Graded Word Sense Assignment

Hu Jingwen

Overview

- Introduction
- Data
- Models for Graded Word Sense Assignment *(WSD/ single, WSD/conf., Prototype)*
- Evaluating Graded Word Sense Assignment *(correlation, divergence, precision and recall)*
- Experiments and Discussion
- Conclusion

Introduction

- Word sense disambiguation (WSD): label a word with best-fitting sense
- Problem: hard to define clear cut sense boundary and word sense strongly depends on the context
- Previous work: create a coarse-grained invertory of word sense for humans and computers by automatically relating or clustering existing word sense

Introduction

In this paper:

- Mark target word in context with graded ratings (scale of 1-5) on senses
- Propose evaluation metrics which is used on graded word sense judgments (GWS)
- Investigate two classes of models (classical WSD models and prototype-based vector space models) perform on GWS
- Provides a novel analysis of annotator performance on the GWS dataset

Data

- Use a subset of the GWS dataset (Erk et al., 2009)
- Three annotators supply ordinal judgments of the applicability of WordNet (v3.0) senses on a 5 point scale (1–completely different, 2– mostly different, 3 – similar, 4 – very similar and 5–identical)
- Random 8 lemmas and 50 sentences per lemma applied

Data

Sentence	Senses 1	2	3	4	5	6	7	Annotator
This can be justified thermodynamically in this case, and	2	3	3	5	5	2	3	Ann. 1
this will be done in a separate **paper** which is being	1	3	1	3	5	1	1	Ann. 2
prepared.	1	5	2	1	5	1	1	Ann. 3
	1.3	3.7	2	3	5	1.3	1.7	Avg

Table 1: A sample annotation in the GWS experiment. The senses are: 1 material from cellulose 2 report 3 publication 4 medium for writing 5 scientific 6 publishing firm 7 physical object

Data

lemma (PoS)	# senses	# training SemCor	SE-3
add (v)	6	171	238
argument (n)	7	14	195
ask (v)	7	386	236
different (a)	5	106	73
important (a)	5	125	11
interest (n)	7	111	160
paper (n)	7	46	207
win (v)	4	88	53
total training sentences		1047	1173

Table 2: Lemmas used in this study

with various sense-tagged datasets (e.g. (Miller et al., 1993; Mihalcea et al., 2004)) for comparison.

Data

- Baseline model: compared by other models as reference, which simply uses training data to obtain probability distribution over sense
- Results: more use of intermediate levels of applicability (2-4) and more positive ratings (3-5)
- Normalization of judgments:

$$normalized\ judgment = (judgment - 1.0)/4.0$$

Models for Graded Word Sense Assignment

Single-best-sense WSD (*WSD/single*):

- Assign a target word in each test occurrence a single best-fitting word sense
- Feedback confidence score 1 to assigned sense and score 0 to all other senses
- Not highly optimized, but fairly standard

Models for Graded Word Sense Assignment

WSD confidence level as judgment *(WSD/conf.)* :

- The same WSD system as *WSD/single*, but predict a judgment for each sense of a target occurrence with confidence level

Models for Graded Word Sense Assignment

Word senses as points in semantic space *(prototype)* :

- Represent word senses as points in a semantic space
- Model graded sense applicability judgments by vector similarity
- Dimensions of vector space-features of WSD system

  Coordinates-raw feature counts
- Measure the distance between feature vectors in space
- *Prototype*- first order feature vector, *Prototype/2*- second order feature vector

Models for Graded Word Sense Assignment

Word senses as points in semantic space *(prototype)* :

- For spurious negative data, prototype models forget competition and ignore that. WSD models fully trust the negative data.

Evaluating Graded Word Sense Assignment

Correlation between gold and predicted judgments:

- Tested by Spearman`s ρ, which uses the formula of Pearson`s coefficient

$$\rho(X,Y) = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

- Measure the abstracts from absolute values of judgments and magnitude of difference
- Judgment represented as 4-tuple < lemma, sense_no, sentence_no, judgment >

Evaluating Graded Word Sense Assignment

Correlation between gold and predicted judgments:

by lemma: for each lemma $\ell \in L$, compute correlation between $G|_{lemma=\ell}$ and $A|_{lemma=\ell}$

by lemma+sense: for each lemma ℓ and each sense number $i \in S_\ell$, compute correlation between $G|_{lemma=\ell,senseno=i}$ and $A|_{lemma=\ell,senseno=i}$

by lemma+sentence: for each lemma ℓ and sentence number $t \in T$, compute correlation between $G|_{lemma=\ell,sentence=t}$ and $A|_{lemma=\ell,sentence=t}$

Evaluating Graded Word Sense Assignment

Correlation between gold and predicted judgments:

- By lemma: test the consistent use of judgments for the same target lemma
- By lemma+sense: test how strongly a given word sense evoked at the same target lemma
- By lemma+sentence: test how strongly a given target lemma occurrence applied for different senses

Evaluating Graded Word Sense Assignment

Divergence:

- Measure the similarity between two probability distributions by Jensen/Shannon (J/S) divergence

$$JS(p,q) = \frac{1}{2}\left(D(p||\frac{p+q}{2}) + D(q||\frac{p+q}{2}\right)$$

- Measure the abstracts from absolute values of judgments NOT from magnitude of difference

Evaluating Graded Word Sense Assignment

Precision and Recall:

- Evaluate the correctness of a pattern recognition algorithm
- Test what degree a model conforms to the absolute judgments given by human annotators
- Precision defined as

$$P_\ell = \frac{\sum_{i \in S_\ell, t \in T} \min(\text{gold}_{\ell,i,t}, \text{assigned}_{\ell,i,t})}{\sum_{i \in S_\ell, t \in T} \text{assigned}_{\ell,i,t}}$$

Evaluating Graded Word Sense Assignment

Precision and Recall:

- Recall defined as

$$R_\ell = \frac{\sum_{i \in S_\ell, t \in T} \min(\text{gold}_{\ell,i,t}, \text{assigned}_{\ell,i,t})}{\sum_{i \in S_\ell, t \in T} \text{gold}_{\ell,i,t}}$$

- F-score: as the harmonic mean of precision and recall

$$F = 2PR/(P + R)$$

Experiments and Discussion

	by lemma			by lemma+sense			by lemma+sentence						
Ann	ρ	*	**	ρ	*	**	ρ	*	**	J/S	P	R	F
Ann.1	0.517	100.0	100.0	0.407	75.0	58.3	0.482	27.3	11.5	0.131	50.6	87.5	64.1
Ann.2	0.587	100.0	100.0	0.403	68.8	58.3	0.612	38.1	17.2	0.153	75.5	62.4	68.3
Ann.3	0.528	100.0	100.0	0.41	77.1	58.3	0.51	21.8	7.8	0.165	82.4	52.3	64.0
Avg	0.544	100.0	100.0	0.407	73.6	58.3	0.535	29.0	12.2	0.149	69.5	67.4	65.5

Table 4: Human ceiling: one annotator vs. average of the other two annotators. *, **: percentage significant at $p \leq 0.05$, $p \leq 0.01$. Avg: average annotator performance

Experiments and Discussion

For Table 4 (shows human performance)

- Annotator 1 different from other two, tends to assign higher ratings, has lower J/S and higher Recall than Precision
- Quite significant correlation for all lemmas, less significant correlation for lemma+sense or lemma+sentence. Smaller p value with worse correlation ($p \leq 0.05 vs. p \leq 0.01$)

Experiments and Discussion

Model	by lemma			by lemma+sense			by lemma+sentence			J/S	P	R	F
	ρ	*	**	ρ	*	**	ρ	*	**				
WSD/single	0.267	87.5	75.0	0.053	6.3	4.2	0.28	2.8	1.8	0.39	58.7	25.5	35.5
WSD/conf	0.396	87.5	87.5	0.177	33.3	18.8	0.401	10.8	3.0	0.164	81.8	37.1	51.0
Prototype	0.245	62.5	62.5	0.053	20.8	8.3	0.396	15.3	2.5	0.173	58.4	78.3	66.9
Prototype/2	0.292	87.5	87.5	0.086	14.6	4.2	0.478	22.8	7.5	0.164	68.2	63.3	65.7
Prototype/N	0.396	100.0	100.0	0.137	22.9	14.6	0.396	15.3	2.5	0.173	82.2	29.9	43.9
Prototype/2N	0.465	100.0	100.0	0.168	29.8	23.4	0.478	22.8	7.5	0.164	82.6	30.9	45.0
baseline	0.338	87.5	87.5	0.0	0.0	0.0	0.355	10.3	3.0	0.167	79.9	34.5	48.2

Table 5: Evaluation: computational models, and baseline. *, **: percentage significant at $p \leq 0.05$, $p \leq 0.01$

Experiments and Discussion

For Table 5 (shows the performance of different models)

- *WSD/single* below baseline
- *WSD/conf* slightly above baseline
- *Prototype* deviate strongly from baseline, have good recall
- *WSD/conf* and *Prototype/2N* show best performance of J/S
- *Prototype/N* and *Prototype/2N* achieve very high correlation by lemma

Experiments and Discussion

Human performance:

- Provide a novel analysis of the GWS dataset
- Show very strong correlation of rankings by lemma and also by lemma+sense
- Low precision and recall indicate that different annotators use the 5-point scale in different ways
- Average judgments NOT influenced by individual annotator

Experiments and Discussion

Evaluation measures:

- Rankings by lemma+sense and by lemma+sentence have potential use in systems
- Measurements of graded precision and recall require a more fine-grained analysis of the performance of models

Experiments and Discussion

Standard WSD models vs. Vector space models:

- *WSD/conf* has the highest correlation by lemma and high precision
- *Prototype* , which tends to assign high ratings, has much better recall and higher F-score
- Compared to *Prototype*, *Prototype/2* (with using second order vectors) is much less sparse and yields better rankings

Experiments and Discussion

Standard WSD models vs. Vector space models:

- *WSD/conf* performs slightly above baseline because of a very familiar picture from standard WSD
- *Prototype/2N* shows best correlation, which pays minimal attention to the negative data, uses normalization and second order vectors
- Low ratings in correlation by lemma+sense indicates models might be limited by the lack of training data for many of the rarer senses

Conclusion

- Study on graded annotator judgments on sense applicability
- Evaluate GWS by correlation, devergence and new extensions (precision and recall)
- Test two types of models: WSD and prototype models
- Investigate features, which is more informative for making graded judgments, in the future

Thank you for your attention!

References

- M. Carpuat and D. Wu. 2007. Improving statistical machine translation using word sense disambiguation. In Proceedings of EMNLP-CoNLL 2007, pages 61–72, Prague, Czech Republic, June. Association for Computational Linguistics.

- Y. S. Chan and H. T. Ng. 2005. Word sense disambiguation with distribution estimation. In Proceedings of IJCAI 2005, pages 1010–1015, Edinburgh,Scotland.

- K. Erk, D. McCarthy, and N. Gaylord. 2009. Investigations on word senses and word usages. In Proceedings of ACL-09, Singapore.

- M. Lapata and C. Brew. 2004. Verb class disambiguation using informative priors. Computational Linguistics, 30(1):45–75.