Amharic Text Chunker using Conditional Random Fields

Birhan Hailu
Department of IT, Assosa University
Assosa, Ethiopia

Birchiko Achamyeleh
Department of IT, Assosa University
Assosa, Ethiopia

Gebeyehu Belay
Department of IT,Bihar dare University
Bihar dare, Ethiopia

ABSTRACT
This paper introduces Amharic text chunkerusing conditional random fields. To get the optimal feature set of the chunker; the researchers' conduct different experiments using different scenarios until a promising result obtained. In this study different sentences are collected from Amharic grammar books, new articles, magazines and news of Walta Information Center (WIC) for the training and testing datasets. Thus, these datasets were analyzed and tagged manually and used as a corpus for our model training and testing. The entire datasets were chunk tagged manually for the training dataset and approved by linguistic professionals. For the identification of the boundary of the phrases IOB2 chunk specification is selected and used in this study. The result of all experiments is reported with the maximum overall accuracy off 97.26%, with a window size of two on both sides, with their corresponding POS tag of each token and the worst performance achieved is 84.57%, with only the window size of one word on both the left and right sides.

Keywords
Amharic text chunker, base phrase chunker, conditional random fields, clause boundary identification

1. INTRODUCTION
Now a day, Natural language processing (NLP) has an important role in our daily life, by enabling computers to understand human languages [1]. Text chunking (TC) is one of the essential tasks in NLP applications. The Information generated by this task can be helpful for many purposes, including automatic summarizing or question answering, information extraction. It can be either shallow parsing or deep parsing [2].

Text chunking or shallow parsing is a kind of NLP task, which is the process of grouping the input text or sentence into syntactically related non-overlapping part of words or chunks like NP (Noun Phrase), VP (Verb Phrase), PP (Prepositional Phrase), AdjP (Adjective Phrase) and so on. Generally, the notion of TC is a word (chunk) that should be a member of one syntactic structure(phrase), word(chunk) can't be a member of two or more syntactic structures(phrase).

For instance, according to [1]: ትንሹልጅትንሽእንጀራበላ "The little boy ate little Injera"

Here the sentence can be segmented into three basic syntactic non-overlapping phrases: a noun phrase ትንሹልጅ "the little boy", another noun phrase ትንሽእንጀራ "little Injera", a verb phrase በላ "ate".

The various studies indicate that many NLP and IR application needs chunking (shallow parsing or partial parsing) rather than full parsing. For instance, to develop a fully-fledged question answering system it needs Chunker as indicated by [3,4,5,6], Finding noun phrases and verb phrases may be enough for the IR system. Commonly, TC can be integrated into information extraction, text mining, automatic summarization, and so on.

Today different researchers investigated text-chunking for the various language of the world using a different methodology, for instance, Arabic[2], Chinese[6], Bengali[3], Amharic [7], Indonesian [8], Urdu [9]. As far as the researchers' knowledge concerned, only one work was done for Amharic. The work done by Abeba [8] is the first Amharic TC, their work didn't include all kinds of Amharic sentences. They used HMM for chunking the text and then error-pruning rules to correct chunks that are incorrectly chunked by the HMM model. However as indicated by [7] and [3], HMM model has a label bias problem. Even if the above works have been done in different languages on different aspects of text chunking, it is not possible to apply this text chunker to the Amharic language context directly. The main reason is the Amharic language is highly inflectional, morphologically rich, and has its own letters and grammatical rules.

Conditional Random Fields (CRFs) are one class of supervised ML, it was applied for different IR and NLP applications, where they are used for classification of sequential data [2].it were first introduced by [11] for the different sequential annotation tasks. Some of them were named entity recognition, POS tag, text chunking, etc., where they get excellent results ([7],[3],[2]). Due to the above reason, we are motivated to investigate Text chunker for Amharic language using them (i.e., CRFs). The rest of the paper is organized as follows. Statement of the problem are explained in section 2. dataset description is explained in section 3, chunk specification are explained in section 4, architecture of CRFs based Amharic text chunker are explained in section 5, Feature representation and description are explained in section 6, experiment and result are explained in section 7, concluding remarks are given in section 8.

2. STATEMENT OF THE PROBLEM
The motivation that initiated the researcher to conduct this research toward Amharic text chunker is the advantageous and application areas of text chunker in different NLP and IR tasks. for instance, in information retrieval task dividing the given search keyword text into different syntactically correlated part of a chunk can increase the performance of the searching time and also in phrase level machine translation chunker can be one component which helps the translation process by dividing the stream of text into a chunk(phrase).

Amharic ranked as the second by the number of the speaker under Semitic language in the world next to Arabic[12]. Despite having a large speaker population, the language has few computational linguistic resources. So, conducting research on such kind of language is advisable to overcome the current shortage of computational linguistic resources. Amharic is one of the under-resourced languages. Hence, there is no well-designed full-fledged Amharic text chunker

that could contribute towards designing NLP. So that conducting Amharic ATC with the best performance will have a contribution to advance research in NLP for the Amharic language.

As described earlier, only one attempt was done by Abeba[1] and further improvement regarding Amharic text chunker has never been tried after the first ATC attempt. As the researcher's knowledge concerned in the Nationwide also there is no attempts for developing a text chunker of other Ethiopian languages. The first attempt to ward Amharic text chunker[1] used Hidden Markov Model (HMM) to develop the Amharic base phrase chunker and bottom-up approach with a transformation algorithm to transform the chunker to the parser. However, the size of the corpus is small, and also as indicated by [3], [7] the model they use (i.e. HMM) has a label bias problem. The researchers minimized the error incorrectly chunked by HMM model using the error pruning rule, but major errors that are not pruned by rules are the tag sequence conflict; by taking these limitations into consideration we are motivated to further investigate the Amharic TC. So, in the proposed system the researcher tried to reconsider features that could increase the accuracy of Amharic text chunker.

Therefore, the research questions are defined and articulated as follows:

1. How to optimize the accuracy of an automatic Amharic text Chunker, and what contexts need to be reconsidering for its best achievement?

2. How to optimize the influential or (determinant) factor (s) to improve the performance of the proposed systems? How it is important than others?

3. DATASET DESCRIPTION

As indicated above we use machine-learning to design Amharic text chunker. Mostly, this approach needs syntactically annotated corpus. Using this approach, the success or failure rate of different NLP applications depends on the quality and availability of appropriate datasets and also selecting optimal features of the target language. The term dataset in computational linguistics mostly called the corpus (plural corpora). Corpora can be categorized into annotated and unannotated corpora. Annotated corpora are a collection of text with some syntactic notation like POS tag, chunk tag, NER tag, etc. While as unannotated corpora are a collection of large amounts of text without any syntactic notations. supervised machine learning often uses annotated corpora while as unsupervised use unannotated corpora. Since our investigation follows a supervised machine learning approach so we use annotated dataset.

Every text chunker which is implemented using a machine-learning approach needs two types of dataset[13]: training and testing dataset. Training dataset used to train and create a model for machine learning components and after completing the training the performance of the model is evaluated using the test dataset.

In this thesis work, we have collected 450 Amharic sentences that has total of 4,020 tokens from WIC corpus, Amharic grammar book and magazines. The sentences which were collected from WIC corpus was tagged by POS tag but the other data which were collected from Amharic grammar book and magazines didn't tagged by POS tag due to that we manually tagged them and receive comment and suggestion on it from linguistic expert. Only POS tagging is not enough for training the model in addition to it, it needs chunk tag. To the best knowledge of the researchers, there is no publicly available chunk tagged documents. For this reason, sentences are chunk tagged manually by the researchers.

To sum up,31 tag sets were used for POS tagging the corpus, and 11 tag sets were used for Chunk tagged the corpus. These sample sentences were transcribed according to the Amharic ፊደል (Fidel or alphabet) Unicode standard. After completing the corpora preparation by the researchers' comment and suggestions received from a linguistic expert.

4. CHUNKING SPECIFICATION

To identify the chunks, it is necessary to find the positions where a chunk can end and a new chunk can begin. The POS tag assigned to every token is used to discover these positions. There are four kinds of complete chunk boundary representations namely IOB1, IOB2, IOE1, and IOE2 [14]. In this study, to identify the boundaries of each chunk in sentences the IOB2 tag set is used for chunk tagged annotated text. Here "I" is a token inside a chunk, "O" is a token outside a chunk and "B" is a token that exists at the beginning of the chunk. See the following example በቤንች-ማጂዞን 30ሺህዝብለመምረጥተመዘገበ "in benchmaj zone thirty thousand people register for voting" with the chunk representation in Table 1.

Table 1: Chunk representation

Word	IOB2 chunk
በቤንች-ማጂ	B-NP
ዞን	I-NP
30ሺ	B-NP
ህዝብ	I-NP
ለመምረጥ	B-VP
ተመዘገበ	I-VP

5. ARCHITECTURE OF CRFs BASED AMHARIC TEXT CHUNKER

The Amharic text chunker is intended in a means that, first it learns properties and parameters associated with chunk tag from the training data. It then receives input words with their POS tag and predicts the possible chunk tags. The architecture has two processes: the learning process and prediction process.

The Training is handled by components in the learning process. Preprocessing is initially performed on the training corpus. The corpus is chunk tagged based on the IOB2 chunk specification format. Then it is passed to the Amharic text chunker (ATC) encoder which identifies a token and its corresponding POS and chunk tag through encoding. The tokens and tag sequence generated by the ATC encoder is handled to the Feature Extractor. Feature Extractor extracts necessary features to identify chunk tag based on the generated token and tag sequence, the extracted features will then be used as an input to the Model Builder. After the above all fulfilled, the builder begins the model building procedure to generate a trained model. Generally, here in the learning process, based on the training corpus we have generated a model that used to predict chunk tag classes of testing.

The testing phase is the final phase in the ATC system. It is a process of recognizing the chunk tag of the given preprocessed tokens. The knowledge in model builder contains features that are extracted and stored during the training phase, are supplied to the recognizer to identify chunk tag from the given Amharic text. Identification of chunk tag is performed based on the calculated probability.

As described above CRFs are probabilistic model, it computes the probability of P(X|Y) of possible outputY={Y1……Yn} Yn for the given input X={X1…….Xn} Xn [21].In case of text chunking Y is related with sequence of chunk tag and X is related with a sequence of word. Linear chain CRFs is a special form of CRFs, which is structured as a linear chain that models the output variables as a sequence. It can be shown using the following mathematical formula [15].

$$Z_{\vec{\lambda}}(\vec{X}) = \sum_{\vec{y} \in Y} exp\left(\sum_{j=1}^{n} \sum_{i=0}^{m} \lambda_i \, f_i(y_{j-1}, y_j, \vec{X}, j) \right)$$

From the above equation, n indicates the length of the sentences, m indicates the number of feature templates, j indicates the position of the input sequence, λi indicates the weights assigned to the different features in the training phase, is normalization factor that makes the probability in the range [0,1], which can be expressed as(Roman & Katrin, 2007) :

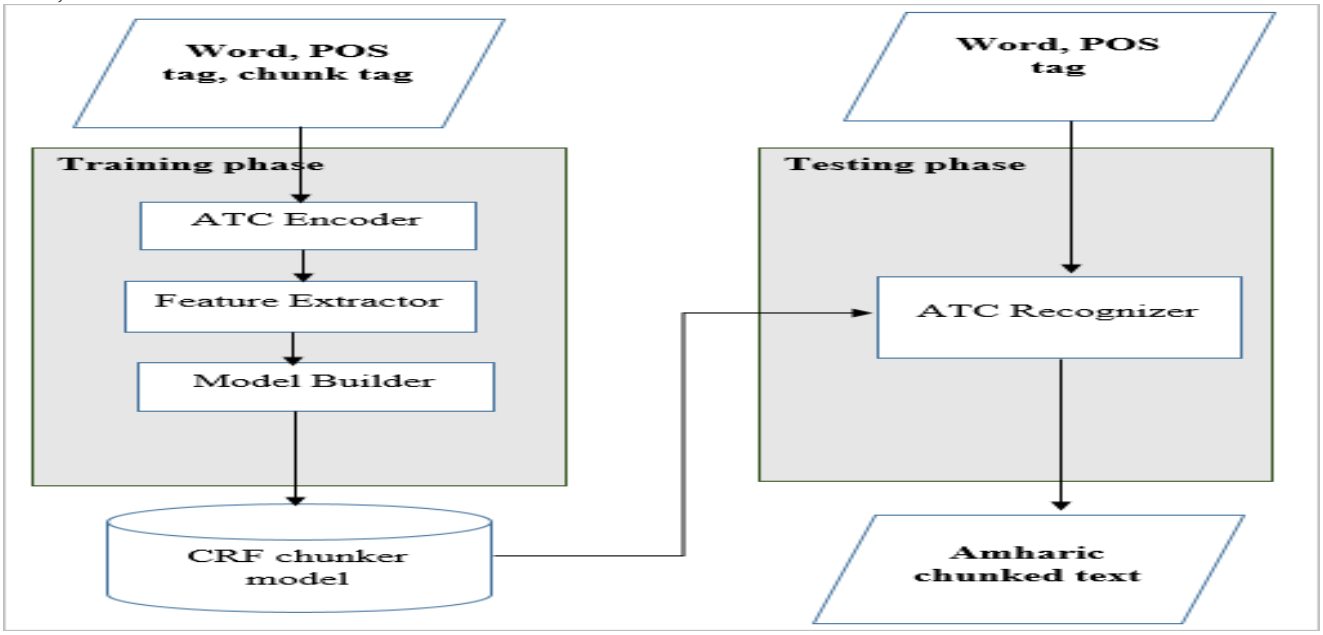

Fig. 1. Architecture of CRFs based ATC

6. FEATURE REPRESENTATION AND DESCRIPTION

Feature selection plays a crucial role in the CRFs framework. Experiments were carried out to find out the most suitable features for the Amharic text chunking task. The main features for text chunking tasks have been identified based on the different possible combinations of available word and tag context. In this study, it has been considered a different combination from the following set for inspecting the best feature set for the Amharic text chunker task. The various features used for developing our system are described below:

- Surrounding word: Preceding and following words of a current word can be used as a feature because surrounding words influence the current word. We have considered different window sizes of the words until we get a promising result.
- Combination of words: combinations such as the preceding/current word and current word/following words are used as features.
- The window for POS of the current word: Preceding and following POSs of a present word are used as a feature because surrounding POSs influence the present word.
- Combination of POSs: combinations such as the preceding/current POS and current/following POS is used as a feature.

Our empirical study found that the following combination of features gives the optimal feature set.

F(best)=[Wi-2Wi-1WiWi+lWi+2,POSi-2POSilPOSiPOSi+1POSi+2]

7.EXPERIMENTAL RESULT ANDDISCUSTION

In this study, a total of 450 sentences were used and manually annotated with chunk tags. From the total 90 % used for model creation and the rest 10% used for testing the model. The detail of corpus statistics is shown in Table 2

To implement and evaluate the model, we have adopted the package CRFsuite. Initially CRFs was introduced by [16], it has a CRFsuite package, which is open source; freely available conditional random fields implementation package, we have used it to develop Amharic text chunker CRFs model.

Table 2: corpus statistics

Number of tokens in the training set	Number of tokens in the testing set	Total number of tokens
3,618 tokens	402 tokens	4,020 tokens

To do the experiment, four different scenarios wereconsidered. In the first scenario, one word left and one word to the right from the current word (wi-1 wi wi+1). The detail of the experiment result is shown in the below table 3.

Table 3. Experiment one

Features used	Extracted Features	Time taken	Accuracy
w_{i-1} w_i w_{i+1}	41,556	3.33 s	84.57%

In experiment two, we used the previous experiment one features and add their corresponding POS tag of each token used as a feature. The accuracy increased by 9.95. the detailof the experiment result is shown in the below table 4.

Table 4 Experiment two

Features used	Extracted Features	Time taken	Accuracy
$w_{i-1}w_i$ $w_{i+1,}$ $POS_{i-1}POS_iPOS_{i+1}$	42,384	3.64 s	94.52%

In experiment three, two words to the left and two wordsfrom the right of the current word used as a feature. The detail is of the experiment resultshown in the below in table 5.

Table 5. Experiment three

Features used	Extracted Features	Time taken	Accuracy
W_{i-2} w_{i-1} w_i w_{i+1} w_{i+2}	66,252	3.04 s	85.32%

In experiment four, we used the above experiment three and add the corresponding POS tag of each token as a feature. The detailof the experiment result is shown in the below table6.

Table 6. Experiment four

Features used	Extracted Features	Time taken	Accuracy
W_{i-2} w_{i-1} w_i w_{i+1} w_{i+2}&$POS_{i-2}POS_{i-1}POS_iPOS_{i+1}POS_{i+2}$	67,656	3.75 s	97.26%

8. CONCLUSION

This work tried to contribute one important tool which plays a role for overcoming some challenges in NLP regarding to Amharic. Having a good Amharic text chunker could be used in information extraction of any domain specific tasks, question answering, information retrieval etc. to come up with a system that could be used in the above application areas, identifying the optimal feature set is the most important task of any text chunking task. This work is also delimited to identifying these optimal features set to recognize Amharic phrases which enhance the accuracy of existing Amharic text chunker.

To do this research, 450 sample sentences were taken from WIC corpus, Amharic grammar book, news article, magazines, and chunk tagged manually. These datasets are classified into 90% training and the rest 10 % used for testing. By conducting different experiments, we have been identified optimal features for ATC, we have conducted four experiments and the roles of each feature in different experiments were tested.

The highest accuracy achieved in this work using CRFs is 97.26%, with a window size of two on both sides, with their corresponding POS tag of each token. The worst performance achieved is 84.57%, with only the window size of one word on both the left and right sides.

From the findings, two words left and two words right from the current words and the corresponding POS tags of each token are the observed optimal feature sets for recognizing Amharic text chunker.

9. REFERENCES

[1] A. Ibrahim, “A Hybrid Approach to Amharic Base Phrase Chunking and Parsing,” Addis Abeba University, 2013.

[2] N. Khoufi, C. Aloulou, and L. H. Belguith, “Chunking Arabic Texts Using Conditional Random,” IEEE, pp. 428–432, 2014.

[3] K. Sarkar and V. Gayen, “Bengali Noun Phrase Chunking Based on Conditional Random Fields,” IEEE, pp. 148–153, 2014.

[4] K. H. AMARE and A, “Tigrigna question answering system for factoid questions,” Addis Abeba University, 2016.

[5] D. Abebaw, “LETEYEQ (ሌጠየቅ)-A Web Based Amharic Question Answering System for Factoid Questions Using Machine Learning Approach,” Addis Abeba University, 2013.

[6] Muhe Seid, “TETEYEQ: Amharic Question Answering System for Factoid Questions,” Addis Abeba University, 2009.

[7] Y. Zhao and T. Zhao, “Exploiting clause boundary information as features for Chinese functional chunk parsing,” IEEE, pp. 874–878, 2016.

[8] A. Ibrahim and Y. Assabie, “Hierarchical Amharic Base Phrase Chunking Using HMM with Error Pruning,” Springer Int. Publ. Switz., vol. 8387, pp. 126–135, 2014.

[9] X. Vwhp, "shallow parsing natural language processing implementation for intelligent automatic customer service," IEEE, pp. 274–279, 2014.

[10] W. Ali, M. K. Malik, S. Hussain, S. Shahid, and A. Ali, "urdu noun phrase chunking," IEEE, pp. 494–497, 2010.

[11] A. M. and F. C. N. P. J. Lafferty, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," Proc. Eighteenth Int. Conf. Mach. Learn. (ICML 2001), pp. 282–289, 2001.

[12] "CSA (Central Statistics Agency), Addis Ababa, Ethiopia: Central Statistics Agency," http://www.csa.gov.et, 2007. .

[13] G. B. Kumar, "UCSG Shallow Parser : A Hybrid Architecture for a Wide Coverage Natural Language Parsing System," 2007.

[14] Taku Kudo, "Machine Learning and Data Mining Approaches to to Practical Natural Language Processing," Nara Institute of Science and Technology, 2003.

[15] K. Roman and T. Katrin, "Classical Probabilistic Models and Conditional Random Fields," Dortmund, 2007.

[16] A. M. and F. C. N. P. J. Lafferty, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," Proc. Eighteenth Int. Conf. Mach. Learn. (ICML 2001), pp. 282–289, 2001.