A rule-based Afan Oromo Grammar Checker

Debela Tesfaye
Information Technology,Jimma Institute of Technology
Jimma, Ethiopia

***Abstract*—Natural language processing (NLP) is a subfield of computer science, with strong connections to artificial intelligence. One area of NLP is concerned with creating proofing systems, such as grammar checker. Grammar checker determines the syntactical correctness of a sentence which is mostly used in word processors and compilers.**
For languages, such as Afan Oromo, advanced tools have been lacking and are still in the early stages. In this paper a rule based grammar checker is presented. The rule base is entirely developed and dependent on the morphology of the language . The checker is evaluated and shown a promising result.

Keywords- afan oromo grammar checker; rule based grammar checker.

I. INTRODUCTION

Natural language processing (NLP) is a subfield of computer science, with strong connections to artificial intelligence. Natural language processing (NLP) is normally used to describe the function of computer system which analyze or synthesize spoken or written language [3]. One area of NLP is concerned with creating proofing systems, such as spell checkers and grammar checkers. A grammar checker looks for grammatical errors and, in many cases, suggests possible corrections. Grammar checking is one of the most widely used tools within language engineering. Spelling, grammar and style checking for English has been an integrated part of common word processors for some years now.

The great challenge of intelligent automatic text processing is to use unrestricted natural language to exchange information with a creature of a totally different nature: the computer. People now want assistance not only in mechanical, but also in intellectual efforts. They would like the machine to read an unprepared text, to test it for correctness, to execute the instructions contained in the text, or even to comprehend it well enough to produce a reasonable response based on its meaning. Human beings want to keep for themselves only the final decisions.

Millions and millions of persons dealing with texts throughout the world do not have enough knowledge and education, or just time and a wish, to meet the modern standards of document processing [4]. For example, a secretary in an office cannot take into consideration each time the hundreds of various rules necessary to write down a good business letter to another company, especially when he or she is not writing in his or her native language. It is just cheaper to teach the machine once to do this work, rather than repeatedly teach every new generation of computer users to do it by themselves.

Grammar checker determines the syntactical correctness of a sentence. Grammar checking is mostly used in word processors and compilers. Grammar checking for application like compiler is easier to implement because the vocabulary is finite for programming languages but for a natural language it is challenging because of infinite vocabulary.

A lot of work has gone into developing sophisticated systems that have gone into widespread use, such as automatic translators and spell checkers. However, most such programs are strictly commercial, and therefore there exists no documentation of the algorithms and rules used. For languages, such as Afan Oromo, advanced tools have been lacking and are still in the early stages. However, one of the most widely used grammar checkers for English, Microsoft Office Suite grammar checker, is also not above controversy [1]. It demonstrates that work on grammar checker in real time is not very easy task; so starting the implementation for language like Afan Oromo is a major feat. In this research, a rule based grammar checker for Afan Oromo is presented.

II. BACKGROUND

Detection and correction of grammatical errors by taking into account adjacent words in the sentence or even the whole sentence are much more difficult tasks for computational linguists and software developers than just checking orthography. Grammar errors are those violating, for example, the syntactic laws or the laws related to the structure of a sentence. In Afan Oromo, one of these laws is the agreement between a noun and an adjective in gender and grammatical number[7]. For example, in *Jarri dhufaa jira*, subject and verb disagree in number. Jarri(*they) which is the subject of the sentence is* plural, and the verb of the sentence *jira* is the indicator for third person singular masculine.

Three methods are widely used for grammar checking in a language; syntax-based checking, statistics-based checking and rule-based checking.

A. Syntax based checking

In this approach, a text is completely parsed, i.e. the sentences are analyzed and each sentence is assigned a tree structure. The text is considered incorrect if the parsing does not succeed.

B. Statistics based checking

In the statistical approach the system is trained on a corpus to learn what is 'correct'. In this approach, a POS-annotated corpus is used to build a list of POS tag sequences. Some sequences will be very common (for example determiner, adjective, noun as in the old man), others will probably not occur at all (for example determiner, determiner, adjective). Sequences which occur often in the corpus can be considered correct, whereas uncommon sequences might be considered as errors. This method has a few disadvantages. One of these is that it can be difficult to understand the error given by the system as there is not a specific error message. This also makes it more difficult to realize when a false positive is given [1].

C. Rule based checking

Using the rule-based approach to grammar checking involves manually constructing error detection rules for the language. These rules are then used to find errors in text that has already been analyzed, i.e. Tagged with a part-of-speech tagger. These rules often contain suggestions on how to correct the error found in the text.

A lot of work has gone into developing grammar checkers for different languages. The most progress, by far, has been made for English. The earliest grammar checkers for English were developed in the 1970s and have gradually been improving over the last decades. Although there is still room for improvement their use is quite widespread as an English grammar checker is built into the most used word processor today, Microsoft Word.

EasyEnglish is a grammar checker developed at IBM especially for non-native speakers. It is based on the English Slot Grammar. It finds errors by "exploring the parse tree expressed as a network" [5]. The errors seem to be formalized as patterns that match the parse tree. Unfortunately [5] does not explain what exactly happens if a sentence cannot be parsed and thus no complete tree can be built.

III. Overview of Afan Oromo

Afan Oromo is one of the major languages that is widely spoken and used in Ethiopia. Currently it is an official language of Oromia state (which is the largest region in Ethiopia). It is used by Oromo people, who are the largest ethnic group in Ethiopia, which amounts to 34.5% of the total population [6]. The language has become the official language in Oromia regional offices *a*nd is also instructional language starting from elementary to university level.

Like a number of other African and Ethiopian languages, Afan Oromo has a very rich morphology [7]. It has the basic features of agglutinative languages where all bound forms (morphemes) are affixes. In agglutinative languages like Afan Oromo most of the grammatical information is conveyed through affixes and other structures.

Therefore the grammatical information of the language is described in relation to its morphology. This makes it very hard to create grammar checker and develop general understanding of the language.

As any other language the grammar of Afan Oromo exhibits gender,number,cases,tenses etc. But the grammatical presentation of the above cases are different from other languages and exhibits its own structure. Unlike English, Afan Oromo gender,number,tense and other cases are described using affix. Therefore the grammatical rule is mostly dependent on the affixation rule of the language.

Example:

1. Inni kalleessa dhufe.(he came yesterday)
2. Isheen kalleessa dhufte.(she came yesterday)
3. Inni dhufaa jira.(he is coming)
4. Isheen dhufaa jirti.(she is coming)

In the above four sentences the gender and tense of the sentences are described through suffix which is attached to the verbs dhuf- and jir-.

IV. Afan Oromo Grammar Checker

As described above, Afan Oromo exhibits its own grammatical structure. Therefore it is not possible to apply and use the grammatical rule of another language for Afan Oromo grammar checker. In this paper different 123 rules were constructed and used in order to identify grammatical error of the language. With the use of these carefully constructed error detection rules, the system can detect and suggest corrections for a number of grammatical errors in Afan Oromo texts. Afan Oromo Grammar Checker has the following components:

A. Tokenizer Module

The tokenizer module splits the input text (paragraphs) from an input file into sentences. The tokenized sentences are further tokenized into words.

B. Parts of Speech (POS) Tagger Module

Part of speech taggers are very important for our approach. In POS tagging of a text, each word in the text is assigned a part of speech. We have used tagger based on Hidden Markov Model which uses a manually tagged corpus for training"unpublished"[8].

C. Stammer Module

A stemming algorithm is a procedure that reduces all words with the same stem to a common form by stripping of its derivational and inflectional suffixes. The stammer module of this checker provides the root and affix for the tagged words. Like a number of other African and Ethiopian languages, Afan Oromo has a very rich morphology [7]. It has the basic features of agglutinative languages where all bound forms (morphemes) are affixes. In agglutinative languages like Afan Oromo most of the grammatical information is conveyed through affixes (prefixes, infixes and suffixes) attached to the roots or stems. Both Afan Oromo nouns and adjectives are highly inflected for number and gender. In contrast to the English plural marker s (-es), there are more than 12 major and very common plural markers in Afan Oromo nouns (example: -oota, -ooli, -wwan, -lee, -an, -een, -oo, etc.) [2]. Afaan Oromo verbs are also highly inflected for gender, person, number and tenses.

The Afan Ormo stammer is based on a series of steps that each removes a certain type of affix byway of substitution rules. These rules only apply when certain conditions hold, e.g. the resulting stem must have a certain minimal length. Most rules have a condition based on the so-called measure. The measure is the number of vowel-consonant sequences (where consecutive vowels or consonants are counted as one) which are present in the resulting stem. This condition must prevent that letters which look like a suffix but are just part of the stem will be removed.

The affix-rules have the following general form:

Affix → substitution measure-condition <additional conditions>

Where:

Affix is a valid Afan Oromo prefix or suffix

In Afan Oromo repetition (plural) is formed by duplicating the first syllabus and it is also considered as prefix.

Substitution is a string which is substituted with a given affix to produce valid stem.

Measure-condition is the number of vowel-consonant sequences (where consecutive vowels or consonants are counted as one) which are present in the resulting stem.

Additional conditions- additional conditions are also designed to cover some specific phenomena. Examples of these conditions are, Endswith Vowel/Consonant.

D. Grammatical Relation Finder

Assigns grammatical relations between subject and verb,subject and adjective,main verb and subordinate verb in terms of number,gender and tense. In this paper 123 different rules are constructed and presented. This rules takes the affixes that are identified and separated from a root word using the stammer module in order to identify the agreement between subject and verb, subject and adjective, main verb and subordinate verb in number, tense, gender and other causes. As explained in section III the grammatical information of the above cases are presented using affix-rules in the language.

E. Suggestions creating module

It provides the correct sentence alternatives. This module provides the alternatives in two way directions. For example in the case of subject verb disagreement it provides one or more alternatives by adjusting the subject and leaving the verbs as they are. Or provide one or more alternatives by adjusting the verbs and leaving the subject as it is. This is based on the assumption that errors can be committed both on the subject and the verb. There for the users must be provided with correct alternative by correcting either of the two one at a time.

The Grammar Checker General Algorithm

The algorithm has five steps as presented in the following section:

step 1.

Tokenize the sentence using '.' or '!' or '?'

step 2.

Identify the part of the speech of each token in the sentence

Step 3:

Identify the root and affixes of each token(word)

Step 4:

4: a:

Forward the affixes to the rule based that checks Subject-verb,subject and adjective,main verb and subordinate verb agreement in terms of number,tense and gender agreement.

4:b:

Check for punctuation errors.

Step 5:

provide grammatically correct sentence suggestions.

The algorithm is illustrated using the following example.

Inni saree ajjeste. Because of grammatical error the sentence is meaningless.

Step 1:

The tokenizer module identify the tokens as :

Inni, saree, and ajjeste.

Step 2:

The part-of-speech tagger tagged the tokens as:

Inni as the subject of the sentence, Saree as the object and ajjeste as the verb.

step 3:

The stammer module identified the root and affix of the word ajjeste as:

ajjes- is the root and -te is the suffix.

Inni has no any suffix.

Saree is the object of the sentence. In Afan Oromo object of the sentence has no any grammatical relation with the subject and verb of the sentence.

step 4:

The subject of the sentence is 1st person singular masculine and the suffix is feminine marker.

The rule

If the subject of the sentence is 3st person singular masculine (Inni) then the verb must end with the masculine marker suffixes –a, and -e.

So the rule-based marks for the subject-verb verb disagreement.

Step 5:

T*he correct suggestions are:*

a. Inni saree ajjese. By changing the suffix from feminine marker to masculine marker e.

b. Isheen saree ajjeste. By changing the subject of the sentence from masculine Inni to feminine Isheen.

Sample rules of the grammar checker

Definitions:

sg.1.p=first person singular

2.p = second person

3.p.m= third person masculine

3.p.f=third person feminine

2..p.pl=second person plural

3.p.pl=third person plural

RV=root verb

There are a total of 123 rules. The rules ranging from 81 to 86 covers past perfect tense as presented in the following.

rule 81. sg. l.p +RV+een+ture

rule 82. 2.p.sg + RV+tee+turte

rule 83. 3.p.m.+RV+ee+ture

RULE 84. 3.p.f.+RV+tee,dee+turte

RULE 85. 2.p.pl +nee++turre

RULE 86. 3.p.pl+ani++turan

Explanation of the rules:

If the subject of the sentence is first person singular, the verb(s) must end with the suffix -een and the sentence must end with ture.

If the subject of the sentence is third person singular feminine, the verb(s) must end with the suffix -tee,dee and the sentence must end with turte.

Example: Callise bira dabruu yaalee ture.(he was trying to pass by silent). In the above example, the subject of the sentence is third person singular masculine, so the verb must end with the suffix -ee and the sentence must end with ture.

V. EVALUATION OF THE CHECKER

Grammar and style checking software have involved measuring the program's error detection capacity in terms of precision (i.e. error detection correctness) and recall (i.e. error coverage) [9]; [10];[11].

In order to check the performance of the system a student graduation thesis text is used. A thesis work of Afan Oromo 1st degree graduate of 2011 were used in order to measure the performance of the checker. Originally, it was thought best to get some sort of text from non-native Afan Oromo speakers as it was assumed that students learning the language might not have the same 'feel for the language' as native speakers and therefore have more grammatical errors in their texts.

Finally, the above data were run through the grammar checker for errors. In order to calculate the performance rate the number of errors in the texts, number of errors found by the Grammar checker and the number of false positives generated by the grammar checker were counted. These numbers were then used to calculate the precision and recall of the system.

The table below shows the precision and recall rates for all the errors in the texts as well as the corresponding rates for each type of error. The rates are found as follows:

$$\text{Precision} = \frac{\text{number of correctly flagged error}}{\text{total number of flagged error}}$$

$$\text{Recall} = \frac{\text{number of correctly flagged error}}{\text{total number of error that occur in the text}}$$

TABLE I. PERFORMANCE RESULT

	Measuring criterias					
	Incorrect flags	*Correct flags*	*Total Number of flags*	Total number of errors in the test set	precision	recall
In NO/ %	100	800	900	1000	88.89%	80.00 %

There are several reasons why a false alarm might occur:

- The stammer identified the root and affix of some words incorrectly.
- A word has been assigned an incorrect part-of-speech tag.
- The rule is not complete and didn't covered every case.

VI. CONCLUSION

In this thesis, Afan Oromo grammar checker has been developed and tested on real-world errors. As can be seen from table 1.1 the performance of the checker is promising. Most of the false flags are related to compound, complex and compound complex sentences as most of the rules are constructed for simple sentences. More rules that handles the above listed types of sentences can be added to the existing rules in order to improve the performance of the grammar checker. There are also sentences that exhibits grammatical errors but not flagged by the checker.

Other than the incompleteness of the rules the part-of-speech tagger component of the checker has also provided incorrectly tagged words for the checker. The incorrectly tagged words lead the checker to not flag errors and generate false flags. The stammer component that separates the root from affix of a word is important since the grammar of the language is mostly described through affixes. Generally, since the grammar rules in the grammar checker is largely dependent on morphology of the language this approach is believed too be used for other languages that are rich in morphology.

REFERENCES

[1] Naber, D. (2003). A Rule-Based Style and Grammar Checker. Diplomarbeit. Technische Fakultät Bielefeld.

[2] Debela Tesfaye & Ermias Abebe(2010). Designing a rule based stemmar for afan Oromo text. International Journal of Computational Linguistics (IJCL), Volume (1): Issue (2)

[3] Peter Jackson and Isabelle Moulinier, natural language processing for online applications: text retrieval, extraction and categorization,1984

[4] Igor Balshagov and Alexander Gelbukh, computational linguistics models resources and applications,2004

[5] Arendse Bernth: EasyEnglish: Grammar Checking for Non-Native Speakers, Proceedings of the Third International Workshop on Controlled Language Applications (CLAW00), Association for Computational Linguistics, April 29-30, Seattle, Washington, pp. 33-42, 2000

[6] Census Report. "Ethiopia's population now 76 million", http://ethiopolitics.com/news, (2008)

[7] Gumii Qormaata Afaan Oromoo. "Caasluga Afaan Oromoo Jildi I", Komishinii Aadaaf Turizmii Oromiyaa, Finfinnee, Ethiopia, pp. 105-220 (1995)

[8] C. G. Mewis. A Grammatical sketch of Written Oromo, Germany: Koln,pp. 25-99 (2001)

[9] Getachew Mamo. Statistical model Part-of-speech Tager for afan Oromo,Addis Abeba University,2009.

[10] Kukich, K. Techniques for automatically correcting words in text. ACM Computing surveys, Vol. 24, No. 4, pp. 377–439 ,1992.

[11] Birn, J. Detecting grammar errors with Lingsoft's Swedish grammar checker. In Proc. 12th Nordic Conference in Computational Linguistics, Nodalida-99. Trondheim, pp. 28–40, 2000.

[12] Richardson, S & Braden-Harder, L. The Experience of Developing a Large-Scale Natural Language Processing System: Critique. In Jensen, K. Heidorn, G. E. Richardson, S. D. (eds.), Natural Language Processing: The PLNLP Approach, pp. 77- 89, 1993.