An In-depth Study on POS Tagging for Assamese Language

Dr. Karabi Kherkatary Boro¹, Dr. Uzzal Sharma²

¹Assistant Professor, Department of Assamese, Pragjyotish College, Guwahati, India
karabboro1185@gmail.com

²Assistant Professor, Department of Computer Applications, Assam Don Bosco University, Guwahati, India
druzzalsharma@gmail.com

Abstract: An automatic POS tagger is an essential component of any Natural Language Processing (NLP) work. It is one of the important steps towards the processing of Natural Language. There are various challenges in the tagging of POS and most of the time these are language-dependent. Assamese is a morphologically rich and free word order language. Because of this, the challenges are even more. In the present paper, the basic concept of the POS tagger and its importance in the NLP is discussed. In the later part of the paper, the overall characteristics of the Assamese language are discussed in short and its various challenge that may arise towards the tagging of POS is discussed. The paper also discusses about the various POS tagging techniques that are commonly used in the tagging of POS.

Keywords: POS Tagging, MRL, Free Word order Language, Word stripping, tagset.

(Article history: Received: 4th August 2020 and Accepted 7th December 2020)

I. INTRODUCTION: WHAT IS PARTS OF SPEECH (POS) AND ITS TAGGING

Words are the constituents of a sentence. Each word in a sentence is classified into different word classes. According to the use and its placement in a sentence, words are divided into different classes called Parts-of-Speech (POS) or word classes or syntactic categories. It is very much necessary to have different classes of words in a sentence to form a sentence and to give its appropriate meaning or semantics. In fact, the parts-of-speech works as the building blocks of the grammar of any language, more specifically, in the linguistic part of the syntax[1]. The grammar of any language is very important and it depends heavily on the parts-of-speech to produce required semantic expressions, both verbal as well as written[2][3]. It can also be said that parts-of-speech may be considered as the life-blood of any language in that, a lack of a sequence of appropriate words representing parts-of-speech would results in the spoken or written words meaningless[4].

Parts-of-Speech are very important as it reveals all the information about a word and its neighbor. In fact, POS helps us to get information about neighboring words knowing the class of a word. For example, if a word is a noun it is preceded by either a determiner or an adjective, if the word is a verb, it is preceded by a noun.

Parts-of-Speech(POS) tagging is a process of classifying each word in a sentence into one of the Parts-of-Speech belongs to that language. In simple words, we can say that tagging is a task, where each word in a sentence is labeled into the appropriate Parts-of-Speech. POS tagging is very much necessary for parsing that is the process of determining the syntactic structure of a sentence. It means parsing is followed by POS tagging. Parsing is one of the very important steps towards the processing of Natural Language. But as a preliminary step, tagging of Parts-of-Speech is carried out, and this is generally performed by the language experts of the language under consideration in manual process or automatically by a program in computer based on the already prepared tagset for that specific language.

II. POS TAGGING AND ITS IMPORTANCE IN NLP

Natural Language Processing (NLP) is the ability of a computer system to process Natural Languages. Natural Languages are languages, which are spoken by human beings. Although, there are grammars associated with each language having different syntactic structures which are all grammatically correct can results in the same semantic meaning[5][6]. Because of not having a one-to-one relationship between syntactic and semantic, the processing of natural languages by a computer system is considered as a real challenge [7]. This is because, to understand and process different syntactic structures, the system is also required to be that much versatile.

Figure1: Natural Language Processing

In NLP, we develop a program, whose purpose is to tag each and every word of the corpus under study. The process of tagging is based on the Parts-of-Speech (POS) of the language under study[8]. It is known as POS tagging which
is a process of automatic assignment of lexical categories or word class markers such as Noun, Pronoun, Verb, Adjective, Adverb, etc. for each word in a sentence according to the grammar of that natural language. In NLP, POS tagging is carried out to remove any kind of disambiguates among the words.

The placement of that word into a lexical category. As an example, we can talk about the Assamese word "কর"/kôr/, which is used as a verb as well as a noun. When it is used as a verb, it means "to do" and when it is used as a noun, it means "hand".

3. The training data used to train the system plays a vital role. If it is small in size, the POS tagger will not be considered as efficient.

III. CLASSIFICATION OF POS TAGGER

There are several approaches for POS tagging. The Figure 4 describes the different POS tagging models normally used.

Although there are many models which can be used for the classification of POS taggers [9][10][11], in the current paper, some of the commonly used are discussed below which are most commonly used in case the Assamese Language:

Supervised models

In these models, a pre-annotated corpus is required which is used to train the model to learn information about the tagset, word-tag frequencies, rule sets, etc. [12][25]. The accuracy of the model depends on the size of the corpus. When the size of the corpus is large, it gives better accuracy and at the same time takes a long time for training and prediction.
Unsupervised models
In comparison to the supervised POS Tagging model, the unsupervised POS Tagging models do not require a pre-annotated corpus. Instead, they are more sophisticated and uses advanced computational techniques such as the Baum-Welch algorithm to automatically induce tagsets, transformation rules, etc. Once the tagsets and information rule is set, they either calculate the probabilistic information which is required by the stochastic taggers or apply contextual rules required by rule-based systems or transformation-based systems [12], [13].

Both the supervised and unsupervised models can be further classified into the following categories.

Rule-based and transformation-based models
In case of the rule-based POS Tagging models, there is a set of handwritten rules and uses contextual information to assign POS tags to words. These rules are often known as context frame rules set by linguistic experts. On the other hand, in the case of the transformation-based model a predefined set of handcrafted rules as well as automatically-induced rules that are generated during training is applied. Apart from it, some model also used morphological rules [14], capitalization and punctuation, etc. [13].

Stochastic models
In case of the stochastic models the frequency, probability, or statistics are used. It is based on different methods such as Hidden Markov Models (HMM), maximum-likelihood estimation (MLE) or n-grams. The HMM-based methods require the evaluation of the "argmax" formula, which is very expensive in terms of time complexity as in this case all the possible tag sequences need to be checked, in order to find the sequence that maximizes the probability. So, in HMM-based approached, a dynamic programming model known as the Viterbi Algorithm is used to find the optimal tag sequence [15].

And lastly, the neural networks [16][17] is applied for both supervised and unsupervised POS Tagging models.

IV. POS IN ASSAMESE LANGUAGE
The grammatical information of any language is determined by the tags. All the tags belongs to a language is represented as a set known as tagset. The size of a tagset is very important as far as the versatility of the tagset is concerned. The efficiency and quality of the research depends on its tagset size. In the table (Table1), the Assamese version of the BIS Tagset is shown along with example of each tag type. Although the Table1 shows a tagset based on BIS tagset, it is not the complete one. There are many tags still need to be added to it and it is an ever-growing process as any language become richer and richer when it come in contact with other languages. In the present research, the authors tries to identify the most important and frequently occurring tags to represent the tagset.

<table>
<thead>
<tr>
<th>Slno</th>
<th>POS</th>
<th>Category</th>
<th>Label</th>
<th>Annotation Conversion</th>
<th>Assamese</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Noun(N)</td>
<td>Common</td>
<td>NN</td>
<td>N_NN</td>
<td>আড়িকাছাঁ বিশেষ্য</td>
<td>মানুহ /manuh/, তৰা /tora/</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Proper</td>
<td>NNP</td>
<td>N_NNP</td>
<td>ব্যক্তিবাচক বিশেষ্য</td>
<td>বম /ram/, ব্যক্তিবাচকবর্ণ /brhmmputro/</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Collective</td>
<td>SN</td>
<td>N_SN</td>
<td>সমাহিতবাচক বিশেষ্য</td>
<td>সভা /xobby/, পুরীডাল /pprioria/</td>
</tr>
<tr>
<td>2</td>
<td>Pronoun(PP)</td>
<td>Personal</td>
<td>PRP</td>
<td>P_PRP</td>
<td>ব্যক্তিবাচক সর্বনাম</td>
<td>মই /moi/, তই /tai/, তুমি /tumi/, আলৈকে /apuni/, আলৈকে /apuni/, হি /xi/, তাই /tai/</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Reflexive</td>
<td>PRF</td>
<td>P_PRF</td>
<td>আয়াচাঁক সর্বনাম</td>
<td>নিউ /nij/, যয় /swayng/, আলৈকে /apuni/, আলৈকে /apuni/, হি /xi/, তাই /tai/</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Relative</td>
<td>PRL</td>
<td>P_PRL</td>
<td>সম্প্রতিবাচক সর্বনাম</td>
<td>হি /ji/, মিহিউ /jih/</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Interogative</td>
<td>PRQ</td>
<td>P_PRQ</td>
<td>প্রশ্নবাচক সর্বনাম</td>
<td>কি /ki/, কিস /kis/, কোন /kon/</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Inclusive</td>
<td>INC</td>
<td>P_INC</td>
<td>সাকলবাচক সর্বনাম</td>
<td>সকলা /xnklo/</td>
</tr>
<tr>
<td>3</td>
<td>Verb (V) ক্রিয়া</td>
<td>Auxiliary</td>
<td>VAUX</td>
<td>V_VAUX</td>
<td>সহায়কী ক্রিয়া</td>
<td>(ই) বই /khai/, পাই /goi/, বহি (বহ-ই) /bohi/, পটিত (পুঠিত) /porhibn/ (/khujisö), কাহা (কথা) /kova//kotha/</td>
</tr>
<tr>
<td>4</td>
<td>Adjective (JJ) বিশেষ</td>
<td>--</td>
<td>JJ</td>
<td>JJ</td>
<td>বিশেষ</td>
<td>গরম /gorrom/, দুয়া অৰু /doyalu/, ফিকিফিয়া /kiskisia/</td>
</tr>
<tr>
<td>5</td>
<td>Adverb (RB) ক্রিয়া বিশেষ</td>
<td>--</td>
<td>RB</td>
<td>RB</td>
<td>ক্রিয়া বিশেষ</td>
<td>থরব/ /khkori/, পোকাকাল /xonkale/, ওছবে-কাজবে /osore - pajore/, তালে /talhe/</td>
</tr>
<tr>
<td>6</td>
<td>Postposition (PSP) অনুদ</td>
<td>--</td>
<td>PSP</td>
<td>PSP</td>
<td>অনুপদ</td>
<td>অর্থ /nrthe/, ঘাবা /dwa/, চাবা /son/, পাই /pai/, হুব /hnbla/, চাসে /sako/</td>
</tr>
<tr>
<td>7</td>
<td>Conjunction (CC) সংযোগক</td>
<td>Co-ordinator</td>
<td>CCD</td>
<td>CC_CCD</td>
<td>সংযোগক</td>
<td>অনুৰ /aru/, অথচ /thnso/, তেলে /tente/</td>
</tr>
<tr>
<td></td>
<td>Subordinator</td>
<td>CCS</td>
<td>CC_CCS</td>
<td>সমন্বয়ক</td>
<td>ধনি /jodio/, তথ্যি /tothapi/, নিছু /kintu/, অথা /nthnba/</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Indeclinable (RP) অবয়</td>
<td>Interjection / Exclamatory</td>
<td>INJ</td>
<td>RP_INJ</td>
<td>ভাবোন্ধক অনুর</td>
<td>বাস: (কি বুদ্বন) /bah/ (/ki kundor/), হাস হাসা (অকাল কাল) /hai hai!/ (/lokalote dhukal/), চিহ্ন (বর্ন লোক) /sibh/ (/bor letera/), আও! (ঈহং বব সাম) /aor/ (/iyar bor xah/</td>
</tr>
</tbody>
</table>
V. POS Tagging and Its Challenges in Assamese Language

The Assamese language is one of the rich languages in terms of the types of word class or POS present in it. At the same time, Assamese is also a Morphologically Rich Languages (MRL). It belongs to the Indo-Aryan class of Languages along with other predominant languages such as Bengali, Hindi, etc[24].

As far as the Assamese language is concerned, there are many challenges, that make the tagging of POS very difficult and complex. Like the other language of Indo-Aryan class of Languages, the Assamese is also a free word order language. A language is said to be a free word order language if the placement of words in the formation of a sentence has negligible influence as far as the semantics of the sentence is concerned. In the Assamese, although, the predominant word order is subject(S), verb (V), and object (O) that is SVO, but we can see some other word orders as well in the construction of sentences. They are: SVO, SOV, VSO, VOS, OVS, and OSV. Out of these six, the first three that is SVO, SOV, VSO are mostly dominant in the formation of the sentence. For example, in the sentence

"মোই ভাত খাও।" /Moi bhat khao/ and
"ভাত মোই খাও।" /Bhat moi khao/, there is hardly any difference in the semantic. This issue may create a problem as far as the tagging of POS is concerned [19][20].

Apart from it, there are some other issues, which makes, the POS tagging a very critical task. Firstly, we can talk about "Ambiguous words". Ambiguous words create a major problem as they can be categorized in different word classes[21]. There are many words, which can have more
than one tag. This incidence is known as "Lexical ambiguity". For example, sometimes the same word such as "mark" in English represents as a 'noun' and 'verb', in a sentence like:

'Mark was marked as a criminal'

Similarly, in Assamese, sentences such as "মুদ্রা একমুদ্রা মুদ্রা না।" /Xund ejon xundor tara/, the word "মুদ্রা" /Xundor/ is categorized into proper-noun (the first "মুদ্রা" /Xundor/) and adjective (the second "মুদ্রা" /Xundor)/). In the same line, the word "কলা" /kola/, may function as Adjective(JJ) as well as Noun(N). For example,

- কলা /kola/ (Adjective(JJ)) - মানুষ কলা। /manuhjon kola/- The man is deaf.
- কলা-কৌশল /kola-kouxol/ (Adjective(JJ)) - সি ভাষা কলা-কৌশল যথাই আছে। /xitar kola-kouxol dekhai ase/- He is showing his skills.
- কলা-সংস্কৃতি/kola-xonskriti/- অসম কলা-সংস্কৃতি/অসম কলা। /xomor kola-xonskriti bareroheniya/- The Assamese culture is versatile.

in the above example, in the first two sentences, the word ‘কলা’ /kola/, the word is used as Adjective(JJ) and in the next two sentences, the same word ‘কলা’ /kola/, is functioning as a noun. So, during the tagging process, the automatic POS tagger, may try to place the word, either in the noun category or in the adjective category. Similarly, for the word ‘কব’ /kor/:

- কব /kor/ (flower bud)/(N) - ফুলের কবটা বড় ধূর্ণীয়। /phulor korte bor dhuniya/- The flower bud is very beautiful.
- কব /kor/ (tax) (N) - কব সময় পরিষোধ কব উচিত। /kor xomoymate porixodh kora usit/- Tax should be paid on time.
- কট-কব (Troublesome) (Adjective(JJ)) - পাহাড় বড়ার বড় কটী-কব।/pahar bogowa bor kosokor/ - The climbing of hill is very difficult.

in the first two sentence, the ‘কব’ /kor/, is functioning as noun and in the third sentence, it is functioning as adjective(JJ).

In the following sentences, the word ‘বিষ’ (bix),

- বিষ - (Noun(NN)) - poison - তেঁতু হাতের বিষ রাখে। /teor hator bix hoise/- He is having pain in his hand.
- বিষ-থাই (Adjective (JJ)) - One which has poison(eg. snake) - সাপদল না বিষ-থাই আছে। /xapdal bor bixodhor asil/- The snake was very poisonous.

is functioning as a noun whereas in the second sentence, the same word is functioning as Adjective(JJ).

In handling the above situations, we need to consider the individual words rather than the sentence, while POS tagging is carried out.

Another challenging task that we can see in the following example for the word "কলা" /kola/,

- কল /kol/(Industry) (NN) - জাগিোদ কাজ কল আছে। /jagirodot kajg kol ase/- There is a paper mill in Jagiroad.
- কল /kol/(fruit) (NN) - কল অতি পুষ্টিকৃত ফল। /kol oti pustikor phol/- Banana is very nutritious fruit.
- কল /kol/ (switch)(NN) - আকাশবাণীর কল টো টিপি দিয়ে। /akabanir koltu tipi dyok/- Switch on the radio.[23]

To identify, the root word in a sentence, Suffix stripping is used[22]. The process basically, strips off the suffix from a word, to extract the root word. For example, the root word ভাত /bharot/ may have a different suffix such as

- ‘ীয়’ /-iyo/ (ভাতীয়) /bharotiya/,
- ‘বাসী’ /-baxi/ (ভাতবাসী) /bharotbaxi/,
- ‘রন্ধ’ /rotno/ (ভাতরন্ধ) /bharotrotno/,

suffix stripping is a step, which is carried out before the POS tagging is done so that the word can be placed in the appropriate class. But as it is done most of the time automatically, there is a chance of incorrect stripping output which is reported by the researchers[Q]. For example, the word মাজুলী /majuli/ may be wrongly stripped of to মাজু /majul/, same is the case with the word ‘দিল’ /dil/ be stripped off as দি /dillo/ [18][22].
Based on the above circumstances, it can be said that the tagging of POS in the Assamese Language is one of the challenging task, which needs to be taken care of very carefully. Otherwise, the tagging process may result in an error.

In the above example, although, they all fall under the same POS tag, that is a noun, but gives different meanings in different contexts. So, while tagging is carried out, the context in which it is used needs to be considered.

VI. CONCLUSION

Although the POS tagging process is carried out automatically, there are many challenges that need attention and human intervention. Especially in languages like Assamese, which is morphologically rich and a free word order language, the challenges are even more. The need for the design of a versatile POS tagger is very much felt. Although, there are many automatic POS taggers for the Assamese language are there, and performing their job very well, but still, the need for a complete automatic POS tagger is always felt, which can handle all the aspects of the language. In the present paper, the authors try to showcase the different challenges that need to overcome, while designing an automatic POS tagger.

REFERENCES


[10] Yair Halevi, “Part of Speech Tagging”, Seminar in Natural Language Processing and Computational Linguistics (Prof. Nachum Dershowitz), School of Computer Science, Tel Aviv University, Israel, April 2006.


AUTHOR PROFILE

Dr. Karabi Kherkatary Boro is presently working as an Assistant Professor in the Department of Assamese at Pargjyotish College, Guwahati, India. She has more than 12 years of teaching and research experience. She has published more than 15 papers in Books and Journals and presented more than 10 papers in International and National conferences. She has a book to her credit as an Author and also co-edited a book.

Dr. Uzzal Sharma is currently working as an Assistant Professor, Stage 2 in the Department of Computer Applications at Assam Don Bosco University, Guwahati, India. He has over 16 years of experience in the field of teaching and research. He has published more than 19 SCOPUS indexed papers and presented many papers in international conferences. He also has a patent to his credit. He has written 5 books as sole author, published by national and international publishers and contributed many book chapters. Dr. Sharma has produced 2 Ph.D. scholars and currently guiding a few scholars for their Ph.D. research.