

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 (AII).

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 Publications	v
List of Figures	xi
List of Tables	xiii
1 Introduction	1
1.1 Background	1
1.2 Motivation	3
1.3 Research Questions and Objectives	4
1.4 Research Innovation and Contributions	9
1.4.1 Research Innovation	9
1.4.2 Research Contributions	10
1.5 Research Significance	12
1.6 Thesis Structure	13
2 Literature Review	17
2.1 Semantic Representation	17
2.1.1 Distributed Representation	17
2.1.2 Contextual Representation	19
2.2 Transfer Learning	21
2.3 NLP Applications	23

3	Knowledge-enhanced Word Representation by Incorporating Concept Convergence and Word Divergence	25
3.1	Introduction	25
3.2	Background	27
3.3	Semantic Structure-based Word Embedding	28
3.4	Experiments and Results	33
3.4.1	Initialization and Parameters	33
3.4.2	Word Similarity and Word Analogy	35
3.4.3	Text Classification	38
3.4.4	Query Expansion	39
3.5	Summary	41
4	Dynamic Meta-Embedding with Domain-Specific Latent Semantic Structures	43
4.1	Introduction	43
4.2	Background	45
4.2.1	Meta-embedding Learning	45
4.2.2	Graph Convolution Networks in NLP	46
4.3	Dynamical Meta-Embedding Method	47
4.3.1	Dynamically Combined Meta-embeddings	48
4.3.2	Domain-specific Knowledge	49
4.3.3	Learning Process	53
4.4	Experiments	53
4.4.1	Implementation	54
4.4.2	Baselines	55
4.4.3	Task I: Text Classification	56
4.4.4	Task II: Relation Extraction	59
4.5	In-depth Analyses	61
4.5.1	Ablation Study	61
4.5.2	Parameter Validation	62

4.5.3	Weight Visualization	63
4.5.4	Case Study	64
4.6	Summary	64
5	Pivot-based Cross-lingual Retrieval With Limited Bilingual Resource	65
5.1	Introduction	65
5.2	Background	69
5.3	Task Description	70
5.4	Pivot-based Candidate Retrieval	71
5.4.1	Filling the Cross-lingual Gap	72
5.4.2	Selective Mechanism	74
5.4.3	Filling the Mention-Entity Gap	76
5.5	Experiments	77
5.5.1	Datasets	77
5.5.2	Baselines	78
5.5.3	Main Results	80
5.5.4	In-depth Analysis	83
5.5.5	Case Study	86
5.6	Summary	87
6	Fuzzy Similarity Measure Based on Refined Semantic Representation	89
6.1	Introduction	89
6.2	Background	91
6.2.1	Top- k Word Selection	91
6.2.2	Association Rules	92
6.2.3	Fuzzy System and Its Application in Similarity Measures	93
6.3	Fuzzy Word Similarity Measure	94
6.3.1	Local Measure: Word Embedding	94
6.3.2	Global Measure: Association Rules	95
6.3.3	Fuzzy Word Similarity Measure	97

TABLE OF CONTENTS

6.4	Experiments	103
6.4.1	Datasets	103
6.4.2	Baselines	105
6.4.3	Implementation and Baselines	106
6.4.4	Results	108
6.5	In-depth Analysis	110
6.5.1	Component-wise Validation	110
6.5.2	Local Measure	112
6.5.3	Expansion Size	113
6.5.4	Parameter λ	114
6.5.5	Case Study	115
6.6	Summary	116
7	Conclusion and Future Study	117
7.1	Conclusions	117
7.2	Future Study	120
	Bibliography	123
	Appendix	147

LIST OF FIGURES

FIGURE	Page
11 Thesis structure	15
31 We mark two types of words in the sentences with blue color and red color, respectively. The underlined words are their context in the corpus. The graphs in the right are their semantic structures generated from WordNet.	26
32 An example of the three-level semantic structures of the word <i>dog</i> in WordNet.	30
33 Performance of the SENSE method with varying parameters of α and β	33
34 Performance over varying parameters on the WordSim 353 dataset.	34
41 An overview of the proposed meta-embedding framework. The dashed lines denote the workflow of modeling contextual information, and the solid lines denote the workflow of modeling semantic structures.	47
42 Top related words of <i>KitchenAid</i> , captured by the co-occurrence matrix \mathbf{M} (left), and the domain-specific graph \mathcal{G} (right). As seen, \mathbf{M} and \mathcal{G} address different levels of semantic context/structures. A comprehensive use of both leads to a more complete understanding of the meaning of words.	51
43 Performance for parameter selection on text classification task for meta-embedding dimension d and α	62
44 Visualization of self-attention-based combination weights (a deeper color signifies a higher weight). <i>CBOW</i> , <i>GloVe</i> , and <i>fastText</i> are the pre-trained source embeddings. <i>Context</i> denotes \mathbf{c} . <i>Target</i> denotes \mathbf{x}	63

51	Comparison of lexicon-based method, semantic-based method, and our pivot-based method.	67
52	An example to illustrate our pivot-based approach.	73
53	R@1000 on QALD to investigate the effectiveness of the NMS component. . .	84
54	Influence of the size of intermediary collection on the QALD dataset. The x-axis shows the size of the intermediary collection, the left y-axis corresponds to the average R@1000 across eight languages, and the right y-axis denotes R@1000 of each language.	85
55	Comparison between our method and Google translator on the QALD dataset. The y-axis denotes R@1000 score of candidate retrieval.	86
56	Examples in the QALD dataset. The red plausible mentions are salient mentions to recall gold entity, marked by human evaluation.	87
61	The framework of the proposed fuzzy word similarity measure based on both local measure and global measure in the corpus.	94
62	An illustration of local and global information in some example sentences. The context window used in word embedding methods is marked by the pale red rectangles. For <i>programming</i> , its local related words and global related words are linked in blue and orange, respectively. Several examples of association rules are listed on the right-hand-side of the figure.	96
63	The component-wise validation performance for each dataset	111
64	The influence of the local measure with different settings in terms of P@5 on the RCV1 dataset. (a) A comparison with varying context window sizes. (b) A comparison between the original Word2Vec model and the its variation NP2Vec, and <i>Local</i> denotes the query expansion method with only a local measure.	112
65	The effect of varying the size of the query expansion in terms of P@5 with different datasets.	113
66	The performance with a varying λ on five development datasets. The Y-axis represents P@5, and the X-axis represents different λ	114

LIST OF TABLES

TABLE	Page
31 Results on the word similarity task and the word analogy task. The word embedding methods are divided into three groups. Bold scores are the best within the groups. Underlined scores are the best overall.	36
32 Evaluation results of multi-class text classification. Bold scores denote the SENSE method outperforms the corresponding baseline methods. Underlined scores are the best overall.	38
33 Performance of different methods for query expansion on the RCV1 dataset. Bold scores denote that the SENSE method outperforms the corresponding baseline methods. Underlined scores are the best overall.	40
41 Symbols and the descriptions. Matrix is denoted with bold capital letter, vector is denoted with bold lowercase letter, and scalar number is denoted with lowercase letter.	48
42 Statistics of the corpora used in our experiments. #Vocab is the size of vocabulary. Avg.Len. is the average length of documents in the corpus.	55
43 Overall performance for the text classification task, conducted on four subsets (i.e., <i>Kitchen</i> , <i>DVD</i> , <i>Book</i> , and <i>Electronics</i>) of <i>Amazon Reviews</i> , <i>Custom Reviews</i> (CR), and <i>TREC</i> datasets. The highest scores are marked in bold.	57
44 The comparison of our method with the pre-trained language models.	58
45 Overall performance on the relation extraction task.	59
46 Evaluation of essential components and different GCNs.	61
47 Top five words related to <i>season</i> in different embedding spaces.	63

51	Terminology and the corresponding description and examples used in the cross-lingual candidate retrieval task.	71
52	Top-1000 recall (R@1000) of different methods on the QALD dataset. #Mentions denotes the number of mentions for each language in QALD.	78
53	Comparison of different methods in terms of average recall on QALD dataset. CR denotes the candidate retrieval in XEL. ED denotes entity disambiguation on the top-1000 candidate entities.	81
54	(R@30) on WIKI-LRL. PBEL_Char and PBEL_BiLSTM denote the PBEL method which encodes entities into vectors using BiLSTM and character-based CNN, respectively.	82
55	R@1000 on the QALD dataset to investigate the influence of character information. OOV denotes the percentage of our-of-vocabulary mentions. Δ denotes the performance improvement.	83
61	Examples of the most similar words selected using word embedding and association rules for two given words (<i>crime</i> and <i>feeling</i>). RCV1 corpus is used here which is detailed in Section IV-A.	100
62	The average percentage of the retrieval gain, neutral, and loss, of 50 collections in the RCV1 dataset (detailed in Section IV-A). Common words are detected by both local measure and global measure.	100
63	The statistics for each dataset used in the experiments.	103
64	Comparison of the proposed fuzzy word similarity measure with other similarity measures in terms of P@5, P@10, and MAP. Avg.Impr(%) is the average percentage of improvement of FWS over other baselines.	108
65	The Top-ten selected words for five queries using the original word embedding method (denoted as <i>Original</i>) and the fuzzy word similarity measure (denoted as <i>FWS</i>) on the RCV1 dataset. The words in bold are contributed by the global measure, and the followed score is S_{global}	115
1	Abbreviations and their explanations.	147