Punjabi Language Stemmer for nouns and proper names

Vishal Gupta
Assistant Professor, UIET
Panjab University Chandigarh
Vishal_gupta100@yahoo.co.in

Gurpreet Singh Lehal
Professor, Department of Computer Science,
Punjabi University Patiala
gslehal@yahoo.com

Abstract

This paper concentrates on Punjabi language noun and proper name stemming. The purpose of stemming is to obtain the stem or radix of those words which are not found in dictionary. If stemmed word is present in dictionary, then that is a genuine word, otherwise it may be proper name or some invalid word. In Punjabi language stemming for nouns and proper names, an attempt is made to obtain stem or radix of a Punjabi word and then stem or radix is checked against Punjabi noun and proper name dictionary. An in depth analysis of Punjabi news corpus was made and various possible noun suffixes were identified like ੀਆਂ īāṃ, ਿਆਂ iāṃ, ੂਆਂ ūāṃ, ਾਂ āṃ, ੀਏ īē etc. and the various rules for noun and proper name stemming have been generated. Punjabi language stemmer for nouns and proper names is applied for Punjabi Text Summarization. The efficiency of Punjabi language noun and Proper name stemmer is 87.37%.

1 Introduction

stemming is the process for reducing inflected or sometimes derived words to their stem, base or root form, generally a written word form. The stem need not be identical to the morphological root of the word, it is usually sufficient that related words map to the same stem, even if this stem is not in itself a valid root. A stemmer for English, for example, should identify the string cats and possibly catlike, catty etc. as based on the root cat, and stemmer, stemming, stemmed as based on stem. A stemming algorithm reduces the words fishing, fished, fish, and fisher to the root word, fish. Stemming is an operation that conflates morphologically similar terms into a single term without doing complete morphological analysis. Stemming (Haidar et al., 2006) is used in information retrieval systems to improve performance. Additionally, this operation reduces the number of terms in the information retrieval system, thus decreasing the size of the index files.

In Punjabi language stemming (Mandeep et al.,2009) for nouns and proper names, an attempt is made to obtain stem or radix of a Punjabi word and then stem or radix is checked against Punjabi noun morph and proper names list. An in depth analysis of Punjabi news corpus was made and various possible noun suffixes were identified like ੀਆਂ īāṃ, ਿਆਂ iāṃ, ੂਆਂ ūāṃ, ਾਂ āṃ, ੀਏ īē etc. and the various rules for noun and proper name stemming have been generated. Punjabi language stemmer for nouns and proper names is applied for Punjabi Text Summarization. Text Summarization is the process of condensing the source text into shorter version. Those sentences containing Punjabi language nouns or proper names are important.

2 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 perhaps the most widely used algorithm for English stemming. Both of these stemmers are rule based and are best suited for less inflectional languages like English. (Goldsmith, 2001) proposed an algorithm for the morphology of a language based on the minimum description length (MDL) framework which focuses on representing the data in as compact manner as possible. (Creutz, 2005) uses probabilistic maximum a posteriori (MAP) formulation for morpheme segmentation.

Not much work has been reported for stemming for Indian languages compared to English and other European languages. The earliest work reported by (Ramanathan and Rao, 2003) used a hand crafted suffix list and performed longest match stripping for building a Hindi stemmer. (Majumder et al., 2007) developed statistical approach YASS: Yet Another Suffix Stripper which uses a clustering based approach based on string distance measures and requires no linguis-

Proceedings of the 2nd Workshop on South and Southeast Asian Natural Language Processing (WSSANLP), IJCNLP 2011, , pages 35–39,
Chiang Mai, Thailand, November 8, 2011.

tic knowledge. They concluded that stemming improves recall of IR systems for Indian languages like Bengali. (Dasgupta and Ng, 2007) worked on morphological parsing for Bengali. (Pandey and Siddiqui, 2008) proposed an unsupervised stemming algorithm for Hindi based on (Goldsmith, 2001) approach.

3 Punjabi Language stemmer for Nouns and Proper names

In Punjabi language stemming (Md. et al., 2007) for nouns and proper names, an attempt is made to obtain stem or radix of a Punjabi word and then stem or radix is checked against Punjabi noun morph and Proper names list. An in depth analysis of corpus was made and the possible noun and proper name suffixes (Praveen et al.,2003) were identified (Table1) and the various rules for Punjabi word noun stemming have been generated.

Table 1. Punjabi language noun/Proper name suffix list

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

Proper names are the names of person, place and concept etc. not occurring in Punjabi Dictionary. Proper Names play an important role in deciding a sentence's importance. From the Punjabi corpus, 17598 words have been identified as proper names. The percentage of these proper names words in the Punjabi corpus is about 13.84 %. Some of Punjabi language proper names are given in Table2.

Table 2. Some of Punjabi language proper names

ਅਕਾਲੀ akālī	ਲੁਧਿਆਣਾ ludhiāṇā
ਬਾਦਲ bādal	ਪਟਿਆਲਾ paṭiālā
ਜਲੰਧਰ jalndhar	ਭਾਜਪਾ bhājapā

Algorithm of Punjabi language stemmer for nouns and proper names is given below:

Stemming Algorithm

The algorithm of Punjabi language stemmer for nouns and proper names proceeds by segmenting the source Punjabi text into sentences and words. For each word of every sentence follow following steps:

- Step 1 : If current Punjabi word ends with ੀਆਂ īāṃ then remove ਆਂ āṃ from end.
- Step 2 : Else If current Punjabi word ends with ਿਆਂ iāṃ then remove ਆਂ āṃ from end.
- Step 3 : Else If current Punjabi word ends with ੂਆਂ ūāṃ then remove ਆਂ āṃ from end.
- Step 4 : Else If current Punjabi word ends with ੀਏ īē then remove ਏ ē from end.
- Step 5 : Else If current Punjabi word ends with ਈ ī then remove ਈ ī from end.
- Step 6 : Else If current Punjabi word ends with ੇ ē then remove ੇ ē from end and add kunna at the end
- Step 7 : Else If current Punjabi word ends with ੀਓ Īō then remove ਓ ō from end.
- Step 8 : Else If current Punjabi word ends with ਿਓ iō then remove ਿਓ iō from end and add kunna at the end
- Step 9 : Else If current Punjabi word ends with ਵਾਂ vāṃ then remove ਵਾਂ vāṃ from end.
- Step 10 : Else If current Punjabi word ends with ਾਂ āṃ then remove ਾਂ āṃ from end.
- Step 11 : Else If current Punjabi word ends with ੋਂ ōṃ then remove ੋਂ ōṃ from end.
- Step 12 : Else If current Punjabi word ends with ੋ ō then remove ੋ ō from end and add kunna at the end
- Step 13 : Else If current Punjabi word ends with ੀਂ īṃ then remove ੀਂ īṃ from end.
- Step 14 : Else If current Punjabi word ends with ਿਉਂ iuṃ then remove ਿਉਂ iuṃ from end and add kunna at the end.

- Step 15: Else If current Punjabi word ends with ੀਆ ā then remove ਆ ā from end.
- Step 16: Else If current Punjabi word ends with ਿਆ ā then remove ਿਆ ā from end and add kunna at the end.
- Step 17: Else If current Punjabi word ends with ਈਆ īā then remove ਆ ā from end.
- Step 18: Else If current Punjabi word ends with ਜ/ਜ਼/ਸ ja/z/s then remove ਜ/ਜ਼/ਸ ja/z/s from end.
- Step 19: Current Punjabi Stemmed word is checked against Punjabi noun morph or Proper names list. If found, It is Punjabi noun or Punjabi Proper name.

Algorithm Input: ਫੁੱਲਾਂ phullāṃ (Flowers) and ਲੜਕੀਆਂ laṛkīāṃ (Girls)

Algorithm Output: ਫੁੱਲ phull (Flower) and ਲੜਕੀ laṛkī (Girl)

Some results of Punjabi language stemmer for nouns and Proper names for various possible suffixes are given in table3.

Table3.Results of Punjabi language Noun/Proper name stemmer

Punjabi Noun/Proper Name word	**Stem word**	**suffix**
ਕਸਾਈਆ Kasāīā	ਕਸਾਈ kasāī	ਈਆ īā
ਫਿਰੋਜ਼ਪੁਰੋਂ phirōzpurōṃ	ਫਿਰੋਜ਼ਪੁਰ phirōzpur	ੋਂ ōṃ
ਲੜਕੀਆਂ laṛkīāṃ	ਲੜਕੀ laṛkī	ੀਆਂ īāṃ
ਫੁੱਲਾਂ phullāṃ	ਫੁੱਲ phull	ਾਂ āṃ
ਲੜਕਿਆਂ laṛkiāṃ	ਲੜਕਾ laṛkā	ਿਆਂ iāṃ
ਮੁੰਡੇ muṇḍē	ਮੁੰਡਾ muṇḍā	ੇ ē
ਲੜਕਿਓ laṛkīō	ਲੜਕਾ laṛkā	ਿਓ iō
ਘਰੀਂ gharīṃ	ਘਰ ghar	ੀਂ īṃ
ਪਰਾਂਦੇ	ਪਰਾਂਦਾ	ੇ
parāndē	parāndā	ē
ਮਾਹੀਆ māhīā	ਮਾਹੀ Māhī	ੀਆ Īā
ਭਾਸ਼ਾਵਾਂ bhāshāvāṃ	ਭਾਸ਼ਾ bhāshā	ਵਾਂ vāṃ
ਆਗੂਆਂ āgūāṃ	ਆਗੂ āgū	ੂਆਂ ūāṃ
ਲੜਕੇ laṛkō	ਲੜਕਾ laṛkā	ੇ ō
ਲੜਕੀਏ laṛkīē	ਲੜਕੀ laṛkī	ੀਏ īē
ਲੜਕੀਓ laṛkīō	ਲੜਕੀ laṛkī	ੀਓ Īō
ਲੜਕਿਆ laṛkiā	ਲੜਕਾ laṛkā	ਿਆ iā
ਮੋਗਿਉਂ mōgiuṃ	ਮੋਗਾ mōgā	ਿਉਂ iuṃ
ਭਾਸ਼ਾਈ bhāshāī	ਭਾਸ਼ਾ bhāshā	ਈ Ī
ਸਟੂਡੈਂਟਸ saṭūḍaiṇṭas	ਸਟੂਡੈਂਟ saṭūḍaiṇṭa	ਸ s

4 Results and Discussions

An In depth analysis of output of Punjabi language stemmer for nouns and proper names has been done over 50 Punjabi documents of Punjabi news corpus of 11.29 million words. The efficiency of Punjabi language noun and Proper name stemmer is 87.37%, which is tested over 50 Punjabi news documents of corpus and is ratio of actual correct results to total produced results by stemmer. Table4 gives accuracy percentage of various rules of stemmer which is ratio of correct results to total results produced under that rule, tested over 50 news documents. Table5 gives the error percentage analysis of various rules of Punjabi language stemmer. Errors are due to rules violation or dictionary errors or due to syntax mistakes. Dictionary errors are those errors in which, after stemming, stem word is not present in noun morph or Proper names list, but actually it is noun. Syntax errors are those errors, in which input Punjabi word is having some syntax mistake, but actually that word falls under any of stemming rules. Overall error percentage, due to rules violation is 9.78%, due to dictionary mistakes is 2.4% and due to spelling mistakes is

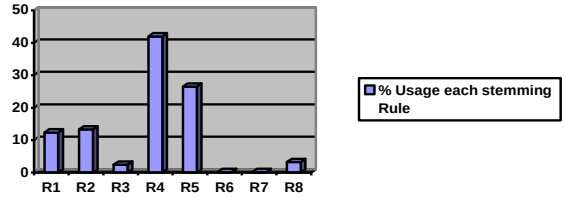
0.45%. Some of rules have not been taken in these table as we have not detected any accurate or in accurate words for those rules in the input Punjabi text.

Table 4. Accuracy %age analysis of rules of Punjabi stemmer for Nouns and Proper names

Punjabi Noun Suffix Rules	**Accuracy Percentage of Correct words detected**
Rule1 ੀਆਂ īāṃ	86.81%
Rule2 ਿਆਂ iāṃ	95.91%
Rule3 ੂਆਂ ūāṃ	94.44%
Rule4 ਾਂ āṃ	92.55%
Rule5 ੇ ē	57.43%
Rule6 ੀਂ īṃ	100%
Rule7 ੋਂ ōṃ	100%
Rule8 ਵਾਂ vāṃ	79.16%

Table 5. Error %age analysis of various rules of Punjabi stemmer for nouns and proper names

Punjabi Noun Suffix Rules	**% age of In Correct words due to rules Violation**	**% age of In Correct words due to dictionary mistakes**	**% age of In Correct words due to spelling mistakes**
Rule1 ੀਆਂ īāṃ	79.7%	20.30%	0%
Rule2 ਿਆਂ iāṃ	86.65%	13.35%	0%
Rule3 ੂਆਂ ūāṃ	0%	100%	0%
Rule4 ਾਂ āṃ	68.71%	18.25%	13.04%
Rule5 ੇ ē	82.21%	17.79%	0%
Rule6 ੀਂ īṃ	0%	0%	0%
Rule7 ੋਂ ōṃ	0%	0%	0%
Rule8 ਵਾਂ vāṃ	89%	11%	0%



Graph 1 Percentage Frequency of Various Stemming Rules

Graph1 depicts the percentage usage of the stemming rules. As can be seen, Rule 4 and Rule 5 are the most frequently used stemming rules. Unfortunately Rule 5 has a low accuracy with 42.57% of words being wrongly stemmed by this rule. Actually some of Punjabi words like ਹੱਸੇ hassē (laugh), ਹਲਕੇ halkē (area), ਮੌਕੇ moukē (oppurtinity) and ਬਦਲੇ badlē (revenge) are not nouns and are not present in noun morph, but they fall under Rule5 of stemmer which makes them noun after stemming, which is not true.If after stemming, root word is still not present in dictionary then, that word may be a proper name or may be syntactically wrong word which can be ignored.

4 Conclusions

In this paper, we have discussed the Punjabi language stemmer for nouns and proper names. Most of the lexical resources used such as Punjabi proper names list, Punjabi noun morph etc. had to be developed from scratch as no work had been done in that direction. For developing these resources an in depth analysis of Punjabi corpus, Punjabi dictionary (Gurmukh et al.,1999) and Punjabi morph had to be carried out using manual and automatic tools. This the first time some of these resources have been developed for Punjabi and they can be beneficial for developing other Natural Language Processing applications in Punjabi. Punjabi language stemmer for nouns and proper names is successfully used in Punjabi language Text Summarization.

References

Creutz, Mathis, and Krista Lagus. 2005. *Unsupervised morpheme segmentation and morphology induction from text corpora using Morfessor 1.0*. Technical Report A81, Publications in Computer and Information Science, Helsinki University of Technology.

Dasgupta, Sajib, and Vincent Ng. 2006. *Unsupervised Morphological Parsing of Bengali*. Language Resources and Evaluation, 40(3-4):311-330.

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.

Goldsmith, John A. 2001.*learning of the morphology of a natural language*, Computational Linguistics, 27(2):153-198.

Gurmukh Singh, Mukhtiar Singh Gill and S.S. Joshi. 1999. *Punjabi to English Bilingual Dictionary*. Punjabi University Patiala.

Majumder, Prasenjit, Mandar Mitra, Swapan K. Parui, Gobinda Kole, Pabitra Mitra, and Kalyankumar Datta. 2007. *YASS: Yet another suffix stripper*. Association for Computing Machinery Transactions on Information Systems, 25(4):18-38.

Mandeep Singh Gill, G.S. Lehal and S.S. Joshi. 2009. *Part of Speech Tagging for Grammar Checking of Punjabi*. The Linguistic Journal Volume 4 Issue 1, 6-21.

Md. Zahurul Islam, Md. Nizam Uddin and Mumit Khan. 2007. *A light weight stemmer for Bengali and its Use in spelling Checker*. Proc. 1st Intl. Conf. on Digital Comm. and Computer Applications (DCCA07), Irbid, Jordan, March 19-23.

Pandey, Amaresh K., and Tanveer J. Siddiqui. 2008. An unsupervised Hindi stemmer with heuristic improvements. In Proceedings of the Second Workshop on Analytics For Noisy Unstructured Text Data, 303:99-105.

Porter, Martin F. 1980. *An algorithm for suffix stripping Program*, 14(3):130-137.

Praveen Kumar, Shrikant Kashyap, Ankush Mittal and Sumit Gupta. 2003. *A query answering system for E learning Hindi documents*. South Asian Language Review, VOL.XIII, Nos 1&2.

Ramanathan, Ananthakrishnan, and Durgesh D. Rao. 2003. *A Lightweight Stemmer for Hindi*, Workshop on Computational Linguistics for South-Asian Languages, EACL.