International Journal of Computer Applications Technology and Research
Volume 10–Issue 03, 51-54, 2021, ISSN:-2319–8656

A Review of Natural Language Processing Techniques: Application to Afan Oromo

Bekele Abera Hordofa
Computer Science Department, College of Science and Technology, Oromia State University, Batu, Oromia
Ethiopia

Shambel Dechasa Degefa
Computer Science Department, College of Science and Technology, Oromia State University
Batu, Oromia
Ethiopia

Abstract: Language is a means of communication and a symbol of national identity. Afan Oromo is one of written and spoken indigenous language in Ethiopia which uses a writing system called Qubee. Natural language processing is automatic or semi-automatic processing of human language that helps computers to understand and process language. NLP techniques involve various linguistic levels to understand and use language. Linguistic levels are an explanatory method for presenting what actually happens within a natural language processing system. This is very important to develop appropriate and desired NLP applications at both higher and lower levels. In this paper, we present a review of techniques, current trends and challenges in NLP application to Afan Oromo.

Keywords: Afan Oromo; Qubee; NLP; NLP Application; Linguistic Level

1. INTRODUCTION

Language is a means of communication and a symbol of national identity [1]. Natural language processing (NLP) is automatic or semi-automatic processing of human language that helps computers to understand and process natural language. It has big role in computer science, because many aspects of the field deal with linguistic features of computation [2].

Afan Oromo is an indigenous Afro-Asiatic language spoken in many parts of Ethiopia and neighboring countries like Kenya, Djibouti and Somalia, which have horn of Africa coverage [3]. Afan Oromo is the second largest Cushitic language in African content next to Hausa. It is spoken and used by 34.5% of the total population of Ethiopia [4]. It is also working language of Oromia regional state, which is one of the largest regional states in Ethiopia.

A number of scholars made huge efforts to transform Afan Oromo from spoken language to a written language [5]. During the Dergue regime writing Afan Oromo in any alphabet, except the Sabean was illegal. Afan Oromo uses a writing system called Qubee. The writing system of Qubee (Latin-based alphabet) has been started since 1842 [3]. The Qubee was accepted unanimously and the first congress of Caffee Oromiyaa put it into a law in 1991 [5]. Since then, Afan Oromo has been a written language, a school language, public media, social issues, religion, political affairs, technology and a working language. Like English, Qubee use constants and vowels (a, e, i, o and u). Every alphabet is pronounced in a clear short/quick or long/stretched sounds. In addition to 26 English alphabets, Qubee uses combination of characters (Qubee dachaa), which is pronounced as single character with the tongue curled back slightly. Examples of Qubee dachaa are ‘ch’, ‘dh’, ‘ny’, ‘ph’, ‘sh’ and ‘ts’. Some examples of words formed from Qubee dachaa, water (bishaan), food (nyaata), butter (dhadhaa), shoe (kophee) and etc.

2. LEVELS OF NATURAL LANGUAGE PROCESSING

Level of natural language processing is the most explanatory method for presenting what actually happens within a natural language processing system. A. Chopra and et al [1] classify phases of linguistic analysis into higher level which corresponds to speech recognition and lower level which corresponds to natural language processing. Linguistics in the science of language classifies level of NLP as shown in Figure 1 [1][2].

Phonology refers to sounds [1]. There are three types of rules used in phonological analysis[1]: phonetic rules, phonemic rules and prosodic rules. Phonetic rules are used for sounds within words. Phonemic rules are used for variations of pronunciation when words are spoken together. Prosodic rules are used for fluctuation in stress and intonation across a sentence. NLP system accepts spoken input, sound waves, analyze it and encode into a digitized signal for interpretation.

Morphology refers to word formation [1] [2]. It is mainly useful for identifying the parts of speech in a sentence and words that interact together. It also describes a set of relations between words’ surface forms and lexical forms [2]. The information gathered at the morphological stage prepares the data for the syntactical stage which looks more directly at the target language’s grammatical structure [2]. Like many local and African languages, Afan Oromo is very rich in morphology. Afan Oromo verbs are highly inflected for gender, person, number and tenses [6]. Both Afan Oromo nouns and adjectives are highly inflected for number and gender [6]. Words can be formed from morphemes in two ways: Derivational Morphology and Inflectional Morphology. Derivational Morphology is concerned with the way words are derived from morphemes through processes such as affixation or compounding while inflectional morphology deals with the combination of a word with a morpheme. Table 1 shows some examples of Afan Oromo morphology.

Table 1: Afan Oromo Morphology (Examples)

Afan Oromo Words	Morphology
Kitaaboota (Books)	Kitaaba[oota]
Alseeruummaa(illegal)	Alseeraa[uummaa]
Namicha (the man)	Nama[icha]
Namoota (men)	Nama[oota]

Syntax refers to the study of structural relationships between words in a sentence [7]. Syntax involves applying the rules of Afan Oromo grammar. It involves analysis of the words in a sentence to depict the grammatical structure of the sentence[2]. In Afan Oromo, a sentence consists of a noun phrase, a verb phrase, and in some cases a prepositional phrase. A noun phrase represents a subject that can be identified by a noun. A verb phrase represents an action. A prepositional phrase modifies a verb or a noun.

Parser is used to convert a sentence into a tree that represents the sentence's syntactic structure. Here words are transformed into structure that shows how the words are related to each other. For example: Chala gave letter to Bontu (Caalaan Boontuudhaf xalayaa kenne). The parser breaks it into noun phrase and verb phrase to determine whether a sentence is valid in relation to the language's grammar rules. This sentence consist of noun phrase: "Caalaan" and verb phrase: "Boontuudhaf xalayaa kenne".

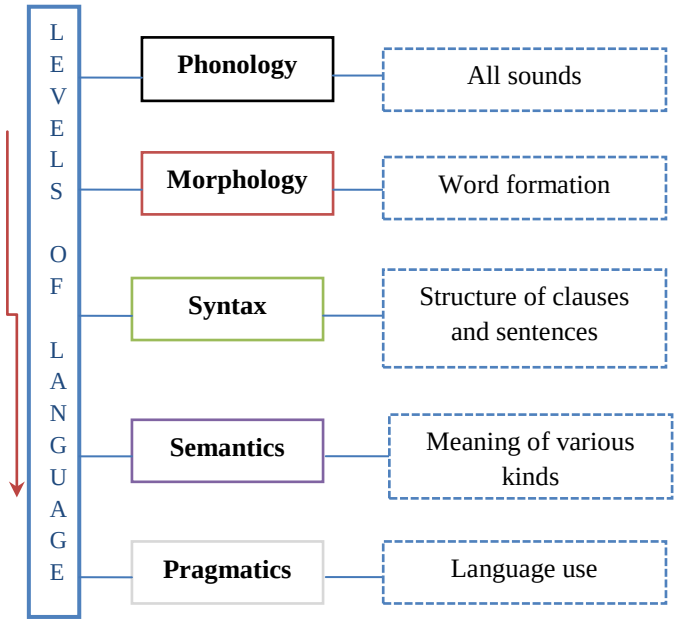

Figure 1: Levels of Natural Language Processing

Semantics are the examination of the meaning of words, phrases and sentences [2][8]. Semantic processing determines the possible meanings of a sentence by focusing on the interactions among word-level meanings in the sentence [1]. All the linguistic levels contribute to determine meaning. Semantic processing includes disambiguation of words with multiple senses. For example: "Bilisumman garaa laafadha". Direct meaning, Bilisumma is kind. Other meaning Bilisumman's stomach is soft. Sense of words and modifiers are used to determine the meaning.

Pragmatics is the analysis of the real meaning of an utterance in a human language, by disambiguating and contextualizing the utterance[9]. This is accomplished by identifying ambiguities encountered by the system and resolving them using one or more types of disambiguation techniques [10][11].

3. NATURAL LANGUAGE PROCESSING TECHNIQUES

Natural language processing has various application areas in different domains. Prakash M. Nadkarni and et al. [8], classified NLP applications into low level and high level tasks. Here in this article, we consider research work done on NLP applications that are important for Afan Oromo.

3.1 Low level NLP Applications

NLP systems need to implement modules to accomplish mainly the lower levels of processing. The lower level application may not require interpretation of the higher levels and relatively more researched and attempted. The lower levels deal with smaller units of analysis like characters, tokens, morphemes, words, and sentences, which are rule-governed. Some of lower level NLP applications are:

Tokenizer: Most text processing works with word and sentence based unit therefore larger blocks of text is split into single words and sentences [12]. Tokenization involves word and sentence boundary detection, and problem specific segmentation. It starts with a sequence of characters to identify the elementary parts of natural language such as words, punctuation marks and separators. In Afan Oromo, like English white space is used to separate words. An end of a statement is marked with full stop (.), while comma (,) is used to separate lists or ideas just like the comma in English.

Morphological Analyzer: Morphological analyzer is a process of returning one or more surface forms from a sequence of morpheme glosses [13]. The most common and widely used approaches for automatic morphological synthesizer are: rule based, machine learning and hybrid approaches. Abebe Abishu [13] designed a rule based morphological synthesizer for Afan Oromo particularly for verbs and nouns.

Part of Speech Tagger (POST): Part-of-speech tagging is the act of assigning each word in sentences a tag that describes how that word is used in the sentences [14]. That means POS tagging assigns whether a given word is used as a noun, adjective, verb, and etc. There are two known approaches that are used to develop part-speech-tagger: Rule based Approach and Stochastic Approach [14]. Getachew Mamo and Million Meshesha [14] attempted Afan Oromo part of speech tagger using Hidden Markov model.

3.2 High Level NLP Applications

Higher level NLP applications are built on bases of the low-level tasks and are usually problem specific.

Spelling and Grammatical Checker: Afan Oromo is written in the way it is spoken, this makes more vulnerable to spelling error [15]. Afan Oromo is morphologically rich language, each root word can combine with multiple morphemes to generate huge number of word forms [15]. Because of these and other reasons explained in Table 2 development of Afan Oromo spell checker is a challenging task.

Grammar checker determines the syntactical correctness of a sentence which is mostly used in word processors and compilers. There are three popular approaches used for grammar checking; syntax-based checking, statistics-based checking and rule-based checking. Debela Tesfaye [6], attempted rule based Afan Oromo grammar checker.

Named Entity Recognition (NER): Named Entity Recognition is an information extraction task aimed at identifying and classifying words of a sentence, a paragraph or a document into predefined categories of named entities [16]. Named entities are categorized into different class of named entity like people, organization, place, time etc. NER is very essential in almost all NLP applications like information extraction, search engines, machine translation and question-answering, etc. N.Kannaiya Raja and et al.[16], attempted a rule based named entity recognition.

Table 2: Afan Oromo spell and grammar checker (cases and examples)

Reasons(cases)	Description	Example(s)
Single consonant (jecha laafaa) vs Double consonant (jecha jabaa)	Single consonant: the sounds are less emphasized. Double consonant: the sounds are more emphasized.	Bad(badaa), highland(badda a), etc.
Single vowels (jecha gabaabaa) vs Double vowels (jecha dheeraa)	Single vowels: the sounds are less stretched or elongated. Double vowels: the sounds are less stretched or elongated.	Earth(lafa), weak(laafaa), etc.
Consonant followed by other consonant (CC), (Jecha irra butaa)	Sound pronounced the air drawn in so that a glottal stop is heard before the following constant begins.	Hand(harka), bag(boorsaa), etc.
Use of ‘ (Jecha hudhaa)	Sound pronounced the air drawn in so that a glottal stop is heard before the following vowel begins.	Month(ji’a), goat(re’ee), etc.
Compound word(Jecha tishoo)	Words formed from two words	Saba + lammii = sablammii or sab-lammii
Morphology	A number of words can be derived or inflated from single root.	deemi, deeme, deema, deemte, deemteetti, etc.

Information Extraction (IE): Information Extraction concerned with the automatic extraction of facts from text and stores them in a database for easy use and management of the data [17]. Most IE systems are domain specific which involves extracting meaningful information from unstructured text data and presenting it in a structured format. There are different approaches to IE; rule-based, supervised machine learning and semi-supervised approach. Sisay Abera and Tesfa Tegegne[17] attempted news domain supervised machine learning Afan Oromo IE model.

Machine Translation (MT): Machine translation is an automatic translation of text from a source language to its counterpart in a target language [18]. Machine translation has its own challenges like translation of low-resource language pairs, translation across domains, translation of informal text and translation form/to morphologically rich languages. Machine translation gives a quick and comprehensive understanding of a text or document written by another language. MT has different approaches, including rule based, corpus based and hybrid approach. Million Meshesha and Yitayew Solomon [18], attempted English-Afan Oromo statistical machine translation.

4. CURRENT TRENDS AND CHALLENGES IN NLP

Currently NLP is hot research areas. More over for under resourced language like Afan Oromo, it is highly realistic to conduct research on NLP applications. An NLP applications demand is growing in an exponential manner. The reason behind this growth is transfer of technology from manual to automated and many other tasks which are required to be automated and involve language at some point. A number of researches in natural language processing have been done or going on Afan Oromo, but applications don’t available publicly. Alongside these researches some prototype were developed to demonstrate the effectiveness of particular applications, but still real implementations of these applications are rare.

Even though there is massive production of raw data on web, the availability of quantity and free dataset is rare for Afan Oromo. This implies that more effort has to be done on preparing corpus.

Text processing, writing is the basic fundamental unit of NLP applications. NLP applications revolves around language’s which refers to words in its basic raw form. The performance of NLP applications is also another issue. This challenge is improved through time. In the future these applications will turn from human-computer interaction to human computer conversation. To accomplish these, there is a necessity of integration of many modern-day technologies such as recognition of human users, sentiment analysis, recommendation analysis and techniques with the engagement in conversations is possible in a dynamic manner.

Another challenge in natural language processing involves speech recognition, natural language understanding, and natural language generation.

NLP researchers are now developing next generation NLP systems that deal reasonably well with general text and account for a good portion of the variability and ambiguity of language. Human level or human readable natural language processing is an AI-complete problem[1]. This challenge is highly tied with advancement of artificial intelligence. Some NLP applications at both lower level and higher level need cloud sourcing.

5. CONCLUSION

Natural language is any ordinary language that is spoken or written by humans for general purpose communication. Natural language processing (NLP) is processing of human language that helps computers to understand and process human language. NLP is a relatively recent area of research and application, as compared to others. With the help NLP applications, we can develop beneficial and successful NLP systems. As lower NLP applications mature it minimize challenges in language use and further it can be embedded into higher level applications. NLP applications will continue to be a major area of research and development in information systems now and far in a future.

6. REFERENCES

[1] Y. Wilks, "Natural Language Processing," Commun. ACM, vol. 39, no. 1, pp. 60–62, 1996, doi: 10.1145/234173.234180.

[2] D. M. P. P. Alpa Reshamwala, "Review on Natural Language Processing," Eng. Sci. Technol. An Int. J., vol. 3, no. 1, pp. 2250–3498.

[3] I. Bedane, "The Origin of Afan Oromo : Mother Language," Glob. J. Hum. Soc. Sci. G Linguist. Educ., vol. 15, no. 12, 2015.

[4] E. C. S. A. (ECSA), "Population and Housing Census of Ethiopia," 2007.

[5] Https://oromiaacademy.wordpress.com/, Oromia Language & Cultural Academy. .

[6] D. Tesfaye, "A rule-based Afan Oromo Grammar Checker," Int. J. Adv. Comput. Sci. Appl., vol. 2, no. 8, pp. 126–130, 2011, doi: 10.14569/ijacsa.2011.020823.

[7] M. Synthesizer and A. Abeshu, "Analysis of Rule Based Approach for Afan Oromo Automatic," vol. 7522, no. 4, pp. 94–97, 2013.

[8] P. M. Nadkarni, L. Ohno-Machado, and W. W. Chapman, "Natural language processing: An introduction," J. Am. Med. Informatics Assoc., vol. 18, no. 5, pp. 544–551, 2011, doi: 10.1136/amiajnl-2011-000464.

[9] D. Khurana, A. Koli, K. Khatter, and S. Singh, "Natural Language Processing: State of The Art, Current Trends and Challenges," no. Figure 1, 2017, [Online]. Available: http://arxiv.org/abs/1708.05148.

[10] W. Tesema, D. Tesfaye, and T. Kibebew, "Designing a Rule Based Disambiguator for Afan Oromo Words," Am. J. Comput. Sci. Inf. Technol., vol. 05, no. 02, pp. 3–6, 2017, doi: 10.21767/2349-3917.100003.

[11] W. Tesema, D. Tesfaye, and T. Kibebew, "Towards the sense disambiguation of Afan Oromo words using hybrid approach (unsupervised machine learning and rule based)," Ethiop. J. Educ. Sci., vol. 12, no. 1, pp. 61-77–77, 2016.

[12] B. A. Hordofa, "Event Extraction and Representation Model from News Articles," vol. 16, no. 3, pp. 1–8, 2020.

[13] A. Abeshu, "Analysis of Rule Based Approach for Afan Oromo Automatic Morphological Synthesizer," Sci. Technol. Arts Res. Journal, ISSN 2226-7522(Print) 2305-3327, vol. V–2, no. I–4, pp. 94–97.

[14] G. Mamo and M. Meshesha, "Parts of Speech Tagging for Afan Oromo," Int. J. Adv. Comput. Sci. Appl., vol. 1, no. 3, pp. 1–5, 2011, doi: 10.14569/special issue. 2011.010301.

[15] G. O. Ganfure and D. Midekso, "Design And Implementation Of Morphology Based Spell Checker," Int. J. Sci. Technol. Res., vol. 3, no. 12, pp. 118–125, 2014.

[16] S. S. N.Kannaiya Raja, Naol Bakala, "NLP: Rule Based Name Entity Recognition," Int. J. Innov. Technol. Explor. Eng. ISSN 2278-3075, vol. Volume-8, no. Issue-11.

[17] T. T. Sisay Abera, "Information Extraction Model for Afan Oromo News Text," in International Conference on Information and Communication Technology for Development for Africa, p. pp 327-340.

[18] M. Meshesha and Y. Solomon, "English-Afan Oromo Statistical Machine Translation," no. 9, pp. 26–31, 2018.