# Semantics-Based Topic Identification for keyphrase Extraction

Kwanrutai Nokkeaw and Rachada Kongkachandra
Department of Computer Science
Thammasat University, Pathum Thani, Thailand
kwanrutai.saclim@gmail.com, rdk@cs.tu.ac.th

***Abstract*—keyphrase are usually used as the representative of a document since they represent the most relevant information contained in the document. However, there are many documents are not provided any keyphrase, an automatic keyphrase extraction is therefore necessary. Topic-Rank is recognized as the good graph-based keyphrase extraction. There are five processes i.e. preprocessing, candidate selection, topic identification, topic ranking, and keyphrase selection. Unfortunately, there are about 15.93% of keyphrase are missing. We hypothesize that the errors are from the topic identification process. Since the topic is identified by grouping candidates having 25% overlapping word stems, some irrelevant candidates having the similar form are selected. This paper proposes the semantic similarity among candidates as criteria for selecting candidates into the same group. The semantic knolwedge base used in this paper is the WordNet. The experiments are performed by comparing the our results with TopicRank in four data sets. The results reveals that the precision, recall and F1-measure are significantly improved.**

***Keywords*—*Keyphrase exteaction, Topic Identification, Topic Ranking, Semantic similarity***

## I. Introduction

Keyphrase are usually used as the representative of a document. Due to the keyphrase compactness, the reader could understand the content without reading the document. However, many documents retrieved from the internet are not provided the keyphrase. These are two way to define a set of keyphrase i.e. keyphrase assignment and keyphrase extraction. Keyphrase assignment is a take that an author words or phrase to express the meaning of a document. [1]. The strength of this approach is all keyphrase are related to the document content. Since keyphrase assignment selects keyphrase by considering their meaning, it needs the human expert and very difficult to make the automatic system. keyphrase extraction is another approach that selects a set of keyphrase from the existing words in the document. To implement an automatic keyphrase extraction system is possible and practical, however the extracted keyphrase may not related to the document meaning. Keyphrase extraction is often used in many applications of natural language processing such as information retrieval, text classification, document clustering and text summarization. The challenges of keyphrase extraction are depended on document length, document structure and document domain. There are several approaches used in keyphrase extraction statistical, linguistic-based and machine learning based. [2]. In statistical-based approach, the occurrences of keyphrase patterns like n-grams are counted and then calculated for finding their scores Term Frequency, Term Frequency-Invers Document Frequency(TF-TDF) and word co-occurrence frequency are offen applied as keyphrase scores. [3] [4]. Although, the statistical approach can yield high prefermances in many works, some keyphrase in the domains as medical and scientific are not occurred frequency. Lingnistic-based keyphrase extraction employs analysis rules in several levels to extract a set of keyphrase such as lexical rules, syntactical rules, semantic rule and pragmatic rules. keyphrase from this approach are usually relevant to the document meaning, however it spends time and experts to define the analysis rules. The latter keyphrase extraction approach, using machine learning, is very popular because it is flexible and adaptive. Many varieties of keyphrase patterns are collected and then used in training process to provide the keyphrase models. There are two sub-catagories i.e. supervised and unsupervised. The examples of keyphrase extraction researches in supervised machine learning are [5]. The limitation of supervised machine learning approach is the difficult and time-consuming to prepare the training data. Both data and their annotations are required. The another machine learning approach, unsupervised, is increasingly applied in many researches [6] [7]. This paper presents an unsupervised machine learning keyphrase extraction. The overall system is similar to extraction system proposed by Adrien Bougouin et all. [7]. Our work attempts to improve the accuracy by considering lexical semantic in Topic Identification process. The paper is organized in five sections. Section I is an introduction and keyphrase extraction reviews. The concept of using lexical semantic in graph-based keyphrase extraction is presented in section II. Methodology of considering semantic in Topic Identification is explained in section III. The experiments is demonstrated and results in section IV. Finally, the conclusion is stated in section V.

## II. Graph-based Keyphrase Extraction

Graph-based keyphrase extracton are developed from TextRank that has disadvantage about overlap of keyphrase. TextRank with a only considers noun and adjective of document. It count a phrase frequency that appears within documents. In TextRank, the relation is the link of two adjacent words within the documents. Then, a random walk algorithm [8] is used to rank all nodes. The top k terms are selected as keyphrase. SingleRank consider of documents in same domain. A document in same domain brings analysis keyphrase extraction. It is with a considerdtion of the long, two nouns together within document. SingleRank in [9] adds weighted score to each edge. The weight is co-occurrence frequency in a window of variable size w gredter than or equal to 2. The k combinations of words from the top ranked node are assigned as the document keyphrase. The main problem of

978-1-5386-2623-8/17 $31.00 © 2017 IEEE
DOI 10.1109/KICSS.2017.35

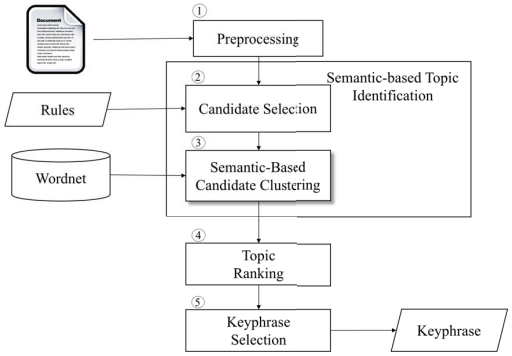


Fig. 1. The overview of the proposed keyphrase extraction system

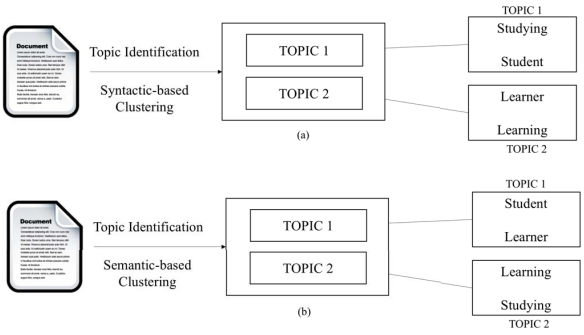


Fig. 2. The limitation of TopicRank candidate clustering

these approaches is the selected keyprases may not cover all topics in the document. TopicRank takes an interest to phrase that have word from in same. All three systems use Graph-Based. The baseline approach of this work is TopicRank proposed by Bougouin et al. in 2004 [7]. From literature reviews, TopicRank outperformed TextRank [6], SingleRank [9] and TextRank. All approaches represent each vertex for a word and each edge for the phrase relation. Bougouin et al. [7] proposed to firstly identify the document topics and then select terms within the top ranked topics as document keyphrase. The processing steps of TopicRank are preprocessing, topic identification, graph-based ranking, and keyphrase selection. The preprocessing step prepares the document in tokenizing, stop word removing, and part-of-speech (POS) tagging. The topic identification process is started by selecting terms tagged as noun and adjective as keyphrase candidates. These candidates are then grouped into clusters (topics) by considering the word stem similarity. After the topics are identified, the significant scores are then calculated based on graph-based ranking algorithm in [6]. The keyphrase located in the top ranked topic are selected. TopicRank can extract the document keyphrase with the 22.68% of F-score, which is the best one compared to the previous works. Even though the baseline approach is outperformed, we found some points to make the approach improved. In this paper, we focus on modify the step of topic identification. Topic Identification consists of two sub-processes: candidate selection and candidate clustering. In candidate selection, we select the candidates by considering In this paper a modification of topicranking is proposed. Figure 1 illustrates the overall process of the proposed graph-based keyphrase extraction.

### A. *Preprocessing*

In this step, a document is sequence into tokens using word-tokenize in NLTK. Next, The stop words are removed. Finally, the remaining tokens are annotated with their part-of-speech(POS) using Stanford parser.

### B. *Semantic-based Topic Identification*

A document is a combination of a small number of topics. The keyphrase extraction should be selected from every topic. Therefore, the topic identification contains two sub-processes i.e. candidate selection and candidate clustering. Step to select of candidate keyphrase has a function choice a word or multiwords from the derived tokens from the preprocessing step and assigned as a candidate. A token tagged as noun and adjective are considered. In this paper, a keyphrase contains 1-5 word. The 15 keyphrase patterns as show in table I. Candidate clustering attempts to categorize all candidate from the previous step. into small number of topic. In [7] all candidates are grouped based on their word stem similarities. Two candidates are matched by stemming the candidates and then syntactical matching the stems. The candidates having 25% overlapping stems are group in the same topic. We called this kind of clustering “form-based candidate clustering”. With the form based clustering some candidates are allocated in the wrong topics. Fig 2. Illustrates the example of incorrect clustering. In Fig2(a), for example: “studying” and “student” are grouped in the same topic because they have the same stem. The same reason is applied to “learner” and “learning”. Actually “student” and “learner” have the same meaning, they should be assigned to the same topic. In this paper, we use semantic similarity to measure “student” and “learner” The similarity score should be high and then the two words are put in the same topic. The detail of semantic similarity is explained

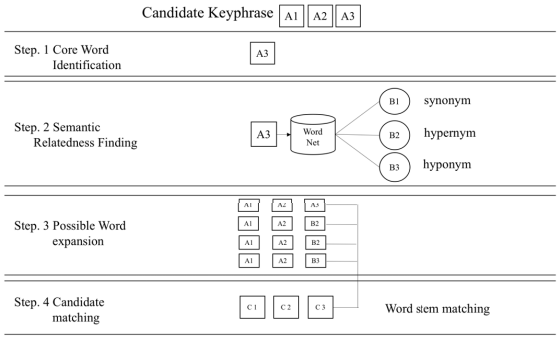


Fig. 3. Semantic Similarity in Candidate Clustering

TABLE I: Example Keyphrase Patterns

| Keyphrase Patterns | | |
|---|---|---|
| No. | Patterns | Keyphrase |
| 1. | NN | problem |
| 2. | NNS | systems |
| 3. | NNP | hyarm |
| 4. | NN-NNS | paper presents |
| 5. | NNP-NN | d47 organization |
| 6. | NN-NN-NN | resource availability second |
| 7. | NN-NNS-NNS | service qos requirements |
| 8. | NNP-NNS-JJ-NN | hyarm yields predictable stable |
| 9. | JJ | real-time |
| 10. | JJ-NN | end-to-end quality |
| 11. | JJ-NNS | operational conditions |
| 12. | JJ-NN-NN | high system performance |
| 13. | JJ-NNS-NNP | resource availability categories |
| 14. | JJ-JJ-NN-NN-NNS | effective adaptive resourcemanagement strategies |
| 15. | NN-JJ-NN-NN-NNP | hybrid adaptive resourcemanagement middleware hyarm |

in the next section.

### C. Topic Ranking

The topic derive from the previous step are builded as complete and undirected graph. Nodes represent topics and they are fully connected. Each link is weighted by the equation 1 which are reflected to the strength of their semantic relations.

$$W_{i,j} = \sum_{C_i \in t_i} \sum_{C_j \in t_j} dis(C_i, C_j) \tag{1}$$

$$dis(C_i, C_j) = \sum_{p_i \in pos(C_i)} \sum_{p_j \in pos(C_j)} \frac{1}{|p_i - p_j|} \tag{2}$$

$$S(t_i) = (1 - \lambda) \times \sum_{t_j \in V_i} \frac{W_{j,i} \times S(t_j)}{\sum_{t_k \in V_j} W_{j,k}} \tag{3}$$

where dist ($C_i$ ,$C_j$) refers to the reciprocal distances between the offset positions of the candidate keyphrase $C_i$ and $C_j$ in the document and where pos($C_i$) represents all the offset positions of the candidate keyphrase $C_i$ [7].Each link is distance by the equation 2 After the topic graph is constructed, all topics are ranked based on the contribution of the topics to their connected topic $T_i$ where $V_i$ are the topics voting for ti and $\lambda$ is a damping factor generally defined to 0.85. Each $S$ ($T_i$) is score by the equation 3 [8].

### D. Keyphrase Selection

This is the last step in graph-based keyphrase extraction. To confirm the extract keyphrase are covered all significant topics in the document. Topic in the document consider semantic word. We select keyphrase form score highest rating of the topic from top five. And we select a phrase ranking highest rating within topic to representative of documents.

## III. Semantic Similarity in Candidate Clustering

In semantic-based candidate clustering, two candidates keyphrase having similar meanings are assigned to the same cluster. In this paper, a candidate, its synonym, its hypernym and its hyponym are assumed that they have same meaning.

TABLE II: Compare Candidate KeyPhrase Clustering IEEE document

| Topic | Form-based | Semantic-based |
|---|---|---|
| Topic 1 | Information | Integral role, users javascript third-party library such |
| Topic 2 | data | Information, data |
| Topic 3 | Integral role | |
| Topic 4 | users | |
| Topic 5 | Javascript third-party library such | |

Figure 3. for example presents the Keyphrase Patterns for clustering candidates based on semantic similarity. From a list of candidates selected from basic structure extract keyphrase patterns process, each candidate keyphrase is picked up and then matched to the remaining candidates.

Algorithm:
1. c := 1
2. t := 1
3. while having candidates do
4. Add candidate(c) to topic(t)
5. remove candidate (c) from candidates
6. consider the next candidate
7. while not last candidate do
8. if topic(t) and candidate(c) have the same meaning then
9. add candidate(c) to topic(t)
10. remove candidate(c) from candidates
11. next topic

Assume that we have a candidate keyphrase containing three words as A1 A2 A3, respectively.
Step 1. Core word identification assigns the last word as core word i.e. A3.
Step 2. Semantic relatedness finding uses the WordNet as semantic resource to retrieve the semantic relatedness words i.e. synonym (B1), hypernym (B2) and hyponym (B3).
Step 3. Possible word expansion will expand the candidate with their semantic related words by replacing the core word with the semantic relatedness words. Additional relevant keyphrase are expanded.
Step 4. Candidate matching compares a candidate and the topic by using word stem matching.

This will later present examble a comparison the documentation between used in TopicRank and Synset. This documents are example from paper TopicRank.

## IV. Experiments and Results

In this section, we test the proposed method in two parts. The first part is to show the execution result for a random document to demonstrate how the method works. The second part is to compare the proposed method with the baseline.

### A. Demonstrative Results

A random document among the dataset was used to demonstrate running results. First, we compare a grouping result of the proposed method (semantic-based) and existing method

IEEE-98 Design and implementation of data visualization in media manuscripts transmission systemAbstract: Data visualizations have become an integral role for management system. It not only allows users to witness, explore and understand large amounts of information, but also convey information more directly expressed. This paper presents how to use JavaScript third-party library such as Flot and D3.js for collection, analysis, processing and visualization of data by using Ajax.

[('ieee-98'), ('design', 'NN'), ('and', 'CC'), (implementation', 'NN'), ('of', 'IN'), ('data', 'NNS'), ('visualization', 'NN'), ('in', 'IN'), ('media', 'NNS'), ('manuscripts', 'NNS'), ('transmission', 'NN'), ('system', 'NN'), ('abstract', 'NN'), (':', ':'), ('data', 'NNS'), ('visualizations', 'NNS'), ('have', 'VBP'), ('become', 'VBN'), ('an', 'DT'), ('integral', 'JJ'), ('role', 'NN'), ('for', 'IN'), ('management', 'NN'), ('system, 'NN'), ('it', 'PRP'), ('not', 'RB'), ('only', 'RB'), ('allows', 'VBZ'), ('users', 'NNS'), ('to', 'TO'), ('witness', 'NN'), ('explore', 'IN'), ('and', 'CC'), ('understand', 'NN'), ('large', 'JJ'), ('amounts', 'NNS'), ('of', 'IN'), ('information', 'NN'), ('but', 'CC'), ('also', 'RB'), ('convey', 'VBP'), (information', 'NN'), ('more', 'RBR'), ('directly', 'RB'), ('expressed', 'VBN'), ('this', 'DT'), ('paper', 'NN'), ('presents', 'NNS'), ('how', 'WRB'), ('to', 'TO'), ('use', 'VB'), ('javascript', 'NN'), ('third-party', 'NN'), ('library', 'NN'), ('such', 'JJ'), ('as', 'IN'), ('flot', 'NN'), ('and', 'CC'), ('d3js', 'NNP'), ('for', 'IN'), ('collection', 'NN'), ('analysis', 'NN'), ('processing', 'NN'), ('and', 'CC'), ('visualization', 'NN'), ('of', 'IN'), ('data', 'NNS'), ('by', 'IN'), ('using', 'VBG'), ('ajax', 'NN')]

Fig. 4. Document of TopicRank compare to Synset

TABLE III: Compare Clustering between TopicRank and Synset

| Topic | Form-Based | Semantic-Based |
|---|---|---|
| Topic 1 | ieee-98 design | design |
| Topic 2 | implementation | implementation |
| Topic 3 | visualization<br>data<br>data visualization<br>data visualizations | data visualization<br>data visualizations<br>visualization |
| Topic 4 | media manuscripts system abstract | media manuscripts system abstract |
| Topic 5 | users | role<br>users<br>javascript third party library |
| Topic 6 | management system | management system |
| Topic 7 | | witness |
| Topic 8 | | understand |
| Topic 9 | large amounts | amounts |
| Topic 10 | information | information<br>data |
| Topic 11 | paper | paper presents |
| Topic 12 | collection | collection analysis processing |
| Topic 13 | javascript third party library such | |
| Topic 14 | d3.js | |
| Topic 15 | processing | |
| Topic 16 | analysis | |
| Topic 17 | ajax | |
| Topic 18 | integral role | |
| Topic 19 | flot | |
| Topic 20 | witness | |

IEEE-98 :::
**Design** and **implementation** of **data visualization** in **media manuscripts transmission system**
**Abstract**: **Data visualizations** have become an **integral role** for **management system.** It not only allows **users** to **witness**, explore and **understand large amounts** of **information**, but also convey **information** more directly expressed. This **paper presents** how to use **JavaScript third-party library such** as Flot and **D3.js** for **collection, analysis, processing** and **visualization** of **data** by using **Ajax.**

Fig. 5. Candidate Extraction

TABLE IV: The Result Missing of Research

| Corpus | Document | | | | Keyphrase | | |
|---|---|---|---|---|---|---|---|
| | Type | Language | Number | Tokens average | Total | Average | Missing |
| SemEval | Abstracts | English | 144 | 85.01 | 1937 | 13.51% | 0.72% |
| IEEE | Abstracts | English | 380 | 72.87 | 5041 | 13.26% | 0.89 % |

(form-based) as shown in Table III. Moreover, the result of candidate extraction of the document using our method is exemplified in Fig. 5 while the generated graph from the example is shown in Fig. 6.

### B. Comparison Results

The datasets in this experiment are SemEval and IEEE. The first dataset from [10]. This contains 144 documents. [10] The second dataset is a collection of an abstract from IEEE publications. They are from 380 papers in a field of data visualization, machine translation, software engineering and data communication in 2014-2015. Hence, there are 522 documents in total.

First, we compare the result of our proposed method to the gold standard result from humans. It result of gold standartd missing keyphrases as 0.92% and 0.93% from SemEval and IEEE In comparison, we found the missing keyphrase as shown in Table IV. From the result, there were very few missing keyphrases as 0.72% and 0.89% from SemEval and IEEE, respectively.

### C. Evaluation

We evaluated the result using precision as a measure of the accuracy of the model [12]. The result is a combination of individual features [11]. An equation to calculate precision score is given in 4.

$$precision = \frac{correct}{output - length} \tag{4}$$

Recall is another evaluation of the accuracy of the model. Recall is a fraction of relevant item that are successfully retrieved. We applied 5 for calculating Recall.

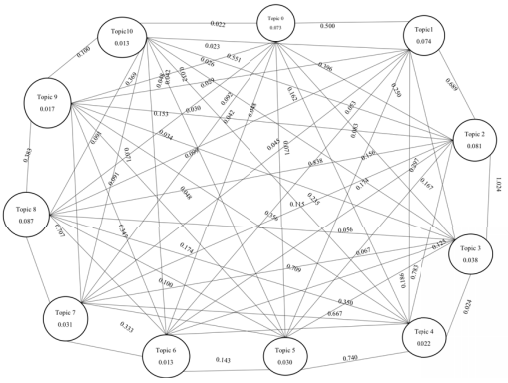


Fig. 6. Complete Graphs

TABLE V: Results are show as a percentage of precision, recall and f-Measure

| Methods | Corpus | Measuring the accuracy | | |
|---|---|---|---|---|
| | | Precisions | Recall | F-Measure |
| Form-Based Clustering | SemEval | 14.90 | 10.30 | 12.10 |
| Semantic-Based Clustering | SemEval | 27.95 | 16.93 | 20.51 |
| Form-Based Clustering | IEEE | 16.40 | 6.14 | 8.66 |
| Semantic-Based Clustering | IEEE | 19.31 | 11.81 | 13.89 |

$$recall = \frac{correct}{reference - length} \tag{5}$$

$$f - measure = \frac{precision * recall}{(precision + recall)/2} \tag{6}$$

Last, F-measure as given in 6 is selected to represent capability of the proposed method.

Second, we evaluate the result using precision, recall and f-measue by (4), (5) and (6), respectively. The comparison results are given in Table V. The results can be implied that our proposed method (semantic-based) received the better results in each measurement.

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

## V. Conclusion

In this paper, we propose a new method to improve graph-based keyphrase extraction by considering semantic in grouping keyphrase candidates. WordNet is chosen to help in detecting synonyms of words. With the synonymous keyphrase in the same group together, a calculation of keyphrase to be extracted and ranked as representatives can be boosted and affect to the better accuracy. With the experimental results, the proposed method obtained higher accuracy in terms of F-measure comparing to the baseline, TopicRank, in all test sets. Moreover, the number of missing keyphrase is lower than the baseline. In the future, we plan to handle another semantic problem i.e. homonym which can cause an error in detecting words based on surface forms. Moreover, we also plan to adjust a method to select for the representative keyphrase since using distance of word position may not represent the importance of words.

## References


[1] S. Beliga, A. Meštrović, and S. Martinčić-Ipšić, "An overview of graph-based keyword extraction methods and approaches," *Journal of information and organizational sciences*, vol. 39, no. 1, pp. 1–20, 2015.

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

[3] V. Gupta and G. S. Lehal, "A survey of text summarization extractive techniques," *Journal of emerging technologies in web intelligence*, vol. 2, no. 3, pp. 258–268, 2010.

[4] K. Sarkar, M. Nasipuri, and S. Ghose, "A new approach to keyphrase extraction using neural networks," *arXiv preprint arXiv:1004.3274*, 2010.

[5] P. D. Turney, "Learning algorithms for keyphrase extraction," *Information retrieval*, vol. 2, no. 4, pp. 303–336, 2000.

[6] R. Mihalcea and P. Tarau, "Textrank: Bringing order into text." in *EMNLP*, vol. 4, 2004, pp. 404–411.

[7] 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.

[8] S. Brin and L. Page., "The anatomy of a large-scale hypertextual web search engine," *Computer Networks and ISDN Systems*, vol. 30, no. 01, pp. 1–7, 1998.

[9] X. Wan and J. Xiao, "Single document keyphrase extraction using neighborhood knowledge." in *AAAI*, vol. 8, 2008, pp. 855–860.

[10] S. N. Kim, O. Medelyan, M.-Y. Kan, and T. Baldwin, "Semeval-2010 task 5: Automatic keyphrase extraction from scientific articles," in *Proceedings of the 5th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, 2010, pp. 21–26.

[11] E. C. Stephens, "Human evaluation on statistical machine translation," Ph.D. dissertation, San Diego State University, 2014.