An Empirical Study of Arabic Formulaic Sequence Extraction Methods

Ayman Alghamdi, Eric Atwell, Claire Brierley
School of Computing, University of Leeds
Woodhouse Lane Leeds, LS2 9JT UK.
scaaa@leeds.ac.uk, e.s.atwell@leeds.ac.uk, c.brierley@leeds.ac.uk

Abstract

This paper aims to implement what is referred to as the collocation of the Arabic keywords approach for extracting formulaic sequences (FSs) in the form of high frequency but semantically regular formulas that are not restricted to any syntactic construction or semantic domain. The study applies several distributional semantic models in order to automatically extract relevant FSs related to Arabic keywords. The data sets used in this experiment are rendered from a new developed corpus-based Arabic wordlist consisting of 5,189 lexical items which represent a variety of modern standard Arabic (MSA) genres and regions, the new wordlist being based on an overlapping frequency based on a comprehensive comparison of four large Arabic corpora with a total size of over 8 billion running words. Empirical n-best precision evaluation methods are used to determine the best association measures (AMs) for extracting high frequency and meaningful FSs. The gold standard reference FSs list was developed in previous studies and manually evaluated against well-established quantitative and qualitative criteria. The results demonstrate that the MI.log_f AM achieved the highest results in extracting significant FSs from the large MSA corpus, while the T-score association measure achieved the worst results.

Keywords: Arabic, Formulaic Sequence, Association Measures

1. Introduction

In recent decades, the phenomenon of formulaic language has witnessed a proliferation from various perspectives within the research community (e.g. linguistic, psycholinguistic natural language processing 'NLP' and language pedagogy 'LP'). Many studies that have been conducted in these areas show the key role that formulaic language plays in our use of everyday languages. Even though most modern languages have benefitted from a large amount of research in this area, the multidisciplinary and heterogeneous nature of this complex linguistic phenomenon requires more researchers' attention using various methodologies borrowed from a range of different related scientific perspectives. This research will ultimately contribute to the improvement of our understanding of the linguistic behaviour of FSs and its implications for language applications such as lexicography, information retrieval, machine translation, and word sense disambiguation.

Research into Arabic FSs is still underdeveloped, particularly research that makes use of MSA corpus-based analysis techniques which enable researchers to build their linguistic assumptions on real used language. In the field of Arabic LP and NLP, there is an urgent need for developing new corpus-based FSs language resources which can be used in various related applications. Therefore, the current study aims to remedy this deficiency by introducing a new corpus-based list of FSs based on an empirical evaluation of Statistical Association Measures (AMs). However, it is worth mentioning here that this study is part of a larger ongoing research project that aims to build an intensive Arabic FSs lexicon for use in LP and NLP.

Extracting the most common and meaningful FSs associated with a frequency based Arabic wordlist - our primary concern in this study - can be seen as the basis for a useful language resource that can be used in various language related applications. The current study uses high frequency and significant AMs scores as reliable predictors of useful FSs' list. Several studies have found a strong link between the high frequency of sequences and holistic processing. For instance, in using an eye-tracking paradigm, Underwood et al. (2004) found an advantage in terms of FSs processing by native speakers. Another study by Durrant (2008) found a significant relationship between the high frequency of occurrence and the mental representation of lexical items in serious lexical decision experiments conducted on adult second language learners. Since the linguistic units we aim to extract in this study are not restricted to any syntactic construction or semantic domain, we use the term FSs based on Schmitt (2010)'s suggestion of using this term as an umbrella one to refer to various types of linguistic units in general. Thus, the current study adopts a practical definition of Arabic FSs which basically concentrates on any type of syntactic construction from different language domains that makes high frequency use of semantically regular phrases.

2. Related work

In the literature there are three main approaches for collocation extraction – data-driven, knowledge-based and hybrid methods. These approaches have been applied in many experimental studies in different languages and different experimental settings. Studies by Smadja (1993), Church and Hanks (1990) and Sinclair (1991) represent the use of data-driven statistical approaches as the main feature in the process of collocation extraction. For instance, Sinclair defined collocation as "…lexical co-occurrence, more or less independently of grammatical pattern or positional relationship" (ibid. p.170), while knowledge-based or linguistic models of collocation extractions emphasise the role of a syntactic relationship between the lexical items in the collocations. Examples of using such an approach can be seen in the work of Choueka (1988), Mel'cuk (1998; 2003) and Bartsch (2004, 76) who defines collocations as:

"…lexically and-or pragmatically constrained recurrent co-occurrences of at least two lexical items which are in a direct syntactic relation with each other".

The third approach used a combination of statistical and linguistic methods in different types of collocation extraction models. In Arabic, several studies have attempted to automatically or semi-automatically extract lists of collocations based on different experimental settings and language domains. For instance, Boulaknadel et al. (2008) developed a programme for multi-word extractions based on linguistic analysis and the evaluation of statistical scores, in which a list of Arabic terms from the environmental domain was used as the gold standard list in the evaluation of four AMs, LLR, T-score, FLR and Mutual Information. The experiment was conducted on an environmental corpus, the extracted terms tested against the reference list, and the result shows that the Log-Likelihood Ratio, and the FLR and t-score measures outperform the MI measure. In another study by Saif and Aziz (2011) using a hybrid method for extracting the collocations from an Arabic corpus that is based on linguistic information and AMs, the evaluation of this study used the n-best method to annotate the extracted collocations. The results show that the Log-Likelihood Ratio is the best association measure in the process of predicting the correct Arabic collocates. In a recent study, Alrabiah et al. (2014) aimed to automatically identify lexical collocations in the Quran and in a large classical Arabic corpus. Eight AMs were used in the evaluation process, and the results demonstrate that the MI.log_f AM achieved the best results in extracting significant Arabic collocations from the Classical Arabic corpora, while mutual information AM achieved the worst results. Since our study is different from the previous study in terms of the targeted lexical items and the used data sets, it is interesting to see the potential findings and compare them with previous research. Therefore, the current study aims to seek answers to the following question, the following question: what are the best AMs that can be used as reliable predictors in extracting semantically regular Arabic formulas?

3. Evaluation of AMs in FSs extraction

This is a preliminary study to explore a range of well-known AMs in the process of extracting meaningful and high frequency Arabic FSs from large MSA corpus, the main objective of this evaluation experiment is to find out the best reliable AM which can be used as a predictor for the right collocates of the lexical items driven from a corpus-based Arabic wordlist.

3.1 Experiment setting

The study uses association scores to rank the FSs candidates extracted from a large corpus and precision scores computed for sets of n-highest-ranking. Thus, the first step in this experiment is to prepare a gold standard list of FSs. However, we adopted an FSs list from a previous study conducted by the researchers which was developed through different processing phases and manually evaluated against well-established quantitative and qualitative criteria (Alghamdi et al., 2015). The Sketch Engine (Kilgarriff et al., 2014) was used in this study to compute six types of well-known AMs which include t-score, mutual information (MI), MI3, logDice, MI.log_f and log-likelihood. Table 1 shows the equations of these AMs along with their references. In his explanation of the number 14 in the logDice AM Rychlý (2008, 9) states that 'theoretical maximum is 14, in case when all occurrences of X co-occur with Y and all occurrences of Y co-occur with X. Usually the value is less than 10'

AMs	Ref	Formula
T-score	(Church et al., 1991)	$\frac{f_{AB-} \frac{f_A f_B}{N}}{\sqrt{f_{AB}}}$
mutual information (MI)	(Daille, 1994)	$log_2 \frac{f_{AB} N}{f_A f_B}$
MI3	(Daille, 1994)	$log_2 \frac{\int_{AB}^{3} N}{f_A f_B}$
MI.log_f	(Rychlý, 2008)	$MI - score \times log_{xy}$
logDice	(Rychlý, 2008)	$logDice = 14 + log_2 D$ $= 14 + log_2 \frac{2fxy}{fx + fy}$
Log-likelihood	(Dunning , 1993)	$-2 \log \frac{L(O_{11,} C_{1,r}).L(O_{12,} C_{2,r})}{L(O_{11,} C_{1,r}).L(O_{12,} C_{2,r_2})}$ $L(k,n,r) = r^k(1-r)^{n-k}$ $r = \frac{R_1}{N}, r_1 = \frac{O_{11}}{C_1}, r_2 = \frac{O_{12}}{C_2}$

Table 1: Algorithms used to measure the association strength of the word pairs

3.2 Data Sets

Two datasets, a sample of 50 high-frequency words and a sample of 50 low frequency words, were selected for this experiment. The words in these data sets were extracted from a newly developed corpus-based wordlist of the most frequent MSA words, based on the overlapping frequency and dispersion in a comprehensive comparison of four large MSA corpora of the total size over more than 8 billion running words, with the final wordlist consisting of more than 5 thousand items.

The lexical units adopted in this wordlist was based on the word lemma which involve all the word forms with the same lemma and its inflectional variants. By the overlapping frequency we mean the sum of the average reduced frequency (ARF) of each lemma in four large corpora which take into account the frequency and the distribution of a lexical unit in the corpora. The final wordlist was based on the highest frequency words in the four corpora. However, more details about the methodology and the full new Arabic wordlist will be published soon in another paper by the researchers. The data sets used in this experiment were randomly selected based on their ARF frequency score in the final version of the new Arabic wordlist.

The new list was automatically lemmatized and

morphologically analysed using the MadaAmira toolkit. Figure 1 shows the distributions of word classes in the new Arabic wordlist.

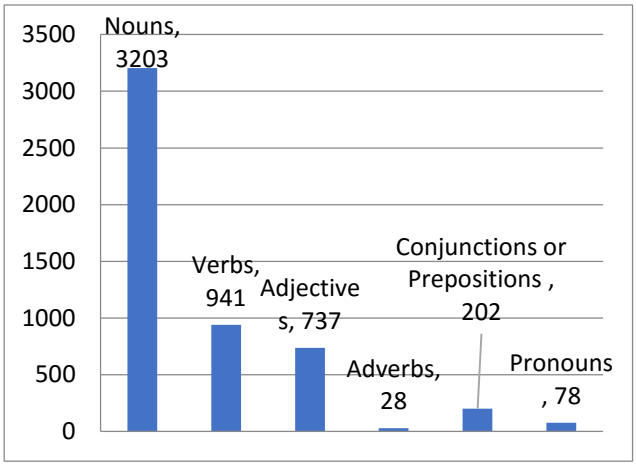

Figure 1: Distribution of word classes in the new corpus-based Arabic wordlist

Each word in the data set has an equivalent FS from a previously developed gold standard FSs list. The reason for dividing the data set into high and low frequency samples is to measures the node word frequency effect on the AMs performance. Tables 2 and 3 show the five highest and the five lowest node words used in this experiment, along with their overlapping ARF frequencies.

Words	POS	ARF Frequency
من man[1] 'from'	prep	184063923
على ʿalā 'on'	prep	94092928
هٰذا hāḏā 'this'	pron	39857940
خاصة ḵāṣa'Private'	verb	11090802
يوم yawm 'day'	noun	6320491

Table 2: The five highest frequency node words

Words	POS	ARF Frequency
التنافس *attanāfs* 'Competition '	noun	124990
قاسية *qāsya* 'Severe'	noun	108866
مدرج *madrj* 'Runway '	noun	91740
يستلزم *yastlzm* 'Require'	verb	86400
حصانة *ḥaṣāna* 'Immunity'	noun	56326

Table 3: the five lowest frequency node words

3.2 Performing the experiment

The study was conducted in two rounds using the high and low frequency data sets using the same procedures in the following steps. First, a threshold with a minimum frequency of 10 per million was selected within a search window of two to four words, and then the six AMs were computed for each node word. The highest identified collocates were recorded and ranked based on different AMs, with the precision of each node word being calculated as shown in the equation

$$precision = \frac{attested\ FSs}{all\ extracted\ FSs}$$

After that, the average precision (AP) for each AM was calculated for each node word, and finally the mean average precision (MAP) for each AM was calculated for all node words. The experiment was performed on the ArTenTen MSA corpus (Belinkov et al 2013) which consists of more than 7.4 billion running words.

4. Results and discussion

Figure 2 shows the MAP scores for each association measure using the high frequency data set in the first round of this experiment. It can be seen that the MI.log_f and MI measures achieved the highest MAP scores with a MAP score of over 0.85, while the t-score and MI3 were the least useful scores in terms of identifying FSs among the high frequency lexical items, with MAP scores below 0.50. Overall, it can be seen that three AMs used with this data set (MI.log_f , MI and logDice) achieved the highest MAP scores, while the other three MAP scores (T-score, MI3 and log-likelihood) achieved the lowest MAP scores. This result coincides with that of Alrabiah et al. (2014) who found that the MI.log_f score outperformed other AMs in predicting the lexical collocations in small and large classic Arabic corpora. However, other studies on Arabic collocations have found that the log-likelihood was the best AM in terms of extracting lexical collocations (e.g. Boulaknadel et al. (2008); Saif and Aziz (2011). However, it is worth mentioning here that these studies did not use the MI.log_f in their evaluation of AMs, which might explain the variations in terms of determining the best AMs in the current experiment.

In the second round of the experiment, dealing with the least frequent lexical items used as the node words in FSs extraction, the MAP scores in Figure 2 with the error bars show an overall drop in the performance of most AMs. This is due to the fact that most AMs usually work better with high frequency data. In addition, it is apparent that MI.log_f and the logDice score outperformed other AMs, with a MAP score of over 0.75. This suggests that they are the best AM predictor when it comes to extracting the collocation of less frequent node words.
Figure 2 also offers a comparison between the findings of the two rounds of the experiment. A slight drop can be noted in the performance of all AMs as can a change in the ranking of the best AMs, in that the MI achieved the second best AMs when using less common node words. The t-score is still the least accurate AMs in terms of predicting FSs, regardless of the level of frequency of the node words.

[1] The writer used the German standard DIN 31636 for rendering Romanized Arabic as described in Appendix 1

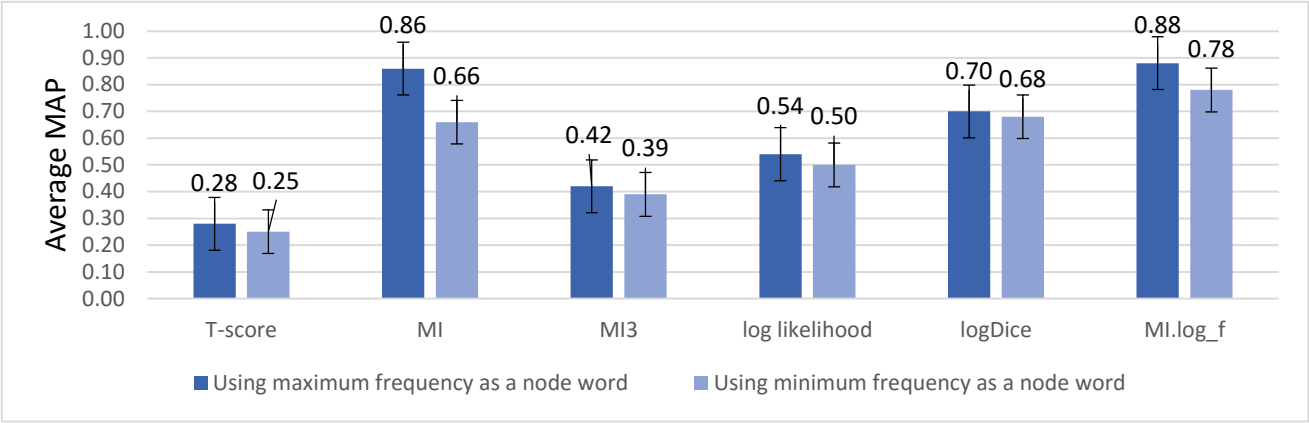

Figure 2: MAP scores of AMs using the tow data sets with the error bars

Figure 3 summarises the results of the AMs evaluation of the two data sets by calculating the average MAP scores for both data sets. It can be seen that the MI.log_f and MI scores ranked as the best AMs for predicting the right collocates of the Arabic keyword list. However, this result is in line with Alrabiah et al. (2014) and also another extensive empirical evaluation of 87 AMs in the automatic extraction of Czech collocations by Pecina (2005), who found the Pointwise MI measures achieved the best result with a 73.0% precision score.

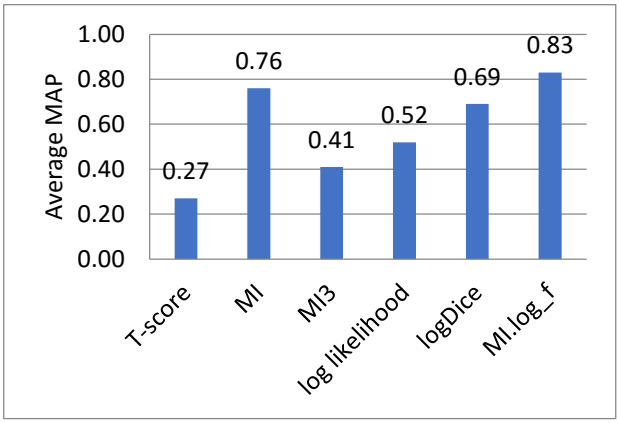

Figure 3: The average MAP scores for both data sets

Table 4 shows an example of the extracted FSs. It can be seen that these bigrams represent various syntactic constructions and semantic fields, as our study was not restricted to syntactic structures or the semantic domain.

FSs	Structures
من أجل man ʾajl *'In order to'*	Prep-Noun
اعتماداً على aʿtmādā ʿalā *'Based on'*	Noun-Prep
التنافس المحموم attanāfs almaḥmūm *'Frenzied competition'*	Noun-Adj
مدرج المطار madrj almaṭār *'Airport Runway'*	Noun-Noun
ظروف قاسية ḍarūf qāsya *'Severe conditions'*	Noun-Adj

Table 4: Examples of extracted FSs with their syntactic structures

5. Conclusion and extension

In this paper we present a brief report on an empirical study that aims to evaluate the best AMs in the process of extracting MSA FSs. This is part of a series of experiments that used a statistical and symbolic approach to extract various types of semantically regular and high frequency FSs, in order to build intensive Arabic FSs language resources for use in LP and NLP. The evaluation of AMs in this study shows a superior predictive result with regard to AMs when using high frequency data. The MI.log_f ,MI and logDice achieved the highest precision scores with regard to FSs extraction from large MSA corpora. Thus these AMs are the best candidates when it comes to predicting useful and meaningful FSs related to frequency based Arabic wordlist. On the other hand, the MAP scores finding illustrates that T-score and MI3 are the worst AMs candidates in predicting a useful FSs, while the Log-likelihood can be seen as an interesting candidate in extracting meaningful FSs. In future work, further experiments will be conducted on the evaluation of other AMs based on larger data sets to extract different types of FSs from a variety of MSA corpora. Our future work also will consider the evaluation of the best AMs with different types of Arabic data sets to examine all the possible factors that might has an impact on the use of various AMs. Durrant (2008) states that knowing two-word collocations is only the first phase in the process of extracting meaningful and useful phrasal items. Therefore, subsequent work will concentrate on extending the current list of bigrams to long FSs which will reflect on the actual use of formulaic language for our different communicative language needs.

6. Bibliographical References

Alghamdi, A., Atwell, E. and Brierley, C. (2015) Constructing a corpus-informed Listing of Arabic formulaic sequences ArFSs for language pedagogy and technology Unpublished paper.

Alrabiah, M., Alhelewh, N., Al-Salman, A. and Atwell, E. (2014) An empirical study on the Holy Quran based on a large classical Arabic corpus. International Journal of Computational Linguistics (IJCL). 5(1), pp.1-13.

Boulaknadel, S., Daille, B. and Aboutajdine, D. (2008) A Multi-Word Term Extraction Program for Arabic Language, the 6th international Conference on Language Resources and Evaluation LREC 2008, Marrakech, Morocco, pp. 1485-1488, 2008.

Church, K., Gale, W., Hanks, P. and Kindle, D. (1991) Using statistics in lexical analysis. In Lexical Acquisition: Using On-line Resources to Builda Lexicon, pages 115–164. Lawrence Erlbaum.

Church, K.W. and Hanks, P. (1990) Word association norms, mutual information, and lexicography. Computational linguistics. 16(1), pp.22-29.

Church, Kenneth W. (2000) Empirical estimates of adaptation: The chance of two Noriegas is closer to p=2 than p2. In Proceedings of COLING 2000, pages 173–179, Saarbrücken, Germany.

Daille, B. (1994) Approche mixte pour l'extraction automatique de terminologie: statistiques lexicales et filtres linguistiques. Ph.D. thesis, Université Paris 7.

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

Durrant, P.L. (2008) High frequency collocations and second language learning. Ph.D. thesis University of Nottingham.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. and Suchomel, V. (2014) The Sketch Engine: ten years on. Lexicography. 1(1), pp.7-36.

Mel'ˇcuk, I. (1998) Collocations and lexical functions. In: Cowie AP (ed) Phraseology. Theory,Analysis, and Applications, Claredon Press, Oxford, pp 23–53.

Mel'cuk, I. (2003) Collocations: définition, rôle et utilité. In: Grossmann F, Tutin A (eds) Les collocations: analyse et traitement, Editions De Werelt, Amsterdam, pp 23–32.

Pecina, P. (2005) An extensive empirical study of collocation extraction methods. In: Proceedings of the ACL Student Research Workshop: Association for Computational Linguistics, pp.13-18.

Rychlý, P. (2008) A lexicographer-friendly association score. Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN. pp.6-9.

Saif, A.M. and Aziz, M.J. (2011). An automatic collocation extraction from Arabic corpus. Journal of Computer Science. 7(1), pp.6-11.

Schmitt, N. (2010) Researching vocabulary: a vocabulary research manual. Basingstoke: Palgrave Macmillan.

Sinclair, J. (1991) Corpus, concordance, collocation. Oxford University Press. Oxford.

Smadja, F. (1993) Retrieving collocations from text: Xtract. Computational linguistics. 19(1), pp.143-177.

Yonatan Belinkov, Nizar Habash, Adam Kilgarriff, Noam Ordan, Ryan Roth, and Vit Suchomel. (2013). arTenTen: a new, vast corpus for Arabic. In Eric Atwell and Andrew Hardie, editors, Proceedings of the Second Workshop on Arabic Corpus Linguistics (WACL2) http://www.comp.leeds.ac.uk/eric/wacl/wacl2proceedings.pdf

Appendix1: The German standard DIN 31636 for rendering Romanized Arabic

No	Original Arabic letter	DIN 31635
1	أ	ʾ
2	ب	**b**
3	ت	**t**
4	ث	**ṯ**
5	ج	**ǧ**
6	ح	**ḥ**
7	خ	**ḫ**
8	د	**d**
9	ذ	**d**
10	ر	**r**
11	ز	**z**
12	س	**s**
13	ش	**š**
14	ص	**ṣ**
15	ض	**ḍ**
16	ط	**ṭ**
17	ظ	**ẓ**
18	ع	ʿ
19	غ	**ġ**
20	ف	**f**
21	ق	**q**
22	ك	**k**
23	ل	**l**
24	م	**m**
25	ن	**n**
26	ه	**h**
27	و	**w**
28	ي	**y**
29	◌َ **(short vowel)**	**a**
30	◌ُ **(short vowel)**	**u**
31	◌ِ **(short vowel)**	**i**
32	ا **(long vowel)**	**ā**
33	و **(long vowel)**	**ū**
34	ي **(long vowel)**	**ī**