word sense disambiguation

Xiaojin Zhu

jerryzhu@cs.wisc.edu

Computer Science Department
University of Wisconsin, Madison
Word sense disambiguation

- bank? plant?
- WordNet
- word sense disambiguation: given a word and its context, decide a sense
- classification
Yarowsky’s algorithm

- assumptions: one sense per collocation (context), and one sense per discourse (document)
- input $x$: keyword and context
  
  ... company said the plant is still operating
  Although thousands of plant and animal species

- label $y$: the sense classes

- a clever self-training algorithm with many twists
Initial labeled data

- hand-label a small set of context words
- assume \( x \) containing certain context word has the sense of that context word

**example:**  
life \( \mapsto A(birdsens) \), manufacturing \( \mapsto B(machinesens) \)

- build a classifier (decision list, features ranked by log likelihood ratio)
Apply to unlabeled data

- apply the classifier to unlabeled data, add most confident predictions to labeled set
- twist: use ‘global knowledge’, i.e. one-sense-per-discourse, to adjust labels

Example: within the same discourse, ‘fill-in’ unknown or even ‘correct’ labels

A ... the existence of plant and animal life ...
A ... classified as either plant or animal ...
? → A Although bacterial and plant cells are enclosed
B → A are protected by plant parts remaining from
Retrain the classifier

- the context words: “life, manufacturing” would rank the highest
- other context words will be detected from the labeled data:
  - animal (±2 – 10 words) $\mapsto A$
  - equipment (±2 – 10 words) $\mapsto B$
  - employee (±2 – 10 words) $\mapsto B$
  - assembly plant $\mapsto B$
  - ...
  - repeat
The final classifier

Initial context words may no longer be at top, their class can even get flipped.

plant growth $\mapsto A$

car (within $\pm k$ words) $\mapsto B$

plant height $\mapsto A$

union (within $\pm k$ words) $\mapsto B$

equipment (within $\pm k$ words) $\mapsto B$

assembly plant $\mapsto B$

nuclear plant $\mapsto B$

flower (within $\pm k$ words) $\mapsto A$

...
Heuristics against pitfalls

self-training has little means to detect mistakes (which may reinforce itself)

- a training point may get ‘unlabeled’ if its classification confidence drops below a threshold
- incrementally increasing the width of the context window (which adds new feature values to shake up the system)
- randomly perturbing the class-inclusion threshold

heuristics. hard to analyze.