I.J. Education and Management Engineering, **2018, 5, 11-17**
Published Online September 2018 in MECS (http://www.mecs-press.net)
DOI: 10.5815/ijeme.2018.05.02

Available online at http://www.mecs-press.net/ijeme

Bruteporter: A Hybrid Approach

Balamurugan Mahalingam, Kannan S, Vairaprakash Gurusamy

Madurai Kamaraj University, Tamilnadu, India

Received: 22 April 2017; Accepted: 11 September 2017; Published: 08 September 2018

Abstract

Stemming fetches the main root word from the inflectional words called stem. Stem gives different meaning when suffix or prefix is added to it. The stem need not give perfect meaning. Lemmatization gives lemma from inflectional words. Lemma should give meaning that in the dictionary form. Natural Language processing, Information retrieval, Text mining are the areas which use the stemming as preprocessing step. Through stemming, the size of the document can be reduced and ambiguity is also removed. It makes the work easy for other process likes information retrieval, semantic analysis, text categorization etc. Though there is a need for linguistic improvements in the existing stemming algorithms, all these algorithms fail in some cases to give an exact Root word and are not able to handle informal verbs. Hence, Bruteporter Hybrid approach is proposed in order to improve the linguistic process of stemming in English Texts. It combines the Wordnet and Modified Porter Algorithm. A Wordnet is a lexical dictionary created by linguistics people. Modified porter algorithm has both suffix removal and suffix substitution functionality. This proposed approach can extract root word from both inflectional words and informal verbs. In this paper, Experiment is conducted on proposed algorithm and the accuracy is calculated.

Index Terms: Porter, Inflection, Wordnet, Stemming.

1. Introduction

Stemming is a Preprocessing step to find the root word from the different word forms. That root word need not to give exact meaning. In Contrast, Lemmatization is the same process like Stemming since it produces the root word of inflectional words called lemma. But Lemma should consider exact dictionary meaning. Lovins (1968) proposed the first stemming algorithm called Lovins stemming. It has two phases named Stemming and Recoding. First words are stemmed by longest matching with the 294 suffixes with 29 rules and after that the

* Corresponding author
E-mail address: apkbala107@gmail.com, skannanmku@gmail.com, vairaprakashmca@gmail.com

resulted stem is checked with 35 rules whether the stem have to modify or not. Martin Porter (1980) developed the Porter algorithm which contains small suffixes. It is an iterative algorithm done with five steps process. Each step has linguistics rules for removing and modifying the suffix. Paice/Husk is also an iterative stemming algorithm created by Paice (1994). It uses the table of rules for stemming and replacing the suffix of inflectional word. Each rule started with ending letter of suffix of inflectional word and it has an appended string, symbol for whether to stop the stemming or go for next iteration and number for how many string has to be removed.

All the Existing stemming algorithms fail to give correct Root word from the inflectional words. They could not process the informal verbs. Some algorithms like Brute force stemming and Lemmatization can produce lemma with time consumption. The proposed approach combines the both modified porter algorithm and brute force algorithm to give the root word with less time consumption. It can handle both inflectional and informal verbs. Porter algorithm has to be modified to remove the suffix of continuous form, plural form and regular verb. It also contains some suffix substitution rules for correcting the incomplete root word. Brute force algorithm gives lemma for given inflectional word if it is available in wordnet database. Hybrid approach has two phases. In the First phase check the words if it is ends with ed/s/ing then apply the porter stemming to stem the word. Otherwise it goes as an input for the second phase to perform brute force stemming. The Algorithm gives root word for the input word if matches are found.

The structure of the paper is as follows: section 2 defines the terms used throughout article. Section 3 presents the related works to this paper done by various other authors. Section 4 describes the proposed work and its advantages. Section 5 contains evaluated result obtained from the experimental. Section 6 concludes the paper.

2. Terminologies

Stemming can be categorized into *Rule based Stemming* and *Statistics based Stemming*. In rule based stemming, Linguistics knowledge is needed to form the rules for stemming algorithm. In Statistics based stemming, there is no need to have linguistics knowledge to produce stemming algorithm. In Stemming process there may be a chance of *unterstemming error* or *Overstemming error* to occur. Unterstemming errors happen when the stemming algorithm produces different stems instead of same stems. Overstemming errors occur when the stems having same meaning instead of different semantic after stemming process. Stemming algorithm can remove either prefix or suffix letters of the inflectional word called *affix removal*. We differentiate the Affix removal stemming as *Inflectional* and *Derivational*. Inflectional affix removal, which is lemma of the inflectional word, is considered after stemming. Here we only remove the plural form, regular verb, and continuous form of word. In derivational removal, any form of the inflectional word has been removed after stemming process. Here stem may not be as lemma. Generally there are two kinds of cutting density available. First one is *Light Stemming*; it shortly cuts the inflectional word and it produces the too large stem which is an understemming error. These types of stemming algorithms are having only few numbers of rules and steps to be followed. Second one is *Heavy Stemming*; it contains large number of rules and largely cuts down the inflections words then produces the too short stem that is overstemming error. All the stemming algorithms may fall under either Light or heavy stemmer (Tomas Brychcin, Miloslav Konopik, 2015).

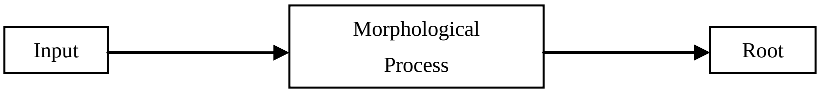

Fig.1. Overview of Morphological Process

In the Fig 1, Input word may be either inflectional word or derivational word. Morphological process represents both Stemming and Lemmatization. The resulted word may be either stem or lemma which is dependent on the morphological process.

3. Related Works

Dawid Weiss (2005) proposed Stempelator, the hybrid algorithm for polish language. The author combines the Heuristic stemmer (Stemper) and Dictionary-based stemmer (Lametyzator) to extract root word from the inflectional word. This algorithm performs two steps process. In the first step, stem input word is looked up in the dictionary if the lemma is found it is returned as stem. Otherwise perform the stemming by using stempel algorithm in the second step. This approach consumes too much time for finding root word to the inflectional word. Because it performs dictionary based approach first which takes too much time to match all the words in the dictionary until match is found. This was time consuming approach.

Kartik suba et al (2011) proposed the Hybrid inflectional stemmer for Gujarati language. It combines the Pos-based files and Linguistic rule based approach. The author creates his own pos files for the language. If the input word is in any pos based file, performing the pos based stemming returns root word. Otherwise apply the linguistics rule to the input word if the rule is exactly matches with input word. Here brute force is applied first so it is time consuming approach to match all the files.

Upendra Mishra et al (2012) Introduces the hybrid approach for Hindi Language. This approach combines the brute force approach and suffix removal approach to extract the root from the inflectional words. It first checks if the input word in the lookup table of the database by using brute force approach. Otherwise suffix removal algorithm performs the stemming on the input word. Finally root word is obtained from the hybrid approach. This is also a time consuming approach. Because it performs the brute force approach first which take too much time to find the root word.

Chandni Dhawan et al (2013) Proposes the hybrid approach for Punjabi Language. This author uses two tables for stemming. One is for root word tables and the other one is for suffix stripping and substituting. This approach combines the brute force algorithm and suffix stripping methodology. First Check the word if it is in the root table then display the root word otherwise apply the suffix stripping and suffix substitution on the word. Disadvantage is considered for time taken in matching process. Because it is done first after that only suffix stripping and substitution will be processed if the word is not in the root word table.

4. Proposed Approach

Our proposed approach has two main Approaches. The First one is suffix removal based approach that is processed by modified porter algorithm to remove suffix from the inflected word if it matches the rule. General porter algorithm contains five rules for suffix removal and recoding. It is a light weight algorithm that has conditions for remove the suffix of some inflectional stems. First it matches the suffix if it is matched then apply the condition on the remaining words to check whether it is satisfied or not. If it is satisfied then remove the suffix, otherwise the keep the word as it is. The same will be followed by Modified porter algorithm but it only focuses on removing Plural form, continuous form and Regular verb form of the inflectional word. It also checks the root word incompletion if it is incomplete then applies the correct suffix substitution. Fig 2 depicts the process of hybrid approach

The second one is wordnet based approach which using brute force method to retrieve the lemma from the inflectional word or derivational word. Princeton university wordnet is used as the dictionary for root word lookup. Wordnet has root word table collection. Brute force methodology is applied to find the root for inflectional or derivational word. When the input word is fed into wordnet, it could be processed by morphological program which is already available in the wordnet application. During the morphological process the input word matches with database of the word net. If the match is found then it will return the root word. The Pseudo code of our approach is mentioned below:

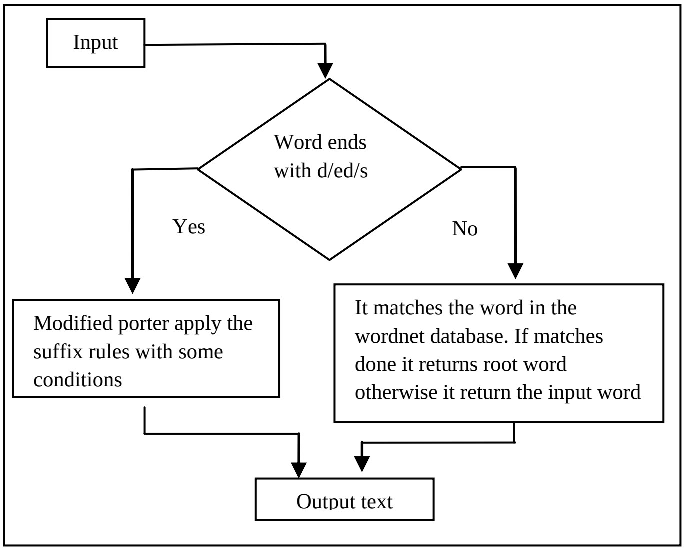
Input
Word ends with d/ed/s
Yes
No
Modified porter apply the suffix rules with some conditions
It matches the word in the wordnet database. If matches done it returns root word otherwise it return the input word
Output text

Fig.2. Overview of Hybrid Approach

4.1. Procedures

Step 1: Read the Word from the file

Step 2: Check the Word if it ends with d or ed or s or ing then remove its suffix by Using modified porter algorithm

Step 3: Otherwise apply the word net stemming for that word. (Look up that Morphological Word whether it is available in the wordnet collection or not).

Step 4: If it is available then return its corresponding root word otherwise return That word.

Our proposed approach has many advantages compare to other stemming and hybrid approaches.
They are listed out below:

- It could perform stemming and lemmatization as well.
- It could handle both inflectional words and informal verbs. Because wordnet database is used to handle the informal verbs.
- It performs suffix stripping and substitution first after that it performs wordnet matching process. So that It gives good performance and consumes less time for process.
- It gives more number of meaningful root words compare to other approach.

5. Experimental Results

This hybrid algorithm has been implemented in java. Princeton wordnet 2.1 is used as lookup table. Wordnet database is connected to java by using jwnl (java wordnet library) api. Porter algorithm is modified for removing only plural form, continuous form and regular verb form. From this setup, porter algorithm is designed to discard inflectional suffixes and wordnet stemming is used to extract the root word from informal verbs as well as inflectional words. Thousand words have tested by using hybrid approach. These words are combination of inflectional and informal words. After testing these words, the overall accuracy could be calculated. Accuracy shows the performance of stemming algorithm. It can be calculated as

Accuracy = (Number of words stemmed correctly/total number of input words) * 100

Table 1. Performance of Algorithms

Algorithm	Accuracy (%)
Porter	70.2
Paice/Husk	67
Lovins	62.3
Hybrid	93.4

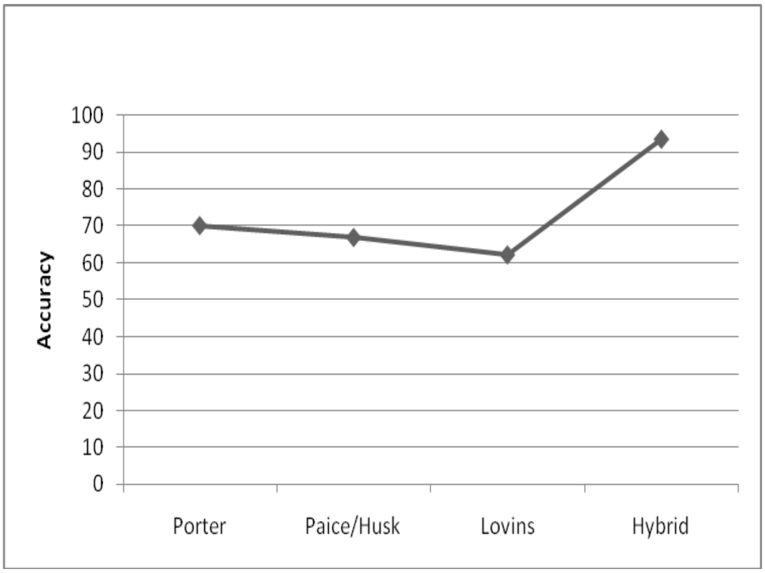

Fig.3. Effect of Different Approaches

Our proposed approach could find 934 meaningful root words and the remaining 66 words did not give meaning. Overall accuracy of this hybrid algorithm is (934/1000)*100 =93.4%. Here, 75 incorrect root words have done by modified porter algorithm. Fig 3 shows the performance of primary stemming algorithms and our approach. From the Table 1, Porter has the highest accuracy among primary stemming algorithms that is 70.2%. So we choose porter algorithm for hybrid approach to suffix stripping. Then probably stemming algorithm has less time consuming than lemmatization. Here we tested brute force algorithm on input words. It gives meaningful root words for all input words. But it takes 3490 milliseconds that are more number of times than stemming algorithms. Porter, Lovins and Paice/Hush only take 11, 44 and 14 milliseconds. Our approach consumes 532 milliseconds. From this, bruteporter approach performance is better than all stemming algorithms and less time consuming than bruteforce approach. Lot of hybrid approach for different languages was time consuming. Because they used brute force approach in first stage and suffix removal stemming for the second stage. These drawbacks have been overcome by the proposed hybrid approach. That is, here both suffix removal algorithm and wordnet are used for handling the inflectional words and derivational words. Suffix removal algorithm is only applied on inflectional words. Informal verbs are handled by wordnet stemming. Our proposed concept can able to save the meaning of root word after stemming with less time consumption. Here both stemming and lemmatization happening depends upon the context of the words.

6. Conclusion

In this article, we developed an efficient hybrid approach that combines the wordnet and modified porter algorithm to extract the root word from both inflectional and informal verbs. It achieves the 93.4 percent stemming performance that saves the meaning of the root word. It could perform the both stemming and lemmatization as well. Our approach has proved less time consumption than other hybrid approach and also gives better performance. In future we have planned to expand our approach for multilingual way that is applied on other languages as well. We will also try to include the additional rules for increase the accuracy of our approach.

Acknowledgement

This research was supported by Department of computer applications, Madurai kamaraj university. I thank my guide who provided insight and expertise that greatly assisted the research.

References

[1] Brychc ń, T., & Konop ḱ, M. (2015). HPS: High recision stemmer. *Information Processing & Management*, *51*(1), 68-91.

[2] Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, *14*(3), 130-137.

[3] Lovins, J. B. (1968). Development of a stemming algorithm.

[4] Paice, C. D. (1994, August). An evaluation method for stemming algorithms. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 42-50). Springer-Verlag New York, Inc..

[5] Weiss, D. (2005). Stempelator: A hybrid stemmer for the Polish language. Institute of Computing Science: Poznan University of Technology Research Report.

[6] Mishra, U., & Prakash, C. (2012). MAULIK: An effective stemmer for Hindi language. *International Journal on Computer Science and Engineering*, *4*(5), 711.

[7] Dhawan, C., Singh, J., & Garg, K. (2013). Hybrid Approach for Stemming in Punjabi. *International Journal of Computer Science & Communication Networks*, *3*(2), 101.

[8] Jiandani, K. S. D., & Bhattacharyya, P. (2011, November). Hybrid inflectional stemmer and rule-based derivational stemmer for gujarati. In *Proceedings of the 2nd Workshop on South and Southeast Asian Natural Language Processing (WSSANLP 2011)* (p. 1).

[9] Wiese, A., Ho, V., & Hill, E. (2011, September). A comparison of stemmers on source code identifiers for software search. In *Software Maintenance (ICSM), 2011 27th IEEE International Conference on* (pp. 496-499). IEEE.

[10] Moral, C., de Antonio, A., Imbert, R., & Ram ŕez, J. (2014). A survey of stemming algorithms in information retrieval. *Information Research: An International Electronic Journal*, *19*(1), n1.

[11] Flores, F. N., & Moreira, V. P. (2016). Assessing the impact of Stemming Accuracy on Information Retrieval–A multilingual perspective. *Information Processing & Management*, *52*(5), 840-854.

Authors' Profile

Balamurugan Mahalingam (born 27 August 1990) he is research scholar in Madurai Kamaraj University india. He is working as a assistant professor in SBK college, tamilnadu, india. His area of interest is Semantic analysis in Natural language processing.

How to cite this paper: Balamurugan Mahalingam, Kannan S, Vairaprakash Gurusamy,"Bruteporter: A Hybrid Approach", International Journal of Education and Management Engineering(IJEME), Vol.8, No.5, pp.11-17, 2018.DOI: 10.5815/ijeme.2018.05.02