XWikiGen: Cross-lingual Summarization for Encyclopedic Text Generation in Low Resource Languages

Dhaval Taunk
IIIT Hyderabad

Shivprasad Sagare
IIIT Hyderabad

Anupam Patil
SCTR's PICT Pune

Shivansh Subramanian
IIIT Hyderabad

Manish Gupta
Microsoft India, IIIT Hyderabad

Vasudeva Varma
IIIT Hyderabad

Abstract

Automated text generation for *low-resource LR languages*, especially on Wikipedia, is a critical problem due to the lack of contributors who can write encyclopedic text. Previous work on Wikipedia text generation has only focused on English language reference articles, which are used to generate English Wikipedia pages. However, for low-resource languages, there is a scarcity of reference articles in same language. To address this problem, this study proposes a new task called XWikiGen, a cross-lingual multi-document summarization approach, which involves generating Wikipedia-style text by summarizing multiple reference articles written in different languages. To enable this task, we have created a benchmark dataset, called XWikiRef, which includes around ~69K Wikipedia articles in five domains and eight languages. We use this dataset to train a two-stage system that takes a set of citations and a section title as input and generates a section-specific summary for low-resource languages. We make our code and dataset publicly available[1].

Keywords: XWikiGen, deep learning, cross-lingual summarization, Wikipedia text generation, low resource NLG

Introduction

Millions of users prefer using Wikipedia as their primary reference for information, but, low-resource (LR) language Wikipedia comprises ~90K articles across 7 languages vs. approximately 6.56 million articles in English. One approach to generating LR Wikipedia content is to translate equivalent English Wikipedia articles, but many local entities of interest lack equivalent pages in English. We propose a new dataset called XWikiRef that includes ~69K Wikipedia articles from eight languages (Bengali (bn), English (en), Hindi (hi), Malayalam (ml), Marathi (mr), Oriya (or), Punjabi (pa) and Tamil (ta)) and five domains (books, films, politicians, sportsmen, and writers). The task of generating long text in a cross-lingual manner is challenging, so we use a two-stage approach involving neural models: the first stage identifies important sentences from reference documents, and the second stage generates section text in an abstractive manner.

[1] https://github.com/DhavalTaunk08/XWikiGen

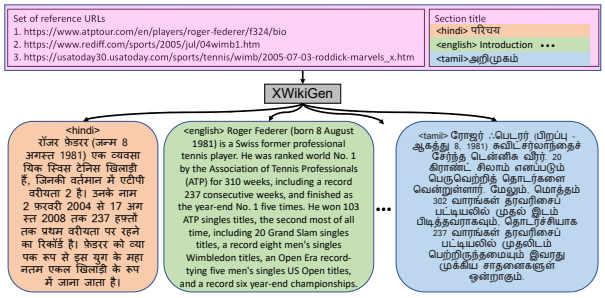

Figure 1: XWikiGen examples: Generating Hindi, English, and Tamil text for the Introduction section from cited references.

Dataset details

The XWikiRef dataset includes Wikipedia articles that are categorized into five different domains and are available in eight different languages including. Each sample in the dataset includes information about the domain, language, section title, a set of reference URLs, and the actual Wikipedia section text. The dataset is divided into three parts, namely train, validation, and test sets, which are stratified by domain and language, with a 60:20:20 ratio. Table 1a shows the distribution of the dataset across the different domains and languages.

Methodology

We follow a 2 stage approach for XWikiGen which consists a extractive summarization followed by a abstractive summarization step.

Extractive Summarization Stage

The extractive stage's objective is to choose a group of sentences from a given set of URLs that provide a summary of the set. Two extractive summarization techniques, Salience and HipoRank, are being tested. In both approaches, the input includes the section title and a list of reference URLs.

Salience based extractive summarization: The salience based extractive summarization aims to identify the most important sentences from the reference URLs, which are relevant to a specific section title. This technique is inspired by a relevance scoring method used by (Yasunaga et al., 2021) & (Taunk et al., 2023) a language model for question-answering, where each answer entity's relevance score is calculated relative to the QA context.

HipoRank based extractive summarization: HipoRank (Dong et al., 2020) is an unsupervised graph-based model designed for extractive summarization of long documents. It creates a directed hierarchical graph with sentence and section nodes and sentence-sentence and sentence-section edges, with asymmetrically weighted edges. The model computes the importance score for a sentence node based on a weighted sum of edges incident on the node and selects the top-K sentences as the extractive summary.

Abstractive Summarization Stage

After the extractive stage, the output is often incoherent and in the reference text language. An abstractive stage is needed to generate coherent summaries in the target language. For this stage, we use two state-of-the-art multi-lingual natural language generation models, mBART-large (Liu et al., 2020) and mT5-base (Xue et al., 2021), which have been shown to be effective across multiple NLP tasks. Both models take the target language ID, article title, section title, and top-K sentences from the extractive stage as input.

Multi-lingual, Multi-domain, and Multi-lingual-Multi-domain setups

Prior research on multi-lingual and cross-lingual natural language generation has demonstrated that multi-lingual models are superior to individual models, particularly for low-resource languages. As this study concentrates on such languages, we experiment with 3 types of configurations namely multi-lingual, multi-domain, and multi-lingual-multi-domain configurations.

Results

Table 1b presents the results of various extractive and abstractive methods, training setups, and evaluation metrics on the test instances in the data set. The best overall results are achieved with the multi-lingual-multi-domain training setup, and the combination of HipoRank with mBART produces the best results within this setup. The detailed results of thie setting is shown in table 2 The HipoRank model is expected to perform well because it combines the knowledge of the pretrained mBERT model with the hierarchical document structure. For the multi-lingual setup, the HipoRank+mBART combination produces the best results, while for the multi-domain setup, Salience+mBART performs better. Additionally, it is worth noting that Salience uses XLM-RoBERTa (270M parameters), while HipoRank uses mBERT (110M parameters), and mT5 and mBART contain 580M and 610M parameters, respectively. Finally, the average output length of the best model is 221 words. Figure 2 shows the predictions of our model on 2 languages and domain pairs.

Conclusion

In this work, we highlighted the need for generating Wikipedia summaries for low-resource languages and introduced a new dataset in eight languages and five domains. We proposed a crosslingual-multidocument summarization based two-stage system that combines an extractive stage using HipoRank with an abstractive stage using mBART. We conducted experiments and found that the multi-lingual-multi-domain model using HipoRank and mBART produced the best results. We have made our dataset and code publicly available for future research in this area.

References

[Dong et al.2020] Yue Dong, Andrei Mircea, and Jackie CK Cheung. 2020. Discourse-aware unsupervised summarization of long scientific documents. *arXiv preprint arXiv:2005.00513*.

[Liu et al.2020] Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

[Taunk et al.2023] Dhaval Taunk, Lakshya Khanna, Pavan Kandru, Vasudeva Varma, Charu Sharma, and Makarand Tapaswi. 2023. GrapeQA: GRaph Augmentation and Pruning to Enhance Question-Answering. *arXiv preprint arXiv:2303.12320*.

[Xue et al.2021] 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.

[Yasunaga et al.2021] Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, and Jure Leskovec. 2021. Qa-gnn: Reasoning with language models and knowledge graphs for question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 535–546.

Domain/ Lang	Books	Film	Politicians	Sportsmen	Writers	Total
bn	313	1501	2006	5470	1603	10893
hi	922	1025	3927	6334	2024	14232
ml	458	2919	2513	1783	2251	9924
mr	87	480	988	2280	784	4619
or	73	794	1060	319	498	2744
pa	221	421	1123	1975	2245	5985
ta	493	3733	4932	2552	1940	13650
en	1467	1810	1628	919	714	6538
Total	4034	12683	18177	21632	12059	**68585**

(a) XWIKIREF: Total #articles per domain per language

	Extractive	Abstractive	ROUGE-L	chrF++	METEOR
Multi-lingual	Salience	mBART	15.59	17.20	10.98
	Salience	mT5	14.66	15.45	8.92
	HipoRank	mBART	**16.96**	**19.11**	**12.19**
	HipoRank	mT5	15.98	17.11	10.08
Multi-domain	Salience	mBART	**19.88**	**22.82**	**15.00**
	Salience	mT5	12.13	13.66	7.27
	HipoRank	mBART	18.87	20.79	14.10
	HipoRank	mT5	12.29	13.93	7.36
Multi-lingual-multi-domain	Salience	mBART	20.50	22.32	14.81
	Salience	mT5	17.31	18.77	11.57
	HipoRank	mBART	**21.04**	**23.44**	**15.35**
	HipoRank	mT5	17.65	19.04	11.74

(b) XWIKIGEN Results across multiple training setups and (extractive, abstractive) methods on test part of XWIKIREF. Best results per block are highlighted in bold. Overall best results are also underlined.

	ROUGE-L					chrf++					METEOR				
	writers	books	sportsmen	politicians	films	writers	books	sportsmen	politicians	films	writers	books	sportsmen	politicians	films
bn	10.61	9.43	15.78	17.46	15.75	14.72	14.19	20.28	21.21	20.03	6.13	5.66	10.56	12.99	10.39
en	13.04	15.62	18.53	13.32	20.15	19.71	18.90	22.80	20.00	24.13	10.65	11.62	13.89	11.47	15.09
hi	33.23	58.71	28.48	53.18	21.46	31.05	51.99	26.99	52.05	19.64	28.49	53.78	21.46	51.65	15.30
mr	15.37	17.00	26.77	20.06	24.15	14.68	16.24	26.84	18.12	21.82	7.40	9.50	20.14	10.74	14.30
ml	8.96	10.93	12.97	14.36	24.19	13.35	12.18	15.42	18.01	26.51	3.92	4.77	6.14	7.73	16.16
or	13.15	12.31	9.38	43.76	26.66	14.44	15.16	10.51	44.17	29.27	5.67	9.14	5.28	40.89	23.30
pa	14.96	12.35	24.54	16.59	17.15	13.42	12.39	21.32	14.02	13.82	8.59	7.48	16.54	9.80	9.63
ta	10.62	11.85	18.94	19.18	24.90	16.43	17.63	23.98	23.77	29.94	4.89	6.29	10.03	11.24	17.05

Table 2: Detailed results for every (domain, language) partition of the test set of our XWIKIREF dataset, for our best XWIKIGEN model: Multi-lingual-multi-domain HipoRank+mBART.

Domain	Reference URLs	Entity/ Section	Lang	Reference Text	Generated Summary
Films	• http://www.thehindu.com/features/cinema/cinema-columns/blast-from-the-past-chandra-mohana-or-samooga-thondu-1936/article5668076.ece	சந்திர மோகனா (திரைப்படம்)/அறிமுகம்	ta	சந்திரமோகனா என்பது 1936 ஆம் ஆண்டு வெளிவந்த தமிழ்த் திரைப்படமாகும். முதுரை மோகன் மூவிடோன் நிறுவனத்தினரின் தயாரிப்பில் வெளிவந்த இத்திரைப்படத்தில் எம். கே. ராதா, எஸ். வி. வெங்கட்ராமன் மற்றும் பலரும் நடித்துள்ளனர். இப்படமானது கொத்தமங்கலம் எஸ். எம். சுப்ரமண்யம் என்று அப்போது அழைக்கப்பட்ட கொத்தமங்கலம் சுப்பு எழுதிய சந்திரமோகனா அல்லது சமுதாயத் தொண்டு என்ற புதினத்தை அடிப்படையாகக் கொண்டு எடுக்கப்பட்டது.	சந்திர மோகனா (Chandra Mohana) 1947 ஆம் ஆண்டு வெளிவந்த தமிழ்த் திரைப்படமாகும். எஸ். எஸ். வாசன் இயக்கத்தில் வெளிவந்த இத்திரைப்படத்தில் எம். கே. ராதா, டி. ஆர். ராமச்சந்திரன் மற்றும் பலரும் நடித்திருந்தனர்.
Sportsmen	• https://www.icc-cricket.com/news/1939383 • https://www.icc-cricket.com/media-releases/1212091 • http://www.espncricinfo.com/story/_/id/25033717/associates-pathway-2023-world-cup-undergoes-major-revamp	2021 पापुआ न्यू गिनी त्रिकोणी सीरीज (मई)/परिचय	hi	2021 पापुआ न्यू गिनी त्रिकोणी सीरीज़ 2019–2023 आईसीसी क्रिकेट विश्व कप लीग 2 क्रिकेट टूर्नामेंट का 8 वां दौर होने वाला था, जो मई 2021 में पापुआ न्यू गिनी में खेला जाना था। यह नामीबिया, पापुआ न्यू गिनी और संयुक्त राज्य अमेरिका की क्रिकेट टीमों के बीच एक त्रिकोणीय राष्ट्र श्रृंखला होती, जिसमें मैच एक दिवसीय अंतर्राष्ट्रीय (वनडे) जुड़नार के रूप में खेले जाते थे। आईसीसी क्रिकेट विश्व कप लीग 2 2023 क्रिकेट विश्व कप के लिए योग्यता मार्ग का हिस्सा है। हालाँकि, 12 फरवरी 2021 को, कोविड-19 महामारी के कारण श्रृंखला को स्थगित कर दिया गया था।	2021 पापुआ न्यू गिनी ट्राई-नेशन सीरीज़ एक क्रिकेट टूर्नामेंट था जो सितंबर 2021 में पापुआ न्यू गिनी में खेला गया था। यह पापुआ न्यू गिनी क्रिकेट टीम और पापुआ न्यू गिनी क्रिकेट बोर्ड (पीएनसीबी) के बीच एक त्रिकोणीय राष्ट्र श्रृंखला थी, जिसमें पापुआ न्यू गिनी और संयुक्त राज्य अमेरिका के बीच एक दिवसीय अंतर्राष्ट्रीय (वनडे) मैच खेला गया था। यह श्रृंखला सितंबर 2021 में होने वाली थी, लेकिन कोविड-19 महामारी के कारण इसे स्थगित कर दिया गया था।

Figure 2: Some examples of XWIKIGEN using our best model