AFRIHG: NEWS HEADLINE GENERATION FOR AFRICAN LANGUAGES

Toyib Ogunremi [1], Serah Akojenu [1], Anthony Soronnadi [1], Olubayo Adekanmbi [1], David Ifeoluwa Adelani [2].
[1] Data Scientists Network
[2] University College London
toyib@datasciencenigeria.ai, d.adelani@ucl.ac.uk

1 INTRODUCTION

Summarization is an established text generation task in NLP, this involves extracting the most important parts or sections that can describe the important or general section of the whole document. There are two popular types of text summarization (Dalal & Malik, 2013): (1) extractive summarization and (2) abstractive summarization. **Extractive** summarization focues on concatenating extracts or excerpts of a document as summary while abstractive summarization focuses on *paraphrasing* or reframing and information compression which capture the main key points in the document. **Abstractive** summarization is popular these days because of its ability to generate diverse and coherent summary of the document being summarized with the application of Neural network architectures (Shukla et al., 2023). In this paper, we focus on **news headline generation**—which can be regarded as a special type of abstractive summarization as it involves carving out information that can serve as the headline of the news or a *one-line summary*. We created AFRIHG— a new dataset with article-headline pair obtained from popular news articles like BBC, VOA, and Isolezwe covering 16 widely spoken languages in Africa. We evaluated the performance of this news headline generation task on two multilingual T5 models: mT5 (Xue et al., 2021) and AfriTeVa V2 (Oladipo et al., 2023). Our evaluation shows a very good performance similar to text summarization, and AfriTeVa V2 performing better than mT5 since it has been pre-trained on many more African languages. However, languages with non-Latin script have a very low (< 4.0 Rouge (Lin, 2004) score 1) on news headline generation unlike the news summarization task where they achieved very good results (> 20.0 Rouge score 1). Lastly, we evaluated Aya LLM (Üstün et al., 2024)—an instruction-tuned mT5-XXL model on several tasks including summarization and headline generation of XL-SUM (Hasan et al., 2021). Our findings shows that AfriTeVa V2 with 313M is competitive to Aya with 13B parameters.

2 DATA AND LANGUAGES COVERED

This paper utilizes news articles and their respective headlines as published by **BBC News**, a division of the British Broadcasting Corporation (BBC). Founded on November 14, 1922, BBC News has evolved into the world's largest broadcast news organization, renowned for its comprehensive global coverage and journalistic integrity. It offers a blend of radio, internet, and television broadcasts, making it an ideal source for diverse and high-quality news content. Notably, BBC News extends its reach beyond high-resource languages like English, French and Arabic, offering content in 10 African languages namely: Amharic, Hausa, Igbo, Kirundi, Nigerian-Pidgin, Oromo, Somali, Kiswahili, Tigrinya, and Yorùbá. Other news sources are Isolezwe for isiXhosa and isiZulu language, and VOA for Shona language.

Benchmark data We collected summarization data from XL-Sum (Hasan et al., 2021)—a dataset containing 1 million news article-summary pairs in 44 languages scraped from BBC news, is the first publicly available abstractive summarization dataset. This dataset also included the headline information for each article. We merged this data with the corpus used for creating MasakhaNEWS corpus (Adelani et al., 2023)—a news topic classification dataset covering 16 languages widely spoken in Africa. We merged these two datasets together and removed duplicates. *We retained the development and test split of the original XL-SUM dataset* for a fair comparison between the tasks

Language	script	Data size train / dev / test	summary mT5-base R1 / R2 / RL	mT5-base R1 / R2 / RL	headline generation AfriTeVa V2 (base) R1 / R2 / RL	Aya R1 / R2 / RL
Amharic (amh)	Ge'ez	16,915 / 719 / 719	20.0 / 7.4 / 18.1	2.7 / 0.1 / 2.6	3.3 / 0.1 / 3.3	**3.4 / 0.1 / 3.4**
Arabic (arb)	Arabic	37,519 / 4,689 / 4689	35.0 / 14.8 / 29.2	1.3 / 0.0 / 1.2	2.0 / 0.1 / 2.0	**2.1 / 0.1 / 2.1**
English (eng)	Latin	311,694 / 11,535 / 11,535	37.6 / 15.2 / 29.9	36.1 / 13.5 / 31.9	**37.9 / 14.5 / 33.4**	32.3 / 11.7 / 28.2
French (fra)	Latin	15,377 / 1,086 / 1,086	35.3 / 16.2 / 28.2	30.6 / 13.7/ 27.2	33.8 / 15.7 / 29.8	**35.9 / 17.5 / 31.0**
Hausa (hau)	Latin	17,174/ 802 / 802	39.4 / 17.7 / 31.7	30.2 / 11.1 / 26.9	**35.0/ 14.9/ 31.5**	34.5 / 14.8 / 30.1
Igbo (ibo)	Latin	8,877 / 522 / 522	31.6 / 10.2 / 24.5	28.5 / 11.2 / 24.6	31.0 / 12.7/ 27.2	**33.4 / 14.9 / 28.7**
Oromo (orm)	Latin	16,417 / 757 / 757	18.7 / 6.2 / 16.2	15.7 / 4.7 / 14.8	18.8 / 6.5 / 17.6	**19.6 / 6.7 / 18.3**
Pidgin (pcm)	Latin	18,214 / 1,151 / 1,151	39.0 / 15.1 / 29.9	31.5 / 10.9 / 27.0	**33.8 / 12.2 / 29.1**	30.5 / 10.5 / 26.2
Kirundi (run)	Latin	9,930 / 718 / 718	32.0 / 14.4 / 25.8	25.2 / 8.9 / 22.5	**29.2 / 11.0 / 25.7**	27.9 / 10.5 / 25.1
Shona (sna)	Latin	9,573 / 1,064 / 1,064	- / - / -	22.6 / 8.1 / 22.1	**25.5 / 9.6 / 24.7**	23.5 / 8.2 / 22.7
Somali (som)	Latin	10,508 / 745 / 745	31.6 / 11.6 / 24.2	24.1 / 7.0 / 21.2	28.2 / 9.6 / 24.7	**28.6 / 10.5 / 24.9**
Swahili (swa)	Latin	18,914 / 987 / 987	37.7 / 17.9 / 30.9	33.0 / 13.4 / 29.1	37.6 / 15.8 / 33.1	**38.9 / 16.7 / 33.9**
Tigrinya (tir)	Ge'ez	12,351 / 681 / 681	25.3 / 8.0 / 21.2	2.6 / 0.1 / 2.5	3.4 / 0.3/ 3.3	**4.0 / 0.1 / 4.1**
Xhosa (xho)	Latin	10,440 / 1,305 / 1,305	- / - / -	13.0 / 4.0/ 12.7	15.2 /5.5 / 14.9	**16.1 / 5.3 / 15.2**
Yoruba (yor)	Latin	15,172 / 793 / 793	31.7 / 11.7 / 25.1	40.0 /14.8 / 31.4	**42.0 / 16.2 / 33.1**	36.0 / 13.5 / 28.3
Zulu (zul)	Latin	14,209 / 1,776 / 1,776	- / - / -	16.0 / 5.0/ 15.5	**17.8 /5.5 / 17.4**	17.0 / 4.9 / 16.2
Average	–	-	- / - / -	22.1 / 7.9 / 19.6	**24.7 / 9.4 / 21.9**	24.0 / 9.1 / 21.2

Table 1: **Rouge score (R1/R2/RL) comparing news summarization and news headline generation**. The baseline summarization model results were obtained from (Hasan et al., 2021). - / - / - indicates the evaluation values that are not available in the XL-Sum

of news summarization and headline generation while also maintaining the 80%-10%-10% TRAIN-DEV-TEST split for the languages not covered by BBC/XL-SUM i.e. chiShona, isiXhosa, and isiZulu. After combining the datasets, we filtered very short articles that are less than 10 characters.

AfriHG dataset We refer to our new dataset as **AfriHG**—it covers 16 languages spoken in Africa, 13 are from the BBC websites, including 10 African languages, Arabic, English, and French. The rest of the languages covered are from the MasakhaNEWS corpus: Shona (obtained from Voice of America) [1], isiXhosa and isiZulu. The last two are from the Isolezwe website [2].

3 BASELINE MODELS

We fine-tune two T5 (Raffel et al., 2020) sequence-sequence models on the news headline generation tasks: mT5-base Xue et al. (2021) and AfriTeVa V2-base (Oladipo et al., 2023). We fine-tune the models using a batch size of 4, number of epochs 3, and the default learning rate of $5e-5$. All models are fine-tuned on Nvidia A10 single GPU using the HuggingFace framework (Wolf et al., 2020). We provide detailed information about the T5-models below:

mt5-base mT5-base is a multilingual T5 model with 580M parameters. It was pretrained on the mC4 corpus with 6.63 trillion tokens covering 101 languages which covers 17 of languages spoken in Africa: Afrikaans, Amharic, Arabic, Chichewa, English, French, Hausa, Igbo, Malagasy, Shona, Sinhala, Somali, Sotho, Kiswahili, isiXhosa, Yorùbá, isiZulu.

Afriteva V2 Afriteva V2 base is an extension of AfriTeVa model(Ogundepo et al., 2022) from 10 African languages to 20 widely spoken languages in Africa.The base model contains about 313M parameters. Similar to mT5-base, it has 12 layers, 12 attention heads and 512 token sequence length. It was pre-trained on WURA dataset [3]—a cleaned version of mC4 corpus with 29.3GB of data. AfriTeVa V2 covers all the languages in the AfriHG dataset.

Aya Aya is an instruction-tuned massively multilingual model (Üstün et al., 2024) available in 101 languages. It is a large language model with 13B parameters built by fine-tuning an XXL-T5 model finetuned on Aya Dataset (Singh et al., 2024) consisting of 25 African languages. Aya covers 12 of the languages evaluated in the AfriHG dataset.

[1] `https://www.voashona.com/`

[2] `https://www.isolezwelesixhosa.co.za/` and `https://www.isolezwe.co.za/`

[3] `https://huggingface.co/datasets/castorini/wura`

4 RESULTS

Table 1 compares the result of news summarization and headline generation. Comparing the two T5 models on headline generation, in all languages we evaluated on, AfriTeVa V2-base was consistently better than the mT5-base with average performance gain in Rouge score of $+2.6/+1.5/+2.3$ (R1/R2/RL). This shows that AfriTeVa V2 is a better pre-trained language model for African languages since it has seen more during pre-training. Probably, the most important result is that we found out that languages with non-Latin scripts like Arabic, Amharic and Tigrinya performed poorly (< 4.0 R1 score) while the other languages that use Latin script have comparable performance similar to the news summarization task. Finally, we evaluated on Aya, and show that it is competitive to AfriTeVa V2-base, achieving the best result for 9 out of the 16 languages. However, it still struggles with non-Latin script. While Aya LLM has a more than 43 times more parameters than AfriTeVa V2, and has seen the same training data, the performance are very similar. This shows that there is still some advantage in using the *fine-tune paradigm* rather than *prompting* when there is abundance of training data in this LLM age.

5 CONCLUSION

In this paper, we presented AfriHG, an extended African news headline generation corpus compiled from XL-SUM dataset and MasakhaNEWS corpus. We added the MasakhaNEWS corpus to widen the language coverage while also providing more news-categories which are not available in the XL-Sum dataset for a more-inclusive evaluation. We performed experiments on this dataset using state-of-the-art multilingual pre-trained language models and demonstrated the capability of these seq2seq models in generating headlines for news articles in various African languages. Future work will focus on investigating why LLMs struggle to perform well on headline generation task for Semitic language families or most importantly non-latin script.

REFERENCES

David Ifeoluwa Adelani, Marek Masiak, Israel Abebe Azime, Jesujoba Alabi, Atnafu Lambebo Tonja, Christine Mwase, Odunayo Ogundepo, Bonaventure F. P. Dossou, Akintunde Oladipo, Doreen Nixdorf, Chris Chinenye Emezue, sana al azzawi, Blessing Sibanda, Davis David, Lolwethu Ndolela, Jonathan Mukiibi, Tunde Ajayi, Tatiana Moteu, Brian Odhiambo, Abraham Owodunni, Nnaemeka Obiefuna, Muhidin Mohamed, Shamsuddeen Hassan Muhammad, Teshome Mulugeta Ababu, Saheed Abdullahi Salahudeen, Mesay Gemeda Yigezu, Tajuddeen Gwadabe, Idris Abdulmumin, Mahlet Taye, Oluwabusayo Awoyomi, Iyanuoluwa Shode, Tolulope Adelani, Habiba Abdulganiyu, Abdul-Hakeem Omotayo, Adetola Adeeko, Abeeb Afolabi, Anuoluwapo Aremu, Olanrewaju Samuel, Clemencia Siro, Wangari Kimotho, Onyekachi Ogbu, Chinedu Mbonu, Chiamaka Chukwuneke, Samuel Fanijo, Jessica Ojo, Oyinkansola Awosan, Tadesse Kebede, Toadoum Sari Sakayo, Pamela Nyatsine, Freedmore Sidume, Oreen Yousuf, Mardiyyah Oduwole, Tshinu Tshinu, Ussen Kimanuka, Thina Diko, Siyanda Nxakama, Sinodos Nigusse, Abdulmejid Johar, Shafie Mohamed, Fuad Mire Hassan, Moges Ahmed Mehamed, Evrard Ngabire, Jules Jules, Ivan Ssenkungu, and Pontus Stenetorp. Masakhanews: News topic classification for african languages, 2023.

Vipul Dalal and Latesh Malik. A survey of extractive and abstractive text summarization techniques. In *2013 6th International Conference on Emerging Trends in Engineering and Technology*, pp. 109–110, 2013. doi: 10.1109/ICETET.2013.31.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 4693–4703, Online, August 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.findings-acl.413.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W04-1013.

Odunayo Jude Ogundepo, Akintunde Oladipo, Mofetoluwa Adeyemi, Kelechi Ogueji, and Jimmy Lin. AfriTeVA: Extending ?small data? pretraining approaches to sequence-to-sequence models. In *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*, pp. 126–135, Hybrid, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.deeplo-1.14. URL https://aclanthology.org/2022.deeplo-1.14.

Akintunde Oladipo, Mofetoluwa Adeyemi, Orevaoghene Ahia, Abraham Owodunni, Odunayo Ogundepo, David Adelani, and Jimmy Lin. Better quality pre-training data and t5 models for African languages. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 158–168, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.11. URL https://aclanthology.org/2023.emnlp-main.11.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

Karishma Shukla, Kartik Barange, Prajakta Shahabade, Akanksha Pandey, and Dnyaneshwar Bavkar. Abstractive text summarization using transformer based approach. 06 2023. doi: 10.55041/IJSREM22369.

Shivalika Singh, Freddie Vargus, Daniel Dsouza, Börje F. Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura OMahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Souza Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergün, Ifeoma Okoh, Aisha Alaagib, Oshan Mudannayake, Zaid Alyafeai, Vu Minh Chien, Sebastian Ruder, Surya Guthikonda, Emad A. Alghamdi, Sebastian Gehrmann, Niklas Muennighoff, Max Bartolo, Julia Kreutzer, Ahmet Üstün, Marzieh Fadaee, and Sara Hooker. Aya dataset: An open-access collection for multilingual instruction tuning, 2024.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Online, October 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.emnlp-demos.6.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 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*, pp. 483–498, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.41. URL https://aclanthology.org/2021.naacl-main.41.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. Aya model: An instruction finetuned open-access multilingual language model, 2024.