SEMANTIC EXTRACTION OF ARABIC MULTIWORD EXPRESSIONS

Samah Meghawry[1],*, Abeer Elkorany[2], Akram Salah[2], and
Tarek Elghazaly[1]

[1]Institute of statistical studies and research, computer science, Cairo University
samah1984@gmail.com, tarek.elghazaly@cu.edu.eg
[2] Faculty of Computers and information, computer science, Cairo University
a.korani@fci-cu.edu.eg, a.salah@fci-cu.edu.eg

ABSTRACT

A considerable interest has been given to Multiword Expression (MWEs) identification and treatment. The identification of MWEs affects the quality of results of different tasks heavily used in natural language processing (NLP) such as parsing and generation. Different approaches for MWEs identification have been applied such as statistical methods which employed as an inexpensive and language independent way of finding co-occurrence patterns. Another approach relays on linguistic methods for identification, which employ information such as part of speech (POS) filters and lexical alignment between languages is also used and produced more targeted candidate lists. This paper presents a framework for extracting Arabic MWEs (nominal or verbal MWEs) for bi-gram using hybrid approach. The proposed approach starts with applying statistical method and then utilizes linguistic rules in order to enhance the results by extracting only patterns that match relevant language rule. The proposed hybrid approach outperforms other traditional approaches.

KEYWORDS

Multiword expressions (MWEs), Statistical Measures, Part of speech tagging (POS), Nominal MWEs, verbal MWEs.

1. INTRODUCTION

Recent research on Multiword Expressions (MWEs) has devoted considerable attention to their identification. One of the problems that these works address is that MWEs can be defined as combinations of words that have idiosyncrasies in their lexical, syntactic, semantic, pragmatic or statistical properties. There is no uniform definition of MWEs. The definition of MWEs given by Sag is "any word combination for which the syntactic or semantic properties of the whole expression cannot be obtained from its parts" [12].In other words, Multiword expressions are groups of words which, taken together, can have unpredictable semantics. MWE is an important task in many applications such as automatic translation [1], ontology engineering and information retrieval [2]. There are two main approaches for extracting MWEs. The statistical approach that uses a set of standard statistical association measures based on frequency and co-occurrence such as T-score [3], log likelihood ratio (LLR) [4], FLR [5] and Mutual Information (MI3) [6] in order

Dhinaharan Nagamalai et al. (Eds) : CCSEA, DKMP, AIFU, SEA - 2015
pp. 21–31, 2015. © CS & IT-CSCP 2015 DOI : 10.5121/csit.2015.50203

to estimate the degree of association between its words. The second approach makes use of the rules of the language such as morphological, syntactic or semantic information implemented in language-specific rules. Alignment-based MWE extraction method, which lends itself to linguistic approach, looks for the sequences of source wordsthat are frequently joined together during the alignment despite the number of target words involved. These MWE candidates may then be automatically validated, and the noisy non-MWE cases among them removed

However, each of those approaches suffers from great limitation [7], for example, statistical approaches "are unable to deal with low-frequency of MWEs". On the other hand, linguistic approaches are "language dependent and not flexible enough to cope with complex structures of MWEs". In order to overcome these weaknesses, a hybrid approach that combines statistical calculus and linguistic information is used. This paper proposes a framework for extracting Arabic Multiword Expressions from unannotated corpus using hybrid model that rely on frequency counts, statistical measures, and linguistic rules in order to create a refined list of candidates MWE. During the first phase of the proposed approach, lexical association measures based on the frequency distribution and co-occurrence patterns is applied in order to extract the first candidate set of MWE. Next, linguistics rules that utilize POS-tagger are applied to exclude specific patterns that match the relevant POS patterns according to Arabic grammar rules. In order to validate the effectiveness of the proposed model, three different Arabic corpuses were used during our experiments. Our experiments confirmed that the proposed approach outperform previous methods. This paper is organized as follows; Section2 presents different approaches applied for extracting MWEs for various languages. In section3 the proposed hybrid framework for Arabic MWE is illustrated. Results of experiment applied using different Arabic corpus are discussed in section4. Finally, section5 concludes the presented work and demonstrate potential future works.

2. RELATED WORK

A considerable amount of research has focused on the identification and extraction of MWEs. Given the heterogeneity of MWEs, different approaches were devised. Unfortunately, unlike in English, there is no capital letters in Arabic to distinguish the compound names and the geographical compound names. Statistical approaches have mostly been applied to bigrams and trigrams, and it becomes more problematic to extract MWEs of more than three words. Pecina evaluates 82 lexical association measures for the ranking of collocation candidates and concludes that it is not possible to select a single best universal measure, and that different measures give different results for different tasks depending on data, language, and the types of MWE that the task is focused on [14]. Similarly, Ramisch investigate the hypothesis that MWEs can be detected solely by looking at the distinct statistical properties of their individual words and conclude that the association measures can only detect trends and preferences in the co-occurrences of words [13]. The linguistic methods are based on linguistic information such as, morphological, syntactic and/or semantic information to generate the types of words. Traboulsi used the local grammar approach to extract person names from Arabic counterparts [11].Harris defines a local grammar as a way of describing syntactic restrictions of certain subsets of sentences, which are closed under some or all of the operations in the language. Frozen expressions may be considered as a subset of sentences that have such syntactic restrictions. One can in fact observe restricted distributions over a number of words. Consider for example: Director of (company + thesis + conscience + *chocolate) (financial + stock + E) market The 20 March (next + 2006 + *bombastic) [10].

Hybrid approaches that combine the statistical approaches with the linguistic rules can cover a large part of the problem of MWEs identification and extraction [9]. Boulaknadel developed a multi-word term (MWT) extraction tool for Arabic. She adopted the standard approach that combined grammatical patterns and statistical score. First, she defined the linguistic specification of MWTs for Arabic language. Then, she developed a term extraction program and evaluated several statistical measures in order to filter the extracted term-like units for keeping the most representative of domain specific corpus [7].Hybrid approaches may also combines the alignment technique with statistical approach like Helna [16] that proposed an approach for the identification of MWEs in a multilingual context, as a by-product of a word alignment process, that not only deals with the identification of possible MWE candidates, but also associates some multiword expressions with semantics.

3. Hybrid Model For Arabic MWE Extraction

The proposed model aims to extracts multi-word expressions from Arabic specialized corpora by combining statistical methods with linguistic rules. The standard approach to MWE identification is n-gram classification. However, our model is limited to multi-words composed of two elements (bigrams). This section discusses three different phases of the proposed model- the preprocessing phase, statistical phase and linguistic phase.

3.1 Preprocessing phase

Text preprocessing is the basic stage needed for MWE. Its main objective is, in one hand to remove all the unnecessary particles and mistyping words and in another hand to transform document contents to a suitable form which can be used easily by different algorithm. Thus during the preprocessing phase, we start by splitting the corpus to set of words, cleaning the corpus from delimiters and symbols, storing each two consecutive words in the corpus into database.

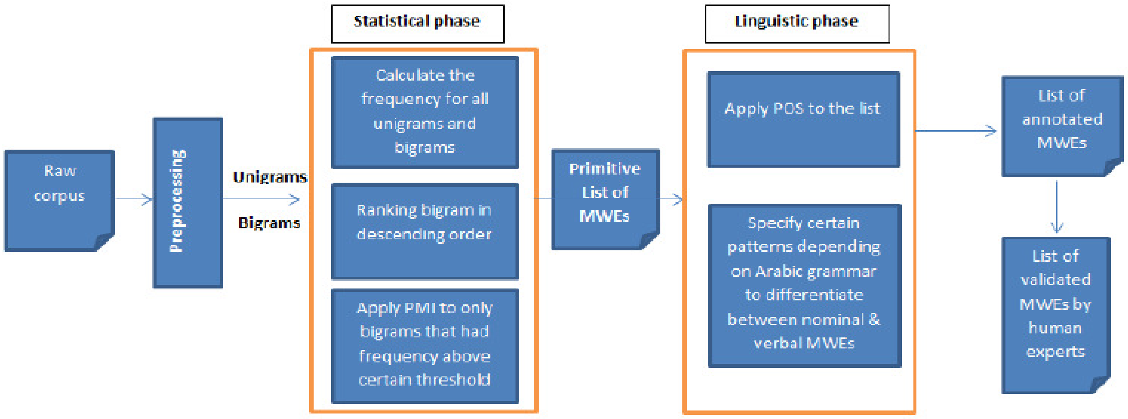

Statistical phase
Linguistic phase
Raw corpus
Preprocessing
Unigrams
Bigrams
Calculate the frequency for all unigrams and bigrams
Ranking bigram in descending order
Apply PMI to only bigrams that had frequency above certain threshold
Primitive List of MWEs
Apply POS to the list
Specify certain patterns depending on Arabic grammar to differentiate between nominal & verbal MWEs
List of annotated MWEs
List of validated MWEs by human experts

Fig.1. Architecture of the proposed hybrid framework for Extracting MWEs

3.2 Statistical phase

Association measures are inexpensive and language-independent means for discovering recurrent patterns, or habitual collocates. Association measures are defined by Pecina[14] as mathematical formulas that determine the strength of the association, or degree of connectedness, between two or more words based on their occurrences and co-occurrences in a text. The higher the connectedness between words, the better the chance they form a collocation. One of widely applied method is Point-wise Mutual Information (PMI) [9] that compares the co-occurrence probability of words given their joint distribution and given their individual (marginal) distributions under the assumption of independence. For two-word expressions, it is defined as:

$$PMI(x,y) = log_2 \frac{p(x,y)}{p(x,*)p(*,y)}$$

Where p(x, y) is the maximum likelihood (ML) estimation of the joint probability (N is the corpus size):

$$p(x,y) = \frac{f(x,y)}{N}$$

And p(x,*), P(*, y) are estimations of marginal probabilities computed in the following manner:

$$p(x,*) = \frac{f(x,*)}{N} = \frac{\sum_y f(x,y)}{N}$$

And analogically for P(*, y).

The following steps were applied during phase1 of the proposed model

1. Calculate the frequency of all the unigrams and bigrams in the corpus.
2. Calculate the PMI to all bigrams that have a frequency above certain threshold
3. Bigrams are ranked in descending order.

Here in this stage we have a list of MWEs with its PMI sorted in descending order.

3.3 Linguistics filtering of Arabic MWE

Extracting MWEs using statistical approach depends on the idea of occurrences and co-occurrences of two words would lead to generate patterns that may not be MWEs such as " ذَيال ◌ٍ◌ًضيو " or "انحكى حول ". Those bigrams repeated many times in the same corpus but are not considered a MWE. Thus, it is important to utilize linguistic rules to identify the correct MWEs

from the ranked list of MWEs generated by the previous statistical phase. These linguistic rules are illustrated in this subsection.

3.3.1 Selected Linguistic rules.

In order to be considered as a multi-word expression, a sequence of words should fulfill syntactic and semantic conditions. In fact, we can distinguish many types of MWEs [15] such as:

- Idioms (e.g.وَر انعهى)
- Phrasal verbs (e.g. عهي ذً يعر)
- Verbs with particles (e.g. عٍ يعفو)
- Compound nouns (e.g. زْاو الاً جزيذج)
- Collocations (e.g. يعزوف مًّ إع)

Furthermore, a compound noun belongs to one of the following categories:

- Annexation compound noun (الاضافي انرزكية): an expression composed of an indefinite noun and one of the following elements:

– A possessive pronoun (e.g. طيارذّ : his car),
– Any simple or compound definite noun (e.g. عهي طيارج : the car of Ali),
– An indefinite adjective compound noun (e.g. يٌ غ رجم طيارج : the car of a rich man).

The first component is called ضًافا ن (first term of annexation) while the second is called ضًاف ان إنيّ (second term of annexation). The definiteness of the compound noun is equal to the definiteness of the second component.

- Adjective compound noun (انوصفي انرزكية): an expression composed of a noun (either simple or compound) which is called "عُوخ ي" (The modified word) and an adjective (ان عُد) having the same definiteness (e.g.يُ غ رجم :a rich man). The gender of the two elements must be agreed.
- Substitution compound noun (انثذل انرزكية): an expression composed of a demonstrative pronoun and a definite noun (e.g.انظيارج ذْ, this car). Such expression is always definite.
- Prepositional compound noun: two nouns linked by a preposition (e.g. انحهواء ي وَع : a kind of sweet).
- Conjunctive compound noun: two nouns linked by a conjunction (e.g.وانفأرا نقظ : the cat and the mouse).
- Compound nouns linked by composite relations: two or more linkers (prepositions and/or conjunctions) are used to link two nouns (e.g. حُ ط نحواني زًار الاطر : To persist for about one year).

Since the proposed framework is applied only for bigram, only linguistic rules for adjective compound noun and substitution compound noun are applied as shown in figure2.

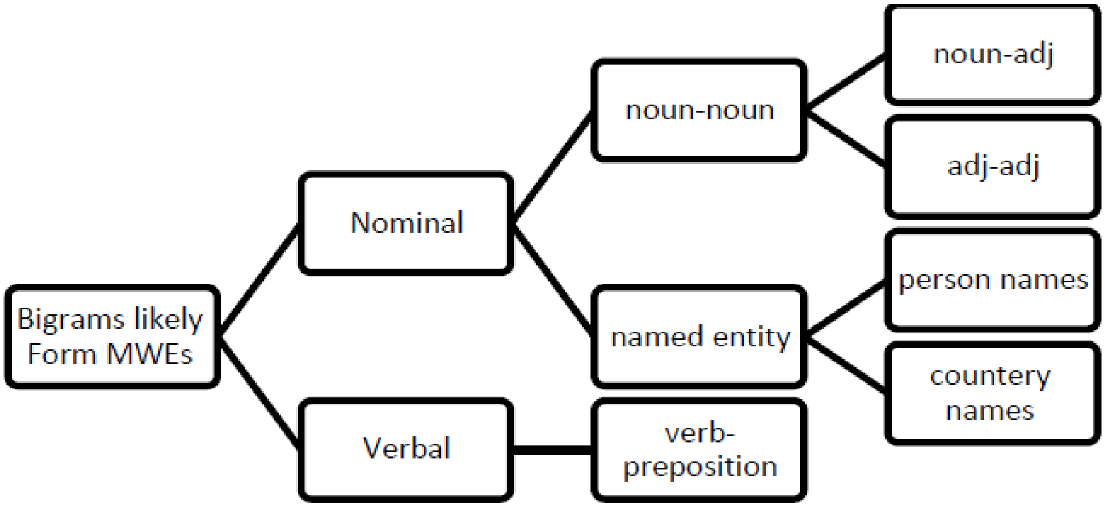


Fig. 2 Sample of used linguistic patterns

Furthermore, we also applied some linguistic rules of verbs such as verbs with particle that represents verb followed by preposition like "إني أدى "," في شارك "," في فشم " or"عهي يضي ".

3.3.2 Filtering Identified pattern

As explained above, the list of ranked bigrams is applied to part of speech tagger (POS) in order to identify the type of the words (noun, verb, preposition or etc.). This framework uses the Stanford POS tagger -a piece of software that reads text in some language and assigns parts of speech to each word (and other token), such as noun, verb, adjective, etc., although generally computational applications use more fine-grained POS tags like 'noun-plural'. Part-of-speech tags are assigned to each single word according to its role in the sentence. Traditional grammar classifies words based on eight parts of speech: the verb (VB), the noun (NN), the pronoun (PR+DT), the adjective (JJ), the adverb (RB), the preposition (IN), the conjunction (CC), and the interjection (UH)- http://www.clips.ua.ac.be/pages/pattern. Next, the linguistic rules illustrated in figure2 are applied to those tagged pattern to extract more meaningful pattern. . It is significant to mention that the main objective of applying linguistic rules after using statistical approach is to limit the scope of the MWE identification process where the experiment yields many bigrams that had a high frequency generated in the statistical phase list like "تعذ ي ". This bigram had a high frequency but did not represent actual MWEs so it was filtered in the linguistic phase according to the patterns specified in fig.2.

4. EXPERIMENT

Three different corpus were used in our experiment. The first one, archives from Omani newspaper Alwatan of the year 2004 [8]- https://sites.google.com/site/mouradabbas9/corpora. The size of the extracted corpus is about 10 millions terms which correspond to 9000 articles, distributed over six topics, in this case: Culture, religion, economy, local news, international news and sports. The second corpus is the Arabic Newswire Part 1This publication contains the Arabic Newswire a Corpus, Linguistic Data Consortium (LDC) catalog number LDC2001T55 and ISBN

1-58563-190-6. The Arabic Newswire Corpus is composed of articles from the Agence France Presse (AFP) Arabic Newswire. The source material was tagged using TIPSTER-style SGML and was transcoded to Unicode (UTF-8). The corpus includes articles from May 13, 1994 to December 20, 2000. There are 209 Mb of compressed data (869 Mb uncompressed) with approximately 383,872 documents containing 76 million tokens over approximately 666,094 unique words. The third one is Named Entity Corpus from Arabic Language Technology Center "ALTEC"https://sites.google.com/site/mouradabbas9/corpora.

4.1 Experiment setup

The following pre-processing steps have been applied for the corpus:

- Cleaning the corpus from punctuations and symbols.
- Splitting it to set of unigrams and bigrams.
- Storing all unigrams and bigrams into database.

4.2 Results of Experiment

The first experiment was applied in order to identify the value of threshold that should be used during phase1 (statistical phase). Thus, we change the frequency used in the statistical phase from 20,30,40 and 50 respectively in order to study the effect of changing the threshold on the accuracy of the result. As shown in figure 3, with decreasing the frequency during statistical phase, the number of candidate MWE increases. As explained earlier, statistical phase did not consider any linguistic features, it only depend on the degree of connectedness between two or more word. Accordingly, increasing the number of obtained MWE from phase1 would lead to provide more set of candidate MWE to be used during linguistic phase and avoid missing any candidate MWE from corpus. However, linguistic phase plays a significant role in enhancing the final results as the number of final MWE dramatically decreased to almost half in all cases as shown in figure3.

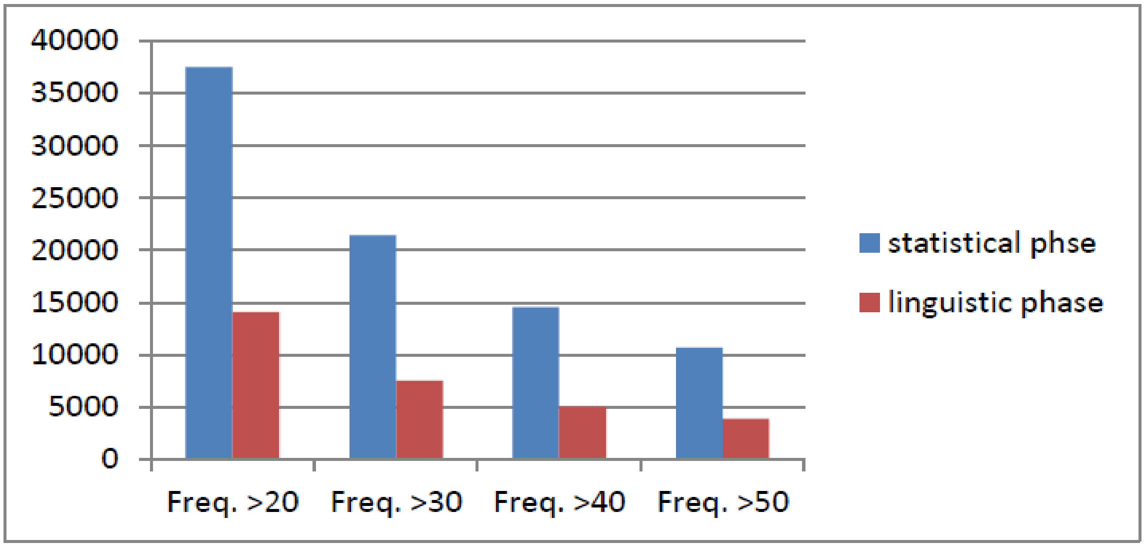
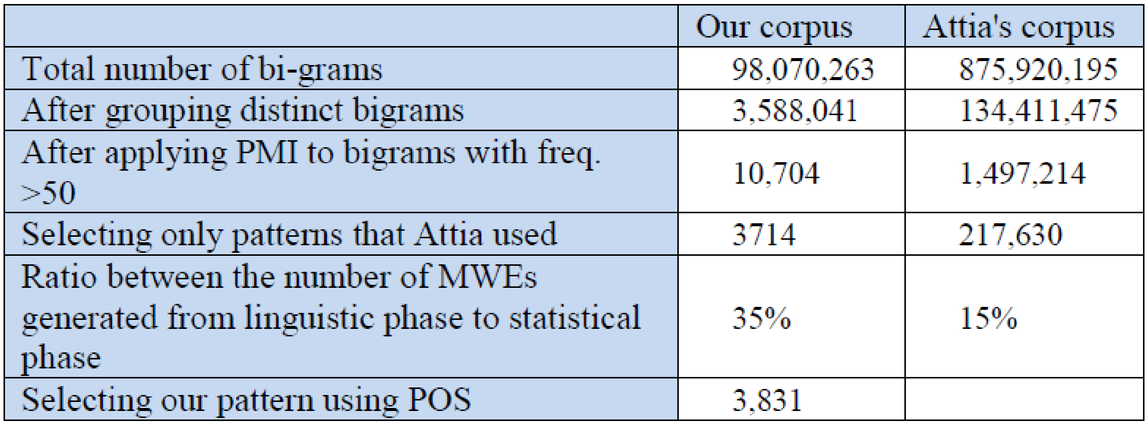


Fig.3 The effect of frequency change on number of MWEs in each phase

The aim of second experiment is to identify the number of candidate MWE after applying each phase and compare the results with the proposed framework by Attia [9]. Therefore, in this experiment during the statistical phase we set the frequency to 50 to be able to compare it with Attia (although this is not matching with the result obtained in the first experiments that recommend setting the frequency to 20). The results summarized in table1 shows that, the number of detected MWE after applying linguistic phase decreased to one-third and the final result outperformed the results obtained when applying statistical phase. According to table1, our proposed framework generate more final MWEs due to applying more linguistic rule (such as those related to verbs) that those proposed by Attia[9] which decrease the possibility of omitting significant MWEs patterns that are not of type (noun-noun, noun- adjective).



	Our corpus	Attia's corpus
Total number of bi-grams	98,070,263	875,920,195
After grouping distinct bigrams	3,588,041	134,411,475
After applying PMI to bigrams with freq. >50	10,704	1,497,214
Selecting only patterns that Attia used	3714	217,630
Ratio between the number of MWEs generated from linguistic phase to statistical phase	35%	15%
Selecting our pattern using POS	3,831	

Table 1. Comparison between the number of generated MWE using proposed model and Attia

Next, ground truth is used to identify the correct set of final list of MWEs. Therefore, we present the final list generated from proposed model as well as the list generaeted when applying Attia model to domain expert to validate the correctness of identified MWEs. Human experts have annotated the list obtained from both models in order to compute the precision of them as shown in table 2.

First word	Second word	MWE(1)/NON-MWEs (0)
الامم/DTNN	المتحدة/DTJJ	1
الولايات/DTNNS	المتحدة/DTJJ	1
فرانس/NNP	برس/NNP	1
وزير/NN	الخارجية/DTNN	1
مجلس/NN	الامن/DTNN	1
اليوم/DTNN	الخميس/DTNN	0
اليوم/DTNN	الأخير/DTNN	1
اطلاق/NN	النار/DTNN	1
اطلع/VBD	على/IN	1
كان/VBD	في/IN	0
لاطلاق/NNP	النار/DTNN	0
اسحق/NNP	رابين/NNP	1
لوس/NNP	انجليس/NNP	1
لحزب/NNP	الله/NNP	1

Table2 sample of the human expert annotation to the final list of MWEs

Finally, precision is calculated in order to compare the accuracy of our proposed with Attia. According to figure4, the value of precision increase to 67% when applying of the whole model compared to Attia (about 34 %). It is significant to mention that applying both nominal and verbal linguistic phase rule during increase the value of precision by 3%. This indicates that the verbal MWEs represent a smaller number of MWEs in comparison with nominal MWEs in Arabic language.

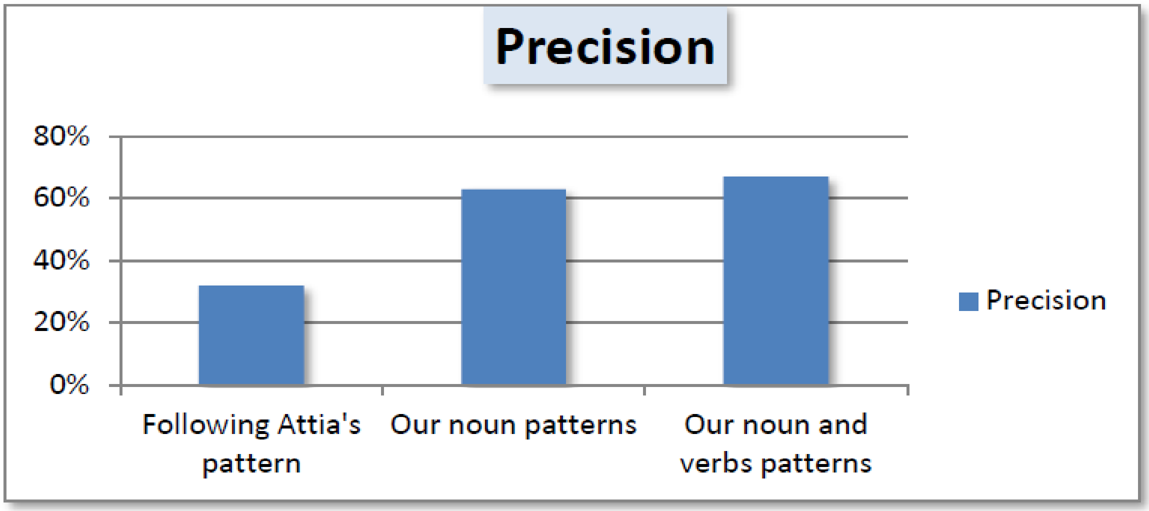


Figure4: Comparison between results of precision when applying Attia and our proposed framework.

5. CONCLUSION

The process of extracting MWEs is a very complicated task to be solved by one single solution. In this paper we develop a framework for extracting Arabic MWEs using hybrid approach that combine the statistical approach with the linguistic rules and the results obtained validated by human experts and the precision differed according to the threshold determined in statistical phase. We find that the more the threshold that set in the statistical phase is low the more we get greater number of MWEs, the statistical approach measures the connectedness of each two consecutive words in the corpus regardless these two words are MWEs or not so the linguistic approach increases the accuracy of the generated MWEs list from the statistical phase by filtering undetermined patterns, after applying our experiment into different data sources we find that the ratio between nominal MWEs and verbal MWEs in the list generated from phase1 and phase2 represents 97:3 respectively.

REFERENCES

[1] O. Kraif, (2003) “Repérage de traduction et commutation interlingue :Intérêt et méthodes”, Traitement Automatique des LanguesNaturelles TALN 2003, Batz-sur-Mer, France, June 11-14, 2003.

[2] V. Malaisé, (2005) “Méthodologie linguistique et terminologique pour la structuration d’ontologies différentielles à partir de corpus textuels”, doctoral thesis, University of Paris 7 – Denis Diderot, 2005.

[3] K.W. Church, W. Gale, P. Hanks, and D. Hindle, (1991) “Using statistics inlexical analysis”. In Lexical Acquisition, Exploiting On-Line Resourcesto Build a Lexicon, Hillsdale, Michigan, USA: Zernik Uri ed., London,Lawrence Erlbaum Associates, 1991, pp.115-164.

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

[5] H. Nakagawa, T. Mori, and H. Yumoto, (2003) “Term Extraction Based on Occurrence and Concatenation Frequency”, Journal of Natural Language Processing, vol. 10 (1), pp.27-45, 2003.

[6] B. Daille, (1994) “Approchemixte pour l’extraction de terminologie : statistiquelexicale et filtreslinguistiques”, doctoral thesis, University of Paris 7, 1994.

[7] S. Boulaknadel, B. Daille and D. Aboutajdine, (2008) “A multi-word term extraction program for Arabic language”, the 6th international Conference on Language Resources and Evaluation LREC 2008, Marrakech, Morocco, 28-30 May 2008, pp. 1485-1488.

[8] 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.

[9] Attia, M., Antonio Toral, Lamia Tounsi, PavelPecina and Josef van Genabith,(2010) “Automatic Extraction of Arabic Multiword Expressions”, In: Proceedings of the Workshop on Multiword Expressions: from Theory to Applications (MWE 2010), pp: 18–26,Beijing, China. 2010.

[10] Z. Harris, (1991) "Theory of Language and Information: A Mathematical Approach", Oxford & New York: Clarendon Press, 1991.

[11] Traboulsi, H,(2009) "Arabic named entity extraction: A local grammar-based approach", In: Proceedings of the International Multiconference on Computer Science and Information Technology, vol. 4, pp. 139–143 (2009) .

[12] Sag, Ivan A., Timothy Baldwin, Francis Bond, Ann Copestake and Dan Flickinger, (2002) "Multiword Expressions: A Pain in the Neck for NLP" In the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002), volume 2276 of Lecture Notes in Computer Science, pp. 1.15, London, UK. Springer-Verlag.

[13] Ramisch, Carlos, Paulo Schreiner, Marco Idiart and Aline Villavicencio, (2008), "An Evaluation of Methodsfor the Extraction of Multiword Expressions", In the Workshop on Multiword Expressions, the 6thInternational Conference on Language Resources and Evaluation (LREC 2008), pp. 50.53. Marrakech, Morocco.

[14] Pecina, Pavel, (2010) "Lexical association measures and collocation extraction", In Language Resources and Evaluation (2010), 44:137-158.

[15] Bounhas, I. and Y. Slimani, (2009) "A hybrid approach for Arabic multi-word term extraction", Proceeding of the International Conference on NLP-KE 2009, Department of Computer Science, University of Tunis, Sept. 24-27, Tunis, Tunisia, pp: 1-8. DOI: 10.1109/NLPKE.2009.5313728.

[16] Helena M. Caseli, Carlos Ramisch, Maria G. V. Nunes, and Aline Villavicencio, (2009) "Alignment-based extraction of multiword expressions", Language resources and evaluation 44 (1-2), 59-77.