OutlineGen: Multi-lingual Outline Generation for Encyclopedic Text in Low Resource Languages

Shivansh Subramanian[1[0000−0001−5647−1790]], Dhaval Taunk[1[0000−0001−7144−4520]], Manish Gupta[1,2[0000−0002−2843−3110]], Vasudeva Varma[1[0000−0003−1923−1725]]

[1] IIIT Hyderabad, India [2] Microsoft, India
{shivansh.s, dhaval.taunk}@research.iiit.ac.in
{manish.gupta, vv}@iiit.ac.in

Abstract. Lack of encyclopedic text contributors, especially on Wikipedia, makes automated text generation for low resource (LR) languages a critical problem. A step towards enabling this is to generate the structural outline of Wikipedia pages for LR languages. Hence, in this work, we propose and study OutlineGen, the problem of generating the outline of Wikipedia pages for LR languages using minimal information in the form of entity name, language and domain. The OutlineGen task is challenging because even within a (language, domain) pair, the outlines vary a lot across entities. Further, given the diversity of Wikipedia editors and audience, the outlines are not consistent across languages. First, we create a dataset, WikiOutlines, which contains Wikipedia section outlines from ∼166K Wikipedia pages across 8 domains and 10 languages. Then, we investigate the effectiveness of non-neural weighted finite state automata as well as Transformer-based methods for this task. We make the code and data publicly available.

Keywords: OutlineGen · WikiOutlines · deep learning · multi-lingual generation, Wikipedia text generation · low resource natural language generation

1 Introduction

Wikipedia has been the most popular source of factual and neutral encyclopedic information for millions of users. Although English Wikipedia is rich with ∼7M articles, number of Wikipedia pages in nine low resource (LR) languages which we consider in this work add to ∼100K. Unfortunately, recent efforts towards enriching LR Wikipedia over the years have also not been as encouraging as for English [25]. Hence, automated text generation for low-resource Wikipedia is critical.

Recently, some studies [25, 26, 8] have focused on generating text corresponding to a specific section for LR Wikipedia. However, to generate an entire Wikipedia page, it is important to first generate the structural outline and then fill the sections with LR language text using these existing methods. In Table 1,

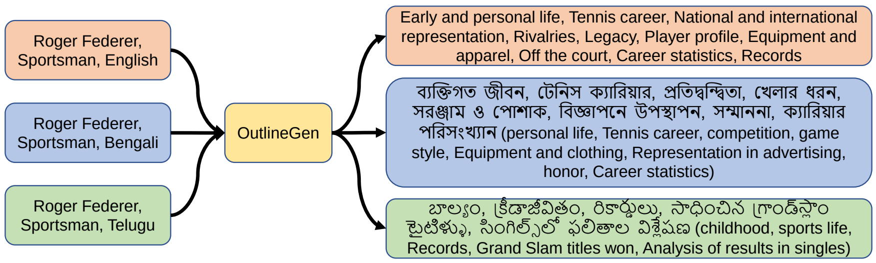

Fig. 1. OutlineGen examples: Generating outlines for the "Roger Federer" entity (which belongs to the sportsmen domain) for English, Bengali and Telugu Wikipedia pages.

we show percentage of articles with outline same as most frequently occurring article outline per (language, domain) for 8 domains and 10 languages. We observe that, on average, only 26.6% of all articles follow the templatized outline. Thus, we need to design a method which generates the Wikipedia section outline by conditioning on (entity, language, domain) triple. Translating the outline from corresponding English Wikipedia page is not effective because (1) several LR pages on Wikipedia do not have equivalent pages on English Wikipedia, and (2) LR Wikipedia pages are written by LR language editors for LR language users and hence their outlines differ significantly from outlines for corresponding Wikipedia pages (if they exist).

Hence, in this paper, we propose the task of Outline Generation, OutlineGen, for Wikipedia articles, which is a novel task to generate Wikipedia-styled outlines given an article's (entity, language, domain) triple. Since our goal is to generate Wikipedia outlines for entities where no Wikipedia page already exists, we take minimal inputs (entity, language, domain) for the task. Fig. 1 shows examples of OutlineGen task for the "Roger Federer" entity (which belongs to the sportsmen domain) for English, Bengali and Telugu Wikipedia pages. These

Table 1. Percentage of articles with outline same as the most-frequent outline of (language, domain) pair

	hi	mr	bn	or	ta	en	ml	pa	kn	te	Avg
politicians	53.44	63.70	36.55	35.25	32.40	9.36	20.50	27.94	17.74	13.00	30.99
cities	17.46	49.82	15.46	22.76	19.98	18.00	37.00	56.08	24.22	10.69	27.15
books	74.09	40.54	17.76	31.03	12.58	10.63	29.73	32.69	30.48	7.91	28.74
writers	33.91	36.77	15.67	15.25	11.91	7.26	14.96	16.66	11.70	8.29	17.24
companies	27.96	50.82	26.51	36.84	25.00	28.75	26.22	45.65	15.57	20.56	30.39
sportsmen	44.23	69.18	22.63	22.79	39.33	14.69	22.73	39.15	16.21	10.47	30.14
films	17.38	20.81	19.95	38.37	17.13	21.80	17.11	27.99	40.90	30.27	25.17
animals	34.96	22.53	18.87	18.31	28.47	12.41	38.41	29.00	16.83	12.68	23.25
Avg	37.9	44.3	21.7	27.6	23.3	15.4	25.8	34.4	21.7	14.2	26.6

Table 2. #samples per domain per language in WikiOutlines

	hi	mr	bn	or	ta	en	ml	pa	kn	te	Total
politicians	6,617	3,815	8,071	1,336	5,885	566	3,405	1,589	699	1,808	33,791
cities	1,048	827	854	268	851	3,550	554	526	256	290	9,024
books	1,428	148	805	87	1,988	762	740	468	105	215	6,746
writers	3,474	1,882	3,605	564	3,005	758	3,475	3,320	1,128	1,339	22,550
companies	683	366	679	38	644	4,546	431	138	212	180	7,917
sportsmen	9,476	11,556	13,154	408	9,808	177	2,583	2,327	660	640	50,789
films	4,959	1,033	3,655	920	6,504	1,165	3,934	618	1,704	3,621	28,113
animals	472	395	1,261	142	1,556	427	2,317	200	315	205	7,290
Total	28,157	20,022	32,084	3,763	30,241	11,951	17,439	9,186	5,079	8,298	166,220

outlines could help human editors to plan the article content better. These outlines could also help improve the quality of the automatically generated text (using methods like [23, 14, 2, 25]) and hence reduce human post-editing efforts.

Lack of availability of a relevant dataset makes the OutlineGen task further challenging. Hence, as part of this work, we contribute a novel dataset, WikiOutlines, which contains Wikipedia section outlines from ∼166K Wikipedia pages across 8 domains and 10 languages. The domains include politicians, cities, books, writers, companies, sportsmen, films and animals. Languages include Hindi (hi), Marathi (mr), Bengali (bn), Oriya (or), Tamil (ta), English (en), Malayalam (ml), Punjabi (pa), Kannada (kn) and Telugu (te). The dataset can serve as a great benchmark for the OutlineGen task.

Overall, we make the following contributions in this paper: (1) We define and motivate the need for the novel OutlineGen task, where the input is minimal (entity, target language, and domain). The output is the wikipedia-style outline. (2) We contribute a novel dataset, WikiOutlines, with outlines for ∼166K articles across 8 domains and 10 languages. (3) We model OutlineGen using Weighted Finite State Automata (WFSA) and multi-lingual Transformer encoder-decoder models. (4) Our experiments show that mT5 [28] outperforms mBART [15] as well as a WFSA-based solution for the outline generation task.

2 Related Work

Wikipedia Text Generation: Initial efforts in the fact-to-text (F2T) line of work focused on generating short text, typically the first sentence of Wikipedia pages using structured fact tuples. Seq-2-seq neural methods [12, 17], such as LSTM-based [27, 24, 19] and pretrained Transformer-based [21] like BART [13] and T5 [20], have been widely adopted for F2T tasks. Most of the previous efforts on F2T focused on English F2T only. Recently, the Cross-lingual F2T (XF2T) problem was proposed in [1] and [22]. Besides generating short Wikipedia text, there have also been efforts to generate Wikipedia articles by summarizing long sequences [14, 7, 9, 8, 26]. For all of these datasets, the generated text is either the full Wikipedia article or text for a specific section [9]. Most of these studies [14, 9, 7] have been done on English only. Compared to all of these pieces of work which have focused on Wikipedia article text generation, the focus of the current paper is on generating outlines.

Wikipedia Outline Generation: Outline Generation for documents was first proposed in [29], where the goal was to identify the boundaries for sections and to generate the required section heading for the specified section. They proposed a model based on LSTMs [10] that performs both tasks simultaneously. Recently, Maheshwari et al. [16] proposed the related task of creating a Table of Contents (or Outline) for long documents like contracts, financial documents etc. given the section distinctions. Both of these efforts focused on English only. Outline generation for multiple low resource languages brings a different set of challenges which we tackle in this paper.

Multi-Lingual Generation: Popular multi-lingual and cross-lingual natural language generation (NLG) tasks include machine translation [5], question generation [6, 18], blog title generation [3], summarization [30, 11], and style transfer [4]. In this work, we propose the novel multilingual Wikipedia outline generation task and propose strong baseline solutions for the task.

3 WikiOutlines: Data Collection, Pre-processing and Analysis

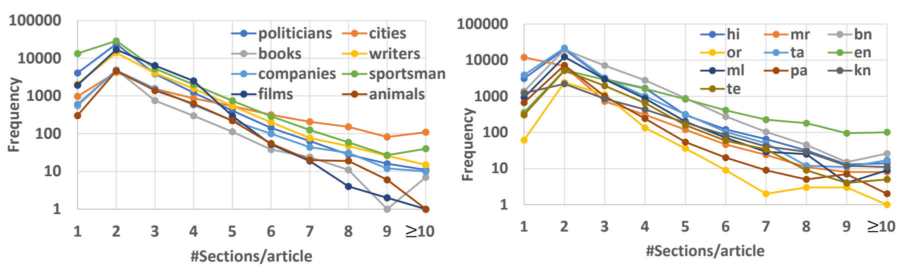

Fig. 2. Distribution of number of sections across various domains and languages in the WikiOutlines dataset

Data Collection and Pre-processing: WikiOutlines contains Wikipedia sections related to eight distinct domains spanning across ten languages. To begin, we utilize the Wikidata API[1] to initially narrow down the domains of interest. Then, we retrieve entities that have corresponding Wikipedia pages in our chosen languages. Afterward, we use Wikipedia's language-specific 20221201 XML dumps, to acquire the Wikipedia pages of the filtered entities. The text in Wikipedia pages follows a standardized structure, with sections and subsections. We extract these sections and subsections from the text. Overall, each sample in the dataset consists of the domain, language, and section title. This dataset

[1] https://query.wikidata.org/

Fig. 3. Word clouds of most frequent Wikipedia section titles per domain. Each word cloud contains titles across all languages. Section titles for one language are shown using a single color. Font size indicates relative frequency.

is then split into train, validation, and test in the 80:10:10 ratio, stratified by domain and language. We make these standard splits publicly available as part of the dataset[2].

Data Analysis: Table 2 shows the overall distribution of the ∼166K samples in the dataset across all domains and languages. The number of samples differ significantly across various (domain, language) pairs. sportsmen and politicians have largest number of samples while books and animals have lowest number of samples. Oriya and Kannada have lowest number of samples from language perspective while Bengali and Tamil are the richest languages. Fig. 2 shows the distribution of number of sections across various domains and languages in the WikiOutlines dataset. Notice that the y-axis is drawn in log scale. The figures show that for every language and every domain, most samples have 2 sections, except for Marathi where most samples have just 1 section. Amongst the languages, the distribution is flatest for English, where the number of samples with ≥10 sections is the highest. Amongst the domains, cities has a similar behavior.

Finally, we show word clouds of the most frequent Wikipedia section titles for each of the eight domains in Fig. 3. Each word cloud contains the five most frequent titles per language. Section titles for one language are shown using a single color. Font size indicates relative frequency. The word clouds show the variety of section titles per (language, domain) pair.

[2] https://github.com/AurumnPegasus/OutlineGen

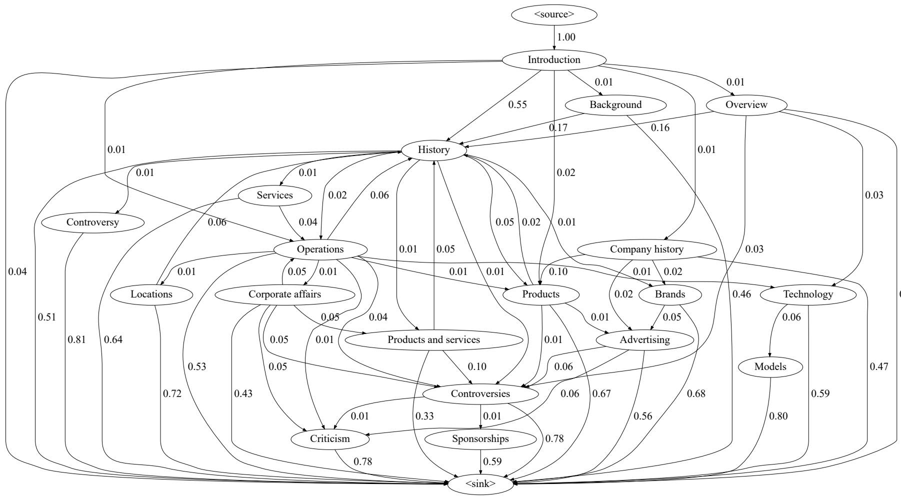

Fig. 4. Examples of generated weighted finite state automata for (en, companies) where section-titles are the nodes, and transition probability is written on the edges.

4 Approaches for OutlineGen Task

Weighted Finite State Automata (WFSA): Table 1 shows that many articles share the same outline. These article outlines are often specific for a language over a particular domain, and the section transition patterns can potentially be found via simple statistical models instead of large generative ones. Hence, instead of defining static outlines based simply on frequency, we learn a weighted finite state automata for all articles belonging to a (language, domain) pair. The source node for the WFSA is ⟨source⟩, and the sink node is represented by ⟨sink⟩. The nodes between the source and the sink contain the section titles, and the transition probability from node A to node B is the conditional probability of section title B following section title A in an article outline. Fig. 4 shows an example of WFSA learned for (en, companies) pair. These are drawn using top 20 most frequent section titles as nodes. Also, only the edges with weight more than 0.005 are shown. WFSA involves two hyper-parameters: (i) a *beam-size* (samples from top-k instead of choosing the most probable next state), and (ii) *token-level* (word or section-title level WFSA).

The WFSA is used for inference as follows. We start from ⟨source⟩ state and select *beam-size* number of next possible states. We base our selection either greedily (selecting the most probable next states) or by sampling them from a probability distribution over the next states. We repeat these instructions recursively (in a breadth-first manner), maintaining the visited part of the outline in a queue and the total accumulated transition probability. Once we reach the ⟨sink⟩ state, we terminate the recursion and store the generated outline with the geometric mean of transition probabilities signifying the probability of that

Table 3. Comparison of WFSA and Transformer-based methods for multi-lingual outline generation.

	XLM-Score	BLEU	METEOR	ROUGE-L
WFSA (section-level)	70.0	45.0	37.1	56.8
WFSA (word-level)	69.1	43.4	36.1	55.9
mBART	70.2	39.1	31.9	52.2
mT5	**76.2**	**48.5**	**40.3**	**59.4**

outline occurring. We terminate when the breadth-first search queue is empty. Outline with the highest probability is selected as the output. Of course, we ensure that the generated outline does not have repeated section titles.

Multi-lingual Transformer encoder-decoder Generative Models: WFSA based methods are not entity-specific. This restricts them to predict the same outline for all entities belonging to the same (language, domain) pair. Hence, we also experiment with popular multi-lingual Transformer encoder-decoder generative models like mT5 [28] and mBART [15]. The language and domain are passed as input with a separator token. The models are fine-tuned to generate outlines. The model is now supposed to automatically decide the number of sections in the outline and the actual section titles in the outline as well.

5 Experiments and Results

We trained both WFSA as well as Transformer-based models using training data and tuned hyper-parameters on the validation set. For WFSA, we found beam size=4 to provide best result on validation set. For mT5 and mBART, we trained using AdamW optimizer for 10 epochs on a machine with 4 NVIDIA V100 GPUs. We used a batch size of 8, a learning rate of 2e-5, and a beam size of 3.

Table 3 shows the comparison of WFSA and Transformer-based methods for multi-lingual outline generation using popular natural language generation metrics like XLM-Score, BLEU, METEOR and ROUGE-L. We observe that (1) mT5 outperforms other methods by large margins across all metrics. (2) WFSA at section-title level leads to better results compared to WFSA at word level.

We show detailed results for our best model (mT5) at a (language, domain) level using the four metrics (XLM-Score, BLEU, METEOR, and ROUGE-L) in Table 4. From these tables, we observe that (1) The model performs best for films and sportsmen domains, and worst for writers and animals domains. This is largely justified because of the large number of training samples in films and sportsmen domains and low number of samples in animals domain. However, it is surprising that the model does not perform well on writers domain inspite of the large number of training samples. (2) The model performs best for Punjabi, and worst for Telugu and Kannada. The worse performance for Telugu and Kannada can perhaps be because of low number of training samples for those languages. Examples of generated outlines using our best method in Table 5 show that our method can generate reasonably usable outlines.

Table 4. XLM-Score, BLEU, ROUGE-L and METEOR scores for mT5 across various (language, domain) pairs.

	en	mr	hi	kn	ta	bn	pa	te	ml	or	Avg
companies	80.4	80.7	64.8	69.1	82.2	73.4	92.2	72.6	69.8	80.4	76.6
writers	73.4	70.7	71.3	63.3	75.2	66.1	86.9	68.4	72.2	84.0	73.2
cities	77.0	72.3	62.9	66.0	78.6	66.9	92.0	63.5	76.1	92.0	74.7
politicians	71.6	81.8	79.2	65.4	81.0	75.2	87.5	69.2	70.5	88.4	77.0
books	71.6	73.3	87.9	70.1	78.9	69.4	89.9	63.4	72.8	83.3	76.1
films	80.7	76.1	72.7	81.1	80.3	71.3	91.1	76.5	72.5	91.8	79.4
animals	71.9	62.0	67.0	69.0	80.9	68.9	88.9	68.2	78.0	81.4	73.6
sportsmen	81.1	88.2	83.0	66.7	86.1	74.2	90.8	64.2	75.0	83.5	79.3
Avg	75.9	75.6	73.6	68.9	80.4	70.7	89.9	68.2	73.4	85.6	76.2

	en	mr	hi	kn	ta	bn	pa	te	ml	or	Avg
companies	52.8	66.7	43.8	44.1	57.9	57.0	67.2	45.6	47.5	37.6	52.0
writers	31.5	46.2	54.3	29.6	37.1	40.6	45.5	33.0	50.2	41.1	40.9
cities	39.4	48.0	28.9	33.3	48.0	35.0	67.7	26.8	60.5	62.4	45.0
politicians	35.3	68.3	67.5	31.9	53.0	58.2	50.1	37.6	46.6	51.9	50.0
books	38.3	54.2	81.5	46.4	48.3	44.4	58.0	38.2	53.9	49.2	51.2
films	52.6	57.8	40.2	64.9	52.0	49.2	62.0	56.8	52.6	55.5	54.4
animals	32.0	37.4	37.3	35.6	54.9	38.6	50.9	44.7	59.3	31.9	42.3
sportsmen	41.9	79.2	72.0	31.7	62.4	45.5	62.9	32.8	53.2	44.4	52.6
Avg	40.5	57.2	53.2	39.7	51.7	46.0	58.1	39.4	53.0	46.7	48.5

	en	mr	hi	kn	ta	bn	pa	te	ml	or	Avg
companies	64.9	80.5	54.3	58.1	64.4	62.6	69.3	57.0	54.6	45.2	61.1
writers	44.1	62.5	64.0	44.8	46.8	51.0	52.6	44.9	56.8	50.0	51.7
cities	59.9	68.5	43.7	48.6	55.9	46.7	73.2	44.9	66.7	77.3	58.5
politicians	44.3	80.9	74.3	52.1	59.8	66.8	57.6	52.0	54.4	64.5	60.7
books	51.1	72.8	84.9	64.3	55.7	51.2	64.4	46.8	58.5	56.1	60.6
films	67.1	66.7	59.0	75.8	57.7	59.1	65.4	63.9	59.6	70.7	64.5
animals	49.8	55.8	47.5	46.7	59.8	50.0	58.1	51.6	65.2	44.0	52.9
sportsmen	73.5	86.7	77.5	54.4	71.9	59.2	68.9	45.7	60.1	53.5	65.1
Avg	56.8	71.8	63.2	55.6	59.0	55.8	63.7	50.9	59.5	57.7	59.4

	en	mr	hi	kn	ta	bn	pa	te	ml	or	Avg
companies	49.3	38.6	35.0	39.9	47.9	45.2	52.8	39.5	32.9	24.9	40.6
writers	28.5	33.5	45.1	26.0	29.5	31.7	32.5	30.6	37.6	29.9	32.5
cities	46.4	30.6	34.5	23.4	37.2	31.1	58.1	25.7	49.8	70.9	40.8
politicians	26.6	39.3	59.0	26.7	42.5	52.7	39.6	36.7	34.3	53.4	41.1
books	40.1	31.7	74.1	28.2	37.0	31.4	46.7	24.1	39.8	37.9	39.1
films	52.3	53.0	54.1	42.8	36.5	36.2	47.2	45.9	37.2	68.4	47.4
animals	43.9	27.1	39.5	25.7	41.7	29.6	41.3	32.0	48.1	22.6	35.1
sportsmen	63.8	46.6	62.9	31.5	58.6	42.8	54.0	25.7	39.6	34.3	46.0
Avg	43.9	37.6	50.5	30.5	41.4	37.6	46.5	32.5	39.9	42.8	40.3

6 Conclusion

In this paper, we proposed OutlineGen, which is the problem of multi-lingual outline generation. We contributed a novel dataset, WikiOutlines, which covers 8 domains and 10 languages. We also investigated effectiveness of two kinds of models: WFSA and Transformer-based encoder-decoders like mT5 and mBART. We found that mT5 provided us good XLM-Score, BLEU, METEOR and ROUGE-L of 76.2, 48.5, 40.3 and 59.4 respectively. This shows that the system is practi-

Table 5. Generated outlines using our best method

Entity Name	Language	Domain	Ground-truth outline	Generated outline
Takin	en	animals	Introduction, Appearance, Distribution and habitat, Behaviour and ecology	Introduction, Taxonomy, Distribution and habitat, Behaviour and ecology
Dresden	en	cities	Introduction, History, Geography, Governance, Culture, Economy, Education and science, Notable people	Introduction, History, Geography, Demographics, Culture, Notable people
इन्साफ का मन्दिर (1969 फ़िल्म)	hi	films	परिचय, मुख्य कलाकार, परिणाम, नामांकन और पुरस्कार	परिचय, मुख्य कलाकार, परिणाम, नामांकन और पुरस्कार
ତତେ ଭଲ ପାଉଛି ବୋଲି	or	films	ପରିଚୟ, ଅଭିନୟ, ଗୀତ ଏବଂ ସଙ୍ଗୀତ	ପରିଚୟ, ଅଭିନୟ, ଗୀତ ଏବଂ ସଙ୍ଗୀତ
জনার্দন নবলে	bn	sportsman	ভূমিকা, আন্তর্জাতিক ক্রিকেট, মূল্যায়ন	ভূমিকা, প্রথম-শ্রেণীর ক্রিকেট, আন্তর্জাতিক ক্রিকেট

cally usable to generate candidate outlines which can be refined further by human editors to bootstrap generation of Wikipedia pages in low-resource languages.

References

1. Abhishek, T., Sagare, S., Singh, B., Sharma, A., Gupta, M., Varma, V.: Xalign: Cross-lingual fact-to-text alignment and generation for low-resource languages. In: WebConf. pp. 171–175 (2022)
2. Banerjee, S., Mitra, P.: Wikiwrite: Generating wikipedia articles automatically. In: IJCAI. pp. 2740–2746 (2016)
3. Bhatt, S.M., Agarwal, S., Gurjar, O., Gupta, M., Shrivastava, M.: Tourismnlg: A multi-lingual generative benchmark for the tourism domain. In: ECIR. p. To appear (2023)
4. Briakou, E., Lu, D., Zhang, K., Tetreault, J.: Olá, bonjour, salve! xformal: A benchmark for multilingual formality style transfer. In: NAACL-HLT. pp. 3199–3216 (2021)
5. Chi, Z., Dong, L., Ma, S., Huang, S., Mao, X.L., Huang, H., Wei, F.: Mt6: Multilingual pretrained text-to-text transformer with translation pairs (2021)
6. Chi, Z., Dong, L., Wei, F., Wang, W., Mao, X.L., Huang, H.: Cross-lingual nlg via pre-training. In: AAAI. vol. 34, pp. 7570–7577 (2020)
7. Ghalandari, D.G., Hokamp, C., Glover, J., Ifrim, G., et al.: A large-scale multi-document summarization dataset from the wikipedia current events portal. In: ACL. pp. 1302–1308 (2020)
8. Giannakopoulos, G., Kubina, J., Conroy, J., Steinberger, J., Favre, B., Kabadjov, M., Kruschwitz, U., Poesio, M.: Multiling 2015: multilingual summarization of single and multi-documents, on-line fora, and call-center conversations. In: SIG on Discourse and Dialogue. pp. 270–274 (2015)
9. Hayashi, H., Budania, P., Wang, P., Ackerson, C., Neervannan, R., Neubig, G.: Wikiasp: A dataset for multi-domain aspect-based summarization. TACL **9**, 211–225 (2021)
10. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
11. Jhaveri, N., Gupta, M., Varma, V.: clstk: The cross-lingual summarization tool-kit. In: WSDM. pp. 766–769 (2019)
12. Lebret, R., Grangier, D., Auli, M.: Neural text generation from structured data with application to the biography domain. In: EMNLP. pp. 1203–1213 (2016)

13. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L.: Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: ACL. pp. 7871–7880 (2020)
14. Liu, P.J., Saleh, M., Pot, E., Goodrich, B., Sepassi, R., Kaiser, L., Shazeer, N.: Generating wikipedia by summarizing long sequences. In: ICLR (2018)
15. Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., Zettlemoyer, L.: Multilingual denoising pre-training for neural machine translation. TACL **8**, 726–742 (2020)
16. Maheshwari, H., Sivakumar, N., Jain, S., Karandikar, T., Aggarwal, V., Goyal, N., Shekhar, S.: Dynamictoc: Persona-based table of contents for consumption of long documents. In: NAACL-HLT. pp. 5133–5143 (2022)
17. Mei, H., Bansal, M., Walter, M.R.: What to talk about and how? selective gen. using lstms with coarse-to-fine alignment. In: NAACL-HLT. pp. 720–730 (2016)
18. Mitra, R., Jain, R., Veerubhotla, A.S., Gupta, M.: Zero-shot multi-lingual interrogative question generation for" people also ask" at bing. In: KDD. pp. 3414–3422 (2021)
19. Nema, P., Shetty, S., Jain, P., Laha, A., Sankaranarayanan, K., Khapra, M.M.: Generating descriptions from structured data using a bifocal attention mechanism and gated orthogonalization. In: NAACL-HLT. pp. 1539–1550 (2018)
20. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J., et al.: Exploring the limits of transfer learning with a unified text-to-text transformer. JMLR **21**(140), 1–67 (2020)
21. Ribeiro, L.F.R., Schmitt, M., Schütze, H., Gurevych, I.: Investigating pretrained language models for graph-to-text generation (2021)
22. Sagare, S., Abhishek, T., Singh, B., Sharma, A., Gupta, M., Varma, V.: Xf2t: Cross-lingual fact-to-text generation for low-resource languages. In: INLG (2023)
23. Sauper, C., Barzilay, R.: Automatically generating wikipedia articles: A structure-aware approach. In: ACL-IJCNLP. pp. 208–216 (2009)
24. Shahidi, H., Li, M., Lin, J.: Two birds, one stone: A simple, unified model for text generation from structured and unstructured data. In: ACL. pp. 3864–3870 (2020)
25. Taunk, D., Sagare, S., Patil, A., Subramanian, S., Gupta, M., Varma, V.: Xwikigen: Cross-lingual summarization for encyclopedic text generation in low resource languages. In: WebConf. pp. 1703–1713 (2023)
26. Tikhonov, P., Malykh, V.: Wikimulti: a corpus for cross-lingual summarization. arXiv:2204.11104 (2022)
27. Vougiouklis, P., Elsahar, H., Kaffee, L.A., Gravier, C., Laforest, F., Hare, J., Simperl, E.: Neural wikipedian: Generating textual summaries from knowledge base triples. J. Web Semantics **52**, 1–15 (2018)
28. Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., Raffel, C.: mt5: A massively multilingual pre-trained text-to-text transformer. In: NAACL-HLT. pp. 483–498 (2021)
29. Zhang, R., Guo, J., Fan, Y., Lan, Y., Cheng, X.: Outline generation: Understanding the inherent content structure of documents. In: SIGIR. pp. 745–754 (2019)
30. Zhu, J., Wang, Q., Wang, Y., Zhou, Y., Zhang, J., Wang, S., Zong, C.: Ncls: Neural cross-lingual summarization. In: EMNLP-IJCNLP. pp. 3054–3064 (2019)