Single Chinese News Article Summarization Based on Ranking Propagation

Mingrong Liu, Yicen Liu, Liang Xiang, Xing Chen, Qing Yang
National Laboratory of Pattern Recognition
Institute of Automation, Chinese Academy of Sciences
95 Zhongguancun Donglu, Beijing 100190, China
{mrliu,ycliu,lxiang,xchen,qyang}@nlpr.ia.ac.cn

Abstract

This paper proposes a news summarization system called NewsSum for simultaneous key entities and sentences extraction from single Chinese news article. NewsSum can provide both query-independent and query-specific news summarization. In this study, NewsSum is implemented by firstly parsing the news text in the preprocessing stage and building a news document graph, then exploiting the ranking propagation algorithm on the graph to extract key entities and sentences from the text as its summary. Furthermore, the quality of NewsSum is compared with state-of-the-art summarization approaches through a user survey.

1. Introduction

Automatic text summarization is an extremely active research field making connections with many other research areas such as information retrieval, natural language processing and machine learning. It has drawn much attention in recent years and it exhibits the practicability in document management and search systems. As the number of documents available on users' desktops and the Internet increases, so does the need to provide high-quality summaries in order to allow the user to quickly locate the desired information. A compelling application of document summarization is the snippets generated by Web search engines for each query result, which assist users in further exploring individual results. Besides, a summarization technique is needed in a complex question answering (Q&A) system to minimize the answer size.

The Information Retrieval (IR) community has largely viewed text documents as linear sequences of words for the purpose of summarization. Although this model has proven quite successful in efficiently answering keyword queries, it is clearly not optimal since it ignores the inherent structure in documents. Furthermore, most summarization techniques are query-independent and follow one of the following two extreme approaches: Either they simply extract relevant passages viewing the document as an unstructured set of passages, or they employ Natural Language Processing techniques. The former approach ignores the structural information of documents while the latter is too expensive for large datasets.

In this paper, we propose a single document summarization system called NewsSum [1], which combines query-independent and query-specific document summarization into one framework. NewsSum provides summarization for news articles (a large portion of information dataset) in Chinese. Figure 1 shows the system architecture of NewsSum. Firstly, at the preprocessing stage, we add structure to every news document, which can then be viewed as an undirected weighted graph, called the news document graph (as shown in Figure 2). The news document graph is built based on the query, the title of the news, and sentences and entities within the news document. The issue of how to construct the document graph will be discussed in detail in section 3.2. Thereafter, ranking propagation is performed on the graph to rank each sentence and candidate entity in the document. For each news document, its summary is made up of several top-ranking sentences and entities.

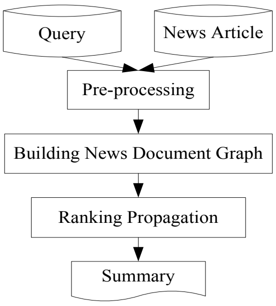

Figure 1. System Architecture.

The rest of this paper is organized as follows. Section

[1] URL: http://v.cindoo.com/system/sum.php

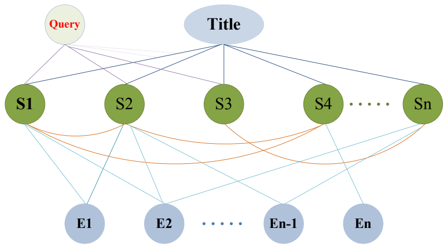

Figure 2. News Document Graph.

2 briefly describes related work on text summarization. In section 3, we deal with the ranking propagation model and develop the NewsSum summarization system. Experiments and user study issues are discussed in section 4. We conclude the paper in section 5 with pointers to future research.

2. Related Work

Document summarization is a key technique for digesting news articles and it has several approaches. Centroid-based summarization (CBS) [3, 2] uses the centroids of the clusters of news articles produced by standard single-pass clustering systems in order to extract sentences central to the topic. Maximal marginal relevance (MMR) [1] is a widely used approach for information retrieval. It ranks documents (including news documents) according to a combined criterion of query relevance and information novelty within a document. It extracts the novel sentences and creates a summary from them. The Columbia summarizer [5] integrates machine learning and statistical techniques to identify similar sentences across the input articles. In [8], the mutual reinforcement principle is employed to iteratively extract key phrases and sentences from a document.

Most recently, the graph-ranking based methods, including TextRank [6] and LexPageRank [4], have been proposed for document summarization. These methods first build a graph based on the similarity relationships between the sentences in a document and then the importance of a sentence is determined by taking into account the global information on the graph recursively, rather than relying only on the local sentence-specific information. X.J. Wan et al. [7] also utilized this spirit, they proposed a method for topic-focused multi-document summarization called Manifold ranking to produce a summary biased to a given topic or user profile.

However, almost all of these pre-mentioned methods focus on only query-specific or query-independent summarization, and furthermore, they ignore the types of the documents (i.e. they all treat each document as a bag of sentences or words). In this study, both query-independent and query-specific summarization are incorporated in the proposed summarization system, NewsSum, which takes the types of documents into account (focusing on news documents), and extracts key entities and significant sentences simultaneously to construct the summaries for news documents.

3. The Proposed NewsSum

3.1. Overview

Given a news document d with its title T, in order to generate summary for d with both key entities and significant sentences, NewsSum firstly parses d to extract some candidate entities and all sentences in d at the pre-computation stage. Then, a news document graph G is constructed based on the news title, all of the sentences and candidate entities. For query-specific summarization, the query Q is also incorporated into G. Thereafter, the proposed ranking propagation algorithm is performed on G to give ranking scores to these sentences. Further more, sentences propagate their ranking scores to candidate entities. The summary of d is generated through selecting several high-ranking sentences as significant sentences and several (three to five typically) top-ranking entities as key entities.

3.2. Building News Document Graph

There are many ways to create and assign weights to a document graph, in this section we present the specific approach to create a news document graph G.

Suppose T is the title of the news article d, and Q is the query keyword used to generate query-specific summary. Let $S = S_1, S_2, ...S_n$ be the sentences in the news article, and $E = E_1, E_2, ...E_m$ represents the extracted candidate entities. We extract all named entities *(People, Locations, Organizations, etc.)* and other nouns as candidate entities since they are essential for news articles. Since this paper focuses on summarization for news in Chinese, we have developed a Chinese text segmenting and parsing tool based on the Hidden Markov Model (HMM), which can automatically segment text into sentences and extract candidate entities.

The news document graph G is constructed as follows. The nodes in G include the following elements: T, Q (for query-specific summarization), S and E. Four types of edges are defined in G: 1. Title-Sentence: an edge between each sentence and the title, 2. Query-Sentence: an edge between each sentence and the query, 3. Sentence-Sentence: for each pair of sentence, if the similarity of them is above a threshold, add an edge between them; 4. Sentence-Entity: if S_i contains an entity E_j, add an edge between them. Notice that the first two types of edges are used to propagate topic-related weights to sentences in G.

Each edge in G is given a weight to evaluate the relationship between the two nodes connected by it. The weight with an edge whose type is Title-Sentence or Sentence-Sentence is the similarity between the two node. The similarity of two sentences (when we compute the similarity between the title and a sentence, the title is also viewed as a sentence) can be determined simply as the number of common tokens between the lexical representations of the two sentences, or it can be run through syntactic filters, which only count words of a certain syntactic category. Moreover, to avoid promoting long sentences, we are using a normalization factor, and divide the content overlap of two sentences with the length of each sentence. Formally, given two sentences S_i and S_j, with a sentence being represented by the set of N_i words that appear in the sentence: $S_i = W_1^i, W_2^i, ...W_{N_i}^i$, the similarity of S_i and S_j is defined as:

$$sim(S_i, S_j) = \frac{|W_k|W_k \in S_i \& W_k \in S_j|}{log(|S_i|) + log(|S_j|)} \tag{1}$$

The weight with an edge whose type is Query-Sentence or Sentence-Entity is computed by the method similar to the BM25 ranking model in the information retrieval literature. Suppose the query Q (an entity can be viewed as a single query word) contains key words $q_1, q_2, ...q_n$, then the similarity between Q and sentence S, $sim(Q, S)$, is defined as:

$$\sum_{i=1}^{n} ISF(q_i) \cdot \frac{f(q_i, S) \cdot (k_1 + 1)}{f(q_i, S) + k_1 \cdot (1 - b + b \cdot \frac{|S|}{avgsl})} \tag{2}$$

where $f(q_i, S)$ is the term frequency of q_i in the sentence S, $|S|$ is the length of the sentence S, and $avgsl$ is the average sentence length in the news document from which sentences are drawn. k_1 and b are free parameters, usually chosen as $k_1 = 2.0$ and $b = 0.75$, respectively. $ISF(q_i)$ is the ISF (inverse sentence frequency) weight of the query term q_i, which is computed as:

$$ISF(q_i) = log\frac{N - n(q_i) + 0.5}{n(q_i) + 0.5} \tag{3}$$

where N is the total number of sentences in the document, and $n(q_i)$ is the number of sentences containing q_i in the specific article.

Naturally, for query-independent summarization, the Query node and Query-Sentence edges are absent in G.

3.3. Ranking Propagation

In this section, all of the sentences and extracted entities in a news document will be ranked through ranking propagation on G.

When choosing sentences, we believe that sentences in the summary of a news article should have the following properties. Firstly, they should have strong relationships with the news title, as we know, for a news article, its title usually summarizes the whole text quite well since it's typically manually generated to help readers quickly get an overview of the topic of the news; Secondly, they should be good representatives of the whole sentences in the news text; Thirdly, they should be related to the query for query-specific summarization. Therefore, all the properties should be taken into account to calculate the ranking scores of the sentences.

Algorithm 1 Ranking Propagation for sentences

```
Require: News document graph $G$.
Ensure: Propagation ranking scores of sentences.
Initialization: $y_i(0) = 1, i = 1, 2, ...n$
for $t = 1, 2, ...$ do
    for $i = 1$ to $n$ do
        $y_i(t+1) = \alpha Q_i + \beta T_i + \gamma \sum_{j=1, j \neq i}^{n} \frac{sim(S_i, S_j)}{outdegree(S_j)} y_j(t)$
    endfor
    normalize $y_i, i = 1, 2, ...n$
until convergence
```

Algorithm 1 shows how sentences are ranked using the ranking propagation. In algorithm 1, $y_i(t)$ is the ranking score of the ith sentence (excluding the title) in d at the iteration time t, which is the weighted sum of three parts. Q_i is the similarity degree between the query and the ith sentence S_i; T_i is the similarity degree between the title and S_i. Q_i and T_i are used to propagate ranking score to S_i reflecting the pre-mentioned two required properties for a good summary. The third part acts similar to the voting mechanism used in web page ranking algorithm such as PageRank. That is, if a sentence is connected with many other sentences which propagate ranking scores to it, then it will get high ranking score and it's a good representative of all the sentences. $outdegree(S_j)$ is the number of sentences connecting (similar) to S_j.

To this step, we describe the process of the ranking propagation for sentences. Each sentence is assigned a ranking score of one initially. In subsequent iterations, the score of a sentence is given by the weighted combination of neighboring sentences plus two stability terms based on its relationship to the title and the query. The weighting parameters α, β and γ control the final ranking of a sentence. If we choose a large $\alpha(\beta)$, then a sentence with strong relation to the query (title) will get high ranking score. If we choose a large γ, then a sentence connected to many other sentences is preferred. It's hard to choose values for these parameters through theoretical mechanics. In our practical system, α, β and γ are empirically set to 2, 0.5 and 0.5 respectively, which we found in our experiments, can produce better summaries. At the end of each iteration, the ranking score of each sentence is normalized to be in the $[0, 1]$ interval, which makes sure the whole iteration can reach convergence.

The iteration procedure stops when the scores of sentences are stable. Due to page limitation and the aim of this paper, the proof of convergence of the iterations is not discussed here. In our experiments, there are typically less than 30 iterations. The output of algorithm 1 is the propagation-ranking score ($PropScore$) corresponding to each sentence.

The proposed NewsSum system also extracts key entities from news documents. Key entities can help people quickly get the knowledge of what are the main subjects involved in the news. Thus, ranking propagation is further performed from sentences to entities in order to rank extracted candidate entities. The basic assumption is that if an entity occurs in many high ranking sentences, it should have high ranking score. In other words, sentences propagate their ranking scores to entities contained in them. Although this spirit is simple, we have found it's effective. Formally, the ranking score of an entity E_i, $RankScore(E_i)$, is computed as:

$$\sum_{j|S_j \in Nei(E_i)} sim(E_i, S_j) \cdot PropScore(S_j) \cdot idf(E_i) \quad (4)$$

Where $Nei(E_i)$ is the set of sentences containing E_i, $idf(E_i)$ is the inverse document frequency of E_i, which is used to give penalty to an entity which appears very frequently in many documents. $idf(E_i)$ is computed in advance based on a large news corpus made of up more than 1 million articles of various news types. $PropScore(S_j)$ is the pre-computed ranking score of S_j, and $sim(E_i, S_j)$ is the similarity between S_j and an entity E_i contained in it, which can be calculated in the same way as equation (1) by taking the entity as a simple sentence.

After all the sentences propagating their ranking scores to the entities, each entity receives its final ranking score, and several entities with highest ranking scores are chosen as key entities in the summary.

3.4. Summary Generation

3.4.1 Significant Sentences

Now, after ranking propagation, each sentence S_i has its propagation ranking score $PropScore(S_i)$. Simply, sentences with high ranking scores may be chosen as the final sentences in the summary. However, there may be much redundancy among the top ranking sentences, since similar sentences tend to get similar ranking scores during the ranking propagation process. Obviously, redundancy is of no good for news summarization, which conflicts with its initial purpose. A greedy algorithm similar to [7] is applied to impose the diversity penalty and compute the final overall ranking scores of the sentences.

After the diversity penalty imposition on sentences, the overall ranking scores are obtained for all sentences, several sentences with highest ranking scores are chosen as significant sentences in the final summary.

3.4.2 Key Entities

At present, all of those key entities extracted by NewsSum are named entities or other nous. We have found that they typically make good sense in summarizing topics of news articles in our experiments.

4. Experimental Study

Evaluation of a text summarization system is not a trivial issue in the literature. In this paper, to evaluate the quality of the results provided by our approach, we have conducted a user survey. We evaluate the summaries based on their quality and size (a longer summary may carries more information but is less desirable). We rank each summarization system by assigning scores. For simplicity, scores include three grades, such as 0, 1 and 2. A ranking score of 2 (0) represents the summaries produced by that summarization system is most (least) descriptive.

4.1. Dataset

The dataset used in this survey consists of six groups of news documents. Each group is made of up five documents with the same type. The six news groups focus on sports, entertainment, politics, education, finance and social issues, respectively. All of the documents were taken from several famous Chinese newswire web sites such as Sina [2], Xinhuanet [3], etc. Numbers of sentences contained in these news documents range from 9 to 55.

4.2. Query-independent Summarization

For query-independent summarization, the proposed NewsSum is compared with two state-of-the-art summarization systems: TextRank [6] and Manifold ranking [7]. Both of them extract several sentences for a document as its summary. We have developed two summarization systems according to [6] and [7] , respectively. Manifold ranking was originally proposed for multi-document summarization, but it can also provide summary for single document with little modification. Given a news document (including the title, which is viewed as an ordinary sentence in TextRank), each summarization system produces its summary made up of three sentences. The NewsSum system also produces three key entities besides the sentences. In addition, we have developed a basic news summarization

[2] URL: http://news.sina.com.cn/

[3] URL: http://www.xinhuanet.com/newscenter/

Table 1: Evaluation of query-independent summarization

♯	Social	Politics	Entertainment
Basic	1.2	1.4	1.4
TextRank	1.4	1.6	1.2
Manifold	1.4	1.8	1.4
NewsSum	1.6	1.8	1.8
NewsSum-Entity	1.4	1.8	1.2
♯	Sports	Finance	Education
Basic	0.6	1.4	1.6
TextRank	1.2	1.6	1.6
Manifold	1.6	1.8	1.4
NewsSum	1.6	1.6	1.4
NewsSum-Entity	1.4	1.2	1.2

Table 2: Evaluation of query-specific summarization

♯	Social	Politics	Entertainment
Entity	1.4	1.6	1.4
Sentence	1.6	1.8	1.6
♯	Sports	Finance	Education
Entity	1.6	1.6	1.2
Sentence	1.6	1.4	1.4

system which extracts the first three sentences of each document as its summary. Table 1 shows the average score assigned by the users for each summarization system.

The users' evaluation results are listed in table 1, each ranking score represents the average of the evaluating scores corresponding to one of the six news type for a specific summarization system. It can be seen from table 1 that NewsSum performs almost as well as the Manifold ranking system, and they get better evaluation than TextRank and the basic system. We found that the results generated by TextRank tend to have much redundancy while NewsSum and the Manifold ranking system have little redundancy. Though it's not good enough, the basic system got not too bad evaluation result compared to the other three systems. We believe that it's due to the fact that the author of a news article usually summarizes the news at the beginning or the end of the text, however, that's not the truth all the times.

In table 1, we list the average user evaluation result for key entities extracted by NewsSum as well. The result indicates that key entities extracted by NewsSum are helpful to quickly get an overview of news topics for the users when they read the news articles.

4.3. Query-specific Summarization

For query-specific summarization, we manually extract an entity or a phrase containing one to three words as the query for each news document. The average evaluation results are shown in table 2.

5. Conclusions and Future Work

We have developed NewsSum, an automatic summarization system to simultaneously extract key entities and sentences to create summary for single news article in Chinese. We firstly create the document graph to represent the hidden semantic structure of the news document and then perform ranking propagation on this graph. Query-specific and query-independent summarization are incorporated into one model in NewsSum. The ability to automatically provide summaries of textual news material will critically aid in effective use of the Internet in order to avoid overload of information. We demonstrate with a user survey that our approach performs well compared with two state-of-the-art approaches.

At present, NewsSum only serves for news articles in Chinese, we plan to extend our work to provide language-independent news summarization. We believe that NewsSum has the portability to a new language such as English when we choose appropriate text parsing tools for that language. Furthermore, we intend to utilize more evaluation methods to evaluate the proposed summarization system.

References

[1] Carbonell, Jaime G. and Goldstein, Jade. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Research and Development in Information Retrieval*, pages 335–336, 1998.

[2] Dragomir R. Radev, Hongyan Jing, and Malgorzata Budzikowska. Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies. In *ANLP/NAACL Workshop on Automatic Summarization*, 2000.

[3] Dragomir R. Radev, S. Blair-Goldensohn, Z. Zhang, R. S. Raghavan. Newsinessence: A system for domain-independent, real-time news clustering and multi-document summarization. In *Human Language Technology Conference*, 2001.

[4] G. Erkan and D. Radev. Lexpagerank: Prestige in multi-document text summarization. In *EMNLP*, 2004.

[5] K. McKeown, R. Barzilay, D. Evans, V. Hatzivassiloglou, J. Klavans, C. Sable, B. Schiffman, and S. Sigelman. Tracking and summarizing news on a daily basis with columbias newsblaster. In *Human Language Technology Conference*, 2002.

[6] R. Mihalcea and P. Tarau. Textrank: Bringing order into texts. In *Conference on Empirical Methods in Natural Language Processing*, 2004.

[7] X. Wan, J. Yang, and J. Xiao. Manifold-ranking based topic-focused multi-document summarization. In *Proceedings of IJCAI 2007.*

[8] H. Zha. Generic summarization and keyphrase extraction using mutual reinforcement principle and sentence clustering. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 113–120. ACM Press, 2002.