SciVerse ScienceDirect

Procedia - Social and Behavioral Sciences 27 (2011) 86 – 94

Procedia
Social and Behavioral Sciences

Pacific Association for Computational Linguistics (PACLING 2011)

Important Sentence Extraction Using Contextual Semantic Network

Jun Okamoto[a], Shun Ishizaki[b] a*

[a] *Keio Research Institute at SFC, Keio University, 5322, Endoh, Fujisawa, Kanagawa, 252-8520, Japan*
[b] *Graduate School of Media and Governance, Keio University, 5322, Endoh, Fujisawa, Kanagawa, 252-8520, Japan*

Abstract

In this paper, we propose a method for calculating important scores of sentences for text summarization. In this method, Contextual Semantic Network is used to calculate scores of importance for sentences included in input documents. The Contextual Semantic Network is constructed by using the Associative Concept Dictionary which includes semantic relations and distance information among the words in the documents. The concept dictionary was built using the results of association experiments which adopted basic nouns as stimulus words in Japanese elementary school textbooks. For evaluating the method, we compared the quality of the important score ranking obtained from our proposed method with that obtained from human subjects and that obtained from a conventional method using term frequency (tfidf). We used eight documents from the Japanese textbooks for the evaluation and carried out an experiment where 40 human subjects chose the five most important sentences from each of the eight documents. The results show that summarization accuracy can be improved by applying our method.

Keywords: Associative Concept Dictionary, Contextual Semantic Network, Important Score of Words, Important Sentence Extraction;

* Corresponding author. Tel.:+81-466-48-6101; fax:+81-466-48-6101
E-mail address: juno@sfc.keio.ac.jp

doi:10.1016/j.sbspro.2011.10.586

1. Introduction

Text summarization technologies for obtaining text contents are crucial in IT because electronic texts in WebPages are rapidly increasing. Text summarization methods generally require deep semantic processing and background knowledge. Many of the previous works, however, have used superficial clues [10] and ad hoc heuristics. They generate a summarized text by arranging the selected sentences in the order of their occurrences in the original text. In summarizing texts, the frequency of words occurrences in the document or the connectionist approach [1] has often been applied for calculating important scores of words. In those methods, the important scores are calculated as sum of scores of the words in the sentence [5][11]. A Summarization method using documents tagged by GDA (Global Document Annotation) can trim sentences in the summary [6]. The method uses propagation on an intra-document network based on GDA tags to calculate the important scores of text elements, for example, word, sentential segment, sentence and paragraph.

Background knowledge concerning input texts is necessary when a computer tries to understand the contents of the text as well as its syntactic and semantic information. In our previous research, an Associative Concept Dictionary was built based on the results of large-scale online association experiments [7]. The dictionary included not only semantic and contextual information about the stimulus words but also conceptual hierarchy information. Conventional concept dictionaries had tree structures for the hierarchy. Distances between the concepts in the dictionaries were calculated using the number of links between them, whereas the Associative Concept Dictionary explicitly includes quantitative distance information between pairs of concepts. This distance was calculated using a linear equation with two parameters, the associated word's frequency, and its order in the experimental results. The parameters are optimized by the linear programming method. In our previous research, the dictionary was shown useful for higher level contextual understanding system such as Word Sense Disambiguation. The Dynamic Contextual Network Model has been developed using the Associative Concept Dictionary which includes semantic relations among concepts/words and the relations can be represented with quantitative distances among them. In this model, the activation values on the Contextual Semantic Network are calculated using the propagation mechanism on the network [9].

In this research, the Contextual Semantic Network is used to calculate important scores for sentences given in the input document. In this paper, the method to extract important sentences is evaluated as follows. The results are compared with those of human summarization experiments using the same input texts. The comparison shows that the system performs well in summarization tasks, though the number of compared documents is few. It also shows effective use of conceptual information from the Associative Concept Dictionary in text summarization.

2. Associative Concept Dictionary

Background knowledge is crucial for computers to understand the contents of the text as well as its syntactic or shallow semantic information from input texts. The Associative Concept Dictionary (hereinafter referred to as ACD) has been built based on the results of large-scale online association experiments, which many subjects can use simultaneously in a campus network at Keio University [7]. In these experiments, the stimulus words were fundamental nouns chosen from Japanese elementary school textbooks and were presented to human subjects. The subjects were requested to associate words from the stimulus words with a given set of semantic relations, hypernym, hyponym, part/material, attribute, synonym, action and situation. All of the associated concepts are, in the ACD, connected to the stimulus words with distances calculated by a linear programming method. The distance D(x, y) between concepts, x and y, is shown by the following formula:

$$D(x, y) = 0.81F(x, y) + 0.27S(x, y)\,, \tag{1}$$

$$F(x,y) = \frac{N_x}{n_{xy} + \delta}\,,\quad \delta = \frac{N_x}{10} - 1,\qquad (N_x \geq 10)\,,\quad S(x,y) = \frac{1}{n_{xy}} \sum_{i=1}^{i=n_{xy}} s_{xyi}\,,$$

N_x denotes a number of the subjects who joined the experiments of stimulus word x, and n_{xy} denotes a number of subjects who input the associated word y with the same semantic relation for a given stimulus word x. Furthermore, δ denotes a factor introduced to limit the maximum value of $F(x, y)$ to 10, and s_{xyi} denotes an order of the associated word y by a subject i for a given stimulus word x.

The ACD is built using the quantified distances and is organized in a hierarchical structure in terms of the hypernym and hyponym. Attribute information is used to explain the features of the given word. In the association experiments, each stimulus word had 50 subjects who were students at Shonan Fujisawa Campus of Keio University. The number of stimulus words is currently 1100. Total number of associated words is about 280,000. And the number of associated words, when the overlapping words are not counted, becomes about 64,000 words. In Figure1, "chair" is a stimulus word for the association. "Furniture" is a hypernym of "chair". The numbers below <1> express the proportion of subjects who gave the same associated word, <2> an average of order of association and <3> a calculated conceptual distances.

	<1>	<2>	<3>
(chair			
(hypernym	↓	↓	↓
(furniture	0.92	1.02	1.09)
(object	0.04	2.50	7.43))
(hyponym			
(sofa	0.48	1.92	1.96)
(rocking-chair	0.28	1.43	2.64))
(part/material			
(wood	0.60	1.20	1.52))
(attribute			
(hard	0.46	1.17	1.82))
(synonym			
(seat	0.02	1.00	8.37))
(action			
(sit down	0.70	1.03	8.37))
(situation			
(school	0.30	2.40	2.78)))

Fig1. Concept dictionary description for a stimulus word "chair" (a part of associated concepts are presented. The stimulus word and associated words are originally in Japanese)

3. Extraction of Important Sentences Based on Word Scores

Text summarization has been accomplished by extracting important sentences from a document based on various superficial cues. In such conventional important sentence extraction methods, for example, the frequency of occurrence of a given word in a document has often been used in calculating the important scores of sentences. In this research, Contextual Semantic Network is developed using the ACD which includes semantic relations among concepts/words and the relations can be represented with quantitative

distances among them. In this method, propagation is used to calculate word's score on the Contextual Semantic Network where the activation values on the network are calculated using the distances.

3.1. Extracting Important Sentences Using Contextual Semantic Network

In the proposed summarization method, the Contextual Semantic Network (hereafter referred to as CSN) is used to calculate important scores of sentences included in the input document. Figure 2 shows construction of CSN according to an input document. We can use not only information obtained from word co-occurrence in their context but also that from comparatively rich network with quantitative distances and contextual information for extracting important sentences. The following steps show a procedure in detail for this network construction.

- Part of speech information for words (nouns, adjectives, adverbs, verbs and so on) and dependency information of each sentence in an input document are obtained by using Cabocha, a Japanese dependency structure analysis software [2].
- The CSN is constructed by using semantic relations extracted among the words from the ACD, where the information is obtained by the dependency structure analysis. When an input word, w_i, is included in the ACD as a stimulus word, a network around the w_i is constructed and added to the CSN. The new network starts from the stimulus word by tracing semantic relation paths until the distance accumulated becomes a certain numerical level
- Several links are added in the network based on a partial dependency structure among words

For example, let the input sentences be "There is a giant tortoise in Galapagos. This turtle is walking around in the island." In this text, "A giant tortoise" is a hyponym of "a turtle". The importance score of "a turtle" is calculated using the distance between "turtle" and "giant tortoise" in the CSN. "A turtle" is a hypernym of "a giant tortoise". "Galapagos" is a situation of "a giant tortoise". "Walk" is a verb concept of "a giant tortoise" and also that of "a turtle". We can construct an intra-document network, because all words are including the ACD. In addition, some hypernym words are added in the CSN such as "living-thing" which is a hypernym of "a turtle".

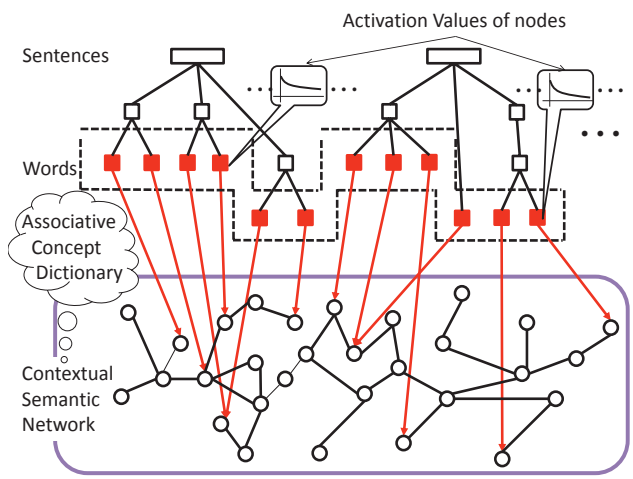

Fig 2. Example of CSN for text summarization

The activation value of each node is calculated by a spreading activation method on the CSN. Initial values ($V_a(0)$) of words in sentence k of input document are calculated by the following equation (2). Next, the activation value of each node N_a is calculated by the equation (3).

$$V_a(0) = 1000 * S(j, N_a), \tag{2}$$

$$V_a(t+1) = V_a(t) - \theta V_a(t) + \sum_{b=1} V_b(t) / D_{ab} \tag{3}$$

where the calculation is repeated until $\Sigma_{a=1}(V_a(t)-V_a(t+a))^2$ reaches a certain small value. The decay parameter θ is assumed to be 0.1. $V_a(t)$ is an activation value of node N_a at time t. $S(j,N_a)$ is a number of node (word) N_a appearing in document j. D_{ab} is a distance between two concepts. $V_b(t)$ is an activation value of the node connected with node N_a . α is a value where the total number of links is divided by the maximum distance. P_{jkl} is score of words (N_a) in sentence k in document j. For each important sentence score (T_{jk}), the sum of the keywords-weights is divided by the number of words (L_{jk}) as shown in the following formula:

$$T_{jk} = \sum_{l=1}^{L_{jk}} P_{jkl} \Big/ L_{jk} \tag{4}$$

$j = 1, 2, \ldots N$, (N is the number of documents)
$k = 1, 2, \ldots M_j$, (M_j is the number of sentences in the document j)
$l = 1, 2, \ldots L_{jk}$, (L_{jk} is the number of words in the kth sentence in the document j).

3.2. Extracting Sentences Based on Word Frequency

We use the following important sentence extraction methods to compare them with our proposal methods. Input texts are analyzed morphologically using the Japanese morphological analysis system Chasen [4]. Next, we correct errors included in the morphologically analyzed data. The important scores are calculated using the root morphemes of nouns, adjectives, adverbs, and verbs. Pronouns and certain nouns (number, counter suffix and so on) are not included in the calculations

In this method, the word, w_{jkl}, is defined as the lth word in the kth sentence in the document j. The important score of word w_{jkl}, P_{jkl}, is calculated by the following formula (5), where F_{jkl} is the frequency of word w_{jkl} in the document j, N is the total number of documents and n_{jkl} is the number of documents where the word w_{jkl} appears.

In those methods, sentence scores (T_{jk}) are calculated by summing the scores of the words in the sentence. The score of sentences k is given by the following formula:

$$P_{jkl} = \sum_{k=1}^{L_{jk}} \sum_{j=1}^{M_j} F_{jlk} * \log N / n_{jkl} \tag{5}$$

$$T_{jk} = \sum_{l=1}^{L_{jk}} P_{jkl} \Big/ L_{jk} \tag{6}$$

$j = 1, 2, \ldots N$, (N is the number of documents)
$k = 1, 2, \ldots M_j$, (M_j is the number of sentences in the document j)
$l = 1, 2, \ldots L_{jk}$, (L_{jk} is the number of words in the kth sentence in the document j).

When we calculated important score of the words in the text, "There is a giant tortoise in Galapagos. This turtle is walking around in the island," shown in section 3.1, the score of both "A giant tortoise" and "a turtle" are high. A word frequency method, however, treats the words, "A giant tortoise" and "a aurtle" as no-related ones.

4. Evaluations and Discussion

We use eight documents extracted from Japanese elementary school textbooks because the ACD is constructed by using basic nouns from these textbooks. Those documents contain about 17 sentences on average (range 10-23). All the documents focus on a single topic of natural science and consist of titles (headings) and documents (bodies) [8].

In this section, in order to evaluate the ranks of importance of sentences in the documents calculated by our method, we carry out an experiment where 40 human subjects chose five most important sentences from each document. We will show the effectiveness of our system using the quantitative conceptual distance information from the ACD. We will compare it with ranks calculated by conventional methods based on word frequencies.

4.1. Important Sentence Extraction by Human Subjects

The human subjects are all students at Keio University who are all native Japanese speakers. We provided questionnaires for each document, asking the subjects to choose the five most important sentences from it. To fill out the questionnaire they are requested to read a title and a text carefully, to arrange the sentences in the order of their importance in the document and to extract the top five sentences. The important scores are from 5 to 1. The important score 5 is given to the most important sentence and the score 1 is given to the fifth one. Next, the important sentences are sorted according to their sum of the important scores given by the subjects.

Next, we calculate degree of coincidences by using Kendall's coefficients of concordance among human subjects result of extracting important sentences. Table 1 shows Kendall's coefficients of concordance (W), and the numbers of sentences in the 8 documents. The coefficients are calculated for the documents. Kendall's coefficients of concordance show a degree of agreement among the sentences ordered by the human subjects. A high W value means that the ordering results by the human subjects are consistent with each other.

The sentences in a document are ordered according to their importance scores as given by the human subjects. Average ranks are given to the other sentences which each human subject did not select within the top five.

The Kendall's coefficients of concordance (W) of three documents (D3, D4 and D7) are relatively low than the one of the other five documents. Therefore the agreements among the sentence ordered by human subjects are relatively low in terms of three documents (D3, D4 and D7) than the one of the other five documents (D1, D2, D5, D6 and D7). However, chi-square statistic suggests that there is not a statistically significant difference. As the result, all the coefficients (W) are found significant level at 0.01.

Table 1 Kendall's coefficients of concordance (W) for the sentence ranks of the documents

Document No.	D1	D2	D3	D4
Number of sentences	15	18	22	21
Kendall's W	0.44	0.45	0.35	0.38
Significance	***	***	***	***

Document No.	D5	D6	D7	D8
Number of sentences	9	9	19	18
Kendall's W	0.57	0.54	0.37	0.48
Significance	***	***	***	***

4.2. Evaluation of Our Method and Conventional Method

In order to show the effectiveness of the ranks of sentences obtained by our system, we compare important sentences extracted by our method with those extracted by the human subjects and those obtained by the conventional methods based on word frequencies automatically.

First, important scores from 10 to 6 are given to the best five sentences, respectively, which the human subjects chose. For a sentence chosen both by the human subjects and by one of the other methods, the value of correspondence (*C*) is calculated by the following formula [8].

$$C = \sum_{i=1}^{5} R_i(hs,m) - \Delta r_i(hs,m) \tag{7}$$

where $R_i(hs,m)$ is important scores (range 10-6) of the best five sentences which human subjects extracted, and $\Delta r_i(hs,m)$ is a value obtained by subtracting the rank of sentences extracted by human subjects from the rank of sentences extracted by the system. The above formula compares the ranks by the methods (our method and the word frequency method) with those by the human subjects.

Table 2 An example of calculating the value of correspondence (*C*)

The rank of sentences	1	2	3	4	5	C
Sentence numbers extracted by human subjects	15	10	2	6	8	-
Sentence numbers extracted by the proposed method	15	8	10	14	5	21

C = 10 + (9 - 1) + (6 - 3) = 21

Correspondence of extracted sentence is 3.

Table2 shows an example of calculating the values of correspondence (*C*) by using the number of sentences extracted by automatic methods which correspond to the sentences extracted by human subjects. The value of correspondence is calculated by formula (7) such as *C = 21*.

Table 3 The value of correspondence C and correspondence of sentences extracted by our method and the word frequency methods

	Our method		tfidf	
Document	C	correspondence of extracted sentences	C	correspondence of extracted sentences
D1	29	4	22	3
D2	26	4	15	3
D3	6	1	6	1
D4	16	3	13	3
D5	20	3	14	3
D6	24	3	25	4
D7	17	2	16	2
D8	21	3	6	1
average	19.9	2.9	14.6	2.5

Table 3 gives a comparison of our method with the word frequency methods. It compares the best five sentences which our method chose by using CSN, and those which the method chose by using word frequencies (tfidf). *C* is a value of correspondence, which is calculated by the formula (7). The correspondence of extracted sentences in the table gives the number of sentences in the best five sentences extracted by automatic methods which correspond to the sentences extracted by human subjects. In Table 3, the value of correspondence (*C*) for our method is high. The results show that this method, using the CSN, is more effective than the word frequency method. The comparison is made using the important sentences extracted by human subjects as a reference. The ability of our system provided better results than the conventional method. The values of correspondence (*C*) of three documents (D3, D4 and D7) are relatively low than the one of the other documents. These results correspond with the result of Kendall's coefficients of concordance (*W*) which are relatively low (see table 2)

5. Future Works

In this paper, we used the distances information between concepts in the ACD to construct the CSN. As a next step, we will improve the CSN by using context sensitive distance information and semantic relations in the ACD. Such improvement will be able to reflect the context dependency of the distances between two concepts in the document.

One of the major problems in summarizing texts is the analysis of anaphora expressions. Another one is keeping consistency among the extracted sentences. We will try to improve the accuracy of our summarization method using various methods by modeling human summarizing methods for these problems. In our future work, we propose a method which can trim sentences for summary by using the important scores which assigned to words, phrase and clause which are of smaller size than sentences.

The CSN will be useful to generate summaries using abstract words and hypernym which are not included in the input documents.

References

[1] K. Hashida, S. Ishizaki, and H. Isahara. 1987. Connectionist approach to the generation of abstracts, natural language generation. New Result in Articial Intelligence Psychology and Linguistics, pages 149-156.

[2] K. Ishikawa, S. Ando, and A. Okumura. 2002. Evaluating text summarization using multiple correct answer summaries. Journal of NLP Vol.9 No.4, pages 33-53.

[3] T. Kudo, Y. Matsumoto, 2002. Japanese Dependency Analysis using Cascaded Chunking, Proceedings of the 6th Conference on Natural Language Learning 2002 (COLING 2002 Post-Conference Workshops), pp 63-69.

[4] Y. Matsumoto, A. Kitauchi, T. Yamashita, Y. Hirano, Matsuda H., and M. Asahara. 1999. Japanese morphological analysis system chasen manual version 2.0 manual 2nd edition (in japanese). Technical Report NAIST-IS-TR99009, Institute of Science and Technology.

[5] H. Mochizuki and M. Okumura. 2000. A comparison of summarization methods based on task-based evaluation. In LREC2000, pages 633-639.

[6] K. Nagao and K. Hashida. 1998. Automatic text summarization based on the global document anotation. In COLING-ACL, pages 917-921.

[7] J. Okamoto and S. Ishizaki. 2001. Construction of associative concept dictionary with distance information, and comparison with electronic concept dictionary. Journal of NLP Vol.8 No.4, pages 37-54.

[8] J. Okamoto and S. Ishizaki. 2003. Evaluating a Method of Extracting Important Sentences using Distance between Entries in an Associative Concept Dictionary, Journal of NLP Vol.10 No.5, pages 139-151.

[9] J. Okamoto and S. Ishizaki. 2010. Homographic Ideogram Understanding Using Contextual Dynamic Network, Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC 2010).

[10] H. Watanabe. 1996. A method for abstracting newspaper articles by using surface clues. In the 16th International conference on Computational Linguistics, pages 974-979.

[11] K. Zechner. 1996. Fast Generation of Abstracts from General Domain Text Corpora by Extracting Relevant Sentences, Proceedings of the 16th International Conference on Computational Linguistics, pages 986-989.