CMPT-882: Statistical Learning of Natural Language

Lecture #2

Anoop Sarkar
anoop@cs.sfu.ca!
http://www.sfu.ca/~anoop

- Unsupervised Word Sense Disambiguation Rivaling Supervised Methods (1995). David Yarowsky. Proceedings of ACL-95. pp. 189-196

Wall Street Journal: Penn Treebank

But other than the fact that besuboru is played
with a ball and a bat , it ’s unrecognizable :
Fans politely return foul balls to stadium
ushers ; the strike zone expands depending
on the size of the hitter ; ties are permitted --
even welcomed -- since they honorably
sidestep the shame of defeat ; players must
abide by strict rules of conduct even in their
personal lives -- players for the Tokyo Giants ,
for example , must always wear ties when
on the road .

Two Constraints

- One sense per collocation
  star $\rightarrow$ celebrity ($f = 1422$) / celestial ($f = 222$) /
  shape of object ($f = 56$)
  Accuracy = 96% (celebrity = 96%; celestial = 95%; shape = 82%)

- One sense per discourse
  plant $\rightarrow$ living / factory
  Accuracy = 99.8%, Applicability = 72.8%

The Algorithm

- Input: corpus, seed collocation rules

- Steps:

  1. Apply seed rules to corpus
  2. Train supervised decision list learner on partially labelled corpus
  3. Apply decision list learner on entire corpus
  4. Use one-sense-per-discourse to obtain new labelled data
  5. If no new classifications are found, Stop. Else go back to Step **??**

Points for Discussion

- Escaping from initial misclassification
- What exactly does the use of the one-sense-per-discourse view provide?

Summary of Experimental Results

- Test data: 12 polysemous words, 3936 avg. number of decisions
- Supervised: The constraint of one-sense-per-collocation provides 96.1% avg accuracy
- Unsupervised: Using just two seed words: 90.6% aa

Summary of Experimental Results

- Unsupervised: Using dictionary defns as seed words: 94.8%
- Unsupervised: Hand corrected dictionary defns: 95.5% aa
- Unsupervised: Using one-sense-per-discourse once at the end: 96.1% aa
- Unsupervised: Using unsupervised algorithm with hand-corrected defns with two views: 96.5% aa

Another Unsupervised Algorithm: Clustering (Schütze 1992)

- For each word w to be disambiguated, pick all the words $\mathbf{v}$ in the document
- Produce a vector $\mathbf{c}$ of the frequency of $\mathbf{v}$
- Use hierarchical clustering based on distance between vectors $\mathbf{c}$

Comparison with (Schütze 1992)

- Since the dimension of these vectors is the number of word types: the sizes are too large for clustering to work properly
- Schütze used Singular Value Decomposition (SVD) to reduce dimensionality of the vectors and to provide a small number of predefined clusters
- Each cluster is then assigned to a particular sense (e.g. based on majority vote using some labelled data)
- Yarowsky (1995) outperforms this kind of clustering algorithm for each polysemous word tested

Other approaches

- Many supervised approaches but relatively few unsupervised algorithms
- Using the EM algorithm to find the right word-sense classification
- Using parallel corpora: exploit the fact that different senses often translate to different words, e.g. Dagan and Itai (1994)
- Depends on accurate alignment between corpora
- Another method of bootstrapping: self-training. Use the output of a single supervised learner to re-train itself.

Points for Discussion

- Are most words polysemous?

- What other kind of view analogous to the one-sense-per-discourse view can be added to the algorithm?