# Analysis of Sentence Ordering Based on Support Vector Machine

Gongfu Peng, Yanxiang He, Ye Tian, Yingsheng Tian, Weidong Wen
*Computer School, Wuhan University*
*Wuhan 430079, P.R.China*
*bluewuhan@163.com, yxhe@whu.edu.cn, tianye@gmail.com, tianyingsheng@gmail.com, wwd@whu.edu.cn*

***Abstract*—In this paper, we present a practical method of sentence ordering in multi-document summarization tasks of Chinese language. By using Support Vector Machine (SVM), we classify the sentences of a summary into several groups in rough position according to the source documents. Then we adjust the sentence sequence of each group according to the estimation of directional relativity of adjacent sentences, and find the sequence of each group. Finally, we connect the sequences of different groups to generate the final order of the summary. Experimental results indicate that this method works better than most existing methods of sentence ordering.**

***Keywords*-Sentence ordering; SVM;**

## I. INTRODUCTION

In multi-document summarization tasks, how to extract sentences from source document is a major work. But for a fluent and readable summary, it is not enough. Recent research indicates that sentence ordering in summary should get more attention. Barzilay has offered empirical evidence that proper order of extracted sentences would greatly improve the readability of a summary [1].

Sentence ordering is much easier in single-document summarization, because single document provides a natural order of sentences in summary based on source document. Differently, in multi-document summarization tasks, multi-documents contribute sentences of different authors and in different writing styles, which means source document can not directly provide ordering criterion in multi-document summarization task.

Obviously, sentence ordering in multi-document summarization task involves two fields, information in source documents and experiential knowledge of human. Neither of them can be easily handled, because both of them involve semantic knowledge more or less, finding feasible methods that suitable for computer is important for sentence ordering. Fortunately, large raw corpus can afford opportunity for quantitative analysis of sentences ordering.

Several methods of sentence ordering are presented in section II. However, there is no ideal strategy to achieve coherent summaries. In this paper, we propose a method based on feature-adjacency to adjust sentence sequences, which is about to discuss the relationship between sentences in multi-document summarization task.

## II. BACKGROUND

There are two major groups in current research: chronological information [2] and cue of raw order of sentences in large corpus [3], [4]. According to these methods, most of the related work in sentence ordering are lead to two groups, chronological ordering and probatilistic ordering. Generally, the articles on newspaper usually contain descriptions of date and occurred events following the publication sequences. Chronological information could be easily achieved from these articles, while it is not ubiquitous in multi-document summary task. However, learning the natural order from large corpus could offer opportunity to analyze in general domain.

Regina Barzilay [1] presented their work using chronological information. They assumed the themes of sentences were the hints of sentences order. If two themes showed the same order in all input texts, then the order was likely to be an acceptable order of the sentences, which contained the themes. According to this, they presented the strategy using different dates of articles which are firstly published at the same time. When two themes have the same date, they are sorted according to their order of presentation in the same article.

Mirella Lapata [3] discussed an unsupervised probabilistic model of text structuring that learned ordering constraints from a large corpus. They considered the transition probability between sentences instead of a knowledge base. The model assumed that sentences were represented by a set of informative features that could be automatically extracted from the corpus without recourse to manual annotation.

They claimed that the model could be used to order the sentences obtained from a multi-document summarizer or a question answering system.

Madnani [8] presented a model containing three rules. The first is the original ordering of sentences in the summary, as written by the author of the summary. The second is a random ordering of the sentences. The third is an ordering created by applying the TSP ordering algorithm [7], which discusses the distance between any pair of adjacent sentences. They proposed TSP ordering algorithm based on two hypotheses, the initial orderings presented to the human subjects have a statistically significant impact on those they created, and the set of individual human reorderings exhibit a significant amount of variability.

Donghong Ji [5] discussed a method based on cluster-adjacent. Firstly, they clustered the sentences of source documents into $K$ clusters, $K$ is the number of summary sentences. Secondly, they analyzed the order of cluster based on feature-adjacency method. They claimed that

978-0-7695-3916-4/09 $26.00 © 2009 IEEE
DOI 10.1109/KESE.2009.14

Table I
ACCURACY OF CLASSIFICATION WITH VARIOUS $\alpha$ AND ITERATE TIMES

| $\alpha$ | Iterate Once | Iterate Twice | Iterate Thrice |
|---|---|---|---|
| 0.2 | 0.38378 | 0.25375 | 0.1225 |
| 0.4 | 0.40000 | 0.25625 | 0.135 |
| 0.5 | 0.52625 | 0.24875 | 0.13 |
| 0.6 | 0.54875 | 0.23125 | 0.1125 |

their model had solved the problem of noise elimination required by the feature-adjacency based ordering.

## III. IMPROVED ALGORITHM

Source documents in multi-document summarization task can not provide order information directly. Being relevant parts of task, source documents definitely contain the clue of order, because each of them describes some aspects of the same topic. There is some reason to believe that the sequence of sentences in source documents could be the reference standards of sentence ordering.

To learn the information of sentences sequence in source documents and predict the order of sentences in summary, we treat it as a classification task. Firstly, we train the model of classification with the position information of representative sentences in source documents. Secondly, we predict the sentence position in summary. Support vector machine (SVM) is a kind of supervised learning method for classification, and we use `libsvm` as classification tool in our model [12].

We gather the first sentence of each paragraph, and put them into training set. For a sentence of summary which is already in training set, we just simply remove it. The label $Sq_i$ is calculated as: $Sq_i = \frac{n_i}{N}$, where $n_i$ is the sequence number of the selected sentence in the source document, and $N$ is the number of all sentences in document. (e.g. document $D$ contains 30 sentences, in 3 paragraphs, and each contains 10 sentences. In this case, $N = 30$ , and 3 sentences are selected, which are $n_1 = 1$, $n_2 = 11$, $n_3 = 21$, respectively).

In nature language processing task TF-IDF (term frequency-inverse document frequency, algorithm provides an effective method to produce vectorization data, and we use TF-IDF scheme in experiment.

We use divide-and-conquer approach to find the order of summary sentences. In each step we divide training data into two group based on the label $Sq_i$ (e.g. training data is divided into two groups: $Sq_i \leq \alpha$ and $Sq_i > 1 - \alpha$, where $\alpha \in (0, 1)$), and we predict the order of summary sentence. The process will be iterated until each sentence of summary gets the position.

After the pre-process, we get the trained data, and expect the model give us good prediction. We use SVM to classify the sentence iterative.

Table I shows that the accuracy of classification decreases greatly as the iterate times increase.

Probabilistic ordering method analyzes the condition probability of given sentence sequence. In the sequence where each sentence is determined only by its previous sentence, the goal of sentence ordering is to find the sentence sequence with the biggest probability [3]. Generally, calculating the sentence adjacency based on adjacency feature of sentence pairs is the major method in sentence ordering task. Notice the attenuation of classification ordering method, we try to combine the classification and probabilistic method. Firstly, we classify the sentences into two groups. Secondly, we sort the sentences of each group by probabilistic method.

In probabilistic ordering method, condition probability $P(S_i|S_{i-1})$ (where $S_i$ is the $i$-th sentence of sequence) is calculated as:

$$P(S_i|S_{i-1}) = \prod_{(a_{(i,j)}, a_{(i-1,k)}) \in S_i \times S_{i-1}} P(a_{(i,j)}|a_{(i-1,k)}) \tag{1}$$

where $a_{(i,j)}$ is the $j$-th feature relevant to sentence $S_i$ and $a_{(i-1,k)}$ is the $k$-th feature of sentence $S_{i-1}$.

The probability $P(a_{(i,j)}|a_{(i-1,k)})$ is calculated as:

$$P(a_{(i,j)}|a_{(i-1,k)}) = \frac{f(a_{(i,j)}, a_{(i-1,k)})}{\sum\limits_{a_{(i,j)}} f(a_{(i,j)}, a_{(i-1,k)})} \tag{2}$$

where $f(a_{(i,j)}, a_{(i-1,k)})$ is the number of times, and feature $a_{(i,j)}$ is preceded by feature $a_{(i-1,k)}$ in the corpus [9].

Intuitively, the formula might contain noisy feature, and it requires higher precision. In our method, we modify formula 1 into 3, and try to avoid the problem.

$$P(S_i|S_{i-1}) = \sum_{k=1}^{n} P(a_{(i,j)}|a_{(i-1,k)}) \tag{3}$$

where $(a_{(i,j)}, a_{(i-1,k)}) \in S_i \times S_{i-1}$, and $n$ is the top most feature pairs of two adjacent sentences, which mean the biggest value of formula 2.

The work of PropBank and FrameNet [6] indicated that semantic representation of sentences could be represented by text structure set (e.g., a verb and its subject, a noun and its modifier). We treat the word and text structure as features of sentences.

## IV. EXPERIMENT

Our task is to produce the ordering of a given sentences set, and evaluation is necessary. Not like multi-document summary work, there is no acknowledged standard in ordering sentence. The general way is to compare it with the human work [10], [11]. Although the order produced by human is coherence and readable, there could be several acceptable orderings by different volunteers or the same one in different period. Barzilay [1] has already indicated that.

In the experiment, 100 summaries were extracted by human based on various topics. Each summary contains 8 sentences, and we used Kendall's $\tau$ [11] as the metric to evaluate the difference between the ordering generated by human and computer, which is defined as below:

$$\tau = 1 - \frac{2(numberofinversions)}{N(N-1)/2} \tag{4}$$

Table II
ORDERING EXAMPLES

| Examples | Criterion | $\tau$ values |
|---|---|---|
| 2 1 3 4 5 6 7 8 | 1 2 3 4 5 6 7 8 | 0.93 |
| 3 2 1 4 5 6 8 7 | 1 2 3 4 5 6 7 8 | 0.71 |

Table III
DIFFERENT EXPERIMENTAL RESULTS

| | StDev | Average | Max | Min |
|---|---|---|---|---|
| n = 1 | 0.28 | -0.11 | 0.57 | -0.93 |
| n = 3 | 0.27 | -0.17 | 0.43 | -0.79 |
| n = 4 | 0.28 | -0.18 | 0.43 | -0.79 |
| n = all | 0.32 | 0.05 | 0.79 | -1.00 |
| Baseline | 0.40 | -0.38 | 0.64 | -1.00 |

where $N$ is the number of sentences to be sorted, and `number_of_inversions` is the minimal number of interchanges of adjacent objects to transfer an ordering into another [5] Here are some examples in Table II.

The value of $\tau$ ranges from -1 to 1, where -1 denotes the worst situation that the sequence of sentences is inverse, and 1, on the contrary, denotes that two orderings are the same.

In the experiment, we choose probabilistic ordering method as the baseline [3]. Experimental results denote that in formula 3 the effect of $n$ is not significant.

As table III shows, when $n \in [1, 4]$ the performance of the method is not quite improved. However, when we choose all feature pairs of two adjacent sentences, the effect of the method is much better than others. These results do not support our hypothesis about noisy feature, while it indicates that we can not judge the noise of feature by just the value of formula 2.

In table III, we notice that the results of our method are better than that of baseline in average and the maximal value of $\tau$, but it is more discrete in $\tau$.

## V. CONCLUSION

Sentence ordering task concerns two fields. First is the clue in source documents. It is obvious that each exacted sentence in multi-document summarization task describes the part of same topic, and source documents are the direct evidence of order. Second is empirical knowledge of human being. A volunteer can give correct sequence of summary sentences without any other information. How to achieve the knowledge of two fields and combine them is what we engage in this paper. Experimental results indicate that our method has good performance, and it is proved to be effective.

The experiment also shows that how to properly set feature of sentences is still unclear. This maybe the key to improve the accuracy of sentence ordering task.

## VI. DISCUSSION

This paper proposed a method to reorder the sentences extracted from multi-document summarization task of Chinese language. The model is designed for general field in summary work which is supported by the corpus of domain-specific. In the experiment, we also wondered how to improve the model to suit for question answering system, and whether it is effective in other fields, just like information retrieving.

In future work, we will focus on modifying the method and enhancing the precision of results run by SVM, to improve the efficiency and effectness.

## ACKNOWLEDGE

The work described in this paper was supported by a grant from National Natural Science Foundation of China (Project No. 60703008).

## REFERENCES

[1] Regina Barzilay, Noemie Elhadad, and Kathleen McKeown, "*Inferring strategies for sentence ordering in multidocument news summarization*", Journalof Artificial Intelligence Research. 17, 2002, pp. 35-55

[2] Naoaki Okazaki, Yutaka Matsuo, and Mitsuru Ishizuka, "*Improving chronological sentence ordering by precedence relation*", In Proceedings of 20th International Conference on Computational Linguistics (COLING 04), 2004, pp. 750-756

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

[4] Regina Barzilay, Lillian Lee, "*Catching the drift: Probabilistic content models, with applications to generation and summarization*", In HLT-NAACL 2004: Proceedings of the Main Conference, 2004, pp. 113-120

[5] Ji Donghong, Nie Yu, "*Sentence Ordering based on Cluster Adjacency in Multi-Document Summarization*", The Third International Joint Conference on Natural Language Processing, 2008, pp. 745-750

[6] Martha Palmer, Daniel Gildea, Paul Kingsbury, "*The proposition bank: An annotated corpus ofsemantic roles*", Computational Linguistics, 2005, 31(1)

[7] J. M. Conroy, J. D. Schlesinger, D. P. O'Leary, J. Goldstein, "*Back to basics Classy 2006*", In Proceedings of DUC'06(2006)

[8] Madnani, Nitin, Rebecca Passonneau, Necip Fazil Ayan, John Conroy, Bonnie Dorr, Judith Klavans, Dianne O'Leary, Judith Schlesinger, "*Measuring variability in sentence ordering for news summarization*", In Proceedings of the 11th European Workshop on Natural Language Generation, Schloss Dagstuhl, Germany, 17-20 June 2007, pp. 81-88

[9] Sogou Labs, `http://www.sogou.com/labs/dl/t.html`

[10] Lebanon, Guy and John Lafferty, "*Combining rankings using conditional probability models on permutations*", In Proceedings of the 19th International Conference on Machine Learning, 2002

[11] Mirella Lapata, "*Automatic Evaluation of Information Ordering: Kendall's Tau*", Association for Computational Linguistics, 2002, pp. 471-484

[12] Chih-Chung Chang and Chih-Jen Lin, *LIBSVM : a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`