New instances classification framework on Quran ontology applied to question answering system

Fandy Setyo Utomo*[1], Nanna Suryana[2], Mohd Sanusi Azmi[3]
[1]Department of Information Systems, STMIK AMIKOM Purwokerto, Purwokerto, Indonesia
[1,2,3]Center for Advanced Computing Technology (C-ACT), Fakulti Teknologi Maklumat dan Komunikasi, Universiti Teknikal Malaysia Melaka, Melaka, Malaysia
*Corresponding author, e-mail: fandy_setyo_utomo@amikompurwokerto.ac.id

Abstract

Instances classification with the small dataset for Quran ontology is the current research problem which appears in Quran ontology development. The existing classification approach used machine learning: Backpropagation Neural Network. However, this method has a drawback; if the training set amount is small, then the classifier accuracy could decline. Unfortunately, Holy Quran has a small corpus. Based on this problem, our study aims to formulate new instances classification framework for small training corpus applied to semantic question answering system. As a result, the instances classification framework consists of several essential components: pre-processing, morphology analysis, semantic analysis, feature extraction, instances classification with Radial Basis Function Networks algorithm, and the transformation module. This algorithm is chosen since it robustness to noisy data and has an excellent achievement to handle small dataset. Furthermore, document processing module on question answering system is used to access instances classification result in Quran ontology.

Keywords: *information retrieval, ontology learning, ontology population, ontology, question answering system*

1. Introduction

Quran is the holy book of Muslims which contains the commandment of Allah remarks. As a holy book, Quran includes rich knowledge, instructions and guidance to humankind in achieving happiness in life in the world and the hereafter, and also scientific facts. Based on our previous research toward recent Quran ontology development [1], we have identified several potential issues with it. Instances classification with the small dataset for Quran ontology is one of the problems which appear in previous research. The taxonomic relationship is an association among classes or instances to classes [2, 3]. Previous research by [4-7] performed instances classification to classify the Quran verses based on their themes or thematic topics in Holy Quran. In these case, themes have a role as classes and Quran verses as instances in the ontology. To accomplish instances classification, all of them conducted with the non-automated process, manually by the human. Manual instance classification by domain experts and knowledge engineers is an expensive, not trivial, and time-consuming task. A different method was performed by [8] to implement verses classification. However, they didn't use an ontology to store the knowledge representation, but with the relational database. They used Surah Al-Baqarah as a dataset and *Backpropagation Neural Network* (BPNN) to classify the Quran verses. Then, they categorise it into three classes, i.e. Fasting, Pilgrimage, and None class. Each class has 50 Quran verses as a dataset. 80% of the dataset is used as a training set and the rest as a testing set. Supervised learning such BPNN *classifier*, commonly necessitate a large training data to determine a classifier that performs well. The drawback from Supervised learning, if the training set amount is small, then the classifier accuracy could decline [9-11]. Unfortunately, Holy Quran has a small dataset. It has 114 surah and 6236 verses. Based on this problem, this research aims to formulate a new framework to perform instances classification on a small dataset. The purpose of instances classification according to thematic topics of Holy Quran is to associate the verses with the principal topic to obtain an entire picture of the theme and for providing an improved comprehension to the users [12, 13]. Moreover, the essential aim of instances (verses) classification is to decrease the searching space by determining the passages of information that are suitable for the specific topic [8], [14, 15]. Furthermore, we also

Received May 4, 2018; Revised October 2, 2018; Accepted October 30, 2018

introduced question answering framework that consumes the Quran ontology which stores the result of instances classification. This Quran ontology used to support question answering system to provide suitable information to the users.

2. Ontology Population Techniques

Ontology population is the methodologies for extracting the instances of concepts from the Natural Language text in the external source, then adding them to the existing domain ontology [16-18]. There are several approaches to perform ontology population, i.e. Lexico-syntactic Patterns, Similarity-based Classification, Supervised learning, Knowledge-based and Linguistic Approaches [19]. Another technique to perform ontology population was conducted by [20]. They used Ontology Design Patterns (ODPs) to develop automated domain ontology bilingual in Indonesian and English in tuberculosis malady. Supervised approaches such as Support Vector Machine, Naive Bayes, Decision Tree, and Artificial Neural Network are machine learning algorithms which are used to classify of a particular instance with a model induced from training data [19], [21]. However, the deficiency from Supervised learning, if the training set amount is small, then the classifier accuracy could decline.

3. Question Answering System

Question Answering systems presenting an interface, where users could state their demand for information in the Natural Language format and the search engine will produce suitable answers to these questions [22-24]. More specific described by [25], question answering system is a technology used to find, extract, and provide a proper answer to the user's query in the natural language format. This system consists of several essential components, i.e. question processing module, document processing module, and answer extraction module [24], [26, 27]. There are several QAS approaches based on the method which is used by the document processing module to retrieve the candidate documents from the data source, i.e. linguistic, statistical, semantic, and rule-based approach [28]. This study applied the semantic approach to our question answering framework. Ontology could store the knowledge representation of Quran. While for providing relevant information to the users, semantic search approaches able to extract the knowledge of Holy Quran from an ontology.

4. New Instances Classification Framework Applied to Question Answering System

Based on our study through the literature review, analysing the current instances classification techniques and question answering system, we proposed the new instances classification framework applied to question answering system as shown in Figure 1. Instances classification framework is described in Figure 1 on the left side, while question answering framework on the right side. Both frameworks could be implemented in every language. However, this study focuses on Tafsir of Indonesian Quran translation as an input in instances classification framework and Indonesian as an input for question answering framework. Each language has a different written format, grammar, vocabulary, and syntax [29-31]. Based on this condition, natural language processing technique applied to perform pre-processing, morphology analysis, and semantic analysis is different for every languages. This Sub-Section is organised as follows. Sub-Section 4.1 describes the new instances classification framework, and sub-section 4.2 presents the question answering framework that consumes the Quran ontology which stores the result from instances classification.

As shown in Figure 1, instances classification framework consists of two parts, i.e. construction of training data and classification stage. Description of each part explained in sub-section 4.1.1 and 4.1.2.

4.1. New Instances Classification Framework

The construction of training data stage purpose is to provide training data that will be used in classification stage, while classification stage aims to classify the instances based on thematic topics on Holy Quran.

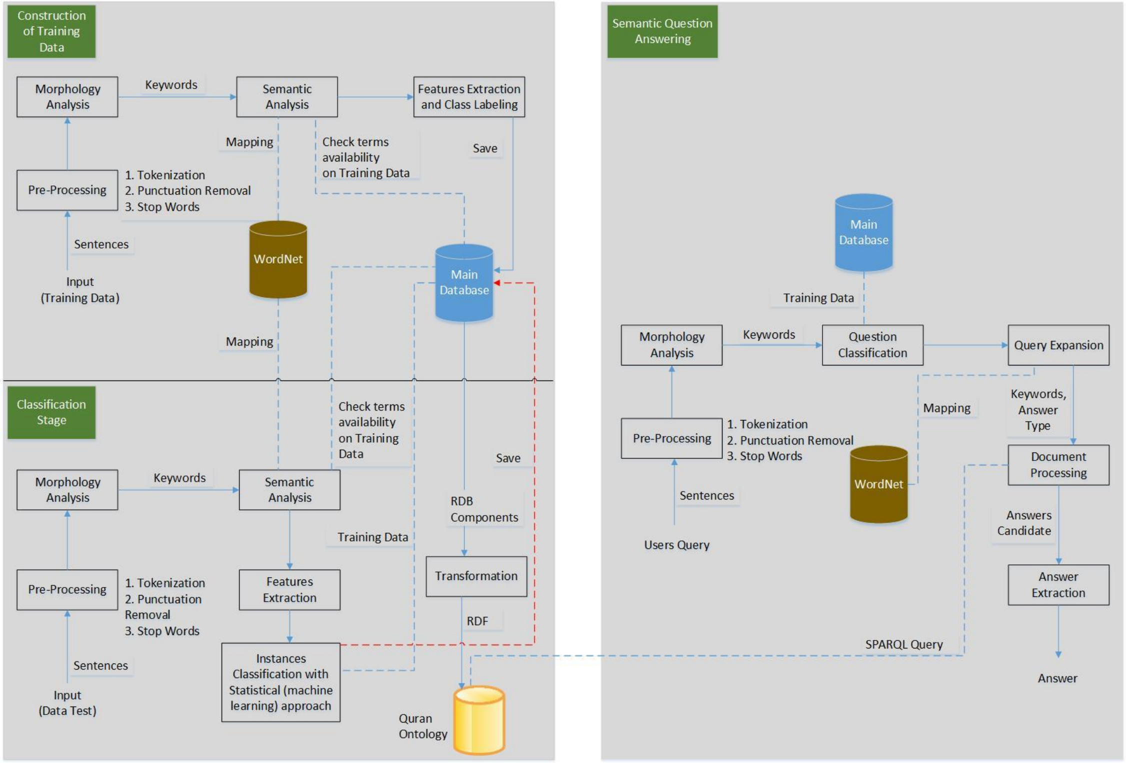

Figure 1. Instances classification framework applied to question answering framework

4.1.1. Construction of Training Data Stage

Tafsir of the Quran verses is used as input on pre-processing phase. The aim of automatic pre-processing phase prepares the words with an appropriate form by tokenisation, disposing the punctuation, and applying stop words removal. The output from the pre-processing phase is used as the input for the morphology analysis phase. At the morphology analysis phase, Tafsir of the verses is processed using stemming operation with stemmer algorithm for Indonesian text from [32] which has been modified. Stemming is an action to modify the words in a document into root word form with a specific rule [33]. More specific, stemming is a process to identify and reduce a derivational or inflectional term to its root form by eliminating all its affixes. We modified stemmer algorithm from [32] since research by [34] has identified several weaknesses in the algorithm. Furthermore, based on our analysis, Adriani stemmer also failed to stem words such as “pemimpi” and “terisi” since derivational suffixes removal are performed first without examining the prefix. Another weakness delivered directly within [32] study, confix pairs “be-. . . –lah”; “be-. . . –an”; “me- . . . –i”; “di-. . . –i”; “pe-. . . –i”; and “te-. . . –i” have incorrectly stemming results since suffixes removal is conducted first before prefixes removal. To solve all the problem that appears, we proposed a modification to [32] algorithm. The flowchart diagram as shown in Figure 2, define the modification for the stemmer algorithm.

The output from morphology analysis phase is essential keywords in root word form. Then, this output is used as input in the next step, semantic analysis. The aim of the semantic analysis stage to expand the keywords representation on Tafsir of Quran verses since words on the dataset (main database) might employ another keyword. Moreover, this stage aim also to minimise and delimit the words representation that appears in the training data. Without semantic analysis on the keywords, then the data set shall have a large number of features. This study used Indonesian WordNet from [35] to conduct semantic analysis on the keywords. Based on verification method and synsets quality, their Indonesian WordNet consists of five categories, i.e. "Y", "O", "M", "L", and "X". Where "Y" and "O" in high quality, "M" in medium quality, "L" probably poor quality, and "X" in poor quality. The algorithm in pseudocode as shown below is used to describe the flow of semantic analysis process for prepare the training data on the dataset.

```
string s = text
string[] words = split the text within s variable by whitespace as text
boundary
string[] new_words = null
foreach (string word in words)
   string[] lemma = null
   if the word not found in the dataset then
      if the word found in the Indonesian WordNet then
         if the word found with "Y" quality then
           string synset_number = get the synset number
            lemma = get the lemma based on synset_number variable
         end if
         if the word found with "O" quality then
            string synset_number = get the synset number
            lemma = get the lemma based on synset_number variable
         end if
         lemma = remove duplicate words in the lemma array variable
         string string_temp = word
         foreach (string term in lemma)
             if term found in the dataset then
                 string_temp = term
                 break
             end if
         end foreach
         add string_temp into new_words variable
      end if
      else
         add word into new_words variable
   end if
   else
      add word into new_words variable
end foreach
```

The output from the semantic analysis phase is the new keywords based on the Indonesian WordNet. This output is used as an input on features extraction and class labelling stage. On this phase, features that extracted are Surah and verses number, keywords, term frequency, and class (themes) labelling for each Tafsir of the Quran verses according to thematic topics on Quran. This labelling manually is conducted by the human. The keywords, term frequency, and class labelling on each Tafsir are used to create the Term Frequency-inverse Document Frequency (TF-IDF) model which will be used in the instances classification stage. All of the features then store in the data set as a training data for classification stage.

4.1.2. Classification Stage

The test data on classification stage is processed with specific components. These components are pre-processing, morphology analysis, semantic analysis, feature extraction, and instances classification with a statistical approach. Tafsir of the verses is used as input on classification stage. Pre-processing, morphology analysis, semantic analysis, feature extraction explanation is equal to these components application on the construction of training data phase. Furthermore, since features extraction phase is finished, thereupon the TF-IDF model is classified by Radial Basis Function Networks (RBFN) algorithm [36]. This model utilised the feature extraction data from the previous and current stage. RBFN has specific benefits. There are hardiness to noisy data [37, 38] and has an excellent achievement to handle small data set [39, 40]. A study by [41] has shown with the F1 measure that RBFN has better performance and more consistent than BPNN to perform text classification on small newsgroup dataset.

The classification results by RBFN then stored in the main database (relational database). Finally, tools such as DataMaster [42] and OntoBase [43] as a plugin for Protégé editor is used to convert the relational database into an ontology. This conversion applied several mapping rules to transform the relational database components such as tables, columns, tuples, and foreign keys into ontology components such as classes, properties, and instances. Figure 3 shows an example of direct mapping from a relational database into an ontology.

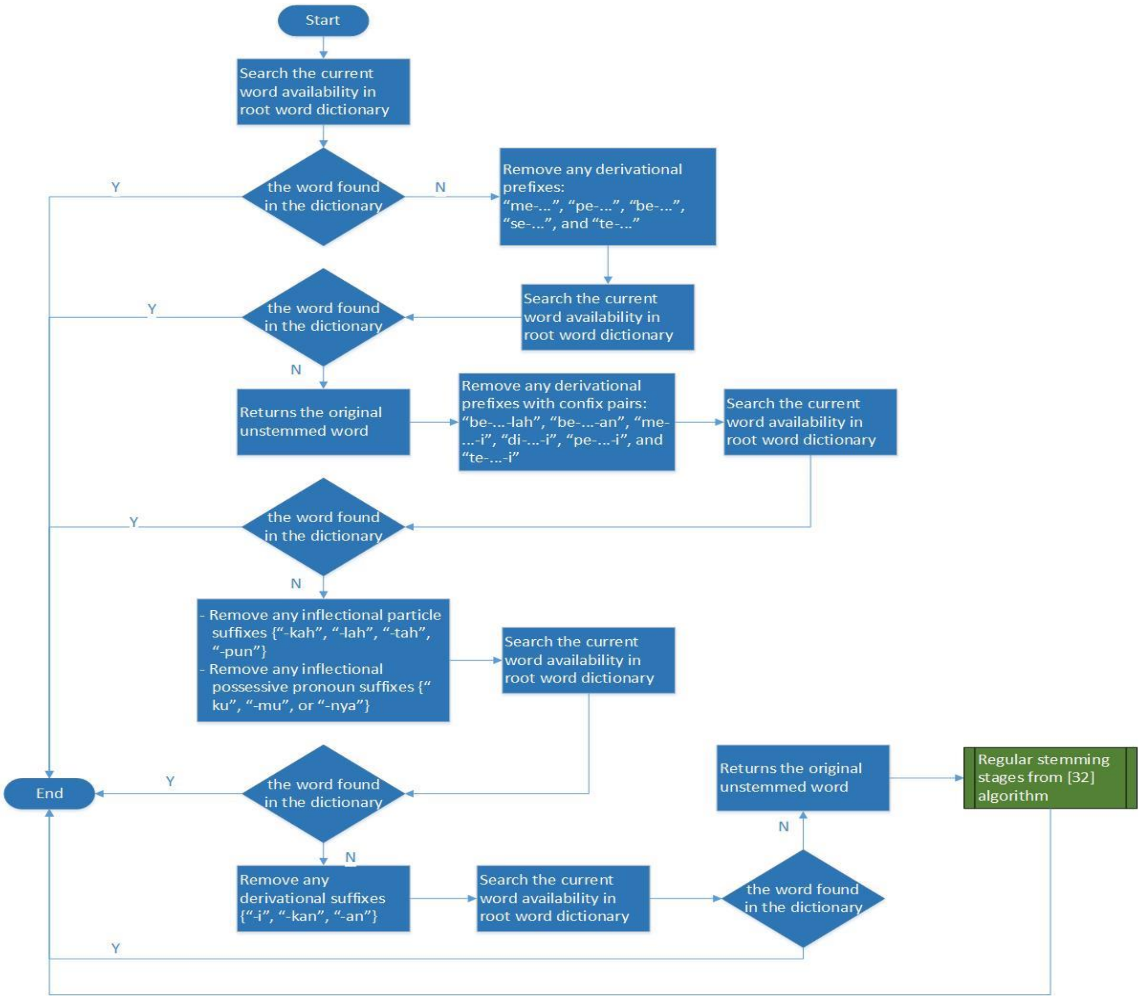
Start
Search the current word availability in root word dictionary
the word found in the dictionary
Y
N
Remove any derivational prefixes: “me-...”, “pe-...”, “be-...”, “se-...”, and “te-...”
Search the current word availability in root word dictionary
the word found in the dictionary
Y
N
Returns the original unstemmed word
Remove any derivational prefixes with confix pairs: “be-...-lah”, “be-...-an”, “me-...-i”, “di-...-i”, “pe-...-i”, and “te-...-i”
Search the current word availability in root word dictionary
the word found in the dictionary
Y
N
- Remove any inflectional particle suffixes {“-kah”, “-lah”, “-tah”, “-pun”}
- Remove any inflectional possessive pronoun suffixes {“ku”, “-mu”, or “-nya”}
Search the current word availability in root word dictionary
End
Y
the word found in the dictionary
Returns the original unstemmed word
Regular stemming stages from [32] algorithm
N
N
Remove any derivational suffixes {“-i”, “-kan”, “-an”}
Search the current word availability in root word dictionary
the word found in the dictionary
Y

Figure 2. Stemmer algorithm modification

As shown as in Figure 3, there are one to many relationships between two tables from a relational database, i.e. *thematic index* to store the topics/themes in Holy Quran; and *verses* to save Quran surah name, verse number, Arabic text of Quran verse, Indonesian translation of Quran verse, Tafsir of the verse, and themes of the verse. This relationship mapped into a subclass of the main ontology class, i.e. *verses* as the subclass and *thematic_index* as the main class. Next, all columns except the foreign key in the relational database are mapped into Datatype Properties. For example, based on Figure 3, *thematic_id* and *thematic_domain_name* as columns on *thematic_index* would be mapped into an integer (represent the *thematic_id*) and string (represent the *thematic_domain_name*) data type as properties for *thematic_index* ontology class. Furthermore, verses classification result which is stored on *thematic_id* column will be mapped into ontology as instances of a class.

4.2. Question Answering Framework

As shown in Figure 1, our question answering framework consists of several components, i.e. pre-processing, morphology analysis, question classification, query expansion, document processing, and answer extraction. The input for the pre-processing stage is user queries in a factoid question, "what" question. Furthermore, this query is processed by pre-processing and morphology analysis phase with the similar technique which utilised on instances classification stage. The output from morphology analysis phase is essential keywords in root word form. Moreover, these keywords are used as an input in question classification stage. This stage aims to extract and determines answer type by performing question classification with RBFN algorithm. The answer type based on thematic topics which available on Holy Quran. Training data which utilised to perform question classification is equal to the data which used in instances classification stage. Furthermore, the keywords from the morphology analysis stage are used as an input in query expansion phase. This phase aims to expand the keywords using query expansion technique. This technique applies Indonesian WordNet [35] to expand the keywords. The output from query expansion stage is the origin keywords and their synonym. Subsequently, this output is used as an input in document processing stage. Moreover, the answer type as the output from question classification phase also used as an input in it.

The document processing stage performed execution of SPARQL query with some parameters, i.e. the answer type and keywords from the previous phase. This answer type would access the class of thematic topics within Quran ontology, and the keywords are used to extract the instances based on the class. This execution generates answers candidate, i.e. Quran verses and their Tafsir. The output from this phase is used as the input in the answer extraction stage. Finally, at the answer extraction stage, words matching scoring technique is applied to rank the verses and their Tafsir. This technique computes the number of similar words between the expanded query and the Tafsir of the verses, then rank all these verses score. The best answer is determined based on the verse with the highest score.

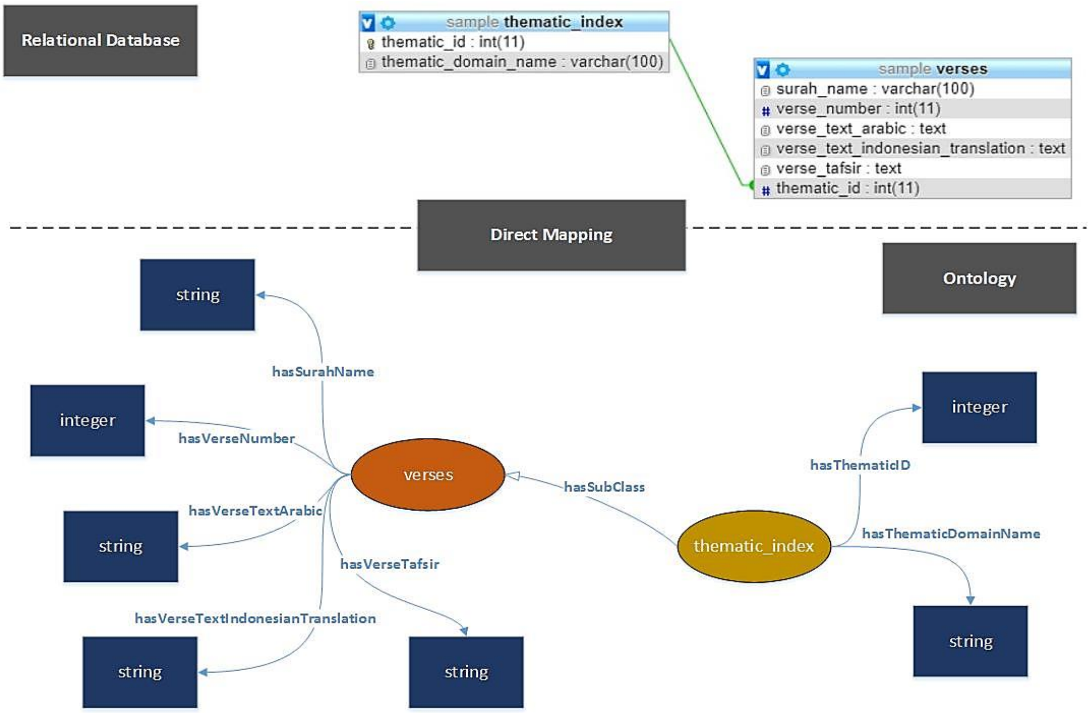

Figure 3. Transformation from relational database to ontology

5. Conclusion

This study introduces an instances classification framework on Quran ontology applied to the semantic question answering system (QAS). The instances classification framework consists of two stages, i.e. construction of training data and classification stage. There are several essential components to classify the instances, i.e. pre-processing, morphology analysis, semantic analysis, features extraction, instances classification with the statistical or machine learning approach: Radial Basis Function Networks (RBFN), and transformation stage. RBFN algorithm is chosen since it robustness to noisy data and has an excellent achievement to handle small dataset. Our proposed framework work for small training corpus like Quran.

Furthermore, document processing stage on question answering system is used to access instances classification result in Quran ontology through SPARQL query execution. This query contains some parameters, i.e. the answer type from question classification phase and keywords from morphology analysis phase on QAS. This answer type would access the class of thematic topics within Quran ontology, and the keywords are used to extract the instances based on the class. For the next research, we will develop and test the instances classification framework. Moreover, we also build and test the semantic question answering system which accesses the Quran ontology. Evaluation against instances classification framework for measuring and improving the framework performance to classify the instances with small training corpus. Furthermore, the evaluation of the QAS is performed to measure the QAS output accuracy for the answer was given to the users.

Acknowledgments

The authors would like to thank the financial support from STMIK AMIKOM Purwokerto; and Center for Advanced Computing Technology (C-ACT), Fakulti Teknologi Maklumat dan Komunikasi, Universiti Teknikal Malaysia Melaka (UTeM) for their assistance in this research.

References

[1] Suryana N, Utomo F S, Azmi M S. Quran Ontology: Review on Recent Development and Open Research Issues. *Journal of Theoretical and Applied Information Technology*. 2018; 96(3): 568-581.

[2] Zong N, Nam S, Eom J-H, Ahn J, Joe H, Kim H-G. Aligning ontologies with subsumption and equivalence relations in Linked Data. *Knowledge-Based Systems*. 2015; 76: 30-41.

[3] Bilgin G, Dikmen I, Birgonul M T. An Ontology-based Approach for Delay Analysis in Construction. *KSCE Journal of Civil Engineering*. 2018; 22(2): 384-398.

[4] Hakkoum A, Raghay S. *Advanced Search in the Qur'an using Semantic modeling*. IEEE/ACS 12th International Conference of Computer Systems and Applications (AICCSA). Marrakech. 2015: 1-4.

[5] Periamalai N S H A R, Mustapha A, Alqurneh A. *An ontology for Juz' Amma based on expert knowledge*. 7th International Conference on Computer Science and Information Technology (CSIT). Amman. 2016: 1-5.

[6] Ta'a A, Abidin S Z, Abdullah M S, Ali A B B M, Ahmad M. *Al-Quran Themes Classification using Ontology*. The 4th International Conference on Computing and Informatics (ICOCI). Sarawak. 2013: 383-389.

[7] Ta'a A, Abdullah M S, Ali A B M, Ahmad M. *Themes-based classification for Al-Quran knowledge ontology*. International Conference on Information and Communication Technology Convergence (ICTC). Busan. 2014: 89-94.

[8] Hamed S K, Aziz M J A. A Question Answering System on Holy Quran Translation Based on Question Expansion Technique and Neural Network Classification. *Journal of Computer Sciences*. 2016; 12(3): 169-177.

[9] Asiaee A H, Minning T, Doshi P, Tarleton R L. A framework for ontology-based question answering with application to parasite immunology. *Journal of Biomedical Semantics*. 2015; 6(31): 1-25.

[10] Deng Zhi-Hong, Luo Kun-Hu, Yu Hong-Liang. A study of supervised term weighting scheme for sentiment analysis. *Expert Systems with Applications*. 2014; 41(7): 3506-3513.

[11] Tomás D, Vicedo J L. Minimally supervised question classification on fine-grained taxonomies. *Knowledge and Information Systems*. 2013; 36(2): 303–334.

[12] Ismail R, Bakar Z A, Rahman N A. Extracting Knowledge From English Translated Quran using NLP Pattern. *Jurnal Teknologi*. 2015; 77(19): 67-73.

[13] Farooqui N.K, Noordin M F. Knowledge Exploration: Selected Works on Quran Ontology Development. *Journal of Theoretical and Applied Information Technology*. 2015; 72(3): 385-393.

[14] Oh Hyo-Jung, Myaeng S.H, Jang Myung-Gil. Semantic passage segmentation based on sentence topics for question answering. *Information Sciences*. 2007; 177(18): 3696-3717.

[15] Khan A, Baharudin B, Lee L.H, khan K. A Review of Machine Learning Algorithms for Text-Documents Classification. *Journal of Advances in Information Technology*. 2010; 1(1): 4-20.

[16] Mitzias P, Riga M, Kontopoulos E, Stavropoulos T.G, Andreadis S, Meditskos G, Kompatsiaris I. *User-Driven Ontology Population from Linked Data Sources*. International Conference on Knowledge Engineering and the Semantic Web. Prague. 2016: 31-41.

[17] Garanina N, Sidorova E, Kononenko I. *A Distributed Approach to Coreference Resolution in Multiagent Text Analysis for Ontology Population*. International Andrei Ershov Memorial Conference on Perspectives of System Informatics. Moscow. 2017: 147-162.

[18] Ganino G, Lembo D, Scafoglieri F. *Ontology Population from Raw Text Corpus for Open-Source Intelligence*. ICWE: International Conference on Web Engineering. Rome. 2017: 173-186.

[19] Cimiano P. Ontology Learning and Population from Text. First Edition. New York: Springer US. 2006: 234-237.

[20] Harjito B, Cahyani D E, Doewes A. An Automatic Approach for Bilingual Tuberculosis Ontology Based on Ontology Design Patterns (ODPs). *TELKOMNIKA Telecommunication Computing Electronics and Control*. 2018; 16(1): 282-289.

[21] Cimiano P, Mädche A, Staab S, Völker J. Ontology Learning. In: Staab S, Studer R. Handbook on Ontologies. First Edition. Berlin: *Springer*, Berlin, Heidelberg. 2009: 245-267.

[22] Chen Chi-Hua, Wu Chen-Ling, Lo Chi-Chun, Hwang Feng-Jang. An Augmented Reality Question Answering System Based on Ensemble Neural Networks. *IEEE Access*. 2017; 5: 17425 - 17435.

[23] Bakis R, Connors D P, Dube P, Kapanipathi P, Kumar A, Malioutov D, Venkatramani C. Performance of natural language classifiers in a question-answering system. *IBM Journal of Research and Development*. 2017; 61(4): 14:1 - 14:10.

[24] Abdi A, Idris N, Ahmad Z. QAPD: an ontology-based question answering system in the physics domain. *Soft Computing*. 2018; 22(1): 213–230.

[25] Pavlić M, Han Z D, Jakupovic A. Question answering with a conceptual framework for knowledge-based system development "Node of Knowledge". *Expert Systems with Applications*. 2015; 42(12): 5264-5286.

[26] Abacha A B, Zweigenbaum P. MEANS: A medical question-answering system combining NLP techniques and semantic Web technologies. *Information Processing and Management*. 2015; 51(5): 570-594.

[27] Diefenbach D, Lopez V, Singh K, Maret P. Core techniques of question answering systems over knowledge bases: a survey. *Knowledge and Information Systems*. 2018; 55(3): 529–569.

[28] Utomo F S, Suryana N, Azmi M S. Question Answering System : A Review on Question Analysis, Document Processing, and Answer Extraction Techniques. *Journal of Theoretical and Applied Information Technology*. 2017; 95(14): 3158-3174.

[29] Andi-Pallawa B, Alam A F A. A Comparative Analysis between English and Indonesian Phonological Systems. *International Journal of English Language Education*. 2013; 1(3): 103-129.

[30] Chiswick B R, Miller P W. Linguistic Distance: A Quantitative Measure of the Distance Between English and Other Languages. *Journal of Multilingual and Multicultural Development*. 2005; 26(1): 1-11.

[31] Rahayu A U. Differences on Language Structure between English and Indonesian. *International Journal of Languages, Literature and Linguistics*. 2015; 1(4): 257–260.

[32] Adriani M, Asian J, Nazief B, Tahaghoghi S M M, Williams H E. Stemming Indonesian: A Confix-Stripping Approach. *ACM Transactions on Asian Language Information Processing*. 2007; 6(4): 1-33.

[33] Ruhwinaningsih L, Djatna T. A Sentiment Knowledge Discovery Model in Twitter's TV Content Using Stochastic Gradient Descent Algorithm. *TELKOMNIKA Telecommunication Computing Electronics and Control*. 2016; 14(3): 1067-1076.

[34] Purwarianti A. *A Non Deterministic Indonesian Stemmer*. IEEE: International Conference on Electrical Engineering and Informatics (ICEEI). Bandung. 2011: 1-5.

[35] Bond F, Lim L T, Tang E K, Riza H. The combined Wordnet Bahasa. *NUSA: Linguistic studies of languages in and around Indonesia*. 2014; 57: 83-100.

[36] Sarimveis H, Doganis P, Alexandridis A. A classification technique based on radial basis function neural networks. *Advances in Engineering Software*. 2006; 37(4): 218-221.

[37] Derks E P P A, Pastor M S S, Buydens L M C. Robustness analysis of radial base function and multi-layered feed-forward neural network models. *Chemometrics and Intelligent Laboratory Systems*. 1995; 28(1): 49-60.

[38] Walczak B, Massart D L. Local modelling with radial basis function networks. *Chemometrics and Intelligent Laboratory Systems*. 2000; 50(2): 179-198.

[39] Lanouette R, Thibault J, Valade J L. Process modeling with neural networks using small experimental datasets. *Computers and Chemical Engineering*. 1999; 23(9): 1167-1176.

[40] Tudu B, Jana A, Metla A, Ghosh D, Bhattacharyya N, Bandyopadhyay R. Electronic nose for black tea quality evaluation by an incremental RBF network. *Sensors and Actuators B: Chemical*. 2009; 138(1): 90-95.

[41] Motwani M, Tiwari A, Sharma S. *Investigation of BPNN & RBFN in text classification by Active search*. IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT). Coimbatore. 2015: 1-6.

[42] Jelokhani-Niaraki S, Tahmoorespur M, Minuchehr Z. An Ontology-Based GIS for Genomic Data Management of Rumen Microbes. *Genomics & Informatics*. 2015; 13(1): 7-14.

[43] Mogotlane K D, Fonou-Dombeu J V. Automatic Conversion of Relational Databases into Ontologies: A Comparative Analysis of Protégé Plug-Ins Performances. *International Journal of Web & Semantic Technology (IJWesT)*. 2016; 7(3/4): 21-40.