Volume 5 ▪ Issue 1 ▪ April 2014

INTERNATIONAL JOURNAL OF COMPUTATIONAL LINGUISTICS (IJCL)

Publication Frequency: 6 Issues / Year

ISSN : 2180-1266

CSC PUBLISHERS

http://www.cscjournals.org

INTERNATIONAL JOURNAL OF COMPUTATIONAL LINGUISTICS (IJCL)

VOLUME 5, ISSUE 1, 2014

EDITED BY
DR. NABEEL TAHIR

ISSN (Online): 2180 - 1266

International Journal of Computational Linguistics (IJCL) is published both in traditional paper form and in Internet. This journal is published at the website http://www.cscjournals.org, maintained by Computer Science Journals (CSC Journals), Malaysia.

IJCL Journal is a part of CSC Publishers

Computer Science Journals

http://www.cscjournals.org

INTERNATIONAL JOURNAL OF COMPUTATIONAL LINGUISTICS (IJCL)

Book: Volume 5, Issue 1, April 2014
Publishing Date: 30-04-2014
ISSN (Online): 2180-1266

IJCL Journal is a part of CSC Publishers
http://www.cscjournals.org

Published in Malaysia

Typesetting: Camera-ready by author, data conversation by CSC Publishing Services – CSC Journals, Malaysia

CSC Publishers, 2014

EDITORIAL PREFACE

The International Journal of Computational Linguistics (IJCL) is an effective medium for interchange of high quality theoretical and applied research in Computational Linguistics from theoretical research to application development. This is the *first* Issue of *Fifth* Volume of IJCL. The Journal is published bi-monthly, with papers being peer reviewed to high international standards. International Journal of Computational Linguistics (IJCL) publish papers that describe state of the art techniques, scientific research studies and results in computational linguistics in general but on theoretical linguistics, psycholinguistics, natural language processing, grammatical inference, machine learning and cognitive science computational models of linguistic theorizing: standard and enriched context free models, principles and parameters models, optimality theory and researchers working within the minimalist program, and other approaches.

IJCL give an opportunity to scientists, researchers, and vendors from different disciplines of Artificial Intelligence to share the ideas, identify problems, investigate relevant issues, share common interests, explore new approaches, and initiate possible collaborative research and system development. This journal is helpful for the researchers and R&D engineers, scientists all those persons who are involve in Computational Linguistics.

Highly professional scholars give their efforts, valuable time, expertise and motivation to IJCL as Editorial board members. All submissions are evaluated by the International Editorial Board. The International Editorial Board ensures that significant developments in image processing from around the world are reflected in the IJCL publications.

IJCL editors understand that how much it is important for authors and researchers to have their work published with a minimum delay after submission of their papers. They also strongly believe that the direct communication between the editors and authors are important for the welfare, quality and wellbeing of the Journal and its readers. Therefore, all activities from paper submission to paper publication are controlled through electronic systems that include electronic submission, editorial panel and review system that ensures rapid decision with least delays in the publication processes.

To build its international reputation, we are disseminating the publication information through Google Books, Google Scholar, Directory of Open Access Journals (DOAJ), Open J Gate, ScientificCommons, Scribd, CiteSeerX Docstoc and many more. Our International Editors are working on establishing ISI listing and a good impact factor for IJCL. We would like to remind you that the success of our journal depends directly on the number of quality articles submitted for review. Accordingly, we would like to request your participation by submitting quality manuscripts for review and encouraging your colleagues to submit quality manuscripts for review. One of the great benefits we can provide to our prospective authors is the mentoring nature of our review process. IJCL provides authors with high quality, helpful reviews that are shaped to assist authors in improving their manuscripts.

Editorial Board Members
International Journal of Computational Linguistics (IJCL)

EDITORIAL BOARD

TABLE OF CONTENTS

Volume 5, Issue 1, April 2014

Pages

An Empirical Study On The Holy Quran Based On A Large Classical Arabic Corpus

Maha Alrabia *msrabiah@gmail.com*
Department of Computer Science
King Saud University
Riyadh, Saudi Arabia

Nawal Alhelewh *drnawalh@gmail.com*
Department of Arabic
Princess Nora bint Abdul Rahman University
Riyadh, Saudi Arabia

AbdulMalik Al-Salman *salman@ksu.edu.sa*
Department of Computer Science
King Saud University
Riyadh, Saudi Arabia

Eric Atwell *e.s.atwell@leeds.ac.uk*
Faculty of Engineering
Leeds University
Leeds, United Kingdom

Abstract

Distributional semantics is one of the empirical approaches to natural language processing and acquisition, which is mainly concerned by modeling word meaning using words distribution statistics gathered from huge corpora. Many distributional semantic models are available in the literature, but none of them have been applied so far to the Quran nor to Classical Arabic in general. This paper reports the construction of a very large corpus of Classical Arabic that will be used as a base to study distributional lexical semantics of the Quran and Classical Arabic. It also reports the results of two empirical studies; the first is applying a number of probabilistic distributional semantic models to automatically identify lexical collocations in the Quran and the other is applying those same models on the Classical Arabic corpus in an attempt to test their ability of capturing lexical collocations and co occurrences for a number of the corpus words. Results show that the MI.log_freq association measure achieved the highest results in extracting significant co-occurrences and collocations from small and large Classical Arabic corpora, while mutual information association measure achieved the worst results.

Keywords: Distributional Lexical Semantics, Quran, Classical Arabic Corpus, Collocation Extraction, Association Measures.

1. INTRODUCTION

The Quran is the divine religious book of Islam. Muslims believe that it contains the actual words of *Allah* (The God), and that it is the last scripture from Him to humankind. It was revealed to Prophet Muhammad 14 centuries ago in pure Classical Arabic language consisting of 77,430 words [1], which are organized in 114 chapters (*Surahs*) that are in turn partitioned into a number of verses (*Ayahs*). The Quranic text is unique and has not encountered any change throughout the previous decades. It is also considered the main source of legislation in Islam since it contains the main rules that govern Muslims in various spiritual, private, social and political aspects of life.

The Quranic text is rich in its vocabulary, morphology and syntactic structures. Its language is unique in its eloquence and style, which differentiate it from any other Classical Arabic text. Some words of the Quran are intended to produce several contextual meanings within the same verses they appear in [2][1]. Other words have different meanings depending on the context in which they appear in. For example, the word *fitnah* (disorder) is mentioned in the Quran with eleven different meanings in different contexts[2]. In addition, the Quranic text is concise; it gives the required meaning with the minimum number of words. In fact, it is common to encounter a phrase or sometimes a word that conveys several concepts[3]. Another form of the unique eloquence of the Quran is the use of allegory and analogy, which indeed adds more dimensions to the meaning[4]. All these features and more make the Quran the optimum candidate for representing Classical Arabic, and the most challenging Arabic text to be understood by machines.

On the other hand, most previous research on understanding the Quran and Classical Arabic in general was dominated by the rationalist approaches to language acquisition, which adopt the principle that human brains are built in with a general module for language, and children relies on already built in rules, procedures and structure of that module in acquiring and generating language. In fact, the rationalist approaches dominated most of the work in linguistics, artificial intelligence and natural language processing during the period between 1960 and 1985. Intelligent systems with lots of hardcoded rules and learning abilities were built to simulate the rationalist view of the human brain. However, these systems were criticized by their reliance on a large amount of human provided rules, and their limited ability to handle large problems [3]. This led scientists to revert to empiricist approaches to language learning, which were the dominant in research through the period between 1920 and 1960, and which assume the absence of such general language module and that the human brain is equipped with intelligent mechanisms such as pattern matching and generalization that allow children to learn the complicated structure of language by applying those mechanisms to the input acquired by their senses [3]. Using this essence, empiricist approaches study real world language usage in order to explore the structure and other characterizing features of language. This real world language use is offered to these approaches through corpora, where a corpus (single form of corpora) is a systematically designed collection of written and/or spoken text that can be used in linguistic studies [4].

One of the well known empirical approaches is distributional lexical semantics which is an approach that is mainly concerned by modeling word meaning using words distribution statistics gathered from very large corpora [5]. It is basically built on the notion of the *Distributional Hypothesis*, which dates back to Zellig Haris in 1970 stating that "difference in meaning correlate with difference in distribution" [6]. In other words, it states that "words which are similar in meaning occur in similar contexts" [7], and hence "words that occur in the same contexts tend to have similar meaning" [8].

Therefore, one step towards understanding the Quran is to attempt to study the meanings of the words used in it through analysis of their distributional semantics in contemporaneous texts, and since the Quranic text was revealed in pure Classical Arabic, which forms the basis of Arabic linguistic theory and which is well understood by the educated Arabic reader. Therefore, it is necessary to investigate the distributional lexical semantics of the Quran's words in the light of similar texts (corpus) that are written in pure Classical Arabic. The absence of such corpus is what is believed to be the obstacle hindering research on the distributional lexical semantics of the Quran and Classical Arabic in general.

---

[1] The Quran 4:157
They slew him not for certain. [Pickthal Translation]

[2] www.islamqa.com

[3] The Quran 2:179
And there is (a saving of) life for you in Al-Qisâs (the Law of Equality in punishment). [Dr. Muhsin translation]

[4] The Quran 19:4
He said: O my Lord! verily the bones of me have waxen feeble, and the head is glistening with hoariness, and have not yet been in my prayer to thee, my Lord, unblest. [Abdul Daryabadi translation]

This article reports the design and compilation of a very large corpus of Classical Arabic that can be used as the basis for studying distributional lexical semantics of the Quran and Classical Arabic. It also presents two linguistic empirical studies; the first study tests the reliability of eight probabilistic distributional semantic models in extracting collocations from the Quranic text. While the second study assess the ability of those models in extracting collocations from the large corpus of Classical Arabic testing whether corpus size influence the performance of those models or not.

The paper is structured as follows. Section 2 provides a brief preview about the construction of the Classical Arabic corpus. Section 3 discusses the first empirical study. The second study is presented in Section 4. Finally, Section 5 discusses the conclusions of the work presented.

2. KING SAUD UNIVERSITY CORPUS OF CLASSICAL ARABIC (KSUCCA)

To the best of the authors knowledge, there exist only two corpora of Classical Arabic; one is part of the King Abdulaziz City for Science and Technology Arabic Corpus (KACST Arabic Corpus)[5], which is not very large and only has a limited number of genres, and the other is the Classical Arabic Corpus (CAC) [4], which is a relatively small corpus with only 5 million words. Both of these corpora are not appropriate for research in distributional semantics, which requires a very large and diverse corpus. Therefore, it was essential to design and compose a new corpus of Classical Arabic that is very large, balanced, and representative so that any result obtained from it can be generalized for Classical Arabic.

2.1. Purpose of KSUCCA

KSUCCA is initially compiled to be used for studying the distributional lexical semantics of the Quran and Classical Arabic in general. However, it is designed as a general corpus analogous to the Brown [9], LOB [10], BNC [11], Corpus of Contemporary Arabic (CCA) [12] and other general corpora that can be used for a variety of Linguistics and Computational Linguistics research, such as building lexicons, studying language change through time, collocations extraction, synonyms detection, etc. [13].

2.2. The Design And Compilation Of KSUCCA

Texts included in KSUCCA are Arabic texts dating back to the period of the pre-Islamic era until the end of the fourth *Hijri*[6] century, which is equivalent to the period from the seventh until early eleventh century CE [14]. The corpus is sampled as "full text", where the whole book or poem text is considered as a sample; this is more appropriate for detecting the linguistic features and meanings that may be distributed throughout the text as suggested by Sinclair [15].

Genre	Number of texts	Number of words	Percentage
Religion	150	23645087	46.73 %
Linguistics	56	7093966	14.02 %
Literature	104	7224504	14.28 %
Science	42	6429133	12.71 %
Sociology	32	2709774	5.36 %
Biography	26	3499948	6.92 %
Total	410	50602412	100 %

TABLE 1: The Content of KSUCCA [13].

KSUCCA has 6 genres: Religion, Linguistics, Literature, Science, Sociology and Biography. These genres are further classified into 27 subgenres. The diversity of genres and the amount of texts in them, as in Table 1, is consistent with the knowledge of the overall writing trends at that

[5] http://www.kacstac.org.sa/Pages/Default.aspx

[6] The *Hijri* calendar is the official calendar for Muslims. Its first year was the year when the *Hijra*, migration, of Prophet Muhammad Peace be upon him from *Makkah* to *Madinah* occurred, which is equivalent to 622 CE.

period of Arab history, and it is an indication of the representativeness and balance of the corpus [13].

3. FIRST EMPIRICAL STUDY: IDENTIFYING QURANIC COLLOCATIONS

This is a preliminary study to explore a range of well-known probabilistic distributional semantic models and how well they work on the Quranic text. The goal is to discover the models that are able to identify the correct collocating words from the Quran as early as possible, which might be an indication of the ability of those models to extract significant collocations from very large corpora of Classical Arabic.

3.1 Setting Up The Study

The first step in preparing for this study is to prepare a gold-standard that can be used to test the accuracy of the obtained results. For this purpose, a manually extracted list of the Quranic collocations from a previous PhD study [16] is prepared, and revised by an expert Arabic linguist. Later on, the list is refined so that it only includes collocations with no morphological variants because they will not be identified. The final version of the list has 59 manually identified, revised and refined collocations.

The study is carried out by Sketch Engine using an authenticated corpus of the Quran. Sketch Engine offers eight different well known probabilistic distributional semantic models, or *association measures*, for collocation extraction. These measures include: raw frequency, t-score [17], mutual information (MI), MI3, minimum sensitivity, logDice, MI.log_freq [18] and log-likelihood [19]. All these measures were included in the study.

3.2 Data Sets

The study is performed with two data sets; one with choosing the least frequent part of the collocation as the node word, and the other with the most frequent part of the collocations as the node word. This will allow also to measure the effect of the node word frequency on the performance of the association measures.

3.3 Performing The Study

In this study it is essential to consider the ranking of the correctly identified collocating word; the association measure that predicted the correct collocating word as early as possible is favored over the others. To do so, the rankings of the identified collocations for each association measure were recorded. After that, the average precision (AP) for each association measure is calculated for each node word. Then the mean average precision (MAP) [20] for each association measure is calculated for all node words.

3.4 Results And Discussion

Figure 1 shows the MAP scores for the first run; it can be noticed that all the association measures achieved good results in identifying the correct Quranic collocations. This is due the fact that the Quranic corpus is relatively small, and mostly the least frequent part of the collocation will correlate with only few number of words, which makes it easier for the association measures to detect the correct collocating word.

The log-likelihood association measure scored the best results when choosing the minimum frequency part of the collocation as the node word. Log-likelihood also gave promising results with modern standard Arabic in previous studies [21, 22, 23]. However, it was noticed that in the situations where the log-likelihood failed to identify the right collocating word, the two collocating words occurred separately so frequent in the Quranic text but only rarely together. In addition, raw frequency, MI3, MI.log_freq and t-score also achieved good results nearly similar to log-likelihood making them also interesting candidates. However, the MI association measure scored the lowest MAP, which is expected due to the fact that MI gives high scores for bigrams with low frequencies compared to high frequency ones, which should not be the case [3].

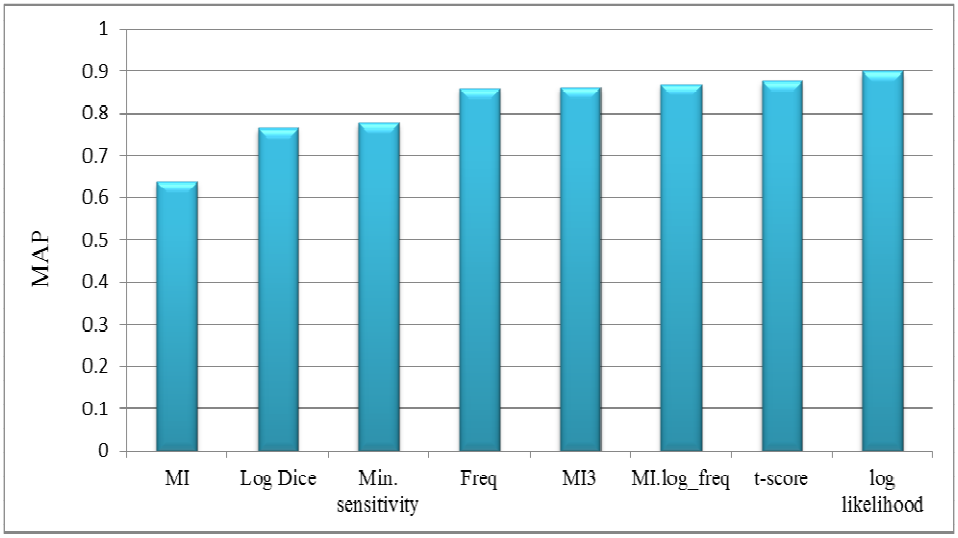

FIGURE 1: MAP scores of the association measures using first data set.

On the other hand, the results of the second run, Figure 2, show that when choosing the most frequent part of the collocation as the node word, the behavior of all the association measures have dropped, which is expected because the most frequent part of the collocation is more likely to be correlated with a large number of words making it more difficult for the association measures to extract the correct collocation. However, it can be noticed that the MAPs have dropped slightly as in the case of logDice, minimum sensitivity and MI, noticeably as in the case of MI3, MI.log_freq and log-likelihood, and drastically as in the case of raw frequency and t-score. This can be of strong indication that the logDice and minimum sensitivity measures are the most immune to data sparseness. In addition, it also indicates that the effect of data sparseness is tolerable on MI3, MI.log_freq and log-likelihood. However, it is not tolerable in the case of the t-score because it tends to ignore co-occurring words that co-occur less than expected, which is confirmed by other studies [24].

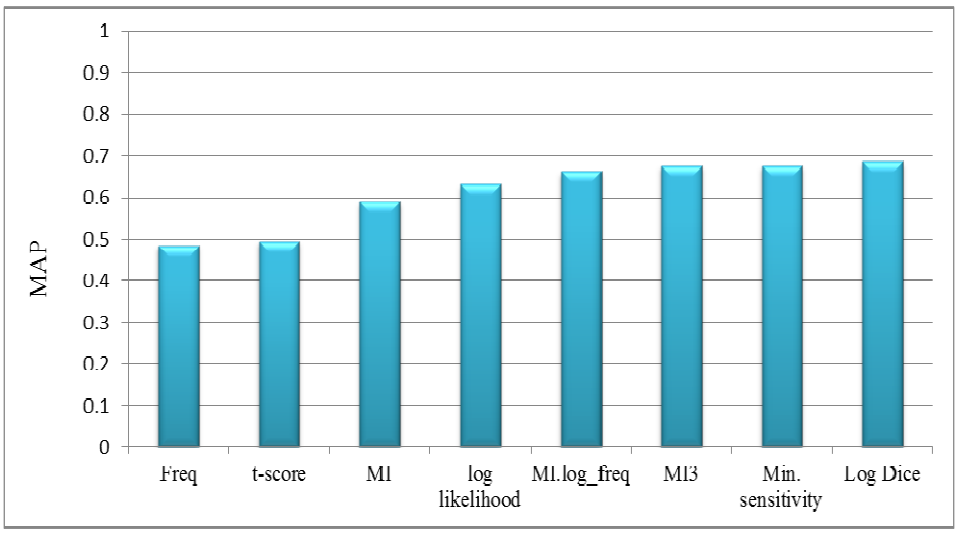

FIGURE 2: MAP scores of the association measures using the second data set.

Looking at the MAP scores of both runs, Table 2 and Figure 3, together with the average MAP scores of each association measure, Table 2 and Figure 4, it can be noticed that logDice and minimum sensitivity are the most stable measures that were not affected very much by the choice of the node word which make them reliable candidates for collocation extraction in relatively small corpora. In addition, log-likelihood, MI.log_freq and MI3 gave acceptable results in the average which also make them interesting candidates. It is also noticeable that raw frequency and t-score tend to have similar behavior in the two tests; and their performance dropped drastically when choosing the maximum frequency part of the collocation as the node word making them bad candidates for collocation extraction. This is expected with the case of raw frequency and also reasonable with the case of t-test since it is directly affected by the frequency of the node word. On the other hand, MI scored the worst in terms of average MAP, which is due to the nature of MI that favors sparse words against common words during collocation extraction and ranking as in [25] making it a bad candidate.

Measure	Minimum frequency	Maximum frequency	Average
MI	64%	59%	61.44%
t-score	87.8%	49.58%	68.7%
raw frequency	86%	48.54%	67.26%
logDice	76.63%	68.66%	72.64%
minimum sensitivity	77.77%	67.78%	72.77%
log-likelihood	90.07%	63.24%	76.66%
MI.log_freq	87%	66.35%	76.68%
MI3	86.26%	67.76%	77.01%

TABLE 2: Comparing the MAP scores for the two data sets.

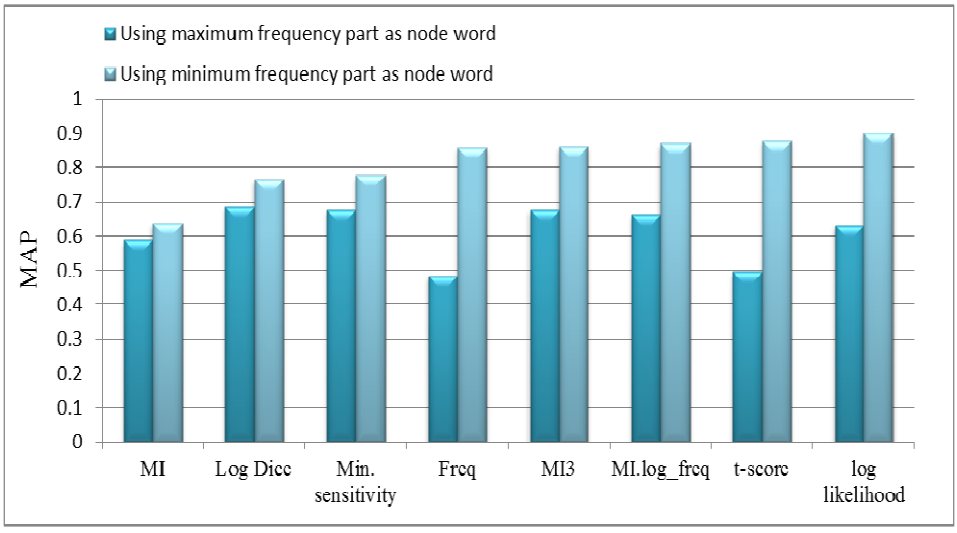

FIGURE 3: Comparing the MAP scores for the two data sets.

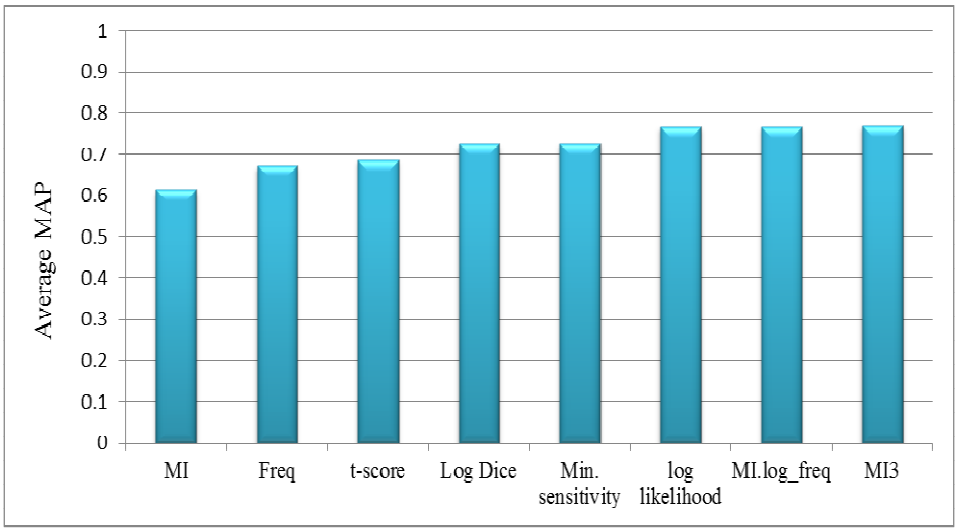

FIGURE 4: The average MAP scores for both data sets.

These obtained results coincide with other reported studies, for example, Boulaknadel et al. [21] tried to extract compound nouns from a small domain specific corpus using four association measures, which are log-likelihood ratio, t-score, MI and FLR [21]. The authors found that the log-likelihood ratio scored the best precision. In addition, Saif and Ab Aziz [22] tried to extract bi-gram collocations from a relatively small corpus of online Arabic newspapers using the log-likelihood ratio, chi-square, MI and the enhanced MI association measures. The results showed that the highest precision was achieved using the log-likelihood ratio.

4. SECOND EMPIRICAL STUDY: IDENTIFYING COLLOCATIONS AND CO-OCCURRENCES FROM KSUCCA

In this study, the results of the previous study are further investigated by applying the same tested association measures to extract collocations from KSUCCA which is considered a very large corpus compared to the Quran. The goal is to assess the performance of the tested association measures and to check whether it is affected by corpus size, and to discover the best association measures that are suitable for extracting significant collocations and co-occurrences from Classical Arabic corpora.

4.1 Setting Up The Study

The first step in this study is to upload the KSUCCA corpus on Sketch Engine, and all the association measures from the first study are going to be tested.

4.2 Data Sets

The study will be performed on a set of 20 node words selected from KSUCCA. Table 3 shows these words and their frequency in KSUCCA.

word	اليد	الوفاء	الليل	الصلاة	بيت	الصبر	العمر	الشمس	□أبا	فر □
frequency	2554	964	14327	22813	12050	2525	893	11934	3396	2844
word	جه□	ليل	عين	الماء	البرق	شعر	البيت	ماء	□لاة	الحياة
frequency	12972	1429	5741	23672	941	5238	19524	13650	12291	3900

TABLE 3: Node words for the second study.

4.3 Performing The Study

During the study, the identified neighboring words within a window of five words surrounding the node word from both sides were recorded for each association measure, and then manually marked by Arabic linguists as a correct collocation, a co-occurrence or an irrelevant word.

4.4 Results And Discussion

The MAP scores for the collocation extraction phase for each association measure are recorded in Table 4 and Figure 5. It can be noticed that the performance of all the association measures have dropped with the increase in corpus size. This is an expected consequence since the increase in corpus size also means the increase of the possibility of finding more correlated words that do not form significant collocations. In addition, MI still has the worst MAP score and this is also expected with the increase in corpus size since, as we mentioned before; it tends to favor sparse words which are unlikely to form significant collocations in a very large corpus.

Measure	MI	Raw frequency	T-score	Log-likelihood	Minimum sensitivity	MI3	LogDice	MI.log_freq
MAP	5.19%	13.39%	14.92%	26.1%	31.72%	34.12%	35.12%	37.02%

TABLE 4: MAP scores for collocation extraction phase.

On the other hand, MI3 outperformed log-likelihood and minimum sensitivity; achieving higher scores together with logDice and MI.log_freq measures. However, the MI.log_freq had the outstanding MAP score.

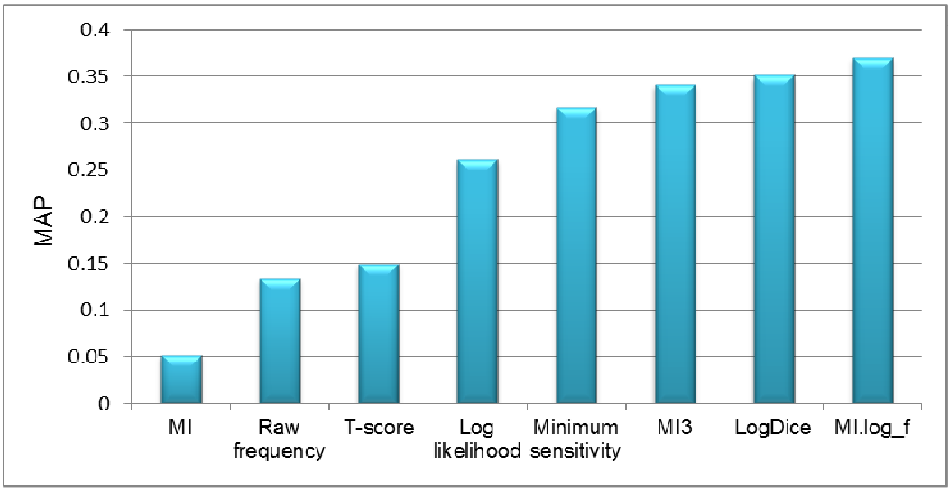

FIGURE 5: MAP scores for the collocation extraction phase.

Looking at the precision scores, Table 5 and Figure 6, it is noticeable that the association measures that had the best ranking were also the ones that retrieved more significant collocations.

Measure	MI	Raw frequency	T-score	Log-likelihood	Minimum sensitivity	MI3	LogDice	MI.log_freq
Precision	12.5%	18.33%	21.33%	33.83%	42.5%	43.5%	45.83%	47.33%

TABLE 5: Precision scores for the collocation extraction phase.

The previous results indicate that MI.log_freq and logDice and MI3 are the best association measures for extracting collocations from both small and very large Classical Arabic corpora in terms of MAP; these findings are also supported by the precision scores of the second study. In addition, the results also indicate that minimum sensitivity tend to produce acceptable results in terms of precision. On the other hand, log-likelihood does not seem to be a good choice when dealing with large corpora.

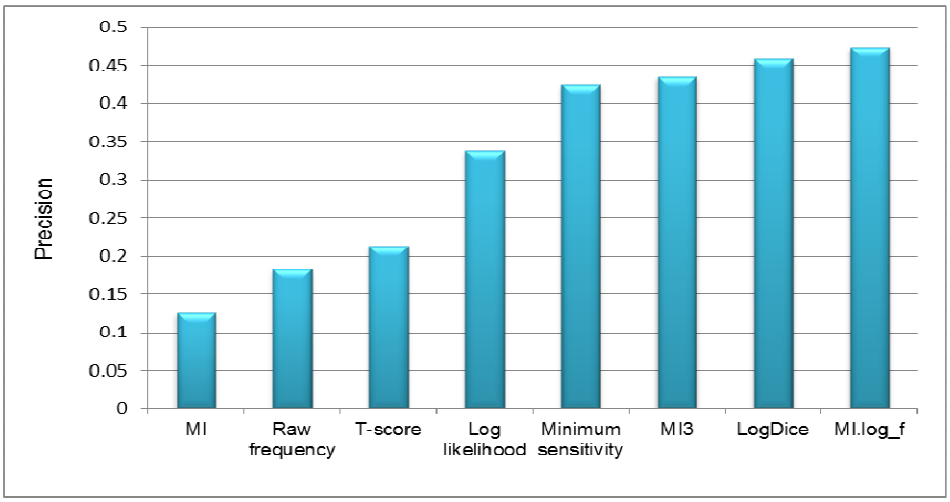

FIGURE 6: Precision scores for the collocation extraction phase.

Regarding the co-occurrence extraction phase, Table 6 and Figure 7, it can be noticed that MI.log_freq achieved the highest results in identifying significant co-occurrences within a window of five words around the node word. logDice also achieved a good MAP score nearly similar to MI.log_freq. In addition, minimum sensitivity and MI3 had nearly similar MAP scores making them also interesting candidates, while MI scored the worst.

Measure	MI	Raw frequency	T-score	Log-likelihood	MI3	Minimum sensitivity	LogDice	MI.log_freq
MAP	8.64%	13.22%	15.09%	28.19%	38.73%	39.04%	42.20%	43.29%

TABLE 6: MAP scores for the co-occurrence extraction phase.

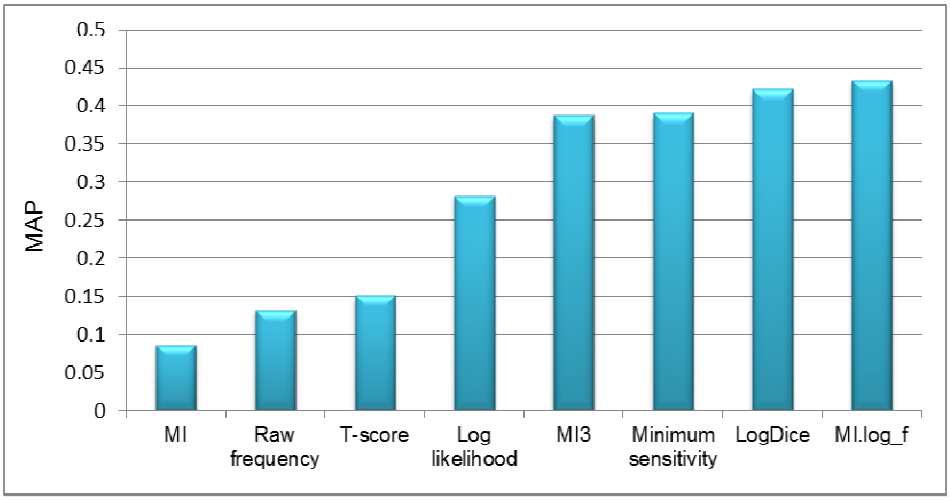

FIGURE 7: MAP scores for the co-occurrence extraction phase.

Looking at precision scores, Table 7 and Figure 8, logDice measure is able to extract the highest number of significant co-occurring words, however, MI.log_freq is also able to extract almost the same number. On the other hand, MI3 and minimum sensitivity achieved nearly similar precisions, while MI had the least one.

Measure	MI	Raw frequency	T-score	Log-likelihood	MI3	Minimum sensitivity	MI.log_freq	LogDice
Precision	21%	27.5%	31.5%	49.66%	61.83%	63.83%	66.83%	67.33%

TABLE 7: Precision scores for the co-occurrence extraction phase.

These results of the co-occurrence extraction phase indicated that MI.log_freq and logDice are the most appropriate candidates, among all the tested association measures, to be used in extracting significant co-occurrences from very large Classical Arabic corpora. Moreover, MI3 and minimum sensitivity are also considered interesting candidates. However, MI showed very poor results in term of MAP and precision, which makes it the worst candidate.

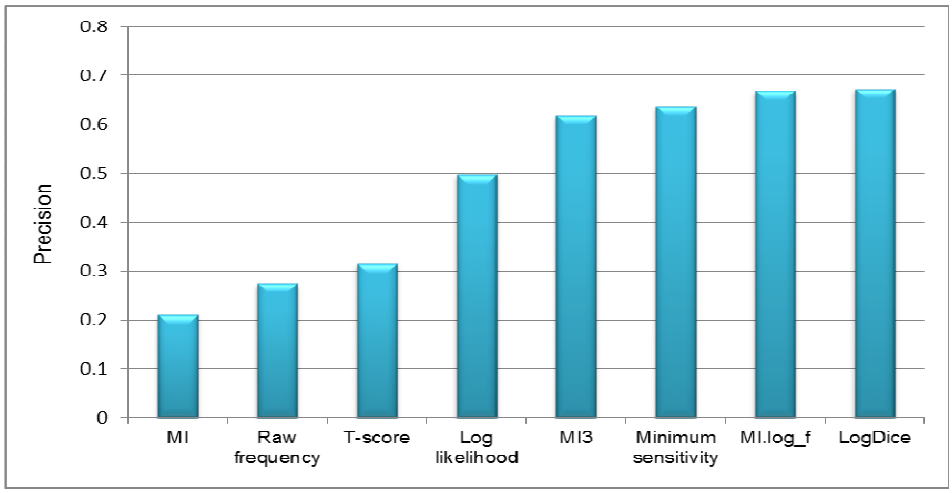

FIGURE 8: Precision scores for the co-occurrence extraction phase.

5. CONCLUSION AND FUTURE DIRECTIONS

In this paper we gave a brief description of KSUCCA; a pioneering 50+ million word corpus of Classical Arabic that captures the culture of a nation. It is a project with scientific, linguistic, cultural, social and religious aspects. KSUCCA was created with the goal of studying distributional semantics of the Quran and Classical Arabic; however, it can also be used for other researches on Linguistics, Computational Linguistics, Literature and History. We believe that KSUCCA will give scholars and researchers a new perspective on the analysis of the language of the Quran and Classical Arabic.

In addition, we have also reported the results of two empirical studies; the first study involves automatic extraction of the Quranic collocations with eight different association measures comparing the results to manually extracted version of the collocations as a gold standard. The study was done in two rounds; in the first round we have chosen the most frequent part of the collocation as the node word, and in the second round we have chosen the least frequent part as the node word. The goal of that study was to determine which of the eight tested association measures is more capable of extracting the most significant collocations from a relatively small corpus of Classical Arabic. In addition to testing how the choice of the node word affects the performance of the association measures. We have concluded that choosing the most frequent part of the collocation as the node word drops the MAP scores of the association measures.

Average MAP scores indicated that the logDice and the minimum sensitivity association measures are the most stable measures that were not affected very much by the choice of the node word which make them reliable candidates for collocation extraction in small corpora of Classical Arabic. In addition, log-likelihood, MI.log_freq and MI3 tend to produce acceptable results on small corpora. On the other hand, the MI, raw frequency and t-score association measures were effected extremely by the choice of the node word, and MI had the worst score in terms of average MAP.

The second study involved extracting significant collocations and co-occurrences from KSUCCA for twenty words using the same association measures in the first study. The goal of this study was to assess the performance of those measures on a very large corpus (KSUCCA) and how they are affected by corpus size, in addition to testing their ability to extract significant co-occurring words. This study had also two rounds; one for collocation extraction and the other for co-occurrence extraction, and the identified collocations and co-occurrences of both rounds were marked by expert linguists. Results for the collocation extraction round indicated that MI.log_freq and logDice are the best association measures for extracting collocations from very large Classical Arabic corpora in terms of MAP scores; they also scored the highest precision scores. In addition, MI3 and minimum sensitivity are also good candidates, while MI had the worst results in terms of MAP and precision scores.

On the other hand, the results of the co-occurrence extraction phase indicated that MI.log_freq and logDice are also the most appropriate candidates, among all the tested association measures, to be used in extracting significant co-occurrences from Classical Arabic corpora. Moreover, MI3 and minimum sensitivity are also considered interesting candidates. However, MI showed very poor results in term of MAP and precision scores, which makes it the worst candidate.

In the future, we plan to explore a wider range of association measures, and try to modify the most accurate ones in order to achieve a measure with higher precision and MAP scores for collocation and co-occurrence extraction from Classical Arabic.

6. REFERENCES

[1] K. Dukes, and N. Habash, (2010). "Morphological annotation of Quranic Arabic." The seventh international conference on Language Resources and Evaluation (LREC-2010), Valletta, Malta, 2010.

[2] A. Ibn Ashoor, *Al-Tahreer wa Al-tanweer,* in Arabic, Dar Sahnoon, Tunisia, 1997.

[3] C.D. Manning, and H. Schuetze, *Foundations of Statistical Natural Language Processing,* 1st ed., The MIT Press, 1999.

[4] A. Elewa, "Did they translate the Qur'an or its exegesis?." 3rd Languages and Translation Conference and Exhibition on Translation and Arbization in Saudi Arabia, Riyadh, Saudi Arabia, 2009.

[5] M. Sahlgren, "The word-space model: using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces." Ph.D. dissertation, Department of Linguistics, Stockholm University, 2006.

[6] Z. Harris, and H. Hiż, "Papers on syntax", Springer, pp. 3-22, 1981.

[7] H. Rubenstein, and J. Goodenough, "Contextual correlates of synonymy." Communications of the ACM, vol. 8, pp. 627–633, 1965.

[8] P. Pantel, "Inducing ontological co-occurrence vectors." In Proceedings of the 43rd Conference of the Association for Computational Linguistics, ACL'05, pp. 125–132, 2005.

[9] W. N. Francis, and H. Kucera, "Brown Corpus Manual: Manual Of Information To Accompany A Standard Corpus of Present-Day Edited American English, for use with Digital Computers." Internet: http://khnt.hit.uib.no/icame/manuals/brown/INDEX.HTM [Feb. 20, 2014].

[10] S. Johansson, E. Atwell, R. Garside and G. Leech, "The Tagged LOB Corpus: Users' manual." ICAME, The Norwegian Computing Centre for the Humanities, Bergen University, Norway, 1986.

[11] L. Burnard, "British National Corpus: User's reference guide for the British National Corpus". Oxford, Oxford University Computing Service, 1995.

[12] L. Al-Sulaiti, and E. Atwell, "The design of a corpus of contemporary Arabic." International Journal of Corpus Linguistics, vol. 11, pp. 135-171, 2006.

[13] M. Alrabiah, A. Al-Salman and E. Atwell, "The design and construction of the 50 million words KSUCCA King Saud University Corpus of Classical Arabic", In Second Workshop on Arabic Corpus Linguistics (WACL-2), Monday 22nd July 2013, Lancaster University, UK, 2013.

[14] M. Eid, *Manifestations Emerging on Arabic*. in Arabic, A'alam Alkutub, Cairo, pp. 20, 1980.

[15] J. Sinclair, "Corpus and Text - Basic Principles." In Developing Linguistic Corpora: a Guide to Good Practice, ed. M. Wynne. Oxford: Oxbow Books, 2005.

[16] H. Duhainah, "Linguistic Collocations and Their Significance in Determining The Semantics of The Holy Quran A Theoretical and Applied Study." in Arabic, PhD dissertation, Al-Azhar University, Cairo, Egypt, 2007.

[17] K. Church, W. Gale, P. Hanks, and D. Hindle, "Using statistics in lexical analysis." In: Uri Zernik (ed.) Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon. Lawrence Erlbaum, New Jersey, pp. 115-164, 1991.

[18] P. Rychly, "A lexicographer-friendly association score". In Sojka, P. & Horák, A. (eds.) Proceedings of Second Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN 2008 , 6-9. Brno: Masaryk University, 2008.

[19] T. Dunning, "Accurate methods for the statistics of surprise and coincidence." Computational Linguistics, vol. 19, pp. 61-74, 1993.

[20] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge, UK: Cambridge University Press, 2008.

[21] S. Boulaknadel, B. Daille and D. Aboutajdine, "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.

[22] A. Saif, and M. Ab Aziz, "An Automatic Collocation Extraction from Arabic Corpus." Journal of Computer Science, vol. 7, pp. 6-11, 2011.

[23] I. Bounhas, and Y. Slimani, "A hybrid approach for Arabic multi-word term extraction." In IEEE, pp. 1-8, 2009.

[24] J. Weeds, and D. Weir, “Co-occurrence Retrieval: A Flexible Framework for Lexical Distributional Similarity.” *Computational Linguistics*, vol. 31(4), pp. 439-475, 2005.

[25] S. Gries, "Useful statistics for corpus linguistics." In Aquilino Sánchez & Moisés Almela (ed.), A mosaic of corpus linguistics: selected approaches, pp. 269-291, 2010.

INSTRUCTIONS TO CONTRIBUTORS

Computational linguistics is an interdisciplinary field dealing with the statistical and/or rule-based modeling of natural language from a computational perspective. Today, computational language acquisition stands as one of the most fundamental, beguiling, and surprisingly open questions for computer science. With the aims to provide a scientific forum where computer scientists, experts in artificial intelligence, mathematicians, logicians, cognitive scientists, cognitive psychologists, psycholinguists, anthropologists and neuroscientists can present research studies, International Journal of Computational Linguistics (IJCL) publish papers that describe state of the art techniques, scientific research studies and results in computational linguistics in general but on theoretical linguistics, psycholinguistics, natural language processing, grammatical inference, machine learning and cognitive science computational models of linguistic theorizing: standard and enriched context free models, principles and parameters models, optimality theory and researchers working within the minimalist program, and other approaches. IJCL is a peer review journal and a bi-monthly journal.

To build its International reputation, we are disseminating the publication information through Google Books, Google Scholar, Directory of Open Access Journals (DOAJ), Open J Gate, ScientificCommons, Docstoc and many more. Our International Editors are working on establishing ISI listing and a good impact factor for IJCL.

The initial efforts helped to shape the editorial policy and to sharpen the focus of the journal. Starting with Volume 6, 2015, IJCL aims to appear with more focused issues related to computational linguistics studies. Besides normal publications, IJCL intend to organized special issues on more focused topics. Each special issue will have a designated editor (editors) – either member of the editorial board or another recognized specialist in the respective field.

We are open to contributions, proposals for any topic as well as for editors and reviewers. We understand that it is through the effort of volunteers that CSC Journals continues to grow and flourish.

IJCL List of Topics:
The realm of International Journal of Computational Linguistics (IJCL) extends, but not limited, to the following:

- Computational Linguistics
- Computational Theories
- Formal Linguistics-Theoretic and Grammar Induction
- Language Generation
- Linguistics Modeling Techniques
- Machine Translation
- Models that Address the Acquisition of Word-order
- Models that Employ Statistical/probabilistic Gramm
- Natural Language Processing
- Speech Analysis/Synthesis
- Spoken Dialog Systems
- Computational Models
- Corpus Linguistics
- Information Retrieval and Extraction
- Language Learning
- Linguistics Theories
- Models of Language Change and its Effect on Lingui
- Models that Combine Linguistics Parsing
- Models that Employ Techniques from machine learning
- Quantitative Linguistics
- Speech Recognition/Understanding
- Web Information

CALL FOR PAPERS

Volume: 6 - **Issue:** 1

i. Paper Submission: November 30, 2014 **ii. Author Notification:** December 31, 2014

iii. Issue Publication: January 2015

CONTACT INFORMATION

Computer Science Journals Sdn BhD

B-5-8 Plaza Mont Kiara, Mont Kiara

50480, Kuala Lumpur, MALAYSIA

Phone: 006 03 6204 5627

Fax: 006 03 6204 5628

Email: cscpress@cscjournals.org

CSC PUBLISHERS © 2014
COMPUTER SCIENCE JOURNALS SDN BHD
B-5-8 PLAZA MONT KIARA
MONT KIARA
50480, KUALA LUMPUR
MALAYSIA

PHONE: 006 03 6204 5627
FAX: 006 03 6204 5628
EMAIL: cscpress@cscjournals.org