ENHANCING THE NATURAL LANGUAGE PROCESSING FOR MALAY LANGUAGE: STEMMING, IDENTIFYING AND CORRECTING MISSPELLED WORDS, IDENTIFYING NEOLOGISM

SURAYAINI BINTI BASRI

FACULTY OF COMPUTING AND INFORMATICS
UNIVERITI MALAYSIA SABAH
2015

ENHANCING THE NATURAL LANGUAGE PROCESSING FOR MALAY LANGUAGE: STEMMING, IDENTIFYING AND CORRECTING MISSPELLED WORDS, IDENTIFYING NEOLOGISM

SURAYAINI BINTI BASRI

THESIS SUBMITTED IN FULLFILMENT OF THE DEGREE OF MASTER OF SCIENCE

FACULTY OF COMPUTING AND INFORMATICS
UNIVERSITI MALAYSIA SABAH
2015

DECLARATION

I hereby declare that the material in this thesis is my own except for quotations, excerpts, equations, summaries and references, which have been duly acknowledged.

25 August 2015

_____

Surayaini Binti Basri
PI20108028

CERTIFICATION

NAME : **SURAYAINI BINTI BASRI**

MATRIC NO. : **PK20108028**

TITLE : **ENHANCING THE NATURAL LANGUAGE PROCESSING FOR MALAY LANGUAGE: STEMMING, IDENTIFYING AND CORRECTING MISSPELLED WORDS, IDENTIFYING NEOLOGISM**

DEGREE : **MASTER OF SCIENCE (COMPUTER SCIENCE)**

VIVA DATE : **15 July 2015**

DECLARED BY

SUPERVISOR

Assoc. Prof. Dr. Rayner Alfred Signature

ACKNOWLEDGEMENT

In the name of Allah, the Most Gracious and the Most Merciful Alhamdulillah, all praises to Allah for the strengths and His blessing in completing this thesis. This research and thesis would not have been possible without the support of many individuals. I hereby would like to express my sincere gratitude to all those who have been directly or indirectly supported me in completing this research and thesis. Special appreciation goes to my supervisor, A.P Dr. Rayner Alfred for his guidance, persistent assistances, and valuable advices. His continuous guidance and support does give me strength to face and overcome difficulties that have been encountered during the period of my research.

I would also like to take this opportunity to acknowledge the contribution from Universiti Malaysia Sabah (UMS) for granting a scholarship to me. With this scholarship, I am able to finish my research study without worrying about the financial burden.

Besides that, I would like to express my heartfelt thanks to my beloved parents, Mr. Basri Durabi and Mrs.Salasia Anong, my beloved husband, Mr. Mohd Nurfaizal Bin Jinin @ Awang Damit, my son Azeem Azfar and all my relatives for their moral and spiritual supports along the way to completing this research and thesis.

Last but not least, I would like to express my thankful to have very supportive friends especially to Miss Florence Sia Fui Sze, Hana Helena Appolonius, Gan Kim Soon, and Leow Ching Leong for their advices, supports and understandings.

Surayaini Binti Basri
25 February 2015

ABSTRACT

Natural Language Processing (NLP) is a field of computer science, artificial intelligence, and linguistics concerned with the interactions between computers and human (natural languages). A good NLP approach is needed because the applications of NLP are used across a wide variety of industries in order to solve critical knowledge problems, such as providing new insights gleaned from massive collection of unstructured content (social media, news, patent filings, financial disclosures, etc.). A weak NLP for a language can cause in irrelevant information being retrieved. The lack of works in building more effective algorithms in performing the stemming process, identifying misspelled words, and identifying neologism has affected the efficiency of retrieving relevant information or articles in Malay language. This is due to the fact that the Malay language is a language that has different and complex morphology structure than other languages and thus, the standard NLP approach used in other languages cannot be easily applied in processing and retrieving relevant information or articles in Malay Language. This work focuses on improving the Malay language stemming process, introducing a new approach in identifying and correcting typo or misspelled words and lastly proposing solution to identify neologism. By improving the Malay stemming process, it will enable the information retrieval process to be performed with more effectively by identifying more affixed word in Malay language because not all affixed words are stored in the standard Malay dictionary. By identifying and correcting typo or misspelled words, it can also prevent the information retrieval system from ignoring several important words just because the words are misspelled. Finally, by identifying neologism, one may assist lexicographer to identify new words that can be considered as part of the lexicon dictionary. Based on the experiments conducted, the proposed approaches are proven to be useful in improving the NLP in Malay language.

ABSTRAK

MENINGKATKAN PEMPROSESAN BAHASA ASLI UNTUK BAHASA MELAYU:STEMMING, MENGENALPASTI DAN MEMBETULKAN PERKATAAN KESILAPAN EJAAN, MENGENALPASTI KATA BARU

Pemprosesan Bahasa Asli (NLP) adalah satu bidang sains komputer, kecerdasan buatan, dan linguistik berkenaan dengan interaksi antara komputer dan manusia (bahasa semula jadi). Pendekatan NLP yang baik diperlukan kerana aplikasi NLP banyak digunakan merentasi pelbagai industri yang luas untuk menyelesaikan masalah pengetahuan kritikal, seperti menyediakan wawasan baru yang dikumpulkan dari koleksi kandungan tidak berstruktur (media sosial, berita, pemfailan paten, pendedahan kewangan , dan lain-lain) yang besar. NLP yang lemah untuk sesuatu bahasa boleh menyebabkan maklumat yang tidak relevan diperolehi. Kekurangan hasil karya dalam membina algoritma yang lebih berkesan dalam melaksanakan proses yang berpunca, mengenal pasti perkataan silap eja, dan mengenal pasti kata baru telah menjejaskan kecekapan pencarian maklumat untuk artikel Melayu. Ini adalah disebabkan oleh hakikat bahawa bahasa Melayu adalah bahasa yang mempunyai struktur morfologi yang berbeza dan kompleks daripada bahasa lain dengan itu pendekatan NLP standard yang digunakan dalam bahasa lain tidak boleh dengan mudah digunakan untuk Bahasa Melayu. Dalam karya ini, peningkatan dalam NLP untuk bahasa Melayu memberi tumpuan kepada meningkatkan bahasa Melayu yang berpunca, memperkenalkan pendekatan baru untuk mengenal pasti dan membetulkan kesilapan menaip atau kesilapan ejaan dan akhir sekali mencadangkan penyelesaian untuk mengenal pasti kata baru. Dengan meningkatkan proses berpunca Melayu, ia akan membolehkan proses IR untuk mengenal pasti sama ada perkataan berimbuhan adalah perkataan Melayu atau bukan kerana tidak semua perkataan berimbuhan disimpan di dalam kamus. Mengenal pasti dan membetulkan kesilapan menaip atau perkataan kesilapan ejaan pula boleh mencegah dari pengabaian perkataan yang penting oleh IR hanya kerana perkataan itu tersilap ejaan. Mengenal pasti kata baru juga boleh membantu ahli kamus untuk mengenal pasti perkataan baru yang perlu dipertimbangkan sebagai sebahagian daripada leksikon. Berdasarkan eksperimen yang dijalankan, pendekatan ini terbukti boleh digunakan untuk meningkatkan NLP dalam bahasa Melayu.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

NLP	Natural Language Processing
IR	Information Retrieval
LD	Levenshtein Distance
IE	Information Extraction
AI	Artificial Intelligent
POS	Part Of Speech
RAO	Rules Application Order
RAO2	Rules Application Order 2
NRAO	New Rules Application Order
RFO	Rules Frequency Order
FSM	Finite State Machine
IDF	Inverse Document Frequency
SVM	Support Vector Machine
REGEX	Regular Expression
MSMW	Malaysia Social Media Week
SMC	Social Media Chamber
NER	Name Entity Recognition

LIST OF APPENDIX

LIST OF PUBLICATIONS

Basri, S., & Alfred, R. 2012. Automatic Spell Checker for Malay Blog. 2012 IEEE *International Conference on Control System, Computing and Engineering. Penang Malaysia*.

Ching Leong, L., Basri, S., & Alfred, R. 2012. Enhancing Malay Stemming Algorithm with Background Knowledge. *Pacific Rim International Conference on Artificial Intelligence. Kuching: Springer*.

CHAPTER 1

INTRODUCTION

1.1 Introduction

Natural Language Processing (NLP) is a subfield of Artificial Intelligence and linguistic, devoted to make computers understand the statements or words written in human languages, (Chopra, Prashar and Sain, 2013). In other words, NLP is a process that assists computer to understand or derive meaningful information from human language. Applications of NLP have been widely used in recent years. For instance, Information Retrieval (IR), Information Extraction (IE), Question-Answering, text mining and machine translation are some of the existing NLP appreciations.

The basic input for NLP applications is word. For a NLP system to be considered robust it must have mechanisms to process unknown words (Thede and Harper, 2010). The appearance of unknown word in NLP applications may cause the unknown word failed to be analysed because the lexicon does not have any information about the unknown word such as its part of speech. The unknown words are the words that appear in sentences, but are not contained within the lexicon for the system. This problem can be worst as the number of internet users who are spreading ideas through online social media and weblog are increasing as people will often use words that a NLP system may not expect. Although there is no exact number of weblog from Malaysia obtained but (Ulicny, 2008) has proved that Malaysian also constitutes a large number of blogs in the blogs statistics. (Colette, 2006) estimated 46% of those online in Malaysia have a blog.

Stemming is a pre-processing step in text mining applications as well as a very common requirement of NLP functions. The main purpose of stemming is to reduce different grammatical forms or word forms of a word like its noun, adjective, verb, adverb and etc. to its root form. In other words, the goal of stemming is to reduce inflectional

forms and sometimes derivationally related forms of a word to a common base form, (Jivani and Ganesh, 2011). Since stemming process can derive a word's root word, thus stemming can be used to identify either a word is known or not. If an unknown word has root, then it is identified as known. Otherwise, it is identified as unknown word. This is because not all affixed words are stored in lexicon. This process may help the NLP system to identify affixed word that has root word in lexicon. Existing Malay stemmer algorithm such as (Abdullah, Ahmad, Mahmod, and T.Sembok, 2009), (Fadzli,Norsalehen, Syarilla, Hasni and M Satar, 2012), (N. Idris and Mustapha, 2001)

Another NLP approach that is used to identify unknown word is a spell checker. A spell checker is a system which is used to detect and correct a misspelled word. A common spell checker works at word level and use a dictionary. Most of the spell checkers share this common process. That is, every word tokenized from the input text is looked up in the speller lexicon. When a word is not listed in the dictionary, it is detected as an error. In order to correct the error, a spell checker searches in the dictionary for alternative words that resemble the erroneous word most. These words are then suggested to the user who chooses the word that was intended. A spell checker is used in various applications like machine translation, search, information retrieval and etc. (Gupta and Mathur, 2012). By correcting the misspelled word or typo word, this process can help NLP to reduce the number of unknown words.

Neologism or newly-coined words, pose problems for NLP systems. Due to the recency of their coinage, neologisms are typically not listed in computational lexicons. Computational lexicon is a dictionary-like resource that many NLP applications depend on. Therefore when a neologism is encountered in a text being processed, the performance of an NLP system will likely suffer due to the missing word-level information (Cook, 2010). Identifying and documenting the usage of neologisms are also needed by lexicographers so they can analyse the unknown word to be included into lexicon or not.

Besides that, the standard NLP approach used in other languages cannot be easily applied in NLP for Malay language because Malay language is a language that has

different and complex morphology structure compared to other languages, (N. Idris and Mustapha, 2001).

In this thesis, an approach to enhance NLP for Malay language is proposed. The aim of this research is to identify Malay words as many as possible so that the NLP processes for Malay language can be performed with effectively and thus a more robust NLP application result can be provided. In order to do that, this work focuses on improving the Malay language stemming process, introducing a new approach to identify and correct any typo or misspelled words found in Malay language and proposing solution to identify neologism, a newly coined word.

1.2 Problem Statement

The information can be obtained from weblogs is not a bit. Weblogs are used in education, politics, entertainment, and etc. This is because weblogs can be seen as a lightweight, cost-efficient and flexible Internet medium, blog attracts growing interests and attention people all over the world (Deng and Yuen , 2007) to use it as a way of expressing their ideas and thoughts. Thus, information retrieval from this kind of medium is a must. Unfortunately, (Janssen, 2007) has mention that, even in thoroughly corrected corpora like newspaper, typo-graphic error does occur. Since the weblog is written by personal individual, the word written in the blog can be informal language. This fact shows that the possibility of unknown words occur in weblog is higher since weblog is usually written by personal individual that tend to write whatever they feel or they do in their blog without any restriction as formal language thus can increase the number of unknown words. For that reason, there is a need to propose a way to identify the unknown words from this kind of medium to assist NLP application become robust.

For NLP applications to be robust, it should be able to process all words contained in any kind of input text (Bali, Chua and Ng, 2007). Thus, the existence of unknown words such as affixed Malay words that do not exist in Malay lexicon, misspelled words or typo words and neologism words may reduce the robustness of NLP applica-

tions. This is because the system does not have any information about the unknown word because it does not exist in the lexicon.

The lack of researches conducted in identifying unknown words, misspelled words and neologism words for Malay language may affect the performance of information retrieval for Malay language text. A research of state of the art Asian language was done by (Huang, Tokunaga and M. Lee, 2006). The research mentioned that, the major language in Asian such as Mandarin Chinese, Hindi, Japanese, Korean, and Thai have benefited from several years of intense language processing research. Also the fast-developing languages (e.g., Filipino, Urdu, and Vietnamese) are gaining ground. However, for many near extinct languages, research and resources are scarce, and computerization represents the last resort for preservation after extinction. This research clearly shows that, research in Malay language still lack in terms of language processing since it is not even mentioned in the research.

The morphology structure for Malay language also differs from another language causes the existing approaches in identifying these words from other languages cannot be directly applied into Malay language. The related research by (Bali et al., 2007) that processing unknown words in Malay language such as abbreviation, proper name, loan word and affixed word also cannot be adapted in identifying unknown words from informal medium such as weblogs because most of the weblogs are written in informal Malay text.

Although there are several researches have been done in Malay stemming algorithm such as (Abdullah et al., 2009), (N. Idris and Mustapha 2001) and etc., however no one is a complete stemmer. This means that, there is no stemmer that is really free from error. Hence, there are still some improvements that are required in order to assist the NLP in Malay language. Thus, a new approach is required in order to identify these types of words in order to enhance the NLP Malay language.

1.3 Research Roadmap

This research introduces a methodology that enhances the processes involved in processing the Malay natural language. The enhancement involves enhancing the Malay stemming performance, proposing a new misspelled and typo words identification and correction algorithm and finally implementing an algorithm that identifies neologism in Malay language. The road map of this methodology is illustrated in Figure 1.1 and it is described as follows:

a. First, a feasibility study is conducted in order to investigate the nature of stemming, spell checker, and neologism identification processes in Malay language.

b. Then, a data collection is achieved by extracting the online Malay blogs and online Malay news.

c. Based on the feasibility study conducted earlier, the three main algorithms are designed and implemented and these algorithms are:
   i. Enhanced Malay stemming using background knowledge
   ii. Malay misspelled words identification and correction
   iii. Neologism identification for Malay language

d. Before any experiments can be conducted, several pre-processing tasks need to be developed and performed in order to clean the dataset. This includes the process of transforming the text into words list, tokenization, symbol elimination, and English words elimination.

e. The experimental setup is then designed so that a comprehensive analysis can be conducted for the enhanced Malay stemming first, then the Malay misspelled words identification and correction algorithm and finally, the Malay neologism words identification algorithm.

f. For each experiment, collect the results obtained. Make a comparison between the initial RFO stemming and the RFO stemming with background knowledge for the enhance Malay stemming algorithm. Calculate the errors produced from the identified and corrected misspelled word for Malay misspelled words identifica-

tion and correction algorithm. For the Malay neologism identification, calculate the number of neologism identified from the experiment. Then, make a discussion for each of the experiment based on the results obtained.

g. Finally the findings and conclusion are presented to conclude this work.

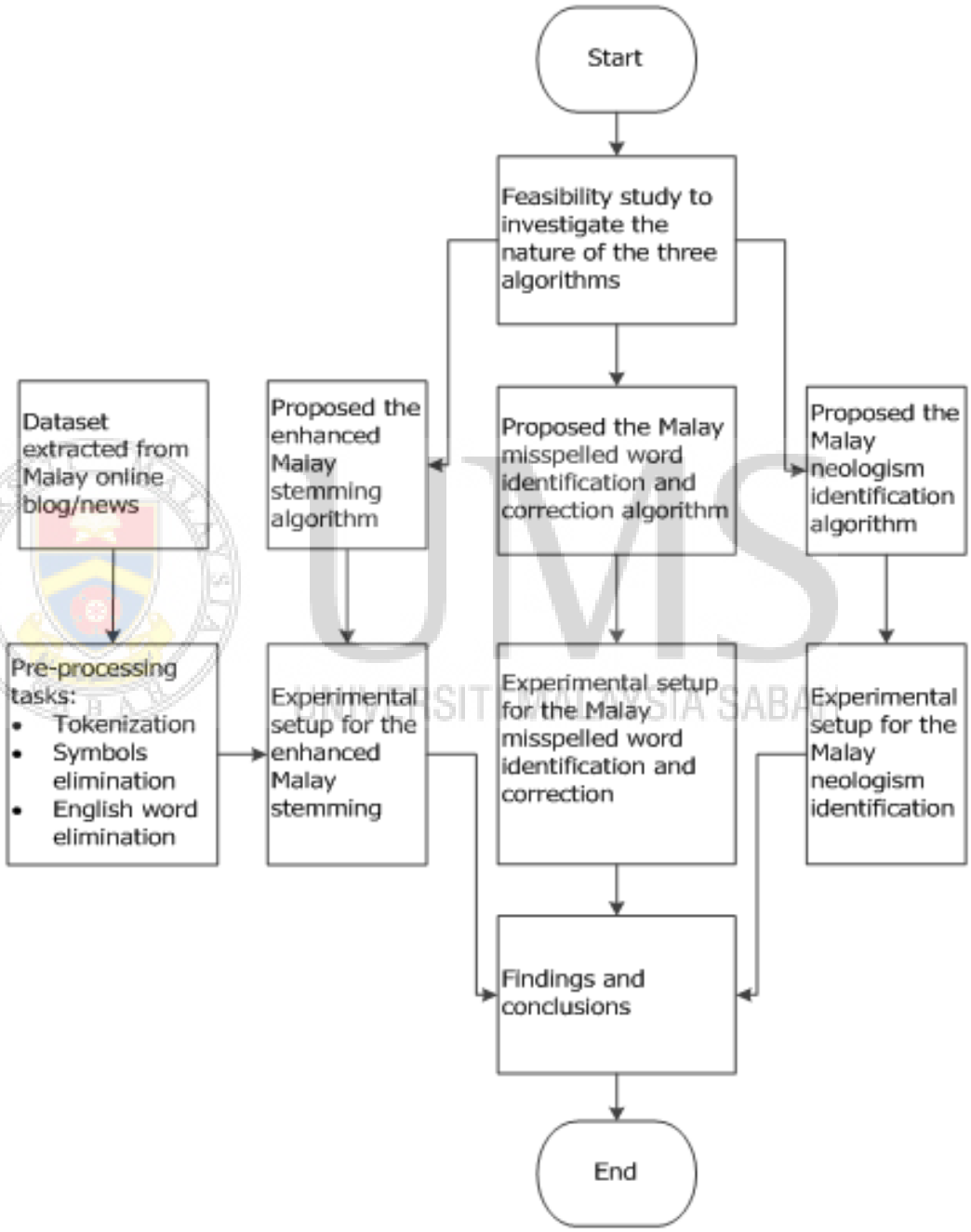

Figure 1.1: Research Roadmap Flow.

1.4 Research Objectives

As mentioned previously, the aim of this work is to enhance the NLP in Malay language. As a result, the objectives of this research are:

a. To enhance Malay stemming algorithm by adding background knowledge.

b. To propose an automatic spell checker for Malay language blog methodology.

c. To develop and assess a neologism identification methodology for Malay language.

1.5 Research Scopes

The scopes of this research are:

a. In evaluating the performance accuracy of the proposed methodology, a total of 5319 words which are obtained from the online news articles were used to evaluate the first methodology which is the methodology to enhance stemming in Malay. The text extracted from the Malay weblogs only used to evaluate the second and third methodology. These Malay weblogs are listed as the finalist in the Malaysia Social Media Week, MSMW (Secretariat Malaysia Social Media, 2013) and (Secretariat Malaysia Social Media, 2014).

b. The first methodology, which is Enhancing the Malay stemmer algorithm by adding background knowledge, is a stemming algorithm that enhances the RFO stemming algorithm introduced by (Abdullah et al., 2009) only.

c. In the second methodology, which is identification and correction misspelled words in Malay language, there are a function called Identifying Selangor Slang words. This function only focus on identifying Malay words that are written in Selangor slang because the Selangor slang words is mostly used in Malay weblog. The observation was made from the collection of popular Malay weblog that are listed in (Secretariat Malaysia Social Media, 2012).

d. The type of misspelled words that can be corrected by the second methodology which is identification and correction misspelled words in Malay language are;
   - Misspelled words formed from short form word,
   - Misspelled words formed from Error in typing

- Misspelled words formed by the influence by Selangor slang

e. In the third methodology, which is identification of neologism, the research only focus on neologism originated from loan word is English language.

1.6 Research Contributions

This research contributes towards an area of Computer Science field that is Artificial Intelligent (AI), specifically in text mining which is mainly focusing on processing data in NLP. Three approaches in identifying words that are unknown to the Malay dictionary are presented and the contributions of this research work are presented as below:

a. The first approach is to enhance the existing Malay stemming by adding background knowledge. This Malay stemming algorithm with background knowledge is able to reduce the error in stemming. Besides that, this enhancement is also able to reduce time consumed in Malay stemming process because the stemmer only processes the root words that have affixes. This enhancement has not been used in any Malay stemming or in any other languages stemming. Thus, by proposing this enhancement, it provides a new alternative to reduce the time consumed and reduce the error of the existing Malay stemming process.

b. The second approach that identifies the misspelled word or typo word in Malay language is also introduced a new methodology in order to identify and correct typo words. Although there are several spell checkers that already exists for Malay language, but the detection and correction of the short form misspelled words only can be done for misspelled words that exist in the predefined database. In other words, the system will not be able to detect and correct any misspelled words that are not listed in the predefined database. As a result, a new robust method is required that is able to perform the spell checker that does not depend entirely on the database. Thus, this approach proposes an approach that can automatically identify and correct misspelled words that do not depend entirely on the predefined database.

c. The third methodology is proposed in order to enhance the NLP in Malay language by identifying neologism words. The identification of neologism in other

languages such as English, Swedish, France, and Japanese are already conducted but not in Malay language. The morphology structure of Malay language is different compared to the other language thus, the approach to identify neologism words in other languages cannot be simply applied in identifying neologism in Malay language. Thus, by proposing this approach, it provides a new way to identify neologism in Malay language. By identifying neologism, it can help NLP to process the new words. Besides that, the identified neologism can be presented to lexicographers so that it can be included as part of the lexicon.

1.7 Thesis Organization

This thesis consists of six chapters. The structure of each chapter is briefly described as below:

a. Chapter 1

This chapter provides an overview of the research work in this thesis. It introduces briefly on the research background and the problem addressed. It also presents the objectives, scopes, and contributions of the research. The organization of the chapters in this thesis is outlined at the end of this chapter.

b. Chapter 2

This chapter presents the literature review related to the research work of this thesis. It begins by presenting an overview of the NLP. Then, some researches that related to the stemming, spell checker, and neologism are discussed in details.

c. Chapter 3

This chapter presents the overall methodology used in the research work. It presents the pre-processes involves which are the symbol elimination, and the English words elimination.

d. Chapter 4

This chapter presents the methodology for enhancing stemming process for Malay language. It also presents the experimental setup and lastly the result and discussion of the approach.