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 দেউতা (*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 কাৰণে (*kArane*), দৰে (*dare*), নিমিত্তে (*nimitte*), হেতু (*hetu*), etc., are placed after pronominal adjective, they are considered conjunction and if placed after

Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, pages 33–36,
Suntec, Singapore, 4 August 2009.

Table 1: Personal definitives are inflected on person and number

Person	Singular	Plural
1^{st} প্রথম	My father মোৰ দেউতা *mor deutA*	Our father আমাৰ দেউতা *aAmAr deutA*
2^{nd} মান্য মধ্যম	Your father তোমাৰ দেউতাৰা *tomAr deutArA*	Your father তোমালোকৰ দেউতাৰা *tomAlokar deutArA*
2^{nd}, Familiar তুচ্ছ মধ্যম	Your father তোৰ দেউতাৰ *tor deutAr*	Your father তহতঁৰ দেউতাৰ *tahator deutAr*
3^{rd} তৃতীয়	Her father তাইৰ দেউতাক *tAir deutAk*	Their father সিহঁতৰ দেউতাক *sihator deutAk*

noun or personal pronoun they are considered particle. For example,

এই কাৰণে মই নগলোঁ।
TF[1] : *ei kArane moi nagalo.*
This + why + I+ did not go.
ET[2] : This is why I did not go.
ৰামৰ কাৰণে মই নগলোঁ।
TF : *rAmar kArane moi nagalo.*
Ram's + because of + I + did not go
ET : I did not go because of Ram.
In the first sentence কাৰণে (*kArne*) is placed after pronominal adjective এই (*ei*); so *kArne* is considered conjunction. But in the second sentence *kArne* is placed after noun ৰাম (*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,

  সিহে গৈছিল। 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,

  ধুনীয়াজনী হৈ আহিছা।
  TF : *dhuniyAjoni hoi aAhisA.*
  ET : You are looking beautiful.
  Here ধুনীয়া (*dhuniyA : beautiful*) is an adjective, but after adding feminine suffix জনী the whole constituent becomes a noun word.

- Even conjunctions can be used as other part of speech.
  হৰি আৰু যদু ভায়েক ককায়েক।
  TF : *Hari aAru Jadu bhAyek kokAyek.*
  ET : *Hari and Jadu are brothers.*
  যোৱাকালি ৰাতিৰ ঘটনাটোৱে বিষয়টোক আৰু অধিক ৰহস্যজনক কৰি তুলিলে।
  TF : *JowAkAli rAtir ghotonAtowe bishoitok aAru adhik rahashyajanak kori tulile.*
  ET : The last night incident has made the matter more mysterious.
  The word আৰু (*aAru : and*) shows ambiguity in these two sentences. In the first, it is used as conjunction (i.e. Hari and Jadu) 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. Avinesh 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. Pammi 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 langauages use

[1]TF : Transliterated Assamese Form
[2]ET : 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, accussative, 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 CN (Common Noun) and probability 1 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.

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

Test set	Size	Average accuracy	UDH accuracy	UPH accuracy
A	992	84.68%	62.8%	42.0%
B	1074	89.94%	67.54%	53.96%
C	1241	86.05%	85.64%	26.47%

Table 3: Comparison of our result with other HMM based model.

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

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

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 form 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.

Acknowledgemnt

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

References

Avinesh PVS & Karthik G. POS tagging and chunking using Conditional Random Field and Transformation based learning. *IJCAI-07 workshop on Shallow Parsing for South Asian Languages*. 2007.

Banko, M., & Robert Moore, R. Part of speech tagging in context. *20th International Conference on Computational Linguistics*. 2004.

Dandapat, S. Part-of-Speech Tagging and Chunking with Maximum Entropy Model. *Workshop on Shallow Parsing for South Asian Languages*. 2007.

Dandapat, S., & Sarkar, S. Part-of-Speech Tagging for Bengali with Hidden Markov Model. *NLPAI ML workshop on Part of speech tagging and Chunking for Indian language*. 2006.

Dermatas, S., & Kokkinakis, G. Automatic stochastic tagging of natural language text. *Computational Linguistics* 21 : 137-163. 1995.

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

Goswami, G. C. Asamīyā Vyākaraṇ Pravesh, Second edition. *Bina Library, Guwahati*. 2003.

http://shiva.iiit.ac.in/SPSAL2007. *IJCAI-07 workshop on Shallow Parsing for South Asian Languages*. Hyderabad, India.

Pammi, S.C., & Prahallad, K. POS tagging and chunking using Decision Forests. *Workshop on Shallow Parsing for South Asian Languages*. 2007.

Rao, D., & Yarowsky, D.. Part of speech tagging and shallow parsing of Indian languages. *IJCAI-07 workshop on Shallow Parsing for South Asian Languages*. 2007.

Rao, P.T., & Ram, S.R., Vijaykrishna, R. & Sobha L. A text chunker and hybrid pos tagger for Indian languages. *IJCAI-07 workshop on Shallow Parsing for South Asian Languages*. 2007.

Saha, G.K., Saha, A.B., & Debnath, S. Computer Assisted Bangla Words POS Tagging. *Proc. International Symposium on Machine Translation NLP & TSS*. 2004.

Sastry, G.M.R., Chaudhuri, S., & Reddy, P.N. A HMM based part-of-speech and statistical chunker for 3 Indian languages. *IJCAI-07 workshop on Shallow Parsing for South Asian Languages*. 2007.

Sharma, U., Kalita, J. & Das, R. K. Acquisition of Morphology of an Indic language from text corpus. *ACM TALIP* 2008.

Singh, S., Gupta K., Shrivastava, M., & Bhattacharyya, P. Morphological richness offsets resource demand-experiences in constructing a POS tagger for Hindi. *COLING/ACL*. 2006.

Toutanova, K., Klein, D., Manning, C.D. & Singer, Y. Feature-Rich part-of-speech tagging with a Cyclic Dependency Network. *HLT-NAACL*. 2003.

Viterbi, A.J. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transaction on Information Theory* 61(3) : 268-278. 1967.