A Socratic Epistemology for Verbal Emotional Intelligence

Abe Kazemzadeh, James Gibson, Panayiotis Georgiou, Sungbok Lee,
Shrikanth Narayanan,

Abstract

In this paper we claim that question asking provides a natural framework for acquiring representing knowledge. Question asking has a well-founded theoretical basis that goes back to ancient times. We develop a framework for empirically observing human question-asking behavior as a means for dealing with subjective information about emotions. Dealing with subjective linguistic knowledge in this way gives insight about how to represent emotions for use in natural langauge technology. First, we build a corpus of people playing emotion twenty questions (EMO20Q), a of twenty questions limited to emotions. Then, we create a question-asking dialog agent that can deal with this subjective information about emotions and test it experimentally. Our results show that the agent is percieved as human-like and can identify a large set of non-prototypical emotions learned from observed human-human data. Even when the agent fails to correctly identify the subject's emotion, it is still percieved as human-like, despite performing worse than humans on the EMO20Q task. We feel that the agent's behavior of searching for knowledge, and not the agent's knowledge itself, is responsible for this subjective performance rating. Exploring this issue of curiosity, we discuss a model of verbal exploration for the agent.

1. Introduction

[Gorgias:] Just as different drugs draw forth different humors from the body – some putting a stop to disease, others to life – so too with words: some cause pain, others joy, some strike fear, some stir the audience to boldness, some benumb and bewitch the soul with evil persuasion" (Gorgias, *Encomium of Helen*, c.415 B.C.).

Socrates: You, Gorgias, like myself, have had great experience of disputations, and you must have observed, I think, that they do not always terminate in mutual edification, or in the definition by either party of the subjects which they are discussing;...Now if you are one of my sort, I should like to cross-examine you, but if not I will let you alone. And what is my sort? you will ask. I am one of those who are very willing to be refuted if I say anything which is not true, and very willing to refute any one else who says what is not true, and quite as

Preprint submitted to Elsevier *May 10, 2012*

ready to be refuted as to refute. (Plato, *Gorgias*, Transl. Benjamin Jowett, 380 B.C.)

In the first quote above, Gorgias, a Sophist rhetorician, describes the effects of words on a person's emotions. Gorgias describes emotions both by using his rhetorical skills and by making reference to the theory of physiological humors. In the second quote, Socrates (as quoted by Plato) cross-examines Gorgias to determine Gorgias' beliefs. In this paper, we describe an experimental methodology that uses a natural language question-asking approach inspired by the Socratic method to acquire knowledge and shared beliefs regarding emotions. This methodology allows us to study the meaning of a specific subset of emotional language, which we call *natural language description of emotion.* Natural language descriptions of emotion are utterances that refer to emotions directly. They can be seen as a subset of the larger phenomenon of emotional language, which also includes emotion or sentiment expressed towards some object, vocal modulation due to emotion, and persuasion and pragmatics. We maintain that studying question-asking dialogs about emotion shows the capability of humans to learn and use a large, rich, subjective vocabulary of emotion words and provides a theory of emotions based on communication that better captures the shared meaning of natural language. The framework we present also differs from other theories of emotion in that it aims to study how people describe emotions, rather than how emotions *should be* described. As such, it can be seen as a descriptive, rather than prescriptive, theory, and hence has commonalities with sociological studies of emotions [1]. The conception of emotions has changed since the time of the ancients, who believed that emotions were generated from bodily "humors", which in turn were derived from alchemical elements. Also, the conception of emotions may vary from culture to culture and person to person. Therefore, a descriptive theory should allow for change over time and a particular instantiation of an emotional theory should be relative to a particular agent or set of agents who hold similar beliefs and communicate using the same language.

The key goals of this paper are to use question-asking to observe natural language descriptions of emotion in a natural context and to computationally model the social processes that support the referential link between language and emotions. In the next section, we will discuss the motivations and theory behind our work. Next, we discuss the data we collected of people playing emotion twenty questions (EMO20Q). The following section describes the computational model and algorithm we used to create an EMO20Q questioner agent. Finally we examine a computational model for curiosity for the agent.

2. Background

2.1. Natural Language Descriptions of Emotions

Just as memory addresses, variables, and URLs refer to electronic resources for computers, so to do words and descriptions identify objects, both physical and conceptual, for humans. When processing natural language by computer, it can help to draw upon these similarities. This is especially the case in the case of affective computing, when the objects we wish to reference are quite abstract and subjective.

In this paper we make a distinction between the emotion *expressed by* the speaker and the emotion *referred to* by the speaker. Currently there has been a great degree of interest in automatically analyzing the emotional content of language. The language used as input to this kind of analysis can be a speech recording or textual representation of language. The goal of such analysis is to determine emotions *expressed by* the speaker or writer, i.e., the emotions that the speaker currently feels. However, analyzing an utterance or document at a gross level, automatic analysis can be confused when a speaker *refers to* emotions that are not his or her own current emotions. Examples of this are quotations, storytelling/gossip, counterfactual reasoning, *post facto* emotional self-report, and abstract references to emotions.

He said that he was mad. (quotation)

Did you see how mad John was? (gossip)

If you eat my ice cream, I will get mad. (counterfactual reasoning)

I was mad when I found my car stolen last year. (self-report)

Anger is one of the seven sins. (abstract references).

In these examples, an automated system would detect anger, but in fact the writer of these sentences is not actually feeling anger at the current time. In many cases, such a subtle distinction might not be pertinent, but the study of natural language descriptions of emotions brings this distinction into focus. The ability to talk about things besides the here-and-now is an important characteristic of language, which has been termed *displacement [2]*. With respect to emotion research, natural language descriptions of emotions have been examined in ethnography, comparative linguistics, and cognitive science [3, 4, 5, 6]. In the natural language processing (NLP) domain we present in this paper, there is a need for analysis that is sensitive to such a distinction. This game the game of Emotion Twenty Questions (EMO20Q), which is described in [7], is primarily about *referring to* emotions: the players pick hypothetical emotions–not necessarily the emotions they feel–to use in a twenty questions-type game. In the game, players also *express* emotions: they are happy or relieved when they correctly guess the emotion and feel mild anguish if they cannot. However, in this paper we focus on they behavior of the players when they *refer to* emotions.

The game of EMO20Q was primarily designed as a way to elicit natural language descriptions of emotion: how people refer to emotions abstractly using language. While it is possible to communicate one's emotions to a first-hand observer by facial expressions, vocal modulation, or lexical choice, in order to communicate emotional information to others who are not first-hand observers it is necessary to describe emotions abstractly rather than to simply express or display the emotions. At the most basic level, natural language descriptions of emotions include words that name emotions, e.g. *angry*, *happiness*, etc. However, due to the productive, generative nature of natural language, it is possible to refine and generalize emotion descriptions to with longer natural language phrases. Furthermore, natural language descriptions of emotions can be related to each other in a similar way that words are related to other words through their dictionary definitions.

In order to communicate using natural language descriptions of emotions, people must be able to come to a shared understanding about the meaning of these descriptions. Russell [8] introduced the notion of *definite descriptions*, a logical device to used to model unique reference in the semantics of languages, both formal and natural. In this paper, we focus on the natural language definite descriptions. Common examples of natural language definite descriptions are proper names and noun phrases with the definite article "the" . Considering natural language descriptions of emotions in this way begs the question of whether natural language descriptions of emotions are definite descriptions or not. By considering terms that refer to emotions as definite descriptions, we are trying to capture the intuition that people mean the same things when they use the same emotion terms. In [9], the question is posed of whether emotions are natural kind terms, to which the paper answered no, i.e., that different emotion terms represent non-unique classes of human behavior rather than fundamentally distinct natural classes. The question of whether emotion terms are definite descriptions can be seen as a less stringent criterion than that of a natural kind. In this paper we formulate natural language descriptions of emotions in terms of definite descriptions and we show how such references to emotions are grounded in natural language dialogs using an adaptation of the mathematical notion of convergence empirically grounded in observed natural language dialogs.

2.2. EMO20Q, Crowd-Sourcing, and Experimental Design

By relying on the wisdom of the masses, we can venture a simple answer to the difficult question, "what is emotion?". The answer, according to crowd-sourcing, is that emotion is what most people say it is. Although this answer side-steps many important issues, such as physiological and psychological descriptions of emotions, it does bring other issues into sharper focus. Currently, there is a movement toward studying non-prototypical emotional data. Non-prototypical emotional data is exemplified by disagreement among annotators when assigning emotional labels to data. We argue that a crowd-sourced description of emotions can effectively deal with non-prototypical emotions. To avoid falling into the *ad populem* logical fallacy, we formulate the answer to the question "what is emotion?" not as a question of truth, but a question of knowledge and belief, i.e., an issue of epistemology, in effect skirting the question of ground truth, but asking another interesting question "what do people believe about emotions and how do they express these beliefs in language?".

When considering people's beliefs about emotions from the point of view of an annotation task, some of the disagreement between annotators with respect to non-prototypical emotion data can actually be seen as an artifact of being forced to choose from a set of "basic" emotions. Moreover, annotation tasks typically ask annotators to assign natural language descriptors to data without the context of natural language interaction, i.e. without the implied communicative goals that are shared with another interlocuter. When the annotation task is set up as a forced choice between one of several labels, it is obvious that any emotional data that does not fit into the preassigned labels will be seen as non-prototypical. However, more fundamentally, even when there is an open choice of emotion labels, the task of annotation is divorced from the full context of natural language interaction, so agreement is defined as a hit-or-miss event. For this reason, many annotation methodologies use a manual or training (ADOS, etc) to establish the context for the shared meaning of the annotation vocabulary. However, in the case of annotating emotions, the technical vocabulary overlaps with the natural language terms, so there is the potential for ambiguity between the meaning established by the annotation standards and the meaning of everday language [7].

The game of EMO20Q addresses the issue of agreement about emotion descriptions when the descriptive process is contextualized in natural language interaction. EMO20Q is played like the traditional game of twenty questions, except that one player, the *answerer*, must choose an emotion term. The other player, the *questioner*, must try to guess the emotion that the answerer picked by posing a series of twenty or fewer questions. Importantly, we do not restrict the set of emotion terms that the players can choose nor the questions they can ask: the set of emotion terms and questions is not limited except by the players' judgment and cooperation.

The EMO20Q game is an effective way to elicit language that people use to state their beliefs about emotions. In terms of the utility of EMO20Q as an experimental design, we distinguish the human-human version of the game, which is played by two human players, and the human-computer version of the game, in which one player is a computer. In this paper we consider the case where the question-asker, or *questioner*, is a computer.

In terms of experimental design, the human-human EMO20Q is a *quasi-experiment* or *natural experiment*, as opposed to a *controlled experiment,* which means that there is not a manipulation of variables made by the experimenters, but rather that the these variables are observed as they vary naturally within the system. Much past work [10, 11, 12, 13, 14, 15, 1] has focused on controlled experiments for studying emotional language, usually eliciting responses from subjects who are presented with words as stimuli. However, the stimuli are predetermined and the responses are constrained, often as a Likert scale or prescribed set of emotion categories. In EMO20Q, potentially any word can be chosen as an emotion word and any question can be asked of it; it is only limited by the game's rules, the subjects' good-faith sportsmanship and the extent to which that one players's model of emotions overlaps with the other player's model. Thus, in contrast with purely elicited experiments, EMO20Q can be presumed to have higher experimental (ecological?) validity and less experimental bias. The player/subjects of EMO20Q are less constrained by the elicitation methodology and, all other factors being equal, we can assume that their honesty and cooperation is comparable to elicitation experiments, or perhaps improved due to the presence of the other player and the players' shared communicative goals. Thus we can assume EMO20Q to have less experimental effects than in experiments with guided/elicited responses. There is the possibility of experimental effects due to interactions between players, which we consider in the analysis of our results.

One drawback of this approach is that it is hard to quantify reliability in the unconstrained interactions of EMO20Q. Reliability can be measured in the amount of agreement between subjects, but this can be difficult because we do not force subjects to pick any particular words, so the words that are in common between users are determined by chance and hence sparse. Determining an appropriate sample size as well as methods to deal with sparsity are important characteristics of this methodology which we discuss later.

Another advantage of the EMO20Q methodology is the potential to provide more experimental sensitivity. Receiving stimuli and giving responses using the same modality, natural language, has the potential to be much more sensitive than Likert scales or restricting user input to fixed choices. This is because we can assume that natural language has the capabilities of expressing most, if not all, of the nuanced distinctions between emotions. Even in cases where one is literally "at a loss for words", there are natural language descriptions, like the quoted phrase, that people use to approximate such an emotion. One exception where the natural language modality could be less sensitive is in the case of children and non-native speakers. In this case, we can imagine less fluent subjects who have conceptual distinctions in their beliefs about emotions that they are not able to verbally express without the aid of elicitation.

The methodological utility of the natural language modality in EMO20Q can be seen in the productivity, ubiquity, and social aspects of language, as well as the relation to engineering solutions, such as natural language processing. From the perspective of natural language processing, the EMO20Q game experiment can be seen as a *Wizard of Oz* experiment that can be used to collect human behavior that can be used to inform an automated agent. Games like EMO20Q can be seen as *games with a purpose* [16] whose purpose is *crowd-sourcing* [17] the collective knowledge and beliefs of the players [7]. The phenomenon of crowd-sourcing is closely tied to the emergent properties of online social communities [18].

The human-computer version of EMO20Q offers more possibilities for experimental control. We examine whether it is possible to control the EMO20Q game using a computer questioner agent and explore different possible ways of controlling such an agent.

2.3. A Socratic Epistemology of Emotions

In [7] we described the construction of a person-specific theories of emotions based on data from EMO20Q, inspired by the definition of a theory from mathematical logic, where a theory is defined as the set of statements in some language that is true of a model. This person-specific theory can be thought of a representation of a person's beliefs about emotions. Though we refer to natural language statements about emotions as being either true or false with respect to a given emotion and particular person, such a proposition or its negation may more appropriately be called a belief of the person, i.e., an epistemological issue. The particular way in which we try to determine the beliefs that people hold about emotions is inspired by the Socratic method of asking questions.

Besides being seen as a method for eliciting language by crowd-sourcing, the questioning asking process can be seen in deeper terms as central to how people understand and describe the world. The logician Charles S. Peirce identified three types of thought processes by which a person can acquire knowledge: deduction, induction, and hypothesis [19]. The third of these, also known as abduction [20], has been compared with the Socratic method of questioning [21]. Socrates applied his method of inquiry to examine concepts that seem to lack any concrete definition, in particular some of the moral concepts of his time like "justice", "knowledge", "piety", "temperance", and "love". We claim that this method of inquiry can shed light on how people define emotional concepts, which also seem to defy concrete definition. In addition to collecting an increasing list of propositions that are true or false of a given emotion for a particular person, the ordering of the questions shows a conceptual continuity that reflects a persons train of thought and the relations of similarity and subsethood between emotion concepts [22] for that person. Moreover, while some questions are better than others and can be evaluated on a per-question basis [23], questioning can also be evaluated at the level of a policy [24] or strategy.

3. Human-Human EMO20Q

The EMO20Q experiments we conducted can be partitioned into human-human and human-computer experiments. This section will examine the human-human experiements and the next will focus on using these results for the human-computer experiments.

3.1. Data and Methodology

We collected a total of 110 matches from 25 players in the human-human EMO20Q experiments. On average this is about 8 matches per player. However, we had two main types of subject: volunteers and paid participants. Most volunteers played two matches per player, one match per role, so two matches per subject was the median. The paid participants were part of a longitudinal experiment that aimed to see the effect of playing the game with each other over a longer period of time, which resulted in a maximum of 57 matches for the two longitudinal experiments. The EMO20Q experiment was is implemented as an online chat application using the Extensible Messaging and Presence Protocol (XMPP) so that the games can be easily recorded and studied.

Early in our pilot studies, we realized that it was difficult to successfully terminate the game when the questioner guessed words that were synonyms of the word the answerer picked. This led us to treat the phenomenon of synonyms with an additional rule that allowed the game to terminate if the answerer could not verbally explain any difference between the two words. In this case, we considered the game to terminate successfully, but we flagged these matches and kept track of both words.

Table 1: Examples of question standardization.

Standardized Question	**Examples**
cause(emptySet,e)	*can you feel the emotion without any external events that cause it?* *is it an emotion that just pops up spontaneously (vs being triggered by something)?*
cause(otherPerson,e)	*is it caused by the person that it's directed at?* *Do you need someone to pull this emotion out of you or evoke it? if so, who is it?*
e.valence==negative	*is it considered a negative thing to feel?* *2) so is it a negative emotion?*
situation(e,birthday)	*would you feel this if it was your birthday?* *is it a socially acceptable emotion, say, at a birthday party?*
e==frustration	*oh, is it frustrated?* *frustration?*

Since the surface forms of the questions vary widely, we used manual preprocessing to standardize the questions to a logical form that is invariant to wording. This logical form converted the surface forms to a pseudo-code language with a controlled vocabulary by converting the emotion names to nouns if possible, standardizing attributes of emotions and the relations of emotions to situations and events. Examples of the standardized questions are shown in Tab. 1. After this semantic standardization, there were a total of 727 question types.

To get a better idea of the relative frequencies of general types of questions, we made the following high level characterization of the questions using the following broad categories: *identity questions* (guessing an emotion), *attribute questions* (asking about dimensional attribute like valence or activation), *similarity/subsethood questions* (asking if the emotion in question is similar to or a type of another emotion), *situational questions* (questions that ask about specific situations associated with the emotion in question), *behavior questions* (questions that are asked about the behavior associated with the emotion in question), *causal questions* (questions about the cause, effect, or dependency of the emotion in question), *social questions* (questions asking about other parties involved in the emotion–this overlaps somewhat with causal questions and situational questions), *miscellaneous questions* (questions that defied classification or had categories with too few examples). Some examples of these categories are given in Tab. 2.

3.2. Results

3.2.1. Successful Game Outcome Rate

Of the 110 matches played between human players, 94 – approximately 85% – terminated successfully with the questioner correctly identifying the emotion that the answerer picked or a word that the answerer felt was a synonym. The mean and median number of questions asked per game was 12.0 and 10, respectively, when failures to correctly guess the emotion were averaged in as 20 questions.

Of the 94 successfully terminated matches, 22 terminated with synonyms. The 16 unsucessfully terminated matches that were considered failures consisted

Table 2: Examples of question categories. NOTE: UPDATE

Question Categories	**Examples**
identity (42%)	*is it angry?* *guilt?*
attribute (13%)	*is it something one feels for long periods of time?* *is it a strong emotion?*
similarity/ subsethood (10%)	*is the emotion a type of or related to content or zen contentment (is that a word?_)* *so it's similar to excited?*
situational (14%)	*is the emotion more likely to occur when you are tired?* *would i feel this if my dog died?*
behavior (3%)	*you can express it in an obvious way by sighing?* *do adults usually try to discourage children from feeling this?*
causal (7%)	*yes. mynext question is can it harm anyone besides the feeler?* *I think I know, but I'll ask one more question...does it ever cause children to wake up and cry?*
social (8%)	*are you less likely to experience the emotion when around good firiends?* *13)would you feel that towards someone who is superior to you?*
miscelaneous (3%)	*i dont' know if this is a valid question, but does it start with the letter D?* *or an aspirational emotion?* *does the word function or can be conjugated as anything eles? i.e. can it be a verb too?*

of several distinct cases. The questioner player could give up early if they had no clue (5/16), they could give up at twenty questions (1/16), or they could pass twenty questions due to losing count or as a matter of pride (6/16). The four remaining cases were considered failures because the answerer inadvertently gave away the answer due to a typing error or giving an unnecessarily generous hint. These four cases were all in the longitudinal experiment.

3.2.2. Emotions

There were unique 71 words that players chose in the human-human games, 61 of which were correctly identified. These are listed in Table 3.

3.2.3. Questions

There was a total of 1228 question-asking events. Of the questions, 1102 are unique (1054 after normalizing the questions for punctuation and case). In Table 4 we list some of the questions that occurred more than once.

4. Human-Computer EMO20Q

Using the human-human data described in the previous section, we built a computer agent to play the questioner in human-computer games.

4.1. Data and Methodology

The model we use for the agent is a sequential Bayesian belief update algorithm that starts with a conditional probability distribution estimated from a corpus of human-human and human-computer EMO20Q. The human-human data was described in the previous section and the human-computer data was

Table 3: Emotion words from human-human EMO20Q matches.

emotions (synonyms)	count	# correct
admiration	1	1
adoration	1	0
affection (love)	2	2
amusement	1	1
anger	2	1
annoyance (irritated)	2	2
anxiety	3	3
apathy (uninterested)	1	1
awe	1	0
boredom	2	2
bravery	1	1
calm	2	2
cheerfulness	1	1
confidence	1	1
confusion	2	1
contempt	1	1
contentment (calm)	2	1
depression (misery)	2	2
devastation	1	0
disappointment	1	1
disgust	2	2
dread (hopelessness)	1	1
eagerness (determination)	1	1
embarrassment	2	2
enthusiasm (eagerness)	3	1
envy (jealosy)	3	3
exasperation	1	1
excitement	1	1
exhilaration (thrill)	1	1
exhaustion	1	1
fear (distress,scared)	2	2
frustration	2	2
fury	1	1
glee	1	0
gratefulness	1	1
grumpiness	1	1
guilt	4	4
happiness	1	1
helplessness	1	1
hope (feeling lucky)	3	3
insecurity (shyness)	1	1
jealousy (envy)	3	3
joy	1	0
loneliness	1	1
love	2	2
madness (anger)	1	1
melancholy	1	1
pity (sympathy)	1	1
pride	2	2
proud	1	1
regret	2	2
relief	5	5
sadness	2	2
satisfaction	1	0
serenity	1	1
shame	1	1
shock	1	1
shyness	1	1
silly	1	1
soberness	1	0
sorrow (sadness)	1	1
stress	1	1
suffering	1	0
surprise	3	3
tense (uncomfortable)	1	0
terror	1	1
thankful	1	0
thrill (entrancement)	2	1
tiredness	2	2
wariness	1	0
worry (anxiety, scared)	3	3
total	110	94

derived from earlier versions of the agent [25]. In this data, there was a set of 105 emotion words that were observed. Let E be this set of 105 emotion words and let $\varepsilon \in E$ be a categorical, Bayesian (i.e., unobserved) random variable distributed over this set. Each question-answer pair from the match of EMO20Q is considered as an observed feature of the emotion being predicted. Thus if Q is the set of questions and A is the set of answers, then a question $q \in Q$ and an answer $a \in A$ together compose the feature $f = (q, a)$, where $f \in Q \times A$. The conditional probability distribution, $P(f|\varepsilon)$, is estimated from the training data using a smoothing factor of 0.5 to deal with sparsity.

In our model we stipulate that the set of answers A are four discrete cases: "yes", "no", "other", and "none". When the answer either contains 'yes' or 'no', it is labeled accordingly. Otherwise it is labeled 'other'. The feature value 'none' is assigned to all the questions that were not asked in a given dialog. 'None' can be seen as a missing feature when the absence of a feature may be important. For example, the fact that a certain question was not asked about a particular emotion may be due to the fact that that question was not relevant at a given point in a dialog.

Similarly, we stipulate that the questions can be classified into some discrete

Table 4: Examples of some of the questions that occurred multiple times (disregarding case and punctuation).

question	count
is it positive?	16
ok is it a positive emotion?	15
is it a positive emotion?	14
is it intense?	13
ok is it positive?	10
is it a strong emotion?	7
is it like sadness?	6
is it sadness?	5
is it pride?	5
is it neutral?	5
is it like anger?	5
is it surprise?	4
is it an emotion that makes you feel good?	4
thrilled?	3
regret?	3
pleased?	3
is it very intense?	3
is it love?	3
is it kinda like anger?	3
is it associated with sadness?	3
...	...
ok is it a negative emotion?	2
ok is it a good emotion?	2
okay is it a strong emotion?	2
is it highly activated?	2
is it directed towards another person?	2
is it directed at another person?	2
is it associated with satisfaction?	2
is it associated with optimism?	2
is it associated with disappointment?	2
is it an emotion that lasts a long time	2
does it vary in intensity?	2

class that is specified through a semantic expression as described in [7]. For example, the question “is it a positive emotion?” is represented as the semantic expression “e.valence==positive”. If the answer to this question was “maybe”, the resulting feature would be represented as `(‘e.valence==positive’,‘other’)`.

Using Bayes’ rule and the independence assumption of the naïve Bayes model, we can formulate the agent’s belief about the emotion vector ε after observing features $f_1 ... f_t$ as

$$P(\varepsilon | f_1, ..., f_t) = \frac{\prod_{i=1}^{t} [P(f_i|\varepsilon)] \, P(\varepsilon)}{\prod_{i=1}^{t} P(f_i)}. \quad (1)$$

When the game begins the agent can start with a uniform prior on its belief of which emotion is likely or it can use information obtained in previously played games. In the experiment of this paper, we use a uniform prior, $P(\varepsilon = e_k) = 1/|E|$, $\forall k = 1...|E|$. We chose to use the uniform prior to start with because our training data contains many single count training instances and because we want to examine how the system performs with less constraints.

In Equation 1, the posterior belief of the agent of emotion e_k at time t, $P(\varepsilon = e_k | f_1, ..., f_t)$ is computed only after the agent has asked the t questions.

In contrast the formulation we use is dynamic in that the agent updates its belief at each time point based on the posterior probability of the previous step, i.e., at time t

$$P(\varepsilon|f_1,...,f_t) \quad = \quad \frac{P(f_t|\varepsilon)P(\varepsilon|f_1,...,f_{t-1})}{P(f_t)}$$

We introducing a new variable $\beta_{t,k} = P(\varepsilon = e_k|f_1,...,f_t)$ for the agent's belief about emotion k at time t and postulate that the agent's current prior belief is the posterior belief of the previous step. Then, the agent's belief unfolds according to the formula:

$$\begin{aligned} \beta_{0,k} &= P(\varepsilon = e_k) = 1/|E| \\ \beta_{1,k} &= \frac{P(f_1|\varepsilon = e_k)}{P(f_1)}\beta_{0,k} \\ \beta_{t,k} &= \frac{P(f_t|\varepsilon = e_k)}{P(f_t)}\beta_{t-1,k} \qquad (2) \end{aligned}$$

Decomposing the computation of the posterior belief allows the agent to choose the best question to ask the user at each turn, rather than having a fixed battery of questions. In this case, we define "best" as the question that is most likely to have a 'yes' answer given ε. This criterion indicates how often the question was observed in the training data in the context of emotions as they are currently weighted by $P(\varepsilon|f_1,...,f_{t-1})$. The agent asks the best question and takes the user's response as input. It then parses the input to classify it into one of {yes, no, other}. This information is then used to update the agent's belief as to which emotion in E in most likely.

Identity questions are a special type of question where the agent makes a guess about the emotion. An affirmative answer to an identity question (e.g., "is it happy?") means that the agent successfully identified the user's chosen emotion. Any other answer to an identity question will set the posterior probability of that emotion to zero because the agent can be sure it is not the emotion of interest. Also, because it is playing a twenty questions game d is set to 20, but this could be changed for the agent to generalize to different question-asking tasks. The pseudo-code for the main loop of the adaptive Bayesian agent is shown in Algorithm .

Using this model and algorithm we made an agent that could play EMO20Q. For more details about the implementation, see [6]. We collected experimental data of fifteen subjects each playing three EMO20Q matches. The subjects were asked to pick 3 emotion words, one easy word, one medium word, and one difficult one, based on how difficult they thought it would be to guess. They were also asked to rate the naturalness of the agent on a 0-10 scale, and were given an opportunity for open-ended comments.

4.2. Results

The results of our usability experiments on fifteen subjects are summarized in Table 4.2 . To compare the agent's performance with human performance,

Algorithm 1 adaptive Bayesian emo20q agent

```
Input: $F = Q \times A$, $E$, and $P(f|\varepsilon)$
$\beta_{0,k} \leftarrow 1/|E|, \forall k = 1...|E|$
for $i = 1$ to $d$ do
    $q^{(i)} = \underset{q \in Q}{\text{argmax}} P((q, \text{‘yes’})|\varepsilon)$
    Print $q^{(i)}$
    $a^{(i)} \leftarrow$ user’s input answer
    $f_i \leftarrow (q^{(i)}, a^{(i)})$
    $\beta_{i,k} \leftarrow \beta_{i-1,k} \cdot P(f_i|\varepsilon = e_k)/P(f_i), \forall k = 1...|E|$
    if ($q^{(i)}$ is identity question for $e_k \wedge a^{(i)} = \text{‘yes’}$) then
        Return: $e^* = e_k$
    end if
    if ($q^{(i)}$ is identity question for $e_k \wedge a^{(i)} = \text{‘no’}$) then
        $\beta_{i,k} \leftarrow 0$
    end if
end for
$k^* \leftarrow \underset{k \in 1...|E|}{\text{argmax}}[\beta_{i,k}]$
$e^* \leftarrow e_{k^*}$
Return: most likely emotion given observations: $e^*$
```

we used two objective measures and one subjective measure. The success rate, shown in column two Table 4.2, is an objective measure of how often the EMO20Q matches ended with the agent successfully guessing the user’s emotion. The number of turns it took for the agent to guess the emotion is the other objective measure. The last column, naturalness, is a subjective measure where users rated how human-like the agent was, on a 0-10 scale.

The emotion words chosen by the subjects as “easy” were recognized by the agent with similar success rate and number of required turns as human-human matches. Some examples of “easy” emotions are anger, happiness, and sadness. However, successful outcomes were fewer in emotions chosen as “medium” and “difficult”. Some examples of “medium” emotions are contentment, curiosity, love, and tiredness. Pride, frustration, vindication, and zealousness are examples of “difficult” emotions. The different classes of words were not disjoint: some words like anger, disgust, and confusion spanned several categories. A complete listing of the words chosen by the subjects of the experiment is given in Table 4.2.

The results in terms of successful outcomes and number of turns required to guess the emotion word are roughly reflected in the percent of words that are in vocabulary. Despite the low performance on emotion words deemed “medium” and “difficult”, there was not a corresponding decrease in the perceived naturalness of the questioner agent.

Table 5: Experimental results.

difficulty	% success	avg. turns	% in vocab.	naturalness
easy	73%	11.4	100%	6.9
medium	46%	17.3	93%	5.5
difficult	13%	18.2	60%	5.8
total	44%	15.6	84%	6.1

Table 6: Observed emotion words by difficulty.

difficulty	examples
easy	happiness, anger, sadness, calm, confusion
medium	anger, confusion, contentment, curiosity, depression, disgust, excitement, fear, hate, irritation, love, melancholy, sorrow, surprise, tiredness
difficult	devastation, disgust, ecstasy, ennui, frustration, guilt, hope, irritation, jealousy, morose, proud, remorse, vindication, zealousness

5. A Computational Model of Curiosity

The previous result, in which the subjects rated the agent as being relatively human-like even in cases when the EMO20Q task performance was low, seems to indicate that there is something human-like in the search for knowledge, rather than having a complete representation of emotions that would allow it to guess correctly and win at EMO20Q more often. In an earlier paper [23], we proposed a model for verbally exploring a sparse graphical representation of an agent's knowledge.

Representing the agent's knowledge as an adjacency graph A, allows us to use methods from collaborative filtering, social network analysis, and spectral graph theory [26, 27]. In this paper, we use the number of zero eigenvalues of the Laplacian of the graph A to determine the number of connected components of the graph. This can be seen as a measure of the sparsity of our data and can be used to identify the questions that must be asked of certain emotions in order to connect the graph components. In this way, we can have a rudimentary representation of curiosity in terms of the agent's exploration of this graph. To construct this graph we built a graph in the following way: questions and answers are nodes in the graph which are linked by edges labelled 1 if the question is answered "yes" of that emotion and -1 if the question is answered "no" of the emotion. Thus the graph A is bipartite in that emotion nodes are connected to question nodes, but there are no edges between any two emotion nodes or between any two question nodes.

We used the Laplacian L of this graph to calculate the connectivity and to identify which nodes were unconnected. The Laplacian L of a signed graph to is calculated by subtracting the absolute adjacency matrix $|A|$ from the diagonal absolute degree matrix $\bar{D}_{ii} = \sum_j |A_{ij}|$:

$$L = \bar{D} - |A|$$

From the matrix L we can tell the number of connected components of A by counting the number of zero eigenvalues. Thus, if there are three eigenvalues that equal zero, the graph is composed of three separate connected components. A graph Laplacian with one zero eigenvalue is a single connected graph. Moreover, the lowest n non-zero eigenvalues can be used to project nodes in the graph into an n-dimensional space (this is similar to principal components analysis (PCA) except that in PCA instances are projected to be farther apart, whereas in the graph projection, connected nodes ought to be projected close by).

If an agent has identified that it has unconnected knowledge, how can it then plan questions to address the knowledge goal of connecting the components of A? To answer this, we must define the notion of a *walk* on a graph. A walk of length l on graph A that joins vertices v_i and v_j is a sequence of vertices $u_0 \dots u_l$ of A such that $v_i = u_0$, $v_j = u_l$, and u_{t-1} and u_t are adjacent for $1 \leq t \leq l$.

According to [28], Lemma 2.5, the number of walks of length l in A that join v_i to v_j is the entry in cell (i, j) of the matrix A^l. Thus, by taking repeated powers of the absolute adjacency matrix $|A|$, we can determine if nodes v_i and v_j are connected by walks of length l. Since the graph is bipartite, the walks from question nodes to other question nodes or from object nodes to other object nodes will always be even length, and conversely, walks between question and object nodes will be odd length. This behavior is undesirable because we wish to preserve connectedness properties across repeated powers of A^l. To remedy this undesirable behavior we can augment the adjacency matrix $|A|$ by adding the identity matrix I to it. At this point, we can say that vertices v_i and v_j are connected by a walk of length l *or less* if the entry (i, j) of $(A + I)^l$ is non-zero. The proof of this, by contradiction, is that that if we imagine that vertices v_i and v_j are connected by some walk of length $k < l$, but not of length l, then there must not be self-loop from v_j to itself after the walk of length k. However, since we added the identity I matrix to A we know that there are in fact self-loops on all of the vertices.

The preceding fact allows us to state an alternative test for connectedness and also allows us to identify the question-object pairs that need to be asked to complete the agent's knowledge. This test can be stated as follows: the graph A is connected if and only if

$$(|A| + I)^{M+N-1}$$

has no zero entries, where M is the number of emotions and N is the number of questions. This is because the length of a walk with distinct steps, a *path*, is at most one less than the number of vertices in the graph, i.e., $M + N - 1$, which would be the case if the graph were a linked list. The question-object pairs that correspond to zero entries in this matrix are precisely the set of candidate questions that need to be asked to connect the agent's knowledge and choosing which one to ask can be determined by using L to project the current state of the dialog into the graph's space.

6. Discussion

more ideas: Push the ground truth issue, non-prototypical emotions
also push the idea of large vocabulary emotion models

7. Conclusion

One of the main results of this study is that when asking questions about emotions we can, in Socrates' words, reach mutual edification. Precisely, this mutual edification happens about 85% of the time after about 12 dialog turns. This represents a high amount of agreement for an emotional classification task, expecially considering that the set of emotional classes was unbounded. The reader may ask, though, what use is it to consider all these emotions, some of which may be synonymous? For example, what use is it to know that for some people, the words "pride" and "proud" may have different connotations?

[regarding "proud" vs. "pride"] because my intuition was that they're different... you know pride sometimes has a negative connotation

or that "anger" is might not always described as "negative"?

[questioner:] so is it a negative emotion?

[answerer:] sort of, but it can be righteous

For some purposes it might be sufficient to consider "anger" to be simply a negative emotion, or that the words "pride" and "proud" to refer to the same emotion. However, from nearly every line of EMO20Q data we can see that the ways that humans describe emotions are much more nuanced. Although analyzing this level of detail is beyond the scope of many current systems, we have shown that it is a task that humans can do with success rates that beat agreement rates on emotional annotations at a much coarser level. We deem that future advances in affective computing can come from studying emotions at this finer grain, as they are described in natural language. We have provided an anonymized version of data we gathered from EMO20Q as well as other resources at http://sail.usc.edu/emo20q .

8. Acknowledgments

players,Prof. Mendel, Prof. Hovy, Ben Alderson-Day, Prof. Tony David Sampson?, Prof. Whissell?, ACII reviewers, Nassos, Thomas E. Forster, Funding..., Ozan, Juanchen

References

[1] P. R. Shaver, U. Murdaya, and R. C. Fraley, “Structure of the indonesian emotion lexicon,” *Asian Journal of Social Psychology*, vol. 4, pp. 201–224, 2001.

[2] C. F. Hockett and S. Altmann, *A note on design features*, pp. 61–72. Indiana University Press, 1968.

[3] B. King, *The Conceptual Structure of Emotional Experience in Chinese.* PhD thesis, Ohio State University, 1989.

[4] Zoltán Kövecses, *Metaphor and Emotion: Language, Culture, and Body in Human Feeling.* Cambridge University Press, 2000.

[5] E. T. Rolls, *What Are Emotions, Why Do We Have Emotions, and What Is Their Computational Basis in the Brain*, ch. 5, pp. 117–146. Oxford University Press, 2005.

[6] A. Kazemzadeh, J. Gibson, J. Li, S. Lee, P. G. Georgiou, and S. Narayanan, “A sequential bayesian agent for computational ethnography,” in *Under review*, 2012.

[7] A. Kazemzadeh, P. G. Georgiou, S. Lee, and S. Narayanan, “Emotion twenty questions: Toward a crowd-sourced theory of emotions,” in *Proceedings of ACII'11*, 2011.

[8] B. Russell, “On denoting,” *Mind*, vol. 14, pp. 479–493, 1905.

[9] L. F. Barrett, “Are emotions natural kinds?,” *Perspectives on Psychological Science*, vol. 1, pp. 28–58, March 2006.

[10] C. M. Whissell, *The Dictionary of Affect in Language*, pp. 113–131. Academic Press, 1989.

[11] C. E. Osgood, G. J. Suci, and P. H. Tannenbaum, *The Measurement of Meaning.* University of Illinois Press, 1957.

[12] P. Oudeyer, “The production and recognition of emotions in speech: features and algorithms,” *J. Hum. Comput. Stud.*, vol. 59, pp. 157–183, 2003.

[13] J. A. Russell and A. Mehrabian, “Evidence for a three-factor theory of emotions,” *Journal of Research in Personality*, vol. 11, pp. 273–294, September 1977.

[14] A. Kazemzadeh, S. Lee, and S. Narayanan, “An interval type-2 fuzzy logic system to translate between emotion-related vocabularies,” in *Proceedings of Interspeech*, (Brisbane, Australia), September 2008.

[15] A. Kazemzadeh, “Using interval type-2 fuzzy logic to translate emotion words from spanish to english,” in *IEEE World Conference on Computational Intelligence (WCCI) FUZZ-IEEE Workshop*, 2010.

[16] L. von Ahn and L. Dabbish, "Labeling images with a computer game," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, 2004.

[17] J. Howe, "The rise of crowdsourcing," *Wired Magazine*, vol. 14.06, June 2006.

[18] N. Zhong, J. Liu, Y. Yao, and S. Ohsuga, "Web intelligence," in *Computer Software and Applications Conference*, 2000.

[19] C. S. Peirce, "Some consequences of four incapacities," *Journal of Speculative Philosophy*, vol. 2, pp. 140–157, 1868.

[20] U. Eco and T. A. Sebeok, eds., *The Sign of Three: Dupin, Holmes, Peirce.* Advances in Semiotics, Indiana University Press, 1988.

[21] J. Hintikka, *Socratic Epistemology: Explorations of Knowledge-Seeking by Questioning.* Cambridge University Press, 2007.

[22] A. Kazemzadeh, S. Lee, and S. Narayanan, "An interval type-2 fuzzy logic model for the meaning of words in an emotional vocabulary," Under review.

[23] A. Kazemzadeh, S. Lee, P. G. Georgiou, and S. Narayanan, "Determining what questions to ask, with the help of spectral graph theory," in *Proceedings of Interspeech*, 2011.

[24] B. Jedynak, P. I. Frasier, and R. Sznitman, "Twenty questions with noise: Bayes optimal policies for entropy loss," *Journal of Applied Probability*, vol. 49, pp. 114–136, March 2012.

[25] A. Kazemzadeh, J. Gibson, P. Georgiou, S. Lee, and S. Narayanan, "Emo20q questioner agent," in *Proceedings of ACII (Interactive Event)*, 2011. The interactive demo is available at http://sail.usc.edu/emo20q/questioner/questioner.cgi.

[26] J. Kunegis, A. Lommatzsch, and C. Bauckhage, "The slashdot zoo: Mining a social network with negative costs," in *World Wide Web Conference (WWW 2009)*, (Madrid), pp. 741–750, April 2009.

[27] Y. P. Hou, "Bounds for the least laplacian eigenvalue of a signed graph," *Acta Mathematica Sinica*, vol. 21, no. 4, pp. 955–960, 2005.

[28] N. Biggs, *Algebraic Graph Theory.* Cambridge University Press, 1974.