|  |
| --- |
| **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

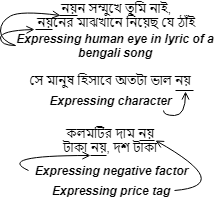


Figure 5-A: Variation of word expression in Bangla sentence

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

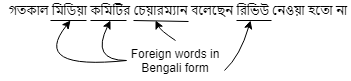


Figure 5-B: Foreign words in Bengali form

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



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.