# Transferable Keyword Extraction and Generation with Text-to-text Language Models

Piotr Pęzik[1,2][0000−0003−0019−5840],
Agnieszka Mikołajczyk[2][0000−0002−8003−6243],
Adam Wawrzyński[2][0000−0002−1698−2390],
Filip Żarnecki[2][0009−0005−1106−408X],
Bartłomiej Nitoń[3][0000−0003−3306−7650], and
Maciej Ogrodniczuk[3][0000−0002−3467−9424]

[1] University of Łódź, Faculty of Philology
[2] VoiceLab, NLP Lab
[3] Institute of Computer Science, Polish Academy of Sciences

**Abstract.** This paper explores the performance of the T5 text-to-text transfer-transformer language model together with some other generative models on the task of generating keywords from abstracts of scientific papers. Additionally, we evaluate the possibility of transferring keyword extraction and generation models tuned on scientific text collections to labelling news stories. The evaluation is carried out on the English component of the POSMAC corpus, a new corpus whose release is announced in this paper. We compare the intrinsic and extrinsic performance of the models tested, i.e. T5 and mBART, which seem to perform similarly, although the former yields better results when transferred to the domain of news stories. A combination of the POSMAC and InTechOpen corpus seems optimal for the task at hand. We also make a number of observations about the quality and limitations of datasets used for keyword extraction and generation.

**Keywords:** keyword extraction · T5 language model · POSMAC · Polish

## 1 Introduction

Author-provided keywords are one of the intrinsic features of scientific articles as a distinct genre of texts. Despite recent advances in information extraction, sets of typically 3 to 5 keywords continue to be widely used to improve automatic retrieval of articles indexed in bibliographic databases. Formally, such keywords are usually noun phrases of varying complexity which may be used verbatim or as variants or derivatives of the wording used in the running text of an article. Some keywords may also denote concepts or descriptors abstracted from the literal content of a text. One implication of this dual nature of scientific keywords is the fact that a purely extractive keyword generation method, which critically depends on the occurrence of keywords in text can rarely produce satisfactorily complete results. Some studies have even proposed a distinction between

Present Keyword Extraction (PKE) and Abstract Keyword Generation (AKG) as separate tasks to reflect those two aspects of labeling texts with keywords [8]. Although this distinction may help evaluate certain solutions optimized for either of these two tasks, it is quite clear that a successful approach to automatically assigning keywords to scientific papers should be both extractive and abstractive as both of these characteristics are used by authors to succinctly describe the topic and domain of such texts. Furthermore, the distinction between PKE and AKG is intrinsically vague as certain keywords are nominalizations or otherwise paraphrased or generalized variants of expressions used in the running text of an article.

This paper focuses on evaluating the performance of the T5 and mBART models on the task of KEG (Keyword Extraction and Generation) from English language scholarly texts. In the initial section of the paper, we discuss the availability of English-language datasets used for KEG and point out some of their peculiarities and limitations. We also introduce the POSMAC corpus, which we believe to be a valuable resource for KEG in English. The subsequent sections of the paper present the evaluation of the aforementioned models on the POSMAC corpus and an extrinsic corpus of news stories.

## 2 Overview of KEG datasets

The top section of Table 1 summarizes many openly available datasets proposed for different variants of keyword extraction and generation tasks. We briefly discuss this selection to show how significantly different such datasets can be and justify the choice of corpora used in this paper.

The NUS dataset [12] seems to be an ad hoc collection of 211 scholarly papers from a variety of domains with mostly extractive keywords assigned by 'student volunteers'.

**Table 1.** A selection of openly available KEG datasets.

| Dataset | Type | Documents | Words | Unique words |
|---|---|---|---|---|
| NUS [12] | Full text | 211 | 1 824 297 | 42 568 |
| SemEval2010 [4] | Full text | 244 | 2 345 689 | 53 923 |
| Inspec [5] | Abstracts | 2 000 | 287 908 | 17 653 |
| Krapivin [6] | Full text | 2 305 | 21 858 324 | 183 976 |
| KP20k [10] | Abstracts | 570 809 | 104 349 114 | 701 706 |
| OAGKX [7] | Abstracts | 22 674 436 | 4 237 931 192 | 18 959 687 |
| InTechOpen | Abstracts | 30 418 | 4 935 962 | 151 598 |
| POSMAC EN | Abstracts | 115 749 | 13 788 880 | 165 168 |
| OAG (AMiner) [19] | Abstracts | 100 000 | 19 252 115 | 699 638 |
| News200 | Articles | 200 | 99 081 | 680 |

The SemEval2010 corpus "consists of a set of 284 English scientific papers from the ACM Digital Library" which were restricted to three subdomains of computer science and a subset of economic papers [4]. The keywords in this dataset are separated into author- and reader-submitted phrases and they seem to be a mixture of abstractive and extractive descriptors.

The Inspec dataset [5] contains abstracts of 2,000 scientific papers representing two subdomains of computer science. The dataset features two sets of keyphrases, including mostly abstractive keywords from a closed-set vocabulary and uncontrolled, mostly extractive keywords.

Krapivin [6] is a similarly sized collection of 2,000 full articles restricted to the domain of computer science. The keywords were assigned to each paper by its authors and verified by reviewers.

The KP20K corpus is a collection of over 570 000 articles scientific articles also representing the domain of computer science. The keywords assigned to these full-length papers are mostly abstractive and provided by their authors.

The Open Academic Graph corpus [19] and its variant processed for the task of keyword extraction and generation (OAGKX) contain over 9 and 22 million abstracts respectively covering a wide variety of scholarly domains. As such, they may appear to be highly relevant to the task of developing and evaluating KEG solutions which render the remaining datasets described here largely insignificant. However, on closer inspection, the overall quality of the keyword assignments available in OAG/OAGKX calls its usefulness into question. First of all, this dataset seems to contain many automatic, low-quality keywords extracted from the text of abstracts. This can be verified by simply comparing the OAG keyword assignments with the corresponding originally published papers available online. In many cases, a single high-level keyword is assigned to a record and in other cases, dozens of keywords are assigned to an equally sized abstract. Additionally, the OAGKX 'edition' of OAG was tokenized with an NLP pipeline which removed all punctuation and casing. In short, although the inconsistent and occasionally clearly erroneous assignments of keywords may improve the retrieval of documents from this collection, it is not obvious whether OAG or OAGKX can be used to train or even evaluate a KEG classifier. To summarize, openly available KEG datasets differ significantly in terms of the number and size of documents (abstracts/ full texts), quality and type of keyword assignments (abstractive/ extractive), the average number of keywords per text, and the total number of distinct keywords. They are also typically restricted to a handful of domains and in some cases, the keywords they offer were not assigned by the authors or reviewers, but rather by volunteers or, even worse, by algorithms. This variability is further summarized in Table 2.

The bottom part of Table 1 lists the datasets used to evaluate the KEG models described in this paper. The most important of them is the English language subset of the newly released Polish Open Science Metadata Corpus (POSMAC) [13], which was developed in the CURLICAT project [4] [17]. The content of POS-

[4] `https://curlicat.eu/`

MAC was acquired from the Library of Science (LoS)[5], a platform providing open access to full texts of articles published in over 900 Polish scientific journals and selected scientific books with bibliographic metadata. Over 70% of the metadata records acquired have author-defined keywords. POSMAC combines high-quality keyword assignments with a fairly wide coverage of scholarly domains including non-technical disciplines such as humanities and social sciences.

In addition to this new resource, we compiled a corpus of over 30 000 abstracts of chapters published in open access books crawled from `https://www.intechopen.com/`. The corpus covers a variety of scholarly disciplines with high-quality keywords assigned by authors of the respective chapters.

The last collection used in this paper is a set of 200 recent news articles published on two websites (`http://euronews.com` and `http://wikinews.org`), whose topics range from health, politics to business and sports. We manually assigned a set of keywords to each of these articles to assess the transferability of KEG models trained on scholarly texts.

**Table 2.** Type of keyword assignment in selected KEG datasets.

| Dataset | Average keywords | Keyword types* | Unique KWs | Annotators |
|---|---|---|---|---|
| NUS | 11 | Extractive | 2 041 | Volunteers |
| SemEval2010 | 15.5 | Abstractive and extractive | 3 220 | Readers and authors |
| Inspec | 9.5 | Extractive | 16,16 | Professional indexers |
| Krapivin | 5 | Abstractive | 8 728 | Authors |
| KP20k | 5 | Abstractive | 760 652 | Authors |
| OAGKX | 4 | Unclear | 18 959 687 | Unclear |
| POSMAC EN | 4.5 | Abstractive | 198 102 | Authors |
| InTechOpen | 4.9 | Abstractive | 90,98 | Authors |
| OAG (AMiner) | 4 | Unclear | 250 899 | Unclear |

*The predominant type of keywords included.

## 3 Evaluation of text-to-text models for KEG

An increasing number of recent approaches to KGE follow the general trend to use deep neural architectures to address NLP problems, cf. GAN [15], TG-Net [3], Para-Net [20], catSeq [1], corrRNN [2], SetTrans [18], KEA. More specifically, transformer-based architectures have also been used to both extract and abstract keywords from scientific texts, cf. BERT-PKE [8], [11], KeyBART [7].

In this paper, we focus on applying generative language models, which have recently been successfully applied to a number of NLP tasks. The first of those

[5] `https://bibliotekanauki.pl/`

**Table 3.** Overall performance of evaluated models on new datasets of scientific and news texts.

| Model | Train set | POSMAC | | | News articles | | |
|---|---|---|---|---|---|---|---|
| | | **P** | **R** | **$F_1$** | **P** | **R** | **$F_1$** |
| mT5-base | POSMAC EN | 0.265 | 0.216 | 0.238 | 0.260 | 0.215 | 0.235 |
| mT5-base | POSMAC EN+InTechOpen | 0.276 | 0.224 | 0.248 | 0.249 | 0.204 | 0.224 |
| mBART-large | POSMAC EN+InTechOpen | 0.270 | 0.236 | **0.252** | 0.237 | 0.213 | 0.224 |
| mT5-large | POSMAC EN+InTechOpen | 0.286 | 0.223 | 0.250 | 0.275 | 0.222 | **0.246** |

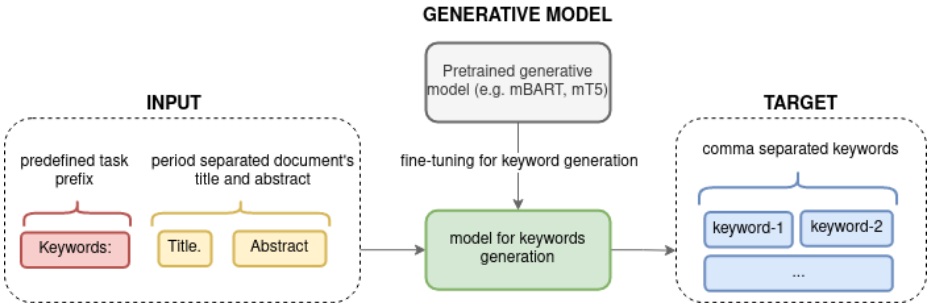


**Fig. 1.** Training procedure for mBART and Text-To-Text Transfer Transformer model for keywords generation.

models is known as T5 [14]. Although its architecture is based on the original encoder-decoder transformer implementation [16], it frames a wide variety of NLP problems as text-to-text operations, where both the input and output are text strings. In the experiments reported in this paper, the input to the mT5[6] variant of T5 is a concatenated title and abstract of a scientific abstract and the text string output is a comma-separated list of lemmatized single- or multiword keywords. For the KGE task at hand, we used an Adam optimizer with 100 warm-up steps, linearly increasing the learning rate from zero to a target of 3e-5. Additionally, we used a multiplicative scheduler that lowered the LR by 0.9 every epoch. The model was trained for ten epochs with a batch size of 32. The maximum input length was set to 512 tokens and the maximum target length was 128. We refer to the resulting KGE model as **mT5kw**. After experimenting with the `no_repeat_ngram_size` and `num_beams` parameters on the development subset of our corpora we found the optimal values of `no_repeat_ngram_size`=3 and `num_beams`=4. The general flow of the mT5 and mBART training procedure is shown in Figure 1.

We compare the results obtained with mT5 with the performance of a KEG model based on mBART, which is a de-noising auto-encoder model pretrained on multiple monolingual corpora [9]. As shown in Table 3 (which lists the average micro-precision and recall scores for each model) there is a noticeable advantage

[6] `https://huggingface.co/docs/transformers/model_doc/mt5`

in using the larger version of the mT5 compared with the base variant. Additionally, the mT5-based model trained on scholarly texts seems to transfer slightly better to the domain of news articles than mBART-large, for which we observed the highest $F_1$ score on the source domain of scientific abstracts. Since the two text-to-text models produced 3-5 keywords, there was no need to artificially limit the number of keywords produced by the model. Our qualitative evaluation of the results shows that many of the keywords absent from the gold set seem relevant to the abstract from the test set. One of the most interesting aspects of the mT5 model is its transferability to other domains. The overall results of this paper confirm the conclusions of a separate study (Anonymized et al. 2022), in which compare a selection of approaches to keyword extraction and generation (KEG) for Polish scientific abstracts and concludes that the T5 outperforms purely extractive and abstractive methods and that it is highly transferable to other domains, including transcripts of spoken language. Another clear advantage of T5 is its ability to learn the truecasing and lemmatization of assigned keyphrases, which is of particular value in morphologically complex languages.

## Acknowledgements

The work reported here was supported by 1) the European Commission in the CEF Telecom Programme (Action No: 2019-EU-IA-0034, Grant Agreement No: INEA/CEF/ICT/A2019/1926831) and the Polish Ministry of Science and Higher Education: research project 5103/CEF/2020/2, funds for 2020–2022) and 2) "CLARIN - Common Language Resources and Technology Infrastructure", which is part of the 2014-2020 Smart Growth Operational Programme, POIR.04.02.00-00C002/19.

## References

1. Chan, H.P., Chen, W., Wang, L., King, I.: Neural keyphrase generation via reinforcement learning with adaptive rewards. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 2163–2174. Association for Computational Linguistics, Florence, Italy (Jul 2019). https://doi.org/10.18653/v1/P19-1208, https://aclanthology.org/P19-1208
2. Chen, J., Zhang, X., Wu, Y., Yan, Z., Li, Z.: Keyphrase generation with correlation constraints. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 4057–4066. Association for Computational Linguistics, Brussels, Belgium (Oct-Nov 2018). https://doi.org/10.18653/v1/D18-1439, https://aclanthology.org/D18-1439
3. Chen, W., Gao, Y., Zhang, J., King, I., Lyu, M.R.: Title-guided encoding for keyphrase generation. CoRR **abs/1808.08575** (2018), http://arxiv.org/abs/1808.08575
4. Hendrickx, I., Kim, S.N., Kozareva, Z., Nakov, P., Ó Séaghdha, D., Padó, S., Pennacchiotti, M., Romano, L., Szpakowicz, S.: SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In: Proceedings of the 5th International Workshop on Semantic Evaluation. pp. 33–38. Association for

Computational Linguistics, Uppsala, Sweden (Jul 2010), https://aclanthology.org/S10-1006
5. Hulth, A.: Improved automatic keyword extraction given more linguistic knowledge. In: Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing. p. 216–223. EMNLP '03, Association for Computational Linguistics, USA (2003). https://doi.org/10.3115/1119355.1119383, https://doi.org/10.3115/1119355.1119383
6. Krapivin, M., Autaeu, A., Marchese, M.: Large dataset for keyphrase extraction (2008), http://eprints.biblio.unitn.it/1671/1/disi09055-krapivin-autayeu-marchese.pdf, University of Trento, Dipartimento di Ingegneria e Scienza dell'Informazione, Technical Report # DISI-09-055
7. Kulkarni, M., Mahata, D., Arora, R., Bhowmik, R.: Learning rich representation of keyphrases from text. CoRR **abs/2112.08547** (2021), https://arxiv.org/abs/2112.08547
8. Liu, R., Lin, Z., Wang, W.: Keyphrase prediction with pre-trained language model. CoRR **abs/2004.10462** (2020), https://arxiv.org/abs/2004.10462
9. Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., Zettlemoyer, L.: Multilingual denoising pre-training for neural machine translation. CoRR **abs/2001.08210** (2020), https://arxiv.org/abs/2001.08210
10. Meng, R., Zhao, S., Han, S., He, D., Brusilovsky, P., Chi, Y.: Deep keyphrase generation. CoRR **abs/1704.06879** (2017), http://arxiv.org/abs/1704.06879
11. Meng, R., Zhao, S., Han, S., He, D., Brusilovsky, P., Chi, Y.: Deep keyphrase generation. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 582–592. Association for Computational Linguistics, Vancouver, Canada (Jul 2017). https://doi.org/10.18653/v1/P17-1054, https://aclanthology.org/P17-1054
12. Nguyen, T.D., Kan, M.Y.: Keyphrase extraction in scientific publications. In: Goh, D.H.L., Cao, T.H., Sølvberg, I.T., Rasmussen, E. (eds.) Asian Digital Libraries. Looking Back 10 Years and Forging New Frontiers. pp. 317–326. Springer Berlin Heidelberg, Berlin, Heidelberg (2007)
13. Pęzik, P., Mikołajczyk, A., Wawrzyński, A., Nitoń, B., Ogrodniczuk, M.: Keyword extraction from short texts with a text-to-text Transfer Transformer. In: Szczerbicki, E., Wojtkiewicz, K., Nguyen, S.V., Pietranik, M., Krótkiewicz, M. (eds.) ACIIDS 2022: Recent Challenges in Intelligent Information and Database Systems. pp. 530–542. No. 1716 in Communications in Computer and Information Science (CCIS), Springer Nature Singapore (2022). https://doi.org/10.1007/978-981-19-8234-7_41
14. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. Journal of Machine Learning Research **21**(140), 1–67 (2020), http://jmlr.org/papers/v21/20-074.html
15. Swaminathan, A., Gupta, R.K., Zhang, H., Mahata, D., Gosangi, R., Shah, R.R.: Keyphrase generation for scientific articles using gans. CoRR **abs/1909.12229** (2019), http://arxiv.org/abs/1909.12229
16. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is All you Need. In: Guyon, I., von Luxburg, U., Bengio, S., Wallach, H.M., Fergus, R., Vishwanathan, S.V.N., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30: Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS 2017). pp. 5998–6008 (2017), https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

17. Váradi, T., Nyéki, B., Koeva, S., Tadić, M., Štefanec, V., Ogrodniczuk, M., Nitoń, B., Pęzik, P., Mititelu, V.B., Irimia, E., Mitrofan, M., Pais, V., Tufiș, D., Garabík, R., Krek, S., Repar, A.: Introducing the CURLICAT corpora: Seven-language domain specific annotated corpora from curated sources. In: Proceedings of the Language Resources and Evaluation Conference. pp. 100–108. European Language Resources Association, Marseille, France (2022), https://aclanthology.org/2022.lrec-1.11
18. Ye, J., Gui, T., Luo, Y., Xu, Y., Zhang, Q.: One2Set: Generating diverse keyphrases as a set. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 4598–4608. Association for Computational Linguistics, Online (Aug 2021). https://doi.org/10.18653/v1/2021.acl-long.354, https://aclanthology.org/2021.acl-long.354
19. Zhang, F., Liu, X., Tang, J., Dong, Y., Yao, P., Zhang, J., Gu, X., Wang, Y., Shao, B., Li, R., Wang, K.: Oag: Toward linking large-scale heterogeneous entity graphs. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery amp; Data Mining. p. 2585–2595. KDD '19, Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3292500.3330785, https://doi.org/10.1145/3292500.3330785
20. Zhao, J., Zhang, Y.: Incorporating linguistic constraints into keyphrase generation. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 5224–5233. Association for Computational Linguistics, Florence, Italy (Jul 2019). https://doi.org/10.18653/v1/P19-1515, https://aclanthology.org/P19-1515