Unsupervised Method for Processing Unstructured Dataset for Multilinguals

N. Vivegapriya.
Department of Computer Science and Engineering.
University College of Engineering
BIT Campus , Trichy District

A. Monika
Department of Computer Science and Engineering.
University College of Engineering
BIT Campus , Trichy District

***Abstract*—NLP –Natural language processing a major domain of processing unstructured dataset (text). One of such type of the task is to extract keyphrase or keyterms or keywords from huge sized text. The term Keywords/Keyphrases/Keyterms are the words which gives the incisive description of the content of the document.In spite of wide researches still this extraction function is in relatively poor performance.partly due to the selection of "correct" set of keyphrases. The fundamental difficulty lies in determining which keyphrases are the *most* relevant and provide the *best* coverage.In this paper, we propose an unsupervised method to extract keywords from a document. This method extracts unigram nouns (candidates) by applying preprocessing steps on the text. Then graph based Unsupervised algorithms is applied to find "frequency of co-occurrence or semantic relatedness" between candidates .It selects a number of keywords from the highest scored candidate.**

Keywords*— *Text mining; keywords extraction; co-occurrence; ranking.

I. INTRODUCTION

Keyword is the smallest unit that can express the meaning of a text. Keywords summarize the content of the document by few selected words [1]. They are easy to define by human, revise, remember and share. Keywords have been used in several tasks, such as information retrieval [2] document retrieval, document clustering [3], document classifying [4], indexing [5], summarization, and topic detection.

Documents such as scientific publications contain a list of keywords explicitly assigned by authors. However, most of other documents have no keywords assigned to them . Manual assignment of keywords is labor intensive, time consuming and error prone. Several automatic keyword extraction methods have been proposed. These methods have been divided into four categories in statistical, linguistic, machine learning and other methods and into three categories in
statistical, linguistic, and mixed methods. The latter categorization is more appropriate because machine learning methods are also based on statistical or linguistic knowledge to learn the model and it is not standalone category.Researchers have devised a plethora of methods for distinguishing between good and bad (or *better* and *worse*) keyphrase candidates. They majorly concern and concentrate o **frequency statistics**, such as TF*IDF or BM25, to score candidates, assuming that a document's keyphrases tend to be relatively frequent within the document as compared to an external reference corpus. However according to researchers important may also less frequently occurred and taking frequency statistics give mediocre in performance.

Our goal is to automatically identify the frequency of cooccurence or semantic relatedness that might be found in text documents.. This ontology can be used in many applications, such as Information Retrieval, Information Extraction, Question answering focusing on computing domain. For this purpose, we propose a methodology, which combine Natural Language Processing (NLP) and Matching Learning.

In this paper we propose Automatic keyphrase extraction as typically a two-step process: first, a set of words and phrases that could convey the topical content of a document with concise description, then these candidates are scored/ranked and the "best" are selected as a document's keyphrases.

Our key contribution are as follows: (i) Candidate identification: a brute-force method might consider *all* words and/or phrases in a document as candidate keyphrases. Moreover according to the rule of thumb and with respect to the computational cost,instead of taking all the ngrams(n umber of words in the document) there are some key terms which convey the concise description equally to the huge text.Common heuristics include removing stop words and punctuation; filtering for words with certain parts of speech or, for multi-word phrases, certain POS patterns; (ii) Keyphrase selection: Unsupervised machine learning methods attempt to discover the underlying structure of a dataset without the assistance of already-labeled examples ("training data").This approach uses the graph based ranking algorithm or method ,in which the words or phrase are scored as important in terms of number of cooccurence or relatedness with other existing words. This method assumes that important pharses or terms are more related with other terms (candidates), and that more of those related candidates are *also* considered important

The rest of this paper is organized as follows: section 2 examines related work and overviews a sample of NLP applications and IE systems; section 3 introduces the proposed methodology; section 4 illustrates the experimental results; section 5 discusses conclusions and future works.

II. RELATED RESEARCH

Information extraction is an important research topic in NLP, especially relevant to extracting semantic-oriented data. Y.Jie et al [6] focused on semantic rules to build Extraction system from LIDAR (Light Detection and Ranging).

F.Gomez et al [7] created a interpreter which relay o semantic relation between the terms. To build the knowledge base on the given unstructured text, grammatic relation are taken as parameter.

G. Kongkachandra et al [8] proposed semantic based key-phrase recovery for domain-independent key-phrase extraction. In this method, he added a key-phrase recovery function as a post process of the conventional key-phrase extractors in order to reconsider the failed key phrases by semantic matching based on sentence meaning.

Z.Goudong et al [9] proposed tree model kernel-based method with rich semantic information structure for the extraction of semantic relations between named entities.

A.B.Abacha et al [10] built a platform MeTAE (Medical Texts Annotation and Exploration). This system allows the extracting and annotating of Medical entities and relationships from Medical text.Pattern of each possible combination of words are constructed to build the medical text files.

A.D.S.Jayatilaka et al [11] constructed ontology from Web pages. He introduced web usage patterns as a novel source of semantics in ontology learning. The proposed methodology crossbars web content mining with web usage mining in the knowledge extraction process.

H.Li et al [12] extract semantic relations between Chinese named entities based on semantic features and the Vector Space Model (VSM).

Those research attempts were meant to identify semantic relations from web documents or text files in order to implement ontology. They either used NLP processing techniques, the statistical method, or the machine learning approach in the ontology learning process. Our research combines Natural Language Processing with Matching Learning for identifying and extracting syntactic, semantic relations between instance data based on domain specific ontology.

III. PROBLEM DEFINITION

The fundamental difficulty lies in determining which keyphrases are the *most* relevant and provide the *best* coverage. Human-labeled keyphrases are generally considered to be the gold standard, humans disagree about what that standard is! .A general heuristics seeks ,that selected keyphrases or keyterms should cover all the topics and should provide good coverage of the content.

IV. SYSTEM DESIGN

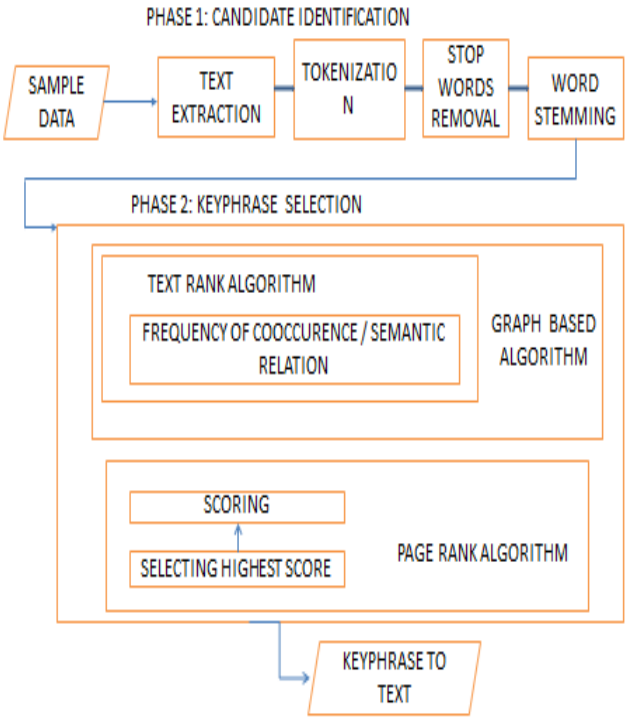

Fig. 1. KeyPhrase Selection Model

V. PROPOSED SYSTEM

There are two main steps involved in our method (see Fig. 2), which are as follows:

A. Candidate Identificatio:Data Collection and Pre-processing

1. Tokenization

Breaking a stream of text up into words, phrases, symbols, or other meaningful elements called tokens. The count of tokens becomes input for further processing such as parsing or text mining. It is also a form of lexical analysis used to form the text segmentation.

2. Stop word Removal

Words which are filtered out before or after processing of natural language data (text).Other search engines **remove** some of the most common **words**—including lexical **words**, such as "want"—from a query in order to improve performance.

3. Stemming

To minimize the replication form of words ,they are derived to their stem (words to words stem) or likely to say as root word or base word.The stem may won't be identical to the morphological root of the word; it is usually sufficient that related words map to the same stem, even if this stem is

not in itself a valid root. Many search engines treat words with the same stem as synonyms as a kind of query expansion, a process called conflation.

All the above preprocessing steps have been achieved by POS tagging: When people do manually assign keywords, the majority of the selected words are either nouns or noun phrases . Therefore, we extract unigram nouns as candidate keywords by applying POS tagging on the text fragments POS assigns parts of speech such as noun, verb, and adjective to each word in the text based on its definition, and relationship with adjacent and related words in a phrase, sentence, or paragraph.In this paper, rather than taking all of the n-grams (where $1 \leq n \leq 5$), we might limit ourselves to only noun phrases matching the POS pattern {(<JJ>* <NN.*>+<IN>)? <JJ>* <NN.*>+}(a regular expression written in a simplified format used by NLTK's RegexpParser()). This matches any number of adjectives followed by at least one noun that may be joined by a preposition to one other adjective(s)+noun(s) sequence, and results candidates identification.

.

B. *KeyPhrase Selection*

In the TextRank algorithm, a text is represented by a graph. Each vertex corresponds to a word type. The weight w(ij) is assigned to the edges between vertices v(i) and v(j).Value is given with respect to the number of times that word cooccurs within the window (W of words) in the associated text.The goal is to (1) compute the score of each vertex, which reflects its importance, and then (2) use the word types that correspond to the highestscored vertices to form keyphrases for the text. The score for v(i) , S(v(i)), is initialized with a default value and is computed formula:

$$S(v(i)) = (1-d) + d * \sum_{v(j)\in Adj(v(k))} \frac{w(ji)}{\sum_{v(k)\in Adj(v(j))} w(jk)} S(v(j))..(2)$$

where Adj(vi) denotes vi's neighbors and d is the damping factor set to 0.85.Intuitively, a vertex will receive a high score if it has many high-scored neighbors. As noted before, after convergence, the T % top-scored vertices are selected as keywords. Scored and selected important words are then outputted as keyphrases.

Example of a text extraction

natural language processing for purposes of automatically extracting structured information from unstructured (text) datasets. One such task is the extraction of important topical words and phrases from documents, commonly known as terminology extraction or automatic keyphrase extraction. Keyphrases provide a concise description of a document's content; they are useful for document categorization, clustering, indexing, search, and summarization; quantifying semantic similarity with other documents; as well as conceptualizing particular knowledge domains.Automatic keyphrase extraction is typically a two-step process: first, a set of words and phrases that could convey the topical content of a document are identified, then these candidates are scored/ranked and the "best" are selected as a document's keyphrases.A brute-force method might consider all words and/or phrases in a document as candidate keyphrases. However, given computational costs and the fact that not all words and phrases in a document are equally likely to convey its content, heuristics are typically used to identify a smaller subset of better candidates. Common heuristics include removing stop words and punctuation; filtering for words with certain parts of speech or, for multi-word phrases, certain POS patterns; and using external knowledge bases like WordNet or Wikipedia as a reference source of good/bad keyphrases.Researchers have devised a plethora of methods for distinguishing between good and bad (or better and worse) keyphrase candidates. The simplest rely solely on frequency statistics, such as TF*IDF or BM25, to score candidates, assuming that a document's keyphrases tend to be relatively frequent within the document as compared to an external reference corpus. Unfortunately, their performance is mediocre; researchers have demonstrated that the best keyphrases aren't necessarily the most frequent within a document. (For a statistical analysis of human-generated keyphrases, check out Descriptive Keyphrases for Text Visualization.) A next attempt might score candidates using multiple statistical features combined in an ad hoc or heuristic manner, but this approach only goes so far. More sophisticated methods apply machine learning to the problem. They fall into two broad categories.Unsupervised machine learning methods attempt to discover the underlying structure of a dataset without the assistance of already-labeled examples. The canonical unsupervised approach to automatic keyphrase extraction uses a graph-based ranking method, in which the importance of a candidate is determined by its relatedness to other candidates, where "relatedness" may be measured by two terms' frequency of co-occurrence or semantic relatedness. This method assumes that more important candidates are related to a greater number of other candidates, and that more of those related candidates are also considered important; it does not, however, ensure that selected keyphrases cover all major topics, although multiple variations try to compensate for this weakness.

Example of a Preprocessing

```
{'automatic',
'categorization',
'concise',
'content',
'datasets',
'description',
'document',
'documents',
'document\xe2\x80\x99s',
'domains',
'extraction',
'important',
'information',
'keyphrase',
'keyphrases',
'knowledge',
'language',
'natural',
'particular',
'phrases',
'processing',
'purposes',
'search',
'semantic',
'similarity'}
```

EXAMPLE OF A KEYPHARSE SELECTION AND RANKING

```
{('candidates', 0.03820017392483226),
('document', 0.022861418617545197),
('document\xe2\x80\x99s', 0.014421997452979665),
('document\xe2\x80\x99s keyphrases', 0.021227557963146956),
('extraction', 0.019684861520724135),
('keyphrases', 0.028033118473314245),
('method', 0.014524792746291534),
('methods', 0.014559632668437891),
('phrases', 0.014302047055286171),
('words', 0.019892350943161358)}
```

VI. TRADING MODEL

A. *Choosing the Best Algorithm*

To analyse the best prediction algorithms in keyphrase extraction various experiments are examined. Many key factors of unstructured are considered to forecasts the keyphrase. The factors are co-occurrence with other words, noun distribution,relatedness with other words . The

performance is measured for the analysed algorithms. The best algorithm which forecasts the results is found and the further design is implemented. The results exhibit that Graph based algorithm is is the best predicting algorithm among the other algorithms in Unsupervised method.

B. Proposing a Prediciton

Tex-tRank's best score on the *Inspec* dataset is achieved when only nouns and adjectives are used to create a uniformly weighted graph for the text under consideration, where an edge connects two word types only if they co-occur within a window of two words. Hence, our implementation of Tex-tRank follows this configuration.

VII. CONCLUSION

We proposed an unsupervised method to extract keywords from text documents based on frequency of cooccurence or semantic relatedness between candidates and distribution of nouns over the text. We conducted various experiments using sets of keywords for each text document that is manually extracted by humans. The results show that our method outperforms (TextRank) by 13 % in precision, 6 % in recall, and 10 % in F-measure but TFIDF only by 11 % in precision, and 6 % in F-measure. We conclude that calculating cooccurence provided the best results. Distribution of nouns over the text is more effective feature than term frequency. Human selection for keywords has an obvious effect on the overall performance, where better F-measure results are achieved when human keywords are precise. Future work may focus on studying the effect of different similarity measures and clustering methods on keywords extraction from web pages.

REFERENCES

[1] G. K. Palshikar, “Keyword extraction from a single document using centrality measures”. In Pattern Recognition and Machine Intelligence. Springer Berlin Heidelberg 2007, pp. 503-510.

[2] X. Wan, J. Yang, and J. Xiao, “Towards an iterative reinforcement approach for simultaneous document summarization and keyword extraction”. In Annual Meeting-Association for Computational Linguistics vol. 45, no. 1, p. 552, June 2007.

[3] S. S. Kang, “Keyword-based document clustering”. In Proceedings of the sixth international workshop on Information retrieval with Asian languages, vol. 11, pp. 132-137. Association for Computational

[4] P. Tonella, F. Ricca, E. Pianta, and C. Girardi,. “Using keyword extraction for web site clustering”. IEEE: In Web Site Evolution. Theme: Architecture. Proceedings, pp. 41-48, September, 2003.

[5] A.Gupta, A. Dixit, and A. K. Sharma, “A novel statistical and linguistic features based technique for keyword extraction”.IEEE: In Information Systems and Computer Networks (ISCON), pp. 55-59, March 2014

[6] Y. Jie, Y. Haiquan, T. Ming, Z. Guoning, "Building Extraction from LIDAR based Semantic Analysis," Geo-Spatial Information Science, vol. 9, no. 4, Sep. 2006.

[7] F. Gomez, C. Segami, "Semantic interpretation and knowledge extraction," Knowledge-Based Systems, vol. 20, no. 1, pp. 51 - 60, July 2006.

[8] G. Kongkachandra, K. Chamnongthai, "Abductive Reasoning for Keyword Recovering in Semantic-based Keyword Extraction," in The Fifth International Conference on Information Technology: New Generations - IEEE, 2008, pp. 714 - 719.

[9] G. Zhou, L. Qian, J. Fan, "Tree kernel-based semantic relation extraction with rich syntactic and semantic information," Information Sciences, vol. 180, no. 8, pp. 1313 - 1325, Dec. 2009.

[10] A.B. Abacha, P. Zweigenbaum, "Automatic Extraction of Semantic Relations between Medical Entities- a rule based approach," Journal of Biomedical Semantics, vol. 2, no. 5, 2011.

[11] A.D.S Jayatilaka, "Knowledge Extraction for Semantic Web Using Web Mining ," in The International Conference on Advances in ICT for Emerging Regions (ICTer 2011) - IEEE, 2011, pp. 89 - 94.

[12] H. Li, X. Wu, Z. Li, G. Wu, "A Relation Extraction Method of Chinese Named Entities based on Location and Semantic Features," Applied Intelligence, vol. 18, no. 1, pp. 1- 14, May 2012.