Out of vocabulary words handling in morphological analysis

Amit Asthana[a], Ganesh Chandra[b]

[a] Department of Computer Science, Babasaheb Bhimrao Ambedkar University, Lucknow, India

[b] Department of Computer Science, Babasaheb Bhimrao Ambedkar University, Lucknow, India

aamitonline@gmail.com, ganesh.iiscgate@gmail.com

KEYWORD

Natural Language Processing; Morphological Analysis;

ABSTRACT

Morphological analysis is the first step in Natural Language Processing (NLP). It paves the way for future analysis and NLP procedures to be completed. Morphological analysis is the act of identifying morphemes in a phrase by studying each word individually. Out of vocabulary (OOV) words are words that are present in a phrase but for which the morphological analyzer is unable to discover a morpheme. In NLP, identifying OOV terms is a challenge. If OOV terms are not detected, it may be difficult to discern the sentence's true meaning. The goal of this research study is to provide a mechanism for identifying OOV words in Hindi during morphological analysis.

1. Introduction

NLP is a phenomenon or a combination of approaches that allows computer system components (hardware or software) to create, analyse, and interpret language in the form of audio or text. The goal of NLP development is to make it possible for a user to communicate successfully with a computer system.

The term "natural" in NLP refers to the common language that humans use to interact with one another, as opposed to formal languages such as those used in computer programming and mathematical symbols. A language may be characterized as a set of symbols combined with a set of rules. The information is built and transformed using a collection of symbols. The rules are applied to a set of symbols in order to create meaningful data. NLP works in different phases like morphological analysis, syntactic analysis, semantic analysis, discourse integration and pragmatic analysis. The output of each phase works as the input for the next phase. So, the morphological analysis phase provides the fuel to ignite the further phases of NLP.

The word morphology is derived from antique Greek word ‘morphe’ which means shape or form. Morphology may be defined as "the study of form or pattern", i.e. the study of the shape and arrangement of parts of the constituents of an object, and how these combined to develop a whole object. In NLP, a morpheme is the meaning root component of word from which the word is derived. So, in terms of NLP, morphology is the study of morphemes and their arrangements in forming words.

Corresponding Author: Amit Asthana, Department of Computer Science, Babasaheb Bhimrao Ambedkar University, Lucknow, India
Email: aamitonline@gmail.com

Morphological analysis is the initial phase in NLP (DebasriChakrabarti, HemangMandalia, RitwikPriya, Vaijayanthi Sarma and Pushpak Bhattacharyya, 2008). In this phase the words present in a sentence is analysed. The analysis of each word is very essential to extract the meaning of the whole sentence. There are many words that may not be available in the corpus that is used for meaning extraction. These are known as the out of vocabulary (OOV) words. If the OOV words are present in a sentence, then it is the morphological analyser will not be able to analyse that word correctly and hence all the further steps will be affected to efficiently translate the sentence and meaning extraction. Therefore, it is essential to find the solution for the OOV words. Hindi is much more complex and lexical rich language. In Hindi language multiple words may be combined to form a new word using a hidden connector like conjunction or preposition. This is known as SAMAAS in Hindi language.

The languages may be divided into two types- Segmented languages and Non-segmented languages (DebasriChakrabarti, HemangMandalia, RitwikPriya, Vaijayanthi Sarma and Pushpak Bhattacharyya, 2008). The segmented languages may be defined as the languages in which words in a sentence are separated by blank spaces like English. The non-segmented languages are the languages in which words are not separated by blank spaces, like Chinese, Japanese, etc. The segmented languages are easier to tokenize in comparison to non-segmented languages because of the available marker i.e. the blank space. Hindi language is mainly a segmented language but there can be formed multiple compound-words by combining (Sandhi) two or more words together. These compound words can be formed that cannot be separated for the extraction of their meaning.

So in this research paper proposes a methodology to extract the meaning of such kind of complex-words that are considered as OOV words due to unavailability into the corpus used.

2. Related Work

As of prior, most machine translation projects included translation between dialects with by and large negligible morphological construction. A few research projects have analyzed the usage of morphology to improve translation quality.

(Debasri Chakrabarti, 2008) gave an algorithm for automatic identification of the verb + verb lexical compound verbs for Hindi. (Mugdha Bapat, 2010) proposed a Paradigm-Based Finite State morphological analyzer for Marathi. Raj Dabre (2013) proposed a compound word analyzer for Marathi and described various possible types of compound words that can be formed in Marathi. (Girish kumar Pinkeye, 2018) attempted to paraphrase noun compounds using prepositions by considering noun compounds and their corresponding prepositional paraphrases as parallelly aligned sequences of words. They encoded a noun compound and its prepositional paraphrase through two different LSTM (Long Short Term Memory) then train a network such that the encodings of a noun compound and its corresponding prepositional paraphrase have high similarity.

3. Methodology

Identifying the words that are out of vocabulary (OOV) is an important task in morphological analysis of Hindi sentence. One way of identifying OOV words is to match the token with a rich corpus that has almost every word in Hindi language and another is proposed in this research paper to identify the words into the token that is OOV word. As it has been found that many words that are unidentified or OOV words are compound words i.e. the combination of more than one word, so I am proposing a method to reduce the count of OOV words during morphological analysis.

Example- There are Hindi words like (xeSahiwakArI) देशहितकारी, (rAmaBakwa) रामभक्त which are compound words having more than one words in it. There are many other words like these, on which it is difficult to apply the morphological analysis as the morphological analyzer will not find any word like it into the corpus.

3.1. Algorithm

The modified algorithm is as follows:

Step 1: The user enters the input in the form of a Hindi word or a sentence.

Step 2: Identifying and matching the root word

2.1 If the user entered a single word then it will be matched with the root word in the corpus and its morphological features are fetched.

2.2 If it is not matched with the words in the corpus, it means that it is not a root word so lemmatizer will be used to extract the root of the entered word.

Step 3: Exception Handling

In this step the output of the lemmatizer is matched with the words in the corpus, if the word matches then it is displayed.

3.2 If the word is not matched with the corpus, the proposed methodology is applied to extract the word(s) into the entered word.

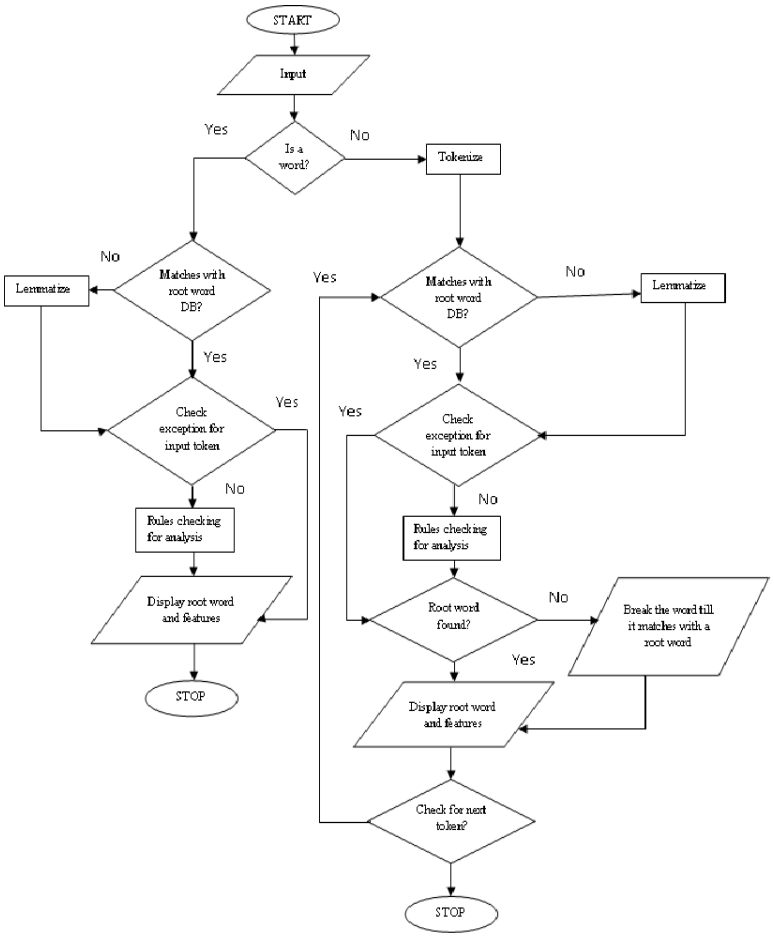

Figure 1: Flowchart to apply the proposed methodology over the compound words.

4. Result

After our experiments of compiling random complex word queries, the results found as shown in fig-2. It has been found that after applying the methodology over the compound words present in the sentence the morphological analyzer extracts better results of morphological analysis.

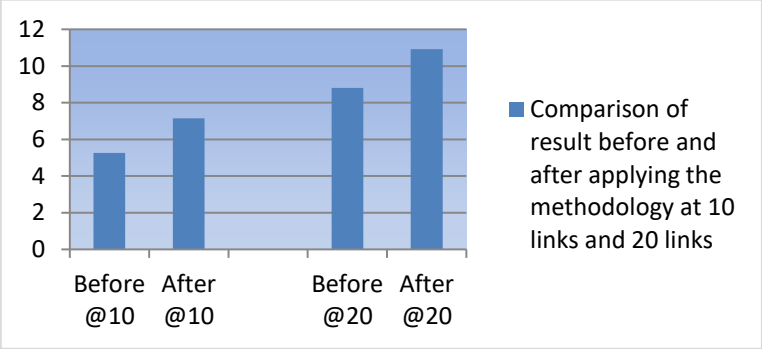

Figure 2: Shows comparison of result before and after applying the methodology at 10 links and 20 links of search result

5. Conclusion and Future Scope

The proposed methodology enables the morphological analyzers to efficiently identify the out of vocabulary words. It certainly reduces the occurrences of out of vocabulary words from the sentences. Out of vocabulary complex words are difficult to extract the morphological information from them but applying the proposed methodology gives the needed improvisation to the corpus to identify the words efficiently that is not available. As per our experiments it is observed that 35.67% of result increased after applying the methodology on considering 10 results and 24.1%

The further enhancement is underlined to identify the relation between the words into the complex word, i.e. known as 'samaas' in Hindi. The hidden meaning identification within a complex word is very essential in order to extract the right meaning of the words.

References

DebasriChakrabarti, HemangMandalia, RitwikPriya, VaijayanthiSarma and Pushpak Bhattacharyya, (2008) Hindi Compound Verbs and their Automatic Extraction, Computational Linguistics (COLING08), Manchester, UK

GirishkumarPonkiya, Kevin Patel, Pushpak Bhattacharyya and Girish Palshikar (2018), Treat us like the sequences we are: Prepositional Paraphrasing of Noun Compounds using LSTM, COLING 2018, Santa Fe, New-Mexico, USA, August 20-26

GirishkumarPonkiya, Rudra Murthy, Pushpak Bhattacharyya and Girish Palshikar; (2020) Looking inside Noun Compounds: Unsupervised Prepositional and Free Paraphrasing using Language Models, In Findings of Int'l Conf. on Empirical Methods in Natural Language Processing (findings of EMNLP),16-20.

GirishkumarPonkiya, Kevin Patel, Pushpak Bhattacharyya and Girish K. Palshikar, (2018), Towards a Standardized Dataset for Noun Compound Interpretation, LREC 2018, Miyazaki, Japan, May 7-12

MugdhaBapat, HarshadaGune and Pushpak Bhattacharyya, (2010) A Paradigm-Based Finite State Morphological Analyzer for Marathi, Workshop on South Asian and South East Asian NLP (part of COLING 2010), Beijing, China

Raj Dabre, ArchanaAmberkar and Pushpak Bhattacharyya, (2013), A Way to Break Them All: A Compound Word Analyzerfor Marathi, ICON 2013, Noida, India, 18-20 December

Yamashita, Tatsuo & Matsumoto, Yuji. (2000). Language Independent Morphological Analysis. 232-238. 10.3115/974147.974179.