Journal of Information Engineering and Applications
ISSN 2224-5782 (print) ISSN 2225-0506 (online)
Vol.4, No.6, 2014


# Normalized Google Distance for Collocation Extraction from Islamic Domain

Hamida Ali Mohamad Salem [1*] Masnizah Mohd [2]

1. Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, Bangi, Selangor 43600, Malaysia
2. Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, Bangi, Selangor 43600, Malaysia

* E-mail of the corresponding author: ibrahim.alhares@yahoo.com

**Abstract**

This study investigates the properties of Arabic collocations, and classifies them according to their structural patterns on Islamic domain. Based on linguistic information, the patterns and the variation of the collocations have been identified. Then, a system that extracts the collocations from Islamic domain based on statistical measures has been described. In candidate ranking, the normalized Google distance has been adapted to measure the associations between the words in the candidates set. Finally, the n-best evaluation that selects n-best lists for each association measure has been used to annotate all candidates in these lists manually. The following association measures (log-likelihood ratio, t-score, mutual information, and enhanced mutual information) have been utilized in the candidate ranking step to compare these measures with the normalized Google distance in Arabic collocation extraction. In the experiment of this work, the normalized Google distance achieved the highest precision value 93% compared with other association measures. In fact, this strengthens our motivation to utilize the normalized Google distance to measure the relatedness between the constituent words of the collocations instead of using the frequency-based association measures as in the state-of-the-art methods.

**Keywords:** normalized Google distance, collocation extraction, Islamic domain

## 1. Introduction

The collocations issue is the linguistic phenomenon that is found in all human languages. It is an important part of many applications, such as machine translation, information retrieval, word sense disambiguation and lexicography. In a bilingual context, collocations are very important for learners of a language to construct meaningful sentences. Usage of the right combinations, being a part of context, generally results in correct language production (speech) at least at the stylistic level.

There is no widely accepted definition of a collocation in the field of computational linguistics. For example, Evert (2004) defined collocation as "a word combination that semantic and/or syntactic properties cannot be fully predicted from those of its components, which therefore has to be listed in a lexicon. Smadja (1993) considered the collocations as "recurrent combinations of words that co-occur more often than expected by chance and that correspond to arbitrary word usages". According to Pecina (2010), there are some restrictions (semantic and/or pragmatic) that must be included in the extraction of collocations in order to produce meaningful and fluent collocations.

The semantic compositionality is to check whether the overall meaning of the collocation is obtained by the composition of the meanings of individual words at the individual word level. In its simple definition, the collocation is defined as the two or more words which appear together and always seem as 'friends'. The collocation is the phenomenon of linguistic high productivity that makes for two words or more, in the confluence of what, attached to each other, combined permanently and does not change because the usage of a particular word. For instance, a noun has a small number of verbs or adjectives that can be combined with this noun to construct a collocation. For example, in English, the noun 'crime' has small number of verbs which may be combined with this noun to indicate the event of 'doing the crime'. The same can apply for an adjective and a verb. There are two verbs 'commit' or 'perpetrate' which can combine with this noun to indicate the action.

Furthermore, this case can be applied in the Islamic domain. If we take the noun 'الإيفاضه' (Ifaadah) in mind, the verbs 'يطوف' (cruising) or 'يمشي' (walking) can be combined with it. The verb 'يسير' (walking) can be used to denote the action, but the expression will be bad. On the other hand, the noun may need an adjective to describe it and constitute the collocation. For example, in English, a suitable adjective that can combine with the noun tea is 'strong'; this noun cannot combine with other adjective, say, 'powerful'. The same situation in Islamic domain;

 

with the noun 'الكعبه'(kaaba) one can combine a limited number of adjectives, such as 'الشريفه' (honorable). This work focuses primarily on collocation within the Islamic domain, which is very important to people who do not know the Arabic language.

## 2. Related Work

This section presents the existing multiword lexical unit extraction works that depend on the linguistic methods, statistical methods, or hybrid methods, and shows the some recent works for extraction some classifications of multiword lexical units in Arabic. It discusses in details the recent methods (Attia 2006; Attia et al. 2010; Boulaknadel et al. 2008; Bounhas & Slimani 2009).

The extraction of MWs in Arabic is a difficult process, because Arabic is a highly complex and ambiguous language. However, some recent works have analyzed this problem by using the linguistic methods or statistical methods. For example, Attia (2006) has presented the semi-automatic method for handling the multiword expressions that is based on lexicon of MWEs constructed manually. He built the MWE transducer in order to complement the morphological transducer, and to interact with other processing and preprocessing components. This transducer is called as a specialized two-sided transducer that uses the finite state regular expression to provide correct analysis on the lexical side and correct generation on the surface side.

However, it can only handle two types of MWEs: the fixed and semi-fixed expressions. The semi-fixed expression is the string that undergoes the morphological variations. Attia (2005) has used the core morphological transducer to obtain all possible forms of certain words in order to handle the morphological flexibility. Nevertheless, his method only focused on generating the compound nouns. The formula for compound nouns in Arabic has been defined as follows: NP [_Compound] -> [N N* A*] & ~N. The structures of MWEs are described as trees that can be parsed to identify the role of each combination. However, this method cannot handle some types of MWEs, such as, verb-particle constructions, and the compound nouns that allow external elements to intervene between the components. It also cannot handle the syntactically-flexible expressions. Additionally, the relevance of the extracted candidates is not computed because the method depends on pure linguistic method and does not use the statistical measures.

Boulaknadel et al. (2008) have designed a multi-word term extraction program for Arabic language. They have used a hybrid method to extract multi-word terminology from Arabic corpus. From linguistic respective, they have used some linguistic information to extract and filter the candidates of multiword terminology. Their method uses the part-of-speech tagging of the corpus that has been assigned by the Diab's tagger (Diab et al. 2004), to be used in the linguistic filter. The linguistic filter identifies the Arabic MWT patterns such as, N ADJ, N N, and N PREP N. In addition, their method takes into account the MWT variations, such as, graphical variants (the graphic alternations between the letters "ha'a" and "Ta'a marbutah"), inflectional variants (the number inflection of nouns, the number and gender inflections of adjectives, and the definite article "ال"), morphosyntactic variants (the synonymy relationship between two MWTs of different structures.), and syntactic variants(the modifications of the internal structure of the base-term, without affecting the grammatical categories of the main item which remain identical). On the other hand, they have used four association measures: log-likelihood ratio (LLR) (Dunning 1993), FLR (Nakagawa & Mori 2003), mutual information (MI3), and t-scores (Church et al. 1990) to order the candidates of MWT. They have reported that, log-likelihood ratio is the best association measure which achieved the highest precision of 85%.

However, this method has been criticized by Bounhas and Slimani (2009) for many reasons. The first reason is that, the POS tagger without morphological step is unable to segment the word (noun or adjective) into its components, such as, affixes, conjunctions, and some prepositions. Furthermore, the POS tagging does not allow take into account many features (gender and number of the component MWT), while defining MWT patterns. The second reason states that, this method is unable to deal with syntactic ambiguities. The final reason, this method does not allow recognizing the internal structure of MWTs. Bounhas and Slimani (2009) have presented a hybrid approach to extract multi-word terminology from Arabic specialized corpora. They have used several tools to identify and extract the compound nouns. Their method used the Arabic morphological analyzer (AraMorph) that has been developed by Hajic et al. (2005). The AraMorph have been used to compute the morphological features required for the syntactic rules.

Similarly to the method of Boulaknadel et al. (2008), this method also uses the part-of-speech tagging of the corpus that has been assigned by the Diab's tagger (Diab et al. 2004) in the linguistic filter. This tool assigns to each word a POS tag based on its context. The context is a window of -2/+2 words centered at the focus token. They have developed the Morpho-POS matcher to integrate the AraMorph and POS tagger. This matcher is used

 

to reduce the morphological ambiguity, by studying the context of each word. The sequence identifier in this method is used to identify the multi-words from sequences and compute the frequency for each entry in the corpus. The entry contains the POS and the morpho-syntactic features that are obtained from the previous tools. The third tool is the syntactic parser that parses the sequence of tokens each represented by a list of solutions.

This tool uses both the syntactic rules, based on the POS and the morphological features, to recognize compound nouns. It identifies sub-sequences that fulfill rules constraints which means that, many ambiguous words in the sequence are ignored. However, it may return more than one parse trees for a sub-sequence. Therefore, they have used the statistical measures to identify the best solution and resolve the ambiguities of the sub-sequence that has more than one parse trees. Their approach only used the log-likelihood ratio that computes the correlation between two terms. They have reported that, the precision value for bi-gram candidates equals to 93%, and is better than the precision value gained by Boulaknadel et al. (2008), who used the same corpus and evaluation method. However, this method is limited for extracting the compound nouns and ignored others types of multi-word lexical units such as verb-particle constrictions, prepositional phrases, and Arabic collocations.

The automatic extraction of Arabic multiword expression has been presented by Attia et al. (2010), to extract and validate the MWEs from Arabic corpora. This expression has used three complementary approaches to extract Arabic MWEs from different corpora. The first approach is the cross-lingual correspondence asymmetry that extracts the MWEs from the Arabic Wikipedia (AWK), which is based on semantic non-decomposable MWE. It has generated all candidates of AWK multiword titles and excluded the titles of disambiguation and administrative pages. After that, all candidates are classified as a MWE (if its translation is a single-word in any of the target languages, or it is found in any WordNet or MINELex), or non-MWE (otherwise). The MINELex is a multilingual lexicon of named entities.

The second approach is the translation-based approach that assumes that, automatic translation of MWEs collected from Princeton WordNet (PWN) into Arabic are high likelihood MWE candidates that which need to be automatically checked and validated. It is a bilingual that is based on English MWEs to extract the Arabic MWEs. From PWN, it has collected the English MWEs and translated them by using an off the-shelf SMT system, namely Google Translate. After that, the extracted candidates are evaluated automatically by using the gold standard PWN.MWEs which are found in English Wikipedia and have a correspondence in Arabic. This approach gives 13,656 real MWEs from the list of 60,292 translations.

The third approach is the corpus-based approach that extracts MWEs from a large raw corpus, relying on statistical measures and POS-annotation filtering. It uses the Arabic Gigaword Fourth Edition corpus. It is a hybrid method that uses the linguistic knowledge and association measures to extract the MWE. From the corpus, this method generates the unigram, bigram, and trigram candidates with their frequency. It used only two association measures: pointwise mutual information, and chi-square to order the candidates of MWE. From linguistic perspective, it has used the lemmatization using MADA (Habash et al. 2009) to collapse all variant forms together and thus created a more meaningful list of candidates. In this approach, they have reported that, the MI is the best association measure for bi-gram candidates that achieved the precision value equals to 71%, but the chi-square is the best association measure for tri-gram candidates with precision value equals to 63%. Saif & Aziz (2011) have presented an Automatic Collocation Extraction from Arabic Corpus, and they have used the hybrid method for extracting the collocation from Arabic corpus. Saif (2011) has presented an Automatic Multiword Lexical Unit Extraction from an Arabic Corpus; he has used four types of association measures: log-1ikelihood ratio, Mutual Information, enhanced mutual information, chi-square to order the candidates of MWLUs. Recently, Saif and Aziz (2011) introduced the hybrid method that depends on both linguistic information and statistical models for collocations extraction from Arabic corpus. This method consists of two main steps: candidate identification and candidate ranking.

From linguistic perspective, Arabic collocations were classified based on structural patterns into five types: a) Noun + Noun, b) Noun + Adjective, d) Verb + Noun, e) Verb + Adverb, f) Adjective + Adverb, and g) Adjective + Noun. The linguistic tools (stemming and part of speech tagging) were utilized with the types of classifications for collocations in order to select the bigram candidate's sets. For candidate ranking, four standard association measures (MI, EMI, LLR, and chi-square) were evaluated on the candidate's sets. Although this method achieved promise evaluation results on corpus of Modern Standard Arabic with the precision values ranging between 70.8% and 83.8%, however the association measures depends mainly on the frequency of the candidates and the performance are directly affected by the frequency of the words in the corpus. The core knowledge sources in Islamic domain are Quran and Hadith. Most collocations in these sources occur with the low frequency. Therefore, applying these standard association measures leads to exclude many common collocations in lower frequencies.

 

## 3. Proposed Model

The research design was built to reach the main research objective, which is extracting Islamic collocations from a corpus in the Islamic domain. It consists of the information within the dataset that has been used for this research. Additionally, it explains the pre-processing steps that were conducted on the dataset, including normalization, the removal of stop-words and stemming. Also, it includes the task of transforming the corpus into statistical information that may later be beneficial for the tasks of filtering and ranking. The task of ranking such candidates using the association measures is then explained. Finally, the evaluation process, which aims to evaluate the extracted collocations, is presented. The proposed model has been created in several stages. These stages will be explained in details below. Figure 1 presents the overall research design.

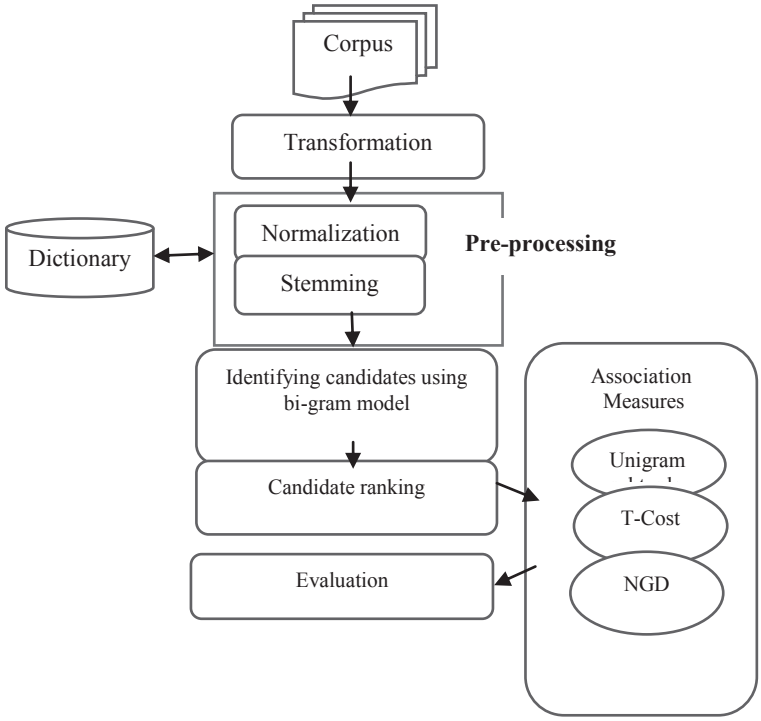


Figure 1. Research design framework

### *3.1 Corpus*

Tafsir ibn Kathir was used as the corpus for this work, which is to extract the Islamic collocations. This book is a classic guide that is used in Sunni Islam tafsir (or commentary of the Qur'an) by a Muslim scholar by the name of ibn Kathir. This book is seen to be a general summary of the earlier tafsir, or commentary, by al-Tabari, Tafsir al-Tabari. Tafsir ibn Kathir is popular due to the fact that it uses the Hadith (words of the prophet Mohammad) to describe, define, and explain every verse of the entire Qur'an. It is considered among the most authentic Tafsir (exegesis) of the Qur'an. Tafsir ibn Kathir, which is the corpus used for this research, consists of 114 documents that correspond to the chapters (Suraat) of the Qur'an (There are 114 different Suraat, or units, in the Qur'an).

### *3.2 Transformation*

During this step, the data is transformed into an internal representation so that it is ready to move to the next step, which is pre-processing. As a matter of fact, the data includes text files of gathered from the book Tafsir ibn Kathir. Due to the unstructured nature of the text gathered, there are several steps that must be conducted for transforming it into a structure that makes it easier to process. The following transformation steps are conducted.

 

*3.3 Preprocessing*

This step attempts to conduct multiple stages, which include normalization and stemming, to change the data to another format, so that statistical processes may be applied. This stage is extremely important and critical in terms of improving the results of extracting noun candidates from multiword units. The stages of this step are presented as follows.

**Normalization Stage**

During this stage, the noisy or undesirable data, such as digits, stop words, and special characters are removed.

A. Remove special characters

Special characters including "_+-*^:"?" are discarded, since it is not used in later steps.

B. Remove non Arabic letters

All non-Arabic letter or characters are removed.

C. Remove digits

All digits and numbers from 0 to 9 are discarded.

D. Remove diacritics

As previously shown, the Islamic text has multiple symbols of diacritics (Table 3.2), and therefore, those diacritics are discarded.

E. Remove definite articles

Definite articles are a collection of letters that may act as a prefix or suffix, which indicates the type of reference that is made by a noun. These are removed.

F. Tokenization

Generally, the task of tokenization is the process of dividing words from text into a set of clusters of sequential morphemes, of which usually correspond to the word stem. The following example shows how tokenizing the texts create a sequence of tokens.

"ان الذين كفروا بايات الله لهم عذاب شديد " becomes " ان_الذين_كفروا_بايات_الله_لهم_عذاب_شديد " after the process of tokenizing.

G. Removal of stop-words

Stop words are words that connect the sentences together, which tend to be irrelevant to the upcoming steps, and must be removed.

H. Filtering Arabic letters

There are several letters in the Islamic domain that come in several forms, and therefore, these letters must be unified to one root letter. Table 3.5 presents a sample of these letters.

**Stemming Stage**

The Islamic domain is generally an inflected, or synthetic, language, and affixes include a different function from non-synthetic languages, such as English. The lemma in the Islamic domain is essentially a stem of a collection of forms (there are thousands of forms in every set) that have common morphological, syntactic or semantic features (Dichy, 2001). For instance , the lemma 'قلم' has multiple forms, including, ' , 'قلمين' , 'الاقلام' , 'القلم' , 'قلمة' , 'قلمها' , 'قلمهم' , 'اقلامهم' , 'قلمنا'.

The stemming step aims extract the root stem of words that are included in the used corpus, to reduce the number of words that are used for further processing. Stemming is a crucial step, which produces a more meaningful list of terms by combining all different forms of every word in the corpus together. There are numerous researches that take into consideration different stemming techniques that focus on morphological structure, including the Buckwalter Arabic morphological analyzer by Buckwalter (2004), Larkey's light stemmer by Larkey et al. (2002), Khoja's root-extraction stemmer by Khoja and Garside (1999), and N-gram stemming technique by Mustafa et al. (2004).

Consequently, this work used a hybrid stemming technique (Alhanini and Ab Aziz, 2011), which has helped to avoid the disadvantages of two types stemmers (dictionary-based and rule-based stemmers) in order to obtain the

 

stem of every word in the used corpus.

An Islamic word consists of a stem of the word, and an affix, that reveal the tense, gender and singular or plural form of the word. Additionally, clitics are concatenated, or attached, to the word. Several clitics are concatenated to the start of the word (such as prepositions and conjunctions), and at the same time, some clitics, like pronouns, are concatenated at the end of the word. The stemming technique is used in order to segment a word into its different components (which are the affix, stem, and clitics) based on Arabic rules. Every word in the corpus has been segmented into the different components it has, based on the formula of the Islamic words. The core formula of Islamic words is explained as follows:

*Clitics + prefix + stem + suffix + clitics*

where the clitics, prefix, and suffix are attached to a word by option.

### *3.4 Creating Term-Term Matrix*

The aim of this step is to produce the terms (which are the words after the stemming process), and their frequencies in the corpus. During this step, a term-term matrix is created to include the terms as rows, and the context in which these terms occur in as columns. Figure 2 shows the term-term matrix. In the term-term matrix, the profile of the term is represented as a vector consisting of the co-occurrences with a collection of context words in its data. The co-occurrence is the number of occurrences of a term with another term $t_i$ (in a context window with a size $k$).

| | **T1** | **T2** | **T3** | . | . | . | . | **TN** |
|---|---|---|---|---|---|---|---|---|
| **T1** | **F11** | **F12** | **F13** | . | . | . | . | **F1N** |
| **T2** | **F21** | **F22** | **F23** | . | . | . | . | **F2N** |
| **T3** | **F31** | **F32** | **F33** | . | . | . | . | |
| . | . | . | . | . | . | . | . | . |
| . | . | . | . | . | . | . | . | . |
| **TN** | **FN1** | **FN2** | . | . | . | . | . | **FNN** |

Figure 2. Term term matrix

### *3.5 Generating Bigram Candidates*

The main aim of this step is to retrieve the candidates from the corpus that consist of two terms. The bi-gram candidate contains of two main units, the first unit is the head word and the second unit is complement word. For every sentence available in the corpus, the sequence terms with a size of two are retrieved as bigram candidates.

For the corpus used in this work, the sentences are the segmented to construct the list of bigram candidates, along with their frequencies. The list constructed is then used in the upcoming step (which is candidate ranking) in order to measure the degree of the association between every bigram candidate and its corresponding constituent terms.

### *3.6 Candidate Ranking*

The candidate ranking is based on the frequencies for the occurrences of words and corresponding co-occurrence in the corpus. In the candidate identification step, the syntactic information and information about the co-occurrence the words in the corpus are collected. The association measures are calculated to the available candidates in the bigram lists and a score of association strength is given for every candidate. For every pair of terms that are extracted, the association score that is assigned is a value that shows the intensity of statistical association between the two terms.

For this work, the three statistical measures, which are the unigram subtuples, T combined cost, and normalized Google distance, have been implemented to calculate the association degree between the bigram candidates and their main terms. Evert (2004) claims that an association measures is “a formula that computes an association score from the frequency information in a pair type’s contingency table”. Both the joint and marginal frequencies for a bigram (u, v) are shown in the contingency table, which is Table 1.

 

Table 1. Contingency table for candidate pair (u, v)

| | Y=v | Y≠ v | |
|---|---|---|---|
| X=*u* | A | B | |
| X≠ *u* | C | D | |
| | $C_1$=a +c | $C_2$=b +d | N=a +b +c +d |

According to the table for the candidate pair (u, v), the association measures were implemented in this work as follows. The initial measure is the Unigram subtuples, which has been created by Blaheta and Johnson (2001) as the follows:

$$USub(u,v) = \log\frac{ad}{bc} - 3.29\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}} \quad (1)$$

The second association measure is the T combined cost, which has been created by Tulloss (1997) as follows:

$$T - cost = \sqrt{U \times S \times R} \quad (2)$$

where U, S and R are three measures that were defined as follows:

$$U = \log(1 + \frac{\min(b,c)+a}{\max(b,c)+a}) \quad (3)$$

$$S = \log(1 + \frac{\min(b,c)+a}{a+1})^{-\frac{1}{2}} \quad (4)$$

$$R = \log(1 + \frac{a}{a+b})\log(1 + \frac{a}{a+c}) \quad (5)$$

The third and last measure is the Normalized Google Distance (NGD) measure, which has been created by Cilibrasi and Vitanyi (2007), in order to calculate the word semantic relatedness between a pair of words. The critical point is that the method analyzes the objects themselves. This precludes comparison of abstract notions or other objects that do not lend themselves to direct analysis, such as emotions and colors. While the previous method that compares the objects themselves is especially appropriate to acquire information about the similarity of entities themselves, irrespective of common beliefs about such similarities. Here, researchers have created an approach that makes use of only the name of an object and obtains knowledge about the similarity of objects, a quantified relative Google semantics, by taking existing informaiton generated by multitudes of Web users. This measure was given between two words based on the Google search engine page counts as the follows:

$$NGD(w_1,w_2) = \frac{\max\{\log f(w_1), \log f(w_2)\} - \log f(w_1,w_2)}{\log N - \min\{log\ f(w_1), \log f(w_2)\}} \quad (6)$$

where *f(w)* is the count of web pages that include the word, *w, f (w1,w2)* is the count of pages that contain the two words, and N is the count of website pages that are indexed by the Google search engine. In order to implement this measure to this research, it is formulated based on the contingency table for the candidate pair (u, v) as follows:

$$NGD(u,v) = \frac{\max\{\log b, \log c\} - \log a}{\log T - \min\{log\ b, \log c\}} \quad (7)$$

 

where T is the total count of words in the entire corpus. The Normalized Google Distance was used as one of the three association measures in order to rank the existing candidates. According to Cilibrasi and Vitanyi (2007), this association measure had performed well after evaluation. This association measure was implemented based on the above formula, so that the MWs are extracted automatically from the text provided. The Normalized Google Distance was implemented using the formula mentioned above as a similarity measure, and is predicted to achieve the highest results.

### *3.7 Evaluation*

In order to evaluate and assess the proposed method, the n-best evaluation technique (Evert 2005) has been implemented. This method makes use of the association score in order to rank candidates of collocations in the Islamic domain. This method generally includes three core steps.

The initial step is the extraction of the n-best list, which attempts to choose the top association scores of the candidate ranking. The second step is the process of using human annotators, which are a number of Arabic teachers, in order to manually choose true collocations from the n-best list with two tags. The first tag is 1, which shows true collocations, and the second tag is 0, which shows false collocations. Finally, the third step is to assess and evaluate the manual annotation of candidates using the equation for *precision,* which is calculated as follows:

$$Precision = \frac{TP}{TEC} \tag{8}$$

Where TP is the count of correctly extracted multiword units. TEC is the count of the total extracted multiword units (the n-value used for n-best list).

## 4. Analysis of Results

This evaluation experiment was conducted to assess and evaluate all of the association measures that were used in ranking the candidates for every MWU bi-gram list. Every MWU is extracted individually by implementing the technique for MWUs in the Islamic domain. Therefore, the extracted lists are separate, based on every kind of MWU. The n-best evaluation technique was used for every list of the individual classifications of the MWU, including noun compounds and collocations.

The core aim of this work was to evaluate the association measures that are implemented for ranking the candidates of the MWUs. Based on the n-best evaluation technique, the data set of the MWUs was obtained the bigram list based on the properties of noun compounds for the Arabic language.

In general, the experiment creates a comparison for the best 500 outputs extracted for every single type of association measure. Every output is a pair word. Based on the best 500 outputs, the n-best evaluation that chooses the n-best set for every association measure was conducted. The n was between 100 and 500, at units of 100.

Table 2 shows the precision values of the unigram subtuples, the T combined cost, and the normalized Google distance after the pre-processing stage. It is clear that the Normalized Google Distance had the highest precision values, followed by the T combined cost, and then the unigram subtuples technique. Figure 3 shows the graph that shows these results, with the x-axis as n, and the y axis as they precision values for all three techniques, which are the unigram subtuples, the T combined cost, and the Normalized Google Distance techniques.

 

Table 2. Precision values for the three techniques

| n | Unigram subtuples | T combined cost | Normalized Google Distance |
|---|---|---|---|
| 100 | 0.79 | 0.81 | 0.93 |
| 200 | 0.76 | 0.8 | 0.93 |
| 300 | 0.75 | 0.77 | 0.91 |
| 400 | 0.75 | 0.72 | 0.84 |
| 500 | 0.74 | 0.71 | 0.82 |

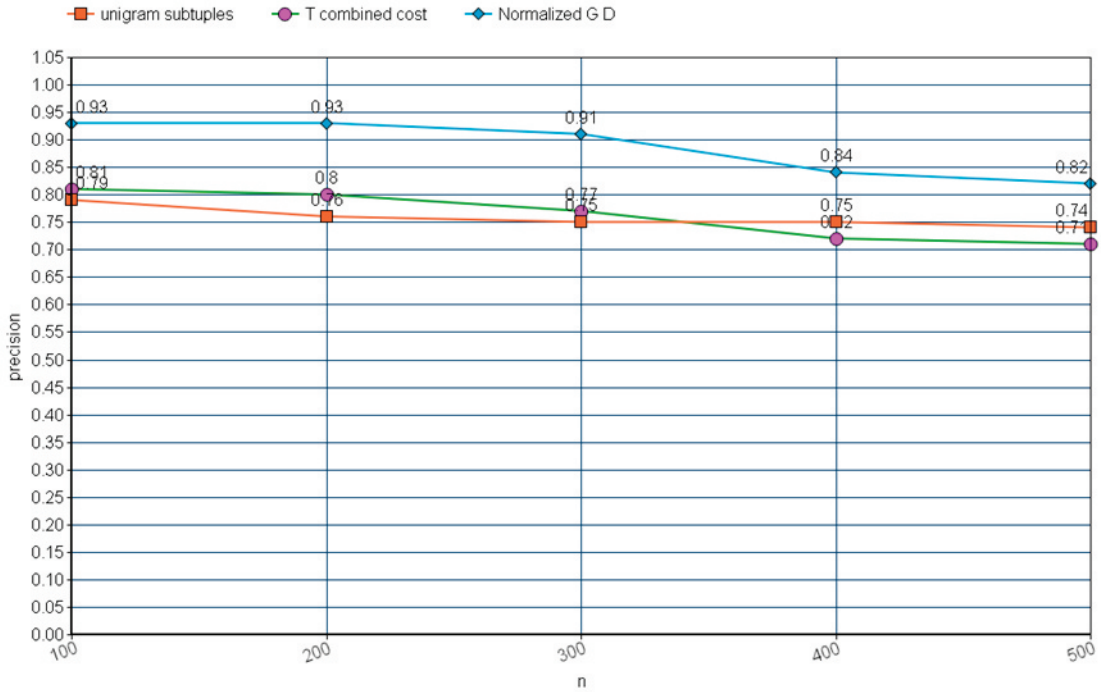


Figure 3. Graph of precision values of the three techniques

A comparison was made between the results achieved from the experiments conducted in this work, and the results from the work of Saif (2011). This work had used several statistical based approaches (unigram subtuples, T combined cost, and Normalized Google Distance) for the extraction of MWs from text written in the Arabic language, and, in particular, in the Islamic domain.

Saif (2011) had used a set of rule based approaches (log-likelihood ratios, mutual information, enhanced mutual information, and Pearson's A2 test) for the extraction of MWs from text written in the Arabic language. A comparison of the precision values achieved for both works when n = 100 to 500 has been carried out.

From these results, it can be concluded that the values for precision for most of the approaches, whether rule based or statistical, had achieved results that were very close, with one exception. The Google Similarity Distance has performed relatively higher than any other approach in both works, with a precision value of 0.886. This shows that the Google Similarity Distance may be further improved to be a promising approach for the extraction of MWs from text written in the Arabic language, specifically for the Islamic domain.

 

## 5. Conclusion

Regarding the results that were achieved by this work, the precision values can be summarized. The average precision values for the three approaches of unigram subtuples the T combined cost, and the Normalized Google Distance were 0.758, 0.762, and 0.886 respectively.

There are several recommendations that must be made for future work in the field of the extraction of MWs from Arabic text in the Islamic domain. These recommendations are as follows:

1. This work has used the Islamic domain when extraction MWs. Other domains may be explored by using the same model, or an enhanced version of the model.

2. This work had made use of three statistical based approaches, namely, unigram subtuples, T combined cost, and Normalized Google Distance. The model that had used these approaches may be enhanced by adding more statistical based approaches, or even rule based approaches.

3. More association measures may be implemented in order to identify MWs and extract them effectively.

4. This model has been constructed for working with the Arabic language. It may be further enhanced to work with other languages, one at a time, or multiple languages at a time.

The extraction of MWs from Arabic text in the Islamic domain was successfully carried out by using the three mentioned statistical association measures (unigram subtuples, T-combined cost, and normalized Google distance), with relatively high precision values after evaluation. These measures are expected to perform better and have a higher precision value if more research was carried out, with taking into consideration the above mentioned recommendations for future work.

## References

Attia, M., Tounsi, L., Pecina, P., Genabith, J. & Toral., A. (2010). “Automatic extraction of Arabic multiword expressions”.

Attia, M. (2006). “Accommodating multiword expressions in an Arabic LFG grammar”, *Advances in Natural Language Processing*.

Attia, M. (2005). “Developing a robust Arabic morphological transducer using finite state technology”, *8th Annual CLUK Research Colloquium*, Manchester, UK.

Appelt, D. (1999). “Introduction to information extraction”, *AI Commun.* 12(3): 161-172.

Alhanini, Y., & Ab Aziz, M. (2011). “The Enhancement of Arabic Stemming by Using Light Stemming and Dictionary-Based Stemming”. *Journal of Software Engineering & Applications*, 4(9).

Baldwin, T. (2005). “Deep lexical acquisition of verb–particle constructions”. *Computer Speech & Language*, 19(4), 398-414.

Blaheta, D., & Johnson, M. (2001). “Unsupervised learning of multi-word verbs”. In Proc. of the ACL/EACL 2001 *Workshop on the Computational Extraction, Analysis and Exploitation of Collocations*. pp. 54-60.

Boulaknadel, S., Daille, B., & Aboutajdine, D. (2008). “A multi-word term extraction program for Arabic language”. *Proceeding of the 6th International Conference on Language Resources and Evaluation*, May 28-30, Marrakech Morocco, pp: 1485-1488.

Bounhas, I., & Slimani, Y. (2009). “A hybrid approach for Arabic multi-word term extraction”, *Proceeding of the International Conference on NLP-KE* 2009, Department of Computer Science, University of Tunis, Sept. 24-27, Tunis, Tunisia, pp: 1-8.

Manning C., & Schütze, H. (1999). “Foundations of Statistical Natural Language Processing”, *MIT Press*. Cambridge MA.

Cunningham, H. (2013). “Developing Language Processing Components With Gate Version 7”*, The University of Sheffield*.

Han, X., & Zhao, J. (2009). “CASIANED: People Attribute Extraction based on Information Extraction”, Anjuran Madrid, Spain.

Manning, C., Raghavan, P., & Schütze, H. (2008). “Introduction to information retrieval”, Vol. 1. Cambridge: *Cambridge University Press*.

 

Pinheiro, V., Furtado, V., Pequeno, T., & Nogueira, D. (2010). "Natural Language Processing based on Semantic inferentialism for extracting crime information from text", *IEEE International Conference on Intelligence and Security Informatics.* ISI. 2010. 19-24.

Ponte, J., & Croft, B. (1998). "A language modeling approach to information retrieval", Anjuran ACM. Melbourne, Australia.

Church K., & Patrick, K. (1990). "Word association norms, mutual information, and lexicography", *Computational linguistics*.

Cilibrasi, R., &. Vitanyi, P. (2007). "The google similarity distance. Knowledge and Data Engineering", *IEEE Transactions* on, 19(3), 370-383

Cilibrasi, R., &. Vitanyi, P. (2004). "The Google Similarity Distance", *arXiv preprint* cs/0412098.

Dias, D., & Vintar, S. (2005). "Unsupervised learning of multiword units from part-of-speech tagged corpora: does quantity mean quality" *Springer Berlin Heidelberg*. pp. 669-679.

Diab, M., Hacioglu, K., & Jurafsky, D. (2004). "Automatic tagging of Arabic text: From raw text to base phrase chunks", *Proceeding of the NAACLHLT*, Boston, USA., pp. 149152.

Dichy, J. (2001). "On lemmatization in Arabic, A formal definition of the Arabic entries of multilingual lexical databases*", In ACL/EACL 2001 Workshop on Arabic Language Processing: Status and Prospects*. Toulouse, France.

Evert, S., & Krenn, B. (2005). "Using small random samples for the manual evaluation of statistical association measures", *Computer Speech & Language*, 19(4), 450-466.

Frantzi, K., Sophia, A., & Hideki, M. (2000). "Automatic recognition of multi-word terms: The C-value/NC value method", *Int. J. Digital Libraries*, 3: 115-130.

Grefenstette, G. (1999). "The World Wide Web as a resource for example-based machine translation tasks", *In Proceedings of the ASLIB Conference on Translating and the Computer* (Vol. 21).

Habash, N., Rambow, O., & Roth, R. (2009). "MADA+ TOKAN: A toolkit for Arabic tokenization, diacritization, morphological disambiguation, POS tagging, stemming and lemmatization", *In Proceedings of the 2nd International Conference on Arabic Language Resources and Tools (MEDAR)*, Cairo, Egypt .pp. 102-109.

Hajic, J., Smrz, O., Buckwalter, T., & Jin, H. (2005). "Feature-based tagger of approximations of functional Arabic morphology", *In Proceedings of the Workshop on Treebanks and Linguistic Theories* (TLT), Barcelona, Spain.

Kilgarriff, A., & Grefenstette, G. (2003). "Introduction to the special issue on the web as corpus". *Computational linguistics*, 29(3), 333-347.

kim. S. (2008). "Statistical Modelling of Multiword Expressions", *PhD thesis, computer sciences and software engineering dept*., University of Melbourne.

Mustafa, S., &. Al-Radaideh, Q. (2004). "Using N-grams for Arabic text searching", *Journal of the American Society for Information Science and Technology*, 55(11), 1002-1007.

Pecina, P. (2009). "Lexical Association Measures Collocation Extraction", *Institute of Formal and Applied Linguistics*. Charles University, Prague. DCU, Dublin.

Ramisch, C., Schreiner, P., Idiart, M., & Villavicencio, A. (2008). "An evaluation of methods for the extraction of multiword expressions", *In Proceedings of the LREC Workshop-Towards a Shared Task for Multiword Expressions*. MWE 2008,pp. 50-53.

Saif, A., & Ab Aziz, M. (2011). "An Automatic Collocation Extraction from Arabic Corpus:, *Journal of Computer Science*, Vol. 7, No. 1, 2011.

Saif, A. (2011). "An Automatic Multiword Lexical Unit Extraction from an Arabic Corpus", *Master Thesis*, UKM, Malaysia.

Seretan, V. (2008). "Collocation Extraction Based on syntactic parsing", *Ph.D. thesis*, University of Geneva.

Scott, M., & Christopher, T. (2006). "Textual Patterns: Key Words and Corpus Analysis in Language Education", Philadelphia: John Benjamins.

Tulloss, R. (1997). "Assessment of Similarity Indices for Undesirable Properties and a New Tripartite Similarity Index", *Mycology in sustainable development: expanding concepts, vanishing borders*, 122.

 

Cruys, T., & Moirón, B. (2006). "Lexico-semantic multiword expression extraction", *In Proceedings of the 17th Meeting of Computational Linguistics in the Netherlands* (CLIN), University of Leuven, Leuven, Belgium. 175.190.

Zhang, Y., Kordoni, V., Villavicencio, A., & Idiart, M. (2006). "Automated multiword expression prediction for grammar engineering", *In Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*. pp. 36-44. Association for Computational Linguistics.

The IISTE is a pioneer in the Open-Access hosting service and academic event management. The aim of the firm is Accelerating Global Knowledge Sharing.

More information about the firm can be found on the homepage:
http://www.iiste.org

## CALL FOR JOURNAL PAPERS

There are more than 30 peer-reviewed academic journals hosted under the hosting platform.

**Prospective authors of journals can find the submission instruction on the following page:** http://www.iiste.org/journals/ All the journals articles are available online to the readers all over the world without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself. Paper version of the journals is also available upon request of readers and authors.

## MORE RESOURCES

Book publication information: http://www.iiste.org/book/

### IISTE Knowledge Sharing Partners

EBSCO, Index Copernicus, Ulrich's Periodicals Directory, JournalTOCS, PKP Open Archives Harvester, Bielefeld Academic Search Engine, Elektronische Zeitschriftenbibliothek EZB, Open J-Gate, OCLC WorldCat, Universe Digtial Library , NewJour, Google Scholar