Word Sense Disambiguation

Predominant Sense Aquisition

Lisa Beinborn

Seminar: Language Processing for different domains and genres

Outline

- Domain Specific Word Sense Disambiguation
  - Idea
- Automatic Method
  - Mc Carthy et al. 2004
- Evaluation

Problem

- Words can have different senses
- Star

Celestial body

Shape

Celebrity

Base solutions

1) Use supervised machine learning with SemCor
   - ♦ SemCor = subset of Brown Corpus
   - ♦ Open-class words are sense-tagged
2) Take most frequent sense
   - ♦ Skewed sense distributions

→Problem: not enough data

Ideas

- One sense prevails in a given discourse
- Most frequent sense often depends on domain

- No domain-specific sense-tagged corpora available

→ Automatically induce predominant sense

Automatic Method [McCarthy et al 2004]

- Get senses s_i for word w from sense inventory

Automatic Method [McCarthy et al 2004]

- Get senses s_i for word w from WordNet
- Rank them
  - ♦ depends on training corpus

Distributional Similarity

- Consider k nearest neighbours
  - Words that appear in the same context
  - *The star revealed…*
  - *The actor revealed…*
- Build thesaurus with k = 50
- "nearest" ≈ distributional similarity score (dss)

Contribution of neighbours

- Different neighbours share different senses with word
  - ♦ actor → celebrity
  - ♦ planet → celestial body
  - ♦ circle → shape
- How can these relations be inferred?

Semantic Similarity

- sss‘ = semantic similarity score
  - Closeness of two senses
- For each neighbour n
  - Get senses s_x
  - Calculate sss‘(s_i, s_x)
  - sss(s_i, n) = max sss‘

Neighbours: {actor, planet, …}

s_x(actor): {role player, worker…}

sss‘(celebrity, role player)= 0.7

Semantic Similarity

- sss‘ = semantic similarity score
  - Closeness of two senses
- For each neighbour n
  - Get senses s_x
  - Calculate sss‘(s_i, s_x)
  - sss(s_i, n) = max sss‘

Neighbours: {actor, planet, …}

s_x(actor): {role player, worker…}

sss‘(celebrity, worker)= 0.5

sss(celebrity, actor)= 0.7

Prevalence Score

$$PrevalenceScore(w, s_i) = \sum_{n_j} dss(w, n_j) * \frac{sss(s_i, n_j)}{normalization}$$

Ranks the senses of word w

50 nearest neighbours

Scored neighbours

Weighted: normalized semantic similarity of sense and neighbour

Prevalence Score

$$Prevalence\,Score(w, s_i) = \sum_{n_j} dss(w, n_j) * \frac{sss(s_i, n_j)}{normalization}$$

Contribution of all neighbours to sense s_i

Contribution of neighbour n_j to sense s_i

Evaluation

- Sense rankings for a sample of nouns
- Corpora
  - BNC
  - Finance
  - Sports

Word Selection F&S

- Only polysemous nouns
- At least one synset (WN) labeled with sports
- At least one synset labeled with economics
- Examples:
  - F&S (17): *manager, record, score, check, return, competition, club, …*
- Manual sense annotation

Sense Distribution

Word	PS BNC	PS FINANCE	PS SPORTS
pass	1 (**accomplishment**)	14 (**attempt**)	15 (**throw**)
share	2 (**portion, asset**)	2	2
division	4 (**admin. unit**)	4	6 (**league**)
head	1 (**body part**)	4 (**leader**)	4
loss	2 (**transf. property**)	2	8 (**death, departure**)
competition	2 (**contest, social event**)	3 (**rivalry**)	2
match	2 (**contest**)	7 (**equal, person**)	2
tie	1 (**neckwear**)	2 (**affiliation**)	3 (**draw**)
strike	1 (**work stoppage**)	1	6 (**hit, success**)
goal	1 (**end, mental object**)	1	2 (**score**)

Additional sets

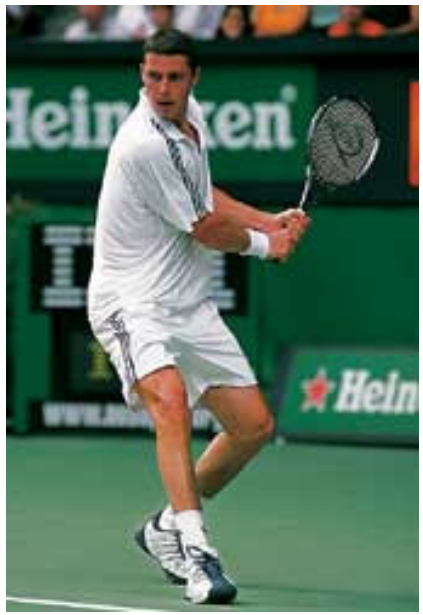
- Selected based on salience
  - most salient words in domain
  - Salience computed by frequency
- Sets
  - S sal (8): *fan, star, transfer, striker, goal, title,…*
  - F sal (8): *package, chip, bank, market, strike,…*
  - eq sal (7): *will, phase, half, top, performance,…*

Sense Distribution

- Even in domain-specific corpora, ambiguity is still present, though it is less than for general text
- The domain specific sense is not always the predominant sense in a domain-specific corpus
  - ♦ but more frequent than in general corpus

Example

- Return = a tennis stroke
  - Not the most frequent sense in SPORTS
  - Frequency = 19
  - Absent in FINANCE and BNC



Results

Training	Testing		
	BNC	FINANCE	SPORTS
BNC	**40.7**	43.3	33.2
FINANCE	39.1	**49.9**	24.0
SPORTS	25.7	19.7	**43.7**
Random BL	19.8	19.6	19.4
SemCor FS	32.0 (32.9)	33.9 (35.0)	16.3 (16.8)

- When applied to corresponding domain, *McCarthy et al. 2004* method beats random baseline and SemCor FS in all cases

Results

Test - Train	F&S cds	F sal	S sal	eq sal
BNC-APPR	33.3	51.5	39.7	48.0
BNC-SC	28.3	44.0	24.6	36.2
FINANCE-APPR	37.0	70.2	38.5	70.1
FINANCE-SC	30.3	51.1	22.9	33.5
SPORTS-APPR	42.6	18.1	65.7	46.9
SPORTS-SC	9.4	38.1	13.2	12.2

- APPR = training on appropriate domain
- SC = SemCor

Results

Test - Train	F&S cds	F sal	S sal	eq sal
BNC-APPR	33.3	51.5	39.7	48.0
BNC-SC	28.3	44.0	24.6	36.2
FINANCE-APPR	37.0	70.2	38.5	70.1
FINANCE-SC	30.3	51.1	22.9	33.5
SPORTS-APPR	42.6	18.1	65.7	46.9
SPORTS-SC	9.4	38.1	13.2	12.2

- Training on appropriate domain makes sense for all words
- Assumption: salient words benefit more

Conclusions

- Automatic acquisition of predominant senses from domain-specific corpora outperforms the automatic acquisition from SemCor for the sample words
- But: still an approximation, lots of problematic cases
- Better: Use local context for disambiguation

Conclusions

- Automatic method is cheaper

→ Use method if there is no manually tagged data available or if the data seems to be inappropriate for the word and domain

Questions?

Thank you!

References

- Diana McCarthy, Rob Koeling,Julie Weeds and John Carroll, 2004. Finding PredominantWord Senses in Untagged Text, *Proceedings of ACL-04*, Barcelona, Spain.

- Rob Koeling, Diana McCarthy and John Carroll, 2005. Domain-Specific Sense Distributions and Predominant Sense Acquisition, *EMNLP 2005*

- Diana McCarthy, Rob Koeling, Julie Weeds and John Carroll, 2007. Unsupervised Acquisition of Predominant Word Senses, *Computational Linguistics 33(4),* pp. 553-590.

References

- Distributional Similarity

  Julie Weeds, 2003. Measures and Applications of Lexical Distributional Similarity. Ph.D. thesis, Department of Informatics, University of Sussex, Brighton, UK.

- Semantic Similarity

  Siddharth Patwardhan, Satanjeev Banerjee and Ted Pedersen. 2003. Using measures of semantic relatedness for word sense disambiguation. In *Proceedings of CICLing 2003*, pp. 241–257, Mexico City, Mexico.