Summarization by Analogy: An Example-based Approach for News Articles

Megumi Makino and **Kazuhide Yamamoto**
Dept. of Electrical Engineering, Nagaoka University of Technology
1603-1 Kamitomioka, Nagaoka, Niigata 940-2188 Japan
{makino,ykaz}@nlp.nagaokaut.ac.jp

Abstract

Automatic summarization is an important task as a form of human support technology. We propose in this paper a new summarization method that is based on example-based approach. Using example-based approach for the summarization task has the following three advantages: high modularity, absence of the necessity to score importance for each word, and high applicability to local context. Experimental results have proven that the summarization system attains approximately 60% accuracy by human judgment.

1 Introduction

The example-based approach generates language by *imitating* instances, which originated in the machine translation method based on the analogy (Nagao, 1984). The idea is derived from the observation that a human being translates according to past translation experiences. In the machine translation task, this approach has been implemented, and has so far achieved efficient results (Sumita, 1998; Imamura, 2004).

In summarization, a human being also summarizes with his own knowledge and experiences. For this reason, we focus on a summarization method which is based on analogy, *example-based summarization*. The example-based method summarizes the input text in three steps. First, it retrieves a similar instance to the input text. Second, it links equivalent phrases between the input text and the similar instance. Finally, a summary is acquired with combination of some corresponding phrases. Here, we employed a Japanese news article as the input text and utilized news headlines as the instances. The news headline consists of one brief sentence which describes the main point.

We assert that the example-based summarization has the following advantages:

(1)High modularity

Easy improvement and maintenance are required to formulate a useful system in general. An example-based framework makes it easy for us to improve a system by only adding instances. Besides, the addition of instances causes few side-effects.

(2)Use of similarity rather than importance

Almost all previous work on summarization has focused on a sentence extraction. These works compute importance for each word to extract a sentence. However, it is difficult to compute the importance which correlates with human sense. Example-based summarization means there is no need to measure the importance, and it computes the similarity instead. We think it is easier to assess the similarity between two expressions rather than the importance of one expression.

(3)High applicability to local context

The statistical method, in general, attempts to compute the probability of each word appearing in the summary corpus (Knight and Marcu, 2002; Witbrock and Mittal, 1999). This may increase difficulties in maintaining local context, since the statistical approach focuses on the global probability. However, the example-based approach attempts to find most locally similar instance out of the instance collection, which may increase the fitness of input contexts.

For the three reasons given above, this paper explains the system which summarizes a Japanese news article to a one-sentence summary by imitating the similar instance.

As related work, Nguyen et al. (2004) have proposed an example-based sentence reduction model. They deal with the compression of one sentence, while we summarize some sentences into a one-sentence summary. Thus, our summarization ratio is inevitably lower than theirs, as it is considered to be more difficult as a summarization task.

Many studies have summarized some sentences, such as a news article, into a one-sentence summary. Most of them extract the important sentence and contract it. In contrast, our method generates a one-sentence summary by combining phrases in some sentences. Consequently, we can obtain high compression summaries that include information from many positions of the source.

2 Instance Collection

Our example-based summarization regards news headlines as the instance collection. A news headline is a short sentence in which the primary point is written. The following example is Japanese news headlines:

Example (1) :
三菱自動車工業は、中国で乗用車を生産へ。
(Mitsubishi Motors Corp. produces passenger cars in China.)

We use Japanese news headlines, like the above examples, as instances. Besides, as we have noted, only news headlines are used as instances; that is, the pairs formed by an original sentence and its summarized sentence are not used.

3 Example-based Summarization

3.1 Overview

Our example-based summarization system summarizes a lengthy news article into a one-sentence summary by using instances. The overall process is illustrated in figure 1. The system is composed of the following three processes in this order:

1. Retrieve a similar instance to an input news article from the instance collection.
2. Align corresponding phrases between the input news article and the similar instance.
3. Combine the corresponding phrases to form a summary.

Detail of each process is described hereafter.

3.2 Retrieval of Similar Instance

The system measures a similarity between the input and each instance in the instance collection when it retrieves a similar instance. If many words are shared between two expressions, we regard two expressions as similar. Hence, the similarity is calculated on basis of the overlaps of content words between the input news article I and the instance E, defined as follows:

$$Sim(E,I) = \sum_{i=1}^{n} Score(i) \cdot \{w \cdot ||Tv_1(E) \cap Tv_i(I)|| + ||To_1(E) \cap To_i(I)||\} \quad (1)$$

where,
- n : the number of sentences in input,
- $Tv_i(\cdot)$: the verbs set in the last phrase of the i-th sentence,
- $To_i(\cdot)$: the set of content words in the i-th sentence,
- $||Tv_1(E) \cap Tv_i(I)||$: the number of overlaps between $Tv_1(E)$ and $Tv_i(I)$.

In the equation, $Score(i)$ and w are designed to give a higher score if words indicating the main topic of the input article are matched with words in the instance. We have found that words have different contributions, depending on the sentence position, to the main topic. Therefore, we apply $Score(i)$ which depends on the sentence position i, and we use the following experimentally-determined score as $Score(i)$.

$$Score(i) = \begin{cases} 5.15 & \text{if} \quad i = 1 \\ 2.78/i^{0.28} & \text{otherwise} \end{cases} \quad (2)$$

The score indicates an agreement rate of content words depending on the sentence position, which is calculated by using 5000 pairs of newspaper's body and its title[1] We have also found that the verbs in the last phrase are appropriate for the main topic of the input article. For that reason, we determine the weight $w = 3$ by our experiment.

Example 2 shows the similar instance obtained by measuring the similarity.

Example (2) :
Input news article
品質管理能力などの再強化策を話し合うものづくり懇談会が 24 日、会合を開催した。(skip the

---

[1] We used the same kind of newspaper as data set in section 4.1 for calculating $Score(i)$.

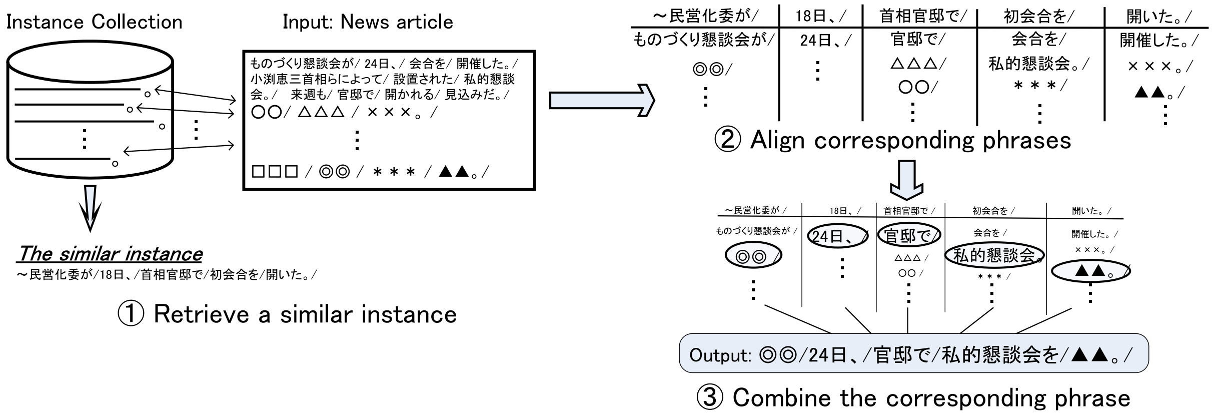
Instance Collection
Input: News article
ものづくり懇談会が/ 24日、/ 会合を/ 開催した。/
小渕恵三首相らによって/ 設置された/ 私的懇談
会。/ 来週も/ 官邸で/ 開かれる/ 見込みだ。/
○○/ △△△ / ×××。/
□□□ / ◎◎ / ＊＊＊ / ▲▲。/
The similar instance
～民営化委が/18日、/首相官邸で/初会合を/開いた。/
① Retrieve a similar instance
～民営化委が/
18日、/
首相官邸で/
初会合を/
開いた。/
ものづくり懇談会が/
24日、/
官邸で/
会合を/
開催した。/
◎◎/
△△△/
私的懇談会。/
×××。/
○○/
＊＊＊/
▲▲。/
② Align corresponding phrases
Output: ◎◎/24日、/官邸で/私的懇談会を/▲▲。/
③ Combine the corresponding phrase

Figure 1: Overview of example-based summarization

rest.)
(The Manufacturing Council held a meeting on the 24th, which discusses the hard-hitting strategy for quality management. ...)

Obtained similar instance
道路公団民営化委が 18 日、首相官邸で初会合を開いた。
(The committee for the privatization of the Public Roads Administration held the first meeting on the 18th at the prime minister's office.)

3.3 Phrase Alignment

We compare the phrases in the input with those in the similar instance, and the system aligns the corresponding phrases. Here, the correspondence refers to the link of the equivalent phrases between the input and its similar instance. The detail of phrase alignment procedures are shown in the following.

To begin with, sentences both in the input and in the similar instance are analyzed using a Japanese syntactic parser CaboCha1). The sentences are split into phrases and named entities (NEs), such as *PERSON*, *LOCATION*, *DATE*, are recognized by the tool.

Then the adnominal phrases in the similar instance are deleted. This is because the adnominal phrases are of many types, depending on the modified noun; accordingly, the adnominal phrase should be used only if the modified nouns are exactly matched between the input and the similar instance.

Finally, the system links the corresponding phrases. Here, phrase correspondence is one-to-many, not one-to-one, and therefore a phrase in a similar instance has some corresponding phrases in the input. In order to compare phrases, the following four measures are employed: (i) agreement of grammatical case, (ii) agreement of NE, (iii) similarity with enhanced edit distance, and (iv) similarity by means of mutual information. The measure of (i) focuses on functional words, whereas the measures of (ii)-(iv) note content words. Let us explain the measures using example 2.

(i) Agreement of Grammatical Case

If there is a phrase which has the same grammatical case[2] in the input and in the similar instance, we regard the phrase as the corresponding phrase. In example 2, for example, the phrases "再強化策 を (the hard-hitting strategy `obj`[3]), 会合 を (the meeting `obj`)" in the input corresponds the phrase "初会合 を(the first meeting `obj`)" in the similar instance.

(ii) Agreement of Named Entity

Provided the input has the same NE tag as the similar instance, the phrase involving its tag links the corresponding phrase. For example, in example 2, the phrase "24 日 [*DATE*] (on the 24th.)" in the input corresponds the phrase "18 日 [*DATE*] (on the 18th.)" in the similar instance.

(iii) Similarity with Enhanced Edit Distance

We adopt the enhanced edit distance to link phrases including the same characters, because Japanese abbreviation tends to include the same characters as the original. For example, the abbreviation of "日

---

[2] Comma is also regarded as grammatical case (i.e., null case) here.

[3] "`obj`" is an object case marker.

本銀行 (Bank of Japan)” is “日銀”. The enhanced edit distance is proposed by Yamamoto et al. (2003). The distance is a measure of similarity by counting matching characters between two phrases. Moreover, the distance is assigned a different similarity weight according to the type of matched characters. We apply 1.0 to the weight only if Chinese-derived characters (Kanji) are matched. We link phrases as corresponding phrases, where the phrases are the top three similar to a phrase in the similar instance.

(iv) Similarity with Mutual Information

We finally compute the similarity with mutual information to link syntactically similar phrases. For example, given the following two expressions: “会議を開く (to hold a meeting)” and “大会を開く (to hold a convention)”, we regard 会議 *(a meeting)* and 大会 *(a convention)* as similar. We use the similarity proposed by Lin (1998). The method uses mutual information and dependency relationships as the phrase features. We extend the method to Japanese by using a particle as the dependency relationships. We link phrases as corresponding phrases, where the phrases are the top three similar to a phrase in the similar instance.

3.4 Combination of the Corresponding Phrases

Our system forms the one-sentence summary by combining the corresponding phrases. Let us explain this process by using figure 2. We arrange the phrase of the input on the node, where the phrases is judged as the correspondence to the phrase in the similar instance. For example, in figure 2, the second nodes *e* and *d* denote the corresponding phrases in the input, which correspond to the second phrase *had* in the similar instance.

We assign the similarity between corresponding phrases as the weight of node. In addition to this, we employ phrase connection score to the weight of edge. The score indicates the connectivity of consecutive two phrases, e.g. two nodes such as node *d* and node *e* in figure 2. If you want to obtain a fine summary, i.e., a summary that contains similar phrases to the similar instance, and that is correct grammatically, you have to search the best path $\hat{W}_p$ for path sequence $W_p = \{w_0, w_1, w_2, \cdots, w_m\}$, where the best path maximizes the score.

$$\hat{W}_p = W_p \quad \text{s.t.} \quad \underset{p}{\text{argmax}}\ Score_p(W_p) \tag{3}$$

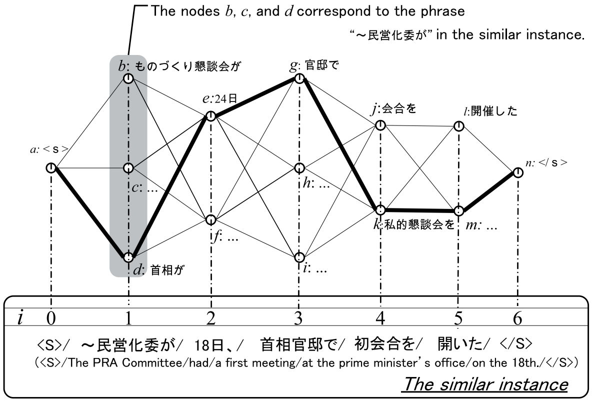
The nodes b, c, and d correspond to the phrase “～民営化委が” in the similar instance.
b: ものづくり懇談会が
g: 官邸で
e:24日
j:会合を
l:開催した
a: < s >
n: </ s >
c: ...
h: ...
k:私的懇談会を
m: ...
f: ...
i: ...
d: 首相が
i 0 1 2 3 4 5 6
<S>/ ～民営化委が/ 18日、/ 首相官邸で/ 初会合を/ 開いた/ </S>
(<S>/The PRA Committee/had/a first meeting/at the prime minister’s office/on the 18th./</S>)
The similar instance

Figure 2: Optimal path problem that depends on combination of the corresponding phrases[4].

The best path $\hat{W}_p$ is a one-sentence summary which is generated by our system. Take the case of the thick line in figure 2, $\hat{W}_p$ is indicated as $\hat{W}_p = \{a, d, e, g, k, m, n\}$, namely, generated summary is formed the phrases $\{a, d, e, g, k, m, n\}$. In eq.3, $Score_p(W_p)$ is given by

$$Score_p(W_p) = \alpha \sum_{i=0}^{m} N(w_i) + (1-\alpha) \sum_{i=1}^{m} E(w_{i-1}, w_i) \tag{4}$$

where α is the balancing factor among the weights of node and edge. We score $\alpha = 0.6$ by our experiment. m indicates the last number of the phrase in the similar instance, $N(w_i)$ is given as follows:

$$N(w_i) = \max \begin{cases} 0.5 & \text{if (grammatical case or} \\ & \quad \text{NE tag is matched)} \\ 1/rank & \text{otherwise} \end{cases} \tag{5}$$

where, *rank* indicates the rank order of the similarity with the enhanced edit distance or mutual information to the phrase w_i. $N(w_i)$ illustrates the similarity between corresponding two phrases. The node score, shown above, is determined by the preliminary experiment. The edge score $E(w_{i-1}, w_i)$ is given by

$$E(w_{i-1}, w_i) = \frac{1}{|loc(w_{i-1}) - loc(w_i)| + 1} \tag{6}$$

where, $loc(w_i)$ denotes where the location of the sentence contains the phrase w_i in the input. The edge score means that if w_{i-1} and w_i are located closely to each other, a higher score is given, since a good connection is expected in this case.

---

[4] The nodes, *a*, *b*, *c*,$\cdots$, *n*, indicate the corresponding phrases to the phrase in the similar sentence. For example, the nodes, *b*, *c*, *d* correspond to *“The PRA Committee.” i* is a phrase number in the similar sentence.

4 Evaluation and Discussion

4.1 The Corpus

We used 26,784 news headlines as instances, which were collected from the Nikkei-goo mail service[2)] for 2001-2006. In order to adjust the weight w in the eq.1 and the balancing parameter α in eq.4, 150 input news articles were used as the tuning set. A different group of 134 news articles were used for evaluation. We used Nihon Keizai Shimbun, a Japanese newspaper[3)], from 1999 through 2000 as tuning and test data.

4.2 Summarization Ratio

To calculate summarization ratio, we have compared the number of characters in the input news articles with that in the output summary. As the result, we obtained a summarization ratio of 5%; namely, 95% characters in the input were reduced. From the summarization ratio, our approach made it possible to summarize sentences into one-sentence summary with high compression.

4.3 Sectional Evaluation

We evaluated each part of our system by human judgment[5]. We first evaluated the process by retrieving similar instance. Next, we evaluated the processes of phrase alignment and the combination by assessing whether the output summaries were appropriate.

· Retrieving Process

An examinee evaluated the similar instances obtained. Given an input news article and the similar instance to the input, the examinee rates the following scale from one to four, based on how similar the similar instance obtained is to the summary which the examinee generated from the input news article:

1) quite similar 2) slightly similar
3) not very similar 4) not similar

Out of 134 input articles, 77 inputs were ranked either 1) quite similar or 2) slightly similar. As a consequence, the accuracy of similar instance obtained is approximately 57%, which indicates that the similarity calculation for obtaining similar instance is feasible.

· Phrase Alignment and Combination

We also evaluated parts of phrase alignment and the combination by human judgment. The examinee compared 77 output summaries with their input. Here, we limited 77 outputs judged as good similar instances in evaluation of the process of retrieving similar instance, because we evaluate specifically the parts of phrase alignment and combination.

The examinee categorized them based on how proper the output summary is to the input news article:

1) quite proper 2) slightly proper
3) not very proper 4) not proper

As a result of judgment, 48 outputs out of 77 are evaluated either 1) quite proper or 2) slightly proper.

Both a statistical method by Knight and Marcu (2002) and an example-based method by Nguyen et al. (2004) contracted one-sentence with a summarization ratio of approximately 60-70%. Both papers indicated that a score of 7-8 on a scale from one to ten was obtained. They deal with the compression of one sentence, while we summarize some sentences into a one-sentence summary. Thus, our summarization ratio is lower than theirs, as it is considered to be more difficult as a summarization task. Despite this, we obtained the ratio that 62% (48 out of 77 results) were judged proper. Although direct comparison of the performance is impossible, it is considered that our proposed method obtains a competitive accuracy.

4.4 Discussions

· Examples of Output Summary

Figure 3 shows some examples of the output summary.

From figure 3, we can see that the similar instances were effectively used, and the appropriate summaries to the input are generated. For example, the second summary in the figure is judged as a fine summary contracting information of two sentences according to the similar instance.

· Analysis of Summarization Errors

In the course of our summarization, we have observed errors due to erroneous correspondences. In Japanese, sometimes two or more phrases are contracted into one phrase, as in the example below. We now only attempt to correspond two phrases one by

---

[5] One examinee judged the parts of our system.

Input news article：
神奈川県警の一連の不祥事のうち、厚木署集団警ら隊の集団暴行事件で起訴された元巡査部長、川野優被告の論告求刑公判が二十一日、横浜地裁で開かれた。検察側はひまを持て余して部下に短銃を突き付けるなど、組織における地位の高さに乗じた悪質な行為などと理不尽な暴力を指弾し、川野被告に懲役一年六月を求刑した。(skip the rest.)
(The prosecution made Kawano's closing arguments on the 21st in the trial at the Yokohama District Court. The ex-sergeant Suguru Kawano is accused of gang-bashing by Atsugi Police Station's patrol group in a string of scandals of Kanagawa Prefectural Police. The prosecutors demanded one and half year in a prison. ...)
Obtained similar instance：
大阪地裁で22日、8人が犠牲となった池田小児童殺傷事件の論告求刑公判が開かれ、検察側は宅間被告に死刑を求刑した。
(The prosecution made Takuma's closing arguments on the 22nd in the trial at the Osaka District Court, and asked for the death penalty.)
Output summary：
横浜地裁で二十一日、論告求刑公判が開かれ、検察側は川野被告に懲役一年六月を求刑した。
(The prosecution made Kawano's closing arguments on the 21st in the trial and demanded one and half years in prison.)

Figure 3: The examples of generated summary

one, and we thus can not deal with many-to-one correspondences.

Example (3) :
前年同月に/[6]比べ → 前年同月比/
(compare with the same month last year)
5月を/メドに → 5月メド/
(in May)

We expect that this kind of phenomenon can be solved by paraphrasing an input summary as well as summary instance. Recently, several works on paraphrasing techniques have been proposed in Japanese, hence such pre-processing before alignment would be feasible.

5 Conclusion and Future Work

We have presented an example-based technique that has been applied to the summarization task. The essence of the proposed method is to generate a one-sentence summary by combining instances each of which imitates the given input.

As the result of human judgment, the retrieval process of a similarity sentence attained 57% accuracy. And our method generated summary in which 62% were judged proper. We have confirmed by our observation that the summaries were generated by combining the phrases in many positions of the input, while those summaries are not given just by common methods such as sentence extraction methods and sentence compression methods.

The sectional evaluation and the inspection of example output show that this system works well. However, larger scale evaluation and comparison of its accuracy remain to be future work.

Tools and language resources

1) CaboCha, Ver.0.53, Matsumoto Lab., Nara Institute of Science and Technology.
http://chasen.org/˜taku/software/cabocha/
2) Nikkei News Mail, NIKKEI-goo,
http://nikkeimail.goo.ne.jp/
3) Nihon Keizai Shimbun Newspaper Corpus, years 1999–2000, Nihon Keizai Shimbun, Inc.

References

Kenji Imamura. 2004. *Automatic Construction of Translation Knowledge for Corpus-based Machine Translation*. Ph.D. thesis, Nara Institute of Science and Technology.

Kevin Knight and Daniel Marcu. 2002. Summarization Beyond Sentence Extraction: A Probabilistic Approach to Sentence Compression. *Artificial Intelligence*, 139(1):91–107.

Dekang Lin. 1998. Automatic Retrieval and Clustering of Similar Words. In *Proceedings of COLING-ACL98*, pages 768–773.

Makoto Nagao. 1984. A Framework of a Mechanical Translation Between Japanese and English By Analogy Principle. In *Artificial and Human Intelligence*, pages 173–180.

Minh Le Nguyen, Susumu Horiguchi, Akira Shimazu, and Bao Tu Ho. 2004. Example-Based Sentence Reduction Using the Hidden Markov Model. *ACM Transactions on Asian Language Information Processing*, 3(2):146–158.

Eiichiro Sumita. 1998. *An Example-Based Approach to Transfer and Structural Disambiguation within Machine Translation*. Ph.D. thesis, Kyoto University.

Michael J. Witbrock and Vibhu O. Mittal. 1999. Ultra-Summarization: A Statistical Approach to Generating Highly Condensed Non-Extractive Summaries. In *Research and Development in Information Retrieval*, pages 315–316.

Eiko Yamamoto, Masahiro Kishida, Yoshinori Takenami, Yoshiyuki Takeda, and Kyoji Umemura. 2003. Dynamic Programming Matching for Large Scale Information Retrieval. In *Proceedings of the 6th International Workshop on Information Retrieval with Asian Languages*, pages 100–108.

---

[6] "/" indicates a phrase boundary.