VIETNAM NATIONAL UNIVERSITY, HANOI
VNU UNIVERSITY OF ENGINEERING AND TECHNOLOGY

SOCIALIST REPUBLIC OF VIETNAM
Independence – Freedom – Happiness

INFORMATION ON DOCTORAL THESIS

1. Full name: Bui Van Tan 2. Sex: Male

3. Date of birth: 12/08/1983 4. Place of birth: Nam Dinh

5. Admission decision number: 654/QĐ-CTSV Dated 05-09-2016

6. Changes in academic process:

The decision to extend the study period is reflected in Decision No. 1127/QD-DT dated October 17, 2019, by the Rector of the VNU University of Engineering and Technology.

7. Official thesis title: *Determining semantic relations based on statistical machine learning.*

8. Major: Computer science 9. Code: Computer science

10. Supervisors: Associate Professor Nguyen Phuong Thai

11. Summary of the **new findings** of the thesis:

This thesis aims to improve the performance of automatic models that identify four semantic relations of words including hypernymy, synonymy, antonym, and semantic similarity. The main results of this thesis are as follows:

Firstly, this thesis has proposed an improvement to the Dynamic Weighting Neural Network (DWN) model proposed by Tuan Luu et al. The improved model called EDWN that is capable of learning specialized word embedding vectors, these embedding vectors are "specialized" with semantic features, thereby being suitable for the problem of determining lexical entailment relation.

Secondly, this thesis has identified subword semantic features and proposed a scheme to extract these features. This thesis proposed the LERC model, which used the input feature combined from the word embedding vector and the semantic feature vector under the word. The experimental results evaluated on a number of standard datasets of both English and Vietnamese have proved that the proposed model in this thesis has significantly higher performance than the best models at the same time.

Thirdly, this thesis has proposed the neural network model DVASNet. This model not only uses the distributional features of words in the corpus but also exploits information about word structure. Experimental results on a number of standard datasets have demonstrated that the DVASNet model has significantly higher performance than the base five models.

Fourthly, this thesis proposes the GraphSim model to improve the performance of measuring the semantic similarity of English word pairs based on the algorithm to find the shortest path on the graph.

Fifthly, this thesis proposed the ExtLeskSim model, which is an improvement of the Lesk algorithm to work more effectively with Vietnamese characteristics.

In addition, this thesis has built four datasets to evaluate semantic relation recognition models, including VLE-999, ViAS-1000, VSimLex-999, and VESim-1000.

12. Practical applicability, if any:

The models for determining semantic relations of words proposed in this thesis can be applied to a number of natural language processing problems, including machine translation, emotion analysis from text ([CT4]), measuring the semantic similarity of sentences ([CT3]), align sentences and build bilingual corpora ([CT11]), thereby improving the performance of models to solve these problems.

13. Further research directions, if any:

Several recent studies are interested in the problem of ranking the membership relationship (Graded Hypernymy). In the next studies, we will exploit the EDWN word embedding model and sub-word semantic features for the membership relation ranking problem. In addition, we are also interested in improving the GraphSim schema in the direction of using the Floyd-Warshall algorithm to find the shortest path on the fuzzy weighted graph. Since semantic distances between words are relative or "fuzzy", it is more natural to use fuzzy numbers to represent this distance than "clear" numbers. Thereby, the shortest path information found on the fuzzy weighted semantic graph can be exploited to more accurately estimate the semantic similarity of word pairs.

14. Thesis-related publications:

[CT1]. **Van-Tan Bui** and Phuong-Thai Nguyen, Van-Lam Pham. Combining Specialized Word Embeddings and Subword Semantic Features for Lexical Entailment Recognition. Data and Knowledge Engineering, 2022. (SCIE, Q2, IF = 1,5).

[CT2]. **Van-Tan Bui** and Phuong-Thai Nguyen. WEWD: A Combined Approach for Measuring Cross-lingual Semantic Word Similarity Based on Word Embeddings and Word Definitions. The 2021 RIVF International conference on computing and communication technologies, pages 1-6, 2021. (Scopus, DBLP).

[CT3]. **Van-Tan Bui** and Phuong-Thai Nguyen. Measuring semantic similarity of Vietnamese sentences based on lexical similarity and distribution semantic similarity. Lecture Notes in Networks and Systems, pages 259-270, 2021. (Scopus).

[CT4]. Hong-Viet Tran, **Van-Tan Bui**, Dinh-Tien Do, Van-Vinh Nguyen. *Combining PhoBERT and SentiWordNet for Vietnamese Sentiment Analysis*. The 13th International Conference on Knowledge and Systems Engineering (KSE), pages 1-5, 2021. (Scopus, DBLP).

[CT5]. **Van-Tan Bui**, Phuong-Thai Nguyen and Khac-Quy Dinh. *Vietnamese Antonyms Detection Based on Specialized Word Embeddings using Semantic Knowledge and Distributional Information*. The 12th International Conference on Knowledge and Systems Engineering (KSE), pages 159-164, 2020. (Scopus, DBLP).

[CT6]. **Van-Tan Bui**, Phuong-Thai Nguyen, Van-Lam Pham and Thanh-Quy Ngo. *A Neural Network Model for Efficient Antonymy-Synonymy Classification by Exploiting Co-occurrence Contexts and Word-Structure Patterns*. International Journal of Intelligent Engineering and Systems, Vol.13, No.1, pages 156-166, 2020. (Scopus).

[CT7]. **Bui Van Tan**, Nguyen Phuong Thai, Pham Van Lam and Dinh Khac Quy. *Antonyms-Synonyms Discrimination Based on Exploiting Rich Vietnamese Features*. 16th International Conference of the Pacific Association for Computational Linguistics, pages 374-387, 2019. (Scopus, DBLP).

[CT8]. **Bui Van Tan**, Nguyen Phuong Thai, Nguyen Minh Thuan. *Enhancing Performance of Lexical Entailment Recognition for Vietnamese based on Exploiting Lexical Structure Features*. The 10th International Conference on Knowledge and Systems Engineering (KSE), pages 341-346, 2018. (Scopus, DBLP).

[CT9]. **Bui Van Tan**, Nguyen Phuong Thai and Pham Van Lam. Hypernymy Detection for Vietnamese Using Dynamic Weighting Neural Network. 19th International Conference on Computational Linguistics and Intelligent Text Processing, 2018. Lecture Notes in computer science. (Scopus).

[CT10]. **Bui Van Tan**, Nguyen Phuong Thai, and Pham Van Lam. Construction of a Word Similarity Dataset and Evaluation of Word Similarity Techniques for Vietnamese.

The 9th International Conference on Knowledge and Systems Engineering (KSE), pages 65-70, 2017. (Scopus, DBLP).

[CT11]. Van-Vinh Nguyen, Ha Nguyen-Tien, Huong Le-Thanh, Phuong-Thai Nguyen, **Van-Tan Bui**, Nghia-Luan Pham, Tuan-Anh Phan, Minh-Cong Nguyen Hoang, Hong-Viet Tran, Huu-Anh Tran. *KC4MT: A High-Quality Corpus for Multilingual Machine Translation*. The 13th Edition of its Language Resources and Evaluation Conference (LREC), 2022. (Rank A, Scopus, DBLP).

[CT12]. **Bui Van Tan**, Nguyen Phuong Thai and Nguyen Minh Thuan. *Enhancement of measurement efficiency for semantic similarity based on wordnet.* Fundamental and Applied IT Research (Fair), 2017.