I Know What You Don't Know: Proactive Learning through Targeted Human Interaction

Socially Interactive Agents Track

Abdelwahab Bourai
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA
abourai@cs.cmu.edu

Jaime Carbonell
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA
jgc@cs.cmu.edu

ABSTRACT

Humans communicate extensively through "meta-information" encoded in emitted non-verbal signals. This meta-information not only allows us to analyze an individual's external emotional state but also certain internal states. For example, humans are able to learn from others thanks to their ability to determine their most knowledgeable peers in a given domain through their interactions with these individuals. As autonomous agents expand into more socially oriented tasks, they must capture and reason through these emitted cues to better understand their human counterparts. In this work, we conduct two experiments. First, we train a model to predict the knowledgeability of speakers using non-verbal features. Next we simulate the process of selecting the most knowledgeable person in a given domain using a proactive learning approach. The results indicate our agent is capable of observing human behavior and using this information to select a specific human for aid on a given question.

CCS CONCEPTS

• **Theory of computation → Active learning**; • **Applied computing → Psychology**; • **Information systems → Sentiment analysis**;

KEYWORDS

Social Agents; Affect; Nonverbal Behavior Understanding; Active Learning

ACM Reference Format:
Abdelwahab Bourai and Jaime Carbonell. 2018. I Know What You Don't Know: Proactive Learning through Targeted Human Interaction. In *Proc. of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2018), Stockholm, Sweden, July 10–15, 2018,* IFAAMAS, 8 pages.

1 INTRODUCTION

Humans learn extensively through observing their surroundings, and learning from other humans specifically is a core aspect of mental development [29]. For example, children learn at a young age to avoid information from unreliable or "inaccurate" people [15]. Humans are able to infer how knowledgeable one is in a domain thanks to their ability to decode emitted non-verbal signals during their interactions with others [16]. This can be done in two ways: if the answer is known to the questioner, then logically they can infer knowledgeability from the correctness of the reply. However, if the answer is not shared between two people, then the questioner is forced to guess based on the non-verbal cues the speaker emits [5]. This process is very noisy, as the speaker may be anxious, stalling to remember a fact, or the questioner may miss subtle signals. Many studies define this internal representation of one's own knowledge as the "feeling of knowing", or FOK [12].

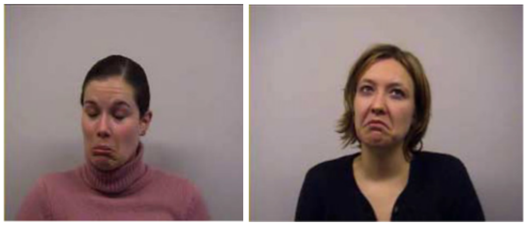

Figure 1: Examples from Swerts and Krahmer's study on Feeling of Knowing (FOK) [28]. Here, speakers are indicating a lack of FOK in response to a question. Certain audio-visual signals were found to be strong indicators of knowledgeability.

This level of interaction and cooperation has not yet been realized in human-computer communication and interaction. Machines have no ability to accurately identify if a data point came from a "knowledgeable" human or from one lacking expertise in a domain. Embedding this level of understanding in a machine learning process is a complex undertaking. It requires modeling a counterpart's "knowledgeability" as a phenomenon that can be identified through a set of non-verbal features. These knowledgeability scores would then need to be matched with topics to create a model that can be used for future tasks where an expert human is needed.

One interesting segment of machine learning that deals with this issue is active learning. In active learning, an agent will work in conjunction with an oracle to classify unlabeled data points, selecting the example it considers most informative for the oracle to label [25]. An extension of this is proactive learning, which seeks to relax certain active learning assumptions [10]. In active learning, there are certain problematic assumptions such as the fact that the annotators are always reliable, always correct, and are cost-insensitive. However, in real-world applications, humans are much more fickle. In Koenig et al's study, children were able

Proc. of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2018), M. Dastani, G. Sukthankar, E. André, S. Koenig (eds.), July 10–15, 2018, Stockholm, Sweden.

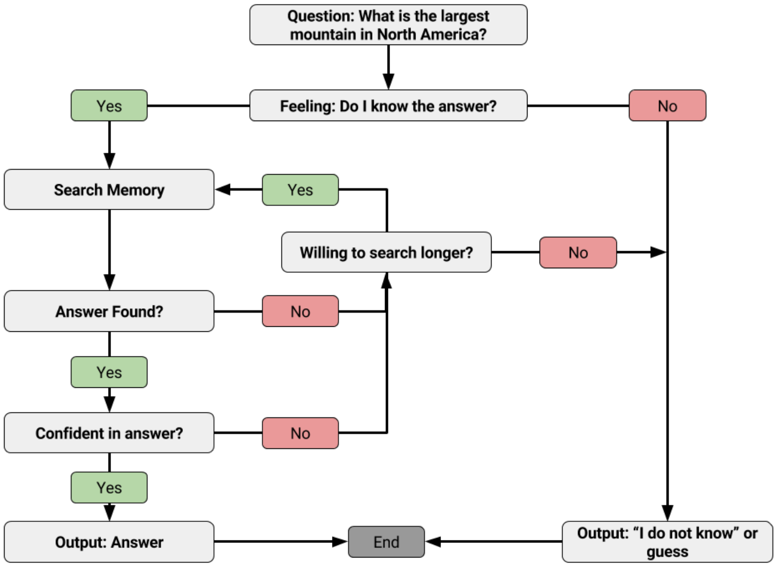
Question: What is the largest mountain in North America?
Feeling: Do I know the answer?
Yes
No
Search Memory
Yes
Willing to search longer?
No
Answer Found?
No
Yes
Confident in answer?
No
Yes
Output: Answer
End
Output: "I do not know" or guess

Figure 2: The process by which humans answer a question from memory is detailed in the above diagram adapted from Nelson et al [20]. In cases where someone is certain they know or do not know a question this process will terminate quickly. However, if they were to search their memory for longer periods of time they may emit involuntary non-verbal signals during this search.

to realize which humans to trust answers from and which humans to ignore, but an active learning agent would have no ability to discriminate between ignorant and knowledgeable oracles [15]. Proactive learning allows an agent to select the optimal oracle based on the likelihood they will label the data point correctly. Donmez and Carbonell were the first to model this by experimenting with scenarios that simulated unreliable, expensive, or reluctant humans [10].

As social agents become increasingly ubiquitous, embedding this ability to analyze complex affective states such as knowledgeability is imperative for agents to successfully interact with humans. Past work in robot tutors has shown that even simple personalization yielded benefits [17]. Having the ability to discern knowledgeability of students through interactions with these robot tutors allows the robot to have a way to immediately recognize a student's understanding of a concept and modify its teaching strategies accordingly. In addition, social agents can learn how to enlist the aid of specific humans on difficult tasks. For example, recent work investigated how robots would seek help from humans in navigation-based tasks [24]. Learning agents such as the Never Ending Language Learning (NELL) system would be able to utilize human interaction for knowledge acquisition as well as textual data on the internet [6].

We propose the following contributions:

(1) As there is no public dataset available from past Feeling of Knowing studies, a novel dataset was recorded to replicate past feeling of knowing studies as closely as possible using video interviews
(2) A proactive learning agent that is capable of modeling its human counterparts' knowledgeability in certain topics and reaching out to the correct "experts" for an answer to a given question
(3) An empirical evaluation of the agent's performance

2 PAST WORK

We divide the past work section into three parts. First, we discuss Feeling of Knowing and the psychological experiments we emulate. We then go over previous work in predicting knowledgeability and uncertainty. Finally, we go over related work from the proactive learning domain.

2.1 Feeling of Knowing and Predicting Knowledgeability

Feeling of knowing can be described as "knowledge of one's knowledge" [21]. When asked a question, humans will quickly determine whether or not they should search their memory for an answer [23]. Once they decide they may know the answer, a search is initiated through their memory [20]. When a human answers a question they have a somewhat robust understanding of, they will usually answer quickly and confidently. However, if they misjudged their initial feeling of knowing, they may emit certain social signals that indicate uncertainty as they search their memory for a suitable answer [27] [20]. A diagram showing this process can be found in Figure 2. Smith and Clark describe these emitted signals as a way to "save face". They set up a study where participants were asked 40 factual questions and gave back spoken answers, and then immediately asked how high their FOK was. Their results showed that participants indicated their FOK was lower after incorrect answers. They found that uncertainty in answers can be detected from signals such as rising intonation and hedge words such as "I guess" as well as the latency of the response [27]. Brennan and Williams extended Smith and Clark's work by instructing participants to listen to a speaker answer a question and rate their knowledgeability [5]. Their results indicate that listeners could accurately predict the knowledgeability of speakers based on similar social signals from Smith and Clark's work such as latencies and rising intonations. Brennan and Williams defined this as the "Feeling of Another's Knowing" (FOAK). We utilize findings that listeners could accurately predict the internal knowledgeability of speakers solely through affective signals to train an agent that can similarly take advantage of these affective signals.

Swerts and Krahmer completed a similar experiment to Brennan and Williams' but also presented video of the speakers rather than speech alone to the participants. They found that including visual and auditory signals boosted participants' FOAK over auditory or visual signals alone [28]. They also corroborated Smith and Clark's findings that FOK is closely correlated with the correctness of the answer (i.e. lower FOK answer is likely incorrect) [28]. They found certain visual features that are significantly correlated to FOK scores as well as the auditory signals previously mentioned by Smith and Clark. Raising the eyebrows, which Bolinger indicated as the visual counterpart to a rising intonation [3], was indicative of lower FOK scores. Other visual features they investigated were "funny faces", smiles, and gaze acts which were also found to be indicative of lower FOK [28]. An example image from their study is shown in Figure 1.

2.2 Automated Prediction

Most of the studies above involved manual annotation and detection of these audiovisual signals. A study done by Bourai et al investigated automated prediction of knowledgeability using emitted non-verbal signals [4]. They collected clips from a trivia show and analyzed predictive facial and speech features. In addition they trained a model to predict the correctness of a speaker above human performance. Pon-Barry et al trained a model to detect uncertainty in speech using features such as pitch, intensity, etc. [22]. Another approach for certainty detection was a multimodal method using EEG and other physical signals [13]. Prediction of valence, arousal, and other emotional states has also been investigated [7][14].

In our knowledgeability prediction approach, we design a study similar to that of Smith and Clark but we also train a model to predict knowledgeability in a similar fashion to [4]. We use a robust representation of visual features using Ekman's Facial Action Coding System (FACS) rather than subjective features such as "funny face" [11]. Bourai et al's study did not focus on linguistic features, instead relying solely on non-verbal features such as speech patterns and facial activations. We take a similar approach but also include certain linguistic features. Finally, we train a model to act as the "listeners" from the above studies to allow an autonomous agent to determine the knowledgeability of human agents. In this study we use the correctness of an answer as proxy for FOK as past studies found a strong correlation between the two [28][27].

2.3 Proactive Learning

An active learning agent works in conjunction with an all-knowing oracle to select the most informative unlabeled data points and asks the oracle for a label. This way, it attempts to maximize accuracy of labeling [25]. Proactive learning is an extension of active learning that attempts to deal with the underlying assumptions of active learning, namely that the oracle is considered to be reliable (always answers), infallible (always right), individual, and insensitive to cost. In real world scenarios where we have multiple oracles, each with differing competencies, active learning would suffer. Thus, proactive learning takes a decision-theoretic approach where we jointly select the optimal example-oracle pair for a given question to improve our learning algorithm while keeping costs low. The utility equation used to select this optimal pairing is defined as

$$U(x,k) = \frac{P(ans|x,k) * V(x)}{C_k}$$

where U is the utility of a labeler $k \in K$ given an unlabeled datapoint x. The cost of each labeler is represented with C_k as a penalty on our utility. $V(x)$ is an active selection criterion that measures the value of an example x. In [10] they select the density weighted uncertainty scoring metric they developed in [9]. We propose a differing scoring value function in a later section.

Donmez and Carbonell showcased the superiority of a proactive learning agent that is capable of jointly optimizing for both the optimal oracle and labeler under three scenarios. The first scenario involves a reliable labeler (always answers) and a reluctant labeler (occasionally does not answer). The second scenario where proactive learning is extremely valuable is when labelers have differing levels of knowledgeability given a query. The third scenario involves oracles with differing costs. For the purposes of this study, the second scenario is the most relevant.

In Donmez and Carbonell's study, they acknowledged that there is no real-world ground truth for the reliability of a labeler in a certain domain ($P(ans|x,k)$) so they used simulated reliability data [10]. However, as mentioned in previous sections, humans have an innate ability to infer reliability and personalities of others in situations ranging from choosing the right lawyer for a case to choosing a partner in marriage [3]. Furthermore, using cues such as audiovisual prosody, humans can determine the "Feeling of Knowing" (FOK) of a speaker in response to a question [28][5]. Thus we seek to focus on the $P(ans|x,k)$ aspect of the proactive learning problem and allow a learning agent to infer human knowledgeability. To do this, we use the outputs of our trained knowledgeabilty model to determine $P(ans|x,k)$ for a given human k as they answer a question.

3 DATASET

We describe the two datasets used for this project. A dataset derived from the BBC's *University Challenge* trivia show was obtained from Bourai et al's previous study on knowledgeability detection [4]. The other dataset was a series of video interviews we conducted with participants answering a standardized set of questions.

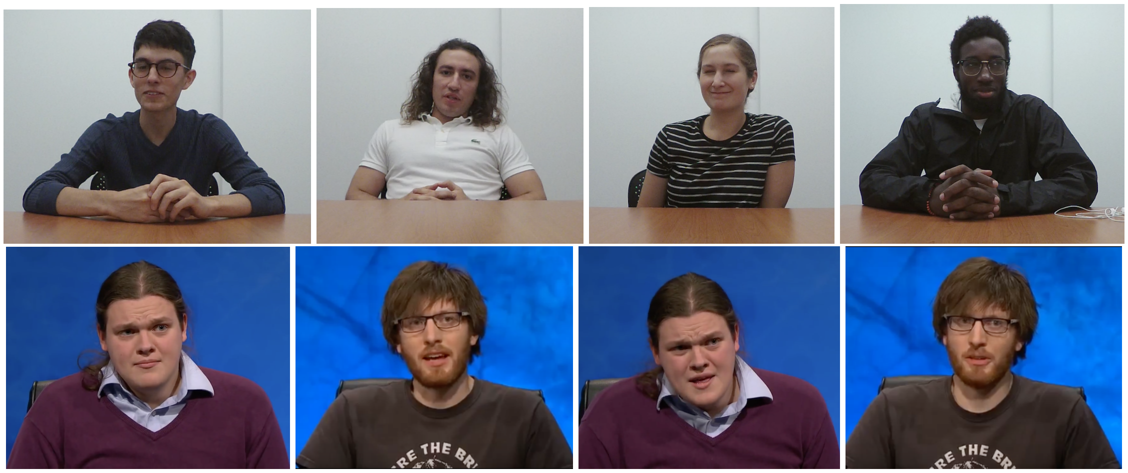

Figure 3: The top row contains answers from the video interview dataset and the bottom row contains answers from the trivia dataset. The first two images in each row contain correct answers while the latter two contain incorrect answers. Note how in the trivia dataset the speakers eye gaze is always fixated on the moderator.

3.1 Collection Methodology

3.1.1 Trivia Setting. The dataset consists of 198 clips from the BBC's *University Challenge* series ranging from 1 to 3 seconds. Each data point contains a contestant answering a question from the moderator. Each answer clip was annotated with either "correct" or "incorrect" labels, based on the moderator's feedback, and then cropped into individual video clips using the `ffmpeg` command line utility tool. An answer is defined as as from the end of the question being posed until the participant completes their answer. Example annotations can be seen in Figure 3. Audio was extracted from the videos using `ffmpeg`. All clips were derived and annotated using the ELAN annotation software [26].

137 clips (69.1%) contain a male participant. The majority of participants in the dataset are college-aged Europeans ranging from 17 to 22 years old there are clips with older individuals. The clips are all in high-definition with many direct camera angles on participants faces as they answer the question.

3.1.2 Interview Setting. Nine subjects were recruited, with eight of them being male and all college-aged students. 40 questions were drafted from four categories: US Presidents, Literature, Sports, and Geography. We attempted to choose as general topics and questions as possible to avoid having any one participant be particularly excellent at that domain (e.g. a section on rock band history would skew heavily towards a very small population). Participants were instructed to answer to the best of their abilities but were not forced to make a guess. Thus, some answer clips contain the participant directly stating their lack of knowledgeability with "I do not know". In addition, no time limit was set for answering a question.

Annotations were also created using the ELAN annotation software. After marking answer boundaries in the interview, each answer clip was tagged with either a "correct" or "incorrect" label. The answer boundaries are identical to those in the trivia dataset. 360 clips were derived from these interviews. From each clip, we can extract useful prosodic features such as speech rate, pitch, and phonation time using Praat [2]. In addition, we extract visual features such as eye gaze, head pose, and facial action unit activations using OpenFace [1].

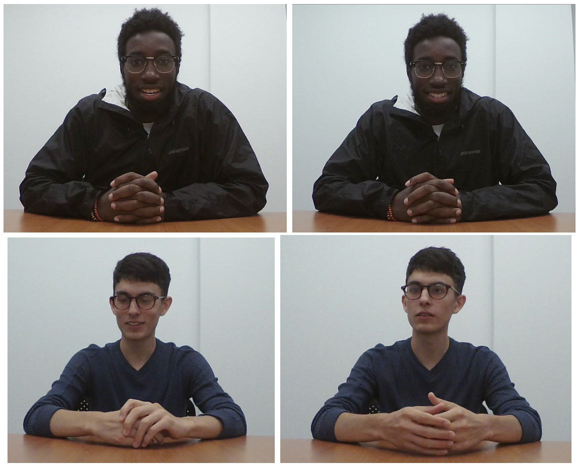

Figure 4: The images on the left contain a speaker indicating they do not know an answer, while the images on the right are a correct answer. Some people indicated they did not know an answer confidently (top left) while others were more expressive (bottom left). Smile events can also be found in both incorrect (bottom left) and correct answers (top right).

3.2 Dataset Comparison

Past work on knowledgeability recognition used clips from the trivia dataset [4]. However, we found some limitations with this dataset. The high-stakes nature of a trivia show may make participants more expressive or anxious and their eyes are usually fixed on the moderator rather than the camera. In addition, certain

Table 1: Significant Audiovisual Features: Trivia vs Interview

Trivia Dataset	Interview Dataset
AU 10 (Upper Lip Raiser)	**AU 6 (Cheek Raiser)**
AU 15 (Lip Corner Depressor)	AU 12 (Lip Corner Puller)
AU 17 (Chin Raiser)	AU 23 (Lip Tightener)
AU 6 (Cheek Raiser)	Answer Duration (latency)
AU 7 (Lid Tightener)	Pitch Range
Pitch Slope	**Pitch Slope**
Speech Rate	-
Number of Pauses	-

features that were found to be strongly correlated to FOK such as latency could not be used due to the artificial time-limit imposed on answers by the game show. Thus, we attempted to replicate the past psychological studies on FOK by recording our own video interviews with participants. A side by side comparison of clips from the two datasets can be seen in Figure 3.

As the datasets from the Swerts and Krahmer and Smith and Clark studies are not publicly available, we cannot directly compare them with ours. However, as we mentioned in the past work section we attempted to replicate their collection procedure as closely as possible by asking all participants the same questions. In the trivia dataset, each answer clip contained a unique question, thus it would not be suitable for our proactive learning experiments.

4 AUTOMATICALLY DETERMINING KNOWLEDGEABILITY

We begin by analyzing which audiovisual cues are most relevant to predicting the correctness of a speaker. We compare the features found in the interview dataset with those found to be relevant under the trivia scenario. Finally we train a model to predict the correctness of a speaker based on these audiovisual features. In our feature analysis section we focus exclusively on the interview dataset.

4.1 Feature Analysis

In past work done by Bourai et al [4] and Swerts et al [28], a combination of audio and visual features were found to be significantly discriminatory for determining correctness. We derived these features using a paired t-test and chose all features with $p < 0.05$. Similar to these past studies, we compare means of features.

Unlike Swerts and Kramer's findings that “funny faces" were indicative of lower FOK, we instead find more granular features to be predictive such as a raising of the cheeks (AU 6) or tightening of the lips (AU 23). The use of more explicit markers such as facial action units allows for better generalization of features as we do not have to rely on a subjective interpretation of what a “funny face" is.

Smile events are particularly interesting as they were singled out by Swerts and Krahmer as a particularly ambiguous event. They found that while smiles were correlated with lower FOK in participants, they were not a statistically significant feature. Their reasoning was some speakers would smile when given an easy answer while others would smile for extremely difficult questions [28]. We did observe smile events in both cases in our dataset, but with the majority of smile events occurring when a participant took a seemingly random guess or simply said “I do not know". Our sample size is smaller than the Swerts study (9 participants vs 20) thus we cannot definitively conclude that smiles are a significant indicator of incorrectness.

Smith and Clark considered answer latency to be highly correlated with FOK, with larger latencies indicative of lower FOK and usually associated with hedge words such as “uh", “um", etc [27]. We assumed the total duration of an answer clip as a measure of answer latency, as all of these questions have short answers. Our analysis indicates latency is also a strong indicator of knowledgeability. The rising intonation, indicated through the pitch slope feature in Table 1, was also found to be a strong indicator of FOK in both Smith and Clark as well as Swerts and Krahmer's studies [27] [28]. However, in our analyses the pitch range was a more indicative feature.

All of the above features are non-verbal cues automatically extracted from our clips. One verbal feature we extracted is a verbal indication of uncertainty, or “I do not know". This feature was added for our proactive learning experiments; if a participant says they do not know an answer then their knowledgeability should be considered lower. However, participants can seem confident when they are saying they do not know an answer.

4.2 Trivia vs Interview Setting Features

A comparison of the significant features found in the trivia and interview datasets can be observed in Table 1. The trivia setting may have made the speakers more anxious and expressive, as visual features were not as prevalent in the interview setting. The speech rate feature was the most significant indicator in the trivia setting, yet was not found useful in the interview setting. However this feature and the number of pauses are similar to the answer duration feature as proxies for latency. Smile events are very strong indicators of incorrectness in the interview dataset but they can be found in correct answers, as shown in Figure 4.

An important factor to consider when observing the differences in feature sets is the trivia and interview datasets contain participants from different cultures. The former contains British participants while the latter is mostly American participants. Past research in cultural effects on emotion has found that there may be non-verbal “accents" specific to certain cultures' emotional expressions [18]. Further work is needed to determine the effect, if any, this may have on our analyses.

4.3 Knowledgeability Model

We use audio and visual features extracted using Praat and OpenFace to predict whether or not a given person is correct. As the features may all be in different scales (i.e. eye gaze and head pose are in 3-D space while Action Units are rated on a continuous scale between 0 and 5 for intensity of observation) we first normalize each feature dimension by subtracting the mean and dividing by twice the standard deviation. We then save these mean and standard deviations from the training set and apply them to the testing set.

We then train a Support Vector Machine (SVM) on visual, audio, and an audiovisual features. [8]. A stratified 5-fold cross-validation is used for hyperparameter tuning due to a slight class imbalance. We validate C in the range of 10^{-5} and 10^{5}.

We also train a two layer neural network on the combined audio and visual feature sets. Both layers' hidden sizes are set to 20 and learning rate is set to 0.001.

In order to ensure the models were not learning person-specific features, the dataset is grouped into person-independent batches, giving us nine total batches with 40 answer clips in each batch from the individuals. We then use a 9-fold testing approach, where one batch is held out for testing and the remaining eight used for training and validation. This allows us to make stronger conclusions about how well the model generalizes across unique samples. Feature selection is completed for each fold by computing ANOVA F-values and selecting the K (hyperparameter) most significant features. We ensure that feature selection was run only on the training set to avoid overfitting. We did not directly use the features we found most predictive in Section 4.1.

4.3.1 Baselines. A majority-class classifier was used to establish a chance baseline. We also compare to a K-Nearest Neighbors model with $k = 2$ to establish a simple baseline.

5 PROACTIVE LEARNING MODEL

We make a few changes to the proactive learning model defined in [10]. Rather than using simulated oracle reliability data, we instead derive $P(ans|x, k)$ from our knowledgeability model. Given a question x and oracle k, the SVM model will return the probability of the answer being correct.

We also modify $V(x)$. In Donmez's proactive learning study, they first group the data into C clusters such that each cluster $c \in C$ has a centroid x_c [10]. In a later study, Moon and Carbonell derived a function to measure multi-class information density [19] that also relies on clustered data with clear centroids. However, we cannot easily cluster our data in this manner as each data point is a question belonging to one of four categories (Geography, US Presidents, Literature, Sports) and there is no easy way to measure similarity between questions in a category. Thus, it is difficult to use a density-based measure.

Since we cannot judge a certain question within a cluster to be more or less informative than another, we instead try to choose the optimal cluster. Once we know the proper cluster, we randomly sample an unlabeled question from within that cluster. We define

$$V(x) = \frac{|c_{UL}|}{|c|} + \text{Mean}_{\text{accuracy}} + \frac{\sigma_{\text{accuracy}}}{\sqrt{N}}$$

where c is the cluster x belongs to and $\frac{|c_{UL}|}{|c|}$ measures the percentage of unlabeled examples within that cluster. The mean and standard deviation of accuracies are determined from the labeled set so far. Recall that our goal in active learning is maximizing classification accuracy while minimizing labeling effort. Thus we wish to select an example x from a cluster that will most improve our model's classification accuracy. We choose the cluster with the highest mean accuracy so far but also one with the highest standard deviation. This indicates that we may have not yet converged to the correct human labelers as there is some variance in our results. However, we also want to add a diversity element to our sampling so the model does not drift towards only labeling clusters that are highly accurate already. We thus add the $\frac{|c_{UL}|}{|c|}$ component which rewards clusters that have not been explored as much by our model. We assume uniform costs across all models, making this method cost-insensitive. A depiction of the algorithm can be found under Algorithm 1.

Table 2: Mean Accuracy Across Folds

Model	Accuracy
Majority Class	0.541
K-Nearest Neighbors	0.627
Neural Network (Audiovisual)	0.703
SVM (Visual Features)	0.540
SVM (Audio Features)	0.686
SVM (Audiovisual Features)	**0.705**
SVM (Audiovisual Features + "I Don't Know")	**0.801**

Algorithm 1 Proactive Learning with Multiple Fallible Oracles

```
Input: classifier $f$, labeled data $L$, unlabeled data $UL$, $k$ oracles
Output: $f$
  procedure TRAIN($L, UL$)
  while$|UL| \neq 0$
(1) $\forall k \in K, x \in UL$ calculate $U(x, k)$
(2) Choose $k* = \text{argmax}_{k \in K} \text{max}_{x \in UL} \{U(x, k)\}$
(3) Choose $x* = \text{argmax}_{x \in UL} \{U(x, k*)\}$
(4) Update $L = L \cup (x*, y*), UL = UL/(x*, y*)$
```

5.1 Empirical Evaluation

The goal of our experiment is to see if the proactive learning agent can pinpoint which human oracles are most likely to respond with a correct answer given a question. We begin each run by randomly selecting three questions from each category and then randomly selecting participants to answer these questions. This will be our initial labeled set L, with the remaining questions considered to be the unlabeled set UL.

At each iteration, the agent selects the question-person pair most likely to increase our accuracy based on $U(x, k)$. It uses the outputs from the knowledgeability prediction model to measure the correctness of a participant answering the question. Once we have $x*$ and $k*$, we add the label $(x*, y*)$ to our labeled set, where $y*$ is the answer provided by person $k*$. The agent continues sampling optimal question-oracle pairs until the unlabeled set is empty. At each iteration, we measure the classification error of the labeled set we are creating. As we began by randomly selecting oracles to answer questions, our initial error will be about the same as random choice.

6 RESULTS

In this section we go over the results of our knowledgeability prediction model and the empirical evaluation results of our proactive learning model.

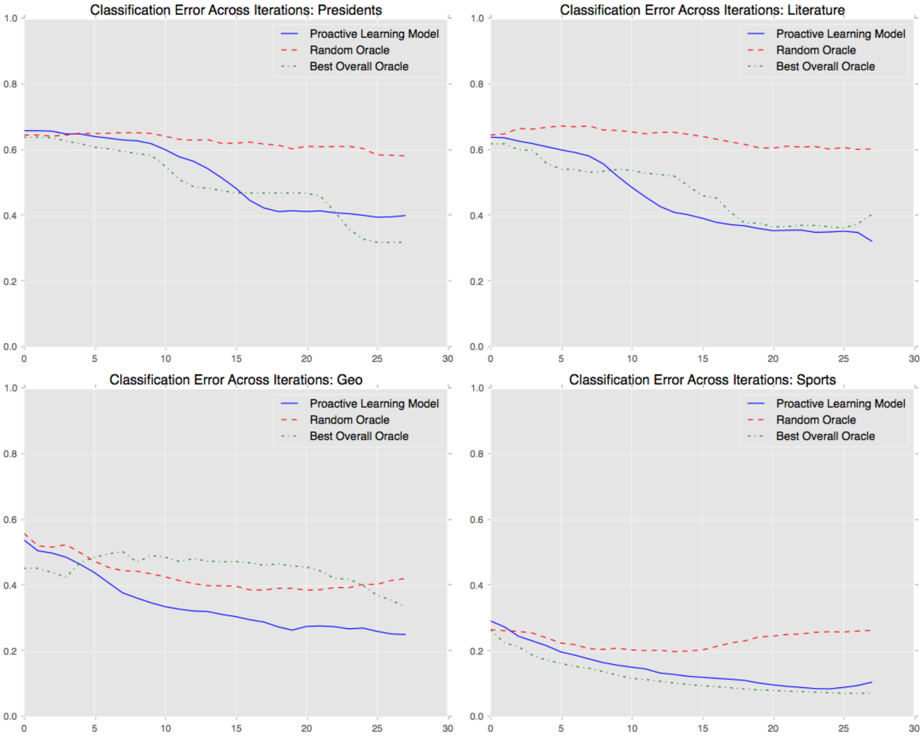

Figure 5: Performance comparison for our proactive learning agent against the most accurate oracle as well as a model that randomly selects oracles. The category is indicated above each plot.

Table 3: Top Oracles per Category

Category	Model's Top Oracle	Best Oracle
Geography	P8	P8
Literature	P1	P1
Presidents	P3	P3
Sports	P6	P8

6.1 Automated Knowledgeability Prediction Results

The results of our experiment in predicting knowledgeability can be found in Table 2. We achieve similar results to work done by Bourai et al as we are able to predict the knowledgeability of speakers above chance level [4]. In addition, audio features outperformed visual features for classification but a combination of the two feature sets yielded better results. This multimodal approach allows the model to capture dependencies between audio and visual features. Swerts and Kramer found that giving human annotators both speech and video of a person answering a question allowed for higher accuracy when predicting knowledgeability [28].

We see that including the “I do not know" feature increases our accuracy. While it is a verbal cue, the increased accuracy of our model is necessary to ensure better results when we are evaluating our proactive learning model. This also indicates that our model is incorrectly classifying non-answers as correct answers. Participants sometimes seem confident indicating their lack of knowledgeability, however as mentioned in the feature analysis section smile events were also prevalent in non-answers. Bourai et al's work did not include this feature of “I do not know" as participants in the trivia setting were required to always give an answer [4].

6.2 Empirical Evaluation of Proactive Learning Agent

We evaluated our proactive learning agent's ability to select the right human oracle for aid on a particular question. Figure 5 contains plots of classification error per category across each iteration. The results presented are averaged across 10 runs. We see that sports and geography immediately experience a drop in error while the presidents and literature categories are smoother at first. This is due to our active sampling function $V(x)$. It favors categories that have mean high mean accuracy and standard deviation but also unlabeled answers. As can be seen in Table 4, literature and presidents were both difficult categories for our labelers, thus the initial labeled set will have lower average accuracy than sports or geography and will not be sampled early. However, as the number of iterations increase and the more accurate categories converge, we can see a significant drop in error for the presidents and literature categories. This is influenced by the $\frac{|c_{UL}|}{|c|}$ term in our active sampling function as the agent will begin to sample from categories that are not fully labeled yet. The plots in Figure 5 also compare with a model that only samples the “best" oracle. This was determined by selecting the participant who had the highest average question accuracy across

Table 4: Comparison of Error

Category	Random	Best Oracle	Proactive Learning
Geography	0.500	0.365	0.248
Literature	0.625	0.413	0.281
Presidents	0.635	0.343	0.372
Sports	0.288	0.082	0.091

all categories. We can see that our agent, with no knowledge about which oracle is an "expert", can still converge to a similar or better average classification error compared to the domain experts by the end of each run. The proactive learning agent relies solely on its interactions with the human oracles to determine which oracle's answer will most likely be correct. The agent is able to capture the non-verbal signals emitted by the humans and use them to make a decision about which oracle to select. In Table 3 we see the oracles the agent selected most often for each category.

In Table 4, we see that the model is able to eventually converge to an error below chance for all categories. It is able to select oracles which are likely to correctly answer the given question. This is true regardless of the initial error of the category. For example, the presidents category had the highest initial error at 0.635 but the model was able to select answers from the correct labelers such that our final error is comparable to if we had only asked the highest performing oracle. In the literature and geography categories, our model actually outperforms even the best oracles. Literature especially was a difficult category for the human oracles, but the proactive learning agent is still able to pinpoint which oracle will most likely be correct. This indicates the strength of combining this decision theoretic approach of proactive learning with our knowledgeability model. We can successfully select the best oracle not just based on their past answers, but also on their direct likelihood to answer this specific question correctly thanks to our knowledgeability model's predictions.

7 CONCLUSIONS

We presented an intelligent agent that is capable of interacting with human oracles and sampling the ones most likely to answer a given question correctly. This interaction was modeled through analyzing speech patterns and facial expressivity to determine the likelihood a speaker will be correct. We also collected a new dataset with video interviews and trained a knowledgeability model that performs above chance and other baselines. Due to variance between the trivia and interview setting, it is clear that a larger study with more participants will be necessary to make definitive conclusions about the types of features we can focus on for a general knowledgeability detector.

Our empirical results indicate that our agent is capable of determining which oracles to ignore or accept answers from. It performs better than chance and is comparable to or better than the best human oracles. The experiment also showcases how proactive learning agents can be used in a real world scenario with human participants and moving away from simulated data. It allows us to model a much more realistic interactive approach to active learning and human-computer cooperation.

REFERENCES

[1] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. 2016. Openface: an open source facial behavior analysis toolkit. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*. IEEE, 1–10.

[2] Paul Boersma and David Weenink. 2001. PRAAT, a system for doing phonetics by computer. *Glot International* 5, 9/10 (2001), 341–345.

[3] Dwight Bolinger. 1982. Intonation and its parts. *Language* (1982), 505–533.

[4] Abdelwahab Bourai, Tadas Baltrušaitis, and Louis-Philippe Morency. 2017. Automatically Predicting Human Knowledgeability Through Non-verbal Cues. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. ACM, 60–67.

[5] Susan E Brennan and Maurice Williams. 1995. The feeling of AnotherâĂš s Knowing: prosody and filled pauses as cues to listeners about the metacognitive states of speakers. *Journal of memory and language* 34, 3 (1995), 383–398.

[6] Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R Hruschka Jr, and Tom M Mitchell. 2010. Toward an Architecture for Never-Ending Language Learning.. In *AAAI*, Vol. 5. 3.

[7] Linlin Chao, Jianhua Tao, Minghao Yang, Ya Li, and Zhengqi Wen. 2015. Long short term memory recurrent neural network based multimodal dimensional emotion recognition. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*. ACM, 65–72.

[8] Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning* 20, 3 (1995), 273–297.

[9] Pinar Donmez and Jaime G Carbonell. 2008. Paired-sampling in density-sensitive active learning. (2008).

[10] Pinar Donmez and Jaime G Carbonell. 2008. Proactive learning: cost-sensitive active learning with multiple imperfect oracles. In *Proceedings of the 17th ACM conference on Information and knowledge management*. ACM, 619–628.

[11] Paul Ekman and Wallace V Friesen. 1977. Facial action coding system. (1977).

[12] Julian T Hart. 1965. Memory and the feeling-of-knowing experience. *Journal of educational psychology* 56, 4 (1965), 208.

[13] Imène Jraidi and Claude Frasson. 2013. Student's uncertainty modeling through a multimodal sensor-based approach. *Journal of Educational Technology & Society* 16, 1 (2013), 219.

[14] Samira Ebrahimi Kahou, Xavier Bouthillier, Pascal Lamblin, Caglar Gulcehre, Vincent Michalski, Kishore Konda, Sébastien Jean, Pierre Froumenty, Yann Dauphin, Nicolas Boulanger-Lewandowski, et al. 2016. Emonets: Multimodal deep learning approaches for emotion recognition in video. *Journal on Multimodal User Interfaces* 10, 2 (2016), 99–111.

[15] Melissa A Koenig and Paul L Harris. 2005. Preschoolers mistrust ignorant and inaccurate speakers. *Child development* 76, 6 (2005), 1261–1277.

[16] Tera D Letzring. 2008. The good judge of personality: Characteristics, behaviors, and observer accuracy. *Journal of research in personality* 42, 4 (2008), 914–932.

[17] Daniel Leyzberg, Samuel Spaulding, and Brian Scassellati. 2014. Personalizing robot tutors to individuals' learning differences. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*. ACM, 423–430.

[18] Abigail A Marsh, Hillary Anger Elfenbein, and Nalini Ambady. 2003. Nonverbal âĂIJaccentsâĂİ cultural differences in facial expressions of emotion. *Psychological Science* 14, 4 (2003), 373–376.

[19] Seungwhan Moon and Jaime G. Carbonell. 2014. Proactive learning with multiple class-sensitive labelers. In *Data Science and Advanced Analytics (DSAA), 2014 International Conference on*. 32–38. https://doi.org/10.1109/DSAA.2014.7058048

[20] Thomas O Nelson. 1990. Metamemory: A theoretical framework and new findings. *Psychology of learning and motivation* 26 (1990), 125–173.

[21] Jasmeet K Pannu and Alfred W Kaszniak. 2005. Metamemory experiments in neurological populations: A review. *Neuropsychology review* 15, 3 (2005), 105–130.

[22] Heather Pon-Barry and Stuart M Shieber. 2011. Recognizing uncertainty in speech. *EURASIP Journal on Advances in Signal Processing* 2011, 1 (2011), 251753.

[23] Lynne M Reder and Frank E Ritter. 1992. What determines initial feeling of knowing? Familiarity with question terms, not with the answer. *Journal of Experimental Psychology: Learning, memory, and cognition* 18, 3 (1992), 435.

[24] Stephanie Rosenthal and Manuela M Veloso. 2012. Mobile Robot Planning to Seek Help with Spatially-Situated Tasks.. In *AAAI*, Vol. 4. 1.

[25] Burr Settles. 2010. Active learning literature survey. *University of Wisconsin, Madison* 52, 55-66 (2010), 11.

[26] Han Sloetjes and Peter Wittenburg. 2008. Annotation by category: ELAN and ISO DCR.. In *LREC*.

[27] Vicki L Smith and Herbert H Clark. 1993. On the course of answering questions. *Journal of memory and language* 32, 1 (1993), 25–38.

[28] Marc Swerts and Emiel Krahmer. 2005. Audiovisual prosody and feeling of knowing. *Journal of Memory and Language* 53, 1 (2005), 81–94.

[29] Lev Semenovich VUIGOTSKY, Eugenia Hanfmann, and Gertrude VAKAR. 1962. *Thought and Language... Edited and Translated by Eugenia Hanfmann and Gertrude Vakar*. Massachusetts Institute of Technology; John Wiley & Sons: New York & London.