International Journal of Advanced Research in

Computer Science Engineering and Information Technology

Volume: 3 Issue: 2 25-Aug-2014,ISSN_NO: 2321-3337

NATURAL LANGUAGE PROCESSING – VALIDATING THE KEYPHRASES USING WIKIPEDIA

[1] G.VIJAY, [2]V.VINOTHA

[1]Assitant Professor, Dept.of.Computer science, MCC college.Pattukottai.

[2]Research Scholar, Dept.of.Computer science, MCC college.Pattukottai.

.

ABSTRACT−*The main objective of the work is to access the vast repository of information that is, Wikipedia's structure and its content using an open source toolkit named as Wikipedia miner. Wikipedia content is a promising resource for natural language processing and many other research areas. An automate process is designed here to validate the available list of key phrases from different domains using this toolkit in two different steps explained below.* **KEA** *(Key phrase Extraction algorithm) is an algorithm for extracting key phrases from text documents. Different combinations of Key phrases are extracted here from a part of the process using KEA. It first cleans the input text, then identifies phrases, and finally stems it. The identified phrases from KEA are then validated by searching it in Wikipedia's content structure. If the phrase is found in the Wikipedia content, then it will be a valid key phrase. It will be useful for developing various applications like thesaurus creation, domain-specific indexing and searching etc.*

Keywords: NLPTOOLKIT..

I. INTRODUCTION

Keyword extraction and validation is done using various algorithms and available resources like thesaurus, dictionary etc. But a new method is proposed here for extracting and validating the key phrases. Different combinations of key phrases are first extracted using an algorithm KEA and the extracted phrases are validated using Wikipedia dump. The complete Wikipedia content dump is processed using Wikipedia miner toolkit.

Natural Language Processing (NLP)

NL refers to the language spoken by people e.g. English, Japanese, Swahili, as opposed to artificial languages, like C++, Java, etc. NLP refers to automatic analysis of human language by computer algorithms.

Natural Language Processing (NLP) is the computerized approach to analyzing text that is based on both a set of theories and a set of technologies. NLP is a theoretically motivated range of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications.

II.EXISTING SYSTEM:

Key phrases provide semantic metadata that summarize and characterize documents. KEA is an algorithm for automatically extracting keyphrases from text. Kea identifies candidate keyphrases using lexical methods, calculates feature values for each candidate, and uses a machine-learning algorithm to predict which candidates are good keyphrases.

Part of the process from this extraction algorithm is used to choose the available combination of phrases from the input text documents. Actually this KEA algorithm has two stages. One is training and another one is extraction. Both stages choose a set of candidate phrases from their input documents, and then calculate the certain attributes for each candidate phrases.

Choosing Candidate keyphrases involves the following three steps.

1. Cleans the input text
2. Identifies candidates
3. Stems and case-folds the phrases

Candidate phrases are limited to a certain maximum length(usually three words). It cannot be proper names (i.e. single words that only ever appear with an

initial capital). It cannot begin or end with a stop word. All contiguous sequences of words in each input line are tested using the above conditions yielding a set of candidate phrases.

III.PROPOSED SYSTEM:

Keyphrases provide a brief summary of a document's contents. As large document collections such as digital libraries become widespread, the value of such summary information increases. Keywords and keyphrases are particularly useful because they can be interpreted individually and independently of each other. They can be used in information retrieval systems as descriptions of the documents returned by a query, as the basis for search indexes, as a way of browsing a collection, and as a document clustering technique.

IV.MODULES

In addition, keyphrases can help users get a feel for the content of a collection, provide sensible entry points into it, show how queries can be extended, facilitate document skimming by visually emphasizing important phrases and offer a powerful means of measuring document similarity. Keyphrases are usually chosen manually. In many academic contexts, authors assign keyphrases to documents they have written. Professional indexers often choose phrases from a predefined "controlled vocabulary" relevant to the domain at hand. However, the great majority of documents come without keyphrases, and assigning them manually is a tedious process that requires knowledge of the subject matter. Automatic extraction techniques are potentially of great benefit.

Several methods have been proposed for generating or extracting summary information from text. In the specific domain of keyphrases, there are two fundamentally different approaches: *keyphrase assignment* and *keyphrase extraction*. Both use machine learning methods, and require for training purposes a set of documents with keyphrases already attached. Keyphrase assignment seeks to select the phrases from a controlled vocabulary that best describe a document. The training data associates a set of documents with each phrase in the vocabulary, and

builds a classifier for each phrase. A new document is processed by each classifier, and assigned the keyphrase of any model that classifies it positively. The only keyphrases that can be assigned are ones that have already been seen in the training data.

Keyphrase extraction, the approach used here, does not use a controlled vocabulary, but instead chooses keyphrases from the text itself. It employs lexical and information retrieval techniques to extract phrases from the document text that are likely to characterize it. In this approach, the training data is used to tune the parameters of the extraction algorithm.

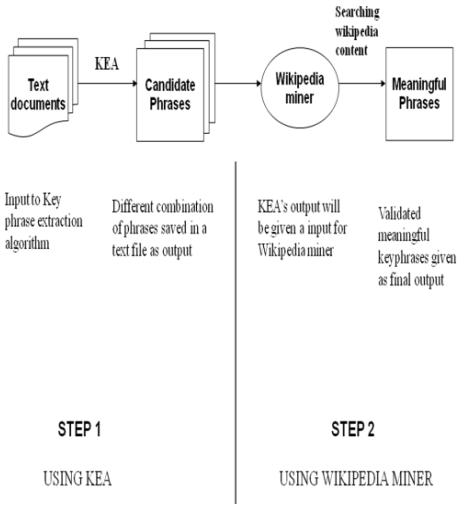
Goal of KEA

The kea algorithm is language-independent (although a stemmer and a stopword list are used, both of which do depend on the language). It provides useful metadata where none existed before. Although evaluates kea's performance by comparing with the author's own keyphrases. If it can extract reasonable summaries from text documents, then it will give a valuable tool to the designers and users of digital libraries.

We then give an example of the prediction model generated by kea and show how it is used to assess a candidate keyphrase. The goals were to assess kea's overall effectiveness, and also to investigate the effects of varying several parameters in the extraction process. It also measured keyphrase quality by counting the number of matches between kea's output and the keyphrases that were originally chosen by the document's author.

There are also differences in the type of keyword extraction that is chosen, which may be broken into three categories: statistical methods, linguistic methods, and mixed methods. Statistical methods, such as those employed in Kea, tend to focus on non-linguistic features of the text such as term frequency, inverse document frequency, and position of a keyword. The benefits of purely statistical

methods are their ease of use, limited computation requirements, and the fact that they do generally produce good results.



Among these two steps, first one is described in this chapter and the remaining process will be explained in the upcoming chapters. The remaining step includes installing the Wikipedia dump and the searching the content for meaning phrases.

The actual KEA training algorithm involves three parts of process for choosing candidate phrases and creating a model based on it. It is shown in the above fig 3.2. But our project uses only the first part, which is choosing candidate phrases and ignored the remaining two parts. It is shown in the following fig 3.3. The extracted candidate phrases are saved in a plain text for further process with Wikipedia miner. It is described in the following chapters.

V. CONCLUSIONS

The candidate phrases extracted from Keyphrase Extraction Algorithm (KEA) is validated as a meaning phrases using Wikipedia content dump. Wikipedia dump contains different type of tables and it is used for various purposes in doing research. Extracting implementing, and mining those table values are quite easy.

But working with content table from the dump is a difficult task. Because the size of the table is several gigabytes and it needs parallelized systems to work on it. So the parts of content table values are used for my research. The validated phrases from Wikipedia are then used for keyphrase indexing and searching in online job searching and recruiting purposes. Online resume writers suggests the job seekers to include these validated keyphrases in their resume for matching against their suitable jobs. Employers used the keyphrase index for searching the suitable candidates in their required domains.

VI. REFERENCES

[1] Paul Buitelaar, Philipp Cimiano, and Bernado Magnini, (2005) *Ontology Learning from Text: Methods, Evaluation and Applications* (DFKI Saarbrucken, University of Karlsruhe, and ITC-irst)

[2] Ian H. Witten, Gordon W. Paynter, Eibe
Frank, (2000) *Practical Automatic Keyphrase Extraction Algorithm*, Dept of Computer Science, University of Waikato, Hamilton, New Zealand.

[3]Takashi Tomokiyo and Matthew Hurst, (2003) Applied Research Center Intelliseek, Inc. Pittsburgh, PA 15213.

[4]Milne, D. and Witten, I.H. (2008) *Learning to link with Wikipedia*, proceedings of the 17th ACM conference on Information and knowledge management (2008). Department of Computer Science, University of Waikato, Hamilton, New Zealand.

[5]Medelyan, O. and Milne, D. (2008) *Augmenting domain-specific thesauri with knowledge from Wikipedia*. In Proceedings of the NZ Computer Science Research Student Conference, NZ CSRSC-2008

[6]Ian H. Witten David Milne, and David M. Nichols. (2007) *A Knowledge-Based Search Engine Powered by Wikipedia*. In Proceedings of the ACM Conference on Information and Knowledge Management (CIKM'2007), Lisbon

[7] Hassan, S. and Mihalcea, R. (2008) *Cross-lingual Semantic Relatedness Using Encyclopedic Knowledge*. In Proceedings of EMNLP 2009, 1192-1201.