Design for Scientific RENAISSANCE

IJASAT

Extraction of Multi-Word Terms and Complex Terms from the Classical Arabic Text of the Quran

Sameer M Alrehaili[1, 2, a] and Eric Atwell[2, b]

Faculty of Information Science and Technology
[1]College of Computer Science & Engineering, Taibah University
Yanbu, Saudi Arabia
[2]School of Computing, University of Leeds
Leeds, United Kingdom
[a]salrehaili@gmail.com, [b]e.s.atwell@leeds.ac.uk

ABSTRACT

The identification of domain-specific terms is a crucial step in many natural language processing applications. Term extraction is a process of obtaining a set of terms that represent the domain of a given text. The majority of term extraction research projects conducted for the Quran have used translated text instead of the original Classical Arabic text of the Quran. The extraction of terms from the original Arabic text rather than a translation may help in retrieving more relevant terms, due to the lack of Islamic equivalents of some Quran terms in other languages. This paper demonstrates a hybrid-based method for the acquisition of a list of domain-specific terms from the Arabic text of the Quran. The produced list of terms was validated using a common evaluation metric for ranked list; precision of up to 0.81 was achieved for the top 200 terms. We discuss the precision that was achieved, in the context of two existing datasets from previous research.

Keywords: term extraction, automatic term recognition, Quran terms

1. INTRODUCTION

Muslims believe the Holy Quran is the word of Allah and the last sacred book of those that were sent down by Allah to his prophets. Muslims consider the Quran to be their primary source of knowledge and guidance; therefore, their daily life is dependent on what is written in the Quran (i.e., the rules of marriage, divorce, inheritance, finance, etc.).

Terms are the basic linguistic units that describe an entity in a domain, using a word or a phrase. Terms can be made of one word (single-word) or a group of words (multi-word). Multi-word terms (MWTs) are believed to be less polysemous than single-word terms (Boulaknadel et al. 2008) and also form the majority of any ontology - approximately 85% of domain-specific terms (Nakagawa & Mori 2002). Multi-word terms or compound terms are important to be extracted because they carry more meaning than the single-word terms. The majority of domain-specific terms are multi-word terms and the specificity of a multi-word term is higher than the specificity of the a single-word term (Ryu & Choi 2005). However, multi-word terms have low appearance in corpora compared to single-word terms (El-Beltagy et al. 2009; Boulaknadel et al. 2008) and this may hinder statistical-based extraction approaches. Multi-word terms are classified into two main categories: simple term and complex term. Complex terms are not only those that are made of two words or more, but they can be nested within other terms as a modifier.

1.1 Arabic Multi-Word Term Variations

English multi-word terms are often sequences of nouns; so a first step in term extraction from English texts is to run a Part-of-Speech (POS) tagger. Arabic multi-word terms are not limited to a specific sequence of POS tags such as $\{N\ N\} or\ \{N\ ADJ\}$. Other POS tags like conjunctions, verbs, prepositions and pronouns can also be a constituent of such terms. Moreover, a term can modify another term or can be nested in another term. Table 1 shows an example of a complex multi-word term in Arabic that is composed of a syntactic pattern of six different POS tags, found in the QurAna project (Sharaf & Atwell 2012).

Table 1: A Multi-word term composed of different Part-of-Speech (POS) information

Transliteration	nn	lm	yHkm	b	mA	Anzl	Allh
English	whoever	Does not	Judge	by	What	Has revealed	Allah
POS	COND	NEG	V	P	REL	V	PN

This term also has a multi-word term nested in it, as can be seen in the Figures 1a and 1b, which show our example of Arabic complex term in two representation models, the term "mA Anzl Allah" is nested in or modifies the main term "mn lm yHkm b mA Anzl Allah". "Allah" is also another nested term in the nested term of the main term.

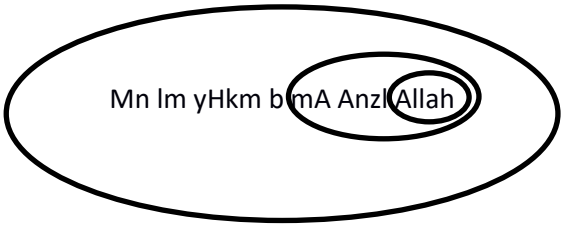

Figure 1a. Complex term represented by Wilson's Nested Model

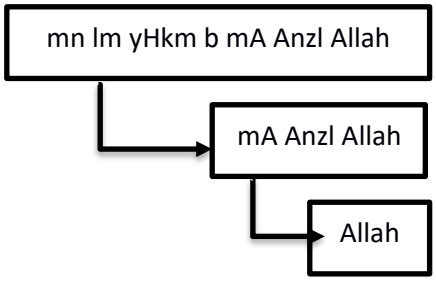

Figure 1b. Complex term represented by a tree structure

Manual extraction of terms can be led by a person who is an expert in the target domain; however, even domain experts misidentify some terms, due to the subjectivity and variation of the decision process from one to another (Nazarenko & Zaegayouna, 2009). In computing, automatic term recognition, also known as term extraction, is a process of obtaining a set of actual terms that are relevant to a given text. According to Cimiano (2006), term extraction is the foremost step, and it is required by further tasks in different complex applications. A variety of natural language processing (NLP) applications, such as automatic labelling of articles, automatic thesaurus construction, ontology learning and machine translation, require terminology extraction.

Term extraction typically involves three steps: generation of candidate terms, scoring of the candidates, and validation. The generation of candidate terms usually begins with preprocessing of the text, for example, part-of-speech (POS) text tagging, followed by the search for a specific set of predefined patterns (i.e.$\{N\ ADJ, N, N\ N\}$). At this step, some filters can also be implemented, such as stopword-list elimination. In the next step, scoring of the candidates, statistical methods are applied to analyse the importance of these candidates and the relevance of the generated terms. The final step is the validation of the correctness of the candidates, and this validation is dependent on the availability of the resources (Norman 2015). For some domains, a gold-standard is publicly available. However, some domains, such as the Quran, are limited to certain parts or a certain level of scope, or are even implemented using translations of the Quran, rather than the original Arabic text (Alrehaili &

Atwell 2014) , for example, ontology made of time nouns by Al-Yahya & Al-Khalifa (2010) and ontology based on living creatures by Ullah Khan et al. (2013). Therefore, for evaluation purposes, it may not be appropriate to choose previous datasets.

This paper is organised as follows. Section 2 presents related methods and research on term extraction from the Quran; Section 3 describes our adapted approach to identifying terms from the Arabic Quranic text; Section 4 presents our results and evaluations; and in Section 5 we draw our conclusion.

2. RELATED WORK

In recent years, there has been an increase in Quran knowledge extraction research. The focus here will be studies that have extracted the ontology of the Quran, particularly those that have used term extraction.

Most existing approaches used for term extraction to date can be divided into three categories: (1) linguistic approaches, (2) statistical approaches, and (3) hybrid approaches. Linguistic approaches exploit NLP techniques, such as tokenisation, morphological analysis, POS tagging and stemming and parsing, for detecting terms from a given text. This method is usually dependent on the selected domain and would not work perfectly in other domains because of its language dependency. For example, linguistic-based methods for extracting medical terms search for some medical characteristics in the text itself, such as abbreviations and doctors' instructions, while other domains do not have the same characteristics.

Statistical approaches overcome language dependency because they rely on independent measures for assessing the importance of extracted candidates; measures such as frequencies, likelihood, term frequency-inverse document frequency (TF-IDF) and mutual information, can be calculated for any domain. However, some statistical methods are incapable of addressing low-frequency terms. In fact, the majority of the words in most corpora have low frequencies, especially MWTs occurring only once or twice. This means that the MWTs are excluded by statistical approaches. Hybrid approaches combine different methods from linguistics and statistics for detecting terms in the text. Firstly, applying the linguistic method to generate term candidates. After that, statistical measures are used for filtering out invalid candidates.

The majority of term extraction research work conducted on the Quran have used translated text instead of the original text, such as Saad & Salim (2008); however, extraction of terms from the original Arabic text may help in retrieving more relevant terms than an extraction from a translation. This is because some Islamic terms have no equivalents in other languages (Ali, Abobaker and Brakhw, M Alsaleh and Nordin, Munif Zarirruddin Fikri Bin and ShaikIsmail 2012; Kashgary 2011). Another reason is that some translations do not consider the meaning of the terms. For example in the Table 1, "whoever does not judge by what has revealed Allah". I can only see one term which is "Allah" in the translation of that term.

Few attempts have been made to use the Arabic Quran in tasks related to term extraction; we found only two, conducted by Harrag et al. (2014) and Alhawarat (2015). Therefore, further research is required in this area. Previous studies that examined the extraction of terms from the Arabic text of the Quran proposed hybrid methods that are based on syntactic patterns and TF-IDF for extracting single-terms (Harrag et al. 2014; Alhawarat 2015).

Harrag et al. (2014) used methods that rely on linguistics and statistical language modelling to generate ontology elements. The concept extraction was based on KP-Miner, which, in turn, uses TF-IDF, which requires some specifications that are not available for the Quran, such as the size of the corpus. No results about the extraction performance were mentioned because the focus was to extract the conceptual relations. Alhawarat (2015) aimed to extract the topic of the chapter no 12 of the Quran using Arabic text. He applied a package called tm from R which implements topic modelling from Latent Dirichlet Allocation (LDA).

We found two previous works that examined the extraction of multi-word terms from Modern Arabic text (El-Beltagy & Rafea 2010; Boulaknadel et al. 2008). Boulaknadel & Daille, (2008) investigated different statistical measures and linguistic techniques for multi-word term extraction for Arabic technical text from environment domain. The authors concluded that Log-likelihood ratio (LLR) achieved the best performance. (El-Beltagy & Rafea (2010) developed a system, called KP-Miner, for key-phrases extraction from both Arabic and English documents. Arabic documents were selected from Agriculture domain. KP-Miner combined TF-IDF with a boosting factor that makes the balance between low frequency of compound words and high frequency of single words in a corpus.

To the best of our knowledge, this paper is the first in extraction complex terms from the Classical Arabic text of the Quran. Our method requires no training data. Moreover, it can successfully capture complex MWTs from Arabic Quran text.

Another work related to the validation process of our method is (Dukes 2012, 2013) which described the Quranic Arabic Corpus (QAC) project, an online Quran that was annotated at several levels, which included an ontology that defines 300 concepts in the Quran, and captures interrelationships using predicate logic. The number of relationships is 350, and the type of relationship between concepts is “Is-a”. The ontology is based on the Tafsir or Quran commentary textbook by Ibn Kathir. The QAC also contains other analyses of the Quran text, such as POS, morphological analysis and dependency parse structure analysis.

Sharaf & Atwell (2012) developed an ontology that encompassed the entire Quran in terms of pronoun tagging called QurAna, whereby each pronoun is linked to its syntactic antecedent or previous reference, and its concept in an ontology of pronoun referrents. The dataset comprises about 24,000 personal pronouns in the Arabic text, each linked to its antecedent and its concept in the ontology. This can be used in ontology extraction in an ontology learning system using anaphora analysis to extract the concepts and relationships. Over 1,000 antecedents which are also domain-specific terms about the text, were arranged in a dataset called QurAna.

Mukhtar et al. (2012) produced a dataset that contains concepts from the second chapter of the Quran, known as the Vocabulary of Quranic Concepts. They used six different English translations of the Quran and applied a domain-independent tool called Termine from Frantzi & Ananiadou (1999) to extract the concepts. However, Termine was designed for the extraction of multi-word terms, while the Quran has numerous single-word concepts (e.g., Allah and Muhammad). Therefore, application of such a method may exclude some important concepts.

3. METHODOLOGY

Our approach aimed to identify a list of terms for the Arabic text of the Quran. We divided this section into two parts: data collection and preparation and term extraction. The first of these provides a brief description of the data that were used in this study, while the second outlines the steps we followed to extract the term list.

3.1 Data Collection and Preparation

We collected data from a range of important resources for the Quran. A wide range of Quran annotations is available in either computer-readable or hand-written formats. Our extraction method used multiple data sources, including the Tanzil Quran project (Zarrabi 2007), the Quranic Arabic Corpus QAC (Dukes 2012, 2013) and Qurany (Abbas 2009, 2013). The source of the Quran text was downloaded from the Tanzil Quran project (Zarrabi 2007), which provided a digital copy of the Quran that has been manually validated by a group of experts against an accepted standard written-text version: Madinah Mushaf. The text of the Quran is stored in a text file, which is composed of 6,236 lines, and each line represents a verse in the Quran. Words, morphemes and POS information was collected from the Quranic Arabic Corpus (Dukes 2012, 2013), which was manually verified.

3.2 Term extraction

In order to generate a list of Quran terms, we adapted the weighting scheme from Kang et al. (2014). The motivation behind taking inspiration from this method was that the test text does not need to be very long. This methodology is based on a set of linguistic patterns, and statistical and domain-specific knowledge. Our extraction method can be explained in a number of steps, as follows:

i. **Preparation of a predefined list of syntactic patterns:** We began by generating a list of predefined syntactic patterns, and we manually extracted all noun and noun phrase patterns from Chapter 29 of the Quran. We extracted approximately 470 nominal items with their syntactic patterns. We tagged all different terms in this chapter including nested terms, and for every verse, and also annotated their sequences of syntactic patterns. Table 2 shows the first 10 lines of our pattern list, and Table 3 shows the most frequent patterns in the entire Quran text.

   The syntactic patterns in the second column of Table 3 include dots between every segment; this feature was required to extract complex terms, such as that in line no. 6 in Table 2, ‘الذين يعملون السيئات - those who do evil deeds.’ Most pronouns and prepositions found in complex terms are attached to the previous part of the term. Therefore, including dot to distinguish the attached ones will help in improve the performance of extraction. Line no. 6 in the Table 2, shows an important type of terms that is composed not only from noun and adjectives, but pronouns and verbs. In addition, this term also contains another term nested in. We did not remove these pronouns and verbs when generating the candidates. Therefore, we kept these dots to ensure that the candidates were correctly generated. Furthermore, we did not miss those terms located as part of other terms, such as ‘السيئات - evil deeds’ in line 7.

Table 2: An example of patterns of terms found in the chapter 29

No.	Syntactic pattern	Transliteration	English translation
1	DET.N	AlnAs	The people
2	REL P N.PRON	Al*yn mn qblhm	Those who were before them
3	PN	Allh	Allah
4	REL V.PRON	Al*yn SdqwA	Those who are the truthful
5	DET.N	AlkA*byn	The liars
6	REL V.PRON DET.N	Al*yn yEmlwn Alsy}At	Those who do evil deeds
7	DET.N	Alsy}At	Evil deeds
8	REL V.PRON	mA yHkmwn	What they judge
9	REL V V N PN	mn kAn yrjwA lqA' Allh	Whoever is hopes meeting Allah
10	N	lqA'	meeting

Table 3: The top 10 most frequently occurring noun phrase syntactic patterns in the Quran

No.	Syntactic pattern	Occurrences
1	N	25,136
2	DET.N	7,488
3	PN	3,911
4	REL V	2,919
5	N N	2,790
6	ADJ	1,961
7	REL V.PRON	1,885
8	N DET N	1,347
9	N V.PRON	805
10	N PN	804

The outcome of this step was a list of syntactic patterns for all of the term candidates in the Chapter 29 which we assume is a good represent for the different types of terms for the Quran.

ii. **Extracting term candidates:** In this step, our aim was to apply regular expressions to chunk a list of term candidates. To this end, we replaced every segment with their POS for tokens that were composed of more than one segment. We used dots between segments, as shown in Figure 1. For example, the 'بسم الله' would be P.N PN. We then applied a regular expression search for all patterns in the predefined list that we obtained in the previous step. The function of regular expression retrieved the position of the first character for every matched pattern or -1 if they were a match. For every position, we extracted the corresponding text from the original text file. The outcome of this step was the initial candidate terms.

```
1  P.N PN DET.ADJ DET.ADJ
2  DET.N P.PN N DET.N
3  DET.ADJ DET.ADJ
4  N N DET.N
5  PRON V CONJ.PRON V
6  V.PRON DET.N DET.ADJ
7  N REL V.PRON P.PRON N DET.N P.PRON CONJ.NEG DET.N
8  INL
9  DEM DET.N NEG N P.PRON N P.DET.N
10 REL V.PRON P.DET.N CONJ.V.PRON DET.N REM.P.REL V.PRON.PRON V.PRON
```

Figure 1. Syntactic patterns for the first 10 verses in the Quran

iii. **Candidate weighting:** In this step, a combination of statistical and domain-specific knowledge was created, based on formula (3), which was proposed by Kang et al. (2014). We chose this method because it works well, even for a small text size. The statistical

knowledge indicates the importance of a candidate in the text; simply, computing the relative frequency a candidate t appeared in the corpus $p(t)$, as explained in formula number (1). The domain-specific knowledge, $w_d(t)$, was the number of times that t appeared as part of glossary list G, as described in equation (2). We chose the dataset of Abbas (2009, 2013) because it is the only topic list that is available in computer-processable form for the Quran.

$$P(t) = \frac{f(t)}{\max\limits_{1 \le i \le |TC|} f(t_i)} \tag{1}$$

Where t was a candidate term, $f(t)$ is the number of times that candidate $t \in TC$ appeared in the corpus D, $\max\limits_{1 \le i \le |TC|} f(t_i)$ is the maximum number of term t that appeared in the corpus, and D, $P(t)$ is the statistical knowledge for a given t.

$$W_d(t) = 1 + \frac{\log\big(df(t)\big)}{\log\left(\max\limits_{1 \le i \le |TC|} df(t_i)\right)} \tag{2}$$

In which $df(t)$ is the number of times that t appeared as part of a term in the glossary list G, $\max\limits_{1 \le i \le |TC|} df(t_i)$ is the maximum occurrences of t as part of another term from G, $W_d(t)$ is the domain-specific knowledge for a given t, and $|t|$ is the length of a term with regard to words number,

$$W(t) = \begin{cases} P(t) \times W_d(t), & if\ |t| = 1 \\ \sum\limits_{i=1}^{|t|} W(t_i), & otherwise \end{cases} \tag{3}$$

where $\sum_{i=1}^{|t|} W(t_i)$ is the sum of the weight of each nested term $t_i\ of\ t$ if it was longer than one word. Computing nested terms is based on recursion technique shown in the Algorithm 1. And $W(t)$ is the total weight of a term.

Algorithm 1: Nested terms weight calculation

Input: t ← a term t = $w_1\ w_2$ ... w_n
function Weight (t)
1: **if** t is one word
2: **return** count(t)
3: **else**
4: c ← P(t) * W_d(t)
5: head ← is the first token of t
6: tail ← is the rest of words
7: **return** c + P(head) * W_d (head) + Weight(tail)

iv. Ranking: Finally, we obtained weighted candidates, and the task was to reorder them in descending order.

4. RESULTS AND EVALUATION

After we had assigned the weights using linguistic and statistical techniques for all candidates, we could then rank them in descending order, as illustrated in Table 4.

Table 4: The top 10 candidates after weighting and ranking

Transliteration of t	$P(t)$	$W_d(t)$	$W(t)$	rank	rel
al-lāh	1	0.836238387	0.836238387	1	1
rabb	0.485334	0.595748247	0.289137104	2	1
yawm	0.230739	0.515237581	0.11888548	3	1
Al Arudh	0.173641	0.418119194	0.072602629	4	1
Qawwam	0.1623	0.446811186	0.072517263	5	1
mā kano	0.149003	0.446811186	0.06657609	6	0
Alnas	0.074306	0.533933782	0.039674391	7	1
Ma kan	0.084083	0.446811186	0.037569185	8	0
Man rabb	0.045757	0.808171929	0.036979318	9	0
Kẖayr	0.074306	0.49475821	0.036763418	10	1

Table 4 shows the top 10 ranked candidates. The first column contains the term; the second contains the statistical information that is explained in equation (1); the third is the domain-specific knowledge, whereby we computed the ratio of term appearance as part of the terms in the Glossary list; the fourth is the total weight that was used to rank the terms; and the last column shows the status of the term, that is, whether it was relative or non-relative. As can be observed, this domain-specific information helped in retrieving multi-word terms, even when they did not occur as single-word terms.

4.1 Evaluation against previous datasets

One previous dataset, namely QurAna (Sharaf & Atwell 2012), primarily relied on pronouns mentioned in the Quran and linked them to a reference list composed of 1,028 concepts. These concepts only encompassed names or things that had been mentioned using pronouns, and did not cover those nouns that were not mentioned by their pronouns. Another dataset for Quranic concepts was established by Dukes & Atwell (2012) and Dukes (2013). By comparing the large extracted list with the small list we obtained low precision, as we had expected. The best performance was observed in comparison to QAC and we achieved 0.62 precision overall.

4.2 Comparison of previous available datasets for a selected chapter

We collected all available terms and concepts from previous work for chapter 29. In addition, we asked two independent annotators to identify the concepts from the same chapter. Table 5 shows the comparison between these datasets in terms of how many of the concepts occurred in each verse. A1 is the annotation made by annotator1, while A2 is the data from annotator2.

Table 5: A comparison of different existing annotations from previous studies and manual annotations

Datasets	Terms	Unique terms
QAC	27	20
QurAna	324	48
Qurany	173	133
A1	497	348
A2	468	299

Table 5 illustrates the total number of terms found with previous datasets and through manual annotations. Manual annotators identified more terms for a certain chapter of the Quran. This is because when we asked the annotators to annotate, we did not tell them to focus on a specific scope or pick certain patterns. QAC, QurAna, and Qurany are specialised for some specific proposes, which reveals why our extraction method did not achieve high precision in comparison.

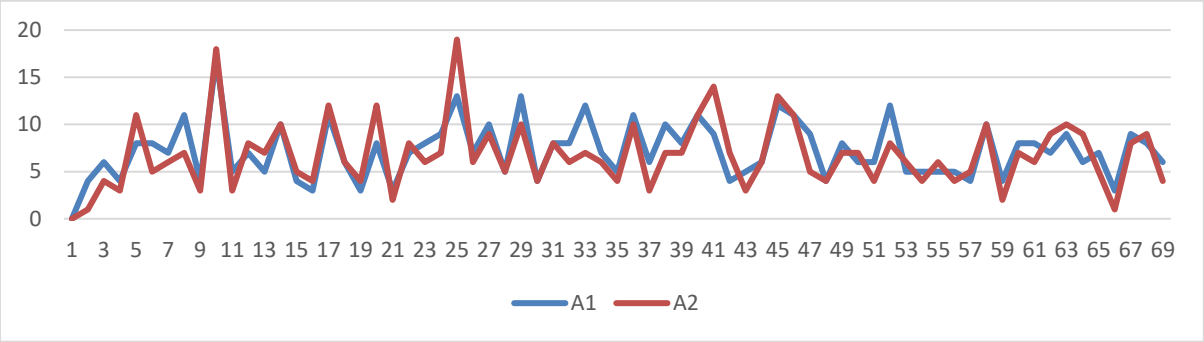

Figure 2. A comparison of hand-annotated terms

Figure 2 shows the agreement between the two annotators who were asked to annotate chapter 29 of the Quran. Although they carried out the task independently, Figure 3 shows that they were very close together in terms of the numbers. However, this does not mean that their extracted terms for a certain verse are similar. We only focused on the number at this stage, to obtain a quick idea of how similar they were to each other.

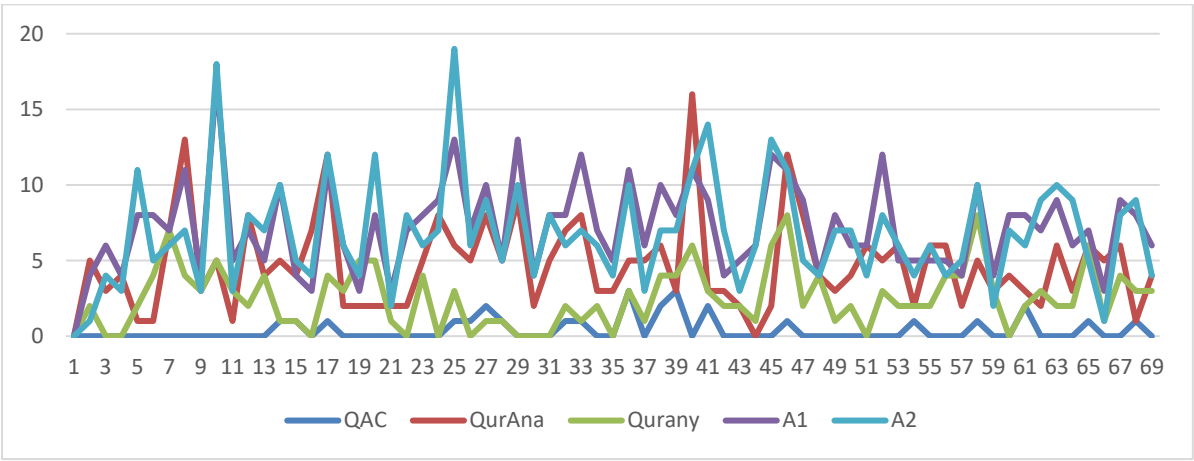

Figure 3. A comparison of hand-annotated and collected terms with those of previous related work

This indicates that these datasets are not complete; therefore it is possible that our method may identify relevant terms that have not been covered in previous datasets. Therefore, we manually validated the extracted terms by a binary judgment that indicated which terms were relevant and which were non-relevant, after which we applied average precision (AvP). AvP is a very popular evaluation metric that is widely used to test the performance of term extraction methods. It is the sum of all precision to rank k over rank number (see equation (4)).

4.3 Evaluation against the average

$$AvP = \frac{\sum_{k=1}^{n}(p(k) \times rel(k))}{n_c} \quad (4)$$

Where $p(k)$ is the precision at cut-off k in the terms list, n means the size of the extracted terms list, n_c is the total number of relevant terms that were retrieved by the method and $rel(k)$ is a binary function that indicated whether or not the retrieved term was relevant. The output of $rel(k)$ is 1 if a $term_k$, which means the term at k, is relevant to the Quran domain and 0 otherwise.

$$R@k = \frac{the\ number\ of\ relevant\ retrieved\ at\ rank\ k}{all\ relevant\ retrieved}$$

$$P@k = \frac{the\ number\ of\ relevant\ retrieved\ at\ rank\ k}{number\ of\ relevant\ and\ non-relevant\ retrieved\ at\ rank\ k}$$

Where $R@k$ is the recall at rank k and $P@k$ is the precision at rank k.

Table 6: The average precision for the first 1,000 terms in our list

k	$recall$	$precision$	AvP
1	0.001789	1	1.000000
50	0.06619	0.74	0.778600
100	0.144902	0.81	0.784534
150	0.211091	0.786667	0.790667
200	0.289803	0.81	0.793123
250	0.357782	0.8	0.796795
300	0.402504	0.75	0.792789
350	0.457961	0.731429	0.785697
400	0.516995	0.7225	0.778722
450	0.567084	0.704444	0.771061
500	0.601073	0.672	0.762724
550	0.65653	0.667273	0.754340
600	0.695886	0.648333	0.746516
650	0.726297	0.624615	0.737926
700	0.749553	0.599428	0.728578
750	0.801431	0.598131	0.719892
800	0.844365	0.590738	0.711766
850	0.876565	0.57715	0.704337
900	0.892665	0.555061	0.696761
950	0.942755	0.555321	0.689081
1,000	1	0.55956	0.682584

This table shows the AvP of the top 1000 extracted terms, and we can clearly observe that those ranked nearest to the top had high precision, which then decreased in accordance with the increase in size. Recall increased as the number of candidates rose, while precision decreased. We obtained an overall precision of 0.81 for the first 200 terms.

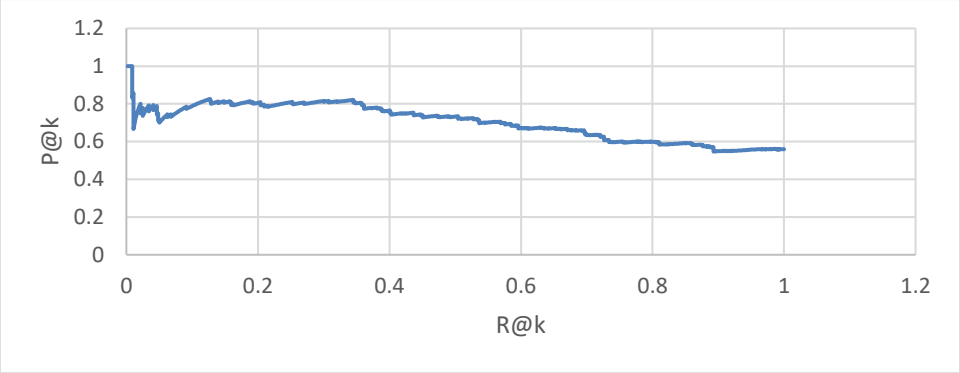

Figure 4. Recall-precision graph for the first 1,000 extracted terms

Figure 4 shows the precision for every k in the list. The precision associated with the candidates at the very top was higher than the precision at the bottom

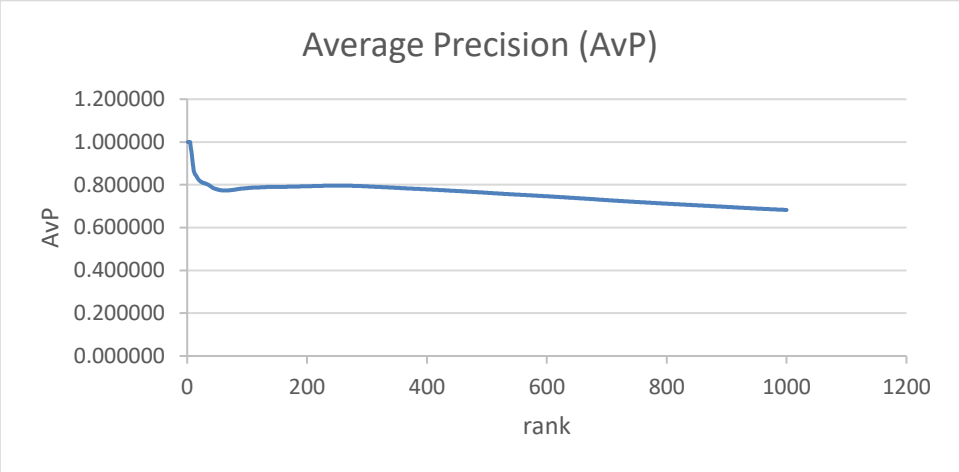

Figure 5. The stages in the terms extraction process

Figure 5 shows the relationship between precision and rank number. Instead of the relationship between recall and precision, as shown in Figure 4, this graph clearly shows that our methods achieved approximately 0.65 as overall precision and 0.8 precision for the first 200 candidates.

5. CONCLUSION

This paper presented a method to identify terms from the Arabic text of the Quran as well as assessing these against three types of evaluation. The datasets from previous studies of the Quran are not complete, and are not appropriate for use in evaluation of the extracted terms, because of the variation in the size and spelling of the text used in each dataset. Moreover, these datasets have been generated to cover only some scope or some parts of words, and agreement between domain experts cannot be guaranteed due to the subjectivity. We evaluated the extracted terms against AvP and achieved precision of up to 0.81 for the top 200 terms. We discussed the limitations when evaluating against two existing datasets.

6. References

Abbas, N.H., 2009. *Quran'search for a Concept'Tool and Website*. Unpublished MSc Dissertation, School of Computing, University of Leeds. Available at:

http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Quran+?Search+for+a+Concept?+Tool+and+Website#0 [Accessed November 26, 2013].

Abbas, N.H. & Atwell, E., 2013. Annotating the Arabic Quran with semantic web content tags. In E. Atwell & A. Hardie, eds. *Proceedings of WACL-2 Second Workshop on Arabic Corpus Linguistics*. Lancaster, UK, pp. 54–55.

Al-Yahya, M. & Al-Khalifa, H., 2010. An Ontological Model for Representing Semantic Lexicons: An Application on Time Nouns in the Holy Quran. *The Arabian Journal for Science and Engineering*, 35(2), pp.21–35. Available at: http://www.researchgate.net/publication/228955782_An_Ontological_Model_for_Representing_Semantic_Lexicons_An_Application_on_Time_Nouns_in_the_Holy_Quran/file/50463516eca79add3d.pdf [Accessed November 26, 2013].

Alhawarat, M., 2015. Extracting Topics from the Holy Quran Using Generative Models. *INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS*, 6(12), pp.288–294.

Ali, Abobaker and Brakhw, M Alsaleh and Nordin, Munif Zarirruddin Fikri Bin and ShaikIsmail, S.F., 2012. Some Linguistic Difficulties in Translating the Holy Quran from Arabic into English. *International Journal of Social Science and Humanity*, 2(6), pp.588–590.

Alrehaili, S.M. & Atwell, E., 2014. Computational ontologies for semantic tagging of the Quran : A survey of past approaches . In *Proceedings of the 2nd Workshop on Language Resources and Evaluation for Religious Texts*. Reykjavik, Iceland, pp. 19–23. Available at: http://www.lrec-conf.org/proceedings/lrec2014/workshops/LREC2014Workshop-LRE-Rel2 Proceedings.pdf.

Boulaknadel, S., Daille, B. & Aboutajdine, D., 2008. A multi-word term extraction program for Arabic language. In *Language Resources and Evaluation Conference (LREC)*. pp. 3–6. Available at: http://pages.cs.brandeis.edu/~marc/misc/proceedings/lrec-2008/pdf/378_paper.pdf.

Cimiano, P., 2006. *Ontology learning from text*, Springer US. Available at: http://medcontent.metapress.com/index/A65RM03P4874243N.pdf [Accessed April 15, 2014].

Dukes, K., 2013. *Statistical Parsing by Machine Learning from a Classical Arabic Treebank*. PhD Thesis, School of Computing, University of Leeds. Available at: http://www.kaisdukes.com/papers/thesis-dukes2013.pdf%5CnAll Papers/D/Dukes 2013 - Statistical Parsing by Machine Learning from a Classical Arabic Treebank.pdf.

Dukes, K. & Atwell, E., 2012. LAMP : A Multimodal Web Platform for Collaborative Linguistic Analysis. *Lrec 2012*, pp.3268–3275.

El-Beltagy, S.R. & Rafea, A., 2010. KP-Miner: Participation in SemEval-2. In *Proceedings of the 5th international workshop on semantic evaluation*. Association for Computational Linguistics, pp. 190–193. Available at: http://www.aclweb.org/anthology/S10-1041 [Accessed March 23, 2017].

El-Beltagy, S.R., Rafea, A. & Melamed, I.D., 2009. KP-Miner: A keyphrase extraction system for English and Arabic documents. *Information Systems*, 34(1), pp.132–144.

Frantzi, K.T. & Ananiadou, S., 1999. The C-value/NC-value domain-independent method for multi-word term extraction. *Journal of Natural Language Processing*, 6(3), pp.145–179.

Hardeniya, N., 2015. *NLTK Essentials Build cool NLP and machine learning applications using NLTK and other Python libraries*, Packt Publishing Ltd.

Harrag, F. et al., 2014. Using association rules for ontology extraction from a Quran corpus. In *Proc. 5th Int. Conf. Arabic Language Process.* pp. 1–8.

Kang, Yong-Bin and Haghighi, Pari Delir and Burstein, F., 2014. CFinder: An intelligent key concept finder from text for ontology development. *Expert Systems with Applications*, 41(9), pp.4494–4504.

Kashgary, A.D., 2011. The paradox of translating the untranslatable: Equivalence vs. non-equivalence in translating from Arabic into English. *Journal of King Saud University - Languages and Translation*, 23(1), pp.47–57. Available at: www.ksu.edu.sa.

Mukhtar, T., Afzal, H. & Majeed, A., 2012. Vocabulary of Quranic concepts: A semi-automatically created terminology of Holy Quran. In *2012 15th International Multitopic Conference, INMIC 2012*. IEEE, pp. 43–46.

Nakagawa, H. & Mori, T., 2002. A Simple but Powerful Automatic Term Extraction Method. In *COLING-02 on COMPUTERM 2002: second international workshop on computational terminology*. COMPUTERM '02. Stroudsburg, PA, USA: Association for Computational

Linguistics, pp. 1–7. Available at: http://dx.doi.org/10.3115/1118771.1118778.

Norman, C., 2015. *Technical Term Extraction Using Measures of Neology*. MSc Dissertation, Department of Computer Science, KTH Royal Institute of Technology,Stockholm, Sweden.

Perkins, J., 2010. *Python text processing with NLTK 2.0 cookbook*, Birmingham, UK: Packt Publishing Ltd.

Ryu, P.-M. & Choi, K.-S., 2005. An Information-Theoretic Approach to Taxonomy Extraction for Ontology Learning. In *Ontology Learning from Text: Methods, Evaluation and Applications*. p. 15. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.101.7975&rep=rep1&type=pdf [Accessed March 22, 2017].

Saad, S. & Salim, N., 2008. Methodology of Ontology Extraction for Islamic Knowledge Text. In *Postgraduate Annual Research Seminar*.

Sharaf, A. & Atwell, E., 2012. QurAna: Corpus of the Quran annotated with Pronominal Anaphora. In *LREC Language Resources and Evaluation Conference*. Istanbul, Turkey, pp. 130–137. Available at: http://www.researchgate.net/publication/228522230_QurAna_Corpus_of_the_Quran_annotated_with_Pronominal_Anaphora/file/60b7d518ab73049436.pdf [Accessed May 15, 2014].

Ullah Khan, H. et al., 2013. Ontology Based Semantic Search in Holy Quran. *International Journal of Future Computer and Communication*, 2(6), pp.570–575. Available at: http://www.ijfcc.org/index.php?m=content&c=index&a=show&catid=43&id=493 [Accessed December 16, 2013].

Zarrabi, H.-Z., 2007. Tanzil Project. Available at: http://tanzil.net/wiki/.