DutchSemCor: Targeting the ideal sense-tagged corpus

Piek Vossen[1], Attila Görög[1], Rubén Izquierdo[3], Antal van den Bosch[4]

[1]VU University Amsterdam, De Boelelaan 1105, 1081HV Amsterdam, The Netherlands

[3]Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands

[4]Radboud University Nijmegen, P.O. Box 9013, 6500 HD Nijmegen, The Netherlands

E-mail: p.vossen@let.vu.nl, a.gorog@let.vu.nl, r.izquierdo@uvt.nl, a.vandenbosch@let.ru.nl

Abstract

Word Sense Disambiguation (WSD) systems require large sense-tagged corpora along with lexical databases to reach satisfactory results. The number of English language resources for developed WSD increased in the past years while most other languages are still under-resourced. The situation is no different for Dutch. In order to overcome this data bottleneck, the DutchSemCor project will deliver a Dutch corpus that is sense-tagged with senses from the Cornetto lexical database. In this paper, we discuss the different conflicting requirements for a sense-tagged corpus and our strategies to fulfill them. We report on a first series of experiments to support our semi-automatic approach to build the corpus.

Keywords: Active Learning, Word Sense Disambiguation, Semantic Annotation, Machine Learning

1. Introduction

State-of-the-art Word Sense Disambiguation (WSD) systems rely on lexical databases and sense-tagged corpora to reach satisfactory results. Building such corpora is a complex and labour-intensive task. The various types of existing sense-tagged corpora have different characteristics, resulting in corresponding WSD systems that perform differently as well. This obscures the training and evaluation of WSD systems: what type of corpus is used for training and how does it relate to the test data? It also makes it difficult to decide on the optimal approach for creating a sense-tagged corpus.

We are building a sense-tagged corpus for Dutch tagged with senses from the Cornetto lexical database (Vossen et al. 2008). In this paper we present a classification of different sense-tagged corpora and describe their pros and cons. We also describe the requirements for an ideal sense-tagged corpus. Sense-tagged corpora should ideally represent all the senses of words (including rare senses), represent the variety of contexts, and provide information on the sense distribution in texts. These three requirements are often contradictory in practice. In this paper we describe our approach of trying to meet these three requirements in the DutchSemCor project. This approach consists of three phases. First we manually create a lexical-sample corpus with double annotations that meets the first requirement (i.e. represent all different senses). In the second phase we semi-automatically extend this corpus through Active Learning. We train a supervised WSD system using the data from the first phase and let the WSD system select more examples to be validated by annotators. In the third phase, we apply clustering techniques to the whole corpus to add more examples to represent the context variety and the sense-distribution reflected in the corpora. The WSD systems built from the final sense-annotated corpus will be tested using an independent all-words corpus.

In this paper, we report on various experiments carried out to determine the optimal approach to meet our requirements. The paper is structured as follows. In the next section, we present a classification of sense-tagged corpora and the requirements for such a corpus. In Section 3 we describe the overall approach that we follow in the DutchSemCor project. In Section 4 we present a series of experiments carried out to fine-tune our approach. In Section 5 we conclude and formulate our next steps on creating groups of fine-grained senses and on the clustering of our background corpus, SoNaR, in order to further meet the other two requirements.

2. Classification of sense-tagged corpora

Roughly speaking, there are two methods to annotate a corpus with senses:

1. sequential tagging: the text is presented in its original order, and each word is tagged in the sequence in which it occurs;
2. targeted tagging: all occurrences of a single target word are listed with a left and right context as in a KWIC index and are annotated through comparison of the contexts.

The two approaches are likely to produce different annotation results for the same text. In the case of sequential tagging, the annotator only reads the text once, but he needs to change focus to different words all the time, repeatedly incurring a substantial cognitive load. In contrast, with targeted tagging the annotator needs to consider the different meanings only once and can apply a more systematic and consistent comparison of the different contexts.

In addition to the annotation method we can also distinguish sense-tagged corpora by their textual coverage:

1. **all-words** corpus: all content words in a selection of texts are annotated with senses;
2. **lexical sample** corpus: a selection of target word occurrences with context are annotated with senses;

Usually, all-words corpora cover a small number of texts, limited to a selection of genres and domains. The advantage is that all content words in the context of a specific word are also annotated. The disadvantage is that the texts usually do not represent all different contexts and meanings of the target word. Target-word corpora, in contrast, are strong in the latter aspects: they cover many different contexts and meanings of the target-word, but lack the context of the whole text and annotation of neighboring words. The most famous example of an all-words corpus is SemCor[1] (Miller et al., 1993), which was created through sequential tagging of parts of the Brown corpus (186 texts have all-words annotation, while in 166 texts only the verbs are annotated). An example of a target-word corpus is the so-called *line-hard-serve* corpus[2] which contains 4,000 instances of the noun *line* (six meanings), 4,000 instances of the verb *serve* (four meanings), and 4,000 instances of the adjective *hard* (three meanings). Another target-words corpus is DSO[3] which has annotations only for the most frequent and ambiguous nouns (121) and verbs (70) in parts of the Brown corpus and a selection of Wall Street Journal articles, but is comparable in size to SemCor. For evaluation purposes, many other small all-words and lexical-sample corpora have been produced (cf. Senseval and SemEval competitions).

2.1 Subdivision of target-word corpora

Lexical-sample and all-words corpora can further differ in the range and selection of their texts. SemCor and DSO partly inherit the balanced nature of the Brown corpus. There are a number of motivations that define the selection of texts for a lexical sample corpus:

1. **balanced-sense** corpus: provide tokens and contexts for words that clearly illustrate the meaning of a word and provide equal numbers of examples for each meaning;
2. **balanced-context** corpus: provide tokens and contexts that represent the different usages of words in a representative corpus;
3. **sense-probability** corpus: provide a representative sample of the true usage of a word meaning in a representative corpus;

In the case of 1, annotators start from the meaning of a word and look for representative examples from any set of sources. In the case of 2, annotators get a selection of tokens based on the structural contextual properties and other meta-data that are available for all tokens. In such a selection, the same context is not annotated twice. In the case of 3 it is sufficient to take a random sample from a corpus that approximates the true population of usage of a word. Such a sample can contain very similar usages and contexts to the extent that they are more probable

[1] http://www.cse.unt.edu/~rada/downloads.html#semcor

[2] http://www.d.umn.edu/~tpederse/data.html

[3] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC97T12

than others.

2.2 Effects on WSD systems

Although a sense-tagged corpus can be used for many different purposes, its main purpose is to train and/or test a word-sense-disambiguation (WSD) system. The above choices to create a sense-annotated corpus also have an effect on the quality and type of WSD system that can be built based on the data. For example, the sense frequency in SemCor is a very strong predictor of senses in unseen text. Likewise, taking the most-frequent-sense (MFS) of SemCor is a baseline that is difficult to beat in many evaluation studies (Aggire and Edmonds, 2006), as long as the test data exhibit a similar frequency distribution. Methods that determine the predominant sense in different domains (McCarthy et al., 2007) attempt to derive sense-frequency correlations in new corpora. Nevertheless it may be infeasible to obtain a variety of domains and contexts in such a way that all different usages and senses can be detected. A sense-probability corpus results in a WSD that can handle frequent senses well but will score low on rare senses. For these reasons, a WSD system that uses the MFS heuristic from SemCor can be expected to perform poorly on texts with different distributional properties. How well does the selection of the Brown corpus in SemCor represent the true population of language as a sample? How well does any corpus represent the true population? Such a corpus may need to be so big that it becomes unaffordable to create an **all-words** corpus.

It may thus be more efficient to construct **lexical-sample** corpora that try to capture all the good properties of the different types of corpora mentioned i.e.: represent all meanings, represent the different contexts in which the words occur, and exhibit realistic probabilities for senses. In the next sections, we describe how we try to harmonize these requirements.

3. Project methodology

The primary goal of the DutchSemCor project is to create a ***balanced-sense lexical-sample*** corpus for the 3,000 most frequent and polysemous Dutch words with about 100 examples for each sense (also for less frequent senses). Since these 3,000 words have about 3-4 senses on average, the final corpus will thus contain about 1 million sense-tagged tokens. This corpus is built partially manually and partially semi-automatically. In the first manual phase 25 examples are collected for each sense. These examples are used to train a supervised WSD system for the second phase. The supervised system searches for the remaining 75 examples of the different senses to complete the corpus. Active learning is used to steer the supervised system in selecting appropriate examples.

3.1 First annotation phase

In the first phase of the project, 282,503 tokens for 2,870 nouns, verbs and adjectives (11,982 senses) were annotated manually by two annotators. The annotators needed to reach a high agreement and were instructed to select 25 diverse examples for each sense. The examples were selected using a sense annotation tool SAT[4] (Görög & Vossen 2010; Van Gompel 2010; Vossen et al. 2011) from the 500 million token SoNaR corpus of written Dutch (Oostdijk et al., 2008) and the CGN Spoken Dutch Corpus (Eerten, 2007). The annotators see a KWIC index of all word occurrences in the corpus. They can sort on the left and right context, and can filter tokens using co-occurrence search. If they do not find a sufficient number of good examples, they can search the Internet for additional instances using a snippet search tool developed for the project. The distribution of annotated examples over the different resources is 67% SoNaR, 5% CGN, and 28% web-snippets. This shows that even a 500-million-token corpus is not big enough to create a balanced-sense corpus, since 28% of the examples needed to come from the Internet.

At the end of the first annotation phase, 80% of the senses got 25 annotated examples or more, and 90% of the lemmas got 25 examples for each sense. A small but significant proportion of senses is not well represented in the corpus even after Internet search. These are mostly very rare senses from specific domains or registers, e.g. the Dutch word *crisis* that refers to a specific critical medical state. Nevertheless, we can conclude that we achieved a satisfactory result on the first quantitative requirement to represent all the senses of the words in the corpus.

The average Inter-Annotator Agreement (IAA)

[4] http://zookma.science.uva.nl/dutchsemcor/

for this corpus is 94%. This high IAA score can be explained by our working method: the annotators did not tag all the tokens presented to them, but were given the instructions to select contexts that clearly represent the senses, and to avoid vague, problematic and unclear cases. This is another indication that the annotated tokens in the corpus represent the senses well.

3.2 Second annotation phase

In the second phase of the project, the *balanced-sense lexical-sample* corpus is extended with more examples using a supervised WSD system and through Active Learning (AL). We designed the following procedure:

1. We build and test a supervised WSD system (Initial Learning WSD, or IL-WSD) using the examples from the first phase.
2. All words with a tested accuracy above 80% are considered done and ready for completing the corpus. That is: the supervised system will be able to find more examples for the senses with sufficiently high confidence and high precision.
3. All words that perform lower than 80% will undergo AL to obtain better results and more examples.

We defined the following AL procedure for all words that are tagged at less than 80% accuracy by IL-WSD:

1. Use IL-WSD to annotate all remaining SoNaR tokens of word *w* in WAL, where WAL is the set of words performing lower than 80% accuracy;
2. Present the annotators a selection of 50 tokens for *w* tagged with senses that perform badly;
3. The annotators tag these tokens with the proper sense, thus creating an Active Learning corpus (AL-corpus). Annotators can assign any of the senses to the tokens, thus also re-assign tokens to senses that already perform well;

The AL-corpus is used to improve the WSD system for words in WAL. This process can be repeated until we attain the desired results for all the words.

There is an important difference between the human annotation performed for IL and for AL. While the IL annotators could search for clear examples and ignore poor examples, the AL annotators are obliged to annotate all the (50) tokens presented to them. The tokens presented to the AL annotators are therefore more determined by the characteristics of the SoNaR corpus. The AL annotators also encountered errors in lemmatization and part-of-speech, figurative and idiomatic usage, and unknown senses, which they explicitly had to mark for exclusion.[5]

The two extreme options when selecting examples for AL from SoNaR are to either select examples that are very similar to the IL set or very different. It is not clear which of these examples will help improve the WSD system most. Choosing similar examples may be good for completing a *balanced-sense* corpus but it will not result in a *balanced-context* corpus. By selecting very different examples we run the risk of selecting deviating material the annotators will often choose to ignore.

Another open issue is how many tokens are needed to achieve sufficient accuracy. With the limited resources available for manual annotation, it is important to know what the minimal set is for each word in WAL and what the different characteristics are of these words. It could very well be that words with fine-grained (metonymic) meanings can never reach an 80% accuracy no matter how hard we try. Rather than adding more training data, these cases are better solved by creating coarse-grained sense groups.

Finally, the corpus created through IL and AL will not have true sense distributions. Although it is clear that SoNaR only represents 67% of the senses it can still be useful to obtain sense-frequency information. We also would like to know if SoNaR contains usages of words that may indicate senses not in our sense database.

In the next sections we describe a number of experiments carried out to find the optimal AL approach (sufficient precision) and the minimally required number of training data for our primary goal: a *balanced-sense lexical-sample corpus* of 1 million tokens.

[5] About 15% of the tokens during AL are being disqualified by the annotators.

4. Experiments and Results

Our WSD system is based on k-Nearest Neighbour classification (Aha et al, 1991), and uses an implementation that has been applied to word sense disambiguation in the past: TIMBL[6] (Hoste et al, 2002; Decadt et al, 2004; Daelemans et al, 2007). The WSD system is an ensemble of classifiers, where each classifier is a word expert that disambiguates among the different meanings of a particular word.

A first series of experiments were carried out on 82 nouns for which the performance of our WSD-system was around the threshold of 80% (accuracy). We measured the improvements of the system and the effect of additional training data in three cycles of AL (see the description of the AL approach above). We chose lemmas scoring around 80% since we assumed that poor scoring lemmas will not improve enough in three cycles and lemmas scoring far above 80% are already sufficiently trained. We will describe in this section the different experiments that have been carried out to adjust our WSD system and achieve the goals of the project.

Active Learning (AL) involves (1) training the WSD system with all manually annotated data (IL set), (2) annotate all the remaining unannotated data, and (3) make a selection of new instances. These new instances are shown to the annotators for validation. After validation an extended set of annotated data is available for training (AL set). After retraining the WSD with the new set of data, the whole process starts again. A key aspect of AL is the method of selecting the new instances, as we explain later.

The goal during the IL annotation process was to generate a **balanced-sense** corpus with very clear examples. A priori we do not know whether the IL set created by double human annotation is representative of the full SoNaR corpus.

To evaluate our WSD system and the evolution of the Active Learning (AL) process, we followed an *n-fold cross-validation* technique. Setting *n*=5, we split folds at the word meaning level, to make sure that the number of instances per word meaning in the five folds is balanced for all 82 nouns. Moreover, when new instances are added during the AL process, the folds are not recalculated, but expanded with the new instances, assuring that the balance between the IL instances and new instances is kept across the folds.

4.1 Feature Set

The first experiments were aimed at analyzing the best set of features for our K-nearest WSD system. Four types of features were selected: words (W), lemmas (L), PoS tags (P) and bag-of-words (B), and different sizes for the context were considered, ranging from one to five. In this experiments the IL example set was used as data. Token accuracy for the combinations of feature types and context sizes are shown in Table 1.

Features	#1	#2	#3	#4	#5
W	78.19	77.46	76.21	76.84	74.26
W B	**81.62**	81.17	80.41	79.79	79.12
W L	76.68	75.64	74.74	73.68	73.16
W L P	76.19	75.96	74.76	73.88	73.61
W L P B	80.35	80.01	79.12	78.68	77.68

Table 1. Experiments with feature sets

The feature set that led to the best performance contains words in a 1-token window around the target word in combination with a bag-of-words representation of sense-discriminating words. This feature set was used in the remaining experiments.

4.2 Increasing training data

As the main idea behind our Active Learning method is to increase the amount of training data to boost our WSD system, we want to first evaluate to what extent our system is affected by the amount of training data. We used again the IL examples as training and evaluation data. In the four experiments in next table, we kept the test folds fixed, and different number of training instances for each word meaning were selected, ranging from 5 to 20. In Table 2 we list the average token accuracy for the 82 lemmas, as well as the average number of total instances per training/testing fold.

[6]http://ilk.uvt.nl/timbl

# Training ex. per meaning	Acc.	Avg # Train per fold	Avg # Test per fold
20	81.62	6913	1728
15	79.68	5185	1728
10	76.37	3456	1728
5	68.61	1728	1728

Table 2. Experiment with different training data sizes

Table 2 shows that more training examples clearly leads to better performance even at the small numbers of examples we operate with, so even the addition of small amounts of training examples may have notable effects.

4.3 High-confidence – Low-distance

Our first idea was to add very similar instances to directly boost the WSD system. In order to achieve this we adjusted our AL method as follows. First our WSD system was trained with the IL set, and all tokens of the 82 lemmas in SONAR were automatically tagged. Then, for each word meaning of these words, all the disambiguated instances were sorted according to the combination (F-score) of the confidence of the WSD system and the distance to the nearest neighbor. For each sense of each of the 82 lemmas we selected 50 instances automatically tagged with high confidence (HC) and low distance (LD). The HC criterion was used to select instances with high probability of being correct, whereas the LD criterion was intended to select *similar* instances to those in the IL set. After the review process, the new instance set was called HC_LD. This new HC_LD set can be divided into two subsets: those instances where the annotators agreed with the WSD system (HC_LD.Agreed) and those where the annotators corrected the output of the system (HC_LD.Disagreed). The first subset was supposed to be even more similar to the IL set, maybe sharing quite structural properties and common features. For this reason, we considered also the HC_LD.Agreed as an independent new set of instances in our evaluation.

In order to evaluate the system with the new instances we extended the training folds used in the initial evaluation (Table 2) with the the new instances (HC_LD and HC_LD.Agreed), and evaluated on the same test folds used in the experiments shown in Table 2, to have a fair comparison. In Table 3 we list the results of this evaluation.

Training	Acc	Avg # Train per fold	Avg # Test per fold
TR_{IL} EV_{IL}	81.62	6913	1728
TR_{IL+HC_LD} EV_{IL}	80.74	11537	1728
$TR_{IL+HC\_LD.Agreed}$ EV_{IL}	82.68	9677	1728

Table 3. Experiment with HC-LD

We can see that the new instances did not increase the performance. The reason could be that the new examples were not as clear and well-formed as the ones used for IL, despite the fact that they have a low distance. This is reinforced by the fact that we do get an improvement when we restrict ourselves to agreed examples.

4.4 High-confidence – High-distance.

Pursuing the second goal of the project to create a balanced-context corpus, we modified our AL module to select examples that have very different contexts from IL. We trained our system on the IL data, tagged the remaining SoNaR corpus, and selected 50 instances per word meaning of the 82 lemmas having a high confidence (HC) and a high distance (HD) to the nearest instance. Again the automatic annotated instances were shown to the annotators for validation, resulting in a new set of instances: HC_HD (and also the subset HC_HD.Agreed as in the previous section). The evaluation of the WSD system considering the new HC_HD set is shown in Table 4[7] .

Training	Acc	Avg#Train/fold	Avg#Test/fold
TR_{IL} EV_{IL}	81.62	6913	1728
TR_{IL+HC_HD} EV_{IL}	79.45	17262	1728
$TR_{IL+HC\_HD.Agreed}$ EV_{IL}	82.61	11631	1728

Table 4: Experiments with HC_HD

[7]The evaluation folds were again the same as in Table 2.

The performance drops 4 points when extending the training data in IL with the new instances, but it increases 1.5 points if only agreed instances are used. Recall that in the creation of IL the annotators were allowed to look for clear examples, while in AL all the instances suggested automatically by the WSD system had to be annotated. The effect is that the WSD system received less clear examples, which were also very different from IL. To check this hypothesis we obtained the performance of the system when training with instances from IL and testing with instances from IL+HC_HD. We wanted to evaluate how useful our IL instances were to disambiguate the new instances HC_HD. The results are shown in Table 5.

System	Accuracy	# instances
TR_{IL+HC_HD} EV_{IL+HC_HD}	76.24	19055
TR_{IL} EV_{IL+HC_HD}	63.73	8641

Table 5. Evaluating IL over HC_HD.

As can be seen in the table, the performance of the system evaluating over IL+HC_HD loses 12.5 points when training only with IL. This confirms our suspicion that the instances in HC_HD are very different and more difficult to disambiguate than the IL instances.

4.5 Evaluating with new data

Finally we evaluated the WSD system considering all the new data obtained (HC_LD and HC_HD) to both training and evaluation. The evaluation was conducted following the same n-fold cross validation, and combining the different data sets. The results are shown in Table 7. An interesting outcome is that when only adding instances for which the annotators agree with AL, performance is boosted up to 85.33, which is a good improvement considering the 81.62 in C1. It appears that high or low distance to training instances is not a usable criterion for adding new instances. In contrast, the addition of instances for which the annotators agree with the suggestion made by the retrained WSD does appear to provide new high-quality information that can boost WSD performance.

System	Accuracy	# instances
TR_{IL} EV_{IL}	81.62	8641
TR_{IL+HC_LD} / EV_{IL+HC_LD}	78.87	13266
$TR_{IL+HC\_LD.Agree}$ / $EV_{IL+HC\_LD.Agree}$	85.02	11405
TR_{IL+HC_HD} / EV_{IL+HC_HD}	76.24	19055
$TR_{IL+HC\_HD.Agree}$ / $EV_{IL+HC\_HD.Agree}$	83.77	13359
$TR_{IL+HC\_HD+HC\_LD}$ / $EV_{IL+HC\_HD+HC\_LD}$	76.74	23692
$TR_{IL+HC\_HD.Agree+HC\_LD.Agree}$ / $EV_{IL+HC\_HD.Agree+HC\_LD.Agree}$	85.33	16123

Table 7. Experiments with all data selected by AL

4. Conclusion and future work

We presented a classification of different sense-annotated corpora and described their (dis)advantages. We proposed a method for meeting the possibly conflicting requirements for such corpora. We demonstrated the feasibility of our approach to efficiently build a balanced-sense lexical-sample corpus in a semi-automatic way. From a manually annotated seed corpus,we can automatically extend the representative annotations through WSD, where we use high-confidence results and active learning for low-performing words. A small proportion of the words and word-senses will always be poorly represented, as their usage can only be found on the Internet or their senses cannot be discriminated. In future work we will further isolate the latter cases by deriving coarse-grained sense groups.

Finally, we will apply independent clustering of all tokens in the corpora that are not annotated in our approach. These clusters are used to extend the context coverage of our corpus and to derive sense-probabilities as reflected in SoNaR. Through these strategies we will eventually obtain a sense-annotated corpus that meets all three requirements. We are currently creating an independent all-words corpus to validate the quality of the WSD system based on our lexical-sample corpus.

5. References

Aha, D. W. and Kibler, D. and Albert, M. K. (1991). Instance-Based Learning AI. In Journal of Machine Learning, number 1, pp. 37—66.

Agirre, E., & Edmonds, P. (Eds.) (2006). Word Sense Disambiguation: Algorithms and Applications, Springer.

Agirre, E. and A. Soroa. (2009). Personalizing pagerank for word sense disambiguation. In Proceedings of EACL-09, pp. 33--41.

Agirre, E., Lopez de Lacalle, O., Ch. Fellbaum, S. Hsieh, M. Tesconi4, M. Monachini, P. Vossen, R. Segers (2010) "SemEval-2010 Task 17: All-words Word Sense Disambiguation on a Specific Domain", in: Proceedings of SemEval-2010: 5th International Workshop on Semantic Evaluations on Kyoto's subtask WSD17: All-words Word Sense Disambiguation on a Specific Domain, workshop collocation: ACL2010, July 11-16, 2010, Uppsala, Sweden, pp. 75--80, Ed. K. Erk & C. Strapparava,, Publ. The Association for Computational Linguistics (ACL).

Chen, J., Schein, A., Ungar, L., Palmer, M., (2006) An empirical study of the behavior of active learning for word sense disambiguation. In: *Proceedings of HLT-NAACL06*

Decadt, B., Hoste, V., Daelemans, W., and Van den Bosch, A., (2004) GAMBL, genetic algorithm optimization of memory-based WSD, In: R. Mihalcea and P. Edmonds (eds.), *Proceedings of the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text (Senseval-3)*, Barcelona, Spain, pp. 108--112.

Daelemans, W., Zravel, J., van der Sloot, K. and van den Bosch, A (2007). TiMBL: Tilburg Memory Based Learner, version 6.1. Reference Guide. ILK Technical Report 07-07

Eerten, L. (2007). Over het Corpus Gesproken Nederlands. In *Nederlandse Taalkunde*, 12 (3) pp. 194--215.

Görög, A., Vossen, P. (2010) "Computer Assisted Semantic Annotation in the DutchSemCor Project." In Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10). Malta, Valletta

Hoste, V.; Hendrickx, I.; Daelemans, W. and Van Den Bosch, A. (2002). Parameter optimization for machine learning of word sense disambiguation. *Nat. Lang. Eng.* 8, 4 (December 2002), pp. 311--325.

McCarthy, D., R. Koeling, J. Weeds and J. Carroll, (2007) Unsupervised Acquisition of Predominant Word Senses. *Computational Linguistics,* 33 (4) pp 553-590

Miller, G. A., Leacock, C., Tengi, R. & Bunker, R.T (1993). A semantic concordance. In: *Proceedings of the ARPA Workshop on Human Language Technology*, 303–308.

Ng, H.T. and Lee, H.B. (1996). Integrating multiple knowledge sources to disambiguate word sense: an exemplar-based approach. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics* (ACL '96). Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 40--47.

Oostdijk, N. et al. (2008). From D-Coi to SoNaR: A reference corpus for Dutch. In: *Proceedings on the sixth international Conference on Language Resources and Evaluation* (LREC 2008), Marrakech, Marocco.

Dan Shen, Jie Zhang, Jian Su, Guodong Zhou and Chew Lim Tan. (2004) Multi-criteria-based active learning for named entity recognition, In: *Proceedings of ACL04*, Barcelona, Spain.

Van Gompel, M. (2010). van UvT-WSD1: A cross-lingual word sense disambiguation system. In *SemEval'10: Proceedings of the 5th International Workshop on Semantic Evaluation*, Uppsala, Sweden, pp. 238--241.

Vossen, P. et al. (2008). Integrating Lexical Units, Synsets, and Ontology in the Cornetto Database. In: *Proceedings on the sixth international Conference on Language Resources and Evaluation* (LREC 2008), Marrakech, Marocco.

Vossen, P. et al. (2011). DutchSemCor: building a semantically annotated corpus for Dutch. In: *Proceedings of Electronic Lexicography in the 21st century: New Applications for new* users (eLEX2011), Bled, Slovenia, November 10-12, 2011

Zhu, J., Hovy, E.H. (2007). Active Learning for Word Sense Disambiguation with Methods for Addressing the Class Imbalance Problem. In: *Proceedings of the EMNLP conference*, Prague, Czech Republic.