A Framework for Evaluating the Suitability of Non-English Corpora for Language Engineering

Avik Sarkar, Anne De Roeck

Centre for Computing Research, The Open University, Walton Hall, Milton Keynes, MK7 6AA, UK
A.Sarkar@open.ac.uk, A.DeRoeck@open.ac.uk

Abstract

In this paper we develop a framework for fast profiling and quality verification of datasets for language engineering and information retrieval research. The profiling steps consist of an initial tokenization of the corpus to produce a frequency list from which some basic statistics are derived. Manual sampling is carried out to detect obvious discrepancies. Two diagnostic tests are performed to check for sparseness related measures. The behaviour of the function words is traced to gauge homogeneity of their distribution in documents.

Introduction

The last decade has seen increased research interest in the computational treatment of non-European languages, in Language Engineering (LE) and in Information Retrieval (IR). One underlying assumption appears to be, that standard techniques and insights from statistical language processing will transfer well to other languages, in spite of the fact that these were mainly developed for English. This is not always so. Harman (1991), for instance, claims that stemming has a negligible effect on recall and precision in retrieval, but Popovic and Willett (1992) show the need to reassess this claim in the context of morphologically complex languages.

When developing corpus based language models for computationally unexplored languages, two questions arise. First, can we assume that results obtained from working with an English corpus of a particular genre and size can be replicated on a comparable corpus in a different language? What does it mean for corpora in different languages to be comparable in their effect on language modeling techniques? Second, corpora[1] play a pivotal role in developing language resources, and in bootstrapping language processing applications for computationally unexplored languages. Corpus quality is an issue, particularly where resources are scarce. Are there any cost effective and fast methods for assessing the *a-priori* suitability of a corpus for language engineering applications?

These questions could be answered by profiling data sets in different languages and comparing some of their characteristics. Developing profiles of this kind has methodological as well as practical advantages. Knowing the profile of a dataset adds a dimension to the significance of experimental results, which can be interpreted in the context of different collections. Profiles can also help researchers and developers in estimating the distance between the type of dataset used for development and evaluation of a system or technique, and the type of dataset on which it is deployed in a practical setting, even across different languages.

This paper aims to develop a framework for fast profiling datasets for language engineering purposes. We will focus on measures that are cheap to obtain, that have relevance in the context of language engineering, and that can highlight important differences between languages.

A Fast Profiling Framework

We seek to develop a methodology in the context of assessing a Bengali corpus for its suitability for language engineering applications. We chose Bengali because its fast emergence as an internet language is creating opportunities for collecting large electronic datasets, making it an ideal language for rapid modeling. Techniques successful for Bengali are also likely to be re-usable for other Indic languages, for which electronic data are not yet present to the same extent.

The methodological framework takes as a starting point Goweder and De Roeck (2001), who fast profiled an 18.6M word electronic collection of Arabic newspaper text[2]. With regard to profiling measures, we will initially repeat their sparseness experiments. In addition, we will gauge homogeneity in the distribution of very frequent terms. Both sparseness and homogeneity are known to curtail the success of language engineering applications (Charniak 1993, Rose et al 1997). They have the advantage that they can be estimated on the basis of term frequency data, which is cheap and fast to collect. We will compare the results obtained for Bengali to those of Arabic and English where possible, and draw some conclusions about the Bengali corpus, and about the methodology.

Two initial observations will help place our findings in context. First, because our interest lies in *a-priori* profiling, we have assumed that no language engineering applications, such as stemmers or morphological analyzers exist. Whereas the basic framework is compatible with more sophisticated pre-processing, all experiments reported here are conducted on raw textual data. Second, measures are based on frequency data derived from tokenized raw text, which may reflect any combination of writing or spelling conventions, including errors and local variations. Findings therefore need careful interpretation.

The Basic Approach

The basic methodology introduced in Goweder and De Roeck (2001) can be summarized as involving three steps:

(a) ***Rough profiling*** gives an overview of the size, content and coverage of the corpus, on the basis of tokenized data. Tokenization involves pre-processing, using a method that will depend on a number of factors, including the script, the intended use of the dataset and the availability (or

[1] By "corpora" we mean a collection of texts.

[2] ELDA dataset W0030 available from http://www.elda.fr

desirability) of morphological processing applications. The output of this stage is a term frequency list.
(b) ***Manual sampling*** creates an opportunity to check the tokenized dataset for obvious idiosyncrasies or to verify that certain phenomenons do occur.
(c) Finally, two ***diagnostic tests*** check the dataset for sparseness related problems through Zipf's law and type-to-token ratios.
We will repeat these steps on a Bengali corpus and extend this basic framework with genre-related measures based on the behavior of function words.

Assessing a Bengali Corpus

Bengali is one of the ten most spoken languages in the world, with almost 200 million speakers. There is a rich literature, but little is available electronically. However, online textual resources are growing and a clear need for Bengali language applications and retrieval systems is emerging. As with other languages with little prior NLP history, development of robust language based applications will require applied LE and IR research and a collection of reasonably balanced textual datasets.
For profiling, we choose the Bengali corpus from the Central Institute of Indian Languages (CIIL) which was developed as a part of the Technology Development of Indian Languages (TDIL) Programme of the Ministry of Information Technology, Government of India. We made this choice because the corpus is freely available on-line[3], it was constructed in the context of developing language applications, and, on cursory investigation, it seemed of good quality.

Rough Profiling
For tokenizing, we used plain ASCII encoding, in order to be compatible with the Cambridge toolkit for extracting term frequency data. After processing, ASCII was translated back into Bengali script for visualisation. This is quite standard for languages with alphabetic scripts e.g. Arabic and the Buckwalter transliteration - Beesley (1997) as a simple alternative to UNICODE encoding.

Corpus Properties	Value
Number of documents	1,270
Corpus length in words	3,052,522
Number of distinct terms	192,007
Average document length	2,403.6
Standard Deviation of document lengths	555.6025
Average number of distinct terms per document	1,149.5
Standard Deviation of number of distinct terms per documents	234.3435

Table 1: Rough profile of tokenised CIIL corpus

Punctuation and numbers were removed. The corpus was tokenized using the CMU-Cambridge Toolkit (Clarkson and Rosenfeld 1997), yielding a term frequency list. Table 1 lists a few rough statistics on the tokenized corpus. This when compared with related information for the TIPSTER collection (De Roeck et al 2004a), the profile points to a small but reasonably sized corpus, containing relatively large documents, and a high proportion of distinct terms.

Manual Sampling
Manual sampling of the frequency lists showed that, to our surprise, this corpus contains a substantial number (8,791) of English words, noted in English script. These constitute a mere 0.29% of terms, but 3.9% of the distinct terms. Most occur only once, and none occur with a frequency higher than 4, so whilst worth noting, their presence is unlikely to skew the statistical profile of the corpus (though it may cause problems for some applications).

Term	Freq.	Term	Freq.
the	4	ultimate	3
ph	4	types	3
world	3	transport	3
wind	3	th	3
war	3	system	3

Table 2: 10 most frequent English terms in the CIIL corpus, with their frequency

Basic Diagnostics

Following Goweder and De Roeck (2001), we adopt two rough, but cheap techniques for *a-priori* profiling of corpus quality. First, we check for obvious imbalances by tracking term distribution patterns against Zipf's Law. Then, we look at type-to-token ratios for different text sizes. Comparing these to other languages gives an indication of language-dependent sparseness.

Zipf's Law.
Zipf's Law is useful as a rough description of the frequency distribution (Manning and Schutze, 1999). The law states that, for a reasonably representative sample of a language, the relationship between rank order (r) and frequency (f) of a term is a constant (c), i.e.

$$r.f = c$$

Set against Zipf's Law, frequency distribution in an actual dataset is a cost effective way of detecting obvious problems with the sample. If a corpus does not comply with Zipf's law, it may not be a suitable source for developing a language model. Non-compliance could be an indicator of excessive sparseness, or idiosyncratic term distribution patterns, as might be caused by a particular sub-language or document type. This would need further investigation, because skewed distribution patterns may render the corpus unsuitable as the basis for developing general language resources.
For this diagnostic, we ranked the term frequencies of the entire corpus, and plotted the frequency against the rank on both normal and logarithmic scale. According to Zipf's Law, for a representative sample, the graph on the logarithmic scale should be a straight line with slope –1 (Figure 1). The results show that term distribution in the CIIL Bengali corpus fits Zipf's law comfortably. Furthermore, the curve progresses smoothly (and not stepwise). As a result, we have no *a-priori* evidence to suspect that this corpus is either imbalanced, or excessively sparse.

[3] http://www.cill.org/

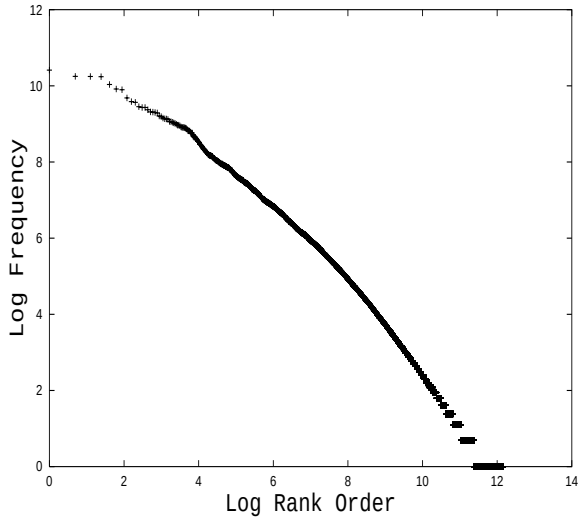

Figure 1: Zipf Curve for CIIL Corpus

Type to Token Ratio

Type-to-token ratios measure the number of "old" words we expect to see in running text before coming across a "new" one. The ratio is easily calculated by dividing the total number of terms in a fragment by the number of distinct terms. Text with a high proportion of new words will have a low type-to-token ratio. The measure is sensitive to sample size, with lower ratios (i.e. a higher proportion of "new words") expected for smaller (and therefore sparser) samples.

Even when assuming a balanced corpus, the factors that influence type-to-token ratios for raw textual data include various morphosyntactic features and orthographic conventions. For instance, the presence of a case system in a language will lead to a comparatively lower type to token ratio. Arabic, a language with a highly inflective morphology, has a very low type-to-token ratio compared to English, as first reported by Yahya (1989). Thai, on the other hand, is not particularly complex morphologically, but the script does not mark word or phrase boundaries consistently (Sukhahuta and Smith 2001), which would have similar effects for Thai data in the absence of segmentation software.

From the perspective of statistical language processing, however, it is important to note that different languages will tend to show different ratios of new to old words for identical text lengths in comparable genres. Whilst needing careful interpretation, this measure suggests that datasets of the same size will experience sparseness related problems to differing degrees, depending on the language. In other words, different languages appear to display different rates of what could be called "inherent sparseness", in the sense that it may take more text in language X than in language Y to inform a language model of the same quality.

Text length	Bengali (CIIL)	English (Brown)	Arabic (Al-Hayat)
100	1.204	1.449	1.190
1600	2.288	2.576	1.774
6400	3.309	4.702	2.357
16000	4.663	5.928	2.771
20000	5.209	6.341	2.875
1000000	10.811	20.408	8.252

Table 3: Type-to-token ratios for fragments of different lengths, from various corpora.

Goweder and De Roeck (2001), for instance, calculated type-to-token ratios for a 19 million word Arabic corpus of newspaper text, and noted a persistently low type-to-token ratio compared to the Brown corpus. We repeat the methodology here for Bengali, and report ratios associated with fragments of up to 1 million words. We picked sample sizes that allow us to compare Bengali and Arabic results with data reported for English on the Brown corpus.

The data for the three languages were derived from corpora with different profiles and genres. However, confidence in the results is supported by the variety and size of all datasets in the comparison. We note that Bengali has a significantly higher proportion of new words per running unit of text than English, and this suggests that larger corpora will be needed to produce similar effects for applications that are sensitive to sparseness.

The Behavior of Function Words

Function words are seen as uninformative in the context of information retrieval, because they occur very frequently in all documents. However, the occurrence and distribution of frequent words has some value in assessing corpus quality, on the assumption that, in a balanced collection, the most frequent terms are function words, and that function words will tend to distribute more homogeneously than content words, whose occurrence is "bursty" (Katz 1996, Church 2000). If evidence is found that very frequent function words do not distribute more homogeneously throughout the collection than less frequent words, then the suitability of the corpus for informing a language model may need closer inspection.

Hence, we will investigate the behavior of very frequent terms in a corpus, by formulating a hypothesis – that very frequent terms distribute homogeneously – and then defeat it. The method is described in detail in De Roeck et al (2004a) and applied to a different type of corpus profiling in De Roeck et al (2004b).

Measuring Homogeneity

Kilgariff (1997) introduces a method for measuring homogeneity, by measuring similarity in term distribution patterns between two halves of a corpus. His method involves the following steps:

(1) Delete document boundaries and divide the corpus into two halves by randomly placing 5000 word chunks of texts in one of two sub-corpora;
(2) Produce a word frequency list for each sub-corpus;
(3) Calculate the χ^2 statistic for the difference in term frequency distributions between the two sub-corpora;
(4) Iterate over successive random halves;
(5) Normalize.

Kilgariff (1997) calculates χ^2 of the N-most frequent terms (reporting the statistic to N-1 degrees of freedom as Chi Square by Degrees of Freedom; CBDF), but presents the results without indication of statistical relevance.

This flavor of homogeneity of term frequency distribution quality checking has been investigated in the context of corpus and genre detection, for instance in determining whether two corpora are of the same language variety and

so could be merged (Kilgariff 1997, Rose and Haddock 1997).

Chunk Size	Number of Terms (N)				
	10	20	50	100	200
5	**0.603** **0.775**	**0.747** **0.733**	**0.789** **0.812**	**0.866** **0.799**	**0.949** **0.669**
10	**1.170** **0.343**	**1.049** **0.453**	**0.987** **0.504**	**1.032** **0.396**	**0.997** **0.497**
50	**1.660** **0.453**	**1.611** **0.297**	**1.587** **0.136**	**1.356** **0.126**	1.261 0.033
100	**1.735** **0.178**	**1.485** **0.101**	**1.402** **0.099**	**1.352** **0.089**	1.425 0.009
1000	4.899 0	3.874 0	3.662 0	2.933 0	3.041 0

Table 4: Homogeneity results for various chunk sizes. The average CBDF values and p-values for a dataset using the N most frequent terms. Values in bold indicate cases where non-homogeneity is not statistically significant ($p>0.05$).

Testing the homogeneity hypothesis for very frequent words requires a more fine-grained tool than simple use of the χ^2 statistic as a homogeneity measure. We are interested in conditions under which non-homogeneity is detected and in verifying whether very frequent words distribute more homogeneously. We adapted Kilgariff's methodology in two ways. First of all, we differentiate results by reporting the p-value as well as the CBDF statistic. Given a null hypothesis (in our case, homogeneity), the p-value allows us to estimate the strength of the evidence offered by the data. Normally, a p-value < 0.05 is considered significant (moderate evidence against the hypothesis), and will here be taken to indicate that evidence of non-homogeneity is statistically significant. The CBDF measure relates to the text and indicates the level of heterogeneity.

Second, we modify the method of partitioning a document set by changing the chunk sizes. Very small chunk sizes introduce a greater randomness element. Our experiments check homogeneity for increasing chunk sizes and compare the point where heterogeneity becomes drastically significant.

Experimental results are shown in Table 4. CBDF and p-values are averaged over iterations. The results show homogeneity for the top most frequent terms at lower chunk sizes and non-homogeneity as the chunk size increases.

We conclude that, for this corpus, there is evidence that very frequent terms (the top 10 terms are all function words) distribute more homogeneously than less frequent terms. These results are in tune with the findings on the TIPSTER dataset (DeRoeck et al 2004a).

Future Work

Type-to-token ratios have being used as a measure of sparseness in this paper. One may note a pattern in which these values increase with increasing text lengths, something we'd like to study in our further research and model it. We'd also like to study this ratio and Zipf's law for different genres of Bengali text.

In the present paper we have develop a framework for fast profiling of a dataset. The present study on the most frequent terms provides some indication about the distributional properties of various terms (function and content). In our future work we aim at modeling term distributions for better profiling of corpora.

References

Beesley, K. (1997) Romanisation, Transcription and Transliteration. Xerox Web publication. http://www.xrce.xerox.com/competencies/content-analysis/arabic/info/romanization.html

Charniak, E. (1993) Statistical Language Learning. MIT Press. Cambridge, Massachussetts.

Church, K. (2000), Empirical Estimates of Adaptation: The chance of Two Noriega's is closer to p/2 than p^2, In Proceedings of Coling, pp. 173-179.

Clarkson P.R. and Rosenfeld R. (1997) Statistical Language Modelling Using the CMU-Cambridge Toolkit. Proceedings ESCA Eurospeech, pp 2707-2710

De Roeck, A., Sarkar, A. and Garthwaite, P.H. (2004a) Defeating the homogeneity assumption. Proceedings of The 7th International Conference on the Statistical Analysis of Textual Data (JADT 2004).

De Roeck, A., Sarkar, A. and Garthwaite, P.H. (2004b) Frequent Term Distribution Measures for Dataset Profiling. Proceedings of The 4th International conference of Language Resources and Evaluation.

Dunning, T. (1993) Accurate Methods for the statistics of surprise and coincidence. Computational Linguistics. 19(1):61-74.

Goweder, A. and De Roeck A. (2001). Assessment of a significant Arabic corpus. Proceedings Workshop on Arabic Language Processing, 39th ACL. Toulouse.

Harman, D. (1991) How effective is suffixing? JASIS, 42:7-15.

Katz, S. (1996) Distribution of content words and phrases in text and language modelling. Natural Language Engineering, 2(1):15-59.

Kilgariff, A. (1997) Using word frequency lists to measure corpus homogeneity and similarity between corpora. Proceedings ACL-SIGDAT workshop on very large corpora, Hong Kong.

Manning, C. and H. Schuetze (1999) Foundations of Statistical Natural Language Processing. MIT Press. Cambridge, MA.

Popovic, M. and P. Willett (1992) The effectiveness of stemming for natural language access to Slovene textual data. JASIS, 43:384-390.

Rose, T., Haddock N. and R. Tucker (1997) The effects of corpus size and homogeneity on language model quality. Proceedings ACL-SIGDAT workshop on very large corpora, pp178-191, Hong Kong.

Sukhahuta, R. and Dan Smith (2001) Information Extraction Strategies for Thai Documents. International Journal of Computer Processing of Oriental Languages, 14:2,153-172

Yahya, A.H. (1989) On the complexity of the initial stages of Arabic Text Processing. Birzeit University, Birzeit, West Bank.