Computational Semantics and Pragmatics

Graded Word Sense Assignment
Katrin Erk and Diana McCarthy

Article given by Raquel Fernández Rovira

Cecilia Chávez Aguilera

University of Amsterdam

December 12, 2012

Agenda

Agenda

- Inter-annotator agreement
  - Fine-grained word senses
    - 69 % HECTOR Dictionary (Krishnamurthy and Nicholls 2000)
    - 78.6% WordNet (Landes et al. 1998)
  - Coarse-grained word senses
    - 90% OntoNotes (Hovy et al. 2006)
- Graded annotation
- Aim: Predict graded judgments of word sense applicability.

- Inter-annotator agreement
  - Fine-grained word senses
    - 69 % HECTOR Dictionary (Krishnamurthy and Nicholls 2000)
    - 78.6% WordNet (Landes et al. 1998)
  - Coarse-grained word senses
    - 90% OntoNotes (Hovy et al. 2006)
- Graded annotation
- Aim: Predict graded judgments of word sense applicability.

- Inter-annotator agreement
  - Fine-grained word senses
    - 69 % HECTOR Dictionary (Krishnamurthy and Nicholls 2000)
    - 78.6% WordNet (Landes et al. 1998)
  - Coarse-grained word senses
    - 90% OntoNotes (Hovy et al. 2006)
- Graded annotation
- Aim: Predict graded judgments of word sense applicability.

- Inter-annotator agreement
  - Fine-grained word senses
    - 69 % HECTOR Dictionary (Krishnamurthy and Nicholls 2000)
    - 78.6% WordNet (Landes et al. 1998)
  - Coarse-grained word senses
    - 90% OntoNotes (Hovy et al. 2006)
- Graded annotation
- Aim: Predict graded judgments of word sense applicability.

- Inter-annotator agreement
  - Fine-grained word senses
    - 69 % HECTOR Dictionary (Krishnamurthy and Nicholls 2000)
    - 78.6% WordNet (Landes et al. 1998)
  - Coarse-grained word senses
    - 90% OntoNotes (Hovy et al. 2006)
- Graded annotation
- Aim: Predict graded judgments of word sense applicability.

- Inter-annotator agreement
  - Fine-grained word senses
    - 69 % HECTOR Dictionary (Krishnamurthy and Nicholls 2000)
    - 78.6% WordNet (Landes et al. 1998)
  - Coarse-grained word senses
    - 90% OntoNotes (Hovy et al. 2006)
- Graded annotation
- Aim: Predict graded judgments of word sense applicability.

- Inter-annotator agreement
  - Fine-grained word senses
    - 69 % HECTOR Dictionary (Krishnamurthy and Nicholls 2000)
    - 78.6% WordNet (Landes et al. 1998)
  - Coarse-grained word senses
    - 90% OntoNotes (Hovy et al. 2006)
- Graded annotation
- Aim: Predict graded judgments of word sense applicability.

- Inter-annotator agreement
  - Fine-grained word senses
    - 69 % HECTOR Dictionary (Krishnamurthy and Nicholls 2000)
    - 78.6% WordNet (Landes et al. 1998)
  - Coarse-grained word senses
    - 90% OntoNotes (Hovy et al. 2006)
- Graded annotation
- Aim: Predict graded judgments of word sense applicability.

Agenda

Graded Word Sense dataset

lemma (PoS)	# senses	# training SemCor	# training 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 : Lemmas used in this study

- The scale used 1:= completely different, 2:= mostly different, 3:= similar, 4:= very similar, 5:= identical.
- It was obtained a single judgment for each sense with a normalized average of the three annotators, with the following normalization:

$$normalized - judgment = \frac{judgment - 1.0}{4.0}$$

- Judgments in the gold standar and assigned judgments can be represented by tuples: $\langle lemma, sense - no., sentence - no., value \rangle$

- The scale used 1:= completely different, 2:= mostly different, 3:= similar, 4:= very similar, 5:= identical.
- It was obtained a single judgment for each sense with a normalized average of the three annotators, with the following normalization:

$$normalized - judgment = \frac{judgment - 1.0}{4.0}$$

- Judgments in the gold standar and assigned judgments can be represented by tuples: $\langle lemma, sense - no., sentence - no., value \rangle$

- The scale used 1:= completely different, 2:= mostly different, 3:= similar, 4:= very similar, 5:= identical.
- It was obtained a single judgment for each sense with a normalized average of the three annotators, with the following normalization:

$$normalized - judgment = \frac{judgment - 1.0}{4.0}$$

- Judgments in the gold standar and assigned judgments can be represented by tuples: $\langle lemma, sense - no., sentence - no., value \rangle$

Example

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

Table : 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

Agenda

Correlation

- Let G be the set of golden tuples, and A the set of assigned tuples; L be the set of lemmas, and S_l the set of sense numbers for lemma l, and T, the set of sentence numbers:
- **lemma** $G_{lemma=l}$ and $A_{lemma=l}$ $\forall l \in L$
- **lemma + sense** $G_{lemma=l,senseno.=i}$ and $A_{lemma=l,senseno.=i}$ $\forall l \in L,\ i \in S_l$
- **lemma + sentence** $G_{lemma=l,sentence=t}$ and $A_{lemma=l,sentence=t}$ $\forall l \in L,\ t \in T$

Correlation

- Let G be the set of golden tuples, and A the set of assigned tuples; L be the set of lemmas, and S_l the set of sense numbers for lemma l, and T, the set of sentence numbers:
- **lemma** $G_{lemma=l}$ and $A_{lemma=l}$ $\forall l \in L$
- **lemma + sense** $G_{lemma=l,senseno.=i}$ and $A_{lemma=l,senseno.=i}$ $\forall l \in L,\ i \in S_l$
- **lemma + sentence** $G_{lemma=l,sentence=t}$ and $A_{lemma=l,sentence=t}$ $\forall l \in L,\ t \in T$

Correlation

- Let G be the set of golden tuples, and A the set of assigned tuples; L be the set of lemmas, and S_l the set of sense numbers for lemma l, and T, the set of sentence numbers:
- **lemma** $G_{lemma=l}$ and $A_{lemma=l}$ $\forall l \in L$
- **lemma + sense** $G_{lemma=l,senseno.=i}$ and $A_{lemma=l,senseno.=i}$ $\forall l \in L,\ i \in S_l$
- **lemma + sentence** $G_{lemma=l,sentence=t}$ and $A_{lemma=l,sentence=t}$ $\forall l \in L,\ t \in T$

Correlation

- Let G be the set of golden tuples, and A the set of assigned tuples; L be the set of lemmas, and S_l the set of sense numbers for lemma l, and T, the set of sentence numbers:
- **lemma** $G_{lemma=l}$ and $A_{lemma=l}$ $\forall l \in L$
- **lemma + sense** $G_{lemma=l,senseno.=i}$ and $A_{lemma=l,senseno.=i}$ $\forall l \in L,\ i \in S_l$
- **lemma + sentence** $G_{lemma=l,sentence=t}$ and $A_{lemma=l,sentence=t}$ $\forall l \in L,\ t \in T$

Spearman's ρ

- Uses the Pearson's coefficient:

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

  computed over rankings
- The rankings are assigned by sorting in ascending order the value of the variables. Equal values get the average of their positions
- Significance of the values is found against a probability p of the observed extreme cases

Spearman's ρ

- Uses the Pearson's coefficient:

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

  computed over rankings
- The rankings are assigned by sorting in ascending order the value of the variables. Equal values get the average of their positions
- Significance of the values is found against a probability p of the observed extreme cases

Spearman's ρ

- Uses the Pearson's coefficient:

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

  computed over rankings

- The rankings are assigned by sorting in ascending order the value of the variables. Equal values get the average of their positions
- Significance of the values is found against a probability p of the observed extreme cases

Agenda

Prototype

Cx/2	until, IN, soft, JJ, remaining, VBG, ingredient, NNS
Cx/50	for, IN, sweet-sour, NN, sauce, NN, . . . , to, TO, a, DT, boil, NN
Ch	OA, OA/ingredient/NNS

Table : Sample features for add in BNC occurrence For sweet-sour sauce, cook onion in oil until soft. **Add** remaining ingredients and bring to a boil. Cx/2 (Cx/50): context of size 2 (size 50) either side of the target. Ch: children of target.

Prototype

- Dimensions: Features, Coordinates: Raw counts
- Vector representation for a sense: centroid of its training occurrences
- Predicted judgment for sentence t, and sense s: similarity of its vectors.
- Like instance based-learners measures the distance between feature vectors but within a single category

Prototype

- Dimensions: Features, Coordinates: Raw counts
- Vector representation for a sense: centroid of its training occurrences
- Predicted judgment for sentence *t*, and sense *s*: similarity of its vectors.
- Like instance based-learners measures the distance between feature vectors but within a single category

Prototype

- Dimensions: Features, Coordinates: Raw counts
- Vector representation for a sense: centroid of its training occurrences
- Predicted judgment for sentence t, and sense s: similarity of its vectors.
- Like instance based-learners measures the distance between feature vectors but within a single category

Prototype

- Dimensions: Features, Coordinates: Raw counts
- Vector representation for a sense: centroid of its training occurrences
- Predicted judgment for sentence t, and sense s: similarity of its vectors.
- Like instance based-learners measures the distance between feature vectors but within a single category

Agenda

Model	by lemma ρ	by lemma $*$	by lemma $**$	by lemma+sense ρ	by lemma+sense $*$	by lemma+sense $**$	by lemma+sentence ρ	by lemma+sentence $*$	by lemma+sentence $**$
WSD/single	0.267	87.5	75.0	0.053	6.3	4.2	0.28	2.8	1.8
WSD/conf	0.396	87.5	87.5	0.177	33.3	18.8	0.401	10.8	3.0
Prototype	0.245	62.5	62.5	0.053	20.8	8.3	0.396	15.3	2.5
Prototype/2	0.292	87.5	87.5	0.086	14.6	4.2	0.478	22.8	7.5
Prototype/N	0.396	100.0	100.0	0.137	22.9	14.6	0.396	15.3	2.5
Prototype/2N	0.465	100.0	100.0	0.168	29.8	23.4	0.478	22.8	7.5
baseline	0.338	87.5	87.5	0.0	0.0	0.0	0.355	10.3	3.0

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

Summary

- Graded annotation was proposed as an alternative view of sense assignment.
- Adequate measures to evaluate performance of graded sense assignment were proposed.
- Evaluation is significant, but system performance is below humans performance
- The authors have already worked a second round of annotation
- The lemma + sense and lemma + sentence correlation measures seem to be the most promising useful measures.
- The GWS should be tested with more sophisticated vector models.

Summary

- Graded annotation was proposed as an alternative view of sense assignment.
- Adequate measures to evaluate performance of graded sense assignment were proposed.
- Evaluation is significant, but system performance is below humans performance
- The authors have already worked a second round of annotation
- The lemma + sense and lemma + sentence correlation measures seem to be the most promising useful measures.
- The GWS should be tested with more sophisticated vector models.

Summary

- Graded annotation was proposed as an alternative view of sense assignment.
- Adequate measures to evaluate performance of graded sense assignment were proposed.
- Evaluation is significant, but system performance is below humans performance
- The authors have already worked a second round of annotation
- The lemma + sense and lemma + sentence correlation measures seem to be the most promising useful measures.
- The GWS should be tested with more sophisticated vector models.

Summary

- Graded annotation was proposed as an alternative view of sense assignment.
- Adequate measures to evaluate performance of graded sense assignment were proposed.
- Evaluation is significant, but system performance is below humans performance
- The authors have already worked a second round of annotation
- The lemma + sense and lemma + sentence correlation measures seem to be the most promising useful measures.
- The GWS should be tested with more sophisticated vector models.

Summary

- Graded annotation was proposed as an alternative view of sense assignment.
- Adequate measures to evaluate performance of graded sense assignment were proposed.
- Evaluation is significant, but system performance is below humans performance
- The authors have already worked a second round of annotation
- The lemma + sense and lemma + sentence correlation measures seem to be the most promising useful measures.
- The GWS should be tested with more sophisticated vector models.

Summary

- Graded annotation was proposed as an alternative view of sense assignment.
- Adequate measures to evaluate performance of graded sense assignment were proposed.
- Evaluation is significant, but system performance is below humans performance
- The authors have already worked a second round of annotation
- The lemma + sense and lemma + sentence correlation measures seem to be the most promising useful measures.
- The GWS should be tested with more sophisticated vector models.