KCDC: Word Sense Induction by Using Grammatical Dependencies and Sentence Phrase Structure

Roman Kern
Know-Center
Graz, Austria
rkern@know-center.at

Markus Muhr
Know-Center
Graz, Austria
mmuhr@know-center.at

Michael Granitzer
Graz University of Technology,
Know-Center
Graz, Austria
mgrani@know-center.at

Abstract

Word sense induction and discrimination (WSID) identifies the senses of an ambiguous word and assigns instances of this word to one of these senses. We have build a WSID system that exploits syntactic and semantic features based on the results of a natural language parser component. To achieve high robustness and good generalization capabilities, we designed our system to work on a restricted, but grammatically rich set of features. Based on the results of the evaluations our system provides a promising performance and robustness.

1 Introduction

The goal of the SemEval-2 word sense induction and discrimination task, see Manandhar et al. (2010), is to identify the senses of ambiguous nouns and verbs in an unsupervised manner and to label unseen instances of these words with one of the induced senses. The most common approach towards this task is to apply clustering or graph partitioning algorithms on a representation of the words that surround an ambiguous target word, see for example Niu et al. (2007) and Pedersen (2007). We followed this approach by employing a clustering algorithm to detect the individual senses, but focused on generating feature sets different to the mainstream approach. Our feature sets utilize the output of a linguistic processing pipeline that captures the syntax and semantics of sentence parts closely related with the target word.

2 System Overview

The base of our system is to apply a parser on the sentence in which the target word occurs. Contextual information, for example the sentences surrounding the target sentence, are currently not exploited by our system. To analyze the sentences we applied the Stanford Parser (Version 1.6.2), which is based on lexicalized probabilistic context free grammars, see Klein and Manning (2003). This open-source parser not only extracts the phrase structure of a given sentence, but also provides a list of so called grammatical relations (typed dependencies), see de Marneffe et al. (2006). These relations reflect the dependencies between the words within the sentence, for example the relationship between the verb and the subject. See Chen et al. (2009) for an application of grammatical dependencies for word sense disambiguation.

2.1 Feature Extraction

The phrase structure and the grammatical dependencies are sources for the feature extraction stage. To illustrate the result of the parser and feature extraction stages we use an example sentence, where the target word is the verb "file":

> Afterward , I watched as a butt-ton of good , but misguided people **filed** out of the theater , and immediately lit up a smoke .

2.1.1 Grammatical Dependency Features

The Stanford Parser provides 55 different grammatical dependency types. Figure 2 depicts the list of the grammatical dependencies identified by the Stanford Parser for the example sentence. Only a limited subset of these dependencies are selected to build the grammatical feature set. This subset has been defined based on preliminary tests on the trial dataset. For verbs only dependencies that represent the association of a verb with prepositional modifiers and phrasal verb particles are selected (`prep`, `prepc`, `prt`). If the verb is not associated with a preposition or particle, a synthetic "missing" feature is added instead (`!prep`, `!prt`). For nouns the selected dependencies are the prepositions (for head nouns that are the object of a preposition) and noun compound modifiers (`pobj`, `nn`).

Proceedings of the 5th International Workshop on Semantic Evaluation, ACL 2010, pages 351–354,
Uppsala, Sweden, 15-16 July 2010.

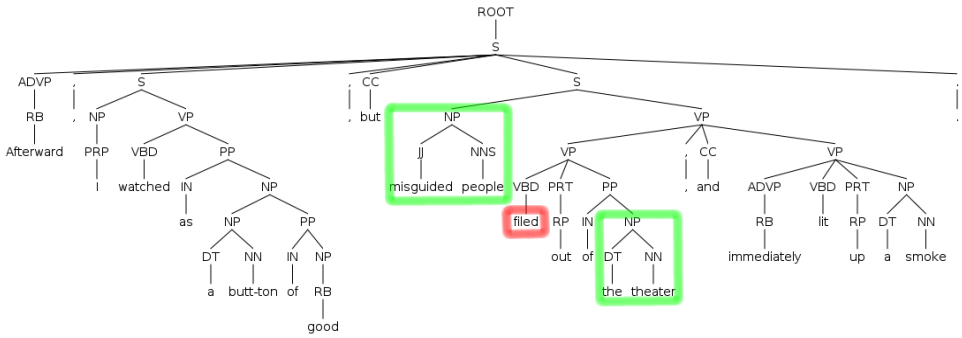

Figure 1: Phrase tree of the example sentence. The noun phrase "misguided people" is connected to the target word via the `nsubj` dependency and the phrase "the theater" is associated with the target verb via the `prep` and `pobj` dependencies.

relation	gov	dep
pobj	as-5	butt-ton-7
det	butt-ton-7	a-6
prep	butt-ton-7	of-8
nsubj	filed-14	people-13
prt	filed-14	out-15
prep	filed-14	of-16
cc	filed-14	and-20
conj	filed-14	lit-22
advmod	lit-22	immediat...
prt	lit-22	up-23
dobj	lit-22	smoke-25
pobj	of-16	theater-18
pobj	of-8	good-9
amod	people-13	misguide...
det	smoke-25	a-24
det	theater-18	the-17
advmod	watched-4	Afterwar...
nsubj	watched-4	I-3
prep	watched-4	as-5
cc	watched-4	but-11
conj	watched-4	filed-14

Figure 2: List of grammatical dependencies as detected by the Stanford Parser.

If the noun is associated with a verb the grammatical dependencies of this verb are also added to the feature set.

The name of the dependency and the word (i.e. preposition or particle) are used to construct the grammatical feature. The different features are weighted. The weights have been derived from their frequencies within the trial dataset and listed in table 1. For the example sentence the extracted grammatical features are:

```
'out', 'of', prep, prt
```

2.1.2 Phrase Term Features

The second set of features are generated from the sentence phrase structure. In figure 1 the parse tree for the example sentence is depicted.

Again we tried to keep the feature set as small as possible. Starting with the target word only phrases that are directly associated with the ambiguous word are selected. To identify these phrases the grammatical dependencies are exploited. For nouns as target words the associated verb is searched at first. Given a verb the phrases containing the head noun of a subject or object relationship are identified. If the verb is accompanied by a preposition, the phrase carrying the object of the preposition is also added. All nouns and adjectives from these these phrases are then collected. The phrase words together with the verb, prepositions and particles are lemmatized using tools also provided by the Stanford Parser project.

Feature	Weight
prepc, prt, nn, pobj	0.9
prep	0.45
!prep, !prt	0.5
'prepositions', 'particles'	0.97

Table 1: Weights of the grammatical features, which were derived from their distribution within the trial dataset.

The weights of the phrase term features are based on the frequency of the words within the training dataset, where N is the total number of sentences and N_f is the number of sentences in which the lemmatized phrase term occurs in:

$$weight_f = log(\frac{N}{N_f + 1}) + 1 \qquad (1)$$

In our example sentence the extracted phrase term features are:

```
of, misguided, file, theater, people, out
```

2.2 Phrase Term Expansion

The feature space of the phrase terms is expected to be very sparse. Additionally different phrase terms may have similar semantics. Therefore the phrase terms are optionally expanded with associated terms, where semantically similar terms should be associated with the same terms.

To calculate the statistics for term expansion we used the training dataset (although other datasets

would be more suitable for this purpose). The dataset is split into sentences. Stopwords and rarely used words, which occur in less than 3 sentences, were removed. The remaining words were finally lemmatized. For a given phrase term the top 100 associated terms are used to build the feature set. The association weight between two terms is based on the Pointwise Mutual Information:

$$weight_{pmi}(t_i, t_j) = \frac{log_2(\frac{P(t_i|t_j)}{P(t_j)})}{log_2(\frac{1}{P(t_j)})} \quad (2)$$

For example the top 10 associated terms for `theater` are:

```
theater.n, movie.n, opera.n,
vaudeville.n, wnxt-abc.n, imax.n,
orpheum.n, pullulate.v, projector.n,
psychomania.n
```

2.3 Sense Induction

To detect the individual senses within the training dataset we applied unsupervised machine learning techniques. For each ambiguous word a matrix - $M_{|Instances|\times|Features|}$ - is created and a clustering algorithm is applied, namely the Growing k-Means, see Daszykowski et al. (2002). This algorithm needs the number of clusters and centroids as initialization parameters, where the initial centroids are calculated using a directed random seed finder as described in Arthur and Vassilvitskii (2007). We used the Jensen-Shannon Divergence function for the grammatical dependency features and the Cosine Similarity for the phrase term feature sets as relatedness function.

For each cluster number we re-run the clustering with different random initial centroids (30 times) and for each run we calculate a cluster quality criterion. The overall cluster quality criterion is the mean of all feature quality criteria, which are calculated based on the set of clusters the feature occurs in - C_f - the number of instances of each cluster - N_c - and the number of instances within a cluster where the feature occurs in - $N_{c,f}$:

$$FQC_f = \frac{weight_f}{|C_f|} * \sum_{c \in C_f} \frac{N_{c,f}}{N_c} \quad (3)$$

$$QC_{run} = \overline{FQC_f} \quad (4)$$

The cluster quality criterion is calculated for each run and the combination of the mean and standard deviations are then used to calculate a stability criterion to detect the number of clusters, which is based on the intuition that the correct cluster count yields the lowest variation of QC values:

$$SC_k = \frac{mean(QC)}{stdev(QC)} \quad (5)$$

Starting with two clusters the number of clusters is incremented until the stability criterion starts to decline. For the cluster number with the highest stability criterion the run with the highest quality criterion is selected as final clustering solution. The result of the sense induction processing is a list of centroids for the identified clusters.

2.4 Sense Assignment

The final processing step is to assign an instance of an ambiguous word to one of the pre-calculated senses. The sentence with the target word is processed exactly like the training sentences to generate a set of features. Finally the word is assigned to the sense cluster with the maximum relatedness.

3 System Configurations & Results

Our system can be configured to use a combination of feature sets for the word sense induction and discrimination calculations: a) *KCDC-GD*: Grammatical dependency features, b) *KCDC-PT*: Phrase terms features, c) *KCDC-PC*: Expanded phrase term features, d) *KCDC-PCGD*: All training sentences are first processed by using the expanded phrase term features and then by using the grammatical dependency features with an additional feature that encodes the cluster id found by the phrase features.

In the evaluation we also submitted multiple runs of the same configuration[1] to assess the influence of the random initialization of the clustering algorithm. Judging from the results the random seeding has no pronounced impact and it influence should decrease when the number of clustering runs for each cluster number is increased.

All configurations found on average about 3 senses for target words in the test set (2.8 for verbs, 3.3 for nouns), with exception of the *KCDC-PT* configuration which identified only 1.5 senses on average. In the gold standard the number of senses for verbs is 3.12 and for nouns 4.46, which shows that the stability criterion tends to underestimate the number of senses slightly.

To compare the performance of the different configurations, one can use the average rank within the evaluation result lists. Judging from the

[1] labeled KCDC-GD-2, KCDC-GDC for configuration 'a' and KCDC-PC-2 for the configuration 'c'

rankings, the configurations that utilize the grammatical dependencies and the expanded phrase terms provide similar performance. The configuration that takes the phrase terms directly as features comes in last, which is expected due to the sparse nature of the feature representation and the low number of detected senses.

Comparing the performance of our system with the two baselines shows that our system did outperform the random baseline in all evaluation runs and the most frequent baseline (MFS) in all runs with the exception of the F-Score based unsupervised evaluation, where the MFS baseline has not been beaten by any system. Although none of our submitted configurations was ranked first in any of the evaluations, their ranking was still better than average, with the exception of the *KCDC-PT* configuration.

Another observation that can be made is the difference in performance between nouns and verbs. Our system, especially the grammatical dependency based configurations, is tailored towards verbs. Therefore the better performance of verbs in the evaluation is in line with the expectations.

When looking at the results of the individual target words one can notice that for a set of words the quality of the sense detection is above average. For 16 of the 100 words a V-Measure of more than 30% in at least one configuration was achieved (average: 7.8%)[2]. This can be seen as indicator that our selection of features is effective for a specific group of words. For the remaining words an according feature set has to be developed in future work.

4 Conclusion

For the SemEval 2010 word sense induction and discrimination task we have tried to build a system that uses a minimal amount of information while still providing a competitive performance. This system contains a parser component to analyze the phrase structure of a sentence and the grammatical dependencies between words. The extracted features are then clustered to detect the senses of ambiguous words. In the evaluation runs our system did demonstrate a satisfying performance for a number of words.

The design of our system offers a wide range of possible enhancements. For example the integration of preposition disambiguation and noun-phrase co-reference resolution could help to further improve the word sense discrimination effectiveness.

[2]The best performing target words are: `root.v`, `presume.v`, `figure.v`, `weigh.v`, `cheat.v`

Acknowledgments

The Know-Center is funded within the Austrian COMET Program - Competence Centers for Excellent Technologies - under the auspices of the Austrian Federal Ministry of Transport, Innovation and Technology, the Austrian Federal Ministry of Economy, Family and Youth and by the State of Styria. COMET is managed by the Austrian Research Promotion Agency FFG. Results are partially funded by the EUROSTARS project 4811 MAKIN'IT.

References

D. Arthur and S. Vassilvitskii. 2007. k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, page 10271035. Society for Industrial and Applied Mathematics Philadelphia, PA, USA.

Ping Chen, Wei Ding, Chris Bowes, and David Brown. 2009. A fully unsupervised word sense disambiguation method using dependency knowledge. *Human Language Technology Conference*.

M Daszykowski, B Walczak, and D L Massart. 2002. On the optimal partitioning of data with K-means, growing K-means, neural gas, and growing neural gas. *Journal of chemical information and computer sciences*, 42(6):1378–89.

M.C. de Marneffe, B. MacCartney, and C.D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *LREC 2006*.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - ACL '03*, pages 423–430.

Suresh Manandhar, Ioannis P. Klapaftis, Dmitriy Dligach, and Sameer S. Pradhan. 2010. SemEval-2010 Task 14: Word Sense Induction & Disambiguation. In *Proceedings of SemEval-2*, Uppsala, Sweden, ACL.

Zheng-yu Niu, Dong-hong Ji, and Chew-lim Tan. 2007. I2R: Three Systems for Word Sense Discrimination, Chinese Word Sense Disambiguation, and English Word Sense Disambiguation. In *Proceedings of the 4th International Workshop on Semantic Evaluations*. ACL.

T. Pedersen. 2007. Umnd2: Senseclusters applied to the sense induction task of senseval-4. In *Proceedings of the 4th International Workshop on Semantic Evaluations*. ACL.