Volume 4, Issue 5, May 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Enhanced Version of Punjabi Stemmer Using Synset

Garima Joshi
Lovely Professional University, (CSE Dept.)
Phagwara- 144411Punjab

Kamal Deep Garg
Assistant Professor (CSE Dept.) Lovely Professional University, Phagwara-144411 , Punjab

Abstract-Stemming is the process of automatic removal of affixes from the word without doing complete morphological analysis. In this process the word having the same stem are reduced to the common form. Stemming is the very first phase of any information retrieval task such as Text Summarization, Word sense disambiguation etc. Also find its use in the search engine optimization so as to reduce the query processing time. The paper presents the Enhanced Punjabi Stemmer which is based on hybridization of the two major algorithms used in Punjabi stemmer so far that are Look up table Algorithm and Rule based algorithm for suffix removal. Synset approach is also incorporated in the stemmer so as to return the list of words that share the same meaning with the valid stem word. A large database will be used so as to improve the accuracy level of the Stemmer

Index Terms - Stemmer, synset ,Disambiguation, Suffix Removal, Punjabi Stemmer

I. INTRODUCTION

Stemming is defined as the process of reducing an Inflected word to its stem, base or root form . . Any Natural Language Processing system requires a stemmer at the very first stage. The basic goal of any stemmer is to standardize the words by reducing it to the base word. Stemming reduces inflected words to their root forms which are referred as stems for ex stemmer, stemming, stemmers are all conflated to single root word stem. Stemmer is available in many languages like English, French, and Arabian and in last few years has been successfully developed for many Indian languages like Punjabi, Hindi, Marathi, Bengali etc. The first paper on the stemmer was published in 1968 which was written by Julie Beth Lovins. A later stemmer was written by Martin Porter which was published in the July 1980. (Willett, P. (2006). This stemmer was very widely used and became the standard algorithm used for English stemming. Most of the stemming algorithms fall in categories of affix removal algorithms, statistical and mixed algorithms. Affix removal stemmers apply set of to each word, so as to remove the known prefixes or suffixes. The first such algorithm was given by J.B. Lovins in 1968. Later some more affix removal algorithms have been suggested. Porter's algorithm published in 1980 was the most frequently used algorithm and the stemming framework Snowball was also developed by Porter. Stemmers can be broadly classified in two types:

- Language Dependent Stemmers
- Language Independent Stemmers

Language Dependent Stemmers: Stemmers that are language dependent are made for a specific language. They are applicable to specific language for which it is designed.
For e.g.: MAULIK which is an effective stemmer for Hindi language. MAULIK is a stemmer designed for only Hindi language so it is language dependent stemmer**.**
Language Independent Stemmers: Language independent stemmers are those which do not depend on a specific language. Language independent stemmers are designed for all languages i.e. it can do stemming any language
For e.g.: Successor variety algorithms are language independent

II. BACKGROUND AND RELATED WORK

The earliest English stemmer was developed by Julie Beth Lovins in 1968. The Porter stemming algorithm (Martin Porter, 1980), which was published later, is the most widely used algorithm for English stemming. These stemmers are rule based and are best suited for less inflectional languages like English.
(Mudassar M. Majgaonker, 2010) proposed and evaluated a rule based stemmer for Marathi language. The rule based approach uses set of suffix removal rules along with an approach which is unsupervised based on n gram splitting approach which automatically learns suffixes from extracted words of Marathi text. The maximum accuracy achieved for this stemmer is 82.5%. [6]

(Vishal Gupta et al.,2011) proposed a basic Punjabi stemmer for proper nouns and proper names. The process of stemming is based on a rule based approach where various different rules are defined depending on the suffixes. When a suffix of the word matches the suffix rule already defined the corresponding rule will be fired and accordingly the suffix will be removed and then substituted if required to obtain the stem word.[10]

(Upendra Mishra, 2012) proposed an effective stemmer for Hindi named as "Maulik". This stemmer is based on the Hybrid approach combining the suffix removal algorithm and the Brute force algorithm in order to assist the task of Information retrieval. [7]

© 2014, IJARCSSE All Rights Reserved

(*Monika Dogra et al.,2013)* proposed and effective stemmer for Devnagri Script. The research work aims at developing Hindi stemmer which is based on rule based approach combined with look up tables in order to strip both suffixes as well as prefixes to derive a stem word.[4]

(*Dr.M.Thangarasu et al. ,2013)* gives a review of the Stemming algorithms of Indian languages. The paper summarizes different stemmers like Kannada morphological analyser and generator by Shambhavi et al., lightweight stemmer for Hindi developed by Ramanathan (2004) et. al, Malayalam stemmer for information retrieval which was developed in 2010 by Vijay Sundar et. al.[5]

III. PUNJABI LANGUAGE DESCRIPTION

Punjabi language is very different from other languages in terms of its grammatical properties and phonetic rules. India .Punjabi is also called as Gurmukhi or Shahmukhi .It was developed in 16th century by first Sikh Guru Shri Guru Nanak DevJi. Gurmukhi means "from the mouth of Guru". Shahmukhi is not spoken in Punjab and majorly spoken in Pakistan and written in Arabic text. Gurmukhi language contains 35 distinct letters. The very first three letters are unique as they form the basis for the vowels. [9]

TABLE I UNIQUE LETTERS

ੳ Ura	ਅ Era	ੲ Iri

ਸ Sussa *Sa*	ਹ Haha *Ha*	ਕ Kukka *Ka*	ਖ Khukha *Kha*	ਗ Gugga *Ga*	ਘ Ghugga *Gha*
ਙ Ungga *Nga*	ਚ Chucha *Ca*	ਛ Chhuchha *Cha*	ਜ Jujja *Ja*	ਝ Jhujja *Jha*	ਞ Yanza *Nya*
ਟ Tainka *Tta*	ਠ Thutha *Ttha*	ਡ Dudda *Dda*	ਢ Dhudda *Ddha*	ਣ Nahnha *Nna*	ਤ Tutta *Ta*
ਥ Thutha *Tha*	ਦ Duda *Da*	ਧ Dhuda *Dha*	ਨ Nunna *Na*	ਪ Puppa *Pa*	ਫ Phupha *Pha*
ਬ Bubba *Ba*	ਭ Bhubba *Bha*	ਮ Mumma *Ma*	ਯ Yaiyya *Ya*	ਰ Rara *Ra*	ਲ Lulla *La*
ਵ Vava *Va*	ੜ Rahrha *Rra*				

TABLE II CONSONANTS

◌ Mukta *a*	◌ਾ Kanna *aa*	◌ਿ Sihari *i*	◌ੀ Bihari *ii*	◌ੇ Lavan *ee*	◌ੈ Dulavan *ai*
◌ੁ Onkar *u*	◌ੂ Dulankar *uu*	◌ੋ Hora *oo*	◌ੌ Kanaura *au*		

TABLE III DEPENDENT VOWELS

TABLE IV INDEPENDENT VOWELS

ਅ *a*	ਆ *aa*	ਇ *i*	ਈ *ii*	ਏ *ee*	ਐ *ai*
ਉ *u*	ਊ *uu*	ਓ *oo*	ਔ *au*		

© 2014, IJARCSSE All Rights Reserved

TABLE V LIST OF PUNJABI SUFFIXES

◌ੀਆਂ īāṃ	◌ਿਆਂ iāṃ	◌ੂਆਂ ūāṃ	◌ਾਂ āṃ
◌ੀਏ īe	◌ੇ ē	◌ੀਓ Īō	◌ਿਓ iō
◌ੇ ō	◌ੀਆ īā	◌ਿਆ Iā	◌ੀਂ īṃ
◌ੀ ī	◌ੋਂ ōṃ	◌ਾਂ vāṃ	◌ਿਉਂ iuṃ
◌ੀਆ īā	ਜ/ਜ਼/ਸ ja/z/s		

Punjabi language has some rules in order to spell a word few of which are listed below [1]

- Kanna, Dulaanv and Kanorha can be used with any consonant except ੳ and ੲ.
- Sihari, Laanv and Bihari can be used with any consonant except ੳ and ਅ.
- Onkarh, Dulankarh and Horha can be used with any consonant except ਅ and ੲ.
- Onkarh cannot come at the end of word except the last consonant of the word used in ੳ and ਸ.
- The consonant ਯ could not be the last character of word.
- The Bindi and Tippi cannot be used with nasal consonant i.e. among ਞ, ਣ when is (nasal consonant) is the last character of the word.
- The Adhak is used only with short vowels and one long vowel namely Dulaanv. It cannot be used with nasal consonant.

IV. PRESENT WORK

We have created three tables in the database Brute Force Table named 'forms' that contains commonly occurring inflated words along with their corresponding stem words , Root Word table that contains all the valid Punjabi root words and a Synonyms table that contains the synonyms of the root words. at present 3500 words are added to the root words table, 2500 words added to 'forms table and 8000 synonyms are added to the Synonyms table. We have also created set of rules based on commonly occurring suffixes in Punjabi words. Rules table which contains the list of rules for suffix removal as well as suffix substitution. If the ending of the rule matched with the ending in the rule table then two actions are taken 'S' and 'R'. Action S stands for substitution and R stands for Removal. So Suffix substitution and suffix removal rules are added to the system. Few of the rules are

- If the word ends with ◌ੀਆਂ Then remove ਆਂ from the end. *(Suffix removal)*
- If the word ends with ◌ੇ Then remove ◌ੇ from end and add ◌ਾਂ at the end.(*Suffix substitution)*

Steps of the algorithm are listed below:

1. The input Punjabi word is checked against the database of the root words to check if the entered word is already a root word or not.

2. If the word is found the database of the root words then the message will be displayed to the user that the entered word is a stem word

3. If the match is not found in the database of the root word then the word will be searched in the Brute force table which consist of inflated words

4. If the word is found in the Brute Force table then the corresponding Root word will be displayed as the output.

5. If the match is not found in The Brute Force table then control passes to the Rule based approach. In this the ending of the word will be matched with various Rules present in the Rules Table. Wherever the match occurs that rule will be fired and the action corresponding to the rule will be taken. There can be two actions Substitution and Removal. Substitution will be done when the suffix is first removed and a new suffix is added at the end .Removal is applicable when only the suffix is removed and the stem word is obtained.

6. The resulting stem word will be once again checked against the database of the root word. If the stemmed word is now found in the database of the root words then the stemmed word is the correct stem word otherwise the output will be an incorrect word

7. For the resulting stem word the shortest synonym corresponding to that stem word is displayed as the preferred synonym

8. The user can also click to view all the synonyms of the word.

© 2014, IJARCSSE All Rights Reserved

9. The stem word along with the synset is displayed on the output screen
10. At the interface level the Punjabi keyboard is provided on the screen so as to give an ease to the user to type the Punjabi word by simply clicking on the buttons.

This algorithm has been implemented in ASP.Net using C# language. The database is built in MS Access. Three different tables were constructed one was for storing the root words called WORDS , one table for storing the inflated words used in Brute force called FORMS and one table for storing the synonyms corresponding to the root words. There was one more table constructed for the rules in order to perform the suffix removal and suffix substitution. Action specified for the Removal is "R" and action specified for substitution is "S" Every rule has an action corresponding to it. When action is R the ending is removed and when action is S then the ending is first removed and then appended with a new ending.

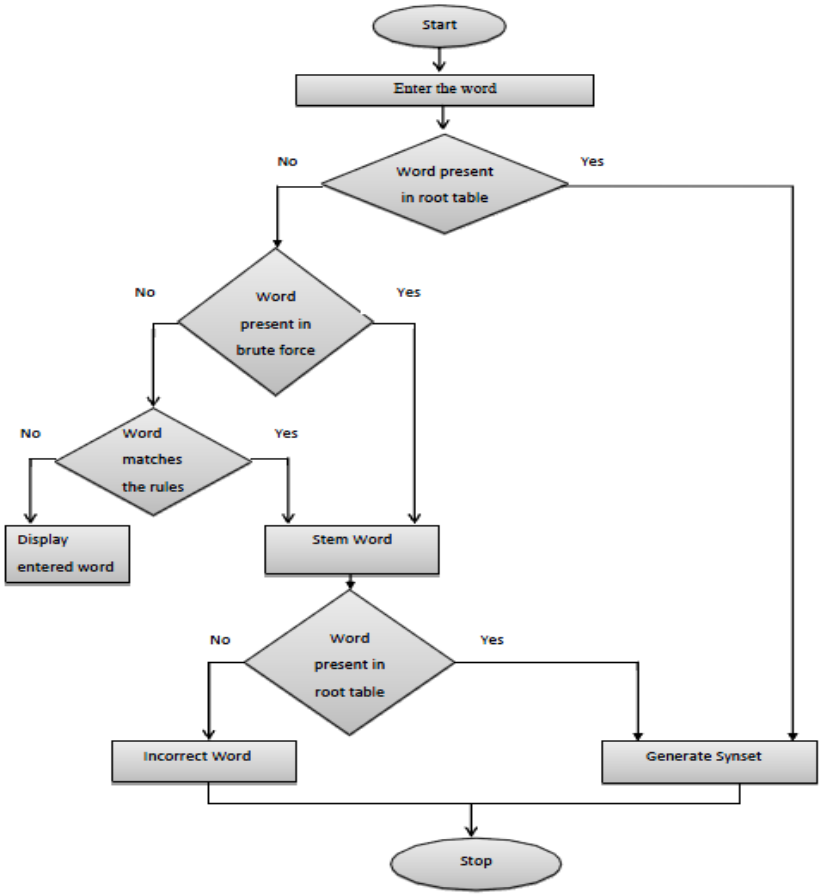

Figure 1. Flowchart

The table below lists some of the stemmed words obtained after stemming. The resulting stemmed word is then checked against the list of all its possible synonyms and the one with the minimum length is displayed as the preferred synonym. In this way the length of the inflated word is further reduced i.e. if entered word is of 8 characters long after stemming it reduces to say 5 characters then further the length can be reduced by using the synonym of the stemmed word which can be 3 or 4 characters. Such examples are listed below.

TABLE VI. STEMMED WORDS AND SYNONYMS

Entered word	**Stem word(length)**	**Preferred Synonym(length)**
ਕਮਜ਼ੋਰੀ	ਕਮਜ਼ੋਰ(5)	ਮਾੜਾ(4)
ਕਰਾਮਾਤਾਂ	ਕਰਾਮਾਤ(6)	ਕਮਾਲ(4)
ਕਰਿਸ਼ਮਿਆਂ	ਕਰਿਸ਼ਮਾ(6)	ਕਮਾਲ(4)
ਕਰਿੰਦੇ	ਕਰਿੰਦਾ(6)	ਸੇਵਕ(4)
ਕਰੀਬੀ	ਕਰੀਬ(4)	ਕੋਲ(3)
ਕਵਿਤਾਵਾਂ	ਕਵਿਤਾ(5)	ਰਚਨਾ(4)
ਕਾਇਦੇ	ਕਾਇਦਾ(6)	ਨਿਯਮ(4)
ਕਿਸ਼ਤੀਆਂ	ਕਿਸ਼ਤੀ(5)	ਬੇੜੀ(4)

© 2014, IJARCSSE All Rights Reserved

ਕਿਸਮਤਾਂ	ਕਿਸਮਤ(5)	ਭਾਗ(3)
ਕੁਕਰਮਾਂ	ਕੁਕਰਮ(5)	ਐਬ(3)
ਕੁਨਬੇ	ਕੁਨਬਾ(5)	ਵੰਸ(3)
ਕੁਰਬਾਨੀ	ਕੁਰਬਾਨ(6)	ਸ਼ਹੀਦ(4)
ਹਮਜੋਲੀਆਂ	ਹਮਜੋਲੀ(6)	ਹਾਣੀ(4)
ਹੁਨਰਮੰਦ	ਹੁਨਰ(4)	ਕਲਾ(3)

V. RESULTS AND DISCUSSIONS

For evaluation of proposed stemmer following parameters are used:

- Correctness of word stemmed
- Effectiveness of Punjabi stemmer
- Performance of Punjabi stemmer

Correctness of stemmer directly depends on the number of words present in the root word table. I have entered approximately 3500 root words in the root word table so its correctness depend on it because even after doing suffix stripping and suffix substitution the word is to be checked in the root word table for its correctness. . All these root words and synonyms are collected from various sources mainly from "ਨੈਸ਼ਨਲ ਪੰਜਾਬੀ ਕੋਸ਼ ਡਾ. ਬਲਦੇਵ ਸਿੰਘ 'ਬੱਦਨ' " Pradeep Punjabi to English Dictionary, www.jagbani.com.

Effectiveness of stemmer is concerned with the behavior of stemmer under abnormal conditions. Under abnormal conditions like when the user enters some word which is not available in the root word table the stemmer is still effective because it will display the message 'Incorrect Word' entered by the user as an output. when the user enters the Stem word itself the stemmer is still effective and displays Entered word is a root word and the system still displays the synonym for the stem word.

Performance of the stemmer can be calculated by the given formula:

Performance=number of words correctly stemmed by the stemmer/number of words entered by the user.

TABLE VII RESULTS

Sr. No.	Words Entered	Correct Stem Words	Percentage
Person 1	50	48	96%
Person 2	50	46	92%
Person 3	50	47	94%
Person 4	50	50	100%
Person 5	50	48	96%

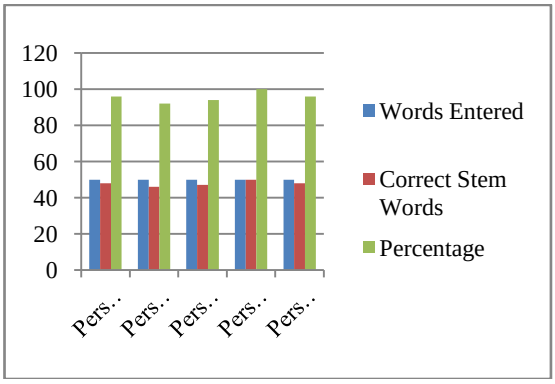

Figure 2. Results Obtained

The system is continually being tested by entering as many words as possible. So far the system has been tested by 250 words giving an average accuracy of 95.6% which is an improvement to the previous stemmer which has an accuracy of 92%.

Few words for which the system did not give correct output are listed here

- ਸੰਵੇਦਨਸ਼ੀਲ does not get stemmed to ਸੰਵੇਦਨ however it gives output as incorrect word.
- ਕਥਾਵਾਚਕ does not get stemmed to ਕਥਾ however it gives output as incorrect word.
- ਸੜੀਅਲ does not get stemmed to ਸੜ੍ਹ however it gives output as incorrect word.
- ਕੱਦਕਾਠ does not get stemmed to ਕੱਦ however it gives output as incorrect word.
- ਕਦਰਦਾਨ does not get stemmed to ਕਦਰ however it gives output as incorrect word.
- ਕਲਪਨਾਤਮਕ does not get stemmed to ਕਲਪਨਾ however it gives output as incorrect word.
- ਕਿਸ਼ੋਰਅਵਸਥਾ does not get stemmed to ਕਿਸ਼ੋਰ however it gives output as incorrect word.

© 2014, IJARCSSE All Rights Reserved

- ਕਾਬਲੀਅਤ does not get stemmed to ਕਾਬਲ however it gives output as incorrect word.

So far these words have not generated the correct output .The system is still being tested as the database of the system is vast. These are the words on which no Rule is applicable and neither these words are added to the Brute force Table. That is why the stem word is not generated and the output is displayed as incorrect word but the word is a valid Punjabi word otherwise.

VI . CONCLUSION AND FUTURE SCOPE

The proposed stemmer is for Punjabi language. Hybrid approach using synset is used for stemming in Punjabi which is combination of brute force approach, Rule Based Approach and Synset approach. Table lookup has benefit of giving accurate results. Due to suffix stripping approach the problem of over-stemming and under-stemming always occur, but in proposed stemmer both of these problems are dealt. This is possible by adding the step of suffix substitution after suffix stripping and at last performing table lookup to ensure that the word is correct root word or not. The performance of stemmer is directly dependent on number of entries in the root word table The size of the stem word can further be reduced by finding the shortest synonym of the stem word.

Table lookup approach can be improved by adding more entries in the database and adding more suffixes in the suffix list. More number of inflated words can be added to the Brute force table so that the words that cannot be stemmed using rules can be stemmed directly through Brute Force. The system can be improved by using the semantic analysis so the shortest synonym can be returned without changing the context in which the word is used. The same system can be embedded in Punjabi Text Summarization in order to improve its accuracy level

VII. ACKNOWLEDGEMENT

I would like to place on record my deep sense of gratitude to Assistant Prof. Kamaldeep Garg of Computer Science Engineering, LPU, and Jalandhar, India for his generous guidance, help and useful suggestions. His assistance was very much beneficial for me to carry out the process of research in this field. Key improvements in the proposed research work would not be possible without the valuable suggestion and the feedback of my mentor. It would not have been possible without the kind support and help of many individuals and organization. I would like to extend my sincere thanks to all of them.

I would also like to thank to Lovely Professional University for the support on academic studies and letting me involve in this study.

REFERENCES

[1] AartiTayal ,Dharamveer Sharma (2011)" *Punjabi Thesaurus- A tool for Natural Language Processing"* RJCSE Vol. 02, Issue 02 June ,2011 pp.100-103

[2] Dr. Baldev Singh Baddn (2001) *National Punjabi Kosh* , National Publishers Ltd., New Delhi.

[3] Dinesh Kumar , Prince Rana "*Stemming Of Punjabi Words Using Brute Force Technique"* International Journal of Engineering Science and Technology (IJEST) Vol. 3 No. 2 Feb 2011 pp. 1351-1358

[4] Monika Dogra, AbhishekTyagi, Upendra Mishra (2013) *"An effective Stemmer in Devnagri Script"* RTCCE 2013 pp. 22-25.

[5] M.Thangarasu, Dr.R.Manavalan (2013) "*A Literature Review: Stemming Algorithms For Indian Languages"* IJCTT- Volume 4 Issue 8 – August 2013 pp. 2582-2584

[6] Mudassar M. Majgaonker et al. "*Discovering suffixes: A case study for Marathi Language"* (IJCSE) International Journal on Computer Science and Engineering Vol. 02, No. 08, 2010, 2716-2720

[7] Upendra Mishra , Chandra Prakash(2012) *"Maulik an Effective Stemmer fr Hindi Language"* International Journal on Computer Science and Engineering (IJCSE) Vol. 4 No. 05 May 2012 pp.711-717

[8] Gupta , Gurpreet Singh Lehal (2013)" *A Survey of Common Stemming Techniques and Existing Stemmers for Indian Languages"* Journal Of Emerging Technologies In Web Intelligence, Vol. 5, No. 2, May 2013 pp 157-161

[9] Vishal Gupta, Gurpreet Singh Lehal "*Pre-processing Phase of Punjabi Language Text Summarization"*

[10] Vishal Gupta and Gurpreet Lehal (2011) "*Punjabi Language stemmer for nouns and proper names"*(WSSANLP), IJCNLP pp 35-39 Chiang Mai, Thailand November 8 2011.

[11] Willet P. (2006)*"The Porter Stemming Algorithm: then and now"* Program: electronic library and information systems, 40 (3). pp. 219-223

© 2014, IJARCSSE All Rights Reserved