SemEval-2010 Task 17: All-words Word Sense Disambiguation on a Specific Domain

Eneko Agirre, Oier Lopez de Lacalle
IXA NLP group
UBC
Donostia, Basque Country
{e.agirre,oier.lopezdelacalle}@ehu.es

Christiane Fellbaum
Department of Computer Science
Princeton University
Princeton, USA
fellbaum@princeton.edu

Shu-Kai Hsieh
Department of English
National Taiwan Normal University
Taipei, Taiwan
shukai@ntnu.edu.tw

Maurizio Tesconi
IIT
CNR
Pisa, Italy
maurizio.tesconi@iit.cnr.it

Monica Monachini
ILC
CNR
Pisa, Italy
monica.monachini@ilc.cnr.it

Piek Vossen, Roxanne Segers
Faculteit der Letteren
Vrije Universiteit Amsterdam
Amsterdam, Netherlands
p.vossen@let.vu.nl,roxane.segers@gmail.com

Abstract

Domain portability and adaptation of NLP components and Word Sense Disambiguation systems present new challenges. The difficulties found by supervised systems to adapt might change the way we assess the strengths and weaknesses of supervised and knowledge-based WSD systems. Unfortunately, all existing evaluation datasets for specific domains are lexical-sample corpora. This task presented all-words datasets on the environment domain for WSD in four languages (Chinese, Dutch, English, Italian). 11 teams participated, with supervised and knowledge-based systems, mainly in the English dataset. The results show that in all languages the participants where able to beat the most frequent sense heuristic as estimated from general corpora. The most successful approaches used some sort of supervision in the form of hand-tagged examples from the domain.

1 Introduction

Word Sense Disambiguation (WSD) competitions have focused on general domain texts, as attested in previous Senseval and SemEval competitions (Kilgarriff, 2001; Mihalcea et al., 2004; Snyder and Palmer, 2004; Pradhan et al., 2007). Specific domains pose fresh challenges to WSD systems: the context in which the senses occur might change, different domains involve different sense distributions and predominant senses, some words tend to occur in fewer senses in specific domains, the context of the senses might change, and new senses and terms might be involved. Both supervised and knowledge-based systems are affected by these issues: while the first suffer from different context and sense priors, the later suffer from lack of coverage of domain-related words and information.

The main goal of this task is to provide a multilingual testbed to evaluate WSD systems when faced with full-texts from a specific domain. All datasets and related information are publicly available from the task websites[1].

This task was designed in the context of Kyoto (Vossen et al., 2008)[2], an Asian-European project that develops a community platform for modeling knowledge and finding facts across languages and cultures. The platform operates as a Wiki system with an ontological support that social communities can use to agree on the meaning of terms in specific domains of their interest. Kyoto focuses on the environmental domain because it poses interesting challenges for information sharing, but the techniques and platforms are

[1] http://xmlgroup.iit.cnr.it/SemEval2010/ and http://semeval2.fbk.eu/

[2] http://www.kyoto-project.eu/

Proceedings of the 5th International Workshop on Semantic Evaluation, ACL 2010, pages 75–80,
Uppsala, Sweden, 15-16 July 2010.

independent of the application domain.

The paper is structured as follows. We first present the preparation of the data. Section 3 reviews participant systems and Section 4 the results. Finally, Section 5 presents the conclusions.

2 Data preparation

The data made available to the participants included the test set proper, and background texts. Participants had one week to work on the test set, but the background texts where provided months earlier.

2.1 Test datasets

The WSD-domain comprises comparable all-words test corpora on the environment domain. Three texts were compiled for each language by the European Center for Nature Conservation[3] and Worldwide Wildlife Forum[4]. They are documents written for a general but interested public and involve specific terms from the domain. The document content is comparable across languages. Table 1 shows the numbers for the datasets.

Although the original plan was to annotate multiword terms, and domain terminology, due to time constraints we focused on single-word nouns and verbs. The test set clearly marked which were the words to be annotated. In the case of Dutch, we also marked components of single-word compounds. The format of the test set followed that of previous all-word exercises, which we extended to accommodate Dutch compounds. For further details check the datasets in the task website.

The sense inventory was based on publicly available wordnets of the respective languages (see task website for details). The annotation procedure involved double-blind annotation by experts plus adjudication, which allowed us to also provide Inter Annotator Agreement (IAA) figures for the dataset. The procedure was carried out using KAFnotator tool (Tesconi et al., 2010). Due to limitations in resources and time, the English dataset was annotated by a single expert annotator. For the rest of languages, the agreement was very good, as reported in Table 1.

Table 1 includes the results of the random baseline, as an indication of the polysemy in each dataset. Average polysemy is highest for English, and lowest for Dutch.

[3] http://www.ecnc.org

[4] http://www.wwf.org

	Total	Noun	Verb	IAA	Random
Chinese	3989	754	450	0.96	0.321
Dutch	8157	997	635	0.90	0.328
English	5342	1032	366	n/a	0.232
Italian	8560	1340	513	0.72	0.294

Table 1: Dataset numbers, including number of tokens, nouns and verbs to be tagged, Inter-Annotator Agreement (IAA) and precision of random baseline.

	Documents	Words
Chinese	58	455359
Dutch	98	21089
English	113	2737202
Italian	27	240158

Table 2: Size of the background data.

2.2 Background data

In addition to the test datasets proper, we also provided additional documents on related subjects, kindly provided by ECNC and WWF. Table 2 shows the number of documents and words made available for each language. The full list with the urls of the documents are available from the task website, together with the background documents.

3 Participants

Eleven participants submitted more than thirty runs (cf. Table 3). The authors classified their runs into supervised (S in the tables, three runs), weakly supervised (WS, four runs), unsupervised (no runs) and knowledge-based (KB, the rest of runs)[5]. Only one group used hand-tagged data from the domain, which they produced on their own. We will briefly review each of the participant groups, ordered following the rank obtained for English. They all participated on the English task, with one exception as noted below, so we report their rank in the English task. Please refer to their respective paper in these proceedings for more details.

CFILT: They participated with a domain-specific knowledge-based method based on Hopfield networks (Khapra et al., 2010). They first identify domain-dependant words using the background texts, use a graph based on hyponyms in WordNet, and a breadth-first search to select the most representative synsets within domain. In addition they added manually disambiguated around one hundred examples from the domain as seeds.

[5] Note that boundaries are slippery. We show the classifications as reported by the authors.

English

Rank	Participant	System ID	Type	P	R	R nouns	R verbs
1	Anup Kulkarni	CFILT-2	WS	0.570	0.555 ±0.024	0.594 ±0.028	0.445 ±0.047
2	Anup Kulkarni	CFILT-1	WS	0.554	0.540 ±0.021	0.580 ±0.025	0.426 ±0.043
3	Siva Reddy	IIITH1-d.l.ppr.05	WS	0.534	0.528 ±0.027	0.553 ±0.023	0.456 ±0.041
4	Abhilash Inumella	IIITH2-d.r.l.ppr.05	WS	0.522	0.516 ±0.023	0.529 ±0.027	0.478 ±0.041
5	Ruben Izquierdo	BLC20SemcorBackground	S	0.513	0.513 ±0.022	0.534 ±0.026	0.454 ±0.044
-	-	*Most Frequent Sense*	-	0.505	0.505 ±0.023	0.519 ±0.026	0.464 ±0.043
6	Ruben Izquierdo	BLC20Semcor	S	0.505	0.505 ±0.025	0.527 ±0.031	0.443 ±0.045
7	Anup Kulkarni	CFILT-3	KB	0.512	0.495 ±0.023	0.516 ±0.027	0.434 ±0.048
8	Andrew Tran	Treematch	KB	0.506	0.493 ±0.021	0.516 ±0.028	0.426 ±0.046
9	Andrew Tran	Treematch-2	KB	0.504	0.491 ±0.021	0.515 ±0.030	0.425 ±0.044
10	Aitor Soroa	kyoto-2	KB	0.481	0.481 ±0.022	0.487 ±0.025	0.462 ±0.039
11	Andrew Tran	Treematch-3	KB	0.492	0.479 ±0.022	0.494 ±0.028	0.434 ±0.039
12	Radu Ion	RACAI-MFS	KB	0.461	0.460 ±0.022	0.458 ±0.025	0.464 ±0.046
13	Hansen A. Schwartz	UCF-WS	KB	0.447	0.441 ±0.022	0.440 ±0.025	0.445 ±0.043
14	Yuhang Guo	HIT-CIR-DMFS-1.ans	KB	0.436	0.435 ±0.023	0.428 ±0.027	0.454 ±0.043
15	Hansen A. Schwartz	UCF-WS-domain	KB	0.440	0.434 ±0.024	0.434 ±0.029	0.434 ±0.044
16	Abhilash Inumella	IIITH2-d.r.l.baseline.05	KB	0.496	0.433 ±0.024	0.452 ±0.023	0.390 ±0.044
17	Siva Reddy	IIITH1-d.l.baseline.05	KB	0.498	0.432 ±0.021	0.463 ±0.026	0.344 ±0.038
18	Radu Ion	RACAI-2MFS	KB	0.433	0.431 ±0.022	0.434 ±0.027	0.399 ±0.049
19	Siva Reddy	IIITH1-d.l.ppv.05	KB	0.426	0.425 ±0.026	0.434 ±0.028	0.399 ±0.043
20	Abhilash Inumella	IIITH2-d.r.l.ppv.05	KB	0.424	0.422 ±0.023	0.456 ±0.025	0.325 ±0.044
21	Hansen A. Schwartz	UCF-WS-domain.noPropers	KB	0.437	0.392 ±0.025	0.377 ±0.025	0.434 ±0.043
22	Aitor Soroa	kyoto-1	KB	0.384	0.384 ±0.022	0.382 ±0.024	0.391 ±0.047
23	Ruben Izquierdo	BLC20Background	S	0.380	0.380 ±0.022	0.385 ±0.026	0.366 ±0.037
24	Davide Buscaldi	NLEL-WSD-PDB	WS	0.381	0.356 ±0.022	0.357 ±0.027	0.352 ±0.049
25	Radu Ion	RACAI-Lexical-Chains	KB	0.351	0.350 ±0.015	0.344 ±0.017	0.368 ±0.030
26	Davide Buscaldi	NLEL-WSD	WS	0.370	0.345 ±0.022	0.352 ±0.027	0.328 ±0.037
27	Yoan Gutierrez	Relevant Semantic Trees	KB	0.328	0.322 ±0.022	0.335 ±0.026	0.284 ±0.044
28	Yoan Gutierrez	Relevant Semantic Trees-2	KB	0.321	0.315 ±0.022	0.327 ±0.024	0.281 ±0.040
29	Yoan Gutierrez	Relevant Cliques	KB	0.312	0.303 ±0.021	0.304 ±0.024	0.301 ±0.041
-	-	*Random baseline*	-	0.232	0.232	0.253	0.172

Chinese

Rank	Participant	System ID	Type	P	R	R nouns	R verbs
-	-	*Most Frequent Sense*	-	0.562	0.562 ±0.026	0.589 ±0.027	0.518 ±0.039
1	Meng-Hsien Shih	HR	KB	0.559	0.559 ±0.024	0.615 ±0.026	0.464 ±0.039
2	Meng-Hsien Shih	GHR	KB	0.517	0.517 ±0.024	0.533 ±0.035	0.491 ±0.038
-	-	*Random baseline*	-	0.321	0.321	0.326	0.312
4	Aitor Soroa	kyoto-3	KB	0.322	0.296 ±0.022	0.257 ±0.027	0.360 ±0.038
3	Aitor Soroa	kyoto-2	KB	0.342	0.285 ±0.021	0.251 ±0.026	0.342 ±0.040
5	Aitor Soroa	kyoto-1	KB	0.310	0.258 ±0.023	0.256 ±0.029	0.261 ±0.031

Dutch

Rank	Participant	System ID	Type	P	R	R nouns	R verbs
1	Aitor Soroa	kyoto-3	KB	0.526	0.526 ±0.022	0.575 ±0.029	0.450 ±0.034
2	Aitor Soroa	kyoto-2	KB	0.519	0.519 ±0.022	0.561 ±0.027	0.454 ±0.034
-	-	*Most Frequent Sense*	-	0.480	0.480 ±0.022	0.600 ±0.027	0.291 ±0.025
3	Aitor Soroa	kyoto-1	KB	0.465	0.465 ±0.021	0.505 ±0.026	0.403 ±0.033
-	-	*Random baseline*	-	0.328	0.328	0.350	0.293

Italian

Rank	Participant	System ID	Type	P	R	R nouns	R verbs
1	Aitor Soroa	kyoto-3	KB	0.529	0.529 ±0.021	0.530 ±0.024	0.528 ±0.038
2	Aitor Soroa	kyoto-2	KB	0.521	0.521 ±0.018	0.522 ±0.023	0.519 ±0.035
3	Aitor Soroa	kyoto-1	KB	0.496	0.496 ±0.019	0.507 ±0.020	0.468 ±0.037
-	-	*Most Frequent Sense*	-	0.462	0.462 ±0.020	0.472 ±0.024	0.437 ±0.035
-	-	*Random baseline*	-	0.294	0.294	0.308	0.257

Table 3: Overall results for the domain WSD datasets, ordered by recall.

This is the only group using hand-tagged data from the target domain. Their best run ranked 1st.

IIITTH: They presented a personalized PageRank algorithm over a graph constructed from WordNet similar to (Agirre and Soroa, 2009), with two variants. In the first (IIITH1), the vertices of the graph are initialized following the ranking scores obtained from predominant senses as in (McCarthy et al., 2007). In the second (IIITH2), the graph is initialized with keyness values as in

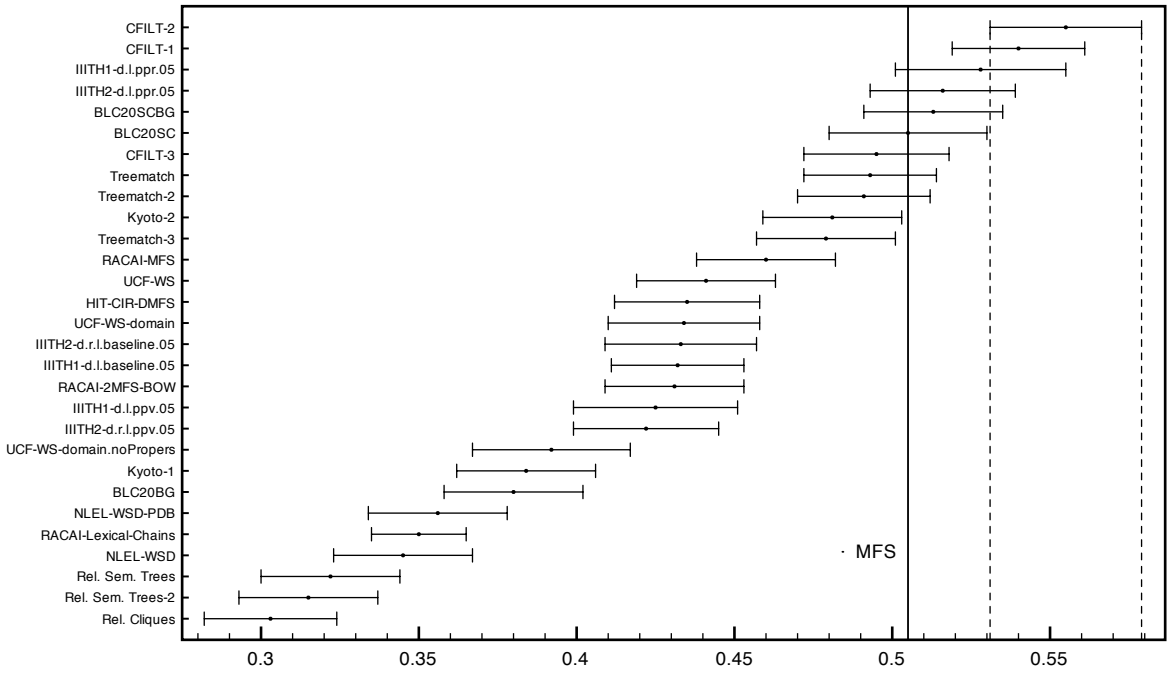

Figure 1: Plot for all the systems which participated in English domain WSD. Each point correspond to one system (denoted in axis $\mathcal{Y}$) according each recall and confidence interval (axis $\mathcal{X}$). Systems are ordered depending on their rank.

(Rayson and Garside, 2000). Some of the runs use sense statistics from SemCor, and have been classified as weakly supervised. They submitted a total of six runs, with the best run ranking 3rd.

BLC20(SC/BG/SCBG): This system is supervised. A Support Vector Machine was trained using the usual set of features extracted from context and the most frequent class of the target word. Semantic class-based classifiers were built from SemCor (Izquierdo et al., 2009), where the classes were automatically obtained exploiting the structural properties of WordNet. Their best run ranked 5th.

Treematch: This system uses a knowledge-based disambiguation method that requires a dictionary and untagged text as input. A previously developed system (Chen et al., 2009) was adapted to handle domain specific WSD. They built a domain-specific corpus using words mined from relevant web sites (e.g. WWF and ECNC) as seeds. Once parsed the corpus, the used the dependency knowledge to build a nodeset that was used for WSD. The background documents provided by the organizers were only used to test how exhaustive the initial seeds were. Their best run ranked 8th.

Kyoto: This system participated in all four languages, with a free reimplementation of the domain-specific knowledge-based method for WSD presented in (Agirre et al., 2009). It uses a module to construct a distributional thesaurus, which was run on the background text, and a disambiguation module based on Personalized PageRank over wordnet graphs. Different WordNet were used as the LKB depending on the language. Their best run ranked 10th. Note that this team includes some of the organizers of the task. A strict separation was kept, in order to keep the test dataset hidden from the actual developers of the system.

RACAI: This participant submitted three different knowledge-based systems. In the first, they use the mapping to domains of WordNet (version 2.0) in order to constraint the domains of the content words of the test text. In the second, they choose among senses using lexical chains (Ion and Stefanescu, 2009). The third system combines the previous two. Their best system ranked 12th.

HIT-CIR: They presented a knowledge-based system which estimates predominant sense from raw test. The predominant senses were calculated with the frequency information in the provided background text, and automatically constructed

thesauri from bilingual parallel corpora. The system ranked 14.

UCFWS: This knowledge-based WSD system was based on an algorithm originally described in (Schwartz and Gomez, 2008), in which selectors are acquired from the Web via searching with local context of a given word. The sense is chosen based on the similarity or relatedness between the senses of the target word and various types of selectors. In some runs they include predominant senses(McCarthy et al., 2007). The best run ranked 13th.

NLEL-WSD(-PDB): The system used for the participation is based on an ensemble of different methods using fuzzy-Borda voting. A similar system was proposed in SemEval-2007 task-7 (Buscaldi and Rosso, 2007). In this case, the component method used where the following ones: 1) Most Frequent Sense from SemCor; 2) Conceptual Density ; 3) Supervised Domain Relative Entropy classifier based on WordNet Domains; 4) Supervised Bayesian classifier based on WordNet Domains probabilities; and 5) Unsupervised Knownet-20 classifiers. The best run ranked 24th.

UMCC-DLSI (Relevant): The team submitted three different runs using a knowledge-based system. The first two runs use domain vectors and the third is based on cliques, which measure how much a concept is correlated to the sentence by obtaining Relevant Semantic Trees. Their best run ranked 27th.

(G)HR: They presented a Knowledge-based WSD system, which make use of two heuristic rules (Li et al., 1995). The system enriched the Chinese WordNet by adding semantic relations for English domain specific words (e.g. ecology, environment). When in-domain senses are not available, the system relies on the first sense in the Chinese WordNet. In addition, they also use sense definitions. They only participated in the Chinese task, with their best system ranking 1st.

4 Results

The evaluation has been carried out using the standard Senseval/SemEval scorer `scorer2` as included in the trial dataset, which computes precision and recall. Table 3 shows the results in each dataset. Note that the main evaluation measure is recall (R). In addition we also report precision (P) and the recall for nouns and verbs. Recall measures are accompanied by a 95% confidence interval calculated using bootstrap resampling procedure (Noreen, 1989). The difference between two systems is deemed to be statistically significant if there is no overlap between the confidence intervals. We show graphically the results in Figure 1. For instance, the differences between the highest scoring system and the following four systems are not statistically significant. Note that this method of estimating statistical significance might be more strict than other pairwise methods.

We also include the results of two baselines. The random baseline was calculated analytically. The first sense baseline for each language was taken from each wordnet. The first sense baseline in English and Chinese corresponds to the most frequent sense, as estimated from out-of-domain corpora. In Dutch and Italian, it followed the intuitions of the lexicographer. Note that we don't have the most frequent sense baseline from the domain texts, which would surely show higher results (Koeling et al., 2005).

5 Conclusions

Domain portability and adaptation of NLP components and Word Sense Disambiguation systems present new challenges. The difficulties found by supervised systems to adapt might change the way we assess the strengths and weaknesses of supervised and knowledge-based WSD systems. With this paper we have motivated the creation of an all-words test dataset for WSD on the environment domain in several languages, and presented the overall design of this SemEval task.

One of the goals of the exercise was to show that WSD systems could make use of unannotated background corpora to adapt to the domain and improve their results. Although it's early to reach hard conclusions, the results show that in each of the datasets, knowledge-based systems are able to improve their results using background text, and in two datasets the adaptation of knowledge-based systems leads to results over the MFS baseline. The evidence of domain adaptation of supervised systems is weaker, as only one team tried, and the differences with respect to MFS are very small. The best results for English are obtained by a system that combines a knowledge-based system with some targeted hand-tagging. Regarding the techniques used, graph-based methods over WordNet and distributional thesaurus acquisition methods have been used by several teams.

All datasets and related information are publicly available from the task websites[6].

Acknowledgments

We thank the collaboration of Lawrence Jones-Walters, Amor Torre-Marin (ECNC) and Karin de Boom (WWF), compiling the test and background documents. This work task is partially funded by the European Commission (KYOTO ICT-2007-211423), the Spanish Research Department (KNOW-2 TIN2009-14715-C04-01) and the Basque Government (BERBATEK IE09-262).

References

Eneko Agirre and Aitor Soroa. 2009. Personalizing pagerank for word sense disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL09)*, pages 33–41. Association for Computational Linguistics.

Eneko Agirre, Oier Lopez de Lacalle, and Aitor Soroa. 2009. Knowledge-based wsd on specific domains: Performing better than generic supervised wsd. In *Proceedigns of IJCAI. pp. 1501-1506.".*

Davide Buscaldi and Paolo Rosso. 2007. Upv-wsd : Combining different wsd methods by means of fuzzy borda voting. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 434–437.

P. Chen, W. Ding, and D. Brown. 2009. A fully unsupervised word sense disambiguation method and its evaluation on coarse-grained all-words task. In *Proceeding of the North American Chapter of the Association for Computational Linguistics (NAACL09)*.

Radu Ion and Dan Stefanescu. 2009. Unsupervised word sense disambiguation with lexical chains and graph-based context formalization. In *Proceedings of the 4th Language and Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, pages 190–194.

Rubén Izquierdo, Armando Suárez, and German Rigau. 2009. An empirical study on class-based word sense disambiguation. In *EACL '09: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 389–397, Morristown, NJ, USA. Association for Computational Linguistics.

Mitesh Khapra, Sapan Shah, Piyush Kedia, and Pushpak Bhattacharyya. 2010. Domain-specific word sense disambiguation combining corpus based and wordnet based parameters. In *Proceedings of the 5th International Conference on Global Wordnet (GWC2010)*.

A. Kilgarriff. 2001. English Lexical Sample Task Description. In *Proceedings of the Second International Workshop on evaluating Word Sense Disambiguation Systems*, Toulouse, France.

R. Koeling, D. McCarthy, and J. Carroll. 2005. Domain-specific sense distributions and predominant sense acquisition. In *Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing. HLT/EMNLP*, pages 419–426, Ann Arbor, Michigan.

Xiaobin Li, Stan Szpakowicz, and Stan Matwin. 1995. A wordnet-based algorithm for word sense disambiguation. In *Proceedings of The 14th International Joint Conference on Artificial Intelligence (IJCAI95)*.

Diana McCarthy, Rob Koeling, Julie Weeds, and John Carroll. 2007. Unsupervised acquisition of predominant word senses. *Computational Linguistics*, 33(4).

R. Mihalcea, T. Chklovski, and Adam Killgariff. 2004. The Senseval-3 English lexical sample task. In *Proceedings of the 3rd ACL workshop on the Evaluation of Systems for the Semantic Analysis of Text (SENSEVAL)*, Barcelona, Spain.

Eric W. Noreen. 1989. *Computer-Intensive Methods for Testing Hypotheses*. John Wiley & Sons.

Sameer Pradhan, Edward Loper, Dmitriy Dligach, and Martha Palmer. 2007. Semeval-2007 task-17: English lexical sample, srl and all words. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 87–92, Prague, Czech Republic.

Paul Rayson and Roger Garside. 2000. Comparing corpora using frequency profiling. In *Proceedings of the workshop on Comparing corpora*, pages 1–6.

Hansen A. Schwartz and Fernando Gomez. 2008. Acquiring knowledge from the web to be used as selectors for noun sense disambiguation. In *Proceedings of the Twelfth Conference on Computational Natural Language Learning (CONLL08)*.

B. Snyder and M. Palmer. 2004. The English all-words task. In *Proceedings of the 3rd ACL workshop on the Evaluation of Systems for the Semantic Analysis of Text (SENSEVAL)*, Barcelona, Spain.

M. Tesconi, F. Ronzano, S. Minutoli, C. Aliprandi, and A. Marchetti. 2010. Kafnotator: a multilingual semantic text annotation tool. In *In Proceedings of the Second International Conference on Global Interoperability for Language Resources*.

Piek Vossen, Eneko Agirre, Nicoletta Calzolari, Christiane Fellbaum, Shu kai Hsieh, Chu-Ren Huang, Hitoshi Isahara, Kyoko Kanzaki, Andrea Marchetti, Monica Monachini, Federico Neri, Remo Raffaelli, German Rigau, Maurizio Tescon, and Joop VanGent. 2008. Kyoto: a system for mining, structuring and distributing knowledge across languages and cultures. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.

[6] `http://xmlgroup.iit.cnr.it/SemEval2010/` and `http://semeval2.fbk.eu/`