|  |
| --- |
| **Chapter 3: Methodology** |

* 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.

* 1. **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

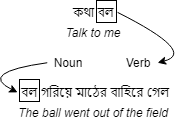


Figure 3-A: Difference of word identity in sports domain

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.

* 1. **Storing the Data into Database**
     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

* + 1. **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.

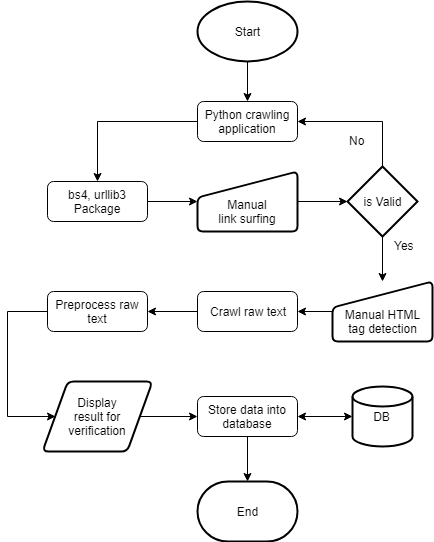


Figure 3-C: Work-Flow of web crawling

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.

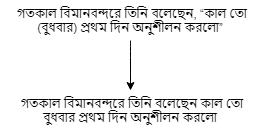


Figure 3-D: Excluding expression for generating tokens

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.

* 1. **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.

* + 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

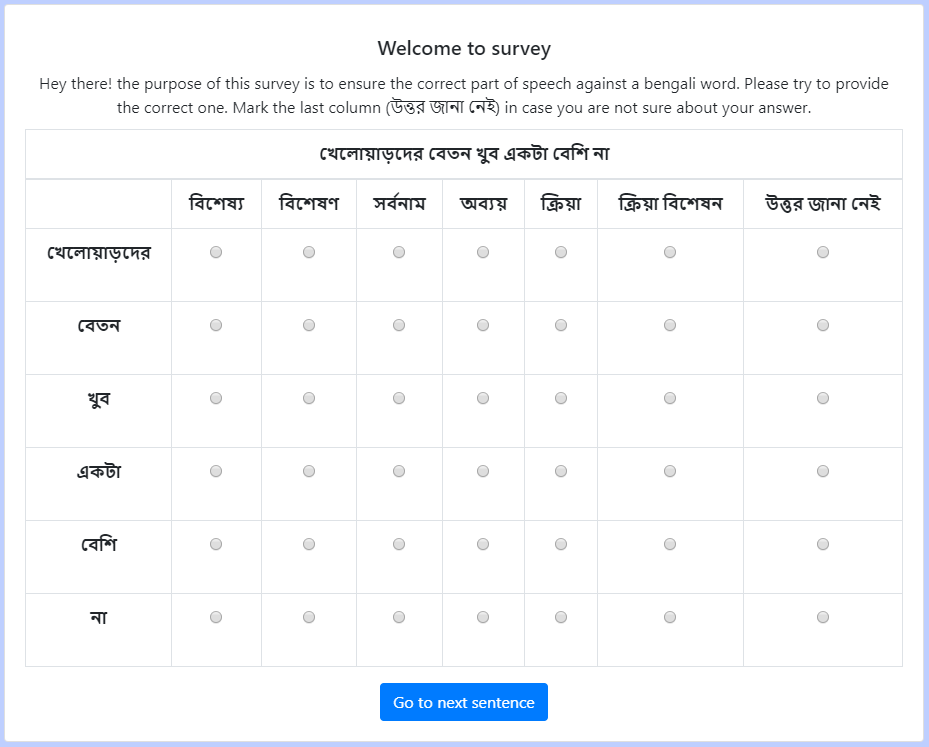


Figure 3-E: Online crowdsourcing GUI

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.

* + 1. **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.

* 1. **Data Modelling**
     1. **Relational Database**

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

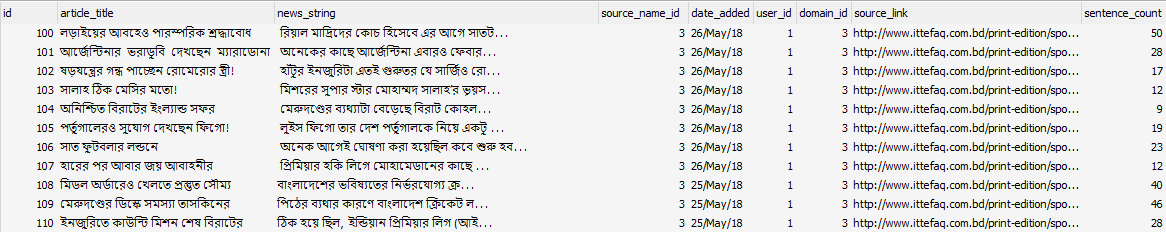
****

Figure 3-F: article table

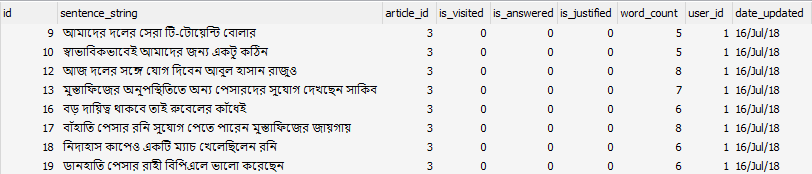
****

Figure 3-G: article\_line table

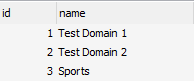
****

Figure 3-H: domain table

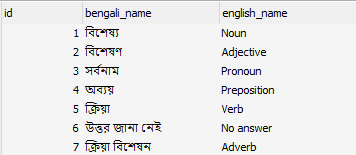
****

Figure 3-I: pos table

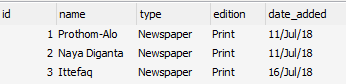
****

Figure 3-J: source\_name table

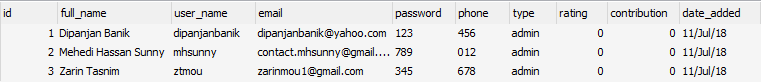
****

Figure 3-K: user\_information table

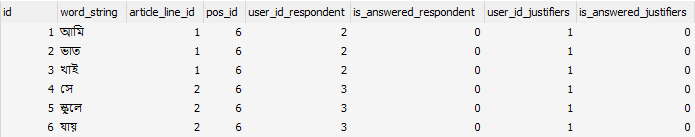
****

Figure 3-L: word table

* 1. **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.

* 1. **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

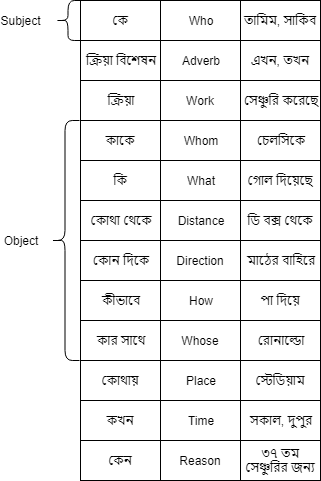


Figure 3-M: Table for sentence identification

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.