Assessment of a Significant Arabic Corpus

Abduelbaset Goweder
Department of Computer Science
University of Essex
Wivenhoe Park, Colchester CO4 3SQ
England
agowed@essex.ac.uk

Anne De Roeck
Computing Department
The Open University
Walton Hall, Milton Keynes MK7 6AA
England
A.DeRoeck@open.ac.uk

Abstract

The development of Language Engineering and Information Retrieval applications for Arabic require availability of sizeable, reliable corpora of modern Arabic text. These are not routinely available. This paper describes how we constructed an 18.5 million word corpus from Al-Hayat newspaper text, with articles tagged as belonging to one of 7 domains. We outline the profile of the data and how we assessed its representativeness.

The literature suggests that the statistical profile of Arabic text is significantly different from that of English in ways that might affect the applicability of standard techniques. The corpus allowed us to verify a collection of experiments which had, so far, only been conducted on small, manually collected datasets. We draw some comparisons with English and conclude that there is evidence that Arabic data is much sparser than English for the same data size.

1 Introduction

There is some evidence that statistical NLP techniques for information retrieval (IR) on European languages do not transfer well to Arabic because of the nature of the language and its writing system (Yahya 1989; Hmeidi et al 1997, De Roeck and Al-Fares 2000). Reasons include its right to left orientation, the diacritization of scripts, omission of vowels and its morphological structure. However, we do not actually know how well these techniques will transfer to Arabic, because experimentation in Arabic language environments has been very new and limited compared to work on English and other European or Asian languages. The main difficulty in collecting evidence is the lack of reliable data for experimentation. Because corpora are hard to find, no extensive findings have been published to confirm that Arabic is problematic for standard techniques.

Most researchers working in ANLP construct their own datasets. For instance, Hmeidi et al (1997) constructed a corpus of 242 abstracts collected from the proceedings of the Saudi Arabian national conference. This is no exception. Datasets are usually small, collected manually and are rarely, if ever, investigated for quality or balance, so it is unclear how well experimental findings would scale up.

What is required are experiments on a very large and representative dataset of common every day Arabic. A corpus of non-sanitised Arabic newspaper text would go some way to allow conclusions on how easily NLP techniques might transfer to Arabic, and confirm or deny the results of previous experiments performed on very small datasets. In order to achieve this, we sought to repeat some standard experiments which highlight problems with Arabic on a significant dataset. As a first step, we describe the way in which we built the corpus, and how we ascertained its quality.

2 Description of Original Dataset

Newspaper text is the most accessible source of modern Arabic. We purchased an original source from Al-Hayat newspaper. The dataset is an electronic archive for the newspaper of the year 1998, presented as windows HTML files.

The collection contains roughly 42,591 articles covering 7 subject categories : (General, News, Economic, Sports, Computers and Internet, Science and Technology, Cars and Business). Each article is saved as a separate file in one of 15 folders. The dataset comes with a search utility for accessing articles by author, issue number, page number, date, country, and/or subject category.

The dataset was not useful for our research in this form, because of the presence of mark-up and code words. The following section will describe the main procedures undertaken to process the dataset into a suitable form.

3 Preliminary Processing

In the first step, a C-program classified the files into subject categories according to the Al-Hayat classification. Seven large HTML files were obtained, one for each domain. The second step converted the HTML files into TEXT format, so they can be used as input data for the Cambridge Toolkit software package. The essential goal of using Cambridge Toolkit is to produce word frequencies which will help us assess our data. The third step was to remove numbers, punctuation, and special symbols to clean up our data. The size of the total file is 268MB. The final dataset comprises 18,639,264 words in 42,591 articles (Table 2).

4 Sampling the Corpus

Following this, we investigated the contents of the dataset, to verify a number of claims made concerning potential problems with Arabic orthography and spelling (Ali 1988; Hmeidi et al 1997; Saliba and Al-Dannan 1989; Omar 1984). Lists of word frequencies were generated by executing the Cambridge Toolkit software package. These we produced for the seven domain files plus the whole-data file.

In the literature, a number of phenomena are predicted, on which standard techniques may differ in their effects from English. We list them in turn and investigate whether they indeed occur in freeflowing newspaper text.

4.1 Common Misspellings

Arabic spelling is quite complex and mistakes are common. We found examples of routine errors such as the preposition “on”, which can be misspelled in Arabic as علـي (pronounced ‘*ali*’) instead of علـى (pronounced ‘*ala*’). Another example is the conjunction *or* occurring as إو (‘*eo*’) instead of او (‘*ao*’) or أو (‘*ao*’). Many of these examples involve erroneous omission or hypercorrection of hamza ء.

4.2 Spelling Conventions

The data contains the usual variation between the proper name “Ali”علـى (‘*ali*’) and علـي (‘*ali*’).

It did not include occurrences of the link character (Arabic Tatweel), which may be inserted for cosmetic purposes and whose effect might be, for instance, to separate instances of the word “document” وثـــيقــة (‘*wathiqa-ton*’) and وثـــيقـــــــــــة (‘*wathiqa-ton*’).

4.3 Word Choice

The corpus includes examples where alternative but root-related words split frequencies. For example, the word “University” could be written as جامعه (‘*jamiah*’) or جامعة (‘*jamia-ton*’).

4.4 Joining Words

It is common practise to join function words. For instance, the conjunction “and” و (‘*wow*’) as well as several common prepositions can be combined with the next word without space (Omar 1984; Saliba and Al-Dannan 1989). For example, the phrase “Computers and Telecommunications” will usually be written in Arabic as الحاســبات والاتصـــــالات (‘*al-haseebat wa al-etesalat*’), with the conjunction و combined with the word “Telecommunications” instead of writing it as a separate token.

By examining the results obtained by running the Cambridge Toolkit on the dataset, we can conclude that most of the potential problems listed above are real and present in our data. Table 1 summarises a sample of the results and gives an indication of phenomena distribution.

Two phenomena did not occur against expectation. Tatweel, a character used for cosmetic purposes, might interfere with frequency counts. Also, modern Arabic tolerates vowelisation if the writer needs to clarify meaning. No vowelisation was found in the dataset.

Table 1. Phenomena Distributions.

ARABIC TERM	TERM FREQ.	ENGLISH TRANSLATION
علـى	281509	The preposition ‘On’ or the proper name ‘Ali’
علـي	7488	
جامعة	3258	University
جامعه	17	
وثـيقـة	989	Document
وثيقــــــــــة	0	
او	25253	The conjunction ‘or’
أو	40087	
آو	2	
إو	3	
الاتصــــالات	2407	Telecommunications
والاتصــــالات	429	and Telecommunications
الإتصــــالات	68	Telecommunications
والإتصــــالات	13	and Telecommunications
الحـاسـبات	27	Computers
و	32606	The conjunction ‘and’
فــي	734621	in
فــى	787	
وفــي	30403	and in
وفــى	31	
الــى	227233	to
الــي	730	
إلــي	308	
ألــي	0	
إلــى	35806	
ألــى	2	

5 Assessment of the Corpus

McEnery and Wilson (2001) describe a modern corpus as any collection of more than one text with four main characteristics: sampling and representativeness, finite size, machine-readable form, and status as standard reference. Two of these are present *a-priori* in our dataset : it is of finite size and machine readable. Clearly it is not yet a standard reference. Our main concern will be assessing sampling and representativeness.

5.1 Sampling and Representativeness

The corpus we have consists entirely of newspaper text, and so cannot claim status as a general model of Arabic text type or style type. However, in the absence of a better alternative, there are some arguments why this dataset is a reasonably representative sample of modern Arabic. First of all, it consists of real text in every day use, which has undergone a minimum of processing. Secondlly, it covers a range of topic areas, each with a credible size of data (Table 2). Also it was written by many authors from a variety of backgrounds and contains pieces of different types (eg. editorials vs sports reports). We conducted some experiments with Zipf’s distribution as a general indicator of quality.

5.2 Basic Diagnostic Tests

Zipf’s Law is useful as a rough description of the frequency distribution of words in human languages (Manning and Schutze, 1999). Set against Zipf’s Law, frequency distribution in an actual dataset is also a reasonable way to gauge data sparseness, and can provide evidence of imbalance in a sample.

Zipf’s Law draws a relationship between the frequency of a word f and its position in the list, known as its rank r (Manning and Schutze, 1999). The law states that: $r.f = c$, where r is the rank of a word, f is the frequency of occurrence of the word, and c is a constant that depends on the text being analysed.

The Cambridge Toolkit was run on all the domains separately, and on the whole dataset, to generate word frequencies. For comparison, we also created a small, 2000 word file from a random selection of articles. In all, nine lists of word frequencies were created. Each was sorted in descending order of frequency. Rank was assigned and the sorted lists were plotted against rank using MATLAB. Table 2 is a summary of all the data used in our experiments. Figures 1 to 9 show the results of the plots on logarithmic scale.

According to Zipf’s Law, for a representative sample the graphs should be a straight line with slope –1. In practise, this may not be the case because many words will have the same frequency but be assigned different rank. As expected, graphs improved as the size of data

Table 2. Sumary of the Dataset.

FILE NAME	SIZE IN MB	NUMBER OF WORDS	NUMBER OF DISTINCT WORDS	WORDS REPEATED 5 TIMES OR LESS
Small file	23.1 KB	2,000	1,179	96.27%
Cars	1.16	86,191	16,279	85.57%
Science	1.76	105,727	23,240	87.20%
Computers	2.2	161,146	26,029	84.24%
Sports	21.9	1,371,035	85,893	77.00%
Economic	39.4	2,434,204	104,135	73.79%
News	89.6	5,714,731	169,381	71.39%
General	111	8,765,930	354,184	74.12%
The whole dataset	268	18,639,264	444,761	71.61%

increased, and the proportion of rare words (frequency ≤ 5) declined. The graphs were skewed at high and low frequencies, and are more distorted at low than at high or middle frequencies, because a large proportion of the words in our data are quite rare. This is the problem of data sparseness which we will discuss in the next section.

The analysis of the graphs (Figures 1 to 9) and the results listed in Table 2 show that there is no reason to believe that the dataset is imbalanced, either overall, or for each subject area.

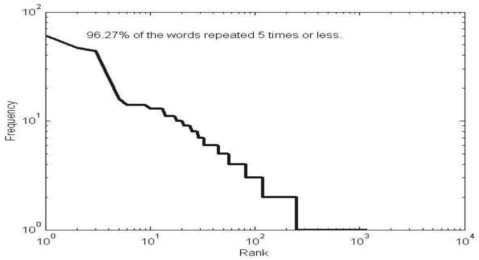

Figure 1. Word Frequency versus Rank: Small File.

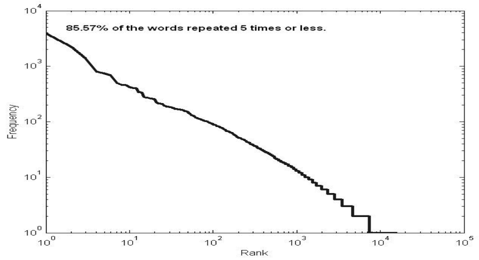

Figure 2. Word Frequency versus Rank: Cars and Buisness.

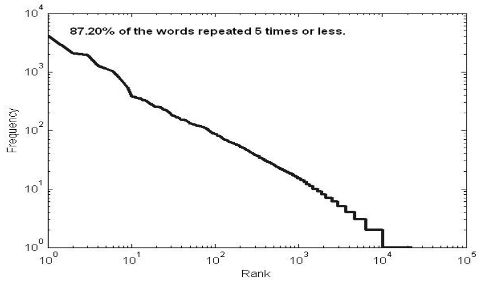

Figure 3. Word Frequency versus Rank: Science and Technology .

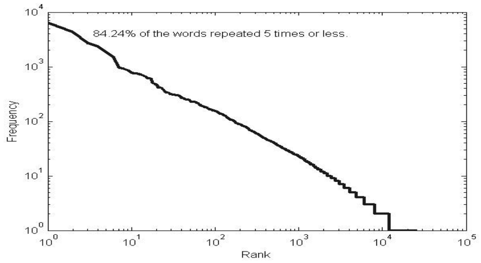

Figure 4. Word Frequency versus Rank: Computers and Internet.

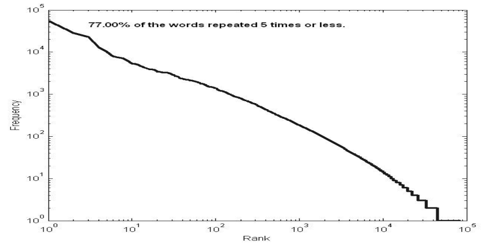

Figure 5. Word Frequency versus Rank: Sports.

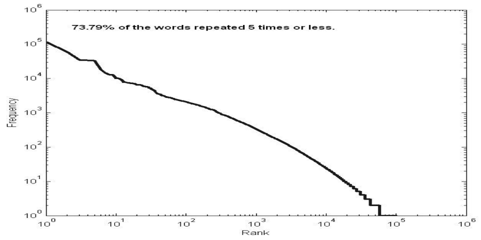

Figure 6. Word Frequency versus Rank: Economy.

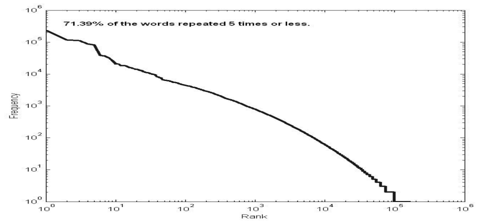

Figure 7. Word Frequency versus Rank: News.

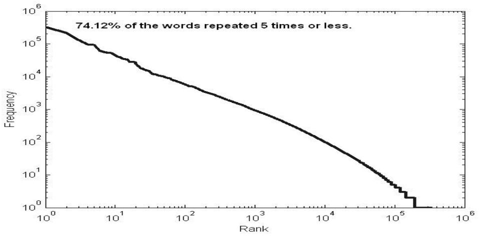

Figure 8. Word Frequency versus Rank: General.

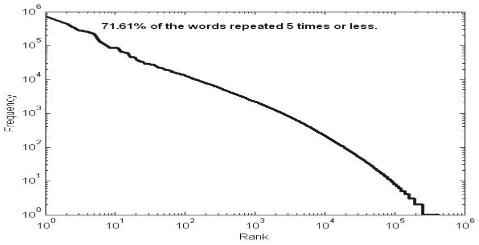

Figure 9. Word Frequency versus Rank: Whole Dataset.

6 Experimenting on Sparseness

6.1 The Issue of Sparseness

There is some evidence that standard NLP methodologies are significantly hampered by Arabic morphology (Hmeidi et al 1997). Since Arabic is very rich in both vocabulary and morphological variation, it follows that any particular word will appear less often than in English for a given text length and type. This hypothesis predicts that Arabic datasets will have a higher degree of inherent sparseness than comparable English counterparts.

Sparseness is an important factor in statistical NLP. In general, it is taken to mean that almost all words in a corpus are rare or infrequent (Manning and Schutze 1999). In the specific, sparseness is also taken to mean that some quite common words, or "reasonable" n-grams, are absent from a particular dataset (Jurafsky & Martin 2000, pp. 206). The associated technical difficulties (eg ensuing null probabilities) have been addressed in a substantial literature on smoothing techniques.

In the context of this paper, the issue of sparseness is of particular interest for two reasons.

The first reason is standard, in that sparseness is related to dataset quality, partly through its relation to size. Although endemic in any NL dataset, sparseness problems become less acute as the dataset gets larger. This is easy to demonstrate. Assuming a balanced, 20M word corpus, and a randomly selected 2000 word subset (see Figure 1 and Table 2) thereof, the latter will be much sparser than the former, and hence less useful as a representative sample of the language. Sparseness metrics can also reflect quality in a more immediate way. Again this is easy to demonstrate with an extreme situation. A dataset consisting of the last 10 versions of the Home Office Internal Telephone and E-mail directory will show comparatively few infrequent words. In other words, given a balanced, representative corpus of reasonable size, we should expect to find a substantial degree of sparseness.

However, the investigation of sparseness in an Arabic dataset may show up additional problems. The second reason why sparseness is an issue here, is that it has been suggested in the literature that some languages may well be inherently sparser than others, in the sense that datasets which are comparable in size and type may be much sparser in one language than in another. Lee et al (2000), for instance, assume this is the case for Korean. For Arabic, an indirect, but related argument has been mounted by Yahya (1989).

Table 3. The Token to Type Ratio.

LENGTH OF TEXT	ARABIC DISTINCT WORDS	ARABIC RATIO	ENGLISH DISTINCT WORDS	ENGLISH RATIO
100	84	1.190	69	1.449
200	149	1.342	124	1.613
400	281	1.423	165	2.424
800	507	1.578	328	2.439
1,600	902	1.774	621	2.576
3,200	1,537	2.082	871	3.674
6,400	2,715	2.357	1,361	4.702
12,800	4,895	2.615	2,337	5.477
16,000	5,775	2.771	2,699	5.928
20,000	6,956	2.875	3,154	6.341

6.2 Yahya’s Experiment

Yahya (1989) conducted a series of word prediction experiments on vowelised and unvowelised Arabic text, up to a maximum of 20,000 words (Table 3). He showed, for texts of variant length, that Arabic behaves differently from English with respect to word occurrence patterns. He measured the token to type ratio, which can be obtained by dividing the number of tokens (text length) by the number of distinct words. Table 3, adapted from his work, summarises his findings. He showed that the English token to type ratio is significantly higher than the Arabic one for the same text length. This implies that English words are repeated more often than Arabic ones for the same text length. This is perhaps to be expected given the comparative morphological complexity of the two languages, but the finding invites the conclusion that Arabic textual data may be inherently sparser than English, for similar text types and sizes. Given the discrepancy between Yahya's Arabic and English ratios, the result may well be sufficiently significant to impact on a range of statistical applications, particularly since most statistical techniques have been developed and tested on English which is a language almost entirely lacking in morphology.

6.3 Repeating Yahya’s Experiment

Yahya’s experiment included text sizes up to 20,000 words, and the differential in ratios may be due to the sample. In order to verify whether the type to token ratio evens out with sample size or type, we conducted an experiment repeating and extending Yahya’s using our dataset. The results are shown in Table 4.

We extracted a 20,000 word fragment from the domain *general*, and divided it into subsamples repeating Yahya (1989). We also extracted a 1M word sample from the same domain, to allow comparison with known data from the Brown Corpus (Allen, 1995). We calculated the ratio for all text in the domain *general*, and, finally, for the whole dataset.

First of all, our results confirm Yahya’s findings for the smaller samples up to 20,000 words, with very similar ratios. Secondly, it appears that the relatively higher incidence of new words for Arabic carries through to larger text samples. The English ratio of the one-million Brown corpus (Allen, 1995) approximately equals 20.408 whereas for Arabic we obtained a ratio of 8.252 for the same text length. Compare also the English ratio at 1M words with the one for Arabic at approximately 8M words. Whereas we still need to compare the ratio for the full 18M corpus with that for English texts of similar size and type, there is some indication that the rate of occurance of new words remains comparatively high in Arabic. This suggests that, for some statistical applications, Arabic datasets may need to be significantly larger than English ones for similar effect.

Table 4. The Token to Type Ratio (Repeated Experiment).

LENGTH OF TEXT	ARABIC DISTINCT WORDS	ARABIC RATIO
100	91	1.099
200	168	1.190
400	305	1.311
800	559	1.431
1,600	988	1.619
3,200	1,741	1.838
6,400	3,451	1.855
12,800	6,377	2.007
16,000	7,607	2.103
20,000	8,986	2.226
1,000,000	121,187	8.252
8,765,930	354,184	24.749
18,639,264	444,761	41.908

7 Conclusion

We collected and processed a sizable Arabic corpus consisting of sundry newspaper text. The resulting dataset is not representative of all Arabic text types and styles, but does provide a large resource of modern, unvowelised, freeflowing Arabic.

We sampled the data and confirmed the presence of a number of phenomena which the literture predicts would occur, and which may be problematic for language engineering applications. We investigated the balance of the corpus by checking Zipf distribution, over each of the sample domains as well as over the dataset as a whole. On the whole, we found no evidence to suggest that the dataset is significantly imbalanced either with respect to frequency distribution, or with respect to the range of ideosyncratic phenomena. In this sense, the corpus is useful as a background for the developments of techniques.

Against this corpus, we repeated Yahya's (1989) experiment which had been conducted on a small sample, and which suggested that Arabic datasets will be much sparser than comparable English ones. This is significant as it may affect the success of standard techniques (eg n-grams) on Arabic data. Our experiment confirmed Yahya's findings for a large dataset.

Acknowledgements

We thank the Libyan Ministry of Education for supporting this work.

References

N. Ali. 1988. *Computers and the Arabic Language.* Al-khat Publishing Press, Ta'reep, Cairo.

James Allen. 1995. *Natural Language Understanding.* The Benjamin/Cummings Publishing Company, Inc.

Anne N. De Roeck and Waleed Al-Fares. 2000. A Morphologically Sensitive Clustering Algorithm for Identifying Arabic Roots. In *Proceedings of the 38th ACL*, Hong Kong.

Ismail Hmeidi, Kanaan Ghassan and Martha Evens. 1997. Design and Implementation of Automatic Indexing for Information Retrieval with Arabic Documents. In *Journal of the American Society for Information Science.* 48(10):867-881.

D. Jurafsky and J.H. Martin. 2000. *Speech and Language Processing*, Prentice Hall, New Jersey.

S. Lee, J. Tsujii and H. Rim. 2000. Hidden Markov Model-Based Korean Part-of-Speech Tagging. In *Proceedings of the 38th ACL*, pp 376-383, Hong Kong.

Christopher D. Manning and Hinrich Schuetze. 1999. *Foundations of Statistical Natural Language Processing.* MIT Press, Cambridge Mass.

Tony McEnery and Andrew Wilson. 2001. *Corpus Linguistics.* Edinburgh University Press 2000.

A. M. Omar (1984). “Al-Naho Al-Asasy.”

B. Saliba and A. Al-Dannan. 1989. Automatic Morphological Analysis of Arabic: A study of content word analysis. In *Proceedings of the First Kuwait Computer Conference*, pp. 3-5, Kuwait.

A. H. Yahya. 1989. On the Complexity of the Initial Stages of Arabic Text Processing. Birzeit University, Birzeit, West Bank.