Available online at www.sciencedirect.com
ScienceDirect
Pattern Recognition Letters 29 (2008) 1366–1371
Pattern Recognition Letters
www.elsevier.com/locate/patrec

An effective sentence-extraction technique using contextual information and statistical approaches for text summarization

Youngjoong Ko [a,*], Jungyun Seo [b]

[a] Department of Computer Engineering, Dong-A University, 840 Hadan 2-dong, Saha-gu, Busan 604-714, Republic of Korea
[b] Department of Computer Science and Interdisciplinary Program of Integrated Biotechnology, Sogang University, Sinsu-dong, Mapo-gu, Seoul 121-742, Republic of Korea

Received 6 August 2007; received in revised form 19 February 2008
Available online 4 March 2008

Communicated by T. Vasilakos

Abstract

This paper proposes an effective method to extract salient sentences using contextual information and statistical approaches for text summarization. The proposed method combines two consecutive sentences into a bi-gram pseudo sentence so that contextual information is applied to statistical sentence-extraction techniques. Salient bi-gram pseudo sentences are first selected by the statistical sentence-extraction techniques, and then each selected bi-gram pseudo sentence is separated into two single sentences. The second sentence-extraction task for the separated single sentences is performed to make a final text summary. Because the proposed method uses the contextual information with the bi-gram pseudo sentences and combines the statistical sentence-extraction techniques effectively, it can achieve high performance. As a result, the proposed method showed better performance than other sentence-extraction methods in both single- and multi-document summarization.

Keywords: Text summarization; Sentence-extraction; Statistical approaches; Single- and multi-document summarization

1. Introduction

With the growing popularity of the Internet and a variety of information services, obtaining the desired information is becoming a serious problem in the information age. Text summarization sets the goal at taking an information source, extracting content, and presenting the most important content to a user in a condensed form and in a manner sensitive to the user's or application's needs (Mani, 2001). To achieve this goal, text summarization systems should identify the most salient information in a document and convey it in less space than the original document. Therefore, the text summarization has been used as the useful tools in order to help users efficiently find useful information from immense amount of information.

Traditional text summarization systems have used linguistic approaches and statistical approaches to extract salient sentences. There are some problems in using both methods for text summarization. Linguistic approaches have some difficulties in using high quality linguistic analysis tools (a discourse parser, etc.) and linguistic resources (WordNet, Lexical Chain, Context Vector Space, etc.) (Barzilay and Elhadad, 1999; Miller et al., 1990; Ko et al., 2003); they would be very useful resources to understand a document for summary-generation, but they require much memory and processor capacity because of additional linguistic knowledge and complex linguistic processing. On the other side, statistical approaches can summarize texts using various statistical features (title, location, etc.). They do not suffer from a memory and processor capacity problem. But all the statistical features are

* Corresponding author. Tel.: +82 51 200 7782; fax: +82 51 200 7783.
E-mail addresses: yjko@dau.ac.kr (Y. Ko), seojy@sogang.ac.kr (J. Seo).

0167-8655/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2008.02.008

not effective for summarization in every case because some features depend on the particular format and the writing style of documents such as title and location. Especially, title cannot be an effective feature any more if a document does not include any title information. Therefore, this paper presents the TF-based query method, which is used in a no-title style document instead of the title method, and it attempts to find an effective hybrid method to combine statistical approaches by using contextual information for text summarization.

In order to develop a more effective sentence-extraction technique, we focus on the two following facts:

1. *Feature sparseness problem*: it is generally caused by extracting features from only one sentence for sentence-extraction. This problem can be partially solved by using bi-gram pseudo sentences; the bi-gram pseudo sentences are generated by binding two consecutive sentences with a sliding window technique.
2. *Hybrid method*: a hybrid method can be applied to sentence-extraction techniques for effectively generating a summary. Because statistical approaches do not require much memory and processor capacity, the combination of several statistical approaches can provide more effective sentence-extraction technique without the burden of memory and processor capacity.

The proposed method carries out two different sentence-extraction tasks for solving the feature sparseness problem. In the first stage, the target of sentence-extraction is bi-gram pseudo sentences which are used for contextual information; the bi-gram pseudo sentences include more features (words) than one sentence because they are composed of two adjacent sentences. After extracting important bi-gram pseudo sentences, each extracted bi-gram pseudo sentence is separated into two-single sentences, and the separated ones become the target of the second sentence-extraction. The hybrid statistical method is applied to both the sentence-extraction tasks. As a result, the proposed method showed better performance than previous statistical and linguistic approaches in both single- and multi-document summarization.

The rest of paper is organized as follows. Section 2 presents previous related work. In Section 3, we explain the proposed hybrid statistical method using the bi-gram pseudo sentences as contextual information. Section 4 is devoted to the analysis of empirical results. Finally, we describe conclusions.

2. Related work

There have been a lot of studies for text summarization in this literature. Most of the researchers have concentrated on not sentence-generation summarization methods (Yang and Zhong, 1998) but sentence-extraction methods in order to create text summary (Kim et al., 2001; Nomoto and Matsumoto, 2001; Wang et al., 2003). There are several reasons: the high complexity, the limitation of natural language processing techniques and knowledge engineering technology in real application fields.

The pioneering works for text summarization studied that most frequent words represent the most important concepts of the text (Luhn, 1959). This representation abstracts the source text into a frequency table.

Edmundson studied that first paragraph or first sentences of each paragraph contain topic information (Edmundson, 1968). He also studied that the presence of words such as *significant*, *hardly*, *impossible* signals topic sentences.

Goldstein et al. proposed a query-based summarization to *generate* a summary by extracting relevant sentences from a document (Goldstein et al., 1999). The criterion for extraction is given as a query. The probability of being included in a summary increases according to the number of words co-occurred in the query and a sentence.

Carbonell and Goldstein presented the MMR (Maximal Marginal Relevance) technique (Carbonell and Goldstein, 1998). In the MMR technique, a sentence has high marginal relevance if it is relevant to the query and contains minimal similarity to previously selected sentences.

Jung et al. proposed a two-step sentence-extraction method for single-document summarization (Jung et al., 2005). This method was improved by applying it to multi-document summarization in our paper.

Barzilay and Elhadad constructed Lexical Chain by calculating semantic distance between words using WordNet (Miller et al., 1990). Strong Lexical Chains are selected and the sentences related to these strong chains are chosen as a summary. The methods which use semantic relations between words depend heavily on manually constructed resources such as WordNet.

Ko et al. constructed Lexical Clusters (Ko et al., 2003). Each Lexical Cluster has different semantic categories, while they are more loosely connected than Lexical Chains. They can choose topic keywords from each cluster and create a text summary using them. Their system is called by DOCUSUM.

Marcu constructed discourse structure of a document. The nuclei of the discourse structure tree for a text determine salience of information (Marcu, 1996). This method is grounded in notions that are very closely related to what abstractors actually do.

3. The hybrid statistical sentence-extraction method using bi-gram pseudo sentences

3.1. General statistical methods

The representative statistical methods in previous text summarization tasks are presented and they are used in the proposed hybrid statistical sentence-extraction method.

3.1.1. The title method

The score of sentences is calculated as how many words are commonly used between a sentence and a title. The

inner product method is exploited for similarity calculation between a sentence and a title query

$$\mathrm{sim}(S_i, Q) = \sum_{k=1}^{n} w_{ik} w_{qk} \tag{1}$$

$$\mathrm{Score}(S_i) = \mathrm{sim}(S_i, Q) \tag{2}$$

where S_i is an ith sentence and Q is a title query. w_{ik} is the binary weight of kth word in ith sentence and w_{qk} is the binary weight of kth word in the title query.

3.1.2. The location method

It has been said that the leading several sentences of an article are important and a good summary (Wasson, 1998). Therefore, the leading sentences in compression rate are extracted as a summary by the location method

$$\mathrm{Score}(S_i) = 1 - \frac{i-1}{N} \tag{3}$$

where S_i is a ith sentence and N is the total number of sentences in the document.

3.1.3. The aggregation similarity method

The score of a sentence is calculated as the sum of similarities with other all sentence vectors in document vector space model (Kim et al., 2001)

$$\mathrm{sim}(S_i, S_j) = \sum_{k=1}^{n} w_{ik} w_{jk} \tag{4}$$

$$\mathrm{Score}(S_i) = a\,\mathrm{sim}(S_i) = \sum_{j=1, j\neq i}^{m} \mathrm{sim}(S_i, S_j) \tag{5}$$

where w_{ik} is the binary weight of kth word in ith sentence.

3.1.4. The frequency method

The frequency of term occurrences within a document has often been used for calculating the importance of sentences (Zechner, 1997). In this method, the score of a sentence can be calculated as the sum of the scores of words in the sentence. The important score w_i of word i can be calculated by the traditional *tf.idf* method as follows (Salton, 1989):

$$w_i = tf_i \times \log \frac{ND}{df_i} \tag{6}$$

where tf_i is the term frequency of ith word in the document, ND is the total number of documents, and df_i is the document frequency of ith word in the whole data set.

3.1.5. The TF-based query method

Identifying topic is so useful to generate a summary. Many text summarization systems have regarded a title as the topic of a document. However, in special cases, it can be hard to extract a title from each document or any kinds of documents do not have a title. A simple method can be used to extract topic words. The TF-based query method extracts high frequent words in a document, and the extracted words are used as a query instead of title. Like the Title method, the inner product metric is used as the similarity measure between a sentence and a TF-based query.

The similarity between a sentence and the TF-based query is calculated by the following equation:

$$\mathrm{sim}(S_i, TfQ) = \sum_{k=1}^{n} w_{ik} w_{\mathrm{TFQ}k} \tag{7}$$

where n is the number of words which is included in a document. w_{ik} is the binary weight of kth word in ith sentence and $w_{\mathrm{TFQ}k}$ is the binary weight of kth word in the TF-based query.

3.2. The effective combination method of statistical approaches using bi-gram pseudo sentences

The proposed method carries out two different sentence-extraction tasks in order to use bi-gram pseudo sentences for text summarization; the sentence-extraction task in the first stage extracts salient bi-gram pseudo sentences from a target document. Then each extracted bi-gram pseudo sentence is separated into two original single sentences. The separated single sentences become the targets of the second sentence-extraction task.

3.2.1. Extracting bi-gram pseudo sentences in the first stage

First of all, the proposed method generates bi-gram pseudo sentences to solve the feature sparseness problem; they are simply made by combining two adjacent sentences by sliding window technique (Ko and Seo, 2004). This window slides from the first sentence to the last sentence of a document with the interval of each window (one sentence). Through a simple experiment, Title and Location methods are chosen for using the combination method due to their high performance. The proposed hybrid method linearly combines these methods as follows:

$$\mathrm{Score}(S_i) = \mathrm{sim}(S_i, Q) + \left(1 - \frac{i-1}{N}\right) \tag{8}$$

where the notations of this equation follow those of Eq. (1) and (3). After all the bi-gram pseudo sentences are scored by Eq. (8), about 50% of them are selected as important bi-gram pseudo sentences.

3.2.2. The final sentence-extraction task for generating summary in the second stage

In the first stage, more important bi-gram pseudo sentences are extracted. Then the extracted bi-gram pseudo sentences are separated into original single sentences. The Aggregation Similarity method is added to the combination method in the first stage (Eq. (8)). Since more important sentences are remained from the first stage, the sentence-extraction task can be improved as using the Aggregation Similarity method. Note that this method regards the sum of similarities with all other

sentences as the important score of a sentence. Actually, we could observe that adding Aggregation Similarity method rather reduced performance when all the sentences were used without removing noisy sentences in the first stage

$$\mathrm{Score}(S_i) = \mathrm{sim}(S_i, Q) + \left(1 - \frac{i-1}{N}\right) + w_a a\,\mathrm{sim}(S_i) \quad (9)$$

where w_a is a weight value reflecting the importance of the Aggregation Similarity method.

If any document does not have title, the TF-based query method is used instead of the Title method as shown in the following equations:

$$\mathrm{Score}(S_i) = \mathrm{sim}(S_i, TfQ) + \left(1 - \frac{i-1}{N}\right) \quad (10)$$

$$\mathrm{Score}(S_i) = \mathrm{sim}(S_i, TfQ) + \left(1 - \frac{i-1}{N}\right) + w_a a\,\mathrm{sim}(S_i) \quad (11)$$

3.3. Applying the proposed method to multi-document summarization

To apply the proposed method to a multi-document summarization task, we constructed a multi-document summarization data set and conducted experiments on the data set. Since the general multi-document summarization has two sentence-extraction processes, our one also has similar processes; the primary sentence-extraction task makes the summary of each document in a document cluster and the secondary sentence-extraction task generates the final summary of the document cluster from the summaries created in the primary sentence extraction. The proposed method for multi-document summarization has two different points from that for single-document summarization; the one is that any title feature of each cluster does not exist, and the other is that the location feature at the secondary sentence-extraction task is useless. Therefore, the TF-based query method is used instead of title method to solve the former problem, and the location feature from the original document is used instead of that from a merged summary for the latter problem.

After two revised points are applied to the proposed method, the following equations are used in primary and second sentence-extraction tasks

$$\mathrm{Score}(S) = \mathrm{sim}(S, TfQ) + \left(1 - \frac{i_d - 1}{N_d}\right) \quad (12)$$

$$\mathrm{Score}(S) = \mathrm{sim}(S, TfQ) + \left(1 - \frac{i_d - 1}{N_d}\right) + w_a a\,\mathrm{sim}(S_i) \quad (13)$$

where i_d denotes the location feature of ith sentence in the document, d; note that the document, d, means the original document with ith sentence. N_d denotes the total number of sentences within the document, d.

4. Empirical evaluation

4.1. Data sets and experimental settings

In our experiments for single-document summarization, we used the summarization data set of KOrea Research and Development Information Center (KORDIC). This data set is composed of 841 news articles. The articles consist of several topics such as politics, culture, economics, and sports. Each document has title, content, 30% summary, and 10% summary. The 30% and 10% summaries of the document are generated by hand.

For multi-document summarization, we constructed a multi-document data set from news articles. This data set has five clusters of 55 documents and 949 sentences; a 10% summary for each cluster and a 20% summary for each document were built by hand. Three people participated in summary-extraction tasks for the 20% summary of each document and the 10% summary of each cluster by voting mechanism. In case they selected different sentences for summary, they conducted their sentence extraction tasks again until they found the point of agreement. The contents of the multi-document summarization data set are shown in Table 1.

We have several parameters: the removing rate of bi-gram pseudo sentences in the first stage, the weight parameter of Aggregation Similarity method in Eqs. (11) and (13), window size in the sliding window technique, and the number of query words in the TF-based query method. For these parameter settings, we used 280 documents as a validation set which are selected at random. By simple experiments, the removing rate of bi-gram pseudo sentences was selected as 50%, the weight parameter of Aggregation Similarity method as 0.4, window size as two sentences (bi-gram pseudo sentence), and the number of query words as one.

The standard *precision*, *recall*, and F_1 measures are used to measure the performance of each sentence-extraction method as follows:

$$\mathit{precision} = \frac{\#\text{ of correct positive predictions}}{\#\text{ of positive predictions}} \quad (14)$$

$$\mathit{recall} = \frac{\#\text{ of correct positive predictions}}{\#\text{ of positive examples}} \quad (15)$$

$$F_1 = \frac{2 \times \mathit{precision} \times \mathit{recall}}{\mathit{precision} + \mathit{recall}} \quad (16)$$

Table 1
The composition of the multi-document summarization data set

Cluster	The number of documents	The number of sentences	Topic
1	8	109	A Korean Actress's Nude Scandal
2	15	190	Dr. Hwang, Woo-suk: the 'Stem Cell' Man
3	14	324	Korean Movies
4	11	166	Spain Terror
5	7	160	North Korea's Nuclear

The F_1 measure combines precision and recall with an equal weight. All the following performances are reported by it.

4.2. Experimental results for single-document summarization

The following experimental results for single-document summarization are reported in two cases; one case, denoted by 'with title', uses the title information and the other case, 'without title', does not use the title information.

4.2.1. Comparing the proposed method to other summarization methods in 'with title'

The proposed method is compared to other methods such as Title, Location, DOCUSUM, and MS word. Especially, DOCUSUM is a linguistic text summarization approach (Ko et al., 2003), and MS word denotes the summarizing results from a commercial software product (Microsoft Word). In addition, the Title & Location method is compared to the proposed method because the method denotes the case not using contextual information (bi-gram pseudo sentence) by only one sentence-extraction stage. The linear combination of Title and Location methods like Eq. (8) is used because the addition of the Aggregation Similarity method decreased the performance.

The experimental results are summarized in Table 2.

As shown in Table 2, the proposed method shows better performance than all other text summarization methods. Although DOCUSUM uses a knowledge resource such as the context vector space for lexical clustering, the proposed method outperformed DOCUSUM in both of 10% and 30% summary experiments. Especially, the proposed method using contextual information achieved better performance than the Title & Location method in both of 10% and 30% summary experiments.

4.2.2. Comparing the proposed method to other summarization methods in 'without title'

For 'without title' case, TF-based query method is used and verified instead of the title method. As you can see in Table 3, TF-based query method achieved significant performance and the proposed method using contextual information also achieved better performance than the Location & TF-based query method.

Table 2
The comparison between the proposed method and other text summarization methods in 'with title' for single-document summarization

	10% summary (F_1 score)	30% summary (F_1 score)
The proposed method	*53.4*	*55.3*
Title & Location	49.5	53.8
DOCUSUM	52.2	50.3
Location	46.6	49.4
Title	43.5	48.8
Aggregation Similarity	22.2	41.5
MS Word	12.8	27.2

Table 3
comparison between the proposed method and other text summarization methods in 'without title' for single-document summarization

	10% summary (F_1 score)	30% summary (F_1 score)
The proposed method	*47.9*	*50.4*
Location & TF-based query method	46.8	48.1
TF-based query method	46.5	45.6

4.3. Experimental results for multi-document summarization

4.3.1. Resetting the number of query words for the TF-based query method

The TF-based query method needs to reset the number of query words to be applied to multi-document summarization because they generally contain more meanings for a document cluster than a single document. Thus, we observed the performances of our system according to the number of query words by using a validation set of the multi-document summarization data set. The results are shown in Fig. 1.

As shown in Fig. 1, our method with five topic words achieved the best performance. Thus we used five topic words in all the experiments for multi-document summarization.

4.3.2. Comparing the proposed method to other methods in multi-document summarization

Since the MMR technique has been used for multi-document summarization systems widely (Kraaij et al., 2001; Dragomir et al., 2000), MMR-based summarization techniques are compared to the proposed method. As you can see in Table 4, we also achieved better performance by using contextual information (bi-gram pseudo sentence).

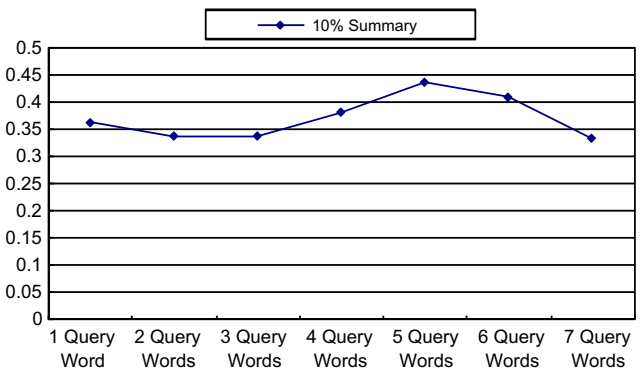

Fig. 1. The performance changes according to the number of query words in multi-documents summarization.

Table 4
The comparison between the proposed method and the other methods in multi-document summarization

	F_1 score
The proposed method	*51.6*
MMR ($\lambda = 0.3$)	48.2
MMR ($\lambda = 0.7$)	48.3
Title & Location	47.9

5. Conclusions

In this paper, we have presented a new hybrid sentence-extraction method using contextual information for single- and multi-document summarization. It used the contextual information (bi-gram pseudo sentence) to solve feature sparseness problem and the combination method of statistical approaches to improve the performance. As a result, the proposed method achieved higher performance than other summarization methods in both single- and multi-document summarization tasks, and the contextual information usage is effective in text summarization.

Moreover, the proposed method is independent on a kind of language. That is, our data sets are written by Korean, but all the statistical methods, which are used in the proposed method, and our hybrid method can be applied to any other languages. This is another strong point of the proposed method.

Acknowledgement

This paper was supported by Dong-A University Research Fund in 2008.

References

Barzilay, R., Elhadad, M., 1999. Using Lexical Chains for Text Summarization, Advances in Automatic Summarization. The MIT Press, pp. 111–121.

Carbonell, J., Goldstein, J., 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In: Proc. 21th ACM SIGIR Internat. Conf. on Research and Development in Information Retrieval.

Dragomir, R.R., Hongyan, J., Malgorzata, B., 2000. Centroid-based summarization of multiple documents: Sentence extraction, utility-based evaluation, and user studies. In: ANLP/NAACL Workshop on Summarization.

Edmundson, H.P., 1968. New methods in automatic extraction. J. ACM 16 (2), 264–285.

Goldstein, J., Kantrowitz, M., Mittal, V., Carbonell, J., 1999. Summarizing text documents: Sentence selection and evaluation metrics. In: Proc. ACM-SIGIR'99, pp. 121–128.

Jung, W., Ko, Y., Seo, J., 2005. In: Automatic Text Summarization Using Two-step Sentence Extraction, LNCS, vol. 3411, pp. 71–81.

Kim, J., Kim, J., Hwang, D., 2001. Korean text summarization using an Aggregation Similarity. In: Proc. 5th Internat. Workshop Information Retrieval with Asian Languages, pp. 111–118.

Ko, Y., Seo, J., 2004. Learning with unlabeled data for text categorization using a bootstrapping and a feature projection technique. In: Proc. 42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004), pp. 255–262.

Ko, Y., Kim, K., Seo, J., 2003. Topic keyword identification for text summarization using lexical clustering. IEICE Trans. Inform. System E86-D (9), 1695–1701.

Kraaij, W., Spitters, M., van der Heijden, M., 2001. Combining a mixture language model and naive bayes for multi-document summarization. In. Document Understanding Conf. (DUC 2001), New Orleans, USA.

Luhn, H.P., 1959. The automatic creation of literature abstracts. IBM J. Res. Develop., 159–165.

Mani, I., 2001. Automatic Summarization. John Benjamins Publishing Co., pp. 1–22.

Marcu, D., 1996. Building up rhetorical structure trees. In: Proc. 13th National Conf. on Artificial Intelligence, vol. 2, pp. 1069–1074.

Miller, G., Beckwith, R., Fellbaum, C., Gross, D., Miller, K., 1990. Introduction to WordNet: An on-line lexical database (special issue). Internat. J. Lexicogr. 3 (4), 234–245.

Nomoto, T., Matsumoto, Y., 2001. A new approach to unsupervised text summarization. In: Proc. ACM SIGIR'01, pp. 26–34.

Salton, G., 1989. Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer. Addison-Wesley Publishing Company.

Wang, J.C., Wu, G.S., Zhou, Y.Y., Zhang, F.Y., 2003. Research on automatic summarization of web document guided by discourse. J. Comput. Res. Develop. 40 (3), 398–405.

Wasson, M., 1998. Using leading text for news summaries: Evaluation results and implications for commercial summarization applications. In: Proc. 17th Internat. Conf. on Computational Linguistics and 36th Annual Meeting of the ACL, pp. 1364–1368.

Yang, J.C., Zhong, Y.X., 1998. Study and realization for text interpretation and automatic abstracting. Acta Electron. Sin. 26 (7), 155–158.

Zechner, K., 1997. Fast generation of abstracts from general domain text corpora by extracting relevant sentences. In: Proc. 16th Internat. Conf. on Computational Linguistics, pp. 986–989.