Truncation of Content Terms for Turkish

Hayri Sever
Computer Engineering Department
Başkent University
06530 Bağlıca, Ankara, TR
sever@baskent.edu.tr

Yaşar Tonta
Information Management Department
Hacettepe University
06532 Beytepe, Ankara, TR
tonta@hacettepe.edu.tr

1. Introduction

It is a well known observation for about last two decades that when exponential growth in storage and processing capacities were combined with rapid developments in computer network technologies, many applications which were impossible to envision before have been either flourished or conceived. Of them, IR (Information Retrieval) techniques for Turkish have become critical since, until second half of 90's, there were no systematic researches on retrieval systems and tools especially focusing on electronic resources. As part of such a systematic research movement, we can list works on metadata [1], stemming [2,3], search engines [4], and statistical nature of Turkish language, such as zipf law and vocabulary growth rate [5].

Turkish as a member of the south-western or Oghuz group of the Turkic family of languages is an agglutinative language with word structures formed by productive affixations of derivational and inflectional suffixes to the root words. In Turkish language, there are approximately 25,000-30,000 stems actively used. When the inflection of the words are included, the number is, however, expressed in millions. Furthermore, the index of synthesis[1] for Turkish language is found to be 2.86. It is conjectured in [2] that when content terms are used for indexing without stemming, space efficieny and effectiveness of IR system decrease substantially. More specifically, it is experimentally shown that the retrieval effectiveness of a typical vector-based IR system is increased by approximately 22% in terms of normalized precision. Early Turkish search engines[2] had many problems, but prominent ones can be listed as lesser Turkish content coverage, novelty, and recency ratios than mega search engines such as yahoo and altavista [4]. Both local and mega search engine groups have something in common: no Turkish stemming.

In this article, we investigate statistical nature of truncated terms for Turkish. In English, it was reported that indexing by truncation of each word to four or five characters yielded almost as good discrimination between relevant and nonrelevant documents as does a system that uses full terms. We test the conjecture that truncation of Turkish words around average word length gives a similar Zipf behavior with one to another, and furthermore, we show that validity of this conjecture can be extended to whole words as well. Upon proving such a

[1] Index of synthesis refers to the amount of affixation in a language, i.e., it shows the average number of morphemes per word in a language.
[2] It is, unfortunately, the case that there is currently no Turkish search engine.

conjecture, we claim that choosing either of truncation schemes involving in taking first five, six, or seven characters for indexing is as good as choosing whole words. Validity of this claim provides us considerable space efficiency for indexing systems without loosing effectiveness.

In retrieval systems it is important to pick up proper content terms for indexing. A right practice for extracting content terms would be to use auxiliary stop list or frequency information. Essence of frequency information for indexing states that a content term being likely to reduce density of document space should have property of medium frequency. In this article, we also provide growth rate constants that determine approximate number of distinct words for a given text size. More generally, we address how to extract statistics of Turkish language including determination of words with medium frequencies. Finally we think that main benefit one can get out of our study is to provide a rapid and feasible text retrieval environment for Turkish.

2. Background

2.1 Zipf's Law

Many man made and naturally occurring phenomena, including city sizes, incomes, word frequencies, and earthquake magnitudes, are distributed according to a *power-law* distribution. A *power-law* implies that small occurrences are extremely common, whereas large instances are extremely rare. This regularity or 'law' is sometimes also referred to as *Zipf*. In a power-law we have $y = C\, x^{-a}$, which can be manipulated as $log(y) = log(C) - a\, log(x)$. So a power-law with exponent "*a"* is seen as nearly inversely linear with slope "*-a"* on a log-log plot. Zipf's Law based on *the principle of least effort* can be used to approximately model human behavior, particularly on the use of a natural language.

2.2 Vocabulary Growth

Heaps' Law gives a general formula for determining the vocabulary size for a corpus containing *n* words. Vocabulary can be defined as the set of distinct words in a corpus. In natural language text, *Heaps' Law* is used to predict the growth of the vocabulary size. This law states that the vocabulary of a text of size, say *V*, with *n* words (total number of words) can be approximately determined by the formula of $V = kn^{B}$, where *k* and *B* depend on the particular text.

3. Method

TurCO is a text corpus of 50,111,828 words [6], which mainly consists of two categories out of six ones: (1) closed-captioned talks at Turkish Parliament (52%) and (2) online edition of newspapers or magazines (44%). Each file in the corpus consists of only 29 lowercase Turkish characters (Consonants: b, c, ç, d, f, g, g, h, j, k, l, m, n, p, r, s, s, t, v, y, z; Vowels: a, e, i, i, o, ö, u, ü) and space character. Most of the documents included in the corpus have been collected from different web sites and there are some samples of Turkish novels and stories (0,13%). In this corpus, average word length is about 6,3.

Truncation Level	Vocabulary Count	Zipf Constant A	Zipf Constant A for rank range [0.1%,10%]
whole word	686804	0,022	0,084
7	259797	0,019	0,094
6	185602	0,018	0,097
5	110351	0,016	0,098
3	10933	0,028	0,185

Table 1. Frequencies of distinct words truncated at different levels and their Zipf constants for both whole corpus and interval between 0.1% and10% of ranks. Note that words are partially sorted with respect to their frequencies in descending order and then, rank number is assigned in ascending order for each row starting from top. Finally Zipf constant is computed by averaging $A(r)=p(r)*r$ values of each row, where r and $p(r)$ indicate rank and probability of that rank, respectively.

4. Discussion

In Table 1, frequencies of truncated terms and their Zipf constants are given. We see that in regard to Zipf constant value the truncated terms around average word length can be treated equally because maximum change factor of Zipf constant in that group is about 18%, i.e., (0,019-0,016)/0,016. When we consider rank range between 300 and 4000, we see that this change factor comes to about 4% i.e., (0,098-0,094)/0,094. This shows when we consider word frequency distribution rank between 300 and 4000, any truncation scheme involving in taking first five, six, or seven becomes as good as another one. This fact can be observed from charts given in Figure 1. Vocabulary growth rates of truncated terms around their average word length and whole words are shown in Figure 2.

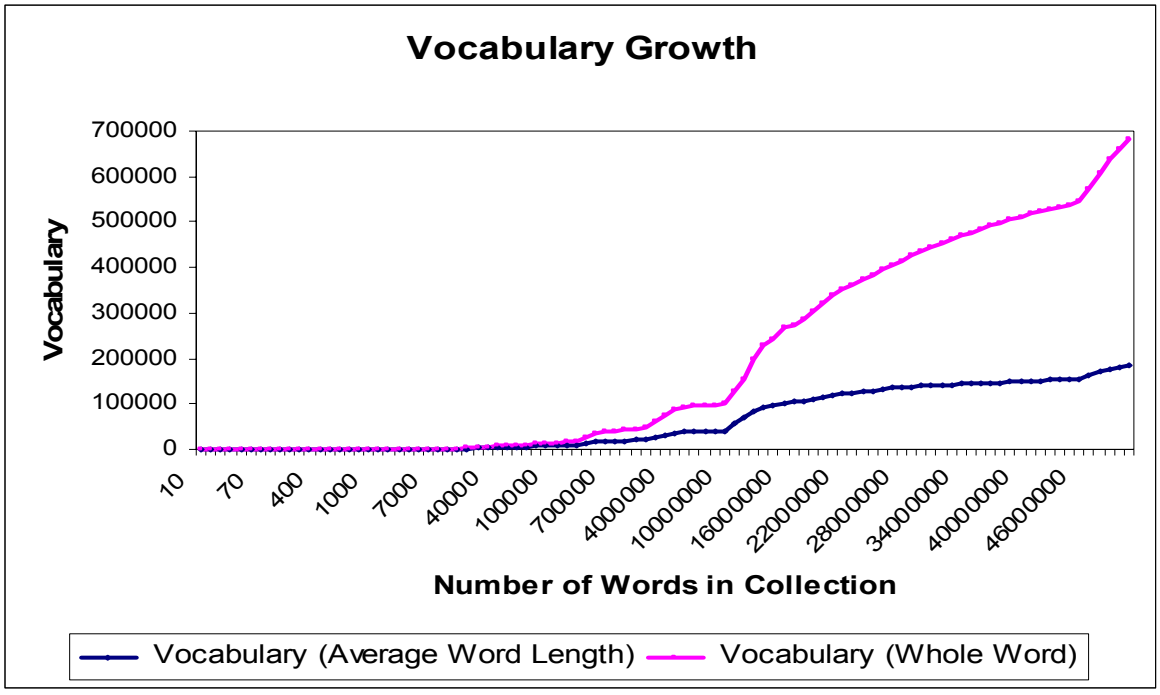

Figure 2. Vocabulary growth rate of truncated words around average word length and of whole words. Last files of TurCO corpus consist of book samples whose novelty ratio can be

regarded as very high, i.e., contribution of distinct words over total number of words of these last files. One can easily see that collection being made up of whole words is very sensitive to changes in novelty ratios, on contrary to the case where the collection is made up of truncated terms around average word length.

5. Conclusion

In this article we investigated feasibility of using different truncation schemes for indexing. We showed that truncation of words around their average length yields better efficiency without loosing effectiveness for Turkish text.

References

[1] M.E. Kucuk, B. Olgun, and H. Sever_, Application of Metadata Concepts to Discovery of Internet Resources, Lecture Notes in Computer Science (LNCS), Springer Verlag, Vol. 1909, pp. 304-13, 2000.

[2] H. Sever and Y. Bitirim. The Analysis and Evaluation of Stemming algorithms for Turkish. 10th International Symposium on String Processing and Information Retrieval (SPIRE'03), Lecture Notes in Computer Science (LNCS), Springer, 2857: 238-51.

[3] B. T. Dinçer and B. Karaoglan. Stemming in Agglutinative Languages: A Probabilistic Stemmer for Turkish. ISCIS 2003: 244-251.

[4] Y. Bitirim, Y. Tonta, and H. Sever. Information Retrieval Effectiveness of Turkish Search Engines. ADVIS'02, Lecture Notes in Computer Science, Springer Verlag, Vol. 2457, pp. 93-103, 2002

[5] Gökhan Dalkiliç and Yalçin Çebi. Zipf's Law and Mandelbrot's Constants for Turkish Language Using Turkish Corpus (TurCo). ADVIS 2004: 273-282.

[6] Gökhan Dalkiliç, Yalçin Çebi. A 300 MB Turkish Corpus and Word Analysis. ADVIS 2002: 205-212.

Acknowledgment

This work is partially funded by TUBİTAK-TİDEB under the grant number of TİDEB3040054.

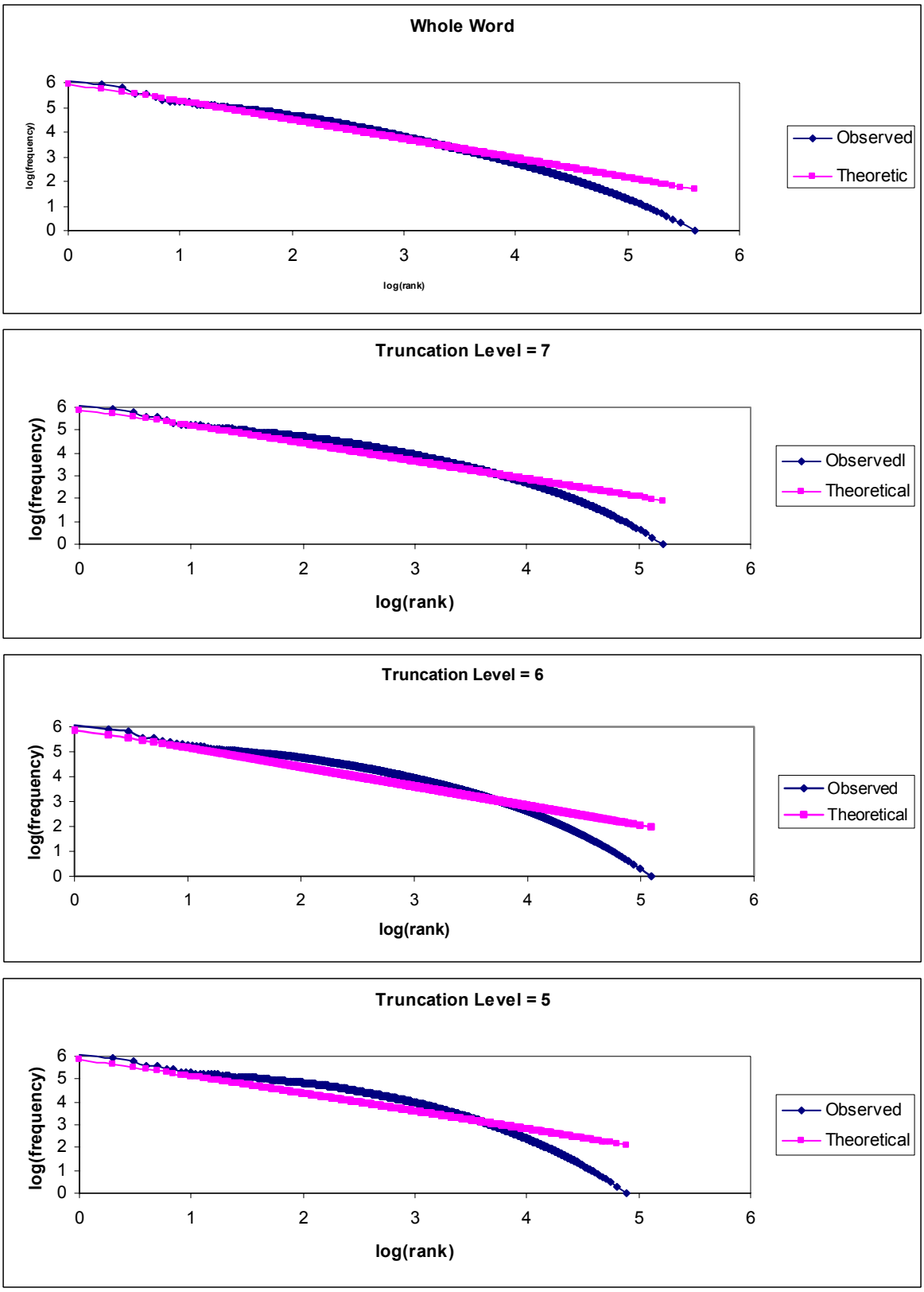

Figure 2. Charts of rank vs frequency on log base for some different truncation levels.