DOI: http://dx.doi.org/10.4314/star.v2i4.16
ISSN: 2226-7522(Print) and 2305-3327 (Online)
Science, Technology and Arts Research Journal
Oct-Dec 2013, 2(4): 94-97
www.starjournal.org

Original Research

Analysis of Rule Based Approach for Afan Oromo Automatic Morphological Synthesizer

Abebe Abeshu

School of Computer Engineering, IT Convergence Engineering, Sungkyunkwan University, South Korea

Abstract

The aim of this study was to design a morphological synthesizer for Afan Oromo particularly for verbs and nouns. In this research rule based approach for Afan Oromo automatic morphological synthesizer was applied. The performance of the synthesizer was evaluated using accuracy as statistical measurement. An average performance of 96.28% for verbs and 97.46% for nouns was obtained. Potential directions are identified to further improve the synthesizer in the future. The result obtained encourages the undertaking of further research in the area, especially with the aim of developing a full-fledged Afan Oromo morphological synthesizer.

Article Information

Article History:
Received : 20-10-2013
Revised : 27-12-2013
Accepted : 29-12-2013

Keywords:
Afan Oromo
morphological synthesizer

***Corresponding Author:**
Abebe Abeshu
E-mail:dinkolas_abeshu@yahoo.com

INTRODUCTION

Language is one of the fundamental aspects of human behavior and it constitutes a crucial component of our lives. In its written form it serves as a means of recording information and knowledge on a long term-basis and transmitting what it records from one generation to the next. In its spoken form it serves as a means of coordinating our day-to-day life with others (Allen, 1996).

Currently, many applications in NLP area require linguistic knowledge to be readily available to successfully implement many NLP systems at various levels of processing. There are, for instance, systems developed for processing natural language at phoneme, word, sentence, and pragmatic levels. These systems are developed in such a way that the output of a lower system can serve as an input for the next higher level. For instance, the output of a morphological synthesizer that works at word level could serve as an input for syntactic and semantic parsers that work at sentence level (Uibo, 2001). In addition, the words generated by morphological synthesizer can be used as a suggestion list for spell checker. Such systems are used as a subcomponent of NLP in applications like machine translation, dictionary (lexicon) development, and spelling and grammar checking, and etc (Salton, 1983). However, limited effort has been made to Afan Oromo language. Afan Oromo belongs to the Cushitic family of languages and one of the most widely spoken natural languages in Ethiopia (Diriba Megersa, 2002; Kula Kekeba, 2008). However, it is among the least researched languages, as there are no morphological processing systems such as morphological synthesizer essential for high level computational models.

Approaches of Morphological synthesis

Morphological synthesis or generation is a process of returning one or more surface forms from a sequence of morpheme glosses. A number of approaches have been proposed for automatic morphological synthesizer. The most common and widely used approaches are: rule-based, machine learning and hybrid approaches (Diriba Megersa, 2002; Mesifin Getachew, 2001).

The rule-based approach is one of the earliest approaches in developing morphological synthesis systems. The rule based approach uses rules created by linguists or by a computer program to generate appropriate word forms from stems. The rules may contain a large number of morphological, lexical and/or syntactical information. They are based on linguistic knowledge about the structure of the specific language.

Machine learning approaches do not strictly follow explicit theory of linguistics. The approaches are completely based on training and testing corpora, which constitute the input data. Approaches in this category use some algorithms to learn, say about the word formation process of a language from a given corpus and perform the synthesis based on this knowledge. Corpus based approaches are further divided into *supervised* and *unsupervised* based on the type of training corpora they use. Unsupervised approaches use heuristics or probability information generated from the test corpora to generate the morphological synthesis system. In this approach, no sample outputs are given (Karttunen, 1994; Tom, 1997; Andrew Roberts, 2009) argues that this approach reduces the cost of browsing annotated requires annotated text corpora. In this case, a teach input is provided which tells the system the outputs required for a given input.

The last category of morphological synthesis is the hybrid approach which tries to combine two or more of the above stated approaches

Morphological Synthesis Attempts Made on Afan Oromo

No research has been conducted so far in the area of automatic morphological synthesis for Afan Oromo. The absence of morphological synthesis systems limits the effort of making computers work comfortable with Afan Oromo. Hence, the aim of this study is to conduct research on morphological synthesis for Afaan Oromo using rule based approach and analyze the effectiveness of the approach.

Design of Afan Oromo Morphological Synthesizer

Afan Oromo synthesizer has 9 components as shown in figure 1.

Lexical level: is the stem to be given to the system as input corpora. Supervised approach, on the other hand, **Stem Presence Checker**: checks whether stem has been stored in the lexicon or not.

Knowledge Base Module: contains lexical information of stems and their tags **POS-guesser**: predicts POS of stem that is not listed in the lexicon

Noun Classifier: identifies the inflectional type or declension of the stems for nouns.

Verb Classifier: classifies the new stem that is not available in the lexicon into one of the preset classes of verbal stems.

Lexicon Updater: update if new word which does not exist in the lexicon is encountered.

Signature Builder: This part lists the set of available suffixes valid for a given stem class.

Boundary Change Handler: This module accesses rules from the knowledge base to replace some pattern by the appropriate stored form.

Synthesizer: generates all possible surface forms from a given stem.

Surface level: The words generated as the output of the synthesizer. The words should be meaningful and acceptable by the speakers of the language.

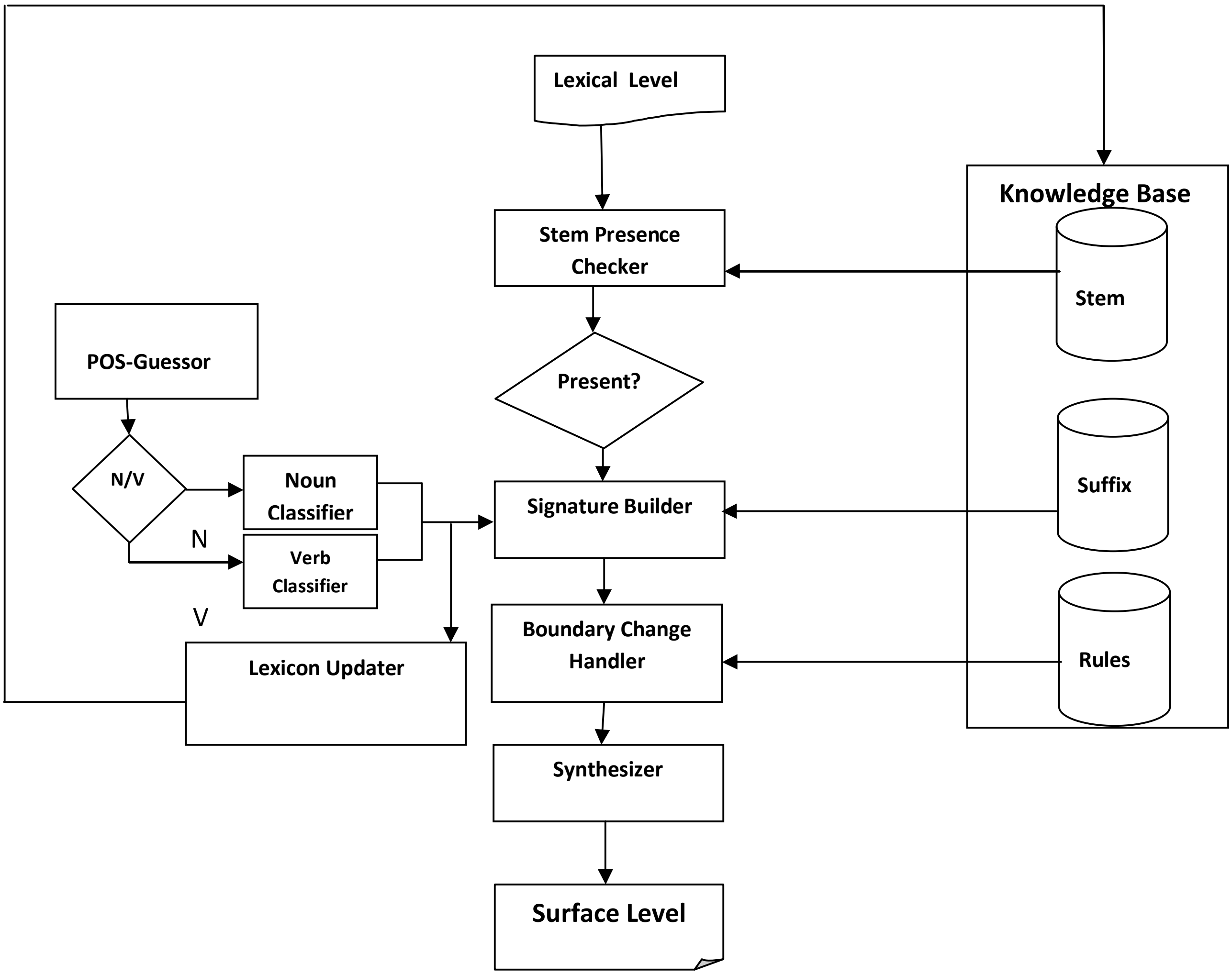

Figure 1: Architecture of Afan Oromo synthesizer.

Implementation

A traditional way to represent the morphology of inflectional languages is through classes of similar stems. Otherwise, one has to make rules for each noun and verb, which is not feasible. Accordingly, the stems in the language are divided into classes. In our classification–based approach, the regularities in the way the stems take person makers for verbs, and plural makers and syllable structure for nouns are used to define classes. Accordingly, six noun classes and ten verb classes were identified.

After listing of stems and affixes, for each stem class there should be corresponding set of affixes with which it can combine to form variants of words. This association is called signature. Signatures are used to organize the stems and suffixes in such a way that the stems in the lexicon or new stems can be organized with appropriate suffixes. After obtaining a complete list of affixes for a given noun class, the concatenation process is done taking care of boundary changes as specified by the rules.

Afan Oromo is different from other languages in morphological properties, patterns of word synthesis and grammatical rules. Thus, the existing algorithms and techniques that are being used to generate word forms of English and other languages are not actually suitable for the language; rather it needs different algorithms and techniques for expected efficiency. Hence, new algorithms have been designed from the scratch. However, some techniques and approaches have also been adopted.

Experiments and Results

Data Sources

To analyze the performance of the synthesizer, we have conducted an experiment. In the experiment, data have been collected from different sources like thesis, reference books, and the internet. After collecting the data, it has been coded in the database in the form suitable for computational scheme. The stems collected have been stored with their tags and subtypes. The suffixes were also compiled according to the type of stem to which they apply.

Different stem types from both verbs and nouns have been selected. The selection was done by linguists according to their representativeness. From each class, stems have been selected to represent the classification criteria such as endings and/or syllable structure for words generation. The generated words are presented to the linguists with the aim of identifying errors.

In this experiment, the error counting approach was adopted to evaluate the word generation algorithm. The number of correctly generated words and incorrectly generated ones are counted for analysis. By preparing questionnaires, the outputs from the synthesizer for verbs and nouns were then checked against the respective valid words by domain experts. The errors were then described in terms of correctness and incorrectness of the produced words. The produced word is said to be correct if it is accepted by speakers of the language in terms of actual and possible words, otherwise incorrect. Throughout the thesis, the statistics used to measure the performance of the system is accuracy. Accuracy refers to the closeness of agreement between a test result and the accepted reference value. The accuracy of the system is calculated as the number of correctly generated words divided by the total number of words generated by the system multiplied by 100%.

Analysis of the Result

As can be seen from Table 1 and Table 2, the algorithms of the synthesizer are tested with 10 verb and 7 noun test stems and the result of each test was checked by linguists [R1,R2,R3,R4] for the validity of generated words. As it can be seen from Table1, the results of the experiments for verbs show an average accuracy of 96.28% correctly generated words. This means that out of about 230 generated words on average, 212 of them are correctly generated. If the average accuracy of synthesized words for each class is considered, we can see that the accuracies are 96.40, 96.68, 97.38, 96.56, 96.80, 95.98, 95.46, 96, 94.46, 97.14 for class1,..., class10 respectively.

Table 1: Test result for selected verb stems.

Stem Name	Class	Number of Words Generated	Number of Correctly Generated words				Number of Wrongly Generated words				Accuracy by each respondent				Average Accuracy
			R1	R2	R3	R4	R1	R2	R3	R4	R1	R2	R3	R4	
beek	V1	222	213	207	206	218	12	9	10	4	95.95	95.95	95.50	98.20	96.40
fedh	V2	286	276	196	223	282	10	8	16	2	96.50	97.20	94.41	98.60	96.68
eeg	V3	200	189	188	191	195	11	4	1	5	94.50	98	99.50	97.50	97.38
barat	V4	254	238	240	245	254	16	12	7	0	93.70	95.28	97.24	100	96.56
barreess	V5	250	236	239	242	243	14	7	4	7	94.40	97.20	98.40	97.20	96.80
awwaal	V6	224	212	202	208	218	12	12	6	1	94.64	94.64	97.32	97.32	95.98
abaar	V7	226	214	199	203	223	13	15	15	3	96.43	93.36	93.36	98.67	95.46
cuuph	V8	200	189	183	181	199	11	9	11	1	94.50	95.50	94.50	99.5	96.00
bushaaw	V9	230	213	165	152	230	17	14	20	0	92.61	93.91	91.30	100	94.46
ajaj	V10	210	196	180	195	210	10	6	4	0	93.33	97.14	98.10	100	97.14
Total Average Accuracy															96.28

From Table 2, we can observe that the average accuracy of generated words for nouns is 97.46% i.e. from the total average of 25 generated words, almost 24 on average are correctly synthesized. The average accuracies of each noun class are 97.41, 97.16, 98.87, 92.92, 100, 100, 100 for class1,… class7 nouns respectively.

Additionally, the experimentation revealed that the major inflectional and derivational categories in Afan

Oromo are verbs. Noun morphology is extremely simpler compared to verb morphology. For instance, a single verb can take on average 230 surface forms whereas a noun takes only about 25 surface forms in average. For both verbs and nouns the incorrect results obtained from this test were due to the fact that suffixes are wrongly applied to wrong stem class. Improper handlings of syllables are worth mentioning.

Table 2: Test result for selected noun stems.

Stem Name	Class	Number of Words Generated	Number of Correctly Generated words				Number of Wrongly Generated words				Accuracy by each respondent				Average Accuracy
			R1	R2	R3	R4	R1	R2	R3	R4	R1	R2	R3	R4	
maatii	N1	29	29	29	27	28	0	0	2	1	100	100	93.10	96.55	97.41
jabbii	N1	29	29	28	28	27	0	1	1	2	100	96.55	96.55	93.10	96.55
beera	N2	22	22	22	21	20	0	0	1	2	100	100	95.45	90.90	96.59
wasiila	N2	22	22	22	20	22	0	0	2	0	100	100	90.90	100	97.73
laga	N3	22	22	22	22	22	0	0	0	0	100	100	100	100	100
mana	N3	22	22	20	22	22	0	2	0	0	100	90.90	100	100	97.73
hiyyeessa	N4	19	19	19	16	17	0	0	2	2	100	100	82.21	89.47	92.92
dureettii	N4	19	18	17	16	17	1	2	3	2	94.44	89.47	82.21	89.47	92.92
Fayyisaa	N7	10	10	10	10		0	0	0	0	100	100	100	100	100
Roobeeraa	N7	10	10	10	10		0	0	0	0	100	100	100	100	100
Halkan	N5	39	39	39	39		0	0	0	0	100	100	100	100	100
foon	N5	39	39	39	39		0	0	0	0	100	100	100	100	100
gamna	N6	27	25	27	25	25	2	0	2	2	92.59	100	92.59	92.59	94.44
dhukkuba	N6	27	26	26	27	27	1	1	0	0	96.30	96.30	100	100	98.15
Total Average Accuracy															97.46

CONCLUSION

To develop word generation algorithm, the knowledge of language morphology is necessarily required. Accordingly, the morphological properties of Afan Oromo have been studied to develop the synthesizer. Then, various techniques to morphological synthesis are reviewed. Rule based approach has been employed as development method because it takes the properties of the language into account. A database consisting of stems, rules and suffixes used as knowledge base has been designed. The boundary change rules such as deletion, epenthesis and assimilation were considered. Most of the algorithms were designed from the scratch as there are no previously designed algorithms for this purpose based on the morphological properties of the language to generate verb and noun forms from an input stem. Some algorithms have also been adopted from other languages. Finally, a prototype morphological synthesizer was developed to evaluate the performance of the designed algorithms.

The analysis of the total number of words generated from single stem especially verbal stem shows that Afan Oromo is morphologically complex language. It has been indicated that verbs are the most productive classes of words in the language. From the experiment, it is possible to say that the performance of the prototype is acceptable. The study has indicated the possibility of developing the synthesizer for Afan Oromo in rule based approach for verbs and nouns. It is pretty easy to extend the system for other parts of speech with minimum effort.

The results obtained from the study encourages the undertaking of further research in the area, especially with the aim of developing a full-fledged Afan Oromo morphological synthesizer. We hope that the performance could have been greater if hybrid would have been applied.

REFERENCES

Allen, J. (1996). Natural Language Understanding. 2nd Ed. California: Redwood, Benjamin/Cummings Publishing Company.

Uibo, H. (2001). On Using the Two-Level Model as the basis of morphological analysis and synthesis of Estonian. Available at www.ut.ee/~heli_u/art/LREC02-Uibo.pdf, 2001.

Salton, G. (1983). Natural Language Processing. Introduction to Modern Information Retrieval. New York: McGraw-Hill, 1983.

Diriba Megersa (2002). Thesis: An automatic sentence parser for Oromo language using Supervised learning technique", Department of Information Science, Addis Ababab University.

Kula Kekeba (2008). Experimentation Report: Evaluation of Oromo-English Cross- Language Information Retrieval", Language Technologies Research Centre IIIT. In IJCAI 2007 Workshop on CLIA, Hyderabad.

Mesifin Getachew (2001). Thesis: Automatic Part of Speech Tagging for Amharic Language: An Experiment Using Stochastic Hidden Makov (HAM) Approach" Department of Information Science, Addis Ababa University.

Karttunen, L. (1994). Constructing Lexical Transducers, In the Proceedings of 15th International Conference of Computational Linguistics. COLING-94. pp. 406–411.

Tom, M. (1997). Mitchell, Machine Learning, McGraw-Hill Publication. ISBN 0070428077.

Andrew Roberts. Machine Learning in Natural Language Processing, http://andy-roberts.net/misc/latex/sessions/bibtex/bib_exampl e_nat.pdf, last visited on Feb, 2009.