Part of Speech Tagger for Assamese Text

Navanath Saharia  
Department of CSE  
Tezpur University  
India - 784028

Dhrubajyoti Das  
Department of CSE  
Tezpur University  
India - 784028

Utpal Sharma  
Department of CSE  
Tezpur University  
India - 784028

Jugal Kalita  
Department of CS  
University of Colorado  
Colorado Springs - 80918

{nava_tu,dhruba_it06,utpal}@tezu.ernet.in  
kalita@eas.uccs.edu

Abstract

Assamese is a morphologically rich, agglutinative and relatively free word order Indic language. Although spoken by nearly 30 million people, very little computational linguistic work has been done for this language. In this paper, we present our work on part of speech (POS) tagging for Assamese using the well-known Hidden Markov Model. Since no well-defined suitable tagset was available, we develop a tagset of 172 tags in consultation with experts in linguistics. For successful tagging, we examine relevant linguistic issues in Assamese. For unknown words, we perform simple morphological analysis to determine probable tags. Using a manually tagged corpus of about 10000 words for training, we obtain a tagging accuracy of nearly 87% for test inputs.

1 Introduction

Part of Speech (POS) tagging is the process of marking up words and punctuation characters in a text with appropriate POS labels. The problems faced in POS tagging are many. Many words that occur in natural language texts are not listed in any catalog or lexicon. A large percentage of words also show ambiguity regarding lexical category. The challenges of our work on POS tagging for Assamese, an Indo-European language, are compounded by the fact that very little prior computational linguistic exists for the language, though it is a national language of India and spoken by over 30 million people. Assamese is a morphologically rich, free word order, inflectional language. Although POS tagged annotated corpus for some of the Indian languages such as Hindi, Bengali, and Telegu (SPSAL, 2007) have become available lately, a POS tagged corpus for Assamese was unavailable till we started creating one for the work presented in this paper. Another problem was that a clearly defined POS tagset for Assamese was unavailable to us. As a part of the work reported in this paper, we have developed a tagset consisting of 172 tags, using this tagset we have manually tagged a corpus of about ten thousand Assamese words.

In the next section we provide a brief relevant linguistic background of Assamese. Section 3 contains an overview of work on POS tagging. Section 4 describes our experimental setup. In Section 5, we analyse the result of our work and compare the performance with other models. Section 6 concludes this paper.

2 Linguistic Characteristics of Assamese

In Assamese, secondary forms of words are formed through three processes: affixation, derivation and compounding. Affixes play a very important role in word formation. Affixes are used in the formation of relational nouns and pronouns, and in the inflection of verbs with respect to number, person, tense, aspect and mood. For example, Table 1 shows how a relational noun (deutA: father) is inflected depending on number and person (Goswami, 2003). Though Assamese is relatively free word order, yet the predominant word order is subject-object-verb (SOV).

The following paragraphs describe just a few of the many characteristics of Assamese text that make the tagging task complex.

- Depending on the context, even a common word may have different POS tags. For example: If রক্ত (kArane), দোহ (dare), নিন্মীত (nimitte), হেতু (hetu), etc., are placed after pronominal adjective, they are considered conjunction and if placed after
nucleus error complex they are considered particle. For example,

এরা বালু খাঁ নামপালী।

TF: ei kArane mohi nagalos.
This + why + I+ did not go.
ET2: This is why I did not go.

রামের বাবা মহি গোলাল।
TF: rAmar kArane mohi nagalos.
Ram’s + because of + I+ did not go
ET: I did not go because of Ram.

In the first sentence বাবা (kArne) is placed after pronominal adjective তিনি (et); so kArne is considered conjunction. But in the second sentence kArne is placed after noun বাম (RAM), and hence kArne is considered particle.

• Some prepositions or particles are used as suffix if they occur after noun, personal pronoun or verb. For example,

দিলে তুবি | TF: sihe goisil.
ET: Only he went.

Actually হে (he : only) is a particle, but it is merged with the personal pronoun সি (si).

• An affix denoting number, gender or person, can be added to an adjective or other category word to create a noun word. For example,

ধুম অজনি হো আহিস আ।
TF: dhuniyAjonI hoi aAhisA.
ET: You are looking beautiful.

Here ধুমিয়া (dhuniyA : beautiful) is an adjective, but after adding feminine suffix হো female
the whole constituent becomes a noun word.

Table 1: Personal definitives are inflected on person and number

<table>
<thead>
<tr>
<th>Person</th>
<th>Singular</th>
<th>Plural</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st Person</td>
<td>My father</td>
<td>Our father</td>
</tr>
<tr>
<td>2nd Person</td>
<td>Your father</td>
<td>Your father</td>
</tr>
<tr>
<td>3rd Person</td>
<td>Her father</td>
<td>Their father</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Personal</th>
<th>Singular</th>
<th>Plural</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plural</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mother</td>
<td>মাতা মোড়ানা</td>
<td>মাতা মোড়ানা</td>
</tr>
<tr>
<td>Sister</td>
<td>বাঈ দুংষ্ট</td>
<td>বাঈ দুংষ্ট</td>
</tr>
<tr>
<td>Brother</td>
<td>ভাই দুংষ্ট</td>
<td>ভাই দুংষ্ট</td>
</tr>
<tr>
<td>Grandfather</td>
<td>বাপের বাপী</td>
<td>বাপের বাপী</td>
</tr>
<tr>
<td>Grandmother</td>
<td>বাপের বাপী</td>
<td>বাপের বাপী</td>
</tr>
</tbody>
</table>

• Even conjunctions can be used as other part of speech.

হবি অফ দুই ভাষকে কন্ট্রোল।

TF: hari aAru Jadu bhaayek kokAyek.
ET: hari and Jado are brothers.

তোরপালি বাতিল ধুমিয়ার ধুমিয়ার আহিক অধিক
বহস্বজনক কবি উল্লেখ।

TF: JowAAli rAir ghotonAtowoe bishoitok
aAru adhik rahashyayanak kori tulile.
ET: The last night incident has made the
matter more mysterious.

The word অফ (aAru : and) shows ambiguity
in these two sentences. In the first, it is used
as conjunction (i.e. hari and Jado) and in the
second, it is used as adjective of adjective.

3 Related Work

Several approaches have been used for building POS taggers. Two main approaches are supervised and unsupervised. Both supervised and unsupervised tagging can be of three sub-types. They are rule based, stochastic based and neural network based. There are number of pros and cons for each of these methods. The most common stochastic tagging technique is Hidden Markov Model (HMM).

During the last two
decades, many different types of taggers have been
developed, especially for corpus rich languages
such as English. Nevertheless, due to relatively
free word order, agglutinative nature, lack of
resources and the general lateness in entering the
computational linguistics field in India, reported
tagger development work on Indian languages
is relatively scanty. Among reported works,
Dandapat (2007) developed a hybrid model of
POS tagging by combining both supervised and
unsupervised stochastic techniques. Avinash and
Karthik (2007) used conditional random field and
transformation based learning. The heart of the
system developed by Singh et al. (2006) for Hindi
was the detailed linguistic analysis of morpho-
syntactic phenomena, adroit handling of suffixes,
accurate verb group identification and learning
of disambiguation rules. Saha et al. (2004)
developed a system for machine assisted POS
tagging of Bangla corpora. Pammari and Prahllad
(2007) developed a POS tagger and chunker
using Decision Forests. This work explored
different methods for POS tagging of Indian
languages using sub-words as units. Generally,
most POS taggers for Indian languages use

---

1TF: Transliterated Assamese Form
2ET: Aproximate English Translation
morphological analyzer as a module. However, building morphological analyzer of a particular Indian language is a very difficult task.

4 Our Approach

We have used a Assamese text corpus (Corpus Asm) of nearly 300,000 words from the online version of the Assamese daily Asomiya Pratidin (Sharma et al., 2008). The downloaded articles use a font-based encoding called Luit. For our experiments we transliterate the texts to a normalised Roman encoding using transliteration software developed by us. We manually tag a part of this corpus, Tr, consisting of nearly 10,000 words for training. We use other portions of Corpus Asm for testing the tagger.

There was no tagset for Assamese before we started the project reported in this paper. Due to the morphological richness of the language, many words of Assamese occur in secondary forms in texts. This increases the number of POS tags that needed for the language. Also, often there are differences of opinion among linguists on the tags that may be associated with certain words in texts. We developed a tagset after in-depth consultation with linguists and manually tagged text segments of nearly 10,000 words according to their guidance. To make the tagging process easier we have subcategorised each category of noun and personal pronoun based on six case endings (viz, nominative, accusative, instumental, dative, genitive and locative) and two numbers.

We have used HMM (Dermatas and Kokkinakis, 1995) and the Viterbi algorithm (1967) in developing our POS tagger. HMM/Viterbi approach is the most useful method, when pretagged corpus is not available. First, in the training phase, we have manually tagged the Tr part of the corpus using the tagset discussed above. Then, we build four database tables using probabilities extracted from the manually tagged corpus- word-probability table, previous-tag-probability table, starting-tag-probability table and affix-probability table.

For testing, we consider three text segments, A, B and C, each of about 1000 words. First the input text is segmented into sentences. Each sentence is parsed individually. Each word of a sentence is stored in an array. After that, each word is searched in the word-probability table. If the word is unknown, its possible affixes are extracted and searched in the affix-probability table. From this search, we obtain the probable tags and their corresponding probabilities for each word. All these probable tags and the corresponding probabilities are stored in a two dimensional array which we call the lattice of the sentence. If we do not get probable tags and probabilities for a certain word from these two tables we assign tag -N (Common Noun) and probability 2 to the word since occurrence of CN is highest in the manually tagged corpus. After forming the lattice, the Viterbi algorithm is applied to the lattice that yields the most probable tag sequence for that sentence. After that next sentence is taken and the same procedure is repeated.

5 Experimental Evaluation

The results using the three test segments are summarised in Table 2. The evaluation of the results require intensive manual verification effort. Larger training corpora is likely to produce more accurate results. More reliable results can be obtained using larger test corpora. Table 3 compares our result with other HMM based reported work. Form the table it is clear that

<table>
<thead>
<tr>
<th>Test Set</th>
<th>Size</th>
<th>Average accuracy</th>
<th>UDH accuracy</th>
<th>UPH accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>992</td>
<td>84.68%</td>
<td>62.8%</td>
<td>42.0%</td>
</tr>
<tr>
<td>B</td>
<td>1074</td>
<td>89.94%</td>
<td>57.54%</td>
<td>53.96%</td>
</tr>
<tr>
<td>C</td>
<td>1241</td>
<td>86.05%</td>
<td>83.64%</td>
<td>26.47%</td>
</tr>
</tbody>
</table>

Table 2: POS tagging results with small corpora. Size of training words : 10000, UWH : Unknown word handling, UPH : Unknown proper noun handling

<table>
<thead>
<tr>
<th>Author</th>
<th>Language</th>
<th>Average accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Toutanova et al.(2003)</td>
<td>English</td>
<td>97.24%</td>
</tr>
<tr>
<td>Banko and Moore(2004)</td>
<td>English</td>
<td>96.55%</td>
</tr>
<tr>
<td>Dandapat and Sarkar(2006)</td>
<td>Bengali</td>
<td>84.37%</td>
</tr>
<tr>
<td>Rao et al.(2007)</td>
<td>Hindi</td>
<td>76.34%</td>
</tr>
<tr>
<td></td>
<td>Bengali</td>
<td>72.17%</td>
</tr>
<tr>
<td></td>
<td>Telegu</td>
<td>53.17%</td>
</tr>
<tr>
<td>Rao and Yarowsky(2007)</td>
<td>Hindi</td>
<td>70.67%</td>
</tr>
<tr>
<td></td>
<td>Bengali</td>
<td>65.47%</td>
</tr>
<tr>
<td></td>
<td>Telegu</td>
<td>65.85%</td>
</tr>
<tr>
<td>Sastry et al.(2007)</td>
<td>Hindi</td>
<td>69.98%</td>
</tr>
<tr>
<td></td>
<td>Bengali</td>
<td>67.52%</td>
</tr>
<tr>
<td></td>
<td>Telegu</td>
<td>68.32%</td>
</tr>
<tr>
<td>Ekbal et al.(2007)</td>
<td>Hindi</td>
<td>71.65%</td>
</tr>
<tr>
<td></td>
<td>Bengali</td>
<td>80.63%</td>
</tr>
<tr>
<td></td>
<td>Telegu</td>
<td>53.15%</td>
</tr>
<tr>
<td>Ours</td>
<td>Assamese</td>
<td>85.64%</td>
</tr>
</tbody>
</table>

Table 3: Comparison of our result with other HMM based model.
Toutanova et al. (2003) obtained the best result for English (97.24%). Among HMM based experiments reported on Indian languages, we have obtained the best result (86.89%). This work is ongoing and the corpus size and the amount of tagged text are being increased on a regular basis.

The accuracy of a tagger depends on the size of tagset used, vocabulary used, and size, genre and quality of the corpus used. Our tagset containing 172 tags is rather big compared to other Indian language tagsets. A smaller tagset is likely to give more accurate result, but may give less information about word structure and ambiguity.

The corpora for training and testing our tagger are taken from an Assamese daily newspaper Asomiya Pratidin, thus they are of the same genre.

6 Conclusion & Future work

We have achieved good POS tagging results for Assamese, a fairly widely spoken language which had very little prior computational linguistic work. We have obtained an average tagging accuracy of 87% using a training corpus of just 10000 words. Our main achievement is the creation of the Assamese tagset that was not available before starting this project. We have implemented an existing method for POS tagging but our work is for a new language where an annotated corpora and a pre-defined tagset were not available.

We are currently working on developing a small and more compact tagset. We propose the following additional work for improved performance. First, the size of the manually tagged part of the corpus will have to be increased. Second, a suitable procedure for handling unknown proper nouns will have to be developed. Third, if this system can be expanded to trigrams or even n-grams using a larger training corpus, we believe that the tagging accuracy will increase.

Acknowledgement

We would like to thank Dr. Jyotiprakash Tamuli, Dr. Runima Choudhary and Dr. Madhumita Barbora for their help, specially in making the Assamese tagset.

References


Ekbal, A., Mandal, S., & Bandyopadhyay, S. POS tagging using HMM and rule based chunking. Workshop on Shallow Parsing for South Asian Languages. 2007.


