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

ABSTRACT Graded Word Sense Assignment (GWSA) data sets [Erk & McCarthy, 2012] provide judgments of applicability for all possible senses of a word in context. In this work, we compare the performance of three knowledge-based Word Sense Disambigation (WSD) algorithms on the task of ranking the senses according to their applicability. In addition, we develop a metric named Adjusted Accuracy which allows for a coarse-grained evaluation similar to the SemEval-2007 task, but in a context-sensitive way.

Graded Word Sense Assignment

Traditional **Word Sense Disambigation (WSD)**: each instance is assigned exactly one sense.
Graded Word Sense Assignment (GWSA): applicability of all possible senses is judged for each instance on a 1-5 scale.
Example: *This can be justified thermo-dynamically in this case, and this will be done in a separate **paper** which is being prepared.*

WordNet sense	Ratings			Avg
#1 A material made of cellulose pulp	5	1	1	2.3
#3 A daily or weekly publication on folded sheets, contains news and articles and advertisements	2	1	3	2
#5 A scholarly article describing the results of observations or stating hypotheses	5	5	5	5
#4 A medium of written communication	5	3	1	3

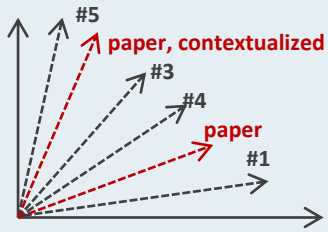
Data: WSsim-1 and WSsim-2 by Erk and McCarthy (2012)

Sense Ranking Evaluation

	WSsim-1		Wssim-2	
	ρ	% sign	ρ	% sign
Average Humans	0.555	30.4	0.641	48.3
Prototype 2/N [Erk&McCarthy]	0.478	22.8	-	-
Sense Frequencies	0.357	10.7	0.245	14.2
VSM [Thater et al.]	0.305	12.7	0.389	21.4
Topic Models [Li et al.]	0.241	11.6	0.256	15.0
PageRank [Sinha et al.]	0.210	4.0	0.097	4.6

Knowledge-based Models

Syntactic Vector SpaceModel (VSM) [Thater et al. 2011]



Sense paraphrase = words in synset + related words, syntactic vectors estimated from Gigaword.

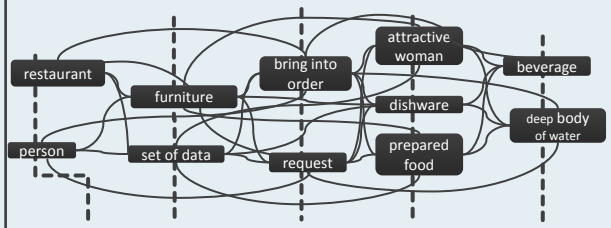
Score of sense for a context = average of cosine similarity of the two synset words most similar to context

Topic Models System [Li et al. 2010]

→ vectors → cosine sim.

Topic Distributions over words.
Context = sentence.
Sense Paraphrases = glosses, example sentences, related synsets.

Graph-based (PageRank & Ext. Lesk) [Sinha et al. 2007]



Why these three models?
(a) knowledge-lean, easy to implement
(b) state-of-the-art performance in SemEval-2007 coarse-grained WSD task

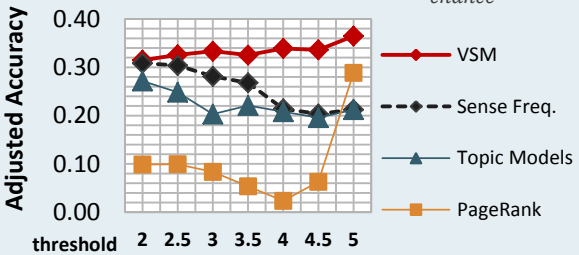
Accuracy-based Analysis

Analysis of concordance of pairwise sense rankings (across all annotators)

Sense 1 \ Sense 2	1	2	3	4	5
5	91.2	86.0	75.4	60.5	33.7
4	68.4	56.7	49.8	32.0	
3	48.2	34.7	43.5		
2	17.9	73.3			
1	88.5				

Adjusted Accuracy:
Does system choose a sense that is applicable at least to some extent?

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



Conclusion

→ Knowledge-based systems show ***positive correlations*** with human judgments of sense applicability, but doing well at WSD does not necessarily imply excellent performance at GWSA.
→ VSM performed best, followed by the Topic Models system. PageRank works well when picking one best-fitting sense, but performs worse when senses do only apply to some extent, as shown by the Adjusted Accuracy plot.
→ The SemEval-2007 coarse-grained WSD task uses pre-defined sense clusters. **Adjusted Accuracy** can be regarded as a ***context-specific*** coarse-grained evaluation.