DISSERTATION INFOMATION

Title:

VIETNAMESE SEMANTIC DOCUMENT RETRIEVAL MODEL

Major:	Computer Science
Major code:	62.48.01.01
PhD student:	Tuyen Thi-Thanh Do
Advisor:	Assoc. Prof. Dang Tuan Nguyen
	Assoc. Prof. Lung Duc Vu
University:	University of Information Technology – VNU-HCM

1. ABSTRACT

Semantic document retrieval is one of the important research fields aimed at exploiting text documents in many very large collections of text document. The problem of semantic document retrieval can be solved in approaches, such as, latent semantic analysis, topic model, distributional semantic with word embedding approximated from deep neural network and sentence semantic parsing. In this dissertation, the semantic document retrieval is aimed at matching document and a query with their semantic represenations at semantic level and not at word and lexicon semantic level according to computational linguistics approach.

In order to complete this objective, the approach of this dissertation is to propose a formal language representing phrasal and sentential semantic. This formal language is defined upon the Universal Dependency representation and its sentences can be generated from semantic dependency graphs. The Vietnamese Lexicon Ontology (VLO) storing word senses and semantic constraints is also proposed and built for sense tagging and checking the dependency between two word senses. Then, a statistic based sense tagging method and a rule-based sentential semantic parsing method are studied to generate the semantic dependency graph of a phrase or a sentence. Finally, a new document retrieval model using two indices, semantic class index (SCI) and semantic relation index (SRI), and Jaccard-Tanimoto based distance measure is proposed to work with the semantic representation of document and query. The experimental results in document retrieval of this approach show that the proposed semantic document retrieval model can be better than the

probabilistic model using BM25 ranking function in both cases of using and not using query expansion technique, the vector space model using TF.IDF ranking function, the vector space model using word embedding and topic model using Latent Dirichlet Allocation (LDA).

All research results of the dissertation are published in international conference proceedings and journals indexed by reputable publisher such as ACM Digital Library, IEEE, DBPL, etc.

2. MAIN CONTRIBUTIONS

The main contributions of the dissertation are:

1) Proposing a formal language representing the semantic of a phrase or a sentence in Universal Dependency representation.
2) A rule-based sentential semantic parsing method for transforming a dependency parse from a parsing model, which use Bidirectional LSMT feature extraction, to a semantic dependency graph.
3) VLO providing semantic labels and semantic constraints which are element units in semantic dependency tranformation.
4) Proposing a new Vietnamese semantic document retrieval model with two indices, Semantic Class Index (SCI) and Semantic Relation Index (SRI), and Jaccard-Tanimoto based ranking function which is better than BM25 ranking function in the experiments of the dissertation.

3. FUTURE WORKS

The Vietnamese semantic document retrieval model has good ranking results and the sentential semantic parsing method can be use in practise. However, there are some future works to improve the results:

- Enriching VLO with word senses and semantic constraints and developing Vietnamese dependency graphbank in order to develop a better semantic dependency parsing method.
- Research in automatically building VLO for reducing VLO building cost.
- Research in training a semantic dependency parsing model on Vietnamese graphbank.

- Research in applying results of Deep Neural Network to improve the components of the proposed Vietnamese semantic document retrieval model.

ADVISOR 1	ADVISOR 2	PHD STUDENT
Assoc. Prof. Dang Tuan Nguyen	Assoc. Prof. Lung Duc Vu	Tuyen Thi-Thanh Do