Extraction of important numerical pairs from text documents and visualization of them

Masaki Murata (NICT), Koji Ichii (Hiroshima Univ.), Qing Ma (Ryukoku Univ.), Tamotsu Shirado, Toshiyuki Kanamaru, Sachiyo Tsukawaki and Hitoshi Isahara (NICT)

1. Introduction

Construction of a system to automatically extract numerical pairs from documents related to a certain topic and display them in graphs.

Graphs are easy to recognize and useful for easily understanding information described in documents.

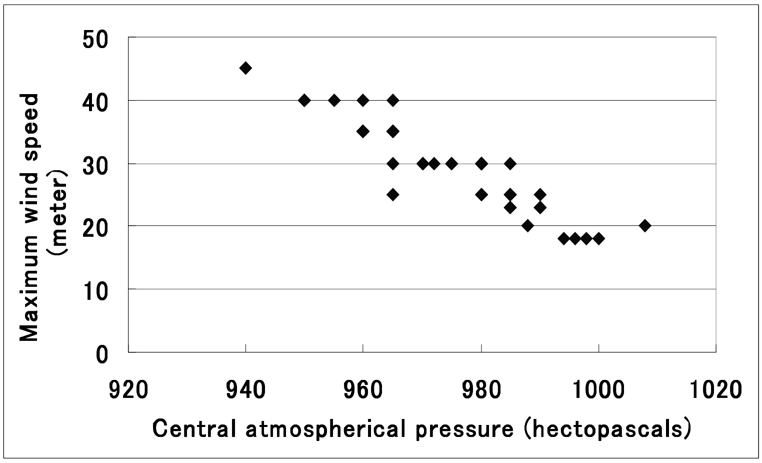
2. Our system's overview

This system first extracts two units and one item expression from documents, it then extracts numerical pairs from sentences including the two units and the item expression; finally, it arranges the pairs and displays them in graphs.

(Ex.) When documents on typhoon are given, the system extracts "hectopascal" and "m/s" as units and "maximum wind speed" as an item expression.

It next extracts values related to "hectopascal" and "m/s".

Finally, it produces a graph where the value related to "hectopascal" is used for the horizontal axis, and the value related to "m/s" is used for the vertical axis.



3. Structure of our system

3.1 Component to extract important expressions

The system extracts important expressions that will be used to extract numerical pairs.

Important expressions belong to one of two categories:

- item units
- item expressions

ChaSen, a Japanese morphological analyzer, is used to extract expressions. Parts of speech in ChaSen output are used.

The system extracts a sequence of nouns adjacent to numerical values as item units.

The system extracts a sequence of nouns as item expressions.

The system next extracts important item units and item expressions that play important roles in the target documents.

Our system used an equation to extract important expressions.

Equation for the TF numerical term in Okapi

$$Score(t) = \sum_{i \in Docs} \frac{TF_i(t)}{TF_i(t) + \frac{l_i}{\Delta}} \tag{1}$$

t: candidate of an important expression. i: ID of a document that includes expression t $Docs$: a set of document IDs. $TF_i(t)$: occurrence number of expression t in document i. l: length of document i. Δ: average document length.

Using the equation, the system calculates the value of each expression.

Expressions with a high value are determined to be important expressions.

3.2 Component to extract numerical pairs

The system identifies the locations in sentences where two item units and an item expression (extracted by the first component) appear in the same sentence and extracts sets of important expressions described in the sentences as a numerical pair.

3.3 Component to extract and display important numerical pairs

The system gathers the extracted numerical pairs and displays them in graphs.

In the graphs, values related to unit items are used for the horizontal and vertical axes.

The system extracts five multiple item units and five item expressions (through the component to extract important expressions) and uses these to make all possible combinations containing two different item units and one item expression.

The system makes a graph for each combination and calculates the number of plots for each graph.

The system judges that a graph having more plots represents a more useful one.

4. Experiments

4.1 Experiments on extracting important expressions

(1) Typhoons: (including "typhoon" and "maximum wind speed")

Typhoon	
Item units	**Item expressions**
gou (No.)	*taihuu* (typhoon)
me-toru (meter)	*saidai huusoku* (maximum wind speed)
kiro (kilo)	*chushinhukin* (near the center)
hekutopasukaru (hectopascal)	*kishouchou* (Japan Meteorological Agency)
miri (mili)	*jisoku*(speed per hour)

"Typhoon" and "maximum wind speed" were extracted as important item expressions

"*gou*" (no.), "*meetoru*" (meter), "*kiro*" (kilo), and "*hekutopasukaru*" (hectopascal), which are related to typhoons, were extracted as important item units.

(2) Movies: (including "movie" and "box-office profits")

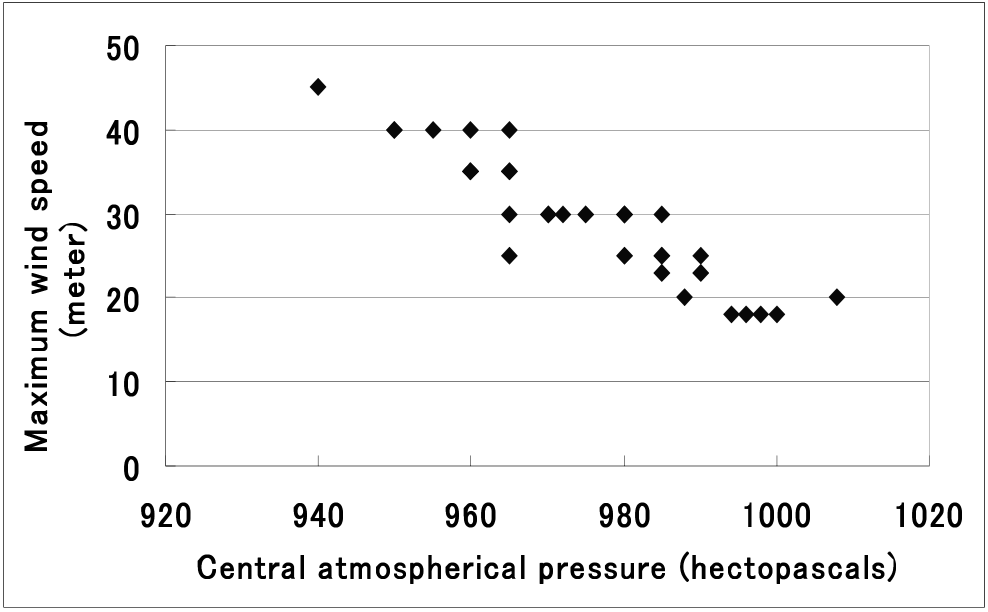
Movies	
Item units	**Item expressions**
en (yen)	*eiga* (movies)
nin (people)	*kougyou shuunyuu* (box-office profits)
doru (dollar)	*sakuhin* (works)
sai (age)	*chihiro* (Chihiro (heroine's name))
hon (piece(s))	*kamikakushi* (Spirited Away)

"Box-office profits", which is strongly related to movies, was extracted as an important item expression

"*en*" (yen) and "*nin*" (unit for number of people) were extracted as important item units.

4.2 Experiments on graphs representing numerical pairs

(1) Graph for the typhoon data set:

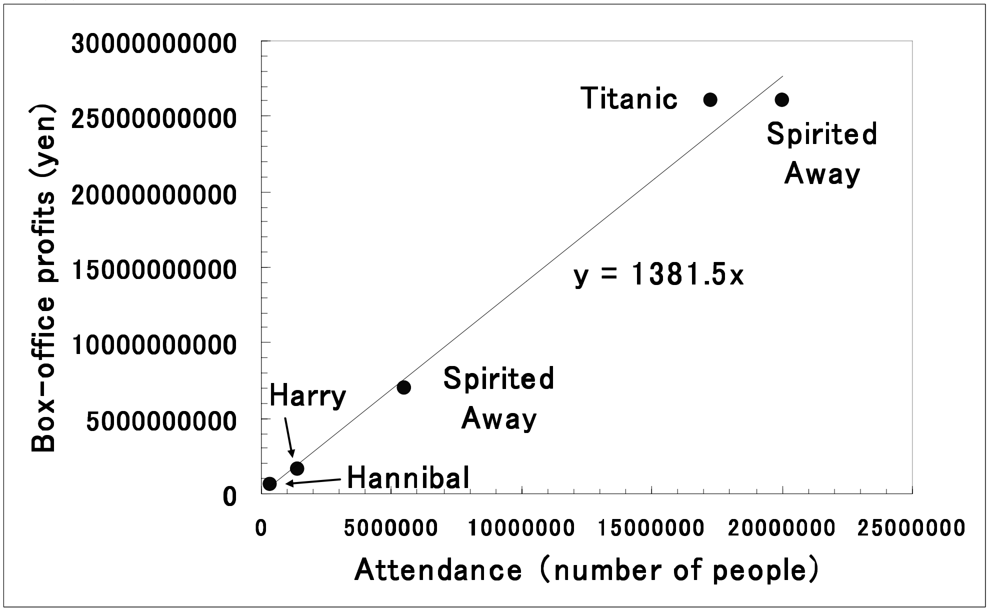


Important expressions

- *me-toru* (meter) — item unit
- *hekutopasukaru* (hectopascal) — item unit
- *saidai huusoku* (maximum wind speed) — item expression

When the pressure was lower, the wind speed was higher. Even at the same pressure, different speeds can occur.

(2) Graph for the movie data set:



Important expressions

- *en* (yen) — item unit
- *nin* (number of people) — item unit
- *kougyou shuunyuu* (box-office profits) — item expression

Expressions for labels corresponding to each point on graph were automatically extracted by extracting expressions surrounded by quotation marks in the documents.

A single regression line was calculated for the points.

Equation for the line indicates that each person pays about 1,400 yen.

Because “Titanic” is above the line and “Spirited Away” (*sen to chihiro no kamikakushi*) is below the line, people pay more than average for “Titanic” (targeted at adults) and less than average for “Spirited Away” (targeted at children).

The two points for “Spirited Away” are due to the system extracting two numerical pairs on different dates.

6. Evaluation

18 document sets (Mainichi newspaper (2000 and 2001))

(not used to construct the system.)

Results

	Eval A	Eval B
• TP1	0.22	0.75
TP5	0.39	0.75

- Eval A — 75% or more of the points on a graph were correct
  Eval B — 50% or more of the points on a graph were correct

Our system is convenient and effective because it can output a graph that includes numerical pairs at these levels of accuracy when given only a set of documents as input.