Chapter 16
Building Language Resources for Emotion Analysis in Bengali

Dipankar Das
National Institute of Technology (NIT), India

Sivaji Bandyopadhyay
Jadavpur University, India

ABSTRACT

Rapidly growing Web users from multilingual communities focus the attention to improve the multilingual search engines on the basis of sentiment or emotion and provide the opportunities to build resources for languages other than English. At present, there is no such corpus or lexicon available for emotion analysis in Indian languages, especially for Bengali, the sixth most popular language in the world, second in India, and the national language of Bangladesh. Thus, in the chapter, the authors describe the preparation of an emotion corpus and lexicon in Bengali. The emotion lexicon, termed Bengali WordNet Affect has been developed from its equivalent version in English by traversing the steps of expansion, translation, and sense disambiguation. In addition to emotion lexicon, a Bengali blog corpus for emotion analysis has also been developed by manual annotators with detailed linguistic expressions such as emotional phrases, intensities, emotion holder, emotion topic and target span, and sentential emotion tags.

INTRODUCTION

In recent times, research activities in the areas of Opinion, Sentiment, and/or Emotion in natural language texts and other media are gaining ground under the umbrella of subjectivity analysis and affective computing.

The Subjectivity Analysis is defined as classifying a given text (usually a sentence) into one of two classes: objective or subjective whereas Affective computing is an area of artificial intelligence that focuses on how emotion is expressed, perceived, recognized, processed, and interpreted in text, speech, dialogue, image, video etc.. Text based emotion analysis relies heavily on Natural Language Processing (NLP), which is mostly focused on understanding the semantics of text. By analyzing the texts and obtaining semantic as well as emotional information, the computer can deal with more interpersonal matters such as

DOI: 10.4018/978-1-4666-3970-6.ch016

understanding the relationships between people. Both affective computing and NLP are needed to reach this goal. NLP algorithms are necessary to understand the semantics or explicit message of text, while affective computing is needed to understand the implicit message in text manifested through emotion (Minato *et al.*, 2008).

The identification of emotional state from texts is not an easy task as emotion is not open to any objective observation or verification (Quirk *et al.*, 1985). Genuine opinion, emotion and sentiment are hard to collect, ambiguous to annotate, and tricky to distribute due to privacy reasons. Different forms of modeling exist, and ground truth is never solid due to the often highly different perception of the mostly very few annotators. Thus, the few available corpora suffer from a number of issues due to the peculiarity of these young and emerging fields.

In order to obtain knowledge and information from emotional text it is necessary to have reliable linguistic resources, such as tagged emotion corpora and emotion dictionaries. As the study of emotion recognition combined with natural language processing is rather new, it is still difficult to obtain such linguistic resources.

Among the social media like e-mails, Weblogs, chat rooms, online forums and even twitter, blog is one of the communicative and informative repository of text based emotional contents in the Web 2.0 (Lin *et al.*, 2007). Thus, we have prepared the emotion annotated corpus from Bengali blog documents.

The proposed corpus annotation task was carried out at sentence and document levels. Three annotators have manually annotated the blog sentences, which were retrieved from an open source Bengali Web blog archive (www.amarblog.com). Ekman's (1993) six basic emotion classes (*anger, disgust, fear*, *happy*, *sad* and *surprise*) were considered to accomplish our tasks. The emotional sentences are annotated with three types of intensities such as *high*, *medium* and *low* as well as the sentences of non-emotional (*neutral*) and multiple (*mixed*) categories were also identified. The emotional words and phrases were marked by fixing the lexical scope of the emotional expressions. Each of the emoticons is also considered as individual emotional expressions. The emotion holder and relevant topics associated with the emotional expressions were annotated by considering the punctuation marks, conjuncts, rhetorical structures and other discourse information whereas the knowledge of the rhetorical structure helps in removing the subjective discrepancies from the writer's point of view. The annotation scheme is used to annotate 123 blog posts containing 4,740 emotional sentences having single emotion tag and 322 emotional sentences for mixed emotion tags along with 7087 *neutral* sentences in Bengali. Three types of standard agreement measures such as Cohen's *kappa* (κ) (Cohen, 1960), Measure of Agreement on Set-valued Items (MASI) (Passonneau, 2004) and *agr* (Wiebe *et al.*, 2005) metrics were employed for the annotated emotion related components. It is observed that the relaxed agreement schemes like MASI and *agr* are specially considered for fixing the lexical boundaries of emotional expressions and topics in the emotional sentences. The inter annotator agreement of some emotional components such as sentential emotions, holders and topics show satisfactory performance whereas the sentences of mixed emotion and intensities of *medium* and *low* show the disagreement. We observed that a preliminary experiment for the word level emotion classification on a small set of the whole corpus yielded satisfactory results.

In this proposed chapter, we also would like to describe the preparation of the Bengali *WordNet Affect*, an emotion lexicon from its equivalent version already available in English (http://www.cse.unt.edu/~rada/affectivetext/). The collection of the *WordNet Affect* synsets was provided as a resource for the *SemEval*-2007 shared task of "Affective Text." The shared task was focused on text annotation by affective tags not from the whole *WordNet Affect* but a part of it being