273. Task 5. Keyphrase Extraction Based on Core Word Identification and Word Expansion

You Ouyang Wenjie Li Renxian Zhang
The Hong Kong Polytechnic University
{csyouyang,cswjli,csrzhang}@comp.polyu.edu.hk

Abstract

This paper provides a description of the Hong Kong Polytechnic University (PolyU) System that participated in the task #5 of SemEval-2, i.e., the *Automatic Keyphrase Extraction from Scientific Articles* task. We followed a novel framework to develop our keyphrase extraction system, motivated by differentiating the roles of the words in a keyphrase. We first identified the core words which are defined as the most essential words in the article, and then expanded the identified core words to the target keyphrases by a word expansion approach.

1 Introduction

The task #5 in SemEval-2 requires extracting the keyphrases for scientific articles. According to the task definition, keyphrases are the words that capture the main topic of the given document. Currently, keyphrase extraction is usually carried out by a two-stage process, including candidate phrase identification and key phrase selection. The first stage is to identify the candidate phrases that are potential keyphrases. Usually, it is implemented as a process that filters out the obviously unimportant phrases. After the candidate identification stage, the target keyphrases can then be selected from the candidates according to their importance scores, which are usually estimated by some features, such as word frequencies, phrase frequencies, POS-tags, etc.. The features can be combined either by heuristics or by learning models to obtain the final selection strategy.

In most existing keyphrase extraction methods, the importance of a phrase is estimated by a composite score of the features. Different features indicate preferences to phrases with specific characteristics. As to the common features, the phrases that consist of important and correlated words are usually preferred. Moreover, it is indeed implied in these features that the words are uniform in the phrase, that is, their degrees of importance are evaluated by the same criteria. However, we think that this may not always be true. For example, in the phrase "video encoding/decoding", the word "video" appears frequently in the article and thus can be easily identified by simple features, while the word "encoding/decoding" is very rare and thus is very hard to discover. Therefore, a uniform view on the words is not able to discover this kind of keyphrases. On the other hand, we observe that there is usually at least one word in a keyphrase which is very important to the article, such as the word "video" in the above example. In this paper, we call this kind of words *core words*. For each phrase, there may be one or more core words in it, which serve as the core component of the phrase. Moreover, the phrase may contain some words that support the core words, such as "encoding/decoding" in the above example. These words may be less important to the article, but they are highly correlated with the core word and are able to form an integrated concept with the core words. Motivated by this, we consider a new keyphrase extraction framework, which includes two stages: identifying the core words and expanding the core words to keyphrases. The methodology of the proposed approaches and the performance of the resulting system are introduced below. We also provide further discussions and modifications.

2 Methodology

According to our motivation, our extraction framework consists of three processes, including

(1) The pre-processing to obtain the necessary information for the following processes;
(2) The core word identification process to discover the core words to be expanded;
(3) The word expansion process to generate the final keyphrases.

In the pre-processing, we first identify the text fields for each scientific article, including its title, abstract and main text (defined as all the section titles and section contents). The texts are then processed by the language toolkit GATE [1] to carry out sentence segmentation, word stemming and POS (part-of-speech) tagging. Stop-words

[1] Publicly available at http://gate.ac.uk/gate

Proceedings of the 5th International Workshop on Semantic Evaluation, ACL 2010, pages 142–145,
Uppsala, Sweden, 15-16 July 2010.

are not considered to be parts of the target keyphrases.

2.1 Core Word Identification

Core words are the words that represent the dominant concepts in the article. To identify the core words, we consider the features below.

Frequencies: In a science article, the words with higher frequencies are usually more important. To differentiate the text fields, in our system we consider three frequency-based features, i.e., **Title-Frequency (TF)**, **Abstract-Frequency (AF)** and **MainText-Frequency (MF)**, to represent the frequencies of one word in different text fields. For a word w in an article t, the frequencies are denoted by

$TF(w)$ = Frequency of w in the title of t;
$AF(w)$ = Frequency of w in the abstract of t;
$MF(w)$ = Frequency of w in the main text of t.

POS tag: The part-of-speech tag of a word is a good indicator of core words. Here we adopt a simple constraint, i.e., only nouns or adjectives can be potential core words.

In our system, we use a progressive algorithm to identify all the core words. The effects of different text fields are considered to improve the accuracy of the identification result. First of all, for each word w in the title, it is identified to be a core word when satisfying

$\{ TF(w) > 0 \ \wedge AF(w) > 0 \}$

Since the abstract is usually less indicative than the title, we use stricter conditions for the words in the abstract by considering their co-occurrence with the already-identified core words in the title. For a word w in the abstract, a co-occurrence-based feature $CO_T(w)$ is defined as $|S(w)|$, where $S(w)$ is the set of sentences which contain both w and at least one title core word. For a word w in the abstract, it is identified as an abstract core word when satisfying

$\{ AF(w) > 0 \ \wedge MF(w) > \alpha_1 \wedge CO_T(w) > \alpha_2 \}$

Similarly, for a word w in the main text, it is identified as a general core word when satisfying

$\{ MF(w) > \beta_1 \wedge CO_{TA}(w) > \beta_2 \}$

where $CO_{TA}(w) = |S'(w)|$ and $S'(w)$ is the set of sentences which contain both w and at least one identified title core word or abstract core word.

With this progressive algorithm, new core words can be more accurately identified with the previously identified core words. In the above heuristics, the parameters α and β are pre-defined thresholds, which are manually assigned[2].

As a matter of fact, this heuristic-based identification approach is simple and preliminary. More sophisticated approaches, such as training machine learning models to classify the words, can be applied for better performance. Moreover, more useful features can also be considered. Nevertheless, we adopted the heuristic-based implementation to test the applicability of the framework as an initial study.

An example of the identified core words is illustrated in Table 1 below:

Type	Core Word
Title	grid, service, discovery, UDDI
Abstract	distributed, multiple, web, computing, registry, deployment, scalability, DHT, DUDE, architecture
Main	proxy, search, node, key, etc.

Table 1: Different types of core words

2.2 Core Word Expansion

Given the identified core words, the keyphrases can then be generated by expanding the core words. An example of the expansion process is illustrated below as

grid ➔ grid service ➔ grid service discovery ➔ scalable grid service discovery

For a core word, each appearance of it can be viewed as a potential expanding point. For each expanding point of the word, we need to judge if the context words can form a keyphrase along with it. Formally, for a candidate word w and the current phrase e (here we assume that w is the previous word, the case for the next word is similar), we consider the following features to judge if e should be expanded to $w+e$.

Frequencies: the frequency of w (denoted by $Freq(w)$) and the frequency of the combination of w and e (denoted by $phraseFreq(w, e)$) which reflects the degree of w and e forming an integrated phrase.

POS pattern: The part-of-speech tag of the word w is also considered here, i.e., we only try to expand w to $w+e$ when w is a noun, an adjective or the specific conjunction "of".

A heuristic-based approach is adopted here again. We intend to define some loose heuristics, which prefer long keyphrases. The heuristics include (1) If w and e are in the title or abstract, expand e to $e+w$ when w satisfies the POS constraint and $Freq(w) > 1$; (2) If w and e are in the main text, expand e to $e+w$ when w satisfies the POS constraint and $phraseFreq(w, e) > 1$.

More examples are provided in Table 2 below.

[2] $(\alpha_1, \alpha_2, \beta_1, \beta_2) = (10, 5, 20, 10)$ in the system

Core Word	Expanded Key Phrase
grid	scalable grid service discovery, grid computing
UDDI	UDDI registry, UDDI key
web	web service,
scalability	Scalability issue
DHT	DHT node

Table 2: Core words and corresponding key phrases

3 Results

3.1 The Initial PolyU System in SemEval-2

In the Semeval-2 test set, a total of 100 articles are provided. Systems are required to generate 15 keyphrases for each article. Also, 15 keyphrases are generated by human readers as standard answers. Precision, recall and F-value are used to evaluate the performance.

To generate exactly 15 keyphrases with the framework, we expand the core words in the title, abstract and main text in turn. Moreover, the core words in one fixed field are expanded following the descending order of frequency. When 15 keyphrases are obtained, the process is stopped.

For each new phrase, a redundancy check is also conducted to make sure that the final 15 keyphrases can best cover the core concepts of the article, i.e.,

(1) the new keyphrase should contain at least one word that is not included in any of the selected keyphrases;

(2) if a selected keyphrase is totally covered by the new keyphrase, the covered keyphrase will be substituted by the new keyphrase.

The resulting system based on the above method is the one we submitted to SemEval-2.

3.2 Phrase Filtering and Ranking

Initially, we intend to use just the proposed framework to develop our system, i.e., using the expanded phrases as the keyphrases. However, we find out later that it must be adjusted to suit the requirement of the SemEval-2 task. In our subsequent study, we consider two adjustments, i.e., phrase filtering and phrase ranking.

In SemEval-2, the evaluation criteria require exact match between the phrases. A phrase that covers a reference keyphrase but is not equal to it will not be counted as a successful match. For example, the candidate phrase “scalable grid service discovery” is not counted as a match when compared to the reference keyphrase “grid service discovery”. We call this the “partial matching problem”. In our original framework, we followed the idea of “expanding the phrase as much as possible” and adopted loose conditions. Consequently, the partial matching problem is indeed very serious. This unavoidably affects its performance under the criteria in SemEval-2 that requires exact matches. Therefore, we consider a simple filtering strategy here, i.e., filtering any keyphrase which only appears once in the article.

Another issue is that the given task requires a total of exactly 15 keyphrases. Naturally we need a selection process to handle this. As to our framework, a keyphrase ranking process is necessary for discovering the best 15 keyphrases, not the best 15 core words. For this reason, we also try a simple method that re-ranks the expanded phrases by their frequencies. The top 15 phrases are then selected finally.

3.3 Results

Table 3 below shows the precision, recall and F-value of our submitted system (**PolyU**), the best and worst systems submitted to SemEval-2 and the baseline system that uses simple TF-IDF statistics to select keyphrases.

On the SemEval-2 test data, the performance of the **PolyU** system was not good, just a little better than the baseline. A reason is that we just developed the PolyU system with our past experiences but did not adjust it much for better performance (since we were focusing on designing the new framework). After the competition, we examined two refined systems with the methods introduced in section 3.2.

First, the PolyU system is adapted with the phrase filtering method. The performance of the resulting system (denoted by **PolyU+**) is given in Table 4. As shown in Table 4, the performance is much better just with this simple refinement to meet the requirement on extract matches for the evaluation criteria. Then, the phrase ranking method is also incorporated into the system. The performance of the resulting system (denoted by **PolyU++**) is also provided in Table 4. The performance is again much improved with the phrase ranking process.

3.4 Discussion

In our participation in SemEval-2, we submitted the PolyU system with the proposed extraction framework, which is based on expanding the core words to keyphrases. However, the PolyU system did not perform well in SemEval-2. However, we also showed later that the framework can be much improved after some

Simple but necessary refinements are made according to the given task. The final PolyU++ system with two simple refinements is much better. These refinements, including phrase filtering and ranking, are similar to traditional techniques. So it seems that our expansion-based framework is more applicable along with some traditional techniques. Though this conflicts our initial objective to develop a totally novel framework, the framework shows its ability of finding those keyphrases which contain different types of words. As to the PolyU++ system, when adapted with just two very simple post-processing methods, the extracted candidate phrases can already perform quite well in SemEval-2. This may suggest that the framework can be considered as a new way for candidate keyphrase identification for the traditional extraction process.

4 Conclusion and future work

In this paper, we introduced our system in our participation in SemEval-2. We proposed a new framework for the keyphrase extraction task, which is based on expanding core words to keyphrases. Heuristic approaches are developed to implement the framework. We also analyzed the errors of the system in SemEval-2 and conducted some refinements. Finally, we concluded that the framework is indeed appropriate as a candidate phrase identification method. Another issue is that we just consider some simple information such as frequency or POS tag in this initial study. This indeed limits the power of the resulting systems. In future work, we'd like to develop more sophisticated implementations to testify the effectiveness of the framework. More syntactic and semantic features should be considered. Also, learning models can be applied to improve both the core word identification approach and the word expansion approach.

Acknowledgments

The work described in this paper is supported by Hong Kong RGC Projects (PolyU5217/07E and PolyU5230/08E).

References

Frank, E., Paynter, G.W., Witten, I., Gutwin, C. and Nevill-Manning, C.G.. 1999. Domain Specific Keyphrase Extraction. Proceedings of the IJCAI 1999, pp.668--673.

Medelyan, O. and Witten, I. H.. 2006. Thesaurus based automatic keyphrase indexing. Proceedings of the JCDL 2006, Chapel Hill, NC, USA.

Medelyan, O. and Witten, I. H.. 2008. Domain independent automatic keyphrase indexing with small training sets. Journal of American Society for Information Science and Technology. Vol. 59 (7), pp. 1026-1040

SemEval-2. Evaluation Exercises on Semantic Evaluation. http://semeval2.fbk.eu/

Turney, P.. 1999. Learning to Extract Keyphrases from Text. National Research Council, Institute for Information Technology, Technical Report ERB-1057. (NRC \#41622), 1999.

Wan, X. Xiao, J.. 2008. Single document keyphrase extraction using neighborhood knowledge. In Proceedings of AAAI 2008, pp 885-860.

System	5 Keyphrases			10 Keyphrases			15 Keyphrases		
	P	R	F	P	R	F	P	R	F
Best	34.6%	14.4%	20.3%	26.1%	21.7%	23.7%	21.5%	26.7%	23.8%
Worst	8.2%	3.4%	4.8%	5.3%	4.4%	4.8%	4.7%	5.8%	5.2%
PolyU	13.6%	5.65%	7.98%	12.6%	10.5%	11.4%	12.0%	15.0%	13.3%
Baseline	17.8%	7.4%	10.4%	13.9%	11.5%	12.6%	11.6%	14.5%	12.9%

Table 3: Results from SemEval-2

System	5 Keyphrases			10 Keyphrases			15 Keyphrases		
	P	R	F	P	R	F	P	R	F
PolyU	13.6%	5.65%	7.98%	12.6%	10.5%	11.4%	12.0%	15.0%	13.3%
PolyU+	21.2%	8.8%	12.4%	16.9%	14.0%	15.3%	13.9%	17.3%	15.4%
PolyU++	31.2%	13.0%	18.3%	22.1%	18.4%	20.1%	20.3%	20.6%	20.5%

Table 4: The performance of the refined systems