VOL. 11, NO. 13, JULY 2016 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

UNSUPERVISED CONCEPT HIERARCHY INDUCTION BASED ON ISLAMIC GLOSSARY

Ammar Abdulateef Ali and Saidah Saad
Center for Artificial Intelligence Technology (CAIT), Faculty of Information Science and Technology, University Kebangsaan Malaysia Bangi, Selangor Darul Ehsan, Malaysia
E-Mail: saidah@ftsm.ukm.my

ABSTRACT

A machine-readable dictionary (MRD) is an electronic dictionary that enables query processing. One of the common processing tasks that has been widely applied is Concept Hierarchy Induction which aims at identifying concepts with its corresponding taxonomies such as named entities, synonyms and hyponyms. The Islamic domain contains a variety of concepts that are associated with numerous taxonomies. The existing concept hierarchy approaches for Islamic domain are using limited linguistic patterns. This study aims to propose an unsupervised concept hierarchy induction for the Islamic domain by extending the patterns and rules. In fact, Term Frequency-Inverse Document Frequency (TF-IDF) was carried out in order to identify the most frequently used concepts. Furthermore, two syntactical features were used including POS tagging and chunk parser in order to identify the tagging for each word (e.g. verb, noun, adjective, etc.) and extracting Noun Phrases (NP). Hence, the proposed extension patterns aim at utilize lexico-syntactic patterns to induce the concept hierarchy. The evaluation was performed using precision method by identifying the number of correctly extracted concepts and relation between them. Moreover, an expert review evaluation was performed by an expert in the Islamic domain. The experimental results showed that the proposed method achieved 82% precision. That demonstrates the usefulness of extending patterns for the Islamic domain.

Keywords: concept hierarchy, terminology extraction, ontology, lexico-syntactic patterns.

1. INTRODUCTION

Domain specific terms (single and multi-word) refer to compound words that have special conceptual meanings in a particular domain. Automatic terminology recognition and extraction is the process of recognizing and extracting domain specific terms.

Automatic terminology extraction is an important task in many natural language processing and knowledge engineering applications such as indexing and information retrieval (IR)(Chowdhury 1999), machine translation (Och & Ney 2000), information extraction (Simoes et al. 2009), domain specific lexicon construction (Hull 2001), and topic extraction (Zhang et al. 2010). In fact, it contributes to all domain-oriented natural language processing domain.

Most studies on automatic terminology extraction have introduced several approaches and techniques to extract and recognize domain specific terms. These approaches can be classified into three categories; linguistic approaches, statistical approaches and combined approaches. The linguistic approaches to domain specific term recognition basically try to identify terms by capturing their syntactic properties as the terms usually have common syntactic structures. The statistical methods use statistical analysis of word usage and also define association criteria to measure correlations between words. A few representative methods are quite commonly used, including Term frequency (TF), Inverse Document frequency (IDF), Mutual Information, Log-Likelihood Ratio, and Dice Factor. In facts, hybrid extraction methods, which make use of a statistical approach that includes some linguistic information, outperform other methods in Automatic terminology extraction and represent the general trend.

In the current study, several automatic terminology recognition and extraction algorithms will be investigated for extracting domain corpus (Islamic dictionary/glossary) single and multi-word terms in the Islamic domain. These algorithms are hybrid approaches based on statistical and linguistic analysis.

In fact, several approaches have been proposed for concept hierarchy induction using several domains such as biomedical, web and social networks (Amsler 1980; Markowitz *et al.* 1986; Montemagni and Vanderwende 1992; Dolan *et al.* 1993). These efforts have treated various kinds of data including unstructured text, machine readable dictionaries and web content. However, Islamic domain has brought many researchers' attentions regarding multiple reasons. Islamic domain contains a huge amount of documents and manuscripts that are not being categorized. In addition, Islamic terms have their own meaning and hard to be explained and could yield multiple meaning. Some machine readable dictionaries have been presented for such purpose, where each Islamic term is being defined properly. Nevertheless, there is an essential demand for providing an automatic concept hierarchy induction from such dictionaries. To do so, the rules used for concept hierarchy induction have to be extended. In particular, Lexico-Syntactic patterns which are specific linguistic rule have been proposed for concept hierarchy induction. These patterns have to be extended in order to include rules that fit the Islamic domain. Therefore, this study aims to propose an extension of Lexico-Syntactic patterns for Islamic concept hierarchy induction using a machine readable dictionary.

2. RELATED WORK

One of the earliest research efforts that have addressed the building of ontology using machine readable dictionary is the one that proposed by Nichols et al. (2005). The authors have extracted definitions from a Japanese readable dictionary and constructed semantic representations among them based on Robust Minimal Recursion Semantic (RMRS) approach.

Saad *et al.* (2008) have proposed a key-phrase extraction approach for Islamic knowledge ontology based on a hybrid method of lexico-syntactic patterns and statistical measure. In fact, the text has been annotated using parser in order to extract noun phrases. Then, using TF-IDF as a statistical measure, the extracted candidates have been ranked based on frequency.

In addition, Saad *et al.* (2011) have proposed a Solat-based ontology prototype for Islamic domain. The authors have used the specific terms related to Solat in order to build the main classes and the relations among them. In fact, the hierarchical taxonomy has been built using a combination of top-down and bottom-up approaches.

Mukhtar *et al*. (2012) have proposed a semi-automatic approach for Quranic terminologies creation. Such approach consists of three main phases. First, the significant terms from Quran have been extracted using NC-value which is a statistical measure that aims to identify head-words or multi-word terms such as 'Judgment day'. Second, a filtering task is being performed by expert in order to filter the candidate terms based on correctness. Third, a lexico-syntactic patterns have been used in order to initiate the hierarchy among the candidate terms.

Saad *et al.* (2013) have proposed a hybrid method of rules and NLP techniques for extracting Quranic knowledge. First, the authors have used a filtering approach that aims to unify the same terms (e.g. replacing the pronouns 'we', 'us', 'he', 'lord', and 'my' into Allah). Second, the NLP techniques including POS tagging and parsing have been used in order to determine the words' tags as well as noun phrases and verb phrases. This task is crucial in terms of applying the rules (lexico-syntactic patterns). Then, the rules are being applied in order to build the taxonomy relations.

3. MATERIALS AND METHODS

The research design of this study consists of five main phases as shown in Figure 1 including the dataset phase, transformation phase, lexical and syntactic Analysisphase, Lexico-Syntactic patterns phase and evaluation. These phases can be explained as follows:

a) **Dataset phase:** the dataset phase is associated with the data that will undergo processing. Such dataset is an Islamic dictionary-glossary which has been collected from the International Islamic University, Malaysia. This glossary contains large amount of Islamic concepts with their corresponding definitions (DEED 2015).

b) **Transformation:** this phase is associated with the tasks that aim at turning the data into an appropriate format which can facilitate the processing. In this manner, sentence splitting task will be performed in order to separate the sentences (i.e. definitions) from the concepts. In addition, tokenization task will be carried out in order to turn the sentences into series of tokens.

c) **Lexical and syntactic analysis:** this phase is associated with the triggers that have the role of building the Lexico-Syntactic patterns. To do so, two sub-tasks should be applied including lexical Analysis and syntactic Analysis. Lexical are the tools that address the words lexical in which the morphology of the word will be analyzed. Lexical Analysis applied using Term Frequency Inverse Document Frequency (TF-IDF). The TF-IDF aims to generate the most frequent terms. On the other hand, syntactical Analysis aim at identifying the grammatical tag of the words where Part-Of-Speech (POS) tagging will be used to provide a tag for each word (e.g. verb, noun, adjective, etc.) and Parser will be used to extract the Noun Phrases (NP).

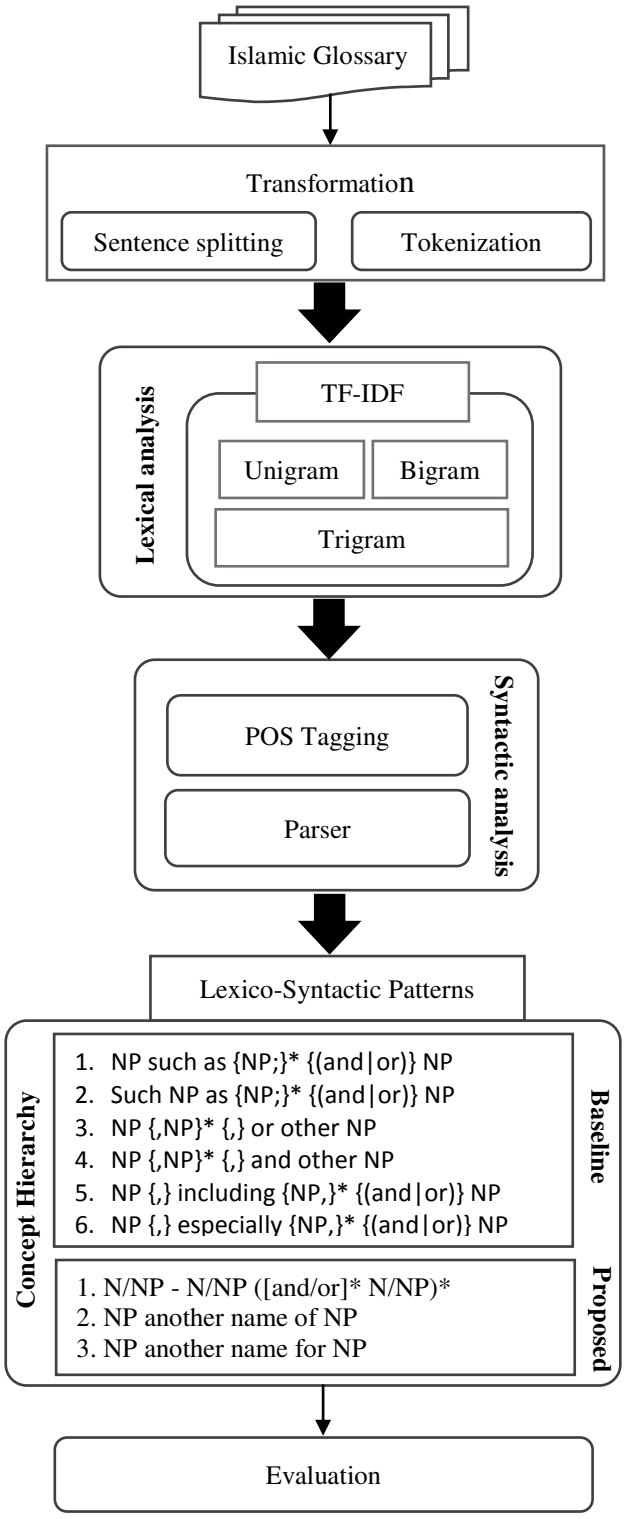

Figure-1. Framework of study.

d) **Lexico-Syntactic Patterns**: this phase aims to utilize the lexical and syntactic Analysis (from the previous phase) to generate specific patterns that would facilitate the process of extracting definitions of certain concepts. Note that, six patterns have been used from the literature, whereas this study propose a pattern extension using two new patterns.

e) **Evaluation**: this phase aims to evaluate the proposed method based on the accuracy where the number of correct extracted definition will be divided on the total number of definitions. Due to the domain that addressed in this study requires expert to assess it, experts in Islamic domain have been consulted in order to assess the extracted definitions based on the language and Islamic criteria.

Dataset (Islamic glossary)

As mentioned earlier, the dataset used in this study is an Islamic dictionary-glossary that contains Islamic concepts with their corresponding definitions. In fact, the data has been collected from the Islamic dictionary of International Islamic University, Malaysia (DEED 2015). Table 1 shows a sample of such data.

Table-1. Sample of the islamic glossary.

Islamic concept	Definition
Aali Imran	The family of Imran. Imran was the father of Mariam (Mary), the mother of the Prophet Isa (Jesus), peace be on them. See Mursaleen. Surah 3 of the Holy Quran.
Abasa	He frowned. The blind man that is referred to in this surah is Abdullah ibn Umm Maktoum. Surah 80 of the Holy Quran.
Afareet	Evil jinns who are large, powerful and very crafty. SingularIfreet. See Holy Quran , An-Naml (27)39.

As shown in Table-1, each Islamic concept is stored in the dictionary with its corresponding definition. The definition consists of multiple sentences for instance, the first concept in the table 'Aali Imran' is refer to 'The family of Imran', 'Imran was the father of Mariam', 'Mariam is the mother of prophet Isa (Jesus)'. Basically, the definition will be separated in the next phase 'transformation' which can be illustrated in the following section.

3.1 Transformation

Transformation phase aims to convert the data into more suitable form which enable processing, extraction and tagging. This phase is a crucial and plays an essential role in terms of providing appropriate tokens. In order to apply transformation, two tasks should be carried out including sentence splitting and tokenization tasks.

In the first place, isolating the concept form its own definition is an important task where the definition will be held for further processing. Since the definition contains multiple sentences thus, it is necessary to accommodate a splitting task for these sentence where each sentence will be separated.

Tokenization aims to turn the sentences resulted from the previous task (sentence splitting) into a series of tokens where each word will be considered as a token. This can facilitate the process of analysing each lexical by identifying the term frequency.

3.2 Lexical and syntactic analysis

This phase aims to exploit lexical and syntactical Analysis in order to provide more comprehensive analysis for the tokenized words. Basically, these tools are the core of generating the Lexico-Syntactic patterns where such patterns are totally depending on such tools. For example, in order to identify the noun words, it is necessary to apply POS tagging which has the ability to provide grammatical tag such as noun and verb. Therefore, this phase has a significant impact on identifying useful patterns.

Whereas syntactic Analysis are associated with grammatical features where exploiting specific tag could lead to identity definitions for specific concepts. For example, exploiting nouns has the ability to provide definitions since most of the definitions are containing nouns. For this manner, POS tagging and Parser have been used in this study as syntactic Analysis. These Analysis will be illustrated in the following subsections.

- **Term frequency - inverse document frequency (TF-IDF)**

Such measure combines both of TF and IDF in order to examine the frequency of a particular term in a specific document over the frequency of the same term in other documents. It can be calculated as follows:

$W_t = TF(t,d).IDF_t$

where TF (t, d) is the term frequency of term t in the document d and IDF is the inverse document frequency of term t. This method is an efficient and has demonstrated a valuable usage by many researchers (Xu & Chen 2010).

According to Saad *et al.* (2008) the most straight-forward approach for extracting concepts relies on the frequent words in which the most significant terms lie on the text are being occurred frequently. Frequency of a term facilitates the process of identifying relations among concepts for instance, the frequent occurrence of the string 'Allah bless him' indicates that the concept is related to a great person.

- **POS Tagging**

This method aims to provide a tag for every word including noun, adjective and verbs. POS tagging is a useful tool that can disambiguate many words with different meaning (Van Gael et al. 2009). For example, the word 'book' may refer to a 'reservation' or 'folder', the key distinguish lies on determining the tag of words where if 'book' is a verb then it means 'reservation', otherwise it would refer to 'folder'. Hence, the use of POS tagging is essential in terms of identifying specific meaning.

- **Parsing**

Parser is a syntactic tool that aims to identify the noun phrases as well as verb phrases (Van Noord 2006). Basically, it handles the sentence as a series of tagged words where a combination of two nouns forms a noun phrase. Whereas, a combination of two verbs forms a verb phrase. In fact, parser has been used in this study in order to extract the Noun Phrases (NP) which probably yield definitions belong to specific concepts.

3.3 Concept hierarchy using lexico-syntactic patterns

This phase represents the contribution of this study where an extension of rules (i.e. patterns) will be carried out in order to extract concepts with their corresponding definitions. According to (Klaussner and Zhekova 2011), there are multiple Lexico-Syntactic patterns that could be used for concept hierarchy. These patterns are illustrated in Table-2.

Table-2. Lexico-syntactic patterns proposed by (Klaussner and Zhekova 2011).

Lexico-Syntactic Patterns	Example	Taxonomy Relation
NP such as {NP;}* {(and|or)} NP	Prophets such as Abraham, Moses and Joseph are.	is-a(Abraham, Prophet) is-a(Moses, Prophet) is-a(Joseph, Prophet)
Such NP as {NP;}* {(and|or)} NP	Such prayers as Dhuhur, Aser or Fajr are.	is-a(Dhuhur, prayer) is-a(Aser, prayer) is-a(Fajr, prayer)
NP {,NP}* {,} or other NP	Haj, Solat, Zakat or other worships are.	is-a(Haj, worship) is-a(Solat, worship) is-a(Zakat, worship)
NP {,NP}* {,} and other NP	Adultery, killing, rubbery and other Permitted are.	is-a(Adultery, Permitted) is-a(killing, Permitted) is-a(rubbery, Permitted)

As shown in Table-2, multiple Lexico-Syntactic patterns have been proposed by the literature. These rules or patterns are useful and shown good accuracy. However, it is limited to specific cases mostly like 'such as', 'or / and'. This increases the restriction of acquiring more definitions. Therefore, this study aims to propose an extension for these rules. Table-3 shows the proposed extension of rules.

Table-3. Proposed extension patterns.

Proposed extension patterns	Example	Taxonomy relation
N/NP - N/NP ([and/or]* N/NP)*	Abu Lahab - fiercest enemy of Islam and prophet Muhammad's uncle	is-a(Abu Lahab, fiercest enemy of Islam) is-a(Abu Lahab, prophet Muhammad's uncle)
N/NP another name of N/NP	Bakkah is another name of Makkah	is-a(Bakkah , Makkah)
N/NP another name for N/NP	Yathreb is another name for Madeena	is-a(Yathreb, Madeena)

As shown in Table-3, the proposed patterns are concentrating on special cases in the Islamic domain where tremendous words could be expressed via other names. In addition, since the Quran and Hadith has been written in Arabic thus, many regular words are located in the dictionary such as 'Khalifa' which has another name of 'prince of believers'. This customization in the Islamic domain makes the proposed rules is needed.

3.4 Evaluation

In order to evaluate the proposed method, a comparison will be established to compare the results of existing patterns against the proposed extension. However, to evaluate these patterns, the number of correct extracted definitions will be divided into the total number of definitions. To do so, precision can be used to identify such computational as shown in the following equation:

$$Precision = \frac{Correct\ extracted\ definitions}{total\ number\ of\ definitions}$$

Since, these extracted definitions are considered as domain-specific expressions. Hence, experts in the Islamic domain have been called to evaluate the extracted definitions. These experts are annotating the wrong definitions by 0, and the correct definitions by 1. Therefore, the precision can be formulated as:

$$Precision = \frac{\#\ of\ definitions\ annotated\ with\ 1}{\#\ total\ number\ of\ definitions}$$

5. RESULT

The final results obtained by both of baseline Lexico-Syntactic patterns against the proposed extension patterns. As mentioned earlier, the results are based on precision. Table-4 shows precision values of both the baseline and the proposed patterns.

Table-4. Concept hierarchy results**.**

Patterns	Total induced concepts	Correct induced concepts	Precision
Baseline Lexico-Syntactic Patterns	62	46	0.741
Proposed Lexico-Syntactic Patterns	167	138	0.820

As shown in Table-4, the results of the baseline patterns have been represented as a 46 correct induced concepts out of 62 total numbers of concepts which leads to 74% of precision. Whereas, the results of the proposed patterns have been represented as a 138 correct induced concepts out of 167 total numbers of concepts which leads to 82% of precision. Apparently, the proposed extension patterns have a superior results compared to the baseline patterns. This can demonstrate the significant impact of the proposed patterns in the field of Islamic conception.

4. CONCLUSIONS

Concept Hierarchy Induction is the process of classifying taxonomies and identifying relationships among the terms. This can be performed by analysing both the lexical and syntactic features of the concept in a form of a rule-based pattern which called Lexico-syntactic patterns. On other hand, Islamic domain is one of the common fields that have been addressed in terms of concept hierarchy induction where the two main sources of Islam including Quran and hadith have been tackled with their concepts. This study has proposed an extension of lexico-syntactic patterns for concept hierarchy from Islamic glossary. Such extension shown superior results compared to the baseline. Generally, there are further lexico-syntactic patterns could be utilized for the Islamic domain in the future work.

ACKNOWLEDGMENT

This work has been supported by the university grant GUP-2015-003.

REFERENCES

[1] Amsler R. A. 1980. The structure of the Merriam-Webster pocket dictionary.

[2] Chowdhury, G. 1999. The Internet and information retrieval research: A brief review. Journal of Documentation. 55(2): 209-225.

[3] DEED D. E. E. D. 2015. Islamic Dictionary-Glossary. http://www.iium.edu.my/deed/quran/glossary.html.

[4] Dolan, W., L. Vanderwende and S. D. Richardson 1993. Automatically deriving structured knowledge bases from on-line dictionaries. Proceedings of the First Conference of the Pacific Association for Computational Linguistics, hlm. pp. 5-14.

[5] Hull D. A. 2001. Software tools to support the construction of bilingual terminology lexicons. D. Bourigault, C. Jacquemin and M.-C. L'Homme Recent Advances in Computational Terminology. Amsterdam/Philadelphia, John Benjamins: pp. 225-244.

[6] Klaussner C. and D. Zhekova. 2011. Lexico-Syntactic Patterns for Automatic Ontology Building. RANLP Student Research Workshop, hlm. pp. 109-114.

[7] Markowitz, J., T. Ahlswede and M. Evens 1986. Semantically significant patterns in dictionary definitions. Proceedings of the 24th annual meeting on Association for Computational Linguistics, hlm. pp. 112-119.

[8] Montemagni S. and L. Vanderwende. 1992. Structural patterns vs. string patterns for extracting semantic information from dictionaries. Proceedings of the 14th conference on Computational linguistics. Vol. 2, hlm. pp. 546-552.

[9] Mukhtar, T., H. Afzal and A. Majeed 2012. Vocabulary of Quranic Concepts: A semi-automatically created terminology of Holy Quran. Multitopic Conference (INMIC), 2012 15th International, hlm. pp. 43-46.

[10] Nichols E., F. Bond and D. Flickinger 2005. Robust ontology acquisition from machine-readable dictionaries. IJCAI, hlm. pp. 1111-1116.

[11] Och F. J. and H. Ney 2000. Statistical machine translation. EAMT Workshop, hlm. pp. 39-46.

[12] Saad S., S. Noah, N. Salim and H. Zainal. 2013. Rules and Natural Language Pattern in Extracting Quranic Knowledge. Advances in Information Technology for the Holy Quran and Its Sciences (32519), 2013 Taibah University International Conference on, hlm. pp. 381-386.

[13] Saad S., N. Salim and N. Omar. 2008. Keyphrase extraction for Islamic Knowledge ontology. Information Technology, 2008. ITSim 2008. International Symposium on, hlm. pp. 1-6.

[14] Saad S., N. Salim, H. Zainal and Z. Muda. 2011. A process for building domain ontology: An experience in developing Solat ontology. Electrical Engineering and Informatics (ICEEI), 2011 International Conference on, hlm. pp. 1-5.

[15] Simoes G., H. Galhardas and L. Coheur 2009. Information Extraction tasks: a survey. Proc. of INForum, hlm.

[16] Van Gael, J., A. Vlachos and Z. Ghahramani. 2009. The infinite HMM for unsupervised PoS tagging. Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2, hlm. pp. 678-687.

[17] Van Noord G. 2006. At last parsing is now operational. TALN06. Verbum Ex Machina. Actes de la 13e conference sur le traitement automatique des langues naturelles, hlm. pp. 20-42.

[18] Zhang D., Q. Mei and C. Zhai 2010. Cross-lingual latent topic extraction. Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, hlm. pp. 1128-1137.