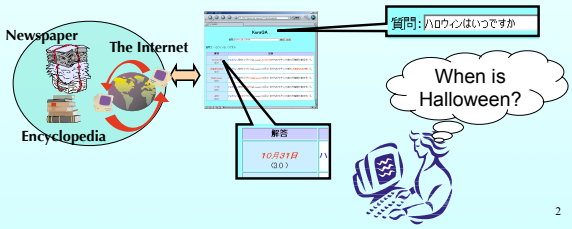
Automatic Extraction of Attribute Relations from Text

Computational Linguistics Lab.
Tetsuro TAKAHASHI

2004/11/25
COE technical presentation

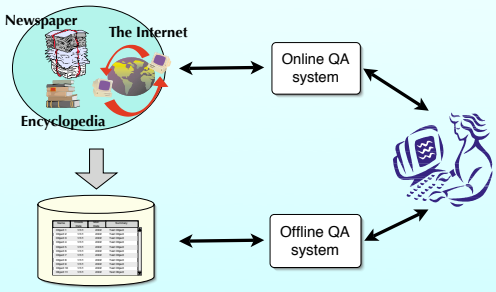
Background

- Efficient Information Access to vast amounts of text data on the Web and in encyclopedias and newspapers



Background

- Online strategy vs. Offline strategy

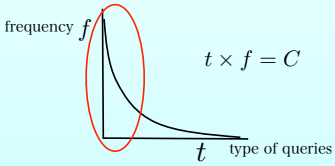


Motivation

- Offline QA strategy

  [Fleischman et al.03, Lin and Katz03]

  - Types of users' questions quantitatively obey Zipf's Law.
  - A small fraction of question types accounts for a significant portion of all question instances.



Motivation

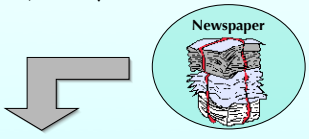
- Offline QA strategy
  - Analysis of questions in QAC2
  - 40% (79/200) of questions asked about Attribute Relations

石ノ森章太郎さんの出身地はどこですか？
Ishinomori Shotaro　birthplace

Attribute Relations

- Objectives
  - Extract attribute relation triplets

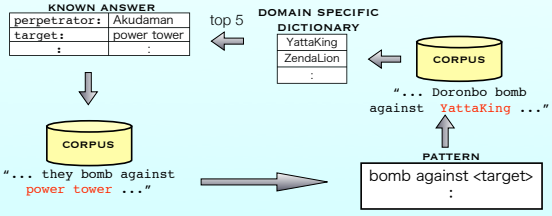
    (Object, Attribute, Value) from text



(Object, Attribute, Value)
(Torino, population, 2.2 million)
(Akashi-Kaikyo-Bridge, length, 3910m)
(Japan subway, parent company, suntory)
:

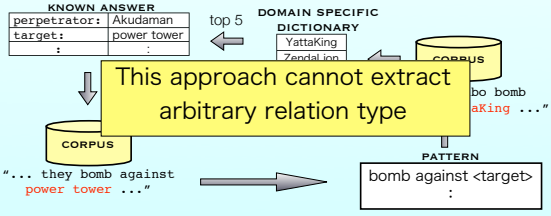
Related Work

- Pattern + Bootstrapping
  - Extracts domain specific words and patterns (Brin98, Riloff99)



Related Work

- Pattern + Bootstrapping
  - Extracts domain specific words and patterns (Brin98, Riloff99)



This approach cannot extract arbitrary relation type

Related Work

- Machine Learning
  - extracts answer candidates using patterns
  - Predicts answer by supervised machine learning

(Zelenko02, Fleischman03)

Training data is required for each type of attribute relation

Approach

- Proposed approach
  - Extract triplet using abstracted patterns
  - Filter out noise using a statistical measure

Birthday : *X* is born on *Y*
Author : author of *X* is *Y*

⇒

X is *Z* on *Y*
Z of *X* is *Y*

Score(O,A,V) = *Score*(Mt.Fuji, height, 3776m)

Approach

$$Score(O,A,V) = S_{OA}(O,A) \times S_{VA}(V,A)$$

Score(Mt.Fuji, height, 3776m)
= S_{OA}(Mt.Fuji, height) × S_{VA}(3776m, height)

S_{OA}(Mt.Fuji, height): Mt.Fuji has an attribute height
S_{VA}(3776m, height): 3776m is a value of height

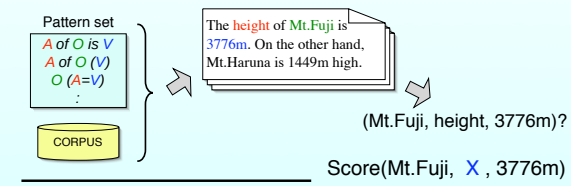
Weighted Mutual Information for both S_{OA} and S_{VA}

$$weighted\_MI = p(x,y) \times \log_2 \frac{p(x,y)}{p(x)p(y)}$$

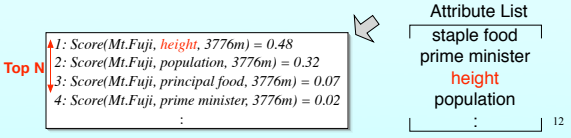
- ☑ Extracts arbitrary relations
- ☑ Training data is not required

Overview

- Estimation of an attribute relation

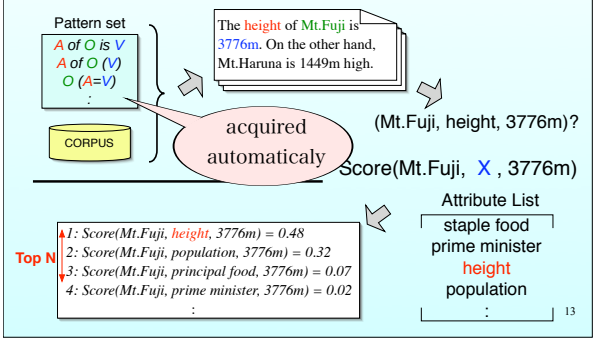


Score(Mt.Fuji, X , 3776m)



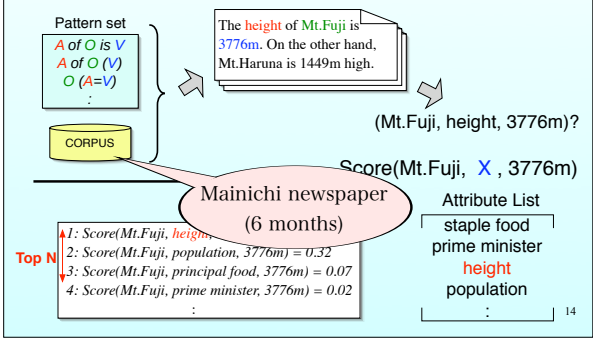
Overview

- Estimation of an attribute relation



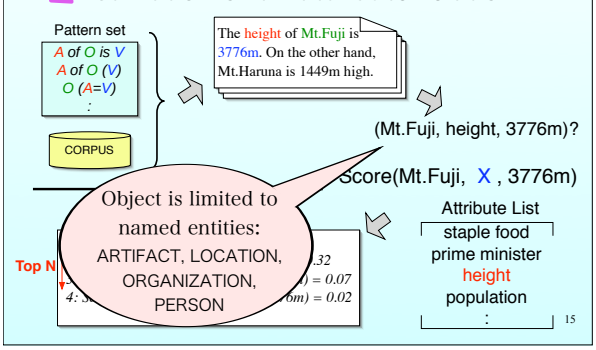
Overview

- Estimation of an attribute relation



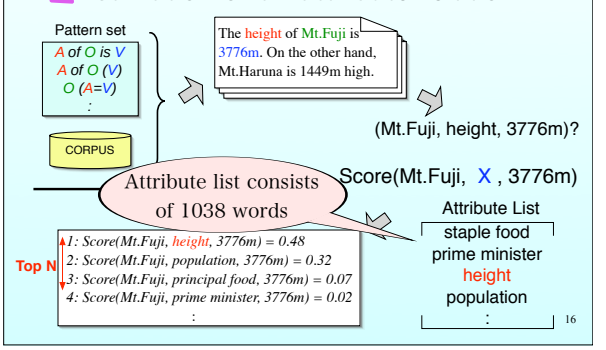
Overview

- Estimation of an attribute relation



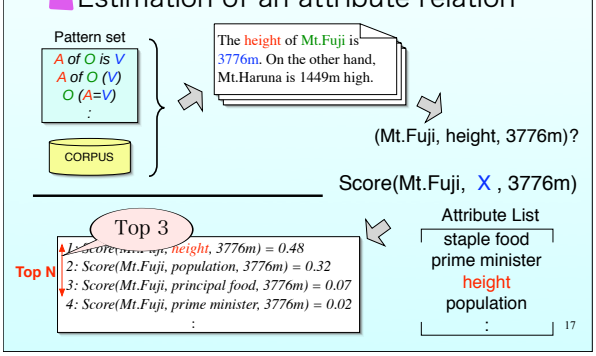
Overview

- Estimation of an attribute relation



Overview

- Estimation of an attribute relation



Evaluation

- comparison
  - without filtering
    (only pattern matching)
    ⇕
  - with filtering
- Evaluation Measure
  - Precision
  - Recall
  - F-measure

Evaluation Results

- Candidates: 19,620
  - Analysis of 500 samples
    - Attribute relations : 216

	w/o Filtering	w/ Filtering
Precision	216/500 (0.43)	210/417 (0.50)
Recall	216/216 (1.00)	210/216 (0.97)
F-measure	0.60	0.66

Improvement of 6% in F-measure

Evaluation Results

- Examples of Attribute estimation

「米朝は朝鮮労働党の金容淳書記が一月訪米，米．．」

1	書記	2.89e-08	6	担当	1.62e-10
2	団長	2.89e-10	7	問題	1.25e-10
3	総書記	2.79e-10	8	代表団	1.06e-10
4	委員長	2.49e-10	9	会議	6.53e-11
5	主席	1.73e-10	10	部長	3.09e-11

object
attribute
value

「ソ連の在外資産・に関する協定．．」

1	問題	2.74e-10	6	大統領	3.58e-11
2	会議	1.34e-10	7	長官	1.86e-11
3	国家	8.39e-11	8	代表	1.06e-11
4	外相	4.30e-11	9	首相	9.48e-12
5	大使	3.93e-11	10	政策	9.29e-12

Evaluation Results

- Almost all candidates are estimated to be in the top 10 attribute candidates.
- The score has high recall.

Correct attributes' ranks

rank	1	2	3	4	7	9	-	Sum
#cand.	164	39	7	2	1	2	1	216

Future work

- Strict cooccurrence measure
  cooccurrence within a sentence
  → cooccurrence in patterns or dependency pairs.
- Smoothing in scoring
  - Introduce a word class based cooccurrence measure into the calculation of the statistical measure.

Summary

- Set a task of extracting triplets of (object, attribute, value)
- Proposed an approach which combines patterns with a statistical measure.
- Improved performance with a 6% increase in the F-measure.