Ontological Approach for Semantic Modelling of Malay Translated Qur'an

Nor Diana Binti Ahmad

Submitted in accordance with the requirements for the degree
of
Doctor of Philosophy

The University of Leeds
School of Computing

January 2022

The candidate confirms that the work submitted is her own and that appropriate credit has been given where reference has been made to the work of others.

Declaration

I declare that the work presented in this thesis is to best of my knowledge of the domain, original, and my own work. Most of the work presented in this thesis have been published.

(Nor Diana Binti Ahmad)

Publication

Chapter 2 - Literature Review

This paper summarizes the search techniques used in existing research on Qur'an. Moreover, this paper also studied the previous research conducted on Qur'an Semantic Search and Qur'an Ontology-Based Search focusing on Malay Qur'an.

1. Ahmad, ND, Bennett, B and Atwell (2016). Semantic-based Ontology for Malay Qur'an Reader. IMAN'2016 4th International Conference on Islamic Applications in Computer Science and Technologies. 20-22 Dec 2016, Khartoum, Sudan. ES orcid.org/0000-0001-9395-3764

Chapter 4 - Issues in Translation between the English, Malay and Arabic languages

The issues in English-Malay Translation presented in Chapter 4 are based on the following paper:

1. Ahmad, N. D., Bennett, B., and Atwell, E. (2017). Retrieval Performance for Malay Quran. International Journal on Islamic Applications in Computer Science and Technology (IJASAT), 5(2), 13-25.

Acknowledgements

First and foremost, I would like to thank Almighty God (SWT) for His Mercy and Compassion, prayer and peace is upon His Prophet Muhammad that enables me to complete this thesis, Alhamdulilah.

I wish to express my sincere gratitude to my supervisors, Dr Brandon Bennett and Professor Eric Atwell, for their guidance and direction in this thesis. During my work on this thesis over the last three and a half years, I benefited immensely from their expert advice on ontology development, corpus annotation, and computational linguistics. They provided me with many interesting, valuable, and real feedback and motivation throughout their supervision. I greatly benefited from their detailed comments and insights, which helped me clarify the ideas presented in "Ontological Approach for Semantic Modelling of Malay Translated Qur'an". Their support and encouragement which has contributed immensely towards the completion of my PhD.

Most of all, I owe my thanks to Ayahanda Walid Al-Malik, my lovely husband, Muhammad Bukhairi Bin Mat Johor, my four beautiful daughters, Nur Aishah Humaira, Nur Aishah Mariah,Siti Nur Athiyyah Sumaiyyah and Nur Hannah Insyirah, my families and my friends for the encouragement and support, for always being there to provide those precious words of encouragement when I needed them the most.

My sincere gratitude and thanks to the staff of Karangkraf Malaysia and Indonesia for granting permission to use the Amazing Al-Quran text as the dataset for this study. I would also like to acknowledge the hard work of the numerous volunteers who contributed their time and effort towards the completion of my PhD.

Lastly, I wish to thank the Ministry of Higher Education of Malaysia and Universiti Teknologi MARA for the financial support awarded to me, and the staff and employees of the University of Leeds, UK.

Thank you very much

Abstract

This thesis contributes to the areas of ontology development and analysis, natural language processing (NLP), Information Retrieval (IR) and Language Resource and Corpus Development.

Research in Natural Language Processing and semantic search for English has shown successful results for more than a decade. However, it is difficult to adapt those techniques to the Malay language, because its complex morphology and orthographic forms are very different from English. Moreover, limited resources and tools for computational linguistic analysis are available for Malay. In this thesis, we address those issues and challenges by proposing MyQOS,the Malay Qur'an Ontology System, a prototype ontology-based IR with semantics for representing and accessing a Malay translation of the Qur'an. This supports the development of a semantic search engine and a question answering system and provides a framework for storing and accessing a Malay language corpus and providing computational linguistics resources. The primary use of MyQOS in the current research is for creating and improving the quality and accuracy of the query mechanism to retrieve information embedded in the Malay text of the Qur'an translation. To demonstrate the feasibility of this approach, we describe a new architecture of morphological analysis for MyQOS and query algorithms based on MyQOS. Data analysis that consisted of two measures; precision and recall, where data was obtained from MyQOS Corpus conducted in three search engines. The precision and recall for semantic search are 0.8409 (84%) and 0.8043(80%), double the results of the question answer search which are 0.4971(50%) for precision and 0.6027 (60%) for recall.The semantic search gives high precision and high recall comparing the other two methods. This indicates that semantic search returns more relevant results than irrelevant ones. To conclude, this research is among research in the retrieval of the Qur'an texts in the Malay language that managed to outline state-of-the-art information retrieval system models. Thus, the use of MyQOS will help Malay readers to understand the Qur'an in better ways. Furthermore, the creation of a Malay language corpus and computational linguistics resources will benefit other researchers, especially in religious texts, morphological analysis and semantic modelling.

List of Tables

List of Figures

Contents

Part I

Introduction and Background

Chapter 1

Introduction

A general overview of the thesis is provided in this chapter. The focus is on the definition of the problems that motivate the interest in the research, an outline of the proposals is developed to address the research problems and the resulting outcomes of the research. Section 1.1 presents the motivation of the research, describing the problems to be addressed and the limitations of the approaches reported in the literature. Section 1.2 defines the scope of the study by stating the addressed research questions, and the central sought goal. Section 1.3 summarizes the specific aimed achievements and contributions of this research to the field, as well as the approach to reach them. Finally, Section 1.4 describes the structure of this thesis.

1.1 Motivation

Qur'an is the holy book of Muslims that contains the words of Allah. Qur'an provides instruction and guidance to humankind in achieving happiness in life in this world and hereafter. As a holy book, the Qur'an contains rich knowledge and scientific facts. Muslims are required to read and learn the meaning of the Qur'an in languages they understand obtaining rewards from Allah and to efficiently help Muslims to perform their daily routines.

Nowadays, the Qur'an has been translated into various languages around the world by Muslim experts. The main aim of the availability of the Qur'an translations is to allow the reader to understand the Qur'an clearly. This is because many people have difficulty in understanding the context of the Qur'an, especially those who are not proficient in Arabic. According [Yusof et al., 2011], a non-Arabic-speaking Muslim would recite the Qur'an even though they do not understand the meaning since it is considered as an act of worship. This has caused them to rely on the translation of the Qur'an written in their native language to understand the content or what she/he has read. Nevertheless, the words used in most of the translations are diverse and it is still quite difficult to understand the meaning of the Qur'an for some readers. Besides, many allegorical or metaphorical words are used in the Qur'an. Allegorical verses are unclear verses and sometimes it has more than one meaning and requires further explanation. For instance, the word الجنة in Arabic can give a different meaning as in Figure 1.1. Here, when the user queries the word الجنة to find the verses related to paradise, the user will retrieve other verses that contain the word الجنة but with a different meaning.

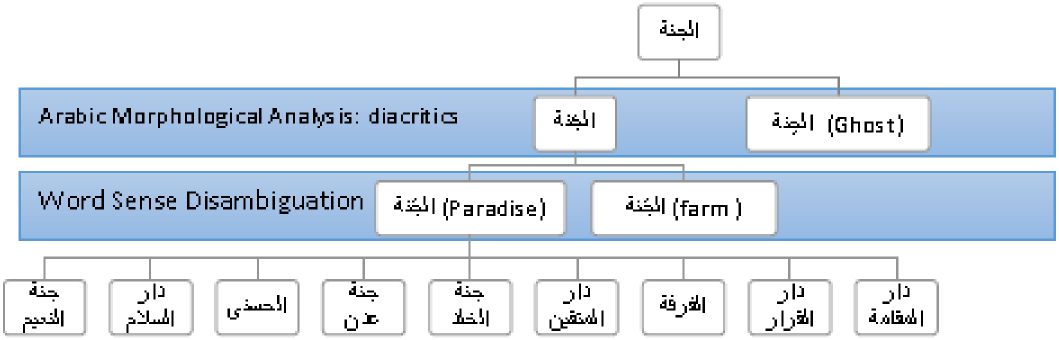


Figure 1.1: The classification of the meaning of word الجنة [Alqahtani and Atwell, 2016]

The massive increase in the amount of contents stored and shared on the Web and other repositories has triggered an extreme demand for tools and techniques that can handle and process data semantically. The present practice in retrieving information from the Qur'an mostly relies on keyword-based search techniques. Nonetheless, such techniques have resulted in the loss of valuable information embedded in the text. The issue of keyword-based search has received considerable critical attention among many computer analysts. According to Sanchez

[Fernández Sánchez, 2009] in his Ph.D. thesis, the limited capabilities of keyword-based search include the failure to describe relations between search terms and the inability to properly handle the linguistic phenomena such as polysemy, synonyms, and homonyms that are found in the Qur'an. Although extensive research has been carried out on the Qur'an, not many studies have been conducted on the translation of the Qur'an in Malay language. The lack of morphological analyses research published in the Malay translated Qur'an creates a need to address the gap in literature on the linguistic characteristics of the language widely used in South East Asia.

Research in information retrieval for the Malay translated Qur'an is a new venture. The first research in information retrieval of the Malay translated Qur'an emerged in 1995 by Ahmad, F. D. in [Ahmad, 1995]. Thereafter, many researchers began conducting studies using Malay translated Qur'an by applying new computational and linguistic approaches. However, research has consistently shown that Malay is a language that is rich with morphology and orthographic forms, which leads to complexity [Mohd Don, 2010]. Moreover, since then, the resources and tools for computational linguistic analysis for this language have remained minimal [Ahmad et al., 2016]. Many attempts have been made to solve this complexity and improve the retrieval of Malay's translated Qur'an data, but most of them use only keyword-based searches. Thus, little work has been done on how semantic retrieval can help to improve the retrieval of information in the Malay translated Qur'an.

Aiming to solve the limitations of keyword-based models in the Malay translated Qur'an, this thesis proposed MyQOS, the Malay Qur'an Ontology System to support semantic retrieval capabilities which can work better than keyword-based search. Although there has been some research that has been carried out in ontology, there are no studies that have been found on ontology-based IR with semantics using Natural language processing (NLP), specifically in the Malay translated Qur'an. Therefore, this thesis makes a major contribution to the research on IR in Malay translated Qur'an by demonstrating a new ontology-based Information Retrieval (IR) with semantics using NLP. Furthermore, this thesis also aims to provide a semantic search engine prototype that enhances the query mechanism to retrieve the information embedded in the text. This approach involves the use of the Malay Qur'an ontology, which is used as the knowledge base for the system. The final result is verses retrieved from the Malay translated Qur'an that answer semantically formulated queries.

1.2 Research Questions

This thesis aims to address the following research questions:

1.2.1 Can morphological analysis help in retrieving accurate information in the Malay translated Qur'an?

Research in morphological analysis for the English language has shown successful results for more than a decade. However, there are difficulties in adapting those techniques in the Malay language. This is reflected by the research outcome from a previous study by Don, Z. M and Abdullah et al. [Mohd Don, 2010, Abdullah et al., 2009] that highlighted that Malay is a language that is rich with morphology and orthographic forms which leads to its complexity. This is the main reason why the morphological analysis in Malay language documents is still far behind. Moreover, there are limited resources, tools for computational linguistic analysis available for the Malay language.

The need of computer understanding in natural languages has been playing a crucial role in many research fields in the past few decades. However, the combination of Natural Language and IR brings a lot of new challenging tasks. The development of Natural language Processing (NLP) with semantic model has shown some improvement on IR. The combination of NLP and semantic model is said will enhance the retrieval quality and accuracy as it will become a key element in the development of the semantic model [Estival et al., 2004, Sheth et al., 2017]. However, only a few studies have been carried out on combining these techniques in the Malay translated Qur'an.

1.2.2 Can ontology-based IR with semantics improve the query mechanism for Malay Translated Qur'an?

The Qur'an is fundamental to all Muslims because it contains comprehensive guidance and knowledge to Muslims in all aspects of life. Nowadays, the Qur'an has been translated into various languages around the world by Muslim experts. The main aim of the availability of the Qur'an translations is to allow the reader to understand the Qur'an in their own native language. With the help of technology, such translations are now available in digital form, including in the form of device applications. However, most of Malay translated Qur'an application offers only the keywords search method for the user to query the information. This type of searching technique results in the failure to retrieve the concise and relevant knowledge in the Qur'an. Thus, one of the key issues in Information Retrieval (IR) is to develop a search engine capable of acquiring knowledge via ontology.

1.3 Original Contribution of the thesis

The main contributions that will be presented in the thesis are :

- The first Malay translated Qur'an corpus of 149,654 words with root word annotation and root dictionary has grammatical categories, ontology of concepts, word-by-word, English translation and synonym relation.
- A new morphological analysis algorithm for Malay Translated Qur'an. This includes a new list of Malay Stop words, a new rule-based stemming algorithm, and a new root words annotation.
- A new Ontology-based IR with semantics search.

1.4 Thesis Outline

This thesis into **THREE(3)** parts with 9 chapters as shown below:

PART I : Introduction and Background

1. Chapter 1: Introduction
2. Chapter 2: Literature Review
3. Chapter 3: Historical Background of Malay Language
4. Chapter 4: Issues in Translation between the English, Malay and Arabic languages

PART II : Results and Finding

5. Chapter 5: Malay Translated Qur'an Corpus Development
6. Chapter 6: Building Malay Ontology for Knowledge Retrieval
7. Chapter 7: Retrieval Evaluation
8. Chapter 8: The Implementation of MyQOS System

PART III : Future Work and Conclusion

9. Chapter 9: Contributions and Future Works

Part I provides the relevant background information. Following the introductory chapter, Chapter 2 contains the literature reviews discussing about IR, recent methodologies in IR, Semantic web technology, Natural language processing in general, and Malay language perspective. Relevant historical background to the Malay linguistic tradition is discussed in Chapter 3. Issues in translation between Malay, English, and Arabic languages are explained in Chapter 4.

Part II presents the results and the findings of the research. Chapter 5 describes the development of Malay translated Qur'an Corpus. This chapter also discusses the NLP pipeline processes involved in developing Malay Translated Qur'an corpus, which started with data collection and preparation. Chapter 6 presents the development of the ontology of concepts derived from Malay Translated Qur'an books. Detailed evaluations of the proposed model and its extensions are reported in Chapter 7. Chapter 8 demonstrates further extensions for the Web environment. The prototype contains the keyword-based search, semantic search, and question-answer search. Besides, it also provides Malay language resources for other researchers who want to embark in this research area.

Part III concludes the thesis findings. The last chapter summarizes the main contributions and presents the recommendations for future research. This chapter 9 concludes with a discussion of the challenges and limitations of the work as well as its implications for theoretical and computational linguistics.

Chapter 2

Literature Review

This chapter covers the background areas and related work necessary to understand the contributions of this thesis. Firstly, we introduce the background of the Semantic Web and its fundamental theory in Section 2.1. After that, we give an overview of what the ontologies are, their underlying formalism, semantic search, and their representation languages. Then, we present a summary of query languages and provide some glimpses of the ones that are relevant to our work. In Section 2.2, we presented the overview of Information Retrieval and its underlying idea. We also focus on how semantics was incorporated into information retrieval search systems. Later, we introduce in Section 2.3 the general background and approaches developed in the last decade of Natural Language Processing (NLP). We include the example of the previous research that is related directly to the work presented in this thesis in Section 2.4. We conclude in Section 2.5 with an outlook for this research area, in particular, our view on the potential directions ahead to realize its ultimate goal: creating and improving the quality and accuracy of the query mechanism to retrieve information embedded in the Malay text of the Qur'an translation.

PRESENTED : Ahmad, ND, Bennett, B and Atwell (2016). Semantic-based Ontology for Malay Qur'an Reader. IMAN'2016 4th International Conference on Islamic Applications in Computer Science and Technologies. 20-22 Dec 2016, Khartoum, Sudan.

2.1 Semantic Web Technology

Semantic web technology development is closely related to the World Wide World (WWW). Both Semantic Web and World Wide World have the same inventor, Sir Tim Berners Lee, and are designed to serve almost the same purpose, to have knowledge widely accessible by enabling applications for searching, browsing and sharing of that knowledge [Hitzler et al., 2009]. However, the Semantic Web was designed as a new generation of the current web, where data is given clear meaning to enable computers to easily work together [Fazzinga and Lukasiewicz, 2010]. The remaining section will describe the semantic web in detail.

2.1.1 Semantic Web

Semantic Web is described as a web of linked data. [Shadbolt et al., 2006] define the semantic web as "an extension of the current version of the web where information is given well-defined meaning, enabling computers and people to work in co-operation". It is designed to overcome the challenges of the current web search systems, which are mainly designed for presenting, organising, and linking data in the form of text, video, audio, and images. The structure of data on the current www is published in such a way that the data can only be understood by humans, computer programs cannot understand its meaning [Moore, 2012]. Web search systems struggle with the aggregation and querying of information without having a consistent way of achieving such tasks, whereas the semantic web concept enables linked documents on the web and assigning better meaning for both human and computer understanding. In other words, it improves the current web structure form of interconnected documents to semantically driven documents that allow better aggregation of information, storage, manipulation and retrieval [Kuck, 2004]. This provides a better enabling environment for promoting good working relationships between humans and computers. The semantic web will be better understood by looking at a graphical representation of the semantic web architecture in figure 2.1 below.

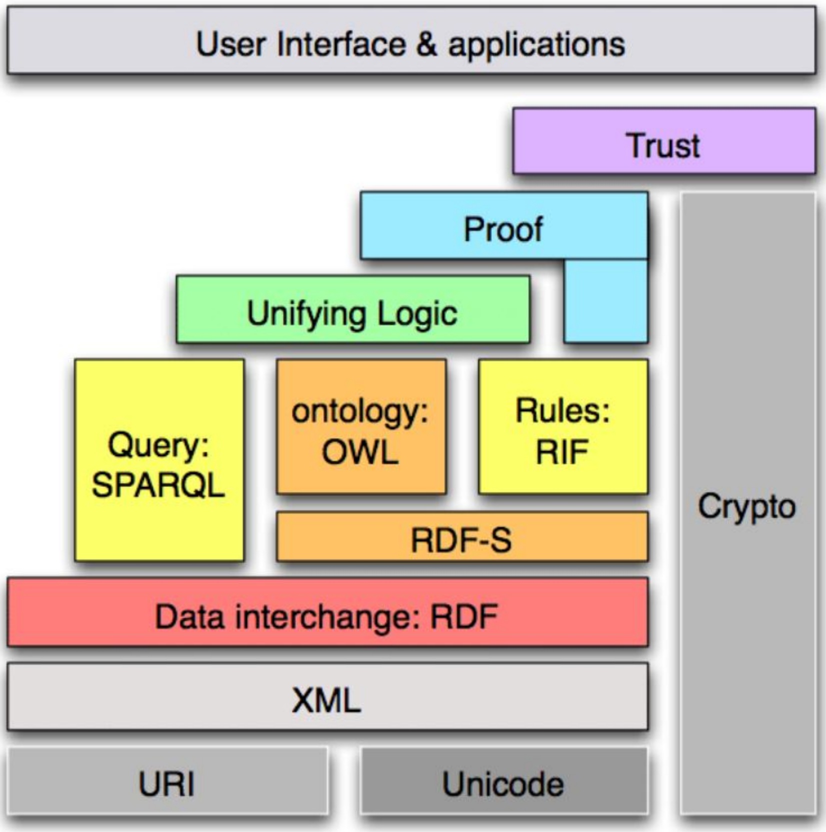


Figure 2.1: Semantic Web Architecture [S. Dandagi and Sidnal, 2016]

Figure 2.1 represents the main architecture of the semantic web, where the whole semantic web. The first layer in the semantic web is the Uniform Resource Identifier (URI). The concept of URI is based on the features of the WWW, where it is the standard form for identifying documents on the web. URI allows the unique tagging of a web document used to uniquely identify the document on the web. The second layer of semantic web architecture is XML, which stands for Extensible Mark-up Language. XML ensures a common syntax is used on the semantic web. XML contains elements that are nested and have attributes and content. It involves an XML name space which allows the specification of various vocabularies in one XML document. An XML schema serves for expressing the schema of the XML document. The main data format of the semantic web is

Resource Description Framework known as RDF. RDF is a data representation format that allows resources or documents to be represented in a graphical format.

Ontology or Web ontology Language (OWL) is a W3C recommended knowledge representation framework for describing web resources. SPARQL Protocol and RDF triple format are W3C standard RDF querying languages. They are the recommended languages for querying semantic web data. SPARQL is a SQL-like query language that uses the RDF Query triple data format. SPARQL is a formal language that matches the user query with the underlying data in the RDF structure to retrieve an effective result. All semantic and rules are executed at the layers below the proof and the result, which are used to prove the deduction. The proof layer presents justification of an inference made, giving a logical ground for the inference. Once a basic logic and proof is set up, there is an environment of trust for conducting transactions. Cryptography is then used for validation and authentication. Finally, the user interface provides the semantic technology to the user.

2.1.2 Web Ontology Language

The main building block of the semantic web is ontology, which transforms web content into a machine-readable format [Ahmed and Gerhard, 2007]. In 1998, [Studer et al., 1998] defined an ontology as a formal, explicit specification of a shared conceptualization. In other words ontology can simply be seen as the study of entities that exist in the real world, and the things they have in common [Lawson, 2004]. Ontology facilitates standards for integrating and sharing data in a conceptual schema. Objects, entities or concepts are identified and annotated with the relationships that exist between them.

In the concept of ontology, an entity or object is referred to as the same thing. This research will be using 'concept' to denote an entity or object, while 'relationship' is considered, the thing's concepts have in common, known as properties. Properties can be classified into object properties and data properties. Object properties represent the semantic relationship between concepts, while data properties define the relationship between a concept and its literals. Annotation of concepts enables a better description of the concepts in the form of metadata, facilitating greater meaning for human and machines to easily process and share.

properties.png properties.bb

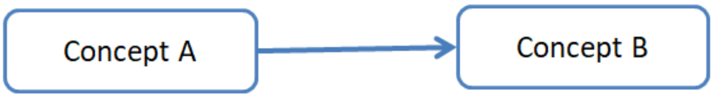


Figure 2.2: Example of Object Property

Figure 2.2 is a graphical representation of an object property, where a semantic mapping between two concepts (Concept A and Concept B) is provided. The semantic mapping gives a better description of the concepts.

properties.png properties.bb

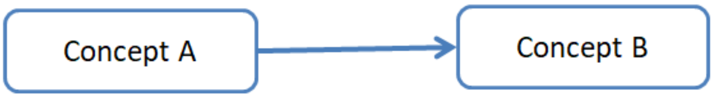


Figure 2.3: Example of Data Property

Figure 2.3 is a graphical representation of the data property, where a concept is mapped to its literal. A literal is a mechanism for describing a concept itself. It gives an additional description of a concept. Figure 2.4 denotes an example of ontology representation for students.

Ontologies can be created in three ways such as automatic, semi-automatic, or manually [Erdmann et al., 2000]. Automatic creation of an ontology involves using an automated tool to automatically generate the ontology of a domain [Balakrishna and Srikanth, 2008]. Semi-automatic ontology creation involves a combination of human effort and automated tools [Balakrishna et al., 2010]. Manual ontology is usually complex and time consuming, especially when dealing with a huge data [Ahmed and Gerhard, 2007]. Manual ontology creation involves the design and creation of an ontology which is completely by a human expert [Tao et al., 2009].

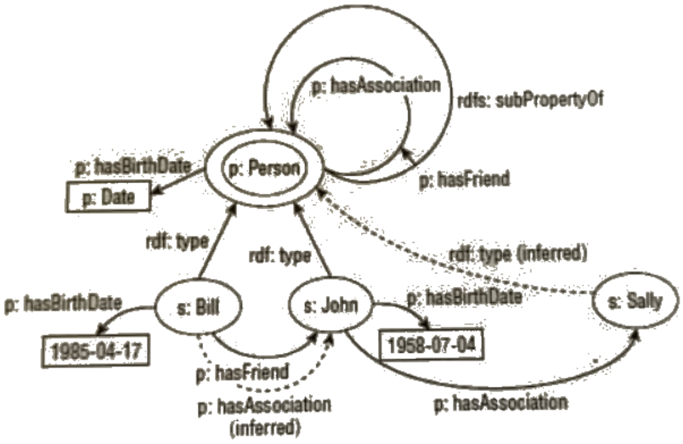


Figure 2.4: Example of Student Ontology Representation

The Semantic Data Model

In the semantic web, data is represented in a formal ontology format. The Ontology model data is in the form of concepts and relationships between concepts. In the semantic web, these concepts and their corresponding relationships are represented in RĐF graphical format. RDF is World Wide Web Consortium's (W3C) standard syntax for representing concepts and relationships between concepts in a graphical format. The World Wide Web Consortium is a group that is working to set the standard format for the semantic web. They provide representations that model the ontology into machine-understandable formats, known as RDF for computer applications to use when making inferences [Tauberer, 2008]

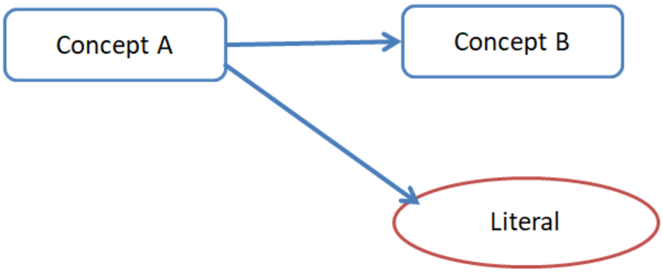


Figure 2.5: Example of an RDF Graph

Figure 2.5 shows an example of an RDF graph, with a directed graph connecting a concept A, concept B, and its literals in a format called 'triple', Concept A, predicate, Concept B. The main building block of the semantic web is RDF statement [Patel-Schneider, 2005]. It is presented in triple as below:

<Subject> <Predicate> <Object>

Subject and object represents an ontology concept, while the predicate represents the relationships between these concepts. RDF triple format is mainly a practical rule language for computers to understand, manipulate, and share data [Decker et al., 2005].Therefore, an RDF graph shows data represented in a format by which computers could understand the meaning and processes. RDF triple formats are stored in a knowledge base, which enables computers to access and manipulate the data.

A knowledge base is a repository that contains data derived from machine-readable formats such as RDF triple [Amsler, 1984]. A knowledge base could be a database where knowledge is stored and manipulated. An example of knowledge based is DBpedia. A knowledge base is a large set of linked data that consists of RDF triples extracted from various sources such as Wikipedia's information boxes, categories, and internal links, among others [Auer et al., 2007]. RDF triples stored in a knowledge base includes knowledge derived from the annotations of ontology concepts, which adds more meaning to the data. Computers can understand the information stored in a knowledge base and therefore applications can be developed to semantically search it. Since information in a knowledge-based is semantically mapped and stored, the search must be done semantically to access such information, known as a semantic search. We will examine the semantic searches in detail in the next section and describe how they are used to access information stored in the knowledge base.

2.1.3 Semantic Search

Semantic searches are a semantic web technology approach to interpreting search queries and resources based on underlying ontologies, labelling some contextual domain knowledge, by connecting web resources to semantic annotations. In other words, it is an attempt to transform the structure of a query to the same structure that can query the knowledge base and other semantically annotated data stores. Data represented in RDF triples, containing structurally annotated ontology concepts, needs the same structured query format to query it. Data is semantically mapped and stored in the knowledge base for further use. To manipulate and access such data, a semantic search mechanism must be put in place. In other words there is a need for an application that will match RDF triple form patterns for retrieval [Fazzinga and Lukasiewicz, 2010].

A semantic search is a data retrieval mechanism that integrates the capabilities of the semantic web and search engines to get more precise results than the current search engine. A semantic search enables computers to think, reason, manipulate data, and provide humans with the information they need in the way that they need it [Kassim and Rahmany, 2009]. It uses semantic web technology to manipulate and interpret a user's natural language queries and match them against information in the knowledge base to extract semantic knowledge [Fazzinga and Lukasiewicz, 2010]. This enables computers to accept complex queries, use semantically annotated documents, make reason and inferences, and finally, present good results to the user.

RDF is a tool for the semantic representation of knowledge. RDF is a simple data model, which describes the objects (resources) and the relations among them with a triple syntax (subject-predicate-object). A knowledge base consisting of these RDF statements is essentially a labelled and directed graph with the nodes being resources while the edges represent the properties. By describing the data in RDF or OWL format, the Semantic Web allows more intelligent search engines to be developed. These search engines can use the metadata associated with the entities to improve search quality. Semantic relations defined in ontologies allow very complex queries to be answered which are not possible otherwise.

The main basic concept of semantic search is that it semantically manipulates and transforms a natural language query to structured formal queries such as Protocol and RDF Query Language (SPARQL), RDF Data Query Language (RDQL) or Sesame RDF Query Language (SeRQL) [Habernal, 2012]. These structured queries enable users to retrieve data from knowledge base and other knowledge base sources.

2.1.4 Representation Language

Ontologies are used to model real-world entities and relations among them in a taxonomic structure. They are nowadays the backbone for the Semantic Web applications. Once we have selected what to model (ontology) and the underlying formalism (DLs) to express it, we must implement it in a representation or implementation language. In this section, we present the most important languages adopted and used in the context of the Semantic Web: RDF (along with RDF-S) and OWL.

RDF (Resource Description Framework) is a language for representing information about resources on the World Wide Web. At first, it was intended for serving metadata (title, date of creation, authorship, etc.) about Web documents; however, by generalizing the notion of resource, it can be used to represent information about anything that can be identified in the Web by a URI (Uniform Resource Identifier). RDF Schema (RDFS) was the first attempt towards developing an ontology language, and it became a W3C recommendation in 2004. RDF-S was built upon RDF. It extends the RDF vocabulary with additional classes and properties such as rdfs: Class and rdfs: subClassOf.

The latest W3C recommendation for ontology languages is the Web Ontology Language (OWL). OWL is based on the DL SROIQ(D). OWL further extends

RDFS by providing additional features such as cardinality constraints, equality, disjoint classes, efficient reasoning support, and much more. OWL language has three sub-languages, which are OWL-Lite, OWL-DL and OWL-Full. OWL-Lite and OWL-Full are not widely used, because the former is too restricted and the latter does not guarantee practical reasoning. OWL-DL provides maximum expression with complete and decidable reasoning support. OWL ontologies have five elements: Individual, concepts (or classes), datatype (or domains), object properties, and data properties. Essentially, concepts are sets of individuals, the datatype is sets of values defined over a specific field (such as integers or dates), object properties are binary relations between individuals, and datatype properties relate individuals and datatype.

There are differences between OWL and RDFS in terms of the ability to add semantics to the RDF. In general, the features of OWL enable one to add more semantics in this regard. According [Martin et al., 2015], when comparing OWL to RDF, we need to express in detail the knowledge. For example, every person has exactly one birth date, or that no person can be both male and female at the same time.

In this thesis, we used OWL to represent all information of the ontology. Using OWL language, it easy to generalize each element of the concept and make the relationship between concepts and instances.

Query Language

In Computer Science, query languages are computer languages that are used to query databases and information systems. Depending on their formality degree, we can classify query languages as informal or formal ones. Informal query languages are more related to information retrieval tasks, where the semantics of the query is not formally defined. Users express with these query languages their information needs, so these languages imply an intermediate step to establish their semantics (query construction) and adapt the query to the underlying data model. On the other hand, formal query languages have their semantics strictly defined, and users express with them queries with an unambiguous interpretation.

1. Informal Query Language

Keyword query language and natural languages query fall into this informal query language group. It does not involve any semantic element in the query.

(a) Natural Language (NL)
Natural Language query enables users to express their information needs in their language. Its ease of use makes it always a possible choice for casual users [Alagha and Abu-Taha, 2015], but processing it correctly is still an open problem. Among others, NL as query language faces the following issues such as word ambiguous and language dependent. Usually, the meaning of the query word depends on user needs. However, sometimes it impossible to interpret the intention or purpose of the query correctly. Another problem with NL is language dependent. Typically, the techniques applied to process NL depend on the language itself. It varies from language to language. For instance, the Malay language is an agglutinative language with rich morphology; morphological analysis is quite difficult to implement. So, processing the Malay text will be quite challenging compared to other languages.

(b) Keyword Query Language
Keyword queries are like to the natural language query. It consists of a set of keywords that represents the user's need. Formerly, many web search engines using this technique to query information in the document. Moreover, users have found it easy and quick to express their information needs using this technique. However, the ease of use of keyword search comes from the simplicity of its query model, whose expressively is low compared with other more complex query models [Ahmad et al., 2017]. The keyword queries are in fact projections of the user's actual information need. This leads to a much more ambiguous context compared to the natural language query. However, despite its inherent ambiguity, keyword-based search interfaces have been adopted by different information systems other than Web search engines as the benefits that they provide in terms of user-friendship and language independence are worthy enough to do so. In this thesis, we aim at overcoming their drawbacks with the help of semantic techniques.

2. Formal Query Language

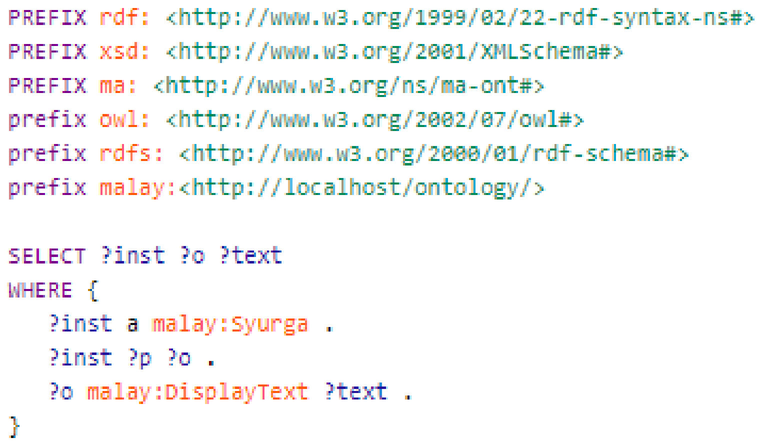
   The formal query language is a language which expresses the information need unambiguously. It contains the semantic element in the query. SPARQL and SQL-like language are the types of query language used in this thesis.

   (a) SPARQL Query Language

   The SPARQL query language is the W3C recommendation for querying RDF documents. It resembles SQL (Structured Query Language) in its syntax, although their queries are expressed in terms of pattern matching over the RDF graphs, instead of dealing with relational tables. There are four different types of SPARQL queries:

   - **SELECT**, which retrieves the answer to a query pattern directly. This answer is modified to the variables used and their binding to the answer value.
   - **CONSTRUCT**, which constructs an RDF graph specified by a template (also in terms of query patterns).
   - **ASK**, which checks whether a query pattern has a feasible solution.
   - **DESCRIBE**, which, given a resource (URI), returns an RDF graph containing information about it.

   The SPARQL query consists of conjunctions and disjunctions of triple patterns like RDF triples. A simple SPARQL query to search *"Syurga"* is shown in Figure 2.6. Despite its simplicity, the usability of SPARQL is limited for the end-user. First, formulating a query requires considerable time and effort even for the simplest query. Secondly, the domain knowledge is required, i.e. the exact names of classes and properties need to be known in advance.



```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX ma: <http://www.w3.org/ns/ma-ont#>
prefix owl: <http://www.w3.org/2002/07/owl#>
prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
prefix malay:<http://localhost/ontology/>

SELECT ?inst ?o ?text
WHERE {
   ?inst a malay:Syurga .
   ?inst ?p ?o .
   ?o malay:DisplayText ?text .
}
```

Figure 2.6: An Example SPARQL Query to Search *"Syurga"*

As we can see in the examples, SPARQL queries resemble SQL queries where the schema has been blurred to the limit of representing everything via RDF tuples. This, along with the use of predefined schemas, provides us with higher levels of flexibility to achieve schema and data interoperability, among other benefits. In this thesis, we used SPARQL query to query OWL file in the semantic search engine.

(b) SQL-Like Language

The notion of SQL-like language is quite broad. SQL-like languages are languages whose syntax resembles the syntax of SQL, which is the most extended language for querying and managing data stored in relational databases. SQL has its formal foundations on relational algebra and consists of a data definition language (to create and maintain the database schema) and a data manipulation language (to insert, update, and query the data in the database). Focusing on SQL as a query language, and independently of the underlying schema, a SQL query has the following general structure:

SELECT [list of attributes]
FROM [list of tables]
WHERE [list of condition]

where,

- List of attributes is the list of attributes/ entities that have to be retrieved as an answer for the query.
- List of tables is the list of the tables that are involved in the query.
- List of conditions is the list of the conditions that the attributes must meet to form part of the answer. This condition can include a different kind of relational operators between tables such as the different kinds of JOIN operators that exist.

Description Logics

Description Logics, languages known as DLs are formal languages for representing knowledge and reasoning about it [Gruber, 1993]. DLs are formed by an intentional layer T called TBox and an extensional layer A called ABox. The TBox is composed of a set of terminological axioms. Axioms are formulas of the form $C \equiv D$ or $C \vee D$, where C and D are concepts. An axiom of the form $C \equiv D$ says that concepts C and D are equivalent, that is, an individual that belongs to C also belongs to D, and vice versa. An axiom $C \equiv D$ is called a concept definition if the left hand of the axiom is a concept name. Axioms of type C v D represent D subsuming that concept C, i.e., any individual in C is in D, but not necessarily vice versa. Concepts are formed using:

1. A set of concepts names N_c conceptualizations of a set of individuals (or instances), for example, Person, Plant, or Animal.

2. A set of roles N_r, which are binary relations between individuals, for instance, hasParent, MentionedIn, or hasGrow

3. Constructors to define new concepts, such as ∪ , ∩, ∃, and ∀. For example, given that we have as concepts Person and Plant, and hasGrow as a role, we can define a new concept to represent people who grow a plant as Person ∩ ∃ hasGrow.Plant.

A general TBox is a finite set of axioms. An example of TBox expressing that Human, LivingCreation and Mother are a woman who have children is:

T = {Human $\vee$ LivingCreation; Mother$\equiv$ Woman $\cap$ $\exists$ hasChild.Human}

An ABox is a set of assertions that describe a specific state of the world represented by the associated TBox. We can express with assertion that Diana is Humaira's mother:

A ={Woman(Diana); Human(Humaira); hasChild(Diana, Humaira)}

DLs are logic and provide a formal framework where rules of inference can be applied to deduce automatically new knowledge from TBoxes and ABoxes. This is done using reasoners, which are programs that can perform several different reasoning tasks over ontologies.

In this thesis, we have used HermiT 1.3.8 which supports OWL, the representation language proposed by W3C for ontology specification, which is overviewed in the following subsection. However, Protégé cannot load a big file of RDF triples. Therefore, we have to store the ontology using Apache Jena with Fuseki server. Jena rules are used to create semantic rules.

2.2 Information Retrieval

Information Retrieval (IR) is a wide area in Computer Science concentrated on providing users with easy access to information of their interest. The goal of the IR discipline is to retrieve information that is useful or relevant to the user. Therefore, a good IR system must interpret the contents of the information items and rank them according to the degree of relevance to the user query. The interpretation of a document context involves extracting syntactic and semantic information from the document text and using the information to match the user information needs. Currently, a significant percentage of information overload is derived from many online and offline sources. Traditionally, finding interesting information in the evolving contextual textual data can be achieved through IR technology and systems. Given a query, the IR system returns a list of potentially relevant documents which will then be scanned by the user to find exactly what he/she wants [Manning et al., 2008].

In the computer-centered view, IR mainly focused on indexes, processing user queries with high performance and accuracy, and developing ranking algorithms to improve the results. Nowadays, research in IR includes modelling, web search, text classification, language and text processing, system architecture, user interfaces, and data visualization. The new embarked in IR and language processing enables the research in this area more diverse. Many researchers start to do research in these areas, especially focusing on their own native language.

In text retrieval, documents are parsed into words. Information Retrieval process begins when a user enters a query into the system. For example, searching the web for a query "Malay Qur'an" may retrieve many results. In this case, the query will not uniquely identify and retrieve relevant information based on the user needs but also retrieve other irrelevant results. Alternatively, many documents could match the query with different degrees of accuracy.

Most Information Retrieval investigations have focused on retrieval effectiveness, which is usually based on document relevance judgment. However, there are some problems associated with relevance judgments such as being subjective and changeable. For example, different judges will give different relevance values to a document retrieved in answer to a given query [Mitra, 2013].

To evaluate the performance of Information Retrieval systems, many measures have been proposed. These measures require a collection of documents and a query. Every document is simply classified to be either relevant or irrelevant to a particular query. The most used measures are recall and precision. Recall in Information Retrieval is the ratio of retrieved and relevant documents for a given query over the total number of relevant documents for that query in the database. Except for small test collections, the result is generally unknown and should be estimated by sampling or some other method. Precision in Information Retrieval is the ratio of retrieved and relevant documents over the total number of documents retrieved. F1 score in Information Retrieval is used to measure the accuracy by considering both the precision and recall. The results for recall, precision, and F1 score values are between 0 and 1.

Precision is a good measure to determine, when the costs of False Positive (fp) is high. For instance, email spam detection. In email spam detection, a false positive(fp) means that an email that is non-spam (actual negative) has been identified as spam (predicted spam). The email user might lose important emails if the precision is not high for the spam detection model. Recall calculates how many of the Actual Positives our models capture through labeling it as Positive (tp). Applying the same understanding, we know that Recall shall be the model metric we use to select our best model when there is a high cost associated with False Negative. For instance, in fraud detection. If a fraudulent transaction (Actual Positive) is predicted as non-fraudulent (Predicted Negative), the consequence can be very bad for the bank. F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. Intuitively it is not as easy to understand as accuracy, but F1 is usually more useful than accuracy, especially when an uneven class distribution. Accuracy works best if false positives and false negatives have similar cost. If the cost of false positives and false negatives are very different, it is better to look at both Precision and Recall.

$$Precision = \frac{tp}{tp + fp} \tag{2.1}$$

$$Recall = \frac{tp}{tp + fn} \tag{2.2}$$

$$F1Score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{2.3}$$

2.2.1 The Processes of Information Retrieval

Most of the applications dealing with IR are based on textual data. When dealing with textual IR, several textual operations and natural language processing are involved in the retrieval steps. Figure 2.7 shows the architecture of textual queries typically performed by an IR engine. The process involved in the architecture are used to retrieve an information. Firstly, the user need is specified via the user interface, in the form of a textual query or keyword. Then, the query is parsed and transformed by a set of textual operations such as pre-processing. Query operation further transforms the pre-processing query into a system-level representation.

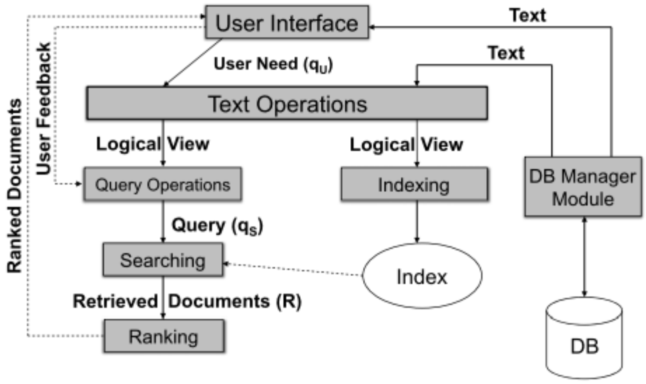


Figure 2.7: Architecture of a textual IR system [Ceri et al., 2013]

User Interface

User interface is one of the most important aspects in IR. The good design of the user interface increases user involvement, perfect functionality, and creates a strong link between user and application. Simple interfaces are easier to use at the cost of resulting unclear queries, while the complex interface is more powerful and usually provides a more detailed and precise query formulation. There are two user interface methods that are widely used today; traditional IR and semantic retrieval. Keyword-based, natural language-based, form-based, graph-based, and image-based interfaces are some of the commonly used interface methods in the literature. In this thesis, we use and compare the retrieval results based on keyword-based and semantic-based interface.

Text Operations

The text operation mainly to reduce the complexity of the document representation and increase the retrieval performance. It is usually happened during the preprocessing phase. There are several useful text operations that can be performed, such as lexical analysis, elimination of stop words, stemming, thesaurus construction, text clustering, text compression, and encryption. Nowadays, there is controversy regarding the potential improvements to retrieval performance generated by stop word elimination, stemming, and index term selection. For instance, the most frequent words appeared in the Malay Translated Qur'an are stop words. Most of the stop words in each text are connection parts of a sentence rather than showing the subject, object, or intent. By eliminating stop words will reduce the complexity of the document and increase the retrieval performance. This research will prove that the use of text operation in preprocessing phase will improve the retrieval performance.

Query Operations

A query in general terms is a statement or series of statements made by a user to a retrieval system for specifying what information is to be retrieved and in what form. Initial query submitted by the user is never good enough to directly fetch the documents. It needs to be expanded and processed before searching. Usually, the query is transformed into an internal form that the system can interpret. The transformation involved several processing tasks such as stop word elimination, stemming, and other application-specific tasks.

Indexing Process

Indexing is an important part in IR systems. The main goal of indexing is to optimize the query performance and improve the response time by storing the query terms in an inverted file structure called inverted index. It stores the text positions for the occurrence of each term. The indexing process involves three basic stages; defining the data sources, transforming the document content to produce a logical view, and building an index of the text on the logical view [Ceri et al., 2013]. Several preprocessing tasks are carried out during the indexing phase like the query and text operation phase, which further improves the performance.

Searching

When the query is submitted by the user, the query term is searched with the inverted index. The documents holding the occurrences of the query terms are retrieved. Inverted indexes are unrivalled in terms of retrieval efficiency because it uses the same term generally occurs in several documents. This can reduce the storage or database requirements. The simplest type of search is that for the occurrence of a single word. This is straightforward using an inverted index. If the query has more than one word, Boolean queries need to be performed. There are two cases of Boolean queries; conjunctive (AND operator) and Disjunctive (OR operator) queries. Conjunctive queries imply searching for the words in the query. Obtaining one inverted list for each word, disjunctive queries imply to search for all words in the query and to gather one inverted list per word [Baeza-Yates and Ribeiro-Neto, 2011].

Ranking

When the document is obtained in the search process, the results are ranked based on co-occurrence of query terms. The document is sorted according to the score, so that the most relevant documents are presented first to the user. Ranking process is highly dependent on the IR model. We will explain in the following section some IR models aimed at producing a ranking function. Finally, the choice of evaluation metrics will be done and it is critical since only a few of them can be used in the ranking information to determine the evaluation score for each retrieval performance.

Content in text information retrieval can be represented according to two bask approaches: the system may search the content represented in the document using natural language text, or using a specific representation language, where the document and queries need to be mapped. An example of language representation is the languages developed in the framework of the semantic web for semantic search, such as the Resource Description Framework (RDF). Language representation enables retrieval systems to incorporate semantics in the information retrieval process. Incorporating some form of semantic processing in information retrieval improves retrieval effectiveness beyond that possible using keywords alone [Rindflesch and Aronson, 1993, Hersh et al., 1992].

In the next section, we will look at semantic information retrieval in detail. We will give comprehensive details about what is semantic in information retrieval and the state of the art in incorporating semantics in information retrieval.

2.2.2 Semantics Embedded in Information Retrieval

Information retrieval technology has contributed greatly towards the success of the web. The concept of web-based indexing and search mechanisms such as Google and Yahoo have profoundly changed the way we access information. The search processes that these search engines use are primarily based on the syntactic matching of the document terms and query representations. The results achieved by these search engines are negatively affected by the problems of natural language ambiguity, however, such as that of polysemy where the same word may have multiple meanings, or synonymy where two or more words may have the same meaning.

Another problem of syntactic searching is that it struggles with queries in phrases which may be represented by complex concepts. Syntactic search approaches do not look at semantic relationships between concepts and queries either. Syntactic search is based on traditional information retrieval systems that possess some inconsistency between the vocabulary of user queries and data representation, and is based on simple keyword matching that exploits the frequency of co-occurrence of terms [Rosario, 2000].

With the development of the semantic web, semantics was incorporated into information retrieval search systems. The concept of semantic IR is centered on retrieving documents and query representations based on semantic analysis of their contents by using natural language processing techniques and retrieving documents by matching these semantic representations [Kharkevich, 2010]. Information retrieval techniques are incorporated in such a way that they can handle semantically annotated data, where systems can make reason and inferences about user queries. This will enable users to retrieve indexed RDF data using the concepts and various properties they possess [Mayfield and Finin, 2003].

The main purpose of both information retrieval technology and semantic web technology is to create a broad network of distributed resources where different multimedia documents are searched using different languages within various structured and indexed data collections. To achieve this, there is a need for a mechanism that can search various repositories despite variations in protocol, format, content, and meaning [Börner, 2003]. In semantic information retrieval systems, searching is performed by interpreting the semantics of keywords, i.e. the meanings of keywords. Information retrieval systems incorporated with such mechanisms have higher precision and recall than those IR systems that are using keyword-based information retrieval approaches, because of the semantics of the keywords [Mustafa et al., 2008]. Semantic information retrieval enables users to retrieve semantically annotated documents such as RDF schema, XML and other structurally formed documents. With the support of the right vocabulary, ontologies, and natural language processing, the semantics incorporated in information retrieval will create a higher level of search retrieval with distinct meaning by involving data and reusing such data to connect across many documents on the web. Ontology provides good metadata schemes that machines read and understand, retrieval is made more effective with less user effort [Schiff, 2011]. Users can retrieve answers that are more effective and efficient because the answers are returned based on computer reasoning and inference from collections of more described data. There is a large amount of literature on semantics incorporated in information retrieval systems, mainly to improve the quality of the results of syntactic approaches.

The idea of incorporating semantics into retrieval systems came into existence at the time when the IR community proposed approaches to extending the classical IR with explicit semantics [Kharkevich, 2010]. Since its inception, different approaches have been proposed by the IR community to deal with the problem of natural language ambiguity which was lacking in traditional IR systems.

2.3 Natural Language Processing (NLP)

Natural language processing (NLP) is a field of computer science, artificial intelligence, and computational linguistics concerned with the interactions between computers and human languages. Natural language processing systems take strings of words (sentences) as their input and produce structured representations capturing the meaning of those strings as their output [Eggebraaten et al., 2014]. There are two problems in processing natural language. These problems are the level of ambiguity that exists in natural languages, and the complexity of semantic information contained even in simple sentences. The ambiguity of the query and returned results can be one of the major constraints. This ambiguity usually happens in understanding the actual meaning of the user's query and giving the right results. Dealing with the natural ambiguity of language, such as polysemy and synonymy, the IR community moves from words that are expressions of natural language to concepts expressed in an ambiguous format in the form of formal language such as ontology, which is generally described as word sense disambiguation.

With the emergence in the 1970s of models of ranked retrieval that process unstructured queries, automatic query systems became a fact. The central philosophy of automatic query systems is that indexing and query formulation should result in a representation that is closer to the actual meaning of the text, ignoring as many of the irregularities of the natural language as possible. A typical approach to indexing and query formulation selects the query terms as follows. First, a tokenization process occurs, then the stop words are removed, and finally the remaining words are stemmed. Additionally, natural language processing modules might provide the identification of phrases or splitting of compounds.

2.3.1 Tokenization

As a first step in processing a document or a query, it has to be determined what the processing tokens are. One of the simplest approaches to tokenization defines word symbols and inter-word symbols. The purpose of tokenization is to split a text into its basic tokens. In the example below, characters that are no letters and no digit are inter-word symbols. The inter-word symbols are ignored during this phase, and the remaining sequences of word symbols are the processing tokens. As a result, it is not possible to search for punctuation marks like, for instance, hyphens and question marks.

- **Surah Al-Baqarah,30: Text before tokenization**
  Behold, thy Lord said to the angels: "I will create a vicegerent on earth." They said: "Wilt Thou place therein one who will make mischief therein and shed blood? - Whilst we do celebrate Thy praise and glorify Thy holy (name)?" He said: "I know what ye know not." (Surah Al-Baqarah, verse 30).

- **Surah Al-Baqarah,30 : Text after tokenization**
  Behold thy Lord said to the angels I will create a vicegerent on earth They said Wilt Thou place therein one who will make mischief therein and shed blood Whilst we do celebrate Thy praise and glorify Thy holy(name He said I know what ye know not (Surah Al-Baqarah, verse 30).

In this thesis, tokenization is required to gather the query terms. However, since the Malay language is an agglutinative language with rich morphology, morphological analysis is quite difficult to implement. One challenge of Malay language that is hardly found in other languages is reduplication. Reduplication is a word-formation process in which meaning is expressed by repeating all or a part of a word. This reduplication is widely used in many Malay texts. Generally, it can be divided into full, such as *"perempuan-perempuan"* "girls" (from *"perempuan"* or "girl"), or partial, such as *"lelaki"* "boys" (from *"laki-laki"* or "boys") or rhyming and chiming, such as the word *"kayu"* "wood" combines with *"kayan"* "wood" to form *"kayu-kayan"* different sorts of wood.

According to Malay linguist Don, Z. M. in [Mohd Don, 2010], only the first word which is considered as the query term and the second word should be removed. However, in this research, the second words will not be discarded as some of the second words are query terms. Considering this scenario, the reduplication process is done in the semi-automatic process. Below is an example of reduplication appeared in Malay translated Qur'an and English translated Qur'an.

- **Malay Translated Qur'an:**
  *"Dan apabila mereka bertemu dengan **orang-orang** yang beriman, mereka berkata, "Kami telah beriman." Tetapi apabila mereka kembali kepada **syaitan-syaitan**(para pemimpin) mereka, mereka berkata, "Sesungguhnya kami bersama kamu, kami hanya **berolok-olok.**"*;(Surah Al-Baqarah, verse 14)

- **English Sahih International:**
  "And when they meet those who believe, they say, "We believe"; but when they are alone with their evil ones, they say, "Indeed, we are with you; we were only mockers." (Surah Al-Baqarah, 14)

2.3.2 Stop word Removal

Stop word is words with little meaning that are removed from the index and the query. Words might carry little meaning from a frequency (or information theoretical) point of view, or a linguistic point of view. Removing a stop word for linguistic reasons can be done by using a stop list that enumerates all words with little meaning, like, for instance "the", "it" and "a". These words do also have a high frequency in English, but most publicly available stop lists are, at least partly, constructed from a linguistic point of view. However, every language has its stopword. The same goes to the Malay language.

For instance, *"yang"*, *"dan"*, and *"mereka"* were the highest stop words appeared in the Malay Translated Qur'an document. By removing the common words, the document scores will not be affected that much. Stopword removal by their frequency can be done quickly by removing the 200-300 words with the highest collection frequencies. For instance, the example below shows before and after the process of elimination of a stopword in Malay text. As we can see, the total number of words before the removal is 30, however, after the withdrawal of the stopword, only 12 are left.

- **Surah Al-Baqarah,14: Text before the elimination of stopword Malay Translated Qur'an:**
  "Dan apabila mereka bertemu dengan orang-orang yang beriman, mereka berkata, "Kami telah beriman." Tetapi apabila mereka kembali kepada syaitan-syaitan(para pemimpin) mereka, mereka berkata, "Sesungguhnya kami bersama kamu, kami hanya berolok-olok.";(Surah Al-Baqarah, verse 14)

- **Surah Al-Baqarah,14: Text after the elimination of stopword Malay Translated Qur'an:**
  bertemu orang-orang beriman, berkata, beriman kembali syaitan-syaitan para pemimpin berkata bersama berolok-olok (Surah Al-Baqarah, verse 14)

2.3.3 Stemming

Stemming is a computational process of reducing a word from its derived form into its root term. Stemming is used in many information retrieval systems to reduce different word forms to common roots. Recognizing searching and retrieving returns more results. This is one of the reasons why this process is needed to integrate search queries and get information. In the stemming algorithm, words with the same root are reduced to a standard form by stripping each word of its derivational and inflectional suffixes.

The advantage of stemming is the indexing time becomes increasingly efficient and index file compression faster, since the index terms are already derived, this operation requires no resources at the search time, and the index file will be compressed. However, the disadvantage of indexing time stemming is that information about the full terms will be lost, or additional storage will be required to store both the stemmed and unstemmed forms.

There are several criteria for judging stemmer: correctness, retrieval effectiveness, and compression performance. Two ways are stemming can be incorrect, which are over-stemming and under-stemming. When a term is over stemmed, too much of it is removed. Over stemming can cause unrelated terms to be conflated. The effect on IR performance is the retrieval of irrelevant documents. Under stemming is the removal of too little of a term. Under stemming will prevent related terms from being conflated. The effect of under-stemming on IR performance is that relevant documents will not be retrieved. Stemmer can also be judged on their retrieval effectiveness, which is usually measured with recall and precision, and on their speed and size. Finally, they can be rated on their compression performance. Stemmer for IR is not usually judged by syntactical correctness, although the stems they produce are generally very similar to root morphemes.

The stemming process is quite complicated for the Malay language compared with other languages because of the unique morphological structure. Malay language affixes consist of four different types of verbal elements:

1. Prefix: attaches itself at the beginning of a word. Example: *'bersalah'* which means guilty, start with *'ber'* which is a typical prefix in Malay language. Another prefix appeared in Malay language such as *'per'*, *'mem'*, *'men'*, *'pen'*, *'ter'*, *'meng'*, and *'juru'*.

2. Suffix: attaches itself at the end of a word. Example: *'memaafkan'* which means to forgive, ends with *'kan'* which is a typical suffix in Malay language.

Another suffix such as *‘an’*, *‘i’* and *‘mu’*.

3. Infix: usually located in the middle of word. Example: *‘gerigi’* which means toothed blade, which is derived from the root word *‘gigi’* (teeth).
4. Circumfix: Prefix-suffix pair where more than one affix that is attached to a word at the same time and usually positioned before and after the root word. Example: *‘kerajaan’* which means kingdom, is derived from root word *‘raja’* (king).

2.4 Computational Research on The Qur’an

Over the years, there has been a growing interest in the content of Islamic knowledge among both Muslims and non-Muslims, especially knowledge of the Qur’an. The Qur’an is the main source of knowledge, wisdom, and law for Muslims. Since the first revelation, the Holy Qur’an has been among the most influential books that exist [Abed, 2015].The first revelation from God was revealed to the Prophet Muhammad (may God’s prayers and peace be with him) through Angel Jibreel. The Qur’an is a book that covers a wide range of knowledge.

The Qur’an is divided into 114 chapters (Suras) of varying sizes, where each chapter is divided into verses (Ayahs). There are 6,234 verses in the Qu’ran. The Qur’an chapters (Suras) are classified into Meccan and Medinan. According to Islamic belief, The Meccan suras are chronologically earlier chapters (suras) of the Qur’an that were revealed any time before the migration of the Prophet Muhammad and his followers from Mecca to Medina. While Medinan suras or Medinan chapters of the Qur’an are the latest 24 suras that were revealed at Medina after Prophet Muhammad hijra’ (migrated) from Mecca, the whole contents represented in Qur’an can be derived as in Figure 2.6.

There are many today’s Qur’an applications, like Quran.com, noblequran.com, and many more. These applications provide many features such as text representation, electronic audio, Qur’an translation, Qur’an indexes, and keyword search. Most of this application has its peculiarities and weakness. Table 2.1 summarises the features of the current available Qur’an online application.

Table 2.1: Summary of the features of the current Qur'an online applications

Qur'an Online Applications	**Features**
www.quran.com	It provides a full translation of the Qur'an in many languages such as English, Arabic, Turkish, Indonesia, Malay and many more. It also provides audio, Arabic text representation, tafsir (interpretation) and keyword based searchable interface for the user to find any verses in different languages.
www.noblequran.com	It has many features such as Qur'an indexes, Qur'an translation, daily hadith, Islamic books, and audio, Arabic text representation. This software provides indexes of the Qur'an rather than a full translation of Qur'an in different languages.
www.globalquran.com	This application provides a keyword-based searchable interface which has been indexed by chapters number. It also shows the Arabic corpus, an annotated linguistic resource which shows the Arabic grammar, syntax and morphology for each word in the Holy Qur'an.
www.quranexplorer.com	It provides full translations of the Qur'an in many languages. It also provides audio and Arabic text representation in desktop and mobile version.

Qur'an Online Applications	**Features**
www.yaquran.com	It has many features such as Qur'an indexes, Qur'an translation, daily hadith, Islamic books, and audio, Arabic text representation. There is keyword based searchable interface for the user to find any verses in different languages but is using traditional keyword searching technique.
Zekr	Zekr is free desktop applications that enable users to read, listen and search Qur'an in many languages. This application is primarily designed to assist people in learning Qur'an. Users can search by querying a word or insert a verse number. When the users query a word, results will return all verses containing the query word. It uses keyword-based searching techniques. For instance, if a user enters the word heaven as a keyword, it will return all verses containing the word heaven.

In recent years, there has been an increasing amount of literature about research that focuses on the computation of Quranic knowledge, such as linguistic processing. information retrieval, semantic search, question answering, and data mining. We are still witnessing a growth in the computerization of Qur'an content and the knowledge it contains. Most of the online Qur'an application offers a different set of features and readers. The features are focused only on natural language or communities. This is something that needs to be corrected to encourage others to use the application. This review will help in giving an idea for the development of the prototype for the final online application for Malay Qur'an.

2.4.1 Linguistic Analysis of the Qur'an

In the linguistic analysis of Quranic content, several research projects have been conducted. [Dukes et al., 2010] proposed a novel linguistic processing of the Qur'an which contains a morphological analysis and part of speech tagging of Quranic Verses. The work presents the Treebank, a syntactic representation of the Qur'an verses. The main objective of the research is to clearly analyze and show the meaning of Quranic texts. The paper described an approach to morphological annotation of Quranic Arabic content which was initially verified manually and then computationally analysed to find a morphological representation of Quranic corpus to enable user searches for the Qur'an verses and see a morphological representation of the Quranic verse selected.

The further study by [Dror et al., 2004] presented a computational system for morphological analysis and annotation of the Qur'an, which is mainly for research and teaching purposes. The work processed several queries from the Quranic text that refer to words and linguistic attributes. The system uses a finite state toolbox to undertake a morphological analysis of the words in the Qur'an. Moreover, [Thabet, 2004] presented a new stemming approach based on a light stemming technique that uses a transliterated version of the Qur'an in western script. QurAna is also a computational research into the Qur'an by [M.Sharaf and Atwell, 2012]. In this work, the Qur'an text was used to develop a large corpus for the Qur'an related knowledge, where personal pronouns are tagged with their antecedence. The corpus can be used by researchers in several Qur'an-related applications, such as for training purposes, extracting empirical patterns and rules for creating new anaphora resolution approaches.

The Qur'an is the most widely read book in the world. The Qur'an has stimulated the interest of many researchers, especially in the field of information systems to assess and automate the extraction of knowledge. This led to the development of some search applications, which aim to provide retrieval of knowledge to facilitate people to understand people in a better way. However, understanding and retrieving the knowledge from the Qur'an is a major research challenge for computer science and artificial intelligence [Dukes et al., 2010]. The search method in Qur'an faces several fundamental problems, such as the inability to retrieve relevant knowledge and verses [Yauri et al., 2012]. The further study proved by [Tian, 2012] in her thesis that most search engines suffer from three common problems in Natural Language Processing such as the synonym problem, the homonym problem, and the wrong granularity problem. The synonym problem appears in the form that the user might send a different term to the search engine than what is contained in a document. For instance, a user query “the last prophet in Islam” might miss documents with Prophet Muhammad (may God’s prayers and peace be with him), even though these two terms are synonymous.

Many digital Quranic databases provide root verse searches. The database processes a morphological analysis, with a query verse as the input and the root verse as output [Shoaib et al., 2009]. The need to search for related words in the Qur’an has resulted in the creation of keyword-based searchable interfaces, indexed by chapter number. The interfaces assist users in browsing the Qur’an and conducting searches using translation and Tafsir. A multi-lingual Quranic software provides Arabic and English Quranic commentaries. Translations into French, German, Spanish, Urdu, Malay, Indonesian, Japanese, Tamil, Hausa, Turkish, and Indonesian, are also available on many websites equipped with a word-based search facility. The software by ’Harf’ provides a subjective search facility, but only in the Arabic language. This software also provides an exact match search for words, terms, parts of verses, and even some following verses. Technically, this software offers the ability to search static files in a way that the verses are prelinked to a topic or subtopic. Thus, semantic search in the Qur’an is sometimes based on Internet searches that reveal some works on the Qur’an [Kayed et al., 2008].

2.4.2 Related work on Linguistic Analysis of the Malay Translated Qur’an

Nowadays, the Qur’an has been translated into various languages around the world by Muslim experts. The main aim of the availability of the Qur’an translations is to allow the reader to understand the Qur’an in more precise ways. As for Malay readers, there are many available Malay Qur’an translations. Nevertheless, there are a few issues regarding the Malay Qur’an translations such as ambiguity of words, lack of word equivalence between Malay and Arabic or Malay with English, and different structures of words, sentences, and discourse in these two languages [Tabrizi and Mahmud, 2013]. Besides, [Wahid, 2011] claimed that Standardization of different versions of Malay Qur’an was challenging due to the need for alignment of meanings. Different versions of Malay Qur’an surfaced due to the manner translations were made, which could likely be from a secondary source for some of the texts. Standardization issues could be reduced if the adaptations were made from the original texts with references to other authentic sources like Hadith or Tafsir.

Most of the research on Malay Qur'an focused on Natural Language Processing area. Each analysis suggested different techniques with the same purpose for improving information retrieval. [Abu Bakar and Abdul Rahman, 2003] proposed stemming and thesaurus to search and retrieve relevant Malay translated Qur'an documents based on user natural query words. A stemming algorithm is an automated procedure as it reduces words with the same stem to a common form, usually by removing derivational and inflectional suffixes from each word. For example, the words study, studies, studied, studying, student or studious are reduced to the root word study. Grouping these words into a common form will increase the need for retrieving relevant documents to a given query. The authors stated by using stemming, the efficiency of document retrieval is increased since the size of index files is reduced by 50% because of grouping many morphological word variants into a single stem word. Based on the experiment, the combination of stemming and thesaurus methods increased recall rate by 60.22% compare to exact match search with just 33%. Although the combination appears to be the finest of all, the retrieve and relevant result is still low which implies that that there are still key terms in Al-Quran documents that are not available in the thesaurus entries.

On the other hand, [Yunus et al., 2010a] presented Stemming Semantic Query (SSQ) as a new approach to improve the retrieval of verses for Qur'an document results. The authors compared the semantic results and stemming semantics results using three different languages such as English, Malay, and Arabic. This research found that the semantic approach of Stemmer contributed to the better performance of retrieving more relevant and related Qur'an documents. Furthermore, [Yahya et al., 2013] proposed a semantic search for the Qur'an based on Cross-Language Information Retrieval (CLIR). In this research, they evaluated a CLIR approach based on domain ontology that used Qur'an Arabic concepts [Dukes, 2013] for disambiguation of the translation of a given query and enhancing dictionary-based query translation.

One Malay Tagger is developed by [Ahmad et al., 2013] which applies the trigram Hidden Markov Model (HMM) method to identify word tags in Malay sentences. The model is tested using a corpus of 18,135 tokens tagged with a set of 21 tags similar to the set of tags used by *Dewan Bahasa dan Pustaka (DBP).* The results show that the best predictions are made with accuracy 67.9% using only prefix information with a fixed prefix length equals to three letters. Similar results with accuracy 66.7% are achieved using a combination of the first and the last three letters of each word. When using suffix information only, the best accuracy achieved is 60% with a suffix length of five letters. These findings show that HMMs are suitable models to be used to predict any Malay word's POS tag. In addition, [Alfred et al., 2013] proposed a rule-based method for identifying Malay POS tags called RPOS. It applies affixing and word relation rules to determine the right word category. Malay words can be formed with prefixes, suffixes, circumfixes, and infixes. In their paper, the authors consider infixes less important and not effective for the task of POS tagging. Different affixes can be categorized in different word categories. When there is more than one possible tag for the word, word relation rules are applied to identify the most suitable POS tag based on the context. If the word is not found in the POS tag dictionary, affixing rules are applied to determine possible tags for the word, and then the word relation rules are applied to solve the ambiguity (if any). The POS dictionary is manually built from Malay Thesaurus by DBP and used to assign all possible tags to each word in a Malay sentence. The results of this rule-based method show that it has higher performance than the statistical POS tagger with an accuracy of 89% for Malay news articles and 86% for Malay biomedical articles. This shows that it can predict unknown words' POS tags with reasonable accuracy. However, this tagger fails to tag words that are borrowed from English and words that have no affixation, especially proper nouns. Richer relation rules are needed to improve the tagging results of RPOS tagger.

Lastly, [Xian et al., 2016] proposed Mi-POS, a Malay language POS tagger that was developed using a probabilistic approach with information about the context. This research used manually built corpus containing 152 articles with a total of 64,534 tokens from Bernama news archive. It was manually tagged by a Malay native speaker who assigned a single POS tag to each word. Then the tagged corpus was verified by two other Malay native speakers to correct mistakes and solve ambiguity if any. There is a total of 13 non-symbols POS types used to tag the training corpus by the Malay native speakers. The authors also compared the result with other Malay POS taggers such as Lazy Man's Tag and Trigram HMM. The final results showed that Mi-POS outperforms other Malay Part-of-Speech taggers in terms of accuracy with an accuracy of 95.16% obtained by tagging new words from the same training corpus type and 81.12% for words from different corpora types.

Although few studies have been done, however, some limited resources and tools are available or made accessible for computational linguistic analysis of the Malay Translated Qur'an.

2.4.3 Related work on Knowledge Representation

Qursim was presented in [Sharaf and Atwell, 2012] and is a system that that linked a related, semantically similar verse which formed a large corpus from the Quranic data. The corpus can be used in computational linguistics and machine translation, problem solving, and another related research. In the work of [Sharaf and Atwell, 2009], they presented a formal knowledge representation of the Qur'an that was corpus-based. The work used FrameNet frame to build a lexical database of verb valences in the Arabic Qur'an. They studied the verbs in their context in the Qur'an and then compared them with matching frames evoked in the English frame.

[Nassourou, 2011] presented a methodology of reconstructing Qur'an's chronology based on machine learning techniques. A hybrid statistical classifier has been employed to get the plausible dates of revelation under the traditional Islamic scholars and western orientalists chronologies. After one year, the author once again produced a new technique for categorizing the Qur'an. In the year 2012, [Nassourou, 2012] proposed a new algorithm using machine learning techniques for classifying the chapters in the Qur'an. In this research, the author used SVM and naïve Bayesian as functional classifiers. The categorization model of the chapters was based on the phases of the messenger ship of Prophet Muhammad. This study helps in arranging the chapters of the Qur'an according to the period of

Prophet Muhammad's messenger ship.

2.4.4 Related work on Information Retrieval

A Quranic information retrieval system based on the use of formal methods is presented by [Al-taani and Al-gharaibeh, 2011] This research describes the use of formal methods for Quranic natural language processing search systems, and uses Z notation to express the formal specifications of the text-based search technique, synonym-based search systems, and stem based search systems used in the Quranic search systems. The system is based on a keyword search that allows users to search using keywords to retrieve relevant verses from the Qur'an.

Besides, [Noordin and Othman, 2006] offered a system for retrieving Quranic content and important knowledge derived from the Qur'an. The system surveys various websites that represent Quranic texts and retrieves Quranic texts and knowledge from them. The focus of system design was on translation texts, recitation, exegesis, hadith, and historical concepts or objects mentioned in the Qur'an. The system helps users identify the meaning of verses in the Qur'an, their proper citation, and the knowledge generated from these verses.

On the other hand, [Nassourou, 2011] reports another system that supports the retrieval of Quranic content. In this system, verses are clustered by chapters, and an arc weight is assigned to each cluster based on the number of verses it contains, so that users can easily identify the most relevant areas and identify places of revelation in the verses. Users can see the complete results and make a selection of the plot to zoom in and click on an indicator to view a table containing verses with their corresponding English translation.

An automatic speech recognition system for the Qur'an was reported in the work of [Mourtaga et al., 2006]. In this research, a system was presented that automatically recognizes speech for Quranic based speakers independent. In this work. they employed a tri-phone Hidden Markov Model (HMM) and Maximum Likelihood Linear Regression (MLLR) for the development of the speech recognizer. They used the 30th chapter of the Qur'an and five of the most famous readers of the Qur'an were used for the training and testing of the data. The results show that that a good speaker independent recognizer can be applied for new readers. The range of accuracy for the tested sample was 68% and 85%. The main drawback of the system is the recognition time required, due to the many number states of the HMM model.

2.4.5 Related work on Semantic Search of the Qur'an

Growing interest among Muslims and non-Muslims to study and understand the Qur'an has created a massive wave of Qur'an ontology development. In the recent past, there have been numerous studies conducted on the development of Qur'an ontology by different researchers around the world. They used many various tools and techniques to either extract, develop or analyze the knowledge from Qur'an. [Ullah Khan et al., 2013] describe an ontological work for searching concepts from the Holy Qur'an using the semantic search technique. In this research, the authors proposed a new framework that represents the semantic search. This framework can be applied to any Islamic document. This research uses living creatures including animals and birds mentioned in Holy Qur'an as the sample domain ontology. Unfortunately, this research is not very clear in explaining the results and how the framework can improve the current ontology-based semantic search in Holy Qur'an. [Uthayan and Anandha Mala, 2015] suggested a new querying mechanism for information retrieval, which integrates ontology queries with the keyword search. This research proposed a new hybrid method based on matching extracted instances from the queries and information field.

Furthermore, [Ta'a et al., 2013] developed an ontology to represent and classify the knowledge of Qur'an using a thematic approach. The thematic or theme-based approach focuses on themes within a story to give narratives a sense of direction and purpose [Hargood et al., 2008]. This article was focusing on the topic of Qur'an as described in Syammil AL-Qur'an Miracle Reference (Qur'an, Akhlak and Iman). This Qur'an ontology helps the user to understand the Qur'an systematically.

On the other hand, [Adhoni and Al Hamad, 2014] developed a cloud-based Qur'an portal using Drupal technology. In their work, the authors build a portal which can be used to search the Qur'an in more than one language. Not only that, the Qur'an Arabic WordNet by [Almaayah et al., 2014] studied on developing WordNet for Qur'an by creating the semantic connection between words to get a better understanding of the meaning of Quranic Words. These studies focus on Classical Arabic of the Qur'an rather than modern standard Arabic. [N H Abbas, 2009] improved on using structured syntax by enabling the use of natural language for querying. This project developed a bilingual (English/Arabic) comprehensive search tool for the Qur'an. Her work mainly involved a keyword search for concrete and abstract concepts that are found in the Qur'an.

The search model has been successfully applied to validate that the verse number in each chapter is correctly written in the Qur'an XML format document. The checker also authorized that XML documents contained precisely the same number of chapters as the holy book by comparing the verse numbers in each chapter and the chapter numbers. Several researchers have described the use of ontology in keyword and key phrase extraction related to the search in Islamic literature [Nguyen and Rusin, 2006]. In their study, [Khalid et al., 2010] argued that a distinct need exists to create automated contextual and thematic associations among heterogeneous and distributed sources in Al-Qur'an and the Book of Hadith. Digital multimedia religious contents available to date are only suitable for online publishing. However, retrieving adequate data, consulting the verse on demand, and integrating all aspects perfectly is difficult for humans. Across the various Quranic resources, authors and scholars use contrasting terminologies to define and describe a concept.

[Alrehaili and Atwell, 2014] reviewed the past approaches or works related to computational ontologies based on nine (9) criteria such as Qur'an Text, Coverage Area, Coverage Proportion, Underlying Format, Underlying Technology, Availability, Concept Number, Relation Type, and Verification Methods. Table 2.2 summarizes these criteria;

Table 2.2: Summary of the criteria of the existing Qur'an online applications

Criteria	**Previous Work**
Qur'an Text	[Ullah Khan et al., 2013], [Saad et al., 2010] and [Ismail et al., 2015] build ontologies using English translation of Qur'an. [Yahya et al., 2013] and [Yunus et al., 2010b] used Malay translation of Qur'an as a source for building the ontology.
Coverage Area	[Ta'a et al., 2013] covered on the themes of Qur'an as describes in Syammil AL-Qur'an Miracle the Reference (Qur'an, Akhlak and Iman). While[Ullah Khan et al., 2013] used living creatures including animals and birds mentioned in Holy Qur'an as the sample domain ontology.
Coverage Proportion	Entire Qur'an: [Muhammad, 2012] in his PhD thesis used all the chapters in the Qur'an. Some parts: There was no research conducted to cover some part of the Qur'an. Specific Topic : Most of the research covered specific topic such as [Shoaib et al., 2009] just focusing on the chapter 2 of the Qur'an.
Underlying format	[Yahya et al., 2013], [Ta'a et al., 2013] and [Ullah Khan et al., 2013] used OWL to build the ontologies. [Yauri et al., 2014] used RDF/XML format.
Underlying technology used	[Ullah Khan et al., 2013] used protégé and SPARQL.
Availability	[Hakkoum and Raghay, 2016a] published his research work online. it is accessible at http://www.quranontology.com/. [Dukes, 2013] also created Qur'an Ontology that can be accessed at http://corpus.quran.com

Criteria	**Previous Work**
Concepts number	[Dukes et al., 2010] defined 300 concepts with 350 relations mainly of type "InstanceOf", in the Qur'an. [Hakkoum and Raghay, 2016a] defined 1181 concepts in the Qur'an
Relations Type	[Shoaib et al., 2009] showed the synonyms relations. [Saad et al., 2009] shows meronyms (partOf).
Verification Method Used	[Saad et al., 2009] and [Saad et al., 2010] used domain experts as the method of verification. [Dukes et al., 2010] and [Muhammad, 2012] used Ibn Kathir as the method of verification.

Most of these applications use keyword search techniques with a few information retrieval methods, as can be noticed in the following reviews. Additionally, I have found many published applications using Semantic Web search technologies. However, numerous researchers have proposed frameworks for the Qur'an semantic search tool based on concepts.[Shoaib et al., 2009] proposed a relational model for semantic search in Qur'an using the WordNet relationships. This relational model creates the taxonomy of the related terms in Surah Al-Baqarah (Chapter 2 from the Qur'an). The model facilitates performing a subject search for Qur'an readers and provides a framework capable of retrieving related verses from the Qur'an. This model of semantic search showed 80% accuracy than the simple search. However, during the retrieving process, some irreverent verses are also retrieved. In this paper, the authors also discussed the problem of the current keyword-based searching and the issues related to semantic search in the Holy Qur'an. This research has contributed to the improvement of semantic search in the Qur'an. In the future works, authors intend to extend this work that can eliminate irrelevant verses. After four years, the researchers have once again proposed new research on ontology-based semantic search. [Tarawneh and AlShawakfa, 2015] presented a new hybrid method called Al-Baheth Searcher of Qur'an Text using the combination of syntactic (keyword) and semantic-based approach to index and search the Holy Qur'an text.

2.5 Summary

This chapter provides a brief idea about the semantic web, in particular what it is and does. The chapter has shown how the semantic web provides a framework for translating data into a structured form in order for computers to read and understand. We have shown how data is represented in structural RDF graphs and how this data is accessed by the system in the form of semantic search using a structure query language known as SPARQL.Moreover we present clearly to retrieve information from a knowledge base, the users prefer natural language, which requires an application, such as SPARQL, that semantically formulates such natural language using structured queries.

We discussed previous work reported by the researchers on semantic query formulation and some of the advantages and challenges of the current semantic query formulation systems. We have looked at how semantic was incorporated into an information retrieval system in general and in the Qur'an domain.Presently, most of the studies were carried out only to extract information from a single ontology. There is a lack of studies focused on matching from multi-lingual ontologies such as English and Malay language. With the increasing number of distributed resources, services, and applications on the web, multi-lingual ontology matching is likely to become essential. Therefore, in this thesis, we used Malay and English language as our dataset. Then, the research will evaluate the impact of using both Malay and English language in getting the maximum relevant results based on the user queries. This will be discussed in chapter 4.

Studies in these chapters indicate a need for a semantic search engine for different domains and languages. Although there is still room for improvement to the above existing techniques, however, there is a need to explore semantic search and ontology development, especially in the Malay language. This language has a complex structure and morphological analysis that need to be examined and discovered.

Chapter 3

Historical Background of Malay Language

Together with the English, Chinese, and Hindi languages, Malay is one of the most widely spoken languages and is rich with linguistic traditions. Although many people use this language, it is still one of the least studied and known among the world community. Thus, this chapter will discuss about the history of Malay language and will highlight the uniqueness of Malay morphology.

3.1 Background of the Malay Language

The Malay language or Bahasa Melayu is part of the Austronesian language family and it is widely used in the South East Asia region that includes Malaysia, Singapore, Indonesia, Brunei, and Thailand by over 290 million people based on June 2017 Statistics in the Internet World Stats page. Austronesian languages are divided into four groups and they are Indonesian, Melanesian, Austronesian, and Polynesian, with Indonesian languages forming the biggest group. Many words in the Malay language have been adopted and adapted from other languages, the earliest being Sanskrit, and later on it was heavily influenced by other languages as well, including Arabic, Tamil, Hindi, and English. The evolution of this language can be broken down into three time periods [Asmah Haji Omar, 2015]:

1. **Old Malay**: Old Malay (6th century to 15th century) is a formal lan-

guage that is based heavily on Sanskrit. The Old Malay is belonging to the archipelago family under the Sumatera language. Various inscriptions were found in South Sumatera by the Tatang River on stone tablets dated between 682 and 689 AD. By that time, the Old Malay had become the lingua franca and the national language due to its easy to be influenced by other languages. The Old Malay language is also not bound by the differences in people's rank and has a simpler system than the Javanese language.

2. **Classical Malay**: The 13th century was the beginning of the transition period in the Malay Archipelago with the expansion of Islam into the region. This has influenced the race and language here, especially the Malay race and language. India's influence was gradually replaced by Islamic and Arab influences. The growing influence of Islam in Southeast Asia in the 13th century also affected the development of the Malay language. The arrival of Islam in Malaya has greatly altered the Malay language system, especially in terms of vocabulary, sentence structure and writing system.

   The splendor of the classical Malay language can be divided into three important periods: the Melaka government period, the Acheh government period and the Johor-Riau government period. At this point, a few sultanates were established in peninsular Malaysia and Borneo. Many legal documents and letters were written mostly in Arabic-based script, as Islam spread from Malacca to the rest of Malaya. Malay became lingua franca (adopted common language) among traders who travelled through Malacca in the 15th century. As Islamic faith and Malay spread throughout the region, many Arabic words were infused into the existing language such as words like halal, *dunia* (world), *falsafah* (philosophy). There were also Persian words like *bandar* (town), *kismis* (raisin), *pasar* (market/bazaar) and Hindi words like *dobi* (laundry), *bendi* (okra) which were adopted from traders.

3. **Modern Malay**:When the Portuguese and Dutch arrived in Malacca, more words were added into the lexicon as they proselytized (converted others to) Christianity. Dutch scholars even converted the bible into Malay, which resulted in more additions to the Malay language. Some Malay words with Dutch origins are *sepanduk* (banner) and *rokok* (cigarette). The same phe-

nomena happened with the English vocabulary when the British colonized the country. After Malaya gained independence in 1957, Malay became the national language. The language spoken and written now is classified as Modern Malay. As far as the evolution of languages go, Malaysia has changed tremendously in a very short of time. A lot of newer words in Malay come from English, possibly due to widely use of English in Malaysia. Hence, when there is a need for a specific word, it is easier to adapt from English. *Kontemporari* (contemporary), *mesej* (message), *naratif* (narrative) are examples of words adopted by Malay directly from English with only changes in spelling.

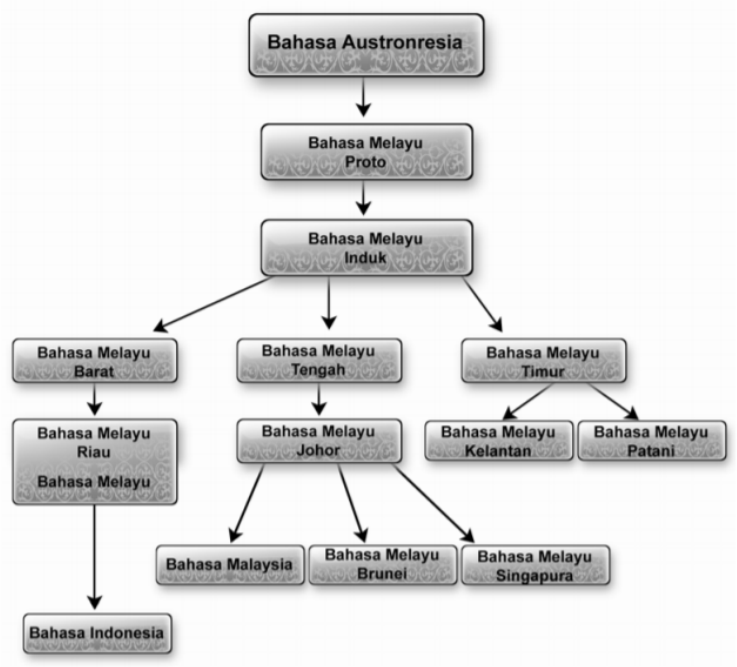


Figure 3.1: Malay Language Family[Asmah Haji Omar, 2015]

3.1.1 The Malay Writing

In the early days of the Malay people, they had no writing skills. All communication is done verbally. The earliest record of old Malay writing was 682 AD on the tombstone found in South Sumatra. Since then, the Malay writing has undergone several changes and used several different types of letters. It started to use many foreign languages in their writing such as Sanskrit, English, Dutch, Javanese, Arabic, and so on.

The Sanskrit language had a great influence on the Malay writing when the Malay language was at an ancient level. Evidence of the influence of Sanskrit in old Malay languages can be found in the inscription left by the Srivijaya government. In Malay, there are 677 words from Sanskrit. The following are some examples of Sanskrit words borrowed into Malay such as *dosa*, *duka*,*dewa*, and *sengsara*. The Javanese and the Malay languages fall into the same language group. The use of the Javanese language happened when a crisis occurs between local people who speak the Malay language with the Javanese immigrants who speak Javanese, which has resulted in the elements of Javanese language being absorbed into the Malay language. The example of the Javanese words in Malay such as *Andong/Kereta Kuda*, *Batok/tempurung*, and *antipati/raja*.

Arabic and Islamic religions greatly influence the development of the Malay language. The status of the lingua franca of the Malay language and its uniqueness has caused Islam to be spread in Malay and not in Arabic. However, the Arabic language through the advent of Islam has influenced the development of the Malay language in several aspects, including vocabulary, pronunciation, and writing. There are some examples of Arabic words borrowed into Malay such as *Abah*, *Kerusi*, *Syukur*, and *Kamus*. We can see the use of this Arabic word in Malay translated Qur'an. For instance, *Subuh*, *qadhi*, *Ajam*, and *Amil*.

In general, there is a significant difference between formal and informal Malay writing. Malay formal writing emphasizes the correctness of spelling based on the Standard Guidelines for Spelling Malay Languages, as well as the use of the grammatical grammar outlined by Dewan Bahasa dan Pustaka (DBP) Malaysia. Besides, in formal writing, writers are encouraged to develop concise short sentences along with the content. The use of words or terms that are borrowed directly from foreign languages is also discouraged if there is a more appropriate word choice. For instance, *bajet* (English: budget) and *diskusi)* (English: discussion). The use of abbreviations in words or phrases or sentences should also be avoided in formal writing style. In addition, any honorary titles such as Dato', Datuk, Datuk Seri, and Tan Sri will also be used in full form. Formal writing styles are commonly found in official documents, government documents, publications, and teaching and learning materials in local schools and universities. It is also commonly found in most mainstream media such as newspapers (Daily News, Utusan Malaysia), and scholarly magazines (Student Council, Community Hall).

Malay informal writing does not emphasize the use of standard language. The spelling correctness aspect, as well as the use of grammar is sometimes ignored, either intentionally or unintentionally. Writers also have more freedom in terms of word selection and tend to emphasize creativity than the correctness of language in sentence development. In addition to the higher frequency of word use, there are also borrowed words from other countries used in this informal writing. Informal writing is not tied to any format. Informal writing styles are commonly found in most mainstream or electronic media socials that are more casual and have a specific target group such as entertainment magazines and teen magazines. It is also used in many printed publications of alternative publishers such as novels and fiction books.

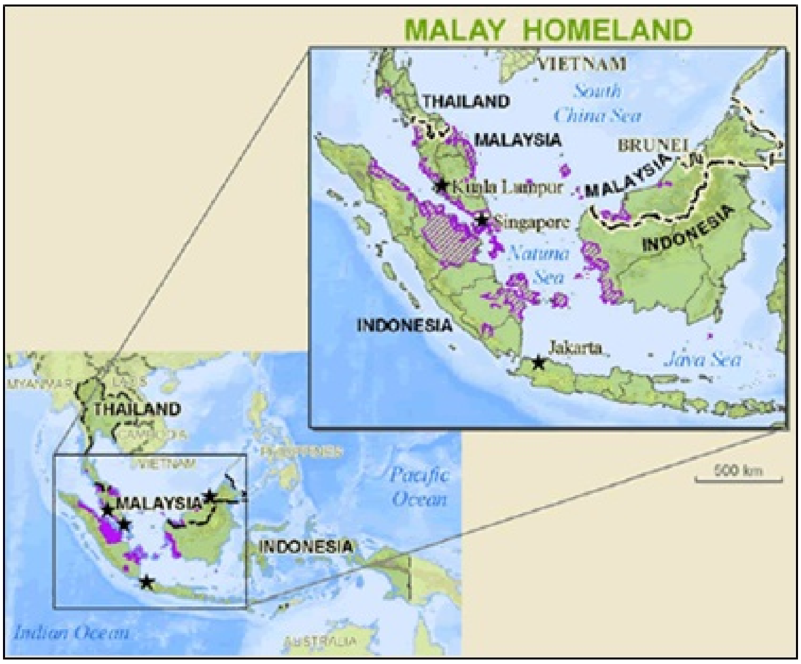


Figure 3.2: Geographical location of Malay Speaker

3.2 Morphology in the Malay Language

Morphology is the field of language structure, form and classification of words. The word structure is the arrangement of the speech or symbolic form (written) that is built up from a smaller meaning-bearing unit called morphemes to become a meaningful language unit. Morphologically, Malay is a language which belongs to the agglutinative language family. It does a lot of affixation, reduplication, and composition (compounding) as well as other rarely used processes (such as deletion or producing acronyms) in word formation [Sharum et al., 2010]. The meaning of words can be changed by adding inflectional morphemes such as prefixes, suffixes, and circumfixes to the root words. For example, the verb ‘makan’ (eat), when added with the suffix ‘-an’, becomes ‘makanan’ (food).

In this research, we used 356 stopword from [Ahmad, 1995]. However, after go through the stopword, we found there are several word that are duplicate and irrelevant. Thus, we have to eliminate those words. Therefore, the total number of stopwords used in this research is 318 words.

3.2.1 Affixation

Affixation is the process whereby a base word may be extended or added by one or more affixes. Affixation is the most common process of the three morphological processes. Affixes can be classified as prefixes, suffixes, infixes, and circumfixes.

1. **Prefix**: attaches itself at the beginning of a word. For example: *‘bersalah’* which mean guilty, start with *‘ber’* which is a typical prefix in Malay language. Another prefix appeared in the Malay language such as *‘per’*, *‘mem’*, *‘men’*, *‘pen’*, *‘ter’*, *‘meng’*, and *‘juru’*.

2. **Suffix**: attaches itself at the end of a word. For example: *‘memaafkan’* which means to forgive, end with *‘kan’* which is a typical suffix in the Malay language. Another suffix such as *‘an’*, *‘i’* and *‘mu’*.

3. **Infix**: usually located in the middle of word.For example: *‘gerigi* which refer to the teeth of the blade, was derived from the root word *‘gigi’* (teeth).

4. **Circumfix**: Prefix-suffix pair where more than one affix is attached to a word at the same time and is usually positioned before and after the root word. For example: *'kerajaan'* which means kingdom, was derived from root word *'raja'* (king).

The difference between Malay and English affixes is English affixes can indicate or produce negative meanings, for example *im-*, *dis-*, *mal-* and *ir-*. These affixes changed positive meaning into negative meanings. For instance, from 'function' to 'malfunction' or 'responsible' to 'irresponsible'. This phenomenon does not exist in Malay. Malays used different or additional words to represent negative meanings. For example, *'berfungsi'* (function) to *'tidak berfungsi'* or *'bertanggungjawab'* (responsible) to *'tidak bertanggungjawab'* (irresponsible). The use of the word *'tidak'* in Malay sentences indicates negative meaning.

Affixation in Malay is highly useful and often accompanied by phonological variation in the word. For example, the combination of the verbal prefixed me- with the word *'lawat'* (visit) produces*'melawat'* (to visit). Besides, multiple affixations are also widely used to form a new single word, with sometimes used up to four affixes. For example, *'diperbanyakkannya'* (made plenty) which consist of *di-* + *per-* + *banyak* (a lot) + *-kan* + *-nya.* This phonological and orthographic variation in the Malay word make the morphological analysis such as stemming and POS-Tagging difficult [Xian et al., 2016]. This can be seen when the affixation method will derive various words that change their syntactic class category from the original word. Contrary to the English language, the syntactic class category remains the same when forming a new word using the inflection method. For instance,

- *'Makan'* (eat) is a verb.
- *'Makanan'* (food) is a noun. (When adding suffix –an)
- *'Pemakanan'* (nutrition) is an adjective. (When adding prefix pe-)
- *'Termakan'* (unintentionally) is a verb. (When adding circumfix –ter and –an)

3.2.2 Reduplication

Reduplication, is a word-formation process in which meaning is expressed by repeating all or part of a word. This reduplication is widely used in many Malay texts. Reduplication is hardly found in other languages. There are 3 basic categories of Malay reduplication; full, partial and rhyming and chiming reduplication. Another reduplication called free-form reduplication is the reduplication's group where their formation is not yet clearly understood, thus undefined.

Full reduplication involves a base word, complex word, or compound word. For example, *'perempuan-perempuan'* (girls) come from base word *'perempuan'* (girl), *'pertubuhan-pertubuhan'* (organizations) come from the complex word per- + *'tubuh'* (body) + -an (*'pertubuhan'* or organization), and *'kakitangan-kakitangan'* (staff) come from compound word *'kaki'* (leg) and *'tangan'* (hand).

Partial reduplication is a process which reduplicates one part of the base word or root word. It can be divided into first syllabic reduplication, root reduplication, and compound reduplication. First, syllabic reduplication is a reduplication process which derives a noun from another noun, by reduplicating the first syllabic and converts the vowel of the copy to a letter 'e' (called *'e-pepet'*) such as *'pepatung'* (dragonfly) from *'pa'* + *'patung'* (statue). Root reduplication is reduplicating the root word of the derived base word. The reduplication can be positioned either in front of or behind the base word. This position usually reflects the meaning of the derived word. The root positioned at the front generates a meaning of a reciprocal act called *'menyaling'* e.g, *'pukul-memukul'* (hitting each other), while a reduplication root positioned at the back generates a meaning of multiple or repeated act e.g, *'berlari-lari'* (running). The difference between root word duplication and reciprocal is just the meaning where it describes repeated and opposing acts. Both are actually root word duplication [Yunus et al., 2010a] Compound reduplication is a partial reduplication when it follows the Distributed Morphology (DM) rules, where the main component of the compound word is preceding the other component e.g, *'alat tulis'* (stationery).

Rhythmic reduplication involves repeating certain forms of the root word such as consonants, syllables, or vowels, which creates symphonic sounds in the pronunciation. For example, 'kayu-kayan' (woods), *'batu-batuan'* (stones) and *'gunung-*

ganang' (mountains). Free-form reduplication is a reduplication which does not belong to any of the categories above. For instance, *'sahabat-handai'* (friends), *'nenek-moyang'* (ancestors) and *'ipar-duai'* (brothers and sisters-in law).

3.2.3 Composition or Compounding

Compounding is a process of linking two or more basic words together into single words and carries a certain meaning. Compound words may be hyphenated, written open as separate words, or written solid. Most of Malay compound words are constructed by nouns and it was modified by other nouns, verbs, and adjectives. For instance,*'adat-istiadat'* (customs and traditions), which linked a single word *'adat'* (customs) with another single word *'istiadat'* (tradition).

3.2.4 Discussion

A major problem in Malay morphological processing of Malay documents is the analysis part. This often happens during the creation of information retrieval applications (stemming) and corpus tagging (glossing). Morphological analysis process is the process of analyzing the lexical form of a word from its root form. For instance, to identify the root where the word originates in stemming or conflation and to identify the underlying word's structure and features for word glossing in corpus tagging. In creating Malay Stemmers, not only all affixes need to be removed, but understanding the variations of different aspects of the four affixes shown above are crucial. Although many researchers have come out with solutions, however, the under-stemming and over-stemming issues remain unresolved [Sharum et al., 2010]. This also causes morphological analysis for the Malay language being still far behind and there are no accessible resources for morphological analysis results to be found.

3.3 Related work on Malay Morphology

3.3.1 Stemming

Stemming is a computational process of reducing a word from its derived form into its root term. Stemming is used in many information retrieval systems to reduce different word forms to common roots. In the stemming algorithm, words with the same root are reduced to a common form by stripping each word of its derivational and inflectional suffixes. The stemming process is quite complicated for the Malay language compared with other languages because of the unique morphological structure.

[Abu Bakar and Abdul Rahman, 2003] proposed stemming and thesaurus to search and retrieve relevant Malay translated Qur'an documents based on user natural query words. A stemming algorithm is an automated procedure as it reduces words with the same stem to a common form, usually by removing derivational and inflectional suffixes from each word. For example, the words study, studies, studied, studying, student or studious are reduced to the root word study. Grouping these words into a common form will increase the need for retrieving relevant documents to a given query. The authors stated by using stemming, the efficiency of document retrieval is increased since the size of index files is reduced by 50% because of grouping many morphological word variants into a single stem word. Based on the experiment, the combination of stemming and thesaurus methods increased recall rate by 60.22% compare to exact match search with just 33%. Although the combination appears to be the finest of all, the retrieve and relevant result is still low which implies that that there are still key terms in Al-Quran documents that are not available in the thesaurus entries.

On the other hand, [Yunus et al., 2010a] presented Stemming Semantic Query (SSQ) as a new approach to improve the retrieval of verses for Qur'an document results. The authors compared the semantic results and stemming semantics results using three different languages, i.e., English, Malay, and Arabic. This research found that the semantic approach of Stemmer contributed to a better performance of retrieving more relevant and related Qur'an document results.

3.3.2 POS-Tagging

Part-of-speech (POS) tagging is an important process that is used to build many Natural Language Processing (NLP) applications. POS-tagged text or sentences with equivalent part-of-of-speech tags based on the word definition and relation. POS tagging is widely adopted for languages such as English (Penn Treebank, CLAWS), German [Schmid, 1999], and Arabic[Al-Omari and Abuata, 2014]. Contrary to the Malay language, not many research has been done on POS-Tagging analysis. However, there is one online Malay POS Tagger developed by PhD students in University of Malaya,Malaysia, [Xian-mo, 2007]. This POS-Tagger can help researchers if they want to tag or carry out computational linguistic analysis.

One Malay Tagger is developed by [Mohamed et al., 2011] which applies Trigram Hidden Markov Model (HMM) method to identify word tags in Malay sentences. Context information other than the surrounding tags, namely, the prefix and the suffix, has been used to predict the correct POS tags. His study measures the effect of using these features individually as well as using a combination of both the prefix and the suffix of each word in the final model's predictions. The model is tested using a corpus of 18,135 tokens tagged with a set of 21 tags similar to the set of tags used by Dewan Bahasa dan Pustaka (DBP). This corpus is tagged automatically by mapping each word to a list of possible tags from a dictionary, and then the ambiguity is solved manually. The results show that the best predictions are made with an accuracy of 67.9% using only prefix information with a fixed prefix length equals to three letters. Similar results with an accuracy of 66.7% are achieved using a combination of the first and the last three letters of each word. Only when using the information on suffixes can the best accuracy be achieved, which is 60% with a suffix length of five letters. These findings show that HMMs are suitable models to be used to predict any Malay word's POS tag.

In addition, [Alfred et al., 2013] proposed a rule-based method for identifying Malay POS tags called RPOS. It applies affixing and word relation rules to determine the right word category. Malay words can be formed with prefixes, suffixes, circumfixes, and/or infixes. In their paper, the authors considered infixes less important and not effective for the task of POS tagging. Different affixes can be categorized in different word categories. For example, the verb type can be identified by the prefix *mem-*, adjective type involves a prefix like *ter-* and the noun type involves prefixes such as pen-. When there is more than one possible tag for the word, word relation rules are applied to identify the most suitable POS tag based on the context. If the word is not found in the POS tag dictionary, affixing rules are applied to determine possible tags for the word, and then the word relation rules are applied to solve the ambiguity (if any). The POS dictionary is manually built from Thesaurus Bahasa Melayu by Dewan Bahasa dan Pustaka (DBP) and used to assign all possible tags to each word in a Malay sentence. The results of this rule-based method show that it has higher performance than the statistical POS tagger with an accuracy of 89% for Malay news articles and 86% for Malay biomedical articles. This shows that it can predict unknown words' POS tags with a more reasonable accuracy. However, this tagger fails to tag words that are borrowed from English and also words that have no affixation, especially proper nouns. Richer relation rules are needed to improve the tagging results of RPOS tagger.

A relational lexical database, MALEX (MALay LEXicon) with the purpose of providing linguistic information for Malay text analysis, has been developed by [Mohd Don, 2010]. It is designed according to the logical relationships among different kinds of linguistic information, and can generate suitable output for a range of computer-based applications. For instance, using grammatical and phonological information, it can output a detailed phonological representation of a text, which has useful applications in speech science. This research work on 800,000 words provided by Dewan Bahasa dan Pustaka (DBP), a newspaper corpus of about 5M words, a corpus of 1.3M words of speeches of the former Malaysian Prime Minister, Dr Mahathir Mohammad and finally some academic text consists of 20,000 words. The data collected are in a form of written text. However, there is a lack of technical discussion on the learning approach. Moreover, the performance of the system remains unclear.

Furthermore, [Yahya et al., 2013] proposed a semantic search for the Qur'an based on Cross-Language Information Retrieval (CLIR). In this research, they evaluated a CLIR approach based on domain ontology that used Qur'an Arabic concepts by [Dukes, 2013] for disambiguation of the translation of a given query and enhancing dictionary-based query translation.

Lastly, [Xian et al., 2016] proposed Mi-POS, a Malay language POS tagger that was developed using a probabilistic approach with information about the context. This research used manually built corpus containing 152 articles with a total of 64,534 tokens from Bernama news archive. It was manually tagged by a Malay native speaker who assigned a single POS tag to each word. Then the tagged corpus was verified by two other Malay native speakers to correct mistakes and solve ambiguity if any. There is a total of 13 non-symbols POS types used to tag the training corpus by the Malay native speakers. The authors also compared the result with other Malay POS taggers such as Lazy Man's Tag and Trigram HMM. The final results showed that Mi-POS outperforms other Malay Part-of-Speech taggers in terms of accuracy with an accuracy of 95.16% obtained by tagging new words from the same training corpus type and an accuracy of 81.12% for words from different corpora types.

The ambiguity issue is the most challenging part in text analysis for POS tagging. The same word may have different meanings in different contexts. This may result in improper text analysis and thus inaccurate POS tags. Therefore, a context-based POS tagger which identifies word categories based on the meaning of a sentence is necessary. On the other hand, another major issue regarding Malay language is the lack of linguistic resources available to be used to train the POS model. It needs to be done manually.

3.4 Summary

This chapter provides historical background on Malay language and Malay linguistic tradition, describing the uniqueness of morphological processes and related research conducted on the Malay language. Many of the kinds of processing we have discussed in this chapter belong to a kind of popular linguistics, current morphological analysis techniques, and the analysis result of text analysis on the Malay language. However, as we can conclude, state-of-the-art text processing systems for Malay Language are still dealing with problems related to lexical, morphological, and syntax analysis. In these circumstances, the emphasis has been on carrying out the task more effectively, e.g. to automate it rather than do it manually, or make a tagger run faster or increase its success rate, rather than on asking whether the right tasks are being done in the most appropriate way. Therefore, this thesis will use sources from across this tradition, the morphological process, and methods as one of the main sources of inspiration for developing the Malay Corpus and a new ontology-based Information Retrieval(IR) with semantics using Natural Language Processing (NLP) techniques. As it will be discussed further in Part II, later works that build on this tradition, such as the analysis on Malay language used in the Malay Translated Qur'an, will be used as the primary reference for the research work.

Chapter 4

Issues in Translation between the English, Malay and Arabic languages

This chapter is based on the work published in the 4th International Conference on Islamic Applications in Computer Science and Technology, 20-22 December 2016, in Sudan. This chapter investigates and discusses how the differences and similarities between the Malay, English, and Arabic languages play important roles in the translation of materials. The characteristics discussed include not only issues at the sentence and word levels, but also the morphemes that make up the two languages.
PUBLISHED: Ahmad, N. D., Bennett, B., and Atwell, E. (2017). Retrieval Performance for Malay Quran. International Journal on Islamic Applications in Computer Science and Technology (IJASAT), 5(2), 13-25.

4.1 Motivation

The translation of documents continues to be carried out today, and one of the most important elements in this document exchange is the role played by the English language. The status of the English language as a major world language has resulted in a growth of works written in English. One of the most distinguished developments arising from this situation is the practice of translating works in English into other languages such as Malay language. The reason behind this attempt is because it can reach a wider audience.

The Malay language, spoken in Malaysia, Indonesia, Brunei, and Singapore, has been used as a lingua-franca for centuries, yet the translation between English and Malay still brings about many challenges for both native English speakers and native Malay speakers because of the vast differences between the two languages [Nurkhalisah Mustapa, 2013].

The Qur'an is fundamental to all Muslims because it contains comprehensive guidance to Muslims in all aspects of life. The language of the Qur'an is Arabic. Nowadays, the Qur'an has been translated into various languages around the world by Muslim experts. The main aim of the availability of the Qur'an translations is to allow the reader to understand the Qur'an in clearer ways. However, there are a few translation issues regarding the Malay Qur'an translation, such as ambiguity of words, lack of word equivalence between Malay and Arabic or Malay with English, and different structures of words, sentences, and discourse in these two languages [Tabrizi and Mahmud, 2013].

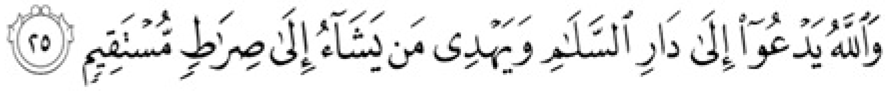
4.2 Issues in Translation between English and Malay

Numerous issues can be identified in translating a text from one language to another language. Two of them are; non-equivalence and and the discourse of the two languages [Mohd Don, 2010]

4.2.1 Non-equivalence

According to [Nurkhalisah Mustapa, 2013], the non-equivalence that occur at word level are one of the toughest parts that most translators face when translating texts. Firstly, the concept is not lexicalized. The concepts or words may be understood in the target language, but there may not be any word to express it. For instance, English does not have any term that distinguishes older and younger siblings and girls because the terms ’brother’ and ’sister’ are general terms. However, in Malay, the word ‘sister’ can be either an older sister (*‘kakak’*) or a young sister (*‘adik’*). This translation of this word should look at the whole form and meaning of the sentence.

Secondly, the source language may be used in more specific terms, while the target language only has a general term. This problem can be seen in the Malay Translated Qur’an as the Malay language appears to have longer sentences compared to English. Figure 4.1 shows an example of the problem.



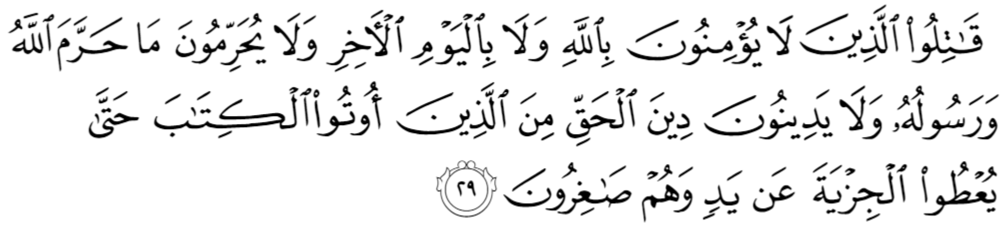
وَٱللَّهُ يَدْعُوٓا۟ إِلَىٰ دَارِ ٱلسَّلَٰمِ وَيَهْدِى مَن يَشَآءُ إِلَىٰ صِرَٰطٍ مُّسْتَقِيمٍ ٢٥

Figure 4.1: Surah Yunus Verse 25

Sahih International: And Allah invites to the Home of Peace and guides whom He wills on a straight path. (Surah Yunus,10:25)
Malay Translation: *“Dan Allah menyeru (manusia) ke Darussalam (syurga) dan memberikan petunjuk kepada orang-orang yang Dia kehendaki ke jalan yang lurus (Islam)”.* (Surah Yunus, 10:25)

As we can see from the example of Surah Yunus, chapter 10, verse 25, the translation in English only used 17 words to translate the Arabic words while in Malay, 21 words were used. Besides, Malay translators tend to explain in detailed using brackets for the specific term. The difference in translation between these two languages has caused a mismatched result when searching for keywords in getting information. This will be discussed further in the experiment section in this chapter.

Thirdly, concepts in the source language may not have other names to describe the concept in the target language. This happens when the source language is a noun or a special case. Although it is rare, but it can be found in the translation of the Qur’an. Normally, the targeted language will use the word in the source language in the translation. For instance, the word ‘jizyah’ in chapter 9 and verse 29 (see figure 4.2) comes from Arabic words. This word does not appearing in the other language, so the translation of the Qur’an just used the same Arabic word in their translation. In this case, when we find this type of word, we will use according to it the source language. We will explain it more in chapter 5.



قَٰتِلُوا۟ ٱلَّذِينَ لَا يُؤْمِنُونَ بِٱللَّهِ وَلَا بِٱلْيَوْمِ ٱلْءَاخِرِ وَلَا يُحَرِّمُونَ مَا حَرَّمَ ٱللَّهُ
وَرَسُولُهُۥ وَلَا يَدِينُونَ دِينَ ٱلْحَقِّ مِنَ ٱلَّذِينَ أُوتُوا۟ ٱلْكِتَٰبَ حَتَّىٰ
يُعْطُوا۟ ٱلْجِزْيَةَ عَن يَدٍ وَهُمْ صَٰغِرُونَ ٢٩

Figure 4.2: Surah At-Tawbah Verse 29

Sahih International: "Fight those who do not believe in Allah or in the Last Day and who do not consider unlawful, what Allah and His Messengers have made unlawful and who do not adopt the religion of truth from those who were given the Scripture - [fight] until they give the jizyah willingly while they are humbled". (At-Tawbah, 9:29)
Malay Translation: *"Perangilah orang-orang yang tidak beriman kepada Allah dan Hari Kemudian, mereka yang tidak mengharamkan apa yang telah diharamkan Allah dan Rasul-Nya dan mereka yang tidak beragama dengan agama yang benar (agama Allah), (iaitu orang-orang) yang telah diberikan kitab, sehingga mereka membayar jizyah dengan patuh sedangkan mereka dalam keadaan patuh".* (At-Tawbah, 9:29)

Additionally, the concept has many other meanings in the source language. This is another challenging problem in translation between the English, Malay, and Arabic languages. Sometimes, the source concepts have either straight or hidden meaning. The problem will become complicated when the concept has hidden meanings. This usually happens in the Qur'an. For instance, the word الجنة in Arabic can give a different meaning as in figure 4.3 below. Here, when the user queries the word الجنة with the intention to find the verses related to paradise, the user will retrieve other verses that contain the word الجنة but with a different meaning.

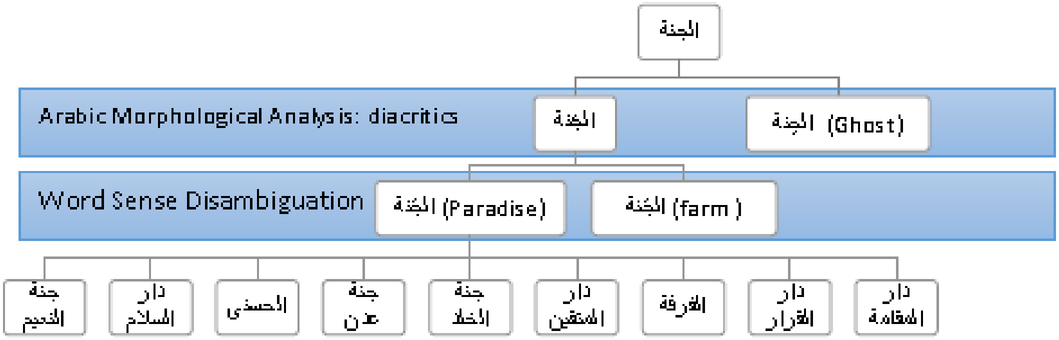


Figure 4.3: The classification of the meaning of word الجنة [Alqahtani and Atwell, 2016]

The ambiguity of the query and returned results can be one of the major constraints. This ambiguity happens usually in understanding the actual meaning of the user's query and giving the right results. The ambiguity issue is the most challenging part in text analysis for POS tagging. The same word or concept may have different meanings in different contexts. This may result in improper text analysis and thus inaccurate POS tags.

Lastly, [Wahid, 2011] claimed that standardization of different versions of Malay Qur'an was challenging due to the need of alignment of meanings. Different versions of Malay Qur'an surfaced due to the manner translations were made, which could likely be from a secondary source for some of the texts. Standardization issues could be reduced if the translations were made from the original texts with references to other authentic sources like Hadith or Tafsir.

4.2.2 The discourse of the two languages

There are many other grammatical differences that exist in the Malay language that have no counterpart in English, making word-for-word translation difficult. For instance, while English typically adds a letter 'S' to the end of a noun to make it plural, Malay often repeats the word. A native English speaker might expect *"batu-batu"* to translate literally as *"stone-stone"* but it is the plural form *"stones"*. Still other Malay words form plurals by only repeating the beginning sound of the noun. The plural *"dedaun"* (leaves) comes from the singular *"daun"* (leaf). Because the plural is considered an entire word by itself and because it does not start the same way as the singular, English speakers may easily find it unrecognizable as the plural form of *"daun"*

4.3 Preliminary Experiment

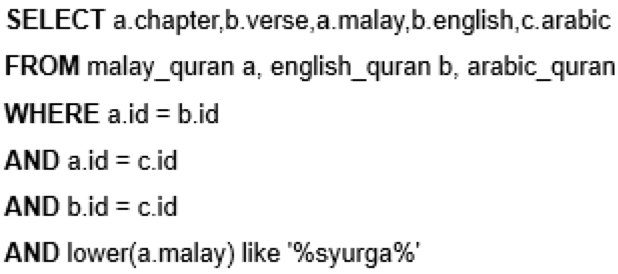
Preliminary experiment was conducted to investigate and prove the issues stated in the previous section between English, Arabic, and Malay translations. Besides, the purpose of this experiment is to see the differences in information retrieval between the three languages. Here, we can measure the accuracy in terms of retrieval between these three languages.

In this experiment, we used the Malay translated Al-Qur'an Amazing book published by *Karya Bestari* [Nursalim et al., 2016] and the English translation of the Qur'an of Abdullah Yusuf Ali (YA) (2003) as our sample data. The reason for adopting the Malay translated Qur'an book is because the translation has been reviewed and verified by JAKIM, Department of Islamic Development, Malaysia. In addition, most of the concepts in this book have been indexed based on chapters and verses. The Qur'an is divided into 114 chapters (Suras) of varying sizes, where each chapter is divided into verses (Ayahs). There are 6,234 verses in the Qur'an. The Qur'anic data is then stored in Oracle 11g database. The detailed explanation about each experiment is as per below.

4.3.1 Experiment 1: Finding the overlapping words between Malay and English words

The aim of this experiment is to find overlapping words between Malay and English words. We want to see the similarities and connection between the words and verses from both English and Malay translations of the Qur'an. Besides, this experiment can also compare the number of retrieved and missed verses retrieved using synonyms in English and Malay languages.

For this experiment, the translation of Surah Al-Baqarah, the largest chapter of the Qur'an has been taken as a sample text. In this experiment, we have chosen the concept of Afterlife which is الجنة 'Jannah' and Hell. According WordReference.com, *‘Jannah’* means heaven or paradise and in Malay language, it is called as *‘syurga’*. The retrieved verse is found by doing a simple query over the database by using the SQL queries in Figure 4.4 below:



```
SELECT a.chapter,b.verse,a.malay,b.english,c.arabic
FROM malay_quran a, english_quran b, arabic_quran
WHERE a.id = b.id
AND a.id = c.id
AND b.id = c.id
AND lower(a.malay) like '%syurga%'
```

Figure 4.4: SQL Query to retrieve the concept from database

The next process is to get the relevant verses of each word. This process is essential, so we have checked it manually by using three sources such as the Tafsir, Ibn Khatir (reference: www.qtafsir.com, online), the list of the concept of N H Abbas in [N H Abbas, 2009] and the thematic index from Uthmani Malay Version of the Qur’an (reference: text book). This thematic index has classified verses of the Qur’an according to the themes in the Malay language. The list of the concept of N H Abbas in [N H Abbas, 2009] has classified the verses according to the Qur’an in English. Then, we cross referenced again the verses that appear on the thematic index with online Tasfir of Ibn-Khatir to verify the relevance of these verses. The result of verses is then classified as relevant. Table 4.2 shows the number of retrieved, relevant, and missed verses retrieved using the keyword search for the concept of *‘Jannah’* in English and Malay.

Table 4.1: The keyword search result in English and Malay

Language	Keyword	No of relevant verse	No of retrieved verse	No of missed verse
Malay	Syurga	12	10	2
English	Heaven	15	11	4
English	Paradise	11	1	10
English	Hell	18	1	17
Malay	Neraka	18	16	2

Based on the results on Table 4.1, there is a big difference between the number of retrieved and relevant verses in English words for Paradise and Hell. For the words Paradise and Hell, only one verse had been retrieved by the database, whereas there were more relevant verses of 11 and 17 times respectively. Besides, we found that the Malay language has missed two verses that cannot be obtained using keywords *'syurga'*. This is because, in the Malay translated Qur'an, it uses nouns to describe *'syurga'* such as *'Darussalam'* or *'Firdaus'*. This shows that using keyword-based search is not enough to capture all the accurate information of a particular keyword.

After that, we analyze each relevant and retrieved verses. This process of analysis is to find the other words or similar word used by both English and Malay translations of the Qur'an that represent the concept of *'Jannah'*. Results (see figure 4.5) show the other words used to describe *'syurga'* in English terms are as Hereafter, The Garden, Paradise, and Home. In this case, The Garden is the most widely used reference in the Qur'an and has high similarities with *'syurga'*. Surprisingly, there is no connection or relation between the words *'heaven'* and *'syurga'*. This is surprisingly unexpected because Malays normally used the word 'heaven' to describe *'syurga'*. However, in this Malay Qur'an translation, the word 'heaven' has been translated as *'langit'* (sky). So, it is not matched with the verses that contain the word *'syurga'*. In this scenario, when the user using heaven as the keyword, they will get irrelevant and incorrect verses.

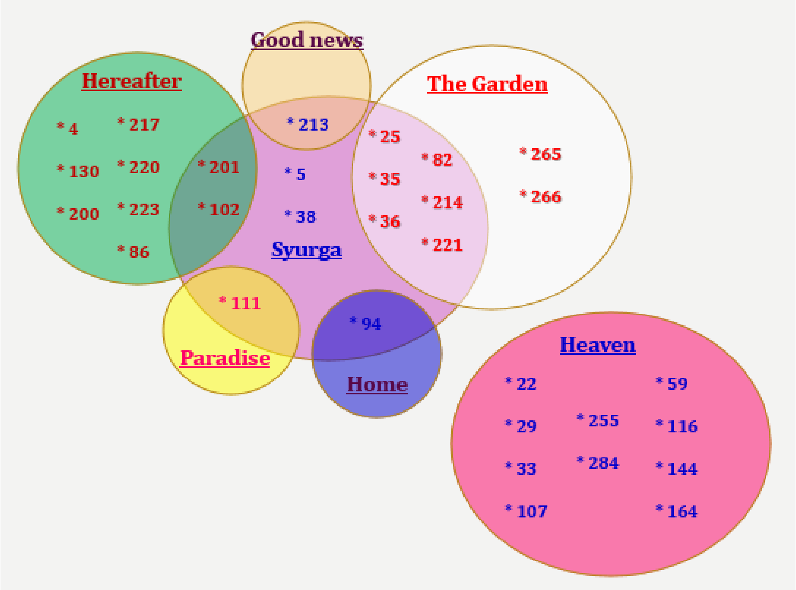


Figure 4.5: The other words used to describe *'syurga'*, heaven and paradise

The other concept of afterlife is Hell. The word hell is often used to describe *'neraka'*. However, the analysis results turned out differently (see Figure 4.6). The most widely used word in this translation is 'The Fire'. Out of 18 verses, 14 were found used this word. In this case, when the user submits the query using the word hell, he or she will get only one verse (verse 206).

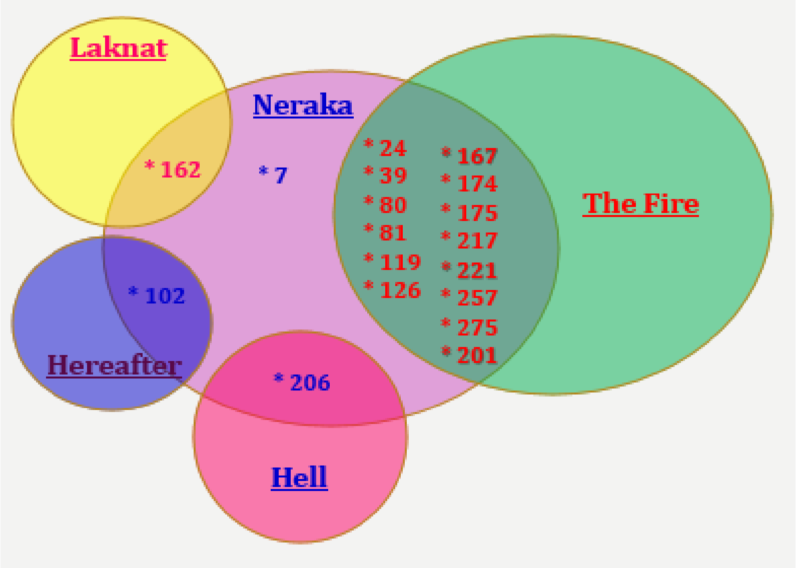


Figure 4.6: The other words used to describe *'neraka'* and hell

In this experiment, we can see in English, there are many other words used to describe 'Hell' such as 'The Fire' and 'Hereafter'. In contrary, in the Malay language translation, only the word *'neraka'* is used to describe hell.

4.4 Summary

As for the conclusion of this experiment, some of the findings are:

1. The translation plays a vital role in providing true and correct interpretation in every respect of the verse in the Qur'an. If not, it can cause confusion, especially if the translation is done with two languages such as English and Malay. The widely spoken words must be checked regularly with the authorizing body whether it is right to be used and whether it is used in a correct meaning.

2. The keyword based search alone is not enough to provide accurate and relevant results. It requires a combination of semantics and it takes from several sources that are authentic such as Tafsir Ibn Khatir.

3. The lack of word equivalence between Malay and English is one of the major problems derived from this experiment.

All of these findings are very important to create a search engine that can provide an accurate information for every search query that users want to search. It becomes the starting point for solving the problems that occur on the Qur'an.

Part II

Implementation of Development

Chapter 5

Malay Translated Qur'an Corpus Development

This chapter discusses the processes involved in creating Malay Translated Qur'an corpus with semantics, which started with data collection and preparation. The steps involved are data collection, data analysis, data preprocessing, and data annotation. A detailed description of each step will be explained below in this chapter. The creation of this corpus is essential for the development of ontology-based IR.

5.1 Introduction

Research in Natural Language Processing and semantic search for English has shown successful results for more than a decade. However, it is difficult to adapt those techniques to the Malay language, because its complex morphology and orthographic forms are very different from English. Moreover, limited resources and tools for computational linguistic analysis are available for Malay. The study in Information Retrieval of Malay documents is relatively new. The first study was conducted by [Ahmad, 1995] and she has laid a solid foundation for further research in this field. After that, many researchers began conducting studies on Malay documents by applying new computational and linguistic approaches. However, since then, resources and tools for computational linguistic analysis for this language are still limited. Moreover, most of the studies that have been conducted only focus on one method of improving information retrieval.

The aim of this chapter is to demonstrate the use of Malay morphological analysis and how it can help in retrieving accurate information.To present the use of Malay morphological analysis, we need to describe a morphological analysis algorithm for the Malay translated Qur'an that allows it to be created, discovered, and queried. This involved the creation of the Malay translated Qur'an Corpus and a new morphological analysis algorithm for Malay Translated Qur'an. This includes a new list of Malay Stop words, a new rule-based stemming algorithm, and a new root words annotation. This new resource will benefit other researchers especially in religious morphological analysis and semantic modelling in research related to the Malay documents.

5.2 Corpus Development

Corpus is a systematic collection of pieces of language text in electronic form that will represent as far as possible language features relevant for computational linguistic research. Most corpus developments are based on scientific reasons required for the research. The principle of selecting the contents of a corpus should be based on the consideration of the communicative function of the text in the community in which it appears [Wynne, 2005]. For instance, in building modern corpus for Arabic by [Dukes et al., 2010] consider a collection of texts that will mostly reflect the reality of the language itself. Thus, building a corpus is necessary to represent the value of the language resources that support research and technology development in language.

There has been a significant development in the process of creating a corpus in some European and Asian country in their native languages. However, there has been no such recognizable attempt to create Malay corpus as such. Presently, Dewan Bahasa dan Pustaka (DBP) Corpus is the only Malay language corpus existed, which consists of 114million of words taken from various sources from modern to classical Malay texts. Regrettably, this corpus is not publicly available and has limitations in studying modern Malay [Mohd Don, 2010]. As for Malay translated Qur'an corpus, the words used in the translation are taken from both modern and classical Malay. However, there are 286 out of 149,654 words used in this translation which are derived from Arabic words (see appendix C). Because of this multilanguage used in this document, there is a need to extract, compile, categorize, and annotate this corpus and reuse the word for other morphological analysis.

In this part, we described the steps and methods used to collect the data for the implementation of the Malay translated Qur'an corpus. The steps involved are data collection, data analysis, data preprocessing, and data annotation. This process is important in the evaluation of retrieval performance for question and answer search. A detailed description of each step will be explained below.

5.2.1 Data Collection

Data collection is an essential part in corpus development. For this study, we were looking for a digital version of Malay translated Qur'an. There are few digital versions of Malay translated Qur'an available online. However, there is no digital version of the Qur'an that meets our criteria. This is because most Malay digital versions of the Qur'an translation are using a combination of Malay and Indonesian language. Therefore, we decided to use the Al-Qur'an Amazing book published by *Karya Bestari* [Nursalim et al., 2016] as our dataset. The purpose of using this Malay Translated Qur'an is because the Malay words used in the translation has been reviewed and verified by JAKIM, Department of Islamic Development, Malaysia. Besides, it contains modern and Classical Malay where both are important in the construction of corpus.Meanwhile, this Malay translated Qur'an has a list of topic indices that can be used as relevant documents in the evaluation phase.

The selection of the Qur'an as dataset is not only because it was widely referred to, but also because many studies have been done by researchers in the field of Quranic studies. The advantages of this book are:

1. The words used in this translation are fully Malay.
2. The process of finding relevant topic or concept for queries can be easily obtained based on the list of topic index in this translation. All these topics

have been reviewed and verified by JAKIM. Each topic or concept is indexed by verse.

3. The relevance of the verses and the topics for a query is based on Tafseer Jalalain and not just a word match.

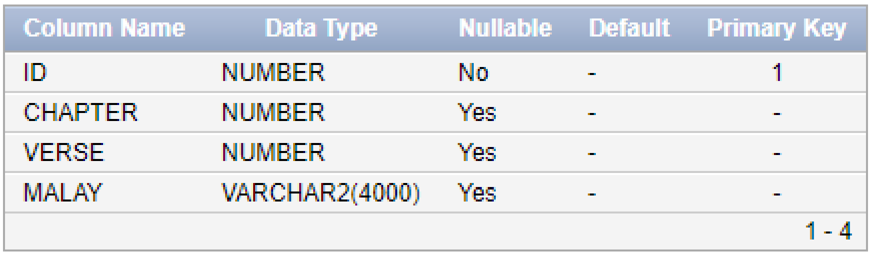
The Qur'an is divided into 114 chapters (Suras) of varying sizes, where each chapter is divided into verses (Ayahs). There are 6,234 verses in the Qur'an. Since the data is from the book, the extraction process must be done. In this study, the extraction process is done manually by extracting every chapter and verse. Because of the extraction process is done manually, the probability of error is high. Thus, data analysis becomes an ongoing iterative process where data is continuously collected and analyzed almost simultaneously.

5.2.2 Data Analysis

Data analysis is intended to validate the data gathered. Data analysis is done to ensure that all chapters, verses, and words are correct. Two levels of validation have been done on the data for this study. The first validation is to analyze and compare the structure and content of Malay Qur'an translation with English and Arabic translations. This is to ensure the structure and words used are synchronized. The second validation is by two experts in the field of Quranic studies. Both experts will check and ensure the chapters, verses, and words used in the translation are correct. The Qur'an analyzed data is then stored in the Oracle 11g database. Table 5.1 shows the statistics of data extracted from the Malay translated Qur'an book. Figure 5.1 shows the database file format used to store the extracted data in database.

Table 5.1: Statistic of extracted data.

Data	Statistic
The total number of chapters	114
The total number of verses	6,236
The total number of words	149,654



Column Name	Data Type	Nullable	Default	Primary Key
ID	NUMBER	No	-	1
CHAPTER	NUMBER	Yes	-	-
VERSE	NUMBER	Yes	-	-
MALAY	VARCHAR2(4000)	Yes	-	-
				1 - 4

Figure 5.1: Database file format

5.2.3 Preprocessing

This phase aims to prepare the Malay translated Qur'an data and its structure to be used in the data annotation process. One of the important parts in morphological analysis is preprocessing. Mostly the processes involved in this phase are from Information Retrieval process to question answering search. Stop word removal, tokenization, reduplication, and stemming are the techniques used in this preprocessing phase. The importance of preprocessing in this research are:

1. Cleaning: This process is to remove unwanted parts of text such as punctuation marks, stopwords, capital letters and other characters that appeared in text.

2. Normalization: This process is important for Information Retrieval process. It will retrieve the base form of the word with reducing the dimensionality of size of the index words usually through Stemming, Lemmatization and other forms of standardization.

3. Analysis: This analysis process usually consists of statistical and visualization of data.

The architecture illustrates the use of NLP techniques in processing Malay translated Qur'an texts. Figure 5.2 shows the architecture for morphological analysis of MyQOS Corpus. To handle this morphological analysis, we divided the analysis process into several different preprocessing stages; performing the simplest analysis process to the most complex analysis. The text preprocessing is built using Python 3.4. Figure 5.3 shows the algorithm for preprocessing text.

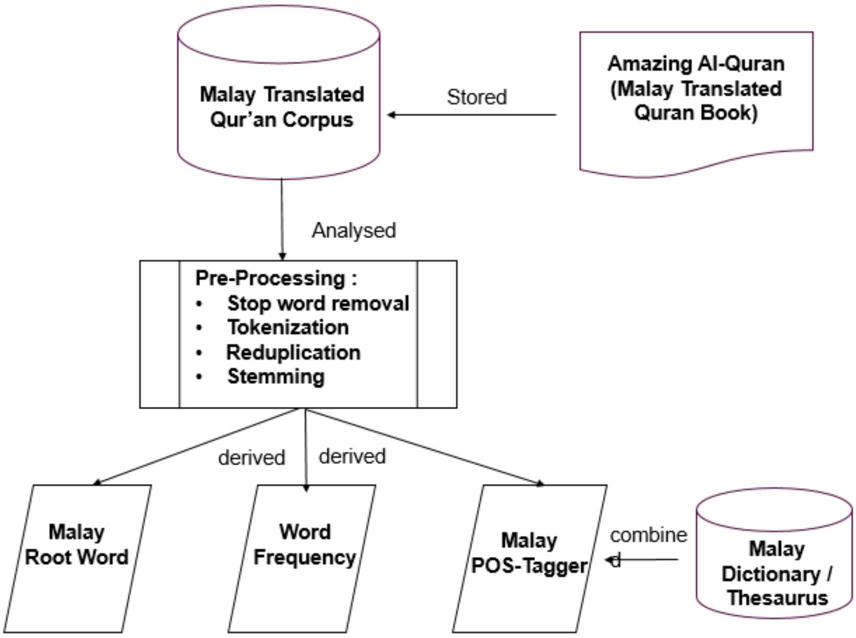
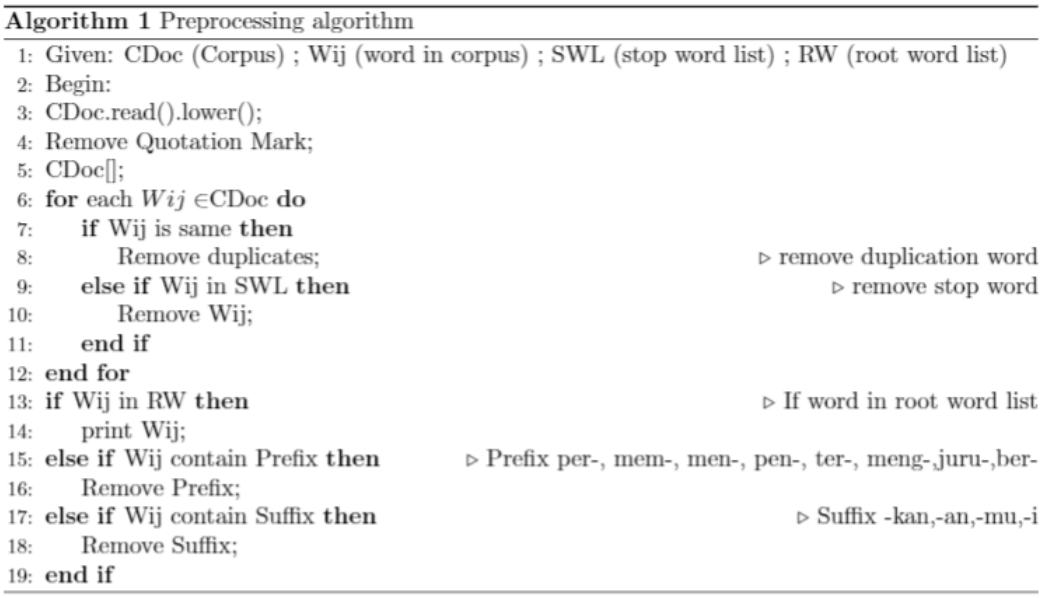


Figure 5.2: The architecture morphological analysis of MyQOS Corpus development



```
Algorithm 1 Preprocessing algorithm
Given: CDoc (Corpus) ; Wij (word in corpus) ; SWL (stop word list) ; RW (root word list)
Begin:
CDoc.read().lower();
Remove Quotation Mark;
CDoc[];
for each $Wij$ ∈CDoc do
    if Wij is same then
        Remove duplicates;                                  ▷ remove duplication word
    else if Wij in SWL then                                 ▷ remove stop word
        Remove Wij;
    end if
end for
if Wij in RW then                                           ▷ If word in root word list
    print Wij;
else if Wij contain Prefix then    ▷ Prefix per-, mem-, men-, pen-, ter-, meng-,juru-,ber-
    Remove Prefix;
else if Wij contain Suffix then                             ▷ Suffix -kan,-an,-mu,-i
    Remove Suffix;
end if
```

Figure 5.3: Pre-processing Algorithm

Capitalization and Character removal.

Firstly, the words in the text are formatted into lowercase. Then, the punctuation marks need to be removed. There are many punctuation marks used in Malay text such as (!, ?, ( ), ., “, *, ”,). Removing punctuation marks is not an easy task since punctuation marks that mark the end of a sentence are often ambiguous. To disambiguate punctuation marks, it often relies on regular expressions. Here, regular expression rules are applied to remove the punctuation marks.

Split the reduplication words.

Secondly, splitting reduplication words appeared in this corpus. Reduplication is a word-formation process in which meaning is expressed by repeating all or a part of a word. Splitting the reduplication words is important to gather the root words. Reduplication word is widely used in many Malay texts. Reduplication is hardly found in other languages. Reduplication is usually found in Austronesian language such as Malay. Reduplication indicate the simple plural or many. In Malay, if the user knows, there is more that one of an object, but does not know or does not wish to specify how many the whole form of the noun may be simply be repeated twice to signal the plural. For example, if there is more than one cat or *'kucing'*, you will say cat or *'kucing-kucing'* in Malay.

Generally, reduplication in Malay can be divided into full, such as *'perempuan-perempuan'* or girls (from *'perempuan'* or girl), or partial, such as *'lelaki'* or boys (from *'laki-laki'* or boys) or rhyming and chiming, such as the word *'kayu'* or wood combines with *'kayan'* or wood to form *'kayu-kayan'* different sorts of wood (see Table 5.2).

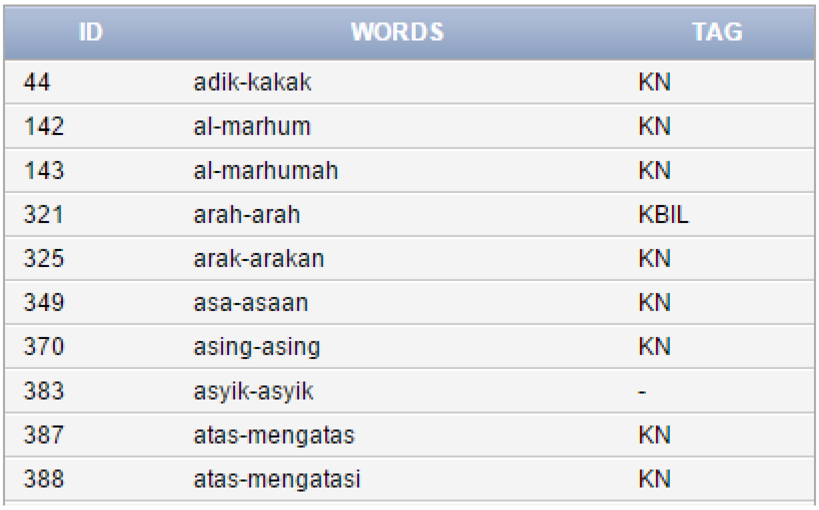
Table 5.2: Example of reduplication in the Malay text

Type of Reduplication	**Example**	**Root words**
Full	*'Perempuan-perempuan'* (girls)	*'Perempuan'* (girl)
Partial	*'Pepatung'* (dragonfly)	*'Pa + 'Patung' (Statue)*
Rhyming or chiming	*'Gunung-ganang'* (mountains)	*'Gunung Ganang'* (mountain)

In this Malay Translated Qur'an, we found 1,587 words is reduplication. According to Malay linguists [Mohd Don, 2010], only the first word which is considered as the root word and the second word should be removed. However, in this research, the second word will not be discarded as some of the second words are root word. Considering this scenario, the reduplication process is done in semi-automatic process. First, we used Python code to split the reduplication words. After that, we checked each word manually and indicate whether the word is a root word or not. Figure 5.4 shows the list of reduplication words found in the Malay translated Qur'an text, while Figure 5.5 shows the results after the splitting of reduplication words. Below shows the example of reduplication appeared in the Malay translated Qur'an and the English translated Qur'an.

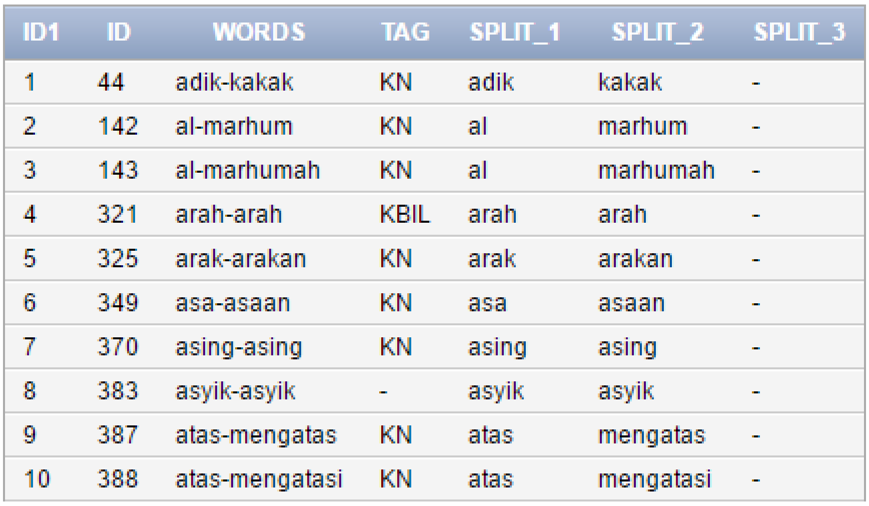
- Malay Translated Qur'an: *"Dan apabila mereka bertemu dengan* ***orang-orang*** *yang beriman, mereka berkata, "Kami telah beriman". Tetapi apabila mereka kembali kepada* ***syaitan-syaitan****(para pemimpin) mereka, mereka berkata,"Sesungguhnya kami bersama kamu, kami hanya* ***berolok-olok****."""* (Al-Baqarah, 14)

- English Sahih International: *“And when they meet those who believe, they say, “We believe”; but when they are alone with their evil ones, they say, “Indeed, we are with you; we were only mockers.””*(Al-Baqarah, 14)



ID	WORDS	TAG
44	adik-kakak	KN
142	al-marhum	KN
143	al-marhumah	KN
321	arah-arah	KBIL
325	arak-arakan	KN
349	asa-asaan	KN
370	asing-asing	KN
383	asyik-asyik	-
387	atas-mengatas	KN
388	atas-mengatasi	KN

Figure 5.4: Example of reduplication words in Corpus

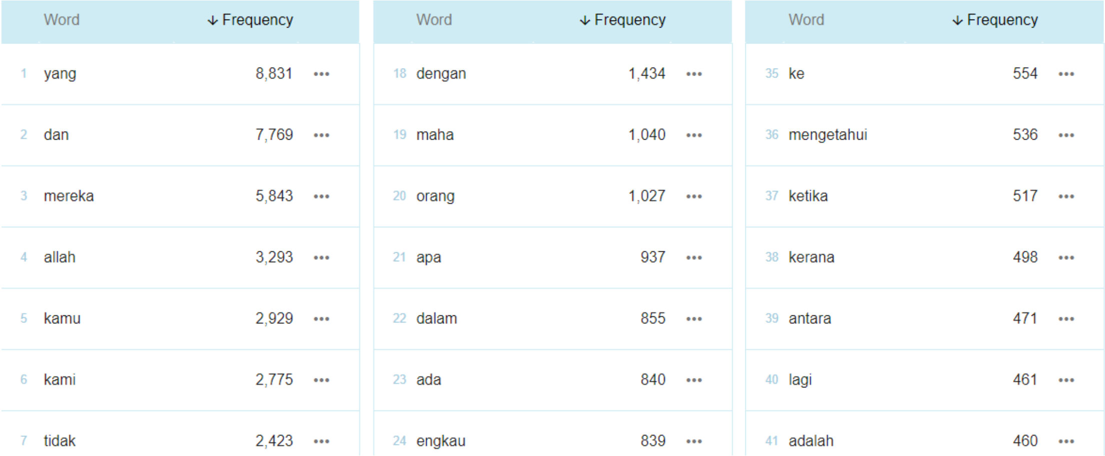


ID1	ID	WORDS	TAG	SPLIT_1	SPLIT_2	SPLIT_3
1	44	adik-kakak	KN	adik	kakak	-
2	142	al-marhum	KN	al	marhum	-
3	143	al-marhumah	KN	al	marhumah	-
4	321	arah-arah	KBIL	arah	arah	-
5	325	arak-arakan	KN	arak	arakan	-
6	349	asa-asaan	KN	asa	asaan	-
7	370	asing-asing	KN	asing	asing	-
8	383	asyik-asyik	-	asyik	asyik	-
9	387	atas-mengatas	KN	atas	mengatas	-
10	388	atas-mengatasi	KN	atas	mengatasi	-

Figure 5.5: After splitting the reduplication words

Tokenization

The following process is tokenization. The purpose of tokenization is to split a text into meaningful units called tokens. In this research, tokenization is required to gather root words to facilitate the information retrieval process [Ahmad, 1995]. There are 149,654 tokens that have been extracted from the corpus. The extraction process used two software such as Sketch Engine and Nvivo 10. Figure 5.6 shows an example of tokens derived from Malay translated Qur'an.



	Word	↓ Frequency		Word	↓ Frequency		Word	↓ Frequency
1	yang	8,831	18	dengan	1,434	35	ke	554
2	dan	7,769	19	maha	1,040	36	mengetahui	536
3	mereka	5,843	20	orang	1,027	37	ketika	517
4	allah	3,293	21	apa	937	38	kerana	498
5	kamu	2,929	22	dalam	855	39	antara	471
6	kami	2,775	23	ada	840	40	lagi	461
7	tidak	2,423	24	engkau	839	41	adalah	460

Figure 5.6: Results from the tokenization process

Stopword removal

The elimination of stopword is performed. A stopword is a word which does not carry meaning in natural language and therefore can be ignored.Stop words are commonly eliminated from many text processing applications because these words can be distracting, non-informative and are additional memory overhead. Removal of stopwords from corpus also leads to its decreased size, which increases the efficiency of any NLP activity [Raulji and Saini, 2017].

In this research, we used 356 stopword from [Ahmad, 1995]. However, after go through the stopword, we found there are several word that are duplicate and irrelevant. Thus, we have to eliminate those words. Therefore, the total number of stopwords used in this research is 318 words.

Fortunately, the most frequent words appeared in this research after the tokenization process are stopwords, and therefore half of the words appearing in a text do not need to be considered. This allows, for instance, a significant reduction in the space overhead of indexes for natural language text. A majority of the stopwords in a given text are connection parts of a sentence rather than showing the subject, object, or intent. Table 5.3 shows the analysis result before and after the elimination of the stop words based on the seven (7) longest chapters in the Qur'an. As we can see from the results, the number of words containing stopwords is more than the words that have subjects or objects. Appendix A is the list of Malay stopword used in this preprocessing phase.

Table 5.3: The Analysis Result of Before and After Elimination of Stopword

Chapter number / documents	**Total number** of words	**Total number of words** (After removed the stopwords)	**Total number of words** (contains stopwords)
(2) Surah Al-Baqarah (The Opening)	11294	4848	6446
(4) Surah An-Nisa (The Women)	7357	3173	4184
(3) Surah Ali' Imran (The Family of Imran)	6548	2770	3778
(7) Surah Al-A'raf (The Heights)	6538	2801	3737
(9) Surah At-Taubah (The Repentance)	4918	2071	2847
(10) Surah Yunus (Jonah)	3387	1394	1993
(12) Surah Yusuf (Joseph)	3357	1455	1902
TOTAL	**43,399**	**18,512**	**28,887**

Stemming

Stemming is a computational process of reducing a word from its derived form into its root term. Stemming is useful for improving retrieval performance because they reduce variants of the same root word to a common concept. Furthermore, stemming has the secondary effect of reducing the size of the indexing structure because the number of distinct index terms is reduced. In the stemming algorithm, words with the same root are reduced to a common form by stripping each word of its derivational and inflectional suffixes [Fadzli et al., 2012]. Malay language affixes consist of four different types of verbal elements:

1. Prefix: attaches itself at the beginning of a word. Example: *'bersalah'* which means guilty, start with *'ber'* which is a typical prefix in Malay language. Another prefix appeared in Malay language such as *'per'*, *'mem'*, *'men'*, *'pen'*,

‘ter’, *‘meng’*, and *‘juru’*.

2. Suffix: attaches itself at the end of a word. Example: *‘memaafkan’* which means to forgive, ends with *‘kan’* which is a typical suffix in Malay language. Another suffix such as *‘an’*, *‘i’* and *‘mu’*.

3. Infix: usually located in the middle of word. Example: *‘gerigi’* which means toothed blade, which is derived from the root word *‘gigi’* (teeth).

4. Circumfix: Prefix-suffix pair where more than one affix that is attached to a word at the same time and usually positioned before and after the root word. Example: *‘kerajaan’* which means kingdom, is derived from root word *‘raja’* (king).

Studies in the stemming algorithm for Malay language are relatively left behind in comparison to other languages such as English and European languages. The availability of Malay information retrieval system is also very limited. The usage of affixes in English and other European languages is less complex than Malay language as it has been found that the stemmers are only concerned with the removal of suffixes. However, in Malay morphology, a stemmed word is produced by removing affixes in the text, document or query. Affix is the verbal element that attaches to the word whether at the beginning of the word (prefix) and at the end of the word (suffix). Besides, more than one affix may also be attached to a word at the same time. The word also can contain both affixes and this is known as prefix-suffix pair, for example as seen in the word *‘pemakanan’*. The root word for this word is *‘makan’* and the prefix is *’pe’* is added at the beginning of the word and the suffix *’an‘* at the end of the word to complete the word *‘pemakanan’*.

English and Malay languages differ in terms of their root words, which are based on their respective morphological structures [Abdullah et al., 2009]. For instance, the English words *‘related’*, *‘relates’*, and *‘relation’*, are derived from the root word *’relate’*, and stemmer can work as suffix removal for English language. Yet, the Malay language has a different stemming process compared to English, due the complexity of its morphological rules. For example, the Malay words *‘pengajaran’*, *‘pembelajaran’*, and *‘pelajar’* are derived from the root word *‘ajar’*, and it is insufficient to use suffix removal to decide on the perfect root word [Ulah Khan et al., 2017].

According to [Hassan, 2002], affixes are used to add a word's meaning. Table 5.4 shows the affixes, type of affixes, and category of words for every affix based on the rules of Alkhawarizmi word labeling. Figure 5.7 shows the flow diagram contains 7 steps of rules applied in this stemming process for Malay translated Qur'an Corpus;

Table 5.4: Affixes labeling based on Alkhawarizmi's Rule

Affixes	**Type of Affixes**	**Category of Words**
pe	prefix	Noun
pen	prefix	Noun
pem	prefix	Noun
peng	prefix	Noun
penge	prefix	Noun
ke	prefix	Noun
per	prefix	Noun
juru	prefix	Noun
an	suffix	Noun
wan	suffix	Noun
wati	suffix	Noun
per+an	infix	Noun
pen+an	infix	Noun
ke+an	infix	Noun
be	prefix	Verb
bel	prefix	Verb
ber	prefix	Verb
per	prefix	Verb
me	prefix	Verb
men	prefix	Verb
mem	prefix	Verb

Continued on next page

Table 5.4 – continued from previous page

Affixes	Type of Affixes	Category of Words
meng	prefix	Verb
menge	prefix	Verb
memper	prefix	Verb
di	prefix	Verb
diper	prefix	Verb
ter	prefix	Verb
i	suffix	Verb
kan	suffix	Verb
diper+i	infix	Verb
me+kan	infix	Verb
me+i	infix	Verb
mem+kan	infix	Verb
mem+i	infix	Verb
meng+kan	infix	Verb
meng+i	infix	Verb
menge+kan	infix	Verb
menge+i	infix	Verb
memper+kan	infix	Verb
memper+i	infix	Verb

The proposed arrangement of the rules applied in the stemming process can be described as follows:

Step 1: *Get the next word until the last word;*

Step 2: *Check if the word is a reduplication word; if yes, choose the first word as the root word.*

Step 3: *Check the word against the dictionary; if it exist, the word is the root word and go to Step 1.*

Step 4: *Check the word spelling with Prefix list; if matched, remove the prefix*

Step 5: *Check the word spelling with Suffix list; if matched, remove the suffix*

Step 6: *If the prefix is removed in Step 4, check the beginning spelling; if a missing letter is found, restore it*
Step 7: *If the suffix is removed in Step 6, check the ending spelling; if the word has a suffix of i; if yes, remove it.*

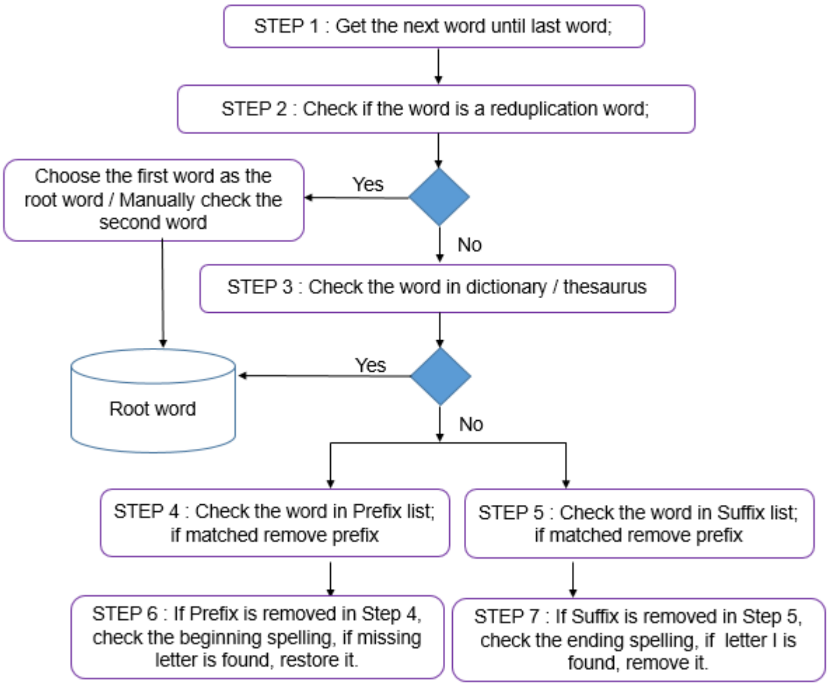

STEP 1 : Get the next word until last word;
STEP 2 : Check if the word is a reduplication word;
Yes
Choose the first word as the root word / Manually check the second word
No
STEP 3 : Check the word in dictionary / thesaurus
Yes
Root word
No
STEP 4 : Check the word in Prefix list; if matched remove prefix
STEP 5 : Check the word in Suffix list; if matched remove prefix
STEP 6 : If Prefix is removed in Step 4, check the beginning spelling, if missing letter is found, restore it.
STEP 7 : If Suffix is removed in Step 5, check the ending spelling, if letter I is found, remove it.

Figure 5.7: Flow diagram of stemming rules

Based on this algorithm, we managed to stem 2,187 root words. Table 5.5 shows the statistics of words after the before and after the stopword removal and stemming processes. As we can see, the number of words after the stemming process is quite low compared to before stemming process. This is due to the use of many affixes in a root word such as ‘mem-’, ‘ber-’, ‘meng-’, ‘-nya’, ‘-kan’ and ‘-I’ in Malay Translated Qur’an. Below is an example of affixes used in Malay Translated Qur’an. Table 5.6 shows the number of occurrences of affixes used in Malay translated Qur’an. According to the table, circumfix is a group of words that are widely used in Malay translated Qur’an.

- Root word: *‘cipta’* (create)
- Prefix: *‘tercipta’* (created), *‘mencipta’* (create)
- Suffix: *‘ciptaan’* (creation), *‘ciptaannya’* (his/her creation)
- Circumfix: *‘menciptakan’* (create), *‘menciptakanku’* (created me), *‘menciptakanmu’* (created you), *‘menciptakannya’* (created it), *‘menciptanya’* (create it)

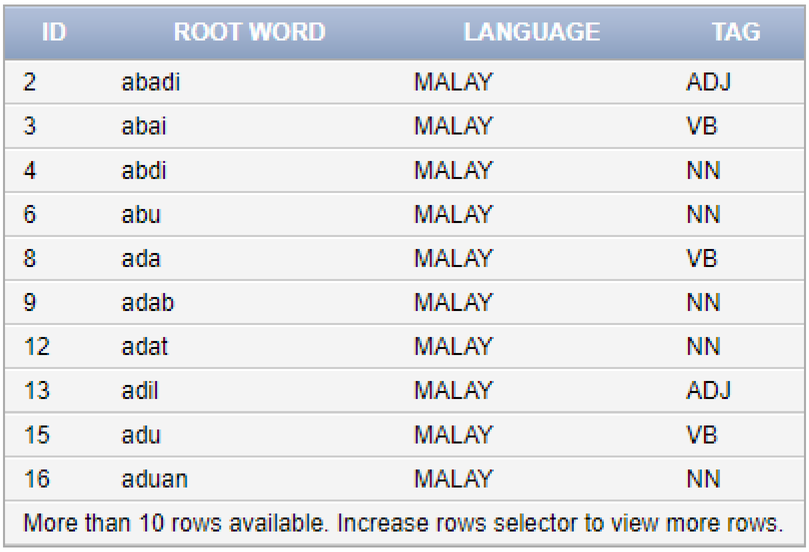
Table 5.5: Statistic of words in Corpus

Words in Corpus	**Total**
Total number of words with stopword	149,654
Total number of words without stopword	63,191
Total number of words after stemming process and without stopword	2,187

Table 5.6: Affixes used in Corpus

Affixes used in Corpus	**Number of Occurrence**
Prefix	10,415
Suffix	7,672
Infix	53
Circumfix	11,800
TOTAL	63,191

Figure 5.8 shows the stemming results stored in Oracle Database. From 63,191 words containing affixes, only 2,817 words are root words. This shows that the Malay language uses many affixes in constructing a sentence. Each word of the affix describes a different meaning. The study by [Ahmad, 1995], [Bakar, 1999] and [Abdullah et al., 2009] showed the accuracy of retrieval can be enhanced if the stopword is removed and stemmed from a word. All the stem words can be used as training resources for other research related to the Malay language, such as semantic annotation, POS-Tagging, Information Retrieval, and Semantic Modelling.



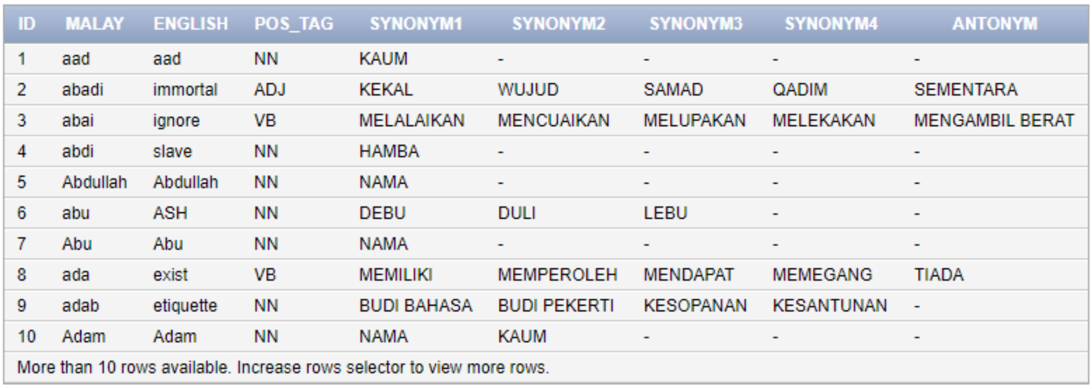
ID	ROOT WORD	LANGUAGE	TAG
2	abadi	MALAY	ADJ
3	abai	MALAY	VB
4	abdi	MALAY	NN
6	abu	MALAY	NN
8	ada	MALAY	VB
9	adab	MALAY	NN
12	adat	MALAY	NN
13	adil	MALAY	ADJ
15	adu	MALAY	VB
16	aduan	MALAY	NN

More than 10 rows available. Increase rows selector to view more rows.

Figure 5.8: Stemming Result stored in Oracle Database

5.2.4 Root Word Annotation

This phase is the core of the study, where the semantic annotation process is carried out for each word in the corpus. The 2,817 root words found in the corpus have been annotated. The annotation process that we use for this study are synonyms and antonyms. These processes were made from scratch because of the lack of digital resources and references. The semantic annotation of synonyms is manually built using Malay Thesaurus by Dewan Bahasa dan Pustaka (DBP), Malay Dictionary and WordNet. These semantic annotations will be used for the creation of a concept for semantic relationship modelling that will be discussed in Chapter 6. Semantic relationship of synonyms is discussed in the next section. Figure 5.9 shows the semantic annotation data stored in Oracle database.



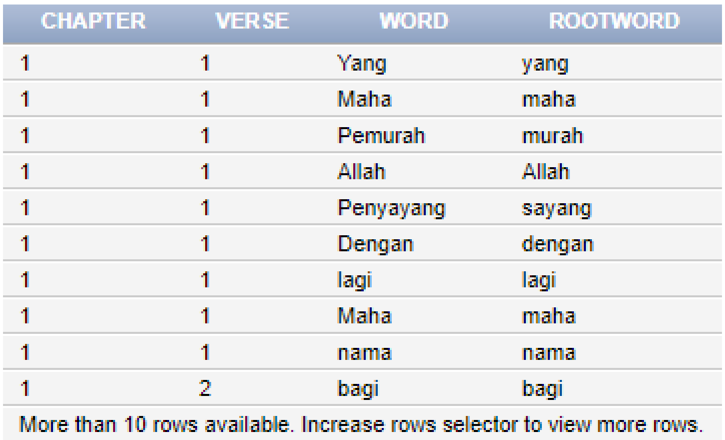
ID	MALAY	ENGLISH	POS_TAG	SYNONYM1	SYNONYM2	SYNONYM3	SYNONYM4	ANTONYM
1	aad	aad	NN	KAUM	-	-	-	-
2	abadi	immortal	ADJ	KEKAL	WUJUD	SAMAD	QADIM	SEMENTARA
3	abai	ignore	VB	MELALAIKAN	MENCUAIKAN	MELUPAKAN	MELEKAKAN	MENGAMBIL BERAT
4	abdi	slave	NN	HAMBA	-	-	-	-
5	Abdullah	Abdullah	NN	NAMA	-	-	-	-
6	abu	ASH	NN	DEBU	DULI	LEBU	-	-
7	Abu	Abu	NN	NAMA	-	-	-	-
8	ada	exist	VB	MEMILIKI	MEMPEROLEH	MENDAPAT	MEMEGANG	TIADA
9	adab	etiquette	NN	BUDI BAHASA	BUDI PEKERTI	KESOPANAN	KESANTUNAN	-
10	Adam	Adam	NN	NAMA	KAUM	-	-	-

More than 10 rows available. Increase rows selector to view more rows.

Figure 5.9: Example of semantic annotation data stored in Oracle database.

Root word annotation is one of the contributions in this thesis. In linguistic, a root word holds the most basic meaning of any word. A root word has no suffix or prefix, it's the heart of word.The root word annotation is an annotation process using a list of root words derived from the affix process. In natural language processing, it is very important to find the real root of a word for information retrieval and document categorization.

In this thesis, we managed to annotate 149,654 words with their root words, synonyms, and antonyms. Each root word is annotated according to the word's position in each chapters and verses in Malay translated Qur'an. Figure 5.10 shows the Malay translated Qur'an Corpus with root word annotation.



CHAPTER	VERSE	WORD	ROOTWORD
1	1	Yang	yang
1	1	Maha	maha
1	1	Pemurah	murah
1	1	Allah	Allah
1	1	Penyayang	sayang
1	1	Dengan	dengan
1	1	lagi	lagi
1	1	Maha	maha
1	1	nama	nama
1	2	bagi	bagi

More than 10 rows available. Increase rows selector to view more rows.

Figure 5.10: Malay translated Qur'an Corpus with root word annotation

Semantic relationship between synonyms

Relationship of two-word pairs can represent the characteristics of the relationships between multi-word groups. To simplify our discussion, we discuss the relationship of two-word pairs based on the result of semantic annotation data. The semantic relationship will be used for constructing semantic modelling in chapter 6. Words used for creating synonyms will have, or have one of the semantic relationships among them. Each word will be annotated with a semantic element to represent the characteristics of the relationships between words. Here, we describe three types of semantic relationships that will be used for the development of semantic relationships using synonyms [Xian-mo, 2007].

1. Embedment
   Embedment is a collection of words in which the meaning of one word (referred to as W1) is totally embedded in the meaning of the other word (referred to as W2). Figure 5.11 shows the relationship of embedment between W1 and W2.

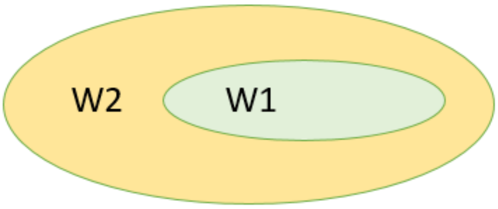


Figure 5.11: Relationship of embedment

For instance, the relationship between ‘Paradise’ and *‘Darussalam’*- this type of relationship is usually known as hyponym by many linguists. ‘Paradise’ is called a superordinate term or an upper term and *‘Darussalam’* is called a hyponym, a subordinate term or a lower term. This type of relationship can be more clearly illustrated by hierarchical tree-diagram. Figure 5.12 shows the relationship between ‘paradise’ and *‘Darussalam’*.

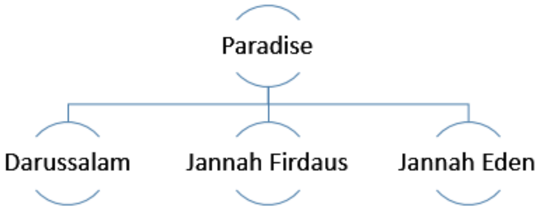


Figure 5.12: Relationship between ‘paradise’ and *‘Darussalam’*

2. Intersection

   Intersection relationship refers to the relationship of the meaning of one word (W1) intersects with the meaning of the other word (W2) to a certain extent. In this relationship, the two words are at the same level. There is no upper term, nor lower term. Figure 5.13 shows the relationship of intersection between W1 and W2.

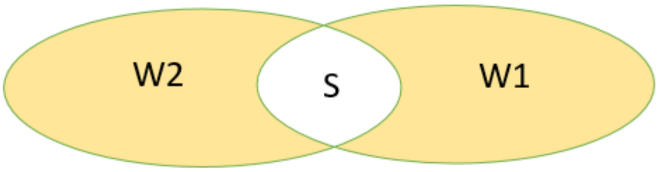


Figure 5.13: Intersection relationship

This type of relationship is most widely found in the Qur'an. For instance, the word 'The Garden' that appeared in the Qur'an can intersect with many other words such as 'paradise', 'Last Home', 'Hereafter', or 'Good news'. In this case, the meaning of all the words can be narrowed to refer only to their intersected part and used as synonymous words. Figure 5.14 shows the example of intersection with the same meaning which appeared in the Qur'an.

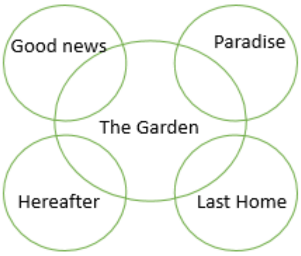


Figure 5.14: Example of intersection with the same meaning appeared in the Qur'an

However, there is another part of intersection that all the words totally have different meaning but they intersect with each other. For instance, the word 'garden' or ’جنة‘ intersected with 'screen' or ’جنة‘. In this case, the meaning of the intersect words are ambiguous and totally have different meanings. Figure 5.15 shows the example of intersection with the different meaning appeared in the Qur'an.

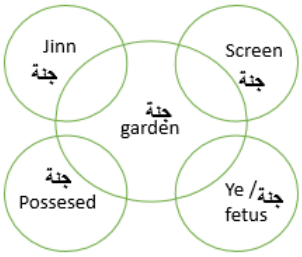


Figure 5.15: Example of intersection with the different meaning appeared in the Qur'an

Most semantic using synonyms are in this relationship. However, to build the relationship of word using synonyms, we need to consider the underlying meaning of each word. This is to avoid mistakes in the semantic annotation process and searching process.

3. Disjoint
   In mathematics, two sets are said to be disjoint sets if they have no element in common. Equivalently, two disjoint sets are a set those intersection is the empty set. The figure 5.16 illustrates the relation between W1 and W2 as the relationship of disjoint.

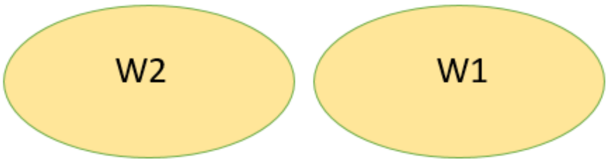


Figure 5.16: Relationship of Disjoint

We conducted a small study among 50 Malaysians to find the words they used in English language to describe the terms we provided in Malay language. One of the words is *'syurga'*. The result is quite fascinating when 80% of them answered 'heaven'. However, in most of Qur'an Qur'an translation, the word ‘heaven’ does not intersect with *‘syurga’*. These two word are disjoint if and only if their intersection ($syurga \cap heaven$) is the empty set. Figure 5.17 shows the relationship between *‘syurga’* and ‘heaven’.

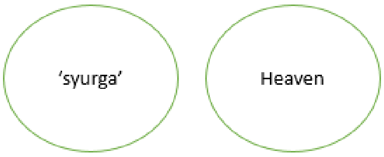


Figure 5.17: The disjoint relationship between *‘syurga’* and ‘heaven’.

To build the semantic relationship concept between synonyms, we used and analyzed three types of the relationship. The semantic relationship is important to expand the meaning of the word to make it more meaningful. In this thesis, we decided to use Malay Thesaurus of Dewan Bahasa dan Pustaka (DBP), Malay Dictionary and WordNet as a resource to generate a semantic relationship concept. This semantic relationship concept will be discussed in Chapter 6.

5.3 Summary

This chapter discusses about the development the first Malay Translated Qur’an corpus using Al-Qur’an Amazing book published by *Karya Bestari* [Nursalim et al., 2016]. The development of Malay translated Qur’an Corpus is an integrated effort involving expert users. The use of Malay translated Qur’an books as main resource of data collection is the best approach rather than using web crawling approach. It reduces the use of mixed languages and dialects. Although there are few studies that have been done, yet there are limited resources and tools for computational linguistic analysis available for the Malay language.

Throughout the preprocessing stage, we created a corpus for Malay language called MyQOS Corpus. Challenges encountered and solutions were also discussed in this chapter. The development of MyQOS Corpus in the thesis will be used for the development of question answer search and concept for ontology development. The use of MyQOS corpus for developing the ontology will be discussed in chapter 6. Besides, MyQOS Corpus will benefit other people, especially Malay researchers who want to do research in natural language processing especially stemming, post-tagging; and improving information retrieval in the Malay translated Qur'an documents.

Chapter 6

Building Malay Ontology for Knowledge Retrieval

This part of the thesis discusses the steps involved in building the ontology for Malay translated Qur'an documents. Detailed descriptions are given on how semantic-based approaches can help improve the performance of traditional keyword-based approaches. We also discuss the potential and limitations of the approach and develop further extensions for the Web environment. Detailed evaluations of the proposed model and its extensions are reported. This research is an inter-disciplinary project benefiting from existing advanced Information Retrieval(IR) techniques coupled with sophisticated Natural Language Processing (NLP) techniques using Semantic-based analysis.
PRESENTED: UK Ontology Network 2018, 30 April 2018, Keele University, UK.

6.1 Malay Ontology

Ontology is one of the emerging technologies in computer science research and semantic web. Gruber defined an ontology as an explicit specification of a conceptualization [Gruber, 1993]. The Ontology represents a domain of knowledge explicitly based on a concept by giving meaning, properties, and relationships to create a knowledge base. Ontology provides a clear and formal way of interpreting data, integration and sharing to help understand the natural language. Ontology is closely associated with Natural Language Processing (NLP), an area of artificial intelligence, computer science, and linguistics.

The use of ontologies to overcome the limitations of keyword-based search has been put forward as one of the motivation of Semantic Web since its emergence in the late 90's. One way to show the semantic aspect of a search engine is to acquire a user query and map it to the formal ontology, expand the query against the Knowledge-Based (KB) ontology, and return a tuple of ontology values that satisfy the query.

The Qur'an is fundamental to all Muslims because it contains comprehensive guidance and knowledge to Muslims in all aspects of life. Nowadays, the Qur'an has been translated into various languages around the world by Muslim experts. The main aim of the availability of the Qur'an translations is to allow the reader to understand the Qur'an in their own native language. As in Malay, there are many versions of Qur'an translation application available online for Malay readers. Based on the observation conducted on existing online Malay translated Qur'an, most of them offered to perform a search using keywords. Retrieving the knowledge from the Qur'an using keywords has several fundamental problems. In many cases, these searching techniques cannot retrieve the relevant knowledge in the Qur'an. For instance, if a user needs information about *'syurga'* or paradise, they will query the word *'syurga'* and get all the verses that contain the word *'syurga'*. However, there are some information that has not been retrieved. If there are information about *'syurga'* but it does not use the word *'syurga'*, it will not be searchable. This is because the query only searches based on the words without other information about the words. Here is an example of extraction of the concept of paradise appeared in Chapter 2, Surah Al-Baqarah, and verse 36.

فَأَزَلَّهُمَا ٱلشَّيۡطَٰنُ عَنۡهَا فَأَخۡرَجَهُمَا مِمَّا كَانَا فِيهِۖ وَقُلۡنَا ٱهۡبِطُواْ بَعۡضُكُمۡ لِبَعۡضٍ
عَدُوّٞۖ وَلَكُمۡ فِي ٱلۡأَرۡضِ مُسۡتَقَرّٞ وَمَتَٰعٌ إِلَىٰ حِينٖ ٣٦

- Malay Translated Qur'an: *"Lalu syaitan memperdayakan mereka sehingga* *mereka dikeluarkan dari (segala kenikmatan) di sana (syurga). Dan Kami berfirman, "Turunlah kamu! Sebahagian kamu menjadi musuh bagi yang lain. Dan bagi kamu ada tempat tinggal dan kesenangan di muka bumi sehingga waktu yang ditentukan.""* (Al-Baqarah, 36)

- English Sahih International: *"But Satan caused them to slip out of it and removed them from that (condition) in which they had been. And We said, "Go down (all of you), as enemies to one another, and you will have upon the earth a place of settlement and provision for a time.""*(Al-Baqarah, 36)

The verse tells about paradise. However, it does not use any paradise word here. Paradise is described implicitly by the word 'condition'. It is very difficult to understand this verse if people do not know the Arabic language and have no knowledge related to the interpretation of the Qur'an. The best way to understand this is to refer it to any hadith and Tafseer book. This indicates that retrieving the knowledge from the Qur'an using keywords will not be able to retrieve the meaning of the word when it uses other words to convey the same meaning.

Besides, one of the key issues in Information Retrieval (IR) is to develop a search engine capable of acquiring knowledge via the ontology. Although there are many search engines that have been developed using ontology. Unfortunately, most of them are made to work with only a small set of widely used languages such as English, Arabic, and Spanish. However, in Malay, it is still lagging behind. There are no tools or search engine and it is a challenge to create ontologies for text written in the Malay language. At present, there is no ontology-based search engine that is being developed for Malay translated Qur'an texts.

Aiming to solve the limitations of keyword-based models in the Malay translated Qur'an, we introduce MyQOS, a new ontology-based IR with semantics using NLP for the Malay translated Qur'an. The creation on MyQOS using ontology-based IR with semantics using Natural language processing (NLP) is new specifically in Malay translated Qur'an. At present, there is no ontology-based search engine that is being developed for Malay translated Qur'an texts. Most of the search engines were developed using keyword-based search. This MyQOS will support semantic retrieval capabilities which can work better than keyword-based search. Therefore, this thesis makes a major contribution to the research on IR in Malay translated Qur'an by demonstrating a new ontology-based Information Retrieval (IR) with semantics using NLP. Furthermore, this thesis also aims to provide a search engine prototype that tailors to improve the quality and accuracy of content retrieval systems. It also facilitates the processing and analysis of unstructured information contained in the text. We also introduce a new ontology-based IR with semantics framework that will guide the development of MyQOS system.

6.2 Methodology

This section will present the methodology of MyQOS semantic search and will illustrate every stage of the design cycle. Figure 6.1 shows the overview of the whole system.

The development of the ontology begins with the definition of the whole concept, the instance or member of the concepts, semantically annotating these concepts with various properties and restrictions in the Malay translated Qur'an. In other words, for each ontology concept, a brief description is stored. A simple description of the ontology is the study of concepts that exist in the real world and the relationships between these concepts. MyQOS semantic search engine is an adaptation of the traditional keyword-based search model. It includes five main processes: extracting and preprocessing, indexing, querying, searching, and ranking. However, as opposed to traditional keyword-based search models, the query in MyQOS search engine is retrieved using an ontology-based query language known as SPARQL and external resources used for indexing and query processing.

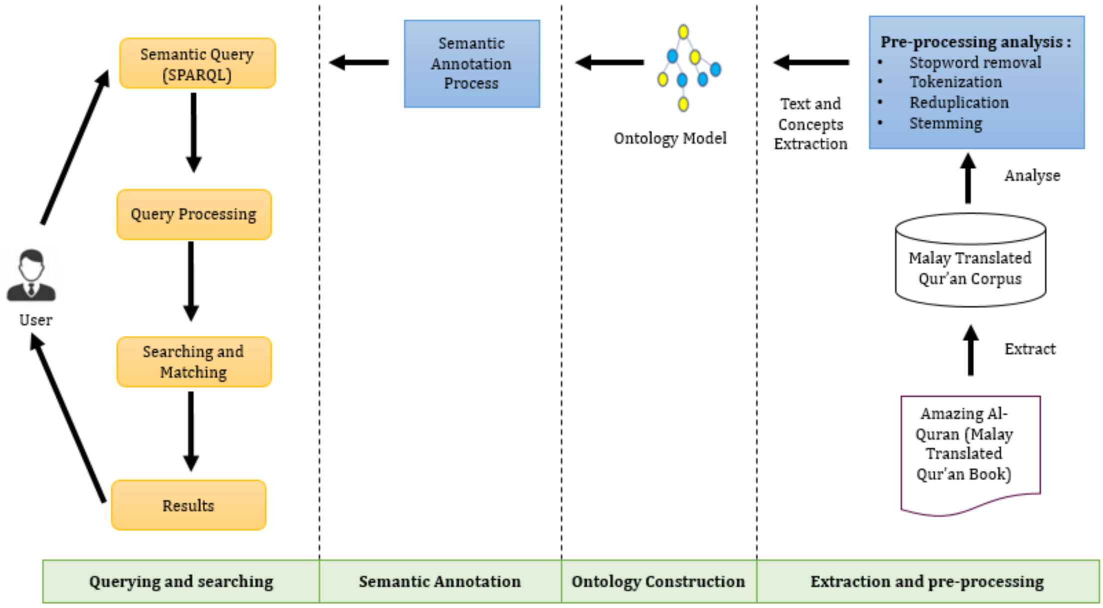


Figure 6.1: MyQOS Ontology-based Search Framework

This framework is divided into four (4) processes. Each processes have their own tasks such as:

1. Extraction and preprocessing
2. Ontology Construction
3. Semantic Annotation
4. Querying and searching

6.2.1 Extraction and preprocessing

Malay language is the tenth most spoken language of the world. Many studies of morphological analysis are based on the English language, but yet there is very little research on the Malay language. The scarcity is mainly due to the incomplete or unavailability of digital resources for the language. Inspired by the scarcity, this research has constructed MyQOS corpus based on the Al-Qur'an Amazing book [Nursalim et al., 2016] as discussed in chapter 5. The construction of MyQOS corpus starts with a preprocessing phase involving stopword removal, tokenization, reduplication, and stemming. Preprocessing is one of the important parts in morphological analysis. In this research, the preprocessing is required to gather root words. The use of root words in this stage is to make semantic annotation using synonyms and antonyms. The 2,817 root words found in the corpus have been annotated. The annotation process that we use for this study are synonyms and antonyms. The semantic annotation of synonyms is manually built using Malay Thesaurus by Dewan Bahasa dan Pustaka (DBP), Malay Dictionary and WordNet. These semantic annotations will be used for the creation of a concept for semantic relationship modelling that will be discussed later in this chapter.

6.2.2 Ontology Construction

For this research, we developed MyQOS semantic search engine by adopting an iterative to ontology development with 6 steps(See figure 6.2):

1. Define the ontology domain and scope
2. Review existing ontologies
3. Enumerate important terms in the ontology
4. Define the classes and the class hierarchy
5. Define the properties of classes and the facets
6. Create instances

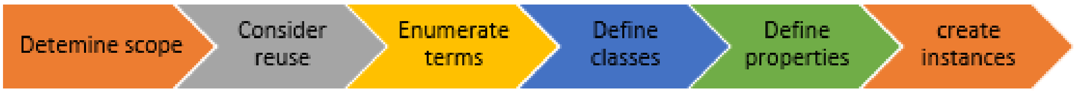


Figure 6.2: Ontology-development process model

Define domain and scope

The concepts are based on the list of topics in the thematic index obtained from the Malay translated Qur'an book. The thematic index has classified verses of the Qur'an according to the Qur'an themes/topics in the Malay language. Besides, the ontology must allow semantic indexing of the Qur'anic contents and the relation between the extracted concepts. The ontology will cover the following subjects: The Qur'an chapters, verses, topics, and each word of the Qur'an and its root words.

Consider the reuse of the existing ontologies

This research adopted the manual knowledge base building approach. Building a knowledge base requires a system administrator or domain expert to build a new, or adopted existing ontology, then semantically annotate the ontology and populate the knowledge base. However, building an ontology from scratch is not a simple task, as it is time-consuming and does not utilise of existing domain-relevant knowledge sources [Kharbat and El-Ghalayini, 2008]. Hence, ontology reuse will be adopted in order to build an ontology for the Malay translated Qur'an. Ontology reuse provides the opportunity of improving the capabilities and knowledge of the existing ontology. Apart from the thematic index in Al-Qur'an Amazing book [Nursalim et al., 2016], the concepts used in this research are referred from three existing Qur'an ontologies.

- The Qurany Concept tools [N H Abbas, 2009] : It is covers nearly 1100 topics and was developed using Google App Engine SDK and the Yahoo! User Interface library. It was available in Arabic and English languages.
- The Qur'an Ontology [Hakkoum and Raghay, 2016b]: It categorizes the topics discussed in the Qur'an verses to a comprehensive index that covering nearly 1100 topics in the Qur'an mostly adapted from The Qurany concepts tools by [N H Abbas, 2009]. It is classified the Qur'an into 15 main themes and sub-themes. It is available in Arabic and English language.
- The Ontology of Quranic Concepts [Dukes, 2013]: It is an annotated linguistic resource, which shows the Arabic grammar, syntax and morphology for

each word in the Qur'an. It contains 300 concepts and was developed using Knowledge Interchange format (KIF). This ontology was translated to OWL and enhanced by designing more relationship and restrictions using sources from the Qur'an, hadith and Islamic websites. This ontology is available in Arabic and English languages.

Since the ontology uses Arabic and English, the process of translation is done manually using dictionary and WordNet.

Enumerate important terms in the ontology

There are two approaches to extract the knowledge from the Qur'an; verse by verse extraction and topic extraction [Hakkoum and Raghay, 2016a]. Verse by verse extraction is difficult to implement because it requires one to cover all the Qur'an otherwise the model will be incomplete. Topic extraction is more reliable in this case which can cover only some topics by only analysing their related verses. This can give more comprehensive results. Therefore, in this research, we used a topic extraction method with the following topics; chapters, verses, words, and synonyms.

To extract these concepts, we used the thematic index in Al-Qur'an Amazing book [Nursalim et al., 2016]. However, because the Qur'an has the knowledge about life, not all knowledge will be taken for this research. The MyQOS ontology focuses mainly on two topics; Location and Living Creation. The selection of those topics is based on preliminary studies that have been conducted and as a result of the preliminary study, we found that there are problems with the queries of these topics (discussed in Chapter 4). In addition, there are existing ontologies created by [Hakkoum and Raghay, 2016a] using the same topics, locations, and Living Creation in Arabic and English. This will make it easier to make comparisons in terms of retrieval results based on Malay language. Aiming to solve this problem, we decided to use these topics as our concepts. Here is the list of terms or concept used in this ontology:

- *'Kedudukan'* (Location)
    - *'Dunia'* (World)

- *’Bandar’* (City)
- *’Gunung’* (Mountain)
- *’Sungai’* (River)
- *’Laut’* (Sea)
- *’Tempat Bersejarah’* (Historical places)

- *’Akhirat’* (Afterlife)
    - *’Syurga’* (Paradise)
    - *’Neraka’* (Hell)

- *’Penciptaan’* (Living Creation)
    - *’Mailaikat’* (Angels)
    - *’Binatang’* (Animals)
        - *’Burung-burung’* (Birds)
        - *’Amfibia’* (Amphibians)
        - *’Mamalia’* (Mammals)
        - *’Reptilia’* (Reptiles)
    - *’Kumpulan Jin’* (Group of Jinn)
    - *’Tumbuhan’* (Plants)
    - *’Manusia’* (Human)
        - *’Golongan manusia’* (Group of people)
        - *’Nabi’* (Prophet)
        - *’Orang Bersejarah’* (Historical People)
- *’Surah’* (Chapter)
- *’Ayat’* (Verse)

From the point of view of querying, all queries for which one concept is relevant are also relevant for the other. Thus for instance “city” is not actually a subconcept of “world” but it is useful to answer queries about the world by including those that mention cities (since they are part of the world). Another concept is “Human”.

Here, the “group of people” is treated as as subconcept of “Human” in the sense that a reference to a group of people is a specialised case of a reference to ‘Human” in general. For example, Al Hawariyun is group of people who followed, supported, and helped Prophet Jesus. So, if we want queries about the concept of “Human”, we would want to consider the set of descriptions that mention a group of people as a subconcept of the set description that involves the general concept of “Human”.

Define the classes and the class hierarchy

There are several approaches in developing a class hierarchy, such as a top-down, bottom-up, and combination approach [Gruninger and Uschold, 1996]. A top-down development process starts with the definition of the most general concepts in the domain and subsequent specialization of the concepts. A bottom-up development process starts with the definition of the most specific classes, the leaves of the hierarchy, with the subsequent grouping of these classes into more general concepts. The combination development process combines the top-down and bottom-up approaches. In this approach, we need to define more significant concepts first and then generalize and specialize them appropriately.

In this research, we used bottom-up process which starts by defining the most specific classes of more general concepts. For example, we start by defining classes for *'syurga'* (paradise) and *'neraka'* (hell) concepts. We then create a common super class from these two classes, *'Akhirat'* (Afterlife), which in turn is a subclass of *'Kedudukan'* (Location).

A class is a concept in the domain. There are several classes created in this ontology model, such as in Table 6.1 below. The main classes of this ontology are: Chapter, Verse, Word, Location, and Living Creation. We establish our ontology based on these main classes and subclasses of these main classes, which are according to sub-topic in the thematic index in Al-Qur’an Amazing book [Nursalim et al., 2016] and other three existing Qur’an ontologies [N H Abbas, 2009, Hakkoum and Raghay, 2016b, Dukes, 2013].

Table 6.1: Description of Class

Class Name	**Description of Class**
Chapter	Represent 144 chapters in the Qur'an
Verse	Represent a verse in each chapters
Concept	Represent a concept discussed in a verse/chapter

We created the ontology model using Protégé and OWL. Protégé is an open source ontology editor and knowledge-based framework. The reason for choosing Protégé is because it is easy to use, well maintained, and has many available and useful plugins.

Define the properties of classes and the facets of the slots

The classes alone will not provide enough information, properties are used to describe resources. It will give more information to a class or subclass. It also provides the relationship between a concept and the data of the concept. Most of the relationships are sub-concept relationships which *is-a* relationships. A sub-concept relationship indicates that a subconcept of another class, as seen in Figure 6.4.

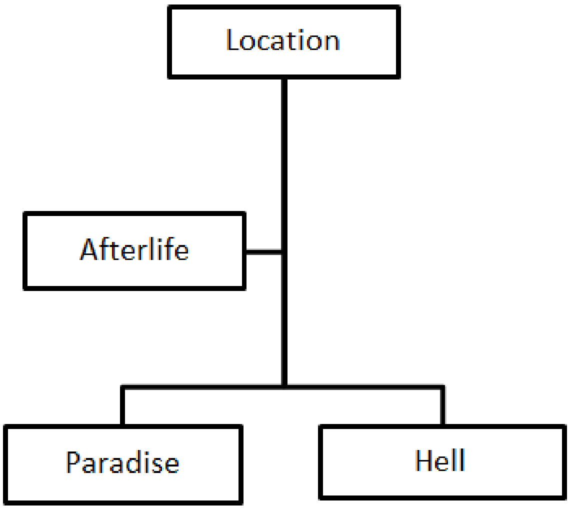


Figure 6.3: Example of a sub-concept taken from MyQOS ontology

Figure 6.3 shows the concept of location, with Afterlife as subconcepts of location and Paradise and Hell as subconcepts of Afterlife. This shows that these concepts are categories of a parent concepts, *Location.* When a user searches for locations that are mentioned in the Qur'an, for example, the system returns the above subconcepts. A user may also search, for example, *Who are the prophets mentioned in the Qur'an?.* Users may also pose more complicated questions than just asking about things that are mentioned in the Qur'an. Users may try to look for solutions to real word problems in a more complex natural language, for example, *Why does Islam allow only eating certain animals?* and this is the type of question we are using for this research.This type of question cannot be answered using the existing knowledge base from Leeds Qur'an Ontology by [Dukes, 2013]. As the result of the limitation in the existing ontologies,this thesis adapted the three ontologies by [N H Abbas, 2009, Hakkoum and Raghay, 2016b, Dukes, 2013].

There are two types of properties; object property and data property. An object property describes the relationship between one concept and another. Data property shows the relationship between concept and its literal meaning. For instance, the value of *MentionedIn* can have multiple values and the values are instances of the class **'Kedudukan'** (Location) and **'Penciptaan'** (Living Creation). We used 27 object properties such as *LocatedIn*, *hasChild*, *hasParent*, *hasBirthMother* and 5 data properties such as *DisplayText*, *VerseCount*, *ChapterIndex* in this ontology. Figure 6.4 shows an example of object properties and data properties used in MyQOS ontology.

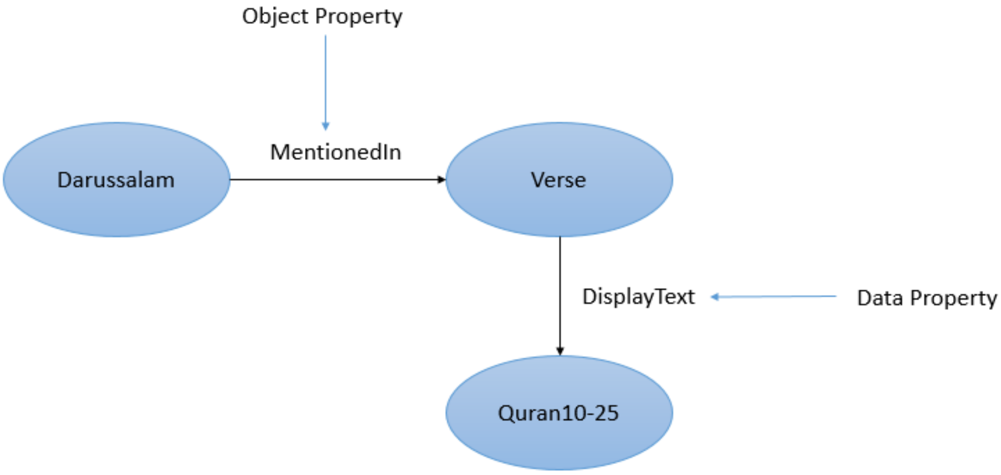


Figure 6.4: Example of Object Property and Data Property

Table 6.2 shows the number of main classes, subclass , object properties, data properties, and instance used in MyQOS ontology. We define the relation between the ontology classes using object property, data property, and cardinality as described in the following entity relationship diagram (ERD). All classes/concepts are related to each other in the system. All classes/concepts are linked to the verses in the Qur'an. One main class/concept can have one or many subclasses/subconcepts. The relation is one-to-many relationship, while many concepts can have many verses. There relation is many-to-many relationship.

In our work, we adapted the idea of designing our ontology by using RDBMS model. The data model of the engine is depicted in Figure 6.5, it contains the following database relations:

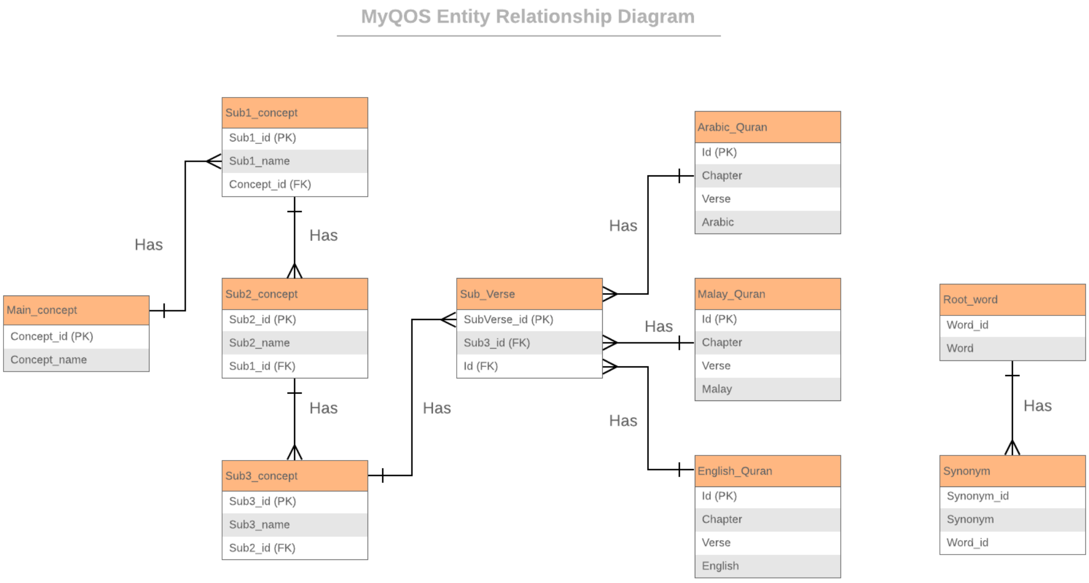


Figure 6.5: Entity Relationship Diagram for MyQOS

- Table 1: Main_concept (Concept_id (PK), Concept_name) – represents all the main concepts or class information.
- Table 2: Sub1_concept (Sub1_id (PK), Sub1_name, Concept_id (FK) – represent 1-tier of sub-concept or sub-class information.
- Table 3: Sub2_concept (Sub2_id (PK), Sub1_name, Sub1_id (FK) – represent 2-tier of sub-concept or sub-class information.
- Table 4: Sub3_concept (Sub3_id (PK), Sub1_name, Sub2_d (FK) – represent 3-tier of sub-concept or sub-class information.
- Table 5: Sub_verse (SubVerse_id (PK), Id (FK), Sub3_id (FK) – act as a bridge to connect the concepts with the verses and chapters in the Qur'an.
- Table 6: Arabi_Quran (Id (PK), Chapter, Verse, Arabic – stores the verses, chapters and Arabic text of the Qur'an.
- Table 7: Malay_Quran (Id (PK), Chapter, Verse, Malay – stores the verses, chapters and Malay text of the Qur'an.

- Table 8: English_Quran (Id (PK), Chapter, Verse, English – stores the verses, chapters, and English text of the Qur'an.
- Table 9: Root_word (Word_id (PK), word – stores the root words derived after the preprocessing process.
- Synonym (Synonym_Id (PK), Synonym Word_id (FK) – stores all the semantics of root words.

In this thesis, we transform the structured data (relational database schema) into a middle model and then create a domain ontology from the model. The use of relational database as the only data source is not practical. Relational database is lack in terms of semantic elements and semantic preserving properties on transforming are proved, but they do not concern the semantic consistency, which is left for the created domain ontology. In this case, more consistency between terms in the ontology will provide more relevant results, and it will increase the precision and recall. The consistency is important to avoid any internal contradiction such as duplication of the term or ambiguous term used in the ontology. This requires OWL specific constraints or rules (Disjunction, Inference) to improve reasoning. Moreover, consistency adequately represents reality.

Table 6.2 shows the ontology metric used in the MyQOS. We have 4 main classes, 23 sub classes, 27 object properties, 5 data properties and 8,198 instances. This will be a knowledge based for the evaluation phase later.

Table 6.2: Ontology Metric

Ontology Metric	**Total**
Main Class	4
Sub Class	23
Object Property	27
Data Property	5
Instances	8,198

Inverse Functional Properties

Figure 6.5 shows the graphical representation of the inverse functional property. It shows that both **Jesus** and **Maryam** are individuals in the Qur'an ontology. The *HasParent* relationship is a data object property, which associates the **Jesus** and **Maryam** individuals. The fact that two individuals are interconnected through a certain property enables inferences to be drawn about the individuals themselves. For example, in Figure 6.5, **Jesus** *HasParent* **Maryam** implies that both individuals are in *Human* concept, and it can be easily inferred that **Maryam** is the parent of **Jesus**. This referred to as an inverse property. Concepts and properties are semantically linked via restrictions, which enable inference.

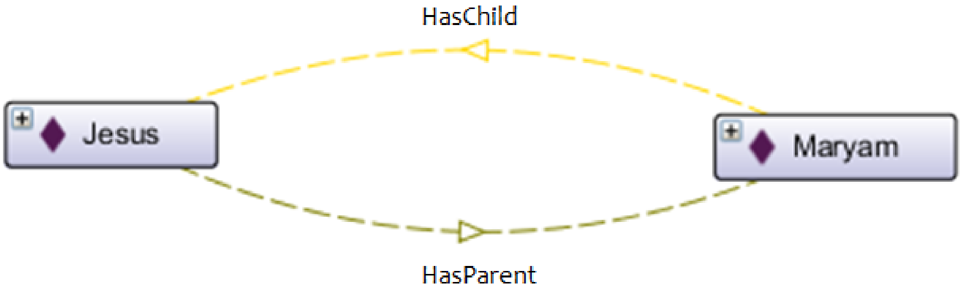


Figure 6.6: Example of Inverse property

Transitive Property

According to [Horridge et al., 2004], in the event that a property is transitive, and the property relates individual **a** to individual **b**, and furthermore individual **a** to the individual **c**, at that point we can surmise that the individual **a** is related with individual **c** by means of property **P**. For example, the *LocatedIn* property between location is transitive. Figure 6.6 shows an example of a transitive property. Individual **Kaabah** is located in **Masjidil al-Haram**, and **Masjidil al-Haram** is located in **Makkah**, then **Kaaba** is also located in **Makkah**

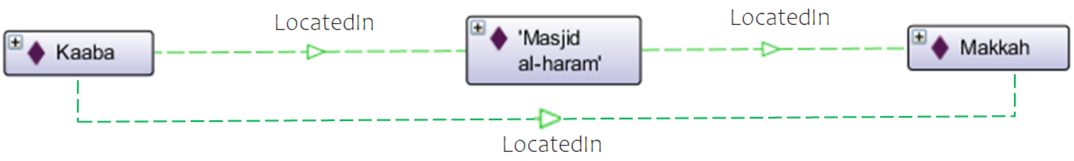


Figure 6.7: Example of transitive property

Symmetric Properties

In the event that a property P is symmetric, and the property relates an individual **a** to individual **b** then individual **b** is related to the individual **a** by means of property **P**. Figure 6.7 shows an example of a symmetric property. Individual **Ishmael** is related to an individual **Isaac** by means of *hasSibling* property, at that point we can induce that **Isaac** should likewise be related with **Ishmael** through *hasSibling* property.

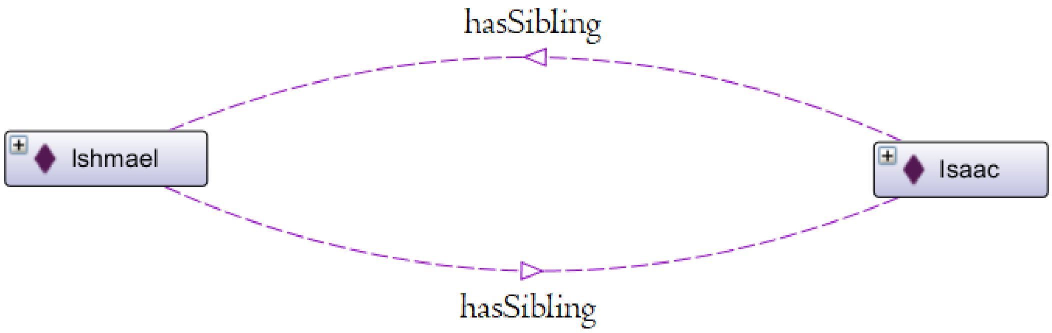


Figure 6.8: Example of symmetric property

Define Facets/Restriction

Describes a constraint on a slot

- Slot cardinality, e.g., one or many values allowed in slot
- Slot value type, e.g. String, Number, Boolean, Enumerated, Instance

Create instances

The last step is creating instances, or individuals of classes in the hierarchy. Instances represent the ground level of the ontology. The combination of an ontology with associated instances is known as a knowledge base. Creating the instances for our ontology was done by extracting data from the thematic index in Al-Qur'an Amazing book [Nursalim et al., 2016] and referring three Qur'an ontologies; the list of concepts from [N H Abbas, 2009], Qur'an ontology by [Hakkoum and Raghay, 2016b] and Qur'anic Ontology by [Dukes, 2013]. Moreover, the Qur'an verses also individuals in this ontology. Each of the concepts in the ontology is assigned a corresponding verse from the Qur'an that discusses such concepts. For example, the concept of Prophet Muhammad is assigned individuals as verse "Quran3-144". "Quran33-40","Quran47-2", "Quran48-29" and "Quran80-1". These individuals are verses in the Qur'an in which Muhammad is mentioned. Table 6.3 presents the number of instances for each main class of the MyQOS ontology.

Table 6.3: Number of Instance according class

Class Name	**Number of Instances**
Chapter	114
Verse	6,236
'Kedudukan' (Location)	86
'Penciptaan' (Living Creation)	1,762
Total	8,198

Figure 6.8 shows the instance for *'syurga'* (paradise) gathered from the four (4) sources mentioned above. From the figure, we can see many names or terms in the Qur'an to represent paradise. Using those names, we created an instance. For example, we created an instance of Paradise, *'Darussalam'* to represent a specific type of paradise. The word *'Darussalam'* is used in Malay translated Qur'an to represent this type of paradise.

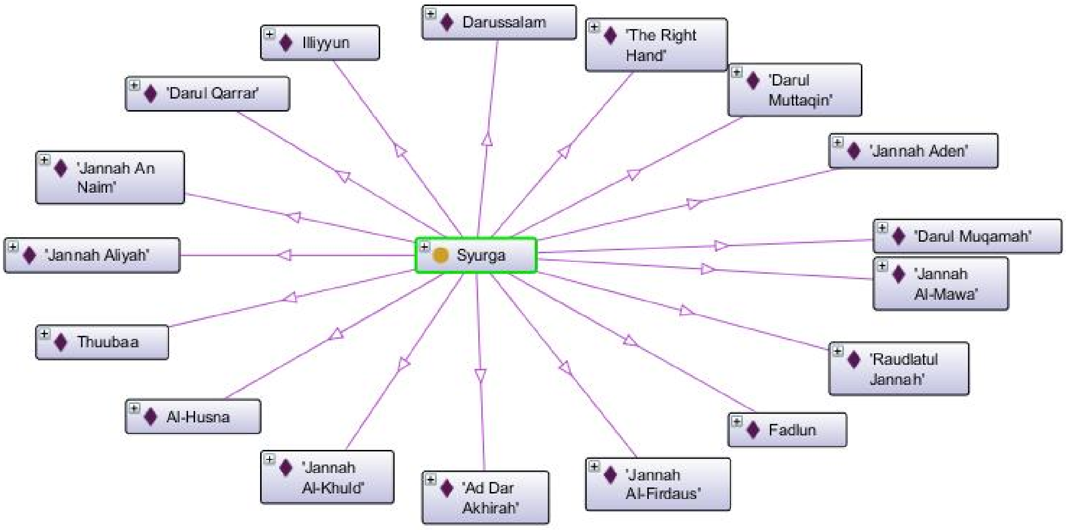


Figure 6.9: Instances for *'syurga'* (paradise)

Ontology Model

MyQOS ontology model was implemented on the Protégé and the ontology schema is stored in an OWL file locally on the computer in RDF/XML syntax. Figure 6.9 shows the MyQOS ontology model.

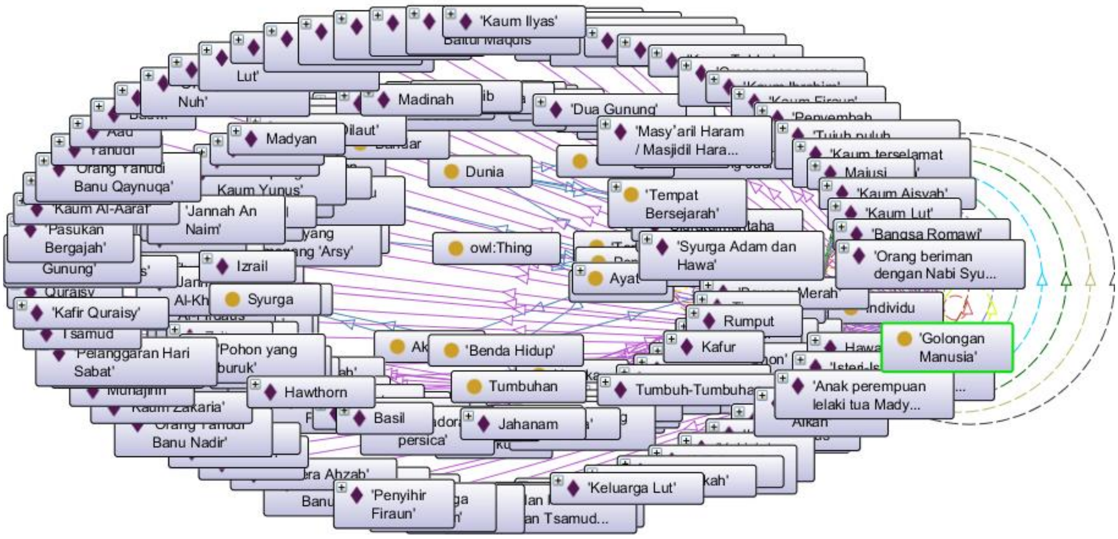
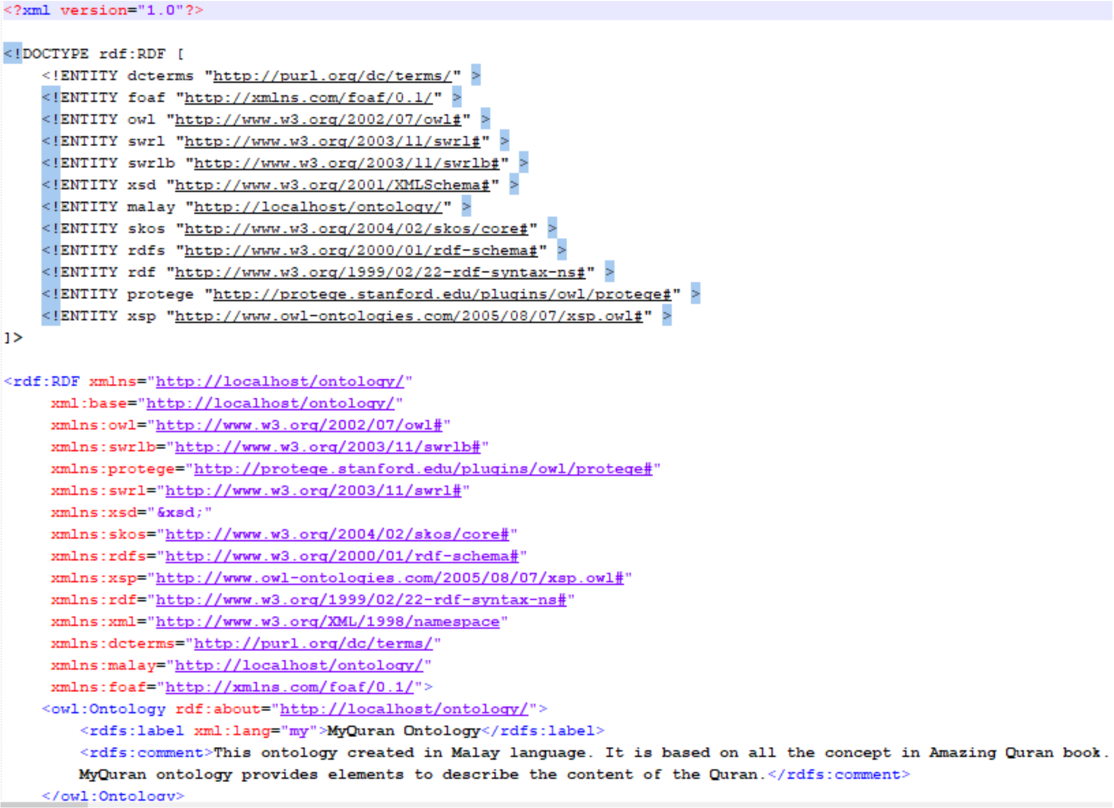


Figure 6.10: The ontology model of MyQOS search engine

The root element of OWL documents for the XML presentation syntax must be an ontology element. The elements and their attributes and element contents in the XML syntax are as in Figure 6.10 below.



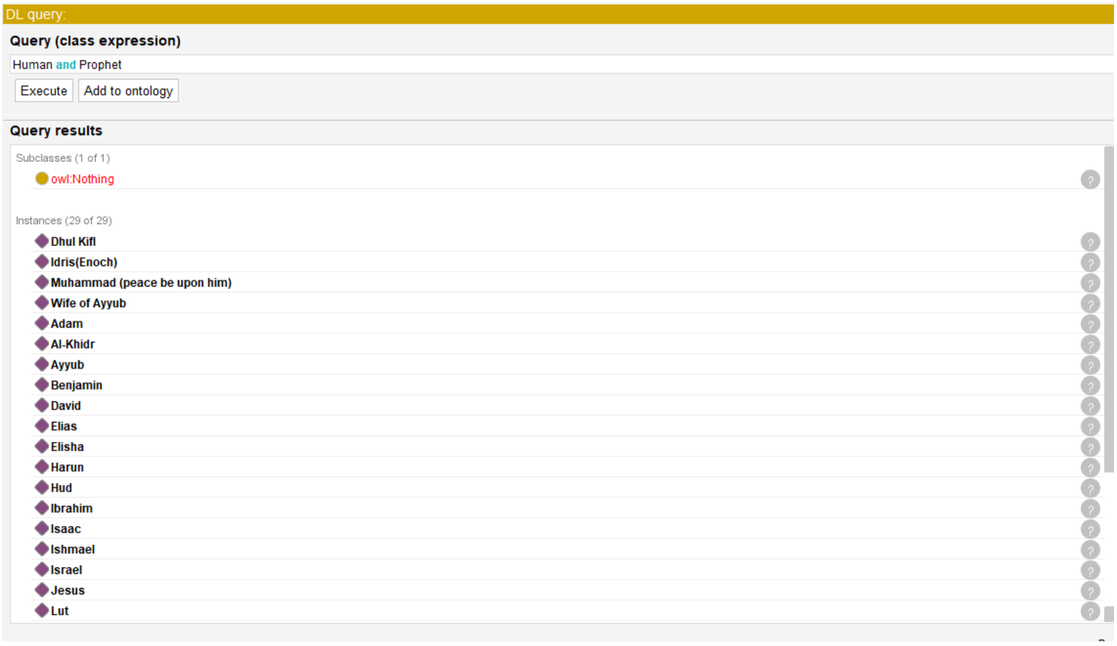
```
<?xml version="1.0"?>

<!DOCTYPE rdf:RDF [
    <!ENTITY dcterms "http://purl.org/dc/terms/" >
    <!ENTITY foaf "http://xmlns.com/foaf/0.1/" >
    <!ENTITY owl "http://www.w3.org/2002/07/owl#" >
    <!ENTITY swrl "http://www.w3.org/2003/11/swrl#" >
    <!ENTITY swrlb "http://www.w3.org/2003/11/swrlb#" >
    <!ENTITY xsd "http://www.w3.org/2001/XMLSchema#" >
    <!ENTITY malay "http://localhost/ontology/" >
    <!ENTITY skos "http://www.w3.org/2004/02/skos/core#" >
    <!ENTITY rdfs "http://www.w3.org/2000/01/rdf-schema#" >
    <!ENTITY rdf "http://www.w3.org/1999/02/22-rdf-syntax-ns#" >
    <!ENTITY protege "http://protege.stanford.edu/plugins/owl/protege#" >
    <!ENTITY xsp "http://www.owl-ontologies.com/2005/08/07/xsp.owl#" >
]>

<rdf:RDF xmlns="http://localhost/ontology/"
     xml:base="http://localhost/ontology/"
     xmlns:owl="http://www.w3.org/2002/07/owl#"
     xmlns:swrlb="http://www.w3.org/2003/11/swrlb#"
     xmlns:protege="http://protege.stanford.edu/plugins/owl/protege#"
     xmlns:swrl="http://www.w3.org/2003/11/swrl#"
     xmlns:xsd="&xsd;"
     xmlns:skos="http://www.w3.org/2004/02/skos/core#"
     xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
     xmlns:xsp="http://www.owl-ontologies.com/2005/08/07/xsp.owl#"
     xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
     xmlns:xml="http://www.w3.org/XML/1998/namespace"
     xmlns:dcterms="http://purl.org/dc/terms/"
     xmlns:malay="http://localhost/ontology/"
     xmlns:foaf="http://xmlns.com/foaf/0.1/">
    <owl:Ontology rdf:about="http://localhost/ontology/">
        <rdfs:label xml:lang="my">MyQuran Ontology</rdfs:label>
        <rdfs:comment>This ontology created in Malay language. It is based on all the concept in Amazing Quran book.
        MyQuran ontology provides elements to describe the content of the Quran.</rdfs:comment>
    </owl:Ontology>
```

Figure 6.11: The RDF/XML presentation of the ontology

Inference and Rules

The formal specification of Web Ontology Language, OWL, is highly influenced by Description Logics (DLs). OWL-DL is designed to be a computationally complete and decidable version of OWL, thus it benefits from a wide range of complete and terminating DL reasoners. For our inference module, we use HermiT 1.4.3, an open-source DL-reasoner, which supports the standard inference services such as consistency checking, concept stability, classification, and realization.

Automatic inference on ontologies expressed in OWL is performed by the inference engine. An inference engine takes a bunch of fact on a specific domain of interest asserted in OWL and determines different facts using axioms and inference rules. All in all, an inference engine makes unequivocal the facts that are just understood from the expressly represented facts. The determined facts are the consequences of the facts in a given fact base and metaphysics that is utilized to communicate the facts. Introductions to any variables that were utilized in the derivation are additionally given. For example, if the inference system knows that Human is a living creation and Prophet is Human, at that point, the system can infer that Prophet is sub-class of living creation. Different inferences can determine the properties of sub-classes dependent on the properties of their super-classes, inferring that an individual is an instance of a specific class, and more.



DL query:

Query (class expression)

Human and Prophet

Query results

Subclasses (1 of 1)

- owl:Nothing

Instances (29 of 29)

- Dhul Kifl
- Idris(Enoch)
- Muhammad (peace be upon him)
- Wife of Ayyub
- Adam
- Al-Khidr
- Ayyub
- Benjamin
- David
- Elias
- Elisha
- Harun
- Hud
- Ibrahim
- Isaac
- Ishmael
- Israel
- Jesus
- Lut

Figure 6.12: The example of inference of Human and Prophet classes

6.2.3 Semantic Annotation

Semantic annotation is the process of attaching additional information to concepts in a given text or any other content. Sometimes, semantic annotations are also known as semantic tagging. When a document is semantically tagged, it becomes a source of information that is easy to interpret, combine, and reuse by computers. Usually, the annotations are used by the retrieval and ranking module. In semantic annotation techniques, a document is analysed to identify its relevant terms and to define the importance of each term.

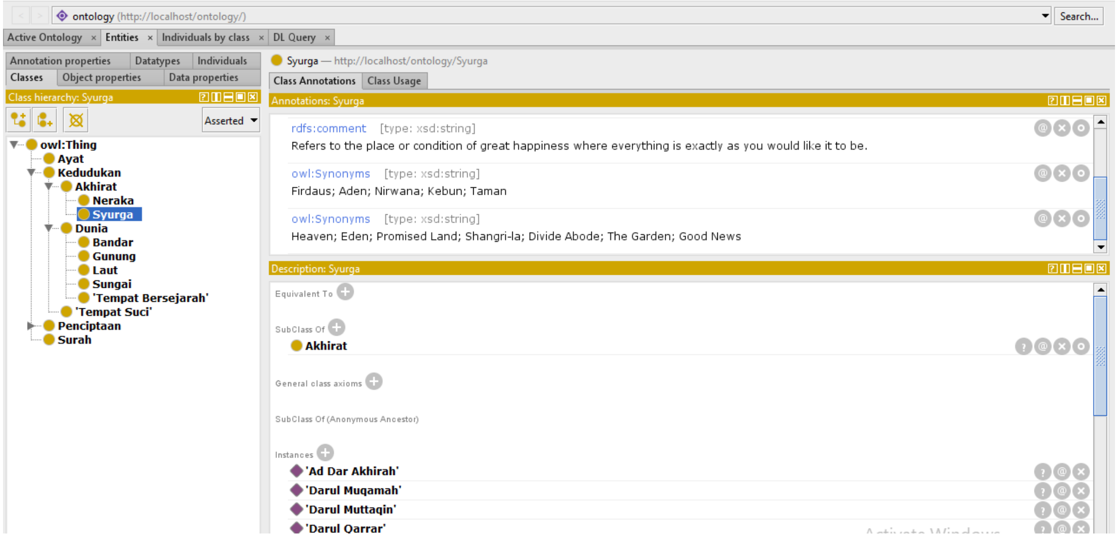
The semantic annotation of a document d consists in linking the terms t in $d = t\_1, t\_2, t\_3 \cdot \cdot\cdot, t\{\_n$ with the entities in the ontology. Namely, let an entity-term pair be (c,t), where c is an entity in the ontology and t is a term of d, so that there is a mapping between the textual descriptions defined in the label $rdfs : label$ of c and t.

Our thesis work proposes to analyse the context of the annotations in order to identify their meaning through the entities such as classes and instances in the ontology. In the extraction of the context, the explicit relationships of each class and instances in the ontology are analysed. Here, we create the annotation by adding labels to the class and instances. In this ontology, we created two annotation properties to represent semantic annotation using $owl : SameAs$ and $owl : Synonyms$. $owl : SameAs$ used for mapping the same concepts from two or more datasets, where each of these concepts can have different features and relations to other concepts whereby $owl : Synonyms$ for mapping the same individuals but in different names.

- <!– http://www.w3.org/2002/07/owl#SameAs –>
  <owl:AnnotationProperty
  rdf:about=”http://www.w3.org/2002/07/owl#SameAs”/>

- <!– http://www.w3.org/2002/07/owl#Synonyms –>
  <owl:AnnotationProperty
  rdf:about=”http://www.w3.org/2002/07/owl#Synonyms”/>

The semantic annotation using a synonym derived from MyQOS Corpus as discussed in Chapter 5. Figure 6.12 shows the screenshot of semantic annotations created in MyQOS.



rdfs:comment [type: xsd:string]
Refers to the place or condition of great happiness where everything is exactly as you would like it to be.

owl:Synonyms [type: xsd:string]
Firdaus; Aden; Nirwana; Kebun; Taman

owl:Synonyms [type: xsd:string]
Heaven; Eden; Promised Land; Shangri-la; Divide Abode; The Garden; Good News

Figure 6.13: The screenshot of semantic annotation in MyQOS

Figure 6.13 shows a fragment of the *'Akhirat'* class containing URI,class, $rdfs : label$, $owl : SameAs$ and $rdfs : Synonyms$.

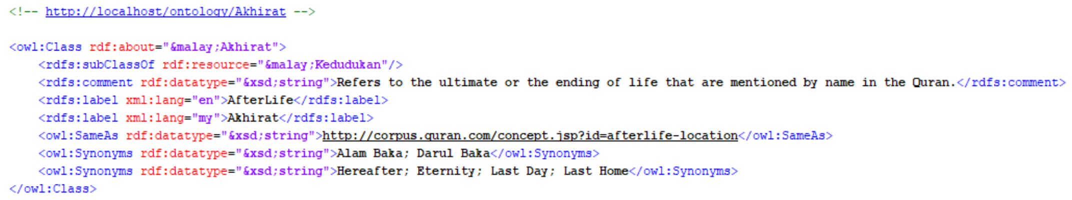

```
<!-- http://localhost/ontology/Akhirat -->

<owl:Class rdf:about="&malay;Akhirat">
    <rdfs:subClassOf rdf:resource="&malay;Kedudukan"/>
    <rdfs:comment rdf:datatype="&xsd;string">Refers to the ultimate or the ending of life that are mentioned by name in the Quran.</rdfs:comment>
    <rdfs:label xml:lang="en">AfterLife</rdfs:label>
    <rdfs:label xml:lang="my">Akhirat</rdfs:label>
    <owl:SameAs rdf:datatype="&xsd;string">http://corpus.quran.com/concept.jsp?id=afterlife-location</owl:SameAs>
    <owl:Synonyms rdf:datatype="&xsd;string">Alam Baka; Darul Baka</owl:Synonyms>
    <owl:Synonyms rdf:datatype="&xsd;string">Hereafter; Eternity; Last Day; Last Home</owl:Synonyms>
</owl:Class>
```

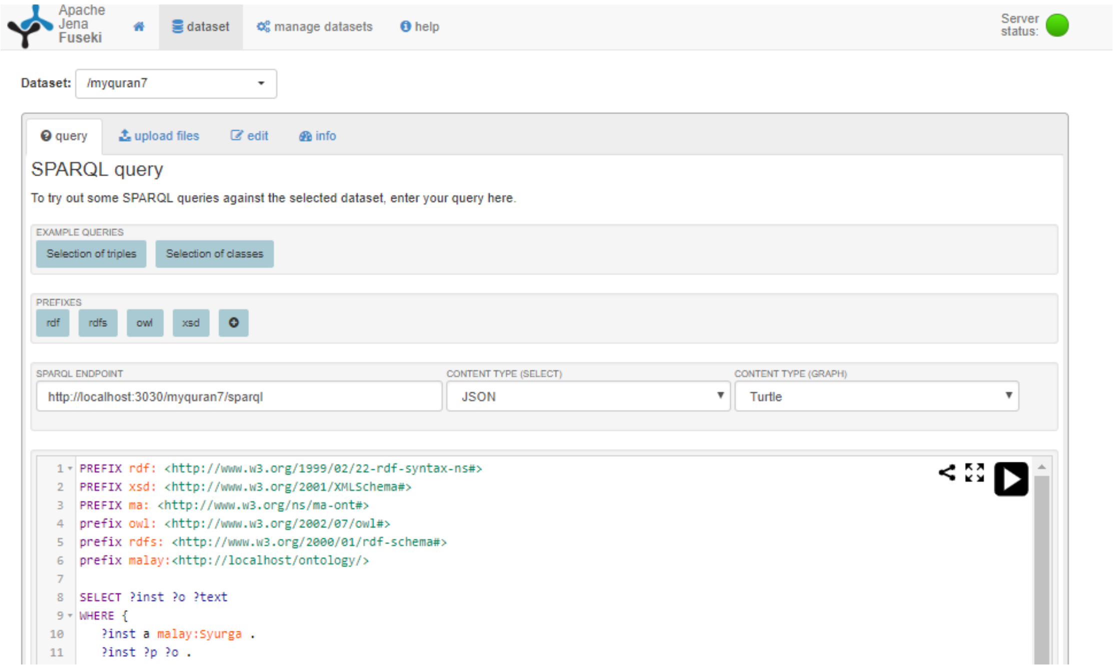
Figure 6.14: The XML/OWL syntax of semantic annotation

6.2.4 Query and Searching

Our system takes as input as a formal SPARQL query. SPARQL is the standard query language of the Semantic Web and can be used to query RDF databases. The main backbone of SPARQL queries is the triple pattern;the subject, predicate and/or object, consist of variables which are parsed for the construction of SPARQL queries. The main idea is to match the triples in the SPARQL queries to RDF triples to retrieve relevant information from the knowledge base. A basic

SPARQL query is composed of a “SELECT” clause that contains the variables we want to return, and a “WHERE” clause that contains the conditions that the variables must meet; these conditions are in the form of a triple. We define in the WHERE clause a graph a pattern where some nodes are known and others are not. When we run the query, it attempts to match the graph pattern to the model and extract the possible values for the unknown nodes.

There are built-in SPARQL queries in the Protégé. However, Protégé cannot load a big file of RDF triples. Thus we have to store the ontology using Apache Jena with Fuseki server. Apache Jena with Fuseki server provides an ontology API that enables to work with ontologies of different formats, like OWL or RDFS. Other than that, it is an open source. After loading the OWL file into Jena Fuseki, we can issue SPARQL queries on the OWL file. Figure 6.14 shows the screenshot of Apache Jena with Fuseki server with MyQOS OWL file.



Dataset: /myquran7

SPARQL query

To try out some SPARQL queries against the selected dataset, enter your query here.

SPARQL ENDPOINT: http://localhost:3030/myquran7/sparql

CONTENT TYPE (SELECT): JSON

CONTENT TYPE (GRAPH): Turtle

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX ma: <http://www.w3.org/ns/ma-ont#>
prefix owl: <http://www.w3.org/2002/07/owl#>
prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
prefix malay:<http://localhost/ontology/>

SELECT ?inst ?o ?text
WHERE {
   ?inst a malay:Syurga .
   ?inst ?p ?o .
```

Figure 6.15: The screenshot of Apache Jena with Fuseki server with MyQOS OWL file

After normalizing the predicate, the variables of the triple are parsed for

SPARQL generation. The system uses the location of the question mark symbol in the parsing variable for query generation. For instance, *?hasBirthMother*, *Maryam*, a SPARQL query is generated and parsed by Jena inference engine, which automatically reasons and infers an answer by looking for the missing variables, which are the subjects and return any subject in the model matches the predicate *hasBirthMonther* and the object Maryam. The inference engine automatically infers that answer and *Jesus* or *Isa* are returned.

The query searches for matches in the ontology. Semantic searching seeks to improve the search accuracy of the search engine by understanding the user needs and the contextual meaning of the query term to retrieve a more relevant result. Here, the SPARQL query will search not only the query term but also the meaning of the term using additional information provided in the ontology. Listing 6.1 is an example of SPARQL queries that query the semantic annotation linked to the ontology.

Listing 6.1: Query the semantic annotation using SPARQL query

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX ma: <http://www.w3.org/ns/ma-ont#>
prefix owl: <http://www.w3.org/2002/07/owl#>
prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
prefix malay:<http://localhost/ontology/>

SELECT ?inst ?o ?text
WHERE {
    ?inst a malay:Syurga .
    ?inst ?p ?o .
    ?o malay:DisplayText ?text .
}
```

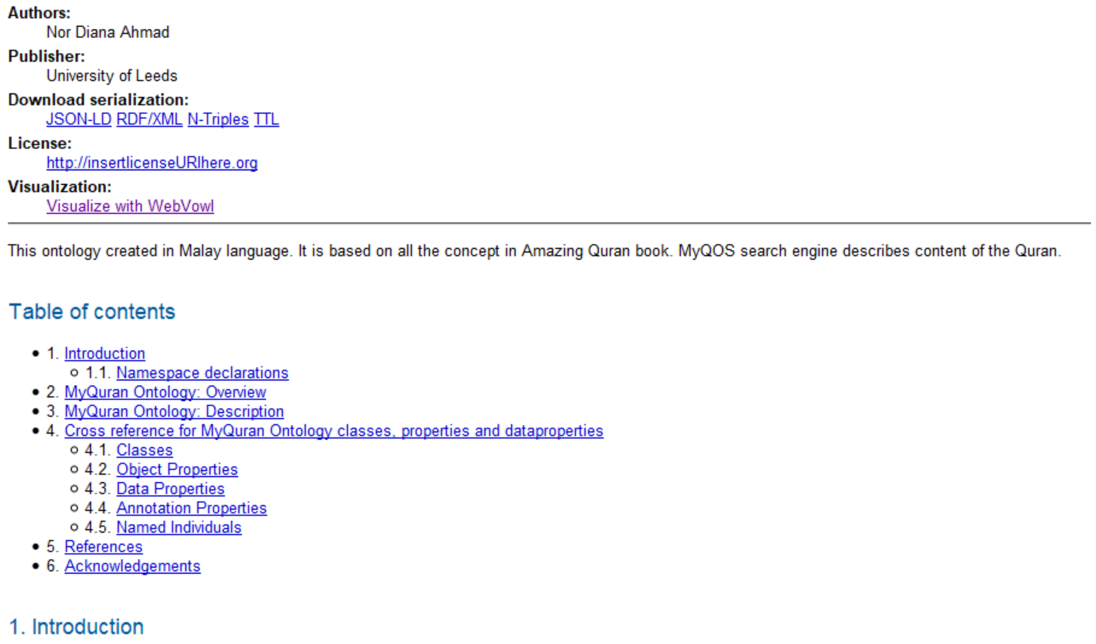
In this query, we are selecting instances and verses related to syurga. The query specifies the relationship of ?inst and malay:Syurga class. Here, it will display all the instances of class Syurga that have the same contextual meaning.

6.3 Implementation and validation

The ontology model needs to publish and evaluate to improve the quality of the ontology. Here, we describe these two activities in the context of MyQOS ontology.

6.3.1 Ontology Publication

In this thesis, we published the ontology by producing an HTML ontology documentation using Widoco[1] tools. Widoco expects an ontology in OWL/RDFS as input and produces an HTML documents by using the ontological definitions, including their relationships, axioms, labels, and descriptions. The resulting HTML document was revised and extended by the ontology development team. Figure 6.15 presents the screenshot of the HTML MyQOS documentation that is available in MyQOS semantic search engine system.



Authors:
Nor Diana Ahmad
Publisher:
University of Leeds
Download serialization:
JSON-LD RDF/XML N-Triples TTL
License:
http://insertlicenseURIhere.org
Visualization:
Visualize with WebVowl

This ontology created in Malay language. It is based on all the concept in Amazing Quran book. MyQOS search engine describes content of the Quran.

Table of contents

- 1. Introduction
  - 1.1. Namespace declarations
- 2. MyQuran Ontology: Overview
- 3. MyQuran Ontology: Description
- 4. Cross reference for MyQuran Ontology classes, properties and dataproperties
  - 4.1. Classes
  - 4.2. Object Properties
  - 4.3. Data Properties
  - 4.4. Annotation Properties
  - 4.5. Named Individuals
- 5. References
- 6. Acknowledgements

1. Introduction

Figure 6.16: The HTML format of MyQOS documentation

[1] https://github.com/dgarijo/WIDOCO/releases/tag/v1.4.7

6.3.2 Ontology Evaluation

Ontology evaluation is one of the important parts in ontology development. This evaluation will indicate the quality of ontology produced. Here, the ontology evaluation approaches can be classified into five categories:

1. Assessment by human. In this approach, the quality of ontology will be assessed by human expert in a particular domain. The human expert will define the criteria and requirement to be met by the ontology [Pinto et al., 1999]

2. Application-based. The quality of the ontology is measured with respect to its suitability for a specific application or task.

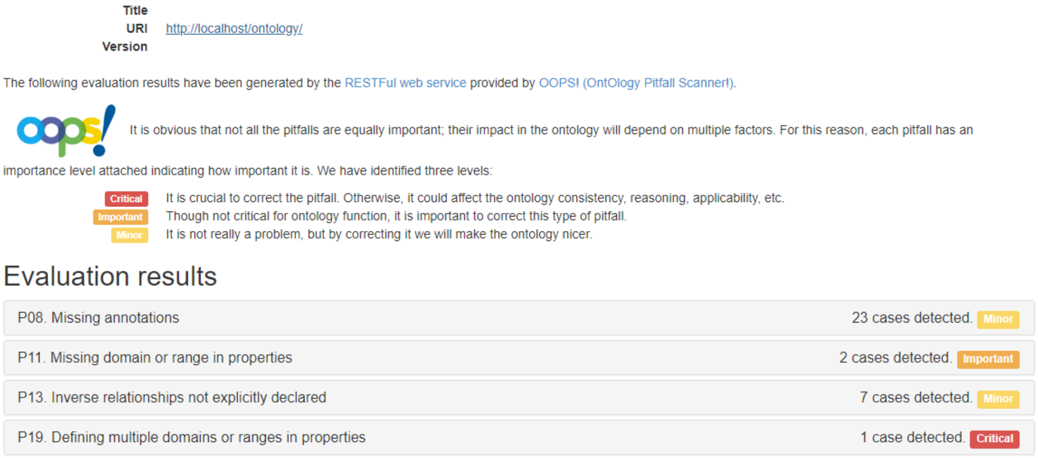
3. Topology-based. The quality of the ontology is assessed by computing a set of measures based on the internal structure of the ontology.

4. Data-driven. The ontology is compared against an unstructured resource representing the problem domain.

5. Gold standard. The ontology is compared against a structured resource representing the problem domain.

For the validation of the MyQOS ontology schema, we used topology-based evaluation using OOPS! (Ontology Pitfall Scanner), a web tool for identifying pitfalls in ontologies to assist developers throughout ontology validation. OOPS! covers the list of pitfalls detected by most up-to-date and accessible approaches and permits choosing set of pitfalls to be analysed in the line with completely different evaluation dimensions. Besides, the system additionally provides an indicator such as critical, important, and minor for every pitfall in step with their potential negatives consequences [Poveda-Villalón et al., 2014].

The iteration process consists of two steps; diagnose and repair. The process begins by diagnosing using an online version of OOPS! to detect pitfalls in the ontology. After that, the repair process will be done to fix any pitfalls identified by OOPS!. The process will continue until the pitfalls are fixed. Figure 6.16 shows a screenshot of the results provided by OOPS! before starting the first diagnose repair iteration.



Title
URI http://localhost/ontology/
Version

The following evaluation results have been generated by the RESTFul web service provided by OOPS! (OntOlogy Pitfall Scanner!).

It is obvious that not all the pitfalls are equally important; their impact in the ontology will depend on multiple factors. For this reason, each pitfall has an importance level attached indicating how important it is. We have identified three levels:

Critical	It is crucial to correct the pitfall. Otherwise, it could affect the ontology consistency, reasoning, applicability, etc.
Important	Though not critical for ontology function, it is important to correct this type of pitfall.
Minor	It is not really a problem, but by correcting it we will make the ontology nicer.

Evaluation results

P08. Missing annotations	23 cases detected. Minor
P11. Missing domain or range in properties	2 cases detected. Important
P13. Inverse relationships not explicitly declared	7 cases detected. Minor
P19. Defining multiple domains or ranges in properties	1 case detected. Critical

Figure 6.17: Summary provided by OOPS! tool before the first iteration for MyQOS ontology

First Iteration

As shown in Figure 6.16, the tools detected four pitfalls which 2 minor, 1 critical, and 1 important. The first iteration focused on repairing all the pitfalls. Below, we describe the pitfalls and the action taken.

- P08-Missing annotations (Minor). The pitfall detected the lack of rdfs:comment annotation to describe the ontology classes and properties. To overcome the missing annotation, we created rdfs:comment for each class and its properties. These are the lists of missing rdfs:comment.

  - http://localhost/ontology/FamilyRelation
  - http://localhost/ontology/Relation
  - http://localhost/ontology/RevealedAfter
  - http://localhost/ontology/AncestorOf
  - http://localhost/ontology/HasGrandParent
  - http://localhost/ontology/HasChild
  - http://localhost/ontology/HasAncestor

- http://localhost/ontology/IsChild
- http://localhost/ontology/FriendOf
- http://localhost/ontology/HasSibling
- http://localhost/ontology/EnemyWith
- http://localhost/ontology/GrandParentOf
- http://localhost/ontology/HasParent
- http://localhost/ontology/HasMember
- http://localhost/ontology/HasSpouse
- http://localhost/ontology/MentionedIn
- http://localhost/ontology/AlsoKnowsAs
- http://localhost/ontology/ContemporaryRelation
- http://localhost/ontology/IsParent
- http://localhost/ontology/MemberOf
- http://localhost/ontology/LocatedIn
- http://localhost/ontology/RevealedBefore
- http://localhost/ontology/AllyTo

- P13- Inverse relationships are not explicitly declared (Minor). This pitfall identified seven owl:inverseOf statement for object properties which has no inverse relationship. Here, we checked again the inverse relationship among the object properties. As we detected, the inverse relationship that we created early is ambiguous. The connection between the two object properties is unclear. Therefore, we correct the relationship and check it again. These are the lists of object properties that need to be corrected.

  - http://localhost/ontology/MemberOf
  - http://localhost/ontology/RevelationPlace
  - http://localhost/ontology/Relation
  - http://localhost/ontology/AlsoKnowsAs

- http://localhost/ontology/MentionedIn
- http://localhost/ontology/LocatedIn
- http://localhost/ontology/HasMember

- P11- Missing domain or range in properties (Important).This pitfall is indicated by a lack of large number of domain and ranges declaration for properties. Here, the majority missing domains indicated by OOPS! corresponds to missing ranges for datatype properties, which are not fully addressed in the ontology. After analysing the missing domains and ranges, we checked again the ontology. The pitfalls affect on the following ontology elements.

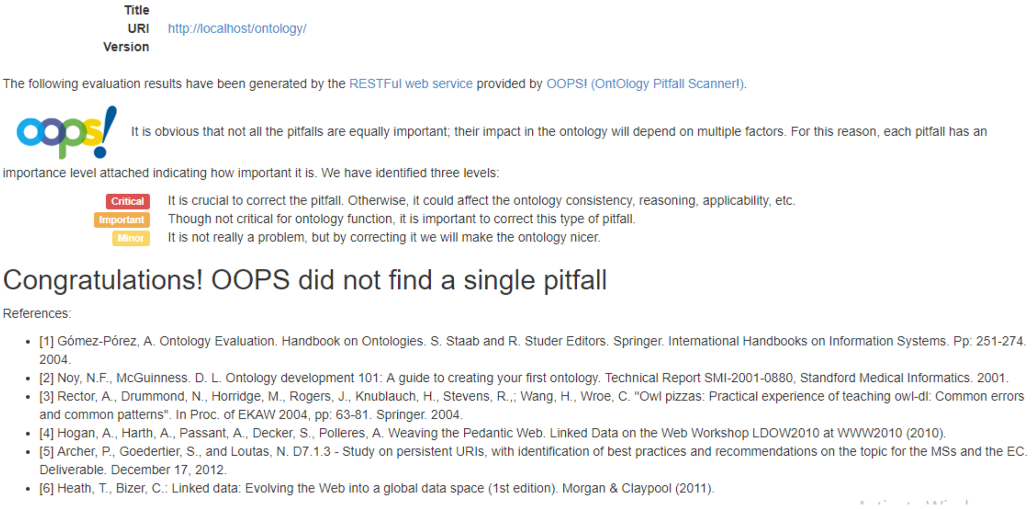
  - http://localhost/ontology/LocatedIn
  - http://localhost/ontology/AlsoKnowsAs

- P19- Defining multiple domains or ranges of properties (Critical). The pitfall indicates there are more than one rdfs:domain or rdfs:range statement for properties. In OWL, multiple rdfs:domain and rdfs:range axioms are allowed, but they are interpreted as conjunctions. It is equivalent to the construct of owl:intersectonOf. After analysing one affected ontology element as described below, we noticed the error in putting the class categories in rdfs:Domain for object property MentionedIn.

  - http://localhost/ontology/MentionedIn

Second Iteration

This iteration was carried out after the publication of the first version of the MyQOS ontology. In this iteration, there are no pitfalls that were found nor any critical issues were found in MyQOS ontology. More importantly, this indicates the sufficient quality of the ontology. This also shows using systematic approaches based on catalogue data and domain expert input provided by the OOPS! tool can produce an ontology with sufficient quality. Figure 6.17 shows the screenshot provided by OOPS! tools after the first iteration is done.



Title
URI http://localhost/ontology/
Version

The following evaluation results have been generated by the RESTFul web service provided by OOPS! (OntOlogy Pitfall Scanner!).

It is obvious that not all the pitfalls are equally important; their impact in the ontology will depend on multiple factors. For this reason, each pitfall has an importance level attached indicating how important it is. We have identified three levels:

Critical	It is crucial to correct the pitfall. Otherwise, it could affect the ontology consistency, reasoning, applicability, etc.
Important	Though not critical for ontology function, it is important to correct this type of pitfall.
Minor	It is not really a problem, but by correcting it we will make the ontology nicer.

Congratulations! OOPS did not find a single pitfall

References:

- [1] Gómez-Pórez, A. Ontology Evaluation. Handbook on Ontologies. S. Staab and R. Studer Editors. Springer. International Handbooks on Information Systems. Pp: 251-274. 2004.
- [2] Noy, N.F., McGuinness. D. L. Ontology development 101: A guide to creating your first ontology. Technical Report SMI-2001-0880, Standford Medical Informatics. 2001.
- [3] Rector, A., Drummond, N., Horridge, M., Rogers, J., Knublauch, H., Stevens, R.,; Wang, H., Wroe, C. "Owl pizzas: Practical experience of teaching owl-dl: Common errors and common patterns". In Proc. of EKAW 2004, pp: 63-81. Springer. 2004.
- [4] Hogan, A., Harth, A., Passant, A., Decker, S., Polleres, A. Weaving the Pedantic Web. Linked Data on the Web Workshop LDOW2010 at WWW2010 (2010).
- [5] Archer, P., Goedertier, S., and Loutas, N. D7.1.3 - Study on persistent URIs, with identification of best practices and recommendations on the topic for the MSs and the EC. Deliverable. December 17, 2012.
- [6] Heath, T., Bizer, C.: Linked data: Evolving the Web into a global data space (1st edition). Morgan & Claypool (2011).

Figure 6.18: Summary provided by OOPS! tool after the first iteration for MyQOS ontology

6.4 Summary

This chapter explains The creation of this first Malay Translated Qur'an ontology known as MyQOS ontology. The MyQOS ontology has collected and combined concepts in the Qur'an with the verses. Concepts entities and relationships of the ontology were formed by analyzing knowledge based on the Qu'ran. An ontology for a specific domain is not a goal in itself. Developing an ontology has per objective to define a set of data which specific programs may use. The propose of this research was to develop the Malay Qur'an ontology to be used as a background knowledge for Quranic research knowledge management systems. This solved the problem with conventional search engines. As conventional search engines cannot interpret the sense of the user's search, not all verses that discuss the concept can be retrieved, the ambiguity of the query leads to the retrieval of irrelevant information.

The quality of the ontology only can be assessed by using it in the applications for which it is designed for. Since an ontology is a basic infrastructure and description of a specific knowledge in a specific domain it is therefore very difficult to construct an ontology without any criteria and guidelines. Criteria and guidelines can facilitate the construction of MyQOS ontology to disambiguate their queries. The MyQOS ontology was evaluated by using the MyQOS search engines in a query enrichment process. The evaluation is carried out by judging how the ontology can help to answer the 30 questions. The following section will analyze and discuss the evaluation process of MyQOS ontology.

The MyQOS will benefit other people who want to do research related to the Malay translated Qur'an. This ontology will facilitate the semantic search. The MyQOS ontology network is the backbone of MyQOS search engines, data sets and services. This thesis makes a major contribution to the research on IR in Malay translated Qur'an by demonstrating a new ontology-based Information Retrieval (IR) with semantics using NLP.

Chapter 7

Retrieval Evaluation

In this chapter, an evaluation of the MyQOS prototype is presented. The evaluation process focused on the accuracy and quality of the data to support the retrieval performance. The evaluation section plays a crucial role to make progress in building a better search engine. The scope of this evaluation is concerned with the aspects of the implemented prototype. These aspects concern about the extent of relevancy between retrieved results and user queries in the search domain. This evaluation significantly depends on the ontology defining the relevant terms for expanding the user query. In this sense, the performance of the semantic search depends on the MyQOS ontology built in chapter 6.

7.1 Query Collection

The user can formulate various forms of a query expression. A simple or keyword query type is a query in the form of a single word. [Lalmas et al., 2002] states that 25% of users interact with search engines by providing single-word inputs. Another type of query is a Boolean form as used in most web search engines.

The question about how many number of queries should be used in the information retrieval experiment has been answered by [Buckley and Voorhees, 2000]. The research they conducted shows that the minimum number of queries that need to be implemented in the information retrieval experiment is 25 and above. The more the number of questions, the better results will be obtained. Meanwhile, most studies in information retrieval using query formulated in the natural language according to sentence structure and language grammar. According to [Popovič, 1991], queries must meet the needs of users and define correctly. While [Sakai et al., 1999] stated, when developing test collections for Japanese, they have categorised queries with six degrees of difficulty from simple word matching to knowledge processing in the context of universal knowledge.

The methods used by [Sakai et al., 1999] and [Buckley and Voorhees, 2000] have been taken into account in the construction of query collections. This study used the Malay query words taken from Pouzi's collection as natural language queries and the English query words are translated from these collections (see Appendix E for query list). From the 40 queries, ten queries have been removed because the queries are not clear in term of the level of difficulty. A simple query example of *"Cerita tentang Qarun"* and a tough query, *"Maklumat tentang kejadian Manusia iaitu Nabi Adam"*. *"Tegahan memakan babi"* though it is easy to list, only documents that have the word *"babi"*, actually have a high degree of difficulty. This is because the word *"Tegahan"* in the query has many other meanings such as *"Jangan"*, *"tidak"*, *"tidak Boleh"*, *"berdosa"* and so on. Some of these meanings are not necessarily based on synonyms but the implied meaning that can only be understood from the context of the sentence in the document.

The development of query collection also takes into account the correct use of spelling. For example, the word *"Qarun"* is a special name that must be spelt out in uppercase letters for the first letter. The study of this thesis involves the labelling of words by which the word *"Qarun"* will be labelled as a noun based on the first letter of the word. Table 7.1 shows the statistics for query collection produced by the systems with results suggested by a human for that same set of queries. Retrieval evaluation will evaluate the quality of the results. Based on the results obtained, the performance of the search system is evaluated.

In the experiments conducted, two aspects of the assessment are taken into account, namely, recall and precision. Recall and precision is one of the metrics for evaluating the retrieval quality of the IR system [C. J. van Rijsbergen, 1979, Chowdhury et al., 1983]. Precision is used to measure the percentage of information returned that is correct, i.e, how many of the retrieved documents were relevant. Recall measures, the percentage of relevant documents were retrieved.

For each document listed, recall and precision will be calculated using the formula in 7.1 and 7.2. The high precision is only retrieved records, but high recall is to find all retrieved records as relevant [Zoghi et al., 2014].

$$Precision = \frac{tp}{tp + fp} \tag{7.1}$$

$$Recall = \frac{tp}{tp + fn} \tag{7.2}$$

$$F1Score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{7.3}$$

The total number of queries is 30 with average words in query 4.5333. This query collection consists of a variety of topics including human events, the history of the past, prophets and their people, living creation, the Hereafter, and many more.

Table 7.1: Statistic of Query Collections

Query Collections	Statistic
Total number of query	30
Number of word	136
Maximum number of word in the query	9
Minimum number of word in the query	2
Average words in query	4.5333

7.2 Relevance Judgment

According to Taylor A [Arthur Taylor, 2009] in his Ph.D. thesis, relevance judgments take place during the information search process, and they are influenced by time, context, and situation. The determination of relevance is dependent on several factors and variables which include the criteria used to make relevance judgments. The standard approach to information retrieval system evaluation revolves relevance around the notion of relevant and non-relevant documents.

In this thesis, we examined criteria for relevance judgments as identified by subjects who were experts in Qur'an in which they were conducting searches. Here, we use Pouzi's query collection [Mohd Pouzi Hamzah, 2006] as one of relevant judgment. Pouzi's query collection formulates the relevant document with the assistance of two Muslim religious experts who specialise in the Qur'an and Arabic studies. This study used the Malay query words taken from Pouzi's collection as natural language queries and the English query words are translated from these collections. On top of that, we do additional checks with another two experts in Quranic studies from Malaysia University. The reason why we conducted additional judgment is that some of the queries are not in Pouzi's collection and we need an expert to determine the relevancy of that particular queries. Moreover, the judgment is conducted not only in Arabic and English documents but it more in Malay documents. So we need a subject who is an expert in the field of Quranic, especially in the Malay translated Qur'an. Subjects then conducted searches and reviewed documents returned from their searches. Lastly, subjects indicated the relevance for the document they examined. All of these approaches are chosen because the determination of the relevance of the document for a query of the Qur'anic interpretation document is most accurately performed by knowledgeable users in this field.

7.3 Retrieval Evaluation

To evaluate an IR system is to measure how well the systems meet the information needs of the users. Retrieval evaluation is a process of systematically associating a quantitative metric to the results produced by an IR system in response to a set of user queries. This metric should be directly related to the relevance of the results to the users. A common approach to measuring such a metric is to compare the results produced by the system with the results suggested by a human for that same set of queries. Retrieval evaluation will evaluate the quality of the results. Based on the results obtained, the performance of the search system is evaluated.

To conduct the retrieval evaluation, we created a dataset as proposed in the MyQOS ontology discussed in Chapter 6. Table 7.2 presents the number of instances for each main class of the MyQOS ontology. This will be a dataset used for the experiment performed in this study. The MyQOS ontology was evaluated by using the MyQOS search engines in a query enrichment process. The evaluation is carried out by judging how the ontology can help to answer the queries questions. The same dataset was used to test the retrieval of the keyword-based search and question answering search. The dataset contained 8,198 instances. A total number of 30 queries related to the dataset was evaluated in this thesis.

Table 7.2: Ontology Metric MyQOS

Class Name	**Number of Instances**
Chapter	114
Verse	6,236
'Kedudukan' (Location)	86
'Penciptaan' (Living Creation)	1,762
Total	8,198

Precision is a good measure to determine when the cost of False Positives is high. For instance, the search identifies 30 documents; 20 are relevant (true positive) and 10 were on irrelevant (False Positive) topics. The search also returned 40 additional relevant pages (False Negatives). Here, we can say that the precision is 20/30 = 0.67 (67% of hits were relevant) and recall is 20/60 = 0.33 (33% of relevant were found). F1 score is the weighted average of precision and recall. Therefore, this score takes both false positives and false negatives into account. Intuitively, it is not as easy to understand as accuracy, but F1 is usually more useful than accuracy, especially when an uneven class distribution. Accuracy works best if false positives and false negatives have similar cost. If the cost of false positives and false negatives are very different, it is better to look at both precision and recall. The precision and recall are the relative and dependent terms for the evaluation. Precision is based on the documents retrieved, but the recall is based on the relevant documents in collection. The high precision is only retrieved records, but high recall is to find the retrieved records as relevant [Zoghi et al., 2014].

7.4 Evaluation Results

To date, many studies on the effects of semantic information on information retrieval have been carried out by researchers for English documents. No similar studies have been done for Malay documents. An experiment was conducted to answer the following questions:

1. Can semantic annotation improve retrieval quality and accuracy compared to a keyword-based search in Malay Translated Qur'an text?

2. What is the effect of using semantic annotation compared to a keyword-based search?

To answer the above two questions, three separate experiments were conducted, and the results of the analysis were reported in the next section. The experiments were set up with two documents with two languages (Malay and English). The experiments were designed to compare the results obtained by three different search approaches:

- Keyword-based search: a conventional keyword-based retrieval model.

- Question answering search: a conventional keyword-based retrieval model using text processing algorithm.

- Semantic-based search: the semantic retrieval model including the combination of keyword-based and ontology-based retrieval results.

Table 7.3 shows the query collection of 30 queries evaluated in this thesis. However, only first ten queries selected among the 30 generated queries were used in this evaluation section. The rest of the query is presented in Appendix E. The focus of this evaluation was to measure the retrieval quality and accuracy using the two measurements used by most researchers, namely, Recall, and Precision. The results will be compared between the methods.

Table 7.3: Query Collection

Query No	**Query (Malay)**	**Query (English)**
Q1	Apa itu Syurga?	What is Paradise?
Q2	Apa itu Neraka?	What is Hell?
Q3	Apa itu Jin?	What is Jinn?
Q4	Syaitan enggan sujud kepada Nabi Adam.	Satan refused to bow to Adam.
Q5	Cerita tentang Qarun.	Story about Qarun.
Q6	Ayat berkaitan kaum Nabi Nuh.	Verse related to People of Noah.
Q7	Cerita mengenai orang tua kaum Madyan.	Story about the old man of Madian.
Q8	Maklumat tentang kejadian manusia iaitu Nabi Adam.	Information on human creation Adam.
Q9	Ayat berkaitan tiupan sangkakala.	Verse related to The Trumpet is blown.
Q10	Kisah Lelaki bersama Nabi Yusuf di penjara.	Story about a man with Joseph in prison.
Q11	Keterangan Hari Mahsyar iaitu hari perhimpunan.	Evidence of the Gathering day is the day of assembly.
Q12	Ayat berkaitan kaum Nabi Lut.	Verse related to People of Lot.
Q13	Siapa itu Israfil?	Who is Israfil?
Q14	Apa yang terjadi pada isteri Al-Aziz?	What happened to the wife of Aziz?
Q15	Cerita berkaitan dengan penciptaan tumbuhan.	Stories related to the creation of plants.
Q16	Apa itu hari Sabat?	What is a Sabbath?
Q17	Tegahan memakan Babi.	Prohibition of eating pig.
Q18	Siapa itu Bunyamin?	Who is Bunyamin?
Q19	Kisah Yakjuj and Makjuj	Stories about Mog and Magog

Q20	Ayat berkaitan kaum Tsamud.	Verse related to People of Thamud.
Q21	Kisah syaitan.	Story about satan.
Q22	Malaikat-malaikat yang disebut didalam Al-Qur'an	The angels mentioned in the Qur'an
Q23	Ayat yang menceritakan berkaitan dengan keluarga Nabi Musa a.s.	The verse which relates to the family of Moses a.s.
Q24	Kisah Nabi Daud a.s	The story of Prophet David a.s
Q25	Apa itu Manna dan Salwa?	What is Manna and Salwa?
Q26	Gunung Sinai	Mount Sinai
Q27	Ayat berkaitan dengan Laut Merah	The verse related to the Red Sea.
Q28	Khasiat buah delima.	Benefits of pomegranate.
Q29	Apa ayat Al-Quran yang menceritakan tentang belalang	What is the Qur'anic verse that tells the locusts?
Q30	Kisah Nabi Ishak a.s.	Story about Prophet Isaac a.s.

7.4.1 Keyword-based Search

To carry out a fair comparison, the MyQOS keyword search is compared with the existing English Translated Qur'an system. In this evaluation, we used Qurany by [N H Abbas, 2009] as the evaluation document for English Translated Qur'an. However, the use of Qurany system is only for keyword-based search evaluation. It is not for the whole evaluation process. This is because Qurany does not provide other search query method except keyword-based. For each search query, all the documents in the MyQOS system was examined, and their relevance to the query was determined.

The recall precision (%) is generated for the evaluation of search results. Considering the same example of the first 10 queries. The precision and recall plots for the 10 queries were discussed in Table 7.3. Based on the result, the average precision and recall achieved by Qurany system is 0.4557 and 0.6688 while MyQOS is 0.5471 and 0.5884. It can be seen that both Qurany and MyQOS have moderate precision and moderate recall. This is because both Qurany and MyQOS were based on a traditional keyword search, which lacked of semantic elements. The fact that the system is based on a keyword search contributes to the moderate precision and moderate recall. The Qurany system attempts to syntactically match a query term with the corresponding term in the Qur'an verse. In this case, any of the Qur'an verse terms that has at least one match with any of the query terms is retrieved. For instance, query Q4 is *"Who is Jinn"* retrieved 152 verses in Qurany and 63 verses in MyQOS. Here, the precision is low even though recall is high because the systems retrieve too many irrelevant verses. (See Appendix C for the list of the results of relevant and retrieved documents)

Table 7.4: Precision and Recall based on Keyword-based Search

QueryNo	**Query (English)**	**Precision**		**Recall**	
		Qurany	**MyQOS**	**Qurany**	**MyQOS**
Q1	What is Paradise?	0.1345	0.1751	0.6531	0.6327
Q2	What is Hell?	0.2115	0.2379	0.8148	0.7901
Q3	What is Jinn?	0.1974	0.5556	0.8108	0.9459
Q4	Satan refused to bow to Adam.	0.5333	1.0000	1.0000	0.3750
Q5	Story about Qarun.	0.8000	0.5000	1.0000	0.5000
Q6	Verse related to People of Noah.	0.2273	0.4375	0.8824	0.8235
Q7	Story about old man of Madian.	1.0000	0.5000	0.2000	0.2000
Q8	Information on human creation. Adam and Eve.	0.0244	0.5455	0.1000	0.6000
Q9	Verse related to The Trumpet is blown.	0.7619	0.7692	0.9412	0.5882

Q10	Story about man with Joseph in prison.	0.6667	0.7500	0.2857	0.4286
	Average	**0.4557**	**0.5471**	**0.6688**	**0.5884**

The plots of recall and precision are considered separately for the data in Table 7.4 and plotted in Figures 7.1 and 7.2. The plots considered define the variation of the calculation of recall and precision after every document retrieval. The evaluation results of Qurany and MyQOS are interpreted based on 10 queries. As Figures 7.1 and 7.2 show, there is a significant difference between the two systems. For instance, Q4 is *"Syaitan enggan sujud kepada Nabi Adam"* (Satan refused to bow to Adam). The precision for Qurany were 0.5333 (53% of hits were relevant) while MyQOS was only 0.8571 (86% of hits were relevant). By using the keyword search, it will search only *"Syaitan"* or *"enggan"* or *"sujud"* or *"Adam"* separately. Thus, it would not be able to yield the desired results of the actual query. For the results, the user has to fire the query and manually merge the results separately.

The precision vs. recall graph provides performance of the search engine. According to [Arora et al., 2016], when the plotted line is in the upper-right portion of the graph, the selected category is performing well. When the plotted line is in the lower-left portion of the graph, this indicates that the category's performance is poor. Figures 7.1 and 7.2 illustrate the precision vs. recall graph performance of the keyword-based search engine for Qurany and MyQOS systems. The curves shown are in the upper-right portion of the graph. This indicates that both search engine performs quite well. Based on the data, consider the calculated values for precision and recall the average, which means that the queries are equally satisfied with the data retrieval mechanism provided to them. The keyword-based search for both systems shows low precision and low recall, which just satisfied the queries between 40% to 50% of relevant and retrieved documents.

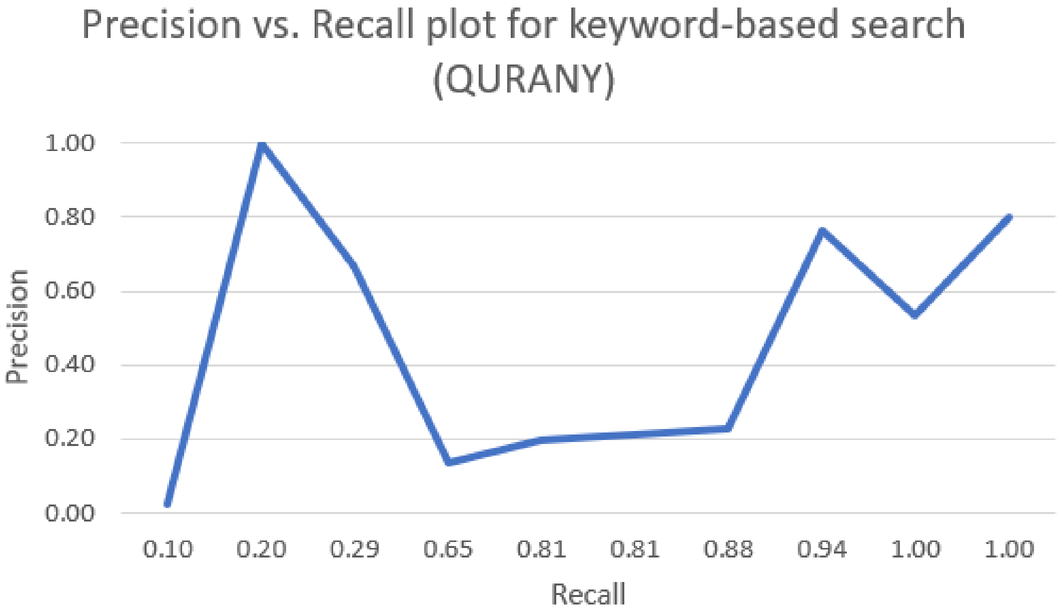


Figure 7.1: Precision vs recall plot for Qurany Keyword-based search

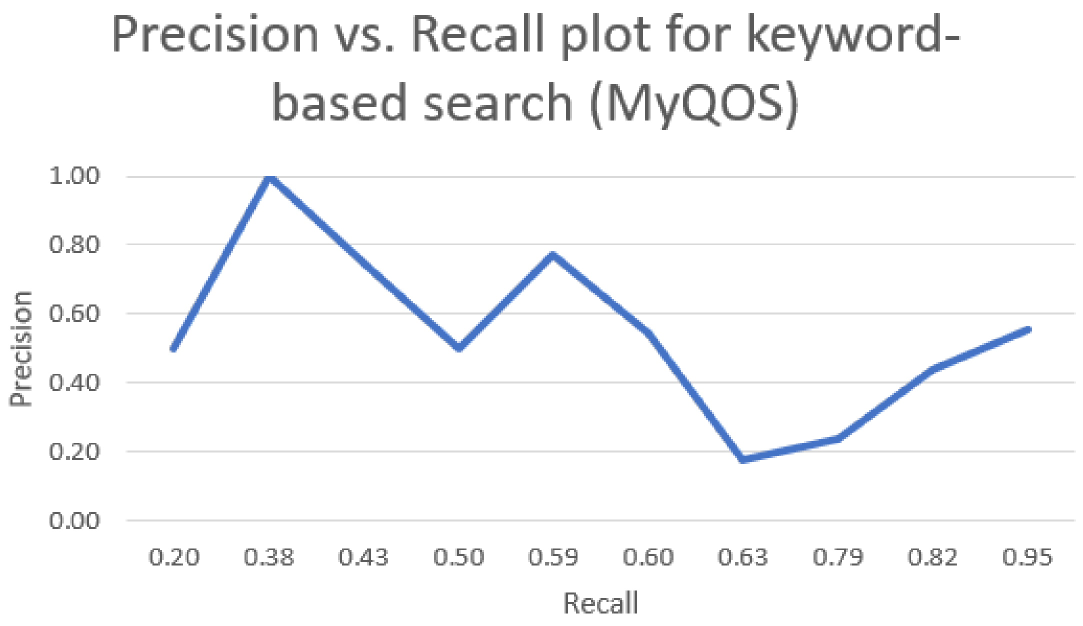


Figure 7.2: Precision vs recall plot for MyQOS Keyword-based search

The overall result of the analysis proves that there is a problem in retrieving information using only keyword-based. It retrieves all verses containing those keywords without considering the answer is relevant or not from the user's context. This is because it is unable to gather complex information. Although the keyword-based search is beneficial in finding specific information, it lacks in finding the meaning of the terms, expressions used in the documents and the relationships between them, especially in Malay documents. The problem comes due to the existence of words which have many meanings known as polysemy and several words having the same meaning also known as synonyms in natural languages. In the next evaluation, the retrieval of the system will be shown with the use of text processing in the natural language queries.

7.4.2 Question Answering Search

In this evaluation, we evaluate two retrieval methods on the MyQOS system, namely, keyword search and question and answering search (use text processing method discussed in chapter 5). In this first evaluation, we made a comparison between the the previous results of keyword-based search with question answering search that use text processing such as stemming, stopword removal, tokenization, and reduplication algorithm . As discussed in chapter 5,text processing will make each word in the query to be root word to have more relevant documents in the results.

Based on the results, the average precision and recall achieved by keyword-based search only is 0.5471 (55%) and 0.5884 (59%), while the question and answering search that uses text processing is 0.4971 (50%) for precision and 0.6027 (60%) for recall. The precision for question answering search is lower than keyword search. Question answering search still gives a lot of unwanted documents in the retrieval result. Although there are only a few differences in these two search engines, we can see that searches with text processing improve in some queries. For instance, query 10 retrieved 100% of precision in both keyword search and question answer search. However, the recall rates are moderate and it just between 40% and 58% due to the large number of retrieved documents which is not relevant. The high precision implies low recall, which means there are only a few documents for which the system can be very certain that they are correct. This performance can be observed from the precision values for some queries. However, if we want to evaluate the performance of a system in terms of retrieving every potentially relevant document, it needs to examine recall.

Table 7.5: Precision and Recall based on Keyword-based vs Question Answering Search

No	**Query (English)**	**Precision**		**Recall**	
		Keyword only	**Question Answering Search**	**Keyword only**	**Question Answering Search**
Q1	What is Paradise?	0.1751	0.1751	0.6327	0.6327
Q2	What is Hell?	0.2379	0.2379	0.7901	0.7901
Q3	What is Jinn?	0.5556	0.5556	0.9459	0.9459

Q4	Satan refused to bow to Adam.	0.2500	1.0000	0.3750	0.3750
Q5	Story about Qarun.	0.5000	0.5000	0.5000	0.5000
Q6	Verse related to People of Noah.	0.4375	0.4375	0.8235	0.8235
Q7	Story about old man of Madian.	0.5000	0.5000	0.2000	0.2000
Q8	Information on human creation. Adam and Eve.	0.5455	0.5455	0.6000	0.6000
Q9	Verse related to The Trumpet is blown.	0.7692	0.7692	0.5882	0.5882
Q10	Story about man with Joseph in prison.	1.0000	1.0000	0.4286	0.5714
	Average	**0.5471**	**0.4971**	**0.5884**	**0.6027**

The plots of recall and precision are considered separately for the data in Table 7.5 and plotted in Figures 7.1 and 7.3. The plots considered define the variation of the calculation of recall and precision after every document retrieval. The curves shown in Figure 7.3 are in the upper-right portion of the graph. This indicates the question answering search performs quite well. The combination of text processing methods in question answering search gives definite improvement at all retrieval points as compared to the keyword only method. This method outperforms previous methods with significant differences. It retrieves all verses containing those keywords and merges the results. For instance, the user fires a query *"Ayat berkaitan kaum Nabi Nuh"* (Verse related to People of Noah). By using a stemming algorithm, it will remove the affixes *"ber"* and *"an"* from the word *"berkaitan"* and search only the root words found in the query such as *"Ayat kait kaum Nabi Nuh"*. Here, the words will be combined to give one meaningful sentence. Thus, it would be able to yield the desired results. Figure 7.3 illustrates the precision vs. recall graph performance of the keyword-based search with stemming algorithm.

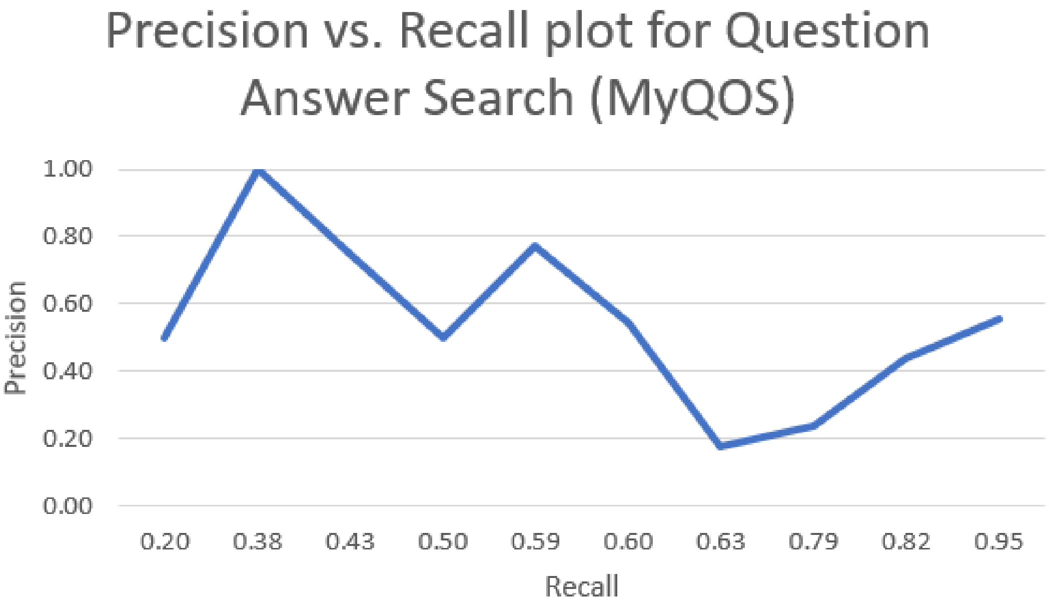


Figure 7.3: Precision vs recall plot for MyQOS Question Answering Search

From the experiment, it is clear that combining methods of text processing with question answering search proves to be more effective in retrieving more relevant documents when compared to keyword search. Although this combination seems to be the best of all, the retrieved and relevant result is still moderate, which is only 60.27%. This implies that that there are still key terms in the Qur'an documents that are not available. The average precision value is the lowest 49.71% due to the enormous number of documents retrieved while using this combination method. The key terms identified are found in the Qur'an documents, but these documents are not listed in the relevance judgment list. This implies that there are still other terms used in the Qur'an that carry the same meaning that have not been retrieved. Performance measures imply limited impact on retrieval performance, possibly due to limited semantic capabilities of expansion terms. One of the most effective techniques to improve the performance of IR systems is expanding the original queries with other terms that can retrieve more relevant documents or can form better queries. In query expansion for information retrieval, if a search does not return enough results, one option is to replace an specific term with a hypernym.

7.4.3 Semantic Search

The basic idea of this thesis is to improve the recall of MyQOS search engine. Search engines retrieve documents related to a specific query term. Generally, the same concepts can be expressed using different terms, thus searching for one of these terms will not retrieve the others. Semantic search can improve the recall since the search query will match an entire term instead of only one or more terms.

An ontology can provide context-aware search capabilities specific to the area of interest. The enhancement, extension, and disambiguation of user query terms become possible with the addition of enriched domain and context specific information. The semantic search can improve information retrieval performance and answer questions which a retrieval system without ontology cannot do. The MyQOS ontology is assessed based on its ability to answer the developed queries. Terms in the ontology were used to query. The retrieval efficiency was measured in terms of its precision and recall. The retrieval evaluation compared a keyword-based search, question answer search, and name entity representation supported with ontology-based query (semantic search).

Table 7.6 shows the comparison of the precision results of the ten queries using three different retrieval methods in MyQOS system.The average precision and recall for semantic search are 0.8409 (84%) and 0.8043(80%), double the results of the question answer which are 0.4971(50%) for precision and 0.6027 (60%) for recall. The semantic search gives high precision and high recall comparing the other two methods. This indicates that semantic search returns more relevant results than irrelevant ones. As conventional search engines cannot interpret the sense of the user's search, not all verses that discuss the concept can be retrieved, the ambiguity of the query leads to the retrieval of irrelevant information. Conventional search engines that match query terms against a keyword-based index will fail to match relevant information when the keywords used in the query are different from those used in the index, despite having the same meaning (synonym). For lack of context, many search engines fail to take into consideration aspects of the user's context to help disambiguate their queries.

Table 7.6: Precision and Recall based on Keyword-based, Question Answering Search and Semantic Search

Query	**Precision**			**Recall**		
	Keyword only	**Question Answering Search**	**Semantic**	**Keyword only**	**Question Answering Search**	**Semantic**
Q1	0.1751	0.1751	0.9800	0.6327	0.6327	1.0000
Q2	0.2379	0.2379	0.9878	0.7901	0.7901	1.0000
Q3	0.5556	0.5556	1.0000	0.9459	0.9459	0.8919
Q4	1.0000	0.2500	0.8571	0.3750	0.3750	0.7500
Q5	0.5000	0.5000	0.4000	0.5000	0.5000	0.5000
Q6	0.4375	0.4375	1.0000	0.8235	0.8235	0.7059
Q7	0.5000	0.5000	0.8333	0.2000	0.2000	1.0000
Q8	0.5455	0.5455	0.7273	0.6000	0.6000	0.8000
Q9	0.7692	0.7692	0.8235	0.5882	0.5882	0.8235
Q10	0.7500	1.0000	0.8000	0.4286	0.5714	0.5714
Average	**0.5471**	**0.4971**	**0.8409**	**0.5884**	**0.6027**	**0.8043**

Based on Table 7.6, Q1,Q2, and Q7 gave 100% of of recall in semantic search. This shows that these queries return the relevant document related to the query. The semantic annotations using synonyms give definite improvement at all retrieval points compared to the keyword method. This method outperforms previous methods with significant differences. It retrieves mostly all relevant documents for every query. For instance, a query about *"Maklumat berkaitan dengan Syurga"* (Information related to Paradise). As we been discussed about this in previous chapters, there are other terms used in the Qur'an that represent Paradise. Using only the keyword-based method, it is impossible to get the relevant document related to the query. However, using semantic annotation methods, all semantic terms related to queries will be annotated. In this example, the other terms that are used in the Qur'an to represent Paradise are *"Darussalam"*, *"Eden"* and many more.

Figure 7.4 illustrates the comparison of precision using a plotted graph between the three methods.

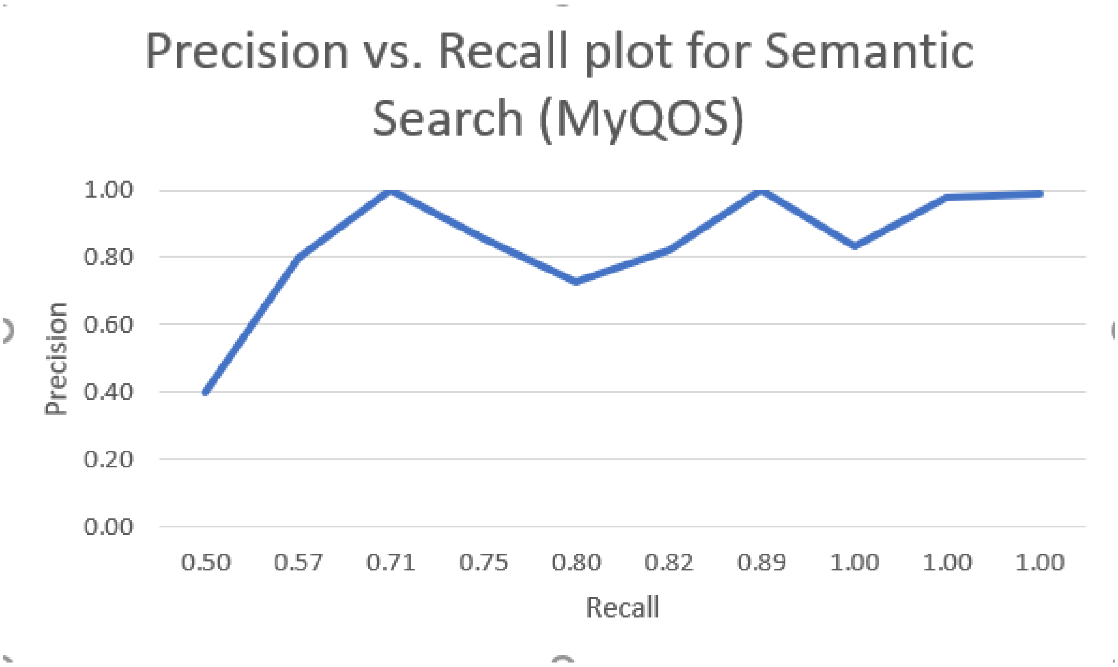


Figure 7.4: The comparison of precision of keyword-based, Question Answering Search and Semantic Search based on 10 queries.

In summary, these results show that semantic annotation has a significant impact on the effectiveness of retrieval. The results obtained from this analysis indicate that semantic search can help improve the information retrieval method for Malay documents. The findings of the current study are consistent with those of [Fernández Sánchez, 2009, Tian, 2012] who said that using semantic annotation can improve retrieval effectiveness and quality. The adaption of this technique in Malay documents was very successful and will become a benchmark for other researchers who wish to conduct research on Malay documents.

Together these results provide important insights into this thesis. The evaluation has demonstrated that our approach in semantic search is better than the other two methods. This study produced results which corroborate the findings of a great deal of previous work in this field. Comparing the results of the semantic search with keyword-based IR and question answer search are summarised below:

1. The MyQOS Semantic search provide more relevant results according to the user query. The creation of MyQOS ontology help to facilitate the semantic search and increase the recall rate.

2. The returned results of the semantic search are vigorous and more relevant to the field of study than keyword-based search and question answering search.

3. The proposed ontology plays an important role to optimise the returned results in semantic search. The semantic annotation with synonyms provides more relevant results.

4. More consistency between terms in the ontology will provide more relevant results, and it will increase the precision and recall.The consistency is important to avoid any internal contradiction such as duplication of term or ambiguous term used in the ontology. This requires OWL specific constraints or rules (Disjunction, Inference) to improve reasoning. Moreover, consistency adequately represents reality.

7.5 Summary

In this chapter, we have presented the evaluation results for the three methods used in MyQOS system. We also have presented the evaluation results of the semantic search in MyQOS with other existing semantic search systems. The results are promising and demonstrate the proof of concept for the approach proposed in this thesis. Based on this analysis, MyQOS provides a step towards the understanding of scalable and effective Semantic Web applications, able to deal with the new layers of complexity introduced by the continuous growth of the Semantic Web. As more information becomes available and the quality of the data improves, it will become possible for MyQOS to focus primarily on precision rather than recall, thus leading to better accuracy. The present results are significant in at least major three respects such as:

1. The meaning of an ontological concept must be precisely defined in the ontology.
2. Semantic annotation using synonym can be used to improve the quality and accuracy of information retrieval rather than the use of keyword-based.
3. MyQOS semantic search outperforms keyword-based search in terms of both precision and recall.

The results of this evaluation have answered the questions mentioned in the previous section:

1. Can semantic annotation improves retrieval quality and accuracy compared to a keyword-based search in Malay Translated Qur'an text?
2. What is the effect of using semantic annotation compared to keyword-based search?

It proves that semantic search is better that keyword-based search and it improves the retrieval quality and accuracy. The evaluations presented in this chapter have been verified against the available contents and the domain ontologies presented in chapter 6. Besides the development and evaluation work, the possibilities for the continuation of the research are manifold. Our work is motivated by the subtleties of semantic retrieval precision and recall, which is variable considering the completeness of the semantic knowledge available for each user request, and do not play an equally important role in all situations. Solving this complexity of the language is inherently difficult, but coping with it to some degree is likely to be key for the robustness and reliability of semantic retrieval systems.

Our approach is new in Malay language area that it combines natural language processing and ontology. The benefit is twofold: the semantic retrieval techniques gain accuracy and quality by each new executed query and the results obtained are filtered, enriched, and made more coherent considering the statistical behaviour of both semantic-based and keyword-based algorithms.

In Chapter 8, we describe the detailed implemented prototype of MyQOS. This prototype will give an overview of how the system works.

Chapter 8

The Implementation of MyQOS System

In this chapter, we briefly mention about MyQOS prototype that we implemented using the framework described in this thesis. Then, we describe the tools that we used for the implementation.

8.1 Prototype Implementation

The prototype implementation is a realisation of a technical specification or algorithm as a program, software component, or other computer systems, achieved through computer programming and deployment. prototype called Malay Translated Qur'an Ontology System or known as MyQOS throughout the thesis. MyQOS provides several features such as searching, resource allocation, and ontology contents. The prototype system is implemented using Java programming language, Jena framework, Java Server Pages (JSP) and HyperText Markup Language (HTML).

8.2 Jena Framework

Jena Semantic Web is an open-source Java. It provides an API to retrieve the data and record RDF/OWL graphs. Graphs are presented as an abstract "model". The model can be derived from data files, databases, and URL-addresses. OWL file, which is an anthology of Malay translated Qur'an derived from using the Protégé-OWL program. The figure shows the Java code used for reading and processing the RDF/OWL file. The figure shows the Java code used for displaying the classes.

8.3 User Interfaces Design (UI)

User interface design is the design of user interfaces for software with the focus on maximising usability and the user experience. The main goal of designing the user interface is to make the user's interaction as simple and efficient as possible. Good user interface design facilitates finishing the task at hand without drawing unnecessary attention to itself. The design process must balance technical functionality and visual elements to create a system that is not only operational but also usable and adaptable to changing user needs.

The primary function of MyQOS is to enable the searching facilities. There are several types of searching methods implemented in MyQOS, such as keyword-based search, question answering system, and Ontology-based search. Each search has its uniqueness and retrieval output. Moreover, MyQOS prototype also provides the Malay language NLP resources that can be used by other people.

8.3.1 Main Interface

Figure 8.1 shows the Home interface of MyQOS prototype. In this home interface, users can search for anything related to the Qur'an in Malay language using the keyword. Besides, there is a list of a menu on the left which can be navigated by the user. The list of a menu are:

1. Home
    - About MyQOS
2. MyQOS Corpus
    - Malay Qur'an Translation
    - List of Malay Stopword
    - List of Malay Root words

- List of Malay POS-Tagging

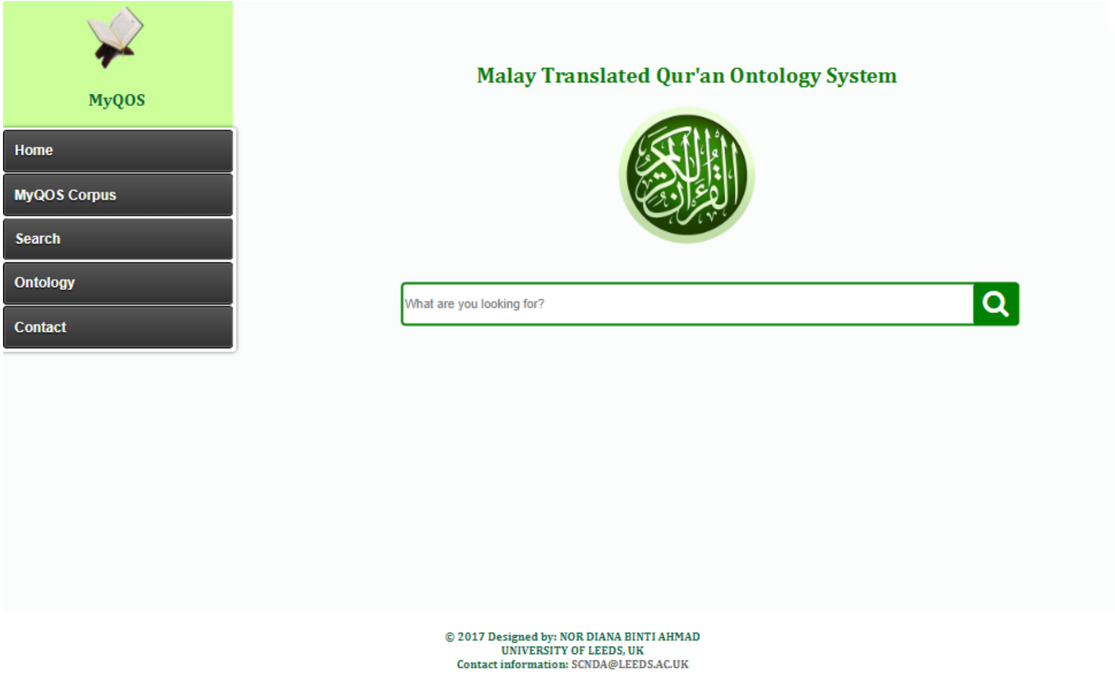
3. Search

    - Keyword-based Search
    - Question-Answering Search
    - Semantic Search

4. Ontology

    - Overview
    - Concept Index

5. Contact Us



Malay Translated Qur'an Ontology System

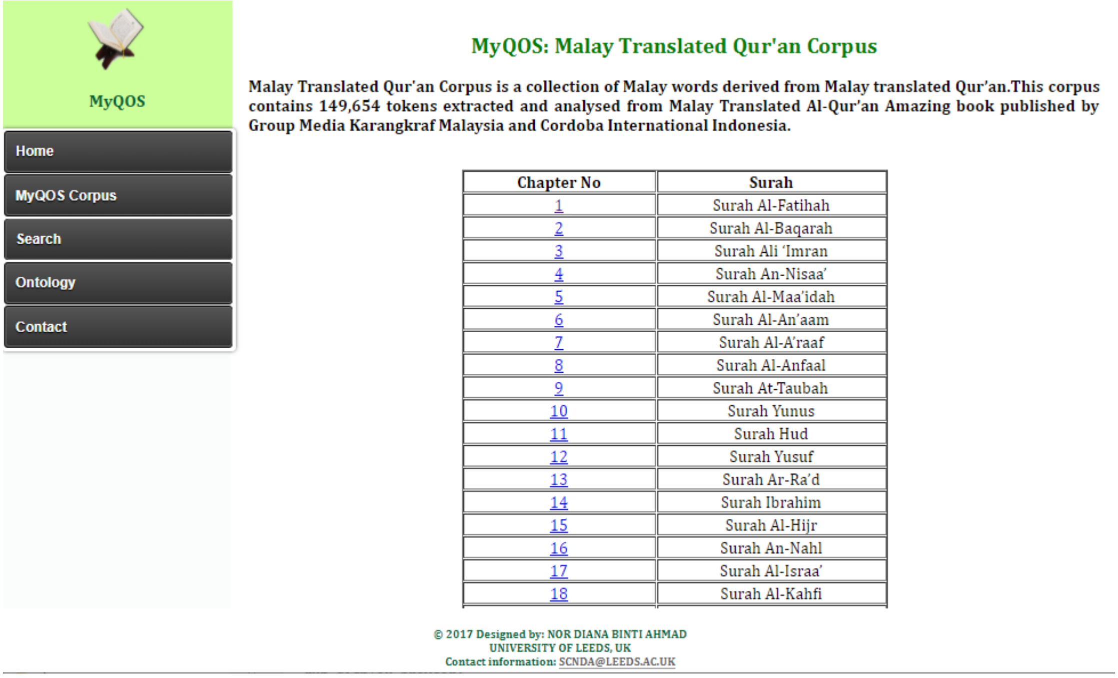
© 2017 Designed by: NOR DIANA BINTI AHMAD
UNIVERSITY OF LEEDS, UK
Contact information: SCNDA@LEEDS.AC.UK

Figure 8.1: Main Homepage

8.3.2 MyQOS Corpus

MyQOS Corpus is an online collection of Malay words derived from Malay translated Qur'an. This corpus contains 149,654 tokens extracted and analysed from Malay Translated Al-Qur'an Al-Qur'an Amazing book [Nursalim et al., 2016]. It is divided into 114 chapters (Suras) of varying sizes, where each chapter is divided into verses (Ayahs). There are 6,234 verses in this corpus.

In this page, we present the MyQOS Corpus which can be accessed online. This corpus is a result of the preprocessing stages that are discussed in Chapter 5. This can be assessed by other research if they want to use this corpus. Using this corpus, users can browse each of the verses by chapter. Figure 8.2 shows the list of Malay translated Qur'an corpus.



MyQOS: Malay Translated Qur'an Corpus

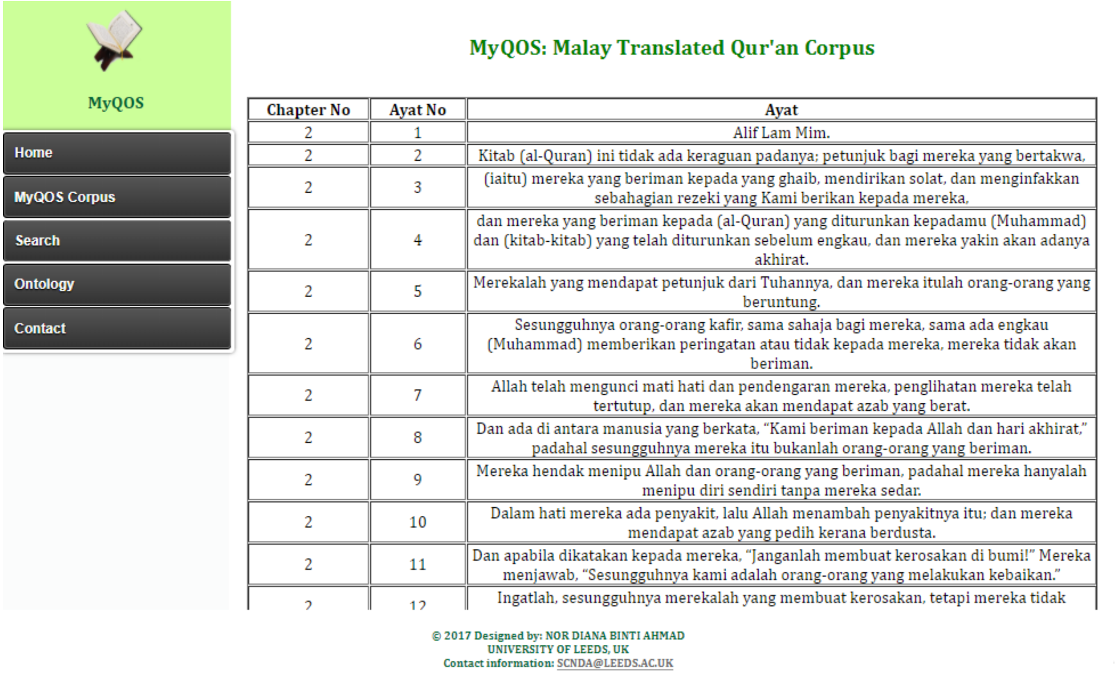
Malay Translated Qur'an Corpus is a collection of Malay words derived from Malay translated Qur'an.This corpus contains 149,654 tokens extracted and analysed from Malay Translated Al-Qur'an Amazing book published by Group Media Karangkraf Malaysia and Cordoba International Indonesia.

Chapter No	Surah
1	Surah Al-Fatihah
2	Surah Al-Baqarah
3	Surah Ali 'Imran
4	Surah An-Nisaa'
5	Surah Al-Maa'idah
6	Surah Al-An'aam
7	Surah Al-A'raaf
8	Surah Al-Anfaal
9	Surah At-Taubah
10	Surah Yunus
11	Surah Hud
12	Surah Yusuf
13	Surah Ar-Ra'd
14	Surah Ibrahim
15	Surah Al-Hijr
16	Surah An-Nahl
17	Surah Al-Israa'
18	Surah Al-Kahfi

© 2017 Designed by: NOR DIANA BINTI AHMAD
UNIVERSITY OF LEEDS, UK
Contact information: SCNDA@LEEDS.AC.UK

Figure 8.2: Display all the chapter of the Qur'an

Once the user selects the desired chapter, it will be linked directly to the details of each chapter. Figure 8.3 shows the detailed of chapter 2, Surah Al-Baqarah.



MyQOS: Malay Translated Qur'an Corpus

Chapter No	Ayat No	Ayat
2	1	Alif Lam Mim.
2	2	Kitab (al-Quran) ini tidak ada keraguan padanya; petunjuk bagi mereka yang bertakwa,
2	3	(iaitu) mereka yang beriman kepada yang ghaib, mendirikan solat, dan menginfakkan sebahagian rezeki yang Kami berikan kepada mereka,
2	4	dan mereka yang beriman kepada (al-Quran) yang diturunkan kepadamu (Muhammad) dan (kitab-kitab) yang telah diturunkan sebelum engkau, dan mereka yakin akan adanya akhirat.
2	5	Merekalah yang mendapat petunjuk dari Tuhannya, dan mereka itulah orang-orang yang beruntung.
2	6	Sesungguhnya orang-orang kafir, sama sahaja bagi mereka, sama ada engkau (Muhammad) memberikan peringatan atau tidak kepada mereka, mereka tidak akan beriman.
2	7	Allah telah mengunci mati hati dan pendengaran mereka, penglihatan mereka telah tertutup, dan mereka akan mendapat azab yang berat.
2	8	Dan ada di antara manusia yang berkata, “Kami beriman kepada Allah dan hari akhirat,” padahal sesungguhnya mereka itu bukanlah orang-orang yang beriman.
2	9	Mereka hendak menipu Allah dan orang-orang yang beriman, padahal mereka hanyalah menipu diri sendiri tanpa mereka sedar.
2	10	Dalam hati mereka ada penyakit, lalu Allah menambah penyakitnya itu; dan mereka mendapat azab yang pedih kerana berdusta.
2	11	Dan apabila dikatakan kepada mereka, “Janganlah membuat kerosakan di bumi!” Mereka menjawab, “Sesungguhnya kami adalah orang-orang yang melakukan kebaikan.”
2	12	Ingatlah, sesungguhnya merekalah yang membuat kerosakan, tetapi mereka tidak

© 2017 Designed by: NOR DIANA BINTI AHMAD
UNIVERSITY OF LEEDS, UK
Contact information: SCNDA@LEEDS.AC.UK

Figure 8.3: Detailed of Chapter 2, Surah Al-Baqarah

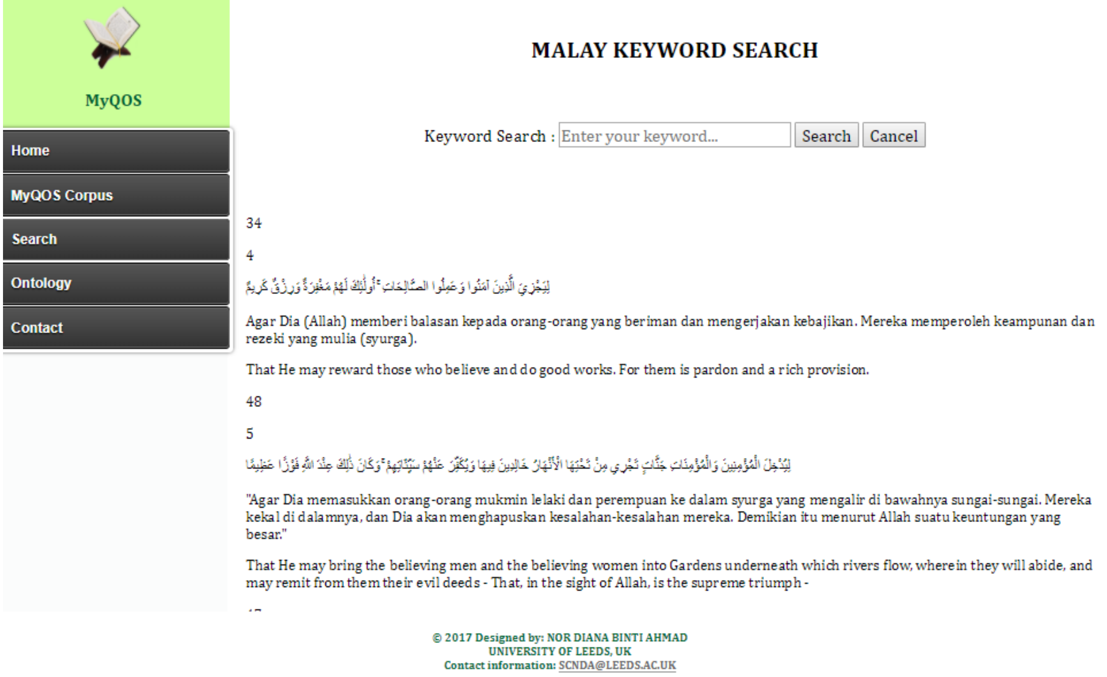
8.3.3 Searching

Searching is the main section of this prototype. The user can search for a certain verse in the Malay translated Qur'an. The focus on the searching method to retrieve relevant and quality results based on the user query. In MyQOS prototype, there are three types of searching methods that were implemented.

- Keyword-based search: a conventional keyword-based retrieval model.
- Question answering search: a conventional keyword-based retrieval model using text processing algorithm.
- Semantic-based search: the completed semantic retrieval model including the combination of keyword-based and ontology-based retrieval results.

Keyword-based Search

Keyword-based search retrieves all verses containing those keywords without considering the fact that the user's context produces an accurate answer. It was developed using SQL query. The user will enter the keyword, and then the system will process the submitted query and show the result in a separate page illustrated in Figure 8.4.



MALAY KEYWORD SEARCH

Keyword Search : Enter your keyword...

34

4

لِيَجْزِيَ الَّذِينَ آمَنُوا وَعَمِلُوا الصَّالِحَاتِ ۚ أُولَٰئِكَ لَهُمْ مَغْفِرَةٌ وَرِزْقٌ كَرِيمٌ

Agar Dia (Allah) memberi balasan kepada orang-orang yang beriman dan mengerjakan kebajikan. Mereka memperoleh keampunan dan rezeki yang mulia (syurga).

That He may reward those who believe and do good works. For them is pardon and a rich provision.

48

5

لِيُدْخِلَ الْمُؤْمِنِينَ وَالْمُؤْمِنَاتِ جَنَّاتٍ تَجْرِي مِنْ تَحْتِهَا الْأَنْهَارُ خَالِدِينَ فِيهَا وَيُكَفِّرَ عَنْهُمْ سَيِّئَاتِهِمْ ۚ وَكَانَ ذَٰلِكَ عِنْدَ اللَّهِ فَوْزًا عَظِيمًا

"Agar Dia memasukkan orang-orang mukmin lelaki dan perempuan ke dalam syurga yang mengalir di bawahnya sungai-sungai. Mereka kekal di dalamnya, dan Dia akan menghapuskan kesalahan-kesalahan mereka. Demikian itu menurut Allah suatu keuntungan yang besar."

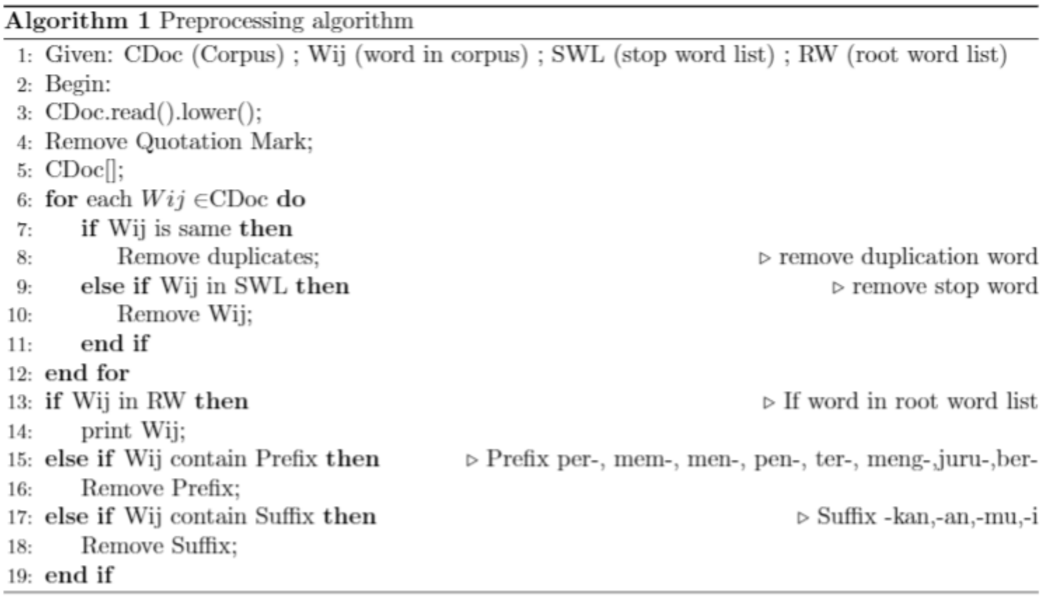
That He may bring the believing men and the believing women into Gardens underneath which rivers flow, wherein they will abide, and may remit from them their evil deeds - That, in the sight of Allah, is the supreme triumph -

© 2017 Designed by: NOR DIANA BINTI AHMAD
UNIVERSITY OF LEEDS, UK
Contact information: SCNDA@LEEDS.AC.UK

Figure 8.4: Keyword-based search interface

Question-Answering (QA) Search

Question Answering (QA) search is an information retrieval system which retrieves point-to-point answers rather than flooding with documents. Question-answering search is developed using the stemming algorithm to get better results from the traditional keyword-based. The query will be processed using NLP techniques as discussed in chapter 5. MyQOS QA search sequentially executes each query, concatenating the results (after the morphological analysis is done) until a result has been achieved. Figure 8.5 shows the algorithm for the preprocessing stage.

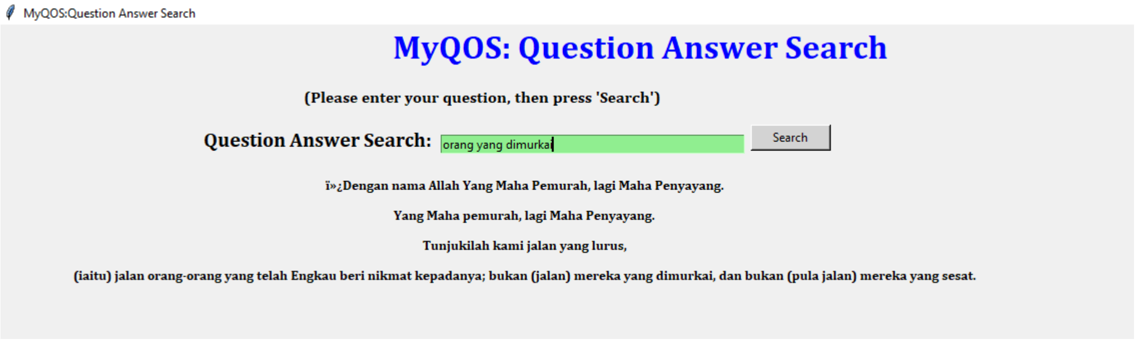


Algorithm 1 Preprocessing algorithm

```
Given: CDoc (Corpus) ; Wij (word in corpus) ; SWL (stop word list) ; RW (root word list)
Begin:
CDoc.read().lower();
Remove Quotation Mark;
CDoc[];
for each $Wij$ ∈CDoc do
    if Wij is same then
        Remove duplicates;                                  ▷ remove duplication word
    else if Wij in SWL then                                 ▷ remove stop word
        Remove Wij;
    end if
end for
if Wij in RW then                                           ▷ If word in root word list
    print Wij;
else if Wij contain Prefix then        ▷ Prefix per-, mem-, men-, pen-, ter-, meng-,juru-,ber-
    Remove Prefix;
else if Wij contain Suffix then                             ▷ Suffix -kan,-an,-mu,-i
    Remove Suffix;
end if
```

Figure 8.5: Preprocessing Algorithm

This prototype has been developed using Python 3.4.3 Tkinter GUI. The prototype preprocesses the text and an input query using a stemming algorithm. The answer extraction process relies on finding the correct stem string from the text, which matches with the user stem question. To obtain the answer, Python functions have been used to read and display the answer. The query process is able to search for multiple strings. At this time, the searching process only uses tokenization, stop word removal and stemming techniques. Extensive work will be done for the POS-tagging methods. Figure 8.6 shows the question-answering prototype.



MyQOS: Question Answer Search

(Please enter your question, then press 'Search')

Question Answer Search: orang yang dimurkai

ï»¿Dengan nama Allah Yang Maha Pemurah, lagi Maha Penyayang.

Yang Maha pemurah, lagi Maha Penyayang.

Tunjukilah kami jalan yang lurus,

(iaitu) jalan orang-orang yang telah Engkau beri nikmat kepadanya; bukan (jalan) mereka yang dimurkai, dan bukan (pula jalan) mereka yang sesat.

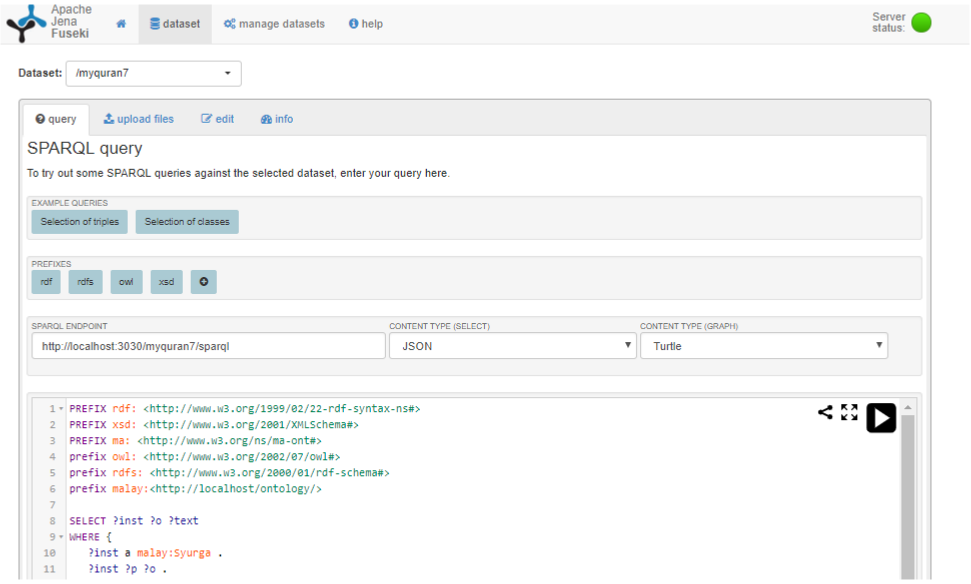
Figure 8.6: Question-Answering Search Interface

After tokenizing the user query, many answers have been displayed, because the system searches for matches in the corpus questions for every keyword found in the user question. After removing the stop words, there are fewer answers than the previous ones and the processing time has improved. The example shown in Figure 8.6 shows the candidate result based on the user' question 'Orang yang dimurkai' (people who are wrecked). The results are based on the matching between the query and answer. The result proved that using the text preprocessing in the query and text file will improve the information retrieval.

8.3.4 Semantic Search

Semantic search requires a search engine to properly interpret the meaning of a user's query and the inherent relations among the terms that a document contains with respect to a specific domain. Traditional search engines do not deal with any domain knowledge, so they do not understand the meaning of a user's search request and the inherent relations among the terms that a web document contains.

In MyQOS Semantic search, we used the SPARQL query to query OWL files in the ontology. Semantic searching seeks to improve the search accuracy of the search engine by understanding the user needs and the contextual meaning of the query term to retrieve the more relevant result. Here, the SPARQL query will search not only the query term but also the sense of the word using additional information provided in the ontology. The use of semantic annotations using a synonym to analyse the context of the annotations to identify their meaning through the entities such as class and instances in the ontology. Figure 8.7 shows the screenshot of Apache Jena with Fuseki server with MyQOS OWL file.



Dataset: /myquran7

SPARQL query

To try out some SPARQL queries against the selected dataset, enter your query here.

SPARQL ENDPOINT: http://localhost:3030/myquran7/sparql

CONTENT TYPE (SELECT): JSON

CONTENT TYPE (GRAPH): Turtle

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX ma: <http://www.w3.org/ns/ma-ont#>
prefix owl: <http://www.w3.org/2002/07/owl#>
prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
prefix malay:<http://localhost/ontology/>

SELECT ?inst ?o ?text
WHERE {
   ?inst a malay:Syurga .
   ?inst ?p ?o .
```

Figure 8.7: The screenshot of Apache Jena with Fuseki server with MyQOS OWL file

The query searches for matches in the ontology. Semantic searching seeks to improve the search accuracy of the search engine by understanding the user needs and the contextual meaning of the query term to retrieve the more relevant result. Here, the SPARQL query will search not only the query term but also the meaning of the word using additional information provided in the ontology. The prototype application is designed and implemented using the Java programming language and Jena Apache framework. However, the prototype of the semantic search is still under development. Listing 8.2 shows the Java code used for reading and displaying the classes.

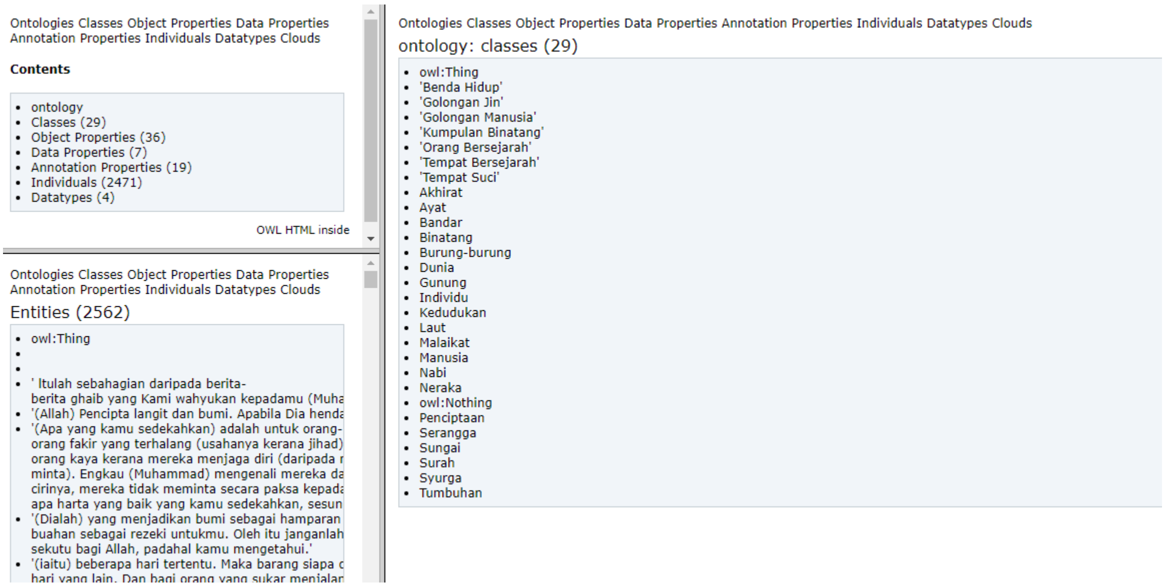
Listing 8.1: Reading and display classes

```
package javardf;
```

```
import org.apache.jena.ontology.OntClass;
import org.apache.jena.ontology.OntModel;
import org.apache.jena.rdf.model.ModelFactory;
import org.apache.jena.util.FileManager;
import org.apache.jena.util.iterator.ExtendedIterator;
import java.io.InputStream;
import java.util.Iterator;

import org.apache.log4j.Logger;
import org.apache.log4j.PropertyConfigurator;

public class GenerateRDF {
static Logger logger = Logger.getLogger(OWLClass.class);
static final String inputFileName =
”C:\\xampp\\htdocs\\jOWLBrowser\\data\\myquran7.owl”;

public static void main(String args[]) {
try {
PropertyConfigurator.configure
(”C://apache-jena-3.6.0//jena-log4j.properties”);
//create the reasoning model using the base
OntModel model = ModelFactory.createOntologyModel();

// use the FileManager to find the input file
InputStream in = FileManager.get().open(inputFileName);
if (in == null) {
throw new IllegalArgumentException
(”File:␣” + inputFileName + ”␣not␣found”);
}

model.read(in, ””);

//to list classes
ExtendedIterator classes = model.listClasses();
while (classes.hasNext()) {
OntClass cls = (OntClass) classes.next();

System.out.println("Classes:␣" + cls.getLocalName());
for (Iterator i = cls.listSubClasses(true); i.hasNext();) {
OntClass c = (OntClass) i.next();
System.out.print("␣" + c.getLocalName() + "\n");
} // end for
}
} catch (Exception e) {
System.out.println(e);
}
}

    private static class OWLClass {

        public OWLClass() {
        }
    }
}
```

8.3.5 Ontology

To display and visualise the concept in the ontology, we used the tools in Protégé. Using these tools, we can easily create OWL documentation and view the information related to the ontology. In the OWL documentation, it has three (3) sections: Contents, Entities (Instances) and Classes. The user can click on each, and it will link to related instances. Figure 8.8 shows the screenshot of MyQOS OWL Documentation.



Ontologies Classes Object Properties Data Properties Annotation Properties Individuals Datatypes Clouds

Contents

- ontology
- Classes (29)
- Object Properties (36)
- Data Properties (7)
- Annotation Properties (19)
- Individuals (2471)
- Datatypes (4)

OWL HTML inside

Ontologies Classes Object Properties Data Properties Annotation Properties Individuals Datatypes Clouds

Entities (2562)

- owl:Thing
-
-
- ' Itulah sebahagian daripada berita-berita ghaib yang Kami wahyukan kepadamu (Muha
- '(Allah) Pencipta langit dan bumi. Apabila Dia henda
- '(Apa yang kamu sedekahkan) adalah untuk orang-orang fakir yang terhalang (usahanya kerana jihad) orang kaya kerana mereka menjaga diri (daripada r minta). Engkau (Muhammad) mengenali mereka da cirinya, mereka tidak meminta secara paksa kepada apa harta yang baik yang kamu sedekahkan, sesun
- '(Dialah) yang menjadikan bumi sebagai hamparan buahan sebagai rezeki untukmu. Oleh itu janganlah sekutu bagi Allah, padahal kamu mengetahui.'
- '(iaitu) beberapa hari tertentu. Maka barang siapa c hari yang lain. Dan bagi orang yang sukar menjalar

Ontologies Classes Object Properties Data Properties Annotation Properties Individuals Datatypes Clouds

ontology: classes (29)

- owl:Thing
- 'Benda Hidup'
- 'Golongan Jin'
- 'Golongan Manusia'
- 'Kumpulan Binatang'
- 'Orang Bersejarah'
- 'Tempat Bersejarah'
- 'Tempat Suci'
- Akhirat
- Ayat
- Bandar
- Binatang
- Burung-burung
- Dunia
- Gunung
- Individu
- Kedudukan
- Laut
- Malaikat
- Manusia
- Nabi
- Neraka
- owl:Nothing
- Penciptaan
- Serangga
- Sungai
- Surah
- Syurga
- Tumbuhan

Figure 8.8: The screenshot of MyQOS OWL Documentation

The OWL documentation is created using Protégé. It presents the elements in the ontology starting from a general class to the list of instances. In addition, in this MyQOS OWL file, we provide two types of languages, namely, Malay and English. However, for search engines, we only focus on the Malay language. We will expand in the future to provide search engines for English as well.

8.4 The limitation of MyQOS prototype

The MyQOS prototype has some limitations. Since the next chapter will show the road map for future works, it requires closer scrutiny because of its potential to improve and develop the real system later.

First, since the user query is based on the ontology we have used to improve semantic search, the user cannot use the queries which are not covered by the ontology. Therefore, the results of the search may be affected. Therefore, as future works, we will add more diverse topics in the knowledge base for the ontological development. These future works will involve the development of the topics or concepts found in the Qur'an.

Secondly, the MyQOS prototype is capable of returning only Malay results and cannot support other languages other than Malay. However, the ontology is created in two languages, namely, Malay and English. We will expand in the future to provide search engines for English as well.

Thirdly, the system performance in terms of speed is not within the scope of this thesis and will not be subject to evaluation. Evaluating the implementation of search engine methods using precision and recall only the measures used in the evaluation stage.

Forth, the semantic search engine needs to be developed. At this time, we only use the SPARQL query to query the contents in the OWL file. To build the search engine, we need to link the SPARQL query with Jena, Apache and Java.

8.5 Summary

The MyQOS prototype system is implemented using Java programming language, Jena framework, Java Server Pages (JSP) and HyperText Markup Language (HTML). This chapter discusses the searching method, which is used in the system is presented. The searching method is based on keyword-based search, question-answering search, and semantic search. This chapter also discusses the limitations of MyQOS prototype.

Part III

Future Work and Conclusion

Chapter 9

Contributions and Future Works

This chapter reviews the work and summarises the research's overall steps to achieve the objectives of this study, the limitations and discusses possible future work in this area.

9.1 Conclusions

The aim of this thesis, as indicated in the beginning, was to solve the limitation of keyword-based models in the Malay translated Qur'an by presented an ontology based IR system with semantics (MyQOS) to facilitate the retrieval of the Malay Translated Qur'an texts. The main research question was can morphological analysis and ontology-based IR with semantics improve the query mechanism for Malay Translated Qur'an. In this thesis,we have focused on the following aspects related to the research question.

- creation of new architecture of morphological analysis for Malay translated Qur'an, Malay translated Qur'an corpus and computational linguistics resources. This includes a new list of Malay Stop words, a new rule-based stemming algorithm, and a new root words annotation,
- representation of information in documents and queries and the mapping of this knowledge into the ontologies,

- improving of the retrieval process by creating an ontology-based IR with semantics search,

9.2 Morphological Analysis

We have discussed about the uniqueness of morphological processes and related research conducted on the Malay language in this thesis. Many of the kinds of processing belong to a kind of popular linguistics, current morphological analysis techniques, and the analysis result of text analysis on the Malay language. However, as we can conclude, state-of-the-art text processing systems for Malay Language are still dealing with problems related to lexical, morphological, and syntax analysis. In these circumstances, the emphasis has been on carrying out the task more effectively, e.g. to automate it rather than do it manually, or make a tagger run faster or increase its success rate, rather than on asking whether the right tasks are being done in the most appropriate way.

This thesis manage to highlight the main challenges in morphological analysis with the creation of new morphological analysis architecture by using Al-Qur'an Amazing book published by *Karya Bestari* [Nursalim et al., 2016] as dataset. The use of Malay translated Qur'an books as main resource of data collection is the best approach rather than using web crawling approach. It reduces the use of mixed languages and dialects. Throughout the preprocessing stage, we created a corpus for Malay language called MyQOS Corpus. The development of Malay translated Qur'an Corpus is an integrated effort involving expert users. The main purpose of MyQOS Corpus in the thesis was for the development of question answer search and concept for ontology development. We used text processing algorithm such as stemming, stopword removal, tokenization, and reduplication algorithm to create the question answering search. The combination of text processing methods in question answering search gives definite improvement at all retrieval points as compared to the keyword only method. This method outperforms previous methods with significant differences. Although the combination appears to be the finest of all, the retrieved and relevant result is still moderate, which is only 60.27%. This implies that that there are still key terms in the Qur'an documents that are not available. The average precision value is the lowest 49.71% due to the enormous number of documents retrieved while using this combination method. The key terms identified are found in the Qur'an documents, but these documents are not listed in the relevance judgment list. This implies that there are still other terms used in the Qur'an that carry the same meaning that have not been retrieved. Thus, to overcome this problem, a semantic search is proposed.

9.3 Ontology Development

The creation on MyQOS using ontology-based IR with semantics using Natural language processing (NLP) is new specifically in Malay translated Qur'an. At present, there is no ontology-based search engine that is being developed for Malay translated Qur'an texts. Most of the search engines were developed using keyword-based search. The MyQOS support semantic retrieval capabilities which can work better than keyword-based search.

The MyQOS ontology focuses mainly on two topics; Location and Living Creation. All topics is extracted from these four sources; thematic index in Al-Qur'an Amazing book [Nursalim et al., 2016], Qur'an Ontology [Hakkoum and Raghay, 2016b], Qurany Concept tools [N H Abbas, 2009], and Ontology of Quranic Concepts [Dukes, 2013]. MyQOS ontology model was implemented on the Protégé and the ontology schema is stored in an OWL file locally on the computer in RDF/XML syntax. Figure 9.1 shows the MyQOS ontology model.

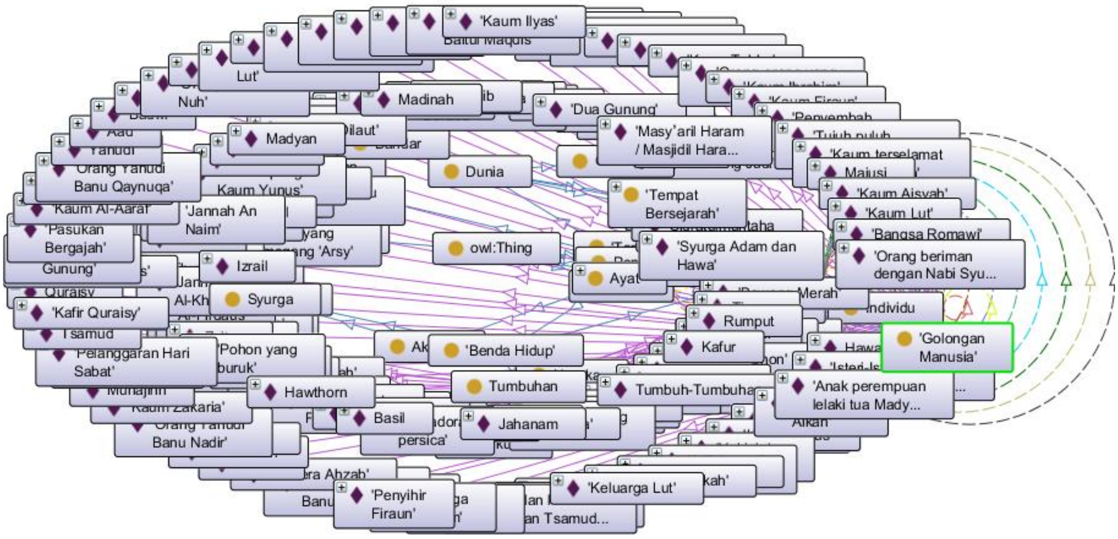


Figure 9.1: The ontology model of MyQOS search engine

The development of MyQOS ontology is to facilitate semantic search. The ontology is a concept that captures knowledge in a widely acceptable standard, and its conceptualization reflects ontology as a notion that identifies entities in the real world

9.4 Semantic search

The basic idea of this thesis is to improve the recall of MyQOS search engine.Search engines retrieve documents related to a specific query term. Generally, the same concepts can be expressed using different terms, thus searching for one of these terms will not retrieve the others. Semantic search improves the recall since the search query will match an entire term instead of only one or more terms. Based on the evaluation results, the average precision and recall for semantic search are 0.8409 (84%) and 0.8043(80%), double the results of the question answer which are 0.4971(50%) for precision and 0.6027 (60%) for recall. The semantic search gives high precision and high recall comparing the other two methods. This indicates that semantic search returns more relevant results than irrelevant ones. The results show that the new ontology-based IR with semantics approach enhanced the precision and recall in all cases. The use of semantic annotation using synonyms gives definite improvement at all retrieval points compared to the keyword method. This method outperforms previous methods with significant differences. It retrieves mostly all relevant documents for every query.

To conclude, this research is among research in the retrieval of the Qur'an texts in the Malay language that managed to outline state-of-the-art information retrieval system models. The results obtained from this analysis indicate that semantic search can help improve the information retrieval method for Malay documents. These findings further support the idea of using semantic annotation improves retrieval quality and accuracy compared to keyword-based search. The results of the current study are consistent with those of [Fernández Sánchez, 2009, Tian, 2012] who said that using semantic annotation can improve retrieval effectiveness and quality. The adoption of this technique in Malay document was very successful and will become a benchmark for other researchers who wish to conduct research on Malay documents.

9.5 Contributions

This research contributes towards the theory and practice of using the ontology of Malay translated Qur'an in information technology. Theoretically, this study adds to the literature and provides insight into the methods used. The research presents an integrated Information System (IS) based on ontology and offers a more systematic approach to Islamic studies. Moreover, the research provides a practical contribution by enabling experts, researchers and readers studying the Qur'an to validate the system. Indeed, the research practically contributes to the evaluation of the relevance of the search and retrieval process related to the verses and knowledge contained in the Qur'an. Additionally, the research ensures that the usage of the Qur'an for searching for related verses through ontology is easier for users.

Secondly, the creation of a prototype of MyQOS, a new ontology-based Information Retrieval (IR) with semantics, is one of the main contributions in this thesis. MyQOS provides a platform that improves the query mechanism to retrieve the information embedded in Malay translated Qur'an. The concepts were adapted and extracted from Malay translated Qur'an Al-Qur'an Amazing book [Nursalim et al., 2016]. The Web Ontology Language (OWL) is the language used to represent the data and SPARQL for querying data. A prototype was implemented using Hypertext Mark-up Language (HTML), Java Server Page (JSP), and Apache Jena.

The other contribution of this thesis is it can offer a new language resource for the Malay translated Qur'an corpus. This resource will help other researchers to build the necessary processing tools for the Malay language. This thesis also develops a question-answer prototype to demonstrate the NLP in processing the text. The prototype preprocesses the Malay translated Qur'an text and an input query using a stemming algorithm and then searches for matches of the query word stem.

The main contributions that have been presented in the thesis are :

- The first Malay translated Qur'an corpus of 149,654 words with root word annotation and root dictionary has grammatical categories, ontology of concepts, word-by-word, English translation and synonym relation.
- A new morphological analysis algorithm for Malay Translated Qur'an. This includes a new list of Malay Stop words, a new rule-based stemming algorithm, and a new root words annotation.
- A new Ontology-based IR with semantics search.

9.6 Limitations and Future Work

Information retrieval relates to assembling knowledge resources that are relevant to an informational need. Searches can be based on metadata, full-text, or other content-based indexing. The ontology-based approach for knowledge retrieval in Malay translated Qur'an has opened some interesting topics for future research in the area of information retrieval. Nevertheless, important research topics still lie ahead, not fully addressed in this thesis or in close relation to the ones we have addressed. In this section, we discuss unsolved limitations, further incremental improvements, as well as new interesting research lines that can be pursued to enhance the current approach:

1. Add more diverse topics in the knowledge base for the ontological development. If the ontology concept does not cover all the aspect of the fields of the study, the results of the search face limitation. This prototype needs to be improved to provide a complete list of concepts and also present precise ontology. These future works will involve the development of all the topics or concepts found in the Qur'an. For this work, this project has received some private grant from one company. They will provide grants to continue to expand the development of this ontology-based IR with the semantic search system.

2. The stemming algorithm needs to be improved. Currently, it only uses tokenization, removing reduplication, removing stopword and stemming. For morphological analysis in Malay especially in stemming, there still a lot of things that need to be done. In this thesis, it relies on Malay root dictionary to maintain the stemming accuracy. Future works will involve introducing additional rules which could eliminate the dictionary dependencies, hence improve the processing speed.

3. There is a limited study conducted to measure the semantic similarity of Malay translated Qur'an text using the Cosine similarity (CS). Measuring the similarity in the text is challenging as it relies heavily on the semantic similarity in meaning. Most of the studies that have been conducted are using synonyms to measure the semantic similarity. However, the standard text similarity measures perform poorly on such tasks especially in handling synonym, especially in the Malay language. This is because the Malay language does not use many synonyms to demonstrate the similarity for a word. It contradicts with the English language. In this thesis, we have surveyed 28 Malaysian seeking to find the list of synonyms of the word that should be used in the development of the ontology. Besides, we also use the Malay thesaurus to get the list of synonym. However, this technique can be improved by using Cosine Similarity (CS) method. Future works will involve creating an automatic semantic similarity measure using the CS method.

4. While most search engine applications allow users to select one factor to rank results, MyQOS prototype also can be improved in the area of ranking the search results according to the user needs. Consequently, the first search results can have a high relevancy than other search results.

5. Rather than use OWL-DL alone, to enhance the expressiveness of the ontology, we can use a set of SWRL (Semantic Web Rule Language) rules. SWRL can then operate over individuals of an OWL-DL ontology. SWRL can be included in OWL ontologies and then exploited by reasoners like Hermit as used in the thesis. SWRL's ability to incorporate user-defined built-in libraries is one of its most useful features. This approach for extending SWRL's expressiveness and boosting the types of information that may be reasoned with using rules is quite strong. This approach can be used to address the problem of data integration, which is one of the Semantic Web's main issues. To solve this problem, a range of mapping technologies must be developed to provide interoperability between the many formats that will be encountered when implementing Semantic Web applications.

This thesis provides the foundation of the Qur'an knowledge representation in the ontology to facilitate the learning of the Qur'an, especially to Malay readers. Therefore, more applications can be developed by using this thesis method to help readers learn and understand the Qur'an in accessible ways, without neglecting the importance of the Qur'an scholars to deliver the truth and accurate knowledge of the Qur'an.

Bibliography

[Abdullah et al., 2009] Abdullah, M. T., Ahmad, F., Mahmod, R., and Sembok, T. M. T. (2009). Rules frequency order stemmer for Malay language. *IJCSNS International Journal of Computer Science and Network Security*, 9(2):433–438.

[Abed, 2015] Abed, Q. A. (2015). *Ontology-based approach for retrieving knowledge in Al-Quran.* PhD thesis, Universiti Utara Malaysia.

[Abu Bakar and Abdul Rahman, 2003] Abu Bakar, Z. and Abdul Rahman, N. (2003). Evaluating the effectiveness of thesaurus and stemming methods in retrieving Malay translated Al-Quran documents. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2911:653–662.

[Adhoni and Al Hamad, 2014] Adhoni, Z. A. and Al Hamad, H. A. (2014). A Cloud Qur'an Application Using Drupal Technology. *International Journal of Web Applications*, 6(1):23–38.

[Ahmad, 1995] Ahmad, F. (1995). *A Malay Language Document Retrieval System: An Experimental Approach And Analysis.* PhD thesis, Universiti Kebangsaan Malaysia (UKM), Malaysia.

[Ahmad et al., 2016] Ahmad, N. D., Bennett, B., and Atwell, E. (2016). Semantic-based Ontology for Malay Qur'an Reader. In *IMAN'2016 4th International Conference on Islamic Applications in Computer Science and Technologies*, page 28.

[Ahmad et al., 2017] Ahmad, N. D., Bennett, B., and Atwell, E. (2017). Retrieval Performance for Malay Quran. *International Journal on Islamic Applications in Computer Science And Technology*, 5(2):13–25.

[Ahmad et al., 2013] Ahmad, O., Hyder, I., Iqbal, R., Murad, m. a. a., Mustapha, A., Sharef, N. M., and Mansoor, M. (2013). A Survey of Searching and Information Extraction on a Classical Text Using Ontology-based semantics modeling: A Case of Quran. *Life Science Journal*, 10(4):1370–1377.

[Ahmed and Gerhard, 2007] Ahmed, Z. and Gerhard, D. (2007). Web to Semantic Web & Role of Ontology. In *National Conference on Information and Communication Technologies (NCICT2007), Baan Pakistan*.

[Al-Omari and Abuata, 2014] Al-Omari, A. and Abuata, B. (2014). Arabic light stemmer (ARS). *Journal of Engineering Science and Technology*, 9(6 (2014)):702–716.

[Al-taani and Al-gharaibeh, 2011] Al-taani, A. T. and Al-gharaibeh, A. M. (2011). Searching Concepts and Keywords in the Holy Quran. In *International Arab Conference on Information Technology*, pages 1–3, Sudan.

[Alagha and Abu-Taha, 2015] Alagha, I. and Abu-Taha, A. (2015). AR2SPARQL : An Arabic Natural Language Interface for the Semantic Web. *International Journal of Computer Applications (0975 –8887)*, 125(6):19–27.

[Alfred et al., 2013] Alfred, R., Mujat, A., and Obit, J. H. (2013). A Ruled-Based Part of Speech (RPOS) Tagger for Malay Text Articles. In *Asian Conference on Intelligent Information and Database Systems*, pages 50–59.

[Almaayah et al., 2014] Almaayah, M., Sawalha, M., and Abushariah, M. A. M. (2014). A Proposed Model for Quranic Arabic WordNet. In *2nd Workshop on Language Resource and Evaluation for religious texts*, pages 9–13.

[Alqahtani and Atwell, 2016] Alqahtani, M. and Atwell, E. (2016). Aligning and Merging Ontology in Al-Quran Domain. In *9th Saudi Students conference in the UK*, pages 8–10, The International Convention Centre, Birmingham.

[Alrehaili and Atwell, 2014] Alrehaili, S. M. and Atwell, E. (2014). Computational ontologies for semantic tagging of the Quran: A survey of past approaches. In *9th Workshop on Language Resources and Evaluation (LREC2014)*, pages 19–23.

[Amsler, 1984] Amsler, R. A. (1984). Lexical knowledge bases. In *ACL '84/COLING '84: Proceedings of the 10th International Conference on Computational Linguistics and 22nd annual meeting on Association for Computational Linguistics*, pages 458–459.

[Arora et al., 2016] Arora, M., Kanjilal, U., and Varshney, D. (2016). Evaluation of information retrieval: precision and recall. *International Journal of Indian Culture and Business Management*, 12(2):224.

[Arthur Taylor, 2009] Arthur Taylor (2009). *Relevance Criterion Choices in Relation to Search Progress.* PhD thesis, The State University of New Jersey.

[Asmah Haji Omar, 2015] Asmah Haji Omar (2015). *Susur galur bahasa Melayu.* Dewan Bahasa dan Pustaka, 2nd edition.

[Auer et al., 2007] Auer, S., Bizer, C., Kobilarov, G., Lehman, J., Cyganiak, R., and Ives, Z. (2007). DBedpia: A Nucleus for a Web of Open Data. In *International Semantic Web Conference (Asian Semantic Web Conference)*, volume 4825, page 722. Springer, Berlin, Heidelberg.

[Baeza-Yates and Ribeiro-Neto, 2011] Baeza-Yates, R. and Ribeiro-Neto, B. (2011). Modern Information Retrieval: The Concepts and Technology behind Search. *Information Retrieval*, 82:944.

[Bakar, 1999] Bakar, Z. A. (1999). *Evaluation of retrieval effectiveness of conflation methods on Malay documents.* PhD thesis, Universiti Kebangsaan Malaysia (UKM) Malaysia.

[Balakrishna et al., 2010] Balakrishna, M., Moldovan, D., Tatu, M., and Olteanu, M. (2010). Semi-automatic domain ontology creation from text resources. In *Proceedings of the 7th International Conference on Language Resources and Evaluation, LREC 2010*, pages 3187–3194.

[Balakrishna and Srikanth, 2008] Balakrishna, M. and Srikanth, M. (2008). Automatic ontology creation from text for National Intelligence Priorities Framework (NIPF). In *Proceedings of 3rd International Ontology for the Intelligence Community (OIC) Conference*, pages 8–12.

[Börner, 2003] Börner, K. (2003). Visual Interfaces for Semantic Information Retrieval and Browsing. In *Visualizing the Semantic Web*, pages 99–115. Springer London.

[Buckley and Voorhees, 2000] Buckley, C. and Voorhees, E. M. (2000). Evaluating evaluation measure stability. *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '00*, 51(2):33–40.

[C. J. van Rijsbergen, 1979] C. J. van Rijsbergen (1979). *Information Retrieval.* Butterworths, 2nd edition.

[Ceri et al., 2013] Ceri, S., Bozzon, A., Brambilla, M., Valle, E. D., Fraternali, P., and Quarteroni, S. (2013). *Web Information Retrieval.* Springer-Verlag Berlin Heidelberg, 1st edition.

[Chowdhury et al., 1983] Chowdhury, G., Salton, G., and McGill, M. J. (1983). *Introduction to Modern Information Retrieval.* McGraw-Hill, Inc. New York, NY, USA.

[Decker et al., 2005] Decker, S., Sintek, M., Billig, A., Henze, N., Harth, A., Leicher, A., Ambite, J.-l., Weathers, M., and Neumann, G. (2005). TRIPLE - an RDF Rule Language with Context and Use Cases. In *W3C Workshop on Rule Languages for Interoperability.*

[Dror et al., 2004] Dror, J., Shaharabani, D., Talmon, R., and Wintner, S. (2004). Morphological Analysis of the Qur'an. *Literary and Linguistic Computing*, 19(4):431–452.

[Dukes, 2013] Dukes, K. (2013). *Statistical Parsing by Machine Learning from a Classical Arabic Treebank.* PhD thesis, University of Leeds, UK.

[Dukes et al., 2010] Dukes, K., Atwell, E., and Sharaf, A.-B. M. (2010). Syntactic Annotation Guidelines for the Quranic Arabic Dependency Treebank. In *International Conference on Language Resources and Evaluation, LREC 2010*, pages 1822–1827.

[Eggebraaten et al., 2014] Eggebraaten, T., Stevens, R., and Will, E. (2014). Natural language processing (NLP).

[Erdmann et al., 2000] Erdmann, M., Maedche, A., Schnurr, H., and Staab, S. (2000). From manual to semi-automatic semantic annotation: About ontology-based text annotation tools. *P. Buitelaar & K. Hasida (eds). Proceedings of the COLING 2000 Workshop on Semantic Annotation and Intelligent Content, Luxembourg.*

[Estival et al., 2004] Estival, D., Nowak, C., and Zschorn, A. (2004). Towards Ontology-based Natural Language Processing. *Proceeedings of the Workshop on NLP and XML (NLPXML-2004): RDF/RDFS and OWL in Language Technology*, pages 59–66.

[Fadzli et al., 2012] Fadzli, S. A., Norsalehen, A. K., Syarilla, I. A., Hasni, H., and Dhalila, M. S. S. (2012). Simple Rules Malay Stemmer. In *The International Conference on Informatics and Applications (ICIA2012)*, pages 28–35.

[Fazzinga and Lukasiewicz, 2010] Fazzinga, B. and Lukasiewicz, T. (2010). Semantic search on the Web. *Semantic Web-Interoperability, Usability, Applicability (SWJ)*, 1:1–7.

[Fernández Sánchez, 2009] Fernández Sánchez, M. (2009). *Semantically enhanced Information Retrieval: an ontology-based approach.* PhD thesis, Universidad Autonoma De Madrid.

[Gruber, 1993] Gruber, T. R. (1993). A translation approach to portable ontologies. *Knowledge Systems Laboratory Technical Report KSL 92-71*, 5(2):199–220.

[Gruninger and Uschold, 1996] Gruninger, M. and Uschold, M. (1996). Ontologies: Principles, Methods and Applications. *Knowledge Engineering Review*, 11(2).

[Habernal, 2012] Habernal, I. (2012). *Semantic Web Search Using Natural Language.* PhD thesis, University of West Bohemia.

[Hakkoum and Raghay, 2016a] Hakkoum, A. and Raghay, S. (2016a). Ontological approach for semantic modeling and querying the Qur ' an. *International Journal on Islamic Applications in Computer Science And Technology*, page 37.

[Hakkoum and Raghay, 2016b] Hakkoum, A. and Raghay, S. (2016b). Semantic Q&A System on the Qur'an. *Arabian Journal for Science and Engineering*, 41(12):5205–5214.

[Hargood et al., 2008] Hargood, C., E.Millard, D., and J.Weal, M. (2008). A thematic approach to emerging narrative structure. In *Proceedings of the hypertext 2008 workshop on Collaboration and collective intelligence*, pages 41–45.

[Hassan, 2002] Hassan, A. (2002). *Tatabahasa bahasa Melayu: morfologi dan sintaksis.* Kuala Lumpur: PTS Publications & Distribution Sdn. Bhd.

[Hersh et al., 1992] Hersh, W. R., Hickam, D. H., and Leone, T. J. (1992). Words, concepts, or both: optimal indexing units for automated information retrieval. In *Proceedings of the 16th Annual Symposium on Computer Applications in Medical Care*, pages 644–648, Baltimore. McGraw-Hill.

[Hitzler et al., 2009] Hitzler, P., Krötzsch, M., and Rudolph, S. (2009). *Foundations of Semantic Web Technologies.* Chapman & Hall/CRC.

[Horridge et al., 2004] Horridge, M., Knublauch, H., Rector, A., Stevens, R., and Wroe, C. (2004). A Practical Guide To Building OWL Ontologies Using The Protege-OWL Plugin and CO-ODE Tools. *The University of Manchester.*

[Ismail et al., 2015] Ismail, R., Abu Bakar, Z., and Abdul Rahman, N. (2015). Extracting knowledge from english translated quran using NLP pattern. *Jurnal Teknologi*, 77(19):67–73.

[Kassim and Rahmany, 2009] Kassim, J. M. and Rahmany, M. (2009). Introduction to semantic search engine. In *Proceedings of the 2009 International Conference on Electrical Engineering and Informatics, ICEEI 2009.*

[Kayed et al., 2008] Kayed, A., Hirzallah, N., Al Shalabi, L. A., and Najjar, M. (2008). Building ontological relationships: A new approach. *Journal of the American Society for Information Science and Technology*, 59(11):1801–1809.

[Khalid et al., 2010] Khalid, H., Baqai, S., Basharat, A., Hassan, A., and Zafar, S. (2010). Leveraging semantic web technologies for standardized knowledge modeling and retrieval from the Holy Qur'an and religious texts. In *7th International Conference on Frontiers of Information Technology*, pages 1–6.

[Kharbat and El-Ghalayini, 2008] Kharbat, F. and El-Ghalayini, H. (2008). Building Ontology from Knowledge Base Systems. In *Data Mining in Medical and Biological Research*.

[Kharkevich, 2010] Kharkevich, U. (2010). *Concept Search : Semantics Enabled Information Retrieval*. PhD thesis, University of Trento.

[Kuck, 2004] Kuck, G. (2004). Tim Berners-Lee's Semantic Web. *SA Journal of Information Management*, 6(1):297.

[Lalmas et al., 2002] Lalmas, M., Cawsey, A., and Roelleke, T. (2002). Seeking Information: Methods from Information Retrieval and Artificial Intelligence.

[Lawson, 2004] Lawson, T. (2004). A Conception of Ontology.

[Manning et al., 2008] Manning, C. D., Raghavan, P., and Schutze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

[Martin et al., 2015] Martin, A., Leon, C., and Lopez, A. (2015). Enhancing semantic interoperability in digital library by applying intelligent techniques. *IntelliSys 2015 - Proceedings of 2015 SAI Intelligent Systems Conference*, pages 904–911.

[Mayfield and Finin, 2003] Mayfield, J. and Finin, T. (2003). Information retrieval on the Semantic Web: Integrating inference and retrieval. *SIGIR Workshop on the Semantic Web*.

[Mitra, 2013] Mitra, M. (2013). Information Retrieval : Models , Techniques , and Evaluation. Technical report, Indian Statistical Institute.

[Mohamed et al., 2011] Mohamed, H., Omar, N., and Aziz, M. J. A. (2011). Statistical Malay Part-of-Speech (POS) Tagger using Hidden Markov Approach.

In *2011 International Conference on Semantic Technology and Information Retrieval*, pages 231–236.

[Mohd Don, 2010] Mohd Don, Z. (2010). Processing Natural Malay texts: A data-driven approach. *Trames*, 14(1):90–103.

[Mohd Pouzi Hamzah, 2006] Mohd Pouzi Hamzah (2006). *Frasa Dan Hubungan Semantik Dalam Perwakilan Pengetahuan: Kesan Terhadap Keberkesanan Capaian Dokumen Melayu.* PhD thesis, Universiti Kebangsaan Malaysia (UKM), Malaysia.

[Moore, 2012] Moore, M. (2012). The semantic web: an introduction for information professionals. *The Indexer: The International Journal of Indexing*, 30(1):38–43.

[Mourtaga et al., 2006] Mourtaga, E., Abdallah, M., Sharieh, A., and Serhan, S. (2006). Quranic based speaker-dependant recognition using triphone/HMM model. *Advances in Modelling and Analysis B.*

[M.Sharaf and Atwell, 2012] M.Sharaf, A.-B. and Atwell, E. (2012). QurSim: A corpus for evaluation of relatedness in short texts. *8th International Conference on Language Resources and Evaluation (LREC'12)*, pages 2295–2302.

[Muhammad, 2012] Muhammad, A. B. (2012). *Annotation of Conceptual Co-reference and Text Mining the Qur'an.* PhD thesis, University of Leeds,UK.

[Mustafa et al., 2008] Mustafa, J., Khan, S., and Latif, K. (2008). Ontology based semantic information retrieval. In *2008 4th International IEEE Conference Intelligent Systems, IS 2008.*

[N H Abbas, 2009] N H Abbas (2009). *Quran 'search for a concept' tool and website.* PhD thesis, University of Leeds,UK.

[Nassourou, 2011] Nassourou, M. (2011). A Knowledge-based Hybrid Statistical Classifier for Reconstructing the Chronology of the Quran. *Webist/Wtm.*

[Nassourou, 2012] Nassourou, M. (2012). Towards a Knowledge-Based Learning System for The Quranic Text.

[Nguyen and Rusin, 2006] Nguyen, N. T. and Rusin, M. (2006). A Consensus-Based Approach for Ontology Integration. In *WI-IATW '06 Proceedings of the 2006 IEEE/WIC/ACM international conference on Web Intelligence and Intelligent Agent Technology*, pages 514–517.

[Noordin and Othman, 2006] Noordin, M. and Othman, R. (2006). An Information Retrieval System for Quranic Texts: A Proposed System Design.

[Nurkhalisah Mustapa, 2013] Nurkhalisah Mustapa (2013). Issues in translation between English and Malay. *Southeast Asia: A Multidisciplinary Journal*, 13:27–34.

[Nursalim et al., 2016] Nursalim, H., Muhtarom, A., Febriadi, S. R., Sanusi, F., Nurhakim, H., Fauzi, H., and Rahman, A. S. (2016). *Al-Quran Amazing*. Karya Bestari, Shah Alam, 6 edition.

[Patel-Schneider, 2005] Patel-Schneider, P. F. (2005). A revised architecture for Semantic Web reasoning. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*.

[Pinto et al., 1999] Pinto, H., Gómez-Pérez, A., and Martins, J. P. (1999). Some issues on ontology integration. *IJCAI-99 workshop on ontologies and problem-solving methods (KRR5)*, (Borst 1997):1–12.

[Popovič, 1991] Popovič, M. (1991). *Implementation of a Slovene language-based free-text retrieval system*. PhD thesis, University of Sheffield, UK.

[Poveda-Villalón et al., 2014] Poveda-Villalón, M., Gómez-Pérez, A., and Suárez-Figueroa, M. C. (2014). OOPS! (OntOlogy Pitfall Scanner!): supporting ontology evaluation on-line. *International Journal on Semantic Web and Information Systems*, 10(2):7–34.

[Raulji and Saini, 2017] Raulji, J. K. and Saini, J. R. (2017). Generating stopword list for sanskrit language. In *Proceedings - 7th IEEE International Advanced Computing Conference, IACC 2017*.

[Rindflesch and Aronson, 1993] Rindflesch, T. C. and Aronson, A. R. (1993). Semantic processing in information retrieval. In *Annual Symposium on Computer Application (SIC) in Medical Care. Symposium on Computer Applications in Medical Care.*

[Rosario, 2000] Rosario, B. (2000). Latent Semantic Indexing : An overview. *Infosys 240.*

[S. Dandagi and Sidnal, 2016] S. Dandagi, V. and Sidnal, N. (2016). Semantic Web (Creating and Querying). *International journal of Web & Semantic Technology*, 7(1):17–27.

[Saad et al., 2009] Saad, S., Salim, N., and Zainal, H. (2009). Pattern extraction for Islamic concept. *Proceedings of the 2009 International Conference on Electrical Engineering and Informatics, ICEEI 2009*, 2(August):333–337.

[Saad et al., 2010] Saad, S., Salim, N., Zainal, H., and Noah, S. A. M. (2010). A framework for Islamic knowledge via ontology representation. In *Proceedings - 2010 International Conference on Information Retrieval and Knowledge Management: Exploring the Invisible World, CAMP'10*, pages 310–314.

[Sakai et al., 1999] Sakai, T., Kitani, T., Ogawa, Y., Ishikawa, T., Kimoto, H., Keshi, I., Toyoura, J., Fukushima, T., Matsui, K., Ueda, Y., Tokunaga, T., Tsuruoka, H., Nakawatase, H., Agata, T., and Kando, N. (1999). BMIR-J2: A test collection for evaluation of Japanese information retrieval systems. *SIGIR Forum (ACM Special Interest Group on Information Retrieval)*, 33(1).

[Schiff, 2011] Schiff, J. (2011). Semantic Web Technologies Effects on Information Retrieval in Digital Libraries : An Annotated Bibliography.

[Schmid, 1999] Schmid, H. (1999). Improvements in Part-of-Speech Tagging with an application to German. In *Natural Language Processing Using Very Large Corpora. Text, Speech and Language Technology.* Springer, Dordrecht.

[Shadbolt et al., 2006] Shadbolt, N., Hall, W., and Berners-Lee, T. (2006). The semantic web revisited. *IEEE Intelligent Systems*, 21(3):96–101.

[Sharaf and Atwell, 2009] Sharaf, A. B. M. and Atwell, E. S. (2009). Knowledge representation of the Quran through frame semantics. In *The Fifth Corpus Linguistics Conference.*

[Sharaf and Atwell, 2012] Sharaf, A. B. M. and Atwell, E. S. (2012). QurSim: A corpus for evaluation of relatedness in short texts. In *Proceedings of the 8th International Conference on Language Resources and Evaluation, LREC 2012.*

[Sharum et al., 2010] Sharum, M. Y., Hamzah, Z. A. Z., Wahab, M. R. A., and Ismail, M. R. (2010). Formal properties and characteristics of Malay rhythmic reduplication. *Procedia - Social and Behavioral Sciences*, 8(5):750–756.

[Sheth et al., 2017] Sheth, K., Bhadka, H., and Jani, A. (2017). Ontology Based Semantic Web Information Retrieval Enhancing Search Significance. *International Journal on Future Revolution in Computer Science & Communication Engineering*, 3(8).

[Shoaib et al., 2009] Shoaib, M., Yasin, M. N., Hikmat Ullah, K., Saeed, M. I., and Khiyal, M. S. H. (2009). Relational WordNet model for semantic search in Holy Quran. *2009 International Conference on Emerging Technologies, ICET 2009*, pages 29–34.

[Studer et al., 1998] Studer, R., Benjamins, V. R., and Fensel, D. (1998). Knowledge Engineering: Principles and methods. *Data and Knowledge Engineering*, 25(1-2):161–197.

[Ta'a et al., 2013] Ta'a, A., Zainal Abidin, S., Abdullah, M. S., Mat Ali, A. B., and Ahmad, M. (2013). Al-Quran Themes Classification Using Ontology. In *4th International Conference on Computing and Informatics, ICOCI 2013*, number 074, pages 383–389.

[Tabrizi and Mahmud, 2013] Tabrizi, A. and Mahmud, R. (2013). Coherence Analysis Issues on English-Translated Quran. *Communications, Signal Processing, and their Applications (ICCSPA)*, pages 1–6.

[Tao et al., 2009] Tao, C., Embley, D. W., and Liddle, S. W. (2009). FOCIH: Form-based ontology creation and information harvesting. *Laender, A.H.F. et al. Conceptual Modeling -ER 2009*, 5829(Springer 2009):346–359.

[Tarawneh and AlShawakfa, 2015] Tarawneh, M. and AlShawakfa, E. (2015). a Hybrid Approach for Indexing and Searching the Holy Quran. *Jordanian Journal of Computers and Information Technology*, 1(1):41.

[Tauberer, 2008] Tauberer, J. (2008). What is RDF and what is it good for?

[Thabet, 2004] Thabet, N. (2004). Stemming the Qur'an. In *Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages*, page 85–88, Geneva, Switzerland. COLING.

[Tian, 2012] Tian, T. (2012). *Using an ontology to improve the web search experience*. PhD thesis, New Jersey Institute of Technology.

[Ulah Khan et al., 2017] Ulah Khan, R., Soraya Mohamad, F., Ulhaq Inam, M., Ahmad Zadi Adruce, S., Nuli Anding, P., Nawaz Khan, S., and Yahya Saleh Al-Hababi, A. (2017). Malay Language Stemmer. *International Journal For Research In Emerging Science and Technology*, 4(12):1–9.

[Ullah Khan et al., 2013] Ullah Khan, H., Muhammad Saqlain, S., Shoaib, M., and Sher, M. (2013). Ontology Based Semantic Search in Holy Quran. *International Journal of Future Computer and Communication*, 2(6):570–575.

[Uthayan and Anandha Mala, 2015] Uthayan, K. R. and Anandha Mala, G. S. (2015). Hybrid Ontology for Semantic Information Retrieval Model Using Keyword Matching Indexing System. *The Scientific World Journal*, 2015:1–9.

[Wahid, 2011] Wahid, FauziahOthman, R. A. (2011). Issues in Evaluating the Retrieval Performance of Multiscript Translation of Al-Quran. International Islamic University Malaysia (IIUM).

[Wynne, 2005] Wynne, M. (2005). *Developing Linguistic Corpora: a Guide to Good Practice*. Oxford: Oxbow Books.

[Xian et al., 2016] Xian, B. C. M., Lubani, M., Ping, L. K., Bouzekri, K., Mahmud, R., and Lukose, D. (2016). Benchmarking Mi-POS: Malay Part-of-Speech Tagger. *International Journal of Knowledge Engineering*, 2(3):115–121.

[Xian-mo, 2007] Xian-mo, Z. (2007). Semantic relationships between contextual synonyms. *US-China education review*, 4(9):452–453.

[Yahya et al., 2013] Yahya, Z., Abdullah, M. T., Azman, A., and Kadir, R. A. (2013). Query translation using concepts similarity based on Quran ontology for cross-language information retrieval. *Journal of Computer Science*, 9(7):889–897.

[Yauri et al., 2012] Yauri, A. R., Kadir, R. a., Azman, A., and Murad, M. A. A. (2012). Quranic-based Concepts: Verse Relations Extraction using Manchester OWL Syntax. In *International Conference on Information Retrieval & Knowledge Management*, number 75, pages 0–1.

[Yauri et al., 2014] Yauri, A. R., Kadir, R. A., Azman, A., and Murad, M. A. A. (2014). Semantic Web Application for Historical Concepts Search in Al-Quran. *International Journal on Islamic Applications in Computer Science And Technology*, 2(2):1–7.

[Yunus et al., 2010a] Yunus, M. A., Zainuddin, R., and Abdullah, N. (2010a). Semantic query with stemmer for quran documents results. *2010 IEEE Conference on Open Systems (ICOS 2010)*, pages 40–44.

[Yunus et al., 2010b] Yunus, M. A., Zainuddin, R., and Abdullah, N. (2010b). Visualizing quran documents results by stemming semantic speech query. *Proceedings - 2010 International Conference on User Science and Engineering, i-USEr 2010*, pages 209–213.

[Yusof et al., 2011] Yusof, R. J. R., Zainuddin, R., and Yusoff, Z. M. (2011). Learning Methods and Problems of Qur ' an Reciters (Malays and Africans). *Centre of Quranic Research international journal*, pages 17–38.

[Zoghi et al., 2014] Zoghi, M., Whiteson, S. A., De Rijke, M., and Munos, R. (2014). Relative confidence sampling for efficient on-line ranker evaluation. In

WSDM 2014 - Proceedings of the 7th ACM International Conference on Web Search and Data Mining, pages 73–82.

Appendix A

List of Malay Stopword

ada
adakah
adakan
adalah
adanya
adapun
agak
agar
akan
aku
akulah
akupun
al
alangkah
amat
antara
antaramu
antaranya
apa
apa-apa
apabila
apakah
apapun
atas
atasmu
atasnya
atau
ataukah
ataupun
bagaimana
bagaimanakah
bagi
bagimu
baginya
bahawa
bahawasanya
bahkan
bahwa
banyak
banyaknya
barangsiapa
bawah
beberapa
begitu
begitupun
belaka
belum
belumkah
berada
berapa
berikan
beriman
berkenaan
berupa
beserta
biarpun
bila
bilakah
bilamana
bisa
boleh
bukan
bukankah
bukanlah
dahulu
dalam
dalamnya
dan
dapat
dapati
dapatkah
dapatlah
dari
daripada
daripadaku
daripadamu
daripadanya
demi
demikian
demikianlah
dengan
dengannya
di
dia
dialah
didapat
didapati
dimanakah
engkau
engkaukah
engkaulah
engkaupun
hai
hampir
hampir-hampir
hanya
hanyalah
hendak
hendaklah
hingga

ia
iaitu
ialah
ianya
inginkah
ini
inikah
inilah
itu
itukah
itulah
jadi
jangan
janganlah
jika
jikalau
jua
juapun
juga
kalau
kami
kamikah
kamipun
kamu
kamukah
kamupun
katakan
ke
kecuali
kelak
kembali
kemudian
kepada
kepadaku
kepadakulah
kepadamu
kepadanya
kepadanyalah
kerana
kerananya
kesan
ketika
kini
kita
ku
kurang
lagi
lain
lalu
lamanya
langsung
lebih
maha
mahu
mahukah
mahupun
maka
malah
mana
manakah
manapun
masih
masing
masing-masing
melainkan
memang
mempunyai
mendapat
mendapati
mendapatkan
mengadakan
mengapa
mengapakah
mengenai
menjadi
menyebabkan
menyebabkannya
mereka
merekalah
merekapun
meskipun
mu
nescaya
niscaya
nya
olah
oleh
orang
pada
padahal
padamu
padanya
paling
para
pasti
patut
patutkah
per
pergilah
perkara
perkaranya
perlu
pernah
pertama
pula
pun
sahaja
saja
saling
sama
sama-sama
samakah
sambil
sampai
sana
sangat
sangatlah
saya
se
seandainya
sebab
sebagai
sebagaimana
sebanyak
sebelum
sebelummu
sebelumnya
sebenarnya
secara
sedang
sedangkan
sedikit
sedikitpun
segala
sehingga
sejak
sekalian
sekalipun
sekarang
sekitar

selain
selalu
selama
selama-lamanya
seluruh
seluruhnya
sementara
semua
semuanya
semula
senantiasa
sendiri
sentiasa
seolah
seolah-olah
seorangpun
separuh
sepatutnya
seperti
seraya
sering
serta
seseorang
sesiapa
sesuatu
sesudah
sesudahnya
sesungguhnya
sesungguhnyakah
setelah
setiap
siapa
siapakah
sini
situ
situlah
suatu
sudah
sudahkah
sungguh
sungguhpun
supaya
tadinya
tahukah
tak
tanpa
tanya
tanyakanlah
tapi
telah
tentang
tentu
terdapat
terhadap
terhadapmu
termasuk
terpaksa
tertentu
tetapi
tiada
tiadakah
tiadalah
tiap
tiap-tiap
tidak
tidakkah
tidaklah
turut
untuk
untukmu
wahai
walau
walaupun
ya
yaini
yaitu
yakni
yang

Appendix B

List of Malay Root word

Aad
abadi
abai
abdi
Abdullah
abu
Abu
ada
adab
Adam
adas
adat
adil
adn
adu
aduk
agama
agung
ahli
ahlulbait
Ahmad
Aikah
air
ais
aisyah
ajaib
ajak
ajal
ajam
ajar
akad
akal
akar
akbar
akhir
akhirat
akhlak
akibat
akrab
akui
alah
alam
alami
alasan
alat
alih
alim
alir
Allah
Al-lata
amal
aman
amanah
amanat
amarah
amaran
ambil
ambing
amil
amin
ampun
Amri
anai-anai
anak
ancam
aneh
aneka
angan
anggap
anggun
anggur
angin
angkasa
angkat
angkuh
aniaya
anjing
ansar
ansur
anugerah
anut
anyaman
api
apung
aqsa
arab
A'raf
arafah
arah
arak
arsy
arus
asa
asak
asal
asap
asas
Asi
asing
Asmaul husna

asuh
atap
Atiq
atur
aurat
awal
awan
awas
ayah
ayat
Ayub
azab
azan
Azar
Aziz
Azlam
babi
Babilonia
baca
badai
badan
Badar
Badwi
baghal
baginda
bagus
bah
bahagi
bahagia
bahan
bahasa
bahaya
bahirah
bahtera
baik
bait
Baitul
Baitullah
Baitulmaqdis
bajak
baju
bakal
bakar
bakhil
baVBah
bakti
bala
balah
balas
baligh
balik
Balqis
balut
bandar
banding
bangka
bangkai
bangkang
bangkit
bangsa
bangun
bangunan
Bani
banjir
bantah
bantal
bantu
banyak
bapa
bara
barang
barangkali
barat
baring
baris
baru
barzakh
basah
basuh
bata
batal
batang
batas
batin
batu
bau
baur
bawa
bawang
baya
bayang
bayar
bayi
bazir
beban
bebas
begini
begitu
bekal
bekas
bela
belah
belajar
belakang
belalai
belalak
belalang
belang
belanja
belas
belayar
belenggu
beli
beliak
belit
belulang
belum
benam
benang
benar
bencana
benci
benda
bendahara
bendera
benderang
bengis
bengkok
benih
bentak
bentang
bentar
benteng
bentuk
berai
berani
berapa
berat
berenang
beri

berita
beritahu
berkah
berkat
bersih
besan
besar
besi
betapa
betina
betis
betul
beza
biak
biar
biara
biasa
bibir
bicara
bidadari
bidang
bidara
bijak
bijaksana
biji
bilah
bilang
bimbang
bin
bina
binasa
binatang
bincang
bingung
bintang
binti
biru
bisa
bisik
bisu
bodoh
bohong
bolak
boleh
bondong
bongkak
bongkar
bongkok
bosan
buah
buai
bual
buang
buas
buat
budi
buih
bujang
buka
bukit
bukti
buku
bulan
bulat
bulu
bumbung
bumi
bunga
bungkus
bunting
bunuh
Bunyamin
bunyi
buru
buruk
burung
busur
buta
butir
cabang
cabar
cabut
cacat
caci
cahaya
cair
cakap
calon
campak
campur
cantik
cantum
capai
cara
cari
catat
cawan
cebur
cecah
ceduk
cegah
cekik
cela
celah
celaka
celik
cemar
cemas
cemerlang
cemeti
cemuh
cenderung
cengang
cepat
cerai
cerca
cerdas
ceria
cerita
ceroboh
cincang
cinta
cipta
ciri
cita
cium
condong
contoh
cuba
cucu
cucuk
cucur
cukup
cukur
culik
cuma
curah
curang
curi
curiga

dada
dadak
daerah
dagang
daging
dagu
dahaga
dahi
dahsyat
dahulu
daki
dakwa
dakwah
dalam
dalih
dam
damai
damping
dangkal
dapat
dapur
darah
darat
dari
darjat
darurat
darussalam
dasar
datang
datar
datuk
Daud
daun
daya
debat
debu
dedah
dekap
dekat
delima
demikian
denda
dendam
dengan
dengar
dengki
depan
derai
deras
derhaka
derita
derma
desa
desak
desis
dewasa
Dhuha
diam
diat
didih
didik
din
dinar
dinding
dingin
dirham
diri
doa
dongak
dongeng
dorong
dosa
dua
duduk
duga
duka
dukacita
dulu
dunia
durhaka
duri
dusta
dusun
dzun
edar
ejek
ekor
elak
elok
emas
embun
empat
empuk
enam
endah
enggan
erti
Esa
esok
faedah
faham
faham
fajar
fakir
fasih
fasik
fatamorgana
fatwa
fidyah
fikir
Fir'aun
firdaus
firman
fitnah
fitrah
fizikal
fulan
furqan
gabung
gadis
gagah
gagak
gagal
gajah
gala
galah
gamak
gambar
ganang
ganas
ganda
gandum
ganggu
ganjaran
ganjil
ganti
garis
gaul
gegas
gejolak

gelang
gelap
gelar
gelas
gelek
geleng
gelimpang
gelincir
gelisah
gelita
gelombang
gelora
gelumang
gelupur
gema
gembala
gembira
gementar
gempa
gempar
gemuk
gemuruh
genap
gendong
generasi
genggam
gentar
gerak
gerbang
gereja
gesa
gesit
getar
ghaib
giat
gigi
gigit
gila
gilir
goda
golong
goncang
gua
gudang
gugur
gugus
gulung
gumpal
guna
guni
gunung
gurau
guruh
Habil
habis
had
hadap
hadas
hadiah
hadir
hadrat
hadyu
hafsah
haid
hairan
haiwan
hajat
haji
hak
hakikat
hakim
hal
hala
halal
halang
halau
halia
halilintar
halus
Haman
hamba
hambat
hambur
hamil
hampar
hancur
hangat
hangus
hanif
hantar
hanyut
hapus
hara
haram
harap
harga
hari
Harithah
harta
haru
harum
harun
harus
harut
hasad
hasil
hasta
hasut
hati
haus
hawa
Hawa
hawariyyun
hawiyah
hayat
hebat
helai
hembus
hendak
hentam
henti
herdik
heret
hias
hibur
hidang
hidap
hidayah
hidung
hidup
hijau
hijr
hijrah
hikmah
hikmat
hilang
himpun
hina
hindar
hingga

hirau
hisab
hitam
hitung
homoseksual
hormat
hubung
hud
hudaibiyah
hujan
hujung
hukum
hulur
hunain
huni
huru
huruf
hurumat
husna
hutamah
hutang
ibadah
ibadat
ibarat
iblis
ibni
ibnu
ibrahim
ibu
idah
idris
ifrit
ihram
ikan
ikat
ikhlas
ikrar
iktikad
iktikaf
ikut
ilham
Illiyyun
ilmu
Ilyas
Ilyasa
iman
Imran
inasan
indah
infak
ingat
ingin
ingkar
injak
injil
insaf
intai
intip
Iram
iri
iring
Isa
Ishak
isi
islam
Ismail
Israel
Israil
istana
isteri
istimewa
istiqamah
istirehat
isu
isyak
isyarat
isytihar
izin
jabatan
jadi
jaga
jahanam
jahat
jahil
jahiliyah
jahim
jalan
jalar
jalin
jalur
jalut
jamah
jamin
jamu
janda
janggut
jangka
janin
janji
jantan
jantung
jarak
jari
jarum
jasa
jasad
jatuh
jauh
jawab
jawat
jaya
jejak
jejas
jelajah
jelang
jelas
jelek
jelita
jemari
jemput
jemu
jengkel
jenis
jerit
jernih
jerumus
jibril
Jibt dan Taghut
jihad
jijik
jilbab
jin
jinak
jiran
jitu
jiwa
jizyah
jua
jual
juang

judi
juga
jujur
julai
julang
julur
jumaat
jumlah
jumpa
junjung
junub
jurang
juru
justeru
kaabah
kabul
kabur
kabut
kaca
kacang
kacau
kadangkala
kadar
kafarah
kafilah
kafir
kafur
kagum
kahwin
kain
kait
kaki
kaku
kala
kalah
kalalah
kalangan
kalau
kali
kali
kalian
kalimah
kalimat
kalong
kalung
Kamal
kamar
kambing
kami
kampung
kamu
kanak
kanan
kandang
kandas
kandung
kapal
karib
karung
kasar
kasih
kasturi
kata
katak
kau
kaum
kawal
kawan
kawasan
kaya
kayu
keadaan
kebajikan
kebun
kecewa
kecil
kecuali
kedekut
kediaman
kedip
kejam
kejap
kejar
keji
kejut
kekal
kelahi
kelakuan
keldai
keliling
kelip
keliru
kelmarin
kelompok
kelopak
keluar
keluarga
keluh
kemalangan
kemaluan
kemarau
kemaslahatan
kembali
kembang
kembara
kemudian
kena
kenal
kenan
kenang
kencang
kendali
kenderaan
kendi
kening
kenyataan
kepala
keping
kepit
kepung
kera
kerabat
kerah
kerajaan
kerap
keras
kerdip
kerikil
kering
kerja
kerongkong
kertas
kerumun
kerut
kesah
kesal
kesan
ketat
ketawa
ketiak

ketika
ketua
keturunan
khabar
khalifah
khamar
khas
khawatir
khazanah
khemah
khianat
khidir
khuatir
khuldi
khusus
khusyuk
khutbah
kiamat
kiasan
kibas
kiblat
kifarat
kilas
kilat
kilau
kira
kiri
kirim
kisah
kita
kitab
kitar
kobar
kolam
kongsi
kontang
kontang-kanting
korban
kosong
kota
kotor
koyak
kristal
kuasa
kuat
kubur
kuda
kufur
kuku
kukuh
kumpul
kunang
kunci
kuning
kunjung
kupu
kurang
kurma
kurnia
kursi
kurun
kurung
kurus
kusta
kutip
kutu
kutuk
labah
labu
labuh
lacur
ladang
lafaz
laga
Lahab
lahir
lailatul
lain
laju
laknat
laksana
laku
lalai
lalat
lali
lalu
lama
laman
lambat
lambung
lampau
lampias
lancar
landa
langgar
langit
langkah
lanjut
lantai
lantar
lapan
lapang
lapar
lapis
larang
larat
lari
lata
latih
lauh-lauh
laut
lawan
layak
layan
layang
layar
lazat
lbn
lebah
lebar
lebat
lebih
lebur
lecur
ledak
lega
leher
leka
lekat
lelah
lelaki
lemah
lemak
lembah
lembaran
lembing
lembu
lembut
lempar
lenang

lengan
lengkap
lenguh
lenyap
lepas
lereng
lesbian
lesu
letak
letih
liar
liat
libat
licin
lidah
lihat
lima
limpah
lindung
lingkar
lingkung
lintas
lipat
liput
lisan
litup
lngat
lni
logam
lompat
lontar
ltu
luap
luar
luas
lubang
luh
luka
luluh
lumat
lumba
lumpuh
lumpur
lumur
lunak
luncur
lupa
luput
luqman
lurus
lut
lutut
maaf
mabuk
macam
madinah
madu
madyan
Maha Suci
mahar
maharaja
maharajalela
mahfuz
mahir
mahjura
mahsyar
mahu
main
majlis
maju
majusi
makam
makan
makhluk
maki
makin
makjuj
maVBah
maklum
makmur
maNNa
makruf
maksiat
maksud
malaikat
malam
malang
malapetaka
malas
malu
mampu
mana
Manat
mandi
mandul
manfaat
mangsa
mani
manis
manna
manusia
maqdis
marah
mari
marjan
martabat
marut
marwah
Maryam
mas
masa
masak
masam
masin
masjid
masjidil
masuk
Masy'aril
mata
matahari
mati
maut
mawar
mayang
mayat
medan
megah
mekah
memang
memphis
menang
menantu
mendung
mengaku
merah
merdeka
mertua
mesir
mesra
mesti

mesyuarat
mewah
mihrab
mikail
milik
mimbar
mimpi
mina
minta
minum
minyak
misal
miskin
moga
mohon
moyang
mualaf
muat
mubahalah
muda
mudah
mudarat
muhajirin
muhammad
muhkamat
muka
mukim
mukjizat
mukmin
mukminin
mula
mulia
mulut
munafik
munajat
mundur
mungkar
mungkin
mungkir
muntah
murah
muram
murka
murni
murtad
musa
musafir
musaharah
musibah
musim
muslihat
muslim
muslimin
musnah
mustahil
musuh
musyrik
musyrikin
mut'ah
mutasyabihat
mutiara
mutlak
nabi
nada
nafas
nafkah
nafsu
nahas
naik
najis
najran
nama
nampak
namun
nanah
nanti
nasab
nasib
nasihat
Nasr
nasrani
nasuha
naung
nazar
negeri
nenek
neraca
neraka
ngantuk
ngeri
niaga
niat
nikah
nil
nilai
nipis
nuh
nun
nurani
nusyuz
nyala
nyaman
nyamuk
nyaris
nyata
nyawa
olah
oleh
olok
ombak
orang
orbit
pada
padahal
padam
padan
padang
padi
padu
pagi
paha
pahala
pahat
pahit
pajak
pakai
pakat
paksa
paku
palestin
paling
palsu
paluan
panah
panas
pancang
pancar
pancut
pandai
pandang

panggang
panggil
pangkal
panik
panjang
panjat
pantas
papan
para
parit
pasak
pasang
pasar
pasir
pasrah
pasti
pasukan
patah
pati
patuh
patung
patut
payah
payau
pecah
pecut
pedih
pedoman
peduli
pegang
pejam
pekak
pekat
pekik
pelamin
pelepah
pelihara
pelik
pelita
peluang
peluk
pena
penat
pencil
pendam
pendek
pendeta
pengaruh
pengetahuan
pengsan
pening
penjara
penting
penuh
perahu
perak
peranan
perang
perangai
peranjat
peras
perasaan
perawan
percaya
percik
perempuan
pergi
peri
perihal
periksa
perinci
perintah
perisai
peristiwa
periuk
perkakas
perkasa
perlahan
perlu
permaidani
permata
pernah
perut
pesan
pesona
pesong
petala
petang
petani
peti
petik
petir
piala
pihak
pijak
pikat
pikul
pilih
pimpin
pinak
pinang
pindah
pinggan
pinggir
pinjam
pintal
pintar
pintu
piring
pisah
pisang
pisau
piutang
pohon
pokok
potong
prasangka
puak
puas
puasa
pudar
puisi
puja
puji
puji
pujuk
pukau
pukul
pula
pulang
puluh
puncak
punggung
pungut
punya
purnama
pusaka
pusing
putar
putera

puteri
putih
putus
qabil
qadar
qaf
qalaid
qarun
qasar
qisas
quraisy
quraizah
quran
qurban
quru
Ra'ina
raba
ragam
ragu
rahbaniyyah
rahib
rahim
rahman
rahmat
rahsia
raja
rakaat
ramadan
ramai
rambut
rampas
rancang
rangkak
rangkul
rantai
ranting
rapat
rapi
rapuh
rasa
rasmi
rasul
rasulullah
rata
ratus
raya
rayap
rayu
rebah
rebut
reda
redha
redup
rehat
rejam
reka
rela
remeh
rencana
rendah
rendang
rentak
rentang
renung
reput
resap
retak
rezeki
ria
riak
riang
riba
ribu
ribut
ringan
rintang
rintih
risalah
risau
riuh
roboh
roh
romawi
rombong
rompak
rongga
rosak
rotan
roti
ruah
ruang
rugi
Ruhul qudus
rujuk
rukuk
rumah
rumput
runding
runtuh
rupa
Sa'ibah
saad
saat
saba
sabar
Sabat
sabiin
sabil
sabit
sabtu
sabut
sadar
saf
safa
sah
sahabat
sahaja
sahaya
saji
sakaratul
sakit
saksi
salah
salam
salamun
saleh
salib
salin
salsabil
salwa
sama
sambar
sambung
sambut
samiri
sampah
sampai
samping
samud
samun

sandang
sandar
sangat
sanggah
sanggup
sangka
sangkakala
sangkal
santun
sapa
sapi
sapih
sapu
saqar
saran
sarang
sari
sasaran
satu
saudara
sawah
sawi
sayang
sayap
sayur
sebab
sebar
sebat
seberang
sebut
sedap
sedar
sedekah
sederhana
sedia
sedih
sedikit
sedu
segala
segar
segera
seimbang
sejahtera
sejuk
sekaligus
sekat
seksa
sekutu
selagi
selamat
selang
selaput
selar
selawat
selera
selesai
selimut
selisih
seludang
seluruh
semakin
semangat
semasa
semayam
sembah
sembahyang
sembelih
sembilan
sembuh
sembunyi
sembur
semoga
sempadan
sempat
sempit
sempurna
semua
semut
senang
senda
sendi
sendiri
sengaja
sengsara
senja
senjata
sentosa
sentuh
senyum
sepakat
sepatu
seperti
sepuluh
serah
serak
serang
serasi
serba
serbu
seret
seri
serigala
sering
serta
seru
sesak
sesal
sesat
sesuai
sesuatu
setara
setelah
setia
setuju
siang
siap
siapa
siar
siasat
sia-sia
sibghah
sibuk
sidr
sidratulmuntaha
sifat
sihat
sihir
sikap
siku
sila
silang
silih
simpan
simpang
simpul
Sinai
sinar
singa
singgahsana
singkap

singkir
singsing
sini
siram
sisa
sisi
sisih
siul
siuman
soal
sodom
sokong
solat
soleh
solehah
sombong
sorak
strategi
suai
suami
suara
suasana
subuh
subur
suci
sudah
sufyan
sujud
suka
sukar
sukarela
sukat
suku
sulaiman
sulbi
sulit
sumbat
sumber
sumpah
sungai
sungguh
sungkur
sunnah
sunyi
surah
surai
suram
surat
suruh
surut
susah
susu
susun
sutera
Suwa’
syafaat
syahid
syahwat
syair
syaitan
syak
syam
syarak
syariat
syeikh
syiar
Syi’ra
syirik
syuhada
syukur
syurga
taat
tabiat
tabir
tabung
tabur
Tabut
tadbir
tadi
Taghut
tagih
tahajud
tahan
tahap
tahu
tahun
tajam
takdir
takjub
takluk
takut
takwa
takwil
talak
tali
tamak
taman
tambah
tampak
tampil
tamu
tanah
tanam
tanda
tandan
tanding
tanduk
tandus
tangan
tangga
tanggal
tangguh
tanggung
tanggungjawab
tangis
tangkai
tangkap
tangkas
tanya
tar
taraf
tarik
tartil
taruh
tasbih
tasik
tasnim
tatah
tatang
tatkala
taubat
taufan
taufik
tauhid
taurat
taut
tawa
tawaf
tawakal
tawan

tawar
tayamum
tebal
tebang
tebar
tebuk
tebus
teduh
tegak
tegap
tegas
teguh
teguk
tegur
teka
tekad
teki
tekun
teladan
telaga
telah
telan
telanjang
telapak
telinga
telingkah
teliti
telungkup
telur
teman
tembaga
tembikar
tembok
tembus
tempang
tempat
tempoh
tempuh
tempur
temu
temurun
tenang
tengah
tenggelam
tenggorokan
tengkar
tengkuk
tentang
tentera
tenteram
tentu
tenung
tepat
tepi
tepuk
terang
terbang
terbit
teriak
terik
terima
terjun
terka
terkam
ternak
tertib
terus
tetap
Thaif
Thalut
Thur
tiang
tiba
tidak
tidur
tiga
tilam
tilik
timba
timbang
timbul
timbun
timpa
timpal
timun
timur
tin
tindak
tindas
tinggal
tinggi
tingkah
tingkat
tinjau
tinta
tipu
tiri
tiru
titah
titis
tiup
tolak
toleh
tolong
tompok
tongkat
tsamud
tua
tuai
tuan
tuang
Tubba’
tubi
tubuh
tubuh
tuduh
tudung
tugas
tuhan
tuju
tujuh
tukang
tulang
tuli
tulis
tulus
tumbang
tumbuh
tumbuk
tumpah
tumpang
tumpas
tumpu
tunai
tunas
tunda
tunduk
tunggang
tunggu
tungku

tunjuk
tuntut
Tursina
turun
turut
tutup
tutur
tuwa
ubah
uban
ubat
ubun
ucap
udara
uhud
uji
ukur
ulama
ulang
ular
ulat
umat
umbi
ummi
ummiyyin
ummul
Ummul qura
umpama
umpat
umrah
umum
umur
undang
undi
undur
unggun
ungkap
unsur
unta
untuk
untung
unzurna
upah
upaya
urai
urat
urus
usaha
usap
usia
usir
usul
usus
utama
utara
utus
uzair
uzur
uzza
Wadd
wafat
wahyu
wajah
wajar
wajib
waktu
walhal
wali
wang
waris
warna
wasiat
wasilah
waspada
wazir
wenang
wujud
wusta
yaakub
Yagus
yahudi
yahya
yakin
yakjuj
yala
yaman
yathrib
yatim
Ya'uq
yunus
yusuf
zabaniyah
zabur
zahir
zaid
zainab
zaitun
zakaria
zakat
zalim
zaman
zaqqum
zarah
zihar
zikir
Az-zikr
zina
zohor
zubur
zulkarnain
zulkifli
zuriat

Appendix C

List of Root word Tagging

WORD	LANGUAGE	TAG	WORD	LANGUAGE	TAG
aad	Arabic	NN	ahli	Malay	NN
abadi	Malay	ADJ	ahlulbait	Arabic	NN
abai	Malay	VB	Ahmad	Arabic	NN
abdi	Malay	NN	aib	Malay	ADJ
Abdullah	Arabic	NN	Aikah	Arabic	NN
abu	Malay	NN	ain	Arabic	NN
Abu	Arabic	NN	air	Malay	NN
ada	Malay	VB	ais	Malay	NN
adab	Malay	NN	aisyah	Arabic	NN
Adam	Arabic	NN	ajaib	Malay	ADJ
adas	Arabic	NN	ajak	Malay	VB
adat	Malay	NN	ajal	Malay	VB
adil	Malay	ADJ	ajam	Arabic	NN
adn	Arabic	NN	ajar	Malay	NN
adu	Malay	VB	akad	Malay	VB
aduan	Malay	NN	akal	Malay	NN
aduk	Malay	VB	akan	Malay	FT
agama	Malay	NN	akar	Malay	NN
agung	Malay	ADJ	akbar	Malay	NN

WORD	LANGUAGE	TAG	WORD	LANGUAGE	TAG
akhirat	Malay	NN	ampun	Malay	NN
akhlak	Malay	NN	Amri	Arabic	NN
akibat	Malay	FT	anai-anai	Malay	NN
akrab	Malay	ADJ	anak	Malay	NN
aku	Malay	NN	ancam	Malay	VB
akui	Malay	VB	andai	Malay	VB
alah	Malay	ADJ	aneh	Malay	ADJ
alam	Malay	NN	aneka	Malay	FT
alami	Malay	VB	angan	Malay	NN
alasan	Malay	NN	anggap	Malay	VB
alat	Malay	NN	anggun	Malay	ADJ
alif	Arabic	NN	anggur	Malay	NN
alih	Malay	VB	angin	Malay	NN
alim	Malay	ADJ	angkasa	Malay	NN
alir	Malay	VB	angkat	Malay	ADJ
Allah	Arabic	NN	angkuh	Malay	ADJ
Al-lata	Arabic	NN	aniaya	Malay	VB
amal	Malay	NN	anjing	Malay	NN
aman	Malay	ADJ	ansar	Arabic	NN
amanah	Malay	ADJ	ansur	Malay	ADJ
amanat	Malay	NN	antara	Malay	FT
amarah	Malay	NN	anugerah	Malay	NN
amaran	Malay	NN	anut	Malay	VB
amat	Malay	FT	anyaman	Malay	NN
ambil	Malay	VB	apa	Malay	FT
ambing	Malay	NN	api	Malay	NN
amil	Arabic	NN	apung	Malay	NN
amin	Arabic	VB	aqsa	Arabic	NN

WORD	LANGUAGE	TAG	WORD	LANGUAGE	TAG
arab	Malay	NN	azab	Malay	NN
A'raf	Arabic	NN	azan	Malay	NN
arafah	Arabic	NN	Azar	Arabic	NN
arah	Malay	NN	Aziz	Arabic	NN
arak	Malay	NN	Azlam	Arabic	NN
aril	Arabic	NN	babi	Malay	NN
arsy	Arabic	NN	Babilonia	Arabic	NN
arus	Malay	NN	baca	Malay	VB
asa	Malay	NN	badai	Malay	NN
asak	Malay	VB	badan	Malay	NN
asal	Malay	NN	Badar	Arabic	NN
asap	Malay	NN	Badwi	Arabic	NN
asas	Malay	ADJ	bagai	Malay	FT
Asi	Arabic	NN	baghal	Arabic	NN
asing	Malay	ADJ	bagi	Malay	VB
Asmaul husna	Arabic	NN	baginda	Malay	BD
asuh	Malay	VB	bagus	Malay	ADJ
atap	Malay	NN	bah	Malay	NN
atas	Malay	FT	bahagi	Malay	VB
Atiq	Arabic	NN	bahagia	Malay	ADJ
atur	Malay	VB	bahagian	Malay	NN
aurat	Malay	NN	bahan	Malay	NN
awal	Malay	NN	bahasa	Malay	NN
awan	Malay	NN	bahaya	Malay	NN
awas	Malay	VB	bahirah	Arabic	NN
ayah	Malay	NN	bahtera	Malay	NN
ayat	Malay	NN	baik	Malay	ADJ
Ayub	Arabic	NN	bait	Malay	NN

WORD	LANGUAGE	TAG	WORD	LANGUAGE	TAG
Baitul	Arabic	NN	bantah	Malay	VB
Baitullah	Arabic	NN	bantal	Malay	NN
Baitulmaqdis	Arabic	NN	bantu	Malay	VB
bajak	Malay	NN	banyak	Malay	ADJ
baju	Malay	NN	bapa	Malay	NN
bakal	Malay	FT	bara	Malay	NN
bakar	Malay	VB	barang	Malay	NN
bakhil	Malay	ADJ	barangkali	Malay	NN
baVBah	Arabic	NN	barat	Malay	NN
bakti	Malay	NN	baring	Malay	VB
bala	Malay	NN	baris	Malay	NN
balah	Malay	VB	baru	Malay	ADJ
balas	Malay	VB	barzakh	Arabic	NN
baligh	Arabic	NN	basah	Malay	ADJ
balik	Malay	VB	basuh	Malay	VB
Balqis	Arabic	NN	bata	Malay	NN
balut	Malay	NN	batal	Malay	NN
bandar	Malay	NN	batang	Malay	NN
banding	Malay	ADJ	batas	Malay	NN
bangka	Malay	ADJ	batin	Malay	NN
bangkai	Malay	NN	batu	Malay	NN
bangkang	Malay	VB	bau	Malay	NN
bangkit	Malay	VB	baur	Malay	VB
bangsa	Malay	NN	bawa	Malay	VB
bangun	Malay	VB	bawang	Malay	NN
bangunan	Malay	NN	baya	Malay	ADJ
Bani	Arabic	NN	bayang	Malay	NN
banjir	Malay	NN	bayar	Malay	VB

WORD	LANGUAGE	TAG	WORD	LANGUAGE	TAG
bayi	Malay	NN	bencana	Malay	NN
bazir	Malay	VB	benci	Malay	VB
beban	Malay	NN	benda	Malay	NN
bebas	Malay	ADJ	bendahara	Malay	NN
begini	Malay	ADJ	bendera	Malay	NN
begitu	Malay	ADJ	benderang	Malay	ADJ
bekal	Malay	VB	bengis	Malay	ADJ
bekas	Malay	NN	bengkok	Malay	ADJ
bela	Malay	VB	benih	Malay	NN
belah	Malay	NN	bentak	Malay	ADJ
belajar	Malay	NN	bentang	Malay	VB
belakang	Malay	VB	bentar	Malay	ADJ
belalai	Malay	NN	benteng	Malay	NN
belalak	Malay	VB	bentuk	Malay	NN
belalang	Malay	NN	berai	Malay	VB
belang	Malay	NN	berani	Malay	ADJ
belanja	Malay	VB	berapa	Malay	VB
belas	Malay	NN	berat	Malay	ADJ
belayar	Malay	NN	berenang	Malay	VB
belenggu	Malay	NN	beri	Malay	VB
beli	Malay	VB	berita	Malay	NN
beliak	Malay	NN	beritahu	Malay	NN
belit	Malay	ADJ	berkah	Malay	NN
belulang	Malay	NN	berkat	Malay	NN
belum	Malay	NEG	bersih	Malay	ADJ
benam	Malay	VB	besan	Malay	NN
benang	Malay	NN	besar	Malay	ADJ
benar	Malay	ADJ	besi	Malay	NN

WORD	LANGUAGE	TAG	WORD	LANGUAGE	TAG
betapa	Malay	FT	biru	Malay	NN
betina	Malay	NN	bisa	Malay	ADJ
betis	Malay	NN	bisik	Malay	NN
betul	Malay	ADJ	bisu	Malay	ADJ
beza	Malay	VB	bodoh	Malay	ADJ
biak	Malay	VB	bohong	Malay	NN
biar	Malay	FT	bolak	Malay	ADJ
biara	Malay	NN	boleh	Malay	FT
biasa	Malay	ADJ	bondong	Malay	ADJ
bibir	Malay	NN	bongkak	Malay	ADJ
bicara	Malay	NN	bongkar	Malay	VB
bidadari	Malay	NN	bongkok	Malay	NN
bidang	Malay	NN	bosan	Malay	ADJ
bidara	Malay	NN	buah	Malay	NN
bijak	Malay	NN	buai	Malay	VB
bijaksana	Malay	NN	bual	Malay	VB
biji	Malay	NN	buang	Malay	VB
bilah	Malay	NN	buas	Malay	ADJ
bilang	Malay	FT	buat	Malay	FT
bimbang	Malay	ADJ	budi	Malay	NN
bin	Malay	NN	buih	Malay	NN
bina	Malay	ADJ	bujang	Malay	NN
binasa	Malay	VB	buka	Malay	VB
binatang	Malay	NN	bukit	Malay	NN
bincang	Malay	VB	bukti	Malay	NN
bingung	Malay	ADJ	buku	Malay	NN
bintang	Malay	NN	bulan	Malay	NN
binti	Malay	NN	bulat	Malay	ADJ

WORD	LANGUAGE	TAG	WORD	LANGUAGE	TAG
bulu	Malay	NN	capai	Malay	VB
bumbung	Malay	NN	cara	Malay	NN
bumi	Malay	NN	cari	Malay	VB
bunga	Malay	NN	catat	Malay	VB
bungkus	Malay	NN	cawan	Malay	NN
bunting	Malay	ADJ	cebur	Malay	VB
bunuh	Malay	VB	cecah	Malay	ADJ
Bunyamin	Arabic	NN	ceduk	Malay	NN
bunyi	Malay	NN	cegah	Malay	VB
buru	Malay	VB	cekik	Malay	VB
buruk	Malay	ADJ	cela	Malay	NN
burung	Malay	NN	celah	Malay	NN
busur	Malay	NN	celaka	Malay	ADJ
buta	Malay	ADJ	celik	Malay	ADJ
butir	Malay	NN	cemar	Malay	VB
cabang	Malay	NN	cemas	Malay	NN
cabar	Malay	VB	cemerlang	Malay	ADJ
cabut	Malay	VB	cemeti	Malay	NN
cacat	Malay	ADJ	cemuh	Malay	ADJ
caci	Malay	VB	cenderung	Malay	VB
cahaya	Malay	NN	cengang	Malay	VB
cair	Malay	ADJ	cepat	Malay	ADJ
cakap	Malay	VB	cerai	Malay	NN
calon	Malay	NN	cerca	Malay	VB
campak	Malay	VB	cerdas	Malay	NN
campur	Malay	VB	ceria	Malay	ADJ
cantik	Malay	ADJ	cerita	Malay	NN
cantum	Malay	VB	ceroboh	Malay	VB

WORD	LANGUAGE	TAG	WORD	LANGUAGE	TAG
cincang	Malay	VB	dahsyat	Malay	ADJ
cinta	Malay	NN	dahulu	Malay	ADJ
cipta	Malay	VB	daki	Malay	VB
ciri	Malay	NN	dakwa	Malay	VB
cita	Malay	NN	dakwah	Malay	NN
cium	Malay	VB	dalam	Malay	ADJ
condong	Malay	NN	dalih	Malay	NN
contoh	Malay	NN	dam	Arabic	NN
cuba	Malay	RB	damai	Malay	ADJ
cucu	Malay	NN	damping	Malay	ADJ
cucuk	Malay	NN	dangkal	Malay	NN
cucur	Malay	NN	dapat	Malay	VB
cukup	Malay	FT	dapur	Malay	NN
cukur	Malay	VB	darah	Malay	NN
culik	Malay	VB	darat	Malay	NN
cuma	Malay	FT	dari	Malay	IN
curah	Malay	VB	darjat	Malay	NN
curang	Malay	ADJ	darurat	Malay	NN
curi	Malay	VB	darussalam	Arabic	NN
curiga	Malay	ADJ	dasar	Malay	NN
dada	Malay	NN	datang	Malay	VB
dadak	Malay	ADJ	datar	Malay	ADJ
daerah	Malay	NN	datuk	Malay	NN
dagang	Malay	ADJ	Daud	Arabic	NN
daging	Malay	NN	daun	Malay	NN
dagu	Malay	NN	daya	Malay	NN
dahaga	Malay	VB	debat	Malay	NN
dahi	Malay	NN	debu	Malay	NN

WORD	LANGUAGE	TAG	WORD	LANGUAGE	TAG
dedah	Malay	VB	dingin	Malay	ADJ
dekap	Malay	VB	dirham	Arabic	NN
dekat	Malay	ADJ	diri	Malay	NN
delima	Malay	NN	doa	Malay	NN
demikian	Malay	FT	dongak	Malay	VB
denda	Malay	NN	dongeng	Malay	ADJ
dendam	Malay	ADJ	dorong	Malay	VB
dengan	Malay	FT	dosa	Malay	NN
dengar	Malay	NN	dua	Malay	VB
dengki	Malay	NN	duduk	Malay	VB
depan	Malay	FT	duga	Malay	VB
derai	Malay	VB	duka	Malay	VB
deras	Malay	ADJ	dukacita	Malay	NN
derhaka	Malay	ADJ	dulu	Malay	ADJ
derita	Malay	NN	dunia	Malay	NN
derma	Malay	NN	durhaka	Malay	ADJ
desa	Malay	NN	duri	Malay	NN
desak	Malay	ADJ	dusta	Malay	NN
desis	Malay	NN	dusun	Malay	NN
dewasa	Malay	ADJ	dzun	Arabic	NN
Dhuha	Malay	NN	edar	Malay	VB
diam	Malay	ADJ	ejek	Malay	VB
diat	Arabic	NN	ekor	Malay	NN
didih	Malay	VB	elak	Malay	VB
didik	Malay	VB	elok	Malay	ADJ
din	Arabic	NN	emas	Malay	NNU
dinar	Arabic	NN	embun	Malay	NN
dinding	Malay	NN	empat	Malay	CPD

WORD	LANGUAGE	TAG	WORD	LANGUAGE	TAG
empuk	Malay	ADJ	gagah	Malay	ADJ
enam	Malay	CPD	gagak	Malay	NN
endah	Malay	ADJ	gagal	Malay	ADJ
enggan	Malay	VB	gajah	Malay	NN
erti	Malay	NN	gala	Malay	NN
Esa	Malay	ADJ	galah	Malay	NN
esok	Malay	ADJ	gamak	Malay	VB
faedah	Malay	NN	gambar	Malay	NN
faham	Malay	VB	ganang	Malay	NN
faham	Malay	NN	ganas	Malay	ADJ
fajar	Malay	NN	ganda	Malay	VB
fakir	Malay	ADJ	gandum	Malay	NN
fasih	Malay	ADJ	ganggu	Malay	VB
fasik	Malay	ADJ	ganjaran	Malay	NN
fatamorgana	Malay	NN	ganjil	Malay	NN
fatwa	Malay	NN	ganti	Malay	VB
fidyah	Arabic	NN	garis	Malay	NN
fikir	Malay	NN	gaul	Malay	VB
Fir'aun	Arabic	NN	gegas	Malay	ADJ
firdaus	Arabic	NN	gejolak	Malay	VB
firman	Malay	NN	gelang	Malay	NN
fitnah	Malay	NN	gelap	Malay	ADJ
fitrah	Malay	NN	gelar	Malay	NN
fizikal	Malay	ADJ	gelas	Malay	NN
fulan	Arabic	NN	gelek	Malay	VB
furqan	Arabic	NN	geleng	Malay	VB
gabung	Malay	VB	gelimpang	Malay	VB
gadis	Malay	NN	gelincir	Malay	VB

WORD	LANGUAGE	TAG	WORD	LANGUAGE	TAG
gelisah	Malay	ADJ	gigit	Malay	VB
gelita	Malay	ADJ	gila	Malay	ADJ
gelombang	Malay	NN	gilir	Malay	VB
gelora	Malay	NN	goda	Malay	VB
gelumang	Malay	VB	golong	Malay	VB
gelupur	Malay	VB	goncang	Malay	ADJ
gema	Malay	NN	gua	Malay	NN
gembala	Malay	VB	gudang	Malay	NN
gembira	Malay	ADJ	gugur	Malay	VB
gementar	Malay	VB	gugus	Malay	NN
gempa	Malay	VB	gulung	Malay	VB
gempar	Malay	ADJ	gumpal	Malay	NN
gemuk	Malay	ADJ	guna	Malay	NN
gemuruh	Malay	ADJ	guni	Malay	NN
genap	Malay	ADJ	gunung	Malay	NN
gendong	Malay	VB	gurau	Malay	VB
generasi	Malay	NN	guruh	Malay	NN
genggam	Malay	VB	habil	Arabic	NN
gentar	Malay	ADJ	habis	Malay	ADJ
gerak	Malay	VB	had	Malay	NN
gerbang	Malay	NN	hadap	Malay	NN
gereja	Malay	NN	hadas	Arabic	NN
gesa	Malay	VB	hadiah	Malay	NN
gesit	Malay	ADJ	hadir	Malay	VB
getar	Malay	VB	hadrat	Arabic	NN
ghaib	Malay	ADJ	hadyu	Arabic	NN
giat	Malay	VB	hafsah	Arabic	NN
gigi	Malay	NN	haid	Malay	ADJ

WORD	LANGUAGE	TAG	WORD	LANGUAGE	TAG
hairan	Malay	VB	hara	Malay	NN
haiwan	Malay	NN	haram	Malay	ADJ
hajat	Malay	NN	harap	Malay	VB
haji	Malay	NN	harga	Malay	NN
hak	Malay	NN	hari	Malay	ADJ
hakikat	Malay	NN	Harithah	Arabic	NN
hakim	Malay	NN	harta	Malay	NN
hal	Malay	NN	haru	Malay	VB
hala	Malay	NN	harum	Malay	ADJ
halal	Malay	ADJ	harun	Arabic	NN
halang	Malay	VB	harus	Malay	FT
halau	Malay	VB	harut	Arabic	NN
halia	Malay	NN	hasad	Malay	ADJ
halilintar	Malay	NN	hasil	Malay	NN
halus	Malay	ADJ	hasta	Malay	NN
Haman	Arabic	NN	hasut	Malay	VB
hamba	Malay	NN	hati	Malay	NN
hambat	Malay	VB	haus	Malay	NN
hambur	Malay	VB	hawa	Malay	NN
hamil	Malay	ADJ	Hawa	Arabic	NN
hampar	Malay	VB	hawariyyun	Arabic	NN
hancur	Malay	ADJ	hawiyah	Arabic	NN
hangat	Malay	ADJ	hayat	Malay	VB
hangus	Malay	NN	hebat	Malay	ADJ
hanif	Arabic	NN	helai	Malay	NN
hantar	Malay	VB	hembus	Malay	VB
hanyut	Malay	VB	hendak	Malay	FT
hapus	Malay	VB	hentam	Malay	VB

WORD	LANGUAGE	TAG	WORD	LANGUAGE	TAG
henti	Malay	VB	hudaibiyah	Arabic	NN
herdik	Malay	VB	hujan	Malay	NN
heret	Malay	VB	hujung	Malay	NN
hias	Malay	VB	hukum	Malay	NNU
hibur	Malay	VB	hulur	Malay	VB
hidang	Malay	VB	hunain	Arabic	NN
hidap	Malay	VB	huni	Malay	VB
hidayah	Arabic	NN	huru	Malay	NN
hidung	Malay	VB	huruf	Malay	NN
hidup	Malay	VB	hurumat	Arabic	NN
hijau	Malay	NN	husna	Arabic	NN
hijr	Arabic	NN	hutamah	Arabic	NN
hijrah	Malay	VB	hutang	Malay	NN
hikmah	Malay	NN	ibadah	Malay	NN
hikmat	Malay	NN	ibadat	Malay	NN
hilang	Malay	ADJ	ibarat	Malay	FT
himpun	Malay	VB	iblis	Malay	ADJ
hina	Malay	ADJ	ibni	Arabic	NN
hindar	Malay	VB	ibnu	Arabic	NN
hingga	Malay	NN	ibrahim	Arabic	NN
hirau	Malay	NN	ibu	Malay	NN
hisab	Malay	NN	idah	Arabic	NN
hitam	Malay	ADJ	idris	Arabic	NN
hitung	Malay	VB	ifrit	Arabic	NN
homoseksual	Malay	ADJ	ihram	Arabic	NN
hormat	Malay	NN	ikan	Malay	NN
hubung	Malay	VB	ikat	Malay	NN
hud	Malay	NN	ikhlas	Malay	ADJ

WORD	LANGUAGE	TAG	WORD	LANGUAGE	TAG
ikrar	Malay	NN	islam	Arabic	NN
iktikad	Arabic	NN	Ismail	Arabic	NN
iktikaf	Arabic	NN	Israel	Arabic	NN
ikut	Malay	VB	Israil	Arabic	NN
ilham	Malay	ADJ	istana	Malay	NN
Illiyyun	Arabic	NN	isteri	Malay	NN
ilmu	Malay	NN	istimewa	Malay	ADJ
Ilyas	Arabic	NN	istiqamah	Arabic	VB
Ilyasa	Arabic	NN	istirehat	Malay	VB
iman	Malay	NN	isu	Malay	NN
Imran	Arabic	NN	isyak	Arabic	NN
inasan	Arabic	NN	isyarat	Malay	NN
indah	Malay	NNP	isytihar	Malay	NN
infak	Arabic	NN	izin	Malay	NN
ingat	Malay	VB	jabatan	Malay	NN
ingin	Malay	VB	jadi	Malay	VB
ingkar	Malay	VB	jaga	Malay	VB
injak	Malay	NN	jahanam	Malay	ADJ
injil	Arabic	NN	jahat	Malay	ADJ
insaf	Malay	VB	jahil	Malay	ADJ
intai	Malay	NN	jahiliyah	Arabic	NN
intip	Malay	VB	jahim	Arabic	NN
Iram	Arabic	NN	jalan	Malay	NN
iri	Malay	NN	jalar	Malay	VB
iring	Malay	ADJ	jalin	Malay	VB
Isa	Arabic	NN	jalur	Malay	VB
Ishak	Arabic	NN	jalut	Arabic	NN
isi	Malay	NN	jamah	Malay	VB

WORD	LANGUAGE	TAG	WORD	LANGUAGE	TAG
jamin	Malay	VB	jemu	Malay	ADJ
jamu	Malay	VB	jengkel	Malay	NN
janda	Malay	NN	jenis	Malay	NN
janggut	Malay	NN	jerit	Malay	VB
jangka	Malay	NN	jernih	Malay	ADJ
janin	Malay	NN	jerumus	Malay	VB
janji	Malay	NN	jibril	Arabic	NN
jantan	Malay	NN	Jibt dan Taghut	Arabic	NN
jantung	Malay	NN	jihad	Arabic	VB
jarak	Malay	VB	jijik	Malay	ADJ
jari	Malay	NN	jilbab	Arabic	NN
jarum	Malay	NN	jin	Malay	NN
jasa	Malay	NN	jinak	Malay	ADJ
jasad	Malay	NN	jiran	Malay	NN
jatuh	Malay	VB	jitu	Malay	ADJ
jauh	Malay	ADJ	jiwa	Malay	NN
jawab	Malay	VB	jizyah	Arabic	NN
jawat	Malay	VB	jua	Malay	RB
jaya	Malay	VB	jual	Malay	NN
jejak	Malay	NN	juang	Malay	VB
jejas	Malay	ADJ	judi	Malay	NN
jelajah	Malay	VB	juga	Malay	NN
jelang	Malay	VB	jujur	Malay	ADJ
jelas	Malay	ADJ	julai	Malay	NN
jelek	Malay	ADJ	julang	Malay	VB
jelita	Malay	ADJ	julur	Malay	ADJ
jemari	Malay	NN	jumaat	Malay	NN
jemput	Malay	VB	jumlah	Malay	NN

WORD	LANGUAGE	TAG	WORD	LANGUAGE	TAG
jumpa	Malay	VB	kalangan	Malay	NN
junjung	Malay	VB	kalau	Malay	FT
junub	Arabic	ADJ	kali	Malay	NN
jurang	Malay	NN	kali	Malay	NNC
juru	Malay	NN	kalian	Malay	NN
justeru	Malay	FT	kalimah	Malay	NN
kaabah	Arabic	NN	kalimat	Malay	NN
kabul	Malay	NN	kalong	Malay	NN
kabur	Malay	NN	kalung	Malay	NN
kabut	Malay	NN	Kamal	Arabic	NN
kaca	Malay	NN	kamar	Malay	NN
kacang	Malay	NN	kambing	Malay	NN
kacau	Malay	ADJ	kami	Malay	NN
kadangkala	Malay	ADJ	kampung	Malay	NN
kadar	Malay	NN	kamu	Malay	NN
kafarah	Arabic	NN	kanak	Malay	NN
kafilah	Arabic	NN	kanan	Malay	NN
kafir	Malay	NN	kandang	Malay	NN
kafur	Arabic	NN	kandas	Malay	ADJ
kagum	Malay	NN	kandung	Malay	NN
kahwin	Malay	VB	kapal	Malay	NN
kain	Malay	NN	karib	Malay	ADJ
kait	Malay	NN	karung	Malay	NN
kaki	Malay	NN	kasar	Malay	ADJ
kaku	Malay	ADJ	kasih	Malay	VB
kala	Malay	NN	kasturi	Malay	NN
kalah	Malay	ADJ	kata	Malay	NN
kalalah	Arabic	NN	katak	Malay	NN

WORD	LANGUAGE	TAG	WORD	LANGUAGE	TAG
kau	Malay	NN	kelmarin	Malay	ADJ
kaum	Malay	NN	kelompok	Malay	NN
kawal	Malay	VB	kelopak	Malay	NN
kawan	Malay	NN	keluar	Malay	VB
kawasan	Malay	NN	keluarga	Malay	NN
kaya	Malay	ADJ	keluh	Malay	VB
kayu	Malay	NN	kemalangan	Malay	NN
keadaan	Malay	NN	kemaluan	Malay	NN
kebajikan	Malay	NN	kemarau	Malay	ADJ
kebun	Malay	NN	kemaslahatan	Malay	NN
kecewa	Malay	ADJ	kembali	Malay	VB
kecil	Malay	ADJ	kembang	Malay	VB
kecuali	Malay	FT	kembara	Malay	VB
kedekut	Malay	ADJ	kemudian	Malay	ADJ
kediaman	Malay	NN	kena	Malay	VB
kedip	Malay	VB	kenal	Malay	VB
kejam	Malay	ADJ	kenan	Malay	VB
kejap	Malay	VB	kenang	Malay	VB
kejar	Malay	VB	kencang	Malay	ADJ
keji	Malay	ADJ	kendali	Malay	VB
kejut	Malay	VB	kenderaan	Malay	NN
kekal	Malay	ADJ	kendi	Malay	NN
kelahi	Malay	VB	kening	Malay	NN
kelakuan	Malay	NN	kenyataan	Malay	NN
keldai	Malay	NN	kepala	Malay	NN
keliling	Malay	NN	keping	Malay	NN
kelip	Malay	VB	kepit	Malay	VB
keliru	Malay	ADJ	kepung	Malay	NN

WORD	LANGUAGE	TAG	WORD	LANGUAGE	TAG
kera	Malay	NN	khazanah	Malay	NN
kerabat	Malay	NN	khemah	Malay	NN
kerah	Malay	VB	khianat	Malay	ADJ
kerajaan	Malay	NN	khidir	Arabic	NN
kerap	Malay	ADJ	khuatir	Malay	VB
keras	Malay	ADJ	khuldi	Arabic	NN
kerdip	Malay	VB	khusus	Malay	ADJ
kerikil	Malay	NN	khusyuk	Malay	ADJ
kering	Malay	ADJ	khutbah	Arabic	NN
kerja	Malay	NN	kiamat	Malay	NN
kerongkong	Malay	NN	kiasan	Malay	NN
kertas	Malay	NN	kibas	Arabic	NN
kerumun	Malay	VB	kiblat	Malay	NN
kerut	Malay	ADJ	kifarat	Arabic	NN
kesah	Malay	VB	kilas	Malay	ADJ
kesal	Malay	ADJ	kilat	Malay	NN
kesan	Malay	NN	kilau	Malay	NN
ketat	Malay	ADJ	kira	Malay	NN
ketawa	Malay	NN	kiri	Malay	NN
ketiak	Malay	NN	kirim	Malay	VB
ketika	Malay	NN	kisah	Malay	NN
ketua	Malay	NN	kita	Malay	NN
keturunan	Malay	NN	kitab	Malay	NN
khabar	Malay	NN	kitar	Malay	NN
khalifah	Malay	NN	kobar	Malay	VB
khamar	Arabic	NN	kolam	Malay	NN
khas	Malay	ADJ	kongsi	Malay	NN
khawatir	Arabic	ADJ	kontang	Malay	ADJ

WORD	LANGUAGE	TAG	WORD	LANGUAGE	TAG
kontang-kanting	Malay	ADJ	kutip	Malay	VB
korban	Malay	NN	kutu	Malay	NN
kosong	Malay	ADJ	kutuk	Malay	VB
kota	Malay	NN	labah	Malay	NN
kotor	Malay	ADJ	labu	Malay	NN
koyak	Malay	ADJ	labuh	Malay	ADJ
kristal	Malay	NN	lacur	Malay	VB
kuasa	Malay	NN	ladang	Malay	NN
kuat	Malay	ADJ	lafaz	Malay	NN
kubur	Malay	NN	laga	Malay	VB
kuda	Malay	NN	Lahab	Arabic	NN
kufur	Malay	NN	lahir	Malay	VB
kuku	Malay	NN	lailatul	Arabic	NN
kukuh	Malay	ADJ	lain	Malay	ADJ
kumpul	Malay	VB	laju	Malay	NN
kunang	Malay	NN	laNNat	Malay	NN
kunci	Malay	NN	laksana	Malay	NN
kuning	Malay	NN	laku	Malay	ADJ
kunjung	Malay	VB	lalai	Malay	ADJ
kupu	Malay	NN	lalat	Malay	NN
kurang	Malay	FT	lali	Malay	ADJ
kurma	Malay	NN	lalu	Malay	VB
kurnia	Malay	NN	lama	Malay	ADJ
kursi	Arabic	NN	laman	Malay	NN
kurun	Malay	NN	lambat	Malay	ADJ
kurung	Malay	NN	lambung	Malay	VB
kurus	Malay	ADJ	lampau	Malay	ADJ
kusta	Malay	NN	lampias	Malay	ADJ

WORD	LANGUAGE	TAG	WORD	LANGUAGE	TAG
lancar	Malay	ADJ	lebat	Malay	NN
landa	Malay	VB	lebih	Malay	FT
langgar	Malay	VB	lebur	Malay	ADJ
langit	Malay	NN	lecur	Malay	VB
langkah	Malay	NN	ledak	Malay	NN
lanjut	Malay	ADJ	lega	Malay	ADJ
lantai	Malay	NN	leher	Malay	NN
lantar	Malay	VB	leka	Malay	ADJ
lapan	Malay	CDP	lekat	Malay	NN
lapang	Malay	ADJ	lelah	Malay	NN
lapar	Malay	ADJ	lelaki	Malay	NN
lapis	Malay	NN	lemah	Malay	ADJ
larang	Malay	RB	lemak	Malay	NN
larat	Malay	VB	lembah	Malay	NN
lari	Malay	VB	lembaran	Malay	NN
lata	Malay	VB	lembing	Malay	NN
latih	Malay	NN	lembu	Malay	NN
lauh-lauh	Arabic	NN	lereng	Malay	NN
laut	Malay	NN	lesbian	Malay	ADJ
lawan	Malay	NN	lesu	Malay	ADJ
layak	Malay	ADJ	letak	Malay	NN
layan	Malay	VB	letih	Malay	ADJ
layang	Malay	NN	liar	Malay	ADJ
layar	Malay	VB	liat	Malay	ADJ
lazat	Malay	ADJ	libat	Malay	VB
lbn	Arabic	NN	licin	Malay	ADJ
lebah	Malay	NN	lidah	Malay	NN
lebar	Malay	ADJ	lembut	Malay	ADJ

WORD	LANGUAGE	TAG	WORD	LANGUAGE	TAG
lempar	Malay	VB	luh	Arabic	NN
lenang	Malay	ADJ	luka	Malay	NN
lengan	Malay	NN	luluh	Malay	ADJ
lengkap	Malay	ADJ	lumat	Malay	ADJ
lenguh	Malay	ADJ	lumba	Malay	NN
lenyap	Malay	VB	lumpuh	Malay	ADJ
lepas	Malay	ADJ	lumpur	Malay	NN
lihat	Malay	VB	lumur	Malay	VB
lima	Malay	CDC	lunak	Malay	NN
limpah	Malay	ADJ	luncur	Malay	VB
lindung	Malay	VB	lupa	Malay	VB
lingkar	Malay	VB	luput	Malay	ADJ
lingkung	Malay	VB	luqman	Arabic	NN
lintas	Malay	VB	lurus	Malay	ADJ
lipat	Malay	ADJ	lut	Malay	NN
liput	Malay	VB	lutut	Malay	NN
lisan	Malay	VB	maaf	Malay	NN
litup	Malay	VB	mabuk	Malay	ADJ
lngat	Malay	VB	macam	Malay	NN
lni	Malay	DT	madinah	Arabic	NN
logam	Malay	NN	madu	Malay	NN
lompat	Malay	VB	madyan	Arabic	NN
lontar	Malay	NN	Maha Suci	Malay	ADJ
ltu	Malay	DT	mahar	Arabic	NN
luap	Malay	VB	maharaja	Malay	NN
luar	Malay	NN	maharajalela	Malay	VB
luas	Malay	ADJ	mahfuz	Arabic	NN
lubang	Malay	NN	mahir	Malay	ADJ

WORD	LANGUAGE	TAG	WORD	LANGUAGE	TAG
mahjura	Arabic	NN	Manat	Arabic	NN
mahsyar	Arabic	NN	mandi	Malay	VB
mahu	Malay	VB	mandul	Malay	ADJ
main	Malay	VB	manfaat	Malay	NN
majlis	Malay	NN	mangsa	Malay	NN
maju	Malay	ADJ	mani	Malay	NN
majusi	Arabic	NN	manis	Malay	ADJ
makam	Malay	NN	manna	Arabic	NN
makan	Malay	VB	manusia	Malay	NN
makhluk	Malay	NN	maqdis	Arabic	NN
maki	Malay	VB	marah	Malay	VB
makin	Malay	FT	mari	Malay	FT
makjuj	Arabic	NN	marjan	Arabic	NN
maVBah	Malay	NN	martabat	Malay	NN
maklum	Malay	VB	marut	Arabic	NN
makmur	Malay	ADJ	marwah	Arabic	NN
maNNa	Malay	NN	Maryam	Arabic	NN
makruf	Arabic	ADJ	mas	Malay	NN
maksiat	Malay	NN	masa	Malay	NN
maksud	Malay	NN	masak	Malay	ADJ
malaikat	Malay	NN	masam	Malay	ADJ
malam	Malay	NN	masin	Malay	ADJ
malang	Malay	ADJ	masjid	Malay	NN
malapetaka	Malay	NN	masjidil	Malay	NN
malas	Malay	ADJ	masuk	Malay	VB
malu	Malay	ADJ	Masy'aril	Arabic	NN
mampu	Malay	VB	mata	Malay	NN
mana	Malay	RB	matahari	Malay	NN

WORD	LANGUAGE	TAG	WORD	LANGUAGE	TAG
mati	Malay	VB	minta	Malay	VB
maut	Malay	NN	minum	Malay	VB
mawar	Malay	NN	minyak	Malay	NN
mayang	Malay	NN	misal	Malay	NN
mayat	Malay	NN	miskin	Malay	ADJ
medan	Malay	NN	moga	Malay	NN
megah	Malay	ADJ	mohon	Malay	VB
mekah	Malay	NN	moyang	Malay	NN
memang	Malay	FT	mualaf	Arabic	NN
memphis	Arabic	NN	muat	Malay	ADJ
menang	Malay	VB	mubahalah	Arabic	NN
menantu	Malay	NN	muda	Malay	ADJ
mendung	Malay	ADJ	mudah	Malay	ADJ
mengaku	Malay	VB	mudarat	Malay	NN
merah	Malay	NN	muhajirin	Arabic	NN
merdeka	Malay	ADJ	muhammad	Arabic	NN
mertua	Malay	NN	muhkamat	Arabic	NN
mesir	Malay	NN	muka	Malay	NN
mesra	Malay	ADJ	mukim	Malay	NN
mesti	Malay	FT	mukjizat	Arabic	NN
mesyuarat	Malay	NN	mukmin	Arabic	NN
mewah	Malay	ADJ	mukminin	Arabic	NN
mihrab	Arabic	NN	mula	Malay	FT
mikail	Arabic	NN	mulia	Malay	ADJ
milik	Malay	NN	mulut	Malay	NN
mimbar	Arabic	NN	munafik	Malay	ADJ
mimpi	Malay	NN	munajat	Malay	VB
mina	Arabic	NN	mundur	Malay	ADJ

WORD	**LANGUAGE**	**TAG**	**WORD**	**LANGUAGE**	**TAG**
mungkar	Malay	VB	nafas	Malay	ADJ
mungkin	Malay	FT	nafkah	Malay	NN
mungkir	Malay	VB	nafsu	Malay	NN
muntah	Malay	NN	nahas	Malay	ADJ
murah	Malay	ADJ	naik	Malay	VB
muram	Malay	ADJ	najis	Malay	NN
murka	Malay	ADJ	najran	Arabic	NN
murni	Malay	ADJ	nama	Malay	NN
murtad	Arabic	NN	nampak	Malay	VB
musa	Arabic	NN	namun	Malay	FT
musafir	Arabic	NN	nanah	Malay	NN
musaharah	Arabic	NN	nanti	Malay	NN
musibah	Arabic	NN	nasab	Malay	NN
musim	Malay	NN	nasib	Malay	NN
muslihat	Malay	NN	nasihat	Malay	NN
muslim	Malay	NN	Nasr	Arabic	NN
muslimin	Arabic	NN	nasrani	Arabic	NN
musnah	Malay	ADJ	nasuha	Arabic	NN
mustahil	Malay	ADJ	naung	Malay	VB
musuh	Malay	NN	nazar	Malay	NN
musyrik	Arabic	NN	negeri	Malay	NN
musyrikin	Arabic	NN	nenek	Malay	NN
mut'ah	Arabic	NN	neraca	Malay	NN
mutasyabihat	Arabic	NN	neraka	Malay	NN
mutiara	Malay	NN	ngantuk	Malay	VBT
mutlak	Malay	NN	ngeri	Malay	ADJ
nabi	Malay	NN	niaga	Malay	VB
nada	Malay	NN	niat	Malay	NN

WORD	LANGUAGE	TAG	WORD	LANGUAGE	TAG
nikah	Malay	VB	paha	Malay	NN
nil	Arabic	NN	pahala	Malay	NN
nilai	Malay	NN	pahat	Malay	NN
nipis	Malay	RB	pahit	Malay	ADJ
nuh	Arabic	NN	pajak	Malay	NN
nun	Arabic	NN	pakai	Malay	VB
nurani	Arabic	NN	pakat	Malay	ADJ
nusyuz	Arabic	NN	paksa	Malay	VB
nyala	Malay	NN	paku	Malay	NN
nyaman	Malay	ADJ	palestin	Malay	NN
nyamuk	Malay	NN	paling	Malay	FT
nyaris	Malay	NN	palsu	Malay	ADJ
nyata	Malay	ADJ	paluan	Malay	NN
nyawa	Malay	NN	panah	Malay	NN
olah	Malay	NN	panas	Malay	ADJ
oleh	Malay	VB	pancang	Malay	NN
olok	Malay	NN	pancar	Malay	VB
ombak	Malay	NN	pancut	Malay	VB
orang	Malay	NN	pandai	Malay	ADJ
orbit	Malay	NN	pandang	Malay	VB
pada	Malay	FT	panggang	Malay	NN
padahal	Malay	NN	panggil	Malay	VB
padam	Malay	VB	pangkal	Malay	NN
padan	Malay	ADJ	panik	Malay	NN
padang	Malay	NN	panjang	Malay	ADJ
padi	Malay	NN	panjat	Malay	VB
padu	Malay	ADJ	pantas	Malay	ADJ
pagi	Malay	NN	papan	Malay	NN

WORD	LANGUAGE	TAG	WORD	LANGUAGE	TAG
para	Malay	FT	pelihara	Malay	VB
parit	Malay	NN	pelik	Malay	ADJ
pasak	Malay	NN	pelita	Malay	NN
pasang	Malay	VB	peluang	Malay	NN
pasar	Malay	NN	peluk	Malay	VB
pasir	Malay	NNU	pena	Malay	NN
pasrah	Malay	ADJ	penat	Malay	ADJ
pasti	Malay	ADJ	pencil	Malay	NN
pasukan	Malay	NN	pendam	Malay	VB
patah	Malay	VB	pendek	Malay	ADJ
pati	Malay	NN	pendeta	Malay	NN
patuh	Malay	VB	pengaruh	Malay	NN
patung	Malay	NN	pengetahuan	Malay	NN
patut	Malay	FT	pengsan	Malay	ADJ
payah	Malay	ADJ	pening	Malay	ADJ
payau	Malay	ADJ	penjara	Malay	NN
pecah	Malay	VB	penting	Malay	ADJ
pecut	Malay	NN	penuh	Malay	ADJ
pedih	Malay	ADJ	perahu	Malay	NN
pedoman	Malay	NN	perak	Malay	NN
peduli	Malay	VB	peranan	Malay	NNU
pegang	Malay	VB	perang	Malay	NN
pejam	Malay	VB	perangai	Malay	NN
pekak	Malay	ADJ	peranjat	Malay	VB
pekat	Malay	ADJ	peras	Malay	VB
pekik	Malay	VB	perasaan	Malay	NN
pelamin	Malay	NN	perawan	Malay	NN
pelepah	Malay	NN	percaya	Malay	VB

WORD	LANGUAGE	TAG	WORD	LANGUAGE	TAG
percik	Malay	VB	piala	Malay	NN
perempuan	Malay	NN	pihak	Malay	NN
pergi	Malay	VB	pijak	Malay	NN
peri	Malay	FT	pikat	Malay	VB
perihal	Malay	NN	pikul	Malay	VB
periksa	Malay	NN	pilih	Malay	VB
perinci	Malay	NN	pimpin	Malay	VB
perintah	Malay	NN	pinak	Malay	NN
perisai	Malay	NN	pinang	Malay	VB
peristiwa	Malay	NN	pindah	Malay	VB
periuk	Malay	NN	pinggan	Malay	NN
perkakas	Malay	NN	pinggir	Malay	NN
perkasa	Malay	NN	pinjam	Malay	VB
perlahan	Malay	ADJ	pintal	Malay	VBI
perlu	Malay	FT	pintar	Malay	ADJ
permaidani	Malay	NN	pintu	Malay	NN
permata	Malay	NN	piring	Malay	NN
pernah	Malay	NN	pisah	Malay	VB
perut	Malay	NN	pisang	Malay	NN
pesan	Malay	NN	pisau	Malay	NN
pesona	Malay	NN	piutang	Malay	NN
pesong	Malay	VB	pohon	Malay	NNC
petala	Malay	NN	pokok	Malay	NN
petang	Malay	NN	potong	Malay	VB
petani	Malay	VBI	prasangka	Malay	NN
peti	Malay	NN	puak	Malay	NN
petik	Malay	VB	puas	Malay	ADJ
petir	Malay	NN	puasa	Malay	NN

WORD	LANGUAGE	TAG	WORD	LANGUAGE	TAG
pudar	Malay	ADJ	qasar	Arabic	NN
puisi	Malay	NN	qisas	Arabic	NN
puja	Malay	VB	quraisy	Arabic	NN
puji	Malay	VB	quraizah	Arabic	NN
puji	Malay	VB	quran	Arabic	NN
pujuk	Malay	VB	qurban	Arabic	NN
pukau	Malay	NN	quru	Arabic	NN
pukul	Malay	VB	Ra'ina	Arabic	NN
pula	Malay	FT	raba	Malay	VB
pulang	Malay	VB	ragam	Malay	NN
puluh	Malay	FT	ragu	Malay	ADJ
puncak	Malay	NN	rahbaniyyah	Arabic	NN
punggung	Malay	NN	rahib	Arabic	NN
pungut	Malay	VB	rahim	Arabic	NN
punya	Malay	VB	rahman	Arabic	NN
purnama	Malay	NN	rahmat	Malay	NN
pusaka	Malay	NN	rahsia	Malay	ADJ
pusing	Malay	VB	raja	Malay	NN
putar	Malay	VB	rakaat	Arabic	NN
putera	Malay	NN	ramadan	Malay	NN
puteri	Malay	NN	ramai	Malay	ADJ
putih	Malay	ADJ	rambut	Malay	NN
putus	Malay	VB	rampas	Malay	VB
qabil	Arabic	NN	rancang	Malay	VB
qadar	Arabic	NN	rangkak	Malay	VB
qaf	Arabic	NN	rangkul	Malay	VB
qalaid	Arabic	NN	rantai	Malay	NN
qarun	Arabic	NN	ranting	Malay	ADJ

WORD	LANGUAGE	TAG	WORD	LANGUAGE	TAG
rapat	Malay	ADJ	reput	Malay	ADJ
rapi	Malay	ADJ	resap	Malay	VB
rapuh	Malay	NN	retak	Malay	NN
rasa	Malay	ADJ	rezeki	Malay	NN
rasmi	Malay	ADJ	ria	Malay	ADJ
rasul	Malay	NN	riak	Malay	ADJ
rasulullah	Arabic	NN	riang	Malay	ADJ
rata	Malay	ADJ	riba	Malay	NN
ratus	Malay	VB	ribu	Malay	NN
raya	Malay	ADJ	ribut	Malay	ADJ
rayap	Malay	VB	ringan	Malay	ADJ
rayu	Malay	VB	rintang	Malay	VB
rebah	Malay	VB	rintih	Malay	VB
rebut	Malay	VB	risalah	Malay	NN
reda	Malay	VB	risau	Malay	ADJ
redha	Malay	VB	riuh	Malay	ADJ
redup	Malay	NN	roboh	Malay	VB
rehat	Malay	VB	roh	Malay	NN
rejam	Malay	VB	romawi	Arabic	NN
reka	Malay	VB	rombong	Malay	NN
rela	Malay	VB	rompak	Malay	VB
remeh	Malay	ADJ	rongga	Malay	NN
rencana	Malay	NN	rosak	Malay	ADJ
rendah	Malay	ADJ	rotan	Malay	NN
rendang	Malay	ADJ	roti	Malay	NN
rentak	Malay	NN	ruah	Malay	NN
rentang	Malay	NN	ruang	Malay	NNC
renung	Malay	VB	rugi	Malay	ADJ

WORD	**LANGUAGE**	**TAG**	**WORD**	**LANGUAGE**	**TAG**
Ruhul qudus	Arabic	NN	sakit	Malay	ADJ
rujuk	Arabic	VB	saksi	Malay	NN
rukuk	Arabic	VB	salah	Malay	NN
rumah	Malay	NN	salam	Malay	NN
rumput	Malay	NN	salamun	Arabic	NN
runding	Malay	VB	saleh	Arabic	NN
runtuh	Malay	NN	salib	Malay	NN
rupa	Malay	NN	salin	Malay	VB
Sa'ibah	Arabic	NN	salsabil	Arabic	NN
saad	Arabic	NN	salwa	Arabic	NN
saat	Malay	NN	sama	Malay	IN
saba	Malay	VB	sambar	Malay	VB
sabar	Malay	NN	sambung	Malay	VB
Sabat	Arabic	NN	sambut	Malay	VB
sabiin	Arabic	NN	samiri	Arabic	NN
sabil	Arabic	NN	sampah	Malay	NN
sabit	Malay	NN	sampai	Malay	NN
sabtu	Malay	NN	samping	Malay	FT
sabut	Malay	NN	samud	Arabic	NN
sadar	Malay	NN	samun	Malay	VB
saf	Malay	NN	sandang	Malay	VB
safa	Arabic	ADJ	sandar	Malay	NN
sah	Malay	ADJ	sangat	Malay	RB
sahabat	Malay	NN	sanggah	Malay	VB
sahaja	Malay	FT	sanggup	Malay	VB
sahaya	Malay	NN	sangka	Malay	NN
saji	Malay	NN	sangkakala	Malay	NN
sakaratul	Arabic	NN	sangkal	Malay	VB

WORD	LANGUAGE	TAG	WORD	LANGUAGE	TAG
santun	Malay	ADJ	sedikit	Malay	ADJ
sapa	Malay	VB	sedu	Malay	VB
sapi	Malay	NN	segala	Malay	NN
sapih	Arabic	NN	segar	Malay	ADJ
sapu	Malay	NN	segera	Malay	RB
saqar	Arabic	NN	seimbang	Malay	ADJ
saran	Malay	NN	sejahtera	Malay	ADJ
sarang	Malay	NN	sejuk	Malay	ADJ
sari	Malay	NN	sekaligus	Malay	ADJ
sasaran	Malay	NN	sekat	Malay	NN
satu	Malay	NN	seksa	Malay	ADJ
saudara	Malay	NN	sekutu	Malay	NN
sawah	Malay	NN	selagi	Malay	NN
sawi	Malay	NN	selamat	Malay	NN
sayang	Malay	ADJ	selang	Malay	NN
sayap	Malay	NN	selaput	Malay	NN
sayur	Malay	NN	selar	Malay	NN
sebab	Malay	CC	selawat	Arabic	NN
sebar	Malay	VB	selera	Malay	NN
sebat	Malay	VB	selesai	Malay	ADJ
seberang	Malay	FT	selimut	Malay	NN
sebut	Malay	NN	selisih	Malay	NN
sedap	Malay	ADJ	seludang	Malay	NN
sedar	Malay	VB	seluruh	Malay	CDI
sedekah	Malay	NN	semakin	Malay	ADJ
sederhana	Malay	ADJ	semangat	Malay	NN
sedia	Malay	ADJ	semasa	Malay	NN
sedih	Malay	ADJ	semayam	Malay	VB

WORD	LANGUAGE	TAG	WORD	LANGUAGE	TAG
sembah	Malay	VB	sepuluh	Malay	CDP
sembahyang	Malay	NN	serah	Malay	VB
sembelih	Malay	VB	serak	Malay	ADJ
sembilan	Malay	CDP	serang	Malay	VB
sembuh	Malay	ADJ	serasi	Malay	NN
sembunyi	Malay	VB	serba	Malay	FT
sembur	Malay	NN	serbu	Malay	VB
semoga	Malay	NN	seret	Malay	VB
sempadan	Malay	NN	seri	Malay	NN
sempat	Malay	ADJ	serigala	Malay	NN
sempit	Malay	ADJ	sering	Malay	ADJ
sempurna	Malay	ADJ	serta	Malay	FT
semua	Malay	FT	seru	Malay	VB
semut	Malay	NN	sesak	Malay	ADJ
senang	Malay	ADJ	sesal	Malay	NN
senda	Malay	NN	sesat	Malay	ADJ
sendi	Malay	NN	sesuai	Malay	ADJ
sendiri	Malay	NN	sesuatu	Malay	NN
sengaja	Malay	VB	setara	Malay	NN
sengsara	Malay	ADJ	setelah	Malay	SC
senja	Malay	NN	setia	Malay	ADJ
senjata	Malay	NN	setuju	Malay	NN
sentosa	Malay	ADJ	siang	Malay	NN
sentuh	Malay	VB	siap	Malay	VB
senyum	Malay	VB	siapa	Malay	NN
sepakat	Malay	NN	siar	Malay	VB
sepatu	Malay	NN	siasat	Malay	VB
seperti	Malay	FT	sia-sia	Malay	ADJ

WORD	LANGUAGE	TAG	WORD	LANGUAGE	TAG
sibghah	Arabic	NN	siuman	Malay	ADJ
sibuk	Malay	ADJ	soal	Malay	VB
sidr	Arabic	NN	sodom	Arabic	NN
sidratulmuntaha	Arabic	NN	sokong	Malay	NN
sifat	Malay	NN	solat	Malay	NN
sihat	Malay	ADJ	soleh	Malay	ADJ
sihir	Malay	NN	solehah	Malay	ADJ
sikap	Malay	NN	sombong	Malay	ADJ
siku	Malay	NN	sorak	Malay	VB
sila	Malay	VB	strategi	Malay	NN
silang	Malay	VB	suai	Malay	VB
silih	Malay	VB	suami	Malay	NN
simpan	Malay	VB	suara	Malay	NN
simpang	Malay	NN	suasana	Malay	NN
simpul	Malay	VB	subuh	Malay	NN
Sinai	Arabic	NN	subur	Malay	ADJ
sinar	Malay	NN	suci	Malay	ADJ
singa	Malay	NN	sudah	Malay	FT
singgahsana	Malay	NN	sufyan	Arabic	NN
singkap	Malay	VB	sujud	Malay	VB
singkir	Malay	VB	suka	Malay	ADJ
singsing	Malay	VB	sukar	Malay	ADJ
sini	Malay	PRL	sukarela	Malay	ADJ
siram	Malay	VB	sukat	Malay	VB
sisa	Malay	NN	suku	Malay	NN
sisi	Malay	NN	sulaiman	Arabic	NN
sisih	Malay	VB	sulbi	Arabic	NN
siul	Malay	VB	sulit	Malay	ADJ

WORD	LANGUAGE	TAG	WORD	LANGUAGE	TAG
sumbat	Malay	NN	syeikh	Malay	NN
sumber	Malay	NN	syiar	Malay	NN
sumpah	Malay	NN	Syi'ra	Arabic	NN
sungai	Malay	NN	syirik	Arabic	NN
sungguh	Malay	ADJ	syuhada	Arabic	NN
sungkur	Malay	VB	syukur	Malay	VB
sunnah	Arabic	NN	syurga	Malay	NN
sunyi	Malay	ADJ	taat	Malay	ADJ
surah	Arabic	NN	tabiat	Malay	NN
surai	Malay	VB	tabir	Malay	NN
suram	Malay	ADJ	tabung	Malay	NN
surat	Malay	NN	tabur	Malay	ADJ
suruh	Malay	VB	Tabut	Arabic	NN
surut	Malay	VB	tadbir	Malay	VB
susah	Malay	ADJ	tadi	Malay	ADJ
susu	Malay	NN	Taghut	Arabic	NN
susun	Malay	VB	tagih	Malay	VB
sutera	Malay	NN	tahajud	Malay	NN
Suwa'	Arabic	NN	tahan	Malay	ADJ
syafaat	Malay	NN	tahap	Malay	NN
syahid	Arabic	NN	tahu	Malay	VB
syahwat	Arabic	NN	tahun	Malay	NN
syair	Malay	NN	tajam	Malay	ADJ
syaitan	Malay	NN	takdir	Malay	NN
syak	Malay	VB	takjub	Malay	ADJ
syam	Arabic	NN	takluk	Malay	VB
syarak	Malay	NN	takut	Malay	VB
syariat	Malay	ADJ	takwa	Malay	VB

WORD	LANGUAGE	TAG	WORD	LANGUAGE	TAG
takwil	Malay	NN	taraf	Malay	NN
talak	Malay	NN	tarik	Malay	VB
tali	Malay	VB	tartil	Malay	NN
tamak	Malay	ADJ	taruh	Malay	NN
taman	Malay	NN	tasbih	Malay	NN
tambah	Malay	VB	tasik	Malay	NN
tampak	Malay	VB	tasnim	Arabic	NN
tampil	Malay	VB	tatah	Malay	NN
tamu	Malay	NN	tatang	Malay	VB
tanah	Malay	NN	tatkala	Malay	FT
tanam	Malay	VB	taubat	Malay	VB
tanda	Malay	NN	taufan	Malay	NN
tandan	Malay	NN	taufik	Malay	NN
tanding	Malay	NN	tauhid	Arabic	NN
tanduk	Malay	NN	taurat	Arabic	NN
tandus	Malay	ADJ	taut	Malay	VB
tangan	Malay	NN	tawa	Malay	VB
tangga	Malay	NN	tawaf	Arabic	NN
tanggal	Malay	VB	tawakal	Arabic	NN
tangguh	Malay	VB	tawan	Malay	VB
tanggung	Malay	VB	tawar	Malay	ADJ
tanggungjawab	Malay	NN	tayamum	Arabic	NN
tangis	Malay	VB	tebal	Malay	ADJ
tangkai	Malay	NN	tebang	Malay	VB
tangkap	Malay	VB	tebar	Malay	VB
tangkas	Malay	ADJ	tebuk	Malay	VB
tanya	Malay	VB	tebus	Malay	VB
tar	Malay	NN	teduh	Malay	ADJ

WORD	LANGUAGE	TAG	WORD	LANGUAGE	TAG
tegak	Malay	ADJ	tempoh	Malay	NN
tegap	Malay	ADJ	tempuh	Malay	VB
tegas	Malay	ADJ	tempur	Malay	VB
teguh	Malay	ADJ	temu	Malay	VB
teguk	Malay	VB	temurun	Malay	NN
tegur	Malay	VB	tenang	Malay	ADJ
teka	Malay	VB	tengah	Malay	FT
tekad	Malay	NN	tenggelam	Malay	VB
teki	Malay	NN	tenggorokan	Malay	NN
tekun	Malay	ADJ	tengkar	Malay	VB
teladan	Malay	NN	tengkuk	Malay	NN
telaga	Malay	NN	tentang	Malay	FT
telah	Malay	FT	tentera	Malay	NN
telan	Malay	VB	tenteram	Malay	ADJ
telanjang	Malay	ADJ	tentu	Malay	FT
telapak	Malay	NN	tenung	Malay	VB
telinga	Malay	NN	tepat	Malay	ADJ
telingkah	Malay	NN	tepi	Malay	NN
teliti	Malay	ADJ	tepuk	Malay	VB
telungkup	Malay	NN	terang	Malay	ADJ
telur	Malay	NN	terbang	Malay	VB
teman	Malay	NN	terbit	Malay	VB
tembaga	Malay	NN	teriak	Malay	VB
tembikar	Malay	NN	terik	Malay	ADJ
tembok	Malay	NN	terima	Malay	VB
tembus	Malay	VB	terjun	Malay	VB
tempang	Malay	ADJ	terka	Malay	VB
tempat	Malay	NN	terkam	Malay	VB

WORD	LANGUAGE	TAG	WORD	LANGUAGE	TAG
ternak	Malay	VB	tingkat	Malay	NN
tertib	Malay	ADJ	tinjau	Malay	VB
terus	Malay	FT	tinta	Malay	NN
tetap	Malay	ADJ	tipu	Malay	NN
Thaif	Arabic	NN	tiri	Malay	NN
Thalut	Arabic	NN	tiru	Malay	VB
Thur	Arabic	NN	titah	Malay	NN
tiang	Malay	NN	titis	Malay	NN
tiba	Malay	VB	tiup	Malay	VB
tidak	Malay	FT	tolak	Malay	VB
tidur	Malay	VB	toleh	Malay	VB
tiga	Malay	cdp	tolong	Malay	VB
tilam	Malay	NN	tompok	Malay	NN
tilik	Malay	VB	tongkat	Malay	NN
timba	Malay	NN	tsamud	Arabic	NN
timbang	Malay	ADJ	tua	Malay	ADJ
timbul	Malay	VB	tuai	Malay	NN
timbun	Malay	ADJ	tuan	Malay	NN
timpa	Malay	VB	tuang	Malay	VB
timpal	Malay	NN	Tubba'	Arabic	NN
timun	Malay	NN	tubi	Malay	ADJ
timur	Malay	NN	tubuh	Malay	NNC
tin	Malay	NN	tubuh	Malay	NN
tindak	Malay	VB	tuduh	Malay	VB
tindas	Malay	VB	tudung	Malay	NN
tinggal	Malay	VB	tugas	Malay	NN
tinggi	Malay	ADJ	tuhan	Malay	NN
tingkah	Malay	NN	tuju	Malay	VB

WORD	LANGUAGE	TAG	WORD	LANGUAGE	TAG
tujuh	Malay	cdp	ubah	Malay	VB
tukang	Malay	NN	uban	Malay	NN
tulang	Malay	NN	ubat	Malay	NN
tuli	Malay	ADJ	ubun	Arabic	NN
tulis	Malay	VB	ucap	Malay	NN
tulus	Malay	ADJ	udara	Malay	NN
tumbang	Malay	VB	uhud	Arabic	NN
tumbuh	Malay	VB	uji	Malay	VB
tumbuk	Malay	VB	ukur	Malay	NN
tumpah	Malay	VB	ulama	Arabic	NN
tumpang	Malay	VB	ulang	Malay	VB
tumpas	Malay	VB	ular	Malay	NN
tumpu	Malay	VB	ulat	Malay	NN
tunai	Malay	VB	umat	Malay	NN
tunas	Malay	NN	umbi	Malay	NN
tunda	Malay	VB	ummi	Arabic	NN
tunduk	Malay	VB	ummiyyin	Arabic	NN
tunggang	Malay	VB	ummul	Arabic	NN
tunggu	Malay	VB	Ummul qura	Arabic	NN
tungku	Malay	NN	umpama	Malay	FT
tunjuk	Malay	VB	umpat	Malay	VB
tuntut	Malay	VB	umrah	Arabic	NN
Tursina	Arabic	NN	umum	Malay	ADJ
turun	Malay	VB	umur	Malay	NN
turut	Malay	ADJ	undang	Malay	VB
tutup	Malay	VB	undi	Malay	VB
tutur	Malay	VB	undur	Malay	VB
tuwa	Malay	NN	unggun	Malay	NN

WORD	LANGUAGE	TAG	WORD	LANGUAGE	TAG
ungkap	Malay	VB	wajib	Malay	ADJ
unsur	Malay	NN	waktu	Malay	NN
unta	Malay	NN	walhal	Malay	FT
untuk	Malay	FT	wali	Malay	NN
untung	Malay	NN	wang	Malay	NN
unzurna	Arabic	NN	waris	Malay	NN
upah	Malay	NN	warna	Malay	NN
upaya	Malay	NN	wasiat	Malay	NN
urai	Malay	NN	wasilah	Malay	NN
urat	Malay	NN	waspada	Malay	ADJ
urus	Malay	VB	wazir	Malay	NN
usaha	Malay	NN	wenang	Malay	NN
usap	Malay	VB	wujud	Malay	VB
usia	Malay	NN	wusta	Arabic	NN
usir	Malay	VB	yaakub	Arabic	NN
usul	Malay	NN	Yagus	Arabic	NN
usus	Malay	NN	yahudi	Malay	NN
utama	Malay	ADJ	yahya	Arabic	NN
utara	Malay	VB	yakin	Malay	ADJ
utus	Malay	VB	yakjuj	Arabic	NN
uzair	Malay	NN	yala	Arabic	NN
uzur	Malay	ADJ	yaman	Arabic	NN
uzza	Arabic	NN	yathrib	Arabic	NN
Wadd	Arabic	NN	yatim	Malay	NN
wafat	Malay	NN	Ya'uq	Arabic	NN
wahyu	Malay	NN	yunus	Arabic	NN
wajah	Malay	NN	yusuf	Arabic	NN
wajar	Malay	ADJ	zabaniyah	Arabic	NN

WORD	LANGUAGE	TAG	WORD	LANGUAGE	TAG
zabur	Arabic	NN	zarah	Malay	NN
zahir	Malay	NN	zihar	Arabic	NN
zaid	Arabic	NN	zikir	Malay	NN
zainab	Arabic	NN	Az-zikr	Arabic	NN
zaitun	Arabic	NN	zina	Malay	NN
zakaria	Arabic	NN	zohor	Arabic	NN
zakat	Malay	NN	zubur	Arabic	NN
zalim	Malay	ADJ	zulkarnain	Arabic	NN
zaman	Malay	NN	zulkifli	Arabic	NN
zaqqum	Arabic	NN	zuriat	Malay	NN

Appendix D

List of Relevant Results

Query No	Chapter/Verse	Query No	Chapter/Verse	Query No	Chapter/Verse
Q1	4:95	Q3	55:15	Q21	2:208
Q1	5:65	Q3	55:33	Q21	2:268
Q1	6:127	Q3	55:39	Q21	2:275
Q1	9:72	Q3	55:56	Q21	3:36
Q1	10:9	Q3	55:74	Q21	3:155
Q1	10:25	Q3	72:1	Q21	3:175
Q1	10:26	Q3	72:5	Q21	4:38
Q1	13:38	Q3	72:6	Q21	4:60
Q1	13:23	Q3	114:6	Q21	4:76
Q1	13:29	Q4	2:34	Q21	4:83
Q1	16:30	Q4	7:11	Q21	5:90
Q1	16:31	Q4	7:12	Q21	5:91
Q1	16:62	Q4	7:27	Q21	6:43
Q1	18:31	Q4	17:61	Q21	6:68
Q1	18:107	Q4	18:50	Q21	6:71
Q1	19:61	Q4	20:116	Q21	6:112
Q1	20:76	Q4	20:117	Q21	6:121
Q1	21:101	Q5	28:76	Q21	6:142
Q1	22:56	Q5	28:79	Q21	7:20
Q1	23:11	Q5	29:39	Q21	7:22

Q1	25:15
Q1	28:83
Q1	31:8
Q1	32:19
Q1	33:47
Q1	35:33
Q1	35:35
Q1	37:43
Q1	38:50
Q1	40:8
Q1	40:39
Q1	41:50
Q1	42:22
Q1	43:35
Q1	56:12
Q1	56:27
Q1	56:38
Q1	56:90
Q1	56:91
Q1	61:12
Q1	68:38
Q1	69:22
Q1	83:19
Q1	83:20
Q1	83:21
Q1	83:22
Q1	88:10
Q1	92:6
Q1	98:8
Q2	2:24
Q2	2:119
Q2	2:126
Q2	2:206
Q5	40:24
Q6	7:59
Q6	7:61
Q6	9:70
Q6	10:71
Q6	10:72
Q6	10:73
Q6	11:25
Q6	11:29
Q6	11:89
Q6	25:37
Q6	26:105
Q6	40:31
Q6	40:5
Q6	50:12
Q6	51:46
Q6	53:52
Q6	54:9
Q7	28:23
Q7	28:25
Q7	28:26
Q7	28:27
Q7	28:28
Q8	3:59
Q8	4:1
Q8	7:11
Q8	7:12
Q8	15:26
Q8	15:28
Q8	15:29
Q8	17:61
Q8	38:174
Q8	38:72
Q21	7:27
Q21	7:30
Q21	7:175
Q21	7:200
Q21	7:201
Q21	8:11
Q21	8:48
Q21	12:5
Q21	12:42
Q21	12:100
Q21	14:22
Q21	15:17
Q21	16:63
Q21	16:98
Q21	17:27
Q21	17:53
Q21	17:64
Q21	18:63
Q21	19:44
Q21	19:45
Q21	19:68
Q21	19:83
Q21	20:120
Q21	21:82
Q21	22:3
Q21	22:52
Q21	22:53
Q21	23:97
Q21	24:21
Q21	25:29
Q21	26:210
Q21	26:221
Q21	27:24

Q2	3:12
Q2	3:162
Q2	3:197
Q2	4:10
Q2	4:55
Q2	5:10
Q2	5:86
Q2	7:145
Q2	8:16
Q2	9:73
Q2	9:113
Q2	11:98
Q2	13:18
Q2	13:25
Q2	14:28
Q2	14:29
Q2	22:4
Q2	22:51
Q2	22:72
Q2	24:57
Q2	25:11
Q2	26:91
Q2	30:10
Q2	31:21
Q2	33:64
Q2	35:6
Q2	37:163
Q2	37:23
Q2	37:55
Q2	37:64
Q2	37:68
Q2	37:62
Q2	37:97
Q9	6:73
Q9	18:99
Q9	20:102
Q9	23:101
Q9	27:87
Q9	36:49
Q9	36:51
Q9	37:19
Q9	39:68
Q9	50:20
Q9	50:42
Q9	69:13
Q9	74:80
Q9	78:18
Q9	79:6
Q9	79:13
Q9	80:33
Q10	12:035
Q10	12:045
Q10	12:042
Q10	12:036
Q10	12:039
Q10	12:041
Q10	12:046
Q11	6:22
Q11	11:103
Q11	17:97
Q11	19:85
Q11	80:37
Q12	7:80
Q12	7:82
Q12	11:70
Q12	11:74
Q21	28:15
Q21	29:38
Q21	31:21
Q21	35:6
Q21	36:60
Q21	37:7
Q21	37:65
Q21	38:37
Q21	38:41
Q21	4:117
Q21	4:119
Q21	4:120
Q21	41:36
Q21	43:36
Q21	43:62
Q21	47:25
Q21	58:10
Q21	58:19
Q21	59:16
Q21	67:5
Q21	81:25
Q22	2:97
Q22	2:98
Q22	2:102
Q22	6:73
Q22	11:69
Q22	11:77
Q22	15:52
Q22	20:102
Q22	23:101
Q22	32:11
Q22	40:7
Q22	40:8

Q2	38:56	Q12	11:89	Q22	43:77
Q2	38:60	Q12	15:67	Q22	66:4
Q2	40:7	Q12	21:74	Q22	74:30
Q2	40:52	Q12	22:43	Q22	96:18
Q2	41:28	Q12	25:40	Q23	20:10
Q2	42:13	Q12	26:160	Q23	27:7
Q2	42:8	Q12	27:54	Q23	28:29
Q2	44:47	Q12	27:56	Q24	2:164
Q2	44:56	Q12	29:28	Q24	6:38
Q2	44:43	Q12	29:31	Q24	8:22
Q2	52:18	Q12	29:34	Q24	8:55
Q2	52:27	Q12	53:53	Q24	11:6
Q2	54:47	Q12	54:33	Q24	11:56
Q2	54:48	Q12	69:9	Q24	16:49
Q2	56:94	Q13	6:73	Q24	16:61
Q2	56:52	Q13	20:102	Q24	22:18
Q2	57:15	Q13	23:101	Q24	24:45
Q2	57:19	Q14	12:30	Q24	27:82
Q2	58:8	Q14	12:51	Q24	29:60
Q2	64:10	Q15	6:99	Q24	31:10
Q2	66:9	Q15	7:58	Q24	34:14
Q2	67:5	Q15	10:24	Q24	35:28
Q2	67:6	Q15	18:45	Q24	35:45
Q2	67:10	Q15	20:53	Q24	42:29
Q2	67:11	Q15	57:20	Q24	45:4
Q2	69:31	Q15	78:15	Q25	2:57
Q2	70:15	Q16	2:65	Q25	20:80
Q2	73:12	Q16	4:47	Q25	7:160
Q2	74:26	Q16	7:163	Q26	2:63
Q2	74:27	Q17	2:173	Q26	2:93
Q2	76:4	Q17	5:3	Q26	4:154
Q2	79:36	Q17	5:60	Q26	7:143
Q2	79:39	Q17	6:145	Q26	7:171

Q2	79:10	Q17	16:115	Q26	19:52
Q2	79:14	Q18	12:8	Q26	20:80
Q2	81:1	Q18	12:59	Q26	23:20
Q2	81:2	Q18	12:63	Q26	28:29
Q2	82:14	Q18	12:64	Q26	28:46
Q2	83:16	Q18	12:70	Q26	52:1
Q2	83:7	Q18	12:76	Q26	95:2
Q2	101:9	Q18	12:87	Q27	2:50
Q2	102:6	Q19	18:94	Q27	7:136
Q2	104:4	Q19	21:96	Q27	7:138
Q2	104:5	Q20	7:73	Q27	10:90
Q3	6:100	Q20	9:70	Q27	20:77
Q3	6:112	Q20	11:61	Q27	20:78
Q3	6:128	Q20	11:68	Q27	26:63
Q3	6:130	Q20	11:95	Q27	28:40
Q3	7:179	Q20	14:9	Q27	44:24
Q3	7:38	Q20	17:59	Q27	51:40
Q3	8:56	Q20	22:42	Q28	6:99
Q3	9:12	Q20	25:38	Q28	6:141
Q3	9:13	Q20	26:141	Q28	55:68
Q3	9:111	Q20	27:45	Q29	7:133
Q3	11:119	Q20	29:38	Q29	54:7
Q3	15:2	Q20	38:13	Q30	2:133
Q3	15:27	Q20	40:31	Q30	2:136
Q3	17:88	Q20	41:13	Q30	2:140
Q3	18:50	Q20	41:17	Q30	3:84
Q3	27:17	Q20	50:12	Q30	4:163
Q3	27:39	Q20	51:43	Q30	6:84
Q3	32:13	Q20	53:51	Q30	11:71
Q3	34:12	Q20	54:23	Q30	12:38
Q3	34:14	Q20	69:4	Q30	12:6
Q3	34:41	Q20	69:5	Q30	14:39
Q3	37:158	Q20	85:18	Q30	19:49

Q3	41:25	Q20	89:9	Q30	21:72
Q3	41:29	Q20	91:11	Q30	29:27
Q3	46:18	Q21	2:14	Q30	37:112
Q3	46:29	Q21	2:36	Q30	37:113
Q3	51:56	Q21	2:102	Q30	38:45
Q3	55:1	Q21	2:168		

Appendix E

Retrieved and Relevant Results

A) Keyword-based Search Results

Query (Malay)	**Query (English)**	**Query No**	**No of Relevant Document**	**Qurany Search System (English)**	**MyQOS Search System (Malay)**
Apa itu Syurga?	What is Paradise?	Q1	49	238	177
Apa itu Neraka?	What is Hell?	Q2	81	312	269
Apa itu Jin?	What is Jinn?	Q3	37	152	63
Syaitan enggan sujud kepada Nabi Adam.	Satan refused to bow to Adam	Q4	8	15	3
Cerita tentang Qarun	Story about Qarun	Q5	4	5	4
Ayat berkaitan kaum Nabi Nuh	Verse related to People of Noah	Q6	17	66	32
Cerita mengenai orang tua kaum Madyan	Story about old man of Madian	Q7	5	1	2
Maklumat tentang kejadian manusia iaitu Nabi Adam.	Information on human creation Adam.	Q8	10	41	11
Ayat berkaitan tiupan sangkakala	Verse related to The Trumpet is blown	Q9	17	21	13

Kisah Lelaki bersama Nabi Yusuf di penjara	Story about man with Joseph who in prison	Q10	7	3	4
Keterangan Hari Mahsyar iaitu hari perhimpunan.	Evidence of the Gathering day is the day of assembly.	Q11	5	1	5
Ayat berkaitan kaum Nabi Lut	Verse related to People of Lot	Q12	18	30	44
Siapa itu Israfil?	Who is Israfil?	Q13	3	1	0
Apa yang terjadi pada isteri Al-Aziz?	What happened to wife of Aziz?	Q14	2	3	3
Cerita berkaitan dengan penciptaan tumbuhan.	Stories related to the creation of plants.	Q15	7	35	15
Apa itu hari Sabat?	What is a Sabbath?	Q16	3	5	3
Tegahan memakan Babi.	Prohibition of eating pig.	Q17	5	5	6
Siapa itu Bunyamin?	Who is Bunyamin?	Q18	7	6	4

Kisah Yakjuj and Makjuj	Stories about Mog and Magog	Q19	2	6	4
Ayat berkaitan kaum Tsamud.	Verse related to People of Thamud.	Q20	25	30	26
Kisah syaitan	Story about satan	Q21	78	132	103
Malaikat-malaikat yang disebut didalam Al-Qur'an	The angels mentioned in the Qur'an	Q22	16	200	150
Ayat yang menceritakan berkaitan dengan keluarga Nabi Musa a.s	The verse which relates to the family of Moses a.s	Q23	19	224	214
Kisah Nabi Daud a.s	The story of Prophet David a.s	Q24	18	19	17
Apa itu Manna dan Salwa	What is Manna and Salwa	Q25	3	4	2
Gunung Sinai	Mount Sinai	Q26	12	11	8
Ayat berkaitan dengan Laut Merah	The verse related with the Red Sea	Q27	10	69	1
Khasiat buah delima	Benefits of pomegranate	Q28	3	3	3

Apa ayat Al-Quran yang menceritakan tentang belalang	What is the Qur'anic verse that tells the locusts?	Q29	2	2	2
Kisah Nabi Ishak a.s	Story about Prophet Isaac a.s	Q30	16	17	18

B) Keyword-based Search Only and Keyword-based search with Stemming Algorithm

Query (Malay)	**Query (English)**	**Query No**	**No of Relevant Document**	**MyQOS : Keyword Search Only**	**MyQOS : Keyword Search + Stemming**
Query (Malay)	Query (English)	Query No	No of Relevant Document	MyQOS : Keyword Search Only	MyQOS : Keyword Search + Stemming
Apa itu Syurga?	What is Paradise?	Q1	49	177	177
Apa itu Neraka?	What is Hell?	Q2	81	269	269
Apa itu Jin?	What is Jinn?	Q3	37	63	63
Syaitan enggan sujud kepada Nabi Adam.	Satan refused to bow to Adam	Q4	8	3	12
Cerita tentang Qarun	Story about Qarun	Q5	4	4	4

Ayat berkaitan kaum Nabi Nuh	Verse related to People of Noah	Q6	17	32	32
Cerita mengenai orang tua kaum Madyan	Story about old man of Madian	Q7	5	2	2
Maklumat tentang kejadian manusia iaitu Nabi Adam.	Information on human creation Adam.	Q8	10	11	11
Ayat berkaitan tiupan sangkakala	Verse related to The Trumpet is blown	Q9	17	13	13
Kisah Lelaki bersama Nabi Yusuf di penjara	Story about man with Joseph who in prison	Q10	7	4	4
Keterangan Hari Mahsyar iaitu hari perhimpunan.	Evidence of the Gathering day is the day of assembly.	Q11	5	5	5
Ayat berkaitan kaum Nabi Lut	Verse related to People of Lot	Q12	18	44	17
Siapa itu Israfil?	Who is Israfil?	Q13	3	0	0

Apa yang terjadi pada isteri Al-Aziz?	What happened to wife of Aziz?	Q14	2	3	3
Cerita berkaitan dengan penciptaan tumbuhan.	Stories related to the creation of plants.	Q15	7	15	15
Apa itu hari Sabat?	What is a Sabbath?	Q16	3	3	3
Tegahan memakan Babi.	Prohibition of eating pig.	Q17	5	6	6
Siapa itu Bunyamin?	Who is Bunyamin?	Q18	7	4	4
Kisah Yakjuj and Makjuj	Stories about Mog and Magog	Q19	2	4	4
Ayat berkaitan kaum Tsamud.	Verse related to People of Thamud.	Q20	25	26	25
Kisah syaitan	Story about satan	Q21	78	103	103
Malaikat-malaikat yang disebut didalam Al-Qur'an	The angels mentioned in the Qur'an	Q22	16	150	150

Ayat yang menceritakan berkaitan dengan keluarga Nabi Musa a.s	The verse which relates to the family of Moses a.s	Q23	19	214	6
Kisah Nabi Daud a.s	The story of Prophet David a.s	Q24	18	17	17
Apa itu Manna dan Salwa	What is Manna and Salwa	Q25	3	2	2
Gunung Sinai	Mount Sinai	Q26	12	8	8
Ayat berkaitan dengan Laut Merah	The verse related with the Red Sea	Q27	10	1	1
Khasiat buah delima	Benefits of pomegranate	Q28	3	3	3
Apa ayat Al-Quran yang menceritakan tentang belalang	What is the Qur'anic verse that tells the locusts?	Q29	2	2	2
Kisah Nabi Ishak a.s	Story about Prophet Isaac a.s	Q30	16	18	18

C) Keyword-based Search Only, Keyword-based search with Stemming and Semantic Search

Query (Malay)	Query (English)	Query No	No of Relevant Document	MyQOS : Keyword Search Only	MyQOS : Keyword Search + Stemming	MyQOS : Semantic Search
Apa itu Syurga?	What is Paradise?	Q1	49	177	177	50
Apa itu Neraka?	What is Hell?	Q2	81	269	269	82
Apa itu Jin?	What is Jinn?	Q3	37	63	63	33
Syaitan enggan sujud kepada Nabi Adam.	Satan refused to bow to Adam	Q4	8	3	12	7
Cerita tentang Qarun	Story about Qarun	Q5	4	4	4	5
Ayat berkaitan kaum Nabi Nuh	Verse related to People of Noah	Q6	17	32	32	12
Cerita mengenai orang tua kaum Madyan	Story about old man of Madian	Q7	5	2	2	6
Maklumat tentang kejadian manusia iaitu Nabi Adam.	Information on human creation Adam.	Q8	10	11	11	11

Ayat berkaitan tiupan sangkakala	Verse related to The Trumpet is blown	Q9	17	13	13	17
Kisah Lelaki bersama Nabi Yusuf di penjara	Story about man with Joseph who in prison	Q10	7	4	4	5
Keterangan Hari Mahsyar iaitu hari perhimpunan.	Evidence of the Gathering day is the day of assembly.	Q11	5	5	5	5
Ayat berkaitan kaum Nabi Lut	Verse related to People of Lot	Q12	18	44	17	18
Siapa itu Israfil?	Who is Israfil?	Q13	3	0	0	3
Apa yang terjadi pada isteri Al-Aziz?	What happened to wife of Aziz?	Q14	2	3	3	3
Cerita berkaitan dengan penciptaan tumbuhan.	Stories related to the creation of plants.	Q15	7	15	15	7
Apa itu hari Sabat?	What is a Sabbath?	Q16	3	3	3	3

Tegahan memakan Babi.	Prohibition of eating pig.	Q17	5	6	6	5
Siapa itu Bunyamin?	Who is Bunyamin?	Q18	7	4	4	7
Kisah Yakjuj and Makjuj	Stories about Mog and Magog	Q19	2	4	4	2
Ayat berkaitan kaum Tsamud.	Verse related to People of Thamud.	Q20	25	26	25	25
Kisah syaitan	Story about satan	Q21	78	103	103	25
Malaikat-malaikat yang disebut didalam Al-Qur'an	The angels mentioned in the Qur'an	Q22	16	150	150	16
Ayat yang menceritakan berkaitan dengan keluarga Nabi Musa a.s	The verse which relates to the family of Moses a.s	Q23	19	214	6	19

Kisah Nabi Daud a.s	The story of Prophet David a.s	Q24	18	17	17	18
Apa itu Manna dan Salwa	What is Manna and Salwa	Q25	3	2	2	3
Gunung Sinai	Mount Sinai	Q26	12	8	8	12
Ayat berkaitan dengan Laut Merah	The verse related with the Red Sea	Q27	10	1	1	10
Khasiat buah delima	Benefits of pomegranate	Q28	3	3	3	3
Apa ayat Al-Quran yang menceritakan tentang belalang	What is the Qur'anic verse that tells the locusts?	Q29	2	2	2	2
Kisah Nabi Ishak a.s	Story about Prophet Isaac a.s	Q30	16	18	18	16

Appendix F

Survey Form : Semantic relationship between synonyms

Semantic relationship between synonyms

Semantics is the study of meaning in language. Synonyms are words with similar meanings.They are listed in a special type of dictionary called a thesaurus.

This assessment is to find the semantic similarity between words using synonym. Please tick (/) for each word based on their relevance to synonyms. The following is the meaning for each of the indicators.

Extremely Far : The meaning of this word is completely different.
Quite Far : The meaning of this word is quite different.
Slightly far : The meaning of this word is slightly different.
Neither: The meaning of this word is neither different or same.
Slightly Close : The meaning of this word is slightly match.
Quite Close: The meaning of this word is quite match.
Extremely Close : The meaning of this word is extremely match.

Thank you

* Required

1. **Gender** *

Mark only one oval.

○ Female

○ Male

2. **Synonym of Bandar** *

Check all that apply.

	Extremely different	Quite different	Slightly different	Neither	Slightly match	Quite match	Extremely match
Kota	□	□	□	□	□	□	□
Pekan	□	□	□	□	□	□	□
Bandaraya	□	□	□	□	□	□	□
Kota raya	□	□	□	□	□	□	□

3. **Synonym of Syurga** *

Check all that apply.

	Extremely different	Quite different	Slightly different	Neither	Slightly match	Quite match	Extremely match
Darussalam	□	□	□	□	□	□	□
Firdaus	□	□	□	□	□	□	□
Kebun	□	□	□	□	□	□	□
Nirwana	□	□	□	□	□	□	□

4. **Synonym of Neraka** *

Check all that apply.

	Extremely different	Quite different	Slightly different	Neither	Slightly match	Quite match	Extremely match
Jahanam	□	□	□	□	□	□	□
Seksaan	□	□	□	□	□	□	□
Azab api	□	□	□	□	□	□	□
Abyss	□	□	□	□	□	□	□

	Extremely different	Quite different	Slightly different	Neither	Slightly match	Quite match	Extremely match
Ganang	□	□	□	□	□	□	□
Jabal	□	□	□	□	□	□	□
Cenuram	□	□	□	□	□	□	□
Banjaran	□	□	□	□	□	□	□
Kemuncak	□	□	□	□	□	□	□

6. **Synonym of Akhirat** *

Check all that apply.

	Extremely different	Quite different	Slightly different	Neither	Slightly match	Quite match	Extremely match
Alam Baqa	□	□	□	□	□	□	□
Alam yang kekal	□	□	□	□	□	□	□
Hari Terakhir	□	□	□	□	□	□	□
Rumah Terakhir	□	□	□	□	□	□	□