Research Article

iMedPub Journals
http://www.imedpub.com/

American Journal of Computer Science and Information Technology
ISSN 2349-3917

2017
Vol.5 No.2:03

DOI: 10.21767/2349-3917.100003

Designing a Rule Based Disambiguator for Afan Oromo Words

Workineh Tesema*, Debela Tesfaye and Teferi Kibebew

Information Science, Jimma University, Jimma, Oromia, Ethiopia

***Corresponding author:** Workineh Tesema, Information Science, Jimma University, Jimma, Oromia, Ethiopia, E-mail: workineh.tesema@ju.edu.et

Received date: March 25, 2017; **Accepted date:** October 14, 2017; **Published date:** October 25, 2017

Citation: Workineh T, Debela T, Teferi K (2017) Designing a Rule Based Disambiguator for Afan Oromo Words. Am J Compt Sci Inform Technol 5: 2. doi: 10.21767/2349-3917.100003

Abstract

This paper presents designing a rule based Afan Oromo Disambiguator. The ultimate aim of this work is to develop a model that identifies the senses of the words. Hence; a word may have multiple senses, the problem is to find out which particular sense is appropriate in a given context. To this end, a rule based approach was used which is designed manually a set of rules. Some ambiguous words were collected from the Oromo society and these words are the most frequently used in the society. Due to under the resource of the language, the work was used 15 natural ambiguous words for the sake of the test. The results of the work were shown that in Afan Oromo language, an ambiguous word have 2 to the n senses (where n unlimited senses; as the number of contexts increased).

Keywords: Rule based; Disambiguator; Afan oromo; Ambiguous word; Word senses

Introduction

The disambiguation is the most challenging at all levels of the natural languages. The ultimate aim of this work was to develop a set of rules that identified the senses of the words. However; the most common way of representing language is *via* of rules [1]. The rule underlies many linguistic theories of the language, which turn into a set of rules [2]. The modifiers and contextual information were the basis of the linguistic properties of Afan Oromo word sense. In this work, the modifier of the ambiguous word is used in order to get the semantic clues of the particular sense in which the ambiguous word is used. To achieve this we analysed the structure of Afan Oromo sentence formation with respect to modifying patterns to develop the rule. The constructed rules used for extracting all modifiers modifying the ambiguous term. The modifiers are words or phrase which provides information about a word and also gives more description about the words it modify. The modifiers (can be a single word or phrase) established for understanding of the ambiguous words.

The motivation behind this work is to allow the users to make ample use of the available technologies in Afan Oromo because ambiguities present in any language provide great difficulty in the use of information technology as words in human language that occur in a particular context can be interpreted in more than one way depending on the different contexts. However, we faced a significant challenge as Afan Oromo has a lack of the resources. So, this work presents a rule based approach came up with an alternate solution to the challenges by obtaining necessary information from the developed set of rules.

The contribution of this paper was towards developing natural language processing applications for Ethiopian languages exhibiting similar patterns with Afan Oromo. Specifically, it increases the scope of the word sense disambiguation research by investigating its applicability for Afan Oromo language. Furthermore, it has been pointed out how Natural Language Processing plays a significant role in enhancing the computer's capability to process word senses. Additionally, IR is also one of the Natural Language Processing (NLP) applications that paybacks from word sense disambiguation which most of the words used to execute queries in IR systems have more than one meaning [3].

Overview of Afan Oromo Language

Afan Oromo is part of the Lowland East Cushitic group within the Cushitic family of the Afro-Asiatic family. Afan Oromo has a very rich morphology like other African and Ethiopian languages [4]. The writing system of Qubee (Latin-based alphabet) has been started since 1842 [4]. It is one of the major African languages that is widely spoken and used in most parts of Ethiopia and some parts of other neighbour countries like Kenya, Tanzania, Djibouti, Sudan and Somalia. It is the second largest Cushitic language in Africa content next to Hausa. Currently, it is an official language of the Oromia state (which is the largest Regional State among the current Federal States in Ethiopia). It is used by Oromo people, who are the largest ethnic group in Ethiopia, which amounts to 50% of the total population [5]. With regard to the writing system, Qubee (a Latin-based alphabet) has been adopted and become the official script of Afan Oromo since 1991 in Ethiopia.

In Afan Oromo, the sense of words fundamentally based on the words preceded by the word (modifies) [6,7]. Hence, the words are described (modified) by the Noun or Verb preceding them. The ambiguous word may appear at the beginning or in

the middle or at the end of a sentence, but modifiers always becomes before the word it modifies. As an example, “Seenaan kaleessa daara bahe”. [Seenaa got cloth yesterday]. From the construction of this sentence, the word “bahe” is modified by the Noun “daara”. According to Afan Oromo structure, the Noun and Verb always appear before the word they modify [8].

Related Works

The rule based approach was important when there is a lack of training data and under resources. The rule-based approach has successfully been used in developing many NLP researches. The research that use rule based approaches are based on a core of solid linguistic knowledge. The rule based approach is for less-resourced languages and for morphologically rich languages like Afan Oromo, which even with the availability of corpora suffer from data sparseness [1].

Rule based approach exploit the hand crafter rule for word sense disambiguation task. The rule based requires extensive work of expert linguists and thus can result in near human accuracy [9]. The Afan Oromo rule based Afan Oromo Grammar Checker, showed a promising result. The results show that rule based is an approach used in the morphologically rich language like Afan Oromo. This rule based approach for languages, such as Afan Oromo, advanced tools has been lacking and are still in the early stages. However, it needs an expert (linguistic knowledge) to develop a set of rules that designing a disambiguator rule in this case. The advantage of this approach is that, it is easy to incorporate domain knowledge into the linguistic knowledge which provides highly accurate results. Furthermore, the linguistic knowledge acquired for one natural language processing system may be reused to build the knowledge required for a similar task in another system [10].

According to Ide et al. [8], the important characteristics of an ambiguous word are grammatical information about the word to be disambiguated, words that are syntactically related, and words that are topically related to the ambiguous word. Since the proposed method relies on the semantic and syntactical information to disambiguate an ambiguous word, for each entry of the sense, it consists of the ambiguous and related words, sense of the ambiguous and related words. Moreover, this information can be converted into understandable rules that best describe the relationship between, the ambiguous word and the related word.

However, a rule based is the most method used in natural language of disambiguation. When there is a lack of available resources and their limitations, the rule based approach was used, which rely on hand-constructed linguistic rule and resources. The use of rule based transformations is based on a core of solid linguistic knowledge [11]. Obviously, the manual creation of rule is an expensive and time consuming effort, which must be repeated every time the disambiguation scenario changes. Knowledge of the linguistic used for word sense disambiguation is either lexical knowledge released to the public or world knowledge learned from a training corpus [12].

Afan Oromo Word Senses

The limited availability of resources in the form of digital corpora and annotated, the rule based method is applied. The linguistic knowledge of the language plays an important role to create the rule. The linguistic knowledge required for the natural language can be obtained in different ways. In this work, the rules were created based on the inherent structure of Afan Oromo in forming sense of the words. However, an effort has been to develop the rule of the language as it discussed above in details.

Proposed rules

In Afan Oromo like other languages, the word sense has its own rule which is manually developed (in our case developed by the researchers of this work). However, the correct sense cannot be only found by choosing the one that is related to another. Promising techniques relied on linguistic knowledge also for extracting semantic features, in our case to mine context of the ambiguous term view of modifiers specializing its sense [13].

The modifiers have a great role to decide on the word sense according to its role in the sentence. The modifiers can appear before the target word (the word, it modifies or describe). Like English, the sentences would be pretty boring without modifiers to provide excitement and intrigue. A modifier adds detail, limits or changes the sense of the other word or phrase.

In Afan Oromo, the words preceding a specific word are more likely to influence the sense of a word.

For example, [Bilisumman sammuu dhahe lafa buuse]. Bilisuma has hits head and make falls on the land.

In this example, the word “[sammuu, lafa]” are modifiers; which give extra information that is part of the sentence. In this case, it is a Noun modifier, because they are modifying the ambiguous word “dhahe”. A modifier should be placed next to the word it describes in Afan Oromo language. Even, the role of modifiers was different, hence, the preceding modifiers carried out more than the following modifiers.

Modifiers of generate a set of rules

Word sense disambiguation was disambiguated based on the surrounding contexts. Additionally, disambiguation is done by analyzing the linguistic features of the word and its preceding word. The rule-based method of our approach disambiguates word automatically using rules in order to complement the features learned from training data. This information is coded in the form of rules. As it discussed in the above sections, the modifiers always precede the target word in Afan Oromo. Based on this notion, the rule was developed by the researcher as follows: Ambiguous Word, Noun has been preceded by Verb modifiers; Ambiguous Word, Verb has been preceded by Noun modifiers; Ambiguous Word has immediately preceded or followed by the modifiers **(Figure 1)**.

The modifiers-rules have the following general form: Modifiers N/V measure-condition <preceding modifiers> Where:

N and V are Noun and Verb modifiers respectively Modifiers are none stop words, numbers and punctuations.

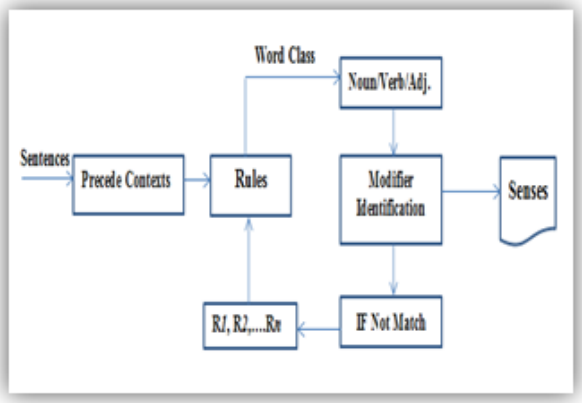

Figure 1: Architecture.

Result

This section was present the result and discussion of the work. The conducted experiment shows that, the semantic has come to the conclusion that the senses of words are closely connected to the words which are modifiers. As shown, the result obtained by rule based approach was great as the semantic information extracted from the distinct set of rules. The most likely reason for this is that our approach relies on automatically assigned immediately preceding words, more reliable and proves to be a most useful linguistic knowledge for word sense disambiguation.

Another issue examined here is the different behaviours of disambiguates for words of different part of speech (verbs and nouns). Out of the 3,656 examples in the complete dataset, 1020 are verb-cases, and 2,546 are noun-cases.

The remaining 90 examples correspond to adjectives and adverbs. The 2,546 nouns-cases represent 538 occurrences of 231 different nouns. Therefore, the linguistic knowledge the best approach to solve word sense disambiguation in Afan Oromo [14,15] as shown in the experiment **(Table 1)**. However, the overall system performance gained thus far is surprising since the developed rules were fresh and still needs to be experts on of the language.

From the finding, the addition of deep linguistic knowledge to a word sense disambiguation system is a significant rise in disambiguation accuracy with the results discussed so far [16,17]. It is especially interesting that using the preceding modifiers of the ambiguous word perform better result. We can conclude that modifiers contain a lot of valuable clues for disambiguation [18,19].

Table 1: Sample test of the experiment.

Target Words	Senses			
	Sense 1	Sense 2	Sense 3	Sense 4
Bahe	Freedom	Get out	Highland	Pass
Handhuura	Center	Gift	Navel	-
Mirga	Direction	Brave	Human	Right
Lookoo	Pretty	Rope	-	-
Dhahe	Hit	Fail	Follow	-
Waraabuu	Fetch	Hyena	Record	-

On the other hand, the set rules were evaluated comparing the result produced by the manually grouped similar contexts of the words in the test set by experts. The evaluation constitutes the following two points:

In order to achieve this we used the following criteria:

How much of the word sense is correct, i.e. to evaluate if all the similar contexts of the ambiguous words are placed in the same group.

Given the number of senses assumed by the words in the test, judge the rule on the basis of the number of senses identified by the rule.

Conclusion

In this work, the rule based approach has improved performance of the disambiguation system. However, it is expensive to develop for local languages which are lack of training corpus. For under resourced Ethiopian language like Afan Oromo the rule approach is recommended. Hence, there is no annotated corpus; rule approach plays a great role to disambiguate. The rule approach relies on hand-constructed rules that are acquired from language specialists rather than automatically trained from data by machines. All the rules described in this work can be a base for this further research and it can support extended disambiguation rules covering most of the terms in the Afan Oromo.

Acknowledgment

I would like to thank Jimma University for their cooperation and financial support. Secondly, I would like to thank all Afan Oromo speakers and Oromia media who have contribute until the end of this work.

References

1. Guya T (2003) CaasLuga Afaan Oromoo: Jildii-1, Gumii Qormaata Afaan Oromootiin Komishinii “Aadaa fi Turizimii Oromiyaa”, Finfinnee.
2. Shao F, Cao Y (2005) A New Real-time Clustering Algorithm, China. Linguistics: Linguistic Studies in Honour of Jan Svartvik, London, Longman.
3. Salton G (2001) The Measurement of Term Importance in Automatic Indexing. J Ame Society Inform Sci.
4. Gragg, Gene B (2006) Oromo of Wollega: non-semetic languages of Ethiopia, East lansing, Michigan state university press.

5. Census Report (2008) "Ethiopia's population now 92 million".
6. Malmkjaar, Kirsten (2005) The linguistics Encyclopedia, New York: Routledge.
7. McCarthy D, Carroll J (2003) Disambiguating Nouns, verbs and adjectives using automatically acquired selectional preferences, Computational Linguistics.
8. Ide N, Veronis J (2008) Introduction to the special issue on word sense disambiguation: the state of the art, Comput, Linguist.
9. Lesk M (2006) Automatic sense disambiguation using machine readable dictionaries, Toronto, Ontario.
10. Rabirra G (2014) Furtuu: Seerluga Afaan Oromoo, Finfinnee Oromiyaa press.
11. Gamta T (2005) Seera Afaan Oromoo, Finfinnee, Boolee Press.
12. Hasan A S (2010) Word Sense Disambiguation and Semantics techniques, Sultan Qaboos University.
13. Tesfaye D (2011) A Rule-based Afaan Oromo Grammar Checker, Addis Ababa, Ethiopia.
14. McCarthy D, Carroll J (2003) Disambiguating nouns, verbs and adjectives using automatically acquired selectional preferences, Computational Linguistics.
15. Megersa D (2002) An Automatic Sentence Parser For Oromo Language Using Supervised Learning Technique, Addis Ababa University, Ethiopia.
16. Tesfaye D (2010) Designing a Stemmer for Afan Oromo Text: A hybrid approach, Master's thesis, School of graduate studies, Addis Ababa University, Ethiopia.
17. Monem A, Shaalan K, Rafea A, Baraka H (2008) Generating Arabic Text in Multilingual Speech-to-Speech Machine Translation Framework, Machine Translation, Springer.
18. Gamta T (2015) Seera Afaan Oromoo, Finfinnee, Boolee Press.
19. Negassa T (2015) word formation: the structure of independent and dependent clauses in oromo, Addis Abeba, Bole.