A Large-Scale Dataset for Biomedical Keyphrase Generation

Mael Houbre, Florian Boudin, Beatrice Daille

A Large-Scale Dataset for Biomedical Keyphrase Generation

Maël Houbre, Florian Boudin and **Béatrice Daille**
Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes, France
first.last@univ-nantes.fr

Abstract

Keyphrase generation is the task consisting in generating a set of words or phrases that highlight the main topics of a document. There are few datasets for keyphrase generation in the biomedical domain and they do not meet the expectations in terms of size for training generative models. In this paper, we introduce kp-biomed, the first large-scale biomedical keyphrase generation dataset with more than 5M documents collected from PubMed abstracts. We train and release several generative models and conduct a series of experiments showing that using large scale datasets improves significantly the performances for present and absent keyphrase generation. The dataset is available under CC-BY-NC v4.0 license at https://huggingface.co/datasets/taln-ls2n/kpbiomed.

1 Introduction

Keyphrase generation aims at automatically generating a set of keyphrases, that is, words and phrases that summarize a given document. Since they distill the important information from documents, keyphrases have showed to be useful in many applications, most notably in information retrieval (Fagan, 1987; Zhai, 1997; Jones and Staveley, 1999; Song et al., 2006; Boudin et al., 2020) and summarization (Zha, 2002; Wan et al., 2007; Qazvinian et al., 2010).

Current models for generating keyphrases are built upon the sequence-to-sequence architecture (Sutskever et al., 2014) and are able to generate absent keyphrases that is, keyphrases that do not appear in the source text. However, training these models require large amounts of labeled data (Meng et al., 2021). Unfortunately, such data is only available for limited domains and languages which greatly limits the applicability of these models (Ye and Wang, 2018). This work addresses this issue and introduces kp-biomed, the first large-scale dataset for keyphrase generation in the biomedical domain.

Creating labeled data for keyphrase generation is a challenging task, requiring expert annotators and great effort (Kim et al., 2010; Augenstein et al., 2017). A commonly-used approach to cope with this task is to collect scientific abstracts and use keyphrases provided by authors as a proxy for expert annotations. Authors provide keyphrases without any vocabulary constraint to highlight important points of their article; whereas indexers use a specific vocabulary and focus on indexing the article within a collection (Névéol et al., 2010). Therefore, keyphrases may differ from MeSH headings which are another indexing resource in the biomedical domain. Fortunately, author keyphrases are becoming increasingly available in the biomedical domain (Névéol et al., 2010), since they can be incorporated into search strategies in PubMed to improve retrieval effectiveness (Lu and Kipp, 2014). Despite this, the largest keyphrase-labeled biomedical dataset that we know of has about 3k abstracts, all of which are labeled with present-only keyphrases (Gero and Ho, 2019). In this paper, we take advantage of the expansive PubMed database to build a sufficiently large dataset to train biomedical keyphrase generation models[1]. We then compare models trained with different training set sizes to highlight the impact of dataset sizes in keyphrase generation. Our contributions are as follows:

- kp-biomed, a large, publicly available dataset for keyphrase generation in the biomedical domain, available through the Huggingface dataset platform[2];
- Transformer-based models for biomedical keyphrase generation, providing open bench-

[1] KP20k is currently considered as the reference dataset size ($\geq 500k$) to train keyphrase generation models
[2] https://huggingface.co/datasets/taln-ls2n/kpbiomed

marks to stimulate further work in the area[3];

- Performance analysis of our models, which provides valuable insights into their generalization ability to other domains.

2 Dataset

We employ the December 2021 baseline set of MEDLINE/PubMed citation records[4] as a resource for collecting abstracts, which contains over 33 million records. We extracted all the records (5.9 million) that include a title, an abstract and some author keyphrases. Records of papers published between 1939 and 2011 only account for a small fraction of these extracted records (3%) and were further filtered out to avoid possible diachronic issues. Last, we went through the remaining records to split the semicolon-separated list of author keyphrases and discard those having keyphrases with punctuation in it. The resulting dataset is composed of 5.6 million abstracts and was randomly and evenly divided by publishing year into training, validation and test splits. To investigate the impact of the amount of training data on the quality of the generated keyphrases, the training split was further divided into increasingly large subsets: small (500k), medium (2M) and large (5.6M). The training splits are also evenly divided by publishing year.

Statistics of the `kp-biomed` dataset are detailed in Table 1 along with other commonly-used datasets for keyphrase generation and extraction. We are aware of only two datasets in the biomedical domain: `NamedKeys` (Gero and Ho, 2019) which is made up of MEDLINE/PubMed abstracts and is therefore mostly included in `kp-biomed`, and Schutz (Schutz, 2008) which is composed of full-text articles from the same source. It is worth noting that these datasets are very limited in size (3k and 1.3k documents respectively) compared to recent keyphrase generation datasets `KP20k` (Meng et al., 2017), `KPTimes` (Gallina et al., 2019) and `LDKP10k` (Mahata et al., 2022). Table 1 shows that thanks to the amount of papers available in MEDLINE/PubMed, `kp-biomed` is the largest of all aforementioned datasets, being more than 10 times larger than `KP20k` which is the current reference dataset for keyphrase generation. The average number of keyphrases per document (#kp) in `kp-biomed` is roughly the same than in `KP20k` and `LDKP10k` which have their keyphrases assigned by authors as well. However, we see that this number is way below the average number of keyphrases assigned by professional indexers like in Inspec (Hulth, 2003) or when authors' keyphrases are combined with readers' as in SemEval-2010 (Kim et al., 2010). The unusually high number of keyphrases per document in `NamedKeys`, despite having author assigned keyphrases, is because of two restrictive criteria. Indeed, each article has at least 5 keyphrases all of which have to occur in the source text. The average number of words per keyphrase (#kp_len) is also comparable for all scientific datasets regardless of the kind of annotators.

Using keyphrases as proxies for indexing or expanding documents with queries composed of words that do not appear in the source text, has been proven more useful to enhance document retrieval than using words occurring in the text (Boudin et al., 2020; Nogueira et al., 2019). In keyphrase generation, we call those keyphrases absent keyphrases, for which several definitions are being used. We refer to the definition from (Meng et al., 2017) "we denote phrases that do not match any contiguous subsequence of source text as absent keyphrases" which was then precised in (Boudin and Gallina, 2021). In (Gero and Ho, 2019) the keyphrase "anesthesia" is considered present if the word "postanesthesia" is in the source text. In our case, it is considered absent which is why `NamedKeys` does not appear with 100% present keyphrases in Table 1. The main difference between `kp-biomed` and `NamedKeys`, despite the number of documents, is the proportion of absent keyphrases. `kp-biomed` contains about 34% of absent keyphrases which is in the same range as scientific datasets `KP20k` and `LDKP10k` that were designed to train neural generative approaches (Meng et al., 2017; Mahata et al., 2022).

3 Experiments

3.1 Models

In keyphrase generation, the architectures are currently mainly based on autoencoders with Recurrent Neural Networks (Meng et al., 2017; Chen et al., 2018, 2019; Chan et al., 2019) or Transformers (Meng et al., 2021; Ahmad et al., 2021).

Following the work of (Meng et al., 2021) that obtained state-of-the-art results with Transform-

[3] https://huggingface.co/datasets/taln-ls2n/kpbiomed-models

[4] https://ftp.ncbi.nlm.nih.gov/pubmed/baseline/

Domain	Dataset	#train	#val	#test	#doc len	#kp	#kp len	P	A
Biomedical	**kp-biomed** (ours)	5.6M	20k	20k	271	5.3	1.9		
	NamedKeys	–	–	3k	276	14.3	1.9		
	Schutz	–	–	1.3k	5.4k	5.4	1.9		
General scientific articles	KP20k	530k	20k	20k	175	5.3	2.1		
	SemEval-2010	144	–	100	192	15.4	2.1		
	Inspec	1k	500	500	138	9.8	2.3		
	LDKP10k	1.3M	10k	10k	4.9k	6.9	2.1		
News	KPTimes	260k	10k	20k	921	5.0	1.5		

Table 1: Statistics of the proposed dataset. For comparison purposes, we also report statistics of commonly-used and other biomedical datasets. Columns P and A are respectively the percentage of keyphrases occurring in the source text and absent ones.

ers, we used two different generative BART models (Lewis et al., 2020) and compared their performances on different domains. However, in this article we did not seek to get state-of-the-art results, but rather introduce kp-biomed to the community with results on well known baselines, which is why we employed pre-trained models that we just fine-tuned for keyphrase generation (Chowdhury et al., 2022). The models are BioBART-base (Yuan et al., 2022) which is already pre-trained on PubMed and BART-base (Lewis et al., 2020) which is pre-trained on news, books and webtext. To the best of our knowledge, there is no generic scientific BART model. Therefore, we chose BioBART for fine-tuning on scientific datasets rather than BART. Models are available via the huggingface platform.

For comparison with extractive approaches, we considered MultipartiteRank (Boudin, 2018) as a baseline, which is state-of-the-art in unsupervised graph-based keyphrase extraction. We used the implementation available in the keyphrase extraction toolkit pke[5] with the default settings.

3.2 Experimental settings

We followed the One2Seq paradigm (Meng et al., 2021) for training which consists of generating the keyphrases of an input article as a single sequence. For each article, we concatenated the ground truth keyphrases as a single sequence with a special delimiter. Following (Meng et al., 2021), present keyphrases were ordered by their first occurrence in the source text followed by the absent ones.

We trained each model for 10 epochs with a batch size of 128. We set the input length limit at 512 tokens for the text and 128 tokens for the reference keyphrase sequence. All the parameters and the training were handled with the huggingface trainer API[6]. Hyperparameters and hardware details are available in appendix A. Training the BioBART-base model on the small training split for 10 epochs took about 9 hours and about 110 hours on the large training split. Once models were trained, we over-generated keyphrase sequences using beam search with a beam width of 20 for evaluation. Inference on test sets took around 50 minutes each.

3.3 Evaluation

We evaluated our models on 3 datasets, kp-biomed for biomedical data, KP20k for generic scientific documents and KPTimes for news articles. We did not use NamedKeys as a test set as we noticed a substantial overlap with our training set. We evaluated present and absent keyphrase generation separately to get better insights of our models' performances. To that end, we only compared each model's output to the present (respectively absent) keyphrases of the ground truth. For present keyphrases we employed F1@M and F1@10. F1@M is the F1 measure applied on the first keyphrase sequence generated by the model whereas F1@10 evaluates the top ten generated keyphrases. We evaluated absent keyphrase generation with R@10 which is the recall on the top 10 generated keyphrases. As F1@10 and R@10 require 10 keyphrases, if we did not have enough unique keyphrases with

[5] https://github.com/boudinfl/pke

[6] Our code is available for reproducibility. https://github.com/MHoubre/kpbiomed

Model	kp-biomed		KP20k		KPTimes	
	F1@10	F1@M	F1@10	F1@M	F1@10	F1@M
MultipartiteRank	15.3	–	12.9	–	16.7	–
BioBART-small	31.4	32.5	25.2	27.1	22.0	24.4
BioBART-medium	32.5†	33.8†	26.2†	28.2†	22.1	24.6
BioBART-large	**33.1**†	**34.7**†	26.9†	28.9†	23.5†	26.2†
BioBART-KP20k	28.2	29.5	**28.6**†	**31.9**†	16.8	19.2
BART-KPTimes	9.1	9.6	3.6	2.7	**29.7**†	**39.4**†

Table 2: Performances of the models on present keyphrase generation. †means significant improvements over BioBART-small. Second best results are underlined.

our over generation, we added the token "<unk>" until we reached 10 keyphrases. The generated keyphrases and the reference were stemmed with the Porter Stemmer to reduce matching errors. To measure statistical significance, we opted for Student's t-test at $p < 0.01$.

3.4 Results

The macro-averaged results of the evaluation are reported in Table 2 and Table 3. BioBART-KP20k (respectively BART-KPTimes) stands for the BioBART (respectively BART) model which has been fine-tuned on KP20k (respectively KPTimes). For BioBART models, we add the size of the kp-biomed training split in the name for clarity.

Model	kp-biomed	KP20k	KPTimes
	R@10	R@10	R@10
BioBART-small	3.3	1.8	2.6
BioBART-medium	3.6†	1.9	2.7
BioBART-large	**4.1**†	1.9	2.1
BioBART-KP20k	2.9	**5.5**†	1.6
BART-KPTimes	1.5	0.8	**39.1**†

Table 3: Performances of the models on absent keyphrase generation. †means significant improvements over BioBART-small. Second best results are underlined.

Transformer based approaches achieve the best results but only on the datasets they were trained on as previously showed for RNN based approaches in (Gallina et al., 2019). For present keyphrase generation, BioBART-large achieves significant improvements compared to its small and medium counterparts in all datasets. This shows that using more data does improve the performances of the generative approaches in predicting present keyphrases in in and out of domain data. The performance drop of BioBART-KP20K on kp-biomed is interestingly much more controlled than BioBART models' on KP20k. Compared to BioBART-small which has been trained on the same amount of data, the drop in F1@M is only of 7.5% relative for BioBART-KP20k when it is of 16.6% relative for BioBART-small. We think that BioBART's pretraining may be beneficial for BioBART-KP20k on kp-biomed. On news articles though, BioBART-KP20k shows a relative drop of 35%, when it is only of 25% relative for BioBART-small. When used on out of domain data, BART-KPTimes performs even worse than MultipartiteRank.

In absent keyphrase generation, models fail in attaining significant improvements outside of their domain. Using more data does not seem to help for out of domain absent keyphrase generation. We can explain the high results of BART-KPTimes on its test set by the fact that many of the absent keyphrases are common to numerous articles.

We also think that the keyphrase order that we chose for training is one reason for the models' poor abstractive results. To verify this hypothesis, we compute the average percentage of the models' predictions appearing in the source text. Results are reported in Table 4. For @10, we removed all the added <unk> tokens before computing. It is clear that the extraction percentage of each model decreases when using top 10 predictions on all datasets. This shows that models prioritize generating present keyphrases which can then lead to low quality absent candidates.

Model	kp-biomed		KP20k		KPTimes	
	@M	@10	@M	@10	@M	@10
BioBART-large	96.3	92.2	94.8	88.5	93.5	84.6
BioBART-KP20k	95.4	84.5	91.8	82.7	83.7	66.6
BART-KPTimes	46.0	31.2	21.4	17.4	65.8	50.7

Table 4: Extraction percentage in top M and top 10 predictions

4 Conclusion

This paper introduces kp-biomed, the first large scale dataset for biomedical keyphrase generation. We hope this new dataset will stimulate new research in biomedical keyphrase generation. Several generation models have been trained on this dataset and showed that having more data significantly improves the performances for present and absent keyphrase generation. However, models still perform very poorly on absent keyphrase generation even when using larger amounts of data. In future work, we will focus on how to use kp-biomed to improve biomedical absent keyphrase generation.

5 Broader Impact and Ethics

kp-biomed contains some abstracts that are part of copyright protected articles. As the "all rights reserved" statement is optional to be copyright protected, removing articles with this statement does not solve the problem (i.e no copyright statement does not mean free of use data). To be able to collect, work with these data and share the dataset to the research community, we complied with the conditions of US fair use and the exceptions from the 2019/79 EU guideline on using copyright content in text and data mining for research purposes. One of those criteria was to not use the data for commercial purposes which is why we opted for the Creative Commons Non Commercial use license CC-BY-NC v4.0.

Acknowledgements

We thank the anonymous reviewers for their valuable input on this article and our colleagues from the TALN team at LS2N for their proofreading and feedback. This work is part of the ANR DELICES project (ANR-19-CE38-0005) and was performed using HPC resources from GENCI-IDRIS (Grant 2022-[AD011013670]).

References

Wasi Ahmad, Xiao Bai, Soomin Lee, and Kai-Wei Chang. 2021. Select, extract and generate: Neural keyphrase generation with layer-wise coverage attention. 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)*, pages 1389–1404, Online. Association for Computational Linguistics.

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.

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 and Ygor Gallina. 2021. Redefining absent keyphrases and their effect on retrieval effectiveness. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4185–4193, Online. 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.

Hou Pong Chan, Wang Chen, Lu Wang, and Irwin King. 2019. Neural Keyphrase Generation via Reinforcement Learning with Adaptive Rewards. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2163–2174, Florence, Italy. Association for Computational Linguistics.

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, Yifan Gao, Jiani Zhang, Irwin King, and Michael R. Lyu. 2019. Title-Guided Encoding for Keyphrase Generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):6268–6275. Number: 01.

Md Faisal Mahbub Chowdhury, Gaetano Rossiello, Michael Glass, Nandana Mihindukulasooriya, and Alfio Gliozzo. 2022. Applying a Generic Sequence-to-Sequence Model for Simple and Effective Keyphrase Generation. *arXiv:2201.05302 [cs]*. ArXiv: 2201.05302.

J. Fagan. 1987. Automatic phrase indexing for document retrieval. In *Proceedings of the 10th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '87, page 91–101, New York, NY, USA. Association for Computing Machinery.

Ygor Gallina, Florian Boudin, and Beatrice Daille. 2019. KPTimes: A large-scale dataset for keyphrase generation on news documents. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 130–135, Tokyo, Japan. Association for Computational Linguistics.

Zelalem Gero and Joyce C. Ho. 2019. Namedkeys: Unsupervised keyphrase extraction for biomedical documents. In *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, BCB '19, page 328–337, 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.

Steve Jones and Mark S. Staveley. 1999. Phrasier: A system for interactive document retrieval using keyphrases. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99, page 160–167, New York, NY, USA. Association for Computing Machinery.

Su Nam Kim, Olena Medelyan, Min-Yen Kan, and Timothy Baldwin. 2010. SemEval-2010 task 5 : Automatic keyphrase extraction from scientific articles. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 21–26, Uppsala, Sweden. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

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

Debanjan Mahata, Navneet Agarwal, Dibya Gautam, Amardeep Kumar, Swapnil Parekh, Yaman Kumar Singla, Anish Acharya, and Rajiv Ratn Shah. 2022. LDKP: A Dataset for Identifying Keyphrases from Long Scientific Documents. *arXiv:2203.15349 [cs]*. ArXiv: 2203.15349.

Rui Meng, Xingdi Yuan, Tong Wang, Sanqiang Zhao, Adam Trischler, and Daqing He. 2021. An empirical study on neural keyphrase generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4985–5007, Online. Association for Computational Linguistics.

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.

Aurélie Névéol, Rezarta Islamaj Doğan, and Zhiyong Lu. 2010. Author keywords in biomedical journal articles. In *AMIA annual symposium proceedings*, volume 2010, page 537. American Medical Informatics Association.

Rodrigo Nogueira, Wei Yang, Jimmy Lin, and Kyunghyun Cho. 2019. Document Expansion by Query Prediction. ArXiv:1904.08375 [cs].

Vahed Qazvinian, Dragomir R. Radev, and Arzucan Özgür. 2010. Citation summarization through keyphrase extraction. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 895–903, Beijing, China. Coling 2010 Organizing Committee.

Alexander Thorsten Schutz. 2008. Keyphrase extraction from single documents in the open domain exploiting linguistic and statistical methods. Master's thesis, Digital Enterprise Research Institute, National University of Ireland, Galway.

Min Song, Il Yeol Song, Robert B. Allen, and Zoran Obradovic. 2006. Keyphrase extraction-based query expansion in digital libraries. In *Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '06, page 202–209, New York, NY, USA. Association for Computing Machinery.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Xiaojun Wan, Jianwu Yang, and Jianguo Xiao. 2007. Towards an iterative reinforcement approach for simultaneous document summarization and keyword extraction. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 552–559, Prague, Czech Republic. Association for Computational Linguistics.

Hai Ye and Lu Wang. 2018. Semi-supervised learning for neural keyphrase generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4142–4153, Brussels, Belgium. Association for Computational Linguistics.

Hongyi Yuan, Zheng Yuan, Ruyi Gan, Jiaxing Zhang, Yutao Xie, and Sheng Yu. 2022. BioBART: Pretraining and evaluation of a biomedical generative language model. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 97–109, Dublin, Ireland. Association for Computational Linguistics.

Hongyuan Zha. 2002. Generic summarization and keyphrase extraction using mutual reinforcement principle and sentence clustering. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '02, page 113–120, 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.

A Training settings

- GPU type: V100 32Go
- Number of GPU: 4
- Trainer: Seq2SeqTrainer
- Text max size: 512
- Reference max size: 128
- Optimizer : AdamW
- Learning rate: 5×10^{-5}
- Other hyperparameters: Seq2SeqTrainer default values