Trend Information Extraction based on Relative Expression participated on MuST T2N Subtask

Yasuhiro Uenishi Fumito Masui Tatsuaki Matsuba Atsuo Kawai Naoki Isu
Mie University

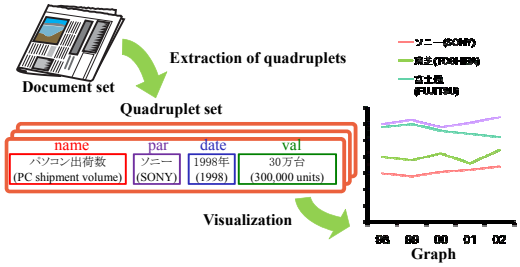
Abstract

This paper describes a system participating in the MuST T2N subtask. To the participating system, we applied the method of implicit trend information extraction utilizing relative expressions such as ``0.1%増(grew 0.1%)", ``前年(previous year)" and ``過去最高(maximum)". Relative differences and numerical changes in trend information can be signified by relative expressions. The system extracts elements of four types by pattern-based rules considering the relative expression. The extracted element is compared with the query word by identifying the synonym of the elements utilizing an EDR dictionary and some synonym databases.

Some experiments were conducted with the MuST T2N formal run test collection. Although the results showed precision of 0.220 and recall of 0.029 totally, the outcomes of additional evaluations suggested the fundamental process performs effectively.

Basic Elements (Quadruplet)

➢On the basis of the MuST definition, we defined a **quadruplet**, which is a set of four basic elements: **name**, **par**, **date** and **val**. In MuST corpus, the four elements are tagged.



Relative Expression and Quadruplet

Document

1999年のパソコン出荷台数は前年比36.7%増の1083万台となった。
(In 1999, PC shipment volume grew 36.7% over the previous year to 10.83 million units.)

Relative Expression

Explicit Quadruplet

name	par	date	val
パソコン出荷数 (PC shipment volume)	Φ	1999年 (1999)	1083万台 (10.83 million units)

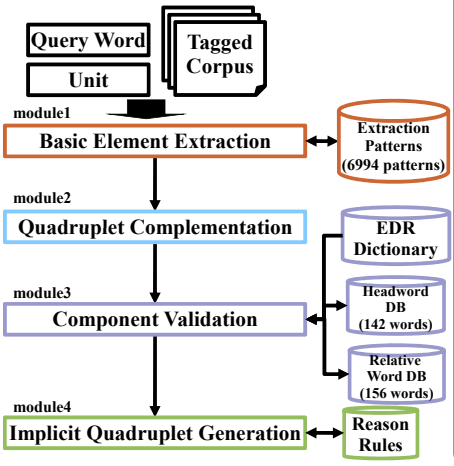
Reasoning the other quadruplet

Implicit Quadruplet

name	par	date	val
パソコン出荷数 (PC shipment volume)	Φ	1998年 (1998)	792万台 (7.92 million units)

System Overview

System Architecture



◆Basic Element Extraction

➢Utilizing extraction patterns, basic elements for quadruplets and a relative expression are extracted from tagged corpus.

Tagged Corpus

日本電子工業振興協会が<date>15日</date>、<date>2006年</date>のパソコン国内実績を発表した。<name>パソコン出荷数</name>は<date>前年</date>比<rel>1.4%</rel>減の<val>1326万台</val>だった。
(On <date>9th</date>, JEIDA reported the PC shipments in <date>2006</date>. <name>The PC shipment volume</name> declined <rel>1.4%</rel> from <date>the previous year</date> to <val>13.26 million units</val>.

Extraction Pattern

<name> は <date> 比 <rel> 減の <val>
(<name> declined <rel> from <date> to <val>)

Quadruplet

name	par	date	val	Relative Expression
パソコン出荷数 (PC shipment volume)			1326万台 (13.26 million units)	前年比1.4%減 (declined 1.4% from the previous year)

◆Quadruplet Complementation

➢The elements which can be extracted using the extraction patterns are complemented.

●Complementation Rules

■**name** : The nearest name element ahead of a part matched with the extraction pattern

■**par** : The nearest par element ahead of a part matched with the extraction pattern from a sentence

■**date** : (1) The nearest date element ahead of a part matched with the extraction pattern in a sentence
(2) The date element which is close to the head of an article
(3) The date when the newspaper article was written

Quadruplet

name	par	date	val	Relative Expression
パソコン出荷数 (PC shipment volume)	Φ	2006年 (2006)	1326万台 (13.26 million units)	前年比1.4%減 (declined 1.4% from the previous year)

◆Component Validation

➢The quadruplets that relate to the query word are selected. To select the quadruplets related to a query word, components of the name element and the query word are validated.

➢We assumed that both a name element and a query word also consist of a headword (a unit of trend) and a specifier (a subject of trend).

➢If the checking components (specifier and headword) succeeds, the name element relates to the query word. Headword is checked using headword database. Specifier is checked using EDR dictionary and relative word database.

➢In the following cases, the element-derived specifier is relevant to the query-derived specifier. If the name element is identical to the headword, the par element becomes a specifier.

- × **Both specifiers are identical concept. (Ex1 and Ex2)**
- × **The element-derived specifier is the hyponym of the query-derived specifier. (EX3)**
- × **The element-derived specifier is the concept relevant to the query-derived specifier. (Ex4)**

If the name element consists of only a headword and a quadruplet does not include the par element, the specifier is checked based on co-occurrence in a sentence.

- × **The query-derived specifier is found ahead of a part matched with the extraction pattern in a sentence. (Ex5)**

Component of Name Element and Query Word

Name Element or Query Word	Specifier	Headword
パソコン出荷台数 (PC shipment volume)	パソコン (PC)	出荷台数 (shipment volume)
ビール出荷数量 (The quantity of beer shipped)	ビール (beer)	出荷数量 (The shipped quantity)
政党支持率 (Approval rating for political parties)	政党 (Political parties)	支持率 (Approval rating)

The Examples which Checking Components Succeeds

Ex1.
query: パソコン出荷台数 (PC shipment volume)
name: パーソナルコンピュータの出荷台数 (Personal computer shipment volume)

Ex2.
query: デジタルカメラ出荷台数 (Digital camera shipment volume)
name: デジカメの出荷台数 (Digital camera shipment volume)

Ex3.
query: 政党支持率 (Approval rating for political parties)
name: 支持率 (Approval rating)
par: 自民党 (LDP)

Ex4.
query: パソコンの出荷台数 (shipment volume of PC)
name: 出荷台数 (shipment volume)
par: NEC

Ex5.
query: パソコンの出荷台数 (shipment volume of PC)
name: 出荷台数 (shipment volume)
sentence: パソコン出荷金額は前年比1.9%減の1兆7360億円で、出荷台数は同4.2%増の1249万台となった。(PC shipment value declined 1.9% from the previous year to 1.736 trillion yen and shipment volume grew 4.2% over the year to 12.49 million units.)

◆Implicit Quadruplet Generation

➢Applying an explicit quadruplet to the reason rules, other implicit quadruplet is generated. **Evaluation Subject of T2N Subtask**

Explicit Quadruplet

name	par	date	val
パソコン出荷数 (PC shipment volume)	Φ	2006年 (2006)	1326万台 (13.26 million units)

Relative Expression: 前年比1.4%減 (declined 1.4% from the previous year)

Applying explicit elements to the reason rules

$$\text{date}_{imp} = 2006\text{年}(2006\ \text{year}) - 1\text{年}(1\ \text{year}) = 2005\text{年}(2005\ \text{year})$$

$$\text{val}_{imp} = \frac{1326\text{万台}(13.26\ \text{million units})}{(1-\frac{1.4}{100})} = 1345\text{万台}(13.45\ \text{million units})$$

Implicit Quadruplet

name	par	date	val
パソコン出荷数 (PC shipment volume)	Φ	2005年 (2005)	1345万台 (13.45 million units)

Experiments

■ **Evaluation:** We evaluated the following performance:
- the performance of our system for the T2N formal run(Table 1)
- the performance of our system for relative expression(Table 2)
- the performance of each module in our system(Table 3)

■ **Data set:** Mainichi newspaper annotated with XML tag from 1998 to 2001 (120 articles)

■ **Query word:** 25 statistics names

■ **Correct data:** 314 of the pairs are contained in the data set.133 of the pairs for relative expression are contained in the data set.

■ **Output:** Our system extracted 30 of the correct pairs.

■ **Measure:** Precision, Recall and F-measure

Table 1. T2N Formal Run Evaluation Results

	Precision	Recall	F-measure
micro-ave.	0.220	0.029	0.051
macro-ave.	0.093	0.021	0.031

Table 2. Evaluation Results for Relative Expressions

	Precision	Recall	F-measure
system1	0.220	0.068	0.108
system2	0.638	0.226	0.333

system1: the system not implementing all functions
system2: the system implementing all functions

Table 3. Detailed Evaluation Results of System2

	Precision	Recall	F-measure
module1	1.000	0.669	0.802
module2	0.718	-	-
module3	0.882	0.455	0.600

module1: Basic element extraction module2: Quadruplet complementation
module3: Component validation

Discussion

There is remarkable difference between the micro-average and macro average in precision in Table 1. The result indicates that the system still has uneven performance.

The result indicates that basic elements are not sufficiently extracted in Table 2.

module1:
In the topics ``digital camera" and ``communication device", the many specific expressions out of the training data appeared in the data set. In the topic of ``gasoline", there is no regular appearance of relative expressions.

module2:
The failure of the complementing of the date element is remarkable in particular. An excessively strict condition for extracting the date element caused the failure of the complement.

module3:
Mainly, identification was unsuccessful in the case where the name element is verbose.