A NOVEL ASPECT TAXONOMY AND ASPECT EXTRACTION METHODOLOGY FOR SCHOLARLY BOOK REVIEWS

Wickramarathna Wengappuli Arachchige Chathur Sajeewan Basuru

(158207C)

Degree of Master of Science

Department of Computer Science & Engineering

University of Moratuwa

Sri Lanka

March 2019

A NOVEL ASPECT TAXONOMY AND ASPECT EXTRACTION METHODOLOGY FOR SCHOLARLY BOOK REVIEWS

Wickramarathna Wengappuli Arachchige Chathur Sajeewan Basuru

(158207C)

Thesis/Dissertation submitted in partial fulfillment of the requirements for the degree
Master of Science in Computer Science and Engineering

Department of Computer Science & Engineering

University of Moratuwa

Sri Lanka

March 2019

DECLARATION

“I declare that this is my own work and this dissertation does not incorporate without acknowledgement any material previously submitted for a Degree or Diploma in any other University or institute of higher learning and to the best of my knowledge and belief, it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Also, I hereby grant to University of Moratuwa the non-exclusive right to reproduce and distribute my dissertation, in whole or in part in print, electronic or other medium. I retain the right to use this content in whole or part in future works (such as articles or books).

Signature: Date:

Name: W.W.A.C.S. Basuru

The above candidate has carried out research for the master’s dissertation under my supervision.

Signature of the supervisor: Date:

Name of the supervisor: Dr. Surangika Ranathunga

Abstract

Many people decide on the quality of a product based on its online reviews, which is also the most commonly used method when purchasing books from online book stores. Compared to other products, a scholarly book is one of the most difficult products to purchase online since customers have limited access to its internal content. Therefore, a customer has to go through multiple reviews in order to get insight on the book. However, the sheer volume of online reviews makes it difficult for a human to process and extract all the meaningful information in order to make an educated purchase. As a result, a requirement for a sentiment analysis system for scholarly book reviews are much needed at this stage. A more accurate opinion of the book can be obtained through aspect-based summarization. This type of summarization of opinions is critical for scholarly book reviews since content, organization, and other features interpret whether the book can be recommended to a customer at a certain education level.

Compared to sentiment analysis on reviews of products/services such as movies or restaurants, there is no well-defined research in aspect extraction or aspect-based sentiment analysis of scholarly book reviews. Not surprisingly for this domain, there is no well-defined aspect taxonomy or an annotated dataset available to extract aspects or to identify aspect categories. Compared to other domains, identifying aspects of book reviews is difficult since aspects such as the quality of the book or the discussed topics always appear implicitly in reviews.

The main contribution of this research is to identify potential aspects and an aspect taxonomy for scholarly book reviews. We also present a (1.) dependency rule-based unsupervised model for aspect extraction, which works better than state-of-the-art unsupervised methods, and (2.) a clustering-based aspect category identification method. Both of these are important first steps for aspect-based sentiment analysis.

The aspect taxonomy for scholarly book reviews is a hierarchical model. Book and Author have been identified as the first level of the taxonomy. Readability, content, worthiness and price, are the next level of aspect taxonomy under the book aspect category. Author expertise has been identified as an aspect category under author. In order to validate the aspect taxonomy, an unsupervised aspect extraction and clustering algorithm is proposed. An existing dependency rule-based aspect extraction algorithm is improved by adding new rules that extract aspects from book reviews. Two existing clustering algorithms for aspect clustering are merged to obtain a new clustering algorithm to discover the categories of aspect terms. The clustering algorithm is able to find the semantic similarity of aspect terms, while considering the sharing words between aspect terms, and groups similar aspects in to a one cluster. After successfully generating an annotated corpus for the scholarly book reviews in the computer science domain with Cohen's kappa statistics of 0.76, the dependency rule-based aspect extractor was able to extract both implicit and explicit aspects with precision 76.04%, recall 75.99% and overall F1-score 76.02%. The proposed semantic similarity based aspect clustering algorithm identifies the aspect in the following categories; book, author, readability, content, worthiness, price and author expertise with rand-index 14.41%, V-measure 36.29%, homogeneity 66.18% and completeness 25%.

Keywords: Aspect based sentiment analysis, Dependency rules, Aspect taxonomy, Clustering, Semantic similarity, Stanford dependency parser, GloVe

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to my supervisors Dr. Surangika Ranathunga for the continuous support given for the success of this research. This would not have been a success without your tremendous mentorship and advice from the beginning.

I would like to thank all staff from the Department of Computer Science and Engineering for their kindness expressed in all occasions. Special thanks to my loving family: my parents, wife, and sisters. Your encouragement always motivated me to do my best. I would also like to thank all my friends who encouraged me all times.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ALGORITHMS

LIST OF ABBREVIATIONS

ABSA	Aspect Based Sentiment Analysis
CINAHAL	Cumulative Index to Nursing and Allied Health Literature
CNN	Convolutional Neural Network
CRF	Conditional Random Field
GFST	Generalized Aspect Sentiment Tree
HMM	Hidden Markov Model
IAC	Implicit Aspect Clue
KNN	K-Nearest Neighbors
LDA	Latent Dirichlet Allocation
MEDLINE	Medical Literature Analysis and Retrieval System Online
MLE	Maximum Likelihood Estimation
MV-RNN	Matrix Vector Recurrent Neural Network
NER	Named Entity Recognition
NLP	Natural Language Processing
NPMI	Normalized Pointwise Mutual Information
OWA	Ordered Weighted Averaging
PAS	Positional Aggregation based Scoring
POS	Part of Speech
RNN	Recurrent Neural Network
RNTN	Recursive Neural Tensor Network

SDM	Sequential Dependence Model
SemEval	International Workshop on Semantic Evaluation
SSWE	Sentiment Specific Word Embedding