Construction of Text Summarization Corpus for the credibility of Information on the Web

Masahiro Nakano, Hideyuki Shibuki, Rintaro Miyazaki, Madoka Ishioroshi, Koichi Kaneko, Tatsunori Mori

Graduate School of Environment and Information Sciences Yokohama National University,
79-7 Tokiwadai, Hodogaya-ku, Yokohama 240-8501, Japan
{nakano, shib, rintaro, ishioroshi, kaneko, mori}@forest.eis.ynu.ac.jp

Abstract

Recently, the credibility of information on the Web has become an important issue. In addition to telling about content of source documents, indicating how to interpret the content, especially showing interpretation of the relation between statements appeared to contradict each other, is important for helping a user judge the credibility of information. In this paper, we will describe the purpose and the way in the construction of a text summarization corpus. Our purpose in the construction of the corpus includes the following three points; to collect Web documents relevant to several query sentences, to prepare gold standard data to evaluate smaller sub-processes in the extraction process and the summary generation process, to investigate the summaries made by human summarizers. The constructed corpus contains six query sentences, 24 manually-constructed summaries, and 24 collections of source Web documents. We also investigated how the descriptions of interpretation, which help a user judge the credibility of other descriptions in the summary, appear in the corpus. As a result, we confirmed that showing interpretation on conflicts is important for helping a user judge the credibility of information.

1. Introduction

Many pages on the Web contain incorrect or unverifiable information. Therefore, there is a growing demand for technologies that enable us to obtain reliable information. However, it would be almost impossible to automatically judge the accuracy of information presented on the Web. In this case, the second-best approach is to develop a supporting method that helps a user judge the credibility of information on the Web (Ando, et al, 2008; Murakami, et al, 2008; Kaneko, et al, 2009; Miyazaki, et al, 2009).

As one of the supporting methods, we attempt to develop an automatic summarization system (Kaneko, et al, 2009) that generates a survey report for helping a user verify the credibility of descriptions in Web documents relevant to a query sentence, which is a statement the user inputted such as "Are diesel engines harmful to the environment?"

If there are a Web document about "Diesel engines are harmful to the environment" and one about "Diesel engines are not harmful to the environment," existing systems for generating general purpose summary will make a summary from both documents, and it will just show two different texts extracted from these documents. However, the summary cannot make a user's judgment about the credibility of them easy because the systems do not explicitly show the contradiction that arises from both documents.

Note that some statements, which appeared to contradict each other at first glance, may be able to coexist under a certain situation. For example, two statements "Diesel engines are harmful to the environment because of more smog-forming oxides of nitrogen emissions" and "Diesel engines are not harmful to the environment because of lower carbon dioxide emissions" are not logically contradictory because they are described from different viewpoints, namely, air pollution and global warming.

Therefore, in addition to telling about content of source documents, indicating how to interpret the content, especially showing interpretation of the relation between statements appeared to contradict each other, is important for helping a user judge the credibility of information. In order to develop such systems, we require a corpus of summaries that contain description of such interpretation. Although there are previous studies on a corpus for summarization (Varasai, et al, 2008; Radev, et al, 2004), their corpora were not designed for summaries that contain descriptions of interpretation. Moreover the purpose of the multi-document summarization tasks in DUC[1], TAC[2] and TSC[3] is also different from supporting a user's judgment about the credibility of information. Therefore, we have to construct a text summarization corpus for the credibility of information.

In this paper, we will describe the purpose in the construction of the corpus and the specifications of the constructed corpus, and investigate the summaries made by human summarizers, especially in terms of descriptions of interpretation.

2. Survey report for the credibility of information

We define a survey report for the credibility of information as a summary that contains description of contents in source documents and ones of interpretation of relations between the contents. The generation of survey report consists of, at least, two processes: extracting important texts from Web documents and making summary text by arranging the extracted texts and adding description of relation between them. Note that some of descriptions of the relations may appear in Web documents, for example, ones in Q&A sites. In such case, the description may be used as a part of summary.

Figure 1 shows an example of a survey report we aim to generate automatically. The survey report consists of the following four parts. The first and second parts respectively show keywords and events relevant to the inputted query statement. The third shows opinions that are grouped into positive/negative clusters about the query

[1] http://duc.nist.gov
[2] http://www.nist.gov/tac/
[3] http://www.lr.pi.titech.ac.jp/tsc/index-en.html

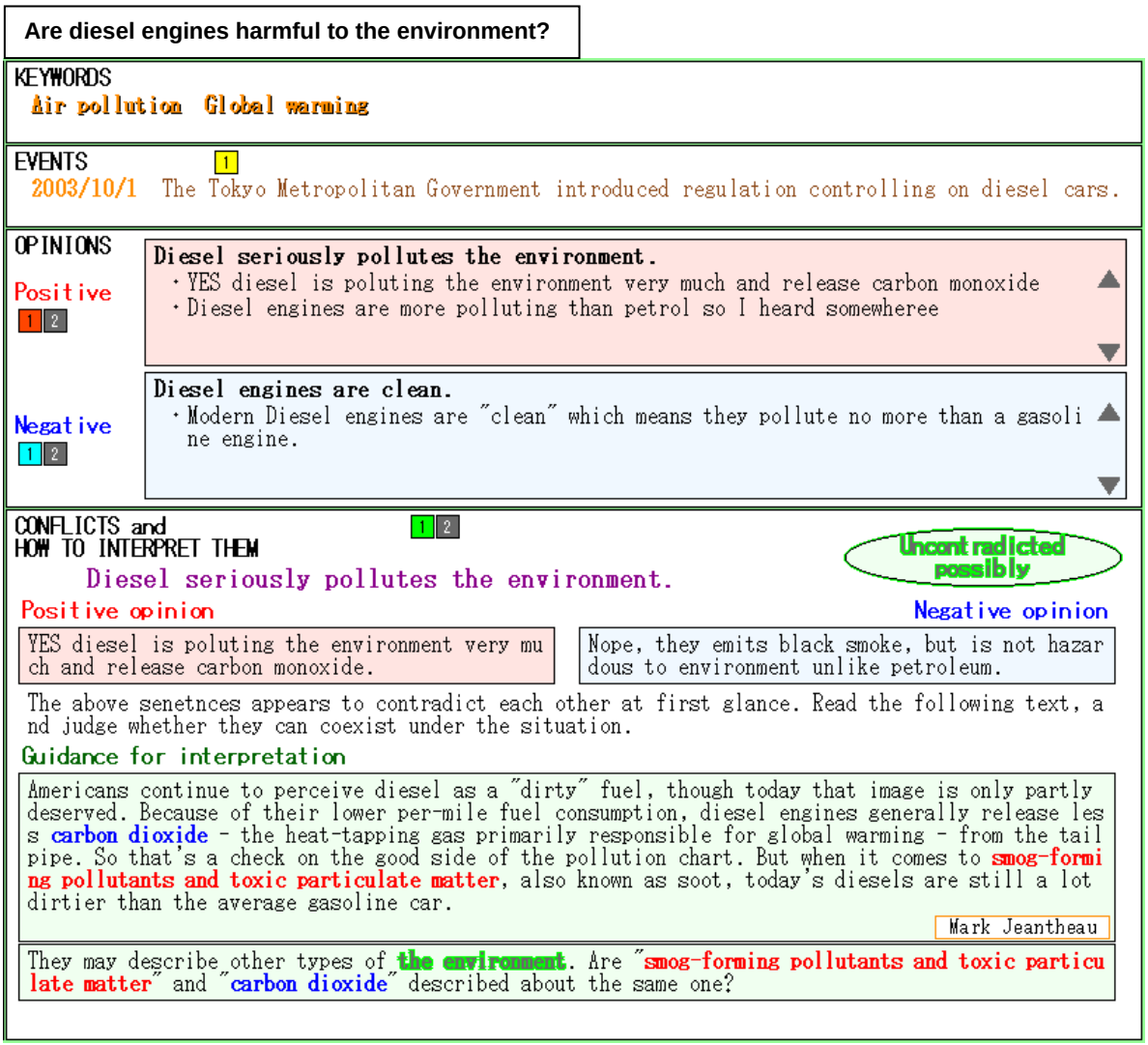

Figure 1: An example of the survey report

statement. These three parts help a user look over main contents of relevant documents on the Web. The fourth part shows pairs of positive/negative opinions and information to help a user interpret relations of the pairs. For example, if opinions appear to be able to coexist under a situation, contexts specifying the situation are described. Otherwise, reasons or evidences supporting each opinion are described.

This research is cooperated with other researches on the timeline analyzer and the statement map generator (Murakami, et al, 2009). The timeline analyzer can find events relevant to a statement from Web documents, and the statement map generator can find semantic relations between statements. The second and third parts in Figure 1 are respectively derived from results of the timeline analyzer and the statement map generator.

Figure 2 shows the outline of the system we attempt to develop in order to generate survey reports. An inputted query statement is passed to the passage retrieval module and the timeline analysis module. In the passage retrieval module, Web documents relevant to the query statement are retrieved via a Web search engine, and the documents are segmented into passages of adequate size as an input of the succeeding processes. Then, passages more relevant to the query statement are extracted and passed to the statement map generator. A name of information sender, which is a person or an organization that posts certain information on the Web (Miyazaki, et al, 2009), of each passage is also extracted. Finally the above results are arranged and summarized into a survey report.

Therefore, our purpose in the construction of the corpus includes the following three points. The first one is to collect Web documents relevant to several query sentences. The query sentences are chosen under the condition that there are opposing statements but there are actual situations in which the statements can coexist. The second one is to prepare gold standard data, or answer data, to evaluate smaller sub-processes in the extraction process and the summary generation process. The third one is to investigate the summaries made by human summarizers, especially in terms of descriptions of interpretation appearing in the corpus.

3. Construction of the corpus

We chose six query sentences: “Does xylitol prevent tooth decay?”, “Is asbestos harmful to human body?”, “Is LASIK operation safe?”, “Is LASIK operation painful?”,

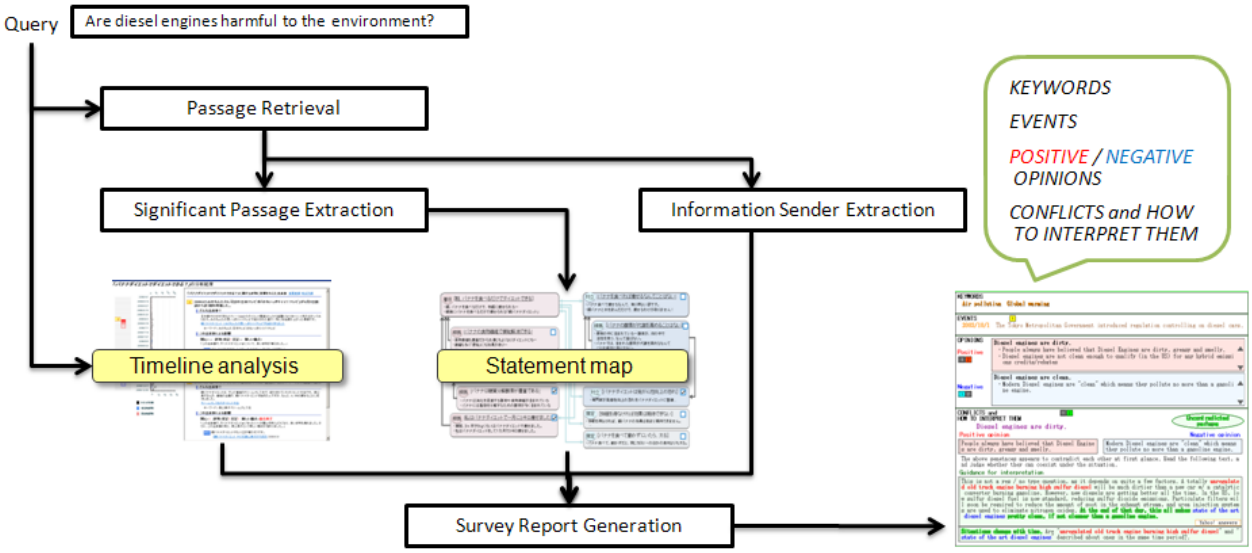

Figure 2 : Outline of the system for generating survey reports

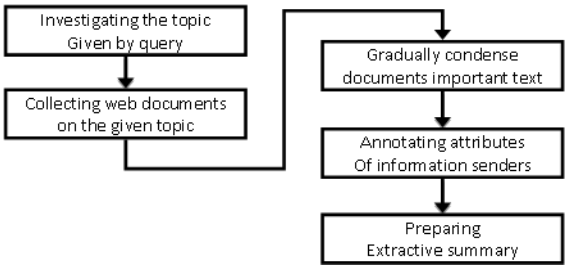

Figure 3 : Stages of the corpus construction

"Does non-washed rice not cause water pollution?" and "Is non-washed rice delicious?"

We instructed 12 human summarizers, who are students of a graduate school and an undergraduate school, about manual summarization as follows. One summarizer was given two query sentences, and was instructed to make a summary for each query sentence in order to support other persons in their judge whether the query sentence is true when they read the summary.

In this paper, we define the *importance* of text in documents as the contribution to such summaries as survey reports. We told summarizers that the summary should be mainly made of extracted parts of source documents. We did not instruct summarizers about how to connect parts and about the length of the summary because we want to investigate how summarizers insert their own expressions and connect parts with them in the summary. Since we would like to observe various strategies for supporting a user's judgment about the credibility of information, we did not explain anything about the description of interpretation. To guarantee the summaries to have good quality as gold standard data, four summarizers were assigned to each query sentence.

For the second purpose described in Section 2, we controlled a flow of manual summarization in terms of five stages as shown in Figure 3. The first stage involves summarizers' investigation of the topic related to a query sentences. The investigation includes searching for argument points, e.g. important sub-topics, in order to ensure balance of argument points and neutrality of the summary.

The second stage involves collecting Web documents so as to cover argument points as much as possible. The collected Web documents are stored query by query in the corpus.

In the first half of this stage, each summarizer freely made 20 search queries from the query statement and the result of investigation of the topic in the first stage. The summarizer filters out search queries that retrieve only irrelevant documents by using the search engine TSUBAKI. TSUBAKI is a search engine that can retrieve documents by using a natural sentence as a search query. Because we consider the process of summarization in this study as one of information access methods, collected documents have to contain argument points which are investigated in stage 1, as much as possible.

In the latter half on this stage, each summarizer constructs the collection of documents to be summarized. The summarizer makes a document set by collecting 100 documents for each search query using TSUBAKI. The summarizer gives document sets priority according to relevance of documents in the set to query statement. The summarizer, adds document sets to the collection of documents according to the priority until the number of document of the collection exceeds 500.

In the third stage, summarizers gradually condensed the documents into important smaller parts of text. The process of narrowing down important descriptions is carried out in three steps: extraction of important documents, extraction of important passages, and extraction of important character strings. Important character strings are used for making summaries and actually appear in summaries. The information of each step is annotated in Web documents by using XML tags as shown in Figure 5. The attribute selected of <File> tag

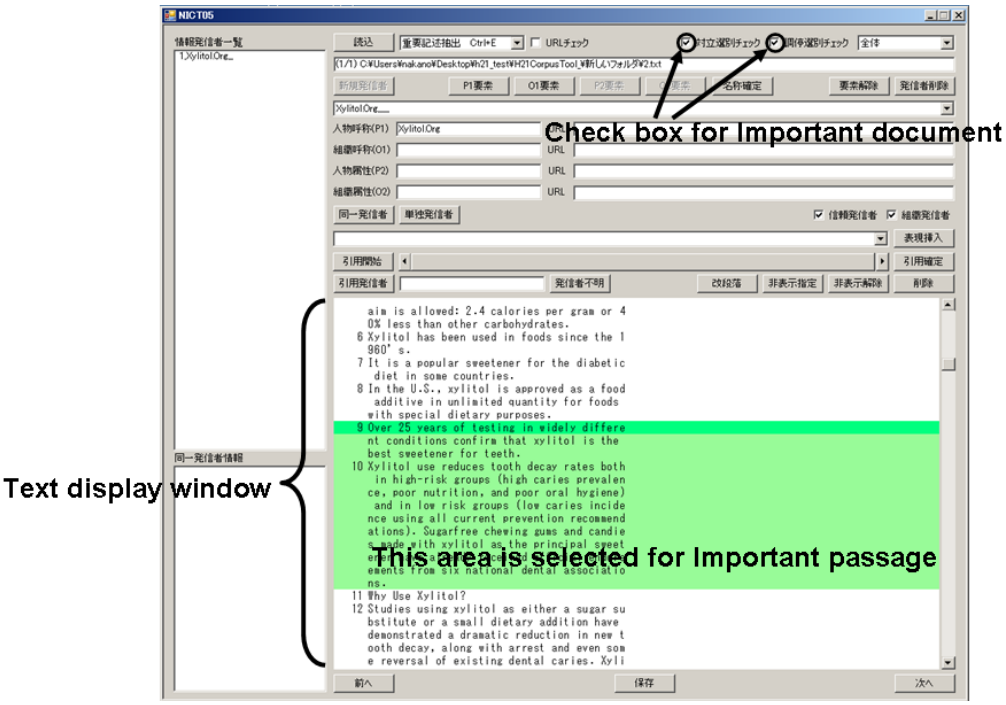

Figure 4 : The general view of the annotation tool

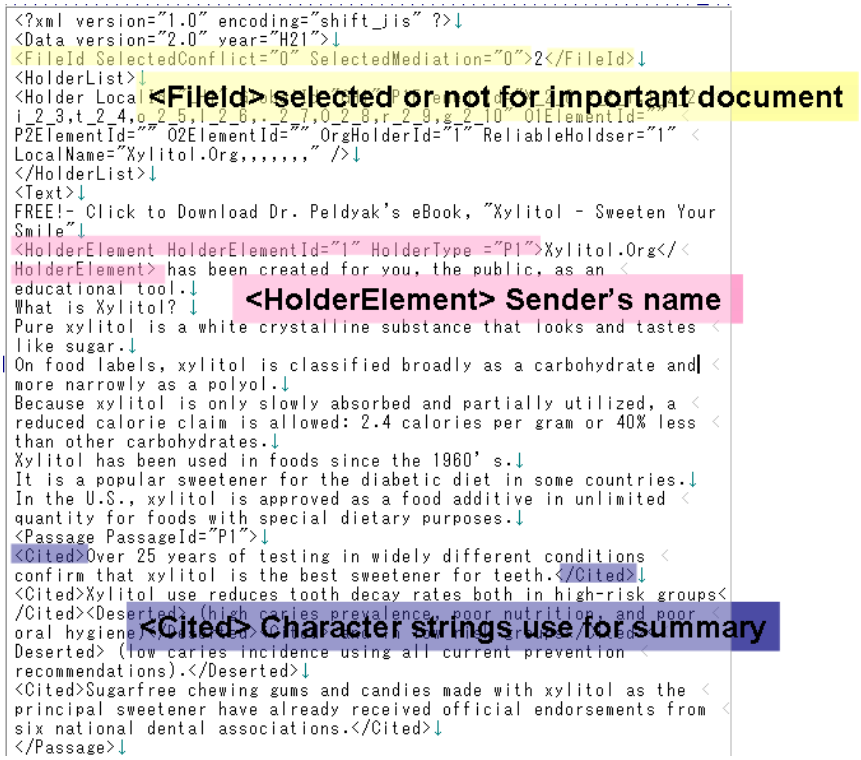

Figure 5 : Annotation in the corpus

indicates the binary importance of document. <Passage> and <Cited> tags indicate an important passage and an important character string, respectively. They are intended to be used for evaluation of the automatic extraction process.

In this and next stage, summarizers annotate with an annotation tool developed for this work. Summarizers were able to annotate easily with the tool through GUI. The over view of the tool showed in Figure 4. The target document of the annotation is displayed on center, and summarizers annotate by the operation of the check to check box and the mouse click.

The fourth stage involves the annotation of the name and the attributes of senders in important documents. The term *sender* refers to a person or an organization providing a description of certain information on the Web.<Holder> and <HolderElement> tags in Figure 5 indicate a name and an attribute of a sender, respectively. The information is

```
<?xml version="1.0" encoding="shift_jis" ?>↓
<Data version="1.2" year="H21">↓
<SurveyReport>↓
<Citation FileId="2" PassageId="P1" BeginPoint="0" HolderID="1" <
HolderName=" 1,Xylitol.Org,,,">Over 25 years of testing in widely <
different conditions confirm that xylitol is the best sweetener for </
Citation><Extra>health of</Extra><Citation FileId="2" PassageId="P1" <
BeginPoint="110" HolderID="1" HolderName=" 1,Xylitol.Org,,,">teeth.<
Xylitol use reduces tooth decay rates both in high-risk groups </<
Citation><Citation>and in low risk groups.</Citation><Citation> <
```

<Citation>
character strings extracted from source document

<Extra>
character strings inserted by summarizer

Figure 6 : annotation in the summary

	Average(hour)	Min~Max (hour)
Stage1	2.4	1.0~ 4.0
Stage2	4.2	2.0~ 8.5
Stage3	13.2	6.5~25.0
Stage5	8.1	4.0~11.0
Stage6	8.2	2.5~19.5
Total	35.6	23.0~47.0

Table 1 : working time of each stage

presented with sentences in survey reports. This annotation is also intended to be used for another study for the credibility of information (Miyazaki, et al, 2009) as gold standard data. In the final stage, a summary is generated for supporting the judgment about the credibility of the given query. The summary consists of two types of character strings: ones extracted from source documents and ones inserted between the extracted strings by the summarizer. We allowed the summarizers to insert any character strings in order to connect between extracted strings. The information is annotated in the summary as shown in Figure 6. <Citation> and <Extra> tags indicate an extracted character string and an inserted character string, respectively. Every summarizer recorded memoranda for perplexed parts of work and their treatments, stage by stage.

4. Analysis of the corpus

The constructed corpus contains six query sentences and Web documents and their summaries are stored in the query-by-query and summarizer-by-summarizer manner. Namely, the corpus contains the 24 collections of Web documents and 24 summaries. The average number of documents collected for each summary was 532.0, and the average number of characters in one summary was 2563.8. Since the average number of characters in the collection of documents for one summary was 2.8 million , the average summary rate was 0.1%. However, the summary rate is overestimated. The number of documents selected as important documents in stage 3 were only 177 on the average, and the number of characters extracted as important passage was 57,121. Therefore, the summary rate calculated by using the number of characters of important passage is 4.5%.

The average, minimum and maximum of working time that summarizers spent for each stages is shown in Table 1. 35.6 hours on the average were spent on one summary, and summarizers took longer time on the stages of T5 and T6 than other stages. Difference of working time among summarizers was especially large on the stage of T6 compared with other stages.

Here, we investigate whether we achieved our purposes described in Section 2.

Our first purpose is to collect Web documents relevant to several query sentences.

We studied the coverage of viewpoints, which are argument points that relate to the query statement. We investigated how the number of the view point included in retrieved documents increased as we increased the number of retrieved documents by using one search query,. Figure 7. shows the result. It has been found that the number of view points that relate to query tend to increase until the first 40 documents and it was saturated after 40 documents. According to the result, we decide that the number of retrieved document from the search engine is 100 for each query sentence in the summarization. As described in Section 3, each summarizer submitted multiple search queries to TSUBAKI. Therefore, the collection of documents may cover main view points, at least, for the

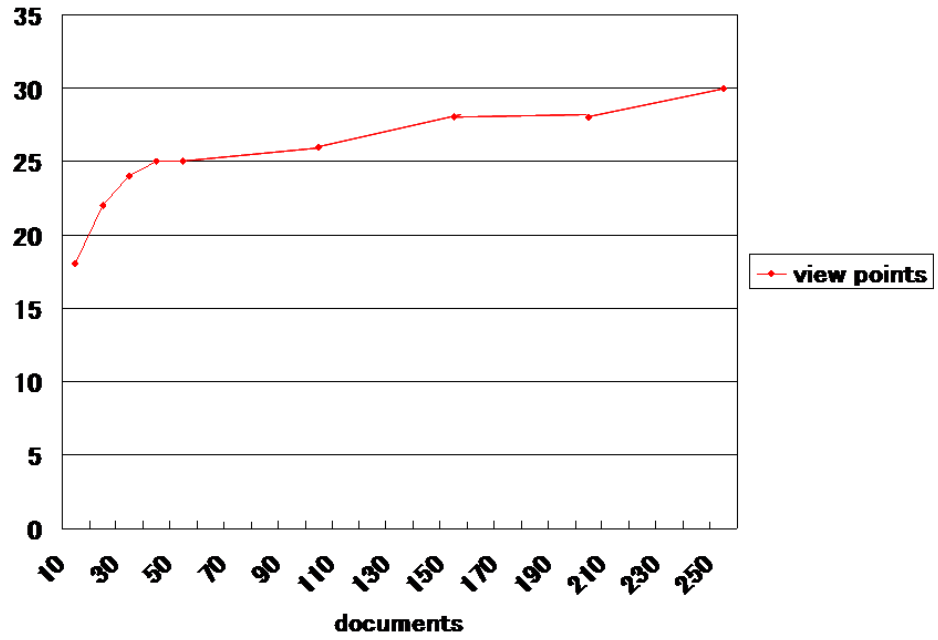

Figure 7 : The number of view points including retrieval documents

submitted search queries. We consider that our first purpose may be achieved to some extent.

Our second purpose is to prepare gold standard data, or answer data, to evaluate smaller sub-processes in the extraction process and the summary generation process. Summarizers gradually condensed the documents into important smaller parts of text in stage 3. The result of narrowing down is recorded by XML tags in the corpus. The information satisfies our second purpose.

We will describe analysis of the recorded data. First, we will describe the rate of agreement in the result of narrowing steps between two summarizers. The calculation of the agreement rate for the step of collecting documents is meaningless because the source of collection is the all of the documents on the Web and the agreement rate may be overestimated. Therefore, we calculated the rate of agreement for the step of important document selection. We adopt the κ value as the measure of the rate of agreements. The κ value is interpreted as the proportion of agreement among raters after chance agreement has been removed. The value ranges 0.4-0.6 and 0.6-0.8 of κ value are considered as moderate agreement and substantial agreement, respectively. The average of the κ value for the step of important document selection was 0.46, which represents moderate agreement.

Second, we will describe the agreement rate in the result of summarization between two summaries. We will use the following two measures to evaluate the rate of agreement. The first one is ROUGE-1, which is a measure of unigram overlap to evaluate the content overlap between summaries. The average value of ROUGE-1 is 0.40. The second one is the number of shared view points. About 20% of the view points are shared in all summaries. 60% of the view points are shared in at least two summaries.

Third, character strings that summarizers inserted are investigated. To distinguish inserted character strings from extraction of source documents in summaries, inserted

Table 2 : Number of types of character strings inserted by summarizer

Types of inserted strings	
Symbol	25
Conjunction,	55
Suffix of word	58
Sentence related to the topic	19
Others	29

character strings are annotated by <Extra> tag. The rate of characters inserted by the summarizer was 2.7% in summaries. Therefore, summarizers constructed summaries by using strings extracted from the source document as much as possible. Character strings are inserted to a summary 8.2 times on the average. The average length of inserted strings is 4.6 characters. .The number of types of inserted character strings is 136. With regard to number of token, the string “mata,(moreover,)” is most frequently inserted. It appears twelve times in 24 summaries.

Table 2 shows the classification of inserted character strings in terms of their grammatical function and meaning. About 75% of the inserted character strings are like the symbols, the conjunctions, and the suffixes of words Inserted symbols are classified to two types according to their functions. The first one is to clarify logical structure of summaries such as bullets in itemization. Another one is to modify sentences in terms of the orthography. 10% of the inserted character strings are sentences related to the topic, which are compiled from gathered fragments of source documents by the summarizer.

Our third purpose is to investigate the summaries, in terms of descriptions of interpretation appearing in the corpus.

Table 3 : Viewpoint patterns in the summaries

	without interpretation	with interpretation	
		implicit	explicit
A	1		
A+D		8	0
A+D+CA		2	2
A+D+CA+CAD		0	1
A+D+CD		1	0
A+D+CD+CAD		0	1
A+CA+CD		0	1
A+CA+CAD		0	1
D+CA		3	0
D+CA+CD		1	0
D+CA+CAD		0	2
Total	1	15	8

We investigated the descriptions of interpretation in summaries as follows. From the viewpoint of agree/disagree with the query sentences, we classified sentences in the summaries into six types: agreement (A), disagreement (D), conditional agreement (CA), conditional disagreement (CD), conditions of both agreement and disagreement (CAD) and others.

According to combination patterns of included sentence types except others, summaries are classified as shown in Table 3, where the number in each cell shows the number of summaries in the combination pattern. Summaries that contain both agreement and disagreement sentences can be regarded as summarizers' intention to include some interpretation on the conflict implicitly ("implicit" in the table). In addition to that, some summarizers introduced expressions that explicitly show some interpretation on the conflict ("explicit" in the table).

As shown in Table 3, almost all summaries include interpretations on the conflict in implicit or explicit ways. This result support our claim in Section 1 that showing interpretation on conflicts is important for helping a user judge the credibility of information.

With regards to conditional sentences, 15 out 24 summaries contain them. Showing conditions of agreement/disagreement is helpful for users to interpret a set of sentences that appeared contradictory. All of CAD sentences include explicit interpretations of conflicts. On the other hand, CA and CD sentences mostly contribute to express implicit interpretations of conflicts.

Moreover, sentences classified into others were described as definition of terms, reasons and examples relevant to the query statement, and so on. We consider that these descriptions should also be included in survey reports.

5. Conclusion

We described the purpose and the specifications of the constructed corpus, which contains six query sentences, 24 manually-constructed summaries, and 24 collections of source Web documents. We also investigated how the descriptions of interpretation, which help a user judge the credibility of other descriptions in the summary, appear in the corpus.

Our future work includes the following three topics: (1) to make new summaries under the condition that summarizers are explicitly instructed to include the descriptions of interpretation on conflicts, (2) to investigate how interpretation on conflicts appears in the summaries, and (3) to develop an automated summarization system that generates a survey report for helping a user verify the credibility.

Acknowledgment

This research is a part of the project "Evaluating Credibility of Web Information" of the National Institute of Information and Communications Technology (NICT), Japan and partially supported by NICT.

References

S. Ando, K. Inui, M. Ishioroshi, Y. Matsumoto, S. Matsuyoshi, R. Miyazaki, T. Mori, K. Murakami, M. Nakano, S. Nakazawa, Y. Okajima, H. Shibuki and T. Suzuki, Information Credibility Survey Reporting: A Prototype System and Project Roadmap, Proc. of ISUC 2008, 2008.

K. Murakami, E. Nichols, S. Matsuyoshi, A. Sumida, S. Masuda, K. Inui and Y. Matsumoto, Statement Map: Assisting Information Credibility Analysis by Visualizing Arguments, Proc. of WICOW 2009, pp.43-50, 2009.

K. Kaneko, H. Shibuki, M. Nakano, R. Miyazaki, M. Ishioroshi and T. Mori, Mediatory Summary Generation: Summary-Passage Extraction for Information Credibility on the Web, Proc. of PACLIC 23, 2009.

R. Miyazaki, R. Momose, H. Shibuki and T. Mori, Using Web Page Layout for Extraction of Sender Names, Proc. of IUCS 2009, 2009.

P. Varasai, C. Pechsiri, T. Sukvari, V. Satayamas and A. Kawtrakul, Building an Annotated Corpus for Text Summarization and Question Answering, Proc. of LREC 2008, 2008.

D. Radev, J. Otterbacher and Z. Zhang, CST Bank: A corpus for the Study of Cross-document Structural Relationships, Proc. of LREC 2004, 2004.