CHAPTER 1 : THE PROBLEM

This chapter discuss about rationale the research, theoretical framework, conceptual framework, statement of the problem, objective, hypotheses, assumption, scope and delimitation, and importance of the study.

1.1 Rationale

Abdul Baquee Muhammad [1] has built Corpus that contained *AlQur'an* domain, *WordNet* and dictionary. He has done initialisation in the development of knowledge about *AlQur'an* and the knowledges about relatedness among texts in *AlQur'an*. To the best our knowledge, the Path based measurement method that proposed by Liu, Zhou and Zheng [3] has never been used in the *AlQur'an* domain. By using *AlQur'an* translation *dataset* in this research, the path based measurement method proposed by Liu, Zhou and Zheng [3] be used to test this method in *AlQur'an* domain to obtains similarity value and to measures its correlation value.

In the study that conducted by Liu, Zhou and Zheng [3] on a semantic similarity using path-based method managed to get a correlation value of 92.6%. The correlation value still has an opportunity to be improved because of the research by Liu, Zhou and Zheng [3] only optimizing the semantic relationship of hypernym and hyponym. In the Semantic relatedness there are many semantic relationships besides hypernym and hyponym such as synonym, antonym, meronym and holonym. Taking advantage of all semantic relationships besides hypernym and hyponym such as synonym, antonym, meronym and holonym is expected to increase the correlation value that has been achieved previously.

To obtains a better correlation value the degree value is proposed to be used in modifying the path based method that proposed by Liu, Zhou and Zheng [3]. Degree Value is the number of links that owned by a lcs (lowest common subsumer) node on a taxonomy. The links owned by a node on the taxonomy represent the semantic relationships that a node has in the taxonomy. By using degree value all of semantic relationships besides hypernym and hyponym such as synonym, antonym, meronym and holonym can be identified in a node. By using degree value to modify the path-based method that proposed by Liu, Zhou and Zheng [3] is expected that the correlation value obtained can increase.

1.2 Theoretical Framework

Semantic is the knowledge which learning about meaning of the word in the language. The semantic related with meaning of the word relation, it's like in synonym, antonym and

hyponym. Johnson [2] stated that semantic theory influence the planning to describe meaning of the word. Brinton [2] added lexical semantic is a study about the meaning of the word individually, Fromkin [2] completed the theory by stating that the lexical semantics are related to the word meaning's and the semantically *relatedness* between the word. So the lexical semantic learns the meaning that related with the word.

Morris and Hirst 2004 [2] stated that semantic *relatedness* explains the power of semantic relationships between two words or concepts are measured. It encompasses a variety of relations between concepts, including the classical relations such as hypernymy, hyponymy, meronymy, antonymy, synonymy, and any other 'non classical relations', Zesch and Gurevych [2] added 'implicit connections' in that definition. Weeds [2] defined that semantic similarity is a specific case of *relatedness*, where the sense of *relatedness* is dependent on the 'degree of synonymy', which is usually accounted by classical relations.

In the taxonomy or lexical hierarchy generally are shaped from hypernym and hyponym words. Hypernym is the words that have general meaning from the other words. Hyponyms are a group of words that are of special significance meaning rather than measured words. Hyponym is a member of group of a hypernym meaning groups, eg hypernym = sport, hyponym = football, basketball, tennis, run.

Semantic similarity between words is often represented by similarity between concepts that associated with the words. Concepts can have different meaning but remain semantically interconnected like sports and balls. Other instance that antonyms are respected has a semantic relationship such as smart and stupid, however they are dissimilar. A way that can be used to find out the value of similarity is to calculate the value of shortest path length between two words. The similarity value is also influenced by the depth of subsumer value of the lowest common subsumer (*lcs*) nodes from two concepts. The similarity value measured between two concepts is increase if the depth of subsumer value of the two concepts also grows higher in the taxonomy.

Refer to the research by Liu, Zhou and Zheng [3], the depth of concepts in the hierarchy also contributes much to the similarity, therefore the depth of concept in lexical hierarchy is used to account the similarity between two words beside the shortest path. Which the shortest path in lexical hierarchy is used to account the different between two words.

The Depth of Concepts is denoted with letter *'d'* and the Shortest Path Length is denoted with letter *'l'*, which the formula of semantic similarity is :

$$S(w_1, w_2) = \frac{f(d)}{f(d) + f(l)} \qquad [1]$$

Where S is semantic similarity, w_1 is the word number 1, w_2 is the word number 2, f is the transfer function for l and d , l is the shortest path length among w_1 and w_2 , d is the depth of *subsumer* in *taxonomi* hierarchy.

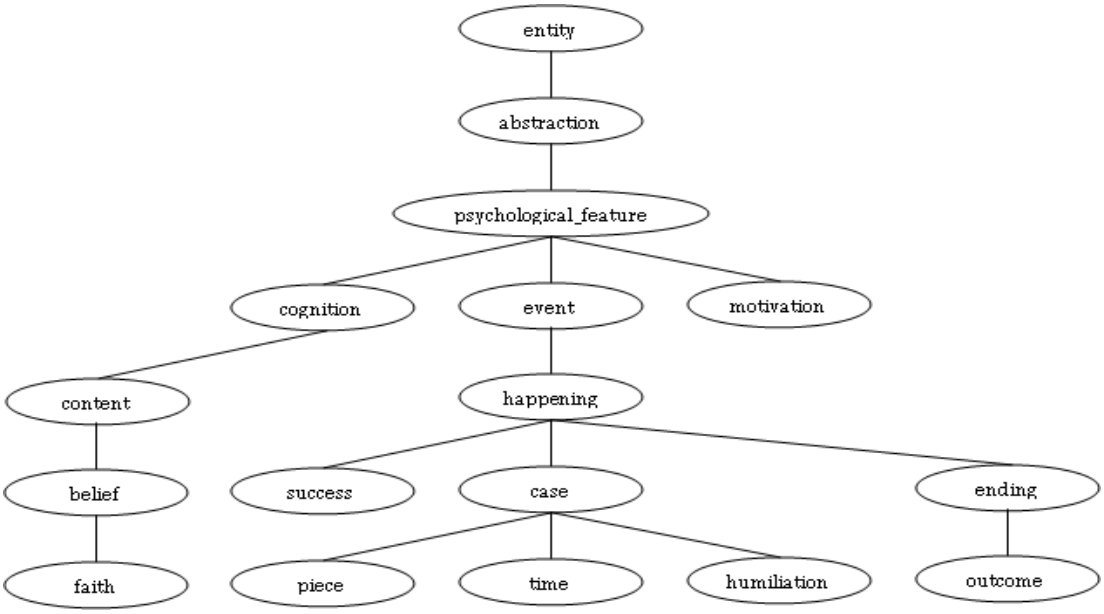

Figure 1 Chart of Taxonomy

For example it use words : *success* and *time* in taxonomy at Figure 1 that can be measured. l is the shortest path length among word$_1$ (*success*) and word$_2$ (*time)* where this value can be got from measurement number of links that exist in paths from node *success* to node *time*. d is the depth of subsumer in taxonomi hierarki. *Lowest Common Subsumer* (*lcs*) or parent node from node *success* and node *time* is node *happening*, so this value (d) can be got from measurement number of links that exist in paths from *lcs* node (*happening)* to root node (*entity*). b is number of links that be owned by *lcs* or parent node so in this case is number of links that be owned by node *happening*.

Table 1 Reference Results

Semantic Similarity Method	Correlation with RG-28	Correlation with MC-28	Correlation with RG-65
WordNet Edges	0.740	0.739	0.787
Hirst-St.Onge	0.671	0.682	0.732
Jiang-Conrath	0.670	0.684	0.731
Leacock-Chodorow	0.801	0.820	0.852
Lin	0.773	0.814	0.834
Resnik	0.706	0.763	0.800
Yang-Powers	0.889	0.921	0.897
Li	0.891	0.883	0.881
Liu-Zhou-Zheng	0.909	0.926	0.891

Table 1 shows the performance of semantic similarity measurement reported by previous researchers. The second column lists the correlation with the mean human ratings from 28 words pair of the RG experiment. The third column is the correlation with the mean human ratings from 28 word pairs of the MC experiment. The correlation with the mean human ratings from 65 word pairs of the RG experiment are listed in the last column [3].

From the Table 1 it can be shown that Liu, Zhou and Zheng [3] method achieves the best correlation 0.926 with the MC's average human value of 28 pairs of words compared with the whole method. Correlation 0.926 is also significantly better than typical human subjects 0.9015. This mean that Liu, Zhou and Zheng [3] method's is more accurates than most individual subjects. Therefore human judgement for semantic similarity can be simulated by the ratio of equal attribute with aggregate attribute that is the sum of equal and different attribute between words.

1.3 Conceptual Framework

Keijo Ruohonen [16] defined that degree of the vertex is the number of edges with vertex as and end vertex. Number of edges that exist at a vertex are the edge that derived from that vertex or towards to the vertex. The degree value can be used to compute the number of the edge that established at the vertex. Each vertex in the vertex network has a degree because each vertex has a relationship with another vertex.

WordNet has a word hierarchy or commonly called taxonomy. In the taxonomy there are relationships among *synsets*, the relationships are synonym, antonym, hypernym, hyponym, meronym and holonym. That relationships among *synsets* are connected by several links to form a word or *synsets* network. Each *synset* in taxonomy can has multiple links depending on the relationships that are formed and owned with other *synsets*.

Therefore the degree value can be obtained from the number of links formed from a node connected to another node through the hypernym, hyponym, meronym and holonym relationships. And degree value which is proposed to modify the formula is the number of edge which exists at the lowest common subsumer (*lcs*) vertex. Therefore the degree value can be proposed to modify the formula derive from Liu, Zhou and Zheng [3], and the formula changed like shown below.

$$S(w_1, w_2) = \frac{b*d}{b*d + l} \quad [2]$$

where *b* is the degree value from the lowest common subsumer (*lcs*) node

Conceptual Framework on Figure 2 showing concept of *WordNet* and path based method usage to find text *relatedness* semantically, that algorithm may solves the problem in this thesis. The usage of *WordNet* and path based method are expected can improve the performance of linkages among texts in *AlQur'an* measurement. Similarity value is the value of performance that be measured.

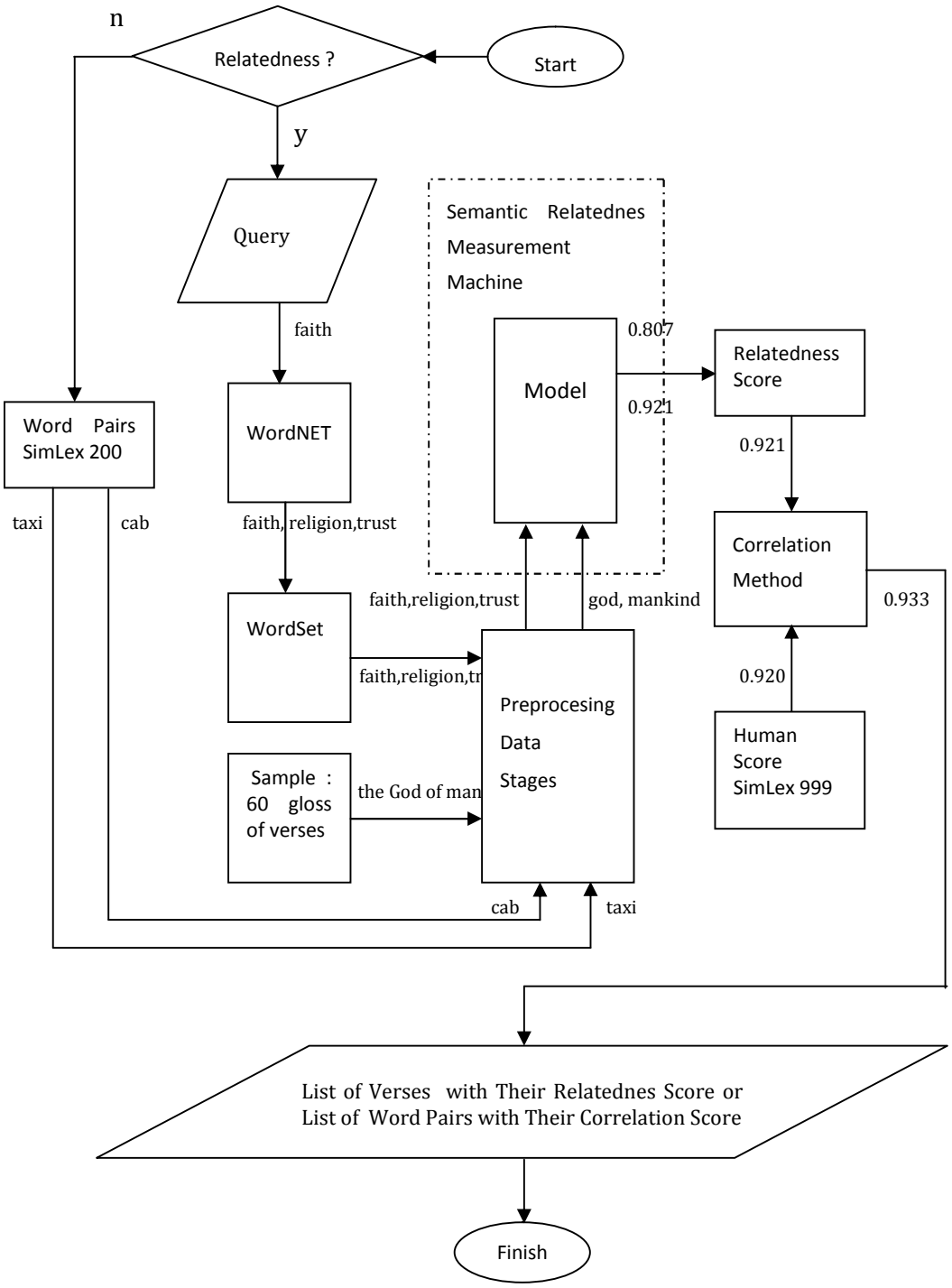

Figure 2 Conceptual Framework

1.4 Statement of The Problem

The previous research for linkage of texts achieve correlation value just about 92.6 % because they use semantic links hypernym and hyponym only, whereas the links beside hypernym and hyponym are not used, so this method make the number of links that accounted just a little and that cause the correlation value is not optimal.

1.5 Objective

Identifing all semantic relationships that owned by a node so that the semantic relationships that owned by a node become more a lot and can increase the correlation value better than before.

1.6 Hypotheses

With use degree value in semantic text relatedness method which capable to identify all relatedness semantic links that owned by a node so with this method is expected to make the number of links that are connected to the node are more a lot and that cause the correlation value is more optimal.

1.7 Assumption

Previous study by Liu, Zhou and Zheng [3] has gained result 92.6 % in correlation value using path based method. It is assumed that using degree value to modify the path based method which capable to identify all links that connected to the node can be used to improve the correlation value.

1.8 Scope and Delimitation

1.8.1 Scope of Knowledges

The knowledges about semantic text *relatedness* that use *Path based* method are useful and appropriate due to it's ability to compare the words that be measured to know the *relatedness* value between them. With account of the depth of lexical and the shortest path length in graph model the similarity and *relatedness* value between words can be measured and determined.

1.8.2 Scope of User

All the mosleem who want to learn about verses of *AlQur'an.*

1.8.3 Scope of Verses in Research

Because of the research time that used to finish the research is short and to make easier for the manual computation so the account of *AlQur'an* verses that be processed in this research is limited in 60 verses, that is from surah *Al-Ashr* until surah *An-Naas* and the research is similarity and *relatedness* among words.

1.9 Importance of The Study

This study can help the *AlQur'an* researcher and the mooslem to enrich their knowledge and augment the optional method that can be used in the *relatedness* among verses of *AlQur'an* research and the *relatedness* among words research.