Enlargement of Clinical Stemmer in Hindi Language for Homoeopathy Province

Pramod P. Sukhadeve and Sanjay K. Dwivedi

Department of Computer Science
Babasaheb Bhimrao Ambedkar University, Lucknow, India
sukhadeve.pramod@gmail.com, skd200@yahoo.com

Abstract. Stemming is an evolution that integrates morphologically corresponding stipulations into a single term devoid of doing inclusive morphological scrutiny. Stemming is utilized in information retrieval systems to acquire enhanced performance. Moreover, this process diminished the number of terms in the information retrieval system. This paper presents and discusses rule-based Hindi language clinical stemmer, which incorporates terms manually extracted affix stripping rules. The proposed stemmer is not domain reliant but we use this stemmer in Homoeopathy language. It converse the well-organized technique in which the affix list was developed, and in attendance the linguistic grounds subsequently including various affixes in the inventory. Analogous performance can be tattered to build stemmer for auxiliary fields such as story books, news articles, etc.

Keywords: Homoeopathy, Hindi Stemmer, prefix, suffix, root words, Grammar, clinical words.

1 Introduction

Stemming is the pre-processing step of collapsing words into their morphological root. For example, the terms दवाइ, दवाइयाँ, दवाइवाला, दवाखाना, दवात might be conflated to their stem nok - The first paper on the stemmer was published in 1968. It was written by Julie Beth Lovins [1] it is a rule based stemmer. This is a single pass context-sensitive, longest match stemmer which make use of a list of about 250 unusual suffixes, and removes the longest suffix attached to the word, that the stem after the suffix has been removed is always at least 3 characters long. Another popular rule based stemmer developed was by Porter in 1980, popularly called as Porter's stemming algorithm [2].

This stemmer was very extensively used and became the genuine standard algorithm used for English stemming. The procedure of stemming is also called conflation. It is also used in indexing and search system. These 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

N. Meghanathan et al. (Eds.): CCSIT 2012, Part III, LNICST 86, pp. 239–248, 2012.

algorithm, suffix stripping algorithm, lemmatization algorithm, stochastic algorithm, N gram analysis, hybrid approaches. These stemming algorithms are different in performance and accuracy. These all techniques have different approaches and methods to stem words.

2 Approaches to Stemming

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, stemming algorithm is used for German languages, and in this stemmer firstly it removes a suffix from the word 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 [3]. In the Dutch stemmer it uses a suffix stripping algorithm and dictionary lookup rule based methods [4]. 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 [5]. 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 [6]. In the lightweight stemmer for Hindi it uses a look up table approach in which word is matched with the words present in the 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 [7]. 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 [8].

3 Language and Grammar

Hindi is a direct descendant of Sanskrit through Prakrit and Apabhramsha. It has been influenced and enriched by Dravidian, Turkish, Farsi, Arabic, Portugese and English. It is a very expressive language. Throughout the world more than hundreds of universities having Hindi as subject in syllabus. Most of the famous composition of Hindi has been translated into foreign languages. In India Delhi, Haryana, Bihar, Uttar Pradesh, Madhya Pradesh, Himachal Pradesh are Hindi heart lands, but non Hindi states like Kerala, Andhra Pradesh, Assam, Manipur, Maharashtra, Gujrat, Calcutta, Nagaland, Tripura etc., are also accumulating Hindi. Today’s in India Hindi is using in terms of communicating language, State language, Mother tongue, official language etc. In India most of the states using Hindi and also in News papers, School books, Advertisements, etc. so Hindi language is India’s dignity.

In Hindi some of the characters and words prevailing two type of forms like;-

Table 1. Two type of forms of words

(शब्द) Words	
गये	गए
नयी	नई
गयी	गई
चाहिये	चाहिए
जायेंग	जाएँगे
अन्त	टंत
सम्बन्ध	संबंध

There are 13 vowels in Hindi which are shown below in table 2.

Table 2. Hindi Vowels

Vowel (मात्रा)

अ	आ	औ	ऋ	अं	इ
ए	अः	ई	ऐ	उ	ओ

There are certain consonents (व्यजंन) in hindi which are shown in table 3.

Table 3. Hindi Consonants

Consonant (व्यजंन)

क	ख	ग	घ	ड	च	छ	ज	झ
त्र	ट	ठ	ड	ढ	ण	त	थ	द
ध	न	प	फ	ब	भ	म	य	र
ल	ट	श	ष	स	ह	क्ष	त्र	ज्ञ

4 Stemmer Characteristics of Hindi Language

A Hindi is morphologically very rich. A single root word may have different morphological variants, for example, words like डॉक्टरों, डॉक्टरी are morphological variants of the word डॉक्टर.

4.1 Prefix

The variants in Hindi are usually formed by adding prefixes to the stem or root word. We can categories the Prefixes found in Hindi language.

Syllable which is added in front of the word and then influenced the meaning of the word, those syllables is called prefix.

Ex. अंगघात = अंग + घात
अंगघाती = अंग + घाती
अंगघातग्रस्त = अंग + घात + ग्रस्त
अंगच्छेदक = अंग + च्छेदक

syllables are Prefixes :- In feasible examples “अंग, आ, उप, प्र, वि, स” are syllables. Prefixes are applied only in the form of syllable.
Prefixes are added in front of words: - most of the time peoples remove syllable to obtain the prefix.

Ex. अकरण = अ + करण
कुशल = कु + शल
बोध = ब + ओध

In the above examples prefixes are correct and have syllables but they are not concert with any words. In the above examples where prefixes are added to the improper words, hence these types of separations are not true.

Prefix influenced the meaning of words: - Prefixes are used to get the new meaning of the words, when prefixes added to the words, then the meaning of the root words are changed. In the above example “आ, उप, प्र, वि, and सं” are prefixes, they changed the meaning of the root word “अंग”.

Hindi Prefixes are shown below in the Table 4.

Table 4. Types of Hindi Prefixes

Prefix	Meaning of prefix	new words
अ	अभाव	अकाल अमर
अन	के बिना	अनजान अनहोनी
अध	आधा	अधमरा अधपका
वु	बुरा	कुपुत्र कुकर्म
भर	पुरा	भरपेट भरमार

Sometimes more than one prefixes are attached with the word examples are shown below in the table 5.

Table 5. Words append with the Prefixes

Prefix	Prefix	Root word	Words
सम्	टा	लेचना	समालोचना
निर्	अभि	मान	निरभिमान
प्रति	डप	कर	प्रत्युकार
सु	सम्	गठित	सुसंगठित

4.2 Suffix

Syllables which are added at the end of the word and then manipulate the meaning of word, those syllables is called suffix.

Ex. धार्मिक = धर्म + इक
पशुता = पशु + ता
लिखावट = लिख + आवट

In Hindi there are certain types of suffixes, some suffixes used in verbs, some are used in Noun, Pronoun and Adjectives. Reversal in the gender form by using some of the suffixes, and some consumed in the separation form. In this way there are certain types of suffixes in Hindi. So, for the purpose of perusal it is divided into three main parts.
I) Secondary suffix (तद्धित प्रत्ययद्धरू Suffixes used at the end of the Noun, Pronoun and Adjective are called Secondary suffix

पशुत्व = पशु + त्व
अपनापन = अपना + पन
बुध्दिमान् = बुध्दि + मान्

a) The vibrant (कर्तृवाचक) secondary suffix: Suffix that makes the realization of a task is called the vibrant suffix.

Table 6. Vibrant suffix

Suffix	Words	Suffix	Words
आर	लुहार	इया	दुखिया
ई	शराबी	एरा	सपेरा
अक	लेखक	कार	कलाकार
आर	छुकानदार	वान	धनवान
वाला	गाड़ीवाला	गर	जादूगर

b) Abstract (भाववाचक) suffix: Suffix that makes the realization of an abstract is called an abstract suffix.

Table 7. Abstract suffix

Suffix	Words	Suffix	Words
आहट	गरमाहट	आई	भलाई
आ	बुलावा	ई	सर्दी
आस	मिठास	त्व	देवत्व
पा	बुढापा	पन	बालपन

c) Relational (संबंध.वाचक) suffix: Suffix that makes the realization of an relation is called an relational suffix.

Table 8. Relational suffix

Suffix	Words	Suffix	Words
आल	स्सुराल	इक	धार्मिक
एरा	च्चेरा	आ	भतीजा

d) Adjectival (विशेषण) suffix: Suffix that makes the realization of a quality is called an adjectival suffix.

Table 9. Adjectival suffix

Suffix	Words	Suffix	Words
आ	अंडा	ई	क्रोधी
ईय	दयनीय	इत	थ्यंतित
ईला	च्मकीला	इल	जटिल

II) Feminine (स्त्रीवाचक) suffix: - suffixes that are used to convert words into feminine are called feminine suffix.

Table 10. List of Feminine Words

Suffix	Words	Suffix	Words	Suffix	Words
ई	लड़की	आ	शिष्या	इन	सुनारिन
नी	मोरनी	आनी	नौकरानी	आइन	ठकुराइन
इया	बिटिया	इका	अध्यापिका	मती	भगवती

III) Krit (कृत्) suffix: suffix using at the end of the things for making noun, adjective and indeclinable is called Krit suffix.

Ex. लिखावट = लिखना + आवट
खिलाड़ी = खेलना + आड़ी

a) Virant (कर्तृ) Krit (कृत्) suffix:- Suffix that makes the realization of a verb and Subject is called an vibrant Krit suffix.

Table 11. Virant Krit

Suffix	Words	Suffix	Words
टक	सहायक	आक	तैराक
आड़ी	अनाड़ी	आलू	दयालु
एरा	लुटेरा	वाला	जानेवाला

b) Object (कर्मवाचक) Krit (कृत्) suffix:- Suffix that makes the realization of verb and object is called objective suffix.

Table 12. Object Krit

Suffix	Words	Suffix	Words
औना	खिलौना	नी	ओढ़नी
आना	खाना	आवना	डरावना

c) Abstract (भावाचक) Krit (कृत्) Suffix:- Suffix that makes the realization of verb and abstract is called objective suffix.

Table 13. Abstract Krit

Suffix	Words	Suffix	Words
आई	लिखाई	आन	उड़ान
टाप	थ्मलाप	आवट	लिखावट
आहट	घबराहट	आव	चढ़ाव

d) Past and Present Krit Suffix:- Suffix that makes the realization of Noun, Indeclinable, and special meaning verbs is called Verbal suffix.

Table 14. Past and Present Krit Suffix

Suffix	Words	Suffix	Words
आ	सूखा	ता	चढ़ता

5 Combination of Suffix and Prefix

Most of the words are made by using prefix and suffix both, for example.

Words		Prefix	+	Root Word	+	Suffix
उपकारक	=	उप	+	कार	+	अक
अभिमानी	=	अभि	+	मान	+	ई
अपमानित	=	अप	+	मान	+	इत
बदचलनी	=	बद	+	चलन	+	ई
दुस्साहसी	=	दुस	+	साहस	+	ई
निर्दयी	=	निर	+	दया	+	ई

6 Pattern of Hindi Words

The variants in Hindi are usually formed by adding prefixes and suffixes to the stem or root word. We can categories the prefixes and suffixes found in Hindi into three types:

I) Plain Prefixes and Suffixes: when prefixes added to the words, then the meaning of the root words are changed. Plain suffixes are also called dependent vowel signs ा, ि,ी, ो,ॆ, ु ,ृ are some of the suffixes which combine with the root word to produce its morphological variants.
For Example, लडका, लडकि, लडके, लडकियाँ.
II) Join word Prefixes and suffixes: Join word prefixes are those which are produced by adding उन, उप, नि, वि, and सु.
उनचास = उन + चास
उपहार = उप + हार

Join word suffixes are those suffixes which are formed by merging two or more consonants and vowels. These join words are formed by merging any of the consonants with the morphological variant

गाड़ीवाला = गाड़ी + वाला
कारीगर = कारी + गर
दुकानदार = दुकान + दार

III) Complex suffixes: Complex Prefixes are formed by adding more than one prefixes to the root word. Like

समालोचना = सम् + आ + लोचना
निरभिमान = निर् + अभि + मान
प्रत्युकार = प्रति + उप + कार

Complex suffixes are formed by combining two or more consonants with the plain suffixes. For example,

भुलक्कड़ = भुल + अक्कड़
मनुष्यत्व = मनुष्य + त्व
नैतिक = नैत + इक
are some of the complex suffixes.

IV) Words follow a specific pattern: Unlike other Indian languages it is found that words in Hindi language follow a specific pattern. The words in the Hindi can be expressed as:

Token = Prefix + Root Word for Prefix
Token = Root Word + Suffix for Suffix
Token = Prefix + Root Word + Suffix

7 Our Approach towards Affix Stripping

The rule-based stemmer extract affix stripping rules based on the morphological characteristics of Hindi.The common stemming patterns found in Hindi are:

{Original word}:= {Plain Prefix} + {SW}
e.g प्रहार = प्र + हार
{Original word} :={ Complex Prefix} + {Root word}
e.g निर्दय = निर् + दय
{Original word} :={ Plain Prefix + join word + Complex prefix}+ {Root word}
e.g. सुसंगठित = स + उ + सम् + गठित
{Original word} :={ SW} + {Plain suffixes}
e.g, पशुता = पशु + ता

{Original word} :={ Root word} + {complex suffixes}
e.g, पशुत्व = पशु + त्व
{Original word} :={ Root word} + { join word + Complex suffixes+ Join Word}
e.g, भुलक्कड+ = भुल + अक् + कड़

The affix (prefix and suffix) stripping rules for the rule-based stemmer are based on these patterns. Fig. 1 depicts steps in the algorithm.

Step 1: Input list of words
उपकारक
Step 2: Eliminate all the complex affixes
e.g, {उप} + {कार} + {अक}
Step 3A: Eliminate the join word suffixes i.e. Eliminate the inflections of consonants like क, द, फ, व, with र्
e.g, {क्र} = {क}+{र्}
Step 3B: Eliminate the join word prefixes i.e. Eliminate the inflections of consonants like प्र, क, उ, प
e.g, {प्रहार} = {प्र}+{हार}
Step 3C:- Eliminate the inflections for consonant j
Step 4A: Eliminate the inflections for consonant y
Step 4B: Eliminate the inflections involving plain affixes.

Fig. 1. Algorithm developed for Hindi Stemmer

8 Conclusion

In this paper we have presented a stemming methodology that is based on a hand crafted rule based system. The rule based system models phenomena of inflectional languages in a linguistic and consistent way. We discussed different approaches of stemming and detailed depiction of Prefix and Suffix. The proposed clinical stemmer incorporates terms manually extracted affixes stripping rules and algorithm, it is not domain relevant. It converse the well categorized technique in which the affixes are detached.

References

1. Lovins, J.B.: Development of a stemming algorithm. Mechanical Translation and Comp. Linguistics 11, 22–31 (1968)
2. Porter, M.F.: An algorithm for suffix stripping Program, vol. 14(3), pp. 130–137 (1980)
3. Caumanns, J.: A fast and simple stemming Algorithm for German Words1, pp. 1–10. Alg. is published in Dept. of comp. sci. at the free univ. of Berlin (1998)
4. Gaustad, T., Bauma, G.: Accurate Stemming of Dutch for Text Classification. Language Computing 45(1), 104–117 (2000)
5. Bal, B.K., Shrestha, P.: A Morphological Analyzer and a Stemmer for Nepali. In: PAN Localization, Working Papers 2004-2007, pp. 324–331 (2004)

6. Zahurul Islam, M., Nizam Uddin, M., Khan, M.: A Light Weight Stemmer for Bengali and Its Use in Spelling Checker. In: Proceedings of 1st International Conference on Digital Communications and Computer Applications (DCCA 2007), Irbid, Jordan, pp. 87–93 (2004)
7. Ramanathan, A., Rao, D.D.: A Lightweight Stemmer for Hindi. In: Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL), Workshop on Computatinal Linguistics for South Asian Languages, Budapest, pp. 42–48 (April)
8. Goweder, A.M., Alhammi, H.A., Rashed, T., Musrati, A.: A Hybrid Method for Stemming Arabic Text. Journal of Computer Science, `http://eref.uqu.edu.sa/files/eref2/folder6/f181.pdf`