

XAlign: Cross-lingual Fact-to-Text Alignment and Generation for Low-Resource Languages

Tushar Abhishek
IIIT Hyderabad
India
tushar.abhishek@research.iiit.ac.in

Shivprasad Sagare
IIIT Hyderabad
India
shivprasad.sagare@research.iiit.ac.in

Bhavyajeet Singh
IIIT Hyderabad
India
bhavyajeet.singh@research.iiit.ac.in

Anubhav Sharma
IIIT Hyderabad
India
anubhav.sharma@research.iiit.ac.in

Manish Gupta
IIIT Hyderabad
India
manish.gupta@iiit.ac.in

Vasudeva Varma
IIIT Hyderabad
India
vv@iiit.ac.in

ABSTRACT

Multiple critical scenarios need automated generation of descriptive text in low-resource (LR) languages given English fact triples. For example, Wikipedia text generation given English Infoboxes, automated generation of non-English product descriptions using English product attributes, etc. Previous work on fact-to-text (F2T) generation has focused on *English* only. Building an effective *cross-lingual F2T (XF2T)* system requires alignment between English structured facts and LR sentences. Either we need to manually obtain such alignment data at a large scale, which is expensive, or build automated models for cross-lingual alignment. To the best of our knowledge, there has been no previous attempt on automated cross-lingual alignment or generation for LR languages. We propose two unsupervised methods for cross-lingual alignment. We contribute XALIGN, an XF2T dataset with 0.45M pairs across 8 languages, of which 5402 pairs have been manually annotated. We also train strong baseline XF2T generation models on XALIGN. We make our code and dataset publicly available¹, and hope that this will help advance further research in this critical area.

CCS CONCEPTS

• Information systems; • Computing methodologies → Neural networks; Natural language generation;

KEYWORDS

XF2T, deep learning, fact-to-text

ACM Reference Format:

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*

¹ <https://github.com/tushar117/XAlign>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

WWW '22 Companion, April 25–29, 2022, Virtual Event, Lyon, France

© 2022 Association for Computing Machinery.

ACM ISBN 978-1-4503-9130-6/22/04...\$15.00

<https://doi.org/10.1145/3487553.3524265>

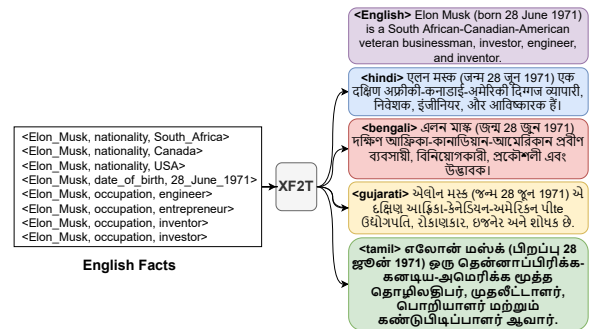


Figure 1: XF2T Example: Generating English, Hindi, Bengali, Gujarati or Tamil sentences to capture semantics from English facts.

Proceedings of the Web Conference 2022 (WWW '22 Companion), April 25–29, 2022, Virtual Event, Lyon, France. ACM, New York, NY, USA, 4 pages. <https://doi.org/10.1145/3487553.3524265>

1 INTRODUCTION

Fact-to-text (F2T) generation [23] is the task of transforming structured data (like fact triples) into natural language. F2T systems are vital in many downstream applications like automated dialog systems, question answering, etc. But most of such F2T systems are English only and not available for low-resource (LR) languages. This is mainly due to lack of training data in LR languages. For example, Wikidata entries for person entities in LR languages are minuscule in number compared to that in English (1.3M in English vs total of ~128K across our 7 LR languages). In addition, the average number of facts per entity in LR language Wikidata (11.3) is almost half compared to English (22.8). Thus, monolingual F2T for LR languages suffers from data sparsity. Hence, in this work, we propose a novel task of *cross-lingual F2T generation (XF2T)* which takes a set of English facts² as input and generates a sentence capturing the fact-semantics in the specified LR language.

One major challenge in training an XF2T system is the availability of aligned data where English facts are well-aligned with semantically equivalent LR text. The manual creation of such a high-quality aligned dataset requires human annotations and is quite challenging to scale. Recently various automatic alignment

² Fact is a triple composed of subject-predicate-object

approaches have been proposed for English F2T alignment like pairing up Wikipedia sentences with Infobox [17], using distant supervision [9], finding the lexical overlap between textual and structural entities [1, 11, 13], etc. Taking inspiration, we propose two novel methods for *cross-lingual* F2T, and use them to create and contribute a new dataset, XALIGN. It consists of sentences from LR language Wikipedia mapped to English fact triples from Wikidata. It contains data for the following languages: Hindi (hi), Telugu (te), Bengali (bn), Gujarati (gu), Marathi (mr), Kannada (kn), Tamil (ta) and English (en). Fig. 1 shows an XF2T example from our dataset.

Overall, we make the following contributions in this work. (1) We propose the problem of XF2T alignment and generation for LR languages. (2) We contribute a high quality XF2T dataset, XALIGN, with overall 0.45M pairs (5402 human-labeled) in 8 languages. (3) We propose transfer learning and distant supervision based methods for cross-lingual alignment. (4) We train multiple multi-lingual XF2T models which lead to a best BLEU score of 25.02. We make the code and dataset publicly available¹.

2 RELATED WORK

Recently, various approaches have been proposed for entity linking (link mention in a sentence to a KG entity) [4] and fact linking (link sentence to a set of facts) [15]. To the best of our knowledge, there has been no prior work on creating the XF2T dataset that requires pairing up a set of English facts with sentences in other languages. Table 1 shows basic statistics of popular F2T datasets. Some previous studies [12] on F2T collected aligned data by crowdsourcing while others have performed automatic alignment [1]. For XF2T dataset creation, we explore two different unsupervised methods to perform automated cross-lingual alignment. Unlike other datasets which are mostly on English, our dataset contains 8 languages and is a cross-lingual dataset. Most of the F2T methods can be classified as (1) template-based methods [8], (2) Seq-2-seq attention networks [17], (3) hierarchical attention networks [18], (4) pretrained Transformer methods [6, 11, 25]. Template-based methods fail for fact triples in a previously unseen domain. Also, all of these methods focus on English F2T only. We focus on XF2T.

3 DATA COLLECTION AND PRE-PROCESSING

We start by gathering a list of 85K person entities from Wikidata each of which have a link to a corresponding Wikipedia page in at least one of our 7 LR languages. This leads to a dataset D where every instance d_i is a tuple $\langle \text{entityID}, \text{English Wikidata facts}, \text{LR language}, \text{LR-language Wikipedia URL for the entityID} \rangle$.

For each language, we use Wikiextractor [2] to extract text from the Wikipedia xml dump dated 2021-05-20. We split this main text into sentences using Indic NLP [16], with a few additional heuristics to account for Indic punctuation characters, sentence delimiters and

Dataset	Languages	A/M	I	F	P	T	X-Lingual
WikiBio	en	A	728K	19.70	1740	26.1	No
E2E	en	M	50K	5.43	945	20.1	No
WebNLG'17	en	M	25K	2.95	373	22.7	No
fr-de Bio	fr, de	A	170K, 50K	8.60, 12.6	1331, 1267	29.5, 26.4	No
TREX	en	A	6.4M	1.77	642	79.8	No
WebNLG'20	en, ru	M	40K, 17K	2.68, 2.55	372, 226	23.7	Yes
KELM	en	A	8M	2.02	663	21.2	No
WITA	en	A	55K	3.00	640	18.8	No
WikiTableT	en	A	1.5M	51.90	3K	115.9	No
GenWiki	en	A	1.3M	1.95	290	21.5	No
XALIGN	en + 7 LR	A	0.45M	2.02	367	19.8	Yes

Table 1: Statistics of popular F2T datasets: WikiBio [17], E2E [19], WebNLG 2017 [12], WebNLG 2020 [10], fr-de Bio [18], KELM [1], WITA [11], WikiTableT [5], GenWiki [13], TREX [9], and XALIGN (ours). Alignment method could be A (automatic) or M (manual). |I|=# instances, |F|=avg fact count, |P|=# unique predicates, |T|=avg word count.

non-breaking prefixes. We prune out (1) other language sentences using Polyglot language detector³, (2) sentences with <5 or >100 words, (3) sentences which could potentially have no factual information (i.e., sentences with no noun or verb⁴). For each sentence per Wikipedia URL, we also store the section information.

We extract facts from the Wikidata dump dated 2020-12-21 for all the 8 languages for each entity in D using the Wikidata API⁵. We gather facts corresponding to Wikidata property types like WikibaseItem, Time, Quantity, Monolingualtext which capture most of the useful factual information for person-type entities. This leads to overall $\sim 0.91M$ facts for $\sim 85K$ entities.

4 F2T ALIGNMENT IN XALIGN

At this point, for every (entity e , language l) pair, the dataset D has a set F_{el} of English Wikidata facts and a set of Wikipedia sentences S_{el} in language l . Next, we build an automatic aligner to associate a sentence in S_{el} with a subset of facts from F_{el} . As shown in Fig. 2, we propose a two-stage system for the automatic alignment. In the first stage (Candidate Generation), we generate (facts, sentence) candidates based on syntactic+semantic match. In the second stage (Selection), we retain only strongly aligned candidates using transfer learning and distant supervision methods.

4.1 Stage 1: Candidate Generation

Given a set of English facts $\{f_i\}_{i=1}^{|f|}$ and set of sentences $\{s_j\}_{j=1}^{|s|}$ in language l , for every (f_i, s_j) pair, we compute a similarity score $sim(f_i, s_j)$ that captures syntactic+semantic similarity. For syntactic match, we use TFIDF by translating either the fact to language l or the sentence to English. For semantic match, we compute cosine similarity between MuRIL [14] representations of the fact and the sentence, or between their translations. Thus, $sim(f_i, s_j)$ is computed as average of these 4 scores: $sim\text{-MURIL}(f_i, s_j)$, $sim\text{-TFIDF}(\text{translate}(f_i, l), s_j)$, $sim\text{-TFIDF}(f_i, \text{translate}(s_j, \text{English}))$, $sim\text{-MuRIL}(\text{translate}(f_i, l), \text{translate}(s_j, \text{English}))$. For translating sentences, we use IndicTrans [22]. When translating the facts, we retain the label of entities within the fact triple for which Wikidata multi-lingual label is available for LR language, and translate only the remaining parts of

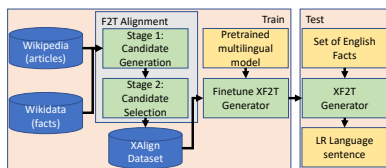


Figure 2: XALIGN: XF2T System Architecture

³ <https://polyglot.readthedocs.io/en/latest/Detection.html> ⁴ For POS tagging, we use Stanza [21] for en, hi, mr, te, ta; LDC Bengali POS Tagger [3] for bn; and [20] for gu. ⁵ <https://query.wikidata.org/>

Lang	V	Train+Validation				Manually Labeled Test			
		I	T	F	κ	A	I	T	F
hi	75037	56582	25.3/5/99	2.0/1/10	0.81	4	842	11.1/5/24	2.1/1/5
mr	49512	19408	20.4/5/94	2.2/1/10	0.61	4	736	12.7/6/40	2.1/1/8
te	60865	24344	15.6/5/97	1.7/1/10	0.56	2	734	9.7/5/30	2.2/1/6
ta	121212	56707	16.7/5/97	1.8/1/10	0.76	2	656	9.5/5/24	1.9/1/8
en	103626	132584	20.2/4/86	2.2/1/10	0.74	4	470	17.5/8/61	2.7/1/7
gu	34605	9031	23.4/5/99	1.8/1/10	0.50	3	530	12.7/6/31	2.1/1/6
bn	130684	121216	19.3/5/99	2.0/1/10	0.64	2	792	8.7/5/24	1.6/1/5
kn	87760	25441	19.3/5/99	1.9/1/10	0.54	4	642	10.4/6/45	2.2/1/7

Table 2: Basic Statistics of XALIGN. |I|=# instances, |T|=avg/min/max word count, |F|=avg/min/max fact count, |V|=Vocabulary size, κ =Kappa score, |A|=# annotators

the fact. We retain a sentence s_j if the most similar fact f_i has $sim(f_i, s_j) > \tau$. For every sentence, we retain at most top- K most similar facts. We set $\tau=0.65$ and $K=10$ for our expts empirically.

4.2 Manual Annotations for Ground-Truth Data

We need manually annotated data for evaluation of Stage 2 output as well as our XF2T generation. We perform manual annotation in two phases. For both the phases, the annotators were presented with (LR language sentence s , K English facts) output by Stage 1. They were asked to mark facts present in s . There were also specific guidelines to ignore redundant facts, handle abbreviations, etc. More detailed annotation guidelines and ethical statement are mentioned here¹. In the first phase, we got 60 instances labeled per language by a set of 8 expert annotators (trusted graduate students who understood the task very well). In phase 2, we selected 8 annotators per language from the National Register of Translators⁶. We tested these annotators using phase 1 data as golden control set, and shortlisted up to 4 annotators per language who scored highest (on Kappa score with golden annotations). We report details of this test part of our XALIGN dataset in Table 2. On average, a sentence can be verbalized using two fact triples.

4.3 Stage 2: Candidate Selection

For every entity and language pair, Stage 1 outputs sentences each associated with a maximum of K facts. We use two different techniques to retain only strongly aligned (fact, sentence) pairs: transfer learning from NLI (Natural language Inference) task and distant supervision from another English-only F2T dataset. For both the methods, the input is “sentence(SEP)subject|predicate|object”.

Transfer learning from NLI: Given a premise and a hypothesis, NLI aims to predict whether the hypothesis entails, contradicts or is neutral to the premise. Fact to sentence alignment is semantically similar to NLI where the sentence and the fact can be considered

⁶ <https://www.ntm.org.in/languages/english/nrtdb.aspx>

Methods		hi	mr	te	ta	en	gu	bn	kn	Avg.
Baselines	KELM-style [1]	49.3	42.6	36.8	45.1	41.0	37.2	43.6	33.8	41.1
	WITA-style [11]	50.7	57.4	51.7	45.9	60.2	50.0	53.5	53.0	52.8
	Stage-1 + TF-IDF	75.0	68.5	69.3	71.8	73.7	70.1	78.7	64.7	71.5
Distant supervision	MuRIL [14]	76.3	68.4	74.0	75.5	70.5	78.5	62.4	67.7	71.7
	XLM-RoBERTa [7]	78.1	69.0	76.5	73.9	76.5	78.5	66.9	72.4	74.0
	mT5 [24]	79.0	71.4	77.6	78.6	76.6	80.0	69.8	70.5	75.4
Transfer learning	MuRIL [14]	71.6	71.7	76.5	75.1	73.4	78.7	79.5	71.8	74.8
	XLM-RoBERTa [7]	77.2	76.7	78.0	81.2	79.0	80.5	83.1	72.7	78.6
	mT5 [24]	90.2	83.1	84.1	88.6	84.5	85.1	75.1	78.5	83.7

Table 3: Stage-2 (Fact, Sentence) Candidate Selection F1.

as the premise and the hypothesis resp. Hence, we could infer (fact, sentence) alignment by directly probing an NLI model. We experiment with multi-lingual NLI models like XLM-R, mT5, MuRIL. We use their Xtreme-XNLI finetuned checkpoints from Huggingface⁷. If the model predicts entailment, we consider the (fact, sentence) pair to be aligned, else not. Thus, from amongst K candidate facts for every sentence, we select a subset of facts, which are compared with the golden fact list to evaluate accuracy of each model.

Distant supervision: Given an (English fact, LR language sentence) pair, we train a binary classifier to predict whether the fact and the sentence are strongly aligned or not, using the Knowledge Enhanced Language Modeling (KELM) [1] dataset. KELM has automatically aligned English (Wikipedia sentence, Wikidata facts) pairs. For a Wikipedia page corresponding to Wikidata entity e , KELM aligns sentence s with a Wikidata fact $f = \langle e, r, e' \rangle$ if s contains subject e and object e' . For every sentence s in KELM, we create a positive instance for every fact aligned with s . For every positive instance, we also create a negative instance as mentioned next. We order all the other sentences on the same Wikipedia page (which contains s) in decreasing order of semantic similarity and choose a sentence s' randomly from top 10. We skip top two sentences as they can be very similar to s . We then use the pair (fact extracted from s' , sentence s) as a negative instance. Overall, the dataset contains $\sim 1.3M$ instances. Since our XALIGN dataset is cross-lingual but KELM is English-only, for inference on output of Stage 1 data, we experiment with cross-lingual, translate-test and translate-train settings. Translate-train performs the best and hence we report results using this setting.

Table 3 shows candidate selection F1 scores across all the languages on our golden annotated dataset. Besides our proposed transfer learning and distant supervision based models, we also compare with the KELM-style [1] and WITA-style [11] alignment baselines. All experiments were run on a machine with four 10GB RTX 2080 GPUs. We finetune for 5 epochs with L2-norm weight decay of 0.001 and dropout of 0.1. We set the learning rate of $1e-5$, $2e-5$ and $1e-3$ for XLM-RoBERTa, MuRIL and mT5 resp. We observe that mT5 with transfer learning performs the best.

5 XALIGN DATASET ANALYSIS AND XF2T

We run mT5-transfer-learning Stage 2 aligner on Stage 1 output to get Train+Validation part of XALIGN. Table 2 shows dataset stats. Fig. 3 shows fact count distribution. Top 10 frequent fact properties across all languages are date of birth, occupation, position held,

⁷ MuRIL does not support vocabulary for all XNLI languages, so we finetuned it for en, hi and ur only.

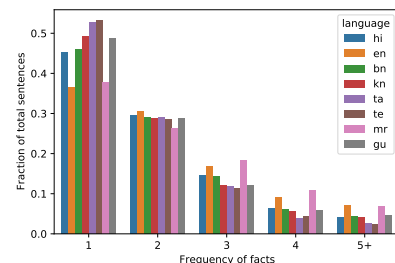


Figure 3: Fact Count Distribution across languages

Lang.	Input	Reference Text	Generated Text
hi	generate hindi : (H) natalie portman (R) place_of_birth (T) jerusalem (QR) country (QT) israel (S) early life	पोर्टमैन का जन्म इजरायल के जेरूसलम में हुआ था।	पोर्टमैन का जन्म जेरूसलम में हुआ था।
mr	generate marathi : (H) carlos i of portugal (R) date_of_birth (T) 28 september 1863 (R) date_of_death (T) 01 february 1908 (R) noble title (T) king of portugal (S) introduction	कार्लोस पहिला (२८ सप्टेंबर, १८६३ - १ फेब्रुवारी, १९०८) हा पोर्तुगालचा राजा होता.	कार्लोस पहिला (२८ सप्टेंबर १८६३ - १ फेब्रुवारी १९०८) हा पोर्तुगालचा राजा होता.
te	generate telugu : (H) hardik pandya (R) date_of_birth (T) 11 october 1993 (R) member_of_sports_team (T) baroda cricket team (R) occupation (T) cricketer (S) introduction	హార్దిక్ పాండ్యా (1993 అక్టోబరు 11 జననం) బరోడా క్రికెట్ టీంకు చెందిన ఖాచ్ ఆటగాడు.	హార్దిక్ పాండ్యా (జననం 11 అక్టోబర్ 1993) ఒక ఖాచ్‌కు చెందిన ఆటగాడు.
ta	generate tamil : (H) n. s. n. nataraj (R) country_of_citizenship (T) india (R) occupation (T) politician (R) position_held (T) member of the tamil nadu legislative assembly (S) introduction	என். எஸ். நடராஜ் என்பவர் ஓர் இந்திய அரசியல்வாதியும், தமிழகத்தின் முன்னாள் சட்டமன்ற உறுப்பினரும் ஆவார்.	என். எஸ். என் நடராஜ் ஓர் இந்திய அரசியல்வாதியும், தமிழக சட்டமன்றத்தின் முன்னாள் சட்டமன்ற உறுப்பினரும் ஆவார்.
en	generate english : (H) george knew, jr. (R) country_of_citizenship (T) united kingdom (R) date_of_birth (T) 05 march 1954 (R) occupation (T) cricketer (S) introduction	George Alan Knew (born 5 March 1954) is a former English cricketer.	George Knew (born 5 March 1954) is a former English first-class cricketer.
gu	generate gujarati : (H) anand bakshi (R) date_of_birth (T) 21 july 1930 (R) date_of_death (T) 30 march 2002 (R) occupation (T) lyricist (R) occupation (T) poet (S) introduction	આનંદ બક્ષી (જુલાઈ ૨૧ ૧૯૩૦ - માર્ચ ૩૦ ૨૦૦૨), ભારતના પ્રસિદ્ધ કવિ અને ગીતકાર હતા.	આનંદ બક્ષી (૨૧ જુલાઈ ૧૯૩૦ - ૩૦ માર્ચ ૨૦૦૨) હિન્દી ફિલ્મોના એક પ્રસિદ્ધ ગીતકાર હતા.
bn	generate bengali : (H) ramprit mandal (R) country_of_citizenship (T) india (R) member_of_political_party (T) janata dal united (R) occupation (T) politician (S) introduction	রামপ্রীত মণ্ডল হলেন একজন ভারতীয় রাজনীতিবিদ, যিনি জনতা দল (সংযুক্ত) এর রাজনীতির সাথে যুক্ত।	রামপ্রী মণ্ডল হলেন একজন ভারতীয় রাজনীতিবিদ, যিনি জনতা দল (সংযুক্ত) এর রাজনীতির সাথে যুক্ত।
kn	generate kannada : (H) barry c. barish (R) award_received (T) henry draper medal (QR) point_in_time (QT) 2017 (S) awards and honours	ಬೆರ್ರಿ ಬ್ಯಾರಿಷ್ ಅವರಿಗೆ ೨೦೧೭ ರ ಹೆನ್ರಿ ಡ್ರೇಪರ್ ಮೆಡಲ್ ನೀಡಲಾಯಿತು.	೨೦೧೭ರಲ್ಲಿ ಅವರು ಹೆನ್ರಿ ಡ್ರೇಪರ್ ಮೆಡಲ್ ಗೆ ಗೌರವಿಸಲ್ಪಟ್ಟರು.

Table 4: Test Dataset Examples with reference text and predictions from our mT5-small model.

Model	hi	mr	te	ta	en	gu	bn	kn	Avg.
Baseline	2.7	2.0	1.0	1.7	1.01	0.6	2.7	0.4	1.5
Transformer	35.4	17.3	6.9	8.8	38.8	13.2	35.6	3.1	19.9
GAT+Trans.	29.5	17.9	4.9	7.2	40.3	11.3	30.2	5.1	18.3
mT5	40.6	20.2	11.4	13.6	43.7	16.6	45.3	8.7	25.0

Table 5: XF2T BLEU scores on XALIGN test set

cast member, date of death, country of citizenship, award received, place of birth, educated at, and member of sports team.

For XF2T generation, we leverage Train+Validation part of our XALIGN dataset to train multiple multi-lingual text generation models. We train mT5-small (finetuned on Xtreme-XNLI), a basic Transformer model, and GAT (Graph Attention Network)+Transformer model (similar to [25]) for the XF2T task. For each of these models, we concatenate the facts with the section header information in the input text. We also compare with a baseline where English facts are translated to LR language and concatenated to generate output. While translating if mapped strings for entities were present in Wikidata they were directly used.

Table 4 shows XF2T prediction examples for our best (mT5) model. Table 5 shows BLEU results across different (model, language) combinations. Naïve translation baseline performs poorly. mT5 performs best with 25.0 average BLEU across all languages. While the BLEU is decent for en, hi and bn, further work needs to be done to improve XF2T generation quality for other LR languages.

6 CONCLUSION

In this work, we proposed a novel XF2T problem, contributed a new XF2T dataset, XALIGN, for 8 languages, proposed two novel F2T alignment methods, and reported BLEU results on strong baseline multi-lingual models. We strongly believe that our alignment models can significantly reduce human annotation costs in low-resource (LR) NLG. We also plan to extend this work to other LR languages. **Acknowledgements:** This research was partially funded by Ministry of Electronics and Information Technology (MeitY), Govt. of India under Sanction Order No: 11(6)/2019-HCC (TDIL).

REFERENCES

[1] O Agarwal, H Ge, S Shakeri, and R Al-Rfou. 2021. Knowledge Graph Based Synthetic Corpus Generation for Knowledge-Enhanced Language Model Pre-training. In *NAACL-HLT*. 3554–3565.
 [2] G. Attardi. 2015. WikiExtractor. <https://github.com/attardi/wikiextractor>.

[3] K Bali, M Choudhury, and P Biswas. 2010. Indian Language POS Tagset: Bengali. *Linguistic Data Consortium, LDC2010T16* (2010).
 [4] J A Botha, Z Shan, and D Gillick. 2020. Entity Linking in 100 Languages. In *EMNLP*. 7833–7845.
 [5] M Chen, S Wiseman, and K Gimpel. 2021. WIKITABLET: A Large-Scale Data-to-Text Dataset for Generating Wikipedia Article Sections. In *ACL-IJCNLP Findings*. 193–209.
 [6] W Chen, Y Su, X Yan, and W Y Wang. 2020. KGPT: Knowledge-Grounded Pre-Training for Data-to-Text Generation. *arXiv:2010.02307* (2020).
 [7] A Conneau, K Khandelwal, N Goyal, V Chaudhary, G Wenzek, F Guzmán, É Grave, M Ott, L Zettlemoyer, and V Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *ACL*. 8440–8451.
 [8] D Duma and E Klein. 2013. Generating natural language from linked data: Unsupervised template extraction. In *IWCS*. 83–94.
 [9] H Elshahar, P Vougiouklis, A Remaci, C Gravier, J Hare, F Laforest, and E Simperl. 2018. T-lex: A large scale alignment of natural language with knowledge base triples. In *LREC*.
 [10] T Ferreira, C Gardent, N Illykh, C van der Lee, S Mille, D Moussallem, and A Shimorina. 2020. The 2020 Bilingual, Bi-Directional WebNLG+ Shared Task: Overview and Evaluation Results. In *WebNLG+*. 55–76.
 [11] Z Fu, B Shi, W Lam, L Bing, and Z Liu. 2020. Partially-aligned data-to-text generation with distant supervision. *arXiv:2010.01268* (2020).
 [12] C Gardent, A Shimorina, S Narayan, and L Perez-Beltrachini. 2017. The WebNLG challenge: Generating text from RDF data. In *INLG*. 124–133.
 [13] Z Jin, Q Guo, X Qiu, and Z Zhang. 2020. Genwiki: A dataset of 1.3 million content-sharing text and graphs for unsupervised graph-to-text generation. In *COLING*. 2398–2409.
 [14] S Khanuja, D Bansal, S Mehtani, S Khosla, A Dey, B Gopalan, D K Margam, P Aggarwal, R T Nagipogu, S Dave, et al. 2021. Muri! Multilingual representations for indian languages. *arXiv:2103.10730* (2021).
 [15] K Kolluru, M Rezk, P Verga, W W Cohen, and P Talukdar. 2021. Multilingual Fact Linking. In *AKBC*.
 [16] A Kunchukuttan. 2020. The IndicNLP Library. https://github.com/anopkunchukuttan/indic_nlp_library/blob/master/docs/indicnlp.pdf.
 [17] R Lebre, D Grangier, and M Auli. 2016. Neural Text Generation from Structured Data with Application to the Biography Domain. In *EMNLP*. 1203–1213.
 [18] P Nema, S Shetty, P Jain, A Laha, K Sankaranarayanan, and M M Khapra. 2018. Generating Descriptions from Structured Data Using a Bifocal Attention Mechanism and Gated Orthogonalization. In *NAACL-HLT*. 1539–1550.
 [19] J Novikova, O Dušek, and V Rieser. 2017. The E2E dataset: New challenges for end-to-end generation. *arXiv:1706.09254* (2017).
 [20] C Patel and K Gali. 2008. Part-of-speech tagging for Gujarati using conditional random fields. In *IJCNLP Workshop on NLP for Less Privileged Languages*.
 [21] P Qi, Y Zhang, Y Zhang, J Bolton, and C D Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *ACL Demos*. <https://nlp.stanford.edu/pubs/qi2020stanzapdf>.
 [22] G Ramesh, S Doddapaneni, A Bheemaraj, M Jobanputra, Raghavan AK, A Sharma, et al. 2021. Samanantar: The Largest Publicly Available Parallel Corpora Collection for 11 Indic Languages. *arXiv:2104.05596* (2021).
 [23] E Reiter and R Dale. 1997. Building applied natural language generation systems. *NL Engineering* 3, 1 (1997), 57–87.
 [24] L Xue, N Constant, A Roberts, M Kale, R Al-Rfou, A Siddhant, A Barua, and C Raffel. 2021. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In *NAACL-HLT*. 483–498.
 [25] C Zhao, M Walker, and S Chaturvedi. 2020. Bridging the structural gap between encoding and decoding for data-to-text generation. In *ACL*. 2481–2491.