

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**
AUSTRALIA'S CAPITAL UNIVERSITY

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 encourage 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 Institute 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

	Page
List of Figures	xvii
List of Tables	xix
Glossary	xxi
Chapter 1: Introduction	1
1.1 Meaning in Context	2
1.2 Representing Meaning	2
1.3 Problem Statement	3
1.4 Proposing Approach	10
1.5 Thesis Contribution	12
1.6 Thesis Outline	12
Chapter 2: Theoretical Background	15
2.1 Introduction	15
2.2 Semantic Representation	16
2.2.1 Feature Generation	16
2.2.2 Document-based Features	17
2.2.3 Word-based Features	19
2.2.4 Pattern-based features	20
2.2.5 Concept-based Features	20
Explicit Concept-based Features	21
Latent Concept-based Features	22
2.2.6 Learning-based Features	23
2.3 Semantic Distance	24

2.3.1	Evaluating Semantic Distance	25
2.4	Approaches to Semantic Distance	27
2.4.1	Knowledge-based Approaches	27
2.4.2	Content-based Approaches	35
2.5	Conclusion	41
Chapter 3:	Relational Feature Analysis For Semantic Interpretation	43
3.1	Introduction	43
3.2	Meaning Representation Using Relations in Context	46
3.2.1	Relation Extraction Algorithm	47
3.2.2	Confidence Function for Extracted Triples	49
3.2.3	Learning Parameters for the Confidence Function	51
3.2.4	The relational Semantic Space of Words	53
3.3	Meaning Representation Using N-Gram Context	54
3.3.1	Building the Word Semantic Space Using N-Gram Context	55
3.3.2	Weighting Filters	56
3.4	Inducing Hidden Concepts from Semantic Space	57
3.5	Experiment	59
3.5.1	Testing Benchmarks	59
3.5.2	Corpus	59
3.5.3	N-Gram Word Features Extraction	60
3.5.4	Relational Feature Extraction	61
3.5.5	Relational Feature Extraction with Confidence	62
3.5.6	Hidden Feature Extraction	63
3.5.7	Semantic Distance	64
3.6	Evaluation	65
3.6.1	Overall results compared to other content-based methods	65
3.6.2	Feature Generation Analysis	67
3.7	Conclusion	70
Chapter 4:	Conceptual Topic Analysis for Semantic Distance	73
4.1	Introduction	73
4.2	Topical Context Analysis	75

4.2.1	Explicit Topic Analysis	76
4.2.2	Extracting Seeds For Wikipedia Concepts	76
4.2.3	Latent Topic Analysis	80
4.3	A Combination Features for Word Meaning Representation	82
4.3.1	Context-based Features and Topical Seeds	82
4.3.2	Context-based Features and Topical Features	84
4.3.3	Evaluation of Feature Combinations	87
4.4	Experiments	88
4.4.1	Testing Benchmarks	88
4.4.2	Text Corpus	89
4.4.3	Extracting latent topics and seeding representatives	89
4.4.4	Extracting explicit topics and seeding representatives	90
4.4.5	Choosing Parameters for Feature Combination	91
4.5	Evaluation	92
4.5.1	Overall results compared to other topic-based methods	92
4.5.2	Latent Topic Features vs. Explicit Topic Features	95
4.5.3	Linear Combination vs. Seed-based Combination	96
4.6	Conclusion	96
Chapter 5:	Multi-way Feature Analysis For Semantic Interpretation	97
5.1	Introduction	98
5.2	Related Work	100
5.2.1	Word Meanings in Distributional Contexts	100
5.2.2	Creating the Word Semantic Space	102
5.2.3	Tensor Analysis	103
5.3	Tensor Analysis for Meaning Representation	104
5.3.1	Multi-way Arrays	104
5.3.2	Modelling the Semantic Space using a Three-way Tensor	106
Memory Tensor	107	
Capturing the Closeness of Triples	112	
Computing on Word Meaning	114	
A Measure of Syntagmatic Association	115	
A Measure of Paradigmatic Association	116	

	A Measure of Semantic Distance	117
5.3.3	Multi-way Latent Feature Analysis	118
	Parallel Factor Analysis	119
	Representing Word Meaning	120
	Measuring Semantic Distance	121
5.4	Experiment	122
5.4.1	Text Repository	123
5.4.2	Extracting Relations as Triples	123
5.4.3	Encoding Triples into Three-way Tensor	124
5.4.4	Measuring Semantic Distance based on Syntagmatic and Paradigmatic Strengths	125
5.4.5	Multi-way Latent Feature Analysis	126
5.5	Evaluation and Discussion	127
5.5.1	Overall Results	127
5.5.2	Syntagmatic Strength vs Paradigmatic Strength	129
5.5.3	Further remarks	131
5.6	Conclusion	132
Chapter 6:	Conclusion	135
6.1	Summaries of the Proposed Approaches	135
6.2	Overall Results from Measuring Semantic Distance	138
6.3	Addressing Research Questions	139
6.3.1	What strategies can be used to utilise relations between words for semantic interpretation?	139
6.3.2	What are the effects of topical analysis for word meaning representation?	140
6.3.3	What are the strategies to use words in contexts and structural information on the task of semantic interpretation?	141
6.3.4	What strategies can be used to improve the performance of content-based methods on the task of measuring semantic distance?	142
6.4	Future Investigations	142
6.5	Future Applications	143
6.6	Conclusion	146

Bibliography 147

LIST OF FIGURES

Figure Number	Page	
4.1	Details of the explicit topic \times seed matrix $M_{exp}[topic, seed]$. Each row of the matrix is a vector of seeds, where only K -best components are selected while the rests are set to zeros	79
4.2	Details of the latent topic \times seed matrix $M_{lat}[topic, seed]$. Each row of the matrix is a representation of a latent topics, which is constituted by a vector of seeds. Only the values of K best seeds are selected, and the rest are set to zeros.	81
5.1	The cube represents a model of a third-order (three-way) tensor: $\mathcal{X} \in \mathbb{R}^{I \times J \times K}$. I, J, and K are the number of dimensions for each mode (way) (Kolda & Bader 2009)	104
5.2	Slices of a third-order tensor. Horizontal slices, lateral, frontal slices are obtained by fixing the first, second, and third model of the tensor respectively.	105
5.3	Fibers of a third-order tensor. Column, Row, or Fiber columns are obtained by varying either the first, second, or the third mode of the tensor respectively and fixing the rest remaining modes (Kolda & Bader 2009)	106
5.4	A three-way tensor as a result of constructing three triples $T_1=(garlic, kill, bacteria)$, $T_2=(garlic, kill, fungi)$, $T_3=(garlic, kill, virus)$. The respective slice for each mode is also depicted when fixing the respective mode, which shows how the encoding values are assigned	113
5.5	Graphical representation of PARAFAC. A tensor \mathcal{X} is approximated into a combination of three loadings matrices: $A_{I \times Z}$, $B_{J \times Z}$, and $C_{K \times Z}$	119
5.6	Procedures to conduct an experiment for the first Tensor model . . .	122
5.7	Procedures to conduct an experiment for the second Tensor model .	125
5.8	Correlation results (%) on each datasets have been presented when changing the trade-off parameter α	130

LIST OF TABLES

Table Number	Page	
2.1	A brief comparison features used for semantic representation	18
3.1	List of features used to identify whether a triple is reliable or unreliable. These features was inherited from the work of Fader et al. (2011) in Open Information Extraction.	50
3.2	Experiment using MTurk for turning parameters. The best Spearman’s correlation score (%) was obtained with $FF = 2$, $IVF = 1$ on NGram word features, and with $r = 0.7$ on the confident threshold. Results of the related works on the same dataset were also presented.	60
3.3	List of features using their weightings updated from the Logistic Regression algorithm on training examples	62
3.4	The experimental results (%) with different content-based methods on WS-353 datasets using Spearman’s rank correlation (ρ). In the first row, results from the related work are presented. The results of our proposed features are showed in two last rows of the table.	65
3.5	Experimental results (%) from various content-based methods on RG-65 datasets using Spearman’s rank correlation (ρ). The first section shows the results from the related works. The results of our proposed features are also demonstrated in two last sections of the table	66
3.6	Correlation results (%) for NGram word features and relational word features when applying a frequency filter ($FF=2$) and information value filter (IVF) with the semantic distance tested on WS-353 dataset using Spearman’s rank correlation (ρ). The best value is bolded and its corresponding parameter value is underlined	68
3.7	Correlation results (%) from relational features when applying different thresholds r on the confidence value for each triple with the semantic distance tested on the WS-353 dataset using Spearman’s rank correlation (ρ). The best value bolded and its corresponding parameter value is underlined.	69

4.1	Experiment on MTruk for tuning parameters. Four different feature combinations that require parameter turning were experimented. For each feature combination, we selected values of parameters that returned the best correlation results during the semantic distance measurements	88
4.2	Comparison results with different content-based methods on WS-353 datasets using Spearman’s rank correlation (ρ). [†] indicates the results using the selected parameters from the independent dataset	93
4.3	Comparison results with different content-based methods on RG-65 datasets using Spearman’s rank correlation (ρ).	95
5.1	An example of indexed and initialized terms from the sentence “ <i>Garlic can also kill harmful bacteria, fungi, and viruses</i> ”	107
5.2	Experiment on MTurk for turning the α parameter. The best Spearman’s correlation score was obtained with $\alpha = 0.6$	126
5.3	Correlation results for different tensor-based features used on the semantic distance measure, and tested on the WS-353 dataset using Spearman’s rank correlation (ρ). Symbol [‡] indicates value of the α parameter selected from a independent test. Symbol \star indicates the best results over a range of selected parameters. The result from the LSA baseline method is also presented.	128
5.4	Correlation results for different tensor-based features used on the semantic distance measure, and tested on the RG-65 dataset using Spearman’s rank correlation (ρ). Symbol [‡] indicates value of the α parameter selected from a independent test. Symbol \star indicates the best results over a range of selected parameters. The result from the LSA baseline method is also presented.	129
6.1	A summary of the best results of our proposed features tested on the WS-353 dataset for the task of semantic distance measurement.	138
6.2	A summary of the best results of our proposed features tested on the RG-65 dataset for the task of semantic distance measurement.	139

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