

# A Study on Position Information in Document Summarization

You Ouyang   Wenjie Li   Qin Lu   Renxian Zhang

Department of Computing, the Hong Kong Polytechnic University

{csyouyang, cswjli, csluqin, csrzhang}@comp.polyu.edu.hk

## Abstract

Position information has been proved to be very effective in document summarization, especially in generic summarization. Existing approaches mostly consider the information of sentence positions in a document, based on a sentence position hypothesis that the importance of a sentence decreases with its distance from the beginning of the document. In this paper, we consider another kind of position information, i.e., the word position information, which is based on the ordinal positions of word appearances instead of sentence positions. An extractive summarization model is proposed to provide an evaluation framework for the position information. The resulting systems are evaluated on various data sets to demonstrate the effectiveness of the position information in different summarization tasks. Experimental results show that word position information is more effective and adaptive than sentence position information.

## 1 Introduction

Position information has been frequently used in document summarization. It springs from human's tendency of writing sentences of greater topic centrality at particular positions in a document. For example, in newswire documents, topic sentences are usually written earlier. A *sentence position hypothesis* is then given as: the first sentence in a document is the most important and the importance decreases as

the sentence gets further away from the beginning. Based on this sentence position hypothesis, sentence position features are defined by the ordinal position of sentences. These position features have been proved to be very effective in generic document summarization. In more recent summarization tasks, such as query-focused and update summarization tasks, position features are also widely used.

Although in these tasks position features may be used in different ways, they are all based on the sentence position hypothesis. So we regard them as providing the sentence position information. In this paper, we study a new kind of position information, i.e., the word position information. The motivation of word position information comes from the idea of assigning different importance to multiple appearances of one word in a document.

As to many language models such as the bag-of-words model, it is well acknowledged that a word which appears more frequently is usually more important. If we take a closer look at all the appearances of one word, we can view this as a process that the different appearances of the same word raise the importance of each other. Now let's also take the order of the appearances into account. When reading a document, we can view it as a word token stream from the first token to the last. When a new token is read, we attach more importance to previous tokens that have the same lemma because they are just repeated by the new token. Inspired by this, we postulate a *word position hypothesis* here: for all the appearances of a fixed word, the importance of each appearance depends on all its following appearances. Therefore, the first appearance of a word is the most important and the importance decreases with the ordinal

positions of the appearances. Then, a novel kind of position features can be defined for the word appearances based on their ordinal positions. We believe that these word position features have some advantages when compared to traditional sentence position features. According to the sentence position hypothesis, sentence position features generally prefer earlier sentences in a document. As to the word position features that attempt to differentiate word appearances instead of sentences, a sentence which is not the first one in the document may still not be penalized as long as its words do not appear in previous sentences. Therefore, word position features are able to discover topic sentences in deep positions of the document. On the other hand, the assertion that the first sentence is always the most important is not true in actual data. It depends on the writing style indeed. For example, some authors may like to write some background sentences before topic sentences. In conclusion, we can expect word position features to be more adaptive to documents with different structures.

In the study of this paper, we define several word position features based on the ordinal positions of word appearances. We also develop a word-based summarization system to evaluate the effectiveness of the proposed word position features on a series of summarization data sets. The main contributions of our work are:

- (1) representation of word position information, which is a new kind of position information in document summarization area.
- (2) empirical results on various data sets that demonstrate the impact of position information in different summarization tasks.

## 2 Related Work

The use of position information in document summarization has a long history. In the seminal work by (Luhn, 1958), position information was already considered as a good indicator of significant sentences. In (Edmundson, 1969), a location method was proposed that assigns positive weights to the sentences to their ordinal positions in the document. Position information has since been adopted by many successful summarization systems, usually in the form of sentence position features. For example, Radev et al. (2004) developed a feature-based system

MEAD based on word frequencies and sentence positions. The position feature was defined as a descending function of the sentence position. The MEAD system performed very well in the generic multi-document summarization task of the DUC 2004 competition. Later, position information is also applied to more summarization tasks. For example, in query-focused task, sentence position features are widely used in learning-based summarization systems as a component feature for calculating the composite sentence score (Ouyang et al, 2007; Toutanova et al, 2007). However, the effect of position features alone was not studied in these works.

There were also studies aimed at analyzing and explaining the effectiveness of position information. Lin and Hovy (1997) provided an empirical validation on the sentence position hypothesis. For each position, the *sentence position yield* was defined as the average value of the significance of the sentences with the fixed position. It was observed that the average significance at earlier positions was indeed larger. Nenkova (2005) did a conclusive overview on the DUC 2001-2004 evaluation results. It was reported that position information is very effective in generic summarization. In generic single-document summarization, a lead-based baseline that simply takes the leading sentences as the summary can outperform most submitted summarization system in DUC 2001 and 2002. As in multi-document summarization, the position-based baseline system is competitive in generating short summaries but not in longer summaries. Schilder and Kondadadi (2008) analyzed the effectiveness of the features that are used in their learning-based sentence scoring model for query-focused summarization. By comparing the ROUGE-2 results of each individual feature, it was reported that position-based features are less effective than frequency-based features. In (Gillick et al., 2009), the effect of position information in the update summarization task was studied. By using ROUGE to measure the density of valuable words at each sentence position, it was observed that the first sentence of newswire document was especially important for composing update summaries. They defined a binary sentence position feature based on the

observation and the feature did improve the performance on the update summarization data.

### 3 Methodology

In the section, we first describe the word-based summarization model. The word position features are then defined and incorporated into the summarization model.

#### 3.1 Basic Summarization Model

To test the effectiveness of position information in document summarization, we first propose a word-based summarization model for applying the position information. The system follows a typical extractive style that constructs the target summary by selecting the most salient sentences.

Under the bag-of-words model, the probability of a word  $w$  in a document set  $D$  can be scaled by its frequency, i.e.,  $p(w)=freq(w)/|D|$ , where  $freq(w)$  indicates the frequency of  $w$  in  $D$  and  $|D|$  indicates the total number of words in  $D$ . The probability of a sentence  $s=\{w_1, \dots, w_N\}$  is then calculated as the product of the word probabilities, i.e.,  $p(s)=\prod_i p(w_i)$ . Moreover, the probability of a summary consisting a set of sentences, denoted as  $S=\{s_1, \dots, s_M\}$ , can be calculated by the product of the sentence probabilities, i.e.,  $p(S)=\prod_j p(s_j)$ . To obtain the optimum summary, an intuitive idea is to select the sentences to maximize the overall summary probability  $p(S)$ , equivalent to maximizing  $\log(p(S)) = \sum_j \sum_i \log(p(w_{ji})) = \sum_j \sum_i (\log freq(w_{ji}) - \log |D|) = \sum_j \sum_i \log freq(w_{ji}) - |S| \cdot \log |D|$ ,

where  $w_{ji}$  indicates the  $i$ th word in  $s_j$  and  $|S|$  indicates the total number of words in  $S$ . As to practical summarization tasks, a maximum summary length is usually postulated. So here we just assume that the length of the summary is fixed. Then, the above optimization target is equivalent to maximizing  $\sum_j \sum_i \log freq(w_{ji})$ . From the view of information theory, the sum can also be interpreted as a simple measure on the total information amount of the summary. In this interpretation, the information of a single word  $w_{ji}$  is measured by  $\log freq(w_{ji})$  and the summary information is the sum of the word information. So the optimization target can also be interpreted as including the most informative words to form the most informative summary given the length limit.

In extractive summarization, summaries are composed by sentence selection. As to the above optimization target, the sentence scoring function for ranking the sentences should be calculated as the average word information, i.e.,  $score(s) = \sum_i \log freq(w_i) / |s|$ .

After ranking the sentences by their ranking scores, we can select the sentences into the summary by the descending order of their score until the length limit is reached. By this process, the summary with the largest  $p(S)$  can be composed.

#### 3.2 Word Position Features

With the above model, word position features are defined to represent the word position information and are then incorporated into the model. According to the motivation, the features are defined by the ordinal positions of word appearances, based on the position hypothesis that earlier appearances of a word are more informative. Formally, for the  $i$ th appearance among the total  $n$  appearances of a word  $w$ , four position features are defined based on  $i$  and  $n$  using different formulas as described below.

**(1) Direct proportion (DP)** With the word position hypothesis, an intuitive idea is to regard the information degree of the first appearance as 1 and the last one as  $1/n$ , and then let the degree decrease linearly to the position  $i$ . So we can obtain the first position feature defined by the direct proportion function, i.e.,  $f(i)=(n-i+1)/n$ .

**(2) Inverse proportion (IP).** Besides the linear function, other functions can also be used to characterize the relationship between the position and the importance. The second position feature adopts another widely-used function, the inversed proportion function, i.e.,  $f(i)=1/i$ . This measure is similar to the above one, but the information degree decreases by the inverse proportional function. Therefore, the degree decreases more quickly at smaller positions, which implies a stronger preference for leading sentences.

**(3) Geometric sequence (GS).** For the third feature, we make an assumption that the degree of every appearance is the sum of the degree of all the following appearances, i.e.,  $f(i) = f(i+1) + f(i+2) + \dots + f(n)$ . It can be easily derived that the sequence also satisfies  $f(i) = 2 \cdot f(i-1)$ . That is, the information degree of each new appearance is

halved. Then the feature value of the  $i$ th appearance can be calculated as  $f(i) = (1/2)^{i-1}$ .

**(4) Binary function (BF).** The final feature is a binary position feature that regards the first appearance as much more informative than the all the other appearances, i.e.,  $f(i)=1$ , if  $i=1$ ;  $\lambda$  else, where  $\lambda$  is a small positive real number.

### 3.3 Incorporating the Position Features

To incorporate the position features into the word-based summarization model, we use them to adjust the importance of the word appearance. For the  $i$ th appearance of a word  $w$ , its original importance is multiplied by the position feature value, i.e.,  $\log freq(w) \cdot pos(w, i)$ , where  $pos(w, i)$  is calculated by one of the four position features introduced above. By this, the position feature is also incorporated into the sentence scores, i.e.,  $score'(s) = \sum_i [\log freq(w_i) \cdot pos(w_i)] / |s|$

### 3.4 Sentence Position Features

In our study, another type of position features, which model sentence position information, is defined for comparison with the word position features. The sentence position features are also defined by the above four formulas. However, for each appearance, the definition of  $i$  and  $n$  in the formulas are changed to the ordinal position of the sentence that contains this appearance and the total number of sentences in the document respectively. In fact, the effects of the features defined in this way are equivalent to traditional sentence position features. Since  $i$  and  $n$  are now defined by sentence positions, the feature values of the word tokens in the same sentence  $s$  are all equal. Denote it by  $pos(s)$ , and the sentence score with the position feature can be written as

$$score'(s) = (\sum_{w \text{ in } s} \log freq(w) \cdot pos(s)) / |s| \\ = pos(s) \cdot (\sum \log_{w \text{ in } s} freq(w) / |s|),$$

which can just be viewed as the product of the original score and a sentence position feature.

### 3.5 Discussion

By using the four functions to measure word or sentence position information, we can generate a total of eight position features. Among the four functions, the importance drops fastest under the binary function and the order is **BF** > **GS** > **IP** > **DP**. Therefore, the features based on the binary function are the most biased to the

leading sentences in the document and the features based on the direct proportion function are the least. On the other hand, as mentioned in the introduction, sentence-based features have larger preferences for leading sentences than word-based position features.

An example is given below to illustrate the difference between word and sentence position features. This is a document from DUC 2001.

1. GENERAL ACCIDENT, the leading British insurer, said yesterday that insurance claims arising from Hurricane Andrew could 'cost it as much as Dollars 40m.'
2. Lord Airlie, the chairman who was addressing an extraordinary shareholders' meeting, said: 'On the basis of emerging information, General Accident advise that the losses to their US operations arising from Hurricane Andrew, which struck Florida and Louisiana, might in total reach the level at which external catastrophe reinsurance covers would become exposed'.
3. What this means is that GA is able to pass on its losses to external reinsurers once a certain claims **threshold** has been breached.
4. It believes this **threshold** may be breached in respect of Hurricane Andrew claims.
5. However, if this happens, it would suffer a post-tax loss of Dollars 40m (Pounds 20m).
6. Mr Nelson Robertson, GA's chief general manager, explained later that the company has a 1/2 per cent share of the Florida market.
7. It has a branch in Orlando.
8. The company's loss adjusters are in the area trying to **estimate** the losses.
9. Their guess is that losses to be faced by all insurers may total more than Dollars 8bn.
10. Not all damaged property in the area is insured and there have been **estimates** that the storm caused more than Dollars 20bn of damage.
11. However, other insurers have **estimated** that losses could be as low as Dollars 1bn in total.
- 12 Mr Robertson said: 'No one knows at this time what the exact loss is'.

For the word "threshold" which appears twice in the document, its original importance is  $\log(2)$ , for the appearance of "threshold" in the 4<sup>th</sup> sentence, the modified score based on word position feature with the direct proportion function is  $1/2 \cdot \log(2)$ . In contrast, the score based on sentence position feature with the

same function is  $9/12 \cdot \log(2)$ , which is larger. For the appearance of the word “estimate” in the 8<sup>th</sup> sentence, its original importance is  $\log(3)$  (the three boldfaced tokens are regarded as one word with stemming). The word-based and sentence-based scores are  $\log(3)$  and  $5/12 \cdot \log(3)$  respectively. So its importance is larger under word position feature. Therefore, the system with word position features may prefer the 8<sup>th</sup> sentence that is in deeper positions but the system with sentence position feature may prefer the 4<sup>th</sup> sentence. As for this document, the top 5 sentences selected by sentence position feature are {1, 4, 3, 5, 2} and the those selected by the word position features are {1, 8, 3, 6, 9}. This clearly demonstrates the difference between the position features.

## 4 Experimental Results

### 4.1 Experiment Settings

We conduct the experiments on the data sets from the Document Understanding Conference (DUC) run by NIST. The DUC competition started at year 2001 and has successfully evaluated various summarization tasks up to now. In the experiments, we evaluate the effectiveness of position information on several DUC data sets that involve various summarization tasks. One of the evaluation criteria used in DUC, the automatic summarization evaluation package ROUGE, is used to evaluate the effectiveness of the proposed word position features in the context of document summarization<sup>1</sup>. The recall scores of ROUGE-1 and ROUGE-2, which are based on unigram and bigram matching between system summaries and reference summaries, are adopted as the evaluation criteria.

In the data sets used in the experiments, the original documents are all pre-processed by sentence segmentation, stop-word removal and word stemming. Based on the word-based summarization model, a total of nine systems are evaluated in the experiments, including the system with the original ranking model (denoted as **None**), four systems with each word position feature (denoted as **WP**) and four systems with each sentence position feature (denoted as **SP**).

<sup>1</sup> We run ROUGE-1.5.5 with the parameters “-x -m -n 2 -2 4 -u -c 95 -p 0.5 -t 0”

For reference, the average ROUGE scores of all the human summarizers and all the submitted systems from the official results of NIST are also given (denoted as **Hum** and **NIST** respectively).

### 4.2 Redundancy Removal

To reduce the redundancy in the generated summaries, we use an approach similar to the maximum marginal relevance (MMR) approach in the sentence selection process (Carbonell and Goldstein, 1998). In each round of the sentence selection, the candidate sentence is compared against the already-selected sentences. The sentence is added to the summary only if it is not significantly similar to any already-selected sentence, which is judged by the condition that the cosine similarity between the two sentences is less than 0.7.

### 4.3 Generic Summarization

In the first experiment, we use the DUC 2001 data set for generic single-document summarization and the DUC 2004 data set for generic multi-document summarization. The DUC 2001 data set contains 303 document-summary pairs; the DUC 2004 data set contains 45 document sets, with each set consisting of 10 documents. A summary is required for each document set. Here we need to adjust the ranking model for the multi-document task, i.e., the importance of a word is calculated as its total frequency in the whole document set instead of a single document. For both tasks, the summary length limit is 100 words.

Table 1 and 2 below provide the average ROUGE-1 and ROUGE-2 scores (denoted as **R-1** and **R-2**) of all the systems. Moreover, we used paired two sample t-test to calculate the significance of the differences between a pair of word and sentence position features. The bolded score in the tables indicates that that score is significantly better than the corresponding paired one. For example, in Table 1, the bolded **R-1** score of system **WP DP** means that it is significantly better than the **R-1** score of system **SP DP**. Besides the ROUGE scores, two statistics, the number of “first sentences<sup>2</sup>” among the selected sentences (**FS-N**) and the

<sup>2</sup> A “first sentence” is the sentence at the first position of a document.

average position of the selected sentences (**A-SP**), are also reported in the tables for analysis.

System	R-1	R-2	FS-N	A-SP
<b>WP DP</b>	<b>0.4473</b>	<b>0.1942</b>	301	4.00
<b>SP DP</b>	0.4396	0.1844	300	3.69
<b>WP IP</b>	<b>0.4543</b>	<b>0.2023</b>	290	4.30
<b>SP IP</b>	0.4502	0.1964	303	3.08
<b>WP GS</b>	<b>0.4544</b>	<b>0.2041</b>	278	4.50
<b>SP GS</b>	0.4509	0.1974	303	2.93
<b>WP BF</b>	<b>0.4544</b>	<b>0.2036</b>	253	5.57
<b>SP BF</b>	0.4239	0.1668	303	9.64
<b>None</b>	0.4193	0.1626	265	10.06
<b>NIST</b>	0.4445	0.1865	-	-
<b>Hum</b>	0.4568	0.1740	-	-

Table 1. Results on the DUC 2001 data set

System	R-1	R-2	FS-N	A-SP
<b>WP DP</b>	0.3728	0.0911	89	4.16
<b>SP DP</b>	0.3724	0.0908	112	2.68
<b>WP IP</b>	<b>0.3756</b>	0.0912	108	3.77
<b>SP IP</b>	0.3690	0.0905	201	1.01
<b>WP GS</b>	<b>0.3751</b>	0.0916	110	3.67
<b>SP GS</b>	0.3690	0.0905	201	1.01
<b>WP BF</b>	<b>0.3740</b>	<b>0.0926</b>	127	3.14
<b>SP BF</b>	0.3685	0.0903	203	1
<b>None</b>	0.3550	0.0745	36	10.98
<b>NIST</b>	0.3340	0.0686	-	-
<b>Hum</b>	0.4002	0.0962	-	-

Table 2. Results on the DUC 2004 data set

From Table 1 and Table 2, it is observed that position information is indeed very effective in generic summarization so that all the systems with position features performed better than the system **None** which does not use any position information. Moreover, it is also clear that the proposed word position features consistently outperform the corresponding sentence position features. Though the gaps between the ROUGE scores are not large, the t-tests proved that word position features are significantly better on the DUC 2001 data set. On the other hand, the advantages of word position features over sentence position features are less significant on the DUC 2004 data set. One reason may be that the multiple documents have provided more candidate sentences for composing the summary. Thus it is possible to generate a good summary only from the leading sentences in the

documents. According to Table 2, the average-sentence-position of system **SP BF** is 1, which means that all the selected sentences are “first sentences”. Even under this extreme condition, the performance is not much worse.

The two statistics also show the different preferences of the features. Compared to word position features, sentence position features are likely to select more “first sentences” and also have smaller average-sentence-positions. The abnormally large average-sentence-position of **SP BF** in DUC 2001 is because it does not differentiate all the other sentences except the first one. The corresponding word-position-based system **WP BF** can differentiate the sentences since it is based on word positions, so its average-sentence-position is not that large.

#### 4.4 Query-focused Summarization

Since year 2005, DUC has adopted query-focused multi-document summarization tasks that require creating a summary from a set of documents to a given query. This task has been specified as the main evaluation task over three years (2005-2007). The data set of each year contains about 50 DUC topics, with each topic including 25-50 documents and a query. In this experiment, we adjust the calculation of the word importance again for the query-focused issue. It is changed to the total number of the appearances that fall into the sentences with at least one word in the query. Formally, given the query which is viewed as a set of words  $Q = \{w_1, \dots, w_T\}$ , a sentence set  $S_Q$  is defined as the set of sentences that contain at least one  $w_i$  in  $Q$ . Then the importance of a word  $w$  is calculated by its frequency in  $S_Q$ . For the query-focused task, the summary length limit is 250 words.

Table 3 below provides the average ROUGE-1 and ROUGE-2 scores of all the systems on the DUC 2005-2007 data sets. The boldfaced terms in the tables indicate the best results in each column. According to the results, on query-focused summarization, position information seems to be not as effective as on generic summarization. The systems with position features can not outperform the system **None**. In fact, this is reasonable due to the requirement specified by the pre-defined query. Given the query, the content of interest may be in any

position of the document and thus the position information becomes less meaningful.

On the other hand, we find that though the systems with word position features cannot outperform the system **None**, it does significantly outperform the systems with sentence position features. This is also due to the role of the query. Since it may refer to the specified content in any position of the

documents, sentence position features are more likely to fail in discovering the desired sentences since they always prefer leading sentences. In contrast, word position features are less sensitive to this problem and thus perform better. Similarly, we can see that the direct proportion (**DP**), which has the least bias for leading sentences, has the best performance among the four functions.

System	2005		2006		2007	
	R-1	R-2	R-1	R-2	R-1	R-2
<b>WP DP</b>	<b>0.3791</b>	<b>0.0805</b>	0.3909	0.0917	0.4158	0.1135
<b>SP DP</b>	0.3727	0.0776	0.3832	0.0869	0.4118	0.1103
<b>WP IP</b>	0.3772	0.0791	0.3830	0.0886	0.4106	0.1121
<b>SP IP</b>	0.3618	0.0715	0.3590	0.0739	0.3909	0.1027
<b>WP GS</b>	0.3767	0.0794	0.3836	0.0879	0.4109	0.1119
<b>SP GS</b>	0.3616	0.0716	0.3590	0.0739	0.3909	0.1027
<b>WP BF</b>	0.3740	0.0741	0.3642	0.0796	0.3962	0.1037
<b>SP BF</b>	0.3647	0.0686	0.3547	0.0742	0.3852	0.1013
<b>NONE</b>	0.3788	0.0791	<b>0.3936</b>	<b>0.0924</b>	<b>0.4193</b>	<b>0.1140</b>
<b>NIST</b>	0.3353	0.0592	0.3707	0.0741	0.0962	0.3978
<b>Hum</b>	0.4392	0.1022	0.4532	0.1101	0.4757	0.1402

Table 3. Results on the DUC 2005 - 2007 data sets

System	2008 A		2008 B		2009 A		2009 B	
	R-1	R-2	R-1	R-2	R-1	R-2	R-1	R-2
<b>WP DP</b>	0.3687	0.0978	<b>0.3758</b>	0.1036	0.3759	0.1015	0.3693	0.0922
<b>SP DP</b>	0.3687	0.0971	0.3723	0.1011	0.3763	0.1031	0.3704	0.0946
<b>WP IP</b>	<b>0.3709</b>	<b>0.1014</b>	0.3741	<b>0.1058</b>	0.3758	0.1030	0.3723	0.0906
<b>SP IP</b>	0.3619	0.0975	0.3723	0.1037	0.3693	0.0994	0.3690	<b>0.0956</b>
<b>WP GS</b>	0.3705	0.1004	0.3732	0.1048	<b>0.3770</b>	0.1051	<b>0.3731</b>	0.0917
<b>SP GS</b>	0.3625	0.0975	0.3723	0.1037	0.3693	0.0994	0.3690	<b>0.0956</b>
<b>WP BF</b>	0.3661	0.0975	0.3678	0.0992	0.3720	<b>0.1069</b>	0.3650	0.0936
<b>SP BF</b>	0.3658	0.0965	0.3674	0.0980	0.3683	0.1043	0.3654	0.0945
<b>NONE</b>	0.3697	0.0978	0.3656	0.0915	0.3653	0.0934	0.3595	0.0834
<b>NIST</b>	0.3389	0.0799	0.3192	0.0676	0.3468	0.0890	0.3315	0.0761
<b>Hum</b>	0.4105	0.1156	0.3948	0.1134	0.4235	0.1249	0.3901	0.1059

Table 4. Results on the TAC 2008 - 2009 data sets

#### 4.5 Update Summarization

Since year 2008, the DUC summarization track has become a part of the Text Analysis Conference (TAC). In the update summarization task, each document set is divided into two ordered sets A and B. The summarization target on set A is the same as the query-focused task in DUC 2005-2007. As to the set B, the target is to write an update summary of the documents in set B, under the assumption that the reader has

already read the documents in set A. The data set of each year contains about 50 topics, and each topic includes 10 documents for set A, 10 documents for set B and an additional query. For set A, we follow exactly the same method used in section 4.4; for set B, we make an additional novelty check for the sentences in B with the MMR approach. Each candidate sentence for set B is now compared to both the selected sentences in set B and in set A to

ensure its novelty. In the update task, the summary length limit is 100 words.

Table 4 above provides the average ROUGE-1 and ROUGE-2 scores of all the systems on the TAC 2008-2009 data sets. The results on set A and set B are shown individually. For the task on set A which is almost the same as the DUC 2005-2007 tasks, the results are also very similar. A small difference is that the systems with position features perform slightly better than the system **None** on these two data sets. Also, the difference between word position features and sentence position features becomes smaller. One reason may be that the shorter summary length increases the chance of generating good summaries only from the leading sentences. This is somewhat similar to the results reported in (Nenkova, 2005) that position information is more effective for short summaries.

For the update set B, the results show that position information is indeed very effective. In the results, all the systems with position features significantly outperform the system **None**. We attribute the reason to the fact that we are more concerned with novel information when summarizing update set B. Therefore, the effect of the query is less on set B, which means that the effect of position information may be more pronounced in contrast. On the other hand, when comparing the position features, we can see that though the difference of the position features is quite small, word position features are still better in most cases.

#### 4.6 Discussion

Based on the experiments, we briefly conclude the effectiveness of position information in document summarization. In different tasks, the effectiveness varies indeed. It depends on whether the given task has a preference for the sentences at particular positions. Generally, in generic summarization, the position hypothesis works well and thus the ordinal position information is effective. In this case, those position features that are more distinctive, such as **GS** and **BF**, can achieve better performances. In contrast, in the query-focused task that relates to specified content in the documents, ordinal position information is not so useful. Therefore, the more distinctive a position feature is, the

worse performance it leads to. However, in the update summarization task that also involves queries, position information becomes effective again since the role of the query is less dominant on the update document set.

On the other hand, by comparing the sentence position features and word position features on all the data sets, we can draw an overall conclusion that word position features are consistently more appreciated. For both generic tasks in which position information is effective and query-focused tasks in which it is not so effective, word position features show their advantages over sentence position features. This is because of the looser position hypothesis postulated by them. By avoiding arbitrarily regarding the leading sentences as more important, they are more adaptive to different tasks and data sets.

## 5 Conclusion and Future Work

In this paper, we proposed a novel kind of word position features which consider the positions of word appearances instead of sentence positions. The word position features were compared to sentence position features under the proposed sentence ranking model. From the results on a series of DUC data sets, we drew the conclusion that the word position features are more effective and adaptive than traditional sentence position features. Moreover, we also discussed the effectiveness of position information in different summarization tasks.

In our future work, we'd like to conduct more detailed analysis on position information. Besides the ordinal positions, more kinds of position information can be considered to better model the document structures. Moreover, since position hypothesis is not always correct in all documents, we'd also like to consider a pre-classification method, aiming at identifying the documents for which position information is more suitable.

**Acknowledgement** The work described in this paper was supported by Hong Kong RGC Projects (PolyU5217/07E). We are grateful to professor Chu-Ren Huang for his insightful suggestions and discussions with us.



## References

- Edmundson, H. P.. 1969. *New methods in automatic Extracting*. Journal of the ACM, volume 16, issue 2, pp 264-285.
- Gillick, D., Favre, B., Hakkani-Tur, D., Bohnet, B., Liu, Y., Xie, S.. 2009. *The ICSI/UTD Summarization System at TAC 2009*. Proceedings of Text Analysis Conference 2009.
- Jaime G. Carbonell and Jade Goldstein. 1998. *The use of MMR, diversity-based reranking for reordering documents and producing summaries*. Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, pp 335-336.
- Lin, C. and Hovy, E.. 1997. *Identifying Topics by Position*. Proceedings of the fifth conference on Applied natural language processing 1997, pp 283-290.
- Luhn, H. P.. 1958. *The automatic creation of literature abstracts*. IBM J. Res. Develop. 2, 2, pp 159-165.
- Nenkova. 2005. *Automatic text summarization of newswire: lessons learned from the document understanding conference*. Proceedings of the 20th National Conference on Artificial Intelligence, pp 1436-1441.
- Ouyang, Y., Li, S., Li, W.. 2007. *Developing learning strategies for topic-based summarization*. Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, pp 79-86.
- Radev, D., Jing, H., Sty's, M. and Tam, D.. 2004. *Centroid-based summarization of multiple documents*. Information Processing and Management, volume 40, pp 919-938.
- Schilder, F., Kondadadi, R.. 2008. *FastSum: fast and accurate query-based multi-document summarization*. Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies, short paper session, pp 205-208.
- Toutanova, K. et al. 2007. *The PYPHY summarization system: Microsoft research at DUC 2007*. Proceedings of Document Understanding Conference 2007.