5GH Foundation

Wuhan University Published Two Duplicated Articles

Recently, two articles [1-2] by authors from Wuhan University are found to be more similar than expected.

Article [1], titled "Analysis of Sentence Ordering Based on Support Vector Machine", were published on proceeding of "2009 Pacific-Asia Conference on Knowledge Engineering and Software Engineering".

Article [2], titled "A Novel Method of Sentence Ordering Based on Support Vector Machine", were published on "Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation"

Both of article [1] and [2] were authored by a same group from Wuhan University: Gongfu Peng, Yanxiang He, Ye Tian, Yingsheng Tian, Weidong Wen (文卫东).

Although the article [2] provides some but not much additional method details and analysis on the results, most of the two articles [1-2] is overlapped:

2009 Pacific-Asia Conference on Knowledge Engineering and Software Engineering

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.

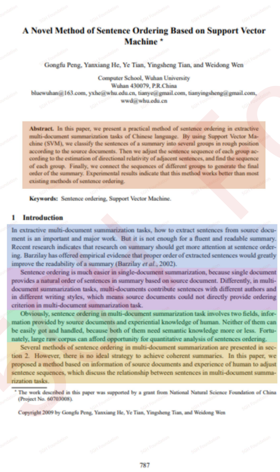
Madnani [8] presented a model containing three rules. The first is the original ordering of sentences in the summary, as written by the author of the summary. The second is a random ordering of the sentences. The third is an ordering created by applying the TSP ordering algorithm [7], which discusses 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

25

IEEE computer society



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).

Copyright 2009 by Gongfu Peng, Yanxiang He, Ye Tian, Yingsheng Tian, and Weidong Wen

787

2009 Pacific-Asia Conference on Knowledge Engineering and Software Engineering

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

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

Table I
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

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:

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

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:

$$P(a_{(i,j)}|a_{(i-1,k)}) = \frac{f(a_{(i,j)}, a_{(i-1,k)})}{\sum_{a_{(i,j)}} f(a_{(i,j)}, a_{(i-1,k)})} \quad (2)$$

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.

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

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 τ [11] as the metric to evaluate the difference between the ordering generated by human and computer, which is defined as below:

$$\tau = 1 - \frac{2(number of inversions)}{N(N-1)/2} \quad (4)$$

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

788

Table I
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

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:

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

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:

$$P(a_{(i,j)}|a_{(i-1,k)}) = \frac{f(a_{(i,j)}, a_{(i-1,k)})}{\sum_{a_{(i,j)}} f(a_{(i,j)}, a_{(i-1,k)})} \quad (2)$$

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.

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

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 τ [11] as the metric to evaluate the difference between the ordering generated by human and computer, which is defined as below:

$$\tau = 1 - \frac{2(number of inversions)}{N(N-1)/2} \quad (4)$$

26

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.

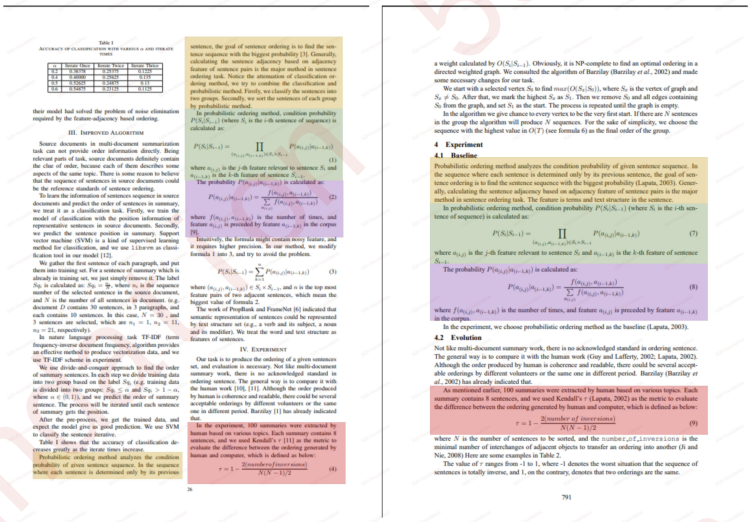
In the first experiment, we notice that the strategy of classification is not good at generating the whole order. Extremely, classification strategy does not work well (e.g. all sentences were classified to one group). Alternately, classification could produce a rough order (e.g. belong to the first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

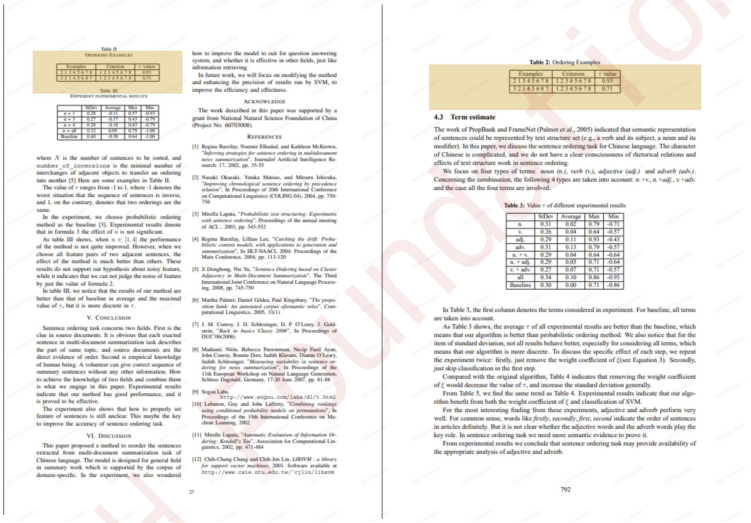
3.2 Generating precise order

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:





[1] 10.1109/KESE.2009.14

[2] PENG GF, et., al., "A Novel Method of Sentence Ordering Based on Support Vector Machine", Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation

---