A Hybrid Multi-Word Terms Extraction System Applied to Topic Detection

[1]Rim Koulali, Abdelouafi Meziane[2]
[1]LaRI Laboratory, Science college, Mohammed I University, Oujda
[2]LaRI Laboratory, Science college, Mohammed I University, Oujda

ABSTRACT

Mutli-word Terms extraction plays an important role in many Natural Language Processing (NLP) tasks. Despite their major importance, few works were dedicated to Arabic multi-word terms extraction. This paper proposes an automatic Arabic multi-word terms (MWTs) extraction system based on two major filtering steps: linguistics filter using a part-of-speech tagger along with morphological patterns and statistical filter based on probabilistic methods, namely: Log-Likelihood Ratio (LLR) and C-value. We evaluate the performances of the realized systems on Wattan; an Arabic oriented topic newspaper corpus. Our system manages to achieve 90.23% in term of multi-word extraction precision. We also study the use of MWTs as features in Arabic Topic Detection. The conducted experiments show good results.

Indexing terms/Keywords

Multi-word Terms Extraction;Topic Detection; C-value; LLR.

Academic Discipline And Sub-Disciplines

Computer Sciences and Engineering.

SUBJECT CLASSIFICATION

Natural language processing; information retrieval.

TYPE (METHOD/APPROACH)

Experimental.

Council for Innovative Research
Peer Review Research Publishing System
Journal: INTERNATION JOURNAL OF COMPUTERS AND TECHNOLOGY
Vol. 13, No. 10
editorijctonline@gmail.com
www.ijctonline.com, www.cirworld.com

INTRODUCTION

The increasing availability of Arabic electronic documents has led to extensive research efforts covering the Arabic Natural Language Processing (ANLP) various fields, taking in consideration, particularities and complex morphological composition of the Arabic language. Controversially, few researches have been undertaken in the field of multi-word terms extraction for Arabic documents.

Although multi-word term has no uniform definition, it can be understood as a sequence of two or more consecutive individual noun words, forming a semantic unit [1]. In fact, the exact meaning of the words composing the MWT cannot be derived separately from the other MWT parts.

MWTs Extraction is an important task of automatic terms recognition and is employed in numerous NLP fields such as: text mining [2], syntactic parsing [3], [4], machine translation [5] and text classification [6]. The MWTs extraction task covers detection and extraction of a consecutive set of semantically related words. The technics used in MWTs extraction can be classified into four categories:

- Statistical approaches based on frequency, probability and co-occurrence measures [7].
- Symbolic approaches using parsers, morphological analysis, MWTs boundaries detection and patterns [8].
- Hybrid approaches combining statistical and morphological methods [9][10].
- Word alignment approaches [11].

The hybrid approaches are wildly used since they combine the benefits of statistical and symbolic methods.

Our work is part of the semantic processing of unvowelized Arabic documents and aims to develop a multi-word terms extraction prototype for Arabic texts based on the hybrid approach using lexical patterns and statistical measures: C-value and LLR. We experimentally investigate the usage of MWTs as features in topic detection.

This paper is organized as follows: In section II we present related work. In section III; we describe the developed MWTs extraction system. Section IV details the conducted experiments and obtained results. Finally we conclude the paper in sectionV and announce future work.

Related Work

Although, MWTs extraction systems and prototypes have been developed for various languages such as: English, French, Chinese, Turkish, Dutch, Urdu…. Only, few researches have been dedicated to MWTs extraction for Arabic language.

The authors of [12] explore three approaches: the first one based on crossing correspondence asymmetries between Arabic Wikipedia titles and titles in 21 different languages, the second approach uses translated English MWTs to Arabic language and proceeds to validation. The last one benefits from large corpora and lexical association measures. These approaches prove to be very efficient for large-scale extraction of Arabic MWTs.

[9] created a MWTs extraction tool by adopting the hybrid approach. The first step is the extraction of the MWTs candidats using a set of 3 syntactic patterns taking into consideration morphological variants. The second step use several statistical scores like: T-Score, FLR, Mutual Information and LLR to rank the extracted MWTs. The authors used an Arabic corpus to calculate the precision of each statistical method and a collection of Arabic MWTs for validation. The experiments shows that the LLR gave the best results: 85% in terms of precision.

A similar work was presented in[13] implement the hybrid approachto extract bigrams . For the linguistic step, the authors used morpho-syntactic analysis to extract two categories of MWTs candidats: sequences of nouns and sequences of nouns separated with a preposition. The statistical filter include the use of the two statistical measures : the C-value and the LLR metrics. A corpus composed of 522845 is used to implement the extraction system. The authors used two methods of validation: the first one consider a MWT is correct if the translation of the MWT candidat is included in a terminologycal database and the second one in the manual validation. The experiments shows that using a combination of the two previous metrics in ranking MWTs, gives better results than using only one of them especially if the number of MWTs is increasing .

MULTI-WORD TERMS EXTRACTION SYSTEM

Our system is based on the hybrid approach and performs in two magor steps:

Linguistic filter

The linguistic filter has a major importance due to its contribution in the very early selection of MWTs candidate terms. The linguistic filter covers the following steps:

1. **Document pre-treatement**: This task covers the unification of documents encoding to avoid any ambiguity, elimination of Latin words, symbols, numbers, Roman numeral, special characters...
2. **Hamza ambiguity**: Although the two words:"امام" and "أمام" have the same meaning, they will be treated as different words due to the ambiguity induced by the letter "أ" . In order to eliminate this ambiguity, we replaced all "أ" ,"آ", "إ" occurrences in the corpus and stop-word list by "ا".

3. **Sentence boundary determination**: To extract MWTs from documents, We implemented a program that breaks up the corpus documents to sentences. The full stop is considered to be the sentence delimiter.
4. **Document POS-tagging**: We assign morphological tickets to the corpus documents sentences using The Stanford Arabic POS Tagger. This step will help us to detect possible MWTs following the patterns bellow:
   - [Noun]+
   - Noun; [Adjective]+
   - Noun; Preposition; Noun.

   In order to extract multi-word terms, the document sentences are scanned for sets of words that conform to one of the patterns above and ordered by their number of occurrences. The linguistic filter allows to extract MWTs candidates with various sizes; Bigrams, Trigrams and Four-grams.
5. **Stop-Word filter**: We eliminate the extracted MWTs beginning with a stop-word using a 600 noisy words list. Since we are using non-stemmed corpus, we implemented a program that assigns connectors like: "ب", "ك", "ف" ,"و", "ل" to stop-words in order to create all the variations of words in the list. For instance, the programm output for the word "هذه" is: "هذه" → "بهذه", "كهذه", " فهذه ", "وهذه ", "لهذه". Thus, MWTs such as:"الأمس القريب" are eliminated.

Statistical filter

To reduce linguistic ambiguities and increase the ratio of correct extracted MWTs, we combined two well known methods for their high effectiveness in MWTs extraction:

- LLR [14] a unithood method used to qualify the association between two words in Bigrams by calculating the ratio between two likelihoods: the probability of observing one component of a collocation given the other is present and the probability of observing the same component of a collocation in the absence of other. TABLE I represents LLR contingency table whereas, TABLE II describes the symbols used in LLR definition.

Table 1: Centingency Table

	V=v	V≠ v
U =u	O11	O21
U≠ u	O12	O22

Table 2: Centingency Parameters

Parameter	Description
U	First word of the bigram
V	Second word of the bigram
O11	Number of compound nouns with U and V .
O12	Number of compound nouns with U but without V .
O21	Number of compound nouns with V but without U.
O22	Number of compound nouns without U and without V. .

The LLR metric is given by the formulas:

$$\begin{cases} R1=O11+O12 \\ R2=O21+O22 \\ N=R1+R2 \\ r=\frac{R1}{N} \\ L(k,n,r)=r^{k}\times(1-r^{n-k}) \\ LLR=-2\log\left\{\frac{L(O11,R1,r)*L(O12,R2,r)}{L(O11,R1,\frac{O11}{R1})*L(O12,R2,\frac{O12}{R2})}\right\} \end{cases} \quad (1)$$

- C-value metric[15] a termhood statistical method based on the frequency of occurrence that gives best results for nested MWTs ranking. The C-value measure comes together with a computationally efficient algorithm, which scores candidate multi-token terms according to the measure. Fig. 2 describes the C-value formula.

$$C-value=\begin{cases}\log 2|a|\, f(a) & \text{, if a is not nested}\\ \log 2|a|\, f(a)-\frac{1}{p(Ta)}\sum_{b\in Ta} f(b) & \text{. Otherwise}\end{cases} \quad (2)$$

Table 3: C-value Parameters

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

We used The C-value metric for the nested words and their variations; the LLR metric was used for the remaining MWTs Bigrams.

EXPERIMENTS AND RESULTS

DATASET

For the set up of our experiments, we used a corpus of over 20.291 articles, collected from the Arabic newspaper Wattan of the year 2004[16]. The corpus contains articles covering the six following topics: culture, economics, international, local, religion and sport. The repartition of documents is described in Table 3.

Table 4: Number of documents and words per topic.

Topics	Number of articles
Culture	2782
Economy	3468
International	2035
Local	3596
Religion	3860
Sports	4450

EVALUATION METHOD

The evaluation of a MWTs extraction system is a very difficult task because of the absence of an evaluation standard of the MWTs, which are language and domaine dependent. In general, two categories of evaluation methods are used :

- **The manual validation** use a humain expert with a linguistic knowledge. The humain judgement is more correct . However this method require more sources and time in the case of large corpora.
- **The use of dictionnaries and standars** is realized automatically based on a comparaison between the output of the MWTs extraction system and the dictionaries. Although, this method is useful in case of large corpora, the lack of standred dictionaries make the comparaison difficult and non objectif.

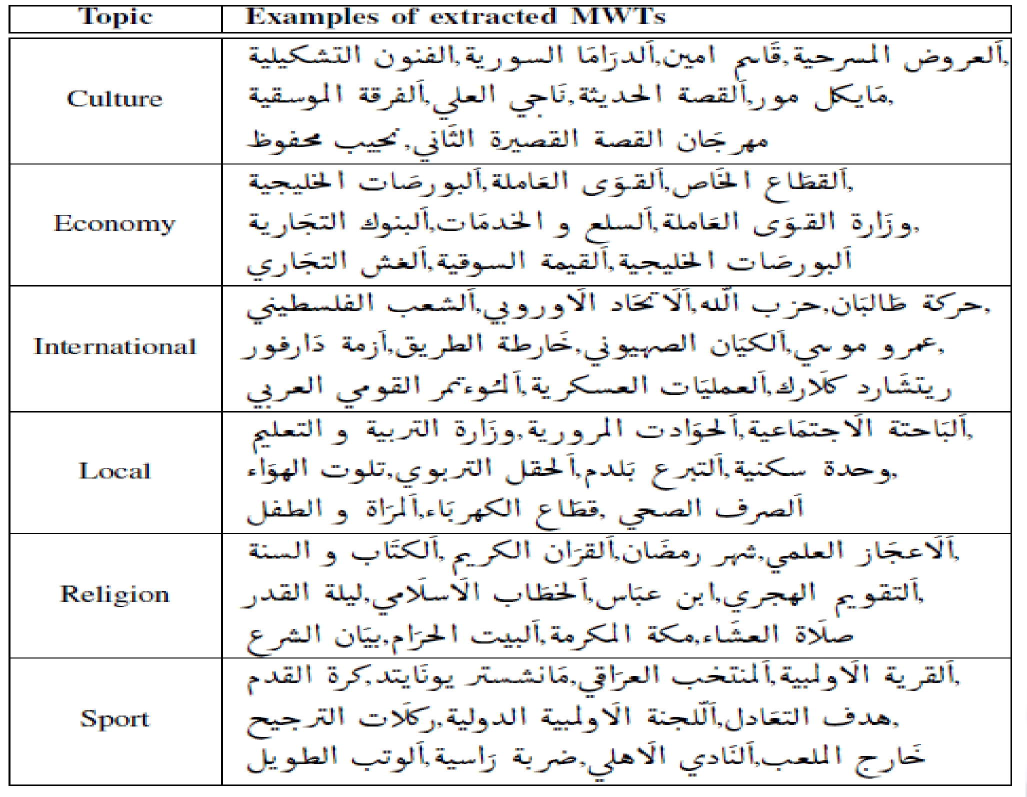
To evaluate the MWTs extraction system developed , we use the manual validation through the n-first muli-words evaluation method [17]. This method works on tree steps: first, the selection of the liste of the n-first MWTs using the list of the MWTs extracted sorted according to their scores obtained using the LLR and the C-value. We consider only the first n MWTs having the best scores. Then, we proceed to the manual evaluation of the n-first MWTs list with the help of a human expert. Finally, the system precision is calculated according to the following formula :

$$Precision = \frac{Number\ of\ correctly\ extracted\ MWTs}{number\ of\ extracted\ MWTs} \quad (3)$$

RESULTS

MWTs Extraction system

The Multi-Word Terms extraction system allow the extraction of terms composed of 2 to 6 words. Fig. 1 gives examples of extracted MWTs for each of the six topics of the corpus:



Topic	Examples of extracted MWTs
Culture	اَلعروض المسرحية,قَاسم امين,اَلدرَامَا السورية,الفنون التشكيلية, مَايكل مور,اَلقصة الحديثة,نَاجي العلي,اَلفرقة الموسقية, مهرجَان القصة القصيرة الثَاني,نجيب محفوظ
Economy	اَلقطَاع الخَاص,اَلقوَى العَاملة,اَلبورصَات الخليجية, وزَارة القوَى العَاملة,اَلسلع و الخدمَات,اَلبنوك التجَارية, اَلبورصَات الخليجية,اَلقيمة السوقية,اَلغش التجَاري
International	حركة طَالبَان,حزب الله,اَلاتحَاد الاَوروبي,اَلشعب الفلسطيني, عمرو موسي,اَلكيَان الصهيوني,خَارطة الطريق,اَزمة دَارفور, ريتشَارد كلَارك,اَلعمليَات العسكرية,اَلمؤتمر القومي العربي
Local	اَلبَاحتة الاَجتمَاعية,اَلحوَادت المرورية,وزَارة التربية و التعليم, وحدة سكنية,اَلتبرع بَلدم,اَلحقل التربوي,تلوث الهوَاء, اَلصرف الصحي ,قطَاع الكهربَاء,اَلمرَاة و الطفل
Religion	اَلاَعجَاز العلمي,شهر رمضَان,اَلقرَان الكريم ,اَلكتَاب و السنة, اَلتقويم الهجري,ابن عبَاس,اَلخطَاب الاَسلَامي,ليلة القدر, صلَاة العشَاء,مكة المكرمة,اَلبيت الحرَام,بيَان الشرع
Sport	اَلقرية الاَولمبية,اَلمنتخب العرَاقي,مَانشستر يونَايتد,كرة القدم, هدف التعَادل,اَللجنة الاَولمبية الدولية,ركلَات الترجيح, خَارج الملعب,اَلنَادي الاَهلي,ضربة رَاسية,اَلوتب الطويل

Fig 1: Examples of extracted MWTs per Topic.

The results showed in Fig. 2 and Fig. 3 give the precision of the developed system for several subsets of extracted MWTs with sizes: 25, 50, 100 and 200 respectively. Fig. 2 illustre the MWTs extraction system precision with a precision average of 89.44%. We observe that the topic “Ecomony” gave the weakest performance in comparaison with the other topics. This can be explained by the nature of the topic which don’t require using a lot of MWTs during documents redaction. The rest of the topics show good performances.

After the elimination of the MWTs containing one or more stopwords, the precision of the system increased to reach an average precision of 90.23% (Fig. 3). The obtained results show that using a stopword filter has improved the system performances. However, using a general and independent stopword list decreased the performances of some topics such as: religion where MWTs like: “ عليه السلام” and “ليال عشر” have been deleted since they contain stopwords. We conclude that using a stopword filter helps to improve the general performance of the system. However, the impact of using this filter depend on the topic nature (literature, scheintific, journalistic, …).

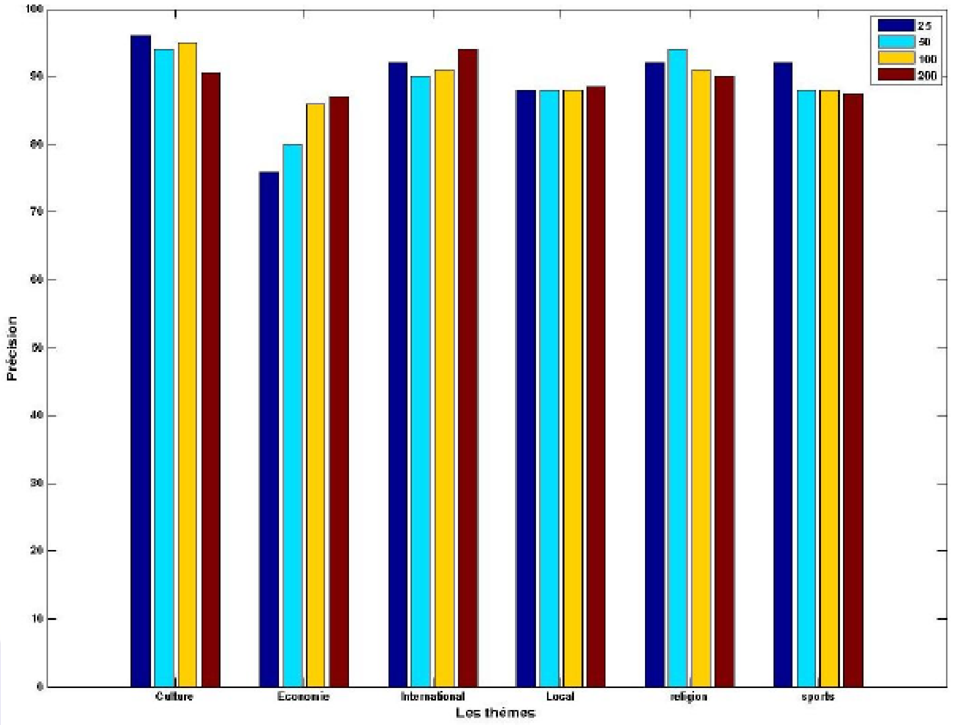


Fig 2: Precision of the MWTs extraction system.

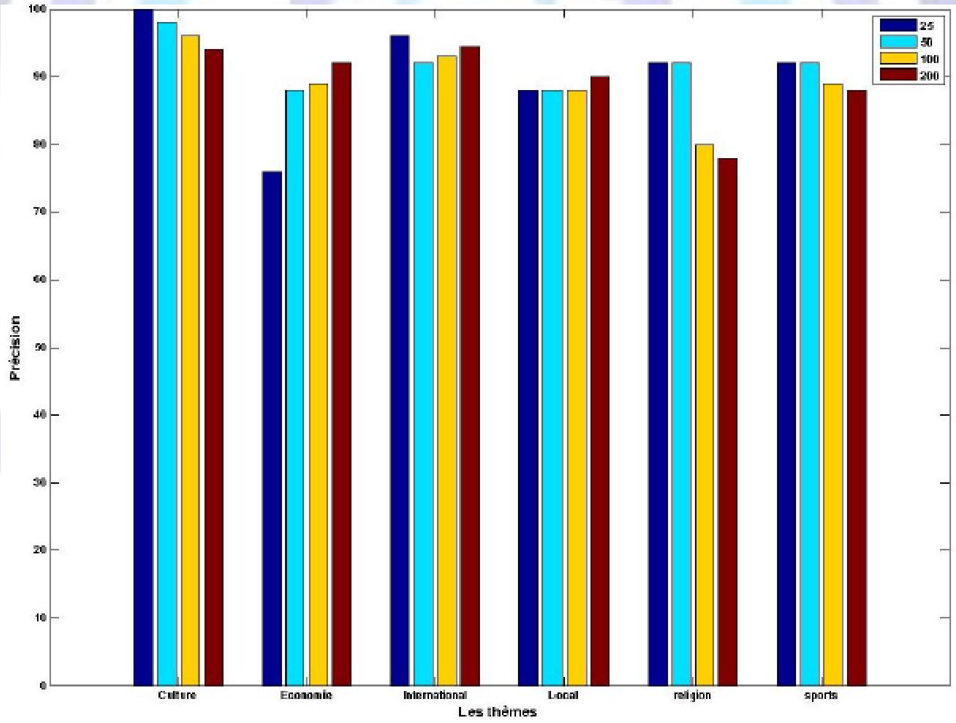


Fig 3: Precision of the MWTs extraction system after using the stoword filter

Topic Dection with MWTs features

Since the MWTs extraction aims to extract specific terms from special copora , we decided to study the impact of using MWTs as feature in the Arabic topic detection. In concordance with an earlier work [19], We built a topic detection system based on Topic Oriented Vocabularies (TOV), Jaccard indicator and an adaptation of the TF-IDF classifier. We conducted experiments using MWTs as features of the TOV. To the best of our knowledge, it's the first time an Arabic detection topic system employs MWTs vocabulary. Fig. 4 shows the results obtained in terme of F1-measure. The average F1-measure of the topic detection system is 83.46%, the average is 84.10% and the average recall is 85.81%, for documents containing MWTs.

As shown in Fig. 4, the system achieves higher performances for: religion and sports topics. This can be explained by the specificity of the MWTs extracted for these topics and the literature nature of the other topics which produces some ambiguity. The topic “Economy” present the lowrest performances in concordance with the results obtained earlier.

We conclude that the performance of the topic detection system based MWts depends on the topics wich confirm that the MWTs depends on the topics and their nature.

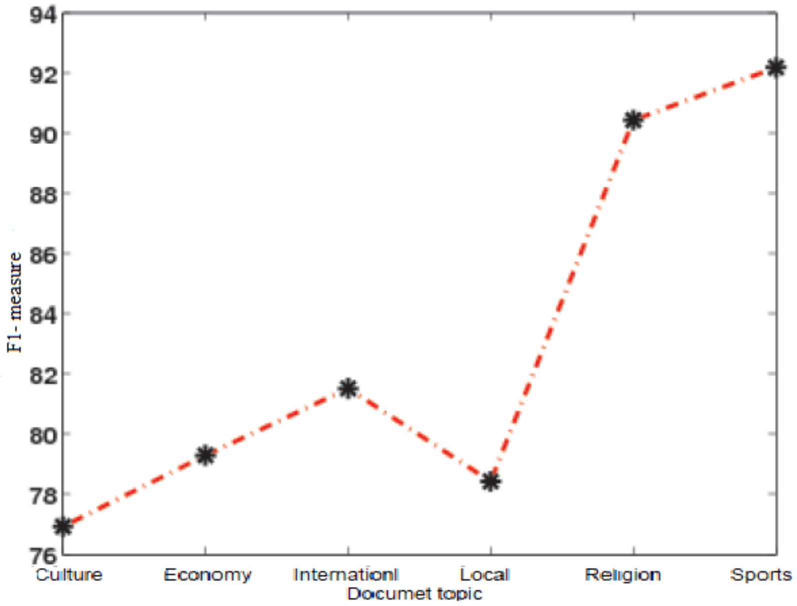


Fig 4: F1-measure of the topic detection system using MWTs based vocabulary.

Conclusion

We developed a multi-word terms extraction system for Arabic electronic documents based on linguistic patterns and the use of two statistical methods: LLR and C-value. We were able to extract words with bigrams, trigrams and four-grams.

We tested our system on an Arabic corpus covering six topics. Our system manages to achieve 90.23% of correctly extracted MWTs in terms of precision. We showed that our system gives higher results for topics: culture, religion and sports.

We also studied the impact of a stopword filter in the MWTs extraction system. The experiments results show an enhancement in the avarage performances of the developed system.However, this filter can decrease the performances according to the topics.

We developed a topic detection system that use MWTs as vocabulary features for Arabic documents. The results obtained show that using the MWts as feature lead to having a good performance of the topic detection system.

In a near future, we plan to experiment novel statistical methods other than LLR and C-Value to enhance our system performances. We also intend to studied the use of a topic-dependent stopword fiter on the sustem performances.

References

[1] Sag, I. A., Baldwin, T., Bond, F., Copestake, A., & Flickinger, D. (2002). Multiword expressions: A pain in the neck for NLP. In *Computational Linguistics and Intelligent Text Processing* (pp. 1-15). Springer Berlin Heidelberg.

[2] SanJuan, E., & Ibekwe-SanJuan, F. (2006). Text mining without document context. *Information Processing & Management*, *42*(6), 1532-1552.

[3] Nivre, J., & Nilsson, J. (2004, May). Multiword units in syntactic parsing. In*Workshop on Methodologies and Evaluation of Multiword Units in Real-World Applications*.

[4] Attia, M. A. (2006). Accommodating multiword expressions in an Arabic LFG grammar. In *Advances in Natural Language Processing* (pp. 87-98). Springer Berlin Heidelberg.

[5] Deksne, D., Skadins, R., & Skadina, I. (2008, May). Dictionary of Multiword Expressions for Translation into highly Inflected Languages. In *LREC*.

[6] Zhang, W., Yoshida, T., & Tang, X. (2008). Text classification based on multi-word with support vector machine. *Knowledge-Based Systems*, *21*(8), 879-886.

[7] Van de Cruys, T., & Moirón, B. V. (2007). Lexico-semantic multiword expression extraction. In *Proceedings of the 17th Meeting of Computational Linguistics in the Netherlands (CLIN)* (pp. 175-190).

[8] Vintar, S., & Fiser, D. (2008). Harvesting Multi-Word Expressions from Parallel Corpora. In *LREC*.

[9] Boulaknadel, S., Daille, B., & Aboutajdine, D. (2008, May). A Multi-Word Term Extraction Program for Arabic Language. In *LREC*.

[10] Duan, J., Zhang, M., Tong, L., & Guo, F. (2009). A hybrid approach to improve bilingual multiword expression extraction. In *Advances in Knowledge Discovery and Data Mining* (pp. 541-547). Springer Berlin Heidelberg.

[11] Moirón, B. V., & Tiedemann, J. (2006, April). Identifying idiomatic expressions using automatic word-alignment. In *Proceedings of the EACL 2006 Workshop on Multi-wordexpressions in a multilingual context* (pp. 33-40).

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

[13] Al Khatib, K., & Badarneh, A. (2010, October). Automatic extraction of arabic multi-word terms. In *Computer Science and Information Technology (IMCSIT), Proceedings of the 2010 International Multiconference on* (pp. 411-418). IEEE.

[14] Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational linguistics*, *19*(1), 61-74.

[15] Frantzi, K. T., & Ananiadou, S. (1996, August). Extracting nested collocations. In *Proceedings of the 16th conference on Computational linguistics-Volume 1*(pp. 41-46). Association for Computational Linguistics.

[16] Abbas, M., Smaili, K., & Berkani, D. (2010). Tr-classifier and knn evaluation for topic identification tasks. *The International Journal on Information and Communication Technologies (IJICT)*, *3*(3), 65-74.

[17] Evert, S., & Krenn, B. (2005). Using small random samples for the manual evaluation of statistical association measures. *Computer Speech & Language*,*19*(4), 450-466.

[18] Basili, R., Moschitti, A., Pazienza, M. T., & Zanzotto, F. M. (2001). A contrastive approach to term extraction. In *Terminologie et intelligence artificielle. Rencontres* (pp. 119-128)..

[19] KOULALI, R., & MEZIANE, A. (2013). EXPERIMENTS WITH ARABIC TOPIC DETECTION. *Journal of Theoretical & Applied Information Technology*, *50*(1).