PIPELINED SUBJECTIVITY CLASSIFICATION AND APPLICABILITY TO DOMAIN SPECIFIC LANGUAGES

\

Committee Signature 1

Committee Signature 2

Committee Signature 3

PIPELINED SUBJECTIVITY CLASSIFICATION AND APPLICABILITY TO DOMAIN SPECIFIC LANGUAGES

by

JASON MICHAEL SWITZER, B.S.

THESIS

Presented to the Faculty of

The University of Texas at Dallas

in Partial Fullfillment

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.

PIPELINED SUBJECTIVITY CLASSIFICATION AND APPLICABILITY 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 is 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

Introduction

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.