Design and Development of a Stemmer for Punjabi

Dinesh Kumar
Assistant Prof. & Head
Department of Information Technology
DAVIET, Jalandhar

Prince Rana
Student – M.Tech
Department of Computer Science
DAVIET, Jalandhar

ABSTRACT

Stemming is the process of removing the affixes from inflected words, without doing complete morphological analysis. A stemming Algorithm is a procedure to reduce all words with the same stem to a common form [20]. It is useful in many areas of computational linguistics and information-retrieval work. This technique is used by the various search engines to find the best solution for a problem. The algorithm is a basic building block for the stemmer. Stemmer is basically used in information retrieval system to improve the performance .The paper present a stemmer for Punjabi, which uses a brute force algorithm. We also use a suffix stripping technique in our paper. Similar techniques can be used to make stemmer for other languages such as Hindi, Bengali and Marathi. The result of stemmer is good and it can be effective in information retrieval system. This stemmer also reduces the problem of over-stemming and under-stemming.

Keywords

Stemmer, Stemming, Brute Force Algorithm, Suffix Striping, Under-stemming, Over-stemming, Stemming Algorithm.

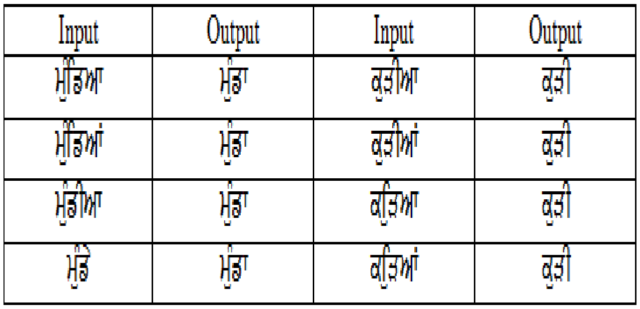
1. INTRODUCTION

Stemming is the process for reducing Inflected words to their stem, base or root form The stem need not be identical to the morphological root of the word; is usual sufficient that related words map the same stem, even if this stem is not in itself a valid root [20]. The first paper on the stemmer was published in 1968.It was written by Julie Beth Lovins [1]. A later stemmer was written by Martin Porter and was published in the July 1980 issue of the journal Program [2]. This stemmer was very widely used and became the de-facto standard algorithm used for English stemming.The process of stemming is also called conflation.

Stemming is also used in indexing and search system. This will handle automatic removal of word endings. Stemming is usually done by removing any attached suffixes, and prefixes from words. Stemmer uses number of techniques like Brute force algorithm, Suffix Stripping algorithm, Lemmatization Algorithm, Stochastic Algorithm, N gram analysis, Hybrid approaches. These stemming algorithms are different in performance and accuracy. If the algorithm is good then output of the stemmer will be good. These all the techniques have their own way of stemming a word. Each has their own pros and cons. Different stemming algorithm is used as per the requirement of the application and the user requirements. These algorithms use different approach and different method to stem a word.

A stemming algorithm reduces the words "asking", "asked" and "asks" to the root word "ask".Example of Stemming also shown in the table 1. Here we have taken two Punjabi words as an example.

Table 1 Stemming examples



Input	Output	Input	Output
ਮੁੰਡਿਆ	ਮੁੰਡਾ	ਕੁੜੀਆ	ਕੁੜੀ
ਮੁੰਡਿਆਂ	ਮੁੰਡਾ	ਕੁੜੀਆਂ	ਕੁੜੀ
ਮੁੰਡੀਆ	ਮੁੰਡਾ	ਕੁੜਿਆ	ਕੁੜੀ
ਮੁੰਡੇ	ਮੁੰਡਾ	ਕੁੜਿਆਂ	ਕੁੜੀ

In the above example it shows that if the word is written in any format then the result of our stemmer always is the same.

2. RELATED WORK

Research in Punjabi Stemming is not done so far as compared to the other languages. The stemmer for other languages like English, Nepali, Bengali and Hindi are present. Mostly the word is done on English language. Algorithm for suffix stripping is used in 1980 by M.F Porter. In this it uses a list of suffixes by which it matches an inflected word and removes the suffix [2]. Stemming algorithm is used for German languages. In this stemmer firstly it removes a suffix from the word and then checks the validity of word. If the word found to be illogical then it substitutes the suffix with the other words [4]. In the Dutch stemmer it uses a suffix stripping algorithm and dictionary lookup rule based methods [5].In the Nepali Stemming it uses a morphological analyzer which determines the given inflected word .In this it also tells about the Dawson stemming algorithm, krowertz algorithm [6].Lightweight Stemmer for Bengali also exists. In which it just strips the affix from the word without doing the complete morphological analysis. It removes suffixes as well as prefixes. This type of approach that is used for stemming is also called affix removal approach [7]. In the lightweight stemmer for Hindi it uses a look up table approach in which word is matched with the words present in the lookup table. Light weight stemmer approach uses affix removal algorithm and n gram stemming algorithm. It also shows the over stemming errors and the under stemming errors [13].There is a hybrid approach which is used for stemming of Arabic text. In this approach it uses a dictionary technique, morphological analysis, affix removal, statistical and translation technique. It also shows the accuracy of this hybrid approach on various areas like economics, science, medical and sport [14].

3. DIFFERENT APPROACHES TO STEMMING

There are different approaches used for stemming. Some of the approaches are explained below

- *Suffix stripping algorithms* do not rely on a lookup table that consists of inflected forms and root form relations. Instead, a smaller list of "rules" is stored which provide a path for the algorithm, given an input word form, to find its root form. Some examples of the rules include: if the word ends in 'es', remove the 'es' .if the word ends in 'ing', remove the 'ing' .Suffix stripping approaches enjoy the benefit of being much simpler to maintain than brute force algorithms.[2] But here over-stemming and under-stemming comes "Over-stemming" occurs when words that are not morphological variants are conflated. For example, in English, if the words compile and compute are both stemmed to comp, it is a case of over-stemming. "Under-stemming" occurs when words that are indeed morphological variants are not conflated. An example of under-stemming, in English, would be: compile being stemmed to comp, and compiling to compil [13].
- *Suffix substitution* is an improvement over suffix stripping. There are number of rules in which suffix is replaces with some other suffix. E.g. 'ies' is replaced with 'ey'. If we have word friendlies then with the help of suffix substitution we can change this word to friendly.
- *Lemmatization algorithm* is another technique to stem a word. This firstly determines a part of speech of a word and applies normalization rules on each part of speech. According to these rules stemming is done.
- *Stochastic algorithms* use probability to identify the root form of a word. These algorithms are trained to develop a probabilistic model. This model is form of complex linguistic rules. Stemming is performed by inputting an inflected form to the trained model and having the model produce the root form according to its internal rule set, which again is similar to suffix stripping and lemmatization [20].
- *Hybrid approaches* are those in which we are using more than one approach for stemming. E.g. if we are using suffix stripping technique and suffix substitution technique together.
- *Affix approaches* are those which are used to remove prefix as well as suffix from the word. E.g. 'indefinitely' is a word in which 'in' is a prefix.
- *N-gram analysis* is an approach which is also used for stemming. Gram is a unit of measurement of text that may refer to a single character, a single syllable, or a single word [20]. Here sequence of two grams is called bigram and sequence of three grams is called trigram. In this approach it uses a table of bigram and trigram words to find the result.
- *Iteration and Longest match algorithms* are used in stemming. In the iteration approach there is an order for classes for suffixes. The last order class which occurs at the end of the word contains suffixes. Iteration algorithms use recursive procedure. This approach removes the string in order class, starts from the endings and then moves towards the beginning. There may be derivational or inflectional suffixes. Here it is us to us to select the number of order classes. In the longest match principle here ending of the word is matched with the list of suffixes. If more than one ending provides a match then the longest match should be removed. E.g. if there are two suffixes "ation" and "ion" will be there then "ation" will be removed [1].

4. APPROACH USED IN STEMMER

Our stemmer uses a brute force approach. Brute force search is also called exhaustive search also known as generate and test this is a systematical approach which search for all the possible solution in the data. This approach employs a lookup table which contains relations between root words and inflected words. This lookup table is used to store the words. Stemming is done by finding a word in the table if the match is found then the associated word is generated. Brute force requires immense amount of storage to create a database but it reduces the problem of under-stemming and over-stemming. Like in English when we are applying suffix stripping then running gives run but not ran. Brute force algorithm requires large number of words in their database. Brute force accuracy depends on the data present in the database. With the brute first technique we can also use production technique, which works as s reverse to brute force technique. In this it generates the table of root and inflected forms. If the word is not found in the database then we are using suffix stripping to handle these words. In this suffix stripping we have a list of suffixes which are removed from the inflected word and the solution for the word is generated. e.g. if we have a number of elements for a solution like [$A=\{a_1,a_2,a_3,\ldots,a_n\}$].Now there are 2^n possible values for A. Brute force algorithm is used to find the best possible solution for these type of problem. It will check for the best result in all the elements and the output is called feasible solution.

5. STEMMING OF PUNJABI

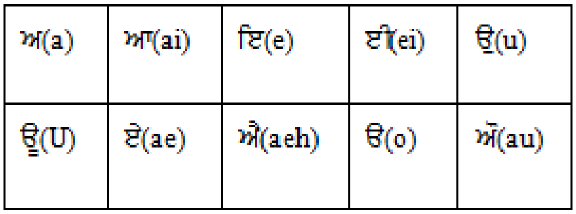
Punjabi is called also Gurmukhi or Shahmukhi. It was developed by Guru Nanak Dev ji (first Sikh Guru) in the 16th Century. Gurmukhi means "from the mouth of the Guru"[20]. Shahmukhi is mostly spoken in Pakistan and written in Arabic text. Arabi has a different syntax than the Punjabi text. Shahmukhi is not so much spoken in Punjab. Shahmukhi have different writing syntax than the Gurmukhi Language. Due to writing syntax in Arabic this is most popular in Pakistan. Gurmukhi has its own features like it is a tonal language with three tones. In this when the consonants are used together, then special symbols are used together to combine the rest part of the word. Here is the list of vowels, laga matra, consonants and the other symbols that are used in Gurmukhi. By using these components you can learn and speak Punjabi easily. In the section 6, it contains the list of symbols and in the brackets these is a text written that means how to pronounce the Punjabi characters.

6. CHARACTER SET FOR PUNJABI

6.1 Vowels

This is a list of vowels that are used in Punjabi represented in Table 2. We have written the vowels in Punjabi and in the brackets it shows how to pronounce it by an English language.

Table 2 Vowel List in Punjabi

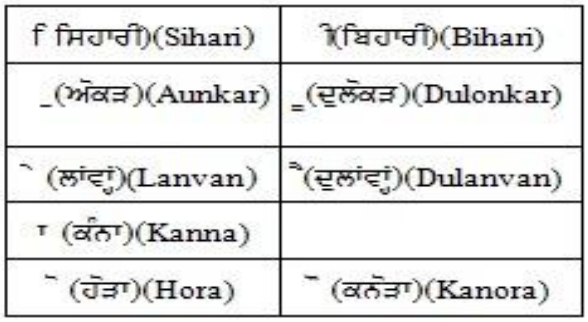


ਅ(a)	ਆ(ai)	ਇ(e)	ਈ(ei)	ਉ(u)
ਊ(U)	ਏ(ae)	ਐ(aeh)	ਓ(o)	ਔ(au)

6.2 Laga Matra (lgw mwqrw)

This is a list of laga matra that are used in Punjabi language. This is like in English language if we want to write a word in English in plural form then we are writing ‘s’ or ‘es’ after the text. These laga matra are used in Punjabi for representing plural, adjective, ad verb of a word. E.g. if we are writing boy in English then for make it plural we are using ‘s’ after boy and it becomes boys. If we are writing ^br in Punjabi then for a plural of ^br we have to add w after a word. Then the word becomes ^brw. These are the helping characters for making a proper word. In the brackets in Table 3, we have shown how to pronounce a Punjabi laga Matra. Laga Matra is shown in following table.

Table 3 Laga Matra in Punjabi

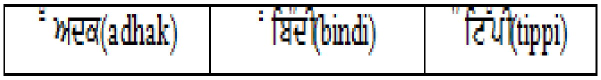


ਿ ਸਿਹਾਰੀ)(Sihari)	ੀ(ਬਿਹਾਰੀ)(Bihari)
ੁ(ਔਂਕੜ)(Aunkar)	ੂ(ਦੁਲੈਂਕੜ)(Dulonkar)
ੇ (ਲਾਂਵਾਂ)(Lanvan)	ੈ(ਦੁਲਾਂਵਾਂ)(Dulanvan)
ਾ (ਕੰਨਾ)(Kanna)	
ੋ (ਹੋੜਾ)(Hora)	ੌ (ਕਨੌੜਾ)(Kanora)

6.3 Other Symbol

This is a list of some other special symbol that is also helpful in writing words in Punjabi. These symbols are used whenever there will be some sound produced in a word. These symbols are represented in table 4. The list of these symbols is listed below in the table:

Table 4 Other symbols used in Punjabi

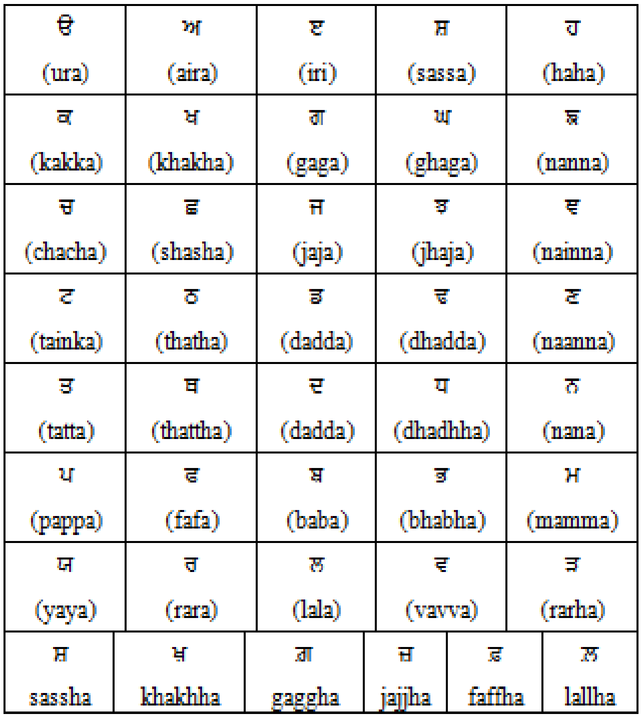


ੱ ਅਦਕ(adhak)	ਂ ਬਿੰਦੀ(bindi)	ੰ ਟਿੰਪੀ(tippi)

6.4 Consonants (ivAMjn)

This is a list of characters that are used to make a Punjabi word .Without understanding and studying these words we could not speak and read Punjabi language. E.g. to study an English language we must have an understanding about A to Z characters. Without these we could not read and speak English language. Consonants are represented in table 5.

Table 5 List of Characters

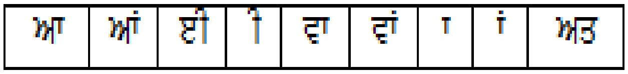


<table>
<tr><td>ੳ
(ura)</td><td>ਅ
(aira)</td><td>ੲ
(iri)</td><td>ਸ
(sassa)</td><td>ਹ
(haha)</td></tr>
<tr><td>ਕ
(kakka)</td><td>ਖ
(khakha)</td><td>ਗ
(gaga)</td><td>ਘ
(ghaga)</td><td>ਙ
(nanna)</td></tr>
<tr><td>ਚ
(chacha)</td><td>ਛ
(shasha)</td><td>ਜ
(jaja)</td><td>ਝ
(jhaja)</td><td>ਞ
(nainna)</td></tr>
<tr><td>ਟ
(tainka)</td><td>ਠ
(thatha)</td><td>ਡ
(dadda)</td><td>ਢ
(dhadda)</td><td>ਣ
(naanna)</td></tr>
<tr><td>ਤ
(tatta)</td><td>ਥ
(thattha)</td><td>ਦ
(dadda)</td><td>ਧ
(dhadhha)</td><td>ਨ
(nana)</td></tr>
<tr><td>ਪ
(pappa)</td><td>ਫ
(fafa)</td><td>ਬ
(baba)</td><td>ਭ
(bhabha)</td><td>ਮ
(mamma)</td></tr>
<tr><td>ਯ
(yaya)</td><td>ਰ
(rara)</td><td>ਲ
(lala)</td><td>ਵ
(vavva)</td><td>ੜ
(rarha)</td></tr>
<tr><td>ਸ਼
sassha</td><td>ਖ਼
khakhha</td><td>ਗ਼
gaggha</td><td>ਜ਼
jajjha</td><td>ਫ਼
faffha</td><td>ਲ਼
lallha</td></tr>
</table>

6.5. List of some Suffixes

This is a list of suffixes that we have used in stemmer .These suffixes are mostly present with the inflected words. This list is shown in

Table 6 List of Suffixes



ਆ	ਆਂ	ਈ	ੀ	ਵਾ	ਵਾਂ	ਾ	ਾਂ	ਅਤ

7. THE PROPOSED STEMMER

In our stemmer the inflected word in entered as an input. Our stemmer will look for the matching word in the database. Our database contains a list of various Punjabi inflected as well as root words. It also contains a list of various probabilities for a word. If the word is found in the database then it will give us the result .The result is in the form of root word. There are number of assumptions are present for an entered word. Our stemmer works on two pointers. One is the text pointer and other is the matching pointer. Text pointer is used to enter the text and carry the text for matching .Text pointer takes an input from the input box and takes this word to the next level for final output. Matching pointer is used to match the word with the database words. If the word found it will give us a result by increment the matching pointer by 1 else decrement the matching pointer. These pointers are helpful for the proper working of system .Because these pointer tells the system whether the word if found or whether the system has to remove the suffix from the word. These pointers also tells how many words are stemmed in one cycle means that whether they are stemmed by using brute force or by suffix removing. Block diagram for proposed stemmer is shown below.

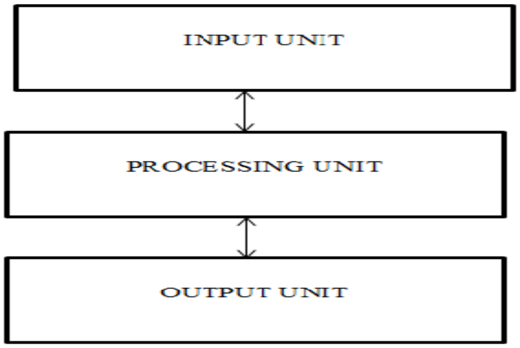
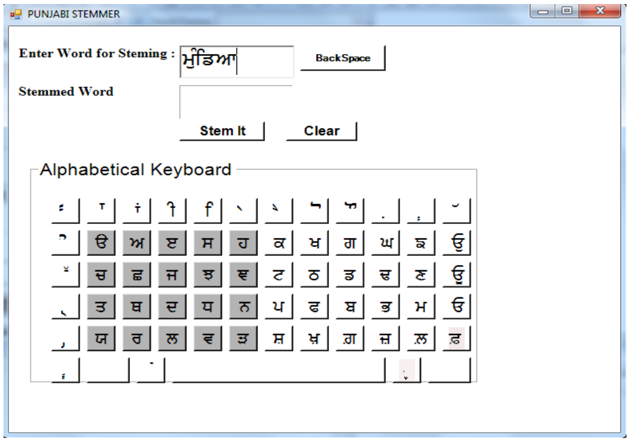


Figure 1 Block Diagram of Stemmer

Block diagram of stemmer have three units. Input unit, processing unit and output unit.

7.1 Input Unit

In the input unit we have to give an input. The input is in the form of any Punjabi word. The input diagram is shown below



PUNJABI STEMMER

Enter Word for Steming : ਮੁੰਡਿਆ

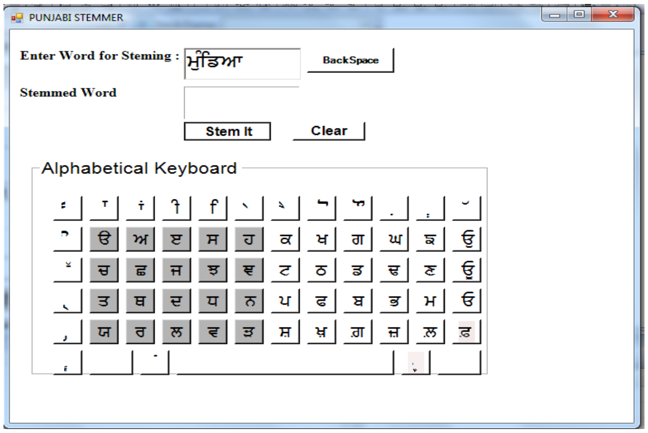
Stemmed Word

Figure 2 Inputs to the Punjabi Stemmer

Here in the designing of this stemmer we also provide a Punjabi keyboard on the screen. With the help of this Punjabi keyboard a new user can write any word easily by just clicking on the buttons. We have used amrboli font in it. Keyboard makes the stemmer user friendly. Here the input is `muMifAw` which we have entered in the input box. Here in the fig it also shows some other controls. These controls are backspace and clear. Clear is used whenever we want to enter a new word for stemming after getting the result of the previous one. It clears both the input and output box and the cursor automatically move to the input box. Backspace is used whenever we want to clear any character to the input box. Backspace clears the character in backward direction**.**

7.2 Processing Unit

Processing of the input will start whenever we click on the "stem it" control button. After pressing the "stem it" control button our stemmer starts searching the input word in the database. It starts the searching by using brute force technique. The blank box which is shown below the input box is output box. In the fig 3 it shows the "stem it "button has been pressed for processing.



PUNJABI STEMMER

Enter Word for Steming : ਮੁੰਡਿਆ

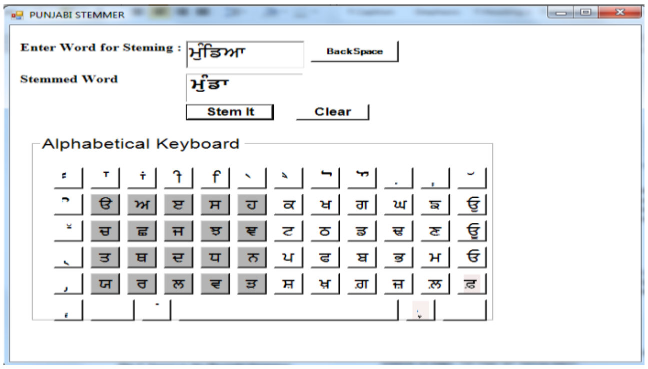
Stemmed Word

Figure 3 Processing of the word for Stemming

Stemmer tries to find the match in the database. If the match for the input word occurs, it displays the result in the output box. Otherwise it checks for the suffix list. If any suffix matches with the word ending it removes the suffix and shows the output in the output box. Flowchart for processing unit is also shown in the fig 5.

7.3 Output Unit

In the output unit stemmer shows the result. The result comes after the processing of word. As we have shown in the fig 2 we are giving input `muMifAw` .In the fig 3 we are pressing "stem it" to get the result . The result is shown in fig 4.



PUNJABI STEMMER

Enter Word for Steming : ਮੁੰਡਿਆ

Stemmed Word ਮੁੰਡਾ

Figure 4 Output of the Punjabi Stemmer

In the output box it shows the result after processing. The result is `muMfw.`

8. STEPS OF STEMMING

The processing of the inflected word for stemming is done in number of steps. The steps are shown below.

1. The inflected word is entered as an input
2. The word is searched by the technique of brute force searching. Brute force search try to match the input word with the number of words present in the list. If the matching word is found then it will give the output as a root word. If the match not found then it look for the list of suffixes. It removes the matching suffix from the end of the word.

3. The root word is a suitable word with the desired input

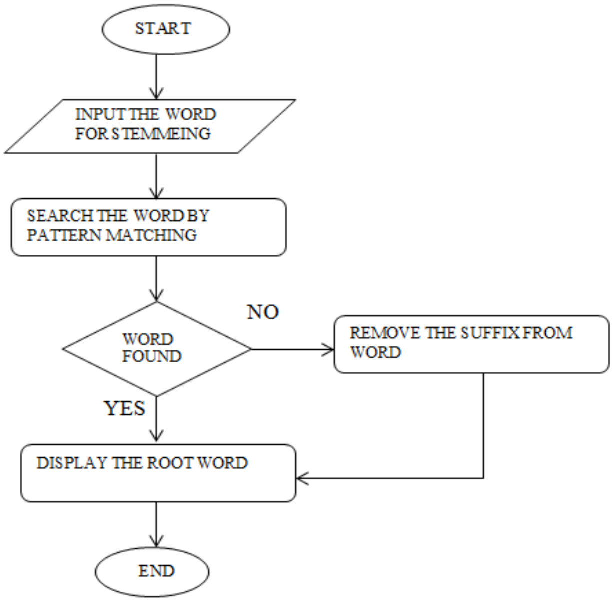

START
INPUT THE WORD FOR STEMMEING
SEARCH THE WORD BY PATTERN MATCHING
WORD FOUND
NO
REMOVE THE SUFFIX FROM WORD
YES
DISPLAY THE ROOT WORD
END

Figure 5 Steps of Punjabi word Stemming

e.g. if we have entered 'boys' and then it search for boys in the list when the suitable match occurs it will displays the word in front of 'boys'. It will result as a 'boy'. If the word not found by the list then there is a list of Punjabi suffixes that are also present as a list. Then the control is transferred to suffix removal technique. Our stemmer will remove the desired suffix from the end of the word. After removing the ending from the word it will give us a word as an output. This will happen only if the word is not found by the brute force technique

9. EVALUATION

For evaluating the proposed stemmer following parameters are used:

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

Correctness [9] of our stemmer is depending on the word present in the look up table. We have entered 28000 words in our lookup table therefore it reduces the chances of going to the second option of suffix stripping. Bigger database causes more correctness. These words are taken from Pardeep Punjabi to English Dictionary, www.shurli.com, www.ajitjalandhar.com, http://forum.desicomments.com/.

Effectiveness [20] of the stemmer depends on the behavior of the system. Behavior means what stemmer will do whenever an abnormal condition occur. Abnormal condition means whenever somebody tries to enter a word which does not exists. Our stemmer is very effective if anybody tries to enter a word that does not exists it just shows that word in the output box. Our stemmer effectively adjusts in different type of environment. Over-stemming and under-stemming problem are also controlled in the stemmer. These errors are more whenever we are using suffix removal technique. By using brute force technique we have controlled these errors. If the word found in the database then there is no chance of these errors to occur.

Performance [21] of stemmer will be high if the result is positive. The conflation class of stemmer is 4.Here in the above calculation it shows that we have 28000 words entered in the list and 7100 words are root type words we can calculate mean number of words [21] by dividing the number of unique words with the number of unique stem after stemming.

MWC=Mean Number of Words.

MWC is given by the following equation.

MWC=N/S --------------- (1)

N=Number of unique words before stemming

S=Number of unique stem after stemming

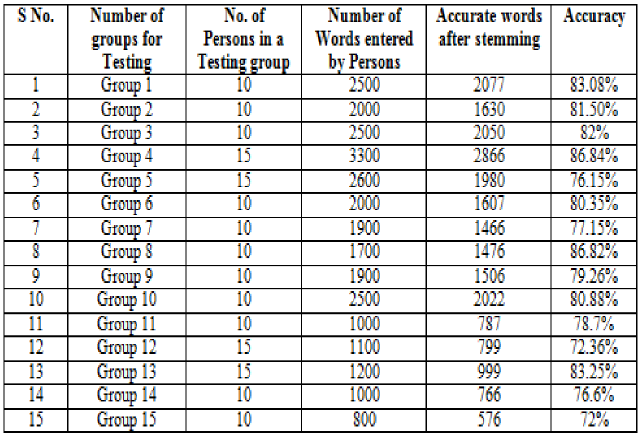
In our stemmer number of unique word before stemming are 28000 and the number of unique words after stemming are 7100. By substituting these values in equation no 1 we get

MWC=28000/7100

MWC=3.94

We have tested our stemmer with the help of ten groups of persons. Results are shown in Table 7.

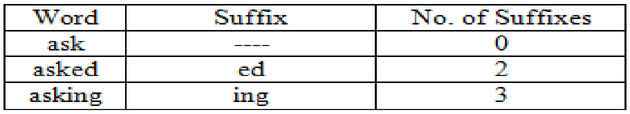
Table 7 Evaluation Results



S No.	Number of groups for Testing	No. of Persons in a Testing group	Number of Words entered by Persons	Accurate words after stemming	Accuracy
1	Group 1	10	2500	2077	83.08%
2	Group 2	10	2000	1630	81.50%
3	Group 3	10	2500	2050	82%
4	Group 4	15	3300	2866	86.84%
5	Group 5	15	2600	1980	76.15%
6	Group 6	10	2000	1607	80.35%
7	Group 7	10	1900	1466	77.15%
8	Group 8	10	1700	1476	86.82%
9	Group 9	10	1900	1506	79.26%
10	Group 10	10	2500	2022	80.88%
11	Group 11	10	1000	787	78.7%
12	Group 12	15	1100	799	72.36%
13	Group 13	15	1200	999	83.25%
14	Group 14	10	1000	766	76.6%
15	Group 15	10	800	576	72%

Here we have taken fifteen groups. Each group has different number of persons. Each group has different number of words to be tested. We have calculated the accuracy for each group and then calculate the accuracy of our stemmer by calculating the average accuracy of the groups. Average accuracy of our stemmer is 80.73%.

Mean removal rate [21] is measured by calculating the number of suffixes attached with any word. E.g. if a word in English like ask is used the various variations for ask is asking, asked. Asking have three suffixes at end, asked have two and ask itself has no suffix therefore total number of suffixes are 3.



Word	Suffix	No. of Suffixes
ask	----	0
asked	ed	2
asking	ing	3

Mean Removal Rate=Sum of all the suffixes/Number of suffixes
Mean Removal rate for word ask is=0+2+3/3 =1.66

10. CONCLUSION

Our stemmer works for Punjabi and this is a simplified version of stemmer. Brute force technique requires no preprocessing of text before stemming a word. It basically emphasizes on brute force techniques. There is a very little influence of suffix stripping algorithm in our stemmer. There is a problem of over-stemming and under-stemming comes under suffix stripping approach. We can also do suffix substitution with suffix stripping to avoid the problem of over-stemming and under-stemming. In future we can also create the stemmer by using some other techniques.

11. REFERENCES

[1] Julie Beth Lovins, (1968)"Development of a Stemming Algorithm*"Mechanical Translation and Computational Linguistics, Vol No.11, Issue No.1, pp 22-31.

[2] M.F Porter (1980)" An algorithm for suffix stripping" Published in Program, Vol No.14, Issue No.3, pp 130-137,URL:http://www.cs.odu.edu/~jbollen/IR04/readings/readings5.pdf.

[3] David A Hull Gregory Grefenstette (1996)" A Detailed Analysis of English Stemming Algorithms "Rank Xerox research Centre 6 chemin day mauperutis, 38240 Melyanfrance,pp 1-16,

[4] Jörg Caumanns (1998) "A Fast and Simple Stemming Algorithm for German Words1"Algorithm is Publish in Department of computer science at the free university of Berlin, pp 1-10,

[5] Tanja Gaustad and Goose Bauma (2000)"Accurate Stemming of Dutch for Text Classification" Language Computing. Vol No.45, Issue No. 1, pp 104-117.

[6] Bal Krishna Bal, Prajol Shrestha (2004)"A Morphological Analyzer and a Stemmer for Nepali" PAN Localization, Working Papers 2004-2007, pp 324-31.

[7] Md. Zahurul Islam, Md. Nizam Uddin and Mumit Khan (2004)" A Light Weight Stemmer for Bengali and Its Use in Spelling Checker" Proceedings of 1st International Conference on Digital Communications and Computer Applications (DCCA2007), Irbid, Jordan, pp 87-93.

[8] Haidar Harmani, Walid Keirouz, & Saeed Raheel (2006) "A rule base extensible stemmer for Information retrieval with application to Arabic" The international Arab journal of information technology, Vol No.3, Issue No.3, pp 265-272.

[9] Ababneh Mohammad , Oqeili Saleh and Rawan A. Abdeen(2006)" Occurrences Algorithm for String Searching Based on Brute-force Algorithm" Jordan Journal of Computer Science Vol No.2,Issue No.1,pp 82-85 .

[10] Jiaul H. Paik and Swapan K. Parui (2008) "A Simple Stemmer for Inflectional Languages" Journal of Documentation, Vol No.61 Issue No.4, pp. 476 – 496.

[11] Muhamad Taufik Abdullah†, Fatimah Ahmad†, Ramlan Mahmod† and Tengku Mohd Tengku Sembok (2009) "Rules Frequency Order Stemmer for Malay Language" IJCSNS International Journal of Computer Science and Network Security, Vol No.9, Issue No.2, pp 433-438.

[12] Miguel E. Ruiz and Bharath Dandala (2010) "Evaluating Stemmers and Retrieval Fusion Approaches for Hindi: UNT at FIRE 2010" URL:http:// www.isical.ac.in/ ~fire/paper_2010 /MiguelRuiz-unt-fire-2010.pdf

[13] Ananthakrishnan Ramanathan and Durgesh D Rao "A Lightweight Stemmer for Hindi" In Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL), on Computatinal Linguistics for South Asian Languages (Budapest, Apr.) Workshop,pp 42-48

[14] Abduelbaset M. GOoweder, Husien A. Alhammi , Tarik Rashed,and Abdulsalam Musrati " A Hybrid Method for Stemming Arabic Text " Journal of computer Science,URL: http://eref.uqu.edu.sa/files/eref2/folder6/f181.pdf.

[15] Samir Abdou and Patrick Ruck and Jacques Savoy (2005) " Evaluation of Stemming, Query Expansion and Manual Indexing Approaches for the Genomic Task "NIST Special Publication :SP 500-266 The Fourteenth Text Retrieval Conference(TREC 2005) Proceedings ,URL: http://trec.nist.gov/pubs/trec14/papers/uneuchatel.geo.pdf.

[16] James Mayfield and Paul McNamee "Single N-gram Stemming" SIGIR 2003: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, July 28 - August 1, 2003, Toronto, Canada,pp 415-416 .

[17] Marie-Claire Jenkins, Dan Smith, "Conservative stemming for search and indexing" School of Computing Sciences.University Of East-Anglia Norwich NR4 7TJUK,URL:www.uea.ac.uk/polopoly_fs/1.85493!stemmer25feb.pdf.

[18] Hayder K. Al Ameed, Shaikha O. Al Ketbi, Amna A. Al Kaabi, Khadija S. Al Shebli,Naila F. Al Shamsi, Noura H. Al Nuaimi, Shaikha S. Al Muhairi "Arabic Light Stemmer: A New Enhanced Approach" The Second International Conference on Innovations in Information Technology (IIT'05) URL:http://www.onlinelibrary.wiley.com/doi/10.1002/asi.21247/pdf.

[19] Ghassan Kanan,Mohammad Ababney,Riyad Al Shalabi,Alla Nal Nobani"Building an effective rule based light Stemmer Arabic language to improve search effectiveness "Arab Academy for banking and financial science,Al balka Applied university,pp. 312-316,URL:www.ccis2K.org/iajit/PDF/vol.3,no.3no.3/12-Haidar.pdf,