TRANSFERABLE KEYWORD EXTRACTION AND GENERATION FROM SCHOLARLY DOCUMENTS WITH TEXT-TO-TEXT LANGUAGE MODELS

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.

Piotr Pęzik	University of Łódź, Faculty of Philology / VoiceLab, NLP Lab
Agnieszka Mikołajczyk	VoiceLab, NLP Lab
Adam Wawrzyński	VoiceLab, NLP Lab
Filip Żarnecki	VoiceLab, NLP Lab
Bartłomiej Nitoń	Institute of Computer Science, Polish Academy of Sciences
Maciej Ogrodniczuk	Institute of Computer Science, Polish Academy of Sciences

INTRODUCTION

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.

We share the three new datasets, i.e. POSMAC EN, InTechOpen and News200 at http://clip.ipipan.waw.pl/POSMAC/datasets.

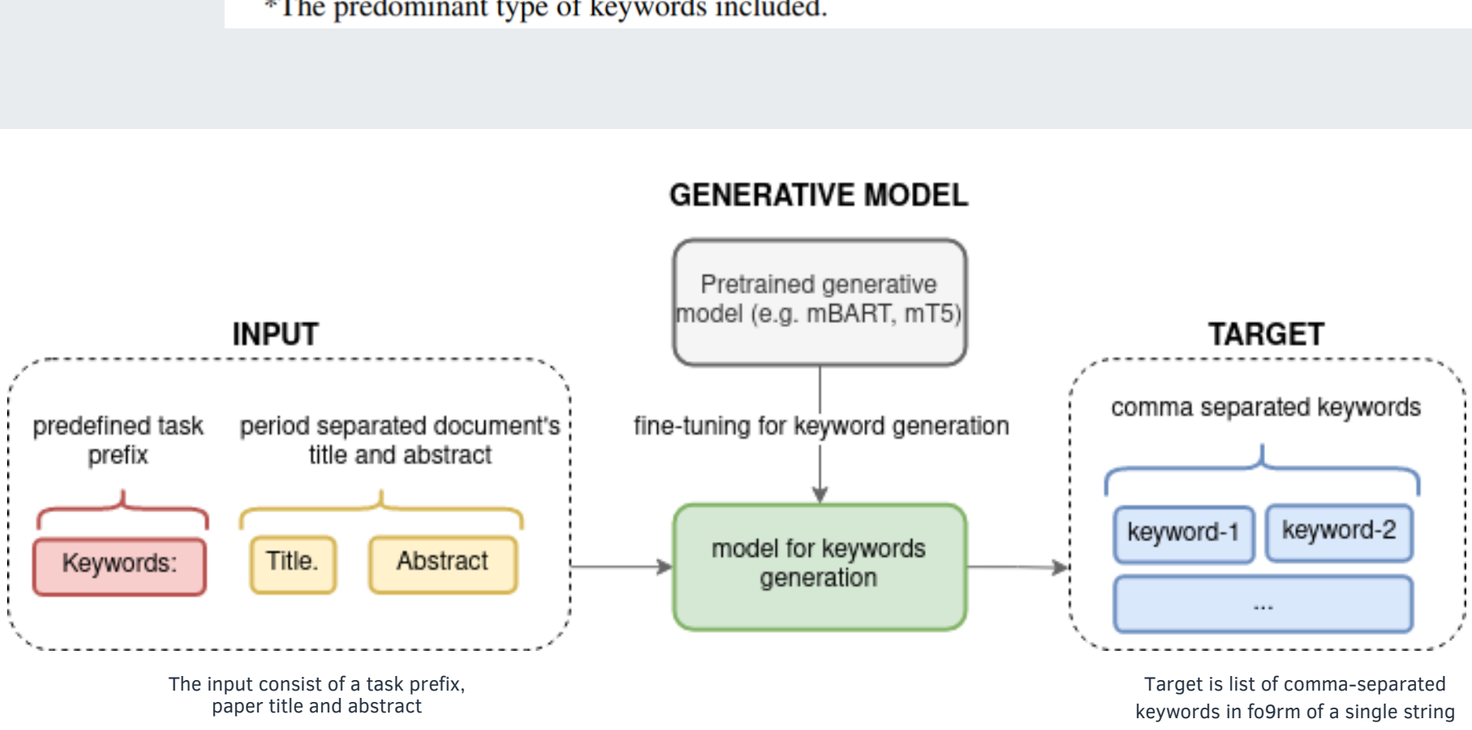
TYPE OF KEYWORD ASSIGNMENT IN SELECTED KEG DATASETS

Dataset	Avg. 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	16 916	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	90 198	Authors
OAG (AMiner)	4	Unclear	250 899	Unclear

*The predominant type of keywords included.

MODEL

We compared the results obtained with mT5 with the performance of a KEG model based on mBART. 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, 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.



OVERALL PERFORMANCE OF EVALUATED MODELS ON NEW DATASETS OF SCIENTIFIC AND NEWS TEXTS

Model	Train set	POSMAC P	POSMAC R	POSMAC F_1	News articles P	News articles R	News articles 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**

CONCLUSION

Our evaluation of a keyword extraction solution based on a T5 model shows that the fine-tuned model 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.

Curated Multilingual Language Resources for CEF.AT

Supported by the EC in the CEF Telecom Programme (Action No: 2019-EU-IA-0034, GA No: INEA/CEF/ICT/A2019/1926831), the Polish Ministry of Science and Higher Education project 5103/CEF/2020/2, funds for 2020–2022 and the National Centre for Research and Development, research grant POIR.01.01.01-00-1237/19.