



Semantic Enhancement for Text Representation

By
Wenfeng HOU

A Thesis Submitted in Partial Fulfillment of
the Requirements for the Degree

Master of Analytics (Research)

University of Technology Sydney
Faculty of Engineering and Information Technology

June 2021

CERTIFICATE OF ORIGINAL AUTHORSHIP

I, Wenfeng HOU declare that this thesis, is submitted in fulfilment of the requirements for the award of Master degree, in the School of Computer Science, 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.

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

Production Note:

Signature: Signature removed prior to publication.

Date: 13 . 06 . 2021

ACKNOWLEDGMENTS

My great gratitude goes first and foremost to my supervisor Professor Longbing Cao for the persistent support of my research and study. Without his help and guidance, this thesis could not have been brought to the present appearance.

I also would like to thank Dr. Yanchang Zhao, who, as a co-supervisor, was always happy to help me for my research and study. His guidance helped me in all the time of research and writing of this thesis.

In addition, I would like to thank the fellow lab mates who study in Advanced Analytics Institute. Without their support, I could not finish my research here.

Last but not least, I would like to thank my family for their unconditional support throughout my whole master study.

CONTENTS

Certificate of Original Authorship	iii
Acknowledgments	v
Contents	vi
List of Figures	ix
List of Tables	xi
Publication	xiii
Abbreviation	xv
Abstract	xvii
1 Introduction	1
1.1 Background	1
1.1.1 Text Representation	1
1.1.2 Short Text Representation and Semantic Enhance- ment	3
1.1.3 Term Weighting Scheme for Text Analysis	6
1.2 Research Issues, Objectives and Contributions	7

1.2.1	Research Issues	7
1.2.2	Objectives	8
1.2.3	Research Contributions	9
1.3	Thesis Structure	9
2	Literature Review	11
2.1	Text Representation and Semantic Enhancement for Short Text Representation	11
2.1.1	Word Representation	11
2.1.2	Short Text Representation and Semantic Enhance- ment for Short Text Representation	15
2.2	Term Weighting Scheme for Text Analysis	23
3	Multiple Aspects-based Short Text Representation with Named Entity, Concept and Knowledge	29
3.1	Introduction	29
3.2	Multiple Aspects-based Short Text Representation with Named Entity, Concept and Knowledge	32
3.2.1	Semantic Information Retrieval Module	33
3.2.2	Feature Extraction Module	35
3.2.3	The Attention Module	38
3.3	Experiments	39
3.3.1	Datasets	40
3.3.2	Data Pre-processing	41
3.3.3	Baselines	42
3.3.4	Parameter Setting	43
3.3.5	Result Analysis	43

3.3.6	Parameter Sensitivity	46
3.3.7	Computational Cost	47
3.4	Conclusions	48
4	A Hybrid Term Weighting for Text Analysis	49
4.1	Introduction	49
4.2	A Hybrid Term Weighting Method for Text Analysis	51
4.2.1	Frequency-based Term Weighting	52
4.2.2	Semantic-based Term Weighting	52
4.3	Experiments	55
4.3.1	Datasets	55
4.3.2	Baseline	56
4.3.3	Parameter Setting	57
4.3.4	Result Analysis	57
4.4	Conclusions	58
5	Conclusions and Future Work	61
5.1	Conclusions	61
5.2	Future Work	62
	Bibliography	65

LIST OF FIGURES

FIGURE	Page
1.1 Thesis' structure	10
2.1 Architecture of CBOW and Skip-gram models [39]	12
2.2 Elmo model architecture ¹	14
2.3 GPT model architecture [44]	15
2.4 Illustration of TransE ¹	20
2.5 Illustration of TransH ¹	21
2.6 Illustration of TransR ¹	21
3.1 The framework of Entity-based Concept Knowledge-Aware (ECKA).	33
3.2 The semantic information retrieval module.	34
3.3 Accuracy distribution	46

LIST OF TABLES

TABLE	Page
2.1 Comparison of current text representation and semantic enhancement models	24
2.2 Comparison of current term weighting models	28
3.2 The twitter data set	40
3.1 The google snippet data set	41
3.3 The ag news data set	41
3.4 Text classification accuracy comparison of different models . .	44
3.5 The text classification accuracy comparison of ECKA variants	44
4.1 The huffpost news data set	56
4.2 The bbc news data set	56
4.3 Text classification accuracy	57

PUBLICATION

Accepted Paper:

W. HOU, Q. LIU, AND L. CAO, Cognitive aspects-based short text representation with named entity, concept and knowledge, *Applied Sciences*, 10 (2020), p. 4893.

ABBREVIATION

BERT - Bidirectional Encoder Representations from Transformers

BiLSTM - Bidirectional Long Short Term Memory

BoW - Bag of Words

CBOW - Continuous Bag-Of-Words

CNN - Convolutional Neural Networks

ECKA - Entity-based Concept Knowledge-Aware model

ELMo - Embedding from Language Models

GPT - Generative Pre-Training

GRU - Gated Recurrent Unit

KG - Knowledge Graph

KNN - K Nearest Neighbor

LSTM - Long Short Term Memory

NLP - Natural Language Processing

RNN - Recurrent Neural Networks

SVM - Support Vector Machine

TF-IDF - Term Frequency - Inverse Document Frequency

TF-ICF - Term Frequency - Inverse Category Frequency

VSM - Vector Space Model

ABSTRACT

Thanks to recent developments in web technology, various textual information can now be found online, including social media, news, product reviews and instant messages. How to automatically classify and organize such texts is currently a topic of great interest. In Natural Language Processing (NLP), text classification is a traditional task and text representation is its foundation. To represent text, we need to obtain a word's representation. The existing language representation models, including Word2vec, ELMo, GPT and BERT, were widely used for word representation. These word representation models were highly successful at processing natural languages. However, they mainly captured implicit representations. Other models that analyzed a text's context can potentially capture richer information which can help deep neural networks gain a better understanding of the text. It is crucial to incorporate semantic information into the text representation because the rich semantics associated with word representations can supplement text representation. New approaches are necessary to represent semantics in combination with existing text representations.

The models presented in this study improved text representation and term weighting by utilizing external knowledge to address the above-mentioned research needs. In contrast to previous work, the models proposed here used multi-level knowledge to facilitate the semantic enhancement of text representation by involving external semantic information.

In Chapter 3, we proposed an Entity-based Concept Knowledge-Aware (ECKA) representation model to incorporate semantic information into short text representations. ECKA is a multi-level short text semantic enhancement model for short text representations which extracts semantic features from the word, entity, concept and knowledge levels by CNN. Since word, entity, concept and knowledge entity in the same short text

have different informativeness for short text classification, attention networks were formed to capture aspects-oriented attentive representations from a text's multi-level textual features. The final multi-level semantic representations were formed by concatenating all these individual-level representations, which were then used for text classification.

In Chapter 4, we proposed a hybrid term weighting method that works by utilizing frequency and semantic similarities for the term weighting calculation. When analyzing a term, we first used the Term Frequency-Inverse Document Frequency (TF-IDF) to calculate term weighting. Next, we used a named-entity-based concept-sense disambiguation process to obtain concepts. Following that, we calculated the term's semantic similarity to the document. The TF-IDF weights were then revised according to the term's semantic similarities to reflect both frequency and semantic similarities of the various terms in the text.

All of these models were applied to the text classification tasks. The proposed models' performance in semantic enhancement were compared with different methods to demonstrate their effectiveness.