IOP Conf. Series: Materials Science and Engineering **180** (2017) 012060 doi:10.1088/1757-899X/180/1/012060

The Usage Evaluation of Official Computer Terms in Bahasa Indonesia in Indonesian Government Official Websites

A Amalia [1], D Gunawan[2], M S Lydia[1] and C Charlie[2]

[1] Department of Computer Science, Faculty of Computer Science and Information Technology, University of Sumatera Utara, Medan, Indonesia

[2] Department of Technology Information Faculty of Computer Science and Information Technology, University of Sumatera Utara, Medan, Indonesia

*Email: amalia@usu.ac.id

Abstract. According to Undang-Undang Dasar Republik Indonesia 1945 Pasal 36, Bahasa Indonesia is a National Language of Indonesia. It means Bahasa Indonesia must be used as an official language in all levels ranging from government to education as well as in development of science and technology. The Government of Republic of Indonesia as the highest and formal authority must use official Bahasa Indonesia in their activities including in their official websites. Therefore, the government issued a regulation instruction called Instruksi Presiden (Inpres) No. 2 Tahun 2001 to govern the usage of official computer terms in Bahasa Indonesia. The purpose of this research is to evaluate the usage of official computer terms in Bahasa Indonesia compared to the computer terms in English. The data are obtained from the government official websites in Indonesia. The method consists of data gathering, template detection, string extraction and data analysis. The evaluation of official computer terms in Bahasa Indonesia falls into three categories, such as good, moderate and poor. The number of websites in good category is 281 websites, the moderate category is 512 websites and the poor category is 290 websites. The authorized institution may use this result as additional information to evaluate the implementation of official information technology terms in Bahasa Indonesia.

1. Introduction

The lack of computer applications that using Bahasa Indonesia has been identified as the cause of the delay in mastered the computer usage for some people in Indonesia. The government of Indonesia had issued the instruction called Instruksi Presiden (Inpres) No. 2 tahun 2001. This regulation is intended to govern the usage of Bahasa Indonesia instead of English for common terms in information technology [1]. Along with the regulation, the government also issues the official computer terms in Bahasa Indonesia called Senarai Padanan Istilah (SPI). SPI consists of 629 terms in Bahasa Indonesia to substitute the commonly used in English. For the examples, the terms such as “unggah”, “unduh”, and “peramban” should be used instead of the terms “upload”, “download”, and “browser” respectively.

Currently, many Indonesia's central and local government have developed official websites as a public service through a network of communication and information [2]. Central and local government use websites as a medium to provide information related to their duties and functions to the society. The official website of government in Indonesia including the central, local and provincial government are marked using the domain go.id in the URL address. The growth of the government official websites in Indonesia has increased. In the year 2005, Indonesia's government has only 564 official websites, in 2016 the number of the websites that using the go.id domain have been increased significantly and reached over 1200 websites. However, until 2016, there has been no mechanism to evaluate whether this government regulation has been successfully socialized to the society. The study aims to evaluate the implementation of Inpres No. 2 Tahun 2001 in Indonesian government official websites.

The result of the study is the distribution percentage of the usage of Bahasa Indonesia official terms compared to English terms in the Indonesian government official websites. The study also yields the most frequently used terms in template of the Indonesian government official websites. The results can be used by Badan Pengembangan dan Pembinaan Bahasa under the Ministry of Education and Culture of the Republic of Indonesia for further socialization and dissemination of the regulation in Inpres No. 2 Tahun 2001 about the official computer terms in Bahasa Indonesia.

2. Related Works

Previous research regarding the evaluation of official computer terms usage in Bahasa Indonesia has been done in 2015 [3]. This research was a linguistic study to observe the responses of the students in Surakarta for the Inpres No. 2 Tahun 2001 about the proper usage of computer terms in Bahasa Indonesia to substitute English. The research used questionnaire method to evaluate the computer terms usage, meanwhile this research evaluates the official government websites and collects data from the Internet.

The other researches about the evaluation of official government website toward e-government paradigm have been conducted [4] [5] [6]. However, these researches do not explicate the usage of official computer terms in Bahasa Indonesia for evaluated websites. Our research might be used as an additional parameter to the evaluation method.

Template detection and data extraction from web sources had been done by [7] [8] [9]. Some works utilized two or more HTML pages at a time in order to find similarities and differences between structure and content page [7] [9]. There is also work with the Document Object Model tree, rather than with raw HTML markup [10]. Meanwhile, another study [11] detected a content segment as the largest body of text on a webpage by counting the number of words. Moreover, the Site-level method is used to detect templates based on several pages from a site [7] [12]. If a particular item is repeated many times, it is counted as a template. The Page-level methods detect templates based on a single page [8] [13]. A page is taken as input and decision on templates is made based on a certain similarity criteria or a threshold value based on the technique used. In this research, we used the page-level method that detect templates based on single page.

3. Methodology

The purpose of our work is to evaluate the usage of official information technology terms in Bahasa Indonesia terms compared by the usage of English terms in Indonesian Government official websites. Calculation is needed to know the ratio quantity of terms in Bahasa Indonesia compare to the quantity of terms in English. We named this calculation as P_{indo}. To get the calculation of P_{indo} for each website, we have to extract all the boilerplate texts from the template of Indonesian government's websites. Many works and techniques are adopted to do web data extraction, such as data gathering, template detection, data extraction. All the steps from our work is described in figure 1.

First of all, URLs seed is need to be determined. URLs seed contains the URL addresses that will be crawled. The list of the URL was obtained by utilizing the search engine. The search results were then logged manually to a list. This process produced 1,275 URLs seed of Indonesian government

websites. After collecting the URLs, we crawled all the URL addresses based on the sequences of the URLs seed. We utilized a web crawler to crawl the website automatically. For every URL, we only downloaded the first page of the website as it was the only page needed. The process of crawling yielded 1,178 pages or about 92.4% of the list. As many as 96 URLs or 7.6% were not successfully crawled because of various reasons such as: the URL was defunct, internal server error, or the first page only consists of images or redirect scripts. All the downloaded pages were stored to local database.

The next step is template detection. Template detection is a process to distinguish template and main content from a HTML page. To get the template of each downloaded page, we parsed the pages and divided them into multiple blocks. These blocks were divided by using HTML tags such as <p>, <td>, <th>, <h1>, <h2>, <h3>, <h4>, <h5>, <h6> and <li> as separators. Afterwards, in each block we counted the text in <a> tags to determine the link density. In case the link density was high (> 0.1), then the block was flagged as a template block. Other blocks were then checked using their string length. If the length was lower than 70 characters, then it was flagged as template block, otherwise it was just a regular block.

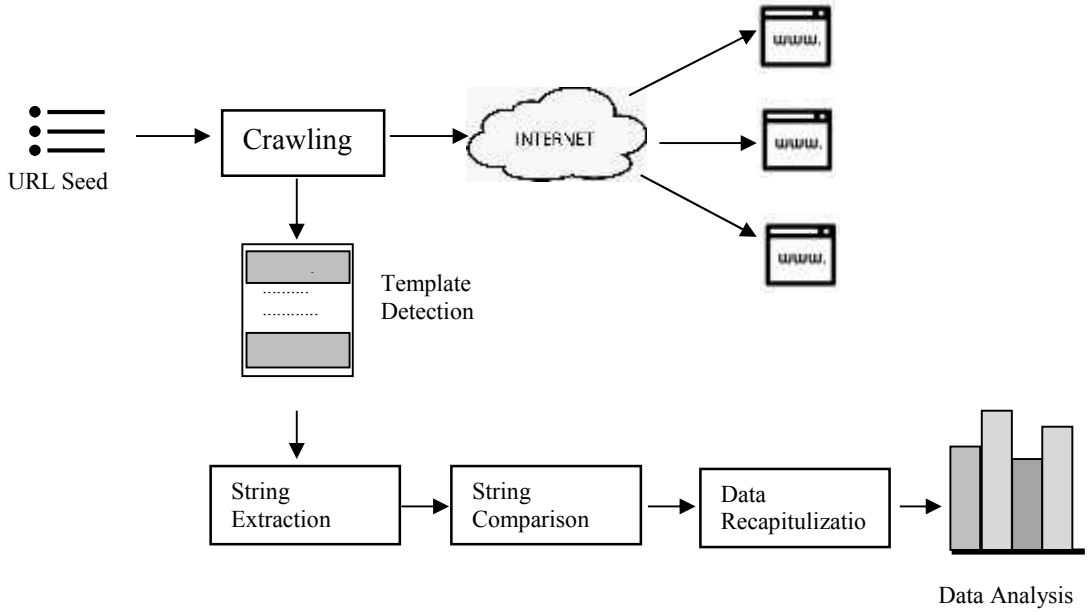

Figure 1. General Architecture

After detecting the template, each of these template blocks is extracted into tokens or strings/terms. In this step, all HTML tags are removed and the strings will be kept. All of the government's websites had represented into strings in this step. These processes produced 85,934 strings.

The next step is calculating the P_{indo} value of each page. Calculation P_{indo} value is done by comparing the string that exists in SPI list. The comparison only applies on the words in computer field and ignore the others. There are 94 websites were excluded from the experiments, because the template of these websites did not contain the words in SPI list.

After removing 94 websites from the experiment, the total number of the websites to be observed was 1084. We found that the lowest result of P_{indo} was 0 and the highest result of P_{indo} was100. The value 0 means the website does not use official computer terms in Bahasa Indonesia in its website. Conversely, value 100 means the website uses official computer terms in Bahasa Indonesia. Therefore, the higher P_{indo} means usage of official computer terms Bahasa Indonesia is better.

The next step is data recapitulation. In this step the websites are classified based on P_{indo}. For websites with the same P_{indo} were classified as one class. To simplify the number of classes, the generated class interval based on class interval by Sturges. The objective of this classification is to make data analysis easier to read. We also, divided Let C is number of class interval, R is the range of the lowest value of ratio to the highest ratio value, and N is number of observed website, so the equation to find C can be described in equation 1 [14].

$$C = \frac{R}{1+3.332\,Log\,N} \quad (1)$$

By the calculation, we got 11 class interval with value range 9 for each class. This class interval is shown in table 1. Based on this calculation, we determine 3 values interval of government websites in implementing official computer terms in Bahasa Indonesia. These values are "good", "moderate" and "poor". We implemented quartile calculation to determine the boundary of these value intervals. The poor category for $P_{indo} < 66.8$, the moderate category for P_{indo} between 66.8 to 78.2 and good category for $P_{indo} > 88.4$.

Table 1. Percentage Usage of Official Computer Terms in Bahasa Indonesia

Class	Number of Websites	P_{indo}
1	35	0 – 8
2	4	9 – 17
3	8	18 – 26
4	12	27 – 35
5	15	36 – 44
6	36	45 – 53
7	80	54 – 62
8	171	63 – 71
9	243	72 – 80
10	237	81 – 89
11	243	90 - 100
Total	1084	

4. Results and Discussions

According to calculation on the data recapitulation step, we found 290 websites out of 1084 Indonesian government websites is in poor category, 512 websites is in moderate category and 281 websites is in good category in implementation of official computer terms in Bahasa Indonesia in their websites. The number of good category should be increase in consider the regulation of using Bahasa Indonesia term than English language has been established from 2001. This phenomenon should need further evaluation by the authorized. Figure 2 shows the ratio of percentage of this category.

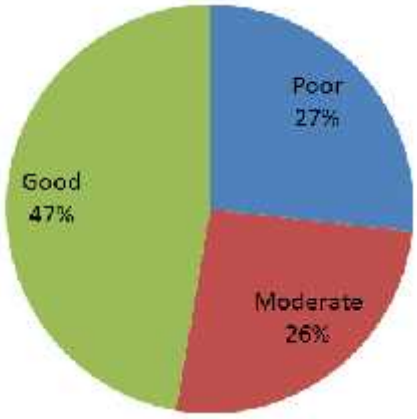

Figure 2. Percentage of The Websites Category

The comparison of English words number and Bahasa Indonesia words number on the Indonesian government websites can be seen in Figure 3. The English terms that are more widely used than the official computer terms in Bahasa Indonesia are "download", "file", "password", "master file", "file", "user" and "fax". Some words such as "password" and "master file" is much more popular than its substitute in Bahasa Indonesia. The most frequent English term are "download". This word is still far more popular than the word "unduh". About 342 websites prefer to use the term "download" and only 34 websites use the term "unduh". Especially for "server" there are two substitutes in Bahasa Indonesia, "peladen" and "server". From the string extraction recapitulation, we find out that "beranda" as Bahasa Indonesia is more widely used in Indonesian Government sites rather than "home". According to the glossary from Presidential Instruction No 2 of 2001, the substitute of "home" is not "beranda", but "pangkal". This finding can be used as an input for the authorized to put "beranda" as the substitute of "home" in the websites. Other Indonesian terms that are more popular than their English terms are "surat" and "hidup".

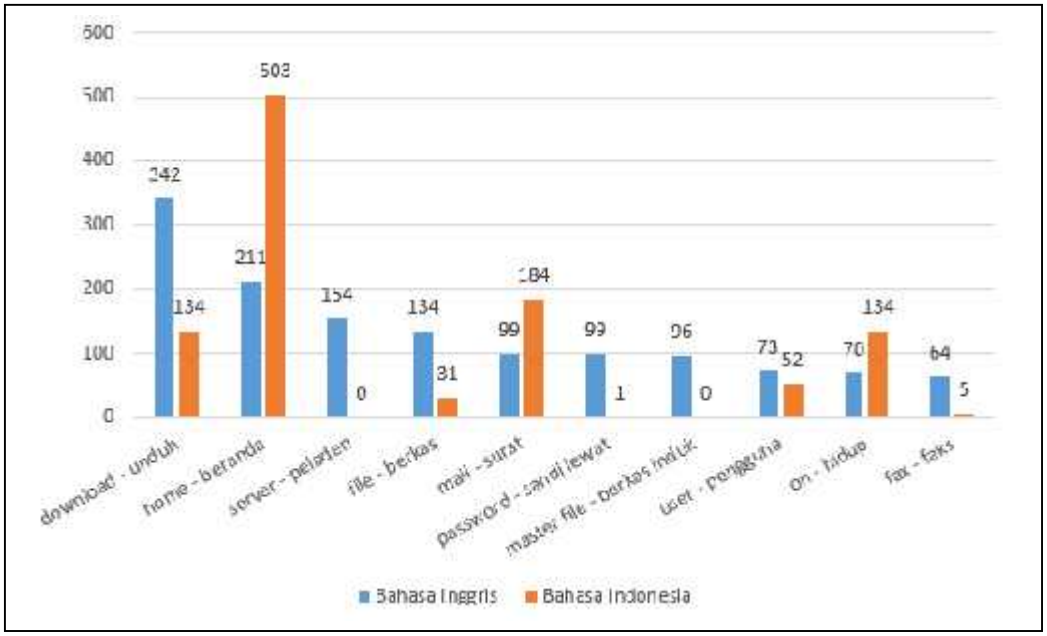

Figure 3. Top Ten Terms Distribution in Indonesian Government Websites

5. Conclusions

The number of websites in good category in the implementation of official computer terms in Bahasa Indonesia is 281 websites, the moderate category is 512 websites and the poor category is 290 websites. The poor category should be decreasing, considering that the regulation has been established from 2001. In addition, the term "*home*" is substitute to "*pangkal*" instead of "*beranda*", although the word "*beranda*" is commonly used terms to indicate the first page of the website. This finding might be used as consideration to determine the proper and common word in Bahasa Indonesia to substitute the English terms. These results need further evaluation by the authorized in order to spread the usage of official computer terms in Bahasa Indonesia more widely.

Acknowledgments

The authors gratefully acknowledge that the present research is supported by Ministry of Research and Technology and Higher Education Republic of Indonesia. The support is under the research grant BP-PTN USU of Year 2016 Contract Number 190/UN5.2.3.1/PPM/SP/2016.

References

[1] Government of Indonesia Presidential Instruction No 2 Year 2001.

[2] Hasibuan Z. A 2007, Langkah-langkah Strategis dan Taktis Pengembangan e-Government Untuk Pemda, *Jurnal Sistem Informasi MTI UI,* **3**.

[3] Sari C. A 2015, Tanggapan Mahasiswa dI Kota Surakarta Terhadap Pengindonesiaan Istilah Asing Bidang Pengkomputeran (Kajian Sosiolinguistik).

[4] Hermana B and Silfianti W 2011, Evaluating E-government Implementation by Local Government: Digital Divide in Internet Based Public Services in Indonesia, *International Journal of Business and Social Science,* **2**.

[5] Sosiawan E. A 2008, Tantangan dan Hambatan dalam Implementasi E-Government dI Indonesia, *Seminar Nasional Informatika,* **1**.

[6] Nurdin N, Stockdale R and Scheepers H 2014, The Role of Social Actors in the Sustainability of E-government Implementation and Use:Experience from Indonesian Regencies, *47th Hawaii International Conference on System Science*.

[7] Bar-Yossef Z and Rajagopalan S 2002, Template Detection Via Data Mining and Its Applications, *Proceedings of The 11th International Conference on World Wide Web*.

[8] Kohlschütter C, Fankhauser P and Nejdl W 2010, Boilerplate Detection Using Shallow Text Feature, *Proc*. *Proceedings of third ACM international conference on web search and data*, (New York: ACM)

[9] Crescenzi V and Mecca G 2004, Automatic Information Extraction from Large Websites, *Journal of the ACM,* **vol. 51**, no. 5, p. 731–779.

[10] Gupta S, Gail K, Neistadt D and Grimm P 2003, DOM-based Content Extraction of HTML Documents, *Proc. of the 12th international conference on World Wide Web*.

[11] McKeown K. R, Hatzivassiloglou V, Barzilay R, Schiffman B, Evans D and Teufel S 2001, Columbia Multi-Document Summarization: Approach and Evaluation, *Proc. of the Document Understanding Conference (DUC01)*.

[12] Kim C and Shim K 2011, Text: Automatic template extraction from heterogeneous web pages, *IEEE Transactions on knowledge and data Engineering ,* **vol. IV**, no. 23, pp. 612-626.

[13] Wang Y, Bingxing F, Cheng X, Guo L and Xu H 2008, Incremental Web Page Template Detection, in *ACM*, Beijing.

[14] Sturges H. A 1926, The Choice of a Class Interval, *Journal of the American Statistical Association*, **vol 21**.