An Abstract Generation System for Social Scientific Papers

Michio Kaneko
Graduate School of Integrated Basic Sciences
Nihon University, Tokyo, JAPAN
m-kaneko@chs.nihon-u.ac.jp

Dongli Han
Department of Information Science,
College of Humanities and Sciences
Nihon University, Tokyo, JAPAN
han@chs.nihon-u.ac.jp

Abstract

Abstracts are quite useful when one is trying to understand the content of a paper, or conducting a survey with a large number of scientific documents. The situation is even clearer for the domain of social science, as most papers are very long and some of them don't even have any abstracts at all. In this work, we narrow our attention down to the social scientific papers and try to generate their abstracts automatically. Specifically, we put weight on three points: important keywords, readability as an abstract, and features of social scientific papers. Experimental results show the effectiveness of our method, whereas some problems remain and will need to be solved in the future.

1 Introduction

Abstracts are expected to help readers who are trying to understand the outline of a paper, or conducting a survey with a large number of scientific documents. The situation is even clearer for the domain of social science, as most papers in this area tend to be very long and some of them don't even have any abstracts at all.

There have been many methods proposed for Japanese summarization (e.g., Ochitani et al. 1997; Hatakeyama et al., 2002; Mikami et al., 1999; Ohtake et al., 1999; Hatayama et al., 2002; Tomita et al., 2009; Fukushima et al. 2011). However, most existing proposals are made towards general text summarization instead of abstract generation for scientific papers. Here, it is important to distinguish between a summary and an abstract. According to a Japanese dictionary, an abstract contains the most important stuffs or the important matter that has been stated in a document, and a summary is a short text transformed from a long text containing all the important points in the original text (Umesao et al., 1995).

With the difference between summaries and abstracts in mind, we attempt to propose a new method to generate abstracts for social scientific papers in this paper. Specifically, we put weight on three points: important keywords, readability as an abstract, and features of social scientific papers.

In this paper, we first describe our proposal in Section 2, 3, 4 and 5. Specifically, Section 2 gives a brief introduction on the necessary language resources for the development of the subsequent modules. Section 3, 4 and 5 describe the sentence processing, importance degree estimation, and abstract generation respectively. Finally, we discuss some experiments conducted to evaluate the effectiveness of our approach in Section 6.

2 Necessary Language Resources

In this work, in order to perform textual analysis and importance degree estimation for words or phrases, we create the following five lexicon-files beforehand.

<Adverb Lexicon>:
created from (Nitta, 2002) containing adverbs describing degrees (like *emphasis*).
<Sentence-End Expression Lexicon>:
extracted from (Morita and Matuki, 1989) containing all expressions functioning similarly to auxiliary verbs in Japanese.
<Conjunction Transformation Lexicon>:

27th Pacific Asia Conference on Language, Information, and Computation pages 57－65

containing the corresponding relations between conjunctive particles and conjunctions.

<Indispensable-case Lexicon>:
generated from EDR [1] containing all the necessary cases of predicates.

<Conjunction Lexicon>:
containing the conjunctions used to expand one affair to multiple affairs, and the copulative conjunctions used to connect two affairs in Japanese as shown in Figure 1 (Ichikawa, 1978).

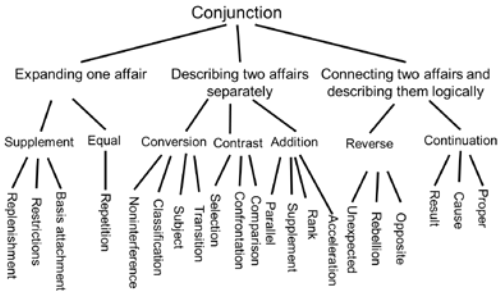

Figure 1. Conjunction classification

Moreover, we have created three lexicons specialized in social science. The first one is a called Keyword Dictionary containing the words extracted from two sociological dictionaries (Uchida et al., 2001; Imamura, 1988).

The second lexicon is the Noun-phrase Dictionary. Based on the idea that noun phrases play important roles in sentences (Minami, 1974), we extract five kinds of noun phrases from a social scientific literature database according to the following definitions

- Expressions ending with the continuous form of a nominalized verb
- Nominalized verb + "カタ", "ブリ" ("ップリ"), "ヨウ", "バ", "バショ", "トコロ" ("ドコロ"), "トキ" ("ドキ"), etc.
- Adjective + "サ"
- Noun + Noun.
- Adnominal form of an inflectable word + noun

The social scientific literature database we have created in advance is composed of 221 social scientific papers obtained from the Web containing 63,056 sentences.

The third lexicon, Mutual-information Table, is also generated from the scientific literature database. It contains mutual information between nouns appearing in each literature. Mutual information between nouns is calculated with Formula 1 (Church , 1990).

$$X(A,B) = \log \frac{P(A,B)}{P(A)P(B)} \quad (1)$$

P(A) and *P(B)* in Formula 1 indicate the occurrence probability of noun A and noun B respectively, and *P(A,B)* indicates the co-occurrence probability of noun A and noun B in the same sentence of the database.

3 Sentence Processing

After conducting a morphological analysis on the input social scientific paper, we execute a series of processing on each sentence of the paper: keyword extraction, parenthesis processing, third-person sentence removing, sentence segmentation, and sentence-information assignment. Here, we describe them in each sub-section respectively.

3.1 Keyword Extraction

Keywords are extracted for subsequent importance degree estimation. Here, words and phrases are extracted from the paper as *Keywords* if they also appear in the Keyword Dictionary. Similarly, the noun-phrases matching the Noun-phrase Dictionary are extracted as *Fkeywords*. Another sort of keyword is called *Nkeywords*, which stands for common noun or compound noun, and has been extracted during the morphological analysis using Mecab [2], a free Japanese morphological analyzer. Meanwhile, the occurrence frequency of each extracted keyword and the place it appears (i.e., the number of paragraph it appears in) are also recorded.

3.2 Parenthesis Processing

Generally, texts enclosed in round parentheses tend to act as supplement or modification to the texts prior to it. Therefore, round parentheses could be simply removed without influencing the basic meaning of the original texts in most cases. However, there is one exception. When the texts contained in the round parentheses are less than 15 characters, they will be extracted as another sort of keyword, *Tkeywords*. Here, the number 15 indicates the maximum keyword-length in the Keyword Dictionary.

[1] http://www2.nict.go.jp/out-promotion/techtransfer/EDR/J_index.html

[2] http://mecab.sourceforge.net/

3.3 Third-person Sentence Removing

One of our goals in this work is to extract the text that expresses the author's opinions most directly and correctly. For this reason, we consider that sentences holding third-person subjects are inappropriate to appear in the final abstract. Sentences fulfilling the following conditions are recognized automatically as third-person subject sentences, and excluded from final sentence candidates for abstract generation.

- sentences
  containing either "は" or "が", and the previous morpheme being a proper personal name.
- sentences
  containing either "は" or "が", the previous morpheme being a suffix, and the morpheme prior to the suffix being a personal name.
- sentences
  containing either "は" or "が", and the previous morpheme being a third-person pronoun such as "彼" (he) or "彼女" (her).

3.4 Sentence Segmentation

Social scientific papers in Japanese often contain long sentences. In most cases, only one part of the sentence is important and expected to be included into the final abstract, whereas the rest part might be unnecessary and redundant. Along this idea, we segment long sentences in accordance with the rules in Table 1.

Original	After Segmentaiton
Verb(+Suffix) +"、"	verbal +"。"+ "そして" +"、"
Conjunctive particle + "、"	verbal +"。"+ Conjunction +"、"

Table 1. Rules for sentence segmentation

Here in Table 1, "、" and "。" indicate comma and period in Japanese, and "そして" means "then" in English.

Moreover, in the lower case of Table 1, i.e., when the original sentence is in a form of "conjunctive particle + comma", a transformation will be executed using the Conjunction Transformation Lexicon described in Section 2. Table 2 shows some examples in the Conjunction Transformation Lexicon.

Conjunctive particle	Conjunction
が	だが
て	そして
で	そして
ので	なので
ば	ならば
や	それに

Table 2. Example rules of the conjunction transformation lexicon

3.5 Sentence-information Assignment

The last process in this module is to assign some required information to sentences: cohesive relation and position information.

A cohesive relation indicates a strong relation lying between two sentences. Specifically, we use the following four patterns to match two sentences where cohesive relations exist in between.

- The sentence containing an interrogative and the subsequent sentence.
- The sentence containing a demonstrative and the preceding sentence.
- Two sentences connected by conjunctions that are used for connecting two affairs logically.
- Two sentences connected by conjunctions that are used to expand and describe the previous affair.

In the first pattern, if the sentence containing an interrogative appears at the end of the paper, no cohesive relation will be assigned. Similarly, in the second pattern, if the sentence containing a demonstrative is the first sentence, or the demonstrative is pointing to something within the current sentence, no cohesive relation will be assigned either. The third and the forth pattern are defined based on the conjunction classification tree in Figure 1.

Position information is associated with the position of the sentence. We have carried out an investigation on 40 social scientific papers with regard to the position where important sentences tend to appear. It turns out that the first paragraph and the last paragraph of each chapter, and the whole last chapter have an inclination to contain important sentences. The system records the number of chapter and paragraph as the position information of the current sentence which will be used for importance degree estimation afterward.

4 Importance degree Estimation

An abstract is expected to contain the most important part of the original paper. In this section, we describe our proposal to estimate the importance degree of each keyword in the first step and that of each sentence in the second step for a particular social scientific literature.

4.1 Importance degree Estimation for Keywords

Four kinds of keywords (i.e., *Keywords*, *FKeywords*, *NKeywords*, and *TKeywords*) are considered as the candidates to be included in the final abstracts. We calculate the importance degree of each keyword (denoted as *K_score* hereafter) using its occurrence frequency and distribution as shown in Formula 2.

$$K\_score = wc \times (\frac{wp}{dp} + 1) + eInf \qquad (2)$$

Here, *wc* indicates the occurrence frequency of the keyword under calculation, *wp* and *dp* indicate the number of the paragraph the keyword appears in and the total number of paragraphs contained in the whole paper. Meanwhile, *eInf*, abbreviated from "extra information" acts to make difference between each kind of keywords.

We have defined two kinds of *eInf* for different keywords. First, for *Keywords*, *FKeywords*, and N*Keywords*, the *eInf* amounts to the occurrence frequency of the keyword within important positions, i.e., the first paragraph and the last paragraph of each chapter, and the whole last chapter. Then, for *TKeywords*, we consider the total number of characters is more informative than the position information, and therefore plug it into *eInf*.

Obtained importance degrees of keywords are recorded and will be used for sentence-importance estimation in Section 4.2.

4.2 Importance degree Estimation for Sentences

This sub-section describes the method for calculating the importance degree of each sentence in a paper. This information will become the basis of abstract generation in Section 5.

The importance degree of a sentence (denoted as *S_score* hereafter) is computed following Formula 3.

$$S\_score = \sum_{i=1}^{n} \{K\_Score(keyword_i)\} \times \alpha^k \qquad (3)$$

Basically, *S_score* can be acquired as the total value of all *K_scores* otained in Section 4.1. Here we denote the total number of keywords in the sentence as *n*. In case shorter keywords are contained in longer keywords, we employ the *longest match principle* and put a high priority on longer keywords.

α in formula 3 is a weighted value for the following four kinds of special expressions.

- emphasis expressions
  existing in the Adverb Lexicon
- sentence-end expressions
  existing in the Sentence-End Expression Lexicon
- theme expressions
  nouns prior to "は"
- cohesive relations

If any of the above expressions is found within the sentence under calculation, the total value of all *K_scores* will be multiplied by α (> 1.0) for *k* times. *k* is the total count of the above expressions contained in the sentence.

5 Abstract Generation

We have obtained the importance degrees for all the sentences in Section 4. However, we still need to cut the unnecessary part in each sentence to keep each sentence in the final abstract appear plain and sophisticated. This function is called sentence simplification in this paper. Then we are going to conduct constituent-sentence acquisition, cohesive sentence insertion, and abstract assembling eventually to generate the final abstract. In this section, we describe each function in detail.

5.1 Sentence Simplification

We attempt to cut the unnecessary part and simplify a sentence using three kinds of information: indispensable cases, dependency relations between segments, and mutual information.

An indispensable case is a necessary case of a predicate, such as "ガ" or "ヲ" expressing *agent* case and *object* case respectively. A sentence tends to appear unnatural if its main predicate lacks one or more indispensable cases. We use the Indispensable-case Lexicon described in Section 2 to put a mark on each segment containing an indispensable case.

Dependency relations are usually obtained with the help of a Japanese dependency analyzer. Here, we use Cabocha [3] to analyze the dependency relations between segments in a Japanese sentence. Figure 2 is the analyzing result of an example sentence, "政治階級という言葉は階級という言葉とともに死語と化したのである" (The word *estate government* turned into the dead language along with the word *estate*).

In Figure 2, there are six segments in the input sentence, and the main segment is "化したのである" (turned). We can also see that three segments are modifying directly, or depending on in other words, the main segment, while the rest two are not.

Our idea is to employ this difference to cut the unnecessary part, i.e., the segments which are not depending on the main segment. However, if an indispensable-case exists in a segment, even the segment is not depending on the main segment directly, it is still left in the sentence otherwise the sentence will appear odd.

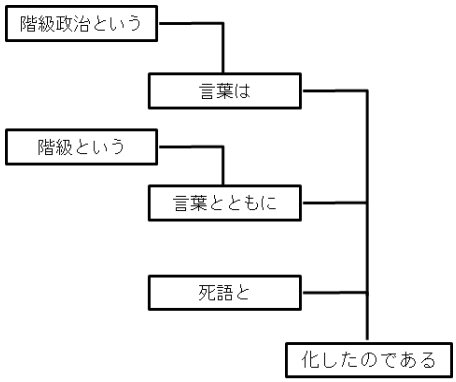

Figure 2. The analyzing result of an example sentence

Meanwhile, if we can find a sufficiently-high mutual information in the Mutual-information Table for a noun (denoted as $noun_a$) in any of the remaining segments, and another noun (denoted as $noun_b$) in the deleted segments, the segment containing $noun_b$ will be left undeleted in the sentence. Table 3 shows some examples from the Mutual-information Table.

All the simplified sentences inherit the importance degrees of the original sentences.

Noun	Noun	Mutual Information
サミット	ミーイズム	1.588042
サミット	世界	0.458759
サミット	論説	2.043721
サミット	各国	1.628684
サミット	自国	2.365649
サミット	利益	0.780687
サミット	形骸	3.365649
サミット	経済	1.687578

Table 3. Some examples from mutual information table

5.2 Constituent-sentence Acquisition

Constituent sentences are the sentences extracted from the original paper to compose the final abstract. Basically, the system just picks out the topmost *n*% simplified sentences based on their importance degrees. Here, *n* stands for the target compression rate which is set by the user before generating the abstract. Three ways have been proposed to determine the total number of constituent sentences or characters. We denote them as NC_1, NC_2, and NC_3 as shown below.

- NC_1
  = *n*% × total number of sentences in the original paper
- NC_2
  = *n*% × total number of characters in the original paper
- NC_3
  = NC_2 + cohesive sentences

NC_1 is the simplest way for determining necessary number of constituent sentences. Unlike with NC_1, NC_2 uses the number of characters to calculate necessary constituent number. For example, if the original paper contains 1000 characters, and *n* has been set to 20, the system will extract simplified sentences in order of their importance degrees until the total number of extracted characters is equal to or larger than 200. The difference between NC_2 and NC_3 lies in the consideration of cohesive sentences. At the time the total number of extracted characters becomes larger than the calculated constituent number (200 in the above example), if the last-extracted sentence is the first sentence of a cohesive sentence pair, the system will extract the second sentence of the pair as well. Otherwise, the last-extracted sentence is removed from the constituent-

[3] http://code.google.com/p/cabocha/

sentence set. We attempt to make the final abstract appear as natural as possible in this way.

We will give a further discussion on the difference among NC_1, NC_2, and NC_3 in Section 6.1.

5.3 Cohesive Sentence Insertion

As stated in Section 5.2, a cohesive sentence pair is composed of two sentences holding strong association in between.

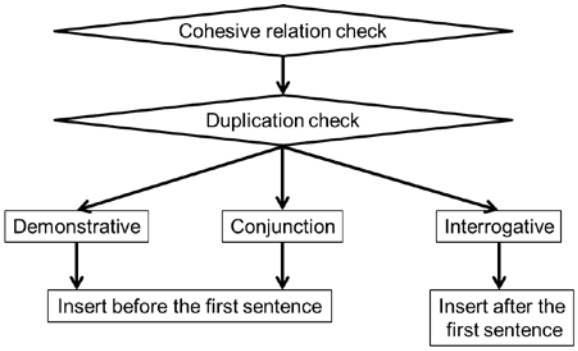

Figure 3. The flow of cohesive sentence insertion

If one and only one sentence has been selected as an abstract constituent, another sentence in the pair should also be extracted and attached to the first sentence in order to keep the final abstract coherent and natural. The appending position is determined according to the type of cohesive relation as shown in Figure 3.

5.4 Abstract Assembling

We have described the procedure to extract constituent sentences so far. The next step is to assemble all the constituent sentences in the order they have appeared in the original paper to compose the abstract. Finally, we conduct the following adjustment to format the abstract.

- connect two sentences coming from the same sentence in the original paper using the rules in Table 2 in the opposite direction.
- replace the theme in the subsequent sentence with a demonstrative if the preceding sentence has the same theme.
- start a new paragraph whenever the chapter changes according to the position information of each sentence.

6 Experiments and Evaluations

We have conducted several experiments to examine the effectiveness of our approach. Here in this section, we first introduce a set of experiments on different manners to determine the number of constituent sentences, then describe a subjective assessment on the system-generated abstract in comparison with another two abstracts. Finally, some discussions are made about the problems and their potential solutions.

6.1 Experiments on the Difference between NC_1, NC_2, and NC_3

In order to figure out the difference between three constituent-extraction manners, we calculate the standard deviations of the total character-number in the generated abstracts with NC_1, NC_2, and NC_3 respectively.

We select six social scientific papers as the experimental objects. Each paper has been input into three prototypes following the definitions of NC_1, NC_2, and NC_3 respectively. The average value of the ratios of the number of characters contained in each generated abstract divided by that of each original paper has been shown in Figure 4, 5 and 6.

A comparison with the target ratio from 5% through 30% has been made to figure out how close the actual number of characters is to the calculated target number.

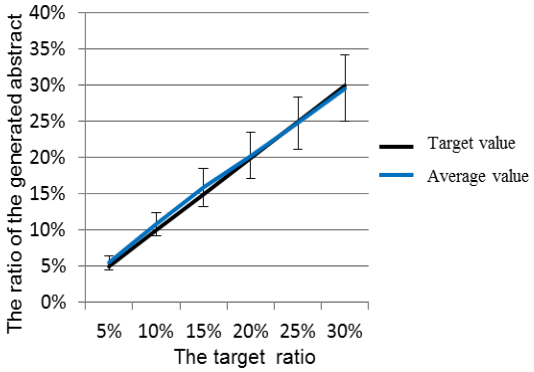

Figure 4. Experimenal results with NC_1

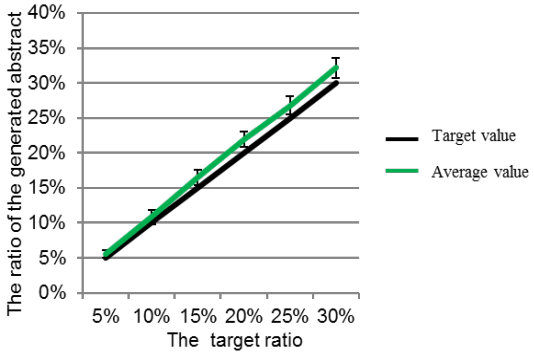

Figure 5. Experimenal results with NC_2

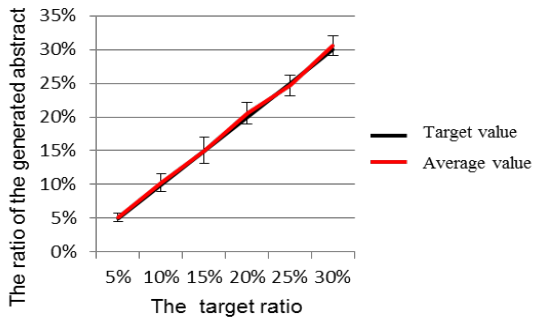

Figure 6. Experimenal results with NC_3

From the above figures, we can see that the average-value curve for NC_3 is the most accurate one. The standard deviation for each constituent-extraction manner has also been calculated. They are 0.92%~4.60% for NC_1, 0.56%~1.40% for NC_2, and 0.66%~1.95% for NC_3. There is little difference between the deviations of NC_2 and NC_3, both of which use a character-based calculation to extract constituent sentences. On the other hand, NC_1 has exhibited relatively more volatility, which shows the instability nature of sentence-based calculation.

As a result, we decide to use character-based calculation to estimate the necessary number of constituents for abstract generation in subsequent processing.

6.2 A Subjective Assessment

We conduct a subjective assessment using three kinds of abstracts.

- The abstract written by the authors (called as *A-abstract* hereafter).
- The abstract created by the system. (called as *S-abstract* hereafter)
- The abstract created by Microsoft Word 2003 (called as *W-abstract* hereafter)

In this experiment, the papers as specified in Table 4 were used.

	Number of paragraphs	***Number of sentences***	***Number of words***	***Publication type***
Paper1	51	448	12138	bulletin
Paper2	38	175	5461	journal article
Paper3	23	155	5514	bulletin

Table 4 . Paper information

Four graduate students and fourteen undergraduate students all majoring in natural language processing have supported us with the subjective assessment. They are divided into five groups each with three or four students. All the three kinds of abstracts are provided to each group without explicit information on which is *A-*, *S-* or *W-abstract*. After 30 minutes' personal reading and 20 minutes' group discussion, each group is asked to rank the three abstract on the following four questions

- Q. 1:
  Is the abstract grammatically natural?
- Q. 2:
  Is the Japanese easy to understand?
- Q. 3:
  Are sentences naturally connected with each other?
- Q. 4:
  Do you think the text is appropriate as an abstract?

The reason we adopt groups' opinions instead of individuals' ones lies in the awareness that examinees tend to be more responsible for the group they belong to, rather than the case when they behave as individuals. Table 5 shows the results of the assessment. Each figure in Table 5 indicates an average evaluation-value of the five groups for Q.1, Q.2, Q3 or Q4 towards one of the three abstracts.

$$aev = \frac{(x \times 3 + y \times 2 + z \times 1)}{5} \qquad (4)$$

An average evaluation value (*aev*) is calculated following Formula 4. Here, *x*, *y*, *z* indicates the number of groups that have assessed the abstract as the first place, second place, or third place respectively in regard to the corresponding question. A larger figure implies a better evaluation.

	A-abstract	***S-abstract***	***W-abstract***
Q. 1	2.8	1.2	1.4
Q. 2	2.6	1.4	1.6
Q. 3	2.6	1.8	1.6
Q. 4	2.4	2.2	1.2

Table 5. Results of the subjective assessment

As we have expected, the abstract written by the authors is the best for all the evaluation items. Also, our system seems to have shown the same or better performance than the summarization function of Microsoft Word 2003. Especially, our system achieves 2.2 for the question *do you think the text is appropriate as an abstract*,

which is almost the same with that from *A-abstract*.

However, there are still some problems remaining. In an interview with the examinees after the assessment, we have got some valuable comments such as "Pronouns are met too frequently" or "Too many long sentences exist in the abstract". In the following sub-section, we are going to make some discussions about these problems and try to conduct a validation.

6.3 Discussions

In regard to the issues observed by the examinees in the subjective assessment, we might have ways to adjust our approach. For example, we can skip the theme replacement function in abstract assembling described in Section 5.4, so that the total number of pronouns will decrease. On the other hand, to get a clearer look at the adequate length of a sentence in the abstract, we have conducted an investigation.

We have randomly selected 20 social scientific papers each with an abstract written by its original authors. Another abstract is produced by the system for each paper with the same number of sentences in the original abstract. The investigation is carried out by measuring the length (i.e., the total number of characters) of sentences in the original abstract, and that of the abstract generated by the system. Figure 7 and Figure 8 show their distributions.

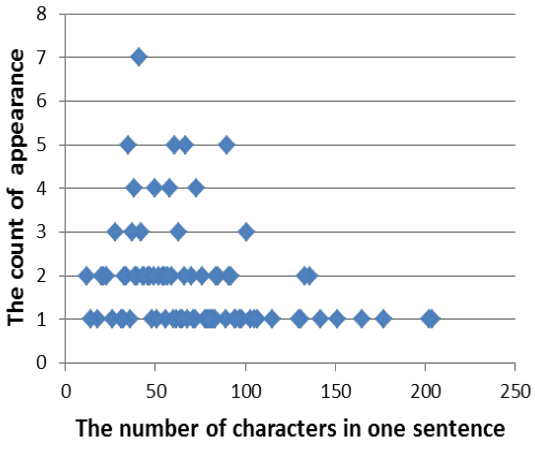

Figure 7. Distribution of the number of characters in original abstracts

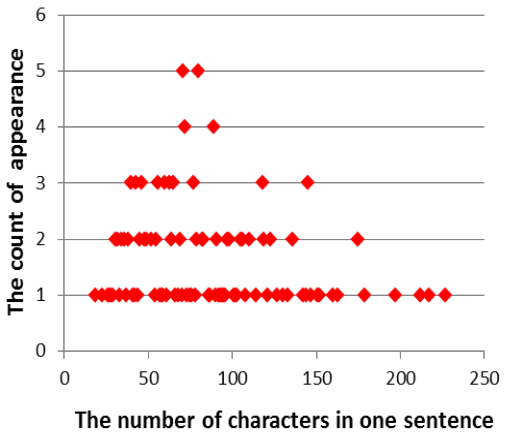

Figure 8. Distribution of the number of characters in abstracts generated by the system

The average numbers of characters in the original abstracts and the system-generated abstracts are 38.5 and 53.3 respctively. Moreover, The median value for the original abstracts is 64.5, whereas the median value for the abstracts generated by the system is 79.0. This might have been the reason of the unsatisfied results in Section 6.2 for Q.1 and Q.2. We could figure out some strategies to cope with this issue. For example, we can leave the cohesive relation out of our consideration when extracting constituent sentences, or just impose a restriction on the number of characters or segments when simplifying a sentence for the abstract.

7 Conclusion

In this paper, we propose a method to generate abstracts for social scientific papers. We put weight on three points: important keywords, readability as an abstract, and features of social scientific papers. Three main modules have been developed in our system to generate the abstract: sentence processing, importance degree estimation, and abstract generation.

Experimental results have shown the effectiveness of our proposal in comparison with another existing summarization tool, especially when we use character-based calculation to estimate the necessary number of constituents for abstract generation.

However, there is still room to improve. Results of an investigation on sentence length exhibit the future possibility to enhance our method and improve the quality of the abstract.

References

Church K. Ward and Hanks Patrick. 1990. Word Association Norms, Mutual Information, And Lexicography. Computational Linguistics, 16(1):22-29.

Fukushima Takahiro, Ehara Terumasa, and Shirai Katsuhiko. 2011. Partitioning long sentences for text summarization. Journal of Natural Language Processing, 6(6):131-147. (in Japaense).

Hatayama Mamiko, Matsuo Yoshihiro, and Shirai Satoshi. 2002. Summarizing Newspaper Articles Using Extracted Informative and Functional Words.Journal of Natural Language Processing, 9(4):55-73. (in Japaense).

Ichikawa Takasi. 1978. Kokugo Kyouiku No Tame No Bunsyouron Gaisetu. Kyouiku-shuppan. (in Japaense).

Imamura Hitoshi. 1988. Gendai Sisou Wo Yomu Ziten. Kodansha Ltd.(in Japaense).

Mikami Makoto, Masuyama Shigeru, and Nakagawa Seiichi. 1999. A Summarization Method by Reducing Redundancy of Each Sentence for Making Captions of Newscasting. Journal of Natural Language Processing, 6(6):65-81. (in Japaense).

Minami Hujio. 1974. Gendai Nihongo No Kouzou. Taishukan Publishing Co., Ltd.(in Japaense).

Morita Yosiyuki and Matsuki Masae. 1989. Nihongo Hyougen Bunkei Yourei Tyuusin Hukugouzi No Imi To Youhou. ALC.(in Japaense).

Nitta Yosio. 2002. Fukushiteki Hyowugen No Shosou. Kurosio Syuppan. (in Japaense).

Ochitani Ryo, Nakao Yoshio, and Nishino Fumihito. 1997. Goal-Directed Approach for Text Summarization. In Proc. of the ACL Workshop on Intelligent Scalable Text Summarization, 47-50.

Ohtake Kiyonori, Funasaka Takahiro, Masuyama Shigeru, and Yamamoto Kazuhide. 1999. Multiple Articles Summarization by Deleting Overlapped and Verbose Parts.Journal of Natural Language Processing, 6(6):45-64. (in Japaense).

Tomita Kohei, Takamura Hiroya, and Okumura Manabu. 2009. A New Approach of Extractive Summarization Combining Sentence Selection and Compression. IPSJ SIG Notes(NL).2009(2):13-20. (in Japaense).

Uchida Mituru, Imamura Hiroshi, Tanaka Aiji, Tanifuji Etsushi, and Yoshino Takashi. 2001. Dictionary of Contemporay Japanese Government and Politics. Brensyuppan(in Japaense).

Umesao Tadao, Kindaichi Haruhiko, Sakakura Atuyosi, and Hinohara Sigeaki. 1995. Nihongo Daiziten Kodansha kara ban dai2han. Kodansha Ltd. (in Japaense).