MATEC Web of Conferences **135**, 00075 (2017) DOI: 10.1051/matecconf/201713500075
ICME'17

Axiomatic Ontology Learning Approaches for English Translation of the Meaning of Quranic Texts

Saidah Saad[1], *Bahari* Idrus[2]

[1,2]Centre for Artificial Intelligence, Faculty of Technology and Information Science, Universiti Kebangsaan Malaysia

Abstract. Ontology learning (OL) is the computational task of generating a knowledge base in the form of an ontology, given an unstructured corpus in natural language (NL). While most works in the field of ontology learning have been primarily based on a statistical approach to extract lightweight OL, very few attempts have been made to extract axiomatic OL (called heavyweight OL) from NL text documents. Axiomatic OL supports more precise formal logic-based reasoning when compared to lightweight OL. Lexico-syntactic pattern matching and statisticsal one cannot lead to very accurate learning, mostly because of several linguistic nuances in the NL. Axiomatic OL is an alternative methodology that has not been explored much, where a deep linguistics analysis in computational linguistics is used to generate formal axioms and definitions instead of simply inducing a taxonomy. The ontology that is created not only stores the information about the application domain in explicit knowledge, but also can deduce the implicit knowledge from this ontology. This research will explore the English translation of the meaning of Quranic texts.

1 Introduction

Ontology learning (OL), which is the process of creating an ontology and populating it, has been the subject of intensive studies for the past decade. Researchers in this field have been motivated by the possibility of automatically building a knowledge base on top of text documents so as to support the extraction of reasoning-based knowledge. Ontology learning (OL) is the computational task of generating a formal knowledge base in the form of an ontology, given an unstructured corpus whose content is in a natural language (NL). This approach is based on the OL layer cake [1], [2]. Several works can be found in this area, most of which are limited to statistical and lexico-syntactic pattern matching-based techniques (lightweight OL) [3]. These techniques do not lead to very accurate learning, mostly because of several linguistic nuances in the NL, which become more challenging due to the complex structure of texts in literary documents [4], [5]. Formal OL is a hard task as it involves an

[1]Corresponding author: saidah@ukm.edu.my

accurate NL, and the understanding and conversion of the domain context into an equivalent formal presentation. This knowledge base formally represents both assertive facts as well as general truth statements expressed in some NL in a textual document. Formal knowledge supports semantic IR [6], [7], Q&A [8], [9] and reasoning services.

Ontologies are knowledge-based and play an important role in merging the semantic web and axioms. An axiom is one of the important components in an ontology that can create and describe the relationship between concepts, either within or across ontologies. Besides the taxonomy and structure of concepts, axioms represent more information about categories and their relationship to each other, as well as constraints on the properties and roles of each category. It is important to note here that there is a similar way of representing a concept or a fine line between an internal concept structure and axioms. First, a category can be represented using a formal frame, with roles and properties represented by the slots of a frame. Second, the same facts can also be expressed using axioms. Axioms represent knowledge in terms of objects, concepts, and roles. Concepts formally describe notions in an application domain, e.g. we could define the concept of being a father as "a man having a child". A knowledge base can be used to store the information we have about the application domain. Besides this explicit knowledge, we can also deduce implicit knowledge from a knowledge base. For example, from the fact that *Fatimah* belongs to the concept of *Woman* and the axiom $Man \equiv \neg Woman$, it can be deduced that *Fatimah* does not belong to the concept of *Man*.

Ontological axioms are assertions in a logical form that put some constraints into an ontology or are used to deduce new information. An example of ontological axioms is the assertion of the concept of subsumption or equivalence.

Axiom construction constitutes one of the layers in the OL layer cake [1] that has become a main building block for fixing the semantic interpretation of the concepts and relations of an ontology. T.Berners Lee said, "For the semantic web to function, computers must have access to structured collections of information and sets of inference rules that they can use to conduct automated reasoning". Currently, there are not that many real-world ontologies that make substantial use of axioms.

Understanding the literary documents that focus on the Quran is a grand challenge in the computational world for knowledge representation and reasoning. It is a challenging task due to the nature of Quranic texts, which have a scattered organization of knowledge, unique patterns of structure and style, and concepts that are interrelated across the entire document. Most of the current research has focused only on the concepts or relations extraction level. However, for these documents, the axiom level is necessary to explicitly define the full meaning of the concept of the Quran. The axiomatic method, which comprises principles for the development of theories, is aimed at the foundation, systematization, and formalization of a field of knowledge about a domain of the world. If knowledge of a domain is assembled in a systematic way, it gives rise to a set of categories that can be stipulated as being primitive or basic. Primitive categories are not defined by explicit definitions but by axioms that define their meaning implicitly.

2 Related Work

Knowledge representation, which is a description of logic-based ontologies, is a key role of semantic technologies. It can be used to describe the intended meaning from document sources and to exploit the powerful description of logic reasoning tools to facilitate the machine understandability of document sources, especially literary documents.

Ontology ideally consists of a TBox and an ABox. A TBox contains intentional knowledge in the form of a terminology, while the ABox contains extensional knowledge

that is specific to the individuals of the domain of discourse. A TBox is a conceptual definition of the upper layer components, while the ABox is more about the components of the lower layer. Meanwhile, the process of ontology creation can be categorized into seven main features (refer to Figure 1) forming increasingly complex subtasks of a "layered cake" [1], [2], which consists of:

i. Relevant terminology acquisition and extraction in the domain.
ii. The identification of synonymous terms or linguistic variants that can be across languages.
iii. The formation of a concept, which can be abstract or concrete, elementary or composite, real or fictitious.
iv. The concept of a hierarchical organization or concept hierarchy (taxonomy)
v. Learning about relations, properties, attributes and also the proper domain and range.
vi. Axiom schemata instantiation using a special logical axiom system such as disjoints, equivalences, symmetries, and others. It can also be the minimal or maximal cardinality of the relation.
vii. A general axiom, which depends strongly on the logical formalism being used in the background knowledge.

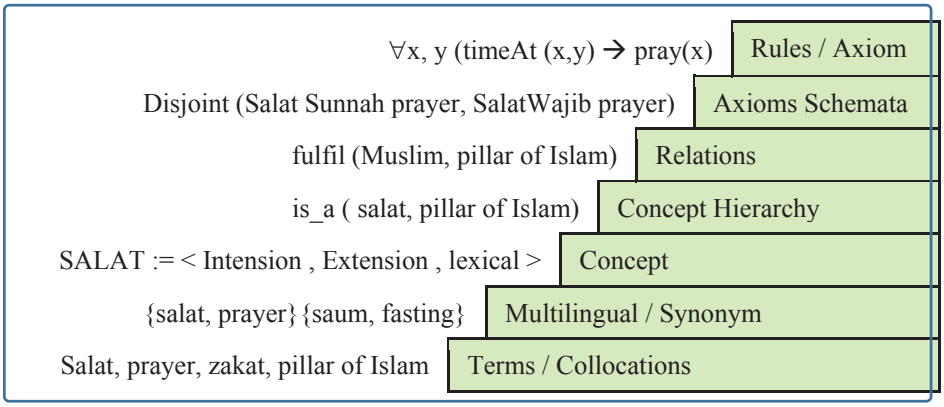

Fig. 1. Layered cake of ontology learning.

Among all those layers, ontology learning based on automatic rule acquisition is probably the least addressed. This rule-based systems are really important applications of semantic where the main reason to narrowing the semantic gap is to complement feature extraction which based on those layer cake with the formalized knowledge using description logic and rule based mechanism and implemented in ontology languages such as RDF, RDFS, Web Ontology Language (OWL) and rule language such as SWRL.At present, rule acquisition is still at a bottleneck,and is an obstacle to the wide propagation of rule-based systems.

State-of-the-art research on the automated learning of ontologies from texts is currently focusing on inexpressive ontologies. The acquisition of complex axioms involving logical connectives, role restrictions, and other expressive features of the OWL remains largely unexplored [10]. At present, the availability of computational techniques for the understanding of Quranic and Islamic knowledge is still limited [11], and most of the current research is just focusing on the level of the extraction of concepts/relations.

In order to provide a higher level of expressiveness to learned ontologies, several approaches have been proposed for extending ontology learning tools. The first approaches were manual, consisting of frameworks and tools such as Prot´eg´e, OntoEdit, NEON, and

KAON, which allowed axioms to be added by users or domain knowledge engineers. The more recent approaches include some kind of automation in order to add axioms under the evaluation and supervision of a knowledge expert. Approaches such as LExO [12], LEDA [13], and ReLExO [14] use a sequence of linguistic analysers. LExO starts by analysing the syntactic structure of an input sentence. The resulting dependency tree is transformed into a set of OWL axioms (concept inclusion, transitivity, role inclusion, role assertions, concept assertions and individual equalities) by means of manually-engineered transformation rules. ReLExO supports the acquisition and refinement of complex class descriptions in order to identify passages from the text that indicate the validity of certain knowledge. Given that the text can contain inconsistencies, LeDA allows the automatic generation of disjointed axioms based on machine learning classifications. The classifier, which determines the disjointedness of any given pair of classes, is trained, based on a gold standard baseline ofmanually-created disjointed axioms.

In [15], an automatic axiom-learning algorithm starts from a set of non-taxonomic relations. It uses the Web as a corpus, and linguistic techniques based on text patterns and a statistical analysis from the distribution of web information. But most of these are limited to axiom schemata provided by OWL, not on the general axiom, which is based on the description logic.

Baumann et al [10] presented an approach to the phenomenon of time. The basic relations are a temporal part of either among time regions or among time boundary regions, a temporal coincidence of time boundaries, and relations that link a time boundary (region) and a time region (including chronoids), declaring the former as a boundary of the latter, such as being the first or the last time boundary of a region. These are formalized by a set of axioms, specifying logical interrelationships between the categories and the relations. Their ontologies are axiomatized as formal theories in first-order logic and are analysed methodologically. They prove the consistency of both ontologies, and the completeness and decidability for one.

At present, the availability of computational techniques for the understanding of Quranic and Islamic knowledge is limited. Kais [16], for example, used Name Entity Extraction from the Quranic text to form the ontology for the Quran. Al-Yahya [17] presented the design and implementation of the ontological model, and the result of this application was the "time nouns" vocabulary of the Quran. Abbas [18] manually augmented a corpus of the Quran with an ontology or index of the key concepts taken from the 'Mushaf Al Tajweed' (Al Tajweed is another name for the Quran). It contains a comprehensive hierarchical index or ontology of nearly 1200 concepts in the Quran. Al-Kabi et al. [19] used an automatic classification technique to classify Quranic verses based on certain surah (chapters), according to the classification made by Islamic scholars. However, the basic ontology of Islamic knowledge in terms of the classification of the main topics in the Quran was already built decades before (based on the Quranic indices). The entire topic was classified manually by domain experts based on their understanding of the context of the Quran. Saidah [11, 21] describes the extraction of ontological component (concept, relation extraction) and the use of natural language pattern in extracting the knowledge of Quranic English translation texts (based on Salah subject) using natural language processing techniques. Research done by Saidah [11, 21] will be our main reference for continuing this research. This knowledge can be used as a general guide to the construction of a computer-automated ontology based on Islamic knowledge.

3 Methodology

The methodology for processing this kind of approach consisted of three phases as shown in Figure 2.

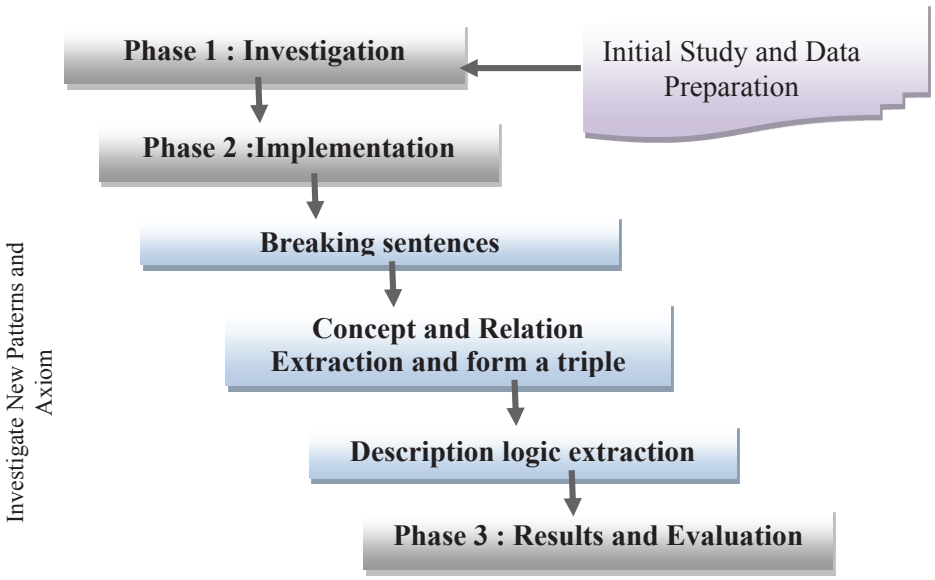

Fig. 2. Research method.

Phase 1 : Investigation

In this phase, the investigation was based on the current state of the:
i. Ontology learning and population
ii. Quranic understanding
iii. Analysis of Quranic content structure and populate Quranic gold standard

This research used the English Extended Quranic Translation Text by Dr. Muhammad Taqi-ud-Din Al-Hilali, and Dr. Muhammad Muhsin Khan [20]. It is quite a complete version that summarizes the information according to the tafsir of At-Tabari, Al-Qurtubi, and Ibn Kathir, with comments based on Sahih Al-Bukhari. Among the features available in this translation are that it defines all the unique words used in the Quran such as Muttaqun, Muzammil and Garden and so on, it gives extra information about certain verses, and it sometimes explains the mutashabihat verses and also the Asbabun Nuzul. This enriched the data and helped in the creation of the axiom rules.

Phase 2 : Implementation. This was based on the following steps (as shown in Figure 3):

Step 1. The document extraction to find the ontological component, as mentioned in Figure 1, began by splitting the document into sentences. The pre-processing stage of the document, such as tagging and identifying the syntax dependencies, was important

before proceeding to the next step. In this stage, a phrase that needed to be parsed had to undergo a temporary definition replacement process to simplify the on-going parsing process. For example: 'Those who believe' or 'You who believe' was replaced with Mukminun, and the words 'He', 'His', 'Our' and 'We' (where the first characters are in capital letters) were replaced by 'Allah'. This was obtained from a predetermined REPLACEMENT list. Next, the natural language pattern of the term had to be identified in order to proceed to the next stage.

In the extraction stage, all the identified patterns that matched the SPO (subject-predicate-object) were extracted. After identifying the ontological components such as concepts/instances, attribute/properties and relations using natural language patterns (SPO) – which based on stages in ontology layer cake from term extraction until relation extraction as mentioned in [11, 21],these were then used to analyse the description logic approach in the computational semantic domain such as lightweight ontology learning and formal ontology learning

Step 2. Triple extraction is a process to transform and convert complex and compound sentences to a set of simple sentences: "The believers, men and women, are Auliya", can be simplified to "The believers men are Auliya" and "The believers women are Auliya". This triple extraction, being a simplification process, helped in the easy translation of NL to DL at a later stage. This transformation can be done using a systematic parsing algorithm such as syntax dependency.

Step 3. All the lexical variations needed to be normalised into proper triple statements. A normalizer automatically identified all the lexical variations of the sentences, before translating them into their standard normal form. This process was necessary to identify the most generalized concept or relation for the mapping process, such as believer: NOUN + suffix (er) was shown as a person, man and woman was also a person; this extra knowledge could be coined from other structured knowledge such as a glossary or WordNet. However, WordNet is just for common terms and is not enough for Islamic terms [5],[11].

Step 4. DL Translator for TBox and ABox components. Those previously processed sentences had to have expressive equivalency in the DL language, as follows:

- $\mathcal{AL}$: Attributive Language supports the atomic concept definition, concept intersection, full value restriction, limited role restriction, and atomic concept negation.
- [$\mathcal{U}$]: Union supports the concept union
- [$\mathcal{E}$]: Existential supports the full role restriction
- [C]: Complement supports the concept negation
- [$\mathcal{H}$]: Role Hierarchy supports the inclusion axioms of roles
- [$\mathcal{O}$]: Nominal supports the concept creation of unrecognized Named Entity
- ($\mathcal{D}$): Data Type supports the range concepts to become the data type

Phase 3 : Every step in phase 2 was supported by a domain expert evaluation in order to get the 'truth' and the right concept according to the understanding of the Quranic domain and the proposed 'gold standard'.

For example, the concept of AlBirr has a different meaning or definition according to the Quran, where:

- Al-Birr is piety, righteousness… (AlBaqarah, 177)
- AlBirr is righteousness (AlMujadilah, 9)
- AlBirrmeans: a person who is the one who believes in Allah and the Last Day, the Angels, the Book, the Prophets, and gives his wealth, in spite of his love for it, to the kinsfolk, to the orphans, and to the Al-Masakin (the poor), and to the wayfarer, and to those who ask, and to set slaves free, performs as-Salat (Iqamat-as-Salat), and gives the Zakat, and who fulfils his covenant when he makes it (AlBaqarah, 177)

So, this information is stored in the form of rules (axioms) to facilitate the extraction of knowledge, and also the translation of the actual concept based on the context of the verse. Examples of knowledge representation in the form of rules (axioms) are the following (as mentioned in [21]):

AlBirr(?a) ⟹piety(?s)
AlBirr(?a) ⟹righteousness(?m)

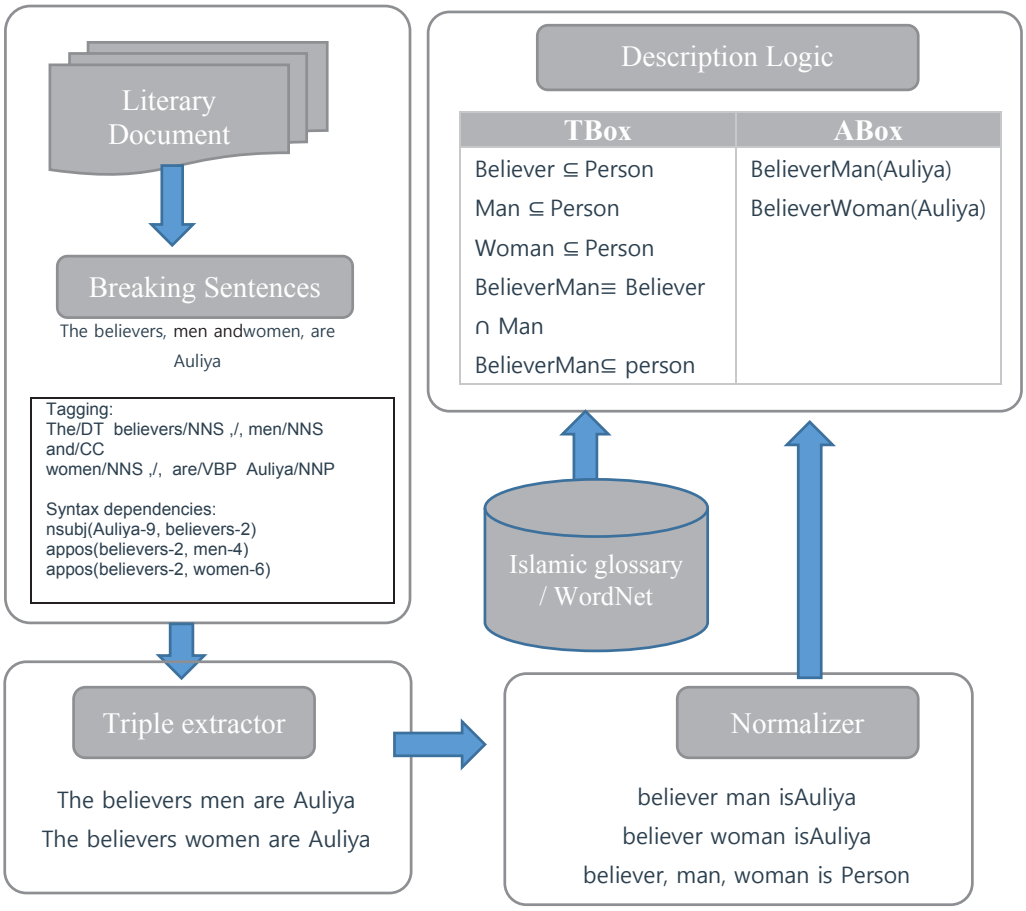

Fig. 3. Research activities.

A concept that has a definition with a combination of several other concepts and instances is:

AlBirr(?a) ⟹ person (?p) ∧ [believe(?p,Allah) ∧ believe(?p,LastDay) ∧ believe(?p,angle) ∧ believe(?p,Book) ∧ believe(?p,Prophet)] ∧ [givewealthto (?p,kinsfolk) ∨givewealthto (?p,orphan) ∨givewealthto(?p,masakin) ∨givewealthto(?p,wayfarer) ∨givewealthto(?p,thosewhoask) ∨givewealthto(?p,setslavefree)] ∧ perform(?p,salat) ∧ give(?p,zakat) ∧ fulfil(?p, covenant)

From other knowledge constructions, part of the concept of AlBirr actually has a formation of other concepts such as iman, which consists of belief in Allah, Last Day, Angel, Book and Prophet, and also the concept of 'sedekah':

- Iman(?c) ⟹ believe(?p,Allah) ∧ believe(?p,LastDay) ∧ believe(?p,angle) ∧ believe(?p,Book) ∧ believe(?p,Prophet)

- Sedekah(?s) ⟹givewealthto (?p,kinsfolk) ⋁givewealthto (?p,orphan) ⋁givewealthto(?p,masakin) ⋁givewealthto(?p,wayfarer) ⋁givewealthto(?p,thosewhoask) ⋁givewealthto(?p,setslavefree)

So, a new knowledge can be deduced from that: AlBirr is a person that has Iman, does sedekah, salat, zakat and fulfils the covenant, as follows:

AlBirr(?a) ⟹ person (?p) ⋀ has(?p, ?c) ⋀ give(?p, ?s) ⋀ perform(?p,salat) ⋀ give(?p,zakat) ⋀fulfill(?p, covenant)

In order to do that, the semantic web rule language (SWRL)can be used. The SWRL was recently proposed to improve the power of ontology. The SWRL rules provide procedural knowledge power to lift the limitations of the ontological inference, particularly in discovering semantic relations among instances. The SWRL is based on a combination of OWL and rule mark-up language. It extends the set of OWL axioms to include Horn-like rules, thereby enabling the SWRL rules to be combined with an OWL knowledge base. The most significant benefits of rules are their ability to chain characteristics and infer the existence of new facts. The SWRL rules are then utilized to infer new facts in accordance with semantic relations between properties inside instances and known facts in ontology [22].

In this research, a sound ontology learning technique that maps English language sentences into their equivalent axiomatic Description Logic (DL) expressions was proposed in order to automatically generate a consistent pair of T-box and A-box, thereby forming both a regular (definitional form) and generalized (axiomatic form) DL ontology.

4 Discussion

In this paper, we described an approach for the automatic development of expressive ontologies from definitions provided in a text of English translation to a meaning of Quranic content, and to automatically generate an axiomatic ontology based on this content. This research is still in progress to identify the automatic construction and pattern of the axiomatic element. The axiomatic method is comprised of principles for the systematization and formalization of a field of knowledge about a domain of this content, and if the knowledge of this domain is assembled in a systematic way, a set of categories is stipulated as primitive or basic. Primitive categories are not defined by explicit definitions but by axioms that define their meaning implicitly. With this approach, all the important concepts defined in the Quran can also be understood and can be related to each other, where one concept not only can be defined by just one other concept but is comprised of several concepts.

Acknowledgement

This research is supported by the UKM University Research Grant GUP-2015-003.

References

1. P. Buitelaar, P. Cimiano, B. Magnini, Ontology learning from text: An overview. ontology learning from text: Methods, evaluation and applications. Frontiers in Artificial Intelligence and Applications Series 123 (2005)
2. P. Cimiano, Ontology Learning and Population from Text. Evaluation, doi:10.1007/978-0-387-39252-3 (2006)

3. S. Dasgupta, A. Padia, K. Shah, R. KaPatel, P. Majumder, DLOLIS-A: Description Logic based Text Ontology Learning. CoRR abs/1303.5929 (2013)
4. S. Saad, N. Salim, H. Zainal, Islamic knowledge ontology creation. Internet Technology and Secured Transactions, 2009. ICITST 2009. International Conference for, hlm.1–6. Inproceedings (2009)
5. A. B. Sharaf, E. Atwell, A Corpus-based computational model for knowledge representation of the Qur'an, Proceedings of the 5th Corpus Linguistics Conference, in proceedings (2009)
6. A. M. Pejtersen, Semantic information retrieval, Communications of the ACM 41 (4) 90–92 (1998)
7. S. Remi, S. C. Varghese, Domain Ontology Driven Fuzzy Semantic Information Retrieval, Procedia Computer Science 46 (2015) 676 – 681 (2015)
8. D. Moldovan, C. Clark, M. Bowden, Lymba's power answer 4 in trec 2007, in: TREC-2007, Association for Computational Linguistics (2007)
9. M. Sheker, S. Saad, R. Abood, M. Shakir, Domain-Specific Ontology-Based Approach For Arabic Question Answering, *Journal of Theoretical And Applied Information Technology*, E-ISSN 1817-3195 / ISSN 1992-8645). Vol **83** (2016)
10. L. Baumanna and Herreb, Axiomatic Theories of the Ontology of Time in GFO. Applied Ontology, IOS Press, **8** (2013).
11. S. Saad, Ontology Learning And Population Techniques For English Extended Quranic Translation Text. PhD Thesis. UTM Skudai. Johor (2014)
12. J. Volker, P. Hitzler, P. Cimiano, Acquisition of owl dl axioms from lexical resources, In Enrico Franconi, Michael Kifer, and Wolfgang May, Editors, The Semantic Web: Research and Applications, volume 4519 of Lecture Notes in Computer Science, pages 670–685. Springer Berlin / Heidelberg, 10.1007/978-3-540-72667-8 47 (2007)
13. J. Volker, D. Vrandeˇci´c, Y. Sure, A. Hotho, Learning disjointness, In Proceedings of the 4th European conference on The Semantic Web: Research and Applications, ESWC '07, pages 175–189, Berlin, Heidelberg, 2007. Springer-Verlag (2007)
14. J. Volker, S. Rudolph, Lexico-logical acquisition of owl - dl axioms. In Raoul Medina and Sergei Obiedkov, editors, Formal Concept Analysis, volume 4933 of Lecture Notes in Computer Science, pages 62–77. Springer Berlin/Heidelberg, 2008. 10.1007/978-3-540-78137-0 5 (2008)
15. L. D. V. Terrientes, A. Moreno, and D. S´anchez, Discovery of relation axioms from the web. In Proceedings of the 4th international conference on Knowledge science, engineering and management, KSEM'10, pages 222–233, Berlin, Heidelberg, Springer-Verlag (2010)
16. K. Dukes, Ontology of Quranic Concepts. Retrieved January 2012, from http://corpus.quran. com/ontology.jsp (2010)
17. M. Al-Yahya, H. Al-Khalifa, A. Bahanshal, I. Al-Odah I., N. Al-Helwah, An Ontological Model For Representing Semantic Lexicons : An Application On Time Nouns In The Quran. The Arabian Journal for Science and Engineering. Volume 35, Number 2C, 21–35 (2010)
18. N. Abbas, Quran 'Search for a Concept' Tool and Website. Master Thesis. School of Computing, University of Leeds (2009)
19. M. N. Al-Kabi, G. Kanaan, R. Al-Shalabi, Statistical Classifier of the Holy Quran Verses (Fatiha and Yaseen Chapters). Journal of Applied Sciences **5**(3): 580-583 (2005)

20. T. Al-Hilali, M. K. Khan, Interpretation of the meaning of the Qur'an in the English Language. King Fahd Quran Printing Complex, Madinah (1998)
21. S. Saad, N. Salim, H. Zainal, Ontology Learning and Population from Quranic Translation Texts. Taibah University International Conference on Advances in Information Technology for the Holy Quran and Its Sciences, pp. 407–415 (2013)
22. Chi & Chen, Ontology and semantic rules in document dispatching, *The Electronic Library*, **27**(4), 694-707 (2009)