UvT-WSD1: a Cross-Lingual Word Sense Disambiguation system

Maarten van Gompel
Tilburg centre for Cognition and Communication
Tilburg University
proycon@anaproy.nl

Abstract

This paper describes the Cross-Lingual Word Sense Disambiguation system UvT-WSD1, developed at Tilburg University, for participation in two SemEval-2 tasks: the Cross-Lingual Word Sense Disambiguation task and the Cross-Lingual Lexical Substitution task. The UvT-WSD1 system makes use of k-nearest neighbour classifiers, in the form of single-word experts for each target word to be disambiguated. These classifiers can be constructed using a variety of local and global context features, and these are mapped onto the translations, i.e. the senses, of the words. The system works for a given language-pair, either English-Dutch or English-Spanish in the current implementation, and takes a word-aligned parallel corpus as its input.

1 Introduction

The UvT-WSD1 system described in this paper took part in two similar SemEval-2 tasks: Cross-Lingual Word Sense Disambiguation (Lefever and Hoste, 2010) and Cross-Lingual Lexical Substitution (Mihalcea et al., 2010). In each task, a number of words is selected for which the senses are to be determined for a number of instances of these words. For each word, a number of samples in context is provided, where each sample consists of one sentence, with the word to be disambiguated marked.

Because of the cross-lingual nature of the tasks, a word sense corresponds to a translation in another language, rather than a sense description in the same language. In the Cross-lingual Lexical Substitution task, the target language is Spanish. The task is to find Spanish substitutes for the English words marked in the test samples. In the Cross-Lingual Word Sense Disambiguation task, we participate for English-Dutch and English-Spanish. The Word Sense Disambiguation task provides training data for all five languages, in the form of the sentence-aligned EuroParl parallel corpus (Koehn, 2005). This is the source of training data the UvT-WSD1 system uses for both tasks.

The system may output several senses per instance, rather than producing just one sense prediction. These are evaluated in two different ways. The scoring type "**best**" expects that the system outputs the best senses, in the order of its confidence. The scoring type "**out of five/ten**" expects five or ten guesses, and each answer weighs the same. These metrics are more extensively described in (Mihalcea et al., 2010). The UvT-WSD1 system participates in both scoring types, for both tasks. The system put forth in this paper follows a similar approach as described in earlier research by (Hoste et al., 2002).

2 System Description

The UvT-WSD1 system uses machine learning techniques to learn what senses/translations are associated with any of the target words. It does so on the basis of a variety of local and global context features, discussed in Section 2.2. At the core of the system are the classifiers, or so called "word experts", one per target word. These are built using the Tilburg Memory Based Learner (TiMBL) (Daelemans et al., 2009), making use of the IB1 algorithm, an implementation of the k-nearest neighbour classifier.

The core of the system can be subdivided into roughly three stages. In the first stage, the word-aligned parallel corpus is read and for each found instance of one of the target words, features are extracted to be used in the classifier. The class consists of the word aligned to the found instance of the target word, i.e. the translation/sense. In this way a word expert is built for each of the target

words in the task, yielding a total amount of classifiers equal to the total amount of target words. The test data is processed in a similar way, for each marked occurrence of any of the target words, features are extracted and test instances are created. Subsequently, the word experts are trained and tested, and on the basis of the training data, a parameter search algorithm (Van den Bosch, 2004) determines the optimal set of classifier parameters for each word expert, including for example the value of k and the distance weighting metric used.

In the last phase, the classifier output of each word expert is parsed. The classifiers yield a distribution of classes per test instance, and these are converted to the appropriate formats for "best" and "out of five/ten" evaluation. For the latter scoring type, the five/ten highest scoring senses are selected, for the former scoring type, all classes scoring above a certain threshold are considered "best". The threshold is set at 90% of the score of the highest scoring class.

2.1 Word-Alignment, Tokenisation, Lemmatisation and Part-of-Speech-tagging

The Europarl parallel corpus, English-Spanish and English-Dutch, is delivered as a sentence-aligned parallel corpus. We subsequently run GIZA++ (Och and Ney, 2000) to compute a word-aligned parallel corpus.

This, however, is not the sole input. The target words in both tasks are actually specified as a lemma and part-of-speech tag pair, rather than words. In the Word Sense Disambiguation task, all target lemmas are simply nouns, but in the Cross-Lingual Lexical Substitution task, they can also be verbs, adjectives or adverbs. Likewise, both tasks expect the sense/translation output to also be in the form of lemmas. Therefore the system internally has to be aware of the lemma and part-of-speech tag of each word in the parallel corpus and test data, only then can it successfully find all occurrences of the target words. In order to get this information, both sides of the word-aligned parallel corpus are run through tokenisers, lemmatisers and Part-of-Speech taggers, and the tokenised output is realigned with the untokenised input so the word alignments are retained. The test data is also processed this way. For English and Spanish, the software suite Freeling (Atserias et al., 2006) performed all these tasks, and for Dutch it was done by Tadpole (Van den Bosch et al., 2007).

2.2 Feature Extraction

The system can extract a variety of features to be used in training and testing. A distinction can be made between *local context features* and *global context features*. Local context features are extracted from the immediate neighbours of the occurrence of the target word. One or more of the following local context features are extractable by the UvT-WSD1 system: word features, lemma features, and part-of-speech tag features. In each case, n features both to the right and left of the focus word are selected. Moreover, the system also supports the extraction of bigram features, but these did not perform well in the experiments.

The global context features are made up of a bag-of-words representation of keywords that *may* be indicative for a given word to sense/translation mapping. The idea is that words are collected which have a certain power of discrimination for the specific target word with a specific sense, and all such words are then put in a bag-of-word representation, yielding as many features as the amount of keywords found. A global count over the full corpus is needed to find these keywords. Each keyword acts as a binary feature, indicating whether or not that particular keyword is found in the context of the occurrence of the target word. The context in which these keywords are searched for is exactly one sentence, i.e. the sentence in which the target word occurs. This is due to the test data simply not supplying a wider context.

The method used to extract these keywords (k) is proposed by (Ng and Lee, 1996) and used also in the research of (Hoste et al., 2002). Assume we have a focus word f, more precisely, a lemma and part-of-speech tag pair of one of the target words. We also have one of its aligned translations/senses s, which in this implementation is also a lemma. We can now estimate $P(s|k)$, the probability of sense s, given a keyword k, by dividing $N_{s,k_{local}}$. (the number of occurrences of a possible local context word k with particular focus word lemma-PoS combination and with a particular sense s) by $N_{k_{local}}$ (the number of occurrences of a possible local context keyword k_{loc} with a particular focus word-PoS combination regardless of its sense). If we also take into account the frequency of a possible keyword k in the complete training corpus ($N_{k_{corpus}}$), we get:

$$P(s|k) = \frac{N_{s,k_{local}}}{N_{k_{local}}}(\frac{1}{N_{k_{corpus}}}) \quad (1)$$

(Hoste et al., 2002) select a keyword k for inclusion in the bag-of-words representation if that keyword occurs more than T_1 times in that sense s, and if $P(s|k) \geq T_2$. Both T_1 and T_2 are predefined thresholds, which by default were set to 3 and 0.001 respectively. In addition, UvT-WSD1 contains an extra parameter which can be enabled to automatically adjust the T_1 threshold when it yields too many or too few keywords. The selection of bag-of-word features is computed prior to the extraction of the training instances, as this information is a prerequisite for the successful generation of both training and test instances.

2.3 Voting system

The local and global context features, and the various parameters that can be configured for extraction, yield a lot of possible classifier combinations. Rather than merging all local context and global context features together in a single classifier, they can also be split over several classifiers and have an arbiter voting system do the final classification step. UvT-WSD1 also supports this approach. A voter is constructed by taking as features the class output of up to three different classifiers, trained and tested on the training data, and mapping these features onto the actual correct sense in the training data. For testing, the same approach is taken: up to three classifiers run on the test data; their output is taken as feature vector, and the voting system predicts a sense. This approach may be useful in boosting results and smoothing out errors. In our experiments we see that a voter combination often performs better than taking all features together in one single classifier. Finally, also in the voter system there is a stage of automatic parameter optimisation for TiMBL.

3 Experiments and Results

Both SemEval-2 tasks have provided trial data upon which the system could be tested during the development stage. Considering the high configurability of the various parameters for feature extraction, the search space in possible configurations and classifier parameters is vast, also due to fact that the TiMBL classifier used may take a wealth of possible parameters. As already mentioned, for the latter an automatic algorithm of parameter optimisation was used (Van den Bosch, 2004), but optimisation of the feature extraction parameters has not been automated. Rather, a selection of configurations has been manually chosen and tested during the development stage.

BEST	UvT-WSD1-v	UvT-WSD1-g
Precision & Recall	21.09	19.59
Mode Prec. & Rec.	43.76	41.02
Ranking (out of 14)	6	9
OUT OF TEN	UvT-WSD1-v	UvT-WSD1-g
Precision & Recall	58.91	55.29
Mode Prec. & Rec.	62.96	73.94
Ranking	3	4

Table 1: UvT-WSD1 results in the Cross-Lingual Lexical Substitution task

The following two configurations of features were found to perform amongst the best on the trial data. Therefore they have been selected and submitted for the contest:

1. **UvT-WSD1-v** (aka *UvT-v*) – An arbiter-voting system over three classifiers: 1) Word experts with two word features and lemma features on both sides of the focus word. 2)Word experts with global features[1]. 3) Word experts with two word features, two lemma features *and* two part-of-speech tag features.

2. **UvT-WSD1-g** (aka *UvT-g*) – Word experts with global features only.

Table 1 shows a condensed view of the results for the Cross-Lingual Lexical Substitution task. Table 2 shows the final results for the Word-Sense Disambiguation task. Note that UvT-WSD1-v and UvT-WSD1-g are two different configurations of the UvT-WSD1 system, and to conserve space these are abbreviated as UvT-v and UvT-g respectively. These are also the names used in both tasks (Lefever and Hoste, 2010; Mihalcea et al., 2010) to refer to our system.

4 Discussion and Conclusion

Cross-Lingual Word Sense Disambiguation and Cross-Lingual Lexical Substitution have proven to be hard tasks, with scores that are relatively close to baseline. This can be attributed to a noticeable trait in the system output to be inclined to assign the same majority sense to all instances.

[1]For the Cross-Lingual Lexical Substitution task only, the parameter to recompute the T_1 threshold automatically was enabled.

Dutch BEST	UvT-v	UvT-g	T3-COLEUR		
Precision & Recall	17.7	15.93	10.72 & 10.56		
Mode Prec. & Rec.	12.06	10.54	6.18 & 6.16		
Dutch OUT OF FIVE	UvT-v	UvT-g	T3-COLEUR		
Precision & Recall	34.95	34.92	21.54 & 21.22		
Mode Prec. & Rec.	24.62	19.72	12.05 & 12.03		
Spanish BEST	UvT-v	UHD-1	UvT-g	T3-COLEUR	FCC-WSD1
Precision & Recall	23.42	20.48 & 16.33	19.92	19.78 & 19.59	15.09
Mode Prec. & Rec.	24.98	28.48 & 22.19	24.17	24.59	14.31
Spanish OUT OF FIVE	UvT-g	UvT-v	FCC-WSD2	UHD-1	T3-COLEUR
Precision & Recall	43.12	42.17	40.76	38.78 & 31.81	35.84 & 35.46
Mode Prec. & Rec.	43.94	40.62	44.84	40.68 & 32.38	39.01 & 38.78

Table 2: UvT-WSD1 results in comparison to other participants in the Word-Sense Disambiguation task

In our system, we used the same configuration of feature extraction, or a voter over a set of configurations, for all word experts. The actual classifier parameters however, do differ per word expert, as they are the result of the automatic parameter optimisation algorithm. Selecting different feature extraction configurations per word expert would be a logical next step to attempt to boost results even further, as been done in (Decadt et al., 2004).

Keeping in mind the fact that different word experts may perform differently, some *general* conclusions can be drawn from the experiments on the trial data. It appears to be beneficial to include lemma features, rather than just word features. However, adding Part-of-speech features tends to have a negative impact. For these local context features, the optimum context size is often two features to the left and two features to the right of the focus word, cf. (Hendrickx et al., 2002). The global keyword features perform well, but best results are achieved if they are not mixed with the local context features in one classifier.

An arbiter voting approach over multiple classifiers helps to smooth out errors and yields the highest scores (see Tables 1 and 2). When compared to the other participants, the UvT-WSD1 system, in the voting configuration, ranks first in the Word Sense Disambiguation task, for the two language pairs in which we participated.

References

Jordi Atserias, Bernardino Casas, Elisabet Comelles, Meritxell Gonzlez, Lluís Padró, and Muntsa Padró. 2006. FreeLing 1.3: Syntactic and semantic services in an open-source NLP library . In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*, Genoa, Italy. ELRA.

W. Daelemans, J. Zavrel, K. Van der Sloot, and A. Van den Bosch. 2009. TiMBL: Tilburg memory based learner, version 6.2, reference guide. Technical Report ILK 09-01, ILK Research Group, Tilburg University.

B. Decadt, V. Hoste, W. Daelemans, and A. Van den Bosch. 2004. GAMBL, genetic algorithm optimization of memory-based WSD. In R. Mihalcea and P. Edmonds, editors, *Proceedings of the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text (Senseval-3)*, pages 108–112, New Brunswick, NJ. ACL.

I. Hendrickx, A. Van den Bosch, V. Hoste, and W. Daelemans. 2002. Dutch word sense disambiguation: Optimizing the localness of context. In *Proceedings of the Workshop on word sense disambiguation: Recent successes and future directions*, pages 61–65, Philadelphia, PA.

V. Hoste, I. Hendrickx, W. Daelemans, and A. Van den Bosch. 2002. Parameter optimization for machine learning of word sense disambiguation. *Natural Language Engineering*, 8(4):311–325.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *In Proceedings of the Machine Translation Summit X ([MT]'05).*, pages 79–86.

Els Lefever and Veronique Hoste. 2010. Semeval 2010 task 3: Cross-lingual word sense disambiguation. In *Proceedings of the 5th International Workshop on Semantic Evaluations (SemEval-2010)*, Uppsala, Sweden.

Rada Mihalcea, Ravi Sinha, and Diana McCarthy. 2010. Semeval 2010 task 2: Cross-lingual lexical substitution. In *Proceedings of the 5th International Workshop on Semantic Evaluations (SemEval-2010)*, Uppsala, Sweden.

Hwee Tou Ng and Hian Beng Lee. 1996. Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach. In *ACL*, pages 40–47.

F.J. Och and H. Ney. 2000. Giza++: Training of statistical translation models. Technical report, RWTH Aachen, University of Technology.

A. Van den Bosch, G.J. Busser, S. Canisius, and W. Daelemans. 2007. An efficient memory-based morpho-syntactic tagger and parser for Dutch. In P. Dirix, I. Schuurman, V. Vandeghinste, , and F. Van Eynde, editors, *Computational Linguistics in the Netherlands: Selected Papers from the Seventeenth CLIN Meeting*, pages 99–114, Leuven, Belgium.

A. Van den Bosch. 2004. Wrapped progressive sampling search for optimizing learning algorithm parameters. In R. Verbrugge, N. Taatgen, and L. Schomaker, editors, *Proceedings of the Sixteenth Belgian-Dutch Conference on Artificial Intelligence*, pages 219–226, Groningen, The Netherlands.