EnAsCorp1.0: English-Assamese Corpus

Sahinur Rahman Laskar, Abdullah Faiz Ur Rahman Khilji, Partha Pakray, Sivaji Bandyopadhyay
Department of Computer Science and Engineering
National Institute of Technology Silchar
Assam, India
{sahinur_rs, abdullah_ug, partha}@cse.nits.ac.in,
sivaji.cse.ju@gmail.com

Abstract

The corpus preparation is one of the important challenging task for the domain of machine translation especially in low resource language scenarios. Country like India where multiple languages exists, machine translation attempts to minimize the communication gap among people with different linguistic backgrounds. Although Google Translation covers automatic translation of various languages all over the world but it lags in some languages including Assamese. In this paper, we have developed EnAsCorp1.0, corpus of English-Assamese low resource pair where parallel and monolingual data are collected from various online sources. We have also implemented baseline systems with statistical machine translation and neural machine translation approaches for the same corpus.

1 Introduction

Assamese (also called as Axomiya) is the predominant language of the Indian state of Assam. It is one of the 23 official languages (22 Indian languages and additional language English) recognized by the Republic of India[1]. It is the most widely used language of the entire north-eastern region (also known as the seven sister states) of India. Assamese is spoken by about 14 million speakers (Barman et al., 2014), they are known as Assamese or Axomiya/Asomiya people. It belongs to the Indo-Aryan language family, and its script is evolved from the Gupta script (Dutta, 2019), also known as Assamese-Bengali script (Mahanta, 2012). In today's digital world, there is a demand for automatic translation of English to Assamese and vice-versa. In a multilingual country like India, machine translation (MT) plays a vital role to minimize the language barrier via automatic translation among human spoken languages (also known as natural languages). For all official modes of communication, Hindi and English languages are used by the government of India. Therefore, to facilitate greater access among the local community, it is necessary to translate the official documents like orders, notices, messages released by the government of India into the regional language of Assam. Moreover, English is a high resource and widely accepted language all over the globe. To establish a better communication in the native languages at the national or international level, MT system for the English-Assamese pair is very much essential. But due to the non availability of a suitable corpus, the MT system for this language pair still remains in beginning stage (Barman et al., 2014; Baruah et al., 2014).

In MT, based on the availability of corpus resource, there are two categories of natural languages: high and low resource languages. High resource languages are those languages which are resource-rich languages like English, German, French, and Hindi. On the other hand, low resource languages are resource-poor like many Indian languages especially found in the north-eastern region of India like Assamese, Boro, Khasi, Manipuri, Kokborok, and Mizo. Data scarcity is one of the major problems in MT for low resource language scenarios. For English-Assamese, there is a lack of standard parallel data and monolingual Assamese data. Because of this reason, English-Assamese can be categorized under low resource pair. The word order and script of both languages are completely different. Unlike English, Assamese follows the subject-object-verb (SOV) word order. Figure 1 depicts an example sentence, where an Assamese sentence is shown with its English translation and transliteration. Moreover, Assamese is known as mor-

[1] https://www.mhrd.gov.in/sites/upload_files/mhrd/files/upload_document/languagebr.pdf

Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages, pages 62–68

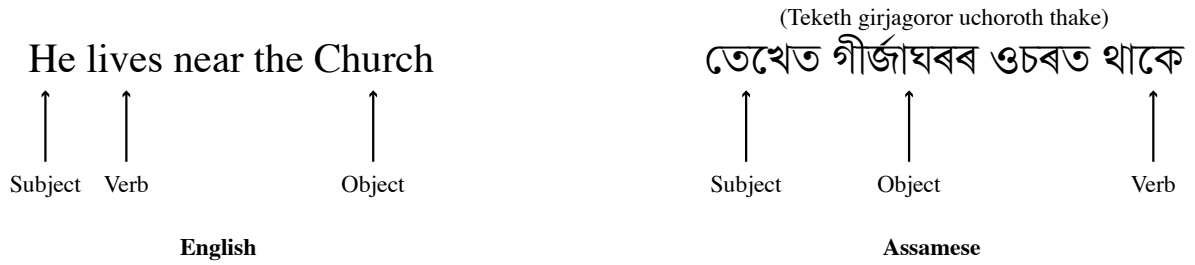


Figure 1: Example of Parallel English-Assamese Sentence.

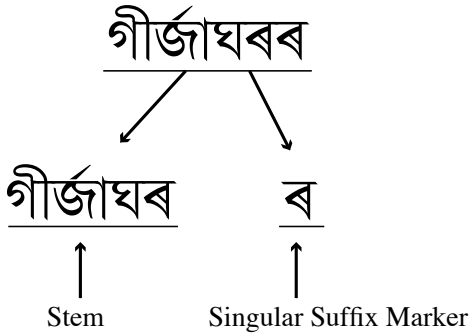


Figure 2: Example of an Assamese Morpheme

phological rich language unlike English. Figure 2 presents an example of Assamese morpheme. There are total of 52 letters in the Assamese alphabet comprising of 41 consonants and 11 vowels. Assamese script and numerals are very different from English and identical to the Bengali language except for the two letters, ৰ (ro) and ৱ (vo) (Mahanta, 2012) In this paper, our focus is to create an English-Assamese corpus that we hope that it will be of benefit to the research community.

2 Corpus

2.1 Role of Corpus in MT

MT has evolved from rule-based methods to a corpus-based approach, where any source language can be translated to any target language. Therefore, MT reached a language-independent stage. Under the corpus-based approach, neural machine translation (NMT) achieves a state-of-the-art over the contemporary statistical machine translation (SMT). But both the approaches need sufficient training corpus which is a challenging issue, especially in the scenario of low resource languages. In MT, there are two types of datasets: parallel and monolingual corpora. The parallel corpus consists of aligned sentences, which are structured set of translated texts between source and target languages. It is required for both SMT and NMT based approaches (Koehn, 2010; Devlin et al., 2014). Such a parallel data is important for the model to learn word alignment between source and target sentences by estimating parameters. This is known as adequacy factor of a good translation. By increasing the amount of parallel corpus, the translation accuracy can be improved in both SMT and NMT. For English-Assamese, it is a challenging task to obtain a parallel corpus since there is lack of standard parallel data. Apart from the parallel corpus, there is a need for monolingual corpus as well, which is used to produce a fluent translation of the target sentence. The fluency is another factor for a good translation. Unlike parallel corpus, monolingual corpus for English and Assamese are available in various online sources, but since Assamese is a low resource language the digitized monolingual corpus is difficult to find. The standard English monolingual data are available from the shared task of MT[2], but there is lack of standard Assamese monolingual data. This is an another challenge to prepare an Assamese monolingual corpora from available online sources. The monolingual corpus plays a vital role in improving the translation quality for both SMT (Koehn, 2010) and NMT in low resource language scenarios (Sennrich et al., 2016).

2.2 Corpus Details

There are very limited available options for English-Assamese parallel data. However, several possible sources have been explored to collect parallel and monolingual data as discussed in Section 2.3 and depicted in Figure 3. The details of the data sources used are described below:

[2] http://www.statmt.org/wmt16/translation-task.html

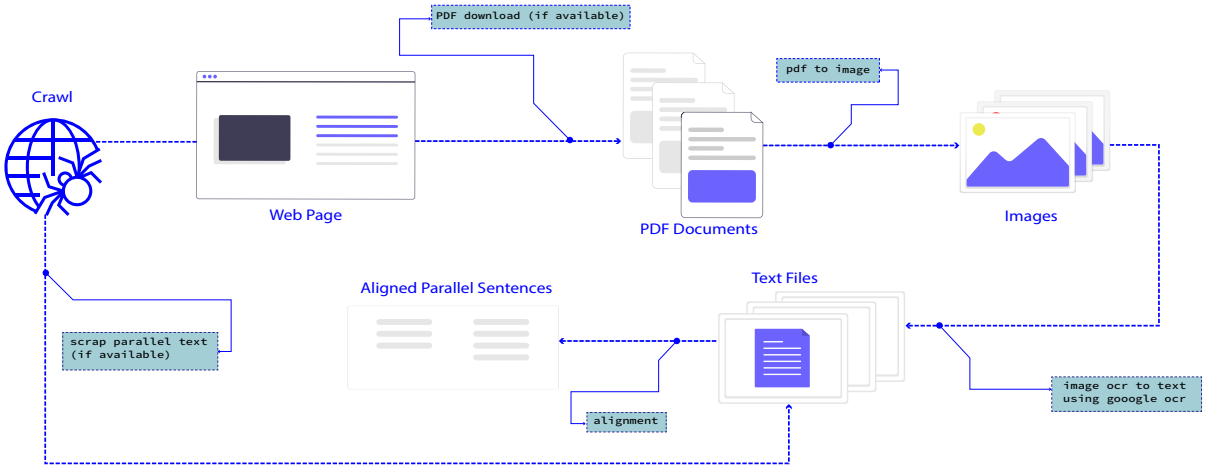


Figure 3: Data Acquisition

2.2.1 Parallel Corpus

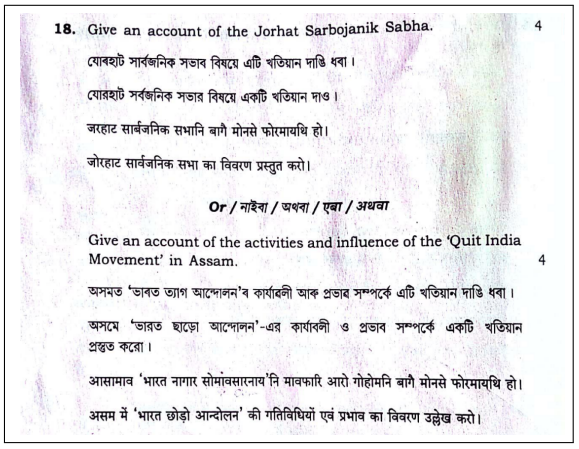
- **Bible:** The holy book Bible is freely and publicly available and the text is available in various languages including Assamese. From this, we have collected 24,199 parallel sentences using scraping technique.

- **Multilingual Online Dictionary:** The multilingual online Dictionary namely, Xobdo and Glosbe, where not only parallel words are found but also example parallel sentences are available. From Xobdo and Glosbe, after removing duplicates, we have collected 15,754 and 200,151 parallel sentences using scraping.

- **SEBA Multilingual Question Paper:** Board of secondary education, Assam (also known as SEBA) is the government recognized authority for conducting the school level examination in Assam. The multilingual SEBA examination question paper written in (English, Assamese, Hindi, and Bengali) contains exact parallel English-Assamese sentences as shown in Figure 4. From here, we have collected 1,000 parallel sentences using two steps. Firstly, we extracted the text from the portable document format (PDF) files of the question paper using Google's optical character recognition (OCR) and stored them into text files and then in the second step, we have manually collected the parallel sentences from the generated text files.

- **Learn-Assamese website:** This site is dedicated to teaching Assamese in terms of linguistic aspects like the alphabet, number, grammar, etc. From here, we have manually collected 188 parallel sentences.

- **PMIndia:** A publicly available parallel data[3](Haddow and Kirefu, 2020) includes 13 Indian languages (Hindi, Bengali, Gujarati, Marathi, Telugu, Tamil, Odia, Malayalam, Punjabi, Kannada, Manipuri, Assamese and Urdu) with English. From this, we have collected 9,732 parallel sentences.


18. Give an account of the Jorhat Sarbojanik Sabha. 4

যোৱহাট সার্বজনিক সভাৰ বিষয়ে এটি খতিয়ান দাঙি ধৰা।

যোৱহাট সৰ্বজনিক সভাৰ বিষয়ে একটি খতিয়ান দাও।

জৰহাট সার্বজনিক সভানি বাগৈ মোনসে ফোৰমাযথি হো।

জোৰহাট সার্বজনিক সভা কা বিবরণ প্রস্তুত করো।

Or / নাইবা / অথবা / এবা / अथवा

Give an account of the activities and influence of the 'Quit India Movement' in Assam. 4

অসমত 'ভাৰত ত্যাগ আন্দোলন'ৰ কার্যাবলী আৰু প্রভাৱ সম্পর্কে এটি খতিয়ান দাঙি ধৰা।

অসমে 'ভারত ছাড়ো আন্দোলন'-এর কার্যাবলী ও প্রভাব সম্পর্কে একটি খতিয়ান প্রস্তুত করো।

আসামাব 'ভারত নাগার সোমাংবসারনায'নি মাবফারি আরো গোহোমনি বাগৈ মোনসে ফোরমাযথি হো।

असम में 'भारत छोड़ो आन्दोलन' की गतिविधियों एवं प्रभाव का विवरण उल्लेख करो।

Figure 4: Example of a Parallel Sentence in SEBA Multilingual Question Paper.

2.2.2 Monolingual Corpus

As we have mentioned in Section 2.1 that standard monolingual data for English is available. Therefore, we have focused on the preparation of Assamese monolingual data only and the same

[3]http://data.statmt.org/pmindia/v1/parallel/

Corpus	Source	Sentences	Tokens	
			English	Assamese
Parallel	Bible	24,199	645,600	512,408
	Xobdo	15,754	209,546	146,316
	Glosbe	200,151	2,329,651	1,907,653
	SEBA Multilingual Question Paper	1,000	7,290	5,810
	Learn-Assamese	188	456	425
	PMIndia	9,732	192,530	164,520
	Total	251,024	3,385,073	2,737,132
Monolingual	Web Pages/Blogs/Holy Books	2,624,828	-	45,900,321

Table 1: Corpus Statistics

English	Assamese
Explain briefly in what circumstances was the "Gandhi-Irwin", pact signed ? Discuss its terms.	কি ৰাজনৈতিক পৰিস্থিতিত 'গান্ধী-আৱউইন' চুক্তি স্ৱাক্ষৰিত হৈছিল? চুক্তিৰ চৰ্তসমূহ আলোচনা কৰা।

Table 2: Example Sentence having Punctuation Marks

Type	Sentences	Tokens		Remark
		English	Assamese	
Train	203,315	2,414,172	1,986,270	
Validation	4,500	74,561	59,677	
Test	2,500	41,985	34,643	2,400 (In-domain) + 100 (Out-domain)

Table 3: Date Statistics for Train, Valid and Test Set

Corpus	English	Assamese	Remark
Parallel	In the beginning God created the heaven and the earth.	আদিতে ঈশ্ৱৰে আকাশ-মণ্ডল আৰু পৃথিৱী সৃষ্টি কৰিলে।	Bible
	The Year which was counted from the birth of Jesus Christ.	যীশুখ্ৰীষ্টৰ জন্মৰপৰা গণনা কৰা বছৰসংখ্যা ।	Xobdo
	Tourists get the birds eye view of the whole city from the top of the dome.	গম্বুজটোৰ শীৰ্ষস্থানৰ পৰা পৰ্যটক সকলে বিহঙ্গম দৃষ্টিৰে নগৰখন দেখা পায় ।	Glosbe
	What is called Shadow Cabinet?	ছায়া মন্ত্ৰীসভা কাক বোলে ?	SEBA Multilingual Question Paper
Monolingual	-	ভাৰতীয় ক্ৰিকেট দলৰ প্ৰাক্তন বলাৰ সুনীল যোশীক ভাৰতীয় ক্ৰিকেট দলৰ মুখ্য নিৰ্বাচক ৰূপে নিযুক্তি প্ৰদান কৰা হৈছে।	Web pages/Blogs
	-	ডাভোছত যোৱা মাহত অনুষ্ঠিত ৱৰ্ল্ড ইক'নমিক ফ'ৰামত এই বিষয়টোক লৈ বিশেষভাৱে সৰৱ হৈ পৰিছিল থুনবার্গ বিজ্ঞানীসকলৰ মতে,	Web pages/Blogs

Table 4: Example Sentences from Various Sources

has been collected from different online web pages/Blogs and Holy books namely, Bible, Quran and Gita. Additionally, we can use the Assamese data from the parallel corpus to increase the Assamese monolingual data.

2.2.3 Corpus Statistics

The overall corpus statistics are shown in Table 1. The total parallel and monolingual sentences 251,024 and 50,086. Basically, there are three types of dataset, training, validation and test set. The training set is required for learning the parameters during the training of a model. During training process, the validation dataset is used to check the performance of a model to select the optimum model generated. The test data is the unseen data which is used to verify the model performance once the training process is completed. For the MT system, the parallel data need to split into train, validation, and test data. Before the split, we have removed duplicate sentences and the count of total parallel sentences got reduced to 210,215. The statistics for train, validation and test data are presented in Table 3. In the test test, we have included 100 out of domain parallel sentences from different blogs/web pages. These 100 parallel sentences include simple sentences which are commonly used in day to day life.

2.3 Data Extraction Techniques

For scraping the data from various online sources, we have utilized Scrapy[4] an open source framework for web crawling. The `xpath` for each element is coded with a certain degree of generalization in order to replicate for multiple web pages. This helped us in crawling a large number of web pages and extracting useful information. In order to extract text from images as well as from the PDF files we employ the use of Google OCR[5] tools. The process is summarized in Figure 3.

As given in the Figure 3, we begin by feeding the URLs of the web pages. The PDF documents are downloaded and the Assamese raw text in the HTML files are extracted directly. According to the format of the data here, the textual data is obtained. If the data is in image or PDF format, it is treated as an image and the Google OCR engine is used to extract the text. The use of Google OCR helped us to extract the data from a wide range of sources without facing any font related issues. The conversion of PDF to images resolves the issues of different fonts, embedded texts or size. The obtained mono-lingual data is kept as is, but the parallel data is aligned by separating them into separate source and target files. This process of alignment and verification, took substantial human effort.

[4] https://scrapy.org/

[5] https://cloud.google.com/vision/

2.4 Data Analysis

Table 4 presents example parallel sentences from each individual sources. We have removed noise i.e. unwanted symbols, too many special characters (....., ####, $$$$), URLs, English text from Assamese text, blank lines, etc from the text. We did not remove punctuation marks and not performed lower-case conversion for the English sentences because if we do this then the contextual meaning of the sentence will be changed as shown in Table 2. Assamese language has single case letters unlike English and hence no case conversion is considered here. Moreover, we split large sentences (more than 120 words) at the end of sentence punctuation mark (।,?).

2.5 Domain Coverage

Our corpus EnAsCorp1.0 covers variety of domains namely, Holy books (Bible, Quran, Gita), daily usage, literature and general domains.

3 Baseline Systems

For the baseline systems, we have trained two models: phrase-based SMT (Koehn et al., 2003) and recurrent neural network (RNN) based NMT with attention (Bahdanau et al., 2015) to estimate benchmark translation accuracy for English to Assamese and Assamese to English translation respectively. We have used the developed dataset EnAsCorp1.0 and English monolingual data about 3 million sentences from WMT16[6].

3.1 SMT Setup

We have trained the phrase-based SMT using the Moses[7] (Koehn et al., 2007) toolkit. GIZA++ and IRSTLM (Federico et al., 2008) are used to generate phrase pairs and language model following default settings.

[6] http://www.statmt.org/wmt16/translation-task.html

[7] http://www.statmt.org/moses/

Translation	System	BLEU
English to Assamese	SMT	3.43
	NMT	5.55
Assamese to English	SMT	4.54
	NMT	7.72

Table 5: Results of Baseline Systems

3.2 NMT Setup

We have trained the NMT using OpenNMT-py toolkit[8] (Klein et al., 2017). The encoder-decoded architecture is of a two-layer network consists of long short term memory (LSTM) cell with 500 nodes in each layer and attention mechanism (Bahdanau et al., 2015) are used. The Adam optimizer with default learning rate and drop-outs 0.3 are used. We have employed pre-trained word vectors using Glove[9] (Pennington et al., 2014) by utilizing the monolingual data to capture semantic word relationships.

3.3 Results

To evaluate the performance of the baseline systems, we have used bilingual evaluation understudy (BLEU) (Papineni et al., 2002) automatic evaluation metric. Table 5 presents the results of both the systems. We have considered the average BLEU score of uni-gram, bi-gram, and tri-gram since the BLEU score tends to diminish after crossing the tri-gram score.

4 Conclusion and Future Work

In this paper, we have presented EnAsCorp1.0 for various tasks of NLP, mainly MT. We have developed both parallel and monolingual corpus for low resource English-Assamese pair. The dataset will be available here: https://github.com/cnlp-nits/EnAsCorp1.0. We have evaluated baseline systems using SMT and NMT. It is observed that the performance accuracy of NMT is better than SMT. In the future, we will increase the size of the corpus and develop MT system by utilizing advanced deep learning techniques for the same pair. The proposed dataset will be available for the benefit of the research community.

[8] https://github.com/OpenNMT/OpenNMT-py
[9] https://github.com/stanfordnlp/GloVe

Acknowledgement

We would like to thank Center for Natural Language Processing (CNLP) and Department of Computer Science and Engineering at National Institute of Technology, Silchar, India for providing the requisite support and infrastructure to execute this work.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Anup Barman, Jumi Sarmah, and Shikhar Sarma. 2014. Assamese WordNet based quality enhancement of bilingual machine translation system. In *Proceedings of the Seventh Global Wordnet Conference*, pages 256–261, Tartu, Estonia. University of Tartu Press.

Kalyanee Kanchan Baruah, Pranjal Das, Abdul Hannan, and Shikhar Kr Sarma. 2014. Assamese-english bilingual machine translation. *International Journal on Natural Language Computing (IJNLC)*, 3.

Jacob Devlin, Rabih Zbib, Zhongqiang Huang, Thomas Lamar, Richard Schwartz, and John Makhoul. 2014. Fast and robust neural network joint models for statistical machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1370–1380, Baltimore, Maryland. Association for Computational Linguistics.

Hemanga Dutta. 2019. *Assamese Orthography: An Introduction and Some Applications for Literacy Development*, pages 181–194. Springer International Publishing, Cham.

Marcello Federico, Nicola Bertoldi, and Mauro Cettolo. 2008. Irstlm: an open source toolkit for handling large scale language models. In *INTERSPEECH*, pages 1618–1621. ISCA.

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

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Philipp Koehn. 2010. *Statistical Machine Translation*, 1st edition. Cambridge University Press, USA.

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.

Philipp Koehn, Franz J. Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 127–133.

Shakuntala Mahanta. 2012. Assamese. *Journal of the International Phonetic Association*, 42(2):217–224.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.