XOutlineGen: Cross-lingual Outline Generation for Encyclopedic Text in Low Resource Languages

Shivansh Subramanian
IIIT Hyderabad

Dhaval Taunk
IIIT Hyderabad

Manish Gupta
Microsoft India,
IIIT Hyderabad

Vasudeva Varma
IIIT Hyderabad

Abstract

One crucial aspect of content organization is the creation of article outlines, which summarize the primary topics and subtopics covered in an article in a structured manner. This paper introduces a solution called XOutlineGen, which generates cross-lingual outlines for encyclopedic texts from reference articles. XOutlineGen uses the XWikiRef dataset, which consists of encyclopedic texts generated from reference articles and section titles. The dataset is enhanced with two new languages and three new domains, resulting in ~92K articles. Our pipeline employs this dataset to train a two-step generation model, which takes the article title and set of references as inputs and produces the article outline.

Keywords: XOutlineGen, deep learning, cross-lingual generation, Wikipedia outline generation, low resource NLG

Introduction

Wikipedia is a global encyclopedia where people can contribute and share knowledge on various subjects. While it is a significant source of information, it has a notable shortage of articles in low-resource languages. To tackle this challenge, XWikiGen (Taunk et al., 2023) was developed to take reference texts in multiple languages as input and generate a coherent Wikipedia style section text in the target language, given the article title and references.

The articles in low-resource languages often contain entities that are specific to the region and not well-known globally. To create an outline for such an article, one approach is to translate or copy the outline from a similar article in English or the same language within the same domain. However, these methods require the user to be familiar with the entity and to be able to identify other similar articles with outlines that can be used as a reference. Additionally, the lack of reference information in low-resource languages is a challenge. This paper introduces a new task called XOutlineGen, which involves generating a cross-lingual outline of a Wikipedia article based on its references.

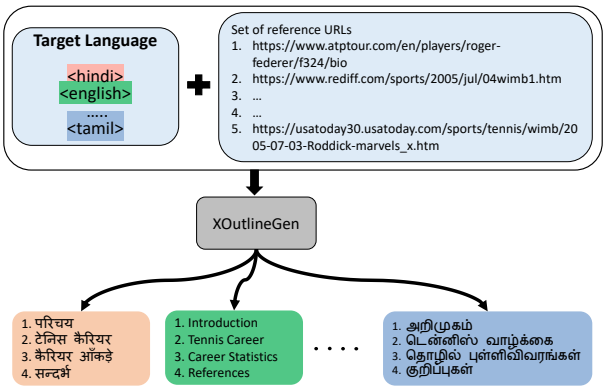

Figure 1: XOutlineGen examples: Generating Hindi, English, and Tamil Outline from cited references.

Overall we make the following contributions in this paper:

- We motivate the need for the problem XOutlineGen, a cross-lingual outline generation task where the input is (article title, reference text) and the output is the outline for a Wikipedia article.
- We extend the XWikiRef dataset to ~92k articles by adding 3 new domains and 2 new languages.
- We model XOutlineGen as a multi-document cross-lingual outline generation problem and propose a two-stage system with reward functions in a Reinforcement Learning based setting.

Dataset details - XWikiRefV2

The XWikiRef (Taunk et al., 2023) dataset included Wikipedia articles that were categorized into five different domains, namely books, films, politicians, sportsmen, and writers, and were available in eight different languages including bn, en, hi, ml, mr, or, pa, and ta. In the second version of the dataset, we are adding two more languages (kn, te) and 3 more domains (animals, cities, companies). Table 1 shows the overall distribution of the dataset across each language and domain pair.

Methodology

The articles on Wikipedia can have references in various languages. To create an outline of an article that is

© Copyright held by the owner/author(s), published under Creative Commons CC BY 4.0 License

cross-lingual, we used a two-stage approach. We first performed extractive summarization in stage 1 to extract the relevant information from reference text. This information was then used in stage two to create the overall outline. Additionally, we included RL based reward functions in stage 2 of the pipeline to ensure that the model produces a coherent and relevant outline.

Extractive Summarization Stage

HipoRank based extractive summarization: HipoRank (Dong et al., 2020) is an unsupervised model that uses graphs to summarize lengthy documents. The model constructs a hierarchical graph comprising sentence and section nodes along with sentence-sentence and sentence-section edges, which are asymmetrically weighted. The model then computes a weighted sum of the edge scores attached to each node and selects the top K most relevant sentences based on these scores.

Outline Generation Stage

The first stage of the pipeline of extracting information may produce output that is incoherent and in the reference text language. Therefore, the second stage is required to create coherent output. We experimented with two different multi-lingual natural language generation models, mBART-large (Liu et al., 2020) and mT5-base (Xue et al., 2021) and compared the performance of both the models. Since the input for stage 2 is reference text, which can have varying styles and subjects, it is important to ensure that the generated outline is compatible with the source text. To achieve this we took inspiration from (Maheshwari et al., 2022) and added 2 reward functions to the generation model in a reinforcement learning setting to ensure that the generated outline meets the desired criteria.

Section-title compatibility reward: To determine if the section title produced is consistent with the input text, we finetuned XLM-RoBERTa (Conneau et al., 2020) based binary classifier, which assesses whether the generated section title and the input reference text are coherent.

Entity Correctness Reward: Entity Correctness Reward looks for hallucinations in the generated title by checking whether the entities in generated section title are present in input text or not by using IndicNER (Mhaske et al., 2022) to extract the named entities from the generated title and the input reference sentences.

Results

We performed out initial experiments on XWikiRef V2 which has ~92K articles. We divide the dataset into train, val and test splits in 70:10:20 ratio and trained our pipeline in a multi-lingual - multi-domain setting. We experimented with both mBART-large and mT5-base models and achieve an average ROUGE-L score of 0.436 and 0.481 respectively on a scale of 0-1. Table 2 and 3 show results on our initial experiments using mBART and mT5 respectively.

Future Work

We presented results on XWikiRefV2 using mBART and mT5 models, and plan to conduct further experiments with hyper-parameter tuning to improve the results and generate a better outline. We hope that our work will help the community generate more useful outlines of articles, aiding in creation of Wikipedia articles in LR languages.

References

[Conneau et al.2020] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale.

[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.

[Maheshwari et al.2022] Himanshu Maheshwari, Nethraa Sivakumar, Shelly Jain, Tanvi Karandikar, Vinay Aggarwal, Navita Goyal, and Sumit Shekhar. 2022. DynamicTOC: Persona-based table of contents for consumption of long documents. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5133–5143, Seattle, United States, July. Association for Computational Linguistics.

[Mhaske et al.2022] Arnav Mhaske, Harshit Kedia, Sumanth Doddapaneni, Mitesh M. Khapra, Pratyush Kumar, Rudra Murthy, and Anoop Kunchukuttan. 2022. Naamapadam: A large-scale named entity annotated data for indic languages.

[Taunk et al.2023] Dhaval Taunk, Shivprasad Sagare, Anupam Patil, Shivansh Subramanian, Manish Gupta, and Vasudeva Varma. 2023. XWikiGen: Cross-lingual Summarization for Encyclopedic Text Generation in Low Resource Languages. *arXiv preprint arXiv:2303.12308*.

[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.

© Copyright held by the owner/author(s), published under Creative Commons CC BY 4.0 License

Domain/Lang	bn	hi	kn	ml	mr	or	pa	ta	te	en	Total
Animals	752	276	136	1459	107	100	104	854	77	2228	6094
Books	493	876	37	435	83	69	210	468	85	3000	5757
Cities	687	693	122	409	188	225	306	598	227	1586	5042
Companies	573	408	146	326	200	35	105	453	133	3000	5380
Films	2886	1525	731	2773	456	754	400	2961	2517	3000	18004
Politicians	3000	3000	343	2387	939	1007	1067	3000	1381	1570	17694
Sportsmen	3000	3000	428	1694	2166	303	1877	2424	433	3000	18326
Writers	2483	1928	474	2138	744	473	2133	1843	796	2634	15647
Total	13875	11707	2418	11622	4883	2967	6203	12602	5650	20019	**91944**

Table 1: XWIKIREF: Total #articles per domain per language

Domain/Lang	bn	en	kn	te	hi	ml	mr	or	pa	ta	Average
animals	0.368	0.359	0.044	0.056	0.219	0.646	0.307	0.177	0.071	0.120	**0.237**
books	0.361	0.487	0.015	0.333	0.914	0.288	0.355	0.686	0.032	0.530	**0.400**
cities	0.125	0.530	0.100	0.064	0.498	0.036	0.250	0.409	0.023	0.066	**0.210**
companies	0.283	0.429	0.100	0.154	0.338	0.158	0.245	0.000	0.000	0.209	**0.192**
films	0.536	0.556	0.566	0.601	0.482	0.650	0.400	0.741	0.083	0.568	**0.518**
politicians	0.576	0.267	0.263	0.319	0.746	0.385	0.402	0.601	0.105	0.408	**0.407**
sportsman	0.321	0.509	0.295	0.227	0.580	0.318	0.577	0.204	0.350	0.298	**0.368**
writers	0.319	0.301	0.211	0.220	0.534	0.217	0.426	0.504	0.122	0.197	**0.305**
Average	**0.361**	**0.430**	**0.199**	**0.247**	**0.539**	**0.337**	**0.370**	**0.415**	**0.098**	**0.300**	**0.436**

Table 2: XOUTLINEGEN results on HipoRank + mBART methodology

Domain/Lang	bn	en	kn	te	hi	ml	mr	or	pa	ta	Average
animals	0.467	0.548	0.053	0.044	0.464	0.572	0.322	0.170	0.071	0.197	**0.291**
books	0.430	0.438	0.018	0.406	0.913	0.337	0.293	0.537	0.190	0.529	**0.409**
cities	0.446	0.566	0.054	0.067	0.523	0.111	0.117	0.144	0.102	0.077	**0.221**
companies	0.420	0.537	0.069	0.174	0.348	0.275	0.339	0.000	0.000	0.211	**0.237**
films	0.641	0.529	0.532	0.646	0.510	0.700	0.439	0.767	0.277	0.599	**0.564**
politicians	0.642	0.346	0.163	0.371	0.730	0.458	0.365	0.581	0.261	0.433	**0.435**
sportsman	0.472	0.526	0.261	0.237	0.593	0.329	0.541	0.168	0.457	0.314	**0.390**
writers	0.405	0.370	0.139	0.262	0.543	0.289	0.450	0.430	0.344	0.212	**0.345**
Average	**0.490**	**0.482**	**0.161**	**0.276**	**0.578**	**0.384**	**0.358**	**0.350**	**0.213**	**0.321**	**0.481**

Table 3: XOUTLINEGEN results on HipoRank + mT5 methodology

© Copyright held by the owner/author(s), published under Creative Commons CC BY 4.0 License