
White Rose Research Online

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk
Universities of Leeds, Sheffield and York

Deposited via The University of Leeds.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/id/eprint/112947/

Version: Accepted Version

Proceedings Paper:
Alrehaili, SM and Atwell, E (2016) A Hybrid-based Term Extraction method on the Arabic text of the Qur'an. In: IMAN'2016 4th International Conference on Islamic Applications in Computer Science and Technologies. IMAN'2016 4th International Conference on Islamic Applications in Computer Science and Technologies, 20-22 Dec 2016, Khartoum, Sudan.

Reuse
Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.


UNIVERSITY OF LEEDS
University of Sheffield
UNIVERSITY of York

A hybrid-based Term Extraction method using the Arabic text of the Qur'an

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

[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 relevant domain terms is a crucial step in numerous 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 Qur'an have used translated text instead of the original text of the Qur'an. 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 equivalence of some Quranic 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 validated a common evaluation for ranked list; precision of up to 0.81 was achieved for the top 200 terms. We discussed the low precision that was achieved, in the context of evaluate the result against two existing datasets from previous research.

***Keywords*: term extraction, automatic term recognition, Quranic terms**

1. Introduction

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

Terms are the basic units that describe an entity in a domain, using a word or a phrase. A person who is an expert in the domain of particular terms may not find it difficult to identify them; however, even domain experts misidentify some terms, due to the subjectivity and variation of the decision process from one to the next (Nazarenko & Zaegayouna, 2009). In computing, Automatic Term Recognition, also known as Term Extraction, is a process of obtaining a set of actual words that are relevant to a given text. According to Cimiano (2006), 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. Terms can be made of one word (single-word) or a group of words (multi-word). Multi-word terms are believed to be less polysemous than single-word terms (Boulaknadel & Daille, 2008) and also form the majority of any ontology - approximately 85% of domain-specific terms (Nakagawa & Mori, 2002).

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 Qur'an, are limited to certain parts or a certain level of scope, or are even implemented using translations of the Qur'an, rather than the original Arabic text (Alrehaili & Atwell, 2014), for example, ontology made of time nouns by Al-Yahya and 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.

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, occurring only once or twice. This means that the major parts, which are multi-word terms, are excluded by statistical approaches. Hybrid approaches combine different methods from linguistics and statistics for detecting terms in the text.

The majority of Term Extraction research studies conducted on the Qu'ran have used translated text instead of the original text; 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 equivalence in other languages (Abobaker et al., 2012; Kashgary, 2011). Few attempts have been made to use Arabic for Qur'anic Term Extraction, and we found only three previously published studies: those conducted by Zaidi et al. (2012), Harrag et al. (2014) and Alhawarat (2015). Therefore, it is obvious that further research is required in this area. The current paper is organised as follows. Section 2 presents related studies on term extraction from the Qur'an; Section 3 describes our adapted approach to identifying terms from the Qur'an text; Section 4 presents our results and evaluations; and in Section 5 we draw our conclusion.

2. Related studies

In recent years, there has been an increase in Qur'anic knowledge representation. The focus here will be studies that have extracted the ontology of the Qur'an, particularly those that have used term extraction. One of the first suggestions for extracting knowledge from the Qur'an was made by Saad and Salim (2008). They based relevant term extraction on an unsupervised algorithm for extraction of key phrases, developed by Huang et al. (2006), using a clustering technique called hybrid particle swarm optimisers for forming concepts,

and followed the process proposed by Quan (2004). Saad and Salim (2008) continued their work on ontology extraction for the Islamic domain, and their approach was based on a particular translation format that includes round brackets to refer to the synonym or meaning, and square brackets for providing further explanation of a term. This approach was limited to a certain format of one of the English translations of the Qur'an.

Only three previous studies that examined the extraction of terms from the Arabic text of the Qur'an proposed hybrid methods that are based on syntactic patterns and TF-IDF for extracting single-terms (Zaidi et al., 2012; Harrag et al., 2014; Alhawarat, 2015). Their methods achieved 0.88 in terms of precision and 0.92 recall was obtained. However, they found some prefixes such as prepositions within the candidate terms, as they focused on the tokens instead of the segment when they collected the terms. A list of words from a lexicon provided by al-mishkat (2012) was used to compare the performance of their methods. This lexicon contained 17,622 single-words for the Qur'an, but these words were not as well segmented, which is why the methods described obtained high precision. For multi-word terms, they fed an open source software for text processing, namely, GATE, with a set of rules {*Noun Noun, Verb Noun, Noun Adjective and Noun Preposition Noun*}, and the precision obtained was 0.5. Some candidates for multi-words, such as '**الجاهل أغنياء**', were incorrectly considered as terms. The second noun here is an objective complement which should be connected as a predicate for the first noun instead of generating the nouns as one term.

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, was based on TF-IDF, which requires some specifications that are not available for the Qur'an, such as the size of the corpus. Alhawarat (2015) used a similar technique for extracting verses topics from the Qur'an.

Dukes et al. (2013) described the Qur'anic Arabic Corpus (QAC) project, an online Qur'an that was annotated at several levels, which included an ontology that defines 300 concepts in the Qur'an, and captures interrelationships using predicate logic. The number of relationships is 350, and the type of relationship between concepts is Part-of or Is-A. The ontology is based on the Tafsir by Ibn Kathir. The QAC also contains other analyses of the Qur'an text, such as POS, morphological analysis and dependency parse structure analysis.

Muhammad (2012) developed an ontology that encompassed the entire Qur'an in terms of pronoun tagging, whereby each pronoun is linked to its syntactic antecedent or previous reference. The dataset comprises 24,500 pronouns in the Arabic text, each linked to its antecedent. This can be used in ontology extraction in an ontology learning system using anaphora analysis to extract the concepts and relationships.

Mukhtar et al. (2012) produced a dataset that contains concepts from the second chapter of the Qur'an, known as the Vocabulary of Quranic Concepts. They used six different English translations of the Qur'an and applied a domain-independent tool called Termine from Franzi et al. (2000) to extract the concept. However, Termine was designed for the extraction of multi-word terms, while the Qur'an 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 Qur'an. 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 Qur'an. A wide range of Qur'an annotations is available in either computer-readable or hand-written formats. Our extraction method used multiple data sources, including the Tanzil Quran project (Zarrabi-Zaden, 2007), the QAC (Dukes et al., 2013) and the Qurany (Abbas, 2009). The source of the Qur'an text was downloaded from the Tanzil Quran project (Zarrabi-Zaden, 2007), which provided a digital copy of the Qur'an that has been manually validated by a group of experts against an accepted standard written-text version: Madinah Mushaf. The text of the Qur'an is stored in a text file, which is composed of 6,236 lines, and each line represents a verse in the Qur'an. Words, morphemes and POS information was collected from the QAC (Dukes et al., 2013), which was manually verified.

3.2 Term extraction

In order to generate a list of Quranic 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 Qur'an. We extracted approximately 470 nominal items with their syntactic patterns. Table 1 shows the first 10 lines of our pattern list, and Table 2 shows the most frequent patterns in the entire Qur'an text.

Table 1: An example of patterns of terms candidates

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 hopes to meet Allah
10	N	lqA'	meeting

We tagged the terms for every verse, and also annotated their sequences of syntactic patterns. The syntactic patterns in the second column include dots between every segment; this feature was required to extract complex terms, such as that in line no. 6 in Table 1, '**الذين يعملون السيئات** - those who do evil deeds.' We could not remove these pronouns and verbs from this term 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: The top 10 most frequently occurring noun phrase syntactic patterns in the Qur'an

No	Syntactic pattern	Occurrences
1	N	25136
2	DET.N	7488
3	PN	3911
4	REL V	2919
5	N N	2790
6	ADJ	1961
7	REL V.PRON	1885
8	N DET N	1347
9	N V.PRON	805
10	N PN	804

The outcome of this step a list of syntactic patterns for all of the term candidates in the Qur'an.

ii. Extracting term candidates.

In this step, our aim was to apply regular expressions to find 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. Part-of-speech sequences for 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 number of times that 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) because it is the only topic list that is available in computer-processable form for the Qur'an.

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

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 for every part of the term if it was longer than one word.
$W(t)$ is the total weight of a term.

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 weight using linguistic and statistical techniques for all candidates, we could then rank them in descending order, as illustrated in Table 3.

Table 3: 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 3 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 terms.

4.1 Evaluation against previous datasets

One previous dataset, namely the QurAna (Muhammad, 2012), primarily relied on pronouns mentioned in the Qur'an 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 (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 studies for chapter 29. In addition, we asked two independent annotators to identify the concepts from the same chapter. Table 4 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 4: 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 4 illustrates the total number of terms for previous datasets and manual annotations. Manual annotators identified more terms for a certain chapter of the Qur'an. 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.

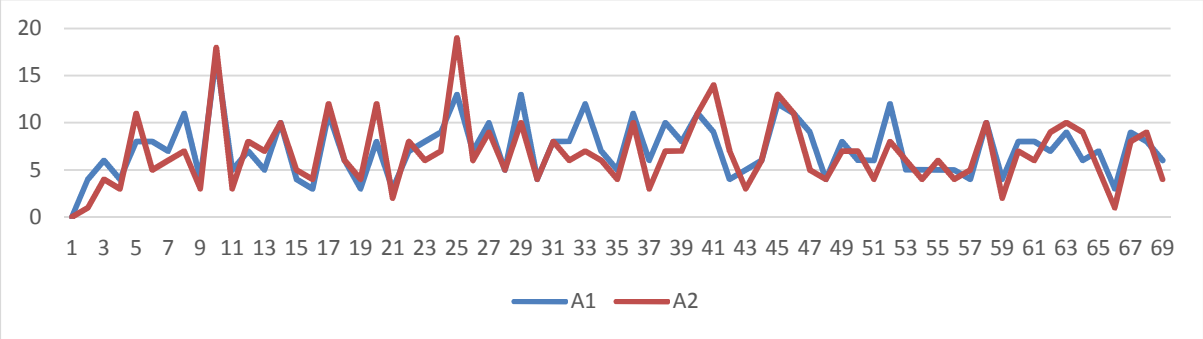

Figure 2. A comparison of hand-annotated terms

Figure 2 shows the agreement between the two annotators who were asked to annotate chapter 29 of the Qur'an. 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.

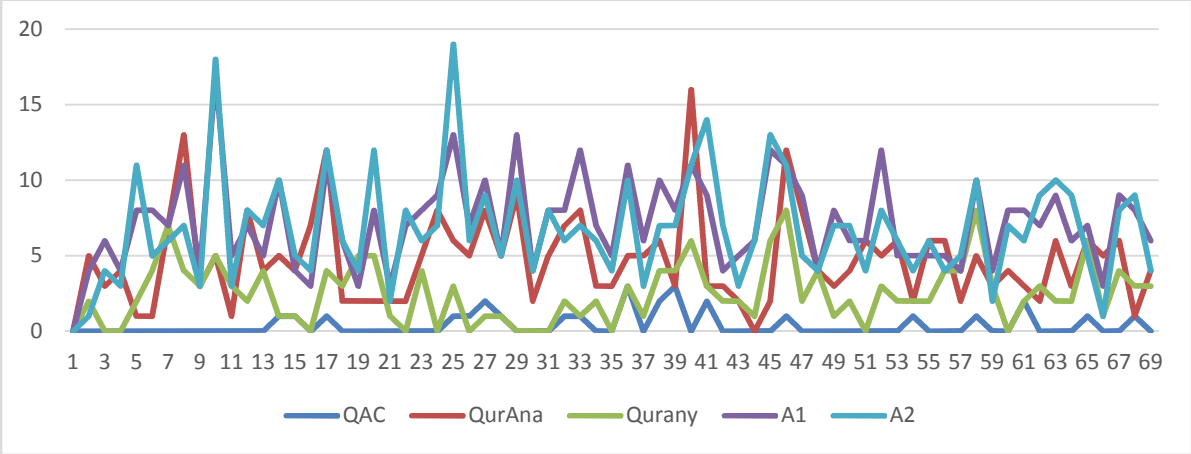

Figure 3. A comparison of hand-annotated and collected terms with those of previous related studies

This proves 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 until rank k over rank number (see equation no 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 Quranic 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 5: 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.

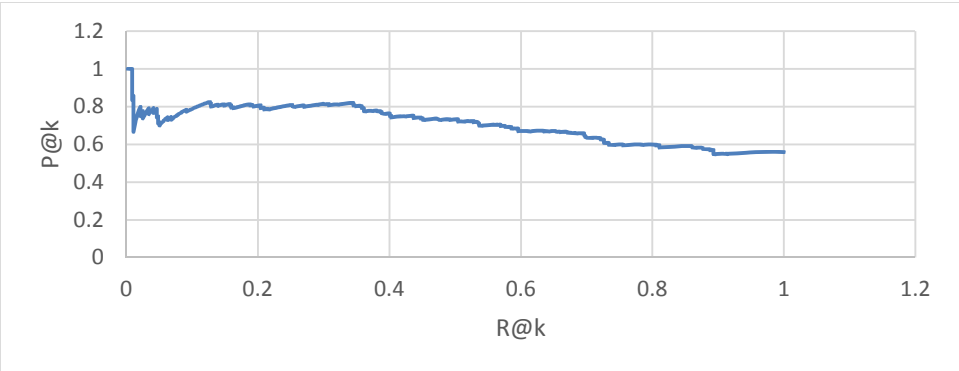

Figure 4. Recall-precision graph for the first 1,000 extracted terms

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

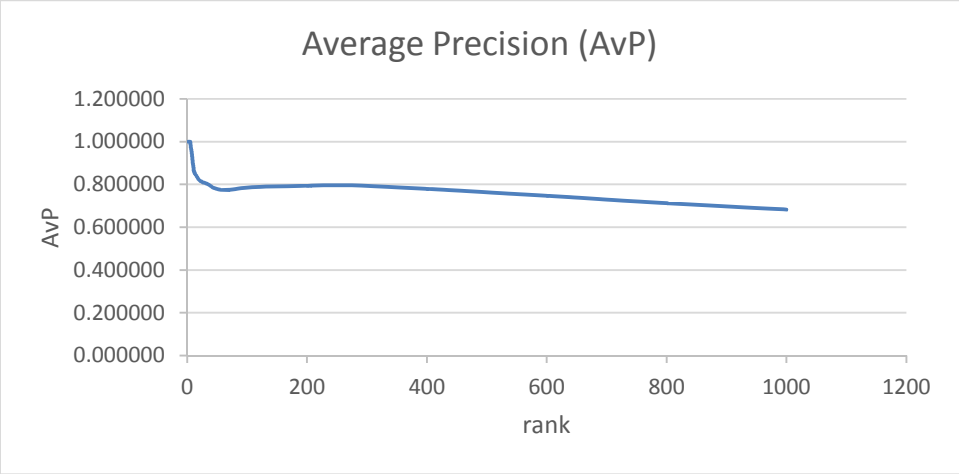

Figure 5. The stages in the terms extraction process

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 Qur'an as well as assessing these against three types of evaluation. The datasets from previous studies of the Qur'an is not complete, and is 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.

References

Abobaker Ali, M Alsaleh Brakhw, Munif Zarirruddin Fikri Bin Nordin and Sharifah Fazliyaton Shaik. Some Linguistic Di_culties in Translating the Holy Quran from Arabic into English. International Journal of Social Science and Humanity, 2(6):588{590, 2012.

Ab Muhammad. Annotation of conceptual co-reference and text Mining the Qur'an. Technical report, 2012. URL http://etheses.whiterose.ac.uk/id/eprint/4160.

al-mishkat, 2012. URL http://www.al-mishkat.com/words/.

Amira D Kashgary. The paradox of translating the untranslatable: Equivalence vs. non equivalence in translating from Arabic into English. Journal of King Saud University -

Languages and Translation, 23:47–57, 2011. doi: 10.1016/j.jksult.2010.03.001. URL www.ksu.edu.sa www.sciencedirect.com.

Chong Huang, Yonghong Tian, Zhi Zhou, Charles X Ling, and Tiejun Huang. Keyphrase Extraction using Semantic Networks Structure Analysis. In The sixth IEEE International Conference on Data Mining (ICDM 2006), pages 275–284, Hong Kong, 2006. IEEE press.

Christopher Norman. Technical Term Extraction Using Measures of Neology. PhD thesis, DEGREE PROJECT, IN , MASTER'S PROGRAMME, COMPUTER SCIENCE SECOND LEVEL STOCKHOLM, SWEDEN 2015 Technical Term Extraction Using Measures of Neology CHRISTOPHER NORMAN KTH ROYAL INSTITUTE OF TECHNOLOGY, 2015.

Cimiano, P. (2006). Ontology Learning and Population from Text: Algorithms, Evaluation and Applications Springer. November 2006.

Fouzi Harrag, Abdullah Al-Nasser, Abdullah Al-Musnad, Rayan Al-Shaya, and Salman Al-Salman. Using association rules for ontology extraction from a Quran corpus. 2014.

FrantziK.T., Ananiadou S., Tsujii, J. (1998).The C-value/NC-value method of Automatic Recognition for Multi-Word Terms. In Christos N. and Staphanidis C. (Eds.) Lecture Notes in Computer Science, LNCS 1513, Springer, 1998, pp. 585-604.

Hikmat Ullah Khan, Syed Muhammad Saqlain, Muhammad Shoaib, and Muhammad Sher. Ontology Based Semantic Search in Holy Quran. International Journal of Future Computer and Communication, 2(6): 570–575, 2013. ISSN 20103751. doi: 10.7763/IJFCC.2013.V2.229. URL http://www.ijfcc.org/index.php?m=content&c=index&a=show&catid=43&id=493.

Hiroshi Nakagawa and Tatsunori Mori. A Simple but Powerful Automatic Term Extraction Method. In COLING-02 on COMPUTERM, COMPUTERM '02, page 1, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics. doi: 10.3115/1118771.1118778. URL http://dx.doi.org/10.3115/1118771.1118778.

K Dukes. Statistical Parsing by Machine Learning from a Classical Arabic Treebank. Technical report, 2013. URL http://www.kaisdukes.com/papers/thesis-dukes2013.pdf\nAll Papers/D/Dukes 2013 - Statistical Parsing by Machine Learning from a Classical Arabic Treebank.pdf.

M Al-Yahya and H Al-Khalifa. 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):21–35, 2010. ISSN 13198025. URL 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.

Mohammad Alhawarat. Extracting Topics from the Holy Quran Using Generative Models. International Journal of Advanced Computer Science and Applications(ijacsa) - See more at:http://thesai.org/Publications/ViewPaper?Volume=6&Issue=12&Code=ijacsa&SerialNo=38#sthash.7kmJYsB9.dpuf, 6(12), 2015.

Mohammed G.H. Al Zamil and Qasem Al-Radaideh. Automatic extraction of ontological relations from Arabic text. Journal of King Saud University - Computer and Information Sciences, 26(4):462– 472, 2014. ISSN 13191578. doi: 10.1016/j.jksuci.2014.06.007. URL http://www.sciencedirect.com/science/article/pii/S1319157814000317.

Nh Abbas. Quran'search for a Concept'Tool and Website. Unpublished Dissertation, Leed, 2009. URL http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Quran+?Search+for+a+Concept?+Tool+and+Website#0.

Philipp Cimiano. Ontology learning and population from text: Algorithms, evaluation and applications. 2006. ISBN 0387306323. doi: 10.1007/978-0-387-39252-3. URL http://medcontent.metapress.com/index/A65RM03P4874243N.pdf.

Saidah Saad and Naomie Salim. Methodology of Ontology Extraction for Islamic Knowledge Text. In Postgraduate Annual Research Seminar, number i, 2008.

Sameer M Alrehaili and Eric Atwell. 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, pages 19–23, Reykjavik, Iceland. ISBN 2951740883. URL http://www.lrec-conf.org/proceedings/lrec2014/workshops/LREC2014Workshop-LRE-Rel2 Proceedings.pdf.

Siham Boulaknadel and Beatrice Daille. A multi-word term extraction program for Arabic language. In Language Resources and Evaluation Conference, pages 3–6, 2008. ISBN 2-9517408-4-0. URL http://pages.cs.brandeis.edu/ marc/misc/proceedings/lrec-2008/pdf/378 paper.pdf.

Soyara Zaidi, Ahmed Abdelali, Fatiha Sadat, and Mohamed-Tayeb Laskri. Hybrid Approach for Extracting Collocations from Arabic Quran Texts. 2012.

Tayyeba Mukhtar, Hammad Afzal, and Awais Majeed. Vocabulary of Quranic concepts: A semi-automatically created terminology of Holy Quran. In 2012 15th International Multitopic Conference, INMIC 2012, 2012. ISBN 9781467322508. doi: 10.1109/INMIC.2012.6511467.

Thanh Tho Quan, Siu Cheung Hui, and Tru Hoang Cao. FOGA : A Fuzzy Ontology Generation Framework for Scholarly Semantic Web. In Knowledge Discovery and Ontologies (KDO-2004), pages 37–48, Pisa, Italy, 2004. URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.91.7752&rep=rep1&type=pdf. 2

Yong Bin Kang, Pari Delir Haghighi, and Frada Burstein. CFinder: An intelligent key concept finder from text for ontology development. Expert Systems with Applications, 2014. ISSN 09574174. doi: 10.1016/j.eswa.2014.01.006.