2010 IEEE International Conference on Data Mining

Multi-Document Summarization Using Minimum Distortion

Tengfei Ma, Xiaojun Wan
Institute of Computer Science and Technology
The MOE Key Laboratory of Computational Linguistics
Peking University
Beijing, China
{matengfei, wanxiaojun}@icst.pku.edu.cn

***Abstract*—Document summarization plays an important role in the area of natural language processing and text mining. This paper proposes several novel information-theoretic models for multi-document summarization. They consider document summarization as a transmission system and assume that the best summary should have the minimum distortion. By defining a proper distortion measure and a new representation method, the combination of the last two models (the linear representation model and the facility location model) gains good experimental results on the DUC2002 and DUC2004 datasets. Moreover, we also indicate that the model has high interpretability and extensibility.**

Keywords - multi-document summarization; information-theoretic summarization; minimum distortion; J-S Divergence; linear representation

I. INTRODUCTION

Document summarization aims to generate a short text from one or more document(s), which conveys the most important information of the original text. With the rapid growth of documents on the Internet, summarization has proved to be an essential task in the area of web data mining. For example, it can be used for news services to compress a group of news articles to a short summary, helping readers to grasp the main points in a short time.

Generally, document summarization can be categorized as abstraction-based or extraction-based. An abstraction-based summary can be seen as a reproduction of the original document(s) in a new way, while the extraction-based summarization focuses on extracting sentences directly from the original document(s). In this paper, we consider generic extraction-based summarization from multiple documents.

Though there is no precise definition about what summary is a good summary, researchers usually follow some common standards [6, 14]:

- **Relevance**: A good summary should contain the most important information, i.e. the extracted sentences should be relevant to main topics of the original documents.
- **Diversity**: The sentences in the summary should be non-redundant.
- **Coverage**: The summary should cover as more topics in the original documents as possible.

Early extractive summarization is based on some heuristic features of the sentences such as their positions in the text, the frequency of the words they contain, or some key phrases indicating the importance of the sentences [19]. More advanced techniques consider the rhetorical structure [21] and semantic relationships [9]. Researchers also leverage these features in some machine learning models [13, 30]. However, these techniques seem to ignore or belittle the redundancy and coverage of the summary.

Carbonell and Goldstein [2] introduce the maximal marginal relevance (MMR) measure which combines query relevance and information novelty in topic-driven summarization. The relevance and the redundancy are simultaneously considered in this model. Moreover, cluster-based [32] and centroid-based techniques [23] have been investigated in these years. They employ clustering to avoid redundancy and then choose the most important or representative sentences by the centrality, LexRank scores [7], and some other features. More recently, Li et al. [14] enhance the diversity, coverage and balance of summaries in supervised single-document summarization.

In this work, we advance a new summarization objective by borrowing a concept in information theory: **distortion** [5]. Our main idea is to use the distortion measures to take place of the above summary standards which cannot be easily quantified integrally, and by minimizing the distortion our final summary can achieve the same or better effects. We see summarization as a data transmission system and assume that the output summary sentences represent the input document sentences. The distortion of the "representation" is used as a measure to evaluate the summary quality. Based on different methods of the representation and the algorithms of minimizing the distortion, we propose three summarization models: p-median model, facility location model and linear representation model. First we adopt the one-to-one representation like the clustering technique (i.e. one original sentence is represented by one summary sentence). Under this assumption we get the first model - p-median model. Then the p-median model is improved by adding constraints or features and we propose the facility location model. At last, we jump out of the idea of clustering and replace the one-to-one representation to many-to-one representation (i.e. we use a linear combination of output sentences to represent one input sentence). Our final approach takes the linear representation model combined with the facility location

1550-4786/10 $26.00 © 2010 IEEE
DOI 10.1109/ICDM.2010.106

model, and gains a Rouge-1 score of 0.39614 on the DUC2004 dataset and 0.35884 on the DUC2002 dataset. The result exceeds most of popular summarization. In addition, we indicate that our final model can be extended to other summarization tasks.

The rest of this paper is organized as follows: Section II describes the previous work of document summarization. Section III and Section IV introduces the motivation and the theoretic foundation. Section V develops cluster-based models to optimize the objective function of distortion. Section VI improves the previous representation method and proposes another optimization model, linear representation model. The experimental results are shown in Section VII and we draw the conclusion and discuss the adaption of our models in Section VIII.

II. Related Work

Extraction-based summarization works by choosing a subset of the sentences in the original documents. It has been an active topic in many research areas, such as Natural Language Processing (NLP) and web data mining. Applications of this technique mainly focus on the domain of news articles, such as Google News [1], Columbia NewsBlaster[2].

The approaches of extraction-based summarization can be categorized as supervised or unsupervised. Supervised extractive summarization approaches generally consider the summarization task as a binary classification problem at the sentence level [13, 31], where the summary sentences are positive samples while the non-summary sentences are negative samples. But this model assumes that the sentences are independent and ignore the relations among the sentences. HMM-based model [4] is then proposed to solve the problem. More recently, CRF-based model [25] is proposed to provide a more effective solution to relax the independence assumption and integrating complex features.

As to unsupervised approaches, early researchers often choose some statistic and linguistic features to score and rank sentences [8]. Gong and Liu [9] employ Latent Semantic Analysis (LSA) to add hidden topic features. Ye et al. [30] argue that the quality of a summary can be evaluated based on how many concepts in the original document(s) that can be preserved after summarization. Besides, some works consider reducing the redundancy in summary. A typical method is based on the criteria of Maximal Marginal Relevance (MMR) [2]. According to MMR, a sentence is chosen for inclusion in summary such that it is maximally similar to the document and dissimilar to the already-selected sentences.

Cluster-based summarization [32] first partitions sentences into topical groups and then ranks sentences by their saliency scores. The centroid-based method [23] leverages the cluster centroids and it has been one of the most popular extractive summarization methods. MEAD is an implementation of the centroid-based method that scores sentences based on sentence-level and inter-sentence features, including the cluster centroids, the position, the TFIDF value, etc. Harabagiu and Lacatusu [10] add information to the clusters via various topic representations. Graph-based model [28, 20] is another extractive summarization model and it is often integrated with cluster-based methods [7, 27]. Erkan and Radev [7] assess the graph-based centrality (LexRank) instead of the centroid score. Wan and Yang [27] improve the Markov random walk model by using cluster-level information.

Our models are unsupervised methods and they propose an optimization objective from the information-theoretic perspective, and then consider the summarization task as an optimization problem. The previous information-theoretic methods for document summarization are based on information distance [17] and entropy estimates [24]. [24] ranks sentences by calculating entropies of symbol units. [17] uses the theory of Kolmogorov complexity and generates summaries that have the minimum information distance with the original documents. It approximates Kolmogorov complexity by compression or the coding theory and then chooses one sentence from each document [18]. The distortion measure in our models is also related to the coding theory, but it is a more accountable concept by seeing the summarization as a data transmission system. In addition, our models are all based on integral optimization of the whole summary sentences, which is different from the information distance method. Summarization models based on optimization methods are rare. Li et al. [14] improve the SVM model by directly optimizing three aspects of good summaries: diversity, coverage and balance, but it is a supervised model and used for single-document summarization.

III. Motivation and Problem Formulation

Given a set of sentences $\Omega = \{x_1, x_2, x_3 ... x_n\}$, where x_i denotes the i^{th} sentence in the documents, the aim of extractive summarization is to select several representative sentences $S = \{\hat{x}_i\} \subset \Omega$.

An intuitive idea of selecting S is to rank the sentences in Ω by some measures and select the sentences with the highest ranks. The ranking system can easily integrate various features of the sentence, but it cannot sufficiently leverage the correlation with the original document(s) if we only consider the word occurrence information, for it calculates the similarity between only one sentence with the whole set. The coverage of the summary is hardly considered in the ranking model either.

Traditional cluster-based models do not avoid the disadvantage. The difference is that they can remove some redundancies and noises by clustering first, hence achieving some extent of diversity. But the subsequent selection in the clusters is essentially another ranking method between one sentence and a subset of the original document(s).

How to develop a model and quantified measures to take advantage of the relevance, or on the contrary, the information loss, between the whole summary sentences and the whole set of original sentences is a problem deserved to

[1] http://news.google.com

[2] http://newsblaster.cs.columbia.edu

be investigated. This problem also has some impact on recent evaluation metrics of the summarization [15].

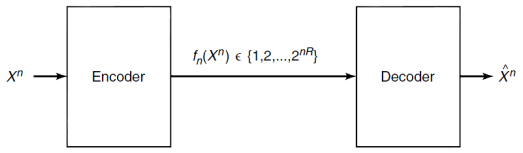

Figure 1. Transmission model (rate distortion encoder and decoder)

As the above discussion refers to the concept of information loss, we develop an information-theoretic model, which sees the summarization as a data transmission system in Fig. 1 [5]. Here the channel is omitted and we should only consider the information loss from the input to the output but ignore the middle process.

The sentences in Ω are represented as values of a variable X , and sentences in S are seen as values of a variable $\hat{X}$. If the original documents are seen as an input of the variable X, the summary is the output of variable $\hat{X}$. Thus, in our approach the summary is a reconstruction of the original documents and every sentence in Ω can be represented by a new value $\hat{x}_i \in S$.

The representation function is defined as:

$$g : \Omega \to S .$$

Now we give a new function defined in the space $\Omega \times S$:

$$d : \Omega \times S \to R^+$$

It is called a distortion function in the information theory, and the distortion $d(x,\hat{x})$ is a measure of the cost of representing the sentence x as the sentence $\hat{x}$ .To measure the sum of the cost, we use expectation of the distortion function:

$$Dis = Ed(X, g(X)) = \sum_{x \in \Omega} p(x) d(x, g(x)) \qquad (1)$$

Using the rate distortion theory [5], the objective of the summarization model in Fig. 1 can be elaborated by the Distortion Rate Function *D(R)*. When the rate *R* (i.e. the R in Figure 1, which can be thought according to the number of sentences in the summary in this case) is limited, our aim is to minimize the expectation of the distortion *Dis* in (1).

IV. DISTORTION MEASURES

While the distortion function is defined differently, the final output will be gained differently. Commonly a sentence *x* can be assumed as a memory-less source of words *Y*. This assumption makes the computation of the distortion more convenient. Actually, it is one of the reasons why we choose data transmission model with the distortion measure.

Then the distortion between sequences X^n and $\hat{X}^n$ can be extended to the following form:

$$d(x^n, \hat{x}^n) = \frac{1}{n} \sum_{i=1}^{n} d(y_i, \hat{y}_i) \qquad (2)$$

where *y* indicate a kind of value of a word. In Hamming distortion *y* indicates the word itself, while in squared error distortion *y* indicates the frequency of the word.

The following are some popular distortion measures.

A. Hamming Distortion:

$$d(y, \hat{y}) = \begin{cases} 0; \ if \ y = \hat{y} \\ 1; \ if \ y \neq \hat{y} \end{cases} \qquad (3)$$

Here y is the word itself. Using (3) in (2), we can see that this distortion mainly evaluate the number of common words between two sentences. In this case, the optimization of the summary can be intuitively explained as sharing the most words with the source without taking into account the weights of words.

B. Squared Error Distortion:

$$d(y, \hat{y}) = (y - \hat{y})^2 \qquad (4)$$

It is the most popular distortion measure used for continuous alphabets [5]. Although there are some disadvantages, it is widely used in image and speech coding. The distortion measure has many useful characteristics: non-negative, non-decreasing, symmetry.

If we see the sentences as points and the distortion as the distance, the optimization of $Dis \propto \sum_x d(x,\hat{x})$ can be seen as a p-median problem. With respect to this assumption, we can use heuristic algorithms for p-median problem to determine which points can be chosen as the reconstruction points. The process will be elaborated in the next section.

[28] has employed the squared error distance to form the clusters , but it is based on the K-means method which is only used to form the clusters and calculate a virtual centroid, instead of real sentences in the original texts.

C. Information Divergence (KLD):

$$d(x, \hat{x}) = D_{KL}(p_x(y) \,||\, p_{\hat{x}}(y)) = \sum_y p_x(y) \log \frac{p_x(y)}{p_{\hat{x}}(y)} \qquad (5)$$

Information divergence is also called K-L divergence (KLD), or relative entropy. It measures the expectation number of extra bits required to code when we use the distribution $p_{\hat{x}}(y)$ to replace $p_x(y)$. Every sentence x is seen as a memory-less source of words *Y*, and the summary is the corresponding output. It is a good measure to evaluate the degree of representation from the information-theoretic perspective. But it has a problem that it is not symmetrical, and it does meet the triangle relation. So it cannot be handled as same as the squared error distortion sometimes.

Besides, the $P_x(y) = P(y \mid x)$ here is finally replaced by $P(x, y)$ in our approach, because we want to add the distortion of each sentence and reflect the integral distortion

between the summary and the source documents, and $P(x, y)$ is more useful to reflect the integral quality.

D. Jensen-Shannon Divergence (JSD)

As the K-L divergence is not symmetrical, the summary based on this measure usually get long sentences. However, the tasks of multi-document summarization are usually limited by the number of words instead of sentences. Thus long sentences may lead to a decrease of the total words in the summary.

One solution is adding length limits to the sentences when optimizing the expectation distortion; the other solution is to replace K-L divergence with J-S divergence, which is a symmetrical measure.

$$D_{JS}(P,Q) = \tfrac{1}{2} D_{KL}(P \| M) + \tfrac{1}{2} D_{KL}(Q \| M);$$
where M is the average of the two distributions, (6)
$$M = \tfrac{1}{2}(P+Q).$$

E. Jensen-Shannon Divergence with Smoothing (JSDS)

The problem using K-L divergence is when an element in the distribution is zero. For example, if a word does not occur in $\hat{x}$ in (5), the $p_{\hat{x}}(y)$ will be zero, and the K-L divergence will be infinite. In this case, we should lead in some smoothing method to solve the problem. Bayes-smoothing [33, 15] is a widely used smoothing method in language models:

$$p = \frac{a_i}{a_0} \rightarrow p = \frac{a_i + \mu p(w_i \mid T)}{a_0 + \mu} \tag{7}$$

where μ is a scaling factor and $p(w_i \mid T)$ is the probability of word *i* occurring in topic T.

J-S divergence can also be improved by the above smoothing method [15].

$$\begin{aligned} d(x,\hat{x}) &= D_{JSDS}(p(x,y) \| p(\hat{x},y)) \\ &= \tfrac{1}{2}\sum_y (p(x,y)\log\frac{p(x,y)}{\tfrac{1}{2}(p(x,y)+p(\hat{x},y))} \\ &\quad + p(\hat{x},y)\log\frac{p(\hat{x},y)}{\tfrac{1}{2}(p(x,y)+p(\hat{x},y))}) \end{aligned} \tag{8}$$

where $p(x,y) = \dfrac{OccurenceInSentence(y) + \mu p(y \mid T)}{OccurenceInDocument(y) + \mu}$, μ take s a value of 2000 following [15] and [33].

F. Other Distortion Measures

Some other distortion measures are also used for data compressing or clustering, such as various divergences introduced in [11]. And if we use the perspective of distortion, the information bottleneck method [26] adopts the loss of mutual information like a distortion measure. [11] demonstrates the rate distortion theory using information divergence distortion is equal to the information bottleneck method in clustering. However, in the document summarization, the two algorithms are not the same. A simple example is that when we represent all the source sentences using the sentence *x* with the highest *T(x,Y)* as follows, the $I(\hat{X}, Y)$ will be the highest, but it is obviously not the best summary.

$$I(X,Y) = \sum_x p(x) \sum_y p(y \mid x) \log \frac{p(y \mid x)}{p(y)};$$

$$T(x,Y) = p(x) \sum_y p(y \mid x) \log \frac{p(y \mid x)}{p(y)}$$

The reason is that it only respects the relation between the global information (*Y*), but ignores the information loss between the sentences. In clustering the new representation is a class which contains the original sentence, but in summarization the representation is a new sentence which has nothing to do with the original sentence if we use the mutual information loss as the distortion.

V. Cluster-based Models using Distortion Measures

A. P-median Clustering Model

As discussed above, we use the data transmission model and the Distortion Rate Function (1) to solve the summarization problem. If the summary has a definite number (*N*) of sentences, there is no need to consider the rate region. So the optimization problem is as follows in (9).

$$\begin{aligned} &\min_S Dis = Ed(X,\hat{X}) = \sum_{x\in\Omega} p(x) d(x, g(x)) \\ &= \sum_{x\in\Omega} p(x) \sum_{\hat{x}\in S} p(\hat{x} \mid x) d(x,\hat{x}) \\ &= p(x) \sum_{\hat{x}\in S} \sum_{x\in\Omega} p(\hat{x} \mid x) d(x,\hat{x}) \\ &= p(x) \sum_{\hat{x}\in S} \sum_{x\in И(\hat{x})} d(x,\hat{x}); \\ &\text{where } p(x) = 1/N,\ p(\hat{x} \mid x) \in \{0,1\}, \\ &\text{and } И(\hat{x}) \text{ denotes the partion generated by } \hat{x}. \end{aligned} \tag{9}$$

The problem can be solved by two heuristic approaches: the agglomerative approach and the interchange approach.

The agglomerative approach first assumes all the sentences in Ω are the representative sentences. Then one sentence is merged into a partition region in every step until the number of the sentences in the summary is *N*. The process is in fact a kind of hierarchical clustering, and this method can also serve as the base clustering method for traditional cluster-based summarization.

The interchange approach randomly chooses *N* sentences as the initial points and then starts an iteration process to replace the former point with a new point and gain a lower cost of the objective. In this approach, the problem is seen as

a p-median problem, and the iteration process is local search [1].

Our final cluster-based model uses the result of the agglomerative approach as the initial point and then searches a local optimization result of the objective function. The process can be interpreted as follows:

1) The Agglomerative Approach

a) Calculate all the distortions between every two sentence, and assign each sentence to a single cluster.

b) Agglomerate the two sentences with the least distortion and form a new cluster. The distance between the new cluster and another cluster is computed by the largest distortion between any two sentences of the two clusters.

c) Agglomerate the two clusters with the minimum distance and form a new cluster. Re-compute the distance between this cluster and other clusters.

d) Repeat c) until the shortest distance has reached the threshold.

e) Calculate the sum distortion of a sentence with others in a cluster. Choose the sentence with the least sum distortion as the deputy (centroid) of this cluster.

2) The Interchange Approach

f) Use the centroids computed by the agglomerative approach as initial points.

g) Assign each sentence to a centroid by choosing the least distortion.

h) Recalculate the centroids as e).

i) Repeat g) and h), until the centroids do not change any more.

B. Facility Location Model

In the former discussion, we can add length constraints when using K-L divergence as the distortion measure. Moreover, in the rate distortion model, if the rate is not a constant, according to the rate distortion theory, the objective function can be written as $\min I(X,\hat{X}) - \beta Dis$.

These problems provide a motivation of adding additional features into the summary cost. The above p-median problem is then converted to a facility location problem [1]. Take the length constraint as an example. If we choose a sentence $\hat{x}$ into the summary, the cost has been changed to:

$$LengthPunish(\hat{x}) + \sum_{x \in H(\hat{x})} d(x,\hat{x}); \tag{10}$$

Here we define the punishment function as

$$LengthPunish(x) = \begin{cases} \beta(Length(x) - \max); \text{if} Length(x) > \max \\ \beta(\min - Length(x)); \text{if} Length(x) < \min \end{cases}$$

The values of max and min depend on the datasets and are described in our experiments.

Then the final objective function is changed as follows:

$$\begin{aligned} \min_S Dis &= E(Cost(X) + d(X,\hat{X})) \\ &= \sum_{x \in \Omega} p(x)(LengthPunish(\hat{x}) + d(x,\hat{x})) \\ &= p(x) \sum_{\hat{x} \in S} \sum_{x \in H(\hat{x})} (LengthPunish(\hat{x}) + d(x,\hat{x})) \end{aligned} \tag{11}$$

This idea complements the shortage of only using information distortion as the standard of summary selection, and it can integrate other features (such as features in the centroid-based method) or constraints in our model. The optimization algorithm is similar to p-median clustering, and we both use the simple local search method. Thus, our model gains a good extensibility without adding much complexity.

VI. Linear Representation Model

A. Motivation

In the above cluster-based models, we assume every original sentence is represented by a new sentence. However, it is not an optimal representation. Intuitively, if a sentence is represented by more sentences instead of a single "center", the information loss may be less.

To formulate this idea, we use the linear combination of $\hat{X}$ to represent X . Thus the distortion function is changed to:

$$d : \Omega \times \lambda(S) \to \mathbb{R}^+;$$

where $\lambda(S)$ denotes the linear generative space of S.

In case of this assumption, the output of the transmission system in Fig. 1 is not changed. The change can be respected as only adopting a different transmission process.

The expectation of distortion then is changed to:

$$\begin{aligned} &Dis = \sum_{x_i \in \Omega} p(x_i) \min_{\{\hat{\lambda}_{ij}\}} d(x_i, \sum_{\hat{x}_j \in S} \hat{\lambda}_{ij} \hat{x}_j); \\ &\text{where } \sum_j \hat{\lambda}_{ij} = 1, \text{ and } \hat{\lambda}_{ij} \in \mathbb{R}^+ \end{aligned} \tag{12}$$

We prove that linear representation is better than one-to-one representation along the distortion measure of J-S divergence in Appendix *A*, i.e.

$$\min_{\{\hat{\lambda}_{ij}\}} d(x_i, \sum_{\hat{x}_j \in S} \hat{\lambda}_{ij} \hat{x}_j) \le d(x_i, \hat{x}_i) , \tag{13}$$

And if there exists

$$\partial d(x_i, \sum_{\hat{x}_j \in S} \hat{\lambda}_{ij} \hat{x}_j) \Big/ \hat{\lambda}_{ij} < \partial d(x_i, \sum_{\hat{x}_j \in S} \hat{\lambda}_{ij} \hat{x}_j) \Big/ \hat{\lambda}_{ii}$$

at $\hat{\lambda}_{ii} = 0$ and $\hat{\lambda}_{ij} = 0 (j \neq i)$, (11) will become:

$$\min_{\{\hat{\lambda}_{ij}\}} d(x_i, \sum_{\hat{x}_j \in S} \hat{\lambda}_{ij} \hat{x}_j) < d(x_i, \hat{x}_i) . \tag{14}$$

However, (12) is hard to compute because we must calculate the set $\{\hat{\lambda}_{ij}\}$ for every sentence. Thus we consider a representation in document-level, i.e. we consider the whole input documents as a word source and the output summary is also a set of words by combining the sentences with different weights.

We expect that the distortion between the whole summary and the whole original documents is smaller than the sum of sentence-level distortion:

$$\min_S \min_{\{\lambda_i\}} \frac{1}{n} d(\sum_{x\in\Omega} x, \sum_{\hat{x}_i\in S} \lambda_i \hat{x}_i) \le \min_S \sum_{x_i\in\Omega} \frac{1}{n} \min_{\{\hat{\lambda}_{ij}\}} d(x_i, \sum_{\hat{x}_j\in S} \hat{\lambda}_{ij}\hat{x}_j);$$

where $\sum_i \lambda_i = n,\ \lambda_i \in \mathbb{R}^+$. n is the number of sentences in Ω. (15)

This assumption is proved to be true in Appendix *B*. Thus we can directly calculate the distortion between the whole summary and the original documents without considering the representation of each sentence. In this way, our final objective function becomes:

$$Dis \propto \min_S \min_{\{\lambda_i\}} d(\sum_{x\in\Omega} x, \sum_{\hat{x}_i\in S} \lambda_i \hat{x}_i);\ \ \frac{1}{n} \text{ is omitted here.} \quad (16)$$

B. The Approach of the Linear Representation Model

1) The Optimization Process

Using the above minimum objective (16), we develop an iterative algorithm based on an interchange process:

a) Choose initial sentences. In our experiment, we adopt the former result of our cluster-based method (the interchange approach).

b) Determine λ_i for corresponding $\hat{x}_i$.

$$Dis = \min_{\{\lambda_i\}} d(\sum_{x\in\Omega} x, \sum_{\hat{x}_i\in S} \lambda_i \hat{x}_i)$$

c) Remove the sentence $\hat{x}_t$ with the lowest λ_t. Add a sentence $\hat{x}_j$ to guarantee that *Dis_new* has the largest decrease.

$$Dis\_new = d(\sum_{x\in\Omega} x, \lambda_t \hat{x}_j + \sum_{\hat{x}_i\in S \,\&\, \hat{x}_i \ne \hat{x}_t} \lambda_i \hat{x}_i)$$

d) Repeat Step b) and c) until the summary set does not change any more.

2) The Algorithm of Assigning λ_i to $\hat{x}_i$.

We use a gradient algorithm to assign λ_i and the proof is given in Appendix *A*.

a) When the initial summary sentences are given, calculate the distortion (or with the punishment weight together) between each sentence *x* and the summary sentences.

b) A sentence *x* is assigned to the region of the summary sentence $\hat{x}_i$, if $d(x,\hat{x}_i) = \min_{\hat{x}_j\in S}\{d(x,\hat{x}_j)\}$.

c) Calculate the size of the $\hat{x}_i$ region, i.e. the number of sentences assigned to $\hat{x}_i$. Then the size is assigned to λ_i as its initial value.

d) Calculate each value of $g_i = \partial d(\sum_{x\in\Omega} x, \sum_{\hat{x}_i\in S} \lambda_i \hat{x}_i) \Big/ \lambda_i$.

e) Find the largest gradient g_i and the smallest one g_j.

$$g_i = g_i - \Delta h, (\Delta h > 0);\ \text{if } g_i > 0.$$

$$g_j = g_j + \Delta h;\ \text{if } g_j < n,$$

where n is the number of sentences in Ω.

In our following experiments, we take $\Delta h = 0.5$. It is a tradeoff between accuracy and computational complexity.

f) Repeat Step d) and e) until $d(\sum_{x\in\Omega} x, \sum_{\hat{x}_i\in S} \lambda_i \hat{x}_i)$ becomes constant or larger than the last step.

C. Comparison with Soft Partition

Someone may notice that in our initial p-median model, the partition is "hard", i.e. $p(\hat{x}\,|\,x) = \{0,1\}$ (see (9)). The model can be improved by using a soft partition (soft clustering method). In this new model, we assume $p(\hat{x}\,|\,x) \in [0,1]$. Thus every sentence can be represented by several sentences with a serial of probabilities:

$$Dis = \sum_{x\in\Omega} p(x) \sum_{\hat{x}\in S} p(\hat{x}\,|\,x) d(x,\hat{x}) = \sum_{x\in\Omega} p(x) \sum_{\hat{x}_i\in S} \lambda_i d(x,\hat{x}_i); \quad (17)$$

where $0 \le \lambda_i = p(\hat{x}_i\,|\,x) \le 1$.

Intuitively, soft partition and the linear representation have similar effects. Now we compare the two ideas. As J-S divergence has similar characteristics with K-L divergence in these inequalities, we need only to take K-L divergence as the example.

(18) indicates that our linear representation model can attain a smaller distortion than soft partition.

$$\begin{aligned}
&\sum_{\hat{x}\in S} p(\hat{x}\,|\,x) d(x,\hat{x}) \\
&= \sum_{\hat{x}\in S} p(\hat{x}\,|\,x) \sum_y p(x,y) \log \frac{p(x,y)}{p(\hat{x},y)} \\
&= \sum_y \sum_{\hat{x}\in S} p(\hat{x}\,|\,x) p(x,y) \log \frac{p(\hat{x}\,|\,x) p(x,y)}{p(\hat{x}\,|\,x) p(\hat{x},y)} \\
&\ge \sum_y \Big(\sum_{\hat{x}\in S} p(\hat{x}\,|\,x) p(x,y)\Big) \log \frac{\sum_{\hat{x}\in S} p(\hat{x}\,|\,x) p(x,y)}{\sum_{\hat{x}\in S} p(\hat{x}\,|\,x) p(\hat{x},y)} \\
&= \sum_y p(x,y) \log \frac{p(x,y)}{\sum_{\hat{x}_i\in S} \lambda_i \hat{x}_i} = d(x, \sum_{\hat{x}_i\in S} \lambda_i \hat{x}_i);
\end{aligned} \quad (18)$$

where $\sum_{\hat{x}\in S} p(\hat{x}\,|\,x) = 1; \lambda_i = p(\hat{x}_i\,|\,x)$.

VII. EXPERIMENTS

A. Data Sets

Document Understanding Conference (DUC [3]) has organized yearly evaluation of document summarization. Generic multi-document summarization is one of the fundamental tasks in DUC2002 and DUC2004. In DUC 2002, 59 document sets of approximately 10 documents each were provided and generic summaries of each document set with lengths of approximately 100 words or less were required to be created. In DUC 2004, 50 document clusters were provided and a short summary with lengths of 665 bytes or less was required to be created.

B. Evaluation Metric

We use the ROUGE [16] evaluation toolkit[4], which is adopted by DUC for automatically summarization evaluation. It measures summary quality by counting overlapping units such as the n-gram, word sequences and word pairs between the candidate summary and the reference summary. ROUGE-N is an n-gram recall measure computed as follows:

$$ROUGE-N = \frac{\sum_{S \in \{RefSum\}} \sum_{n-gram \in S} Count_{match}(n-gram)}{\sum_{S \in \{RefSum\}} \sum_{n-gram \in S} Count(n-gram)}$$

where n stands for the length of the n-gram, and $Count_{match}(n\text{-}gram)$ is the maximum number of n-grams co-occurring in a candidate summary and a set of reference summaries. $Count(n\text{-}gram)$ is the number of n-grams in the reference summaries.

According to [16], among the evaluation methods implemented in ROUGE, ROUGE-N (N=1, 2) is relatively simple and works well in most cases. Therefore, in our work we employ ROUGE-1 and ROUGE-2 to score the summaries.

C. Experimental Results

We evaluate all the proposed models with different distortion measures on the DUC2004 dataset, and the results on the DUC2002 dataset further show the improvement of the models. Table I and Table II list the comparison results generated by our models.

As the Agglomerative approach of the p-median model is most simple and its result can provide initial sentences for other approaches, we take it as the baseline. The Interchange approach (p-median) is then used to improve the Agglomerative approach. In these two approaches, four distortion measures are employed on the DUC2004 dataset: Hamming, Squared Error, K-L divergence with smoothing (KLDS) and J-S divergence with smoothing (JSDS). In the agglomerative process, the cluster threshold is empirically set to 0.9 for Hamming and Squared Error distortion; and when two sentences have less than two common words, we assign the KLDS and JSDS with a large value (1.0) and stop the clustering process according to this value. The results in Table I show that JSDS is the best measure in the interchange approach, while in the agglomerative approach different distortion measures achieve similar results. On DUC2002, we do not test all distortion measures but only use the best measure-JSDS to demonstrate the effectiveness of the improved models.

To limit the lengths of summary sentences when using KLDS, we add the length punishment function and solve the optimization problem using the facility location model. We also tried this model with the JSDS measure. On the DUC2004 dataset, we punish sentences whose lengths are more than 100 bytes or less than 50 bytes. And on the DUC2002 dataset, we assume the length of a good sentence is between 7 and 20 words. In the listed results, we find that in most cases the length constraint leads to performance improvement. The model can be further extended by adding more features like the positions and structure features; however, we do not investigate other features in this work. Our main aim here is to demonstrate the extendibility of our model.

The final improvement comes from the usage of linear representation. As our method usually gains a local optimization, the selection of the initial sentences is crucial and it can greatly impact the final result. Fortunately, we always obtain performance improvement when using results of former runs (the interchange approach of the p-median model and the facility location model) as the initial sentences. We do not conduct experiments using distortion measures other than KLDS and JSDS at this step.

The linear representation model with length punishment method (i.e. we use the result of facility location model to initiate the linear representation model and add length punishment to the summary sentences.) achieves the best performance on both DUC2004 and DUC2002. This indicates the effectiveness of the two techniques, and also proves the distortion is a good standard to estimate the quality of summaries.

We also compare our results with some other popular models in Table III and Table IV. (As all the results of these models are cited from their original papers which maybe experiment only on one of our datasets, finally we have different control groups on DUC2002 and DUC2004, and "-" indicates there is no reported score in this term.) First, we list the best performance values of the DUC2002 and DUC2004 participants. Moreover, on DUC2004, the human summaries are also evaluated and the official ROUGE scores are given. In comparison with the results provided by [7], we can see the advantage of our model over the traditional cluster-based models, such as MEAD and LexRank. The topic theme method [10], the language independent graph-based model [22], and a semi-supervised model [29] are also included,

[3] http://duc.nist.gov/

[4] We use ROUGEeval-1.4.2 downloaded from http://www.haydn.isi.edu/ROUGE/

and we use the toolkit of "Information Distance" [17] to experiment[5] too.

TABLE I. EXPERIMENTAL RESULTS ON DUC2004 DATA

		DUC2004 Task2	
		Rouge-1	***Rouge-2***
P-Median Model (Agglomerative)	Hamming	0.36756	0.07755
	Squared Error	0.36703	0.07813
	KLDS	0.36583	0.07571
	JSDS	0.36599	0.07495
P-Median Model (Interchange)	Hamming	0.37132	0.08150
	Squared Error	0.37413	0.07845
	KLDS	0.36791	0.07823
	JSDS	0.38235	0.08364
Facility Location Model (Length Punish)	KLDS	0.37208	0.07868
	JSDS	0.38429	0.09107
Linear Representaion Model	KLDS	0.38095	0.08262
	JSDS	0.38599	0.08345
Linear Representaion Model with Length Punishment	KLDS	0.37996	0.07749
	JSDS	**0.39614**	**0.09179**

TABLE II. EXPERIMENTAL RESULTS ON DUC2002 DATA

		DUC2002 Task2	
		Rouge-1	***Rouge-2***
P-Median Model (Agglomerative)	JSDS	0.33923	0.07224
P-Median Model (Interchange)	JSDS	0.34625	0.07262
Facility Location Model (Length Punish)	JSDS	0.35262	0.07418
Linear Representaion Model	JSDS	0.35021	0.07673
Linear Representaion Model with Length Punishment	JSDS	**0.35884**	**0.07752**

TABLE III. COMPARISON WITH OTHER MODELS ON DUC2004

	ROUGE-1	**95% confidence**	**ROUGE-2**
Best Human	0.41828	0.40193 - 0.43463	0.10500
Worst Human	0.38902	0.36793 - 0.41011	0.08595
Team65	0.38232	0.37034 - 0.39278	**0.09219**
Team104	0.37436	0.36502 - 0.38568	0.08544
Team35	0.37427	0.36074 - 0.38664	0.08364
Our best model	**0.39614**	0.38244 - 0.41220	**0.09179**
Our Interchange Approach (JSDS)	0.38235	0.37028 - 0.39744	0.08364
Centroid	0.3670	0.3580-0.3767	-
Cont. LexRank	0.3758	0.3617-0.3826	-
Topic Theme[a]	About 0.37	-	About 0.08
Semi-supervised	0.329	-	0.073

a. Its average result is shown in a figure without accurate number values, so we use "about" here to indicate the values are estimated by the figure.

From the results in Table III and Table IV, we can see that our final approach (Linear Representaion Model with Length Punishment) exceeds most of popular models and the participating systems. Especially, we have achieved a result close to the human-annotated result on the DUC2004 dataset. The result of our interchange approach is better than the centroid method, and LexRank (a graph-based method). It shows that the traditional selection methods in a cluster are not good enough and our optimization approach is a better choice, for our method conveys more integral information from the perspective of information theory. The information distance model is not very effective on the DUC2002 dataset, the reason may be that it is a model which is more suitable for topic-focused summarization.

TABLE IV. COMPARISON WITH OTHER MODELS ON DUC2002

	ROUGE-1	**ROUGE-2**
Our best Model	**0.35884**	**0.07752**
Team26	0.35151	0.07642
Team19	0.34504	**0.07936**
Team28	0.34355	0.07521
Pagerank^{W}-U	0.3552	-
Information Distance	0.29216	0.05478

VIII. CONCLUSION AND FURTURE WORK

In this paper, we propose three new models based on the optimization of an information theoretic measure: distortion. The p-median model respects the optimization as a p-median problem and conveys as more information between the whole summary and the whole original documents as possible. The facility location model adds features to the p-median model, and the linear representation model jumps out of the idea of clustering, and modify the representation method. Linear representation is proved to be effective both theoretically and experimentally. The experimental results of our final model which combine the facility location model and linear representation model exceed most of current popular models on the DUC2002 and DUC2004 datasets.

The proposed models are basic models for generic multi-document summarization without adding many features. We focus on their theoretic foundation and extensibility. Though the final model gains a good result in the comparison with other systems, however, more summary features (e.g. positions and structural features) should be integrated in the model in future.

Recently, the tasks of summarization have turned to topic-focused and updated summarization. Researchers must take into account more factors in the new tasks. Fortunately, it is convenient to adapt our model to these new tasks. In the topic-focused task, topic relevance can be added to the cost in the facility location model. In the updated summarization, we can follow the solution in information distance method [17] .

$$D(B_i, S \mid A) = D(B_i^A, S \mid A) = D(B_i^A, S)$$

where a document B_i is mapped to B_i^A under the condition of the document A, and B_i^A is a document set of sentences which can be chosen by a distortion threshold.

[5] Its tool is available at http://www.csai.tsinghua.edu.cn/~hml/resources/summarizer. As it limits the number of summary words, we do not experiment it on DUC2004.

ACKNOWLEDGMENTS

This work was supported by NSFC (60873155), Beijing Nova Program (2008B03) and NCET (NCET-08-0006).

REFERENCES

[1] V. Arya , N. Garg , R. Khandekar , K. Munagala , V. Pandit, "Local search heuristic for k-median and facility location problems," Proceedings of the thirty-third annual ACM symposium on Theory of computing (STOC01), pp.21-29, July 2001, Hersonissos, Greece.

[2] J. Carbonell, and J. Goldstein, "The Use of MMR, diversity-based reranking for reordering documents and producing summaries," in Proceding of SIGIR'98, pp. 335-336, New York, USA, 1998.

[3] K. Chaudhuri, A. McGregor, "Finding Metric Structure in Information Theoretic Clustering," in Proceedings of COLT, 2008.

[4] J. M. Conroy and D. P. Oa´rleary, "Text Summarization via Hidden Markov Models," in SIGIR2001, pp. 406–407, New York, NY, USA, 2001. ACM.

[5] T. Cover and J. Thomas, "Elements of Information Theory", John Wiley & Sons, New York, USA, 1991.

[6] D. Das, and A. Martins, "A Survey on Automatic Text Summarization," Literature Survey for the Language and Statistics II Course at CMU, 2007.

[7] G. Erkan and D. Radev, "LexRank: Graph-based centrality as salience in text summarisation," Journal of Artificial Intelligence Research, vol. 22, pp. 457-479.

[8] J. Goldstein, M. Kantrowitz, V. Mittal, and J. Carbonell, "Summarizing Text Documents: Sentence Selection and Evaluation Metrics," in SIGIR'99.

[9] Y. Gong and X. Liu, "Generic text summarization using relevance measure and latent semantic analysis," in SIGIR'01, pp. 19-25, 2001.

[10] S. Harabagiu, and F. Lacatusu, "Topic Themes for Multi-Document Summarization," in SIGIR'05.

[11] P. Harremoes, N. Tishby, "The Information Bottleneck Revisited or How to Choose a Good Distortion Measure," in Proc. of the IEEE Int. Symp. on Information Theory (ISIT), 2007.

[12] K. S. Jones, "Automatic summarising: The state of the art," Information Processing and Management, vol. 43, pp. 1449-1481, 2007.

[13] J. Kupiec, J. Pedersen, and F. Chen, "A trainable document summarizer," in SIGIR, pp. 68-73, 1995.

[14] L. Li, K. Zhou, G-R. Xue, H. Zha, and Y. Yu, "Enhancing Diversity, Coverage and Balance for Summarization through Structure Learning," in WWW 2009.

[15] C-Y. Lin, G. Cao, J. Gao, and J-Y. Nie, "An Information-Theoretic Approach to Automatic Evaluation of Summaries," in NAACL06.

[16] C.-Y. Lin and E. Hovy, "Automatic evaluation of summaries using n-gram co-occurrence statistics," in HLT-NAACL, 2003, pp. 71-78.

[17] C. Long, M. Huang, X. Zhu, and M. Li, "Multi-Document Summarization by Information Distance," in ICDM2009.

[18] C. Long, S. Chen, Y. Yu, F. Jin, L. Qin, M. Huang, X. Zhu, "Tsinghua University at the summarization track of TAC 2008," in TAC2008.

[19] H. P. Luhn, "The Automatic Creation of Literature Abstracts," IBM Journal of Research and Development, vol. 2, no. 2, pp. 159-165, 1958.

[20] I. Mani and E. Bloedorn, "Multi-document summarization by graph search and matching," In AAAI/IAAI, pp. 622-628, 1997.

[21] D. Marcu, "From Discourse Structures to Text Summaries," in ACL'97/EACL'97 Workshop on Intelligent scalable Text Summarization, pages 82-88, 1997.

[22] R. Mihalcea and P. Tarau, "A language independent algorithm for single and multiple document summarization," in Proceedings of IJCNLP2005.

[23] D. R. Radev, H. Jing, M. Stys, and D. Tam, "Centroid-based summarization of multiple documents," Information Processing and Management, vol. 40 (2004), pp. 919-938.

[24] G. Ravindra, N. Balakrishnan, K. R. Ranmakrishnan, "Multi-Document Automatic Text Summarization Using Entropy Estimates," in SOFSEM2004, pp. 289-300.

[25] D. Shen, J-T. Sun, H. Li, Q. Yang, and Z. Cheng, "Document Summarization using Conditional Random Fields," in IJCAI07.

[26] N. Tishby, F. C. Pereira, W. Bialek, "The information Bottleneck Method," in The 37th annual Allerton Conference on Communication, Control, and Computing, Sep 1999: pp. 368–377.

[27] X. Wan and J. Yang, "Multi-document Summarization Using Cluster-based Link Analysis," in SIGIR08.

[28] X. Wan and J. Yang, "Improved affinity graph based multi-document summarization," in Proceedings of HLT-NAACL2006.

[29] K-F. Wong, M. Wu, and W. Li, "Extractive Summarization Using Supervised and Semi-supervised Learnsing," in Coling 2008, pp. 985-992.

[30] S. Ye, T-S. Chua, M-Y. Kan, and L. Qiu, "Document concept lattice for text understanding and summarization," Information Processing and Management, vol. 43, pp. 1643-1662, 2007.

[31] J-Y. Yeh, H-R. Ke, W-P. Yang, and I-H. Meng, "Text summarization using a trainable summarizer and latent semantic analysis," IPM, 41(1):75-95, 2005.

[32] H. Zha, "Generic Summarization and Keyphrase Extraction Using Mutual Reinforcement Principle and Sentence Clustering," in SIGIR'02.

[33] C. Zhai and J. Lafferty, "A Study of Smoothing Methods for Language Models Applied to Information Retrieval," ACM Transactions on Information Systems, Vol 22, No. 2, April 2004, pp. 179-214.

APPENDIX

A. *Proof 1: Linear representaion is better than one-to-one representation.*

First, we can take $\hat{\lambda}_{ii} = 1; \hat{\lambda}_{ij} = 0 (j \neq i)$ in $\sum_{\hat{x}_j \in S} \hat{\lambda}_{ij} \hat{x}_j$. Thus, $\min_{\{\hat{\lambda}_{ij}\}} d(x_i, \sum_{\hat{x}_j \in S} \hat{\lambda}_{ij} \hat{x}_j) \leq d(x_i, \hat{x}_i)$.

Then, we calculate the partial derivative of $f = d(x_i, \sum_{\hat{x}_j \in S} \hat{\lambda}_{ij} \hat{x}_j)$ with respect to $\hat{\lambda}_{ij}$. Here J-S divergence is taken as an example of the distortion measure.

$$\frac{\partial f}{\partial \hat{\lambda}_{ij}} = \sum_{y} \frac{y_{ij}}{Count} \log \frac{\hat{\lambda}_{ij} y_{ij}}{y_{ij} + \sum_{j} \hat{\lambda}_{ij} y_{ij}} \leq 0$$

Assuming there exists $\hat{\lambda}_{ij}$ subject to $\frac{\partial f}{\partial \hat{\lambda}_{ij}} < \frac{\partial f}{\partial \hat{\lambda}_{ii}}$, the new representation, $\sum_{\hat{x}_j \in S} \hat{\lambda}_{ij} \hat{x}_j; \hat{\lambda}_{ii} = 1 - \triangle h, \hat{\lambda}_{ij} = \triangle h, \hat{\lambda}_{ik} = 0 (k \neq i, j)$, will have a smaller distortion.

$$d_{new} = d_{old} + (\frac{\partial f}{\partial \hat{\lambda}_{ij}} - \frac{\partial f}{\partial \hat{\lambda}_{ii}}) \triangle h < d_{old}$$

Though we cannot guarantee the assumption is always contented, our experiments show the effectiveness of the method later.

B. ***Proof 2*** *: The summary-level distortion is better than the sum of sentence-level distortion.*

For simple elaboration, let us see the case of K-L divergence first.

$$Dis = 1/n \sum_{x_i} D_{KL}(p(y,x_i) \| p(y, \sum_{\hat{x}_j \in S} \hat{\lambda}_{ij} \hat{x}_j))$$

$$= 1/n \sum_{y} \sum_{x_i} p(y,x_i) \log \frac{p(y,x_i)}{p(y, \sum_{\hat{x}_j \in S} \hat{\lambda}_{ij} \hat{x}_j)}$$

$$\geq 1/n \sum_{y} (\sum_{i} p(y,x_i)) \log \frac{\sum_{i} p(y,x_i)}{\sum_{i} p(y, \sum_{\hat{x}_j \in S} \hat{\lambda}_{ij} \hat{x}_j)}$$

(according to the log sum inequality [5])

$$= 1/n \sum_{y} p(y,X) \log \frac{p(y,X)}{p(y, \sum_{\hat{x}_j \in S} \lambda_j \hat{x}_j)}$$

$$= 1/n \, D_{KL}(p(y,X) \| p(y, \sum_{\hat{x}_j \in S} \lambda_j \hat{x}_j))$$

Then, it is easy to see that the log sum equality is also correct in J-S divergence and we can gain the same conclusion when using J-S divergence as the distortion measure.

$$Dis = 1/n \sum_{x_i} D_{JS}(p(y,x_i) \| p(y, \sum_{\hat{x}_j \in S} \hat{\lambda}_{ij} \hat{x}_j))$$

$$= 1/n \sum_{x_i} 1/2 \{D_{KL}(p(y,x_i) \| 1/2 (p(y,x_i) + p(y, \sum_{\hat{x}_j \in S} \hat{\lambda}_{ij} \hat{x}_j)))$$

$$+ D_{KL}(p(y, \sum_{\hat{x}_j \in S} \hat{\lambda}_{ij} \hat{x}_j) \| 1/2 (p(y,x_i) + p(y, \sum_{\hat{x}_j \in S} \hat{\lambda}_{ij} \hat{x}_j)))\}$$

(according to (6))

$$\geq 1/n * 1/2 \{D_{KL}(p(y,X) \| 1/2 (p(y,X) + p(y, \sum_{\hat{x}_j \in S} \lambda_j \hat{x}_j)))$$

$$+ D_{KL}(p(y, \sum_{\hat{x}_j \in S} \lambda_j \hat{x}_j) \| 1/2 (p(y,X) + p(y, \sum_{\hat{x}_j \in S} \lambda_j \hat{x}_j)))\}$$

$$= 1/n \, D_{JS}(p(y,X) \| p(y, \sum_{\hat{x}_j \in S} \lambda_j \hat{x}_j))$$