Predicting and Eliciting Addressee's Emotion in Online Dialogue

Takayuki Hasegawa*
GREE Inc.
Minato-ku, Tokyo 106-6101, Japan
takayuki.hasegawa@gree.net

Nobuhiro Kaji and **Naoki Yoshinaga**
Institute of Industrial Science,
the University of Tokyo
Meguro-ku, Tokyo 153-8505, Japan
{kaji,ynaga}@tkl.iis.u-tokyo.ac.jp

Masashi Toyoda
Institute of Industrial Science,
the University of Tokyo
Meguro-ku, Tokyo 153-8505, Japan
toyoda@tkl.iis.u-tokyo.ac.jp

Abstract

While there have been many attempts to estimate the emotion of an addresser from her/his utterance, few studies have explored how her/his utterance affects the emotion of the addressee. This has motivated us to investigate two novel tasks: predicting the emotion of the addressee and generating a response that elicits a specific emotion in the addressee's mind. We target Japanese Twitter posts as a source of dialogue data and automatically build training data for learning the predictors and generators. The feasibility of our approaches is assessed by using 1099 utterance-response pairs that are built by five human workers.

Figure 1: Two example pairs of utterances and responses. Those responses elicit certain emotions, JOY or SADNESS, in the addressee's mind. The addressee in this example refers to the left-hand user, who receives the response.

1 Introduction

When we have a conversation, we usually care about the emotion of the person to whom we speak. For example, we try to cheer her/him up if we find out s/he feels down, or we avoid saying things that would trouble her/him.

To date, the modeling of emotion in a dialogue has extensively been studied in NLP as well as related areas (Forbes-Riley and Litman, 2004; Ayadi et al., 2011). However, the past attempts are virtually restricted to estimating the emotion of an addresser[1] from her/his utterance. In contrast, few studies have explored how the emotion of the addressee is affected by the utterance. We consider the insufficiency of such research to be fatal for computers to support human-human communications or to provide a communicative man-machine interface.

With this motivation in mind, the paper investigates two novel tasks: (1) prediction of the addressee's emotion and (2) generation of the response that elicits a prespecified emotion in the addressee's mind.[2] In the prediction task, the system is provided with a dialogue history. For simplicity, we consider, as a history, an utterance and a response to it (Figure 1). Given the history, the system predicts the addressee's emotion that will be caused by the response. For example, the system outputs JOY when the response is *I hope you feel better soon*, while it outputs SADNESS when the response is *Sorry, but you can't join us today*

*This work was conducted while the first author was a graduate student at the University of Tokyo.

[1]We use the terms *addresser*/*addressee* rather than a speaker/listener, because we target not spoken but online dialogue.

[2]We adopt Plutchik (1980)'s eight emotional categories in both tasks.

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 964–972,
Sofia, Bulgaria, August 4-9 2013.

(Figure 1).

In the generation task, on the other hand, the system is provided with an utterance and an emotional category such as JOY or SADNESS, which is referred to as *goal emotion*. Then the system generates the response that elicits the goal emotion in the addressee's mind. For example, *I hope you feel better soon* is generated as a response to *I have had a high fever for 3 days* when the goal emotion is specified as JOY, while *Sorry, but you can't join us today* is generated for SADNESS (Figure 1).

Systems that can perform the two tasks not only serve as crucial components of dialogue systems but also have interesting applications of their own. Predicting the emotion of an addressee is useful for filtering flames or infelicitous expressions from online messages (Spertus, 1997). The response generator that is aware of the emotion of an addressee is also useful for text completion in online conversation (Hasselgren et al., 2003; Pang and Ravi, 2012).

This paper explores a data-driven approach to performing the two tasks. With the recent emergence of social media, especially microblogs, the amount of dialogue data available is rapidly increasing. Therefore, we are taking this opportunity to building large-scale training data from microblog posts automatically. This approach allows us to perform the two tasks in a large-scale with little human effort.

We employ standard classifiers for predicting the emotion of an addressee. Our contribution here is to investigate the effectiveness of new features that cannot be used in ordinary emotion recognition, the task of estimating the emotion of a speaker (or writer) from her/his utterance (or writing) (Ayadi et al., 2011; Bandyopadhyay and Okumura, 2011; Balahur et al., 2011; Balahur et al., 2012). We specifically extract features from the addressee's last utterance (e.g., *I have had a high fever for 3 days* in Figure 1) and explore the effectiveness of using such features. Such information is characteristic of a dialogue situation.

To perform the generation task, we build a statistical response generator by following (Ritter et al., 2011). To improve on the previous study, we investigate a method for controlling the contents of the response for, in our case, eliciting the goal emotion. We achieve this by using a technique inspired by domain adaptation. We learn multiple models, each of which is adapted for eliciting one specific emotion. Also, we perform model interpolation for addressing data sparseness.

In our experiment, we automatically build training data consisting of over 640 million dialogues from Japanese Twitter posts. Using this data set, we train the classifiers that predict the emotion of an addressee, and the response generators that elicit the goal emotion. We evaluate our methods on the test data that are built by five human workers, and confirm the feasibility of the proposed approaches.

2 Emotion-tagged Dialogue Corpus

The key in making a supervised approach to predicting and eliciting addressee's emotion successful is to obtain large-scale, reliable training data effectually. We thus automatically build a large-scale emotion-tagged dialogue corpus from microblog posts, and use it as the training data in the prediction and generation tasks.

This section describes a method for constructing the emotion-tagged dialogue corpus. We first describe how to extract dialogues from posts in Twitter, a popular microblogging service. We then explain how to automatically annotate utterances in the extracted dialogues with the addressers' emotions by using emotional expressions as clues.

2.1 Mining dialogues from Twitter

We have first crawled utterances (posts) from Twitter by using the Twitter REST API.[3] The crawled data consist of 5.5 billion utterances in Japanese tweeted by 770 thousand users from March 2011 to December 2012. We next cleaned up the crawled utterances by handling Twitter-specific expressions; we replaced all URL strings to 'URL', excluded utterances with the symbols that indicate the re-posting (`RT`) or quoting (`QT`) of others' tweets, and erased `@user_name` appearing at the head and tail of the utterances, since they are usually added to make a reply. We excluded utterances given by any user whose name included 'bot.'

We then extracted dialogues from the resulting utterances, assuming that a series of utterances interchangeably made by two users form a dialogue. We here exploited 'in_reply_to_status_id' field of each utterance provided by Twitter REST API to link to the other, if any, utterance to which it replied.

[3] `https://dev.twitter.com/docs/api/`

# users	672,937
# dialogues	311,541,839
# unique utterances	1,007,403,858
ave. # dialogues / user	463.0
ave. # utterances / user	1497.0
ave. # utterances / dialogue	3.2

Table 1: Statistics of dialogues extracted from Twitter.

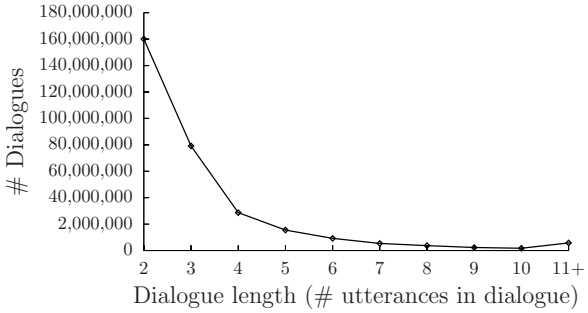

Figure 2: The number of dialogues plotted against the dialogue length.

Utterance	Emotion
A: Would you like to go for dinner with me?	
B: Sorry, I can't. I have a fever of 38 degrees.	
A: Oh dear. I hope you feel better soon.	SURPRISE
B: Thanks. I'm happy to hear you say that.	JOY

Table 2: An illustration of an emotion-tagged dialogue: The first column shows a dialogue (a series of utterances interchangeably made by two users), while the second column shows the addresser's emotion estimated from the utterance.

Table 1 lists the statistics of the extracted dialogues, while Figure 2 plots the number of dialogues plotted against the dialogue length (the number of utterances in dialogue). Most dialogues (98.2%) consist of at most 10 utterances, although the longest dialogue includes 1745 utterances and spans more than six weeks.

2.2 Tagging utterances with addressers' emotions

We then automatically labeled utterances in the obtained dialogues with the addressers' emotions by using emotional expressions as clues (Table 2). In this study, we have adopted Plutchik (1980)'s eight emotional categories (ANGER, ANTICIPATION, DISGUST, FEAR, JOY, SADNESS, SURPRISE, and TRUST) as the targets to label, and manually tailored around ten emotional expressions for each emotional category. Table 3 lists examples of the emotional expressions, while the rest are mostly their spelling variations.[4]

Emotion	Emotional expressions
ANGER	frustrating, irritating, nonsense
ANTICIPATION	exciting, expecting, looking forward
DISGUST	disgusting, unpleasant, hate
FEAR	afraid, anxious, scary
JOY	glad, happy, delighted
SADNESS	sad, lonely, unhappy
SURPRISE	surprised, oh dear, wow
TRUST	relieved, reliable, solid

Table 3: Example of clue emotional expressions.

Emotion	# utterances	Precision	
		Worker A	Worker B
ANGER	190,555	0.95	0.95
ANTICIPATION	2,548,706	0.99	0.99
DISGUST	475,711	0.93	0.93
FEAR	2,671,222	0.96	0.96
JOY	2,725,235	0.94	0.96
SADNESS	712,273	0.97	0.97
SURPRISE	975,433	0.97	0.97
TRUST	359,482	0.97	0.98

Table 4: Size and precision of utterances labeled with the addressers' emotions.

Because precise annotation is critical in the supervised learning scenario, we annotate utterances with the addressers' emotions only when the emotional expressions do not:

1. modify content words.
2. accompany an expression of negation, conditional, imperative, interrogative, concession, or indirect speech in the same sentence.

For example, *I saw a frustrated teacher* is rejected by the first condition, while *I'll be happy if it rains* is rejected by the second condition. The second condition was judged by checking whether the sentence includes trigger expressions such as 'ない (*not/never*)', 'たら (*if-clause*)', '?', 'けど (*(al)though*)', and 'と (*that-clause*)'.

Table 4 lists the size and precision of the utterances labeled with the addressers' emotions. Two human workers measured the precision of the annotation by examining 100 labeled utterances randomly sampled for each emotional category. The inter-rater agreement was $\kappa = 0.85$, indicating almost perfect agreement. The precision of the annotation exceeded 0.95 for most of the emotional categories.

[4]Note that the clue emotional expressions are language-specific but can be easily tailored for other languages. Here, Japanese emotional expressions are translated into English to widen the potential readership of the paper.

3 Predicting Addressee's Emotion

This section describes a method for predicting emotion elicited in an addressee when s/he receives a response to her/his utterance. The input to this task is a pair of an utterance and a response to it, e.g., the two utterances in Figure 1, while the output is the addressee's emotion among the emotional categories of Plutchik (1980) (JOY and SADNESS for the top and bottom dialogues in Figure 1, respectively).

Although a response could elicit multiple emotions in the addressee, in this paper we focus on predicting the most salient emotion elicited in the addressee and cast the prediction as a single-label multi-class classification problem.[5] We then construct a one-versus-the-rest classifier[6] by combining eight binary classifiers, each of which predicts whether the response elicits each emotional category. We use online passive-aggressive algorithm to train the eight binary classifiers.

We exploit the emotion-tagged dialogue corpus constructed in Section 2 to collect training examples for the prediction task. For each emotion-tagged utterance in the corpus, we assume that the tagged emotion is elicited by the (last) response. We thereby extract the pair of utterances preceding the emotion-tagged utterance and the tagged emotion as one training example. Taking the dialogue in Table 2 as an example, we obtain one training example from the first two utterances and SURPRISE as the emotion elicited in user A.

We extract all the n-grams ($n \leq 3$) in the response to induce (binary) n-gram features. The extracted n-grams could indicate a certain action that elicits a specific emotion (*e.g.*, '*have a fever*' in Table 2), or a style or tone of speaking (*e.g.*, '*Sorry*'). Likewise, we extract word n-grams from the addressee's utterance. The extracted n-grams activate another set of binary n-gram features.

Because word n-grams themselves are likely to be sparse, we estimate the addressers' emotions from their utterances and exploit them to induce emotion features. The addresser's emotion has been reported to influence the addressee's emotion strongly (Kim et al., 2012), while the addressee's emotion just before receiving a response can be a reference to predict her/his emotion in question after receiving the response.

To induce emotion features, we exploit the rule-based approach used in Section 2.2 to estimate the addresser's emotion. Since the rule-based approach annotates utterances with emotions only when they contain emotional expressions, we independently train for each emotional category a binary classifier that estimates the addresser's emotion from her/his utterance and apply it to the unlabeled utterances. The training data for these classifiers are the emotion-tagged utterances obtained in Section 2, while the features are n-grams ($n \leq 3$)[7] in the utterance.

We should emphasize that the features induced from the addressee's utterance are unique to this task and are hardly available in the related tasks that predicted the emotion of a reader of news articles (Lin and Hsin-Yihn, 2008) or personal stories (Socher et al., 2011). We will later confirm the impact of these features on the prediction accuracy in the experiments.

4 Eliciting Addressee's Emotion

This section presents a method for generating a response that elicits the goal emotion, which is one of the emotional categories of Plutchik (1980), in the addressee. In section 4.1, we describe a statistical framework for response generation proposed by (Ritter et al., 2011). In section 4.2, we present how to adapt the model in order to generate a response that elicits the goal emotion in the addressee.

4.1 Statistical response generation

Following (Ritter et al., 2011), we apply the statistical machine translation model for generating a response to a given utterance. In this framework, a response is viewed as a translation of the input utterance. Similar to ordinary machine translation systems, the model is learned from pairs of an utterance and a response by using off-the-shelf tools for machine translation.

We use GIZA++[8] and SRILM[9] for learning translation model and 5-gram language model, re-

[5] Because microblog posts are short, we expect emotions elicited by a response post not to be very diverse and a multi-class classification to be able to capture the essential crux of the prediction task.

[6] We should note that a one-versus-the-rest classifier can be used in the multi-label classification scenario, just by allowing the classifier to output more than one emotional category (Ghamrawi and McCallum, 2005).

[7] We have excluded n-grams that matched the emotional expressions used in Section 2 to avoid overfitting.

[8] http://code.google.com/p/giza-pp/

[9] http://www.speech.sri.com/projects/srilm/

spectively. As post-processing, some phrase pairs are filtered out from the translation table as follows. When GIZA++ is directly applied to dialogue data, it frequently finds paraphrase pairs, learning to parrot back the input (Ritter et al., 2011). To avoid using such pairs for response generation, a phrase pair is removed if one phrase is the substring of the other.

We use Moses decoder[10] to search for the best response to a given utterance. Unlike machine translation, we do not use reordering models, because the positions of phrases are not considered to correlate strongly with the appropriateness of responses (Ritter et al., 2011). In addition, we do not use any discriminative training methods such as MERT for optimizing the feature weights (Och, 2003). They are set as default values provided by Moses (Ritter et al., 2011).

4.2 Model adaptation

The above framework allows us to generate appropriate responses to arbitrary input utterances. On top of this framework, we have developed a response generator that elicits a specific emotion.

We use the emotion-tagged dialogue corpus to learn eight translation models and language models, each of which is specialized in generating the response that elicits one of the eight emotions (Plutchik, 1980). Specifically, the models are learned from utterances preceding ones that are tagged with emotional category. As an example, let us examine to learn models for eliciting SURPRISE from the dialogue in Table 2. In this case, the first two utterances are used to learn the translation model, while only the second utterance is used to learn the language model.

However, this simple approach is prone to suffer from the data sparseness problem. Because not all the utterances are tagged with the emotion in emotion-tagged dialogue corpus, only a small fraction of utterances can be used for learning the adapted models.

We perform model interpolation for addressing this problem. In addition to the adapted models described above, we also use a general model, which is learned from the entire corpus. The two models are then merged as the weighted linear interpolation.

Specifically, we use `tmcombine.py` script provided by Moses for the interpolation of translation models (Sennrich, 2012). For all the four features (i.e., two phrase translation probabilities and two lexical weights) derived from translation model, the weights of the adapted model are equally set as α ($0 \leq \alpha \leq 1.0$). On the other hand, we use SRILM for the interpolation of language models. The weight of the adapted model is set as β ($0 \leq \beta \leq 1.0$).

The parameters α and β control the strength of the adapted models. Only adapted models are used when α (or β) $= 1.0$, while the adapted models are not at all used when α (or β) $= 0$. When both α and β are specified as 0, the model becomes equivalent to the original one described in section 4.1.

5 Experiments

5.1 Test data

To evaluate the proposed method, we built, as test data, sets of an utterance paired with responses that elicit a certain goal emotion (Table 5). Note that they were used for evaluation in both of the two tasks. Each utterance in the test data has more than one responses that elicit the same goal emotion, because they are used to compute BLEU score (see section 5.3).

The data set was built in the following manner. We first asked five human worker to produce responses to 80 utterances (10 utterances for each goal emotion). Note that the 80 utterances do not have overlap between workers and that the worker produced only one response to each utterance.

To alleviate the burden on the workers, we actually provided each worker with the utterances in the emotion-tagged corpus. Then we asked each worker to select 80 utterances to which s/he thought s/he could easily respond. The selected utterances were removed from the corpus during training.

As a result, we obtained 400 utterance-response pairs (= 80 utterance-response pairs $\times$ 5 workers). For each of those 400 utterances, two additional responses are produced. We did not allow the same worker to produce more than one response to the same utterance. In this way, we obtained 1200 responses for the 400 utterances in total.

Finally, we assessed the data quality to remove responses that were unlikely to elicit the goal emotion. For each utterance-response pair, we asked two workers to judge whether the response elicited the goal emotion. If both workers regarded the

[10] `http://www.statmt.org/moses/`

Goal emotion: JOY
U: 16歳になりました，これからもよろしくお願いします！ (I'm turning 16. Hope to get along with you as well as ever!)
R1: 誕生日おめでとうございます！ (Happy birthday!)
R2: おめでとう！今度誕生日プレゼントあげるね. (Congratulations! I'll give you a birthday present.)
R3: おめでとうー!! 幸せな一年を！ (Congratulations! I hope you have a happy year!)

Table 5: Example of the test data. English translations are attached in the parenthesis.

Emotion	**# utterance pairs**
ANGER	119,881
ANTICIPATION	1,416,847
DISGUST	333,972
FEAR	1,662,998
JOY	1,724,198
SADNESS	436,668
SURPRISE	589,790
TRUST	228,974
GENERAL	646,429,405

Table 6: The number of utterance pairs used for training classifiers in emotion prediction and learning the translation models and language models in response generation.

response as inappropriate, it was removed from the data. The resulting test data consist of 1099 utterance-response pairs for 396 utterances.

This data set is submitted as supplementary material to support the reproducibility of our experimental results.

5.2 Prediction task

We first report experimental results on predicting the addressee's emotion within a dialogue. Table 6 lists the number of utterance-response pairs used to train eight binary classifiers for individual emotional categories, which form a one-versus-the rest classifier for the prediction task. We used opal[11] as an implementation of online passive-aggressive algorithm to train the individual classifiers.

To investigate the impact of the features that are uniquely available in a dialogue data, we compared classifiers trained with the following two sets of features in terms of precision, recall, and F_1 for each emotional category.

RESPONSE The n-gram and emotion features induced from the response.

[11] http://www.tkl.iis.u-tokyo.ac.jp/~ynaga/opal/.

Emotion	**RESPONSE**			**RESPONSE/UTTER.**		
	PREC	REC	F_1	PREC	REC	F_1
ANGER	0.455	0.476	0.465	**0.600**	**0.548**	**0.573**
ANTICIPA.	0.518	0.526	0.522	**0.614**	**0.637**	**0.625**
DISGUST	0.275	**0.519**	0.359	**0.378**	0.511	**0.435**
FEAR	**0.484**	**0.727**	**0.581**	0.459	0.706	0.556
JOY	0.690	0.417	0.519	**0.720**	**0.590**	**0.649**
SADNESS	**0.711**	0.467	0.564	0.670	**0.562**	**0.611**
SURPRISE	0.511	0.348	0.414	**0.584**	**0.437**	**0.500**
TRUST	**0.695**	0.452	0.548	0.682	**0.514**	**0.586**
average	0.542	0.492	0.497	**0.588**	**0.563**	**0.567**

Table 7: Predicting addressee's emotion: Results.

		PREDICTED EMOTION								
		ANGER	ANTICIPA.	DISGUST	FEAR	JOY	SADNESS	SURPRISE	TRUST	total
CORRECT EMOTION	ANGER	**69**	0	26	20	0	8	2	1	126
	ANTICIPA.	1	**86**	11	7	13	0	6	11	135
	DISGUST	25	1	**68**	18	2	8	7	4	133
	FEAR	3	0	22	**101**	1	5	9	2	143
	JOY	1	28	9	4	**85**	1	7	9	144
	SADNESS	6	3	25	14	5	**77**	5	2	137
	SURPRISE	7	10	9	32	5	7	**59**	6	135
	TRUST	3	12	10	24	7	9	6	**75**	146
	total	115	140	180	220	118	115	101	110	1099

Table 8: Confusion matrix of predicting addressee's emotion, with mostly predicted emotions **bold-faced** and mostly confused emotions underlined for each emotional category.

RESPONSE/UTTER. The n-gram and emotion features induced from the response and the addressee's utterance.

Table 7 lists prediction results. We can see that the features induced from the addressee's utterance significantly improved the prediction performance, F_1, for emotions other than FEAR. FEAR is elicited instantly by the response, and the features induced from the addressee's utterance thereby confused the classifier.

Table 8 shows a confusion matrix of the classifier using all the features, with mostly predicted emotions bold-faced and mostly confused emotions underlined for each emotional category. We can find some typical confusing pairs of emotions from this matrix. The classifier confuses DISGUST with ANGER and vice versa, while it confuses JOY with ANTICIPATION. These confusions conform to our expectation, since they are actually similar emotions. The classifier was less likely to confuse positive emotions (JOY and ANTICIPATION) with negative emotion (ANGER, DISGUST, FEAR, and SADNESS) vice versa.

Goal emotion: ANGER (predicted as SADNESS)
U: 毎日通話してるなんなの羨ましいわ (You have phone calls every day, I envy you.) **R:** 君の方こそ誰からも電話こないから暇で羨ましいよ。 (I envy you have a lot of time 'cause no one calls you.)
Goal emotion: SURPRISE (predicted as FEAR)
U: 黒髪がモテるってマジか。 (Is it true that dark-haired girls are popular with boys?) **R:** ８０％くらいの男子は黒髪が好きらしい。 (About 80% of boys seem to prefer dark-haired girls.)

Table 9: Examples of utterance-response pairs to which the system predicted wrong emotions.

We have briefly examined the confusions and found the two major types of errors, each of which is exemplified in Table 9. The first (top) one is sarcasm or irony, which has been reported to be difficult to capture by lexical features alone (González-Ibáñez et al., 2011). The other (bottom) one is due to lack of information. In this example, only if the addressee does not know the fact provided by the response, s/he will surprise at it.

5.3 Generation task

We next demonstrate the experimental results for eliciting the emotion of the addressee.

We use the utterance pairs summarized in Table 6 to learn the translation models and language models for eliciting each emotional category. We also use the 640 million utterances pairs in the entire emotion-tagged corpus for learning general models. However, for learning the general translation models, we currently use 4 millions of utterance pairs sampled from the 640 millions of pairs due to the computational limitation.

Automatic evaluation

We first use BLEU score (Papineni et al., 2002) to perform automatic evaluation (Ritter et al., 2011). In this evaluation, the system is provided with the utterance and the goal emotion in the test data and the generated responses are evaluated through BLEU score. Specifically, we conducted two-fold cross-validation to optimize the weights of our method. We tried α and β in $\{0.0, 0.2, 0.4, 0.6, 0.8, 1.0\}$ and selected the weights that achieved the best BLEU score. Note that we adopted different values of the weights for different emotional categories.

Table 10 compares BLEU scores of three methods including the proposed one. The first row represents a method that does not perform model adaptation at all. It corresponds to the special case (i.e., $\alpha = \beta = 0.0$) of the proposed method. The second row represents our method, while the last row represents the result of our method when the weights are set as optimal, i.e., those achieving the best BLEU on the test data. This result can be considered as an upper bound on BLEU score.

System	**BLEU**
NO ADAPTATION	0.64
PROPOSED	1.05
OPTIMAL	1.57

Table 10: Comparison of BLEU scores.

The results demonstrate that model adaptation is useful for generating the responses that elicit the goal emotion. We can clearly observe the improvement in the BLEU from 0.64 to 1.05.

On the other hand, there still remains a gap between the last two rows (i.e., proposed and optimal). We think this is partly because the current test data is too small to reliably tune parameters.

Human evaluation

We next asked two human workers to manually evaluate the generation results.

In this evaluation, the baseline (no adaptation in Table 10) and proposed method generated a response for each of the 396 utterances in the test data. For the resulting 792 utterance-response pairs, the two workers manually assessed the appropriateness of the response. Each response was judged whether it is grammatical and meaningful. If the response was regarded as so by either of the workers, it was further judged whether it elicits the goal emotion or not. To make the comparison fair, we did not expose to the workers which system generated the response. In addition, the responses generated by the two systems were presented in a random order.

As the result, 147 and 157 responses of the baseline and proposed method were regarded as appropriate, i.e., ecliting the goal emotion, by either of the workers; 74 and 92 responses were regarded as appropirate by both of the workers. These results suggest the effectiveness of the proposed method. Especially, we can confirm that the proposed method can generate responses that elicit addresee's emotion more clearly. We investigated the agreement between the two workers in this evaluation. We found that the κ coefficient is 0.59, which indicates moderate agreement. This supports the reliability of our evaluation.

Goal emotion:	JOY
Input:	写真コンクールで優秀賞とったーーーー！2 月 7 日表彰式！ (I wooooon the outstanding award at the photo competition! The ceremony is on Feb. 7!)
NO ADAPTATION:	残念ながら 1 日目だけですのー (Sorry to say, only the first day.)
PROPOSED:	金賞受賞おめでとう !!! おめでとうなのー (Congratulations on winning the gold prize!!! Congrats.)
Goal emotion:	TRUST
Input:	自暴自棄になっているの？ (Do you get desperate?)
NO ADAPTATION:	楽しみ！ (I'm looking forward to it!)
PROPOSED:	きっとまだ大丈夫 (Maybe still OK.)
Goal emotion:	ANTICIPATION
Input:	だよね！なんとかなるよね！ww (Huh! It's gonna be all right! lol) あ，わたしグッズ買わなきゃなのでその時間だけは取ってくれるとうれしい (´ ∀ `) (I gotta buy the goods, so I'll be glad if you can take the time :-))
NO ADAPTATION:	私はグッズ買ってないから不安ですね (Since I've not bought it, I feel worried.)
PROPOSED:	いいですね！私も買いますね !!! (Good! I'll buy it too!!!)

Table 11: Examples of the responses generated by the two systems, NO ADAPTATION and PROPOSED.

Examples

Table 11 illustrates examples of the responses generated by the no adaptation baseline and proposed method. In the first two examples, the proposed method successfully generates responses that elicit the goal emotions: JOY and TRUST. From these examples, we can consider that the adapted model assigns large probability to phrases such as *congratulations* or *OK*. In the last example, the system also succeeded in eliciting the goal emotion: ANTICIPATION. For this example, we can interpret that the speaker of the response (i.e., the system) feels anticipation, and consequently the emotion of the addressee is affected by the emotion of the speaker (i.e., the system). Interestingly, a similar phenomenon is also observed in real conversation (Kim et al., 2012).

6 Related Work

There have been a tremendous amount of studies on predicting the emotion from text or speech data (Ayadi et al., 2011; Bandyopadhyay and Okumura, 2011; Balahur et al., 2011; Balahur et al., 2012). Unlike our prediction task, most of them have exclusively focused on estimating the emotion of a speaker (or writer) from her/his utterance (or writing).

Analogous to our prediction task, Lin and Hsin-Yihn (2008) and Socher et al. (2011) investigated predicting the emotion of a reader from the text that s/he reads. Our work differs from them in that we focus on dialogue data, and we exploit features that are not available within their task settings, e.g., the addressee's previous utterance.

Tokuhisa et al. (2008) proposed a method for extracting pairs of an event (e.g., *It rained suddenly when I went to see the cherry blossoms*) and an emotion elicited by it (e.g., SADNESS) from the Web text. The extracted data are used for emotion classification. A similar technique would be useful for prediction the emotion of an addressee as well.

Response generation has a long research history (Weizenbaum, 1966), although it is only very recently that a fully statistical approach was introduced in this field (Ritter et al., 2011). At this moment, we are unaware of any statistical response generators that model the emotion of the user.

Some researchers have explored generating jokes or humorous text (Dybala et al., 2010; Labtov and Lipson, 2012). Those attempts are similar to our work in that they also aim at eliciting a certain emotion in the addressee. They are, however, restricted to elicit a specific emotion.

The linear interpolation of translation and/or language models is a widely-used technique for adapting machine translation systems to new domains (Sennrich, 2012). However, it has not been touched in the context of response generation.

7 Conclusion and Future Work

In this paper, we have explored predicting and eliciting the emotion of an addressee by using a large amount of dialogue data obtained from microblog posts. In the first attempt to model the emotion of an addressee in the field of NLP, we demonstrated that the response of the dialogue partner and the previous utterance of the addressee are useful for predicting the emotion. In the generation task, on the other hand, we showed that the

model adaptation approach successfully generates the responses that elicit the goal emotion.

For future work, we want to use longer dialogue history in both tasks. While we considered only two utterances as a history, a longer history would be helpful. We also plan to personalize the proposed methods, exploiting microblog posts made by users of a certain age, gender, occupation, or even character to perform model adaptation.

Acknowledgment

This work was supported by the FIRST program of JSPS. The authors thank the anonymous reviewers for their valuable comments. The authors also thank the student annotators for their hard work.

References

Moataz El Ayadi, Mohamed S. Kamel, and Fakhri Karray. 2011. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44:572–587.

Alexandra Balahur, Ester Boldrini, Andres Montoyo, and Patricio Martinez-Barco, editors. 2011. *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*. Association for Computational Linguistics.

Alexandra Balahur, Andres Montoyo, Patricio Martinez Barco, and Ester Boldrini, editors. 2012. *Proceedings of the 3rd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*. Association for Computational Linguistics.

Sivaji Bandyopadhyay and Manabu Okumura, editors. 2011. *Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology*. Asian Federation of Natural Language Processing.

Pawel Dybala, Michal Ptaszynski, Jacek Maciejewski, Mizuki Takahashi, Rafal Rzepka, and Kenji Araki. 2010. Multiagent system for joke generation: Humor and emotions combined in human-agent conversation. *Journal of Ambient Intelligence and Smart Environments*, 2(1):31–48.

Kate Forbes-Riley and Diane J. Litman. 2004. Predicting emotion in spoken dialogue from multiple knowledge sources. In *Proceedings of NAACL*, pages 201–208.

Nadia Ghamrawi and Andrew McCallum. 2005. Collective multi-label classification. In *Proceedings of CIKM*, pages 195–200.

Roberto González-Ibáñez, Smaranda Muresan, and Nina Wacholder. 2011. Identifying sarcasm in twitter: a closer look. In *Proceedings of ACL*, pages 581–586.

Jon Hasselgren, Erik Montnemery, Pierre Nugues, and Markus Svensson. 2003. HMS: A predictive text entry method using bigrams. In *Proceedings of EACL Workshop on Language Modeling for Text Entry Methods*, pages 43–50.

Suin Kim, JinYeong Bak, and Alice Haeyun Oh. 2012. Do you feel what I feel? social aspects of emotions in Twitter conversations. In *Proceedings of ICWSM*, pages 495–498.

Igor Labtov and Hod Lipson. 2012. Humor as circuits in semantic networks. In *Proceedings of ACL (Short Papers)*, pages 150–155.

Kevin Lin and Hsin-Hsi Hsin-Yihn. 2008. Ranking reader emotions using pairwise loss minimization and emotional distribution regression. In *Proceedings of EMNLP*, pages 136–144.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of ACL*, pages 160–167.

Bo Pang and Sujith Ravi. 2012. Revisiting the predictability of language: Response completion in social media. In *Proceedings of EMNLP*, pages 1489–1499.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of ACL*, pages 311–318.

Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. In *Emotion: Theory, research, and experience: Vol. 1. Theories of emotion*, pages 3–33. New York: Academic.

Alan Ritter, Colin Cherry, and William B. Dolan. 2011. Data-driven response generation in social media. In *Proceedings of EMNLP*, pages 583–593.

Rico Sennrich. 2012. Perplexity minimization for translation model domain adaptation in statistical machine translation. In *Proceedings of EACL*, pages 539–549.

Richard Socher, Jeffrey Pennington, Eric H. Huang, Andrew Y. Ng, and Christopher D. Manning. 2011. Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of EMNLP*, pages 151–161.

Ellen Spertus. 1997. Smokey: Automatic recognition of hostile messages. In *Proceedings of IAAI*, pages 1058–1065.

Ryoko Tokuhisa, Kentaro Inui, and Yuji Matsumoto. 2008. Emotion classification using massive examples extracted from the Web. In *Proceedings of COLING*, pages 881–888.

Joseph Weizenbaum. 1966. ELIZA — a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45.