IUG Journal of Natural and Engineering Studies Vol 22, No 2, 2014, pp 30-44, ISSN 1726-6807
Peer-reviewed Journal of Islamic University-Gaza IUGNES

Automatic Domain-Relevant Collocation Extraction from Arabic Corpus

Rebhi S. Baraka[1,*], Manar S. Fayyad[2]

[1]Faculty of Information Technology, Islamic University of Gaza, Gaza Strip, State of Palestine

[2]Info. and Comm. Technology Center, Al Quds Open University,Gaza Strip, State of Palestine

Received on (20-10-2013) Accepted on (8-1-2014)

Abstract

An approach for automatic domain-relevant collocation extraction from Arabic text corpus is proposed. It uses naïve linguistic and statistical methods to extract collocations and relate them to specific domains depending on prevalence and tendency collocation ranking mechanism. In order to realize the proposed approach we use a corpus separated into ten domains. The proposed approach starts with preprocessing this corpus, then extracting candidate collocations. After that, it ranks the candidate collocations depending on the distributional behavior of candidate collocations within the domain and across the rest of the corpus. Then we distribute the candidate collocations over the domains depending on their rank values to get domains' term matrix. Finally, we evaluate the resulting collocation matrix by using it to classify the domain of a number of documents. The results are encouraging in most domains such that the achieved rate of accuracy exceeded 90%.

Keywords Preprocessing, Stemming, light stemming, Arabic Collocation Extraction, Collocations, Domain-Relevant Collocation Extraction.

استخراج المصطلحات العربية المرتبطة بمجال معين من مجموعة نصوص عربية آلياً

ملخص

تم اقتراح طريقة آليةلاستخراج المصطلحات العربية المرتبطة بمجال معين من مجموعة نصوص عربية. استخدمت هذه الطريقة الأساليب اللغوية والإحصائية لاستخراج المصطلحات ذات الصلة بمجال محدد وإسنادها إلى هذا المجال. من أجل تحقيق الطريقة المقترحة استخدمنا مكنزاً (Corpus) عربيا مقسما إلى عشرة مجالات. الطريقة المقترحة تقوم بمعالجة هذه المستندات معالجة لغوية خفيفة (Light stemming) ثم تستخرج المصطلحات المرشحة. بعد ذلك يتم تقييم كل مصطلح من المصطلحات المرشحة (Candidate terms) بناءً على مدى انتشار المصطلح داخل المجال المحدد وخارجه ومدى ارتباطه بهذا المجال. ومن ثم يخصص المصطلح المرشح للمجال ذو الوزن الأكبر لنحصل بعدها على مصفوفة مصطلحات المجالات (Domains term matrix).ولاختبار مدى فاعلية هذه الطريقة تم استخدام هذه المصفوفة في عملية تصنيف بعض المستندات أو النصوص وتحديد مجالاتها مع العلم بان مجالاتها كانت محددة مسبقا وقد تم تصميم مصنف يعتمد على مصفوفة مصطلحات المجال وكانت النتائج ممتازة في أغلب المجالات بحيث حققت نسبة دقة تجاوزت 90%.

كلمات مفتاحية: المعالجة الأولية، التجذير، التجذير الخفيف، استخراج المصطلحات العربية المنتظمة، استخراج المصطلحات المنتظمة، استخراج المصطلحات المنتظمة ذات الصلة بمجال.

* Corresponding author e-mail address: rbaraka@iugaza.edu.ps

1. Introduction

Collocation is any text sequence that appears together frequently and collocation extraction is the process of extracting collocations automatically from a corpus[1]. Automatic domain relevant collocation extraction from text corpus is rather important in natural language processing studies and applications. Moreover, it is an essential component in many lingual system models. The resulting collocations are commonly used in NLP tasks like information retrieval, text mining, and document summarization [1].

Any corpus participating in the collocation extraction process needs to be preprocessed by removing non letters, stop word, and others [2]. The main stages for collocation extraction are: the extraction of candidate collocations and the validating and ranking of these collocations [3].

There are several methods for extracting candidate collocations like linguistic filtering that uses the linguistic patterns like "N ADJ, N N, and N PREP N" for filtering the tagged corpus [4]. Also the noun phrase which take any sequence of word following a noun can be used [3]. Other approaches use a local grammar approach that uses a role for extracting a collocation like the telling role in [5]. The n-gram sliding window could be used for candidate collocation extraction with n length [6, 7].

There are several ranking methods for validating the extracted collocation. They Are categorized into two categories, Unithood and Termhood (TH) [8]. Unithood is the degree of strength or stability of syntagmatic combinations and collocations [9]. It is calculated only for complex collocations. Samples of these measures are T-Score, Normalized Google Distance(NGD), mutual information and log-likelihood, and relies simply on the occurrence and co-occurrence frequencies from domain corpora as the source of evidence [10]. On the other hand Termhood measures the degree to which these stable lexical units are related to domain-specific concepts like C-value, NC-value, TF/IDF and others [11]. Some ranking methods use both of them like Termhood.

Our aim in this paper is to develop an approach for automatic domain-relevant collocation extraction from Arabic multiple domains corpus. Existing works concentrate on a single domain [4]. This approach depends on the prevalence and tendency measures for ranking the extracted candidate collocation on the target domain and across the rest of the corpus. We expect to have pure domain-relevant collocations matrix as an output of the approach. This matrix could be helpful in classifying documents, automatic library indexing, and other lingual application. Depending on the type of the corpus, this approach could be used in generating spam mail matrix for spam mail detection.

The rest of this paper is organized as follows: Section 2 presents some related works, Section 3 presents the detailed development of the approach, Section 4 presents and discusses the results, and Section 5 concludes the paper and suggests future work that improves the developed approach.

2. Related Works

A lot of work in the field of domain-relevant term extraction is done in non-Arabic languages. For example ExATOlp [12] is a software tool that extracts domain-relevant terms of syntactic annotated corpus. It uses both linguistic and statistical approaches to extract and select significant terms from a domain represented by the annotated corpus. The system extracts the noun phrases from xml documents and counts the iteration of each phrase and save a list of them.

Term Extractor [13] is a high-performing technique for automatic extraction of shared terminology from available documents in a given domain. It identifies relevant terms based on two steps: first, a linguistic processor is used to parse text and extract typical terminological structures, like compounds, adjective-noun and noun preposition noun sequences. Then, the list of terminological candidates is purged according to domain pertinence, domain consensus, lexical cohesion, structural relevance and miscellaneous filters to give a list of terms.

For the Arabic language several works are available for term extraction, but little work is done in the domain-relevant term extraction. A few approaches for single domain as well as for multiple domains automatic term extraction is done. These works mostly use what is called Field Association (FA) to classify terms related to a specific domain [14]. The pre-processing step is very important in the Arabic language because it is highly inflectional. Moreover, special stemmer is designed depending on the topic of the research and the methods that are used. In information retrieval light stemming is widely used to keep the information value within the terms and words [15-17].

In building a word vector, Al-shalabi and Kanaan [18] designed and implements a system for building an Arabic lexicon. The stemming process they use is likely more accurate. Other light stemmer approaches like the one tested in [19] have low results, and the tool proposed by [20] could be merged with other tools to enhance the preprocessing stage. In our work, we try to test several preprocessing methods to choose the best for our approach.

A multi-word term extraction program for Arabic language is designed in [21]. They take into consideration the linguistic specifications of Arabic word like graphical, inflectional, morpho-syntactic and syntactic variants. They use the N ADJ, N1 N2 and N1 PREP N2 patterns. They rank the Multi Word Term like (MWT-like) units by means of statistical techniques, Log-Likelihood Ratio (LLR), Mutual Information (MI) and t-scores.

A two-step approach is proposed in [16] for extracting candidate MWTs. First, using a Part of Speech (POS) linguistic filter to extract candidate MWTs then using a bigram compound noun patterns. Second, they assign each candidate MWT a score depending on the combination of both the C-value ranking method and the LLR ranking method [22-24].

A model for automatic Collocation Extraction is proposed in[4]. They define collocation as*: "A word combination whose semantic and/or syntactic properties cannot be fully predicted from those of its components and which therefore has to be listed in a lexicon"*. They use the following structural patterns of Arabic collocation (N+N, N+ADJ, V+N, V+ADV, ADJ+ADV and ADL+N). They use the joint tagging and segmenting algorithm that used for Arabic tagging by [25] and produce a bigram collocation depending on POS and previous patterns. Then, they select four association measures (LLR, Chi-Square (X^2), MI, and Enhanced Mutual Information (EMI)). They conclude that the log-likelihood ratio clearly outperforms the other association measures.

Most of the works reviewed above are dealing with one domain. This could give a false indicator of the relation between the term and the domain. On the other hand, the number of domains in the corpus increases the representatives of the extracted terms for the domains. The number of the domains increases the probability of the term to appear in several domains and competition of the domains for the term increases. Moreover these works depend on dedicated patterns for extracting candidate terms. This could exclude a large number of terms that might have a significant relation to the domain. They use ranking methods that quantify the term depending on one domain. These approaches for term candidate ranking might be inappropriate for multi domain corpus. Ranking candidate terms should depend on both domain and cross domain.

3. Development of the Approach

The approach is shown in Figure 1 and consists of the following stages:

1. Select a corpus suitable for our research.
2. Preprocess the corpus, extract candidate collocation, and count the iteration,
3. Rank the candidate collocations, and
4. Distribute the ranked collocations over the domains.

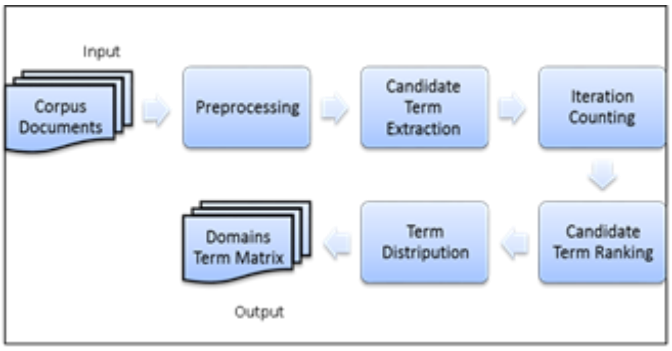

Figure 1 *The primitive model*

Next, we elaborate each of these stages.

3.1 Corpus selection stage

The approach should facilitate the extraction of the domain relevant collocations from Arabic corpus. Therefore, it needs to handle a corpus with the following properties:

– A big corpus that could give a good distributional behavior for the collocations.
– The corpus should be separated into domains.

There are several corpuses on the Internet which can beused for collocation extraction. We review them depending on the above properties and select the OSAC corpus [26]. The corpus collected from various websites and includes 22,429 text documents distributed over ten domains: Economics, History, Entertainments, Education and Family, Religious and Fatwa's, Sports, Health, Astronomy, Law, Stories, and Cooking Recipes). The corpuses contain around 18,183,511 (18M) words and 449,600 keywords after stop words removal.

3.2 Preprocessing, collocation extraction, and iteration counting stage

The second stage in the approach is preprocessing, collocation extraction, and iteration counting. It is clarified by tracing the following example:

"BBC Arabic مِنْ المنتظر أَنْ يَكْتَمِلَ مَشْرُوعُ خَطٍّ أنابيب نابوكو ، الْبالِغَ طُولَهُ 3300 كيلومترا."

As shown in Figure 1 this stage consists of three processes (Preprocessing, collocation extraction, and iteration counting)beginning with preprocessing which uses light stemmer that removes diacritics, punctuations, none Arabic letters, the definite article, and stop words as shown in Table 1. This process is repeated for all the domains.

The stemmed word vector matrix then passes to the candidate collocations extraction process, which extracts the collocations from the stemmed word vector depending on a sliding window with length from one to four saving them to candidate collocation vector matrix as shown in Table 2. This process is also repeated for all the domains.

After that, we count the iteration of each candidate collocation and the number of document the candidate collocation appears in. The result of applying this process on the previous example is shown in Table 3.

Where CI in Table 3 is the number of times the collocation appears in the corpus whereas DI in Table 3 is the number of document where the collocation appears in. 1WC, 2WC, 3WC, and 4WC are the collocation lengths from one to four words. This process is repeated for all domains.

Table 1 *Results of the preprocessing step.*

original text	Remove definite article	Remove diacritics	Remove punctuation	Remove non Arabic letters	Remove stop words
BBC	BBC	BBC	BBC		
Arabic	Arabic	Arabic	Arabic		
مِنْ	مِنْ	من	من	من	
المنتظر	منتظر	منتظر	منتظر	منتظر	منتظر
أَنْ	أَنْ	أن	أن	أن	
يَكْتَمِلَ	يَكْتَمِلَ	يكتمل	يكتمل	يكتمل	يكتمل
مَشْرُوعُ	مَشْرُوعُ	مشروع	مشروع	مشروع	مشروع
خَطَ	خَطَ	خط	خط	خط	خط
أنابيب	أنابيب	أنابيب	أنابيب	أنابيب	أنابيب
نابوكو	نابوكو	نابوكو	نابوكو	نابوكو	نابوكو
،	،	،			
الْبالِغَ	ْبالِغَ	بالغ	بالغ	بالغ	بالغ
طُولَهُ	طُولَهُ	طوله	طوله	طوله	طوله
3300	3300	3300	3300	3300	
كيلومترا	كيلومترا	كيلومترا	كيلومترا	كيلومترا	كيلومترا

Table 2 *Collocation Extraction with one to four words*

One word collocation	Two words collocation (2WC)	Three words collocation (3WC)	Four words collocation (4WC)
منتظر	يكتمل مشروع	يكتمل مشروع خط	يكتمل مشروع خط أنابيب
يكتمل	مشروع خط	مشروع خط أنابيب	مشروع خط أنابيب نابوكو
مشروع	أنابيب نابوكو	خط أنابيب نابوكو	
نابوكو	بالغ طوله		
بالغ			
طوله			
كيلومترا			

Table 3 *The iteration matrix for economy domain*

1WC	CI	DI	2WC	CI	DI	3WC	CI	DI	4WC	CI	DI
منتظر	64	57	يكتمل مشروع	1	1	مشروع خط أنابيب	1	1	مشروع خط أنابيب نابوكو	1	1
يكتمل	9	7	مشروع خط	5	5	خط أنابيب نابوكو	1	1			
مشروع	937	336	أنابيب نابوكو	1	1						
نابوكو	2	2									
بالغ	227	191									
طوله	7	5									
كيلومترا	9	7									

3.3 CANDIDATE COLLOCATION RANKING STAGE

Candidate collocation ranking is the third stage of this approach. In this stage, we give a value for each candidate collocation and this value is used in the evaluation of the relevancy of the collocation to the domain. These values are stored in a matrix with two columns for each domain; one for the collocation and the other for the rank value. The ranking methodology used is due to Wilson Wong et. al. [11, 27,28] and is stated as follows:

The termhood of a collocation a (TH(a)) is the final ranking value of the collocation and is stated in Equation 1. The rank value depends on the candidate evidence in the form of the discriminative weight of the collocation ($DW(a)$ Equation 1) and the adjusted contextual contribution of this collocation ($ACC(a)$ Equation 7) contextual evidence.

The discriminative weight is measured in Equation 2. As shown in the equation, this measure depends on cross-domain distributional behavior (domain tendency of the collocation$DT(a)$ and intra-domain distribution domain prevalence of the collocationDP(a)).

The domain tendency of the collocation is measured depending on the frequencies of a collocation within the domain and frequencies of a collocation outside the domain as shown in Equation 3.

The domain prevalence of the collocation depends on the collocation itself for simple collocation (one word collocation). It is measured using Equation 4 and for complex collocation (more than one word collocation) it is measured using Equation 5. The prevalence for simple collocation is measured depending on the frequencies of the collocation over the domain and across the rest of the corpus and the total frequencies of it to the total collocations iterations.

The prevalence for complex collocation depends on the prevalence for the header of the collocation and the value of the modifier evidence of the collocation.

The modifier evidence of collocation (in the form of modifier factor) is calculated using Equation 6. The modifier factor depends on the summation of frequencies of all the modifiers of the collocation over the domain and across the rest of the corpus.

The adjusted contextual contribution of the collocation ACC(a) as contextual evidence is calculated using Equation 7 where the adjusted contextual contribution depends on the adjustment of the contextual discriminative weight and the discriminative weight itself.

The adjusted contextual discriminative weight of the collocation ACDW(a) is calculated using equation 8 and it depends on discriminative weight of all the context words of the collocation and the similarity between the collocation and its context words Equation 9.

The similarity is calculated using Google normalized distance (NGD(a, c)) as stated in Equation 10. It depends on the number of the documents the collocation and its context words appear within.

Table 4 shows the result of applying this ranking methodology on the example sentence.

From the preceding, we find that the ranking method we use quantifies the three types of linguistic evidences derived from the prevalence and tendency measures in the form of *candidate evidence*, *modifier evidence*, and *contextual evidence*. In addition, to adjust the contribution of the contextual weight to the overall termhood, we can employ two measures, the *adjusted contextual contribution* and the *Normalized Google Distance*.

$$TH(a) = DW(a) + ACC(a) \quad (1)$$

$$DW(a) = DP(a)DT(a) \quad (2)$$

$$DT(a) = \log_2\left(\frac{f_{ad}+1}{f_{a\bar{d}}+1}+1\right) \quad (3)$$

Where f_{ad} represents the frequencies of a collocation within the domain and $f_{a\bar{d}}$ represents the frequencies of a collocation outside the domain.

$$DP(a) = \log_{10}(f_{ad}+10)\log_{10}\left(\frac{F_{TC}}{f_{ad}+f_{a\bar{d}}}+10\right) \quad (4)$$

$$DP(a) = \log_{10}(f_{ad}+10)\,DP(a^h)MF(a) \quad (5)$$

Where F_{TC}represents the frequencies summation of all collocations,f_{ad}represents the frequencies of a collocation within the domain,$f_{a\bar{d}}$ represents the frequencies of a collocation outside the domain,$MF(a)$ is the modifier factor, and$DP(a^h)$is the domain prevalence of the collocation header.

$$MF(a) = \log_2\left(\frac{\sum_{m\in M_a\cap TC} f_{m\bar{d}}+1}{\sum_{m\in M_a\cap TC} f_{m\bar{d}}+1}+1\right) \quad (6)$$

Where Ma represents all the modifiers of collocation a, and TC is the entire collocation candidate.

$$ACC(a) = ACDW(a)\left(\frac{e^{\left(1-\frac{ACDW(a)+1}{DW(a)+1}\right)}e^{\left(1-\frac{DW(a)+1}{ACDW(a)+1}\right)}}{\log_2\frac{ACDW(a)+1}{DW(a)+1}+1}\right) \quad (7)$$

Where $ACDW(a)$ is the average contextual discriminative weight and$DW(a)$ is the discriminative weight.

$$ACDW(a) = \left(\frac{\sum_{c\in C_a} DW(c)*sim(a,c)}{|C_a|}\right) \quad (8)$$

$$sim(a,c) = 1 - NGD(a,c)*\text{è} \quad (9)$$

Where $C_a$represents all the context words of collocation a, $|C_a|$ is the number of words,$sim(a,c)$ is the similarity between a and$c$, and θ is a constant for scaling the distance value of NGD (Normalized Google Distance).

$$NGD(x,y) = \frac{\max\{\log f(x), \log f(y)\} - \log f(x,y)}{\log \mathrm{M} - \min\{\log f(x), \log f(y)\}} \quad (10)$$

Where M is the total number of documents $f(x)$, $f(y)$is the number of document x, y appears in and $f(x,y)$ is the number of document booth x and y appears in.

Table 4 *Collocation ranking matrix for one domain*

One word collocation	Rank value	Two word collocation	Rank value	Three word collocation	Rank value	Four word collocation	Rank value
منتظر	0	يكتمل مشروع	0	مشروع خط أنابيب	137	مشروع خط أنابيب نابوكو	137
يكتمل	0	مشروع خط	263	خط أنابيب نابوكو	57		
مشروع	0	أنابيب نابوكو	318				
نابوكو	161	بالغ طوله	0				
بالغ	0						
طوله	0						
كيلومترا	0						

3.4 COLLOCATION DISTRIBUTION STAGE

The fourth stage in this approach is collocation distribution over the domains. This process is done by assigning each collocation in the candidate collocation matrix to a specific domain depending on the rank value of the collocation. This process is needed to construct a matrix for domain collocations used in a classifier for testing the accuracy of collocation extraction approach.

In this stage, we use a simple method for collocation distribution. If the collocation exists in several domains, we put the collocation in the domain with the highest rank value and remove it from the other domains. This is because the domain with the highest rank value usually includes this collocation. The collocation is removed from the other domains with lower ranks because it is most likely not included in these domains.

Depending on the same example above and after ranking the example collocation vector to the ten domains we get rank values stated in Table 5.

It can be noticed from Table 5that rank values for domain 1 are higher than the other domains. This is due to taking the example from domain 1.

In addition, we notice that there are some collocations that do not appear in the other domains. These collocations are marked as (no). Other collocations with rank value0 indicates that the collocation is weakly relevant to the domain. Value 0 indicates that the collocation seldom occurs in the domain and is not an essential part of it.

Some collocations are ranked over several domains, like "بالغ"is ranked for domain 4 and domain 5. The winning domain is the domain with higher rank value. Some domains do not rank any collocation of the example although they exist in the candidate collocations of the domain. This means that not all the collocations of the example are related to these domains.

Collocations like "منتظر"are not ranked in its original domain (domain 1) and are ranked in another domain (domain4). This means that the collocation is strongly related to the other domain.

The complex collocations (collocations with two and more words) are stronger than the collocations with one word to the target domain because complex collocations are less frequent than simple collocations. Finally, the strongest relation between a collocation and a domain are found in complex collocations. The resulting collocation matrix is shown in Table 6. It shows that terms are placed under their respective domains based on their rank values, e.g., collocations like "نابوكو" and "مشروع خط" are placed under domain 1 because their rank

values under domain 1 are 161 and 263 respectively while they do not appear in the other domains.

Table 5 *Rank results of candidate collocations from the sample over the domains*

Collocation	Rank values for the domains									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
منتظر	0	0	0	161	0	no	no	0	0	no
يكتمل	0	0	0	0	0	0	no	0	0	no
مشروع	0	0	0	0	0	0	0	0	0	no
نابوكو	161	no	0	no	no	no	no	no	no	no
بالغ	0	0	0	161	128	0	0	0	0	0
طوله	0	0	0	0	0	0	0	0	0	0
كيلومترا	0	0	0	no	0	0	0	no	0	no
يكتمل مشروع	0	no	no	no	no	no	no	no	no	no
مشروع خط	263	no	no	no	no	no	no	no	no	no
أنابيب نابوكو	318	no	no	no	no	no	no	no	no	no
بالغ طوله	0	0	no	no	no	no	no	no	no	no
مشروع خط أنابيب	137	no	no	no	no	no	no	no	no	no
خط أنابيب نابوكو	57	no	no	no	no	no	no	no	no	no
مشروع خط أنابيب نابوكو	137	no	no	no	no	no	no	no	no	no

Table 6 *Sample of Domain Collocation Matrix*

1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	
نابوكو			منتظر						
بتكلفة			بالغ						
مليارات									
مشروع خط									
أنابيب نابوكو									
مشروع خط أنابيب									
خط أنابيب نابوكو									
مشروع خط أنابيب نابوكو									

4. Results nd Discussion

We have implemented the approach and tested it over a real corpus. Table 7 shows the domains of the tested corpus from 0 to 9, the word vector size for each domain, and the size of the candidate collocation for each collocation length (1 to 4).

Table 8 shows the domains of the tested corpus from 0 to 9, the word vector size for each domain, and the size of the domain collocation for each collocation length (1 to 4).

We can conclude from Tables 7 and 8 that the one word collocation is less relevant to the domains than collocations that contain two, three, and four words length. Figure 2 depicts this effect and shows the results of candidate collocations compared to the distributed collocations for one word length.

Figure 3 depicts the domain relevancy with collocation size effect. It represents the economy domain and this effect is true for the other domains. As shown in the graph when the size of the collocation increases the excluded collocations reduced.

We have tested the classifier on the testing corpus that contains 4670 document distributed into ten domains (0 to 9) as shown in Table 9.

The results of the classification process are described in the confusion matrix in Table 10. The confusion matrix is used for evaluating the performance of a system using the data in the matrix. The confusion matrix contains information about actual and predicted classifications done by a system [29].

As shown in Table 10 the numbers from 0 to 9 represent the domains. REL represents the reliability of the classifier to classify the document domain. ACC represents the accuracy of the classifier to classify the document domain. For example the result of calculating the ACC of domain 6 is as follows:

Astronomy domain		predicted	
		negative	positive
actual	Negative	4548	0
	Positive	1	121

Accuracy	ACC=	0.9199786
Recall	R=	0.9041803
Specificity	TN=	1
Precision	REL =	1
	FP=	0
	FN=	0.008197
	G-mean1=	0.995893
	G-mean2=	0.995893

Rows 4,6, 7 and 9of Table 10 which represent the domains Sport, Astronomy, Law, and Cooking Recipes respectively could be classified with more than 90% reliability that this document is not a member of the other domains. This is due to the nature of these domains and the kind of words that are used in them.

The rest of the domains are also highly reliable except for the History domain (row 1). The History domain does not have unique collocations that could represent it clearly.

All the domains are highly accurate except the Stories domain(row 8) as the number of wrong classifications is high in respect to the Stories' tested documents. The majority of wrong classifications are due to History domain as the Stories and History domains are close to each other.

Table 10 reveals that the classifier is reliable for classifying all the domains except for the History domain. It also reveals that the classifier is highly accurate for all the domains except for the Law domain (row 7).

The accuracy of the classifier is calculated using the following equation and it reaches 92%:

$$Accuracy = \frac{\sum_{i=1}^{n}(true\ classification)}{total\ number\ of\ cases}$$

Where i is the class number and n is the total number of classes.

When we have reviewed the corpus and the errors, we found that the errors are due to the weakness of the corpus because of the small number of websites the corpus is grabbed from.

Table 7 *Number of candidate terms for the domains*

Code	Domain	Word vector size	Collocation candidate size			
			1	2	3	4
0	Economics	1618618	63035	435188	442312	339321
1	History	3668139	154943	789543	627274	411164
2	Education & Family	2241672	122038	500072	383418	251896
3	Religious & Fatwa	1527183	58452	201014	160079	108847
4	Sports	1266928	47198	231434	235817	188543
5	Health	1490953	46942	157712	124271	84680
6	Astronomy	275469	22892	63914	52381	37312
7	Law	619292	28977	77772	61927	43573
8	Stories	2065902	101488	323691	230145	146663
9	Cooking Recipes	268387	14997	62530	68563	54507

Table 8 *Number of distributed terms over the domains*

Code	Domain	Word vector size	Domain collocation size			
			1	2	3	4
0	Economics	1618618	24281	400464	433403	333269
1	History	3668139	94630	728401	610012	401045
2	Education & Family	2241672	60425	447830	370287	244885
3	Religious & Fatwa	1527183	17281	170256	153029	105068
4	Sports	1266928	16623	209520	228048	181979
5	Health	1490953	16800	139945	119822	81926
6	Astronomy	275469	6079	53738	48490	34777
7	Law	619292	7316	66600	58972	41791
8	Stories	2065902	44111	282002	218437	140094
9	Cooking Recipes	268387	6594	56600	64668	51641

Table 9 *The number of documents to be classified for the domains*

Code	Domain	Number of documents
0	Economics	647
1	History	615
2	Education & Family	712
3	Religious & Fatwa	713
4	Sports	522
5	Health	425
6	Astronomy	122
7	Law	213
8	Stories	173
9	Cooking Recipes	528

Table 10 *The classifier confusion matrix for the domains*

		Real domain											
		0	1	2	3	4	5	6	7	8	9	**Sum**	**REL**
Classified domains	0	634	1	0	0	6	1	0	0	0	1	**643**	**0.99**
	1	13	583	8	35	22	1	1	6	35	2	**706**	**0.83**
	2	0	27	682	2	0	5	0	0	1	5	**722**	**0.94**
	3	0	1	10	676	0	0	0	0	0	0	**687**	**0.98**
	4	0	0	0	0	494	0	0	0	1	0	**495**	**1.00**
	5	0	1	0	0	0	418	0	0	0	15	**434**	**0.96**
	6	0	0	0	0	0	0	121	0	0	0	**121**	**1.00**
	7	0	0	0	0	0	0	0	207	0	0	**207**	**1.00**
	8	0	2	12	0	0	0	0	0	136	0	**150**	**0.91**
	9	0	0	0	0	0	0	0	0	0	505	**505**	**1.00**
	Sum	**647**	**615**	**712**	**713**	**522**	**425**	**122**	**213**	**173**	**528**	**4670**	
	ACC		**0.95**	**0.96**	**0.95**	**0.95**	**0.98**	**0.99**	**0.97**	**0.79**	**0.96**		

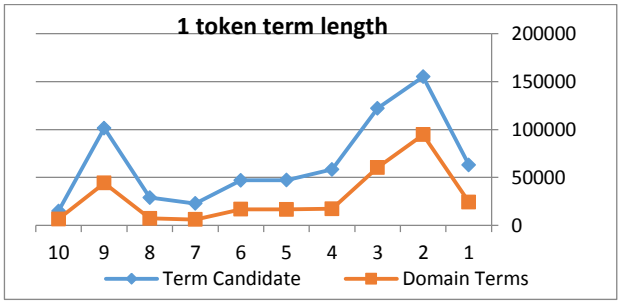

Figure 2 *Candidate collocations and distributed collocations for 1 word*

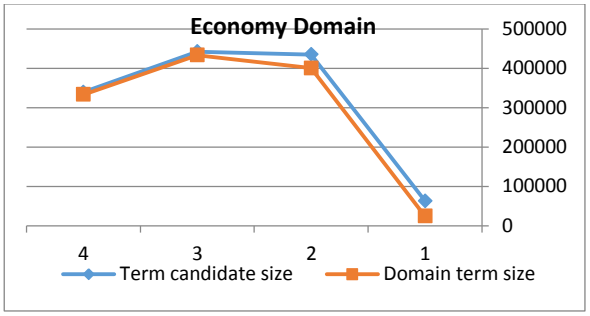

Figure 3 *Collocation candidate and domain collocations over collocation size for economy domain*

5. Conclusion and Future Work

We developed an approach for automatic domain-relevant collocation extraction from Arabic multiple domains corpus. The approach takes several criteria into consideration such as the specification of the corpus, the collocation extraction method used in extracting the candidate collocations. The ranking methodology used by the approach for ranking the collocations. The distribution method it uses for distributing collocations over the domains and the

evaluation methods and tools used for evaluating the approach.

It deals with several domains so the corpus should be separated into domains while most of existing works deal with general corpus and others with one domain specific corpus.

The approach uses the sliding window method for candidate collocation extraction while existing works deal with several methods for collocation extraction like NLP patterns, Local grammar approach, or syntactic patterns. We use this method because the other methods depend on the taggers and the existing taggers have low accuracy – nearly 25% of the words not identified by the tagger [25] which affect the accuracy of the models.

The ranking method we used depends on several domains which measure the collocation prevalence and tendency over the domain and across the rest of the corpus.

We used a simple method for collocation distribution over the domains to generate the domain relevant collocation matrix which depends on the ranking value for the collocation over all the corpuses and assign the collocation to the domain with high rank. Other works deal with one domain and this differentiation does not exist in other works.

Finally we designed the classifier depending on the domain relevant term matrix to classify a domain known document and use a confusion matrix for evaluating the approach.

There are several ways for improving our approach:

- Use several corpuses and study the effect of the corpus change on the results.
- In the preprocessing stage we could evaluate several preprocessing options and compare the effect of each option.
- In the collocation extraction stage we could use other methods for candidate collocation extraction like pattern passed, local grammar or other NLP methods and examine the approach for these options.
- For the collocation ranking stage we could experiment several ranking methods and compare the implementation results.

REFERENCES

[1] T. Vu, A. Aw, and M. Zhang, "Term extraction through unithood and termhood unification," in International Joint Conference on Natural Language Processing - IJCNLP, pp. 631–636, 2008.

[2] M. Syafrullah and N. Salim, "Improving Term Extraction Using Particle Swarm Optimization Techniques," JOURNAL OF COMPUTING, vol. 2, no. 2, pp. 116–120, 2010.

[3] R. Mitkov, G. Corpas, and others, "Mutual terminology extraction using a statistical framework," Procesamiento del lenguaje Natural, vol. 41, no. Section 2, pp. 107–112, 2008.

[4] A. Saif, M. Aziz, "An Automatic Collocation Extraction from Arabic Corpus," Journal of Computer Science, vol. 7, no. 1, pp. 6–11, 2010.

[5] J. Nam, "A Local-Grammar-based Approach to Recognizing of Proper Names in Korean Texts," in the 5th Workshop on Very Large Corpora (WVLC-5), pp. 273–288, 1997.

[6] J. Foo, "Term extraction using machine learning," Linköping University, LINKÖPING, 2009.

[7] S. Katz, "Distribution of content words and phrases in text and language modelling," Natural Language Engineering, vol. 2, no. 1, pp. 15–59, 1996.

[8] W. Wong, "Determining termhood for learning domain ontologies in a probabilistic framework," In Proceedings of the sixth Australasian conference on Data mining and analytics, vol. 07, pp. 51–60, 2007.

[9] J. Foo, "Exploring termhood using language models," in NEALT PROCEEDINGS SERIES, vol. 12, pp. 32–35, 2011.

[10] W. Wong, W. Liu, and M. Bennamoun, "Determining the unithood of word sequences using mutual information and independence measure," in Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics (PACLING), pp. 246-254, 2007.

[11] W. Wong, W. Liu, and M. Bennamoun, "Determining termhood for learning domain ontologies using domain prevalence and tendency," in Proceedings of the sixth Australasian conference on Data mining and analytics, vol. 70, pp. 47–54, 2007.

[12] L. Lopes, P. Fernandes, and R. Vieira, "ExATOlp-an automatic tool for term extraction from Portuguese language corpora," Proceedings of the fourth language and technology conference: human language technologies as a challenge for computer science and linguistics, LTC'09, pp. 427–431, 2009.

[13] F. Sclano, R. La, and P. Velardi, "Termextractor: a web application to learn the shared terminology of emergent web communities," Enterprise Interoperability II, pp. 287–290, 2007.

[14] E. Atlam, M. Fuketa, K. Morita, and J. Aoe, "Automatic building an extensive Arabic FA terms dictionary," Proceedings of World Academy of Science, Engineering and Technology, vol. 44, pp. 719–725, 2010.

[15] L. Larkey, L. Ballesteros, and M. Connell, "Light stemming for Arabic information retrieval," Arabic computational morphology, vol. 38,pp. 221–243, 2007.

[16] K. Al Khatib and A. Badarneh, "Automatic extraction of Arabic multi-word terms," Computer Science and Information Technology (IMCSIT), pp. 411–418, 2010.

[17] M. Beseiso, A. Ahmad, and R. Ismail, "A Survey of Arabic language Support in Semantic web," International Journal of Computer Applications IJCA, vol. 9, no. 1, pp. 24–28, Nov. 2010.

[18] R. Al-shalabi and G. Kanaan, "Constructing an automatic lexicon for Arabic language," international journal of computing &information sciences, vol. 2, no. 2, pp. 114–128, 2004.

[19] H. Al Ameed, S. Al Ketbi, A. Al-Kaabi, K. Al Shebli, N. Al Shamsi, N. Al Nuaimi, and S. Al Muhairi, "Arabic light stemmer: A new enhanced approach," The Second International Conference on Innovations in Information Technology (IIT'05), pp. 1–9, 2005.

[20] M. Saad and W. Ashour, "Arabic Morphological Tools for Text Mining," 6th International Symposium on Electrical and Electronics Engineering and Computer Science(EEECS'10). European, Lefke, North Cyprus, pp. 112–117, 2010.

[21] S. Boulaknadel, B. Daille, and D. Aboutajdine, "A multi-word term extraction program for Arabic language," In Proceeding of the Sixth LREC, pp. 1485–1488, 2008.

[22] T. Naseem and B. Snyder, "Multilingual part-of-speech tagging: Two unsupervised approaches," Journal of Artificial Intelligence Research, vol. 36, pp. 1–45, 2009.

[23] S. Mansour, K. Sima'an, and Y. Winter, "Smoothing a lexicon-based pos tagger for Arabic and Hebrew," Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources, 2007.

[24] G. Kanaan, R. AL-SHALABI, and M. Sawalha, "Full automatic Arabic text tagging system," proceedings of the International Conference on Information Technology and Natural Sciences, pp. 258–267, 2003.

[25] S. AlGahtani, W. Black, and J. Mcnaught, "Arabic part-of-speech tagging using transformation-based learning," in Proceedings of the Second International Conference on Arabic Language Resources and Tools, pp. 66–70, 2009.

[26] M. Saad and W. Ashour, "OSAC: Open Source Arabic Corpora," 6th International Symposium on Electrical and Electronics Engineering and Computer Science(EEECS'10). European, Lefke, North Cyprus, pp. 1-6, 2010.

[27] R. Basili, A. Moschitti, M. T. Pazienza, and F. M. Zanzotto, "A contrastive approach to term extraction," International Conference on Terminology and Artificial Intelligence (TIA-2001), 2001.

[28] A. Hippisley, D. Cheng, and A. Khurshid, "The head-modifier principle and multilingual term extraction," Natural Language, vol. 11, no. 2, pp. 129–157,2005.

[29] R. Kohavi and F. Provost, "Special Issue on Applications of Machine Learning and the Knowledge Discovery Process," Machine Learning, vol. 30, no. 2/3, pp. 127–271, 1998.