HOANG Cong Duy Vu

Last Update: April 22, 2018

Contact Information

School of Computing and Information Systems
Melbourne School of Engineering
The University of Melbourne, VIC, Australia
Homepage: https://duyvuleo.github.io/

Mobile Phone: (+61) 459 496 556

E-mail: vhoang2@student.unimelb.edu.au

Personal Information

- Born on March 09, 1984
- Nationality: Vietnamese

Research Interests

My primary research interests lie on Natural Language Processing (NLP) and Applied Machine Learning (ML). I am particularly enthusiastic in exploring NLP and ML techniques to solve real-world problems relating to language technologies.

My current research focus is on Deep Learning models (e.g., sequence to sequence learning/inference) applied to structured prediction problems such as: Machine Translation, Abstractive Summarisation, Parsing. My PhD work involves in two parallel directions:

For training, I have been developing different techniques for incorporating abundant prior or external knowledge for neural sequence models. Also, I am particularly interested in semi-supervised methods which aim to utilise the availability of abundant monolingual data to boost the performance, e.g., in Neural Machine Translation.

For inference, I have been working on better inference methods for structured prediction problems, e.g., global inference in Neural Machine Translation (see my recent paper) within a relaxation framework with bidirectional and bilingual inference; or inference with constraints (work in progress) for complex problems such as: word ordering or constituency/dependency parsing.

Education

The University of Melbourne (UniMelb), VIC, Australia

Doctor of Philosophy (PhD), Engineering, June 2015 - Sep 2018 (expected)

- Dissertation Topic: Incorporating Prior and External Knowledge in Neural Sequence Models
- Supervisors: Assoc. Prof. Trevor Cohn (UniMelb) & Dr. Reza Haffari (Monash University)

National University of Singapore (NUS), Republic of Singapore

Master of Science (MSc), Computer Science, August 2008 - December 2010

- Dissertation Topic: "Towards Automated Related Work Summarization"
- Supervisor: Assoc. Prof. Min-Yen Kan

Vietnam National University at Ho Chi Minh City (VNUHCM), University of Science (HCMUS), Vietnam

Bachelor of Science (BSc) (First Class Hons), Computer Science, October 2002 - August 2006

- GPA: 8.35/10 (Excellent)
- Final Year Thesis: "Topic-based Retrieval of Vietnamese On-line News Articles" (Grade: 10/10)
- Supervisor: Assoc. Prof. Dinh Dien

Honors and Awards

- Google Australia PhD Travel Scholarship for EMNLP 2017
- ASEAN Outstanding Engineering Achievement Award/IES Prestigious Engineering Achievement Award for the project "Speak to Me in My Language", awarded by Prime Minister's Office and Institute of Engineers Singapore (IES) (joint awards with my colleagues at Machine Translation Lab, Institute for Infocomm Research (I^2R), A*STAR, Singapore) (2015)
  Awardees: Prof. Li Haizhou, Aw Ai Ti, **Hoang Cong Duy Vu**, Sharifah Mahani Aljunied, Wang Xuancong, Wu Kui, Nina Zhou, Yeo Boon Hong
  This project will represent Singapore in the ASEAN congress.
- Melbourne International Fee Remission Scholarship and Melbourne International Research Scholarship for PhD study at the University of Melbourne, Australia (2015)
- Japanese Government (MONBUKAGAKUSHO) MEXT Scholarship for PhD Program in Informatics at National Institute of Informatics, Department of Informatics, The Graduate University for Advanced Studies, Tokyo, Japan (2008)
- NUS Graduate Research Scholarship, National University of Singapore, Republic of Singapore (2008)
- Outstanding Young Lecturer Award at Vietnam National University, University of Science, Ho Chi Minh City, Vietnam (2008)
- 1st Prize of the "Festival Challenge IT 2006" Competition (member of **XChange** team) organized by the Faculty of Information Technology - Vietnam National University, University of Science, Ho Chi Minh City, Vietnam (2006)
- Scholarships for excellent students at Vietnam National University, University of Science, Ho Chi Minh City, Vietnam (2002-2006)
- 1st Prize of the Amateur Young Informatics Contest, Ba Ria Town, Ba Ria - Vung Tau Province, Vietnam (2001)

Research Experience

NAVER LABS Europe (previously Xerox Research Centre Europe), Grenoble, France
Research Intern **Mar, 2018 - June 2018**

- I work with Dr. Marc Dymetman on the project "Globally-driven Training Techniques for Neural Machine Translation". I have been developing a new framework based on moment matching technique for incorporating prior and external knowledge in Neural Machine Translation.

Carnegie Mellon University (CMU), Pittsburgh, PA, USA
Language Technologies Institute (LTI)
Visiting Scholar **June, 2017 - August 2017**

- Attended 2017 CMU Summer School on Human Language Technology.
- Participated The 2017 Annual Jelinek Memorial Workshop on Speech and Language Technology (JSALT 2017). My research focus is with the research team in Neural Machine Translation conditioned on low/zero resources.

Institute for Infocomm Research (I^2R), Agency for Science, Technology and Research (A*STAR), Republic of Singapore
Human Language Technology Department (advised by AiTi Aw and Prof. Haizhou Li)
Senior Research Engineer I **April, 2014 - June, 2015**
Research Engineer **March, 2011 - April, 2014**
Research Associate with Baidu-I^2R Research Centre **July, 2012 - June, 2015**

Annual Performance Rating: 4/5 (2011−2012); 5/5 (2012−2013); 4/5 (2013−2014); 4/5 (2014−2015)

Research Projects Attended
Research Project: "VIVA II - Vietnamese to English Machine Translation Project" (in collaboration

with Centre for Strategic Infocomm Technologies (CSIT))
Duration: April, 2011 - January, 2014
Objective:

- Researched & developed robust spelling error detection and correction system for Vietnamese OCR-scanned text documents
- Researched & developed Rule Definition Language (RDL) to collect translation user feedbacks
- Researched & developed hybrid MT system to integrate RDL into state-of-the-art SMT system
- Achievement: My Vietnamese-to-English SMT system achieves competitive performance with Google Translate and significantly outperforms Bing Translate (using BLEU, NIST, METEOR evaluation measures).

Research Project: "Chinese to English Machine Translation Project" (in collaboration with Ministry of Manpower (MOM))
Duration: January, 2014 - March, 2014
Objective:

- Researched & developed a framework that integrates rule-based methods into standard phrase statistical machine translation (SMT) system

Research Project: "NIST OpenMT 2015 Challenge for Chinese to English Translation in informal genres (SMS, Chat, Conversational Telephone Speech - CTS)"
Duration: April, 2014 - January, 2015
Objective:

- Researched & developed I^2R's hybrid SMT system for NIST OpenMT 2015 Challenge, including: data preparation & processing, feature engineering, system development, R&D for state-of-the-art technologies.
- Achievement: Overall, our system was **ranked 3rd** (top 3) out of 20 systems for all categories (2nd for CTS; 3rd for SMS and Chat) in the NIST OpenMT 2015 competition[1].

Research Topic: "Improving I2R's Malay-to-English Statistical Machine Translation System using Hybrid Approach (with Rule-based MT) and Malay-based Factorization Approach"
Duration: January, 2015 - May, 2015
Objective:

- Developed a standalone Malay tokenization tool, including advanced features: processing X-X cases; processing nya-ending words; handling Malay varied numbers; handling others (URL, emails, ...).
- Integrated Rule-based MT system into existing phrase-based statistical machine translation system.

National University of Singapore (NUS), Republic of Singapore
Department of Computer Science, School of Computing

Research Assistant **August, 2008 - December 2010**
Includes Masters level coursework and research projects
Member of the Web Information Retrieval & Natural Language Processing Group (WING)
Advisor: Assoc. Prof. Min-Yen Kan

Research Projects Attended
Research Project: "Automatic Related Work Summarization for Scientific Articles"
Duration: April, 2009 - December 2010
Objective:

- Researched and developed robust methods to automating the related work summarization

Research Project: "Co-training NLP Systems and Language Learners"

[1] `ftp://jaguar.ncsl.nist.gov/mt/mt2015/openmt15results.html`

Duration: January, 2009 - April, 2009
Objective:
- Studied and surveyed about human computation
- Unified human-based computation frameworks for data annotation in NLP

Vietnam National University at Ho Chi Minh City (VNUHCM), University of Science (HCMUS), Vietnam
Department of Knowledge Engineering, Faculty of Information Technology

Assistant Lecturer **August, 2008 - present (on leave)**

Teaching Assistant **October, 2006 - August, 2008**
Teaching courses (as assistant):
- C/C++ Programming on Windows
- Data Structures & Algorithms
- Natural Language Processing

Research Assistant **October, 2006 - August, 2008**
Member of the Vietnamese Computational Linguistics (VCL) Group led by Assoc. Prof. Dinh Dien

Research Projects Attended
Research Project (attended): Government-funded Machine Translation Project
Advisor: Assoc. Prof. Dinh Dien
Objective:
- Mined the web for the bilingual texts
- Integrated Lexical Morphology Knowledge into the English-Vietnamese Statistical Machine Translation (SMT) system
- Proposed a dependency-based word reordering algorithm for English-Vietnamese SMT system
- Developed EVTranslator - an English-Vietnamese SMT system

Research Project (PI): "Research Issues in Vietnamese to English and English to Vietnamese Statistical Machine Translation"
Completed in 2011 - Rated: Excellent (9/10)

National Institute of Informatics (NII), Tokyo, Japan

Research Intern **August, 2007 - January, 2008**
Attended in BioCaster project (`http://biocaster.nii.ac.jp`)
Advisor: Assoc. Prof. Nigel Collier
Objective:
- Researched term matching & analysis for bio-named entities
- Mined time series data

Professional Activities

Program Committee/Reviewers: EMNLP 2013 (Summarization & Generation Track); ACL-IJCNLP 2015 (external); COLING 2016 (Information Retrieval Track); ALTA 2016; ACL 2017 & ACL 2017 Neural Machine Translation Workshop; IJCNLP 2017 (Machine Learning and Machine Translation Tracks); Transactions on Asian and Low-Resource Language Information Processing (TALLIP'17); LREC 2018; ACL 2018 & ACL 2018 workshops (The 2nd Workshop on Neural Machine Translation and Generation; Deep Learning Approaches for Low Resource Natural Language Processing)
Members: The Association for Computational Linguistics (ACL) (2012, 2013, 2014); Asia Pacific Signal and Information Processing Association (APSIPA) (2014-)

Supervision

Undergraduate Projects at HCMUS, VNU-HCM, Vietnam

- Nguyen Thi Thu Suong and Tran Thi Ngoc Thanh - Mining English-Vietnamese Comparable Texts from the Web (completed in 2007, Rated: 8.5/10)
- Luong Viet Thang - Enriching Language Model in Statistical Machine Translation (completed in 2013, Rated: 9.5/10)

Recent Toolkits/Datasets GitHub

- **Transformer-DyNet**: An implementation of purely attention - based neural machine translation toolkit, developed in 2017.
- **Mantidae**: A C++ Lightweight Neural Machine Translation Toolkit, developed from 2016.
- **RWSData**: A Dataset for Related Work Summarization, compiled in 2010.
- **VNTC**: A Large-scale Dataset for Classification of Vietnamese Online News, compiled in 2006.
- more in my github.

Publications (Google Scholar, DBLP)

Journal Papers/Articles

Aobo Wang, **Cong Duy Vu Hoang**, Min-Yen Kan. "*Perspectives on Crowdsourcing Annotations for Natural Language Processing*". Special Issue on Collaboratively Constructed Language Resources, Language Resources and Evaluation (JLRE), 2012. (submitted 2010, accepted 2011, published 2012)

Conference/Workshop Papers

Cong Duy Vu Hoang, Gholamreza Haffari, Trevor Cohn. "*Towards Decoding as Continuous Optimization in Neural Machine Translation*". In Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP), September 7-11, 2017, Copenhagen, Denmark.

Cong Duy Vu Hoang, Gholamreza Haffari, Trevor Cohn. "*Improving Neural Translation Models with Linguistic Factors*". In Proceedings of the 14th Annual Workshop of The Australasian Language Technology Association (ALTA2016), Melbourne, Australia. (Best Student Paper Award)

Cong Duy Vu Hoang, Gholamreza Haffari, Trevor Cohn. "*Incorporating Side Information into Recurrent Neural Network Language Models*". In Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2016) (short paper), June 2016, San Diego, USA.

Trevor Cohn, **Cong Duy Vu Hoang**, Ekaterina Vymolova, Kaisheng Yao, Chris Dyer, Gholamreza Haffari. "*Incorporating Structural Alignment Biases into an Attentional Neural Translation Model*". In Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2016) (long paper), June 2016, San Diego, USA. (Code)

Xuancong Wang, **Cong Duy Vu Hoang**, Kui Wu, Nina Zhou, Boon Hong Yeo, AiTi Aw, Haizhou Li. "*I^2R Chinese-English Translation System for OpenMT 2015*". In Proceedings of the NIST Open Machine Translation evaluation (OpenMT15), Washington DC, USA, June 25−27, 2015.

Cong Duy Vu Hoang, AiTi Aw, Hong-Nhung Nguyen Thi. "*A Rule−Augmented Statistical Phrase−based Translation System*". In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics), System Demonstration Session, ACL 2014, pages 73−78, Baltimore, USA, June 22-27, 2014.

Cong Duy Vu Hoang, AiTi Aw. "*An Unsupervised and Data-Driven Approach for Spell Checking in Vietnamese OCR-scanned Texts*". In Proceedings of the Workshop on Innovative Hybrid Approaches to the Processing of Textual Data (HYBRID 2012), EACL 2012, pages 36−44, Avignon, France, April 23 2012.

Cong Duy Vu Hoang, Min-Yen Kan. "*Towards Automated Related Work Summarization*". In Proceedings of The 23rd International Conference on Computational Linguistics (COLING 2010), pages 427−435, 23−27 August, Beijing, China.

Vu Hoang, Mai Ngo, Dien Dinh. "*A Dependency-based Word Reordering Approach for Statistical Machine Translation*". In Proceedings of The 6th IEEE International Conference on Research, Innovation, and Vision for the Future - Information and Telecommunication Technologies (IEEE-RIVF'08), pages 120−127, July 13-17, 2008, Ho Chi Minh City, Vietnam.

Vu Hoang, Nguyen Nguyen, Dien Dinh, Nigel Collier. "*Topic-based Vietnamese News Document Filtering in The BioCaster Project*". In Proceedings of The 6th International Conference on Advanced Language Processing and Web Information Technology (ALPIT 2007), IEEE CS, pages 224-229, 22-24 August 2007, Luoyang, China.

Vu Cong Duy Hoang, Nguyen Le Nguyen, Hung Quoc Ngo, Dien Dinh. "*A Comparative Study on Vietnamese Text Classification Methods*" . In Proceedings of The 5th IEEE International Conference on Research, Innovation, and Vision for the Future - Information and Telecommunication Technologies (IEEE-RIVF'07), pages 267-273, March 05-09, 2007, Ha Noi, Vietnam.

Hoang Cong Duy Vu, Van Chi Nam, Dinh Dien. "*English-to-Vietnamese and Vietnamese-to-English Statistical Machine Translation with Vietnamese Lexical Language Knowledge*". In Proceedings of National Conference on Information Technology, August 2007, Ha Noi, Vietnam. (in Vietnamese)

Technical/Unpublished Reports

Aobo Wang, **Cong Duy Vu Hoang**, Min-Yen Kan. "*Perspectives on Crowdsourcing Annotations for Natural Language Processing*". Technical Report at Department of Computer Science, School of Computing, National University of Singapore, TRB7/10, Jul 2010. (online version: http://dl.comp.nus.edu.sg/dspace/handle/1900.100/3266)

Cong Duy Vu HOANG. "*Automatic Scientific Related Work Summarization*". Graduate Research Paper, Nov 2009, National University of Singapore.

Vu HOANG. "*Experiments in Grounding Task for Bio-Named Entities*". Final Internship Report at National Institute of Informatics, Jan 2008, Tokyo, Japan.

Theses

Cong Duy Vu HOANG. "*Towards Automated Related Work Summarization*". MSc Thesis, Dec 2010, National University of Singapore.

Hoang Cong Duy Vu. "*Topic-based Retrieval of Vietnamese On-line News Articles*". Bachelor Thesis (in Vietnamese), Aug 2006, University of Science, Vietnam National University at Ho Chi Minh City, Vietnam.

Computer Skills

- Programming Languages: .NET (C# & VB), C/C++ (with STL and Boost), Java, Perl, Python, and Unix shell scripts.
- Deep Learning toolkits: DyNet, PyTorch, Tensorflow
- Data Analysis: Weka, Matlab/Octave.
- Databases: SQL.
- Technologies: DLL, MFC/STL, COM, DCOM, COM+, XML, UML, Windows Services.
- Applications: LaTeX, MindMap.
- Operating Systems: Linux, Windows.

Languages

- Vietnamese (Native)
- English (Professional)

References

Assoc. Prof. Trevor Cohn

Associate Professor
Computing and Information Systems, The University of Melbourne, VIC, Australia
Email: tcohn@unimelb.edu.au

Dr. Gholamreza (Reza) Haffari

Senior Lecturer
Faculty of Information Technology, Monash University, VIC, Australia
Email: gholamreza.haffari@monash.edu

Ms. AiTi Aw

Senior Research Manager
Institute for Infocomm Research (I^2R), A*STAR, Singapore
Email: aaiti@i2r.a-star.edu.sg

Assoc. Prof. Dinh Dien

Research Associate Professor
Vietnam National University (VNU-HCM) at Ho Chi Minh City, University of Science (HCMUS)
Email: ddien@fit.hcmus.edu.vn

Prof. Haizhou Li

Director (Research)
Institute for Infocomm Research (I^2R), A*STAR, Singapore
Email: hli@i2r.a-star.edu.sg

Assoc. Prof. Min-Yen Kan

Associate Professor
School of Computing, National University of Singapore
Email: kanmy@comp.nus.edu.sg