Part of Speech Tagger for Amharic with Shallow Stemming

Abel Gadissa
HiLCoE, Computer Science Programme, Ethiopia
Derba Transport PLC, Ethiopia
homygg@gmail.com

Wondwossen Mulugeta
HiLCoE, Ethiopia
School of Information Science, Addis Ababa University, Ethiopia
wondisho@yahoo.com

Abstract

Natural Language Processing has impact on creating the ability for computers to understand human language. It is a branch of artificial intelligence that deals with analyzing, understanding and generating the languages that humans use naturally in order to interface with computers in both written and spoken form. In this process, part of speech tagging, PoS, is used to achieve the goal of making computers understand the contribution of words in sentences.

In this regard there have been several researches in English and other European languages. There have been many attempts to apply the approaches used for other languages on Amharic. Amharic is the official language of the Federal Democratic Republic of Ethiopia and the second most spoken Semitic language next to Arabic.

PoS Tagging for morphologically rich languages like Amharic is very challenging. There are various attempts employing a number of techniques for Amharic PoS tagging. While some are rule based approaches, others tried to see the possibility of using machine learning methods for the same task. In this research, we have tried to employ shallow stemming for part of speech tagging on Amharic text written using the Ethiopic script. It is assumed that the attempt will contribute for an improved usage of Amharic language in the field of Natural Language Processing.

Natural Language Tool Kit (NLTK) is used in integration with Python programming to process human language data. NLP with Python provides practical introduction to programming for language processing. NLTK can be used to tokenize and tag texts, identify name entities and display a parse tree. From the modules available in NLTK, nltk tag for HMM has been used for the processing of data. There is a class in the NLTK tag module called Hidden Markov model that is a generative model for labelling data. News articles from Walta Information Center are used for this research. Shallow stemming was applied on the data to emerge with accuracy score. In the overall tagging process, several pre- and post-processing tasks are undertaken. These processes include removal of prepositions and conjunctions, extraction of lexicons (unique words), corpus manipulation, applying tenfold cross validation, training and testing the tagger in order to measure the performance of the system and finally, checking the accuracy of the output using Confusion Matrix.

The accuracy of the tagger has been evaluated based on the words that are correctly tagged and the words that are wrongly tagged which are 90.65% and 9.34% respectively. The result encourages the application of the PoS tagger for other similar languages that use the Ethiopic Script (Amharic Fidel) such as Tigrigna, Geez and others with little modification based on the availability of corpus.

Keywords: Morphology; Part of Speech; Amharic; Tagger; Fidel; Corpus

1. Introduction

Natural Language Processing plays a great role in the effort to make computers understand human language and respond accordingly. Natural Language Processing (NLP), sometimes referred to as Human Language Technology, is a branch of artificial intelligence that deals with analyzing, understanding and generating the languages that humans use naturally in order to interface with computers in both

written and spoken form using natural languages. The implementation of Natural Language Processing backdates to the 20th Century [1].

In the past few years, there have been attempts to work on computational linguistics on Amharic. Researching on Amharic is essential as it is the second most widely spoken Semitic language next to Arabic. Amharic is the official language of the Federal Democratic Republic of Ethiopia and the working language of many regional governments in Ethiopia with more than 30 million estimated speakers [2].

In the field of Natural Language Processing, Part of Speech (POS) Tagging plays a significant role in achieving the goal of computers understand and process human language. Words are traditionally grouped into equivalence classes called parts of speech (PoS). Part of Speech is also referred to as lexical categories, word classes, morphological classes, and lexical tags. It is also a linguistic category of words (or more precisely lexical items), which is generally defined by the syntactic or morphological behavior of the lexical item in question [3].

Before further addressing other issues, it seems important to discuss what morphologically rich language means. Morphologically rich languages are characterized by highly productive morphological processes (inflection, derivation, compounding, etc.) that may produce a very large number of word forms from a given root form.

Part of Speech (POS) Tagging is the process of identifying the word class of the words that are found in a sentence. According to Girma Awgichew and Mesfin Getachew, there are around eleven basic word classes in Amharic [4]. These are Noun, Preposition, Interjection, Pronoun, Conjunction, Punctuation, Verb, Adverb, Adjective, Numeral, and Unclassified.

The most cited classical work on Amharic grammar, written in Amharic by Blata Merseahazen Woldeqirqos, claims that the word classes of Amharic are eight [5]. It is further minimized to only five Amharic word classes by categorizing some of the previous categories into sub class according to a recent work of Baye Yimam [6].

Working on POS Tagging in a morphologically rich language is very challenging. Even though it is hard, there are previous attempts to work on such research but still there are observed problems in addressing the words in the appropriate manner. Most of the previous researches conducted consider every morphologically derived word as distinct word [1]. This might be valid for languages which are not morphologically complex like English. However, languages like Amharic where a single word might represent a complete sentence that has subject-verb-object-tense-preposition, etc., is difficult to consider as a distinct word. When considering such words as one word in one POS becomes challenging and is a source for errors. On previous works, researchers took words in a corpus as they are and if the words are taken as they are, they maximize the number of unique words to be processed by the tagger taking into consideration that Amharic is morphologically rich Language. One observed problem in previous researches was that words were given to the tagger without pre-processing the affixes and we believe that has affected the accuracy of the tagger. Applying shallow stemming on the corpus is believed to increase its accuracy.

2. Related Work

There have been researches conducted in POS Tagging for Amharic and other languages. Concerning the research conducted on Amharic Part of Speech Tagging, the research work in [7] is one of the pioneers in experimenting Automatic Part of Speech Tagging. The tagging algorithm chosen was Hidden Markov Model. In the experiment, the authors organized a single page of Amharic corpus to train and test the tagger. In this research the authors added own tagsets together with the basic tags and came up with a total of 25 tags. In relation to the tagset, there were comments made by some

researchers on the usage of additional tags as it can be considered as sub-classes and can be easily reduced back to the basic tags like N, V, ADJ, etc. Other researchers suggested that context of usage would help achieve a better result by handling affixes that are attached with words rather than coming up with modified tags like NPC, which is to represent a Noun with Preposition and Conjunction, it is possible to obtain the context of its usage from the tagset [8].

Another research used a small annotated corpus of 1000 words and applied Conditional Random Fields to the task of word segmentation and POS Tagging [9]. Some of the categories proposed in [8] were collapsed to avoid some of the tags that should have been included in the tagset like Verbal Noun, Pronoun and Relative Verb. Moreover, some tags like Adposition were merged and cardinal and ordinal numbers were considered collectively as numerals. The POS Tags used are Noun (N), Verb (V), Auxiliary verb (AUX), Numeral (NU), Adjective (AJ), Adverb (AV), Adposition (AP), Interjection (I), Residual (R), and Punctuation (PU). It is mentioned that the reason for working with such abstract POS Tags is due to lack of large annotated corpus. So, working with 25 POS tags is considered to make the data sparse and the outcome unreliable [9]. By adding the previously mentioned features, the author reported an accuracy of 74% in POS tagging in which it is stated that the application of morphological feature didn't bring as much anticipated result whereas using bi-gram feature increased the performance by 3%.

A medium sized Amharic corpus was organized with the aim to manually annotate 210,000 prosodic words [4]. For this a total of 1065 news items from a local news agency (Walta Information Center) were taken with technical and financial assistance provided by the University of Stockholm for manually annotating Amharic news items. Nouns (N), pronouns (PRON), adjectives (ADJ), adverbs (ADV), verbs (V), prepositions (PREP), conjunctions (CONJ), interjections (INT), punctuations (PUNC) and unclassified/difficult to tag words (UNC) were the classes and sub-classes identified as the basic POS tagset during the research process.

In order to obtain the final result, pre-processing was carried out on the news text. In order to make manual tagging, they identified and briefed annotators with linguistic background on the nature of Amharic POS and how to annotate the news items. Accordingly, the annotators made manual tagging on each news document.

Gambäck also worked on Amharic POS tagging to verify, correct and retag a corpus of Amharic news texts. For this purpose, a total of 8715 Amharic texts from the news domain were gathered from Walta Information Center web archives. The total corpus (1065 articles; 210,000 words) was then morphologically analyzed and manually annotated [10].

The work in [8] applied tokenization on Amharic texts. The resulting words represent the concatenation of morphemes, each capable of having its own POS tag unlike English. Following tokenization, feature extraction was carried out and different features are used in POS tagging each token. Of all, the novel features are the vowel patterns and radicals.

From this work we have observed that the improvement in performance is attributed to a combination of the POS tagged corpus cleaned up to minimize the pre-existing tagging errors and inconsistencies, the vowel patterns and the roots, which are characteristics of Semitic languages, which have been used to serve as important elements of the feature set and finally, state-of-the-art machine learning algorithms have been used and parameter tuning has been done whenever necessary and as much as possible. Hence, all these processes and steps helped to meet the attempted purpose, which is the improvement in the performance of POS tagging for Amharic texts.

3. The Proposed Solution and Experiments

This work is an attempt to model Amharic Part-of-Speech tagger that uses Amharic Fidel script and

implements machine learning approach by applying shallow stemming. The acquisition, preparation and analysis of the corpus for the experiment is very crucial. The publicly available Amharic corpus used for this research is the Walta Information Center's news articles, which is manually tagged by the Ethiopian Language Research Center of Addis Ababa University. The corpus has 210,000 prosodic words. In this research, it was necessary to do data verification that requires thorough review of the content of the corpus.

To manipulate the data, Python programming is used, as it is easy to work with strings. Moreover, the NLTK module, which is developed to be integrated with Python, made the usage of the tool more preferable. Starting with the statistical analysis on the corpus, identifying the number of words tagged with the specific tag and other additional data pre-processing tasks have been performed. First, all the words in the corpus were counted using a Python program and we found 206,929 words including punctuations. This figure shows a variation with the initial claim made by the tagging team of the corpus. There were also 32,480 lexica in the corpus. Among the 206,929 words, around 63,560 words are found to have phrasal structure. This is 30% of the total corpus. This can be taken as an indication that even though a probabilistic tagger can tag 70% of the tokens in the corpus correctly, the result that would be obtained is going to be poor compared to previously conducted PoS Tagging researches on Amharic. From the 30% of the words with phrasal structure, 52% are tagged as Noun with Preposition (NP). This implies that if there was a pre-processing task to perform shallow stemming, the same amount of Nouns with a tag N could have been added to the corpus decreasing the phrasal tag instances. From the 63,560 words with phrasal structure, 3,389 are words that have both suffix and prefix with tags like Pronoun with a proclitic preposition and an enclitic conjunction (PRONPC), Verb including relative verbs and auxiliaries with a proclitic preposition and an enclitic conjunction (VPC), Numeral (cardinal or ordinal) with a proclitic preposition and an enclitic conjunction (NUMPC), Noun Proposition with Cardinal (NPC), Adjective with a proclitic preposition (ADJPC), and Preposition with conjunction (PREPC). The rest have one of the two affixes: Suffix or Prefix. The majority of the phrasal words are tagged with compound tag that has Preposition and other basic word classes. There are 55,007 tokens that are tagged with tags conjoined with prepositions. These tags are: NP, PRONP, NUMP, ADJP, and VP. The rest of the tokens that are having phrasal structure are tokens tagged with basic tag combined with Conjunction. These tokens are accounted for about 5,164 of the tokens with the phrasal structure. The tags that are combined together with conjunctions alone are Noun Conjunction (NC), Pronoun attached with conjunction (PRONC), Adjective with Conjunction (ADJC), Numeral (cardinal or ordinal) attached with conjunction (NUMC), and Verb attached with conjunction (VC).

Segmenting such words is important in order to tag each and every token with a basic tag without any prepositions or conjunctions. By doing so, all previously identified words as having phrasal structure are segmented and retagged in the corpus. This will eventually remove the 16 compound tags that are used by ELRC to annotate words with the phrasal structure.

After going through pre-processing the corpus in a way we thought would bring a better result in the experiment, we have removed the compound tags that are used and came up with 15 tags. The newly suggested tags, which do not include tags for combined word classes as it was used in the ELRC tagset are presented in Table 1.

During this stage of data pre-processing, the corpus was available in Latin script. Words in the corpus are represented using Latin script and have to be transliterated back to Amharic Fidel. The words are placed in an Amharic format holding a character combination of their exact equivalence in Latin script. The transliteration was based on a system for Ethiopic Representation in ASCII (SERA).

Table 1: The Newly Suggested Tagset

New Tag Suggested	*Description*	*New Tag Suggested*	*Description*
ስ	ስም	ስነ	ስርአተነጥብ
ቅ	ቅጽል	ረግ	ረዳትግስ
ተጠግ	ተውላጠግስ	መጻ	መስተጻምር
መቁ	መቁጠሪያቁጥር	መደቁ	መደበኛቁጥር
ግ	ግስ	ተጠስ	ተውላጠስም
መዋ	መስተዋድድ	ቃአ	ቃልአጋኖ
ግስም	ግሳዊስም	አም	አልተለየም
ተከግ	ተውሳከግስ		

Phrasal tagset was removed from the tagset received to be used for corpus training. The untagged corpus is also generated from the tagged corpus by removing the tags. The words in the corpus are made to be organized into sentences and made to be delimited by a new line at the end of each sentence for the purpose of training and testing the tagger.

Tenfold (10 fold) Cross validation method is applied to test the performance of the tagger. Sentences to be used for the testing were removed from the training data set. The tagset, the lexicon and the training corpus were used in the process of training the tagger. After the training, the untagged sentences extracted from the initial corpus were provided to the tagger for testing and the result is compared with the original/manually tagged corpus.

The shallow stemming process resulted in an increased number of tokens in the corpus. After the shallow stemming, the resulting corpus contained a total of 270,523 words with 23,882 unique words.

Having the tagged and untagged list in the corpus, the words are converted into sentences. The corpus was converted into a list of 8,076 sentences. The algorithm used for the sentence construction is based on the sentence delimiter tag using the initial corpus.

When the 10 fold cross validation is applied on the corpus, in each iteration 7269 tagged sentences were used to train the tagger and the remaining 807 sentences are used for testing the tagger.

The Hidden Markov Model is one of the most important machine learning models in speech and language processing. In order to define it properly, Markov chain must be firstly introduced, sometimes called the observed Markov model. Markov chains and Hidden Markov Models are both extensions of the finite automaton. A finite automaton is defined by a set of states, and a set of transitions between states that are taken based on the input observations. A weighted finite-state automaton is a simple augmentation of the finite automaton in which each arc is associated with a probability, indicating how likely that path is to be taken. Hidden Markov Models (HMMs) are largely used to assign the correct label sequence to sequential data or assess the probability of a given label and data sequence. These models are finite state machines characterized by a number of states, transitions between these states, and output symbols emitted while in each state. The HMM is an extension to the Markov chain, where each state corresponds deterministically to a given event. In HMM, the observation is a probabilistic function of the state. HMMs share the Markov chain's assumption, being that the probability of transition from one state to another only depends on the current state, i.e., the series of states that led to the current state are not used. They are also time invariant [3].

After organizing the corpuses for training and testing accordingly, the tagged corpus with 7269 sentences together with the lexicon and tagset is given to the tagger for training. Then the previously extracted 807 untagged sentences that are not available in the tagged corpus are given to the Hidden Markov Model (HMM) tagger to test it.

The result that was gained from the tagger is compared with 807 sentences from the original/manually annotated corpus. The same procedure is followed to go through all the rest of the training and testing process. After the comparison is made, the accuracy of the system is documented. Comparison was done using a confusion matrix. The process checks the golden corpus with a tagset and presents a value of order for the current output and intended output. The confusion matrix contains

precision, recall and false positive as headers. The precision is calculated by dividing the sum of actual positives and actual negatives with the sum of predicted and actual values taken from the initial corpus. The recall is calculated by dividing the actual positives with the sum of predicted negatives and actual positives whereas the false positive is calculated by dividing the predicted positives with the sum of actual negatives and predicted positives. The confusion matrix, represented in percentage form, is shown in Table 2.

Table 2: Confusion Matrix in Percentage

	Precision	*Recall*	*FP*
መቁ	0.5126	0.0390	0.0373
መዋ	0.5024	0.2256	0.2244
መደቁ	0.5596	0.0016	0.0013
መጻ	0.5041	0.0365	0.0360
ስ	0.5242	0.4084	0.4010
ስነ	0.5001	0.0517	0.0517
ረግ	0.5571	0.0036	0.0029
ቃአ	1.0000	0.0000	0.0000
ቅ	0.5764	0.0425	0.0319
ተከግ	0.5610	0.0111	0.0087
ተጠስ	0.5713	0.0079	0.0059
ተጠግ	0.5774	0.0280	0.0208
አም	0.9102	0.0006	0.0001
ግ	0.5425	0.1050	0.0915
ግስም	0.6264	0.0194	0.0117

4. Discussion

The result of this research, which is 90.65% of the overall accuracy, is found to be encouraging for Amharic using HMM model. With all its pros and cons of the HMM Model, the obtained result by implementing shallow stemming to the corpus not only improved the previously obtained result by other researchers who used the HMM model but also paves the way for similar researches that use different taggers in achieving more accuracy in the tagging process. In previous works, there have been considerable attempts that have moved this process one step forward and we believe this research contributes an additional step in the process of achieving a PoS Tagger for Amharic with more accurate result similar to PoS taggers for languages like English.

5. Conclusion and Future Work

This work used Walta Information Center's news corpus that is manually tagged by Ethiopian Language Research Center of Addis Ababa University. Analysis was made on the corpus to determine the impact of modified tags in the corpus. Shallow stemming was also carried out to separate words that are with Preposition, Conjunction or both and then attempted to tag each token with a basic tag without prepositions or conjunctions. The compound tags that are used to tag words with phrasal structure were removed and 15 tags out of the 31 were left for use.

There are several modules available in NLTK that can be used for tagging purpose. In our case, we used the HMM module to train and tag our corpus. While conducting the research, after applying Shallow Stemming to separate words with phrasal structure, lexicons are generated from the pre-processed corpus. Tenfold cross validation method was applied. After the words have been converted into sentences, the tagset, part of the tagged corpus and lexicons are used in the process of training the tagger.

Once the training was completed, the tagging (testing) process continued and the accuracy of the obtained result from the tagging process is compared with the manually tagged corpus.

Having all said, being able to conduct the research in Amharic Fidel has an advantage in terms of getting the actual result without any complications that arises with transliteration.

As part of future work, working with other taggers like TnT can be considered as TnT tagger uses the assumption of 2nd order of Markov Chain.

References

[1] Webopedia, Available at: http://www.webopedia.com/TERM/N/NLP.html, Last Accessed on 9 June 2014].

[2] cia.gov, Available at: https://www.cia.gov/library/publications/the-world-factbook/geos/et.html, Last Accessed on 17/06/2014.

[3] D. Jurafsky and J. H. Martin, “Speech and Language Processing”, Alan Apt, 1999.

[4] Girma Awgichew and Mesfin Getachew, "Manual Annotation of Amharic News Items with Part-of-Speech Tags and its Challenges", Ethiopian Languages Research Center Working Papers, pp. 1-16, 2006.

[5] Blata Merseahazen Woldeqirqos, “Yamargna Sewasew”, Addis Ababa, 1955.

[6] Baye Yimam, “yamariññа säwasiw (Amharic Grammar)”, Addis Ababa: EMPDA, 1987 E.C.

[7] Mesfin Getachew, "Automatic Part of Speech Tagger for Amharic Language," Addis Ababa, 2001.

[8] B. G. Gebre, "Part of Speech Tagging for Amharic," Unpublished MSc Thesis, Wolverhampton, 2010.

[9] Sisay Adafre, "Part of Speech Tagging for Amharic using Conditional Random Fields," in Proceedings of the ACL Workshop on Computational Approaches to Semitic Language, Ann Arbor, Michigan, June 2005.

[10] B. Gambäck, "Tagging and Verifying an Amharic News Corpus," in Workshop on Language Technology for Normalization of Less-Resourced Languages, 2012.