Generation of Summaries that Appropriately and Adequately Express the Contents of Original Documents Using Word-Association Knowledge

Kazuki Takigawa[a], Masaki Murata[b], Masaaki Tsuchida[c], Stijn De Saeger[c],
Kazuhide Yamamoto[a], and Kentaro Torisawa[c]

[a] Deparment of Electrical Engineering,Nagaoka University of Technology
1603-1, Kamitomioka, Nagaoka City, Niigata, 940-2188, Japan
{takigawa,yamamoto}@jnlp.org
[b] Graduate School of Engineering, Tottori University
4-101, Koyama-Minami, Tottori, 680-8550, Japan
murata@ike.tottori-u.ac.jp
[c] National Institute of Information and Communications Technology
3-5, Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0289, Japan
{m-tsuchida,stijn,torisawa}@nict.go.jp

Abstract. In this study, our purpose was to make a short summary for sentences. For example, we aimed to make a short summary "terror" for sentences "A bomb went off. Some people were killed. This was triggered by rebel campaign." In this study, we proposed a new method that generates summaries that can appropriately and adequately express the contents of their respective original documents using word-association knowledge. In this method, we assumed that a good summary comprises words that can express the contents of the original document and does not contain words that are unable to express the contents of the original document. Using statistical tests, we confirmed that the use of elements in our method was beneficial. Our method obtained 0.75 as the ratio where the top 10 summaries for each document include a correct summary and 0.45 as the mean reciprocal rank (MRR) in the "lenient" case of experiments.

Keywords: Summarization, Word-Association Knowledge, Precision, Recall, Generation

1 Introduction

Summarization is one of the important techniques in natural language processing more so because of the development of internet-based technologies and the existence of many documents and many kinds of information on the Web (Hovy and Mareu, 1988; Mani and Maybury, 1999; Barzilay and Elhada, 1997; Goldstein et al., 1999; Marcu, 2000; Ker and Chen, 2000; Hongyan, 2000; Radev et al., 2001; Barzilay and Lee, 2004; Radev et al., 2004; Kato et al., 2005).

In this study, our purpose was to make a short summary for sentences. For example, we aimed to make a short summary "terror" for sentences "A bomb went off. Some people were killed. This was triggered by rebel campaign." In this paper, we have proposed a new method that generates summaries that appropriately and adequately express the contents of the original documents using word-association knowledge. In our proposed method, a system judges that a candidate summary that satisfies the following criteria as much as possible is a suitable summary: (i) The contents of the original document are conveyed by the candidate. (ii) The content that is not described in the original document is not conveyed by the candidate.

In this study, we use co-occurring words as word-association knowledge. By using co-occurring words, a summary can be generated containing words that do not appear in the original document (summarization by paraphrasing). In this aspect, our method is completely different

from the existing methods that extract some sentences or words from the original document for generating its summary (Knight and Marcu, 2002; Lin and Hovy, 2003; Kang et al., 2008).

In terms of related studies pertaining to summarization by paraphrasing, Kondo et al. proposed a method of paraphrasing plural verbs in the original document into a verb having superordinate concepts including plural verbs concepts by using definition sentences from a word dictionary (Kondo and Okumura, 1997). This method was useful for handling only verbs and could not handle nouns. Another problem is that it cannot handle summarization of the information that is not described in the definition sentences of a word dictionary. In contrast, we can handle parts of speech other than verbs by using co-occurring words. We can also perform paraphrasing of various kinds of information by using co-occurring words. Banko et al. generated newspaper headlines by using statistical machine translation (Banko et al., 2000).

In this study, we handled only the case in which the output summary is one word for simplicity; however, we aim to generate a summary comprising sentences as the output in the future. The case where the output summary comprises two words or sentences can be handled by using an extended version of this method. In this paper, we have only described our idea for handling two-word summaries in Section 4. The summarization process that outputs one-word summaries can also be used to categorize documents, because that one word is representative of the main content and theme of the document. We handled the summarization of documents written in Japanese in this study.

2   Proposed Method

2.1   Basic Procedure

The procedure of our proposed method is as follows:

1. Obtain nouns from an input document
2. Obtain words co-occurring with each of the obtained nouns and handle them as candidate summaries
3. Calculate the score for each candidate by using the criteria (i) and (ii) described in Section 1
4. Output the candidate having the highest score as the summary

2.2   Co-occurring Words

We used the top 50 nouns that appear most frequently with a particular noun on the basis of 50 million Web pages as the co-occurring words of that noun.

2.3   Extraction of Nouns from the Input Document

We obtain nouns from the input document.[1] We used JUMAN (Kurohashi and Kawahara, 2009) for tagging parts of speech.

2.4   Extraction of a Candidate Summary

In this study, we used words co-occurring with the nouns from an input document as candidate summaries. This is because we believe that a noun that is related to the nouns from the input document can be appropriate as the summary.

2.5   Calculation of Candidate Scores

To select an appropriate candidate among the possible candidates, we defined candidate scores on the basis of the two criteria described in Section 1. Assuming that the content of the input document is a set of correct data items, criterion (i) indicates the maximum number of correct data items a candidate can relate to without missing correct data items by using word-association knowledge such as related words. This is thus similar to *recall*. Criterion (ii) indicates the least number of incorrect data items a candidate can relate to by using word-association knowledge. This is thus similar to *precision*.

---

[1] In Japanese, important information is expressed using nouns, so we used nouns for co-occurring words and for words extracted from an input document.

On the basis of the above idea, we set the following candidate scores in this study:

$$Recall(c) = \frac{|RW(c) \cap IW|}{|IW|} \tag{1}$$

$$Precision_1(c) = \frac{|RW(c) \cap IW|}{|RW(c)|} \tag{2}$$

$$F\text{-}measure_x(c) = \frac{2 \times Recall(c) \times Precision_x(c)}{Recall(c) + Precision_x(c)} \tag{3}$$

Here, c is a candidate summary, IW is a set of nouns obtained from an input document, and $RW(x)$ is a set of related words of a word x. In this study, we used co-occurring words as related words.

Words other than the nouns from an input document can include the content of the document. For example, the co-occurring words of the nouns from an input document can include the content of the document. Therefore, we can use the co-occurring words as correct data items in criterion (ii). From this idea, we set the following candidate score:

$$Precision_2(c) = \frac{|RW(c) \cap (IW \cup (\cup_{i \in IW} RW(i)))|}{|RW(c)|} \tag{4}$$

2.6   Arrangement of Candidates

We arrange candidates in the descending order of the candidate scores, and output the candidate summary that has the highest candidate score.

We used the following four methods of arrangement:

- Method 1: We use $Precision_1(c)$ for $Precision_x(c)$ and one of Equations 1 to 3 for arrangement.[2]
- Method 2: We use $Precision_2(c)$ for $Precision_x(c)$ and $Recall(c)$ for arrangement.
- Method 3: We use $Precision_2(c)$ for $Precision_x(c)$ and $Precision_2(c)$ for arrangement.
- Method 4: We use $Precision_2(c)$ for $Precision_x(c)$ and $F\text{-}measure_x(c)$ for arrangement.

2.7   An Example of Generating a Summary

We demonstrate an example of using Method 2. We assume that an input document is as follows:

yoi	*kigyou ni*	*naitei wo*	*morau tame,*	*mensetsu no*	*renshu wo*
(good)	(company)	(official job offer)	(to get)	(interviewing)	(practice)

mainichi	*okonau.*
(every day)	(perform)

(To get an official job offer from good companies, I practice interviewing every day)

The system obtained six nouns, *kigyo* (company), *naitei* (official job offer), *mensetsu* (interviewing), etc. from the input document. The system obtained 50 co-occurring words for each obtained noun. Some of these co-occurring words are as follows: *kaisha* (corporation), *shoukai* (introduction), *kyaku* (customers), etc., as co-occurring words of *kigyou* (company); and *kigyo* (company), *shuushoku katsudo* (job hunting), *jouhou* (information), etc., as co-occurring words of *naitei* (official job offer). The co-occurring words are candidate summaries. Next, the system obtained co-occurring words of each candidate to calculate candidate scores for those candidates. Finally, the system arranged the candidates in the descending order of the obtained candidate scores. Because we used Method 2, the candidate score of $Recall(c)$ was used first, and the candidate score of $Precision_2(c)$ was used next. The candidate score of *shuushoku katsudo* (job hunting)was the highest. Consequently, the system outputted *shuushoku katsudo* (job hunting) as the summary for the input document.

---

[2] In this study, we used 50 words for $RW(c)$, thus, $|RW(c)|$ is a constant (50). When an input document is determined, $|IW|$ is also a constant. Therefore, even if we use any of $Recall(c)$, $Precision_1(c)$, and $F\text{-}measure_1(c)$ (Equations 1 to 3) for arrangement, we obtain the same results.

3 Experiments

3.1 Evaluation Procedure

We manually created 24 input documents for evaluation. Each of the input documents comprised a simple sentence whose content can be expressed using a single noun. The evaluation was performed by a test subject.

In evaluation, a test subject evaluated the top 10 arranged candidates. We used Top1, Top5, Top10, and MRR as evaluation measures. TopX indicates the ratio when one of the top X candidates is correct. MRR indicates the average for a score of $1/r$, which is given when the r-th candidate is correct. We used two kinds of evaluation standards, "strict," and "lenient." In "strict," we evaluated only correct candidates as the correct summary. In "lenient" evaluation, we evaluated candidates similar to a correct candidate as the correct summary.

We obtained a high Kappa value of 0.78 and 0.75 for "strict" and "lenient," respectively, when we evaluated the results using two test subjects in a preliminary experiment (Landis and Koch, 1977).

3.2 Results

Evaluated results are shown in Tables 1 and 2. Methods 1 to 4 are described in Section 2.6.

Method 5 is a comparison method that uses definition sentences in the EDR word dictionary (EDR, 1993). In Method 5, we selected a candidate word whose definition sentence contained a noun that was also present in the input document; we used these words in the definition sentence of a word x as the related words of the word x. The subsequent working is the same as that of our proposed method. Since Method 5 used a definition sentence, it is related to Kondo et al.'s studies (Kondo and Okumura, 1997). However, Kondo et al. did not propose the method of selecting a candidate from plural candidates. We used our candidate scores that were obtained on the basis of criteria (i) and (ii) for selecting a candidate among the possible candidates.

Method 2 obtained the best scores in all the cases. We used a one-sided binomial sign test for TopX and a statistical test, one-sided t-test, for MRR. All results where the difference with Method 2 is statistically significant are marked using asterisks. The number of asterisks reflects the confidence level, respectively "*" for 0.15, "**" for 0.1, "***" for 0.05 and "****" for 0.01. We confirmed that Method 2 was better than all the other methods for an evaluation measure at a significant level of 0.15. This indicates that $Recall(c)$ is good for obtaining candidate scores used for arrangement, and $Precision_2(c)$ is good for obtaining $Precision_x(c)$.

The fact that Method 2 was better than Method 5 indicates that it is better to use co-occurring words instead of definition sentences for word-association knowledge.

It was confirmed by a statistical test that the use of elements in our method (use of $Precision_2(c)$, use of Method 2 (use $Recall(c)$ and then use $Precision_x(c)$), and use of co-occurring words) was beneficial.

We show some examples of summaries outputted by Method 2 as follows. A tag "s" as in "*sekyuritii* (security)^s" indicates a word that is judged as a correct one when using the evaluation standard,"strict." A tag "l" indicates a word that is judged as a correct one when using the evaluation standard, "lenient." "No.1: An Input document" is a sample of the results where the first outputted summary is judged as a correct one using "strict." "No. 2: An Input document" is a sample of the results where the top 10 outputted summaries include a correct one for "strict." "No. 3: An Input document" is a sample of the results where the first outputted summary is judged as a correct one using "lenient." "No. 4: An Input document" is a sample of the results where the top 10 outputted summaries include a correct one for "lenient."

No. 1: An Input document:

puraibashii wo	*mamoru tame,*	*kojin*	*jouhou wo*	*hogo suruyouni*
(privacy)	(defend)	(individual)	(information)	(protect)

Table 1: Results (Strict)

	Candidate score used for arrangement	Used for $Precision_x(c)$	Top1	Top5	Top10	MRR
Method 1	-	$Precision_1(c)$	0.13	0.13**	0.29	0.15*
Method 2	$Recall(c)$	$Precision_2(c)$	**0.17**	**0.29**	**0.38**	**0.22**
Method 3	$Precision_x(c)$	$Precision_2(c)$	0.00***	0.17*	0.25*	0.06***
Method 4	$F\text{-}measure_x(c)$	$Precision_2(c)$	0.08	0.25	0.33	0.16**
Method 5	$Recall(c)$	$Precision_2(c)$	0.04	0.08**	0.08***	0.05**

Table 2: Results (Lenient)

	Candidate score used for arrangement	Used for $Precision_x(c)$	Top1	Top5	Top10	MRR
Method 1	-	$Precision_1(c)$	0.29	0.50	**0.75**	0.41
Method 2	$Recall(c)$	$Precision_2(c)$	**0.33**	**0.58**	**0.75**	**0.45**
Method 3	$Precision_x(c)$	$Precision_2(c)$	0.00****	0.25****	0.46****	0.10****
Method 4	$F\text{-}measure_x(c)$	$Precision_2(c)$	0.17**	**0.58**	0.71	0.34***
Method 5	$Recall(c)$	$Precision_2(c)$	0.08**	0.21***	0.29****	0.15***

settei wo   okonatta.
(setting)   (perform)

(In order to defend privacy, I performed setting so as to protect individual information.)

The correct summary: *sekyuritii* (security)

The top 10 outputted summaries:

1. *sekyuritii* (security)[s], 2. *herupu* (help), 3. *sekyuritiii* (security), 4. *porishi* (policy), 5. *hogo* (protect), 6. *riyou kiyaku* (terms of service), 7. *rinku* (link), 8. *shiyou* (use), 9. *riyou* (use), 10. *chosakuken* (copyright).

No. 2: An Input document:

keisatsu ga   yougisha wo   tsukamaeta.   hikitsuzuki   kousoku wo   okonau youda
(police)   (suspect)   (catch)   (continue)   (hold)   (perform)

(Police catched a suspect. Police will continue to hold him.)

The correct summary: *taiho* (arrest)

The outputted summaries:

1. *migara* (body), 2. *kousoku* (hold), 3. *yougi* (suspicion), 4. *taiho* (arrest)[s], 5. *jijouchoushuu* (interview), 6. *utagai* (doubt), 7. *chikan* (molestation), 8. *gennkouhan taiho* (on-the-spot arrest), 9. *kyoujutsu* (statement), 10. *sousa* (investigation).

No. 3: An Input document:

jikken no   mokuteki ya   riron wo   matome   toshokan wo   riyou shite
(experiment)   (purpose)   (theory)   (summarize)   (library)   (use)

kadai wo   sakusei suru
(task)   (make)

(I summarized the purpose and theory in the experiment and made the results of the task in a library.)

The correct summary: *report sakusei* (to make reports)

The outputted summaries:

1. *jisshu* (practical training)[l], 2. *kenkyu* (research), 3. *kousatsu* (examination), 4. *bunseki* (analysis), 5. *shiryo* (written material), 6. *keikaku* (plan), 7. *jikken* (experiment), 8. *saabisu* (service), 9. *shuhou* (method), 10. *kenshou* (test).

Table 3: Comparison between the proposed method and the method using a definition sentence (Strict)

Method	Top1	Top5	Top10	MRR
The proposed method	**0.17**	**0.29**	**0.38**	**0.22**
The method using a definition sentence	0.04	0.08**	0.08***	0.05**

Table 4: Comparison between the proposed method and the method using a definition sentence (Lenient)

Method	Top1	Top5	Top10	MRR
The proposed method	**0.33**	**0.58**	**0.75**	**0.45**
The method using a definition sentence	0.08**	0.21***	0.29****	0.15***

No. 4: An Input document:

tsukue ni	*mukai*	*kyoukasho to*	*jugyou no*	*nouto wo*	*hiraita.*
(desk)	(against)	(textbook)	(lesson)	(notebook)	(open)

kyou no	*hukushuu to*	*asu no*	*yoshu wo*	*suru*	*hitsuyou ga*	*aru*
(today)	(review)	(tomorrow)	(preparation)	(do)	(necessary)	(is)

(I opened a textbook and a notebook for a lesson at the desk. I must review today's content and prepare tomorrow's content.)

The correct summary: *benkyou* (to study)

The outputted summaries:
1. *yoshu* (prepare, study in advance), 2. *shukudai* (homework)[l], 3. *junbi* (prepare), 4. *rika* (science), 5. *moshi* (mock examination), 6. *kyoukasho* (textbook), 7. *fukushu* (review), 8. *kougi* (lecture), 9. *suugaku* (mathematics), 10. *gakkou* (school).

3.3   Additional Experiments

Furthermore, we made additional experiments using a word dictionary. Our proposed method we used co-occurring words as word-association knowledge.

Here, we tried using another method for word-association knowledge. In this method, we used a definition sentence in a word dictionary to use word-association knowledge. In the method, $RW(x)$ is a set of words that appeared in the definition sentence a word x. The other part is the same as the method described in Section 2. We call this method the method using a definition sentence.

We made experiments comparing the proposed method described in Section 2 and the method using a definition sentence. The results are shown in Tables 3 and 4. In the experiments, we used Method 2 for both methods, because Method 2 obtained the highest accuracy in the experiments mentioned in Section 3. We used definition sentences in the EDR word dictionary (EDR, 1993) to handle the method using a definition sentence.

In these tables, Top1, Top5, Top10, MRR, "strict," and "lenient" are the same as in 1. All results where the difference with the proposed method is statistically significant are marked using asterisks. The number of asterisks reflects the confidence level, respectively "*" for 0.15, "**" for 0.1, "***" for 0.05 and "****" for 0.01.

From the tables, we found that the proposed method, which use co-occurring words as word-association knowledge, was significantly better than the method using a definition sentence as word-association knowledge.

4   Summarization Outputting a Summary Containing Multiple Words

In the previous sections, we only handled a summary comprising one word. We can also handle a summary that comprises more than one word using our method. This method is useful in cases when a one-word outputted summary cannot cover all the contents of an input document. In the new method, at first, we generate a one-word summary c_1 using our method. Next, we detect candidate words that are related to nouns from an input document. Each candidate word ec was

evaluated using our proposed method. For the calculation of candidate scores, we make a set of c_1 and ec and use it as c in our equations (Equations 1 to 4). Among a set of ec, the candidate word c_2 having the highest candidate score is added to the summary. A set of nouns, c_1 and c_2, is outputted as the final summary for the input document. A summary comprising more than two words can be generated in the same way. Furthermore, we can extend this method to generate sentences for a summary by adding verbs to candidate summaries by using generation techniques for connecting plural words and by using word n-grams or dependency information in candidate score equations.

We performed the experiments for generating a two-word summary. We show an example of the experimental results as follows:

An Input document:

han'nin ga	*juu wo*	*tsukai,*	*juudan ga*	*higai*	*sha ni*	*atari*
(criminal)	(gun)	(use)	(bullet)	(suffered)	(person)	(hit)

satsugai sareta.
(be killed)

(A criminal used a gun, and a suffered person was hit by a bullet and killed.)

Nouns from the input document:

han'nin (criminal), *juu* (gun), *juudan* (bullet), *higai* (suffered), *satsugai* (be killed)

The correctly outputted summary:

The firstly used word c_1: *happou* (shooting or fire)

The secondly used word c_2: *satsugai* (be killed)

happou (shooting or fire) was obtained as a candidate for the first word of the summary c_1 using co-occurring words of the word, *juu* (gun). It had the highest candidate score, and was selected as the first summary word. It covered a content of words *han'nin* (criminal), *juu* (gun), *juudan* (bullet), and *satsugai* (be killed). Among the words from the input document, the remaining word was only *higai* (suffered). *satsugai* (be killed), which was a co-occurring word of a word, *han'nin* (criminal), was a candidate in the second pass of the summarization process, and was selected as the second word of the summary because it had *higai* (suffered) as its co-occurring word; *han'nin* (criminal) also had the highest candidate score. Finally, the outputted summary was a set of the two words *happou* (shooting or fire) and *satsugai* (be killed).

5   Conclusion

In this paper, we proposed a new method for generating summaries that appropriately and adequately express the contents of the original documents using word-association knowledge. In the proposed method, we assumed that a good summary comprises words that can express the contents of the original document and does not contain words that do not express the contents of the original document. Using statistical tests, we confirmed that the use of elements in our method was beneficial. We found that $Recall(c)$ is good for a candidate score used for arrangement, and $Precision_2(c)$ is good for $Precision_x(c)$. We also found that co-occurring words are better than definition sentences when used for word-association knowledge. Our proposed method (Method 2) obtained 0.75 as the ratio where the top 10 summaries for each document included a correct summary, and 0.45 as MRR in the "lenient" case of experiments. Furthermore, we described our idea in the case of a two-word summary, and showed a sample experiments using that idea.

References

Michele Banko, Vibhu O.Mittal, and Michael J.Witbrock. 2000. Headline generation based on statistical translation. *Proceedings of 38th Meeting of Association for Computation Linguistics*, pages 318–325.

Madrid. Barzilay and Micheal Elhada. 1997. Using lexical chains for text summarization. *ACL/EACL Workshop on Intelligent Scalable Text Summarization*, pages 10–17.

Regina Barzilay and Lillian Lee. 2004. Catching the drift: Probabilistic content models, with applications to generation and summarization. *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 113–120.

EDR, 1993. *EDR Electronic Dictionary Technical Guide*. EDR (Japan Electronic Dictionary Research Institute, Ltd.).

Jade Goldstein, Mark Kantrowitz, Vibhu Mittal, and Jaime Carbonell, 1999. *Summarizing Text Documents: Sentence Selection and Evaluation Metrics*, pages 121–128. Berkeley.

Jing Hongyan. 2000. Sentence reduction for automatic text summarization. *Proceedings of ANLP 2000*, pages 310–315.

Eduard Hovy and Daniel Mareu. 1988. Automated text summarization. *Tutorial in 17th ACL*, 33(11):29–36.

Moonyoung Kang, Sourish Chaudhuri, Mahesh Joshi, and Carolyn P. Rose. 2008. Side : The summarization integrated development environment. *Proceedings of the ACL HLT Demo Session (Companion Volume)*, pages 24–27.

Tsuneaki Kato, Mitsunori Matsushita, and Noriko Kando. 2005. MuST: A workshop on multimodal summarization for trend information. *Proceedings of the Fifth NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access*, pages 556–563.

Sue J. Ker and Jen-Nan Chen. 2000. A text categorization based on summarization technique. *Proceedings of the ACL-2000 workshop on Recent advances in natural language processing and information retrieval*, pages 79–83.

Kevin Knight and Daniel Marcu. 2002. Summarization beyond sentence extraction: A probabilistic approach to sentence compression. *Artificial Intelligence*, 139(1):91–107.

Keiko Kondo and Manabu Okumura. 1997. Summarization with dictionary-based paraphrasing. *Proceedings of Natural Language Processing Pacific Rim Symposium (NLPRS'97)*, pages 649–652.

Sadao Kurohashi and Daisuke Kawahara, 2009. *Japanese Morphological Analysis System JUMAN version 6.0*. Department of Informatics, Kyoto University.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, pages 159–174.

Chin-Yew. Lin and Eduard Hovy. 2003. The potential and limitations of automatic sentence extraction for summarization. *Proceedings of the HLT/NAACL Workshop on Automatic Summarization*, pages 73–80.

Inderjeet Mani and Mark T. Maybury. 1999. *Advances in automatic text summarization*. The MIT Press.

Daniel Marcu. 2000. *The Theory and Practice of Discourse Parsing and Summarization*. The MIT Press.

Dragomir R. Radev, Sasha Blair-Goldensohn, Zhu Zhang, and Revathi Sundara Raghavan. 2001. NewsInEssence: a system for domain-independent, real-time news clustering and multi-document summarization. In *Proceedings of the First International Conference on Human Language Technology Rresearch*, pages 1–4.

Dragomir R. Radev, Hongyan Jing, Ma Gorzata Sty, and Daniel Tam. 2004. Centroid-based summarization of multiple documents. *Information Processing & Management*, 40(6):919–938.