Prepositional Phrase Attachment through a Backed-off Model

Paper by Michael Collins and James Brooks

Presentation by Paul Chen

PP Attachment - Problem

- I saw the man with the telescope

  $[V]\ [_{NP}\quad N\ ]\ [_{PP}\ P\qquad N\quad]$

- Meaning:
  - I used the telescope to see the man?
  - I saw the man carrying the telescope?

PP Attachment - Disambiguate

- Accuracy
  - Always noun attachment – 59%
  - Most likely for each preposition – 72.2%
  - Human (looking at the 4 words) – 88.2%
  - Human (with whole sentence) – 93.2%

PP Attachment - Disambiguate

- “I saw the man with the telescope”
  - Input: 4 words

    (v=saw, n1=man, p=with, n2=telescope)

  - Output:
    - if Noun attachment -> 1
    - if Verb attachment -> 0

Maximum Likelihood Estimate

- If p(1|v,n1,p,n2) >= 0.5, PP attach to Noun
- p(1|v,n1,p,n2) = $\dfrac{f(1,v,n1,p,n2)}{f(v,n1,p,n2)}$
- Problem: lots of 0 counts (sparse data ☹)

Backed-off Estimate

- f(v,n1,p,n2) = 0,
- but maybe f(n1,p,n2) > 0 ?
- 6 triples to look at:
  - (v,n1,p), (v,p,n2), (v,n1,n2), (n1,p,n2),…
- 6 doubles:
  - (v,n1),(v,n2),(v,p),(n1,p),(n1,n2),(p,n2)
- 4 singles…

Backed-off Estimate

- Use a combination of triples

$$p(1|v,n1,p,n2) = \frac{f(1,v,n1,p) + f(1,v,p,n2) + f(1,n1,p,n2)}{f(v,n1,p) + f(v,p,n2) + f(n1,p,n2)}$$

- Only use the ones containing the preposition
  - But lower accuracy if we use (v,n1,n2) and such

Backed-off Estimate

- Combination of doubles

$$p(1|v,n1,p,n2) = \frac{f(1,v,p) + f(1,n1,p) + f(1,p,n2)}{f(v,p) + f(n1,p) + f(p,n2)}$$

- Singles: just use the preposition

$$p(1|v,n1,p,n2) = \frac{f(1,p)}{f(p)}$$

Backed-off Model - Results

- Corpus – WSJ Treebank
  - Training: 20801
  - Testing: 3097

Stage	Total number	Number correct	Percent correct
quadruples	148	134	90.5%
triples	764	688	90.1%
doubles	1965	1625	82.7%
singles	216	155	71.8%
defaults	4	4	100.0%
totals	**3097**	**2606**	**84.1%**

Morphological Analysis

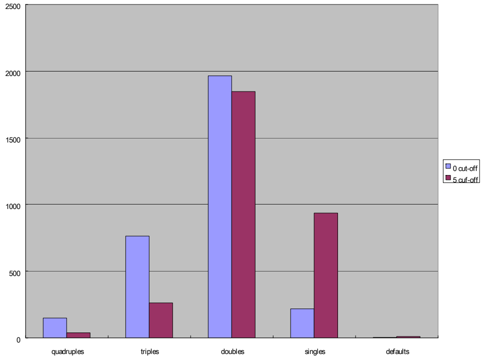
- Getting a little more accurate by preprocessing the data
  - Replace 4-digit numbers with 'YEAR'
  - Replace numbers with 'NUM'
  - First-letter-capitalized word becomes 'NAME'
  - Verbs reduced to their stem form (running becomes run)
- Result… 84.5% (0.4% increase)

Comparison with other work

- HR93– an unsupervised approach
  - ~80%
- RRR94 – Maximum entropy model
  - 81.6%
- BR94 – Greedy search for transformation rules
  - 81.9%
- Backed-off with morphological analysis
  - 84.5%

Increasing the cut-off count

- Increasing the cut-off count to 5.
  - Accuracy down to 81.6% (-2.5%)



- Smoothing?