Machine Translation of Sentences with Fixed Expressions

Naoto Katoh[1] **Teruaki Aizawa**
NHK Science and Technical Research Laboratories
1-10-11, Kinuta, Setagaya-ku,
Tokyo 157, Japan
{katonao, aizawa}@strl.nhk.or.jp

Abstract

This paper presents a practical machine translation system based on sentence types for economic news stories.

Conventional English-to-Japanese machine translation (MT) systems which are rule-based approaches, are difficult to translate certain types of Associated Press (AP) wire service news stories, such as economics and sports, because these topics include many fixed expressions (such as compound words or collocations) which are difficult to be processed by conventional syntactic analysis and/or word selection methods.

The proposed MT system, an economic-news stories machine translation system (**ENTS**), can translate economic news sentences with fixed expressions. The system consists of three processes, to handle different types of sentences, fixed type, economics-specific type and general type. This paper focuses mainly on the translation method for fixed-type sentences, which is a kind of example-based approach. In this translation method, fixed sentence translation (**STRA**) data plays a key role. The STRA data is a set of bilingual templates, which is built automatically from fixed English sentences and their Japanese translation equivalents. The fixed English sentences are extracted automatically from the AP corpus.

A series of experiments to evaluate ENTS using economic news in the AP news stories showed the translation accuracy was about 50 % higher than with our conventional rule-based MT method.

1 Introduction

We are developing an English-to-Japanese machine translation (MT) system to produce real-time rough translations for Associated Press (AP) wire service news stories. With some news topics, troubles with fixed expressions lows the translation accuracy of the MT system. Economic news stories in particular are difficult to translate by conventional rule-based methods, because they contain many fixed expressions sharing two major characteristics:

c1) The fixed expressions produce economics-specific syntactic structure.

c2) Equivalents of the fixed expressions require Japanese economic jargons.

These characteristics respectively cause two major bottlenecks for the conventional rule-based MT system:

b1) General-purpose grammatical rules are not sufficient to yield correct analysis of economic news stories. (Simple addition of grammatical rules increases syntactic ambiguities.)

b2) It is difficult to select the appropriate Japanese words for the translation.

Actually, these problems reduce the translation accuracy of our rule-based MT system to only 20%, which is too low for practical use.

This paper presents a new English-to-Japanese MT system for economic news stories, which is called **ENTS** (Economic News stories machine Translation System), to process fixed expressions effectively. ENTS consists of three sequential processes (as shown in Fig. 1), based on the three basic types of economic news sentence. Process 1 is a kind of example-based approach, while Processes 2 and 3 are rule-based ones that differ in grammatical rules.

This paper focuses mainly on Process 1, which is composed of fixed sentence translation, compound word translation, fixed sentence translation data production and fixed sentence extraction. Fixed sentence translation data (**STRA data**), which is a kind of bilingual template, plays a key role in the fixed sentence translation. The STRA data is built automatically from fixed English sentences extracted from a large corpus and their corresponding Japanese translations.

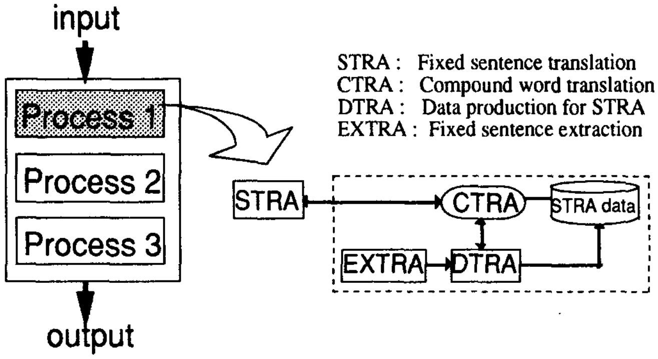
input
Process 1
Process 2
Process 3
output
STRA : Fixed sentence translation
CTRA : Compound word translation
DTRA : Data production for STRA
EXTRA : Fixed sentence extraction
STRA
CTRA
STRA data
EXTRA
DTRA

Figure 1 An overview of ENTS

[1] Currently he is working at ATR Interpreting Telecommunications Research Laboratories.

Recently, several example-based MTs were proposed for processing fixed expressions [Nagao84][Sumita91]. Furuse proposed a cooperative method using tightly woven combination of example- and rule-based approaches [Furuse92]. In contrast to their approach, we use the two methods independently. Therefore, the translation accuracy of our example-based method is guaranteed to be 100%.

Creating an example-based MT requires bilingual translation data. Kaji proposed acquiring the bilingual translation data from bilingual texts [Kaji92]. However that would require a complete syntactic analysis of bilingual texts. Our method is more robust, because it requires only a partial analysis.

Section 2 describes some relevant features of economic news stories. In Section 3, we present an overview of ENTS. The following sections describe the fixed sentence translation method in ENTS, and the results of experiments using ENTS for AP economic news stories.

2 Features of economic news sentences

The AP delivers about 350 wire-service news stories a day, of which about 50 are concerned with economics. Each news story has its own title related to the contents. Because the titles on economic news stories are fixed, such stories can be selected easily. Most sentences in these economic news stories have fixed expressions comprised of compound words and/or collocations.

Example of fixed expressions

e1) **compound words**

"5 cents", "17.76 dollars per kilo",

and "The U.S. dollar"

e2) **collocations**

"Malaysian tin closed at", "The U.S. dollar opened",

and "as share prices rose"

Based on the fixed expressions, the sentences in economic news stories, called economic sentences, are classified into three types:

Type I : Fixed sentences

Example 1

1-1) "In Kuala Lumpur, Malaysian tin closed at 17.76 dollars per kilo, up 5 cents."

1-2) "In Kuala Lumpur, Malaysian tin closed at 16.83 dollars per kilo, up 19 cents."

1-3) "In Kuala Lumpur, Malaysian tin closed at 16.40 dollars per kilo, down 8 cents."

Type II : Economics-specific sentences

Example 2

2-1) "The U.S. dollar opened slightly higher against the Japanese yen Tuesday morning in Tokyo, while share prices inched up."

2-2) "The U.S. dollar drifted lower against the Japanese yen Wednesday morning, while share prices on the Tokyo Stock Exchange rose sharply."

2-3) "The U.S. dollar opened higher against the Japanese yen in Tokyo Thursday, as share prices rose in early trading."

Type III : General sentences

Example 3

3-1) "Kagawa added, however, that the market still anticipates a rising dollar."

3-2) "Shigeru Sato, an analyst with Sanyo Securities, said the index fell some 65 points at one point in the afternoon, but last-minute arbitrage buying pulled it back up."

3-3) "But Tobo said the market's basic sentiment remained bearish because of a lack of incentives to focus on."

(1) **Type I**

The sentences in Type I contain fixed expressions in which the words change a little form day to day. The parts of speech of the translation equivalents of these fixed expressions are nouns in Japanese. For example, the translation equivalents of compound words like "17.76 dollars per kilo" are nouns, as are those of "up" and "down"in 1-1, 2, 3. The verb in a Type I sentence, such as "close" in examples 1-1, 2, 3, is fixed.

(2) **Type II**

Although each sentence in Type II has a unique style with fixed expressions, there is a greater variety of fixed expressions than that in Type I sentences. For example:

"opened" and "drifted"

or "slightly higher", "lower" and "higher"

The parts of speech of their translation equivalents of these fixed expressions are verb or adjective in Japanese. Therefore, their translation equivalences require a production method of their inflections in Japanese generation process of MT.

(3) **Type III**

The Type III sentences have no features that make MT appropriate. Most of the general sentences are dealers' comments.

3 Outline of ENTS

ENTS consists of three translation methods corresponding to the types of economic sentences. ENTS processing follows the flow in Fig. 2.

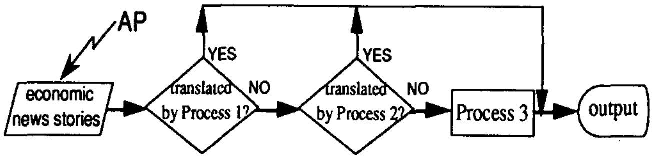

Figure 2 ENTS flow chart

(1) **Process 1**

Process 1 translates fixed sentences (Type I) using bilingual templates that directly handle fixed expressions.

(2) **Process 2**

Process 2 translates sentences of Type II using a conventional rule-based approach with grammatical rules tuned to economic sentences obtained from two data worth of AP stories [Aizawa93]. The grammatical rules are built reflecting features of fixed expressions. These economics-specific grammatical rules total about 500, which is 1/5 of the number of rules for general sentences. Therefore, there are few ambiguities in syntactic structure.

(3) **Process 3**

Process 3 translates those sentences not processed by Process 1 or Process 2. It is a rule-based MT with general-purpose grammatical rules.

4 A translation method of fixed sentence

In our translation method, **STRA** (a fixed Sentence TRAnslation method), the bilingual templates in which translation equivalents of the fixed expressions are represented as variables are created using **STRA data**. That data is built automatically by **DTRA** (a Data production method for STRA) from fixed English sentences and their corresponding Japanese translations. The fixed English sentences are extracted automatically from a corpus by **EXTRA** (a fixed sentence EXTRAction method). **CTRA** (a Compound word TRAnslation method) plays a main role in STRA and DTRA.

Fig. 3 visually summarizes the translation system.

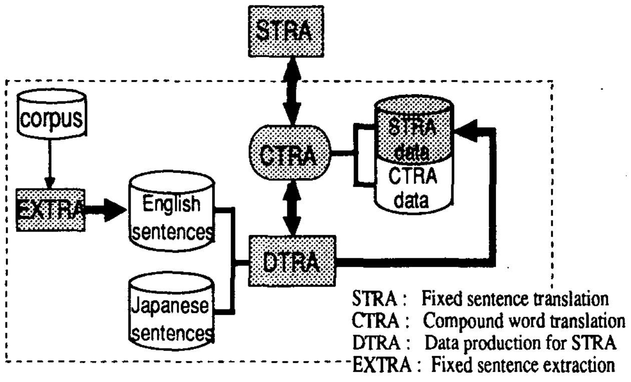
STRA
corpus
CTRA
STRA data
CTRA data
EXTRA
English sentences
Japanese sentences
DTRA
STRA : Fixed sentence translation
CTRA : Compound word translation
DTRA : Data production for STRA
EXTRA : Fixed sentence extraction

Figure 3 Fixed-sentence translation method

4.1 Compound word translation (CTRA)

The compound word translation module (**CTRA**) translates compound words in fixed expressions [Katoh91]. In STRA and DTRA, CTRA is the main processing unit, while it is used in one step of analysis in Processes 2 and 3.

In our MT system used in Processes 2 and 3, the CTRA step occurs between morphological and syntactic analyses as shown in Fig. 4.

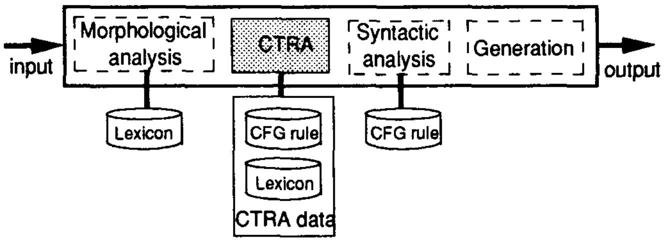

Figure 4 Our rule-based MT system with CTRA

CTRA extracts fixed expressions and defines their appropriate translation equivalents, the parts of speech and the semantic markers. For example, fixed expressions in example 1-1 are processed as:

idiomatic expression	translation equivalents	part of speech	semantic marker
"17.76 dollars per kilo"	"1キロ17.76 ドル"	noun	unit expression
"5 cents"	"5 セント"	noun	unit expression

In CTRA, English analysis is done by CHART parser based on CFG rules which represent fixed expressions. On the other hand, Japanese generation is not based on a rule-based method, but conducted by substituting the translation equivalents of the English words for variables in Japanese templates. Fig. 5 shows examples of CFG rules and their corresponding Japanese templates. Both these CFG rules and Japanese templates are named as **CTRA data**.

English fixed expressions	Japanese templates
1: S --> UNTEXP	「#1#」
2: UNTEXP --> UNTEXP PER UNIT	「1 #3##1#」
3: --> NUMEXP UNIT	「#1##2#」
4: UNIT --> "dollar", "cents", "kilo", "yen", etc	「ドル」,「セント」,「キロ」, 「円」…
5: PER --> "per", "a"	「」
6: NUMEXP --> "1", "12", etc	「1」,「12」…
7: CMA --> ","	「」
8: UPDW --> "up", "down"	「アップ」,「ダウン」
9: CITY --> "Kuala Lumpur", "Tokyo", etc.	「クアラルンプール」,「東京」…

(where part of the #i# denotes the translation equivalent of the ith symbol in the right-hand of CFG rule, and 「(null)」 means the rule has no corresponding translation equivalent.)

Figure 5 Sample CTRA data

4.2 Fixed sentence translation (STRA)

The fixed sentence translation module (**STRA**) is an expanded CTRA with added CTRA data (named as **STRA data**) for translating not only fixed expressions but also fixed sentences. STRA data is produced automatically, as described in next section.

An example of STRA data used to translate example 1-1 in Section 2, are shown in Fig. 6.

English fixed expressions	Japanese templates
1: S --> PAT1 CITY CMA PAT2 UNTEXP CMA UPDW UNTEXP	「#2#でマレーシアのすずは，#8##7#の#5#でひけた」
2: PAT1 --> "In"	「」
3: CITY --> "Kuala Lumpur"	「クアラルンプール」
4: CMA --> ","	「」
5: PAT2 --> "Malaysian tin closed at"	「」
6: UPDW --> "up"	「アップ」
7: UNTEXP --> "17.76 dollars per kilo"	「1 キロ17.76 ドル」
8: --> "5 cents"	「5 セント」

(UNTEXP is obtained by CTRA)

Figure 6 STRA data for example 1-1 in Section 2

At the top of the CFG rules in Fig. 6 is an English template of example 1-1, and its corresponding translation with variables is a Japanese template. The CFG rules are based not on English grammar but on an English sentence pattern, although they represent the word order of a fixed sentence. For example, "Malaysian tin closed at", which is arranged in one phrase, cannot usually be represented as one grammatical category according to English grammar.

The STRA data is flexible in its ability to translate fixed sentences. For example, the STRA data shown in Fig. 6 and the CTRA data in Fig. 5 can translate :
1-4) "In Tokyo, Malaysian tin closed at 1941 yen per kilo, down 19 yen."
into Japanese:

「東京でマレーシアのすずは，19円ダウンの１キロ1941円でひけた」

because, the fixed expressions in the examples are matched:

Kuala Lumpur	<--->	Tokyo
17.76 dollars per kilo	<--->	1941 yen per kilo
up	<--->	down
5 cents	<--->	19 yen

To make STRA data flexible, the words used in fixed expressions, such as "Kuala Lumpur", "Tokyo", "cents" and "yen", should be registered in CTRA data. These words are selected by hand, referring to frequently appearing fixed expressions collected from corpora.

4.3 Data production for STRA (DTRA)

A data production module for STRA (**DTRA**) builds STRA data automatically from English fixed sentences and their Japanese equivalent sentences. In DTRA, CFG rules are constructed by transforming English fixed sentences, and Japanese templates are made by replacing fixed expressions in their Japanese equivalent sentences with variables. DTRA's algorithm is as follows:

[DTRA's Algorithm]

STEP 0
Translate a fixed sentence w1...wn into Japanese by hand.

STEP 1
CTRA makes candidate variables for a bilingual template.

STEP 2
Define weights for the candidates by the algorithm shown in Fig. 7.

STEP 3
DP selects the optimal set of candidates.

STEP 4
Make CFG rule.

STEP 5
Make Japanese template.

```
for i := 0 to n-1 do
   for j := i+1 to n do
      if (there is a non-active edge including w i ... wj.)
      & (there is an equivalent in the Japanese sentence.)
         then
            weight(i, j) = $3^{j-i}$
      else then
            weight(i, j) = 0
```

Figure 7 Algorithm for calculating relative weights of positions

In STEP 0, a fixed sentence w1...wn is translated into Japanese by hand. STEP 1 collects candidates for variables in the Japanese sentence. Actually, the fixed sentence is analyzed by CTRA, and various fixed expressions are extracted as symbols (pre-terminal or terminal symbols) used in non-active edges. STEP 2 is to calculate the weights of the symbols by the algorithm shown in Fig. 7 to select an optimal set of fixed expressions. If the translation equivalent of a symbol exist in the Japanese equivalent sentence, its weight is defined according to the number of words in edges, otherwise it is zero. STEP 3 selects an optimal set of edges by calculating the maximum in sums of the weights between positions 0 and n by Dynamic Programming (DP). STEP 4 produces pre-terminal symbols for the word sequences not selected in STEP 3, and lines up the symbols in order of their appearance to make CFG rules. In STEP 5, each translation equivalent of the edges in the optimal set is replaced with a variable in the Japanese equivalent sentence to make Japanese templates.

DTRA is illustrated by processing sentence 1-1.

STEP 0 translates the sentence into Japanese by hand:

「クアラルンプールでマレーシアのすずは，5セントアップの１キロ17.76ドルでひけた」

The non-active edges obtained by CTRA in STEP 1 are shown in Fig. 8.

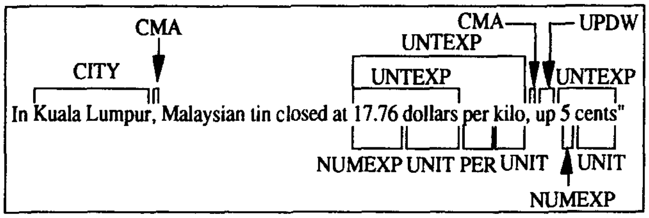
CMA
CMA
UPDW
UNTEXP
CITY
UNTEXP
UNTEXP
In Kuala Lumpur, Malaysian tin closed at 17.76 dollars per kilo, up 5 cents"
NUMEXP
UNIT
PER
UNIT
UNIT
NUMEXP

Figure 8 Non-active edges in sentence 1-1 (by CTRA)

STEP 2 calculates the weights of the non-active edges as shown in Fig. 9. For example, the weight of "Kuala Lumpur" is 9.

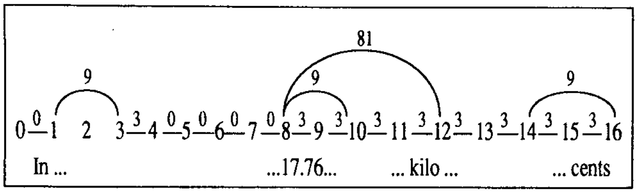

Figure 9 Weights of non-active edges in sentence 1-1

STEP 3 has DP select the maximum in sum of the weights between edge 0 and edge 16. In Fig. 9, the maximum is 108 and the optimal set of edges is selected as {"Kuala Lumpur", ",", "17.76 dollars per kilo", ",", "up", "5 cents"}.

In STEP 4, the word sequences not selected in STEP 3 are given pre-terminal symbols automatically:

"In"	PAT1
"Malaysian tin closed at"	PAT2

and setting these symbols in a line, the CFG rule is:

```
S --> PAT1 CITY CMA PAT2 UNTEXP CMA UPDW UNTEXP
    (  1     2    3    4     5     6    7     8    )
```

The variables are defined as:

#2# = 「東京」
#5# = 「1キロ17.76ドル」
#7# = 「アップ」
#8# = 「5セント」

STEP 5 replaces their translation equivalents of the selected edges in the Japanese sentence with variables:
「#2#でマレーシアのすずは，#8##7#の#5#でひけた」

4.4 Fixed sentence extraction (EXTRA)

A method of extracting fixed sentences (**EXTRA**) collects fixed sentences for DTRA from a corpus using the fixed pattern ratio (FPR) defined below.

The first step in EXTRA is to extract the fixed-word sequences which appear in a corpus most frequently, ignoring differences of days of the week (e.g., Monday and Tuesday) and digits (e.g., 123 and 1000). The fixed-word sequences are not only compound words such as "[DIGIT] dollars per kilo"(where [DIGIT] denotes digits) and "on condition of anonymity", but also some parts of fixed expressions, such as "said in a" and "condition of anonymity". The fixed-word sequences are called "fixed patterns" and the compiled fixed patterns are called "fixed pattern data".

Using fixed patterns, FPR is defined as follows:

$$\text{FPR} = \frac{\text{sum of words in fixed sequences of a sentence}}{\text{the total number of words in a sentence}}$$

Example

4-1) The NYSE's composite index rose 0.39 to 196.61.
4-2) The NYSE's composite index edged up 0.33 to 186.51.

FPD1) "The NYSE's composite index rose [DIGIT] to [DIGIT]" (8 words),
FPD2) "The NYSE's composite index"(4 words),
FPD3) "[DIGIT] to [DIGIT]"(3 words)

FPD1, FPD2 and FPD3 are assumed to be in fixed pattern data. Thus,

The FPR of example 4-1 = 8/8 =1.0, because 4-1 itself is FPD1.

The FPR of example 4-2 = (4+3)/9 = 0.78, because 4-2 includes FPD2 and FPD3.

Fixed sentences are defined as those with FWR values above a certain threshold. EXTRA analyze each sentence in a corpus and extracts the sentences with sufficiently high FPR as fixed sentences.

EXTRA has three parameters:

P1) range of fixed patterns
P2) frequency of fixed patterns
P3) threshold of FPR

5 Experiments

5.1 Extracting fixed sentences

The parameters for EXTRA were selected as:

P1) 3 to 6 words in fixed patterns
P2) more than 10 times
P3) 0.8

To satisfy conditions P1 and P2, about 92,000 fixed patterns were collected from AP wire-service news stories from a two-year period, which include about 1.6 million sentences. Using these fixed patterns, about 21,000 fixed sentences were extracted under the condition P3. The experiment was not limited to economic news stories. Examples of the extracted results are shown in Appendix. Since most of the sentences are economic ones with many idiomatic expressions, EXTRA would be a good method enough to extract fixed sentences.

5.2 Production of STRA data

The 388 most frequently occurring economics-related fixed sentences were manually sampled from the 21,000 fixed sentences. After manually translating them into Japanese, STRA data was produced by DTRA.

While most of CFG rules in the STRA data include variables, a few do not, such as for "Gold prices were mixed." The STRA data produced was as simple as:

S --> PAT225 「#1#」
PAT225 --> "Gold prices were mixed"
「金価格はまちまちだった」

The total number of symbols given in STEP 4, such as PAT1 and PAT2, are approximately 230.

5.3 Experiment for ENTS

A series of experiments was conducted using the STRA data discussed in Section 5.2 to evaluate the accuracy of ENTS.

Table 1 and 2 show each process's volume and translation accuracy, respectively for two data sets: Datal includes 193 economic sentences used to tune to the CFG rules of Process 2, and data2 includes 167 sentences which were not used in the tuning.

Table 1 Processing volume (%)

	Process 1	Process 2	Process 3	Total
data1	29.1	61.3	9.6	100
data2	31.2	57.5	11.3	100

Table 2 Translation accuracy (%)

	Process 1	Process 2	Process 3	Total
data1	100	70.1	10.2	73.0
data2	100	58.7	22.2	66.3

About 30% of each data set is translated in Process 1 and its translation accuracy is 100% for both cases. The translation accuracy of Process 2 for data2 is so high as for data1, although Process 2 is not tuned to data2. The overall translation accuracy increases from about 20% with our conventional MT system to about 70%.

6 Conclusion

We described our new machine translation system (ENTS), which is an economics-specific MT system for processing fixed expressions. We focused mainly on the method of translating fixed sentences in ENTS. The results of experiments show that translation accuracy increases from 20% with our conventional MT to 70% with ENTS. We conclude ENTS will be effective for translating AP economic news stories into Japanese.

The processing rate in Process 1 will be improved by increasing CFG rules in CTRA data and by collecting more fixed sentences and their translation equivalents in DTRA. Moreover we intend to apply the translation method to sports and general news stories.

References

[Aizawa93] Aizawa, T., Katoh, N. and Kamata, M.: "Tuning of a Machine Translation System to Wire-Service Economic News", Proc. of PACLING-93, pp. 304-308 (1993).

[Furuse92] Furuse, O. and Iida, H.: "Cooperation between Transfer and Analysis in Example-Based Framework", Proc. of COLING-92, pp. 645-651 (1992).

[Kaji92] Kaji, H. and Morimoto, Y.: "Learning Translation Templates from Bilingual Text", Proc. of COLING-92, pp. 672-678 (1992).

[Katoh91] Katoh, N., Uratani, N. and Aizawa, T.: "Processing Proper Nouns in machine Translation for English News", Proc. of International Conference on Current Issues in Computational Linguistics, pp. 431-439 (1991).

[Nagao84] Nagao, M.: "A Framework of a Mechanical Translation between Japanese and English by Analogy Principle", in Elithorn, A. and R. Bernerji (eds.) Artificial and Human Intelligence, North-Holland, pp. 173-180 (1984).

[Sumita91] Sumita, E. and Iida, H.: "Experiments and Prospects of Example-Based Machine Translation", Proc. of ACL-91, pp. 185-192 (1991).

[Uratani91] Uratani, N., Katoh, N. and Aizawa T.: "Extraction of Fixed Patterns from AP Economic News" Proc. of 42nd Annual Convention of IPSJ, 6E-4 (1991) ; in Japanese.

Appendix

(FPR: extracted sentence)

1.000: He did not elaborate.

1.000: No injuries were reported.

1.000; The U.S. dollar opened at 159.97 yen on the Tokyo foreign exchange market Monday, up from last Friday's close of 157.65 yen.

1.000; The Federal Reserve Board's index measuring the value of the dollar against 10 other currencies weighted on the basis of trade was 97.46 Tuesday, off 0.74 points or 0.74 percent from Monday's 98.20.

0.970: The average price for strict low middling 1 1-16 inch spot cotton declined 99 points to 78.64 cents a pound Wednesday for the seven markets, according to the New York Cotton Exchange.

0.952: The Nikkei Stock Average closed at 25,194.10, down 48.30 points, or 0.19 percent on the Tokyo Stock Exchange Wednesday.