Automatic Stemming of Words for Punjabi Language

Vishal Gupta

Abstract. The major task of a stemmer is to find root words that are not in original form and are hence absent in the dictionary. The stemmer after stemming finds the word in the dictionary. If a match of the word is not found, then it may be some incorrect word or a name, otherwise the word is correct. For any language in the world, stemmer is a basic linguistic resource required to develop any type of application in Natural Language Processing (NLP) with high accuracy such as machine translation, document classification, document clustering, text question answering, topic tracking, text summarization and keywords extraction etc. This paper concentrates on complete automatic stemming of Punjabi words covering Punjabi nouns, verbs, adjectives, adverbs, pronouns and proper names. A suffix list of 18 suffixes for Punjabi nouns and proper names and a number of other suffixes for Punjabi verbs, adjectives and adverbs and different stemming rules for Punjabi nouns, verbs, adjectives, adverbs, pronouns and proper names have been generated after analysis of corpus of Punjabi. It is first time that complete Punjabi stemmer covering Punjabi nouns, verbs, adjectives, adverbs, pronouns, and proper names has been proposed and it will be useful for developing other Punjabi NLP applications with high accuracy. A portion of Punjabi stemmer of proper names and nouns has been implemented as a part of Punjabi text summarizer in MS Access as back end and ASP.NET as front end with 87.37% efficiency

Keywords: Punjabi Stemming, Punjabi Noun Stemming, Punjabi Verb Stemmer, Punjabi Names Stemmer, Punjabi Adjective Stemmer.

1 Introduction

Stemming [1] is a method that reduces morphologically similar variant of words into a single term called stems or roots without doing complete morphological

Vishal Gupta
University Institute of Engineering & Technology,
Panjab University Chandigarh, India
e-mail: vishal@pu.ac.in

S.M. Thampi, A. Gelbukh, and J. Mukhopadhyay (eds.), *Advances in Signal Processing and Intelligent Recognition Systems*, Advances in Intelligent Systems and Computing 264,
DOI: 10.1007/978-3-319-04960-1_7,

analysis. A stemming method stems the words talking, talked, talk and talks to the word, talk. The major task of a stemmer is to find root words that are not in original form and are hence absent in the dictionary. The stemmer after stemming finds the word in the dictionary. If a match is not found in dictionary, then it can be some incorrect word or a name, otherwise the word is correct. For any language in the world, stemmer is a basic linguistic resource required to develop any type of application in Natural Language Processing (NLP) with high accuracy like: machine translation, document classification, document clustering, text question answering, topic tracking, text summarization and keywords extraction etc. In information Retrieval system the Stemming is used to enhance performance. When a particular user searches the word aligning, he may want to find all documents containing words align and aligned as well.

The stemmer's development is based on language and it needs some linguistic knowledge about the language and spelling checker for that language. Most of the simple stemmer involves suffix stripping using a list of endings while some complex stemmer requires morphological knowledge of the language to find root word.

In comparison with English and other European languages very little work has been done for Indian regional languages in the area of stemming. Punjabi language has rich inflectional morphology in contrast to English. A verb in Punjabi has nearly 48 verb forms that are based on gender, number, tense, person and aspect value but a verb in English has only five different inflectional forms. For example different verb forms of 'eat' in English are eat, ate, eaten, eats and eating. Based on syntax and semantics Punjabi language completely differs from other languages of the world.

This paper concentrates on complete automatic stemming of Punjabi words covering Punjabi nouns, verbs, adjectives, adverbs, pronouns and proper names. A suffix list of 18 suffixes for Punjabi nouns and proper names and a number of other suffixes for Punjabi verbs, adjectives and adverbs and different stemming rules for Punjabi nouns, verbs, adjectives, adverbs, pronouns and proper names have been generated after analysis of Punjabi corpus. Some examples of stemmer for names/nouns in Punjabi with distinct endings are: **ਕੁੜੀਆਂ** kurīāṃ "girls" → **ਕੁੜੀ** kuri "girl" with ending ੀਆਂ īāṃ, **ਲੜਕੇ** larkē "boys" → **ਲੜਕਾ** muṇḍā "boy" with ending ੇ ē, ਫਿਰੋਜ਼ਪੁਰੋਂ phirōzpurōṃ → ਫਿਰੋਜ਼ਪੁਰ phirōzpur with ending ੋਂ ōṃ and **ਰੁੱਤਾਂ** ruttāṃ "seasons" → **ਰੁੱਤ** rutt "season" with ending ਾਂ āṃ etc. It is first time that complete Punjabi stemmer covering Punjabi nouns, verbs, adjectives, adverbs, pronouns, and proper names has been proposed and it will be useful for developing other Punjabi NLP applications with high accuracy. A portion of Punjabi stemmer for proper names and nouns has been implemented as part of Punjabi text summarizer in MS Access as back end and ASP.NET as front end with 87.37% efficiency

2 Related Work

Porter (1980) [1] proposed a rule based method for suffix stripping which is used widely for stemming of English words. The rule based removal of suffixes is suitable for less inflectional languages such as English and is useful in the area of retrieval of Information. The removal of suffixes leads to reduction in the total terms in Information Retrieval system which reduces the data complexity and is hence quiet useful. Jenkins and Smith (2005) [2] also proposed a rule based stemmer that works in steps and perform conservative stemming for both searching and indexing. It detects the word that need not be stemmed like proper nouns and stems only orthographically correct words. Mayfield and McNamee (2003) [3] suggested that N-gram can be selected as a stem is quiet useful and efficient method for certain languages. It analyzes the distribution of N-grams in the text where the value of N is selected empirically and even some high values of N such as 4 or 5 is selected. Massimo and Nicola (2003) [4] suggested a Hidden Markov based statistical method for stemming of words. It makes use of unsupervised training at the time of indexing and does not require any linguistic knowledge. The transition functions of finite-state automata of HMMs are described with the help of probability functions. Goldsmith (2001) [5] described the text in compact manner and suggested a method for morphology of language with the help of minimum description length (MDL). Creutz and lagus (2005) [6] performed morpheme segmentation using statistical maximum a posteriori (MAP) technique. The stemmer takes a corpus and gives segmentation of various word in it as an output which is similar to the linguistic morpheme segmentation

In comparison with English and other European languages very little work has been done for Indian regional languages in the area of stemming. Rao and Ramanathan (2003) [7] suggested a Hindi lightweight stemmer that performs longest match stripping of suffixes based on manually created suffix list. Islam et al. (2007) [8] proposed an approach similar to the approach used in Hindi lightweight stemmer for Bengali language that can also be used as spelling checker. It also performs longest match suffix stripping by making use of 72 endings for verbs, 22 for nouns and 8 for adjectives for Bengali language. Majumder et al. (2007) [9] proposed an approach YASS: Yet Another Suffix Stripper. It calculates four different string distances and then on the basis of these distances makes clusters of lexicon such that each cluster points to a single root word. Their stemmer did not require any prior language knowledge and suggested that stemming enhances recall of Information Retrieval Systems for regional languages. Dasgupta and Ng (2006) [10] performed unsupervised morphological analysis of Bengali wherein the words are segmented into suffixes, prefixes, stems without any knowledge of morphology of the language. Pandey and Siddiqui (2008) [11] proposed split-all method based unsupervised stemmer for Hindi. EMILLE corpus has been used for unsupervised training. The words are divided to provide n-gram (n=1, 2, 3 … k) endings, where k is the word length. And then endings and stem probability is calculated and compared. Majgaonker and

Siddiqui (2010) [12] proposed combination of stripping, rule based and statistical stripping for generation of suffix rules and developed an unsupervised stemmer for Marathi language.

3 Stemming of Punjabi Words

Punjabi stemmer finds the stem/root of the word and then checks it in the Punjabi, morph/dictionary, Punjabi noun dictionary or names dictionary for finding Punjabi nouns, verbs, adjectives, pronouns, or Punjabi names. Punjabi noun corpus has 37297 nouns. First the Punjabi Language stemmer segments the Punjabi text that is entered into words. Analysis of Punjabi corpus containing nearly 2.03 lakh distinct words with a total of 11.29 million words from famous Punjabi newspaper Ajit has been done.

3.1 Stemming for Punjabi Verbs

Thirty six (36) suffixes for Punjabi verbs were found after analyzing Punjabi news corpus. These suffixes are given in Table1.

Table 1 Suffix List for Punjabi verbs

Sr. No.	Punjabi verb suffix	Example Punjabi verb	Punjabi verb root word
1	ਾਉਦਿਆਂ	ਗਾਉਦਿਆਂ (singing) Gender: Feminine; Plural	ਗਾ (sing)
2	ਉਂਦਾ	ਪੜ੍ਹਾਉਂਦਾ (Teaches) Gender: Masculine; Singular	ਪੜ੍ਹਾ (Teach)
3	ਉਂਦੀ	ਪੜ੍ਹਾਉਂਦੀ (Teaches) Gender: Feminine; Singular	ਪੜ੍ਹਾ (Teach)
4	ਉਂਦੇ	ਪੜ੍ਹਾਉਂਦੇ (Teach) Gender: Masculine; Plural	ਪੜ੍ਹਾ (Teach)
5	ਉਣੀਆਂ	ਪੜ੍ਹਾਉਣੀਆਂ (Teach) Gender: Feminine; Plural	ਪੜ੍ਹਾ (Teach)
6	ਉਣਾ	ਪੜ੍ਹਾਉਣਾ (Teach) Gender: Masculine; Singular	ਪੜ੍ਹਾ (Teach)
7	ਉਣੀ	ਪੜ੍ਹਾਉਣੀ (Teach) Gender: Feminine; Singular	ਪੜ੍ਹਾ (Teach)
8	ਉਣੇ	ਪੜ੍ਹਾਉਣੇ (Teach) Gender: Masculine; Plural	ਪੜ੍ਹਾ (Teach)
9	ਂਦਾ	ਪੜ੍ਹਾਂਦਾ (Teaches) Gender: Masculine; Singular	ਪੜ੍ਹਾ (Teach)
10	ਂਦੀ	ਪੜ੍ਹਾਂਦੀ (Teaches) Gender: Feminine; Singular	ਪੜ੍ਹਾ (Teach)
11	ਂਦੇ	ਪੜ੍ਹਾਂਦੇ (Teach) Gender: Masculine; Plural	ਪੜ੍ਹਾਂ (Teach)
12	ਿਆ	ਭੱਜਿਆ (ran) Gender: Masculine;: Singular	ਭੱਜ (run)
13	ੀਆਂ	ਭੱਜੀਆਂ (ran) Gender: Feminine; Singular	ਭੱਜ (run)
14	ਇਆ	ਪੜ੍ਹਾਇਆ (Taught) Gender:X; Singular/Plural	ਪੜ੍ਹਾ (Teach)

Table 1 *(continued)*

15	ਈ	ਪੜ੍ਹਾਈ (studying) Gender: X; Singular	ਪੜ੍ਹਾ (study)
16	ਏ	ਪੜ੍ਹਾਏ (Taught) Gender: X; Singular/Plural	ਪੜ੍ਹਾ (Teach)
17	ਵਾਂ	ਪੜ੍ਹਾਵਾਂ (will teach) Gender: Masculine; Singular	ਪੜ੍ਹਾ (Teach)
18	ਵੀਂ	ਪੜ੍ਹਾਵੀਂ (will teach) Gender: Feminine; Singular	ਪੜ੍ਹਾ (Teach)
19	ਵੇਂ	ਪੜ੍ਹਾਵੇਂ (will teach) Gender:X; Plural	ਪੜ੍ਹਾ (Teach)
20	ਣਾ	ਪੜ੍ਹਣਾ (read) Gender:X; Singular	ਪੜ੍ਹ (read)
21	ਣੀ	ਖਾਣੀ (ate) Gender:Feminine; Plural	ਖਾ (eat)
22	ਣੇ	ਖਾਣੇ (eat) Gender:Masculine; Plural	ਖਾ (eat)
23	ਦਾ	ਪੜ੍ਹਦਾ (reads) Gender:Masculine; Singular	ਪੜ੍ਹ (read)
24	ਦੀ	ਪੜ੍ਹਦੀ (reads) Gender:Feminine; Singular	ਪੜ੍ਹ (read)
25	ਦੇ	ਪੜ੍ਹਦੇ (read) Gender:X; Plural	ਪੜ੍ਹ (read)
26	ੀ	ਭੱਜੀ (ran) Gender: Feminine; Singular	ਭੱਜ (run)
27	ਾ	ਭੱਜਾ (ran) Gender: Masculine; Singular	ਭੱਜ (run)
28	ੇ	ਭੱਜੇ (should run) Gender: X; Plural	ਭੱਜ (run)
29	ੇ	ਭੱਜੇ (escaped) Gender:Masculine; Plural	ਭੱਜ (escape)
30	ੇਗਾ	ਭੱਜੇਗਾ (will run) Gender: Masculine; Singular	ਭੱਜ (run)
31	ੇਗੀ	ਭੱਜੇਗੀ (will run) Gender: Feminine; Singular	ਭੱਜ (run)
32	ੇਗੇ	ਭੱਜੇਗੇ (will run) Gender: X; Plural;	ਭੱਜ (run)
33	ਣਗੇ	ਭੱਜਣਗੇ (will run) Gender: Masculine; Plural	ਭੱਜ (run)
34	ਣਗੀਆਂ	ਭੱਜਣਗੀਆਂ(will run) Gender: Feminine; Plural	ਭੱਜ (run)
35	ਏਗਾ	ਗਾਏਗਾ (will sing) Gender: Masculine; Singular	ਗਾ (sing)
36	ਏਗੀ	ਗਾਏਗੀ (will sing) Gender: Feminine; Singular	ਗਾ (sing)

Algorithm for Punjabi verb stemming

Algorithm Input:
ਪੜ੍ਹਦੇ, ਭੱਜਣਗੇ, ਗਾਏਗਾ

Algorithm Output:
ਪੜ੍ਹ, ਭੱਜ, ਗਾ

Step 1: The Punjabi word is given as input to Punjabi verb stemmer.

Step 2: If the entered word matches with a word in Punjabi dictionary then it is returned as output. Otherwise if suffix of the entered word lies in any of the verb suffixes: ਾਉਂਦਆਂ, ਉਂਦਾ, ਉਂਦੀ, ਉਂਦੇ, ਉਣੀਆਂ, ਉਣਾ, ਉਣੀ, ਉਣੇ, ਂਦਾ, ਂਦੀ, ਂਦੇ, ਿਆ,

ੀਅਾਂ, ਇਆ, ਈ, ਏ, ਵਾਂ, ਵੀਂ, ਵੇਂ, ਣਾ, ਣੀ, ਣੇ, ਦਾ, ਦੀ, ਦੇ, ੀ, ਾ, ੇ, ੇ, ੇਗਾ, ੇਗੀ, ੇਗੇ, ਣਗੇ, ਣਗੀਆਂ, ਏਗਾ, ਏਗੀ then delete ending from end and search the stemmed Punjabi word in Punjabi dictionary. If resultant word is found then it is returned as output.

Step 3: If the resultant Punjabi word is not found in Punjabi dictionary then verb stemmer is unable to locate the actual verb and error message is displayed.

3.2 Stemming for Punjabi Adjectives/Adverbs

Five suffixes for Punjabi adjectives/adverbs were found after analyzing the Punjabi news corpus. The suffix list is given below:

Table 2 Suffix List for Punjabi Adjectives/Adverbs

Sr. No.	Punjabi Adjective / Adverb suffix	Example Punjabi Adjective/Adverb	Punjabi Adjective/Adverb root word
1	ਆਂ	ਚੰਗੀਆਂ(good) Gender: Feminine; Plural	ਚੰਗੀ (good)
2	ਈਆਂ	ਉਚਾਈਆਂ (Heights) Gender:X; Plural	**ਉਚਾਈ** (Height)
3	ਿਆ	ਸੋਹਣਿਆ (Handsome)Gender: Masculine; Singular	ਸੋਹਣਾ (Handsome)
4	ਿਏ	ਸੋਹਣਿਏ (Beautiful) Gender: Feminine; Singular	ਸੋਹਣੀ (Beautiful)
5	ਿਓ	ਸੋਹਣਿਓ (Beautiful) Gender: X; Plural	ਸੋਹਣਾ (Beautiful)

Algorithm for Punjabi adjective/adverb stemming

Algorithm Input:
ਚੰਗੀਆਂ, ਸੋਹਿਣਆ, ਸੋਹਿਣਏ

Algorithm Output:
ਚੰਗੀ, ਸੋਹਣਾ, ਸੋਹਣੀ

Step 1: The Punjabi word is given as input to Punjabi adjective/adverb stemmer.

Step 2: If the entered Punjabi word matches with a word in Punjabi dictionary then it is returned as output. Otherwise if ending of entered word is ਆਂ or ਈਆਂ then remove the respective ending from the end and then search the stemmed Punjabi word in Punjabi dictionary. If the word is found then it is returned as output.

Step 3: If ending of entered word is ਿਆ or ਿਓ then delete the ending and add ਾ at the end and search the stemmed Punjabi word in Punjabi dictionary. If resultant word matches with a word in the dictionary then it is returned as output.

Step 4: If ending of entered word is ਿਏ or ਿਓ then delete and add ੀ at the end and search the stemmed Punjabi word in Punjabi dictionary. If resultant word is found then it is returned as output.

Step 5: If resultant word is not found then adjective/adverb stemmer is unable to locate the actual adverb/adjective and error message is displayed.

3.3 Stemming for Punjabi Pronouns

13 suffixes for Punjabi pronouns were found after analyzing the Punjabi news corpus these are as given in Table 3.

Table 3 Suffix List for Punjabi Pronouns

Sr. No.	Punjabi Adjective/Adverb suffix	Example Punjabi Adjective/Adverb	Punjabi Adjective/Adverb root word
1	ੇ	ਇਹੇ (This one) Gender: X; Singular	ਇਹ (This)
2	ੇ	ਸਾਡੇ (ours) Gender:X; Plural	ਸਾਡਾ (our)
3	ਾਂ	ਆਪਾਂ (we) Gender:X; Plural	ਆਪ (ourselves)
4	ਆਂ	ਤੁਹਾਡੀਆਂ (Your's) Gender:X; Singular	ਤੁਹਾਡੀ (Your)
5	ਦਾ	ਉਹਦਾ (His) Gender:Masculine; Singular	ਉਹ (He)
6	ਦੀ	ਉਹਦੀ (Her) Gender:Feminine; Singular	ਉਹ (She)
7	ਦੇ	ਉਹਦੇ (His/her) Gender:X ; Plural	ਉਹ (He/She)
8	ਜੀ	ਆਪਜੀ (Your's) Gender:X; Singular	ਆਪ (Your)
9	ੀ	ਇਹੀ (This only) Gender:X; Singular	ਇਹ (This)
10	ਨੀ	ਇਹਨੀ (Their) Gender:Feminine; Singular	ਇਹ (This)
11	ਨਾ	ਇਹਨਾ (Their) Gender:Masculine; Singular	ਇਹ (This)
12	ਨੇ	ਇਹਨੇ (His) Gender: Masculine; Singular	ਇਹ (This)
13	ਨਾਂ	ਇਹਨਾਂ (Theirs) Gender:X; Plural	ਇਹ (This)

Algorithm for Punjabi adjective/adverb stemming

Algorithm Input:
ਇਹਨਾ, ਆਪਣੇ, ਆਪਜੀ

Algorithm Output:
ਇਹ, ਆਪਣਾ, ਆਪ

Step 1: The Punjabi word is entered as input to Punjabi Pronoun stemmer.

Step 2: If entered word is found in dictionary then it is returned as output. Otherwise if ending of entered word is ੇ or ੍ਹੇ or ਆਂ or ਦਾ or ਦੀ or ਦੇ or ਜੀ or ੀ or ਨੀ or ਨਾ or ਨੇ or ਨਾਂ then remove the ending from the end and then search the resultant stemmed Punjabi word in Punjabi dictionary. If word is found then it is returned as output.

Step 3: If ending of entered word is ੇ then remove the ending from the end and add ੍ਹ at the end and then search the stemmed Punjabi word in Punjabi dictionary. If stemmed word is found then it is returned as output.

Step 4: If resultant word is not found in dictionary then Punjabi pronoun stemmer is unable to locate the actual pronoun and error message is displayed.

3.4 Stemming for Punjabi Proper Names and Nouns

Punjabi stemmer for names/nouns [14][15], the goal is to locate the root words and then the root words are matched with the words in dictionary for Punjabi names and nouns dictionary. 18 endings were listed for Punjabi names/nouns after analyzing the Punjabi corpus like ੍ਹੇ āṃ, ਿਆਂ iāṃ, ੁਆਂ ūāṃ and ੀਆਂ īāṃ etc. and distinct rules for Punjabi name/noun stemming have been developed. Some examples of Punjabi names/nouns for distinct endings are:

ਕੁੜੀਆਂ kurīāṃ "girls" → **ਕੁੜੀ** kuri "girl" with ending ੀਆਂ īāṃ,, ਫਿਰੋਜ਼ਪੁਰੋਂ phirōzpurōṃ → ਫਿਰੋਜ਼ਪੁਰ phirōzpur with ending ੋਂ ōṃ, **ਲੜਕੇ** larkē "boys" → **ਲੜਕਾ** muṇḍā "boy" with ending ੇ ē and **ਰੁੱਤਾਂ** ruttāṃ "seasons" → **ਰੁੱਤ** rutt "season" with ending ੍ਹੇ āṃ etc.

Algorithm for Punjabi Nouns/ Proper Names:

Algorithm Input:
ਲੜਕੇ larkē "boys" and **ਰੁੱਤਾਂ** ruttāṃ "seasons"
Algorithm Output:
ਲੜਕਾ muṇḍā "boy" and **ਰੁੱਤ** rutt "season"

Step 1: The Punjabi word is given as input to Punjabi noun/proper name stemmer.

Step 2: If entered Punjabi word matches with a word in names dictionary/ nouns dictionary then it is returned as output. Else If ending of entered word is ਆਂ āṃ (

in case of ੂਆਂ ūāṃ, ਿਆਂ iāṃ and ੀਆਂ īāṃ), ਏ ē (in case of ੀਏ īē), ਓ ō (in case of ੀਓ Īō), ਆ ā (in case of ੀਆ ā, ਈਆ īā), ਵਾਂ vāṃ, ਈ ī , ਾਂ āṃ, ੀਂ īṃ, ਜ/ਜ਼/ਸ ja/z/s and ੋਂ ōṃ then remove the ending and then go to Step 5.

Step 3: If entered has ending ੇ ō, ਿਓ iō, ੇ ē, ਿਆ iā and ਿਉਂ iuṃ then remove the ending and add kunna at the end and then go to Step 5.

Step4: Else entered word is either incorrect word or incorrect word.

Step 5: Resultant Stemmed word is matched with the words of Punjabi names-dictionary/noun-morph. If it matches, then it is Punjabi-name or noun.

3.5 Overall Algorithm for Complete Stemming of Punjabi Words

Algorithm Input:
ਮੁੰਡੇ (boys), ਫੁੱਲਾਂ (flowers), ਇਹਨਾ (Theirs), ਆਪਣੇ (yours), ਪੜ੍ਹਦਾ (reads), ਭੱਜਣਗੇ (will run),
Algorithm Output:
ਮੁੰਡਾ (boy), ਫੁੱਲ (flower), ਇਹ (this), ਆਪ (your), ਪੜ੍ਹ (read), ਭੱਜ (run)

The overall algorithm for Complete stemming of Punjabi words is given below:

Step 0: The Punjabi word is given as input to Punjabi Stemmer.

Step 1: If entered word is found in dictionary/ noun morph/ names dictionary then it is returned as output. Else go to step 2.
Step 2: Apply stemmer for Proper Names and Nouns [14] [15]. If the stemmed word after stemming is located in the Punjabi names dictionary/ noun morph then it is given as output. Else go to step 3.

Step 3: Apply stemmer for Punjabi verbs. If the stemmed word after stemming is located in the Punjabi dictionary then it is given as output. Else go to step 4.

Step 4: Apply stemmer for Punjabi Adjectives/Adverbs. If the stemmed word after stemming is located in the Punjabi dictionary then it is given as output. Else go to step 5.

Step 5: Apply stemmer for Punjabi pronouns. If the stemmed word after stemming is located in the Punjabi dictionary then it is given as output. Else stemmer is unable to locate the actual root word and error message is displayed.

4 Results and Discussions

Punjabi stemming algorithm has been applied on fifty documents of news corpus of Punjabi and efficiency is found to be 87.37%. Table 4 shows efficiency and error percentage analysis of Punjabi verb stemmer, adverb/adjective stemmer, noun/proper name stemmer, pronoun stemmer and overall Punjabi stemmer.

Table 4 Accuracy and Error Percentage Analysis of Punjabi Stemmer

Punjabi Verb Stemmer	Adjective/Adverb Punjabi Stemmer	Punjabi Pronoun Stemmer	Punjabi Nouns/Proper Names	Overall Accuracy of Punjabi Stemmer
Accuracy: 84.67% Errors:15.33%	Accuracy: 85.87% Errors: 14.13%	Accuracy: 90.88% Errors: 9.12%	Accuracy: 87.37% Errors: 12.63%	Accuracy: 87.2% Errors: 12.8%

Errors are because of non existence of some of suffixes in Punjabi stemmer and also due to absence of certain root words in Punjabi dictionary. The efficiency of stemmer is words that are stemmed correctly is to the total number of words stemmed by stemmer. In the same way efficiency of each rule of stemmer is number of correct outputs of that rule is to total outputs produced by the rule.

In case of stemmer for proper names and nouns [14][15], three categories of errors are possible: 1) Syntax mistakes 2) Dictionary errors 3) Violation of stemming-rules. Syntax errors are the errors due to incorrect syntax and it occurs due to typing mistakes. Dictionary errors are the errors due to dictionary i.e. the stemmed word is not present in names/noun dictionary but it is a noun or a name actually. The typing

Errors are 0.45%, dictionary errors are 2.4% and rules violation errors are 9.78%.

Examples of rules violation errors are as given below:

The word ਹਲਕੇ halkē “light weight” is an adjective and ਬਦਲੇ badlē “in lieu of” is an adverb. The words are not matched with the words in names/nouns dictionary, though they come under ੇ ē rule which takes them as noun, which is not true.

Dictionary errors examples are given below:

The words like ਮੁਨਾਫ਼ਿਆਂ munāphaiāṃ “profits” and ਪ੍ਰਦੇਸਾਂ pradēsāṃ “foreign” are nouns but are absent in Punjabi dictionary /noun morph. The words come under ਾਂ āṃ rule and ਿਆਂ iāṃ rule of stemmer and after noun stemming their root words ਮੁਨਾਫ਼ਾ munāphā “profit” and ਪ੍ਰਦੇਸਾਂ pradēsāṃ “foreign” are also absent in dictionary /noun morph.

Syntax errors examples are as given below:

It is possible that the spellings of certain noun words are typed incorrectly, words such as ਆਕ੍ਰਤੀ ākrtī “shape” and ਚਿਡੀਆ chiḍīā “sparrow” instead of their correct spellings ਚਿੜੀਆ chiṛīā “sparrow” and ਆਕ੍ਰਿਤੀ ākritī “shape” respectively.

5 Conclusions

In this paper, I have proposed the Complete Punjabi language stemmer covering Punjabi verbs, Punjabi nouns, Punjabi proper names, Punjabi adjectives/adverbs, and Punjabi pronouns. The resources used like Punjabi noun morph, Punjabi proper names list etc. had to be done from scratch because no work had been done in the area. A detailed analysis of news corpus is done using manual and automatic tools for development of these resources. The resources for Punjabi language are developed for first time and they can be useful in a lot of Punjabi NLP applications. A part of stemmer in Punjabi covering stemmer for Punjabi proper names and nouns [14][15] is used successfully in Text Summarization for Punjabi language [16].

References

1. Porter, M.: An Algorithm for Suffix Stripping Program 14, 130–137 (1980)
2. Jenkins, M., Smith, D.: Conservative Stemming for Search and Indexing. In: Proceedings of SIGIR 2005 (2005)
3. Mayfield, J., McNamee, P.: Single N-gram stemming. In: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 415–416 (2003)
4. Massimo, M., Nicola, O.: A Novel Method for Stemmer Generation based on Hidden Markov Models. In: Proceedings of the Twelfth International Conference on Information and Knowledge Management, pp. 131–138 (2003)
5. Goldsmith, J.A.: Unsupervised Learning of the Morphology of a Natural Language. Computational Linguistics 27, 153–198 (2001)
6. Creutz, M., Lagus, K.: Unsupervised Morpheme Segmentation and Morphology Induction from Text Corpora using Morfessor 1.0. Publications of Computer and Information Science, Helsinki University of Technology (2005)
7. Ramanathan, A., Rao, D.D.: A Lightweight Stemmer for Hindi. In: Proceedings of Workshop on Computational Linguistics for South-Asian Languages, EACL (2003)
8. Islam, M.Z., Uddin, M.N., Khan, M.: A Light Weight Stemmer for Bengali and its Use in Spelling Checker. In: Proceedings of. 1st Intl. Conf. on Digital Comm. and Computer Applications (DCCA 2007), Irbid, Jordan, pp. 19–23 (2007)
9. Majumder, P., Mitra, M., Parui, S.K., Kole, G., Datta, K.: YASS Yet Another Suffix Stripper. Association for Computing Machinery Transactions on Information Systems 25, 18–38 (2007)
10. Dasgupta, S., Ng, V.: Unsupervised Morphological Parsing of Bengali. Language Resources and Evaluation 40, 311–330 (2006)

11. Pandey, A.K., Siddiqui, T.J.: An Unsupervised Hindi Stemmer with Heuristic Improvements. In: Proceedings of the Second Workshop on Analytics For Noisy Unstructured Text Data, vol. 303, pp. 99–105 (2008)
12. Majgaonker, M.M., Siddiqui, T.J.: Discovering Suffixes: A Case Study for Marathi Language. Proceedings of International Journal on Computer Science and Engineering 2, 2716–2720 (2010)
13. Suba, K., Jiandani, D., Bhattacharyya, P.: 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) IJCNLP 2011, Chiang Mai, Thailand, pp. 1–8 (2011)
14. Gupta, V., Lehal, G.S.: Punjabi Language Stemmer for Nouns and Proper Names. In: Proceedings of the 2nd Workshop on South and Southeast Asian Natural Language Processing (WSSANLP) IJCNLP 2011, Chiang Mai, Thailand, pp. 35–39 (2011)
15. Gupta, V., Lehal, G.S.: Preprocessing Phase of Punjabi Language Text Summarization. In: Singh, C., Singh Lehal, G., Sengupta, J., Sharma, D.V., Goyal, V. (eds.) ICISIL 2011. CCIS, vol. 139, pp. 250–253. Springer, Heidelberg (2011)
16. Gupta, V., Lehal, G.S.: Automatic Punjabi Text Extractive Summarization System. In: Proceedings of International Conference on Computational Linguistics COLING 2012, pp. 191–198 (2012)