Decisive Factors in the Annotation of Emotions for Spoken Dialogue Systems

Zoraida Callejas and Ramón López-Cózar

Dept. Languages and Computer Systems, University of Granada
18071 Granada Spain
{zoraida,rlopezc}@ugr.es

Summary. The recognition of human emotions is a very important task towards implementing more natural computer interfaces. A good annotation of the emotional corpora employed by researchers is fundamental to optimize the performance of the emotion recognizers developed. In this paper we discuss several aspects to be considered in order to obtain as much information as possible from this kind of corpora, and propose a novel method to include them automatically during the annotation procedure. The experimental results show that considering information about the user-system interaction context, as well as the neutral speaking style of users, yields a more fine-grained human annotation and can improve machine-learned annotation accuracy by 24.52%, in comparison with the classical annotation based on acoustic features.

1 Introduction

Accurate annotation is a first step towards optimized detection and management of emotions, which is a very important task in order to avoid significant problems in communication, as for example misunderstandings and user dissatisfaction, which end up in very low task completion rates. Some studies, e.g. [5], have shown that once the user is in this type of states, it is difficult to guide him out. Furthermore, these bad experiences can also discourage users from employing the system again.

In despite of its benefits, annotation of emotions in spoken dialogue systems has restrictions that issue some important problems to be faced. For example, all information must be gathered through the oral modality and in some systems where the dialogue is less flexible, the length of the user prompts can be too small to use other knowledge sources like linguistic information. As our aim is to use context information even with restricted interactions (e.g. with systems that use system-directed initiatives for dialogue management), we suggest the inclusion of two new different context sources: neutral speaking style of users and dialogue history. The former, provides information about how users talk when they are not conveying any emotion, which can lead to a better recognition of the user non-neutral emotional states. The latter, involves using information about the current dialogue state in terms of dialogue length and number of confirmations and repetitions, which gives a reliable clue about which is the emotional state of the user in each moment.

M. Kurzynski et al. (Eds.): Computer Recognition Systems 2, ASC 45, pp. 747–754, 2007.
springerlink.com

We have applied this contextual information to the annotation of three negative user emotions: *doubtful*, *angry* and *bored*. The first is useful to know how the dialogue context influences the user certainty about what to do next; whereas the second and third must be recognized before the user gets too much frustrated because of system malfunctions. We consider useful to distinguish between the three because they would involve different dialogue management strategies once the recognizer is implemented.

The rest of the paper is structured as follows. In Section 2, we measure the impact of the proposed contextual information sources over human annotation. In Section 3, we evaluate the performance of our approach with machine learning approaches and compare the results to the ones obtained by the human annotators. To automatically classify emotions, we introduce a novel method in two steps which enhances negative emotion annotation with automatically generated context information. The first step introduces dialogue context and allows the distinction between *angryORbored* and *doubtful* categories; whereas the second calculates users' neutral speaking style, which we use to classify emotions into *angry* or *bored*. Finally, in Section 4 we discuss the conclusions extracted from the experimental results and point out some future work guidelines.

2 Human Annotation Results

To obtain rigorous annotations the most reliable way is to recruit specialized annotators, for example psychologists who are trained to recognize human emotions. Unfortunately, in most cases expert annotators are difficult to find and thus the annotation must be done by non-expert annotators [6]. We employed nine non-expert annotators, which is much more than what is typically reported in previous studies [2] [3]. The segment considered for the assignment of emotions was the whole utterance because it was not useful to employ smaller segmentation units (i.e. words) in our case, given that our goal was to analyze the emotion as a whole response to a system prompt, and track its effect on the subsequent interaction, and not studying the change in the emotion within a user utterance.

The utterances corpus employed in our experiments was collected from real users interacting with the UAH (Universidad Al Habla - University On the Line) dialogue system [1]. It is comprised of 85 dialogues, which contain 422 user turns, with an average of 5 user turns per dialogue. The corpus has a similar size to other real emotional speech corpora like those used by [2] (10 dialogues, 453 turns) or [4] (391 user turns). The corpus was annotated twice by every annotator, firstly in an ordered style and secondly in an unordered style. In the first mode the annotators had information about the dialogue context and the system's user speaking style and in the second they did not, so in the unordered style their annotations were based only on acoustic information. The annotation result in both ordered and unordered schemes, was the emotion annotated by more than 4 annotators. If the result was not the same for the ordered and the unordered annotation, then a non-neutral emotion was preferred as global annotation result. In the case in which the results were both non-neutral but

still different, the utterance was discarded. Global annotation results were the tags used for the machine learning approaches.

Emotional corpora extracted from real users interacting with spoken dialogue systems, are usually very unbalanced [2]. In our experiments, the 87.28% of the utterances in the UAH corpus were annotated as *neutral* in the ordered case, whereas in the case of unordered annotation, the corpus was even more unbalanced: 90.68% of utterances were annotated as *neutral*. As shown in Figure 1, the ordered annotation style yielded a greater percentage for the *bored* category, concretely 39% more than in the unordered style. In addition, the *angry* category was substantially affected by the annotation style (i.e. ordered vs. unordered), concretely 70.58% more *angry* annotations were found in the ordered annotation style. On the contrary, the *doubtful* category was practically independent from the annotation style: only 2.75% more doubts were found in the unordered annotation.

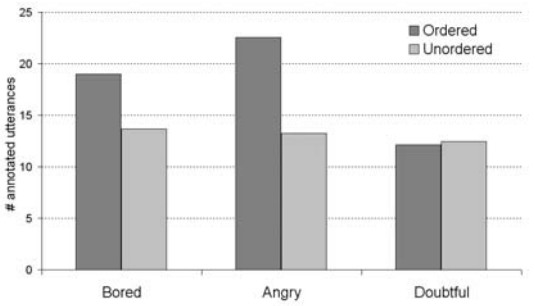

Fig. 1. Proportion of non-neutral annotated utterances

The reason for these results is that taking into account context in the ordered case causes a more stable annotation per dialogue; for example if anger is detected in one prompt then the next one is probably also annotated as *angry*. Besides, the context allowed the annotators to have information about the user speaking style and the interaction history. On the contrary, in the unordered case, they only had information about the current prompt. Hence, sometimes they could not distinguish whether the user was angry or he normally spoke loud and fast. Thus, it is an important fact to be taken into account when annotation is carried out by non-expert annotators. Furthermore, when listening to the corpus in order, the annotators had information about the position of the current user turn inside the whole dialogue, which also gave a reliable clue about the user state.

3 Machine-Based Annotation Results

Our first set of experiments was carried out to try to classify emotional prompts (i.e. not tagged as *neutral*), considering acoustic information only. This is a

classification problem that can be solved by different machine learning algorithms that receive as input tuples of features related to acoustic information. We decided to use 60 features after a literature survey to find the most employed by authors [3] [4]. These are utterance-level statistics corresponding to the four groups set out in Table 1.

Table 1. Acoustic features used for classification

Category	Features
Fundamental frequency (F0)	Min, max, range, mean, median, standard deviation, slope, correlation coef., regression error, value at first voiced segment, value at last voiced segment
F1, F2, B1, B2	Min, max, range, mean, median value at first voiced segment, value at last voiced segment
Energy	Min, max, range, mean, median, standard deviation, slope, correlation coef., regression error, value at first voiced segment, value at last voiced segment
Rhythm	Rate, voiced duration, unvoiced duration, value at first voiced, number of unvoiced segments

The first group was comprised of pitch features, which are significant indicators for emotional speech when compared to neutral conditions. We calculated all the pitch features in the voiced portion of speech. All the duration parameters (e.g. slope) were normalized by the utterance duration to obtain comparable results for all the utterances in the corpus. The second group was comprised of features related to the first two formant frequencies (F1 and F2) and their bandwidths (B1 and B2). Different speaking styles produce variations of the typical positions of formants. In the particular case of emotional speech, the vocal tract is modified by the emotional state. Energy was considered in the third group of features, it is related to the arousal level of emotions. The variation of energy of words or utterances can be used as a significant indicator for various speech styles, as the vocal efforts and ratio (duration) of voiced/unvoiced parts of speech changes. For these features, we only used non-zero values of energy, similarly as what we did for pitch. Finally, the fourth group was composed by rhythm features. Rhythm and duration features can be good emotion indicators as the duration variance decreases for most domains under fast stress conditions.

In our case the most frequent emotion category was *angry*, so the first machine learning approach that we used for comparison purposes was a baseline that always annotated user turns with this label. Secondly, we used the feature vectors as an input to a multilayer perceptron (MLP) classifier, which we used for our experiments following a 10-fold cross-validation strategy.

3.1 Classification Based on Audio Features

When we used the traditional classification based on audio features, the emotion recognition rate was 51.62% for the baseline, whereas for the perceptron

was 35.48%. These results are comparable to the case where human annotators labeled user turns unordered.

It was possible that not all the features employed for classification (60 in total) were very informative. As using irrelevant features makes the learning process slower and increases the dimensionality of the problem, we carried out a feature selection. Three methods were employed for feature selection: a forward selection algorithm, a genetic search, and finally a ranking of the features (instead of finding a subset) using the information gain as a ranking filter. The optimal subset, as it appeared with non zero gain in all the three approaches, seemed to be comprised of B1 in the last voiced segment, and energy maximum. However, we obtained no improvements using only the selected features as the accuracy was again 35.48% for the multilayer perceptron.

3.2 Classification Based on Normalized Audio Features

When humans annotated the corpus in ordered style they had information about previous user turns. Therefore, they could know e.g. that user 'A' always speaks very fast and loudly, whereas user 'B' always speaks relaxed. Therefore, some acoustic features may be the same for 'A' neutral as for 'B' angry. Hence, if the algorithms automatically learn that these features values correspond to the *angry* category, the classification will fail for user 'A'.

To solve this problem we had the user context into account, and normalized the features around the neutral voice of the user. To do this, we calculated the user's neutral voice features in each dialogue and subtracted those from the feature's values obtained in the rest of the utterances. In the experiments we considered that the first utterance of the user was neutral, assuming that he was initially in a non-negative emotional state. This assumption is feasible as employing other approaches like average value of utterances is impossible to calculate in an emotion recognizer working on real time. Thus, the first utterance of each dialogue was not taken into account for the classification experiments, that is why the baseline accuracy obtained is different across them, even when we used the same dataset. In this case we obtained 45.45% correctly classified utterances for the baseline and 54.54% with MLP.

Using the features selected in the previous section (B1 in the last voiced segment and energy maximum) we obtained 68.18% correctly classified utterances (13.64% more compared to no feature selection). In the non-normalized case the feature selection did not introduce any improvement. Thus, using normalized acoustic features (68.18% success rate) yielded an improvement of 16.56% compared to the best case in non-normalized classification (51.62% success rate), which is achieved with the baseline algorithm (Figure 2).

A study of the confusion matrices of all the described experiments showed that the *doubtful* category is often confused with the *angry* or *bored* categories, with percentages above 20% in most cases. Thus, automatically learned annotations are affected by the context information in the same way than human annotations, as for them ordered annotation did not improve the annotation of the *doubtful*

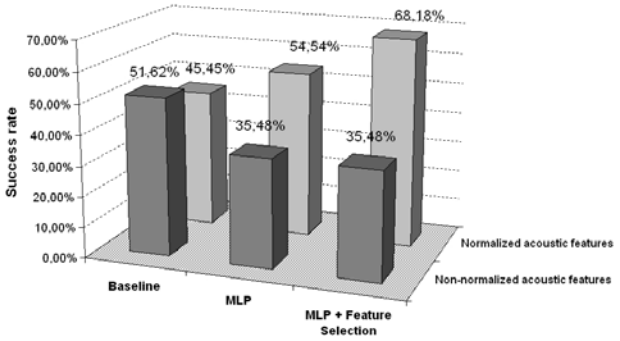

Fig. 2. Percentage of correctly annotated turns with normalized vs. non-normalized acoustic features

emotion. In contrast, for human annotators adding context information (ordered case) lead to better results in the annotation of *angry* and *bored*.

To confirm that better results can also be obtained with automatic approaches by deciding only between *angry* and *bored* categories, we classified only *bored* and *angry* utterances with the multilayer perceptron and obtained 71.42% of correctly classified utterances. To improve these results we carried out a new feature selection using the forward algorithm and obtained a subset comprised of three features: F0 median, energy maximum and duration of the longest voiced segment. The classification accuracy was 85.71%.

With these experiments we have shown that normalized acoustic features are preferable to non-normalized, as these yield 16.56% improvement (68.18% vs. 51.62% success rate). This is due to the information about the neutral style of the speaker. Besides, these results can be improved by 17.53% if we only distinguish between *bored* and *angry* emotions (85.71% with *bored* and *angry* after feature selection vs. 68.18% with three emotions). Thus, we have 85.71% correctly annotated utterances in the best case (MLP classification of *bored* and *angry* with information about the acoustic neutral after feature selection) and 35.48% in the worst case (MLP classification of non-normalized acoustic features regardless of feature selection).

3.3 Dialogue Context Annotation

We carried out dialogue context annotation considering two labels: *depth* and *width*. The former indicates the length of the dialogue, whereas the second denotes the number of user turns necessary to obtain a particular piece of information. To obtain the information about the dialogue context we employed the dialogue history, using the system prompt to automatically calculate the value for *depth* (D) and *width* (W).

The following annotation scheme was employed: D was initialized to 1 (0 would mean that the user hangs up the telephone before he says anything) and

incremented by 1 for each new user turn and each time the interaction went backwards (e.g. to the main menu). W was initialized to 0 and incremented by 1 for each user turn generated to confirm, repeat data, disambiguate input or ask the system for help. For classification purposes we used an *accumulated width* (A), so that in dialogue turn i, A(i) was the summatory of the W values from the first utterance to i. This way, confirmation and repetition subdialogues in which the user had been involved through the dialogue had always a negative impact on the user emotional state, even if he was not currently in these subdialogues.

An exhaustive study of our corpus showed that in the corpus the distribution of the *angry* and *bored* emotions regarding *depth* and *width* was rather random, e.g. we find users angry or bored with a high *depth* value. Because of this reason, we decided to take into account only two emotion categories: *doubtful* and *angryORbored*, and for classification we implemented an algorithm based on a threshold. The classification algorithm was calculated using the equation: T = D + A, where D denotes *depth* and A the *accumulated width*. A value of T greater or equal than the threshold indicated *angry* or *bored* emotional states, whereas a smaller indicated *doubtful*. Several values for the threshold were studied, obtaining that T = 4 was the optimal, for which 70% utterances were correctly annotated.

4 Conclusions and Future Work

We have carried out several experiments to study the annotation of human emotions in a corpus collected from real (non-acted) interactions with a dialogue system. This is a very difficult task, as even human beings may consider different emotions for a particular utterance. Because of this, may previous studies have focused on the recognition of emotions expressed by actors, who tend to emphasize them.

The experiments consider both a manual annotation from nine human annotators, and automatic annotation with different machine learning approaches. The results show that traditional annotation methods, based solely on acoustic features, yield to worse results in terms of classification error (in the case of automatic annotation) and decrease by 3.4% the number of non-neutral emotions annotated (in the case of human annotation). For machine-learned classification methods, the experimental results show that similarly as what happened with human annotators, the emotion annotation is substantially improved when adding information about the user neutral voice and the dialogue history. The dialogue context is useful to distinguish between *doubtful* and *angryORbored* categories with a 70% success rate. Once an utterance is classified as *angryORbored*, the normalized acoustic features let us distinguish between *bored* and *angry* with 85.71% success. Thus, 60% classification rate can be attained for the three emotions (*angry*, *bored* and *doubtful*), which is 24.52% better than the case in which no context information was used for the annotation.

As our classification method is automatic and can be employed during the running of a dialogue system, main future work guideline will be the design of

an emotion recognizer following this scheme for the UAH system, as an attempt to better adapt automatically its behavior considering the recognized emotional state of the user.

References

[1] Callejas, Zoraida, & Ramn Lpez-Czar 2005. Implementing modular dialogue systems: a case study. In Proceedings of the ASIDE 2005.
[2] Forbes-Riley, Kate, & Diane J. Litman 2004. Predicting emotion in spoken dialogue from multiple knowledge sources. In Proceedings of the HLT-NAACL 2004, pages 201–208.
[3] Lee, Chul Min, & Shrikanth S. Narayanan 2005. Toward detecting emotions in spoken dialogs. IEEE transactions on speech and audio processing, 13(2):293–303.
[4] Morrison, Donn, Ruili Wang, & Liyanage C. De Silva 2006. Ensemble methods for spoken emotion recognition in call-centers. Speech communication. In Press.
[5] Riccardi, Giuseppe, & Dilek Hakkani-Tr 2005. Grounding Emotions in Human-Machine Conversational Systems. Lecture Notes in Computer Science, pages 144–154.
[6] Vidrascu, Laurence, & Laurence Devillers 2005. Real-life emotion representation and detection in call centers data. Lecture Notes on Computer Science, 3784:739–746.