A Web Based Bengali News Corpus for Development of Bangla POS Tagging

Thesis

Submitted By

14-25770-1	Tasnim, Zarin
14-25947-1	Sunny, Mehedi Hassan
14-25672-1	Banik, Dipanjan

Department of Computer Science

Faculty of Science & IT

American International University Bangladesh

2018-2019, Fall Semester

06 December, 2018

Declaration

We declare that this thesis is our original work and has not been submitted in any form for another degree or diploma at any university or other institute of tertiary education. Information derived from the published and unpublished work of others has been acknowledged in the text and a list of references is given.

_____

Tasnim, Zarin
14-25770-1
B.Sc. CSE

_____

Sunny, Mehedi Hassan
14-25947-1
B.Sc. CSE

_____

Banik, Dipanjan
14-25672-1
B.Sc. CSSE

Approval

The thesis titled "**A Web Based Bengali News Corpus for Development of Bangla POS Tagging**" has been submitted to the following respected members of the board of examiners of the department of computer science in partial fulfilment of the requirements for the degree of Bachelor of Science in Computer Science on (06-DECEMBER-2018) and has been accepted as satisfactory.

_____

Fahad Ahmed
Assistant Professor & Supervisor
Department of Computer Science
American International University-Bangladesh

_____

Dr. Tabin Hasan
Head (Graduate and MIS) & External
Department of Computer Science
American International University-Bangladesh

_____

Dr. Mahbub Chowdhury Mishu
Assistant Professor & External
Department of Computer Science
American International University-Bangladesh

_____

Dr. Mahbubul Syeed
Head (Undergraduate)
Faculty of Science & Information Technology
American International University-Bangladesh

_____

Professor Dr. Tafazzal Hossain
Dean
Faculty of Science & Information Technology
American International University-Bangladesh

_____

Dr. Carmen Z. Lamagna
Vice Chancellor
American International University-Bangladesh

Acknowledgement

First of all, we would like to be grateful to the almighty **ALLAH** who gives us the effort to work on the project. Thanks to our respectable parents for their outstanding support at every point in our life. We would like to thank our supervisor Mr. **FAHAD AHMED** for his enormous support, inspiration and helpful criticism. We are very grateful to him for giving us opportunity to work with him. Without this continuous support, it would be very difficult for us to complete this project. We would also like to thank all the faculty members for their guideline for making proper documentation of our project. We convey our thanks to our honorable Dean, **PROF. DR. TAFAZZAL HOSSAIN** for encouragement. We convey our thanks to our honorable Vice Chancellor, **DR. CARMEN Z. LAMAGNA** for encouragement. We wish to express our gratitude American International University-Bangladesh (AIUB) providing an excellent environment for research.

Abstract

In this paper we have presented a rule-based POS tagger for Bengali news corpus. We have identified the sentence structure and proposed a word identification model which accept almost all types of Bangla sentences. We have focused our news corpus domain specifically on sports section of online newspaper. More than 300-thousand-word token was collected and we have considered the limitation of human error which we have found during the manual annotation process. We hope that our work will impact on the baseline for building a domain-based news corpus for further development in POS tagging of Bangla language.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Part-of-speech (POS) tagging, also known as grammatical tagging and it is the common form of corpus annotation. POS tagging is the one of the main components of Natural Language Processing (NLP). POS tagging label the word with their appropriate Parts-of-Speech and then it expresses the formation of words in a sentence. In computational application POS tags usually help us for develop different NLP techniques (e.g. sentimental analysis, speech recognition) and it also maintain a link between human and modern artificial machines.

1.1 Objective

This paper developed a POS tagging technique for online Bengali newspaper-based data source. A supervised rule-based technique was proposed for identify Bangla sentence and its corresponding POS tags. Our main goal is to tag the parts of speech of a Bangla word which based on the pre-annotated corpus and train the dataset for extend its functionalities. Our paper was mainly focused on the sports-based news corpus. We have also identified the characteristics of Bengali grammar for this domain and several limitation and work procedure will be discussed later in this paper.

1.2 Background of Study

Bengali also known by its endonym Bangla (বাংলা), is an Indo-Aryan language primarily spoken by the Bengalis in the Indian subcontinent and the whole part of the Bangladesh [24]. Geographically Bengali language is one of the most popular language in the world. Around approximately 250–300 million people in the world speaks with this language [25]. Except some part in the Indian continent more than 90% people of Bangladesh are speaks with Bangla language. Although use of vast majority of this language is still popular and ranked 7^{th} in the languages of native speakers but still we are far behind from the modern science of language development process.

Part-of-speech categorization is taught in school-age children to construct a meaningful sentence. Bangladesh is a sports-oriented country where 80% of the people's daily conversation contains related to sports. So, in this case we have narrowed down our working domain in sports section of Bengali newspaper. Wh

1.3 Scope

Our thesis scope was bound to find a POS tagging mechanism from which we can find out the POS tags against e sentence. We focused to determine that the domain we choose whether it was correct for this particular area or not. Traversing the paper, we will be

discussing how Bengali sentence for a news corpus can be annotated and the different findings that we found analyzing corpus. We will be trying to demonstrate the behavior of a Bengali sentence in the sports domain, how they occur, how they are placed and vary from each other and list out the findings on that and come to a decision against the findings we got.

1.4 Brief summary of Parts-of-Speech tagging

Part-of-Speech explains how a word is used in a sentence. There are several grammatical categories in a POS tagger. A POS tag is a special label assigned to each word in a text corpus. POS tags are used in corpus searches and in text analysis tools and algorithms. A corpus is a large collection of texts which is usually labelled with information about part-of-speech and grammatical category. POS tagging is used in natural language processing (NLP) which is mainly use for machine learning. POS tags are extremely useful since they provide linguistic signal on how a word is being used within the scope of a phrase, sentence, or document.

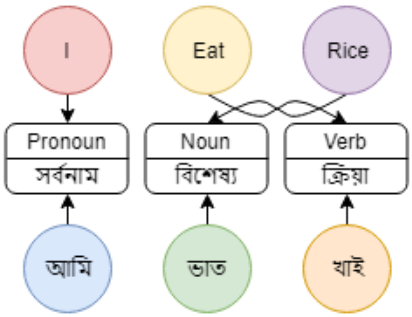

Figure 1-A: POS Tagging

A POS consists different types of grammatical categories such as tense, number etc. All kind of POS tags have a several different tag set and different language tag sets are typically different from each other. Grammatical categories are subdomain to POS tag set. Tags are useful for building parse trees which shows the relations between words. Some set of specific rules, conventions, and principles which dictate how phase, clause and words combined into sentences. There are some lexical categories which words are assigned based on their syntactic context and role. The most common part-of-speech are: noun (বিশেষ্য), pronoun (সর্বনাম), adjective (বিশেষণ), verb (ক্রিয়া), adverb (ক্রিয়া বিশেষন), preposition (অব্যয়).

There are several POS tags set available to build a linguistic corpus. Penn treebank tag sets are very much popular and it is widely used for building POS tagger. Besides that, it got some sub categories too. In part-of-speech tagging by computer, it is typical to distinguish from those sub categories. For simplicity we will restrict to the common part-of-speech that are known to general people. Identifying part of speech tags is much more complicated than simply mapping words to their part of speech tags. This is because POS tagging is not something that is generic. It is quite possible for a single word to have a different part of

speech tag in different sentences based on different contexts. That is why it is impossible to have a generic mapping for POS tags.

Sometimes it is not the same case that one word represents one part-of-speech. Multiple words can represent a single POS tag and each of them also represent their own POST tag as individually they are different but with combined format it may indicate different meaning. It is a common finding that the 'Verb' tag almost always consists of several words. As is represent the act of a subject, it can't always be representable by a single word. And therefore, multiple words need to consist a verb.

POS-tagging algorithms fall into two distinctive groups:

- **Rule-Based POS Taggers**
- **Stochastic POS Taggers**

Rule-Based Tagging

Automatic part of speech tagging is an area of natural language processing where statistical techniques have been more successful than rule-based methods.

Typical rule-based approaches use contextual information to assign tags to unknown or ambiguous words. Disambiguation is done by analyzing the linguistic features of the word, its preceding word, its following word, and other aspects.

Stochastic Part-of-Speech Tagging

The term 'stochastic tagger' can refer to any number of different approaches to the problem of POS tagging. Any model which somehow incorporates frequency or probability may be properly labelled stochastic.

The simplest stochastic taggers disambiguate words based solely on the probability that a word occurs with a particular tag. In other words, the tag encountered most frequently in the training set with the word is the one assigned to an ambiguous instance of that word. The problem with this approach is that while it may yield a valid tag for a given word, it can also yield inadmissible sequences of tags.

Chapter 2: Background Research Review

2.1 Previous Research

There are several different techniques used for POS tagging. Both supervised and unsupervised [10], [1] model of the POS tagging techniques were used in previous researches. Hammad [1] proposed an unsupervised way of building a POS tagger. The author claims that supervised model is time consuming rather than an unsupervised model and Total 54 tag set was used. The author used the pre-tagged corpus and in and it follows the Baum-Welch algorithm which a modified model of HMM but the performance evaluation was not shown in this paper. Rule based POS tagging [2], [3], [15], [18] also shows the significant improvement for Bangla NLP research. CFG and CSG grammar-based parts-of-tagging [3], [6], [9] technique also verified by some authors. We have learned that for Bangla we cannot identified any Bengali sentence just consider the simple verb and simple noun [2]. Beside identifying Bangla sentence [2] proposed and well developed a very good CFG rule to identify the syntax of Bangla sentence. CFG and CSG categories have difference which is showed in [3] and it need to be considered before start of making a rule-based parts-of-speech tagger. Debsari [4] proposed a 4 layer of hybrid parts-of-speech tagging for Bangla which will achieve accuracy closer to the English and French language and also proposed a rule-based system for detect word ambiguity for noun.

Stemming and Lemmatization are the most important part of reducing inflectional and derivationally related forms of a word. In [20], a language independent and supervised model was developed using FIRE Bengali corpus and Tagor's short stories which identifies the lemma of Bangla sentence. Their process involves of Word sense disambiguation with the result of 69.57% accuracy. Newspaper based Bengali corpus also involved in developing for identify stem of words [5]. Total 1000 sentence along with 10 tag set they [4] achieve 74.6% accuracy. A rule based [18] stemming process which achieves 88% accuracy but is limited to verb and noun. The author implied that it can be extensible for other parts-of-speech tags.

In [22], different types of language model for Bangla parts-of-speech tagging was shown. With the 41 tag set and 4048 token the rule-based Brill model gives better performance for making parts-of-speech tagging corpus for Bangla. Although statistical based [21] method with 26 POS tags and 72341 wordform gives more accuracy (90.3%) than Brill method. Hybrid model of parts-of-speech tagging [10] is far better than any other of method we have analyzed in our literature review. Using Hidden Markov Model corpus-based approach was followed and it give 96.28% accuracy. For development of POS tagging Indian language have a specific corpus named Anncorra [16] which is gives guideline for make any rule based parts-of-tagging system in any Indian language. In [10], it also reviewed that for supervised technique we need a pre-annotated corpus which is rather optional in

unsupervised technique. But if a POS tagger make with CFG approach it may give us a good result in morphological analysis but word ambiguity probability is much higher in this process [6].

Universal Networking Language (UNL) is a declarative formal language specifically designed to represent semantic data extracted from natural language texts [26]. The UNL is a common language to exchange information through computers which can deal with natural languages [27]. In [7], a inflectional morphology was proposed to produce word from another words maintain the same POS category. Using UNL word sense ambiguity detection also possible to identify [8]. Correction of Bangla grammar can be a difficult task in machine translation. But CFG based grammar identification [15] using HPSG structure can overcome this problem to make error free grammar rule for POS tagger.

2.2 Related Work

Research of Bengali newspaper-based corpus [11], [12], [13] have been done previously with maximum accuracy of 91.23% accuracy [13] using SVM. In [13], 34 million wordform has been recorded with total 108305 of news document. Total 26 POS tags were used for lexicon development and both SVM and HMM method used to build this POS tagging. Another research [12] which used knowledge-based AI technique which contains 74698 words form. This research is identical to our thesis and but the word forms are low in size compared to our corpus. Both [11] and [12] used Prothom-Alo as their primary data source but in [11] author compares the web-based news corpus with the CIIL corpus and they found the less amount of foreign words in their statistical analysis. Although [11] have done only the statistical analysis and they have not proposed any model for identify the accuracy level of their corpus.

Chapter 3: Methodology

3.1 Data Source Selection

For Bengali language it is quite difficult to find a reliable data source. It is not possible to collect data from some document (e.g. novel, article, magazine) because OCR technology for Bangla still in development process. Also, it not possible to manually take the data from some source because it will take a long time to develop a POS tagger for Bangla. Later on, this paper we have mentioned our limitation for finding a reliable data source. Although for Bengali language we have a huge resource in internet which is mostly newspaper based online portal and some social networking blog. Several researches [1], [12], [13] was conducted from those online newspapers. So, we have selected our primary data source based from the previous research.

We have analyzed several different types of newspaper from online. Our primary target was to select a well-developed website from which we can collect all the raw data. From our observation we found that individual website contains their own format of HTML documentation. Bad format of HTML documentation can cut up our development speed and also makes thing difficult to understand. Finally, we have considered the Daily Ittefaq as our data source.

3.2 Domain Selection

Newspaper contains several different kinds of categories (e.g. Politics, Economics, World, Technology, Science) and different types of news to share but we narrowed down our domain from everything else to sports. We believe that selecting a specific domain can improve the overall result of making a POS tagger.

After reviewed the several newspaper websites from online we found that one of the main problems in sport section is there are lot of foreign word exist in a simple Bengali sentence. Previously no other survey or analysis was done within this domain and we are a sports-oriented country where most of the people of Bangladesh talking contains related to sports. We identified that in sports section of every newspaper the meaning, context, and expression are different from other part of the newspaper. For example, **বল** which express in

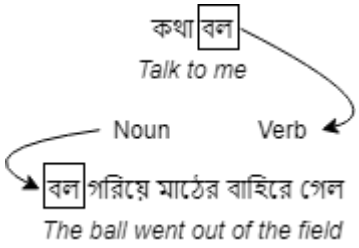

Figure 3-A: Difference of word identity in sports domain

sports domain as Ball which is a playing instrument of Cricket and act as noun but if we express this same word in other sentence we found that it contains different word meaning.

So later on, this paper we will discuss about the sports related contents and development of POS tags in this particular domain.

3.3 Storing the Data into Database

3.3.1 Database Creation

After selection of data source and specific domain our next step was to store the data into database. We have created an extensive database model which can extend later for further research. We have added user roles so that we can track the user of this database. We have used the open source MySQL database management system for store all the data.

The database consists 7 tables for store the all kinds of information related to POS tagging. We have stored the 4 months data into this database along with annotated POS tags. Newspaper's article stored in a way so that we can find its source, published date and number of sentence that article contains. A domain table was created for store the news categories (e.g. Politics, World) so that in future rather than sports we can also add other types of newspaper article.

All the sentences of an article then split and stored into the article lines table. Here we counted and stored the actual number of words we can get by the generated by the lines of articles. We have also made a system to mark a sentence whenever it is justified or answered by the respondent.

Our word table contains all the word token generated by the article sentences. Here we stored the word annotation information for both justified answer and respondent answer. Mainly this table stores the actual information of POS tags of words. We can store multiple answer for each individual word and this will help us to get us more accuracy for building a lexical resource of corpus.

We have used total 7 POS tags which are বিশেষ্য (Noun), বিশেষণ (Adjective), সর্বনাম (Pronoun), অব্যয় (Preposition), ক্রিয়া (Verb), ক্রিয়া বিশেষন (Adverb) and a NULL tag which represent the not answered information both respondent and the justifiers. All the POS information in pos table so that we can later add more tags. There are several POS treebanks available but for Bengali language there is no suitable tag set defined for parse a sentence. An Indian [16], [4] universal tag set was defined for all of the Indian language which also consists Bangla language. Their tags are well developed for this language. If we use some other tag set which uses for other language then word definition can be more difficult to identify. Examining all the tags for Bengali sentence we finally decided to strict with 7 most used POS tags in Bengali grammar.

Subcategories of POS tags are not considered for this research because it will make both our respondent and justifiers confused to annotate the corpus. So all the subcategories of parts-of-speech tags are considered as main root form of their corresponding tags.

A user information table was created for control the database. For use this dataset we decided that only the registered entity can use this for the research. A contribution rate was added for identify the level of participation.

At final step we verify the database model with some test cases before our work procedure and after that we took our next step to crawl the data from online resource.

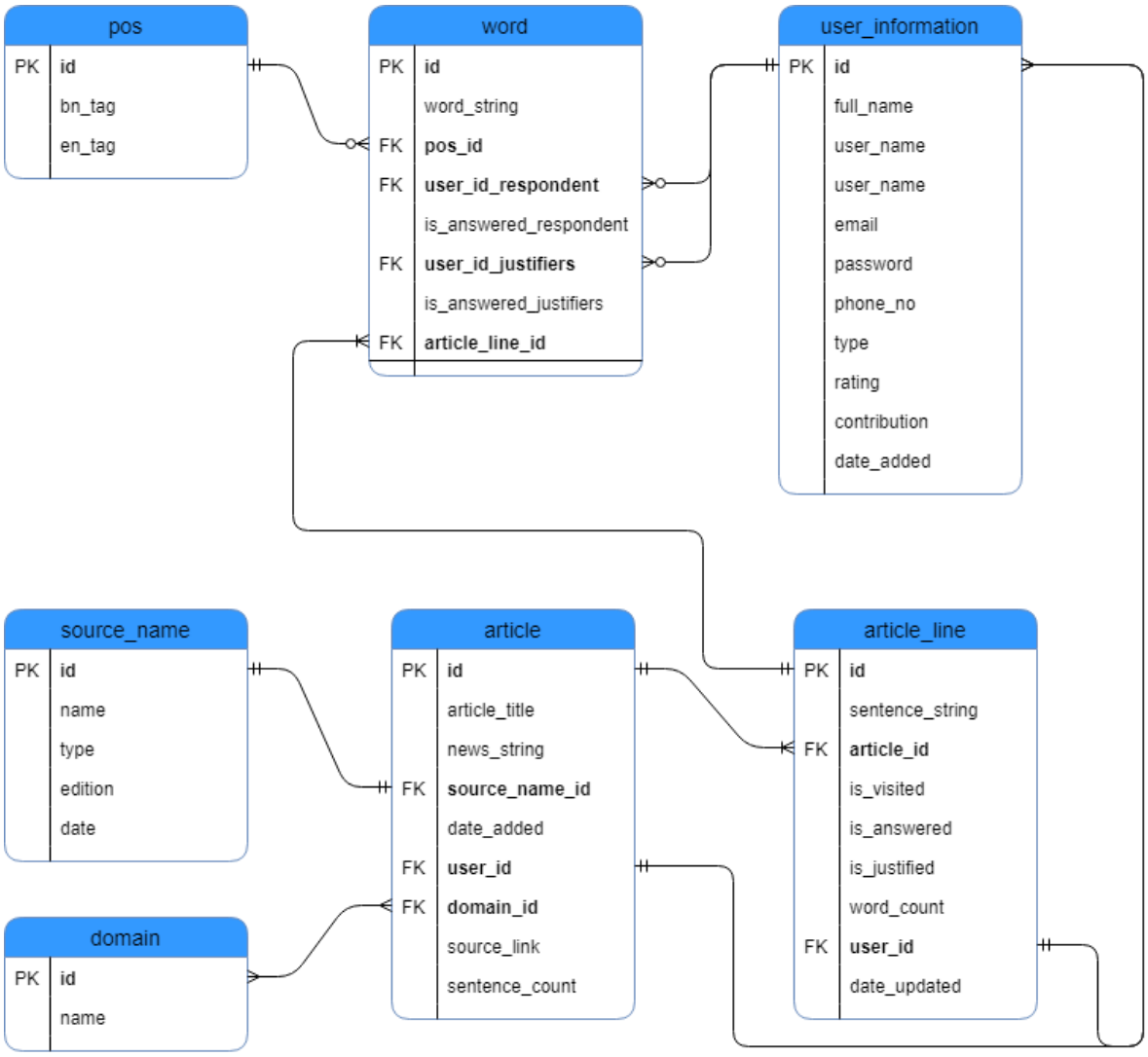

Figure 3-B: ER Diagram of Database

3.3.2 Web Crawling

After the creating relational database we started to collect data from online source. We have collected all the data from online newspaper (Daily Ittefaq) which have discussed earlier part of this section. Our web crawler was built from scratch and we use python programming language for robustness. Two types of python package were used for crawl the data from web source and its was done by semi-automatic manner.

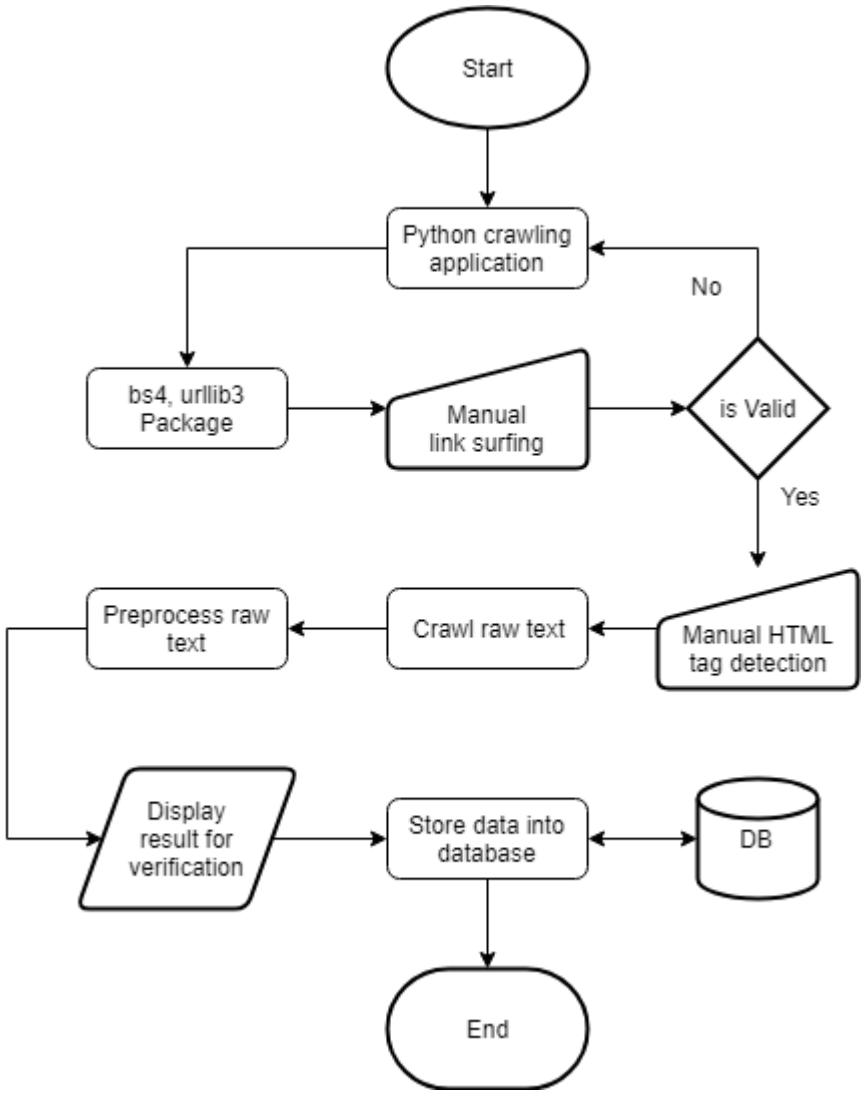

Figure 3-C: Work-Flow of web crawling

We have collected total of 4-month data (from March, 2018 to June, 2018) which includes the all of the important information (e.g. source link, title of the article) we have found on the internet. First, copy the link of the website then we transfer the

link into our crawling application. urllib package parse the link into our application and with the help of beautifulsoup package we manually identify the HTML tag of text extraction from the website. If any website found invalid we then again check with another link and the process continuously goes on.

Before storing the data into database, we preprocessed the text for easier use of making a POS tag. With the help of regular expression, we eliminated the commas, quotations, brackets and other types of expression from the sentence which helps us to generate the word tokens.

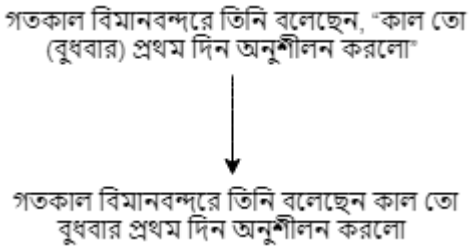

Figure 3-D: Excluding expression for generating tokens

From our data collection we have gathered total 1475 articles which have total 28531 sentences and from the baseline of this sentences we have estimated that maximum 300138 word token can be generate for building lexical resource for a corpus.

Before stored the data into database we have manually checked the result which we have found form our application and unverified data checked again and verified data then stored into the database.

3.4 Data annotation

After data collection and storing the data into database we have done the manual annotation. Our annotation was done with 2 phases. 1st phase annotation was done by online with the help of some intermediate level student. We analyzed that school level student will be perfect for our research because their knowledge is much similar with this case. 2nd phase annotation was done by some expert justifiers who were finally checked the 1st phase annotation answers and also corrected them.

3.4.1 Crowdsourcing

1st phase annotation was done with online crowdsourcing. A modern web base GUI

was developed for easier understanding to our respondent. More than 30 students took participation in this crowdsourcing. We have made it into a responsive GUI so that students can use this application anytime and anywhere. Our data collection

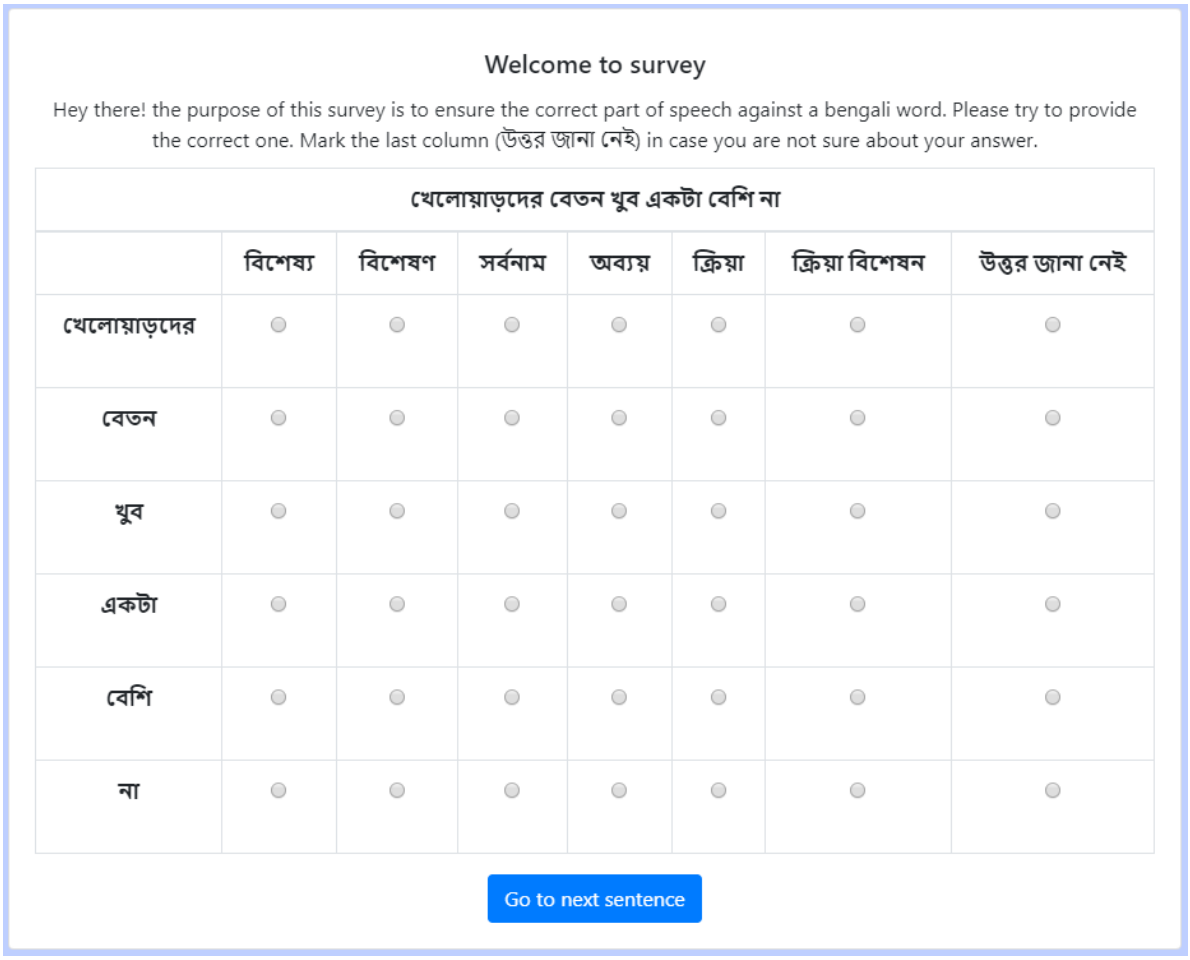

Figure 3-E: Online crowdsourcing GUI

scope was bounded to not greater than 1 month. Because of time constraint after 1 month of data collection we moved into next phase annotation. 1st phase annotation gives us total 4571 annotated sentence and at the later stage we manually check those sentences by experts.

3.4.2 Manual Annotation

2nd phase annotation was done by 3 justifiers who are expert in this field and experienced with this type of work. We took 1000 sentence for justification and this set of data we considered as a training set for our thesis. Rather than online survey expert judges' answers were took with printed document and then it manually

entered into the database. Our 2nd phase annotation takes around 15 days to complete and it took 5 days to entered the manually into the database again.

3.5 Data Modelling

3.5.1 Relational Database

Below figures are the synopsis data of our relational database model –

id	article_title	news_string	source_name_id	date_added	user_id	domain_id	source_link	sentence_count
100	লড়াইয়ের আবহেও পারস্পরিক শ্রদ্ধাবোধ	রিয়াল মাদ্রিদের কোচ হিসেবে এর আগে সাতটি...	3	26/May/18	1	3	http://www.ittefaq.com.bd/print-edition/spo...	50
101	আর্জেন্টিনার ভরাডুবি দেখছেন ম্যারাডোনা	অনেকের কাছে আর্জেন্টিনা এবারও ফেবার...	3	26/May/18	1	3	http://www.ittefaq.com.bd/print-edition/spo...	28
102	ষড়যন্ত্রের গন্ধ পাচ্ছেন রোমেরোর স্ত্রী!	হাঁটুর ইনজুরিটা এতই গুরুতর যে সার্জিও রো...	3	26/May/18	1	3	http://www.ittefaq.com.bd/print-edition/spo...	17
103	সালাহ ঠিক মেসির মতো!	মিশরের সুপার স্টার মোহাম্মদ সালাহর ভয়স...	3	26/May/18	1	3	http://www.ittefaq.com.bd/print-edition/spo...	12
104	অনিশ্চিত বিরাটের ইংল্যান্ড সফর	মেরুদণ্ডের ব্যথ্যাটা বেড়েছে বিরাট কোহল...	3	26/May/18	1	3	http://www.ittefaq.com.bd/print-edition/spo...	9
105	পর্তুগালেরও সুযোগ দেখছেন ফিগো!	লুইস ফিগো তার দেশ পর্তুগালকে নিয়ে একটু ...	3	26/May/18	1	3	http://www.ittefaq.com.bd/print-edition/spo...	19
106	সাত ফুটবলার লন্ডনে	অনেক আগেই ঘোষণা করা হয়েছিল কবে শুরু হব...	3	26/May/18	1	3	http://www.ittefaq.com.bd/print-edition/spo...	23
107	হারের পর আবার জয় আবাহনীর	প্রিমিয়ার হকি লিগে মোহামেডানের কাছে ...	3	26/May/18	1	3	http://www.ittefaq.com.bd/print-edition/spo...	12
108	মিডল অর্ডারেও খেলতে প্রস্তুত সৌম্য	বাংলাদেশের ভবিষ্যতের নির্ভরযোগ্য ক্র...	3	25/May/18	1	3	http://www.ittefaq.com.bd/print-edition/spo...	40
109	মেরুদণ্ডের ডিস্কে সমস্যা তাসকিনের	পিঠের ব্যথার কারণে বাংলাদেশ ক্রিকেট ল...	3	25/May/18	1	3	http://www.ittefaq.com.bd/print-edition/spo...	46
110	ইনজুরিতে কাউন্টি মিশন শেষ বিরাটের	ঠিক হয়ে ছিল, ইন্ডিয়ান প্রিমিয়ার লিগ (আই...	3	25/May/18	1	3	http://www.ittefaq.com.bd/print-edition/spo...	28

Figure 3-F: article table

id	sentence_string	article_id	is_visited	is_answered	is_justified	word_count	user_id	date_updated
9	আমাদের দলের সেরা টি-টোয়েন্টি বোলার	3	0	0	0	5	1	16/Jul/18
10	স্বাভাবিকভাবেই আমাদের জন্য একটু কঠিন	3	0	0	0	5	1	16/Jul/18
12	আজ দলের সঙ্গে যোগ দিবেন আবুল হাসান রাজুও	3	0	0	0	8	1	16/Jul/18
13	মুস্তাফিজের অনুপস্থিতিতে অন্য পেসারদের সুযোগ দেখছেন সাকিব	3	0	0	0	7	1	16/Jul/18
16	বড় দায়িত্ব থাকবে তাই রুবেলের কাঁধেই	3	0	0	0	6	1	16/Jul/18
17	বাঁহাতি পেসার রনি সুযোগ পেতে পারেন মুস্তাফিজের জায়গায়	3	0	0	0	8	1	16/Jul/18
18	নিদাহাস কাপেও একটি ম্যাচ খেলেছিলেন রনি	3	0	0	0	6	1	16/Jul/18
19	ডানহাতি পেসার রাহী বিপিএলে ভালো করেছেন	3	0	0	0	6	1	16/Jul/18

Figure 3-G: article_line table

id	name
1	Test Domain 1
2	Test Domain 2
3	Sports

Figure 3-H: domain table

id	bengali_name	english_name
1	বিশেষ্য	Noun
2	বিশেষণ	Adjective
3	সর্বনাম	Pronoun
4	অব্যয়	Preposition
5	ক্রিয়া	Verb
6	উত্তর জানা নেই	No answer
7	ক্রিয়া বিশেষন	Adverb

Figure 3-I: pos table

id	name	type	edition	date_added
1	Prothom-Alo	Newspaper	Print	11/Jul/18
2	Naya Diganta	Newspaper	Print	11/Jul/18
3	Ittefaq	Newspaper	Print	16/Jul/18

Figure 3-J: source_name table

id	full_name	user_name	email	password	phone	type	rating	contribution	date_added
1	Dipanjan Banik	dipanjanbanik	dipanjanbanik@yahoo.com	123	456	admin	0	0	11/Jul/18
2	Mehedi Hassan Sunny	mhsunny	contact.mhsunny@gmail....	789	012	admin	0	0	11/Jul/18
3	Zarin Tasnim	ztmou	zarinmou1@gmail.com	345	678	admin	0	0	11/Jul/18

Figure 3-K: user_information table

id	word_string	article_line_id	pos_id	user_id_respondent	is_answered_respondent	user_id_justifiers	is_answered_justifiers
1	আমি	1	6	2	0	1	0
2	ভাত	1	6	2	0	1	0
3	খাই	1	6	2	0	1	0
4	সে	2	6	3	0	1	0
5	স্কুলে	2	6	3	0	1	0
6	যায়	2	6	3	0	1	0

Figure 3-L: word table

3.6 Test Set and Training Set Selection

After two phase annotation we have selected 1000 sentence as training set and from this selection we took 300 sentences for validate the data and rest of 27531 sentence we marked as test set. From the training set we have found 7679 word token where 3400 identified as noun, 958 as adjective, 620 as pronoun, 186 as preposition, 1273 as verb, 141 as adverb and 1101 words are tagged as not answered.

3.7 Generating Rules for Identifying POS Tags

We have created a rule driven program to identify the POS tags. We analyzed that any kinds of sentence can be identified with this rule. The table below shows that a sentence needs to

Subject	কে	Who	তামিম, সাকিব
	ক্রিয়া বিশেষন	Adverb	এখন, তখন
	ক্রিয়া	Work	সেঞ্চুরি করেছে
Object	কাকে	Whom	চেলসিকে
	কি	What	গোল দিয়েছে
	কোথা থেকে	Distance	ডি বক্স থেকে
	কোন দিকে	Direction	মাঠের বাহিরে
	কীভাবে	How	পা দিয়ে
	কার সাথে	Whose	রোনাল্ডো
	কোথায়	Place	স্টেডিয়াম
	কখন	Time	সকাল, দুপুর
	কেন	Reason	৩৭ তম সেঞ্চুরির জন্য

Figure 3-M: Table for sentence identification

be parsed by wh-questions to reveal its identity properly. Generally, we tried to develop a rule from the traditional Bengali grammar text book and from our understanding and some help of the teachers we created this rule specifically for the sports domain of news corpus. For other cases it may work but we have not performed any analysis on other domain on news corpus. If we can identify the subject and object with our wh-questions rule it is possible to identify the verb and by identifying the verb we will get the actual sentence expression and corresponding POS tags.

Chapter 4: Findings

4.1 Sentence Identification

Since we have created rules specifically for the sports domain our queries and examples are focused on the players, tournaments and all types of area where sport are the primary category.

From the wh-question we can identify the POS tags which describes in the previous section this paper. In below example, if we can identify the sentence by whom, when and what we can get the result. Our algorithm was not efficient enough to identify all the words but small types of sentences can be identified and get the result of POS tags and its expression.

A full parse tree with details breakdown shown below -

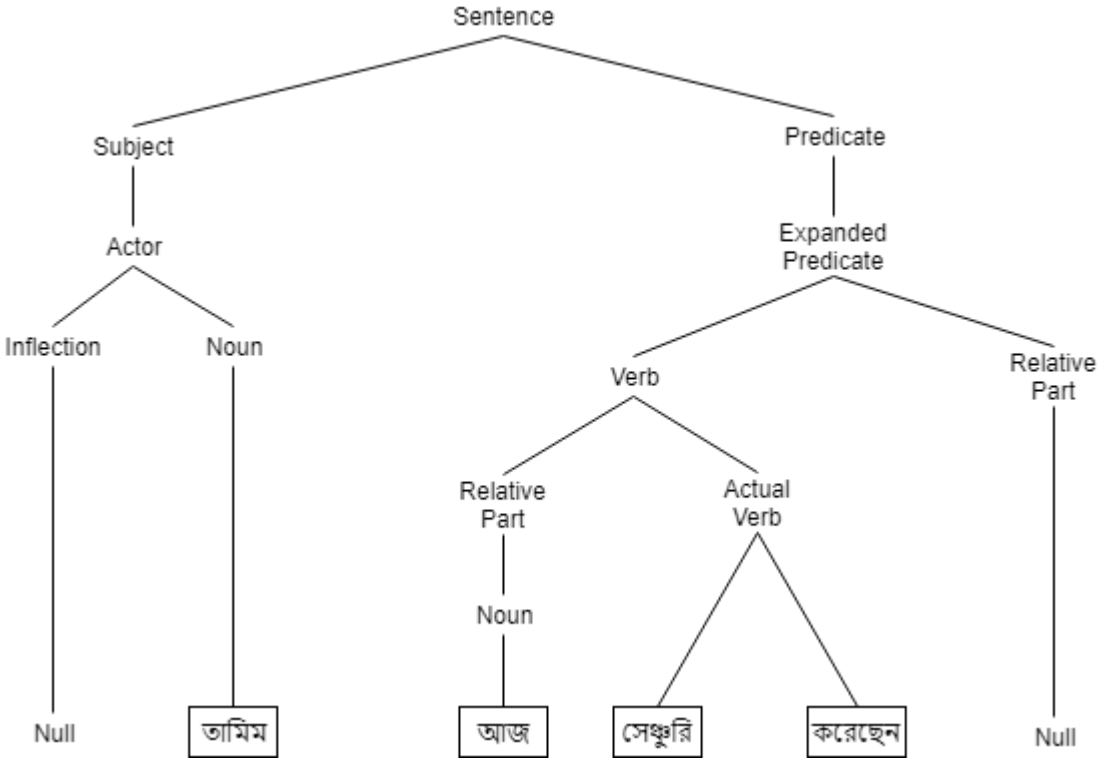

Figure 4-A: Sentence identification parse tree

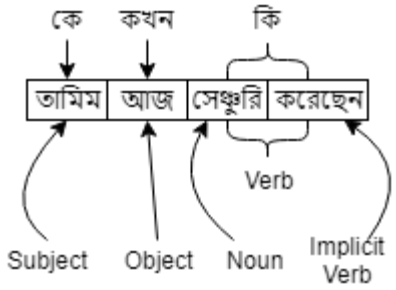

Figure 4-B: Sentence identification

4.2 Position of Verb in Bangla Sentence

Verb is the most difficult to detectable part-of-speech in Bangla sentence [5]. We have found that almost every Bangla sentences verb position hold the last position. It is quite easy to

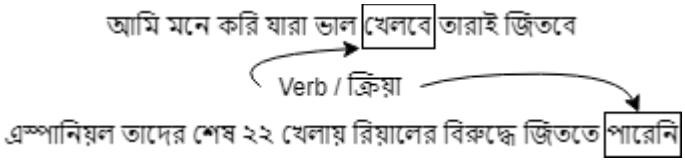

Figure 4-C: Position of verb in Bengali sentence

identify the verb but in Bengali verbs are formed with 2 or more words. So, it's very difficult to understand the expression and also harder to make a link in machine learning. In English this verb position varied from sentence to sentence and way of we express itself.

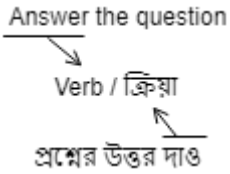

Figure 4-D: Verb position in Bangla & English

4.3 Data Analysis

For our thesis we have done quantitative analysis using various methods of statistics. We have done the total statistical analysis of our corpus, word frequency analysis and the most frequent used raw words in the corpus

Number of sources	3
Number of source type	1
Total raw words in DB	300138
Total number of sentences	28531
Total number of articles	1475
Data collection range	4 months
Total number of POS tags	6
Total number of respondents who were participated in 1st phase annotation	30-40
Total number of respondents who were participated in 2nd phase annotation	3

Table 4-A: Overall Statistics

Noun	61.94
Verb	15.94
Adjective	7.61
Preposition	1.35
Adverb	4.71
Pronoun	6.39
Undefined	2.06

Table 4-B: Most frequent POS tags in corpus

Chapter 5: Limitations

The major problems in the process of POS tagging are: Ambiguous words and unknown words. The first and foremost problem is with those words whose more than one tag can exist. This problem can be solved by emphasizing on context rather than single words. These can an easy task for humans but not so for the automatic word taggers. In the process of tagging we can sometimes get such words that have different tag categories when they are used in different context. Thus, it is a very tedious job. This phenomenon is known as lexical ambiguity. But while occupying the same part of speech many words can have multiple meanings.

Ambiguous words are the major problem in the part of speech tagging. Many words can have tags which are more than one. Some words can have different meaning in different

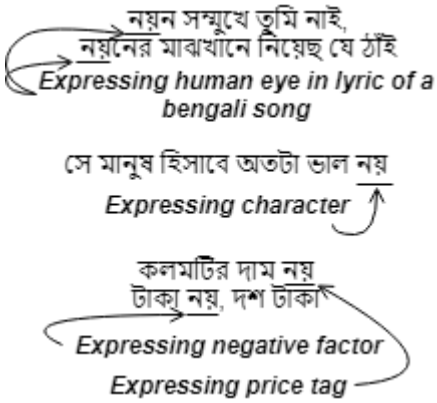

Figure 5-A: Variation of word expression in Bangla sentence

context but they have same POS. In order to solve such problem single word is considered rather than the context.

Our data source were some Bengali newspaper and all the sentences and words we got wasn't in the correct format. So, we had to make them structured and represent as such a format that human can get understood about the scenario. So, there might be some cases we don't do it the proper way as the process was done manually.

Again, we were highly dependent on our justifiers. Though Bengali is our mother tongue, but a few people work with the core structure of it. So, after the data collection and the survey, we took the sample annotations to our experts from different university and colleges. They helped us making the corrections on the data we annotated. There might be a chance that the experts made a mistake justifying the right form.

We got a lot of words that are not from the origin of Bengali. As the area of our research was based on sports we got such words that were introduced a few days ago. Most of the

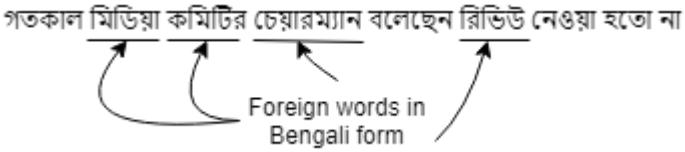

Figure 5-B: Foreign words in Bengali form

words from our source was a mixed-up origin. So, it was difficult to find the correct POS tag as some of them were new to us.

As our domain was based on sports, most of the tag or word we found was noun. So, the POS tagging result wasn't that much accurate as we expected. We expected to have a variation of different POS tag, but it didn't according to the plan as we got huge number of

সিলেট সিক্সার্সের মালিক অর্থমন্ত্রী আবুল মাল আব্দুল মুহিতের পরিবার

Figure 5-C: Noun case sentences in sports domain

noun tag while finding the tags. For example, in above example we could not detect any other POS tags rather than noun but the translation of this sentence in English (*Finance minister Abul Mal Abdul Muhit's family are the owner of Sylhet Sixers*) clearly shows other kinds of POS tags which is much difficult to find in Bengali sentence but both of the English and Bangla grammar are correct to express its form.

Some of the sentences finished up with no meaningful context. Some of them were assertive. It was difficult to generate the rule we got applied in our working process as it works for the general format of the sentence. Therefore, for the assertive sentences and the sentences with no subject visible may not result out in our expected way.

Chapter 6: Future work & Conclusion

6.1 Future Work

This thesis addresses the technique to improve the quality of POS tagging by finding the possible changes of a word. In the others direction how, a word may vary in a sentence. The main limitation of the approach presented here is that it will not result well if the corpora is too little. That means if the number of word variation of a word is few or little, the result may not come according the expected output.

There are a number of possible directions for future work, based on the findings in this thesis. Some of the directions are given bellow -

Increasing the size of parallel corpora always help to improve the accuracy of the system. Adding different sentence structures and handling the idioms and phrases externally would help to improve the system.

The reordering rules by which a subject of a sentence can be measured as well as the POS tags suggested in this thesis are relatively simple, and do not perform very well on large sentences. It would be possible to replace the handcrafted rules with automatically learned rules.

Word prediction of possible word can be detect from our research. Generating rules with stemming and lemma development can also help detect similar wordform.

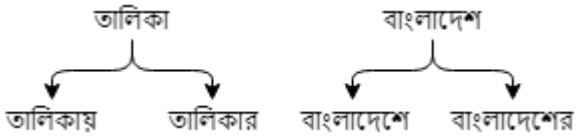

Figure 6-A: Word possibilities detection

In future, advancement of this system lies in integrating with speech recognition systems. Additionally, there is a possibility to convert this system into mobile environment for translating simple sentences.

It would be useful to perform a thorough error analysis of the translation output. Such an analysis would give improvements in future.

6.2 Conclusion

In this paper, we have investigated the possible way of Bengali POS tagging. Our implementation of the model was not effective due to our choice of training sets as well as the cleanliness of our data. Automatic POS tagging makes errors because many high frequency words of part-of-speech are ambiguous. Statistical tagging assigns a word its most likely tag, based on the n-set values frequencies in a training corpus.

References

[1] Ali, Hammad. "An unsupervised parts-of-speech tagger for the bangla language." *Department of Computer Science, University of British Columbia* 20 (2010): 1-8.

[2] Mehedy, Lenin, SM Niaz Arifin, and M. Kaykobad. "Bangla syntax analysis: A comprehensive approach." In *Proceedings of International Conference on Computer and Information Technology (ICCIT), Dhaka, Bangladesh*, pp. 287-293. 2003.

[3] Hoque, Mohammed Moshiul, and Muhammad Masroor Ali. "Context-sensitive phrase structure rule for structural representation of Bangla natural language sentences." *Proceedings of ICCIT, Dhaka* (2004): 615-620.

[4] Chakrabarti, Debasri, and Pune CDAC. "Layered parts of speech tagging for bangla." *Language in India* 11, no. 5 (2011).

[5] Das, Amitava, and Sivaji Bandyopadhyay. "Morphological stemming cluster identification for Bangla." *Knowledge Sharing Event-1: Task* 3 (2010).

[6] Chowdhury, Md Shahnur Azad, NM Minhaz Uddin, Mohammad Imran, Mohammad Mahadi Hassan, and Md Emdadul Haque. "Parts of speech tagging of bangla sentence." In *Proceeding of the 7th International Conference on Computer and Information Technology (ICCIT), Bangladesh*. 2004.

[7] Ali, Md Nawab Yousuf, SM Abdullah Al-Mamun, Jugal Krishna Das, and Abu Mohammad Nurannabi. "Morphological analysis of bangla words for universal networking language." In *Digital Information Management, 2008. ICDIM 2008. Third International Conference on*, pp. 532-537. IEEE, 2008.

[8] Mridha, Muhammad F., Aloke Kumar Saha, and Jugal Krishna Das. "Solving Semantic Problem of Phrases in NLP Using Universal Networking Language (UNL)." *International Journal of Advanced Computer Science and Applications (IJACSA), Special Issue on Natural Language Processing (NLP)* (2014).

[9] Anwar, Md Musfique, Mohammad Zabed Anwar, and Md Al-Amin Bhuiyan. "Syntax analysis and machine translation of Bangla sentences." *International Journal of Computer Science and Network Security* 9, no. 8 (2009): 317-326.

[10] Dandapat, Sandipan, Sudeshna Sarkar, and Anupam Basu. "A Hybrid Model for Part-of-Speech Tagging and its Application to Bengali." In *International conference on computational intelligence*, pp. 169-172. 2004.

[11] Majumder, Khair Md, and Yasir Arafat. "Analysis of and observations from a Bangla News Corpus." (2006).

[12] Chaudhury, Tasnim Haider, Abdul Matin, M. S. Hossain, Asie Uzzaman, and Md Masum. "Annotated Bangla News Corpus and Lexicon Development with POS Tagging and Stemming." *Global Journal of Research In Engineering* (2017).

[13] Ekbal, Asif, and Sivaji Bandyopadhyay. "Web-based Bengali news corpus for lexicon development and POS tagging." *Polibits* 37 (2008): 21-30.

[14] Piantadosi, Steven T. "Zipf's word frequency law in natural language: A critical review and future directions." *Psychonomic bulletin & review* 21, no. 5 (2014): 1112-1130.

[15] Islam, Md Asfaqul, KM Azharul Hasan, and Md Mizanur Rahman. "Basic HPSG structure for Bangla grammar." In *Computer and Information Technology (ICCIT), 2012 15th International Conference on*, pp. 185-189. IEEE, 2012.

[16] Bharati, Akshar, Dipti Misra Sharma, Lakshmi Bai, and Rajeev Sangal. "Anncorra: Annotating Corpora." *LTRC, IIIT, Hyderabad* (2006).

[17] " The Natural Language Processing Dictionary ". 2018. *Cse.Unsw.Edu.Au*. Accessed December 23 2018. http://www.cse.unsw.edu.au/~billw/nlpdict.html.

[18] Mahmud, Md. Redowan, Mahbuba Afrin, Md. Abdur Razzaque, Ellis Miller, and Joel Iwashige. "A Rule Based Bengali Stemmer." *2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 2014. doi:10.1109/icacci.2014.6968484.

[19] Haque, Md, Md Habib, and Md Rahman. "Automated word prediction in bangla language using stochastic language models." *arXiv preprint arXiv:1602.07803* (2016).

[20] Chakrabarty, Abhisek, Akshay Chaturvedi, and Utpal Garain. "A Neural Lemmatizer for Bengali." In *LREC*. 2016.

[21] Ekbal, Asif, Rejwanul Haque, and Sivaji Bandyopadhyay. "Bengali part of speech tagging using conditional random field." In *Proceedings of the seventh International Symposium on Natural Language Processing, SNLP-2007*. 2007.

[22] Hasan, Fahim Muhammad, Naushad UzZaman, and Mumit Khan. "Comparison of different POS Tagging Techniques (N-Gram, HMM and Brill's tagger) for Bangla." In *Advances and innovations in systems, computing sciences and software engineering*, pp. 121-126. Springer, Dordrecht, 2007.

[23] Ekbal, Asif, Rejwanul Haque, and Sivaji Bandyopadhyay. "Maximum entropy based bengali part of speech tagging." *A. Gelbukh (Ed.), Advances in Natural Language Processing and Applications, Research in Computing Science (RCS) Journal*33 (2008): 67-78.

[24] "Bengali Language". 2018. *En.Wikipedia.Org*. Accessed December 23 2018. https://en.wikipedia.org/wiki/Bengali_language.

[25] "List of Languages By Number Of Native Speakers". 2018. *En.Wikipedia.Org*. Accessed December 23 2018 https://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers.

[26] "Universal Networking Language". 2018. *En.Wikipedia.Org*. Accessed December 23 2018. https://en.wikipedia.org/wiki/Universal_Networking_Language.

[27] "UNL.RU :: Introduction Of The UNL System". 2018. *Unl.Ru*. Accessed December 23 2018. http://www.unl.ru/system.html.

Appendix: Acronyms and Abbreviations

AI	Artificial Intelligence
CFG	Context Free Grammar
CIIL	Central Institute of Indian Languages
CSG	Context Sensitive Grammar
FIRE	Forum for Information Retrieval Evaluation
HMM	Hidden Markov Model
HPSG	Head-Driven Phrase Structure Grammar
LFG	Lexical Functional Grammar
NLP	Natural language processing
POS	Part-of-speech
SVM	Support Vector Machine
UNL	Universal Networking