Supervised Learning of Keyphrase Extraction Utilizing Prior Summarization

Tingyi Liu and Mizuho Iwaihara(✉)

Graduate School of Information, Production and Systems, Waseda University, Kitakyushu 808-0135, Japan
tyliu@akane.waseda.jp, iwaihara@waseda.jp

Abstract. Keyphrase extraction is the task of selecting a set of phrases that can best represent a given document. Keyphrase extraction is utilized in document indexing and categorization, thus being one of core technologies of digital libraries. Supervised keyphrase extraction based on pretrained language models are advantageous thorough their contextualized text representations. In this paper, we show an adaptation of the pertained language model BERT to keyphrase extraction, called BERT Keyphrase-Rank (BK-Rank), based on a cross-encoder architecture. However, the accuracy of BK-Rank alone is suffering when documents contain a large amount of candidate phrases, especially in long documents. Based on the notion that keyphrases are more likely to occur in representative sentences of the document, we propose a new approach called Keyphrase-Focused BERT Summarization (KFBS), which extracts important sentences as a summary, from which BK-Rank can more easily find keyphrases. Training of KFBS is by distant supervision such that sentences lexically similar to the keyphrase set are chosen as positive samples. Our experimental results show that the combination of KFBS + BK-Rank show superior performance over the compared baseline methods on well-known four benchmark collections, especially on long documents.

Keywords: Keyphrase extraction · Supervised learning · Pretrained language model · Extractive summarization · Document indexing

1 Introduction

Keyphrase extraction is a natural language processing task of automatically selecting a set of representative and characteristic phrases that can best describe a given document. Due to its clarity and practical importance, keyphrase extraction has been a core technology for information retrieval and document classification [1]. For large text collections, keyphrases provide faster and more accurate searches and can be used as concise summaries of documents [2, 18].

For keyphrase extraction, unsupervised methods have played an important role, because of corpus independence and search efficiency. However, compared with supervised methods, unsupervised methods only use statistical information from the target document and the document set. The performance of unsupervised is limited due to the

H.-R. Ke et al. (Eds.): ICADL 2021, LNCS 13133, pp. 157–166, 2021.
https://doi.org/10.1007/978-3-030-91669-5_13

lack of information on the contexts surrounding candidate phrases. Supervised methods can learn contextual information on where keyphrases are likely to occur, but they require training datasets.

In this paper, we discuss supervised keyphrase extraction based on finetuning pretrained language model BERT [6]. Our proposed method consist of two parts. First, Keyphrase-Focused BERT Summarization (KFBS) is applied for prior-summarization, which extracts important sentences that are likely to contain keyphrases. We utilize distant supervision for training of KFBS, such that sentences that contain words lexically similar to reference keyphrases are used as golden summaries for training.

After prior-summarization, part-of-speech (POS) tagging is applied to extract candidate noun phrases. BERT Keyphrase-Rank (BK-Rank) has a cross-encoder architecture which attends over the pair of the extracted summary sentences and a candidate phrase, and scores the candidate phrase. Top-ranked phrases are chosen as keyphrases. Our rigorous experimental evaluations show that our proposed method of KFBS+BK-Rank outperforms the baseline methods in terms of F1@K, by a large margin. The results also show that prior-summarization by KFBS improves the results of BK-Rank alone, especially on long documents.

2 Related Work

KP-Miner [7] is a keyphrase extraction system that considers various types of statistical information beyond the classical method TF-IDF [18]. YAKE [4] considers both statistical and contextual information, and adopts features such as the position and frequency of a term, and the spread of the terms within the document.

TextRank [14], borrowing the idea of PageRank [3], uses part-of-speech (POS) tags to obtain candidates, creates an undirected and unweighted graph in which the candidates are added as nodes and an edge is added between nodes that co-occur within a window of N words. Then the PageRank algorithm is applied. SingleRank [22] is an extension of TextRank which introduces weights on edges by the number of co-occurrences.

Embedding-based methods train low-dimensional distributed representations of phrases and documents for evaluating importance of phrases. EmbedRank [2] extracts candidate phrases from a given document based on POS tags. Then EmbedRank uses two different sentence embedding methods (Sent2vec [17] and Doc2vec [11]) to represent the candidate phrases and the document in the same low-dimensional vector space. Then the candidate phrases are ranked using the normalized cosine similarity between the embeddings of the candidate phrases and the document embedding. SIFRank [20] combines sentence embedding model SIF [1] which is used to explain the relationship between sentence embeddings and the topic of the document, and autoregressive pretrained language model ELMo [19] is used to compute phrase and document embeddings, and achieves the state-of-the-art performance in keyphrase extraction for short documents. For long documents, SIFRank is extended to SIFRank+ [20] by introducing position-biased weighting.

3 Methodology

3.1 Motivations

This section discusses motivations and backgrounds that lead us for designing a new keyphrase extraction method.

Context. Context information is vital in determining whether a phrase is a keyphrase. Local contexts often give clues on whether an important concept is stated or not. Also, phrases that are co-occurring with the main topic of the document can be regarded as representative. EmbedRank [2] utilizes context information through document embeddings, and SIFRank [20] adopts the pretrained language model Elmo [19] for context-aware embedding. Both EmbedRank and SIFRank are unsupervised method. On the contrary, BERT [6] captures deep context information through the multi-head self-attention mechanism. We design a BERT Keyphrase-Ranker, called BK-Rank, where keyphrase extraction is formulated as a phrase ranking problem.

Keyphrase Density. The number of keyphrases annotated by human annotators for a document is around 10–15 in average, as shown in the benchmark document collections in Table 1, which include both short documents, such as abstracts and news articles, and long documents such as scientific papers. This means that the density of keyphrases in long documents is relatively lower than in short documents. Also, long documents contain more diverse phrases that are apart from the main topic of the document. As a consequence, long documents are more difficult in finding keyphrases than short documents.

Considering the above analysis, we propose a new approach that integrates document summarization and keyphrase extraction. Extractive summarization [15] is a task to select sentences from a given target document such that the summary well represents the target document. We adopt the following assumption: Keyphrases are more likely to occur in representative sentences. We remove non-representative sentences from the document before keyphrase extraction, as *prior-summarization*. Our approach has the following expected effects:

1. Prior-summarization can reduce phrases that are remotely related to the topic of the document, while the summary retains local contexts of keyphrases that are utilized for final keyphrase extraction.
2. In a summary, keyphrases are more densely occurring than the original document, so that relations between phrases are more easily captured by the attention mechanism of BK-Rank.
3. Prior-summarization will be especially effective for long documents.

We propose a supervised keyphrase extraction method, based on finetuning pretrained language models for both prior-summarization and final keyphrase extraction. Our proposed method of KFBS+BK-Rank, illustrated in Fig. 1, consists of the following steps:

1. For a given document, prior-summarization is performed by KFBS, which is trained to extract important sentences that are lexically similar to the list of golden keyphrases, so that the selected important sentences are more likely to contain keyphrases.
2. Candidate phrases are extracted which are noun phrases based on POS tagging from prior-summarization.
3. BK-Rank is finetuned by binary cross-entropy loss on keyphrases and non-keyphrases, and used to score candidate phrases occurring in important sentences selected by KFBS.
4. The top-N phrases ranked by BK-Rank are selected as the keyphrases.

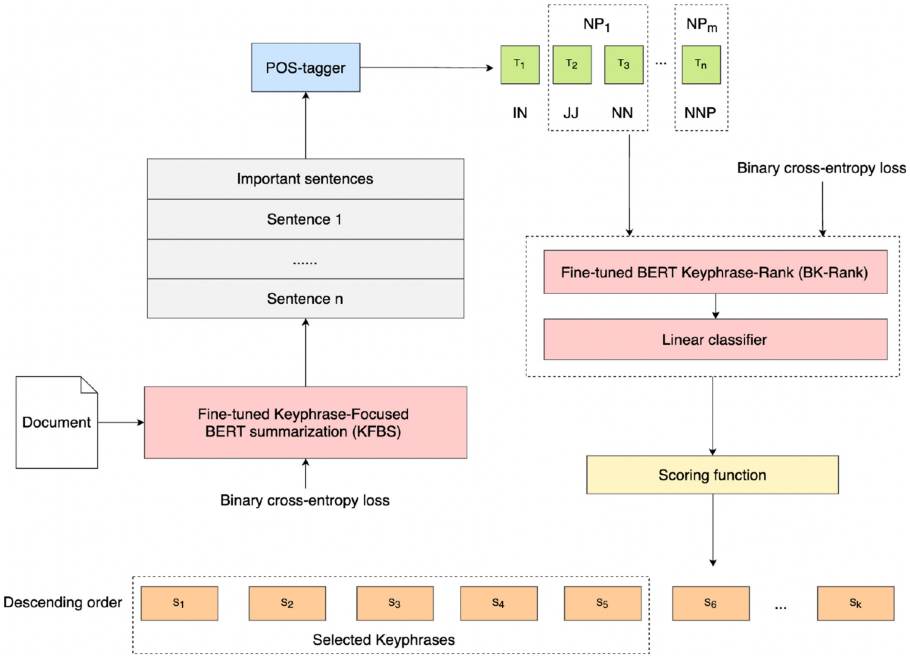

Fig. 1. The framework of our proposed method KFBS + BK-Rank.

3.2 Candidate Phrase Selection

In this stage, we apply **Keyphrase-Focused BERT Summarization (KFBS)** to select important sentences from a document that can represent the document and are more likely to contain keyphrases as the concise summary of this document.

BERTSUM [12] is an extractive summarization method, which changes the input of BERT by adding a CLS-token and a SEP-token at the start and end of each sentence respectively. The output vector at each CLS-token is used as a sentence embedding and

entered to the succeeding linear layer, and a fixed number of highly scored sentences are selected as the output summary. When the document exceeds the length limit of 512 tokens of BERT, the leading part of the document is used. In case the given document is already short, prior-summarization is skipped.

To train an extractive summarization model, we need reference summaries. However, since our target task is keyphrase extraction, only reference keyphrases are available as training samples. Therefore, we take the approach of distant supervision such that sentences that contain words or subwords of the reference keyphrases are regarded as quality sentences, and used as positive samples for training the extractive summarization model.

To evaluate overlapping words and subwords between sentences and keyphrases, we utilize the ROUGE-N score, which quantifies the overlap of N-grams. We score the sentences of the target document by the sum of ROUGE-1 + ROUGE-2, and choose the top-ranked sentences as important sentences for training. Binary cross-entropy loss is used for the model to learn the important sentences.

For short documents of length within 200 tokens, KFBS avoids extraction and returns the input document as the final output.

Part-of-speech (POS) Tagging. Keyphrases chosen by humans are often noun phrases that consist of zero or more adjectives followed by one or more nouns (e.g., communication system, supervised learning, word embedding). Thus we utilize part-of-speech (POS) tagging to extract candidate noun phrases as candidate phrases from the prior-summarization performed by KFBS, which are not allowed to end with adjectives, verbs, or adverbs, etc.

3.3 BERT Keyphrase-Rank (BK-Rank)

For final selection of keyphrases from candidate phrases, we construct a BERT model with two inputs: the prior-summarization text and a candidate phrase. We utilize a cross-encoder [9] which computes self-attention between the prior-summarization text and the candidate phrase, to capture relationship between these two parts. Figure 2 shows the configuration of BERT Keyphrase-Rank (BK-Rank). For keyphrase scoring, the classification outcome is whether or not a candidate phrase is a golden keyphrase. So we adopt binary cross-entropy loss for finetuning BK-Rank with a classifier which generates a scalar between 0 and 1. We note that the training documents as well as the target documents receive prior-summarization by KFSB, which needs to be trained before BK-Rank.

4 Experiments

In this section, we report our experimental evaluations of our proposed models, compared with baseline methods, on four commonly used datasets. F1@K is used for evaluating results.

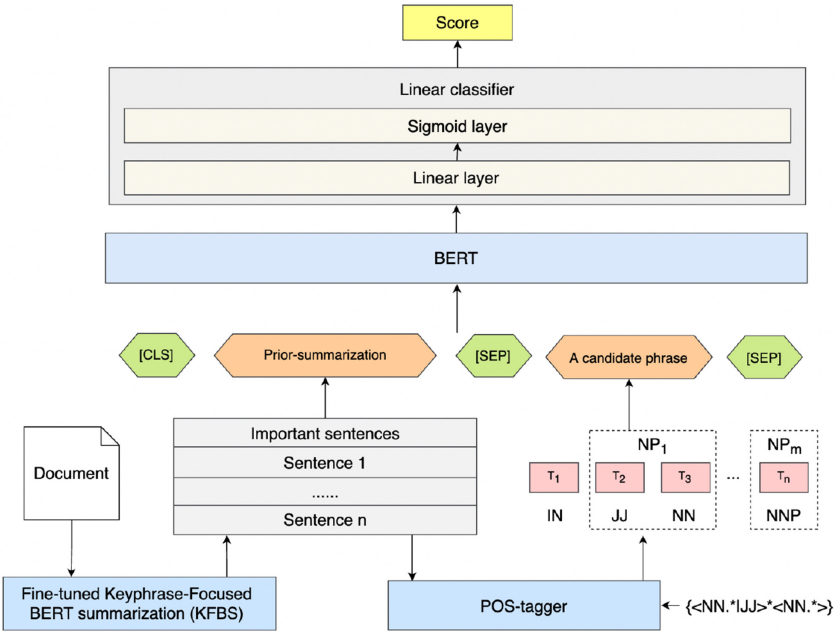

Fig. 2. The configuration of BERT Keyphrase-Rank (BK-Rank).

Table 1. Statistics of four datasets.

Dataset	Documents		Keyphrases			
	Number (test set)	Average tokens	Total	Average	Missing in doc	Missing in candidates
Inspec	500	134.28	4913	9.83	14.46%	38.94%
DUC 2001	123	800.63	1010	8.21	4.11%	10.38%
SemEval 2010	100	7662.42	1467	14.67	15.11%	16.36%
NUS	100	8765.93	1106	11.06	5.68%	11.86%

4.1 Datasets

Table 1 shows the statistics of the four benchmark datasets.

- **Inspec** [8] consists of 2,000 short documents from scientific journal abstracts in English. The training set, validation set, and test set contain 1,000, 500, and 500 documents, respectively.
- **DUC 2001** [22] consists of 308 newspaper articles which are collected from TREC-9, where the documents are organized into 30 topics. The golden keyphrases we used are annotated by X. Wan and J. Xiao. Here we use 145 for training and 123 for test.

- **SemEval 2010** [10] consists of 284 long documents which are scientific papers, 144 documents for training, 100 documents for test and 40 for validation.
- **NUS** [16] consists of 211 long documents which are full scientific conference papers of between 4–12 pages. Here we use 111 for training and 100 for test.

Table 2. Comparison of our method and baseline methods, by F1@K (%).

Method	Inspec		DUC 2001		SemEval 2010		NUS	
	F1@5	F1@10	F1@5	F1@10	F1@5	F1@10	F1@5	F1@10
Baseline method								
TextRank	14.72	15.28	15.17	15.24	3.23	6.55	3.21	6.56
EmbedRank doc2vec	31.51	37.94	24.02	28.12	2.28	3.53	2.35	3.58
EmbedRank sent2vec	29.88	37.09	27.16	31.85	3.31	5.33	3.39	5.42
SIFRank	29.11	38.80	24.27	27.43	3.05	5.43	–	–
SIFRank +	28.49	36.77	30.88	33.37	10.47	12.40	–	–
Proposed method								
BK-Rank	41.99	46.53	42.08	46.88	9.49	13.54	11.46	16.55
KFBS (Top-3) + BK-Rank	38.89	44.22	40.44	44.11	11.12	13.30	**17.60**	**17.70**
KFBS (Top-4) + BK-Rank	38.15	44.31	41.17	45.53	**11.18**	**15.59**	17.41	15.96
KFBS (Top-5) + BK-Rank	**42.01**	**46.62**	**42.16**	**46.93**	10.05	13.46	17.24	16.23

4.2 Baseline Methods

We compare our proposed method with the following baseline methods: **TextRank** [14], **EmbedRank** [2], and **SIFRank/SIFRank+** [20]. SIFRank is an unsupervised method which combines sentence embedding model SIF [1] and pretrained language model ELMo [19] to generate embeddings. For long documents, SIFRank is upgraded to SIFRank + by position-biased weight.

4.3 Experimental Details

In the experiments, we use StandfordCoreNLP [21] to generate POS tags and use AdamW [13] as the optimizer. For training KFBS, which is used to select important sentences, we finetune the model with learning rate in {5e−5, 3e−5, 2e−5, 1e−5}, dropout rate 0.1, batch size 256, and warm-up 5% of the training steps. We finetune

BERT Keyphrase-Rank (BK-Rank) with a batch size of 32, learning rate in {5e−5, 3e−5, 2e−5, 1e−5}, weight decay 0.01, and warm-up 10% of the training data. Then we save the models which achieve the best performances. For the pretrained language models, we use bert-base-uncased model for both BK-Rank and KFBS.

4.4 Performance Comparison

For evaluation, we use the common metrics of F1-score (F1). Table 2 shows the results. KFBS (Top-k) means top-k sentences selected by KFBS are used as important sentences, on which KB-Rank is applied. Due to hardware limitations, SIFRank and SIFRank+ are not obtained on NUS, so we do not report their results.

As shown in Table 2, the performance of KFBS + BK-Rank shows the best results on all the four datasets, both on short documents and long documents, achieving superior performance over the compared baseline methods. When we select top-5 sentences by KFBS, KFBS + BK-Rank achieves the best results on Inspec and DUC 2001 for F1@5 and F1@10. KFBS (Top-4) + BK-Rank achieves the best results on F1@5 and F1@10 on SemEval 2010. On NUS, KFBS (Top-3) + BK-Rank achieves the best results on F1@5 and F1@10.

Prior-summarization by KFBS is improving the results of BK-Rank by 0.02 to 6.17 points. The results show that selecting important sentences before candidate phrase selection by BK-Rank is effective, especially on long document collections of SemEval 2010 and NUS. Prior-summarization by KFBS is effectively removing sentences that are unlikely to contain keyphrases, which also benefits finetuning of BK-Rank. We notice that on Inspec and DUC 2001, KFBS (Top-k) with $k = 5$ is better than $k = 3$ or 4, while on SemEval 2010 and NUS, $k = 5$ is falling behind of $k = 3$ and 4. This can be explained by keyphrase density such that for short documents, keyphrases are relatively evenly occurring in sentences, while for long documents, more selective summarization is advantageous.

5 Conclusion

In this paper, we proposed a supervised method for keyphrase extraction from documents, by combining BERT Keyphrase-Rank (BK-Rank) and Keyphrase-Focused BERT Summarization (KFBS). We introduce KFBS to select important sentences from which candidate phrases are extracted and also used for finetuning BK-Rank. BK-Rank fully exploits contextual text embeddings by the cross-encoder reading a target document and candidate phrase. KFBS is trained by distant supervision to extract important sentences that are likely to contain keyphrases. Our experimental results show that our proposed method has superior performance on this task over the compared baseline methods. Diversity on keyphrases is necessary to avoid the situation that similar keyphrases occupy the result. BK-Rank can be extended to incorporate Maximal Marginal Relevance (MMR) [5] for enhancing diversity.

References

1. Arora, S., Liang, Y., Ma, T.: A simple but tough-to-beat baseline for sentence embeddings. In: ICLR (2016)
2. Bennani-Smires, K., Musat, C.C., Hossmann, A., et al.: Simple unsupervised keyphrase extraction using sentence embeddings. In: Conference on Computational Natural Language Learning (2018)
3. Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine. Comput. Netw. ISDN Syst. **30**(1–7), 107–117 (1998)
4. Campos, R., Mangaravite, V., Pasquali, A., Jorge, A.M., Nunes, C., Jatowt, A.: YAKE! Collection-independent automatic keyword extractor. In: Pasi, G., Piwowarski, B., Azzopardi, L., Hanbury, A. (eds.) ECIR 2018. LNCS, vol. 10772, pp. 806–810. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-76941-7_80
5. Carbonell, J., Goldstein, J.: The use of MMR, diversity-based reranking for reordering documents and producing summaries. In: Proceedings of 21st Annual International ACM SIGIR Conference on on Research and Development in Information Retrieval, pp. 335–336 (1998)
6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL-HLT 2019, pp. 4171–4186 (2019)
7. El-Beltagy, S.R., Rafea, A.: KP-miner: participation in SemEval-2. In: Proceedings of. 5th Int. Workshop on Semantic Evaluation, pp. 190–193 (2010)
8. Hulth, A.: Improved automatic keyword extraction given more linguistic knowledge. In: Proceedings of 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 216–223 (2003)
9. Humeau, S., Shuster, K., Lachaux, M.A., et al.: Poly-encoders: architectures and pre-training strategies for fast and accurate multi-sentence scoring. In: International Conference on Learning Representations (2019)
10. Kim, S.N., Medelyan, O., Kan, M.Y., et al.: Semeval-2010 task 5: automatic keyphrase extraction from scientific articles. In: Proceedings of 5th International Workshop on Semantic Evaluation, pp. 21–26 (2010)
11. Le, Q., Mikolov, T.: Distributed representations of sentences and documents. In: International Conference on Machine Learning, pp. 1188–1196 (2014)
12. Liu, Y.: Fine-tune BERT for Extractive Summarization. arXiv preprint arXiv:1903.10318 (2019)
13. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2018)
14. Mihalcea, R., Tarau, P.: TextRank: bringing order into text. In: Proceedings of 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 404–411 (2004)
15. Moratanch, N., Chitrakala, S.: A survey on extractive text summarization. In: 2017 International Conference on Computer, Communication and Signal Processing (ICCCSP). IEEE (2017)
16. Nguyen, T.D., Kan, M.-Y.: Keyphrase extraction in scientific publications. In: Goh, D.-L., Cao, T.H., Sølvberg, I.T., Rasmussen, E. (eds.) ICADL 2007. LNCS, vol. 4822, pp. 317–326. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-77094-7_41
17. Pagliardini, M., Gupta, P., Jaggi, M.: Unsupervised learning of sentence embeddings using compositional n-gram features. In: Proceedings of NAACL-HLT, pp. 528–540 (2018)
18. Papagiannopoulou, E., Tsoumakas, G.: A review of keyphrase extraction. Wiley Interdisc. Rev. Data Min. Knowl. Discov. **10**(2) e1339 (2020)

19. Peters, M.E., et al.: Deep contextualized word representations. arXiv preprint arXiv:1802.05365 (2018)
20. Sun, Y., et al.: SIFRank: a new baseline for unsupervised keyphrase extraction based on pre-trained language model. IEEE Access **8**, 10896–10906 (2020)
21. Toutanova, K., et al.: Feature-rich part-of-speech tagging with a cyclic dependency network. In: Proceedings of 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (2003)
22. Wan, X., Xiao, J.: Single document keyphrase extraction using neighborhood knowledge. In: AAAI Conference on Artificial Intelligence (AAAI-08), pp. 855–860 (2008)