TEXT-BASED TAGGING OF MALAY HANSARD DOCUMENT

MOHD RAZIF BIN ABD JALIL

BACHELOR OF SCIENCE (HONS) COMPUTER SCIENCE
FACULTY OF COMPUTER SCIENCE AND MATHEMATICS
UNIVERSITI TEKNOLOGI MARA
MALAYSIA

JANUARY 2012

Acknowledgement

Bissmillahirrahmanirrahim,

Alhamdulillah. Thanks to Allah SWT, who with His willing giving me the opportunity to complete this Final Year Project which is title "Text-Based Tagging of Malay Hansard Document".

Firstly, I would like to express my deepest thanks to, Sir. Abdul Rahman Mohamad Gobil, the course coordinator and also to my project supervisor, Prof. Dr. Zainab Abu Bakar who had guided me a lot along these semesters, especially during the production of this project, and given helpful information, suggestions and guidance in the compilation and preparation of this project. Also thanks to my project examiner and the SIG who will evaluate this project, report and presentation.

Deepest thanks and appreciation to my parents and family for their cooperation, encouragement, constructive suggestion and full of support for the project completion, from the beginning till the end. Also thanks to all of my friends for the greatest support and to everyone who have been contributed by supporting my work and help myself during the final year project progress till it is fully completed.

From the bottom of my heart, thank you very much.

Abstract

In natural language processing, part-of-speech tagging plays a vital role. It is a significant condition for putting a human language on the computer science track. Before developing a part-of-speech tagger, a tag set is required for that language. This project is about the rule based part-of-speech tagging system for Malay language in Malay hansard document and a tag set that helps in the development of a Parser for the said language. The tagged word will compare with a text with manually tagging each word. The context free grammar will attach with the word that have more than one possible word class to perform a better result of tagging. A very simple architecture is applied that gives reasonably good accuracy. The result shows that 1.37 percent of hansard dictionary with highest frequency helps to tagging more than 55 percent words in hansard document.

Table of Contents

CHAPTER 1

Introduction

1.0 Introduction

This chapter states the overview of this project and discuss about project background, problem statement, objectives, scope and significance of the project.

1.1 Introduction of Project

Part-Of-Speech tagging is a process that particular tag assigned to each word sentences to show the function of words in specific contexts. POS tagging is considered as one of the basic tools and components required for any robust Natural Language Processing infrastructure of a given language. It is needed in various areas of language processing, starting from the simpler ones as text phrasing and alignment, to the more elaborate ones as syntax and semantic analysis and ending up with linguistic processes that is heavy as machine translation. Native speakers of a language perform grammatical and semantic analysis innately, and trained linguists can identify the grammatical parts of speech to multiple fine degrees depending on the tagging