interspeech 2008 incorporating SST 08

Leveraging Emotion Detection using emotions from Yes-no answers

Narjès Boufaden [1], *Pierre Dumouchel* [1,2]

[1] Centre de recherché informatique de Montréal, Québec, Canada
[2] École de technologie supérieure, Québec, Canada

Narjes.Boufaden@crim.ca, Pierre.Dumouchel@etsmtl.ca

Abstract

We present a new approach for the detection of negative versus non-negative emotions from Human-computer dialogs in the specific domain of call centers. We argue that it is possible to improve emotion detection without using additional information being linguistic or contextual. We show that no-answers are emotional salient words and that it is possible to improve the accuracy of the classification of Human-computer dialogs by taking advantage of the high accuracy achieved on no-answer turns. We also show that stacked generalization using neural networks and SVM as base models improves the accuracy of each model while the combination of the no-model and the dialog model improves the accuracy of the dialog-model alone by 13%.

Index Terms: emotion detection, Human-computer system dialogs, machine learning

1. Introduction

The detection of user satisfaction from human-computer dialogs is a challenging problem with a growing interest as human-computer dialog systems have become more widely used. Since the late 90's, more and more research tackled this issue and two main streams emerged: the first one addressing user satisfaction by using statistics drawn from the system logs on the whole dialog [14,3]. This approach showed interesting results and was particularly useful to evaluate the impact of the dialog system components' performance on the user satisfaction.

The second approach addresses the issue by monitoring the user emotional state during the dialog. Several approaches were developed for the detection of emotions that can be divided into acoustic based approaches and acoustic+linguistic based approaches. While the latter approaches have the advantage of achieving better detection accuracy they require the output of automatic speech recognition system and are sensitive to word error rate [9]. Another direction worth the investigation is the use of conversational knowledge such us the illocutionary force of yes-no answers.

In this paper we address user satisfaction from an emotion detection perspective using prosodic and voice quality features. We present a new approach to leverage state of the art approaches in emotion detection from Human-computer dialogs in the specific domain of call centre applications.

We take into account characteristics of Human-computer dialogs such as the prominence of yes-no answers and the illocutionary force attached to them. We argue that yes-no answers are isolated emotionally salient words from which it is easy to detect emotions.

We first develop a model (no-model) to detect negative versus non-negative emotions from a no-answer corpus. We use stacked generalization algorithm with neural networks and support vector machine as base models to achieve high accuracy. Then we combine the no-model with a second model (dialog-model) trained on the dialog corpus using stacked generalization. We show that by choosing the highest prediction from the predictions generated by the no-model and the dialog-model we improve the dialog-model classification by 15% (eq. 1).

$$P(E \mid x) = \max_k P(E_k \mid x) \quad (eq.1)$$

Where k represent the no-model and the dialog-model, x is an utterance, E_k the emotion generated by the model k and E the emotion maximizing the probability $P(E_k|x)$.

In what follows, we review some current work in emotion detection from real Human-computer dialogs. Section three describes the corpora used for the experiments. In section four, we present the features and approach used for the detection of emotions. Section five presents three experiments in which we evaluate three algorithms to detect emotions from no-answers, dialogs and a combined model including the no-model and dialog-model. In section six we discuss the results and describe future work.

2. Related work

Emotion detection has been addressed using a variety of knowledge sources mainly acoustic cues [11, 17] but also linguistic knowledge sources such as word n-grams, semantic information such as domain-concepts, pragmatic knowledge such as dialog acts and the context [1, 12].

While most of the work in emotion detection have been done on artificial corpus such as the LDC Emotional Prosody some studies have been reported on the DARPA Communicator and private corpora such as the HMIHY corpus of AT&T labs (see [4] for an overview).

For instance, Lee [9] detected emotions on Human-computer dialogs obtained from real user. They combined 21 acoustic cues including fundamental frequency, energy, duration and formant frequencies, from which they selected a set of features using principal component analysis. They were able to achieve around 82.15% classification accuracy for a binary classification with negative versus non negative emotion classes using linear discriminant classifiers. By adding language model trained on emotionally salient words they were able to increase the accuracy by 7.3%.

Ang [1] reported around 78% classification accuracy for classification into two classes: *annoyance* + *frustration* versus *else* on a subset of the DARPA Communicator corpus [12] mainly composed of yes-no answers using over 16 prosodic cues modeled with a decision tree. By adding n-grams, repetition/correction and speaking style features they increased their accuracy by 7.4% using manually transcribed words.

A comparison of the different approaches shows that the best accuracy using prosodic cues on Human-computer dialogs obtained from real user varies between 78% and

83%. In order to improve the accuracy all the authors used linguistic features that need automatic speech transcription. In what follows, we show that similar improvement can be achieved by combining a model trained on acoustic features extracted from yes-no answers.

3. Corpora

For our experiments we used two corpora obtained from real users engaged in spoken dialog with a system agent over the telephone using a commercially-deployed call center application. The first corpus (no-corpus) is composed of 5000 No-answer turns and no-linguistic variation such as the words "*nop"* and "*wrong"*.

The second corpus (dialog-corpus) is composed of 5000 English dialogs (34.154 turns) representing only the costumer part of dialogs. The no-corpus was taken from a set of dialogs different from those composing the dialog-corpus. For confidentiality issues all sensitive information such as credit card numbers was removed.

3.1. Classes of emotions

While a definite set of emotions is yet to come, there is some consensus on what are basic emotions. For instance, Eckman [7] defined a set of 5 basic emotions composed of *anger*, *disgust*, *fear*, *happiness*, *sadness* and *surprise* and Banse [2] defined a finer grained list of 14 emotions including *frustration*, *cold anger, hot anger*, *boredom* and *interest*.

For the purpose of our experiment we found that the set of 14 emotions was more appropriate for the domain of call-centers. We used a subset of five emotional states: *frustration*, *cold anger*, *hot anger, interest* and *boredom* + *neutral* to model the space of emotions observed in our corpora.

In addition, because we are more interested in detecting negative emotional states without taking into account degrees of negative emotions and to ensure a higher annotator's agreement we grouped *frustration*, *cold anger* and *hot anger* into a negative emotion class and *interest*, *neutral* and *boredom* into a non-negative emotion class. Despite the fact that *boredom* is intrinsically considered a negative emotion, we found that in most dialogs, speakers were bored just because they were dealing with a virtual agent and since we are interested in emotional states that can jeopardize the user satisfaction; we considered *boredom* as a "non-negative" emotional state.

3.2. Annotation

An important part of this work was the annotation of the corpora. For that purpose twenty four annotators regrouped in teams of three were recruited for this task. The choice of three annotators was meant to ensure the quality of the annotations. The first team composed of two students and the first author have annotated the no-answer corpus. While the remaining teams composed of a call center employee's annotated the dialog corpus.

All the annotators were asked to read an annotator's guide were the class of emotions considered for this experiment were described. Then, they had to start the annotation of 100 turns after which they were asked to give their feedback and discuss the experiment. For each team, we evaluated the annotator's agreement for the 100 annotated turns. To evaluate the annotator's agreement, we used the pair-wise kappa [4] defined as follows:

$$kappa = \frac{P(A) - P(E)}{1 - P(E)}$$

Where, *P(A)* is the observed agreement and *P(E)* is the agreement by chance.

The kappa is a way to measure the reproducibility of a process, in this case the annotation. It can also be interpreted as a measure of the quality of the corpus and the degree of noisiness.

After evaluating the annotator's agreement, we presented the results and gave recommendations about the type of targeted emotions.

The average kappa obtained on the dialog-corpus was 0.4 and the kappa on the no-corpus was 0.52. The annotator's agreement was higher for the no-corpus than for the dialog-corpus. At least two reasons explain the differences; first we noticed that annotators did not use only prosodic cues to detect emotions they also take into account the content of the utterance. Hence turns where people display irony were sometimes annotated as negative despite the fact that the emotion conveyed was rather neutral. The second reason is the natural bias towards the perception of a negative emotion when a question is answered with a *no*. In this case the content influences the perception of emotions and results in a more uniform bias towards negative emotions than it is the case for dialogs.

Each corpus was divided into an all-agreement corpus where only turns with complete agreement on the emotion class are retained and a consensus-corpus where the emotion class is the one chosen by at least two annotators. In table 1 we show the distribution of the negative class and non-negative class for the (all-agreement/consensus) no-corpus and the (all-agreement/consensus) dialog-corpus.

Emotion classes	Consensus		All-agreement	
	No	Dialog	No	Dialog
Negative	950	2092	447	996
Non Negative	3986	30164	3162	27629
All Data	4936	34154	3603	28625
%Negative	19.3%	6.4%	12.3%	3.4%
%Non Negative	80.7%	93.5%	87.7%	96.6%
% All data	100%	100%	73%	88.7%

Table 1: Number of negative and non negative turns and their percentage in the all agreement corpus and the consensus corpus

From the class distributions, we can see that negative emotions occur more often in the no-corpus with 19.3% in the consensus no-corpus, respectively 12.3% in the all-agreement no corpus in comparison to 6.4% in the consensus dialog-corpus, respectively 3.4% in the all-agreement dialog-corpus.

4. Approach

In these experiments our main interest was on the validation of our hypothesis: that it is possible to improve detection of emotions from dialogs by detecting emotions form no-answers. Hence, we only used five acoustic features that have been shown to correlate with negative and non-negative emotion classes [11, 9]. All the features were extracted at the frame level and have been fitted to a Legendre Polynomial to reduce data sparseness and to generate fixed length vectors.

4.1. Feature extraction

For these experiments we used three prosodic features and two quality voice features:

- Pitch contour of each frame as a function of time with interpolation of the values on the unvoiced frames.
- Energy contour for each frame of the utterance as a function of time
- Spectral energy of each frame as a function of time.
- We also used the bandwidth of the first and second formants for each frame as a function of the time.

All these features were extracted with the snack sound toolkit (http://www.speech.kth.se/snack/)

4.2. Smoothing

All the features extracted at the frame level were fitted to a function f_t based on a combination of Legendre Polynomials of degree 7 (equation. 2). We used dynamic time warping [11] to shift all the values into the same time interval of [-1, 1] to reduce data sparseness. This approach has been shown to improve the results for speaker and language identification [5,6].

$$f_t = \sum_{i=1}^{M} a_i . P_i(t)$$

f_t is a function of time, M the highest LP degree (in our case 7) and a_i are the LP coefficients. The LP coefficients generated for each feature were used to generate a vector of dimension 35 (7x5), where each vector represents the feature values at the turn level.

5. Classification and results

We conducted three experiments. The first experiment aimed to detect emotions from the no-corpus. Different settings were explored using uniform class distributions and three algorithms for both corpora: the all-agreement-no-corpus and the consensus-no-corpus.
In the second experiment we reproduced the same setting for the dialog-corpus. In the third experiment we evaluated the combination of the no-model and dialog-model to improve emotion detection from dialogs.

5.1. Emotion detection from no-answers

In this experiment we used neural networks (NN), support vector machine (SVM) and generalized stacking [16] with the NN and SVM as base models. Morisson [10] showed that combining algorithms using stacked generalization can outperform individual performances of the base models. In these experiments we used linear regression as the meta-learner. In all our experiments we used the Weka data mining library (http://www.cs.waikato.ac.nz/ml/weka/). All the evaluations were done using a stratified 10 fold cross-validation on 1000 no-answers obtained by sub sampling the original all-agreement-no-corpus in order to have a uniform distribution over the emotion classes.

Model	NN		SVM		Stacking-No	
Emotion	Neg	NNeg	Neg	NNeg	Neg	NNeg
Rec.	0.90	0.90	0.90	0.90	0.92	0.92
Prec.	0.89	0.89	0.90	0.91	0.92	0.92
F-sc.	0.89	0.90	0.90	0.90	0.92	0.92
Rate	89.21%		90.49%		91.95%	

Table 2: Recall, Precision, F-score and error rate for the detection of negative and non-negative emotions from the all-agreement-no-corpus with uniform distribution using down sampling.

Table 2 shows the result of the classification of negative versus non-negative emotions on the all-agreement-no-corpus using a uniform distribution. We see that neural networks and support vector machine achieved almost the same performance with slightly better results for the SVM. We also see that the stacked generalization outperformed both models.

We conducted the same experiments on the consensus-no-corpus, with the same features and uniform distribution. However this time the corpus was bigger with twice the size of the all-agreement-no-corpus. The results obtained by the support vector machine and neural networks are shown in table 3. We can see that the results are far worse (about 8-10% less) than the results achieved on the all-agreement-no-corpus. These results can be explained by the fact that the all-agreement-co-corpus represents only 73% of the whole data meaning that 27% of the added data contains a lot of noisy data.

Model	NN		SVM	
Emotion	Neg	NNeg	Neg	NNeg
Rec.	0.80	0.79	0.81	0.83
Prec.	0.70	0.80	0.83	0.82
F-sc.	0.80	0.80	0.82	0.82
Rate	79.94%		82.10%	

Table 3: Recall, Precision, F-score and error rate for the detection of negative and non-negative emotions from the consensus-no-corpus with the unbiased distribution using down sampling

5.2. Emotion detection from Human-computer dialogs

We repeated the same experiments wit the same set of features and algorithms on the all-agreement-dialog corpus. All evaluations were done using a stratified 10 fold cross validation on the 2100 turns obtained by sub sampling the original all-agreement-dialog corpus in order to have a uniform distribution of the emotion classes.

Model	NN		SVM		Stacking-Turn	
Emotion	Neg	NNeg	Neg	NNeg	Neg	NNeg
Rec.	0.72	0.76	0.72	0.72	0.74	0.75
Prec.	0.75	0.73	0.72	0.72	0.75	0.74
F-sc.	0.74	0.75	0.72	0.72	0.74	0.75
Rate	74.17%		71.67%		74.63%	

Table 4: Recall, Precision, F-score and error rate for the detection of negative and non-negative emotions from the all-agreement-dialog-corpus with the unbiased distribution using down sampling

Table 4 shows the results for each classifier. In these experiments, neural networks outperformed support vector machine and the stacked generalization didn't improve significantly the accuracy.

In order to evaluate the model performances on noisy data, we repeated the experiment with the same setting using uniform distribution by sub sampling the consensus-dialog-corpus. The size of this corpus was about 4000 turns. The

same decrease of performance is observed with about 10% less accuracy for both models (Table 5).

Model	NN		SVM	
Emotion	Neg	NNeg	Neg	NNeg
Rec.	0.67	0.69	0.60	0.69
Prec.	0.62	0.64	0.66	0.63
F-sc.	0.64	0.67	0.63	0.66
Rate	65.47%		64.68%	

Table 5: Recall, Precision, F-score and error rate for the detection of negative and non-negative emotions from the consensus-dialog-corpus with the unbiased distribution using down sampling

In the remaining experiments, we use only the models trained on the all-agreement data.

5.3. Combining the two models

In this experiment we show that it is possible to improve the overall classification by taking advantage of the good results achieved for the classification of acknowledgment words.

We combine the predictions of the no-model and the dialog-model and take the prediction with the highest probability (see equation 1). This time the no-model was trained on the whole all-agreement-no-corpus while the dialog-model was trained with stratified cross validation. The test corpora generated for each step of the cross validation was evaluated by the no-model and the dialog-model and the highest prediction was taken as result. Table 6 shows the results of the classification with the combined model.

Model	Stacking-No		Stacking-Turn		combination	
Emotion	Neg	NNeg	Neg	NNeg	Neg	NNeg
Rec.	0.92	0.92	0.74	0.75	0.88	0.88
Prec.	0.92	0.92	0.75	0.74	0.88	0.88
F-sc.	0.92	0.92	0.74	0.75	0.88	0.88
Rate	91.95%		74.63%		87.8%	

Table 6: Recall, Precision, F-score and error rate for the detection of negative and non-negative emotions from the all-agreement-dialog-corpus uniform distribution using highest prediction

As we expected, the combination of the two models improved the classification of the dialogs by 13%. Both f-scores for the negative and non-negative emotions were also improved comparatively to the performances observed for the dialog-model.

6. Future work

This paper presented a new approach to improve the detection of emotions from call-center Human computer-dialogs using only acoustic features. We showed that by combining prosody from no-answers we were able to increase accuracy by 13%. Most of the improvement made by adding other knowledge sources such as linguistic sources was around 7-8%. While the focus of this work is on exploring new simple ways to improve emotion detection, little effort was made to experiment other acoustic features. We intend to improve the classification by adding new features such as the zero-crossing rate and duration. An ongoing experiment showed that the use of the crossing rate for the detection of emotions from dialog increased our model by more than 2%.

In future work, would like to explore further the usefulness of yes-no answers by developing a model for the detection of emotions from yes-answers. Finally, we plan to integrate the emotion detection model in a system developed to predict user satisfaction given dialog information extracted from the logs generated by system dialog.

7. Acknowledgements

This project was funded by NSERC and Bell University labs. We would like to thank all the people from Bell who helped in the annotation of the corpora. A special thank goes to Mariana-Damova, Jean-Pierre Croteau and Marie-Hélène Talon who were actively involved in this work.

8. References

[1] Ang J, Radjip Dhillon, Ashley Krupsky, Elisabeth Shriberg, and Andreas Stolcke, “Prosody based automatic detection of annoyance and frustration in human computer dialog”, In Proceedings of the conference ICSLP, 2002.

[15] Banse R., & Scherer K. R., “Acoustic profiles in vocal emotion expression”, Journal of Personality and Social Psychology, 70(3), 614-636, 1996.

[2] Boufaden N., Truong H and Dumouchel P. “Détection et prédiction de la satisfaction des usagers dans les dialogues Personnes-Machine”, Proceedings of Traitement Automatique du Language Naturel. 2007.

[16] Cohen J. “A coefficient of agreement for nominal scale”. Educational and Psychological Measurement, 20:37-46. 1960.

[3] Dehak N., P. Dumouchel and P. Kenny. “Modeling Prosodic Features with Joint Factor Analysis for Speaker Verification” in IEEE Transactions on Audio, Speech and Language Processing, 2007.

[4] Devillers L., Vidrascu L. and Lamel L. “Challenges in real-life emotion annotation and machine learning based detection”, Neural Networks, 18(4), May 2005.

[5] Eckman. “Basic Emotions”, in T. Dalgleish and T. Power (Eds.) The Handbook of Cognition and Emotion, p.45-60, Sussex, U.K.: John Wiley & Sons, Ltd., 1999.

[6] Huang R., Ma Ch. “Toward a Speaker-Independant Realt-time Affect Detection System”, Proceedings of the 18th International Conference on Pattern Recognition, 2006.

[8] Chul Min Lee and Shrikanth Narayanan, “Towards detecting emotions in spoken dialogs”, IEEE Transactions on Speech and Audio Processing, 13(2), pp 293-302, 2004.

[9] Lin C-Y. and H-C.Wang, “Language identification using pitch contour information”, In Proceedings of ICASSP, pp.601–604, 2005.

[7] Lugger M. and B. Yang, “Classification of different speaking groups by means of voice quality parameters”, ITG-Sprach-Kommunikation, 2006.

[10] Morisson D., Ruili W.,De Silva L.C., “Ensemble methods fr spoken emotion recognition in call-centres”, Speech Communication 49(2), pp 98-112, 2007.

[11] Myers C. S. and Rabiner L. R. “A comparative study of several dynamic time-warping algorithms for connected word recognition”. The Bell System Technical Journal, 60(7), p 1389-1409, September 1981.

[12] Walker M. and al. “Darpa communicator dialog travel planning systems: The June 2000 Data Collection”. In EuroSpeech 2001, Aalborg, Scandinavia, 2001.

[13] Walker M., Langkidle J.W.I., Gorin A., and Litman D. “Learning to predict Problematic Situations in a Spoken Dialog System: experiments with how May I Help You”, In Proceedings of the NAACL conference, 2000.

[17] Wolpert, D, “Stacked generalization”, Neural Networks, 5, 241-260, 1992.

[14] Yacoub S., Simske S., Xiaofan L. and Burns J., “Recognition of emotions in interactive voice response systems”, In proceedings of EuroSpeech, pp. 729-732, 2003.