A categorical annotation scheme for emotion in the linguistic content of dialogue

Richard Craggs and Mary McGee Wood

Computer Science Department
The University of Manchester
Manchester, England
M13 9PL

Abstract. If we wish to implement dialogue systems which express emotion, dialogue corpora annotated for emotion would be a valuable resource. In order to develop such corpora we require a reliable annotation scheme. Here we describe an annotation scheme for emotion in dialogue using categorical labels to complement previous work using dimensional scales. The most difficult challenge in developing such a scheme is selecting the categories of emotions that will yield the most expressive yet reliable scheme. We apply a novel approach, using a genetic algorithm to identify the appropriate categories.

There is increasing recognition of a need to incorporate an understanding of emotion into dialogue systems, since this understanding can greatly enhance their performance. For example, artificial communicative agents can be made to converse more naturally and appear more engaging by having characters behave emotionally [1, 2].

For gaining an understanding of the relationship between emotion and dialogue, which would allow us to implement such systems, dialogue corpora annotated for emotional content would be a valuable resource. To develop such corpora, it is first necessary to develop an annotation scheme that yields rich and reliable results. In this paper we will describe such a scheme which employs the common approach of descriptive labels applied to segments of dialogue.

1 Annotating emotion in transcribed speech

To date, much of the work studying the relationship between speech and emotion has concentrated on the prosodic properties of speech. Although much emotional expression is conveyed in speech's acoustic realisation, it is equally important to understand how it is conveyed in the linguistic content. For example, if we are to generate speech that will be perceived to be emotional, it is not sufficient to construct a semantically and syntactically suitable, yet impassive utterance and then impose emotional intonation onto it. In order to be realistic, the content of speech must also express emotion to some degree.

Besides complementing prosodic expressions of emotion in multi-modal dialogue systems, an understanding of the relationship between emotion and linguistic content will facilitate emotional dialogue systems where only the textual content is necessary or available.

Since emotion is easier to identify with audio or visual context, it is tempting to annotate multi-modal dialogue and then draw conclusions about the relationship between the annotation and the linguistic content. However, since the annotation was not based solely on the information found in the linguistic content, the reliability of any conclusions drawn about this relationship must be doubted. In order to draw justifiable conclusions, the annotation must be applied to transcribed dialogue using a scheme evaluated in this context.

2 Previous studies

Some studies have endeavoured to annotate emotion in dialogue. Some annotated general emotional states such as `Positive` and `Negative` [3, 4], while others used specific labels for emotions [5].

Since part of the aim of the work using emotional states was to develop systems to detect the emotion of speech based on its acoustic features, it was necessary to make only very coarse distinctions. In order to study the relationship between emotion and speech it would be valuable to make finer distinctions, allowing more detailed analysis. Choosing specific emotion categories for an annotation scheme is a difficult task. Laurence Devillers' scheme [5] contained labels that were specific to the call centre application for which it was developed, and it is unclear whether these labels would be useful in a broader context.

The scheme that we propose incorporates both these types of annotation, but differs from previous attempts by allowing annotators to make finer distinctions for emotional states, and by including labels that can most easily be identified within transcribed dialogue.

3 Expressions of emotion

Since *emotion* is a very general term that may refer to a number of different things, we need to describe more specifically what we wish to annotate. Roddy Cowie distinguishes two types of descriptions of emotions for speech studies, cause-type and effect-type [6]. This is similar to Austin's distinction between Illocutionary acts and their Perlocutionary effects [7].

Cause-type descriptions relate to the 'internal states and external factors that caused a person's speech to have particular characteristics'. Analysis of dialogue annotated for this type of phenomena would be useful for detecting a speaker's emotion based on their speech. An example of a type of dialogue system that would benefit from this understanding would be those that wish to recognise user emotion and adapt their behaviour accordingly. If we annotate emotion in this form, we are attempting to guess the emotion that the speaker was experiencing whilst speaking. Since humans can successfully disguise their emotions or falsely exhibit others, this would be a very difficult task, and there is no way of knowing if our annotation is correct.

Effect-type descriptions ‘describe what effect [the characteristics of speech] would be likely to have on a typical listener’. We could annotate this type of phenomenon by labelling speech for the emotion that we perceive as being expressed, regardless of whether the speaker was experiencing, or even *trying* to express that emotion. An understanding of the relationship between the content of speech and the emotion that listeners perceive within it, would be especially useful for automatically generating emotional speech. Furthermore, since annotators are labelling the effect that the speech had on them rather than guessing the emotion experienced by the speaker, the annotation is more objective and intrinsically valid for each annotator.

Because of the practical uses of effect-type annotation and also because it is likely to produce more valid and reliable results, our annotation scheme is used to label the *perceived expression of emotion* in dialogue.

4 Developing an annotation scheme for emotion in dialogue

Designing annotation schemes and evaluating their quality is rarely a trivial task. For subtle, rare and subjective phenomena the task becomes substantially more difficult. The two main challenges that must be overcome are –

Reliability The data that an annotation scheme produces must be shown to be reliable before its analysis is valid. Broadly, reliability reflects the clarity of the mapping of units of data onto categories which describe that data. In turn, this reflects the degree to which there is a shared understanding of the meaning of the phenomena in question. Reliability can be inferred from the level of agreement achieved by a number of coders labelling the same data. An overview of agreement statistics for discourse and dialogue coding can be found in [8] and [9].
Labelling emotion is frequently referred to as a ‘subjective’ process, meaning that the mapping of data to categories is abstruse. The greatest challenge when developing a scheme for emotion in dialogue is designing it in such a way that the data that it produces is sufficiently reliable. We tackle this problem by attempting to identify labels for emotions upon which annotators can agree.

Coverage Coverage refers to the proportion of units of data that have labels applied to them during the annotation process. Since obtaining dialogue corpora is difficult and annotation time can be expensive, it is desirable for the application of an annotation scheme to result in as much labelled data as possible. Since episodes of strong emotion are rare in spoken dialogue, obtaining sufficiently high rates of coverage is another challenge in developing an annotation scheme for this phenomenon.

In order to assess how difficult it would be to develop a scheme which overcame these challenges, we developed a trial scheme containing labels for emotions used in psychological research; *Courage, Dejection, Sadness, Disgust, Aversion, Shame, Anger, Surprise, Guilt, Wonder, Hate, Affection, Happiness, Desire, Contempt and Fear* [10]. Four annotators used this scheme to label a dialogue containing 400 utterances, from our corpus of Cancer Patient/ Nurse conversations [11]. The results were discouraging, with only an average of 16% of the utterances labelled with an overall agreement level of 0.17[1].

The disappointing results from this trial led us to adopt an alternative approach to describing emotion, using abstract numerical scales. The results of this were much more encouraging. A description of this scheme was published in [13] and is summarised here.

4.1 A two dimensional annotation scheme for emotion in dialogue

This scheme is based on the notion that properties of emotions can be described as points on a numerical scale. Our approach is similar to the Activation–Evaluation space coding [14] used in the Feeltrace application [15] to track the emotion of speakers in multi-modal dialogue. In this application the perceived emotion of a speaker is tracked in two dimensions; Activation, which describes the degree to which that emotion inspires action in humans, and Evaluation which describes how *positive* or *negative* one might consider that emotion to be.

In order to produce a practical and usable scheme for dialogue we adapted this approach in a number of ways. Firstly, dialogues are segmented into utterances and values are applied to each individual utterance. One benefit of employing utterances is that the majority of other annotation schemes also use these as their basic unit, and aligning our scheme with others will allow us to make comparisons between layers of annotation. Also, since utterances are a convenient unit for the generation of dialogue, labelling at this granularity makes the results of analysis easier to apply. Although expressions of emotion do not always align with utterance boundaries, asking annotators to segment dialogue into smaller units would increase the complexity of the coding process, especially considering that the boundaries of emotional expression are rarely clear.

When applying our scheme, annotators are restricted to applying one pair of values for each utterance. Although it is possible to express more than one emotion within a single utterance it is relatively rare for speakers to do so. In an experiment that we conducted, in which annotators labelled emotions in a dialogue using labels of their own choosing, around only 3% of utterances required more than one label. Allowing annotators to apply more than one pair of values per utterance increases the complexity of the annotation process for little benefit.

[1] Agreement was measured using Krippendorff's alpha statistic [12]. The value should be interpreted as a level of agreement between 0 and 1, where 1 represents perfect agreement and 0 suggests that the coders did not understand the task and behaved randomly.

The next adaptation was that instead of *Activation*, our scheme used an *Intensity* dimension. Whereas Activation refers to the arousal of the person experiencing the emotion, it is not clear how this relates to the perceived expression of emotion within speech. Intensity describes the overall level of expression within an utterance, and this conceptually simpler dimension should be more easily understood and applied by annotators. It is also likely that intensity will serve as a more useful parameter for dialogue generation systems in which the level of emotional expression can be adjusted.

Finally, we wished to bestow some meaning on the values applied by the annotators. During the development of Feeltrace, it was recognised that coders performed more reliably when the locations of specific emotions were placed on their two dimensional plane. The introduction of reference points implies some meaning to the values within the dimensional space. We also suggest that we need to introduce some notion of scale to the dimensions, without which, values away from the reference points become difficult to interpret.

Reference points and scale are introduced into our two dimensional scale implicitly by dividing the continuous dimensional space into Likert-scale style sets of discrete values. *Level* can be a value from 0 to 4 and *evaluation*, -3 to $+3$. This allows us to describe the meaning of each value and give examples of circumstances in which that value should be used (see section 6). Making the meaning of each value explicit also should reduce the subjectivity in the coding process.

4.2 Why the need for a categorical scheme for annotating emotion in dialogue?

The dimensional model was used to create our annotation scheme because of the difficulty we observed in developing a categorical scheme which would elicit satisfactory reliability and coverage. While we believed that agreement could be increased by concentrating on a selection of emotions that could be reliably identified by annotators, this would reduce coverage to intolerable levels.

The prospect of a categorical annotation scheme for expression of emotion remains attractive. If we can understand how a small number of emotions are expressed in speech, this could be exploited to make artificial communicative agents more realistic. There are cases in which it is possible to identify specific emotions, ‘surprise’ being a common example. It is a shame that even though annotators may agree that an utterance expresses a specific emotion they are limited to using numeric values to describe it.

Now that we have a scheme that allows annotators to label every utterance for its emotional expression regardless of how subtle it may be, we may augment that scheme with nominal labels for emotions that can be reliably identified within transcribed dialogue.

4.3 Developing the categorical annotation scheme

A categorical annotation scheme for emotion may be considered to be a collection of labels for emotions which annotators may apply to individual utterances. A wide range of different lists of emotions has been proposed in psychological research, from the popular *big six* (anger, fear, happiness, sadness, surprise, and disgust) to larger lists of emotional related states. Choosing an appropriate list for an annotation scheme would seem a daunting process. However we can be guided by the proposed application of the analysis of the annotated data (developing an understanding of how to express emotion in speech). We might consider that the most useful emotions to annotate would be those that can most easily and reliably identified by humans.

To derive a suitable list of emotions, an experiment was conducted in which four annotators labelled three dialogues each of around 400 utterances. Instead of being prescribed a set of labels, they were allowed to use any labels that they believed appropriately described the emotions expressed. They were also allowed to use more than one label per utterance if necessary. The results look something like this –

```
N1: Right, how are you this morning []
P2: Not too brilliant [despondency]
N3: No? What's the problem [interest]
P4: I've had some right bad breathing do's [anxiety]
⋮
```

Over 100 different labels were invented by the annotators. Permitting them to choose labels for individual utterances resulted in the chosen labels being very fine grained and specific (e.g. `Disgruntled, Reflective, Resolute`). Some of them were inappropriate for our annotation scheme because what they describe is not clearly an emotion (e.g. `Friendly, Confirmation`). In order to arrive at our final list of emotions we can group these fine–grain labels into broader categories and ignore the ones that are not appropriate or cause too much disagreement. For example, if the labels *Anxiety, Concern* and *Worry* cause disagreement between annotators we could achieve greater reliability by combining the three into a single label.

This grouping and ignoring can be considered as a search procedure, where we are attempting to find an arrangement of labels into groups which results in the highest level of agreement. An appropriate way of finding suitable arrangements is to employ the artificial intelligence heuristic search procedure known as a genetic algorithm.

Genetic algorithms can explore very large search spaces by applying a 'survival of the fittest' evolution process to candidate solutions to a problem. Solutions to our problem consist of an arrangement of labels into groups, including an *ignore-me* group for labels that will not appear in the final scheme. Our algorithm generates a random population of candidate solutions and at each iteration, discards a proportion of the weakest solutions then refills the population by combining and mutating the better solutions.

We applied the state of the art genetic algorithm SPEA2 [16], which simply required us to write a function that would return a value denoting the fitness of any given candidate solution. The simplest function for this purpose could return the level of agreement that would be achieved, if instead of using the individual labels, annotators used a single label, common to all members of the group in which the label was placed. Unfortunately this simplistic function encouraged the algorithm to ignore lots of labels and over-zealously group other labels into very large collections.

In order to yield more useful results, the algorithm was was asked to satisfy multiple objectives –

Agreement Solutions were rewarded for resulting in high levels of agreement between annotators.

Coverage Higher scores were awarded when greater numbers of utterances were labelled using the solution's scheme. This discouraged the algorithm from ignoring too many labels.

Entropy Solutions which balanced the frequency with which each group of label were used, were preferred to those with an unequal distribution. This discouraged the algorithm from forming *super-groups* by bundling many labels together.

Coherence In an attempt to balance the distribution of groups to satisfy the entropy objective, the algorithm tended to merge groups of unrelated emotions. To avoid this, each individual emotion was labelled using the dimensional scale described in section 4.1 and this was used in the calculation of a group's coherence. A solution's coherence score is penalised for placing disparate labels within a group.

When searching for high quality solutions using more than one objective, we are attempting to find the *Pareto optimal solutions*. These are only those solutions for which there is no other solution which is superior for every objective.

For each arrangement into groups we can derive a list of emotions by selecting a label for each group which adequately describes the emotions within it. The choice as to which label would best describe the group was a personal decision but for most cases the appropriate label was obvious. For example if it was suggested that we group `Anxiety, Worry, Concern` and `Trepidation` we may decide to use the label `Worry` to describe that group.

5 The categorical annotation schemes for emotion in dialogue

Rather than selecting one definitive list of labels that are to be used in our annotation scheme, we have chosen three, each with fewer, coarser grained categories than the last. These are shown in table 1 and are hereafter referred to as the fine-grained, medium–grained and coarse–grained schemes.

The labels that constitute the fine-grained group schemes are those that appeared most frequently in the Pareto optimal set of solutions. It also includes

`Affection` and `Anger` since for the experiment described in section 4 these two emotions caused considerably less disagreement than the others. The medium–grained and coarse–grained schemes were created by conflating categories that were sometimes grouped by the genetic algorithm. Specifically, `Worry` and `Fear` became `Worry`; `Contentment` and `Joy` became `Happiness`; `Dislike` and `Misery` became `Sadness`; `Positivity, Happiness` and `Affection` became `Happiness` and finally `Sadness` and `Anger` became `Anger`. `Frustration` and `Positivity` were removed from the fine and medium grained schemes respectively as part of the reduction process.

Fine	**Medium**	**Coarse**
Anger	Anger	Happiness
Affection	Affection	Sadness
Contentment	Hope	Worry
Dislike	Happiness	Hope
Frustration	Positivity	Surprise
Fear	Sadness	
Hope	Surprise	
Joy	Worry	
Misery		
Positivity		
Surprise		
Worry		

Fig. 1. The three candidate collections of labels

5.1 Evaluating the annotation schemes

We next assessed the quality of the proposed annotation schemes by measuring the reliability of the data that they produce using an inter–rater agreement measure, namely Krippendorff's Alpha[2] [12].

For each scheme eleven annotators (10 for the medium–grained scheme) annotated a dialogue of 52 utterances. They followed written instructions describing how to perform the annotation with a definition of each label (see section 6). This dialogue was distilled from anonymised extracts of our corpus. While the results of Alpha are not a function of the data being annotated, it was important to encourage the use of as many different labels as possible so that the overall reliability of the scheme can be inferred from the agreement results. For this

[2] Since these schemes allow more than one label to be applied to each utterance, neither Kappa, nor Alpha in its original form will correctly calculate agreement between annotators applying them. Together with Klaus Krippendorff, we have designed and implemented an extension of Alpha which works in these circumstance. The details of this measure are yet to be disseminated.

reason the extracts that made up to dialogue were those that contained a range of emotions.

The results of the Alpha test on each of the scheme were as follows –

Fine Grained	$\alpha = 0.329$
Medium Grained	$\alpha = 0.438$
Coarse Grained	$\alpha = 0.443$

Reliability is inferred from the level of agreement observed in an annotation based on the degree to which we are willing to rely on imperfect data [17]. It has become common practise in computational linguistics to measure reliability against Krippendorff's criterion, with schemes that surpass agreement figures of 0.667 being considered 'reliable' and worthy of use. This is an unfortunate and dangerous mis-interpretation of Krippendorff's work.

The appropriate way in which reliability should be inferred is that the attained level of agreement should dictate the applications to which the resulting annotated data can be applied. Although the agreement shown for our schemes is not strong, this does not mean that the schemes should not be used, only that any conclusions made from resulting data must be based on strong evidence to counter the imperfections in that data. It is likely that the reliability that there schemes display could be increased by making improvements to the coding manual and by training the annotators.

6 A closer look at our final scheme

As described previously, the collection of labels that have been chosen are used to supplement the numerical scheme for annotating emotion in dialogue. The procedure for annotating dialogue, segmented into utterances, using this hybrid scheme is as follows –

1. For each utterance label the overall level and polarity of the expression of emotion using the following guidelines –

   Level
   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"*

2. If one or more of the following labels apply to the expression of emotion in the utterance then add those label to the annotation. If none of the labels apply then leave the utterance unlabelled. (The actual emotions to list depends on the chosen granularity of the scheme, all descriptions are given below)

   Anger - The speaker expresses that a certain situation or person has upset them such that they feel passionately about it.

   Affection - The speaker expresses a liking or love for something.

   Hope - The speaker expresses emotion due to the anticipation of something good happening.

   Happiness - The speaker expresses a generally positive feeling.

   Positivity - The speaker expresses a wish to avoid sadness caused by a situation. This includes the like of bravery, desire and determination.

   Sadness - The speaker expresses that a situation, person, memory etc. is making them unhappy without necessarily being motivated to do anything about it.

   Surprise - The speaker expresses that something unexpected has affected them.

   Worry - The speaker expresses that uncertainty about the future is negatively affecting them.

   Contentment - The speaker expresses satisfaction or gratification.

   Dislike - The speaker expresses disapproval or aversion toward a situation, person or object without necessarily being motivated to do anything about it.

   Frustration - The speaker expresses that their inability to achieve something is upsetting them.

   Fear - The speaker expresses a disposition caused by anticipation of something bad happening.

   Joy - The speaker expresses a positive feeling which they intend to enjoy.

   Misery - The speaker expresses that a situation or person is making them unhappy without necessarily being motivated to do anything about it.

7 Conclusion

In this paper we have described an annotation scheme for labelling expressions of emotion in dialogue. We recognised that reliably identifying emotion is a difficult task, but by combining a categorical annotation scheme with another using dimensional scales we could include a select group of labels which can most easily be identified by human annotators.

We proposed three different schemes, each with a different level of granularity. Although the overall level of agreement for each of these schemes was well below ideal, it was evident that the finer the distinctions between different emotions, the more difficult it was for annotators to agree. Under these circumstances, having more than one scheme allows us to choose a scheme which is most appropriate for each task, based on the level of granularity required and the degree to which we are willing to rely on imperfect data.

By annotating dialogue corpora for emotional expression we hope that it is possible to gain an understanding of the factors that contribute to listeners believing that emotion is being expressed in speech. This understanding may be applicable to a range of tasks, and we identify the generation of emotional speech by artificial communicative agents as a potential beneficiary.

References

1. Piwek, P.: A flexible pragmatics-driven language generator for animated agents. In: Proceedings of EACL03. (2003)
2. Padgham, L., Taylor, G.: A system for modelling agents having emotion and personality. In: PRICAI Workshop on Intelligent Agent Systems. (1996) 59–71
3. Litman, D., Forbes, K.: Recognizing emotions from student speech in tutoring dialogues. In: Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop ASRU. (2003)
4. Narayanan, S.: Towards modeling user behavior in human-machine interaction: Effect of errors and emotions. In: Proceddings of ISLE. (2002)
5. Devillers, L., Vasilescu, I., Lamel, L.: Annotation and detection of emotion in a task-oriented human-human dialog corpus. In: Proceedings of ICSLP 2002. (2002)
6. Cowie, R.: Describing the emotional states expressed in speech. In: SpeechEmotion-2000. (2000)
7. Austin, J.: How to do things with words. Oxford University Press (1962)
8. Carletta, J.: Assessing agreement on classification tasks: The kappa statistic. Computational Linguistics **22** (1996) 249–254
9. Craggs, R., Wood, M.: Agreement and reliability in discourse and dialogue annotation. In: Submitted to the ACL workshop on discourse annotation. (2004)
10. Ortony, A., Turner, T.J.: What's basic about basic emotions? Psychological Review **97** (1990) 315–331
11. Wood, M., Craggs, R.: Rare dialogue acts common in oncology consultations. In: Proceedings of SIGdial3. (2002)
12. Krippendorff, K.: Content Analysis: An Introduction to its Methodology. Sage Publications, Beverly Hills, CA (1980)
13. Craggs, R., Wood, M.M.: A 2 dimensional annotation scheme for emotion in dialogue. In: Proceedings of AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications, Stanford University (2004)
14. Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., Taylor, J.: Emotion recognition in human-computer interaction. IEEE Signal Processing Magazine **18** (2001)
15. Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., Schröder, M.: Feeltrace: An instrument for recording perceived emotion in real time. In: ISCA Workshop on Speech and Emotion, Belfast (2000)

16. Zitzler, E., Laumanns, M., Thiele, L.: SPEA2: Improving the Strength Pareto Evolutionary Algorithm. Technical Report 103, Swiss Federal Institute of Technology ETII, Zurich, Gloriastrasse 35, CH-8092 Zurich, Switzerland (2001)
17. Krippendorff, K.: Reliability in content analysis: Some common misconceptions and recommendations. To appear in Human Communication Research (2004)