A Socratic epistemology for verbal emotional intelligence

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

Signal Analysis and Interpretation Laboratory, University of Southern California, Los Angeles, CA, United States

ABSTRACT

We describe and experimentally validate a question-asking framework for machine-learned linguistic knowledge about human emotions. Using the Socratic method as a theoretical inspiration, we develop an experimental method and computational model for computers to learn subjective information about emotions by playing emotion twenty questions (EMO20Q), a game of twenty questions limited to words denoting emotions. Using human–human EMO20Q data we bootstrap a sequential Bayesian model that drives a generalized pushdown automaton-based dialog agent that further learns from 300 human–computer dialogs collected on Amazon Mechanical Turk. The human–human EMO20Q dialogs show the capability of humans to use a large, rich, subjective vocabulary of emotion words. Training on successive batches of human–computer EMO20Q dialogs shows that the automated agent is able to learn from subsequent human–computer interactions. Our results show that the training procedure enables the agent to learn a large set of emotion words. The fully trained agent successfully completes EMO20Q at 67% of human performance and 30% better than the bootstrapped agent. Even when the agent fails to guess the human opponent's emotion word in the EMO20Q game, the agent's behavior of searching for knowledge makes it appear human-like, which enables the agent to maintain user engagement and learn new, out-of-vocabulary words. These results lead us to conclude that the question-asking methodology and its implementation as a sequential Bayes pushdown automaton are a successful model for the cognitive abilities involved in learning, retrieving, and using emotion words by an automated agent in a dialog setting.

INTRODUCTION

Epistemology is the branch of philosophy that deals with knowledge and belief. According to basic results in epistemology, knowledge is defined as true, justified belief. This paper was inspired by reflecting on how humans justify their beliefs about emotions. This reflection led to an experimental method for collecting human knowledge about emotions and a computational model that uses the collected knowledge in an automated dialog agent.

The logician Charles S. Peirce identified three types of thought processes by which a person can justify their beliefs and thereby acquire knowledge: induction, deduction, and hypothesis (*Peirce, 1868*). Whereas induction is primarily involved with observational data, deduction and hypothesis have a linguistic, propositional component. The third of these, hypothesis (also known as abduction (*Eco & Sebeok, 1988*)), has been compared with the Socratic method of question-asking dialogs (*Hintikka, 2007*). The Socratic method was named after the ancient Greek philosopher Socrates who applied his method of inquiry to examine concepts that seem to lack any concrete definition, in particular some of the complex moral and psychological concepts of his time like "justice", "knowledge", "piety", "temperance", and "love". We claim that this method of inquiry can shed light on how people justify beliefs about emotional concepts, which also seem to defy concrete definition.

Question-asking allows people to learn about things without directly experiencing them. Since a computer agent cannot directly experience emotions as a human would, question-asking can be leveraged for the computer agent to learn about emotional concepts. Question-asking has also been proposed as a stage in child development responsible for rapid learning and language acquisition (*Frazier, Gelman & Wellman, 2009*). Likewise, a computer agent can use question-asking to acquire knowledge and vocabulary. We call the approach of using question-asking to interactively acquire linguistic knowledge about emotion by a computer dialog agent a *Socratic epistemology for verbal emotional intelligence*.

The knowledge acquired by the Socratic epistemology for verbal emotional intelligence is an informal, social type of knowledge. This informal knowledge about emotions is important because although there has been much recent progress toward understanding the underlying biological basis for emotion, humans have been able to understand emotions informally since ancient times. We call this informal, language-based understanding of emotions *natural language description of emotion* (*Kazemzadeh, 2013*). Natural language descriptions of emotion are utterances that refer to emotions, as opposed to utterances that express emotions. This phenomenon can be seen as a specific 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. Studying language that deals with referential statements about emotions is a novel focus that is distinct from the prevailing trends of studying the expressive characteristics of emotional language.

The framework we present also differs from other computational 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 (*King, 1989*; *Shaver, Murdaya & Fraley, 2001*; *Mihalcea & Liu, 2006*; *Kamvar & Harris, 2009*; *Kazemzadeh et al., 2012*) and crowd-sourcing. Although there is clearly a biological basis for emotions, the way that people describe emotions in natural language varies over time from culture to culture and person to person. A descriptive theory of how people describe emotions in

natural language should reflect this variation. Thus, a particular instantiation of a theory of natural language description of emotion should be relative to a particular agent or set of agents at a given time who hold similar beliefs and communicate using the same language. Rather than trying to establish single, particular definitions for a predetermined set of emotion words, we study how shared meaning can be established through natural language dialog behavior even when people and agents may have different ways of experiencing and describing an undetermined set of emotion words.

The key goals of this paper are to use question-asking to observe natural language descriptions of emotion in the context of interactive dialogs and to computationally model the communicative behavior that supports the referential link between language and emotions. To do this, we use an experimental game called emotion twenty questions (EMO20Q). EMO20Q is played like the traditional game of twenty questions, except that one player, the *answerer*, must choose a word that names an emotion. 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 words that the players can choose nor the questions they can ask: the set of emotion words and questions is not limited except by the players' judgment and cooperation. An example of a human–computer EMO20Q game is given below:

Agent: *Let me know when you have picked an emotion word.*

User: *Ok*

Agent: *Is it very intense?*

User: *No*

Agent: *Is it positive?*

User: *Yes*

Agent: *Is it an emotion that just pops up spontaneously (vs. being triggered by something)?*

User: *Depends*

Agent: *Is it caused by anything in particular*

User: *Well, yes, you need the right conditions*

Agent: *Is it relaxation?*

User: *Yes*

Agent: *So did I get it right?*

User: *Yes, you got it.*

Agent: *Awesome! Would you like to play again?*

User: *No, thanks.*

Agent: *Thank you for playing*

The data from the human–human EMO20Q games allows us to observe what words people consider to denote emotions, what questions people ask about emotions, what

factors determine the length of question/answer exchanges, and what is the probability that the emotion words will be correctly guessed. These results are used to create the computer agent and test its performance with respect to human performance. For the automated agent presented in this paper, we focused on the case where the computer agent plays the role of the questioner while future work for an answerer agent is discussed in 'Discussion'.

The paper is organized as follows. 'Background' discusses the motivations and theory behind our work. 'Model' describes the computational model and algorithm we used to create an EMO20Q questioner agent. 'Experiments' discusses experiments we conducted of humans and computers playing EMO20Q. 'Results' describes the results of testing the agent. Finally 'Discussion' and 'Conclusion' propose future work and provide discussion and links to open source software implementations.

BACKGROUND

Natural language descriptions of emotions

Just as memory addresses, variables, and URLs refer to electronic resources for computers, so 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 helpful in the case of affective computing, when the objects we wish to refer to, emotions, are 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 emotional expression in 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. The language used as input to this kind of analysis can be a speech recording or textual representation of language. However, automatically analyzing the emotions expressed in an utterance or document is problematic when a speaker *refers to* emotions that are not his or her own current emotions. Some examples of this include 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)
> *I was mad when my car got stolen last year.* (self-report)
> *Anger is one of the seven sins.* (abstract reference)

In these examples, a naïve automated analysis would detect anger, but in fact the writer of these sentences is not actually feeling anger at the current time. In many cases, such as task-driven dialogs like ordering airline tickets from an automated call center, this distinction might not be pertinent. However, for open-ended dialog systems the distinction between expression and reference of emotions could be relevant, for example an automated agent for post-traumatic stress disorder therapy. The study of natural language descriptions of emotions brings the distinction between emotion expression and reference into focus.

The ability to talk about things beyond the here-and-now has been termed *displacement* (*Hockett & Altmann, 1968*). Displacement is an important characteristic that distinguishes human language from animal communication. In the context of this research, the ability to talk about an emotion without it being physically present is a key component of natural language description of emotion. Natural language description of emotion has been examined in ethnography, comparative linguistics, and cognitive science and it is beginning to be studied in the domain of natural language processing (*King, 1989*; *Kövecses, 2000*; *Rolls, 2005*; *Kazemzadeh et al., 2012*).

At the most basic level, natural language description of emotion includes 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 with longer natural language phrases. 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 (1905)* introduced the notion of *definite descriptions*, a logical device used to model unique reference in the semantics of languages, both formal and natural. In this paper, we focus on natural language definite descriptions. Common examples of natural language definite descriptions are proper names and noun phrases with the definite article "the". Indefinite descriptions, on the contrary, are prefaced with indefinite articles, such as "a", "some", or "every".

We maintain that natural language descriptions of emotions are definite descriptions when they are used in natural language interaction that terminates in mutual agreement. By considering terms that refer to emotions as definite descriptions, we are trying to capture the intuition that different people mean the same things when they use the same emotion terms. In *Barrett (2006)*, the question is posed of whether emotions are natural kind terms, to which the paper answered no, i.e., that emotion words in general represent non-unique classes of human behavior rather than fundamentally distinct biological classes. The question of whether emotion terms are definite descriptions can be seen as a less stringent criterion than that of whether they are natural kinds. The theoretical notion of definite descriptions captures the experimental observations in this paper, which show that there is a high degree of consensus in how humans describe emotions when they play EMO20Q. Specifically, we consider a series of questions that successfully terminates a game of EMO20Q to be an interrogative formulation of a definite description of the emotion that was successfully guessed, for the two players.

EMO20Q, crowd-sourcing, and experimental design

The game of EMO20Q was designed as a way to elicit natural language descriptions of emotion. Posing the experiment as a game leverages past results in crowd-sourcing and games with a purpose. From the perspective of natural language processing, the EMO20Q game can be seen as a *Wizard of Oz* experiment that collects human behavior to train the behavior of an automated agent. Games like EMO20Q can be seen as *games with a purpose* (*Von Ahn & Dabbish, 2004*) whose purpose is *crowd-sourcing* (*Howe, 2006*) the collective knowledge and beliefs of the players (*Kazemzadeh et al., 2011*). The

Kazemzadeh et al. (2016), *PeerJ Comput. Sci.*, DOI 10.7717/peerj-cs.40

5/23

phenomenon of crowd-sourcing is closely tied to the emergent properties of online social communities (*Zhong et al., 2000*).

By relying on the wisdom of the masses, we venture a simple answer to the difficult question, "what is emotion?". The answer, according to crowd-sourcing, is that emotion is what 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. There has been a trend toward studying *non-prototypical* emotional data (*Mower et al., 2009*). Non-prototypical emotional data is exemplified by disagreement among annotators when assigning emotional labels to data. We argue that our methodology provides a crowd-sourced description of emotions that 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 as described in 'Introduction', in effect skirting the question of ground truth, but asking other interesting questions: "what do people believe about emotions, how do they express these beliefs in language, and how do they justify their beliefs through question-asking behavior?"

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 these variables are observed as they vary naturally within the system. In the field of emotion research, two prevalent types of controlled experiments are annotation of recorded data and elicitation of experimental subjects. With annotation tasks, experimenters can control the vocabulary of annotation labels and with elicitation tasks experimenters can control the stimuli that are presented. With this control, experiments are more easily repeated. In EMO20Q, we did not control what emotion words or questions the subjects picked. Thus, for another population the results could vary, resulting in less experimental reliability. However, trading off control and reliability leads to more experimental sensitivity and naturalness and less experimental bias. In EMO20Q subjects can choose any words or questions they want and they communicate in a natural dialog setting. This way of characterizing emotion is closer to natural communication and more sensitive to nuances of meaning. For example, when forced to annotate using a fixed vocabulary of emotion words, subjects are experimentally biased toward using that vocabulary.

The automated dialog agent is one way to enforce more experimental control for EMO20Q. Because the agent's behavior is programmed we can use this as a way to better control and replicate experiments. Another way we aimed to improve experimental reliability is by prompting users to pick emotion words from three different difficulty classes. 'Model' and 'Experiments' further describe the computational model for the agent's behavior and our experimental design.

MODEL

Bayesian models have been successfully applied to a wide range of human cognitive abilities (*Griffiths, Kemp & Tenenbaum, 2008*), including inductive inference of word

meaning from corpora (*Steyvers, Griffiths & Dennis, 2006*) and experimental stimuli (*Xu & Tenenbaum, 2005*) and powering affective dialog agents (*Carofiglio, De Rosis & Novielli, 2009*). To our knowledge, this work is the first application of Bayesian cognitive models for learning emotion words from dialog interaction.

The model we use for the EMO20Q questioner agent is a sequential Bayesian belief update algorithm. This model fits the framework of Socratic epistemology, as described in the introduction, because it combines the notion of belief and question-asking. Intuitively, this algorithm instantiates an agent whose *semantic knowledge* is based on data from previous EMO20Q games. The agent begins a new game of EMO20Q with a uniform *belief* about the emotion word to be guessed. Based on the previous semantic knowledge, the agent asks questions and updates its belief based on each *observation* of the user's answers to the questions. While the EMO20Q game is played, the observations are stored in the agent's *episodic buffer* (*Baddeley, 2000*), also known as working memory. After the game, the agent updates its semantic knowledge using the results of the game, clears its episodic buffer, and is then ready to play again. The words in italics are high-level abstractions used to create a cognitive model for the agent, which is underlyingly implemented as a sequential Bayesian statistical model.

The *semantic knowledge* described above is implemented as the conditional probability of observing a set of question–answer pairs given a hidden variable ranging over emotion words. This conditional probability distribution is estimated from the corpus of past human–human and human–computer EMO20Q games as follows. Let E be the set of emotion words and let $\varepsilon \in E$ be this categorical, Bayesian (i.e., unobserved) random variable distributed over the set E. The cardinality of set E grows as the agent encounters new emotion words, from 71 words after the human–human experiments to 180 after the human–computer experiments. The probability of ε, $P(\varepsilon)$ is the *belief* about the emotion word to be guessed. Each question–answer pair from the game of EMO20Q is considered as an *observation* or 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 value $f = (q, a)$, i.e., $f \in Q \times A$. The conditional probability distribution, $P(f|\varepsilon)$, which represents semantic knowledge, is estimated from the past dialogs by counting the frequency that a question q was answered with answer a for a given emotion word.

The estimation of the conditional distribution $P(f|\varepsilon)$ used a smoothing factor of 0.5 to deal with sparsity and to prevent probabilities from becoming zero if the players answered in a nonstandard way. Besides the theoretical motivation of using a Bayesian model to represent the agent's belief state, the smoothing had the practical benefit of preventing the agent from getting stuck on an incorrect line of questions, for example, if we were to have used a decision tree model as an alternative baseline to the sequential formulation of naïve Bayes.

In this model we stipulate that the questions are classified into discrete classes that are specified through a semantic annotation of the human–human data, as described in 'Human-human EMO20Q'. 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 dimension would be represented as (`'e.valence==positive'`,`'other'`). More examples of the question features can be seen in Table 4. This set of question classes Q, of cardinality 727, was held constant in the human–computer experiments.

Similarly, we stipulate that the set of answers A are four discrete cases: "yes", "no", "other", and "none". Initially, the feature vector consists of "none" values corresponding to each question. After each turn, the feature dimension corresponding to the turn's question is updated with that turn's answer, either "yes", "no", or "other". When the answer either contains "yes" or "no", it is labeled accordingly. Otherwise it is labeled "other". The feature value "none" remains 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.

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 $f_1 \ldots f_t$, in one single batch, as opposed to sequentially (which will be formulated next):

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

This is simply the formulation of naïve Bayes, where in this case $P(\varepsilon)$ is the prior probability of a player choosing a specific emotion word, $\prod_{i=1}^{t}\big[P(f_i|\varepsilon)\big]$ is the likelihood of seeing question–answer pairs given specific emotion words, and $\prod_{i=1}^{t}P(f_i)$ is the probability of observing a set of question–answer pairs in general.

In terms of the high-level cognitive model, the set of observational feature vectors $f_1 \ldots f_t$ is what was described as the agent's *episodic buffer*. $P(f|\varepsilon)$ is the agent's *semantic knowledge* that relates question–answer features to emotion words. $p(\varepsilon)$ and $P(\varepsilon|f_1,\ldots,f_t)$ are the agent's initial/prior and final/posterior *beliefs*, respectively.

In Eq. (1), the posterior belief of the agent about emotion e_k at time t, $P(\varepsilon = e_k|f_1,\ldots,f_t)$ is computed only after the agent has asked all t questions. This model is known as naïve Bayes. In contrast the sequential Bayes model that we use is dynamic: the agent updates its belief at each time point based on the posterior probability of the previous step, i.e., at time t

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

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 experiments of this paper, we use a uniform prior, $P(\varepsilon = e_k) = 1/|E|, \forall k = 1\ldots|E|$. We chose to use the uniform prior to initialize the agent because our training data contains

many single count training instances and because we want to examine how the system performs with fewer constraints.

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

$$\beta_{0,k} = P(\varepsilon = e_k) = 1/|E|$$
$$\beta_{1,k} = \frac{P(f_1|\varepsilon = e_k)}{P(f_1)}\beta_{0,k}$$
$$\beta_{t,k} = \frac{P(f_t|\varepsilon = e_k)}{P(f_t)}\beta_{t-1,k}. \tag{2}$$

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.

We define "the best question" at time t to be the question class that is most likely to have a "yes" answer given the posterior belief at time $t - 1$, $P(\varepsilon | f_1, \ldots, f_{t-1})$:

$$\text{argmax}_{q \in Q} P((q, \text{'yes'}))|\varepsilon)P(\varepsilon | f_1, \ldots, f_{i-1}).$$

This next-question criterion is a heuristic motivated by considering "yes" answers to be positive feedback that the agent is on the right track. While this heuristic worked well in practice, other next-question criteria may be possible in future research. After the question class has been chosen it is converted from its logical form (e.g., `e.valence==positive`) to a concrete realization from the human–human data (e.g., "is it positive?") by choosing a random question that had been annotated with the logical form. In this way, the questions are more varied and might even include slang or typos that occurred in the human–human data. In terms of natural language generation, the choice of question class can be seen as pragmatic planning and the choice of surface form can be seen as a form of lexical access.

At time t 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 posterior belief $\beta_{t+1,k}$ about each emotion $e_k \in E$, which will then be used as the prior probability in the following step. The unfolding of variable β in Eq. (2) models the update of belief as it is justified by the agent's question-asking and the user's answers. It is this computational model of question-asking and belief update that represents the Socratic epistemology for verbal emotional intelligence in a software agent. Table 1 shows an example interaction between the automated EMO20Q questioner agent and a human user, along with a trace of the agent's belief state that shows the justification of beliefs by question-asking.

Identity questions are a special type of question where the agent makes a guess about the emotion that the other player has picked. Identity questions are chosen with the same best question criteria as other questions and the belief state is initially updated in the usual way. However, after the belief update there is a transition to a different dialog

Table 1 Example dialog from EMO20Q questioner agent where the human interlocutor has chosen "relaxation".

Speaker/Event	Utterance/Content
Agent	Let me know when you are ready …
User	Ok
Belief update	(Uniform distribution)
Agent	Is it very intense?
User	No
Belief update	Relaxation:0.027, hope:0.027, boredom:0.027, …, anger:0.0024
Agent	Is it positive?
User	Yes
Belief update	hope:0.15, relaxation:0.12, cheerfulness:0.091, …
Agent	Is it an emotion that just pops up spontaneously (vs. being triggered by something)?
User	depends
Belief update	hope:0.33, relaxation:0.23, cheerfulness:0.15, …
Agent	Is it caused by something in particular?
User	Well, yes, you do need the right conditions
Belief update	relaxation:0.50, hope:0.29, cheerfulness:0.097, …
Agent	Is it relaxation?
User	Yes
Agent	So did I get it right?
User	Yes, you got it
Agent	Awesome! Would you like to play again?
User	No, thanks.
Agent	Thank you for playing

state. An affirmative answer to an identity question (e.g., "is it happy?") means that the agent successfully identified the user's chosen emotion and the game is then terminated successfully. 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. The usual belief state update is performed even in the case of identity questions for two reasons. First, identity questions can be ambiguous: the example "is it happy?" can be interpreted as "Did you pick the emotion 'happiness'?" (i.e., an identity question) or "Did you pick a happy emotion?" (i.e., an ordinary question). In this case, the usual Bayesian belief update deals with the former interpretation. Second, identity questions, even when answered negatively, contain information that can be useful for updating the belief state beyond setting the posterior probability of the negatively identified emotion to zero.

The pseudo-code for the main loop of the adaptive Bayesian agent is shown in Algorithm 1. This automated, data-driven component was framed within a manually designed dialog graph, as shown in Fig. 1. The dialog graph is implemented as a generalized pushdown transducer. Recall that a pushdown transducer is a transducer that determines its output symbol and next state based on its current state, the input symbol, and the top of its stack (*Allauzen & Riley, 2012*). A generalized pushdown transducer is a
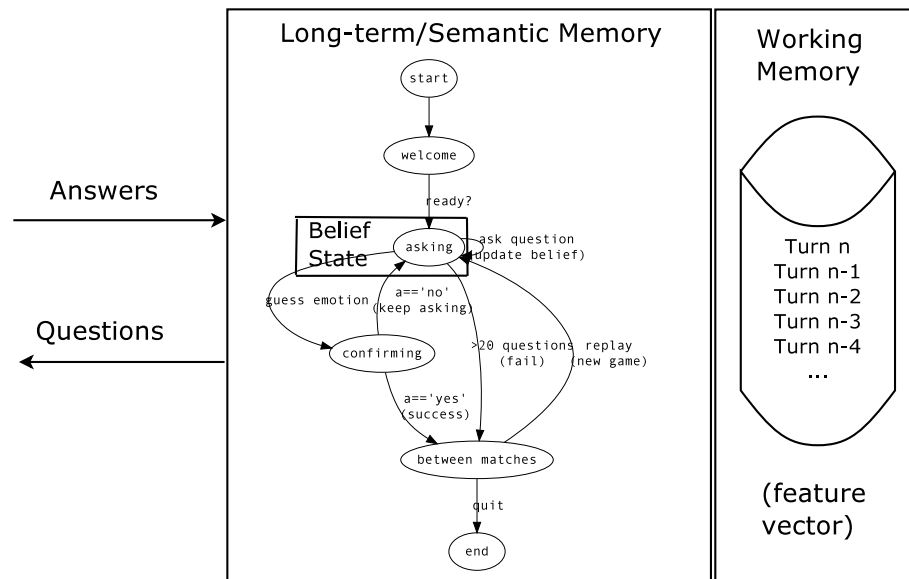
Figure 1 Dialog graph for the EMO20Q questioner agent. The loop labelled "asking" represents the functionality described by the sequential Bayesian belief model of Eq. (2) and Algorithm 1. The dialog graph is implemented as a generalized pushdown automaton, where the stack represents the agent's working memory of question–answer turns.

pushdown transducer that is not limited to only the top of the stack when determining the output and next state. This aspect is important in the question-asking loop because the stack represents the episodic memory, which stores the question–answer observations. Otherwise, the agent could be implemented as a plain pushdown transducer.

EXPERIMENTS

The EMO20Q experiments we conducted can be partitioned into human–human and human–computer experiments. 'Human-human EMO20Q' examines the data from human–human experiments, which was the initial corpus used to train the EMO20Q question-asking agent and which was the source of the agent's set of questions. 'Human-computer EMO20Q' focuses on experiments with the question-asking agent described in 'Model'.

Human-human EMO20Q

The human–human EMO20Q results are described in an earlier conference paper (*Kazemzadeh et al., 2011*) but we include a brief description because it is important for understanding the development of the automated agent.

We collected a total of 110 games from 25 players in the human–human experiments in which EMO20Q was played over text chat. The EMO20Q experiment was implemented as an online chat application using the Extensible Messaging and Presence Protocol (XMPP) and logged so that the games could 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 an emotion word that was a synonym of the word that

---

Algorithm 1 Sequential Bayesian EMO20Q agent. F is the observed question–answer features, E is the set of previously seen emotion words, $P(f|\varepsilon)$ is the semantic knowledge relating the observed question–answer pairs to emotion words, and $\beta_{t,k}$ is the belief about the emotion word indexed by k at time t. Because the agent is playing a twenty questions game, d is set to 20 and the final return statement represents the agent's last guess as the final question.

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**

    $q^{(i)} = \underset{q \in Q}{\mathrm{argmax}}\, P((q, \text{'yes'})|\varepsilon)P(\varepsilon|f_1, ..., f_{i-1})$

    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)} = $ 'yes') **then**

        Return: $e^* = e_k$

    end if

    if ($q^{(i)}$ is identity question for $e_k \wedge a^{(i)} = $ 'no') **then**

        $\beta_{i,k} \leftarrow 0$

    end if

end for

$k^* \leftarrow \underset{k \in 1...|E|}{\mathrm{argmax}}[\beta_{i,k}]$

$e^* \leftarrow e_{k^*}$

Return: most likely emotion given observations: e^*

---

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 games and kept track of both words.

Of the 110 games played between the 25 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 games, 22 terminated with synonyms. The 16 unsuccessfully terminated games consisted of several distinct cases of failure. The questioner could give up early if they had no clue (5/16), they could give up at twenty questions (1/16), they could continue past twenty questions (6/16), for example due to losing count or as a matter of pride or competitiveness, or the answerer could inadvertently give away the answer early due to a typing error or by giving an unduly generous hint (4/16).

There were 71 unique words that players chose in the human–human games, 61 of which were correctly identified. These are listed in Table 2.

PeerJ Computer Science

Table 2 Emotion words from human–human EMO20Q games. Games that terminated with words that the answerers felt to be synonyms are denoted by parentheses.

Emotions (synonyms)	Count	# correct	...	Emotions (synonyms)	Count	# correct
Admiration	1	1		Guilt	4	4
Adoration	1	0		Happiness	1	1
Affection (love)	2	2		Helplessness	1	1
Amusement	1	1		Hope (feeling lucky)	3	3
Anger	2	1		Insecurity (shyness)	1	1
Annoyance (irritated)	2	2		Jealousy (envy)	3	3
Anxiety	3	3		Joy	1	0
Apathy (uninterested)	1	1		Loneliness	1	1
Awe	1	0		Love	2	2
Boredom	2	2		Madness (anger)	1	1
Bravery	1	1		Melancholy	1	1
Calm	2	2		Pity (sympathy)	1	1
Cheerfulness	1	1		Pride	2	2
Confidence	1	1		Proud	1	1
Confusion	2	1		Regret	2	2
Contempt	1	1		Relief	5	5
Contentment (calm)	2	1		Sadness	2	2
Depression (misery)	2	2		Satisfaction	1	0
Devastation	1	0		Serenity	1	1
Disappointment	1	1		Shame	1	1
Disgust	2	2		Shock	1	1
Dread (hopelessness)	1	1		Shyness	1	1
Eagerness (determination)	1	1		Silly	1	1
Embarrassment	2	2		Soberness	1	0
Enthusiasm (eagerness)	3	1		Sorrow (sadness)	1	1
Envy (jealousy)	3	3		Stress	1	1
Exasperation	1	1		Suffering	1	0
Excitement	1	1		Surprise	3	3
Exhilaration (thrill)	1	1		Tense (uncomfortable)	1	0
Exhaustion	1	1		Terror	1	1
Fear (distress, scared)	2	2		Thankful	1	0
Frustration	2	2		Thrill (entrancement)	2	1
Fury	1	1		Tiredness	2	2
Glee	1	0		Wariness	1	0
Gratefulness	1	1		Worry (anxiety, scared)	3	3
Grumpiness	1	1		**Total**	110	94

There was a total of 1,228 question-asking events. Of the questions, 1,102 were unique (1,054 after normalizing the questions for punctuation and case). In Table 3 we list some of the questions that occurred more than once. Since the surface forms of the questions vary widely, we used manual preprocessing to annotate the questions to a logical form that is invariant to wording. This logical form converted the surface form of the player's

Kazemzadeh et al. (2016), *PeerJ Comput. Sci.*, DOI 10.7717/peerj-cs.40　13/23

Table 3 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

input to a pseudo-code language by converting the emotion names to nouns if possible, standardizing attributes of emotions, and relating emotions to particular situations and events. Examples of the standardized questions are shown in Table 4. After this semantic standardization, there were a total of 727 question types.

Human-computer EMO20Q

Using the human–human data described earlier in 'Human-human EMO20Q' and the computational model and algorithm described in 'Model', we built a computer agent to play the questioner role in EMO20Q games. The EMO20Q dialog agent was implemented

Table 4 Examples of semantic annotations for 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?

using a server-side web application that maintained the belief state and episodic buffer for each open connection. The belief state was serialized to EmotionML (*Schröder et al., 2012*; *Burkhardt et al., 2014*) and saved in a session database between each question–answer turn.

To test the proposed model of Socratic epistemology for verbal emotional intelligence, we conducted two experiments to assess the performance of the agent. The first experiment was a small pilot study of 15 subjects who each played three games against the agent (*Kazemzadeh et al., 2012*). In the pilot study, the subjects were recruited locally. Subjects were asked to pick three emotion words, one that they thought was "easy", one that was "medium", and a third that was "difficult". These difficulty ratings were described in terms of a person's maturity and vocabulary: an "easy" emotion word was one that a child could guess, whereas a "difficult" word was one that would require maturity and a sophisticated vocabulary to guess. The pilot study was designed to assess the feasibility of the agent design but did not use training data beyond the original human–human data used as bootstrap.

The second experiment was a larger experiment that forms the key experimental contribution reported by this paper. It followed the same methodology as the pilot study, but with 101 subjects recruited from Amazon Mechanical Turk. These subjects were selected to come from the United States, speak English fluently, and have high past acceptance rates as Mechanical Turkers.

In the second experiment, the parameters of the model were updated every ten subjects. Thus, there were ten batches of ten subjects, each playing three games against the automated agent, which yielded 300 games. After each ten subjects, the model described in 'Model' was updated based on the total counts of question–answer pairs for each emotion word in the corpus to that point. Thus, the human–human data, the human–computer pilot data, and the preceding data from the past batches of the Mechanical Turk experiments comprised the corpus used to train the model used in the current batch. In addition to updating the probabilities of the model's semantic knowledge (likelihoods), new emotion words were added to the vocabulary if encountered.

Kazemzadeh et al. (2016), *PeerJ Comput. Sci.*, DOI 10.7717/peerj-cs.40

15/23

Table 5 Experimental results for 15 subject pilot study (45 EMO20Q games).

| 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

RESULTS

The results of our pilot experiments on fifteen subjects are summarized in Table 5. To compare the agent's performance with human performance, we used two objective measures and one subjective measure. The success rate, shown in column two of Table 5, is an objective measure of how often the EMO20Q games 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.

In the pilot study, the agent obtained a performance of 44% successful outcomes (where the emotion word was correctly guessed). This performance was much less than in the human–human experiments, where successful outcomes occurred in 85% of EMO20Q games. However, the results indicated that this performance was due to sparsity of data. 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 games. 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. Overall, 28 new emotion words were encountered in the pilot study.

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 rated "medium" and "difficult", there was not a corresponding decrease in the perceived naturalness of the questioner agent. This led us to believe that the model could reproduce somewhat natural behavior, but that the data we had was insufficient due to the amount of out-of-vocabulary words in the medium and difficult classes, which motivated us to perform the second, larger-scale experiment with 100 players from Mechanical Turk.

In the larger scale Mechanical Turk experiment, we aimed to improve performance by retraining the model after each batch of 10 subjects. This strategy did in fact increase the successful outcome rate and reduced the length of the EMO20Q dialogs (number of questions), as can be seen from comparing Tables 5 and 6, which are visualized in Fig. 2. Across all three difficulty classes, the successful outcome rate improved. The "difficult" class had the largest relative improvement in successful outcomes, increasing from 13% to 25%, and the overall successful outcome increased from 44% to 57%. The increase in successful outcomes was significant at $p < .05$ as measured by a two-sided proportion
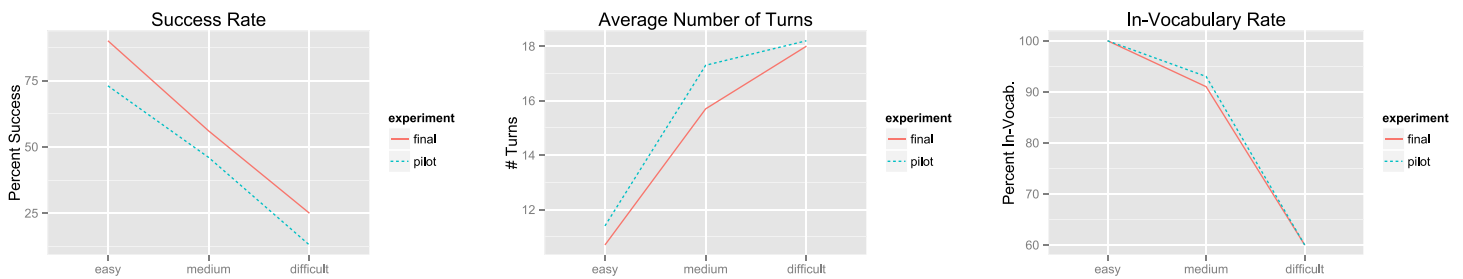
Figure 2 Results of initial automated agent pilot compared to the final experiment of 300 matches on Mechanical Turk, in which the agent was retrained every 30 matches.

Table 6 Experimental results for 100 subject Mechanical Turk study (300 EMO20Q games).

Difficulty	% success	Avg. turns	% in vocab.
Easy	90%	10.7	100%
Medium	56%	15.7	91%
Difficult	25%	18.0	60%
Total	57%	14.8	83.7%

test. The lengths of the EMO20Q dialogs decreased most for the medium difficulty class, resulting in an average of 1.6 less turns for this class, and overall the dialog length decreased from 15.6 to 14.8 turns. However, the observed differences in dialog length and in-vocabulary rate were not statistically significant. Also, the improvement from the beginning to the end of the Mechanical Turk experiment was not significant. Overall, 81 new emotion words were encountered in the Mechanical Turk study.

One surprising result was that even after collecting data from 300 EMO20Q dialogs (more than doubling the earlier human–human data), the out-of-vocabulary rate stayed nearly the same. We had expected out-of vocabulary-words to become fewer as more data had been seen. However, with each round of the Mechanical Turk experiment, we continued to receive new emotion words rather than converging to a closed vocabulary. For the Mechanical Turk experiment, we did not ask subjects about the perceived naturalness of the agent in order to save on time, and hence costs to pay the Turkers, so unfortuntately we cannot say whether the perceived naturalness increased.

Of the 101 subjects, only one was rejected, due to misunderstanding the task by choosing the words "easy", "medium", and "difficult" instead of emotion words. This level of acceptance, approximately 99% is rather high for Mechanical Turk, showing a high degree of cooperation. Several users commented that we could have paid less because the task was fun.

A listing of the words chosen by the subjects of the Mechanical Turk experiment is given in Table 7. It can be seen that there are a wide variety of words. A few (those marked by "?") were questionable in the authors' intuitions, but otherwise the words showed a high level of understanding and cooperation by the Mechanical Turkers. The three difficulty classes of words were not disjoint: some words like anger, disgust, love, and confusion spanned several categories. It can be concluded that these three difficulty levels do not form

Table 7 Observed emotion words from the Mechanical Turk experiment, by difficulty. Words that were attested but which did not fit the authors' broad intuitions are marked with '?'. The same words in multiple categories indicate that different subjects had differing opinions about difficulty.

Difficulty	Examples
Easy	Happiness, anger, sadness, calm, confusion, love, mad, hate, joy
Medium	Anger, confusion, contentment, curiosity, depression, disgust, excitement, fear, hate, irritation, love, melancholy, sorrow, surprise, tiredness, envy, outrage, elation, suffering, jealousy, nervousness, sympathy, thrill, upset, joy, anxiety, frustration, flustered, enjoyment, exhaustion, fury, bordom, delight, cold, apathy, hostility, loved, annoyance, playfulness, downtrodden, stupor, despair, pissed, nostalgia, overjoyed, indifference, courage
Difficult	Devastation, disgust, ecstasy, ennui, frustration, guilt, hope, irritation, jealousy, morose, proud, remorse, vindication, zealousness, elation, mischievous, usure, angst, patience, despise, inspired, euphoria, exuberance, worrying, melancholy, ambivalence, love, loneliness, exacerbated(?), avarice, stress, envy, disillusionment, maudlin, depression, confusion, maniacal, ambiguity, concern, pleasure, shame, indifference, anger, suicidal, pessimism, annoyance, sense of failure, educated(?), manic, overwhelmed, astounded, discontent, energetic, introspective, appalled, serenity, dissatisfaction, anxiety, lust, conflicted, perplexed, jubilance, disappointment, satisfaction, remorse, embarrassment, downcast, guilty, enamored, alienation, exotic(?), hate, caring, resentment, pity, aversion, quixotic, infuriation

precise, natural classes of emotion words, but the levels do show a trend toward a smaller basic vocabulary and a wider open vocabulary. The difficulty levels also served as a method to elicit a diverse set of words. The original human–human dialogs identified 71 unique emotion words, after the pilot study there were 99 unique emotion words, and after the large-scale Mechanical Turk experiment there were 180 unique emotion words.

DISCUSSION

The human–human EMO20Q data abounds in highly nuanced natural language descriptions of emotion. For example, one human–human EMO20Q game ended with a discussion of whether "pride" and "proud" refer to the same emotion:

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

In another human–human EMO20Q dialog, a player had difficulty answering whether "anger" was a negative emotion:

> [questioner:] so is it a negative emotion?

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

In one human–computer game, one player differentiated the emotion of loving from the emotion of being loved and another player picked the emotion "maudlin", which the authors needed to look up in a dictionary.

Given the highly nuanced, idiosyncratic descriptions in the human–human data, we were surprised at the amount of successful outcomes in the human–human EMO20Q

games and we were initially unsure whether devising an automated agent would be feasible. Although analyzing this level of detail is beyond the scope of many current systems, we saw that it is a task that humans can do with high success rates. In fact the successful outcome rates in the human–human EMO20Q games are comparable to agreement rates on emotional annotations at a much coarser level, such as labeling data with nine basic emotion labels (*Busso et al., 2008*).

The human–computer results showed us that it is possible for a computer agent to perform well at the questioner role of EMO20Q and moreover that the agent can learn new vocabulary items and improve its performance past the pilot study, which only used annotated human–human bootstrap data, by using unannotated human–computer data. The fully trained agent successfully completed 57% of the EMO20Q games, which is 67% of human–human performance and 30% better than the bootstrapped agent of the pilot study. The agent's emotion word vocabulary nearly doubled after the Mechanical Turk experiment. Normally larger emotion vocabularies would result in less agreement in annotation tasks but this showed that in the EMO20Q dialog task, vocabulary size is not a weakness but rather a strength. Even when the agent fails to guess the human opponent's emotion word in the EMO20Q game, the agent's behavior of searching for knowledge makes it appear human-like, which enables the agent maintain user engagement and learn from new, out-of-vocabulary words.

Other methodologies could be used for the EMO20Q agent and decision trees would seem to have been a natural choice. However, we saw the potential pitfall of decision trees, in their default formulation, to get stuck on the wrong branch of a tree if a subject answered a question in a non-standard way. In contrast, as our formulation of sequential Bayes extends naïve Bayes to deal with sequential observation of features, decision trees would similarly need to be updated to deal with sequentially observed features. We dealt with this issue in our methodology by smoothing the counts of the sequential Bayes model and by dynamically choosing the next question after each dialog turn. The sequential Bayes model proved capable of dealing with variability in users' answers and providing natural behavior in the case of out of vocabulary words. Other methodologies for an EMO20Q agent would need to deal with these challenges.

The ground truth issue involved in annotating recorded data with descriptive labels is a challenge that the Socratic epistemology can shed light upon. The traditional annotation task seeks to have human annotators assign one of a number of labels to data. In the case of emotion research, usually the labels are a controlled vocabulary of several emotion descriptors, like "angry", "happy", "sad", "disgusted", "fearful", "surprised", and "neutral". The problem with this approach is that these labels often do not fit realistic emotional data. Theoretically, our approach addresses the issue of ground truth in the annotation task with the notion of epistemology, which frames the issue as justification of belief rather than ground truth. Practically, our approach addresses the issue of non-prototypical emotions by enabling a more nuanced representation where the description is not a small, closed set of alternatives but rather an interactive process of communication over a large, open set of natural language descriptions. Though this more nuanced view brings with it new

challenges, we have shown that the design of an intelligent dialog agent is a feasible way of dealing with these challenges.

We plan to continue this research in several ways. First, we hope to see the effect of modality on how people describe emotions in natural language. The current work was limited to text-based chat, so the paralinguistic data that may help to convey emotional information was minimized. Including audio and video data may allow greater convergence of the players to agree upon the unknown emotion in EMO20Q. Another area of future research will be to model the answerer role. The current research focused on the questioner role, but the answerer role will offer additional challenges and insights. In particular, automating the answerer role will require more robust natural language understanding because it will need to process new, unseen questions from users, whereas the questioner used a fixed set of questions and only had to process answers to yes/no questions. The answerer would also likely require a different model than the Socratic, question-asking model presented in this paper. A successful answerer agent would allow a pleasing closed-loop simulation where both roles of EMO20Q are played by computer. There are also further areas to explore for the questioner agent, in particular, the criterion for choosing each question. We also think that this approach can improve emotion annotation and other annotation tasks, such as coding behavioral data for psychological assessment. In these tasks human annotators are asked to label data using a controlled vocabulary of words and agreement is established statistically between isolated annotators. However, we have shown that humans are able to communicate with high accuracy using a large, subjective vocabulary and we think that allowing natural language descriptions in an interactive, question-asking setting will allow for more accurate and less constrained annotations. Moreover, using a similar agent to elicit and identify items in open-ended vocabularies in other domains is another area for future work. We have found that in the case of emotion words, the EMO20Q agent was able to elicit a growing vocabulary and successfully identify items from the vocabulary, which could be possible for other domains. For example, eliciting "things users are passionate about" or medical illnesses are possible open-ended domains that may be hard to enumerate in an a priori exhaustive list, but which might be identified by a similar dialog agent.

CONCLUSION

The main goals of this paper were to formulate a theoretical and computational model for a subset of human emotional language. We called this model the Socratic epistemology for verbal emotional intelligence because it uses question-asking to justify beliefs about emotions in a natural language dialog context. We presented the emotion twenty questions (EMO20Q) game and showed that the level of human performance was high despite not limiting the players to any predefined emotion vocabulary. We also presented an automated agent that can play the question-asking role of EMO20Q. This agent uses a sequential Bayesian belief update algorithm to simulate a cognitive process by which the agent updates its belief state of candidate emotion words over time. This framework was inspired by a method of question-asking that was proposed by the ancient philosopher Socrates and the field of epistemology:

[**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 ready to be refuted as to refute (Plato, *Gorgias*, 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 by making reference to the theory of physiological humors. Humankind's 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, but our conception of emotion is still largely expressible through language.

In the second quote, Socrates (as quoted by Plato) cross-examines Gorgias to determine Gorgias' beliefs. Socrates applied his method of question-asking to understand beliefs about complex abstract concepts that were disputed in ancient times. Two millennia later we have used a computational implementation of this method to make a dialog agent better understand human beliefs about emotional concepts.

We have provided an anonymized version of data we gathered from EMO20Q, source code for the experiments, demos, and other resources at http://sail.usc.edu/emo20q.

ADDITIONAL INFORMATION AND DECLARATIONS

Author Contributions

- Abe Kazemzadeh conceived and designed the experiments, performed the experiments, analyzed the data, contributed reagents/materials/analysis tools, wrote the paper, prepared figures and/or tables, performed the computation work, reviewed drafts of the paper.
- James Gibson conceived and designed the experiments, performed the experiments, analyzed the data, contributed reagents/materials/analysis tools, wrote the paper, prepared figures and/or tables, reviewed drafts of the paper.
- Panayiotis Georgiou, Sungbok Lee and Shrikanth Narayanan conceived and designed the experiments, analyzed the data, contributed reagents/materials/analysis tools, wrote the paper, prepared figures and/or tables, reviewed drafts of the paper.

Data Availability

The following information was supplied regarding data availability:

Emotion Twenty Questions: http://sail.usc.edu/emo20q/

Github: https://github.com/abecode/emotion-twenty-questions.

REFERENCES

Allauzen C, Riley M. 2012. A pushdown transducer extension for the openfst library. In: *Proceedings of the conference on implementation and application of automata.*

Baddeley AD. 2000. The episodic buffer: a new component of working memory? *Trends in Cognitive Science* **4(11)**:417–423 DOI 10.1016/S1364-6613(00)01538-2.

Barrett LF. 2006. Are emotions natural kinds? *Perspectives on Psychological Science* **1(1)**:28–58 DOI 10.1111/j.1745-6916.2006.00003.x.

Burkhardt F, Becker-Asano C, Begoli E, Cowie R, Fobe G, Gebhard P, Kazemzadeh A, Steiner I, Llewellyn T. 2014. Application of emotionML. In: *Emotion, social signals, sentiment, and linked open data (ES3LOD) 2014.* Reykjavik: European Language Resources Association, 1–5.

Busso C, Bulut M, Lee C-C, Kazemzadeh A, Mower E, Kim S, Chang J, Lee S, Narayanan S. 2008. IEMOCAP: interactive emotional dyadic motion capture database. *Journal of Language Resources and Evaluation* **42(4)**:335–359 DOI 10.1007/s10579-008-9076-6.

Carofiglio V, De Rosis F, Novielli N. 2009. Cognitive emotion modeling in natural language communication. In: Tao J, Tan T, eds. *Affective information processing.* London: Springer, 23–44.

Eco U, Sebeok TA (eds.) 1988. The sign of three: Dupin, Holmes, Peirce. *Advances in semiotics.* Bloomington: Indiana University Press.

Frazier BN, Gelman SA, Wellman HM. 2009. Preschoolers' search for explanatory information within adult-child conversation. *Child Development* **80(6)**:1592–1611 DOI 10.1111/j.1467-8624.2009.01356.x.

Griffiths TL, Kemp C, Tenenbaum JB. 2008. Sun R, ed. *The Cambridge handbook of computational psychology.* Cambridge: Cambridge University Press, 99–101.

Hintikka J. 2007. *Socratic epistemology: explorations of knowledge-seeking by questioning.* Cambridge: Cambridge University Press.

Hockett CF, Altmann S. 1968. Sebeok TA, ed. *Animal communication: Techniques of study and results of research.* Bloomington: Indiana University Press, 61–72.

Howe J. 2006. The rise of crowdsourcing. *Wired Magazine* **14**(6).

Kamvar S, Harris J. 2009. *We feel fine: an almanac of human emotion*. New York: Scribner.

Kazemzadeh A. 2013. Natural language description of emotion. PhD thesis, University of Southern California.

Kazemzadeh A, Georgiou PG, Lee S, Narayanan S. 2011. Emotion twenty questions: toward a crowd-sourced theory of emotions. In: *Proceedings of ACII'11*.

Kazemzadeh A, Gibson J, Li J, Lee S, Georgiou PG, Narayanan S. 2012. A sequential Bayesian agent for computational ethnography. In: *Proceedings of interspeech*.

King B. 1989. The conceptual structure of emotional experience in Chinese. PhD thesis, Ohio State University.

Kövecses Z. 2000. *Metaphor and emotion: language, culture, and body in human feeling*. Cambridge: Cambridge University Press.

Mihalcea R, Liu H. 2006. A corpus-based approach to finding happiness. In: *AAAI spring symposium on computational approaches to weblogs. Available at http://www.cse.unt.edu/~rada/papers/mihalcea.aaaiss06.pdf*.

Mower E, Metallinou A, Lee C-C, Kazemzadeh A, Busso C, Lee S, Narayanan S. 2009. Interpreting ambiguous emotional expressions. In: *ACII special session: recognition of non-prototypical emotion from speech—the final frontier?* Amsterdam: IEEE, 1–8.

Peirce CS. 1868. Some consequences of four incapacities. *Journal of Speculative Philosophy* **2**:140–157 *Available at http://www.iupui.edu/~peirce/writings/v2/w2/w2_22/v2_22.htm*.

Rolls ET. 2005. What are emotions, why do we have emotions, and what is their computational basis in the brain. In: Fellous J-M, Arbib MA, eds. *Who needs emotions? The brain meets the robot*. Oxford: Oxford University Press, 117–146.

Russell B. 1905. On denoting. *Mind* **14**:479–493 DOI 10.1093/mind/XIV.4.479.

Schröder M, Baggia P, Burkhardt F, Pelachaud C, Peter C, Zovato E. 2012. W3C candidate recommendation: emotion markup language (EmotionML) 1.0. *Available at http://www.w3.org/TR/emotionml/*.

Shaver PR, Murdaya U, Fraley RC. 2001. Structure of the Indonesian emotion lexicon. *Asian Journal of Social Psychology* **4**:201–224 DOI 10.1111/1467-839X.00086.

Steyvers M, Griffiths TL, Dennis S. 2006. Probabilistic inference in human semantic memory. *Trends in Cognitive Sciences* **10**(7):327–334 DOI 10.1016/j.tics.2006.05.005.

Von Ahn L, Dabbish L. 2004. Labeling images with a computer game. In: *Proceedings of the SIGCHI conference on Human factors in computing systems*.

Xu F, Tenenbaum JB. 2005. Word learning as bayesian inference: evidence from preschoolers. In: *Proceedings of the twenty-seventh annual conference of the cognitive science society*.

Zhong N, Liu J, Yao YY, Ohsuga S. 2000. Web intelligence. In: *Computer software and applications conference*.