Chapter 3

Text Summarization Challenge: An Evaluation Program for Text Summarization

**Hidetsugu Nanba, Tsutomu Hirao, Takahiro Fukushima,
and Manabu Okumura**

Abstract In Japan, the Text Summarization Challenge (TSC), the first text summarization evaluation of its kind, was conducted in 2000–2001 as a part of the NTCIR (NII-NACSIS Test Collection for IR Systems) Workshop. The purpose of the workshop was to facilitate collecting and sharing text data for summarization by researchers in the field and to clarify the issues of evaluation measures for summarization of Japanese texts. After that, TSC has been held every 18 months as a part of the NTCIR project. In this chapter, we describe our TSC series, the data used, and the evaluation methods for each task, and the features of TSC evaluation.

3.1 What is Text Summarization?

The ever-growing amount of information forces us to read through a great number of documents in order to extract relevant information from them. To cope with this situation, research on text summarization has attracted much attention recently, producing many studies in this field.¹

¹Many survey papers are now available on text summarization, e.g., Gambhir and Gupta (2017), Allahyari et al. (2017), Nazari and Mahdavi (2019).

H. Nanba

Faculty of Science and Engineering, Chuo University, Tokyo, Japan

e-mail: nanba@kc.chuo-u.ac.jp

T. Hirao

NTT Communication Science Laboratories, Kyoto, Japan

e-mail: tsutomu.hirao.kp@hco.ntt.co.jp

T. Fukushima

Department of International Liberal Arts, Otemon Gakuin University, Ibaraki, Japan

e-mail: fukusima@otemon.ac.jp

M. Okumura (✉)

Tokyo Institute of Technology, Tokyo, Japan

e-mail: oku@pi.titech.ac.jp

As research on text summarization is a hot topic in Natural Language Processing (NLP), we also see the needs to discuss and clarify issues of how to evaluate text summarization systems. In Japan, the Text Summarization Challenge (TSC), the first text summarization evaluation of its kind, was conducted in 1999–2000 as a part of the NTCIR (NII-NACSIS Test Collection for IR Systems) Workshop. The aim of TSC was to facilitate collecting and sharing text data for summarization by researchers in the field and to clarify the issues of evaluation measures for summarization of Japanese texts.

Since that time, TSC² was held twice more, every 18 months, as a part of the NTCIR project. Multiple document summarization as one of the tasks was included for the first time at the TSC2 in 2002.

As we mention in Sect. 3.5, the contributions of our TSC can be considered as follows:

- We proposed a new evaluation method, evaluation by revision, that evaluates summaries by measuring the degree of revisions of the system results.
- We proposed a new evaluation method for multiple documents summarization that enables us to measure the effectiveness of redundant sentence reduction in the systems.

In the following sections, we first introduce the types of summarization and the evaluation methods in general. Then, we describe our TSC series, the data used, and the evaluation methods for each task. Finally, we summarize the contributions of the TSC evaluations.

3.2 Various Types of Summaries

Text summarization is a task of producing a shorter text from the source, while keeping the information content of the source. Summaries are the results of such a task. Perhaps, one of the most widely used summaries in the world today is the snippets that Web search engines display for each Web page. Sparck Jones (1999) discussed several ways to classify summaries. The following three factors are considered to be important for text summarization research:

Input factors: text length, genre, and single versus multiple documents,

Purpose factors: who the user is, and the purpose of summarization,

Output factors: running text or headed text, etc.

Summaries can be classified with respect to the number of the source texts (single document versus multiple document summarization), and with respect to whether they are tailored to particular users. Early research in summarization was primarily based on single-document summarization, in which systems produced a summary from a single-source document. However, another task has been later introduced into

²<http://www.lr.pi.titech.ac.jp/tsc/index-en.html>.

text summarization, that is based on multiple source documents. In multi-document summarization, several documents sharing a similar topic are taken as the input. The task of multi-document summarization can be considered more difficult than the single-document one, because the systems would need to remove any redundancies across multiple documents and then make the contents from multiple documents into a coherent summary.

If summaries are targeted for specific users, they are called user focused, and if they are intended for users in general, they are called generic. Query-focused summaries are another name for user focused summaries. In query-based summarization, the summary is generated by selecting sentences that correspond to the user's query (Tombros and Sanderson 1998). Sentences that are relevant to the query have a higher chance to be extracted for the final summary. In terms of summarization purpose, summaries can be either indicative or informative. Users can make use of indicative summaries before referring to the source, e.g., to judge relevance of the source text. On the other hand, users may use summaries in place of the source text (informative summaries). The snippets of Web search engines are a good example of indicative and query-focused summaries.

As pointed out by Mani and Maybury (1999), summaries can be also classified into extracts and abstracts, depending on how they are composed. Conventional text summarization systems produce summaries by using sentences or paragraphs as a basic unit, giving them a degree of importance, sorting them based on the importance, and gathering the important sentences. In short, summaries that are constructed of a set of important sentences extracted from the source text are called extracts. In contrast, summaries that may contain newly produced texts are called abstracts. Therefore, abstractive summarization can be much more complex than extractive summarization.

3.3 Evaluation Metrics for Text Summarization

Evaluation methods for text summarization can be largely divided into two categories: intrinsic and extrinsic. The quality of summaries can be judged directly based on some norms; typically, ideal summaries are produced by hand, or important sentences are selected by hand. Then, the quality of summaries is evaluated by comparing them with the human-produced summaries (intrinsic evaluation). The quality of a summary can also be judged by measuring how it influences the achievement of some other task (extrinsic evaluation). Mani and Maybury (1999) stated such tasks can be question-answering, reading comprehension, as well as relevance judgement of a document to a certain topic indicated by a query.

Relevance judgement: determines whether it is possible to judge whether the presented document is relevant to a user's topic, that can be indicated by her query, by reading the summary.

Reading comprehension: determines whether it is possible to correctly complete a multiple-choice test after reading the summary.

There are two measures for intrinsic evaluation: Quality and informativeness (Gambhir and Gupta 2017). The first measure checks the summary for grammatical errors, redundant information, and structural coherence. Here, the linguistic aspects of the summary are considered. In the Document Understanding Conference (DUC) and Text Analysis Conference (TAC), five questions based on linguistic quality are employed for evaluating summaries, which are non-redundancy, focus, grammaticality, referential clarity, and structure and coherence. Human assessors evaluate summaries manually by assigning a score to the summary, on a five-point scale.

For intrinsically evaluating the informativeness of a summary, the most popular metrics are precision, recall, and F-measure; they measure the overlap between human-made summaries and automatically generated machine-made summaries.

Precision: determines what fraction of the sentences selected by the system are correct.

Recall: determines what proportion of the sentences chosen by humans are selected by the system.

F-measure: is computed by combining recall and precision.

3.4 Text Summarization Evaluation Campaigns Before TSC

The first conference where text summarization systems were evaluated was held at the end of the 90's and was named the TIPSTER Text Summarization Evaluation (SUMMAC) (Mani and Maybury 1999). At that time, text summaries were evaluated using two extrinsic and one intrinsic methods. Two main extrinsic evaluation tasks were defined: *adhoc* and *categorization*. In the *adhoc* task, the focus was on indicative summaries which were tailored to a particular topic, and they were used for relevance judgement. In the *categorization* task, the evaluation sought to find out whether a generic summary could effectively present enough information to allow people to quickly and correctly categorize a document. The final task, a question-answering task, involved an intrinsic evaluation, where a topic-related summary for a document was evaluated in terms of its "informativeness".

Another important conference for text summarization was DUC, which was held every year from 2001 to 2007 (Gambhir and Gupta 2017). All editions of this conference contained newswire documents. Initially, in DUC-2001 and DUC-2002, the tasks involved generic summarization of single and multiple documents; they later extended to query-based summarization of multiple documents in DUC-2003. In DUC-2004, topic-based single and multi-document cross-lingual summaries were evaluated.

3.5 TSC: Our Challenge

Another evaluation program, NTCIR, formed a series of three Text Summarization Challenge (TSC) workshops—TSC1 in NTCIR-2 from 2000 to 2001, TSC2 in NTCIR-3 from 2001 to 2002, and TSC3 in NTCIR-4 from 2003 to 2004. These workshops incorporated summarization tasks for Japanese texts. The evaluation was done using both extrinsic and intrinsic evaluation methods.

3.5.1 TSC1

In TSC1, newspaper articles were used, and two tasks for a single article with intrinsic and extrinsic evaluations were performed (Fukushima and Okumura 2001; Nanba and Okumura 2002). We used newspaper articles from the Mainichi newspaper database of 1994, 1995, and 1998. The first task (Task A) was to produce summaries (extracts and free summaries) for intrinsic evaluation. We used recall, precision, and F-measure for evaluation of the extracts, and content-based as well as subjective methods for the evaluation of free summaries. The second task (Task B) was to produce summaries for the information retrieval task. The measures for evaluation were recall, precision, and F-measure for the correctness of the task, as well as the time that it takes to carry out the task. We also prepared human-produced summaries for the evaluation. In terms of genre, we used editorials and business news articles in the TSC1 dry-run evaluation, and editorials and articles on social issues in the formal run evaluation. As shareable data, we gathered summaries not only for the TSC evaluation but also for the researchers to share. By spring 2001, we collected summaries of 180 newspaper articles. For each article, we had the following seven types of summaries: important sentences (10, 30, 50%), summaries created by extracting important parts in sentences (20, 40%), and free summaries (20, 40%).

The basic evaluation design of TSC1 was similar to that of SUMMAC. The differences were as follows:

- As the intrinsic evaluation in Task A, we used a ranking method in subjective evaluation for four different summaries (baseline system results, system results, and two kinds of human summaries).
- Task B was basically the same as one of the SUMMAC extrinsic evaluations (the adhoc task), except the documents were in Japanese.

The following points were some of the features of TSC1. For Task A, we used several summarization rates and prepared the texts of various lengths and genres to use for evaluations. Their lengths varied at 600, 900, 1200, and 2400 characters, and the genres included business news, social issues, as well as editorials. As for Task A, because it was difficult to perform intrinsic evaluation on informative summaries, we presented the evaluation results as materials for discussions, at NTCIR workshop 2.

3.5.2 TSC2

TSC2 had two tasks (Okumura et al. 2003): single-document summarization (Task A) and multi-document summarization (Task B). In Task A, we asked the participants to produce summaries in plain text to be compared with human-prepared summaries from single texts. This task was the same as Task A in TSC1. In Task B, more than one (multiple) texts were summarized for the task. Given a set of texts, which has been manually gathered for a pre-defined topic, the participants produced summaries of the set in plain text format. The information that was used to produce the document set such as queries and summarization lengths were also given to the participants.

We used newspaper articles of the Mainichi newspaper database from 1998 and 1999. As the gold standard (human prepared summaries), we prepared the following types of summaries:

Extract-type summaries: We asked annotators, captioners who were well experienced in summarization, to select important sentences from each article.

Abstract-type summaries: We asked the annotators to summarize the original articles in two ways. First, to choose important parts of the sentences in extract-type summaries. Second, to summarize the original articles freely without worrying about sentence boundaries and trying to obtain the main ideas of the articles.

Both types of abstract-type summaries were used for Task A. Both extract-type and abstract-type summaries were made from single articles.

Summaries from more than one article: Given a set of newspaper articles that has been selected based on a certain topic, the annotators produced free summaries (short and long summaries) for the set. Topics varied from a kidnapping case to the Y2K problem.

We used summaries prepared by humans for evaluation. The same two intrinsic evaluation methods were used for both tasks. They were evaluated by ranking the summaries and by measuring the degree of revisions.

Evaluation by ranking: This is basically the same method as the one we used for Task A in TSC1 (subjective evaluation). We asked human judges, who are experienced in producing summaries, to evaluate and rank the system summaries from two points of views:

1. **Content:** How much the system summary covers the important content of the original article?
2. **Readability:** How readable the system summary is?

Evaluation by revision: It was a newly introduced evaluation method in TSC2 to evaluate summaries by measuring the degree of revisions of the system results. The judges read the original texts and revised the system summaries in terms of content and readability. The revisions were made by only three editing operations (insertion, deletion, and replacement). The degree of the human revisions, which we call “edit distance”, was computed from the number of revised characters divided by the number of characters in the original summary. As a baseline for

Task A, human-produced summaries, as well as lead-method results, were used. Also, as a baseline for Task B, human-produced summaries, lead-method results, and the results based on the Stein method (Stein et al. 1999) were used. The lead-method extracts a few first sentences of news articles. The procedure of the Stein method is roughly as follows:

1. Produce a summary for each document.
2. Group the summaries into several clusters. The number of clusters is adjusted to be less than the half of the number of the documents.
3. Choose the most representative summary as the summary of the cluster.
4. Compute the similarity among the clusters and output the representative summaries in such order that the similarity of neighboring summaries is high.

We compared the evaluation by revision with the ranking evaluation, which is a manual method used in both TSC1 and TSC2. To investigate how well the evaluation measure recognizes slight differences in the quality of the summaries, we calculated the percentage of cases where the order of edit distance of two summaries matched the order of their ranks given by the ranking evaluation by checking the score from 0 to 1 at 0.1 intervals. As a result, we found that the evaluation by revision is effective for recognizing slight differences between computer-produced summaries (Nanba and Okumura 2004).

3.5.3 TSC3

In a single document, there are few sentences with the same content. In contrast, in multiple documents with multiple sources, there are many sentences that convey the same content with different words and phrases, or even with identical sentences. Thus, a text summarization system needs to recognize such redundant sentences and reduce this redundancy in the output summary.

However, we have no ways of measuring the effectiveness of such methods of reducing redundancy in the corpora for DUC and TSC2. The gold standard in TSC2 was given as abstracts (free summaries) with the number of characters less than a fixed number. It was therefore difficult to use for repeated or automatic evaluation and for the extraction of important sentences. Moreover, in DUC, where most of the gold standard was abstracts with the number of words less than a fixed number, the situation was the same as in TSC2. At DUC 2002, extracts (important sentences) were used, and this allowed us to evaluate sentence extraction. However, it was not possible to measure the effectiveness of redundant sentence reduction because the corpus was not annotated to show sentences with the same content.

Because many of the current summarization systems for multiple documents were based on sentence extraction, in TSC3, we assumed that the process of multiple

document summarization should consist of the following three steps. We produced a corpus for evaluating the system at each of these three steps³ (Hirao et al. 2004).

- Step 1 Extract important sentences from a given set of documents,
- Step 2 Minimize redundant sentences from the results of Step 1,
- Step 3 Rewrite the results of Step 2 to reduce the size of the summary to the specified number of characters or less.

We have annotated not only the important sentences in the document set, but also those among them that have the same content. These are the corpora for Steps 1 and 2. We have prepared human-produced free summaries (abstracts) for Step 3. We constructed extracts and abstracts of thirty sets of documents drawn from the Mainichi and Yomiuri newspapers published between 1998 to 1999, each of which was related to a certain topic.

In TSC3, because we had the gold standard (a set of correct important sentences) for Steps 1 and 2, we conducted automatic evaluation using a scoring program. We adopted *intrinsic* evaluation by human judges for Step 3. Therefore, we used the following *intrinsic* and *extrinsic* evaluation. The *intrinsic* metrics were “Precision”, “Coverage”, and “Weighted Coverage.” The *extrinsic* metric was “Pseudo Question-Answering,” i.e., whether a summary has an “answer” to the question or not. The evaluation was inspired by the question-answering task in SUMMAC. Please refer to Hirao et al. (2004) for more details of the intrinsic metrics.

3.6 Text Summarization Evaluation Campaigns After TSC

In DUC-2005 and DUC-2006, multi-document query-based summaries were evaluated whereas in DUC-2007, multi-document update query-based summaries were evaluated. These conferences also provided standard corpora of documents and gold summaries.

After 2007, DUC was succeeded by TAC, in which summarization tracks were presented (Gambhir and Gupta 2017). The 2008 summarization track consisted of two tasks: update task and opinion pilot. The update summarization task aimed to produce a short summary (around 100 words) from a collection of news articles, assuming that the user has already gone through a collection of previous articles. The opinion pilot task aimed to produce summaries of opinions from blogs. The 2009 summarization track had two tasks: update summarization, which was the same as in 2008, and Automatically Evaluating Summaries of Peers (AESOP). AESOP was a new task that was introduced in 2009; AESOP computes a summary’s score with respect to a particular metric that is related to the summary’s content, such as overall responsiveness and pyramid scores. The 2010 summarization track had two tasks: guided summarization and AESOP. The guided summarization task aimed

³This is based on a general idea of a summarization system and is not intended to impose any conditions on a summarization system.

to generate a 100-word summary from a collection of 10 news articles pertaining to a specific topic; each topic belongs to a previously defined category. The 2011 summarization track consisted of three tasks: guided summarization, AESOP, and multilingual pilot.

3.7 Future Perspectives

We described our TSC series, the data used, and the evaluation methods for each task, and the features of TSC evaluation. As we mentioned in Sect. 3.5, the contributions of our TSC can be considered as follows:

- We proposed a new evaluation method, evaluation by revision, that evaluates summaries by measuring the degree of revisions of the system results.
- We proposed a new evaluation method for multiple document summarization that enables us to measure the effectiveness of redundant sentence reduction in the systems.

More than 15 years have passed since our last evaluation challenge. Today, the text summarization field has changed a lot in that a huge amount of summarization data is now available in the field and neural models have prevailed and dominated the field. While we now have a variety of large summarization datasets such as Gigaword Corpus, New York Times Annotated Corpus, CNN/Daily Mail dataset, and NEWSROOM dataset (Grusky et al. 2018), it becomes difficult to compare systems on the datasets, against our expectations, because we do not necessarily have a standard dataset to compare them with. Even for the same dataset, the performance might change depending on differently sampled test data. Therefore, we can say that the current evaluation of summarization systems might not necessarily be reliable. In the future, we should construct a good standard dataset, against which we could compare summarization systems. For this purpose, it is necessary to investigate the properties of a variety of datasets that will enable us to sample test data to create a good evaluation dataset.

References

- Allahyari M, Pouriyeh S, Assef M, Safaei S, Trippe ED, Gutierrez JB, Kochut K (2017) Text summarization techniques: a brief survey. *Int J Adv Comput Sci Appl (IJACSA)* 8(10)
- Fukushima T, Okumura M (2001) Text summarization challenge: text summarization evaluation in Japan. In: *Proceedings of NAACL workshop on automatic summarization*, pp 51–59
- Gambhir M, Gupta V (2017) Recent automatic text summarization techniques: a survey. *Artif Intell Rev* 47(1):1–66
- Grusky M, Naaman M, Artzi Y (2018) Newsroom: a dataset of 1.3 million summaries with diverse extractive strategies. In: *Proceedings of the 2018 conference of the north american chapter of the association for computational linguistics: human language technologies*, vol. 1 (long papers), pp 708–719

- Hirao T, Fukusima T, Okumura M, Nobata C, Nanba H (2004) Corpus and evaluation measures for multiple document summarization with multiple sources. In: Proceedings of the 20th international conference on computational linguistics, pp 535–541
- Mani I, Maybury M (eds) (1999) Advances in automatic text summarization. MIT Press, Cambridge
- Nanba H, Okumura M (2002) Some examinations of intrinsic methods for summary evaluation based on the text summarization challenge (TSC). In: Proceedings of the 3rd international conference on language resources and evaluation (LREC2002), pp 739–746
- Nanba H, Okumura M (2004) Comparison of some automatic and manual methods for summary evaluation based on the text summarization challenge 2. In: Proceedings of the 4th international conference on language resources and evaluation, pp 1029–1032
- Nazari N, Mahdavi MA (2019) A survey on automatic text summarization. J AI Data Min 7(1):121–135. <https://doi.org/10.22044/JADM.2018.6139.1726>
- Okumura M, Fukushima T, Nanba H (2003) Text summarization challenge 2 text summarization evaluation at NTCIR workshop 3. In: Proceedings of the HLT/NAACL 2003 workshop on text summarization, pp 49–56
- Sparck Jones K (1999) Automatic summarizing: factors and directions. In: Mani I, Maybury M (eds) Advances in automatic text summarization. MIT Press, Cambridge, pp 1–12. <http://xxx.lanl.gov/ps/cmp-lg/9805011>
- Stein G, Strzalkowski T, Wise G (1999) Summarizing multiple documents using text extraction and interactive clustering. In: Proceedings of the PACLING'99, pp 200–208
- Tombros A, Sanderson M (1998) Advantages of query biased summaries in information retrieval. In: Proceedings of the 21st annual international ACM-SIGIR conference on research and development in information retrieval, pp 2–10

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.