BILINGUAL PUBLISHING CO.
Pioneer of Global Academics Since 1984

Journal of Computer Science Research
https://ojs.bilpublishing.com/index.php/jcsr

REVIEW

Natural Language Processing and Its Challenges on Omotic Language Group of Ethiopia

Girma Yohannis Bade*

Department of Computer science, School of Informatics, Wolaita Sodo Uninveristy, Wolaita, Ethiopia

ARTICLE INFO

Article history
Received: 23 August 2021
Accepted: 26 September 2021
Published Online: 13 October 2021

Keywords:
Omotic group
NLP
Challenges
Application

ABSTRACT

This article reviews Natural Language Processing (NLP) and its challenge on Omotic language groups. All technological achievements are partially fuelled by the recent developments in NLP. NPL is one of component of an artificial intelligence (AI) and offers the facility to the companies that need to analyze their reliable business data. However, there are many challenges that tackle the effectiveness of NLP applications on Omotic language groups (Ometo) of Ethiopia. These challenges are irregularity of the words, stop word identification problem, compounding and languages 'digital data resource limitation. Thus, this study opens the room to the upcoming researchers to further investigate the NLP application on these language groups.

1. Introduction

1.1 Natural Language Processing (NLP) and its Application

You probably saw the news on the latest digital assistants that can book your next any appointment over the phone. And heard about the Artificial Intelligence (AI) algorithm that can answer eighth grade elementary science questions better than humans. You may have even interacted with a chatbot that can answer your simple banking questions. You are possibly carrying a mobile phone that can translate your sentences to 100 different languages in real time. All these technological achievements are partially fueled by the recent developments in NLP [1]. NPL is an application of artificial intelligence and offers the facility to companies that need to analyze their reliable business data. According to [2], the NLP's market is expected to grow 14 times in 2025 than it was in 2017. Some of the application of NLP are as follows according to [2]:

Market Intelligence

Marketers can use natural language processing to understand their customers in a better way and use those insights in creating effective strategies.

Sentiment Analysis

Humans have the gift of being sarcastic and ironic during conversations. With sentiment analysis, NPL helps to manage the mentions on social media and tackle them before they disseminate.

Hiring and Recruitment

We all will agree that the HR department performs one

**Corresponding Author:*
Girma Yohannis Bade,
Department of Computer science, School of Informatics, Wolaita Sodo Uninveristy, Wolaita, Ethiopia;
Email: girme2005@gmail.com

DOI: https://doi.org/10.30564/jcsr.v3i4.3614

of the most crucial tasks for the company: With the help of Natural Language Processing, this task can be done more easily.

Text Summarization

This is one of the NLP application and used to summarize the most important information from the vast content to reduce the process of going through the whole data in news, legal documentation and scientific papers.

Survey Analysis

The problem arises when a lot of customers take these surveys leading to exceptionally large data size. All of it cannot be comprehended by the human brain. That's where natural language processing enters the canvas. These methods help the companies to get accurate information about the customer's opinion and improve their performance.

Machine Translation

Machine Translation is one of the applications of NLP and uses a neural network to translate low impact content and speed up communication with its partners.

Email Filters

NLP makes use of a technique called text classification to filter emails. It refers to the process of classification of a piece of text into predefined categories.

Grammar Check

Yes, this natural processing technique is here to stay. Tools like Grammar provide tons of features in helping a person write better content. It is one of the most widely used applications of NLP that helps professionals in all job domains create better content.

Stemming algorithms

Stemming algorithms are commonly known in a domain of Natural Language Processing (NLP) and which has a positive impact on Information Retrieval (IR) system and Morphological Analysis. Now a day, information technology has contributed a great availability of recorded information to exist. The mass production of electronic information, digitalized library collections and the awareness of society, increased the demand for storing, maintaining and retrieving information in a systematic way [3]. Information retrieval (IR) one of such a systematic ways is designed to facilitate the access to stored information [4]. To enhance the effectiveness of IR performance, the suffix stripping process (stemmer) helps by reducing the different variants of terms into common forms as conflating the variants of words [9].

1.2 Capabilities NLP

Sentences segmentation: identifies where one sentence ends and another begins. Punctuation often marks sentence boundaries.

Tokenization: identifying individual words, numbers, and other single coherent constructs. Hashtags in Twitter feeds are example of constructs consisting of special and alphanumeric characters that should be treated as one coherent token. Languages such as Chinese and Japanese do not specifically delimit individual words in a sentence, complicating the task of tokenization.

Part-of-speech tagging: assigns each word in a sentence its respective part of speech such as a verb, noun, or adjective.

Parsing: derives the syntactic structure of a sentence. Parsing is often a prerequisite for other NLP tasks such as named entity recognition.

Name entity recognition: identifies entities such as persons, locations, and times within documents. After the introduction of an entity in a text, language commonly makes use of references such as 'he, she, it, them' instead of using the fully qualified entity. Reference resolution attempts to identify multiple mentions of an entity in a sentence or document and marks them as the same instance.These methods can tell us what people are saying, feeling, and doing or determine where documents are relevant to transactions. Companies need a new approach to combine the structured and unstructured components—the old ways don't really work—they just aren't effective.

1.3 Omotic Languages

Ethiopians are ethnically diverse [5], with the most important differences on the basis of linguistic categorization. Ethiopia has 86 different languages that can be classified into four major groups. The vast majority of languages belong to the Semitic, Cushitic, Omotic Group and Nilo-Sahara, these all are part of the Afro-Asiatic family [7].

The Omotic languages are predominantly spoken between the Lakes of southern Rift Valley and Southwest of Ethiopia around River Omo (hence their name). These language groups have 28 languages. However, they are little studied and the Afro-Asiatic membership of Omotic is controversial being regarded by some as an independent family. Omotic have affinities with Cushitic, another branch of Afro-Asiatic. The following table shows the

 DOI: https://doi.org/10.30564/jcsr.v3i4.3614

lists of Omotic languages.

Table 1. Lists of Omotic languages

Anfillo	Dime	Kachama-Ganjule	Nayi
Ari	Dizzi	Kara	Oyda
Bambassi	Dorze	Kefa	Shakacho
Basketto	Gamo-Gofa	Kore	Sheko
Bench	Ganza	Male	Welaita
Boro	Hammer-Banna	Melo	Yemsa
Chara	Hozo	Mocha	Zayse-Zergulla

1.4 Morphology in Omotic Language Groups

There is no grammatical gender. The main identification is between animate and inanimate. In animate nouns, gender is determined by sex. Inanimate nouns are inflected like masculine nouns. Only definite nouns are marked for plural, and the singular is unmarked. Omotic distinguishes subject and object by case suffixes as well as by tonal inflection. In some languages the subject case is marked (nominative) while the object remains unmarked (i. e., identical to the quotation form of the noun). In other languages the object is marked (accusative) while the subject is unmarked. The Ometo group shows a predominance of marked-nominative languages, whereas other North Omotic languages and the South Omotic ones have, mostly, accusative systems [8].

Morphology is the study of word structure [9]. All languages have word and morphemes. Morphemes are the minimal units of words that have a morphological meaning and cannot be subdivided further. There are two main types of morphemes: free and bound. Free morphemes can occur alone and bound morphemes must occur with another morpheme [9]. For the word "badly", an example of a free morpheme is "bad", and an example of a bound morpheme is "ly". The morpheme "ly" is bound because although it cannot stand alone. It must be attached to another morpheme to produce a word.

The Omotic Language has 29 consonant phonemes, including voiced glotalized consonant, which have been analyzed as consonant clusters. It also has five vowel phonemes, which can be combined to long vowels and diphthongs. In Ometo, there are two ways of forming words, affixation and compounding. Among three types of affixes (prefix, infix, suffix); suffixation, adding suffix (morpheme) to the word at the end is common. This process, in Ometo makes word lengthy [5].

Basically three types of affixes are there, prefixes, infixes, and suffixes.

Prefix:-is a morpheme that can be attached at the beginning of the word.

Infix:-this morpheme can be found at between of a word.

Suffix:-this can be added at the last of the word.

However, among these three morphemes, prefix and infix morphemes do not exist in Omotic group, the only morpheme that exists in Omotic group is suffix.

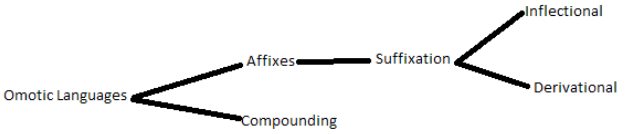

Figure 1. Omotic group morphology

2. Challenges of NLP

General challenges

NLP is a powerful tool with huge benefits, but there are still a number of Natural Language Processing limitations and problems:

- Contextual words and phrases
- Synonyms
- Irony and sarcasm
- Ambiguity
- Errors in text or speech
- Colloquialism and slang
- Domain specific languages
- Low resource languages
- Lack of research and development

Ambiguity

Ambiguity in NLP refers to sentences and phrases that potentially have two or more possible interpretations.

Lexical ambiguity

A word that could be used as a verb, noun, or adjective.

Semantic ambiguity

The interpretation of a sentence in context. For example, I saw the boy on the beach with my binoculars. This could mean that I saw a boy through my binoculars or the boy had my binoculars with him.

Syntactic ambiguity: In the sentence above, this is what creates the confusion of meaning. The phrase with my binoculars could modify the verb, "saw," or the noun, "boy."

Even for humans this sentence alone is difficult to interpret without the context of surrounding text. POS (part of speech) tagging is one NLP solution that can help solve the problem, somewhat.

DOI: https://doi.org/10.30564/jcsr.v3i4.3614

NLP Specific challenges on Omotic languages

a)Irregularity

Some of words in Omotic group are irregular. For instance, addussa 'long', adduqqees 'is getting long', and addussatetta 'length'are basically one word 'Long' classes but the stemmer results it in two forms'adduss-' and 'adduqq-'in Wolaita language which is one the popular Omotic language group. Even if it is possible in view of linguistic, it has a different sense in Information Retrieval point of view.

b)Stop words Concept

Stop words are functional (non-content bearing) words. They give sense for other words. The most suggested function words or stop words are propositions, conjunctions, articles and such likes. For example, the man jumped down. Here bolded terms are stop words. The issue of stop words is also worth mentioning in relation to retrieval effectiveness. The removal of stop words from indexing and query, results in effectiveness of retrieval by reducing storage requirement and increasing the matching of a query with index terms of a document [6]. Stop words in English, known and identifiable but in Omotic group they very confusing to identify them from others, for instance in Wolaita "I ba bala qottis" means He hide his mistakes. Form this statement "ba " has two literal meaning. One is "to go" and other indicate as stop word 'his'. SO identifying stop words in Omotic language group is very challenging task.

c)Compound word concept

Compounding is the second main word formation process in Omotic language group. Even if compound morphemes are rare in Omotic language, their formation process is irregular. As a result, it is difficult to determine the stem of compounds from which the words are made. The example below shows compounding in Wolaita which accounts majority of Ometo.

wora-kanna 'jackal', literally 'the dog of the forest'
demba hariya 'zebra, literally 'field donkey'
keetta-'asa 'family', literally 'the people of a house'
mache-'isha 'brother-in-law', literally 'wife's brother'
aaye-michchiyo 'aunt', literally 'mother's sister'

One of the application of NLP is stemming but somebody may face confusion which one to stem to get root word.

d)Limitation of digital datasets

As Omotic language is one of the least resourced language in Ethiopia, it suffers from the limitation of digital datasets. If somebody wants to make research on one of Omotic languages, he or she has to collect the real time data. There is no collected online accessible Omotic digital data.

Other challenges for NLP researches [10]:

Improvement of the performance of individual analyzers, especially at the semantic/pragmatic level,i. e having single actor but in different place, coreference system would consider them two different entities.

Domain adaptation methods to tune generic NLP processors to deal with process descriptions in a specific organization or sector. This may require the creation/ acquisition of tailored ontologies that help specifying with the right terms important parts of the process and relations among these relevant domain concepts.

Definition of new tasks, such as the detection of exclusivity, parallelism/concurrency, decision points, or iteration of tasks described in the text.

Use of world knowledge to improve the results

Each natural language has its own characteristics and features. So, it's quite difficult to follow the rules of ones for other languages. This is because of different prefixes and suffixes, and exceptions needs a special handling and a careful formation of frame with specific norms. These issues common for all languages not only for Omotic language groups.

3. Conclusions

Omotic language groups are morphologically rich language. This effect is due to its inflectional and derivational morphologies. Suffixes in Omotic language plays a great role in forming many variants of words. As a result, more than one combination of the suffixes can be appended to the root and; thus the length of the words in Omotic language is very long. Stemmer one of the NLP applications in information retrieval (search engines) is increasing recall without decreasing precision, because both document indexes and queries use stems.

Even though there are many capabilities in NLP like sentences segmentation, part-of-speech tagging, tokenization etc, there is a challenges in NLP researches like ambiguity, synonyms, contextual words, improvement in individual analyzer, and definition of new task. Event the above challenges are common to all languages, the most affecting challenges for NLP application on Omotic language groups are irregularity, stop words identification, compounding and limitation of digital datasets.

References

[1] https://www.software.slb.com/blog/natural-lan-

 DOI: https://doi.org/10.30564/jcsr.v3i4.3614

guage-processing---the-new-frontier?gclid=EAIaIQobChMI3NKS3ISu8gIVBertCh2AVA1yEAAYASAAEgL96vD_BwE.
[2] https://www.mygreatlearning.com/blog/trending-natural-language-processing-applications/.
[3] Lemma Lessa. "Development of stemming algorithm for wolaytta text."M.Sc. Thesis, Addis Ababa University, Department of Information Science, Addis Ababa,(2003).
[4] Salton,G. & McGill, N."Introduction to Modern Information Retrieval". New York: McGraw-Hill, (1983).
[5] O'Grady, W., (1997). Contemporary Linguistics: An Introduction. London: Longman.
[6] Savoy, Jacques. (1993). "Stemming of French Words Based on Grammatical Categories." In Journal of American Society for Information Science, 44(1):PP. 1-9.
[7] https://www.britannica.com/topic/Omotic-languages.
[8] http://www.languagesgulper.com/eng/Omotic.html.
[9] O'Grady, W., (1997). Contemporary Linguistics: An Introduction. London: Longman.
[10] Challenges and Opportunities of Applying Natural Language Processing in Business Process Management.
[11] https://tanzu.vmware.com/content/blog/3-key-capabilities-necessary-for-text-analytics-natural-language-processing-in-the-era-of-big-data.

DOI: https://doi.org/10.30564/jcsr.v3i4.3614