**XF2T: Cross-lingual Fact-to-Text Generation for Low-Resource Languages**

**Shivprasad Sagare, Tushar Abhishek, Bhavyajeet Singh, Anubhav Sharma**  
**Manish Gupta and Vasudeva Varma**  
IIIT-Hyderabad, India  
{shivprasad.sagare,tushar.abhishek}@research.iiit.ac.in  
{bhavyajeet.singh,anubhav.sharma}@research.iiit.ac.in  
{manish.gupta,vv}@iiit.ac.in

**Abstract**

Multiple business scenarios require an automated generation of descriptive human-readable text from structured input data. This has resulted into substantial work on fact-to-text generation systems recently. Unfortunately, previous work on fact-to-text (F2T) generation has focused primarily on English mainly due to the high availability of relevant datasets. Only recently, the problem of cross-lingual fact-to-text (XF2T) was proposed for generation across multiple languages along-with a dataset, XALIGN for eight languages. However, there has been no rigorous work on the actual XF2T generation problem. We extend XALIGN dataset with annotated data for four more languages: Punjabi, Malayalam, Assamese and Oriya. We conduct an extensive study using popular Transformer-based text generation models on our extended multi-lingual dataset, which we call XALIGNV2. Further, we investigate the performance of different text generation strategies: multiple variations of pretraining, fact-aware embeddings and structure-aware input encoding. Our extensive experiments show that a multi-lingual mT5 model which uses fact-aware embeddings with structure-aware input encoding leads to best results (30.90 BLEU, 55.12 METEOR and 59.17 chrf++) across the twelve languages. We make our code and dataset publicly available\(^1\), and hope that this will help advance further research in this critical area.

1 **Introduction**

Fact-to-text (F2T) is a natural language generation (NLG) task where input is structured data (like facts\(^2\)) and output is its natural language description. F2T systems have been shown to be effective in many applications like automated dialog systems (Wen et al., 2016), domain-specific chatbots (Novikova et al., 2017), open domain question answering (Chen et al., 2020), authoring sports reports (Chen and Mooney, 2008), financial reports (Plachouras et al., 2016), news reports (Leppänen et al., 2017), etc. Recently, several English F2T systems have been proposed, but lack of training data in low-resource languages (LRLs) implies that there are hardly any such systems for LRLs.

Across many business domains, there is abundance of facts (or key-value stores) in English, and consumers want to access that information in their own regional languages. For example, users want product descriptions, weather report, match report, financial report in various LRLs. Another related problem is to automatically populate first sentence for LRL Wikipedia pages using facts from English Wikidata. If such facts were in LRLs and there were models to do F2T in those LRLs, we could leverage those. However, neither exist. Even LRL facts on Wikidata are very sparse. Another approach could be to do F2T in English and then translate the output to LRLs. But our experiments show that this leads to poor quality primarily due to lack of robust translation systems for LRLs.

Specifically, we focus on the F2T problem of...
Table 1: Statistics of popular Fact-to-Text datasets: WikiBio (Lebret et al., 2016), E2E (Novikova et al., 2017), WebNLG 2017 (Gardent et al., 2017), WebNLG 2020 (Ferreira et al., 2020), fr-de Bio (Nema et al., 2018), KELM (Agarwal et al., 2021), WITA (Fu et al., 2020), WikiTableT (Chen et al., 2021), GenWiki (Jin et al., 2020), TREX (Elsahar et al., 2018), XAlign (Abhishek et al., 2022), and XA_LIGN V2 (ours). Alignment method could be A (automatic) or M (manual). |I|=number of instances. F/I=number of facts per instance. |P|=number of unique relations. |T|=average number of tokens per instance.

| Dataset   | Languages | A/M | |I| | F/I | |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 2017 | en   | M   | 25K | 2.95 | 373 | 22.7| No  |
| fr-de Bio | fr, de    | A   | 170K, 50K| 8.60, 12.6| 1331, 1267| 29.7, 26.9| No  |
| TREX      | en        | A   | 6.4M| 1.77 | 642 | 79.8| No  |
| WebNLG 2020 | 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 | 2.20 | 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  |
| XA_LIGN   | en + 7 LR | A   | 0.45M| 2.02 | 367 | 19.8| Yes |
| XA_LIGN V2 | en + 11 LR | A   | 0.35M| 1.98 | 374 | 19.7| Yes |

Generating LRL person biographies (like a sentence on Wikipedia page) from English Wikidata facts. While millions of English person entities exist on Wikidata, there are a total of only 168K (non-unique) person Wikidata entries across 11 LRLs of our interest. As an extreme, Assamese has only 1.7K person entries! Even worse, average number of facts per entity on Wikidata in LRLs (10.39) is less than half of that of English (22.8). Monolingual F2T for LRLs suffers from lack of training data. Translating English output (using English F2T) to LRLs leads to poor results. This necessitates us to build cross-lingual F2T generation (XF2T) systems, wherein the input is a set of English facts and output is a sentence capturing the fact-semantics in the specified LR language, as introduced in our previous work (Abhishek et al., 2022).

In (Abhishek et al., 2022), we proposed transfer learning and distance supervision based methods for cross-lingual alignment for aligning English Wikidata facts with equivalent text from LRL Wikipedia pages. In that paper, we used such alignment methods to contribute the XA_LIGN dataset which consists of sentences from LR language Wikipedia aligned with English fact triples from Wikidata. It contains data for the following eight languages: Hindi (hi), Telugu (te), Bengali (bn), Gujarati (gu), Marathi (mr), Kannada (kn), Tamil (ta) and English (en). In that paper, we focused on dataset creation and not much on the XF2T task. In this paper, we extend this dataset to four more LR languages: Punjabi (pa), Malayalam (ml), Assamese (as) and Oriya (or). Fig. 1 shows an XF2T example from our extended dataset, XA_LIGN V2. Further, we rigorously investigate models for the XF2T problem. First, we experiment with standard existing Transformer-based multi-lingual encoder-decoder models like the vanilla Transformer, IndicBART and mT5. Next, we explore performance across various training setups: bi-lingual, translate-output, translate-input and multi-lingual. Further, we systematically explore various strategies for improving XF2T generation like multi-lingual data-to-text pre-training, fact-aware embeddings, and structure-aware encoding. Overall, we make the following contributions in this work.

- We extend the XA_LIGN dataset with annotated XF2T data corresponding to four more LR languages, leading to a new dataset, XA_LIGN V2.
- We rigorously experiment with multiple encoder-decoder models, training setups, pre-training methods, and input representations toward building a robust XF2T system.
- We find that a multi-lingual mT5 model which uses fact-aware embeddings along with structure-aware input encoding leads to best results. Our best small-scale model achieves an average BLEU of 29.27, METEOR of 53.64, and chrF++ of 57.30 for XF2T across 12 languages. We make the code and dataset publicly available.

2 Related Work

Multi-lingual and Cross-lingual NLG: Recently there has been a lot of work on multi-lingual and cross-lingual NLG tasks like machine translation (Chi et al., 2021; Liu et al., 2020), question generation (Chi et al., 2020; Mitra et al., 2021),
news title generation (Liang et al., 2020), and summarization (Zhu et al., 2019; Taunk et al., 2023) thanks to models like XNLG (Chi et al., 2020), mBART (Liu et al., 2020), mT5 (Xue et al., 2021), etc. In this work, we investigate effectiveness of multiple modeling techniques for the XF2T task. Further, from a knowledge graph (KG) and text linking perspective, our work is related to tasks like entity linking (link mention in a sentence to a KG entity) (Botha et al., 2020) and fact linking (linking sentence to a set of facts) (Kolluru et al., 2021). As against this, XF2T is the problem of generating a sentence given a set of facts. XF2T is also related to graph-to-text (Ribeiro et al., 2021) where our fact triples about an entity can be mapped to a star-like graph, but no cross-lingual graph-to-text methods exist unfortunately.

**F2T Datasets**: Several F2T datasets have been proposed in the literature: WikiBio (Lebret et al., 2016), E2E (Novikova et al., 2017), WebNLG 2017 (Gardent et al., 2017), WebNLG 2020 (Ferreira et al., 2020), fr-de Bio (Nema et al., 2018), KELM (Agarwal et al., 2021), WITA (Fu et al., 2020), WikiTableT (Chen et al., 2021), GenWiki (Jin et al., 2020), TREX (Elsahar et al., 2018) and XAlign (Abhishek et al., 2022). These datasets contain text from various domains like people, sports, restaurants, airports, politicians, artists, etc. Also, these datasets vary widely in terms of statistics like the number of instances, number of facts per instance, number of unique relations and average number of tokens per instance. All of these are English only except fr-de Bio (which has French and German), WebNLG 2020 (which has English and Russian) and XAlign (which has English and 7 other LR languages). Both fr-de Bio and WebNLG 2020 propose multi-lingual but not cross-lingual F2T tasks. Unlike other datasets, XALIGN and our dataset, XALIGNV2 are cross-lingual. Our proposed dataset, XALIGNV2, contains 12 languages, has 0.55M instances, 374 unique relations, avg 19.7 tokens/instance and avg 1.98 facts/instance. Table 1 shows basic statistics of popular F2T datasets.

**F2T Generation**: Training F2T models requires aligned data with adequate content overlap. Some previous studies like WebNLG (Gardent et al., 2017) collected aligned data by crowdsourcing, while others have performed automatic alignment by heuristics like TF-IDF. In (Abhishek et al., 2022), we explored two unsupervised methods to perform a cross-lingual alignment. We leverage the “transfer learning from Natural Language Inference task” based method for this work.

Initial F2T methods were template-based and were therefore proposed on domain-specific data like medical (Bontcheva and Wilks, 2004), cooking (Cimiano et al., 2013), person (Duma and Klein, 2013), etc. They align entities in RDF triples with entities mentioned in sentences, extract templates from the aligned sentences, and use templates to generate sentences given facts for new entities. Template-based methods are brittle and do not generalize well. Recently, Seq-2-seq neural methods (Lebret et al., 2016; Mei et al., 2016) have become popular for F2T. These include vanilla LSTMs (Vougouklis et al., 2018), LSTM encoder-decoder model with copy mechanism (Shahidi et al., 2020), LSTMs with hierarchical attentive encoder (Nema et al., 2018), pretrained Transformer based models (Ribeiro et al., 2021) like BART (Lewis et al., 2020) and T5 (Raffel et al., 2020). Vougouklis et al. (2018) proposed a method which uses feedforward neural networks to encode RDF triples and concatenate them as the input of the LSTM decoder. Variations of LSTM encoder-decoder model with copy mechanism (Shahidi et al., 2020) or with hierarchical attention encoder (Nema et al., 2018) have also been proposed. Recently, pretrained Transformer based models like BART (Lewis et al., 2020) and T5 (Raffel et al., 2020) have been applied for monolingual English Fact-to-Text (Ribeiro et al., 2021).

Richer encoding of the input triples has also been investigated using a combination of graph convolutional networks and Transformers (Zhao et al., 2020), triple hierarchical attention networks (Chen et al., 2020), or Transformer networks with special fact-aware input embeddings (Chen et al., 2020). Some recent work also explores specific F2T settings like plan generation when the order of occurrence of facts in text is available (Zhao et al., 2020) or partially aligned F2T when the text covers more facts than those mentioned in the input (Fu et al., 2020). However, all of these methods focus on English fact to text only. Only recently, we proposed the XF2T problem in our previous paper Abhishek et al. (2022) but in that work, our focus was on problem formulation and dataset contribution. In this paper, we extensively evaluate multiple methods for the XF2T generation task.
Table 2: Basic Statistics of XALIGNV2. |I|=# instances, |T|=avg/min/max word count, |F|=avg #facts, |V|=Vocab. size, κ=Kappa score, |A|=#annotators. For Train+Validation, min and max fact count is 1 and 10 resp across languages.4

3 XALIGNV2: Data Collection, Pre-processing and Alignment

Data Collection and Pre-processing: We start by gathering a list of ∼95K person entities from Wikipedia each of which has a link to a corresponding Wikipedia page in at least one of our 11 LR languages. This leads to a dataset D where every instance di is a tuple (entityID, English Wikidata facts, LRL, LRL Wikipedia URL for the entityID). We extract facts (in English) from the 20201221 WikiData dump for each entity in D using the WikiData API.3 We gathered facts corresponding to only the following Wikidata property (or relation) types that capture most useful factual information for person entities: WikibaseItem, Time, Quantity, and Monolingualtext. We retain any supporting information associated with the fact triple as a fact qualifier. This leads to overall ∼0.55M data instances across all the 12 languages. Also, for each language, we gather sentences (along with section information) from 20210520 Wikipedia XML dump using same pre-processing steps as described in (Abhishek et al., 2022).

Fact-to-Text Alignment: For every (entity e, language l) pair, the pre-processed dataset has a set Fel of English Wikidata facts and a set of Wikipedia sentences Sel in that language. Next, we use a two-stage automatic aligner as proposed in (Abhishek et al., 2022) to associate a sentence in Sel with a subset of facts from Fel. We run this aligner for the new four LR languages to obtain the corresponding Train+Validation part of XALIGNV2.

Manual Annotations for Ground-Truth Data: We need manually annotated data for evaluation of our XF2T generation. Again, we follow the same procedure as outlined in (Abhishek et al., 2022) to get annotations for the new four languages in XALIGNV2. Detailed annotation guidelines are also mentioned here.1 Our annotator pool is selected from the National Register of Translators5. Annotators were in age range 25 to 40 years; 46% females and 54% males; occupations varied as linguists, editors, translators, freelancers; qualifications varied as BA, MA, MSc, LLB, PhD. We report details of this test part of our XALIGNV2 dataset in Table 2. On average, a sentence can be verbalized using ∼2 fact triples.

XALIGNV2 Dataset Analysis: Table 2 shows the dataset statistics. Figs. 2 and 3 show fact count distribution. We observe that a large percent of sentences contain more than one fact across languages. Also, the distribution is similar across languages and data subsets. Finally, Table 3 shows top 10 frequent fact relations across all the languages.

4 XF2T Approaches

In this section, we first discuss our input representation. Next, we discuss various Transformer-based methods, different training setups, multiple pretraining methods, and discussion on fact-aware embeddings.

Structure-aware Input encoding: Each input instance consists of multiple facts $F = \{f_1, f_2, \ldots, f_n\}$ and a section title $t$. A fact $f_i$ is a tuple composed of subject $s_i$, relation $r_i$, object $o_i$ and $m$ qualifiers $Q = q_1, q_2, \ldots, q_m$. Each qualifier provides more information about the fact. Each of the qualifiers $\{q_j\}_{j=1}^m$ can be linked to the fact using a fact-level property which we call as qualifier relation $qr_j$. For example, consider the sentence: “Narendra Modi was the Chief Minister of Gujarat from 7 October 2001 to 22 May 2014, preceded by Keshubhai Patel and succeeded by Anandiben Patel.” This can be represented by a fact where subject is “Narendra Modi”, relation is “position held”, object is “Chief Minister of Gujarat” and there are 4 qualifiers each with their qualifier relations as follows: (1) $q_1$=“start time”, (2) $q_2$=“end time”, (3) $q_3$=“Keshubhai Patel”, $qr_3$=“replaces”, and (4) $q_4$=“Anandiben Patel”, $qr_4$=“replaced by”.

---

3https://query.wikidata.org/
4For κ, κ is not reported since we did not get redundant judgments done due to lack of available annotators. For as, κ is not reported since we had only one annotator.
5https://www.ntm.org.in/languages/english/nrtdb.aspx
Each fact \( f_i \) is encoded as a string and the overall input consists of a concatenation of such strings across all facts in \( F \). The string representation for a fact \( f_i \) is 
\[
\langle S \rangle s_i \langle R \rangle r_{i1} \langle O \rangle o_{i1} \langle R \rangle r_{i2} \langle O \rangle o_{i2} \cdots \langle R \rangle r_{im} \langle O \rangle o_{im} \]
where \( \langle S \rangle, \langle R \rangle, \langle O \rangle \) are special tokens. Finally, the overall input with \( n \) facts is obtained as follows: “\( \text{generate [language] } f_1 f_2 \cdots f_n \langle T \rangle \langle t \rangle \)” where “[language]” is one of our 12 languages, \( \langle T \rangle \) is the section title delimiter token, and \( \langle t \rangle \) is the section title.

**Standard Transformer-based Models:** For XF2T generation, we train multiple popular multilingual text generation models on Train+Validation part of our XALIGN dataset. We use a basic Transformer model, mT5-small, and the IndicBART (Dabre et al., 2021) for the XF2T task. We do not experiment with mBART (Liu et al., 2020) and Muril (Khanuja et al., 2021) since their small sized model checkpoints are not publicly available. We train these models in a multi-lingual cross-lingual manner. Thus, we train a single model using training data across languages without any need for translation.

**Bi-lingual, Multi-lingual & Translation models:** Next, we experiment with different training setups. We first build bilingual models, where input is in English and output could be in any of the 12 languages. A drawback with this approach is the need to maintain one model per language which is cumbersome.

Further, we also train two translation based models. In the “translate-output” setting, we train a single English-only model which consumes English facts and generates English text. The English output is translated to desired language at test time using IndicTrans (Ramesh et al., 2021). In the “translate-input” setting, English facts are translated to LR language and fed as input to train a single multi-lingual model across all languages. While translating if mapped strings for entities were present in Wikidata they were directly used. A drawback with these approaches is the need for translation at test time.

**Pretraining approaches:** Pretraining has been a standard method to obtain very effective models.
even with small amounts of labeled data across several tasks in natural language processing (NLP). Domain and task specific pretraining has been shown to provide further gains (Gururangan et al., 2020). We experiment with the following four pretraining strategies on top of the already pretrained encoder-decoder model before finetuning it on XALIGN V2 dataset. (1) Multi-lingual pretraining: Wang et al. (2021) provide a noisy, but larger corpus (542192 data pairs across 15 categories) crawled from Wikipedia for English F2T task. The dataset is obtained by coupling noisy English Wikipedia data with Wikidata triples. We translate English sentences from the Wikipedia-based Wang et al. (2021)'s data to our LR languages. Thus, the multi-lingual pretraining data contains ∼6.5M data pairs. For translating sentences, we use IndicTrans (Ramesh et al., 2021). (2) Translation-based pretraining: Translation is a preliminary task for effective cross-lingual NLP. Thus, in this method, we pretrain mT5 on translation data corresponding to English to other language pairs with ∼0.25M data instances per language. (3) Two-stage pretraining: This combines the above two methods. In the first stage, we do translation-based pretraining. In the second stage, we perform multi-lingual pretraining. (4) Multi-task pretraining: This method also involves training for both translation as well as XF2T tasks. Unlike the two-stage method where pretraining is first done for translation and then for XF2T (multi-lingual pretraining), in this method we perform the two tasks jointly in a multi-task learning setup.

Fact-aware embeddings: The input to mT5 consists of token embeddings as well as position embeddings. For XF2T, the input is a bunch of facts. Facts contain semantically separate units each of which play a different role: subject, relation, object. We extend the standard mT5 input with specific (fact-aware) role embeddings. Specifically, we use four role IDs: ROL1 for subject, ROL2 for relation and qualifier relation, ROL3 for object and qualifier tokens, and ROL0 for everything else, as shown in Fig. 4. These are randomly initialized and learned while training. We hope that this explicit indication of the role played by each token in the input facts, will help the model for improved XF2T generation.

We also experimented with (1) separate role embeddings for qualifier relation and qualifier, and (2) adding fact id embeddings, i.e., if the input contains $K$ facts, we have $K$ fact IDs, and all tokens corresponding to a fact gets the same fact ID embedding. However, these did not lead to better results and thus we do not report those results.

5 Experiments

Implementation Details for Reproducibility:
We closely follow Abhishek et al. (2022)'s data-collection and XF2T alignment method for the creation of cross-lingual fact-to-text dataset for four additional languages. All XF2T generation approaches were run on a machine equipped with four 32GB V100 GPUs. For all experiments, we use IndicNLP (Kakwani et al., 2020) to convert the low-resource languages of XALIGN V2 to the unified Devanagari script. All Transformer models have 6 encoder and 6 decoder layers. For Vanilla Transformer, we follow the standard architecture and hyper-parameters suggested by Vaswani et al. (2017). For other methods, we optimize cross entropy loss using AdamW with constant learning rate of 3e-5 with L2-norm weight decay of 0.001, batch size of 20 and dropout of 0.1. We closely follow (Dabre et al., 2021) for finetuning IndicBart.

When applicable, we pretrain for 7 epochs. For multi-lingual pretraining, we use full validation set. In two-stage pretraining, we save best checkpoint of first stage (translation task) on validation set of translation task and use it to initialize model parameters for second stage. For multi-task pretraining, we create new validation set by combining validation set of translation task and XF2T task. We finetune for 30 epochs and use beam search with width of 4.

Evaluation Metrics: We use overall BLEU scores (Ramesh et al., 2021) for evaluating the multi-lingual models for English-Indic fact-sentence pairs. Following previous work, we also use METEOR (Banerjee and Lavie, 2005) and chrF++ (Popović, 2017). PARENT (Dhingra et al., 2019) relies on the word overlap between input and the prediction text. Since the input and prediction in XF2T are in different languages, we cannot compute PARENT scores.

6 Results and Analysis
Since XF2T is a very recently proposed task, there are not many baseline methods to compare with. In this section, we will present results using methods described in Section 4. Due to lack of space, we show per language results only for our best model,
but present language-wise results for other models in the Appendix. For other comparisons and analysis, we show average across all languages while pointing out any interesting per-language insights.

| Method                  | BLEU | METEOR | chrF++ |
|-------------------------|------|--------|--------|
| Vanilla Transformer     | 21.93| 50.21  | 50.89  |
| IndicBART               | 23.78| 50.80  | 53.88  |
| mT5                     | 26.13| 53.54  | 57.27  |

Table 4: XF2T scores on XALIGN V2 test set using standard Transformer-based encoder-decoder models. The best results are highlighted.

| Method                                      | BLEU | METEOR | chrF++ |
|---------------------------------------------|------|--------|--------|
| Bi-lingual mT5 (12 models)                  | 55.88| 59.91  | 52.88  |
| Translate-Output mT5 (1 model)              | 18.91| 42.83  | 49.10  |
| Translate-Input mT5 (1 model)               | 20.53| 32.24  | 55.32  |
| Multi-lingual mT5 (1 model)                 | 28.12| 33.24  | 57.27  |

Table 5: XF2T scores on XALIGN V2 test set using bi-lingual, multi-lingual and translation-based variants of mT5 model. Best results are highlighted.

**Bi-lingual, Multi-lingual & Translation models:** Table 5 shows results when mT5 model is trained using various bi-lingual, multi-lingual and translation-based settings. We observe that across all settings, the initial setting of training a single multi-lingual cross-lingual model is the best on average across all metrics. That said, for Bengali, a bi-lingual model, i.e., a model specifically trained for en→bn, is much better. Translate-output and translate-input settings lead to slightly improved models for English and Tamil respectively. On average, translate-output setting performs the worst while the multi-lingual setting performs the best. Although we use the state-of-the-art translation method, we believe low accuracy for translate-output setting is mainly due to poor translation quality.

**Pretraining approaches:** Table 6 (lines 1 to 5) shows results using different pretraining strategies. We observe that multi-lingual pretraining leads to improvements compared to no XF2T specific pretraining across 2 of the 3 metrics. Two-stage pretraining is slightly better than translation-based pretraining but not as good as multi-lingual pretraining. Finally, multi-task performs better than two-stage. For English and Bengali, we found that two-stage pretraining provided best results. However, multi-lingual pretraining is the best on average across languages, with biggest wins for Malayalam and Oriya.

**Fact-aware embeddings:** Table 6 (line 6) shows that fact-aware embeddings lead to improvements over the vanilla mT5 method without fact-aware embeddings (line 1).

In summary, we note that both the proposed methods (multi-lingual pretraining, fact-aware embedding) lead to improvements over the vanilla mT5. We also experimented with combinations

---

6Even later we observe that translation-only pretraining helps improve Bengali performance. We hypothesize this is because of huge influence English has had over Bengali historically.
| Lang. | Input | Reference Text | Generated Text |
|-------|-------|----------------|----------------|
| हिंदी | भारतीय जनता पार्टी के सदस्य (जन्म 25 जून 1969) | सुषिमा सेन | सुषिमा सेन |
of these approaches but did not observe better
results. Amongst these, multi-lingual pretraining
performs the best on two of the metrics (METEOR
and chrF++) while fact-aware embeddings perform
best on BLEU. Hence, we present language-wise
detailed comparison across these three models in
Table 7. As expected, since we have relatively
more training data for bn, en and hi, overall the
models perform well on these languages. We also
observe that the models do not perform very well
for languages of the Dravidian family (te, ta and
kn) even though ta has as many training examples
as hi. Oriya is an exception – models perform
almost as good as on Hindi, even with almost one
third amount of training data. We hypothesize that
this is limited variety in terms of types of person
entities in Oriya compared to that in Hindi.

Fig. 5 shows BLEU, METEOR and chrF++
scores for the best model across languages for test
instances with a specific number of facts. Number
of facts per instance range from 1 to 9. We observe
that the model performs best on instances with 2–4
facts across languages and across all metrics.

Table 8 shows XF2T prediction examples for our
fact-aware embedding model. In general, across
examples, we observe that the generated text is
fluent and correct. Most of the input facts are
covered by the generated sentence. Sometimes, though, the
model hallucinates and brings in extra information
in the output, e.g., for English, “Uttar Pradesh” is
not mentioned as part of input facts.

**Scaling study**: So far we presented results using
small-scale models. For the fact-aware embedding
model, we also train a large scale checkpoint with
12 encoder and 12 decoder layers. We observe that
it leads to a BLEU of 30.90, METEOR of 55.12
and chrF++ of 59.17 which is significantly better
compared to the small model as expected.

**Human Evaluation Results**: Finally, we obtain
human annotations to evaluate the perceived qual-
ity of the generated text. Table 10 shows results
for our best model across three metrics: fluency,
coverage and hallucination in the generated output.
Higher the better. The evaluation has been done
on 100 samples for 7 languages on a 5-point Likert
scale per metric. The table shows values averaged
across judgments from three annotators. Fluency
checks for coherence and grammar correctness of
generated output. Coverage verifies if most facts
are captured in the sentence correctly. Absence
of extra information verifies if the model does not
generate any hallucinated information. Fluency,
coverage and hallucination are 4.71, 4.31, 4.37 on
average for our best model respectively.

Further, we observed that even though our models
generate reasonable results, sometimes they are
wrongly penalized using automated metrics for
multiple reasons as shown in Table 9.

Table 9: Some examples of Hindi sentences generated by our best model which got low BLEU scores but are actually better than ground truth.

| Input | Reference text | Generated text | Human | Language |
|-------|----------------|----------------|-------|----------|
| (::) rajendra pratap singh <R> place_of_birth <O> 20 october 1954 <T> early life and education | राजेंद्र प्रताप सिंह का 20 अक्टूबर 1954 को भारत में प्रजामूर्ति हुआ था। | The generated output covers more facts from input | Better transliteration of dates | en |
| (::) simam-ud-dah <R> date_of_birth <O> 12 december 1995 <T> place_of_birth <O> lahore <T> personal life | झल्की में 12 दिसंबर 1995 को झल्की में हुआ था। | Better transliteration of dates | Better transliteration of dates | IN |
| (::) usan mellor <R> country_of_originship <O> south africa <T> occupation <O> cricketer <T> introduction | उसन मेल्लर का जन्म 12 दिसंबर 1983 में आफ्रिका में हुआ। | Better transliteration of dates | Better transliteration of dates | IN |

Table 10: Human Evaluation Results for our best model

| Fluency | Coverage | Hallucination |
|---------|----------|--------------|
| in | 4.89 | 4.75 | 4.37 |
| mr | 4.87 | 4.92 | 4.73 |
| ta | 4.85 | 4.97 | 4.59 |
| te | 4.65 | 4.18 | 4.14 |
| pa | 4.69 | 4.35 | 4.28 |
| ur | 4.70 | 4.35 | 4.44 |
| hi | 4.69 | 4.17 | 4.29 |

7 Conclusion

In this paper, we worked on the XF2T problem. We
contributed the XALIGNV2 dataset which has inst-
ances with English facts aligned to 12 languages.
We investigated several multi-lingual Transformer
methods with different training setups, pretraining
setups and input representations. We obtained
models with best metrics of 30.90 BLEU, 55.12
METEOR and 59.17 chrF++ for XF2T. We make our
code and dataset publicly available to empower
future research in this critical area.

23
8 Ethical Concerns

We do not foresee any harmful uses of this technology. In fact, F2T generation systems are vital in many downstream Natural Language Processing (NLP) applications like automated dialog systems (Wen et al., 2016), domain-specific chatbots (Novikova et al., 2017), open domain question answering (Chen et al., 2020), authoring sports reports (Chen and Mooney, 2008), etc. We believe that these systems will be useful for powering business applications like Wikipedia text generation given English Infoboxes, automated generation of non-English product descriptions using English product attributes, etc.

As part of this work, we collected labeled data as discussed in Section 3. The dataset does not involve collection or storage of any personally identifiable information or offensive information at any stage. Human annotators were paid appropriately while performing data collection according to the standard wages set by National Translation Mission (https://www.ntm.org.in/) and mutually agreed upon. The data is publicly released under MIT Open-Source License. The annotation exercise was approved by the Institutional Review Board of our institute.

Usage of XALIGN dataset: Our usage was consistent with its intended use. The dataset was made available to us by the authors under MIT Open-Source License.

References

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 The World Wide Web Conference, pages 171–175.

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, pages 3554–3565.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

K Bontcheva and Y Wilks. 2004. Automatic report generation from ontologies: the miakt approach. In Conf. on application of natural language to info. systems, pages 324–335.

J A Botha, Z Shan, and D Gillick. 2020. Entity linking in 100 languages. In EMNLP, pages 7833–7845.

D L Chen and R J Mooney. 2008. Learning to sportscast: a test of grounded language acquisition. In ICML, pages 128–135.

M Chen, S Wiseman, and K Gimpel. 2021. Wikitable: A large-scale data-to-text dataset for generating wikipedia article sections. In ACL-IJCNLP Findings, pages 193–209.

W Chen, Y Su, X Yan, and W Y Wang. 2020. Kcpt: Knowledge-grounded pre-training for data-to-text generation. arXiv:2010.02307.

Z Chi, L Dong, S Ma, S Huang, X-L Mao, H Huang, and F Wei. 2021. Mt6: Multilingual pretrained text-to-text transformer with translation pairs.

Z Chi, L Dong, F Wei, W Wang, X-L Mao, and H Huang. 2020. Cross-lingual natural language generation via pre-training. In AAAI, volume 34, pages 7570–7577.

P Cimiano, J Lükner, D Nagel, and C Unger. 2013. Exploiting ontology lexica for generating natural language texts from rdf data. In European Workshop on NLG, pages 10–19.

Raj Dabre, Himani Shrotriya, Anoop Kunchukuttan, Ratish Pudupully, Mitesh K Kapraha, and Pratyush Kumar. 2021. Indicbart: A pre-trained model for natural language generation of indic languages. arXiv preprint arXiv:2109.02903.

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.

D Duma and E Klein. 2013. Generating natural language from linked data: Unsupervised template extraction. In IWCS, pages 83–94.

H Elsahar, P Vougiouklis, A Remaci, C Gravier, J Hare, F Laforest, and E Simperl. 2018. T-rex: A large scale alignment of natural language with knowledge base triples. In LREC.

Thiago Ferreira, Claire Gardent, Nikolai Ilinykh, Chris Van Der Lee, Simon Mille, Diego Moussallem, and Anastasia Shimorina. 2020. The 2020 bilingual, bi-directional webnlg+ shared task overview and evaluation results (webnlg+ 2020). In Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+).

Z Fu, B Shi, W Lam, L Bing, and Z Liu. 2020. Partially-aligned data-to-text generation with distant supervision. arXiv:2010.01268.
C Gardent, A Shimorina, S Narayan, and L Perez-Beltrachini. 2017. The webnlg challenge: Generating text from rdf data. In INLG, pages 124–133.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don’t stop pretraining: Adapt language models to domains and tasks. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8342–8360.

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, pages 2398–2409.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, NC Gokul, Avik Bhattacharyya, Mitesh M Khapra, and Pratyush Kumar. 2020. Indicenpsuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for indian languages. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 4948–4961.

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. Murlit: Multilingual representations for indian languages. arXiv:2103.10730.

K Kolluru, M Rezk, P Verga, W W Cohen, and P Talukdar. 2021. Multilingual fact linking. In AKBC.

R Lebret, D Grangier, and M Auli. 2016. Neural text generation from structured data with application to the biography domain. In EMNLP, pages 1203–1213.

L Leppänen, M Munezero, M Granroth-Wilding, and H Toivonen. 2017. Data-driven news generation for automated journalism. In INLG, pages 188–197.

M Lewis, Y Liu, N Goyal, M Ghazvininejad, A Mohamed, O Levy, V Stoyanov, and L Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pretraining for natural language generation, translation, and comprehension. In ACL, pages 7871–7880.

Y Liang, N Duan, Y Gong, N Wu, F Guo, W Qi, M Gong, L Shou, D Jiang, G Cao, et al. 2020. Xglue: A new benchmark dataset for cross-lingual pre-training, understanding and generation. arXiv:2004.01401.

Y Liu, J Gu, N Goyal, X Li, S Edunov, M Ghazvininejad, M Lewis, and L Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. TACL, 8:726–742.

H Mei, M Bansal, and M R Walter. 2016. What to talk about and how? selective gen. using lstms with coarse-to-fine alignment. In NAACL-HLT, pages 720–730.

Rajarshee Mitra, Rhea Jain, Aditya Srikanth Veerubhotla, and Manish Gupta. 2021. Zero-shot multilingual interrogative question generation for “people also ask” at bing. In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, pages 3414–3422.

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, pages 1539–1550.

J Novikova, O Dušek, and V Rieser. 2017. The e2e dataset: New challenges for end-to-end generation. arXiv:1706.09254.

V Plachouras, C Smiley, H Bretz, O Taylor, J L Leidner, D Song, and F Schilder. 2016. Interacting with financial data using natural language. In SIGIR, pages 1121–1124.

Maja Popović. 2017. chrf++: words helping character n-grams. In Proceedings of the second conference on machine translation, pages 612–618.

C Raffel, N Shazeer, A Roberts, K Lee, S Narang, M Matena, Y Zhou, W Li, and P J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. JMLR, 21:1–67.

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.

Leonardo FR Ribeiro, Martin Schmitt, Hinrich Schütze, and Iryna Gurevych. 2021. Investigating pretrained language models for graph-to-text generation. In Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI, pages 211–227.

H Shahidi, M Li, and J Lin. 2020. Two birds, one stone: A simple, unified model for text generation from structured and unstructured data. In ACL, pages 3864–3870.

Dhaval Taunk, Shyprasad Sagare, Anupam Patil, Shivansh Subramanian, Manish Gupta, and Vasudeva Varma. 2023. Xwikigen: Cross-lingual summarization for encyclopedic text generation in low resource languages. In Proceedings of the ACM Web Conference 2023, pages 1703–1713.

A Vaswani, N Shazeer, N Parmar, J Uszkoreit, L Jones, A N Gomez, L Kaiser, and I Polosukhin. 2017. Attention is all you need. In NIPS, pages 5998–6008.

P Vougiouklis, H Elsahar, L-A Kaffee, C Gravier, F Laforest, J Hare, and E Simperl. 2018. Neural wikipedian: Generating textual summaries from knowledge base triples. J. Web Semantics, 52:1–15.

Qingyun Wang, Semih Yavuz, Xi Victoria Lin, Heng Ji, and Nazneen Rajani. 2021. Stage-wise fine-tuning for graph-to-text generation. In Proceedings of the
Table 12 shows detailed results per language using various bi-lingual, multi-lingual and translation-based settings.

A Limitations

In this work, we performed XF2T for a total of 12 languages. Clearly, the work can be extended to include many more low resource languages. Further, the amount of training data per language varies significantly. Gathering more labeled data across languages is difficult but should help improve accuracy of the trained models.

For some languages, finding qualified annotators was very difficult. For Assamese, we could obtain only one annotator. For Oriya, we found two annotators but due to their limited bandwidth, we did not get overlapping samples annotated by them and hence cannot compute inter-annotator agreement. While our annotation guidelines are clear, and inter-annotator agreement is high on most languages, we acknowledge that the annotation quality may have suffered for Assamese and Oriya.

The best automatic evaluation results from our models as well as human evaluation results show that there is a lot of scope for further work in this area.

B Detailed results

Table 11 shows detailed results per language. We observe that IndicBART performed exceptionally well for Bengali but is exceptionally poor on English.
| Language | Vanilla Transformer | IndicBART | mT5 |
|----------|---------------------|-----------|-----|
|          | BLEU | METEOR | chrF++ | BLEU | METEOR | chrF++ | BLEU | METEOR | chrF++ |
| hi       | 35.04 | 63.46 | 60.85 | 40.44 | 66.41 | 66.27 | 44.65 | 68.58 | 68.49 |
| mr       | 18.28 | 50.66 | 49.87 | 28.08 | 55.35 | 57.73 | 26.47 | 56.85 | 59.17 |
| te       | 6.95  | 30.17 | 41.30 | 12.67 | 41.82 | 50.40 | 14.46 | 43.45 | 52.58 |
| ta       | 14.67 | 44.84 | 53.03 | 19.37 | 45.78 | 56.63 | 18.37 | 46.15 | 57.42 |
| en       | 37.12 | 65.32 | 59.69 | 40.47 | 62.35 | 63.35 | 46.94 | 70.60 | 65.20 |
| gu       | 15.66 | 47.70 | 46.29 | 19.16 | 47.92 | 49.30 | 22.69 | 50.31 | 51.36 |
| hi       | 48.35 | 71.18 | 75.68 | 55.90 | 79.29 | 80.51 | 40.38 | 61.71 | 68.71 |
| Mr       | 4.78  | 28.96 | 37.60 | 10.30 | 33.55 | 46.65 | 10.66 | 32.58 | 46.92 |
| te       | 41.07 | 66.15 | 65.57 | 24.88 | 55.91 | 54.48 | 41.98 | 66.14 | 66.47 |
| mr       | 16.74 | 49.36 | 48.40 | 20.62 | 46.87 | 52.23 | 24.90 | 54.56 | 57.25 |
| te       | 12.23 | 37.85 | 44.94 | 14.13 | 38.69 | 50.36 | 13.11 | 40.83 | 49.64 |
| hi       | 18.37 | 46.57 | 57.10 | 3.36  | 30.41 | 46.35 | 19.23 | 45.68 | 57.54 |
| Mr       | 45.79 | 69.90 | 63.79 | 50.31 | 70.47 | 65.43 | 45.12 | 69.88 | 64.11 |
| en       | 12.29 | 37.71 | 37.01 | 18.25 | 42.25 | 46.27 | 20.84 | 48.71 | 49.90 |
| el       | 53.04 | 75.42 | 78.12 | 20.57 | 46.58 | 56.00 | 40.56 | 67.75 | 71.36 |
| Mr       | 8.71  | 31.02 | 41.16 | 7.93  | 27.58 | 44.47 | 7.75  | 30.82 | 41.44 |
| en       | 24.28 | 55.37 | 55.49 | 18.60 | 47.39 | 51.47 | 26.16 | 56.49 | 57.22 |
| el       | 21.92 | 51.10 | 47.82 | 26.24 | 53.18 | 51.57 | 24.42 | 51.64 | 49.28 |
| el       | 45.53 | 62.91 | 65.30 | 7.37  | 29.40 | 33.80 | 43.43 | 64.12 | 65.20 |
| as       | 9.76  | 26.48 | 29.80 | 7.15  | 25.25 | 32.19 | 10.89 | 30.77 | 35.00 |
| Avg      | 25.88 | 50.91 | 52.88 | 18.91 | 42.83 | 49.10 | 26.53 | 52.24 | 53.32 |

Table 11: XF2T scores on XALIGNV2 test set using standard Transformer-based encoder-decoder models. Best results for a (metric, language) combination are highlighted.

| Language | Bi-lingual (12 models) | Translate-Output (1 model) | Translate-Input (1 model) | Multi-lingual (1 model) |
|----------|-------------------------|---------------------------|---------------------------|-------------------------|
|          | BLEU | METEOR | chrF++ | BLEU | METEOR | chrF++ | BLEU | METEOR | chrF++ | BLEU | METEOR | chrF++ |
| hi       | 41.07 | 66.15 | 65.57 | 24.88 | 55.91 | 54.48 | 41.98 | 66.14 | 66.47 | 44.65 | 68.58 | 68.49 |
| mr       | 16.74 | 49.36 | 48.40 | 20.62 | 46.87 | 52.23 | 24.90 | 54.56 | 57.25 | 26.47 | 56.85 | 59.17 |
| te       | 12.23 | 37.85 | 44.94 | 14.13 | 38.69 | 50.36 | 13.11 | 40.83 | 49.64 | 14.46 | 43.45 | 52.58 |
| hi       | 18.37 | 46.57 | 57.10 | 3.36  | 30.41 | 46.35 | 19.23 | 45.68 | 57.54 | 18.37 | 46.15 | 57.42 |
| Mr       | 45.79 | 69.90 | 63.79 | 50.31 | 70.47 | 65.43 | 45.12 | 69.88 | 64.11 | 46.94 | 70.60 | 65.20 |
| en       | 12.29 | 37.71 | 37.01 | 18.25 | 42.25 | 46.27 | 20.84 | 48.71 | 49.90 | 22.69 | 50.31 | 51.36 |
| el       | 53.04 | 75.42 | 78.12 | 20.57 | 46.58 | 56.00 | 40.56 | 67.75 | 71.36 | 40.38 | 61.71 | 68.71 |
| Mr       | 8.71  | 31.02 | 41.16 | 7.93  | 27.58 | 44.47 | 7.75  | 30.82 | 41.44 | 10.66 | 32.58 | 46.92 |
| en       | 24.28 | 55.37 | 55.49 | 18.60 | 47.39 | 51.47 | 26.16 | 56.49 | 57.22 | 26.22 | 56.71 | 57.01 |
| el       | 21.92 | 51.10 | 47.82 | 26.24 | 53.18 | 51.57 | 24.42 | 51.64 | 49.28 | 26.96 | 54.82 | 52.33 |
| el       | 45.53 | 62.91 | 65.30 | 7.37  | 29.40 | 33.80 | 43.43 | 64.12 | 65.20 | 47.17 | 67.32 | 71.20 |
| as       | 9.76  | 26.48 | 29.80 | 7.15  | 25.25 | 32.19 | 10.89 | 30.77 | 35.00 | 12.61 | 34.93 | 36.91 |
| Avg      | 25.88 | 50.91 | 52.88 | 18.91 | 42.83 | 49.10 | 26.53 | 52.24 | 53.32 | 28.13 | 53.54 | 57.27 |

Table 12: XF2T scores on XALIGNV2 test set using bi-lingual, multi-lingual and translation-based variants of mT5 model. Best results for a (metric, language) combination are highlighted.