AUTOMATIC ENGLISH TO INDONESIAN LEXICAL MAPPING USING LATENT SEMANTIC ANALYSIS

UNDERGRADUATE THESIS

ELIZA MARGARETHA

120400030Y

FACULTY OF COMPUTER SCIENCE

UNIVERSITY OF INDONESIA

DEPOK

2008

AUTOMATIC ENGLISH TO INDONESIAN LEXICAL MAPPING USING LATENT SEMANTIC ANALYSIS

UNDERGRADUATE THESIS

Proposed as one of the prerequisites to obtain undergraduate degree

ELIZA MARGARETHA

120400030Y

FACULTY OF COMPUTER SCIENCE

UNIVERSITY OF INDONESIA

DEPOK

2008

Halaman Pengesahan

Judul Tugas Akhir:

Automatic English to Indonesian Lexical Mapping using Latent Semantic Analysis

Nama : Eliza Margaretha

NPM: 120400030Y

Laporan tugas akhir ini telah diperiksa dan disetujui.

Depok, Juli 2008

Pembimbing Tugas Akhir

Dr. Hisar Maruli Manurung

Acknowledgement

I wish to express my sincere gratitude to my super kind supervisor, Pak Ruli, who has taught me so many things, even English. I especially thank for your guidance, support, and patience. I do appreciate every valuable idea and advice in our interesting discussions.

I would like to acknowledge my examiner, Pak Chan, for giving some valuable suggestions for future work and numerical analysis. Also, I would like to acknowledge my other examiner and our head of Information Retrieval laboratory, Bu Mirna, for giving me opportunity to use some essential resources. I also thank for good occasions to share our research in the lab. I do appreciate your advice and support.

I am very thankful to my research colleague, Franky, for great and fun teamwork. Thank you for being my right hand in every need. I am also thankful to every member of Information Retrival laboratory and Arfan for good fellowship and amusing friendship. Especially to Desmond, thanks for helping in building bilingual dictionary and computing Fleiss Kappa values.

My special thanks to Ardi, Enrico, Martin, and Mulki for supporting me constantly. To Aristo, JP, and Ius, thanks for the suggestions in technical writing. Also, to every member of Fasilkom UI family, thanks for every help, support, and consideration towards my thesis.

Above all, I am very grateful to my family, who always prays for me and shows me heartfelt love and care. And especially to my little nephew, Bryan, who has brought me joy in fatigued moments.

The work presented in this thesis is supported by an RUUI (Riset Unggulan Universitas Indonesia) 2007 research grant from DRPM UI (Direktorat Riset dan Pengabdian Masyarakat Universitas Indonesia).

Abstrak

WordNet (Fellbaum, 1998) adalah suatu *lexical resource* yang kaya akan informasi linguistik yang sangat bermanfaat bagi berbagai macam aplikasi, khususnya aplikasi-aplikasi yang berhubungan dengan linguistik, pemrosesan bahasa alami, dan kecerdasan buatan. Dewasa ini, WordNet telah dibangun untuk lebih dari 40 bahasa, tetapi WordNet untuk bahasa Indonesia belum tersedia. Oleh karena pengembangan WordNet secara manual membutuhkan sumber daya yang tidak sedikit, penelitian yang dipaparkan dalam laporan tugas akhir ini bermaksud untuk membangun WordNet secara otomatis.

Penelitian ini mencoba untuk membuat *synset* (*synonym set*) untuk bahasa Indonesia dengan melakukan pemetaan konsep dwibahasa secara otomatis antara konsep bahasa Inggris yang diambil dari Princeton WordNet dan konsep bahasa Indonesia yang diambil dari Kamus Besar Bahasa Indonesia (KBBI). Tugas lain, yaitu pemetaan kata dwibahasa, diperkenalkan untuk memetakan kata-kata bahasa Inggris ke kata-kata bahasa Indonesia secara otomatis. Kedua pemetaan tersebut dilakukan dengan mengaplikasikan metode *Latent Semantic Analysis* (Landauer, Foltz, & Laham, 1998) pada korpora paralel berupa teks.

Awalnya, pemetaan kata dwibahasa dimaksudkan untuk melakukan verifikasi proses di balik pemetaan konsep dwibahasa. Namun, hasil pemetaan kata tidak memuaskan karena performa model kemiripan vektor lebih baik dari pada model LSA. Di sisi lain, hasil dari pemetaan konsep dwibahasa, menunjukkan kemampuan LSA untuk menangkap informasi semantik yang terkandung secara implisit dalam suatu korpus parallel. Walaupun LSA belum berhasil mencapai tingkat yang setara dengan pemetaan yang dilakukan manusia, secara umum LSA lebih baik dari pada *random baseline*.

Abstract

WordNet (Fellbaum, 1998) is a lexical resource containing rich linguistic knowledge, which is very useful for a wide variety of applications, especially for applications related to linguistics, natural language processing, and artificial intelligence. Recently, WordNets have been built for more than 40 languages, but not yet in Indonesian. Since building a WordNet manually is complex and expensive, the work presented in this thesis considers building an Indonesian WordNet automatically.

This work attempts to construct Indonesian *synsets* (*synonym set*) by conducting automatic bilingual concept mapping between English concepts derived from Princeton WordNet and Indonesian concepts derived from Kamus Besar Bahasa Indonesia (KBBI). Another task, namely bilingual term mapping, is introduced to map English terms to their Indonesian analogues automatically. Both mappings are conducted by applying Latent Semantic Analysis (Landauer, Foltz, & Laham, 1998) on parallel corpora of text.

Bilingual term mapping was intended to verify the underlying process of bilingual concept mapping. However, the results are unsatisfactory suggesting that vector model similarity performs better than the LSA model. The results of bilingual concept mapping, on the other hand, show some capability of LSA to capture some semantic information implicit within a parallel corpus. Although LSA is not yet able to attain levels comparable to human judgements, it is generally better than random baseline.

Table of Contents

List of Tables

List of Figures