Assessing the Stability of Text-to-Text Models for Keyword Generation Tasks

Tomasz Walkowiak[0000−0002−7749−4251]

Wroclaw University of Science and Technology
tomasz.walkowiak@pwr.edu.pl

Abstract. The paper investigates the stability of text-to-text T5 models in keyword generation tasks, highlighting the sensitivity of their results to subtle experimental variations such as the seed used to shuffle fine-tuning data. The authors advocate for incorporating error bars and standard deviations when reporting results to account for this variability, which is common practice in other domains of data science, but not common in keyphrase generation. Through experiments with T5 models, they demonstrate how small changes in experimental conditions can lead to significant variations in model performance, particularly with larger model sizes. Furthermore, they analyze the coherence within a family of models and propose novel approaches to assess the stability of the model. In general, the findings underscore the importance of considering experimental variability when evaluating and comparing text-to-text models for keyword generation tasks.

Keywords: keywords extraction · T5 · model stability

1 Introduction

Keywords play a crucial role in a summary and retrieval of documents, providing the reader with a brief overview of the content of the document. Maintaining high-quality keywords is vital to improve content visibility and allow users to easily find relevant information. Various techniques are available to automate keyword extraction [14], including unsupervised and supervised methods. One of the most promising approaches is the text-to-text generation technique, which generates tags dynamically based on the input text. This method allows us to capture not only the extracted keywords but also relevant terms that may not be explicitly mentioned in the text. Techniques in this field employ transformer-based neural networks such as T5 [13] and BART [6] to achieve cutting-edge results[9–11].

In many fields of data science, it is considered best practice to incorporate error bars when presenting findings, accounting for variations stemming from different random seeds across multiple experiments. This methodology is frequently endorsed by various conferences, as evidenced by the guidelines outlined in the NeurIPS checklist [1]. However, in the field of keyphrase generation, it is not

[1] https://neurips.cc/Conferences/2022/PaperInformation/PaperChecklist

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63759-9_14

customary to repeat experiments with different seeds and subsequently report standard deviations [11].

In this study, our objective is to examine this conventional practice. Our inquiry is prompted by the experiments described in [12], where two models exhibit disparities in their results, even starting from the same base model and probably trained on the same dataset. We seek to investigate the degree of sensitivity of the T5 models to subtle alterations under experimental conditions.

2 Related work

There are various techniques to automate keyword extraction [14], including unsupervised [17, 9] and supervised methods. The text-to-text approach to keyword extraction utilized in this paper falls within the supervised domain, which can be further classified into two groups: (i) classification with a closed set of labels [8, 12], and (ii) classification with an open set of labels [18, 7], both of which have been applied to this task. Recently, there has been increasing attention among researchers towards the use of T5 for keyword extraction [9–11]. The sensitivity of text-to-text models to subtle experimental conditions, to the best of the authors' knowledge, has not been thoroughly explored in the literature. However, the impact of minor experimental variations has been studied in other domains of machine learning [4], including natural language processing [1], image recognition [2], and out-of-distribution detection [15].

3 Method

The T5 [13] models are trained for the keyword extraction task using transfer learning. We start with a pre-trained base model and fine-tune it to generate a sequence of keywords from the original text, which is prefixed with a prompt. Our chosen prompt is "generate keywords: ". During experiments, we used base models for English [13] and Polish (pltT5) [3]. Evaluation of fine-tuned models requires metrics that compare the results obtained with those of the target ones. We utilize three metrics: F1 score, a custom metric based on set similarity and semantic similarity. In this paper, we use the "micro" F1 score, which is calculated globally by summing the total true positives, false negatives, and false positives. One of the drawbacks of the F1 metric is its lack of symmetry, which means that its value can change when we replace targets with the results obtained. To address this, we propose a modification of the F1 metric, denoted here as *keysim*. The concept originates from the idea of similarity between sets (the target and the obtained keywords). We propose to calculate the similarity between sets a and b as the number of common elements divided by the number of elements in the first set (a). Since this metric is asymmetric, we symmetrize it using the harmonic mean (similar to the F1 metric). Finally, averaging these similarities, among the n examples, we can derive the desired metric:

$$keysim = \frac{1}{n}\sum_{i=1}^{n}\frac{2|a_i \cap b_i|}{|a_i| + |b_i|}, \tag{1}$$

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63759-9_14

where a_i, b_i represent the targets and results.

The F1 and *keysim* evaluate keywords as discrete symbols, without considering the existence of synonyms in the language and the subjective nature of keywords. Therefore, it is valuable to evaluate keyword generators by incorporating semantic similarity between them. Once again, we will average the similarities for each sample. The semantic similarity between two words is commonly calculated using word embeddings [5]. We will follow this approach by initially defining the similarity between a keyword x and a set of keywords a as the maximum cosine similarity between an embedding of the word ($word2vec(x)$) and all keyword embeddings from the set ($word2vec(a_i)$), i.e.:

$$sim(x, a) = \max_{i=1...|a|} cos\left(word2vec(x), word2vec(a_i)\right). \tag{2}$$

In the experiments reported, we generated embeddings for keywords using fastText models [5] for English and Polish. With the word-to-set similarity established, we can define the similarity between two sets as the average of similarities calculated using Eq. 2. As the compared sets of keywords may have different numbers of elements, we must average the similarities between set A and set B and between set B and set A. Finally, by averaging over all samples (pairs of keyword sets), we obtain the semantic similarity metric, i.e.:

$$semsim = \frac{1}{2n}\sum_{i=1}^{n}\left(\frac{1}{|a_i|}\sum_{j=1}^{|a_i|} sim(a_{i,j}, b_i) + \frac{1}{|b_i|}\sum_{j=1}^{|b_i|} sim(b_{i,j}, a_i)\right). \tag{3}$$

4 Experiments and Results

4.1 Datasets

The data used to train and test the T5 models were corpora developed within the CURLICAT project [16], exactly the Polish Open Science Metadata Corpus (POSMAC) [9, 10] which is part of CURLICAT. POSMAC contains English and Polish text annotated with keywords. For Polish texts, two versions were used: a smaller subset S-PL(accessible at[2]) and an almost three times larger one L-PL from [10].

4.2 Stability for Different Family of Models

T5 models are available in various sizes, typically classified as small, base, and large, and for different languages [13]. The fine-tuned T5 model retains the structure of the base model, but its weights are adjusted during fine-tuning. For clarity, in the following analysis, we will refer to a group of models derived from the same base model and tuned on the same training set as a "family of models."

[2] `https://clip.ipipan.waw.pl/POSMAC`

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63759-9_14

The train data set is typically randomly shuffled. So, if we keep the learning rate scheduler method unchanged, it seems as the only source of randomness that can affect the final model and hence its quality metrics.

To test this hypothesis, we performed a series of experiments. We used three data sets: two for Polish with different sizes (details are given in Section 4.1) and one for English. We train each each model on the same data set five times and show the results not as a single value, but as the mean and standard deviation. The results of these experiments are shown in Table 1. It is evident that the metrics vary slightly across each run.

A striking observation is the poor performance of large models, which exhibit a wide range of results. We conducted a deeper investigation of this phenomenon. They showed that for three runs of the large model, we observe notably poor results. However, for the other two runs, the results are significantly better compared to those of small models. Further examination revealed that these underperforming models could be identified during the training phase, as they exhibited significantly higher losses (even after the first epoch) than the runs, yielding better results on the training sets. Similar patterns were observed with the L-PL and EN datasets. This suggests that large models tend to struggle during training, but such behavior can be detected early. Therefore, it is recommended to identify such runs (based on random seed values) during training and exclude such models from further analysis.

4.3 Coherence of the Model Family

The findings of earlier sections reveal that simply changing the order of the training data can lead to different models. Having metrics to gauge the similarities between these models would be advantageous. The results presented thus

		keysim [%]			F1 [%]			*semsim* [%]		
dataset	model	average	std	range	average	std	range	average	std	range
S-PL	small	14.34	0.84	2.39	17.76	1.03	3.03	60.56	0.59	1.65
	base	20.28	1.10	2.96	24.34	0.64	1.80	63.69	0.11	0.31
	large	15.28	6.87	16.54	20.66	6.28	15.90	61.93	2.82	6.65
L-PL	small	15.50	0.27	0.72	19.05	0.26	0.72	61.83	0.12	0.36
	base	21.42	1.04	2.96	25.72	0.78	2.13	64.87	0.31	0.84
	large	15.07	5.29	15.98	21.17	5.32	16.20	61.30	2.92	8.91
EN	small	11.25	0.76	2.03	14.53	0.76	2.02	57.89	0.47	1.26
	base	16.67	0.46	1.25	21.02	0.40	0.99	60.88	0.23	0.65
	large	13.68	3.40	8.84	17.69	3.30	8.65	60.17	1.62	4.18

Table 1. Stability analysis of T5 models of varying sizes and fine-tuned on diverse datasets. Each row corresponds to a model family. The results are the averages of the quality metrics computed from five runs with different shuffle seeds. Standard deviations and ranges (computed as the difference between the maximum and minimum values) are also provided. Low values of standard deviation and range indicate higher stability of the model family.

.

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63759-9_14

far offer some insight by revealing the similarity between the model output and the target, making the standard deviation a potential indicator of stability, especially when it exhibits low values. However, the standard deviation of quality metrics might be overlooked because it primarily estimates the distances to the target values rather than assessing the similarities between models within the same family. Given that the *keysim* and *semsim* metrics are symmetric, we can utilize them to measure the similarities between two models. By treating one model as the target and then averaging over all pairs of models within a family, we can obtain an indicator of the family's coherence (with larger values indicating greater coherence).

The results depicting the similarity between models, obtained for the test datasets, are shown in Table 2. For a family of models comprising five fine-tuned models (derived from five runs with different seeds), we have a total of ten pairs. Therefore, we show the average, standard deviation, and range for each metric. To maintain clarity, we omit the results for large models, as discussed in the preceding section. Higher metrics values (with a maximum of 100%) indicate greater coherence of a model family. We observe that the metrics values are higher than those in Table 1, suggesting that the models are more similar to each other than to the training targets. However, the values are still relatively small, indicating significant differences between the models obtained. For the smaller datasets (S-PL and EN), we notice that the base models are more stable (with higher values of metrics) than the small models. However, for the L-PL model, the trend is the opposite.

4.4 Coherence of the Family of Models for Texts Further Away from the Original Training Set

A crucial feature of a keyword generator by text-to-text models is its ability to generate keywords for text that span various domains [9, 11]. Since the training set typically covers a limited domain, researchers evaluate keyword generators in various topical domains and text genres that differ from those used during

		keysim [%]			*semsim* [%]		
data	model	average	std	range	average	std	range
S-PL	small	60.08	10.88	24.51	85.88	3.99	9.04
	base	66.41	6.03	15.77	87.86	1.52	4.18
L-PL	small	72.77	1.10	3.92	90.95	0.38	1.40
	base	68.20	4.19	11.91	88.72	1.25	3.88
EN	small	66.99	5.25	14.07	87.16	2.11	5.64
	base	69.65	2.29	7.56	87.88	0.83	2.93

Table 2. The mean value, standard deviation, and range of the *keysim* and *semsim* metrics calculated for all pairs of models within the same family. We suggest utilizing these metrics as indicators of family model coherence, with higher average values suggesting greater coherence.

.

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63759-9_14

model family		evaluation datasets			
data	model	test	novels	French	Lorem Ipsum
S-PL	small	60.08±10.88	55.71±9.17	45.41±11.62	42.46±13.19
	base	66.41±6.03	65.67±3.60	57.72±3.00	24.41±17.70
L-PL	small	72.77±1.10	66.03±1.21	55.99±1.17	44.49±4.49
	base	68.20±4.19	60.21±2.69	55.59±2.04	16.44±15.32
EN	small	73.16±1.34	64.23±1.88	65.89±0.96	36.85±3.16
	base	72.99±0.66	67.23±0.73	62.71±1.01	29.01±4.72

Table 3. The mean value and standard deviation of the *keysim* metric calculated for all pairs of models within the same family, considering test datasets and texts originating farther from the original training data. The results indicate a consistent decrease in the mean values, suggesting that models within the same family tend to diverge increasingly as texts are drawn from further domains.

.

training [9, 10], or even for languages not used during training [11]. That is why it is worth examining the coherence of each family of models against data sets that are increasingly distant from the original training set in the sense of topic area and languages. The benefit of the approach outlined in the preceding section lies in its independence from the need for targets to evaluate the coherence of the model family on a given dataset. We evaluated various model families using their respective testing sets (so texts within an identical domain). Next, we used texts from novels (in the same language as the training data). So texts from other domain. It is followed by texts in French and finally random texts (Lorem Ipsum). The results are shown in Table 3. It is noticeable that the coherence metric, defined as the average *keysim* score between all pairs of models from each family, declines as texts originate from increasingly distant domains. This suggests a degradation in the coherence of model outputs, indicating a trend toward randomness as the input data strays further from the training set.

5 Conclusion

We have demonstrated that text-to-text models fine-tuned for key generation downstream tasks are sensitive to experimental factors, such as the order of training data. These nuances influence the achieved metric values. Therefore, we propose that the research community conducting key generation tasks should replicate experiments using different seeds and subsequently report the results as mean values and standard deviations. Although this approach is a de facto standard in many areas of AI, it is not yet adopted in the domain of key generation [11].

We conducted an in-depth analysis of the coherence within a family of models. We analyzed models derived from the same training dataset and base model, differing only in the order of their training sets used in fine-tuning, to examine how their results vary. To assess this difference, we used the standard deviation of quality metrics. Additionally, we proposed an original approach that analyzes

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63759-9_14

output between models without requiring a target dataset to evaluate model coherence. The results illustrate how model coherence decreases across datasets from increasingly diverse domains. This demonstrates how the responses of models within the same family, expected to be highly similar, become increasingly distinct for texts that are farther removed from the training set.

Acknowledgments. The work was financed as part of the investment: "CLARIN ERIC – European Research Infrastructure Consortium: Common Language Resources and Technology Infrastructure" (period: 2024-2026) funded by the Polish Ministry of Science and Higher Education (Programme: "Support for the participation of Polish scientific teams in international research infrastructure projects"), agreement number 2024/WK/01.

References

1. Belz, A., Agarwal, S., Shimorina, A., Reiter, E.: A systematic review of reproducibility research in natural language processing. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. pp. 381–393 (2021)
2. Bouthillier, X., Laurent, C., Vincent, P.: Unreproducible research is reproducible. In: International Conference on Machine Learning. pp. 725–734. PMLR (2019)
3. Chrabrowa, A., Dragan, Ł., Grzegorczyk, 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 Thirteenth Language Resources and Evaluation Conference. pp. 4374–4394. European Language Resources Association, Marseille, France (2022), https://aclanthology.org/2022.lrec-1.466
4. Gundersen, O.E., Coakley, K., Kirkpatrick, C.: Sources of irreproducibility in machine learning: A review. arXiv preprint arXiv:2204.07610 (2022)
5. 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
6. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L.: 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. pp. 7871–7880. Association for Computational Linguistics, Online (2020). https://doi.org/10.18653/v1/2020.acl-main.703
7. Martinc, M., Škrlj, B., Pollak, S.: TNT-KID: Transformer-based neural tagger for keyword identification. Natural Language Engineering **28**(4), 409–448 (2022). https://doi.org/10.1017/S1351324921000127
8. Morales-Hernández, R.C., Juagüey, J.G., Becerra-Alonso, D.: A comparison of multi-label text classification models in research articles labeled with sustainable development goals. IEEE Access **10**, 123534–123548 (2022)
9. Pezik, P., Mikolajczyk, A., Wawrzynski, A., Niton, B., Ogrodniczuk, M.: Keyword extraction from short texts with a text-to-text transfer transformer. In: Recent

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63759-9_14

Challenges in Intelligent Information and Database Systems - 14th Asian Conference, ACIIDS 2022, Ho Chi Minh City, Vietnam, November 28-30, 2022, Proceedings. Communications in Computer and Information Science, vol. 1716, pp. 530–542. Springer (2022). https://doi.org/10.1007/978-981-19-8234-7"˙41

10. Pezik, P., Mikolajczyk, A., Wawrzynski, A., Zarnecki, F., Niton, B., Ogrodniczuk, M.: Transferable keyword extraction and generation with text-to-text language models. In: Computational Science - ICCS 2023 - 23rd International Conference, Prague, Czech Republic, July 3-5, 2023, Proceedings, Part II. Lecture Notes in Computer Science, vol. 14074, pp. 398–405. Springer (2023)
11. Piedboeuf, F., Langlais, P.: A new dataset for multilingual keyphrase generation. In: Advances in Neural Information Processing Systems. vol. 35, pp. 38046–38059. Curran Associates, Inc. (2022)
12. Pogoda, M., Oleksy, M., Wojtasik, K., Walkowiak, T., Bojanowski, B.: Open versus closed: A comparative empirical assessment of automated news article tagging strategies. In: Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 27th International Conference KES-2023, Athens, Greece, 6-8 September 2023. Procedia Computer Science, vol. 225, pp. 3203–3212. Elsevier (2023). https://doi.org/10.1016/J.PROCS.2023.10.314
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. Song, M., Feng, Y., Jing, L.: A survey on recent advances in keyphrase extraction from pre-trained language models. In: Vlachos, A., Augenstein, I. (eds.) Findings of the Association for Computational Linguistics: EACL 2023. pp. 2153–2164. Association for Computational Linguistics, Dubrovnik, Croatia (May 2023). https://doi.org/10.18653/v1/2023.findings-eacl.161, https://aclanthology.org/2023.findings-eacl.161
15. Szyc, K., Walkowiak, T., Maciejewski, H.: Why out-of-distribution detection experiments are not reliable - subtle experimental details muddle the OOD detector rankings. In: Uncertainty in Artificial Intelligence, UAI 2023, July 31 - 4 August 2023, Pittsburgh, PA, USA. Proceedings of Machine Learning Research, vol. 216, pp. 2078–2088. PMLR (2023)
16. Váradi, T., Nyéki, B., Koeva, S., Tadić, M., Štefanec, V., Ogrodniczuk, M., Niton, B., Pezik, P., Barbu Mititelu, V., Irimia, E., Mitrofan, M., 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 Thirteenth Language Resources and Evaluation Conference. pp. 100–108. European Language Resources Association, Marseille, France (2022)
17. Wan, X., Xiao, J.: Single document keyphrase extraction using neighborhood knowledge. In: Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 2. p. 855–860. AAAI'08, AAAI Press (2008)
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 (2021). https://doi.org/10.18653/v1/2021.acl-long.354

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63759-9_14