# Parts of Speech Tagging for Afaan Oromo

Getachew Mamo Wegari
Information Technology Department
Jimma Institute of Technology
Jimma, Ethiopia

Million Meshesha (PhD)
Information Science Department
Addis Ababa University
Jimma, Ethiopia

*Abstract*—**The main aim of this study is to develop part-of-speech tagger for Afaan Oromo language. After reviewing literatures on Afaan Oromo grammars and identifying tagset and word categories, the study adopted Hidden Markov Model (HMM) approach and has implemented unigram and bigram models of Viterbi algorithm. Unigram model is used to understand word ambiguity in the language, while bigram model is used to undertake contextual analysis of words.**

**For training and testing purpose 159 sentences (with a total of 1621 words) that are manually annotated sample corpus are used. The corpus is collected from different public Afaan Oromo newspapers and bulletins to make the sample corpus balanced. A database of lexical probabilities and transitional probabilities are developed from the annotated corpus. These two probabilities are from which the tagger learn and tag sequence of words in sentences.**

**The performance of the prototype, Afaan Oromo tagger is tested using tenfold cross validation mechanism. The result shows that in both unigram and bigram models 87.58% and 91.97% accuracy is obtained, respectively.**

***Keywords-Natural Language processing; parts of speech tagging; Hidden Markov Model; N-Gram; Afaan Oromo.***

## I. INTRODUCTION

At the heart of any natural language processing (NLP) task, there is the issue of natural language understanding. However, the process of building computer programs that understand natural language is not straightforward. As explained in [1], natural languages give rise to lexical ambiguity that words may have different meanings, i.e. one word is in general connected with different readings in the lexicon. Homograph, the phenomenon that certain words showing different morpho-syntatic behavior are identically written. For instance, the word 'Bank' has different meanings; Bank (= financial institute), Bank (= seating accommodation), etc.

In other words, words match more than one lexical category depending on the context that they appear in sentences. For example, if we consider the word miilaa 'leg' in the following two sentences,

| | |
|---|---|
| Lataan kubbaa miilaa xabata. | 'Lata plays football'. |
| Lataan miilaa eeraa qaba. | 'Lata has long leg'. |

In the first sentence, miilaa 'leg' takes the position of adjective to describe the noun kubbaa 'ball'. But in the second sentence, miilaa is a noun described by eeraa 'long'.

Besides ambiguity of words, inflection and derivation of the language are other reasons that make natural language understanding very complex. For instance, tapha 'play' contains the following inflection in Afaan Oromo language.

| | |
|---|---|
| tapha-t | ' she plays' |
| tapha-ta | 'he plays' |
| tapha-tu | 'they play' |
| tapha-ta-niiru | ' they played' |
| tapha-chuu-fi | 'they will play' |

In the above particular context suffixes are added to show gender {–t, --ta}, number { –tu/--u} and future {--fi}.

To handle such complexities and use computers to understand and manipulate natural language text and speech, there are various research attempts under investigation. Some of these include machine translation, information extraction and retrieval using natural language, text to speech synthesis, automatic written text recognition, grammar checking, and part-of-speech tagging. Most of these approaches have been developed for popular languages like English [3]. However, there are few studies for Afaan Oromo language. So, the study presents the investigation of designing and developing an automatic part-of-speech tagger for Afaan Oromo language.

## II. PART-OF-SPEECH TAGGING

Part-of-speech (POS) tagging is the act of assigning each word in sentences a tag that describes how that word is used in the sentences. That means POS tagging assigns whether a given word is used as a noun, adjective, verb, etc. As Pla and Molina [4] notes, one of the most well-known disambiguation problems is POS tagging. A POS tagger attempts to assign the corresponding POS tag to each word in sentences, taking into account the context in which this word appears.

For example, the following is tagged sentence in Afaan Oromo Language.

Leenseen\NN kaleessa\AD deemte\VV 'Lense went yesterday'.

In the above example, words in the sentence, Leensaan kaleessa deemte, are tagged with appropriate lexical categories of noun, adverb and verb respectively. The codes NN, AD, VV are tags for noun, adverb and verb respectively. The process of tagging takes a sentence as input, assigns a POS tag to the word or to each word in a sentence or in a corpus, and produces the tagged text as output.

There are two efficient approaches that have been established to develop part-speech-tagger [14].

### A. Rule based Approach

Rule based taggers use hand coded rules to determine the lexical categories of a word [2, 13]. Words are tagged based on the contextual information around a word that is going to be tagged. Part-of-speech distributions and statistics for each word can be derived from annotated corpora - dictionaries. Dictionaries provide a list of word with their lexical meanings. In dictionaries there are many citations of examples that describe a word in different context. These contextual citations provide information that is used as a clue to develop a rule and determine lexical categories of the word.

In English language, for instance, a rule changes the tag from modal to noun if the previous word is an article. And the rule is applied to a sentence, the/art can/noun rusted/verb. Brill's rules tagger conforms to a limited number of transformation types, called templates. For example, the rule changes the tag from modal to noun if the previous word is an article, corresponds to template. The following table shows sample template that is used in Brill's rule tagger [2].

TABLE I. SAMPLE TEMPLETE BRILL'S RULE

| Rules | Explanation |
|---|---|
| alter(A, B, prevtag(C)) | Change A to B if preceding tag is C |
| alter(A, B, nexttag(C)) | Change A to B if the following tag is C |

Where, A, B and C represent lexical categories or part-of-speech.

### B. Stochastic Approach

Most current part-of-speech taggers are probabilistic (stochastic). It is preferred to tag for a word by calculating the most likely tag in the context of the word and its immediate neighbors [15, 16]. The intuition behind all stochastic taggers is a simple generalization of the 'pick the most-likely tag for this word' approach based on the Bayesian framework. A stochastic approach includes most frequent tag, n – gram and Hidden Markov Model [13].

HMM is the statistical model which is mostly used in POS tagging. The general idea is that, if we have a sequence of words, each with one or more potential tags, then we can choose the most likely sequence of tags by calculating the probability of all possible sequences of tags, and then choosing the sequence with the highest probability [17]. We can directly observe the sequence of words, but we can only estimate the sequence of tags, which is 'hidden' from the observer of the text. A HMM enables us to estimate the most likely sequence of tags, making use of observed frequencies of words and tags (in a training corpus) [14].

The probability of a tag sequence is generally a function of:

- the probability that one tag follows another (n-gram); for example, after a determiner tag an adjective tag or a noun tag is quite likely, but a verb tag is less likely. So in a sentence beginning with the run…, the word 'run' is more likely to be a noun than a verb base form.
- The probability of a word being assigned a particular tag from the list of all possible tags (most frequent tag); for example, the word 'over' could be a common noun in certain restricted contexts, but generally a preposition tag would be overwhelmingly the more likely one.

So, for a given sentence or word sequence, HMM taggers choose the tag sequence that maximizes the following formula [14]:

$$\underbrace{P(\text{word/tag })}_{\substack{\text{Most frequent}\\ \text{tag}\\ \text{(likelihood)}}} * \underbrace{P(\text{tag/previous n tags})}_{\substack{\text{N-gram (a}\\ \text{prior)}}}$$

## III. AFAAN OROMO

Afaan Oromo is one of the major languages that is widely spoken and used in Ethiopia [6]. Currently it is an official language of Oromia state. It is used by Oromo people, who are the largest ethnic group in Ethiopia, which amounts to 34.5% of the total population according to the 2008 census [19].

With regard to the writing system, since 1991 Qubee (Latin-based alphabet) has been adopted and become the official script of Afaan Oromo [12]. Currently, Afaan Oromo is widely used as both written and spoken language in Ethiopia. Besides being an official working language of Oromia State, Afaan Oromo is the instructional medium for primary and junior secondary schools throughout the region and its administrative zones. It is also given as the department in five universities in Ethiopia. Thus, the language has well established and standardized writing and spoken system [7].

## IV. RELATED RESEARCHES

To use computers for understanding and manipulation of Afaan Oromo language, there are very few researches attempted. These attempts include text-to-speech system for Afaan Oromo [8], an automatic sentence parser for Oromo Language [9] and developing morphological analyzer for Afaan Oromo text [10].

There are also other related researches that were conducted on other local language. Specially on Amharic language, two researches were conducted on POS tagging by [5] and [11], but to the best of our knowledge there is no POS tagging research conducted for Afaan Oromo language.

## V. APPLICAION OF THE STUDY

The output of POS tagger has many applications in many natural language processing activities [4]. Morpho-syntactic disambiguation is used as preprocessor in NLP systems. Thus, the use of a POS tagger simplifies the task of syntactic or semantic parsers because they do not have to manage ambiguous morphological sentences. Thus parsing cannot proceed in the absence of lexical analysis, and so it is necessary to first identify and determine part-of-speech of words.

It can also be incorporated in NLP systems that have to deal with unrestricted text, such as information extraction, information retrieval, and machine translation. In this modern world, huge amount of information are available on the Internet in different languages of the world. To access such information we need machine translator to translate into local languages. To develop a machine translation system, the lexical categories of the source and target languages should be analyzed first since a translator translates, for example, nouns of the source language to the nouns of the target language. So, POS tagger is one of the key inputs in machine translation processes.

A word's part-of-speech can further tell us about how the word is pronounced. For instance, the word 'content' in English can be a noun or an adjective. It is pronounced as 'CONtent' and 'conTENT' respectively. Thus, knowing part-of-speech can produce more natural pronunciations in a speech synthesis system and more accuracy in a speech recognition system [8].

All these applications can benefit from POS tagger to improve their performance in both accuracy and computational efficiency.

## VI. METHODOLOGY

### A. Algorithm Design and Implementation

HMM approach is adopted for the study since it does not need detail linguistic knowledge of the language as rule based approach [14]. Viterbi algorithm is used for implementing the tagger.

The Viterbi algorithm is a dynamic programming algorithm that optimizes the tagging of a sequence, making the tagging much more efficient in both time and memory consumption. In a naïve implementation it would calculate the probability of every possible path through the sequence of possible word-tag pairs, and then select the one with the highest probability. Since the number of possible paths through a sequence with a lot of ambiguities can be quite large, this will consume a lot more memory and time than necessary [18].

Since the path with highest probability will be a path that only includes optimal sub paths, there is no need to keep sub paths that are not optimal. Thus, the Viterbi algorithm only keeps the optimal sub path of each node at each position in the sequence, discarding the others.

### B. Test and Evaluation

The prototype tagger is tested based on the sample test data prepared for this purpose. The performance evaluation is analyzed based on correctly tagged once by the prototype tagger. The performance analysis is using tenfold cross validation. Ten fold cross validation divides a given corpus in to ten folds. And nine folds are used for training and the tenth fold is used for testing. It provides an unbiased estimate of value of prediction error and preferred for small sample corpus [20].

## VII. AFAAN OROMO TAGSET AND CORPUS

### A. Afaan Oromo Tagsets

Since there is no tagset prepared for natural language processing purpose for Afaan Oromo language, seventeen tags have been identified for the study as indicated in Table II.

TABLE II. TAGSETS

| Tags | Description |
|---|---|
| NN | A tag for all types of nouns that are not joined with other categories in sentences. |
| NP | A tag for all nouns that are not separated from postpositions. |
| NC | A tag for all nouns that are not separated from conjunctions. |
| PP | A tag for all pronouns that are not joined with other categories. |
| PS | A tag for all pronouns that are not separated from postpositions. |
| PC | A tag for all pronouns that are not separated from conjunctions. |
| VV | A tag for all main verbs in sentences. |
| AX | A tag for all auxiliary verbs. |
| JJ | A tag for all adjectives that are separated from other categories. |
| JC | A tag for adjectives that are not separated from conjunction. |
| JN | A tag for numeral adjectives. |
| AD | A tag for all types of adverbs in the language. |
| PR | A tag for all preposition/postposition that are separated from other categories. |
| ON | A tag for ordinary numerals. |
| CC | A tag for all conjunctions that are separated from other categories. |
| II | A tag for all introjections in the language. |
| PN | A tag for all punctuations in the language. |

### B. Corpus

The collected corpus for the study was manually tagged by experts of linguists in the field. The tagging process is based on the identified tagset and corpus that is manually tagged, considering contextual position of words in a sentence. This tagged corpus is used for training the tagger and evaluates its performance. The total tagged corpus consists of 159 sentences (the total of 1621 tokens).

## VIII. THE LEXICON

Lexicon was prepared from which the two probabilities are developed for the analysis of the data set.

TABLE III. SAMPLE OF LEXCON

| words | NN… | PP… | VV… | JJ… | AD… | Total |
|---|---|---|---|---|---|---|
| nama | 2 | 0 | 0 | 1 | 0 | 3 |
| Yeroo | 0 | 0 | 0 | 0 | 9 | 9 |
| . | . | . | . | . | . | . |
| . | . | . | . | . | . | . |
| . | . | . | . | . | . | . |
| Total | 334 | 100 | 351 | 226 | 81 | 1621 |

### A. *Lexicon probability*

The lexical probabilities have been estimated by computing the relative frequencies of every word per category from the training annotated corpus. All statistical information, that enables to develop probabilities, are derived automatically from a hand annotated corpus (the lexicon).

For instance, the lexical probability of the word *Oromoon* tagged with *NN* is calculated as:

C(Oromoon, NN) = 7

C(NN) = 334

So, P(Oromoon/NN) = C(Oromoon, NN)/C(NN)

= 7/334

= 0.0206

Where, C and P are count of and Probability, respectively.

TABLE IV. SAMPLE LEXICAL PROBABILITY

| Words with given lexical probability | Probability |
|---|---|
| P(Oromoon/NN) | 0.0206 |
| P(jedhaman/VV) | 0.0052 |
| P(kabajaa/AD) | 0.02174 |
| P(ayyaanichaafi/NC) | 0.11111 |
| P(amma/AD) | 0.04348 |
| P(yeroo/AD) | 0.10869 |

### B. *Transition Probability*

In transitional probabilities, the information of one part-of-speech category preceded by other categories is developed from training lexicon corpus. For this study, bigram is used. Bigram considers the information of the category (t-1) preceded the target category (t).

That means, P(t/t-1), where t is – part-of-speech category.

For example, C($S) = 157

C(NN,$S) = 79

P(NN/$S) = C(NN, $S)/C($S)

= 79/157

= 0.5032

TABLE V. SAMPLE TRANSITION PROBABILITY

| Bigram Category | Probability |
|---|---|
| P(NN/$S) | 0.5032 |
| P(VV/$S) | 0.0063 |
| P(NN/VV) | 0.1538 |
| P(NN/PN) | 0.0063 |
| P(JJ/NN) | 0.2695 |
| P(JJ/$S) | 0.1465 |
| P(PP/NN) | 0.1018 |

## IX. AFAAN OROMO PARTS OF SPEECH TAGGER

The tagger learns from the two probabilities to label appropriate tag to each word in sentences. The tagger for the study is developed from Viterbi algorithm of hidden Markov model.

### A. *Performance Analysis of the tagger*

TABLE VI. AVERAGE TAGGER RESULTS

| Unigram | Bigram |
|---|---|
| 87.58% | 91.97% |

In the performance analysis, the tagger is repeatedly trained and tested following tenfold cross validation.

The algorithms of the tagger are tested with a corpus of 146 Afaan Oromo words in average in each test set and that is trained on the training set of 1315 words, and the result of each test are compared with a copy of the test set that is hand annotated. As a result, the results of the experiments for both bigram and unigram algorithms show an accuracy of 91.97% and 87.5% correctly tagged words in average respectively.

With this corpus, the distributions of accuracy performance in both models are not as far from each other. The maximum variation in the distribution of bigram and unigram models is 8.97 and 11.04 respectively. If the corpus is standardized, this variation will reduce since standardized corpus consist relatively complete representative of words for the language and fair distribution of words in training set and test are observed.

In bigram model, the statistical accuracy is performed more than unigram model. Bigram model uses probability of contextual information besides the highest probability of categories given a word in a sentence to tag the word. The difference accuracy rate from bigram to unigram is 4.39% with this dataset.

This indicates, contextual information (the position in which the word appear in sentence) affects the determination of word categories for Afaan Oromo language.

## ACKNOWLEDGMENT

The support of Jimma University, Ethiopia, is greatly acknowledged. The authors would also like to acknowledge the help and support from Tigist Mazgabu, Dr. Mengesha Mamo and Mr. Fituma Tefera.

## REFERENCES

[1] Hermann Helbig. Knowledge representation and the semantics of natural language. Springer-Verlg Berlin Heidelberg, Germany, 2006.

[2] James Allen. Natural language Understanding. The Benjamin/Cummings Publishing company, Redwood City, Canada, 1995

[3] Gobinda G. Chowdhury. Natural Language Processing: Department of Computer and Information Sciences, University of Strathclyde, Glasgow G1 1XH, UK, http://www.cis.strath.ac.uk/cis/research/publications/papers/strathcis_publication_320.pdf

[4] Ferran Pla and Antonio Molina. Natural Language Engineering: Improving part of- speech tagging using lexicalized HMMs_ 2004. Cambridge University Press, United Kingdom, 2004

[5] Mesfin Getachew. Automatic part-of-speech tagging for Amharic language an experiment using stochastic Hidden Markov Approach. MSc. Thesis. School of Graduate Studies, Addis Ababa University, 2001.

[6] Abara Nefa. Long Vowels in Afaan Oromo: A Generative Approach. M.A. Thesis. School of Graduate Studies, Addis Ababa University, 1988. Unpublished.

[7] Kula K. T., Vasudeva Varma and Prasad Pingali. Evaluation of Oromo-English Cross-Language Information Retrieval. In IJCAI 2007 Workshop on CLIA, Hyderabad (India), 2007.

[8] Morka Mekonnen. Text to speech system for Afaan Oromo. MSc Thesis. School of Graduate Studies, Addis Ababa University, 2001.Unpublished

[9] Diriba Magarsa. An automatic sentence parser for Oromo language. MSc Thesis. School of Graduate Studies, Addis Ababa University, 2000. Unpublished

[10] Assefa W/Mariam. Developing morphological analysis for Afaan Oromo text. MSc Thesis. School of Graduate Studies, Addis Ababa University, 2000. Unpublished

[11] Yenewondim Biadgo. Application of multilayer perception neural network for tagging part-of-speech for Amharic language. MSc Thesis. School of Graduate Studies, Addis Ababa University, 2005. Unpublished

[12] Gumii Qormaata Afaan Oromo. Caasluga Afaan Oromo. Komoshinii Aadaafi Tuurizimii Oromiyaa, 1996. Unpublished

[13] Pierre M. Nugues. An Introduction to Language Processing with Perl and Prolog. Springer-Verlag Berlin Heidelberg, Germany, 2006

[14] Daniel Jurafsky and James H. Martin. Speech and Language Processing: An Introduction to Natural Speech Recognition. Prentice-Hall, Inc., 2000.

[15] Sandipan Dand, Sudeshna Sarkar, Anupam Basu. Automatic Part-of-Speech Tagging for Bengali: An Approach for Morphologically Rich Languages in a Poor Resource Scenario. Department of Computer Science and Engineering Indian Institute of Technology Kharagpur, 2007

[16] Frank Van Eynde. Part-of-speech Tagging and Lemmatisation for the Spoken Dutch Corpus, Center for Computational Linguistics Maria-Theresiastraat 21 3000 Leuven, Belgium, 2000

[17] Roger Garside and Nicholas Smith. A Hybrid Grammatical Tagger: CLAWS4, http://ucrel.lancs.ac.uk/papers/HybridTaggerGS97.pdf

[18] Simon STÅHL. Part-of-Speech Tagger for Swedish, Computer Science, Lund University, 2000

[19] Census report: Ethiopia's population now 76 million. December 4th, 2008. http://ethiopolitics.com/news

[20] Yoshua Bengio and Yves Grandvalet. No Unbiased Estimator of the Variance of K-Fold Cross-Validation. Dept. IRO, Universit´e de Montr´eal C.P. 6128, Montreal, Qc, H3C 3J7, Canada, 2004 http://www.faqs.org/faqs/ai-faq/neural-nets/part3/section-12.html