Procesamiento del Lenguaje Natural, núm. 35 (2005), pp. 195-202 recibido 02-05-2005; aceptado 01-06-2005

Exploring the construction of semantic class classifiers for WSD *

Luis Villarejo y **Lluís Màrquez**
TALP Research Center,UPC
C/Jordi Girona, 1-3,
08034 Barcelona
lluism@lsi.upc.edu

German Rigau
IXA Group
Euskal Herriko Unibersitatea
649 pk. 20.080, Donostia
rigau@si.ehu.es

Resumen: El objetivo de este artículo es dar a conocer la metodología, los resultados y las futuras lineas de investigación que se derivan de una novedosa aproximación a la tarea de Word Sense Disambiguation(WSD). Dicha aproximación consiste, a grandes rasgos, en la construcción de clasificadores para distintas clases semánticas y su posterior combinación. De esta forma, esperamos contar no sólo con sistemas de WSD con una granularidad más gruesa que la ofrecida comunmente por los sistemas basados en sentidos de WordNet, sino también con sistemas que ofrezcan distintas perspectivas del problema de forma que su combinación resulte beneficiosa para el resultado final de la tarea.

Palabras clave: WSD, ontologías, aprendizaje automático, Semantic Fields, SUMO

Abstract: The aim of this paper is describing the experiments, results achieved and further work, in a novel approach to the Word Sense Disambiguation(WSD) task. This novel approach consists, mainly, in the learning and combination of several semantic class classifiers. So we can not only get WSD systems with coarser granularity than the traditionally offered by WordNet senses, but also systems showing different views of the task which allows us to improve the overall task results.

Keywords: WSD, ontologies, machine learning, Semantic Fields, SUMO

1. *Introduction and Background*

Word Sense Disambiguation (WSD) is an intermediate Natural Language Processing (NLP) task which consists in assigning the correct semantic interpretation to each word in a text. One of the most successful approaches in the last years is the *supervised learning from examples*, in which statistical or Machine Learning classification models are induced from semantically annotated corpora. Generally, supervised systems have obtained better results than the unsupervised ones, as shown by experimental work and international evaluation exercises such as Senseval[1]. These annotated corpora are usually manually tagged by lexicographers with word senses taken from a particular lexical semantic resource –most commonly WordNet (Miller, 1990) (Fellbaum, 1998).

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. Thus, some work has been focused on constructing sense clusters to overcome this fine-grained distinction (Peters and Peters, 2000) (Mihalcea and Moldovan, 2001) (Agirre and Lopez de Lacalle, 2003). However, not so much attention has been paid to learn semantic classes such as: Lexicographical Files (or Semantic Fields of WordNet), WordNet Domains (Magnini and Cavaglia, 2000), SUMO Ontology (Niles and Pease, 2001), Base Concepts, and Top Concept Ontology (Vossen et al., 1998) (Atserias, Climent, and Rigau, 2004). All these semantic classes hold a coarser granularity than the one that WorNet offers and are already available and integrated with WordNet in the Multilingual Central Repository (Atserias et al., 2004), which is a result of the MEANING project (Rigau et al., 2002). At the same time, these resources group senses that are related at some level of generality using different

* This research has been supported by the European project MEANING (IST-2001-34460). The authors would also like to thank Gerard Escudero for letting us use the Feature Extraction module and Jesus Giménez for his helpful collaboration.

[1] http://www.senseval.org

ISSN: 1135-5948

semantic criteria and properties that could be of interest for WSD. This led us to think that the combination of them could improve the overall results since they offer different perspective of the task. As far as we know, only one previous work has been published on learning semantic classes(Segond et al., 1997). But, tt only concerned one isolated semantic class and no combination with other semantic class classifiers was explored. In that work, Hidden Markov Models were used to obtain similar results, although not directly comparable to ours because of differences in training and test sets used, in learning a tagger for the Semantic Fields.

In this paper we present the work done in a first approach to build Semantic Class Classifiers via supervised learning and combine them. The methodology to carry out these experiments is the same for both semantic classes, but both are developed in an independent manner. These experiments have been faced in an *all-words* way, trying to assign to each word in a text its correct semantic label. In order to avoid the data sparseness problem, the approach proposed here is training class-experts classifiers, instead of word-experts. This implies increasing the amount of available training data for each classifier but also having to deal with different words involved in each classifier.

This paper is organised as follows. The next section, briefly describes the semantic classes considered for this study. In section three, the source data and its codification is explained. Section four presents some upperbounds for the task and performs an initial evaluation and analysis of the results. And finally, in section five we proceed to discussion and section six presents the conclusions and further work.

2. Semantic Classes

Our experiments have been adressed to Semantic Fields and SUMO Ontology type classes (as a first exploration to the building of classifiers for the rest of the reported types). These classes have been chosen because of their significantly different granularities and the fact that every synset has an unique label assignment in each of them, making easier the learning process.

The Semantic Fields (or WordNet Lexicographical Files) consist of 45 categories in which WordNet synsets are organised with respect to syntactic categories and logical groupings (e.g., verb.motion, noun.person, noun.food, etc.).

The SUMO (Suggested Upper Merged Ontology) ontology (Niles and Pease, 2001) has been created as part of the IEEE Standard Upper Ontology Working Group. The goal of this Working Group is to develop a standard upper ontology that will promote data interoperability, information search and retrieval, automated inferencing, and natural language processing. An ontology is similar to a dictionary or glossary, but with greater detail and structure that enables computers to process its content. An ontology consists of a set of concepts, relations, and axioms that formalise a field of interest. An upper ontology is limited to concepts that are meta, generic, abstract or philosophical, and hence are general enough to address (at a high level) a broad range of domain areas. Concepts specific to particular domains are not included in an upper ontology, but such an ontology does provide a structure upon which ontologies for specific domains (e.g., medicine, finance, engineering, etc.) can be constructed. The current version of SUMO consists of 1,019 terms (all of them connected to WordNet 1.6 synsets), 4,181 axioms and 822 rules.

Each of these semantic classes hold a coarser granularity with respect to the one from WordNet. This leads to a lower level of ambiguity per semantic class, as can be seen in table 1, and thus to higher rates of predictability for each single problem to be learned. In addition, any of these classes would serve as a different criteria to stablish a coherent but diverse semantic class partition.

Resource	Average ambiguity
WordNet	6.81
SemFile	2.43
Sumo	1.42
Top Ontology	1.79

Cuadro 1: Average word ambiguity over SemCor-1.6.

Label	Number of examples	Average polysemy degree	Number of words	class-MFC accuracy	word-MFC accuracy
noun.animal	3015	3.11	269	20.0 %	88.2 %
noun.artifact	21575	3.46	1960	100 %	84.5 %
noun.food	2185	2.62	184	63.6 %	95,4 %
noun.person	13533	3.06	1520	86.3 %	95.3 %
verb.communication	17354	3.25	730	67.8 %	78.6 %
verb.competition	5456	5.42	157	14.0 %	58.0 %
verb.motion	11276	5.64	615	26.4 %	83.6 %
verb.possession	17406	6.73	275	9.3 %	70.0 %
...	...	...	...	...	...
Averages	9877	3.82	572	63.5 %	84.1 %

Cuadro 2: SemCor-1.6 data and baselines related information over a subset of labels.

3. Training data and feature representation

SemCor-1.6 corpus has been used as the training data to feed machine learning algorithms. Documents in SemCor-1.6 corpus were distributed into ten folders. In each of them, diversity on the topics of its documents was maintained. Each word appearing in SemCor was labelled with its semantic class tags making use of the Multilingual Central Repository. As we stated in the previous section, the learning approach taken in this work is training class-experts rather than word-experts. We choose this approach to avoid the data sparseness problem faced when developing word-experts for tasks such as an all-words with little corpus like SemCor for training. Examples for each class where arranged in the following manner. Each training example of word W tagged with label L is given as a positive example to the L-class and as a negative example to the rest of possible classes of the word W. In this way, an independent training set was generated for each Semantic Class label present in SemCor-1.6 corpus. This makes a total of 45 independent learning problems in the case of Semantic Fields, and 657 for the SUMO Ontology. Figures about the corpus average polysemy degree, number of words and number of examples are shown in table 2.

Regarding feature representation of the training examples, we have designed two different approaches. Both of them were applied to experiments with Semantic Fields and SUMO Ontology. The first one consists of a rich feature set (RICH-FS) obtained using the Feature Extraction module of the TALP team in the Senseval-3 English lexical sample (Escudero, Màrquez, and Rigau, 2004) and all-words task (Villarejo et al., 2004). The feature set includes the classic window–based pattern features extracted from a local context (up to a maximum of ten words on left and right sides of the target word) and the “bag–of–words” type of features taken from a broader context (the preceding and following sentences). It also contains, making use of MiniPar (Lin, 1994) which is a broad-coverage parser for the English language, a set of features representing the syntactic relations involving the target word, and semantic features of the surrounding words extracted from the Multilingual Central Repository of the MEANING project. The second one, which was designed to set a baseline to learn the most frequent class classifier, is a reduced feature set (RED-FS) consists of three features relative to the target word: one containing its lemma, another one holding its POS and the last one standing for its most frequent semantic class tag calculated over the training folders in SemCor-1.6.

4. Evaluation and Results

To face the task of building the classifiers, we have used two different supervised learning systems:

- **Support Vector Machines** (SVM) is a technique based on finding the hyperplane (in a high dimensional feature space) that separates with maximum margin the positive examples from the negatives, i.e., the maximal margin hyperplane. We used SVMlight (Joachims, 1999), which is the freely available implementation by Joachims, T., linear kernels, and one-vs-all binarization.
- **AdaBoost** (ADB) (Schapire and Singer, 1999) is a method for learning an ensemble of weak classifiers and combine

them into a strong global classification rule. The software we used (Carreras and Màrquez, 2001) implements the ADB algorithm with real-valued confidence weak rules, which are decision trees with user-defined maximum depth.

Test corpus for these classifiers has been a randomly chosen folder containing 19,700 target word occurrences.

Two baselines, based on frequencies and extracted from the nine training folders, have been calculated for each Semantic Class. The first baseline is a class-based most frequent class (MFC) which has been calculated by building a ranking on the frequencies of each semantic label in the nine training folders. To apply this baseline to a word, all the possible semantic labels for this word are obtained and the one with a higher frequency is chosen. The second one, a word-based most frequent class, has been calculated by obtaining the most frequent semantic label for each word appearing in the nine training folders. Table 3 endorses with results the idea that the word-based MFC should perform better than the class-based for both semantic classes because of its more informed construction.

	Class-based MFC	Word-based MFC
SemFile	63.5 %	84.1 %
SUMO	56.7 %	77.9 %

Cuadro 3: Accuracies achieved by the baselines on the SemCor-1.6 test folder.

Regarding the learning algorithms, several experiments have been performed over the test corpus with different parameters. For simplicity reasons, table 4 shows only the results on the most promising learning algorithm configurations. These are, for SVM, a linear kernel with a 0.01 value for the c regularisation parameter. And, for ADB, a configuration consisting of learning stumps (depth of the weak rules equal to zero) and using just two weak rules to classify for the RED-FS and up to two thousand weak rules for the RICH-FS. The fact of using just two weak rules for the reduced feature approach is motivated by its particular codification of the examples and the fact that we want to learn the MFC which is held in just one of the codified attributes. First thing we realize when looking at table 4 is the contrast in results between the two different feature representations. The difference between the two features set results is made by the most frequent class attribute. In principle, although not including an attribute like the most frequent class, a richer set of features should be able to help in getting higher accuracies than the ones from the RED-FS. This is not shown in these first experiments, where the RED-FS performance is better than the richer one. This leads us to think that redundancy and irrelevance of some attributes must be controlled in order to provide a better feature set. Now, our RICH-FS contains a huge amount of attributes not specifically designed for this experiments but for the lexical sample task in Senseval. Also note in table 4 that, this big amount of features seems to be managed much better by the ADB algorithm than by the SVM.

	RICH-FS		RED-FS	
	SVM	ADB	SVM	ADB
SemFile	70.8 %	76.3 %	**82.5 %**	80.0 %
SUMO	59.9 %	68.3 %	**71.9 %**	68.7 %

Cuadro 4: Accuracies achieved by the learning algorithms on the test folder.

Partial figures on the accuracies, over the test folder, of the class-based and the word-based MFC's for some of the Semantic Fields labels[2] are shown in table 2. Note that, as justified before, in most cases the word-based MFC outperforms the class-based one. In table 5, accuracies of the RED-FS approach using, for both learning algorithms, the best configurations on the test folder is also shown in detail. In contrast with the RICH-FS approach, SVM seems to perform better than AdaBoost when few attributes are used. To have an idea on how difficult is each single problem, figures on the average polysemy degree are shown in table 2. As expected, there is a well defined correlation between the polysemy degree of a class and its accuracy. As can be seen in figure 1 for Semantic Fields, the more the polysemy inside a class, the worse the accuracy. Finally, in table 7 results for Semantic

[2] Only overall, but not detailed, results on SUMO labels are provided in this paper for clarity reasons and due to the fact that they don't seem to behave different from Semantic Fields labels.

Fields from table 5 can be found grouped by POS and directly contrasted with figures on the polysemy degree inside each class (note that adverbs are not shown since their accuracy is always maximum because there is only one label for them). Looking carefully at figures in this table, it might be surprising the fact that, for adjectives, which hold the lowest average polysemy degree, algorithms perform worse than for nouns and verbs holding higher average polysemy degree. This is explained due to the fact that the average polysemy degree shown in the table is calculated among all the adjectives, polysemic and monosemic.

Label	SVM	AdaBoost
noun.animal	87.0 %	87.0 %
noun.artifact	80.1 %	76.9 %
noun.food	95.4 %	95.4 %
noun.person	95.6 %	95.3 %
verb.communication	76.5 %	63.1 %
verb.competition	56.0 %	54.0 %
verb.motion	64.8 %	61.0 %
verb.possession	43.3 %	26.7 %
...	...	...
Total acc.	82.5 %	79.9 %

Cuadro 5: Detailed accuracies using the RED-FS over the test folder.

Combining the output of ranked classifiers on the same word (for both semantic classes, Semantic Fields and SUMO), we can realize that, for some words, the assigned combination of both labels is not present in any of the possible senses of the word (we will call this combination an incompatibility for that word). In table 6 we summarize all the experiments done and show the percentage of words instances that have been asigned an incompatibility. A percentage equal to 0 %, like in the Word-based MFC, means that for all the words the predicted combination of Semantic Fields and SUMO labels is possible. A percentage of incompatibilities above zero stands for the cases in which one or both classifiers have predicted incorrectly its semantic label because this combination of Semantic Field and SUMO labels is not possible for any of the senses of this word. So, improvements in the overall results can be achieved by not allowing incompatibilities in the outputs.If we discard every incompatible combination and assign the next more probable combination of labels predicted, results in both semantic classes improve in 1.5 %. For Semantic Fields, this means a 83.9 % accuracy, nearly reaching the 84.1 % of the word-based MFC. This process, in which incompatibilities are erased from the results, can be applied on the top of the output of any kind of system or heuristic performing ranked predictions on both semantic classes.

Outputs		Incompatibilities
Red. feat. set	SVM	11.6 %
(RED-FS)	ADB	15.4 %
Rich. feat. set	SVM	26.5 %
(RICH-FS)	ADB	21.3 %
Word-based MFC		0 %
Class-based MFC		29.6 %

Cuadro 6: Incompatibilities inside the combined predicted outputs.

5. *Discussion*

In the reduced approach, two examples of the same word (and not necessarily tagged with the same Semantic Class label) are codified exactly in the same way. Thus, the job to be carried by the learning algorithm with this approach is classifying words instead of senses, predicting then the same label for a given word, despite its context. So, the upper bound for this RED-FS should be the word-based MFC, unless the distribution of the test folder would not match the word-based MFC. As it is shown in tables 3 and 4, in these first experiments performance with the RED-FS clearly outperforms the class-based MFC but is a little under the word-based MFC. In concrete, for Semantic Fields is 1.6 points below and for SUMO is 6 points below. This can be due to the fact that the discriminative patterns for each word inside a semantic class are very different and must be learned from a unique training set. This would mean that is very difficult to generalise across words inside a class and thus, grouping words in the class-based approach only would add difficulty to the learning problem. Also note that, SVM performs better than ADB in this RED-FS. More efforts in this direction have been carried out by introducing new attributes able to capture the context, or at least, producing different codifications for examples of the same word and thus, differentiating them. These attributes have been: bigrams of lemmas and POS, bag of words of lemmas (in a 10 and 20 word window), attributes on the rest of semantic classes and syntactic relations involving the target word.

POS	Polysemy	avPolyDegree	SVM Acc-P	SVM Acc-T	ADB Acc-P	ADB Acc-T
adj	10.3 %	2.00	61.1 %	95.9 %	51.0 %	94.9 %
noun	71.1 %	3.54	70.1 %	78.7 %	66.7 %	76.3 %
verb	86.6 %	4.45	66.9 %	71.3 %	60.9 %	66.1 %
Weighted total	55.7 %	3.82	68.6 %	82.5 %	64.0 %	79.9 %

Cuadro 7: Detailed accuracies on Semantic Field, grouped by POS, using the RED-FS over the test folder. *Acc-P* stands for 'accuracy inside polysemic words' and *Acc-T* stands for 'total accuracy'.

Results on this new codification did not outperform the ones obtained with the RED-FS, leading our conclusions to the need for new features which could help to generalise better in order to overcome, if possible, the difficulty of generalising across-words inside a class. In fact, comparing the reduced approach with the word-based MFC is not fair because we are training class classifiers and not word classifiers. A more fair comparison is contrasting the RED-FS approach (or even the RICH-FS approach, which lacks from the most frequent class attribute) with the class-based MFC. And, as can be seen in table 3 and 4, in this cases our learning clearly outperforms the baselines.

6. Conclusion and Further Work

We have presented a novel approach to the WSD task based in learning semantic class classifiers for two semantic classes and combining them to improve overall results. Experiments presented here show how classifiers with very little information nearly reach the baselines for the task, which are difficult to outperform, and how combination of these two different classifiers improve results on both semantic classes.Building the classifiers we choosed a class-based approach because it allowed us to avoid the data sparseness problem we would have with the word-based approach. But this approach also increases the complexity of each classifier by having to generalise for hundreds of different words in a single classifier as can be seen in table 2 for the Semantic Fields. Thus, the word-based approach, as it is draw by the performance of the two baselines, is more likely to offer a higher performance than the offered by the class-based approach as long as a considerable number of words have enough training data in the corpus.

Next steps to be taken now are: 1) enrich the RED-FS with more carefully chosen attributes to try to overcome the difficulty we have found in generalising across words inside a class and leading to an best performance over the word-based MFC, 2) explore the word-based approach building classifiers for every word, and 3) combine the class and word-based approaches substituting the word-based classifier for the most frequent classes when not enough training data is found for an specific word.

References

Agirre, E. and O. Lopez de Lacalle. 2003. Clustering wordnet word senses. In *Proceedings of the Conference on Recent Advances on Natural Language Processing (RANLP'03)*, Borovets, Bulgary.

Atserias, J., S. Climent, and G. Rigau. 2004. Towards the meaning top ontology: Sources of ontological meaning. In *Proceedings of the Fourth International Conference on Language Resources and Evaluations (LREC'04)*, Lisbon, Portugal.

Atserias, Jordi, 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.

Carreras, Xavier and Luís Màrquez. 2001. Boosting trees for clause splitting. In *Proceedings of CoNLL-2001*, pages 73–75. Toulouse, France.

Escudero, G., L. Màrquez, and G. Rigau. 2004. Talp system for the english lexical sample task. In Rada Mihalcea and Phil Edmonds, editors, *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 113–116, Barcelona, Spain, July. Association for Computational Linguistics.

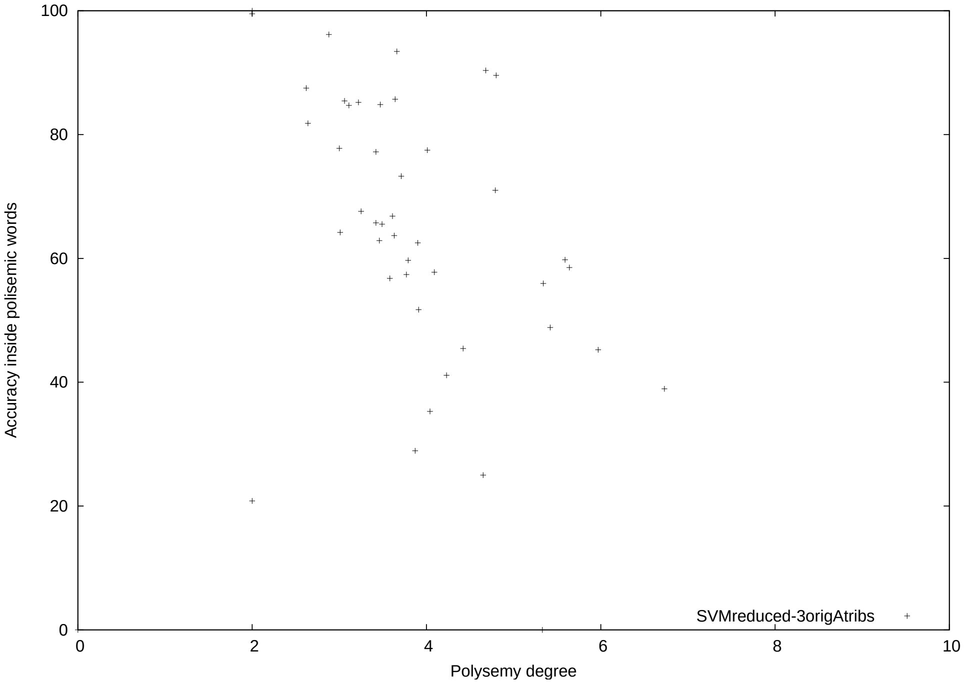

Figura 1: Relation between SVM-reduced-feature-set accuracy inside polysemic words and Polysemy degree for Semantic Fields.

Fellbaum, C. 1998. Wordnet: An electronic lexical database. MIT Press, Cambridge, MA, USA.

Joachims, Thorsten. 1999. Making large-scale svm learning practical. In B. Schlkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*. MIT Press.

Lin, D. 1994. Principar - an efficient, broad-coverage, principle-based parser. In *COLING-94*, pages 42–48, Kyoto, Japan.

Magnini, B. and G. Cavaglia. 2000. Integrating subject field codes into wordnet. In M. Gavrilidou, G. Crayannis, S. Markantonatu, S. Piperidis, and G. Stainhaouer, editors, *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, pages 1413–1418, Athens, Greece.

Mihalcea, R. and D. Moldovan. 2001. Automatic generation of a coarse grained wordnet. In *Proceedings of the Workshop on WordNet and Other Lexical Resources(NAACL'01)*.

Miller, G. A. 1990. Wordnet: An on-line lexical database. *International Journal of Lexicography*, 3(4):235–312.

Niles, I. and A. Pease. 2001. Towards a standard upper ontology. In Chris Welty and Barry Smith, editors, *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS'01)*, Ogunquit, Maine, October 17-19.

Peters, W. and I. Peters. 2000. Automatic sense clustering in eurowordnet. In *Proceedings of the International Conference on Language resources and Evaluation (LREC'00)*, Granada, Spain.

Rigau, G., B. Magnini, E. Agirre, P. Vossen, and J. Carroll. 2002. Meaning: A roadmap to knowledge technologies. In *Proceedings of COLING'02 Workshop*, Taipei, Taiwan.

Schapire, R. E. and Y. Singer. 1999. Improved boosting algorithms using confidence-rated predictions. *Machine Learning*, 37(3):297–336.

Segond, F., A. Schiller, G. Grefenstette, and J-P. Chanod. 1997. An Experiment in Semantic Tagging Using Hidden Markov Model Tagging. pages 78–81.

Villarejo, L., L. Màrquez, E. Agirre, D. Martínez, B. Magnini, C. Strapparava, D. McCarthy, A. Montoyo, and A. Suárez.

2004. The "meaning"system on the english all-words task. In Rada Mihalcea and Phil Edmonds, editors, *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 253–256, Barcelona, Spain, July. Association for Computational Linguistics.

Vossen, P., 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, EuroWordNet, LE2-4003.