Word Representation with Transferable Semantics

by Qian Liu

Thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

under the supervision of Distinguished Prof. Jie Lu and Associate Prof. Guangquan Zhang

University of Technology Sydney
Faculty of Engineering and Information Technology

May 2021

Certificate of Original Authorship Template

Graduate research students are required to make a declaration of original authorship when they submit the thesis for examination and in the final bound copies. Please note, the Research Training Program (RTP) statement is for all students. The Certificate of Original Authorship must be placed within the thesis, immediately after the thesis title page.

Required wording for the certificate of original authorship

CERTIFICATE OF ORIGINAL AUTHORSHIP

I, *Qian Liu* declare that this thesis, is submitted in fulfilment of the requirements for the award of *Doctor of Philosophy*, in the *Faculty of Engineering and Information Technology* at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.
**If applicable, the above statement must be replaced with the collaborative doctoral degree statement (see below).*

**If applicable, the Indigenous Cultural and Intellectual Property (ICIP) statement must be added (see below).*

This research is supported by the Australian Government Research Training Program.

Signature: Production Note:
Signature removed prior to publication.

Date: 18/05/2021

Collaborative doctoral research degree statement

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of the requirements for a degree at any other academic institution except as fully acknowledged within the text. This thesis is the result of a Collaborative Doctoral Research Degree program with *Beijing Institute of Technology*.

Indigenous Cultural and Intellectual Property (ICIP) statement

This thesis includes Indigenous Cultural and Intellectual Property (ICIP) belonging to *N/A*, custodians or traditional owners. Where I have used ICIP, I have followed the relevant protocols and consulted with appropriate Indigenous people/communities about its inclusion in my thesis. ICIP rights are Indigenous heritage and will always remain with these groups. To use, adapt or reference the ICIP contained in this work, you will need to consult with the relevant Indigenous groups and follow cultural protocols.

Abstract

Semantic representation aims to encode the meaning of text (e.g., words) in a form which can be stored and processed by a machine, such as real-valued vectors or neural networks with well-trained parameters. In particular, semantic knowledge is expected to be embodied in representations. For example, words with similar meanings are expected to be close to each other when they are represented as vectors. Semantic representation is the basic block of neural networks, and it should have better expression ability to support downstream natural language processing applications.

Although recent research on semantic representation has shown a reasonable ability to represent textual data by only using large-scale raw text, most research is incomplete and biased as it only models the surface co-occurrence information of corpora but ignores deep semantic and syntactic information. In addition, most research focuses on modeling generic semantics, while disengaging from task requirements. Hence, existing semantic representation methods still face several unsolved and challenging problems in the real world.

This thesis aims to design better representation learning methods by utilizing *transferable* semantics extracted from source domains, which are resourceful and beyond raw text. More specifically, this thesis aims to address four problems faced by existing semantic representation methods: 1) how to reliably transfer semantics from a structural knowledge base to an unstructured representation space; 2) how to reliably transfer semantics from multiple source domains to a low-resource target domain; 3) how to achieve the reliable and low-cost cross-lingual transfer of semantics; and 4) how to adapt semantic representations for specific applications.

To address Problem 1), this thesis designs two assumptions to model semantic structures in knowledge bases and proposes a new semantic structure-based semantic representation method (Chapter 3). It leverages the human-defined relationships among words from structured knowledge bases as transferable semantics to improve its representation ability. Instead of using the relations between word-pairs, our method uses whole semantic structures which have proven to be more effective in semantic representation.

To address Problem 2), this thesis proposes a dynamical meta-embedding method to leverage the semantics from multiple source domains (Chapter 4). It leverages latent knowledge from multiple source embeddings to improve representation learning for a low-resource domain. Considering domain shifts and quality discrepancy, it dynamically aggregates multiple source embeddings by a differentiable attention module, instead of using them equally. It is proven to be more suitable to transfer true required semantics from multiple source domains to a low-resource domain.

To address Problem 3), this thesis proposes a new method to bridge the cross-lingual semantic gap with limited bilingual resource reliance (Chapter 5). Based on multilingual embeddings, it learns a pivot set which is semantically related to a low-resource language and lexically related to a high-resource language. With the learned pivots, our method is useful to help models trained on high-resource languages to be adapted on low-resource languages.

To address Problem 4), this thesis proposes a fuzzy word similarity measure to adapt general semantic representations according to the need of a specific task (Chapter 6). It takes task-oriented features into consideration and adapts general semantics to the specific tasks, which alleviates the problem of disengaging from task requirements.

To conclude, this thesis proposes a set of effective methods to improve semantic representation by exploring and modeling knowledge beyond raw text and places an emphasis on encoding task-specific features for real-world applications.

Acknowledgments

It is an exciting journey at University of Technology Sydney (UTS) for pursuing my Ph.D. degree in the past four years. I am sincerely grateful to the people who inspired and helped me in many ways. I would like to express my foremost and deepest gratitude to my principal supervisor, Distinguished Professor Jie Lu. Her decisiveness and sharp insights continuously motivated me when I got lost or afraid about the future. Her confidence and enthusiasm inspired me to do the right thing even when the road got tough. She placed considerable trust in my research ability and unconditionally support me in pursuing my own research interests. Her wisdom and immense knowledge always enlightened me to go further and deeper in my research. I felt extremely honored to be guided by such a rigorous researcher as well as an enthusiastic mentor. What she taught me and what I learned from her in the past four years has benefited my Ph.D. study and will be a great treasure throughout my life.

Meanwhile, I am greatly indebted to my co-advisor, A./Professor Guangquan Zhang. Without his patience and encouragement, I would not have been able to complete this Ph.D. program. He taught me step by step how to become a qualified researcher from its beginning. He always led me to the right research direction with his expert knowledge of theory and abundant research experience. Without his critical comments, I would waste my time on trivial research ideas. Discussion with him greatly improves the scientific aspect and quality of my research. He helped me to build my confidence in my research outcomes and to be hopeful when faced with any difficulty, from academic to living.

I would like to express my thankfulness to every member of the Decision Systems & e-Service Intelligence Lab (DeSI) in the Australian Artificial Intelligence Institute (AAII).

It was a wonderful experience to spend four years with these dedicated researchers. I especially thank Dr. Junyu Xuan, Dr. Tao Shen, Dr. Yi Zhang, Dr. Fan Dong, Dr. Feng Gu, Dr. Anjin Liu and Dr. Hua Zuo who provided insightful comments related to my research problem during my Ph.D. candidature; Dr. Ruiping Yin, Dr. Hang Yu, Dr. Guanjin Wang, Dr. Chenlian Hu, Dr. Feng Liu, Dr. Yiliao Song, Dr. Adi Lin, and Bin Wang who have shared their opinions and comments with me; Dr. Dan Shang, Dr. Shan Xue, Dr. Guanjin Wang, and Dr. Xiaohang Xu, who shared my joys and sadness.

Last, I would like to express my heartfelt appreciation and gratitude to my parents and families for their love and support.

List of Publications

1. **Qian Liu**, Xiubo Geng, Jie Lu, Daxin Jiang. Pivot-based Candidate Retrieval for Cross-lingual Entity Linking. *Proceedings of the 30th The Web Conference (WWW-21)*, Ljubljana, Slovenia, April 12-16, 2021. [ERA: A, CORE: A*]

2. **Qian Liu**, Xiubo Geng, Tao Qin, Heyan Huang, Jie Lu, Daxin Jiang. MGRC: An End-to-End Multi-Granularity Reading Comprehension Model for Question Answering. *IEEE Transactions on Neural Networks and Learning Systems (IEEE-TNNLS)*, 2021. [ERA&CORE: A*, JCR Q1]

3. **Qian Liu**, Jie Lu, Guangquan Zhang, Tao Shen, Zhihan Zhang, Heyan Huang. Domain-specific meta-embedding with latent semantic structures. *Information Sciences*, Vol. 555, pp. 410-423, May 2021. DOI: 10.1016/j.ins.2020.10.030. [CORE: A, JCR Q1]

4. **Qian Liu**, Heyan Huang, Junyu Xuan, Guangquan Zhang, Jie Lu. A Fuzzy Word Similarity Measure for Selecting Top-k Similar Words for Query Expansion. *IEEE Transactions on Fuzzy Systems (IEEE-TFS)*, 2020. DOI: 10.1109/TFUZZ.2020.2993702. [ERA&CORE: A*, JCR Q1]

5. **Qian Liu**, Heyan Huang, Guangquan Zhang, Yang Gao, Junyu Xuan, Jie Lu. Semantic Structure based Word Embedding by Incorporating Concept Convergence and Word Divergence. *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18)*, pp.5261-5268, New Orleans, LA, USA, 2018. [ERA: A, CORE: A*]

6. **Qian Liu**, Heyan Huang, Jie Lu, Yang Gao, Guangquan Zhang. Enhanced Word Embedding Similarity Measures using Fuzzy Rules for Query Expansion. *Proceedings of the 2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE-17)*, pp. 1-6, Naples, Italy, 9-12 July, 2017. [ERA&CORE: A]

7. **Qian Liu**, Jie Lu, Guangquan Zhang. Towards Zero-shot Cross-lingual Entity Linking. *IEEE Transactions on Knowledge and Data Engineering (IEEE-TKDE)*, 2021. [CORE: A*, ERA: A, JCR Q1] (Under Review)

8. Yi Zhang, Jie Lu, Feng Liu, **Qian Liu**, Alan Porter, Hongshu Chen, Guangquan Zhang. Does deep learning help topic extraction? A kernel k-means clustering method with word embedding. *Journal of Informetrics*, Vol. 12, no. 4, pp. 1099-1117, 2018. [ERA&CORE: A, JCR Q1]

Note: Chapter 3 relates paper 5, Chapter 4 relates paper 3, Chapter 5 relates paper 1, and Chapter 6 relates paper 4.

Table of Contents

List of Figures

List of Tables