# Building Large Scale Text Corpus for Joint Word Segmentation and Part-of-Speech Tagging of Myanmar Language

Dim Lam Cing, Khin Mar Soe

Natural Language Processing Lab (NLP), University of Computer Studies, Yangon, Myanmar

**Abstract.** In Natural Language Processing (NLP), Word segmentation and Part-of-Speech (POS) tagging are fundamental tasks. The POS information is also necessary in NLP's preprocessing work applications such as machine translation (MT), information retrieval (IR), etc. Currently, there are many research efforts in word segmentation and POS tagging developed separately with different methods to get high performance and accuracy. Word segmentation and Part-of-speech tagging is one of the important actions in language processing. Against this, while numerous models are provided in different languages, few works have been performed for Myanmar language. This paper describes the building of Myanmar Corpus to use for joint word segmentation and part-of-speech tagging of Myanmar Language. In our research, the corpus contains 51207 sentences and 839161words. The corpus is created using 12 tags. To evaluate the accuracy of the corpus, HMM model is trained on different data size and testing is done with closed test and opened test. Results with 94% accuracy in the experiments show the appropriate efficiency of the built corpus.

**Keywords:** Natural Language Processing, POS, HMM, Corpus

## 1. Introduction

Language corpora are widely used in linguistic research and language technology. A tremendous interest has arisen in recent years in building and developing computerized language corporations. Studying the electronic corpora of different languages provides learners and researchers with the opportunity to work with language information with a variety of tools and techniques in analytical procedures and programs [1].

POS tagged corpus is a structured textual database that serves as a reference material for further NLP work as well as a learning repository for machine translation algorithms and other applications for code. Building syntactically classified corpus requires a sequence of procedures such as text preprocessing, tokenizing sentences and POS tagging. Also it is influenced on all areas of NLP such as information retrieval, text-to-speech, parsing, information extraction and any linguistic research for corpora [2].

While many words can be unambiguously associated with one POS or tag, other words match multiple tags, depending on the context that they appear in [3]. Therefore, the accuracy of a tagger depends on its learning database or its training data. The larger the corpus size, the better the accuracy for tagging. Also, an automatic part-of-speech tagger requires a large corpus because hand annotating is tedious task and also assigning POS tags to each word is very time consuming [4].

In this paper, we start by hand annotating raw text to build a tagged corpus. Then we process by preparing training data from the manually tagged corpus. Next, we automatically assign POS tags to each word of raw text using our proposed POS tagger. Then, we analyze result of tagged text and refine manually. We conclude with the result that POS tagged corpus for Myanmar Language is annotated by stochastic method of POS tagging.

Myanmar Language is a common language of the national languages of Myanmar and is part of the family of the Sino-Tibetan language. It is spoken as first language by about 33 million people and as second language by 10 million people [5]. The truth is that Myanmar Language has only a small amount of linguistic computational capital. On this language, there are a few computational works. Researchers have recently started to engage in the creation and enrichment of Myanmar Language's language in the Natural Language Processing (NLP) sector. These NLP activities included the need to build a large amount of language-based corporations.

The term "corpus" is used to refer to a collection of linguistic records (masking spoken and written records) in a language for certain unique functions, and to save, take care of and translate those facts in virtual format. A corpus, as an example, can be quite small, consisting of 50,000 words or texts, or very large, such as millions of words. Corpus is the premise for linguistic research of a wide variety. The corpus range is huge. The fields of corpus-based totally studies are : grammatical research of unique linguistic production, building reference grammar, lexicography, language variation and dialectology, ancient linguistics, studies of transcription, language acquisition, language pedagogy, and processing of natural language, etc.

The need of language corpora has caused to the study of corpus linguistics. It is not a branch in linguistics, however a method that helps to carry out linguistic studies. The development of computer software program for corpus evaluation has been closely related to modern-day corpus linguistics from the very beginning. In modern corpus linguistics, linguists and computer scientists share a common goal that in order to perform any kind of linguistic analysis, it is necessary to rely on actual or real language knowledge (speech or writing). It is also an approach that addresses two main goals: how people use language in daily communication and how to create intelligent systems to communicate with people [1].

## 2. Related Work

To date, numerous methods of POS tagging have been presented in some languages such as English, often based on rules or statistics. But in this field there are few activities that have been done over the past few years [6]. Jabbari and Allison are doing one of the latest works on POS tagging [7]. Their strategy is based on transformation, and Brill and Hepple used it previously in English [8, 9]. Creating this tagger requires a professional learner computer that provides approximate rules. They actually applied Error-Driven Transformation Based Learning implementation. They believe their approach is 93 % accurate [10] presents an HMM model for POS tagging in Manipuri. Since Manipuri has no tagged corpus, the system uses Tagger's small set of tagged phrases based on Manipuri Rule. The system can assign tags to most of the lexical items of the test set. This tagger will be very useful in language processing applications such as text-based information retrieval, speech recognition and machine translation, etc. The proposed system can be made more efficient by applying the bigram probability to trigram probability and it gives 92 % accuracy.

There are many other POS tagging system by using Hidden Markov Model in other languages[11, 12]. The HMM method is applied in Persian POS tagging to determine the reliability of the proposed approach in simulations performed on both homogeneous and heterogeneous Persian corpus. Obtaining 98.1 % accuracy results in the experiments shows the adequate effectiveness of the Persian corpus approach proposed [11]. The research is based on a statistically based approach by measuring the probability of the tag sequence and the term likelihood of the given corpus, where the linguistic data is automatically extracted from the annotated corpus in which the tagging process is carried out. For known words, more than 90% of the accuracy can be achieved by the current tagger [12].

## 3. Tagged Corpus Generation

In this paper, we are mainly concerned with building the framework of Myanmar Language un-annotated raw corpus consisting of 839,161 total words and also attempting to highlight the problems faced during the construction process. A huge collection of texts would be useful for language and non-linguistic research, cross-linguistic correlations and all other communication technologies.

There are different problems related to corpus design, development and management. These issues differ depending on the corpus' form and usefulness. In fact, the development of speech corpus is different from the development of text corpus.

Myanmar language tagged corpus is essential in any applications of Natural Language Processing. There are several steps to create tagged corpus using stochastic method that are Collecting Raw Data, Manually Tagged, Preparing Training Data and Increasing Data size in the Tagged Corpus.

### 3.1. Collecting Raw Text

The collection of data is a vital activity to build a corpus. A great deal of raw text must be assembled from a variety of sources. Also raw text is checked for morphological and syntactic errors so as to be ready to annotation.

In case of this work, bunch of raw text are collected from online journals, newspaper and e-books. Myanmar text are copied and saved in text files.

### 3.2. Manually Tagged

The collecting Myanmar texts in the corpus are tagged manually by hand and have training data for statistical method.

### 3.3. Preparing Training Data

Currently we prepared over 50,000 sentences as training data. Then we will develop a HMM model by calculating the probabilities of the tag of each word, counting word frequency. These functions help us to analyze on tagged corpus.

### 3.4. Increasing Data Size in the Tagged Corpus

The corpus is enlarged by assigning POS tag automatically to unprocessed text files. POS tagger runs and assigns POS tag to each word by using the Hidden Markov Model (HMM) by selecting the maximum POS tag for each word automatically on the untagged text.

After generating tagged text, we have to analyze and refine manually to unknown tag and wrong tag. Finally, we can use these correct texts in the corpus so that our corpus size can be enlarged.

## 4. Experimental Evaluation

The accuracy of any Part of Speech tagger is measured in terms of the following accuracy:

$$Recall, R = \frac{Number\ of\ correct\ POS\ tag\ assigned\ by\ the\ system}{Number\ of\ words\ in\ the\ test\ set} \tag{1}$$

$$Precision, P = \frac{Number\ of\ correct\ POS\ tag\ assigned\ by\ the\ system}{Number\ of\ POS\ tag\ assigned\ by\ the\ system} \tag{2}$$

$$F_{score}, F = \frac{2PR}{P+R} \tag{3}$$

### 4.1. Statistic of the Dataset

For evaluation of the proposed tagger [13], a corpus having texts from different genres were used. In our corpus, consists of the Asian Language Treebank (ALT) corpus, is one part from the ALT Project and the UCSY corpus, is created by the NLP Lab, University of Computer Studies, Yangon (UCSY), Myanmar. The other data of the corpus is collected from Myanmar Grammar Book and websites that contain economic, social, art, culture, sport and religious. It aims to promote word segmentation and POS tagging research on Myanmar language. The statistic of the Dataset is described in Table 1. Although the News dataset is dominant, the data coverages the various topics.

### 4.2. Processing

Myanmar Language has no space between words. So, in our research, we preliminary segment the sentence in syllable using the syllable break tool [14]. And then, segment by using N-grams model (5-grams) and select the maximum POS tag for the word by using HMM. This is implemented by using Python programming code.

Table1: Statistic of the Dataset

| Data Type | No. of Sentences | No. of Words |
|---|---|---|
| UCSY | 5017 | 58301 |
| ALT | 1500 | 27340 |
| Web News | 35725 | 612875 |
| Short Stories | 1300 | 25375 |
| Novels | 5142 | 72656 |
| Books | 2523 | 42614 |
| Total | 51207 | 839161 |

## 5. Result and Discussion

To evaluate the generated corpus, two different experiments are performed. We collect 500 new sentences for closed test and open test. In our experiments, we compare four different training data size as described in Table 2. According to this experiment, the overall obtained accuracy of the training data is over 92% .

Table 2: Comparison of the accuracy on Closed Test and Open Test on different data size

| **Corpus Size (Total Words)** | **Closed Test** | | | **Open Test** | | |
|---|---|---|---|---|---|---|
| | **R** | **P** | **F** | **R** | **P** | **F** |
| 547969 | 76% | 78% | 77% | 69% | 70% | 69% |
| 690258 | 81% | 83% | 82% | 78% | 79% | 78% |
| 740495 | 91% | 92% | 91% | 89% | 90% | 89% |
| 839161 | 93% | 94% | 93% | 92% | 93% | 92% |

Nevertheless, the results of the experiments show that the greater the training data, the greater the reliability. And there's only a little gap between the closed test and the open test in the last corpus. Finally, the accuracy of these tests in terms of accuracy show the success of building the large corpus for joint word segmentation and POS tagging for Myanmar Language.

## 6. Error Analysis

Some errors occurred in the experiments especially in syllable break that cannot break in some consonant appeared consecutive in the sentence. As the consequences of the syllable break error, the segmentation error occurred. The unknown words occurred because of segmentation error and person names, locations that are not containing in the corpus. The segmentation is performed by N-gram and the tagging also performed on the longest (5-grams) words, so some wrong tagging occurred. Some words of Myanmar Language have more than one POS tag. So, some POS tagging in words may cause ambiguity.

## 7. Conclusion

In this paper the building of Myanmar POS Corpus is described. Experimental results show that there are differences in the accuracy rate on different training data. By using a large training, joint word segmentation and the assignment of POS tagging is more accurate and reduced the unknown words, incorrect tag and ambiguous words. The paper has shown that the training corpus is efficient for joint word segmentation and POS tagging in Myanmar Language. High accuracy rate (94%) is got in closed testing of the experiment.

## 8. References

[1] S. Sarma, H. Bharali, A. Gogoi, R. Deka, A. Barman. A Structured Approach for Building Assamese Corpus: Insights, Applications and Challenges , *Proceedings of the 10th Workshop on Asian Language Resources*, pages 21–28, COLING 2012, Mumbai, December 2012

[2] A. Ganbold, P. Jaimai, Integrative Tools for Part-of-Speech Tagged Corpus, *School of IT, National University of Mongolia*, Ulaanbaatar, Mongolia.

[3] J. Diesner, Part of Speech Tagging for English Text Data, *School of Computer Science, Carnegie Mellon University, Pittsburgh.*

[4] F. M. Hasan, N.UzZaman , M. Khan, Comparison of Unigram, Bigram, HMM and Brill's POS Tagging Approaches for some South Asian Languages, *Proc. Conference on Language and Technology (CLT07),* Pakistan, 2007

[5] https://en.wikipedia.org/wiki/Burmese_language

[6] M. Mohseni, H. Motalebi, B. Minaei-bidgoli, M. Shokrollahi-far, A farsi part-of-speech tagger based on markov, *In the proceedings of ACM symposium on Applied computing*, Brazil (2008).

[7] S. Jabbari, B. Allison, Persian Part of Speech Tagging, *In the Proceedings of Workshop on Computational Approaches to Arabic Script-Based Languages (CAASL-2),* USA (2007).

[8] E. Brill, Transformation-Based Error-Driven Learning and Natural Language Processing: *A case Study in Part of Speech Tagging, Computational Linguistics*, USA (1995).

[9] M. Hepple, Independence and Commitment: Assumptions for Rapid Training and Execution of Rule-based Partof-Speech Taggers, *In Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL)*. Hong Kong (2000).

[10] K. R. Singha, B. S. Purkayastha, K. D. Singha, Part of Speech Tagging in Manipuri with Hidden Markov Model, *IJCSI* International *Journal of Computer Science Issues*, Vol. 9, Issue 6, No 2, November 2012

[11] M. Okhovvat , B.M. Bidgoli. A Hidden Markov Model for Persian Part-of-Speech Tagging, *Procedia Computer Science* 3 (2011) 977–981

[12] A.J.P.M.P. Jayaweera, N.G.J. Dias. Hidden Markov Model Based Part of Speech Tagger for Sinhala Language, *International Journal on Natural Language Computing (IJNLC)* Vol. 3, No.3, June 2014

[13] D.L.Cing, K.M.Soe. Joint Word Segmentation and Part-of-Speech (POS) Tagging for Myanmar Language, *17th International Conference on Computer Application*, Yangon, 27-28, February,2019

[14] https://github.com/ye-kyaw-thu/sylbreak