A Pattern for Concept Identification from English Translated Quran

Rohana Ismail[1,*], *Nurazzah* Abd Rahman[2], *Zainab* Abu Bakar[3]

[1]Faculty of Informatics and Computing, Universiti Sultan Zainal Abidin, Kuala Terengganu, Terengganu.
[2]Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Shah Alam, Selangor.
[3]Faculty of Computer and Information Technology, Al-Madinah International University, Shah Alam Selangor.

Abstract. Ontology development is time consuming and tedious task. The task can be minimized by automatic or semi-automatic ontology development. This minimizing task is a field of ontology learning. Ontology learning will be able to extract ontological elements to form ontology. Concept identifying is one of the important activities in ontology learning. Various methods can be used to find concepts. Thus, this experiment used the n-grams and JAPE pattern in identifying concepts. As the term Allah occurs very frequent in English translated Quran, this experiment considers the word surrounding the term Allah to find other related concepts. It is important because the occurrences produced the term Allah as a concept in ontology but at the same time ignore other related terms to term Allah. The strength connection between words surrounding term Allah has been analysed. Results show the significant terms related to term Allah can be extracted. Later, the term can be used as concepts in ontology.

1 Introduction

The availability of Semantic Web technologies makes the process of understanding Quran is much easier. It is because the Semantic Web technologies can help representing content from Quran as ontology. Ontology is a formal, explicit specification of shared conceptualization [1]. It can be used to form a knowledge based in order to increase the chances of knowledge sharing. Ontology is widely explored in Quran study [2][3][4][5][6]. For example, the ontology from Arabic Quran focusing on time purposely develop to represent semantic meaning of an Arabic word time [2]. It used semantic field theory and componential analysis. Another example is the form of Quran Ontology from Arabic Language Computing Research at the University of Leeds [5]. The comprehensive ontology is focusing on concepts, relations and verse that exist in Quran. However the ontology has insufficient concepts. For example, the Hajj is a concept and just be a part of a calendar event. It is actually many concepts exist related to Hajj.

[1] Corresponding author: rohana@unisza.edu.my

Currently, ontology from Quran is only cover for a few domains such as women, solat, time and animal [7][8][9]. Most of the ontology development is done manually. Manual extraction is a time consuming task, hence there is a need to used automatic or semi-automatic approaches to develop ontology [10][11]. This approach is called ontology learning. The ontology learning approach is able to produce concepts, relations and sometimes axioms from text [12][11]. Few researches attempts to develop ontology from Quran using ontology learning [7][9][13]. For instance, a research used semi-ontology building from Arabic corpus [13]. The research used the WordNet ontology as a model in their ontology development. Another example is using ontology learning for solat ontology development [9]. The ontology used English translated Quran as their corpus. It used statistical and linguistic methods in extracting concepts and relations.

Identifying concept is important in ontology learning. Concept can be a single term or multi term [14][15]. Linguistic and statistical methods are popular methods in ontology learning [12]. Such statistical method is Term Frequency (TF). The TF is simple but shows significant to find concepts [9]. But, sometimes it will ignore the multi term regarding the concepts. For example, the TF produced the term *Allah* as a concept but it will ignore multi terms regarding the term *Allah*. Hence, term surrounding *Allah* needs to be analyzed for getting more relevant terms to be concepts.

This paper proposed techniques to extract concepts related to term *Allah*. The concepts can be identified based on terms collocation with *Allah* and by using the JAPE pattern. The Log Likelihood measurement has been used in N-grams to analyzed terms collocations. Meanwhile, the JAPE pattern has been design to extract the concepts. Result shows the performance of the pattern extraction and the collocation term extraction.

2 Related Work

Extractions of ontological elements from text and building ontology from them is a field of ontology learning [12][16]. Ontology provides conceptual sharing of formal represented knowledge [1]. Ontology learning can be categorized into ontology learning layer cake that consists of layers of activities [17]. These layers consist of identifying terms, synonym, concepts, concept hierarchies, relations and rules. Various methods from established field such as Natural Language Processing and Information Retrieval could be used in ontology learning.

Concepts have no clear definition, but it should provide the intentional part of a domain for leading the way the structure of corresponding ontology[17]. In concepts identification from Quran study, it has varies either single terms or multi terms such as *Allah Bounties, mountain*, *Hajj*, *raging fire* [14]. Single terms can be identified using various statistical methods such as term frequency, tfidf, glossex, Average tf and C-Value [9]. Meanwhile, the multi terms can be identified using N-grams [18] or using linguistic method [19]. Linguistic method depends on Part Of Speech tagger [12]. A hybrid of both methods are also used to identify concept [9][18].

The use of linguistic technique in an ontology learning system such as CRCTOL, mainly depends on pattern and POS rule to detect multi terms from text to be concepts [15]. The performance of the extraction has been compared to other ontology learning system which is Text2Onto. The multi terms also can be identified by identifying terms collocation. The collocation will determine a sequence of terms that frequently appear more together than would be expected by chance collocation [20]. For example the noun *Fire* can be a collocation term when it occurs frequently with the term *raging*. The combinations of terms *raging Fire* that occurs frequently can be a collocation. However, not all the

combination of terms can be a collocation. It can be analyzed using the statistical methods such as Log Likelihood.

Collocation can be identified using linguistic patterns such as verb-noun, noun-noun, adjective-noun and also using the N-grams method. For example, ontological elements from Curriculum Vitae (CV) has been extracted using patterns such as Noun-Noun, Adjective-Noun, Noun-Adjective, Noun-Preposition-Noun collocations [21]. Another example is the used of N-grams method. The method has been used in finding terms collocation in Arabic Quran corpus [22]. Based on the sequence of Part of Speech, the experiment has extracted collocation of N-gram words ranging from 2 to 6 grams Tagger. One more example in collocation extraction is the use of noun-adjective collocation. The collocation has been extracted from Arabic corpus [18]. The experiment used JAPE pattern in GATE application for extracting terms. These terms will be used for ontology building. Collocation has been widely explored to extract multi terms. It is considered as a phraseological perspective and from a frequency-based perspective [23]. Mostly, three criteria is used to identify collocations which is the distance, frequency and exclusivity [20]. The frequency identification is a simple method to determine the collocation of terms. In statistical, learning collocation has different methods [24]. The methods include i) Mutual Information, ii) Log-Likelihood Measures, and iii) T-test. The methods, used to measure the association between those extracted terms. It calculates the probability that a collocation would occur.

This paper proposed a hybrid of linguistic and statistical approaches. The proposed method has been tested on Hajj corpus. Apart from that, the proposed method tries to analyze specific terms regarding to *Allah*. It is proposed due to the highest occurrence of the term *Allah* in Quranic text translation. The highest term make the term *Allah* is a concept in ontology but at the same time can reduce the chances to get other relevant concept. Therefore, sequence of terms that co-occur with the term *Allah* has been analyzed to see the term collocation. Later, the collocation will be used in building ontology. In the initial extraction, the specific linguistic pattern has been produced for verb- *Allah* collocation. This is differ from extracting terms collocation in Arabic Quran corpus that used the N-grams method [22]. The pattern has been design in JAPE for extracting the terms and the Log Likelihood has been used as an association measure in learning collocation. The measurement used to filter the extracted terms.

3 Methodology

Experiment used English translated Quran. Among eight translations available for the English translation, Hillali Khan Translation has been selected because it contains more elaboration on verses. As the purpose is to construct the Hajj ontology, verses related to Hajj have been selected. The verses or data have gone through for a few processes to produce data such as preprocess, bi-gram application and GATE process.

In pre-processing the data, removing certain symbols from the data is required to get more accurate results when using NLP process. Apart from that, replacing the exist pronoun in the corpus in necessary. The detail process of extraction can be shown as in Fig. 1.

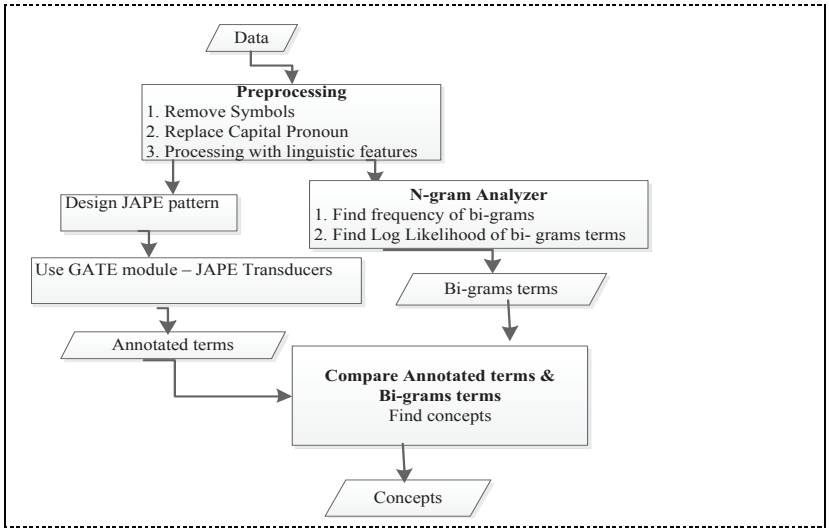

Fig. 1. The detail process of concept extractions.

The pre-processing includes the components as follows:

- Removing symbols, hyphen marks and unnecessary words.
  This component removes punctuation marks such as *Ka'bah, Fir'aun, 'Umrah to Kabah, Firaun, Umrah*. On top of that, it eliminates the hyphen marks such as *Al-Birr* to *AlBirr*. Lastly, it removes the unnecessary words such as *"etc."* and *"i.e"*.

- Replacing the capital pronouns exist in the middle of verses. Pronouns such as *Him, Me, My, Us, We, His Lord* were replaced by term *Allah*.

- Processing with linguistic features such as:
  Tokenizer: splits the text into tokens
  Sentence Splitter: splits the text into sentences
  POS Tagger: adds part-of-speech information to tokens based on Stanford tagger

After the pre-processing has been done, data needs to be run on two paths.

1. Run the data in GATE application.
   In GATE application, a module called JAPE transducers has been called to process the data. The JAPE transducer can annotates the data according to design patterns. Therefore, a pattern regarding the term *"Allah"* is designed using JAPE pattern, as shown in *Fig. 2*. The pattern matches a variant of verbs that have been identified via POS tagger from previous process, followed by term *"Allah"*.

```
Rule: AllahRule
(
({Token.category==VBZ}|{Token.category==VBP}
|{Token.category==VBG}
|{Token.category==VB, Token.kind==word})
({Token.string == "Allah", Token.kind==word})
)
```

Fig. 2. The JAPE pattern.

2. Run the data in N-gram application.
   The N-gram application will process the data into unigrams, bi-grams, tri-grams, 4grams and 5grams terms. The terms will be analysed based on their frequencies and log likelihood ratios. However, this experiment only cover bi-grams analysis since the intended output is bi-grams. Based on the bi-grams terms, frequencies and Log likelihood ratios are being employed in the terms to find collocation terms.

Next, the terms are being compared with the JAPE patterns to find concepts.

4 Results and Discussions

53 verses from Hajj domain has been used in this experiment. The verses contain different topics of Hajj. Among the topics, two topics have the highest number of verses. The topics are "*duty of Hajj and its moral"* which contains 42 verses and *"the honoured Makkah"* which contains 16 verses. Previous experiment produced 6,689 unique terms from 213,679 total terms of extracted from a whole English translated Quran. Meanwhile, 794 unique terms has been extracted from 3125 total terms from Hajj corpus. Based on that, the term Allah shows to be appeared as the highest term in whole English translated corpus (4761) and Hajj corpus (134). The highest term of Allah make the term Allah can be excluded as significant term. But this experiment shows that considering the term Allah can produce the significant terms related to term Allah.

Since the main purpose is to extract concepts for Hajj corpus, only terms related to *Allah* in Hajj corpus has been analyzed. The extractions produced 145 terms that relate to term *Allah*. Based on the total extracted terms, the data has pass through each steps shown as in fig 2. The first path is to use the n-grams analyser. It produces the extracted terms as shown in Table 1.

Table 1. Result of the extracted terms related to Allah.

Terms	No. of terms
Left term of Allah	76
Right term of Allah	69
verb -Allah	4
Allah-verb	38

Terms surrounding Allah which is the left and right terms of Allah has been analysed. Left terms of Allah produced highest number in the extraction which is 76(52.41%) and the rest are the right terms of Allah which is 69(47.58%). Apart from that, the extraction produced 4 terms from left term that is related to verb-Allah and 38 terms from right term that related to Allah-verb. Even though the Allah-verb is higher than verb-Allah, the used of the Allah-verb is common in text such as *Allah destroy, Allah made.* Thus, it cannot be used as a concept. The top 10 extracted terms is shown as in Table 2. The terms are sorted by Log Likelihood and frequency.

Table 2. Results of The top 10 terms related to Allah.

Terms	Log Likelihood (LL)	Frequency

fear Allah	50.4	9
Allah have	36.33	10
Allah has	21.6	6
Allah is	17.67	10
to Allah	16.24	11
of Allah	15.74	16
Allah Messenger	14.37	3
remember Allah	14.34	4
obey Allah	12.65	2
Allah wills	12.65	2

The above table shows that, the highest Log Likelihood value is the term *fear Allah*. Although the frequency of the term *fear Allah* is lower compared to the term *Allah have*, the Log Likelihood ratio will rate the term *fear Allah* as the highest. The Log Likelihood measure the collocation of the term or the strength between words to the term *Allah*. The higher the value, the more the term can be consider as significant terms to be concepts. Based on the extraction, only 3 terms are selected to be concepts which are *fear Allah*, *remember Allah* and *obey Allah*. The second path is using the GATE application. The extracted terms using the JAPE pattern is shown as in Table 3.

Table 3. Results of the pattern extraction.

Terms	Frequency	Terms	Frequency
fear Allah	8	ask Allah	1
remember Allah	4	glorifying Allah	1
have Allah	3	invoking Allah	1
obey Allah	2	magnify Allah	1
reaches Allah	2	purify Allah	1
unto Allah	2	sanctify Allah	1

Based on the table, it shows that the pattern is able to extract concepts that relate to term Allah. The extracted process has extracted 12 candidate concepts based on their frequency. However, only 6 terms are being selected as related terms and the rest are false extracted. The terms such as fear Allah, remember Allah, obey Allah, ask Allah is selected as concepts and invoking Allah, magnifying Allah as an instance of concept remember Allah. Other terms are false extracted because it relates to another word in the verse. For example, the glorifying Allah is actually related to glorifying Allah praises, purify Allah and sanctify Allah are related to purify Allah House and sanctify Allah House.

From both paths, it shows that the used of the statistical method such as Log Likelihood is able to extract collocation terms to be concepts but it cannot extract other relevant terms when having minimum frequency such as 1. Thus, the terms such as invoking Allah, magnify Allah cannot be extracted. Therefore, the Jape patterns can extract more the collocation terms to be concepts. The JAPE pattern can be used for even a small corpus.

5 Conclusions

The technique used to extract relevant terms related to Allah has been presented. The n-grams technique with Log Likelihood and the JAPE pattern are used to extract terms for ontology development. Both techniques are able to extract significant terms. Among the techniques, the JAPE pattern demonstrates that it is able to extract more collocation terms to be concepts. Future works includes extracting relations in the corpus to build hierarchy of the concepts.

References

1. T. R. Gruber, A translation approach to portable ontology specifications, *Knowl. Acquis.*, **5**, April, pp. 199–220 (1993)
2. M. Al-yahya and H. Al-khalifa, An Ontology Model for Representing Semantic Lexicons: An Application on Time Nouns in the Holy Quran, **35,** no. 2, pp. 21–35 (2010)
3. R. Iqbal and A. Mustapha, An experience of developing Quran ontology with contextual information support, **7,** no. 4, pp. 333–343 (2013)
4. H. S. Al-Khalifa, M. M. Al-Yahya, A. Bahanshal, and I. Al-Odah, SemQ: A proposed framework for representing semantic opposition in the Holy Quran using semantic web technologies, in *Proceedings of the 2009 International Conference on the Current Trends in Information Technology, CTIT 2009*, 2009, pp. 44–47 (2009)
5. K. Dukes, Ontology of Quran concepts. [Online]. Available: http://corpus.quran.com/ontology.jsp. [Accessed: 12-Jun-2014].
6. F. Harrag, A. Al-Nasser, A. Al-Musnad, R. Al-Shaya, A. S. Al-Salman, Using association rules for ontology extraction from a Quran corpus (2013)
7. S. Alrehaili and E. Atwell, Computational ontologies for semantic tagging of the Quran: A survey of past approaches, *Lr. 2014 Proc.* (2014)
8. A. J. Petiwala and S. S. Sathya, A Multi-Agent System to Learn Literature Ontology : An Experiment on English Quran Corpus, pp. 46–51 (2011)
9. S. Saad, N. Salim, and S. Tiun, Concept Extraction on Quranic Translation Text, *Int. J. Islam. Appl. Comput. Sci. Technol.*, **2,** no. 1, pp. 1–9 (2014)
10. M. F. Lopez, A. Gomez-Perez, J. P. Sierra, and A. P. Sierra, Building a chemical ontology using Methontology and the Ontology Design Environment, *IEEE Intell. Syst. their Appl.*, **14,** no. 1, pp. 37–46 (1999)
11. A. Maedche and S. Staab, The text-to-onto ontology learning environment, *Softw. Demonstr. ICCS-2000, Eight Int. Conf. Concept. Struct.* (2000)
12. W. Wong, W. Liu, and M. Bennamoun, Ontology learning from text, *ACM Comput. Surv.*, **44,** no. 4, pp. 1–36 (2012)
13. B. E. Benaissa, D. Bouchiha, A. Zouaoui, and N. Doumi, Building Ontology from Texts, *Procedia Comput. Sci.*, **73,** no. Awict, pp. 7–15 (2015)
14. S. Saad, N. Salim, and H. Zainal, Islamic knowledge ontology creation, *2009 Int. Conf. Internet Technol. Secur. Trans.*, no. NOVEMBER 2009, pp. 1–6 (2009)
15. X. Jiang and A. H. Tan, CRCTOL: A semantic-based domain ontology learning system, *J. Am. Soc. Inf. Sci. Technol.*, **61,** pp. 150–168 (2010)
16. M. Shamsfrad and A. A. Barforoush, The state of the art in ontology learning : a framework for comparison, *Knowl. Eng. Rev.*, **18,** pp. 293–316 (2003)
17. P. Buitelaar, P. Cimiano, and B. Magnini, Ontology Learning from Text : An Overview, pp. 1–10, 2004.
18. S. Zaidi, M. T. Laskri, and A. Abdelali, Arabic collocations extraction using gate, *2010 Int. Conf. Mach. Web Intell. ICMWI 2010 - Proc.*, pp. 473–475 (2010)

19. S. Ghadfi, N. Bechet, and G. Berio, Building ontologies from textual resources: A pattern based improvement using deep linguistic information, *CEUR Workshop Proc.*, **1302** (2014)
20. V. Brezina, T. Mcenery, and S. Wattam, Collocations in context A new perspective on collocation networks*, *Int. J. Corpus Linguist.*, **202,** pp. 139–173 (2015)
21. M. Roche and Y. Kodratoff, Pruning Terminology Extracted from a Specialized Corpus for CV Ontology Acquisition, *OTM Confed. Int. Conf. Move to Meaningful Internet Syst.*, pp. 1107–1116 (2006)
22. W. Alromima, I. F. Moawad, R. Elgohary, and M. Aref, Extracting N-gram terms collocation from tagged Arabic corpus, *2014 9th Int. Conf. Informatics Syst. INFOS 2014*, pp. NLP10–NLP15 (2015)
23. J. Parkinson, Noun-noun collocations in learner writing, *J. English Acad. Purp.*, **20,** pp. 103–113 (2015)
24. F. Xu and D. Kurz, Text Mining for the Extraction of Domain Relevant Terms and Term Collocations, *Proc. Int. Work. Computational Approaches to Collocations.* (2002)