Keyphrase Extraction as Topic Identification Using Term Frequency and Synonymous Term Grouping

Kwanrutai Nokkaew and Rachada Kongkachandra
Department of Computer Science,
Faculty of Science and Technology, Thammasat University,
Pathumthani, Thailand
kwanrutai.saclim@gmail.com and rdk@cs.tu.ac.th

***Abstract*—Keyphrase are usually used as a representative of in the document. This paper presents a method to improve keyphrase extraction by using synonymous term grouping. Topic identification is recognised by term frequency for keyphrase extraction. We utilize a language model including linguistic patterns and language knowledge such as morphology syntax. The language model is a probability of word sequence. The focus unit is a pattern of noun adjective combination The proposed method consist of five processes i.e. preprocessing, candidate selection, semantic-based topic clustering, topic ranking, and keyphrase selection. This experimental result has precision value 54.44 from dataset of IEEE and 39.99 from dataset of SamEval.**

Keywords—Topic Identification; Keyphrase; linguistic model; TF-IDF

I. Introduction

The keyphrases are a combination of words significantly representing of important concepts of a document. A task to automatically extract keyphrases from a document called keyphrase extraction is to select or generate a word or multi-word that represents significant concepts from the content within document. A keyphrase extraction is i.e. keyphrase assignment and keyphrase extraction. Assignment of keyphrase summarize contents and generate terms from summerization. As a result of this, the given keyphrase may exist in the given content. On the other hand, keyphrase extraction focuses only on the words in the document and selects many them as representatives.

This paper studies on a task of keyphrase extraction using as unsupervised approach [1]. In the past, several works. TextRank, The work of the graph based ranking consider the frequency of words that appear in the document. The word derived from the document and the graph based ranking is applied. The frequency of words maybe not be different. Keywords that are more important maybe not a few appear little. So this is the weakness of the textrank. [1] From the work of textrank, the word or phrases that have a candidate keyphrase a lot of many because the textrank does not have a group of words. However, in Topicrank's fix to this problem of textrank by grouping words with have suface the same into the same group. Because of this, there are weaknesses. If document have words same meaning. But that words are assigned may have different group. And may not be selected as a keyphrase. [3] And since topicrank's work has been resolved by focusing on words that have the same meaning assign into same group and can give to higher results. [2] were proposed to this task. Those existing works mainly applied frequency of terms in a document along with additional features such as position of words and co-occurrence of nearby terms. However, their accuracy is relatively low due to complexity of terms in context such as synonymous words. Recently, [2] proposed an improvement version that considers a semantic meaning of words in grouping keyphrase using relation from WordNet, and it was reported to yield better accuracy. Unfortunately, some issues remain for including the missing keyphrase candidates in extraction process and inappropriate ranking of the selected keyphrase.

In this work, we propose a method to solve the issues. Linguistic knowledge of noun phrase is applied to disambiguate phrase boundary in extraction. This is expected to help on scoping a candidate for extracting keyphrase and preventing inclusion of improper terms in phrase scoring. Moreover, keyphrase ranking method is adjusted for improving a ranking result of the found keyphrase to properly rank significant keyphrase from highest. The rest of this paper is organized as follows. Section 2 provides relevant works in a task of keyphrase extraction. Section 3 gives a detail on the proposing method on using linguistic knowledge to improve semantic-based keyphrase extraction. Experimental results are given in Section 4. Section 5 provides a summary of this paper.

978-1-4673-9077-4/17/$31.00 ©2018 IEEE

II. Literature review

In this section, we provide related knowledge to the proposed method. First, existing works on keyphrase extraction are reviewed especially a semantic based keyphrase extraction that we aim to improve. Second, a technique that we apply in this work is also summarized including term frequency and semantic similarity.

A. *Existing keyphrase Extractions*

In a keyphrase extraction task, there are two main processes as an extraction process and ranking process. An extraction process focuses on detecting terms given in a document and selecting the significantly important terms among them as keyphrase. A ranking process is then to calculate the selected terms into a rank to determine the highest important terms as document representatives. Most of the existing works applied unsupervised method because of a great difference of used terms in documents resulting in unspecified scope of possible words.

The approach in overview methods of Keyphrase Extraction. Automatic keyphrase extraction have three brand. The first is approaches of keyphrase extraction. Then, it is represent of a data. Finally, it is scope of the data for extract. These three methods are described in the following order. The first is linguistic approaches use expert knowledge and linguistic rules. It develops simple because it uses rules. But mostly not covered all the used terms. Then, statistical Approaches use the information that appears in the document. The numerical model is derived from the document. Such as frequency of term, position of term no need to rely on language proficiency. Keyphrase extraction use the information contained in the document only. Finally, Machine learning approaches use information to learn. Trying to model the keyphrase of a document with learning of documentation and it's need a lot of document. Overview methods of Keyphrase Extraction part two is representation of data. Then, there are statistical representation, are vector and graph. Finally, it is a scope of extraction. The scope can be a working on extracting from single document or extracting from a corpus. They are given in Figure 1

The famous works in this task include Topic Rank and Graph-based Ranking while recently semantic based keyphrase extraction was proposed. The Topic Rank is a classical baseline in the task. [3] It detects terms in given documents in a stem level to reduce variation of terms from suffixes and inflections. Statistical frequency of terms plays important role in term selection and ranking. The work can find keywords and phrases as topic to represent a document based on terms in surface level and frequency; however, its accuracy is yet unsatisfied comparing to human evaluation. Later on, Graph-based Ranking was developed. The core method of the work is to create a graph to represent term relation in a graph. In a graph, terms are linked based on frequency and co-occurrence position as a clue for determining a group of terms and co-occurrence. But in terms of distance value between can't explain the meaning of the words significantly. In this research, we want to study to the extract keyphrase which consider combine between frequency and grouping semantic candidate keyphrase.

Recently, semantic based keyphrase extraction was developed. The work attempted to improve on grouping different surface words that may be synonymous or have related meanings. WordNet was exploited as semantic resource to expand words into meanings. [4] Thus, a grouping of terms in a similar meaning is managed as a keyphrase. The remaining issues from this work are as follows. First, it does not specify scope of targeted phrases in a document; thus all words in a document are processed with WordNet to expand their meaning. This results in too many candidates for a keyphrase and increasing in selection complexity. Moreover, when different terms with similar meaning are grouped, a representative of the group has to be carefully chosen as a core concept, but a method to do so is yet invented and applied.

This paper aims to solve the remaining issues from Semantic-Based for keyphrase Extraction to improve an accuracy. A scope of phrases is defined to limit and reduce terms for consideration. A method to select keyphrase sentence within from document. It is representative by a group is proposed.

This work applies two main components including WordNet and Term-Frequency (TF). In this section, they are summarized as a background knowledge. [5]. WordNet is an English lexical resource that informs relation of words in a semantic level. In WordNet, entries are given as a set of words with the same meaning (synonym) called 'Synset'. Synsets are mainly linked to another to form a hierarchical structure. From the structure, a hypernym-hyponym relation can be derived. Hypernym (superset) refers to a concept that is more generalized while hyponym (subset) is a concept in specification.

Term-Frequency (TF) is a popular method to find important terms in a set of documents by considering a frequency of co-occurred terms. Segmented terms, which can be in syllable, word or phrase, in a document are generated in a vector for defining their frequency. Generally, terms with more frequency in a certain document are more important than the low frequent terms [6].

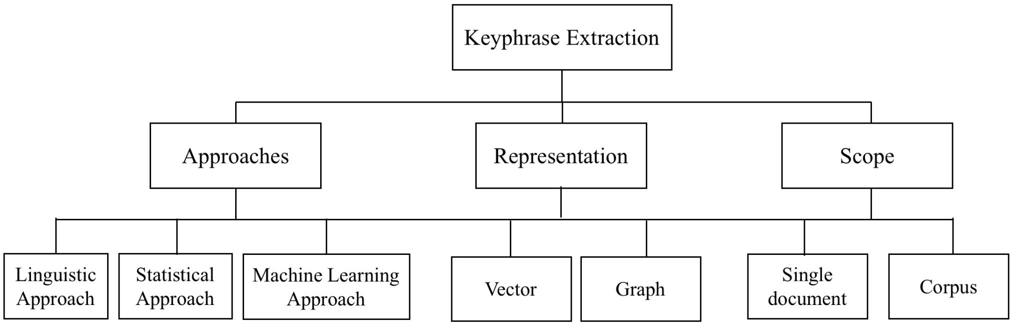


Figure 1: Overview methods of Keyphrase Extraction

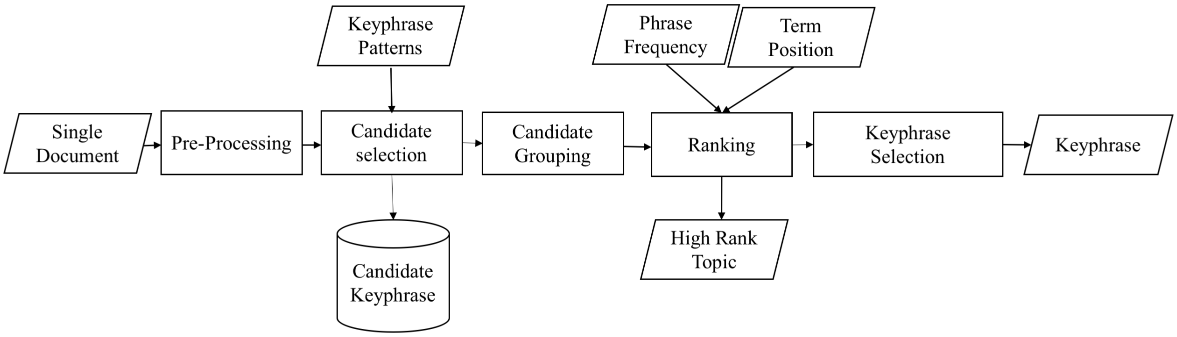
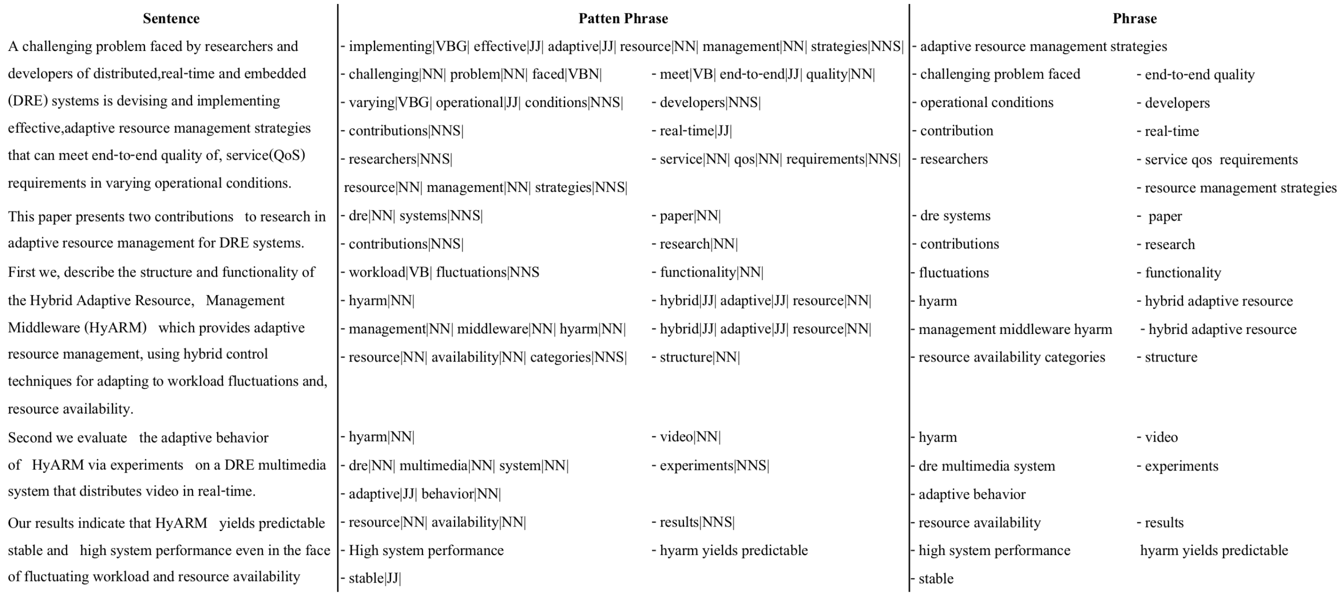


Figure 2: Overview System



Sentence	Patten Phrase		Phrase	
A challenging problem faced by researchers and	- implementing|VBG| effective|JJ| adaptive|JJ| resource|NN| management|NN| strategies|NNS|		- adaptive resource management strategies	
developers of distributed,real-time and embedded	- challenging|NN| problem|NN| faced|VBN|	- meet|VB| end-to-end|JJ| quality|NN|	- challenging problem faced	- end-to-end quality
(DRE) systems is devising and implementing	- varying|VBG| operational|JJ| conditions|NNS|	- developers|NNS|	- operational conditions	- developers
effective,adaptive resource management strategies	- contributions|NNS|	- real-time|JJ|	- contribution	- real-time
that can meet end-to-end quality of, service(QoS)	- researchers|NNS|	- service|NN| qos|NN| requirements|NNS|	- researchers	- service qos requirements
requirements in varying operational conditions.	resource|NN| management|NN| strategies|NNS|			- resource management strategies
This paper presents two contributions to research in	- dre|NN| systems|NNS|	- paper|NN|	- dre systems	- paper
adaptive resource management for DRE systems.	- contributions|NNS|	- research|NN|	- contributions	- research
First we, describe the structure and functionality of	- workload|VB| fluctuations|NNS	- functionality|NN|	- fluctuations	- functionality
the Hybrid Adaptive Resource, Management	- hyarm|NN|	- hybrid|JJ| adaptive|JJ| resource|NN|	- hyarm	- hybrid adaptive resource
Middleware (HyARM) which provides adaptive	- management|NN| middleware|NN| hyarm|NN|	- hybrid|JJ| adaptive|JJ| resource|NN|	- management middleware hyarm	- hybrid adaptive resource
resource management, using hybrid control	- resource|NN| availability|NN| categories|NNS|	- structure|NN|	- resource availability categories	- structure
techniques for adapting to workload fluctuations and,				
resource availability.				
Second we evaluate the adaptive behavior	- hyarm|NN|	- video|NN|	- hyarm	- video
of HyARM via experiments on a DRE multimedia	- dre|NN| multimedia|NN| system|NN|	- experiments|NNS|	- dre multimedia system	- experiments
system that distributes video in real-time.	- adaptive|JJ| behavior|NN|		- adaptive behavior	
Our results indicate that HyARM yields predictable	- resource|NN| availability|NN|	- results|NNS|	- resource availability	- results
stable and high system performance even in the face	- High system performance	- hyarm yields predictable	- high system performance	hyarm yields predictable
of fluctuating workload and resource availability	- stable|JJ|		- stable	

Figure 3: Selection of keyphrase candidate

Table I: Example Group of Candidate keyphrase

Topic	Group of Keyphrase
Systems	[real-time, real-time systems, dre systems, n multimedia system, system, high system, high system performance, system performance, applications organization, organization,dre multimedia, multimedia]
Resource	[adaptive resource, adaptive resource management, resource, resource management, resource availability]
Control	[hybrid control, hybrid control techniques, control, control techniques]
Service	[service, service qos, availability, availability categories]
Management	[management, management strategies, management middleware]

III. Improving keyphrase Extraction using semantic

A. Preprocessing

The proposed system is divided into three sub-process i.e. tokenization, lemmatization, and part-of-speech tagging. Tokenization is firstly used for separating a sentence into terms using punctuations. Then, lemmatization is appliced to convert terms into their original forms. Finally, each term is labeled with its word function such as noun, verb, adjective by using Stanford tagging. They are given in Figure 3 [7]

B. Candidate Selection

For prevention of discrimting the same meaning terms, WordNet is applied to help in grouping terms with the same conceptual meaning. WordNet can explain limits on the synset of term. (both hypernym and hyponym). For example of hierarchical relation, ”Hotel”, ”Resort” and ”Guesthouse” are related in WordNet as synset. ”Van” and ”Vehicle” are related in WordNet as hypernym and hyponym but not synonym. The relation can help to indicant that these two terms are closely similar, the writing document can use interchangeably. Thus, the relation from WordNet can into groups semantically the terms. This process is designed to group candidates with same meaning and later process counts frequency of term. They are given in Figure 2

Once grouping, candidate keyphrase are selected within the topic. The selection perform best on the term appearance in a document. The first appeared and most frequency candidate. The chosen as representative of the topic. From the document, a list of candidate keyphrase groups is noun created. Examples are given in Table 1.

C. Ranking by Term Frequency

Ranking in this paper used frequency of frequency keyphrase Candidate occurrence within a document. The procedure of calculation is divided into third step. The first step counts frequency keyphrase Candidate of each keyphrase within group. The second step calculates a sum of frequency keyphrase Candidate for representation score of group. The third step calculates adjustment normalization about score of keyphrase Candidate of each keyphrase within group.

$$d = (p_1, p_2, \ldots, p_n) \tag{1}$$

$$d = (t_1, t_2 \ldots, t_k) \tag{2}$$

d = The document within the corpus n = number of phrases in document. k = number of topics in document. p = The phrase derived from defining feature within a document.

$$Tj = (Fjp_1, Fjp_2 \ldots, Fip_n) \tag{3}$$

FP = The frequency of the phrase that appears in the document. Sum Score is sum the scores for each phrase that appears in the document and within the topic.

$$Scoreoftopic = \frac{\sum_{i\epsilon 1}(Fjpi)}{len(t_j)} \tag{4}$$

$$Fjpi = \{0, phrase\ f, frequency \tag{5}$$

Score of topic is sum score divide by the total number of phrases within the topic.

D. Keyphrase Selection

This step is to find keyphrase for representative a document. It selects five best groups. We consider keyphrase candidate to selects first appeared term is chosen document. [3]

$$KeyphraseSelection = (2 \times frequency + position) \tag{6}$$

IV. Experiment

A. Experiment Setting

1) Data set: We use two standard evaluation datasets which are English abstract of journal papers. The first corpus is SemEval dataset composing of 144 document by Kim et al (2010) Second corpus is IEEE dataset composing of 380 document

2) Evaluation metrics: The paper uses precision, recall, f-measure calculate (7,8,9) A result is examined for the missing values compare between the Answers Human and Answer Automatic from System. This is to see system can answer or extraction keyphrase as close to the answer Human.

$$precision = \frac{correct}{output - length} \tag{7}$$

Recall is another evaluation of the accuracy of the model. Recall is a fraction of relevant item that are successfully retrieved.

$$recall = \frac{correct}{reference - length} \tag{8}$$

Last, F-measure is selected to represent capability of the proposed method.

$$f - measure = \frac{precision * recall}{(precision + recall)/2} \tag{9}$$

Correct : number of cases correctly identified as system Output-length : the number overall of cases correctly identified by system reference-length : the number overall of cases correctly identified by human

B. Experimental Result

1) Missing Value: The goal standard is answers of human. Human answer will be answer the phrase appearing inside the document and may come to be the human answer summarized from the document. And they are maybe without this phrase within the document. So, they are difference between system answers and goal standard. It is missing value. The problem-solving missing value can improve to good. It will make high accuracy. This experiment has result improvement. The example are given in Table 2.

2) Precision and Recall: Precision and Recall: Automatic Evaluation compute similarity between goal standard and answer of system. Precisions will be focus the number of phrases extracted from system. This experimental result has precision value 54.44 from dataset of IEEE and 39.99 from dataset of SamEval. And recall value is 55.51 from dataset of IEEE and 57.28 from dataset of SamEval. The results of experimental provide better. When compare between of our experimental and Semantic-Based Clustering. The example are given in Table 3.

Table II: Missing Value between IEEE and SemEval corpus

Corpus	Document			Keyphrase		
	Methods	Number of documents	Tokens average	Total	Average	Missing
IEEE EnglishAbstract	Topic Ranking	380	72.87	14607	38.41	42.00
	Our work	380	68.72	16570	43.21	34.35
SemEval EnglishAbstract	Topic Ranking	144	85.01	6251	43.41	18.87
	Our work	144	98.45	7456	48.34	9.09

Table III: Precision and Recall Result

Corpus	Algorithm	Precision	Recall	F-measure
IEEE	Semantic-sum within Topic	54.44	51.51	51.43
	Semantic-Based Clustering	19.07	4.93	7.46
SemEval	Semantic-sum within Topic	39.99	57.28	46.05
	Semantic-Based Clustering	21.20	11.97	14.88

V. Conclusion

This paper is improved of semantic-based keyphrase extraction. We propose a method of extract keyphrase with apply statistical and linguistic knowledge approache. This paper focus of noun phrase structure. The candidates keyphrase consider frequency with 3 or more occurrences in a document. We will consider the last word of the phrase or called core noun. The core noun for find synonym words. That apply by WordNet. The words have similar of meaning will keep in the same group. Grouping by means of words with similar meaning are the same group. The solution a words written differently but have the same meaning. This grouping method considers both the same written words and the different written words. We select the representative of the group choose from frequency top five and the word appears first in the document. This experimental result has precision value 54.44 from dataset of IEEE and 39.99 from dataset of SamEval. The recall value is 55.51 from dataset of IEEE and 57.28 from dataset of SamEval. And f-measure value is 51.43 from dataset of IEEE and 46.05 from dataset of SamEval. Moreover, the number of missing keyphrase is apparently lower. The missing values better than Topicrank- Graph-based topic ranking for keyphrase extraction. and Semantic-Based for Topic Identification Keyphrase Extraction. Sometime keywords have few frequently appear in the document. It may not be selected from this method. Which can solve the problem in the next order.

References

[1] K. S. Hasan and V. Ng, "Automatic keyphrase extraction: A survey of the state of the art." in *ACL (1)*, 2014, pp. 1262–1273.

[2] R. K. Kwanrutai Nokkaew, "Semantics-based topic identication for keyphrase extraction." pp. 138–142, 2017.

[3] A. Bougouin, F. Boudin, and B. Daille, "Topicrank: Graph-based topic ranking for keyphrase extraction," in *International Joint Conference on Natural Language Processing (IJCNLP)*, 2013, pp. 543–551.

[4] R. Navigli, “Word sense disambiguation: A survey,” *ACM Computing Surveys (CSUR)*, vol. 41, no. 2, p. 10, 2009.

[5] K. Chen, Z. Zhang, J. Long, and H. Zhang, “Turning from tf-idf to tf-igm for term weighting in text classification,” *Expert Systems with Applications*, vol. 66, pp. 245–260, 2016.

[6] M. Lan, C. L. Tan, J. Su, and Y. Lu, “Supervised and traditional term weighting methods for automatic text categorization,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 4, pp. 721–735, 2009.

[7] A. Naseem, M. Anwar, S. Ahmed, A. Jan, and A. K. Malik, “Reusing stanford pos tagger for tagging urdu sentences,” in *2017 13th International Conference on Emerging Technologies (ICET)*, Dec 2017, pp. 1–6.