Extracting Characteristic Sentences from Related Documents

Naoaki Okazaki [*†] Yutaka Matsuo [‡] Naohiro Matsumura [*†] Hironori Tomobe [*]
Mitsuru Ishizuka [*]
Hongo 7-3-1, Bunkyo-ku, Tokyo, 113-8656 Japan.

Abstract.
More and more information is available recently. To find a chance i.e., an important event for decision-making, we have to be prepared for the chance. Recent progress of automatic summarization may contribute to Chance Discovery in that it helps a user read a lot of documents easily and be prepared for the chance. In this paper, we develop a new method for multi-document summarization which extracts a set of characteristic sentences that maximizes the coverage of an original content and minimizes the redundancy of a summary. On top of the summary result, we provide a word cooccurrence graph and show why the result is obtained.

1 Introduction

Text summarization is the process of distilling the most important information from a source (or sources) to produce an abridged version for a particular user and task [1]. Although the measure of 'importance' in this definition varies from one user or task to another, most of the current researches hardly pay attention to that.

From chance discovery point of view, a summary should be useful for a particular user to make decisions, or at least to understand the content for making decisions. Importance of a sentence or a set of sentences is dependent on who makes the decision (ie. a user) and what kind of decision one makes (ie. a task). We are aiming at user-task-dependent summarization, however, in this paper we show a general-purpose summarization system. The important feature of our system is that we provide a word cooccurrence graph of the original documents on top of the summary result and show a user why the summary is obtained. The explanation of the summary result may help a user convince the summary result, and attract new interests. It makes a great benefit in a user's decision-making (even if the algorithm is currently general-purpose).

To enable the visual presentation of a summary result, our method is based on a graph representation. We transform the summarization task into an edge covering problem on the graph. Because the edge covering problem is NP-complete, we use a fast hypothetical reasoning solver to obtain the result immediately.

[*]Graduate School of Information Science and Technology, University of Tokyo
[†]Japan Science and Technology Corporation
[‡]Cyber Assist Research Center, National Institute of Advanced Industrial Science and Technology

The rest of the paper is organized as follows: In the following section, we overview the summarization. Then we show how to compile the summarization into a hypothetical reasoning problem. In Section 4, the examples are shown. We discuss the future works and conclude this paper.

2 Summarization overview

Considering the human's process of summarization, we (1)understand the content of a passage, (2)pick up sentences or phrases that are considered as significant, and (3)synthesize an outputting summary together with splicing the extracted textual segments. Since (2) is relatively easier for computer to deal with than (1) and (3), extracting significant units has researched since 1950s [2, 3] and been the basis of automatic summarization.

Extracting significant units is a method of evaluating textual units in the source document(s). Most of the current research is not necessary for catching on what the text actually said to estimate the significance, but takes advantage of the surface phenomenon behind a text.

The extension of single-document summarization to collections of related documents is called multi-document summarization. We frequently meet related documents, for example, a collection of documents retrieved from a search engine with some queries, messages on an internet discussion board or mailing list, etc. As the related documents have some similar contents or expressions, extracting the significant textual units often results in a redundant summary. Multi-document summarization should therefore be capable of identifying common and different parts and removing redundancy in a summary.

3 Extracting the best combination of sentences

3.1 Formulation of extracting characteristic sentences

Based on the above discussion, we created a system extracting a set of sentences from related multi-documents. It uses a word cooccurrence graph to extract the best combination of sentences. We formulate this multi-summarization problem as follows.

First, we make a word cooccurrence graph from documents. Figure 1 shows a word cooccurrence relation between terms in a set of articles about "hybrid car." A node stands for a term, and we link nodes when a pair of terms is appeared in the same sentences more than twice.

What kind of sentences is characteristic in the graph? Each sentence in a document presents the relations between terms [4]. That is equivalent in the graph to covering several links. As a consequence, we should choose a set of sentences that covers as many links as possible in the graph. It is useless to pick up a sentence which covers the same links as the previously selected sentence does. Therefore we obtain an edge covering problem defined as the following optimization problem;

$$\min f = \sum_{i \in K} cost_i x_i \tag{1}$$

where K is a set of links, $cost_i$ is a penalty cost when link i is not included in the summary, and x_i is a 0–1 boolean variable whether link i is included(1) or not(0).

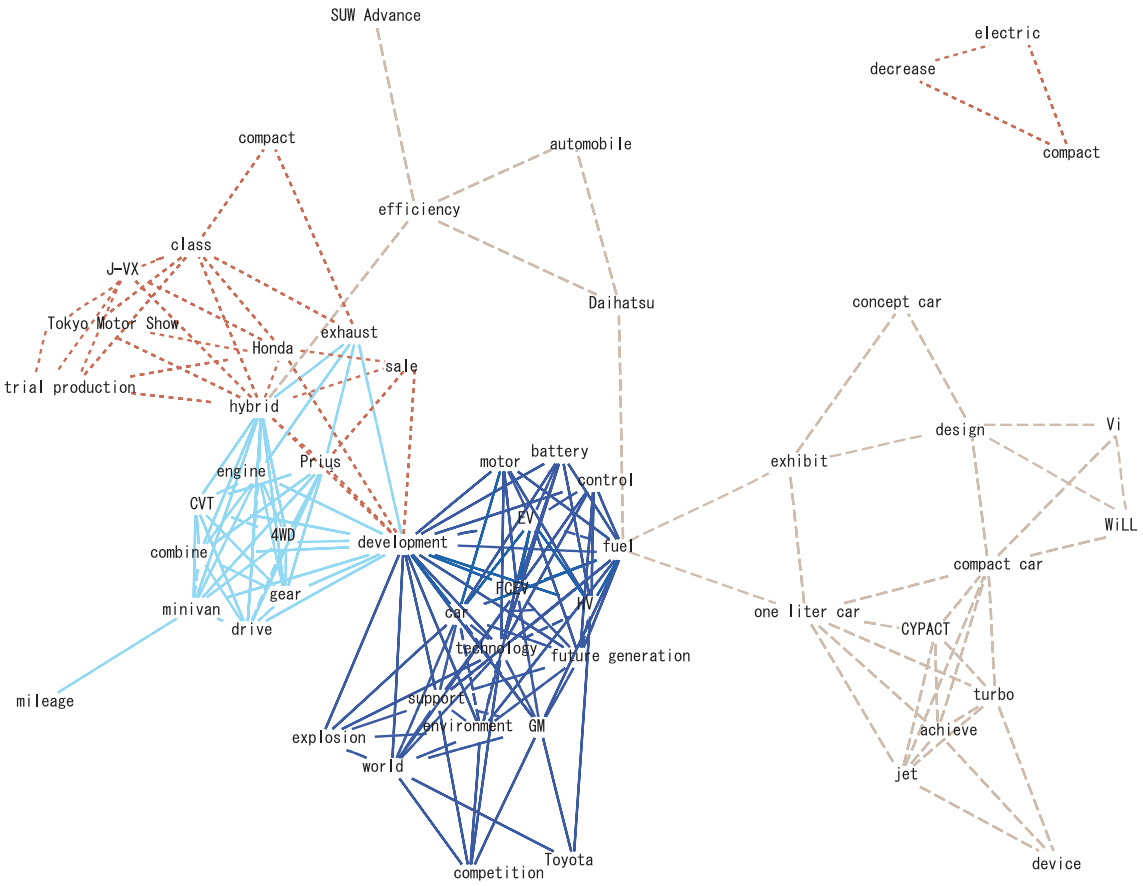

Figure 1: A word cooccurrence graph of a set of news articles. The source articles are a set of news articles about hybrid car written by the Mainichi newspaper(originally written in Japanese). The distance between nodes(terms) is about in inverse proportion to the times of cooccurrence. A line style corresponds to an article.

3.2 Transformation of the optimization problem into cost-based hypothetical reasoning

We solve the optimization problem by applying cost-based hypothetical reasoning as follows. We denote k as the total number of links and m as the total number of sentences, We define goal G to represent *"All links are taken into consideration"* as follows.

$$G \leftarrow x_1, x_2, ..., x_k \tag{2}$$

A hypothesis h_{s_j} represents *"sentence j is selected"* and has the corresponding cost with its length. For example, if sentence 1 has link#13, link#220, link#223, then we get the following rules.

$$x_{13} \leftarrow h_{s_1}, x_{220} \leftarrow h_{s_1}, x_{223} \leftarrow h_{s_1} \tag{3}$$

For unselected link i, on the other hand, we introduce a hypothesis h_{emp_i} to represent *"sentence i is not included in the summary"* and the following rules.

$$x_i \leftarrow h_{emp_i} (i = 1, ..., k) \tag{4}$$

We annotate h_{emp_i} with penalty cost 1.

At last we can decide a set of sentences by generating knowledge base and finding a combination of sentence that proves goal G. We use a fast hypothetical reasoning method [5]

Toyota was the first to develop the hybrid car and started selling the Prius in December of last year. October of last year at the Tokyo Motor Show, Honda presented the J-VX, a 1000cc class experimental hybrid engine car.
On the 19th, Toyota and General Motors (GM) announced at the same time their plans to cooperate in automobile technology advancement for environmentally friendliness in automobiles with the prospect of making next generation low-pollution vehicles such as the awaited Fuel Cell Electric Vehicle (FCEV). They will collaborate in the research of chemical reactions of hydrogen from fuel and oxygen from air to generate power, plus a wide variety of technologies which should bring about the production of more hybrid vehicles (HV) like the FCEV which join a gasoline engine and an electric motor, and electric vehicles (EV) whose motors are run on a storage battery.
Up to now, the only one to combine gasoline and electricity has been the Prius' 1500cc engine, but for a minivan that needs more power, they revealed to newspapers the plans for the first hybrid with four-wheel drive, a Continuously Variable Transmission (CVT) vehicle which combines a 2400cc engine and motor.
Nissan has also achieved CYPACT, a three-liter compact car with a jet-fueled diesel turbo engine .

Figure 2: An example of summary. The source is a collection of 4 articles about "hybrid car" in the Mainichi. (Translated from Japanese for purposes of illustration)

which solves a hypothetical reasoning problem quickly by transforming the problem into two continuous optimization problems.

3.3 Implementation

First, we analyze the source text into morpheme and identify part of speech of each term by using Chasen [6]. Sorting nouns and verbs from terms, we enumerate cooccurrence relations between the terms in the same sentence. And then we make and solve a summarization problem described above.

We participate in a competition of summarization, TSC(Text Summarization Challenge) [7] task organized by NTCIR-3 project and we used a collection of the Mainichi articles for an experiment. Current system is only for Japanese news paper articles because of the morphlogicalanalysis, but the core algorithm can be easily applied to other languages.

4 Discussion and Conclusion

Figure 2 and 3 are two examples of summaries. The source articles are omitted due to limitations of space. Figure 2 is produced from Figure 1, a word cooccurrence graph in a collection of articles about "hybrid car." As can be seen from the summary, our system depicts various efforts of makers toward bybrid cars. Several brand names in Figure 1 indicate hybrid cars, ie. *Prius*, *J-VX*, and *SUW Advance*, which are located near 'hybrid' node. *WiLL* and *CYPACT* are located, on the other hand, far from 'hybrid' node. They are not actually hybrid cars.

Figure 3 is a summary of the article collection about earning gold medals of Japanese athletes. In addition to prompt reports, we must not miss that it includes some anecdotes about the win. It is not novel in the summary that Japanese athletes won the games because the ar-

On the 10th, day four of the eighteenth winter olympics in Nagano, male speed skater Shimizu won the first gold medal for Japan in the men's 500 meter competition held at Nagano city's M-Wave. From this competition on, scores are determined by the combined event times of two races, skating on the outside and inside courses.
On the 11th, women's freestyle skiing and mogul competitions were held. Satoya, on a foreign expedition, was shaken at Masaaki's illness when she was called just before. She said with strong conviction, "I skiied for him as well as for myself."
On the 15, day nine at the large hill (120m) ski jump in Hakuba village, Kazuyoshi Funaki (of Descente) won the gold, and Masahiko Harada (of Snow Brand) took third place.
On the 15th at the individuals large jump competition, Harada, the 25th jumper in his second jump, jumped 136 meters, but his result as combined with his first jump were not immediately displayed on the electrical scoreboard, but they were finally shown about ten minutes after finishing, after the calculation and confusion over the reporting of the winners.
The Japanese ski jump team won their first olympic competition, and Japanese athletic teams have secured one-hundred gold medals in all in the summer and winter olympics.

Figure 3: Another example of summary. The source is a collection of 7 articles retrieved with queries, "Nagano Olympic, Japan, gold, win" in the Mainichi. (Translated from Japanese for purposes of illustration)

ticles were collected intentionally with queries, *"Nagano Olympic, Japan, gold, win."* These queries often appear in the same sentence and have close cooccurrence ralations. Because our summarization strategy tends not to bring such same cooccurrence relations into a summary, it chose instead some secret stories of which some users might not know.

In conclusion, we have developed a new summarization algorithm which solves the transformed edge covering problem. By showing a word cooccurrence graph, we can make the summary more understandable for a user. We are going to extend our method to deal with user- and context-dependent summarization system.

References

[1] Mani, I. *Automatic Summarization.* John Benjamins Publishing Company, 2001.

[2] Luhn, H. P. The automatic creation of literature abstracts. *IBM journal of Research and Development*, Vol. 2, No. 2, pp. 159–165, 1958.

[3] Salton, G. *Automatic Text Processing*. Addison-Wesley, 1989.

[4] Halliday, M.A.K, Hansa, R, *Cohesion in English*, Langman, 1976.

[5] Matsuo, Y., Ishizuka, M. Two Transformation of Clauses into Constraints and their Properties for Cost-based Hypothetical Reasoning, PRICAI-02, to appear.

[6] Chasen's Homepage: `http://www.chasen.org/`

[7] TSC2's Homepage: `http://lr-www.pi.titech.ac.jp/tsc/index-en.html`