Data Augmentation for Less Resourced Summarization

Anonymous ACL submission

Abstract

To improve automatic text summarization for less-resourced languages, we explore fine-tuning multilingual pre-trained models in each language with additional data beyond the human-written summaries. We explore three data augmentation strategies to make use of unlabeled Wikipedia articles as additional synthetic training data. We find that the addition of comparatively small amounts of extra data leads to an improvement in ROUGE scores and that the models trained using extractive target summaries maintain novelty above that of models trained on non-extractive targets. We show that the data augmentation strategies lead to improvements in ROUGE scores for each language, and that the best performing augmentation strategy differs across languages.

1 Introduction

Automatic text summarization in higher resource languages, like English, has achieved high scores in automated metrics (Al-Sabahi et al., 2018; Liu et al., 2022; Zhang et al., 2020a). However, for many less resourced languages the task remains challenging. While there are datasets that have coverage for multilingual summarization in less-resourced languages (Giannakopoulos et al., 2015, 2017; Palen-Michel and Lignos, 2023; Hasan et al., 2021), these datasets often still have relatively few examples compared to their higher resourced counterparts. There is often some amount of additional text data available for less-resourced languages, but it is often not annotated.

While prior work has focused on building multilingual summarization models which take advantage of multilingual transfer (Palen-Michel and Lignos, 2023; Hasan et al., 2021), this work focuses on improving on the performance of multilingual pre-trained models fine-tuned using data for only a single language. Multilingual transfer has proven to be a useful strategy for less resourced languages (Wang et al., 2021); however, other works have shown that multilingual models have limits and given enough data, fully monolingual models can perform better (Virtanen et al., 2019; Tanvir et al., 2021). This work takes one step towards exploring how to acquire enough monolingual summarization data for monolingual training to outperform multilingual models. We also examine how to best make use of the additional unlabeled data from Wikipedia and find that for all languages, there's an improvement over baseline performance, but it is not always the same augmentation strategy that does best.

Our contributions are the following: 1) a comparison of different methods for making use of unlabeled data for summarization of less-resourced languages, 2) new state of the art ROUGE scores on the LR-Sum dataset for Sorani Kurdish and Khmer and higher scores for ROUGE1 and ROUGE-L for Armenian and Georgian and closing the gap between monolingual models and multilingual models for other languages, and 3) an analysis of the quality of the model generated summaries using mean novelty scores.

2 Background

The two main approaches to automatic summarization have been extractive and abstractive methods. Extractive models select important sentences in the source article to use as summaries (Luhn, 1958; Radev et al., 2001; Christian et al., 2016). Abstractive models typically cast the problem as a sequence to sequence problem and apply a neural language model (Rush et al., 2015; See et al., 2017; Hsu et al., 2018; Zhang et al., 2020a). Abstractive neural models typically require larger amounts of training data to train. Summarization is largely scored using ROUGE-1, ROUGE-2, and ROUGE-L metrics (Lin, 2004) for evaluation.

Prior work on multilingual summarization has

largely focused on newswire text from higher resourced languages or covers more languages but with very limited data (Scialom et al., 2020; Giannakopoulos et al., 2015, 2017).

Some of the languages have little to no work in summarization, like Armenian (Avetisyan and Broneske, 2023). Others, like Georgian, have been studied in cross-lingual summarization (Turcan et al., 2022) but appear to be underexplored for monolingual summarization. There is a recent effort to create a Kurdish summarization dataset (Badawi, 2023). The Global Voices summarization dataset (Nguyen and Daumé III, 2019) contains some examples of Macedonian. MassiveSumm (Varab and Schluter, 2021) has greater coverage of languages, but is automatically created and recall oriented and has more complicated redistribution requirements, so we did not make use of it in this work.

3 Datasets

For experiments, we use LR-Sum (Palen-Michel and Lignos, 2023). LR-Sum contains summarization data for 40 languages, many of which are also less-resourced. LR-Sum is built using the description field from the Multilingual Open Text corpus (Palen-Michel et al., 2022) and is similar in approach and content to XL-Sum (Hasan et al., 2021). For this work we focused on a small set of languages from LR-Sum which had the very fewest number of examples in the corpus.

As seen in Table 1, many of the languages we work with have fewer than 1,000 examples, which presents a challenge for neural abstractive summarization systems, which typically require large amounts of training data. Despite XL-Sum providing coverage for many other less resourced languages, the languages we examine here are not covered by XL-Sum. While there is little summarization training data for these languages, there is unlabeled text data available in Wikipedia. However, as seen in Table 1, many Wikipedia articles for less resourced languages are quite short in length. After filtering wikipedia articles less than five sentences long, for many of the languages there is substantially less data available than may appear in raw counts of Wikipedia articles. Specifically, Khmer surprisingly has nearly four times as many training examples available in LR-Sum than there are suitable Wikipedia articles.

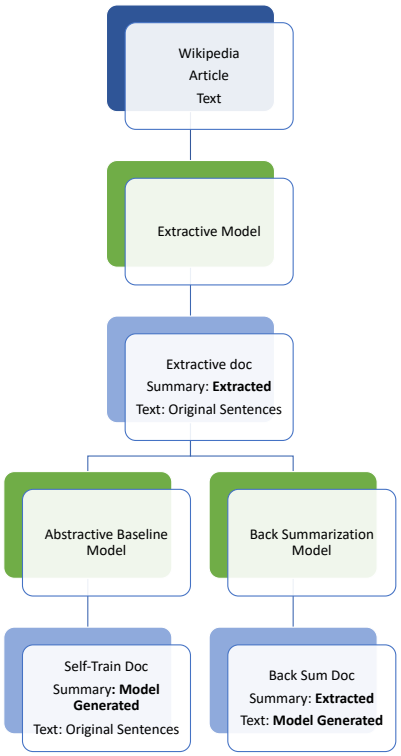

Figure 1: Methodology for generating additional training examples from Wikipedia articles

4 Methodology

We use three approaches for making use of Wikipedia articles as extra synthetic training data for summarization. The approach to creating these extra synthetic training documents is described by Figure 1. We train a baseline sequence to sequence abstractive model using mT5 (Xue et al., 2021). We use the same set of hyperparameters across all experiments. Hyperparameter descriptions can be found in Appendix A.

Extractive-Training: For augmented data first, we use the LexRank LexRank (Erkan and Radev, 2004) extractive summarization algorithm as implemented in sumy[1]. We then directly use these extracted summaries as target summaries alongside the original Wikipedia text.

Self-Training: Second, after fine-tuning an abstractive sequence to sequence model using mT5 as the underlying model, we use it to generate summaries on Wikipedia articles. Self-training approaches of varying levels of complexity have been shown to be useful with other tasks and datasets (Du et al., 2021; Karamanolakis et al., 2021; Meng et al., 2021). These generated summaries and the

[1] https://miso-belica.github.io/sumy/

Language	ISO 639-3	Lang. Family Family	Train LR-Sum	Approximate Wikipedia	Wikipedia Length Filtered
Sorani Kurdish	ckb	Indo-European (Indo-Iranian)	1,230	52,000	18,139
Haitian Creole	hat	French Creole	452	70,200	15,758
Armenian	hye	Indo-European (Isolate)	920	303,000	33,602
Georgian	kat	Kartvelian	511	170,000	105,446
Khmer	khm	Mon-Khmer	3,888	12,000	1,094
Kurmanji Kurdish	kmr	Indo-European (Indo-Iranian)	791	63,100	13,290
Macedonian	mkd	Indo-European (Slavic)	1,223	140,000	103,676

Table 1: Language families and size of training data and available additional data from Wikipedia articles

original Wikipedia text are used for the self training experiment.

Back-summarization: Third, we train a model that when given a summary generates the article associated with the summary. This is motivated by the experiments in Parida and Motlicek (2019), which used a similar approach for German summarization. The approach is also similar to the concept of back-translation (Sennrich et al., 2016) for machine translation where inference is done in the opposite direction to create additional synthetic labeled data.

For each of the three experiments we train on a concatenation of the original training dataset with up to 6k of the synthetic training examples. We choose to use only a subset of available Wikipedia articles in part to have a better balance of synthetic data and real data and also partly for faster experiments.

5 Results

As shown in Table 2, all languages have higher ROUGE scores with the inclusion of additional synthetic training data. Sorani Kurdish, Kurmanji Kurdish, and Armenian in particular have the most substantial increases in ROUGE scores. Armenian using the back sum approach is the only language has a worse score when using augmented data. Of the different strategies for making use of the additional Wikipedia articles, none stands out as being particularly stronger than the others across all languages. Self-training seems to have better scores for ROUGE2 and ROUGE-L when it outperforms the other methods, but the difference tends to be small with the exception of Kurmanji. Khmer had the smallest amount of augmented data since the Khmer Wikipedia articles were quite small and had a relatively small increase in scores.

Hasan et al. (2021) and Palen-Michel and Lignos (2023) found multilingual models to generally perform better than individually trained models. We compare the performance of the best augmented training approach with the reported multilingual model scores from LR-Sum. The best performing augmented training models outperform the multilingual model for Sorani, Khmer, Armenian, and Georgian. It is notable that Armenian and Georgian still have lower R2 scores despite ROUGE1 and ROUGE-L being higher for augmented training individual models.

It is notable that the reported scores for Khmer from Palen-Michel and Lignos (2023) are much lower than the baseline scores we saw for the language. We suspect the reported score from LR-Sum for Khmer may be in error or a difference in tokenization as our baseline method approach was similar and much higher.

6 Discussion

How extractive or abstractive are the summaries? While models trained on synthetic data have an advantage in ROUGE score over the baselines trained on only the human written summaries, it is possible that summaries produced by these models are still lacking in certain ways despite having higher scores. In particular, models trained on Extract-Train or Back-Sum data are being trained on summaries generated from extractive models. One concern could be that these models only learn to copy material from the text rather than synthesizing a novel summary. We further probe this issue by computing mean novelty scores for each summary. This score is the percentage of tokens that do not appear in the article text.

As seen in Table 4, the test set reference summaries have somewhat high novelty. Each model generally has lower mean novelty than the test set. We may have expected model trained on extractive summaries to be generally less novel than those trained on self-training; however this does not appear to be the case. This also shows a hint at why Armenian has low ROUGE scores for the back-sum

	Baseline			Extract-Train			Self-Train			Back-Sum		
Lang.	R1	R2	RL	R1	R2	RL	R1	R2	RL	R1	R2	RL
ckb	13.73	3.69	12.32	**20.39**	**7.27**	**18.54**	18.21	6.15	16.59	17.71	5.63	16.04
hat	19.96	6.21	16.26	**22.93**	6.70	17.95	22.26	**7.23**	**17.96**	22.85	6.99	17.91
hye	17.02	4.37	14.66	**22.51**	7.56	**19.63**	22.10	**7.71**	19.52	7.21	0.16	6.09
kat	11.80	3.18	10.88	13.22	5.02	12.18	15.02	6.98	14.26	**15.22**	**7.14**	**14.41**
khm	22.70	4.82	19.51	22.54	4.71	19.25	23.11	4.75	**20.08**	**23.21**	**5.06**	20.07
kmr	15.99	3.94	14.14	22.19	7.99	19.08	**22.55**	**9.35**	**19.88**	20.91	7.36	17.95
mkd	19.10	6.16	15.74	19.22	5.77	15.73	19.27	6.11	16.00	**19.62**	**6.29**	**16.31**

Table 2: Results of data augmentation experiments for each language

	Best Augmented			Multilingual Reported		
Lang.	R1	R2	RL	R1	R2	RL
ckb	**20.4**	**7.3**	**18.5**	16.6	5.4	15.1
hat	22.3	7.2	17.9	**24.1**	**8.5**	**19.0**
hye	**22.5**	7.6	**19.6**	20.5	**8.5**	17.5
kat	**15.22**	7.1	**14.4**	13.2	**7.2**	12.6
khm	**23.2**	**5.1**	**20.1**	3.7	1.2	3.6
kmr	22.6	9.4	19.9	**25.4**	**12.4**	**22.1**
mkd	19.6	6.3	16.3	**21.3**	**7.6**	**18.0**

Table 3: Comparison between best performing model trained on augmented data and the reported scores from LR-Sum's multilingual model

	LR Sum	Base-line	Extract-Train	Self-Train	Back-Sum
ckb	38.9	3.0	4.6	1.7	2.0
hat	18.1	6.7	2.6	1.4	2.0
hye	35.4	5.7	6.8	4.5	66.0
kat	22.3	19.1	4.2	1.4	1.9
khm	7.8	6.6	7.6	6.9	6.2
kmr	16.3	15.8	3.9	4.1	1.2
mkd	31.4	4.7	6.4	3.7	3.9

Table 4: Mean Novelty for summaries generated by each model and the summaries of the test set (LR-Sum)

approach. With such a high mean novelty score, there is evidence the model is generating a large amount of irrelevant words.

Does augmentation strategy affect length? We examined mean length of generated summaries across all approaches. The mean lengths are shown in Table 6 in Appendix C. Summaries generated from the model trained on synthetic data using an extracted summary tended to have a higher mean length. This is not surprising since extractive summaries being composed of sentences from the original document may have a tendency to be longer and a model trained on this may mimic longer summaries. Between the baseline and self-training, mean lengths varied between being longer or shorter for different languages.

Does bigram mean novelty show different patterns? We also examined mean novelty using bigrams shown in 5 in Appendix C. Bigram mean novelty tends to still be lower than the reference summaries for summaries generated for each model. Again with bigram novelty, the model trained on extractive output surprisingly does not have lower novelty than the baseline in many cases.

What augmentation approach works best? Overall, we found that each data augmentation approach showed an increase in ROUGE scores over the baseline, but there was not one that proved to be definitely better than any other across languages.

7 Conclusion

We have demonstrated three options for generating additional synthetic summaries from Wikipedia articles for summarization in less resourced languages. By filtering Wikipedia for less resourced languages to articles with a suitable length, we noted how large numbers of Wikipedia articles are too short to be used as articles for the task of summarization. We demonstrated that the models trained using extractive target summaries maintain novelty above that of models trained on non-extractive synthetic summaries. In experiments on different varieties synthetic data, we found that the addition of comparatively small amounts of extra data leads to an improvement in ROUGE scores leading to new high scores for some languages in LR-Sum. We did not observe a clear advantage of one method of generating synthetic data over another.

For future work, it would be useful to conduct further experiments strategies for combining synthetic data across augmentation strategies and by combining languages. Other promising directions include determining at what point additional synthetic data leads to diminishing gains in scores and further analysis of the quality of summaries.

8 Limitations and Ethical Considerations

An important limitation to this work is that the evaluation is done entirely with ROUGE score. Limitations to ROUGE score are known and human evaluation is preferred. However, human evaluation can be expensive and especially difficult for less-resourced langauges. Other alternatives to ROUGE have been proposed such as BERTScore (Zhang et al., 2020b), but BERTScore also faces its own challenges (Hanna and Bojar, 2021).

Like any text generation model, automatic summarization is based on statistical properties of language and is likely to sometimes generate statements that may be false. The models and approaches described in this work are primarily for research purposes and summaries generated by these models are only intended to be used to aid human creation of summaries and should be viewed with skepticism regarding their factual content.

References

Kamal Al-Sabahi, Zhang Zuping, and Mohammed Nadher. 2018. A hierarchical structured self-attentive model for extractive document summarization (hssas). *IEEE Access*, 6:24205–24212.

Hayastan Avetisyan and David Broneske. 2023. Large language models and low-resource languages: An examination of Armenian NLP. In *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023 (Findings)*, pages 199–210, Nusa Dua, Bali. Association for Computational Linguistics.

Soran Badawi. 2023. Kurdsum: A new benchmark dataset for the kurdish text summarization. *Natural Language Processing Journal*, 5:100043.

Hans Christian, Mikhael Pramodana Agus, and Derwin Suhartono. 2016. Single document automatic text summarization using term frequency-inverse document frequency (tf-idf). *ComTech: Computer, Mathematics and Engineering Applications*, 7(4):285–294.

Jingfei Du, Edouard Grave, Beliz Gunel, Vishrav Chaudhary, Onur Celebi, Michael Auli, Veselin Stoyanov, and Alexis Conneau. 2021. Self-training improves pre-training for natural language understanding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5408–5418, Online. Association for Computational Linguistics.

Günes Erkan and Dragomir R. Radev. 2004. Lexrank: graph-based lexical centrality as salience in text summarization. *J. Artif. Int. Res.*, 22(1):457–479.

George Giannakopoulos, John M Conroy, Jeff Kubina, Peter A Rankel, Elena Lloret, Josef Steinberger, Marina Litvak, and Benoit Favre. 2017. MultiLing 2017 overview. *MultiLing 2017*, page 1.

George Giannakopoulos, Jeff Kubina, John Conroy, Josef Steinberger, Benoit Favre, Mijail Kabadjov, Udo Kruschwitz, and Massimo Poesio. 2015. MultiLing 2015: multilingual summarization of single and multi-documents, on-line fora, and call-center conversations. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 270–274.

Michael Hanna and Ondřej Bojar. 2021. A fine-grained analysis of BERTScore. In *Proceedings of the Sixth Conference on Machine Translation*, pages 507–517, Online. Association for Computational Linguistics.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online. Association for Computational Linguistics.

Phan Viet Hoang. 2020. Khmer natural language processing tookit. `https://github.com/VietHoang1512/khmer-nltk`.

Wan-Ting Hsu, Chieh-Kai Lin, Ming-Ying Lee, Kerui Min, Jing Tang, and Min Sun. 2018. A unified model for extractive and abstractive summarization using inconsistency loss. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 132–141, Melbourne, Australia. Association for Computational Linguistics.

Giannis Karamanolakis, Subhabrata Mukherjee, Guoqing Zheng, and Ahmed Hassan Awadallah. 2021. Self-training with weak supervision. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 845–863, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yixin Liu, Pengfei Liu, Dragomir Radev, and Graham Neubig. 2022. BRIO: Bringing order to abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2890–2903, Dublin, Ireland. Association for Computational Linguistics.

Hans Peter Luhn. 1958. The automatic creation of literature abstracts. *IBM Journal of research and development*, 2(2):159–165.

Yu Meng, Yunyi Zhang, Jiaxin Huang, Xuan Wang, Yu Zhang, Heng Ji, and Jiawei Han. 2021. Distantly-supervised named entity recognition with noise-robust learning and language model augmented self-training. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10367–10378, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Khanh Nguyen and Hal Daumé III. 2019. Global Voices: Crossing borders in automatic news summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 90–97, Hong Kong, China. Association for Computational Linguistics.

Chester Palen-Michel, June Kim, and Constantine Lignos. 2022. Multilingual open text release 1: Public domain news in 44 languages. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2080–2089, Marseille, France. European Language Resources Association.

Chester Palen-Michel and Constantine Lignos. 2023. LR-sum: Summarization for less-resourced languages. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6829–6844, Toronto, Canada. Association for Computational Linguistics.

Shantipriya Parida and Petr Motlicek. 2019. Abstract text summarization: A low resource challenge. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5994–5998, Hong Kong, China. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Dragomir R Radev, Sasha Blair-Goldensohn, and Zhu Zhang. 2001. Experiments in single and multidocument summarization using MEAD. In *First document understanding conference*, pages 1–7.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389, Lisbon, Portugal. Association for Computational Linguistics.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2020. MLSUM: The multilingual summarization corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8051–8067, Online. Association for Computational Linguistics.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Hasan Tanvir, Claudia Kittask, Sandra Eiche, and Kairit Sirts. 2021. EstBERT: A pretrained language-specific BERT for Estonian. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 11–19, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Elsbeth Turcan, David Wan, Faisal Ladhak, Petra Galuscakova, Sukanta Sen, Svetlana Tchistiakova, Weijia Xu, Marine Carpuat, Kenneth Heafield, Douglas Oard, and Kathleen McKeown. 2022. Constrained regeneration for cross-lingual query-focused extractive summarization. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2668–2680, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Daniel Varab and Natalie Schluter. 2021. MassiveSumm: a very large-scale, very multilingual, news summarisation dataset. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10150–10161, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Antti Virtanen, Jenna Kanerva, Rami Ilo, Jouni Luoma, Juhani Luotolahti, Tapio Salakoski, Filip Ginter, and Sampo Pyysalo. 2019. Multilingual is not enough: Bert for finnish.

Danqing Wang, Jiaze Chen, Hao Zhou, Xipeng Qiu, and Lei Li. 2021. Contrastive aligned joint learning for multilingual summarization. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2739–2750, Online. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Lang.	LR Sum	Base-line	Extract-Train	Self-Train	Back-Sum
ckb	63.94	9.16	11.15	5.17	5.72
hat	50.46	12.74	9.64	5.58	6.53
hye	69.08	17.78	21.68	16.95	97.33
kat	44.02	32.14	11.04	4.22	6.56
khm	25.81	67.44	68.92	68.48	66.78
kmr	41.26	43.38	10.77	8.90	4.06
mkd	66.74	12.99	18.70	11.87	12.22

Table 5: Mean novelty scores using bigrams.

Lang.	LR Sum	Base-line	Extract-Train	Self-Train	Back-Sum
ckb	23.3	25.1	27.9	25.0	26.8
hat	26.7	20.9	31.4	26.9	29.4
hye	24.6	22.2	19.9	16.9	16.8
kat	14.7	17.9	16.7	14.7	15.4
khm	31.6	71.2	74.9	69.0	71.1
kmr	20.2	15.9	27.4	21.2	22.6
mkd	20.0	21.6	21.2	19.9	21.5

Table 6: The mean lengths for all summaries in terms of tokens.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020a. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 11328–11339. PMLR.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020b. Bertscore: Evaluating text generation with bert.

A Hyperparameters

All models used mT5-base as the underlying pretrained model. All models were trained for 3 epochs with 100 warmup_steps. We used a label smoothing factor of 0.1, a beam size of 4, weight_decay of 0.01, a max_target_length 512, max_source_length of 1024, an effective batch size of 32 and a learning rate of 5e-4. Hyperparameters were chosen largely following those suggested in XL-Sum (Hasan et al., 2021) and LR-Sum (Palen-Michel and Lignos, 2023).

B Tokenizers

For Haitian Creole, Georgian, Macedonian, and both varieties of Kurdish, we used utoken[2]. For Armenian, we used Stanza (Qi et al., 2020). and we used khmernltk (Hoang, 2020) for Khmer. The tokenizers used in this work matters both for calculating ROUGE scores and for determining the mean novelty score. For non-latin scripts, using the rouge package in huggingface's evaluate[3] can result in zero or near zero scores for non-latin script languages without explicitly supplying a tokenizer.

C Analysis

We conducted further analysis of generated summaries using bigrams to compute mean novelty and also include the mean length of summaries. We include them here due to space constraints in the paper. Table 5 shows the mean novelty scores for summaries computed using bigrams. Table 6 shows the mean lengths for summaries in the dataset and for summaries generated by each model.

[2] https://github.com/uhermjakob/utoken
[3] https://github.com/huggingface/evaluate