MULTI-WORD TERM EXTRACTION BASED ON NEW HYBRID APPROACH FOR ARABIC LANGUAGE

Meryeme Hadni[1], Abdelmonaime Lachkar[2] and Said Alaoui Ouatik[1]

[1]L.I.M, FSDM, USMBA, FEZ, MOROCCO, [2]L.S.I.S, ENSA, USMBA, FEZ, MOROCCO

[1]meryemehadni@gmail.com,s_ouatik@yahoo.com,
[2]abdelmonaime_lachkar@yahoo.fr

ABSTRACT

Arabic Multiword Term are relevant strings of words in text documents. Once they are automatically extracted, they can be used to increase the performance of any text mining applications such as Categorisation, Clustering, Information Retrieval System, Machine Translation, and Summarization, etc. This paper introduces our proposed Multiword term extraction system based on the contextual information. In fact, we propose a new method based a hybrid approach for Arabic Multiword term extraction. Like other method based on hybrid approach, our method is composed by two main steps: the Linguistic approach and the Statistical one. In the first step, the Linguistic approach uses Part Of Speech (POS) Tagger (Taani's Tagger) and the Sequence Identifier as patterns in order to extract the candidate AMTWs. While in the second one which includes our main contribution, the Statistical approach incorporates the contextual information by using a new proposed association measure based on Termhood and Unithood for AMWTs extraction. To evaluate the efficiency of our proposed method for AMWTs extraction, this later has been tested and compared using three different association measures: the proposed one named NTC-Value, NC-Value, and C-Value. The experimental results using Arabic Texts taken from the environment domain, show that our hybrid method outperforms the other ones in term of precision, in addition, it can deal correctly with tri-gram Arabic Multiword terms.

KEYWORDS

Multiword Term extraction,Part Of Speech, Categorisation, Clustering, Information Retrieval, Summarization.

1. INTRODUCTION

Like many other languages such as English, European, Chinese, Hindi Languages, etc, the term in Arabic Language may be a single term composed by one word, or a multiple words named Multiword Term. Note that, a multiword term may carry more meaning than a single-word term and can represent documents more currently. Therefore, once they are automatically extracted, they can be used to increase the performance of any Text mining applications such as Categorisation, Clustering, Information Retrieval System, Machine Translation, and Summarization, etc. Automatic Multiword term (MWT) extraction has gained the interest of

Dhinaharan Nagamalai et al. (Eds) : CSE, DBDM, CCNET, AIFL, SCOM, CICS, CSIP - 2014
pp. 109–120, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4410

many researchers and has applications in many kinds of NLP tasks. The aim of Extraction term is to automatically extract relevant terms from a given corpus.

There are three approaches for MWTs: Linguistic Approach, Statistical Approach and Hybrid Approach. In the Linguistic approach, there exists a variety of previous researches based on morphological, syntactic or semantic information implemented in language-specific rules or programs. These methods are limited by the experience of the specialists who manually select the grammatical patterns. As examples of tools based on this approach we can cite ACABIT [8], Nomino [9] OntoLearn [10] and Lexter[11]. Many researches on MWT focus on methods that are based on Statistical Filters. The methods of T-Score [5], Log-Likelihood Ratio (LLR) [6], FLR [7], Mutual Information (MI) [1] and C-Value [4] are the widely used. Note that, the Mutual Information, Log-Likelihood Ratio and T-Score are proposed to be used in order to measure the Unithood from the strength of inner unity. While the C-Value has been used to measure the Termhood from the strength of marginal variety.

From the above presented approaches, we can conclude that linguistic and statistical approaches present some drawbacks and weakness when they are used alone: On one hand, the statistical approach is unable to deal with low-frequency of MWTs. On the other hand, the linguistic one is language dependent and not flexible enough to cope with complex structures of MWTs.

To avoid the weaknesses of the two filters a commonly recognized solution is to propose a hybrid approach that combines statistical calculus and linguistic Filters [13, 14, and 15]. The T-Score, C-Value and Part-of-speech tags are used as features for compound extraction.

In this paper we present a hybrid Arabic Multi-Word Term extraction method based on two main filters. In the Linguistic Filters we used the method which has been proposed by A. Taani [10], it consists of three levels: the Lexicon Analyzer, the Morphological Analyzer and the Syntactic Analyzer. The Statistical Approach, we adopted to use a new method based on the Unithood and the Termhood measure. The Unithood is to estimate whether a string is a complete lexical unit, and it is measured by the strength of inner unity and marginal variety. The Termhood is to investigate whether the lexical unit is used to refer to a specific concept in a specific domain. we take into account the combination between Termhood and Unithood measures, where we introduce a novel statistical measure, the NTC-Value, that unifies the contextual information and both Termhood and Unithood measure. This measure is applied to another language such as English, French. But not used by Arabic Language.

The remainder of this paper is organized as follows. In the next section, we present the related work. Section 3 describes the proposed method to extract MWTs. In section 4, we present the experimental result. Section 5 concludes this work and presents some perspectives.

2. RELATED WORKS

A lot of works has been done to extract MWT in many languages. These latter have been proposed by using linguistic filter, statistical methods, or both as a hybrid approach. Attia et al. [12] presented a pure linguistic approach for handling Arabic MWTs. It is based on a lexicon of MWTs constructed manually. Then the system tries to identify other variations using a morphological analyzer, a white space normalize and a tokenized. Precise rules allow taking into account morphological features such as gender and definiteness to extract MWTs. The MWTs structures are described as trees that can be parsed to identify the role of each constituent. However some types of MWTs are ignored such as substitution compound nouns. Besides on, the relevance of the extracted candidates is not computed because the lack of statistical measures.

However, the majority of the recently proposed MWT extraction systems have adopted the hybrid approach, because it has given better results than using only linguistic filters or statistical methods [11]. Bouleknadel et al. [13] 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} using available part of speech tagger. In the second step is ranking the extract MWT-like units using association measures, these measures are: Log-Likelihood Ratio, FLR, Mutual Information, 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.

Bounhas et al. [14] have followed a hybrid method to extract compound nouns. In the linguistic side, they combined two types of linguistic approaches discussed above. In the one hand, they detect compound noun boundaries and identify sequences that are like to contain compound nouns. On the other hand, they use syntactic rules to handle MWTs. These rules are based on linguistic information: morphological analyzer and a POS tagger. 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 [13]. Their results were promising especially with bigram MWTS [14]. Recently, another system has been proposed by Khalid El-Khatib et al. [15] based on Linguistic and Statistical Filters to extract Arabic MWTs. (i) The Linguistic Filter, where propose new patterns for syntactic patterns based on definite and indefinite types of nouns. Secondly the extraction of the candidate MWTs takes account the sequence of nouns, as well sequences of nouns that connected by a preposition.(ii) In the statistical filter, the Unithood measure was considered by choosing LLR measure because it gives good results with Arabic MWT extraction [14]. For the Termhood they adopted C-Value measure because it has a wide acceptance as a valuable method to rank candidate MWTs. LLR method can be used efficiently as significance of association measure between the two words in the bigram.

Note that, the most recent work in our knowledge, that has been done by our research team [20], this latter consists to combine the linguistic method that used a part-of-speech (POS) tagger named AMIRA to extract candidate MWTs based on syntactic patterns. It propose a novel statistical measure, the NLC-value, that unifies the contextual information and both Termhood and Unithood measures.

The most proposed previous works present some drawback and weakness that can be summarized as follow: the method proposed in [13], many critics can be addressed to this approach. First, the approach does not include a morphological analysis step. The used POS tagger [16] is unable to separate affixes, conjunctions and some prepositions from nouns and adjectives. The lack of a morphological analysis step obliged the authors to identify in a second step- variant of the already identified MWT. Thus, they identify graphical variants, inflectional variants, morph syntactic and syntactic variants. Second, POS tagging does not allow taking into account many features while defining MWT patterns. For example, we cannot impose constraints about the gender and/or the number of the MWT constituents. Third, this approach does not deal with syntactic ambiguities. In [12], the relevance of the extracted candidates is not computed because the lack of statistical measures. Other work [14] produces results that were promising but only using bi-grams MWTs.

The most hybrid methods presented previously are suitable to use only bi-grams. They have been evaluated the top-ranked does not exceed 100 real terms.

In this investigation, we propose a new method for MWT based ob hybrid approach extraction that can be deal with the previous problems. This proposed method composed of two main stages: the linguistic Filters and the statistical Filter. The linguistic filters operate on the POS–tagged, making use of different kinds of linguistic analyze. The POS–tagged text, obtained with the tagger described in Taani (2009), is searched for on the basis of a set of rules. Specifically, for

each multi–word term to be identified in texts, it passes through three levels of analysis. The lexical analyzer, morphological analyzer and a syntax analyzer. As a statistical filter, we proposed a new method based en C-Value, NC-Value and T-Score. The C-Value method aims at bringing out those terms which tend to occur as nested terms, then; the NC-Value incorporates context information to the C-Value, aiming at improving term extraction in general. The T-Score is used to measure the adhesion between two words in a corpus. The new measure is NTC-Value.

The main novelty of the proposed approach lies in the fact that, differently from previous studies, we incorporate contextual information for each words and frequency analysis on words association. The advantages of this idea, is to reduce the execution time and the size of the vector.

3. PROPOSED APPROACH

In this section we present our proposed multi-word term extraction system based hybrid approach. The system includes two components (fig.1): A Linguistic Filter which uses Part Of Speech (POS) Tagger and Sequence identifier to extract candidate MWTS. The Statistical Filters which unifies the contextual information and both Termhood Estimation and Unithood Estimation

.

3.1. Linguistic Filter

We decide to adopt the approach which has been proposed by Ahmad Taani and al [10], there are two reasons for that. First, this approach is simple and accurate. Therefore, it is able to keep one of the metrics of our syntactic patterns, which is the simplicity. Second, this approach has a morphological analyzer phase. The architecture of adopted approach for words classification contains three main phases. He 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 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.

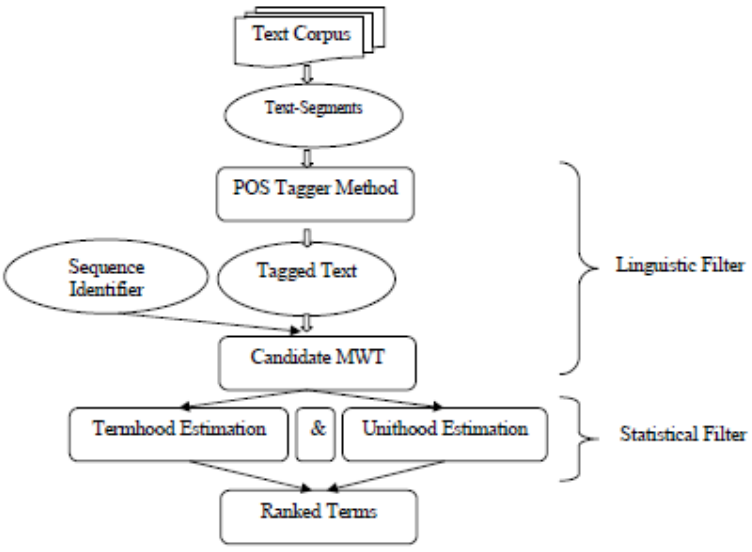

Figure1: Proposed Multiword Term Extraction System

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 second component of linguistic filter is the sequence identifier, using the list of syntactic patterns as follows:

-Noun Prep Noun.
-Noun Noun.

The linguistic filtering performs a morphological analysis and takes into account several types of variations. We followed the typology suggested by [13].

3.1.1. Graphical Variants

By graphical variants, we mean the graphic alternations between the letters ي and ى. Table1 shows some examples of graphic alternations.

Table 1: Graphical variats

Variant	Arabic MWT	Translation
ي/ى	التلوث الكيميائي /التلوث الكيميائى	Chemical pollution

3.1.2. Inflectional Variants

Inflectional variants include the number inflection of nouns, the number and gender inflections of adjectives, and the definite article that is carried out by the prefixed morpheme (Al).Table 2 shows some examples of inflectional variants.

Table 2: Inflectional Variants

Variant	Arabic MWT	Translation
Number	تلوث المحيطات/ تلوث المحيط	Ocean pollution
Definitude	تلوث هوائي /التلوث الهوائي	The Air pollution

3.1.3. Morphosyntactic and Syntactic Variants

Morphosyntactic variants refer to the synonymy relation-ship between two MWTs of different structures. The example below shows synonymic terms of N1 PREP N2 structures (Table3).

Table 3: Morphosyntactic Variants

Variant	Arabic MWT	Translation
N1prepN2	بئر من النفط/بئر نفطي	Oils wells

The syntactic variants modify the internal structure of the base-term, without affecting the grammatical categories of the main item which remain identical. We distinguish modification and coordination variants.

Table 4 shows some examples of syntactic variants.

Table 4: Syntactic Variants

Variant	Arabic MWT	Translation
Insertion	الغلاف الجوي للأرض/ الغلاف للأرض	Atmosphere of Earth
Postposition	الغلاف الجوي المتحرك/ الغلاف الجوي	the atmosphere moving
Expansion	تلوث المحيط و البيئة/تلوث البيئة	pollution of Ocean and environment
Tête	المخاطر و الوقاية من التلوث/المخاطر من التلوث	Risks and prevention of pollution

The next step to extract the candidate MWTs is extraction of sequence of nouns and verb. In this step, we consider each sentence as a separated unit, and using the word's classification approach to extract sequences of nouns. However, this is the last step before using the statistical method to rank the terms.

Specifically, we identify sequences of patterns in order to cover most of the Arabic multi–words structures, using the following pattern: N1N2, N1prepN2. The term candidates are passed to the second step.

3.2. Statistical Filter

In a statistical filter, a term is evaluated using two types of feature: Termhood and Unithood [8]. In C-NC method, the features used to compute the term weight are based on Termhood only. In this paper, we introduce a Unithood feature, T-Score, to the C-NC method.

3.2.1. T-Score

The T-Score is used to measure the adhesion between two words in a corpus. It is defined by the following formula [19]:

$$TS(w_i, w_j) = \frac{P(w_i, w_j) - P(w_i).P(w_j)}{\sqrt{\frac{P(w_i, w_j)}{N}}} \quad (1)$$

Where,

$P(w_i, w_j)$ Is the probability of bi-gram w_i, w_j in the corpus, $P(w)$ is the probability of word w in the corpus, and N is the total number of words in the corpus. The adhesion is a type of Unithood feature since it is used to evaluate the intrinsic strength between two words of a term.

3.2.2. The C-Value/NC-Value Method

The NC-Value measure [4] [6], aims at combining the C-Value score with the context information. A word is considered a context word if it appears with the extracted candidate terms. The first part, C-value enhances the common statistical measure of frequency of occurrence for term extraction, making it sensitive to a particular type of multi-word terms, the nested terms. The second part, NC-value, gives: 1) a method for the extraction of term context words (words that tend to appear with terms), 2) the incorporation of information from term context words to the extraction of terms.

- ✓ *C-Value*

The C-Value calculates the frequency of a term and its sub-terms. If a candidate term is found as nested, the C-Value is calculated from the total frequency of the term itself, its length and its frequency as a nested term; while, if it is not found as nested, the C-Value, is calculated from its length and its total frequency.

$$\text{CValue}(a) = \begin{cases} \log_2|a|.f(a) & \text{if a is not nested} \\ \log_2|a|.\left(f(a) - \frac{1}{\text{P}(\text{T}_a)}\sum_{b\in T_a} f(b)\right) & \text{otherwise} \end{cases} \quad (2)$$

Where, f (a) is the frequency of term a with |a| words, T_a is the set of extracted candidate terms that contain a and P(T_a)is the total number of longer candidate terms that contain a . The formula $\frac{1}{P(T_a)}\sum_{b\in T_a} f(b)$ will have value 0 when T_a is empty.

- ✓ *NC Value*

The NC-Value measure [6] aims at combining the C-Value score with the context information. A word is considered a context word if it appears with the extracted candidate terms. The algorithm extracts the context words of the top list of candidates (context list), and then calculates the N-Value on the entire list of candidate terms. The higher the number of candidate terms with which a word appears, the higher the likelihood that the word is a context word and that it will occur with other candidates. If a context word does not appear in the extracted context list, its weight for such term is zero. Formally, given w as a context word, its weight will be:

$$weight(b) = \frac{\text{t(b)}}{\text{n}} \quad (3)$$

Where t (b) is the number of candidate terms b appears with, and n is the total number of considered candidate terms; hence, the N-Value of the term t will be

$$\text{NValue} = \sum_{b\in C_a} f_a(b) * \text{weight(b)} \quad (4)$$

where $f_a(b)$ is the frequency of b as context word of a, and C_a is the set of distinct context words of the term t. Finally, the general score, NC-Value, will be:

$$\text{NCValue}(a) = 0.8.\text{CValue}(a) + 0.2.\text{NValue}(a) \quad (5)$$

From the above formula, we find that NC-Value is mainly weighted by C-Value .It treats the term candidate as a linguistic unit and evaluates its weight based on characteristics of the Termhood,

i.e. frequency and context word of the term candidate. The performance can be improved if feature measuring the adhesion of words within the term is incorporated.

3.2.3. The NTC-Value

Theoretically, the C/NC method can be improved by adding Unithood feature to the term weighting formula. Based on the comparison of [18], we explore T-Score, a competitive metric to evaluate the association between two words, as a Unithood feature.

Our idea here is to combine the frequency with T-Score, a Unithood feature. Taking the example in Table 5, the candidates have similar rank in the output using C/NC Termhood approach.

Table 5. Example of context MWT

MWT	Translation
وزارة التعليم العالي	Ministry of Higher Education
التعليم العالي بالمغرب	Higher Education in Morocco
سلامة التعليم العالي	the Safety of Higher Education
التعليم العالي الجامعي	the Higher Education University

Example for Environmental domain:

Table 6. Example of context MWT

MWT	Translation
ملوثات الغلاف الجوي	Atmospheric pollutants
الغلاف الجوي للأرض	Earth's atmosphere
الغلاف الجوي و الطقس	the atmosphere and the weather
توازن الغلاف الجوي	The balance of the atmosphere
غازات الغلاف الجوي	Atmospheric gases

To give better ranking and differentiation, we introduce T-Score to measure the adhesion between the words within the term. We use the minimum T-Score of all bi-grams in term a, minTS(a), as a weighted parameter for the term besides the term frequency.

For a term $a = w_1 . w_2 \ldots v$, the minTS(a) is defined as :

$$minTS(a) = \min\{TS(w_i, w_{i+1})\}, i = 1 \ldots (n -$$

Table 7. Term with Minimum T-Score value

MWT	Translation	minTS(MWT)
وزارة التعليم العالي	Ministry of Higher Education	3.53
التعليم العالي بالمغرب	Higher Education in Morocco	2.64
سلامة التعليم العالي	the Safety of Higher Education	9.78
التعليم العالي الجامعي	the Higher Education University	1.73

Table8. Term with Minimum T-Score value

MWT	Translation	minTS (MWT)
ملوثات الغلاف الجوي	Atmospheric pollutants	9.65
الغلاف الجوي للأرض	Earth's Atmosphere	3.74
الغلاف الجوي و الطقس	the atmosphere and the weather	6.28
توازن الغلاف الجوي	The balance of the atmosphere	1.72
غازات الغلاف الجوي	Atmospheric gases	3.54

Table 7 and Table 8 shows the minTS(MWT) of the different terms in table 5 and Table 6 respectively. Since $minTS(a)$ can have a negative value, we only considered those terms with $minTS(a) > 0$ and combined it with the term frequency. We redefine C-Value to TC-Value by replacing f (a) using F(a), as follows:

$$F(a) = \begin{cases} f(a) & if\ minTS(a) \leq 0 \\ f(a) * \ln(2 + minTS(a)) & if\ minTS(a) > 0 \end{cases} \quad (6)$$

$$TCValue(a) = \log_2|a|.\left(F(a) - \frac{1}{P(T_a)}\sum_{b \in T_a} F(b)\right) \quad (7)$$

The final weight, defined as NTC-Value, is computed using the same parameter as NC-Value.

$$NTCValue(a) = 0.8.TCValue(a) + 0.2.NValue(a) \quad (8)$$

4. EXPERIMENT AND RESULT

4.1. The Corpus Collection

The lake of Arabic specialized domain corpora forced the research to build new corpora to evaluate their approaches. The texts are taken from the environment domain and are extracted from the web site "Al-Khat Alakhdar"[1]. The corpus contains 1.013 documents and 470.175 words.

4.2. Evaluation

Evaluation of MWT approaches is a complex task, there are no specific standards for evaluate and compare different MWT approaches. However, the most of the approaches have used one of two evaluation steps: reference list and validation. In the first step, we attest that a term is relevant to the environment domain if it has already been listed in existing terminology database AGROVOC[2]. The second methods, if the term not exists in AGROVOC we search his translation in database IATE[3] (InterActive Terminology for Europe).

Table 9 shows the comparison result of the origin C-value, NC-value and NTC-value on the ranking for the MWT candidates. We evaluate the performance based on the k best candidates from 100-500 at intervals of 100.

[1] http://www.greenline.com.kw
[2] www.fao.org/agrovoc/
[3] http://iate.europa.eu/iatediff/SearchByQueryLoad.do?method=load

We attested that a term is relevant if it has been listed in existing database AGROVOC and IATE.

Table 9. Precision for C-Value, NC-Value, and NTC-Value

Top terms	*C-Value*	*NC-Value*	*NTC-Value*
100	66,0%	74,0%	86,0%
200	63,0%	69,0%	75,0%
500	58,0%	66,0%	71,0%

Furthermore, the combination of the context information and the C-Value improves the performance of the process of MWT extraction because the NC-Value outperforms the C-Value for each considered MWT list. The Unithood feature NTC-Value outperforms the C-Value/NC-Value as expected from previous studies. Figure 2 illustrates the precision obtained for the C-Value/NC-Value and the NTC-Value.

Figure 2 expresses the same information as table 9, as a graph. In the horizontal axis, the number of candidate term for the three methods are shown, while in the vertical axis, the precision for number of these intervals is provided.

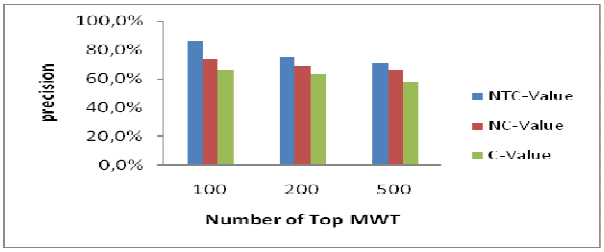

Figure 2. Precision Obtained for NC-value, C-value and NTC-value

The integration of contextual information and the T-score Unithood measure to the C-Value improves the performance of MWT acquisition, since the NTC-Value has better precision than the C-Value\NC-Value, as illustrated in Figure 2.

In figure3, we present some obtained results for Arabic MWTs extraction using the three methods: C-Value, NC-Value and NTC-Value

Figure 3. Sample of extracted MWT using: C-Value, NC-Value and NTC-Value

Arabic MWT	Translation	NTC-Value	NC-Value	C-Value
الخط الأخضر	Green Line	99.2	99.2	89.26567206138039
الطاقة الشمسية	Solar Energy	99.2	89.44804089867734	38.21052858969426
تلوث المنتجات	Contamination of products	99.2	50.88460312771606	3.806022677747762
المناطق المناخية	Climatic zones	99.2	76.76391432416152	76.17764754947734
مستوى تذبذبات	Level fluctuations	99.2	76.59618564676952	22.97811394253883
المسطحات الطينية	Mudflats	99.2	76.5961856467652	22.97811394253883
التوريد والتوزيع	Supply and distribution	99.2	50.91157110023308	3.1699250014423126
فصيلة الفيلة	Platoon elephants	99.2	50.91157110023308	3.806022677747762
الطاقة المتجددة	Renewable energy	99.2	50.91157110023308	23.071735871528716
الأمم المتحدة	United Nations	99.2	23.63345998892698	11.200000000000001
البترول الوطنية	National Petroleum	99.2	76.30461513232898	89.60000000000001
بلدية الكويت	Kuwait Municipality	99.2	76.17764754947734	5.065613907861519

محتوى المعادن	Metal content	96.80000000000001	4.754887502163469	2.53727333451947
المتجددة بالصين	China's renewable	96.80000000000001	80	38.06022677747762
التمثيل الضوني	Photosynthesis	89.60000000000001	50.85613907861519	89.44804089867734
الانبعاثات الغازية	Emissions	89.26567206138039	38.21385730048871	76.59618564676952
النظام البيني	Ecosystem	89.26567206138039	38.21385730048871	4.800000000000001
الغازات السامة	Toxic gases	89.26567206138039	95.09775004326938	5.188460312771606
مخاطر التلوث	The risk of contamination	89.26567206138039	50.85613907861519	23.200000000000003
ملوثات الهواء	Air pollutants	89.26567206138039	38.222773488520145	89.60000000000001

are performed for bi-grams and tri-grams on an Arabic Texts taken from the environment domain. In conclusion, the efficiency of our proposed method for AMWTs extraction has been tested and compared using three different association measures: the proposed one named NTC-Value, NC-Value, and C-Value. The experimental results show that our hybrid method outperforms the other ones in term of precision; in addition, it can deal correctly with tri-grams Arabic Multiword terms.

In the future work we are considering to integrate evaluation by an expert, because there's words that not exist in AGROVOC or in IATE and there are correct. For example; "فصيلة الفيلة" "A platoon elephants" and "توريد والتوزيع" "Supply and distribution". Then study the impact of POS tagging on AMWTs extraction.

REFERENCES

[1] B. Daille, (1994) " Approche mixte pour l'extraction de terminologie : statistique lexicale et filtres linguistiques ", doctoral thesis, University of Paris 7.

[2] K. Church & W. Gale & P. Hanks,&D. Hindle(1991), "Using statistics in lexical analysis," in Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon. U. Zernik, pp. 115–164.

[3] Hiroshi Nakagawa,& Tatsunori Mori.(2002)." A Simple but Powerful Automatic Term Extraction Method". 2nd International Workshop on Computational Terminology,ACL.

[4] Katerine& T. Frantzi,& Sophia Ananiadou, & Junichi Tsujii,(1998)." The C-Value/NC-Value Method of Automatic Recognition for Multi-word terms". Journal on Research and Advanced Technology for Digital Libraries.

[5] Hideki Mima& Sophia Ananiadou(2001)." An Application and Evaluation of the C/NC-Value Approach for the Automatic Term Recognition of Multi-Word Units in Japanese". International Journal on Terminology.

[6] Spela Vintar.(2004)." Comparative Evaluation of C-value in the Treatment of Nested Terms", Memura 2004 –Methodologies and Evaluation of Multiword Units in Real-World Applications. Proceedings of the International Conference on Language Resources and Evaluation 2004, pp. 54-57.

[7] E. Milios& Y. Zhang& B. He,&L. Dong. (2003). "Automatic Term Extraction and Document Similarity in Special Text Corpora". Proceedings of the 6th Conference of the Pacific Association for Computational Linguistics (PACLing'03), Halifax, Nova Scotia, Canada, pp. 275-284.

[8] Kyo Kageura (1996)." Methods of Automatic Term Recognition - A Review". Terminology, 3(2): 259 – 289.

[10] A.T Al-Taani&&S. Abu-Al-Rub(2009)," A rule-based approach for tagging non-vocalized Arabic words". The International Arab Journal of Information Technology, Volume6 (3): 320-328,

[11] M. Tadi&, K. Sojat(2003)," Finding multiword term candidate in Croatian". In the Proceeding of IESL2003 Worksop, pp. 102-107.

[12] Attia& M.A(2008),"Handling Arabic Morphological and Syntactic Ambiguity within the LFG Framework with a view to Machine Translation", doctoral thesis, University of Manchester, Faculty of Humanities.

[13] S. Bouleknadel& B.Daille & D. Aboutajdine(2008),"A multi-word term extraction program for Arabic language", In the 6th international Conference on language resources and evaluation LREC, pp. 1485-1488.

[14] I. Bounhas& Y. Slimani,(2009),” 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.

[15] K. El Khatib& A. Badarenh. (2010).” Automatic Extraction of Arabic Multi-word Term”. Proceedings of the International Multiconference on Computer Science and Information Technology, pp.411-418.

[16] M. Diab& K. Hacioglu & D. Jurafsky,(200’), "Automatic Tagging of Arabic Text: From raw text to Base Phrase Chunks", in the 5th Meeting of the North American Chapter of the Association for Computational Linguistics/Human Language Technologies Conference (HLT-NAACL04), Boston, Massachusetts, May 2-7.

[17] C. Manning & H. Schuetze. (1999).” Foundations of StatisticalNatural Language Processing”. MIT Press Cambridge, Massachusetts.

[18] Evert& S. & B. Krenn. (2001). “Methods for Qualitative Evaluation of Lexical Association Measures”. Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics, pages 369 – 381.

[19] Vu Thy& Ai Ti Aw& Min Zhang(2008),” Term extraction throught unithood and termhood unification”. In proceeding of the 3rd International Joint Conference on Natural Language Processing.

[20] A.EL Mehdaoui&S. EL Alaoui Ouatik & E. Gaussier,(2013),” A Study of Association Measures and their Combination for Arabic MWT Extraction”, published in "Terminology and Artificial Intelligence, Paris : France .