IJCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Vol.-7, Issue-1, Jan 2019 E-ISSN: 2347-2693

A Study of Research Papers on Punjabi stemming with special reference to Brute Force approach

Harjit Singh

APS Neighbourhood Campus, Punjabi University Patiala, Punjab, India

**Corresponding Author: hjit@live.com, Tel.: +91-98551-79078*

Available online at: www.ijcseonline.org

Accepted: 24/Jan/2019, Published: 31/Jan/2019

***Abstract*:** For processing a natural language, the individual words of the text need to be processed. But the suffixes and prefixes attached to the words arises ambiguity problems and make the process more difficult. Stemming is the process of removing any suffixes or prefixes from the word to get the root word. Global languages like English took the advantage that much research has been done for stemming. Punjabi is an Asian language used by people of India's Punjab state and Pakistani Punjab state. Shahmukhi script is used in Pakistan and Gurumukhi script is used in India to write Punjabi. Although there are limited resources available for this language, yet some efforts have been made to develop stemmers for Punjabi language. This paper puts a light on those research papers which are published on Punjabi stemming with a special reference to Brute Force approach since it is used in almost all the research papers discussed here.

Keywords: Punjabi Stemmer, Brute Force Approach, Rule Based Stemming, Suffix Striping.

I. INTRODUCTION

A single root word can be used in a number of ways in a sentence by adding affixes to it. The suffixes and prefixes are attached to a word to make the sentence meaningful and grammatically correct. A grammar is a set of rules that need to be followed while writing properly formatted sentences in any language. But each language has its own grammar and hence has its own rules. Natural Language Processing (NLP) is a research area that deals with processing of natural languages so that the machines or computers can understand them in some way. All this is basically to make human-computer communication possible in natural languages.

For processing a natural language, the individual words of the text need to be processed. But the affixes attached to the root word create ambiguities and make the processing more cumbersome. So these affixes are removed to get root word and this procedure is called stemming. For stemming, various approaches are used such as Brute Force approach, Rule based approach, Statistical approach etc.

This paper will provide the basic information about almost all the research papers published on Punjabi stemming and it will motivate readers to develop more ideas in this area. It will contribute to explore the basic building blocks of Natural Language Processing and will be very helpful to new researchers for generating new ideas to develop more robust Punjabi language stemmers in future.

In this paper, Section I is the Introduction, Section II is about Brute Force Approach, Section III provides Review of Research papers published on Punjabi stemming and Section IV concludes the contents of paper.

II. BRUTE FORCE APPROACH

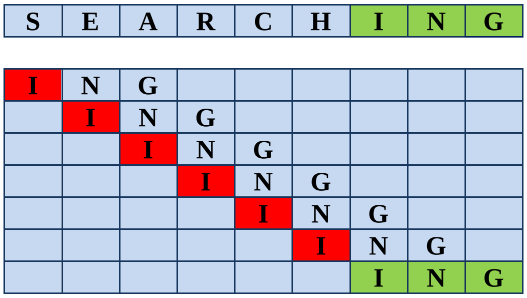
It is a simple approach of problem solving which works directly based on the statement of problem and its concepts. For example, Brute Force can be applied to calculate factorial of a number [1]. The definition of factorial states that factorial of n can be calculated as n x (n-1)! We use a loop to iterate and multiply the next number with factorial of previous number to get the factorial of next number.

A. Brute Force for String Matching

When Brute Force is used for string matching, we define a 'pattern' which is a substring of characters we need to search for. The 'text' is a search space consisting of longer string of characters in which we need to perform the search. The problem is to match the 'pattern' in the 'text' [1]. If 'pattern' has length m and 'text' has length n then, the Brute Force algorithm works something like:

```
Begin
For i=0 to n-m
        For j=0 to m-1
                If $text_{i+j}$ <> $pattern_j$ then
                        Break
                End if
        Next
        If j=m then
                Display “Match Found”
        Else
                Display “No Match”
        End if
Next
End
```

The ‘pattern’ is matched character by character with the initial substring of the ‘text’. Then substring position is shifted one character right to again match the ‘pattern’ with it. Shifting one character right in the ‘text’ is continued until a match occurs or all the characters in the ‘text’ are matched with ‘pattern’ but no match found. In figure-1, Brute Force is used to match the pattern ‘ING’ in the text ‘SEARCHING’.



S	E	A	R	C	H	I	N	G

I	N	G						
	I	N	G					
		I	N	G				
			I	N	G			
				I	N	G		
					I	N	G	
						I	N	G

Figure-1: Brute Force for String Matching

The variation of above algorithm which uses a table lookup is widely used by researchers in the research papers discussed in this paper.

III. Study of Research Papers

Dinesh Kumar et al. (2010) [2] in the paper “Design and Development of a Stemmer for Punjabi” proposed Punjabi language stemmer based on Brute Force Approach. They used a database of root words and their corresponding inflected words. The system searches for the word in the database using Brute Force search technique. If a matched word is found in the database, then the corresponding root word is given as output root word. If the match does not occur, the system creates the root word by simply removing the suffix from the inflated word. For suffix removal, the system uses a list of possible suffixes of Punjabi language Gurumukhi script. So it is a two step process to get the stem word from a Punjab inflected word. The first step finds the word already present in the database table and second step removes the ending part of the word to generate the stem word. If first step is successful in finding the right word then there is no need to execute the second step. The success of first step depends on the number of words stored in the database table.

Dinesh Kumar et al. (2011) [3] in the paper “Stemming of Punjabi Words by using Brute Force Technique” proposed Punjabi language stemmer which uses the same Brute Force technique. They stored the inflected words and their corresponding root words in a database. The system searches for the inflected word in the database through Brute Force search approach. If a matched inflected word is present in the database, then the associated root word is considered as output root word. If the system fails to find a match in inflected words, the system makes the root word by just removing the suffix from the end of the inflected word. For suffix removal, the system is given a list of possible suffixes used in Punjabi language Gurumukhi script. The system is divided into two modules. First module deals with brute force technique to search for the root word in database. It is prefect method to find the correct root word and is an efficient approach. But if the word to be searched in not in database then the second module works to generate the root word by removing the end part of the word. The result may be correct but it may also be incorrect.

Vishal Gupta et al. (2011) [4] in the paper “Punjabi Language Stemmer for Nouns and Proper Names” proposed a Punjabi language stemmer to find root words of those words which are not present in dictionaries. The words present in the dictionaries are language genuine words but the words which do not appear in dictionaries are either proper names or invalid words. The algorithm removes the suffixes and sometimes attaches another suffix to generate the root words. Proper names may be names of places, persons or concepts etc. A list of proper names and nouns is generated by the researchers to check the validity of final result by searching the list for a match. The system used a set of rules to find and remove a suffix from the end of the word to get root word. The system not only uses suffix removal technique but also uses suffix substitution technique to attach another suffix (if required) to get the correct root word.

Chandni Dhawan et al. (2013) [5] in the paper “Hybrid Approach for Stemming in Punjabi” proposed a stemming technique based on hybrid approach of using more than one approaches together. In this approach suffix stripping algorithm along with suffix substitution algorithm is used and to some extent Brute Force approach is also incorporated. This stemmer approach makes use of two database tables. One database table contains approximately 250000 Punjabi root words and another database table contains the list of suffixes to be removed and related suffixes to be substituted (if any). Initially, the word to be stemmed is searched in the database table of root words.

Only root words are available in the database, so it is tested whether the word to be stemmed is already a root word or not. If a match occurs, same word is the output root word. If that word is not found in the database table of root words then suffix stripping is applied to remove the suffix and suffix substitution is applied to append the suffix (if available). After suffix stripping and substitution, the resultant word is again searched in the database table of root words to make sure it is correctly stemmed. The problem of over-stemming and under stemming is overcome by ensuring the resultant word is present in the database table of root words. If a word is not found, it is not considered as correctly stemmed word.

Vishal Gupta (2014) [6] in the paper "Automatic Stemming of Words for Punjabi Language" proposed a method for complete automatic stemmer for Punjabi language words including nouns, pronouns, verbs, adverbs, adjectives and proper names. The paper provides separate lists for verb suffixes, adverb suffixes, adjective suffixes and pronoun suffixes. The algorithm removes the suffixes and sometimes attaches another suffix to generate the root words. Proper names may be names of places, persons or concepts etc. A Punjabi dictionary is generated by the researchers to check the validity of final result by searching the list for a match. If unable to find a match then an error message is displayed.

Garima Joshi et al. (2014) [7] in the paper "Enhanced version of Punjabi Stemmer using Synset" proposed a Punjabi stemmer based on hybrid approach of two major algorithms i.e. Table lookup and Rule based algorithms. Table lookup is actually the same Brute Force search algorithm. The stemmer makes use of Punjabi Synset to output the synonyms of stemmed word. Four database tables are used to make the stemmer work. Root Word table is a collection of 3500 root words of Punjabi language. Forms table is a collection of 2500 inflected words along with associated root words. These words are taken from Pardeep Punjabi to English Dictionary, National Punjabi Kosh Dr. Baldev Singh 'Badhan' and jagbani.com. Synonyms table is a collection of 8000 synonyms. The fourth table is Rules table which is a collection of rules to be applied for suffix stripping and suffix substitution. The Punjabi word to be stemmed is searched in Root Word table. If it is there, then the word is already a root word. If the word is not present in Root Word table then it is searched in inflected word stored in the Forms table. If a match occurs, the associated root word is returned. If the word is not in inflected word list, then Rule based approach is used. The word ending is matched with suffix rules of Rules table. If a match is found, that rule is applied for suffix removal and suffix substitution, whatsoever the rule specifies. The stemmed word is again searched in the Root Word table to ensure the correctness. The stemmer also returns the shortest length synonym of the stemmed word. Correctness of this stemmer is dependent on the number of words stored in the Root Word table, because the output root word of stemmer is finally checked for presence in Root Word table. The stemmer is effective to output only the correct root word. A stemmed word if not present in Root Word table is considered as incorrect word. If the input word is already present in the Root Word table, then the same word is returned as output root word. The performance of stemmer is calculated by dividing the number of correctly stemmed words to the total number of word inputted. Five persons were given 50 different words each to test the system.

Puneet Thapar (2014) [8] in the paper "A Hybrid Approach used to Stem Punjabi Words" proposed Punjabi language stemmer based on Naïve algorithm. He used a lookup table of inflected words and its corresponding root words. The system searches for the inflected word in the lookup table using Naive approach. If a matched inflected word is found in the database, then the associated root word is given as output root word. If the system fails to find a match in inflected words, the system makes the root word by just removing the suffix from the end of the inflected word. For suffix removal, the system is given a list of possible suffixes used in Punjabi language Gurumukhi script.

Rajeev Puri et al. (2015) [9] in the paper "Punjabi Stemmer using Punjabi Wordnet Database" proposed a revised stemmer which makes use of Punjabi Wordnet database. It is an improvement over the stemmer proposed by Vishal Gupta et al. (2011). The algorithm removes the suffixes and sometimes attaches another suffix to generate the root words. A Punjabi dictionary is generated by the researchers to check the validity of final result by searching the list for a match. If unable to find a match then an error message is displayed.

Abdul Mateen et al. (2017) [10] in the paper "A Hybrid Stemmer of Punjabi Shahmukhi Script" proposed a stemmer using hybrid technique for Punjabi Shahmukhi script. They used a table of words and their corresponding stem words. The system searches for the word to be stemmed in the lookup table. If a matched word is found in the table, then the associated stem word is given as output stemmed word. If the system fails to find a match in the table of words, the system makes the stem word by removing the suffix from the end of the word and if required appends another suffix to make the correct stem word. For suffix removal and substitution, the system is given a list of possible suffixes used in Punjabi language Shahmukhi script.

IV. CONCLUSION AND FUTURE SCOPE

The affixes are attached to a word to make the sentence grammatically correct, but these affixes attached to the words arise ambiguity problem and make the processing more cumbersome. Stemming is the process of removing any affixes from the word to get the root word. Although there are limited resources available for this Punjabi language, yet some efforts have been made to develop stemmers for Punjabi language. Through this paper it is tried to put some light on those research papers which has been published on Punjabi stemming with a special reference to Brute Force approach since it is used in all these research papers. From the discussion it is clear that approaches adopted by researchers do not vary so much from one to another and the research in this area is still in initial stage. This paper provided the basic information about almost all the research papers published on Punjabi stemming and it will motivate readers to develop more ideas in this area. It contributes to explore the basic building blocks of Natural Language Processing and will be very helpful to new researchers for generating new ideas to develop more robust Punjabi language stemmers in future.

REFERENCES

[1] Anany Levitin, "*Introduction to the Design & Analysis of Algorithms*", 2nd ed., Pearson Addison-Wesley, Chapter **3**, **2007**

[2] Dinesh Kumar, Prince Rana, "*Design and Development of a Stemmer for Punjabi*", International Journal of Computer Applications (ISSN: 0975 – 8887) Volume **11**– No.**12**, pp. **18-23**, December **2010**

[3] Dinesh Kumar, Prince Rana, "*Stemming of Punjabi Words by using Brute Force Technique*", International Journal of Engineering Science and Technology (IJEST) ISSN : 0975-5462, Vol. **3** No. **2**, **1351-1358**, Feb **2011**

[4] Vishal Gupta, Gurpreet Singh Lehal, "*Punjabi Language Stemmer for nouns and proper names*", Proceedings of the 2nd Workshop on South and Southeast Asian Natural Language Processing (WSSANLP), IJCNLP 2011, Chiang Mai, Thailand, pp. **35–39**, November 8**, 2011**

[5] Chandni Dhawan, Jashanpreet Singh, Kamaldeep Garg, *Hybrid Approach for Stemming in Punjabi*, International Journal of Computer Science & Communication Networks (ISSN:2249-5789),Vol **3(2)**, pp. **101-104**, **2013**

[6] Vishal Gupta, "*Automatic Stemming of Words for Punjabi Language*", Advances in Signal Processing and Intelligent Recognition Systems-Vol. 264, Springer International Publishing Switzerland, DOI: 10.1007/978-3-319-04960-1_7, pp. **73-84**, **2014**

[7] Garima Joshi, Kamal Deep Garg, "*Enhanced Version of Punjabi Stemmer Using Synset*", International Journal of Advanced Research in Computer Science and Software Engineering (ISSN: 2277-128X), Volume **4**, Issue **5**, pp. **1060-1065**, May **2014**

[8] Puneet Thapar, "*A Hybrid Approach used to Stem Punjabi Words*", International Journal of Computer Science and Mobile Computing (ISSN 2320–088X), Vol. **3**, Issue. **11**, pp.**1 – 9**, November **2014**

[9] Rajeev Puri, R. P. S. Bedi, Vishal Goyal, "*Punjabi Stemmer Using Punjabi WordNet Database*", Indian Journal of Science and Technology, Vol **8(27)**, DOI:10.17485/ijst/2015/v8i27/82943, pp. **1-5**, October **2015**

[10] Abdul Mateen, M. Kamran Malik, Zubair Nawaz, H. M. Danish, M. Hassan Siddiqui, Qaiser Abbas, "*A Hybrid Stemmer of Punjabi Shahmukhi Script*", International Journal of Computer Science and Network Security, Vol.**17** No.**8**, pp. **90-97**, August **2017**

Authors Profile

Mr. Harjit Singh received the MCA (Master in Computer Applications) degree from IGNOU (Indira Gandhi National Open University), New Delhi, India. Alongwith pursuing Post Graduation he worked as a Web Developer. He acquired M.Phil.(CS) degree alongwith working as a Web professional. Presently he is working as Assistant Professor (Senior Scale) in Computer Science at Punjabi University Neighbourhood Campus Dehla Seehan, Sangrur, India and is pursuing his Ph.D. degree. His current research interests include Natural Language Processing, Machine Translation, Artificial Intelligence.