Asian Journal of Information Technology 15 (3): 372-375, 2016
ISSN: 1682-3915

Key Phrase Extraction Using Naive Bayes' in Question Generation System

P. Pabitha, S. Suganthi and Raja Ram
Anna University, Chennai, India

Abstract: Automatic Question Generation (AQG) is a challenging task which involves many difficulties. The major aspects of automatic question generation are selecting the target content (what to ask), question type (who, why, how) and actual question generation. The problem encountered in the existing system was that some of the definition sentences are extracted from Wikipedia which were implicit or matched with multiple rules from different key phrase categories. Another limitation is that it is domain dependent and may not apply this approach to other applications such as reading comprehension. The proposed system overcomes the problems by using supervised learning approach. It also extends its work to applications like reading comprehension. The computers can read the submitted documents. The proposed system initially stems the document. The system extracts the key phrases from the documents through its knowledge. Each key phrase is matched with the database.

Key words: Key phrases, supervised machine learning, naive bayes, stemming, automatic question generation

INTRODUCTION

Formally, semantic web can be described as an extension of the present web which is able to describe things in a way that computers can understand. Key phrases are phrases that represent the important part of the document. So, key phrases are needed to understand the document with little time. But, many articles that is available today are not assigned with key phrases. Assigning key phrases to the entire article manually is not also possible now a days. There is a need of automation in assigning key phrases to the entire document. The proposed system assigns key phrases to the submitted document based on the supervised machine learning after stemming is done. Supervised machine learning is a one of the learning technique which initially trains the system through the training documents and it makes the system to extract key phrases for the new document based on the training data. Naive Bayes is supervised probabilistic classifier and it works based on the Bayes' Theorem with many independent assumptions. These key phrases are the input to the next step.

Literature review: Stemming is the process for reducing inflectional or derived words to its stem or root form. The automatically removing suffixes of word in English in specific interest of information retrieval. The algorithm (Porter, 1980) used for removing the suffix stripping is described which has been implemented as a short fast program in BCPL. A theoretical and practical attributes of stemming algorithm and a new version of context-sensitive and longest matching stemming algorithm (Megala *et al*., 2013) for English is proposed, though developed for use in library information of transfer systems. The improvement of a stemming algorithm to produce efficient and meaningful stem such that information retrieval system retrieves the information from a collection of documents which is need by the user and satisfy the user in this case information retrieval system using stemming algorithm in the pre-processing stage convert the word into a root form.

Academic journals and articles with a list of key words also called as phrases as it contains two or more words. Supervised learning (Turney, 2000) is used to generate key phrases. The key phrases are documented by classifying the whole document into a set of positive and negative examples. Two sets of experiments are applied one with the C4.5 decision tree algorithm and the other with the GenEx algorithm. A fuzzy set theoretic approach, fuzzy n-gram indexing, is used to extract n-gram key words. It is noticed that n-gram keyword renders a better result as compared to mono-gram keyword but, for some documents the most relevant keyword is mono-gram. This approach neither requires a dictionary or thesaurus nor does it depend on the size of text document. The Lingo algorithm (Osinski *et al*., 2004), another unsupervised approach, is generally used for clustering web search results. It is based on Singular Value Decomposition (SVD). Barker and Cornacchia (2000) ranked noun phrases extracted from a document by using simple heuristics based on the length and the frequency of their head noun. Another widely adopted

Corresponding Author: P. Pabitha, Anna University, Chennai, India

unsupervised approach for key phrase extraction is to use graph-based ranking methods. Mihalcea and Tarau (2004) represented a document as a term graph based on term relatedness; a graph based ranking algorithm is then used to assign importance scores to each term.

MATERIALS AND METHODS

Proposed system: The proposed system architecture is shown in Fig. 1. It consists of following components: stemmer, key phrases extractor, database and question generator. Stemmer does the job of stemming the words in the given text document. Supervised learning is a machine learning technique in which the systems are trained and produces an inferred function which can be used for mapping new examples. The proposed system initially extracts key phrases from the documents using the key phrase extractor. The extractor does the extraction of key phrase based on the model. For reading comprehension, the questions are generated from the extracted key phrases. Therefore, the proposed method generates question for the submitted documents automatically through the supervised learning Approach.

Algorithm for key phrase extraction:

INPUT: Text Document (.txt)
OUTPUT: Key Phrases (.key)
BEGIN
Train ( )
Create Model ()
Testing ( )
Getting the Input file with extension of ".txt".
ExtractKeyphrase ( )
END

Modules and components

Stemming: The process of reducing the derived words to root words is called stemming. Stemmer has great influence in the process of key phrase extraction. A stemming algorithm reduces the words "fishing", "fished" and "fisher" to the root word "fish". This avoids confusion in extracting key phrases because it extracts same key phrase for many times if stemming is not applied. Therefore, stemming is applied before the key phrase extraction. The given text document is initially stemmed by the stemmer and the stemmed file is given to the key phrase extractor for key phrase extraction.

Key phrase extractor: Key phrase extractor extract the key phrase from the stemmed file. Supervised Naive Bayes' approach is used to extract key phrases. Key phrases provide a short view about the document. Today, there are many documents in the Internet, It is essential to assign key phrases for that document. The significance of key phrases lies here. Key phrases are particularly useful

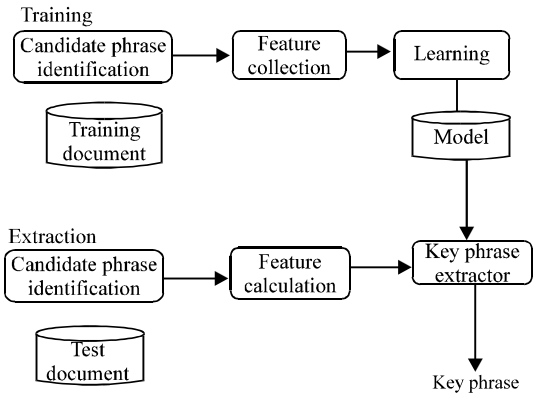

Fig. 1: Proposed system architecture

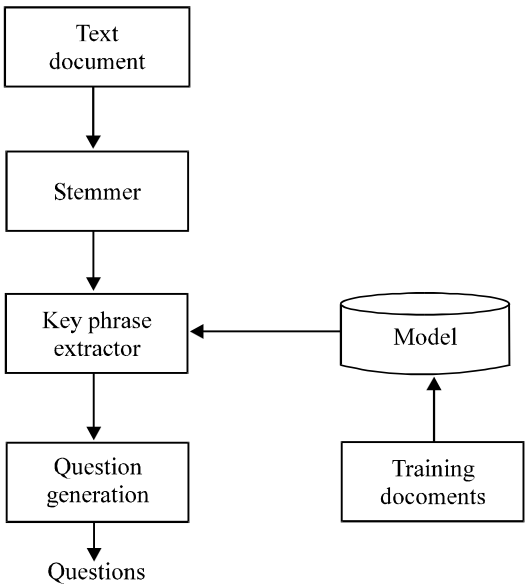

Fig. 2: Block diagram for key phrase extraction

because, they can be computed individually and independently of each other. In addition, key phrases can help users get a feel for the content of a collection. Key phrases are usually chosen manually. researchers assign key phrases to documents they have written. However, the great majority of documents come without key phrases, and assigning them manually is a tedious process that requires knowledge of the subject matter. Automatic extraction techniques are potentially of great benefit. Key phrase extraction chooses key phrases from the text itself. In this approach, the training data is used to train the key phrase extractor. The block diagram for key phrase extractor is shown in Fig. 2. The proposed Key Phrase Extraction algorithm has following works.

Training: Create a model for identifying key phrases using training documents where the training documents contain the key phrases for document.

Extraction: Extract key phrases from a new document using the model that is created. Both stages choose a set of candidate phrases from their input documents and then calculate the values of certain attributes for each candidate.

Supervised Naive Bayes' method: Naive Bayes is a supervised probabilistic classifier and it works based on Bayes theorem. This classifier is chosen because it is an independent feature model. Bayes method assumes that the presence or absence of a particular feature is unrelated to the presence or absence of any other feature given the class variable. The advantage of this method is it requires only a small amount of training data. To calculate the parameter necessary for classification.

Bayes' theorem:

$$P(C|F1,....,Fn) = \frac{p(C)p(F1,...,Fn|C)}{p(F1,...,Fn)}$$

Where:
C = Dependent class variable
F1,..,Fn = Features

Applying this theorem to key phrase extraction:

$$p(kp1) = P(kp1)\times p(lang|\ kp1)\ p(sw|kp1)\times p(maxpl\ |kp1)\ p(minpl\ |kp1)\ p(minno\ |kp1)/All\ features$$

$$p(kp2) = P(kp2)\times p(lang|\ kp2)\ p(sw|kp2)\times p(maxpl\ |kp2)\times p(minpl\ |kp2)\times p(minno\ |kp2)/All\ features$$

Where:
kp1 = Key phrase 1
kp2 = Key phrase 2
lang = Language
sw = Stop word
maxpl = Maximum phrase length
minpl = Minimum phrase length
minno = Minimum no of occurences
All features = (kp1) p(lang| kp1) p(sw| kp1) p(maxpl |kp1) p(minpl |kp1) p(minno |kp1)+(kp1) p(lang| kp1) p(sw| kp1) p(maxpl |kp1) p(minpl |kp1) p(minno |kp1)

Software description:
Eclipse kepler: Eclipse kepler is multi-language Integrated Development Environment (IDE) for software development. It comprises of an in-built text editor, several compilers for various languages and also, an extensible plug-in system. It can support the development and deployment of web services. And also supports some of the more popular languages such as C/C++, PHP, Java, Python and Android SDK.

Implementation: The central aim of designing the system is to generate questions from the given word document. Implementation is done by using Eclipse Kepler. The input given to the system is text document. Key phrases are to be extracted from the text document. To extract key phrases, supervised Naive Bayes method is used. In supervised learning, training should be given to the system. The training dataset used here contains 25 text files and 25 key phrase files.

The parameters used in extracting key phrases are language, stop word, maximum phrase length, minimum phrase length and minimum number of occurrences. Supervised naive Bayes algorithm works based on Bayes algorithm. It uses the five parameters to calculate the probability for each key phrase. Naive Bayes method is distinct from other supervised learning methods in the aspect of calculating the probability for individual key phrases.

The dependent Class variable (C) in the Bayes theorem is the set of all key phrases whose individual probability has to be calculated. The features (F1,...,Fn) are the parameters to extract the key phrases where the parameter language is constrained to only English, stop words like is, or, of, the are omitted maximum and minimum phrase length and minimum no of occurrences are already given to the system.

The probabilities calculated for the key phrase with the parameters is done using hash map. The <key value, feature> is the representation of hash map. In the system the hash mapping is done by <key phrase, parameter>. Finally, the key phrases with highest probabilities are chosen to generate.

The key phrases generated for a text file called "a.txt" will be stored in "a.key". The key phrases retrieved from that text file are salinity, salt tolerance, irrigation, desertification, developed countries, scientists, crops, planting, agriculture and India which are shown as generated key phrases:

- Salinity
- Salt tolerance
- Irrigation
- Desertification
- Developed countries
- Scientists
- Crops
- Planting
- Agriculture
- India

Table 1: Performance of the key phrase extractor

Key phrases extracted	Average matches with author key phrases
5	0.93
10	1.39
15	1.68

RESULTS AND DISCUSSION

The proposed system extracts the key phrases after stemming is done so there is no confusion in the extraction. So, the extracted key phrases will be free of duplication. The key phrases which are extracted have higher probability than other. Key phrases are extracted based on the parameter such as language, stop words, maximum phrase length, minimum phrase length, minimum number of occurrence. The extracted key phrases are specific because the system is constructed using the Naive Bayes' approach supervised probabilistic classifier is an independent model. In rare cases, some unimportant key phrases may be extracted. Table 1 shows the performance of the key phrase extractor.

CONCLUSION

Thus, the proposed system has so far implemented the stemming part using the Porter algorithm. Porter stemmer utilizes suffix stripping. The stemmer stems using the set of rules, transformations. The system then using the stemmed file starts extracting the key phrases based on the some features such as number of occurrences, maximum phrase length, maximum phrase length, stop word, language.

In the future research, the system will access the database to know more information about the key phrase. Using the matched study, the important sentences are extracted from the article based on the predefined phrases such as definition, challenges, improvements, recent development, objectives, references, pros and cons, comparing to, overcomes, implementation, problem addressed, purpose of, author, outcome, main idea, uses of, techniques, applications, difficulty, categorize into, etc. This process is called summarization. The sentences are pos-tagged based on the dynamic approach. Finally, questions are generated using the tagged sentences.

The computer-generated questions were perceived to be as pedagogically useful as human supervisor questions and more useful than generic questions. The questions are intended to prompt students to reflect on key concepts in their area of study.

REFERENCES

Barker, K. and N. Cornacchia, 2000. Using noun phrase heads to extract document Keyphrases. Proceedings of the 13th Biennial Conference of the Canadian Society for Computational Studies of Intelligence, May 14-17, 2000, Canada, pp: 40-52.

Megala, S., A. Kavitha and A. Marimuthu, 2013. Improvised stemming algorithm-TWIG. Int. J. Adv. Res. Comput. Sci. Software Eng., 3: 168-171.

Mihalcea, R. and P. Tarau, 2004. TextRank: Bringing order into texts. Proceedings of the Conference on Empirical Methods in Natural Language Processing. http://www.citeulike.org/user/johnkork/article/430523.

Osinski, S., J. Stefanowski and D. Weiss, 2004. Lingo: Search results clustering algorithm based on singular value decomposition. Proceedings of the International Conference on Intelligent Information Systems (IIPWM), May 17-20, 2004, Zakopane, Poland, pp: 359-367.

Porter, M.F., 1980. An algorithm for suffix stripping. Program Electron. Lib. Inform. Syst., 14: 130-137.

Turney, P.D, 2000. Learning algorithms for keyphrase extraction. J. Inform. Retriev., 2: 303-336.