

# Structural Ambiguity and Conceptual Relations

---

Philip Resnik and Marti A. Hearst

Presented by Yun Niu

Department of Computer Science

University of Toronto

# The ambiguity in PP attachment

---

- a. Eventually, Mr. Stoll was invited to both the CIA and NSA [to brief [high-ranking officers *on computer theft*] ].
- b. Eventually, Mr. Stoll was invited to both the CIA and NSA [to brief [high-ranking officers] [*on computer theft*] ].

# Structurally-based preference strategies

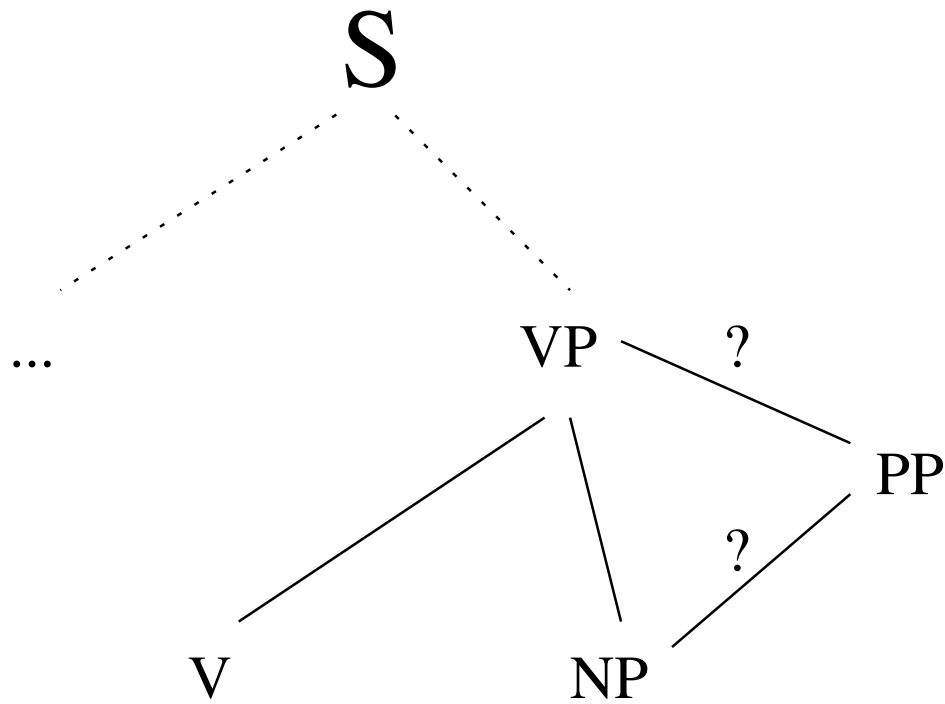
---

- Right association: a constituent tends to attach to another constituent immediately to its right (Kimball 1973).
- Minimal attachment: a constituent tends to attach so as to involve the fewest additional syntactic nodes (Frazier 1978).

Problem: not adequate.

# Structurally-based preference strategies

---



# Preference semantics

---

Wilks et al., 1985: Represent the various meanings of the preposition in terms of:

- (a) the preferred semantic class of the noun or verb that proceeds the preposition (e.g., move, be)
- (b) the case of the preposition (e.g., instrument, time)
- (c) the preferred semantic class of the head noun of the prepositional phrase (e.g., event)

Problem: hand-encoded preference information; hard to evaluate the strengths of the preferences.

# Corpus-based preference identification (Hindle and Rooth, 1991)

---

Lexical attachment preference on the basis of lexical co-occurrence statistics

- The *4-tuple* representation: (brief, officer, on, theft):  
 $(v, n_1, p, n_2)$
- The attachment strategy:  $Pr(p | n_1)$  versus  $Pr(p | v)$ .  
T-test: a measure of dissimilarity (Church et al. 1991)

$$t = \frac{Pr(p | n_1) - Pr(p | v)}{\sqrt{\sigma^2(Pr(p | n_1)) + \sigma^2(Pr(p | v))}}$$

# Corpus-based preference identification (Hindle and Rooth, 1991)

---

Problem: ignore  $n_2$ ?

Britain [reopened [its embassy] [*in December*] ].

Britain [reopened [its embassy *in Teheran*] ].

Another strategy:  $Pr(p, n_2 | n_1)$  versus  $Pr(p, n_2 | v)$ ?

No, sparse data problem.

# Corpus-based preference identification using conceptual relation

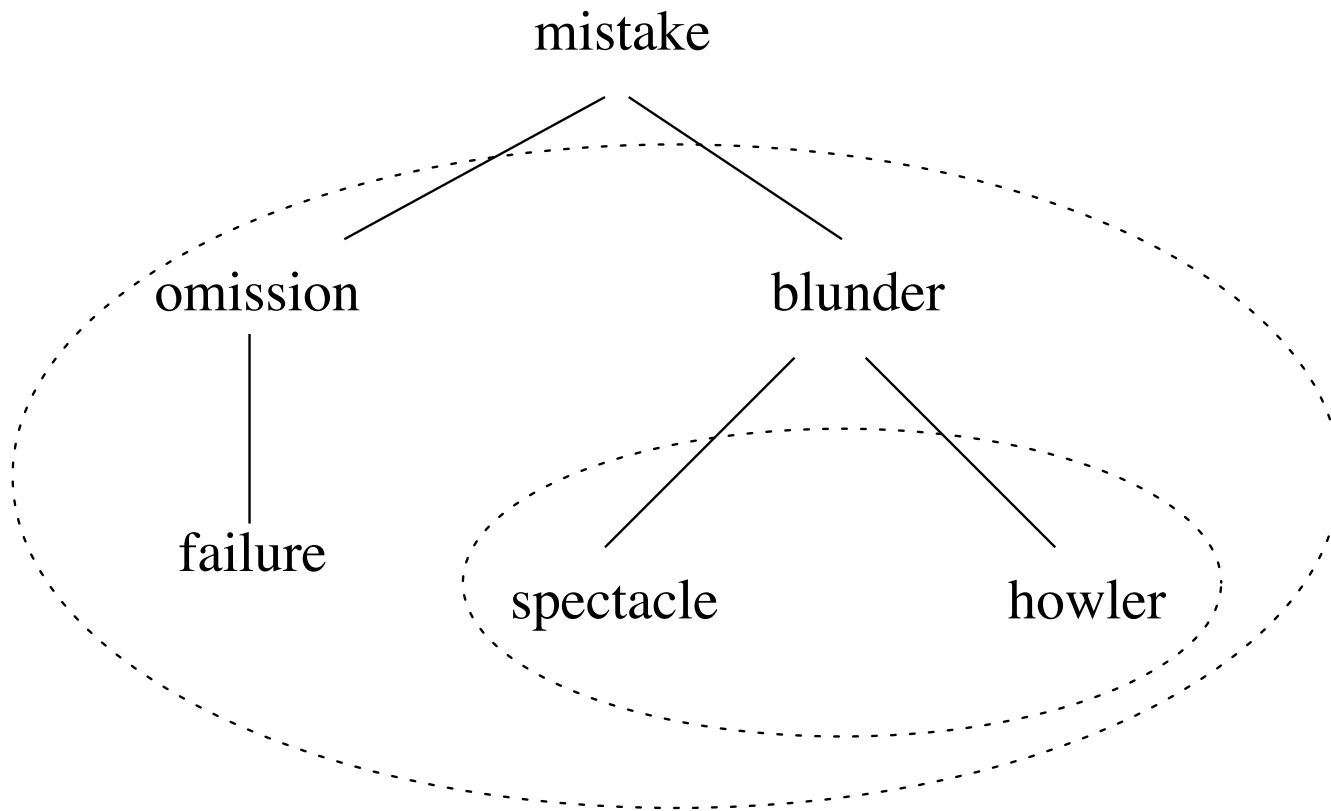
---

- Simulate lexical relations by corresponding class relations
- A generalized model using classes  
words(*c*): the set of all words that are a member of class *c* or any subordinate class.

$$f(c) = \sum_{n \in \text{words}(c)} f(n)$$

# Corpus-based preference identification using conceptual relation

---



# The Algorithm

---

1. Let  $C_1 = \{c \mid n_1 \in \text{words}(c)\}$

Let  $C_2 = \{c \mid n_2 \in \text{words}(c)\} = \{c_{2,1}, \dots, c_{2,N}\}$

2. For  $i$  from 1 to  $N$ ,

$$c_{1,i} = \underset{c \in C_1}{\operatorname{argmax}} I(c; p, c_{2,i})$$

$$I_i^n = I(c_{1,i}; p, c_{2,i})$$

$$I_i^v = I(v; p, c_{2,i})$$

3. For  $i$  from 1 to  $N$ ,

$$S_i^n = \text{freq}(c_{1,i}, p, c_{2,i}) I_i^n$$

$$S_i^v = \text{freq}(v, p, c_{2,i}) I_i^v$$

4. Compute a paired samples t-test for a difference of the means of  $S^n$  and  $S^v$ . Let “confidence” be the significance of the test with  $N - 1$  degrees of freedom.
5. Select attachment to  $n_1$  or  $v$  according to whether  $t$  is positive or negative, respectively.

# Experiment 1

---

Corpus: 1988-89 Wall Street Journal (for training and testing).

Ambiguous PP attachments: 174 instances.

Table 1:

	LA	CA	Combined
% Correct	81.6	77.6	82.2

# Experiment 1

---

Table 2: Answer only when confident

Strategy	Answered%	Accuracy%
LA	44.3	92.8
CA	67.2	84.6

# Experiment 2

---

Supposition: a strategy based upon a domain-independent semantic taxonomy would provide a greater degree of robustness, reducing dependence of the attachment strategy on the training corpus.

New test set: 173 instances of ambiguous PP attachments.

Table 3:

	LA	CA	Combined
% Correct	69.9	72.3	72.8

# Experiment 2

---

Table 4: Answer only when confident

Strategy	Answered%	Accuracy%
LA	31.8	80.0
CA	49.7	77.9

# Experiment 3

---

Training set:

Penn Treebank's parsed version of the Brown corpus.

Test set:

174 instances in experiment 1.

Table 5:

	LA	CA	Combined
% Correct	77.6	73.6	79.3

# Experiment 3

---

Table 6: Answer only when confident

Strategy	Answered%	Accuracy%
LA	35.6	85.5
CA	59.2	81.6

# Conclusions

---

- Improve the coverage.
- Helpful in the sparse data problem.
- Mutual information is effective?
- Combining evidence using the paired t-test is problematic.

# Conclusions: cont.

---

- Incorporation of structurally-based attachment strategies along with lexical and conceptual association.
- Application of similar techniques to other problems, e.g., noun-noun modification.