Co-occurrence Cluster Features for Lexical Substitutions in Context

Chris Biemann
Powerset (a Microsoft company)
475 Brannan St Ste. 330
San Francisco, CA 94107, USA
cbiemann@microsoft.com

Abstract

This paper examines the influence of features based on clusters of co-occurrences for supervised Word Sense Disambiguation and Lexical Substitution. Co-occurrence cluster features are derived from clustering the local neighborhood of a target word in a co-occurrence graph based on a corpus in a completely unsupervised fashion. Clusters can be assigned in context and are used as features in a supervised WSD system. Experiments fitting a strong baseline system with these additional features are conducted on two datasets, showing improvements. Co-occurrence features are a simple way to mimic Topic Signatures (Martínez et al., 2008) without needing to construct resources manually. Further, a system is described that produces lexical substitutions in context with very high precision.

1 Introduction

Word Sense Disambiguation (WSD, see (Agirre and Edmonds, 2006) for an extensive overview) is commonly seen as an enabling technology for applications like semantic parsing, semantic role labeling and semantic retrieval. Throughout recent years, the Senseval and Semeval competitions have shown that a) WordNet as-is is not an adequate semantic resource for reaching high precision and b) supervised WSD approaches outperform unsupervised (i.e. not using sense-annotated examples) approaches. Due to the manual effort involved in creating more adequate word sense inventories and sense-annotated training data, WSD has yet to see its prime-time in real world applications.

Since WordNet's sense distinctions are often too fine-grained for allowing reliable distinctions by machines and humans, the OntoNotes project (Hovy et al., 2006) conflated similar WordNet senses until 90% inter-annotator agreement on sense-labelling was reached. The SemEval 2007 lexical sample task employs this ”coarse-grained” inventory, which allows for higher system performance.

To alleviate the bottleneck of sense-labelled sentences, (Biemann and Nygaard, 2010) present an approach for acquiring a sense inventory along with sense-annotated example usages using crowdsourcing, which makes the acquisition process cheaper and potentially quicker.

Trying to do away with manual resources entirely, the field of Word Sense Induction aims at inducing the inventory from text corpora by clustering occurrences or senses according to distributional similarity, e.g. (Veronis, 2004). While such unsupervised and knowledge-free systems are capable of discriminating well between different usages, it is not trivial to link their distinctions to existing semantic resources, which is often necessary in applications.

Topic Signatures (Martínez et al., 2008) is an attempt to account for differences in relevant topics per target word. Here, a large number of contexts for a given sense inventory are collected automatically using relations from a semantic resource, sense by sense. The most discriminating content words per sense are used to identify a sense in an unseen context. This approach is amongst the most successful methods in the field. It requires, however, a semantic resource of sufficient detail and size and a sense-labeled corpus to estimate priors from the sense distribution. Here, a similar approach is described that uses an unlabeled

Proceedings of the 2010 Workshop on Graph-based Methods for Natural Language Processing, ACL 2010, pages 55–59,
Uppsala, Sweden, 16 July 2010. ©2010 Association for Computational Linguistics

corpus alone for unsupervised topic signature acquisition using graph clustering, not relying on the existence of a WordNet. Unlike in previous evaluations like (Agirre et al., 2006), parameters for word sense induction are not optimized globally, but instead several parameter settings are offered as features to a Machine Learning setup.

Experimental results are provided for two datasets: the Semeval-2007 lexical sample task (Pradhan et al., 2007) and the Turk bootstrap Word Sense Inventory (TWSI[1], (Biemann and Nygaard, 2010)).

2 Cluster Co-occurrence Features

2.1 Graph Preperation and Parameterization

Similar to the approach in (Widdows and Dorow, 2002), a word graph around each target word is constructed. In this work, sentence-based co-occurrence statistics from a large corpus are used as a basis to to construct several word graphs for different parameterizations. Significant co-occurrences between all content words (nouns, verbs, adjectives as identified by POS tagging) are computed from a large corpus using the tinyCC[2] tool. The full word graph for a target word is defined as all words significantly co-occurring with the target as nodes, with edge weights set to the log-likelihood significance of the co-occurrence between the words corresponding to nodes. Edges between words that co-occur only once or with significance smaller than 6.63 (1% confidence level) are omitted.

Aiming at different granularities of usage clusters, the graph is parameterized by a size parameter t and a density parameter n: Only the most significant t co-occurrences of the target enter the graph as nodes, and an edge between nodes is drawn only if one of the corresponding words is contained in the most significant n co-occurrences of the other.

2.2 Graph Clustering Parameterization

As described in (Biemann, 2006), the neighborhood graph is clustered with Chinese Whispers. This efficient graph clustering algorithm finds the numbers of clusters automatically and returns a partition of the nodes. It is initialized by assigning different classes to all nodes in the graph. Then, a number of local update steps are performed, in which a node inherits the predominant class in its neighborhood. At this, classes of adjacent nodes are weighted by edge weight and downweighted by the degree (number of adjacent nodes) of the neighboring node. This results in hard clusters of words per target, which represent different target usages.

Downweighting nodes by degree is done according to the following intuition: nodes with high degrees are probably very universally used words and should be less influential for clustering. Three ways of node weighting are used: (a) dividing the influence of a node in the update step by the degree of the node, (b) dividing by the natural logarithm of the degree + 1 and (c) not doing node weighting. The more aggressive the downweighting, the higher granularity is expected for the clustering.

It is emphasized that no tuning techniques are applied to arrive at the 'best' clustering. Rather, several clusterings of different granularities as *features* are made available to a supervised system. Note that this is different from (Agirre et al., 2006), where a single global clustering was used *directly* in a greedy mapping to senses.

2.3 Feature Assignment in Context

For a given occurrence of a target word, the overlap in words between the textual context and all clusters from the neighborhood graph is measured. The cluster ID of the cluster with the highest overlap is assigned as a feature. This can be viewed as a word sense induction system in its own right.

At this, several clusterings from different parameterizations are used to form distinct features, which enables the machine learning algorithm to pick the most suitable cluster features per target word when building the classification model.

2.4 Corpora for Cluster Features

When incorporating features that are induced using large unlabeled corpora, it is important to ensure that the corpus for feature induction and the word sense labeled corpus are from the same domain, ideally from the same source.

Since TWSI has been created from Wikipedia, an English Wikipedia dump from January 2008 is used for feature induction, comprising a total of 60 million sentences. The source for the lexical sample task is the Wall Street Journal, and since the

[1] full dataset available for download at http://aclweb.org/aclwiki/index.php?title=Image:TWSI397.zip

[2] http://beam.to/biem/software/TinyCC2.html

76,400 sentences from the WSJ Penn Treebank are rather small for co-occurrence analysis, a 20 Million sentence New York Times corpus was used instead.
For each corpus, a total of 45 different clusterings were prepared for all combinations of t={50,100,150,200,250}, n={50,100,200} and node degree weighting options (a), (b) and (c).

3 Experimental Setup

3.1 Machine Learning Setup

The classification algorithm used throughout this work is the AODE (Webb et al., 2005) classifier as provided by the WEKA Machine Learning software (Hall et al., 2009). This algorithm is similar to a Naïve Bayes classifier. As opposed to the latter, AODE does not assume mutual independence of features but models correlations between them explicitly, which is highly desirable here since both baseline and co-occurrence cluster features are expected to be highly correlated. Further, AODE handles nominal features, so it is directly possible to use lexical features and cluster IDs in the classifier. AODE showed superior performance to other classifiers handling nominal features in preliminary experiments.

3.2 Baseline System

The baseline system relies on 15 lexical and POS-based nominal features: word forms left and right from target, POS sequences left and right bigram around target, POS tags of left and right word from target, and POS tag of target, two left and two right nouns from target, left and right verbs from target and left and right adjectives from target.

3.3 Feature Selection

To determine the most useful cluster co-occurrence features, they were added to the baseline features one at the time, measuring the contribution using 10-fold cross validation on the training set. Then, the best k single cluster features for k={2,3,5,10} were added together to account for a range of different granularities. The best performing system on the lexical sample training data resulted in a 10-fold accuracy of 88.5% (baseline: 87.1%) for k=3. On the 204 ambiguous words (595 total senses with 46 sentences per sense on average) of the TWSI only, the best system was found at k=5 with a 10-fold accuracy of 83.0% (baseline: 80.7%, MFS: 71.5%). Across the board, all single co-occurrence features improve over the baseline, most of them significantly.

System	F1
NUS-ML	88.7% ± 1.2
top3 cluster, optimal F1	88.0% ± 1.2
top3 cluster, max recall	87.8% ± 1.2
baseline, optimal F1	87.5% ± 1.2
baseline, max recall	87.3% ± 1.2
UBC-ALM	86.9% ± 1.2

Table 1: Cluster co-occurrence features and baseline in comparison to the best two systems in the SemEval 2007 Task 17 Lexical Sample evaluation (Pradhan et al., 2007). Error margins provided by the task organizers.

4 Results

4.1 SemEval 2007 lexical sample task

The system in the configuration determined above was trained on the full training set and applied it to the test data provided bt the task organizers.
Since the AODE classifier reports a confidence score (corresponding to the class probability for the winning class at classification time), it is possible to investigate a tradeoff between precision and recall to optimize the F1-value[3] used for scoring in the lexical sample task.

It is surprising that the baseline system outperforms the second-best system in the 2007 evaluation, see Table 1. This might be attributed to the AODE classifier used, but also hints at the power of nominal lexical features in general.
The co-occurrence cluster system outperforms the baseline, but does not reach the performance of the winning system. However, all reported systems fall into each other's error margins, unlike when evaluating on training data splits. In conclusion, the WSD setup is competitive to other WSD systems in the literature, while using only minimal linguistic preprocessing and no word sense inventory information beyond what is provided by training examples.

[3] $F1 = (2 \cdot precision \cdot recall)/(precision + recall)$

	Substitutions		
	Gold	**System**	**Random**
YES	469 (93.8%)	456 (91.2%)	12 (2.4%)
NO	14 (2.8%)	27 (5.4%)	485 (97.0%)
SOMEWHAT	17 (3.4%)	17 (3.4%)	3 (0.6%)

Table 2: Substitution acceptability as measured by crowdsourcing for TWSI gold assignments, system assignments and random assignments.

coverage	**YES**	**NO**
100%	91.2%	5.4%
95%	91.8%	3.4%
90%	93.8%	2.9%
80%	94.8%	2.0%
70%	95.7%	0.9%

Table 3: Substitution acceptability in reduced coverage settings. SOMEWHAT class accounts for percentage points missing to 100%.

4.2 Substitution Acceptability

For evaluating substitution acceptability, 500 labeled sentences from the overall data (for all 397 nouns, not just the ambiguous nouns used in the experiments above) were randomly selected. The 10-fold test classifications as described above were used for system word sense assignment. The three highest ranked substitutions per sense from the TWSI are supplied as substitutions.

In a crowdsourcing task, workers had to state whether the substitutions provided for a target word in context do not change the meaning of the sentence. Each assignment was given to three workers.

Since this measures both substitution quality of the TWSI and the system's capability of assigning the right sense, workers were also asked to score the substitutions for the gold standard assignments of this data set. For control, random substitution quality for all sentences is measured.

Table 2 shows the results for averaging over the worker's responses. For being counted as belonging to the YES or NO class, the majority of workers had to choose this option, otherwise the item was counted into the SOMEWHAT class.

The substitution quality of the gold standard is somewhat noisy, containing 2.8% errors and 3.4% questionable cases. Despite this, the system is able to assign acceptable substitutions in over 91% of cases, questionable substitutions for 3.4% at an error rate of only 5.4%. Checking the positively judged random assignments, an acceptable substitution was found in about half of the cases by the author, which allows to estimate the worker noise at about 1%.

When using confidence values of the AODE classifier to control recall as reported in Table 3, it is possible to further reduce error rates, which might e.g. improve retrieval applications.

5 Conclusion

A way to improve WSD accuracy using a family of co-occurrence cluster features was demonstrated on two data sets. Instead of optimizing parameters globally, features corresponding to different granularities of induced word usages are made available in parallel as features in a supervised Machine Learning setting.

Whereas the contribution of co-occurrence features is significant on the TWSI, it is not significantly improving results on the SemEval 2007 data. This might be attributed to a larger number of average training examples in the latter, making smoothing over clusters less necessary due to less lexical sparsity.

We measured performance of our lexical substitution system by having the acceptability of the system-provided substitutions in context manually judged. With error rates in the single figures and the possibility to reduce error further by sacrificing recall, we provide a firm enabling technology for semantic search.

For future work, it would be interesting to evaluate the full substitution system based on the TWSI in a semantic retrieval application.

References

Eneko Agirre and Philip Edmonds, editors. 2006. *Word Sense Disambiguation: Algorithms and Applications*, volume 33 of *Text, Speech and Language Technology*. Springer, July.

Eneko Agirre, David Martínez, Oier L. de Lacalle, and Aitor Soroa. 2006. Evaluating and optimizing the parameters of an unsupervised graph-based wsd algorithm. In *Proceedings of TextGraphs: the Second Workshop on Graph Based Methods for Natural Language Processing*, pages 89–96, New York City. Association for Computational Linguistics.

Chris Biemann and Valerie Nygaard. 2010. Crowdsourcing WordNet. In *Proceedings of the 5th Global WordNet conference*, Mumbai, India. ACL Data and Code Repository, ADCR2010T005.

Chris Biemann. 2006. Chinese whispers - an efficient graph clustering algorithm and its application to natural language processing problems. In *Proceedings of the HLT-NAACL-06 Workshop on Textgraphs-06*, New York, USA.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA data mining software: An update. *SIGKDD Explorations*, 11(1).

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: The 90% solution. In *Proceedings of HLT-NAACL 2006*, pages 57–60.

David Martínez, Oier Lopez de Lacalle, and Eneko Agirre. 2008. On the use of automatically acquired examples for all-nouns word sense disambiguation. *J. Artif. Intell. Res. (JAIR)*, 33:79–107.

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, June. Association for Computational Linguistics.

Jean Veronis. 2004. Hyperlex: lexical cartography for information retrieval. *Computer Speech & Language*, 18(3):223–252.

G. Webb, J. Boughton, and Z. Wang. 2005. Not so Naive Bayes: Aggregating one-dependence estimators. *Machine Learning*, 58(1):5–24.

Dominic Widdows and Beate Dorow. 2002. A graph model for unsupervised lexical acquisition. In *Proceedings of the 19th international conference on Computational linguistics*, pages 1–7, Morristown, NJ, USA. Association for Computational Linguistics.