

**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 Retrieval 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 paralel. 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

Title	i
Halaman Pengesahan	ii
Acknowledgement.....	iii
Abstrak	iv
Abstract	v
Table of Contents.....	vi
List of Tables.....	x
List of Figures.....	xiii
Chapter 1 Introduction	1
1.1 Motivation.....	1
1.2 Problem Statement.....	2
1.3 Objective.....	2
1.4 Scope of This Work	2
1.5 Research Methodology	3
1.6 Structure of This Thesis.....	3
Chapter 2 Literature Study	5
2.1 WordNet.....	5
2.1.1 History of WordNet	5
2.1.2 Design of WordNet.....	7

2.1.3	Contents of WordNet	9
2.1.4	Relations in WordNet	11
2.1.5	Applications of WordNet.....	13
2.2	Singular Value Decomposition	14
2.2.1	Properties of SVD.....	14
2.2.2	Construction of SVD.....	15
2.2.3	Rank Approximation.....	19
2.3	Latent Semantic Analysis	20
2.3.1	LSA as a Theory of Meaning	21
2.3.2	Similarity Approximation of LSA	22
2.3.3	Mathematical Properties of LSA	23
2.3.4	Applications of LSA	30
2.4	Automatic Sense Disambiguation.....	30
2.4.1	Word Sense Disambiguation.....	31
2.4.2	Cross Language Information Retrieval Using LSA.....	33
2.4.3	WSD Using LSA	36
Chapter 3 Design.....		37
3.1	Task Definitions.....	37
3.1.1	Task of Bilingual Term Mapping	39
3.1.2	Task of Bilingual Concept Mapping.....	40
3.2	How LSA is Applied.....	41
3.2.1	Building Bilingual Term-Document Matrix	42
3.2.2	Building Bilingual LSA Matrix	44
3.3	Design of Bilingual Term Mapping	45
3.4	Design of Bilingual Concept Mapping	48
Chapter 4 Implementation Details		51
4.1	Overview.....	51
4.2	Implementing LSA	52

4.3	Building Conceptual Semantic Matrix.....	57
4.4	Conducting Bilingual Mapping.....	60
Chapter 5 Result and Discussion.....		63
5.1	Bilingual Term Mapping Experiment.....	63
5.1.1	Existing Resources.....	64
5.1.2	Variables	65
5.1.3	Sample of Bilingual Term Mapping Result	66
5.1.4	Effect of Source Terms Selection and Frequency Weighting.....	68
5.1.5	Effect of Collection Size and Rank Approximation	70
5.1.6	Effect of Stopwords	74
5.1.7	Effect of Target Term Mapping Selection	75
5.1.8	Effect of Text Alignment Granularity.....	76
5.1.9	Discussion.....	77
5.2	Bilingual Concept Mapping Experiment	79
5.2.1	Existing Resources.....	80
5.2.2	Variables	81
5.2.3	Sample Bilingual Concept Mapping Result.....	81
5.2.4	Comparison of Level of Agreement.....	86
5.2.5	Effect of Target Concept Mapping Selection	88
5.2.6	Effect of Frequency Weighting.....	88
5.2.7	Effect of Frequency Weighting and Concept Mapping Selection	89
Chapter 6 Conclusions.....		90
6.1	Summary	90
6.2	Conclusions.....	91
6.2.1	Conclusions of Bilingual Term Mapping Experiments	91
6.2.2	Conclusions of Bilingual Concept Mapping Experiments.....	93
6.3	Limitations	94
6.4	Future Work.....	95

Appendix A Sample Resources.....	98
C.1 Sample of Parallel Corpus 1	99
A.1.1 Sample of Article Pairs	99
A.1.2 List of Random Set Terms	102
A.1.3 List of Top Score Set Terms	102
C.2 Sample of Parallel Corpus 2	103
C.2.1 Sample of Bible Chapters and Verses.....	103
C.2.2 List of Top Score Set Terms	106
Appendix B Experiment Results.....	108
B.1 Bilingual Term Mapping Results.....	109
B.1.1 Parallel Corpus 1 for Random Set	109
B.1.2 Parallel Corpus 1 for Top Score Set	111
B.1.3 Random Baseline for Top Score Set of Parallel Corpus 1	117
B.1.4 Parallel Corpus 2 for Top Score Set	118
B.1.5 Text Alignment Granularity Experiment Results	120
Appendix C Full Paper in Proceedings of the Second International Malindo Workshop 2008, Malaysia.....	123
Bibliography.....	131

List of Tables

Table 2.1 Example of Data: Titles for Topics on Internet and Baking.....	23
Table 2.2 Term-Document Matrix with Term Frequencies Corresponding to the Titles in Table 2.1	24
Table 2.3 SVD of Term-Document Matrix Represented in Table 2.2.....	25
Table 2.4 Two-Dimensional Reconstructed Term-Document Matrix Represented in Table 2.2 based on Two Shaded columns of SVD matrices.....	26
Table 2.5 Correlation Coefficient between Titles in the Original Term-Document Matrix	27
Table 2.6 Correlation Coefficient between Titles in the Two-Dimensional Reconstructed Term-Document Matrix	27
Table 3.1 Sample of Concepts and Terms in <i>Ne</i> and <i>Ni</i>	39
Table 3.2 Example of Bilingual Term-Document Matrix.....	43
Table 3.3 Sample of Documents: Titles for Topics on Time and River	46
Table 3.4 Sample of English-Indonesian Term-Document Matrix Based on Table 3.3...	47
Table 3.5 Sample of English-Indonesian LSA Matrix with Rank-2 Approximation Based on Table 3.4.....	47
Table 3.6 Sample of English-Indonesian Term Mapping Result Based on Table 3.5	48
Table 3.7 Sample of Set of Textual Context and Conceptual Semantic Vector Based on English-Indonesian LSA Matrix in Table 3.5	49

Table 3.8 Sample of Bilingual Concept Mapping Result Based on Table 3.7	50
Table 5.1 Sample of Bilingual Term Mapping for English terms (a) film and (b) billion	67
Table 5.2 Comparison of Mapping Results	68
Table 5.3 Comparison of Source Terms Selection	69
Table 5.4 Comparison of Weighting Usage.....	70
Table 5.5 Unique Term Statistics.....	71
Table 5.6 Effect of Collection Size.....	72
Table 5.7 Effect of Varying Rank Approximation	73
Table 5.8 Effect of Stopwords	74
Table 5.9 Effect of Varying Target Term Mapping Selection.....	75
Table 5.10 Sample of LSA Experiment Results using Chapters and Verses.....	76
Table 5.11 Effect of Text Alignment Granularity.....	77
Table 5.12 Comparison of Level of Agreement	86
Table 5.13 Effect of Concept Mapping Selection.....	88
Table 5.14 Effect of Weighting	89
Table 5.15 Effect of Weighting and Concept Mapping Selection	89
Table A.1 List of Random Set Terms of P_{100}	102
Table A.2 List of Top Score Set Terms of P_{100}	102
Table A.3 English Bible: Book Genesis Chapter 1.....	103
Table A.4 Indonesian Bible: Book Kejadian Chapter 1.....	105
Table A.5 List Top Score Set Terms of P_{1000}	106
Table B.1 Bilingual Term Mapping Results for Random Set using Parallel Corpus 1 Including Stopwords and No Weighting.....	109

Table B.2 Bilingual Term Mapping Results for Random Set using Parallel Corpus 1 Excluding Stopwords and No Weighting	110
Table B.3 Bilingual Term Mapping Results for Top Score Set using Parallel Corpus 1 Including Stopwords and No Weighting	111
Table B.4 Bilingual Term Mapping Results for Top Score Set using Parallel Corpus 1 Excluding Stopwords and No Weighting	112
Table B.5 Bilingual Term Mapping Results for Top Score Set using Parallel Corpus 1 Including Stopwords and with TF-IDF Weighting	113
Table B.6 Bilingual Term Mapping Results for Top Score Set using Parallel Corpus 1 Excluding Stopwords and with TF-IDF Weighting	114
Table B.7 Bilingual Term Mapping Results for Top Score Set using Parallel Corpus 1 Including Stopwords and with Log Entropy Weighting.....	115
Table B.8 Bilingual Term Mapping Results for Top Score Set using Parallel Corpus 1 Excluding Stopwords and with Log Entropy Weighting	116
Table B.9 Random Baseline Results for Top Score Set of Parallel Corpus 1	117
Table B.10 Bilingual Term Mapping Results for Top Score Set using Parallel Corpus 2 Including Stopwords and No Weighting	118
Table B.11 Bilingual Term Mapping Results for Top Score Set using Parallel Corpus 2 Excluding Stopwords and No Weighting	119
Table B.12 Text Alignment Granularity Experiment Results for Bible Verses and Chapters Including Stopwords and No Weighting	120
Table B.13 Text alignment granularity Experiment Results for Bible Verses and Chapters Excluding Stopwords and No Weighting	121
Table B.14 Comparison of Mapping Selection and Text Alignment Granularity	122
Table B.15 Comparison of Rank Approximation and Text Alignment Granularity.....	122
Table B.16 Comparison of Stopwords Usage and Text Alignment Granularity	122

List of Figures

Figure 2.1 WordNet SQL version 3.0 Scheme (Bou, 2007).....	10
Figure 2.2 Two Dimensional LSA Semantic Space for Internet-Baking Titles Documents Showing Similarity of Meaning of the Documents	28
Figure 2.3 Two Dimensional LSA Semantic Space for Internet-Baking Titles Documents Showing Similarity of Meaning of the Terms	29
Figure 3.1 Constructing English-Indonesian Term-Document Matrix.....	43
Figure 3.2 Construction of Rank k Approximation Term-Document Matrix.....	45
Figure 4.1 Pseudocode for Creating a Document Collection.....	53
Figure 4.2 Pseudocode for Building a Term-Document Matrix	54
Figure 4.3 Pseudocodes for <i>TF-IDF</i> and <i>Log-Entropy</i> Weighting	55
Figure 4.4 Pseudocode for Building an LSA Matrix	56
Figure 4.5 Sample of Query Result Acquired from Princeton WordNet Database	57
Figure 4.6 Textual Context Set for Concept in Figure 4.5.....	58
Figure 4.7 Sample of Query Result from KBBI Database.....	58
Figure 4.8 Pseudocode for Building Conceptual Semantic Matrix	59
Figure 4.9 Pseudocode for Building Similarity Matrix.....	61
Figure 5.1 Sample of Bilingual Concept Mapping Result.....	82
Figure 5.2 Example of Successful Mapping for English Common-Based Concept.....	84

Figure 5.3 Example of Unsuccessful Mapping for English Common-Based Concept.....	85
Figure A.1.1 Indonesian Article for Good Translation Sample	99
Figure A.1.2 English Article for Good Translation Sample	100
Figure A.1.3 Indonesian Article for Poor Translation Sample	101
Figure A.1.4 English Article for Poor Translation Sample.....	101

