Yassine Benajiba, Paolo Rosso
Universidad Politécnica de Valencia, Spain

Towards a Measure for Arabic Corpora Quality

Abstract

In this paper we present a statistical measure which for the first time is used to evaluate the quality of Arabic corpora. This measure is entirely based on statistical data and language-independent. However, the values which might be obtained in the experiments could be very different for corpora written in different languages. Our experiments were conducted using Arabic corpora. We have chosen four corpora of different types in order to determine the corpus charcteristics reflected by our quality measure. The preliminary results show that the measure is significantly correlated with the writing style and the nature of the text.

***Keywords*: Arabic**, Natural Language Processing, Zipf's law, Corpus Variety, Corpus Complexity, Corpus Quality, Kullback-Leibler distance.

1. Introduction

Lately, Statistical Natural Language Processing (NLP) researches helped significantly to ease the use of internet and to automatically do tasks which were tedious and expensive to do manually. Namely search engines like Google[1] and Yahoo[2] which allow to all kinds of internet users to locate the relevant documents and web pages to their query. Automatic translators like Babel Fish,[3] Google translator,[4] WorldLingo[5] which allow users to get information from documents written in foreign languages. Also, some first steps were made to build a Question Answering (QA) system[6] able to answer the user questions written in natural language.

Both the design and evaluation of NLP systems need reliable corpora. The choice of the measures to use in order to evaluate a corpus depends mainly on the type of corpus. Nevertheless, some measures need to be taken into consideration for any kind of corpora, e.g. size, ratio of token to types, vocabulary growth, classes distribution and annotation coherence (in case of annotated corpora), and so forth. In the literature, very few are the attempts to measure the complexity, the variety and the correctness of the corpus words frequency distribution (especially Arabic corpora). In (Goweder, 2001) a test of correspondence of word frequency distribution of different corpora to Zipf's law is used to evaluate different Arabic corpora. However, the authors aimed mainly at the importance of the data sparseness factor whereas in this paper we use the same technique together with other factors in order to find out a reliable evaluation technique for Arabic corpora.

[1] http://www.google.com/
[2] http://www.yahoo.com/
[3] http://babelfish.altavista.com/
[4] http://www.googl.coml/language_tools/
[5] http://www.worldlingo.com/
[6] http://start.csail.mit.edu/

The rest of this paper is organized as follows. Section Two is devoted to present some characteristics of the Arabic language which are closely related to the topic of the paper. In the third section we introduce the Zipf's law. A detailed explanation of our evaluation technique is given in the fourth section. In Section Five we present our preliminary experiments and results, and we give further interpretion and sicussion of the obtained results in the sixth section. Finally, in the last section we draw some conclusions and the future works to be done.

2. Peculiarities of the Arabic Language

Following we present some of the characteristics of the Arabic language which need to be taken into consideration in almost every NLP task:

(i) *Short vowels*:

Nowadays, all the newspapers articles and books do not use short vowels in the text. Thus, the text becomes even more ambiguous and a robust context-based sense disambiguation technique is absolutely needed to determine the sense of a word.

(ii) *Absence capital letters*:

Such as for many of the Semitic languages, the Arabic language does not use capital letters. This absence of capital letters hardens significantly tasks like Named Entity Recognition (Benajiba, 2007).

(iii) *Very complex morphology*:

To form an Arabic word consists basically of concatenating the root morpheme with other affixes in order to obtain the desired meaning (Figure 1 shows an example of the composition of an Arabic word). This particular strategy to from the words in Arabic explains why Arabic corpora generally have a very low ratio of tokens to types. Consequently, for all the NLP tasks which require a training phase a very large training corpus is needed. In order to overcome this obstacle two solutions are possible. The first solution is to perform a light-stemming which consists of omitting from every word the affixes in order to keep only the root morpheme.

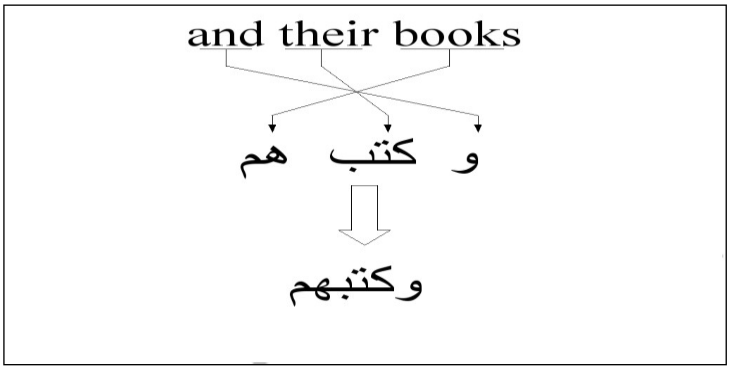


Figure 1: A simple example of the composition of an Arabic word.

This solution is very convenient for tasks where prepositions are not needed such as the Information Retrieval and Question Answering tasks where prepositions are considered stop words and are generally omitted in order to increase the systems quality. However, other NLP tasks such as Named Entity Recognition (NER), Information Extraction, Word Sense Disambiguation, etc. need to preserve the prepositions because in such tasks they represent essential contextual information. Therefore, in order to separate the root morpheme from the affixes and keep these ones in the text at the same, a very good solution is to perform a pre-processing "text-segmentation" of the text.

As we mentioned before, in this paper we describe our preliminary experiments to obtain a reliable Arabic corpora quality measure. For this reason we wanted to keep the preposition for the first experiments (see Section Five for more details).

3. Introduction to Zipf's Law

Zipf's law is an empirical law named after the Harvard linguist George Kingsley Zipf. It is based on the observation that the frequency of occurrence of some events is a function of its rank in the frequency table. This function can be expressed by the following equation:

$$freqI = C / r^{\alpha} \qquad (1)$$

where r is the rank, C is the highest observed frequency and α is a constant usually close to 1. This equation states that the most frequent event will occur twice as often as the second most frequent word. Its graphical representation in a log-log scale is a straight line with a negative slope. The different equations which might express Zipf's law are widely discussed in the literature and presenting an overview of the above goes beyond the scope of this paper and can be found in (Manning, 1999).

Moreover, many phenomena of different types proved to have that Zipf's law pattern: (i) city populations (Knudsen, 2002); (ii) internet traffic data (Adamic and Huberman, 2002); (iii) company sizes (Axtell, 2001); (iv) science citation (Osareh, 1996); (v) Finally, words frequency in natural language corpora (Manning, 1999). In this paper we are only concerned by the fifth phenomenon.

4. Our Approach to Measure Arabic Corpora Quality

In order to measure the quality level of an Arabic corpus we need to compute three factors introduced in (Makagonov, 2000): *complexity*, *variety* and *Correctness of the corpus words frequency distribution*. These three factors can be defined as follows:

(i) $Complexity = N \,.\, log(M)$

where N is the average number of characters in a word and M is the average number of words ina sentence. The complexity is a factor which is very related to the nature of the corpus.

(ii) $Variety = n \,/\, log(N)$

where n is the number of types and N is the number of tokens. This factor gives an idea about the variety of expressions is a document. It is most of all related to the author style and the nature of the document, e. g. a low variety is expected in a scientific document or any document which covers only one topic whereas a high variety is expected in a poem or a collection of newspapers articles of different topics. Other considerations should be taken in consideration for this factor, namely the language, e. g. in Arabic the use of synonyms is appreciated as a good writing style which would significantly raise the variety.

(iii) *Correctness of the corpus words frequency distribution*: In (Zipf, 1949), the author reports that a text which had Zipf's law pattern requires less effort from its reader to be understood. In order to measure this correspondence of the word frequency distribution of a text to the Zipf's law distribution we have chosen the Kullback-Leibler distance measure. This distance is asymmetric and measure the distance from a "true" probability distribution P (Zipf's law distribution) to an arbitrary probability distribution Q (word frequency distribution). Following is the definition of this measure:

$$D_{KL}[P,Q[= \Sigma_i P[i[. \log \frac{P[i[}{Q[i[} \qquad (2)$$

In the next section we give an overview on the preliminary experiments that we have carried out to evaluate the reliability of the measure we mentioned above. We also give an idea about the corpora we have used and discuss the obtained results.

5. Experiments and Results

We have used four different corpora in order to have a first idea about the order of magnitude of the three factors (see Section Four) for different kinds of corpora.

The corpora we have chosen are the following:

- *Corpus 1:* more that 66,000 words (more than 360kB) of Abu-Taïb Al-Moutanabbi poetry taken from the internet[7] (We have automatically removed all the short vocals from this corpus). This corpus was chosen to reprensent corpora with high variety of the vocabulary.
- *Corpus 2:* a collection of 111 news paper articles (more than 50,000 words. almost 260kB) of different topics.
- *Corpus 3:* A Linux Red Hat installation tutorial book (more than 55,000 words, almost 126,000kB). It was chosen in order to study the measures which might be obtained for scientific copora.
- *Corpus 4:* An extract of almost 65,000 words (more than 460kB) from a religious book of the Imam Ibnu Qayyim El Jawziyah[8].This corpora has two particularities: first it is a one-topic corpus and second it is known for its quality writing style.

[7] http://www.adab.com/

[8] http://www.almaktba.com/

Before computing the measure factors for each of the corpora we have performed a text-segmentation as a pre-processing step. For this purpose we have used Mona Diab's tokenizer (freely available on her website[9]). Table 1 shows the complexity an variety observed in our preliminary experiments. Figure 2, Figure 3, Figure 4 and Figure5 represent the words frequency distributions of Corpus 1, Corpus 2, Corpus 3 and Corpus 4, respectively, together with the Zipf's law distribution. Finally, in Table 2 we sum up the Kullback-Leibler distance computed for each corpus.

Table 1. Variety and complexity measures for four Arabic corpora.

	Variety	***Complexity***
Corpus 1	3.34	1.3
Corpus 2	2.32	6.21
Corpus 3	1.13	3.89
Corpus 4	1.74	5.18

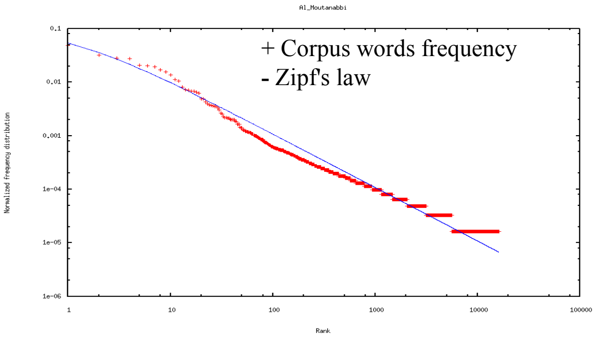
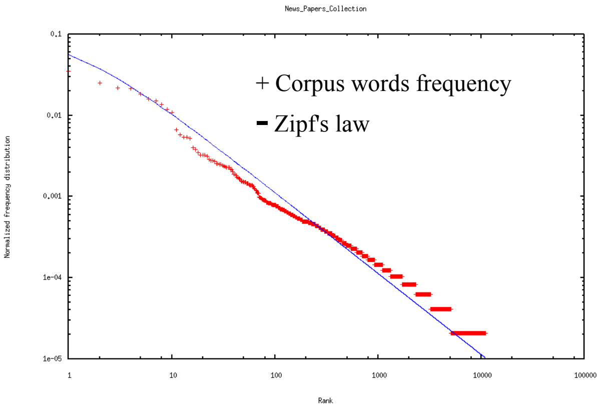


Figure 2: Word frequency distribution and Zipf's law disitribution for Corpus 1 law disitribution for Corpus 1

[9] http://www1.cs.columbia.edu/~mdiab/



ncy distribution and Zipf's law distribution for Corpus 2

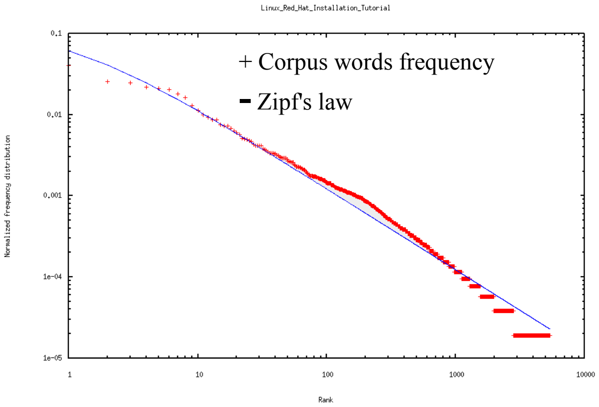


Figure 4: Word frequency distribution and Zipf's law disitribution for Corpus 3

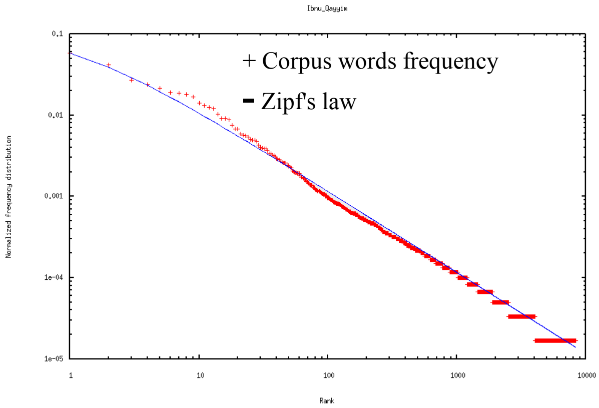


Figure 5. Word frequency distribution and Zipf's law disitribution for Corpus 4

Table 2. Kullback-Leibler measures for four corpora

	Kullback-Leibler distance
Corpus 1	22120.32
Corpus 2	32836.98
Corpus 3	41983.44
Corpus 4	28870.38

6. Interpretation and Discussion of Results

The complexity obsereved for Corpus 1 should not be taken in consideration because classical Arabic poetry is organized in columns which reduces significantly the average of words per sentence and thus reduces the complexity. However, for the rest of the corpora we can see that the complexity is very low for corpus where the content is more important than the writer style (e.g. the scientific corpus). Moreover, we can observe that the variety factor is able to show the size of the vocabulary used in a text. The lowest variety was obtained for the scientific corpus whereas the highest one was obtained for the poetry where the use of synonyms is very good appreciated. The Kullback-Leibler distance from the Zipf's law distribution to the word frequency distribution seems to be lower for corpora where the writing style is more important: i.e., the poetry corpus obtains the lowest distance whereas the scientific text obatins the highest one.

Finally, figures 2, 3, 4 and 5 show that a words frequency distribution of a corpus can be generally separated in three different zones:

(i) High frequencies zone: The study of this zone is useless because very high frequency words are normally stop words and their frequency depends more on the size of the corpus than any other characteristic;
(ii) Middle frequencies zone: The middle frequencies are the ones close to the Transition Point (Pinto, 2006). The words located in this zone represent the words which are most related to the topic of the corpus. For this reason we observe that middle frequencies are lower in Figure 2 and Figure 3 (corpora without a specific topic) than in Figure 4 and Figure 5 (corpora concerning a specific topic).
(iii) Low frequencies: The more words we have in this zone the more variety of vocabulary we have. Therefore, the width of this zone is very related to the writting style.

7. Conclusions and Future Works

We have presented in this paper the definition of Arabic corpora quality measure supported by some preliminary experiments.
This measure consists in three different factors:
(i) The complexity which reflects the size of words and sentences used in a corpus. This factor was observed to be high for corpora more focused on the content than on the writing style.
(ii) The variety which shows the size of vocabulary used in a corpus. Similarily to the previous factor, the variety of a corpus is lower for the scientific corpus.
(iii) The Kullback-Leibler distance from the Zipf's law distribution to the corpus words frequency distribution which is related to the quality of the writing style employed in the corpus. Moreover, we have shown that a graphical analysis of these distributions can reveal other characteristics of the corpus such as the unicity of the topic and the variety of the vocabulary.
In the next future, we plan to confirm these results by further experiments on bigger corpora and making automatical evaluation of this measure by annotating each corpus.

Acknowledgments
The research work of the first author was supported partially by MAEC - AECI. We would like to thank the PCI-AECI A/7067/06 and MCyT TIN2006-15265-C06-04 research projects for partially funding this work.

References

Adamic, L.A., Huberman, B.A.: 2002, Zipf's law and the internet. *Glottometrics 3, 143- 150.*

Axtell, R.L.: 2001, Zipf distribution of U.S. firm sizes. *Science 293, 1818-1820.*

Benajiba, Y., Rosso, P., Benedí-Ruiz, J.M.: 2007, ANERsys: An Arabic Named Entity Recognition System Based on Maximum Entropy. In: *Proceeding of CICLing-2007.* Lecture Notes in Computer Science 4394, Springer-Verlag, pp. 143-153.

Goweder, A. and De Roeck, A.: 2001, “Assessment of a Significant Arabic Corpus” *ACL 2001. Arabic language Processing.* pp. 73-79, 2001.

Knudsen, T.: 2001, Zipf's law for cities and beyond – the case of Denmark. American *Journal of Economics and Sociology 60, 123-146.*

Makagonov, P. and Alexandrov, M.: 1999**,** Some Statistical Characteristics for Formal Evaluation of the Quality of Text books and Manuals. In: *Computing Research*: *Selected papers.* A. Guzmán and R. Menchaca (Eds.). CIC-IPN, Mexico, 2000, pp. 99-103.

Manning, C.D., Schütze, H.: 1999, Foundations of Statistical Natural Language Processing. Cambridge, Massachusetts: *MIT Press*.

Osareh, F.: 1996, Bibliometrics, citation analysis and co-citation analysis: a review of literature **I**. *Libri 46, 149-158;* **II**: *Libri 46, 217-225.*

Pinto, D., Jimenez-Salaraz, H., Rosso, P. (2006) Clustering Abstracts of Scientific Texts Using the Transition Point Technique. In: *Proceedings of CICLing-2006*. Lecture Notes in Computer Science 3878, Springer-Verlag, pp. 536-546.

Zipf, G. K.: 1949, Human Behaviour and the Principle of Least-Effort. Cambridge MA: *Addison-Wesley*.