Emotion and Its Triggers in Human Spoken Dialogue: Recognition and Analysis

Nurul Lubis†‡, Sakriani Sakti†, Graham Neubig†, Tomoki Toda†, Ayu Purwarianti‡, and Satoshi Nakamura†

Abstract Human communication is naturally colored by emotion, triggered by the other speakers involved in the interaction. Therefore, to build a natural spoken dialogue system, it is essential to consider emotional aspects, which should be done not only by identifying user emotion, but also by investigating the reason why the emotion occurred. The ability to do so is especially important in situated dialogue, where the current situation plays a role in the interaction. In this paper, we propose a method of automatic recognition of emotion using support vector machine (SVM) and present further analysis regarding emotion triggers. Experiments were performed on an emotionally colorful dialogue corpus. The result shows performance that surpasses random guessing accuracy.

1 Introduction

Communication between humans is extensively colored and strongly affected by emotions of the speakers. By nature, humans adjust their responses based on the actions of their dialogue partner in a certain emotional way – responding sadly if they're down, happily if they're nice, and angrily if they're rude. This results in a dynamic and rich communication experience – an aspect yet to be completely replicated in human-machine dialogue.

Though a number of research efforts have been performed to carry this experience to human-computer interaction [3, 2, 9], they are still mainly focused in estimating the emotion from the human's utterances, but not why these emotions occurred in the first place. Since emotion plays a two-way role in communication, knowing the reason behind displayed emotion is crucial in imitating dialogues between humans. With this information, it would be possible for machines to provide emotion-triggering responses in real world situation, leading to a form of interaction closer to the dynamic and rich communication experience between human.

Recently, the study by [7] addressed this issue by predicting and eliciting addressees emotion in online dialogue. However, the study was limited to written text form in online human communication using Twitter data. In this paper, we present a study of emotion and its triggers based on spoken utterances in emotionally colored human-human dialog. In particular, we perform: (1) automatic emotion recognition

N. Lubis · S. Sakti · G. Neubig · T. Toda · S. Nakamura
†Nara Institute of Science and Technology (Japan) e-mail: {nurul-l,ssakti,neubig,tomoki,s-nakamura}@is.naist.jp

N. Lubis · A. Purwarianti
‡Institut Teknologi Bandung (Indonesia) e-mail: 13510012@std.stei.itb.ac.id, ayu@stei.itb.ac.id

Proceedings of 5th International Workshop on Spoken Dialog Systems
Napa, January 17-20, 2014

based on SVM and (2) an analysis of the correlations of emotions and the manner in which its triggered. The overview of the system is shown in Fig 1.

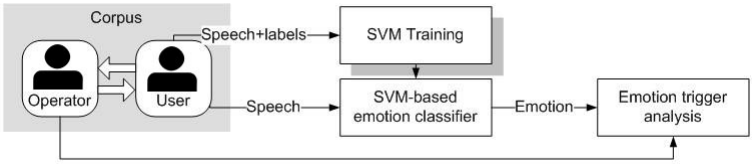

Fig. 1 System overview.

2 Emotion Definition, Recognition, and Analysis

For centuries, experts have argued on the definition of emotion and proposed systems to classify or structure emotions. The problem is, emotion is too broad a class of events be described or assessed as a single structure [10]. Amidst the thorny and intense debate, in this paper, we adopt the description of emotions using 4 dimensions as proposed in [5]. According to the level of importance, these dimensions are *valence, power, arousal*, and *expectancy*.

Two of these dimensions are then used to simplify the emotions even more into emotion classes. We define 4 emotion classes from the combinations of positive-negative of valence and active-passive of arousal, and assign them with common terms; *happiness* for positive-active, *anger* for negative-active, *sadness* for negative-passive, and *contentment* for positive-passive. It needs to be emphasized that this paper uses these terms as generalization for simplification only. The richness of each emotion class can be further explored in [5].

The emotion contained in speech can be linked to its speech features characteristics. These acoustics features are predicted on a frame basis, resulting in hundreds or even thousands of features for a single utterance. Given the complexity of the domain at hand, we employ an SVM using RBF Kernel for learning as it can implicitly handle high-dimensional feature spaces [8].

On triggers, instead of performing prediction, we attempt to tell the correlation between the emotion they contain and the emotion they trigger. From the two-way dialogue in the corpus, we identify the most common affect characteristic of the triggers of each emotion class. Deeper textual analysis makes use of the triggers' unigram, arranged according to score as calculated in Equation (1). The score of a word w for emotion e is the sum of its occurrence in e divided by its total occurrence in the corpus C.

$$Score(w,e) = freq(w,e) \div freq(w,C) \tag{1}$$

3 Experimental Setup

The experiments are performed on an emotionally colorful dialogue corpus: the SEMAINE Database [11]. The SEMAINE Database consists of dialogue between a user and an operator using Sensitive Artificial Listener (SAL) scenario, where operator shows colorful emotional expressions and manages basic aspects of conversation, such as turn taking and back-channeling, based on their observation towards

the user's condition [12]. There are 4 SAL to interact with; Poppy the optimist, Prudence the sensible, Obadiah the depressed, and Spike the angry. All interactions are then divided into a training set, development set, and test set according to session mapping in AVEC 2012 [14], though due to missing annotations several sessions are excluded. This part of the corpus will be used for construction of the emotion recognition model using LIBSVM [1]. As appropriate emotion labels for the user's utterances are provided by the corpus annotation, we use them accordingly. After segmenting the user's utterances into words and sentences, we extracted features as defined in the INTERSPEECH 2009 [13] using openSMILE [4].

For affect dimensions value estimation, we choose word as speech unit to avoid emotion fluctuation. For emotion estimation and trigger analysis, we choose each dialogue turn as speech unit to keep the dialogue context. After pre-processing, for the affect dimensions value estimation, we have 13,628 segments as training data and 10,014 as testing data, and for emotion prediction we have 947 segments as training data and 864 as testing data. Features of each segment is described in Tab 1.

Table 1 Details of extracted acoustic features.

LLD (16×2)	Functionals (12)
(Δ) ZCR	mean
(Δ) RMS Energy	standard deviation
(Δ) F0	kurtosis, skewness
(Δ) HNR	extremes: value, relative position, range
(Δ) MFCC 1-12	linear regression: offset, slope, MSE

Taking advantage of the obvious SAL characteristic in the corpus, after estimation and classification, we take a look at who provokes each emotion class most. Utterances of the most emotion-provoking character are then used in textual analysis of triggers as mentioned in Sect. 2.

4 Experiment Result

This section presents in detail the results of all the experiments performed. The explanation is broken down according to two main tasks performed in this research.

4.1 Recognizing Emotion

Emotion recognition in an emotional utterance is done based on its speech features. We confirm our SVM-based automatic emotion recognition system by comparing it to the official baseline results of AVEC 2012's emotion recognition system, and performing the automatic recognition of the emotion classes after. The result is explained below.

4.1.1 Recognition of Affective Dimensions

For each dimension, we train a regressor using RBF kernel on the training set and evaluate it on the test set. The performance of the models is measured in cross-correlation coefficient between predicted and ground-truth ratings. Tab. 2 shows the

performance of our system in comparison to the official baseline results of AVEC 2012's emotion recognition system on word-level sub-challenge (audio only), which uses Histogram kernel [14].

Table 2 The system performance on the AVEC 2012 test set measured in cross-correlation averaged over all sequences (best performance is boldfaced)

System	Valence	Arousal	Power	Expectancy	Mean
AVEC 2012 system	0.040	0.014	0.016	0.038	0.027
Proposed system	**0.338**	**0.361**	**0.088**	**0.193**	**0.245**

The result presented in Tab. 2 shows significantly better performance (p-value$<$0.001) compared to the baseline. For both systems, it appears that the power dimension is the most difficult to model. The proposed model reached the best correlation for valence dimension, the dimension claimed to be the most important on [5].

4.1.2 Recognition of Emotion Class

We performed the experiment in three schemes; one-against-one, one-against-all, and multiclass classification. The first two schemes are done to more thoroughly analyze the distinction between all emotion classes, executed in 5-fold-cross validation manner using the training set alone. The experimental result is presented below.

Table 3 One-against-one and one-against all speech-based experiment result with 95% confidence level. For each experiment, the highest accuracy is boldfaced and the lowest is underlined.

		Happiness	Anger	Sadness	Contentment
one-against-one	Happiness		81.51±4.02%	77.15±3.91%	<u>72.13±3.26%</u>
	Anger			77.17±5.55%	**86.85±2.95%**
	Sadness				76.46±3.43%
one-against-all		74.15±2.89%	84.42±2.40%	**92.90±1.77%**	<u>66.84±3.11%</u>

The one-against-one experiment result in Tab. 3 shows that some pairs of emotions are more difficult to distinguish than the rest. The system showed the best performance for contentment-anger, two classes of emotion with contrasting valence and arousal. On the one-against-all experiment, it was shown that contentment is the most difficult emotion to distinguish, followed by happiness, both of which are associated with a high valence value.

When all classes are considered for the classification task, system achieves a performance of 52.08±3.30%, surpasses the 25% accuracy of random guessing. This lower accuracy compared to the first two schemes is suspected to be due to the confusing pairs of emotion mentioned previously, affecting the overall classification accuracy. Human recognition scores 69% on the same test set.

4.2 Analyzing Emotion Triggers

As each operator in the corpus has obvious characteristics, we correlate triggered emotions with each operator. Shown in Fig. 2, natural correlations are formed be-

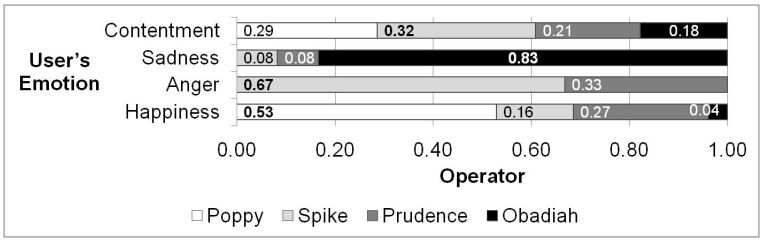

Fig. 2 Correlation of triggered emotion with operators. (highest percentage is boldfaced).

tween emotion and the operators – Happiness is most triggered by Poppy the optimist, anger by Spike the angry, and sadness by Obadiah the depressed. The figure also shows that contentment, or positive-passive emotion, is the most difficult to trigger, indicated by relatively even distribution of operator triggers, unlike the other three emotions that are dominated by a specific operator.

Table 4 Frequent trigger words.

Emotion	Trigger words
Happiness	christmas, beaches, (laugh), family, great
Anger	idiots, foolish, rage, glad, annoy
Sadness	days, miss, stressful, dog, worst
Contentment	rude, ought, never, try, actually

Further textual analysis gives us high-scoring trigger words presented in Tab. 4. Closer observation on the words points that the trigger words for contentment doesn't show any prominent characteristics, unlike those of happiness, sadness, and anger. This adds to the evidence that positive-passive emotions are trickier to trigger, while for the other three emotion classes, triggers can simply be in the identical emotion and use words related to the emotion. Tab. 5 presents dialogue example from the corpus.

Table 5 Emotion-triggering dialogues.

Spike	People can be very rude
User	Unfortunately so *(contentment)*
Obadiah	Oh that sounds nice. Sounds like you're having a good day
User	Well yeah it's going well so far, I still have things to do in the afternoon, but *(contentment)*
Spike	What's your response to those idiots?
User	Again like I said I'm never very good at telling people that they've annoyed me, so most of the time I said nothing *(anger)*

5 Conclusion and Discussion

We present a study of emotion and its triggers in human spoken dialogue–a construction of SVM based emotion recognition model and an analysis of the correlations of emotions and the manner in which it's triggered. In recognizing emotion, we

paid close attention to the characteristics of emotions and observed how it affect the recognition process. Upon trigger analysis, we drew connection between emotions and their most common cause.

As well as the emotion itself, this paper tries to draw attention to another aspect of emotion–the trigger. This aspect will play an important role in real-time systems that wants to further engage in user's emotional state, such as sensitive dialogue systems, creating dynamic two-way emotional interaction between the system and the user. These findings open the possibility of dialogue system that can cheer user up or calm them down, among other emotion-diverting acts, through incorporation of words and speech characteristic that triggers a certain emotion most in a response.

The overall performance of the system is widely open for improvements. More data from various sources can be used for training and development of the regressor and classifiers as well as experiments with kernels and SVM parameters. Further study on triggers should involve more advanced analysis using N-grams with longer context, more thoughtful scoring, as well as visual cues.

Acknowledgements Part of this research is supported by Japan Student Services Organization (JASSO) scholarship.

References

1. Chang C, Lin C (2011) LIBSVM: A library for support vector machine. In: ACM Transactions on Intelligent Systems and Technology, 2:27:1–27:27
2. Chuang Z, Wu C (2004) Multi-modal emotion recognition from speech and text. In: Computational Linguistics and Chinese Language Processing Vol 9, 4:45–62
3. Dellaert F, Polzin T, Waibel A (1994) Recognizing emotion in speech. Carnegie Mellon University. Pennsylvania, US
4. Eyben F, Woeller M, Schuller B (2010) openSMILE – The Munich versatile and fast open-source audio feature extractor. In: Proc. Multimedia (MM):1459–1462
5. Fontaine et al (2007) The world of emotion is not two-dimensional. In: Psychological Report Vol.18, 12:1050–1057
6. Frijda N (1986) The emotions. Cambridge University Press, Cambridge, UK
7. Hasegawa et al (2013) Predicting and eliciting addressee's emotion in online dialogue. In: Proc. of the 51st annual meeting of the association for computational linguistics, Vol. 1:964–972
8. Hearst M (1998) Support Vector Machines. In: Intelligent Systems and their Applications, IEEE Vol. 13, 4:18–28
9. Petrantonakis P, Hadjileontiadis L (2010) Emotion recognition from EEG using higher order crossings. In: Information Technology in Biomedicine, IEEE Transactions Vol. 14, 2:186–197
10. Russell J, Barrett L (1999) Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant. Journal of Personality and Social Psychology 1999, Vol 76, 5:805–819
11. McKeown G, Valstar M, Cowie R, Pantic M, Schroeder M (2012) The SEMAINE database: annotated multimodal records of emotionally coloured conversations between a person and a limited agent. IEEE Transactions on Affective Computing 3:5–17
12. Schroeder et al. (2012) Building Autonomous Sensitive Artificial Listeners. IEEE Transactions of Affective Computing Vol.3, 2:165–183
13. Schuller B, Steidtl S, Batliner A (2009) The INTERSPEECH 2009 Emotion Challenge. In: Proc. Interspeech, Brighton, UK:312–315
14. Schuller B, Valstar M, Eyben F, Cowie R, Pantic M (2012) AVEC 2012 – The continuous audio/visual emotion challenge. In: Proc. ACM Int'l Conf. Multimodal Interaction:449-456