A NOVEL METHOD FOR ARABIC MULTI-WORD TERM EXTRACTION

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

[1]LIM, USMBA, Fez, Morocco
[2]LSIS, ENSA, Fez, Morocco

ABSTRACT

Arabic Multiword Terms (AMWTs) are relevant strings of words in text documents. Once they are automatically extracted, they can be used to increase the performance of any Arabic Text Mining applications such as Categorization, Clustering, Information Retrieval System, Machine Translation, and Summarization, etc. Mainly the proposed methods for AMWTs extraction can be categorized in three approaches: Linguistic-based, Statistic-based, and hybrid-based approach. These methods present some drawbacks that limit their use. In fact they can only deal with bi-grams terms and their yield not good accuracies. In this paper, to overcome these drawbacks, we propose a new and efficient method for AMWTs Extraction based on a hybrid approach. This latter is composed by two main filtering steps: the Linguistic filter and the Statistical one. The Linguistic Filter uses our proposed Part Of Speech (POS) Tagger and the Sequence identifier as patterns in order to extract candidate AMWTs. While the Statistical filter incorporate the contextual information, and a new proposed association measure based on Termhood and Unithood Estimation named NTC-Value.

To evaluate and illustrate the efficiency of our proposed method for AMWTs extraction, a comparative study has been conducted based on Kalimat Corpus and using nine experiment schemes: In the linguistic filter, we used three POS Taggers such as Taani's method based Rule-approach, HMM method based Statistical-approach, and our recently proposed Tagger based Hybrid –approach. While in the Statistical filter, we used three statistical measures such as C-Value, NC-Value, and our proposed NTC-Value. The obtained results demonstrate the efficiency of our proposed method for AMWTs extraction: it outperforms the other ones and can deal correctly with the tri-grams terms.

KEYWORDS

Multiword Terms extraction, contextual information, Part Of Speech, Termhood Estimation , Unithood Estimation

1. INTRODUCTION

Automatic Multiword term (MWT) extraction has gained the interest of many researchers and has applications in many kinds of NLP tasks, such as Information Retrieval, Information Extraction, Text Categorization and Automatic Domain Ontology Construction in the last few years. The aim of Extraction term is to automatically extract relevant terms from a given corpus.

There is a variety of previous researches that focus on the Linguistic Filters, such as 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

DOI : 10.5121/ijdms.2014.6304

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 (MI3) [1] and C-Value [4]. Are widely used Mutual Information, log-likelihood ratio and T-score measures the Unithood from the strength of inner unity, and C-Value is used to get more accurate terms, especially those nested terms.

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 approaches a commonly recognized solution is to propose a hybrid approach that combines statistical 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 a comparison with three taggers: Taani's Rule-Based method [10], HMM method [20] and Hybrid method[21]. The Statistical Filters, we adopted to uses 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 rest of the paper is structured as follows. Some of the related work is described in section2. Section 3 presents a method of POS Tagger .In section4 we describes our proposed approach to extract MWTs. Section5 shows the experiments and the results of applying the extraction approach. The last section contains the conclusion and the future work.

2. RELATED WORK

A lot of work has been done to extract MWT in many languages. These works has been proposed by using linguistic filter, statistical methods, or both as a hybrid approach. However, the majority of the last MWT extraction systems have adopted the hybrid approach, because it has given better results than using only linguistic filters or statistical methods [11]. Some recent works which dealt with this problem use either pure linguistic or hybrid approaches. For example, Attia [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.

Bouleknadel and 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 and al. [14] have followed a hybrid method to extract multiword terminology from Arabic corpora. 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] to extract multi-word terms form Arabic corpus. They concentrated on compound nouns as in important type of MWT and select bi-gram term. The approach relies on two filters. (i) 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 [22], 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 on hybrid approach extraction that can be deal with the previous problems. This proposed method composed of two main stages: the Linguistic Filter and the Statistical Filter. The Linguistic Filter operates on the POS–tagged, making use a comparison with different method of Tagger such as Taani's Rule-Based method [10], HMM Method [20] and Hybrid Method [21].As a statistical filter; we proposed a new method based en C-Value, NC-Value and T-Score, to derive a new measure, NTC-Value.

3. POS TAGGER

Part-Of-Speech (POS) tagging is known as a necessary work in many areas Natural Language Processing (NLP) systems like information extraction, parsing of text and semantic processing. The POS tagging is known as assigning grammatical tags to words and symbols making a text which include a large amount of lexical information and captures the relationship between these words and their adjacent related words in a sentence, or paragraph [1][2][3].

The most methods of POS tagging can be classified in three categories: Rule-Based approach (e.g. Taani's method), Statistical method (e.g. HMM method), and Hybrid method. A brief description of each method is presented subsequently.

3.1. Taani's method: Rule-Based approach

The Taani's Rule-Based tagging method [10] allows labelling the words in a non-vocalized Arabic text to their tags. It is constituted of three main phases: the lexicon analyzer, the morphological analyzer, and the syntax analyzer.

Lexicon Analyzer: 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. Else, it passes to the next step.

Morphological Analyzer: Each word which has not been tagged in the previous phase will immigrate to this phase. A set of the affixes of each word are extracted. After that, these affixes and the relations between them are used in a set of rules to tag the word into its class.

Syntax Analyzer: This phase can help in tagging the words which the previous two phases failed to tag. It consists of two rules: sentence context and reverse parsing. The sentence context rule is based on the relation between the untagged words and their adjacent. The reverse parsing rule is based on Arabic context-free grammar. The authors propose a set of rules which are used frequently in Arabic language.

3.2. HMM Method: Statistical -Based approach

This section covers the use of a Hidden Markov Model (HMM) to do part-of-speech tagging can be seen as a special case of Bayesian inference [20]. It can be formalized as follows: for a given sequence of words, what is the best sequence of tags which corresponds to this sequence of words? If we represent an entered text (sequence of morphological units in our case) by $W = (w_i)_{1 \le i \le n}$ and a sequence of tags from the lexicon by $T = (t_i)_{1 \le i \le n}$, we have to compute:

$$\max_{T}[P(T|W)] \quad (1)$$

By using the Bayesian rule and then eliminating the constant part, $P(T)$ the equation can be transformed to this new one:

$$\max_{T}[P(T|W) * P(T)] \qquad (2)$$

Where $P(T)$ represents the probability of the tag sequence (tag transition probabilities), and can be computed using an N-gram model, as follows:

$$P(T = t_1 t_2 \dots t_n) = \prod_{i=1}^{n} P(t_i | t_{i-n} \dots t_{i-2} t_{i-1}) \qquad (3)$$

A tagged training corpus is used to compute$P(t_i | t_{i-n} \dots t_{i-2} t_{i-1})$, by calculating frequencies of N-gram as follows:

$$P(t_i | t_{i-n} \dots t_{i-2} t_{i-1}) = (f(t_{i-n} \dots t_{i-2} t_{i-1}) | f(t_{i-n} \dots t_{i-2} t_{i-1})) \quad (4)$$

However, it can happen that some trigrams (or bigrams) will never appear in the training set; so, to avoid assigning null probabilities to unseen trigrams (bigrams), we used a deleted interpolation developed by [20]:

$$\lambda_1 P(t_i | t_{i-n} \dots t_{i-n-1}) + \dots + \lambda_{n-2} P(t_i | t_{i-2} t_{i-1}) + \lambda_{n-1} P(t_i | t_{i-1}) + \lambda_n P(t_i) \quad (5)$$

Where $\lambda_1 + \lambda_2 + \dots + \lambda_n = 1$

Then, for calculating the likelihood of the word sequence given tag $P(T|W)$, the probability of a word appearing is generally supposed to be dependant only on its own part-of-speech tag. So, it can be written as follows:

$$P(W|T) = \prod_{i=1}^{n} P(w_i | t_i) \qquad (6)$$

In addition, a tagged training set has to be used for computing these probabilities, as follows:

$$P(w_i | t_i) = \frac{f(w_i, t_i)}{f(t_i)} \qquad (7)$$

Where $f(w_i, t_i)$and $f(t_i)$ represent respectively how many times $w_i$is tagged as t_i and the frequency of the tag $t_i$itself.

Tag sequence probabilities and word likelihoods represent the HMM model parameters: transition probabilities and emission (observation) probabilities. Once these parameters are set, the HMM model can be used to find the best sequence of given a sequence of input words. The Viterbi algorithm can be used to perform this task.

3.3. Our proposed Tagger based-Hybrid approach

Our proposed Tagger [21] solves the problem of misclassified and unanalyzed words generate by rule-based method [10] using the statistical method that is the Hidden Markov Model [20]. Figure 1 show the hybrid method for POS Tagging.

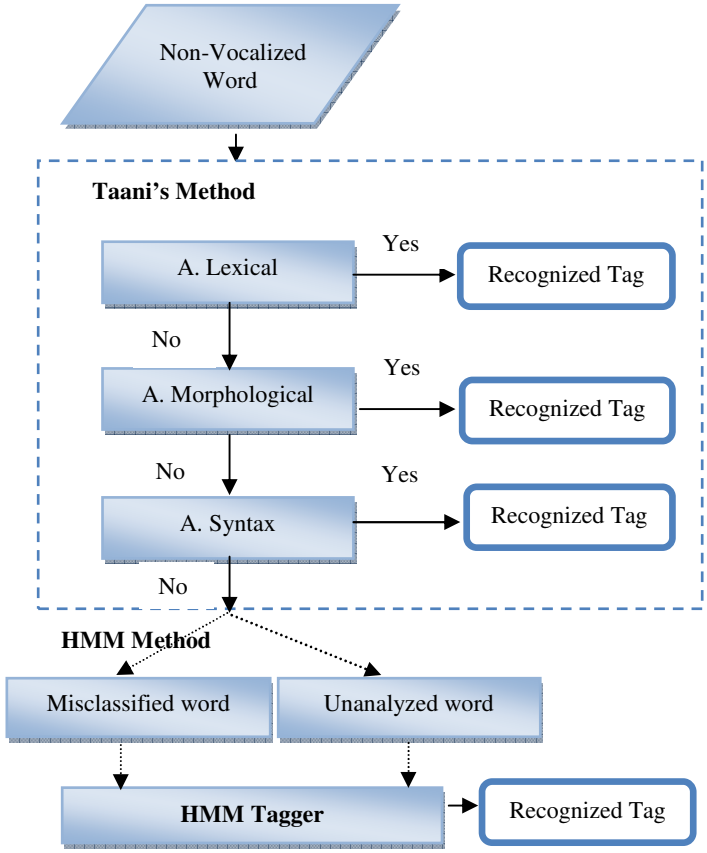

Figure1. Flowchart of the Hybrid Arabic POS Tagger

4. PROPOSED METHOD FOR ARABIC MULTIWORD TERMS EXTRACTION

In this section we present our proposed multi-word term extraction system based hybrid approach. The approach requires the following steps (Figure2): A Linguistic Filter which uses Part Of Speech (POS) Tagger mention in the previous section and the Sequence identified tokenizes tagged files of the corpus and uses syntactic patterns in order to identify candidate terms that fit the rules of the grammar, as follows: *Noun Prep Noun., Noun Noun* . The Statistical Filters which unifies the contextual information and both Termhood Estimation and Unithood Estimation.

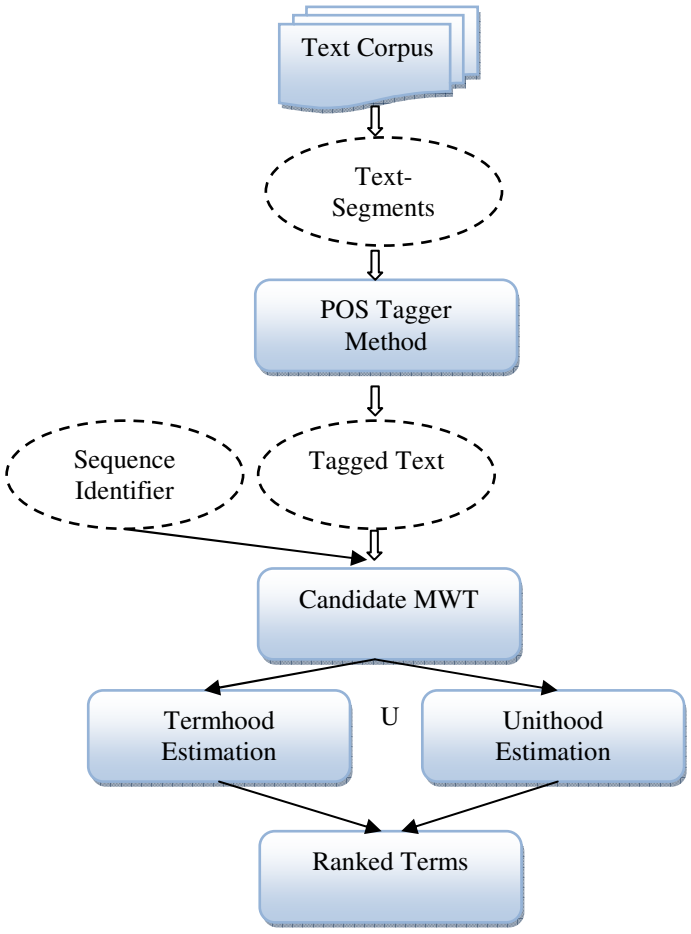

Figure2: Proposed Multiword Term Extraction System

4.1. The Linguistic Filter

The Linguistic Filtering performs a morphological analysis and takes into account several types of variations: graphical, inflectional, morph syntactic and syntactic variants.

4.1.1. Graphical variants

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

Table1. Graphical variants

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

4.1.2. 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.

Table2. Inflectional variants

Variant	*Arabic MWT*	*Translation*
Number	تلوث المحيطات/ تلوث المحيط	Ocean pollution
Definitude	انشطة ترفيهية / الانشطة الترفيهية	Entertainment activities

4.1.3. 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).

Table3. 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.

Table4. syntactic variants

Variant	*Arabic MWT*	*Translation*
Insertion	لجنة الشؤون المالية / لجنة المالية	Finance Committee
Postposition	الامين العام للجبهة / الامين العام	Secretary General
Expansion	اسلحة الطب والدمار / اسلحة الدمار	Weapons of mass
Tête	المخاطر و الوقاية من التلوث/المخاطر من	Risks and prevention of pollution

4.2. Statistical Filter

In the next step, we apply a number of statistical measures to rank the list of candidate MWTs extracted by the linguistic filter. In C-Value/NC-Value method, the features used to compute the term weight are based on Termhood only. In the rest, we introduce a Unithood feature, T-Score, to the C-NC method.

4.2.1. 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.

4.2.2. The C-Value/NC-Value measures

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.

- ✓ *The C-Value measure*

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.

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

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 $\frac{1}{P(T_a)}\sum_{b\epsilon T_a} f(b)$ will have value 0 when T_a is empty.

- ✓ *The NC-Value measure*

The NC-Value measure (Frantzi et al., 1999) 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:

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

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:

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

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.

4.2.3. The NTC-Value

Theoretically, the C-Value/NC-Value 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.

Table5. 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

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 $a = w_1.w_2 \ldots w_n$, the minTS(a) is defined as : $minTS(a) = \min\{TS(w_i, w_{i+1})\}, i = 1 \ldots (n-1)$

Table6. 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

Table 6 shows the minTS(MWT) of the different terms in table 5. 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:

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

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

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

$$\mathrm{NTCValue(a)} = 0.8.\,\mathrm{TCValue(a)} + 0.2.\,\mathrm{NValue(a)} \quad (8)$$

5. EXPERIMENT & RESULT

5.1. The Corpus Collection

The corpus built contains 20.291 documents, the texts are taken from the Kalimat Corpus[1].It cover various topic such as culture, economic, international, local, religion, sports, and international.

In this section we assess the results.

5.2. 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 if it has already been listed in existing terminology database AWN[2]. The second method, if the term not exists in AWN we search the translation in included 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, and with different method of POS Tagger. We evaluate the performance based on the k best candidates from 100-500.

[1] http://bit.ly/16jO3Ks
[2] http://sourceforge.net/projects/awnbrowser/
[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 AWN and IATE.

$$\text{precision} = \frac{\text{attested multiword term}}{\text{all extracted sequence}}$$

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.

The hybrid method of POS Tagging improves the result of multiword term extraction relative to statistical and rule-based methods, combined to NTC-Value.

Table7. Precision: NC-Value, C-Value and NTC-Value for different method of POS Tagging

Top terms	*Taani's Method*			*HMM Method*			*Hybrid Method*		
	C-Value	*NC-Value*	*NTC-Value*	*C-Value*	*NC-Value*	*NTC-Value*	*C-Value*	*NC-Value*	*NTC-Value*
100	65,40%	79,00%	89,00%	90,10%	91,00%	92,40%	91,60%	92,20%	94,40%
200	63,00%	64,00%	73,40%	87,00%	87,40%	89,50%	88,20%	90,10%	91,10%
500	57,50%	65,60%	68,20%	85,20%	84,60%	85,60%	86,00%	88,20%	89,00%

Figure 3, 4, 5 expresses the information as table 7, 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.

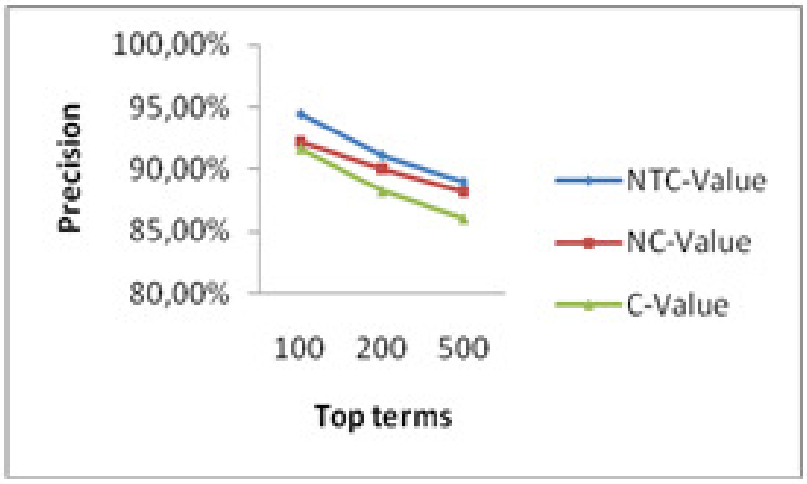

Figure3. Precision Obtained for NC-value, C-value and NTC-value for Hybrid POS tagging

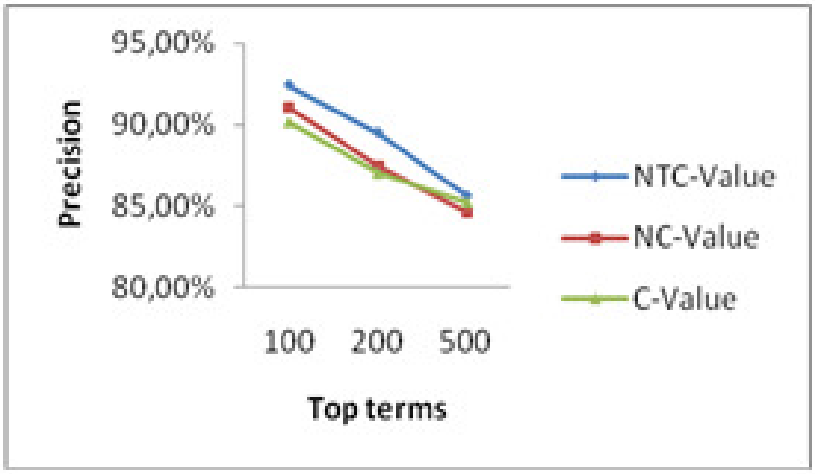

Figure4. Precision Obtained for NC-value, C-value and NTC-value for HMM Method

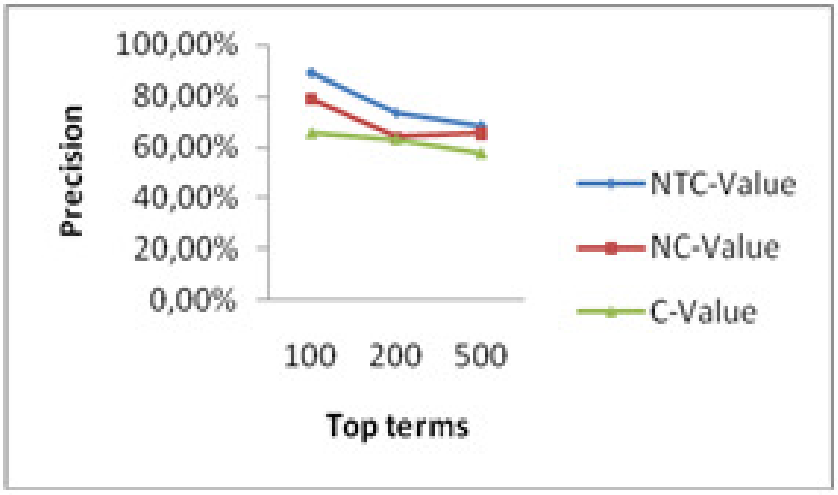

Figure5. Precision Obtained for NC-value, C-value and NTC-value for Taani's Method

The integration of contextual information and the T-score Unithood measure to the C-Value improves the performance of MWT acquisition with the combination to Hybrid method for POS Tagging, since the NTC-Value has better precision than the C-Value\NC-Value, as illustrated in Figure3, 4, 5.

6. CONCLUSION AND FUTURE WORKS

In this paper we presented our proposed Multiword term extraction system based on the contextual information. Our hybrid method is composed by two main steps: the Linguistic approach and the Statistical one. In the first step, we apply the Linguistic approach to extract the candidate MWTs based on Part Of Speech (POS) Tagger using a comparison with three methods such as Taani's Rule-based method, HMM method and Hybrid method, and syntactic pattern using a Sequence Identifier. In the second approach 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.

Experiments are performed for bi-grams and tri-grams on Arabic Texts taken from the Kalimat corpus. In conclusion, the efficiency of our proposed method for AMWTs extraction has been tested and compared three method of POS Tagging and 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 are words that not exist in AWN or in IATE and there are correct.

REFERENCES

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

[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". 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. Tadié & 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"e, 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 (2004), "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 2004.

[17] C. Manning & H. Schuetze(1999). “Foundations of Statistical Natural 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] M. Albared & O.Nazlia(2010),” Automatic Part of Speech Tagging for Arabic: An Experiment Using Bigram Hidden Markov Model “,Springer-Verlag Berlin Heidelberg, LNAI 6401, pp. 361–370.

[21] M.hadni, S.ouatik El Alaoui & A.Lachkar (2013),”Hybrid Part-Of-Speech Tagger for Non-Vocalized Arabic Text”, International Journal on Natural Language and Computing,Vol.2, No.6.

[22] A. El Mahdaouy S. El Alaoui Ouatik & E. Gaussier (2013)“A Study of Association Measures and their Combination for Arabic MWT Extraction”, In Proceedings 10th International Conference on Terminology and Artificial Intelligence, pp. 45-52.