Toward a Unified Framework for Standard and Update Multi-Document Summarization

HONGLING WANG and GUODONG ZHOU, Soochow University

This article presents a unified framework for extracting standard and update summaries from a set of documents. In particular, a topic modeling approach is employed for salience determination and a dynamic modeling approach is proposed for redundancy control. In the topic modeling approach for salience determination, we represent various kinds of text units, such as word, sentence, document, documents, and summary, using a single vector space model via their corresponding probability distributions over the inherent topics of given documents or a related corpus. Therefore, we are able to calculate the similarity between any two text units via their topic probability distributions. In the dynamic modeling approach for redundancy control, we consider the similarity between the summary and the given documents, and the similarity between the sentence and the summary, besides the similarity between the sentence and the given documents, for standard summarization while for update summarization, we also consider the similarity between the sentence and the history documents or summary. Evaluation on TAC 2008 and 2009 in English language shows encouraging results, especially the dynamic modeling approach in removing the redundancy in the given documents. Finally, we extend the framework to Chinese multi-document summarization and experiments show the effectiveness of our framework.

Categories and Subject Descriptors: I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Text analysis*

General Terms: Algorithms, Experimentation, Languages, Performance

Additional Key Words and Phrases: Multi-document summarization, topic modeling, latent Dirichlet allocation, dynamic modeling

ACM Reference Format:
Wang, H. and Zhou, G. 2012. Towards a unified framework for standard and update multi-document summarization. ACM Trans. Asian Lang. Inform. Process. 11, 2, Article 5 (June 2012), 18 pages.
DOI = 10.1145/2184436.2184438 http://doi.acm.org/10.1145/2184436.2184438

1. INTRODUCTION

Multi-document summarization (MDS) aims to produce a single summary from a set of documents. Usually, it can be described as a three-step process: selection of salient portions of text, aggregation or abstraction of the salient information, and presentation of the summary text [Jones 1999; Mani and Bloedorn 1999].

Over the past few years, MDS has drawn more and more attention and made much progress. However, various evaluations indicate that MDS is highly complex and demanding, and there is still a long way for automatic summarizers to catch up with

This work is supported by the National Natural Science Foundation of China (90920004, 60970056, 60873150).

Author's address: G. Zhou, School of Computer Science and Technology, Soochow University, Suzhou, China; email: gdzhou@suda.edu.cn.

DOI 10.1145/2184436.2184438 http://doi.acm.org/10.1145/2184436.2184438

human beings [Dang and Owczarzak 2008]. Currently, there exist two major issues with respect to MDS:

— How to detect and select salient information to be included in the summary;
— How to control the redundancy in the final summary, given a set of documents.

To address these two issues, this article presents a purely statistical framework for both standard and update MDS in an extractive way. In particular, a topic modeling approach is employed for salience determination using the topic distribution information and a dynamic modeling approach is proposed for redundancy control using the similarities between various kinds of text units, such as sentence, summary, document, and documents. For salience detection, the intuition behind the topic modeling approach is that the summary's topic probability distribution should be similar to the documents' topic probability distribution. For various topics contained in the documents, the central topic represents the main theme and the other topics support around the central one. Together, the central topic and the other ones form a topic probability distribution for the documents. Here, we ignore the summary focus (a short description of desired summary) and a topic is defined as a weighted "bag-of-words" rather than the summary focus while the topic modeling approach is employed to derive the hidden topics from a set of documents, either the given documents or a related corpus. In this way, we can represent various kinds of text units, such as word, sentence, document, documents, and summary, using a single vector space model via their corresponding probability distributions over the derived topics, and calculate the similarity between any two text units via their topic probability distributions. For redundancy control, the dynamic modeling approach is proposed for both standard and update summarization. In particular, for standard summarization, we consider the similarity between the summary and the given documents, and the similarity between the sentence and the summary, besides the similarity between the sentence and the given documents while for update summarization, we also consider the similarity between the sentence and the history documents or summary. Finally, the summary is generated greedily by selecting the sentences according to linear interpolation of these similarity scores in a dynamic way.

The rest of this article is as follows. Section 2 gives an overview of related work. Section 3 gives a brief introduction to topic modeling, in particular LDA, and presents the topic-driven similarity between any two text units. Section 4 presents our MDS framework in details. Section 5 and Section 6 present experimental results on both English and Chinese MDS respectively. Finally, Section 7 draws a conclusion.

2. RELATED WORK

At present, the literature of summarization has grown to a level which is very hard to overview in detail [Jones 2007]. However, we can still identify some critical commonalities in the way of extracting and producing salient information into output summaries. Generally, summarization methods can be categorized into either extraction-based or abstraction-based. Since this article deals only with extraction-based summarization, we only give an overview of the literature in extraction-based summarization.

In the literature, the development of summarization has been largely promoted by Document Understanding Conferences (DUC) and Text Analysis Conferences (TAC)[1]. Recently, new summarization tasks, such as update summarization [Dang and Owczarzak 2008] and comparative summarization [Wang et al. 2009] have been proposed. Here, update summarization is a task relative to standard summarization and

[1] See http://duc.nist.gov/ and http://www.nist.gov/tac/.

the update summary is generated from a set of documents, under the assumption that the user has already read a set of history documents. In the literature, the standard summary is also called main summary in DUC 2007 and initial summary in TAC 2008 and 2009.

In the following paragraphs, we will overview the state-of-the-art in summarization from the view points of salience determination and redundancy control. Besides, we will also make a brief overview for the literature of summarization in Chinese language.

2.1. Salience Determination

To determine the salience of information, researchers have used various positional and structural properties of the judged sentences with respect to the source texts. These properties can be the sentence position in a document, the fact that a sentence is part of the title or the abstract of a document [Edmundson 1969; Radev et al. 2000], and the relation of sentences with respect to a user-specific query or a specified topic [Park et al. 2006; Varadarajan and Hristidis 2006].

Following the bag-of-words assumption, a sentence in the given documents is often represented as a vector of features, such as the sentence position, the sentence length, the cosine similarity between the sentence and the document title, and the sentence's TF-IDF value, in summarization [Torralbo et al. 2005]. In other cases, further analysis is performed, aiming to reduce the dimensionality and produce a vector in a latent topic space [Bhandari et al. 2008; Steinberger and Jezek 2004]. The vector representation can be exploited for measuring the semantic similarity between information chunks by using various measures, such as the cosine distance and Euclidean distance between vectors. Such a vector representation can be effectively exploited to measure the salience of a sentence.

In this article, different kinds of text units, such as word, sentence, document, documents, and summary, are all represented as a probabilistic distribution vector over the inherent topics in the given documents or a related corpus. Here, topic modeling is deployed as a clustering tool to derive the inherent topics. In this way, each sentence can be represented accordingly.

Similar to our work, Haghighi and Vanderwende [2009], Arora and Ravindran [2008b], and Bhandari et al. [2008] use a topic model to represent a sentence. Haghighi and Vanderwende [2009] utilize a hierarchical LDA-style model to represent a topic as a hierarchy of topic vocabulary distributions. Each sentence is then represented using three kinds of topic distributions, that is, the background vocabulary distribution, the content distribution, and the document-specific vocabulary distribution. Arora and Ravindran [2008a] use LDA to find different topics in the documents and include the probability of a topic as a feature in sentence presentation. Arora and Ravindran [2008b] further improve the performance by using SVD to find a better topic representation for a sentence. Bhandari et al. [2008] use PLSI, another popular topic modeling tool, to divide a document into several topics and include as a feature the single probability that the sentence occurs in the major topic.

Among others, Nastase [2008] defines a topic using a set of sentences or questions (i.e. query). The disadvantage of this approach is that it assumes a large amount of (related) manual summaries in advance, which is impractical for automatic summarization. In comparison, this article automatically derives the topic probability distribution from a document collection using a topic modeling tool, which is scalable to automatic summarization on a large number of documents.

With the progress of graph-based learning in the machine learning community, graph-based methods have been drawing more and more attention in recent years

[Erkan and Radev 2004; Mihalcea 2005]. Graph-based methods view each sentence as a node in a graph and the similarity between sentences as links between corresponding nodes. In particular, the sentences are ranked using some graph ranking algorithms such as HITS [Kleinberg and Authoritative 1998] and PageRank [Brin and Page 1998]. However, such graph ranking algorithms tend to give the highest rank to the sentences which are related to the central topic in the documents. Therefore, these algorithms tend to choose those sentences related with the central topic and ignore other sentences related with other topics. This makes the extracted summary fail to cover other useful topics in the documents. This problem is particularly serious in multi-document summarization. Another drawback is that such algorithms are normally very complicated, especially given a large number of documents for summarization.

2.2. Redundancy Control

As the core component in MDS, sentence selection has drawn most attention in the literature. There are two main methods to sentence selection: ranking and clustering [Dang and Owczarzak 2008]. In the first method, the sentences in the given documents are ranked and those top-ranked sentences are selected to form the summary. In the second method, the sentences are clustered and a central sentence is extracted from each cluster to form the summary.

One key problem here is how to control the redundancy. In fact, the research on redundancy control has given birth to MR (Marginal Relevance) and MMR (Maximal Marginal Relevance) sentence selection criteria [Carbonell and Goldstein 1998]. For example, MMR chooses the sentences according to the weighed combination of their general relevance with the document and their redundancy with the sentences already chosen. Alternatively, CSIS (Cross-Sentence Informational Subsumption) determines whether and in what degree a sentence has been contained in another sentence already chosen in the summary [Radev et al. 2000]. Among others, Allan et al. [2003] use statistical features of the judged sentences with respect to the sentences already chosen in summary to avoid repetition.

In particular, among topic-driven approaches, Haghighi and Vanderwende [2009] greedily augment the sentences to form a summary via minimizing the KL-divergence between the summary and the documents. Arora and Ravindran [2008b] use SVD to find the most orthogonal sentences over the topic distribution to form the summary. Bhandari et al. [2008] extract the sentences from the highest-ranked topic as the basic summary and gradually combine the sentences from different topics to form the final summary. Although this approach is successful in single document summarization, it may not be readily applied to multi-document summarization. The reason is that it is difficult for this approach to cope with the huge redundancy in the given documents.

Among others, Nastase [2008] generates a summary based on the probability distribution of unigrams in human summaries. The disadvantage of this approach is that it assumes a large amount of (related) manual summaries in advance, which is impractical for automatic summarization.

In this article, a dynamic modeling approach is proposed to control the redundancy for both standard and update summarization given a set of documents. For standard summarization, we extend Haghighi and Vanderwende [2009] by considering both the similarity between the sentence and given documents, and the similarity between the sentence and the summary, besides the similarity between the summary and given documents. For update summarization, we also consider the similarity between the sentence and the history documents or summary. In particular, the KL-divergence is employed to measure the similarity between any two text units. Finally, the sentences are chosen greedily to form a summary according to linear interpolation of these similarity scores.

In summary, although some studies have used the topic distribution information in summarization, our study extends the idea to use such information in redundancy control via a dynamic modeling approach for both standard and update MDS in a unified framework.

2.3. MDS in Chinese Language

Compared with the large number of studies on English MDS, there are only a few on Chinese MDS, much due to the lack of publically-available annotated corpora and evaluation criteria. Representative works include Xu et al. [2007] and Liu et al. [2006].

Inspired by Radev et al. [2001], Xu et al. [2007] propose a multi-document framework, which first simplifies the traditional multi-document representation model in the cross-structure theory and then supplements the change and distribution information about event topics absent from the information fusion theory.

Liu et al. [2006] propose a MDS model to both maximize the coverage of topics and minimize the redundancy of contents. They first extract a set of concepts by combining semantic analysis and various statistical techniques to represent the information content of documents and then calculate the relevance between sentences based on concept cohesion. In this way, the relationship among sentences and documents are established.

3. TOPIC MODELING AND TOPIC-DRIVEN SIMILARITY

This section briefly introduces topic modeling and presents the topic-driven similarity between any two text units using the topic model derived from a topic modeling approach.

3.1. Topic Modeling

Among various topic modeling approaches, Latent Dirichlet Allocation (LDA) [Blei et al. 2003] has drawn the most attention recently in the NLP community and has been applied successfully in topic detection. In principle, LDA is a generative three-level hierarchical Bayesian probabilistic model for analyzing the content of documents and the meaning of words. Similar to other topic models, such as LSA, PLSA, and PLSI, LDA assumes that documents are mixtures of topics and a topic can be represented as a probability distribution over words. In this article, we use LDA to capture the topics in the documents. In particular, we employ the LDA toolkit[2] developed by Phan and Nguyen in our multi-document summarization system. Since LDA assumes a topic to be a weighted "bag-of-words" and we don't involve any grammatical information, our approach is purely statistical.

Mostly, the basic LDA model will be extended to a smoothed version to gain better result. The plate notation is shown in Figure 1. α is the parameter of the uniform Dirichlet prior on the per-document topic distributions. β is the parameter of the uniform Dirichlet prior on the per-topic word distribution. θ_i is the topic distribution for document i, z_{ij} is the topic for the jth word in document i, and w_{ij} is the specific word. w_{ij} is the only observable variable, and the other variables are latent variables. K denotes the number of topics considered in the model and ϕ is a K*V (V is the dimension of the vocabulary) Markov matrix each line of which denotes the word distribution of a topic.

[2] See http://sourceforge.net/projects/jgibblda.

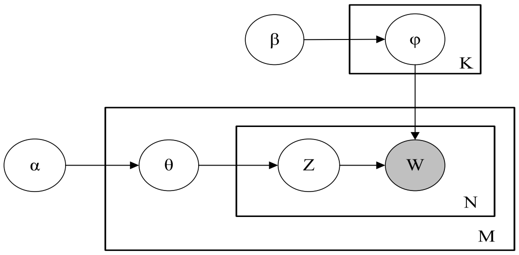

Fig. 1. The plate annotation for smoothed LDA.

In this article, we investigate two ways to apply LDA. One is to derive a single topic model using all the document sets in the given corpus (called Generic Topic Model). Another way is to derive a topic model for each particular document set under consideration (called Specific Topic Model). Here, both LDA models use the following default parameters: $\alpha = 50/K$, $\beta = 0.1$, where K stands for the number of topics. In particular, K is fine-tuned to the number of document sets for the generic topic model and the number of documents in the particular document set for the specific topic model while the number of iterations is fine-tuned to 2000, using the DUC 2007 English evaluation corpus as the development data. For details, please refer to Section 5.1.

3.2. Topic-Driven Similarity

Given any two text units a and b, the topic-driven similarity between them is calculated in this paper as,

$$TSim(a,b) = -(D_{KL}(P_a||P_b) + D_{KL}(P_b||P_a)), \tag{1}$$

where P_a and P_b are the probability distributions of text units a and b over the derived topics respectively. Here, $D_{KL}(P_a||P_b)$ measures the Kullback-Leibler (KL) divergence [Kullback and Leibler 1951] between two probability distributions P_a and P_b as follows:

$$D_{KL}(P_a||P_b) = \sum_i P_a(i) \log \frac{P_a(i)}{P_b(i)} \tag{2}$$

Since the KL divergence is asymmetric, both $D_{KL}(P_a||P_b)$ and $D_{KL}(P_b||P_a)$ are included to guarantee the symmetry of the topic-driven similarity measure.

Given any text unit, its topic probability distribution can be easily determined from the topic model using the topic distribution P(Z|D), where D indicates the given document set and Z indicates the K derived topics, and the word distribution for each topic $z_i \in Z \, (i = 1, 2 ... K)$, that is, the topic-word distribution $P(W|z_i)$, where W stands for the set of words (i.e., bag-of-words) considered in the topic model. For example, assume X is any text unit and P_X represents X's probability distribution over the K derived topics, that is,

$$P_X = (P(z_1|X), P(z_2|X), \ \dots \ , P(z_K|X)), \tag{3}$$

where $P(z_i|X)$ stands for the probability of text unit X belonging to topic z_i and can be computed as

$$\begin{aligned} P(z_i|X) &= \tfrac{P(X,z_i)}{P(X)} = \tfrac{P(X|z_i)P(z_i|D)}{P(X)} \\ &= \frac{\prod\limits_{j=1...|X|} P(x_j|z_i)P(z_i|D)}{P(X)} \end{aligned} \tag{4}$$

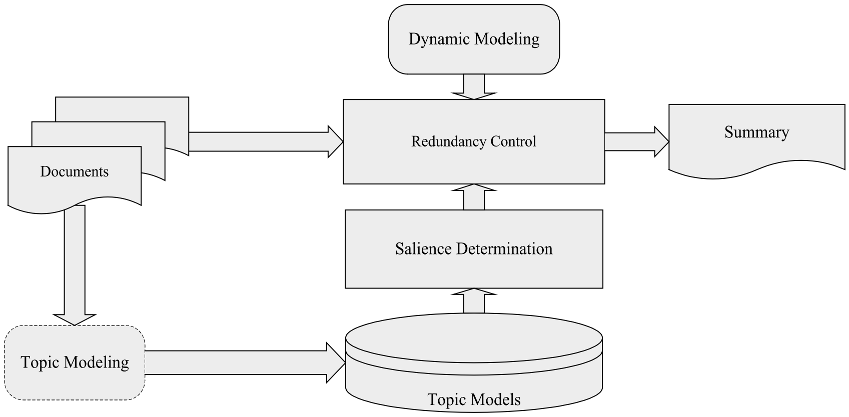

Fig. 2. The unified MDS framework.

Where x_j is the j-th token in text unit X and P(X) is a normalization factor, which can be computed as

$$P(X) = \sum_{i=1...K} \prod_{j=1...|X|} P(x_j|z_i)P(z_i|D) \tag{5}$$

4. UNFIED FRAMEWORK FOR STANDARD AND UPDATE MDS

Figure 2 illustrates the unified framework for standard and update MDS. Please note that various kinds of preprocessing, such as stemming, stop-word removal, word segmentation, and sentence segmentation are called upon beforehand. This section focuses on salience determination in how the salience of various kinds of text units is determined using the topic model derived from the topic modeling approach, and standard/update summary generation in how a dynamic modeling approach is employed to remove the redundancy.

4.1. Salience Determination

Given the topic-driven similarity between any two text units as described in Section 3.2, the salience of a sentence *s* can be easily measured by the topic-driven similarity between the sentence and the given documents, *TSim(s, Docs)*.

In the literature, some popular features have been widely used and proven important for the success of MDS [Dang and Owczarzak 2008], for example, the sentence position, the sentence length, the cosine similarity between the sentence and the document title (using the bag-of-words representation), and the sentence's TF-IDF value. Due to the failure of our topic-driven similarity to cover such useful information, we adjust the salience score of a sentence via linear interpolation of those popular feature scores with the original salience score as follows, for example, the topic-driven similarity between the sentence S_r and the given documents *Docs*:

$$\begin{aligned} Score\ (S_r) = {} & TSim\ (S_r, Docs) + w_1 Pos\ (S_r) \\ & + w_2 Len\ (S_r) + w_3 CosSim\ (S_r, Title) \\ & + w_4 Tfidf\ (S_r) \end{aligned} \tag{6}$$

where w_1, w_2, w_3, and w_4 are relative weights of the following popular feature scores.

(1) $Pos(S_r)$. For the sentence position, early sentences are considered more likely to contain focused and important information. If the document has n sentences, the position score of ith sentence can be computed as $Pos(S_i) = (n - i + 1)/n$.
(2) $Len(S_r)$. For the sentence length, since very short and very long sentence are considered unlikely to be useful, we evaluate the score of sentence length using a normal distribution, which measures the probability of a sentence length with the average length of all sentences as the mean.
(3) $CosSim(S_r, Title)$. The title of a document often contains important information. So we calculate the cosine similarity between the sentence and the title.
(4) $Tfidf(S_r)$. The TF-IDF score has also been widely used in information retrieval to measure the similarity between a query (here, a sentence) and a document.

In this article, w_1, w_2, w_3, and w_4 are fine-tuned to 2, 0.3, 0.9, and 0.3, respectively, using the DUC 2007 English evaluation corpus as the development data. For details, please refer to Section 5.1.

4.2. Standard Summary Generation

In this article, we explore two approaches in generating the standard summary: a static modeling approach and a dynamic modeling approach. Compared with the former baseline approach, the latter further controls the redundancy in a dynamic way.

4.2.1. Static Modeling. The static modeling approach extracts the sentences with the highest salience scores to produce the summary in the following way.

(1) Run LDA to construct the topic model.
(2) Compute the salience score of each sentence, i.e. the similarity between the sentence and the given documents with possible adjustment using the four popular feature scores.
(3) Order the sentences according to their salience scores.
(4) Pick up the sentences with the highest salience scores in the descending order and include them in the summary until the summary reaches the size limitation.

4.2.2. Dynamic Modeling. The dynamic modeling approach is proposed to remove the redundancy in the given documents. It incrementally chooses a sentence and augments it into the summary to maximize the similarity between the augmented summary and the given documents in a dynamic way. In other words, whenever a sentence is added to the summary, the similarity between the augmented summary and the given documents is calculated. In particular, the sentence which makes the summary most similar to the given documents is picked up to augment the summary. In this way, redundant information is avoided effectively. Compared with the above static modeling approach, the dynamic modeling approach generates the summary in the following dynamic way.

(1) Run LDA to construct the topic model, compute the salience score of each sentence and order the sentences according to their salience scores, in the same way as the static modeling approach.
(2) Pick the top-scoring sentence as the initial summary.
(3) Pick the sentence which maximizes the similarity score between the augmented summary and the given documents.
(4) Repeat last step until the summary has reached the size limitation.

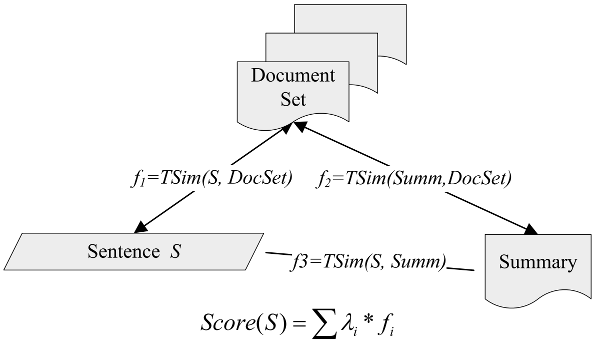

Fig. 3. The dynamic model for standard summarization.

Figure 3 illustrates the dynamic model for standard summary generation, which calculates the salience score of a sentence S via linear interpolation as

$$Score(s) = \sum \lambda_i * f_i \tag{7}$$

Here, besides the similarity between the augmented summary and the given documents (f_2), we also consider both the similarity between the sentence S and the given documents (f_1) (i.e., the salience score as shown in Equation (6)), and the similarity between the sentence S and the augmented summary (f_3). Obviously, it becomes a static model when $\lambda_2 = 0, \lambda_3 = 0$, and the simple dynamic model as described in Haghighi and Vanderwende [2009] when $\lambda_1 = 0, \lambda_3 = 0$. In this article, the parameters are fine-tuned to $\lambda_1 = 1.5, \lambda_2 = 0.4, \lambda_3 = -0.1$ in standard summarization using the DUC 2007 English evaluation corpus as the development data. For details, please refer to Section 5.1.

4.3. Update Summary Generation

Since an update summary is generated under the assumption that the user has already read a history set of documents, we extend the dynamic model in standard summary generation to consider the similarity between the sentence and the given history documents or simply the corresponding history summary for update summary generation.

Figure 4 illustrates the idea, where DocSet A, Summary A, DocSet B, and Summary B indicate the history set of documents, the history summary (for DocSet A), the update set of documents and the current summary (for DocSet B), respectively. In this model, a sentence is scored by linearly interpolating the similarity between the sentence S and the current documents (f_1), the similarity between the current summary and the current documents (f_2), the similarity between the sentence S and the current summary (f_3) and the similarity between the sentence S and the history documents (f_4) or the history summary A (f_5).

Obviously, this dynamic model becomes the static model for standard summary generation when$\lambda_2 = 0, \lambda_3 = 0, \lambda_4 = 0$, and the dynamic model for standard summary generation when$\lambda_4 = 0$. Besides, we could replace f_4 with f_5 by simply considering the similarity between the sentence S and the history summary (dynamic update with history summary) instead of the similarity between the sentence S and the history documents (dynamic update with history documents). In this article, all the history reference summaries generated by the human annotators are included.

In this article, the parameters are fined-tuned to$\lambda_1 = 0.4, \lambda_2 = 1.5, \lambda_3 = -0.1, \lambda_4 = -0.1$ for dynamic update summarization with history summary and $\lambda_1 = 0.4, \lambda_2 = 1.5, \lambda_3 = -0.1, \lambda_5 = -0.2$ for dynamic update summarization with history documents,

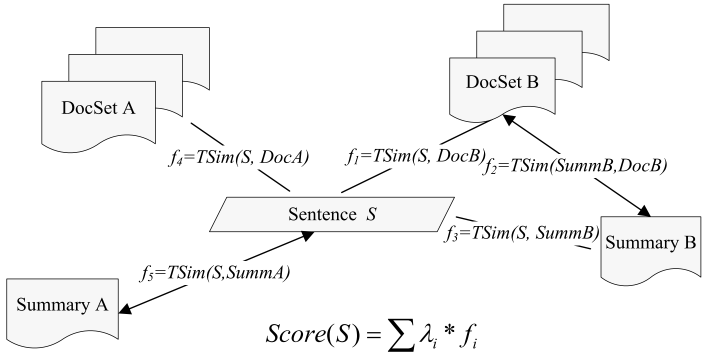

Fig. 4. The dynamic model for update summarization.

using the DUC 2007 English evaluation corpus as the development data. For details, please refer to Section 5.1.

5. EXPERIMENTATION ON ENGLISH MDS

In this article, we have systematically evaluated our unified framework on TAC 2008 for both standard and update summary generation. Besides, we also report the performance on TAC 2009.

5.1. Experimental Setting

The update summarization task at TAC 2008 consisted of two kinds of summaries: a history summary and an update summary. Here the history summary is a standard summary. The documents for TAC 2008 are retrieved from the AQUAINT-2 collection of newswire articles. There are 48 document sets, each of which contains 20 AQUAINT-2 documents. Besides, the retrieved documents in each document set describe the same theme and are ordered chronologically and divided into two sets of 10 documents each, such that Set B follows Set A in the temporal order, with Set A for standard summarization and Set B for update summarization. The task is to produce a 100-word summary for each set, guided by a statement describing the reader's need for information (not explored in this article). Therefore, a total of 96 summaries will be produced, half of them as standard summaries and the other half as update summaries. Similar to TAC 2008, TAC 2009 is also an update summarization task and includes 44 document sets, each of which contains 20 documents. An example statement, including a title and a narrative, is shown here.

> **num:** D0842G
> **title:** Natural Gas Pipeline
> **narrative:** Follow the progress of pipelines being built to move natural gas from Asia to Europe. Include any problems encountered and implications resulting from the pipeline construction.

For evaluation, we use the ROUGE toolkit [Lin and Hovy 2003] provided by TAC 2008[3]. Rouge-1 (recall against unigrams), Rouge-2 (recall against bigrams) and Rouge-SU4 (recall against skip-4 bigrams) scores at the 95% confidence level are computed by running ROUGE-1.5.5, where Rouge-2 and Rouge-SU4 are automatic ROUGE evaluation scores in TAC 2008. In order to compare our model with others, we extract a 100-word summary as required by TAC 2008 and TAC 2009.

[3] See http://www.nist.gov/tac/data/index.html.

In all the experiments, all the documents are pre-processed using an in-house, state-of-the-art sentence segmentation and word tokenization toolkit. Besides, stop words are removed using a popular stop word list[4] and all the remaining words are stemmed using the Porter stemmer[5]. For model development, we utilize the DUC 2007 English evaluation corpus (including 45 document sets each of which contains 25 documents) as the development data to fine-tune all the free parameters, including the number of topics K, coefficients ws in Equation (6) for adjusting the salience score via linear interpolation with the popular feature scores, and coefficients λs in the dynamic models for both standard and update summarization.

In particular, the number of topics K is fine-tuned to the number of document sets in the given corpus for the generic topic model and the number of documents in the particular document set for the specific topic model[6], respectively, using the DUC 2007 English evaluation corpus as the development data. This is consistent with our intuition that there should be at least one particular topic for each document set in the generic topic model and each document in the specific topic model since each of them should have its own specialty. For MDS, it is reasonable to set the number of topics according to the number of document sets for the generic topic model and the number of documents in the particular document set for the specific topic model[7].

5.2. Experimental Results on TAC 2008

Table I presents various Rouge scores using different models on TAC 2008 Set A (standard summarization). It shows that the specific topic model performs better than the generic topic model. It indicates that the obtained topic probability distribution is more accurate when LDA is applied on a small-scale highly-relevant data. This may be related to the structure of the corpus, which contains 48 document sets with 20 documents each. Each document set has a central topic and thus the entire corpus has 48 central topics. When applying LDA on each document set, the derived topic probability distribution may better reflect the natural distribution of all the topics for the given document set. As expected, Table I shows that the dynamic model outperforms the static model due to redundancy control. Since our framework performs better when applying LDA on each document set using the dynamic model, we use the specific topic model and the dynamic model by default in all the following experiments.

Tables II and III compare various static and dynamic models in redundancy control on Set A and Set B. Here, “standard” (for Set B) means that we treat the update summarization task as the standard summarization task without considering the history documents or summary. Moreover, the content in parentheses indicates the exact set of topic-driven similarities considered in the corresponding model. As expected, it shows that the dynamic model also works well for update summarization. It also shows that the dynamic update model with history documents much outperforms the one with history summary. From the statistical point of view, a summary always contains only

[4] See http://www.lextek.com/manuals/onix/stopwords1.html.

[5] See http://drupal.org/project/porterstemmer.

[6] For MDS, the given document set (e.g., the one in the TAC corpora) for generating a summary is normally highly related with similar topics. That is why the amount of data is sufficient to derive the topics in the specific topic model. Besides, we can suppose that for MDS, each document in the given set has a different sub-topic. From this respect, the goal of LDA is to differentiate these sub-topics in the given document set from each other. Of course, it would be beneficial to retrieve more relevant data from other resources (e.g., web) for better topic models. We will explore this extension in the near future.

[7] For the generic topic model to work with regard to K, a reasonable number of document sets is necessary. Besides, this may not work for the specific topic model when there are too few documents in a document set. In this case, we may segment a document into several disjoint portions using a text segmentation toolkit, such as TextTiling, and count the number of disjoint portions derived from the document set as K.

Table I. Performance on TAC 2008 Set A Using Different Models

	Rouge-1	Rouge-2	Rouge-SU4
Generic Topic Model			
Static	0.34713	0.08067	0.12386
Dynamic	0.35673	0.08266	0.12818
Specific Topic Model			
Static	0.36227	0.09258	0.12864
Dynamic	0.36991	0.09610	0.13316

Table II. Performance of Redundancy Control on TAC 2008 Set A (Standard Summarization)

Model	Rouge-1	Rouge-2	Rouge-SU4
static standard (f_1)	0.36227	0.09258	0.12864
static standard (f_2)	0.36781	0.09534	0.13129
static standard (f_3)	0.36325	0.09347	0.12946
dynamic standard (f_1, f_2, f_3)	0.36991	0.09610	0.13316

Table III. Performance of Redundancy Control on TAC 2008 Set B (Update Summarization)

Model	Rouge-1	Rouge-2	Rouge-SU4
static standard (f_1)	0.36176	0.10023	0.13393
static standard (f_2)	0.36132	0.10113	0.13454
static standard (f_3)	0.36434	0.10205	0.13548
dynamic standard (f_1, f_2, f_3)	0.36459	0.10254	0.13571
dynamic update with history summary (f_1, f_2, f_3, f_4)	0.36462	0.10263	0.13583
dynamic update with history documents (f_1, f_2, f_3, f_5)	0.36603	0.10334	0.13707
dynamic update with history summary and documents (f_1, f_2, f_3, f_4, f_5)	0.36739	0.10412	0.13798

part of information in a set of documents. Moreover, it shows that considering both history summary and documents further improves the performance. In all the following experiments on update summarization, we employ the dynamic model with history summary and documents by default.

In order to better evaluate the advantage of the dynamic model in redundancy control on standard summarization, we compare it with a state-of-the-art statistical model, which uses various statistical characteristics of the sentence with respect to sentences already included in the summary to avoid repetition [Larkey et al. 2003]. Here, two features are used: one is the number of words repeated and another is the cosine similarity of TF-IDF between two sentences. Table IV compares the statistical model with our dynamic model in removing the redundancy of the summary. It shows that, while the statistical model only slightly improves the performance, the dynamic model much outperforms the statistical one. It is also interesting to notice that the statistical model fails to complement the dynamic one via linear interpolation. It may be due to that the two features employed in the statistical model have been captured by the dynamic model in a better (more systematic) way.

Figure 5 evaluates the performance when different numbers of words are considered in the summary. As expected, all the Rouge scores increase monotonously as the summary length increases and almost double when increasing the summary length from 100 words to 400 words. From the statistical point of view, the more the number

Table IV. Comparison of Statistical and Dynamic Models in Redundancy Control on TAC 2008 Set A (Standard Summarization)

	Rouge-1	Rouge-2	Rouge-SU4
Statistical	0.36283	0.09304	0.12882
Dynamic	0.36991	0.09607	0.13316
Dynamic+ statistical	0.36986	0.09603	0.13305

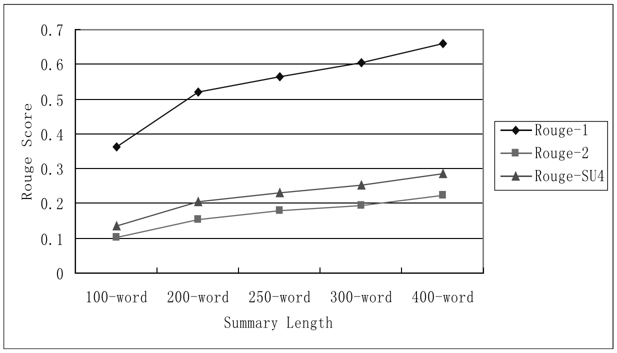

Fig. 5. Learning curves of various Rouge scores over summary lengths in words on TAC 2008 Set B (update summarization).

Table V. Contribution of the Popular Features and the Topic-Driven Similarity on TAC 2008 Set A (Above, Standard Summarization) and Set B (Below, Update Summarization)

	Rouge-1	Rouge-2	Rouge-SU4
popular features	0.34503	0.08101	0.11771
	0.34007	0.08484	0.11999
topic-driven similarity	0.36682	0.09294	0.13055
	0.34856	0.08169	0.12107
linear interpolation	0.36991	0.09610	0.13316
	0.36739	0.10412	0.13798

of words included in a summary, the more similar the probability distribution of a summary to that of documents over the derived topics is expected.

In the preceding experiment, we adjust the salience score by linearly interpolating the topic-driven similarity score with the popular feature scores, as shown in Equation (6). Table V examine the complementary nature of the topic-driven similarity and the popular features. It shows that, although the popular features contributes much less than the topic-driven similarity on most experiments, they are much complementary to each other and linear interpolation of the topic-driven similarity score with the popular feature scores much improves the performance in all experiments. This is due to the failure of the topic-driven similarity to cover such popular features.

Finally, Table VI compares several widely-used similarity measures. It shows that the KL-divergence performs best while the cosine measure performs worst.

5.3. Experimental Results on TAC 2009

Table VII presents the performance of various static and dynamic models on the TAC 2009 update summarization task, which is consistent with our experiments on the

Table VI. Comparison of Various Similarity Measures on TAC 2008 on Set A (Above, Standard Summarization) and Set B (Below, Update Summarization)

	Rouge-1	Rouge-2	Rouge-SU4
KL	0.36991	0.96070	0.13316
	0.36739	0.10412	0.13798
Cosine	0.33494	0.07438	0.11093
	0.33088	0.07958	0.11627
JS	0.36378	0.09463	0.12989
	0.36336	0.10054	0.13515

Table VII. Performance on TAC 2009

	Rouge-1	Rouge-2	Rouge-SU4
TAC 2009 Set A			
static standard (f_1)	0.37036	0.09679	0.13417
static standard (f_2)	0.37891	0.10563	0.14234
static standard (f_3)	0.37827	0.10565	0.14216
dynamic standard (f_1, f_2, f_3)	0.38681	0.10984	0.14658
TAC 2009 Set B			
static standard (f_1)	0.35513	0.08426	0.12577
static standard (f_2)	0.36282	0.09347	0.13362
static standard (f_3)	0.36155	0.09174	0.13268
dynamic standard (f_1, f_2, f_3)	0.36334	0.09369	0.13382
dynamic update with history summary (f_1, f_2, f_3, f_4)	0.36283	0.09376	0.13244
dynamic update with history documents (f_1, f_2, f_3, f_5)	0.36362	0.09452	0.13406
dynamic update with history summary and documents (f_1, f_2, f_3, f_4, f_5)	0.36428	0.09583	0.13534

Table VIII. Comparison of our System with the Best-Reported Ones on the TAC 2008 and 2009 Update Summarization Tasks (Set B). Please Note that the Organizers Only Report Rouge-2 and Rouge-SU4 Scores on Both Tasks

	Rouge-2	Rouge-SU4
TAC 2008		
Gillick et al. (2008)	0.10381	0.13625
Ours	0.10412	0.13798
TAC 2009		
Gillick et al. (2009)	0.10386	0.13948
Ours	0.09473	0.13354

TAC 2008 update summarization task. This suggests the effectiveness of the dynamic model in both standard and update summarization.

5.4. Comparison with Related Work

Table VIII compares our system with the best-reported systems on the TAC 2008 and 2009 update summarization tasks.

Compared to the official experimental results (Rouge-2 and Rouge-SU4) published by the TAC 2008 organizers [Dang and Owczarzak 2008], our system outperforms the best system on the TAC 2008 update summarization task [Gillick et al. 2008] (0.10412 vs. 10381 in Rouge-2, 0.13798 vs. 0.13625 in Rouge-SU4 score). Although our system performs worse than the best system on the TAC 2009 update summarization task [Gillick et al. 2009], our system is much simpler and thus much more flexible for further improvement.

However, the performance gap is still large compared to human beings. One major reason may be that, although we consider the sentence length, our approach still prefers long sentences. This raises the necessity of effective summary compression, which is worth exploring in the future. Actually, Gillick et al. [2008, 2009] much benefit from a sentence compression module in post-processing.

6. EXPERIMENTATION ON CHINESE MDS

Since our proposed MDS framework is purely statistical and language-independent, it can be easily applied to MDS in Chinese language. However, due to the lack of annotated corpora on update summarization in Chinese language, we only focus on standard summarization in Chinese language.

6.1. Experimental Setting

Currently, there are few publically available corpora on Chinese MDS. For fair comparison, we use the corpus described in Xu et al. [2007]. Actually, this is the only publically available corpus with reported performance in Chinese MDS. This corpus comes from online news reports with topics covering sports, economic, emergency, etc., and contains 19 document sets, each of which includes 5–10 documents.

In particular, golden summary sentences in each document set are marked as well as candidate sentences which could replace summary sentences but couldn't be concurrent with those golden summary sentences. Moreover, each candidate sentence is given a corresponding weight value between (0, 1) according to its degree of replacement. Based on such annotation, three criteria are used to measure the quality of the evaluation summary: precision (P), redundancy (R) and total quality (T).

$$precision = \left(\sum\nolimits_{i=1}^{k_1} \omega_i\right)/K$$

$$redundancy = \left(\sum\nolimits_{i=1}^{k_1} \sum\nolimits_{j=i+1}^{k_1} \phi\left(s_i, s_j\right)\right)/K \tag{8}$$

$$total = precision - redundancy$$

Where K is the total number of sentences in the evaluation summary, k_1 is the number of golden summary sentences in the evaluation summary; $(\omega_1, \omega_2, ..., \omega_k)$ is the weight of the sentence (annotated manually in the corpus); and $\phi(s_i, s_j)$ is a binary discrimination function, 1 if s_i and s_j are the same class summary sentences, 0 otherwise.

In this article, we adopt the ICTCLAS 2009 system[8] for word segmentation. Due to the failure to having another publically available Chinese MDS corpus as the development data to fine-tune the free parameters, we simply adopt the experimental setting fine-tuned using the DUC 2007 English evaluation corpus as the development data.

[8] See http://ictclas.org.

Table IX. Performance of our MDS Framework

Summary length (sentence number)	Static model			Dynamic model		
	P (%)	R	T (%)	P (%)	R	T (%)
5	68.73	4.38	64.35	76.84	1.25	75.59
10	71.89	10.46	61.43	74.47	5.94	68.53
20	76.16	9.75	66.41	78.90	6.67	72.23

Table X. Performance of the Upper Bound System and MDF, Both Described in Xu et al. [2007]

Summary length (sentence number)	upper bound system			MDF		
	P (%)	R	T (%)	P (%)	R	T (%)
5	88.125	0[9]	88.125	70.31	0	70.31
10	90.625	4.68	85.945	68.13	9.37	58.76
20	86.17	5.94	80.23	72.66	7.81	64.85

6.2. Experimental Results

Table IX presents the performance of our MDS framework. It shows that the dynamic model always outperforms the static model. This indicates the effectiveness of the dynamic model in Chinese summarization, similar to English summarization. It also shows the effectiveness of the dynamic model in removing the redundancy of Chinese summarization.

Table X gives the performance of an upper bound system and a state-of-the-art multi-document framework (MDF), both described in Xu et al. [2007], on the same Chinese data. In the upper bound system, the summary is extracted based on some information marked by hand, while all of the information in MDF is automatically generated. Comparison of Table IX and Table X shows that our MDS framework (using the dynamic model) much outperforms MDF.

7. CONCLUSION

In this study, we present a unified framework for both standard and update summarization, which adopts a topic modeling approach for salience determination and a dynamic modeling approach for redundancy control. In particular, the topic modeling approach is based on the probability distribution over the derived topics, which applies to various kinds of text units, such as word, sentence, document, multi-documents and summary. The dynamic modeling approach considers both the similarity between the sentence and given documents, the similarity between the sentence and the summary, besides the similarity between the summary and given documents, for standard summarization, and is further extended to consider the similarity between the sentence and the history documents or summary, for update summarization. Evaluation on TAC 2008 and 2009 shows promising results. In addition, we also evaluate our framework to Chinese summarization. Evaluation shows the effectiveness of our framework on Chinese summarization. In particular, the dynamic modeling approach is very effective at redundancy control on both English and Chinese summarization.

The main contribution of this article lies in the dynamic modeling approach for redundancy control in MDS, which is simple but effective.

[9] These redundancy figures are problematic since such figures should NOT be zero in practice. We have contacted the authors for this issue several times, but no reply. However, we include these figures here just for reference because it's the only reported work that we can find on this Chinese MDS corpus.

In the future, we will explore better ways in integrating various similarity scores between different kinds of text units. In addition, we will explore summary compression to further improve the quality of both English and Chinese MDS.

REFERENCES

ALLAN, J., WADE, C., AND BOLIVA, A. R. 2003. Retrieval and novelty detection at the sentence level. In *Proceedings of the 26th Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR'03)*. 314–321.

ARORA, R. AND RAVINDRAN, B. 2008a. Latent Dirichlet allocation based multi-document summarization. In *Proceedings of the 2nd Workshop on Analytics for Noisy Unstructured Text Data (ANUTD'08)*. 91–97.

ARORA, R. AND RAVINDRAN, B. 2008b. Latent Dirichlet Allocation and Singular Value Decomposition-Based Multi-Document Summarization. In *Proceedings of the International Conference on Data Mining (ICDM'08)*. 713–718.

BHANDARI, H., SHIMBO, M., ITO, T., AND MATSUMOTO, Y. 2008. Generic text summarization using probabilistic latent semantic indexing. In *Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP'08)*. 133–140.

BLEI, D. M., NG, A. Y., AND JORDAN, M. I. 2003. Latent Dirichlet allocation. *J. Mach. Learn. Res. 3*, 993–1022.

BRIN, S. AND PAGE, L. 1998. The anatomy of a large-scale hypertextual Web search engine. *Comp. Netw. 30*, 1–7, 107–117.

CARBONELL, J. AND GOLDSTEIN, J. 1998. Use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR'98)*. 335–336.

DANG, H. T. AND OWCZARZAK, K. 2008. Overview of the TAC 2008 update summarization task. In *Proceedings of the 1st Text Analysis Conference (TAC'08)*.

EDMUNDSON, H. P. 1969. New methods in automatic extracting. *J. ACM 16*, 2, 264–285.

ERKAN, G. AND RADEV, D. R. 2004. LexPageRank: Prestige in multi-document text summarization. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'04)*. 365–371.

GILLICK, D., FAVRE, B., AND HAKKANI-TUR, D. 2008. The ICSI summarization system at TAC 2008. In *Proceedings of the 1st Text Analysis Conference (TAC'08)*.

GILLICK, D., FAVRE, B., HAKKANI-TUR, D., BOHNET, B., LIU, Y., AND XIE, S. 2009. The ICSI/UTD summarization system at TAC 2009. *In Proceedings of the 2nd Text Analysis Conference (TAC'09)*.

HAGHIGHI, A. AND VANDERWENDE, L. 2009. Exploring content models for multi-document summarization. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (ACL'09)*. 362–370.

JONES, K. 1999. Automatic summarizing: Factors and directions. In *Advances in Automatic Text Summarization,* MIT Press, 1–12.

JONES, K. 2007. Automatic summarizing: The state of the art. *Inf. Proc. Man. 43*, 6, 1449–1481.

KLEINBERG, J. AND AUTHORITATIVE, M. 1998. Sources in a hyperlinked environment. In *Proceedings of the Annual ACM-SIAM Symposium on Discrete Algorithms (SIAM'98)*. 668–677.

KULLBACK, S. AND LEIBLER, R. A. 1951. On information and sufficiency. *Annals Math. Stat. 22*, 1, 79–86.

LARKEY, L. S., ALLAN, J., CONNELL, M. E., BOLIVAR, A., AND WADE, C. 2003. UMass at TREC 2002: Cross Language and Novelty Tracks. *Nat. Inst. Stand. Tech.* 721–732.

LIN, C. Y. AND HOVY, E. H. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the Language Technology Conference (HLT-NAACL'03)*.

LIU, D., WANG, Y., LIU, C., AND WANG, Z. 2006. Multiple documents summarization based on genetic algorithm. *Fuzzy System and Knowledge Discovery*, Lecture Notes in Computer Science, vol. 4223, 355–364.

MIHALCEA, R. 2005. Language independent extractive summarization. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions (ACL'05)*. 49–52.

MANI, I. AND BLOEDORN, E. 1999. Summarizing similarities and differences among related documents. *Inf. Retriev. 1*, 1, 35–67.

NASTASE, V. 2008. Topic-driven multi-document summarization with encyclopedic knowledge and spreading activation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'08)*. 763–772.

PARK, S., LEE, J. H., AHN, C. M., HONG, J. S., AND CHUN, S. J. 2006. Query based summarization using non-negative matrix factorization. In *Proceeding of International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES'06)*. 84–89.

RADEV, D. R., JING, H., AND BUDZIKOWSKA, M. 2000. Centroid-based summarization of multiple documents: Sentence extraction, utility-based evaluation, and user studies. In *Proceedings of the ANLP-NAACL Workshop on Summarization (ANLP-NAACL'00)*.

RADEV, D. R., JING, H., AND BUDZIKOWSKA, M. 2001. Experiments in single and multiple documents summarization using MEAD. In *Proceedings of the Document Understanding Conference (DUC'01)*.

STEINBERGER, J. AND JEZEK, K. 2004. Using latent semantic analysis in text summarization and summary evaluation. In *Proceedings of ISIM'04*. 93–100.

TORRALBO, R., ALFONSECA, E., GUIRAO, J. M., AND MORENO-SANDOVAL, A. 2005. Description of the UAM system at DUC-2005. In *Proceedings of the Document Understanding Conference Workshop at HLT/EMNLP 2005 (HLT/EMNLP'05)*.

VARADARAJAN, R. AND HRISTIDIS, V. 2006. A system for query-specific document summarization. In *Proceedings of the 15th ACM International Conference and Information and Knowledge Management (CIKM'06)*. 622-631.

WANG, D., ZHU, S., LI, T., AND GONG, Y. 2009. Multi-document summarization using sentence-based topic models. In *Proceedings of the International Joint Conference on Natural Language Processing Conference Short Paper (INCNLP'09)*. 297–300.

XU, Y. D., XU, Z. M., AND WANG, X. L. 2007. Multi-document automatic summarization technique based on information fusion. *Chin. J. Comp. 30, 11,* 2048–2054.

Received November 2010; revised February 2011, June 2011; accepted July 2011