Chapter 1

Introduction

This chapter expresses the motivation of the work presented in this thesis. It states the research problem and the objective of the research. In the last sections, the methodology used and the structure of this thesis are briefly described.

1.1 Motivation

In this information era, the need of information does increase rapidly. Not only does the number of information rise, but there is also a need for wider coverage of information across many languages. As a result, cross language resources turn out to be vital to support cross language systems. For certain applications, there is also a need for very specific mappings which are more refined than bilingual dictionary. For example, concept or word sense mappings are crucial to acquire high accuracy and quality in machine translation. Considering these needs, the work in this thesis attempts to map English terms and concepts to their Indonesian analogues.

WordNet (Fellbaum, 1998) is an example of lexical resources arranged in terms of word senses. It contains rich linguistic knowledge, including *synsets*, lexical relations, and semantic relations. Primarily, WordNet was established for English and it has been very useful for a wide variety of applications, especially for applications related to linguistics, information retrieval, natural language processing, and artificial intelligence (see Section 2.1). Recently, WordNets have been built for more than 40 languages, but not yet in Indonesian. Concerning the importance of WordNet for processing a language, the work in this thesis considers building an Indonesian WordNet.

1.2 Problem Statement

Basically, a WordNet can be built manually or automatically. However, building a WordNet manually is certainly complex and expensive, i.e. time consuming and requires contributions of many people, e.g. (Darma Putra, Arfan, & Manurung, 2008). To meet high quality contents, it even requires extensive work by linguistic experts, e.g. (Lim & Hussein, 2006). For these reasons, the work presented in this thesis considers building an Indonesian WordNet automatically, which requires less human effort and time.

Accordingly, the research problem is how to build an Indonesian WordNet automatically. In particular, a specific method, namely Latent Semantic Analysis or LSA (Landauer, Foltz, & Laham, 1998) is observed whether it can be used to resolve the problem. Thus, the specific research problem is whether LSA can be used to build Indonesian WordNet automatically.

1.3 Objective

As described in the previous section, the work presented in this thesis is intended to automate the process of building an Indonesian WordNet. Specifically, the objectives are:

- To make a preparation for building Indonesian WordNet, specifically to provide Indonesian *synsets* automatically.
- To discover whether LSA can be used to provide the Indonesian *synsets* automatically by employing parallel corpora of text.
- To map English concepts derived from Princeton WordNet to Indonesian concepts derived from the Kamus Besar Bahasa Indonesia (KBBI) using LSA.

1.4 Scope of This Work

Some constraints which restrict the scope of this work are:

- This work does not seek to construct a complete Indonesian WordNet, but it makes some preparation, that is to establish mappings between English and Indonesian concepts.
- This work is restricted with respect to the availability of resources. For concept mapping, English concepts derived from Princeton WordNet version 3.0 are mapped to Indonesian concepts derived from KBBI.

- The only method used in this work is LSA applied on parallel corpora of text.

1.5 Research Methodology

The work in this thesis was conducted via experimental research. The purpose of the study is explanatory or descriptive. Explicitly, the research was conducted as follows:

1. **Literature Study**

   First of all, some literatures were studied so as to obtain some information related to the research problems. The information observed includes topics on WordNet, LSA, Singular Value Decomposition (SVD), Cross Language Information Retrieval (CLIR), and Word Sense Disambiguation (WSD). Additionally, former research employing LSA was observed.

2. **Design**

   In design, the research objectives were elucidated as tasks to be accomplished, namely bilingual term mapping and bilingual concept mapping. Subsequently, a model employing LSA was defined to carry the tasks out. The resources required and the variables which would be experimented were also defined.

3. **Implementation and Experiments**

   Based on the design and the available resources, the LSA model for bilingual term and concept mappings was implemented. Afterwards, some experiments with various variable configurations were performed. Some baselines were also computed as comparison.

4. **Results Interpretation and Analysis**

   Finally, the experiment results were interpreted and analysed. Also, the results of LSA and the baselines were compared.

1.6 Structure of This Thesis

This thesis is structured as follows:

Chapter 1 describes the motivation of this work and the research problem to be solved. Then, it states the objective and the scope of the work. It also explains the research methodology used and the outline of the chapters in this thesis.

Chapter 2 discusses the information gathered during the literature study including topics on WordNet, Latent Semantic Analysis (LSA), Singular Value Decomposition (SVD), Cross Language Information Retrieval (CLIR), and Word Sense Disambiguation (WSD).

Chapter 3 presents the design of the work in detail. Firstly, it explains the task definitions. Secondly, it explains how LSA is applied to accomplish the tasks. Finally, it explains the design of the term and concept mappings.

Chapter 4 describes how the design was implemented. It explains how to implement the LSA model, i.e. how to build a term-document matrix, and how to apply SVD of that matrix so as to get an LSA matrix. Specifically for bilingual concept mapping, it explains how to build a conceptual semantic matrix. Lastly, it explains how to conduct both bilingual term and concept mappings.

Chapter 5 reports the results of the experiments along with their interpretations. Some discussions about the characteristics of LSA and the variables experimented are given.

Chapter 6 presents a summary of the work and some conclusions that can be drawn. It lists some limitations encountered in this work and gives some suggestions for future work.