RESEARCH ARTICLE OPEN ACCESS

Keyphrase Extraction From Punjabi Corpus

Preeti[1] and Brahmaleen Kaur Sidhu[2]

[1] Research Scholar, University College of Engineering, Punjabi University, Patiala, Punjab, Pin Code-147002, India

[2] Assistant Professor, University College of Engineering, Punjabi University, Patiala, Punjab, Pin Code-147002, India

ABSTRACT
Punjabi, the official language of the Indian Punjab, is an Indo-Aryan language. Keyphrase extraction is a Natural Language Processing task to identify keyphrases in a document. Keyphrases are groups of words that describe the meaning of a document, provide a brief summary and characterize the document's contents. This paper describes an automated system for extraction of keyphrases from Punjabi corpus.

I. Introduction

Natural Language Processing is a field that deals with languages. Language refers to a body of words and the systems for their use common to a people who are of the same community or nation, the same geographical area, or the same cultural tradition. It is the primary means of communication used by particular groups of human beings [1]. It is the tool used to express our ideas and emotions. A field of linguistics called "Computational Linguistics" deals with the application of "linguistic theories" which identify rules that describe the structure of language and "computational techniques" for Natural Language Processing. Natural Language Processing is a technique where machine can become more human by reducing the distance between human and the machine [1]. It is an approach to analyzing text and representing naturally occurring texts for human like language processing [2]. NLP is a research area and application used to understand natural language text. NLP researchers aim to explore how human beings understand and use language [3].

A phrase is a group of two or more words grammatically linked with each other [4]. It is a group or words that express a concept and is used as a unit within a sentence [5] and is a collection of words that may have nouns or verbals, but it does not have a subject doing a verb [6]. A phrase is a group of words which acts as a single unit in meaning and in grammar, and is not built round a verb. A phrase is an expansion of one of the words inside it, which is called its head [7].

The following examples of Punjabi language show phrases which are highlighted.

- **s~jx isMG mMjy** ipAw ipAw dUr in~kI DI dy irSqy dI Bwl c guAwicAw ipAw sI[
- ieh sB **s~jxisMG frwmW dyKx** leI mjbUr KVw sI [
- Ardws dy frwmy qoN bAwd **mu~KI bwbw boilAw,** lY BeI krmW vwilAw qyrI borI dI Ardws ho cu~kI hY [
- **ishq KojkwrW ny** isgryt pIx vwlIAW lokW iDAwn iK~cx leI ie~k nvW rsqw iqAwr kr ilAw hY [
- skwtlYNf iv~c strilMg dy XUnIvristI iv~c KojkwrW ny hr **vwr F~kn KolHdw** ie~k sunyhw irkwrf iqAwr kIqw hY [

Keyphrases are useful tools for searching large amount of documents in short time. Majority of documents come without keyphrases and assigning them manually is a difficult process. Many Keyphrase extraction systems are available for English language, but no such system is available for Punjabi language.

Keyphrase extraction can highly improve the efficiency of search operations in Punjabi documents. Search engines don't have to traverse the complete document if keyphrases have been listed, thus leading to better results in shorter time. Keyphrases are representative of the complete document. Irrelevant results can be reduced if search is based on keyphrases.

II. Keyphrase Extraction

Keyphrases are linguistic descriptors of documents used as features in many text-related applications. They describe document's contents helping the readers to decide whether it is relevant for them or not. Keyphrases can be single keywords or multiword key terms.

Extensive amount of information is available in the form of books, articles in digital media, making the task of finding a specific item very difficult. There is a need of an efficient

information extraction system that identifies terms that reflect the meaning of given document. Keyphrases express the theme of the entire document and may be augmented with many language processing applications such as text summarization, information retrieval, web searches etc.
Keyphrase Extraction system are available for English, Hindi, Chinese languages. Keywords extraction system is available for Punjabi language, but not keyphrase extraction.

Example:

- KEA: practical automatic keyphrase extraction [8]
- An ontology–based approach for key phrase extraction [9]
- Domain-specific keyphrase extraction [4]
- Extracting keyphrases from chinese news articles [10]

1. Existing Approaches

The manual extraction of keyphrases is slow, expensive and bristling with mistakes. Therefore, most algorithms and systems to help people perform automatic keyphrase extraction have been proposed. There are two different ways of approaching the problem: keyphrase assignment and keyphrase extraction [11][12].

(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 can be divided into four categories: simple statistics, linguistics, machine learning approaches [13][14].

Simple Statistics Approaches

These methods are simple, have limited requirements and don't need the training data. They tend to focus on non-linguistic features of the text such as term frequency, inverse document frequency, and position of a keyphrase. The statistics information of the words can be used to identify the keyphrases in the document. Cohen uses N-Gram statistical information to automatic index the document [15]. Other statistics methods include word frequency, TF*IDF [16], word co-occurrences [17][14], etc. The benefits of purely statistical methods are their ease of use and the fact that they do generally produce good results.

Linguistics Approaches

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 [18][19][11].

Machine Learning Approaches

Keyphrase 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 keyphrases from new documents[20][21][22].

Other approaches

Other approaches about keyphrase 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 [23].

III. ALGORITHM

The proposed keyphrase extraction system is based on following algorithm.

STEP 1: INPUT TEXT

In the first step, Punjabi text is input in Unicode format. Unicode characters can be inserted in two ways: from the screen by means of an applet from which one can select the character, or by certain key sequence on the keyboard. Unicode is a computing industry standard for the consistent encoding, representation and handling of text expressed in most of the world's writing systems. It is the one-size-fits-all character encoding standard designed to clean up the mess of dozens of mutually incompatible ASCII extensions and special encodings and to allow the computer interchange of text in any of the world's writing systems.

STEP 2: TEXT RECONSTRUCTION

Text is scanned to filter out special tokens such as \\, ||, (,), [,] *, {,},!, ^, , +, -,. Several modifications are made: punctuation marks, brackets, and numbers are replaced by blank space. Apostrophes are removed; Hyphenated words are split in two.

STEP 3: WORD SEGMENTATION

Word segmentation module is used to identify and separate the tokens (words) present in the text in such a way that every individual word will be a different token. After tokenization the words are

entered into a database that forms the input to part of speech tagging module.

STEP 4: PART OF SPEECH TAGGING

The output of the segmentation module is taken as input by the part of speech tagging module. Part of Speech tagging is the process of assigning a part of speech (a part of speech, a word class, a lexical class, or a lexical category, is a linguistic category of words, which is generally defined by the syntactic or morphological behaviour of the lexical item, e.g. noun, verb etc) or other lexical class marker to each word in a corpus. Tags are also usually applied to punctuation markers; thus tagging for natural language is the same process as tokenization for computer languages. Each word is built using the words of various word classes like pronoun, pronoun, adjective etc. After POS tagging, the part of speech tags are added into the database. This forms in input to phrase identification.

STEP 5: PHRASE IDENTIFICATION

A phrase is a group of two or more words grammatically linked with each other. The structure followed in Punjabi for phrase is subject-object-verb. Punjabi phrases can be classified into two types – Noun Phrases and Verb Phrases. Noun Phrases usually function as subject or object in sentences, and Verb Phrases function as verb in those sentences. The proposed system identifies phrases from database using the rule subject-object-verb. The generated list of candidate phrase is input to the final step of keyphrase extraction.

STEP 6: EXTRACTING KEYPHRASES

After identification of phrases, the list of phrases is generated as output. The frequency of every phrase is calculated. The most frequently occurring phrases are selected as keyphrases.

The above explained algorithm is diagrammatically illustrated below.

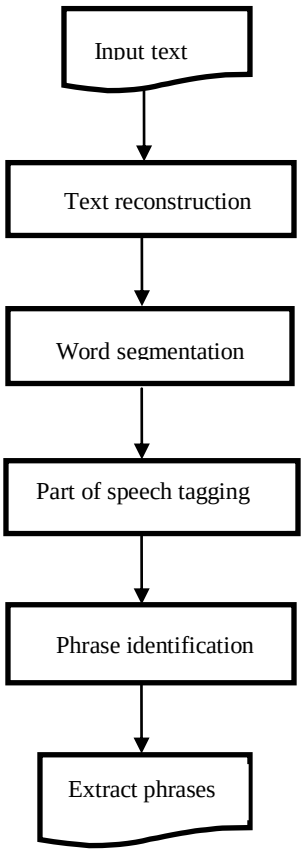

Fig. keyphrase extraction flowchart.

IV. Results and Conclusion

The results of the proposed system are discussed below. The keyphrase extraction system effectively extracts keyphrases from various categories of Punjabi text corpus such as stories, news, articles etc. The system generates list of keyphrases which are representative of the complete text and can be used in various searching and summarization activities.

TYPES OF TEXT	SIZE OF INPUT TEXT (NO. OF WORDS)	AVERAGE NO. OF KEYPHRASES EXTRACTED
Stories	2375	11
Articles	3574	9.6
News	1215	6.6

Table. extracted keyphrases

V. Future scope

The proposed system is integrated with such NLP applications such as text summarization,

searching, text clustering, document similarity analysis, text indexing. The system focuses on noun phrase and verb phrase. Thus, the algorithm may be improved by including more complex phrases. The system may be adopted for extracting variable length keyphrases. The performance can be improved by adding more nouns and verbs in the database used in POS tagging step of the extraction algorithm.

References

[1] U. S. Tiwary, Tanveer Siddiqui *Natural Language Processing and Information Retrieval* (Oxford Higher Education 2008).
[2] http://en.wikipedia.org/wiki/Natural_ langu age _processing
[3] Chowdhury, G. G.: Dept of Computer and Information Sciences, University of Strathclyde, Glasgow G1 1XH, UK.
[4] WWW.google.com
[5] http://examples.yourdictionary.com/exam pl es/phrase-examples.html
[6] http://web.cn.edu/kwheeler/gram_ clauses _n_phrases.html
[7] http://www.phon.ucl.ac.uk/home/dick /tta/ phrases/phrases.html
[8] Witten, I.H., Paynter, G.W., Frank, E., Gutwin, C., and Nevill- Manning, C.G. "Kea: Practical automatic keyphrase extraction". *In Proc. of the 4th ACM Conference on Digital Libraries, 1999.*
[9] Chau Q. Nguyen, Tuoi T. Phan "An Ontology–Based Approach for Key Phrase Extraction" , *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 181–184, Suntec, Singapore, 4 August 2009.
[10] Weiming Lianga, Chang-Ning Huangb , Mu Lib , and Bao-Liang Lu, "Extracting Keyphrases from Chinese News Articles Using TextRank and Query Log Knowledge " , *23rd Pacific Asia Conference on Language, Information and Computation*, pages 733–740
[11] 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.*
[12] Zhang, C. (2008): *Automatic keyword extraction from documents using conditional random fields. Journal of computational and information systems 4:3,1169-1180.*
[13]. Michael J. Giarlo. *A comparative analysis of keyword extraction techniques. Rutgers,* The State University of New Jersey
[14]. Chengzhi Zhang, Huilin Wang, Yao Liu, Dan Wu, Yi Liao, Bo Wang. *Automatic Keyword Extraction from Documents Using Conditional Random Fields. Journal*
[15]. J. D. Cohen. *Language and domain-independent automatic indexing terms for abstracting. Journal of the American Society for Information Science, 1995*
[16] 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.*
[17]. Y. Matsuo, M. Ishizuka. *Keyword extraction from a single document using word co-ocuurrence statistical information. International Journal on Artificial Intelligence Tools, 2004*
[18] Xinghua u and Bin Wu, "*Automatic Keyword Extraction Using Linguistics Features ", Sixth IEEE International Conference on Data Mining(ICDMW'06), 2006.*
[19] Miller, J. (1990): Wordnet: *An online lexical database. International Journal of Lexicography,Vol.3(4).*
[20] Jianga, X. (2009): A ranking approach to keyphrase extraction, Microsoft Research Technical report (MRT'09)
[21] Turney, P. (2000): *Learning Algorithms for keyphrase extraction. Information retrieval-INRT National research council, Vol.2, No.4,303-336.*
[22] 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).*
[23] WWW.wikipedia.org.