

Annotating emotion in dialogue – Issues and Approaches

Richard Craggs

Department of Computer Science
University of Manchester
craggs@cs.man.ac.uk

Abstract

Adopting statistical tests from content analysis allows dialogue analysts to evaluate their work and make comparisons with other research in the field. However, since some aspects of communicative behaviour are more difficult to study than others, such broad comparisons may lead us to avoid interesting areas of research. Emotion in dialogue is one such area.

1 Dialogue analysis and annotation

There are a wealth of applications for which we might want a computer to be able to understand human dialogue. These include natural language interfaces to information sources of services (Gorin et al., 1997) and development of communicative agents for educational, entertainment or commercial purposes (Piwek, 2003). Even applications which are not concerned solely with dialogue such as speech-to-speech translation can benefit from an understanding of how humans behave during a dialogue (Reithinger and Maier, 1995). To facilitate such applications, the role of dialogue analysis is to investigate, formalise, and document communicative behaviour, allowing us to study it empirically, statistically and computationally.

One of the tools available to a dialogue analyst is *Dialogue Annotation*. Dialogue annotation is the process of labelling segments of dialogue with labels that describe some property of that segment. For example, a dialogue may be split up into utterances with each utterance being labelled for the topic to which it refers e.g. –

- A. How was your holiday? [**Holiday**]
B. I had to cancel it because I was ill. [**Holiday**]
A. Illness is a real drag. [**Illness**]

There is a consensus that the richness and complexity of dialogue can best be studied using a multi-layered approach with each layer labelling a separate property of the dialogue's content. Using this approach a single dialogue corpus can be used to investigate a number of different phenomena and also

relationships between different layers can be studied.

2 Emotion in dialogue

As our understanding of dialogue and communicative behaviour increases, and as the applications we develop from that understanding become more sophisticated, it becomes important to investigate more subtle and possibly more interesting properties of dialogue. One such area of investigation is emotion.

There are number of applications which would benefit from an understanding of how emotion influences the way we communicate. For example, those developing communicative agents either to talk to each other or to humans, are interested in making the agents more believable and the speech that they generate more natural (Ball and Breeze, 1998; André et al., 1998; Piwek, 2003). Also, for automated call centres it can be important to identify when a caller is becoming agitated or frustrated with the system. Clearly, an understanding of the relationship between the emotion of a speaker and the speech that they produce would help achieve these aims, and others.

One way in which this understanding of emotion could be gained is by the development of dialogue corpus annotated for the emotion expressed by the speakers. In order to do this an appropriate annotation scheme must be developed. How this aims may be achieved is discussed in section 5.

3 Annotating subtle, rare, or subjective phenomena

To date, dialogue analysis has almost exclusively been concerned with the identification of objective and quantifiable aspects of its content. The reasons for this are best conveyed based on an understanding of how dialogue annotation is planned and executed.

3.1 Developing annotation schemes

To annotate a dialogue corpus, labelling is usually performed by human annotators following an annotation scheme. These schemes describe the labels which can be applied and the circumstances in

which to apply them. Examples of dialogue annotation (coding) schemes include one for coding children's speech in the CHILDES project (MacWhinney, 1998) and many schemes for coding dialogue acts (Core and Allen, 1997; Di Eugenio et al., 1998; Jurafsky et al., 1997).

The design of these schemes is frequently motivated by some theory about dialogue or language in general. For example, Searle's theory of speech acts (Searle, 1969) which describes how actions are performed by speech, has led to the idea of dialogue acts which are used to describe the function of an utterance. In turn, the choice of which labels to include in a scheme is often motivated by a particular domain or problem. For instance, the above mentioned CHILDES scheme uses labels which can be used to describe the utterances used by children. A scheme used to identify positive and negative behaviour of a health professional attempting to elicit the concerns of a cancer patient (Heaven and Maguire, 1997) uses labels which highlight such behaviour.

The next important step in developing an annotation scheme is to prove that it is appropriate for the task for which it is developed. For a scheme to be valid it is important that it produces reliable results. Klaus Krippendorff proposed three measures of reliability — *stability*, *reproducibility* and *accuracy*. For our purposes, stability means that the same scheme applied more than once to the same data at different points in time will give the same results, reproducibility means that more than one annotator applying the scheme should yield similar results and accuracy means that these results should match some 'correct' standard. Usually, validity for dialogue annotation schemes is assessed using *inter-rater reliability*, which is a reproducibility test. We shall return to the troublesome subject of inter-rater reliability later, but suffice to say that for restricted domains and clearly defined labels, satisfactory levels of reliability can often be obtained.

3.2 Issues regarding subtler phenomena

Alongside these types of annotation, a multi-layered annotation may be augmented by layers describing phenomena which are more subtle or subjective, such as emotion or speaker intention. Consider the above process being applied in order to develop schemes for layers such as these. Since emotion and intention are less closely associated with linguistics than, say, dialogue acts, choosing labels that apply to individual portions of a dialogue is consequently a more difficult task. Furthermore describing these types of phenomena is harder than describing, for example, the topic of conversation. For these reasons, creating a suitable list of labels from which to construct an annotation scheme for subtle, subjective or even non-linguistic phenomena can pose a serious obstacle on the path to annotating it in dialogue.

If after some persistence a list of labels is constructed and an annotation manual produced the real difficulty of proving the validity of the scheme is upon us. Results for tests of stability and reliability are heavily influenced by the ease with which an annotator is able to identify the circumstances under which each label is applicable. For example when labelling utterances according to whether they were successfully completed as one might do when applying the *DAMSL* scheme (Core and Allen, 1997) identifying which utterances were completed and which were not is a fairly trivial task e.g. —

- A. Have you seen the new secretary? [**Complete**]
- B. Yeah, what's her name? [**Complete**]
- A. It's errr.... [**Abandoned**]
- B. Oh never mind. [**Complete**]

The same cannot be said for labelling the intention of a speaker, since this facet of communicative behaviour is much less evident from the content of someone's speech. Harder still is establishing whether any given annotation of a dialogue is 'correct' as one would be required to do to pass the third of Krippendorff's tests for reliability — accuracy.

Increasingly there is a reliance on inter-rater reliability tests to measure the quality of dialogue annotation research. Of course, establishing the validity of any scheme and striving to develop schemes which are as reliable as possible is an important part of dialogue analysis. However insisting on a minimum 'score' from a certain reliability test before a scheme may be accepted by the field would make the honourable aim of researching difficult (and therefore interesting) aspects of dialogue very difficult.

4 Inter-rater reliability

In 1996 Jean Carletta recognised that in order to improve the consistency of computational linguistics research and facilitate comparisons between different researchers' results a more rigorous, common approach to evaluation was required (Carletta, 1996). This certainly applies to dialogue annotation schemes.

4.1 Selecting an appropriate reliability test

In that paper it is suggested that the Kappa statistic (Seigal, 1988) is a suitable measure of agreement with which to validate annotation. The Kappa statistic measures the level of agreement between any number of annotators assigning nominal labels to objects, awarding scores of between 0 and 1 where 1 denotes perfect agreement and 0 is equivalent to the level of agreement that could be expected from random behaviour by the annotators. However, the suitability of Kappa for this purpose has been questioned because of issues regarding the accuracy with

which it measures the level of agreement expected by chance (Krippendorff, 1980).

An alternative agreement measure that was also mentioned in (Carletta, 1996) is Krippendorff's Alpha statistic (Krippendorff, 1980). Alpha has similar properties to Kappa but considers the frequency with which labels are used, when calculating the level of agreement that could be expected from chance coding, which was the source of Kappa's inaccuracy.

Considering this, in most cases it would appear sensible to use Alpha in place of Kappa.

The researcher interested in annotating the type of subtle phenomena discussed in section 3 may be called upon to use more sophisticated labelling approaches than simple categories. For instance, we could use a numerical scale to label the level of confidence that a speaker has in an assertion, which could have applications in natural language interfaces to knowledge bases systems. It is necessary to be able to measure the level of agreement between raters using these types of labels. Both Kappa and Alpha accommodate this – a generalised form of Kappa, *weighted Kappa* allows partial agreement between pairs of labels and disagreement in Alpha is calculated according to a difference function which acts in a similar way. Although this allows researchers to develop tests for agreement which are suited to the form of their annotation schemes, the accuracy and comparability of the results depends largely on the metric that they choose and how it is implemented.

4.2 Interpreting agreement results

Given a result of between 0 and 1 from Alpha, Kappa or any of the similar tests, how can that be used to evaluate the validity of a scheme? The assumed wisdom is that values of 0.8 represent high levels of agreement and values between 0.67 to 0.8 are acceptable if not desirable. In (Landis and Koch, 1977) it was recognised that these values may not be suitable for all domains. A table is drawn up showing the relationship between Kappa scores and strength of agreement which they believe is more applicable in medicine. This table suggests that values between 0.41 and 0.6 provide moderate levels of agreement and only values less than 0.2 should be considered poor.

In (Carletta, 1996) it was suggested that since the phenomena under observation in computational linguistics are more difficult to study than that for which the values '0.67 to 0.8' were proffered, then obtaining satisfactory levels of agreement may not be possible. It would be unfortunate if researchers were wary of studying difficult aspects of dialogue in case the results of their experiments did not reach those exacting standards. Identifying subjective phenomena such as emotion is implicitly difficult, human beings frequently mis-read other peoples' feeling, but

that does not stop it from being an important thing for us to try to achieve. I believe that the field of dialogue analysis should be more open-minded toward research in these areas and not preclude work for which these high levels of reliability are not obtained. This is not to say that we should not all endeavour to attain the highest level of reliability possible, only that we should not abandon or ignore work for which these high levels of agreement are beyond our reach.

5 Developing annotation schemes for emotion in dialogue

When developing an annotation scheme for any purpose the first step must be to find some mechanism for describing your chosen phenomena. Since emotion is such a complicated matter, a number of ways that have been developed to describe it. Here we shall look at two of the most useful for our purposes.

5.1 Categorical labels

The most obvious way describing many types of phenomena is to use descriptive labels. This is also true of emotion.

Since the word 'emotion' can describe a very broad range of concepts, it is important that researchers studying it describe exactly what they mean by emotion in their own context. To this end, Roddy Cowie provides a useful distinction between cause-type descriptions of emotion which are used to describe the properties of speaker's internal state, and effect-type descriptions which refer the effect that emotional characteristics of speech have on the listener (Cowie, 2000). In cases where we are interested in emotions that people are able to detect from speech, which is true when developing annotation schemes, the second of these is more appropriate.

In the same article it is also noted that while descriptive, categorical labels such as 'Anger' and 'Happiness' are easy to understand and capable of describing complex emotional states, describing their exact meaning and compiling a suitable list of labels which covers the vast range of human emotion while remaining distinct from each other is a difficult task. Furthermore, if we are to restrict ourselves to emotions that we can identify in speech, the task becomes even more arduous.

A test in which conversational dialogues were annotated using labels for emotions identified in psychological research, namely –

Fear, Anger, Happiness, Sadness, Surprise, Disgust, Love, Acceptance, Hate, Contempt, Aversion, Dejection, Desire, Courage, Wonder, Interest, Shame and Guilt

gave us a level of agreement between four annotators of $Alpha = 0.167$, which is unacceptable low.

More success is likely to be achieved by discovering which emotions are most clearly evident from the content of speech. This may mean that we are limiting the range of emotions which we are capable of studying. However, if it allows us to develop an annotation scheme which is both informative and valid then it is an acceptable concession to make. Also, there are cases in which these types of emotions are exactly those in which we are interested. For example, if the ultimate goal of some research is to understand the relationship between emotion and speech, so that conversational agents are able to exhibit emotion in their output, we would only need concern ourselves with emotions that the other collocutor could identify.

Work to identify which emotions these are, and the level of granularity it is possible to distinguish (e.g. can we distinguish joy and happiness) is one of our current research goals.

5.2 Numerical representations

Another useful descriptive mechanism which can be used to annotate emotion is to describe emotional states using values from numerical scales. For example, we could ascribe a value to an emotional state according to the degree which it could be considered positive or negative. We might regard *fear* as strongly negative and *contentment* as mildly positive.

This approach is attractive since all possible emotional states can be described, to some degree, in these terms. (Cowie et al., 2001) describes how a two dimensional scale, ‘Activation–Evaluation space’ can be used to describe emotion in speech. In (Craggs and Wood, 2004) we have shown that a similar approach yields promising results for annotating emotion in transcribed dialogue.

From a pragmatic point of view there are a couple of drawbacks to this approach. Firstly, the degree to which different emotions can be distinguished depends largely on which dimensional scales are used. For example, as Cowie points out, in Activation–Evaluation space, fear and anger are indistinguishable. This is because as complex emotions are reduced to two simple values, inevitably information is lost. Secondly, it is not obvious how an understanding of the relationship between emotions described using numerical values and language production could be used in the applications described in section 2. Perhaps a hybrid scheme which allows the broad coverage of numerical representation while allowing annotators to use category labels when distinguishable emotions are observed would be beneficial.

5.3 Other differences between emotion and less subtle phenomena

Besides those previously mentioned, there are many other differences between emotion and the simpler

phenomena which are more frequently annotated in dialogue, two of which we’ll now consider.

Our first difference relates to the units to which labels are applied. Since most annotation schemes are founded on some theory of language, the units to which those schemes apply is implicit or obvious. The clearest example is part of speech tagging where words form natural units which can be labelled. This is not the case for emotional episodes which exist over an indistinct period of time, fading in and out and subtly changing throughout a dialogue. However, unless we restrict ourselves to labelling some recognised unit, such as utterances, it will become difficult to make comparisons or investigate relationships with other layers.

Another difference is that whereas many other annotation layers require exactly one label per unit this is not the case for emotion. There may be cases when no emotion is detectable from the content of speech or that it is impossible to identify what the emotion is. There may be cases when more than one label is applicable per unit. For instance, it is conceivable that a person may convey fear and sadness in a single utterance. This raises the question of how we would test levels of agreement for cases of absence of, or multiple labels per unit. For absence of labels we have the choice of using Krippendorff’s notion of ‘missing’ data when applying the Alpha statistic or allowing an ‘Unlabelled’ category which is likely to be abundant, having a large impact on the level of agreement obtained.

Assessing the level of agreement in cases where more than one label is allowed to be applied to a single unit is not a straightforward task. Neither Alpha or Kappa accommodate this in their original forms. The following example of two annotators labelling three utterances with perfect agreement shows how the results given by the original forms of Kappa and Alpha deviate from the expected value of one –

Annotator	Utt-1	Utt-2	Utt-3
A1:	Anger	Fear, Sadness	Misery
A2:	Anger	Fear, Sadness	Misery

$Alpha = 0.57$

$Kappa = 1.6$

Development of appropriate tests for annotations which allow multiple labels per unit is a necessary step in developing annotation schemes for emotion using categorical labels.

6 Summary

In this paper we have investigated ways in which annotating subtle, rare or subjective phenomena is more difficult than those currently used in dialogue analysis. It was suggested that although adopting rigorous statistical tests developed in other fields will

help us assess the quality of our research and allow comparisons to be made, applying their standards to our work may not be constructive. Restricting ourselves to studying properties of dialogue for which we can obtain levels of agreement comparable to content analysis or medicine would lead us to ignore some of the more interesting aspects of communication.

Taking emotion as particular example we discussed how we might need to investigate a number of ways of describing a particular facet of communicative behaviour in order to develop annotation schemes. We also showed that standard reliability tests such as Kappa and Alpha may not be suitable for these schemes without modification.

References

- E André, T Rist, S van Mulken, and S Klesen, M nad Baldes. 1998. Emotion and personality in a conversational character. In *Proceedings of the Workshop on Embodied Conversational Characters*, pages 83–86, Lake Tahoe, CA.
- G Ball and J Breeze. 1998. Emotion and personality in a conversational character. In *Proceedings of the Workshop on Embodied Conversational Characters*, pages 83–86, Lake Tahoe, CA.
- Jean Carletta. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254.
- Mark G Core and James F Allen. 1997. Coding dialogs with the damsl annotation scheme. In *AAAI Fall Symposium on Communicative Action in Humans and Machines*.
- R Cowie, E Douglas-Cowie, N Tsapatsoulis, G Votsis, S Kollias, W Fellenz, and J Taylor. 2001. Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine*, 18.
- Roddy Cowie. 2000. Describing the emotional states expressed in speech. In *In SpeechEmotion-2000*.
- Richard Craggs and Mary McGee Wood. 2004. A 2 dimensional annotation scheme for emotion in dialogue. In *To appear in, AAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications*, Stanford University, March.
- B Di Eugenio, P.W Jordan, and L Pylkkanen, 1998. *The COCONUT project: Dialogue Annotation Manual*. ISP Technical Report 98-1.
- A Gorin, G Riccardi, and Wright J. 1997. How may I help you? *Speech Communication*, 23(1/2):113–127.
- C Heaven and P Maguire. 1997. Disclosure of concerns by hospice patients and their identification by nurses. *Palliative medicine*, 11(283-290).
- D Jurafsky, L Shriberg, and D Biasca, 1997. *Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual*. University of Colorado, draft 13 edition.
- Klaus Krippendorff. 1980. *Content Analysis: An Introduction to its Methodology*. Sage Publications, Beverly Hills, CA.
- Robin J. Landis and G. G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33:159–174.
- B MacWhinney, 1998. *The CHILDES Project: Tools for Analysing Talk*. Carnegie Mellon University.
- P Piwek. 2003. A flexible pragmatics-driven language generator for animated agents. In *Proceedings of EACL03*.
- Norbert Reithinger and Elisabeth Maier. 1995. Utilizing statistical dialogue act processing in VERB-MOBIL. In *Meeting of the Association for Computational Linguistics*, pages 116–121.
- J Searle. 1969. *Speech Acts: An essay in the philosophy of language*. Cambridge Press.
- S Seigal. 1988. *Nonparametric statistics: Second edition*. McGraw-Hill.