Concept Extraction on Quranic Translation Text

Saidah Saad[1,a], Naomie Salim[2,b], Sabrina Tiun[1,c]
[1]Schoolof Information Science, Universiti Kebangsaan Malaysia , Malaysia
[2]School of Computing, Universiti Teknologi Malaysia, Malaysia
asaidah@ftsm.ukm.my, [b]naomie@utm.my, [c]sabrinatiun@ftsm.ukm.my

ABSTRACT
The Semantic knowledge that based on ontology learning technology introduced is to reduce the overall time of ontology construction. Ontology construction process includes several aspects and layers, and extraction domain concept is one of the most important aspects. This step becomes a prerequisite process in developing ontology and also becomes a seed to the next step. In this paper, we carried out several experiments based on linguistics, statistical and hybrid approaches in order to identify which are the best techniques and approaches to extract terms from Quranic translation text. For linguistic approach, we used POS pattern, for statistical approaches, we choose seven frequency-based as the techniques to choose frequency terms and for hybrid, we combined both linguistic and statistical approaches. The results obtained show that the hybrid approach is the best in identifying and filtering relevant concepts in Quranic domain corpus.

***Keywords*: concept extraction, term extraction, statistical and machine learning techniques, Quranic translation text**

1. Introduction

Term extraction (TE) is a fundamental processing step preceding more complex tasks such as semantic search and ontology learning, as mentioned in ontology layer cake (Cimiano, 2006). Term extraction is a process of recognizing the important entities in a specific domain and it often relies on domain-specific knowledge in order to upgrade and improve the system performance.

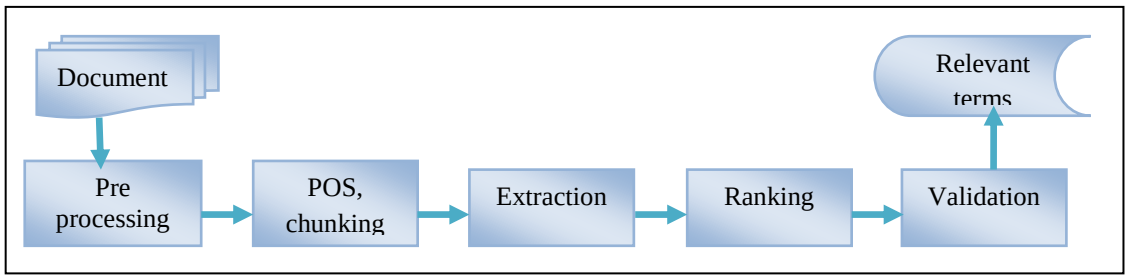

Fig 1. Stages in Terms Extraction Process

In general, terms are considered as words used in domain-specific contexts. More specific and purposeful interpretations of terms do exist for certain applications such as ontology learning. There are two types of terms, namely, simple terms (single-word terms or unithood)

and complex terms (multi-word terms or termhood). Collectively, terms constitute to what is known as terminology. In order to extract these two types of terms, normally, there are two types of properties used, which are terms based on linguistic properties and terms based on statistical properties. For extracting a relevant term or terms, there are several stages that need to be done as shown in Fig.1.

In the pre-processing stage, a phrase that needs to be parsed must undergo a definition temporary replacement process to simplify the ongoing parsing process. For example: *Those who believe* or *You who believe* is replaced with Mu'minun and the word 'He', 'His', 'Our' and 'We' (where the first characters are in capital letters) will be replaced by 'Allah'. This is obtained from the list of predetermined REPLACEMENT. The second stage is to identify the part of speech (POS) of the term in order to proceed to the next stage. In the extraction stage, all the identified patterns that match with the POS are extracted and the weight of every term is calculated. The ranking process will take place in the next stage, which is then followed by the validation process before the term can be justified as a relevant term.

This paper investigates NLP techniques for term extraction as a candidate concept, using a combination of rule-based approaches and machine learning. A method for term recognition using linguistic and statistical techniques is described, making use of contextual information to bootstrap learning. Thus, in section 2, which is the research method, we divide it into three main sections which will be describing the three main approaches in TE, and they are:(i) TE based on linguistic properties, (ii) TE based on statistical properties, and lastly, (ii) TE based on the hybrid of statistical and linguistic properties. The analysis and results of those TE will be discussed in the same section as well. Finally, we then conclude our paper in section 3.

2. Research Method: TE Approaches, Analysis and Results

In this section, we will explain in details how each of the approaches we have mentioned before in extraction terms for concepts candidates. Starting with TE based on linguistic properties, follows by statistical properties and finally, the hybrid approach.

2.1 TE based on Linguistic Properties

Majority of researchers who study automatic TE use the linguistic approach such as POS tagger and phrase chunking for extracting the noun and noun phrase (NP), in which, most of the terms tend to be nominal. This approach can also be used to filter out the stop words and to carry out the stemming process. In extracting term, NP is given more attention, whereas the verbs, and adjectives can be domain-specific.

The TE using linguistic processing is based on three steps: (i) first, the candidate terms are extracted using identified pattern such as:

$$((Adj|Noun)^{+}|((Adj|Noun)^{*}(Noun)^{?})(Adj|Noun)^{*})Noun$$

Thus, the longest NP is extracted. (ii) Secondly, each of the candidate terms is separated when conjunctions and other operators such as 'and', 'or', '(', ')' and ',' are found. (iiI) In the

last step, all the candidate terms undergo the stemming process to remove the affixes and also the stop words which occur in the phrases.

In order to identify the variations in term (defined as "an utterance which is semantically and conceptually related to an original term"(Daille et al., 1996)), such as *Tahajjud optional prayer Nawafil* as a variant to the phrase *optional prayer Nawafil* and *optional prayer*, the variation can be identified base on the pattern in Table 1 below.

Table 1: Patterns for NP extraction

Pattern	Meaning
$N_0(N_n)$	N_n is the other variant of term (Where N = noun and A = adjective)
N_n of N_0	
A N	

These variations are based on the rule that content words are based on two words; and the length of word that is greater than two are formed on the basis of base word through (Daille et. al, 1994): (i) Over composition - *Islamic Monotheism Hanifa* :*Islamic +Monotheism Hanifa→ Islamic Monotheism Hanifa*, (ii) Modification insertion of modifying adjectives or adverbs, or post-modification; e.g. *evil deed→evil wicked deed*, which in English is an insertion. (ii) Coordination e.g. *evil deeds / evil sins→evil deeds* and *sins.*

2.2 TE based on Statistical Properties

Statistical calculation provides great help in ranking all the extracted candidate terms based on a criterion that is able to distinguish between the most relevant and less relevant terms. Terms with the highest score tend to be more relevant compared to the lower score. The basic statistical approach is by using frequency in a corpus. In order to evaluate term using statistical approach, there are two types of properties that can be used, either termhood or unithood as mentioned in Kageura & Umino (1996), and Wong et al (2008). Termhood is a linguistic unit that is related to domain specific concepts which is a peculiar characteristic of terms and single word, whereas unithood is the syntagmatic linguistic unit where by definition, it characterizes complex linguistic units called collocations which are composed by words with strong association, such as compound words, idiomatic and complex terms. Unithood is significant only for multiword terms (Pazienza et al. 2005).

Our experiments on TE based on statistical approaches are divided into two approaches: keyword-based and N-gram approach.

2.2.1 Keyword Approach

The keyword-based approach measures the termhood circumscribed to frequency-based approaches and the use of reference corpora. Such approaches are like the classic TD-IDF (Salton & Buckle, 1987; Medelyan&Witten, 2006), the Average Term Frequency or also known as AveTF (Zhang et.al, 2008) and Residual IDF or known as RIDF (Church&Gale,

1995; Rennie, 2005).

We used the above approaches on our Quranic document dataset and compared how those approaches performed. First, we retrieved terms (see Fig 1) are ranked based on their weight decreasing order. Next, the recall and precision values are calculated from the ranked term listing. We used the he standard recall $\mathcal{R}$ and precision $\mathcal{P}$ are defined as in Eq. 1and Eq. 2, respectively (Salton & McGill, 1983) for the evaluation.

$$\mathcal{R} = \frac{\text{number of items retrieved and relevant}}{\text{total relevant in collection}} \quad (1)$$

$$\mathcal{P} = \frac{\text{number of items retrieved and relevant}}{\text{total retrieved from collection}} \quad (2)$$

Figure 2 illustrates the recall-precision graphs pertaining to these average recall and precision values for four different types of term extraction statistical techniques for text representation.

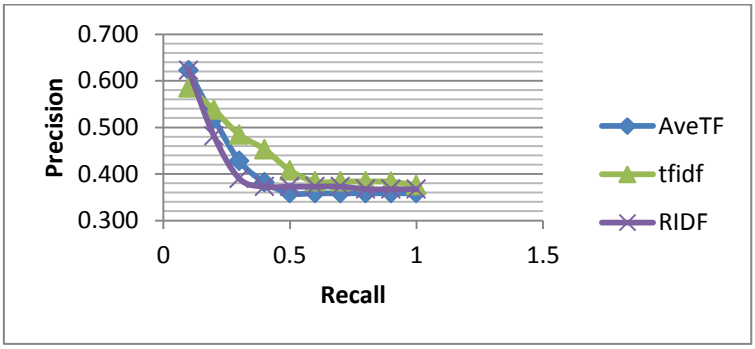

Fig 2. Average Recall-Precision Termhood Statistical Techniques for Text Representation

From the Fig. 2, AveTF performs better at the first recall. The other two remaining techniques also show the same results, but slightly lower than AveTF. For the overall performance, TF-IDF performs better with an average precision of 43.8%. The results in Figure 2 indicates that: (i)TF-IDF performed better than the other statistical techniques and thus, it is the best option to retrieve relevant terms from Quranic translation document compared to AveTF and RIDF, and (ii) termhood alone cannot give a good result for TE to retrieve more relevant candidate terms.

2.2.2 N-gram Approach

By definition, the word N-gram is a vector of n words where each word is indexed by the precision values that separate it from its associated pivot word. Consequently, an N-gram can be contiguous or non-contiguous, regardless of whether the words involved in the N-gram represent or do not represent a continuous sequence of words in the corpus.

In our experiment, we extract the unithood where n <=5. The maximum n=5 is based on the corpus where the maximum length of terms that were identified by the expert in the Quranic translation text were 5. We measure the unithood based on N-gram using the frequency-based approaches of; TF, TFIDF, AveTF, and RIDF. We also investigate the

frequency-based measures of weirdness, C-value and Glossex method. **C-Value** is a weight that calculates the smallest unit size of a word, which is bigrams. This means that C-Value is not applicable to single word-unit term extraction. The nested terms are important in this calculation (Frantzi et al, 1998). In the case of non-nested terms, the C-value takes into account the length of the term candidate and the number of occurrences (see Eq. 3).

$$C-value(a) = \begin{cases} log_2|a| \cdot f(a) & if\ a\ is\ not\ nested \\ log_2|a| \cdot \left(f(a) - \frac{1}{|T_a|}\sum_{b \in T_a} f(b)\right) & otherwise \end{cases} \quad (3)$$

In Eq

. 3 above, where a and b are the candidate terms, f is a frequency of candidate terms and $T_a$is set of candidate terms that contain a. **Weirdness** (see Eq. 4) is a weight that is based on the idea of the distribution of terms within a specialized corpus and general corpus which significantly differ. The equation expressed is as follows; where $f_s(i)$is a frequency of word i in specialized corpus (domain) and $f_g(i)$in general corpus (background knowledge); $n_s$and $n_g$is the total number of words in the respective corpora.

$$Weirdness(i) = \frac{\frac{f_s(i)}{n_s}}{\frac{f_g(i)}{n_g}} \quad (4)$$

Glossex is a weight that is based on two heuristics. The first measures or calculates the domain specificity (TD - same as weirdness), while the second is the idea of term cohesion where $|t| = n$ is a number of words forming term t. This equation, Eq. 5.2, is as follows:

$$TC_{D_i}(t) = \frac{n.tf_{tD_i} \cdot log\ tf_{tD_i}}{\sum_{j=0}^{n} tf_{w_j D_i}} \quad (5.1)$$

$$Glossext = \alpha \cdot TD(t) + \beta \cdot TC(t); \quad (5.2)$$

In order to investigate whether the pure statistical techniques that are based on N-gram do influence the performance results, the additional ranked-cutoff procedure at position recall 10% to 100% is employed. The recall-precision graphs pertaining to these average recall and precision values for the seven different techniques are shown in Figure 3.

From Figure 3, one can observe that Glossex performs the best among the seven techniques in identifying the relevant terms of the domain using the N-gram approach. Figure 3 reveals that there is a performance difference between Glossex and TF. The performance of TF-IDF is nearly the same as TF. C-Value performs the worst by achieving an average precision of only 12.3% respectively.

The results in the Figure 3, indicates the following: (i) Both Glossex and TF perform and show better results than other statistical approaches. On average, they are able to retrieve relevant terms from the Quranic translation text corpus. While the Glossex achieves an average precision of 24.5%, the latter, TF, is slightly less with an average precision value of 23.3%. (ii) TermEx shows the lowest performance among the statistical approaches in

extracting the relevant terms with an average precision of 9.3%, and (iii) Using the N-gram techniques or statistical techniques can only cause data sparseness where they increase the recall value and decrease the precision value. Any data will be extracted and not focusing on the concept formation.

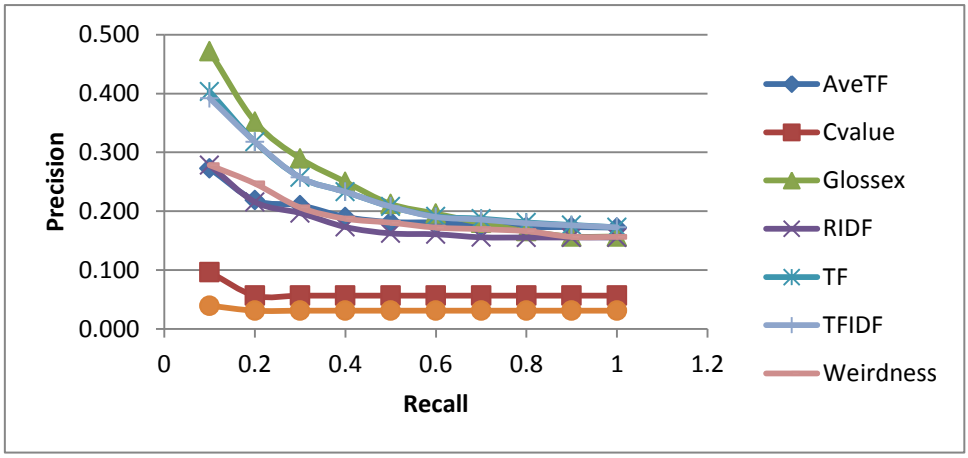

Fig 3. Average Recall-Precision graph for the seven types of unithood statistical techniques for text representation

2.3 TE based on Hybrid Approach

The trend in recent research is to use hybrid approach, in which linguistic patterns and statistical properties are combined to produce a unified indicator. In our approach, the linguistic analysis is carried out before the application of statistical properties. The TE based on linguistic patterns are used to select the most likely to be the concepts candidates, then statistical properties will be used to rank those concept candidates according to a specific measure.

The first approach of this hybrid technique is by; (i) using the open NLP to extract normal noun phrases in a corpus as a candidate terms, and (ii) these extracted NP or candidates are selected or filtered based on different statistical method (see section 2.3).

In this hybrid approach or we call it as NL approach, we also propose NP extraction, in which, prepositional phrases are also included as part of the extracted term. To be precise, in the extraction of NP, the candidate terms are extracted based on the following pattern where the prepositional phrases are also included as part of the extracted term.

$((Adj|Noun)^{+}|((Adj|Noun)^{*}(Noun|Prep)^{?})(Adj|Noun)^{*})Noun$

We called the hybrid approach as the NL approach. This approach uses the Stanford parser for syntactic pattern-extraction and ranks by term frequency (this is based on the previous approach which shows that the term frequency technique is among the best approaches in identifying the relevant candidate term). This also involves another preprocessing of the document. The text is filtered and all the pronouns with capital letters such as 'We', 'Us', 'He', 'Lord', 'My' into ‘Allah’ and 'O you who believe' are converted into ‘Mu’minun’ to increase the precision value. This approach also combines the termhood and unithood extraction. This is carried out in order to identify the variation in terms as mentioned before, conveying the

meaning of the term augmented with other specific semantic information. This variation can be identified based on the pattern shown in Table 2.

Table 2: Variation of patterns for NP extraction

Pattern	Meaning
$N_0(N_n)$	N_n is the other variant of term (Where N = noun and A = adjective)
N_n of N_0	
A N	
N_0 P N_n	N_0 and N_n is the other variant of term (where P = prepositional term N_0 and N_n can be defined as above pattern)

Experiment is conducted in order to evaluate the performance of NL technique with the other techniques. The results of this experiment are analyzed using recall and precision as a measure of the performance effectiveness and efficiency of the extracted terms.

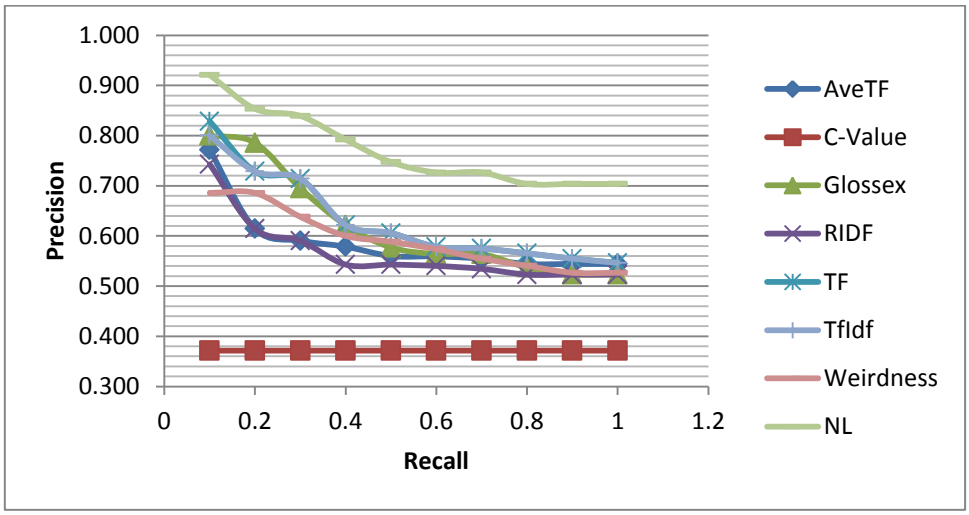

Fig 4. Average Recall-Precision Graph for the Eight Types of Hybrid Approach for Term Extraction

Once the extraction has been performed according to what was mentioned before, the results are then compared with the expert-defined relevant term towards the eight different lists of the results. Figure 4 shows the average recall and precision values for all eight techniques performed.

From Figure 4, we can see that NL (based on the new style of extraction) performs much better than other techniques at all recall levels. There is an improvement of 140% from NL to the second best technique and 197% to the overall techniques. These results clearly indicate that the NL approach is able to retrieve a higher percentage of relevant terms from a document compared to others.

When these results are compared to the result of normal natural language approach (see section 2.2) and statistical approaches only (see section 2.3), the followings can be observed:

(i) The hybrid approach which combines termhood, unithood and statistical will perform much better than single approaches. (ii) The modification on pattern extraction and the implementation of Stanford parser on syntactical analysis, performs much better than other techniques and approaches by 77% at all levels of recall; (iii) For Quranic translation text corpus, the TF, TF-IDF and Glossex of statistical techniques give better results in all techniques and approaches applied; (iv) The unithood which applies the N-gram approach (based on statistical analysis only) performs the worst compared to others, where it is able to achieve only an average precision of 24%.

3. Conclusion

Several techniques and approaches have been presented in this study to tackle the problem of extracting concepts from the corpus. This step becomes a prerequisite process in developing ontology and also becomes a seed to the next step. A few experiments have been proposed in order to identify the best techniques and approaches for extracting the texts from Quranic translation texts as domain knowledge. The work presented in this section thus represents a significant contribution of the state-of-the-art aspect in the field of term relevant extraction based on Quranic translation texts.

In this paper of concept or instances extraction, 3 different types of approaches have been used: (i) First, the linguistic approaches; linguistic approaches are based on linguistic information in the form of annotation sets. The necessary information is retrieved according to the linguistic rules. In most cases, linguistic rules need to be constructed manually by a domain expert. Extraction can be based on shallow linguistic information, namely the morphosyntactic, derivational and compound analysis, including POS classification or deeper information, for example phrases, chunk analysis and clauses. Most of these approaches result in a rather high efficiency, but their major drawbacks are the amount of supervision and huge effort required for handcrafting linguistic rules. (ii) Second, the statistical approaches: Frequency-based methods are fairly simple to implement and evaluate. They produce reasonable results but do not allow for fine-grained control over the extraction process. (iii) Third, the hybrid approaches: the combination of both techniques mentioned above. The linguistic approaches act as a filter for extracting relevant terms which will become a candidate of a concept and instances. Statistical approaches are used to rank the relevant terms in order to identify which term is the most relevant among the others. The hybrid approach is the best in identifying and filtering relevant concepts in this domain corpus. Term frequency (TF) and term frequency-inverse document frequency (TF-IDF) show the best results on extracting those concepts.

References

Church, K. and Gale, W. (1995). Inverse Document Frequency (IDF): a measure of deviations from Poisson, in D. Yarowsky and K. Church (Eds), Third Workshop on very large corpora, ACL, MIT, pp. 121–130

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

Daille B., Gaussier E., Lange J, (1994). Towards Automatic Extraction of Mono lingual and Bilingual Terminology. Proceeding of COLING 94. 515-524.

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.

Gerard, S. and Chris, B. (1987). Term Weighting Approaches in Automatic Text Retrieval. Technical Report.Cornell University, Ithaca, NY, USA.

Kageura, K, and Umino, B. (1996). Methods of automatic term recognition: a review. Terminology 3(2):2590–289.

Medelyan, O., Witten, Ian H., Thesaurus based automatic keyphrase indexing, Digital Libraries, 2006. JCDL '06. Proceedings of the 6th ACM/IEEE-CS Joint Conference on , vol., no., pp.296,297, June 2006

Pazienza, M. T, Pennacchiotti, M., Zanzotto F M. (2005). Terminology extraction: an analysis of linguistic and statistical approaches. In: S. Sirmakessis (ed.) Knowledge Mining. Series: Studies in Fuzziness and Soft Computing, Vol.185, Springer Verlag.

Rennie J. (2005). Using term informativeness for named entity detection. In Proceeding SIGIR '05 Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval.Pages 353-360.

Salton, G. and McGill, M. J., (1983). Introduction to Modern Information Retrieval. McGraw-Hill. Printed in New York.

Wong, W., Liu, W., Bennamoun, M. (2008).Determination of unithood and termhood for term recognition. In: Handbook of research on text and web mining technologies. IGI Global (2008).

Zhang, Z., Iria, J., Brewster, C., Ciravegna, F. (2008).A Comparative Evaluation of Term Recognition Algorithms. In Proceedings of The sixth international conference on Language Resources and Evaluation, (LREC 2008), May 28-31, 2008, Marrakech, Morocco.