A Comparison of Knowledge-based Algorithms for Graded Word Sense Assignment

Annemarie Friedrich *Nikos Engonopoulos* *Stefan Thater* *Manfred Pinkal*
Department of Computational Linguistics
Saarland University
{afried, nikolaos, stth, pinkal}@coli.uni-saarland.de

ABSTRACT
Standard word sense disambiguation (WSD) data sets annotate each word instance in context with exactly one sense of a predefined inventory, and WSD systems are traditionally evaluated with regard to how good they are at picking this sense. Recently, the notion of graded word sense assignment (GWSA) has gained attention as a more natural view of the contextual specification of word meaning; multiple senses may apply simultaneously to one instance of a word, and they may be applicable to different degrees. In this paper, we apply three different WSD algorithms to the task of GWSA. The three models belong to the class of knowledge-based models in the WSD terminology; they are unsupervised in the sense that they do not depend on annotated training material. We evaluate the models on two recently published GWSA data sets. We find positive correlations with the human judgments for all models, and develop a metric based on the notion of accuracy that highlights differences in the behaviors of the models.

KEYWORDS: lexical semantics, graded word senses, knowledge-based disambiguation.

Proceedings of COLING 2012: Posters, pages 329–338,
COLING 2012, Mumbai, December 2012.

1 Introduction

The problem of *word sense disambiguation* (WSD) is a central topic in computational linguistics, with a long-standing, rich history of research (see McCarthy, 2009; Navigli, 2009). Typically, the WSD task is designed such that each target word in context is assigned a single word sense from a predefined sense inventory. However, several word senses may be simultaneously present in a contextual instance of a word, which holds in particular in connection with fine-grained sense inventories, like the one provided by WordNet (Fellbaum, 1998). The single-sense restriction typically leads to a somewhat arbitrary overspecification of word meaning, which may be detrimental to the use of WSD systems in practical applications. Moreover, both agreement between human annotators and accuracy of WSD systems tend to be rather low, which stands in contrast to the strong intuition that words in context generally have a well-understood meaning.

Recently, the notion of *graded word sense assignment* (GWSA) has been brought into discussion by Erk et al. (2009, 2012), and two closely related GWSA data sets are now available. The underlying assumption of GWSA is that a word in context may in fact evoke more than one sense, and the different senses may participate in the meaning of the word to different degrees. To produce the aforementioned data sets, annotators were presented target instances, i.e., lemmas in the context of a sentence, and asked to assign a value, which indicates the applicability of the sense in the context, on a scale from 1 to 5 to each WordNet sense of the lemma independently. The annotation method allows more than one word sense for a given target instance to be assigned a high applicability score, and it induces an ordering of the word senses on the level of single instances. Erk et al. (2009) give the example of "paper" occurring in a sentence which clearly identifies a scientific context. All three annotators agree that the WordNet sense *scholarly article* fully applies and consistently assign a score of 5. However, the senses *essay* and *medium for written communication* are also assigned high scores by some of the annotators. This reflects these annotators' intuitions that several senses apply simultaneously, and induces an ordering of the senses' applicabilities.

A first, supervised, computational model for GWSA is presented by Erk and McCarthy (2009). In this paper, we explore models that are unsupervised in the sense that they do not depend on annotated training material; in the WSD terminology, they belong to the class of knowledge-based WSD systems. More specifically, we address the task of ranking the WordNet senses of a lemma for each of its instances, according to the degree of applicability of the respective senses in context. We evaluate our models against the data sets provided by Erk et al. (2009, 2012), and use the ranking induced by the average scores for each word sense as a gold standard. We carry out the evaluation for three different systems: two related models, which are based on the individual similarity scores between the contextualized vector representation of a target word in context and vector representations computed for the respective word senses (Thater et al., 2011; Li et al., 2010), plus a reimplementation of the approach of Sinha and Mihalcea (2007), a representative of the larger class of graph-based approaches to WSD. Our major findings are first, that the knowledge-based systems show positive correlation with the human judgments, and second, that there are interesting differences in performance between the different types of systems according to our metric of Adjusted Accuracy.

2 Related Work

The only WSD system that has been evaluated on the full GWSA data set of Erk et al. (2009) so far is the supervised model of Erk and McCarthy (2009). Thater et al. (2010) describe an approach to unsupervised GWSA on the basis of a syntactically informed distributional similarity

model. The evaluation was carried out for three selected verb lemmas, and therefore has the character of a case study only. The study of Jurgens (2012), which explores the application of word sense induction techniques to GWSA, has a similar status: Since he needs a large part of the GWSA data set as a sense mapping corpus, only a very small amount of data is left for evaluation.

3 Modeling

This section reviews the three knowledge-based WSD algorithms that we use in our study, and which we chose for the following reasons: (1) They are knowledge-lean, i.e., the only resource required is a semantic lexicon (such as WordNet), and they can be implemented quickly. (2) They exhibit state-of-the-art performance on the SemEval-2007 coarse-grained WSD task.

3.1 Vector Space-based WSD System

We use the vector-space model (VSM) of Thater et al. (2011), which is closely related to the models of Thater et al. (2010) and Erk and Padó (2008). The general idea behind VSMs of word meaning is to represent words by vectors in a high-dimensional space. These vectors record co-occurrence statistics with context words in a large unlabeled text corpus, and their relative directions are taken to indicate semantic similarity. The particular model used in our experiments is the one of Thater et al. (2011), which provides context-specific (contextualized) vectors for words in their syntactic context. It can be applied to WSD and GWSA in a straightforward way: given a target word in a sentential context, we extract a set of *sense paraphrases* for each sense of the target from WordNet. We then compute the cosine similarities of all sense paraphrases and the contextualized vector of the target word, and set the similarity of the sense to be the average of the best two sense paraphrases. In the case of standard WSD, the VSM predicts the sense with the highest score; in the case of GWSA, the scores assigned to the senses induce a ranking. In rare cases, the VSM fails to make predictions, i.e., when the dependency tree for the input sentence does not assign the correct POS to the target word, or when no useful sense paraphrases can be extracted from WordNet.

3.2 Topic Model-based WSD System

Li et al. (2010) use topic models (Blei et al., 2003), which represent text corpora using generative probability distributions, as the central component of their WSD system. Topics are distributions over words and each document is modeled as a mixture of latent topics. Li et al. (2010) extract one *sense paraphrase* per word sense from WordNet. The topic model is used to estimate a vector of the topic distribution for the context of the target word (usually the sentence in which it occurs) and a vector for the sense paraphrase of the candidate sense. The cosine between these vectors is taken as the final score for the word sense. This algorithm naturally produces a ranking of word senses. We closely follow the experimental settings (for Model II) reported by Li et al. (2010), but we were not able to fully reproduce their system. For SemEval-2007, Li et al. (2010) report an F1-measure of 79.99% for their Topic Model system. Our reimplementation achieved an F-measure of 71.7%. Hence, the Topic Models approach might yield better performance using different parameter settings. We noticed that due to the sampling step inside the algorithm, the results varied by small, but non-negligible, amounts. We thus sum up the scores produced by the system across multiple (ten) runs in order to predict a more reliable ranking. This results in a slight increase of performance.

3.3 Graph-based WSD System

To date, many graph-based WSD algorithms have been proposed, (among others by Sinha and Mihalcea, 2007; Agirre and Soroa, 2009; Navigli and Lapata, 2010; Tsatsaronis et al., 2010; Ponzetto and Navigli, 2010). We chose to reimplement the approach of Sinha and Mihalcea (2007) for several reasons. First, it is based on the PageRank algorithm, which is easy to understand and implement; second, a reference implementation was made available by the authors, which allowed for clarification in several issues; and third, its performance is robust. The algorithm consists of the following steps, which we illustrate using Figure 1. (1) *Construction of the graph.* When disambiguating a word (e.g. "order"), a graph is built using a context of N (2 in the example) content words on either side of the word. For each content word, the admissible word senses are added to the graph as nodes. Undirected edges are introduced between nodes that were not introduced for the same word and whose content words are not more than M (2 in the example) content words apart in the surface string. The edge weights are determined using the Extended Lesk Similarity (Banerjee, 2003) between the two synsets of the two nodes' word senses.[1] The setting we used for the SemEval-2007 experiments was N=6 and M=3; for the GWSA task, we report results for N=2 and M=2. The parameters were tuned on the respective data sets. (2) *Scoring using a graph-based centrality algorithm* (ten iterations of PageRank). (3) *Assignment of word senses.* In a standard WSD setting, the system picks the sense of the target word whose node has been assigned the highest score. In GWSA, we simply assign the scores of the respective nodes to the senses.

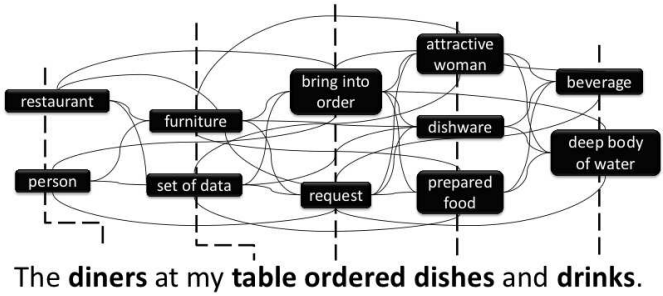

Figure 1: Example graph used in the PageRank algorithm.

4 Evaluation

4.1 Data Sets

WSsim-1: (Erk and McCarthy, 2009) present the first data set for the evaluation of GWSA. A total of 430 sentences for 11 different lemmas were extracted from SemCor and Senseval-3. Three untrained annotators provided judgments of the applicability of word senses of the lemmas in the context of the sentence on a scale from 1 to 5, where 1 means that the sense is not present at all in the sentence and 5 means that the sense totally matches the meaning of the word in the context. We refer to the task of ranking the senses of a word (lemma) in the context of a particular sentence as the *lemma-sentence ranking task*.

[1] We are using the WordNet::Similarity toolkit of (Pedersen et al., 2004). We also experimented with other sense similarity measures, but the method suggested by Sinha and Mihalcea (2007) worked best with PageRank.

WSsim-2: In this round of data collection, eight annotators judged the applicability of the senses of 26 lemmas in 10 sentences each, resulting in a set of 260 sentences (Erk et al., 2012). Otherwise, the annotation procedure was identical to WSsim-1.

4.2 Correlation Analysis of Sense Ranking

Erk and McCarthy (2009) propose Spearman's rank correlation coefficient (ρ) as a measure of a system's performance on the GWSA task. ρ compares two rankings while abstracting away from the absolute values of judgments. We used the R mathematical package to compute ρ for the rankings of the senses for each lemma-sentence task and then average across all sentences (see Table 1). As an upper bound, we report the correlations achieved by the human annotators (compare to Tables 9 and 10 of (Erk et al., 2012)). Significance is hard to show due to the small number of senses to be ranked (on average 6.1 senses per lemma in WSsim-1 and 10.6 in WSsim-2). The upper part of Table 1 shows the performance of the supervised system reported by Erk and McCarthy (2009), Prototype 2/N, as well as a sense frequencies baseline, whose sense frequencies have been estimated on SemCor and the training part of Senseval-3 (minus the sentences used in WSsim-1), while the lower part of the table shows the correlations achieved by our implementations of knowledge-based systems. Erk and McCarthy test their system only on the sentences for 8 out of the 11 lemmas of WSsim-1 and the numbers are therefore not directly comparable. The supervised system (Prototype 2/N) performs best, but the knowledge-based systems also show meaningful correlations with the human judgments. The VSM performs surprisingly well; the Topic Models system outperforms the PageRank system. It is worth noting that the sense frequencies baseline performs much better on WSsim-1 than on WSsim-2 for the lemma-sentence task, the reason being that the frequencies have been estimated in-domain for WSsim-1.

	WSsim-1		WSsim-2	
Model	ρ	sign.	ρ	sign.
Average of humans*	0.555	30.4	0.641	48.3
Prototype 2/N (E&K)	0.478	22.8	-	-
Sense Frequencies (SF)	0.357	10.7	0.245	14.2
VSM (Thater et al.)$\star$	0.305	12.7	0.389	21.4
Topic Models (Li et al.)$\bowtie\dagger$	0.241	11.6	0.256	15.0
PageRank (Sinha et al.)$\dagger$	0.210	4.0	0.097	4.6

Table 1: Spearman's rank correlation coefficient (ρ) by lemma-sentence compared to the average scores of all human annotators. The columns labelled "sign." show the percentage of the sentences in which the sense ranking correlation was significant. *Correlation of scores of one annotator with the average scores of the other annotators (omitting cases where annotators did not produce valid rankings). $\star$Performance of the VSM is reported on the 99% (WSsim-1) and 93% (WSsim-2) of the sentences for which the model creates a ranking. †Our reimplementations.

5 Analysis

5.1 Analysis of Data

As we have seen in section 4.2, the correlation between the human annotators is by no means perfect: it is hard to quantify the actual degree of applicability on the scale proposed by Erk et al. (2012) in many cases. In order to gain some more understanding about how the human

annotators use the scale and to what extent the (correlation) analysis of systems using the WSsim-2 data set is meaningful, we created the plot shown in Figure 2.

In the lemma-sentence task, two annotators define the same ranking for a pair of senses if one assigns the scores 3-4 and the other assigns 4-5. For this reason, we look at the scores given to a sense pair by one annotator, and whether the ranking of these two senses is concordant with the ranking of the average of all other annotators. Each pair of senses of the ranking of one annotator is sorted into one of the diagram's cells depending on the scores assigned to the two senses; if there is no tie, we find the position on the y-axis using the higher score and the position on the x-axis using the lower score, thus producing a diagonal matrix. We then compare the ranking of the first annotator to the ranking resulting from the average of the other annotators and increment the cell's count if the pair is concordant. In each cell, we add up the numbers of concordant pairs over all the annotators. Finally, we divide each cell's count by the total number of pairs that fell into the cell in order to decrease the bias caused by score combinations that occur more often. From this analysis, we can conclude that humans agree more often on the ordering of two senses if they assign scores at the two ends of the scale (the cell 5-1 has the highest proportion of concordant pairs), but that they use the intermediate scores rather interchangeably. There is high agreement for cell 1-1 (88.5%) out of the 100,217 pairs that fell into this cell. Cell 2-2 also shows high agreement, but note that only 1,684 pairs fell into this cell. However, we can see that in WSsim-2, annotators seem to make a clear distinction whether a sense applies to some extent (scores 2-5) or does not apply at all (score 1). Based on this analysis, we propose a new way of judging a system's performance from an application point of view in section 5.2.

	1	**2**	**3**	**4**	**5**
5	0.912 *16,914*	0.860 *2,034*	0.754 *1,480*	0.605 *1,084*	0.337 *838*
4	0.684 *6,984*	0.567 *1,283*	0.498 *963*	0.320 *394*	
3	0.482 *11,024*	0.347 *1,868*	0.435 *800*		
2	0.179 *14,433*	0.733 *1,684*			
1	0.885 *100,217*				

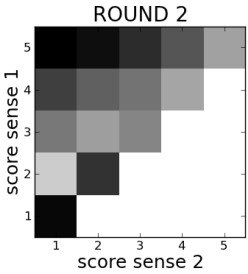

Figure 2: Analysis of the percentage of concordant pairs for sense pairs given particular scores in the data set for GWSA by (Erk et al., 2012). Normalized concordance matrix, summed over all annotators. The total counts of (concordant and discordant) pairs per cell, summed over all annotators, are reported in italic.

5.2 Accuracy-based Analysis using Graded Annotations

As we have seen above, human annotators use the scores 2-5 as an indicator that the sense is present at least to some extent in WSsim-2. Using fine-grained sense inventories such as WordNet, it is very hard even for humans to argue which of the senses is more present in a particular context. From a practical point of view, it may be sufficient if the sense to which a system assigned the highest score is present at least to some extent according to the humans' annotations. We propose to evaluate systems – in addition to the correlation analysis – according to this

criterion. This analysis allows to treat the GWSA data set as the gold standard for an evaluation similar to the coarse-grained WSD task of SemEval-2007, with the difference that SemEval-2007 uses predefined sense clusters, while in the GWSA data set, clusters are formed per context. Erk et al. (2009) show that it is not possible to form clusters out of the GWSA annotations that are applicable across the instances. Hence, we believe that the GWSA data sets are a valuable resource for evaluating WSD system performance in a coarse-grained but context-sensitive way.

For each treshold from 2 to 5 (in steps of 0.5), we create a gold standard in which all senses that received an average score $\geq$ the threshold are counted as correct, and then we evaluate accuracy as the percentage of lemma-sentence tasks in which the sense scored highest by a system is in this set of correct senses. For lower thresholds, the probability of picking a correct sense is higher as the set of correct senses is larger. Hence, we adjust our measure of accuracy inspired by Cohen's κ (Cohen, 1960). For each threshold t and for each lemma-sentence task i, we partition the set of graded senses S_i into two sets $S_{i,score \geq t}$ and $S_{i,score<t}$. Then, the probability of choosing a correct sense by chance for this task becomes

$$P^{t,i}_{chance} = \frac{|S_{i,score \geq t}|}{|S_i|}.$$

The average chance of picking a correct sense at threshold t is

$$P^{t}_{chance} = \frac{\sum_{i=1}^{N^t} P^{t,i}_{chance}}{N^t}.$$

where N^t is the total number of lemma-sentence tasks at threshold t in which $P^{t,i}_{chance} > 0$. We exclude the cases where the set of true positives is empty because the system cannot possibly pick a "correct" sense. The accuracy of a system at threshold t is computed as

$$Acc^t = \frac{\sum_{i=1}^{N^t} 1_{s^i \in S_{i,score>=t}}}{N^t}$$

with $\mathbf{1}$ being the indicator function and s^i being the sense that was scored highest by the system for the lemma-sentence task i. We then compute the Adjusted Accuracy at threshold t, which is plotted in Figure 3 , as

$$AdjAcc^t = \frac{Acc^t - P^t_{chance}}{1 - P^t_{chance}}.$$

The threshold-accuracy plots show how much better than chance a system is at predicting a sense that has a score above a certain threshold. As an upper bound, for each annotator, we regard the average of the other annotators as the gold standard and compute the Adjusted Accuracies, which range from 62% (for $t = 4.5$) to 76% (for $t = 2$). For $t = 5$, humans achieve a remarkable average Adjusted Accuracy of 73%.

5.3 Discussion

Referring to Figure 3, it is interesting to note that the shape of the curve for PageRank is much lower for all $t < 5$ than the other systems' curves. It shows a sharp increase when setting $t = 5$, which suggests that unlike the other systems, the graph-based PageRank algorithm is better suited for the standard WSD task of picking one best-fitting sense[2], while its ranking ability is not as good past the top rank as the other systems'. These findings are also supported by the observation that PageRank outperforms our reimplementation of the Topic Models approach on

[2]Sinha and Mihalcea report an F1-measure of 52.55% on the fine-grained WSD task of Senseval-2, and our PageRank system achieves an F1-measure of 76.0% on the coarse-grained WSD task of SemEval-2007.

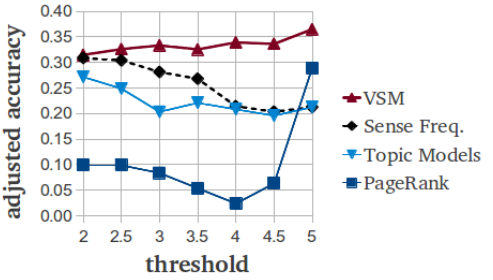

Figure 3: Adjusted Accuracy of picking a "correct" sense as the highest-ranked sense at various thresholds. Computed using the WSsim-2 data set. The numbers of sentences for which the set of "correct" senses $S_{i,score \geq t}$ is non-empty at the respective thresholds are 259, 256, 252, 238, 204, 150, and 46.

the SemEval-2007 coarse-grained task (F-measure of PageRank: 76% vs. Topic Models: 71.7%). A possible reason for this behavior might be the interaction of all the senses of the target word in one graph in PageRank. In contrast, the Topic Models and the VSM methods score only one sense at a time. When comparing to PageRank, the Topic Models system correlates more closely with the human average judgments per lemma-sentence. The same holds for a comparison using our metric of Adjusted Accuracy. We would like to note that none of the algorithms were tuned specifically for the GWSA task, with the exception of setting M and N of PageRank.

The VSM approach outperforms the two other knowledge-based systems of our study in all metrics presented in this paper. The VSM method has been developed mainly for tasks involving fine-grained lexical distinctions and has shown excellent performance on other lexical semantic tasks as well. Our comparison suggests that the model is good at capturing subtle distinctions between senses. It is also worth noting that the VSM is the only system that does not rely on WordNet's glosses, which in some cases contain examples that may be misleading for a system looking for topical information.

6 Conclusion

We explored the applicability of three knowledge-based WSD systems to the task of graded word sense assignment. We found a positive rank correlation between each of the systems' outputs and the human annotators' judgments. However, the performance levels of the individual systems were quite different. The most successful model (Thater et al., 2011) does not reach the supervised approach, but outperforms a sense frequencies baseline on the WSsim-2 data set. In addition, we showed that systems that are good at standard WSD (like the PageRank-based system) are not necessarily strong on the GWSA ranking task. We conclude that the use of the GWSA data sets with correlation and accuracy analyses as presented in this paper sheds a different light on the performance of WSD systems, by providing an in-depth analysis of their ranking behavior instead of treating WSD as a standard classification problem.

Acknowledgment

We are grateful to Diana McCarthy and Katrin Erk for providing us with the data and for clarifying several questions. We want to thank Alexis Palmer and the anonymous reviewers for their insightful comments. Thanks also go to Moinuddin Mushirul Haque. This work was supported by the Cluster of Excellence "Multimodal Computing and Interaction", funded by the German Excellence Initiative.

References

Agirre, E. and Soroa, A. (2009). Personalizing PageRank for word sense disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, pages 33–41, Stroudsburg, PA, USA. Association for Computational Linguistics.

Banerjee, S. (2003). Extended gloss overlaps as a measure of semantic relatedness. In *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence*, pages 805–810.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022.

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37.

Erk, K. and McCarthy, D. (2009). Graded Word Sense Assignment. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, EMNLP 2009, 6-7 August 2009, Singapore, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 440–449.

Erk, K., McCarthy, D., and Gaylord, N. (2009). Investigations on Word Senses and Word Usages. In Su, K.-Y., Su, J., and Wiebe, J., editors, *ACL/AFNLP*, pages 10–18. The Association for Computer Linguistics.

Erk, K., McCarthy, D., and Gaylord, N. (2012). Measuring word meaning in context. *Computational Linguistics (to appear)*.

Erk, K. and Padó, S. (2008). A structured vector space model for word meaning in context. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 897–906, Honolulu, Hawaii. Association for Computational Linguistics.

Fellbaum, C., editor (1998). *WordNet: An Electronic Lexical Database*. The MIT Press, Cambridge, MA ; London.

Jurgens, D. (2012). An Evaluation of Graded Sense Disambiguation using Word Sense Induction. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 189–198, Montréal, Canada. Association for Computational Linguistics.

Li, L., Roth, B., and Sporleder, C. (2010). Topic Models for Word Sense Disambiguation and Token-Based Idiom Detection. In Hajic, J., Carberry, S., and Clark, S., editors, *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010), July 11-16, 2010, Uppsala, Sweden*, pages 1138–1147. The Association for Computational Linguistics.

McCarthy, D. (2009). Word sense disambiguation: An overview. *Language and Linguistics Compass*, 3(2):537–558.

Navigli, R. (2009). Word sense disambiguation: A survey. *ACM Comput. Surv.*, 41(2):10:1–10:69.

Navigli, R. and Lapata, M. (2010). An Experimental Study of Graph Connectivity for Unsupervised Word Sense Disambiguation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(4):678–692.

Pedersen, T., Patwardhan, S., and Michelizzi, J. (2004). Wordnet::Similarity – Measuring the Relatedness of Concepts. In *Proceedings of the Nineteenth National Conference on Artificial Intelligence (AAAI-04)*, pages 1024–1025.

Ponzetto, S. P. and Navigli, R. (2010). Knowledge-Rich Word Sense Disambiguation Rivaling Supervised Systems. In Hajic, J., Carberry, S., and Clark, S., editors, *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010), July 11-16, 2010, Uppsala, Sweden*, pages 1522–1531. The Association for Computer Linguistics.

Sinha, R. S. and Mihalcea, R. (2007). Unsupervised Graph-based Word Sense Disambiguation Using Measures of Word Semantic Similarity. In *ICSC*, pages 363–369. IEEE Computer Society.

Thater, S., Fürstenau, H., and Pinkal, M. (2010). Contextualizing semantic representations using syntactically enriched vector models. In Hajic, J., Carberry, S., and Clark, S., editors, *ACL*, pages 948–957. The Association for Computer Linguistics.

Thater, S., Fürstenau, H., and Pinkal, M. (2011). Word Meaning in Context: A Simple and Effective Vector Model. In *Proceedings of 5th International Joint Conference on Natural Language Processing (IJCNLP 2011)*.

Tsatsaronis, G., Varlamis, I., and Nørvåg, K. (2010). An Experimental Study on Unsupervised Graph-based Word Sense Disambiguation. In Gelbukh, A. F., editor, *CICLing*, volume 6008 of *Lecture Notes in Computer Science*, pages 184–198. Springer.