Keyphrase Generation for Scientific Document Retrieval

Florian Boudin, Ygor Gallina, Akiko Aa Aizawa

Keyphrase Generation for Scientific Document Retrieval

Florian Boudin **Ygor Gallina**
LS2N, Université de Nantes, France
first.last@univ-nantes.fr

Akiko Aizawa
National Institute of Informatics, Tokyo
aizawa@nii.ac.jp

Abstract

Sequence-to-sequence models have lead to significant progress in keyphrase generation, but it remains unknown whether they are reliable enough to be beneficial for document retrieval. This study provides empirical evidence that such models can significantly improve retrieval performance, and introduces a new extrinsic evaluation framework that allows for a better understanding of the limitations of keyphrase generation models. Using this framework, we point out and discuss the difficulties encountered with supplementing documents with –not present in text– keyphrases, and generalizing models across domains. Our code is available at https://github.com/boudinfl/ir-using-kg.

1 Introduction

With the exponential growth of the scientific literature (Bornmann and Mutz, 2015), retrieving relevant scientific papers becomes increasingly difficult. Keywords, also referred to as keyphrases, provide an effective way to supplement paper indexing and improve retrieval effectiveness in scientific digital libraries (Barker et al., 1972; Zhai, 1997; Gutwin et al., 1999; Lu and Kipp, 2014). However, only few documents have assigned keyphrases, and those who do were, for the most part, self-labeled by their authors, thus exhibiting annotation inconsistencies (Strader, 2011; Suzuki et al., 2011). This has motivated an active line of research on automatic keyphrase extraction (see Hasan and Ng (2014) for an overview) and, more recently, keyphrase generation (Meng et al., 2017), where the task is to find a set of words and phrases that represents the main content of a document.

Although models for predicting keyphrases have been extensively evaluated on their ability to reproduce author's keywords, it still remains unclear whether they can be usefully applied in information retrieval. One reason for this lack of evidence may have been their relatively low performance discouraging attempts at using them for indexing (Liu et al., 2010; Hasan and Ng, 2014). Yet, recently proposed models not only achieve much better performance, but also display a property that may have a significant impact on retrieval effectiveness: the capacity to generate keyphrases that do not appear in the source text. These *absent* keyphrases do not just highlight the topics that are most relevant, but provide some form of semantic expansion by adding new content (e.g. synonyms, semantically related terms) to the index (Greulich, 2011). The goal of this paper is two-fold: to gather empirical evidence as to whether current keyphrase generation models are good enough to improve scientific document retrieval, and to gain further insights into the performance of these models from an extrinsic perspective. Our contributions are listed as follows:

- We report significant improvements for strong retrieval models on a standard benchmark collection, showing that keyphrases produced by state-of-the-art models are consistently helpful for document retrieval, even, to our surprise, when author keywords are provided.
- We introduce a new extrinsic evaluation framework for keyphrase generation that allows for a deeper understanding of the limitations of current models. Using it, we discuss the difficulties associated with domain generalization and absent keyphrase prediction.

2 Methodology

This section presents our methodology for assessing the usefulness (§2.3) of keyphrase generation (§2.2) in scientific document retrieval (§2.1).

2.1 Scientific Document Retrieval

Here, we focus on the task of searching through a collection of scientific papers for relevant docu-

ments. All of our experiments are conducted on the NTCIR-2 test collection (Kando, 2001) which is, to our knowledge, the only available benchmark dataset for that task. It contains 322,058 documents[1] (title and abstract pairs) and 49 search topics (queries) with relevance judgments. Most of the documents (98.6%) include author keywords (4.8 per doc. on avg.), which we later use to investigate the performance of keyphrase generation models.

Documents cover a broad range of domains from pure science to social sciences and humanities, although half of the documents are about engineering and computer science. Queries are also categorized into one or more research fields (e.g. science, chemistry, engineering), the original intent being to help retrieval models in narrowing down the search space. We follow common practice and use short[2] queries with binary relevance judgments (i.e. without "partially relevant" documents).

We consider two standard *ad-hoc* retrieval models to rank documents against queries: BM25 and query likelihood (QL), both implemented in the Anserini IR toolkit (Yang et al., 2017). These models use unsupervised techniques based on corpus statistics for term weighting, and will therefore be straightforwardly affected when keyphrases are added to a document. We further apply a pseudo-relevance feedback method, known as RM3 (Abdul-Jaleel et al., 2004), on top of the models to achieve strong, near state-of-the-art retrieval results (Lin, 2019; Yang et al., 2019). For all models, we use Anserini's default parameters.

To verify the effectiveness of the adopted retrieval models, we compared their performance with that of the best participating systems in NTCIR-2. Retrieval performance is measured using mean average precision (MAP) and precision at 10 retrieved documents (P@10). MAP measures the overall ranking quality and P@10 reflects the number of relevant documents on the first page of search results. Documents are indexed with author keywords, same as for participating systems. Results are presented in Table 1. We see that the considered retrieval models achieve strong performance, even outperforming the best participating system by a substantial margin. Note that the two best-performing systems use pseudo-relevance feedback, and that the second-ranked system is based on BM25.

[1] Scientific abstracts and summaries of research results.
[2] <description> field of topic description.

Model	MAP	P@10
BM25+RM3	**35.17**	**38.57**
QL+RM3	33.00	34.90
1st (Fujita, 2001)	31.93	37.35
BM25	31.38	36.33
2nd (Murata et al., 2001)	31.31	36.12
QL	30.63	34.08
3rd (Chen et al., 2001)	26.24	33.88

Table 1: Retrieval effectiveness of the considered models and the best participating systems on NTCIR-2.

2.2 Keyphrase Generation

Keyphrase generation is the task of producing a set of words and phrases that best summarise a document (Evans and Zhai, 1996). In contrast with most previous work that formulates this task as an extraction problem (a.k.a. keyphrase extraction), which can be seen as ranking phrases extracted from a document, recent neural models for keyphrase generation are based on sequence-to-sequence learning (Sutskever et al., 2014; Bahdanau et al., 2014), thus potentially allowing them to generate any phrase, also beyond those that appear verbatim in the text. In this study, we consider the following two neural keyphrase generation models:

seq2seq+copy (Meng et al., 2017) is a sequence-to-sequence model with attention, augmented with a copying mechanism (Gu et al., 2016) to predict phrases that rarely occur. The model is trained with document-keyphrase pairs and uses beam search decoding for inference.

seq2seq+corr (Chen et al., 2018) extends the aforementioned model with correlation constraints. It employs a coverage mechanism (Tu et al., 2016) that diversifies attention distributions to increase topic coverage, and a review mechanism to avoid generating duplicates.

We implemented the models in PyTorch (Paszke et al., 2017) using AllenNLP (Gardner et al., 2018). Models are trained on the KP20k dataset (Meng et al., 2017), which contains 567,830 scientific abstracts with gold-standard, author-assigned keywords (5.3 per doc. on avg.). We use the parameters suggested by the authors for each model.

To validate the effectiveness of our implementations, we conducted an intrinsic evaluation by counting the number of exact matches between predicted and gold keyphrases. We adopt the standard

metric and compute the f-measure at top 5, as it corresponds to the average number of keyphrases in KP20k and NTCIR-2, that is, 5.3 and 4.8, respectively. We also examine cross-domain generalization using the KPTimes news dataset (Gallina et al., 2019), and include a state-of-the-art unsupervised keyphrase extraction model (Boudin, 2018, henceforth mp-rank) for comparison purposes. This latter baseline also provides an additional relevance signal based on graph-based ranking whose usefulness in retrieval will be tested in subsequent experiments. Results are reported in Table 2. Overall, our results are consistent with those reported in (Meng et al., 2017; Chen et al., 2018), demonstrating the superiority of well-trained neural models over unsupervised ones, and stressing their lack of robustness across domains. Rather surprisingly, seq2seq+corr is outperformed by seq2seq+copy which indicates that relevant, yet possibly redundant, keyphrases are filtered out by the added mechanisms for promoting diversity in the output.

Model	KP20k	NTCIR-2	KPTimes
s2s+copy	**27.75**	**23.90**	**16.47**
s2s+corr	23.78	22.27	11.73
mp-rank	14.67	18.10	14.59

Table 2: f-measure at top-5 predicted keyphrases. Stemming is applied to reduce the number of mismatches.

2.3 Extrinsic Evaluation Framework

Our goal is to find out whether the keyphrase generation models described above are reliable enough to be beneficial for document retrieval. To do so, we contrast the performance of the retrieval models with and without automatically predicted keyphrases. Two initial indexing configurations are also examined: title and abstract only (T+A), and title, abstract and author keywords (T+A+K). The idea here is to investigate whether generated keyphrases simply act as a proxy for author keywords, or instead supplement them.

Unless mentioned otherwise, the top-5 predicted keyphrases are used to expand documents, which is in accordance with the average number of author keywords in NTCIR-2. We evaluate retrieval performance in terms of MAP and omit P@10 for brevity. We use the Student's paired t-test to assess statistical significance of our retrieval results at $p < 0.05$ (Smucker et al., 2007).

3 Results

Results for retrieval models using keyphrase generation are reported in Table 3. We note that indexing keyphrases generated by seq2seq+copy, which performs best in our intrinsic evaluation, significantly improves retrieval effectiveness for all models. More interestingly, gains in effectiveness are also significant when both keyphrases and author keywords are indexed, indicating they complement each other well. This important finding suggests that predicted keyphrases are consistently helpful for document retrieval, and should be used even when author keywords are provided. Another important observation is that while both keyphrase generation models perform reasonably well in our intrinsic evaluation on NTCIR-2 (cf. Table 2, column 3), their impact on retrieval effectiveness are quite different, as only s2s+copy reaches consistent significance. This finding advocates for the importance of using document retrieval as an extrinsic evaluation task for keyphrase generation.

Index	BM25	+RM3	QL	+RM3
T+A	29.16	31.93	28.98	31.47
+ s2s+copy	30.54†	**34.30**†	30.58†	33.26†
+ s2s+corr	30.30†	33.24	29.76	31.38
+ mp-rank	29.24	32.27	29.57	32.29
T+A+K	31.38	35.17	30.63	33.00
+ s2s+copy	31.55	**36.53**‡	31.70‡	35.15‡
+ s2s+corr	31.37	35.84	31.14	33.65
+ mp-rank	31.38	35.18	31.23	33.47

Table 3: MAP scores for retrieval models using various indexing configurations. † and ‡ indicate significance over T+A and T+A+K, respectively.

Overall, BM25+RM3 achieves the best retrieval effectiveness, confirming previous findings on *ad-hoc* retrieval in limited data scenarios (Lin, 2019). For clarity and conciseness, we focus on this model in the rest of this paper. Encouraging diversity in keyphrases seems not to be appropriate for retrieval, as seq2seq+corr consistently gives lower results than seq2seq+copy. It is also interesting to see that the effectiveness gains of query expansion (RM3) and document expansion are additive, suggesting that they provide different but complementary relevance signals. Moreover, our results show that query expansion is more effective, which is in line with past work (Billerbeck and Zobel, 2005).

One hyper-parameter that we have deliberately left untouched so far is the number N of predicted

keyphrases that directly controls the precision-recall trade-off of keyphrase generation models. To understand how this parameter affects retrieval effectiveness, we repeated our experiments by varying N within the range $[0, 9]$, and plotted the results in Figure 1. Without author keywords, we observe that all models achieve gains, but only seq2seq+copy does yield significant improvements. With author keywords, seq2seq+copy is again the only model that achieves significance, while the others show mixed results, sometimes even degrading scores. One likely explanation for this is that these models produce keyphrases that cause documents to drift away from their original meaning. We note that results are close to optimal for $N = 5$, supporting our initial setting for this parameter.

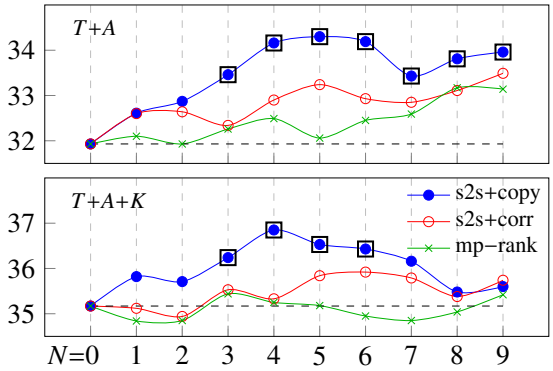

Figure 1: MAP scores for BM25+RM3 w.r.t. the number N of predicted keyphrases. □ denotes significance.

From our experiments, it appears that unsupervised keyphrase extraction is not effective enough to significantly improve retrieval effectiveness. The fact that keyphrase generation does so, suggests that the ability to predict absent keyphrases may be what enables better performance. Yet counter-intuitively, we found that most of the gains in retrieval effectiveness are due to the high extractive accuracy of keyphrase generation models. Results in Table 4 show that expanding documents with only absent keyphrases is at best useless and at worst harmful, while using only present keyphrases brings significant improvements. We draw two conclusions from this. First, absent keyphrases may not be useful in practice unless they are tied to some form of domain terminology to prevent semantic drift. Second, as generation does not yield improvements, keyphrase extraction models may be worth further investigation. In particular, supervised models could theoretically provide similar results while being easier to train.

Model	*T*+*A* (cf. 31.93)		*T*+*A*+*K* (cf. 35.17)	
	pres.	**abs.**	**pres.**	**abs.**
s2s+copy	**34.17**†	32.14	**36.30**†	34.97
s2s+corr	32.97	31.96	36.09	34.77

Table 4: MAP scores for BM25+RM3 using the top-5 present or absent keyphrases. † indicates significance over indexing without predicted keyphrases.

Neural models for keyphrase generation exhibit a limited generalization ability, which means that their performance degrades on documents that differ from the ones encountered during training (cf. Table 2, columns 3 and 4). To quantify how much this affects retrieval effectiveness, we divided the queries into two disjoint sets: *in-domain* for those that belong to research fields present in KP20k, and *out-domain* for the others. Results are presented in Table 5. The first thing we notice is the overall lower performance of *out-domain* queries, which may be explained by the unbalanced distribution of domains in the NTCIR-2 collection. Most importantly, *out-domain* queries on full indexing (i.e. *T*+*A*+*K*) is the only configuration in which no significant gains in retrieval effectiveness are achieved. This last experiment shows that expanding documents using existing keyphrase generation models may be ineffective in the absence of in-domain training data, and stresses the need of domain adaptation for keyphrase generation.

Model	*T*+*A*		*T*+*A*+*K*	
	I (32.70)	**O** (30.99)	**I** (36.18)	**O** (33.93)
s2s+copy	**35.40**†	**32.96**†	**38.13**†	**34.55**
s2s+corr	33.49	32.92	37.13	34.25
mp-rank	32.73	31.71	36.74	33.26

Table 5: MAP scores for BM25+RM3 on *in-domain* (I) and *out-domain* (O) queries. † indicates significance over w/o keyphrases whose scores are in parentheses.

4 Conclusion

We presented the first study of the usefulness of keyphrase generation for scientific document retrieval. Our results show that keyphrases can significantly improve retrieval effectiveness, and also highlight the importance of evaluating keyphrase generation models from an extrinsic perspective. Other retrieval tasks may also benefit from using keyphrase information and we expect our results to serve as a basis for further improvements.

References

Nasreen Abdul-Jaleel, James Allan, W Bruce Croft, Fernando Diaz, Leah Larkey, Xiaoyan Li, Mark D Smucker, and Courtney Wade. 2004. Umass at trec 2004: Novelty and hard. *Computer Science Department Faculty Publication Series*, page 189.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Frances H Barker, Douglas C Veal, and Barry K Wyatt. 1972. Comparative efficiency of searching titles, abstracts, and index terms in a free-text data base. *Journal of Documentation*, 28(1):22–36.

Bodo Billerbeck and Justin Zobel. 2005. Document expansion versus query expansion for ad-hoc retrieval. In *Proceedings of the 10th Australasian Document Computing Symposium*, pages 34–41. Citeseer.

Lutz Bornmann and Rüdiger Mutz. 2015. Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology*, 66(11):2215–2222.

Florian Boudin. 2018. Unsupervised keyphrase extraction with multipartite graphs. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 667–672, New Orleans, Louisiana. Association for Computational Linguistics.

Aitao Chen, Fredric C Gey, and Hailing Jiang. 2001. Berkeley at ntcir-2: Chinese, japanese, and english ir experiments. In *Proceedings of the Second NTCIR Workshop on Research in Chinese & Japanese Text Retrieval and Text Summarization*.

Jun Chen, Xiaoming Zhang, Yu Wu, Zhao Yan, and Zhoujun Li. 2018. Keyphrase generation with correlation constraints. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4057–4066, Brussels, Belgium. Association for Computational Linguistics.

David A. Evans and Chengxiang Zhai. 1996. Noun phrase analysis in large unrestricted text for information retrieval. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, pages 17–24, Santa Cruz, California, USA. Association for Computational Linguistics.

Sumio Fujita. 2001. Notes on the limits of clir effectiveness: Ntcir-2 evaluation experiments at justsystem. In *Proceedings of the Second NTCIR Workshop on Research in Chinese & Japanese Text Retrieval and Text Summarization*.

Ygor Gallina, Florian Boudin, and Béatrice Daille. 2019. Kptimes: A large-scale dataset for news keyphrase generation. In *Proceedings of the 12th International Conference on Natural Language Generation*. Association for Computational Linguistics.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. AllenNLP: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6, Melbourne, Australia. Association for Computational Linguistics.

Walter Greulich. 2011. Scientific texts and the indexer. *The Indexer: The International Journal of Indexing*, 29(3):114–122.

Jiatao Gu, Zhengdong Lu, Hang Li, and Victor O.K. Li. 2016. Incorporating copying mechanism in sequence-to-sequence learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1631–1640, Berlin, Germany. Association for Computational Linguistics.

Carl Gutwin, Gordon Paynter, Ian Witten, Craig Nevill-Manning, and Eibe Frank. 1999. Improving browsing in digital libraries with keyphrase indexes. *Decis. Support Syst.*, 27(1-2):81–104.

Kazi Saidul Hasan and Vincent Ng. 2014. Automatic keyphrase extraction: A survey of the state of the art. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1262–1273, Baltimore, Maryland. Association for Computational Linguistics.

Noriko Kando. 2001. Overview of the second ntcir workshop. In *Proceedings of the Second NTCIR Workshop on Research in Chinese & Japanese Text Retrieval and Text Summarization*.

Jimmy Lin. 2019. The neural hype and comparisons against weak baselines. *SIGIR Forum*, 52(2):40–51.

Zhiyuan Liu, Wenyi Huang, Yabin Zheng, and Maosong Sun. 2010. Automatic keyphrase extraction via topic decomposition. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 366–376, Cambridge, MA. Association for Computational Linguistics.

Kun Lu and Margaret E.I. Kipp. 2014. Understanding the retrieval effectiveness of collaborative tags and author keywords in different retrieval environments: An experimental study on medical collections. *Journal of the Association for Information Science and Technology*, 65(3):483–500.

Rui Meng, Sanqiang Zhao, Shuguang Han, Daqing He, Peter Brusilovsky, and Yu Chi. 2017. Deep keyphrase generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 582–592, Vancouver, Canada. Association for Computational Linguistics.

Masaki Murata, Masao Utiyama, Qing Ma, Hiromi Ozaku, and Hitoshi Isahara. 2001. Crl at ntcir2. In *Proceedings of the Second NTCIR Workshop on Research in Chinese & Japanese Text Retrieval and Text Summarization*.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch. In *NIPS 2017 Workshop Autodiff*.

Mark D. Smucker, James Allan, and Ben Carterette. 2007. A comparison of statistical significance tests for information retrieval evaluation. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, CIKM '07, pages 623–632, New York, NY, USA. ACM.

C Rockelle Strader. 2011. Author-assigned keywords versus library of congress subject headings. *Library resources & technical services*, 53(4):243–250.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3104–3112. Curran Associates, Inc.

Takafumi Suzuki, Kiyoko Uchiyama, Ryota Tomisaka, and Akiko Aizawa. 2011. Analyzing the characteristics of academic paper categories by using an index of representativeness. In *Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation*, pages 587–596, Singapore. Institute of Digital Enhancement of Cognitive Processing, Waseda University.

Zhaopeng Tu, Zhengdong Lu, Yang Liu, Xiaohua Liu, and Hang Li. 2016. Modeling coverage for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 76–85, Berlin, Germany. Association for Computational Linguistics.

Peilin Yang, Hui Fang, and Jimmy Lin. 2017. Anserini: Enabling the use of lucene for information retrieval research. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, pages 1253–1256, New York, NY, USA. ACM.

Wei Yang, Kuang Lu, Peilin Yang, and Jimmy Lin. 2019. Critically examining the "neural hype": Weak baselines and the additivity of effectiveness gains from neural ranking models. In *Proceedings of the 42Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, pages 1129–1132, New York, NY, USA. ACM.

Chengxiang Zhai. 1997. Fast statistical parsing of noun phrases for document indexing. In *Fifth Conference on Applied Natural Language Processing*, pages 312–319, Washington, DC, USA. Association for Computational Linguistics.