ISSN 1746-7659, England, UK
Journal of Information and Computing Science
Vol. 8, No. 3, 2013, *pp*. 227-234

Sentence Ordering Algorithm with Subject Criterion for Automatic Multi-Document Summarization

Naser Jawas, Randy Cahya Wihandika, and Agus Zainal Arifin

Department of Informatics, Faculty of Information Technology,
Institut Teknologi Sepuluh Nopember (ITS) Surabaya, Indonesia.

(*Received January 03, 2013, accepted May 3, 2013*)

Abstract. In multi-document summarization, order of sentences in summarization result must be coherence and it must represent information in correct steps to make it easy to understand by the reader. Problem arises when some subject of sentences are represented using pronouns. The subject pronoun in an incorrect sentence order will confuse the reader as the pronoun can refer to more than one subject. In this paper, we propose a new subject criterion for sentence ordering strategy to complement the existing ordering strategy. Sentences will be clustered based on its subject and will be ordered with respect to subject levels. We test the system using Document Understanding Conference summarization data and compare it with results from existing algorithm without using subject criterion. The accuracy of ordering with all criterions including subject criterion is 83%. The result shows that there is a slight improvement in ordering accuracy when subject criterion is included.

Keywords: information retrieval, multi-document summarization, sentence ordering.

1. Introduction

Sentence ordering is one of important task in information retrieval. It found a place in many applications that need ordering information such as document summarization. Document summarization can be divided into 2 general applications by the number of documents to be summarized, namely single-document summarization and multi-document summarization. In single-document summarization, the summary is extracted from one document only while in multi-document summarization, there are more than 1 source document.

Creating summary from single document is quite straightforward, where the the sentences are ordered based on the occurrence in the original document. However, in multi-document summarization, order of sentences must be coherence and must represent information in correct steps to make it easy to understand by the readers. One of the problems in multi-document summarization is that the source documents are written by different authors in different time and different perspective. Each author may also have different level of knowledge. This problem does not appear in single-document summarization because the sentences in single-document summarization can be ordered based on the order in original document.

There are some previous researches concerned in ordering sentences for document summarization. [1] proposed a combination of machine learning and statistical technique to find similar paragraphes and then order the sentence chronologically. [2] improved the chronological sentence ordering with adding topical relatedness. Each sentence are grouped in same topics and ordered in each groups.

A probabilistic approach was presented by [3]. The system learns which sequence of sentences that commonly appear together and makes prediction based on known orders. It uses a large corpus. Each sentence is extracted into a set of informative features that are automatically collected from the corpus.

[4] proposed an improvement for chronological ordering with sentence precedence relation. Relation precedence refines order from chronological ordering with information of segment orders in source documents. They assumed this case in newspaper articles that authors usually arrange information using time series.

A semi-supervised sentence classification and historical ordering proposed by [5]. Their method is divided into 3 parts. First, create summary sentence neighborhood network. The network is based on

similarity between summary sentences with weights on edge are considered as transition probabilities. Second, make classification of document sentences with class label is taken from summary sentences. Third, they extract sentences from the network and order it based on original position of their partners in the same class.

[6] combines 4 criterions on sentence ordering. The criterions are chronological, topical closeness, precedence and succession. These 4 criterions are combined into one criterion using Support Vector Machine (SVM). The paper shows good results in sentence ordering. The problem arises when subject of sentences are represented using pronouns. Subject pronoun in an incorrect sentence order will confuse the reader as the pronoun can refer to more than one subject. The example is shown in Figure 1. If the second and third sentence is shown before the first sentence, it will confuse the reader because the pronouns do not show the correct subjects they refer to.

In this paper, we propose a new subject criterion as a criterion for sentence ordering strategy to complement the existing ordering criterions by [6]. The method clusters sentences based on its subject and then sentences will be ordered with respect to subject levels in the clusters.

Barack Obama nominated Hillary Rodham Clinton as his secretary of state on Monday.

He chose her because she had foreign affair experience as a former First Lady.

Although Hillary was Obama's rival as Democratic Nomination, she decided to support Obama's campaign to make Democratic win over Republic.

Fig. 1: Example of sentence ordering using subject pronoun.

2. Previous Ordering Criterion

Firstly, we define the standard convention for this paper. The small letters are used to describe the sentence and capital letters is for segment. Segment is a set of sentences that written in order. The order is represented by ≻. For example, there are 5 sentences in 1 order a≻b≻c≻d≻e. These sentences are divided into 2 segments A and B. For example, segment A contains a≻b and segment B contains c≻d≻e. The orders of segments are also represented with ≻. The segment order in A≻B means segment A comes before segment B.

There are 4 previous criterions that are used by [6]. The criterions are chronology, topical closeness, precedence, and succession. Each criterion has an output value between 0 and 1. The value 0 means the sentences are in the wrong order and 1 means they are in the correct order. Support vector machine are used for combining the criterions.

2.1. Chronology Criterion

This criterion is the most applied criterion in sentence ordering systems. It is commonly used in news summarization. Chronological criterion idea originally came from [1]. It uses timestamp from news documents as key point to decide order of summary sentences. This criterion will not work good if some of source documents do not have timestamp. However, a number of researches earlier have proposed ways to overcome this problem.

In [6], the chronological criterion orders sentences with four conditions. It gives a value for each condition. Let a be the first sentence and b the second sentence. The order will be a≻b if their original documents are different and timestamp of documents a is less than documents b. If their original documents are the same, the order will stay a≻b if line number of a is less than b. If original documents are different but they have same timestamp, the sentences are not ordered. When none of condition above are matching, the order will be reversed b≻a. The author uses a value for each condition between 0 and 1. Score 0 is for reverse action (b≻a), score 0.5 for not ordering action, and score 1 for order a≻b. Here are formulas for doing the chronological ordering above:

$$f_{chro}(A \succ B) = \begin{cases} 1, & T(a) < T(b), \\ 1, & D(a) = D(b) \wedge N(a) < N(b), \\ 0.5, & T(a) = T(b) \wedge D(a) \neq D(b), \\ 0, & otherwise, \end{cases} \quad (1)$$

where T(s) is the timestamp of original document, D(s) is the document id and N(s) is the line number of the sentence in the original document.

2.2. Topical Closeness Criterion

Topical closeness idea is proposed by [2]. In [6], this criterion is applied with cosine similarity. It works by grouping sentences based on the similarity. The grouping process uses similarity value. For example, if there are three sentences a, b and c. Similarity between a and b is calculated and also similarity of a and c. If a≻c is more similar to a≻b then the order will be a≻c≻b, otherwise the order will be a≻b≻c.
The similarity of two sentences is calculated from vector of word frequencies in each sentence. [6] use the following formula for the topical closeness criterion given the sentence a and b:

$$f_{topic}(A \succ B) = \frac{1}{|B|}\sum_{b \in B} \max_{a \in A} sim(a, b), \quad (2)$$

here sim(a,b) is similarity calculation using cosine similarity.

2.3. Precedence Criterion

This criterion idea originally comes from [4]. [6] puts this criterion on his proposed method. The idea is to find a near similar sentence which occurs before sentence b in the original document. If sentence a is supposed to be placed before sentence b, there should be a high similarity between sentence a and the sentences preceding sentence b in the original document. For example, compare sentence a and sentence b, search similar sentence with sentence a in original document where sentence b taken from. If there is a sentence with high similarity with sentence a and it is located before sentence b, then the sentences a≻b is in the correct order. Formula for this criterion is:

$$f_{pre}(A \succ B) = \frac{1}{|B|}\sum_{b \in B} \max_{a \in A, p \in P_b} sim(a, p), \quad (3)$$

where p is the sentence that is taken from sentence b original document.

2.4. Succession Criterion

Succession criterion idea is nearly same as precedence criterion. While precedence criterion looks for similar sentences before the compared sentence, the succession criterion finds the similarity between sentence b and the sentences that occurs after sentence a in the original document. Cosine similarity is also taken a part here for calculating similarity between sentences. [6] use the following formula for calculating the succession:

$$f_{suc}(A \succ B) = \frac{1}{|A|}\sum_{b \in B} \max_{s \in S_a, b \in B} sim(s, b), \quad (4)$$

here s is the sentences taken from original document of sentence a.

2.5. Support Vector Machine

[6] use support vector machine (SVM) to combine all criterion functions to find the best mixture function which can result close to human summary. The strategy is to make a partition between sentences. For example, there are four sentences a≻b≻c≻d. First, put the separator between sentence a and b, and the following segments are obtained: {(a), (b)}, {(a), (b≻c)} and {(a), (b≻c≻d)}. Second, put the separator between c and d. This produces segments: {(b), (c)}, {(a≻b), (c)}, {(a≻b), (c≻d)}, Third, move the separator between c and d gives these segments: {(c), (d)}, {(b≻c), (d)}, {(a≻b≻c), (d)}. Total segments for four sentences are 10 pairs. Generally, this process produce n(n2 - 1) / 6 segments.

Pair of segments is calculated using all functions criterion and assigned to positive class if the pair in correct order and negative class if the pair in wrong order. The SVM output gives final order direction. For example, with pair A and B, set the class to be positive class (+1) for the set of vector obtained from all functions in correct A≻B order: $f_{chro}(A \succ B)$, $f_{topic}(A \succ B)$, $f_{pre}(A \succ B)$, $f_{suc}(A \succ B)$. The opposite order (wrong

order) B≻A must be given negative class (-1). The vectors for training phase are:

$$+1 = \left[f_{chro}(A \succ B), f_{topic}(A \succ B), f_{pre}(A \succ B), f_{suc}(A \succ B)\right], \quad (5)$$
$$-1 = \left[f_{chro}(B \succ A), f_{topic}(B \succ A), f_{pre}(B \succ A), f_{suc}(B \succ A)\right]. \quad (6)$$

3. Methodology

This research uses the summarization data from Document Understanding Conference (DUC) 2005. DUC 2005 has summary results from 32 automatic multi-document summarization systems. The summaries have been grouped by topics and summarization systems along with the original documents. For each summary, it has between 8 to 15 sentences. The original documents are also used in some criterion. Input full documents are parsed into sentences in preprocessing step.

Subject criterion is proposed to refine sentence ordering by [6]. The main idea is to extract subject of sentences and compare it with subject from different sentences. The sentence can be an active form or passive form.

The output for this criterion is a subject similarity value. Similarity between 2 subjects from different sentences can be calculated using following formula:

$$f_{subj}(a \succ b) = 2 \times \frac{mw}{tw}. \quad (7)$$

where mw is number of match word in subject of a and b, two is total words from both subject.

Subject extraction from a sentence is done with Stanford NLP (nlp.stanford.edu/software/corenlp.shtml). The library can detect the main subject of a sentence. The formula (7) is used to calculate similarity between two sentences after subject words are extracted. There are some rules applied to get the main subject. Stanford NLP uses dependency tree to display correlation of words in a sentence. The tree contains information about the roles of each word in a sentence. From the tree, subject phrase of verb are extracted. Since there may be more than one verb in a sentence, the subject of the first verb is assumed to be the subject of sentence if the sentence in active form. If the sentence is in passive form, the subject is found after 'by' word. If it is not found, then the passive subject is used as subject. For example, the sentence is "The work is done by Budi". Word "Budi" is selected as the subject because it is the main actor. But, if the sentence is just "The work is done", then "The work" is considered as the main subject.

SVM is used as learning strategy to combine all the criterions. Criterion outputs are combined as input vector. There is a difference between [6] and our proposed method in SVM learning. [6] uses segment as the input for each criterion. However, we use sentence for input in each criterion. Therefore, only 2 sentences are compared in one input. Here is the updated formula for the changes:

$$f_{topic}(a \succ b) = sim(a, b), \quad (8)$$
$$f_{pre}(a \succ b) = \max_{p \in B} sim(a, p), \quad (9)$$
$$f_{suc}(a \succ b) = \max_{s \in A} sim(s, b), \quad (10)$$

p is the precedence sentence in document B, s is succession sentence in document A. $sim(a, b)$ is similarity calculation using cosine similarity. The chronology criterion is not changed. It can be used with 2 sentences.

The input vectors for SVM are as follows:

$$+1 = \left[f_{chro}(a \succ b), f_{topic}(a \succ b), f_{pre}(a \succ b), f_{suc}(a \succ b), f_{subj}(a \succ b)\right], \quad (11)$$
$$-1 = \left[f_{chro}(b \succ a), f_{topic}(b \succ a), f_{pre}(b \succ a), f_{suc}(b \succ a), f_{subj}(b \succ a)\right], \quad (12)$$

positive class is given to the correct order and negative class is given to the wrong order.

All sentences from 200-word DUC summaries are parsed into sentences to form pairs of sentences. For example, from five sentences a, b, c, d, and e, the sentence pairs are (a≻b), (a≻c), (a≻d), (a≻e), (b≻c), (b≻d), (b≻e), (c≻d), (c≻e), and (d≻e). Those sentence pairs represent the correct sentence ordering, thus used for the positive class. Obviously, for the negative class, the reversed version of the pairs is used, that represent the wrongly ordered sentences. In the next step, all criterion functions are used to calculate order value for each sentence pairs. In the training phase, each of these combinations is calculated twice for positive and negative class using all the above mentioned criterions.

The experiments scheme uses the summary results from multi-document summarization system. The results are parsed and the sentence order is randomized. The experiments are in two schemes. First, the combination of previous criterion with subject criterion is tested. And the second experiment uses previous criterion without subject criterion. The results of these experiments are compared to measure the effectiveness of the proposed subject criterion.

4. Result and Discussion

The experiments are in 2 schemes. First experiment is sentence ordering with all criterions including subject criterion. The second experiment is sentence ordering with all criterions without subject criterion.

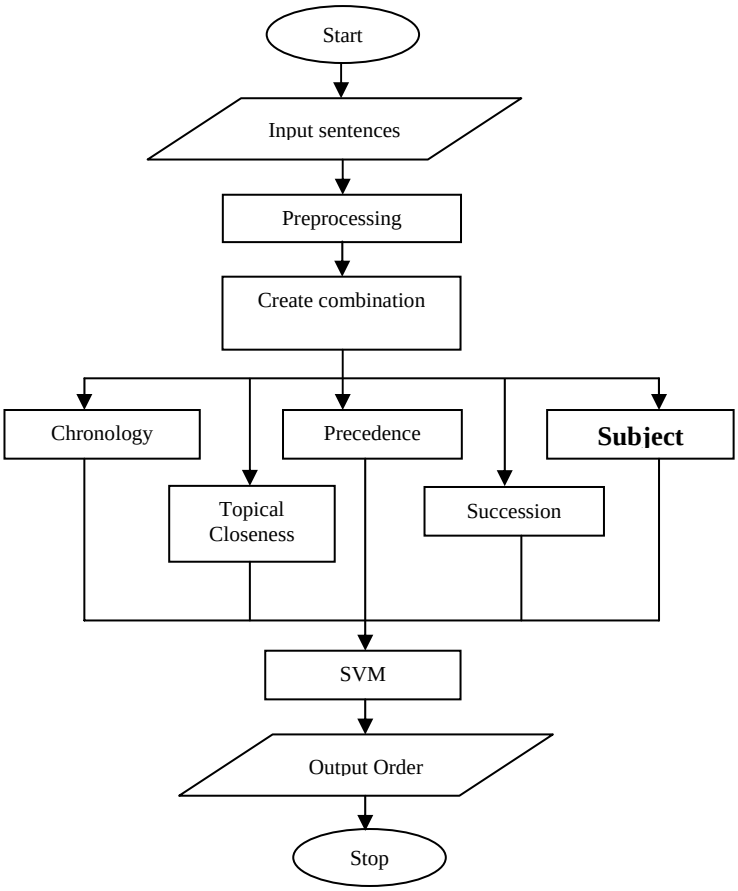

Fig. 2: The Overall System and Main Contribution

4.1. Result of sentence ordering with all criterions

In this experiment, sentences in each summary are ordered with all five criterions including the proposed subject criterion to measure the performance of the overall system. Figure 3 depicts an example of subject extraction from two sentences. The bold-typed words are the extracted noun phrase as the sentence subject. It is worth noting that the subject criterion does not provide score for a single sentence as in chronology criterion. Therefore, it cannot produce a total sentence ordering. Instead, it provides score for each pair of sentence since it works as a similarity measure between subjects in two sentences. Thus, this criterion alone can only produce sentence clusters with similar subjects in each cluster. Here, agglomerative clustering method is used. In addition, this criterion also cannot determine the arrangement of the sentence clusters. Figure 4 depicts the result of the sentence subject clustering where sentences with similar subjects are placed next to each other.

The next step on this first experiment is to try to calculate the sentence orders using all criterions including the new subject criterion. All criterions are used in both training and testing of the classifier. Support vector machine is used as the classifier with linear kernel function. The C parameter is set to 0.1 as

the trade-off between margin maximization and error minimization. The testing result for this experiment gives 81% accuracy.

Christopher Parkes examines the personalities and issues in what one German judge said could be 'potentially the biggest-ever case of industrial espionage'.
If charges were made, it would be 'a case of industrial espionage of unbelievable proportions,' **Mr Hughes** said.

Fig. 3: The result of subject extraction.

At least 158 Colombians have been arrested in Mexico on drug charges over the last six years, **Mexican officials** say.
More commonly, **Mexican officials** say, the planes fly nonstop from Colombia up the Pacific and Gulf coasts of Mexico, then cut inland to isolated landing strips in the states of Chihuahua, Nuevo Leon, Sonora, Sinaloa and Baja California.
The **Mexican air force** will be employed to chase down the craft.
Federal investigators believe the cartel has purchased real estate in the United States, something the bank records could corroborate, according to Djinis.
A **federal grand jury** in Tampa, also in February, 1988, indicted Noriega on three counts of marijuana trafficking.
Spanish police arrested Ochoa at U.S. request, but a court ordered him extradited to Colombia, where he was soon released.
Spanish authorities confiscated 3,461 kilos of the total, more than half of it from ships' cargo.
In another raid on a busy weekend, **Spanish police** seized 109 kilos of Turkish heroin in southern Spain and arrested two Turks and a British woman.

Fig. 4: Sentences arranged using subject similarity.

Table 1. Sentences with wrongly extracted subject.

No	Sentence	Subject
1	The defence, by calling current and former government officials and requiring them to read documents they wrote on Gen Noriega's efforts in drug law enforcement, is attempting to counter the prosecution claim that Gen Noriega turned Panama into a haven for Colombian drug dealers.	The calling requiring defence
2	But letters and memoranda by Mr Sedillo and Mr Bramble also referred to cooperation between the DEA and Panamanian forces under Gen Noriega's command.	memoranda Sedillo Bramble letters
3	Hundreds of kilogrammes of Asian heroin, mostly from India, Thailand and the Golden triangle, and cocaine from South America pass through Lagos for storage and repackaging before onward shipment to the US and Europe.	cocaine kilogrammes Thailand Hundreds

Table 2. Sentence ordering result.

All criterions with subject criterion	All criterions without subject criterion
0.83	0.81

4.2. Result of sentence ordering without subject criterion

Since the subject criterion is new in this method, it needs to be evaluated to know how well it contributes to the overall performance in addition to other criterions. The second experiment is done similarly with the previous one, but by leaving out the subject criterion in both the training and testing process. The parameters of classifier are set to be the same as the first experiments. The result is then compared with the result produced using all criterions as shown in Table 2. The accuracy without subject criterion is 81%. It is lower than the accuracy with subject criterion.

4.3. Discussion

Figure 3 shows the successful subject extraction. Since this part is using Stanford NLP, the subject extraction is based on the main rule of it. However, in some experiment we cannot get the desired words, as shown in Table 1. Therefore, we put additional rule to make sure the subject words are extracted. The rule of active-passive is used to fulfill the need of subject in every sentence in this criterion. These rules run well with assumption that subject of passive sentence can be found after 'by'. If there is no 'by' word found in the sentence, it selects the passive subject as the main subject.

Figure 4 shows the result of subject clustering. This result shows example of successful subject clustering process. First and second sentence have exactly the same subject. The third sentence has 1 similar word with the first and second sentence. These sentences are the first cluster. The second cluster is build from the fourth and fifth sentence. They share one similar word. The third cluster is made from sixth, seventh and eighth sentence which have one word in common. The subject criterion does not determine the order, therefore the third cluster is not ordered as the same subject.

The successful of subject extraction determines the result of subject criterion. The subject extraction sometimes do not give a good result, therefore it still need some rules to ensure all the sentence subject are extracted correctly.

The classification accuracy is measured simply as the fraction of sentences correctly ordered. Table 2 shows the classification results using all criterions and without the subject criterion. The sentence ordering without using the subject criterion yields a slightly lower accuracy. It indicates that the subject criterion increases the classification accuracy although by a very small score. It will probably give a better contribution for other forms of data.

We have attempted to do the classification using other kernel functions such as radial basis function (RBF) and polynomial but they do not give satisfactory results as given by the linear function. Therefore, the linear function is used in all our experiments.

5. Conclusion

This paper presents the new criterion for multi-document summarization ordering. It uses subject as order criterions. The subject of sentences is extracted using StanfordNLP and the sentences are clustered based on their subject similarity. The experiments are done with two schemes. The first scheme is to order the sentences with all criterions including subject criterion and the second scheme is without the subject criterion. The result shows that there is a slight improvement in ordering accuracy when subject criterion is included.

6. References

[1] K.R. McKeown, J.L. Klavans , V. Hatzivassiloglou, R. Barzilay, E. Eskin, *Towards Multidocument summarization by reformulation: progress and prospects*, American Association for Artificial Intelligence, (1999), pp. 453-460.

[2] R. Barzilay, N. Elhadad, K. McKeown. *Inferring strategies for sentence ordering in multidocument news summarization*, Journal of Artificial Intelligence Research 17 (2002), pp. 35-55.

[3] M. Lapata, *Probabilistic text structuring: Experiments with sentence ordering*, Proc. of the annual meeting of ACL, pp. 545-552, 2003.

[4] N. Okazaki, Y. Matsuo, M. Ishizuka, *Improving chronological sentence ordering by precedence relation*, Proc. of 20th international conference on computational linguistics, pp. 750-756, 2004.

[5] P.D. Ji, S. Pulman, *Sentence ordering with manifold-based classification in multi-document summarization*. Proc. of empirical methods in natural language processing, pp. 526-533, 2006.

[6] D. Bollegala, N. Okazaki, M. Ishizuka, *A bottom-up approach to sentence ordering for multi-document summarization*. Information Processing and Management 46 (2010), pp. 89-109.