Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 3417–3424
Marseille, 20-25 June 2022

The LTRC Hindi-Telugu Parallel Corpus

Vandan Mujadia, Dipti Misra Sharma
MT-NLP Lab, LTRC, KCIS, IIIT-Hyderabad, India
{vandan.mu@research., dipti@}iiit.ac.in

Abstract

We present the Hindi-Telugu Parallel Corpus of different technical domains such as Natural Science, Computer Science, Law and Healthcare along with the General domain. The qualitative corpus consists of 700K parallel sentences of which 535K sentences were created using multiple methods such as extract, align and review of Hindi-Telugu corpora, end-to-end human translation, iterative back-translation driven post-editing and around 165K parallel sentences were collected from available sources in the public domain. We present the comparative assessment of created parallel corpora for representativeness and diversity. The corpus has been pre-processed for machine translation, and we trained a neural machine translation system using it and report state-of-the-art baseline results on the developed development set over multiple domains and on available benchmarks. With this, we define a new task on Domain Machine Translation for low resource language pairs such as Hindi and Telugu. The developed corpus (535K) is freely available for non-commercial research and to the best of our knowledge, this is the well curated, largest, publicly available domain parallel corpus for Hindi-Telugu.

Keywords: Machine Translation, Domain Machine Translation, Parallel Corpora, Indian Languages, Hindi, Telugu

1. Introduction

Generally, in multilingual society such as India, people use one language for ethnic identity, another for business transactions, another for official dealings and yet another for entertainment, rituals and so on(Annamalai, 2001). The language used by a group (eg. Dakkhini[1], a variety of Hindi, spoken in the Deccan region of India) living in among another group (e.g. Telugu) is an illustrative case for linguistic convergence(Vasanta et al., 2010). According to the Census of India of 2011[2], Hindi[3] is one of the major languages spoken primarily in the Indian subcontinent. It has around 528 million native speakers and around 700 million total speakers. It is the official language of many North and Central states in India and acts as lingua franca for many Indians. In contrast, Telugu, the largest member of the Dravidian language family, is the official language[4] of the states of Telangana and Andhra Pradesh. It has 95 million native speakers in India and around the world. Hindi and Telugu both are part of 22 official languages recognized by India. For such a diverse group of language users, in order to have an appropriate communication or to provide educational content in multiple languages, such as Hindi and Telugu, the need for translation systems to translate content in the respective languages becomes essential.

Hence, there is immense potential for Hindi-Telugu machine translation and yet, the qualitative parallel corpora available between these two languages are limited and specially for the domains. The non-availability of domain parallel data is a major issue in the training of domain specific machine translation models i.e educational domain (technical domain) such as science. Therefore, our work is an effort where we create, align, collect quality parallel corpora for Hindi-Telugu for **General** (named for type of corpora such as news, blogs, etc) as well as **Technical** domains such as Natural science, computer science, **Law** and **Healthcare** and developed corpora (535K-[5]) will be freely available for non-commercial research to the community.

2. Related Work

In recent times, neural machine translation (NMT) shows high performance gain in terms of output fluency and translation quality, when large amounts of parallel data are available (Barrault et al., 2020). Unfortunately, for most language pairs, parallel data is either scarce or non-existent. When it comes to Indian languages, in recent times, the machine translation community has witnessed increasing interest. However, this demands availability of parallel corpora for these languages (Wang et al., 2021). In this regards, Multilingual Machine Translation shared task (WAT-2021 MultiIndicMT) (Nakazawa et al., 2021) compiled many existing parallel corpora involving Indian languages (10 Indian Languages, Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Oriya, Punjabi, Tamil and Telugu) and English and placed them at one location[6]. Earlier to this, The Indian government TDIL program and the Indian Language Corpora Initiative developed a parallel corpora for English and 16 Indian languages (Jha, 2010). (Mayer

[1] https://en.wikipedia.org/wiki/Deccani_language
[2] https://censusindia.gov.in/2011Census/pes/Pesreport.pdf
[3] https://en.wikipedia.org/wiki/Hindi
[4] https://en.wikipedia.org/wiki/Telugu_language
[5] https://ltrc.iiit.ac.in/nlpmt/download.php
[6] http://lotus.kuee.kyoto-u.ac.jp/WAT/indic-multilingual/

and Cysouw, 2014) and (Resnik et al., 1999) presented efforts to create a massively parallel Bible corpus with over 900 translations in more than 830 language varieties. The IITB English-Hindi parallel corpora was an effort (Kunchukuttan et al., 2017) where a compilation of English-Hindi parallel corpora available in the public domain were collected and made available for the research community. OdiEnCorp 2.0 (Parida et al., 2020), the parallel corpora between English and Odia was developed. (Philip et al., 2019) enhanced the IIT-B English-Hindi corpora with back-translation. (Siripragada et al., 2020) and (Haddow and Kirefu, 2020) compiled several multilingual Indian government websites such as PMIndia[7], PIB[8] and automatically aligned Indian languages with the English and released for the community. However most of these efforts were English and Indian languages centric and do not involve Indian to Indian language parallel corpora.

To overcome this scarcity of sizable parallel data, as an alternative, quasi-parallel or comparable parallel training corpora provides an important resource for training machine translation systems for resource scared language pairs. Recently released, Samanantar(Ramesh et al., 2021), provides significantly large parallel corpora for Indic languages contains 49.7 million sentence pairs between English and 11 Indic languages (Assamese, Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Oriya, Punjabi, Tamil and Telugu). It was developed using web-crawled monolingual corpora, document OCR for extracting sentences from scanned documents, multilingual representation models for aligning sentences, and approximate nearest neighbour search for searching in a large collection of sentences. Further, they extract Indic to Indic language parallel corpora using English as the pivot language. The efforts were similar to (Schwenk et al., 2019) where multilingual sentence embeddings were used to automatically extract parallel sentences from the content of Wikipedia articles in 96 languages. Since such comparable corpora are not of as high quality as manually annotated parallel data, using them for training can have a negative effect(Dakwale and Monz, 2019)(Muischnek and Müürisep, 2018) on the translation performance of an NMT model. To overcome the impact of noisy corpora, (Dakwale and Monz, 2019) proposed distillation as a remedy from the teacher network trained on the clean parallel corpora. Hence, development of human curated parallel corpora remains valuable resource for high quality machine translation.

3. Methods for Parallel Corpora Creation

In General, well curated corpus construction is a difficult and laborious task. Usually, a corpus is developed according to specific objectives and purposes using several methods. The objective of a corpus construction would ask details such as kinds of domain involved, size of the corpora, and sources of corpora, etc. To answer this, as mentioned earlier, we aim to create quality machine translation parallel corpora involving Hindi-Telugu language pairs for general (named for type of corpora such as news, blogs, etc) as well as different technical domains such as Natural Science, Computer Science, Law and Healthcare. In our approach we decided to use several methods such as, Collect, Clean and Review existing parallel corpora, Extracting and aligning available Hindi-Telugu corpora, end-to-end translation and in-house adaptive back-translation driven post-editing to create sizable parallel corpora for Hindi-Telugu. Each of these methods are described in the following subsections.

3.1. Collect, Clean and Review Parallel Corpora from Existing Sources

As a part of our effort to Hindi-Telugu parallel data creation, we collected parallel data from different sources, accounted for 165K parallel sentences. The sources are from OPUS[9](Tiedemann, 2012) such as: bible-uedin, globalvoices, gnome, kde4, opensubtitles, tanzil, tatoeba, ubuntu, wikimedia. We also included ILCI(Jha, 2010) Hindi-Telugu parallel corpora after prepossessing, cleaning and thorough sampled human review. Apart from this, recently, Samanantar(Ramesh et al., 2021) released around 2425K Hindi-Telugu parallel sentence corpora by pivoting English language. We analyse some properties of this corpora and its impact on translation quality in comparative analysis and result section.

3.2. Extract, Align and Review Hindi-Telugu data

We identified few data sources of the technical domain which provides maths, science, social science and environment content in Hindi and Telugu. Usually, these texts are aligned page by page and passage by passage across the languages. We extracted, cleaned the collected text, converted it to Unicode format for both languages by maintaining respective page number information. Then we semi-automatic aligned the extracted text at sentence level. Similarly NPTEL[10] is the largest online repository in the world of courses in engineering, basic sciences, humanities and social sciences subjects. Some of these lecture texts are available in Hindi and Telugu as a translation from English. We collected these texts by maintaining page and other alignment information such as figure ID, table ID. Using these alignments and LASER[11] embedding(Artetxe and Schwenk, 2019) we automatically aligned textual content sentence by sentence and then carried out hu-

[7] https://www.pmindia.gov.in/
[8] https://www.pib.gov.in/
[9] https://opus.nlpl.eu/
[10] https://nptel.ac.in/
[11] https://github.com/facebookresearch/LASER

man verification for all collected text. We discarded miss-aligned sentence pairs identified in human vetting. To perform mentioned human verification for the aligned sentences, we hired a couple of persons with decent bilingual understanding for both of these languages.

Domain	Train-Sentences	Dev-Sentences
Collected (3.1)	165000	-
Technical	229309	1000
Law	158812	1000
Healthcare	90901	-
General	56334	840
Total	700356	2840

Table 1: Statistics of created Parallel Corpora (Train and Development) : Hindi-Telugu. Here, General named for type of corpora such as news, blogs, etc

3.3. End to End Translation

We hired 3 translation agencies to translate the provided text of Hindi - Telugu (both directions) by following below described guidelines. As the task involves technical domain translation, we particularly asked agencies to maintain certain standards in translation.

- Read the source text (complete passage) carefully before starting the translation process.
- Try to fully capture the conveyed meaning of the source text (sentence by sentence) in your translation as much as possible and make translations as understandable as the Source.
- Do not be satisfied with just one reading. Read the translation several times, silently and aloud. This will help to check if the words sound clear and harmonious. You may find that some of the words you use are not appropriate.
- Use the technical term and expression consistently throughout a translation of a text. While writing a topic related to translations or other repetitive content, re-use the same target side terminology as much as possible. Translate Terminology only if we see their usage in technical books such as NCERT[12] otherwise they can be transliterated.
- Compare the translation to the original text and see that you convey the same natural message and context.

We employed an in-house team of experienced translators to validate each translated sentence by the agencies to maintain the quality and reject the translations if there are any discrepancies.

[12] https://ncert.nic.in/textbook.php

3.4. Iterative Back-translation driven Post-Editing

In post-editing, a human translator works on the translation produced by the machine in order to provide a final translated version. Here, we hired an in-house team of translators to post-edit pre-filled machine translation output for a given source sentence. For this task, translators can directly accept the machine translation output or make appropriate edits when necessary. We also give liberty to reject particular sentence if machine generated translation requires total rewriting. As always, translators need to maintain the resulting translation to be accurate and read naturally in the target language.

Here, to generate machine translation output for post-editing we trained initial Hindi-Telugu and Telugu-Hindi MT model using available Hindi-Telugu parallel corpora as mentioned in section-3.1. As we aim to develop domain parallel corpora, we exploit in-domain monolingual data with the iterative-back-translation(Hoang et al., 2018). Back-translation(Sennrich et al., 2015) is a method to create synthetic source texts from clean target texts by using an MT model that is trained in the target–to–source direction. We can build a better MT model by combining back-translated domain parallel data with the original clean parallel data, and then we repeat this process for other language direction as described in (Hoang et al., 2018). The next iteration of entire process utilises better MT model to back-translate domain data further, and use this generated data in order to build an even better system and so on. The final system benefits from domain monolingual data in both the source and target languages. Further we also periodically added created parallel corpora in this iterative back-translation model creation process to further improve quality.

After a few rounds of iterative-back-translation and after reaching a certain quality of MT output, We decided to do sampled validation (50% of overall corpora development allocated to this method) of translations obtained using Adaptive Back-translation driven Post-Editing method.

Each of these methods have their own advantages as end-to-end translation produces high quality parallel corpora while adaptive back translation driven post editing reduces the turnaround time with qualitative translations.

We identified several online publicly available Hindi and Telugu sources such as government and other websites/bulletins containing Technical, Health and Law domain as our source language texts. After completing the translation (or post-editing) and the required review process, we globally randomised the order of parallel corpora such that one would not be able to reconstruct the original source text.

Method	Parallel Sentences
Collect, Clean and Review Parallel Corpora from Existing Sources	165K
Extract, Align and Review	90K
End-To-End Translation	230K
Iterative Back-translation driven Post-Editing	215K
Total	700K

Table 2: Statistics of Created and Collected Parallel Corpora : Hindi-Telugu

Domain	Domain	Hindi	Telugu
#Type	Collected (3.1)	124094	347922
	Technical	176573	278286
	Law	100179	164941
	Healthcare	43164	65078
	General	48276	106561
	Overall	273926	450789
#Token	Collected(3.1)	3221645	2171959
	Technical	4533873	3072882
	Law	2564943	1692163
	Healthcare	1594433	1124547
	General	976334	647838
	Overall	9563141	9563141

Table 3: Type Token statistics of created Parallel Corpora : Hindi-Telugu

Analysis	#
% of translation picked as it is	30.05
Avg key-strokes per translation	14.33
Avg no of words Edited per translation	2.57

Table 4: Key-Strokes Analysis for Iterative Back-translation driven Post-Editing (6445 Sentences)

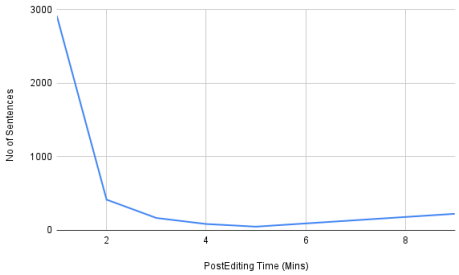

Figure 1: Average Time taken for post-editing a sentence : Iterative Back-translation driven Post-Editing (6445 Sentences)

4. Corpus Statistics

Table-1 describes our overall efforts for Hindi-Telugu parallel corpora development across the domains. We considered all created corpora as train sets and developed development sets by manually translating Hindi sentences into Telugu. We developed around 43% of data for the technical domain while 30% and 17% of overall parallel corpora for the law and healthcare domains.

Table-2 shows overall distribution of developed corpora in-terms of methods. We see that over 40% of parallel sentences were created using iterative back-translation and while 43% of sentences were developed by hiring external agencies following an end-to-end translation across the domains. Over 16% of the parallel corpora were created by manually aligning Hindi-Telugu sentences as mentioned in section-3.2.

Table-3 shows type and token stats over the complete corpora as well as across the domains. Here, high type numbers across the corpora suggest that the created parallel corpora is diverse in nature. We observe that #type and #token highly dependent on corpora size. Technical domain has highest types while healthcare has comparatively low types across all the domains. This must be due to in-domain variability, as the technical domain contains diverse sub-domains such as natural-sciences, computer sciences, management, etc while healthcare contains text only around Covid-19 related topics. We also see that Telugu shows high number of types compare to Hindi which justifies that Telugu is more agglutinative[13] language than Hindi.

4.1. Involved Human Efforts in Post-Editing

As described in section-3.4, we carried overall back-translation driven post-editing by hiring a team of Hindi-Telugu translators (4 translators). During the post-editing, we also captured and measured post-editing efforts in-terms of keyboard keystrokes and time spent over post-editing a machine translation output as explained in (Ahsan et al., 2021).

Table-4 shows keystrokes analysis over 6.5K sentences of healthcare domain for Hindi to Telugu direction. The average source token length of these sentences is 19.23. Here, we observed that around 30% of translation outputs are taken as it is without any edit. For remaining sentences we observed that on average 14 keystrokes were applied to correct a machine output sentence in post-editing. This accounts, on average, editing of 2.5 words in this sample post-editing stats. Figure-1 shows the temporal analysis over the same post-editing experience. Here we observed that around 3K sentences were post-edited within a minute while most of the post-editing completed within 4 minutes. We also ob-

[13] https://en.wikipedia.org/wiki/Agglutinative_language

Attributes	Corpus	Hindi	Telugu
Avg Sentence Length (Token)	Collected (165K - this work)	18.95	12.44
	Created (535K - this work)	18.10	12.24
	Created (56K - this work : General Domain)	17.33	11.49
	Samanatar (2425K - (Ramesh et al., 2021))	11.94	8.51
	Monolingual (18M-(Kakwani et al., 2020))	20.14	-
	Monolingual (11M-(Kakwani et al., 2020))	-	11.72
Avg Sentence Length (Char)	Collected (165K - this work)	84	87
	Created (535K - this work)	92	90
	Created (56K - this work : General Domain)	18.10	12.24
	Samanatar (2425K - (Ramesh et al., 2021))	53	56
	Monolingual (18M-(Kakwani et al., 2020))	94	-
	Monolingual (11M-(Kakwani et al., 2020))	-	79
Type Token Ratio (TTR)	Collected (165K - this work)	0.031	0.15
	Monolingual (165K-(Kakwani et al., 2020))	0.04	-
	Monolingual (165K-(Kakwani et al., 2020))	-	0.11
Type Token Ratio (TTR)	Created (535K - this work)	0.028	0.07
	Monolingual (535K-(Kakwani et al., 2020))	0.025	-
	Monolingual (535K-(Kakwani et al., 2020))	-	0.07
Type Token Ratio (TTR)	Created (56K - this work : General Domain)	0.049	0.16
	Monolingual (56K-(Kakwani et al., 2020))	0.061	-
	Monolingual (56K-(Kakwani et al., 2020))	-	0.079
Type Token Ratio (TTR)	Samanatar (2425K - (Ramesh et al., 2021))	0.0071	0.042
	Monolingual (2425K-(Kakwani et al., 2020))	0.013	-
	Monolingual (2425K-(Kakwani et al., 2020))	-	0.046

Table 5: Attributes of different Corpora (parallel and monolingual) : Hindi-Telugu

served outliers where the logged time was more than 10 minutes which must be due to several external reasons.

5. Comparative Analysis

In this section, we present analysis on different parallel corpora and compare it with respective monolingual corpora on multiple aspects to assess representativeness and diversity. We considered naive aspects(textinspector, 2022) such as average sentence length and Type-Token Ratio (TTR) across corpora as measures. Average sentence length (#token/#characters) over corpora indicates how representative the corpora is and considered as one of essential language properties. TTR or the Type-Token Ratio of text, determine whether that text is of good quality or not. TTR is a measure of the lexical diversity (and some say, hence quality) of a text(Litvinova et al., 2017).

Table-5 shows average sentences length in terms of tokens and characters across the parallel corpora and sizable monolingual corpora(Kakwani et al., 2020) for both Hindi and Telugu. We observe that average sentence length are 18.95, 18.10 11.94 and 20.14 for collected (this work), created (this work), samanatar(Ramesh et al., 2021) and sizable monolingual corpora respectively for Hindi and 12.44, 12.24, 8.51 and 11.72 for Telugu. The difference is minimal between developed corpora (this work) and monolingual corpora for both languages while recently released samanatar(Ramesh et al., 2021) and monolingual corpora shows difference of average 9 and 3 tokens for Hindi and Telugu respectively. Similar pattern can be observe on character based average sentence length, here we also observed average 41 and 23 character difference between samanatar and monolingual corpora for Hindi and Telugu.

Table-5 shows variations in TTR across corpora. Since, TTR is known to depend on the length of the analysed text and hence the comparison makes sense where the size of corpora is same. Therefore we randomly picked the same number of sentences from large monolingual corpora to match the size of parallel corpora for fair comparative analysis. High TTR indicates that the observed text is more diverse and rich in-terms of vocabulary. We can see from Table-5 that TTR for created corpora stands high for both Hindi and Telugu compared to samanatar and stands close to the same sized monolingual corpora. To nullify the impact of domain, we calculated TTR on collected general domain corpora which also shows similar trends.

The analysis on average sentence length and TTR indicates that created and collected corpora (this work) show representativeness and diversity that one can observe in natural language text for both Hindi and Telugu.

Domain	Trained on	BLEU (Hi-Te)↑	TER (Hi-Te)↓	BLEU (Te-Hi)↑	TER (Te-Hi)↓
Chemistry (500)	Collected (165K)	5.83	0.81	7.78	0.78
	Samanatar (2425K)	6.56	0.74	9.23	0.77
	Created (535K)	**16.33**	**0.65**	**20.47**	**0.62**
Computer Science(500)	Collected (165K)	4.67	0.99	11.8	0.79
	Samanatar (2425K)	10.75	0.9	16.43	0.76
	Created (535K)	**16.06**	**0.86**	**22.73**	**0.70**
Law (1000)	Collected (165K)	9.65	0.98	8.95	0.83
	Samanatar (2425K)	10.47	0.9	16.43	0.72
	Created (535K)	**14.85**	**0.87**	**28.94**	**0.63**
General (1000)	Collected (165K)	9.2	0.85	12.33	0.79
	Samanatar (2425K)	10.47	0.80	16.17	0.71
	Created (535K)	**30.07**	**0.6**	**31.46**	**0.60**
Flores101 (1012)	Collected (165K)	6.18	0.89	8.56	0.80
	Samanatar (2425K)	6.64	0.84	10.08	0.77
	Created (535K)	**13.99**	**0.72**	**19.02**	**0.70**

Table 6: Bleu and TER scores for Hindi (Hi) - Telugu (Te) baseline MT systems (both direction) and comparison of MT systems trained using different available corpora. Here brackets value indicated size of parallel corpora (i.e Collected (165K) indicated that 165K sentences of collected corpora, chemistry(500) indicates validation data includes 500 sentences)

6. Baseline Systems

We trained baseline machine translation models using created parallel corpora with state-of-the-art neural machine translation methods as a baseline and tested on different domains for both Telugu-Hindi and Hindi-Telugu directions. We also train translation systems on different available training corpora for comparative analysis. Following subsections give details on pre-processing and exact training configuration used for training the neural machine translation engines.

6.1. Data Pre-Processing

For data preprocessing, we used IndicNLP Tool[14] with in-house tokenizer to tokenize and clean both Hindi and Telugu corpora (train, validation and test corpora) as a first step. Following subsections explain other pre-processing steps.

6.2. Morph + BPE Segmentation

From token/type ratio, Telugu is morphologically richer compared to Hindi from Table-1. Translating from morphologically-rich agglutinative languages is more difficult due to their complex morphology and large vocabulary(Mujadia and Sharma, 2021). We address this issue with a segmentation method which is based on morphology and BPE segmentation(Sennrich et al., 2016) as a pre-processing step as prescribed in (Mujadia and Sharma, 2020). We utilised unsupervised Morfessor (Virpioja et al., 2013) by training it on monolingual data for Hindi and Telugu. We then applied this trained Morfessor model on our corpora (train, test, validation) to get meaningful morpheme segmented sub-tokens for each word in a sentence. Subsequently, we applied the subword segmentation on top of the morph segmentation.

[14]http://anoopkunchukuttan.github.io/indic nlp library/

6.3. Training Configuration

Throughout all experiments, we used Transformer sequence to sequence architecture with the following configuration for constrained and unconstrained experiments. For these experiments, we used shared vocab across training and used Opennmt-py(Klein et al., 2020) toolkit with following configuration.

- Morph + BPE based subword segmentation, Embedding size : 512 Transformer for encoder and decoder, RNN_size 512, heads 8 encoder - decoder layers : 6, label smoothing : 1.0, dropout : 0.30, Optimizer : Adam, Beam size : 4 (train) and 10 (test), training steps : 20K

The results are discussed in following Result section.

7. Results

Table-6 shows performance of trained systems on different training corpora in terms of BLEU (Papineni et al., 2002) and TER scores for Hindi-Telugu and Telugu-Hindi respectively on the validation and Test data. We used SACREBLEU(Post, 2018) to calculate BLEU and TER scores. As described in section-4, the technical domain test sets are further divided into chemistry and computer science as sub-domains along with Law and General domains. We Further evaluated trained models on recently released The FLORES-101 Evaluation Benchmarks(Goyal et al., 2021). We achieved baseline BLEU scores as 16.33, 16.06, 14.85, 30.07 and 13.99 for respectively chemistry, computer science, Law , General and on Flores sets respectively

for Hindi-Telugu. Similarly, The baseline BLEU scores are 16.33, 16.06, 14.85, 30.07 and 13.99 respectively for chemistry, computer science, Law , General and on Flores sets respectively for Telugu-Hindi. TER scores show similar patterns across domains (Table-6).

Table-6 also shows BLEU, TER comparison of different translation models trained on different corpora and these parallel corpora vary in size and methods by which they developed as discussed in section-3. Here we find that trained models with 535K created data (this work) show best and state-of-the-art results.

8. Future work and Conclusion

We present a new 535K Hindi-Telugu parallel corpus of different technical domains such as Natural Science, Computer Science, Law and Healthcare along with the General domain. We also compiled, cleaned, reviewed 165K of Hindi-Telugu parallel corpora from different sources. We presented different methods for parallel corpora creation that we followed in this work. Particularly we observed that over 30% of translations were taken as it is (without single edit) by the translators and we recommend that Iterative Back-translation driven Post-Editing can be used for similar parallel corpora creation work. We present the state-of-the-art baselines and models for Hindi-Telugu and Telugu-Hindi across the domains. The results suggest that carefully created and curated parallel corpora boost the translation performance even with the lower parallel corpora size. The corpora and baseline models will be available under a Creative Commons Licence. We carried out average sentence length and TTR based analysis to assess the quality of parallel corpora. We find that our created parallel corpora matches the numbers that one observes in the natural text. In future, we plan to enhance the Hindi-Telugu corpus using described parallel corpora creation methods and plan to study and investigate domain adaptation for Hindi-Telugu domain dependent machine translation.

Acknowledgement

We thank the reviewers for their insightful comments. We thank Radhika Mamidi, Pruthwik Mishra for helping on various corpora aspects. This work was supported by the Ministry of Electronics and Information Technology, Government of India.

9. Bibliographical References

Ahsan, A., Mujadia, V., and Sharma, D. M. (2021). Assessing post-editing effort in the english-hindi direction. *arXiv preprint arXiv:2112.09841*.

Annamalai, E. (2001). *Managing multilingualism in India: Political and linguistic manifestations*. SAGE Publications Pvt. Limited.

Artetxe, M. and Schwenk, H. (2019). Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Barrault, L., Bojar, O., Bougares, F., Chatterjee, R., Costa-jussà, M. R., Federmann, C., Fishel, M., Fraser, A., Graham, Y., Guzman, P., et al. (2020). Proceedings of the fifth conference on machine translation. In *Proceedings of the Fifth Conference on Machine Translation*.

Dakwale, P. and Monz, C. (2019). Improving neural machine translation using noisy parallel data through distillation. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, pages 118–127.

Goyal, N., Gao, C., Chaudhary, V., Chen, P.-J., Wenzek, G., Ju, D., Krishnan, S., Ranzato, M., Guzman, F., and Fan, A. (2021). The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *arXiv preprint arXiv:2106.03193*.

Haddow, B. and Kirefu, F. (2020). Pmindia–a collection of parallel corpora of languages of india. *arXiv preprint arXiv:2001.09907*.

Hoang, V. C. D., Koehn, P., Haffari, G., and Cohn, T. (2018). Iterative back-translation for neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 18–24.

Jha, G. N. (2010). The tdil program and the indian langauge corpora intitiative (ilci). In *LREC*.

Kakwani, D., Kunchukuttan, A., Golla, S., Gokul, N., Bhattacharyya, A., Khapra, M. M., and Kumar, P. (2020). inlpsuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for indian languages. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 4948–4961.

Klein, G., Hernandez, F., Nguyen, V., and Senellart, J. (2020). The opennmt neural machine translation toolkit: 2020 edition. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (AMTA 2020)*, pages 102–109.

Kunchukuttan, A., Mehta, P., and Bhattacharyya, P. (2017). The iit bombay english-hindi parallel corpus. *arXiv preprint arXiv:1710.02855*.

Litvinova, T., Seredin, P., Litvinova, O., and Zagorovskaya, O. (2017). Differences in type-token ratio and part-of-speech frequencies in male and female russian written texts. In *Proceedings of the Workshop on Stylistic Variation*, pages 69–73.

Mayer, T. and Cysouw, M. (2014). Creating a massively parallel bible corpus. *Oceania*, 135(273):40.

Muischnek, K. and Müürisep, K. (2018). Impact of corpora quality on neural machine translation. In *Human Language Technologies–The Baltic Perspective: Proceedings of the Eighth International Conference Baltic HLT 2018*, volume 307, page 126. IOS Press.

Mujadia, V. and Sharma, D. M. (2020). Nmt based similar language translation for hindi-marathi. In *Proceedings of the Fifth Conference on Machine Translation*, pages 414–417.

Mujadia, V. and Sharma, D. M. (2021). Low resource similar language neural machine translation for tamil-telugu. In *Proceedings of the Sixth Conference on Machine Translation*, pages 288–291.

Nakazawa, T., Nakayama, H., Ding, C., Dabre, R., Higashiyama, S., Mino, H., Goto, I., Pa Pa, W., Kunchukuttan, A., Parida, S., Bojar, O., Chu, C., Eriguchi, A., Abe, K., Oda, Y., and Kurohashi, S. (2021). Overview of the 8th workshop on Asian translation. In *Proceedings of the 8th Workshop on Asian Translation (WAT2021)*, pages 1–45, Online, August. Association for Computational Linguistics.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Parida, S., Dash, S. R., Bojar, O., Motlicek, P., Pattnaik, P., and Mallick, D. K. (2020). OdiEnCorp 2.0: Odia-English parallel corpus for machine translation. In *Proceedings of the WILDRE5– 5th Workshop on Indian Language Data: Resources and Evaluation*, pages 14–19, Marseille, France, May. European Language Resources Association (ELRA).

Philip, J., Siripragada, S., Kumar, U., Namboodiri, V., and Jawahar, C. (2019). Cvit's submissions to wat-2019. In *Proceedings of the 6th Workshop on Asian Translation*, pages 131–136.

Post, M. (2018). A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium, October. Association for Computational Linguistics.

Ramesh, G., Doddapaneni, S., Bheemaraj, A., Jobanputra, M., AK, R., Sharma, A., Sahoo, S., Diddee, H., Kakwani, D., Kumar, N., et al. (2021). Samanantar: The largest publicly available parallel corpora collection for 11 indic languages. *arXiv preprint arXiv:2104.05596*.

Resnik, P., Olsen, M. B., and Diab, M. (1999). The bible as a parallel corpus: Annotating the 'book of 2000 tongues'. *Computers and the Humanities*, 33(1):129–153.

Schwenk, H., Chaudhary, V., Sun, S., Gong, H., and Guzmán, F. (2019). Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia. *arXiv preprint arXiv:1907.05791*.

Sennrich, R., Haddow, B., and Birch, A. (2015). Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*.

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725.

Siripragada, S., Philip, J., Namboodiri, V. P., and Jawahar, C. (2020). A multilingual parallel corpora collection effort for indian languages. *arXiv preprint arXiv:2007.07691*.

textinspector. (2022). Statistics and readability scores — text inspector. `https://textinspector.com/help/statistics-readability/`. (Accessed on 01/16/2022).

Tiedemann, J. (2012). Parallel data, tools and interfaces in opus. In *Lrec*, volume 2012, pages 2214–2218. Citeseer.

Vasanta, D., Suvarna, A., Sireesha, J., and Raju, S. B. (2010). Language choice and language use patterns among telugu-hindi/urdu-english speakers in hyderabad, india. In *Proceedings of the International Conference on Language, Society and Culture in Asian Contexts*, pages 57–67. Mahasarakam University.

Virpioja, S., Smit, P., Grönroos, S.-A., Kurimo, M., et al. (2013). Morfessor 2.0: Python implementation and extensions for morfessor baseline.

Wang, R., Tan, X., Luo, R., Qin, T., and Liu, T.-Y. (2021). A survey on low-resource neural machine translation. *arXiv preprint arXiv:2107.04239*.