Tsinghua University at TAC 2009: Summarizing Multi-documents by Information Distance

Chong Long, Minlie Huang, Xiaoyan Zhu
State Key Laboratory of Intelligent Technology and Systems, Tsinghua National Laboratory for Information Science and Technology, Department of Computer Science and Technology, Tsinghua University, China

Abstract

This paper presents our extractive summarization systems at the update summarization track of TAC 2009. This system is based on our newly developed document summarization framework under the theory of conditional information distance among many objects. The best summary is defined in this paper to be the one which has the minimum information distance to the entire document set. The best update summary has the minimum conditional information distance to a document cluster given that a prior document cluster has already been read. Experiments on the TAC dataset have proved that our method has got a good performance in many categories.

1 Introduction

We participated in the update summarization track of TAC 2009. The update summarization task is to write a short (not more than 100 words) summary of a set of newswire articles, under the assumption that the user has already read a given set of earlier articles. The summaries will be evaluated for readability and content (based on Columbia University's Pyramid Method) [1]. We firstly proposed information distance based approach in TAC 2008. This year we have developed a framework in which multi-document summarization can be modeled by the information distance theory. The best summary is defined as having the minimal information distance (or conditional information distance) to the entire document set (if a prior document set is given).

The paper is organized as follows. Section 2 introduces our method in TAC 2008.

Our newly developed theory is described in Section 3.1. Section 3 presents the summarization method under the new theory and experiments in Section 4 emphasize the advantages of our work. Conclusions and future work are outlined in Section 5.

2 Overview of Our Method in TAC 2008

In TAC 2008, we firstly proposed to use information distance to solve the summarization problem [2]. Fix a universal Turing machine U. The Kolmogorov complexity [3] of a binary string x conditioned to another binary string y, $K_U(x|y)$, is the length of the shortest (prefix-free) program for U that outputs x with input y. It can be shown that for a different universal Turing machine U', for all x, y

$$K_U(x|y) = K_{U'}(x|y) + C,$$

where the constant C depends only on U'. Thus $K_U(x|y)$ can be simply written as $K(x|y)$. We write $K(x|\epsilon)$, where ϵ is the empty string, as $K(x)$. It has also been defined in [4] that the energy to convert between x and y to be the smallest number of bits needed to convert x to y and vice versa. That is, with respect to a universal Turing machine U, the cost of conversion between x and y is:

$$E(x,y) = \min\{|p| : U(x,p) = y, \\ U(y,p) = x\} \quad (1)$$

The following theorem has been proved in [4]:

Theorem 1 $E(x,y) = \max\{K(x|y), K(y|x)\}$.

Thus, the max distance was defined in [4]:

$$D_{\max}(x,y) = \max\{K(x|y), K(y|x)\}. \quad (2)$$

TAC update summarization task is to write a short summary S of n newswire articles $B_1, B_2, \ldots, B_n$, under the assumption that the user has already read a given set of earlier m articles $A_1, A_2, \ldots, A_m$. In TAC 2008, we use the following criteria to select the best summary S:

$$\min D_{\max}(S, B_1 B_2 \ldots B_m | A_1 A_2 \ldots A_m), \\ |S| \leq \theta \quad (3)$$

S is selected from sentences of articles $A_1, A_2, \ldots, A_m$. However, it is more or less intuitive method.

This year we have set up a relatively complete information distance summarization framework. Our new summarization model in TAC 2009 is based on our newly developed theory instead of an empirical formula(Equation 3) in TAC 2008. Next we will introduce this new framework.

3 New Summarization Framework

Our new framework is based on our newly developed theory of conditional information distance among many objects. In this section we will firstly introduce our newly developed theory and then our summarization model based on the new theory.

3.1 New Theory

In [5], the authors generalize the theory of information distance to more than two objects. Similar to Equation 1, given strings $x_1, \ldots, x_n$, they define the minimal amount of thermodynamic energy needed to convert any x_i to any x_j as:

$$E_m(x_1, \ldots, x_n) = \min\{|p| : U(x_i, p, j) = x_j \text{ for all } i, j\} \quad (4)$$

Then it is proved in [5] that:

Theorem 2 *Modulo to an $O(\log n)$ additive factor,*

$$\min_i K(x_1 \ldots x_n | x_i) \leq E_m(x_1, \ldots, x_n) \leq \min_i \sum_{k \neq i} D_{\max}(x_i, x_k) \quad (5)$$

In update summarization, the summary should contain new information which former documents have not mentioned, so we extended Equation 5 in paper [6] to be:

Theorem 3 *Modulo to an $O(\log n)$ additive factor,*

$$\min_i K(x_1 \ldots x_n | x_i, c) \leq E_m(x_1, \ldots, x_n | c) \leq \min_i \sum_{k \neq i} D_{\max}(x_i, x_k | c) \quad (6)$$

where c is the conditional sequence that is given for free to compute from sequence x to y and from y to x.

Given n objects and a conditional sequence c, the left-hand side of Equation 6 may be interpreted as the most comprehensive object that contains the most information about all of the others. The right-hand side of the equation may be interpreted as the most typical object that is similar to all of the others.

3.2 Modeling

We have developed the theory of conditional information distance among many objects. In this subsection, a new summarization model be built based on our new theory.

3.2.1 Modeling Traditional Summarization

The task of traditional multi-document summarization can be described as follows: given n documents $B = \{B_1, B_2, \ldots, B_n\}$, the task requires the system to generate a summary S of B. According to our theory, the conditional information distance among $B_1, B_2, \ldots, B_n$ is $E_m(B)$.

However, it is very difficult to compute E_m. Moreover, E_m itself does not tell us how to generate a summary. Equation 5 has provided us a feasible way to approximate E_m: the most comprehensive object and the most typical one are the left and right of Equation 6, respectively. The most comprehensive object is long enough to cover as much information in B as possible, while the most typical object is a concise one that expresses the most common idea shared by those objects. Since we aim to produce a short summary to represent the general information, the right-hand side of Equation 5 should be used. The most typical document is the B_j such that

$$\min_j \sum_{i \neq j} D_{\max}(B_i, B_j)$$

However, B_j is far from enough to be a good summary. A good method should be able to select the information from B_1 to B_n to form a best S. We view this S as a document in this set. Since S is a short summary, it does not contain extra information outside B. The best traditional summary S_{trad} should satisfy the constraint as:

$$S_{trad} = \arg\min_S \sum_i D_{\max}(B_i, S) \qquad (7)$$

In most applications, the length of S is confined by $|S| \leq \theta$ (θ is a constant integer) or $|S| \leq \alpha \sum_i |B_i|$ (α is a constant real number between 0 and 1).

3.2.2 Modeling Update Summarization

Given a set of earlier m articles $A = \{A_1, A_2, \ldots, A_m\}$, the update summarization task is to summarize new contents presented by a document set $B = \{B_1, B_2, \ldots, B_m\}$. This earlier article set A can be viewed as a precondition. Thus this task can be well modeled by the conditional version of information distance. The best summary S_{best} should satisfy the constraint as follows:

$$S_{best} = \arg\min_S \sum_i D_{\max}(B_i, S|A) \qquad (8)$$

If $m = 0$ ($A = \phi$), it will be a traditional multi-document summarization problem. If $m > 0$ ($A \neq \phi$), it will be a multi-document update summarization problem. Therefore, the traditional summarization can be viewed as a special case of formula 8.

According to [7], from Equation 8 we can get:

$$D_{\max}(B_i, S|A) = D_{\max}(B_i^A, S)$$

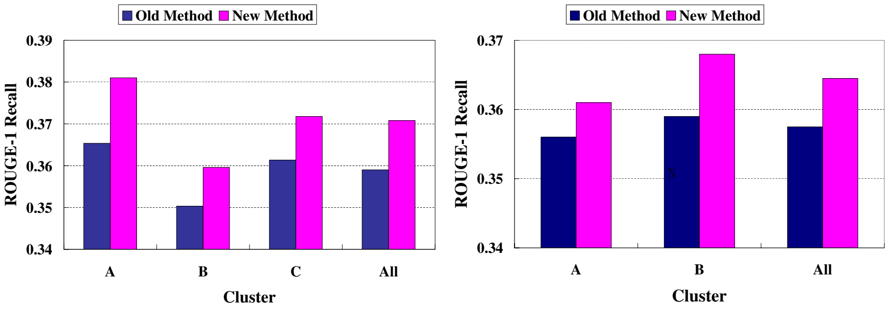
where B_i is mapped to B_i^A under the condition of A. Then for a document B_i and a document set A, B_i^A is a set of B_i's sentences ($B_{i,k}$s) which are different from all the sentences in A_1 to A_m:

$$B_i^A = \{B_{i,k} | \forall\ sen \in \bigcup_i A'_i, \\ D_{\max}(B_{i,k}, sen) > \varphi\} \qquad (9)$$

where A'_i is the sentence set of a document A_i and φ is a threshold.

We have already developed a framework for summarization. However, the problem is that neither $K(.)$ nor $D_{max}(.,.)$ is computable. we can use frequency count, and use Shannon-Fano code [8] to encode a phrase which occurs in probability p in approximately $-\log p$ bits to obtain a short description.

This approximation method can deal with a sentence in word and phrase granu-



DUC 2007

TAC 2008

Figure 1. Comparisons

Table 1. Evaluation Results

Cluster	Traditional			Update		
Evaluation Method	Best	Ours	Rank	Best	Ours	Rank
AVG Modified Score	0.383	0.311	9	0.307	0.296	4
MacroAVG Modified Score with 3 Models	0.377	0.316	9	0.303	0.292	4
AVG Linguistic Quality	5.932	5.682	3	5.886	5.886	1
AVG Overall Responsiveness	5.159	4.955	2	5.023	5.023	1

larities. Therefore, firstly we divide a sentence into semantic elements; then information distance between two sentences is estimated through their semantic element sets [6].

Semantic element extraction method were simply implemented in TAC 2008 [2] by using named entity recognition and counting the overlap of the words and entities. However, an entity may have different names. For example, “George Bush” and “George W. Bush” were viewed as different entities; “May 15th, 2008”, “May 15, 2008” and “5/15/2008” were recognized as different dates in our TAC 2008 system.

We add coreference resolution to our system this year. Firstly named entities are normalized using wikipedia [9], then different writing styles of dates such as “May 15th, 2008”, “May 15, 2008” and “5/15/2008” are normalized into the same date through regular expressions. Experiment results showed in [6] have proved the effectiveness of our coreference resolution method.

4 Experimental Results

In this section, we will firstly compare our two different summarization method (developed in TAC 2008 and 2009) and then provide the evaluation results on TAC 2009.

4.1 Comparison with TAC 2008's Method

Firstly our newly developed method (called "new method") is compared with the original one in TAC 2008 [2](called "old method"). We compare these two methods on the DUC 2007 and the TAC 2008 update datasets under the ROUGE-1 recall criterion. We can see from the Figure 1 the figure that our system has a got much better performance after using the method based on the newly developed theory framework.

4.2 Results of TAC 2009

Finally our new method is tested on the TAC 2009 dataset. The experiment results under pyramid evaluation methods are shown in Table 1. The results of traditional summarization (Cluster A) and update summarization (Cluster B) are listed separately. "Best" means the best result among all 52 submissions. "Ours" means our system's result. "Rank" means the rankings of our result. We can see from this table that our system performs better on update datasets than on traditional datasets. Our system has got the best result under average linguistic quality and average overall responsiveness on update datasets.

5 Conclusion and Future Work

In this paper, we have built up a document summarization framework based on the theory of information distance. Experiments show that our approach performs well on the TAC 2009 dataset. In future work, we will further study our framework and develop a better information distance approximation method.

Acknowledgment

The work was supported by NSFC under grant No.60803075, the National Basic Research Program (973 project in China) under grant No.2007CB311003. The work was also supported by IRCI from the International Development Research Center, Canada.

References

[1] A. Nenkova, R. Passonneau, and K. Mckeown, "The pyramid method: Incorporating human content selection variation in summarization evaluation," *ACM Transactions on Speech and Language Processing*, vol. 4, no. 2, 2007.

[2] S. Chen, Y. Yu, C. Long, F. Jin, L. Qin, M. Huang, and X. Zhu, "Tsinghua university at the summarization track of tac 2008," in *TAC*, 2008.

[3] M. Li and P. M. Vitányi, *An Introduction to Kolmogorov Complexity and its Applications*. Springer-Verlag, 1997.

[4] C. H. Bennett, P. Gács, M. Li, P. M. Vitányi, and W. H. Zurek, "Information distance," *IEEE Transactions on*

Information Theory, vol. 44, no. 4, pp. 1407–1423, July 1998.

[5] C. Long, X. Zhu, M. Li, and B. Ma, “Information shared by many objects,” in *CIKM*, 2008, pp. 1213–1220.

[6] C. Long, M. Huang, X. Zhu, and M. Li, “Multi-document summarization by information distance,” in *Accepted by ICDM*, 2009.

[7] X. Zhang, Y. Hao, X. Zhu, and M. Li, “Information distance from a question to an answer,” in *SIGKDD*, August 2007.

[8] R. L. Cilibrasi and P. M. Vitányi, “The google similarity distance,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, no. 3, pp. 370–383, March 2007.

[9] F. Li, Z. Zheng, Y. Tang, F. Bu, R. Ge, X. Zhu, X. Zhang, and M. Huang, “Thu quanta at tac 2008 qa and rte track,” in *TAC*.