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 $\succ$. For example, there are 5 sentences in 1 order $a \succ b \succ c \succ d \succ e$. These sentences are divided into 2 segments A and B. For example, segment A contains $a \succ b$ and segment B contains $c \succ d \succ e$. The orders of segments are also represented with $\succ$. The segment order in $A \succ B$ 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 $a \succ b$ 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 $a \succ b$ 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 $b \succ a$. The author uses a value for each condition between 0 and 1. Score 0 is for reverse action ($b \succ a$), score 0.5 for not ordering action, and score 1 for order $a \succ b$. Here are formulas for doing the chronological ordering above: