A two dimensional annotation scheme for emotion in dialogue

Richard Craggs, Mary McGee Wood
Department of Computer Science
University of Manchester
Manchester M13 9PL U.K.
{craggs, mary}@cs.man.ac.uk

Abstract

An increased recognition of how an understanding of emotion could benefit dialogue systems means that corpora annotated for emotion would be a valuable resource. In this paper we describe an annotation scheme for expressions of emotion in dialogue using abstract dimensions, discuss how to evaluate such a scheme and use those techniques to assess the scheme's reliability.

Dialogue analysis and emotion

To date dialogue annotation has been conducted almost exclusively on objective, measurable aspects of a dialogue's content. This is partly understandable due to the difficulty associated with reliably identifying subjective phenomena.

However, recently there has been an increased interest in how an understanding of some of the more subtle facets of communicative behaviour might benefit this field. Emotion is an example of the type of behaviour that it would be beneficial to incorporate into dialogue systems. For instance, artificial communicative agents which conduct dialogues either between themselves or with humans can be allowed to converse more naturally and be more engaging if they appear to act emotionally (Bates 1997; Piwek 2003).

In order to gain an understanding of how emotion affects the way we communicate or how emotional expression is perceived by a listener, we could study corpora annotated for emotional content. In this paper we shall discuss the difficulties that hinder the development of annotation schemes for this purpose and describe a suitable scheme that overcomes these difficulties.

Emotion in transcribed dialogue

There has been considerable work studying the relationship between emotion and prosody, e.g. (Lee, Narayanan, & Pieraccini 2002b; Dellaert, Polzin, & Waibel 1996). Indeed much of the emotional expression in speech is conveyed in its acoustic realisation. However, emotion is also expressed in the language used by a speaker and it is equally important to understand how emotion affects the content of what we say. Knowing this would allow us to work with emotion in cases where only the linguistic content was available or necessary and this understanding would also complement work on acoustic correlates.

Since identifying emotion from transcribed speech can prove more difficult than from an audio recording, annotation schemes designed to be used with audio are likely to prove unsuitable for working from transcriptions. This is because these schemes ask annotators to make distinctions or apply labels for emotions which it is not possible to detect in text. Furthermore, a scheme which is shown to be reliable for annotators listening to the dialogue may not be when the vocal delivery is not available.

The scheme that we describe here is designed to be applied to transcribed dialogue and is evaluated in that context.

Previous studies of emotion in dialogue

There have been a handful of studies which have developed annotation schemes for emotion in dialogue, including the following examples –

- (Narayanan 2002; Lee, Narayanan, & Pieraccini 2002a) This scheme was applied to call centre dialogues and included two labels, *Negative* and *Positive*. Given the simplicity of the scheme and the presence of audio during the annotation, the reliability of the scheme was low (Kappa = 0.465), which reflects the difficulty of the task.
- (Devillers, Vasilescu, & Lamel 2002) Another study using call centre dialogues, this time the scheme includes labels for specific emotions; *Fear*, *Anger*, *Excuse* and *Satisfaction*. The choice of labels was tailored to their domain although no evaluation of the reliability of this scheme was conducted.
- (Litman, Forbes, & Silliman 2003; Litman & Forbes 2003) This scheme was applied to transcribed tutoring dialogues between humans. This scheme included labels for *Negative* and *Positive* utterances and was proven to be reliable (Kappa = 0.624).

Later we shall discuss how these approaches can be improved upon in order to develop a scheme that is suited to capturing the complexities of emotion while remaining reliable.

Designing an annotation scheme

Before we describe the design of our annotation scheme it can be useful to consider how annotation schemes are designed in general. Here we describe a number of design decisions that must be made and highlight any differences between emotion and more frequently annotated phenomena that will affect the decisions we make.

Dialogue Units The are a number of unit types into which a dialogue can be segmented and to which annotation can be applied. Since many annotation schemes are motivated by some theory about language or communication the choice of unit is often implicit or obvious. This is not the case for emotion. Emotional episodes exist over an indistinct period of time, fading in and out and subtly changing throughout a dialogue.

Descriptive Mechanisms There are a number of ways with which to describe aspects of communication (e.g. descriptive labels, numeric values). Again, the appropriate way of describing phenomena used in dialogue annotation is often obvious and is most commonly descriptive categorical labels. For emotion this would be manifest as a collection of labels such as *'Anger'*, *'Fear'*, *'Sadness'* etc. For many purposes this is an inadequate way of describing emotion and subsequently a number of different approaches have been proposed (Cowie & Cornelius 2003).

Annotation Scheme For the chosen descriptive mechanism it remains to decide upon the details of how the annotation is to be conducted. For categorical labels this involves producing a taxonomy of categories based on the required level of complexity, the necessary distinctions and the range of phenomena it is desirable to identify. Similar decisions need to be made for other descriptive mechanisms.

Developing an annotation scheme for emotion in dialogue

With an understanding of some of the important factors for consideration in developing an annotation scheme, we now describe the development of our scheme for emotion in dialogue.

What exactly are we annotating?

To simply state that we are annotating *emotion* in dialogue is not adequate given the complexity of the phenomenon. In order for an annotated corpus to be useful we must be more specific about what the annotation entails. To this end, we draw upon a valuable distinction given in (Cowie 2000) between cause-type and effect-type descriptions of emotion.

Cause-type descriptions of emotion relate to the 'internal states and external factors that caused a person's speech to have particular characteristics'. To adopt this approach would mean that the dialogue is annotated for the emotion that the speaker is experiencing as they speak. A corpus annotated as such would allow us for example, to attempt to ascertain whether a customer conversing with an automated call centre was experiencing frustration.

The major difficulty of working with cause-type description is that we have to be confident that we have correctly identified the emotion that a speaker is experiencing. Given that speakers can successfully disguise their emotions or falsely exhibit them, asking annotators to guess how a speaker is feeling is likely to lead to inaccurate results. In this case, inter-rater agreement offers no indication of validity and the only reliable way of ensuring a dialogue is correctly labelled is for the annotation to be performed by the speakers themselves, so-called 'self-reporting'.

Effect-type descriptions of emotion are used to 'describe what effect [emotional] characteristics would be likely to have on a typical listener'. Using this approach annotation can be performed by asking annotators to label the emotion that they believe is being expressed. A corpus annotated using this approach allows us to study the type of language use that causes a listener to believe that emotion is being expressed. This is especially useful for automatically generating speech that will be perceived as emotional.

Obtaining an annotated corpus for effect-type emotion is a much simpler affair than for cause-type since identifying the emotion that we believe is being expressed is a far more objective exercise than guessing the speaker's emotional state. Our annotation scheme uses the former of these and is thus applied by identifying the emotion that is expressed in its content.

Units of consideration

For our annotation scheme, dialogues are segmented into *Utterances* as described in (Traum & Heeman 1996) as "Speech by a single speaker, speaking without interruption by speech of the other constituting a single turn".

For practical reasons this is a sensible choice since their use is common in dialogue annotation. Adopting these units will allow us to benefit from tools and techniques developed for other schemes and also allow comparisons to be made between emotion and other phenomena. Utterances also appear to form appropriate units of emotional expression.

Descriptive mechanism

As previously mentioned the most frequently adopted descriptive mechanism for dialogue annotation is categorical labels. Applying this approach to emotion is likely to prove problematic. There is little consensus regarding a suitable taxonomy of emotional labels, with many psychologists proposing different lists (Ortony & Turner 1990).

If we were to develop a categorical scheme we would need to select labels from the hundreds available, between which annotators are able to distinguish and reliably identify in transcribed dialogue. Restricting ourselves to these emotions will almost certainly lead to the majority of utterances in any corpus to being left unlabelled.

To demonstrate the difficulty of developing a scheme using categorical labels, we conducted an experiment in which four subjects annotated a dialogue using a list of labels gleaned from psychological research [1]. This dialogue was

[1] Courage, Dejection, Sadness, Disgust, Aversion, Shame, Anger, Surprise, Guilt, Wonder, Hate, Love, Happiness, Desire,

taken from our corpus of conversations between nurses and cancer patients (Wood & Craggs 2002). In the resulting annotation only 16% of the utterances were labelled with an overall agreement level of Alpha = 0.14[2]. This is clearly unsatisfactory and an alternative approach must be employed.

A more fruitful approach may be to describe emotions using abstract dimensions, a technique in which emotions can be described by points in dimensional space. Litman and Forbes' scheme (Litman & Forbes 2003) which labels utterances as *Positive*, *Negative* or *Neutral* could be considered to adopt this approach, with utterances plotted on one of three discrete points on a valence scale.

This technique can be made more expressive if utterances are labelled using more dimensions and with finer granularity. Our scheme is based on an adaptation of *Activation–Evaluation space* (Cowie *et al.* 2001) in which an emotion is described using two values, one for the degree to which that emotion inspires action in humans and another which describes how *positive* or *negative* one might consider that emotion to be.

Constructing an annotation scheme using abstract dimensions

In order to make activation–evaluation space appropriate for annotation of transcribed dialogue, a number of adaptations were made.

Firstly, we replaced the 'activation' dimension with a simpler 'intensity'. Although activation–evaluation may be a better way of describing emotions, they are less suited to describing an instance of emotional expression. The intensity dimension is used to describe the level of emotion expressed in utterance. For example –

`'I like this song'`

would be given a lower emotional intensity score than –

`'Oh wow, this song is fantastic!'`

This adaptation should make corpora annotated using our scheme no less useful since for many applications, an understanding of the relationship between language use and the intensity of emotion could prove more valuable than for activation.

The second adaptation relates to the granularity of distinctions that can be made along each scale. Each dimension can be considered to be a continuous scale where any level of distinction is possible. Reliably making such fine grained distinctions based on transcribed speech is not possible and allowing this level of discrimination is likely to perplex annotators.

For this reason our annotation scheme condenses the possible values for each dimension into a likert–scale style set of discrete values; *intensity* can be a value from 0 to 4 and *evaluation*, -3 to $+3$. This permits us to prescribe to annotators a description of the circumstances in which each values should be used including examples for each.

The final adaptation limits annotators to ascribing one value in each dimension for the *overall* emotional expression in any utterance. Although it is obviously possible to express more than one emotion per utterance, it is relatively rare for a speaker to do so. In an experiment in which four annotators labelled 1075 utterances from dialogues in our corpus using categorical labels for emotions of their own choosing, only 3.2% of the utterances required more than one label to be applied.

Allowing annotators to apply more than one value for each dimension would increase the scheme's complexity, incurring greater disagreement and lengthening annotation time, for little benefit.

A two dimensional annotation scheme for emotion in dialogue

What follows is a description of our scheme and instructions regarding how it should be applied –

For each utterance, label the overall intensity and polarity of the expression of emotion using the following guidelines.

Intensity

0 No emotion or it is impossible to tell – *"So how are you?"*

1 Not totally lacking in emotion, (a hint of) – *"I suppose so"*

2 low level, but apparent – *"I'm not feeling too great"*

3 Clear expression of emotion – *"Oh she's annoying that girl"*

4 Strong expression of emotion – *"I can't bear to talk about it"*

Evaluation

-3 Wholly/Strong negative – *"It was the most awful feeling"*

-2 Clearly negative – *"He tries, but he keeps messing it up"*

-1 Perhaps negative (but not positive) – *"You know, the stupid one"*

neutral Neutral or impossible to tell – *"He's moved to Blackburn"*

+1 Perhaps positive (but not negative) – *"Oh yes, that new show"*

+2 Clearly positive – *"That's a nice view"*

+3 Wholly/Strong positive – *"Oh that is wonderful news"*

For example the utterance –

`'Oh how we laughed ...'`

may be labelled as *intensity=3, evaluation=+2* meaning that it contains a clear level of clearly positive emotion.

Contempt, Fear (Ortony & Turner 1990)

[2] Agreement statistics for annotation are discussed in detail later in this paper.

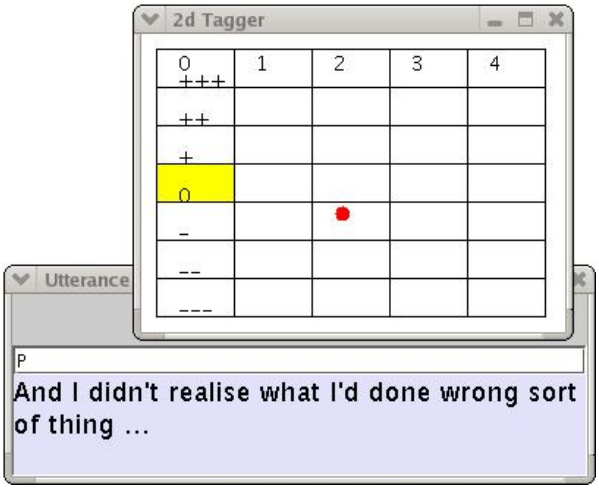

Figure 1: Dimensional annotation tool

Evaluating the annotation scheme

In order for analysis of annotated corpora to be valid the schemes used to develop them must be shown to be reliable. The evaluation of our scheme was conducted on a sample of two dialogues from our corpus. Although these dialogues do come from a specialist domain (cancer care) the conversations within them contain a large number of topics and styles. Due to the circumstance in which they are conducted they also contain a large amount of emotional speech.

Eleven annotators applied our scheme to an extract of 34 utterances from one dialogue while ten different annotators worked on an extract containing 35 utterances from another dialogue. The samples were chosen for the variability of the emotional expression within them, thus representing a serious test with which to conduct the evaluation.

Each annotator was given written instructions on how to apply the scheme with the definitions and examples presented in the previous section. The annotations were performed using a tool which presented each utterance in turn, waiting for the annotator to select the appropriate values on a two dimensional grid before continuing onto the next utterance (see figure 1).

Nominal inter–rater reliability: Kappa and Alpha

Efforts have been made to standardise the way in which annotation schemes are evaluated (Carletta 1996) with the consensus being on the use of Kappa (Siegal 1988) as an inter–rater agreement test.

Kappa measures the agreement between a number of annotators by comparing the number of times they agree upon a label for an object (utterances in our case) against the number of disagreements. The result is a value between 0 and 1, where zero equates to the level of agreement that might be expected if annotators behaved randomly and one represents perfect agreement.

There is prevailing confusion about the suitability of Kappa as a method of evaluating reliability (Di Eugenio & Glass 2004). The problem resides in the way that Kappa accounts for the unequal preference by coders for certain categories. This is especially worrisome when working with subtle or rare phenomena since the distribution of label use is likely to be very uneven.

Fortunately, there are alternative statistical tests that do not suffer this inadequacy. One such test, Krippendorff's Alpha (Krippendorff 1980; 2004) is highly appropriate for testing reliability of this kind of data, partly because its results and their interpretation are similar to Kappa's and also because its flexibility allows us to tailor its computation to best evaluate data annotated in the manner described.

In order to perform this test on our data we must consider each pair of values given for an utterance to be one of 29 classes, one class for each possible pair of intensity–evaluation values excluding positive and negative evaluation for intensity 0 (annotators were advised that if they judge an utterance to have no level of emotion then it makes no sense to allocate it an evaluation level of anything other than zero).

Alpha can also be performed on individual dimensions by considering each value as a separate class. The table below shows the results for performing the Alpha test on each dimension and the combined dimensions for each of the two dialogues in our trial.

Transcript and Test	Alpha
Transcript 1: Intensity–Activation	0.145
Transcript 1: Intensity	0.25
Transcript 1: Evaluation	0.26
Transcript 2: Intensity–Activation	0.13
Transcript 2: Intensity	0.211
Transcript 2: Evaluation	0.217

The results suggest a low level of reliability. However, this has more to do with the suitability of the test to qualitative annotation than the ability of the annotators to apply the scheme.

The slight increase in the level of agreement when applying Alpha to individual dimensions provides a clue as to why categorical agreement tests penalise annotation using numerical values. Take for example the utterance –

`'Right, there's nothing on your mind at the minute?'`

In the intensity dimension three people rated this as 'no emotion' (intensity 0), seven people said 'hint of emotion' (intensity 1) and one rated it 'low level emotion' (intensity 2). There is agreement that this utterance contains a hint of emotion although a few people considered it slightly above or below this level. This 'diffusion' of the values around the one generally agreed upon, weakens the results.

Now, also considering the evaluation dimension, one person thought it was 'possibly positive' (evaluation $+1$), eight people said 'neutral' and two considered it 'possibly negative' (evaluation -1). Again there is reasonable agreement between all annotators with slight variation. When you combine these two dimensions only four people agreed on the most popular label even though overall, there was a reasonable consensus as to the intensity and evaluation level of this utterance.

This problem is caused by the fact that our annotators are labelling using numerical scales. Neither Kappa nor Alpha in its original form distinguish between the cases of two people disagreeing wildly about an utterance and people agreeing on one dimension but being slightly in disagreement on the other.

In order to fairly test the reliability of our scheme, we must employ another, more suitable statistical test. We shall consider two alternatives, interval and ordinal agreement measures.

Interval inter–rater reliability

Testing agreement using intervals is performed by rewarding annotators for assigning similar values for objects while punishing disparity. This would appear to be a more appropriate test for our scheme but again it is not suitable. The reason is that agreement testing using intervals is only valid when the difference between each adjacent pair of values is equal.

This is not the case with our scales and neither is it desirable to make it so. Given that full blown emotional episodes are rare, it is important to be able to make subtler distinctions for lower levels of emotion, ergo the difference between intensity 0 (no emotion) and intensity 1 (hint of emotion) is finer than between intensity 3 (clear) and intensity 4 (strong).

Ordinal inter–rater reliability

When objects are rated on a scale which cannot be assumed to be linear the reliability of the scheme used to generate those ratings can best be assessed using ordinal agreement measures.

These measures test agreement by assessing the similarity between the order in which judges rank objects. In our case this tests the degree to which annotators agree on the order of utterances from the least to the most emotional and from the most positive to the least positive.

This is especially appropriate for our scheme since insights about the relationship between emotional expression and communicative behaviour ought to be made in relative terms, e.g. –

"The higher the level of emotion, the smaller the lexical diversity."

rather than in absolute terms, e.g –

"For intensity three emotion, people swear."

One suitable test for assessing the level of agreement for rank ordered data is Alpha using an ordinal metric ($_{ordinal}\alpha$) in which the level of disagreement between any two annotators for any utterance is given by the distance between the ranks into which they are placed, relevant to the number of objects placed in each rank. The benefit of using $_{ordinal}\alpha$ as our measure of agreement is that it allows us to judge reliability against the standards adopted by the field when applying Kappa.

Applying $_{ordinal}\alpha$ to each dimension for both extracts we get the following results –

Transcript and Test	$_{ordinal}\alpha$
Transcript 1: Intensity	0.6
Transcript 1: Evaluation	0.55
Transcript 2: Intensity	0.49
Transcript 2: Evaluation	0.34

These results are less than those reported for the scheme using the labels *Negative* and *Positive* described in (Litman, Forbes, & Silliman 2003), although we believe that this is more than compensated for by the far greater expressive ability of our scheme.

Intra–rater reliability

Another important test of a scheme's reliability is whether the results of a single annotator using it over a period of time remain constant. To assess this, our scheme was applied six times to the 34 utterances of the first dialogue extract over a period of four months. Applying $_{ordinal}\alpha$ to each dimension gave the results: Intensity = 0.9, Evaluation = 0.78, both of which show satisfactory levels of agreement.

Agreement for a finer grained scale

The decision regarding the number of discrete units into which our scale was condensed was based on a desire to make each level easily distinguishable from its neighbours. In order to justify this decision, we tested the effects of increasing the granularity by conducting another set of annotations using a scheme with six values for the *intensity* dimension, introducing an additional level, thus –

0 No emotion or it is impossible to tell – *"So how are you?"*

1 Not totally lacking in emotion, (a hint of) – *"I suppose so"*

2 low level, but apparent – *"I didn't want to believe it at first"*

3 Moderate expression of emotion – *"Hopefully I should be out next week"*

4 Clear expression of emotion – *"Oh she's annoying that girl"*

5 Strong expression of emotion – *"I can't bear to talk about it"*

This time a different set of annotators, four for each of the two transcripts, applied the scheme to dialogue. The results of applying $_{ordinal}\alpha$ showed a systematic fall in the level of agreement –

Transcript and Test	$_{ordinal}\alpha$
Transcript 1: intensity	0.52
Transcript 2: intensity	0.37

This represents a reduction of around 20% which the increased expressive power of the scheme would not justify.

Conclusion

In this paper we have shown that despite the perceived difficulty in developing annotation schemes for subtle or subjective phenomena, a reliable scheme with strong expressive powers has been created for emotion in transcribed dialogue. Employing abstract dimensions in order to annotate emotion has the advantage of allowing all utterances in a dialogue to be labelled rather just those that display distinct emotion. Also, the scheme is equally applicable to dialogues conducted in different domains and languages.

In developing a general purpose scheme for expressions of emotion in dialogue we hope to enable and hasten studies into the relationship between emotion and communicative behaviour. This understanding will be valuable for improving the functionality of dialogue systems.

Acknowledgements

The authors would like to acknowledge Ian Fletcher and Peter Maguire for supplying the corpus on which the agreement tests were conducted, Cancer Research UK for the funding of the development of that corpus and Manchester University for funding Richard Craggs. We would also like to express our appreciation of the efforts of the volunteers who participated in our trials.

References

Bates, J. 1997. The role of emotion in believable agents. *Communications of the ACM* 37(7):122–125.

Carletta, J. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics* 22(2):249–254.

Cowie, R., and Cornelius, R. R. 2003. Describing the emotional states that are expressed in speech. *Speech Communication* 40(1):5–32.

Cowie, R.; Douglas-Cowie, E.; Tsapatsoulis, N.; Votsis, G.; Kollias, S.; Fellenz, W.; and Taylor, J. 2001. Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine* 18.

Cowie, R. 2000. Describing the emotional states expressed in speech. In *SpeechEmotion-2000*.

Dellaert, F.; Polzin, T.; and Waibel, A. 1996. Recognizing emotions in speech. In *Proc. ICSLP '96*, volume 3, 1970–1973.

Devillers, L.; Vasilescu, I.; and Lamel, L. 2002. Annotation and detection of emotion in a task-oriented human-human dialog corpus. In *Proceedings of ICSLP 2002*.

Di Eugenio, B., and Glass, M. 2004. The kappa statistic: a second look. *Computational Linguistics*.

Krippendorff, K. 1980. *Content Analysis: An Introduction to its Methodology*. Beverly Hills, CA: Sage Publications.

Krippendorff, K. 2004. *Content Analysis: An Introduction to its Methodology: Second Edition*. Beverly Hills, CA: Sage Publications.

Lee, C.; Narayanan, S.; and Pieraccini, R. 2002a. Combining acoustic and language information for emotion recognition. In *Proceedings of ICSLP*.

Lee, C.; Narayanan, S.; and Pieraccini, R. 2002b. Recognition of negative emotions from the speech signal. In *Proceedings of Automatic Speech Recognition and Understanding*.

Litman, D., and Forbes, K. 2003. Recognizing emotions from student speech in tutoring dialogues. In *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop ASRU*.

Litman, D.; Forbes, K.; and Silliman, S. 2003. Towards emotion prediction in spoken tutoring dialogues. In *Proceedings of the Human Language Technology Conference: 3rd Meeting of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL)*.

Narayanan, S. 2002. Towards modeling user behavior in human-machine interaction: Effect of errors and emotions. In *Proceddings of ISLE*.

Ortony, A., and Turner, T. J. 1990. What's basic about basic emotions? *Psychological Review* 97:315–331.

Piwek, P. 2003. A flexible pragmatics-driven language generator for animated agents. In *Proceedings of EACL03*.

Siegal, S. 1988. *Nonparametric statistics: Second edition*. McGraw–Hill.

Traum, D. R., and Heeman, P. A. 1996. Utterance units in spoken dialogue. In *ECAI Workshop on Dialogue Processing in Spoken Language Systems*, 125–140.

Wood, M., and Craggs, R. 2002. Rare dialogue acts common in oncology consultations. In *Proceedings of SIGdial3*.