

Multi-Document Summarization Using Cross-Language Texts

Jung-Min Lim, In-Su Kang, Jong-Hyeok Lee

Div. of Electrical and Computer Engineering

Pohang University of Science and Technology (POSTECH)

Advanced Information Technology Research Center (AITrc)

San 31, Hyoja-dong, Nam-gu, Pohang, 790-784, Republic of Korea

{beuett, dbaisk, jhlee}@postech.ac.kr

Fax: +82-54-279-5699

Abstract

Without a summarization system in source language, we try to generate a summary in source language, using translated documents by a machine translator and a summarization system in target language. For summarizing multiple documents translated by a machine translator, we extract important sentences, and remove redundant sentences using an improved term-weighting method. It assigns weights to words, using syntactic information. According to the score of the extracted sentence, we choose sentences, and map them to Japanese sentences in original documents. Finally, we arrange Japanese sentences in chronological order, and report them as the result of our system. We submitted both a short and long type of summary, and the evaluation of our results is not good. However, our approach shows the possibility of multi-documents summarization, using cross-language texts.

Keywords: Multi-document summarization, Translated documents, Redundancy measuring, Sentence extraction.

1. Introduction

In NTCIR Text Summarization Challenge (TSC) 3, all documents given to participants are written in Japanese, and the result of systems also should be written in Japanese. Although we do not have a Japanese summarization system, we have developed a Korean summarization system, and a Japanese-to-Korean machine translator (MT). We cannot directly summarize Japanese documents. However we translate original Japanese documents to Korean ones, produce a sentence-level summary in target language, and generate Japanese summaries by mapping extracted sentences to original sentences. Generating a summary by cross-language texts, we could participate in TSC 3. Thus, we focused our goal to produce a summary in a source language, using multiple target documents translated by an MT. We use a rule-

based Japanese-to-Korean MT in TSC 3. However, if we have bilingual corpus and use a statistical MT, Our approach can be applied to generate a source language summary without a summarization system of a source language.

In the following sections, we first provide a system description, focusing on the method of extracting sentences and removing redundant sentences. Then, we discuss the evaluation results, and conclude this paper.

2. System overview

Figure 1 shows the overall architecture of our summarization system.

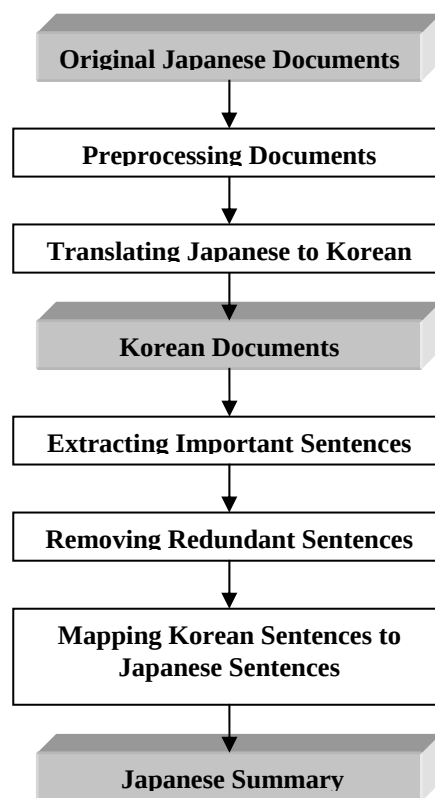


Figure 1. System Architecture

To summarize multiple documents written in Japanese, we first translate Japanese documents into Korean ones by an MT. Next, we summarize Korean documents, based on extracting sentences. For extracting sentences, we use well-known traditional features: sentence position, length, stigma term, keywords in a headline, and global terms in whole documents. For removing redundant sentences, we develop an improved term-weighting strategy. It assigns weights to terms that have syntactically important role in a sentence. For generating a Japanese summary, we map Korean sentences extracted by our system to Japanese sentences in original documents with a sentence-ID, and a document-ID. A summarization system usually contains a module that shortens extracted sentences or revises generated sentences. However we generate a summary by mapping translated sentences to original sentences, and do not have linguistic knowledge for revising Japanese sentences, thus we cannot implement it in our approach.

2.1 Preprocessing documents

Before translating the original documents, we remove irrelevant sentences that describe photographs, and contain the name of a newspaper or writer. We also remove some special characters, such as ■, ◆, ▲ in sentences.

2.2 Translating Japanese to Korean

To translate the original Japanese document to Korean, we use COBALT-JK (Collocation-Based Language Translator from Japanese to Korean) [1]. It is a Japanese-to-Korean machine translation (MT) system that has been developed by Knowledge and Language Engineering (KLE) Lab of POSTECH. It adopts a direct machine translation approach, and contains a bilingual dictionary (Japanese to Korean) with 178,300 entries from a general vocabulary, and 134,500 entries of person and place nouns. When we translate Japanese document to Korean ones, we can obtain useful information for summarizing documents without other efforts. The syntactic role and concept code of each word are obtained and unknown words are distinguished by an MT. This information is able to contribute on the phase of extracting sentences or removing redundancy.

2.3 Extracting important sentences

To identify important contents in multiple documents, we use two types of extraction method. First, we use an extraction method for single document summarization. In TSC3, target documents consist of newspaper articles. Thus, we mainly rely on sentence position. Also, for identifying the significance of sentences, we use sentence length, stigma terms, and

lead words in a headline and lead sentences in the phase of extracting sentences [2].

Second, considering multiple documents, we construct a global term cluster [3]. It is used to identify which sentences are central to the topic of multiple documents rather than individual articles. Combining both single and multiple document summarization method, we extract important sentences in each document. To consider multiple newspaper documents, we assign higher weights to a position-score, and a global term cluster. The score of an i -th sentence S_i is calculated as follows.

$$S(S_i) = w_{position} \times S_{pos}(S_i) - Pen(S_i) - w_{stigma} \times |S_i \cap P| + w_{lead} \times |S_i \cap L| + w_{gts} \times \sum_{t \in S_i \cap G} f_t \quad (1)$$

Where

- $S_{pos}(S_i)$: a position score of S_i
- $Pen(S_i)$: a length penalty of S_i
- L : a set of lead words
- P : a set of stigma terms
- G : a global word set
- f_t : the frequency of a word t in G

In formula (1), $w_{position}$, w_{stigma} , w_{lead} , and w_{gts} are weights of a position score, stigma terms, lead terms, and global terms, respectively. In our system, we select 40% of sentences that have a higher score at each document, and use them as the input of the next phase. The following subsections describe each weighting scheme.

2.3.1 Sentence position

A sentence position was used to find important contents since the late 1960s [4]. We assign a score to sentences according to their position. A sentence that is located at the beginning of the document is given a higher score than others. However, some important contents can be found at the end of a document. Thus, we also give additional weights to sentences that are located at the end of a document [5]. We assume that sentences in the middle are not important, and the difference of their score does not critically depend on their position. Thus, we assign the same value 0.5 for sentences in the middle of a document. When the number of sentence is n in a document, the position-score $S_{pos}(S_i)$ of a i -th sentence S_i is calculated as follows.

$$S_{pos}(S_i) = \max\left(1 - \frac{(i-1)}{n}, 1 - \frac{(n-i+1)}{n}, 0.5\right)$$

2.3.2 Sentence length

We give a length penalty $Pen(S_i)$ to a sentence S_i which is too short or too long. In Korean, the length

of important sentences that are selected by humans in newspaper articles is usually between 10 and 30 *eojeols*¹ [6]. The length of an *eojeol* is usually five to six syllables. Thus, we assign a penalty 0.5 to sentences of which length is shorter than 50 syllables or longer than 180 syllables.

2.3.3 Stigma terms

When some sentences contain quotation marks, they can be redundant contents in a summary [7]. Therefore we reduce the score of sentences by 0.4, when sentences include quotation marks.

2.3.4 Lead words

A headline is the most simple and concise in terms of delivering information about news articles. The basic idea is that a sentence that contains words in a headline will deliver important contents about the article. Also, the main contents are typically located at the beginning of a news article. Thus, we use words in a headline and lead sentences to identify important sentences in a document [2]. These lead words in each document also are used to construct a global term cluster.

2.3.5 A global term cluster

To find topic words in multiple documents, we construct a global cluster [3]. We calculate the frequency of lead words from each document, and construct a global term cluster with lead terms that have higher frequency than a threshold value T_g . A threshold value is empirically determined to 0.4. The frequency of each word is used as a weight.

2.4 Removing redundant sentences

After extracting the important sentences in multiple documents, we group them by their similarity. We add a module that prevents similar contents from including in a final result. Our system is based on sentence-extraction. Thus, we check redundant contents on the sentence-level, and estimate the similarity between sentences. The similarity value is used to construct the cluster of semantically similar sentences. Our system basically calculated the Dice coefficient as a similarity measure based on the number of words. We develop an improved term-weighting method that assigns weights to words, using syntactic information [8].

Measuring similarity between sentences, we do not rely on term frequency (TF), and inverse document frequency (IDF), because they can not distinguish words that are more syntactically important from

others. When we compare two sentences, we expect that syntactically important words will obtain a higher score than others. Basically, main clauses will deliver more important information than sub clauses. In addition, we believe that subjects, objects, and verbs are syntactically important, compared to others in a sentence. Therefore, we give weights to each word according to its syntactic role and the type of sentences that it locates.

When comparing words, we use not only the surface form of a word, but also the concept code of it. In particular, we use the concept code only for predicates. For conceptual generalization, we use the concept codes of the Kadokawa thesaurus, which has a 4-level hierarchy of 1,110 semantic classes, as shown in Figure 2 [9]. Concept nodes in level L1, L2 and L3 are further divided into 10 subclasses and nodes in level L4 have 3-digit code between 000 and 999. Formally, the similarity between two sentences S_1 and S_2 is calculated as follows.

$$Sim(S_1, S_2) = \frac{\sum_{(t_i, t_j) \in S_1 \cap S_2} W(t_i, S_1) + W(t_j, S_2)}{\sum_{t \in S_1} W(t, S_1) + \sum_{t \in S_2} W(t, S_2)}$$

Where

- $S = \{t_1, \dots, t_n\}$
- $W(t, S) = W_{st}(t, S) + W_{gr}(t, S)$
- $W_{st}(t, S)$: A weight function by a term t is located which type of a clause.
- $W_{gr}(t, S)$: A weight function for a term t by its grammatical role.
- $\cap = \cap + \cap$
- $S_1 \cap S_2 = \{(t_i, t_j) | t_i \in S_1, t_j \in S_2, t_i = t_j\}$
- $S_1 \cap S_2 = \{(t_i, t_j) | t_i \in S_1, t_j \in S_2, C(t_i) \cap C(t_j) \neq \emptyset\}$
- $C(t)$: The semantic code set of term t

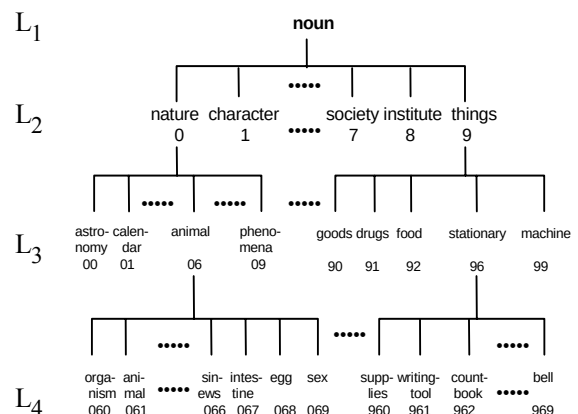


Figure 2. Hierarchy of Kadokawa thesaurus

¹ An *eojeol* is a one or more morphemes and identified with a preceding and following space. It is similar to the notion of a word in English.

We use single-link clustering algorithm. Sentence pairs having similarity value higher than a threshold value T_r are regarded as similar, and are included in a redundant cluster. A threshold value T_r is set to 0.5. After clustering, from each cluster that is expected to have several redundant sentences, we choose only one sentence that has a highest score. In our system, we use a Korean dependency parser to obtain syntactic information.

If a term t is located in a main clause, its score is set as 0.8, otherwise 0.2. When a term is a subject, object or verb in its sentence, 0.5 is added to its score. Unknown words are usually a new term, a name of a person or place, thus it has higher probability to represent important contents in newspaper articles. Thus, we add an additional score 0.5 to the score of an unknown word.

2.5 Mapping Korean sentences to Japanese sentences

After removing redundant sentences, we arrange the remaining extracted sentences by their score. As the final result must be written in Japanese, we map Korean sentences generated by our system to Japanese sentences from original documents. We continue to add a Japanese sentence to a final summary, until the size of the summary is satisfied. In TSC3, some source documents contain a published date, and others do not. However, the name of documents implicitly represents a date when they were published. We considered it as the time information of documents. Using it, we rearrange sentences in the final summary.

3. Evaluation result

In NTCIR TSC3, we submit 2 types of an abstract summary: 'Short' and 'Long'. Although we did not submit the result of an extraction, we evaluate the result of a sentence extraction, using the scoring tool given to participants from NTCIR. The ranking of each table is counted among nine systems except for LEAD system and Human summary.

3.1 Content for abstraction

Table 1 shows the results on readability evaluation by human. Our system performs better in the short type of a summary rather than in long type of it.

Table 1. Results on content evaluation

ID	Short	Long
LEAD	0.160	0.159
HUMAN	0.385	0.402
Our system	0.222	0.210
Ranking	5/9	9/9

3.2 Pseudo Question-Answering for abstraction

In the result of pseudo Question-Answering, our system obtains the similar ranking to the evaluation of contents. Our results are higher than the result of Lead method. However, the difference in performance between the Lead method and our system is small.

Table 2. Results on pseudo Q.A. evaluation

ID	Short		Long	
	Exact	Edit	Exact	Edit
LEAD	0.300	0.589	0.275	0.602
HUMAN	0.461	0.716	0.426	0.721
Our system	0.321	0.601	0.313	0.611
Ranking	6/9	6/9	6/9	7/9

3.3 Readability for abstraction

In TSC3, the readability of the results is evaluated by using Quality Questions. A q00 measures how many redundant or unnecessary sentences are in the result of a system. For removing redundant contents, our system shows good performance. As our summary consists of original sentences, we obtain a high ranking in some questions that do not require shortening or revising of sentences and words, such as q02, q05 q06, q12, q13 and q15. A q08 measures whether summary has wrong chronological ordering. Our system using a published data obtains the highest ranking. Table 3 shows the results of the readability evaluation.

Table 3. Results on readability evaluation

	Short			Long		
	Hum	Result	Rank	Hum	Result	Rank
q00	0.033	0.100	3	0.033	0.333	4
q01	0.267	1.067	6	0.167	1.567	7
q02	0.000	0.433	2	0.100	1.067	4
q03	0.000	0.400	6	0.000	0.533	8
q04	0.433	2.433	6(2)	1.133	4.567	5
q05	0.400	0.500	2	0.467	1.000	1
q06	0.400	0.567	2	0.433	0.933	2
q07	0.000	0.867	6(2)	0.067	1.967	8(2)
q08	0.933	0.200	1	0.800	-0.13	1
q09	0.500	0.267	7	0.567	0.200	6
q10	0.033	1.633	5	0.000	3.300	6
q11	0.000	0.100	7(2)	0.033	0.000	1(2)
q12	0.000	0.000	1(8)	0.000	0.000	1(5)
q13	0.033	0.000	1(4)	0.000	0.000	1(3)
q14	0.033	0.067	5(4)	0.033	0.033	2(6)
q15	0.033	0.100	1(4)	0.100	0.100	1

- (number) is the number of systems with the same ranking

3.4 Precision and coverage for extraction

In the extraction task, 11 systems are evaluated. Our system is ranked at 7 or 8 out of 11 systems. We implement our system based on sentence-level extraction, thus the results of the extraction task is similar to the results of the abstraction task. As with the evaluation of the abstraction, our system does not perform well.

Table 4. Results on extraction task

ID	Short		Long	
	Cov.	Prec.	Cov.	Prec.
F0301(a)	0.315	0.494	0.355	0.554
F0301(b)	0.372	0.591	0.363	0.587
F0303(a)	0.222	0.314	0.313	0.432
F0303(b)	0.293	0.378	0.295	0.416
F0304	0.328	0.716	0.327	0.535
F0306	0.283	0.496	0.341	0.528
F0307	0.329	0.567	0.391	0.680
Our system	0.283	0.433	0.302	0.475
F0309	0.308	0.505	0.339	0.585
F0310	0.181	0.275	0.218	0.421
F0311	0.251	0.476	0.247	0.547
LEAD	0.212	0.422	0.259	0.539
Ranking	8/11	7/11	8/11	8/11

4. Additional Experiment

In TSC 3, our result was not good. Our system was even below a LEAD system in long type results on extraction. After a formal run, we tried to increase a performance of our system. Because our system is based on the sentence-extraction, increasing the results on extractions should be essential. By several experiment, we found optimal parameter values such a threshold value for redundant measuring, weight for scoring function. We increased a parameter for redundant measuring, weights for sentence-position and for global terms. Although our approach must contain errors in phase of translating documents, the performance of our system is highly increased after resetting some parameters.

Table 5. Results on extraction of additional experiments

	Short		Long	
	Cov.	Prec.	Cov.	Prec.
Formal Run	0.283	0.433	0.302	0.475
New Results	0.309	0.528	0.344	0.587
LEAD	0.212	0.422	0.259	0.539
Ranking	5/11	3/11	4/11	2/11

Increasing the performance of extraction results does not mean the better abstract summary. However

our system is based on sentence-extraction, thus we think that the results on abstraction will be better than the results of a formal run.

5. Conclusion

In NTCIR TSC3, we generate a summary in a source language, using multiple target documents translated by an MT. Our approach is possible to produce a summary without a summarization system in a source language. Using a statistical MT and bilingual corpus, our approach can be applied to any language pair. Compared with other systems, the evaluation of results shows that our system does not perform well in all evaluation metrics. For extracting sentences, we use traditional features for considering multiple newspaper articles. Our approach basically contains translation errors, thus the results of our evaluation are not good. However after additional experiments, our system shows good performance compared the results in TSC 3. We expect that our approach shows the possibility of multi-document summarization using cross-language texts.

Acknowledgements

This work was supported by the Korea Science and Engineering Foundation (KOSEF), through the Advanced Information Technology Research Center (AITrc).

References

- [1] C. J. Park, J. H. Lee, G. B. Lee, and K. Kakechi, *Collocation-Based Transfer Method in Japanese-Korean Machine Translation*. Transaction of Information Processing Society of Japan, 1997, 38(4), page 707-718.
- [2] B. Schiffman, A. Nenkova, K. McKeown, *Experiments in Multidocument Summarization*. HLT conference, 2002.
- [3] D. R. Radev, H. Jing, M. Budzikowska, *Centroid-Based Summarization of Multiple Documents*. ANLP/NAACL Workshop, 2000.
- [4] H.P. Edmundson, *New Method for the Statistics of Surprise and Coincidence*. Computational Linguistics 19, page 61-74, 1969.
- [5] C. Nobata, S. Sekine, K. Uchimoto, H. Isahara, *A summarization system with categorization of document sets*. Proceedings of The Third NTCIR Workshop3, 2002.
- [6] J. M. Yoon, *Automatic summarization of newspaper articles using activation degree of 5W 1H*, Master's thesis POSTECH, 2002.

- [7] C. Y. Lin and E. Hovy, *Form Single to Multi-document Summarization: A Prototype System and its Evaluation*. Proceedings of the 40th Annual Meeting of the ACL, p.457-464, 2002.
- [8] J. M. Lim, I. S. Kang, J. H. Bae, J. H. Lee, *Measuring Improvement of Sentence-Redundancy in Multi-Document Summarization*. Proceedings of the 30th KISS fall conference, 2003.
- [9] S. Ohno and M. Hamanishi, *New synonyms Dictionary*, Kadokawa Shoten, Tokyo, 1981.

