NEGATIVE EMOTIONS DETECTION AS AN INDICATOR OF DIALOGS QUALITY IN CALL CENTERS

Christophe Vaudable[1-2], Laurence Devillers[1-3]

[1]Department of Human-Communication, LIMSI-CNRS, France
[2]Department of Computer Sciences, University Paris 11
[3]University Paris-Sorbonne PIV

ABSTRACT

Negative emotions such as anger recognition in particular can deliver useful information to both the customer and the agent of Interactive Voice Response platforms. The state-of-the-art of emotion detection is characterized as not taking into account real-life emotion behavior but "realistic" induced emotion. This study is part of the French project Voxfactory (Cap Digital). The aim is to analyze the quality of the interactions collected in call centers by using the topics of the dialogs, but also informations on opinions and emotions. A corpus of 18 hours of real dialogs between agent and customer collected in a service of complaints of French company EDF (power supply) has been annotated with emotional labels. We describe experiments on detection of three emotional states during calls. Full speaker independent test set has been used in order to be closer to a real life situation. The novelty of this paper is the analysis of full conversations (including turns with low confidence in emotion annotation and noisy turns) and the impact on the detection score. The idea is to see how far we are from a system adapted to a real life situation.

Index Terms— emotion detection, natural language processing, speech processing

1. INTRODUCTION

In this study we focus our attention on human-human conversations collected in a complaint service. An obvious interest to use data collected in call centers is the fact that the audio channel is the only channel to communicate, thus it is the only way to conveys emotions.

In the past we have worked on several corpora collected in such call centers [1-2] (medical, stock exchange, etc.) for emotion detection from speech. The work presented is done in the context of the French national VoxFactory project[1]. The aim of this project is to analyze the quality of the interactions by using the topics of the dialogs, but also the information on the opinions and the emotions.

In order to achieve robust performance under naturalistic conditions, automatic emotion detection systems must draw upon all information sources. Several studies have been carried out on data collected in call centers using both acoustic and lexical features [3]. In some papers like [4], the authors compare the obtained scores on three different databases. Two databases are taken from Interactive Voice Response (IVR) customer care domain, another database accounts for a Wizard-of-Oz (WoZ) data collection. All corpora are in "realistic" speech condition. The results of the authors are that acoustic modeling clearly outperforms linguistic modeling. Other like [5] shows that in some cases linguistic detection provides better performances. In a previous study comparing lexical model and paralinguistic model [3] on a 4 emotions classification task, we observed that the two channels were complementary, some emotions such as Fear being better recognized with the paralinguistic model and others such as relief by the linguistic one.

This paper presents a system for emotion detection using lexical and paralinguistic cues, as well as the fusion of these two types of cues for real tests: automatic segmentation, automatic transcription and unknown speakers. Few studies use automatic transcription. As it has been noted in [6], the segmentation into emotion unit is one of the most important issues if we aim at real applications but has been "largely unexplored so far". Paralinguistic, linguistic and fusion of both extracted from the conversations are explored for characterizing the emotions during the conversations in order to detect the quality of the interaction. As a first indicator of this quality, we compute the proportion of negative and positive segments in each dialog for each speaker and evaluate the problematic dialogs.

Section 2 describes the corpus (collection and annotation) used for this study. Section 3 presents protocol of experiments and different kinds of feature and method used for this study. Results are given in section 3, conclusion and future work in section 4.

2. HUMAN-HUMAN CONVERSATION

Two recording campaigns have been conducted within Callsurf [7] and VoxFactory projects respectively in two

[1] VoxFactory : Cap digital French national project founded by FUI6

EDF (French electricity utility) call centers with the same recording machine. All the dialogs have been automatically indexed by transcription obtained via an automatic speech recognition (ASR) module described in [8].

The sparseness of emotional content in real-life data and the cost of annotations led us to select for emotions annotation a subset of the corpora (which contains more than 1500 hours in total) that we hope represents the more emotional parts of the collection. To select these calls, in a first step two listeners have used keywords to find potentially emotional dialogs (eg: terms relative to power cut, verbs of complaint, etc. in order to select problematic dialogs). In a second step, the dialogs have been listened and chosen based on paralinguistic and lexical content on basis of mutual agreement of the listeners. The selected sub-corpus is composed of 115 calls (about 18 hours of dialogs) and 139 different speakers (115 clients and 24 agents). The duration of calls is between 1 and 30 minutes**.**

2.1. Segmentation, transcription and annotation

Dialogs segmentation have been considered in two different ways, automatic and manual. In the manual way we designate by "emotional turn" a unit with homogeneous emotional content which represents one or less than one speaker turn. The manual transcription of speaker turns follows standard protocol used in speech recognition area. In order to conduct tests in realistic conditions, we also use automatic segmentation and transcription of dialogs. To automatically transcribe the call center conversations, a speech recognizer has been used [8]. Considering automatic segmentation way we called turn a unit given by the ASR (Automatic Speech Recognition). Segmentation is mainly based on the speaker's change and pause (silence) during the conversations.

The emotion annotation group is composed of two experimented persons (2 female coders of about 34 years old) who have already worked on emotions annotation. Two tags have been used: agent or client for the speaker role. Three tags have been considered for the speech turns annotation: "clean", "dirty", "noisy". The "dirty" turns contain more than one speaker (most often agent-client but also two agents discussing together). Overlapped speech turns but also turns with backchannels are considered "dirty". The "noisy" turns include several noises such as the answering machine or animals' noise for example. Clean turns are considered as good signal quality and labelled only with an agent or client tag. For the emotion annotation, we have used fined-grained annotations and then regrouped the classes in 3 macro-classes: neutral (no emotion is expressed), negative (containing anger, disappointment, and negative-surprise), positive (satisfaction, positive-surprise).

2.2. Corpus

The total number of turns for the manual annotation is 11798 whereas for the automatic segmentation the number of turns is 7094. The first difference that we can notice is the average length of turns: 5.3 seconds (min 1s – max 45s) for automatic segmentation and 2.7 seconds (min 1s - max 14s) for manual segmentation. On average the length of neutral turns is quite similar between automatic and manual segmentation (2 seconds for the automatic part and 2.1 second for the manual one) but larger for negative (3.9 seconds vs 5.1 seconds for automatic) and positive turns (2seconds for the manual part and 5seconds for automatic turns) than for neutral.

For the train set we only use dialogs with a manual segmentation and transcription and emotional turns with high confidence (i.e. having the same classification for the two coders). This subset contains a total of 3684 turns (1236 positives turns, 1265 negatives and 1182 neutrals). The train set contains 66% of the turn's number and 33% for the test set. There are 139 different speakers (115 clients and 24 agents), 83 different speakers in the train set and 56 in the test set.

We have computed the kappa measure κ by Cohen's for each class (Negative/Positive/Neutral) in order to observe the rate of agreement for the two coders [9]. The results can be seen in Table 1. The number of "noisy" turns is very small and is not considered in the tests.

Annotation agreement κ with linear weighting	Manual Segmentation	Automatic Segmentation
Positive /Neutral	0,47	0,32
Negative / Neutral	0,77	0,42
Negative/Positive/Neutral	0,58	0,40

Table 1: Kappa for manual and automatic test set

We obtain κ = 0,58 for the manual segmentation and κ = 0,40(see table 1) for the automatic segmentation which represents a moderate agreement [10]. The worst scores are obtained for the annotation of Positive/Neutral emotions. The score obtained with the automatic segmentation is quite lower than the one obtained with manual segmentation. This result can be explained by the length of the turns which are longer and more difficult to evaluate for the coder but also by the presence of "dirty" turns (ie. turns with overlapped speech, backchannels, etc.) which have been also emotionally annotated.

3. PROTOCOL, EXPERIMENTS AND DISCUSSION

3.1. Acoustic and lexical cues, learning and experimental protocol

For paralinguistic cues extraction (mfcc, f0, formants, energy, etc.), the Praat program [11] is used. Then, we use our own library of functional including min/max, mean and higher order statistics to compute 374 features on voiced parts. For training model, we use a classical approach for

emotion classification: SVM classifier with radial basis function using [12]. For lexical we use bag of words to represent text in a numeric feature space. According to [13], we choose only words tagged as adjectives, nouns, verbs and interjections. Each feature thereby represents the occurrence of a specific word in a sentence. In order to reduce the amount of features in a meaningful way, stemming is applied, and a minimum occurrence frequency fmin(2) is set to discard very rare words. The occurrence frequency, referred to term frequency (tf). Another measure that is widely used for document retrieval is the inverse document frequency transformation (idf). The idea is that a sentence is characterized by words that often appear in it, except for words used in almost every sentence which are useless as discriminators. The tf and the idf transform can be combined, resulting in the tf-idf transform. We finally have 493 words in the training vocabulary against 2615 if we consider all the dialogs vocabulary of the train set for 83 different speakers.

3.2. Comparison with prototypical data

In order to highlight differences between real-life data as VoxFactory corpus and prototypical data, we have computed the relative difference between anger-mean and all-mean of each remaining features according to [14]. This relative difference is called distance. We first compute and compare the distance of negative emotions expressed in Voxfactory with other emotions of Voxfactory. For comparison we proceed the same operation with prototypical corpus JEMO presented in [14] and described as "a portrayed emotion corpus". The distance gives an indication on the difference between negative emotions and other emotions present in the corpus. The higher is the score, the more important is the difference between negative and other emotions. Results are given in Table 2 below:

	VoxFactory	JEMO
Distance between features of negative emotions and features of other emotions	25.85	134.27

Table 2: Distance between negatives emotions features and other emotions features

We can see that the distance obtained for voxfactory is considerably lower between negative emotions and other emotions (positives and neutral) compared to these obtained for JEMO. This measure provides a first indicator of the difficulty of these data. Emotions are much more shaded in the corpus Voxfactory than emotion in the prototypical corpus JEMO.

3.3. Results on paralinguistic, linguistic and fusion

In order to evaluate the reliability of the models, we use three different models based respectively on the acoustic clues only, lexical clues only and fusion on decision level of both clues. Models have been used on a speaker independent test set of 56 speakers. We use F-score as evaluation measure which is defined as a harmonic mean of precision (P) and recall(R): F = (2*PR) / P + R. We can see the results on Table 3.

	56 Speakers		
	Paraling. F-score	Linguistic F-score	Fusion F-score
Manual Segmentation (pos/neg/neu)	0.59	0.55	**0.64**
Automatic Segmentation (pos/neg/neu)	0.54	0.45	**0.56**

Table 3: F-score for paralinguistic, linguistic and fusion performance

As we can see manual segmentation outperforms automatic segmentation using the same emotional classes (pos/neg/neu). With this segmentation we obtained significant gains with the fusion model. This increase is clearly lower with models using automatic segmentation partly due to the fact that the linguistic model obtains very low score with automatic segmentation in "3 classes' mode". As a comparison we obtained a score of 0.71 with manual segmentation if we use only the acoustic channel but with a classical 10 fold cross validation approach. This difference is due to the fact that the same speaker can be presents in both train and test sets and induce bias in the experiments results.

3.4. From artificial to real life detection

Previous results show that the acoustic score are better for this task than linguistic. In order to be closer to a realistic situation we conduct all the experiments on the three classes' automatic test set. We then keep only acoustic features for the rest of the experiments. The main objective of the VoxFactory project is the detection of problematic calls. In order to detect these calls, we need to examine the whole dialogs including turns with high confidence, noisy turns, and turns with low confidence. As mentioned in section 2 we had selected 41 dialogs for the test: 26 dialogs which are globally tagged with negative emotions and 15 with positive. As a first indicator, we compute the proportion of emotional turns (negative and positive) that can be detected by our system. If a majority of negative turns compared to positive turns is detected in the dialogs the call is affected to the negative class. On the contrary if a majority of positive turns are detected, the call will be classified as positive. In the first step of this experiment we use only "clean" data of our automatic test set by selecting

turns with high confidence without "overlap", backchannel" (we call the test set: subset1). In the second step we add to subset1 "dirty" turns with high confidence (we call it subset 2). In the last step we select the whole data containing turns of subset 1 and 2 plus turns with low confidence (we call it subset 3). Initially we have computed F-score only for "dirty" turns (overlap, backchannel) with high confidence in order to see the potential degradation of model's performances (Table 4). Results for the three subsets are presented in table 5.

	F-score ("clean" turns)	F-score ("dirty" turns)
Automatic segmentation (pos/neg/neu)	0.54	0.352

Table 4: F-score's comparison for "clean" and "dirty" turns

	Positive dialogs	Negative dialogs	F-score
Subset 1 = "clean" turns (high confidence)	7/15	22/26	0,66
Subset 2 = "dirty" and "clean" turns with high confidence	7/15	23/26	0,68
Subset 3 = "dirty" and "clean" turns with high AND low confidence	4/15	21/26	0,51

Table 5: F-score for the complete dialogs analysis

The addition of "dirty" turns with high confidence (subset 2) doesn't seem to have consequences on model's performances (F-score = 0.68). The addition of turns with low confidence seems to be more problematic. While the classification of negative dialogs is correct (recall > 0.80), it is not the case for positive dialogs.

The first indicator used follows the following simple rules:

IF #negative seg > #positive seg THEN negative dialog

IF # positive seg > #negative seg THEN positive dialog

A lot of other indicators will be tested in future experiments taking into account for example the presence of negative emotions at the beginning or end of a dialog.

4. CONCLUSION AND FUTURE WORK

We have successively explored in this study some characteristics of real-life data. First, we have used a distance measure to differentiate real life data and prototypical data. Secondly, we have compared manual segmentation and automatic segmentation in order to evaluate the degradation of performances. We have used an independent test set to be closer of a real life situation. Even with a training corpus containing 83 different speakers, the performance difference between cross validation method (speaker dependent) and separated test set (speaker independent) is important (especially considering a 3 emotions detection system)).

Results show that full automatic approach is from the beginning more difficult to treat. In order to classify a complete dialog as problematic or not, we compute a first simple indicator. All the segments of the dialog tests are emotionally tagged and the global percentages of positive, negative and neutral emotions are computed. If the number of negative turns is superior to the number of positive turns, the dialog is automatically labeled as negative. While the classification of negative dialogs is correct (recall > 0.80), it is not the case for positive dialogs. In our next experiments we will test other indicators which can be defined and calculated by using the combination of emotions detected with paralinguistic and linguistic models and the presence of the affect bursts and disfluences. The positions of the negative emotions detected (beginning, middle or ending of the dialog) as well as the role of the speaker who provides negative emotions will be also studied in order to create a better indicator of the quality of the interaction.

5. REFERENCES

[1] Devillers L., Vidrascu L. and Lamel L., *Emotion detection in real-life spoken dialogs recorded in call center*. Journal of Neural Networks, numéro spécial 2005. volume 18.

[2] Devillers L., Vaudable C. and Chastagnol C., *Real-life emotion-related states detection in call centers*, in *InterSpeech 2010*. 2010: Makhuari, Japan.

[3] Devillers L. and Vidrascu L., *Real-life emotions detection with lexical and paralinguistic cues on Human-Human call center dialogs*, in *Interspeech 2006*. 2006, ISCA: Pittsburg, USA.

[4] Polzehl T., Schmitt A., Metze F. and Wagner M., *Anger recognition in speech using acoustic and linguistic cues.* Speech Communication, 2011. 53(9-10).

[5] Gupta P. and Rajput N., *Two-Stream Emotion Recognition For Call Center Monitoring*, in *Interspeech 2007*. 2007: Antwerp, Belgium.

[6] Batliner A., Seppi D., Steidl S. and Schuller B., *Segmentic into adequate units for automatic recognition of emotio-related episodes : A speech-based Approach.* Advances in human-computer interaction, 2010. 10.

[7] Garnier-Rizet M., Adda G., Cailliau F., Gauvain J.-L., Guillemin-Lanne S., Lamel L., Vanni S. and Waast-Richard C., *CallSurf - Automatic transcription, indexing and structuration of call center conversational speech for knowledge extraction and query by content*, in *LREC*. 2008: marrakech.

[8] Gauvain J.-L., Lamel L., schwenk H., Adda G., chen L. and Lefevre F., *Conversational telephone speech recognition*, in *ICASSP*. 2003: Hong Kong. p. 212-215.

[9] Cohen J., *A coefficient of agreement for nominal scales.* Educational and Psychological Measurement, 1960. **20**: p. 27-46.

[10] Craggs R., *Annotating emotion in dialogue – Issues and Approaches*, in *7th Annual CLUK Research Colloquium* 2004: University of Birmingham.

[11] Boersma P. and Weenink D., *Praat: doing phonetics by computer*. 2009.

[12] Chih-Chung, Chang and Lin C.-J., *LIBSVM : a library for support vector machines*. 2001.

[13] Turney P., D, *Thumbs up or thumbs down ? semantic orientation applied to unsupervised classification of reviews*, in *40th Annual Meeting of the Association for computationnal Linguistics*. 2002: Philadelphia.

[14] Tahon M. and Devillers L., *Acoustic measures characterizing anger across corpora collected in artificial or natural context*, in *Speech Prosody*. 2010: Chicago.