Proceedings of the International Multiconference on Computer Science and Information Technology pp. 411–418
ISBN 978-83-60810-27-9
ISSN 1896-7094

Automatic Extraction of Arabic Multi-Word Terms

Khalid Al Khatib
Department of Computer Science
Jordan University of Science and Technology
Irbid 22110, Jordan
Khalid_ikh@yahoo.com

Amer Badarneh
Department of Computer Information Systems
Jordan University of Science and Technology
Irbid 22110, Jordan
amerb@just.edu.jo

***Abstract*—Whereas a wide range of methods has been conducted to English multi-word terms (MWTs) extraction, relatively few studied have been applied to Arabic MWTs extraction. In this paper, we present an efficient approach for automatic extraction of Arabic 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 two statistical methods (log-likelihood ratio and C-value) are used to rank candidate MWTs. Many types of variations (e.g. inflectional variants) are taken into consideration to improve the quality of extracted MWTs. We obtained promising results in both coverage and precision of MWTs extraction in our experiments based on environment domain corpus.**

I. Introduction

AUTOMATIC term recognition (ATR) is still playing an important role in many Natural Language Processing (NLP) applications such as machine translation, book and digital library indexing, hypertext linking, and text categorization. The main objective of ATR is extracting domain specific terms from special language corpora [1].

One of the most important types of ATR is extraction of multi-word terms (MWTs); this comes from the advantages of using MWTs in machine translation, summarization, question answering systems, and many important computational linguistic applications. MWT can be defined simply as a group of words, which are consecutive and constitute a semantic unit [2].

There are three main approaches for extracting MWTs. The first one uses a linguistic filter that depends on syntactic patterns or MWT boundaries detection. The second approach uses a statistical filter to specify the probability of each sequence of words to constitute a MWT. The last one is a hybrid approach of the two previous approaches, this approach extracts candidate MWTs using a linguistic filter, and then it assigns each candidate MWT a score depending on some statistical methods [3].

Most of the statistical methods for MWT extraction concentrate on one of two features: the *unithood* which is "*the degree of strength or stability of syntagmatic combinations or collocation*" [4], and the *termhood* which is "*the degree to which a linguistic unit is related to domain-specific concepts*" [4]. Many methods have been proposed as a *unithood* measure such as mutual information [5], log-likelihood ratio (LLR) [6], and left/right entropy [7], while there is the C-value method [8] as an example of *termhood* measure.

In this paper, we adopt the hybrid approach to extract MWTs from an Arabic corpus. Needless to say that there is a rapid development of computational linguistic applications for Arabic language nowadays, Arabic is the official language of 22 countries, it is spoken by more than 200 million, and it has a very high esteem in the Muslim world [9]. The proposed approach includes two main steps: the linguistic filter and the statistical filter. In the first step, we propose syntactic patterns and use simple part of speech (POS) tagger to extract candidate MWTs. In the second step, two statistical methods are used to rank the candidate MWTs: the LLR method and the C-value method. We consider some related issues like morphological and syntactic ambiguities. For evaluation purpose, we use an environment domain Arabic corpus, the results indicated that our approach is effective, and can be used in many related NLP applications efficiently.

The contribution of our work includes two main points: in the linguistic side, we made an enhancement to the syntactic patterns to be simple and able to exclude a number of wrong candidate MWTs. Moreover, in the statistical side, we take into account both *termhood* and *unithood* measures, since we use a combination between the LLR method and the C-value method in the ranking process.

The paper is structured as follows. Some of the related work is described briefly in section two. In section three we present our proposed approach to extract MWTs. Section four explains how the approach treats the term variations. Section five shows the experiments and the results of applying the extraction approach. The last section contains the conclusion and the future work.

II. Related Work

A lot of work has been done to extract MWT in many languages. This work has been proposed by using linguistic filters, statistical methods, or both as a hybrid approach. However, the majority of the latest MWT extraction systems have adopted the hybrid approach, because it has given better results than using only linguistic filters or statistical methods [10].

As far as we know, there are a few MWTs extraction systems of Arabic language, one of them is the work which has been presented by Attia, M. A. [11], he has adopted the linguistic approach by doing manual and semi-automatically extraction of Arabic MWTs.

In another work, Boulaknadel, S.; Daille, B.; and Aboutajdine, D. [12] have adopted the hybrid approach to extract Arabic MWTs. The first step of their system is extraction of MWT-like units, which fit the follow syntactic patterns: {noun adjective, noun1 noun2, noun1 preposition noun2} using available part of speech tagger, taken into consideration graphical, inflectional, morphosyntactic, and syntactic variants. The second step is ranking the extract MWT-like units using association measures, these measures are: log-likelihood ratio, FLR, Mutual Information (MI^3), and t-score. The evaluation process includes applying the association measures to an Arabic corpus and calculating the precision of each measure using a collected reference list of Arabic terms [12].

Another system has been proposed by Bounhas, I. and Slimani, Y. [13]; they have proposed a hybrid approach to extract compound nouns. In the linguistic side, they used matcher between POS tagger and morphological analyzer to produce sequences of tokens, each token could be represented by a number of solutions, and then using a syntactic parser to extract candidates of compound nouns. In the statistical side, they applied the LLR method. In the evaluation step, they used almost the same corpus and reference list which have been used in [12]. Their results were promising especially with bigram MWTs [13].

III. Proposed Approach

The proposed MWTs extraction approach has the following features: (i) the system is simple as much as possible to avoid performance and complexity problems, (ii) accurate since there are previous systems [12] and [13] which have got good results, and (iii) able to cover the importance of MWT to increase the possibility for using it in other NLP systems such as summarization and machine translation systems.

The proposed approach for extracting Arabic MWTs is composed of two main steps: (i) the linguistic filter, where we extract candidate MWTs, and extract bigrams from candidate MWTs (ii) the statistical filter, where we rank bigrams by the LLR and C-value scores. In the following subsections, we cover the two steps in more details.

A. Linguistic Filter

There are many types of MWT, such as idioms, phrasal verbs, verbs with particles, compound nouns, and collocations [13]. Our choice was similar to [12] and [13] where we chose to deal with compound nouns, since we agree with the fact that nouns can represent document's subject efficiently [13]. To extract candidate MWTs, we left the syntactic patterns which have been used in many systems like [12], and propose new patterns based on definite and indefinite types of nouns. Table 1 and Figure 1 show our syntactic patterns.

Table 1.
Syntactic patterns of MWT

(1)	definite noun ⇒ one or more definite nouns
(2)	indefinite noun ⇒ one or more indefinite nouns
(3)	indefinite noun ⇒ one or more definite nouns
(4)	(1) or (2) or (3) ⇒ preposition ⇒ (1) or (2) or (3)

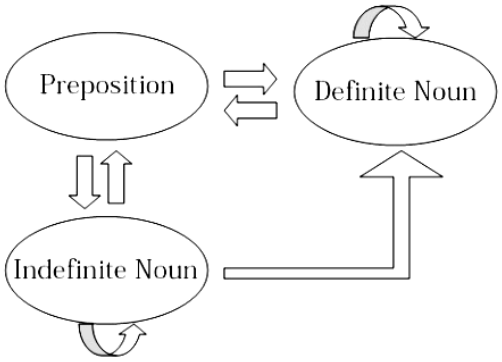

Fig.1: Graphical model of syntactic patterns

These syntactic patterns have some advantages over the patterns used in other extraction systems such as [12]. Firstly, it doesn't require an advanced part of speech tagger. It needs simple one with just three categories (noun, verb, particles). Obviously, this means better performance because most of the POS taggers still have problems in differentiation between nouns and adjectives [11], and less complexity since we don't need many words' classes such as adjectives and adverbs, which advanced POS taggers try to determine. Secondly, these patterns include the entire correct candidate MWTs which might be extracted by patterns in [12], and exclude a collection of wrong candidate MWTs which patterns in [12] may extract. Table 2 shows examples of MWTs extracted from proposed patterns and patterns proposed in [12].

The extraction of candidate MWTs starts with the preprocessing step, which includes four sub steps. The first one is the tokenization, where we separate text into main tokens (words). Words are always separated by white spaces or punctuation marks in Arabic language. The second sub step is the stemming, where the stem of each word is extracted using an available stemmer proposed by khoja, S. [14].

TABLE 2.
EXAMPLES OF MWTS EXTRACTED FROM PROPOSED PATTERNS AND PATTERNS PROPOSED IN [12]

Candidate MWT	New patterns	Patterns used in [12]	Is it correct MWT?
التنوع الحيوي biodiversity	yes	yes	yes
تقطير الماء water distillation	yes	yes	yes
تلوث حراري thermal pollution	yes	yes	yes
السماء غائمة the sky is cloudy	no	yes	no

This stemmer specifies the word's stem and type, the types which the stemmer can define are {stemmed word, stop word, strange word}, the stemmer has its own list of stop words which we modify through adding additional stop words. The third sub step is the frequency calculation. In this sub step, we calculate the frequency of each word as well as the frequency of each stem. This sub step is important in dealing with variations. The last sub step is the sentence segmentation. We use simple method for this purpose, where special punctuation marks are used to determine the boundaries of the sentences.

The second step after the preprocessing one is the word's classification (by means of Simple POS tagger). There are three main classes of the word in Arabic: noun, verb, and particle. What we care about in this step is distinguishing the nouns from other classes, since our patterns primarily depend on nouns. Although the available POS taggers can help us very well in this step, we decided to ignore them and adopted the approach which has been proposed by Ahmad T. and Salah A. [15].There are two reasons for that. First, this approach is simple and accurate. Therefore, it is able to keep one of the merits of our syntactic patterns, which is the simplicity. Second, this approach has a morphological analyzer phase. This phase is helpful on dealing with term variations.

The architecture of the adopted approach for words' classification contains three main phases. The first phase is the lexicon analyzer. In this phase a lexicon of stop lists in Arabic language is defined. This lexicon includes prepositions, adverbs, conjunctions, interrogative particles, exceptions, questions and interjections. All the words have to pass this phase, if the word is found in the lexicon, it is considered as tagged to one of the previous closed lists. The next phase is the morphological analyzer. Each word which has not been tagged in the previous phase will immigrate to this phase. In this phase, firstly, the affixes of each word are extracted, the affix is a set of prefixes, suffixes and infixes. After that, these affixes and the relation between them are used in a set of rules to tag the word into its class. It is important to say that this phase is the core of the system, since it distinguishes the major percentage of untagged words into nouns or verbs. The last phase is the syntax analyzer. This phase can help in tagging the words which the previous two phases failed to tag. It is consisting of two rules: sentence context and reverse parsing. The sentence context rule is based on the relation between the untagged words and their adjacent, where Arabic language has some types of relations between adjacent words. These relations can help in tagging the words into its corresponding class. The reverse parsing rule is based on Arabic context-free grammar. There are ten rules, which are used frequently in Arabic language.

The Third step to extract the candidate MWTs is extraction of sequences of nouns, as well as sequences of nouns that connected by a preposition. In this step, we consider each sentence as a separated unit, and using the words' classification approach to extract sequences of nouns. For the sequences of nouns that connected by prepositions, we had two types of prepositions: prepositions which constitute a separated word and prepositions which are stuck with another word. We deal with the two types because our syntactic patterns consider prepositions from the two types. Table 3 shows examples of prepositions' types. Table 4 shows examples of extracted sequences of nouns.

TABLE 3.
EXAMPLES OF PREPOSITIONS' TYPES

separated proposition	من	e.g. التخلص من النفايات disposal of wastes
stuck preposition	ب	e.g. الري بالتنقيط drip irrigation

TABLE 4.
EXAMPLES OF EXTRACTED SEQUENCES OF NOUNS

sequence of nouns	e.g. منظمة الأرصاد الجوية العالمية world meteorological organization
sequences of nouns that connected by a preposition	e.g. التحكم عن بعد remote control
	e.g. التعبير بالإشارة the expression by reference

The last step is testing each extracted sequence based on MWTs syntactic patterns, the sequences which fit the patterns will be considered as candidate MWTs. Figure 2 shows the main steps for extraction of candidate MWTs using the linguistic filter.

MWT might be classified based on the number of words. Bigram term is the term of two words. We decide to consider bigrams and discard the other terms which consist of more than two words. Simply, we noted from the terminology databases that the major percentage of compound nouns is bigrams.

In our work, we extract the bigrams from each candidate MWT. We noted that some bigrams are MWT while others are not. However, this is the last step before using the statis-

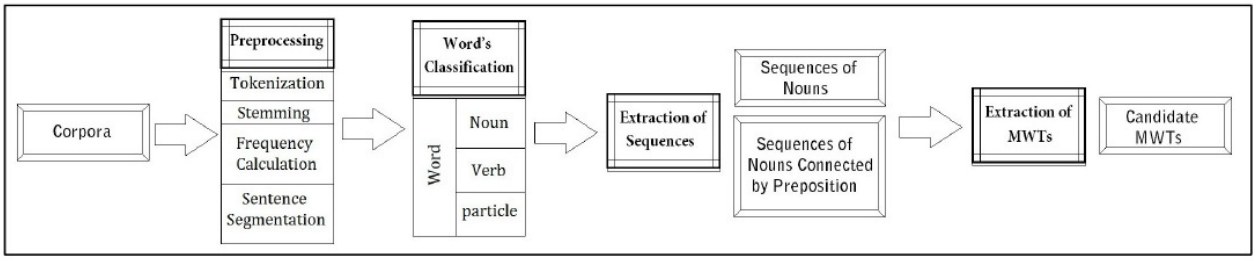

Fig. 2: The main steps for extraction of candidate MWTs using the linguistic filter

tical methods to rank the terms. Table 5 shows an example of bigrams' extraction.

TABLE 5.
EXAMPLE OF BIGRAMS' EXTRACTION

Candidate MWT	برنامج الولايات المتحدةلبحوث القطب الجنوبي united states Antarctic research program	
Bigrams	برنامج الولايات states program	NOT MWT
	الولايات المتحدة united states	MWT
	المتحدةلبحوث united for research	NOT MWT
	لبحوث القطب for Antarctic research	NOT MWT
	القطب الجنوبي Antarctic	MWT

B. *Statistical Filter*

Using statistical methods can help with morphological and syntactic ambiguities and therefore, increasing the quality and the quantity of correct extracted MWTs. In this step, we consider both *termhood* and *unithood* measures to get better results than using only one measure type [16].

To consider the *unithood*, we chose LLR method because it gives good results with Arabic MWTs extraction [12]. For the *termhood* we adopted C-value method because it has a wide acceptance as a valuable method to rank candidate MWTs [16]. LLR method can be used efficiently as significance of association measure between the two words in the bigram [17]. Regarding the C-value method, it requires simple modification to be able to rank the extracted bigrams. We list the entire candidate MWTs and the extracted bigrams as the first step, and then we apply the C-value equation only to the bigrams. Note that we would not be able to calculate the C-value score for the bigrams without some information about the candidate MWTs which contain those bigrams.

Practically, we make a list of bigrams ranked by the LLR. We make another list, which is ranked by the C-value method. Lastly, we combine the two lists to get a new list of-bigrams ranked by the two statistical methods. Figure 3 shows the Log-Likelihood ratio equations, and Figure 4 shows C-value method equation. Figure 5 shows the algorithm of proposed statistical filter.

Contingency table

	$V=v$	$V\neq v$
$U=u$	$O11$	$O12$
$U\neq u$	$O21$	$O22$

U: first word of the bigram. V: second word of bigram
$O11$: #compound nouns with U and V.
$O12$: #compound nouns with U but without V.
$O21$:#compound nouns with V but without U.
$O22$:#compound nouns without V and without U.

$$R1 = O11 + O12 \qquad C1 = O11 + O21$$
$$R2 = O21 + O22 \qquad C2 = O12 + O22$$
$$N = O11 + O12 + O21 + O22 = R1 + R2 = C1 + C2$$
$$LLR = -2\log\left\{\frac{L(O11, C1, r) * L(O12, C2, r)}{L(O11, C1, r1) * L(O12, C2, r2)}\right\}$$
$$L(k, n, r) = r^k \times 1 - r^{n-k}$$
$$r = \frac{R1}{N} \qquad r1 = \frac{O11}{C1} \qquad r2 = \frac{O12}{C2}$$

Fig.3: Log-Likelihood ratio equations

$$C-Value = \begin{cases} \log_2|a|.f(a) & \text{If a is not nested} \\ \log_2|a|\left(f(a) - \frac{1}{p(Ta)}\sum_{b\in Ta} f(b)\right) & \text{Otherwise} \end{cases}$$

a : The candidate MWT.
b : Longer candidate MWTs.
$|a|$: Length of the candidate MWT.
$f(a)$: Frequency of occurrence of a in the corpus.
Ta : Set of extracted candidate MWTs that contain a.
$P(Ta)$: Number of candidate terms in Ta.
$f(b)$: Frequency of occurrence of longer candidate MWT b in the corpus.

Fig.4: C-value method equation

Input : List of Candidate multi-word terms

output : List of Ranked Bigrams

Method :

Extract bigrams from List of Candidate multi-word terms
For each bigram b do{

Calculate log-likelihood ratio value{
Input: Bigram *b*
Output: LLR value of bigram *b*
Method: use equations in Fig.3
}

Calculate C-method value {
Input: List of Candidate multi-word terms
+
Bigram *b*
Output: C-method value of bigram *b*
Method: use equation in Fig.4
}

}

Sort (Bigrams, by LLR value, descending)
Sort (Bigrams, by C-method value, descending)
Make a list of bigrams sorted by LLR
where:
The index of the bigram represents its rank

Make a list of bigrams sorted by C-value method
where:
The index of the bigram represents its rank

Sort (Bigrams, by LLR rank+ C-value rank, descending)

return List of bigrams ranked by C-value method
+
log- likelihood ratio
where:
The index of the bigram represents its rank

Fig. 5: The statistical filter Algorithm

IV. Term Variations

When we try to extract MWT, term variation is one of the significant factors that should be studied, in other words, it is important to show how the proposed approach deals with different types of variations.

In our proposed approach, we started dealing with the term's variations in the statistical step. As we mentioned before, the input of the statistical step is the extracted bigrams which we try to rank using the statistical methods. Obviously, these methods use the frequency as a primary factor of weighting. What we did here is using the stem's frequency of nouns instead of word's frequency, it's clear that verbs are excluded from the stem's frequency calculating process.

To clarify the point, suppose we have the following bigrams which have graphical and inflectional variants:

(Environmental pollution)

تلوث البيئة	تلوث البيئه	بتلوث البيئة
لتلوث البيئة	تلوث البيئات	التلوث البيئي

The first word in all bigrams has the same stem, and we can say the same about the second word. This means that the statistical methods will consider these bigrams as identical, and it will give them the same score. After completing the list of ranked bigrams, an enhancement process will be available for this list; all the bigrams with graphical and inflectional variants will be removed except the best one, we consider the best choice is the bigram which has the smallest number of common affixes that might be existed in different inflectional forms. Moreover, some prefixes of words in bigrams are removed before choosing the best bigram. The best choice for the previous bigrams is [تلوث البيئة].

Morphosyntactic and syntactic variants are more complex than the previous ones, and need advanced linguistic processing to deal with, our proposed system can deal efficiently only with some types of morphosyntactic variants, as well as syntactic variants from modification type and postposition sub-type. Table 6 shows examples of variants' types that our approach deals with.

Table 6.
Examples of variants that our approach deals with

graphical variants	e.g. تيارات حرارية /تيارات حراريه (thermals)
inflectional variants	e.g. سماد نباتي (vegetable mould) أسمدة نباتية (vegetable moulds)
morphosyntactic variants modification/postposition	e.g. مدى الرؤية (visibility) مدى الرؤية الرأسي(vertical visibility)
syntactic variants	e.g. تلوث إشعاعي/ التلوث بالإشعاع (radioactive pollution)

V. Experiments and Results

A. The corpus

The lake of Arabic specialized domain corpora forced the researcher to build new corpora to evaluate their approaches. In fact, using different corpus from different terminology extraction approaches has a negative impact of the ability to compare between them.

To keep our system comparable (as much as possible) with previous work on Arabic MWTs, we used corpus with some similar properties to that which used in [12] and [13].

The corpus belongs to the environment domain and collected from number of websites. The website[1] which has been used in [12] and [13] is part of the corpus. Table 7 shows some information about the corpus.

TABLE 7.
STATISTICAL INFORMATION OF THE CORPUS

number of words (tokens)	522845
number of stemmed words	495618
number of nouns	281531
number of sequences of nouns	62761
number of candidate MWTs	43018

B. Evaluation and Results

Evaluation of ATR approaches is a complex task, basically, there are no specific standards for evaluate and compare different ATR approaches. However, the most of the approaches have used one of two evaluation methods (and sometimes both): reference list and validation [17].

For the evaluation purpose we decided to evaluate our approach using two methods. In the first one, we used the same way used in [12] and [13]. We consider the MWT is correct, if its translation is included in Eurodicautom[2] (terminological database). Unfortunately, Agrovoc[3] (terminological database includes Arabic terms) is not available currently. The second method is the manual validation of the terms. In fact, we found many correct MWTs which are not included in the used terminology database.

The results of our approach are given in Table 8. Obviously, the results show that using C-value method gives better results than using LLR method, while using the combination between the two methods gives us the best results. For the results of LLR method, we can explain the differences between our results and the results obtained in [12] to the difference of the used corpus.

Indeed, it is important to say that we count the terms with basic singular-plural and definitude variation as correct terms, since most of ATR studies allow for these kinds of variations [4]. Figure 6 shows the results of the proposed approach. Figure 7 shows sample of extracted Arabic MWTs ranked by the combination between C-value and LLR methods.

VI. CONCLUSION

In this paper, we have presented a hybrid approach to extract Arabic MWTs. We have concentrated on compound nouns as an important type of MWT, and chose to extract bigram terms, which constitute a high percentage of compound nouns. Extraction of MWT required substantial software development effort. The proposed approach started with the linguistic filter step, this step contains: preprocessing, word's classification, extraction of nouns' sequences, as well as nouns' sequences that connected by prepositions, testing each extracted sequence based on MWTs syntactic patterns, and finally, extraction of bigrams from candidate MWTs.

The next step is the statistical filter. This step includes rank the bigrams based on LLR and C-value methods, and this step follows by dealing with different types of term vari-

TABLE 8.
THE RESULTS OF THE STATISTICAL METHODS

# terms / Method	Top 25				Top 50			
	correct	**not correct**	**with variations**	**precision**	**correct**	**not correct**	**with variations**	**precision**
LLR	23	2	0	92%	43	7	0	86%
C-value	22	3	0	88%	45	5	0	90%
LLR+C-value	23	2	0	92%	47	3	0	94%

# terms / Method	Top 100				Top 150			
	correct	**not correct**	**with variations**	**precision**	**correct**	**not correct**	**with variations**	**precision**
LLR	78	19	3	78%	117	30	3	78%
C-value	86	11	3	86%	128	18	4	85%
LLR+C-value	94	5	1	94%	133	16	1	89%

[1] http://www.greenline.com.kw
[2] http://iate.europa.eu
[3] www.fao.org/agrovoc/

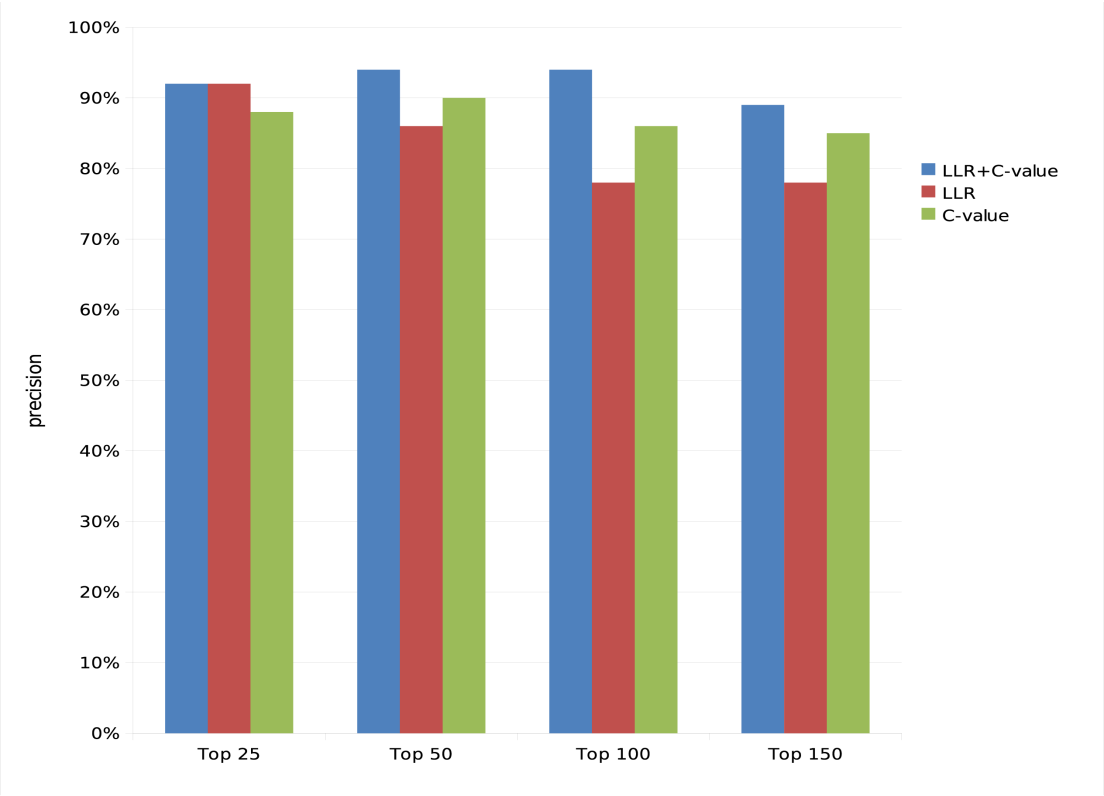

Fig.6: The results of the statistical methods

Search Results

Multi-Word Term	LLR value	C-value	LLR Rank	C-value Rank	Rank
تغير المناخ	[404216.65402181563]	[528.3333333333334]	1	1	1
الأمم المتحدة	[400939.662959421]	[394.7816091954023]	2	3	2
درجة الحرارة	[396669.47281038004]	[352.42857142857144]	4	4	3
أكسيد الكربون	[400922.17808933905]	[342.5625]	3	5	4
مكافحة التصحر	[394124.67433174915]	[148.17391304347825]	6	9	5
الانبعاث الغازي	[395017.99484742293]	[131.0]	5	13	6
الغلاف الجوي	[393694.47276309785]	[148.17391304347825]	9	11	7
الاحتباس الحراري	[392252.77781336545]	[218.9607843137255]	15	8	8
سبيل المثال	[393540.8072412432]	[110.66666666666667]	10	19	9
طبقة الأوزون	[392190.3032954277]	[126.94736842105263]	19	15	10
الكرة الأرضية	[392040.9501259899]	[111.78947368421052]	25	18	11
الشرق الأوسط	[392274.59782239]	[82.17647058823529]	14	29	12
المعادن الثقيلة	[394080.1211716642]	[70.72727272727273]	7	40	13
أشعة الشمس	[392884.538885498]	[60.0]	11	50	14
الاتحاد الأوروبي	[392228.7656031804]	[60.3]	18	46	15
المجلس الوزاري	[391875.752025566]	[69.33333333333333]	36	42	16
المرأة الريفية	[392240.3465985796]	[48.333333333333336]	16	68	17
الوقود الاحفوري	[392159.0324369316]	[40.0]	20	95	18
القطب الجنوبي	[391890.71445463924]	[38.5]	33	100	19
الشعب المرجانية	[392302.3853131083]	[34.25]	13	129	20

Fig.7: Sample of extracted Arabic MWTs ranked by the combination between C-value and LLR methods

ation. The results show that our approach of using a combination between LLR and C-value methods in the ranking process gave better results than using only one of them. In general, we obtained promising results in both coverage and precision of MWT extraction in our experiments based on environment domain corpus.

In the future, we will work to enhance the linguistic filter to be able to extract more complex types of MWTs, use more combinations of statistical methods to rank the candidate MWTs, and extend our method to deal with n-grams MWTs.

REFERENCES

[1] Korkontzelos, I.; Klapaftis, I. P.; and Manandhar, S.: *Reviewing and Evaluating Automatic Term RecognitionTechniques.* In Proceedings of the 6th international Conference on Advances in Natural Language processing, 2008.

[2] Zhang, W.; Yoshida, T.; and Tang, X: *A Study on Multi-word Extraction from Chinese Documents.* In Advanced Web and Network technologies, and Applications: Apweb, 2008.

[3] Koeva, S.: *Multi-word term extraction for Bulgarian.* In Proceedings of the Workshop on Balto-Slavonic Natural Language Processing, 2007.

[4] Kageura,K.;and Umino, B.: *Methods of Automatic Term Recognition A Review,* Termonology 3(2), 259-289.1996.

[5] Church, K. W.; Hanks, P.: *Word association norms, mutual information, and lexicography.* Computational Linguistics 16(1), 22–29, 1990.

[6] Dunning, T.: *Accurate Methods for the Statistics of Surprise and Coincidence.* Computational Linguistics, vol. 19(1), pp. 61-74, 1994.

[7] Patry, A.; Langlais, P.: *Corpus-based Terminology Extraction.* In the 7th International Conference on Terminology and Knowledge Engineering, pp. 313– 321 , 2005.

[8] Frantzi, K. T.; Ananiadou, S.; Mima, H.: *Automatic Recognition of Multi-Word Terms: the C-value/NC-value method.* International Journal on Digital Libraries Vol. 3, No. 2, pp.115–130, 2000.

[9] http://www.un.org/depts/OHRM/sds/lcp/Arabic/

[10] Tadić, M.; Šojat, K.: *Finding multiword term candidates in Croatian.* In the Proceedings of IESL2003 Workshop, pp. 102-107, 2003.

[11] Attia, M. A.: *Handling Arabic Morphological and Syntactic Ambiguity within the LFG Framework with a View to Machine Translation,* doctoral thesis, University of Manchester, Faculty of Humanities, 2008.

[12] Boulaknadel, S.; Daille, B.; and Aboutajdine, D.: *A multi-word term extraction program for Arabic language,* In the 6th international Conference on Language Resources and Evaluation LREC, pp. 1485-1488,2008.

[13] Bounhas, I.; Slimani, Y.: *A hybrid approach for Arabic multi-word term extraction,* NLP-KE 2009. International Conference on Language Processing and Knowledge Engineering, vol., no., pp.1-8, 24-27, 2009.

[14] http://zeus.cs.pacificu.edu/shereen/research.htm#stemming.

[15] Al-Taani, A. T.; Abu-Al-Rub, S.: *A rule-based approach for tagging non-vocalized Arabic words.* The International Arab Journal of Information Technology, Volume 6 (3): 320-328 , 2009.

[16] Thuy Vu; Ai Ti Aw; and Min Zhang: *Term extraction through unithood and termhood unification.* In Proceedings of the 3rd International Joint Conference on Natural Language Processing, 2008.

[17] Pazienza, M. T.; Pennacchiotti, M.; Zanzotto, F. M.: *Terminology extraction: an analysis of linguistic and statistical approaches.* In Knowledge Mining, Springer Verlag, 2005.