

Keyword Extraction from Short Texts with a Text-To-Text Transfer Transformer

Piotr Pęzik^{1,2}[0000-0003-0019-5840],
Agnieszka Mikołajczyk²[0000-0002-8003-6243],
Adam Wawrzyński²[0000-0002-1698-2390],
Bartłomiej Niton³[0000-0003-3306-7650], and
Maciej Ogrodniczuk³[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. The paper explores the relevance of the Text-To-Text Transfer Transformer language model (T5) for Polish (pT5) to the task of intrinsic and extrinsic keyword extraction from short text passages. The evaluation is carried out on the new Polish Open Science Metadata Corpus (POSMAC), which is released with this paper: a collection of 216,214 abstracts of scientific publications compiled in the CURLICAT project. We compare the results obtained by four different methods, i.e. pT5kw, extremeText, TermOPL, KeyBERT and conclude that the pT5kw model yields particularly promising results for both frequent and sparsely represented keywords. Furthermore, a pT5kw keyword generation model trained on the POSMAC also seems to produce highly useful results in cross-domain text labelling scenarios. We discuss the performance of the model on news stories and phone-based dialog transcripts which represent text genres and domains extrinsic to the dataset of scientific abstracts. Finally, we also attempt to characterize the challenges of evaluating a text-to-text model on both intrinsic and extrinsic keyword extraction.

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

1 Keyword Extraction and Generation

The main NLP problem discussed in this paper can be described as keyword extraction or generation from short text passages. More specifically, given a span of text such as a concatenated title and abstract of a research paper, the task is to generate a small set of words or multiword phrases (usually nominal phrases) which succinctly describe its content. Approaches to this problem can be purely *extractive* or partly *abstractive*. In the former case, keywords are extracted and possibly normalized more or less directly from the text of the sample. Abstractive methods can assign labels that may not have occurred in the original text sample or a restricted vocabulary of known keywords.

Despite the long tradition of keyword extraction as a distinct NLP task, ”no single approach (...) effectively extracts keywords from different data sources” [5]⁴. Extraction of keywords from longer texts can be approached similarly to term extraction and partly facilitated by considering word frequency distributions and identifying significantly frequent phrases as potential keywords. This paper focuses mostly on showcasing the applications of the T5 generative language model [13] and comparing its performance to text classification (extremeText / fastText) [17][10] and statistical terminology extraction (C / NC-values) [7] as baseline methods. Although the complementarity of statistical and transformer-based approaches to keyword extraction has been explored before [8], we are not aware of any published assessment of text-to-text generative models on this task.

From the point of view of model evaluation, it should be noted that the manual assignment of keywords to scientific articles, which are used as groundtruth annotations in our analysis, is far from deterministic and quite different from text classification or labelling based on a closed-set taxonomy. Authors usually draw a small set of terms from a largely uncontrolled vocabulary. Such descriptors may be terminological items used in the text of their paper, but they can also be more abstract or at least hyperonymic descriptors of its content. Synonyms, hyperonyms and abbreviated forms contribute to the apparent sparsity of the vocabulary, which over time tends to grow in a large collection of abstracts at a sublinear rate. This in turn has implications for building and evaluating automatic keyword extraction solutions. Firstly, the distribution of keywords as distinct class labels in many datasets is rather sparse, which means that the recall of rare keywords is unlikely to be high, at least in any supervised text classification scenario. Secondly, the evaluation of automatically assigned keywords is problematic as the ‘ground truth’ assignments are neither consistent nor exhaustive. The latter problem could be systematically addressed by measuring inter-rater agreement in datasets which are explicitly developed for NLP purposes. However, the corpus of scientific abstracts used in this study has been adapted from metadata sources which were not globally curated and checked for consistency.

Despite these methodological limitations, we believe that our evaluation of keyword generation approaches provides fresh insights into the transferability of a T5 model to loosely related topical domains and text genres. As shown in the last section of this paper, a model tuned on a high-quality corpus of scientific abstract extracts surprisingly accurate keywords from news stories and even spoken dialogue transcripts. Thus, the novelty of this paper consists in the fact that a) we test the relevance of text-to-text transfer transformers to the task of keyword generation and b) we evaluate and release a non-obvious dataset which shows significant potential of transferability to extrinsic domains and languages.

⁴ This paper also offers an up-to-date review of keyword extraction methods.

2 The Polish Open Science Metadata Corpus

The source dataset used in this study was developed in CURLICAT⁵, an international project aimed at delivering rich metadata monolingual corpora in seven EU languages, including Polish, and representing different topical domains and text genres. The Polish subset of CURLICAT (released for the first time with this paper) named Polish Open Science Metadata Corpus (POSMAC)⁶ — contains a new source of valuable corpus data acquired from the Library of Science (LoS)⁷, a platform providing open access to full texts of articles published in over 900 Polish scientific journals and full texts of selected scientific books together with extensive bibliographic metadata. More than 70 % of the metadata records included in the resulting corpus contain keywords describing the content of the indexed articles. Since authors of the respective articles typically enter such keywords themselves, their selection is relatively uncontrolled. After lowercasing and ASCII-folding (i.e., removing Polish diacritics due to their inconsistent use), we found a total of 256 139 distinct keywords used in the corpus, with only 69 266 (ca. 27%) used more than once and 10 074 keywords assigned to 10 or more articles. Syntactically, the vast majority of keywords are lemmatized noun phrases whose length typically ranges from 1 to 3 words (mean=2.39, sd=1.22). A single article record is tagged with an average of 4.76 keywords (median=4).

Table 1. Top 10 scientific domains represented in the POSMAC.

Domains	Documents	With keywords
Engineering and technical sciences	58 974	57 165
Social sciences	58 166	41 799
Agricultural sciences	29 811	15 492
Humanities	22 755	11 497
Exact and natural sciences	13 579	9 185
Humanities, Social sciences	12 809	7 063
Medical and health sciences	6 030	3 913
Medical and health sciences, Social sciences	828	571
Humanities, Medical and health sciences, Social sciences	601	455
Engineering and technical sciences, Humanities	312	312

⁵ <https://curlicat.eu/>

⁶ <http://clip.ipipan.waw.pl/POSMAC>

⁷ <https://bibliotekanauki.pl/>

3 Approaches

3.1 T5, plT5 and plT5kw

T5 stands for the Text-To-Text Transfer Transformer model proposed by [13]. In terms of its architecture, the model is based on the original encoder-decoder transformer implementation [16]. Unlike popular transformer-based language models used in classification tasks, T5 frames all NLP problems as text-to-text operations, where both the input and output are text strings. Although this approach may only seem natural for selected NLP problems such as question answering, translation or summarization, it has been demonstrated to apply to other tasks such as classification or regression tasks. In this study, the input to a T5 model is a concatenated title and abstract of a scientific paper and the text string output is a comma-separated list of lemmatized single- or multiword ‘keywords’⁸. In the case of morphologically rich languages such as Polish, such lemmatization may additionally involve number, case and gender agreement operations on the resulting multiword keywords. This requirement is particularly important for out-of-vocabulary (OOV) keywords which need to be lemmatized and formatted on demand.

For the extraction of Polish keywords, we used the plT5-base model [3]⁹ trained on six reference corpora of Polish. More specifically, we train the model to predict comma-separated keywords from article abstracts concatenated with titles. 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 0.7 every epoch. We trained the model for ten epochs with a batch size of 8. The maximum input length was set to 512 tokens and the maximum target length was 128. We refer to the resulting keyword extraction model as **plT5kw**.

We experimented with `no_repeat_ngram_size` and `num_beams` parameters on *dev* subset of datasets to find out the best configuration. During evaluation on the test subset, we set `no_repeat_ngram_size` to 3 and `num_beams` to 4.

3.2 FastText and extremeText

FastText is a popular text classification library which uses vector representations of (sub)words as input to a relatively simple neural network [10]. Despite the obvious differences between supervised text classification and unsupervised keyword extraction, the assignment of keywords attested in a representative collection of tagged texts can be treated as a text labelling task. In the comparison described in this paper we used an extension of FastText called extremeText [17], which uses a Probabilistic Labels Tree loss function (PLT) to optimize the assignment of labels from very large taxonomies, such as the set of over 200 000 distinct keywords found in POSMAC. Using the PLT loss function, we trained

⁸ We use the traditional term *keyword* to refer to potentially multiword phrases found in the *Keywords* section of a scientific abstract.

⁹ <https://huggingface.co/allegro/plt5-large>

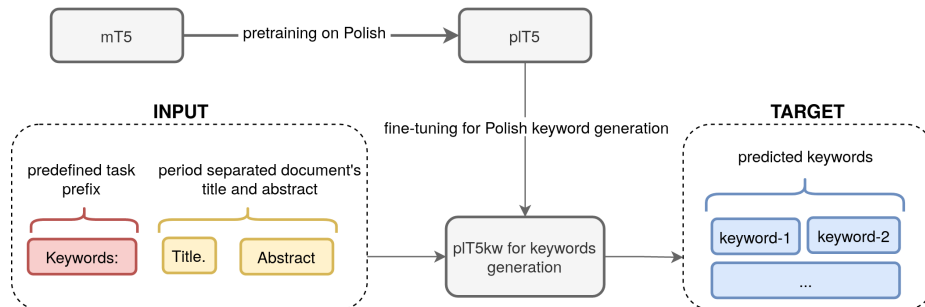


Fig. 1. Training procedure for our Text-To-Text Transfer Transformer model for keywords generation.

a keyword classifier with 300 dimensions in the hidden layer for 50 epochs to obtain the results reported below.

3.3 TermoPL

TermoPL [11] is a statistical terminology extraction tool designed to identify recurrent words and multiword combinations in domain corpora of Polish. It identifies, lemmatizes and scores recurrent noun phrases as potential terminological items using a ranking function proposed by [7]. We include this approach to measure the upper bound of purely extractive keyword identification. In other words, we estimate the maximum recall of simply extracting and lemmatizing all noun phrases contained in any given abstract.

3.4 KeyBERT

KeyBERT [9] is a keyword extraction library utilizing BERT representations. For each document it creates a representation vector using a transformer-based language model. Next, word representations for each n-gram found in a given text are compared with the document vector using cosine vector distance scores. The most similar phrases are selected as those that represent the document content. Additionally, two methods are used to increase the diversity of the generated phrases: Maximum Sum Similarity and Maximal Marginal Relevance. We used KeyBERT with the *distiluse-base-multilingual-cased-v1* model from the Sentence Transformers library [14]. The filtering method used in the experiments was Maximal Margin Relevance, and the diversity factor was set to 0.7 with an n-gram range of (1,2). The other parameters were used with their default values.

4 Intrinsic Evaluation

To evaluate the above-mentioned set of complementary keyword generation approaches intrinsically (i.e. on the original POSMAC dataset), abstracts annotated with keywords were split into a training and test set with a ratio of 70/30%.

To ensure consistent distribution of labels in the training and test set, we used an implementation of the iterative stratification algorithm¹⁰ for multilabel data, originally proposed by [15]. The relevance and coverage of keyword assignments are evaluated in terms of micro- and macro- precision and recall values, as well as their harmonic means (F_1 -scores) averaged over the documents in the test set. These scores are measured at several ranks ($k=1, 3, 5$ and more) for each approach in two different scenarios: a) using the full set of keywords assigned in the training and test set and b) training and/or evaluating only on keywords which occur at least 10 times in the stratified dataset.

It should be noted that in addition to evaluating the four main approaches described in this paper, we also assessed several baseline keyword extraction approaches, including FirstPhrases and TopicRank [2], PositionRank [6], MultipartiteRank [1], TextRank [12], KPMiner [4] and TfIdf with some adjustments aimed at boosting their performance (such as lemmatizing input text). The results obtained for all of those methods were less than 0.025 F_1 on all ranks, which is why we excluded them from the detailed comparisons included below.

Table 2 compares the performance of extremeText and plT5kw on the task of extracting keywords from the full set of more than 200 000 items found in POSMAC. The highest average F_1 (harmonic mean of precision and recall) is obtained for both approaches at rank 5, although plT5kw clearly outperforms extremeText on this task both in terms of precision and recall at all the corresponding ranks. We include rank 10 for the extremeText classifier but not for plT5kw, because the former can be requested to produce a ranked list of keywords used in the training set of any length while the latter is implicitly trained to produce text strings with up to 4 commas on average. The results for KeyBERT for Polish keyword extraction are very poor both in terms of recall and precision, neither of which increases over 0.03 at any rank measured (see the qualitative explanations below). As signalled above, TermoPL is meant to be used on longer texts as it ranks terms by a score which is partly derived from their frequency in a sufficiently large corpus of texts. Nevertheless, looking at its recall when no rank limit is applied, it is interesting to observe that more than 33 percent of all keywords in our corpus are actually some form of nominal phrases found in the text of the abstracts.

¹⁰ See <https://vict0rs.ch/2018/05/24/sample-multilabel-dataset/>.

Table 2. Results of evaluation on the full set of POSMAC keywords. The upper part of the table presents the results for all keywords present in the dataset, while the lower part refers to experiments conducted with the rejection of keywords occurring less than 10 times.

Method	Rank	Micro			Macro		
		P	R	F ₁	P	R	F ₁
extremeText	1	0.175	0.038	0.063	0.007	0.004	0.005
	3	0.117	0.077	0.093	0.011	0.011	0.011
	5	0.090	0.099	0.094	0.013	0.016	0.015
	10	0.060	0.131	0.082	0.015	0.025	0.019
plT5kw	1	0.345	0.076	0.124	0.054	0.047	0.050
	3	0.328	0.212	0.257	0.133	0.127	0.129
	5	0.318	0.237	0.271	0.143	0.140	0.141
KeyBERT	1	0.030	0.007	0.011	0.004	0.003	0.003
	3	0.015	0.010	0.012	0.006	0.004	0.005
	5	0.011	0.012	0.011	0.006	0.005	0.005
TermoPL	1	0.118	0.026	0.043	0.004	0.003	0.003
	3	0.070	0.046	0.056	0.006	0.005	0.006
	5	0.051	0.056	0.053	0.007	0.007	0.007
	all	0.025	0.339	0.047	0.017	0.030	0.022
extremeText	1	0.210	0.077	0.112	0.037	0.017	0.023
	3	0.139	0.152	0.145	0.045	0.042	0.043
	5	0.107	0.196	0.139	0.049	0.063	0.055
	10	0.072	0.262	0.112	0.041	0.098	0.058
plT5kw	1	0.377	0.138	0.202	0.119	0.071	0.089
	3	0.361	0.301	0.328	0.185	0.147	0.164
	5	0.357	0.316	0.335	0.188	0.153	0.169
KeyBERT	1	0.018	0.007	0.010	0.003	0.001	0.001
	3	0.009	0.010	0.009	0.004	0.001	0.002
	5	0.007	0.012	0.009	0.004	0.001	0.002
TermoPL	1	0.076	0.028	0.041	0.002	0.001	0.001
	3	0.046	0.051	0.048	0.003	0.001	0.002
	5	0.033	0.061	0.043	0.003	0.001	0.002
	all	0.021	0.457	0.040	0.004	0.008	0.005

The lower part of Table 2 shows the results of evaluating the four approaches on a set of 10,083 distinct keywords which were assigned to at least 10 different abstracts in the stratified dataset. As expected, the results for extremeText are slightly better in this run as they are not affected by very rare or unknown

keywords in the test set. However, the tuned plT5kw is again significantly better in this scenario.

Table 3 reveals an interesting property of plT5kw. Its precision increases up to 0.425 at rank 1 when the predictions are limited to keywords found in the training set. The improvement in precision comes at the expense of recall, which drops by nearly 10 percentage points. This observation also means that unlike extremeText or any text classification model plT5kw is capable of assigning relevant keywords which were not seen in the training set. The transferability of plT5kw is further discussed in the next section of this paper.

Table 3. Evaluation of plT5kw on the set of keywords found in the training set.

Method	Rank	Micro			Macro		
		P	R	F ₁	P	R	F ₁
plT5kw	1	0.425	0.093	0.153	0.086	0.074	0.080
	3	0.415	0.212	0.281	0.165	0.158	0.161
	5	0.412	0.227	0.293	0.172	0.167	0.169

5 Transfer to Other Domains

Although the results of intrinsic evaluation of keyword extraction from scientific abstract reported above may seem moderately useful, the plT5kw model trained on a rather narrowly defined source domain seems to produce surprisingly precise (although incomplete) keywords for samples of other topical domains and text genres. In this section, we explore the relevance of the plT5kw model trained on POSMAC to the domain of news stories and transcripts of conversational speech. We also compare the type of keywords produced by the four extraction approaches discussed above in more qualitative terms.

5.1 News Stories

Table 4 shows a set of shorthand English translations of headlines of recent news stories published in Polish web-based media outlets. The full text of each story is linked to the shorthand headline. The next four columns of the table show samples of keywords generated for the full text of each article by the four respective extraction methods described in this paper.

The overall quality of the extracted keywords can be considered from a number of perspectives. The **importance** of an extracted keyword (also known as **keyness**) refers to its potential to express the most important aspects of a text passage. Although all important keywords are also **relevant** (related to the content of the text sample), not all relevant keywords are equally important and usually some limit on the **complete** set of relevant keywords is required. The **abstraction** aspect of keywords pertains to the degree to which they can

describe the content without necessarily relying on the verbatim word combinations used in a given text passage. The **transferability** of a keyword generation method refers to its ability to produce good quality keywords for texts from domains which are different from those of the originally labelled datasets. Finally, the **formatting** quality of a keyword refers to the correct lemmatization, true-casing or abbreviation of the extracted keywords.

The transferability of the keywords produced by **extremeText**, a closed-label set classifier is clearly limited. The predictions are far from complete and they are only remotely relevant to the texts from the news domains in cases where some overlap exists between the original domain of scientific abstracts and a given news story. The labels are static, i.e. they are not dynamically reformatted or adjusted to the text samples.

The results produced by the **KeyBERT** model are probably the least convincing in this comparison. The model shows a clear preference for longer n-grams, which may not be syntactically complete nominal phrases or any regular phrases for that matter. The results are not lemmatized or properly cased, but their relevance is at least relatively easy to judge as they can be traced back to the exact span of text from which they were extracted.

As a terminology extraction solution **TermoPL** produces lemmatized although not always correctly cased noun phrases. Some of the results are clearly complementary to the keywords produced by plT5kw, but the choice of the most important items remains an issue as this solution requires a larger body of text to score its suggestions. The solution is domain-independent, but this also means that it does not transfer any knowledge about the desired keyword format or level of abstraction from other domains.

The keywords produced by **plT5kw** are mostly relevant and well abstracted although occasionally also too generic. For example the meeting of the Polish PM with the German Chancellor is tagged as *Poland* and *Germany* and the story about a gas explosion in Sicily gets the rather generic tag of *Italy*. The model does not seem to have a bias toward single- or multiword expressions. It produces correctly lemmatized, syntactically agreed and well-cased phrases. It seems to have transferred the skill of identifying, formatting and sometimes abstracting comma-separated nominal phrases without over-fitting excessively to the topics represented in the source domain. Needless to say, the recall of the model is far from perfect, but its precision is reliably stable.

5.2 Customer Support Dialogues

The promising results obtained on a sample corpus of news stories have led us to test plT5kw on the completely different domain of phone conversation transcripts which were sampled from the DiaBiz corpus.¹¹ The following excerpt from the DiaBiz corpus is a translation of a dialogue between a support agent informing a client about an outstanding electricity bill which has resulted in an

¹¹ DiaBiz is a corpus developed in the CLARIN-Biz project. It contains some 4,000 phone-based customer support calls covering a range of topics and business processes.

Table 4. Comparison of keywords generated for 5 news stories.

#	Short title	PI15kw	KeyBERT	TermPL	ExtremeText
1	Polish PM meets new German Chancellor	Niemcy, Polska, Unia Europejska, Nord Stream 2, bezpieczeństwo, kryzys migracyjny, polityka	premiera morawieckiego, migracyjne energetyczne, powiedzieli czego, będzie bardzo, serii spotkań	nowy kanclerz Niemiec, instrument szantażu, kwestia migracyjna, kanclerz Niemiec, Olaf Scholz, ... (237 in total)	rosja, polska, unia-europejska
2	Gas pipe explosion on Sicily	gaz, wybuch, Włochy	wybuchu gazociągu, czterech osób, miejscowości ravanusa, jak poinformowała, pomocze przeszukiwaniu	dziesiątek osób, miejsc wybuchu, wybuch gazociągu, włoska wyspa, ... (71 in total)	paleontologia, skamieniałości, fauna-kopalna
3	Person wounded in Kielce knife attack	Kielce, policja, rana, rynek	policji kielcach, rynku 31, wyjaśniają okoliczności, narzędziem przez, letni mężczyzna	kielceki rynek, poszukiwanie napastnika, ostre narzędzie, ... (44 in total)	bezpieczeństwo, kontrola, policja
4	Omicron coronavirus variant in UK	Covid, Omikron, Wielka Brytania, koronawirus, hospitalizacja	przypadki hospitalizacji, powiedzmy 50, minister edukacji, brytania pierwsze, liczba potwierdzonych	nowy wariant, wariant Omikron, wielka Brytania, ... (126 in total)	migracja, bezpieczeństwo, transport-kolejowy
5	Tatry National Park landmark vandalized	Tatry, wandalizm, szlak turystyczny	pomalowali granitowy, narodowy film, tatrzański park, morskiego oka, sprawę gdzie	morskie oko, tatrzański park narodowy, granitowy głaz, ... (88 in total)	historia, ochrona, bezpieczeństwo

energy supply disconnection. Our pT5kw model trained on scientific abstracts labels the following passage with two noun phrases: *loan* and *financial advisory* even though there is only a handful of abstracts with these keywords in the POSMAC:

Customer: I can't imagine how I could live without electricity. I need to use my fridge and washing machine. I guess I have no... I don't even know where I could borrow some money. Is there any way I could pay my debt in some kind of installments. How should I go about it?

Agent: It's alright. I understand and... I'm really sorry about your situation. Before we continue, however, we need to sort out a few formal issues. Can I ask you again to state your name and email address?

Table 5 shows frequencies of keywords assigned to a set of 50 DiaBiz transcripts representing scenarios from 6 different customer support domains. The recurrent keywords seem to accurately (if not exhaustively) summarize the underlying conversations. *Logistyka* (*logistics*) is the only potentially irrelevant keyword in this subset which may have resulted from some over-fitting of the model on the original domain of scientific abstracts.

Table 5. Frequent keywords generated by pT5kw for phone dialogues in different customer support domains and their English translations.

Domain	Keywords PL	Keywords EN
medical	gastroskopia (4)	gastroscopy (4)
	diagnostyka medyczna (3)	medical diagnosis (3)
	numer PESEL (2)	National Identification Number (2)
	diagnostyka (2)	diagnosis (2)
	lekarz POZ (1)	GP (1)
tourism	hotel (9)	hotel (9)
	turystyka (6)	tourism (6)
	Afryka (2)	Africa (2)
	Turcja (2)	Turkey (2)
	atrakcje (2)	attractions (2)
insurance	naprawa (6)	repair (6)
	uszkodzenie (5)	damage (5)
	ekspres do kawy (3)	coffee machine (3)
	reklamacja (3)	complaints (3)
	ubezpieczenie (2)	insurance (2)

Domain	Keywords PL	Keywords EN
banking	identyfikacja (6)	identification (6)
	bankowość internetowa (4)	online banking (4)
	logistyka (3)	logistics (3)
	Bank Narodowy S.A. (3)	National Bank S.A. (3)
	logowanie (2)	login (2)
energy	licznik (6)	meter (6)
	zerwanie plomby (4)	seal break (4)
	plomba na liczniku (8)	seal on meter (8)
	PESEL (2)	National Identification Number (2)
	weryfikacja tożsamości (1)	identity verification (1)

6 Summary and Future Work

Our evaluation of a keyword extraction solution based on a text-to-text transformer shows that the fine-tuned model called plT5kw outperforms the other approaches when tested on the original dataset of scientific abstracts. Furthermore, a preliminary analysis of keywords assigned to text from very different domains (news stories and speech transcripts) shows that the proposed solution is capable of generating relevant, properly formatted and well-abstracted keywords on extrinsic text samples. One of the limitations of this study stems from the fact that manual keyword annotations are intrinsically biased against high recall evaluations as authors are artificially restricted to assign a limited number of terms to each text. Therefore, a more systematic quantitative evaluation on extrinsic domains would require manually annotated datasets verified for inter-rater agreement. We envisage further challenges which need to be addressed in future research on this problem. For example, it seems reasonable to assume that open-set keyword extraction could benefit from distributional vector-based techniques of normalizing semantically equivalent keywords. Also, there are potential benefits of zero- or few-shot fine-tuning of text-to-text keyword extraction models to the target domain, which need to be considered more systematically. Finally, the results obtained in this study may vary for different languages, which requires further evaluation, possibly on multilingual variants of the T5 model used.

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) the National Centre for Research and Development, research grant POIR.01.01.01-00-1237/19.

References

1. Boudin, F.: 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). pp. 667–672. Association for Computational Linguistics, New Orleans, Louisiana (2018), <https://aclanthology.org/N18-2105>
2. Bougouin, A., Boudin, F., Daille, B.: TopicRank: Graph-based topic ranking for keyphrase extraction. In: Proceedings of the Sixth International Joint Conference on Natural Language Processing. pp. 543–551. Asian Federation of Natural Language Processing, Nagoya, Japan (2013), <https://aclanthology.org/I13-1062>
3. Chrabrowa, A., Dragan, Ł., Grzegorzczak, K., Kajtoch, D., Koszowski, M., Mroczkowski, R., Rybak, P.: Evaluation of Transfer Learning for Polish with a Text-to-Text Model. In: Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022). pp. 4374–4394. European Language Resources Association, Marseille, France (2022), <http://www.lrec-conf.org/proceedings/lrec2022/pdf/2022.lrec-1.466.pdf>
4. El-Beltagy, S.R., Rafea, A.: KP-miner: Participation in SemEval-2. In: Proceedings of the 5th International Workshop on Semantic Evaluation. pp. 190–193. Association for Computational Linguistics, Uppsala, Sweden (Jul 2010), <https://aclanthology.org/S10-1041>
5. Firoozeh, N., Nazarenko, A., Alizon, F., Daille, B.: Keyword Extraction: Issues and Methods. *Natural Language Engineering* **26**(3), 259–291 (2020)
6. Florescu, C., Caragea, C.: PositionRank: 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). pp. 1105–1115. Association for Computational Linguistics, Vancouver, Canada (Jul 2017), <https://aclanthology.org/P17-1102>
7. Frantzi, K., Ananiadou, S., Mima, H.: Automatic Recognition of Multi-word Terms: The C-value/NC-value Method. *International Journal on Digital Libraries* **3**(2), 115–130 (2000). <https://doi.org/10.1007/s007999900023>
8. Giarelis, N., Kanakaris, N., Karacapilidis, N.: A Comparative Assessment of State-of-the-art Methods for Multilingual Unsupervised Keyphrase Extraction. In: Maglogiannis, I., Macintyre, J., Iliadis, L. (eds.) *Artificial Intelligence Applications and Innovations*. pp. 635–645. Springer International Publishing, Cham (2021)
9. Grootendorst, M.: KeyBERT: Minimal Keyword Extraction with BERT (2020). <https://doi.org/10.5281/zenodo.4461265>
10. Joulin, A., Grave, E., Bojanowski, P., Mikolov, T.: Bag of Tricks for Efficient Text Classification. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers. pp. 427–431. Association for Computational Linguistics, Valencia, Spain (2017), <https://aclanthology.org/E17-2068>
11. Marciniak, M., Mykowiecka, A., Rychlik, P.: TermoPL — a Flexible Tool for Terminology Extraction. In: Calzolari, N., Choukri, K., Declerck, T., Grobelnik, M., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S. (eds.) *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. pp. 2278–2284. European Language Resources Association (2016), http://www.lrec-conf.org/proceedings/lrec2016/pdf/296_Paper.pdf
12. Mihalcea, R., Tarau, P.: TextRank: Bringing order into text. In: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing. pp. 404–

411. Association for Computational Linguistics, Barcelona, Spain (2004), <https://aclanthology.org/W04-3252>
13. 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>
 14. Reimers, N., Gurevych, I.: Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. pp. 4512–4525. Association for Computational Linguistics (2020), <https://aclanthology.org/2020.emnlp-main.365/>
 15. Sechidis, K., Tsoumakas, G., Vlahavas, I.: On the Stratification of Multi-label Data. In: Gunopulos, D., Hofmann, T., Malerba, D., Vazirgiannis, M. (eds.) *Machine Learning and Knowledge Discovery in Databases (ECML PKDD 2011)*. pp. 145–158. *Lecture Notes in Computer Science* vol. 6913, Springer Berlin Heidelberg (2011). https://doi.org/https://doi.org/10.1007/978-3-642-23808-6_10
 16. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., 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. Wydmuch, M., Jasinska, K., Kuznetsov, M., Busa-Fekete, R., Dembczyński, K.: A No-regret Generalization of Hierarchical Softmax to Extreme Multi-label Classification. In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems (NeurIPS 2018)*. pp. 6358–6368. Curran Associates Inc. (2018), <https://proceedings.neurips.cc/paper/2018/hash/8b8388180314a337c9aa3c5aa8e2f37a-Abstract.html>