Development of Corpus for Affect Analysis in Japanese

Michal Ptaszynski* Pawel Dybala* Rafal Rzepka* Kenji Araki*
*Department of Information Science and Technology, Hokkaido University
Sapporo, 060-0814, Japan
Tel: +81-11-706-7389, Fax: +81-11-709-6277
E-mail: {ptaszynski, paweldybala, kabura, araki}@ist.hokudai.ac.jp

Abstract— **This paper presents a design of a dialogue corpus required for the research on affect analysis in Japanese. The research on analyzing and estimating emotional states of users during their interaction with robots and generally perceived cyber-environment has been focused on developing sophisticated methods rather than gathering a relevant material for evaluation. In this paper we explain an urgent need for the change of this tendency and propose a design of a corpus meant to support research in this field. We also propose the first application of the corpus in the evaluation method proposed by us in the former research.**

I. INTRODUCTION

The research on mechanical processing of emotions, including analysis and recognition of affect, has been gathering popularity of researchers since being initiated a little over ten years ago [1]. Recent years have brought research on emotions into focus in Artificial Intelligence and Natural Language Processing [2].

However, after over ten years of development of the field, there have been no significant or reliable methods of objective evaluation for emotion recognition methods. Unreliable measures might put the fairness of the results in question or doubt, despite of intricate mechanisms used in the proposed recently methods for emotion estimation [3, 4]. To solve this problem Ptaszynski et al. [5] proposed Double Standpoint Evaluation Method (DSEM) - a method of evaluation designed especially for the affect analysis systems and showed that even with a small collection of material for evaluation the method is more objective than the popular methods used in the field today and in detail reveals which parts of a system need improvement. The method was designed to be the more objective the more participants took part in the creation of the evaluation material. However, in its primary shape, DSEM was based only on a small collection of utterances.

Below we briefly describe DSEM and present a design of a large corpus of dialogues with multifaceted annotations of emotive information to support the evaluation method and provide a reliable research material for the future research in the field of affect analysis.

II. DOUBLE STANDPOINT EVALUATION METHOD (DSEM)

DSEM is a method where a system for affect analysis is evaluated from two different standpoints: recognitive (the first person evaluation) and commonsensical (the third person evaluation). As a method aiming to be objective, it assumes that neither do people themselves understand their emotional states with 100% reliability, nor do other people perceive the emotional states of their interlocutors with a perfect accuracy. DSEM looks rather for a balance between these two approaches. In the method, the system is first evaluated on whether it can appropriately recognize the emotional states of users. After that, another evaluation is performed by an objective group of third party evaluators to check how much the system's procedures agree with human commonsense about other person's emotions. For both sets of results, the better situation is when the results of both first- and third-person evaluation are higher, and the more balanced the both results are.

This research is partially supported by a Research Grant from the Nissan Science Foundation and The Global Centers of Excellence Program founded by Japan's Ministry of Education, Culture, Sports, Science and Technology.

III. PREVIOUS EVALUATIONS

In the former research the minimal unit of interest was one whole utterance. The evaluation was based on a collection of 90 natural utterances gathered through an anonymous survey. In the survey participated 30 people of different ages and social groups. Each of them was to imagine or remember a conversation (or conversations) with any person (or persons) they know and write three sentences from that conversation: one free, one emotive, and one non-emotive. After that the participants tagged their own utterances with emotive information, such as whether or not an utterance was emotive, or what were the specific emotion types conveyed in the emotive utterances. Next, the utterances were annotated in the same manner by a group of third party annotators.

An evaluation experiment of ML-Ask, a system proposed by Ptaszynski et al [6], using DSEM conducted with the collection of utterances mentioned above, showed that the evaluation method is more objective than the popular methods used in the fields today [5]. DSEM is also potentially capable to reveal in detail which parts of a system for affect analysis need improvement. In an evaluation experiment of ML-Ask supported with Web mining [7] it was also possible to specify, which of the four versions of the system's procedure is the most balanced.

IV. CORPUS FOR EVALUATION OF AFFECT ANALYSIS SYSTEMS

For the deep evaluation provided by DSEM there is a need for an appropriate large corpus with multi-faceted annotations of emotive information. However for utterances taken out of context it is difficult for the third party annotators to decide about presumable emotive information. Therefore, aside of gathering new utterances, there is a need to gather a corpus of dialogues and continue annotating using not only the separate sentences, but the whole conversations.

A. General Assumptions for the Corpus Design

As a general premise for the design of the corpus, the more participants take part in creating it the more accurate become the annotations of emotive information. Constructing of the corpus is divided into three parts. Firstly, we gather a large collection of the raw data of conversations and process them according to the needs of a certain research. The raw data of conversations can be recorded as audio data or audio-video data. However, since the usefulness of the corpus in the future research will strongly depend on its universality and richness of information in contains, it is strongly required that the conversations were recorded as the latter, including as far as possible also other information, such as changes of blood pressure, or fluctuations of brain waves, since methods for estimating emotional states using information such

as biometric data has also been proposed lately [8]. The ML-Ask system analyses the textual layer of speech, therefore for the need of this research the audio data will have to be converted into text.

B. Preparing the Conversations

Although emotions are expressed in dialogues as well as in conversations between more than two people, the systems for estimating emotions are usually designed to be implemented in robots interacting with their users in natural dialogues. Therefore the conversations should be planned as dialogues. In one conversation set two people will perform a dialogue for either a specified amount of time (e.g. 3 minutes), or till they decide that that the conversation is finished. Every participant will take part in three conversations. The first conversation will be a free conversation not charged with any presupposed emotions. Before the second and the third conversation the participants will be induced with positive and negative emotions respectively using such methods as IAPS [9] or by having a short conversation with a psychologist assistant to remind in the participant of their pleasant or unpleasant experiences from the past. After the conversations, the recorded data will be prepared for the annotation process.

C. Three Levels of Annotations

The participants will annotate the script of the conversation adding emotive tags to the utterances of the conversations. In the first person annotation, the participants will annotate mainly their own utterances. In the third person annotation every participant will annotate all of the utterances of a conversation.

The annotation process will be performed in three stages. First, the annotators will decide whether an utterance is emotive or not. Next, for the utterances described as emotive they will decide how strong were the emotions conveyed in the utterance, or set the emotive value (on a scale 0-5). Finally, emotive utterances will be annotated with specific emotion types.

D. Classification of Emotion Types

As markers for the specific emotion type annotations we use Nakamura's classification of emotions [10]. Nakamura, after a thorough study on emotions in Japanese, proposed a classification of emotions into 10 types - said to be the most appropriate for the Japanese language. That is: 喜 (*ki, yorokobi* – joy, delight), 怒 (*do, ikari* – anger), 哀 (*ai, aware* – sorrow, sadness), 怖 (*fu, kowagari* – fear), 恥 (*chi, haji* – shame, shyness, bashfulness), 好 (*kou, suki* – liking, fondness), 厭 (*en, iya* – dislike, detestation), 昂 (*kou, takaburi* – excitement), 安 (*an, yasuragi* – relief) and 驚 (*kyou, odoroki* – surprise, amazement).

E. First Person Annotations

Since too long time gap between the conversations and the annotation might cause the participants to forget about the emotions they had and influence the annotation process, it is desirable that the data was prepared shortly after the conversations or during the conversations in the real time. The participants will add the emotive annotations to the conversations they took part in with a special attention paid on their own utterances. This process corresponds to the self-assessment interviews performed widely in psychology.

F. Third Person Annotations

In the second part of the annotation process a group of third party annotators, unrelated to the conversation participants, will annotate the emotive information to all utterances from the conversation scripts. In this part, the more participants evaluates one script the more vivid will be the tendencies in the general commonsense (see section II) about the emotive load of the conversations. In former experiments the minimum number of participants was 8, and the average was 10 annotators per one evaluation unit (before – one utterance, now – one conversation). Although in other research in the field, even in the ones presented on COLING [4] or in scientific journals [3] the researchers tend to use only few participants or sometimes even one, after our hitherto research we noticed that at least 8 and desirably 10 or more participants should take part in the third party annotation process. The general commonsense will be calculated as an inter-agreement between all the annotators.

G. Application of the Corpus in Evaluation of ML-Ask

The gathered and fully annotated corpus will be put into practice in the proposed evaluation method to perform a thorough evaluation of the ML-Ask system. ML-Ask will analyze the corpus and the results of the system will be compared to the annotations provided by, first - the authors of the utterances and second - the third party evaluators.

V. Conclusions and Future Work

In this paper we presented a design of a natural dialogue corpus thoroughly annotated with information about emotional states conveyed in the conversations. The annotation is performed firstly by the conversation participants and secondly by a third party annotators and consist of the three level emotive descriptions of the utterances appeared in the conversations. First, whether the utterance is emotive or not; second, about the strength of the conveyed emotions; and third, about the particular emotion types. The corpus tagged this way is to be used in deep evaluation of affect analysis systems. For the first application of the corpus we propose its use as a base for DSEM - the evaluation method proposed in our former research. The method supported with the corpus will be used to perform a thorough evaluation of ML-Ask - the system for affect analysis constructed by us. Although in the first application we plan to use only the textual layer of the conversations in the corpus, in the process of gathering the corpus we plan to gather the audio as well as video information and other information, such as biometric data for the need of future research in the field of affect analysis.

References

[1] R. W. Picard, "Affective Computing." *MIT Technical Report #321*, MIT Media Laboratory. 1995.

[2] James G. Shanahan, Yan Qu, Janyce Wiebe (Eds.), *Computing Attitude and Affect in Text: Theory and Applications*, Springer, 2006.

[3] K. Matsumoto, F. Ren, S. Kuroiwa, S. Tsuchiya, "Emotion Estimation Algorithm Based on Interpersonal Emotion Included in Emotional Dialogue Sentences", *LNCS 4827*, pp. 1035-1045, 2007.

[4] R. Tokuhisa, K. Inui, Y. Matsumoto, "Emotion Classification Using Massive Examples Extracted from the Web", *Proceedings of Coling 2008*, pp. 881–888, 2008.

[5] M. Ptaszynski, P. Dybala, R. Rzepka and K. Araki, "Double Standpoint Evaluation Method for Affect Analysis System." *Proceedings of JSAI*, 2008.

[6] M. Ptaszynski, P. Dybala, R. Rzepka and K. Araki, "Effective Analysis of Emotiveness in Utterances Based on Features of Lexical and Non-Lexical Layers of Speech." *Proceedings of NLP14*, pp.171-174. 2008

[7] M. Ptaszynski, P. Dybala, S. Wenhan, R. Rzepka and K. Araki, "How to find love in the Internet? Applying Web mining to affect recognition from textual input", *Proceedings of 2008 EMALP Workshop, PRICAI'08*, pp. 67-79, 2008.

[8] J. Teixeira, V. Vinhas, E. Oliveira, L. P. Reis, "A New Approach to Emotion Assessment Based on Biometric Data", *HAI 2008 Workshop, in Proceedings of WI-IAT'08*, pp. 505-511, 2008.

[9] P. Lang, M. Bradley, "International Affective Picture System (IAPS): Affective rating of pictures and instruction manual." *Technical Report A-6*. 2005.

[10] A. Nakamura, *Kanjo hyogen jiten* [Dictionary of Emotive Expressions] (in Japanese), Tokyodo Publishing, Tokyo. 1993.