Annotation of Complex Emotions in Real-Life Dialogues: The Case of Empathy

Morena Danieli
Department of Information Engineering and Computer Science, University of Trento, Italy
danieli@disi.unitn.it

Giuseppe Riccardi
Department of Information Engineering and Computer Science, University of Trento, Italy
riccardi@disi.unitn.it

Firoj Alam
Department of Information Engineering and Computer Science, University of Trento, Italy
alam@disi.unitn.it

Abstract

English. In this paper we discuss the problem of an-notating emotions in real-life spoken conversations by investigating the special case of empathy. We propose an annotation model based on the situated theories of emotions. The annotation scheme is directed to ob-serve the natural unfolding of empathy during the conversations. The key component of the protocol is the identification of the annotation unit based both on linguistic and paralinguistic cues. In the last part of the paper we evaluate the reliability of the annotation model.

Italiano. In questo articolo illustriamo il problema dell'annotazione delle emozioni nelle conversazioni reali, illustrando il caso particolare dell'empatia. Proponiamo un modello di annotazione basato sulla teoria situazionale delle emozioni. Lo schema di an-notazione è diretto all'osservazione al naturale dipanamento dell'empatia nel corso della conversazione. La componente principale del protocollo è l'identificazione dell'unità di annotazione basata sul contenuto linguistico e paralinguistico dell'evento emozionale. Nell'ultima parte dell'articolo riportiamo i risultati relativi all'affidabilità del modello di annotazione.

1 Introduction

The work we present is part of a research project aiming to provide scientific evidence for the situated nature of emotional processes. In particular we investigate the case of complex social emotions, like empathy, by seeing them as relational events that are recognized by observers on the basis of their unfolding in human interactions. The ultimate goals of our research project are a) understanding the multidimensional signals of empathy in human conversations, and b) generating a computational model of basic and complex emotions. A fundamental requirement for building such computational systems is the reliability of the annotation model adopted for coding real life conversations. Therefore, in this paper, we will focus on the annotation scheme that we are using in our project by illustrating the case of empathy annotation.

Empathy is often defined by metaphors that evoke the emotional or intellectual ability to identify another person's emotional states, and/or to understand states of mind of the others. The word "empathy" was introduced in the psychological literature by Titchener in 1909 for translating the German term "Einfühlung". Nowadays it is a common held opinion that empathy encompasses several human interaction abilities. The concept of empathy has been deeply investigated by cognitive scientists and neuroscientists, who proposed the hypothesis according to which empathy underpins the social competence of reconstructing the psychic processes of another person on the basis of the possible identification with his/her internal world and actions (Sperber & Wilson, 2002; Gallese, 2003).

Despite the wide use of the notion of empathy in the psychological research, the concept is still vague and difficult to measure. Among psychologists there is little consensus about which signals subjects rely on for recognizing and echoing empathic responses. Also the uses of the concept by the computational attempts to reproduce empathic behavior in virtual agents seem to

be suffering due to the lack of operational definitions.

Since the goal of our research is addressing the problem of automatic recognition of emotions in real life situations, we need an operational model of complex emotions, including empathy, focused on the *unfolding* of the emotional events. Our contribution to the design of such a model assumes that processing the discriminative characteristics of acoustic, linguistic, and psycholinguistic levels of the signals can support the automatic recognition of empathy in situated human conversations.

The paper is organized as follows: in the next Section we introduce the situated model of emotions underlying our approach, and its possible impact on emotion annotation tasks. In Section 3 we describe our annotation model, its empirical bases, and reliability evaluation. Finally, we discuss the results of lexical features analysis and ranking

2 Situated theories of emotions and emotion annotation

The theoretical model of situated cognition is an interesting framework for investigating complex emotions. Recently, both neuropsychologists and neuroscientists used the situated model for experimenting on the emotional experiences. Some results provided evidences supporting the thesis that complex emotions are mental events which are construed within situated conceptualizations (Wilson-Mendenhall et al. 2011). According with this view, a subject experiences a complex emotion when s/he conceptualizes an instance of affective feeling. In other terms, experiencing and recognizing an emotion is an act of categorization based on embodied knowledge about how feelings unfold in situated interactions (Barrett 2006). In this view experiencing an emotion is an event emerging at the level of psychological description, but causally constituted by neurobiological processes (Barrett & Lindquist 2008; Wambach and Jerder, 2004).

The situated approach is compatible with the modal model of emotions by Gross (Gross 1998; Gross & Thompson 2007), which emphasizes the attentional and appraisal acts underlying the emotional process. According to Gross, emotions arise in situations where interpersonal transactions can occur. The relevant variables are the behavior of the participating subjects, including their linguistic behavior, and the physical context, including the physiological responses of the participating speakers. The situation compels the attention to the subject, implies a particular meaning for the person, and gives rise to coordinated and malleable responses.

The framework mentioned above has important implications for our goal because it focuses on the process underlying the emotional experience. Actually one of the problems of annotating emotions is related with the difficulty of capturing how the emotional events feel like and how they arise in verbal and non-verbal interactions.

In the field of spoken language processing we have several collections of annotated emotional databases. Rao and Koolagudi (2013), and El Ayadi (2011) provide well informed survey of emotional speech corpora. From their analysis it results that there is a significant disparity among such data collections, in terms of explicitness of the adopted definitions of emotions, of complexity of the annotated emotions, and of definition of the annotation units. Most of the available emotional speech databases have been designed to perform specific tasks, e.g. emotion recognition or emotional speech synthesis (Tesser et al. 2004; Zovato et al. 2004), and the associated annotation schemes mostly depend from the specific tasks as well. A common feature shared by many emotional corpora is their focus on discrete emotion categorizations. To the best of our knowledge no one provides specific insights for annotating the process where emotions unfold. Also more comprehensive models either base their annotation schemes on sets of basic emotions, like the one developed within the HUMAINE project (Douglas-Cowie *et al.* 2003), or they present data collected in artificial human-virtual agent interactions, like the SEMAINE corpus (McKeown et al. 2007).

In the field of human computer interaction, the present models of empathy aim to identify different "sentiment features" such as affect, personality and mood (Ochs et al. 2007). Few, if any, of those works investigate the differential contribution of speech content and emotional prosody to the recognition of empathy, in spite of the evidences that the interplay between verbal and non-verbal features of behavior are probably the best candidate *loci* where human emotions reveal themselves in social interactions (a view supported by many studies, including Magno-Caldognetto 2002; Zovato *et al.* 2008; Danieli, 2007; Kotz & Paulmann 2007; Brück *et al.* 2012; Gili-Fivela & Bazzanella, 2014 among others).

3 Annotation scheme for complex emotions

We argue that the difficult problem of providing guidelines for complex emotion annotation can benefit from focusing the annotators' attention on the emotional process. This requires the identification of the annotation units that are more promising from the point of view of supporting the observer's evaluation on when and how a given emotion arises.

3.1 In search of the annotation units

For pursuing the research described in this paper we investigated if any of the available psychometric scales or questionnaires were usable in our data analysis, both for empathy and for other complex social emotions like satisfaction, and frustration.

As for empathy, we found that among psychologists there are some fundamental concerns about the adequacy of the various scales. For example, no significant correlation was found between the scores on empathy scales and the measurement of empathic accuracy (Lietz *et al*. 2011). The *de-facto* standardized available tests, such as the one referenced in Bahron-Cohen et al. 2013, seem to be effective mostly for clinical applications within well-established experimental settings. However, they can hardly be adapted to judge the empathic abilities of virtual agents and to evaluate human empathic behavior in everyday situations by an external observer.

Given the problematic applicability of psychological scales and computational coding schemes, for capturing in real-life conversations the unfolding of the emotional process, we chose to focus on the interplay between speech content and voice expression. It is well known that the paralinguistic features of vocal expression convey a great deal of information in spoken interactions. In different kinds of interpersonal communication, the accessibility to the facial expressions (in terms of visual frames) is not available. In such cases we usually rely on spoken content and on the paralinguistic events of the spoken utterances. Therefore, in our research we focused on acoustic, lexical and psycholinguistic features for the automatic classification of empathy in conversations, but we chose to rely only on the perception of affective prosody for the annotation task.

3.2 *The empirical base*s

For designing the annotation scheme we made an extensive analyses on a large corpus of real human-human, dyadic conversations collected in a call center in Italy. Each conversation length was around 7 minutes. An expert psycholinguist, Italian native speaker, listened to one hundred of such conversations. She focused on dialog segments where she could perceive emotional attitudes in one of the speakers. The expert annotator's goal was to pay attention to the onset of prosodic variations and judge their relevance with respect to empathy. In doing that she evaluated the communicative situation in terms of appraisal of the transition from a neutral emotional state to an emotional connoted state. Let us clarify this with a dialogue excerpt from the annotated corpus. The fragment is reported in Figure 1, where "C" is the Customer, and "A" is the Agent. The situation is the following: C is calling because a payment to the company is overdue, he is ashamed for not being able to pay immediately, and his speech is plenty of hesitations. This causes an empathic echoing by A: that emerges from the intonation profile of A's reply, and from her lexical choices. For example in the second question of A's turn, she uses the hortatory first plural person instead of the first singular person. Also the rhetorical structure of A's turn, i.e., the use of questions instead of assertions, conveys her empathic attitude.

C:	Senta … ho una bolletta scaduta di 833 euro eh… vorrei sapere se … come posso rateizzarla?
A:	Ma perché non ha chiesto prima di rateizzarla? Proviamo a farlo adesso, ok? […]

Figure 1: An excerpt of a conversation

The expert annotator thus perceived the intonation variation, and marked the speech segment corresponding to the intonation unit outlined in the example, where the word "proviamo" (*let us try*) is tagged as onset of the emotional process.

The results of this listening supported the hypothesis that the relevant speech segments were often characterized by significant transitions in the prosody of speech. As expected, such variations sometimes co-occurred with emotionally connoted words, but also with functional parts of speech like Adverbs and Interjections. Also phrases and Verbs, as in the example, could play

the role of lexical supports for the manifestation of emotions.

On the basis of those results, we designed the annotation scheme for empathy by taking into account *only* the acoustic perception of the variations in the intonation profiles of the utterances. Two expert psychologists, Italian native speakers, performed the actual annotation task. They were instructed to mark the relevant speech segments with empathy tags where they perceived a transition in the emotional state of the speaker, by paying attention to the speech melody, the speaker's tone of voice and only limited attention to the semantic content of the utterance. In the analyzed corpus 785 calls were tagged with respect to the occurrence of empathy. The annotators used the EXMARaLDA Partitur Editor (Schmidt 2004) for the annotation task.

3.3 *Evaluation*

To measure the reliability of this coding scheme we calculated inter-annotator agreement by using the Cohen's kappa statistics, as discussed in Carletta, 1996. For the evaluation, two psychologists worked independently over a set of 64 spoken conversations. We found reliable results with kappa = 0.74. In particular, the comparison showed that 31.25% of the annotated speech segments were exactly tagged by the two annotators at the same positions of the time axis of the waveforms. 53.12% was the percentage of cases where the two annotators perceived the empathic attitude of the speaker occurring in different time frames of the same dialog turns. No other disagreement was reported.

4. Lexical feature analysis and ranking

For the feature analysis, we extracted lexical features from manual transcription consisting of a lexicon of size 13K. Trigram features were extracted to understand whether there are any linguistically relevant contextual manifestations while expressing empathy. For the analysis of the lexical features we used Relief feature selection algorithm (Kononenko, 1994), which has been effective in personality recognition from speech (Alam & Riccardi 2013). Prior to the feature selection we have transformed the raw lexical features into bag-of-words (vector space model), which is a numeric representation of text that has been introduced in text categorization (Joachims, 1998) and is widely used in behavioral signal processing (Shrikanth *et al.* 2013). Each word in the text can be represented as an element in a vector in the form of either Boolean zero/one or frequency. In case of using frequency, it can be transformed into various forms such as logarithmic term frequency (tf), inverse document frequency (idf) or combination of both (tf-idf). For this study, the frequency in the feature vector was transformed into tf-idf, the product of tf and idf. After that, feature values were discretized into 10 equal frequency bins using un-supervised discretization approach to get the benefits in feature selection and classification. Then, we used Relief feature selection algorithm and ranked the features, based on the score computed by the algorithm. In Table 1, we present a selection of the top ranked lexical features selected using the Relief feature selection, which are highly discriminative for the automatic recognition of empathy.

Lexical Features	Score
posso aiutarla	0.17
se lei vuole	0.10
assolutamente sì	0.10
vediamo	0.07
sicuramente	0.06

Table 1: Excerpt from top-ranked lexical features using Relief feature selection algorithm.

As we can see from Table 1, the selected lexical features highlight the type of sentences that are commonly used in customer care services by the Agents, like "posso aiutarla" (*can I help you*), but also less common phrases like "se lei vuole" (*if you want*, including the courtesy Italian pronoun *lei*), and the use of the first plural form of Verbs, like "vediamo" (*let us see*).

5. Conclusions

In this paper we propose a protocol for annotating complex social emotions in real-life conversations by illustrating the special case of empathy. The definition of our annotation scheme is empirically-driven and compatible with the situated models of emotions. The difficult goal of annotating the unfolding of the emotional processes in conversations has been approached by capturing the transitions between neutral and emotionally connoted speech events as those transitions manifest themselves in the melodic variations of the speech signals.

Acknowledgements

The research leading to these results has received funding from the European Union - Seventh Framework Program (FP7/2007-2013) under grant agreement n 610916 SENSEI.

References

Alam, F., Riccardi, G. 2013. Comparative Study of Speaker Personality Traits Recognition in Conversational and Broadcast News Speech, *Proceedings of Interspeech-2013*

Alba-Ferrara, L., Hausmann, M., Mitchell, R.L., and Weis, S. 2011. The Neural Correlates of Emotional Prosody Comprehension: Disentangling Single from Complex Emotions. *PLoS ONE* 6(12): e28701. doi: 10.1371/journal.pone.0028701

Baron-Cohen, S., Tager-Flusberg, H., and Lomberdo, M.(Eds). 2013. *Understanding Other Minds*. London: Oxford Univ. Press

Barrett, L. F. (2006). Solving the emotion paradox: Categorization and the experience of emotion. *Personality and social psychology review*, 10(1), 20-46.

Barrett, L. F., & Lindquist, K. A. (2008). The embodiment of emotion. In Semin, G..R. & Smith, E.R. (Eds) *Embodied grounding: Social, cognitive, affective, and neuroscientific approaches*, Cambridge Univ. Press, New York, 237-262.

Brück, C., Kreifelts, B., & Wildgruber, D. 2012. From evolutionary roots to a broad spectrum of complex human emotions: Future research perspectives in the field of emotional vocal communication. Reply to comments on. *Physics of Life Reviews*, 9, 9-12.

Carletta, J., 1996. "Assessing agreement on classification tasks: the kappa statistics", *Computational linguistics* 22.2: 249-254.

Danieli, M. 2007. "Emotional speech and emotional experience". In Turnbull, O., & Zellner, M. 2010. International Neuropsychoanalysis Society: Open Research Days, 2002-2009. *Neuropsychoanalysis: An Interdisciplinary Journal for Psychoanalysis and the Neurosciences*, 12(1), 113-124.

Douglas-Cowie, E., Cowie, R., Schröder, M., 2003. The description of naturally occurring emotional speech. In: *Proceedings of the 15th International Conference on Phonetic Sciences*, Barcelona, Spain, pp. 2877-2880

El Ayadi, M., Kamel, M. S., & Karray, F. 2011. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3), 572-587.

Gallese, V. (2003). The roots of empathy: the shared manifold hypothesis and the neural basis of intersubjectivity. *Psychopathology*, 36(4), 171-180.

Gili Fivela, B., & Bazzanella, C. 2014. The relevance of prosody and context to the interplay between intensity and politeness. An exploratory study on Italian. *Journal of Politeness Research*, 10(1), 97-126.

Gross, J. J. 1998. The emerging field of emotion regulation: An integrative review. *Review of general psychology*, 2(3), 271.

Gross, J. J., & Thompson, R. A. 2007. Emotion regulation: Conceptual foundations. In J.J. Gross (Ed) *Handbook of emotion regulation.* The Guildford Press, New York.

Joachims, T., 1998. *Text categorization with support vector machines: Learning with many relevant features*, Springer.

Kononenko, I. 1994. "Estimating Attributes: Analysis and Extensions of RELIEF", *European Conference on Machine Learning*, 171-182.

Kotz, S.A., and Paulmann, S. 2007. When Emotional Prosody and Semantics Dance Cheek-to-Cheek: ERP Evidence. *Brain Research*, vol. 1151, pages 107-118.

Lietz. C., Gerdes, K.E, Sun, F., Geiger Mullins J., Wagaman, A., and Segal, E.A. 2011. The Empathy Assessment Index (EAI): A Confirmatory Factor Analysis of a Multidimensional Model of Empathy, *Journal of the Society for Social Work and Research*. Vol. 2, No. 2, pp. 104-124, 2011.

Magno Caldognetto, E. 2002. I correlati fonetici delle emozioni. In C. Bazzanella & P.Kobau, 2002. *Passioni, emozioni, affetti*, Mc Graw Hill, Milano, 197-213.

McKeown, G., Valstar, M., Cowie, R., Pantic, M., & Schroder, M. 2012. The SEMAINE database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *Affective Computing, IEEE Transactions on*, 3(1), 5-17.

Ochs, M., Pelachaud, C., & Sadek, D. 2007. Emotion elicitation in an empathic virtual dialog agent. In *Prooceedings of the Second European Cognitive Science Conference (EuroCogSci).*

Rao, K. S. and Koolagudi, S.G. 2013. *Emotion Recognition Using Speech Features*. Springer. New York.

Schmidt, T. 2004. Transcribing and annotating spoken language with EXMARaLDA, *Proc. of LREC 2004 Workshop on XML-based Richly Annotated Corpora*.

Shrikanth S. Narayanan and Panayiotis G. Georgiou. 2013. Behavioral Signal Processing: Deriving Human Behavioral Informatics from Speech and Language, *Proceedings of IEEE*, 101 (5) : 1203-1233.

Sperber, D., & Wilson, D. 2002. Pragmatics, modularity and mind-reading. *Mind & Language*, 17(1 - 2), 3-23.

Tesser, F., Cosi, P., Drioli, C., Tisato, G., & ISTC-CNR, P. 2004. Modelli Prosodici Emotivi per la

Sintesi dell'italiano. *Proc. of AISV 2004*: http://www2.pd.istc.cnr.it/Papers/PieroCosi/tf-AISV2004.pdf

Titchener, E. B. 1909. *Experimental Psychology of the Thought Processes*. Macmillan, London.

Wambach, I.J.A., and Jerger, J.F.. 2004. Processing of Affective Prosody and Lexical Semantics in Spoken Utterances as Differentiated by Event-Related Potentials. *Cognitive Brain Research*, 20 (3): 427-437.

Wilson-Mendenhall, C. D., Barrett, L. F., Simmons, W. K., & Barsalou, L. W. (2011). Grounding emotion in situated conceptualization. *Neuropsychologia*, 49(5), 1105-1127.

Zovato, E., Sandri, S., Quazza, S., & Badino, L. 2004. Prosodic analysis of a multi-style corpus in the perspective of emotional speech synthesis. *Proc. ICSLP 2004*, Vol. 2: 1453-1457.

Zovato, E., Tini-Brunozzi, F., and Danieli, M. 2008. "Interplay between pragmatic and acoustic level to embody expressive cues in a text-to-speech system", *Proc. Symposium on Affective Language in Human and Machine*, AISB 2008.