Novel approach in multilingual and mixed English-Arabic test collection

Mohammed M. Ali

Department of Information Technology,
Faculty of Computers and Information Technology,
University of Tabuk,
Tabuk, Kingdom of Saudi Arabia
Email: mmustafa@ut.edu.sa

Mohammed M. Abu Shquier*

Faculty of Computer Science and Information Technology,
Jerash University,
Jerash, Jordan
Email: shquier@jpu.edu.jo
*Corresponding author

Afag Slah Eldeen

Sudan University of Science and Technology,
Khartoum, Sudan
Email: afagSalah@hotmail.com

Mohamed E. Zidan

Faculty of Science,
University of Tabuk,
Tabuk, Kingdom of Saudi Arabia
Email: mohamed@nassp.uct.ac.za

Ra'ed M. Al-Khatib

Department of Computer Sciences,
Faculty of Information Technology and Computer Sciences,
Yarmouk University,
Irbid, Jordan
Email: raed.m.alkhatib@yu.edu.jo

Abstract: Mixing languages together in text and in talking is a major feature in non-English languages in developing countries. This mixed grammar is also emerging in SMS, Facebook communication, searching the web and any future attempts also may increase the footprint of such a mixed language knowledge base. Traditional information retrieval (IR) and cross-language information retrieval (CLIR) systems do not exploit this natural human tendency as the

underlying assumption is that user query is always monolingual. Accordingly, the majority of the text collections are either monolingual or multilingual. This paper explores the trends of mixed-language querying and writing. It also shows how the corpus is validated statistically and how an Arabic lexicon can be extracted using co-occurrence statistics. Results showed that the distribution of frequencies of words in the corpus is very skewed the vocabulary growth is a good fit. The results of how to handle mixed queries are also summarised.

Keywords: multilingual; monolingual; multilingualism characteristic; retrieval of documents.

Reference to this paper should be made as follows: Ali, M.M., Abu Shquier, M.M., Eldeen, A.S., Zidan, M.E. and Al-Khatib, R.M. (2020) 'Novel approach in multilingual and mixed English-Arabic test collection', *Int. J. Computing Science and Mathematics*, Vol. 11, No. 3, pp.291–304.

Biographical notes: Mohammed M. Ali holds a PhD in Computer Science. His research interests include artificial intelligence and text classification.

Mohammed M. Abu Shquier holds a PhD in Computer Science. His research interests focuses on developing novel Arabic machine translation (rule and statistical-based). He also interested in conducting research in the areas of computational linguistics, information retrieval and Arabic natural language processing in general. In addition to that he had conducted good research in the area of Arabic morphology, syntactic and Symantec.

Afag Slah Eldeen holds a PhD in Computer Science. His research interests include artificial intelligence and natural language processing.

Mohamed E. Zidan holds a PhD in Computer Science. His research interests include artificial intelligence of text mining.

Ra'ed M. Al-Khatib holds a PhD in Computer Science. His main research interests include artificial intelligence, natural language processing (NLP) and machine learning. He also interested in conducting research in the areas of computational bioinformatics, information retrieval, IoTs, high parallel computing (HPC), biometrics and Arabic NLP in general. In addition to that he had conducted good research in the area of wireless sensor networks (WSNs).

This paper is a revised and expanded version of a paper entitled 'Mixed language Arabic-English information retrieval' presented at 16th International Conference on Computational Linguistics and Intelligent Text Processing, CICLing 2015, Cairo, Egypt, 14–20 April 2015.

1 Introduction

Mixing languages together in term of speaking and writing is a widespread phenomenon in many countries where, their citizens speak more than one language. However, in these particular communities; natives are expressing some phrases differently. This mixed-language trend is known as code-switching (Rieh and Rieh, 2005; Gupta et al., 2014; Bhat et al., 2014). However, the common factor is the use of English as a pivot/second language. This is because of the dominance of English language. Hence,

non-English native speakers, i.e., (Arabic speakers), they often search for mixture of languages, in order to approach their needs more precisely, rather than using monolingual queries written in their native-tongue. This new type of search can be identified as mixed or multilingual querying. It is also referred to as the bilingual query (Mustafa and Suleman, 2015). A mixed query is defined as a query which has been written in more than one language – usually bilingual, for example, 'مفهومال polymorphism, (meaning: concept of polymorphism). English portions in mixed queries are often the most significant keywords. In the same context, a mixed or a multilingual document is defined as a document, whose text is often presented/scattered in terms of terms/portions/snippets/phrases/paragraphs in both primary and some other languages (Mustafa, 2013; Fung et al., 1999). The issue of mixed-language in querying and writing has attained little attention in the information retrieval (IR) studies.

This paper is an on going study to a work begins earlier, to present the first-phase, i.e., (building the corpus) – to illustrate the trends of mixed-language querying and writing, from an IR perspective, with special focus on Arabic/English multilingual and mixed texts in scientific domains. The paper shows the main features of the corpus after increasing its size and the statistical tests that have been conducted.

2 Related work

In the field of text-retrieval research, several text collections for many languages have been developed. In terms of their languages, corpora are either classified as monolingual or multilingual. They can be also categorised into general or specialised corpora from a genre prospective. In terms of their vocabularies, corpora can be synchronic or diachronic (McEnery et al., 2006). Synchronic corpora are often used to compare regional varieties, whereas diachronic, or historical, corpora are usually used to compare vocabulary during different time periods (McEnery et al., 2006). An example of a monolingual collection, is the Arabic Agence France Presse (AFP) (Graff and Walker, 2001), which is an Arabic (monolingual) newswire collection. The different editions of TREC, NII and CLEF test collections which are multilingual. Arabic has been included in TREC in 2001, in the same cross-lingual track. An example of specialised Arabic corpus is Hmeidi corpus (Hmeidi et al., 1997), who has built an Arabic corpus with 242 abstracts gathered from the proceedings of the Saudi Arabian national computer science conference. Abdelali et al. (2005) constructed a large synchronic corpus in modern standard Arabic (MSA), which is a modern version of the Arabic language; that is usually used in formal communications, from different regional Arabic newspapers. A similar approach was also conducted by Gamallo et al. (2016).

Su et al. (2017) collected a unique prototype corpus on Arabic texts chosen by children. The corpus is monolingual and covers a variety of children genres, i.e., (fictional characters and classical tales). The corpus contains only 2 million words. Results showed that naive Bayesian classifiers perform well; if they are applied on in-domain. Different methods are employed to gathering corpora (Su et al., 2017). For those languages with wide computational resources, it is often to collect their corpora using three major approaches. These are: automatic crawling and harvesting, which is based on a pre-defined list of URLs; automatic and/or manual downloading based on manual and/or automatic submission of queries to search engines and manual collection

of documents. On the other hand, for languages with small amounts of readily available data (i.e., in the web), crowdsourcing approaches have become more efficient (Su et al., 2017). Crowdsourcing is the task of distributing burden to a group of users via the internet in a collaborative manner (Helmy et al., 2016). Inspired by sentiment analysis, Kiritchenko and Mohammad (2016) built an Arabic twitter lexicon containing single Arabic words and simple negated expressions from the Arabic tweets. The researchers reported that it is possible to obtain consistent annotation results via annotating and ranking of words using crowdsourcing paradigm. A similar approach has been also used by Helmy et al. (2016) for extracting key phrases.

3 Building the test collection

Recall, the three approaches, which were discussed in the related work, i.e., (automatic crawling; automatic and/or manual downloading; and manual collection); were also used to gather the corpus documents collection from the web. The primary reason for this variety of approaches; was due to the fact that computer science documents in more than one language are not always available, with adequate mass and varied contents placed into a suitable electronic format, to be collected by crawlers. In the first approach, i.e., (automatic crawling), a list of URLs was prepared and utilised for seeding a web crawler – namely WebReaper. The process was run intermittently, rather than continuously, and from time to time to avoid congesting target servers. A manual collection of data was also considered. In particular, a group of 100 Arabic native students/tutors, at a different academic levels in some Arabic universities, were asked to collect documents on common computer science topics. Contents of the corpus are mainly collected from references, papers, websites, books, online help, students essays and articles, software documentation, forums, patents, etc., but yet all of them are from computer science fields. All the collected pages and documents were merged together into a single pool. This results in a total size of 14.2 GB of raw text data with a total number of documents equals 90,583.

3.1 Corpus processing

After gathering the document collection, it was processed in order to create a cleaned HTML format corpus. The cleaning process was conducted through two phases: in the first pass of processing, pages in different HTML formats (i.e., SHTML, HTML), were automatically processed, while preserving the same format of pages. Throughout this cleaning, formulae, ellipses, figures, mathematical symbols, images, HTML tags and punctuations were not discarded, on the other hand, only weird symbols (®, §, ^{T}M), fixed comments and navigational data, were removed. An application was also developed to create HTML files from documents that were created by word processor software, i.e., (doc, RTF, txt, etc.). An Adobe Acrobat Reader edition with Semitic languages support was also used to convert PDF files into HTML format.

Categories within the computer science discipline, were manually identified in each HTML and/or created HTML file. Regional variants in the Arabic texts in the collection were preserved as they appear in documents. Regional varieties were extracted automatically to create a lexicon, using some co-occurrence measures – as it will be described later in this paper.

In the second pass of processing, which was developed to create the textual version of the corpus in HTML format, with HTML extension, Jericho application has been used to generate high level manipulation of HTML files. Jericho also has the ability to recognise all types of server tags (ASP, JSP, PSP, PHP, etc.) and, thus, HTML files were parsed properly. Case-folding in English text were kept. On the other hand, a very limited normalisation process for Arabic texts was carried out, i.e., (removal of diacritical marks and Arabic kasheeda). Recall, every word/phrase/portion/paragraph/document, was marked with a language tag attribute, using a simple language identifier. Furthermore, pure textual documents in HTML format with a single codeset and a size of 797 MB (0.8 GB) were produced. The corpus has been named multilingual and mixed English Arabic corpus (MULMIXEAC).

3.2 Corpus statistics

In order to obtain the essential information needed for the corpus, the Lucene IR system was used. So during the indexing process, appropriate terms were extracted (without stemming) and populated in the Lucence index. However, Lucence tools and a developed application code were used to extract some statistics about the corpus. Table 1 shows these statistics. So, it can be observed in Table 1, that English is still the dominant language in common computer science domain, at least in terms of preferences of Arabic scholars. Monolingual Arabic documents on computer science are very scarce. This is due to the fact that Arabic speakers, especially scholars, do not know the proper translations and/or exact meanings for most terminology in their native language.

Table 1 The most frequent 20 unigrams in each language (top 40 words) in the corpus

Rank	*Arabic*				*English*			
	Token	*Freq.*	*%*	*Pr(i)* rank$_i$*	*Token*	*Freq.*	*%*	*Pr(i)* rank$_i$*
1	من	119,147	0.285	0.003	the	2,242,811	5.366	0.054
2	يف	111,064	0.266	0.005	of	900,770	2.155	0.043
3	على	72,013	0.172	0.005	to	876,674	2.097	0.063
4	و	65,141	0.156	0.006	a	859,291	2.056	0.082
5	نأ	40,445	0.097	0.005	and	694,178	1.661	0.083
6	وأ	36,467	0.087	0.006	is	603,148	1.443	0.087
7	إلى	34,116	0.082	0.006	in	592,150	1.417	0.099
8	التي	30,574	0.073	0.007	for	409,930	0.981	0.078
9	هنه	30,505	0.073	0.007	The	348,493	0.834	0.075
10	هنا	26,769	0.064	0.007	this	278,881	0.667	0.067
11	عن	25,897	0.062	0.007	be	252,061	0.603	0.066
12	البينات	25,549	0.061	0.007	are	224,024	0.536	0.064
13	مح	19,788	0.047	0.008	as	217,454	0.52	0.068
14	هو	18,927	0.045	0.007	you	216,721	0.519	0.073
15	لا	18,610	0.045	0.007	by	213,440	0.511	0.077

Note: *Freq. = Frequency.

As shown in Table 1, we can tell that the total number of tokens in the corpus is relatively high (approximately 42 million words). This is true when it is compared to the number of tokens in many standard collections, especially the Arabic ones. For instance, the 2001 LDC Arabic AFP collection (Graff and Walker, 2001) contains 76 million tokens, which is close to the words in the MULMIXEAC corpus, despite the big difference in number of documents in each collection. The number of documents in the 2001 LDC collection is 383,872 (about 5.5 times larger than the constructed corpus). This phenomenon of larger number of words in the created corpus is mainly caused by its genre type. The same phenomenon of higher numbers of words is also observed when the unique words were extracted beside, the wide use of regional variety of the Arabic vocabulary in computer science, additionally, Arabic characteristics, i.e., (Arabic grammatical rules, orthography, large number of affixes) contribute also in increasing number of distinct words in the created collection.

3.3 Corpus assessment

Usually, whenever a corpus is collected, both corpus major features and statistical nature should be explored to indicate whether the corpus is valid and appropriate to serve as a test-bed or not. One of the most important measures to study the characteristics of texts, is the statistical models of word occurrence. Next section shows statistical test that were applied.

3.3.1 Zipf's distribution

Form statistical point of view, the distribution of frequencies of words in text is predicted to be very skewed (Croft et al., 2010; Christopher et al., 2008). This means that only small number of words, usually the most common, would have very high frequencies, whereas many words would have low frequencies. Thus, frequencies reduce rapidly with their ranks after the frequencies of the most common words. This statistical distribution is usually described by the Zipf's law, which is a commonly used model for describing the frequency distribution of words in a language or a collection. Given a corpus in a natural language, Zipf's law states that the frequency *freq* of any word in a collection (collection frequency of a given term) is proportional to the inverse of its position in the word list or its rank in the same corpus. Alternatively, the frequency of a word *freq* times its rank is approximately a constant *k*:

$$k = freq * rank \tag{1}$$

Ideally, when log(*freq*) is drawn against log(*rank*) in a graphical representation, a straight line with a slope of –1 is obtained. Sometimes the frequency freq of a given term *i* at rank $rank_i$ is substituted by its probability Pr(*i*), which is computed as the frequency of that term over the total number of terms in the collection.

To apply Zipf's law in the MULMIXEAC corpus, the unigram language model (the frequency of each token) is employed. Hence, unique words with their ranks and frequencies are firstly extracted. Table 1 illustrates the most frequent 20 words in each language in MULMIXEAC along with their frequencies and their percentages of appearance (converted probability). Apparently, in both the two languages, most frequent words are prepositions, particles, definite articles and stopwords.

It is noticed that the frequencies of words begin with very high values (see the frequencies of the words at the 1st rank in each language). Then, the frequencies in both languages begin to decrease rapidly as new words are ranked and after the appearance of the few frequent words.

Figure 1 A log-log Zipf's curves (actual and predicted) for the 675,008 unigrams in the corpus (see online version for colours)

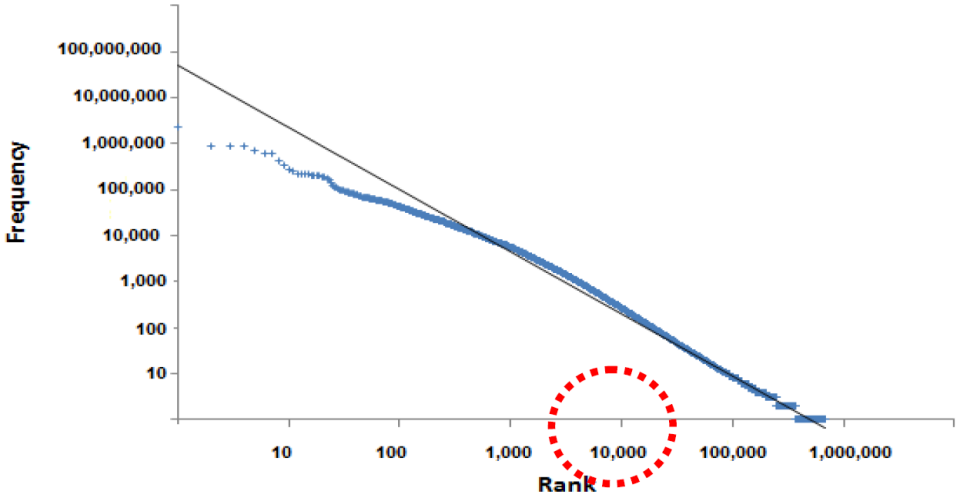


Figure 1 shows Zipf's law applied to the unigrams in the MULMIXEAC corpus. The blue curve illustrates the actual relationship between the word ranks and their frequencies, whereas the straight line in black shows the predicted relationship between them using the least square method that calculates the best fitting line to data.

Form Figure 1, it is observed that the actual Zipf's curve and the fit of the data to the law from the corpus is quite close. In particular, the curve clearly reveals that frequencies of words decrease rapidly with rank (skewed distribution). Thus, it appropriately predicts words frequencies in the corpus with a slope of –1, approximately, except for some most frequent words. In particular, Figure 1 reveals that the actual curve is inaccurate for, approximately, the words in the first 300 ranks.

From the dashed circle in Figure 1, it is noted that the majority of the words in the corpus are hapaxes (words that occurring only once). This is due to that MULMIXEAC is a special corpus in common computer science, whose vocabulary is expected to be diverse. The Arabic morphology contributes to these hapaxes, as well.

3.3.2 Vocabulary size estimation

The size of the vocabulary in corpora is usually estimated by Heap's law. The Heap's law is used to predict vocabulary growth in a certain collection (Croft et al., 2010). In particular, the law is a power function states that the number of distinct words d (vocabulary size) in a given collection with M words (corpus size) is approximately $sqrt(M)$. Formally:

$$d = a * M^{\beta} \tag{2}$$

where a and β are parameters that vary from a certain corpus to another. The typical values for the parameters a and β are: $10 \leq a \leq 100$ and $\beta \approx 0.5$ (between 0.4–0.6). The reason behind the quite large range in the variability of the a parameter is that it depends on factors like stemming, case-folding and spelling errors (Christopher et al., 2008). For instance, spelling errors are directly proportional to the growth rate.

Figure 2 The predicted and the actual vocabulary growth in the corpus using the Heap's law (see online version for colours)

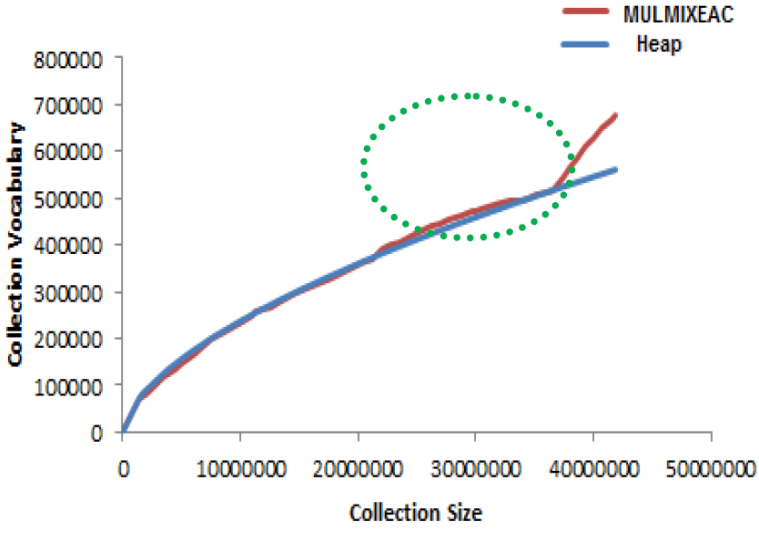


Figure 2 shows both the predicted vocabulary growth, corresponding to the blue curve in Figure 2, and the actual Heap's curve (with $k = 15$ and $\beta = 0.6$), represented in the figure with a red curve, for the MULMIXEAC collection. Figure 2 shows clearly that the growth of the vocabulary in the corpus is a good fit. Thus, new words result in a rapid increase in vocabulary when the collection size is small. However, when the corpus size increases, more new words would still increase the vocabulary size but, at slower rates.

Nevertheless, as the number of words reaches approximately more than 35 million (which is approximately the total number of the English words) it begins again to increase rapidly, instead of steadily – see the green dotted oval in Figure 2. There is a possible explanation for this observation, which is mainly caused by the multilingual characteristic of the corpus.

It is usually observed that English documents are named with an English file name, whereas the names of Arabic documents are mostly in Arabic, although there are many names that are mixed (begin with Arabic or English letters). This fact causes English documents, and thus English words, to be ranked ahead and before the Arabic documents, as the Arabic letters often have a higher codeset and thus, lower ranks, when the application program begins to accumulate both the number of words and the distinct words. Thus, when English vocabulary begins to grow at slower rate (after the rapid increase at the beginning), Arabic documents appear and they begin to accumulate their vocabulary and thus, the curve begins to jump, approximately after more than 35 million words. Meanwhile, it is possible to randomise the document selection after applying a numbering mechanism for documents. However, another scenario was applied, that is to

implement Heap's law for each language separately in the MULMIXEAC. As an example, the Heap's law was applied to Arabic text only.

Figure 3 The predicted and the actual vocabulary growth in the corpus using the Heap's law (see online version for colours)

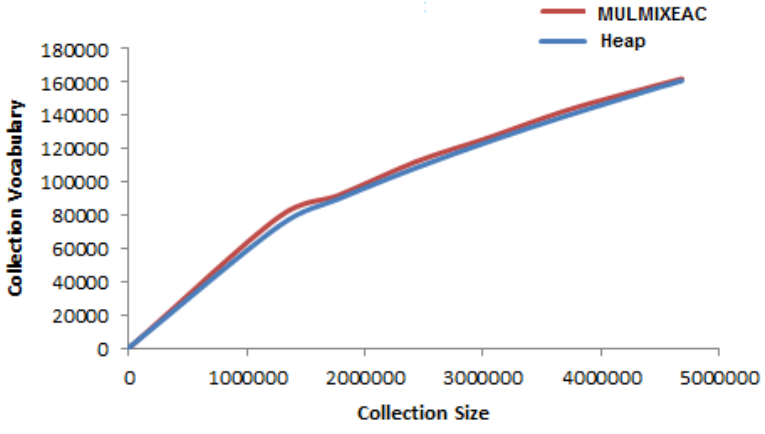


Figure 3 shows the predicted vocabulary growth along with the actual growth in the corpus for Arabic texts. The curve is a good fit. This good prediction is clear at different points. For instance, in the first 1,368,222 words in the corpus, Heap's law estimates that the number of distinct words is 76,880, whereas the actual value is 77,991. Furthermore, in 4,683,724 words, Heap's law predicts 160,870, whereas the actual number is 162,032, which is very close to the predicted value. This example confirms what was concluded above when the entire corpus is analysed together.

3.3.3 Token-to-type ratio

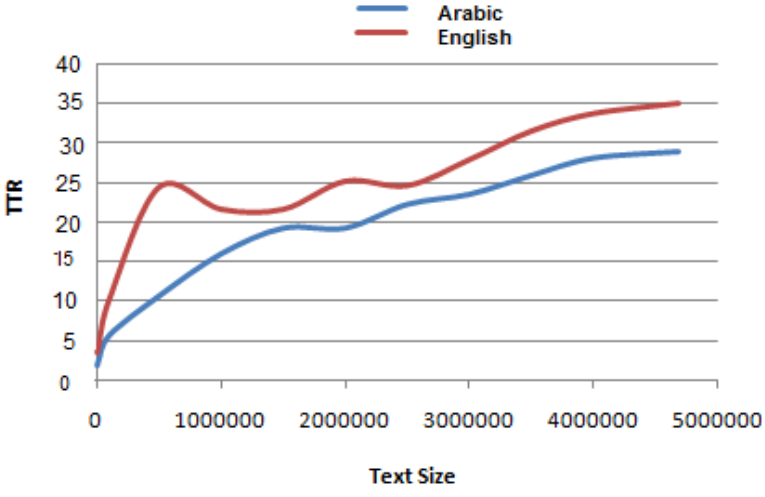
Token-to-type ratio, known as the TTR, is a lexical variety measure for text often used to evaluate the richness of collections and their adequacy for a specific task, i.e., in IR (Mike, 2014). As stated in Abdelali et al. (2005), this measure reflects mainly sparseness of data. The TTR is computed as the number of occurrences of words (tokens) divided by the number of unique words (token types). The TTR tends to be influenced by many factors, including corpus genre, orthography, stemming and case-folding. For instance, orthography usually results in several token types and, thus, it is inversely proportional to the TTR. Hence, high orthography results in a comparatively low TTR.

The TTR is informative if it is used with a corpus comprising lots of equal-sized text fragments (Mike, 2014). Therefore, in MULMIXEAC, different and equal text length(s) (points) for both Arabic and English are used after the corpus is processed. This was done by accumulating words at the selected points, regardless of their positions inside documents.

Figure 4 plots the TTR curves for both languages in the corpus. In the figure, the x-axis represents the different fragments of the texts that had been selected for the two languages, whereas the y-axis presents the computed TTR, according to (tokens/token types). However, since there are about around 4 million tokens in the Arabic texts, the

same number of tokens had been extracted from the English tokens. This is important for both consistency and comparison.

Figure 4 Token-to-type ratios, computed for the first 4,683,724 tokens, for both Arabic and English texts in the MULMIXEAC corpus (see online version for colours)



Although similar fragments' sizes were taken, it is observed in Figure 4 that Arabic has more distinct words in the corpus than English and thus resulting in lower TTR ratios at all different text sizes. This is caused by the Arabic morphology, richness of Arabic vocabulary and Arabic orthography, including phonological orthography like regional variants across scientific terminology. The findings that the TTR of Arabic is lower than for English, had been concluded by many researchers (Alotaiby et al., 2009; Goweder and De Roeck, 2001).

Figure 5 Ratio of Arabic word types to English word types, computed for the first 4,683,724 tokens in each language in the corpus (see online version for colours)

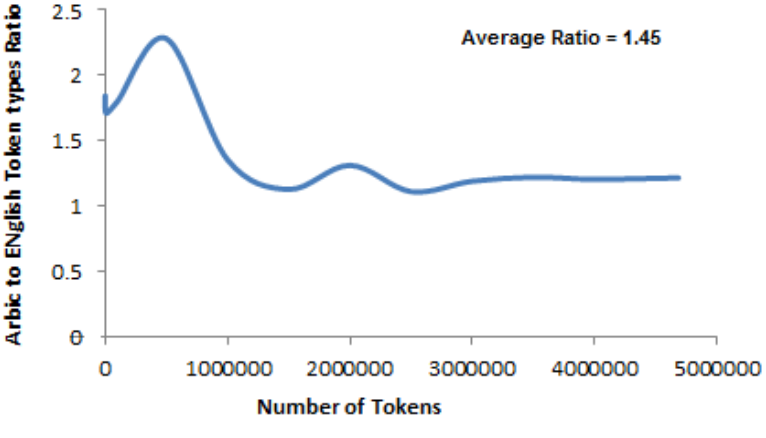


Figure 5 plots the ratios of the Arabic TTR to English TTR in the corpus, shown at the same different text sizes in Figure 4. The average ratio is 1.45. This finding may be useful for building thesauri, lexicon, automatic translation and text summarisation in computer science field. This is because it shows the needed token type sizes when Arabic and English languages are considered, specifically, for 1×10^6 token types in English, one might need about $14{,}500 \times 10^3$ token types in Arabic so as to achieve same contents.

3.3.4 Building a lexicon with Arabic regional variants

Regional variations in the corpus have been extracted automatically. However, language models were not used in this study, alternatively, we have built an Arabic regional-varieties lexicon-like. Similar approach has been used by Cheng et al. (2004) and Nie (2010). The basic idea is based on the fact that Arabic terms, tend to co-occur with the same English translated term. From that perspective, if the Arabic neighbouring terms are properly extracted and the duplicates are removed, it is likely to obtain regional variants. Hence, a list of technical English terms were firstly prepared. Afterward, each term has been automatically submitted to the MULMIXEAC corpus and the result lists were obtained. For each result list, its snippet is added to a small corpus of result snippets. Thus, the translation candidates were added to the corpus. To extract Arabic variants, a combined approach of the Chi-square and the context vector methods was developed in order to estimate association of the English term under question with its co-occurred terms. After the ranked Arabic variants are extracted, they were manually revised to extract regional variants. The details of the utilised approach were provided in Cheng et al. (2004). An example of this created lexicon is the entry of term hashing, which was found to have the following regional variants: التشتت، البعثرة and الفرم.

3.4 Query set

Queries for experimental purposes can be created using different approaches. One realistic approach when moving from experimental systems to realistic systems is to use queries that are collectively represent queries posted by the users of the target application (Croft et al., 2010). Such queries may be acquired either from a query log from a similar application or from potential users directly. Such an approach (asking potential users for sample queries) provides more realistic results and fills the gap between the real environment and the environment of the developed algorithms. Moreover, the approach has being used in creating query sets in many well-known forums, for example TREC query track. Therefore, this approach is followed to create the query set for the MULMIXEAC collection, although, the scenario here is somewhat different. This is because in the task of TREC query track, users are usually asked to submit examples for queries after being shown the texts of topics, which are the information needs. However, in creating a query set for MULMIXEAC another approach was applied, as will be discussed next.

4 Reporting experiments

Several experiments using the created corpus have been conducted. The details of these experiments are reported in Mustafa and Suleman (2015), Mustafa (2013) and they are concluded in this section. In that work, it was shown that current search engines and traditional cross-language information retrieval (CLIR) systems perform poorly when handling mixed-language queries and documents. In most cases, their result lists are dominated by mixed documents. In particular, current approaches tend to perform exact matching between queries and documents, regardless of the languages presence, rather than retrieving the most relevant documents. Hence, the weights of the mixed queries in documents are often computed from the entire mixed query regardless of the document language. Accordingly, there may be many monolingual highly relevant documents that are poorly ranked. Thus, the result list is biased towards mixed documents. The study concluded that the majority of algorithms, as well as the test collections, are optimised for monolingual queries, even if they are translated.

Recall these claims, the major goal of citemustafa2015mixed, Mustafa (2013) study, was to develop an IR system that can handle mixed queries and mixed documents effectively. The experiment, firstly, reports that the best approach of indexing mixed documents, is the use of a single index – rather than traditional distributed indices. The study concluded also that the weighting components should be adjusted to fit this feature of multi-lingualism. Secondly, the experiment reports that the use of a monolingual approach of weighting handles each term and its translation as two different terms. This means that two different weighting (one for the source term and the other for its translation) are computed when finding documents scores. Hence, this would result in biasing term frequency and document frequency statistics. The basic idea behind the developed approach is based on suppressing the impact of the co-occurred terms in different language in the same documents by handling them as if they are a single term or synonymous across languages. Thus, any technical source query term can be reconsidered as a language-aware, by obtaining its translations firstly and then grouping all the candidate translations together with the source term itself, resulting in cross-lingual synonyms. Thus, term frequency, document frequency and document length components were re-estimated using this proposed cross-lingual re-weighted approach. The results concluded that the proposed approach improves the performance significantly and could empower and present a route for future search engines, which should allow multilingual users to retrieve relevant information created by other multilingual users.

5 Conclusions and future work

Most existing test collections, and most CLIR collections, are mainly focusing on rapid use of general-domain news stories. Furthermore, specialised corpora are limited for many languages, including Arabic. For this purpose a multilingual and mixed Arabic-English test collection, named MULMIXEAC, on common computer science vocabulary domain has, therefore, been created. The corpus is primarily gathered from the web as using such a resource is cheap and allows building a large amount of data in any genre within a relatively short time. However, issues like dissemination of information and copyright permission are major difficulties that prevent collecting much larger data. The corpus could be considered as a valuable resource and a useful collection

for research in text retrieval, text classification, machine translation and natural language processing community in general. Although some useful statistical tests were applied to the corpus, we believe that this corpus is not fully representative for common computer science domain, mainly, because the copyrights issues prevent having a truly representative corpus in such specialised domain. Future work will focus on extending the corpus in terms of size and to develop more algorithms to handle the multi-linguality feature of mixed documents. Other co-occurrence measures will be tested to extract translation for those technical terms that are not found in dictionaries. With information globalisation and moving towards an international community, it becomes essential not to constrain non-English speakers, such as Arabic users to single languages in searching.

References

Abdelali, A., Cowie, J. and Soliman, H. (2005) 'Building a modern standard Arabic corpus', in *Workshop on Computational Modeling of Lexical Acquisition. The Split Meeting*, Croatia, 25–28 July.

Alotaiby, F., Alkharashi, I. and Foda, S. (2009) 'Processing large arabic text corpora: preliminary analysis and results', in *Proceedings of the Second International Conference on Arabic Language Resources and Tools*, Cairo, Egypt, pp.78–82.

Bhat, I.A., Mujadia, V., Tammewar, A., Bhat, R.A. and Shrivastava, M. (2014) 'IIIT-H system submission for fire 2014 shared task on transliterated search', in FIRE 2014 Proceedings of *the Forum for Information Retrieval Evaluation*, ACM, pp.48–53.

Cheng, P-J., Pan, Y-C., Lu, W-H. and Chien, L-F. (2004) 'Creating multi-lingual translation lexicons with regional variations using web corpora', in *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, Association for Computational Linguistics, p.534.

Christopher, D.M., Prabhakar, R. and Hinrich, S. (2008) *Introduction to Information Retrieval. An Introduction to Information Retrieval*, Vol. 151, No. 177, p.5.

Croft, W.B., Metzler, D. and Strohman, T. (2010) *Search Engines: Information Retrieval in Practice*, Addison-Wesley, Reading.

Fung, P., Xiaohu, L. and Shun, C.C. (1999) 'Mixed language query disambiguation', in *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, Association for Computational Linguistics, pp.333–340.

Gamallo, P., Alegria, I., Pichel, J.R. and Agirrezabal, M. (2016) 'Comparing two basic methods for discriminating between similar languages and varieties', in *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*, pp.170–177.

Goweder, A. and De Roeck, A. (2001) 'Assessment of a significant Arabic corpus', in *Arabic NLP Workshop*, ACL/EACL.

Graff, D. and Walker, K. (2001) 'Arabic newswire part 1', *Linguistic Data Consortium*, Philadelphia, LDC Catalog number LDC2001T55 and ISBN, 1-58563.

Gupta, P., Bali, K., Banchs, R.E., Choudhury, M. and Rosso, P. (2014) 'Query expansion for mixed-script information retrieval', in *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, ACM, pp.677–686.

Helmy, M., Basaldella, M., Maddalena, E., Mizzaro, S. and Demartini, G. (2016) 'Towards building a standard dataset for Arabic keyphrase extraction evaluation', in *2016 International Conference on Asian Language Processing (IALP)*, IEEE, pp.26–29.

Hmeidi, I., Kanaan, G. and Evens, M. (1997) 'Design and implementation of automatic indexing for information retrieval with Arabic documents', *JASIS*, Vol. 48, No. 10, pp.867–881.

Kiritchenko, S. and Mohammad, S.M. (2016) 'Capturing reliable fine-grained sentiment associations by crowdsourcing and best-worst scaling', in *HLTNAACL*, pp.811–817.

McEnery, T., Xiao, R. and Tono, Y. (2006) *Corpus-based Language Studies: An Advanced Resource Book*, Taylor & Francis, Routledge, London; New York.

Mike, S. (2014) *Type/token Ratios and the Standardised Type/Token Ratio* [online] http://www.lexically.net/downloads/version5/HTML/index.html?type_token_ratio_proc.htm (accessed 15 November 2017).

Mustafa, A.M. (2013) *Mixed-language Arabic-English Information Retrieval*, PhD thesis, University of Cape Town.

Mustafa, M. and Suleman, H. (2015) 'Mixed language Arabic-English information retrieval', in *Computational Linguistics and Intelligent Text Processing*, Springer, pp.427–447.

Nie, J-Y. (2010) 'Cross-language information retrieval', *Synthesis Lectures on Human Language Technologies*, Vol. 3, No. 1, pp.1–125.

Rieh, H-y. and Rieh, S.Y. (2005) 'Web searching across languages: preference and behavior of bilingual academic users in Korea', *Library & Information Science Research*, Vol. 27, No. 2, pp.249–263.

Su, R., Shi, S., Zhao, M. and Huang, H. (2017) 'Utilizing crowdsourcing for the construction of Chinese-Mongolian speech corpus with evaluation mechanism', in *International Conference of Pioneering Computer Scientists, Engineers and Educators*, Springer, pp.55–65.