Arabic Collocation Extraction Based on Hybrid Methods

Alaa Mamdouh Akef[1], Yingying Wang[1], Erhong Yang[1](✉)

[1] School of Information Science, Beijing Language and Culture University, Beijing 100083, China
alaa_eldin_che@hotmail.com

Abstract. Collocation Extraction plays an important role in machine translation, information retrieval, secondary language learning, etc., and has obtained significant achievements in other languages, e.g. English and Chinese. There are some studies for Arabic collocation extraction using POS annotation to extract Arabic collocation. We used a hybrid method that included POS patterns and syntactic dependency relations as linguistics information and statistical methods for extracting the collocation from Arabic corpus. The experiment results showed that using this hybrid method for extracting Arabic words can guarantee a higher precision rate, which heightens even more after dependency relations are added as linguistic rules for filtering, having achieved 85.11%. This method also achieved a higher precision rate rather than only resorting to syntactic dependency analysis as a collocation extraction method.

Keywords: Arabic collocation extraction; dependency relation; hybrid method.

1 Introduction

Studies in collocation have been advancing steadily since Firth first proposed the concept, having obtained significant achievements. Lexical collocation is widely used in lexicography, language teaching, machine translation, information extraction, disambiguation, etc. However, definitions, theoretical frameworks and research methods employed by different researchers vary widely. Based on the definitions of collocation provided by earlier studies, we summarized some of its properties, and taking this as our scope, attempted to come up with a mixed strategy combining statistical methods and linguistic rules in order to extract word collocations in accordance with the above mentioned properties.

Lexical collocation is the phenomenon of using words in accompaniment, Firth proposed the concept based on the theory of “contextual-ism”. Neo-Firthians advanced with more specific definitions for this concept. Halliday (1976: 75) defined collocation as “linear co-occurrence together with some measure of significant proximity”, while Sinclair (1991: 170) came up with a more straightforward definition, stating that “collocation is the occurrence of two or more words within a short space of each other in a text”. Theories from these Firthian schools emphasized the recurrence (co-occurrence) of collocation, but later other researchers also turned to its other properties. Beson (1990) also proposed a definition in the BBI Combinatory Dictionary of English, stating that “A collocation is an arbitrary and recurrent word combination”, while Smadja (1993) considered collocations as “recurrent combinations of words that co-occur more often than expected by chance and that correspond to arbitrary word usages”. Apart from stressing co-occurrence (recurrence), both of these definitions place importance on the “arbitrariness” of collocation. According to Beson (1990), collocation belongs to unexpected bound combination. In opposition to free combinations, collocations have at least one word for which combination with other words is subject to considerable restrictions, e.g. in Arabic, “خلف” (breast) in “خلف الناقة” (the breast of the she-camel) can only appear in collocation with “الناقة” (she-camel), while “خلف” (breast)cannot form a correct Arabic collocation with “البقرة” (cow) and “المرأة” (woman), etc.

In BBI, based on a structuralist framework, Benson (1989) divided English collocation into grammatical collocation and lexical collocation, further dividing these two into smaller categories, this emphasized that collocations are structured, with rules at the morphological, lexical, syntactic and/or semantic levels.

We took the three properties of word collocation mentioned above (recurrence, arbitrariness and structure) and used it as a foundation for the qualitative description and quantitative calculation of collocations, and designed a method for the automatic extraction of Arabic lexical collocations.

2 Related work

Researchers have employed various collocation extraction methods based on different definitions and objectives. In earlier stages, lexical collocation research was mainly carried out in a purely linguistic field, with researchers making use of exhaustive exemplification and subjective judgment to manually collect lexical collocations, for which the English collocations in the Oxford English Dictionary (OED) are a very typical example. Smadja (1993) points out that the OED’s accuracy rate doesn’t surpass 4%. With the advent of computer technology, researchers started carrying out quantitative statistical analysis based on large scale data (corpora). Choueka et al. (1983) carried out one of the first such studies, extracting more than a thousand English common collocations from texts containing around 11,000,000 tokens from the New York Times. However, they only took into account collocations’ property of recurrence, without putting much thought into its arbitrariness and structure. They also extracted only contiguous word combinations, without much regard for situations in which two words are separated, such as “make-decision”.

Church et al. (1991) defined collocation as a set of interrelated word pairs, using the information theory concept of “mutual information” to evaluate the association strength of word collocation, experimenting with an AP Corpus of about 44,000,000 tokens. From then on, statistical methods started to be commonly employed for the extraction of lexical collocations. Pecina (2005) summarized 57 formulas for the calculation of the association strength of word collocation, but this kind of methodology can only act on the surface linguistic features of texts, as it only takes into account the recurrence and arbitrariness of collocations, so that “many of the word combinations that are extracted by these methodologies cannot be considered as the true collocations" (Saif, 2011). E.g. "doctor-nurse" and "doctor-hospital" aren't collocations. Linguistic methods are also commonly used for collocation extraction, being based on linguistic information such as morphological, syntactic or semantic information to generate the collocations (Attia, 2006). This kind of method takes into account that collocations are structured, using linguistic rules to create structural restrictions for collocations, but aren't suitable for languages with high flexibility, such as Arabic.

Apart from the above, there are also hybrid methods, i.e. the combination of statistical information and linguistic knowledge, with the objective of avoiding the disadvantages of the two methods, which are not only used for extracting lexical collocations, but also for the creation of multi-word terminology (MWT) or expressions (MWE). These methods use part-of-speech tagging as linguistic rules for extracting candidates for multi-word terminology, and then calculating the C-value to ensure that the extracted candidate is a real MWT. There are plenty of studies which employ hybrid methods to extract lexical collocations or MWT from Arabic corpora (Attia, 2006; Bounhas and Slimani, 2009).

3 Experimental Design for Arabic collocation extraction

We used a hybrid method combining statistical information with linguistic rules for the extraction of collocations from an Arabic corpus based on the three properties of collocation. In the previous research studies, there were a variety of definitions of collocation, each of which can't fully cover or be recognized by every collocation extraction method. It's hard to define collocation, while the concept of collocation is very broad and thus vague. So we just gave a definition of Arabic word collocation to fit the hybrid method that we used in this paper.

3.1 Definition of Collocation

As mentioned above, there are three properties of collocation, i.e. recurrence, arbitrariness and structure. On the basis of those properties, we define word collocation as combination of two words (bigram[1]) which must fulfill the three following conditions:

a. One word is frequently used within a short space of the other word (node word) in one context.

This condition ensures that bigram satisfies the recurrence property of word collocation, which is recognized on collocation research, and is also an essential prerequisite for being collocation. Only if the two words co-occur frequently and repeatedly, they may compose a collocation. On the contrary, the combination of words that occur by accident is absolutely impossible to be a collocation (when the corpus is large

[1] It is worth mentioning that the present study is focused on word pairs, i.e. only lexical collocations containing two words are included. Situations in which the two words are separated are taken into account, but not situations with multiple words.

enough). As for how to estimate what frequency is enough to say “frequently”, it should be higher than expected frequency calculated by statistical methods.

b. One word must get the usage restrictions of the other word.

This condition ensures that bigram satisfies the arbitrariness property of word collocation, which is hard to describe accurately but is easy to distinguish by native speakers. Some statistical methods can, to some extent, measure the degree of constraint, which is calculated only by using frequency, not the pragmatic meaning of the words and the combination.

c. A structural relationship must exist between the two words.

This condition ensures that bigram satisfies the structure property of word collocation. The structural relationships mentioned here consist of three types on three levels: particular part-of-speech combinations on the lexical level; dependency relationships on the syntactic level, e.g. modified relationship between adjective and noun or between adverb and verb; semantic relationships on the semantic level, e.g. relationship between agent and patient of one act.

To sum up, collocation is defined in this paper as a recurrent bound bigram that internally exists with some structural relationships. To extract collocations according to the definition, we conducted the following hybrid method.

3.2 Method for Arabic collocation extraction

The entire process consisted of data processing, candidate collocation extraction, candidate collocation ranking and manual tagging.

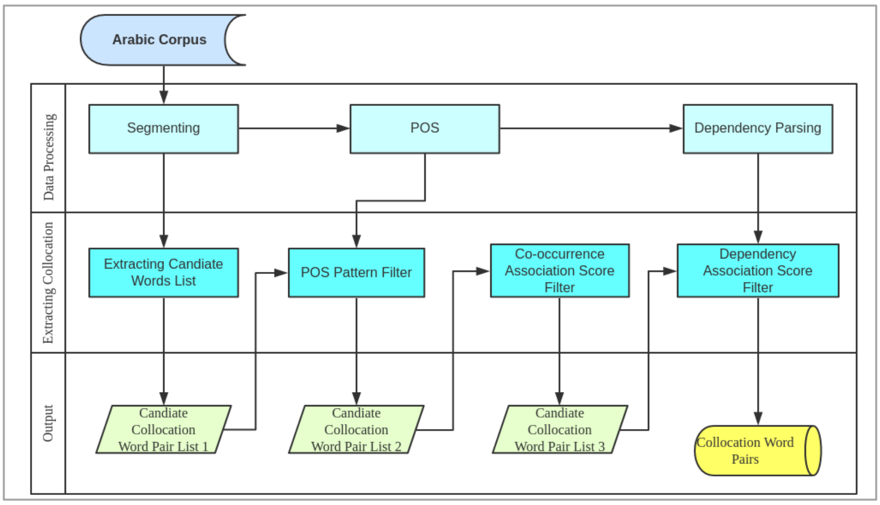
Arabic Corpus
Data Processing
Segmenting
POS
Dependency Parsing
Extracting Collocation
Extracting Candiate Words List
POS Pattern Filter
Co-occurrence Association Score Filter
Dependency Association Score Filter
Output
Candiate Collocation Word Pair List 1
Candiate Collocation Word Pair List 2
Candiate Collocation Word Pair List 3
Collocation Word Pairs

Fig. 1. Experimental flow chart

Data processing. We used the Arabic texts from the United Nations Corpus, comprised of 21,090 sentences and about 870,000 tokens. For data analysis and annotation, we used the Stanford Natural Language Processing Group’s toolkit. Data processing included word segmentation, POS tagging and syntactic dependency parsing.

Arabic is a morphologically rich language. Thus, when processing Arabic texts, the first step is word segmentation, including the removal of affixes, in order to make the data conform better to automatic tagging and analysis format, e.g. the word “يدعمها” (to support something), after segmentation “يدعم+ها”. POS tagging and syntactic dependency parsing was done with the Stanford Parser, which uses an "augmented Bies" tag set. The LDC Arabic Treebanks also uses the same tag set, but it is augmented in comparison to the LDC English Treebanks’ POS tag set, e.g. extra tags start with "DT", and appear for all parts of speech that can be preceded by the determiner "Al" (ال). Syntactic dependency relations as tagged by the Stanford Parser, are defined as grammatical binary relations held between a governor (also known as a regent or a head) and a dependent, including approximately 50 grammatical relations, such as “acomp”, “agent”, etc. However, when used for Arabic syntactic dependency parsing, it does not tag the specific types of relationship between word pairs. It only tags word pairs for dependency with “dep(w1, w2)”. We extracted 621,964 dependency relations from more than 20,000 sentences.

This process is responsible for generating, filtering and ranking candidate collocations.

Candidate collocation extracting. This step is based on the data when POS tagging has already been completed. Every word was treated as a node word and every word pair composed between them and other words in their span were extracted as collocations. Each word pair has a POS tag, such as ((w1, p1), (w2, p2)), where w1 stands for node word, p1 stands for the POS of w1 inside the current sentence, w2 stands for the word in the span of w1 inside the current sentence (not including punctuation), while p2 is the actual POS for w2. A span of 10 was used, i.e. the 5 words preceding and succeeding the node word are all candidate words for collocation. Together with node words, they constitute initial candidate collocations. In 880,000 Arabic tokens, we obtained 3,475,526 initial candidate collocations.

After constituting initial candidate collocations, taking into account that collocations are structured, we used POS patterns as linguistic rules, thus creating structural restrictions for collocations. According to Saif (2011), Arabic collocations can be classified into six POS patterns: (1) Noun + Noun; (2) Noun + Adjective; (3) Verb + Noun; (4) Verb + Adverb; (5) Adjective + Adverb; and (6) Adjective + Noun, encompassing Noun, Verb, Adjective and Adverb, in total four parts of speech. However, in the tag set every part of speech also includes tags for time and aspect, gender, number, as well as other inflections (see to table 1 for details). Afterwards, we applied the above mentioned POS patterns for filtering the initial candidate collocations, and continued treating word pairs conforming to the 6 POS patterns as candidate collocations, discarding the others. After filtering, there remained 704,077 candidate collocations.

Table 1. Arabic POS tag example.

POS	**POS tag**
Noun	DTNN, DTNNP, DTNNPS, DTNNS, NN, NNP, NNS, NOUN
Verb	VB, VBD, VBN, VBG, VBP, VN
Adjective	ADJ, JJ, JJR
Adverb	RB, RP

Candidate collocation ranking. For this step, we used statistical methods to calculate the association strength and dependency strength for collocations, and sorting the candidate collocations accordingly.

The calculation for word pair association strength relied on frequency of word occurrence and co-occurrence in the corpus, and for its representation we resorted to the score of Point Mutual Information (PMI), i.e. an improved Mutual Information calculation method, and also a statistical method recognized for reflecting the recurrent and arbitrary properties of collocations, and being widely employed in lexical collocation studies. Mutual Information is used to describe the relevance between two random variables in information theory. In language information processing, it is frequently used to measure correlation between two specific components, such as words, POS, sentences and texts. When employed for lexical collocation research, it can be used for calculating the degree of binding between word combinations. The formula is:

$$\mathrm{pmi}(\mathrm{w}_1, \mathrm{w}_2) = \log \frac{\mathrm{p}(\mathrm{w}_1,\mathrm{w}_2)}{\mathrm{p}(\mathrm{w}_1)\mathrm{p}(\mathrm{w}_2)} \tag{1}$$

$p(w_1,w_2)$ refers to the frequency of the word pair (w_1, w_2) in the corpus. $p(w_1)$, $p(w_2)$ stands for the frequency of word occurrence of w_1 and w_2. The higher the frequency of co-occurrence of w_1 and w_2, the higher $p(w_1, w_2)$, and also the higher the $pmi(w_1, w_2)$ score, showing that collocation (w_1, w_2) is more recurrent. As to arbitrariness, the higher the degree of binding for collocation (w_1, w_2), the lower the co-occurrence frequency between w1 or w2 and other words, and also the lower the value of $p(w_1)$ or $p(w_2)$. This means that when the value of $p(w_1, w_2)$ remains unaltered, the higher the $pmi(w_1, w_2)$ score, which shows that collocation (w_1, w_2) is more arbitrary.

The calculation of dependency strength between word pairs relies on the frequency of dependency relation in the corpus. The dependency relations tagged in the Stanford Parser are grammatical relations, which means that dependency relations between word pairs still belong to linguistic information, constituting thus structural restrictions for collocations. In this paper, we used dependency relation as another linguistic rule (exception of the POS patterns) to extract Arabic collocation. Furthermore, the amount of binding relations that a word pair can have is susceptible to statistical treatment, so that we can utilize the formula mentioned above to calculate the Point Mutual Information score. We used the score to measure the degree of binding between word pairs, but the $p(w_1, w_2)$ in the formula refers to the frequency of dependency relation of (w_1, w_2) in the corpus, whilst $p(w_1)$,$p(w_2)$ still stand for the frequency of word occurrence of w_1 and w_2. The higher the dependency relation of w_1 and w_2, the higher the value of $p(w_1, w_2)$, and also the higher the $pmi(w_1, w_2)$ score, meaning that collocation (w_1, w_2) is more structured.

This step can be further divided into two stages. First we calculated the association score (as) for all collocation candidates and sorted them from the highest to the lowest score. And then we traverse all $((w_1, p_1), (w_2, p_2))$ collocate candidates, and if (w_1, w_2) possessed a dependency relation in the corpus, then we proceeded to calculate their dependency score (ds), so that every word pair and the two scores composed a quadruple $AC((w_1, p_1), (w_2, p_2), as, ds)$. If (w_1, w_2) do not have a dependency relation in the corpus, then ds is null. After calculating the strength of dependency for all 621,964 word pairs and sorting them from the highest to the lowest score, the word pairs and strength of dependency constitute a tripe $DC(w_1, w_2, ds)$.

Manual Tagging. In order to evaluate the performance of the collocation extraction method suggested in the present paper, we extracted all collocation candidates for the Arabic word "تنفيذ execute" (AC quadruples where all w_1 is "تنفيذ" or its variants[1]) and all dependency collocations (DC triples where all w_1 is "تنفيذ" or its variants), obtaining a total of 848 AC quadruples and 689 DC triples. However, only word pairs in 312 of the AC quadruples appear in these 689 DC triples. This happens because the span set in the methods for collocation candidates in quadruples is 10, while analysis of the scope of syntactical dependency analysis comprises the whole sentence. Thus, words outside of the span are not among the collocation candidates, but might have a dependency relation with node words. Afterwards, each word pair in AC quadruples and DC triples were passed on to a human annotator for manual tagging and true or false collocation.

4 Results and Analysis

The tables 2&3 below present the proportional distribution of the results from the collocation candidates for "تنفيذ", as well as their precision rate. "True collocations" refer to correct collocations selected manually, while "false collocations" refer to collocation errors filtered manually. "With Dependency relation" indicates that there exists one kind of dependency relation between word pairs, while "Without Dependency relation" indicates word pairs without dependency relation. So "With Dependency relation" indicates collocations selected by the hybrid method presented in this paper, "true collocation" and "With Dependency relation" stand for correct collocations selected using hybrid methods. As to precision rate, "Precision with dependency relation" in table 2 represents the precision rate of the hybrid method which comprises POS patterns, statistical calculation and dependency relations. "Precision without dependency relation" represents the precision rate using POS patterns and statistical calculation, without dependency relations. "Precision with dependency relation only"

[1] One Arabic word could have more than one from in corpus because Arabic morphology is rich, so "تنفيذ" has 55 different variants.

in table 3 represents the precision rate of the method only using dependency relations[1].

Table 2. The numerical details about extracted collocations using the hybrid method.

			as>0	as>1	as>2	as>3	as>4	as>5	as>6
	Percent of candidate collocation	100	78.89	77.36	45.28	30.90	17.57	9.79	6.49
Percent of true collocations	**With Dependency relation**	23.11	19.93	19.69	12.85	8.25	4.72	2.00	1.53
Percent of false collocations		13.68	7.31	6.72	2.71	1.77	0.83	0.71	0.35
Percent of rue collocations	**Without Dependency relation**	17.10	15.92	15.80	11.32	8.14	4.25	2.83	1.53
Percent of false collocations		46.11	35.97	35.14	18.40	12.74	7.78	4.25	3.07
	Precision with dependency relation	62.82	73.16	74.55	82.58	82.35	85.11	73.91	81.25
	Precision without dependency relation	40.21	45.44	45.88	53.39	53.05	51.01	49.40	47.27

Table 3. The numerical details about extracted collocations using dependency relation.

		ds>0	ds>1	ds>2	ds>3	ds>4	ds>5	ds>6	ds>7
Percent of candidate collocation	100.0	92.29	83.00	70.71	56.57	40.43	29.29	17.86	11.29
Percent of true collocations	38.14	37.57	36.00	31.86	25.71	18.57	13.00	7.29	4.71

[1] Bigrams sorted by their dependency score (ds), which actually is the Point Mutual Information Score.

Percent of false collocations	61.86	54.71	47.00	38.86	30.86	21.86	16.29	10.57	6.57
Precision with dependency relation only	38.14	40.71	43.37	45.05	45.45	45.94	44.39	40.80	41.77

From the tables above, we can see that the precision of the hybrid method has been significantly improved compared to the precision of method without dependency relation and with dependency relation only. More concretely, we can find that in the set of candidate collocations (bigrams) extracted and filtered by POS patterns, PMI score and dependency relations, true collocations have much higher proportions than false collocations. But the result is completely opposite in the set of candidate collocations (bigrams) extracted without dependency relations, i.e. false collocations have much higher proportions than true collocations. This data illustrates that the potential is very great for one word collocation internally exists with some kind of dependency relation, but not all collocations do. Thus the results are enough to illustrate that it is reasonable to use dependency relation as a linguistic rule to restrict collocation extraction. However, when we only use dependency relation as a linguistic rule to extract collocations, just as the data showed in table 3, false collocations also have much higher proportions than true collocations. This data illustrates that dependency relation is not sufficient enough, and that POS patterns are also necessary to restrict collocation extraction.

There is an example that illustrates the effect of dependency relation as a linguistic rule to filter the candidate collocations. The bigram (تنفيذي, رتبة) has a very high frequency in the Arabic corpus, ranking second, “رتبة” meaning “the level of”. And when the two words co-occur in one sentence of the corpus, which mostly means "the level of the executive (organ or institution)", there is no dependency relation between the two words, so the bigram is filtered out. There are so many situations like this bigram that can be successfully filtered out, which can significantly improve the precision rate of the hybrid method of collocation extraction.

As mentioned above, not all collocations have an internal dependency relation and not all bigrams that have internal dependency relation are true collocations. Such as bigram (تنفيذ, قرر), which means “decide to implement”, “تنفيذ” (implementation) is the object of “قرر” (decide), there is a dependency relation between the word pair. But we can annotate the bigram as a “false collocation” without hesitation. These kinds of bigrams result in the error rate of the hybrid method. Beyond this, another reason for error rate can be the incorrect dependency result analyzed by the Stanford Parser. The hybrid method of this paper only uses dependency relation as one linguistic rule, without being entirely dependent on it, so the precision of the hybrid method is much higher than the method only using dependency relations.

To sum it all up, the hybrid method presented in this paper can significantly improve the precision of collocation extraction.

5 Conclusion

In this study, we have presented our method for collocation extraction from an Arabic corpus. This is a hybrid method that depends on both linguistic information and association measures. Linguistic information is comprised of two rules: POS patterns and dependency relations. Taking the Arabic word "تنفيذ" as an example, by using this method we were able to extract all the collocation candidates and collocation dependencies, as well as calculating its precision after manual tagging. This experiment's results show that by using this hybrid method for extracting Arabic words, it can guarantee a higher precision rate, which heightens even more after dependency relations are added as rules for filtering, achieving 85.11% accuracy, higher than by only resorting to syntactic dependency analysis as a collocation extraction method.

References

1. Attia, M. A.: Accommodating Multiword Expressions in an Arabic LFG Grammar. Advances in Natural Language Processing, International Conference on NLP 2006, vol. 4139, pp. 87-98. DBLP, Fintal (2006).
2. Benson, M.: Collocations and general-purpose dictionaries. International Journal of Lexicography, 3(1), 23-34 (1990).
3. Bounhas, I., Slimani, Y.: A hybrid approach for Arabic multi-word term extraction. International Conference on Natural Language Processing and Knowledge Engineering 2009, Nlp-Ke vol. 30, pp. 1-8. IEEE (2009).
4. Church, K. W., Hanks, P., Hindle, D.: Using Statistics in Lexical Analysis. Lexical Acquisition (1991).
5. Frantzi, K., Sophia A., Hideki, M.: Automatic recognition of multi-word terms: The C-value/NC-value method. Int. J. Digital Libraries 3, 115-130 (2000).
6. Halliday, M. A. K.: Lexical relations. System and Function in Language. Oxford University Press, Oxford (1976).
7. Pecina. An Extensive Empirical Study of Collocation Extraction Methods. ACL 2005, Meeting of the Association for Computational Linguistics, pp. 13-18, University of Michigan, USA (2005).
8. Saif, A. M., Aziz, M. J. A.: An automatic collocation extraction from Arabic corpus. Journal of Computer Science 7(1), 6 (2011).
9. Sinclair, J.: Corpus, concordance, collocation. Oxford University Press, Oxford (1991).
10. Smadja, F.: Retrieving collocations from text: extract. Computational Linguistics, 19(19), 143-177 (1993).