Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia - Social and Behavioral Sciences 27 (2011) 233 – 240

Procedia
Social and Behavioral Sciences

Pacific Association for Computational Linguistics (PACLING 2011)

Frequencies determination of characters for Bahasa Melayu: results of preliminary investigations

Asadullah Shah[a,*], Aznan Zuhid Saidin[b], Imad Fakhri Taha[a], Akram M. Zeki[b]

[a]*Department of Computer Science, Kulliyyah of ICT, International Islamic University Malaysia, 53100, Kuala Lumpur, Malaysia*
[a]*Department of Information Systems, Kulliyyah of ICT, International Islamic University Malaysia, 53100, Kuala Lumpur, Malaysia*

Abstract

Bahasa Melayu (Malay language) is a language spoken in Malaysia and many countries around it. It has rich literature and deep roots in culture. Bahasa Melayu language uses roman character set (i.e. A-Z) identical to English language. The written language uses the character set as building blocks to build word, sentences and phrases along with special punctuations and signs to create documents of interest. In this paper, results of preliminary investigation of Malay text documents are provided. For this purpose scanning of articles written upon various topics in Malay were carried out. Approximately 31 thousand characters from different articles are scanned. Preliminary observations indicate that on average, character "A" occurs 19%, character "N" occur 10%, character "E" occur "9%"and character "I" 8% in text. However, it is also observed from the data that, these are the characters from over all set with highest frequencies of occurrences and it is expected that during further investigation they will remain as higher frequency occurring characters. Furthermore, the results indicate that for Bahasa Melayu characters appearance in text is very close in character frequencies of Bahasa Indonesia, but having different appearance of characters than English language. The investigation also indicate that these two languages, Bahasa Melayu and Bahasa Indonesia share close phonetic structure but not English, though all three use same character set.

Keywords: Bahasa Melayu ; Bahasa Indonesia ; English language ; absolute frequencies ; relative frequencies ; running average

* Corresponding author. Tel.: +6-03-6196-5627; fax: +6-03-6196-5179.
E-mail address: asadullah@kict.iium.edu.my.

doi:10.1016/j.sbspro.2011.10.603

1. Introduction

Text documents for any language are prepared from its standard symbol set, may it be Malay, English, Arabic and Indonesian or any other language. The set of the symbols in all of these languages is a limited number of characters. Each one of these characters is used for text document preparations and words are formulated, from words sentences are built and phrases, passages and then the full documents are prepared. Some characters are more frequently used than others. For example, the character "E" in English appears in text with highest frequency. Similarly, in Bahasa Indonesia, the character "A" has the highest frequency of occurrence in Indonesian text documents [1].

Bahasa Melayu (Malay language) also use character set same as English "a-z". The frequencies of English language characters, such as, absolute and relative are reported in [2][3] and for Indonesian language in [1], and their determination is reported in the literature, however, for, Bahasa Melayu, its character frequencies and other statistics is not known. Therefore its character frequencies determinations are considered in this research.

Determination of absolute frequencies usually needs a great data and documents to be scanned, whether using computer program or scanning them document by document manually. In past some work has been reported [4] in which the first-order, second-order and third order of Malay printed documents is reported. However, determination of the relative frequencies, absolute frequencies and other statistical characteristic of the Bahasa Melayu are not reported in literature so far. In this study, the preliminary results about the relative and absolute frequencies of each individual character are reported. Mostly, the preliminary investigation is based upon the scanning of 31 thousand characters from newspapers with different editorials and other standard text documents and results are reported in this paper.

2. Frequencies of Bahasa Melayu characters

The frequencies of characters can be of two types: relative frequencies, depending upon the current document under observation and absolute frequencies of the characters, those are the frequencies determined based upon an aggregation of frequencies found in all separate documents under observation and averaged out. There is another simple measure used in conjunction with the absolute and relative frequencies that is the running average of frequency for each character. For Bahasa Melayu these frequencies of characters can then be utilized for various computer applications such as text document compression and cryptography; which in turn has may other benefits for users of these applications. In this study, relative frequencies, absolute frequencies and running averages are observed and reported.

3. Methodology

Included: For this study 24 articles of newspapers are considered. The size of these articles varies from minimum of 1270 characters to maximum of 3665 characters per document. The titles chosen for the research involve different topics, like current affairs, politics, crime, social issues, science and technology, business news, sports and entertainment. In this study approximately 31thousand characters are scanned. Each word is manually scanned and character frequencies are recorded. The words are as small as two characters to 15 characters long. For each document the relative frequencies are observed. The characters in upper and lower case are both considered. After knowing the relative frequencies, the absolute frequencies are calculated for all characters from A to Z.

Not included: The short forms, numbers and numerical figures, special punctuations, signs, and spaces are not considered for experimental observations.

4. Relative probabilities of characters A, N, E and I

Relative probabilities are the probabilities of characters within each document. In probabilistic representation the relative probability is the count of a particular character, for example "A" within a document of observation divided by the total number of character in that document.

5. Results of observation

Initial results of all 24 documents scanned indicate that four characters with the highest occurrence in Bahasa Malay are characters "A", "N" and "E" and "I" are shown in Figures 1-4, respectively.

It is shown in the Figure 1, that for character "A", the relative probability varies between 0.145 to 0.249 that also indicate that it varies between 14% to 24%.

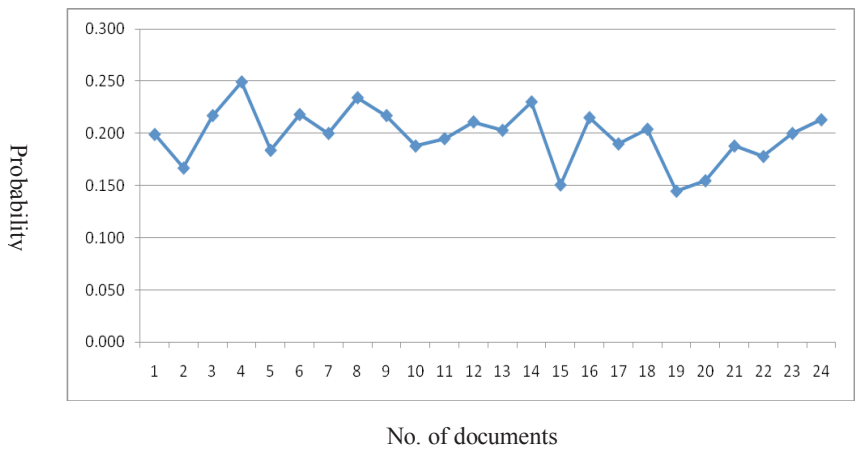

Fig. 1: Relative probabilities of character "A"

For character "N" the relative probability is between 0.086 to 0.124 and shown in Figure 2: This means that the frequency varies between a minimum of 8% to a maximum of 12%, respectively.

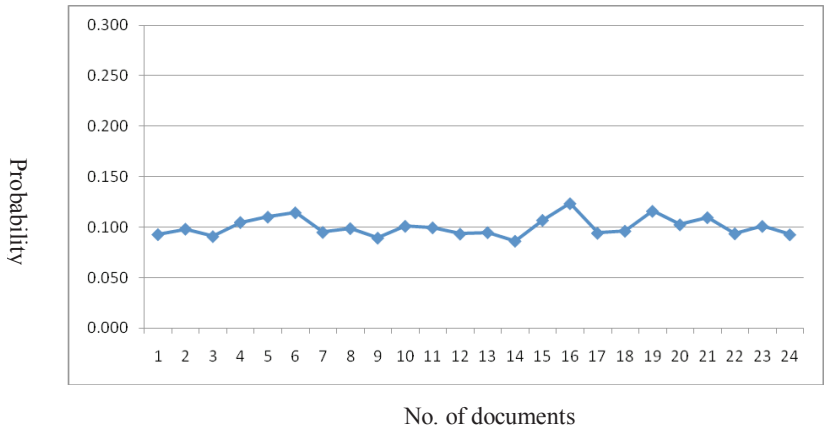

Fig. 2 : Relative probabilities of character "N"

For character "E", the relative probability is between 0.059 to 0.111 and shown in Figure 3: It frequency varies between 5% minimum to a maximum of 11%.

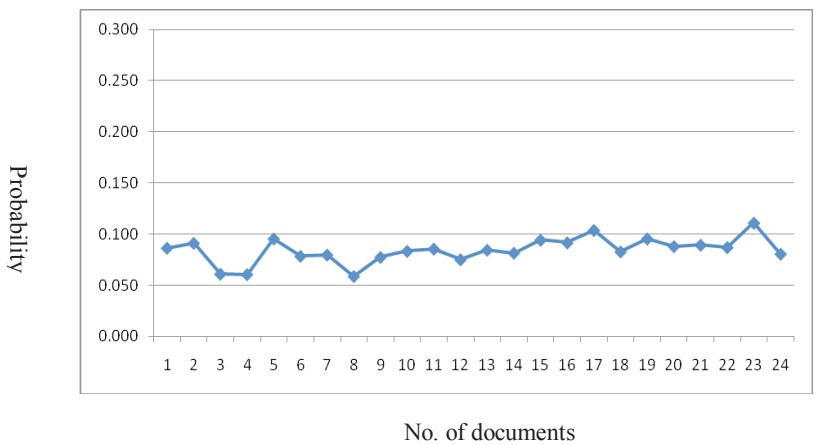

Fig.3 : Relative probabilities of character "E"

Similarly for character "I", the relative probability varies between 0.062 to 0.105 and shown in Figure 4: Its frequency varies between 6% minimum to 10% maximum.

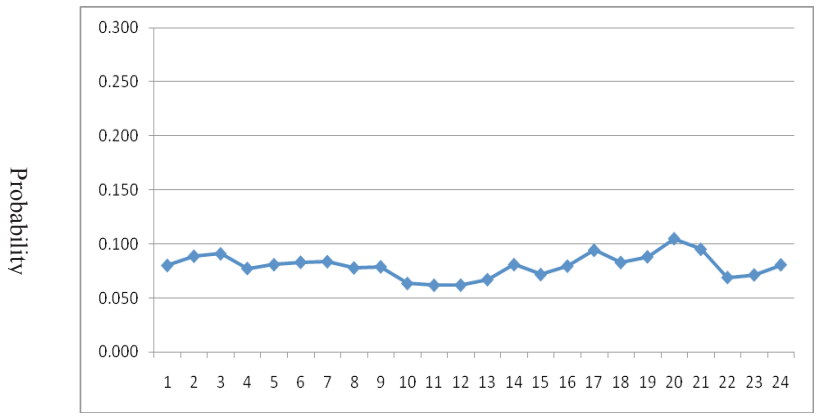

Fig.4 : Relative probabilities of character "I"

6. Running average

The running average of the characters' frequency is calculated to know and determine the frequencies of characters in every new document along with the previously scanned documents. This allows the variation of frequencies for each character to be stabilized and determine the absolute frequencies out of all the documents scanned. The running average is calculated by dividing the total frequency of a character with the number of documents in which that character is observed. Running average is also called moving average or cumulative average; the formula for moving average is reported in [5] and given as below:

$$CA_i = \frac{x_1 + \cdots + x_i}{i}.$$

Where each x is represents a total of an individual character in a document and i represent the total number of all documents.

The running average frequency for character "A" is shown in Figure 5, which shows that it is stabilizing at the frequency of around 230. This also indicate that most probability in more documents the running average frequency of character "A" will somewhere stay around 230 in all 24 documents scanned so far.

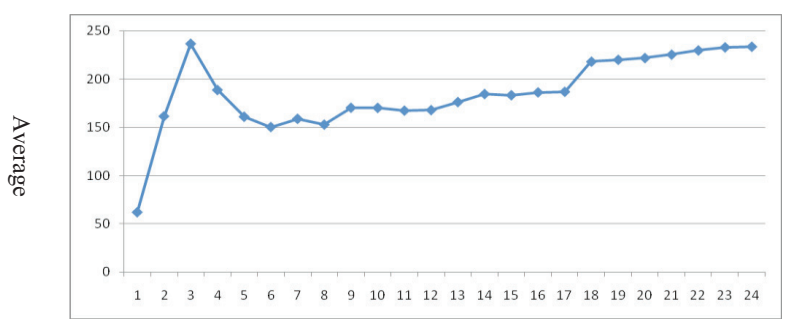

Fig. 5 : Running average for character "A"

The running average frequency for character “N” is shown in Figure 6, stabilizing at the frequency of around 120.

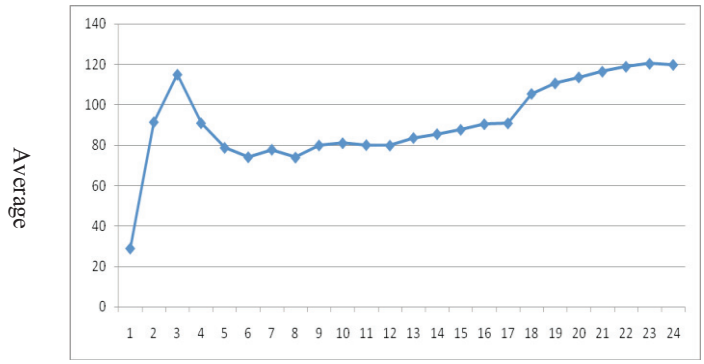

Fig. 6 : Running average for character “N”

The running average frequency for character “E” is shown in Figure 7, stabilizing at the frequency of around 100.

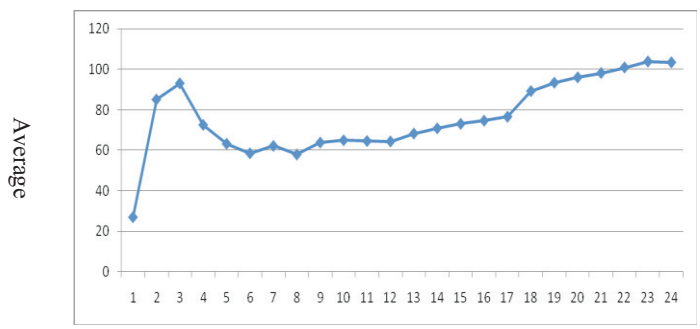

Fig. 7 : Running average for character “E”

The running average frequency for character “I” is shown in Figure 8, stabilizing at the frequency of around 98.

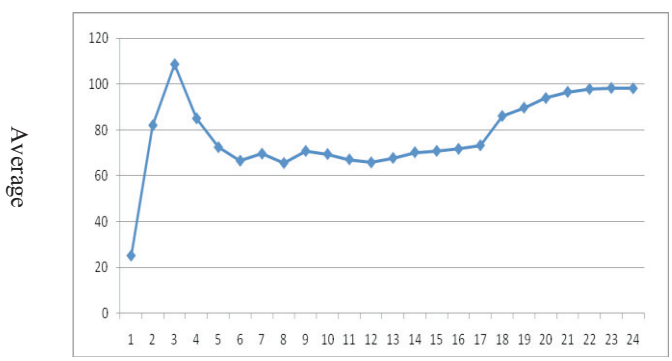

Fig. 8 : Running average for character “I”

By increasing the number of documents in observations, the frequency of each character will be reaching to its saturation value. This measure can be applied to all characters in the Bahasa Melayu to determine their running average frequencies.

After a great number of documents scanned, a stage will arrive where this running average for all characters will be saturated at a fixed value. That will be the value for each character to be fixed and determined for any use or implementation in applications such as text compressions.

7. Absolute frequency pie chart for all characters from A-Z

Based on total characters of all 24documents investigated, the character with the highest frequency is "A" (19%), followed by "N" (10%), then "E" (9%) and "I" (8%). The Characters Q and X are not occurred at all and PIE chart indicate 0% occurrence. Other characters with lower frequencies are "W", "V", "Z", "C", "F", "J" and "O". All the rest, other than above mentioned characters, the frequencies are between 3% to 5%. Frequencies for all characters are shown in Figure 9 in the pie chart.

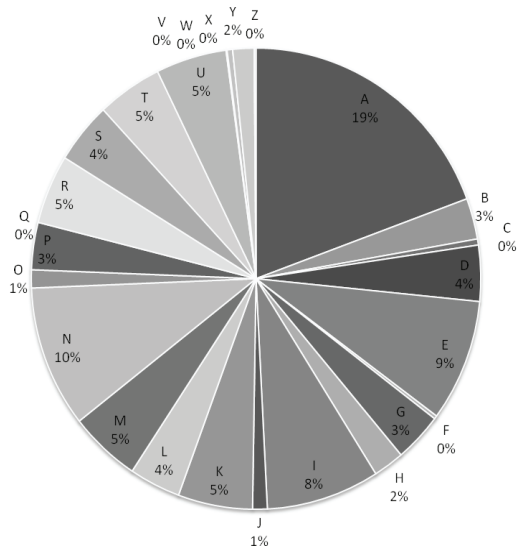

Fig. 9 : Absolute frequency for characters "A" to "Z"

8. Conclusion

The results show that the character with the highest frequency is "A", followed by "N", "E" and "I" respectively. The frequency of character "A", "N" and "E" are closely resembling to that of Bahasa Indonesia. It is predicted that this will remain true even if the number of documents are increased based on the frequency of "A" being significantly higher than that of "N" or "E". Characters "Q" and "X" have shown zero frequency in this study of 31 thousand characters scanned so far. Words in Bahasa Melayu having these characters are rare, mostly occurring in words originating from other languages which are used in certain limited context, such as "aqidah" (from Arabic, concerning Islamic religious belief) and "x-ray" (from English). Issues mentioned here warrant for further studies.

Acknowledgement

This study was funded by the International Islamic University under the Research Endowment Grant (Type A) EDW A11-084-0875.

References

[1] S. Trost. (2011, March 14) Character frequency: Indonesian (Bahasa) [Online] Available: http://www.sttmedia.com/characterfrequency-indonesian.

[2] Wikipedia. (2011, February 20) *Letter frequency* [Online] http://en.wikipedia.org/wiki/Letter_frequency.

[3] WorldLingo. (2011, March 14) *Letter frequency* [Online] http://www.worldlingo.com/letter_frequency

[4] A. Wang, " First, second and Third order Entropies of Printed Malay", *Indian Journal of Statistics*, ser. B, vol. 46, pt 3, pp 372-376, 1984.

[5] Wikipedia. (2011, April 1) *Moving Average* [Online] http://en.wikipedia.org/wiki/Moving_average.