An Ontology–Based Approach for Key Phrase Extraction

Chau Q. Nguyen
HCM University of Industry
12 Nguyen Van Bao St, Go Vap Dist,
HCMC, Vietnam
chauqn@hui.edu.vn

Tuoi T. Phan
HCMC University of Technology
268 Ly Thuong Kiet St, Dist 10,
HCMC, Vietnam
tuoi@cse.hcmut.edu.vn

Abstract

Automatic key phrase extraction is fundamental to the success of many recent digital library applications and semantic information retrieval techniques and a difficult and essential problem in Vietnamese natural language processing (NLP). In this work, we propose a novel method for key phrase extracting of Vietnamese text that exploits the Vietnamese Wikipedia as an ontology and exploits specific characteristics of the Vietnamese language for the key phrase selection stage. We also explore NLP techniques that we propose for the analysis of Vietnamese texts, focusing on the advanced candidate phrases recognition phase as well as part-of-speech (POS) tagging. Finally, we review the results of several experiments that have examined the impacts of strategies chosen for Vietnamese key phrase extracting.

1 Introduction

Key phrases, which can be single keywords or multiword key terms, are linguistic descriptors of documents. They are often sufficiently informative to allow human readers get a feel for the essential topics and main content included in the source documents. Key phrases have also been used as features in many text-related applications such as text clustering, document similarity analysis, and document summarization. Manually extracting key phrases from a number of documents is quite expensive. Automatic key phrase extraction is a maturing technology that can serve as an efficient and practical alternative. In this paper, we present an ontology-based approach to building a Vietnamese key phrase extraction system for Vietnamese text. The rest of the paper is organized as follows: Section 2 states the problem as well as describes its scope, Section 3 introduces resources of information in Wikipedia that are essential for our method, Section 4 describes extraction of titles and its categories from Wikipedia to build a dictionary, Section 5 proposes a methodology for the Vietnamese key phrase extraction model, Section 6 evaluates our approach on many Vietnamese query sentences with different styles of texts, and finally the conclusion is presented in Section 7.

2 Background

The objective of our research is to build a system that can extract key phrases in Vietnamese queries in order to meet the demands associated with information searching and information retrieving, especially to support search engines and automatic answer systems on the Internet. For this purpose, we provide the following definition:

Key phrases in a sentence are phrases that express meaning completely and also express the purpose of the sentence to which they are assigned.

For an example, we have a query sentence as follows:*"Laptop Dell E1405 có giá bao nhiêu?"*. That means "*How much does a Dell E1405 laptop cost?* ".

Key phrases are "*Laptop Dell E1405*", "*giá*", and "*bao nhiêu*". In this case, the interrogative word "*bao nhiêu*" is used to add a meaning for the two rest noun phrases, making the query of users clear, wanting to know the numeral aspect about the "*price*" of a "*Laptop Dell E1405*".

3 Wikipedia

Wikipedia is a multilingual, web-based, freely available encyclopedia, constructed as a collaborative effort of voluntary contributors on the web. Wikipedia grows rapidly, and with approximately 7.5 million articles in more than 253 languages, it has arguably become the world's largest collection of freely available knowledge.

Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, pages 181–184,
Suntec, Singapore, 4 August 2009. ©2009 ACL and AFNLP

Wikipedia contains a rich body of lexical semantic information, the aspects of which are comprehensively described in (Zesch et al., 2007). Additionally, the redirect system of Wikipedia articles can be used as a dictionary for synonyms, spelling variations and abbreviations.

A PAGE. A basic entry in Wikipedia is a page that represents either a normal Wikipedia article, a redirect to an article, or a disambiguation page. Each page object provides access to the article text (with markup information or as plain text), the assigned categories, the ingoing and outgoing article links as well as all redirects that link to the article.

A LINK. Each page consists of many links which function not only to point from the page to others, but also to guide readers to pages that provide additional information about the entries mentioned. Each link is associated with an anchor text that denotes an ambiguous name or is an alternative name, instead of a canonical name.

CATEGORY. Category objects represent Wikipedia categories and allow access to the articles within each category. As categories in Wikipedia form a thesaurus, a category object also provides means to retrieve parent and child categories as well as siblings and all recursively collected descendants.

REDIRECT PAGE. A redirect page typically contains only a reference to an entry or a concept page. The title of the redirect page is an alternative name for that entity or concept.

DISAMBIGUATION PAGE. A disambiguation page is created for an ambiguous name that denotes two or more entities in Wikipedia. It consists of links to pages that define different entities with the same name.

4 Building a dictionary

Based on the aforementioned resources of information, we follow the method presented in (Bunescu and Pasca, 2006) to build a dictionary called ViDic. Since our research focuses on Key phrases, we first consider which pages in Wikipedia define concepts or objects to which key phrases refer. The key phrases are extracted from the title of the page. We consider a page has key phrases if it satisfies one of the following steps:

1. If its title is a word or a phrase then the title is key phrase.
2. If its title is a sentence then we follow the method presented in (Chau and Tuoi, 2007) to extract key phrases of the sentence.

Following this method, the ViDic is constructed so that the set of entries in the ViDic consists of all strings that denote a concept. In particular, if c is a concept, its key phrases, its title name, its redirect name and its category are all added as entries in the ViDic. Then each entry string in the ViDic is mapped to a set of entries that the string may denote in Wikipedia. As a result, a concept c is included in the set if, and only if, the string has key phrases which is extracted from the title name, redirect name, or disambiguation name of c.

Although we utilize information from Wikipedia to build the ViDic, our method can be adapted for an ontology or knowledge base in general.

5 Proposed method

We consider the employment of a set of NLP techniques adequate for dealing with the Vietnamese key phrase extraction problem. We propose the following general Vietnamese key phrase extraction model (see Figure 1).

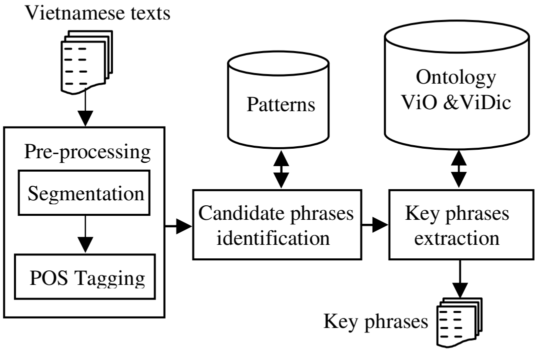

Figure 1. The general Vietnamese key phrase extraction model.

5.1 Pre-processing

The input of pre-processing is user’s queries and the output is a list of words and their POS labels. Because of the effectiveness and convenience associated with integrating two stages of word segmentation and POS tagging, we proposed two modules for the pre-processing stage. The purposes of two modules are as follows:

- **Word Segmentation**: The main function of this segmentation module is to identify and separate the tokens present in the text in such a way that every individual word, as well as every punctuation mark, will be a different token. The segmentation module considers words, numbers with decimals or dates in numerical format in order not to separate the dot, the comma or the slash (respectively) from the preceding and/or following elements.

- **POS tagging**: The output of the segmentation module is taken as input by the POS tagging module. Almost any kind of POS tagging could be applied. In our system, we have proposed a hybrid model for the problem of Vietnamese POS Tagging (Chau and Tuoi, 2006). This model combines a rule-based method and a statistical learning method. With regard to data, we use a lexicon with information about possible POS tags for each word, a manually labeled corpus, syntax and context of texts.

5.2 Candidate phrases identification

The input of the candidate phrase identification is a list of words and their POS labels, and the output is a list of words and their chunking labels. The idea underlying this method (Chau and Tuoi, 2007) for the Vietnamese key phrase extraction is based on a number of grammatical constructions in Vietnamese. The method consists of pattern-action rules executed by the finite-state transduction mechanism. It recognizes entities such as noun phrases. In order to accomplish the noun phrases recognition, we have developed over 434 patterns of noun phrase groups that cover proper noun constructs.

5.3 Key phrases extraction

In this section, we focus on the description of a methodology for key phrase extraction. This method combines a pattern-based method and a statistical learning method. Both methods will complement each other to increase the expected performance of the model. In particular, the method has the following steps:

- Step 1: We propose a method that exploits specific characteristics of Vietnamese (Chau and Tuoi, 2007). At the heart of this method is the idea of building a Vietnamese words set that reflects semantic relationships among objects. For example, consider the sentence that follows:

“Máy tính này có dung lượng RAM lớn nhất là bao nhiêu ?” that means “*What is the largest RAM capacity for this computer?”*

In this sentence, we have two objects “*Máy tính”(this computer)* and “*RAM”* in real world. Respectively, two noun phrases are *“Máy tính”(this computer)* and “*dung lượng RAM lớn nhất” (the largest RAM capacity).* We consider the meanings of words per the above example; we will recognize “*cố*”, a meaning word in our meaning word set, which reflects a possessive relationship between *“Máy tính”* and “*dung lượng RAM lớn nhất”*. This has identified “*dung lượng RAM lớn nhất”* representing the meaning of the sentence.

This meaning word-based approach provides a set of semantic relationships (meaning words) between phrases to support key phrase extraction, which does not require building a hierarchy or semantic network of objects in the Vietnamese language.

- Step 2: In case the sentence has no meaning word among phrases, the key phrase extracting process is based on the ViO ontology via concept matching. In particular, this step has the following phases:

1. every candidate phrase in the sentence is matched to an entry in the VicDic dictionary especially when new phrases are not a concern or do not exist in the dictionary. Because a partial matching dilemma usually exists, we apply several strategies to improve the matching process, including maximum matching, minimum-matching, forward-matching, backward-matching and bi-directional matching.
2. if the matching process is successful, then we retrieve categories for the entries respectively via the category system in the ViO ontology; if the candidate phrase has the most specific category, then the phrase is the key phrase of the sentence indicated in Step 3.
3. if the matching process is not successful, then we find a semantic similarity concept in the ViO ontology as Step 4. After that, the key phrase extracting process will go to phase 2.

- Step 3: The idea of the most specific category identification process based on the ViO ontology is shown as pseudo-code, such as

Algorithm: the most specific category identification

```
  - Input: $C_1$, $C_2$ categories, and  the ViO Ontology
  - Output: $C_1$ or $C_2$ or both $C_1$ and $C_2$
begin
if $C_1$& $C_2$  have a synonyms relationship in ViO
then  $C_1$ & $C_2$ are the most specific categories
else if $C_1$ has isa relationship of $C_2$ then $C_1$ is the
    most specific category.
to traverse the ViO ontology  from $C_1$ & $C_2$  to
    find the nearest common ancestor node (C’).
    Calculate  the distance between $C_1$ and C’ ($h_1$),
    distance $C_2$  and C’ ($h_2$).
if  $h_1 > h_2$ then $C_1$ is the most specific category
else if  $h_1 < h_2$ then $C_2$ is the most specific
```

```
    category
else $C_1$ & $C_2$ are the most specific categories
end;
```

- Step 4: To find the semantic similarity concept for each concept *t* that is still unknown after phase 2, we traverse the ontology hierarchy from its root to find the best node. We choose the semantic similarity that was described as in (Banerjee and Pederson, 2003). However, we do not use the whole formula. In particular , we use a similar formula that is specified as follows:

$$Acu\_Sim(w, c) = Sim(w, c) + \sum Sim(w, c')$$

in which, w is the phrase that needs to be annotated, c is the candidate concept and c' is the concept that is related to c.

At the current node c while traversing, the similarity values between t and all children of c are calculated. If the maximum of similarity values is less than similarity value between t and c, then c is the best node corresponding to t. Otherwise, continue the procedure with the current node as the child node with the maximum similarity value. The procedure stops when the best node is found or it reaches a leaf node.

6 Evaluation

To evaluate the result of the proposed model, we use **recall** and **precision** measures that are defined as in (Chau & Tuoi, 2007). In order to test the model we selected a questions set from sources on the web as follows:

- TREC (Text REtrieval Conference) (http://trec.nist.gov/data/): TREC-07 (consisting of 446 questions); TREC-06 (consisting of 492 questions); and TREC-02 (consisting of 440 questions).
- The web page www.lexxe.com: consisting of 701 questions.

After that, the question set (consisting of 2079 questions) is translated into a Vietnamese questions set, we called D_1 dataset. All key phrases of the D_1 dataset are manually extracted by two linguists for the quality of the dataset. Then we have two versions respectively, V_1 and V_2. The results of our system is shown as follows:

Ver	R	A	Ra	Precision	Recall
V_1	3236	3072	2293	74.6%	70.8%
V_2	3236	3301	2899	**89.6%**	**87.8%**

Table 1. Results of Vietnamese key phrase extraction.

7 Conclusion

We have proposed an original approach to key phrase extraction. It is a hybrid and incremental process for information searching for search engines and automatic answer systems in Vietnamese. We achieved precision of around 89.6% for our system. The experimental results have show that our method achieves high accuracy.

Currently, Wikipedia editions are available for approximately 253 languages, which means that our method can be used to build key phrase systems for a large number of languages. In spite of the exploitation of Wikipedia as a Vietnamese ontology, our method can be adapted for any ontology and knowledge base in general.

Furthermore, we had to construct all necessary linguistic resources and define all data structures from scratch, while enjoying some advantages derived from the many existent methodologies for morpho-syntactic annotation and the high consciousness of a standardization tendency. Specifically, we built a set with 434 noun phrase patterns and a rules set for Vietnamese key phrase identification. Our patterns and rules set can be easily readjusted and extended. The results obtained lay the foundation for further research in NLP for Vietnamese including text summarization, information retrieval, information extraction, etc.

References

Bunescu, R., Pasca, M. 2006. Using encyclopedic knowledge for name entity disambiguation. In *Proceedings of the 11th Conference of EACL*:9-16.

Banerjee S.,Pederson T., 2003. Extended Gloss Overlaps as a Measure of Semantic Relatedness, In *Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI-03)*: 805–810.

Chau Q.Nguyen, Tuoi T.Phan. 2007. A Pattern-based Approach to Vietnamese Key Phrase Extraction, In *Addendum Contributions of the 5th International IEEE Conference on Computer Sciences- RIVF'07*: 41-46.

Chau Q.Nguyen, Tuoi T.Phan. 2006. A Hybrid Approach to Vietnamese Part-Of-Speech Tagging. In *Proceedings of the 9th International Oriental COCOSDA Conference (O-COCOSDA'06)*, Malaysia:157-160.

Zesch, T., Gurevych, I. 2007. Analysis of the Wikipedia Category Graph for NLP Applications. In *Proceedings of the TextGraphs-2 Workshop (NAACL-HLT 2007)*:1–8.