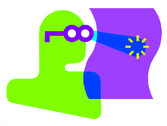
Statistical Analysis for Thesaurus Construction using an Encyclopedic Corpus

Yasunori OHISHI[1], Katunobu ITOU[2], Kazuya TAKEDA[1], Atsushi FUJII[3]
[1]Nagoya University, Japan, [2]Hosei University, Japan, [3]University of Tsukuba, Japan

MOTIVATON

- Thesaurus
  - Expanding the range of possible terms in information retrieval
  - Example-based machine translation system
  - Most technical terminology is not included
- Automatic thesaurus construction
  - Determining whether a word pairs has a hierarchical relation or a disrelation



Word X ? Word Y

PREVIOUS WORK

- To extract hyponyms, synonyms, and hypernyms,
  - Sentences that have specific syntactic patterns (“a part of” “is a” “such as”) (Marti, 1992; Tsurumaru, 1991)
  - Descriptions in a dictionary (Suzuki, 2003)
  - Specific document structure (Shinzato, 2004)

  are used

⇨ It is difficult to cover such an enormous vocabulary

Less frequent words are not expected in the desired expressions

THE CYCLONE CORPUS

- Collection procedure (Fujii, 2005)

① Searching the Web for pages including a target term
② Analyzing the layout of the pages and identifying paragraphs that potentially describe the target term
③ Classifying multiple paragraphs into predefined domains

Headword (Target term)
- **Description 1** .....
- **Description 2** .....

New terms and rarer terms
Many descriptions for each headword
- Semantic relations between a headword and terms in its description

⇨ **Extracting hierarchical relations effectively**

HIERARCHICAL RELATION ESTIMATION METHOD

- Descriptions of words have “directionality”

Lion
a large wild animal of the cat family with yellowish brown fur …

Animal
a living creature such as a dog or cat that is generally distinguished from plants …

A collection of descriptions for a headword shares common hypernyms but may not share hyponyms

- Target function $H(X|Y) = C(X|Y) - C(Y|X)$

$C(X|Y)$: Probability of X , given the descriptions of Y

- Semantic expansion technique for descriptions

Persian cat
a type of cat with long hair, short thick legs and a round face

Cat
a small four-legged furry animal with a tail and claws …

Hypernym — Hyponym

Animal —— Cat —— Persian Cat

1st – order probability matrix $\mathbf{A}$

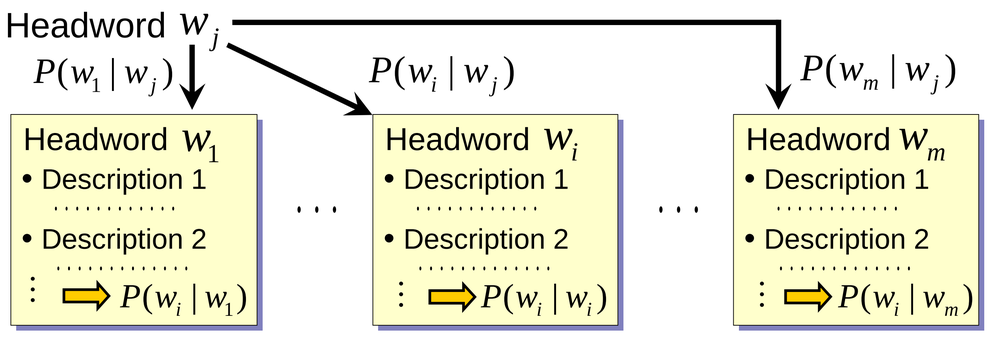
$$\mathbf{A}_{i,j} = P(w_i \mid w_j) \qquad i, j = 1, \ldots, m$$

Relative frequency :

$$P(w_i \mid w_j) = \frac{F(w_i \mid w_j)}{\sum_{k=1}^{m} F(w_k \mid w_j)}$$

m is the number of headwords

2nd – order probability matrix $\mathbf{A}^2$



$$\mathbf{A}^2_{i,j} = P(w_i^{(2)} \mid w_j) = \sum_{all\ k} P(w_i \mid w_k) P(w_k \mid w_j)$$

nth – order probability matrix is defined as a square matrix $\mathbf{A}^n$

Expanded probability matrix

$$C = \sum_{i=1}^{N} \left[\alpha_i \mathbf{A}^i \right] \qquad \left[N \text{ is the maximum order of expansion} \right]$$

Identifying the relation of a word pair (w_i, w_j) by calculating $H(w_i \mid w_j)$

EVALUATION

- Word pairs in the computer-related domain (2074 words)

Correctness of description	Average number of descriptions per headword	Test sets: Hierarchical relation data	Test sets: Disrelation data
A	6.6	136 pairs	301 pairs
A+B	10.4	168 pairs	366 pairs
ALL	80.7	206 pairs	497 pairs

A: Correct
B: Partially correct

Example

Hierarchical relation data: Hypernym	Hierarchical relation data: Hyponym	Disrelation data	
ISO	MPEG	Bit error rate	MS-DOS
Operations research	Linear programming	Polish notation	10BASE-T
OS	UNIX	Compiler	Zoned decimal
Protocol	TCP/IP	Hexadecimal number	CATV
Memory	DRAM	Dual system	Graph theory

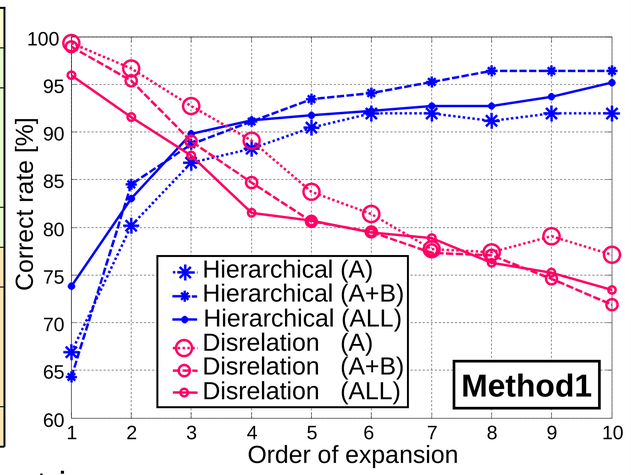
- To determine weights $\alpha_i, (i = 1, \cdots, N)$

Method1 Estimation using linear discrimination analysis (proposed method)
Method2 Exponential weight method (Suzuki, 2003)

$$C = \lim_{k \to \infty} b(aA + a^2A^2 + \cdots + a^kA^k) \qquad 0 < a < 1,\ b = \frac{1-a}{a}$$

- Hierarchical relation or disrelation ?

Test set	A	A+B	ALL
Method 1			
(Hierarchical) (%)	86.8	83.9	90.8
(Disrelation) (%)	92.7	94.8	86.1
(order)	3	2	3
Average (%)	89.7	89.4	88.4
Method 2			
(Hierarchical) (%)	80.4	80.9	76.9
(Disrelation) (%)	96.5	95.1	96.4
(Weight a)	0.6	0.5	0.3
Average (%)	88.5	88.0	86.7



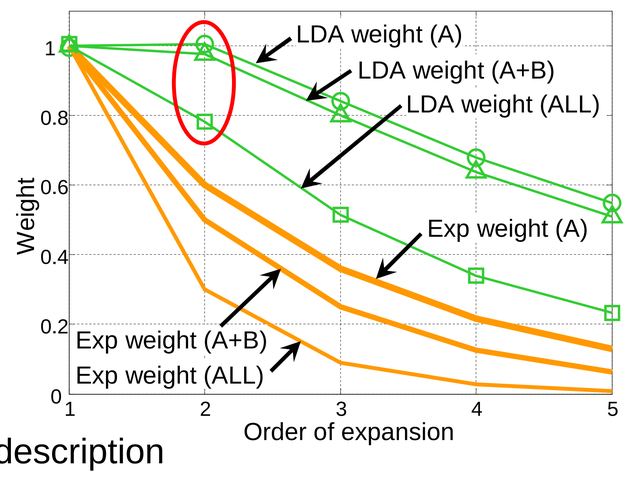
- Weights for the expanded matrices

(Weights are normalized by each first-order weight)

For method 1, a weight on the second-order matrix is a high value

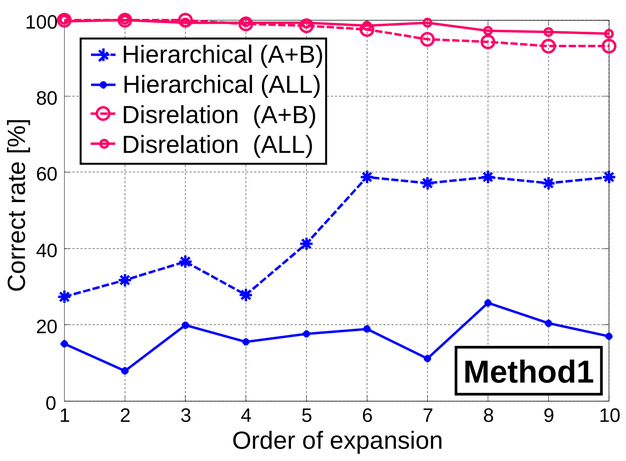
⇕

For method 2, a weight on the first-order matrix is the highest value



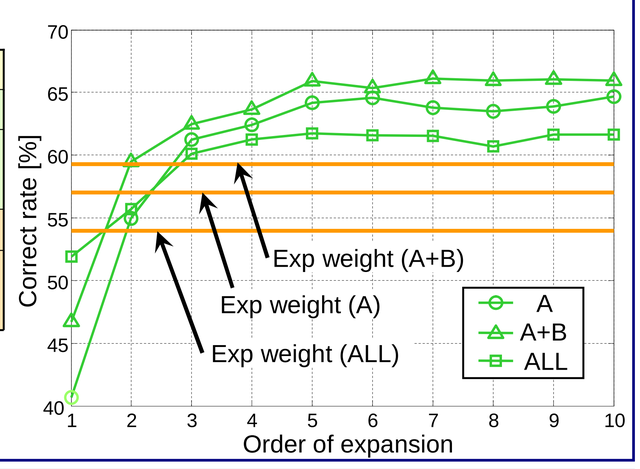
- Using only the top correct description

Test set	A+B	ALL
Method 1		
(Hierarchical) (%)	58.7	25.7
(Disrelation) (%)	97.5	97.2
(order)	6	8
Average (%)	78.1	61.4
Method 2		
(Hierarchical) (%)	49.4	29.2
(Disrelation) (%)	98.4	98.4
(Weight a)	0.99	0.999
Average (%)	73.9	63.8



- Hierarchical discrimination

Test set	A	A+B	ALL
Method 1			
Proposed (%)	64.6	66.1	61.7
(order)	6	7	5
Method 2			
Exponential (%)	57.0	59.3	54.0
(Weight a)	0.6	0.5	0.3



Effectiveness of using multiple and correct descriptions

CONCLUSION AND FUTURE WORK

- Conclusion
  - Discrimination for the hierarchical relation of a word pair using an encyclopedic corpus called the Cyclone corpus
  - In order not to miss an indirect relationship, a semantic expansion technique for descriptions is used
  - The proposed method is able to detect 66.1% of relations
- Future work
  - Discrimination between hierarchical and synonymous relation