ONTOLOGY POPULATION FROM QURANIC TRANSLATION TEXTS BASED ON A COMBINATIONOF LINGUISTIC PATTERNS AND ASSOCIATION RULES

[1]TAHER WEAAM,[2]SAIDAH SAAD

Center for Artificial Intelligence Technology (CAIT)

E-mail: [1]vip_twm@yahoo.com,[2]saidah@ukm.edu.my

ABSTRACT

With the increasing volume of English translation of Islamic documents available on the web, there is a need to retrieve and extract important information in order to fully understanding these documents. Understanding the Quran is a grand challenge for society, for western public education, for Muslim-world education, for knowledge representation and reasoning and for knowledge extraction from text. Ontology learning from the Quran text is challenging task due to the nature of the Quran text which has scattered organization of knowledge and its unique feature. This paper illustrates an ontology learning based on a hybrid method which combines lexico-syntactic patterns and association rules for English translation of the meaning of the Quran text. First, this paper designs a new two layers of filtering method which combine linguistic and statistical methods for concept extraction. Second, this work designs a new hybrid method based on lexico-syntactic patterns and association rules method for relation extraction. The results showed that using the two layers of extraction prove to be adequate and efficient measures for automatic extraction of Quranic concepts with an overall F-measure of 85.3%. In addition, the results obtained indicate that the used methods are very suitable technique for extracting relation from with an overall F-measure of 87.3% and 88.3% respectively.

Keywords: *Ontology Learning, Statistical Methods, Pattern Extraction, Association Rules, Quran*

1. INTRODUCTION

In recent years, the notion of ontology has emerged and become common for researchers and scientists in the world. Web Ontologies can be ranged from large taxonomies categorizing Web sites (such as on Yahoo, Ask and AOL) to categorizations of products for sale and their features (such as on Amazon.com and ebay.com). Ontology provides a solution to capture information about concepts and relations between those concepts in the same domain. Gruber [1] defined ontology as “A formal explicit specification of a shared conceptualization.” Ontology can be represented in a declarative form by using web semantic languages such as RDF, OWL or XML. There are many potential benefits of using ontology in representing and processing knowledge, including the separation of domain knowledge from application knowledge, sharing of common knowledge of subjects among human and computers, and the reuse of domain knowledge for a variety of applications.

This paper aims to create an ontology extraction tool for Holy Quran using the association rule algorithm. This involves investigating the most proper methods for dealing with the Quran’s text collection to extract concepts and semantic relations between them. Using this Quran ontology, various text mining tasks including information extraction, text categorization, concept linkage, and discovery of associations and patterns can be tested and evaluated. The paper considers investigating the literal meaning of Quran text as a pilot benchmark before focusing later on the entire Islamic studies collection to build a fully functional ontology for Islam knowledge. The ontology will provide a powerful representation of Quran knowledge, with the rule schemas giving a more expressive representation of Quran relations in term of rules.

With rapid advance of internet technologies, the amount of electronic Islamic documents has drastically increased worldwide. As consequence, it this makes the process of understanding and extracting useful information using normal search engine a very difficult task. Understanding the Islamic documents is a grand challenge for society, for Muslim-world education, for knowledge representation and reasoning, for knowledge extraction from text, for systems robustness and correctness, for online collaboration. The creation of a complete Islamic named entity, recognition for the Hadith and Islamic Knowledge can help in fully understanding of the Quran.

2. RELATED WORK

In the recent years, Several studies have been made to understand the Quranic and Islamic text and extract knowledge from it using computational linguistics .Dukes [2] uses Named Entity Extraction from the Quranic Text in order to extract concept/instances and associate it in the ontological concept. Al-Yahya, et al. [3] presents the design and implementation of the ontological model and the results of its application is on the “Time nouns” vocabulary of the Quran. Abbas [4]is a Quranic corpus where it is augmented with an ontology or index of key concepts, manually done, taken from the 'Mushaf Al Tajweed'. It contains a comprehensive hierarchical index or ontology of nearly 1200 concepts in the Quran. Al-Kabi, et al. [5] uses the automatic classification technique on the Quranic verses based on certain surahs (chapters) according to the classifications made by the Islamic Scholars. The automatic extraction is carried out just for the terms and concept layers[2, 3]. The classification of those concepts/terms is done manually. Access to the Quran has traditionally been through the text.

Saad, et al. [6] use patterns to identify the key concepts that exist in the Islamic domain, their properties and the relationships that hold between them.

Saad and Salim [7]evaluates several methods for extraction keyword and key phrase candidate in order to develop ontology for Islamic Knowledge. They extract candidate keywords and key phrases using lexico-syntactic method and statistical method.

Harrag, et al. [8]use association rules to extract the ontology of prophetic narrations (Hadith). Their approach involves investigating the use of association rules to identify frequent item sets over concepts that are related to Islamic jurisprudence (Fiqh) from the Sahîh Al-Bukhârî documents by computing correspondence relations using the A priori algorithm. In particular, the semantic structure of the Sahîh Al-Bukhârî as a knowledge source is exploited to extract a specific domain ontology, while the conceptual relations embedded in this knowledge source are modeled based on the notion of association rules. Azmi and bin Badia [9]automatically generates the narrators’ chain of a given Hadith and graphically displays it.

This process involves parsing and annotating the Hadith text and recognizing the narrators’ names. They also use shallow parsing along with a domain specific grammar to parse the Hadith content. Moreover, a transformation mechanism based on semantic web ontology to represent the narration chain in a standard format and then graphically render its complete tree. They parse a plain Hadith text and automatically generate the full narration tree. It involves parsing and annotating the Hadith text and recognizing the narrators’ names. Harrag [10]detect and extract passages or sequences of words containing relevant information from the prophetic narrations text using finite state transducers-based system. Experimental evaluation results demonstrated that our approach is feasible. Their system achieved precision and recall rates, the overall precision and recall are 71% and 39% respectively.

More computational analysis of the Quran will be of great advantage in helping Moslems to understand the Quran, yet there has been little analysis performed on it. At present, the most promising techniques in ontology creation are the combination of rule-based approach using NLP and the machine learning [11, 12].

3. PROPOSED SYSTEM

This work provides a hybrid method based on linguistic pattern sand association rules machine learning for ontology extraction from English translation of Quran text. The designed method covers multi-word and single words semantics - domain concepts. In this part, we give an overview of the architecture of Islamic ontology extraction system and describe the functionality of each component in Islamic ontology extraction system. A general architecture for the proposed methodology of the system is depicted in Figure 1. However, as shown in Figure 1, the general architecture of the Quran ontology learning consists of the following phases:

1. The preprocessing phase
2. Concept extraction and filtering phase
3. Relations extraction phase
4. Evaluation phase.

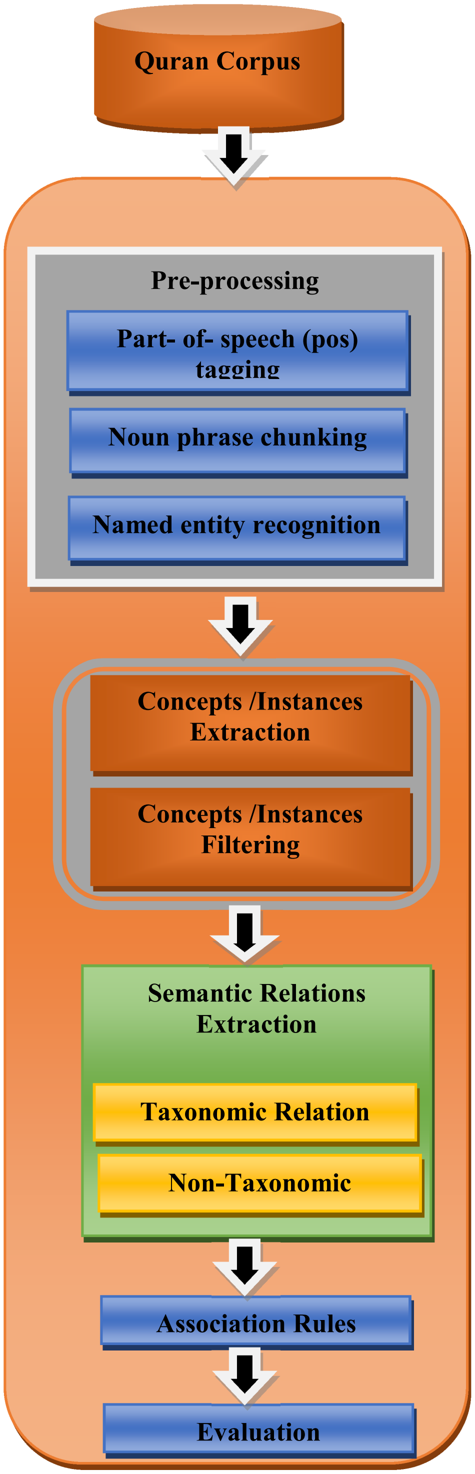

Figure 1. Proposed method framework

3.1. Preprocessing Phase

The Preprocessing is a task of ensuring a dataset is ready for the concept and relation extraction. Each document must pass through the preprocessing phase. In this work, each new Quran verse goes through the several pre-processing tasks. The preprocessing tasks for ontology learning include:

3.1.1 Normalization

It is of essential usefulness to normalize the noisy text before employing NLP tasks. In this work, the goal of text normalization is to remove "noise" from translation of Arabic names such as Ka'aba is converted to Kaaba. Some other non-alphabetic symbols (numbers, special characters, etc.) are to be removed to obtain clear and ready dataset.

3.1.2 Tokenization

Tokenization: Tokenization is the first step of any natural language processing task which identifies individual word units separated by a space or other punctuation characters. Many NLP tasks works on individual words in the corpus. For example, the part of speech tagger assigns a part of speech tag to each individual word, perhaps taking the sentence context into account.

3.1.3 POS tagging

Part of Speech (POS) tagging is the ability to computationally assign grammatical word categories to individual words or other tokens such as numbers. Words can be divided into categories that behave similarly. Traditionally, there are eight parts of speech: noun, verb, pronoun, preposition, adverb, conjunction, adjective and article A tag set defines how certain features are represented. For example in the Penn Treebank tag set, a singular common noun is tagged NN, while a plural common noun is tagged NNS. This work adopted Part-Of-Speech tagging system that is developed by The Stanford Natural Language Processing Group Stanford [13], which is based on Penn Treebank Tag set for English.

3.1.4 Named Entity Recognition

Named entity Recognition: is the process of identification of instances of world entities such as a person, location or organizations, referenced by name in the dataset. However, NER is very important for identifying ontological concepts and/or instances and for identifying attributes and relations of these instances. This work adopted named entity recognition system that is developed by The Stanford Natural Language Processing Group, which is based on three classes person, location or organization for English.

3.2. Concepts Extraction and Filtering

Term Extraction is a fundamental prerequisite for all aspects of more complex tasks such as semantic search and ontology learning, as mentioned in the ontology layered cake[14]. The aim of this phase is building the ontology lexicon. This work describes new two layers of filtering method which combine linguistic and statistical methods for concept extraction. After selecting the potential candidate terms, two layers of filtering linguistics and statistical are used for ranking and selecting Quran specific concept from all candidate concepts extracted from the corpus. The following describe the two layers of filtering:

3.2.1 Linguistic Concept Filtering: The statistical information, without any linguistic filtering, is not enough to produce useful results. Without any linguistic information, undesirable strings such as of the, is a, etc., would also be extracted. First, Terms consist mostly of nouns and adjectives and sometimes prepositions. The linguistic filter is used to extract noun phrases that constitute multi-word terms discarding such undesirable strings. Shallow parsing and Part-of-Speech (POS) is applied prior to linguistic filters. This work implemented all the filters below using the regular expression pattern:

1. Noun$^+$ Noun
2. (Adj | Noun)$^+$ Noun
3. ((Adj |Noun)$^+$| ((Adj | Noun)*(Noun Prep)?)(Adj | Noun)*) Noun

The output of this phase is the entire linguistic-acceptable candidate terms over which will be passed to the statistical filtering.

3.2.2 Statistical Concept Filtering: In this work, a combination of two statistical approaches is used to rank terms based on their occurrence in the corpus. This method is suitable for extracting single and multi-word terms from a domain corpus, assigning a numeric value to each candidate string, where a high value indicates important candidates probable concepts The procedure for Statistical Concept Filtering is described below.

For each concept t, compute a linear combination

$LN(t)=\alpha_1 TFIDF\ (t)+ \alpha_2 TIM(t)$

Where $\alpha_{1,}\ \alpha_2 \in[0,1]$is a weighting parameter and

$\alpha_1+ \alpha_2=1$.

1. TFIDF (term frequency inverse document frequency)) : the method assigns a weight to a term based on two measures: (1) tf the frequency ofoccurrence of a term within a single verse , and (2) df thenumber of versesin the Quran corpus which contain the giventerm. N is the total number of verses

$$tfidf(c) = \quad tf(c) * \log(\frac{N}{df(cw)})$$

2. TIM : the method assigns aweigh to a term based on two measures: (1) tf the frequency ofoccurrence of a concept, and (2) ndf thenumber of verses in the corpus which contain the given term. max(tf)is the maximumfrequency among all terms frequencies.

$$TIM(t) = \frac{ndf(t)}{2}\left(\frac{tf(t)}{\max(tf)} + 1\right)$$

3.3. Taxonomic Relations Extraction

These relations give a hierarchical structure to ontology. There are some suggested methods for extracting this relation from text. To extract relations between extracted concepts, this work adopted:

IS-A Patterns: Several lexico-syntactic patterns are adopted for extracting "IS-A" relations are depicted in Table 1. This work measure the confidence value for each matches lexico-syntactic patterns. This confidence is computed through the following formula:

$$Conf(Pattern) = \frac{freq(NP1, NP2)}{freq(NP1)}$$

Table 1. IS-A Relation Pattern

IS-A PATTERN	AUTHOR
<NP>s such as <NP >	Hearst[15]
such <NP>s as <NP >	Hearst[15]
<NP>, (especially | including) <NP >	Hearst[15]
< NP > (and | or) other < NP >	Hearst[15]
the < NP >< NP >	Cimiano[14, 16]

the < NP >< NP >	Cimiano[14, 16]
<NP >, a <NP>	Cimiano[14, 16]
<NP > is a <NP>	Cimiano[14, 16]
<NP > a/as a <NP>	Saad[17]
< NP > verb-to-be<NP>	Saad[17]

Compound Nouns Relations: Several lexico_synatic patterns are used to detect compound nouns relations. This work uses the patterns adopted in[17] for learning these relations:

1. The N1N0
2. N0'sN1
3. NP→JJ, NP0
4. head of Noun where N0 is part-of N1N0 /The N0 of N1
   For example “funeral prayer” is part of “pray”

3.4. Non-Taxonomic Relations Extraction

Non-Taxonomic relations are any other relations between two concepts other than taxonomic relations. These relations may convey possession, causation, synonymy, antonymy or any other kind of relations. This work present two kinds of methods which are association rules and linguistic patterns:

Subject Verb Object Relations: In addition, this work focus on learning non-hierarchical relations of form<C_i, V, C_j>, Here V is a non-hierarchical relation, and C_i, C_j are concepts.

- <Allah, sent, messengers with glad tidings and warnings >
- <Allah, give, guidance , life and death >
- <Allah, made , the Kaba, the Sacred House>

The proposed algorithm for finding <SUBJECT VERB OBJECT> relations is as follows: The algorithm takes as input the syntax dependency parsing from the full parser:

- Parse the text using full parsing, get the syntax dependency parsing information and extract the patterns like “Subject verb Object”. For example, from the Syntax dependency parsing: nsubj(made-2, Allah-1) , dobj(made-2, Kaba-4),
- For each verb, extract (subject, object) tuples.
- For each node which represents a relation a confidence value should be assigned. The confidence value is computed using the following formula:

$$Conf(relation < sub, verb, obj >) = \frac{df(verb)}{N} * \frac{df(sub, verb, obj)}{freq(verb)}$$

Where df (verb) stands for the number of documents in which verb is appeared in. Also N stands for the total number of documents in the dataset, df(sub, verb, obj)is the number of times which(sub, verb, obj) tuple is appeared. freq(verb)isthe total frequency of this verb.

AssociationRules: Association rules are commonly used to discover data, text elements or patterns that co-occur frequently within a dataset. Such patterns can be used to make predictions on data. In ontology learning, transactions are defined in terms of words occurring together in certain syntactic dependencies. If the rule X → Y has been generated and stored, we can conclude that there is a relationship between the concepts in X and the concepts in Y (Gulla&Brasethvik, 2008). The main reasons of usingAssociation rules in this work is to find the implicit relations between concepts which cannot be discovered using the predefined lexico-syntactic patterns. Given a set of concepts $A = \{C_1, C_2, \dots\dots, C_n\}$ and a setof documents $d = \{d_1, d_2, \dots\dots, d_m\}$,X and Y are set of concepts which are subsets of A $X \subseteq A \; and \; Y \subseteq A$, the association rule Rof(X) and(Y)is computed as follows:

$$C(X,Y) = \frac{(X,Y).COUNT}{X.COUNT}$$

$C(X,Y)$is called the confidence ofR, with respect to collection T. C(X,Y) is the probability of existing a concept set Xin document if there is already concept set Y in the same document. For example if the two concepts (Allah, prophet) appear together in 24 verses and “Allah” appears in 48 verses so$C(X,Y) = \frac{24}{48} = 0.5$.

4. PERFORMANCE MEASURES

The performance of the machine learning algorithms is measured on manually labeled Islamic corpus. The evaluation was performed on a corpus of English translation of Quran text. In order to evaluate ontology learning techniques to extract ontology from the Quran text. The datasets which are used as the sources for the information-extraction ontology are the English Quranic translation text by Yusuf Ali [18]. In addition, this work is evaluated using the English extended Quranic translation text, which summarizes the information according to the tafsir of At-Tabari, Al-

Qurtubi and Ibn Kathir, with comments based on Sahih Al-Bukhari by Dr. Muhammad Taqiud-Din Al-Hilali, and Dr. Muhammad Muhsin Khan [19]Selected 100 Qurans verses from each dataset are used. All Qurans are in English language and available in txt format. The text was placed in a directory as the input corpus. The corpus was imported into an ontology learning prototype. The most common performance metrics used for information extraction, which are similar to the information retrieval systems, are precision, recall and F-measure.

Precision measure is used to measure the accuracy of the information extraction system, as the percentage of information extracted correctly by the system.

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

Recall measure is defined as the percentage of relevant information which can be extracted by a system, or the ability of the system to retrieve all the relevant items in the corpus.

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

5. RESULTS

In this experiment, several experiments have been performed to empirically evaluate the two layer filtering (linguistics and statistical) algorithm for ranking and selecting Quran specific concept from all candidate concepts extracted from the corpus. The linguistic filter is used to extract noun phrases that constitute multi-word terms discarding such undesirable strings. A combination of statistical approaches (*discussed in section 3.2.2*) TFIDF and TIM are used for ranking these multi-word terms. The two layers of filtering method are suitable for extracting single and multi-word terms from a domain corpus. The performance (precision, recall and F-measure) values metrics for the three statistical methods are shown in Table 2 and Table 3.

Table 2. The Performance (Precision, Recall And F-Measure) For The Three Statistical Approaches (English Extended Quranic Translation Text, By Al-Hilali, And Khan)

Top % concept	Precision	Recall	F-measure
TFIDF	0.89	0.81	0.86
TIM	0.88	0.79	0.89
Linear Combination	0.90	0.87	0.87

Table 3.3 The Performance (Precision, Recall And F-Measure) For The Three Statistical Approaches (The English Quranic Translation Text By Yusuf Ali)

Top % concept	Precision	Recall	F-measure
TFIDF	0.92	0.86	0.89
TIM	0.91	0.87	0.89
Linear Combination	0.93	0.87	0.90

It is quite prominent from the results in Table 2 and Table 3 that the Linear Combination of both TFIDF (term frequency inverse document frequency) and TIM prove to be good measures for automatic extraction of Quran concepts from both Quran translations. Results obtained also show that using combination method out performs that obtained using all individual statistical methods. These results indicate that the classifier combination methods are most suitable technique for extraction of hadith concept. The Linear Combination method achieves the highest result with 89.9% resulted in a precision of 86,67% which shows that the output is highly reasonable. Formulating an ontology lexicon with reasonable terms is an important perquisite for the determination of relations that model appropriately the domain in question.In addition, several experiments have been performed to empirically evaluate the linguistic patterns and the association rules for the relation extraction. The performance (precision, recall and F-measure) values metrics for the proposed methods on both Quranic translation datasets are shown in Table 4 and Table 5.

Table 4. The Relation Extraction Results Based On Precision, Recall (The English Quranic Translation Text By Yusuf Ali)

Task	Precision	Recall
Taxonomic relations	85.8	88.2
Non Taxonomic Relation	84.34	89.5

Table 5. The Relation Extraction Results Based On Precision, Recall (English Extended Quranic Translation Text, By Al-Hilali, And Khan)

Task	Precision	Recall
Taxonomic relations	83.3	85.7
Non Taxonomic Relation	81.34	84.9

These results indicate that the used method is suitable technique for extracting both non-taxonomic relation extraction and taxonomic relation extraction. The experiments in this study have generally shown highly promising results that clearly demonstrate the appropriateness of the application of linguistic patterns and association rules for ontology extraction from Islamic dataset.

6. CONCLUSION

This paper presented an Islamic ontology extraction system based on a hybrid method which combines lexico-syntactic patterns and association rules for English translation of the meaning of the Quran text. First, this paper used a new two layers of filtering method which combine linguistic and statistical methods for concept extraction. Second, this work used a new hybrid method based on lexico-syntactic patterns and association rules method for relation extraction. The results showed that using the two layers of filtering method prove to be adequate and efficient measures for automatic extraction of Quranic concepts with an overall F-measure of 85.3%.In addition, the results obtained indicate that the used methods are very suitable technique for extracting relation from with an overall F-measure of 87.3%and 88.3%respectively.. Future research may be targeted at developing a large Islamic corpus and designing a general framework for Islamic information extraction and analysis.

7. ACKNOWLEDGMENT

This research has been supported by the UKM university research Grant Gup-2015-003

REFERENCES

[1] Tom Gruber, "What Is An Ontology," Ed.

[2] Kais Dukes, "Ontology Of Quranic Concepts," Ed, 2010.

[3] Maha Al-Yahya, Hend Al-Khalifa, Alia Bahanshal, Iman Al-Odah, And Nawal Al-Helwah, "An Ontological Model For Representing Semantic Lexicons: An Application On Time Nouns In The Holy Quran," *Arabian Journal For Science And Engineering,* Vol. 35, P. 21, 2010Retrieved From.

[4] Noorhan Hassan Abbas, "Quran'search For A Concept'Tool And Website," Citeseer, 2009Retrieved From.

[5] Mohammed Al-Kabi, Ghassan Kanaan, Riyad Al-Shalabi, K Nahar, And B Bani-Ismail, "Statistical Classifier Of The Holy Quran Verses (Fatiha And Yaseen Chapters)," *Journal Of Applied Sciences,* Vol. 5, Pp. 580-583, 2005Retrieved From.

[6] Saidah Saad, Naomie Salim, And Hakim Zainal, "Pattern Extraction For Islamic Concept," In *Electrical Engineering And Informatics, 2009. ICEEI'09. International Conference On*, 2009, Pp. 333-337.Doi:Retrieved From.

[7] Saidah Saad And Naomie Salim, "Methodology Of Ontology Extraction For Islamic Knowledge Text," In *Postgraduate Annual Research Seminar*, 2008Retrieved From.

[8] F Harrag, A Alothaim, A Abanmy, F Alomaigan, And S Alsalehi, "Ontology Extraction Approach For Prophetic Narration (Hadith) Using Association Rules," *International Journal On Islamic Applications In Computer Science And Technology,* Vol. 1, Pp. 48-57, 2013Retrieved From.

[9] Aqil Azmi And Nawaf Bin Badia, "An Application For Creating An Ontology Of Hadiths Narration Tree Semantically And Graphically," 2010Retrieved From.

[10] Fouzi Harrag, "Text Mining Approach For Knowledge Extraction In Sahîh Al-Bukhari," *Computers In Human Behavior,* Vol. 30, Pp. 558-566, 2014Retrieved From.

[11] Diana Maynard, Yaoyong Li, And Wim Peters, "Nlp Techniques For Term Extraction And Ontology Population," In *Proceeding Of The 2008 Conference On Ontology Learning And Population: Bridging The Gap Between Text And Knowledge*, 2008, Pp. 107-127.Doi:Retrieved From.

[12] Ritesh Shah And Suresh Jain, "Ontology-Based Information Extraction: An Overview And A Study Of Different Approaches," *International Journal Of Computer Applications,* Vol. 87, 2014Retrieved From.

[13] Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J Bethard, And David Mcclosky, "The Stanford Corenlp Natural Language Processing Toolkit," 2014Retrieved From.

[14] Philipp Cimiano, *Ontology Learning From Text*: Springer, 2006.

[15] Marti A Hearst, "Automatic Acquisition Of Hyponyms From Large Text Corpora," In *Proceedings Of The 14th Conference On Computational Linguistics-Volume 2*, 1992, Pp. 539-545.Doi:Retrieved From.

[16] P. Cimiano, *Ontology Learning And Population From Text: Algorithms, Evaluation And Applications*: Springer, 2006.

[17] Saidah Saad, "ONTOLOGY LEARNING AND POPULATION TECHNIQUES FOR ENGLISH EXTENDED QURANIC TRANSLATION TEXT," Phd, Faculty Of Computing Universiti Teknologi Malaysia, 2013Retrieved From.

[18] Abdullah Yusuf Ali, "The Holy Quran: Text And Translation," *Kuala Lumpur: Islamic Book Trust,* 1994Retrieved From.

[19] Muhammad Al-Hilali And Mohammad Khan, *The Noble Qur'an: English Translation Of The Meanings And Commentary*: King Fahd Complex, 1997.