Lexical Scoring System of Lexical Chain for Quranic Document Retrieval

Hamed Zakeri Rad
Jerald0000030@yahoo.com
Faculty of Information Science and Technology,
Universiti Kebangsaan Malaysia

Sabrina Tiun
sabrinatiun@ukm.edu.my
Faculty of Information Science and Technology,
Universiti Kebangsaan Malaysia

Saidah Saad
saidah@ukm.edu.my
Faculty of Information Science and Technology,
Universiti Kebangsaan Malaysia

ABSTRACT

An Information Retrieval (IR) system aims to extract information based on a query made by a user on a particular subject from an extensive collection of text. IR is a process through which information is retrieved by submitting a query by a user in the form of keywords or to match words. In the Al-Quran, verses of the same or comparable topics are scattered throughout the text in different chapters, and it is therefore difficult for users to remember the many keywords of the verses. Therefore, in such situations, retrieving information using semantically related words is useful. In well-composed documents, the semantic integrity of the text (coherence) exists between the words. Lexical cohesion is the results of chains of related words that contribute to the continuity of the lexical meaning found within the text are a direct result of text being about the same thing (i.e. topic, etc.). This indicates that using an IR system and lexical chains are a useful and appropriate method for representing documents with concepts rather than using terms in order to have successful retrieval based on semantic relations. Therefore, a new Lexical Scoring System is proposed in this study, in addition to determining the semantic relation that exists between words whereby WordNet was used as the semantic knowledge base. The proposed scoring system helped to retrieve 86.58% of the total relevant documents in the Al-Quran based on the relevance judgment, using the lexical chain approach. Based on the findings, the study concludes that, the proposed approach on representing verses using lexical chains is appropriate and suitable for a Quranic IR system.

Keywords: lexical chain; information retrieval (IR); semantic retrieval; lexical scoring system; Quranic semantic retrieval system

INTRODUCTION

Information retrieval (IR), according to Manning, Raghavan and Schütze (2009), refers to the activity of acquiring the resources of information which are related to the information that one requires from a broader source of information. Accordingly, this indicates how one can utilise the information need from the information retrieval system during the search. Presumably, one can imagine the vast amount of information stored in a specific area, which one is specifically entitled to the query. For instance, if a user, who is uncertain about a particular topic, intends to find specific information in the Al-Quran, he/she must request that information from the retrieval system. Thus, the way in which the verses are determined and

eISSN: 2550-2131
ISSN: 1675-8021

www.manaraa.com

retrieved is what Quranic IR is concerned with, implying that the information requested by the user is related to the retrieved verses.

The Quranic IR initiates the process of retrieving information once a query is entered into the system by any user. Queries refer to the formal statements of information that a user requires for a specific task and is mostly in the form of keywords. Emphatically, a query does not identify a single title and object in the Al-Quran. The object refers to the entity stored in the database related to the information. When a query is entered by the user, several different titles and objects are revealed which can be relevant or irrelevant. This is called matching, which signifies the matching of the queries and the related verses. Figure 1 displays the Quranic IR system.

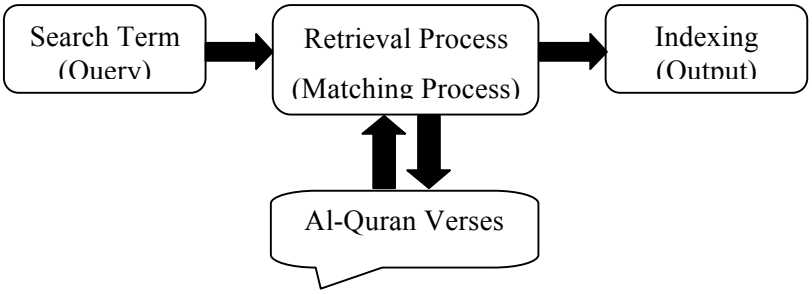

FIGURE 1. General Quranic IR system architecture

Retrieving information based on semantic shows better result than the exact word. Lexical chain is a method to represent the documents based on semantic. Every lexical chain algorithm introduced in the literature has some form of scoring system to determine the level of importance that the created chain has. Jarmasz and Szpakowicz (2012) built lexical chains using a computerised version of the 1987 edition of Penguin's Roget's Thesaurus (Roget, 1977) of English Words and Phrases. To select the candidate words, they used two relations regarding the 1987 Roget's structure; repetition of the same word and words in the same Head. Ruas and Grosky (2017) explore lexical chain structure using WordNet (Miller & Fellbaum, 1998), proposing a different algorithm. Hirst and St-Onge (1998) used all words that appear in the noun section of WordNet as candidate words. Indeed, this means that all the nouns in a document are considered candidates, as well as all words that have a noun form. Further, Barzilay and Elhadad (1999) used a small number of simple semantic relations to build lexical chains; reiteration, synonymy, antonymy, hypernymy/hyponymy, and holonymy/meronymy. In this case, each relation is assigned a weight, where a greater weight indicates a stronger semantic relation. In a separate study, Enss (2006) used the score function to measure the strength of the relation between any two terms in a document. Accordingly, the score function used by Enss (2006) is the same as the score function used by Silber and Mccoy (2000, 2000, 2002).

Scoring the words in the chain more accurately can lead to create more accurate IR system. In this study, a new scoring system based on the lexical chain will be proposed to index retrieved information more accurately.

LEXICAL COHESION

When reading a text, it is apparent that it is not merely made up of a set of unrelated sentences, but these sentences are in fact connected to each other using two linguistic phenomena's; cohesion and coherence. Cohesion relates to the fact that text elements tend to 'hang together'(Halliday & Hasan 2014), while coherence refers to the fact that there is a

eISSN: 2550-2131
ISSN: 1675-8021

www.manaraa.com

'sense (or intelligibility)' in a text. Lexical cohesion, on the other hand, is the result of chains of related words that contribute to the continuity of the lexical meaning, or in other words, lexical cohesion is the cohesion that arises from semantic relationships between words. In addition to that, the lexical cohesion also arises from a writer's chosen of words surrounding the topic to which the author is working on (Haris & Yunus, 2014). These lexical chains are a direct result of units of text being about the same thing, whereas, finding a text structure involves finding the units of text about the same thing.

In any text, lexical chains are said to be the related words that continuously contribute to the lexical meaning. Accordingly, the chains are about the units of text which are indicative of the same category. Therefore, if a person intends to find the text structure, it may mean finding the same thing. Morris and Hirst (1991) were the first researchers to suggest the use of lexical chains to determine the structure of texts, as a general example, consider {house, kitchen, home, building, room} as a lexical chain. In this example 'house' and 'home' are synonyms; 'kitchen' and 'room' are meronym of 'house' or in another words 'house' is a holonym of 'kitchen' and 'room'; while 'building' is a hypernym of 'house'; and in the 'house' holonym tree 'kitchen' and "room" are siblings. In this regard, the chains and text structure are correspondents which indicate the significance of the text structure for its meaning. Further, lexical chains identify the relation between words by means of identifying the cohesion links. Since lexical chains and semantics are directly related, we can, therefore, call two parts of a text related if they both have the same words and are related based on semantics. In this way, one can assume that the two pieces of text are lexically related, not only if they use the same words, but also if they use semantically related words. This is a way to identify the specific structure of a text through considering the content distribution. The following sentences are examples, explaining how these sentences can be related semantically:

Example 1: Using the same noun: "*The garden is near my house. The garden is full of trees.*" The noun 'garden' which is used in two mentioned sentences is the same.

Example 2: Using synonyms: "*The child that reads a book. That kid runs fast.*" Two noun instances 'child' and 'kid' are synonyms.

Example 3: Using hypernym/hyponym relation between two sentences: "*I have a toy. It is a doll.*" In this example, 'doll' is a hyponym of 'toy' or in another word 'toy' is a hypernym of 'doll'.

Example 4: Using siblings: "*Toyota is fast. Ferrari is faster.*" In this example, 'Toyota' and 'Ferrari' are siblings in their hypernymy tree. Both are hyponym of 'car'.

LEXICAL CHAIN

Conceptually, sequences of related words regarding cohesion and semantic relations are what Morris (1991) termed as lexical chains. A word in the chain and the words that co-occur on a given span are to some extent related. While lexical chains are not constrained by sentence boundaries but can connect a pair of adjacent words or span over an entire text.

There are two major reasons for the significance of lexical chaining computational text understanding systems.

A: Lexical chains offer an easily determined context to resolve ambiguity and in narrowing to a specific meaning of a word.

B: Lexical chains offer a clue to determine the coherence and discourse structure, and thus the larger meaning of the text.

eISSN: 2550-2131
ISSN: 1675-8021

www.manaraa.com

Lexical chains offer a limited and easily determined representation of context to consider semantic distance and offer a clue for determining coherence and discourse structure. When a chunk of text forms a unit in discourse, their related words tend to be utilised; therefore, if lexical chains can be established, they are likely to show the structure of the text. Discourse structure and cohesion are closely connected. Further, there can be apparent cohesion in a sequence of sentences but with no coherence. Similarly, a set of sentences can have coherence without apparent cohesion. In general, cohesion is evident when sentences are related coherently, and this trait can be taken advantage of by utilising cohesion relations to identify coherent parts of the text. Five types of cohesion relationships were defined by Halliday and Hasan (2014):

- **Conjunction:** Usage of conjunctive structures like 'and' to present two facts cohesively. In the sentence, "*I have a dog, and his name is Wolfy*," two facts are connected with the conjunctive 'and'.
- **Reference:** Usage of pronouns for entities. In the example, "*Tom lives in Moscow. He is a businessman.*" the pronoun 'he' in the second sentence refers to 'Tom' in the first sentence.
- **Lexical Cohesion:** Usage of related words. In the example sentence, "*Earth is the third Planet from the Sun*," the words 'Earth', 'Planet' and 'Sun' are semantically related.
- **Substitution:** Using an indefinite article for a noun. In the example, "*As soon as Tom was given a chocolate ice cream cup, Lucy wanted one too.*" The word 'one' refers to the phrase 'chocolate ice cream cup.'
- **Ellipsis:** Implying noun without repeating. In the example sentence, "*Do you have a pen? No, I don't.*" The word 'pen' is implied without repeating in the second sentence.

From these cohesion structures, lexical cohesion is the most definite and easiest to find. Thus, like most of the research on cohesion, this current study focuses on lexical cohesion. Moreover, cohesion is based on the relationships between units of the document, and in the case of lexical cohesion, these units are words and phrases. The phrases in a document should be semantically related, and this is called lexical cohesion. Forming the lexical cohesion depends on determining the semantic relationships between words which are understood by humans and quickly recognised if the vocabulary used is familiar to the learner. There are five different types of commonly used lexical cohesion: synonymy; generalisation and specialisation (Hypernym / Hyponym); whole-part and part-whole (Meronymy / Holonymy). The semantic relationships used in this study are synonymy and similarly; hyponymy and hypernymy; meronymy and holonymy; siblings, which are divided into two groups; close siblings and distant siblings, and finally verb conjugation, which will be briefly explained, in the following section along with examples.

- **Synonyms and Similar:** These terms are used when one intends to refer to two or more words which have similar meanings.
- **Hyponymy and Hypernymy:** When words or phrases have a common or similar field of meaning, they are linguistically and semantically labelled as such. Thus, it can be said that the two hyponyms have the same relationship together. For example, 'sparrow', 'parrot' and 'eagle' are the hyponyms of 'bird'.
- **Meronymy and Holonymy:** These terms specify the semantic relation among words. So, a meronym means that two or more words are members of something or they constitute one part. This can be clarified in this example. If 'A' is a meronym of 'B', therefore 'A' is part of 'B'; or 'A' is a meronym of 'B' if 'A' is a member of 'B'. In other words, 'toe' is a meronym of 'foot', or a 'toe' is part of a 'foot'. In the same way, 'roof' is a meronymy of

eISSN: 2550-2131
ISSN: 1675-8021

www.manaraa.com

'house'. Knowing this, it can be said that meronymy means 'part of' while Holonymy signifies the opposite. For example, 'foot' is Holonymy of 'toe' and 'house' is Holonymy of 'roof'.

- **Siblings:** When two words have the same Holonymy/Hypernymy, they are called siblings. According to this definition, for example, 'arm' and 'leg' are siblings in their holonymy tree, which is 'body'.
- **Verb Conjugation:** Conjugation is a set of various inflectional forms of a verb. Different forms of the same verb are used depending on the situation or time. For example, we can change the verb 'walk' to 'walks' to 'walked,' and so on.

RELATED WORKS

According to Muslim belief, the Al-Quran is the holy book of God which was revealed to the Prophet Muhammad (peace be upon him). The Al-Quran provides a complete code of life for Muslims to live a life that pleases God. It is a fundamental Muslim belief that success in the afterlife depends on adhering to the commandments found in the Al-Quran. In today's world of advancing knowledge, knowledge of the Quran is not just an area of interest for Muslims but has significantly attracted the interest and attention of millions of people from other faiths as well (Iqbal et al., 2013).

In recent years, workings on the Al-Quran have grown among Muslim researchers at both syntactic and semantic levels. Shoaib, Yasin, Hikmat and Khiyal (2009) proposed a model for semantic search in the Al-Quran, which means using word sense disambiguation to identify only one sense for the query term and using synonymy to perform a search against each and every synonym of the sense. The proposed model was implemented on Al-Baqarah, the longest chapter of the Al-Quran. In the work of Al-Ghafour, Awal, Zainuddin and Aladdin (2017), they investigated the near-synonyms words in Al-Quran in terms of its denotative and expressive meaning. Whereas, in Mohamed and Tiun (2015) and Tiun, Zakr, Mohd, Abidin and Hisham (2013), investigated the use of word sense for Al-Quran IR system.

On the other hand, Yauri, Kadir, Azman and Murad (2013) proposed a model that uses semantics to model the Al-Quran domain knowledge using ontology which consists of statements that define concepts, relationships and constraints. Moreover, their proposed system comprises the Al-Quran concepts semantic search model. Abdelnasser et al. (2014) proposed a new Question Answering (QA) system named Al-Bayan for the Al-Quran, which aims at understanding the semantics of the Al-Quran and answering users' questions using reliable Quranic resources. Whereas, Alrehaili and Atwell (2014) investigated a range of existing studies on the Al-Quran ontology. In a separate study, Ayed and Atwell (2017) also used question and answering system where they applied their model to the Al-Quran written in the Classical Arabic language. A Quranic QA system of Hamed and Aziz (2016) used the well-known semantic knowledge base (WordNet) to expand questions, in order, to increase the number of relevant answers which can help in increasing the number of accurate answers.

Ta'a, Abed and Ahmad (2017) used the semantic search approach for searching knowledge of the Al-Quran by developing ontology for the Al-Quran. Indeed, their research employed an ontology approach representing concepts of the Al-Quran that can be classified and organised according to a specific theme. Not with standing, the proposed approach consists of two sub-stages: development of the Al-Quran ontology, and development of a semantic searching method. According to Ta'a, Abed and Ahmad (2017), their method facilitates searching in the Al-Quran ontology by using a semantic approach. Yunus, Mustapha and Samsudin (2017) conducted empirical experiments of 12 retrieval processes to investigate the performance between keywords and query words based on the total number

eISSN: 2550-2131
ISSN: 1675-8021

www.manaraa.com

retrieved and relevant for each retrieval process, while Zakariah, Khan, Tayan and Salah (2017) presented a holistic survey and analysis approach for Digital Quran Computing. Alsmadi and Zarour (2017) used hashing algorithms to search information and web pages via the Internet for any possible intentional and unintentional changes or instances of fraud regarding the Al-Quran verses while Eljazzar, Hassan and AlSharkawy (2017) presented a solution for partitioning common Tafsir videos into short videos according to search queries and sharing these videos on social networks.

Alqahtani and Atwell (2017) evaluate different search tools based on 13 criteria which includes search features; output features; precision of the retrieved verses; recall database size; and the types of database content. Based on their survey, they conclude that most of the existing Al-Quran search tools such as desktop applications and online applications cannot solve the ambiguity problem in the retrieved results because they use traditional query analysis, thereby making limited use of the Al-Quran ontology. Alqahtani and Atwell (2017) highlighted several deficiencies such as the limitations of existing Al-Quran search tools for retrieving all requested information as most search tools only use a unique source or part of the existing Al-Quran ontologies that affected the accuracies of the retrieved results. As for Quranic semantic based ontology's and approaches, Khan et al. (2017) presented a refined ontology for the Quranic nature domain, named QNature.

Therefore, it is evident that the vast contributions of different researchers have only focused on the semantic aspects of the Al-Quran text. Notably, the linguistic aspects of the Al-Quran text can be used for finding semantically related verses. Accordingly, most of the Quranic research on the semantic aspects of the Al-Quran text proves this claim. Therefore, the focus of this study, like previous researchers is on the semantic aspects of the Al-Quran text. As the semantic aspects of Al-Quran, the previous researchers used Quranic ontology whereas this study uses the semantic relations that exist between English words using WordNet, and using the lexical cohesion that exist between the words to build lexical chain and proposes new lexical scoring system to weigh the chains more accurately.

METHOD

The primary objective of this study is to use the lexical cohesion that exists between words in the Al-Quran for creating a lexical chain and to propose a method to assign values (scores) to each aspect of that cohesion. The focus of this study is on the scoring system for the created lexical chains which include five phases: (I) ITF process phase, which will determine the rarity of the word in the Al-Quran; (II) Importance Score phase, which will determine the importance of the word in the verse; (III) Relation Score phase, which will determine the relation of the word with the next word in the verse; (IV) chain score phase, which gives the total sum of the previous phases; and (V) Chain weight phase which calculates the weight of each chain based on the calculated total sum.

eISSN: 2550-2131
ISSN: 1675-8021

www.manaraa.com

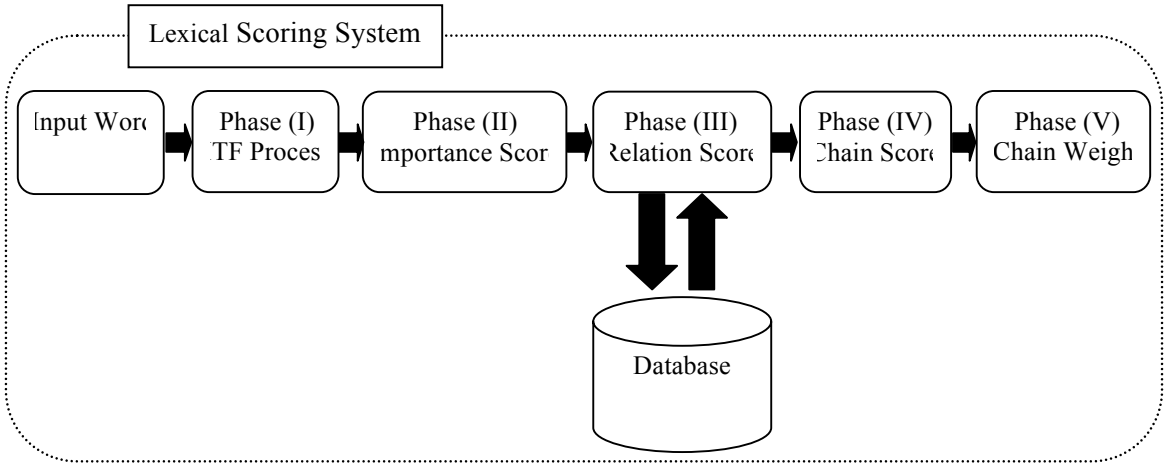

FIGURE 2. Diagram of the scoring process for the proposed Lexical Scoring System

DATASET

The dataset used in this study is the Al-Quran which is regarded as the book having the widest spread knowledge for the benefit of humanity. The Al-Quran consists of 114 chapters including a total of 6,236 verses which are different in number for each chapter. According to Islamic belief, the Al-Quran holds the words of Allah, and until the Day of Judgments; it is absolutely protected by Him. As the Holy Book of Muslims, the Al-Quran is read, recited and referred to globally thereby being available in all places and at all times. Also, the translations of the Al-Quran are available in other languages. In this study, the English translation of the Al-Quran by Abdullah Yusuf Ali is used.

The dataset required for the experimental test should consist of three predominant aspects:

- **A collection of documents:** The Al-Quran is used to represent the collection of documents. Although most verses in each chapter are connected to each other, each verse in this paper is considered as a separate document. This makes a document collection with 6,236 documents.
- **A set of queries:** A query is the user's need for a statement, which is verbalised and is shown via a statement or a question using natural language. It can also include terms which are manually selected from an indexing language (Belal 2001). The query sets for the Al-Quran dataset were developed by Fatimah (1995). Furthermore, they were written by 25 Muslim students in their final year of completing a Bachelor of computer science, only 36 queries out of an initial number of 70 queries were released.
- **Set of related documents with Relevance judgments to the query:** Relevance is a term referring to various methods by which the corresponding query and the user's query are matched. Each verse required by the user is retrieved based on an individual's enquiry, which is performed through a relevance assessment. The set of relevance assessments for the 36 queries on the Al-Quran are also provided by Fatimah (1995) which will be used for comparing the retrieved results and the evaluation of the system.

SCORING SYSTEM

For a much better understanding and explanation, Verse102 of Chapter 2 (Al-Baqrah) will be used as the example for the application of this method. Figure 3 displays the example verse.

eISSN: 2550-2131
ISSN: 1675-8021

www.manaraa.com

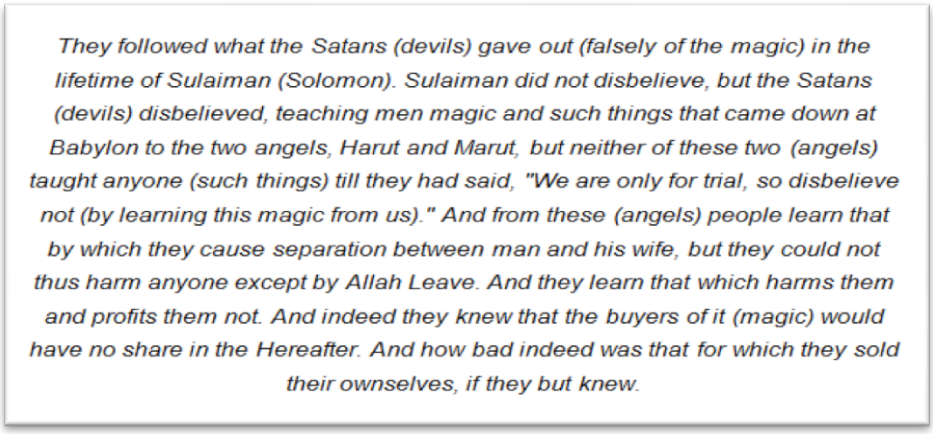
They followed what the Satans (devils) gave out (falsely of the magic) in the lifetime of Sulaiman (Solomon). Sulaiman did not disbelieve, but the Satans (devils) disbelieved, teaching men magic and such things that came down at Babylon to the two angels, Harut and Marut, but neither of these two (angels) taught anyone (such things) till they had said, "We are only for trial, so disbelieve not (by learning this magic from us)." And from these (angels) people learn that by which they cause separation between man and his wife, but they could not thus harm anyone except by Allah Leave. And they learn that which harms them and profits them not. And indeed they knew that the buyers of it (magic) would have no share in the Hereafter. And how bad indeed was that for which they sold their ownselves, if they but knew.

FIGURE 3. Verse 102 of Chapter 2, (Al-Baqrah) from the Al-Quran

The extracted chains for figure 3, the example verse are listed in Table 1. Only the first chain will be used as an example for all phases.

TABLE 1. Created chains for the example verse (Figure 3)

{follow,learn,learn,learn,separate,leave,profit,buy,share,teach,teach}
{satan,devil,satan,devil,angel,angel,angel}
{magic,magic,magic,magic}
{man,people,man,wife}
{sulaiman,sulaiman}
{harm,harm}
{solomon}
{babylon}
{lifetime}
{harut}
{marut}
{trial}

PHASE I: INVERSE TERM FREQUENCY, SCORE OF EACH TERM

The first phase of the experiment is to determine the term 'frequency' of the word in the Al-Quran. Within this phase, we try to find the most important terms in the Al-Quran and score them accordingly. In this phase, the rarest words are given a higher score than the more frequent words. For example, in this query: "*the condition of heaven*", we have two nouns 'condition' and 'heaven'. The word 'condition' is more frequently used than the word 'heaven', therefore, the verses which contain the word 'heaven' are more related to the query than the verses which contain the word 'condition'. Therefore, the score of the word 'heaven' should be higher than the score of the word 'condition'.

In order to calculate the Inverse Term Frequency (ITF) score the well-known Inverse Document Frequency (IDF) is used with small modification to calculate the term frequency of each word in the entire Al-Quran. This modification is done by replacing the document frequency (DF) (the frequency of the documents in the Al-Quran which contains the term) with term frequency (TF) of the word through the entire Al-Quran. For example, in order, to calculate IDF score for the term 'A', IDF will return the inverse value for the verses which contain term 'A' regardless of the repetition number of the term 'A'; by using the term frequency instead of document frequency, the inverse value for the repetition of term 'A' will be return through the entire Al-Quran regardless of the verses which contain the term 'A'.

eISSN: 2550-2131
ISSN: 1675-8021

www.manaraa.com

PHASE II: IMPORTANCE SCORE OF EACH WORD

During the chain creation, the longer chain always holds more related words in the document, which means that the most related document can often be determined by word members of their longest chain. Therefore, the size of the chain matters (Barzilay & Elhadad 1999; Hirst & St-Onge 1995; Hirst & St-Onge 1998). For obtaining an important score for each word, the size of the chain and number of iterations of each word and the total number of distinct words in the chain should be considered.

During the chain creation for a document, there are always some chains that include only one word or repetition of one word. In such cases, the importance of the word cannot be determined because there is no different word in the chain to compare it with and to make one word more important than the other word. In capturing such a scenario, consideration of the distinct words is therefore needed instead of the number of words available in the chain. In order, to calculate the importance of the word over another word, the word itself should not be considered as a distinct word. The number of distinct words in the chain apart from the word itself can also determine whether the chain has only one word and in-turn eliminate that chain. Also, the distinct words member of the longer chain should have the greater importance of the distinct words member of the shorter chain, to ensure that the total number of words presented in the chain is added to the equation. Table 2 depicts the Importance Score of the candidate words in the first chain.

TABLE 2. Importance Score of the candidate words in the first chain of the example verse (Figure 3)

{follow,learn,learn,learn,separate,leave,profit,buy,share,teach,teach}

Words	Total frequency of the word in the chain	Total Chain Words	Total Distinct Words	Importance Score
follow	1	11	8	12.72
learn	3	11	8	13
separate	1	11	8	12.72
leave	1	11	8	12.72
profit	1	11	8	12.72
buy	1	11	8	12.72
share	1	11	8	12.72
teach	2	11	8	12.86

PHASE III: RELATION SCORE OF EACH WORD

In this phase, the relation of each word to the next word is scored. A score is assigned to each relation based on the relation distance. Five types of relationship are used: Synonyms and Similar; Hyponymy and Hypernymy; Meronymy and Holonymy; and Siblings. Also, there are two kinds of siblings which are siblings from Holonymy and siblings from Hypernymy. The distance between these two siblings are different, hence, are categorised as Close-Siblings for Holonymy siblings and Distant-Siblings for Hypernymy siblings.

The relation score is assigned in this phase based on the relation between the selected word and the next word. To calculate the relation score, firstly, the distance score for each relation is calculated. The strongest relations are Similar and Synonyms. Similar to Barzilay and Elhadad (1999), in the present study, 10 is assigned as the distance score to these kinds of relations as the highest distance score and calculate the distance score for the other relations (Hyponymy/Hypernymy; Meronymy/Holonymy; Close-Siblings; and Distant-Siblings) are based on the strongest distance score which is 10. The distance steps between each semantic relation are illustratedin Figure 4.

eISSN: 2550-2131
ISSN: 1675-8021

www.manaraa.com

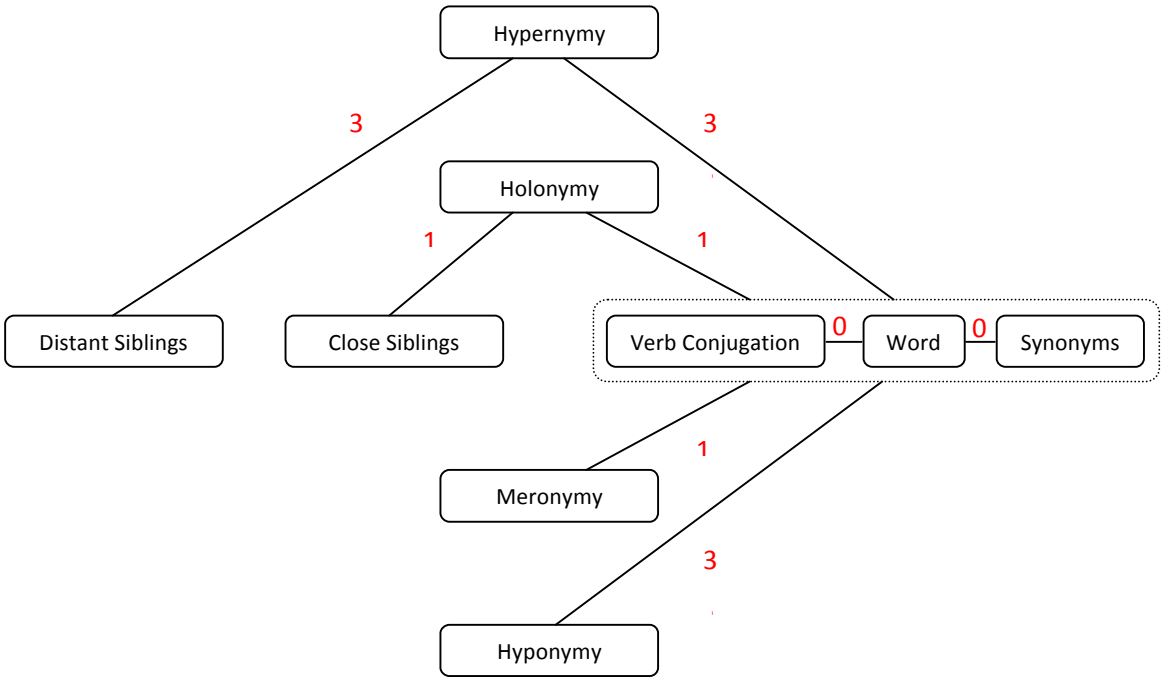

FIGURE 4. Distance steps between the semantic relations

- **0 Step** for Verb Conjugation, Synonyms and Similar; because there is no semantic difference between them. For example, for Synonymy: 'beautiful' and 'handsome', for Similar: 'beautiful' and 'beautiful' and for Verb Conjugation: 'walks' and 'walk'.
- **1 Step** for Holonymy and Meronymy: because they are the closest semantic relations to the words after Verb Conjugation, Synonyms and Similar as shown in Figure 4.
- **2 Steps** for Close-Siblings: since the semantic distance between the word and its Holonymy is 1 step, therefore, for all the words which are sharing the same Holonymy, the distance steps will be 2. For example, consider 'A' and 'B' as two words which have 'C' as the same Holonymy, the distance steps between 'A' and 'C' is 1, and the distance steps between 'B' and 'C' is also 1. Therefore the distance between 'A' and 'B' will be 2.
- **3 Steps** for Hypernymy and Hyponymy: since the semantic distance between Close-Siblings is closer than the semantic distance between the word and its Hypernymy as shown in Figure 4. For example, consider 'lion' as the selected word, 'lion', and 'tiger' are Close-Siblings in their holonymy tree which is 'Feliformia', hence, the distance between 'lion' and 'tiger' is shorter than the distance between 'lion' and 'mammal' which is the Hypernymy of 'lion'.
- **6 Steps** for Distant-Siblings: since the same rule for Close-Siblings is applied here. Thus, the semantic distance between word and its Hypernymy is 3 steps. Therefore, the distance steps between the words which have same Hypernymy will be 6.

To calculate the distance score between each relation we simply need to reduce the distance steps of the semantic relation from the top score which is 10. Therefore, the distance score of Verb Conjugation, Similar and Synonym is 10, (10-0 = 10). The distance score of Holonymy and Meronymy relation is 9, (10-1 = 9). For Close-Sibling relation the distance is 8, (10-2 = 8). Hypernymy and Hyponymy relation have a distance score of 7, (10-3 = 7) and the distance score of Distant-Siblings will be 4, (10-6 = 4).

Every word in the chain has a different importance score based on the term frequency and chain length as mentioned in phase 2. The relation score of the words with high importance should be higher than the relation score of the words with lower importance. To calculate the relation score for each word, the importance score of the word will be multiplied

eISSN: 2550-2131
ISSN: 1675-8021

www.manaraa.com

by the distance score which exists in between the word and the next word, and the distance score of the last word is 1 (there are no more words in the chain to determine the distance score).

When the algorithm does not find any word in the candidate word list that has any kind of relationship with the selected word when constructing the chains, the algorithm will select a word in the chain and check it against the words in the candidate list. If it finds any relation between the words, the word will be added at the end of the chain. These kinds of words do not have any relation with the word that precedes them in the chain, but they have a relation with one of the previous words in the chain (These words are indicated by *). Therefore, when there is no relation, the relation score will be 1. Table 3 displays the relation score for the first chain of the example verse (refer Figure 3).

TABLE 3. Chain relation score for the first chain of the example verse (Figure 3)

{follow,learn,learn,learn,separate,leave,profit,buy,share,teach,teach}				
Word (w)	**Semantic Relation with the next word**	**Distance Score**	**Importance Score**	**Relation Score**
follow	**Distant-Siblings**	**4**	**12.72**	**50.88**
learn	Similar	10	13	130
learn	Similar	10	13	130
learn	*	1	13	13
separate	Distant-Siblings	4	12.72	50.88
leave	Distant-Siblings	4	12.72	50.88
profit	Distant-Siblings	4	12.72	50.88
buy	Distant-Siblings	4	12.72	50.88
share	*	1	12.72	12.72
teach	Similar	10	12.86	128.6
teach	*	1	12.86	12.86

The relation score for each word in the chain, in the table 3, is calculated by multiplying the Important Score (calculated in phase II for each word in the chain) by, the calculated Distance Score. For example, for the word 'follow', the relation of the word 'follow' with the next word 'learn' is Distant-Siblings which has 6 step, by reducing the steps from the top Distance Score which is 10, the distance score between the word 'follow' and 'learn' will be 4. To calculate the Relation Score, the Importance Score of the word 'follow' which is 12.72 will be multiply by the Distance Score which is 4. Thus, the Relation Score for the word 'follow' will be 50.88.

PHASE IV: CHAIN SCORE

In this phase, each word is scored by the sum of the previous scores and assigns the score to the word as the word score. To calculate the chain score, the word score of all presented words in the chain will be added together as the chain score. Table 4 illustrates the word score and the chain score for the first chain of the figure 3 example verse. Table 4 shows the calculated score of each phase on the words presented in the first chain which was created for the figure 3 example verse. The total word score for each word will be the summation of the previous scores: ITF Score, relation Score and Importance Score. Then calculation of the chain score will be the sum of the total word score of each word presented in the chain. The chain score for the first chain is 881.36.

eISSN: 2550-2131
ISSN: 1675-8021

www.manaraa.com

TABLE 4. Word score and chain score of the first chain created for the example verse (Figure 3)

Chain {follow,learn,learn,learn,separate,leave,profit,buy,share,teach,teach}				**Chain Score** 881.36
Word (w)	**ITF Score**	**Relation Score**	**Importance Score**	**Total Word Score**
follow	3.24503	50.88	12.72	66.85
learn	5.96551	130	13	148.97
learn	5.96551	130	13	148.97
learn	5.96551	13	13	31.97
separate	5.96551	50.88	12.72	69.57
leave	3.4754	50.88	12.72	67.8
profit	5.56004	50.88	12.72	69.16
buy	6.94633	50.88	12.72	70.55
share	5.07453	12.72	12.72	30.52
teach	4.90945	128.6	12.86	146.37
teach	4.90945	12.86	12.86	30.63

PHASE V: CHAIN WEIGHT

Each chain based on the word score of each presented word can have a different chain score. The chain score is important for the retrieval process because it shows the strength of the chain. Therefore, a scale measure is required to determine the strength of each chain in each verse and to weight the chains accordingly.

To achieve this goal, the chain score is then calculated based on the percentages of 0 to 100. By assigning the lowest chain score in the document to 0 and the highest chain score in the document to 100, the weigh is determined for all chains between them. If there is only one chain in the document, it will be considered as the strongest chain, and the weight will be 100%. If there are two chains in the verse, the chain with the highest score is the strongest chain (weight of 100%), and the chain with low score will be the weakest chain (weight of 0%). Table 5 shows the chain score and the chain weight of the created chains (Figure 3).

For example, by comparing the Chain Scores in the table 5, the Chain Score for the longest chain (the first chain) is 881.36 which is the highest score of all created chains for the example verse (Figure 3), this chains will be consider as the strongest chain which gets the chain weight of 100%, the chain with the lowest score (the last chain, {trial}) has a chain score of 5.21173 which will be consider as the weakest chain and gets the Chain weight of 0%. Calculating the other chains will be based on the strongest chain and weakest chain.

TABLE 5. Chain Score of all created chains for example verse (Figure 3) on the scale 100%

Chain	**Chain Score**	**Chain Weight**
{follow,learn,learn,learn,separate,leave,profit,buy,share,teach,teach}	**881.36**	**100 %**
{satan,devil,satan,devil,angel,angel,angel}	391.3565	44.07%
{magic,magic,magic,magic}	20.731	1.77%
{man,people,man,wife}	173.4843	19.2%
{sulaiman,sulaiman}	11.1201	0.67%
{harm,harm}	8.9792	0.43%
{solomon}	5.56004	0.04%
{babylon}	8.73809	0.4%
{lifetime}	7.3518	0.24%
{harut}	8.73809	0.4%
{marut}	8.73809	0.4%
{trial}	**5.21173**	**0%**

eISSN: 2550-2131
ISSN: 1675-8021

www.manaraa.com

RESULTS

This section investigates experimentally how well the proposed scoring system for lexical chain performs at retrieving verses. The scoring system is applied to the lexical chain that was created for the Al-Quran IR system. Figure 5 illustrates the number of queries with correctly retrieved verses, categorised in 4 different percentage states: 100% of relevant verses retrieved, above 90% of relevant verses retrieved, above 80% of relevant verses retrieved and less than 80% of relevant verses retrieved.

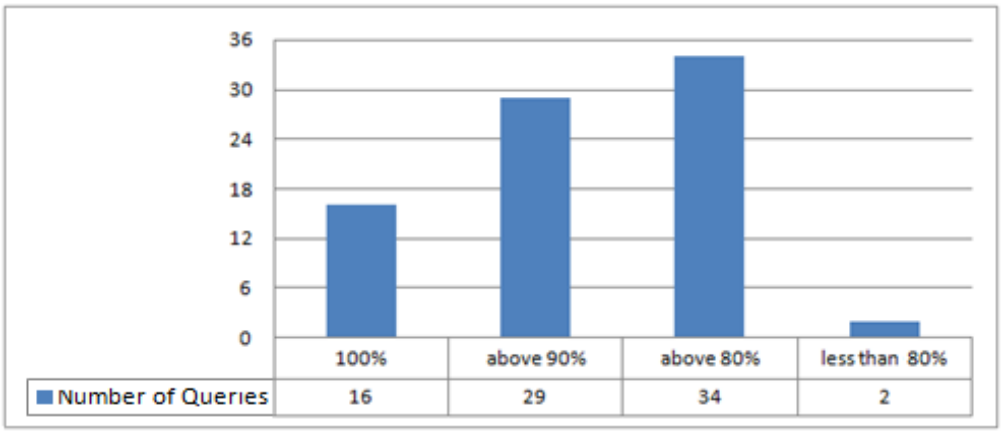

FIGURE 5. Number of queries with the percentage of correctly retrieved verses

The total number of relevant documents retrieved compared to the total number of relevance judgments provided by Fatimah (1995) for each query is shown in Table 6. The total number of related verses in Al-Quran for each query based on the relevance judgments provided by Fatimah (1995) are shown in the Relevance Judgement column and the total related verses correctly retrieved by this study's retrieval system are shown in the retrieved column. The total related verses in the Al-Quran for all 36 queries are 3,697 verses and the total correctly retrieved verses by this study's retrieval system are 3,201 verses.

TABLE 6. Number of retrieved verses in compare to Relevance Judgment

Query	Relevance Judgment	Retrieved
1	19	19
2	74	62
3	26	25
4	3	3
5	47	42
6	10	10
7	13	13
8	8	8
9	10	10
10	6	6
11	39	37
12	1,252	1,148
13	8	8
14	3	3
15	51	43
16	8	7
17	7	7
18	18	18
19	43	38
20	14	14
21	102	98

eISSN: 2550-2131
ISSN: 1675-8021

www.manaraa.com

22	11	1
23	77	76
24	31	29
25	1	1
26	395	369
27	225	210
28	550	292
29	114	108
30	90	87
31	17	17
32	294	265
33	8	8
34	77	74
35	36	36
36	10	9
Total	**3,697**	**3,201**

Figure 6 shows the system interface for the Al-Quran IR. Notably, it is not possible to include and discuss all 36 queries in this study; therefore, as ~~the~~ example, only 3 queries are included with the top 3 retrieved documents for each query. The similarity between the verse and the query is based on the total number and the strength of the relation between the semantically related words that exist between the query and the verses. The score of these relations is calculated based on phases of the scoring system.

The similarity measure between query and verses varies between 0.0 and 1.0. The closer similarity value to 1.0, between the query and verse shows the higher relatedness between them. Figure 6 shows the result for example query 1 and the top 3 verses that were retrieved.

FIGURE 6. The system interface of the Al-Quran IR

Example Query1: *From which verse/chapter can be found about the **importance** of **knowledge?***

Retrieved Verse 1: "**taught** man that which he **knew** not."

Retrieved Verse 2: "and among them are men who listen to thee, but in the end, when they go out from the, they say to those who have received knowledge, "what is it he said just then?" such are men whose hearts Allah has sealed, and who follow their own lusts."

eISSN: 2550-2131
ISSN: 1675-8021

www.manaraa.com

Retrieved Verse 3: "he grants wisdom to whom he pleases, and he to whom wisdom is granted receive indeed a benefit overflowing, but none will grasp the message but men of understanding."

The first document retrieved for example 1 is Verse 5 of Chapter 96 with the similarity of 0.983973. The second document is Verse 16 of Chapter 47 with the similarity of 0.968774, and the third document is Verse 269 of Chapter 2 with the similarity of 0.965463. Here, the related words extracted from query-1 are 'important' and 'knowledge'. The first true retrieved verse only contains the related words 'taught' and 'knew'; words by common sense related to the word 'knowledge'. Therefore, this shows how the lexical chain representation is able to relate to the words from the query with the words in the verses.

Example Query 2: *Information retrieval about the pillars of that is the **testimony**.*
Retrieved Verse 1: "o prophet! truly we have sent thee as a **witness**, a bearer of glad tidings, and warner."
Retrieved Verse 2: "o ye who believe! stand out firmly for justice, as witnesses to Allah, even as against yourselves, or your parents, or your kin, and whether it be (against) rich or poor: for Allah can best protect both. follow not the lusts (of your hearts), lest ye swerve, and if ye distort (justice) or decline to do justice, verily Allah is well-acquainted with all that ye do."
Retrieved Verse 3: "We have truly sent thee as a **witness**, as a bringer of glad tidings, and as a warner."

The first document retrieved for example 2 is Verse 45 of Chapter 33 with the similarity of 0.984574, the second document is Verse 135 of Chapter 4 with the similarity of 0.984401, and the third document is Verse 8 of Chapter 48 with the similarity of 0.910211. Again, from this example, applying query-2, and taking the word 'testimony' the system relates to the words 'witness', thus showing the situation where the hyponym relation of the word 'testimony' is captured. This phenomenon obviously demonstrates the benefits of representing the verses and query in lexical chain form.

Example Query 3: *Information retrieval about the pillars of Islam that is the **charity**.*
Retrieved Verse 1: "so he gave nothing in **charity**, nor did he pray!-"
Retrieved Verse 2: "by no means shall ye attain righteousness unless ye give (freely) of that which ye love; and whatever ye **give**, of a truth allah knows it well."
Retrieved Verse 3: "know they not that allah doth accept repentance from his votaries and receives their gifts of charity, and that allah is verily he, the oft-returning, most merciful?"

The first document retrieved for example 3 is Verse 31 of Chapter 75 with the similarity of 0.990901, the second document is Verse 92 of Chapter 3 with the similarity of 0.965237, and the third document is Verse 104 of Chapter 9 with the similarity of 0.963426. From this example, applying query-3, and taking the word 'charity' the system relates to the words 'charity' and 'give' in the related retrieved verses.

In the information retrieval process, the retrieved verses are then evaluated for relevance. The evaluations are based on the proportion of relevant verses returned (i.e. retrieved) and the proportion of verses returned that are relevant (Powers, 2003), which are named Recall and Precision. The recall represents one aspect of search performance; the

eISSN: 2550-2131
ISSN: 1675-8021

www.manaraa.com

effectiveness of the algorithm in retrieving the relevant verses, with regards to the fraction of relevant verses of the Al-Quran. Specifically, it is calculated as the number of relevant items (i.e. verses) retrieved as a proportion of all relevant items (verses) that may potentially be retrieved (Bautista-Gomez et al. 2016). To calculate recall, the number of relevant verses retrieved by the system is divided by the total number of relevant verses in the Al-Quran, and to obtain the percentage, the result will be multiplied by 100. Equation 1 shows the calculation.

$$\text{The number of relevant documents retrieved by the system}\text{The total number of relevant documents in the corpus} \times 100 \quad (1)$$

In contrast, Precision accounts for both the retrieval of relevant verses and the exclusion of non-relevant verses which measures the quality of providing relevant verses. It is calculated as the number of relevant items (i.e. verses) retrieved as a proportion of all items (verses) retrieved, as defined by Walters (2009). Precision and recall can usually be traded off in an IR algorithm; when recall increases the precision decreases, and vice versa (Berry, Ferrari & Gnesi, 2017). In order to calculate precision, the number of relevant verses retrieved by the system is divided by the total number of retrieved verses, and to obtain the percentage, the result will be multiplied by 100. Equation 2 shows the calculation.

$$\text{The number of relevant documents retrieved by the system}\text{The total number of retrieved documents by the system} \times 100 \quad (2)$$

The result of precision and recall are usually shown in the form of a graph named as the "precision versus recall curve". The curve is the result of averaging the results of the retrieved verses for all 36 queries. Such an average curve is normally used to compare the retrieval performance of distinct retrieval systems. In the case of this study, the precision versus recall curve is used to illustrate the retrieval performance of the proposed scoring system of lexical chains for the Al-Quran. Figure 7 illustrates the precision versus recall curve for the retrieved verses of the Al-Quran, in which it shows that the performance of the retrieval quran system is considered good since it is near to curve where a perfect retrieval system should have. A perfect curve should have a flat horizontal line at the highest point of Precision (0.7) that meets with the flat vertical line at farthest point of Recall (1.0).

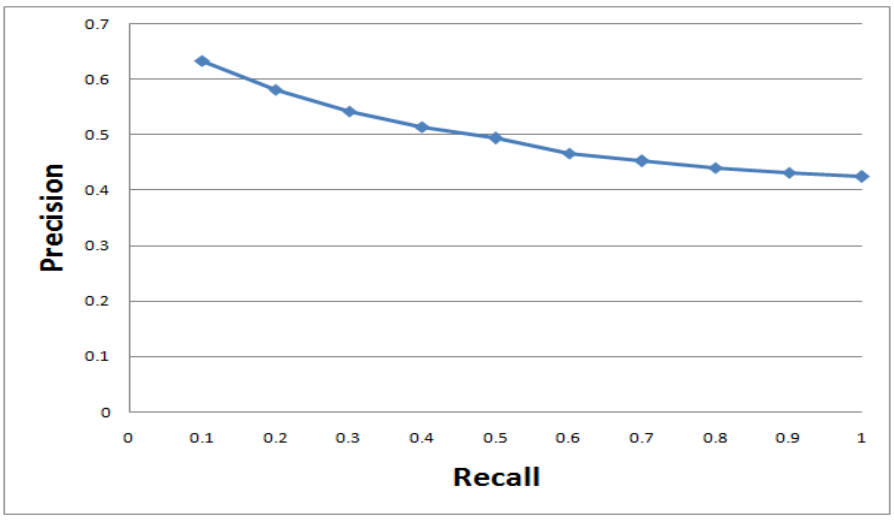

FIGURE 7: Precision versus recall curve comparison for the retrieved verses of Al-Quran using the proposed scoring system on Lexical Chain.

eISSN: 2550-2131
ISSN: 1675-8021

www.manaraa.com

DISCUSSION

The lexical scoring system for Quranic information retrieval proposed in this paper is based on the lexical chain approach. To create lexical chains and to find the semantic relations between the words in the verses, WordNet (Miller & Fellbaum, 1998) has been used as the lexical database. The semantic relation which is used to build the lexical chains is based on the relationship between the related words found in WordNet. The results and analysis revealed in the present study draw attention to the significance of using lexical chain for representing the Al-Quran verses by using the semantic relations that exists between the words in the verse. The findings indicate that the scoring system for lexical chain proposed and used in our Quranic IR, score and weight the words more accurately by scoring all aspects of semantic relations between the words.

The range of chain score for the verses is not predictable and depends on the length of the verse and relation between the words in that verse. Therefore, there is a necessary step to scale the scores. The scale weight calculation is performed by assigning the lowest chain score in the verse to 0% and the highest chain score in the verse to 100%. By identifying the strongest and the weakest chain in the chain collection of each verse, the weight for all chains between them can be determined. This calculation has the following condition. If there is only one chain in the verse, the chain will be considered as the strongest chain, and the weight will be 100%. If there are only two chains in the verse, the highest scored chain is the strongest chain and is given a weight of 100%. The low score chain will be the weakest chain and given a weight of 0%.

By examining how the IR system behaves (i.e. human's judgment), the retrieval result for all 36 queries was evaluated against the relevance judgment provided by Fatimah (1995). Table 6 shows the result of the total number of retrieved verses in comparison to the relevance judgment. The result indicates that the total number of relevance judgments for all queries in the Al-Quran is 3,697 and the total number of retrieved verses using our Quranic IR with the proposed Lexical Scoring System is 3,210. Therefore, this indicates that 86.58% of total relevant verses are retrieved. The proposed lexical scoring system successfully retrieved 100% of relevant verses based on 16 queries, and above 80% of retrieved verses based on 34 queries and only 2 queries retrieved less than 80% of total related verses.

In a perfect retrieval system, the precision versus recall curve will be displayed as a straight line with the precision of 1.0 shown on the graph. In the existing retrieval systems, the end of curve always moves down as recall increases. This downward curve is caused by the retrieval of unrelated documents in the system. Depending on the retrieval system, the end of the curve may shift down all the way which shows a high-level of unrelated documents retrieved. In the retrieval system with the higher related documents retrieved at the top ranks (top retrieved documents), the curve tends to remain up. The precision verses recall curve shown in Figure 7 illustrates that the Quranic IR system in this study indicating better retrieval of related documents (verses) at the upper top rank.

CONCLUSION

In this study, a new scoring system for lexical chain is proposed. The proposed system consists of five different phases, and within each phase, one aspect of the word in the chain is scored. The aspects include: (1) The Inverse Term Frequency of the word which produces the rarity of the word in the corpus. Words with high frequency are given a lower score whereas words with low frequency are given a higher score. (2) The importance of the word in the chain. The distinct words are based on the number of repetitions of the word, and the total length of the chain that the word is presented in will be scored. (3) The relation which the

eISSN: 2550-2131
ISSN: 1675-8021

www.manaraa.com

word has with the next word in the chain. The relations that are considered for this aspect are: Synonyms and Similar; Hyponymy and Hypernymy; Meronymy and Holonymy; and Siblings which are divided into two groups, Close-Siblings (words with the same holonymy) and Distant-Siblings (words with the same hypernymy). (4) The total score for the chain by summing up the previous score of the words, and (5) calculating the chain weight on a scale of 0 to 100%.

The experiment in this study employed the proposed Lexical Scoring System for scoring the created lexical chains and was tested on the Al-Quran. The results indicated 86.58% of total relevant verses were successfully retrieved based on the relevance judgments provided by Fatimah (1995). Furthermore, the results show that 16 queries retrieved 100% of the relevant verses, 29 queries retrieved above 90% of the total relevant verses, 34 queries retrieved above 80% of the total verses and only 2 queries retrieved less than 80% of the relevant verses. By referring to the results, the conclusion is that a lexical chain is an appropriate approach for Quranic IR.

FUTURE WORK

The Natural language words in general are highly ambiguous; a unique interpretation can usually be determined only by taking into an account the constraining influence of the context in which the word occurred. When an ambiguous word is written in a sentence, it is possible to select the correct sense of that word. However, in any application where a computer has to process natural language, ambiguity is a problem. Ambiguity in information retrieval can cause the retrieval of irrelevant documents, while different words which represent the same concept can cause the retrieval of unrelated documents. For example, the word 'bat' in a sentence, can be translated as: "an implement used in sports to hit balls" or "flying mammal". These problems can decrease the information retrieval performance system. It seems reasonable to assume that an information retrieval system will improve its performance if the documents it retrieves are represented by word senses rather than words. Since ambiguity is one of the factors that decreases retrieval system performance, retrieval performance can be improved if ambiguity can be solved, and one of the well-known solution is the application of Word Sense Disambiguation (WSD) technique. WSD is a process to define the sense/meaning of an ambiguous word. The main aim of WSD is to determine the most probable sense of polysemous word among the possible set of sense candidates in a given context. WSD in the semantic based applications such as lexical chain can be beneficial. Disambiguating the words sense in the documents is done before creating the lexical chains and using the disambiguated sense to help create the lexical chains. Thus, WSD will the main future direction of this study.

ACKNOWLEDGEMENT

This project is funded by MoHE under research code FRGS/1/2016/ICT02/UKM/02/14.

eISSN: 2550-2131
ISSN: 1675-8021

www.manaraa.com

REFERENCES

Abdelnasser, H., Ragab, M., Mohamed, R., Mohamed, A., Farouk, B., El-Makky, N. & Torki, M. (2014). Al-Bayan: An Arabic Question Answering System for the Holy Quran. *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, 25–29 October, Doha, Qatar: 57-64.

Abdul-Ghafour, A. K. M., Norsimah Mat Awal., Intan Safinaz Zainudin & Ashinida Aladdin. (2017). Meanings of Near-Synonyms and Their Translation Issues in the Holy Qur'ān. *GEMA Online® Journal of Language Studies. Vol. 17*(4), 258-273.

Alqahtani, M. & Atwell, E. (2017). Evaluation Criteria for Computational Quran Search. *International Journal on Islamic Applications in Computer Science And Technology Vol. 5*(1), 12-22.

Alrehaili, S. M. & Atwell, E. (2014). Computational Ontologies for Semantic Tagging of the Quran: A Survey of Past Approaches. *LREC 2014 Proceedings*, 26-31 May, Reykjavik, Iceland:19-23.

Alsmadi, I. & Zarour, M. (2017). Online Integrity and Authentication Checking for Quran Electronic Versions. *Applied Computing and Informatics*. *Vol. 13*(1), 38-46.

Ayed, M. a. H. & Atwell, E. (2017). Quran Question Answering System Using Arabic Number Patterns (Singular, Dual, Plural). *International Journal on Islamic Applications in Computer Science and Technology (IJASAT)*. *Vol. 5*(2), 1-12.

Barzilay, R. & Elhadad, M. (1999). Using Lexical Chains for Text Summarization. In I. Mani and M. T. Maybury, (Eds.). *Advances in Automatic Text Summarization* (pp. 111-121), Cambridge: The MIT Press.

Bautista-Gomez, L., Benoit, A., Cavelan, A., Raina, S. K., Robert, Y. & Sun, H. (2016). Coping with Recall and Precision of Soft Error Detectors. *Journal of Parallel and Distributed Computing*, *Vol. 9*(8), 8-24.

Belal, M. H. A. (2001). An Arabic Stemming Algorithm Based on Extensive Rules Application (Area) for Information Retrieval: Its Development and Performance Measures. Unpublished PhD thesis, University Kebangsaan Malaysia, Malaysia.

Berry, D. M., Ferrari, A. & Gnesi, S. (2017). Assessing Tools for Defect Detection in Natural Language Requirements: Recall Vs Precision. Retrieved 18 May, 2018 from https://pdfs.semanticscholar.org/bb70/310c2fad1648cdc31e9799733a52f6711311.pdf

Eljazzar, M. M., Hassan, A. & Alsharkawy, A. A. (2017). Towards a Time Based Video Search Engine for Al Quran Interpretation. *Computer Science*. Retrieved 18 May, 2018 from https://arxiv.org/ftp/arxiv/papers/1701/1701.09138.pdf.

Enss, M. J. R. (2006). An Investigation of Word Sense Disambiguation for Improving Lexical Chaining. Unpublished master thesis, University of Waterloo, Canada.

Fatimah Ahmad. (1995). A Malay Language Document Retrieval System: An Experimental Approach and Analysis. Unpublished PhD thesis, University Kebangsaan Malaysia, Malaysia.

Halliday, M. A. K. & Hasan, R. (2014). *Cohesion in English*, New York: Routledge.

Hamed, S. K. & Mohd Juzaiddin Ab Aziz. (2016). A Question Answering System on Holy Quran Translation Based on Question Expansion Technique and Neural Network Classification. *Journal of Computer Sciences. Vol. 12*(3), 169-177.

Hirst, G. & St-Onge, D. (1995). Lexical Chains as Representations of Context for the Detection and Correction of Malapropisms. In C. Fellbaum (Ed.), *WordNet: An Electronic Lexical Database* (pp. 305-332), : MIT Press.

Hirst, G. & St-Onge, D. (1998). Lexical Chains as Representations of Context for the Detection and Correction of Malapropisms. *WordNet: An Electronic Lexical Database. Vol. 3*(5), 305-332.

eISSN: 2550-2131
ISSN: 1675-8021

www.manaraa.com

Iqbal, R., Mustapha, A. & Mohd. Yusoff, Z. (2013). An Experience of Developing Quran Ontology with Contextual Information Support. *Multicultural Education & Technology Journal*. *Vol. 7*(4), 333-343.

Jarmasz, M. & Szpakowicz, S. (2001). Roget's Thesaurus: A Lexical Resource to Treasure. *Proceedings of NAACL Workshop*, 3-4 June, Pittsburgh: 186 - 188.

Khan, S. Z., Rahman, M. M., Sadi, A. S., Anwar, T., Mohammed, S. & Chowdhury, S. (2017). The Quranic Nature Ontology: From Sparql Endpoint to Java Application and Reasoning. *International Journal of Innovative Computing*. *Vol. 7*(2), 13-20.

Manning, C. D., Raghavan, P. & Schütze, H. (2009). *An Introduction to Information Retrieval*. England: Cambridge University Press.

Miller, G. & Fellbaum, C. (1998). *Wordnet: An Electronic Lexical Database*, Cambridge: MIT Press.

Mohamed, O. J. & Sabrina Tiun. (2015). Word Sense Disambiguation based on Yarowsky Approach in English Quranic Information Retrieval System. *Journal of Theoretical and Applied Information Technology. Vol. 82*(1), 163-171.

Morris, J. & Hirst, G. (1991). Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text. *Computational Linguistics*. *Vol. 17*(1), 21-48.

Powers, D. M. (2003). Recall & Precision Versus the Bookmaker. *International Conference on Cognitive Science*, 13-17 July, Sydney, Australia: 539-534.

Roget, P. M. (1977). *Roget's International Thesaurus*. New York: *HarperCollins Publishers.*

Ruas, T. & Grosky, W. (2017). Exploring and Expanding the Use of Lexical Chains in Information Retrieval (Technical Report). Retrieved 18 May, 2018 from https://deepblue.lib.umich.edu/bitstream/handle/2027.42/136659/LexicalChainsReport.pdf.

Sabrina Tiun, Zakr, H., Masnizah Mohd, Norazlinda Zainal Abidin & Ahmad Irfan Ikmal Hisham. (2013). Word Sense Disambiguation on English Quranic IR System. *Proceedings of Taibah University International Conference on Advances in Information Technology for Holy Quran and Its Science (NOORIC 1435/2013)*, 19-22 December 2013, Madinah, Saudi Arabia: 214-217.

Shoaib, M., Yasin, M. N., Hikmat, U. K., Saeed, M. I. & Khiyal, M. S. H. (2009). Relational Wordnet Model for Semantic Search in Holy Quran. *International Conference on Emerging Technologies*, 19-20 October, Islamabad, Pakistan:29-34.

Silber, H. G. & Mccoy, K. F.(2000). An Efficient Text Summarizer Using Lexical Chains. *Proceedings of the first international conference on Natural language Generation,* 12-16 June, Mitzpe Ramon, Israel: 268-271.

Silber, H. G. & Mccoy, K. F. (2002). Efficiently Computed Lexical Chains as an Intermediate Representation for Automatic Text Summarization. *Computational Linguistics. Vol. 28*(4), 487-496.

Siti Nor Fatimah Haris. & Melord Md Yunus. (2014). The Use of Lexical Cohesion among TESL Post Graduate Students in Academic Writing. *Journal of Education and Human Development*. *Vol. 3*(2), 847-869.

Ta'a, A., Abed, Q. & Ahmad, M. (2017). Al-Quran Ontology Based on Knowledge Themes. *Journal of Fundamental and Applied Sciences*. *Vol. 9*(5), 800-817.

Walters, W. H. (2009). Google Scholar Search Performance: Comparative Recall and Precision. *portal: Libraries and the Academy*. *Vol. 9*(1), 5-24.

Yauri, A. R., Kadir, R. A., Azman, A. & Murad, M. A. (2013). Quranic Verse Extraction Base on Concepts Using Owl-Dl Ontology. *Research Journal of Applied Sciences, Engineering and Technology*. *Vol. 6*(23), 4492-4498.

eISSN: 2550-2131
ISSN: 1675-8021

www.manaraa.com

Yunus, M. a. M., Mustapha, A. & Samsudin, N. A. (2017). Analysis of Translated Query in Quranic Malay and English Translation Documents with Stemmer. *MATEC Web of Conferences*. *Vol. 135*, 00069.

Zakariah, M., Khan, M. K., Tayan, O. & Salah, K. (2017). Digital Quran Computing: Review, Classification, and Trend Analysis. *Arabian Journal for Science and Engineering. Vol. 42*(8), 3077-3102.

ABOUT THE AUTHORS

Hamed Zakeri Rad: Ph.D candidate at the Faculty of Information Science and Technology (FTSM) in Universiti Kebangsaan Malaysia (UKM). Under Knowledge Computing (KC) research group under the Center for Artificial Intelligence Technology (CAIT), FTSM. Under supervision of Dr. Sabrina Tiun and Dr. Saidah Saad.

Dr. Sabrina Tiun: Senior lecturer at the Faculty of Information Science and Technology in Universiti Kebangsaan Malaysia. Range of research interests are from Natural Language Processing to Speech Processing and Information Retrieval. Member of the Knowledge Computing research group, under the Center for Artificial Intelligence Technology, FTSM.

Dr. Saidah Saad: Senior lecturer at the Faculty of Information Science and Technology in Universiti Kebangsaan Malaysia. Range of research interests are from Semantic Web - Ontology Creation, Natural Language Processing to Information Retrieval. Member of the Knowledge Computing research group, under the Center for Artificial Intelligence Technology FTSM.

eISSN: 2550-2131
ISSN: 1675-8021

www.manaraa.com

Copyright of GEMA Online Journal of Language Studies is the property of GEMA Online Journal of Language Studies and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.

www.manaraa.com