Language-based text Segmentation using Word Frequency Statistics- an Experimental Analysis

Nicholas Akosu, Obasan Adebola & Norafida Ithnin

Abstract— In text processing it is often necessary to process a large amount of text in order to achieve Segmentation. The main objective of this research is to study the word frequency statistics of Universal Declaration of Human Rights (UDHR) translations in fifteen (15) languages with a view to systematically reduce the quantity of text to be processed in order to achieve a specific task. This will form part of an extended study to investigate possibilities of utilizing the Pareto principle to attain enhanced performance for language identification algorithms.

Key words: Word frequency, frequency statistics ,information retrieval, Pareto principle , Zipf’s law, Segmentation, Pareto analysis

.

1. INTRODUCTION

Interesting challenges exist in the area of information retrieval and text categorization in a multilingual environment such as the Internet. These include issues concerning language identification of resource-scarce languages, classification of textual data for purposes of Spam Filtering, Identification of Document Genre, and Authorship Attribution among others. Both statistical and linguistic methods have been employed for these tasks. The decision to use one kind of method over the other often depends on preference but most times information availability constraints influence the choice. For example in cases where large corpora exist statistical methods perform well in that there is enough data for statistical training of the required models. Where such are not available or not available in sufficient quantity, linguistic methods hold greater promise. In such cases it becomes interesting to discover such attributes as the anatomy of sentences in a particular writing or language, the statistical composition of the text by word frequency, word length and stop word content in order to consider how to take advantage of the data in the course of processing. For example, how can we determine if the Pareto principle is applicable in any given situation? Also, Can the Pareto principle be used to improve on the performance of language identification techniques for resource-scarce languages? These, and similar questions prompted this research.

According Newman [8], the Pareto principle (also known as the 80–20 rule, the law of the vital few, and the principle of factor sparsity) states that, for many events, roughly 80% of the effects come from 20% of the causes. On the other hand, [11] stated that if we analyze a corpus in any natural language, the frequency of any word is inversely proportional to its rank in the frequency table. This means that the most frequent word will appear approximately twice as often as the second most frequent word, thrice as often as the third most frequent word, and so on. For example, it has been shown that “the” is the most frequent word in the BROWN CORPUS and it accounts for about 7% of all words (69,971words out of slightly over 1 million words) in the said corpus. The next word is, “of”, accounting for about 3.5% of words, (with 36,411 occurrences), followed by “and”, (with 28,852 instances). It is interesting to note that only 135 word types account for half the Brown Corpus [3].

However, it is important to note that Zipf’s law was derived from large text corpora, which means that it may not immediately become obvious if the text under consideration is small. However, it gives us a lead to study languages in a slightly different fashion. In this research we study the relationship between word frequency and text composition with respect to the size of the document. Such information could become critical in other applications such as optimization of language identification techniques.

The rest of the paper is structured as follows: Section 2 examines related work and section 3 presents the methodology while section 4 explains the experimental setup. Section 5 discusses the results and section 6 concludes the paper.

Corresponding Author: Obasan Adebola & Nickolas Akosu are currently pursuing Ph.D degree program in computer science Universiti Teknologi Malaysia,Malaysia,FSKSM,81310 Skudai.

2. RELATED WORK

In Computer science and engineering control theory, the Pareto principle is often applied to optimization efforts [7]. For example, Microsoft has noted that by fixing the top 20% of the most reported bugs, 80% of the errors and crashes were often eliminated [10]. Vilfredo Pareto was one of the first people to analyze economic problems using mathematics; in the late 1800s, he observed that 80 percent of the land in Italy was owned by 20 percent of the population. After Pareto developed his formula, many other researchers observed similar phenomena in their own field of research [4], [9], [1]. According to Dr. Juran’s observation of the “vital few and trivial many,” 20 percent of the tasks are always responsible for 80 percent of the results [2]. Even though, Pareto originally applied the concept to distribution of wealth (20 percent of the people owned 80 percent of the wealth) and Juran to quality (20 percent of the defects cause 80 percent of the problems), over the years the 80/20 rule has been expressed in a number of different ways. Some of these are:

(a) 80 percent of the results are achieved by 20 percent of the group.
(b) 20 percent of your effort will generate 80 percent of your results.
(c) In any process, few elements (20 percent) are vital and many elements (80 percent) are trivial.
(d)If you have to do ten things, two of those are usually worth as much as the other eight put together.
(e) 20 percent of the tasks account for 80 percent of the value.

Pareto analysis can be effectively employed for separating the major causes (the “vital few”) of a problem, from the minor ones (the “trivial many”). Such an analysis focuses one’s attention to tackling the major causes of the problem at hand rather than wasting time on the minor ones. According to [6], the 80/20 rule has been tried and tested over the years but it has withstood the stringent scrutiny that it has been subjected to.

In this research we experimentally study the word frequency distribution on translations of a single document (UDHR) in 15 languages with a view to understanding how the results relate to the Pareto principle and Zipf’s law. The results, if properly harnessed, may well contribute to enhancing the performance of techniques for language identification and other text retrieval systems. This is important in view of the large amount of data that is often processed to achieve meaningful outcomes in these applications.

3. METHODOLOGY

Word frequency statistics are generated by first running a frequency distribution on the entire dataset. This tags all word types with their frequency of occurrence. Thereafter we loop through the entire document (wordlist) and aggregate all words with their frequency thereby determining its fractional component in the document. For example if the word “can”, appears 20 times in a document of size 100 then its percentage is (20/100) 10%. This will also automatically mean that the word “can”, will be included in a statistic of words occurring “20 or more” times.

4. EXPERIMENTAL SETUP

In this research we investigate the word frequency statistics on the UDHR translations in 15 languages namely, Asante, Akuapem, Ndebele, Malay, Indonesian, Croatian, Serbian, Tiv, Yoruba, Igbo, Hausa, Zulu, English, Swahili and Slovakian. For each language we run a frequency distribution on its word list and then stratify the words according to the following categories: words occurring 1-2 times, words that appear 3-4 times, words that appear 5-9 times, words that occur 10-14 times, words that appear 15-19 times, words that occur between 20 -24 times and finally, words that occur 25 or more times.

Thereafter we compute for each category the number of tokens, the number of word types and the percentage of the document comprising that category.

5. RESULTS AND DISCUSSION

Results of the frequency distribution for four languages are shown in Tables 1a and 1b. In table 1 we see that the highest frequency words actually hold the greater percentage of the document. For Hausa, words occurring 25 or more times form 27.1% of the document, while Tiv is 39%, English is 38.5% while Malay is 25.3%. Apart from this we observe something else that is very interesting. For Hausa, the 27.1% of the document is only comprising 8 word types!, for Tiv, 39% is formed by only 15 word types, English 38.5% has only 12 types while Malay’s 25.3% consists of 10 word types only. This is very significant because for any spell checking operations we can process 39% of the Tiv language document by checking only 15 words. Similarly we can process 27.1% of the Hausa document by checking 8 words, 38.5% of the English document by checking only 12 words and 25.3% of the Malay language document will be done with by checking 10 words. The practical implications of this result in terms of savings in processing time will be enormous; perhaps this could come along with some space savings as well.

Table 1a: Frequency statistics for Hausa and Tiv languages

Hausa (1826 Tokens)				Tiv (1803 Tokens)			
Frequency	No of tokens	Types	% of Doc	Frequency	No of tokens	Types	% of Doc
>=25	495	8	27.1	>=25	703	15	39
20-24	132	6	7.2	20-24	107	5	6
15-19	117	7	6.4	15-19	34	2	1.9
10-14	113	12	7.3	10-14	189	16	10.4
5-9	380	55	20.8	5-9	189	46	16.6
3-4	203	61	11.1	3-4	195	58	10.8
1-2	366	286	20.0	1-2	275	217	15.2

Table 1b: Frequency statistics for Hausa and Tiv languages

English (1576 Tokens)				Malay (1306 Tokens)			
Frequency	No of tokens	Types	% of Doc	Frequency	No of tokens	Types	% of Doc
>=25	606	12	36.5	>=25	331	10	25.3
20-24	0	0	0	20-24	0	0	0
15-19	37	2	2.3	15-19	92	6	7
10-14	118	10	7.5	10-14	160	14	12.3
5-9	209	32	13.2	5-9	155	24	11.9
3-4	154	47	9.8	3-4	164	49	12.6
1-2	452	380	28.7	1-2	404	336	30.9

Table 2: Frequency statistics for 15 languages

Language	Frequency	No of Tokens	No of Types	% of Doc
Hausa	>=25	495	8	27.1
Tiv	>=25	703	15	39.0*
English	>=25	606	12	38.5*
Malay	>=25	331	10	25.3
Zulu	>=5	214	14	21.2
Swahili	>=25	629	12	37.6*
Ndebele	>=5	266	23	27.7
indonesian	>=25	338	9	25.1
Croatian	>=15	320	12	23.4
Serbian	>=15	342	13	23.9
Slovak	>=10	343	16	25.7
Igbo	>=25	827	16	43.1*
Yoruba	>=25	1145	19	72.5*
Asante	>=25	891	17	46.3*
Akuapem	>=25	1167	19	58.9*

Table 3: Frequency statistics for 15 languages (Adjusted)

Language	Frequency	No of Tokens	No of Types	% of Doc
Hausa	>=25	495	8	27.1
Tiv	>=45	384	5	21.3
English	>=55	368	4	23.4
Malay	>=25	331	10	25.3
Zulu	>=5	214	14	21.2
Swahili	>=55	368	4	22.0
Ndebele	>=5	266	23	27.7
indonesian	>=25	338	9	25.1
Croatian	>=15	320	12	23.4
Serbian	>=15	342	13	23.9
Slovak	>=10	343	16	25.7
Igbo	>=55	402	4	20.9
Yoruba	>=85	345	3	21.8
Asante	>=55	495	6	25.7
Akuapem	>=85	473	4	23.9

Table 2 shows the summary statistics for all the languages studied. Here we observed that some of the languages reveal a much larger percentage of the document in the category of words occurring 25 or more times. This can be seen in Igbo, Yoruba, Asante, Akuapem, Tiv, English and Swahili. For these languages we expanded the frequency table to include words occurring up to 85 or more times (Yoruba and Akuapem) and others for smaller ranges. This then gave us the final statistics for what will be required to process approximately 25% of each document. The results for this are shown in Table 3. Notice that in each of these cases we have succeeded in reducing the number of words to be processed to attain about 20% completion.

6. CONCLUSION AND FUTURE WORK.

From the results obtained in our experiments we confirmed that Zipf's law appears to have been obeyed even as the size of the dataset is small. We further confirmed that the word frequency statistics show interesting results that may be useful when applied to optimization of language identification techniques for resource-scarce languages. In future research we plan to investigate the viability of the word frequency statistics in enhancing the performance of language identification algorithms.

7. ACKNOWLEDGMENT

The authors would like to thank Universiti Teknologi Malaysia, UTM for providing a suitable environment for research and the significant contributions of FSKSM members of staff .

References:

[1] Adamic, L. A. & Huberman, B. A. (2000). The nature of markets in the World Wide Web. Quarterly Journal of Electronic Commerce 1, 512.

[2] Bunkley, N. (2008), Joseph Juran, 103, Pioneer in Quality Control, Dies, New York Times. Retrieved 7th Nov. 2012 from http"//www.nytimes.com/2008/03/03/business/03juran.html.

[3] Dahl, H. (1979). Word Frequencies of Spoken American English. Verbatim, Essex, CT

[4] Li, W. (1992). Random texts exhibit Zipf's-law-like word frequency distribution. IEEE Transactions on Information Theory 38, 1842-1845.

[5] Library of Congress (2008). The subject headings manual. Washington, DC.: Library of Congress, Policy and Standards Division. (Sheet H 180: "Assign headings only for topics that comprise at least 20% of the work.")

[6] Newman, M. E.J.(2012) Power laws, Pareto distributions, and Zipf's law. Retrieved 10th Nov. 2012 from http://arxiv.org/ps_cache/cond-mat/pdf/0412/0412004v3.pdf

[7] Gen, M.; & Cheng, R. (2002), Genetic Algorithms and Engineering Optimization, New York: Wiley

[8] Newman, M. E. J. Forrest, S. & Balthrop, J. (2002) Email networks and the spread of computer viruses. Phys. Rev. E 66, 035101.

[9] Redner, S. (1998). How popular is your paper? An empirical study of the citation distribution. Eur. Phys. J. B 4, 131-134.

[10] Rooney, P. (2002), Microsoft's CEO: 80-20 Rule Applies to Bugs, Not Just Features. Retrieved 11th Nov. 2012, from http://www.crn.com/news/security/18821726

[11] Zipf, G.K.(1949). Human Behavior and the Principle of Least Effort. Addison-Wesley, Reading, MA.