ATLANTIS PRESS
Advances in Computer Science Research, volume 87
3rd International Conference on Mechatronics Engineering and Information Technology (ICMEIT 2019)

Research on Summary Sentences Extraction Oriented to Chinese Patent

Lei Wang[1, 2, a], Xueqiang Lv[1, b] and Xindong You[1, 2, c, *]

[1]Beijing Key Laboratory of Internet Culture and Digital Dissemination Research, Beijing Information Science & Technology University, Beijing 100101, China

[2]Computer School, Beijing Information Science and Technology University, Beijing 100101, China

[a]wanglei2xf@163.com, [b]lxq@bistu.edu.cn, [c, *] iccdstudents@163.com

Abstract. In this paper, it describes our system oriented to single document summarization task at Chinese patent. We treat the task as a typical document summarization based on sentence extraction and decide to formulate the task in a supervised learning to rank framework, utilizing both common sentence features including term frequency, sentence position, sentence length for generic document summarization and specially designed semantic weigh feature. Summary sentence are selected according the scores by the LTR model we trained from the patent specification. Evaluation results show that our method is indeed appropriate for this task, outperforming several baseline methods in different aspects.

Keywords: Chinese patent summarization; learning to rank; semantic weigh; word2vec.

1. Introduction

Nowadays, there are a huge number of patent document released on mass media services. It is urgently demanding to generate a short summary for a given Chinese patent specification. The short summary can help people quickly understanding the patent and deciding whether to read the full patent specification.

In this paper, the Chinese patent summarization task is considered as the ordinary extractive single document summarization. We formulate the task in a learning to rank framework, exploring both common features for document summarization and semantic features based on word2vec model during supervised learning. Our system can automatically estimate the significance of the sentence. The top relevant sentences will be extracted to generate the summary of patent.

The rest of the paper is organized as follows. Section 2 reviews related work. Section 3 describes the process of constructing patent short summary by sentence extraction. Section 4 presents experiment and results. Finally, Section 5 concludes the paper and suggests the future work.

2. Related Work

Existing research on single document summarization largely relies on extractive approaches. Various approaches exist to challenge the document summarization task, including centroid based methods, link analysis, graph-based algorithms, combinatorial optimization techniques, supervised models and regression [1]. Erkan et al [2] train a LexRank model to computes sentence importance based on the concept of eigenvector centrality in a graph representation of sentences. Ziheng Lin [3] proposed a method which combine a graph model with time-stamped and the MMR technique to extract the summary sentences. Mendoza et al [4] proposed a method of extractive single-document summarization based on genetic operators and guided local search. Supervised learning approaches have been successfully applied in single document summarization, where the training data is available or easy to build [5]. The most straightforward way is to regard the sentence extraction task as a binary classification problem. Kupiec et al. [6] developed a trainable summarization system which adopted various features and used a Bayesian classifier to learn the feature weights. Learning to rank or machine-learned ranking is the application of machine learning in information retrieval area [7]. Learning to rank combines information retrieval techniques and machine learning theory, and its goal

is to obtain a ranking model from the training set using various algorithms and ranking documents in the test set [8].

In this paper, we propose a summarization method based on learning to rank model. In our method, several kinds of features are developed to describe the sentence weight for sentence ranking. We use learning to rank to obtain the feature weight and the top ranked sentences are collected to generate the summary of patent.

3. Constructing Patent Short Summary Via Sentence Extraction

3.1 Feature Selection

3.1.1 Common Feature

We extract common features which have been widely used for generic document summarization. Three kinds of features at sentence level, including the sentence position in the patent specification, the length of the sentence, and the term frequency are introduced into our system.

3.1.2 Semantic Feature

We regard weight of each sentence based on the Word2vec model and the principle of TextRank as semantic features. This specification will use word2vec to compute sentence similarity and combine TextRank ideas to generate semantic features of each sentence. We build a prediction model based on Skip-gram [9]. The process of train is shown in Fig.1.

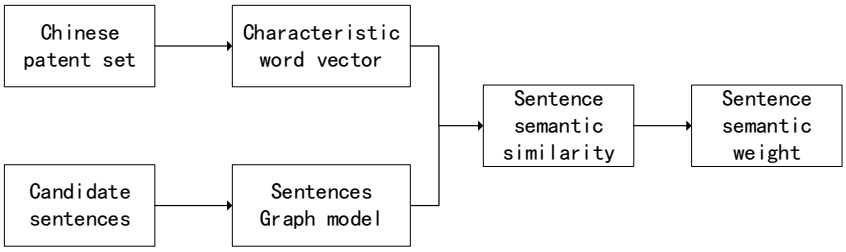

Fig. 1 The calculation process of sentence semantic weight

The cosine similarity between two sentence vectors is computed as the weight of the edge of the sentence node. The similarity is computed as the following formula (1):

$$sim(\vec{t}_1, \vec{t}_2) = \frac{\vec{t}_1 \cdot \vec{t}_2}{\left\|\vec{t}_1\right\| * \left\|\vec{t}_2\right\|} \tag{1}$$

Where $\vec{t}_1$ and $\vec{t}_2$ represent sentence vector by the word2vec model. In this step, the weights of each node are computed iteratively until the convergence is achieved by using the initial weights of nodes and the weights between sentence nodes. In order to obtain the final weight of the sentence nodes, this paper uses TextRank's improved PageRank formula to compute iteratively until the iteration converges. The formula is as follows:

$$Weight(S_i) = (1-d) + d * \sum_{S_j \in Con(S_i)} \frac{Sim(S_i, S_j)}{\left\|Con(S_{ij})\right\|} * Weight(S_j) \tag{2}$$

Where the Si represents the sentence in graph, D is the damped coefficient, generally set to 0.85. The Con (Sij) is a set of sentences that linked to Si. Ultimately, each node gets a score that reflects its semantic feature weight.

3.2 Extraction Model

In the paper, we adopted an extractive summary technique to generate a summary of Chinese. Given a patent specification including some sentences, the system will grade those sentences and select the most significant sentences. We regarded the process as a sentence ranking problem. We apply the Listwise approach learning to rank to the task of summarization. in the extractive document summarization task, we regard rouge-2 F-scores as a performance measure for evaluating summaries.

For given the sentences in Chinese patent specification, Supervised sentence scoring models based on learning to rank require input training data in the format of (x_i, y_i). There is n sentence for each specification, and there is a feature vector list $x_i = (x_1, x_2, ..., x_n)$. Each list of features x_i and the corresponding list of scores $y_i = (y_1, y_2, ..., y_n)$ then form an instance for train. Given x_i, ranking function f will calculate a relevance score $f(x_i)$. For x_i, we can get a relevance score list $z_i = (z_1, z_2, ..., z_n) = (f(x_i,1), f(x_i,2), ..., f(x_i,n_i))$. The goal of learning to rank is to minimizing the sum of the training set's loss function $G(x)$ is as follows:

$$G(x) = \sum_{i=1}^{n} Loss(y_i, z_i) \tag{3}$$

To obtain the objective function, this paper adopts a gradient descent algorithm to optimize loss function. In the Listwise approach the method named ListNet. We develop our summarization system based on Lambda Mart [10], implemented in RankLib, a typical Listwise learning to rank method. Our modification of Lambda Mart is inspired by adopting cost sensitive loss function to differentiate sentence from the Chinese patent.

4. Experiment and Results

4.1 Data Preparation

In the training data provided by EAST LINDEN (http://eastlinden.com/ch/index.aspx), there are 10000 Chinese patents with summary. Among those data, we select 8000 patents having the description summary to train the LTR model. We used the model to test the data set containing 2000 patent specifications.

4.2 Result and Analysis

We use the ROUGE metric F-scores in ROUGE-1, ROUGE-2 and ROUGE-SU4 to evaluate the result of this experiment comprehensively.

To verify the effect of the semantic weigh feature, we conducted some comparative experiment while other experimental parameters being unchanged. The extraction results are shown in Table 1:

Table 1. Comparative results of different feature

Feature	ROUGE-1	ROUGE-2	ROUGE-SU4
TF+Len+Pos	0.46852	0.35765	0.19803
TF+Len+Pos+Weight	0.48741	0.36854	0.20317

As seen in Table1, we can find that the system can effectively improve with the semantic weigh feature. We conducted some experiment on different methods and results are shown in Table 2:

Table 2. Comparative results of different methods

Method	ROUGE-1	ROUGE-2	ROUGE-SU4
TextRank	0.45408	0.34652	0.17603
First sentence	0.47952	0.36983	0.19873
Learning to rank	0.48741	0.36854	0.20317

From the results, our system based on learning to rank show statistically significant improvements over general-purpose summarization based on TextRank. It is a remarkable fact that the first sentence method has some effect on the task. For Chinese patent the first sentence of specification usually is the thesis statement of the patent.

Through the above comparing experiments, it can be concluded that the method based on learning to rank model has made the good performance compared with other methods in the field of Chinese patent summarization.

5. Conclusions and Future

In this paper, we utilize semantic weigh while also keep common features such as term frequency, length and position. Experimental results show that our framework indeed outperforms other summarization baselines, while still having much room for improvement. Some research on the abstractive summarization can generate the patent title via Recurrent Neural Networks. We will do some exploration on the area in future work.

Acknowledgements

This work is supported by National Natural Science Foundation of China under Grants No. 61671070, National Language Committee of China under Grants ZDI135-53, and Project of Cycle Economy and Knowledge Management Based on Big Data in Promoting the Developing University Intension-Disciplinary Cluster No. 5111823517.

References

[1]. Zhang J, Yao J, Wan X: Towards Constructing Sports News from Live Text Commentary. In: 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers, v3(2016), p. 1361-1371.

[2]. Günes Erkan and Dragomir R Radev. Lexrank: Graph-based lexical centrality as salience in text summarization. Journal of Artificial Intelligence Research, 2004, p. 457 – 479.

[3]. Lin Z, Chua T S, Kan M Y, et al. NUS at DUC 2007: Using evolutionary models of text[C]//Proceedings of Document Understanding Conference (DUC), (2007).

[4]. Mendoza, M., Bonilla, S., Noguera, C., et al. Extractive single-document summarization based on genetic operators and guided local search. Expert Systems with Applications 41(9), 2014, p. 4158 – 4169.

[5]. Chao Shen and Tao Li. Learning to rank for query-focused multi-document summarization. In Data Mining (ICDM), 2011 IEEE 11th International Conference on, p. 626 – 634.

[6]. J. Kupiec, J. Pedersen, and F. Chen. “A trainable document summarizer” in Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 1995, pp. 68 – 73.

[7]. Yue Shang, Huihui Hao, Jiajin Wu, et al. Learning to rank-based gene summary extraction. From IEEE International Conference on Bioinformatics and Biomedicine, (2013).

[8]. Cao Z, Qin T, Liu TY, et al. Learning to rank: from pairwise approach to listwise approach. Proceedings of the 24th International Conference on Machine Learning, 2007, p. 129-136.

[9]. Mikolov T, Dean J. Distributed representations of words and phrases and their compositionality [J]. Advances in neural information processing systems, (2013).

[10]. Qiang Wu, Christopher JC Burges, Krysta M Svore, and Jianfeng Gao. Adapting boosting for information retrieval measures. Information Retrieval, 2010, 13(3):254 – 270.