Using Context to Improve Emotion Detection in Spoken Dialog Systems

Jackson Liscombe

Columbia University
Computer Science Department
New York City, New York, USA
jaxin@cs.columbia.edu

Giuseppe Riccardi, Dilek Hakkani-Tür

AT&T Labs
Research
Florham Park, New Jersey, USA
{dsp3,dtur}@research.att.com

Abstract

Most research that explores the emotional state of users of spoken dialog systems does not fully utilize the contextual nature that the dialog structure provides. This paper reports results of machine learning experiments designed to automatically classify the emotional state of user turns using a corpus of 5,690 dialogs collected with the "How May I Help You[SM]" spoken dialog system. We show that augmenting standard lexical and prosodic features with contextual features that exploit the structure of spoken dialog and track user state increases classification accuracy by 2.6%.

1. Introduction

Consider a situation in which you are a customer of a company that uses an automated agent to partially or fully interact with its customers. Suppose you call this company with a particularly strongly felt grievance. Or, consider a slightly different situation in which you call the company for information with no complaint in mind but in the course of interacting with the automated agent you become increasingly frustrated with the system's seeming inability to understand you. In both these situations you are likely to express some sort of emotion – most probably a negative one – either intentionally or unintentionally. While human agents are well-equipped to perceive the emotional state of the person with whom they are interacting, there is no commercially deployed state-of-the-art automated agent that can detect the emotional state of the caller.

In general, if an automated system can detect a problematic point in a conversation then it can modify its dialog strategy in an attempt to repair the problem or transfer the call to a human operator. Surely, one indicator of a problem is the display of negative emotion on the part of the user. We feel that tracking the emotional state of callers will improve customer satisfaction and increase the number of successful interactions.

This paper presents research on the classification of the emotional state of users of a commercially deployed customer care call-center. Section 2 describes other work in this area. Section 3 presents the corpus used in this study and describes some annotation done on it. The experimental design, feature descriptions, and results are detailed in Sections 4 and 5. Section 6 summarizes the implications of this research and outlines future directions.

2. Related Work

There is an established body of research that attempts to characterize the emotional state of human speech. This research, gaining momentum over the past 10 years, has attempted to characterize 'classic' emotions such as anger, fear, joy, and sadness using emotionally charged data elicited from actors. Research has tended to focus on either on lexical ([1]) or prosodic/acoustic ([2], [3], [4], [5], [6], [7]) cues such as intonation, speaking rate, and loudness.

The aforementioned studies use non-naturally occurring speech devoid of context in an attempt to classify a wide range of often extreme emotions. On the other hand, research in the field of emotion detection in spoken dialog systems, such as [8], [9], [10], [11], and [12] attempts to classify the more subtle naturally-occurring emotions of actual systems users. For this reason, many researchers in this area use lexical and prosodic features as a basis for emotion classification, but then augment their feature sets with additional features usually designed to take advantage of the conversational nature of their data. An additional difference between natural corpus-based research and research using actor-elicited speech is that the former often attempts only to identify the emotional valency, or positive/negative affect, of a person's speech. In other words, in the place of several emotion labels such as *anger, fear, sadness* and *joy* – which are often hard for labelers to agree upon and which are largely non-categorical – researchers tend to adopt a binary classification such as *negative* versus *positive* affect.

As an example of this latter area of research, [9] attempted to automatically detect annoyance/frustration in user turns in data collected from the DARPA Communicator Project, a travel reservation system. By including discourse features such as turn location within the conversation and dialog acts of the current turn (*repeat, repair, neither*), they found that classification accuracy using decision trees showed a relative improvement of up to 4% over the use of prosodic and lexical features alone.

In a study similar in spirit, [10] used data from a commercially deployed call-center to automatically classify caller turns as conveying either *negative* or *non-negative* emotion. This study used an extensive set of lexical and prosodic/acoustic features including a novel lexical feature in the form of an emotional salience word score that expresses the mutual information between a specific word and emotion. In addition, the researchers introduced the dialog act of the user turn as a discourse-level feature. A dialog act could be one of the following: *rejection, repeat, rephrase, ask-start over, none of the above*. It can be calculated from their results that the addition of discourse information added approximately 3% relative improvement over the use of lexical and prosodic features alone.

Many researchers, including [11] and [12], have not attempted to detect explicit emotions, but rather the points in a human-computer dialog at which "problems" occur. One study, [12], used data from an appointment scheduling dialog system. They constructed Wizard-of-Oz experiments in order to simu-

System:	How may I help you?
User:	I need to find out about a number that I don't recognize.
System:	Would you like to look up a number you don't recognize on your bill?
User:	Yes I would.
System:	Are you calling from your home phone?
User:	Yes I am.
System:	...

Figure 1: Sample dialog from the HMIHY 0300 Corpus.

late human-machine communication in which users would become angry and/or frustrated with the dialog system. In addition to lexical and prosodic features, they introduced discourse information in the form of dialog acts such as *introduce, request* and *suggest* and observed 1.2% relative improvement in the prediction of problematic units over using prosodic and lexical features alone.

The research presented in this study extends the use of discourse-level features in the automatic prediction of emotion in user turns of a commercially-deployed spoken dialog system. In addition to extracting the types of features described in the studies above, we have greatly expanded discourse-level features in such a way that more of the dialog history is used by tracking user state across previous turns. Using this extended feature set we observe an increase in prediction accuracy of user state.

3. Corpus Description and Annotation

"How May I Help You[SM]", AT&T's natural language human-computer spoken dialog system, enables callers to interact verbally with an automated agent. Users can ask for their account balance, help with AT&T rates and calling plans, explanations of certain bill charges, or identification of numbers on bills that they don't recognize and they can expect that the automated agent will understand their requests and route them to the correct information. If the system wants to confirm or clarify a customer's response, the dialog manager asks for more information; if it is still not clear, it routes the caller to a service representative. Speech data from the deployed "How May I Help You[SM]" system has been assembled into a corpus referred to as HMIHY 0300 [13]. Figure 1 presents a transcription of an example dialog from the corpus.

For a study by [14], 5,147 user turns sampled from 1,854 HMIHY 0300 calls were annotated with one of seven emotional states: *positive/neutral, somewhat frustrated, very frustrated, somewhat angry, very angry, somewhat other negative, very other negative*. Cohen's Kappa statistic, measuring inter-labeler agreement, was calculated on a subset of the data consisting of 627 user turns. A score of 0.32 was reported using the full emotion label set whereas a score of 0.42 was observed when the classes were collapsed to *positive/neutral* versus *other*.

We were primarily interested in studying user behavior over entire calls; thus, we increased the size of the corpus to 5,690 complete dialogs that collectively contain 20,013 user turns. Each new user turn turn was labeled with one of the emotion labels mentioned above. We used this expanded corpus for the experiments presented in this paper.

4. Automatic Emotion Classification

Our experiments apply the machine learning program BOOSTEXTER to the automatic classification of the emotion conveyed in each user turn. BOOSTEXTER is a boosting algorithm that forms a classification hypothesis by combining the results of several iterations of weak learner decisions [15]. For all experiments reported here we ran 2,000 such iterations. BOOSTEXTER allows input features to take both continuous and discrete values.

The corpus was divided into training and testing sets. The training set contained 15,013 user turns (75% of the corpus) and the test set was made up of the remaining 5,000 turns. The corpus was split using temporal information; the user turns in the training set occur at dates prior to those in the testing set. In addition, no dialogs were split between training and test sets. The corpus was divided in this way in order to simulate actual system development in which training data is first collected from the field, a system is then constructed using this data, and finally performance is evaluated on the newly-deployed system.

To apply BOOSTEXTER, the user turns in the corpus were encoded as a set of classes and a set of input features used as class predictors. The classes were chosen based on the seven emotions described in Section 3. However, due to the non-uniform distribution of the emotion labels (73.1% were *positive/neutral*), we adopted a binary classification scheme: *positive/neutral* was re-labeled as *non-negative* and all remaining emotions from Section 3 were collapsed to *negative*.

Each user turn was characterized by a set of 80 features that were either automatically derived or annotated by hand. As described in the following subsections, the features were grouped into the following four coherent feature sets: lexical features (LEX), prosodic features (PROS), dialog acts (DA), and contextual features (CONTEXT).

4.1. Lexical Features

The LEX feature set contains only 1 feature: the manual transcription of each user utterance. BOOSTEXTER was configured such that all unigrams, bigrams, and trigrams for each user transcription were considered in a "bag of words" fashion. In addition to lexical items, transcriptions also contained non-speech human noise such as laughter and sighs.

In the training corpus we noticed that certain words found in the user transcriptions correlated with emotional state. While these correlations were slight (the highest was less than 0.2), they were very significant ($p < 0.001$). This would seem to indicate that the words people say play a part in their emotional state, although they may not be the only indicators. Some of the more interesting correlations with negative user state are words that mention some feature of their bill ("dollars", "cents", "call") and those that indicate that the caller wishes to be transferred to a human operator ("person", "human", "speak", "talking", "machine"). Also, the data show that filled pauses such as "oh" and non-speech human noises such as sighs are also correlated with negative user state.

4.2. Prosodic Features

The PROS feature set includes 17 features designed to capture acoustic, prosodic, and voice quality information of the user turn. The motivation for the features in this feature set was an attempt to capture the way a user speaks an utterance as an indication of their emotional state.

The following 10 features were automatically extracted

over the entire user turn using Praat, a program for speech analysis and synthesis [16]: overall energy minimum, maximum, median, and standard deviation, to approximate loudness information; overall fundamental frequency (f_0) minimum, maximum, median, standard deviation, and mean absolute slope, to approximate pitch contour; and ratio of voiced frames to total frames, to approximate speaking rate.

The remaining 7 features in this set were semi-automatically extracted. Phones and silence were identified via forced alignment with manual transcriptions of user turns using a special application of AT&T WATSON, a real-time speech recognizer [17]. These features included: f_0 slope after the final vowel, intended to model turn-final pitch contour; mean f_0 and energy over longest normalized vowel, to approximate pitch accent information; syllables per second, mean vowel length, and percent internal silence, to approximate speaking rate and hesitation; and local jitter over longest normalized vowel, as a parameter of voice quality. The normalized length of each vowel was conditioned upon durational and allophonic context found in the training corpus.

The extraction techniques produce raw feature values that are often too specific for the application of a generalizing learning algorithm. Therefore, we considered a few different normalizing techniques. The optimal solution would have been to normalize by speaker. However, due to the fact that the average dialog only contained 3.5 user turns, this created a data sparsity problem. However, we felt that normalizing over the entire corpus would be too broad. Therefore, we settled on a middle ground in which we normalized by gender. Normalized feature values were expressed in units of standard deviations from the mean (z-scores). The information necessary for normalization (means and standard deviations) were only calculated over the training corpus.

4.3. Dialog Act Features

The DA feature set includes 1 feature indicating the dialog act of the current user turn. Dialog acts can be considered the function an utterance plays within the context of a dialog and as such may represent the current state of a human-computer interaction. There are different ways to label dialog acts and they range from generic to specific. For this study we used the pre-annotated call-types of the HMIHY 0300 corpus. These are somewhat specific, domain-dependent dialog act tags. Each user turn is labeled with one or more call-type from a set of 65. A few examples of the most frequent call-types in the corpus are: *Yes*, when the caller confirms a system-initiated question; *Customer_Rep*, when the caller requests to speak with a customer representative; and *Account_Balance*, when the caller requests to hear information regarding their account balance.

4.4. Contextual Features

The CONTEXT feature set was introduced as a way to model phenomena at a level that extends beyond the present user turn. User turns are embedded in a larger structure – a dialog – and it therefore seemed natural to use past evidence of user activity to help inform the emotion classification of the present user turn. Because the dialogs are relatively short in our corpus, we decided to use contextual information that extended to the previous two user turns. This feature set contains 61 features designed to track how the features described in the other feature sets compare to those of previous turns.

Feature Sets Used	Accuracy	Improvement over BASELINE
BASELINE (majority class)	73.1%	0.0%
LEX+PROS	76.1%	4.1%
LEX+PROS+DA	77.0%	5.3%
LEX+PROS+DA+CONTEXT	79.0%	8.1%

Table 1: Classification accuracy of user emotional state given different feature sets as well as relative performance improvement over the baseline.

4.4.1. Prosodic Context

Thirty four (34) features record first order differentials, or rate of change, of the PROS feature set. Half of these record the rate of change between current user utterance n and previous user utterance $n-1$. The other half record rate of change between utterances n and $n-2$. An additional 17 features calculate the second order differential between each feature in the PROS feature set for the current and previous user turns.

4.4.2. Lexical Context

An additional four features record the history of lexical information within the dialog. Two features list the manual transcriptions of the previous two user turns. Two additional features calculate the Levenshtein edit distance between the transcriptions of user turns n and $n-1$ as well as n and $n-2$. Edit distance was used as an automatic way to represent user repetition, a common indicator of misunderstanding on the part of the automated agent and, often, negative user state.

4.4.3. Discourse Context

Four features were designed to capture dialog act history. Two features record the dialog acts of user turns $n-1$ and $n-2$. In addition, two features were introduced to record the dialog acts of the system prompts that elicited user turns n and $n-1$. The HMIHY 0300 system prompts are predetermined and consist of the following dialog acts: *greeting, closing, acknowledgment, confirmation, specification, disambiguation, informative, reprompt, help, apologetic*. The final two features of the CONTEXT feature set were the emotional state of the previous two user turns. For this experiment we used hand-labeled emotions rather than predicting them.

5. Results

We present results of several experiments in the automatic classification of user emotional state using different combinations of the features sets described in Section 4. Table 1 summarizes the results.

The baseline performance on the first row of Table 1 represents classification using the majority class. Since 73.1% of all user turns are *non-negative* this is the accuracy we can achieve without looking at any features at all and simply guess this class for every user turn.

The remaining rows of Table 1 show classification results using different features set combinations as input. The second row shows the result of using both lexical and prosodic features. These features sets were combined because they are the most common features used when attempting to build a spoken language emotion classification system. As we can see, a

classification accuracy of 76.1% for LEX+PROS performs 4.1% better than the baseline.

The third row shows the classification accuracy of user emotion when we incorporate the dialog acts of the present user turn. Here we observe classification accuracy of 77.0%, which perfroms better than the baseline by 5.3% and also better than using lexical and prosodic features alone.

The last row of Table 1 lists the accuracy of classifying user state given all of our features sets combined. As we can see, the addition of contextual features boosts accuracy to 79.0%, which is an 8.1% relative increase over baseline accuracy and outperforms all other experiments as well.

6. Discussion

In this study we explored the automatic classification of the emotional state of user turns collected from a naturally-occurring human-machine spoken dialog system. The observed experimental results were largely what we anticipated. The use of lexical information coupled with prosodic features aided in emotion classification over the baseline performance. However, most researchers in the field tend to agree that classification accuracy is still sub-optimal and that lexical and prosodic features of isolated user turns do not exploit the structure inherent to spoken dialog. Indeed, we have cited several research groups who have seen emotion classification accuracy improve by 1-4% after incorporating the dialog act of a user turn as a feature. We, too, report such an improvement. Dialog acts improved classification accuracy by 1.2% over lexical and prosodic features alone.

Due to the fact that dialog acts encode the function a turn plays within the context of a dialog, it is a natural stepping off point for utilizing dialog-specific features. However, we feel that there is much more that can be exploited from dialog structure than this. In this research we attempted to utilize dialog history by including as features contextual information such as the dialog acts and lexical characteristics of previous user turns as well as monitoring prosodic information and tracking how it changes over the course of the spoken interactions. Our results show that these additional features do indeed aid emotion classification. A system trained with all the features (LEX+PROS+DA+CONTEXT) exhibited a relative improvement of 2.6% over a system trained without contexutal information (LEX+PROS+DA). We feel that this finding lends credence to the notion that regardless of the performance of a particular emotion classifier it could always be improved upon by adding contextual information whenever available.

Among possible future avenues of exploration, we intend to test classification accuracy using only automatically-derived features. For example, ASR output instead of hand transcriptions and predicted dialog acts instead of hand-labeled ones. In addition, we intend to push the use of contextual features even further in an ongoing effort to monitor the emotional state of the user throughout the course of a dialog.

7. References

[1] Z. J. Chuang and C. H. Wu, "Emotion recognition from textual input using an emotional semantic network," in *Proceedings of ICSLP*, Denver, Colorado, USA, 2002, pp. 2033–2036.

[2] P. Oudeyer, "Novel useful features and algorithms for the recognition of emotions in human speech," in *Proceedings of Speech Prosody*, Aix-en-Provence, France, 2002, pp. 547–550.

[3] S. Mozziconacci and D. J. Hermes, "Role of intonation patterns in conveying emotion in speech," in *Proceedings of ICPhS*, San Francisco, California, USA, 1999.

[4] J. R. Davitz, *The Communication of Emotional Meaning*. New York: McGraw-Hill, 1964, ch. 8: Auditory Correlates of Vocal Expression of Emotional Feeling, pp. 101–112.

[5] C. Pereira, "Dimensions of emotional meaning in speech," in *Proceedings of the ISCA Workshop on Speech and Emotion*, Belfast, Northern Ireland, September 2000.

[6] J. Yuan, L. Shen, and F. Chen, "The acoustic realization of anger, fear, joy, and sadness in chinese," in *Proceedings of ICSLP*, Denver, Colorado, USA, 2002, pp. 2025–2028.

[7] E. Zetterholm, "Emotional speech focusing on voice quality," in *Proceedings of FONETIK: The Swedish Phonetics Conference*, Gothemburg, Sweden, 1999, pp. 145–148.

[8] D. Litman and K. Forbes-Riley, "Predicting student emotions in computer-human tutoring dialogues," in *Proceedings of the 42nd Annual Meeting of the Association for Compuational Linguistics (ACL)*, Barcelone, Spain, 2004.

[9] J. Ang, R. Dhillon, A. Krupski, E. Shriberg, and A. Stolcke, "Prosody-based automatic detection of annoyance and frustration in human-computer dialog," in *Proceedings of ICSLP*, Denver, Colorado, USA, 2002, pp. 2037–2039.

[10] C. M. Lee and S. Narayanan, "Towards detecting emotions in spoken dialogs," *IEEE Transactions on Speech and Audio Processing*, in press, 2004.

[11] M. A. Walker, I. Langkilde-Geary, H. W. Hastie, J. Wright, and A. Gorin, "Automatically training a problematic dialogue predictor for a spoken dialogue system," *Journal of Artificial Intelligence Research*, vol. 16, pp. 293–319, 2002.

[12] A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Nöth, "How to find trouble in communication," *Speech Communication*, vol. 40, pp. 117–143, 2003.

[13] A. L. Gorin, G. Riccardi, and J. H. Wright, "How may I help you?" *Speech Communication*, vol. 23, pp. 113–127, 1997.

[14] I. Shafran, M. Riley, , and M. Mohri, "Voice signatures," in *Proceedings of The 8th IEEE Automatic Speech Recognition and Understanding Workshop (ASRU 2003)*, St. Thomas, U.S. Virgin Islands, November 2003.

[15] R. E. Schapire and Y. Singer, "BoosTexter: A boosting-based system for text categorization," *Machine Learning*, vol. 39, no. 2/3, pp. 135–168, 2000.

[16] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot International*, vol. 5, no. 9/10, pp. 341–345, 2001. [Online]. Available: http://www.praat.org

[17] V. Goffin, C. Allauzen, E. Bocchieri, D. Hakkani-Tr, A. Ljolje, S. Parthasarathy, M. Rahim, G. Riccardi, and M. Saraclar, "The AT&T WATSON speech recognizer," in *Proceedings of IEEE ICASSP-2005*, Philadelphia, PA, USA, 2005.