Journal of Theoretical and Applied Information Technology
31st December 2013. Vol. 58 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

ARABIC TERM EXTRACTION USING COMBINED APPROACH ON ISLAMIC DOCUMENT

[1]ALI MASHAAN ABED, [2]SABRINA TIUN, [3]MOHAMMED ALBARED

Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, Malaysia

E-mail: [1]ali.alani1985@gmail.com

ABSTRACT

While a wide range of methods has been conducted to English terminology extraction, relatively few studies have been applied to Arabic terms extraction in Islamic corpus. In this paper, we present an efficient approach for automatic extraction of Arabic Terminology (SWTs, MWTs). The approach relies on two main filtering steps: the linguistic filter, where simple part of speech (POS) tagger is used to extract candidate MWTs matching given syntactic patterns, and the statistical filter where several statistical methods (PMI, Kappa, CHI-squire, T-test, Piatersky- Shapiro and Rank Aggregation) are used to rank candidate MWTs and we applied IF.IDF to rank the SWTs candidate. Our approach extracted the bi-gram candidates of MWTs Islamic term from corpus and evaluated the association measures (STWs and MWTs) by using the n-best evaluation method.
Keywords: *Term Extraction, SWTs, MWTs, Association measures, n-best evaluation*

1. INTRODUCTION

Automatic terminology extraction is an important task in many NLP and knowledge engineering applications, such as indexing and information retrieval (IR) machine translation [15], information extraction, domain specific lexicon construction [16], and topic extraction [18]. In fact, automatic terminology extraction contributes to all domain-oriented NLP domains. Generally the term extraction (TE) process consists of two necessary steps: (1) identifying term candidates (either single or multi–word terms) from text, and (2) filtering through the candidates to separate terms from non-terms. In order to execute these two steps, term extraction systems make use of different degrees of linguistic filtering and statistical measures ranging from raw frequency to Information Retrieval (IR) measures such as Term Frequency/Inverse Document Frequency (TF.IDF) [20], up to more complex methods such as the C-NC Value method (Frantzi et al., 1999), or mutual information. Others make use of extensive semantic resources or lexical association measures like log likelihood [12], but as underlined in [4], such methods face the hurdle of portability to other domains. Current and past researches on automatic terminology extraction have introduced several approaches and techniques to extract and recognize domain specific terms [8]. These approaches can be classified into three categories; (1) linguistic approaches, (2) statistical approaches and (3) combined approaches. In this study, several automatic terminology recognition and extraction algorithms will be investigated for extracting domain specific MWTs in the Islamic domain. These algorithms are hybrid approaches based on statistical and linguistic analysis. In addition, we also develop an algorithm based on TF.IDF weighting for automatic SWTs recognition. Extraction algorithms will be investigated for extracting domain specific multiword terms in the Islamic domain. The interesting challenge tackled this study is to adapt the techniques that work well on general-purpose texts to handle domain-specific texts. Furthermore, recently, religious domain has become an interesting and challenging area for NLP and text mining. In the following section, we will discuss the related work of SWTs and MWTs.

2. RELATED WORK

2.1 SWTs Related Work

There are few approaches are available for single-word term extraction, which rely on frequency-related information. For instance, Xu et al. [22] have used the TF-IDF method to extract single-word terms from three domains: financial management succession, stock market, and crime-drug domain. They have stated that, TF-IDF approach is very appealing for extracting single-word terms, even though it failed to deliver particular data, pertaining to its efficiency. Few other researchers have compared the frequency of incidence of a word in a domain-specific corpus with its frequency of occurrence in a reference corpus [5] [19]. For instance, Baroni [5] has made use of a small list of seed terms, which represent some specific domains. They have arbitrarily merged these seed terms and employed the collaboration as a Google query string. The top *n* pages returned for each query were retrieved and formatted as text. Subsequently, new unigrams have been extracted from the corpus of retrieved pages, by evaluating the frequency of incidence of each word in this collection, with its frequency of incidence in a reference corpus, and then they compared the frequencies by employing the log odds ratio measure.

2.2 MWTs Related Work

This section presents the existing multiword term (MWTs) extraction works in different languages, which based on hybrid method. One of them is the work which has been presented by Bounhas [6] who have proposed a hybrid method to extract multiword terminology from Arabic corpora. They have applied several tools to extract and identify the compound nouns. Their approach used the Arabic morphological analyser (AraMorph), proposed by Hajic et al. [17]. The AraMorph has been applied to compute the morphological features, required for the syntactic rules.

Al khatib et al. [1] has proposed a hybrid approach to extract multi-word terms from Arabic corpus. They concentrated on compound nouns as in important type of MWT and select bi-gram term .The approach relies on two main filters: (i) linguistic filter, where simple part (POS) tagger is use to extract candidate MWTs .This step contains; prepositions word's classification, extraction of noun's sequence and noun's sequence that associated by prepositions, testing each extracted sequence based on MWTs syntactic patterns. (ii) Statistical method where log-likelihood ratio and C-value are used to rank bi-gram candidate MWTs.

Frantzi et al. [15] have proposed a hybrid approach to extract multi-word terms from English corpus combining linguistic and statistical. From linguistic point of view, their approach extracts the candidates of multi-word terminology based on some linguistic information such as POS tagging of the corpus, which is then utilized in the linguistic filter. The linguistic filter includes all kinds of terminologies, and generates beneficial outcome. The stop-list also prevents the extraction of candidates which are less likely to be terminology, and enhances the precision of the output list. Furthermore, the C-value is utilized to ensure that, the extracted outcome is basically a multiword terminology. The C-value measure has been utilized for resolving the problem of nested terms. Generally, the terms used chemistry documents, automotive, and biomedical articles, follow by a specific pattern of combined nouns and adjectives. In the syntactic point of view; generally they are compound nouns and associated constructions. For example, the terms found in biomedical abstracts of the Genia corpus are names of diseases and drugs, chemical elements, anatomy and other names. Consequently, methods for automatic identification of compound nouns, when used on domain-specific data, might be effective in extracting multiword terminologies aimed at comparing MWTs interpretation and classification. SanJuan et al [25] have merged three linguistic resources and techniques such as: statistical associations, WordNet, and clustering, to construct a hierarchical model, as against manual annotation of terms of the corpus. They have compared their model with traditional clustering algorithms, and with the manually created ontology of the corpus.

Chen [24] have proposed a novel automatic statistical method for determining multi-word terms depending on co-related text-segments present in a range of documents. The proposed method applied a novel and efficient statistical strategy for determining multi-word terms. Chen have presented the multi-word term extraction system, a new automatic statistical approach to identify multi-word terms based on co-related text-segments existing in a set of documents. The suggestion was used a new and effective statistical method for identifying multi-word terms. The above system consists of four components: (i) text-segment generator which utilizes a small pre-defined stop-list as an preliminary input, to classify a set of text documents into text-segments; (ii) text segment-weigher, which computes the segment-weight for each created text-segments, (iii) text segment-segmenter, which segments all the text-segments depending on their segment-weights to create new text-segments-term candidates; the term candidates shall be re-input for additional segmentation, or directly input to the subsequent component, (vi) term identifier, which recognizes the resulting term candidates to become terms, according to a specific threshold.

3. OUR PROPOSED APPROCH

Our proposed of terminology extraction approach requires the following steps, the first step is pre-processing; the algorithms can have optimal performance with minimum noise. All the steps described in this section are based on heuristics and simple transformation rules. The text pre-processing methods which are applied as the followings; (i) Removal of digits, punctuation marks and diacritics for each text in the Arabic dataset as example remove (', ',': ','? ',' \ ' …), (ii) Normalization of some Arabic letters by normalizing letter (Replace آ, إ and أ with ا , Replace ة with ه , ي withى). (iii) Splitting the text into tokens which generally consists only letters (القطرات_القليلة_تصنع_جدولاً). (iv) stop word removal example (’’من”,” عن”,”الى”,”على” ,”في”) [2] .

3.1 SWTs Method

It is difficult to identify single terms because they do not carry precise meanings such as multi term, however, we cannot do without them because of their importance to represent a particular area. In order, to extract SWTs, adopted statistical approach used, frequency word-Reverse frequency document (TF.IDF) way is used, and summed up as follows: After pre-processing we use the result in the second phase, which is in the calculation of frequencies each word in every document of the selected code. We then assign the weight of each word, is calculated using the TF.IDF We arrange these words in descending order by weight and experimentally to determine the threshold that separate the words accepted and which are rejected words.Term Frequency (TF) weighting is also recognized as a simple method for term weighting, $w_{i=TF_i}$.

According to this method, there is an equality of the weight of a term in a document and the number of times of appearance of this term in the document, i.e. to the raw frequency of the term in the document.It is pointed that Boolean weighting and TF weighting do not take the frequency of the term into consideration throughout all the documents in the document corpus. Term Frequency × Inverse Document Frequency (TF.IDF) weighting is seen as the most popular method used for term weighting since it considers this property. By using this approach, assigning the weight of term i in document d to the number of times the term appears in the document is proportional, and it is in inverse proportion to the number of documents in the corpus in which the term appears.as equation (1) :

$$w_{i=TF_i}.\log(\frac{N}{ni}) \qquad (1)$$

TF.IDF weighting approach gives weight to the frequency of a term in a document with a factor discounting its importance in case when the appearance of it is found in most of the documents. For example, this can be applicable to the case in which the term is assumed to have little discriminating power.

3.2 MWTs Method

The proposed approach for extraction Islamic MWTs consist of two main steps; (i) the linguistic filter, where we extract candidate MWTs, and extract bi-grams from candidate MWTs, (ii) the statistical filter, where we rank bi-grams by association measure .we will cover the two steps in more details in step (A) and (B).

A. Generating Bi-gram list

This step generates the list of candidates, which is created from two words from the corpus. This is the initial phase of our technique, where linguistic filters syntactic pattern and simple part of speech (POS) tagger is employed to extract candidate MWTs, by using Bi-gram list. The Bi-gram list candidate comprises of two parts: (i) head word by using unigram list, (ii) complement word [21]. The unigram includes all words in corpus with their frequency and the linguistic classes. The linguistic categories for each word in the list are the additional part of unigram list. This part needs linguistic processes, such as; Lemmatization and/or the Arabic POS tagging.

Lemmatization has been employed for all the words in corpus. Prior to the extraction of the Islamic MWTs, it is essential to determine the word-class of the elements of each candidate. The word-class (linguistic category) is described by the syntactic or morphological behaviour of the lexical item in questions. There are several linguistic classes such as, noun, verb, and others. Lemmatization is the method of obtaining the lemma (lexeme) for a given word. This method might involve complicated tasks like, comprehending context and identifying the part of speech of a word in a phrase. However, there are some challenges in searching for words as baseforms (lemmas); these complications differ from one language to the other based on grammatical changes such as, tense or plurality. For instance, in English the lemma 'walk' can appear in the context in many forms such as, 'walks', 'walked', 'walking' and so on. On the other hand, Arabic is an inflected (synthetic) language, where ambles have a distinct purpose as against non-synthetic languages like English. The lemma in Arabic is actually a stem of a set of forms (hundreds or thousands of forms in each set), which share the same morphological, syntactic or semantic features [10].

The bi-gram candidate is composed of two sections; the first section is known as the 'head Word' and the next section is referred to as 'complement word'. Based on our discussion earlier, the unigram list includes every one of the words in corpus along with their statistical information. Essentially, the pair of word, which is most likely recurrent in corpus, should be recurrent in a minimum of one of its components. Because of this, the words, which use to create the bi-gram candidates, need to have increased frequency. Consequently, in this phase, the head word of large candidate is chosen from unigram list, which have high frequency.

The Bi-gram candidates is created by several stages as follows: first and for most, the words in unigram list are arranged in accordance with their frequency i.e. highest to lowest ; the word stem, which has frequency less that 2 are overlooked in the next phases . From the unigram list, each and every word is chosen as the head word in the bi-gram candidate. For this head word, the second phase has to create all Bi-gram candidates from corpus, where their head word matches to this head word; furthermore, for the bi-gram candidates the frequency of each bi-gram candidate in the corpus is measured .There are two possibilities of linguistic groups based on the combination of the head word and complement word; both of the head and complement words have only linguistic category, or one of them has more than one linguistic category. In case of the first possibility, the bi-gram candidate is directly kept with its frequency and linguistic category in the bi-gram list. On the other hand, in case of the second possibility, the phrase that consists of the bi-gram candidate should be ascertained by using the disambiguation POS.

B. Candidate Ranking

The second step of MWTs is the candidate ranking .The main aim of this step is to compute the association measures for each candidate in all bi-gram lists, and to rank the candidates according to their association scores. The candidate ranking relies on frequency information about word occurrence and co-occurrence in a corpus. In the previous steps, the candidate pairs are identified from the bi-gram list according to their syntactic structures (patterns). For each couple of extracted words from a corpus, association score is a single real value that point to the amount of (statistical) association between the two words. Many association measures are based on statistics assumption tests while some others are purely heuristic combinations of the observed joint and marginal frequencies In our work, the following association measures ;t-test, chi-square, point wise mutual information, kappa, Piatersky-Shapiro , Rank Aggregation, have been selected. These measures have sturdy association according to selected recent researches for terminology extraction ([15][23]. we will discuss more detail of all association measure were used in our approach. In our study, we selected four association measures that have strong association, according to some recent methods for STWs. The first association measure is

Chi-square used for MWTs by many researchers [3] Where it compares between the observed and expected frequencies. It is calculated for bi-gram (x, y) as follows:

$$\mathrm{x}^2 = \frac{\left(\mathrm{f}_{\mathrm{xy}}\text{-}\left(\frac{\mathrm{f}_\mathrm{x}\mathrm{f}_\mathrm{y}}{\mathrm{n}}\right)\right)^2}{\left(\frac{\mathrm{f}_\mathrm{x}\mathrm{f}_\mathrm{y}}{\mathrm{n}}\right)} \qquad (3)$$

T-test is the second association measure used to compare between two population means where you have two samples in which observations in one sample can be paired with observations in the other sample
It is calculated for bi-gram (w_1, w_2) as follows:

$$\mathrm{t}= \frac{\frac{freq\,(w1\,,r,w2\,)}{N} - \frac{freq\,(w1\,)}{N} \cdot \frac{freq\,(w2\,)}{N}}{\sqrt{\frac{freq\,(w1\,,r,w2\,)}{N^{\,2}}}} \qquad (4)$$

Pointwise mutual information (PMI) is third association measure and this measure has been used as an association measure to rank the candidates of collocation by Church and Hanks [8] PMI is calculated for collocation (w_1, w_2) as follows:

$$PMIf\ (w^1,\ldots,w^n) = \log_2 \frac{f\,(w^1,\ldots,w^n)}{N^{n-1}\prod_{1}^{n} p(w_i)} \qquad (5)$$

The fourth association measure is the kappa, where coefficient (Cohen) [7] is generally regarded as the statistic of choice for measuring agreement on ratings made on a nominal scale. Cohen's kappa measures the agreement between two raters who each classify *N* items into *C* mutually exclusive categories. The equation (6) for κ is:

$$= \frac{\mathrm{p(w_1w_2)}+p(\overline{w1w2})-\mathrm{p(w_1*)p(*w_2)}-\mathrm{p(\overline{w_1}*)}-\mathrm{p(*\overline{w_2})}}{1-\mathrm{p(w_1*)p(*w_2)}-\mathrm{p(\overline{w_1}*)}-\mathrm{p(*\overline{w_2})}}$$
(6)

Piatersky- shapiro is the fifth association measure it the use to rank the candidates of collocation $\mathrm{Piatersky-Shapiro}$ is calculated for bi-gram (w_1, w_2) as follows:

$$\mathrm{Piatersky - Shapiro} = \mathrm{p(w_1w_2) - p(w_1\,*)p(*\,w_2)} \qquad (7)$$

Finally, each of the above association measures methods gives a ranked list. We tried the following approach to combine these ranked lists:

Rank Aggregation (RA): The aim is to combine ranked lists produced by several association measures using information of the ordinal ranks of the elements in each list. The weighted combination method has proved to give better results the individual association measure [11] [9]. Given multiple ordered lists **$\mathbf{L_1}$, $\mathbf{L_2}$...$\mathbf{L_k}$** of CNs, the rank aggregation problem is to combine these lists into a single ranked list. We use the following rank aggregation heuristic which is called Borda’s positional ranking:

Given lists **$\mathbf{L_1}$, $\mathbf{L_2}$...$\mathbf{L_m}$** , where m ≤k for each candidate c ∊ NNCs and list **$\mathbf{L}_i$**, the score $B_{L_i}(c)$ is the number of candidates ranked below c in **$\mathbf{L}_i$**. The total Borda score is:

$$B\,(c) = \sum_{i=1}^{m} B_{L_i}\,(c) \qquad (8)$$

The candidates are then sorted by descending Borda scores.

4. EVALUATION METHOD

This section present the quantitative evaluation method according to (Evert 2005) that assesses the statistical association measures in term extraction (SWTs, MWTs). This method is called *n*-best evaluation that uses association measure to rank the extracted terms candidates from a text corpus, and computes the precession for sets of highest-ranking candidates, called n-best lists. The *n*-best evaluation method involves of three steps to evaluate the statistical association measures for the extraction terms from corpus; selection the *n*-best list, manual annotation, precision calculation, to select n-best list the unigram list for SWTS and bi-gram MWTs that contains the candidates with their association scores. The candidates are organized from highest to lowest score for each single association measure. Finally, the highest n candidates from SWTs and SMTs ranking are selected as n-best list. The second step is the manual annotation the phase is important to check the candidates in n-best list are real MWTs terms or not. In this step, from the n-best list of each association measure, each candidate is passed on to human annotators for manual selection of the true MWTs term. Each candidate is noticeable as one of the four following tags, T (true term), N (not true), NT (cannot decide), Err (ambiguous mean). Finally, Precision calculation use to compute the precision of association measure it is next the manual annotation of candidates in n-best list, as follows:

$$\mathrm{Precision} = \frac{number\ of\ true\ MWLU\ term}{total\ number\ of\ MWLU\ term} \qquad (9)$$

5. DISCUSSION AND RESULT

In our experiment, we have used the Islamic corpus. Our corpus is an electronic corpus containing Classic Arabic (CA) and modern standard Arabic (MSA) collected from online Islamic newspaper archives, including shamela.ws, and islamweb.net.

In order to evaluate the association measure, firstly, we have computed the precision for each n-best list. In this experiment, the n-best has been selected set from data set for each association measure with n ranging from 100-500 at intervals of 100. For each association measure, we have computed the precision for n-best list both of SWTs and MWTs.

5.1 Experiment SWTs

The main objective of this experiment is to assess the association measure (TF.IDF) that is used for ranking the single-word terms candidates according to the n-best evaluation method.

We will shows the (precision, recall and f-measure) values of the n-best lists on the single word term with (N=100, 200, 300, 400, 500) that are ranked according to TF.IDF.

Table 1: Experiment result of TF.IDF

Evaluation	Precision	Recall	f-measure
N=100	**0.88**	0.03	0.08
N=200	0.85	0.08	0.14
N=300	0.83	0.18	0.28
N=400	0.82	0.35	0.46
N=500	0.78	0.51	0.59

From table 1 the TF.IDF achieved the height ratio in that the ratio of precision be high value in (88% for N=100), the ratio gradually decrease until it reaches to its lowest rate in (77% for N=500).

2. Experiment MWTs

The main objective of this experiment is to assess the association measure (Chi-square, T-test, PMI, P-Shpiro-kappa, and RA) that is used for ranking the multi-word terms candidates according to the n-best evaluation method with (N=100,200,300,400,500).

Table 2: The precision values for n-best of MWTs

	PMI			Chi-square			T-test		
Evaluation	Precision	Recall	F-measure	Precision	Recall	F-measure	Precision	Recall	f-measure
N=100	0.76	0.04	0.07	0.73	0.04	0.07	0.79	0.04	0.08
N=200	0.74	0.08	0.14	0.72	0.07	0.13	0.76	0.08	0.14
N=300	0.72	0.18	0.29	0.7	0.18	0.28	0.7	0.18	0.28
N=400	0.69	0.35	0.46	0.67	0.34	0.45	0.69	0.35	0.46
N=500	0.68	0.52	0.59	0.66	0.5	0.57	0.68	0.51	0.59
	piatersky-shpiro			Kappa			Combination		
Evaluation	Precision	Recall	F-measure	Precision	Recall	F-measure	Precision	Recall	f-measure
N=100	0.79	0.04	0.08	0.78	0.04	0.08	**0.80**	0.04	0.08
N=200	0.74	0.08	0.14	0.77	0.08	0.14	0.76	0.08	0.14
N=300	0.7	0.18	0.28	0.69	0.18	0.28	0.72	0.18	0.29
N=400	0.66	0.34	0.45	0.68	0.35	0.46	0.70	0.35	0.47
N=500	0.65	0.5	0.56	0.67	0.51	0.58	0.68	0.52	0.59

From table 2 the Rank Aggregation (RA) achieved the height ratio value precision in (80% for N=100), in N=500 the PMI achieved the height ration precision in (69% for N=500).

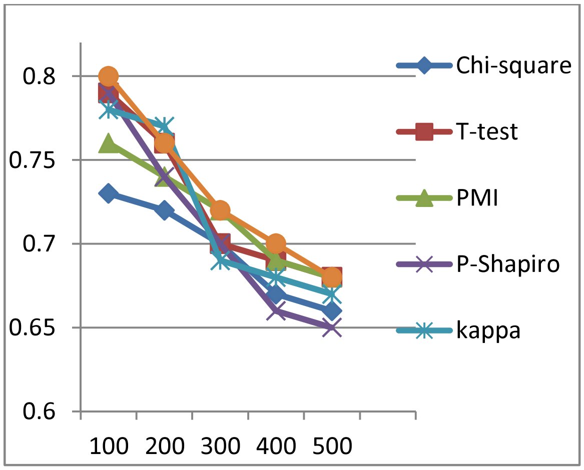

Figure 1 The N-Best Precision For The Association Measure On Mwts Data Set

From Figure 1 as we said before the rank combination measures that achieved the highest precision rate for all n-best lists on MWTs in Islamic data set. The Chi-Square is the worst association measure that achieved the lowest precision value in the most n-best list (73% when the n=100).

6. CONCLUSION

In this paper, we have presented our method for terminology extraction (SWTs, MWTs) from Islamic corpus. This method is a hybrid method that depends on both linguistic information and statistical filter. In our system the TF.IDF was applied to extract SWTs, which achieved good ratio in Islamic corpus.

We have concentrated on MWTs as an important type of terminology, and chose to extract bi-gram, which constitute a high percentage of compound nouns. Extraction of MWTs required substantial software development effort. The proposed approach started with linguistic filter step by use POS tagger for extract candidate. The next step for extract MWTs is statistical filter it includes rank bi-gram we used five association measures to rank candidates based on (t-test, chi-squarer, point wise mutual information, kappa, Piatersky-Shapiro). Furthermore, to enhance our method we applied Rank Aggregation (RA) to combination all the five association measures.

Our method has been applied in-house to collected corpus from Islamic newspaper archives and Islamic website. In order to evaluate association measures, we used the n-best evaluation method that selects n-best set for each association measure and annotates the extracted candidates manually. In our experiment, the Rank Aggregation (RA) has proved to be the best association measure that has achieved the highest precision value 80% in the n-best list with n=100.

ACKNOWLEDGEMENTS

This paper is partially funded by Malaysian Government under research grant ERGS/1/2013/ICT07/UKM/03/1

REFRENCES

[1] Al Khatib, K. & Badarneh, A. 2010. Automatic Extraction of Arabic Multi-Word Terms. *Computer Science and Information Technology (IMCSIT), Proceedings of the 2010 International Multiconference on*, PP. 411-418.

[2] Al-Shammari, E. & Lin, J. 2008. A Novel Arabic Lemmatization Algorithm. *Proceedings of the second workshop on Analytics for noisy unstructured text data*, pp. 113-118

[3] Attia, M., Tounsi, L., Pecina, P., van Genabith, J., & Toral, A. (2010). Automatic extraction of arabic multiword expressions.

[4] Basili, V. R. & Boehm, B. 2001. Cots-Based Systems Top 10 List. *Computer* 34(5): 91-95.

[5] Baroni, M. & Bernardini, S. 2004. Bootcat: Bootstrapping Corpora and Terms from the Web. *Proceedings of LREC*, hlm. 1313-1316.

[6] Bounhas, I. & Slimani, Y. 2009. A Hybrid Approach for Arabic Multi-Word Term Extraction. *Natural Language Processing and Knowledge Engineering, 2009. NLP-KE 2009. International Conference on*, pp. 1-8.

[7] Cohen, J. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and psychological measurement* 20(1): 37-46.

[8] Church, K. W. & Hanks, P. 1989. Word Association Norms, Mutual Information, and Lexicography. *Proceedings of the 27th annual meeting on Association for Computational Linguistics*,pp. 76-83.

[8] Chung, T.M. (2003). A corpus comparison approach for terminology extraction.*Terminology*, *9*(2), 221-246.

[9] Chakraborty, T., Even-Dar, E., Guha, S., Mansour, Y. & Muthukrishnan, S. 2010. Approximation Schemes for Sequential Posted Pricing in Multi-Unit Auctions. Dlm. (pnyt.). *Internet and Network Economics*, pp. 158-169. Springer.

[10] Dichy, J. 2001. On Lemmatization in Arabic, a Formal Definition of the Arabic Entries of Multilingual Lexical Databases. *ACL/EACL 2001 Workshop on Arabic Language Processing: Status and Prospects.Toulouse, France*, pp

[11] Dandapat, S., Mitra, P. & Sarkar, S. 2006. Statistical Investigation of Bengali Noun-Verb (Nv) Collocations as Multi-Word-Expressions. *Proceedings of Modeling and Shallow Parsing of Indian Languages (MSPIL)* 230-233

[12] Dunning, T. 1993. Accurate Methods for the Statistic of Surprise and Coincidence. *Computational linguistics* 19(1): 61-74.

[13] Evert, S. 2005. The Statistics of Word Cooccurrences. Tesis Dissertation, Stuttgart University

[14] Frantzi, K. T. & Ananiadou, S. 1999. The C-Value/Nc Value Domain-Independent Method for Multi-Word Term Extraction.

[15] Frantzi, K., Ananiadou, S. & Mima, H. 2000. Automatic Recognition of Multi-Word Terms:. The C-Value/Nc-Value Method. *International Journal on Digital Libraries* 3(2): 115-130.

[15] Gao, J., Nie, J.-Y. & Zhou, M. 2006. Statistical Query Translation Models for Cross-Language Information Retrieval. *ACM Transactions on Asian Language Information Processing (TALIP)* 5(4): 323-359

[16] Hull, D. A., Bourigualt, D., Jacquemin, C. & L'homme, M. 2001. Software Tools to Support the Construction of Bilingual Terminology Lexicons. *D. Bourigault, C. Jacquemin and M.-C. L'Homme Recent Advances in Computational Terminology. Amsterdam/Philadelphia, John Benjamins* 225-244.

[17] Hajic, J., Smrz, O., Buckwalter, T. & Jin, H. 2005. Feature-Based Tagger of Approximations of Functional Arabic Morphology. *Proceedings of the Fourth Workshop on Treebanks and Linguistic Theories (TLT 2005)*, pp. 53-64.

[18] Katz, B., Bilotti, M. W., Felshin, S., Fernandes, A., Hildebrandt, W., Katzir, R., Lin, J. J., Loreto, D., Marton, G. & Mora, F. 2004. Answering Multiple Questions on a Topic froHeterogeneous Resources. *TREC*,pp.

[19] Kit, C. & Liu, X. 2008. Measuring Mono-Word Termhood by Rank Difference Via Corpus Comparison. *Terminology* 14(2): 204-229

[20] Salton, G. & Buckley, C. (1988). Term-weighing approache sin automatic text retrieval. In Information Processing & Management, 24(5): 513-523.

[21] Saif, A. M. & Aziz, M. J. 2010. An Automatic Collocation Extraction from Arabic Corpus. *Journal of Computer Science* 7(1): 6-11.

[22] Xu, F., Kurz, D., Piskorski, J. & Schmeier, S. 2002. A Domain Adaptive Approach to Automatic Acquisition of Domain Relevant Terms and Their Relations with Bootstrapping. *Proc. of LREC*.

[23] Zhang, X. 2011. Enhanced Term Extraction Based on Probabilistic Estimation from Syntactic Parse Tree

[24] Chen, J., Yeh, C.-H. & Chau, R. 2006. A Multi-Word Term Extraction System. Dlm. (pnyt.). *Pricai 2006: Trends in Artificial Intelligence*, pp. 1160-1165.Springer

[25] Sanjuan, E., Dowdall, J., Ibekwe-Sanjuan, F. & Rinaldi, F. 2005. A Symbolic Approach to Automatic Multiword Term Structuring. *Computer Speech & Language* 19(4): 524-542.