SEMI-SUPERVISED SUBJECTIVITY CLASSIFICATION AND APPLICATION TO DOMAIN SPECIFIC LANGUAGES

\

Committee Signature 1

Committee Signature 2

Committee Signature 3

SEMI-SUPERVISED SUBJECTIVITY CLASSIFICATION AND APPLICATION TO DOMAIN SPECIFIC LANGUAGES

by

JASON MICHAEL SWITZER, B.S.

THESIS

Presented to the Faculty of

The University of Texas at Dallas

in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE IN COMPUTER SCIENCE

THE UNIVERSITY OF TEXAS AT DALLAS

November 2010

ACKNOWLEDGEMENTS

I would like to thank everyone who has helped me in my journey. I would like to thank God for giving me the ability, passion, and opportunity to achieve. I would like to thank all of my instructors, professors and classmates who have pushed, taught, and assisted me when it was needed most. I would like to thank my family who has believed in me over the years. Lastly, I would like to thank my wife for having patience, understanding, love, and encouragement these many years. Without her, I would have been lost.

SEMI-SUPERVISED SUBJECTIVITY CLASSIFICATION AND APPLICATION TO DOMAIN SPECIFIC LANGUAGES

Jason Michael Switzer

The University of Texas at Dallas, 2010

Supervising Professor: Dr. Latifur Khan

Semantic analysis of a corpus consisting mostly of domain specific words and phrases introduces problems not addressed by most corpuses. Modern semantic analysis relies heavily on data from the web, such as blogs, or heavily edited sources, such as the New York Times. These corpuses lack words and phrases that are specific to a certain domain or topic. This paper will present techniques that can be used to train a semantic model towards a corpus consisting of domain specific language. Specifically, this paper will address the subjectivity identification of words and phrases and their presence within the NASA flight log corpus, which draws heavily on phrases and jargon used by pilots. We will do this by creating a pipelined architecture of semi-supervised estimators based on manually labeled clustered datasets, such as thesauruses. Then, this paper will show that even a small set of manually labeled data can greatly improve the performance of all subsequent estimators. This paper will show that each node in the hypothesis pipeline can be boosted to further improve performance. Lastly, this paper will discuss the findings within the NASA flight log corpus and how such findings can improve semantic analysis.

TABLE OF CONTENTS

List of Figures

List of Tables

CHAPTER 1

INTRODUCTION

FIXME

1.1 The Approach

There is a vast amount of subjective information available on the Internet and within data set created by a number of organizations. Since English is a rapidly expanding language, automatic discovery of subjective terms is highly desirable. This thesis presents a novel approach of combining several semi-supervised learning algorithms to train a large and accurate lexicon. A small general purpose lexicon based solely on manually classified examples is insufficient when applied to corpus that uses a domain specific languages.

Subjectivity classification is determined by a number of Subjectivity Lexicon Learners (SLL). Each SLL operates independently of every other SLL and may be either semi-supervised or fully supervised. Each SLL will classify and discover instances differently. For example the Moby Seed Lexicon will use the manually labelled synonym sets (or synonym clusters), whereas the WordNet algorithm will explore WordNet synsets to discover and label instances. All semi-supervised learning algorithms take advantage of the unified lexicon format and classify the existing and new instances based on the SLL who's output is fed as the new input. This pipeline

concept can be visualized as a Bayesian network of SLL algorithms, though it will not be shown that this is mathematically sound.

1.2 Experimental Context

The related research has shown that thesaurus based classification can achieve a higher level of accuracy when compared to the authoritative General Inquirer lexicon. This learning algorithm is mixed with a variety of other semi-supervised learners and shows certain combinations have favorable results.

To test the semi-supervised learners, each possible combination is run to build a distinct lexicon file. Since the portions of the training data to the individual Subjectivity Lexicon Learners (SLL) is remains constant, each SLL can be trained independently. Each SLL is then retrained and Boosted to create a similar set of SLL output files. The lexicon file for a single combination is then compared to the General Inquirer and our manually created lexicon for accuracy. The result is 35 lexicons are created six times: once without boosting, once with AdaBoost applied for 3 iterations, once with AdaBoost applied for 10 iterations, once with InvBoost applied for 3 iterations, and once with InvBoost applied for 10 iterations. There are a total of 210 lexicons created and checked for accuracy against two reference lexicons.

1.3 Contributions

This thesis explores a variety semi-supervised learning algorithms for subjectivity classification. The main contribution lies with the pipelined combinations of the semi-supervised learning algorithms and the utility of the best performing lexicon against the domain specific terminology of the NASA flight log corpus. Another main contribution is the application of AdaBoost and a modified AdaBoost, known as InvBoost, as a stage of post processing to improve the performance of the individual SLL algorithms.

1.4 Thesis Outline

The rest of the thesis is organized in the following manner. Chapter 2 outlines the prior work that this thesis is based upon, emphasizing the overtly marked lexicon from dictionary and thesaurus data sources. Chapter 3 discusses the variety of training data used by the various semi-supervised learners. Chapter 4 explores boosting algorithms and how their application to the SLL algorithms. Chapter 5 introduces the full set of SLL algorithms and their strengths and weaknesses. Chapter 6 interprets the results from from the SLL combination experiments. Chapter 7 covers an overview of the NASA flight log corpus and the difficulties of semantic analysis. Chapter 8 explores the application of the subjectivity lexicons within the domain of the NASA flight log corpus.

Related Work

Manually Annotated Lexicons

Boosting

Pipeline Subjectivity Classification System

Domain Specific Languages

Results

Future Work

Bibliography

VITA

Jason Switzer was born in Austin, Texas on May 5, 1982, the son of Paul and Patricia Switzer. After graduating with Honors from Round Rock High School, Round Rock, Texas in 2000, he entered the University of Texas at San Antonio. In July 2003, he took a software development position at Secorp Technologies, where he worked full-time while pursuing his degree full-time as well. He received his Bachelor of Science in August 2004, majoring in Computer Science. In May 2005, he took a position as a Software Engineer at L-3 Communications working in the Special Systems group developing the state of the art Human-Computer Interface systems. In December 2010, he will receive his Masters of Science in Computer Science in the field of Intelligent Systems. In May 2011, he will become a father to his first born child.