: Test Senseval3 (9257 gloss)

However, this version is not free of errors and inconsistencies. For instance, XWN has 724 word tagged senses not belonging to WN. Three of them labelled as gold.

Furthermore, as we can see in table 3 the synset distributions of the test data and WN2.0 are very different. In particular for adjective and adverbs. Being the test data not representative of WN2.0, this test set missleads the global results and the final goal of the task.

	WordNet 2.0		Test Data	
POS	Gloss	%	Gloss	%
Noun	79689	69.0	6706	72.4
Verb	13508	11.7	773	8.4
Adj	18563	16.0	94	1.0
Adv	3664	3.2	1684	18.2
Total	115424	100	9257	100

Table 3: Synset distributions of WN2.0 and Test

4 Results

The final results of both TALP systems are presented in table 4. Obviously, the performance of PRE-TALP is lower than PRE-XWN because it uses a different preprocess. The difference in the amount of words must to the preprocess (different tokenization, PoS tagging). From the ten systems presented at the task, PRE-XWN has obtained the first position of recall (10,363 correct gold assigments, 68.3%) and PRE-TALP the third position (8,466 correct gold assigments, 55.8%).

PRE-XWN					
	Noun	Verb	Adj	Adv	Total
Correct	7788	1191	134	1246	10363
Attempted	10981	2105	263	1717	15102
Total	10985	2105	263	1826	15179
Precision	70.9%	56.6%	51.0%	72.6%	68.6%
Recall	70.9%	56.6%	51.0%	68.2%	68.3%
% Attemp.	100%	100%	100%	94.0%	99.5%

PRE-TALP					
	Noun	Verb	Adj	Adv	Total
Correct	6076	979	130	1260	8466
Attempted	9339	2000	253	1746	14757
Total	10985	2105	263	1826	15179
Precision	65.1%	48.9%	51.4%	72.2%	57.4%
Recall	55.3%	46.5%	49.4%	69.0%	55.8%
% Attemp.	85.0%	95.0%	96.2%	95.6%	97.2%

Table 4: Results of **PRE-XWN** and **PRE-TALP**

Table 5 presents the final results per heuristic. As espected, each heuristic has different behaviour of (P) precision and (R) recall. However, none of them has higher performance than its combination.

PRE-XWN					
Heuristic	Corr	Attem	P	R	Attem
Monos	131	140	93.6%	0.9%	0.9%
MostFre	7741	13903	55.7%	51.0%	91.6%
Hyper	2003	2271	88.2%	13.2%	15.0%
Relations	5243	8054	65.1%	34.5%	53.1%
Domains	2873	4119	69.7%	18.9%	27.1%
Pattern	709	753	94.2%	4.7%	5.0%
LexPar	756	1360	55.6%	5.0%	9.0%
Sumo	2334	4181	55.8%	15.4%	27.5%
Category	38	64	59.4%	0.3%	0.4%
Bigram	1903	3305	57.6%	12.5%	21.8%
SenseOne	8338	15093	55.2%	54.9%	99.4%

PRE-TALP					
Heuristic	Corr	Attem	P	R	Attem
Monos	8	86	9.3%	0.1%	0.6%
MostFre	6061	12340	49.1%	39.9%	81.3%
Hyper	1845	2082	88.6%	12.2%	13.7%
Relations	4538	7443	61.0%	29.9%	49.0%
Domains	1856	3096	59.9%	12.2%	20.4%
Pattern	712	757	94.1%	4.7%	5.0%
LexPar	590	1143	51.6%	3.9%	7.5%
Sumo	2172	3942	55.1%	14.3%	26.0%
Category	34	62	54.8%	0.2%	0.4%
Bigram	1361	2692	50.6%	9.0%	17.7%
SenseOne	6538	13403	48.8%	43.1%	88.3%

Table 5: Per heuristic results

5 Conclusions and Future Work

It is our belief, following (McRoy, 1992) and (Rigau et al., 1997), that full-fledged lexical ambiguity resolution should integrate several information sources and techniques. Our heuristics used most of the information content coherently integrated within the Multilingual Central Repository (MCR) of MEANING (Atserias et al., 2004), one of the richest and largest multilingual lexical knowledge base in existence.

In order to improve the current systems, we plan to enrich the current set of heuristics using other knowledge uploaded into the MCR with a more robust preprocessing schema.

Acknowledgments

This research has been partially funded by the European Commission (Meaning Project, IST-2001-34460), and by the Spanish Research Ministry (Hermes Project: TIC2000-0335-C03-02), Generalitat de Catalunya (2002FI 00648) and Universidad Tecnológica Metropolitana (Chile).

References

E. Agirre, O. Ansa, X. Arregi, J.M. Arriola, A. Diaz de Ilarraza, E. Pociello, and L. Uria. 2002. Methodological issues in the building of the basque wordnet: quantitative and qualitative analysis. In *Proceedings of the first International WordNet Conference in Mysore*, India, 21-25 January.

J. Atserias, S. Climent, X. Farreres, G. Rigau, and H. Rodríguez. 1997. Combining multiple methods for the automatic construction of multilingual wordnets. In *Proceedings of RANLP'97*, pages 143–149, Bulgaria.

Jordi Atserias, Luís Villarejo, German Rigau, Eneko Agirre, John Carroll, Bernardo Magnini, and Piek Vossen. 2004. The meaning multilingual central repository. In *Proceedings of the Second International Global WordNet Conference (GWC'04)*, Brno, Czech Republic, January. ISBN 80-210-3302-9.

E. Brill. 1995. Transformation-based error-driven learning and natural language processing: A case study in part of speech tagging. *Computational Linguistics*, 21(4).

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

A. Gangemi, R. Navigli, and P. Velardi. 2003. Axiomatizing wordnet glosses in the ontowordnet project. In *Proceedings of 2nd International Semantic Web Conference Workshop on Human Language Technology for the Semantic Web and Web Services*, Sanibel Island, Florida.

S. Harabagiu, G. Miller, and D. Moldovan. 1999. Wordnet 2 - a morphologically and semantically enhanced resource. In *Proceedings of ACL on Standardizing Lexical Resources (SIGLEX'99)*, Maryland, MD.

B. Magnini and G. Cavagli. 2000. Integrating subject field codes into wordnet. In *In Proceedings of the Second Internatgional Conference on Language Resources and Evaluation LREC'2000*, Athens. Greece.

Susan Weber McRoy. 1992. Using multiple knowledge sources for word sense discrimination. *Computational Linguistics*, 18(1):1–30.

R. Mihalcea and D. Moldovan. 2001. extended wordnet: Progress report. In *Proceedings of NAACL Workshop on WordNet and Other Lexical Resources*, Pittsburgh, PA.

G. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. Miller. 1990. Five Papers on WordNet. *Special Issue of International Journal of Lexicography*, 3(4).

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

A. Novischi. 2002. Accurate semantic annotations via pattern matching. In *Florida Artificial Intelligence Research Society (FLAIRS'02)*, Pensacola, Florida.

E. Pianta, L. Bentivogli, and C. Girardi. 2002. Multiwordnet: developing an aligned multilingual database. In *First International Conference on Global WordNet*, Mysore, India.

G. Rigau, J. Atserias, and E. Agirre. 1997. Combining unsupervised lexical knowledge methods for word sense disambiguation. In *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.

G. Rigau. 1998. *Automatic Acquisition of Lexical Knowledge from MRDs*. Ph.D. thesis, Departament de LSI. Universitat Politecnica de Catalunya.

D. Yarowsky. 1993. One sense per collocation. In *Proceedings, ARPA Human Language Technology Workshop*, Princeton.