A Novel Method of Sentence Ordering Based on Support Vector Machine [⋆]

Gongfu Peng, Yanxiang He, Ye Tian, Yingsheng Tian, and 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 extractive 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, Support Vector Machine.

1 Introduction

In extractive multi-document summarization tasks, how to extract sentences from source document is an important and major work. But it is not enough for a fluent and readable summary. Recent research indicates that research on summary should get more attention at sentence ordering. Barzilay has offered empirical evidence that proper order of extracted sentences would greatly improve the readability of a summary (Barzilay *et al.*, 2002).

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 with different authors and in different writing styles, which means source documents could not directly provide ordering criterion in multi-document summarization task.

Obviously, sentence ordering in multi-document summarization task involves two fields, information provided by source documents and experiential knowledge of human. Neither of them can be easily got and handled, because both of them need semantic knowledge more or less. Fortunately, large raw corpus can afford opportunity for quantitative analysis of sentences ordering.

Several methods of sentence ordering in multi-document summarization are presented in section 2. However, there is no ideal strategy to achieve coherent summaries. In this paper, we proposed a method based on information of source documents and experience of human to adjust sentence sequences, which discuss the relationship between sentences in multi-document summarization tasks.

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

2 Related Work

There are two major groups in current research: chronological information (Okazaki *et al.*, 2004) and cue of raw order of sentences in large corpus (Lapata, 2003; Barzilay and Lee, 2004). Also, the methods for sentence ordering are divided into two groups: chronological ordering and probabilistic ordering. Generally, the articles on newspaper usually contain descriptions of date and 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 sequences in general domain.

Regina Barzilay (Barzilay *et al.*, 2002) presented their work using chronological information. They assumed the themes of sentences were the hints of sentences order. According to this, they presented the strategy by using the dates of different articles, which were firstly published as the order of the sentences. When two themes have the same date, they are sorted according to their order of presentation in the same article.

Mirella Lapata (Lapata, 2003) 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 (Madnani *et al.*, 2007) 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 (Conroy *et al.*, 2006), 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 (Ji and Nie, 2008) 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 their model had solved the problem of noise elimination required by the feature-adjacency based ordering.

3 Model Construction

3.1 Generating rough order

Source documents in multi-document summarization tasks could not provide order information directly. But they definitely contain the clue of order, because each of them describes some aspects of the same topic. There are some reasons to assure 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 (Chang and Lin, 2001).

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 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) (Salton *et al.*, 1983), the algorithm provided an effective method to produce vectorization data, and we use TF-IDF scheme in experiment.

The paper uses 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.

In our work, we collect 100 topics based on various fields, each topic contains 8 ˜ 12 documents. Several volunteers extract 8 sentences from relevant documents as summary for each topic, and put them in proper sequence manually.

After the pre-process, we get the trained data from 100 topics, and expect the model give us good prediction. We use SVM to classify the sentence iterative.

Table 1: Accuracy of classification with various α and iterate times

α	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

Table 1 shows that the accuracy of classification decreases greatly as the iterating times increase. Experiment denotes that classification strategy does not suit for ordering whole summary. We checked sentences of each summary carefully, and found that most sentences extracted from the beginning ($Sq_i < 35\%$) or ending ($Sq_i > 80\%$) of source documents. Some summaries even only contain sentences of the beginning, which means that all sentences were classified to one group and it decreases the average accuracy of experiment in the first iteration greatly.

In the first experiment, we notice that the strategy of classification is not good at generating the whole order. Extremely, classification strategy does not work well (e.g. all sentences were classified to one group). Alternately, classification could produce a rough order (e.g. belong to the first half of summary or latter).

3.2 Generating precise order

Barzilay indicates that different volunteers generate different orders in one sentence set of a summary, but within the multiple orderings of a set, some sentences always appear together (Barzilay *et al.*, 2001). These sentences behave like a combination, and Barzilay defines them as blocks. From the observation they found that these blocks contain the units of text dealing with the same subject. In other words, block is the group that contains related themes.

The conception ‘block’ enlightens us that the sentence order of a summary may be concerned with the relative degree of the theme. The idea of evaluating the similarity between adjacent sentences of a summary could provide an effective way to generate the order of summarized sentences.

Cosine similarity is a traditional method to measure the similarity between the pair of text. It produces static and constant scores, and can be defined as:

$$\cos(S_i, S_j) = \frac{\sum_{k=1}^{n} S_{i_k} S_{j_k}}{\sqrt{\sum_{k=1}^{n} S_{i_k}^2}\sqrt{\sum_{k=1}^{n} S_{j_k}^2}} \tag{1}$$

Where S_i and S_j is the i-th and j-th sentence of the summary, S_{i_k} is the TF-IDF value of the k-th term of the sentence S_i.

There is a problem in formula 1. The method of cosine similarity is symmetrizing, which means $\cos(S_i, S_j)$ is equal to $\cos(S_j, S_i)$. Intuitively, the measure of cosine similarity needs a parameter to indicate the order between two adjacent sentences.

Conditional entropy is a concept in the information theory. It is a measurement of the information entropy, which shows the uncertainty of a random variable.

The definition of conditional entropy is described as follows:

$$H(Y|X) = -\sum_{x \in X, y \in Y} p(x, y) \log p(y|x) \tag{2}$$

Given the value of the second random variable X, the conditional entropy quantifies the remaining entropy of the first random variable Y.

Then we define the following weight coefficient function:

$$\xi(S_i|S_{i-1}) = \frac{H(S_i|S_{i-1})}{H(S_i|S_{i-1}) + H(S_{i-1}|S_i)} \tag{3}$$

Where $H(S_i|S_{i-1})$ is the conditional entropy of S_i and S_{i-1}, and it can be written as follows:

$$\begin{aligned} &H(S_i|S_{i-1}) = \\ &- \sum_{S_{(i-1)_m} \in S_{i-1}, S_{i_n} \in S_i} p(S_{i_n}, S_{(i-1)_m}) \cdot \log p(S_{i_n}|S_{(i-1)_m}) \end{aligned} \tag{4}$$

where S_{i_n} is the n-th term of the sentence S_i, and $S_{(i-1)_m}$ is the m-th term of the sentence S_{i-1}. Then, we define the order weight function as:

$$O(S_i|S_{i-1}) = \cos(S_i, S_{i-1})\xi(S_i|S_{i-1}) \tag{5}$$

With the above method, we can estimate the weight of the order for a certain sentence sequence based on similarity and relative position:

$$\begin{aligned} O(T) &= O(S_1 \ldots S_n) \\ &= O(S_2|S_1) + O(S_3|S_2) + \ldots + O(S_n|S_{n-1}) \\ &= \sum_{i=1}^{n} O(S_i|S_{i-1}). \end{aligned} \tag{6}$$

3.3 Generating global order

In previous sections, we proposed two methods to generate the order of the summarized sentences in varying degrees of precision. Now we combine them to produce global order.

Firstly, we use SVM to classify the sentences into two groups (the first half of order and latter). Secondly, we estimate the order of each group with formula 6, and select the sequence with maximal value of $O(T)$ as the order of the group. After these steps, we connect the first group(classify into the first half of summary) and second group (classify into latter summary) as global order of summary.

In the second step, we notice that there are $N!$ orders for N sentences. A complete graph can represent them, and every sentence corresponds to a vertex in the graph, each edge $S_{i-1} \rightarrow S_i$ has a weight calculated by $O(S_i|S_{i-1})$. Obviously, it is NP-complete to find an optimal ordering in a directed weighted graph. We consulted the algorithm of Barzilay (Barzilay *et al.*, 2002) and made some necessary changes for our task.

We start with a selected vertex S_0 to find $max(O(S_x|S_0))$, where S_x is the vertex of graph and $S_x \neq S_0$. After that, we mark the highest S_x as S_1. Then we remove S_0 and all edges containing S_0 from the graph, and set S_1 as the start. The process is repeated until the graph is empty.

In the algorithm we give chance to every vertex to be the very first start. If there are N sentences in the group the algorithm will produce N sequences. For the sake of simplicity, we choose the sequence with the highest value in $O(T)$ (see formula 6) as the final order of the group.

4 Experiment

4.1 Baseline

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 (Lapata, 2003). Generally, calculating the sentence adjacency based on adjacency feature of sentence pairs is the major method in sentence ordering task. The feature is terms and text structure in the sentence.

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:

$$P(S_i|S_{i-1}) = \prod_{(a_{(i,j)},a_{(i-1,k)})\in S_i\times S_{i-1}} P(a_{(i,j)}|a_{(i-1,k)}) \tag{7}$$

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:

$$P(a_{(i,j)}|a_{(i-1,k)}) = \frac{f(a_{(i,j)},a_{(i-1,k)})}{\sum\limits_{a_{(i,j)}} f(a_{(i,j)},a_{(i-1,k)})} \tag{8}$$

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.

In the experiment, we choose probabilistic ordering method as the baseline (Lapata, 2003).

4.2 Evolution

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 (Guy and Lafferty, 2002; Lapata, 2002). 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 (Barzilay *et al.*, 2002) has already indicated that.

As mentioned earlier, 100 summaries were extracted by human based on various topics. Each summary contains 8 sentences, and we used Kendall's τ (Lapata, 2002) as the metric to evaluate the difference between the ordering generated by human and computer, which is defined as below:

$$\tau = 1 - \frac{2(number\ of\ inversions)}{N(N-1)/2} \tag{9}$$

where N is the number of sentences to be sorted, and the `number_of_inversions` is the minimal number of interchanges of adjacent objects to transfer an ordering into another (Ji and Nie, 2008) Here are some examples in Table 2.

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

Table 2: Ordering Examples

Examples	Criterion	τ value
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

4.3 Term estimate

The work of PropBank and FrameNet (Palmer *et al.*, 2005) 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). In this paper, we discuss the sentence ordering task for Chinese language. The character of Chinese is complicated, and we do not have a clear consciousness of rhetorical relations and effects of text structure work in sentence ordering.

We focus on four types of terms: *noun (n.)*, *verb (v.)*, *adjective (adj.)* and *adverb (adv.)*. Concerning the combination, the following 4 types are taken into account: *n.+v.*, *n.+adj.*, *v.+adv.* and the case all the four terms are involved.

Table 3: Value τ of different experimental results

	StDev	Average	Max	Min
n.	0.31	0.02	0.79	-0.71
v.	0.26	0.04	0.64	-0.57
adj.	0.29	0.11	0.93	-0.43
adv.	0.31	0.13	0.79	-0.57
n. + v.	0.29	0.04	0.64	-0.64
n. + adj.	0.29	0.03	0.71	-0.64
v. + adv.	0.27	0.07	0.71	-0.57
all	0.34	0.10	0.86	-0.93
Baseline	0.30	0.00	0.71	-0.86

In Table 3, the first column denotes the terms considered in experiment. For baseline, all terms are taken into account.

As Table 3 shows, the average τ of all experimental results are better than the baseline, which means that our algorithm is better than probabilistic ordering method. We also notice that for the item of standard deviation, not all results behave better, especially for considering all terms, which means that our algorithm is more discrete. To discuss the specific effect of each step, we repeat the experiment twice: firstly, just remove the weight coefficient of ξ(see Equation 3). Secondly, just skip classification in the first step.

Compared with the original algorithm, Table 4 indicates that removing the weight coefficient of ξ would decrease the value of τ, and increase the standard deviation generally.

From Table 5, we find the same trend as Table 4. Experimental results indicate that our algorithm benefit from both the weight coefficient of ξ and classification of SVM.

For the most interesting finding from these experiments, adjective and adverb perform very well. For common sense, words like *firstly*, *secondly*, *first*, *second* indicate the order of sentences in articles definitely. But it is not clear whether the adjective words and the adverb words play the key role. In sentence ordering task we need more semantic evidence to prove it.

From experimental results we conclude that sentence ordering task may provide availability of the appropriate analysis of adjective and adverb.

Table 4: Value τ of experiment without ξ

	StDev	Average	Max	Min
n.	0.30	0.00	0.71	-0.64
v.	0.33	0.01	0.93	-0.79
adj.	0.32	0.13	1.00	-0.57
adv.	0.28	0.09	0.86	-0.64
n. + v.	0.32	0.10	0.93	-0.50
n. + adj.	0.29	0.05	0.86	-0.57
v. + adv.	0.29	0.02	0.93	-0.64
all	0.32	0.08	1.00	-0.57
Baseline	0.30	0.00	0.71	-0.86

Table 5: Value τ of experiment without SVM

	StDev	Average	Max	Min
n.	0.30	0.00	0.57	-0.71
v.	0.33	0.02	0.93	-0.93
adj.	0.30	0.15	1.00	-0.57
adv.	0.27	0.09	0.86	-0.50
n. + v.	0.31	0.02	0.79	-0.79
n. + adj.	0.29	0.02	0.93	-0.57
v. + adv.	0.32	0.06	0.93	-0.93
all	0.31	0.04	1.00	-0.71
Baseline	0.30	0.00	0.71	-0.86

5 Conclusion

The aim of this paper is to provide a method for the sentence ordering in multi-document summarization task, and this method proceeds from the idea that sentence ordering task involve information in source documents and experiential knowledge of human.

The information of source documents is helpful because each exacted sentence in multi-document summarization task describes the part of same topic, and source documents are the direct evidence of the order. For empirical knowledge of human being, a volunteer can give correct sequence of summary sentences without any other information. The common sense supports our hypothesis of sentence ordering.

To implement two hypothesizes, firstly, we use SVM method to learn the information of source documents, and separate sentences of summary into two groups (the first half of order and latter), secondly, we propose a method to estimate directional relativity of different sentence sequences based on the information of raw corpus.

Experimental results indicate that our method has good performance, and it is prove that the classification and the estimation of directional relativity are both work.

The experiment also shows that adjective and adverb perform better than other terms and their combination generally. To analysis the adjective and adverb maybe the key to improve the accuracy of sentence ordering task.

6 Discussion

This paper proposed a method to reorder the sentences extracted from multi-document summarization tasks of Chinese language. The model is designed for general field in summary work which is supported by the corpus of domain-specific.

Although the experiments prove the effect of our algorithm, it still lacks the support from semantic knowledge. Semantic knowledge may be the key to discuss the coherence and readability.

In future work, we will focus on improving the method and try to import semantic knowledge for sentence ordering task to enhance the efficiency and effect.

References

Barzilay, Regina, Noemie Elhadad and Kathleen R. McKeown. 2001. Sentence Ordering in Multi-document Summarization. In *Proceedings of the 1st Human Language Technology Conference*, pp.1-7.

Barzilay, Regina, Noemie Elhadad and Kathleen McKeown. 2002. Inferring strategies for sentence ordering in multidocument news summarization. *Journal of Artificial Intelligence Research*, Volume 17, pp.35-55.

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

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

Conroy, J. M., J. D. Schlesinger, D. P. O'Leary and J. Goldstein. 2006. Back to Basics: Classy 2006. In *Proceedings of DUC'06*.

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

Ji, Donghong and Yu Nie. 2004. Sentence Ordering based on Cluster Adjacency in Multi-Document Summarization. *The Third International Joint Conference on Natural Language Processing*, pp.745-750.

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

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

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

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

Palmer, Martha, Daniel Gildea and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1).

Salton, G. and M. J. McGill, 1983. *Introduction to Modern Information Retrieval*. McGraw-Hill Book Company.