Lexical Relations in Coreference

Yannick Versley
project ELKFED

August 15, 2007

Why Lexical and Encyclopedic knowledge?

- Coreference cases with no matching substrings:

instance	IBM — the company
hypernymy	the computer maker — the company
alias	IBM — Big Blue
(quasi-)synonymy	Album — CD

- Lexical relations need to be combined with salience / compatibility
- Use ontologies or methods from ontology learning

WordNet vs. Patterns vs. Wikipedia

	WordNet	Patterns	Wikipedia
hypernymy	good	fair	—
(quasi-)synonymy	fair	—	—
instance	—	good	good
alias	—	—	good

- Wordnet (and wordnets in general) helps with hypernymy, often also with (quasi-)synonymy, but has poor coverage of proper names
- Use unannotated text and look for patterns that indicate the relation we're interested in
- successful approaches:
  - use the Web as a huge, unannotated corpus
  - exploit a hand-written encyclopedia (e.g. Wikipedia)

Pattern Search on the Web

- Combine multiple patterns
- MI-based threshold: assume the relation if any pattern found and

$$\sum_{rel} \log \frac{N_{\text{found}}(rel, X, Y)}{N_{\text{exp}}(rel, X, Y)} \geq z$$

  (z can be used to adjust precision/recall ratio)
- Want binary feature to combine with salience features (Monsanto – Pioneer – *the company*)

The web pattern feature

- Use two best-performing web patterns (X and other Ys, Ys such as X)
- only for NN→NE, dist $\leq$ 4 sentences
- avoid queries for very rare items
- Use adjective-noun mapping for prenominals (Iraqi → Iraq)

Does it work?

Positive:

- the Chinese government — China
- the Internet bookseller — Amazon.com
- the sprawling archipelago — Indonesia
- the Meryll-Lynch analyst — Bernstein

Does it work?

Positive:

- the Chinese government — China
- the Internet bookseller — Amazon.com
- the sprawling archipelago — Indonesia
- the Meryll-Lynch analyst — Bernstein

Negative:

- the combined company — Compaq
- the Iraqi capital — Moscow
- the largest (. . .) Arab country — United States
- the D.C. area — Prince George

Does it work?

Positive:

- the Chinese government — China
- the Internet bookseller — Amazon.com
- the sprawling archipelago — Indonesia
- the Meryll-Lynch analyst — Bernstein

Negative:

- the combined company — Compaq
- the Iraqi capital — Moscow
- the largest (. . .) Arab country — United States
- the D.C. area — Prince George

Missing:

- the Indonesian businessman — James Riady
- the fifth BTG vice president — Steve Baldwin
- the Washington rheumatologist — Raymond Scalettar

Results on ACE02

Older results (using SVM and older preprocessing)

	bnews	npaper	nwire
recall	+1.3	+0.5	+0.5
precision	−0.4	−0.1	−0.3
F1	+0.3	+0.2	+0.0

- decision tree classifier does not select this feature
- need to do experiments with recent preprocessing

Downsides of using the Web

- Being dependent on Google/Yahoo/Altavista/MSN
  Everything stops working if [*Search engine*]
  discontinues their SOAP API
- Limited reproducability
- Search engines do things (approximate counts, stemming) that gets in our way
- Essentially fixed, limited rate of doing Web queries
  cannot scale up if it's too slow

. . . but still better than crawling the Web yourself

Alternatives

- large corpora (BNC, ≈100M tokens; English Gigaword, ≈1G tokens)
  - pretty big by most standards
  - pattern approach doesn't work, need to do smart stuff
- even larger corpora (ukWaC, ≈5G tokens)
  - bigger, but still much smaller than the Web
- Google 1T 5-gram data
  - still smaller than the web
  - only *n*-grams, harsh frequency cutoff (≥ 40)

But, with such corpora

- Even with a large number of patterns, we get recall problems
- Higher-recall methods
  (distributional similarity, association measures, etc.)
  are not precise enough