Prepositional Phrase Attachment through a Backed-Off Model

Michael Collins and James Brooks

Department of Computer and Information Science

University of Pennsylvania

Presented by Jane Li

March 4, 2004

Problem Outline

- **Data format**

  element quadruple $\{V = v, N_1 = n1, P = p, N_2 = n2\}$
  e.g., $\{V = joined, N_1 = board, P = as, N_2 = director\}$

- **Attachment decision**

$$A = \begin{cases} 1 & : \quad attach\ to\ noun; \\ 0 & : \quad attach\ to\ verb. \end{cases}$$

Strategy

- **Probability decision**

$$\hat{p}(A = 1 \,|\, V = v, N_1 = n1, P = p, N_2 = n2)$$

  - if $\hat{p}(1 \,|\, v, n1, p, n2) \geq 0.5$

    PP attaches to noun $\langle n1 \rangle$

  - else

    PP attaches to verb $\langle v \rangle$

Maximum Likelihood Estimation

- **Notation**: f – the number of times a particular tuple occurs in training data. **e.g.**

  1. $f(1, is, revenue, from, research)$

     number of times of (*is, revenue, from, research*) with noun attachment.

  2. $f(1, P = from)$

     number of times of (*P* = *from*) occurs with noun attachment.

  3. $f(V = is, N2 = research)$

     number of times of (*V* = *is, N2 = research*) occurs with either attachment.

- **Maximum likelihood estimation**

$$\hat{p}(1 \mid v, n_1, p, n_2) = \frac{f(1, v, n_1, p, n_2)}{f(v, n_1, p, n_2)}$$

Backing off Estimate

- **N-gram model for speech recognition**

$$\hat{p}(w_n \mid w_1, w_2, ..., w_{n-1}) = \frac{f\left(w_1, w_2, ..., w_n\right)}{f\left(w_1, w_2, ..., w_{n-1}\right)}$$

- **Katz'87 backing off**

  Hypothesis: If the current level n-gram frequency count is lower than a given threshold, then the count at this level is regarded to be too low to give an accurate estimate.

Backing off Estimate (cont.)

- **If** $f\left(w_1, w_2, \ldots, w_{n-1}\right) > c_1$
  $\hat{p}\big(w_n \,|\, w_1, w_2, \ldots, w_{n-1}\big) = \frac{f\left(w_1, w_2, \ldots, w_n\right)}{f\left(w_1, w_2, \ldots, w_{n-1}\right)}$

- **Else if** $f\left(w_2, w_3, \ldots, w_{n-1}\right) > c_2$
  $\hat{p}\big(w_n \,|\, w_1, w_2, \ldots, w_{n-1}\big) = \alpha_1 \times \frac{f\left(w_2, w_3, \ldots, w_n\right)}{f\left(w_2, w_3, \ldots, w_{n-1}\right)}$

- **Else if** $f\left(w_3, w_4, \ldots, w_{n-1}\right) > c_3$
  $\hat{p}\big(w_n \,|\, w_1, w_2, \ldots, w_{n-1}\big) = \alpha_1 \times \alpha_2 \times \frac{f\left(w_3, w_4, \ldots, w_n\right)}{f\left(w_3, w_4, \ldots, w_{n-1}\right)}$

Backing off Estimate (cont.)

- **Backing off in PP attachment**

  Hypothesis: Among backing off tuples, ones containing the observed preposition are particularly important to the attachment decision.

- **Backing off to triples**

  $$(v, n_1, p, n_2) \rightarrow ((v, n_1, p), (v, p, n_2), (n_1, p, n_2), (v, n_1, n_2))$$

  $$\hat{p}_{triple}(1 \mid v, n_1, p, n_2) = \frac{f(1,v,n_1,p)+f(1,v,p,n_2)+f(1,n_1,p,n_2)}{f(v,n_1,p)+f(v,p,n_2)+f(n_1,p,n_2)}$$

- **Backing off to pairs**

  $$(v, n_1, p, n_2) \rightarrow ((v, p), (n_1, p), (p, n_2), (v, n_1), (v, n_2), (n_1, n_2))$$

  $$\hat{p}_{pair}(1 \mid v, n_1, p, n_2) = \frac{f(1,v,p)+f(1,n_1,p)+f(1,p,n_2)}{f(v,p)+f(n_1,p)+f(p,n_2)}$$

PP attachment Backing off Algorithm

- **No backing off:** If $f(v, n_1, p, n_2) > 0$

$$\hat{p}(1 \mid v, n_1, p, n_2) = \frac{f(1, v, n_1, p, n_2)}{f(v, n_1, p, n_2)}$$

- **Backing off to triples:**

  Else if $f(v, n_1, p) + f(v, p, n_2) + f(n_1, p, n_2) > 0$

$$\hat{p}(1 \mid v, n_1, p, n_2) = \frac{f(1, v, n_1, p) + f(1, v, p, n_2) + f(1, n_1, p, n_2)}{f(v, n_1, p) + f(v, p, n_2) + f(n_1, p, n_2)}$$

PP attachment Backing off Algorithm (cont.)

- **Backing off to pairs:**

  Else if $f(v, p) + f(n_1, p) + f(p, n_2) > 0$

$$\hat{p}(1 \mid v, n_1, p, n_2) = \frac{f(1, v, p) + f(1, n_1, p) + f(1, p, n_2)}{f(v, p) + f(n_1, p) + f(p, n_2)}$$

- **Backing off to singles:**

  Else if $f(p) > 0$

$$\hat{p}(1 \mid v, n_1, p, n_2) = \frac{f(1, p)}{f(p)}$$

- **Default:**

  Else $\hat{p}(1 \mid v, n_1, p, n_2) = 1.0$

  (default is noun attachment.)

Lower and Upper Bounds on Performance

- **IBM data**: training and test data

Method	Percentage Accuracy
Always noun attachment	59.0
Most likely for each preposition	72.2
Average Human (4 head words only)	88.2
Average Human (whole sentence)	93.2

Results

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
Total	**3097**	**2606**	**84.1**

Results with “Morphological Analysis”

- **Morphological replacement**
  - 4-digit number $\Longrightarrow$ ’YEAR’
  - Other number strings $\Longrightarrow$ ’NUM’
  - Verb, prep. fields $\Longrightarrow$ lower case
  - Strings with capital letter $\Longrightarrow$ ’NAME’
  - NAME-NAME $\Longrightarrow$ ’NAME’
  - Morphological stem

Results with "Morphological Analysis" (cont.)

- **Result**

Stage	Total Number	Number Correct	Percent Correct
Quadruples	242	224	92.6
Triples	977	858	87.8
Doubles	1739	1433	82.4
Singles	136	99	72.8
Defaults	3	3	100
Total	**3097**	**2617**	**84.5**

Results Comparison

- **Comparison with others**

Method	Percentage Accuracy
Ratnaparkhi et al.'94 (words only)	77.7
Ratnaparkhi et al.'94 (words and classes)	81.6
Brill and Resnik'94 (words only)	81.9
Backing off (no processing)	84.1
Backing off (morphological processing)	84.5

Results Comparison (cont.)

- **Closer comparison with Hindle et al.'93**
  - Hindle et al.'s data (v, n_1, p)
  - Hindle et al.'s decision

$$\begin{cases} \frac{f(n_1,p)}{f(n_1)} \geq \frac{f(v,p)}{f(v)} & : \quad attach\ to\ noun; \\ Otherwise & : \quad attach\ to\ verb. \end{cases}$$

- **Leave out n_2 to match Hindle et al.'s data**
  - $(v, n_1, p, n_2) \ \rightarrow \ (v, n_1, p)$
  - Decision

$$\begin{cases} \frac{f(1,n_1,p)}{f(1,n_1)} \geq \frac{f(0,v,p)}{f(0,v)} & : \quad attach\ to\ noun; \\ Otherwise & : \quad attach\ to\ verb. \end{cases}$$

Results Comparison (cont.)

– Backing off estimates

$$\hat{p}(1\,|v, n_1, p) = \frac{f(1, v, p) + f(1, n_1, p)}{f(v, p) + f(n_1, p)}$$

- Results compared with Hindle et al.

Hindle's method: 82.1%

Backing off stategy: 86.5%

Closer Look at Backing off

- **Low counts matter?**
  - N-gram language model: (< 5)
  - Backing off in the paper: ($= 0$)

Stage	Total Number	Number Correct	Percent Correct
Quadruples	39	38	97.4
Triples	263	243	92.4
Doubles	1849	1574	85.1
Singles	936	666	71.2
Defaults	10	5	50.0
Total	**3097**	**2526**	**81.6**

Closer Look at Backing off (cont.)

- **What kinds of tuples are better?**

Triples		Doubles		Singles	
Tuple	**Accuracy**	**Tuple**	**Accuracy**	**Tuple**	**Accuracy**
n1 p n2	90.9	n1 p	82.1	p	72.1
v p n2	90.3	v p	80.1	n1	55.7
v n1 p	88.2	p n2	75.9	v	52.7
v n1 n2	68.4	n1 n2	65.4	n2	47.4
		v n1	59.0		
		v n2	53.4		

Conclusion

- Better than other methods
- Surprising result: the significance of low count events
- Conceptually simple and computationally inexpensive to implement

What's Happened since 1995?

- **Semantic-class-based approaches**
  - **Niemann'98:** Semantic role determines the appropriate attachment of the preposition.

    e.g.

    a) I saw the girl [with a basketball].

    b) the hand [with the broken finger]

    1. Hand coded each noun, verb, adjective, and adverb
    2. Group nouns and verbs according to WordNet synsets
    3. Find attachment patterns for each preposition
    4. Use pattern occurrence time to determine PPs in future

What's Happened since 1995? (cont.)

- **Pantel&Lin'00:** Compare attachment scores for tuples $(v, p, N_2), (N_1, p, N_2)$ and choose the higher one.
  1. $v, N_2 \leftarrow$ contextually similar words
  2. Compute the average adverbial attachment score
  3. $N_1, N_2 \leftarrow$ contextually similar words
  4. Compute the average adjectival attachment score
  5. Compute the final decision score by combination
  6. Make attachment decision by the final decision score.

- **Statistical approach**
  - **Volk'01:** WWW-based statistical method