Shahjalal University of Science and Technology
Department of Computer Science and Engineering

Available Corpora of Low-resource Bangla Language for Machine Translation

MASUM AHMED

Reg. No.: 2016331028

4^{th} year, 1^{st} Semester

MD SHAMIHUL ISLAM KHAN

Reg. No.: 2016331078

4^{th} year, 1^{st} Semester

Department of Computer Science and Engineering

Supervisor

MOHAMMAD ABDULLAH AL MUMIN

Professor

Department of Computer Science and Engineering

23rd July, 2022

Available Corpora of Low-resource Bangla Language for Machine Translation

A Thesis submitted to the Department of Computer Science and Engineering, Shahjalal University of Science and Technology, in partial fulfillment of the requirements for the degree of Bachelor of Science in Computer Science and Engineering.

By

Masum Ahmed
Reg. No.: 2016331028
4^{th} year, 1^{st} Semester

Md Shamihul Islam Khan
Reg. No.: 2016331078
4^{th} year, 1^{st} Semester

Department of Computer Science and Engineering

Supervisor
MOHAMMAD ABDULLAH AL MUMIN
Professor
Department of Computer Science and Engineering

23rd July, 2022

Recommendation Letter from Thesis Supervisor

The thesis entitled *Available Corpora of Low-resource Bangla Language for Machine Translation* submitted by the students

1. Masum Ahmed
2. Md Shamihul Islam Khan

is under my supervision. I, hereby, agree that the thesis/project can be submitted for examination.

Signature of the Supervisor:
Name of the Supervisor: Mohammad Abdullah Al Mumin
Date: 23rd July, 2022

Certificate of Acceptance of the Thesis

The thesis entitled *Available Corpora of Low-resource Bangla Language for Machine Translation* submitted by the students

1. Masum Ahmed
2. Md Shamihul Islam Khan

on 23rd July, 2022, hereby, accepted as the partial fulfillment of the requirements for the award of their Bachelor Degrees.

Head of the Dept. Mohammad Abdullah Al Mumin Professor Department of Computer Science and Engineering	Chairman, Exam. Committee Dr. Farida Chowdhury Associate Professor Department of Computer Science and Engineering	Supervisor Mohammad Abdullah Al Mumin Professor Department of Computer Science and Engineering

Abstract

Machine Translation translates texts of one natural language into texts of another automatically. State of the art MT approaches use parallel corpora as their training data. This study aims at finding available Bangla ↔ English parallel corpora as well as measuring quality of them by using common linguistics features, evaluating well-known metrics by following statistical machine translation techniques. This paper reviews of existing publicly available Bangla ↔ English parallel corpora till 2020. Thirteen out of fourteen corpora have been founded publicly available. Most of the corpora were constructed using open source domain and preprocessed of their own way.

Keywords: Machine Translation, Low-resource, Available Corpora, Bangla to English, English to Bangla, Statistical Machine Translation

Acknowledgements

First, we express our deepest gratitude to Allah, the Almighty, for pouring His blessings on us throughout our research, enabling us to finish the research successfully. We would really like to express our heartiest gratitude to our research supervisor, Professor Mohammad Abdullah Al Mumin, for providing us with the chance to conduct research and for providing us with essential guidance throughout this research. His vision, genuineness, and determination have all left an indelible impression on us. He has instructed us on how to do research and how to convey the results of our investigation as simply as possible. It was an incredible pleasure and honor to be able to work and learn under his supervision. We are very appreciative for all he has provided for us. We would also want to express our gratitude to him for his friendly, sensitivity, and wonderful sense of humour.

We would like to convey our heartfelt thanks to our mentor, Md Zobaer Hossain, for his guidance and support during this thesis work.

Finally, we would want to thank everyone who has helped us in completing the research project, whether directly or indirectly.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

In the year 2012, the Encyclopedia Britannica declared that they will no longer print any publications. 244 years old publisher has moved out from the printing to publishing online as the technological revolution made access to the internet so easy [1]. This news is nine years old, recent statistics are more astounding. The amount of data produced in the past two years has surpassed all prior human history. The exponential growth of the internet has made this possible. Internet subscribers of Bangladesh have reached 93.702 million users in 2019 at an increased rate of 15-16% per year [2]. This all proves how we have engaged ourselves in these modern technologies.

72.1 percent of internet users prefer to access sites in their native language, according to Common Sense Advisory [3]. However, using human translations on billions of sites or contents is not a viable option. Without any automatic translations system it is not possible to make the internet easier. Machine Translations(MT) is introduced to solve this problem. MT automatically converts the text of one language to another. Statistical machine translations(SMT) and Neural machine translations(NMT) are the two very common approaches that have been used in MT. Both SMT and NMT are data-driven techniques in this case. That is, in order to construct an automated translations system, both strategies need corpora.

Bangla is the world's fifth most widely spoken native language and the world's seventh most widely spoken language, with roughly 228 million native speakers and additional 37 million as second language speakers [4]. However, it is also true that in terms of machine translation, Bangla is a low-resource language. The number of studies and required resources related to Bangla MT is very low.

1.1 Motivation

Machine translation system development is a time-consuming and complex process that requires a lot of effort. Languages are complicated, therefore it is vital to examine word knowledge for each language. The meanings of many terms are not always clear, and they may be translated in a number of ways. The result of machine translation may be evaluated automatically. Various metrics, such as BLEU [2], NIST [3], TER [4], have been suggested. It is insufficient to rely just on automated assessment of translation output. As a result, a human examination is required to identify the key issues with a machine translation output.

1.2 Goals of the Project

Since corpora are pre-requisite of MT systems, a large scale corpora will enhance the MT system in Bangla language. However, there are some publicly available corpora on Bangla ↔ English. According to our knowledge, There have been no systematic analyses of the parallel corpora that have been collected so far. An analysis on corpora construction along with their data distribution and quality comparison will create an opportunity for the researchers on this field.

In this paper, we reviewed all the available Bangla ↔ English parallel corpora. While reviewing these corpora we explored the system they followed to make their corpora. In addition, we have also taken a look at the data distribution of these corpora. In addition, we employed a state-of-the-art Statistical Machine Translation system to assess their machine translation performance.

1.3 Structure of the Thesis

We outline this paper as follows:

Chapter 2 In this chapter, we have discussed about background study and literature reviews.

Chapter 3 Research methodology is discussed in this chapter.

Chapter 4 Data Collection, preprocessing, and evaluation are described here.

Chapter 5 In this chapter, we have discussed about Results and Discussion.

Chapter 6 This chapter concludes this report in brief.

Chapter 2

Background Study

2.1 Machine Translation

Machine translation is the use of computer software to translate a text from one language to another without the assistance of a person. Machine translation may be divided into three categories:

Transfer-based MT: This technique is based on the source and target languages, syntactic, morphological, and semantic analyses.

Interlingual MT: The original language is converted into a linguistically representation that is an optional part of the semantics of the text.

Direct MT: Without any intermediary stages in the translation process, the source language is immediately translated into the destination language. The morphological inflections are removed from the source to get the basic form, which is then matched in a bilingual dictionary.

The level of the source language analysis is the key distinction between these three methodologies. The Vauquois Triangle *Figure 2.1* breaks down the procedure into manageable segments. The methodology requires more analysis on the source language side and more generation on the target language side with each level higher. The methodology requires more analysis on the source language side and more generation on the target language side with each level higher. The direct MT, which represents the lowest level in the triangle, is a lexical translation from the source to the target language. Each phrase is turned into an abstract representation at the top level, which symbolizes the Interlingual.

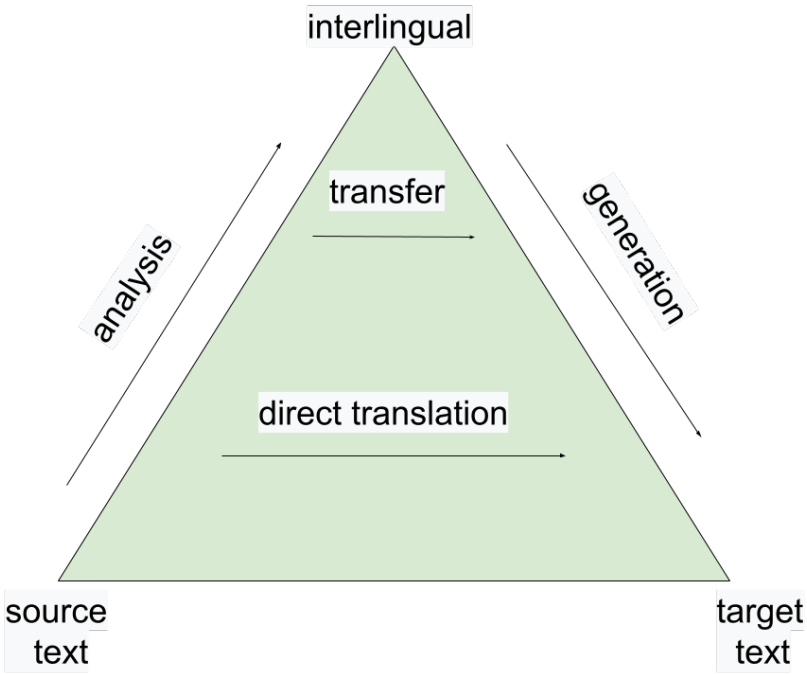

Figure 2.1: Vauquois Triangle [1]

2.2 SMT

In the field of machine translation, statistical machine translation (SMT) is a paradigm in which translations are created on the basis of statistical models, the parameters of which are determined via the examination of bilingual text corpora. The statistical approach to machine translation differs from rule-based and example-based methods to machine translation [5]. The concept of statistical machine translation is derived from the field of information theory. When a document is translated, the probability distribution $p(e|f)$ is used to determine if a string e in the target language is a translation of a string f in the source language.

By examining previous human translations, statistical machine translation (SMT) is a kind of machine translation that learns how to translate (known as bilingual text corpora). In contrast to the Rules-Based Machine Translation (RBMT) approach, which is mostly word-based, most current SMT systems are phrase-based and construct translations utilizing overlap phrases. The aim of phrase-based translation is to get beyond the constraints of word-based translation by translating long sequences of words. Rather than being language phrases, phrase sequences are word sequences that have been found using statistical methods from multilingual text corpora.

2.3 NMT

In the field of machine translation, neural machine translation (NMT) is a methodology that makes use of a convolutional neural network to determine the probability of a sequence of words, generally modeling full sentences in a highly centralized model [6]. NMT differs from phrase-based statistical techniques that rely on subcomponents that are individually created and assembled. NMT (neural machine translation) is not a radical departure from statistical machine translation (SMT). The use of vector representations ("continuous space representations", "embeddings") for words and internal states is the fundamental departure. The models have a simpler structure than phrase-based models.

NMT is a groundbreaking technique to language translation and localization that use deep neural networks and artificial intelligence to develop neural models. With a major transition from SMT to NMT in only three years, NMT has swiftly become the dominant technique to machine translation. Statistical Machine Translation techniques often provide lower-quality translations with less fluency and adequacy than Neural Machine Translation systems.

The memory used by neural machine translation is a fraction of that required by classic Statistical Machine Translation (SMT) models. This NMT methodology varies from traditional translation SMT systems in that all portions of the neural translation model are trained together (end-to-end) to provide the best translation results.

2.4 Systematic Review

Systematic reviews are impartial descriptions of what has been published and discovered on a certain subject. This is particularly useful in broad research areas when there are several publications, each focused on a certain part of the topic. Traditional literature reviews are substantially different from systematic reviews.

Chapter 3

Methodology

In this section, we will discuss the following topics: the review's purpose, searching parallel corpus, finding out available relevant corpora, data sources and distribution, quality appraisal and data evaluation.

In order to help us through this process, we developed seven research questions, which are shown in *Table 3.1*, to serve as a guide for this review's effort to find out all available Bangla ↔ English Parallel corpora, tools, algorithms and systems work.

3.1 Research Questions

Research questions to go through our systemic review of available corpora of low resource Bangla language for Machine translation are included in Table (3.1).

ID	Questions
RQ1	How many Bangla ↔ English parallel corpora were found?
RQ2	Are all the founded corpora publicly available?
RQ3	What are the sources of each available parallel corpus?
RQ4	How data are preprocessed to build corpus?
RQ5	How data are distributed in all available parallel corpora?
RQ6	How each corpus is evaluated via Statistical Machine Translation System?
RQ7	Analyze the statistics of both linguistics features and SMT of available corpora.

Table 3.1: Research Questions.

Steps
1) Purpose statement
2) Databases, search engines used
3) Search limits
4) Inclusion and exclusion criteria
5) Search terms
6) Exact searches per database, search engine and the results
7) Document final number of search results

Table 3.2: 7-step framework for an effective search strategy

3.2 Search Strategy

For developing an effective search strategy, we used a 7-step framework by following the 12-step framework developed by Kable [7] . As the steps are mentioned in *Table 3.2*, in addition to serving as a useful tool for recording a systematic review's search strategy, the 7-step framework also serves to guide researchers through the process of identifying and finding relevant literature. Each of the seven stages is presented sequentially in this part in order to improve readability and make it easier for readers to locate specific stages quickly.

3.2.1 Purpose Statement

In 1991, the first attempt was made to develop *Anglabharti* [8], an Bangla ↔ English Machine Translation system as part of the English-to-Indian Languages system. After that, the number of corpora had increased slowly and reached at fourteen. The search did not include any book chapters. The purpose was to discover as well as a thorough analysis of available corpora of row-resource Bangla language developed during the years 1991 to 2021.

3.2.2 Search Limits

To study Machine Translation, and Systemic literature review, we limited searches to peer-reviewed journals. As Bangla is a low-resource language, to study Machine translation in Bangla ↔ English, and collect available corpora, we did not limit our search. We studied all articles and collected all datasets we got.

Inclusion Criteria	Exclusion Criteria
Papers written in English	Papers written in other languages rather than English
Available Bangla ↔ English parallel corpus	All others parallel corpus without Bangla or Bengali
Bangla↔English bilingual corpus	All kinds of monolingual corpus

Table 3.3: Inclusion and Exclusion Criteria defined for finding out relevant corpus and paper.

3.2.3 Relevance Assessment

To find out available relevant corpora and their papers, we used the criteria indicated in *Table 3.3* for inclusion and exclusion, to be explicit about the corpora we considered in our review. We found fourteen relevant corpora. Only nine of them had papers.

3.2.4 Search Terms

Prof. Dr. Abdullah Al Mumin, the supervisor, helped to design the search keywords. Previously, he had worked in the Machine Translation area, and he was better equipped to create successful search phrases than the author would have been on his own, owing to the author's lack of experience in the area of study. We have used Machine translation as a core text to search by adding *Bangla*, *Bengali*, *row-resources*, *corpus*, *corpora*, *review papers*.

3.2.5 Documentation of Search Process

The searches were carried out in five different digital libraries, ACM Digital Library [5], IEEE Explore [6], Science Direct Elsevier [7], Springer Link [8] and ACL [9]. In order to complete this work, a conceptual research string comprising the primary keyword of the topic was created. Using search strings in each digital library, we ran the searches on February 18, 2020, and the results are provided in *Table 3.4*.

In ACM library's search engine, we searched all of the strings mentioned in *Table 3.4* by filtering 'Machine Translation' domain in common. If we search without having any domain, engine will show different results and the number of irrelevant papers will be huge because of their searching algorithm.

In SpringerLink, we narrowed the search contents based on the parameters provided on the

	Digital Library					
Term Search	ACM	IEEE	Springer	Elsevier	ACL	Total
Machine Translation	1,129	4201	76,474	283	387	82474
Machine Translation of Low Resources	738	60	18,079	2,656	9	21542
Machine Translation of Bangla	6	32	141	2,656	0	2835
Machine Translation of Bengali	7	21	213	2,656	0	2897
Statistical Machine Translation of parallel corpus	1129	121	1,703	2,656	2	5611
Neural Machine Translation of parallel corpus	1129	30	708	2,656	2	4525
SMT of Bangla parallel corpus	990	1	10	2,656	0	3657
NMT of Bangla parallel corpus	978	0	1	2,656	0	3635
Review paper of Low Resources	1,056	313	57,213	2,669	1	61252
Review paper of Bangla MT	1,052	0	6	2,669	0	3727
Review paper of Parallel Corpora MT	1,085	0	236	2,670	1	3992

Table 3.4: Following the execution of search strings, the number of papers retrieved in each Digital Library.

search page, and only conference papers were counted.

Both ACM and SpringerLink's search engine searches by words in contrast to full strings which is a major problem for getting weird results. For this reason, a huge number of papers come up on the search results, where a maximum of them are not relevant according to the full strings.

Elsevier's searching algorithms seems messy to us because it showed the same results, again and again, despite changing search strings.

3.2.6 Retrieved Datasets and Articles at End of the Search Process

Initially, huge amount articles were found in the search processes *Table 3.4*. Excluding duplicates, irrelevant, we finally came to a stand of 30. Those papers also include datasets' papers. As Bangla is a low-resource language, we only found fourteen parallel corpora.

3.3 Datasets

We have used thirteen Bangla ↔ English Parallel corpora. Nine of them are found on Opus [10] named *GlobalVoices*, *Gnome*, *JW300*, *KDE4*, *OpenSubtitles*, *QED*, *Tanzil*, *Tatoeba*, *Ubuntu*. Rest are *SUPara* [9], *EMILLE* [10], *ILMPC* (Indic Language Multilingual Parallel Corpus) ([11], [12]), *Pan Treebank* Bangla-English parallel ([13], [14]), *SIPC* (Six Indian Parallel Corpora) [15].

Chapter 4

Data Collection, Preprocessing and Evaluation

Statistical Machine Translation (SMT) needs considerably good corpus having large amounts of text data to produce good translations. By searching through different libraries, websites for Bangla ↔ English parallel corpora, we merely found thirteen available corpora. Corpus construction phases are given below.

4.1 Sources of Data

By searching through different libraries, websites for Bangla English parallel corpora but we merely found thirteen Bangla ↔ English available parallel corpora. Some corpora follow newspaper as a corpora source, and some follow the website and Wikisource. Overall we noticed maximum data are collected from online. We made *Table 5.2* to observe a detailed overview of sources of available parallel corpora.

4.2 Data Preprocessing

Because there were so many different sources of similar texts, it required a great amount of time and effort to bring them all together into a common framework. We couldn't get more details about preprocessing. From *Table 5.3*, we get an overview of corpora of how they preprocessed

themselves.

4.3 Data Distribution

We attempted to discover numerous linguistic parameters in each individual corpus, such as sentence fragments, unique words, total words, average words per sentence, words ratio, lexical diversity, and so on. With this analysis, we were able to observe ins and out overall qualities, quantities, difficulties, characteristics of all available corpora.

4.4 Evaluation Metrics

By using Statistical Machine Translation System, we evaluated translation results by three metrics: BLEU, NIST, TER. BLEU scores were calculated using *multi-bleu.perl* in the Moses toolkit [16]. NIST scores were calculated using *mteval-v13a.pl* in the Moses toolkit. To get the proper scores of NIST, you must have to wrap translated file into sgm format by using *wrap-xml.perl*. The amount of editing that a person would have to do to update a system output so that it perfectly matches a reference translation is measured by TER (Translated Edit Rate). TER scores were calculated using *tercom.version.jar* [11] which use JDK and JRE to run.

In this paper, we reviewed all available corpora of Bangla ↔ English we had found. The following section is devoted to data sources, distribution, evaluation. Thereafter, we discuss provided tools and, finally, we present our plans for future work.

4.5 Automatic Evaluation

Human assessment is one of many approaches for evaluating machine translation. It is vast, but it is also quite costly, requires a large number of human resources, and may take months to complete. As a result, an automated assessment is required. It should be low-cost, rapid, consistent, and have a good correlation with human assessment. BLEU, NIST, TER, and other automated assessment techniques for machine translation have been used. The Bleu technique is the most popular.

4.5.1 BLEU

Bleu is an acronym for "Bilingual Evaluation Understudy" [2]. Using this algorithm, automatically evaluates the quality of a machine translation on a constant basis. The hypothesis translation is compared to one or more reference translations in this form of comparison. A hypothesis translation, also known as a candidate translation, is a translation that must be examined before it is accepted. In most cases, the reference translation is the accurate translation used to evaluate and contrast the hypothesis. When the candidate translation has a large number of words and phrases that are identical to those in the reference translations, translation is considered to be satisfactory. The following is the formula used by the Bleu to calculate the score:

$$Bleu = BP \times exp(\sum_{n=1}^{4} w_n \times log p_n) \tag{4.1}$$

The Bleu-Score is calculated using an adjusted n-gram accuracy. To arrive at the final result, multiply the number of words from the system translation that occur in any one reference translation by the total number of words in the candidate translation. The program determines the number of 1-, 2-, 3-, and 4-grams in each text using a mathematical formula. The BP (Brevity Penalty) is used to account for the differences between the hypothesis and reference translations when the hypotheses translation is lower. The Bleu-Score can be a positive value from 0 to 1. This score indicates the functional similarity between the method and the reference. When the result is around 1, the machine translation contains a big number of sentences that are comparable to those in the reference translation, indicating that the translation is excellent.

4.5.2 NIST

Other recent endeavors, such as investigations of the association between various human evaluations and various automated measurements, as done in recent WMT workshops [17], [18], demonstrate the interest in enhancing MT metrology. The NIST [3] Measurements for Machine Translation Challenge (MetricsMATR) has the unique purpose of focusing only on MT metrology research, bringing together many research initiatives in the area of MT metrology, and assisting in the creation of automated metrics. Researchers may discuss ideas on MetricsMATR [3].

To effectively capture the strengths and drawbacks of MT measures, they must be analyzed

across vast and diverse data sets. For example, one would wish to compare the relative performance of measures depending on certain factors like the source language, the kind of MT system (statistical, rule-based, or hybrid), and the data genre. MetricsMATR makes use of a variety of data sets collected by NIST-coordinated MT assessments. Each data collection includes one or more reference translations, one or more machine translations, and one or more human evaluation types.

The analysis for MetricsMATR was based on the principle that the closest an automated measure resembles human assessors, the better the metric. As a result, human evaluation is critical in this task. Human evaluations of many forms are available and will be used to compare metrics scores against. Finding the optimal approach to conduct human evaluations is a huge scientific problem in and of itself. One problem is achieving appropriate intra- and inter-annotator agreement; another is devising evaluations that can be completed in a fair amount of time and effort. The IWSLT 2006 evaluation campaign [19], the WMT-07 [17] and WMT-08 [18] workshops, and the NIST OpenMT 2009 assessment are all recent projects that investigated intra- and/or inter-annotator agreement and revealed a need for improvement. WMT-08 focused on enhancing human evaluation techniques by enhancing intra- and inter-annotator agreement and minimizing assessment time (by assessing at the sub-sentential element level, as opposed to MetricsMATR's approach of assessing at the sentence level).

4.5.3 TER

It is a statistic for evaluating the output of a machine translation program that is automatically generated [4]. It counts the number of modifications that a person would have to do in order to convert a system output into one of the reference values. Replacements of words, deletions, changes of a word sequence, and insertions are all examples of changes that may be required. The term TER is defined as follows:

$$TER = \frac{number \quad of \quad edits \quad needed}{average\ number\ of\ reference\ words}$$

Chapter 5

Results and Discussion

All of the research questions are described and explained in this chapter one by one from *Table 3.1* with a detailed analysis.

5.1 RQ1 - *How many Bangla ↔ English parallel corpora were found?*

Corpus Name	**Available on (as *tmx*, *moses*, or *txt* format)**
EMILLE	Personally Collected
GlobalVoices	Opus, Casmacat (website)
Gnome	Opus
ILMPC	Workshop on Asian Translation
JW300	Opus
KDE4	Opus
OpenSubtitles	Opus
QED	Opus, QCRI
SUPara	supara-github
SIPC	sipc-github
Tanzil	Opus, TanzilNet
Tatoeba	Opus
Ubuntu	Opus
Pan Treebank*	Not publicly available

Table 5.1: List of websites where corpora were available for download (Sorted Alphabetically)

We have found fourteen Bangla ↔ English Parallel corpora by searching through different libraries. Nine of them are found on Opus named *GlobalVoices*, *Gnome*, *JW300*, *KDE4*, *OpenSubtitles*, *QED*, *Tanzil*, *Tatoeba*, *Ubuntu*. Opus attempts to help by collecting new data sets on a big

scale in order to offer data for a wide range of languages and topics that are frequently underserved. The Opus project's general aim is to make parallel resources publicly accessible, with a particular focus on low-density languages which is really helpful for the researcher [20]. Others are given below.

SUPara corpus [9] is distributed through the Computer Science and Engineering (CSE) department of Shahjalal University of Science and Technology (SUST).

EMILLE corpus [10] developed through the "Enabling Minority Language Engineering" project, which was undertaken by the universities of Lancaster and Sheffield.

ILMPC (Indic Language Multilingual Parallel Corpus) is introduced on Workshop on Asian Translation (WAT 2018) ([11], [12]).

Pan Treebank Bangla-English parallel corpus ([13], [14]) (PTB) is developed by PAN Localization Project [12].

SIPC (Six Indian Parallel Corpora) is constructed via CrowdSoucring by using Amazonâs Mechanical Turk (MTurk) [15].

5.2 RQ2 - *Are all the founded corpora publicly available?*

All of the founded corpora except *Pan Treebank*, are available free of charge for educational and research purposes, however, the license allows collecting statistical data and making short citations.

Maximum corpora are found as *tmx*, *moses* and *txt* formats in different websites. Two of them are not available on internet. So we collected them personally from respected Authors. Some of the corpora have software that uses their own corpus to translate from one language to another. All the information are provided in the *Table 5.2*. All datasets' links are provided in Appendix at the last of the paper.

5.3 RQ3 - *What are the sources of each available parallel corpus?*

Here is a discussion about the source of the corpora that we've worked with. All source link is appended in Appendix B.

In *JW300* corpus [21], data collected from online website *jw.org*. The vast majority of texts

	Corpora											
Sources	Global Voices	GNOME	ILMPC	JW300	KDE4	Open Subtitles	QED	Supara	SIPC	Tatoeba	Tanzil	Ubuntu
Amazon's Mechanical Turk									✓			
Budget speech								✓				
Commercial policy								✓				
Essential interfaces								✓				
European Central Bank		✓		✓	✓						✓	✓
Educational video subtitles							✓					
Education policy								✓				
Government official website								✓				
Holy Book				✓							✓	
Khan Academy							✓					
Movie subtitles						✓						
Magazines				✓			✓					
Novel and Feature								✓				
Newspaper	✓	✓		✓	✓			✓		✓	✓	✓
OpenSubtitles						✓						
Online sources.	✓	✓	✓				✓			✓		✓
Online educational content								✓				
Prime Minister’s speeches								✓				
TED							✓					
Udacity							✓					
WikiSource	✓	✓		✓	✓					✓	✓	✓
Website	✓	✓		✓		✓	✓		✓	✓	✓	✓
Wikipedia and Banglapedia								✓	✓			

Table 5.2: Sources of all available parallel corpora

collected from the magazines *Awake!* and *Watchtower*. In this corpus, multilingual articles are mainly translated from the bible to 300 languages.

In *SIPC* corpus [15], they apply an established protocol for using *Amazonâs Mechanical Turk* (MTurk) to collect parallel data to train and evaluate translation systems for six Indian languages. They investigate the relative performance of syntactic translation models and explore the impact of training data quality on the quality of the resulting model. Finally, they release the corpora to the research community under the Creative Commons Attribution-Sharealike 3.0 Unported License.

QED or *AMARA* [22] corpus is an open multilingual collection of subtitles for educational videos and lectures collaboratively transcribed and translated over the AMARA web-based platform. Here maximum contents are collected from *Khan Academy*, *Coursera*, *Udacity*, *TED Talks* and *ELRA* website.

In *OpenSubtitles* corpus [23], The collection is a database dump of the *OpenSubtitles.org* subtitle repository, which contains 3.36 million subtitle files in over 60 languages. This is a very upgraded version of the subtitle gathering that has better sentence alignment and language checks.

SUPara [9] have used texts from some sources that are either publicly available or granted permission from respective copyright holders. They collect data from novel and feature. Some Documents containing Prime Minister's speeches, budget speech, commercial policy, etc are obtained from the government official website. Several essayistic texts are collected from different online newspapers, e.g., Bangladesh Sangbad Sangstha(BSS), BDNews24.com. Many other documents are obtained from websites of several companies like Grameen Phone, Bangladesh Parjatan Corporation and so on. They also generated some documents by mining data from *Wikipedia* and *Banglapedia*.

Tanzil [20] is a collection of Quran translations, and all data are collected from the Quran and their own *website*.

GlobalVoices [20] collected data mainly from *Global Voice* website. *Tatoeba* [20] collected data from *Tatoeba* website.

KDE4, *Ubuntu*, *GNOME* corpora are originally mentioned in OPUS's *website*. We didn't get much idea about its source from [20] paper.

For all available corpora from Opus, website also provides pre-compiled word alignments and phrase tables, bilingual dictionaries, frequency counts, and these files are found there.

	Corpus Name				
Preprocessing Terms	GlobalVoices GNOME JW300 Tanzil Ubuntu	OpebSubtitles	QED	Supara	ILMPC
Cleaning up Documents				✓	
Encoding and Markup				✓	
Alignment				✓	
Tools				✓	
Subtitle conversion		✓			
Sentence segmentation	✓	✓	✓		✓
Tokenization	✓	✓	✓		✓
Correction of OCR	✓	✓			
Spelling errors	✓	✓			
Inclusion of meta-data		✓			
Training			✓		✓

Table 5.3: Statistics overview of Corpora of how they preprocessed themselves.

5.4 RQ4 - *How data are preprocessed to build corpus?*

We couldn't find more details on preprocessing on the corpora. In the following *Table 5.9*, we tried to give a rough idea of the few preprocessing terms corpora that follow. Here, the following steps have been applied as preprocessing on the English and Bangla documents for the available corpora:

Cleaning up Documents - This cleaning up means that the various formats, for example, rtf, doc, and pdf, are converted to plain text files. Tagged files like HTML and PHP files are normalized by deleting tags and then converted to plain text files.

Encoding and Markup - The texts are encoded according to international standards by using UTF8 (Unicode). For Bangla documents, they have used the 'Nikosh' converter to encode all formats into Unicode.

Alignment - The alignment of translated segments with source segments is essential for building parallel corpora. Each file in a sub-directory is aligned separately with its translation to keep alignment errors at a low level.

Tools - Some corpora used various tools for preprocessing the data. In particular, they applied various types of open-source software and free research tools. These tools include sentence splitter,

word histogram generator, Unicode converter, etc.

Subtitle conversion - Only the OpenSubtitles corpus follows these preprocessing words, as seen in *Table 5.3*. No specific encoding format is required for the subtitles submitted by its users, and OpenSubtitles does not impose any such requirement. In order to avoid this, it is necessary to find the most probable encoding for the file using a variety of heuristics. When there are numerous acceptable alternative encodings available, the 'chardet' library is used to select the most probable encoding based on the file content [24].

Sentence segmentation and tokenization - In *Table 5.3*, we can see most of the corpora follow these preprocessing terms except the Supara corpus.

Correction of OCR and spelling errors - Because a large percentage of subtitles in our dataset are automatically generated from video streams using Optical Character Recognition (OCR), there are a significant number of OCR errors. In *Table 5.9*, we can see most of the corpora follow these preprocessing terms except the Supara corpus, the QED corpus, the ILMPC corpus.

Inclusion of meta-data - This preprocessing phase's goal is to create the meta-data that will be connected with each individual subtitle.

Corpus Name	Languages	Files	Sentence Fragments	Tokens
Emille	14			67M
GlobalVoices	41	224,096	4.93M	88.71M
Gnome	187	113,344	58.12M	267.27M
ILMPC	8	3,200		4.64M
JW300	**380**	1,285,939	105.11M	1.95G
KDE4	92	75,535	8.89M	60.75M
OpenSubtitles	62	**3,735,070**	**3.35G**	**22.10G**
QED	225	271,558	30.93M	371.76M
SUPara	2	80	0.02M	0.45M
SIPC	6		0.15M	1.6M
Tanzil	42	105	1.01M	22.33M
Tatoeba	309	309	7.82M	57.55M
Ubuntu	244	30,959	7.73M	29.84M

Table 5.4: *Bird's-eye-view of founded Parallel Corpora (Sorted Alphabetically)*

	Sentences	Words		Unique Words		Average Words per Sentence	
Corpus Name	En or Bn	En	Bn	En	Bn	En	Bn
Emille	6,375	89,027	90,062	6,636	10,816	13.965	14.127
Global Voices	137,620	2,536,451	2,269,045	**88,602**	**130,606**	18.4308	16.4877
Gnome	132,481	637,363	619,182	6,869	7,147	4.811	4.674
ILMPC	337,428	2,260,636	1,840,722	42,510	83,418	6.710	5.455
$JW300_1$	366,972	5,180,241	5,074,551	43,781	80,447	14.116	13.828
$JW300_2$	370,948	**5,890,630**	**5,083,847**	57,461	80,575	2.332	13.705
KDE4	36,381	149,273	129,159	14,174	10,869	4.103	3.550
OpenSubtitles	**413,602**	2,401,653	1,974,181	42,432	84,343	5.807	4.773
QED	2,043	71,695	245,548	8,167	26,674	**35.093**	**120.190**
SUPara	21,158	244,539	202,866	14,571	22,456	11.56	9.59
SIPC	20,788	290,972	240,077	16,594	13,501	13.997	11.549
Tanzil	187,052	4,391,125	4,185,894	17,600	17,925	23.475	22.378
Tatoeba	5,120	24,656	22,321	2,082	3,345	4.816	4.360
Ubuntu	5,634	21,791	17,900	5,619	4,053	3.868	3.177

Table 5.5: *Statistics Analysis of Available English ↔ Bangla Parallel Corpora (Sorted Alphabetically)*

5.5 RQ5 - *How data are distributed in all available parallel corpora?*

5.5.1 Coverage of Language

JW300 is the largest having huge coverage of languages, almost 380 languages mentioned in *Table 5.4*. Tatoeba covers the second most languages. But Bangla ↔ English parallel corpus is tiny having 5,120 sentences mentioned in *Table 5.5*. Though GlobalVoices has the lowest language coverage among other corpora of Opus, Bangla ↔ English side is pretty huge having 137,620 sentences. The number of language coverages of other Opus corpora is given in *Table 5.4*. All available corpora of Opus are bidirectional. Which means, Bangla to English or English to Bangla translation both are possible.

ILMPC corpus covers eight languages. These are Bangla, Hindi, Malayalam, Tamil, Telegu, Sinhalese, Urdu and English. This corpus is used for the pilot as well as multilingual English-Indic or Indic-English Languages sub-tasks. It is a collection of 7 bilingual parallel corpora of varying sizes, one for each Indic language and English. The parallel corpora are also accompanied by monolingual corpora from the same domain.

SUPara corpus is an Bangla↔English parallel corpus consisting of more than 0.45M words in

either languages, which is the largest among freely released corpus of its kind.

EMILLE corpus consists of a series of monolingual corpora for fourteen South Asian Languages and a parallel corpus of English and five of these languages.

SIPC is a collection of parallel corpora between English and six languages from the Indian subcontinent: Bangla, Hindi, Malayalam, Tamil, Telugu, and Urdu.

5.5.2 Arrangement of Files

There are 152,939 movies or TV episodes covered by OpenSubtitles (as determined by IMDb identifier). 8% of the IMDb identifiers are associated with subtitles in at at least 20 languages, 28% with at least 10 languages, 44% with at least 5 languages, 70% with at least 2 languages. However, the reason for the large OpenSubtitles files is because, in the event of numerous CDs, each movie or TV show may contain numerous files.

QED or AMARA corpus covers 44,620 online educational videos, TV shows such as Khan Academy, TED Talks, Udacity, Coursera and generates around 271,558 files from their subtitle files. JW300 generates files from Bibel, New Testament, Awake! and Watchtower magazines by segmenting each chapter or article into a reasonable size. Tanzil builds 80 files by splitting the Holy Quran into 80 sections. SUPara corpus contains altogether 80 document pairs that consist of literature, journalistic, instructive, administrative, external communication. Other corpora mentioned in *Table 5.4* follow the same rule to generate files.

5.5.3 Count of Sentences or Fragments

Sentence fragment is a punctuated word, phrase, or dependent clause that lacks a subject, verb, or both. While sentence fragments may be used for impact in certain kinds of writing, they are seldom utilized in academic or professional writing. Some of the corpora especially Opus used fragments of sentences. Others (such as ILMPC, SUPara) used full sentences, instead of using a fragment.

We found two different corpora of JW300, shown in *Table 5.5*. Every corpus except $JW300_1$ and $JW300_2$ [13] have same number of sentences or fragments of sentence. These two corpora are not even properly aligned with each other which is so much important for translation. For example, Bangla document's fifth line is not aligned with the English document's fifth. There

exists excessive new lines (empty lines) instead of proper alignment.

OpenSubtitles has the maximum number of sentence fragments around 3.35G because of the availability of online subtitles' sources. Also, Bangla ↔ English side of OpenSubtitles has the largest number of sentences. Overall count of sentence fragments for EMILLE and ILMPC are not mentioned in the dataset's description. SUPara has only 0.02M sentences because it covers only two languages. But SUPara is one of the finest datasets of it's kind. Count of sentences or fragments of other corpora is mentioned in *Table 5.5*.

5.5.4 Smallest Unit

Token is the smallest unit that each corpus divides to. Typically each word form and punctuation (comma, dot) is a separate token. Therefore, corpora contain more tokens than words. Spaces between words are not tokens. But in *Table 5.5*, the term "Token" actually refers to the total number of words in the corpus, of course excluding punctuations. OpenSubtitles has the largest number of tokens nearly 22.1G because of its data collection policies. SUPara has 0.45Million tokens because it covers only two languages. Others are mentioned in *Table 5.4*.

"Words" which is mentioned in *Table 5.5*, denotes the total number of barely Bangla words in Bangla side and English words in English side corpus individually. Some corpora (such as Gnome, Ubuntu, QED) have Bangla sentences having a lot of English words. We excluded English words from them and counted only Bangla words. $JW300_2$ [13] has the highest number of words in both English and Bangla side. QED or *AMARA* corpus [22] has 71,695 English words in contrast to 245,548 Bangla words. This means, almost every sentence of Bangla corpus has more words than the aligned English sentence. Translated English sentence is incomplete for almost every sentences of Bangla. That's why QED has the highest "average words per sentence" in Bangla side which is 120.190 whereas aligned English side has an average 35.093 words per sentence. Number of "Unique Words" of each corpus are given in *Table 5.5*. First, we tokenized them into merely Bangla words for Bangla side and English words for English side and then used set to find out unique words. GlobalVoices has the highest unique words for both in Bangla and English side.

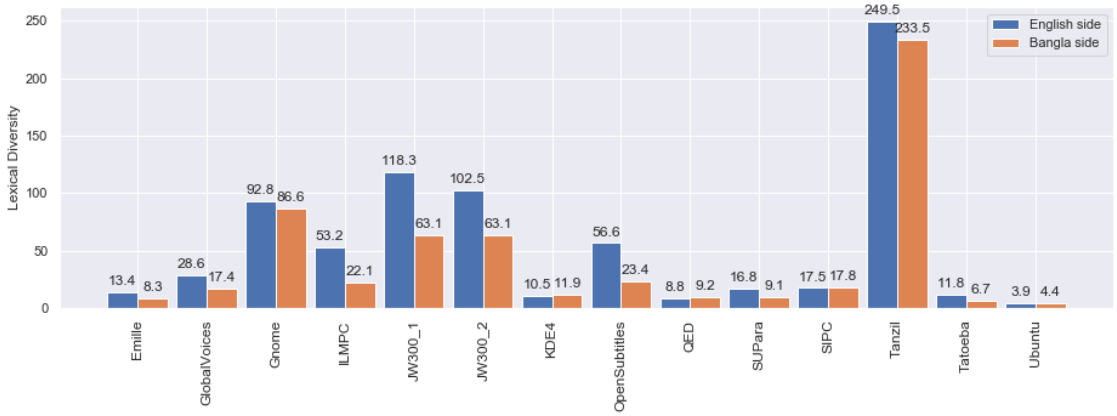

Figure 5.1: Lexical Diversity of available corpora

5.5.5 Diversities

Lexical diversity refers to the ratio of various distinct word stems (types) to the total number of words, which is one element of "lexical richness". It refers to âthe range of different words used in a text, with a greater range indicating a higher diversityâ. Details of lexical diversity of available corpora are shown in *Figure 5.1*. Big score of *Tanzil* refers that on average each vocabulary item appears very frequently for both English and Bangla sides. Ubuntu got the lowest score which means items appear very rarely on sentences for both English and Bangla side.

5.6 RQ6 - *How each corpus is evaluated via Statistical Machine Translation System?*

Here, we have gone through two major Machine Translation paradigms: Phrase-based Statistical Machine Translation (SMT) and recent adventâ Neural based Machine Translation (NMT). In MT research, SMT [25] has been considered as the state-of-the-art technology until the advent of NMT ([26], [27], [28]) recently, which shows an improved result for many high-resource language pairs [29]. However, for low-resource settings, SMT is still considered as state-of-the-art technology since NMT failed to show improved performance in this scenario ([30], [31], [32]).

We followed [32] for the preprocessing, training, tuning and system configuration.

Preprocessing - We normalized both English and Bangla sides using standard *normalize-punctuation.perl* of the Moses toolkit [16]. We further tokenized English side with *tokenizer.perl*

[14] in the Mosesdecoder and Bangla side by using our own tokenizer [15]. Then we used *clean-corpus-n.perl* to create clean corpus of both Bangla and English.

Training - We built word alignments using *train-model.perl* and symmetrized them using grow-diag-final-and heuristic [16]. We extracted phrase pairs up to a maximum length of eight words. We scored these phrase pairs using maximum likelihood, As a result, a phrase table with typical translation model characteristics for each phrase pair. Using *msd-bidirectional-fe*, we also constructed a lexicalized reordering model.

For language modeling, we trained a separate 5-gram LM model for both English and Bangla sides of the training bitext using *lmplz* to *arpa* format and then using *build_binary* to *blm* format. For Bangla, we used SUMono[33], a monolingual corpus. And for English, we used Europarl[34]. Finally, we built a large joint log-linear model, which used standard SMT feature functions: language model probability, word penalty, the parameters from the phrase table, and those from the reordering model.

Tuning - We tuned the weights in the log-linear model by using SUparadev dataset. We did not have much time to use ilmpc-dev or sipc-dev *Table 5.7* for tuning purpose. These tasks will be accomplished in further research.

For translation, we utilized the Moses toolkit's [16] phrase-based SMT model and provided evaluation results from three test sets, suparatest in *Table 5.7*, ilmpc-testest in *Table 5.8* and sipctestset in *Table 5.8*.

SUPara corpus got the best result while using suparatestset and sipctestset for the statistical machine translation system. But with suparatestset, deviation of BLEU scores among corpora is really high in both En $\leftrightarrow$ Bn and Bn $\leftrightarrow$ En side. Some corpora got 0.00 BLEU scores because of their unorganized dataset, ratio of Bangla and English words per sentence, high lexical diversity, etc. By using ilmpctestset, OpenSubtitles got the best result. But still got the huge deviation problem in BLEU scores section.

Corpus Name	Data sets	Sentences	Tokens(EN)	Tokens(BN)
ILMPC	Development	500	8,050	6,786
	Test	1,000	14,755	12,548
SUPara	Development	500	9,865	7,902
	Test	500	9,926	7,938
SIPC	Development	914	13,363	10,545
	Test	1,001	15,134	11,756

Table 5.6: Statistics of available development and test sets among founded corpora

	En → Bn			Bn → En		
Corpus Name	BLEU ↑	NIST ↑	TER ↓	BLEU ↑	NIST ↑	TER ↓
EMILLE	2.44	2.13	101.96	2.97	2.82	87.25
Gnome	0.92	1.64	107.43	1.65	2.18	88.95
GlobalVoices	9.80	4.21	79.21	10.58	4.73	75.51
ILMPC	4.79	3.06	88.47	7.70	3.89	81.27
KDE4	1.34	1.63	109.10	1.56	2.21	90.97
OpenSubtitles	5.15	3.052	87.51	7.83	1.00	80.89
QED	0.00	1.12	105.16	0.70	1.59	93.81
SUPara	**15.24**	**5.06**	**72.47**	**16.85**	**5.70**	**68.51**
SIPC	3.29	2.62	90.13	5.10	3.22	84.71
Tatoeba	0.74	1.21	108.74	1.19	0.87	89.00
Tanzil	1.35	1.64	96.11	2.48	2.20	93.26
Ubuntu	1.04	1.39	108.95	1.30	1.89	92.08

Table 5.7: Evaluation scores of all corpora using *sumono.5-gram.blm.bn* as language model for English → Bangla and *europarl.5-gram.blm.en* as language model for Bangla → English translation (Sorted Alphabetically) [SUParadev for tuning and **SUParatestset** for evaluating]

	En → Bn			Bn → En		
Corpus Name	BLEU ↑	NIST ↑	TER ↓	BLEU ↑	NIST ↑	TER ↓
EMILLE	1.58	1.81	101.25	2.64	2.11	101.79
Gnome	2.70	1.41	101.14	1.71	1.79	97.90
GlobalVoices	3.94	2.87	84.02	7.01	3.53	85.34
ILMPC	6.40	3.23	80.52	11.92	4.33	79.59
KDE4	2.82	1.40	103.20	1.71	1.67	110.42
OpenSubtitles	**19.16**	**4.69**	**65.07**	**33.18**	**6.66**	**58.67**
QED	0.71	1.39	94.84	2.23	1.92	106.53
SIPC	1.27	1.71	95.08	2.74	2.25	93.63
SUPara	4.18	2.93	85.01	8.04	3.78	83.27
Tatoeba	1.75	1.90	95.12	3.34	2.30	90.86
Tanzil	1.56	1.63	92.65	2.51	2.15	103.68
Ubuntu	1.72	1.17	102.57	1.29	1.52	102.00

Table 5.8: Evaluation scores of all corpora using *sumono.5-gram.blm.bn* as language model for English → Bangla and *europarl.5-gram.blm.en* as language model for Bangla → English translation (Sorted Alphabetically) [SUParadev for tuning and **ILMPCtestset** for evaluating]

	En → Bn			Bn → En		
Corpus Name	BLEU ↑	NIST ↑	TER ↓	BLEU ↑	NIST ↑	TER ↓
EMILLE	0.37	1.14	113.12	1.26	1.70	93.35
Gnome	0.77	1.15	113.56	0.82	1.48	93.32
GlobalVoices	7.35	3.35	86.04	8.66	4.13	82.40
ILMPC	2.90	2.24	98.49	4.50	3.08	88.31
KDE4	0.67	1.08	117.09	1.31	1.55	94.46
OpenSubtitles	2.54	2.15	98.78	4.44	3.11	88.20
QED	0.00	0.68	116.09	0.36	1.12	95.98
SIPC	7.36	3.29	85.60	7.02	3.56	87.07
SUPara	**8.87**	**3.54**	**85.24**	**9.88**	**4.49**	**79.45**
Tatoeba	0.28	0.68	118.39	0.38	0.81	95.43
Tanzil	0.00	0.84	109.87	1.16	1.33	96.23
Ubuntu	0.00	0.84	119.05	0.68	1.17	94.61

Table 5.9: Evaluation scores of all corpora using *sumono.5-gram.blm.bn* as language model for English → Bangla and *europarl.5-gram.blm.en* as language model for Bangla → English translation (Sorted Alphabetically) [SUParadev for tuning and **SIPCtestset** for evaluating]

5.7 RQ7 - *Analyze the statistics of both linguistics features and SMT of available corpora*

OpenSubtitles has the maximum number of sentences or fragments of sentences mentioned in *Table 5.5*. But not all sentences are full, there exist fragments of sentences that affect Bangla English words ratio. $JW300_2$ has the largest number of "Words" in both English and Bangla sides although this corpus is not properly aligned with Bangla and English side's sentences. *GlobalVoices* has the largest number of "Unique Words" in both English and Bangla sides. Although *QED* has the maximum number of "Average Words per sentence" in both English and Bangla sides, Bangla↔English isn't fully translated. Bangla→English words ratio is 3:1. Bangla sentence has more words and too long whereas English sentence is very short and doesn't have a full translation of that Bangla sentence. Ubuntu has the best lexical diversity score but it is the smallest among available corpora.

We used SUParadev for tuning purposes. By using SUParatestset for evaluation of three metrics mentioned in Table, Statistical Machine Translation system generates the best results for SUPara in both En to Bn and Bn to En sides. With ilmpctestset, OpenSubtitles achieved the best results mentioned in Table as well as by using sipctestset, SMT again generates the best performance for SUPara.

Chapter 6

Conclusion

In this study, we described the present state of existing available Bangla ↔ English parallel corpora. This review found fourteen corpora. Except for Pan Treebank, other thirteen corpora are publicly available free of charge for educational and research purposes, however, the license allows collecting statistical data and making short citations. Available corpora were analyzed by finding common features listed them into tables and also into some figures to visualize the quality of each corpus.

We were able to inform about paradigms and their methods, our results with three evaluation metrics by following phrased based statistical machine translation methods. This helped our way to dig deeper into the corpus and find better one. We used three different test sets to make sure the correctness of our SMT model.

For future research, we would like to train all corpora with ILMPCdev and SIPCdev, and then evaluate BLEU, NIST and TER. Also we want to calculate machine translation errors using SMT and NMT results. We would like to add "Word embedding" for each corpus which uses continuous-space representation for natural language.

References

[1] B. J. Dorr, E. H. Hovy, and L. S. Levin, "Machine translation: Interlingual methods," 2004.

[2] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, 2002, pp. 311–318.

[3] M. Przybocki, K. Peterson, S. Bronsart, and G. Sanders, "The nist 2008 metrics for machine translation challengeâoverview, methodology, metrics, and results," *Machine Translation*, vol. 23, no. 2-3, pp. 71–103, 2009.

[4] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, "A study of translation edit rate with targeted human annotation," in *Proceedings of association for machine translation in the Americas*, vol. 200, no. 6, 2006.

[5] P. Koehn, *Statistical machine translation*. Cambridge University Press, 2009.

[6] K. Wołk and K. Marasek, "Neural-based machine translation for medical text domain. based on european medicines agency leaflet texts," *Procedia Computer Science*, vol. 64, pp. 2–9, 2015.

[7] A. K. Kable, J. Pich, and S. E. Maslin-Prothero, "A structured approach to documenting a search strategy for publication: A 12 step guideline for authors," *Nurse education today*, vol. 32, no. 8, pp. 878–886, 2012.

[8] R. Sinha, K. Sivaraman, A. Agrawal, R. Jain, R. Srivastava, and A. Jain, "Anglabharti: a multilingual machine aided translation project on translation from english to indian languages,"

in *1995 IEEE International Conference on Systems, Man and Cybernetics. Intelligent Systems for the 21st Century*, vol. 2. IEEE, 1995, pp. 1609–1614.

[9] M. A. Al Mumin, A. A. M. Shoeb, M. R. Selim, and M. Z. Iqbal, "Supara: a balanced english-bengali parallel corpus," *SUST Journal of Science and Technology*, pp. 46–51, 2012.

[10] P. Baker, A. Hardie, T. McEnery, H. Cunningham, and R. J. Gaizauskas, "Emille, a 67-million word corpus of indic languages: Data collection, mark-up and harmonisation." in *LREC*, 2002.

[11] T. Nakazawa, N. Doi, S. Higashiyama, C. Ding, R. Dabre, H. Mino, I. Goto, W. P. Pa, A. Kunchukuttan, S. Parida *et al.*, "Overview of the 6th workshop on asian translation," in *Proceedings of the 6th Workshop on Asian Translation*, 2019, pp. 1–35.

[12] T. Banerjee, A. Kunchukuttan, and P. Bhattacharyya, "Multilingual indian language translation system at wat 2018: Many-to-one phrase-based smt," in *Proceedings of the 5th Workshop on Asian Translation (WAT2018)*, 2018.

[13] M. A. Hasan, F. Alam, S. A. Chowdhury, and N. Khan, "Neural vs statistical machine translation: Revisiting the bangla-english language pair," in *2019 International Conference on Bangla Speech and Language Processing (ICBSLP)*. IEEE, 2019, pp. 1–5.

[14] ——, "Neural machine translation for the bangla-english language pair," in *2019 22nd International Conference on Computer and Information Technology (ICCIT)*. IEEE, 2019, pp. 1–6.

[15] M. Post, C. Callison-Burch, and M. Osborne, "Constructing parallel corpora for six indian languages via crowdsourcing," in *Proceedings of the Seventh Workshop on Statistical Machine Translation*. Association for Computational Linguistics, 2012, pp. 401–409.

[16] P. Koehn, M. Federico, W. Shen, N. Bertoldi, O. Bojar, C. Callison-Burch, B. Cowan, C. Dyer, H. Hoang, R. Zens *et al.*, "Open source toolkit for statistical machine translation: Factored translation models and confusion network decoding," in *Final Report of the Johns Hopkins 2006 Summer Workshop*, 2007.

[17] C. Callison-Burch, C. S. Fordyce, P. Koehn, C. Monz, and J. Schroeder, "(meta-) evaluation of machine translation," in *Proceedings of the Second Workshop on Statistical Machine Translation*, 2007, pp. 136–158.

[18] ——, "Further meta-evaluation of machine translation," in *Proceedings of the third workshop on statistical machine translation*, 2008, pp. 70–106.

[19] R. Zhang, H. Yamamoto, M. Paul, H. Okuma, K. Yasuda, Y. Lepage, E. Denoual, D. Mochihashi, A. Finch, and E. Sumita, "The nict-atr statistical machine translation system for the iwslt 2006 evaluation," in *International Workshop on Spoken Language Translation (IWSLT) 2006*, 2006.

[20] J. Tiedemann, "Parallel data, tools and interfaces in opus," in *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, N. C. C. Chair), K. Choukri, T. Declerck, M. U. Dogan, B. Maegaard, J. Mariani, J. Odijk, and S. Piperidis, Eds. Istanbul, Turkey: European Language Resources Association (ELRA), may 2012.

[21] Ž. Agić and I. Vulić, "JW300: A wide-coverage parallel corpus for low-resource languages," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 3204–3210. [Online]. Available: https://www.aclweb.org/anthology/P19-1310

[22] A. Abdelali, F. Guzman, H. Sajjad, and S. Vogel, "The amara corpus: Building parallel language resources for the educational domain." in *LREC*, vol. 14, 2014, pp. 1044–1054.

[23] P. Lison and J. Tiedemann, "Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles," 2016.

[24] S. Li and K. Momoi, "A composite approach to language/encoding detection," in *Proc. 19th International Unicode Conference*, 2001, pp. 1–14.

[25] P. Koehn and K. Knight, "Feature-rich statistical translation of noun phrases," in *proceedings of the 41st Annual Meeting of the association for Computational Linguistics*, 2003, pp. 311–318.

[26] N. Kalchbrenner and P. Blunsom, "Recurrent continuous translation models," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 2013, pp. 1700–1709.

[27] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, 2014, pp. 3104–3112.

[28] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," *arXiv preprint arXiv:1409.1259*, 2014.

[29] R. Sennrich, B. Haddow, and A. Birch, "Edinburgh neural machine translation systems for wmt 16," *arXiv preprint arXiv:1606.02891*, 2016.

[30] P. Koehn and R. Knowles, "Six challenges for neural machine translation," *arXiv preprint arXiv:1706.03872*, 2017.

[31] R. Östling and J. Tiedemann, "Neural machine translation for low-resource languages," *arXiv preprint arXiv:1708.05729*, 2017.

[32] M. Mumin, M. Seddiqui, M. Iqbal, and M. J. Islam, "shu−torjoma: An english↔bangla statistical machine translation system," *Journal of Computer Science*, vol. 15, pp. 1022–1039, 08 2019.

[33] M. A. Al Mumin, A. A. M. Shoeb, M. R. Selim, and M. Z. Iqbal, "Sumono: A representative modern bengali corpus," *SUST Journal of Science and Technology*, vol. 21, pp. 78–86, 2014.

[34] P. Koehn, "Europarl: A parallel corpus for statistical machine translation," in *MT summit*, vol. 5. Citeseer, 2005, pp. 79–86.

Appendices

Appendix A

Necessary links mentioned in this report

A.1 List of dataset's data source links

Source Name	Website link
Amazonâs Mechanical Turk	https://www.mturk.com/
Wikipedia	https://www.wikipedia.org/
Khan Academy	https://www.khanacademy.org
Coursera	https://www.coursera.org
Udacity	https://www.udacity.com
TED Talks	http://www.ted.com
ELRA	http://www.elra.org
Banglapedia	https://www.banglapedia.org/
Awake and Watchtower	https://www.jw.org/en/library/magazines/

Table A.1: List of datasets' data source links.

A.2 List of all necessary links mentioned by numbers

Name	Link address
CMSWire [1]	www.cmswire.com/customer-experience/why-machine-translation-matters-in-the-modern-era
Internet in Bangladesh [2]	$en.wikipedia.org/wiki/Internet-in-Bangladesh$
Omniscien [3]	$omniscien.com/?faqs=why-do-i-need-machine-translation$
Bengali Language [4]	$en.wikipedia.org/wiki/Bengali_language$
ACM [5]	https://dl.acm.org
IEEExplore [6]	https://ieeexplore.ieee.org/Xplore/home.jsp
Elsevier [7]	https://www.elsevier.com
Springer [8]	https://link.springer.com
Anthology [9]	https://www.aclweb.org/anthology/
Opus [10]	opus.nlpl.eu
Tercom [11]	http://www.cs.umd.edu/šnover/tercom/
Pan tree bank [12]	https://www.panl10n.net/
$JW300_2$ [13]	$JW300_2$ has 370,948 English and 2,525,512 Bangla Lines. It isn't properly aligned with sentence by sentence.
Mosesdecoder [14]	https://github.com/moses-smt/mosesdecoder/tree/RELEASE-2.1.1/scripts/tokenizer/tokenizer.perl
ReviewOfCorpora [15]	github.com/masumahmedeesha/reviewOfCorpora

Table A.2: List of all necessary links mentioned by numbers.

Appendix B

Datasets Links

Corpus Name	Dataset link
Opus/global-voices	http://opus.nlpl.eu/GlobalVoices.php
Opus/gnome	http://opus.nlpl.eu/GNOME.php
gnome-org	https://l10n.gnome.org/
WAT/indic-multilingual	http://lotus.kuee.kyoto-u.ac.jp/WAT/indic-multilingual/index.html
Opus/jw300	http://opus.nlpl.eu/JW300.php
Opus/kde4	http://opus.nlpl.eu/KDE4.php
Opus/open-subtitles	http://opus.nlpl.eu/OpenSubtitles.php
Opus/open-subtitles-alt	http://opus.nlpl.eu/OpenSubtitles-alt-v2018.php
Opus/qed	http://opus.nlpl.eu/QED.php
QCRI	http://alt.qcri.org/resources/qedcorpus/
supara-github	https://github.com/maamumin/SUPara
sipc-github	https://github.com/joshua-decoder/indian-parallel-corpora
Symfony	https://symfony.com/legacy
Opus/tanzil	http://opus.nlpl.eu/Tanzil.php
TanzilNet	http://tanzil.net/trans/
TranslationTanzil	http://tanzil.net/#19:1
Opus/tatoeba	http://opus.nlpl.eu/Tatoeba.php
Opus/ubuntu	http://opus.nlpl.eu/Ubuntu.php
Translation-LaunchPad	https://translations.launchpad.net/

Table B.1: List of dataset links to download available parallel corpora