UNIVERSITI TEKNOLOGI MARA

ENHANCEMENT OF COMPOUND WORD EXTRACTION IN MALAY SENTENCES USING MODIFIED LINGUISTICS APPROACHES

ZAMRI BIN ABU BAKAR

Thesis submitted in fulfilment
of the requirements for the degree of
Doctor of Philosophy
(Computer Science)

College of Computing, Informatics and Mathematics

September 2023

ABSTRACT

Malay compound word is defined as a form of words that exists when two or more words are combined into a single syntax, and it gives a specific meaning. Thus, this extraction of compound words is significant for the following research, which is text summarization, grammar checker, sentiments analysis and machine translation. The aim of this study is to propose a new extraction technique using linguistic approaches that combines many features and rules. There are many research efforts that have been proposed in extracting compound word using linguistic approaches. However, the result for this approach still produces some problems in giving a better result. Overall, this study has three objectives; to identify new rules in detecting the Malay compound word, to construct an improved compound word extraction technique (algorithm) that combines many rules for Malay sentences using linguistic approaches, and lastly to evaluate the accuracy of proposed technique from using the standard evaluation of Recall, Precious and F-Measure. To achieve the objective, this research explores a linguistic method for extracting compound word from standard Malay corpus. A Malay news dataset was used to extract compound word in this research. Therefore, an improvement for the effectiveness of the compound word extraction is needed because the result can be compromised. Thus, this study proposed a modification of linguistic approach to enhance the extraction of compound word processing. Several pre-processing steps were involved which include normalization, tokenization, stemming and tagging. Finally, this study described several rules-based and modified the rules to get the most relevant relation between the first word and the second word in order to assist this study in solving the problems. The result showed that the proposed technique outperforms the previous techniques on all measures, such as Precision, Recall and F-Measure with 1000, 2000, 3000, 5000, 10000, 15000, 30000, 60000, 80000, 10000, 15000, 30000, 60000, 80000 and 100000 sentences used in the experiment. Precision, Recall and F-Measure respectively had an average value of 88.23, 49.13 and 68.7, compared to the previous technique of Suhaimi's Noun Phrase Frame Structure result which respectively had an average value of 85.71, 46.46 and 66.1. As a conclusion, the new Enhancement of Compound Word Extraction Using Modified Linguistic Approaches has the potential to improve the results of Malay Natural Language Processing.

ACKNOWLEDGEMENT

First and foremost, I wish to praise and thank God, the Almighty for giving me the opportunity and blessing throughout my research work to embark on my PhD successfully even it was a very long journey and too challenging to complete the process of writing. Secondly, I would like to express my deep and sincere gratitude to my main research supervisor, Associate Professor Dr. Normaly Kamal Ismail for the continuous support of my PhD study. He always gives me a lot of motivation and keeps encouraging without feeling tired and weary along my PhD journey. Besides my main supervisor, my sincere thanks also go to Professor Dr. Zainab Abu Bakar as my first main supervisor before she retired in the first year of my PhD journey. I also want to say a thousand thanks to Dr. Mohd Izani Muhamed Rawi for his willingness to be my second supervisor who gave me a lot of guidance and sharing with me the way to succeed in PhD learning. Without their guidance, inspiring suggestions and perseverance, I may not be able to complete this thesis according to the schedule plan.

Primarily, I would like to express my sincere gratitude to my main supervisor Associate Professor Dr. Normaly Kamal Ismail for his patience, motivation, immense knowledge, continuous encouragement and giving too much passion during my PhD study. I could not have imagined having a better supervisor for my PhD study.

My appreciation goes to all my colleagues who sat together with me in the faculty lab research assistant. They were always giving me reminders and support to complete this thesis as soon as possible, and not forgetting to all faculty staff who provided the facilities and place in order for me complete this thesis. Special thanks to all my friends and colleagues who kept on helping me during the process of completing this research and along my PhD study.

Finally, this thesis is dedicated to the loving memory of my dearest wife, my late father and also to my lovely mother for the vision, goal and determination to educate me. This piece of victory is dedicated to the three of you. Alhamdulillah.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

1.1 Introduction

Everyone in the world realizes the importance of language usage for daily activities particularly in making interactions among their friends and family. Today, language has become one of the important mediums to ensure that people can communicate with the community. If we count the actual number of natural languages existing in the whole world, they are countless. According to Dewan Bahasa dan Pustaka (DBP) in Malaysia, it is estimated that there are about 3000 to 8000 languages spoken in the world. Natural Language is used by people for completing their learning and teaching process. Other than conveying information, communicating and presenting the material to the public also use the natural language. Referring to Ethnologue (2022), at present, the spoken language used by human beings in the world is about 7139 speakers. However, Mandarin, English, Hindi, Spanish, Arabic, Russian, Malay, Portuguese, Bengali and French are the ten biggest speaking languages in regard to the total number of speakers. The Malay language is 7th ranked with approximately 259 million speakers. The rough estimation for the total of Malay language speakers in 2009 was based on Indonesian citizens which reached 230 million and Malaysian citizens with 28 million. Another 338 million of Brunei citizens and small portions of Thailand, Singapore and Timor Timur also contributed to the estimation (Ethnologue, 2022). Among those biggest speaking languages, the Malay language has been chosen as the main topic or focus of this research. It is because according to Zuraidah (2010), research in Malay language documents is still underdone and too few.

In Malaysia, Dewan Bahasa dan Pustaka (DBP) was established as a government body to coordinate and empower the Malay language as the national language and official language of the country. When it comes to the research of the Malay language, there are several topics that can be discussed. For example, grammar pattern, Named-entities recognition (NER), extraction of compound word, morphology, synthesis, semantics, ambiguity, ontology and etc. (Shengnan et al., 2022; Jurafsky and Martin, 2000). Several researchers (Hassan Mohamed et al., 2023; Abdullah Hassan,