PROCESSING LARGE ARABIC TEXT CORPORA: PRELIMINARY ANALYSIS AND RESULTS

Fahad A. Alotaiby

Department of Electrical Engineering,
College of Engineering,
King Saud University
P.O. Box 800 Riyadh 11421
Saudi Arabia
falotaiby@hotmail.com

Ibrahim A. Alkharashi

Computer and Electrical Research
Institute, King Abdulaziz City for
Science and Technology
P.O. Box 6086 Riyadh 11442
Saudi Arabia
kharashi@kacst.edu.sa

Salah G. Foda

Department of Electrical Engineering,
College of Engineering,
King Saud University
P.O. Box 800 Riyadh 11421
Saudi Arabia
sfoda@ksu.edu.sa

Abstract

Important research areas such as Automatic Speech Recognition (ASR), Optical Character Recognition (OCR) and Information Retrieval (IR) heavily depend on the presence of a good statistical representation of the used language. A more precise representation leads to more accurate systems. On the other hand, Arabic is a quite richer and more complex language than English. This raises the need to study the key statistics of Arabic language and the statistical differences between Arabic and English on a large scale. For the purpose of this study, two large and comprehensive Arabic and English corpora are used. They are "Arabic Gigaword Third Edition" (Graff, 2007) and "English Gigaword Third Edition" (Graff *et al.*, 2007), respectively. In this paper, we are going to use these two corpora to perform our preliminary analysis and show the results for Arabic language in conjunction with English. The aim of this paper is to present statistics about token and paragraph length distribution, punctuation marks and unigrams for Arabic and English. Preliminary processing considerations and issues are discussed throughout the paper.

Introduction

One of the first steps of processing any text corpora is to divide the input text into proper units. These units could be characters, words, numbers, sentences or any other appropriate unit. The definition of a word here is not the exact syntactic form, that is why we call it a 'token'. A token could refer to a syntactic word, a number or, as in Arabic, a whole grammatical phrase (e.g. وسنساعدهم "and we shall help them"). The process of extracting tokens is called tokenization (Attia, 2008; Lee et al, 2003). The simplest way used in tokenization is extracting any alphanumeric string between two white spaces, which will be used in this research. Moreover, finding out the boundary of a sentence automatically is not a simple task. It is important to detect the shortest complete-sentence length, particularly in areas like automatic parsing or language modeling (Diab, Hacioglu, & Jurafsky, 2004).

In English, considerations and issues regarding these research areas have been well studied. Unfortunately, some researchers have not paid good attention to the special characteristics of the Arabic language. For example, in morphologically rich languages, such as Arabic, the Out-Of-Vocabulary problem is worse (Heintz, 2008). To overcome the problem of richness of Arabic language, many researchers provided different algorithm for stemming (removing prefixes and suffixes) (Kadri & Nie 2006; Majdi & Eric 2008; Rogati, McCarley & Yang 2003).

Studying major statistical differences and similarities between Arabic and English languages would provide a good assistance when processing Arabic, especially if it is done on a large scale. A comprehensive statistical study and comparison between English and Chinese has been introduced in (Yang *et al.*, 2007) based on a corpus of 100 million web pages for each language. In contrast, Arabic language has been studied on a relatively small scale. Al-Kadi (1996) presented a statistical study of frequencies of Arabic language based on a 700,000 word. With the availability of large corpora and machines with greater processing capability, it is now easier to discover the statistical relations between Arabic and English languages.

Arabic Language

Arabic language is one of the oldest languages that are still widely used. It is a Semitic language and is written from right to left (Seikaly, 2007). Ancient Arabic writing system was originally consonantal. Every letter in the 28 Arabic alphabets represents a single consonant. Late in the seventh century, "Abu Al-Aswad Al-Du'ali" invented the Arabic diacritics, which are graphical signs that distinguish the different pronunciations of consonants. Short vowels are indicated by diacritics, but they are very often omitted from modern written text. Arab readers could differentiate between word with the same writing form (homographs) by the context of the script (Alotaiby, 2002).

Processed Corpora

Arabic Gigaword (Graff, 2007) and English Gigaword (Graff *et al.*, 2007) are archives of newswire text data from Arabic and English news sources that have been collected over several years by the Linguistic Data Consortium (LDC) at the University of Pennsylvania. After preprocessing both corpora, the Arabic Gigaword corpus was found to have two

million documents with nearly 600 million tokens, while the English Gigaword was found to have 7,150,000 documents with three billion tokens. To be consistent, only 600 million tokens are extracted from random English documents. Sources of Arabic Gigaword can be categorized into two classes. The first consists of articles from four newspapers and the second consists of newswire from two press agencies. Articles extracted from press agencies are edited with less care, since there will be more editing prior to publishing.

Characters and Punctuation Marks

One of the most useful features in detecting sentences boundaries and tokens is punctuation marks. They were introduced to the Arabic writing system for the first time in 1912 by an Egyptian linguist named "Ahmed Ali Pasha". For centuries, Arabic text had been written without punctuation marks or paragraphing. Therefore, punctuations use is not consistent in Arabic language typography. In fact, punctuation marks are considered by some as redundant cosmetic marks.

After analyzing both corpora, it is remarkable that Arabic documents have inconsistent way of using punctuation marks and symbols. Actually, total number of punctuation marks and symbols used in Arabic corpus was 134, while in the corresponding English corpus only 54 punctuations and symbols were used. Besides, many Arabic characters used in Farsi and Urdu languages were included in some documents.

Table 1 shows only the 15 most frequent punctuation marks and symbols. Because of the peculiar way of using punctuation marks in Arabic script, there were many difficulties in processing Arabic corpus. For instance, in the Arabic corpus two space characters were used (0x0020) and (0x00A0). In addition, different characters have been used for comma, quotation, question mark and period. In some documents, number zero (٠) has been used as a period. Even worse, in many documents space character has been omitted between words that ends with graphically non-connecting characters as in (ومكافحةانتشارالاسلحةالنوويةمشيراً) which is a phrase of five connecting words.

Unigrams in Arabic and English

Unigrams represent the frequency in which a certain token has been written or uttered. The question is "Does the Arabic language have more unigrams than English?" On one hand, Arabic morphology is far more complex and richer than English; which may increase number of unigrams. On the other hand, homographs (i.e. a single token representing different words) are more frequent in the Arabic language especially with the absence of diacritics; which may decrease number of unigrams.

To show the difference between the number of "word tokens" and the number of "word types", the word 'في' appeared in the Arabic corpus 21,141,537 times as a word token, but it is counted one time as a word type.

In the 600 million tokens Arabic and English corpora, the total number of word types in the Arabic corpus is 2,207,637 word types, while in the English corpus it is 1,257,112 word types.

To get close to the statistical differences in the unigram level of the corpora, the two corpora were processed at different counts of word tokens ($n10^m$; $1 \le n \le 9$, $1 \le m \le 8$) where the total number does not exceed 600 million. The number of word types was produced for every set of word tokens, and this was done for both corpora. Figure 1 shows the difference in the number of word types of both Arabic and English languages.

This shows that the total number of Arabic word types needed in any application is more than the number of English word types needed for the same application. To see the difference in numbers, Figure 2 shows the ratio of the Arabic word types to the English word types in the used corpora. The average ratio is 1.76. This means that to cover the same linguistic content that has been covered using 50,000 word types in English language; one may need about 88,000 word types in Arabic language. This result is a very important result for research areas that might need a lexicon such as speech recognition, optical character recognition, automatic translation and automatic headline generation.
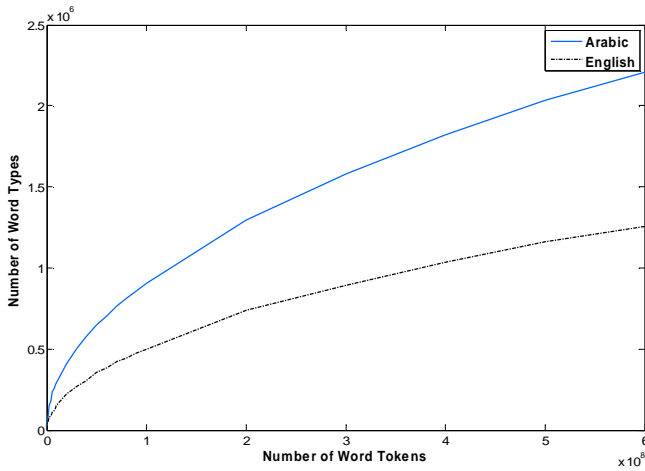
Arabic			English		
Frequency	(Hex)	Ch	Frequency	(Hex)	Ch
2,414,627	002F	/	848,843	0009	TAB
2,708,135	003A	:	1,664,305	003B	;
3,187,882	00AB	«	1,715,585	005F	_
3,199,269	00BB	»	2,230,403	003A	:
3,541,866	0028	(	3,201,203	0029	)
3,546,591	0029	)	3,210,434	0028	(
4,120,782	002D	-	4,031,581	0022	"
6,068,738	064B	ً	7,726,127	0060	`
6,701,103	0640	ـ	14,569,156	002D	-
7,712,759	0022	"	16,393,719	000A	LF
15,521,936	000A	LF	16,393,719	000D	CR
15,521,936	000D	CR	16,901,500	0027	'
17,402,128	060C	،	33,540,565	002C	,
25,683,122	002E	.	34,753,866	002E	.
552,229,058	0020	SP	544,581,507	0020	SP

Table 1: The 15 most frequent punctuation marks and symbols in Arabic and English corpora.

Figure 1: Number of word types in the Arabic and English corpora.
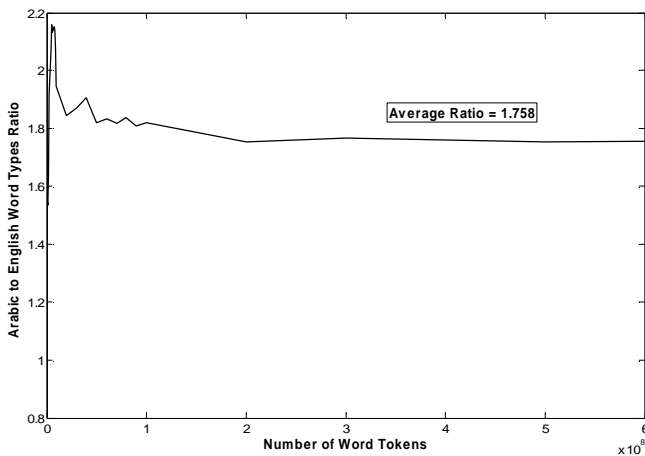
Figure 2: Ratio of Arabic to English word types.

of the word type to its rank, and this number roughly reflects the richness of the language (Manning & Schütze, 1999). Using the Arabic and English corpora, k_{Ar} = 4,300,300 for Arabic language and k_{En} = 2,426,600 for English language (calculated in the stable area from 10^3 to 10^5 only). Note that k_{Ar}/ k_{En} = 1.77 which almost equals the ratio calculated above.

| Arabic			English		
Token	Frequency	(%)	Token	Frequency	(%)
في	21,141,537	3.52	the	30,126,524	5.02
من	12,763,354	2.13	to	14,682,631	2.45
ان	8,912,777	1.49	of	14,659,938	2.44
على	8,670,263	1.45	and	12,638,594	2.11
الى	7,285,936	1.21	a	12,322,616	2.05
التي	3,437,140	0.57	in	11,481,175	1.91
عن	3,236,308	0.54	that	5,264,798	0.88
الذي	2,517,443	0.42	for	5,227,499	0.87
مع	2,504,989	0.42	said	4,705,064	0.78
فى	2,225,489	0.37	The	4,534,766	0.76
لا	2,140,904	0.36	on	4,397,759	0.73
ما	2,041,200	0.34	is	4,316,691	0.72
هذا	1,957,484	0.33	with	3,527,605	0.59
هذه	1,831,502	0.31	was	3,448,414	0.57
بين	1,722,750	0.29	at	3,051,347	0.51
اليوم	1,563,204	0.26	by	2,897,787	0.48
بعد	1,505,344	0.25	as	2,841,353	0.47
لم	1,233,767	0.21	from	2,593,827	0.43
كان	1,198,945	0.20	he	2,459,732	0.41
أن	1,119,355	0.19	it	2,416,102	0.40

Table 2: Most frequent 20 word types in the Arabic and English corpora.
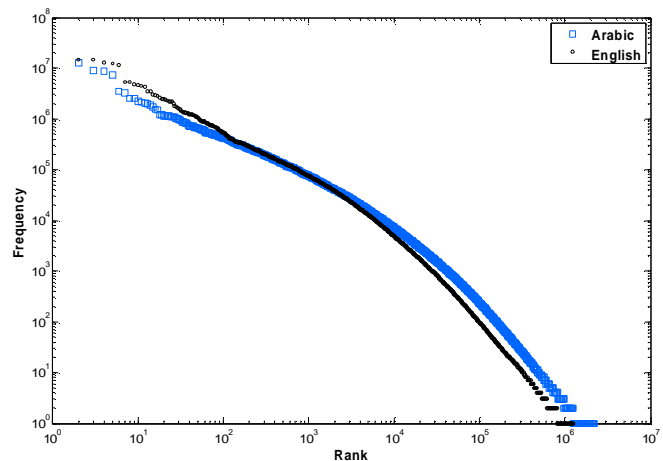
Figure 3: Zipf's law, the rank of the word type versus frequency of correspoding word using a logarithmic scale in bothe axes

As a result, the unigram language models of every corpus were produced. Unigram language model represents the probability of occurrence of every word. The procedure of producing such a model consumes a significant amount of time and effort. Furthermore, browsing such file as a whole is not applicable due to its huge size (6.4 GB)

Table 2 shows the most frequent 30 word types in both corpora and their frequencies and percentages of appearance. It is notable that the majority of the most frequent words are prepositions and they have no direct relation to the idea of the document. However, they play a large role in binding words together. Furthermore, spelling errors are common in the Arabic corpus to the extent that the miss-spelled "ان" and "الى" is more frequent than the correct "أن" and "إلى" respectively.

In contrast, Figure 3 and Figure 4 show frequency of words versus their ranks, and they are ordered in a descending direction (the most frequent word type is in rank 1). Zipf's law says that there is a constant k that relates the frequency
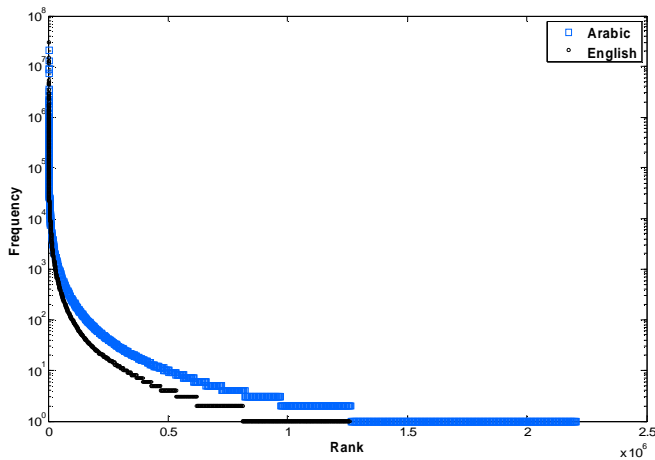
Figure 4: Zipf's law, the rank of the word type versus frequency of corresponding word using a logarithmic scale in the Y axis only.

One of the important factors that the Arabic language is richer than the English language is the large amount of clitics (proclitics and enclitics). A clitic is an element that can be added to a word to construct another word and has a restricted syntactic distribution like "They're" and "don't". In fact, in Arabic language a single word token may represent a whole sentence like "ودرسناها" "and we studied it". For instance, in a smaller version Arabic corpus there are more than 110,000 word types (in a 28,239,779 word corpus) that start with the letter "و". In most of them, the letter "و" represents the word "and".

Word and Paragraph Length Distributions

Every paragraph in the Arabic Gigaword and English Gigaword corpora is marked up with a tag. Using the paragraph tag to divide document into paragraph yielded 17,298,414 paragraphs in the Arabic corpus, and 17,657,120 paragraphs in the English corpus. Figure 5 shows the paragraph length distribution in both Arabic and English corpora. The two humps appearing in Figure 5 could be a result of having two distinguishable groups of long and short documents in the corpora.

It is remarkable that some documents in the Arabic corpus consist of one long paragraph with a single period at the end. Obviously, these very long paragraphs need a special attention.

On the other hand, Figure 6 shows the word length distribution (in character) in the Arabic and English corpora. The average word length in Arabic is 5, while it is 3 in English. However, Arabic Gigaword contains a lot of connected Arabic words causing long tokens.
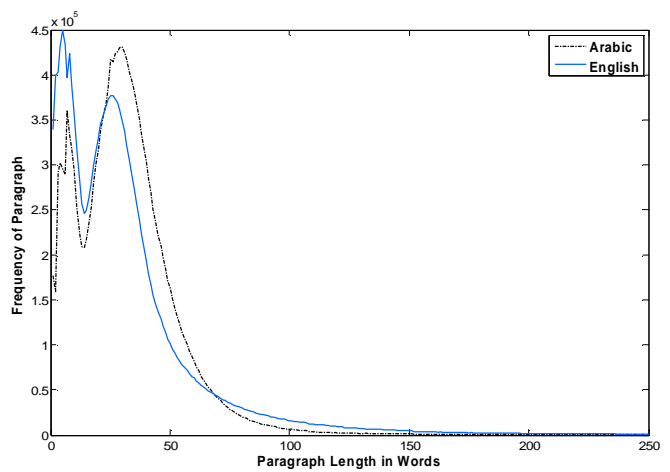

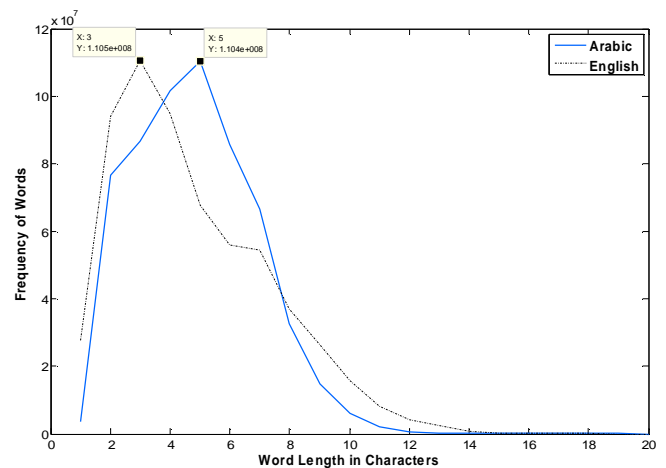Figure 5: Paragraph length distribution in the Arabic and English corpora.

Figure 6: Word length distribution in Arabic and English corpora.

Conclusions

In this paper basic statistical differences between Arabic and English languages have been presented on a large scale. Results have been presented by utilizing Arabic and English Gigaword. The unigrams of 600 million words in Arabic and English languages have been produced. It has been shown that the number of Arabic word types is 76% more than in English. This can be explained in the sense that Arabic is statically a richer language than English. Also, statistical distributions of the word length and paragraph length have been presented for Arabic and English corpora.

On the other hand, Arabic documents suffers miss-spellings, bad usage of punctuation marks and careless organization. Therefore, preprocessing might be an important step before using it.

Acknowledgements

References

Al-Kadi, I. (1996). Study of Information-theoretic Properties of Arabic Based on Word Entropy and Zipf's Law. Journals of King Saud University, (10), 1--14.

Alotaiby, F. (2002), Speaker-Dependant Continuous Arabic Speech Recognition. Master Thesis in Engineering, King Saud University.

Attia, M. (2007). Arabic Tokenization System. ACL-Workshop on Computational Approaches to Semitic Languages. Prague, Czech Republic.

Diab, M., Hacioglu, K. & Jurafsky, D. (2004). Automatic Tagging of Arabic Text: From Raw Text to Base Phrase Chunks. In Proceedings of Human Language Technology-North American Association for Computational Linguistics.

Graff, D. (2007). Arabic Gigaword Third Edition. Linguistic Data Consortium. Philadelphia, USA.

Graff, D., Kong, J., Chen, K. &Maeda, K. (2007). English Gigaword Third Edition. Linguistic Data Consortium. Philadelphia, USA.

Heintz, I. (2008). Arabic Language Modeling with Finite State Transducers. In Proceedings of the ACL-08: HLT Student Research Workshop, Companion Volume (pp. 37—42). Columbus.

Kadri, Y., & Nie, J. Y. (2006). Effective Stemming for Arabic Information Retrieval. In Proceedings of the challenge of Arabic for NLP/MT Conference. The British Computer Society. London, UK.

Lee, Y.S., Papineni. K., Roukos., S., Emam, O. & Hassan, H. (2003). Language Model Based Arabic Word Segmentation. In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (pp. 399—406). Sapporo, Japan. Association for Computational Linguistics.

Majdi, S. & Eric, A. (2008). Comparative Evaluation of Arabic Language Morphological Analysers and Stemmers. in Proceedings of COLING 2008 22nd International Conference on Computational Linguistics. Manchester, UK.

Manning, C.D. & Schütze, H. (1999). Foundations of Statistical Natural Language Processing. The MIT Press.

Rogati, M., McCarley, S. & Yang, Y. (2003). Unsupervised Learning of Arabic Stemming using a Parallel Corpus. In Proceedings of the second workshop on Analytics for noisy unstructured text data (pp. 113--118). Singapore. ACM.

Seikaly, Z.A. (2007). The Arabic Language: The Glue that Binds the Arab World. AMIDEAST, America-Mideast Educational and Training Services, Inc. http://www.amideast.org/publications/arabic-language.pdf

Yang, S., Zhu, H., Apostoli, A. & Cao, P. (2007). N-gram Statistics in English and Chinese: Similarities and Differences. In Proceedings of IEEE International Conference on Semantic Computing (pp. 454--460). Irvine, California, USA.