Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 81 (2016) 208 – 213

Procedia
Computer Science

5th Workshop on Spoken Language Technology for Under-resourced Languages, SLTU 2016,
9-12 May 2016, Yogyakarta, Indonesia

Dictionary-based Word Segmentation for Javanese

Dipta Tanaya, Mirna Adriani*

Universitas Indonesia, Depok, 16424, Indonesia

Abstract

Word segmentation is the first step to process language that written in non-Latin letters such as such as Javanese script. In this study, we report our work on word segmentation based on dictionary approach. In the first phase, we generate all possible segmented word series using a word dictionary. The correct word is selected based on the last character in a word, the last two characters in a word, the difference of two consecutive words, and the frequency of the word in the additional corpus. The experimental results show that identifying words using the frequency of words in the additional corpus yield the best accuracy that is 91.08%.

Keywords: javanese character; word segmentation

1. Introduction

Javanese is one of the local languages exists in Indonesia. According to www.ethnologue.com, the number of its speaker is 84.3 million [1]. There is around more than eighty million people who speak Javanese which is the biggest spoken language in Java Island. Javanese has non-Latin characters which are known as Hanacaraka Javanese script. Javanese script is considered as abugidas, which is a kind of a way of writing characters by combining consonants and vowels into a single syllable in a unit[2]. The Javanese script has some punctuation marks such as commas, period, colon, and other characters to mark the end of the sentence.

Generally the non-latin characters do not contain any delimeter to separate words so we need to develop an approach to recognize the words. For example, in the sentence “*aku maca koran”* or “I read a newspaper”, written

* Corresponding author.
E-mail address: mirna@cs.ui.ac.id

doi:10.1016/j.procs.2016.04.051

as (ꦲꦏꦸꦩꦕꦏꦺꦴꦫꦤ꧀) A*ku* or "I" (ꦲꦏꦸ), *maca* or read (ꦩꦕ), and *koran* (ꦏꦺꦴꦫꦤ꧀) or the newspaper, are not separated by a space.

As mentioned in[3], in the area of Natural Language Processing (NLP), prior to further processing, text that is written in non-Latin characters need to be to be split into units of words. This process is known as segmentation stage. Applications that require this process include the automated machine translation, POST Tagger, sentence parser etc. Developing a segmentation method for a non-Latin text faces a challenge such as the ambiguity problem. The ambiguity occurs when we need to segment words or sentences because there can be more than one possible segmentation.

Research on word segmentation in foreign languages such as Japanese, Chinese, Thai language, has been done with various techniques. Haruechaiyasak et al[3] compare some word segmentation techniques in Thai such as dictionary based approaches (longest matching algorithm and maximal matching) and machine learning based approaches (Naive Bayes (NB), decision tree, Support Vector Machine (SVM) and Conditional Random Field (CRF)). The results show that the best performance is achieved using CRF algorithm, with precision 95.79% and recall 94.98%. Then Tepdan, Haruechaiyasak, and Kongkachandra work further to improve the word segmentation algorithm in Thai[4] by combining Named Entity Recognition (NER) algorithm on the TLex (Thai Lexeme Analyzer) model. Norbu et al[5] conduct a research on the word segmentation in Dzongkha. In this study they use a combination of maximal matching algorithm and bigram techniques. The best accuracy of this study is 91.5% using 714 words. Bi and Taing conduct a research on word segmentation in Khmer. In this study they use Bidirectional Maximal Matching (BIMM) which is a variation of the maximum matching algorithm[6]. Some work on Chinese word segmentation also have done[7,8,9,10, 11].

An earlier work on Javanese word segmentation has been done by Prabantoro [12] using a dictionary and character statistics, such as the number of characters on the longest word, the number of characters on the shortest word, and the average of word's length.

In this study we work on ~~a~~ Javanese word segmentation algorithm based on a dictionary. In addition, it is intended to overcome the deficiencies that exist in previous research[12]. In the previous study [12] it uses some parameter values from dictionary to modify segmented words. Thus, the algorithm's performance will depend on those parameters which derived from the dictionary.

2. Javanese Word Segmentation Approach

2.1. Character Statistics

We begin our study with analyzing the character statistics using a corpus written in Javanese and Latin characters. The statistical analysis is used to develop the word segmentation algorithm for Javanese. The corpus contains 3.164 articles (it has 2.576.821 words) from *Panjebar Semangat*[†].

Table 1 shows the number of words that end with consonant or vowel. We calculate the percentage of number of words that end with consonant by

$$Percentage = \frac{\#words\ ended\ by\ a\ character\ type}{\#words\ in\ corpus} \qquad (1)$$

For example, we calculate percentage of consonant-ended character. Thus, it will be 1.385.405 (total of consonant-ended words) divided by 2.2576.821 total words in the corpus. The result is 54.76% (0,5476).

Table 1 Proportion of last character in corpus written in Latin Character

Last character	#words	Percentage
Consonant	1.385.405	53,76%
Vowel	1.191.416	46,23%

[†] http://panjebarsemangat.co.id