Journal of Computer Science 7 (1): 6-11, 2011
ISSN 1549-3636

An Automatic Collocation Extraction from Arabic Corpus

Abdulgabbar Mohammad Saif and Mohd Juzaiddin Ab Aziz
Department of Computer Science, Faculty of Information Science and Technology,
National University of Malaysia, Malaysia

Abstract: Problem statement: The identification of collocations is very important part in natural language processing applications that require some degree of semantic interpretation such as, machine translation, information retrieval and text summarization. Because of the complexities of Arabic, the collocations undergo some variations such as, morphological, graphical, syntactic variation that constitutes the difficulties of identifying the collocation. **Approach:** We used the hybrid method for extracting the collocations from Arabic corpus that is based on linguistic information and association measures. **Results:** This method extracted the bi-gram candidates of Arabic collocation from corpus and evaluated the association measures by using the n-best evaluation method. We reported the precision values for each association measure in each n-best list. **Conclusion:** The experimental results showed that the log-likelihood ratio is the best association measure that achieved highest precision.

Key words: Collocation extraction, hybrid methods, collocation variations, Association measures, morphosyntactic, graphical variants, n-best evaluation

INTRODUCTION

The collocations issue is the linguistic phenomenon that is found in all the human languages. It is an important part in many applications, such as, machine translation, information retrieval, word sense disambiguation and lexicography. In a bilingual context, collocations are very important for learners of a language to construct the meaningful sentences. Usage of the right combinations, being a part of context, results in correct language production (speech) at least at the stylistic level.

There is no widely accepted definition of a collocation in the field of computational linguistics. For example, Evert defined the collocation as "A word combination who semantic and/or syntactic properties cannot be fully predicted from those of its components and which therefore has to be listed in a lexicon" (Evert, 2004). Another researcher, (Smadja, 1993) considered the collocations as " recurrent combinations of words that co-occur more often than expected by chance and that correspond to arbitrary word usages". According to (Pecina, 2010), there are some restrictions (semantic and/or pragmatic) that must be included in the extraction of collocations in order to produce the meaningful and fluent collocation. The semantic compositionality is to check whether the overall meaning of the collocation is obtained by the composition of the meanings of individual words.

In its simple definition, the collocation is defined as the two or more words which appear together and always seems as comrades. The collocation is the phenomenon of linguistic high productivity that makes for two words or more, in the confluence of what, attached to each other, combined permanently and does not change because the usage of a particular word. For instance, a noun has a small number of verbs or adjectives that can combine with this noun to construct the collocation. For example, in English, the noun crime has small number of verbs which combines with this noun to indicate the event of 'doing the crime'. The same can apply for an adjective and a verb. There are two verbs 'commit' or 'perpetrate' which can combine with this noun to indicate the action. As well as, this case can be applied in Arabic. If we take the noun الشعر in mind, the verbs حلق or قص can be combined with it. The verb 'قطع' can be used to denote the action, but the expression will be bad. On the other hand, the noun may need an adjective to describe it and constitute the collocation. For example, in English, the adjective that can combine with the noun tea is 'strong'; this noun can not combine with other adjective like powerful. The same situation in Arabic; with the noun قول one can combine a limited number of adjectives like سديد, صائب.

Corresponding Author: Abdulgabbar Mohammad Saleh Saif, Department of Computer Science,
Faculty of Information Science and Technology, National University of Malaysia, Malaysia

Related work: For collocation extraction, there are three main methodologies: the statistical, linguistic and the hybrid methods. Statistical methods (Smadja, 1993; Dunning, 1993; Van de Cruys and Moiron, 2006) are using frequency scores of candidate patterns to extract collocations from text. In general, those methodologies use the text in corpora and only require the association measure between the words in texts. However, many of the words that are extracted by using these methodologies cannot be considered as the true collocations although it may be useful to identify the textual associations in the context of their usage. The linguistic methods (Attia, 2006) are based on linguistic information such as, morphological, syntactic and/or semantic information to generate the collocations. However, they cannot deal with the flexibility of language and generate some type of collocations that have no productivity. The hybrid methods (Boulaknadel *et al*., 2008; Duan *et al*., 2009; Frantzi *et al*., 2000) are the combination of statistical information and linguistic knowledge. They have been proposed in order to avoid the disadvantages of the two methods. For example, Frantzi *et al*. (2000) present a hybrid approach to extract multi-word terminology from English corpora combining linguistic. From linguistic respective, their approach extracts the candidates of multiword terminology by using some linguistic information, such as, part-of-speech tagging of the corpus to use in the linguistic filter, the linguistic filter to cover all types of terminologies and produce useful result and the stop-list to avoid the extraction of candidates that are unlikely to be terminology and improving the precision of the output list. In addition, the C-value is used to ensure that the extracted candidate is real a MWLU. Their technique was compared with raw frequency filtering though they failed to take advances in MWE association measures into account. In Arabic, there are a few works that extract the MWT from corpus (Attia, 2006; Boulaknadel *et al*., 2008; Bounhas and Slimani, 2009). Attia (2006) presented the semi-automatic linguistic method for extracting some types of MWE. He used the regular expressions to identify the candidates of Arabic MWE and presented some linguistic variations such as, morphological, lexical and syntactic variations. Also, Boulaknadel *et al.* (2008) designed a multi-word term extraction program for Arabic language. They used a hybrid method to extract multi-word terminology from Arabic corpus. From linguistic respective, they used some linguistic information to extract and filter the candidates of multiword terminology. Their method uses the part-of-speech tagging of the corpus that has been assigned by the Diab *et al*. (2004) to use in the linguistic filtering. The linguistic filter is to identify the Arabic MWT patterns, such as, N ADJ, N N and N PREP N. In addition, their method takes into account the MWT variations, such as, graphical variants (the graphic alternations between the letters "ha'a" and "Ta'a marbutah"), Inflectional variants (the number inflection of nouns, the number and gender inflections of adjectives and the definite article (AL)), Morphosyntactic variants (the synonymy relationship between two MWTs of different structures.) and syntactic variants(the modifications of the internal structure of the base-term, without affecting the grammatical categories of the main item which remain identical). On the other hand, they used four association measures: log-likelihood ratio, FLR, Mutual Information (MI^3) and t-scores to order the candidates of MWT. In this paper, we will discuss some aspects of collocation for Arabic language and use the hybrid method for extracting the collocation from Arabic corpus.

Collocation variations: The automatic collocation extraction requires the determining of variations on the candidates extracted in order to improve the accuracy of the results. In this paper, we take into account three types of variation as the following: (1) Graphical variants: according to (Boulaknadel *et al*., 2008), the graphical variants are the graphic alternations between the letters 'haa' and 'taa marbutah'. The graphic alternations between these letters occur only when one of the letters is in the end of the word. For examples, 'المكتبة' or 'السياسه'; (2) The morphological variations: the morphological variations of noun include the number inflection of noun(singular, dual, or plural) , gender inflections and the definite article 'AL' that appears as the prefix of the noun. The same can count for adjective. The morphological variations of verb include the tense (present, past, or imperative), the number and the object pronoun; (3) Syntactic variations: the syntactic variations include the internal modification in Arabic expressions that allow external elements to intervene between the components. In most cases, the external elements may be the preposition or the conjunction that appears between the two words. In other cases, the external elements are the complement word that appears either in the beginning, middle, or at the end of the expression.

The structural patterns of Arabic collocation: In Arabic, the structural patterns of Arabic collocations can be classified into the following patterns (based on POS): (1) Noun + Noun: this is the expression that consists of two nouns with a space. For examples, 'رئيس الوزراء', or 'مكة المكرمة'; (2) Noun + Adjective: this type

corresponds to the adjective constituent (‘التركيب الوصفي’), in which, the first component is called ‘الموصوف’ and the second component is called ‘الصفة’. The components of this type have the same definiteness (without, or with the definite article for both). Also, they are inflected for number and gender. For examples, ‘الارض المقدسة’, or ‘علاج طبيعي’; (3) Verb + Noun (V+N): this is the expression that consists of verb and noun to form the collocation. The noun may be either subject or object. For examples, ‘ذكرت التقارير’, ‘اضاف البيان’, or ‘اشار المصدر’; (4) Verb + Adverb (V+ADV): ‘انتقد بشدة’, ‘اتصل هاتفيا’; 5) Adjective + Adverb (ADJ+ADV): ‘ صعب بعيدة’, ‘مستمر تماما’, ‘للغاية’; 6) Adjective + Noun (ADJ+N): ‘شديد اللهجة’, ‘المدى’.

MATERIALS AND METHODS

The steps of collocation extraction: pre-processing, candidate identification and candidate ranking. Pre-processing: Generally speaking, the corpus that is used to extract the collocation has to include the POS tagging for each lexeme in the corpus. But, there is no free available Arabic corpus to use for collocation extraction. So, we have collected an in-house corpus from online Arabic newspaper archives, including Almotamar.net and Al-Jazeera.net. The pre-processing step is responsible for the filtering corpus and generating unigram list. The filtering of corpus includes the normalization of different forms of (*hamza*) to (*alef*) and removing the all non-Arabic words and symbols from the corpus. The unigram list contains all words in corpus with their frequency and linguistic categories for each word.

Candidate identification: the candidate identification depends on linguistic analysis tools such as lemmatizers, POS taggers and/or parsers in order to cope with morphological and syntactic variations. In the current method, the candidate identification relies on lemmatization and POS in order to filter the candidates and determine the variations. This step includes two phases: generating the candidates and filtering. The first phase is to generate all bi-gram candidates from corpus. From the unigram list, we select only the words that their linguistic categories are corresponded to the first part of the structural patterns of collocations. These words with their frequency and linguistic categories are stored in the new list (called enhanced unigram list). From enhanced unigram list, for each word, we select all possible combinations of this word with another word from corpus to represent the bi-gram candidates. The linguistic categories of second part in the bi-gram candidate have to correspond to the second part of the structural patterns of collocations. Through this combination, if the first and the second word have only one linguistic category, the combination of two words is stored in the bi-gram list without any more processing; but if one of the word has more than one linguistic categories, we use POS tagger to disambiguate from the linguistic categories. There are many works in Arabic POS tagging, such as, a hybrid technique of statistical and rule-based with a morph-syntactic tagset by Khoja (2001), POS using Support Vector Machine (SVM) by Diab *et al*. (2004), hybrid method for tagging Arabic text (Tlili-Guiassa, 2006), and Arabic Part-Of-Speech Tagging Using Transformation-Based Learning (AlGahtani *et al*., 2009). Additionally, Albared *et al.* (2010) presented the smoothing algorithm with hidden markov model to solve the problem of data sparseness. In this stage, we used the joint tagging and segmenting algorithm that used for Arabic tagging by AlGahtani *et al*. (2009). The output of this phase is the Bi-gram list that contains the bi-gram candidates, the component of bi-gram candidates (the first word, second word) with their frequency, the frequency of bi-gram and POS tag for the bi-gram.

The second phase includes also the filtering of the bi-gram candidates according to the morphological and syntactic variations. To increase the statistical measures for the extracted candidates, we sum the frequency of all the forms that variety morphologically from the main candidates (the exact collocation). Of course, this process overcomes ignoring some candidates that have low frequency and low association measures.

Candidate ranking: The second step of collocation extraction is the candidate ranking. The candidate ranking relies on frequency information about word occurrence and co-occurrence in a corpus. As we have observed, the Bi-gram list also contains the information related to the candidate’s occurrence in the corpus. For the candidate pairs identified, the candidate identification step collects both syntactic information and information about their occurrence in the corpus. In this step, the association measures are computed to the identified candidates in bi-gram list that assigns to each candidate a score of association strength. For each pair of words extracted from a corpus, association score is a single real value that indicates the amount of (statistical) association between the two words. Some of association measures are based on statistical hypothesis tests and supported with mathematical proof, while others are heuristic combinations.

In our study, we selected four association measures that have strong association, according to some recent methods for collocation extraction (Evert and Krenn, 2005; Ramisch *et al*., 2008; Pecina, 2008; Zhang *et al*., 2009; Boulaknadel *et al*., 2008). The first association measure is the Log-Likelihood Ratio (LLR) that was introduced by Dunning (1993). The log-likelihood is calculated with a formula adjusted for co-occurrence contingency table as follows. For a given pair of words W_1 and W_2 and a search window W, let a be the number of windows in which W_1 and W_2 co-occur, let b be the number of windows in which only W_1 occurs, let c be the number of windows in which only W_2 occurs and let d be the number of windows in which none of them occurs. The LLR is defined as the following:

LLR = 2 ((alna + blnb + clnc + dlnd + (a+b+c+d) ln (a+b+c+d)) –((a+b) ln(a+b) +(a+c)ln(a+c) + (b+d)ln(b+d)+ (c+d) ln(c+d))) (1)

The second association measure is the chi-square. It compares between the observed and expected frequencies (Pecina 2010). It is calculated for bi-gram (x, y) as follows:

$$x^2 = \frac{(F_{x,y}\left(\frac{f_x f_y}{n}\right))^2}{\left(\frac{f_x f_y}{n}\right)} \quad (2)$$

The third association measure is the Pointwise Mutual Information. This measure has been used as an association measure to rank the candidates of collocation by Zhang *et al*. (2009). It was calculated as follows. For given two words x and y, P(x) is the occurrence probability of word x and P(y) is the occurrence probability of word y in the corpus, the formula of Mutual Information (MI) as the following:

$$MI(x,y) = \log_2 \frac{p(x,y)}{p(x)p(y)} \quad (3)$$

According to Zhang *et al*. (2009), the reason for using this measure in the candidate ranking for collocation extraction is that MI has the support from information theory and mathematical proof. The last association measure is the Enhanced Mutual Information (EMI). This association measure is used by Zhang *et al*. (2009) to cope with the problem as unsymmetrical co-occurrence. The mathematic formula of EMI is defined as:

$$EMI(x,y) = \log_2 \frac{p(x,y)}{(p(x)p(x,y))(p(y)-p(x,y))} \quad (4)$$

Evaluation method: In our method, we used the n-best evaluation method (Evert, 2005) that uses association scores to rank the collocation candidates extracted from a text corpus. This method consists of three major steps: selection the n-best list, manual annotation and computation the precision. From the bi-gram list, this method selects the sets of *n* highest-ranking candidates according to the association scores for each association measure, called *n*-best lists. In the second step, from the n-best list of each association measure, each candidate is passed on to human annotators for manual selection of the true collocations. Each candidate is marked as one of the four following tags: T: the true collocation; N: not collocation; NT: cannot decide (incomplete); Err: This expression is not a collocation (an error of the morphological disambiguation). After the manual annotation of candidates in n-best list, we computed the precision of each association measure that defines as the following:

$$\text{Precision} = \frac{TP}{TEC} \quad (5)$$

Where:

TP = The number of correct extracted collocations
TEC = The total number of extracted collocations (the n value for n-best list)

RESULTS

In our experiment, we have used the Arabic corpus. Our corpus is an electronic corpus of Modern Standard Arabic that was collected from online Arabic newspaper archives. Table 1 provides the numerical details about the Arabic corpus used in the method for collocation extraction.

Dataset: Table 2 shows the number of extracted bi-gram candidates for each structural pattern.

Table 1: Statistics on the corpus used in extraction

Statistics	Value
Size (MB)	12.300000
Files	100.000000
Words	2.325,152
Sentences	102.356000

Table 2: The number of candidate pairs in collocations

Patterns	Freq>10	Freq<10
Noun+ Noun	1284	53726
Noun + Adjective	1651	31888
Noun + Verb	286	8521
Verb + Adverb	521	6523
Adjective + Adverb	365	7852
Adjective + Noun	985	9564
Collocation	5092	150534

Table 3: The precision values for n-best collocations

AMs	N =100 (%)	200 (%)	300 (%)	400 (%)	500 (%)
MI	83	78.0	77.0	77.75	70.8
EMI	90	75.0	76.6	75.50	76.8
LLR	95	93.5	90.3	85.50	83.8
X^2	89	85.5	85.6	81.50	80.0

Precision: In order to evaluate the association measures, we have computed the precision for each n-best list. In this experiment, we selected the n-best set from data set for each association measure, with n ranging from 100-500 at intervals of 100. For each association measure, we have computed the precision for five n-best list. The Table 3 shows the precision values for n-best collocation bi-gram.

DISCUSSION

From Table 3, the log-likelihood ratio clearly outperforms the others association measures with n> = 100. As expected from the results of other studies (Evert, 2008; Boulaknadel *et al*., 2008) the precision of MI is significantly lower than that of log-likelihood ratio. Also, we can observe that the precision of all association measures (except EMI) decreases when the n increases. This means the extracted bi-gram candidates that have high association score have also high probability to be the collocation.

CONCLUSION

In this study, we have presented our method for collocation extraction from Arabic corpus. This method is a hybrid method that depends on both linguistic information and association measures. We have discussed some linguistic information for Arabic collocation that has been used for candidate identification, such as, the structural patterns of collocation, morphological and syntactic variations. We used four association measures to rank the candidates according to association scores. We applied our method for an in-house collected corpus from Arabic newspaper archives. In order to evaluate association measures, we used the n-best evaluation method that selects n-best set for each association measure and annotates the extracted candidates manually. In our experiment, the log-likelihood ratio has proved to be the best association measure that has achieved the highest precision value 83.8% in the n-best list with n=500. The hybrid extraction method that we demonstrated on bigram collocation can be straightforwardly expanded to extract trigram and n-gram Arabic collocations.

REFERENCES

Albared, M., Omar, N., Ab Aziz, M.J., and Nazri, M., 2010. Automatic part of speech tagging for Arabic: An experiment using bigram hidden markov model. Lecture Notes Comput. Sci. Springer, 6401: 361-370. DOI: 10.1007/978-3-642-16248-0_52

AlGahtani, S. and W. Black and J. Mc-Naught., 2009. Arabic part-of-speech-tagging using transformation-based learning. Proceedings of the 2nd International Conference on Arabic Language Resources and Tools, Apr. 22-23, Cairo, Egypt, The MEDAR Consortium, pp: 66-70. http://www.elda.org/medar- conference/pdf/43.pdf

Attia, M., 2006. Accommodating multiword expressions in an Arabic LFG grammar. In: Advances in Natural Language Processing, Tapio, S., F.G.S. Pyysalo and T. Pahikkala (Eds.). Springer-Verlag, Berlin, Heidelberg, pp: 87-98. DOI: 10.1007/11816508_11

Boulaknadel, S., B. Daille and D. Aboutajdine, 2008. A multi-word term extraction program for Arabic language. Proceeding of the 6th International Conference on Language Resources and Evaluation, May 28-30, Marrakech Morocco, pp: 1485-1488. http://www.lrec-conf.org/proceedings/lrec2008/pdf/378_paper.pdf

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

Diab, M., K. Hacioglu and D. Jurafsky, 2004. Automatic tagging of Arabic text: From raw text to base phrase chunks. Proceeding of the NAACL-HLT, Boston, USA., pp: 149152.

Duan, J., M. Zhang, L. Tong and F. Guo, 2009. A Hybrid Approach to Improve Bilingual Multiword Expression Extraction. Proceeding of the 13th Paci_c-Asia Conference on Knowledge Discovery and Data, Apr. 27-30, Bangkok, Thailand, Lecture Notes in Computer Science 5476 Springer 2009, pp: 541-547. DOI: 10.1007/978-3-642-01307-2_51

Dunning, T., 1993. Accurate methods for the statistics of surprise and coincidence. Comput. Linguistics, 19: 61-74.

Evert, S. and K. Brigitte, 2005. Using small random samples for the manual evaluation of statistical association measures. Comput. Speech Language Spec. Issue Multiword Exp., 19: 450-466. DOI: 10.1016/j.csl.2005.02.005

Evert, S., 2004. The statistics of word cooccurrences: word pairs and collocations. Ph.D. Thesis, University of Stuttgart. http://elib.uni-stuttgart.de/opus/volltexte/2005/2371/pdf/Evert2005phd.pdf

Evert, S., 2008. A lexicographic evaluation of German adjective-noun collocations. Proceeding of the LREC 2008 Workshop Towards a Shared Task for Multiword Expressions, May 28-30, Marrakech, pp: 3-6. http://cogsci.uni-osnabrueck.de/~severt/PUB/Evert2008_MWE_Resource.pdf

Frantzi, K., A. Sophia and M. Hideki, 2000. Automatic recognition of multi-word terms: The C-value/NC-value method. Int. J. Digital Libraries, 3: 115-130.

Khoja, S., 2001. APT: Arabic part-of-speech tagger. Proceeding of the (NAACL2001), Carnegie Mellon University, Pennsylvania.

Pecina, P., 2008. A machine learning approach to multiword expression extraction. Proceeding of the LREC Workshop Towards a Shared Task for Multiword Expressions, May 28-30, Marrakech, Morocco, pp: 54-57. http://ufal.mff.cuni.cz/~pecina/publications/mwe-2008-shared-task.pdf

Pecina, P., 2010. Lexical association measures and collocation extraction. Language Res. Evaluat., 44: 137-158. DOI: 10.1007/s10579-009-9101-4

Ramisch, C., S. Paulo, M. Idiart and A. Villavicencio, 2008. An evaluation of methods for the extraction of multiword expressions. Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions, May 28-30, Marrakech, Morocco, pp: 50-53. http://www.lrec-conf.org/proceedings/lrec2008/workshops/W20_Proceedings.pdf

Smadja, F., 1993. Retrieving collocations from text: Xtract. Comput. Linguistics, 19: 143-77.

Tlili-Guiassa, Y., 2006. Hybrid method for tagging Arabic text. J. Comput. Sci., 2: 245-248.

Van de Cruys, T. and B.V. Moiron, 2006. Lexico-Semantic Multiword Expression Extraction. In: Computational Linguistics in the Netherlands, Dirix, P. *et al*. (Eds.). University of Leuven, Leuven, Belgium, pp: 175-190. http://lotos.library.uu.nl/publish/articles/000196/bookpart.pdf.

Zhang, W., Yoshida, T., Tang, X., Ho, T.B., 2009. Improving effectiveness of mutual information for substantival multiword expression extraction. Exp. Syst. Appli., 36: 10919-10930. DOI: 10.1016/j.eswa.2009.02.026