A New Re-ranking Method for Generic Chinese Text Summarization and Its Evaluation

Xiaojun Wan and Yuxin Peng*

Institute of Computer Science and Technology,
Peking University, Beijing 100871, China
{wanxiaojun, pengyuxin}@icst.pku.edu.cn

Abstract. In this paper a new EMD-MMR (EMD: earth mover's distance; MMR: maximal marginal relevance) re-ranking method is proposed for generic Chinese text summarization. Our extraction-based summarization approach first ranks the sentences in a document by their weight calculated based on word frequency and position, and then re-ranks a few highly weighted sentences by the EMD-MMR method for sentence extraction. The proposed re-ranking method adopts a novel EMD-based similarity metric instead of the Cosine metric into the MMR approach. The EMD-based similarity metric can naturally take into account the semantic relatedness between words and compute the semantic similarity between texts with a many-to-many matching among words. We evaluate the performance of the proposed approach with a novel *nk-blind* method and the results demonstrate its effectiveness.

1 Introduction

Automated generic text summarization has drawn much attention in recent years and a generic summary should contain the main topics of the document while keeping redundancy to a minimum. The summarization methods can be categorized into two categories: extraction-based methods and abstraction-based methods. Extraction is much easier than abstraction because extraction is just to select existing sentences while abstraction needs understanding and rewriting sentences. Extraction-based methods usually assign each sentence a score and then rank the sentences in the document. Statistical and linguistic features, including word frequency, position, cue words, stigma words, topic signature, etc., have been employed for scoring sentence.

Our summarization approach takes two steps to extract summary sentences. In the first step, sentences are ranked by their weight calculated based on word frequency and position, and then a few salient sentences (10 sentences in the experiments) are reserved as candidate sentences. The weight based on word frequency is computed as the sum of the *tf*idf* weights of words in the sentence. The weight based on position is computed with *1-((i-1)/n)*, where *i* is the sequence of the sentence and *n* is the total number of sentences in the document After the above weights are calculated for each sentence, we linearly combine the weights and normalize the sum by the length of the sentence to get the final score. The normalization aims to avoid favoring long sentences.

[*] Contact author.

E.A. Fox et al. (Eds.): ICADL 2005, LNCS 3815, pp. 171 – 175, 2005.

The second step is a redundancy-removing process and in this step those candidate salient sentences are re-ranked by the proposed EMD-MMR (EMD: earth mover’s distance; MMR: maximal marginal relevance) method and the summary is produced by extracting several top sentences. The EMD-MMR re-ranking method adopts a novel EMD-based similarity metric instead of the Cosine metric into the popular MMR approach [1]. The EMD-based similarity metric can naturally take into account the semantic relatedness between words and get the semantic similarity between texts with a many-to-many matching among words. The EMD-MMR method is described in detail in next section.

2 The EMD-MMR Re-ranking Method

The maximal marginal relevance (MMR) method strives to maximize relevant novelty in summarization. A sentence is selected into the summary as follows:

$$\mathrm{MMR} \underset{\mathrm{def}}{=} \mathrm{Arg} \max_{s_i \in D \backslash S} \left[\lambda (\mathrm{sim}_1(s_i, q) - (1-\lambda) \max_{s_j \in S} \mathrm{sim}_2(s_i, s_j)) \right], \tag{1}$$

where q is a query representation; D is the set of sentences in the document; S is the set of sentences in the summary, which is a sub set of D; $D\backslash S$ is the set difference, i.e. the set of as yet unselected sentences in D; sim_1 is the similarity metric for calculating the similarity between the query q and a sentence s_i. sim_2 is the similarity metric for calculating the similarity between two sentences s_i and s_j. λ is a weighting parameter. In the experiments, we use all the occurrences of top 50 words with the largest *tf*idf* values in the document as the query representation. The parameter λ is set to 0.7.

The similarity metrics *sim1* and *sim2* are usually the widely-used standard Cosine measure and the terms are weighted by *tf*idf* value. Texts are usually represented by a bag of words (or phrase) and then the similarity is calculated between the lists of words. In the Cosine metric, different words are usually assumed to be semantically independent and in the similarity calculation process one word in a text can only be matched to the same word in another text. However, different words could express the same or similar meanings due to the synonym phenomenon. An example of synonyms is the words “cat” and “feline”. In Chinese language, “战斗” and “战役’ represent almost the same meaning. There is also other semantic relatedness between different words, such as hypernymy/hyponymy, and all these phenomena in natural language argue that words are not independent with each other in reality. Extremely, a text containing one set of words might be semantically similar to another text containing a different set of words. The proposed EMD-based similarity metric can naturally consider the semantic relatedness between words and adopt it into the MMR method for re-ranking. We denote the MMR method with the EMD-based similarity metric as EMD-MMR.

In the EMD-based similarity metric, the semantic distance (the contrary metric of semantic relatedness) between words is required to be calculated and then EMD is employed to measure text similarity with a many-to-many matching among words. In this study, we extracts sense explanation of each word from a Chinese dictionary and builds a feature vector for the word, and then the semantic relatedness s of two words is calculated by applying the Cosine metric on the two vectors. In the feature vector, each word is weighted by *tf*idf*. The semantic distance between the two words is gotten by 1-*s*, which is between 0 and 1. The more "similar" two words are, the smaller the semantic distance is. The semantic distances of all pairs of words in the test document set are calculated beforehand. For example, the semantic distance between "战斗" and "战役" is 0.257, and 0.455 for "战斗" and "战火".

The Earth Mover's Distance (EMD) [2] is a method to evaluate dissimilarity between two multi-dimensional distributions in some feature space where a distance measure between single features, which we call the *ground distance* is given. The EMD "lifts" this distance from individual features to full distributions. Computing the EMD is based on a solution to the well-known transportation problem. In our context, the distributions are the word distributions of texts, and a weighted graph is constructed to model the similarity between two texts, and then the EMD is employed to compute the minimum cost of the weighted graph as the similarity value between two texts. The problem is formalized as follows:

In our context, the distributions are the word distributions of texts, and a weighted graph is constructed to model the similarity between two texts, and then EMD is employed to compute the minimum cost of the weighted graph as the similarity value between two texts. The problem is formalized as follows:

Given two texts *A* and *B*, a weighted graph *G* is constructed as follows:

- Let $A=\{(t_{a1},w_{a1}),(t_{a2},w_{a2}),\ldots,(t_{am},w_{am})\}$ as the representation of text *A*, t_{ai} represents a unique word in text *A* and w_{ai} is the word's *tf*idf* value.
- Let $B=\{(t_{b1},w_{b1}),(t_{b2},w_{b2}),\ldots,(t_{bn},w_{bn})\}$ as the representation of text *B*, t_{bj} represents a unique word in text *B* and w_{bj} is the word's *tf*idf* value.
- Let $D=\{d_{ij}\}$ as the distance matrix where d_{ij} is the semantic distance between words t_{ai} and t_{bj}. In our case, d_{ij} has been computed beforehand.
- Let $G=\{A, B, D\}$ as a weighted graph constructed by *A, B* and *D*. $V=A\cup B$ is the vertex set while $D=\{d_{ij}\}$ is the edge set.

In the weighted graph *G*, we want to find a flow $F=\{f_{ij}\}$, where f_{ij} is the flow between t_{ai} and t_{bj}, that minimizes the overall cost

$$WORK\quad(A,B,F)=\sum_{i=1}^{m}\sum_{j=1}^{n}f_{ij}d_{ij}\ , \tag{2}$$

subject to the following constraints:

$$f_{ij}\geq 0 \quad 1\leq i\leq m \quad 1\leq j\leq n \qquad (3) \qquad\qquad \sum_{j=1}^{n}f_{ij}\leq w_{ai} \qquad 1\leq i\leq m \qquad (4)$$

$$\sum_{i=1}^{m}f_{ij}\leq w_{bj} \qquad 1\leq j\leq n \qquad (5) \qquad\qquad \sum_{i=1}^{m}\sum_{j=1}^{n}f_{ij}=\min\left(\sum_{i=1}^{m}w_{ai},\sum_{j=1}^{n}w_{bj}\right) \qquad (6)$$

Constraint (3) allows moving words from *A* to *B* and not vice versa. Constraint (4) limits the amount of words that can be sent by the words in *A* to their weights. Constraint (5) limits the words in *B* to receive no more words than their weights. Constraint (6) forces to move the maximum amount of words possible. We call this amount the *total flow*. Once the transportation problem is solved, and we have found the optimal flow F, the earth mover's distance is defined as the resulting work normalized by the total flow:

$$EMD\ (A, B) = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij} d_{ij}}{\sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij}} \cdot \tag{7}$$

The normalization factor is introduced in order to avoid favoring shorter text in the case of partial matching. Finally, the similarity between texts A and B is defined as

$$Sim_{EMD}\ (A, B) = 1 - EMD\ (A, B) \cdot \tag{8}$$

$Sim_{EMD}(A,B)$ is normalized in the range of [0,1]. The higher the value of $Sim_{EMD}(A,B)$, the more similar the texts A and B.

The above EMD-based similarity metric allows for many-to-many matches among words according to their semantic relatedness. For example, the word "战斗" in text A can match both the words "战役' and "战火" in text B.

Efficient algorithms for the transportation problem are available, which are important to compute EMD efficiently.

4 Evaluation with the *nk-blind* Method

For Chinese document summarization, there are no gold standard data set for evaluation. So we downloaded 30 Chinese news articles from *news.sina.com.cn,* one of the most famous news portals in China, and those articles include political news, sports news and recreational news. Then eight students are employed to extract five sentences from each article and produce a summary for that article. The inter-human agreement between students is low by our analysis. So in fact there is no ideal summary for each document and we cannot evaluate system summaries based on any single annotated summary. The traditional metric for evaluating extraction-based summaries, such as precision and recall, can not be applied directly.

To resolve the issue of low inter-human agreement, we introduce a so-called *nk-blind* method which has been used to evaluate Chinese word segmentation systems [3]. This method is based on an intuitive idea of "majority win". Given a document, n human-annotated summaries (or n judges) were created independently. Then, the system-generated summary is compared against the annotated ones: for each sentence in the system-produced summary, a sentence is considered to be correct if at least k of the n human-annotated summaries contain the sentence. The precision increases with smaller k. If k=1, it is sufficient for any judge to sanction a sentence selection. If k=n, the sentence must be shared by all human-annotated summaries. Given k, the precision for each system-produced summary is calculated and then the values are averaged across all summaries. So a precision rate can be given under any chosen *(n, k)* setting under the *nk-blind* method. This result can be plotted as an *n-k* curve which is similar to *p-r* curve. We can compare two summarizers via their *n-k* curves.

In the experiments, those system-produced five-sentence summaries are compared with the human-annotated five-sentence summaries. We use an in-house tool for

Chinese word segmentation. The baseline system is a lead baseline system, which takes the first five sentences in the document as the summary.

All results reported in Figure 1 give the precision values for n=8 judges with all values of k between 1 and n. "w/o re-ranking" means that the system selects top five sentences in the candidate sentence set generated in the first step and produces the summary without the second step. "w/ re-ranking (MMR)" means that the traditional MMR re-ranking method with the Cosine metric is taken to re-rank the candidate sentences and then the summary is produced. "w/ re-ranking (EMD-MMR)" means that the EMD-MMR re-ranking method is taken to re-rank the candidate sentences. Seen from Figure 1, the lead baseline method performs worst. The re-ranking step does benefit the summarization performance in that it can remove redundancy in the summary. The EMD-MMR re-ranking method outperform the traditional MMR re-ranking method, which proves that the EMD-based similarity metric has a better ability to measure semantic similarity between texts than the Cosine metric. The many-to-many matching between words plays the key role for the performance improvement. From human's perspective, someone judges whether two texts are similar enough not by the word occurrences but by the semantic similarity between the texts.

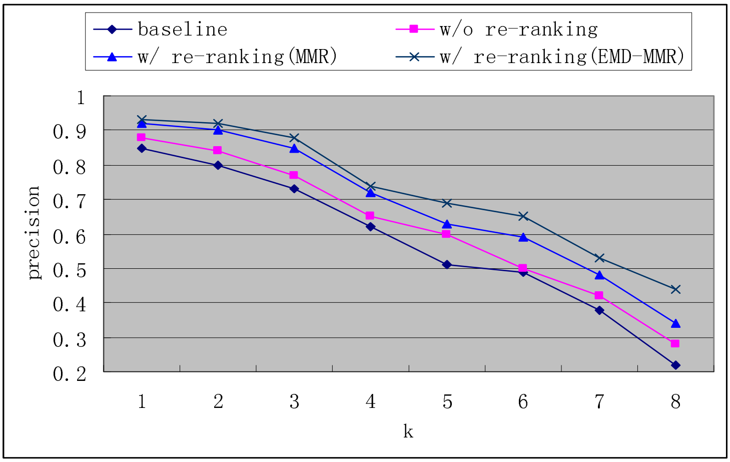

Fig. 1. Comparison of *nk-blind* precisions

References

1. Carbonell, J., Goldstein, J.: The Use of MMR, Diversity-based Reranking for Reordering Documents and Producing Summaries. In Proceedings of SIGIR'98 (1998)
2. Rubner, Y., Tomasi, C. and Guibas, L.: The Earth Mover's Distance as a Metric for Image Retrieval. *I*nt. Journal of Computer Vision **40-2** (2000) 99-121
3. Wu, D., Fung, P.: Improving Chinese Tokenization With Linguistic Filters On Statistical Lexical Acquisition. In Proceedings of ANLP'94 (1994) 180-181