Chapter 1

Introduction

This chapter presents background of the research, problem statement, objectives, delimitations, research methodology and a brief description of this book structure.

1.1 Background

Quran is the holy book for Muslims –people who believe and follow the religion of Islam. This Holy Book is currently being used as the guidebook of life to 1.6 billion Muslims in the world living today [3]. According to them, Quran is considered to be verbatim words of God, and they value it as the source of authority, knowledge, wisdom and law.

Unfortunately, this important manuscript has not been gained its popularity in computational linguistics community as a research object. Many groundwork data sets and linguistics preprocessing methods studies are yet still unable to be seen, for example, the study of semantic textual similarity that uses Quran manuscript as the data set.

Semantic textual similarity is a linguistics study that plays a major role in text mining and natural language processing. Semantic similarity, or sometimes being called as attributional similarity [4], takes two language units as arguments (be it in a form of paragraph, sentence, concept, or word) and look at how much these two language units share common attribute or property. The more this pair of linguistics unit have common characteristics, the more semantically similar it is.

In this study, we wanted to build an automatic system that would be able to measure the degree of semantic similarity between two concepts within a Quran manuscript. The original form of Quran is spoken Classical Arabic and these days, the Holy Book has been recorded into a form of book, being digitized and translated into at least 44 languages as a supplement to help Muslims who don't speak Arabic understand the knowledge [5]. In this prior study, we didn't use the Arabic form of Quran, but rather, we used the English translation by Saheeh International as the data set. This won't be a problem, because the property of concept is language independent [6].

Along with building the system, we also wanted to generate a gold standard data set that consists of pairs of concepts gathered from the Quran and its similarity score based on human annotations. The degree of similarity would be measured in the continuous range of [0, 10]. The score 10 denotes maximum similarity, whereas 0 expresses no similarity at all. The challenge of this study is to better mimic human intuition in measuring similarity on semantic domain.

The automatic concept semantic similarity measurement that we built is a knowledge based system that uses the knowledge gathered from WordNet; a lexical database that is widely being used in lexical semantic similarity study [7] [4], and the quantity and range of vocabulary of this database is the best that we currently have. We used the equation from Yuhua Li to manipulate the features of WordNet into similarity score [8], since it was proven to be working good in WordNet and has been used as the baseline by many researchers in similar study.

As many previous research in semantic similarity, performance of the proposed system would later be evaluated using intrinsic evaluation, by computing the Spearman's and Pearson's correlations between the output score given by system and manually annotated score set by human experts (gold standard).

Later in the future, this research can be brought to even more interesting problems related to natural language processing and information retrieval of Quranic text. For example, question answering in Islamic topic, text simplification of Islamic texts to better understand the context, and Islamic scripts summarization.

1.2 Problem Statement

The fundamental purpose of this study are as follow:

- How to develop a Quranic concepts gold standard?
- How to measure semantic similarity between two Quranic concepts by making use of WordNet as the source of knowledge.

1.3 Research Objectives

The two main objectives of this conducted study are as follow:

- to develop a Quranic concepts gold standard,

- to apply Yuhua Li's algorithm to measure semantic similarity between two Quranic ceoncepts by making use of path length and depth of lowest common subsumer between two concepts that are gathered from WordNet, and
- to find the best value of α and β in Yuhua Li's equation to get the best correlation between the gold standard and the system's output.

1.4 Delimitations

Below are the delimitations of this conducted research:

- The term *concept* is roughly defined as word or phrase in this research.
- In this pilot study, we focused only on studying concepts that belong to the noun category. This is because noun is the only part of speech in WordNet that forms a tree (with one singular root), while other part of speech forms a separate tree [9].
- Similarity problem may be asymmetric (similarity(concept1, concept2) and similarity(concept2, concept1) might result in different score), but previous research have shown that its impact is insignificant and can be ignored [7] [10].
- The concepts that were gathered from Quran were being translated into English according to a translation by Saheeh International [11]. Semantic similarity measurement involving concepts in English and not in the classical Arabic form.
- In this initial small study, annotators were not expected to have expertise in Islamic or Quranic Study, nor lingustics.
- The similarity score is in the range of [0.00, 10.00]. The score 10.00 denotes maximum similarity, whereas 0.00 expresses no similarity at all.

1.5 Research Methodology

Below were the methodology of this study:

1. Research problem identification
   In this initial step, research problem was defined and clarified. The method

used in order to identify the research problem was by literature study. Previous related works were studied to get an insight of previous attempt to the problem.

2. Data gathering
   Data sets were collected from some available sources. This proposed research also requires some data sets which are not available yet. Those data sets will be collected and prepared in this step.

3. Implementation
   The system will be implemented in this process. Architecture, method and other components will be built in order to meet the specifications listed on previous problem identification step.

4. Analyzing the results
   Performance of system will be evaluated in this step. The results are then will be analyzed to answer the research question.

1.6 Book Structure

This book is divided into five chapters structured this way:

1. Introduction
   In this chapter, background, problem statement, objectives and delimitations of this conducted study were described, along with research methodology and this book structure section.

2. Literature Review
   This second chapter will explore the prior relevant studies and theories.

3. System Design
   In this chapter, the architecture of system and data gathering is described further.

4. Testing and Analysis
   This chapter provides system result and analysis of performance and some interesting findings.

5. Conclusion
   This last chapter lists all the conclusions from this study.