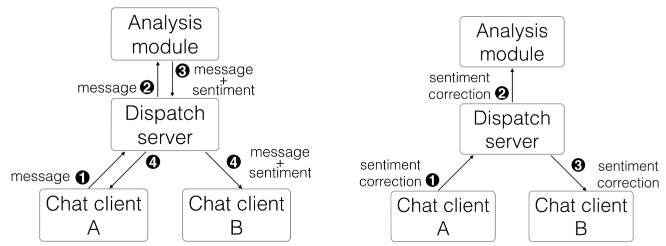
Chatmood: A Mixed-initiative Labeling Approach to Building Natural Dialogue Corpus for Sentiment Analysis

Chi-Chia Huang[1], **Yi-Ching Huang**[1] and **Jane Yung-jen Hsu**[2,3]
[1]Graduate Institute of Networking and Multimedia
[2]Department of Computer Science and Information Engineering
[3]Intel-NTU Connected Context Computing Center
National Taiwan University
{d01944003, d00944010, yjhsu}@csie.ntu.edu.tw

Abstract

Natural dialogue corpus is a useful resource in sentiment analysis research. Scientists could make computer better understand human intension with it, but it is difficult to obtain. The ordinary way to build a corpus is to recruit experts or workers to label the collected conversion log. However, experts or workers may not truly understand the sentiment contained in the log. Furthermore, some feelings are hard to reconstruct after the moment passed. In this work, we propose a mix-initiative labeling approach which is designed to let people contribute labels during the conversation with minimum efforts. We also conducted a pilot study to examine the design for further improvement.



(a) Data flow of messages and sentiment classification results.

(b) Data flow of user-made sentiment correction.

Figure 1: Chatmood data flow

Introduction

After Apple released Siri as a successful product, both Google and Microsoft launched their personal assistants to compete for market. So far, all these tools are still question answer agents, far from the artificial intelligent in science fiction for which people wish. In order to make computer understand people's thoughts and answer like a real person, where emotion plays a important role, we must teach computer with natural dialogue corpus with sentiment label.

In previous work, researchers collect dialogue-like data from microblog. Such behavior that users put emoticons in microblog posts to convey their emotions gives researcher a great opportunity to build a corpus for sentiment analysis (Pak and Paroubek 2010; Wu, Song, and Huang 2015). However, the content from microblog is usually a collection of independent short messages, missing the context in natural dialogue. We still need to figure a way to build a natural dialogue corpus. The difficulty comes from conflicting demands that we want the the dialogue remain natural state while asking people to contribute sentiment annotations.

To deal with the difficulty in building natural dialogue corpus, we propose a mixed-initiative labeling tool called Chatmood. This tool make computer and human work together. It predicts the sentiment labels for every messages in the online chatting, and presents the sentiment labels with emoticons. What human need to do is to verify (tacitly agree) the generated emoticons and make correction if necessary. We hope that this design could minimize the impact on conversation. A pilot study has been conducted to examine our design and to explore the potential to inject the tool in real online chatting.

Chatmood System

We design Chatmood system to let human and computer cooperate in building natural dialogue corpus. Computer is responsible to make rough emoticon predictions for every messages (Figure 2a). On the other hand, human play a role of producing dialogue content and verifying the system generate emoticons. In this system, human can tacitly agree the generated emoticons or make corrections if necessary. To make a correction, just click on the system generated emoticon, and then pick the appropriate emoticons in the expanded menu as showed in Figure 2b. A user can make corrections not only for his own emoticons, but also those belong to the opposite side. When a correction is made, it will be logged and delivered to the opposite side (Figure 2c). The corrections could be used to improve emoticon prediction. With this system, we can collect two kinds of labels: the user-tacitly-agreed emoticons and corrections.

Chatmood consists of three components, client side agents, message dispatch server, and sentiment analysis module. The client side agents log every messages in the Facebook chat and send to message dispatch server. Each message will then be delivered to sentiment analysis module for sentiment classification. The result will be returned to the server and dispatched to all client side agents involve in

(a) The system generates emoticons for every messages in the chat.

(b) User can make modification to the system generated emoticon.

(c) The modification will be passed to the opponent as well.

Figure 2: Facebook chat with Chatmood chrome extension.

the chat. Figure 1 shows the data flow of Chatmood system. The classification results are displayed as emoticons in the chat. There are seven emoticons in this system, angry, sad, neutral, delighted, shy, astonished, and fear.

In our current progress, the sentiment analysis module has not been implement yet. We took Wizard of Oz approach in the pilot study. Twenty percent of labels were generated randomly to simulate a real classification model.

Pilot Study

In order to observe the user reaction to the system, we invited 10 Facebook users to our pilot study, ranging in age from 20 to 26. None of them have computer science background, and none of them knew about our research in advance. Two subjects form a group. The group members did not know each other before this experiment. We set up 4 topics, history, travel, movie, and design. For each topic, subjects were asked to read an article about it and have a ten minutes online chat using Chatmood system. Subjects were interviewed after finishing chatting.

In the interview, all of the subjects agree with that Chatmood doesn't influence the online chat at all. The bad thing is that they may ignore the system generated emoticons when they groove on chatting. Two subjects mentioned that there were too many emoticons in the chat, because not all messages carry sentiments. Even though there is a emoticon for "neutral" in the system, the subjects still feel that it is meaningless to attach emoticon to every message. In addition, most of the subjects have encountered the same situation that sometimes they can not find an appropriate emoticon when making corrections. They suggested that we should put more emoticons in the system to fit usual online chatting. While we told them that can make modification on the partner's emoticons, only three subjects really did so, but the modifications are not noticed by their partners.

Discussion and Future Work

According to the pilot study, we found that to draw user attention to system generated emotions is the most critical issue. We put too much information in the interface, and some of them are not necessary, e.g. "neutral" emotions. The "neutral" emoticon can be replaced with an inconspicuous icon. Other emotions should then become more obvious. The second issue is to make users notice the change of emoticons. We hope it can be resolved by adding an animation effect to sentiment correction events. Since Facebook Messenger[1] has large space to display more messages than in Facebook chat, a modification is more likely to be seen, we will move to Facebook Messenger in following experiments. The last issue is the insufficient emoticons. Due to space limitation, we propose to build a context-aware dynamic menu to help users to find an appropriate emoticon easily while adding more emoticons to this system.

Our next step is to conduct a larger scale experiment to collect dialogue data and sentiment labels. We will recruit Facebook users to chat online under some preset topics (mainly in Traditional Chinese). The experiment will focus on analyzing the system performance and the quality of collected labels. If the quality of collected labels meets our expectation, we will publish the this dataset and hope it could facilitate the research in this area. We are also going to release the source code of Chatmood to let people from different cultures build natural dialogue corpus in their languages.

Acknowledgements

This work was supported in part by the Ministry of Science and Technology, National Taiwan University, and Intel Corporation under Grants MOST 103-2911-I-002-001, NTU-ICRP-104R7501, and NTU-ICRP-104R7501-1.

References

Pak, A., and Paroubek, P. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the the 7th Conference on International Language Resources and Evaluation*, volume 10 of *LREC '10*, 1320–1326. Valletta, Malta: European Language Resources Association (ELRA).

Wu, F.; Song, Y.; and Huang, Y. 2015. Microblog sentiment classification with contextual knowledge regularization. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI '15, 2332–2338. Palo Alto, California: AAAI Press.

[1] https://www.messenger.com