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.

Signature: Production Note: 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

List of Figures

List of Tables

PUBLICATION

Accepted Paper:

W. HOU, Q. LIU, AND L. CAO, Cognitive aspects-based short textrepresentation 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 gained 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 abovementioned 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.