ABSTRACT

Abstract of thesis presented to the Senate of Universiti Malaysia Terengganu in fulfillment of the requirements for the Degree of Master of Science

MALAY WORD SENSE DISAMBIGUATION ALGORITHM USING HYBRID METHODOLOGY

MOHD ARIZAL SHAMSIL BIN MAT RIFIN

AUGUST 2020

Supervisor : Associate Professor Mohd Pouzi Hamzah, Ph.D.

School/Institute : Faculty of Ocean Engineering Technology and Informatics

The ambiguity of word meaning is a major issue in the field of information retrieval and Natural Language Processing (NLP). The problem is that computers cannot determine the true meaning of a sentence based on the context of the sentence contained in a document. This is because of the nature of a word that has more than one meaning called a polysemy word. This problem is encountered in all languages including Malay. However, studies on this subject in the context of the Malay language are still lacking. Therefore, the thesis focuses on Malay word sense disambiguation and methods for word sense disambiguation of Malay words. There are several approaches that can be used to solve multiple word problems namely supervised, unsupervised and knowledge based approaches. In this study unsupervised and knowledge based approaches are combined and used to improve the Malay word sense disambiguation algorithms. In addition, a prototype that applied Malay Word Sense Disambiguation (MWSD) algorithm has been developed and a numerous testing have been carried out to evaluate the effectiveness of this algorithm. The results of the evaluation show that the developed algorithm obtained a good accuracy value of 0.723183673 compared to Google Translate 0.672938776, Yarowsky 0.543020408 and the Lesk algorithm 0.492571429.

ABSTRAK

Abstrak tesis yang dikemukakan kepada Senat Universiti Malaysia Terengganu sebagai memenuhi keperluan untuk Ijazah Sarjana Sains

ALGORITMA PENGECAMAN ERTI KATA PERKATAAN POLISEMI BAHASA MELAYU MENGGUNAKAN KAEDAH HIBRID

MOHD ARIZAL SHAMSIL BIN MAT RIFIN

OGOS 2020

Penyelia utama : Profesor Madya Mohd Pouzi Hamzah, Ph.D.

Pusat Pengajian : Fakulti Teknologi Kejuruteraan Kelautan dan Informatik

Masalah kekaburan makna perkataan merupakan satu isu yang besar dalam bidang capaian maklumat dan pemprosesan bahasa semula jadi. Masalah yang timbul ialah kerana komputer tidak dapat menentukan maksud sebenar sesuatu perkataan berdasarkan konteks ayat yang terdapat dalam sesuatu dokumen. Hal ini terjadi kerana sifat sesuatu perkataan itu yang mempunyai lebih daripada satu makna yang disebut sebagai perkataan polisemi. Masalah ini dihadapi oleh semua bahasa di seluruh dunia termasuklah bahasa Melayu. Namun begitu kajian mengenai perkara ini dalam konteks bahasa Melayu masih kurang dilakukan. Oleh itu, tesis ini memfokuskan kepada pengecaman makna perkataan bahasa Melayu dan kaedah bagi pengecaman makna perkataan bahasa Melayu. Terdapat beberapa pendekatan yang boleh digunakan dalam menyelesaikan masalah perkataan polisemi iaitu pendekatan diselia, pendekatan tidak diselia dan pendekatan berasaskan pengetahuan. Pendekatan tidak diselia dan pendekatan berasaskan pengetahuan digabungkan dan digunapakai dalam membangunkan algoritma pengecaman makna perkataan bahasa Melayu yang baharu. Selain itu, satu prototaip yang menggunakan algoritma Malay Word Sense Disambiguation (MWSD) telah dibangunkan dan sebilangan pengujian telah dijalankan bagi menguji keberkesanan algoritma ini. Hasil daripada ujian menunjukkan algoritma Malay Word Sense Disambiguation (MWSD) memberikan keputusan nilai ketepatan yang baik iaitu 0.723183673 berbanding Google Translate 0.672938776, Yarowsky 0.543020408 dan algoritma Lesk 0.492571429.