Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 35 (2014) 464 – 473

Procedia Computer Science

18th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems - KES2014

Extraction Japanese slang from weblog data based on script type and stroke count

Kazuyuki Matsumoto*, Kyosuke Akita, Xielifuguli Keranmu, Minoru Yoshida, Kenji Kita

The University of Tokushima, Minami Josanjima-cho 2-1, Tokushima City, 770-8506, Japan

Abstract

Young people commonly use slang in the texts for weblogs or Social Networking Sites. How to treat such slang words properly is one of the problems in the field of text mining. In this paper, we examined several methods to extract Japanese slang called "*Wakamono Kotoba*," which is particularly used by young people, by focusing on its script type and stroke count. In the evaluation experiment, a high precision was obtained when we adopted script type for extraction.

Peer-review under responsibility of KES International.
Keywords: Japanese slang; *Wakamono Kotoba*; Conditional Random Fields(CRF); script type; stroke count.

1. Introduction

Recently, availability of smart phones and tablet type PCs has expanded with an explosive pace, increasing the number of users of Social Networking Site (SNS), bulletin board and weblog sites. Smart phone is being used particularly by young people from the teenagers to those in their thirties, making weblog or SNS more popular among those young people.

Young people tend to use unique words distinctive to their generations (i.e. "*Riaju* (having fruitful life)", "*Komyusho* (having trouble in communication)," etc.), which are called *Wakamono Kotoba* in Japanese. They often use such *Wakamono Kotoba* also in the communication on weblog or SNS. Because *Wakamono Kotoba* is a word closely connected with daily life and is subject to change of trend, it is generated daily but falls into disuse in short period. Therefore, it is difficult to construct a dictionary that covers *Wakamono Kotoba*. Besides, some *Wakamono Kotoba* are used only among a limited communication group with a specific interest. And, people who do not belong to the communication group sometimes cannot understand the meaning of *Wakamono Kotoba* used in the group.

In the recent text mining technique that uses a collective wisdom, social data on Web are being used to identify human relationships or to conduct opinion mining or reputation analysis. However, in these data, slang is often included. It is considered that slang causes errors to the basic language analysis tools for morphological analysis,

* Corresponding author. Tel.: +0-88-656-7654.
E-mail address: matumoto@is.tokushima-u.ac.jp

Peer-review under responsibility of KES International.
doi:10.1016/j.procs.2014.08.127

syntactic parsing or semantic analysis, badly influencing the mining accuracy. In this paper, we focused on Japanese slang of *Wakamono Kotoba* on Web. We discussed on usage and application of *Wakamono Kotoba* in text mining, and proposed a method to extract *Wakamono Kotoba* automatically from a text data.

There were various approaches that have been made to extract automatically unknown words such as slang. Asahara et al.[4] proposed the character-based method to identify Japanese unknown word. Ling et al.[5] also proposed the character-based tagging and chunking method for Chinese unknown word. Ritter et al.[6] identified named entity by using LabeledLDA. Tsuchiya et al.[7] proposed the judgment method of alphabet abbreviation word that are not registered into a dictionary by using association mechanism.

Hadi et al.[8] constructed a slang dictionary for opinion words, and, Taysir et al.[9] also constructed an emotion polarity dictionary by automatically extracting slangs from arabic text using Support Vector Machine. Murawaki et al.[10] researched how to acquire unknown words automatically without manual supervision and register these words into a morphological analysis dictionary.

Many researchers studied on the problems of unknown word. However, there are a few research aiming to extract slangs using slang-specific features like the stroke numbers. In this paper, we report the effect of using such slang-specific features on the accuracy of slang extraction.

2. Japanese slang(*Wakamono Kotoba*)

Wakamono Kotoba(*WK*) is a Japanese language particularly used among the younger generation from teenagers to those in their late twenties[1]. They tend to use *WK* especially on Internet. Some examples of *WK*, their original words and meanings are listed below.

- *Kebai* . . . *Kebakeashii* (floozy)
- *Kuripa* . . . *Kurisumasu party* (Christmas party)
- *Dotakyan* . . . *Dotanba Cancel* (last-minute cancellation)

As described in the previous section, *WK* is sometimes limitedly used among a certain group of communication. There are varieties of such groups and varieties of WK used in the group, making it very difficult to collect all *WK*.

There are a lot of researches on how to process unknown words. However, there are a few that focused on *WK* because of the following reasons:

1. Varieties of notations
2. Short duration of use

It is hard to obtain its sense or usage because *WK* has many different notations and it is used relatively for short period. As the result, it is difficult to create a dictionary of *WK* by manually.

Matsuo et al.[2] proposed a method to extract *WK* written in *Katakana* by preparing a template: "word prior to *WK*" + "*WK*" + "word after *WK*," and by matching words with the template. As the result of the evaluation experiment, this method obtained 42.4% accuracy of extraction of *WK* written in *Katakana*. They analyzed the randomly chosen 217 extracted *WK* and found that 45 unknown *WK* had been successfully extracted. However, in the character strings that they analyzed, spelling mistakes and proper nouns such as personal name were included 57.6%. They concluded that they need to reduce the noise such as proper noun or spell-miss and to consider template using sentence structure.

Mori et al.[3] assumed that if the words have the same part of speech, the words have similar character strings before and after the words. Based on this assumption, they proposed a method to extract a word from a corpus and estimate its part of speech at the same time. This method defined the environment of part of speech and character string as conditional probability distribution of backward and forward character strings of the target character string in the corpus. In the experiment, by considering frequency of the character strings, the precision of 96.8% was obtained when threshold was set at 0.1. Without considering frequency of character string, the precision of 86.2% was obtained. The method successfully extracted 268 unknown words and proved its effectiveness.