Extraction of Trend Information from Newspaper Articles: Hiroshima City University at NTCIR-7 MuST

Hidetsugu Nanba
Hiroshima City University
3-4-1 Ozukahigashi, Hiroshima 731-3194 JAPAN
Phone: +81-82-830-1584
nanba@hiroshima-cu.ac.jp

Abstract

Trend information is a summarization of temporal statistical data, such as changes in product prices and sales. We propose a method for extracting trend information from multiple newspaper articles. Our group participated in the T2N Subtask at NTCIR-7 MuST (*Multimodal Summarization for Trend Information*)*. Our goal was to evaluate the effectiveness of our rule-based system using the data provided by NTCIR-7 MuST. From the experimental results, we obtained precision of 62.9% and recall of 7.0%.*
Keywords: *Trend Information, Visualization, Newspaper article.*

1 Introduction

Trend information is a kind of summarization of temporal statistical data, such as changes in product prices and sales. We have constructed a system that extracts trend information from multiple newspaper articles and blogs [Nanba et al. 2007]. We modified this system to improve its performance in trend information extraction, and submitted four results to the T2N Subtask at NTCIR-7 MuST (Multimodal Summarization for Trend Information) [Kato and Matsushita 2008]. In this paper, we report and discuss the evaluation results in this subtask.

The remainder of this paper is organized as follows. Section 2 explains how we extracted trend information from newspaper articles. Section 3 reports on the experiments, and discusses the results. We present some conclusions in Section 4.

2 Extraction of Trend Information

To extract trend information from newspaper articles, we modified our previous system [Nanba et al. 2007]. We explain the system in Section 2.1, and we point out its problems and their improvement in Section 2.2.

2.1 Extraction of Trend Information from Newspaper Articles and Blogs

Nanba et al. [2007] constructed a system that extracts trend information from newspaper articles and blogs. The system consists of two modules: (1) a module for temporal expressions extraction and (2) a module for statistical values extraction. In the following, we describe both modules.

2.1.1 Module for Temporal Expressions Extraction

The module used CaboCha[1] [Kudo and Matsumoto 2003] for extracting temporal expressions from texts. CaboCha is a statistical syntactic parser for Japanese texts, and also identifies eight kinds of named entity, such as date, organization, and location, in texts. Here, not all temporal expressions in texts are used to generate a graph. For example, there are three temporal expressions in the sentence in Figure 1, but "1994" is the only expression that is used in generating a graph of "the number of births for one year".

一九九四年の年間出生数は前年より四万七千人も多い百二十三万五千人を記録し，二十一年ぶりに大幅増に転じた．
(The number of births recorded in 1994 was 1,235,000, which was 47,000 greater than in the previous year, and this significant increase was the first in 21 years.)

Figure 1. Example of a sentence that contains more than one temporal expression

To eliminate unrelated temporal expressions from texts, we focused on some cue phrases, some of which are shown in Figure 2. In the figure, cue phrases are underlined. If these cue phrases appear before or after temporal expressions, they are eliminated.

前年より (in comparison with the previous year)
10 年ぶり (for the first time in 10 years)
昨年以来 (since last year)
3 年連続で (for the third year in a row)

Figure 2. Examples of cue phrases to eliminate unrelated temporal expressions

In the next stage, the extracted temporal expressions in the prior stage are converted into a specific format "<DATE>YYYY-MM-DD</DATE>", where YYYY, MM, and DD indicate year, month, and day, respectively. For example, if "昨日" (yesterday) is extracted from a text written on January 25, 2007, then the expression "yesterday" is replaced by "<DATE>2007-01-24</DATE>". If the exact date is not provided, such as "昨年 12 月" (last December), this expression is converted as "<DATE>2006-12-??</DATE>".

[1] http://chasen.org/~taku/software/cabocha/

2.1.2 Module for Statistical Values Extraction

The module extracts statistical values in the following four steps.

1. Split a sentence into separate statistical values.
2. Annotate "NUM" and "UNIT" tags.
3. Eliminate unrelated statistical values.
4. Extract statistical values.

(Step 1) Split a sentence into separate statistical values

Some sentences contain more than one statistical value; for example:

今日のレギュラーガソリンは１３０円，ハイオクは１５０円だった．
(Today's price of regular gasoline was 130 yen, and that of premium gasoline was 150 yen.)

From this sentence, "150 yen" might be extracted mistakenly as the price of regular gasoline, or "130 yen" as that of premium gasoline. We therefore split sentences for their statistical values. First, we analyse the dependency structure of the sentence using CaboCha. Second, we integrate two *bunsetsus* (segments) that have a modification relation and are adjacent to each other. We explain this process using the example shown in Figure 3. This figure is the result from CaboCha for the sentence given above. "Chunk" tags with ID numbers are assigned to each *bunsetsu*. Here, the attribute values in each chunk tag indicate the ID numbers of *bunsetsus* that have a modification relation. *Bunsetsus* 0 and 1 have a modification relation, and are adjacent to each other. Therefore, these *bunsetsus* are integrated. By repeating the process to the end of the sentence, it is split into two parts: (1) "Today's price of regular gasoline was 130 yen", and (2) "and that of premium gasoline was 150 yen". In this step, the system also adds an "NP" tag to all noun phrases in the text, based on the result from CaboCha.

<chunk id="0" link="1"><NP>今日</NP>の (Today's)</chunk>
<chunk id="1" link="2"><NP>レギュラーガソリン</NP>は(price of regular gasoline was)</chunk>
<chunk id="2" link="4">130 円，(130 yen,)</chunk>
<chunk id="3" link="4"><NP>ハイオク</NP>は(and that of premium gasoline was)</chunk>
<chunk id="4" link="–1">150 円だった．(150 yen.)</chunk>

⇩

<chunk><NP>今日</NP>の<NP>レギュラーガソリン</NP>は 130 円，(Today's regular gasoline was 130 yen,)</chunk>
<chunk><NP>ハイオク</NP>は 150 円だった．(and premium gasoline was 150 yen.)</chunk>

Figure 3. Splitting a sentence for statistical values

(Step 2) Annotate "NUM" and "UNIT" tags

The system adds "NUM" and "UNIT" tags to all candidate statistical values using the pattern "number + (noun phrase | postfix | counter suffix)". For example, NUM and UNIT tags are annotated for an expression "150 円" (yen) as "<NUM>150</NUM><UNIT>円</UNIT>".

(Step 3) Eliminate unrelated statistical values

Not all expressions assigned "NUM" and "UNIT" tags are used to generate a graph of trend information. For example, there are two statistical values in the sentence in Figure 4, but "９２円" (92 yen) is the only expression that is used to generate a graph.

石油情報センターが２３日発表した給油所石油製品市況調査によると，６月のガソリン価格は全国平均でレギュラー１リットル当たり<NUM>９２</NUM><UNIT>円</UNIT>となり，前月比で<NUM>２</NUM><UNIT>円</UNIT>上昇した．
(According to the Oil Information Centre's survey of market conditions for products sold through service stations on (June) 23, the price of gasoline reached a national average of <NUM>92</NUM><UNIT>yen</UNIT> per litre, regular; <NUM>2</NUM><UNIT>yen</UNIT> higher than the average price of last month)
(June 24, 1999, *Mainichi* newspaper article)

Figure 4. Example of an analysis of a newspaper article

In the same way as extraction of temporal expressions, unrelated statistical values are eliminated from texts using cue phrases, some of which are shown in Figure 5. Here, cue phrases are shown as underlines. If they appear before or after statistical values, they are eliminated.

2 円高い (2 yen higher)
昨年比 45.2%上昇 (up 45.2% from the previous year)
0.28 パーセント下落 (down 0.28 percent)

Figure 5. Example of cue phrases to eliminate unrelated statistical values

(Step 4) Extract statistical values

Pairs of statistical values and their nearest anterior temporal expressions are extracted[2] when one of the keywords (the name of the statistic) and its associated unit appear in the same integrated *bunsetsu*, and the keyword is a prefix of one of the noun phrases in the *bunsetsu*. For example, "130 yen" was extracted from the sentence in Figure 3, when "レギュラー" (regular) and "円" (yen) were given to the system in advance, because "レギュラー (regular)" is a prefix of

[2] Nanba et al. [2007] reported that adequate temporal expressions corresponding to each statistical value were correctly extracted in all cases.

a noun phrase "レギュラーガソリン (regular gasoline)". We call this method the "prefix-search method".

2.2 Problems of our Previous System and their Improvement

2.2.1 Problem with the Temporal Expressions Extraction Module

Most of the errors in our previous system were caused by misconversion of temporal expressions, such as "前月" (the previous month), into "YYYY-MM-DD" form, although temporal expressions were detected correctly in most cases. We therefore employed another method, in addition to the method in Section 2.1.1. The method considers the date that the article was written as the temporal information corresponding to all the statistical values in the article. We call this method the "publication-date method".

2.2.2 Problem with the Statistical Values Extraction Module

Another typical error in our previous system was caused by the fourth step in statistical values extraction. Although the prefix-search method obtained a high precision score, many statistical values were not extracted. We therefore propose another approach, called the "DB-matching method". This method extracts statistical values when a keyword and a unit appear in the same integrated bunsetsu, and similarity between the keyword and one of noun phrases exceeds a threshold value. To calculate the similarity, we employed dynamic programming (also known as "edit distance") [Needleman and Wunsch 1970], and used 0.9 as the threshold value, which was determined by preliminary experimental results using Nanba's data [Nanba et al. 2007].

3 Evaluation

3.1 Submission

We submitted four systems: "HCU1", "HCU2", "HCU3", and "HCU4" to the formal run of the T2N Subtask at NTCIR-7 MuST [Kato and Matsushita 2008]. The differences between these systems are summarized in Table 1.

Table 1. Summary of submitted systems

	Statistical values extraction		Temporal expressions extraction	
	prefix search	DP-matching	rule-based	publication date
HCU1	○		○	
HCU2	○			○
HCU3		○	○	
HCU4		○		○

3.2 Data and Evaluation

We used the data provided by the MuST organizers for evaluation. They consist of eight topics including 120 newspaper articles in total. All the submitted systems were evaluated in terms of precision, recall, and F-measure [Kato and Matsushita 2008].

3.3 Results

We show the evaluation results for each system in Table 2. The results show that HCU1 and HCU2 obtained higher precision scores than HCU3 and HCU4. Both HCU1 and HCU2 employed the prefix-search method to extract of statistical values.

3.4 Discussion

"Rule-based method" vs. "publication-date method"

To compare the methods for the extraction of temporal expressions, we count the number of errors for each system. The results are shown in Table 3. The number of errors for HCU2 and HCU4 are smaller than those for HCU1 and HCU3, which indicates that the publication-date method is superior to the rule-based method.

Table 3. The number of errors for the extraction of temporal expressions

HCU1	HCU2	HCU3	HCU4
12	3	36	23

"Prefix-search method" vs. "DP-matching method"

Generally, the numbers of statistics extracted by HCU3 and HCU4 are much larger than those extracted by HCU1 and HCU2. We compared recalls of HCU1 and HCU3, both of which employed the same method (rule based) for the extraction of temporal expressions. The recall of HCU3 is 7.7% better than that of HCU1. We also compared recalls of HCU2 and HCU4. The recall of HCU4 is 8.9% greater than that of HCU2. However, the precisions of both HCU3 and HCU4 are smaller than those of HCU1 and HCU2. From these comparisons, we can conclude that the DP-matching method contributes to improve recall, while it impairs precision.

Evaluation results for each topic

Precisions of HCU1 and HCU2 for the topics "Communication Device (T2N0103)", "I-mode (T2N0106)", and "Digital Camera (T2N0107)" are much better than those of HCU3 and HCU4. The common feature of these topics is that two or more kinds of statistical values are mentioned in each article. For example, the number of "Personal Handy-phone System (PHS)" subscribers is compared with those of "cellular phone" and "land phone" in the "Communication Device" topic. In such cases, names of statistics are often shortened. Figure 6 is a typical example in which the DP-matching method could not extract statistical values correctly. In this article, three

statistics, "the number of cellular phone and PHS subscribers", "the number of cellular phone subscribers", and "the number of PHS subscribers", were mentioned. However, the latter two are expressed as "cellular phone" and "PHS". In such cases, the DP-matching method could not contribute to improve recall, but rather impaired precision.

総務省が１０日発表した４月の携帯電話とＰＨＳの加入者数は、前月比１・６％、１０９万３０００人増の６７８７万７０００人となった。
(略)
携帯電話が１・８％増の６２０３万７０００台、ＰＨＳが０・０３％減の５８４万台。

According to the report by the Ministry of Internal Affairs and Communications on May 10, the number of cellular phone and PHS subscribers became 67,877,000, which is 1.6% (1,093,000) larger than in the previous month.
(snip)
Cellular phone is 62,037,000 (1.8% up), while PHS is 5,840,000 (0.03% down).

Figure 6. Example that the DP-matching method could not extract correctly

To improve the low precision by the DP-matching method, the following procedure may be useful.

1. When more than one statistic is mentioned in a newspaper article,
   Apply the prefix-search method.

2. In other cases,
   Apply the DP-matching method.

4 Conclusions

We have proposed a method for extracting trend information from multiple newspaper articles. To confirm the effectiveness of our system, we participated in the T2N Subtask at NTCIR-7 MuST, and submitted several results provided by four different systems. From the experimental results, we obtained precision of 62.9% and recall of 7.0% with the prefix search method for statistical values extraction and the publication date method for temporal expressions extraction.

5 Acknowledgements

The authors would like to express their gratitude to the organizers of the MuST.

References

T. Kato and M. Matsushita. Overview of MuST at the NTCIR-7 Workshop - Challenges to Multi-modal Summarization for Trend Information - . In Proceedings of the 7th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access, 2008.

T. Kudo and Y. Matsumoto. Fast Methods for Kernel-based Text Analysis. In Proceedings of the 41st Annual Meeting on Association for Computational Linguistics, pp.24-31, 2003.

H. Nanba, N. Okuda, and M. Okumura. Extraction and Visualization of Trend Information from Newspaper Articles and Blogs. In Proceedings of the 6th NTCIR Workshop, pp.414-419, 2007.

S. B. Needleman and C.D. Wunsch. A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins. Journal of Molecular Biology, Vol. 48, pp.443-453, 1970.

Table 2. Evaluation results for each system

Topic	System	Precision	Recall	F-measure
Gasoline (T2N0101)	HCU1	0.13 (1/8)	0.03(1/33)	0.05
	HCU2	0.38 (3/8)	0.09 (3/33)	0.15
	HCU3	0.10 (1/10)	0.03 (1/33)	0.05
	HCU4	0.30 (3/10)	0.09 (3/33)	0.14
Personal Computer (T2N0102)	HCU1	0.00 (0/0)	0.00 (0/19)	0.00
	HCU2	0.00 (0/0)	0.00 (0/19)	0.00
	HCU3	0.00 (0/3)	0.00 (0/19)	0.00
	HCU4	0.33 (3/9)	0.16 (3/19)	0.22
Communication Device (T2N0103)	HCU1	0.36 (4/11)	0.08 (4/50)	0.13
	HCU2	0.62 (8/13)	0.16 (8/50)	0.25
	HCU3	0.24 (5/21)	0.10 (5/50)	0.14
	HCU4	0.27 (4/15)	0.08 (4/50)	0.12
Political Trend (T2N0104)	HCU1	0.00 (0/0)	0.00 (0/85)	0.00
	HCU2	0.00 (0/0)	0.00 (0/85)	0.00
	HCU3	0.12 (22/181)	0.26 (22/85)	0.16
	HCU4	0.13 (24/181)	0.28 (24/85)	0.18
Industry (T2N0105)	HCU1	0.00 (0/0)	0.00 (0/27)	0.00
	HCU2	0.00 (0/0)	0.00 (0/27)	0.00
	HCU3	0.00 (0/0)	0.00 (0/27)	0.00
	HCU4	0.00 (0/0)	0.00 (0/27)	0.00
I-mode (T2N0106)	HCU1	0.25 (1/4)	0.02 (1/45)	0.04
	HCU2	0.75 (3/4)	0.07 (3/45)	0.13
	HCU3	0.06 (1/17)	0.02 (1/45)	0.03
	HCU4	0.24 (4/17)	0.08 (4/45)	0.12
Digital Camera (T2N0107)	HCU1	1.00 (1/1)	0.06 (1/18)	0.11
	HCU2	1.00 (1/1)	0.06 (1/18)	0.11
	HCU3	0.33 (2/6)	0.11 (2/18)	0.17
	HCU4	0.50 (3/6)	0.17 (3/18)	0.25
Center Exam (T2N0108)	HCU1	0.67 (6/9)	0.16 (6/37)	0.26
	HCU2	0.78 (7/9)	0.19 (7/37)	0.31
	HCU3	0.38 (6/16)	0.16 (6/37)	0.23
	HCU4	0.63 (10/16)	0.27 (10/37)	0.38
Total	HCU1	0.371	0.041	0.074
	HCU2	0.629	0.070	0.126
	HCU3	0.146	0.118	0.130
	HCU4	0.197	0.159	0.176