

KEYWORD EXTRACTION FOR PUNJABI LANGUAGE

KAMALDEEP KAUR

University Institute of Engineering & technology, Panjab University,
Chandigarh, 160014, India
kamal.gndec@gmail.com

VISHAL GUPTA

University Institute of Engineering & Technology, Panjab University,
Chandigarh, 160014, India
vishal@pu.ac.in

Abstract

This paper introduces keyword extraction for Punjabi language. Keywords are the index terms that contain the most important information about the contents of the document. Automatic keyword extraction is the task to identify a small set of words, keyphrases or keywords from a document that can describe the meaning of document. Not much work has been done in keyword extraction for Indian languages in general and Punjabi in particular. Adequate annotated corpora are not yet available in Punjabi. The paper represents the Automatic keyword extraction system for Punjabi language to find words from a document which convey the complete meaning of the text. First we survey about the various approaches available for keyword extraction, then represent our hybrid approach for Punjabi. A number of features are used to extract keywords effectively. The experimental results are shown.

Keywords: NLP, Text Mining, Keywords.

1. Introduction

Keyword extraction is a text mining task that is covered under the concept of Natural Language Processing. Natural Language processing (NLP) is a field of computer science and linguistics concerned with the interactions between computers and human (natural) languages. In theory, natural-language processing is a very attractive method of human-computer interaction. Natural-language understanding is sometimes referred to as an AI-complete problem, because natural-language recognition seems to require extensive knowledge about the outside world and the ability to manipulate it. [17]

NLP is an area of research and application that explores how computers can be used to understand and manipulate natural language text or speech to do useful things. NLP researchers aim to gather knowledge on how human beings understand and use language so that appropriate tools and techniques can be developed to make computer systems understand and manipulate natural languages to perform the desired tasks. The foundations of NLP lie in a number of disciplines, viz. computer and information sciences, linguistics, mathematics, electrical and electronic engineering, artificial intelligence and robotics, psychology, etc. [3]

Text mining is a new area of computer science which fosters strong connections with natural language processing, data mining, machine learning, information retrieval and knowledge management. It seeks to extract useful information from unstructured textual data through the identification and exploration of interesting patterns.[11]

The field of text mining has received a lot of attention due to the always increasing need for managing the information that resides in the vast amount of available documents [12]. The goal is to discover unknown information, something that no one yet knows.

[9]The problem introduced by text mining is obvious: natural language was developed for humans to communicate with one another and to record information, and computers are a long way from comprehending

natural language. Humans have the ability to distinguish and apply linguistic patterns to text and humans can easily overcome obstacles that computers cannot easily handle.

2. Keyword Extraction

As the rapid growth of textual information online, information retrieval becomes more important than ever. Keyword extraction as a foundational technique for documents description has been a focal point of research in this area. Keywords can help the users quickly skim over the documents to determine whether they are worth reading or not. [14]

So effective keywords are a necessity. Everyday thousands of books, papers are published which makes it very difficult to go through all the text material; instead there is a need of good information extraction or summarization methods which provide the actual contents of a given document. Since keyword is the smallest unit which express meaning of entire document, many applications can take advantage of it such as automatic indexing, text summarization, information retrieval, classification, clustering, filtering, cataloging, topic detection and tracking, information visualization, report generation, web searches etc. [2]

2.1. Approaches

There are two existing approaches to automatic keyword indexing [10] [16]

(1) Keyword extraction:

Words occurred in documents are analyzed to identify apparent significant ones, on the basis of properties such as frequency and length. Here aim is to extract keywords with respect to their relevance in text without prior vocabulary.

(2) Keyword Assignment:

Keywords are chosen from a controlled vocabulary of terms and documents are classified according to their content into classes that correspond to elements of vocabulary. This approach is also called Text Categorization. There is a prior set of vocabulary and aim is to match them to texts in a set.

Existing methods for Automatic Keyword Extraction can be divided into four categories:

2.1.1. Simple Statistical Approach

[5] These methods are simple and do not need the training data. The statistical information of the words can be used to identify keywords in the document. Cohen uses N-Gram statistical information to automatically index the document. Other statistical methods include word frequency, TF*IDF, word co-occurrence etc.

2.1.2. Linguistics Approach

[8][10] These approaches use the linguistic features of the words mainly sentences and documents. The linguistic approach includes the lexical analysis, syntactic analysis, discourse analysis etc.

2.1.3. Machine Learning Approach

[6][7][13] 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.

2.1.4. Other Approaches

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

On the basis of above approaches, a number of different methodologies can be used for effective keyword extraction:

(1) Identify Noun Phrase for Keyword Extraction

[9][17] The algorithm adopted by this approach needs certain language dependent components, a morphological analyzer and simple noun phrase grammar in order to determine the keywords. During the morphological analysis, word segmentation, Part of speech tagging, Stemming, Unigram frequency

calculation, is performed. Noun phrases are scored and clustered. Centroids from the top scoring clusters are chosen to be the keywords.

(2) Position Weight Algorithm

[15]This is based on the fact that words in different positions carry different importance, according to which weights are assigned to them. More important are the words in the introduction and conclusion paragraphs, title, summarization sentences.

(3) Informative Feature Selection for Keyword Extraction

[1]The forms of writing provide substantial information about the importance of words. The various informative features can be

- Words emphasized by application of bold, italic or underline fonts
- Words typed in upper case
- Normalized sentence length, which is the ratio of number of words occurring in sentence over number of words occurring in the longest sentence of the document.
- Cue phrases are sentences beginning with summary phrase and transition phrase like however, but, yet, nevertheless.

(4) Term Frequency Inverse Document Frequency

[18]It is the very basic statistical approach to extract keywords. It evaluated how important a word is to a document in a collection or corpus. Importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of word in a corpus.

(5) Keyword extraction using conditional random field model

[19]This model works on document specific features. CRF model is a new probabilistic model for segmenting and labeling sequence data. It is an undirected graphical model that encodes a conditional probability distribution with a given set of features. It uses CRF++ tool to extract keywords.

2.2. Keyword extraction for Punjabi Language

Punjabi is the language of Punjab, spoken mainly in Northern parts of India. Punjabi is highly inflectional and agglutinating language providing one of the richest and most challenging sets of linguistic and statistical features resulting in long and complex word forms. Each word in Punjabi is inflected for a large number of word forms. It is primarily a suffixing language. An inflected word starts with a root and may have several suffixes added to the right. It is a free word order language.

Punjabi, like other Indian languages, is a resource poor language- annotated corpora, name dictionaries, good morphological analyzers; POS taggers are not yet available in the required measure. Although Indian languages have a very old and rich literary history, technological developments are of recent origin. Web sources for name lists are available in English, but such lists are not available much in Punjabi.

2.2.1. Approach

Since all the existing approaches have certain significant features or advantages that are helpful in extraction of keywords. The approach being used for Keyword extraction for Punjabi language in our experiment is 'Hybrid approach'. The hybrid approach combines the salient features used in these approaches which are applicable to Punjabi language. The extracted keywords are most effective and efficient because the system utilizes a large number of features and does not rely on a single technique. As large corpus for Punjabi is not yet available, so it is not possible to train a system to generate rules automatically. That is why; these approaches are being used for the experiment.

- So keywords are extracted based on mixed approach
- Candidate keywords extracted by each approach are weighted
- The highly weighted keywords chosen by all techniques are merged to form final set of keywords

2.2.2. Application

The keyword extraction for Punjabi language being done in this experiment is used for the TOPIC TRACKING IN PUNJABI LANGUAGE in our next experiment. Topic tracking task is to detect news of a known topic, by monitoring a stream of news stories and finding out those which discuss the same topic described by a few positive samples. That is, the system will determine whether two Punjabi news documents describe the same topic or not. The various keyword features extracted in this experiment will be used in topic tracking by expressing them in the form of collection of event vectors. The event vectors representing the two news documents will be compared to match by at least a predefined threshold value in order to track the same topic or event.

2.2.3. System features

For our system, various keyword extraction approaches are merged that include

- Noun chunking
- Term frequency or number of occurrences of a term
- Term's existence in title
- Cue phrases like words within a sentence boundary containing 'summary phrase' or 'transition phrase'.

Using these features in combination provide better results in comparison to sticking to a single technique. The features can be detailed as:

- Text Segmentation: The written text is divided into meaning units, such as words, sentences or topics. It is the ability to break words into individual syllables.
- Stop words: These words are filtered out. These words are not indexed. For example, a few Punjabi stop words include ਅਤੇ (atē), ਆਪਣਾ (āpaṇā), ਇਕ (ik), ਇਸ (is), ਇਹ (ih), ਉਸ (us), ਉਹ (uh), ਗਈ (gāi) etc.
- Term frequency: Each word is counted to get the term frequency. The highly occurred word has priority to be treated as a keyword since it depicts the meaning of entire text file content. Term frequency describes the count or the number of times a word occurs in a given file and its ranking with respect to other words which are also present in the same text file.
- Noun frequency: Since term frequency alone cannot be considered as the complete measure to find keywords. Those words having part of speech as NOUN are also the important ones to be treated as keywords because nouns are very crucial in any text material. So the terms which are having the highest occurring frequency and are noun as well, ranked as highly preferred keywords or considered as the most appropriate ones.
- Title: Title of the text file plays a very crucial role in deciding the final keywords since from title itself one could judge or guess the complete content of a file.
- Cue phrase: Cue phrases such as summary phrase or transition phrase are considered as the most important for depicting the main theme of whole text or it could be the turning point if some transition phrase has been encountered in a particular sentence of a file. So those words which are given within a sentence containing cue phrases are given high importance and are treated as final keywords. Some of Punjabi cue phrases include ਨਤੀਜਾ (natījā), ਅੰਤ (ant), ਸਿੱਟਾ (siṭṭā), ਨਿਚੋੜ (nicōṛ), ਸੋ (sō), ਅਖੀਰ (akhīr) etc.

2.2.4. Gazetteer list

The lists created in the database as gazetteer lists include

- Punjabi Dictionary with part of speech tagging
- Punjabi Nouns
- Cue phrases words

2.2.5. Methodology

The algorithm for the system can be given as:

- (1) Punjabi source file is preprocessed for stop word elimination.
- (2) Term frequency for each remaining word is calculated.
- (3) Noun feature is applied to select terms which are both noun and are highly frequent.
- (4) Title words are taken as important keywords after removal of unnecessary words such as stop words.
- (5) The words which are present within a sentence containing a Cuephrase are chosen as important keywords.
- (6) All features are incorporated together to get heavily weighted desired keywords.

The system architecture can be shown as:

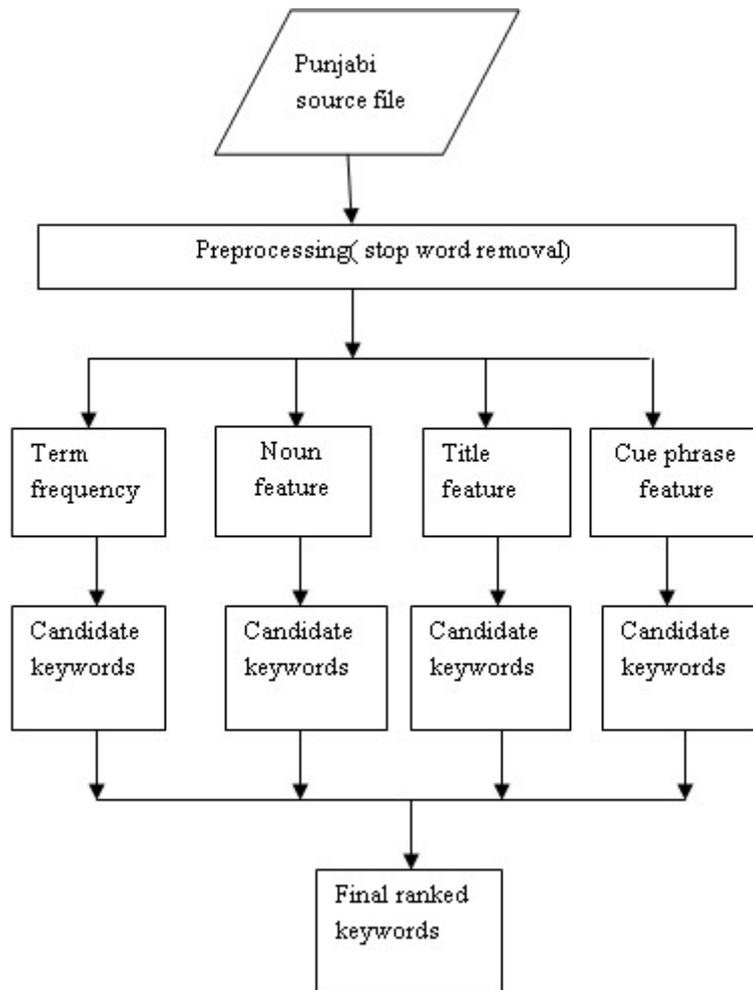


Fig. 1. System Architecture for Keyword Extraction

2.2.6. Evaluation Metrics

The performance of keyword extraction system is measured using precision(P), recall(R) and F-measure.

The precision measures the number of correct keywords, obtained by the system, over the total number of keywords extracted by the system. That is, $P = \text{number of correct keywords} / \text{total number of keywords}$

The recall measures the number of correct keywords, obtained by the system over the total number of keywords in a text that have been used for testing.

$R = \text{number of correct keywords} / \text{total number of keywords in a text}$

The F-measure represents harmonic mean of precision and recall.

$F = 2RP / R + P$

2.2.7. Implementation Details

The system has been implemented using vb.net platform and gazetteers lists are stored as tables in the database. The experiment required input documents. Such test documents are taken from Punjabi web sources such as punjabispectrum.org, likhari.org, ajitweekly.com, jagbani.com, sahitkar.com etc.

2.2.8. Experimental results

The experimental results reported in table show that

Table 1. The results for keyword extraction system

Feature	P(%)	R(%)	F(%)
Title	97.04	97.58	97.30
Cue phrase	100	97.56	98.41
Noun	97.81	75.48	83.39
Total	98.28	90.19	93.03

The system shows good results for all the features individually. And the total result of the keyword extraction system is also good. The little lack in the percentage is due to the reason that some common nouns in Punjabi are also names due to which wrong words are found sometimes. And after the removal of stop words, the system can not find important abbreviations that contain that stop word. But results can be improved by incorporating more features to the existing and considering the issues that lower the results.

2.3. Conclusion & Future Scope

Not much work has been done in keyword extraction for Punjabi and other Indian languages. In this paper, we have reported our work on keyword extraction for Punjabi. We have prepared a 'hybrid system', with the combination of various approaches, i.e. combining a number of features to generate effective automatic keyword extraction system. The language dependent features are formed and analyzed to extract the keywords for Punjabi document. The approach uses the gazetteer lists created from the dictionary with part of speech tagging. Hence, keywords from title, cue phrase and high frequency noun are extracted. The system shows good results for all features independently and the total results for the system are improved with the combination of these features resulting in effective keyword extraction.

Future works include incorporating more features to improve the existing results. The features that can be included are position weight algorithm which weighs words according to their position of occurrence in the document, length of the word by giving more importance to the lengthy words as compared to the short words, informative feature selection such as bold, italic, underlined words. As many first names in Punjabi are also common nouns, this limitation lowers the performance of the system. This issue can be considered to improve the system.

References

- [1] Alguliev, R. M.; Aliguliyev, R. (2005). Effective Summarization Method of Text Documents. Proceedings of International Conference on Web Intelligence(WI'05),IEEE.
- [2] Bracewell, D. B.; Ren F. (2005). Multilingual Single Document Keyword Extraction for Information Retrieval. Proceedings of NLP-KE, pp. 517-522.
- [3] Chowdhury, G. G.: Dept of Computer and Information Sciences, University of Strathclyde, Glasgow G1 1XH, UK.
- [4] Dubey, A. (2006): A deterministic technique for extracting keyword based grammar rules from programs. In the proceedings of 2006 ACM symposium on Applied Computing.
- [5] Jia, H. (2007): Chinese keyword extraction based on word platform. In the proceedings of fourth international conference on fuzzy systems and knowledge discovery, (FSKD'07), Vol.2.
- [6] Jianga, X. (2009): A ranking approach to keyphrase extraction, Microsoft Research Technical report (MRT'09).
- [7] Liu, F.; Liu, Y. (2008):Automatic keyword extraction for the meeting corpus using supervised approach and bigram expansion. In proceedings of the University of Texas at Dallas, Institute of electrical and electronics engineers (IEEE).
- [8] Miller, J. (1990): Wordnet: An online lexical database. International Journal of Lexicography, Vol.3(4).
- [9] Navathe; Shamkant, B.; Elmasri, R. (2000): Data warehousing and Data mining. In Fundamentals of Database systems, Pearson education pvt. Inc., Singapore, 841-872.
- [10] Ogawa, Y. (1993): Simple word strings as compound keywords: An indexing and ranking method for Japanese texts. Proceedings of 16th annual international ACM-SIGIR Conference on Research and development in information retrieval.
- [11] Radovanovic, M.;Ivanovic, M. (2008): Text Mining: Approaches and Applications. Math, J., Vol.38,No. 3:227-234.
- [12] Stavrianon, A.; Andritsos, P; Nicoloyannis, N. (2007):Overview and semantic issues of text mining. Sigmoid record,Vol.36,No.3.

- [13] Turney, P. (2000): Learning Algorithms for keyphrase extraction. Information retrieval-INRT National research council, Vol.2,No.4,303-336.
- [14] Wenchao, M.; Lianchen, L.; Ting, D. (2009): A modified approach to keyword extraction based on word similarity. National CIMS Engineering Research Center, Tsinghua University, Beijing, China, IEEE.
- [15] Xinghua, U.;Bin, W. (2006): Automatic keyword extraction using linguistics features. Sixth IEEE International conference on Data mining (ICDMW'06).
- [16] Zhang, C. (2008): Automatic keyword extraction from documents using conditional random fields. Journal of computational and information systems 4:3,1169-1180.
- [17] http://en.wikipedia.org/wiki/Natural_language_processing
- [18] <http://en.wikipedia.org/wiki/Tf-idf>
- [19] CRF++ Yet another. CRF toolkit. <http://chasen.org/~takn/software/CRF++>