What a neural net needs to know about emotion words

R. COWIE, E. DOUGLAS-COWIE, B. APOLLONI, J. TAYLOR, A. ROMANO and W FELLENZ

School of Psychology, Queen's University, Belfast, UK
Department of Mathematics, King's College, London, UK

Abstract: We describe an empirical approach to identifying the kind of task that an emotion recognition system could usefully address. Three levels of information are elicited – a basic emotion vocabulary, a basic representation in 'evaluation-activation space' of the meaning of each word, and a richer 'schema' representation. The results confirm that an approach of this kind is feasible.

Key-Words: - emotion recognition, semantics, neural network *CSCC'99 Proceedings*, Pages:5311-5316

1 Introduction

This paper describes work being done as part of the PHYSTA project. PHYSTA is concerned with an increasingly high profile problem in IT, which is to develop artificial systems that are capable of detecting signs emotion in a human user, and of reacting accordingly. PHYSTA will use hybrid technology, involving neural net and symbolic techniques.

One of the challenges facing the project is to identify the kind of knowledge about emotion that is relevant, and appropriate techniques for expressing it. The obvious approach is to select a few well-known emotion terms and to construct classification systems which try to assign these terms correctly to records of human performance. However, that is unsatisfactory for a multitude of reasons.

1. The selection of terms defines the problem to be solved, and that is not something that should be left to the experimenter's casual intuitions.
2. The selection criteria should take account of the states that the system is likely to enounter - not just of examples which are salient, and perhaps theoretically interesting, but rare.
3. Emotion terms are discrete, but emotional states form a continuum. Discrete terms need to be embedded in continuous representations that at least permit interpolation, and that ideally allow for the kind of shaded judgement that people can achieve (e.g. by qualifying emotion terms).
4. Real inputs often give partial information about an emotion - e.g. showing arousal without making it clear whether the person is happy or angry. An effective system should be able to use that kind of information rather than being forced to make a classification that goes far beyond the evidence.
5. Classifying emotion states is of no use in and of itself. Use depends on associating the terms with a semantics which indicates what the user is likely to do, and what interventions might be appropriate.
6. Intuitively it seems quite likely that these issues are linked - e.g. dealing with either incomplete information, or intermediate / compound states, involves assessing the underlying dimensions of a person's state directly rather than via the categories for which there are convenient emotion terms.
7. Last but not least, simple classification is not particularly interesting intellectually. Using a practical IT problem to develop deeper ideas about emotion, and emotion words, is.

Considerations like these led us to develop techniques for eliciting from human beings the kind of knowledge that neural net and hybrid systems need if they are to detect and respond to human emotions in a way that is intellectually satifying and practically useful. The core aims are first to provide a rationale for selecting the emotion-related terms that the system should be able to use, and second to define the kind of representation that we want the system to produce in response either to an emotion-related word, or to a sample of emotionally coloured behavior.

The ideas that we have used are rooted in the psychological and biological literature on emotion [1]. Our contribution has been to translate these ideas into a form that lends itself to IT applications, and particularly to training systems with a neural net component. The result has three main elements

1.1 A Basic English Emotion Vocabulary

Research on emotion is dogged by ad hoc selections of emotions to work with. There is no agreed benchmark, in the form of a range of emotion terms that a competent system should be able to apply. Without that, it is impossible to assess the performance of emotion detection systems in a meaningful way. Investigators describe innumerable tests or variables that are claimed to be relevant to various specific distinctions with no reference to the importance of the distinction, or the way the test would function when other possibilities had to be considered.

One approach to this problem has been explored by theorists in biology and psychology since Descartes - attempting to identify a set of 'primary' emotions, the pure elements underlying the various compounds that tend to occur in everyday life. Despite several centuries of predominance, that approach has not produced an agreed set of primaries [1].

We have developed a second approach, which complements the traditional one and is more immediately relevant to the IT issue. It involves trying to identify the main compounds (if such they are) that actually occur in everyday life. We have done that by trying to identify a relatively small vocabulary of words that people regard as sufficient to describe most emotional states and events that are likely to occur in everyday life. We call it a Basic English Emotion Vocabulary - BEEV for short.

As regards theory, our approach offers a way of defining the everyday 'compounds' (if that is what they are) that should be deriveable from a proposed set of elements - and without that, it is difficult to see how variants of the traditional approach can be evaluated. As regards practice, the natural goal for IT is to develop a system that can use what humans agree is a basic emotion vocabulary.

1.2 A 2-D emotion space

Many authors agree that emotions can be organised roughly into a two-dimensional space whose axes are evaluation (i.e. how positive or negative the emotion is) and activation (i.e. the level of energy a person experiencing the emotion is likely to display) [2]. That provides a useful basic continuum in which to embed emotion words. We have developed techniques that allow informants to assign co-ordinates in evaluation-activation space to both words and expressions of emotion (through the face, voice, or music).

These techniques serve several functions. They define a non-categorical targets that networks can naturally be trained to emulate. They also provide a very basic kind of semantics for emotion words - a machine that could reliably assess activation and evaluation levels from audio-visual images would have at least some basis for making appropriate responses. The techniques also allow people to record aspects of their response to emotion-related displays that are difficult to capture in categorical terms. In particular, the 2-D space can be used to record how emotion-related judgements change continuously over time - capturing an aspect of human judgement that it would certainly be useful for a machine to emulate, but that is difficult to record satisfactorily using categorical descriptions.

1.3 An emotion schema

Evaluation-activation space captures a good proportion of distinctions between emotion-related terms, but there are many that it fails to capture. For example, fear and anger tend to be placed nearby in the space. The important difference between them involves a different kind of dimension altogether. Higher order spaces are required to capture that kind of distinction. They reflect the fact that, as many authors have pointed out, emotion is closely linked to the way the organism is disposed to act (e.g. fear involves a disposition to flee, whereas in anger the disposition is to attack), and also to the way the organism appraises the situation.

Drawing on a range of psychological theories [2,3,4], we have constructed questions designed to capture a wider range of distinctions in a systematic way. These express a simple but reasonably powerful semantics for emotion terms, in a quantitative format that lends itself to implementation in neural nets. The dimensions divide into three blocks.

Questions in the first block deal with the broad kind of action that someone in a given emotion-related state would be likely to take - engage, withdraw, seek information. Questions in the second ask whether the emotion has an object - i.e. whether saying that an individual is in a given emotion-related state implies that they are reacting to or thinking about a particular person or situation; and if so, whether the relevant person or situation is present at the time, or located in the past, or the future, or in the individual's mind. Questions in the third block ask about the broad characteristics of any situation - present, past, future, or mental - that is directly relevant to the emotion. They deal first with the individual's own perceived standing in the situation - powerful or powerless, well-informed or lacking information, morally sound or not - and with relevant characteristics of the situation or person – human or not, powerful or not, appealing or not.

An important feature of the system is that it does not require a rating on every dimension. It is always allows a question to be answered by saying that the word being considered does not provide information about that issue. The intended effect is to identify for each word a relatively compact range of features on which it does carry information, and at the same time to acknowledge that other features, which are important for other emotion words, are not particularly relevant.

The structure that we have outlined defines a form in which a partial but useable understanding of emotions can be couched. The content has been derived in the traditional psychological way, by experiment. Subjects have been asked to give their ratings on the various dimensions. That approach ensures that the information we provide is not simply an expression of our own personal theories on the subject. It also ensures that a net which is trained uing our data will use emotion terms in a way that is broadly consistent with people who are hopefully a reasonably representative sample of potential users. The possibility is clearly open to use the same elicitation techniques with other groups or in other languages if so that the system can be adapted for different users.

2 Method

We have developed a program called BEEVer which asks subjects to carry out the various ratings associated with the scheme outlined above. The study has been carried out in two phases. Phase 1 dealt with only the first two elements, identifying a Basic English Emotion Vocabulary (BEEV) and providing ratings in the evaluation-activation space. It provided a basis for refining both those elements and the choice of words from which the BEEV was to be selected. Phase 2 presented modified versions of those elements and the emotion schema.

BEEVer uses an initial vocabulary of emotion-related words from which subjects select 16 that they regard as consituting an acceptable basic emotion vocabulary. For Phase 1, the initial vocabulary consisted of words that feature in published lists that are meant to summarise the main types of emotion, plus additional terms needed to describe emotions that occur regularly in material that we have recorded for use as a database in the PHYSTA project. That produced an initial vocabulary of 45 terms. The initial vocabulary was revised for phase 2 by excluding terms that at most one phase 1 subject included in his or her selection of 16, and adding terms that subjects suggested should have been present. That produced an initial vocabulary of 40 terms in stage 2.

For the sake of readability, specific descriptions of the tasks are presented along with the results they produced.

This kind of exercise depends on ensuring that subjects understand what they are being asked to do. Much of the effort in phase 1, and smaller pilot studies for the schema, was devoted to that issue. The resulting procedure in phase 2 incorporated oral instructions, particularly on the use of evaluation-activation space; a preliminary program using four practice terms and incorporating written instructions on the use of the schema; and a proviso that subjects would be dropped if the experimenter was not satisfied that they had understood the task.

Eighteen subjects took part in each phase. Two phase 2 subjects were dropped because it was not certain that they had understood the task.

3 Results

3.1 A Basic English Emotion Vocabulary

The elicitation of the basic vocabulary had two stages. The first two questions about each word asked subjects to rate how common or rare the state was, and how psychologically simple or complex. Those ratings were used to create a preliminary division into a set of 16 candidates for a basic vocabulary and a residue of less basic words. Candidates were entered on the basis of the lower of the two ratings, i.e. they were likely to be included if they were either common or psychologically simple.

The second stage occurred at the end of the whole exercise. At that stage, after subjects had rated all the words in the initial vocabulary, they were presented with a screen showing the 16 current candidates for a basic vocabulary on the left, and the residue on the right. They were then allowed to switch words from one side to the other until the words on the left formed the best basic emotion vocabulary that they could construct.

Figure 1 summarises the results. The words on the left hand axis are the initial vocabulary used in phase 2. Their order is determined by the frequency with which they were chosen as basic emotion terms in phase 1. That arrangement shows that there is a good deal of stability in the outcome despite quite large procedural differences between the two phases. The most frequently selected words in phase 1, from 'disappointed' downwards, are also the 16 most frequently selected in the whole data set - with the exception of 'satisfied', which clearly ought to be incorporated into a balanced vocabulary.

It is not the aim of this paper to set out a definitive Basic English Emotion Vocabulary - not least because the data do not show a sharp cut-off. However, the results provide an empirical basis for assessing how defensible alternative lists are. Many lists in the literature, based on a priori judgements, quite clearly fare poorly by empirical criteria: they omit terms that empirically appear important, and include others that very few subjects regard as particularly useful. From an IT point of view, that is a non-trivial point. If research allows itself to be guided by that kind of list, and the intuitions underlying it, then it risks producing systems that are expert in emotion judgements that are almost never needed, and incapable of judgements that are.

For practical purposes, it is useful to identify definite groups. It is worth distinguishing two - an inner group of seven terms, chosen by a clear majority of subjects, consisting of happy, angry, sad, interested, pleased, relaxed, and worried; and an outer group of ten, chosen by half of the subjects or slightly more, consisting of affectionate, afraid, content, excited, bored, confident, amused, loving, disappointed, and satisfied. Using the words in the inner group in particular is a reasonable target for IT.

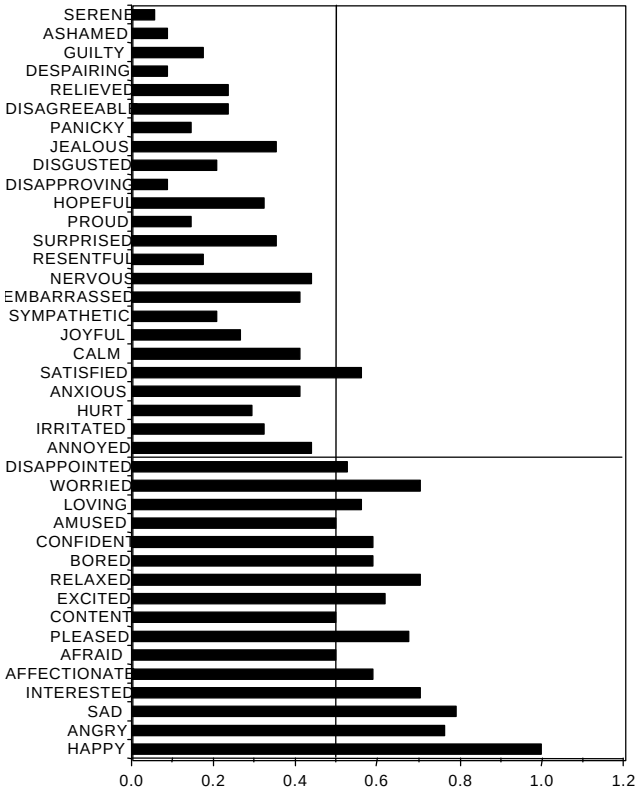

Fig. 1: Probability that each of 40 words will be included in a basic emotion vocabulary chosen by subjects (based on a sample of 36 people).

3.2 Evaluation-activation space

Ratings were made using a representation associated with Plutchik, Russell and others, in which possible emotions are arranged in a circle. Strong emotions lie at the periphery: an emotion-free state of alert neutrality lies at the centre. The vertical axis of the circle represents activation level, the horizontal axis evaluation - positive emotions are on the right, negative on the left. Key emotion are arranged round the periphery to provide landmarks and help subjects to orient themselves within the space. Subjects rated a word by clicking with a mouse at an appropriate point in the circle. They were allowed to revise their initial choice if they wanted to.

Phase 1 data suggested that subjects had not understood the significance of distance from the centre, and had chosen relatively peripheral points irrespective of emotion strength. Hence for phase 2, adjustments were made to the landmarks round the periphery (ensuring that they all referred to extreme emotions, in line with their distance from the centre) and to the instructions (making explicit the meaning of distance from the centre). The resulting axes and landmarks are shown in figure 2.

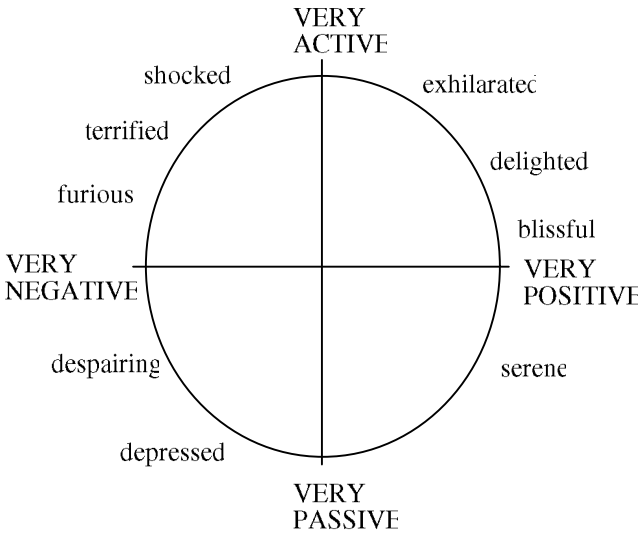

Fig. 2: Axes and landmark items of evaluation / activation space as presented to subjects.

Figure 3 shows mean positions for the frequently selected items - ratings for the inner group of seven on the left, and for the outer group of ten on the right. The plots illustrate both the strength and the weakness of the evaluation-activation system.

The strength of the representation is that it captures a good deal of the information contained in basic emotion terms by way of a medium that is simple,

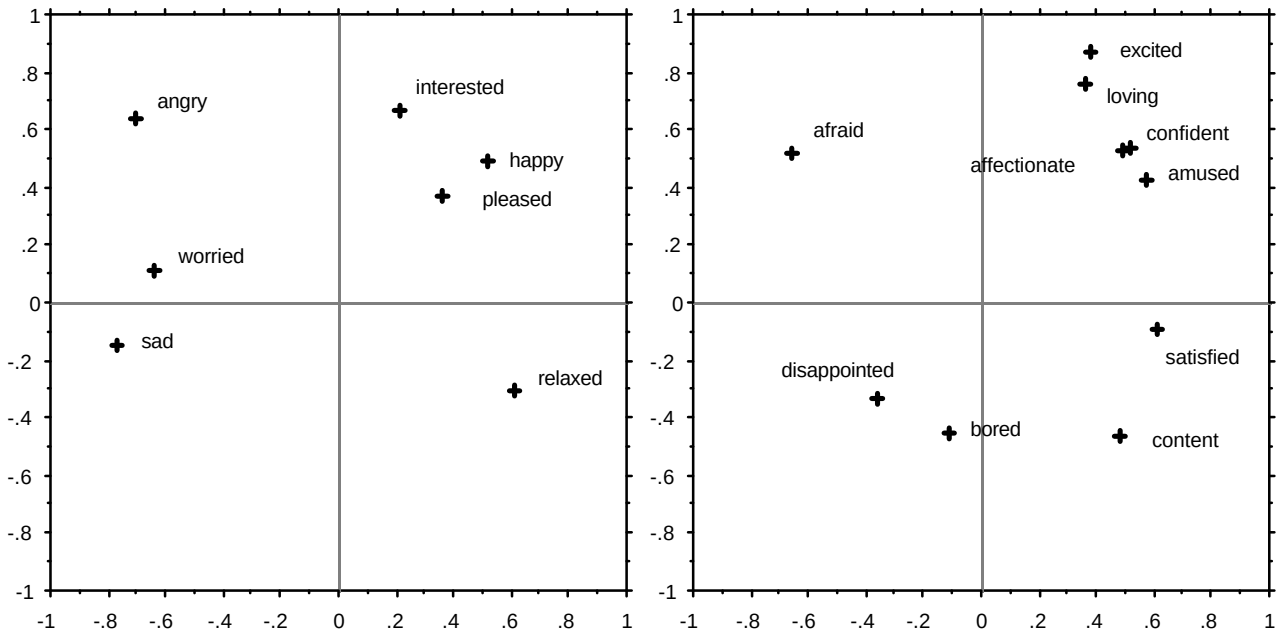

Fig.3 Mean ratings in evaluation/activation space. The horizontal axis is evaluation, the vertical axis activation.

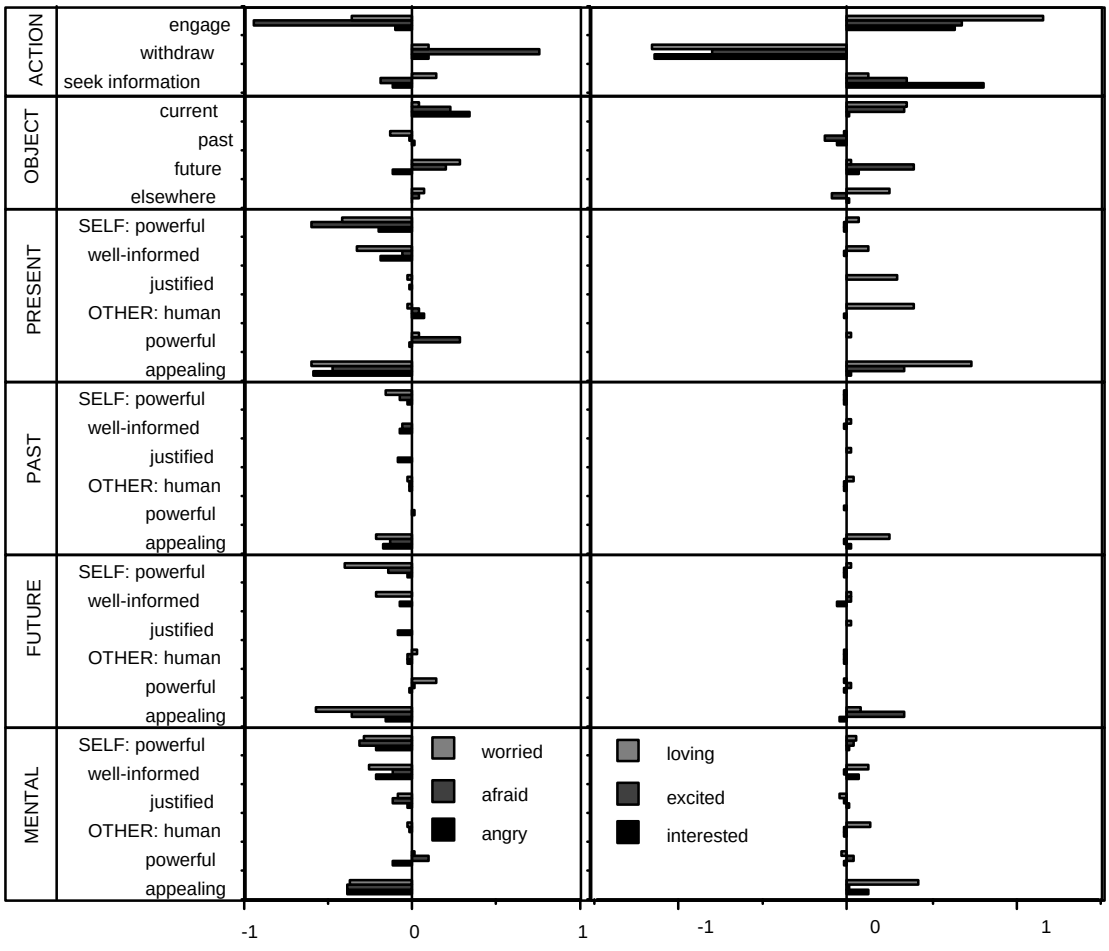

Fig. 4: Samples of profiles obtained from the schema element of BEEVer

meaningful, and convenient for both training and response. Also, people find it easy to place samples of emotional behaviour in this framework, which makes it convenient for generating training samples.

The weakness is that some discriminations that matter are not well drawn in this space. One example has already been given - fear and anger more or less coincide in this space. Another is apparent on the positive side of the graph - happy, pleased, confident, amused, and affectionate are all essentially together. If they are to be discriminated, additional dimensions are needed. It is, of course, useful to have a this kind of indication that different kinds of discrimination may be needed to manage anything beyond a minimal emotion vocabulary.

3.3 Schema representations

The schema element is the medium that we have used to provide more sophisticated discrimination. Figure 4 illustrates the kind of information that it provides, taking as an illustration two clusters of items that are not effectively separable in evaluation/activation space.

The left hand panel of the figure deals with terms that lie relatively close together in the upper left hand quadrant of evaluation-activation space. The right hand panel deals with terms that lie in the upper right hand quadrant of evaluation-activation space, and towards the upper end of it. There are gross differences between the profiles which reflect that broad contrast, as one might expect. The top panel shows that likely actions are biased towards withdrawal in the first group, and engagement in the second. The lower panels, describing relevant situations, show predominantly negative appraisals in the first group, and predominantly positive appraisals in the second. More interesting, though, are the differences within the two clusters that evaluation-activation space is not well suited to capture.

Anger can be considered as the simplest of the terms on the left. It does not indicate any particular course of action very strongly. It suggests reaction to a situation which is present, and whose main feature is that it contains something that is definitely not appealing. Fear is more distinctive. It is associated with a strong disinclination to engage, and an almost equally strong inclination to withdraw. It also carries a distinctive appraisal of the present situation - there is a powerful other involved. Worry carries an inclination to seek information, and an orientation towards future events rather than current ones.

Of the terms on the right, excited and interested are relatively simple. The only distinctive feature of 'excited' is an orientation towards a future situation with some - uspecified - appealing characteristic. Interest implies a disposition to seek information. Loving, in contrast, implies an object, present or in the mind, which is human and appealing.

The point of these summaries is not that they are in any way surprising. It is simply that the schema appears to capture the obvious implications of the terms in a format that is straightforward, intuitive, and empirical. A system which registered implications like these could reasonably be said to have a rough grasp of what the terms meant. That is why the schema provides a useful kind of input to a system that is to learn how to use emotion terms.

4 Conclusion

We have described an empirical approach to identifying the kind of task that an emotion recognition system could usefully address. The results confirm that an approach of this kind is feasible. We are currently engaged in extending it, in three main ways. First, we will obtain responses from a much larger sample of English speaking informants. Second, we will extend the approach to other European languages. Third, we are developing methods of allowing subjects to rate examples of emotional behaviour in terms of these dimensions.

The approach lends itself to a particular style of implementation. It suggests that the domain of emotion understanding can be represented as a network involving nodes of many kinds. Emotion terms are one kind of node, but only one. Also involved are highly compressed representations of situations, built round evaluations of the main agents and forces; and of the actions that these situations are likely to evoke. Various kinds of evidence may be relevant to activating the nodes in the network, and activation in various combinations of nodes may serve to activate high order representations, such as the nodes associated with emotion terms.

References:

[1].*Review of existing techniques for human emotion understanding* Report for TMR Physta project Research contract FMRX – CT97 – 0098 (DG12-BDCN)

[2]. Plutchik, R. *The psychology and biology of emotion*. Harper Collins, New York, p.58, 1994

[3] Lazarus, R.S. *Emotion and adaptation*. Oxford University Press, New York, 1991

[4] Oatley, K & Jenkins, J. *Understanding Emotions*. Blackwell, Oxford: 1996