Keyphrase Extraction Based on Prior Knowledge

Guoxiu He, Junwei Fang, Haoran Cui, Chuan Wu, Wei Lu
School of Information Management, Wuhan University
Wuhan, China
{guoxiu.he,junwei.fang,haoran.cui,wu.chuan,weilu}@whu.edu.cn

ABSTRACT

Keyphrase is an important way to quickly get the topic of a document by providing highly-summative information. The previous approaches for keyphrase extraction simply rank keyphrases according to statistics-based model or graph-based model, which ignore the influence of external knowledge. In this paper, we take prior knowledge, which contains controlled vocabulary of keyphrases and their prior probability, into consideration to enhance previous methods. First, we build a controlled vocabulary of keyphrases introduced by keyphrases from existing collections and a keyphrase candidate set is filtered from a given document by it. Then, we use prior probability to represent the importance of keyphrases candidate with TF-IDF and TextRank. Finally, a supervised learning algorithm is used to learn optimal weights of these three features. Experiments on four benchmark datasets show the great advantages of prior knowledge on keyphrase extraction. Furthermore, we achieve competitive performance compared with the state-of-the art methods.

CCS CONCEPTS

• **Information systems** → Information extraction; • **Computing methodologies** → Supervised learning by classification;

KEYWORDS

Keyphrase Extraction; Prior Knowledge; TF-IDF; TextRank; Supervised Learning Algorithm

ACM Reference Format:
Guoxiu He, Junwei Fang, Haoran Cui, Chuan Wu, Wei Lu. 2018. Keyphrase Extraction Based on Prior Knowledge. In *JCDL '18: The 18th ACM/IEEE Joint Conference on Digital Libraries, June 3–7, 2018, Fort Worth, TX, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3197026.3203869

1 INTRODUCTION

Nowadays, keyphrase extraction for documents becomes a great demand in automatically understanding the topic of academic literature which generally includes two steps: keyphrases candidate selection and keyphrases candidate ranking[2]. The first step usually uses some enlightening rules such as n-grams or noun phrases with certain part-of-speech patterns to identify potential candidates. The second step is to rank the keyphrases candidate based on their importance. Either supervised or unsupervised machine learning methods with a set of manually-defined features are used for ranking.

Existing keyphrase extraction methods consider the importance of keyphrase candidate through the frequency of keyphrases or co-occurrence relation of keyphrases within the documents only, which ignore the influence of existing knowledge of keyphrases and development trend of documents in specific domains. Ideally, when annotating keyphrases, you should consider keyphrases that have already been used for documents and are always used by specific domain.

Hence, we extract controlled vocabulary of keyphrases and their prior probability as prior knowledge and then use a supervised learning algorithm to learn optimal weights for features which are TF-IDF, TextRank and prior probability.

2 METHODOLOGY

Given a collection of N samples, the $i-$th sample (d_i, K_i) contains one document d_i and M_i keyphrases $K_i = \{k_{i,1}, k_{i,2}, \cdots, k_{i,M_i}\}$. Both the document d_i and keyphrase $k_{i,j}$ are sequence of words represented as $d_i = \{w_{i,1}, w_{i,2}, \cdots, w_{i,L_{d_i}}\}$ and $k_{i,j} = \{w_{i,j,1}, w_{i,j,2}, \cdots, w_{i,j,L_{k_{i,j}}}\}$, where L_{d_i} and $L_{k_{i,j}}$ denote the length of word sequence of d_i and $k_{i,j}$ respectively and w represents a word.

2.1 Keyphrase Candidate Selection

Building a controlled vocabulary of keyphrases is an important part in this work. We collect existing keyphrases from document collections and get rid of duplicates to get controlled vocabulary represented as $KV = \{k_1, k_2, \cdots, k_O\}$, where k_i is a keyphrase and O is the vocabulary size.

In addition, this paper quantifies the use of keyphrases as a prior probability of candidate keyphrases and uses them as external knowledge in keyphrase extraction. For the detailed calculation process of this probability, see the section of keyphrase candidate ranking.

The keyphrases candidate C_i of a document d_i is selected by the controlled vocabulary of keyphrases. That is, given a document d_i, candidate keyphrases are the largest match pattern of word sequence according to the controlled vocabulary of keyphrase.

2.2 Candidate Keyphrases Ranking

In order to rank keyprhases candidate, we extract prior probability combined by TF-IDF and TextRank and use a supervised learning algorithm to learn the optimal weights of them.

2.2.1 Feature Extraction. Based on keyphrases candidate C_i of document d_i, keyphrases in keyphrases candidate set need to be scored and ranked accordingly. Our method considers the importance of candidate keyphrases based on three features in a document d_i i.e. TF-IDF (TF), TextRank (TR) and Prior Probability (PP).

JCDL '18, June 3–7, 2018, Fort Worth, TX, USA

ACM ISBN 978-1-4503-5178-2/18/06.
https://doi.org/10.1145/3197026.3203869

Table 1: The Result of Experiment

Method	Inspec		Krapivin		Nus		Ke20K	
	F1@5	F1@10	F1@5	F1@10	F1@5	F1@10	F1@5	F1@10
TF-IDF[5]	0.221	0.313	0.129	0.160	0.136	0.184	0.102	0.126
TextRank[6]	0.223	0.281	0.189	0.162	0.195	0.196	0.175	0.147
RNN[5]	0.085	0.064	0.135	0.088	0.169	0.127	0.179	0.189
CopyRNN[5]	0.278	0.342	0.311	0.266	0.334	0.326	0.333	0.262
Controlled Vocabulary+TF-IDF	0.395	0.363	0.298	0.257	**0.360**	**0.303**	0.315	0.247
Controlled Vocabulary+TextRank	0.330	0.341	0.252	0.247	0.296	0.264	0.271	0.235
Prior Knowledge+TF-IDF	0.401	0.365	0.301	0.261	0.271	0.258	0.299	0.245
Prior Knowledge+TextRank	0.321	0.367	0.205	0.224	0.205	0.231	0.286	0.242
Prior Knowledge+TF-IDF+TextRank	**0.403**	**0.371**	**0.309**	**0.267**	0.324	0.276	**0.326**	**0.252**

Features such as TF ad TR only consider the importance of candidate keywords from the perspective of mutual information in a document, but utilization of knowledge outside the document. We introduce the use of keyphrases in the specific domain as a background feature named prior probability. According to statistics of the given collection and their keyphrases, the prior probability PP is defined as

$$PP_{k_i} = \frac{A_{i1} + A_{i2}}{A_{i1} + A_{i3}} \quad (A_{i1} + A_{i3} > 0)$$

Where A_{i1} represents the number of times k_i appears in documents and is also selected by the author as the keyphrase. A_{i2} represents the number of times k_i does not appear in documents but is selected as the keyphrase. A_{i3} represents the number of times k_i is in documents but not selected as a keyphrase.

In addition, in order to ensure that the magnitude difference of these three features does not affect the importance of them, they are normalized in the same way. Taking PP_{k_i} as an example, the PR_{k_i} of each candidate keyphrase do the following process:

$$PP'_{k_i} = \frac{PP_{k_i} - PP_{min}}{PP_{max} - PP_{min}}$$

where PP_{max} is the maximum and PP_{min} is the minimum.

2.2.2 Supervised Learning Algorithm. We use logistic regression as ranking model in this step. For a phrase k_i, TF_{k_i}, TR_{k_i} and PR_{k_i} are three features as input of model:

$$y = sigmoid(w_1 \cdot TF_{k_i} + w_2 \cdot TR_{k_i} + w_3 \cdot PR_{k_i} + b)$$

where w_1, w_2, w_3 and b are parameters will be learned by Stochastic Gradient Decline (SGD)[1] and y is the output.

3 EXPERIMENT RESULT AND ANALYSIS

We conduct experiments on four public available datasets: Inspec[3], Krapivin[4], NUS[7] and Ke20K[5]. To evaluate the performance, we adopt the F1 score, which is the primary metrics used in keyphrase extraction. To further evaluate the effectiveness of prior knowledge, we compare our methods with some baselines shown as Table 1 whose settings are same as[5].

It is observed that unsupervised learning methods such as TF-IDF and TextRank combined with the controlled vocabulary of keyphrase which is a part of prior knowledge achieve a significant improvement compared with TF-IDF and TextRank. That is, it is very important and necessary to take the external knowledge of existing keyphrases into consideration. The supervised learning algorithm based on entire prior knowledge which contains controlled vocabulary and their prior probability achieve the best performance on all dataset except Nus. The main reason is that supervised learning algorithm needs a large training set to learn the optimal weight but there are only 200 samples in Nus.

Meanwhile, our methods achieve the new state-of-the art performance on Inspec in terms of both F1@5 and F1@10, Krapivin in terms of F1@10 and Nus in terms of F1@5. And the best results obtained from our methods are comparable with the state-of-the art model named Copy Recurrent Neural Network (CopyRNN).

4 CONCLUSIONS

To the best of our knowledge, this is the first try to consider prior knowledge in keyphrase extraction. And the effectiveness of prior knowledge is introduced by the empirical analysis.

5 ACKNOWLEDGEMENTS

We would like to thank National Demonstration Center for Experimental Library and Information Science Education, Wuhan University for sharing the GPU devices. This work is partially supported by the National Natural Science Foundation of China under Grant No.71473183.

REFERENCES

[1] Yoshua Bengio, Patrice Simard, and Paolo Frasconi. 1994. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks* 5, 2 (1994), 157–166.
[2] Eibe Frank, Gordon W Paynter, Ian H Witten, Carl Gutwin, and Craig G Nevill-Manning. 1999. Domain-specific keyphrase extraction. In *IJCAI 99*, Vol. 2. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 668–673.
[3] Anette Hulth. 2003. Improved automatic keyword extraction given more linguistic knowledge. In *EMNLP 2003*. Association for Computational Linguistics, 216–223.
[4] Mikalai Krapivin, Aliaksandr Autaeu, and Maurizio Marchese. 2009. *Large dataset for keyphrases extraction.* Technical Report. University of Trento.
[5] Rui Meng, Sanqiang Zhao, Shuguang Han, Daqing He, Peter Brusilovsky, and Yu Chi. 2017. Deep Keyphrase Generation. In *ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*. 582–592.
[6] Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *EMNLP 2004*.
[7] Thuy Nguyen and Min-Yen Kan. 2007. Keyphrase extraction in scientific publications. *Asian Digital Libraries*. (2007), 317–326.