Cross-lingual Multi-Sentence Fact-to-Text Generation: Generating factually grounded Wikipedia Articles using Wikidata

Bhavyajeeet Singh IIIT Hyderabad
Aditya Hari IIIT Hyderabad
Rahul Mehta IIIT Hyderabad
Tushar Abhishek Microsoft, IIIT Hyderabad,
Manish Gupta Microsoft, IIIT Hyderabad
Vasudeva Varma IIIT Hyderabad

Abstract

Fact-to-text generation can allow for the generation of high-quality, informative texts such as Wikipedia articles. Cross-lingual fact-to-text generation (XF2T) involves using facts available in a language, typically English, and generating texts in a different language based on these facts. This is particularly relevant for low and medium-resource languages, which have relatively structured informative content. This work explores the problem of XF2T for generating long text from given facts with a specific focus on generating factually grounded content. Unfortunately, previous work either focuses on cross-lingual facts to *short text* or *monolingual* graph to text generation. In this paper, we propose a novel solution to the multi-sentence XF2T task, which addresses these challenges by training multilingual Transformer-based models with coverage prompts and rebalanced beam search, and further improving the quality by defining task-specific reward functions and training on them using reinforcement learning.

Keywords: XF2T, text generation, cross-lingual, NLG evaluation, low resource NLG

Introduction

Fact-to-text generation is a data-to-text generation problem in which natural language descriptions are generated from structured facts. This is important for generating informative texts such as Wikipedia articles. However, monolingual fact-to-text generation tends to suffer from the problem of data sparsity for low-resource languages. To address this, cross-lingual fact-to-text generation (XF2T) is proposed, which involves generating text using facts available in a different language, typically a high-resource language like English. A highly desirable aspect of the XF2T task is producing reliable content which is grounded in the facts. Large language models, including the recent ones like GPT-3 or chatGPT, have been prone to hallucination or diverging from the input facts. Thus, it becomes crucial to investigate ways to address this problem. In this work, we investigate multi-sentence XF2T and its associated challenges. We present a dataset for multi-sentence XF2T based on the existing XAlign dataset(Abhishek et al., 2022), with a high-quality test dataset partitioned based on coherence and coverage. Further, we investigate various methods, such as explicit clustering of facts, using coverage prompts and grounded decoding, which attend to the source and show their impact in improving performance.

Language	**Train**	**Val**	**Test**
Assamese	799	159	111
Bengali	14,858	2,968	1,984
English	32,176	6,427	4,292
Gujarati	901	179	121
Hindi	9,266	1,850	1,239
Kannada	2,026	404	273
Malayalam	8,363	1,671	1,117
Marathi	5,394	1,077	722
Odia	1,742	348	237
Punjabi	5,454	1,085	731
Tamil	10,026	2,004	1,340
Telugu	2,820	563	379
Total	**93,825**	**18,735**	**12,546**

Table 1: Train, validation and test splits for each language

Dataset

We use the existing XAlign dataset to construct our dataset. The sentences available for an entity are concatenated as per their order in the original Wikipedia article to create multi-sentence descriptions. In total, the dataset contains 125,106 paragraphs across 12 different languages. This is summarized in table 1.

To create a high-quality test dataset, the instances were partitioned based on two metrics - coherence and coverage. Coherence is defined as the quality of being ”logical and consistent”. Since no dataset for coherence exists for Indian languages, a coherence classifier was trained using the next sentence prediction task. Sentence pairs were extracted from featured Wikipedia articles, with continuous sentence pairs chosen as positive samples and randomly permuted sentence pairs as negative samples. A coverage classifier was trained using manually annotated data. The coherence classifier achieved an overall F1 of 0.71

and the coverage classifier achieved an F1 of 0.896.

Methodology

Fact Organisation

The grouping of facts and the order in which they appear in the text is used as a feature for text generation. Two approaches are explored for identifying the fact clusters - statistical and end-to-end with mT5.

Statistical: Here, spectral clustering is performed based on the affinity matrices of pairs of facts occurring in the same sentence. The number of clusters is predicted by a classifier trained on the ground truth facts and the number of sentences.

End-to-end: Here, the problem of identifying fact clusters and ordering them is treated as a text-to-text problem, and an mT5 model is used to identify and order the clusters in an end-to-end manner.

Coverage prompts

The complete factual information of a sentence may not be represented by the aligned facts. Coverage prompts are used to provide the model with additional information regarding the quality of a training sample. During training, each example is provided to the model with a prompt of low, medium, or high obtained using the coverage classifier. During inference, only the high prompt is provided to generate texts that align well with the provided input.

Grounded Decoding

To reduce hallucination, we use a decoding strategy that attends to the source, similar to (Tian et al., 2020). The coverage score of each candidate token is added to its probability score to obtain a final score. To improve the quality of generation and increase vocabulary overlap, selective script unification is used. Specifically, the Indo-Aryan languages are all expressed in the Devanagari script, while the Dravidian languages are expressed in the Malayalam script.

Reinforcement Learning

Similar to the approach used in (Lai et al., 2021), RL-based rewards are used to improve the quality of generation. Specifically, we propose two reward scores - coverage with respect to the source facts and coverage with respect to the target sentence.

Cross-lingual PARENT score

Evaluation metrics for text generation like BLEU and ROUGE rely on the reference text. This is problematic when the reference and the source do not align entirely. PARENT (Dhingra et al., 2019) addresses this by aligning the n-grams from generated and reference text to the semi-structured input before scores. We use a modified version of PARENT - XPARENT, that can be used for cross-lingual settings. Instead of exact string matching which can only be used for mono-lingual settings, cosine similarity based matching of token embeddings is used. This shows a greater correlation with human evaluation and more explainability in terms of precision and recall.

Results

The performance of the proposed methods and existing baselines is summarized in table **??**. It can be observed that all components of the pipeline contribute towards better generation and an mT5 model trained with RL rewards, leveraging fact clustering, coverage prompts and grounded decoding attains the best performance for the weighted average scores. Note that the numbers reported are for mT5 clustering, since it resulted in the best performance in our experiments.

Conclusions

In this work we explored the problem for XF2T for generation of multi-sentence paragraphs. We create a dataset with a high quality test partition. We explore different methods such as explicit clustering of facts, coverage prompting, grounded decoding which both improve the quality of generation and address the problem of hallucination. These approaches can be used to directly generate Wikipedia like long text from structured data. We also define XPARENT score for evaluation of cross-lingual data-to-text problem which is of particular relevance for divergent references .

References

[Abhishek et al.2022] Tushar Abhishek, Shivprasad Sagare, Bhavyajeet Singh, Anubhav Sharma, Manish Gupta, and Vasudeva Varma. 2022. Xalign: Cross-lingual fact-to-text alignment and generation for low-resource languages. In *Companion Proceedings of the Web Conference 2022*, WWW '22, page 171–175, New York, NY, USA. Association for Computing Machinery.

[Dhingra et al.2019] Bhuwan Dhingra, Manaal Faruqui, Ankur Parikh, Ming-Wei Chang, Dipanjan Das, and William Cohen. 2019. Handling divergent reference texts when evaluating table-to-text generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4884–4895, Florence, Italy, July. Association for Computational Linguistics.

[Lai et al.2021] Huiyuan Lai, Antonio Toral, and Malvina Nissim. 2021. Thank you BART! rewarding pre-trained models improves formality style transfer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 484–494, Online, August. Association for Computational Linguistics.

[Tian et al.2020] Ran Tian, Shashi Narayan, Thibault Sellam, and Ankur P. Parikh. 2020. Sticking to the facts: Confident decoding for faithful data-to-text generation.

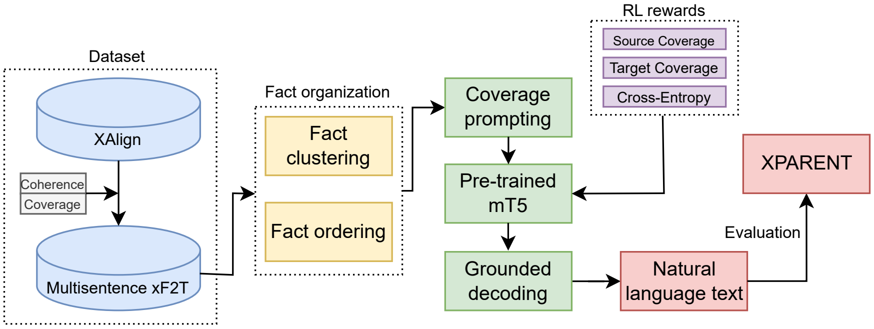

Figure 1: Proposed pipeline for cross-lingual, multi-sentence fact-to-generation.

	Singlesent		Multisent		mT5-C		mt5-C+CP+GD		mt5-C+CP+GD+RL	
Language	BLEU	xPARENT	BLEU	xPARENT	BLEU	xPARENT	BLEU	xPARENT	BLEU	xPARENT
Assamese	5.09	26.79	5.85	27.86	6.58	35.00	7.26	38.95	7.93	40.50
Bengali	16.46	43.50	21.95	53.75	23.10	59.93	24.69	62.08	25.34	63.05
English	22.21	56.55	29.01	66.70	26.89	63.97	28.92	66.63	30.65	68.17
Gujarati	6.62	29.20	7.04	33.25	9.12	36.58	11.16	41.14	12.96	43.44
Hindi	14.54	43.32	16.86	52.64	23.73	55.48	24.57	57.76	25.85	58.85
Kannada	4.28	21.89	3.62	25.72	5.94	30.81	7.37	36.03	7.56	37.92
Malayalam	6.55	24.74	5.94	30.40	9.42	34.14	10.55	37.02	10.50	37.29
Marathi	22.53	40.66	17.46	50.20	26.04	50.17	28.96	54.34	30.82	56.58
Odia	17.63	42.94	21.27	38.13	23.59	47.98	25.30	49.74	26.69	50.99
Punjabi	10.94	37.21	11.13	44.84	14.21	51.50	15.30	52.57	15.80	53.05
Tamil	6.64	22.95	6.88	28.07	10.38	34.46	12.03	36.53	9.60	34.89
Telugu	3.86	24.25	2.74	25.39	6.74	32.89	8.01	36.52	8.44	38.24
Average	15.51	42.20	18.66	50.34	20.39	52.68	22.06	55.27	22.82	56.22

Table 2: The table shows the BLEU and PARENT scores for all the approaches. Singlesent - generation using mT5 trained on single-sentence XF2T; multisent - generating using mT5 trained on end-to-end multi-sentence XF2T (our dataset); mT5-C - multisent with mT5 based clustering and ordering; mT5-C+CP +GD- previous with coverage prompts and grounded decoding; and mt5-C+CP+GD+RL is previous with RL rewards. While the BLEU score for end-to-end multisentence approach is better than the single sentence approach, it performs worse on XPARENT, indicating a greater degree of hallucination.

Facts	**Reference**	**Generated**
⟨H⟩dilipkumar mansukhlal gandhi ⟨R⟩date_of_birth ⟨T⟩09 may 1951 ⟨R⟩member_of_political_party ⟨T⟩bharatiya janata party ⟨R⟩occupation ⟨T⟩politician ⟨R⟩position_held ⟨T⟩member of the 13th lok sabha ⟨QR⟩electoral_district ⟨QT⟩ahmednagar lok sabha constituency ⟨R⟩position_held ⟨T⟩member of the 15th lok sabha ⟨QR⟩electoral_district ⟨QT⟩ahmednagar lok sabha constituency ⟨QR⟩start_time ⟨QT⟩2009	Dilipkumar Mansukhlal Gandhi (9 May 1951 – 17 March 2021) was an Indian politician and a member of the Bharatiya Janata Party (BJP) political party . In 1999 he was elected to the 13th Lok Sabha from Ahmednagar constituency in Maharashtra state . In 2009 , he was elected to the 15th Lok Sabha from the same constituency .	Dilipkumar Mansukhlal Gandhi (born 9 May 1951) is an Indian politician from the Bharatiya Janata Party (BJP) .He was elected to the 13th Lok Sabha from Ahmednagar (Lok Sabha constituency). He was elected to the 15th Lok Sabha from Ahmednagar constituency in 2009 .
⟨H⟩vasudev nirmal ⟨R⟩languages ⟨T⟩sindhi ⟨R⟩notable_work ⟨T⟩vijoon vasan aayoon ⟨R⟩award_received ⟨T⟩sahitya akademi award ⟨QR⟩point_in_time ⟨QT⟩2007 ⟨QR⟩for_work ⟨QT⟩vijoon vasan aayoon	वासुदेव निर्मल सिन्धी भाषा के विख्यात साहित्यकार हैं । इनके द्वारा रचित एक एकांकी . संग्रह विजूण् वसण आयूं के लिये उन्हें सन् 2007 में साहित्य अकादमी पुरस्कार से सम्मानित किया गया ।	वासुदेव निर्मल सिन्धी भाषा के विख्यात साहित्यकार हैं । इनके द्वारा रचित एक कविता . संग्रह विजुन वासन अयों के लिये उन्हें सन् 2007 में साहित्य अकादमी पुरस्कार से सम्मानित किया गया ।

Table 3: Some examples of generation using the best performing model in English and Hindi