Redefining Absent Keyphrases and their Effect on Retrieval Effectiveness

Florian Boudin
LS2N, Université de Nantes, France
`florian.boudin@univ-nantes.fr`

Ygor Gallina
LS2N, Université de Nantes, France
`ygor.gallina@univ-nantes.fr`

Abstract

Neural keyphrase generation models have recently attracted much interest due to their ability to output *absent keyphrases*, that is, keyphrases that do not appear in the source text. In this paper, we discuss the usefulness of absent keyphrases from an Information Retrieval (IR) perspective, and show that the commonly drawn distinction between present and absent keyphrases is not made explicit enough. We introduce a finer-grained categorization scheme that sheds more light on the impact of absent keyphrases on scientific document retrieval. Under this scheme, we find that only a fraction (around 20%) of the words that make up keyphrases actually serves as document expansion, but that this small fraction of words is behind much of the gains observed in retrieval effectiveness. We also discuss how the proposed scheme can offer a new angle to evaluate the output of neural keyphrase generation models.

1 Introduction

Searching the scholarly literature for documents of interest is becoming frustratingly difficult and time-consuming as the volume of published research grows exponentially. One promising approach to address this problem and improve the retrievability of documents is to supplement paper indexing with automatically generated keyphrases (Zhai, 1997; Gutwin et al., 1999; Boudin et al., 2020). Traditionally, keyphrases are defined as a short list of terms that represent the main concepts in a document (Turney, 2000). In recent years, this definition was further refined to differentiate between keyphrases that are present in the source document or not, and in turn, proposed models for producing keyphrases were divided into extractive (Florescu and Caragea, 2017; Boudin, 2018; Sun et al., 2019; Wang et al., 2020; Santosh et al., 2020, *inter alia*) and generative models (Meng et al., 2017; Zhao and Zhang, 2019; Chen et al., 2020; Bahuleyan and

El Asri, 2020, *inter alia*) based on their ability to output absent keyphrases.

Obviously, keyphrases have different effects on retrieval models depending on whether or not they occur in the document: *present keyphrases* highlight important parts of the input and make weighting terms easier, while *absent keyphrases* add new terms to the input and provide some form of document expansion. Intuitively, assigning absent keyphrases is more appealing since it may alleviate the *vocabulary mismatch* problem between query terms and relevant documents (Furnas et al., 1987), hence enabling the retrieval of relevant documents that otherwise would have been missed. This is especially true for scholarly collections, in which documents are mostly short texts (i.e. scientific abstracts) due to licensing issues and/or resource limitations (Huang et al., 2019). Yet, the extent to which present and absent keyphrases contribute to improved retrieval effectiveness has not been thoroughly explored. Worse still, there is no unique and rigorous definition of what exactly makes a keyphrase absent.

Although not stated explicitly, many recent studies adopt the definition by (Meng et al., 2017), in which keyphrases that do not match any contiguous subsequence of source text are regarded as absent. From an Information Retrieval (IR) perspective where stemmed content words are used to index documents, this definition is not sufficiently explicit, as demonstrated by the example shown in Figure 1. We see that, under this definition, some absent keyphrases can have all of their words occurring in the source document, and therefore act no differently from present keyphrases on indexing. In fact, only a fraction of the words that compose these absent keyphrases are genuinely expanding the document, which in our example are the set of words ⌈*retrieval*, *behavior*, *support*⌋. From a keyphrase generation point of view, this definition is not entirely satisfactory either, since training a

Study on the Structure of Index Data for Metasearch System

This paper proposes a new technique for Metasearch system, which is based on the grouping of both keywords and URLs. This technique enables metasearch systems to share information and to reflect the estimation of users' preference. With this system, users can search not only by their own keywords but by similarity of HTML documents. In this paper, we describe the principle of the grouping technique as well as the summary of the existing search systems.

Present kps: Metasearch – Search System

Absent kps: Information Sharing – Information Retrieval – User's Behavior – Retrieval Support

Reordered   Mixed   Mixed   Unseen

Figure 1: Sample document (title, abstract) from the NTCIR-2 test collection (docid: gakkai-e-0001384947). Author-assigned keyphrases are divided into present and absent using token-level matching with stemming. Finer-grained categories for absent keyphrases (i.e. Reordered, Mixed and Unseen) are also outlined.

model to produce absent keyphrases from an output vocabulary, while some of these might actually be reconstructed from the source document, is arguably overkill. Here, we argue that this may be one reason behind the poor performance of current sequence-to-sequence models in generating absent keyphrases (Gallina et al., 2020).

In this paper, we advocate for a stricter definition of absent keyphrases and propose a fine-grained categorization scheme that reflects how many new words are introduced within each keyphrase. Through this scheme, we shed new light on the effect of absent keyphrases on document retrieval effectiveness, and provide insights as to why current models for keyphrase generation are unable to accurately produce absent keyphrases. As a by-product, we introduce a new benchmark dataset for scientific document retrieval through the task of context-aware citation recommendation, that is composed of 169 manually extracted queries with relevance judgments and a collection of over 100K documents on topics related to IR.

2   (Re)defining Absent Keyphrases

Telling absent and present keyphrases apart may seem quite easy at first, but actually there are several intricacies to the process that should be noted. Starting from Meng et al. (2017)'s definition, "*we denote phrases that do not match any contiguous subsequence of source text as absent keyphrases, and the ones that fully match a part of the text as present keyphrases*", it is apparent that simple string matching between keyphrases and source document is not acceptable since it produces false positives (e.g. "*supervised learning*" matches

"*unsupervised learning*"). Instead, token-level sequence matching is to be used and combined with stemming to deal with different inflectional forms of the same word. Using stemming is critical here since it is carried out as a standard procedure in indexing documents for IR, but also in evaluating the precision of keyphrase generation models against gold standard annotations (Hasan and Ng, 2014).

Looking back at our example in Figure 1, we see that absent keyphrases can be further divided into three sub-categories depending on the proportion of present words they contain. Indeed, some absent keyphrases have some, or even all, of their constituent words (in stemmed forms) present in the text, while others are composed entirely of unseen words. Accordingly, we propose the following fine-grained categorization scheme (illustrated with the example from Figure 1 and explained in more depth with pseudo-code in Appendix A):

Present: keyphrases that match contiguous sequences of words in the source document (e.g. "*Search System*").

Reordered: keyphrases whose constituent words occur in the source document but not as contiguous sequences (e.g. "*Information Sharing*").

Mixed: keyphrases from which some, but not all, of their constituent words occur in the source document (e.g. "*Information Retrieval*").

Unseen: keyphrases whose constituent words do not occur in the source document (e.g. "*Retrieval Support*").

In contrast to the previously-used binary classification (i.e. present or absent), this finer-grained

categorization scheme draws a distinction between keyphrases that expand the document (i.e. mixed and unseen) and those that don't (i.e. present and reordered). It thus allows us to better understand how keyphrases affect the retrieval process by making it possible to numerically quantify the contribution of each category to the overall retrieval effectiveness. At the same time, this scheme provides a new angle to evaluate the ability of keyphrase generation models to output absent keyphrases by contrasting their PRMU distributions against those observed in the gold standard annotations. In other words, a model has to mimic the distribution of absent keyphrases in manual annotation in order to perform well.

3 Experiments

Here, we outline our experimental setup (§3.1), examine the distribution of keyphrases in commonly-used datasets with respect to the proposed categorization scheme (§3.2), show the influence of each category on the retrieval effectiveness (§3.3), and explore how these categories fit into the outputs of neural keyphrase generation models (§3.4).

3.1 Experimental settings

Experiments in *ad-hoc* document retrieval are carried out on the NTCIR-2 test collection (Kando, 2001) which is, to our knowledge, the only available benchmark dataset for that task. It includes 322,058 scientific abstracts in English annotated with author-assigned keyphrases (4.8 per doc. on avg.), and 49 search topics (queries) with relevance judgments. Documents cover a wide range of domains from pure science to humanities, although half of the documents are about computer science.

Given the rather limited size of the NTCIR-2 test collection, we conducted additional experiments in context-aware citation recommendation (He et al., 2010) which is the task of retrieving citations (documents) for a given text (query). Since no publicly available keyphrase-annotated collection exists for that task, we created one by collecting documents (BIBTEX entries) from the ACM Digital Library. Our dataset contains 102,411 documents in English on topics related to IR[1], most of which (69.2%) have author-assigned keyphrases (4.5 per doc. on avg.). We then followed the methodology proposed in (Roy, 2017), and selected 30 open-access sci-

entific papers[2] from which we manually extracted 169 citation contexts (queries) and 481 cited references (relevant documents). The resulting dataset, named ACM-CR, is publicly available[3].

For both retrieval tasks, we rank documents against queries using the standard BM25 model implemented in the Anserini[4] open-source IR toolkit (Yang et al., 2017), on top of which we apply the RM3 query expansion technique (Abdul-Jaleel et al., 2004) 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. We evaluate retrieval effectiveness in terms of mean average precision (mAP) on the top 1,000 retrieved documents for *ad-hoc* document retrieval, and in terms of recall at 10 retrieved documents for context-aware citation recommendation as recommended in (Färber and Jatowt, 2020). We use the Student's paired t-test to assess statistical significance of our retrieval results at $p < 0.05$ (Smucker et al., 2007).

		absent keyphrases			
Dataset	%P	%R	%M	%U	%uw
NTCIR-2	61.9	8.1	16.5	13.5	21.4
ACM-CR	53.6	11.7	19.3	15.4	25.5
KP20k	60.2	9.5	15.4	15.0	22.3
	term-weighting		doc. expansion		

Table 1: Proportion of Present, Reordered, Mixed and Unseen keyphrases in datasets. We also report the ratio of unique, unseen words in M+U keyphrases (%uw).

3.2 Distribution of gold-standard keyphrases under the PRMU scheme

Table 1 shows the proportion of gold-standard, author-assigned keyphrases for each category in the different datasets. We also report results for the KP20k dataset (Meng et al., 2017), which is used as training data by most neural keyphrase generation models. We observe very similar distributions across datasets, with absent keyphrases accounting for about 40% of the total number of keyphrases. Interestingly, most of the absent keyphrases belong to the mixed and unseen categories, and therefore

---

[1] We use the SIGs IR, KDD, CHI, WEB and MOD sponsored conferences and journals as a means to filter documents.

[2] Papers published in SIGIR, CHIIR, ICTIR or WSDM 2020 conferences.

[3] https://github.com/boudinfl/redefining-absent-keyphrases/blob/main/data/acm-cr/acm-cr.v1.tar.gz

[4] http://anserini.io/

	⌈ NTCIR-2 (mAP) ⌉			⌈ ACM-CR (recall@10) ⌉		
index	BM25	+RM3	#kp	BM25	+RM3	#kp
title & abstract	29.55	32.83	-	35.64	34.09	-
+ Present	30.74†	33.47	2.9	36.02	34.09	2.4
+ Reordered	29.79	33.48	0.4	35.43	33.40	0.5
+ Mixed	**30.80**†	33.85	0.8	36.22	33.41	0.9
+ Unseen	29.67	**33.94**	0.7	**36.24**	33.78	0.8
+ Absent (R+M+U)	30.77†	34.87†	1.9	36.62	34.10	2.1
+ Highlight (P+R)	30.64†	33.82	3.3	35.82	32.36	2.9
+ Expand (M+U)	**30.83**†	**34.34**	1.5	**37.21**	33.38	1.6
+ all (P+R+M+U)	31.92†‡	35.48†‡	4.8	36.65	32.88	4.5

Table 2: Retrieval effectiveness of BM25 and BM25+RM3 using various indexing configurations. We also report the average number of keyphrases (#kp). $\dagger$ and $\ddagger$ indicate significance over title & abstract indexing and Present, respectively.

should provide some form of semantic expansion. To have a precise idea of how many new words are actually added when indexing absent keyphrases, we compute the ratio (%uw) of unique words from keyphrases that do not occur in their corresponding documents. We find that only about 20% of the words included in keyphrases contribute to expanding documents. This surprisingly low percentage indicates that absent keyphrases play a much smaller role on document expansion than previously thought. Yet, as we will see next, this small fraction of new words is behind much of the gains observed in retrieval effectiveness.

3.3 Effect of indexing PRMU keyphrases on retrieval effectiveness

Table 2 presents the results of retrieval models on documents supplemented with keyphrases from PRMU categories. We see that adding keyphrases systematically improves retrieval effectiveness on both datasets, but a closer look reveals that the largest gains are obtained with Mixed and Unseen keyphrases. This observation, combined with the fact that the number of Mixed and Unseen keyphrases is comparatively small (less than one on average), demonstrate that expanding documents is more effective than highlighting salient phrases for improving document retrieval performance. The higher scores achieved when combining Mixed and Unseen keyphrases, compared to when combining Present and Reordered keyphrases, further confirm this conclusion. Surprisingly, coupling query expansion (+RM3) with appending keyphrases yields

conflicting results, which we attribute to the narrow set of topics (all related to IR) in ACM-CR that limits the vocabulary mismatch problem and makes it sensitive to semantic drift. Another reason may be the incomplete nature of the relevance judgments, i.e. that do not include uncited, yet relevant documents. Here, the use of a co-cited probability metric as in (Livne et al., 2014) may bring some new insights.

Model	%P	%R	%M	%U	F@5
s2s+copy	96.9	1.3	0.9	0.9	24.0
s2s+corr	89.7	7.1	2.5	0.8	22.1

Table 3: Proportion of Present, Reordered, Mixed and Unseen at the top-5 keyphrases on NTCIR-2. The f-measure against gold standard is also reported (F@5).

3.4 Analysis of keyphrase generation outputs under the PRMU scheme

In this last experiment, we explore how the proposed categories fit into the outputs of neural keyphrase generation models. Table 3 shows the distributions over PRMU categories for two strong baseline models: *s2s+copy*, a sequence-to-sequence model with attention and copying mechanisms (Meng et al., 2017), and *s2s+corr* which extends the aforementioned model with a coverage mechanism (Chen et al., 2018). We observe that the output distributions are heavily skewed towards the Present category, indicating that the models have trouble producing keyphrases made

up of new words. Accordingly, the overall performance of these models is quite poor (about 20% in f-measure), and mainly capped by the number of present keyphrases in the gold standard. This advocates for more focus on training generative models to expand documents, rather than to imitate author-assigned annotation.

4 Related Work

Until recently, most previous models for predicting keyphrases were doing so by extracting the most salient noun phrases from documents (Hasan and Ng, 2014). Keyphrase extraction models are usually divided into supervised models that cast keyphrase extraction either as a binary classification problem (Turney, 2000; Witten et al., 1999; Hulth, 2003; Nguyen and Kan, 2007; Medelyan et al., 2009; Sterckx et al., 2016) or as a sequence labelling problem (Augenstein et al., 2017; Xiong et al., 2019; Alzaidy et al., 2019), and unsupervised models that rely predominantly on graph-based ranking approaches (Mihalcea and Tarau, 2004; Litvak and Last, 2008; Wan and Xiao, 2008; Bougouin et al., 2013; Tixier et al., 2016; Boudin, 2018). Note that none of these models can produce absent keyphrases.

A related line of research focuses on keyphrase assignment, that is, the task of selecting entries from a predefined list of keyphrases (i.e. a controlled vocabulary) (Leung and Kan, 1997; Dumais et al., 1998; Medelyan and Witten, 2006). Here, predicting keyphrases is treated as a multi-class classification task, and models can produce both present and absent keyphrases. Further in that direction is (Bougouin et al., 2016) that jointly performs keyphrase extraction and assignment using an unsupervised graph-based ranking model.

Also closely related to our work is previous research on document expansion (Tao et al., 2006; Efron et al., 2012), and particularly recent work on supplementing document indexing with automatically generated queries (Nogueira et al., 2019; Nogueira and Lin, 2019). These latter models augment texts with potential queries that, just as keyphrases, mitigate vocabulary mismatch and re-weight existing terms (Lin et al., 2020). On the term weighting side, recent work shows that deep neural language models, in this case BERT (Devlin et al., 2019), can be successfully applied to estimate document-specific term weights (Dai and Callan, 2020).

5 Conclusion

In this paper, we investigated the usefulness of absent keyphrases for document retrieval. We showed that the commonly accepted definition of absent keyphrases is not sufficiently explicit in the context of IR, and proposed a finer-grained categorization scheme that allows for a better understanding of their impact on retrieval effectiveness. Our code and data are publicly available at `https://github.com/boudinfl/redefining-absent-keyphrases`.

Acknowledgements

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.

Rabah Alzaidy, Cornelia Caragea, and C. Lee Giles. 2019. Bi-lstm-crf sequence labeling for keyphrase extraction from scholarly documents. In *The World Wide Web Conference*, WWW '19, page 2551–2557, New York, NY, USA. Association for Computing Machinery.

Isabelle Augenstein, Mrinal Das, Sebastian Riedel, Lakshmi Vikraman, and Andrew McCallum. 2017. SemEval 2017 task 10: ScienceIE - extracting keyphrases and relations from scientific publications. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 546–555, Vancouver, Canada. Association for Computational Linguistics.

Hareesh Bahuleyan and Layla El Asri. 2020. Diverse keyphrase generation with neural unlikelihood training. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5271–5287, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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.

Florian Boudin, Ygor Gallina, and Akiko Aizawa. 2020. Keyphrase generation for scientific document

retrieval. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1118–1126, Online. Association for Computational Linguistics.

Adrien Bougouin, Florian Boudin, and Béatrice Daille. 2013. TopicRank: Graph-based topic ranking for keyphrase extraction. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 543–551, Nagoya, Japan. Asian Federation of Natural Language Processing.

Adrien Bougouin, Florian Boudin, and Béatrice Daille. 2016. Keyphrase annotation with graph co-ranking. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2945–2955, Osaka, Japan. The COLING 2016 Organizing Committee.

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.

Wang Chen, Hou Pong Chan, Piji Li, and Irwin King. 2020. Exclusive hierarchical decoding for deep keyphrase generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1095–1105, Online. Association for Computational Linguistics.

Zhuyun Dai and Jamie Callan. 2020. Context-aware term weighting for first stage passage retrieval. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 1533–1536, New York, NY, USA. Association for Computing Machinery.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Susan Dumais, John Platt, David Heckerman, and Mehran Sahami. 1998. Inductive learning algorithms and representations for text categorization. In *Proceedings of the Seventh International Conference on Information and Knowledge Management*, CIKM '98, page 148–155, New York, NY, USA. Association for Computing Machinery.

Miles Efron, Peter Organisciak, and Katrina Fenlon. 2012. Improving retrieval of short texts through document expansion. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, page 911–920, New York, NY, USA. Association for Computing Machinery.

Michael Färber and Adam Jatowt. 2020. Citation recommendation: Approaches and datasets. *arXiv preprint arXiv:2002.06961*.

Corina Florescu and Cornelia Caragea. 2017. Position-Rank: An unsupervised approach to keyphrase extraction from scholarly documents. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1105–1115, Vancouver, Canada. Association for Computational Linguistics.

G. W. Furnas, T. K. Landauer, L. M. Gomez, and S. T. Dumais. 1987. The vocabulary problem in human-system communication. *Commun. ACM*, 30(11):964–971.

Ygor Gallina, Florian Boudin, and Béatrice Daille. 2020. Large-scale evaluation of keyphrase extraction models. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, JCDL '20, page 271–278, New York, NY, USA. Association for Computing Machinery.

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.

Qi He, Jian Pei, Daniel Kifer, Prasenjit Mitra, and Lee Giles. 2010. Context-aware citation recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, page 421–430, New York, NY, USA. Association for Computing Machinery.

Chien-yu Huang, Arlene Casey, Dorota Głowacka, and Alan Medlar. 2019. Holes in the outline: Subject-dependent abstract quality and its implications for scientific literature search. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*, CHIIR '19, page 289–293, New York, NY, USA. Association for Computing Machinery.

Anette Hulth. 2003. Improved automatic keyword extraction given more linguistic knowledge. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, pages 216–223.

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*.

Chi-Hong Leung and Wing-Kay Kan. 1997. A statistical learning approach to automatic indexing of controlled index terms. *J. Am. Soc. Inf. Sci.*, 48(1):55–66.

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

Jimmy Lin, Rodrigo Nogueira, and Andrew Yates. 2020. Pretrained transformers for text ranking: Bert and beyond.

Marina Litvak and Mark Last. 2008. Graph-based keyword extraction for single-document summarization. In *Coling 2008: Proceedings of the workshop Multi-source Multilingual Information Extraction and Summarization*, pages 17–24, Manchester, UK. Coling 2008 Organizing Committee.

Avishay Livne, Vivek Gokuladas, Jaime Teevan, Susan T. Dumais, and Eytan Adar. 2014. Citesight: Supporting contextual citation recommendation using differential search. In *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '14, page 807–816, New York, NY, USA. Association for Computing Machinery.

Olena Medelyan, Eibe Frank, and Ian H. Witten. 2009. Human-competitive tagging using automatic keyphrase extraction. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1318–1327, Singapore. Association for Computational Linguistics.

Olena Medelyan and Ian H. Witten. 2006. Thesaurus based automatic keyphrase indexing. In *Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '06, page 296–297, New York, NY, USA. Association for Computing Machinery.

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.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, Barcelona, Spain. Association for Computational Linguistics.

Thuy Dung Nguyen and Min-Yen Kan. 2007. Keyphrase extraction in scientific publications. In *Asian Digital Libraries. Looking Back 10 Years and Forging New Frontiers*, pages 317–326, Berlin, Heidelberg. Springer Berlin Heidelberg.

Rodrigo Nogueira and Jimmy Lin. 2019. From doc2query to docttttquery.

Rodrigo Nogueira, Wei Yang, Jimmy Lin, and Kyunghyun Cho. 2019. Document expansion by query prediction.

Dwaipayan Roy. 2017. An improved test collection and baselines for bibliographic citation recommendation. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, page 2271–2274, New York, NY, USA. Association for Computing Machinery.

T.y.s.s Santosh, Debarshi Kumar Sanyal, Plaban Kumar Bhowmick, and Partha Pratim Das. 2020. SaSAKE: Syntax and semantics aware keyphrase extraction from research papers. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5372–5383, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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.

Lucas Sterckx, Cornelia Caragea, Thomas Demeester, and Chris Develder. 2016. Supervised keyphrase extraction as positive unlabeled learning. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1924–1929, Austin, Texas. Association for Computational Linguistics.

Zhiqing Sun, Jian Tang, Pan Du, Zhi-Hong Deng, and Jian-Yun Nie. 2019. Divgraphpointer: A graph pointer network for extracting diverse keyphrases. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, page 755–764, New York, NY, USA. Association for Computing Machinery.

Tao Tao, Xuanhui Wang, Qiaozhu Mei, and ChengXiang Zhai. 2006. Language model information retrieval with document expansion. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 407–414, New York City, USA. Association for Computational Linguistics.

Antoine Tixier, Fragkiskos Malliaros, and Michalis Vazirgiannis. 2016. A graph degeneracy-based approach to keyword extraction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1860–1870, Austin, Texas. Association for Computational Linguistics.

Peter D. Turney. 2000. Learning algorithms for keyphrase extraction. *Inf. Retr.*, 2(4):303–336.

Xiaojun Wan and Jianguo Xiao. 2008. CollabRank: Towards a collaborative approach to single-document keyphrase extraction. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 969–976, Manchester, UK. Coling 2008 Organizing Committee.

Yansen Wang, Zhen Fan, and Carolyn Rose. 2020. Incorporating multimodal information in open-domain web keyphrase extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1790–1800, Online. Association for Computational Linguistics.

Ian H. Witten, Gordon W. Paynter, Eibe Frank, Carl Gutwin, and Craig G. Nevill-Manning. 1999. Kea: Practical automatic keyphrase extraction. In *Proceedings of the Fourth ACM Conference on Digital Libraries*, DL '99, page 254–255, New York, NY, USA. Association for Computing Machinery.

Lee Xiong, Chuan Hu, Chenyan Xiong, Daniel Campos, and Arnold Overwijk. 2019. Open domain web keyphrase extraction beyond language modeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5175–5184, Hong Kong, China. 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, page 1129–1132, New York, NY, USA. Association for Computing Machinery.

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.

Jing Zhao and Yuxiang Zhang. 2019. Incorporating linguistic constraints into keyphrase generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5224–5233, Florence, Italy. Association for Computational Linguistics.

A Computing fine-grained PRMU categories

The python code below (Figure 2) showcases how PRMU keyphrase categories proposed in this paper are computed. Note that `kw` and `doc` are preprocessed lists of tokens (lowercased and in stemmed forms), and that title is separated from the abstract using a special token in order to prevent the last word of the title to be contiguous with the first word of the abstract.

These functions are written to be an understandable reference, the code shared with this work uses a different implementation.

```python
def contains(small, big):
    # Checks whether `small` appear
    #  contiguously in `big`
    for i in range(
            len(big) - len(small) + 1):
        match_len = 0
        for j in range(len(small)):
            if big[i + j] == small[j]:
                match_len += 1
            else:
                break

        if match_len == len(small):
            # Every elements were found
            return True
    return False


def kw_category(kw, doc):
    if contains(kw, doc):
        return 'P'  # Present
    else:
        abs_words = [w for w in kw
                     if w not in doc]
        if len(abs_words) == 0:
            return 'R'  # Reordered
        elif len(abs_words) < len(kw):
            return 'M'  # Mixed
        elif len(abs_words) == len(kw):
            return 'U'  # Unseen
```

Figure 2: Python code computing the fine-grained PRMU category of a keyphrase (`kw`) with respect to a given document (`doc`).