International Journal of Academic Multidisciplinary Research (IJAMR)
ISSN: 2643-9670
Vol. 5 Issue 10, October - 2021, Pages: 72-75

Automatic Sentence Parser for Amaharic Language Using SVM

Birchiko Achamyeleh[1], Birhan Hailu[2], Getachew Mamo[3]

[1]Department of IT, Assosa University, Assosa,, ,Ethiopia.
Email- Birchikoa@gmail.com
[2]Department of IT, Assosa university ,Assosa Ethiopia,
email:Berhanhailu1@gmail.com
[3]Department of Computing, Jimma university, Jimma, Ethiopia
Email- Getachew.mamo@yahoo.com

Abstract— *Natural language processing is a research area which is becoming increasingly popular each day for both academic and commercial reasons. Higher level NLP systems (e.g., machine translation) are materialized only when the lower ones (e.g., part-of-speech tagger, syntactic parser) are successfully built. The objective of this research is to build and evaluate Automatic Amharic sentence parser using supervised machine learning approach specifically by using support vector machine. In order to conduct the experiment, we adopt LIBSVM package, which is a library for Support Vector Machines (SVMs). To do this research, 370 sample sentences were taken from WIC corpus, Amharic grammar book, news article, magazines; manually parsed all the sentence by the researcher and comment and suggestion given from linguistics expert. These datasets are classified into 90% is for training and the rest 10 % is used for testing. Experiments have been conducted in this study using the training set and test set. Finally, the overall model accuracy of the mode that we have got is 98.913.*

***Keywords*—** Parser, Amharic text Parser using SVM ,text parser, parser.

I. INTRODUCTION

Natural language processing have an important role in our daily life, by enabling computers to understand human languages. Sentence parsing is also one of the task in NLP. it is the process of identifying the syntactic structure of a specific sentence. Or deals with analyzing a sentence that generally consists of segmenting a sentence into words, grouping these words into a certain syntactic structural units.

Sentence parsing serves as intermediate component for different higher level nlp application like:-grammar checking, question answering word sense disambiguation and so on. now a day parsing underlies most of the applications in natural language processing. .a sentence parser outputs a parse structure that could be used as a component in many applications

STATEMENT OF PROBLEM

as described before, Sentence parsing serves as intermediate component for different higher level NLP applications like grammar checking It is also helpful for language experts in teaching the structure of the language. As indicated by [Daniel gochel, 2003][1] Spell checker, grammar checker question answering and word sense disambiguation are among the applications that require sentence parser for successful and full-fledged implementation. As indicated by [Abeba,2014][2], [Daniel Gochel , 2003], [Atalech,2002][3]
Sentence parsing plays a significant role in reducing overall system complexity Even if parser has a lot of advantage for higher-level NLP application, Amharic language has not benefited from it due to lack of good sentence parser. Few researchers where done for sentence parsing in Amharic language by rule based and statistical approach example [atalech,2002] The study attempted to develop a simple automatic parser for Amharic sentences They were used Inside Outside algorithm with bottom up chart parsing strategy. In this research, a small sample corpus (i.e., 100 sentences) was selected for their experiment. The limitation of Atalech work are as follow:-

The parser developed to parse only 4-word long Amharic simple sentences.
The sentences were only from similar type of sentences, i.e., simple declarative type of sentences which are composed of four words

In addition to statistical tagger the corpus size they use is too small Hence the development of these sentence parsing will have a vital role in the implementation of higher-level NLP application

Research questions

This study attempts to address the following research questions:

- ✓ Does the designed model will bridge the gap encountered in Amharic sentence parsing? .
- ✓ How to build a sentence parser which parse all type of Amharic sentence except imperative and interrogative sentences?

Objective of the study

General objective

The general objective of the research is to investigate automatic sentence parser for Amharic sentences using machine learning approach

Specific objective

- ➢ To review related literature in order to understand the state of the art and different methods that are used to solve the problem

- To study Amharic language in order to have good understand on the syntactic structure of Amharic grammar
- To collect data from different source which used for experimentation
- To design generic sentence parser model using SVM
- To develop the parser prototype and
- To test the developed prototype

Scope and delimitation of the study

The scope of this research mainly focused on investigating Amharic sentence parser using supervised Machin learning approach (SVM) The study investigates the optimal feature set for Amharic sentence parser based on SVM Parsing imperative and interrogative sentences and integrating the parser with different NLP and IR application is out of this scope

Literature Review

In order to achieve the objective of the research books, journal articles and conference papers on automatic sentence parse was reviewed. As indicated before few works has been done for Amharic sentence parser for instance by Atelach Alemu Argaw [3]. The study attempted to develop a simple automatic parser for Amharic sentences Inside Outside algorithm with bottom up chart parsing strategy. In their research, a small sample corpus 100 sentences was selected Among **100** sentences, **80** sentences were used to train the system while **20** sentences were used to test the system accuracy of their result

by Daniel Gochel [1]The study is just an extension of Atalech work [3] specifically focused on complex Amharic sentences In their research, **350** different types of sample sentences were used. Among **350** sentences, **280** sentences were used to train the system while **70** sentences were used to test the system. The overall accuracy of their result were **89.6 on** train and 81.6 on the test set by Ababa Ibrahim [2]

The main objective of the research was to extract different types of Amharic phrases and transform the chunkier to full parser.

The researcher used (HMM) to develop the chunkier and bottom-up approach to transform the chunkier to the parser. however, As indicated by many scholars transform the chunkier to full parser is time consuming and reduce system performance. by Abdurrahman Dawood [4**]** The main objective of their research work was to design and implement a top-down chart parser for Amharic sentences The researcher selects rule-based approaches. Since. Requirement of huge linguistic knowledge, very large number of rules to cover all the features of a language, highly language dependent it is impossible to adapt for other language easily, it consume too much time to develop the rule and difficulty to extend or scale-up the system**.**

Text parsing for other language

Text parser for Oromo Language Was aimed to parse simple sentences using supervised learning The study has been conducted using the chart algorithm with the grammar formalism (HPSG) compiled into left to right table.[Al-Taani, Ahmad T., Mohammed M. Msallam, and Sana A. Wedian 2012] [5] It is designed for parsing simple Arabic sentences,

In the implementation of the parser, top-down parsing technique with recursive transition network grammar has been used[Thant, Win, Tin Myat Htwe, and N. L. Thein 2011] [6]

The paper discus about parse simple and complex Myanmar sentences of 5 to 50 words length The study used CFG formalism to represent production rules whereas top-down parsing is used to build the parse tree [8,9,10].

The researchers collected nearly 3000 training sentences and 530 testing sentences from Myanmar religious books, short stories and newspapers. The researchers annotated the corpus for part of speech, chunk, and function tags relationship between the words in the sentences.

Application of the result

As indicated by many scholars some of NLP application are beneficiary from sentences parsing, Phrase recognition, Conceptual parsing, Machine translation, Spell checker, Text summarization [7], etc.…

Data collection

We have used Walta Information Center (WIC) corpora, in addition to we collect data from Amharic grammar book, news article and magazines. The collected dataset is annotated manually to make suitable for training and testing the model.

II. PROPOSED ALGORITHM

The proposed Algorithm to Implementation the system was used the tool is Support vector machine (SVM) adopted to implement sentence parser

For our experiments, we were used the adopted software package LibSVM. It is open source freely available SVM implementation tool[11,12,13,14].

Weka 3.8 used for the development of the prototype and implementation of the parsing algorithm.This selected for the various features it provides.

One of the reasons is that Weka contains tools for data preprocessing classification and it is also suitable for developing new machine learning.

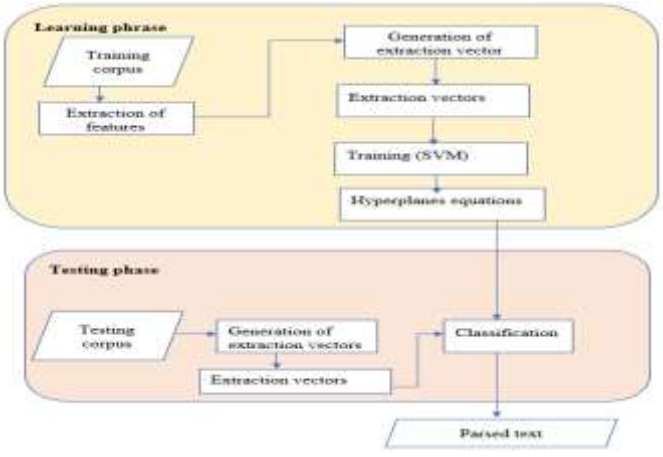

Figure 1. Architecture of the System

III. EXPERIMENT AND RESULT

In this study, , a total of 370 sentence were used and manually hand parsed and morphologically analyzed all the sentence, with close guidance linguistics expert From the total 90 % used for model creation and the rest 10% used for testing the model. .

Table 1:courpus statistics

Number of sentences in the training set	Number of sentences in the testing set	Total number of sentences
333sentences	37 sentences	370 sentences

The evaluation of our analyzer the following result using the Weka tool

Table 2:experimental result

Precision	**Recall**	**F-score**
0.989	**0.989**	**0.989**

The performance of the developed model evaluated using the accuracy measure we obtained an overall accuracy of 98.913

IV.CONCLUSION

- This work tried to contribute one important model which plays a role for overcoming some challenges in NLP regarding for Amharic.
- Having a good Amharic sentences parser could be used in information extraction of any domain specific tasks, question answering, spell checker.
- To do this research,370 sentence were taken from WIC corpus, Amharic grammar book and magazines and parse tag manually.
- Even though different researcher investigates toward Amharic sentence parser using different approach on a different tool, corpus size and feature set. Due to these reasons, it is difficult to compare this work with the previous one.
- But from the finding we got, a model achieved a promising result for Amharic sentence parser.

Recommendation

The following points are some of our observation that should be taken into consideration for future works of ASP(Amharic sentence parser).

- Replicate this work using large data and incorporating all types of sentences , and also it can be an extend work in other local languages
- Integrating our parser results with other NLP tasks such as machine translation, text categorization, IR etc
- Experimenting using other machine leaning algorithm such as Conditional random field (CRFs), deep neural network, etc. may give better result than this

REFERENCE

1. Daniel Gochel. (2003). An Integrated Approach to Automatic Complex Sentence Parsing for Amharic Text, Unpublished, Master Thesis at School of Information Studies for Africa, Addis Ababa university.
2. Abeba Ibrahim. (2013). A Hybrid Approach to Amharic Base Phrase Chunking and Parsing, Unpublished, Master's thesis, Addis Ababa University,
3. Atelach Alemu. (2002). "Automatic Sentence Parsing for Amharic Text: An Experiment using Probabilistic Context Free Grammars." Unpublished Master 's Thesis, School of Graduate Studies, Addis Ababa University ,2002.
4. Abdurehman Dawud Moh. (2015). A Top-Down Chart Parser for Amharic Sentences, Unpublished Master 's Thesis, School of Graduate Studies, Addis Ababa University.
5. Al-Taani, Ahmad T., Mohammed M. Msallam, and Sana A. Wedian. (2012). A top-down chart parser for analyzing Arabic sentences. Int. Arab J. Inf. Technol, *Vol. 9, No. 2*, pp 109-115.
6. Thant, Win, Tin Myat Htwe, and N. L. Thein. (2011). Context free grammar-based top-down parsing of myanmar sentences." International conference on computer science and information technology, Pattaya.
7. Lu, Wenpeng, Heyan Huang, and Chaoyong Zhu. (2012). Feature words selection for knowledge-based word sense disambiguation with syntactic parsing." Electrical Review, pp 82-87.
8. C. T. Rekha Raj and P. C. Reghu Raj. (2015). Text chunker for Malayalam using Memory-Based Learning," *IEEE*, pp. 595–599.
9. A. Molina and F. Pla.(2002). Shallow parsing using specialized HMMs," *J Mach. Learn. Res.*, vol. 2, pp. 595–613.
10. U. Jain and J. Kaur. (2015). A Review on Text Chunker for Punjabi Language,vol. 4, no. 7, pp. 116–119.
11. K. Sarkar and V. Gayen. (2014). Bengali Noun Phrase Chunking Based on Conditional Random Fields, *IEEE*, pp. 148–153.
12. A. Ibrahim and Y. Assabie. (2014). Hierarchical Amharic Base Phrase Chunking Using HMM with Error Pruning," Springer Int. Publ. Switz., vol. 8387, pp. 126–135.

13. K. Sarkar and V. Gayen. (2014). "Bengali Noun Phrase Chunking Based on Conditional Random Fields," IEEE, pp. 148–153.

14. W. Ali, M. K. Malik, S. Hussain, S. Shahid, and A. Ali. (2010). Urdu noun phrase chunking," IEEE, pp. 494–497.