IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 5, No 3, September 2011
ISSN (Online): 1694-0814
www.IJCSI.org

327

# Automatic Keywords Extraction for Punjabi Language

**Vishal Gupta[1] and Gurpreet Singh Lehal[2]**

**[1] Assistant Professor Computer Science & Engineering, UIET,**
**Panjab University Chandigarh, UT, Pin Code-160014,India**

**[2] Professor Department of Computer Science,**
**Punjabi University Patiala, Punjab, Pin Code-147002, India**

## Abstract

Automatic keywords extraction is the task to identify a small set of words, key phrases, keywords, or key segments from a document that can describe the meaning of the document. Keywords are useful tools as they give the shortest summary of the document. This paper concentrates on Automatic keywords extraction for Punjabi language text. It includes various phases like removing stop words, Identification of Punjabi nouns and noun stemming, Calculation of Term Frequency and Inverse Sentence Frequency (TF-ISF), Punjabi keywords as nouns with high TF-ISF score and title/headline feature for Punjabi text. The extracted keywords are very much helpful in automatic indexing, text summarization, information retrieval, classification, clustering, topic detection and tracking and web searches etc.

*Keywords: Punjabi keywords extraction, Keywords, Key phrases, TF-ISF*

## 1. Introduction

Keywords [2] are set of significant words in a document that give high-level description of the content for investigating readers and are useful tools for many purposes. They are used in academic articles to give an insight about the article to be presented. In a magazine, they give clue about the main idea about the article so that the readers can determine whether the article is in their area of interest. In a textbook they are useful for the readers to identify the main points in their mind about a particular section. They can also be used for search engines in order to return more precise results in shorter time. Since keywords describe the main points of a text, they can be used as a measure of similarity for text categorization. In summary, keywords are useful tools for scanning large amount of documents in short time. The extracted keywords are very much helpful in automatic indexing, text summarization, information retrieval, classification, clustering, topic detection and tracking [7] and web searches etc.

Despite the usefulness of the keywords, very few of the current documents include them. In fact many authors are not intended to extract keywords and do not denote them unless they are not explicitly instructed to do so. Extracting keywords manually is an extremely difficult and time consuming process, therefore it is almost impossible to extract keywords manually even for the articles published in a single conference. Therefore there is a need for automated process that extracts keywords from documents. Existing methods about Automatic Keyword Extraction can be divided into four categories [6]:-

1) Simple Statistics Approach: These methods are simple and do not need the training data. The statistical information of the words can be used to identify the keywords in the document. Cohen uses NGram statistical information to automatically index the document. N-Gram is language and domain independent. Other statistical methods include word frequency, TF*IDF [1], word co-occurrence [4][8], etc.

2) Linguistics Approach [3]: These approaches use the linguistic features of the words mainly sentences and documents. The linguistic approach includes the lexical analysis, syntactic analysis discourse analysis and so on.

3) Machine Learning Approaches: Keyword Extraction can be seen as supervised learning, Machine Learning approach employs the extracted keywords from training documents to learn a model and applies the model to find keywords from new documents. This approach includes Naïve Bayes, Support Vector Machine, etc.

4) Other approaches: Other approaches about keyword extraction mainly combines the methods mentioned above or use some heuristic knowledge in the task of keyword extraction, such as the position, length, layout feature of words, html tags around of the words, etc.

Various extraction methods discussed are for single document but these can further applied to multiple documents as per their suitability [5].In Automatic Keywords extraction system for Punjabi language, we are using combination of statistical and linguistics approaches for Punjabi language.

IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 5, No 3, September 2011
ISSN (Online): 1694-0814
www.IJCSI.org

328

## 2. Automatic Keywords Extraction for Punjabi Language

Various phases of automatic keywords extraction for Punjabi language are: 1)Removing stop words 2)Identification of Punjabi nouns and noun stemming 3)Calculation of Term Frequency and Inverse Sentence Frequency (TF-ISF) [1] 4) Punjabi keywords as nouns with high TF-ISF score 5)Title/Headline feature.

### 2.1 Removing Stop words from Punjabi text

Punjabi language Stop words are most frequently occurring words in Punjabi text like: ਦੇ dē, ਹੈ hai, ਨੂੰ nūṃ, ਨਾਲ nāl, ਤੋਂ tōṃ… etc. We have to eliminate these words from the original text otherwise, sentences containing them can get influence unnecessarily. We have made a list of Punjabi language stop words by creating a frequency list from a Punjabi corpus. Analysis of Punjabi corpus taken from popular Punjabi newspapers has been done. This corpus contains around 11.29 million words and 2.03 lakh unique words [9]. We manually analyzed these unique words and identified 615 stop words. In the corpus of 11.29 million words, the frequency count of these stop words is 5.267 million, which covers 46.64% of the corpus.

Some of the most commonly occurring stop words are displayed in Table1

Table 1. Punjabi language Stop words list

| ਦੀ dī | ਤੋਂ tōṃ | ਹੋ hō | ਸਨ san |
|---|---|---|---|
| ਨੂੰ nūṃ | ਵੀ vī | ਉਹ uh | ਕੀਤੀ kītī |
| ਹੈ hai | ਕਿ ki | ਉਸ us | ਜਿਸ jis |
| ਨੇ nē | ਅਤੇ atē | ਕਰ kar | ਵਾਲੇ vālē |
| ਸੀ sī | ਹਨ han | ਪਰ par | ਕਰਕੇ karkē and so on………. |

### 2.2 Identification of Punjabi nouns and Stemming

Input words are checked in Punjabi noun morph for possibility of nouns. Usually the words which are nouns with high TF-ISF scores are treated as keywords. Punjabi noun morph is having 74592 noun words in different forms. Examples of Punjabi nouns are shown in table2.

Table2. Punjabi Nouns list

| ਪਹੀਆ pahīā | ਟੱਬਰ ṭabbar | ਸਿੰਗ siṅg |
|---|---|---|
| ਪਰਛਾਂਵਾਂ parchāṃvāṃ | ਘਰ ghar | ਹੱਥ hatth |
| ਪਲਾਟ palāṭ | ਖੰਭ khambh | ਆਰਾ ārā and so on… |

The purpose of stemming [10][11] is to obtain the stem or radix of those words which are not found in dictionary. In Punjabi language noun stemming [12][13][14], an attempt is made to obtain stem or radix of a Punjabi word and then stem or radix is checked against Punjabi noun morph for the possibility of noun. An in depth analysis of corpus was made and the possible noun suffixes [16] were identified (Table 3) and the various rules for Punjabi word noun stemming have been generated. Results of Punjabi language noun stemmer [16] are given in table 4.

Table 3. Punjabi language nouns 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 |
| ੀਆ īā | -- | -- | -- |

Table 4. Results of Punjabi language Noun stemmer

| Punjabi Noun word | Stem word | suffix | Punjabi Noun word | Stem word | suffix |
|---|---|---|---|---|---|
| ਕਸਾਈਆ Kasāīā | ਕਸਾਈ kasāī | ਈਆ īā | ਮਾਹੀਆ māhīā | ਮਾਹੀ māhī | ੀਆ īā |
| ਘਰੋਂ gharōṃ | ਘਰ ghar | ੋਂ ōṃ | ਭਾਸ਼ਾਵਾਂ bhāshāvāṃ | ਭਾਸ਼ਾ bhāshā | ਵਾਂ vāṃ |
| ਲੜਕੀਆਂ laṛkīāṃ | ਲੜਕੀ laṛkī | ੀਆਂ īāṃ | ਆਗੂਆਂ āgūāṃ | ਆਗੂ āgū | ੂਆਂ ūāṃ |
| ਫੁੱਲਾਂ phullāṃ | ਫੁੱਲ phull | ਾਂ āṃ | ਲੜਕੋ laṛkō | ਲੜਕਾ laṛkā | ੋ ō |

IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 5, No 3, September 2011
ISSN (Online): 1694-0814
www.IJCSI.org

329

| ਲੜਕਿਆਂ | ਲੜਕਾ | ਿਆਂ | ਲੜਕੀਏ | ਲੜਕੀ | ੀਏ |
|---|---|---|---|---|---|
| laṛkiāṃ | laṛkā | iāṃ | laṛkīē | laṛkī | īē |
| ਮੁੰਡੇ | ਮੁੰਡਾ | ੇ | ਲੜਕੀਓ | ਲੜਕੀ | ੀਓ |
| muṇḍē | muṇḍā | ē | laṛkīō | laṛkī | īō |
| ਲੜਕਿਓ | ਲੜਕਾ | ਿਓ | ਲੜਕਿਆ | ਲੜਕਾ | ਿਆ |
| laṛkīō | laṛkā | iō | laṛkiā | laṛkā | iā |
| ਘਰੀਂ | ਘਰ | ੀਂ | ਦਰਵਾਜਿਉੰ | ਦਰਵਾਜਾ | ਿਉੰ |
| gharīṃ | ghar | īṃ | darvājium̐ | darvājā | ium̐ |
| ਪਰਾਂਦੇ | ਪਰਾਂਦਾ | ੇ | ਭਾਸ਼ਾਈ | ਭਾਸ਼ਾ | ਈ |
| parāndē | parāndā | ē | bhāshāī | bhāshā | ī |

An In depth analysis of output of Punjabi noun stemmer has been done over 50 Punjabi documents of Punjabi news corpus. The efficiency of Punjabi language noun stemmer is 82.6%.

## 2.3 Calculation of Term Frequency-Inverse Sentence Frequency TF-ISF

The basic idea of TF-ISF [1] [15] score is to evaluate each word in terms of its distribution over the document. Indeed, It is obvious that words occurring in many sentences within a document may not be useful for topic segmentation purposes. It is used to evaluate the importance of a word within a document based on its frequency within a given sentence and its distribution across all the sentences within the document. The TF-ISF measure of a word w in a sentence s, denoted TF-ISF(w,s), is computed by:

TF-ISF(w,s)= TF(w,s)* ISF(w)   where the term frequency TF(w,s) is the number of times that word w occurs in sentence s, and the inverse sentence frequency ISF(w) is given by the formula:

ISF(w) = log(|S|/ SF(w)) , where the sentence frequency SF(w) is the number of sentences in which the word w occurs. Top scored Punjabi words (Top 20%) with high value of TF-ISF scores are candidates for keywords from this phase.

## 2.4 Punjabi keywords as nouns with high TF-ISF score

In this phase, Punjabi keywords are extracted by performing intersection of noun keywords and keywords with high TF-ISF score (Top 20%) from previous phases. Those Punjabi words which are Punjabi nouns and with high TF-ISF scores are candidates for Punjabi Keywords.

## 2.5 Punjabi language Title/Headline Feature

Noun words appearing in title/headline (after removing stop words) are always more important. These words are treated as keywords. The union of these keywords and keywords coming from previous phase (noun words with high TF-ISF scores) are treated as final Punjabi keywords.

## 2.6 Algorithm for Punjabi Keywords Extraction

**Step1:-** From input Punjabi text remove stop words.

**Step2:-** Check the input words in Punjabi noun morph for the possibility of Punjabi nouns and if necessary, perform noun stemming.

**Step3:-** Calculate TF-ISF score of each remaining Punjabi word    TF-ISF(w,s)= TF(w,s)* ISF(w)   where TF(w,s) is the number of times that word w occurs in sentence s, and the inverse sentence frequency ISF(w) = log(|S|/ SF(w)) , where the sentence frequency SF(w) is the number of sentences in which the word w occurs.

**Step4:-** Top scored words (top 20%) with high TF-ISF scores are candidates for keywords from this phase.

**Step5:-** Perform intersection of Punjabi noun keywords and keywords with high TF-ISF scores. Punjabi nouns with high TF-ISF score are candidates of Punjabi keywords from this phase.

**Step6:-** Treat the noun words appearing in title/headlines as keywords (After removing stop words).

**Step7:-** The Union of keywords coming from step5 and step6 are final Punjabi keywords.

## 3. Results and Conclusions

An In depth analysis of output of Punjabi keyword extraction has been done over 50 Punjabi documents of Punjabi news corpus. The Precision, Recall and F-Score of Punjabi language keywords extraction are 80.4%, 90.6% and 85.2% respectively. 14.8% of errors are due to absence of certain Punjabi noun words in noun morph, dictionary mistakes, input text syntax mistakes and certain rules violations of noun stemming.

**The Example Input Punjabi text is as follows:-**

ਉਪ ਮੁੱਖ ਮੰਤਰੀ ਸੁਖਬੀਰ ਬਾਦਲ ਦੇ ਦੌਰੇ ਨੂੰ ਲੈ ਕੇ ਭੁੱਚੋ ਮੰਡੀ ਦੇ ਵਾਸੀਆਂ ਨੇ ਕੀਤੀ ਮੀਟਿੰਗ

ਭੁੱਚੋ ਮੰਡੀ, 8 ਜਨਵਰੀ ( ਜਸਪਾਲ ਸਿੰਘ ਸਿੱਧੂ)- ਪੰਜਾਬ ਦੇ ਉਪ ਮੁੱਖ ਮੰਤਰੀ ਸ: ਸੁਖਬੀਰ ਸਿੰਘ ਬਾਦਲ ਅਤੇ ਲੋਕ ਸਭਾ ਮੈਂਬਰ ਬੀਬਾ ਹਰਸਿਮਰਤ ਕੌਰ ਬਾਦਲ ਦੇ 10 ਜਨਵਰੀ ਦੇ ਭੁੱਚੋ ਮੰਡੀ ਦੌਰੇ ਨੂੰ ਲੈ ਕੇ

IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 5, No 3, September 2011
ISSN (Online): 1694-0814
www.IJCSI.org

330

ਮੰਡੀ ਨਿਵਾਸੀਆਂ ਨੇ ਨਗਰ ਕੌਸਲ ਭੁੱਚੋ ਮੰਡੀ ਦੇ ਦਫ਼ਤਰ ਵਿਖੇ ਵਿਸ਼ਾਲ ਮੀਟਿੰਗ ਕੀਤੀ। ਇਸ ਮੀਟਿੰਗ ਵਿਚ ਮੰਡੀ ਦੀਆਂ ਮੰਗਾਂ ਬਾਰੇ ਖੁੱਲ੍ਹ ਕੇ ਵਿਚਾਰ ਵਟਾਂਦਰਾ ਕੀਤਾ ਗਿਆ। ਪਤਾ ਲੱਗਾ ਹੈ ਕਿ ਸ: ਸੁਖਬੀਰ ਸਿੰਘ ਬਾਦਲ ਇਸ ਦਿਨ ਮੰਡੀ ਦੇ ਆਮ ਲੋਕਾਂ ਨੂੰ ਮਿਲਕੇ ਮੰਡੀ ਦੀਆਂ ਸਾਂਝੀਆਂ ਸਮੱਸਿਆਵਾਂ ਨੂੰ ਮੌਕੇ 'ਤੇ ਹੀ ਦੂਰ ਕਰਨ ਦਾ ਯਤਨ ਕਰਨਗੇ।

up mukkh mantrī sukhbīr bādal dē daurē nūṃ lai kē bhuccō maṇḍī dē vāsīāṃ nē kītī mīṭiṅg

bhuccō maṇḍī, 8 janvarī (jaspāl siṅgh siddhū)- pañjāb dē up mukkh mantrī sa: sukhbīr siṅgh bādal atē lōk sabhā maimbar bībā harsimrat kaur bādal dē 10 janvarī dē bhuccō maṇḍī daurē nūṃ lai kē maṇḍī nivāsīāṃ nē nagar kauṃsal bhuccō maṇḍī dē daftar vikhē vishāl mīṭiṅg kītī. is mīṭiṅg vic maṇḍī dīāṃ maṅgāṃ bārē khullh kē vicār vaṭāndrā kītā giā. patā laggā hai ki sa: sukhbīr siṅgh bādal is din maṇḍī dē ām lōkāṃ nūṃ milkē maṇḍī dīāṃ sāñjhīāṃ samssiāvāṃ nūṃ maukē 'tē hī dūr karan dā yatan karnagē.

**Output Keywords of Punjabi Keywords extraction are as follows:-**

ਮੁੱਖ (mukkh)

ਮੰਤਰੀ (mantrī)

ਮੰਡੀ (maṇḍī)

ਵਾਸੀਆਂ (vāsīāṃ)

ਮੀਟਿੰਗ (mīṭiṅg)

ਪੰਜਾਬ (pañjāb)

ਨਿਵਾਸੀਆਂ (nivāsīāṃ)

ਮੰਗਾਂ (maṅgāṃ)

ਵਿਚਾਰ (vichār)

ਲੋਕਾਂ (lōkāṃ)

ਸਮੱਸਿਆਵਾਂ (samssiāvāṃ)

ਯਤਨ (yatan)

Now in the conclusion, in this paper, we have discussed the Automatic Keywords extraction for Punjabi language. Punjabi nouns with high TF-ISF scores are candidates for Punjabi Keywords. Noun words appearing in the title/headlines are directly treated as keywords. The extracted keywords are very much helpful in automatic indexing, text summarization, information retrieval, classification, clustering, topic detection and tracking and web searches etc. Most of the lexical resources used in pre-processing such as Punjabi Stop words list and Punjabi noun stemmer 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 and Punjabi morph had to be carried out. This the first time some of these resources have been developed for Punjabi and they can be beneficial for developing other NLP applications in Punjabi.

## References

[1] Neto, Joel al., "Document Clustering and Text Summarization", In: Proc. of 4th Int. Conf. Practical Applications of Knowledge Discovery and Data Mining, London, 2000, pp. 41-55.

[2] David B. Bracewell and Fuji REN, " Multilingual Single Document Keyword Extraction For Information Retrieval", Proceedings of NLP-KE, 2005,pp. 517-522.

[3] Xinghua u and Bin Wu, " Automatic Keyword Extraction Using Linguistics Features ", Sixth IEEE International Conference on Data Mining(ICDMW'06), 2006.

[4] Chengzhi Zhang, " Automatic Keyword Extraction From Documents Using Conditional Random Fields ", Journal of Computational and Information Systems, 2008.

[5] www. wikipedia.org.

[6] Jasmeen Kaur and Vishal Gupta, " Effective Approaches for Extraction of Keywords", International Journal of Computer Science Issues IJSCI, Vol.7, Issue 6. Non 2010, pp. 144-148.

[7] Sungjick Lee, Han-joon Kim, " News Keyword Extraction For Topic Tracking ", Fourth International Conference on Networked Computing and Advanced Information Management, 2008 ,pp. 554-559.

[8] Y. Matsuo and M. Ishizuka, " Keyword Extraction from a Single Document using Word Co-occurrence Statistical Information ", International journal on Artificial Intelligence Tools, vol.13, no.1, 2004,pp.157-169.

[9] Punjabi Unique word Corpus.

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

[11] Praveen Kumar, Shrikant Kashyap, Ankush Mittal and Sumit Gupta, "A query answering system for E-learning Hindi documents", South Asian Language Review, VOL.XIII, Nos 1&2, January-June, 2003.

[12] Mandeep Singh Gill, G.S. Lehal and S.S. Joshi, "Part of Speech Tagging for Grammar Checking of Punjabi", The Linguistic Jornal Volume 4 Issue 1, 2009, pp.6-21

[13] www.advancedcentrepunjabi.org/punjabi_mor_ana.asp

[14] Ananthakrishnan Ramanathan and Durgesh Rao, "ALightweight Stemmer for Hindi", Workshop on Computational Linguistics for South-Asian Languages, EACL, 2003.

[15] Rasim M. Alguliev and Ramiz M. Aliguliyev," Effective Summarization Method of Text Documents ", Proceedings of International Conference on Web Intelligence, IEEE, 2005.

[16] Vishal Gupta and Gurpreet Singh Lehal," Preprocessing Phase of Punjabi Language Text Summarization", International Conference on Information Systems for Indian Languages Communications in Computer and Information

IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 5, No 3, September 2011
ISSN (Online): 1694-0814
www.IJCSI.org

331

Science    ICISIL2011,  Volume  139,  Part2,  Springer-Verlag
Berlin  Heidelberg, 2011, pp. 250-253.

**First Author's Biodata**

Vishal Gupta is Assistant Professor in Computer Science &
Engineering at University Institute of Engineering & Technology,
Panjab university Chandigarh. He has done MTech. in computer
science & engineering from Punjabi University Patiala in 2005. He
is among University toppers. He has done BTech. in Computer
Science & Engineering from Govt.  Engineering College Ferozepur
in   2003. He is also pursuing his PhD in Computer Science &
Engineering from University College of Engineering, Punjabi
University Patiala, under the supervision of Dr. Gurpreet Singh
Lehal. He has developed a number of research projects in field of
natural  language  processing  including  synonyms  detection,
automatic question answering and text summarization etc.

**Second Author's Biodata**

Professor Gurpreet Singh Lehal received
undergraduate degree in Mathematics in
1988 from Panjab University, Chandigarh,
India, and Post Graduate degree in
Computer Science in 1995 from Thapar
Institute of Engineering & Technology,
Patiala, India and Ph. D. degree in
Computer Science from Punjabi University,
Patiala,  in  2002.  He  joined  Thapar
Corporate R&D Centre, Patiala, India, in 1988 and later in 1995 he
joined Department of Computer Science at Punjabi University,
Patiala. He is actively involved both in teaching and research. His
current areas of research are- Natural Language Processing and
Optical Character recognition. He has published more than 25
research papers in various international and national journals and
refereed conferences. He has been actively involved in technical
development of Punjabi and has to his credit the first Gurmukhi
OCR, Punjabi word processor with spell checker and various
transliteration software. He was the chief coordinator of the project
"Resource Centre for Indian Language Technology Solutions-
Punjabi", funded by the Ministry of Information Technology as well
as the coordinator of the Special Assistance Programme (SAP-
DRS) of the University Grants Commission (UGC), India. He was
also awarded a research project by the International Development
Research Centre (IDRC) Canada for Shahmukhi to Gurmukhi
Transliteration Solution for Networking.