INFORMATION OF DOCTORAL DISSERTATION

The Doctoral Dissertation:

A VIETNAMESE KEY PHRASE INFORMATION EXTRACTION MODEL

- Field: COMPUTER SCIENCE
- Field code: 62.48.01.01
- Ph.D Candidate name: CHAU QUANG NGUYEN
- Supervisors: ASS. PROF. DR. TUOI THI PHAN
- Institution: University of Technology, Vietnam National University of Ho Chi Minh City

1. Dissertation Abstract

AUTOMATIC KEY PHRASE EXTRACTION is the task of automatically selecting a set of phrases describing the content of a simple sentence. That a key phrase is extracted means that it is present verbatim in the sentence to which it is assigned. Furthermore, accurate key phrase extraction is fundamental to the success of many recent digital library applications and a difficult and essential problem in Vietnamese natural language processing (NLP) including text summarization, semantic information retrieval and information extraction. In this dissertation, we propose a novel method for key phrase extracting of Vietnamese text that combines assignment and extraction approaches. To achieve about goal, the dissertation solves three main problems as follows:

Problem 1 - Constructing a Vietnamese Key phrase Extraction model for Assignment approach (ViKEa). The dissertation not only presents the proposed model based on the language knowledge (as rule based and Ontology) and pre-processing problems (as word segmentation and part-of-speech tagging), The dissertation proposes also a method that exploits specific characteristics of the Vietnamese language and exploits the Vietnamese Wikipedia as a Vietnamese ontology for key phrase ambiguity resolution.

Problem 2 - Constructing a Vietnamese Key phrase Extraction model for Extraction approach (ViKEe). Exploring machine learning models so that the dissertation proposes the ViKEe model using a support vector machines (SVMs) approach for Vietnamese key phrase extraction.

Problem 3 - Constructing a hybrid model ViKE (Vietnamese Key phrase Extraction) that combines the ViKEa and ViKEe models for a maximum accurate exploiting of Vietnamese key phrase extracting problem for each model in problem 1 and problem 2.

2. Contributions of the dissertation

2.1 Scientific contributions

The main contributions of the research presented in this dissertation are summarized as follows:
The first contribution is the dissertation proposed a Vietnamese key phrase assignment model (ViKEa) of Vietnamese sentence which is based on a pattern-matched method and exploiting the Vietnamese Wikipedia as an ontology.

The second contribution is the dissertation proposed a Vietnamese key phrase extraction model (ViKEe) of Vietnamese sentence using a support vector machines (SVMs) approach with proposed features as follows: (1) the location of word in the sentence; (2) a part-of-speech (POS) of word; (3) structure of phrase; (4) relation word among phrases.

The third contribution is the dissertation proposed a novel model for key phrase extracting of Vietnamese sentence that combines the ViKEa and ViKEe models.

The fourth contribution is the proposal of a method for key noun phrase extracting of Vietnamese sentence. Then, the dissertation also provided a completed architecture for Vietnamese key phrase extracting application.

The fifth contribution is a proposed method that exploits the Vietnamese Wikipedia as an ontology not only for key noun phrase extraction but also recommendation of a solution for a gap in Vietnamese corpus of many natural language processing research works.

2.2 Realistic contributions

The results obtained of the dissertation lay the foundation for further research in NLP for Vietnamese including text summarization, information retrieval and information extraction. They also support to develop applications in practice.

3. Recommendations of the future works

- To exploit all semantic relationships of categories via language Ontology (as Wiktionary, WordNet,...) for Vietnamese key phrase extracting in the key phrase assignment approach.

- To exploit the example training stage of the machine learning model in specific domains, and also update the training corpus to improve the precision of the system in the key phrase extraction approach.

- To develop and apply proposed models of the dissertation for other languages.

Supervisor Ph.D Candidate

Ass. Prof. Dr. Tuoi Thi Phan Chau Quang Nguyen