INFORMATION ON DOCTORAL THESIS

1. Full name : Le Quang Hung2. Sex: Male

3. Date of birth: 10-Oct-1981 4. Place of birth: QuangBinh

5. Admission decision number: 1142/QĐ-CTSV Dated 28-Dec-2011

6. Changes in academic process: Changes supervisors, admission decision number 48/QĐ-ĐT, dated 25-Jan-2013.

7. Official thesis title: Mining parallel knowledge and applications in English -

Vietnamese machine translation.

8. Major: Computer Science 9. Code: 62 48 01 01

10. Supervisors:

- Assoc. Prof. Le Anh Cuong

- Assoc. Prof. Huynh Van Nam

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

- Adding bilingual data: we proposed several methods to build the bilingual corpus for SMT from the Web and e-book. For sources from the Web, we proposed two new methods of designing content-based features: (i) based on cognation, (ii) based on identifying translation segments. For the source of the e-book, we first use a machine translation system to translate English books into Vietnamese. Then, we connect parallel paragraphs by using some patterns. Finally, we align sentences for these paragraphs.
- Improving alignment model based on constraints: we proposed some improvements to IBM alignment models, which is based on the constraint approach, including anchor constraint, distance constraint, Part-of-Speech constraint, and phrase constraint. These constraints are then used to estimate the parameters of the model in the Expectation-Maximization algorithm.
- Identifying English Vietnamese bilingual phrases for SMT: we proposed a new method to determine the bilingual phrases for SMT. We use a set of syntactic patterns at each language to detect mono-phrases (source phrases); then, we find the translation of a source phrase by a constrained word alignment model.

12. Practical applicability, if any:

The results of the thesis are to increase the quality of English – Vietnamese SMT system. Furthermore, the results have been published in journals and international conferences, thus contributing to science in the field of natural language processing, especially for the development of machine translation systems have the ability to use in reality.

13. Further research directions, if any:

- Firstly, for building the bilingual corpus, we will extend our work on extracting smaller parallel components such as paragraphs, sentences or phrases. This work will also be interesting in the case the quality of translation between bilingual web pages is not good.
- Secondly, the problem of word alignment for SMT: we will handle exceptions occur for constraints as proposed.
- Thirdly, to identify English Vietnamese parallel phrases for SMT: we will use much more syntactic patterns to increase the number of extracted parallel phrases. Also, we plan to use the proposed method for building a dictionary of parallel phrases for the language pair English-Vietnamese and will improve the quality of English-Vietnamese statistical machine translation by using this dictionary.
- 14. Thesis-related publications:
 - [1] Le Quang Hung and Le Anh Cuong (2010), "Extracting parallel texts from the web", *Proceedings of the Second International Conference on Knowledge and Systems Engineering, IEEE Computer Society*, pages 147-151.
 - [2] Le Quang Hung and Le Anh Cuong (2012), "Improving Word Alignment for Statistical Machine Translation Based on Constraints", Asian Language Processing (IALP), International Conference on, IEEE Computer Society, pages 113-116.
 - [3] <u>Le Quang Hung</u> and Le Anh Cuong (2012), "Statistical Word Alignment with Part-of-Speech Constraint", *Kỷ yếu hội thảo Quốc gia lần thứ XV "Một số vấn đề chọn lọc của Công nghệ thông tin và Truyền thông"*, trang 410-416.
 - [4] <u>Quang-Hung LE</u>, Duy-Cuong NGUYEN, Duc-Hong PHAM, Anh-Cuong LE, and Van-Nam HUYNH (2013), "Paragraph Alignment for English-Vietnamese Parallel E-Books", In *Knowledge and Systems Engineering, Springer International Publishing*, pages 251-259.
 - [5] <u>Quang-Hung LE</u>, Anh-Cuong LE, and Van-Nam HUYNH (2013), "Parallel phrase extraction from English-Vietnamese parallel corpora", In *Computing and Communication Technologies, Research, Innovation, and Vision for the Future* (*RIVF*), 2013 IEEE RIVF International Conference on, pages 175-179.

- [6] Le Quang Hung and Le Anh Cuong (2013), "An effective method to sentence alignment for the English-Vietnamese parallel e-book", Kỷ yếu hội thảo Quốc gia lần thứ XVI "Một số vấn đề chọn lọc của Công nghệ thông tin và Truyền thông", trang 12-16.
- [7] <u>Le Quang Hung</u> (2014), "A new approach to extract parallel corpus", *Tap chí khoa học Trường Đại học Quy Nhơn*, Số 4 Tập VIII, trang 12-24.
- [8] <u>Quang-Hung LE</u> and Anh-Cuong LE (2014), "Syntactic pattern based Word Alignment for Statistical Machine Translation", *The International Journal of Knowledge and Systems Science (IJKSS) IGI Global Publishing*, Volume 5 Issue 3, pages 36-45.

Supervisor

Phd Student

Date: 09-Sep-2015 Signature: Date: 09-Sep-2015 Signature:

Full name: Le Anh Cuong

Full name: Le Quang Hung