A Survey on Various Stemming Techniques for Hindi and Nepali Language

Biraj Upadhyaya, Kalpana Sharma, and Sandeep Gurung

Abstract Stemming is considered to be an important step in the field of natural language processing. Any word may be represented in various forms as per its use in a sentence or a paragraph. The various forms of the word are derived from one root word. This root word is called stem and the process of extracting the root word is known as stemming. As an example, words like singing and singer have the root word sing. The stemming technique particularly finds its use in the field of information retrieval where it contributes to efficient retrieval. The underlying algorithm used by various search engines like Google uses stemming for mapping a user's query which is put in various word forms to a base word or stem. This paper highlights several stemming techniques which have been developed for languages like Hindi and Nepali.

Keywords Stem · Root · Stemming · Information retrieval

1 Introduction

Stemming is the method by which the inflected or derived words are reduced to a stem, base or root form. The software which produces the stem or the root word is referred to as a stemmer [1]. A text in a natural language may contain several variants of the same word. These morphologically similar words have their meanings rooted in the stem or the root word [2, 3]. Today, several stemmers have been developed for many languages across the world, mainly for English and other European languages

B. Upadhyaya (✉) · K. Sharma · S. Gurung
Sikkim Manipal Institute of Technology, Sikkim Manipal University, Sikkim, India
e-mail: upadhayaya.biraj@smit.smu.edu.in

K. Sharma
e-mail: kalpana.s@smit.smu.edu.in

S. Gurung
e-mail: sandeep.gu@smit.smu.edu.in

S. Dhar et al. (eds.), *Advances in Communication, Devices and Networking*,
Lecture Notes in Electrical Engineering 776,
https://doi.org/10.1007/978-981-16-2911-2_14

as well [4]. However, not much work in the field of stemming has been carried out for South Asian languages like Hindi and Nepali [5].

Languages like Hindi and Nepali are written using a writing methodology known as Devanagari script which is considered to be a descendant of Brahmi script. It is written and read from left to right direction and there is no distinction between upper case and lower case letters [6]. The Hindi and the Nepali alphabets share many similar characteristics in the way they are written, as both are written using the Devanagari script. The Devanagari script is the fourth most widely adopted writing system in the world and is composed of 47 primary characters, including 14 vowels and 33 consonants [7]. In order to develop a stemmer for any language, it is very important to understand the word constructs pattern for that language. Nepali is an Indo-Aryan language written in Devanagari script. It follows a Subject + Object + Verb pattern in sentences which is different for languages like English. From the aspect of applicability, stemming techniques find their use in various information retrieval tasks used by various search engines like Google, Yahoo etc. The stemming techniques also find their use in finding domain vocabularies related to a particular domain of interest.

1.1 Types of Stemmer

Stemming algorithms can be broadly classified into the following three categories.

1.2 Rule-Based Stemming

It is a structural stemming approach that utilizes the structure of the words in a language as well as the morphological rules to identify the stem of each word. The rules are written based on the morphology of the language and its word derivation structure [8].

1.3 Statistical Stemming

Rule-based stemmers have the disadvantage of being reliant on a fixed set of rules for carrying out the stemming operation. Rule-based stemmers also require an adequate amount of language expertise [9]. On the other hand, statistical stemming approaches do not require language expertise and use statistical information from a large corpus of a given language to learn the morphology of words [10].

1.4 Hybrid Stemming

This approach combines the features of a rule-based stemmer and a statistical stemmer for stemming. Combining the features of both stemmer helps to increase the accuracy of the stemming algorithm [11].

2 Challenges in Stemming

Stemming algorithms are often challenged with two kinds of errors that occur frequently depending upon the nature of algorithms used by the stemmer. The types of resulting errors are given below.

2.1 Over-Stemming

Over-stemming happens when two words with different stems are derived from the same root. Over-stemming may also be known to be false-positive.

2.2 Under-Stemming

Under-stemming happens when two words that do not have separate stems are derived from the same root. It is possible to interpret under-stemming as false-negative.

3 History of Stemming

Initially, the only language where stemming was carried out was the English language. The first-ever algorithm [1] for stemming was proposed by Julie Beth Lovins in the year 1968 and published in the Journal of Mechanical Translation and Computational Linguistics. An inflected word is a result of a combination of the word with prefix or suffix or both. The stemmer was a rule-based stemmer basically aimed at extracting the word by suffix removal. The Lovins stemmer has 294 endings, 29 conditions and 35 transformation rules [1].

Inspired by the work of Lovins, an algorithm [2] for stemming was proposed by Martin Porter in the year 1980 and published in the journal named Program. This stemmer is the most widely used stemmer in the world and is the most cited paper on

stemming [7]. It is a de facto standard for stemming. The algorithm used was a rule-based approach. However, the stemmer had a lesser number of rules as compared to Lovins stemmer when it was derived.

The third significant work on stemming was carried by Christopher D Paice in the year 1990. The paper [3] was published in the proceedings of the conference on Special Interest Group on Information Retrieval. The stemmer was a rule-based stemmer with an added advantage of the inclusion of a subroutine for index compression in the algorithm. It was faster but produced a relatively large number of over-stemming errors [3] compared to the previous algorithms on stemming.

4 Stemming Techniques for Hindi Language

Various algorithms have been proposed for stemming based on the Hindi language. Ramanathan and Rao [4] proposed a lightweight stemmer for Hindi which uses handcrafted set of suffix list and looks for longest match stripping. They have used the name "Light stemming" as the algorithm is used for tripping of a small set of either prefixes or suffixes or both, without trying to deal with infixes, or recognize patterns and find roots. Out of 35,977 words used as input to the algorithm, 4.6% of words were found to be under-stemmed while 13.8% were found to be over-stemmed [4].

Pandey and Siddiqui [5] proposed an unsupervised Hindi stemmer with the aim of improving the combining various prefix and suffix rules based on heuristics. This paper focuses on the development of an unsupervised stemmer for Hindi and the evaluation of the approach using manually segmented words. The algorithm was evaluated on 1000 words randomly extracted words from the Hindi WordNet-1 database [12]. The training data has been constructed by extracting 106,403 words extracted from EMILLE (Enabling Minority Language Engineering) 2 corpus [13]. The observed accuracy was found to be 89.9% after applying some heuristic measures. The F-score was 94.96% [5].

Ganguly et al. (6) proposed two separate rule-based stemmers for the Bengali and Hindi languages. In this paper, linguistic knowledge was used to manually craft the rules for removing the commonly occurring plural suffixes for Hindi and Bengali. A baseline was fixed by choosing words randomly from websites of news articles written in Hindi on the web. The improvement obtained with the incorporation of new rules for stemming by handling some exceptional words which were not a part of the baseline was 5.03% [6].

Mishra and Prakash [7] proposed a stemmer named "Maulik" for the Hindi language. This stemmer is purely based on Devanagari script and it uses the hybrid approach which combination of brute force and suffix removal approach. A lookup table with 15,000 words was maintained in the database. The average accuracy of the stemmer was obtained to be 91.59% [7].

Anand et al. [8] proposed a semi-supervised approach for stemming text written in the Hindi language. This paper uses an algorithm to find the stem of a word in Hindi. The proposed algorithm uses word2vec, which is a semi-supervised learning

algorithm, for finding the 10 most similar words from a corpus. Also, a mathematical function is used to find the stem. The results are verified by selecting a set of 1000 Hindi words randomly taken from a corpus [8].

5 Stemming Techniques for the Nepali Language

Bal and Shrestha [9] proposed a morphological analyzer and a stemmer for the Nepali language. This earliest stemming technique did not handle words formed as a result of the combination of two free morphemes. This paper discusses the design and implementation issues as well as the linguistic aspects of a morphological analyzer and a stemmer for the Nepali language [9].

Sitaula [10] proposed a hybrid algorithm for stemming Nepali text. The hybrid Nepali stemming algorithm uses affix stripping in conjunction with a string similarity function and reports a recall rate of 72.1% on 1200 words. The handwritten rules comprised 150 suffixes and around 35 prefixes were considered. The accuracy of this hybrid algorithm is 70.10% [10].

Paul et al. [11] proposed an affix removal stemmer for the Nepali language. This work has a rule base of 120 suffixes and 25 prefixes and a root lexicon of over 1000 words and reports an overall accuracy of 90.48% [11].

Shrestha and Dhakal [14] proposed a new stemmer for the Nepali language and classifies suffixes into three categories and stem them according to different criteria. The proposed algorithm was implemented in Ruby and was tested in a data set of 5000 words, extracted from a corpus containing E-Kantipur news. The accuracy of the algorithm is obtained as 88.78% [14].

Koirala and Shakya [15] proposed a Nepali rule-based stemmer and analyzed its performance on different NLP applications. The corpus contained articles from various different areas, including news, sports, politics, literature etc. Corpus contained a total of 438 news articles with a total word count of 11,813 and a total unique word count of 11,346. Each news article, on average, contained 269 total words and 181 unique words. The F1 score was 0.79 [15].

References

1. Lovins JB (1968) Development of a stemming algorithm. J Mech Trans Comput Linguist, 22–31
2. Porter M (1980) An algorithm for suffix stripping Program. Program 14(3):130–137
3. Paice CD (1990) Another stemmer. Proc SIGIR Forum 24(3):56–61
4. Ramanathan A, Rao D (2003) A lightweight stemmer for hindi. Proceedings of workshop on computational linguistics for south asian languages, 10th conference of the European chapter of association of computational linguistcs., pp 42–48
5. Pandey AK, Siddiqui TJ (2008) An unsupervised hindi stemmer with heuristic improvements. In: Proceedings of the second workshop on analytics for noisy unstructured text data, vol 303, pp 99–105

6. Ganguly D, Leveling J, Jones GJF (2013) Rule-based stemmers for bengali and hindi. Post-Proceedings of the 4th and 5th workshops of the forum for information retrieval evaluation, New Delhi, India, December 4–6, 2013
7. Mishra U, Prakash C (2012) MAULIK: an effective stemmer for hindi language. Int J Comput Sci Eng, ISSN: 0975-3397, Vol 4 No 05 May 2012
8. Anand A, Chatterji S, Bhattacharya S (2019) Semi-supervised approach for hindi stemming. 8th International conference on natural language processing (NLP 2019), Vol 9, No 12, September 28–29, 2019, Copenhagen, Denmark
9. Bal BK, Shrestha P (2004) A morphological analyzer and a stemmer for Nepali. PAN Localization, working papers 2007, pp 324–31
10. Sitaula C (2013) A hybrid algorithm for stemming of Nepali text. Intelligent information management
11. Paul A, Dey A, Purkayastha BS (2014) An Affix Removal stemmer for natural language text in Nepali. Int J Comput Appl
12. https://catalog.ldc.upenn.edu/, LDC2008L02 Bhattacharyya, Pushpak, Prabhakar Pande, and Laxmi Lupu. Hindi WordNet LDC2008L02. Web Download. Philadelphia: Linguistic Data Consortium, 2008
13. http://www.emille.lancs.ac.uk/, The EMILLE (Enabling Minority Language Engineering) project, Lanchester University, United Kingdom (Website Last Accessed on 12.10.2020)
14. Ingroj Shrestha and Shreeya Singh Dhakal, "A new stemmer for Nepali language", International Conference on Advances in Computing and Communication, 2016
15. Pravesh Koirala and Aman Shakya, "A Nepali Rule Based Stemmer and its performance on different NLP applications", arXiv preprint arXiv:2002.09901, 2020 arxiv.org