Information Retrieval Based on Linguistic Structure

Takashi MIYATA (CREST, JST)
Kôiti HASIDA (CARC, AIST)

Japanese-German Workshop on NLP:
NLP for Information Management and Semantic Web
4-5/July, 2003 (Sapporo)

Target and Requirement

- **Requests that specify 'content' clearly**
  - × documents on robots
  - ○ documents reporting **robots build houses**
  - ⋆ *Find documents, not "standard" web sites*
- To treat such requests, the followings are needed:
  - Clarification of users' intention
  - Implement highly accurate document search
  - **Interaction between user and computer to compensate recall**

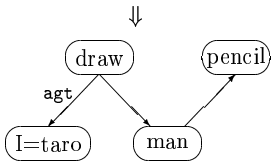
Background

- Large amount of machine-readable documents
  ⇒ provide research motivation and resource
- Recent improvement of parsing technology
  [Charniak 2000][Kudo&Matsumoto 2000]
  ⇐ require large amount of **annotated** corpora

Popularization of annotation by providing useful application would promote NLP research, and vice versa.

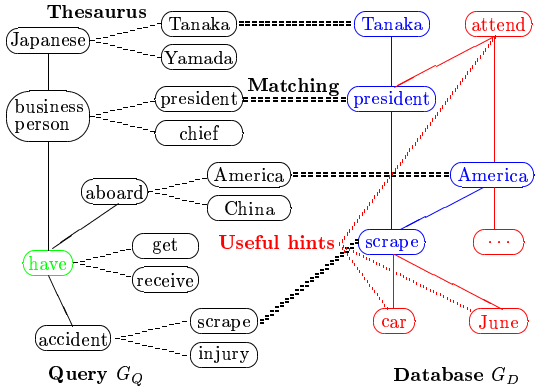
Annotation and Semantic Structure

```
<su><np opr="agt" eq="taro">I</np>
    <v>drew</v>
    <np><n>a man</n>
        <adp>with a pencil</adp></np>
    </su>
```

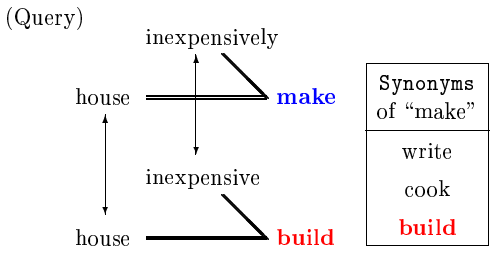


Annotation for "I drew a man with a pencil" (above) and derived semantic structure (below)

Semantic Structure in IR



Structure based Similarity Measure



(Subgraph in database (candidate))

Prefer synonyms fitting in the context of the query and the database

Graph Embedding

- NP-hard problem [Zhang et al 1996]
- Query graph can be assumed small.
  ⇒ Enumerate candidates and their scores by dynamic programming (not always return strict solution)
- Undirected graphs with unlabeled edges are assumed for simplicity.

Example (1/2)

Suppose to find articles that report:

"building houses with lower cost by robots"

1. Input keywords: `build`, `house`, `cost`, `use`, `robot`
2. and synonyms: `estate` for house, `machine` for robot, etc
3. Input edges: `build`–`house`, `use`–`robot`, etc
4. Find related/synonymous keyword `construct` for build, which happened to be less preferred in thesaurus.

Demonstration

Example (2/2)

- Thinking of related/synonymous words is quite difficult.
- Preparation of general and complete thesaurus in advance is impossible.
  ⇒ semantic structure can complement thesaurus.

Evaluation

- 100,000 articles (Mainichi Newspaper) in 1994 are converted into semantic graphs by KN Parser [Kurohashi 1996]
- 8 subjects perform 4 tasks each.

Database Statistics		
# nodes	13,652,694	
# edges	10,928,259	
thesaurus size	149,270	(distinct words)

Tasks

Each subject finds the 4 kinds of articles that report:

1. *A boy who beat Prime Minister Major in a vote*
2. *Subsidiary that will be set up in future is evaluated better than its parent company.*
3. *Area in China where people obtain capitals from aboard*
4. *Phone calls rush in when the party appears in mass media.*

(for practice) *building houses with lower cost by robots*

Settings

Subjects can make use of:

A. Keywords and thesaurus

B. Keywords, structure, and thesaurus

C. Keywords, structure, and thesaurus augmented by word co-occurrences in a sentence

D. Keywords, structure, and thesaurus augmented by neighborhood in semantic graph

Experimental Design

	Task 1	Task 2	Task 3	Task 4
Sbj 1	A	B	C	D
Sbj 2	A	B	D	C
Sbj 3	B	A	C	D
Sbj 4	B	A	D	C
Sbj 5	C	D	A	B
Sbj 6	C	D	B	A
Sbj 7	D	C	A	B
Sbj 8	D	C	B	A

Performance

	Task 1	Task 2	Task 3	Task 4	Total
Sbj 1	NG	**OK**	**OK**	NG	2
Sbj 2	**OK**	**OK**	**OK**	**OK**	4
Sbj 3	**OK**	NG	**OK**	**OK**	3
Sbj 4	**OK**	**OK**	**OK**	NG	3
Sbj 5	**OK**	NG	NG	**OK**	2
Sbj 6	**OK**	**OK**	NG	**OK**	3
Sbj 7	NG	NG	NG	NG	0
Sbj 8	NG	**OK**	**OK**	**OK**	3
Total	5	5	5	5	20

Details in 20 Success

	rank	time(min)	# pages	# operations
A	14.00(12.90)	15.45(14.68)	30.00(26.07)	†**29.50(9.63)**
B	24.50(32.53)	13.88(7.92)	25.33(21.76)	16.00(7.37)
C	1.50(0.76)	∗37.85(47.90)	16.83(16.00)	†**11.17(3.44)**
D	4.75(4.82)	8.72(5.18)	12.00(8.22)	12.00(12.55)

(means and standard deviations)

(∗) One of the subjects extremely took time (136.3 minutes). Without this data, the average becomes 21.65.

(†) Difference is statistically significant(t-test, 5%).

Details in Operations

	∗node (add/del)		∗edge (add/del)		syn (add/del)	
A	†**7.50(2.29)**	3.00(3.08)	0.00(0.00)	0.00(0.00)	‡**17.00(7.31)**	2.00(2.45)
B	2.17(2.11)	1.33(1.70)	2.33(2.21)	1.50(1.80)	6.83(3.02)	1.17(2.19)
C	†**1.67(1.80)**	0.17(0.37)	2.00(1.63)	0.83(1.07)	‡**5.50(1.50)**	0.33(0.75)
D	2.00(2.45)	0.50(0.87)	1.50(2.06)	0.75(0.83)	6.25(6.46)	0.25(0.43)

(means and standard deviations)

(∗) Excluding those in initially generated graphs by parser.

(†,‡) Differences are statistically significant(t-test, 5%).

Summary

- Searching in small and medium sized sets of documents based on semantic structure is already feasible within the current technology.
- Semantic structure can provide useful hints for query revision.
- Making users learn and understand 'tips' of search with semantic structure would be needed.