OroRoots: Rule-Based Root Generation System for Afaan Oromo

Wegari, Getachew Mamo

Abstract— A root generator system for concatinative languages could be a starting point for natural language processing related works particularly for morphological analysis and information retrieval systems.Rule-based method was investigated to develop a root generation system for Afaan Oromo. Afaan Oromo is one of widely used language in Ethiopia as well as in Northern Kenya. The proposed system was evaluated with testing wordlist and has been experimentally shown that it improves the performance of state-of-the-art methods for the language.

Index Terms— Afaan Oromo, rule-based, root-generation, affix-sequence, morphological properties

1 Introduction

AFAAN Oromo is an indigenous Ethiopian language which belongs to Afro-Asiatic language family. It is spoken and used by 34.4% of the total population of Ethiopia [2]. It is the official language of Oromia regional state, which is the biggest regional state in Ethiopia, and it has also spoken and used in Kenya and Somalia.

Words in Afaan Oromo are generated from roots in concatenation form by adding suffixes except for some plural forms of adjectives and intensive verbs that are formed by duplication of the first syllable.Thus, morphologically, Afaan Oromo is concatenativesuffix-based language.

Having a root generator for concatinative language could be a starting point for natural language processing related works particularly for morphological analysis and information retrieval systems. A great deal of attention should be paid to words which are the building blocks of structural linguistic levels such as syntax, semantic and discourse. Hence, understanding informative structures of words enables to exploit their informative structures in different application domains.

An analysis of a rule-based stemmer of Afaan Oromo has been tried based on the design principles of the Porter stemmer [7], [6]. However, it is not as such effective since languages are morphologically different from one another. A finite state approach morphological analyzer has been modeled which also indicate roots of words for three widely used languages in Ethiopia including Afaan Oromo [4] but it is not evaluated for Afaan Oromo due to the great variation in the use of double consonants and vowels by Oromo writers as reported on the article. There is no complete work which provides a detailed computational analysis of roots of Afaan Oromo and this paper tries to do that.

The organization of the rest of the paper is as follows. Next section gives brief morphological properties of Afaan Oromo. Section 3 presents root identification rules. Results of the performed tests are presented together with their analysis in Section 4. At last, discussion and conclusion are described in Section 5.

2 Basic Morphological Properties of Afaan Oromo

Morphologically, Afaan Oromo has concatenative form of suffixes and its roots of lexical words (nouns, verbs, adjectives and adverbs) end either in single or double consonants[5]. A root is "the base form of a word which cannot be further analysed without total loss of identity" [1]. Table 1 shows sample lexical words that their roots end with single and double consonants.

TABLE 1: SAMPLE LEXICAL WORDS THAT THEIR ROOTS END WITH SINGLE OR DOUBLE CONSONANTS.

Sample words that their roots end with single consonants	
deemaniiru	**deem**-aniiru
baraareera	**baraar**-eera
jireenya	**jir**-eenya
Sample words taht their roots end with double consonants	
arganiiru	**arg**-aniiru
arjoomuu	**arj**-oomuu
eebbisuu	**eebb**-isuu

Most roots of lexical words of Afaan Oromo add suffixes without changing their forms.

2.1 Root Generation Rules

As indicated in Table 1, roots of the lexical words end either with single or double consonants so that consonants that come after the first character of a word can be candidates that indicate the boundary of the root of the word.For example, the candidate root boundary of the word **biteera** can be:

b i t e e r a

There is no lexical word that contains single character in Afaan Oromo.

International Journal of Scientific & Engineering Research, Volume 8, Issue 3, March-2017
ISSN 2229-5518

In another hand, most suffixes of Afaan Oromo are composed of vowels. Consonants that are used in suffixes are ch, dh, f, l, m, n, r, s, t, w and yandmost of them are used with certain suffixes. These are the rationale that inspired us to develop root generation system using rule-based approach.

Based on the general morphological properties of the language mentioned above, Afaan Oromo root generation rules were developed based on the following core steps.

Step One: The lexical words of the language were categorized into five groups based on their forms that they characterize from their first character to the next consonant to minimize the complexity of rules.

- The first group iswords that form VC where V and C stand for vowel and consonants respectively. For example, arge, ogeessa.
- The second group is words that form VVC and CVC. For example, aaruu, rafuu.
- The third group is words that form CCVC, CVCC and CVVC. For example, dhala, hidhuu, caaluuf.There are five consonant clusters (ch, ny, dh, sh, ph) that are considered as one character in Afaan Oromo. CC is stands for these consonant clusters.
- The fourth group is words that form CCVVC, CCVCC and CVVCC. For example, nyaara, dhadhaa, daadhii.
- The fifth group is words that form CCVVCC. For example, dhaadhessuu, shaashii.

Step Two: Based on step one, a word was segmented into two: firstPart and secondPart. The firstPart is definitely parts of a root so that it is not processed further. The secondPart is that could be further processed to identify a root boundary. For example, the word ogeessa can be segmented as og(firstPart)andeessa(seconPart).And the general algorithm is as follow.

Algorithm: Generation of Roots

1: A word is taken as input
2: The word is segmented into two based on step one: firstPart and secondPart
3: ThesecondPart's characters are iterated from right to left
4: for i in range(length (secondPart)):
5: if secondPart[i]==w and [wwan] in secondPart:
6: pass to next
7: elif w or d[i] == m and [am oroom orumm ormaat orum] in secondPart:
8: pass to next
9: elif secondPart[i] ==l and [olii or olee] in second Part:
10: pass to next
11: elif secondPart[i] == r and [teer or aniir or neer oreer] in secondPart:
12: pass to next
13: elif secondPart[i] == h and [d or c] is found at immediate left side of h:
14: pass to next
15: elif secondPart[i] == t and [a or e or i or o or u or t or r] is found at immediate left side of t:
16: pass to next
17: elif secondPart[i] == f and [a or e or i or o or u] is found at immediate left side of f:
18: pass to next
19: elif secondPart[i] == s and [a or e or i or o or u or s or n or m] is at immediate left side of s:
20: pass to next
21: elif secondPart[i] == n and [a or e or i or o or u or t or n or m or r or s] is at immediate left side of n:
22: pass to next
23: elif secondPart[i] == y and [n] is at immediate left side of y:
24: pass to next
25: elif secondPart[i] in [a or e or i or o or u]:
26: pass to next
27: else:
28: consider a character as the root boundary of a word.
29: break

---

3 EXPERIMENTS

This section is to measure the effectiveness of the methodology described in section 3. The methodology was compared with some state -of-art to make the experiments stronger.

3.1 TESTING

A testing wordlist that was usedfor the experiments was taken from [8]. The wordlist contains 2000 segmented unique lexical words. The evaluation was taken place in terms of precision that means the number of correctly indicated root boundaries divided by the total number of root boundaries indicated by the model. The output of the model was classified as:

- Correctly indicated root boundaries (CRIB)
- Over-segmented (OS)
- Under-segmented (US)

CIRB is the number of root boundaries that were correctly indicated by the models. Both OS and US are root boundaries that were incorrectly indicated by the model.

International Journal of Scientific & Engineering Research, Volume 8, Issue 3, March-2017
ISSN 2229-5518

3.2 RESULTS

TABLE 2: RESULTS OF AUTOMATIC ROOT GENERATION OF USING A BASELINE ALGORITHM (THE SECOND COLUMN) AND THE ALGORITHM PROPOSED IN THIS PAPER (THE LAST COLUMN).

Output Categories	Probabilistic and Grouping Methods	OroRoots
CIRB	70.05%	98.4%
OS	3.55%	0%
US	1.55%	1.6%

The results of the models are summarized in Table 2. The proposed system performs in better competitive than the baseline algorithm (Probabilistic and Grouping Methods).

4 DISCUSSION

Since majority of the proposed rules are based on suffixes, the most errors of the system is due to strings that are found both as parts of roots and suffixes in the language. For instance, for the word rukuta, the system indicated ruk as its root instead of rukut since ut is commonly used as parts of suffixes of the language so that all errors are under segmented (US). In spite of these limitations, experiments on Afaan Oromo have showed that the presented method performed well compared to the-state-of-the-art.

5 CONCLUSSION

In our investigation, we have been shown that rule-based method can be used to develop root generation system for Afaan Oromo. The system mainly used as starting point to develop a complete morphological analysis and information retrieval for the target language. The experimental results show that the methodology proposed is effective in identifying root boundaries.

REFERENCES

[1] Crystal, D. (2008). A Dictionary of Linguistics and Phonetics. 6th Edition. Blackwell.
[2] CSA (Central Statistical Authority) (2007). Population and Housing Census of Ethiopia.
[3] Central Statistical Office. Addis Ababa, Ethiopia, pp. 73.
[4] Gasser, M.(2011). HornMorpho: A System for Morphological Processing of Amharic, Oromo, and Tigrinya. Conference on Human Language Technology for Development, Alexandria, Egypt.
[5] GQAO (Gumii Qormaata Afaan Oromo), (1996). Caasluga Afaan Oromo. Komoshinii Aadaafi Tuurizimii Oromiyaa, Itoopiyaa.
[6] Porter, M.F.(1980). An algorithm for suffix stripping.Program, Vol. 14, no.3, pp. 130−137.
[7] Tesfaye, D. and Abebe, E.(2010). Designing a Rule-based Stemmer for Afaan Oromo Text. In: International journal of Computational Linguistics, Vol. 1, no. 2. pp. 1-11.
[8] Wegari, Getachew M., Melucci, M. and Teferra, S. (2016). Probabilistic and grouping methods for morphological root identification for Afaan Oromo. 6th International Conference - Cloud System and Big Data Engineering (Confluence), pp. 12 – 15, IEEE.