Computation on Meanings:
Content-based Feature Analysis for Semantic Interpretation

DAT TAN HUYNH

A dissertation submitted in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy

in

Information Sciences and Engineering

University of Canberra

February, 2015

University of Canberra

Abstract

Computation on Meanings: Content-based Feature Analysis for Semantic Interpretation

DAT TAN HUYNH

Building a computer system that has extraordinary capacity to understand human language has become one of the greatest challenges. One of the barriers in building such a system is to construct a semantic interpretation of human language.

Many prior approaches, which have been proposed to address the issue, can be categorized into knowledge-based approaches and content-based approaches. While knowledge-based approaches utilise human-crafted knowledge repositories to construct semantic interpretation, content-based approaches analyse a large amount of unstructured text data available. Although knowledge-based approaches produce outstanding performance compared to content-based approaches on semantic interpretation as well as semantic distance,they are limited themselves within in particular text knowledge domains. On the other hand, despite content-based approaches have the ability to process unstructured text data on different domains and languages, there are certain aspects that are worthy of considerations.

First, most of the prior content-based approaches popularly use a single type of

feature aspects to construct word meaning representation. This raises a concern how multiple feature aspects can be used to model meaning representation. Secondly, experiments of the content-based approaches using various sets of features were undertaken, their performances tested on the task of measuring semantic distance are still under expectation.

The main focus of this research is to propose new content-based approaches for semantic interpretation of human language. By undertaking semantic analysis on large amount of unstructured text available, our proposed methods have presented new sets of features for semantic interpretation. On the one hand, multiple aspect sets of features have been proposed, which help to cover semantic meanings of words in different angles. On the other hands, feature transformations and combinations have been proposed to not only reduce size of feature dimensions, but also encorage the interaction between different aspects of word features.

The contributions of this research are listed as the followings:

– Relational Feature Analysis For Semantic Interpretation : this approach constructs representing features of a word by considering its relations extracted from the word's local contexts as well as the hidden aspects built from the sets of relations. The effectiveness of generated features is evaluated based on the task of measuring semantic distance. Experimental results have demonstrated the promising capacity of the relation-based features in modelling word mean-

ing compared to traditional context-based features when tested on the same benchmarks.

– Conceptual Topic Analysis For Semantic Distance: this approach introduces a new way to construct a semantic profile of word meanings and measuring semantic distance by using topical clues from surrounding contexts to characterise meanings of a word. With the experiment on various standard benchmarks, the method demonstrates outstanding performance compared to related methods using topical information.

– Multi-way Feature Analysis for Semantic Interpretation: this approach proposes a tensor-based technique for semantic interpretation by building meaning representation of words directly from text and does not require pre-existing linguistic knowledge. Taking in to account structural information such as word order and syntactic information, the method that utilises tensor analysis to build representation of word meaning. This content-based model demonstrates significantly improved performance when compared to a robust baseline model on a number of semantic distance measures.

The success of semantic interpretation of words contributes for building a reliable metric for semantic distance, which involves in most tasks of natural language processing and understanding.

ACKNOWLEDGMENTS

I am grateful to many people who have assisted me throughout my research journey.

First, I am thankful for the generous support provided by the chair of supervision panel, Associate Professor Dat Tran. I am honoured to have an opportunity to work with him and thankful for his endless guidance, support, and encouragement.

I would also like to express my sincere gratitude to supervisor Assistant Professor Wanli Ma, for his supports and feedbacks, collaborations during my PhD journey. I am grateful to Professor Dharmendra Sharma, Mr Hanh Huynh, Dr Kim Le for consistent helps and supports within the Faculty of Information Science and Engineering and The Faculty of ESTEM.

I specially acknowledge the University of Canberra and family of the late W. J. Weeden for their postgraduate research scholarship to conduct this research. I also thank the Turing Centre at the University of Washington, Stanford NLP Group for providing me tools, data, and resource to conduct experiments during my research, and special thank Wikipedia for such a valuable data

repository. I also thank the Institude for Applied Ecology for providing access to the high end cluster during my experiments.

I would like to thank everyone in the Faculty of ISE, ESTEM–especially colleagues in CIKADA group including Dr Tuan Hoang, Phuoc Nguyen, Khoa Nguyen, Tan Vo, Dung Pham, Dinh Phung–for providing helpful assistances, discussions and productive collaborations. I also specially thank Harriet Searcy for taking the hard part in proofreading my thesis.

Finally, I specially thank my wife and my families back home in Vietnam for all their love and support, for letting me to follow my dreams.

DEDICATION

To my wife, and my parents

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

GLOSSARY

KB: Knowledge-based

CB: Content-based

FF: Frequency Filter

IVF: Information Value Filter

LSA: Latent Semantic Analysis

HAL: Hyperspace Analogue to Language

ESA: Explicit Semantic Analysis

VSM: Vector Space Model

NMF: Non-negative Matrix Factorization

SVD: Singular Vector Decomposition

SSA: Salient Semantic Analysis

LDA: Latent Dirichlet Allocation

DF: Document Frequency

ESA: Explicit Semantic Analysis

IDF: Inverse Document Frequency

IG: Information Gain

MI: Mutual Information

PMI: Pointwise Mutual Information

TFIDF: A term weighting approach based on Term Frequency and Inverse Document Frequency

WWW: World Wide Web

WS-353: Word Similarity dataset with 353 pairs of words.

RG-65: A dataset created by (Rubenstein & Goodenough 1965) containing 65 pairs of words.

MTURK: A dataset created by (Radinsky et al. 2011) containing 287 pairs of words.

PARAFAC: Parallel Factor Analysis