EM Corpus: a comparable corpus for a less-resourced language pair Manipuri-English

Rudali Huidrom **Yves Lepage** **Khogendra Khomdram**
IPS, Waseda University The Sangai Express
Kitakyushu, Japan Manipur, India
{rudali.huidrom@ruri.,yves.lepage@}waseda.jp khogen.kh@gmail.com

Abstract

In this paper, we introduce a sentence-level comparable text corpus crawled and created for the less-resourced language pair, Manipuri (mni) and English (eng). Our monolingual corpora comprise 1.88 million Manipuri sentences and 1.45 million English sentences, and our parallel corpus comprises 124,975 Manipuri-English sentence pairs. These data were crawled and collected over a year from August 2020 to March 2021 from a local newspaper website called 'The Sangai Express.' The resources reported in this paper are made available to help the low-resourced languages community for MT/NLP tasks[1].

1 Introduction

The web is immense, free, and available to all (Kilgarriff and Grefenstette, 2003). Several studies have proposed the use of the web as a corpus for teaching and research (Rundell, 2000; Robb, 2003; Fletcher, 2001, 2004; Kilgarriff and Grefenstette, 2003). Languages such as English and Chinese are widely published and are well-equipped with resources and tools. Availability of data for low-resource languages on the web is increasing day by day (Schryver, 2002) contributing hugely to bridge the gap between high-resource and low-resource languages. In addition, it is important to mention the language in discussion states #BenderRule (Bender, 2019) to minimize the existing divide of languages in NLP. In this paper, our work is to equip a less-resourced language pair, Manipuri–English with resources.

Our objective is to increase the size of available data for Manipuri–English language pairs. Our goal is to build a sentence-level comparable corpus for Manipuri–English[2] from a newspaper website called "*The Sangai Express*".[3] We introduce the creation of a comparable corpus named "***Ema-lon Manipuri Corpus***", (translation: our mother tongue Manipuri Corpus) abbreviated as the **EM Corpus**, of the low-resourced language pair, Manipuri–English. We report on the method for creating the comparable corpus. We also tried to extract parallel corpus from our comparable data. Additionally, we provide the table that maps the corresponding glyph points to its Unicode codepoints for Manipuri.

The structure of the paper is as follows. Section 2 describes previous work. Sections 3 and 4 describes the characteristics of the language and the data. Section 5 presents the methodological aspects. Section 6 provides the details of the experiment and its analysis. Section 7 concludes and proposes future directions.

2 Related Work

Several works on the web as a corpus (Rundell, 2000; Robb, 2003; Fletcher, 2001, 2004; Kilgarriff and Grefenstette, 2003) for many languages have been reported from the past decades (Schryver, 2002). The use of web-based Manipuri corpus has been reported by Singh and Bandyopadhyay (2010) for the identification of reduplicated multi-word expression (MWE) and multi-word named entity recognition (NER). PMIndia is yet another crawled data set of 13 Indian languages with English. This data set includes Manipuri–English language pair data. IndicCorp, sourced from news crawls, is a large monolingual corpus of 11 Indian languages from two different language families (Indo-Aryan branch and Dravidian) (Kakwani et al., 2020). Some of the familiar datasets obtained from web crawls are The Leipzig corpus (Goldhahn et al., 2012), CommonCrawl, and The OSCAR project (Ortiz Suárez et al., 2019), none of which contains the Manipuri–English language pair in it.

[1]The corpus is available from http://lepage-lab.ips.waseda.ac.jp/en/projects/meiteilon-manipuri-language-resources/

[2]The codes from ISO 639-2 for these languages are as follows: Manipuri (mni) and English (eng)

[3]https://www.thesangaiexpress.com/

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 60–67
July 5–10, 2020.
https://doi.org/10.26615/978-954-452-056-4_008

Fung and Cheung (2004) analyses different types of bilingual corpora, ranging from parallel, noisy parallel, comparable, very-non-parallel corpora. Types of comparable corpus includes: (i) non-sentence-aligned, non-translated bilingual documents that are topic-aligned. Example, newspaper articles that are published on the same date in different languages, and (ii) non-aligned sentences that are mostly bilingual translations of the same document. Our work is close to the former.

3 Manipuri (Meiteilon)

Manipuri, locally known as Meiteilon, is an Indian language from the Sino-Tibetan language family. It is highly agglutinative in nature. Manipuri follows the SOV (Subject-Object-Verb) syntax structure. As the predominant language of the Indian state Manipur, Manipuri has about two million native speakers. As a language classified as 'vulnerable language' by UNESCO (Moseley and Nicolas, 2010), it is one of the two Indian languages listed in the 8th Schedule of the Indian Constitution as endangered.

Manipuri has two writing systems: Eastern Nagari Script (also known as the Bengali Script) and Meitei Mayek. We use Manipuri written in Eastern Nagari Script for all of our works. Again, Manipuri is a low-resourced language that has not been explored much in computational linguistics. One of the reasons being the limited amount of available resources. In this paper, we aim to bridge this gap by sharing our resources publicly.

4 Ema-lon Manipuri Corpus (EM Corpus)

The amount of resources for Manipuri–English language pair is limited for performing Machine Translation (MT)/Natural Language Processing (NLP) tasks (Huidrom and Lepage, 2020). For example, there are 41,669 sentences monolingually and 7,419 parallel sentences with English in the open-sourced monolingual and parallel data from the pmindia dataset[4] (Haddow and Kirefu, 2020). Other sources include TDIL-DC[5], where the data is available upon an undertaking agreement. A standard site such as OPUS[6] (Tiedemann, 2012) is limited in the coverage of low-resource Asian and South Asian Languages including, Manipuri. This motivated us to create our comparable corpus.

EM Corpus is built for Manipuri–English language pair. This corpus is created by collecting news articles daily from a newspaper website known as "The Sangai Express," which is available in both languages. An average of 14,000 sentences is crawled for this language pair daily. The reported data is being collected from August 2020 to March 2021, as shown in Figure 1. The domain of the EM Corpus includes general articles, news on state, national and international affairs, sports and entertainment news, and the editorial. The English articles are topic-aligned with the Manipuri articles, however, they are not the exact bilingual translation of each other but rather the summary or the gist of the Manipuri news.

The monolingual datasets contain 1.88 million Manipuri sentences and 1.45 million English sentences and the parallel corpora contain 124,975 sentences. The number of words per sentence in Manipuri and English is reported to be 17 and 23 monolingually and, 21 and 26 in parallel. It is to note that the number of word types in each language reflects the number of sentences and the structure of the language: it is natural that the more the sentence pairs, the higher the number of word types as reported in Table 1. It is reported that the average word length of Manipuri is more than that of English monolingually and in parallel, however, the average word types length is the same for both the languages.

5 Methodology

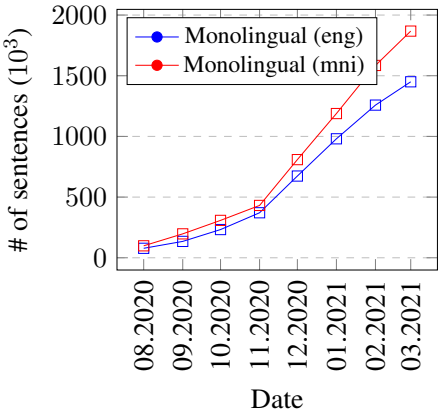
In this section, we introduce the creation of EM corpus and extraction of parallel corpus from the comparable data.

- **Crawling and Extraction.** The news articles which are available in both languages were crawled and extracted on a daily basis. The news updates in 'The Sangai Express' are available in a section-based format and, each section contains articles in an infinite scroll format. The request for the lists of URLs follows a simple form, and so we source our data with a web scraper for each language which we built. Since the class in the HTML of each article corresponds to each other, document alignment was straightforward. Figure 1 shows the statistics of the data collection obtained per month from August 2020

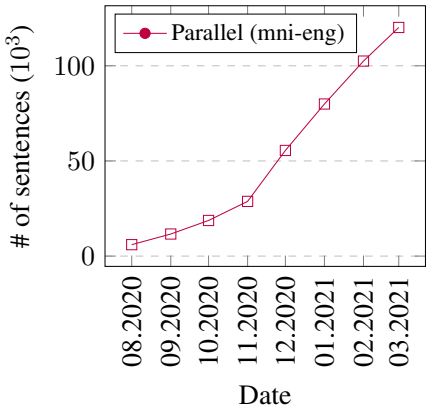
[4] http://data.statmt.org/pmindia/
[5] https://www.tdil-dc.in/
[6] http://opus.nlpl.eu/index.php

Data set	Language pair	sentences	words	words / sent.	word length	word types	word types length
Monolingual	Manipuri	1,880,035	31,124,061	17	6	95,380	8
	English	1,450,053	33,667,493	23	5	108,812	8
Parallel	Manipuri	124,975	2,589,109	21	6	74,516	8
	English		3,289,671	26	5	64,501	8

Table 1: Detailed statistics on the EM Corpus.



(a) Statistics of the monolingual data collected from August 2020 to March 2021.



(b) Statistics of the parallel data collected from August 2020 to March 2021.

Figure 1: Statistics of the monolingual and parallel data from the EM Corpus.

to March 2021. To extract the content of the articles from the HTML, we use BeautifulSoup[7] (Richardson, 2007) which is a rich python library for parsing HTML/XML documents, which on inspection performs well in extracting the body of the articles. Additionally, we use cronTab (Reznick, 1993) to automate our news crawl.

[7] https://www.crummy.com/software/BeautifulSoup

- **Text Processing.** The data crawled for Manipuri encoded in nature as the website uses its custom web font file for Manipuri. To obtain the correct text for Manipuri, we map the glyph points to the exact Unicode codepoints. We identified the corresponding matches in this process manually. After obtaining the precise format of the font for Manipuri, we split the articles into sentences for sentence alignment using Moses splitter (Koehn et al., 2007) by taking into account about the sentence delimiter, punctuation, and list items of Manipuri in Eastern Nagari script (Bengali script).

- **Sentence Alignment.** We use Hunalign (Varga et al., 2005), a sentence aligner that aligns bilingual text based on the heuristics of sentence-length information and a bilingual dictionary (if available). It is to be noted that Hunalign does not deal with changes of sentence order like most sentence aligners. Due to the absence of the dictionary for Manipuri, we use the automatic dictionary built based on the alignment. We retain 1-1 alignments obtained from filtering sentences with a threshold that discards score lower than 0.3.

6 Experiment and result

The paper discusses the creation of a comparable corpus from scratch and extracts parallel sentences from the comparable data. As mentioned earlier, the nature of the sentences in the two documents is such that the English news provides a summary of the Manipuri news. Although our documents are topic-aligned, the sentences are not present in a one-to-one correspondence. This explains the difference in the number of sentences monolingually.

Further, we use Hunalign to extract the parallel sentences from EM corpus. We wanted to study the relevance of the parallel sentences extracted

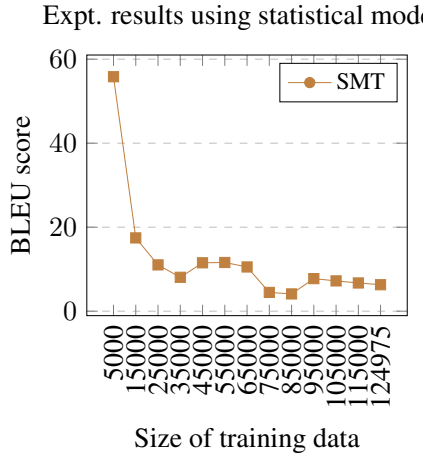


Figure 2: The results are from the SMT experiments (square and brown).

from the comparable data of the aforementioned nature. We designed a simple experiment on statistical machine translation. The models were trained on 5,000 sentences from PMIndia dataset and incremented by 10,000 parallel sentences from our EM corpus in each iteration. Our validation and test data are from the PMIndia (Haddow and Kirefu, 2020) dataset whose domain is the official documents from the Prime Minister Office of India. Figure 2 shows the result of the experiment.

As we progress with the adding of more training data, we observe a decrease in the BLEU score which is expected. It is to be noted that the decrease is not linear in nature. The data that we add other than the baseline are obtained from the news crawls which are not standardised translated data. Although, the sentences are aligned, the parallel sentences are not exact translations of one another, instead comparable. The sudden increase of the BLEU score could be the result of seeing similar sentences crawled from the news articles related to the Prime Minister Office while training.

7 Conclusion

This work provided an insight into corpus creation for Manipuri–English language pair. Firstly, we studied the creation of the comparable corpus, EM corpus for the low-resourced language pair Manipuri–English. Secondly, we discussed the nature of the comparable corpus for Manipuri-English language pair. We report the statistics on these data which is built by collecting from the web for over a year, from August 2020 to March 2021. The appendices provide information on mapping the glyph points to the Unicode codepoints for Manipuri. This is a necessary step due to the nature of the news articles that were crawled. The Sangai Express uses its custom web font file. This table can be referred if you are crawling independently to build your own corpus.

In the future, we would like to inspect the possibility of increasing the size of data by using data-augmentation techniques. Also, we welcome everyone in improving and contributing to these resources.

References

Emily Bender. 2019. The benderrule: On naming the languages we study and why it matters. *The Gradient*.

William H. Fletcher. 2001. Concordancing the web with KWiCFinder. In *In Third North American Symposium on Corpus Linguistics and Language Teaching*, pages 1–16, Boston, MA.

William H. Fletcher. 2004. Making the web more useful as a source for linguistic corpora. *Applied Corpus Linguistics: A multi-dimensional perspective*, 52:191–205.

Pascale Fung and Percy Cheung. 2004. Mining very-non-parallel corpora: Parallel sentence and lexicon extraction via bootstrapping and E. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 57–63, Barcelona, Spain. Association for Computational Linguistics.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building large monolingual dictionaries at the Leipzig corpora collection: From 100 to 200 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 759–765, Istanbul, Turkey. European Language Resources Association (ELRA).

Barry Haddow and Faheem Kirefu. 2020. Pmindia - a collection of parallel corpora of languages of india. *ArXiv*, abs/2001.09907.

Rudali Huidrom and Yves Lepage. 2020. Zero-shot translation among Indian languages. In *Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages*, pages 47–54, Suzhou, China. Association for Computational Linguistics.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: 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, Online. Association for Computational Linguistics.

Adam Kilgarriff and Gregory Grefenstette. 2003. Introduction to the special issue on the web as corpus. *Computational linguistics*, 29(3):333–347.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Christopher Moseley and Alexandre Nicolas. 2010. *Atlas of the world's languages in danger*, 3 edition. UNESCO, France.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. In *Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019*, pages 9 – 16, Mannheim. Leibniz-Institut für Deutsche Sprache.

Larry Reznick. 1993. Using cron and crontab. *Sys Admin*, 2(4):29–32.

Leonard Richardson. 2007. Beautiful soup documentation. *April*.

Thomas N. Robb. 2003. Google as a corpus tool? *ETJ Journal*, 4(1):20–21.

Michael Rundell. 2000. The biggest corpus of all. *Humanising language teaching*, 2(3).

Gilles-Maurice de Schryver. 2002. Web for/as Corpus: A Perspective for the African Languages. *Nordic Journal of African Studies*, 11(2):266–282.

Thoudam Doren Singh and Sivaji Bandyopadhyay. 2010. Web based Manipuri corpus for multiword NER and reduplicated MWEs identification using SVM. In *Proceedings of the 1st Workshop on South and Southeast Asian Natural Language Processing*, pages 35–42, Beijing, China. Coling 2010 Organizing Committee.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

D. Varga, L. Németh, P. Halácsy, A. Kornai, V. Trón, and V. Nagy. 2005. Parallel corpora for medium density languages. In *In Proceedings of the RANLP 2005*, pages 590–596, Prague, Czech Republic.