A NOVEL APPROACH TO MANAGE LSA'S SYTACTICAL BLINDNESS PROBLEM

Mohsin Hassan Khan

Enrollment No: 01-244151-039

Supervisor: Dr Raja M.Suleman

A thesis submitted to the Department of Software Engineering, Faculty of Engineering Sciences, Bahria University, Islamabad in the partial fulfillment for the requirements of a Master's degree in Software Engineering

March 2017

ABSTRACT

Natural language processing (NLP) is a computerized technique that is used for analyzing and representing human language automatically. NLP has been employed in many applications such as information retrieval, information processing, translations of language, automated answer grading and many more. The main problem with NLP is high level of uncertainty in natural language. High uncertainty in natural language makes automated analyses and extraction of useful information very difficult. Several approaches have been developed for automated grading. Latent Sematic Analysis (LSA) is one of the widely used approaches for automated text matching. LSA is a corpus based approach that evaluates similarity on the basis of semantic relations among words and ignores the structural composition of sentence. The structure blindness of LSA treats a logically wrong answer as a correct answer. LSA cannot recognize sentences that are semantically related but inverse of each other [8]. Furthermore, LSA cannot handle “gaming the system”, where user provides only the list of keywords without proper sentence structure.

The target of our research is to develop an algorithm Extended Latent Sematic Analysis (xLSA) which focuses on synthetic composition of a sentence and overcome LSA’s syntactic blindness problem. xLSA examine sentences and identifies that proper sentence structure exists to cater “gaming the system” problem. xLSA analyzes text inputs to recognize their dependency structure and then decompose each sentence to identify subject, verb and object. Sentences are then compared and an approximation of synthetic and semantic space is generated for similar texts. xLSA compute semantic similarity score of two sentences and also identifies inverse sentences, negative sentences and “gaming the system”.

We have tested xLSA with 200 semantically similar sentences from two corpuses [28] [29]. Results show xLSA outperforms then traditional LSA and identifies inverse sentences, negative sentence and list of keywords without having proper sentence structure.

DEDICATION

This thesis is dedicated to my beloved parents, Maqbool Hussain and Rashida Maqbool, for being role models for me and my brother Marghoob Ahmed and my sister Sobia Maqbool for their continuous support and encouragement regarding my goals.

ACKNOWLEDGMENTS

First of all, I thank Allah Almighty who endowed my potential and ability to complete this dissertation. I would like to extend my humble gratitude to my supervisor, Dr. Raja M.Suleman for offering his best possible support and guidance all the way through. It has been an honor to work under his adept supervision. I am grateful for his precious time, ideas and knowledge that made my research an unforgettable experience for me. Without his motivation and guidance it would have been impossible to remain firm in obscure situations. His enthusiasm towards research was motivational for me during tough times in my research. His perseverant and encouraging behavior always boosted my morale up

. Last, but not the least, I would like to thank my beloved parents and my other family members who practically freed me from all responsibilities and who constantly prayed for me throughout my academic career, that, in consequence, made better accomplishment of this dissertation possible. I am also grateful to my friends, especially Adil Arif, Sohail Ashraf, Irfan Muhammad Khan, Khurrum Mustafa Abbasi, Sofyan Aslam, and Rizwan Ghani for the concern, help and motivation regarding this research.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATION

HMM	Hidden Markov Model
NLP	Natural Language Processing
NLU	Natural Language Understanding
POS	Part of Speech
PEG	Project Essay Grade
AMS-SAE	Automatic Marking System for Short Answers Examination
AEE	Automated Essay Evaluation
GLSA	Generalize Latent Semantic Analysis
HAL	Hyperspace Analogue to Language
PMI-IR	Pointwise Mutual Information - Information Retrieval
SCO-PMI	Second-order co-occurrence pointwise mutual information
LSA	Latent Semantic Analysis
SVD	Singular Value Decomposition
STS	Semantic Text Similarity
MRLSA	Multi-Relational Latent Semantic Analysis
PILSA	Polarity Induced Latent Semantic Analysis
SELSA	Syntactically Enhanced LSA
ASAS	Automated Short Answer Scoring
AES	Automated Essay Scoring
PEG	Project Essay Grade
ATM	Automated Text Marker
xLSA	Extended Latent Semantic Analysis