Available online at www.sciencedirect.com

SciVerse ScienceDirect

Computer Speech and Language 28 (2014) 38–55

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Automatically annotating a five-billion-word corpus of Japanese blogs for sentiment and affect analysis☆

Michal Ptaszynski[a,*], Rafal Rzepka[b], Kenji Araki[b], Yoshio Momouchi[c]

[a] *Department of Computer Science, Kitami Institute of Technology, Japan*
[b] *Graduate School of Information Science and Technology, Hokkaido University, Japan*
[c] *Department of Electronics and Information Engineering, Faculty of Engineering, Hokkai-Gakuen University, Japan*

Received 7 August 2012; received in revised form 15 April 2013; accepted 24 April 2013
Available online 18 May 2013

Abstract

This paper presents our research on automatic annotation of a five-billion-word corpus of Japanese blogs with information on affect and sentiment. We first perform a study in emotion blog corpora to discover that there has been no large scale emotion corpus available for the Japanese language. We choose the largest blog corpus for the language and annotate it with the use of two systems for affect analysis: ML-Ask for word- and sentence-level affect analysis and CAO for detailed analysis of emoticons. The annotated information includes affective features like sentence subjectivity (emotive/non-emotive) or emotion classes (joy, sadness, etc.), useful in affect analysis. The annotations are also generalized on a two-dimensional model of affect to obtain information on sentence valence (positive/negative), useful in sentiment analysis. The annotations are evaluated in several ways. Firstly, on a test set of a thousand sentences extracted randomly and evaluated by over forty respondents. Secondly, the statistics of annotations are compared to other existing emotion blog corpora. Finally, the corpus is applied in several tasks, such as generation of emotion object ontology or retrieval of emotional and moral consequences of actions.

Keywords: Emotion corpora; Corpus annotation; Sentiment analysis; Affect analysis

1. Introduction

There is a lack of large corpora for Japanese applicable in sentiment and affect analysis. Although there are large corpora of newspaper articles, like Mainichi Shinbun Corpus[1], or corpora of classic literature, like Aozora Bunko[2], they are usually unsuitable for research on emotions since spontaneous emotive expressions either appear rarely in these kinds of texts (newspapers), or the vocabulary is not up to date (classic literature). Although there exist speech corpora, such as Corpus of Spontaneous Japanese[3], which could become suitable for this kind of research, due to the difficulties with compilation of such corpora they are relatively small. In research such as the one by Abbasi and Chen

☆ This paper has been recommended for acceptance by Prof. R. K. Moore.
* Corresponding author. Tel.: +81 157 26 9327.
E-mail address: ptaszynski@cs.kitami-it.ac.jp (M. Ptaszynski).
[1] http://www.nichigai.co.jp/sales/mainichi/mainichi-data.html.
[2] http://www.aozora.gr.jp/.
[3] http://www.ninjal.ac.jp/products-k/katsudo/seika/corpus/public/.

0885-2308/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.csl.2013.04.010

(2007) it was proved that public Internet services, such as forums or blogs, are a good material for affect analysis because of their richness in evaluative and emotive information. One kind of these services are blogs, open diaries in which people encapsulate their own experiences, opinions and feelings to be read and commented by other people. Recently blogs have come into the focus of opinion mining or sentiment and affect analysis (Aman and Szpakowicz, 2007; Quan and Ren, 2010). Therefore creating a large blog-based emotion corpus could help overcome both problems: the lack in quantity of corpora and their applicability in sentiment and affect analysis. There have been only a few small Japanese emotion corpora developed so far (Hashimoto et al., 2011). On the other hand, although there exist large Web-based corpora (Erjavec et al., 2008; Baroni and Ueyama, 2006), access to them is usually allowed only from the Web interface, which makes additional annotations with affective information difficult. In this paper we present the first attempt to automatically annotate affect on YACIS, a large scale corpus of Japanese blogs. To do that we use two systems for affect analysis of Japanese, one for word- and sentence-level affect analysis and another especially for detailed analysis of emoticons, to annotate on the corpus different kinds of affective information (emotive expressions, emotion classes, etc.).

The outline of the paper is as follows. Section 2 describes the related research in emotion corpora. Section 3 presents our choice of the corpus for annotation of affect- and sentiment-related information. Section 4 describes tools used in annotation. Section 5 presents detailed data and evaluation of the annotations. Section 6 presents tasks in which the corpus has already been applied. Finally the paper is concluded and future applications are discussed.

2. Emotion corpora

Research on affect analysis has resulted in a number of systems developed within several years (Aman and Szpakowicz, 2007; Ptaszynski et al., 2009c; Matsumoto et al., 2011). Unfortunately, most of such research ends in proposing and evaluating a system. The real world application that would be desirable, such as annotating affective information on linguistic data is limited to processing a usually small test sample in the evaluation. The small number of annotated emotion corpora that exist are mostly of limited scale and are annotated manually. Below we describe and compare some of the most notable emotion corpora. Interestingly, six out of eight emotion corpora described below are created from blogs. The comparison is summarized in Table 1. We also included information on the work described in this paper for better comparison (YACIS).

Quan and Ren (2010) created a Chinese emotion blog corpus **Ren-CECps1.0**. They collected 500 blog articles from various Chinese blog services, such as sina blog (http://blog.sina.com.cn/), qq blog (http://blog.qq.com/), etc., and annotated them with a large variety of information, such as emotion class, emotive expressions or polarity level. Although syntactic annotations were simplified to tokenization and POS tagging, this corpus can be considered a state-of-the-art emotion blog corpus. The motivation for Quan and Ren is also similar to ours – dealing with the lack of large corpora for sentiment analysis in Chinese (in our case – Japanese).

Wiebe et al. (2005) report on creating the **MPQA** corpus of news articles. The corpus contains 10,657 sentences in 535 documents.[4] The annotation schema includes a variety of emotion-related information, such as emotive expressions, emotion valence, intensity, etc. However, Wiebe et al. focused on detecting subjective (emotive) sentences, which do not necessarily convey emotions, and classifying them into positive and negative. Thus their annotation schema, although one of the richest, does not include emotion classes.

A corpus of Japanese blogs, called **KNB**, rich in the amount and diversification of annotated information was developed by Hashimoto et al. (2011). It contains 67 thousand words in 249 blog articles. Although it is a rather small scale corpus, it developed a certain standard for preparing corpora, especially blog corpora for sentiment and affect-related studies in Japan. The corpus contains all relevant grammatical annotations, including POS tagging, dependency parsing or Named Entity Recognition. It also contains sentiment-related information. Words and phrases expressing emotional attitude were annotated by laypeople as either positive or negative. One disadvantage of the corpus, apart from its small scale, is the way it was created. Eighty-one students were employed to write blogs about different topics especially for the need of this research. It could be argued that since the students knew their blogs will be read mostly by their teachers, they selected their words more carefully than they would in private.

[4] The new MPQA Opinion Corpus version 2.0 contains additional 157 documents, 692 documents in total.