Automatic assessment of student "reasoning" processes in face-to-face interactions using speech data

Gahgene Gweon[1], Pulkit Agrawal[2], Nitish Srivastava[2], Shantanu Agrawal[3], Abhilash Jindal[2], Marietta Sionti[3], Bhiksha Raj[1], Carolyn Rose[1]

[1] Carnegie Mellon University {ggweon, bhiksha, cprose}@cs.cmu.edu
[2] Indian Institute of Technology, Kanpur {pulkit, ajindal, nitishs}@iitk.ac.in
[3] Indian School of Mines {shantanu_ag}@ismu.ac.in
[4] University of Athens {marietsionti}@gmail.com

Abstract Assessment of student reasoning processes is one of the holy grails of intelligent tutoring. One way in which students display their reasoning is through conversation. In this paper we present work towards detecting where students are displaying "reasoning" in conversational speech. Such technology would add to the body of work in educational data mining another means of monitoring student work. Because the concept of reasoning is somewhat abstract, we first discuss how we have operationalized it, achieving an agreement of 0.67 kappa between human raters. We then discuss how we have used machine learning technology to predict whether a given speech segment contains a reasoning statement using features that can be extracted automatically using simple audio signal processing techniques. The result is promising with an f-score of 0.63.

Keywords: Machine learning, reasoning assessment, audio signal processing

1 Introduction

Assessment of student reasoning processes is one of the holy grails of intelligent tutoring. Much prior work in the educational data mining community [2, 5] has focused on the analysis of logfiles from intelligent tutoring systems in order to distinguish patterns of behavior that are indicative of shallow involvement (e.g., gaming the system) versus patterns of behavior that are indicative of deeper engagement with the material. Other work has focused on the assessment of group processes from speech, where the goal has been to assess the extent to which students are participating equally in a conversation [6]. In this paper, we bring together these two lines of work, presenting an approach to assessment from speech where the goal is to distinguish segments of speech that are indicative of deep engagement with the material (e.g., explaining why or how something works) versus segments of speech that operate more on the surface (e.g., stating known facts). We refer to segments of speech that display this deeper engagement with material explicit reasoning displays.

An obvious application of such technology within the intelligent tutoring community would be for monitoring self-explanation during problem solving [3].

However, it would also be useful in the case of groups of students interacting with an intelligent tutoring system in order to determine which students are deeply engaged with the task versus those who are talking, but not contributing substantively to the intellectual work of the group. In this paper we work with data collected in the midst of a group problem solving task, and thus we focus on the second of these two scenarios.

In the remainder of the paper we first situate our work in the midst of current directions in speech processing. Next we discuss our approach to operationalizing reasoning displays. We then move on to a discussion of the technological contribution of the paper, which is an approach for automatically distinguishing between speech segments that contain reasoning displays and those that do not. We then present an evaluation of our approach and conclude with a discussion of current directions.

2 Motivation and Background

The research goal of this work is to develop a technique to distinguish between segments of speech that contain an explicit display of reasoning from those that do not. In the midst of group work, these explicit reasoning displays are important for achieving some of the important benefits of collaborative learning. From a Piagetian perspective, for example, when a student participates in a discussion with other students who have different perspectives, it may provide opportunities for experiencing cognitive conflict if that students compares the perspectives of the other students to his own and notices the inconsistencies between them [4]. If students are not articulating their reasoning, then they won't have the opportunity to do this valuable mental model comparison. Thus, while we do not believe the articulation of reasoning is a sufficient condition for learning either in the student who articulated the reasoning or the other students who are listening, it can be seen as providing potential opportunities for learning.

The task of distinguishing speech segments that contain reasoning displays from those that do not is a new task. In this section, we discuss research on the current state of using machine learning technology and audio processing techniques for a variety of other speech processing tasks that can be seen as in league with our assessment problem, for example, detecting flirting [8] and emotion [1]. We see these as similar because what we are trying to predict from speech is related to "how" the words are spoken rather than the content of the words. Such structural aspects of speech reflect the speech style (e.g., prosody) rather than content. For instance, one can extract basic acoustic features such as variation and values of pitch, intensity of speech, amount of silence and duration of speech. Such feature choice reflects everyday observations about conversational speech. For example, increased variation in pitch might indicate that the speaker wants to deliver his ideas more clearly. Likewise, volume and duration of speech may signal that the speaker is explaining his ideas in detail, and is presenting his point of view about the subject matter.

3 Operationalization of the reasoning process

Our work focuses on discussions involving university students working together on a problem solving task. In this section we describe the data, how we operationalized reasoning displays, and how we annotated the corpus using this operationalization.

3.1 Data

Our corpus was collected in a laboratory setting while students worked face-to-face in groups of three. We are collecting data from a large number of groups as part of a formal group work study. In this paper, we focus on a subset of the data that is being collected, which has already been transcribed and annotated. The specific task the students are engaged in is to design a contraption to protect an egg when falling the distance of two flights of stairs. This task involves applying a variety of principles of physics. The data we focus on is a 30 minute discussion portion of each 3-student group work session when the participants were designing and building the egg holder.

In order to collect clean speech with each student on a separate channel, each student wore a microphone. Nevertheless, although it is possible to clearly identify the main speaker from an audio file, crosstalk, which is the other participants' voices, could still be heard in the background. For each audio file, the main thirty-minute discussion sessions were transcribed and manually segmented for further analysis. The segmentation was conducted according to the following two rules.

1. Begin a segment when the main speaker starts talking. If there is silence at the beginning of the file when the main speaker is silent, this means that there will be an "empty" segment in the beginning.
2. A segment should contain the main speaker's continuous speech. If there is an interruption (silence or crosstalk) that lasts for more than 1 second, a new segment should be created. When you create a new segment, there should be two boundaries – one that marks the end of the main speaker's first utterance, and another that marks the start of the next utterance after the pause.

Once the meetings were segmented, the corpus contained 619 segments, including data from all three participants, from the first meeting and 721 from the second meeting.

3.2 Operationalization of Reasoning Displays

We are considering that there is a certain amount of information that has been given to the students, in the form of a task statement and training materials. The displayed reasoning that we are interested in capturing is what goes beyond what is given and displays some understanding of a causal mechanism since typically some causal mechanism would be referenced in a discussion of how something works or why something is the way it is. One purpose in segmenting student talk and identifying

which segments display reasoning is so that amount of reasoning displayed can be quantified. What we are coding is attempts at displayed reasoning. Thus, we need to allow for displays of incorrect, incomplete, and incoherent reasoning to count as reasoning, as long as in our judgment we can believe an attempt at reasoning was going on. That will necessarily be quite subjective – especially in the case of incoherent explanations.

Our formulation of what counts as a reasoning display comes from the Weinberger & Fischer's [11] notion of what counts as an "epistemic unit", where what they look for is a connection between some detail from a scenario (which in their case is the object of the case study analyses their students are producing in their studies) with a theoretical concept (which comes from the attribution theory framework, which the students are applying to the case studies). When they have seen enough text that they can see in it mention of a case study detail, a theoretical concept, and a connection between the two, they place a segment boundary. Occasionally, a detail from a case study is described, but not in connection with a theoretical concept. Or, a theoretical concept may be mentioned, but not tied to a case study detail. In these cases, the units of text are considered degenerate, not quite counting as an epistemic unit.

We have adapted the notion of an epistemic unit from Weinberger & Fischer, rather than using it the same way Weinberger et al. did in their work because the topic of our conversations is very different in nature. We consider that the basic requirements for a unit of talk to count as a reasoning display is that it has to contain evidence of a connection between some detail from the problem the students are trying to solve, such as a choice of materials, or a way of combining materials, and relevant concept, such as from physics, which could be a principle that justifies a design choice. As mentioned, we would like to distinguish this from just parroting what they have heard. In our current formulation, we are considering the task and training materials that the experimenter has provided as what is given. We would like to make a distinction between what is given and what the students contribute beyond that.

Now we will make more concrete what our operationalization of reasoning looked like. First, examine a segment of a conversation where we have highlighted the instances of displayed reasoning using bold italics.

s1: ***i think we'll need only one rubber band because the rubber band is circular. We can just break it off right***
s3: oh yeah. that's a good idea.
s2: See what are the weights
s1: ***it is some significant difference***
s2: ***Yeah this is heavier. So this could be on top***
s3: ***yeah cause if we did that then that would fall on the bottom, right? It might do some spinning.***

The simple way of thinking about what constitutes a reasoning display is that it has to communicate an expression of some causal mechanism. Often that will come in the form of an explanation, such as X because Y. However, it can be more subtle than that, for example "Increasing the tension makes the spring springier." The basic

premise was that a reasoning statement should reflect the process of drawing an inference or conclusion through the use of reason. Note that in the example with the spring, although there is no "because" clause, one could rephrase this in the following way, which does contain a "because" clause: "The spring will be springier because we will increase the tension." Reasoning statements stand in contrast to mere information sharing statements, which can be thought of as sharing of rote knowledge.

Concepts. The basic building block of a reasoning statement is a concept. We identified 5 types of concepts relevant for our domain, namely theoretical concepts, prior knowledge, physical system properties, emergent system properties, and goals. For each concept, the definition and an example are illustrated in table 1.

Table 1. Definition and examples for the 5 concepts. The examples are from our dataset described in section 3.1, where students are discussing a best approach to build an egg holder. Note that the "system" in this case is the egg holder, plus any materials that are available for use.

Type	Definition	Example
Theoretical concept	principles (i.e. physics principle) and theories	when an object is falling, the force of impact when it hits the ground can be decreased by slowing down the speed.
Prior Knowledge	information based on common sense	Using a small amount of tape would not be enough to hold two bowls together
Physical system properties	elements and characteristics of elements that are available for the system	paper bowl is round, straws are flexible
Emergent system properties	characteristics of elements that appear in a process	stability of an egg holder which emerges as a result of using certain materials
Goal	general believes/ perspectives, anything associated with strong expectations related to points of view	aesthetics of an egg holder, i.e. trying to make the egg holder aesthetically pleasing

Relationship. The presence of multiple concepts in a statement by itself does not determine whether a statement contains reasoning. Rather, the relationship between multiple concepts is the determining factor. For example, a simple list of concepts (e.g. this cup is round, and it is also white) is information sharing, and not reasoning. We identified two types of relationships that signal a reasoning process; 1. compare and contrast, 2. cause and effect.

1. juxtapositions, compare and contrast, tradeoff: When the speaker compares two concepts, the speaker is making a judgment, which involves thinking about how two concepts are related to another.

    a. The speaker compares two materials ("that" & "rubber band") for his solution.
    "I am thinking that might work <u>better than</u> a lot of rubber bands."

2. Cause and effect: When the speaker uses a cause-and-effect relationship, this process involves establishing the relationship between two concepts through a reasoning process. The general relation in this category is "doing x helps you achieve y" There are three main types of causal relationship a)cause and effect b)in order to c)analogy. Example for each of the three types are illustrated below

    a. Let's do A because of B.
    "Let's use bubble wrap <u>because</u> it cushions the fall"
    b. Let's do A in order to achieve B.
    "Let's use rubber bands <u>for</u> tying the bag onto the bowl."
    c. When a speaker makes an analogy, he is making a link due to the similarity between two concepts. Some of the keywords that signal analogies are "like", "as".
    "Oh, you're trying to use the bowl <u>as</u> a parachute."

3.3 Reliability of Annotation

Two coders were initially trained using a manual that describes the above operationalization of reasoning displays in detail along with an extensive set of examples. After each coding session, the coders discussed disagreements and refined the manual as needed. Most of the disagreements were due to the interpretation of what the students meant rather than the definition of reasoning itself. Therefore, later efforts focused more on defining how much context of a statement could be brought to bear on the interpretation and how. In a final evaluation of reliability, we calculated kappa agreement of 0.67 between two coders over all the data. After calculation of the kappa, disagreements were settled by discussion between the two coders.

4 Automatic assessment of reasoning processes

The purpose of our investigations with speech technology were to determine the extent to which it is possible to use current machine learning technology to predict whether a given speech segment contains a reasoning statement or not using features extracted by means of simple audio processing techniques. We first describe our approach. In the subsequent section, we detail our promising results.

4.1 Methods

The goal of this portion of our investigation was to distinguish reasoning statements from non reasoning statements using machine learning technology. This procedure consists of three main stages, namely, preparing the audio data, extracting features, and applying machine learning. The details of each stage are presented below.

The first step involved "cleaning up" the audio data and segmenting it into units for analysis. For each meeting participant, we collected an audio file containing his speech. Although each audio file mainly contained the main speaker's speech, it also contained crosstalk, which is voice of other participants in the background. Since each audio file was used to analyze the main speaker's talking behavior, we were only interested in the main speaker's audio data. Therefore, we cleaned up the audio files by removing the crosstalk by using a relatively simple algorithm. Since the crosstalk was relatively low in volume, we set a threshold parameter of 0.1 over the volume. The range of the main speaker's regular signal lies in the interval [1-, 1]. To remove the cross talk, we first generated speech segmentation boundaries using the threshold. The choice of threshold value is detailed in the next paragraph. Next, we zero out all signals that were marked as non-speech using this threshold.

To find the threshold value, we experimented with threshold values between 0.05 and 0.2, with increments of 0.01. First, for any value that was below the threshold of 0.1, i.e. values between -0.1 and 0.1, the value was reduced to 0. For each threshold value, we generated segment boundaries by scanning the signal from left to right after applying the threshold. Next, we set up a cost function, which computed the cost of aligning these segment boundaries to the gold standard (human segmentation described in section 3.1). The threshold of 0.1 was the value that minimized the cost function. We also varied the value of the threshold to see how much of the actual speech was lost, and 0.1 was found to minimize the loss of speech while removing most of the background noise. In summary, this method of choosing the threshold value was relatively simple, yet computationally inexpensive and effective.

Once the crosstalk was removed, the audio file was segmented according to the following 4 steps. First, the signal was scanned from left to right using a window of 500ms. Second, at each time when the signal became non-zero, it was marked as a start time. Third, the window kept on moving across the signal until all the values inside the window became zero. Finally, a backwards search was done to locate the exact end point within the last 500ms window.

After the initial stage of cleaning up and segmenting the data into units, the second stage involved transforming each segmented unit into a set of feature-value pairs. Each segment was labeled using the gold standard label described in section 3.2. For the feature set, a total of 50 features were initially extracted for each segment, including, speaker id, duration of the segment, 40 Mel Frequency Cepstral Coefficients (mfcc), 4 amplitude features, and 4 pitch features. However, to train the model, only 5 features of the 50 were used; the speaker id, duration, and the top three features from principal component analysis (PCA) for the rest of the 48 features. The decision to include only 3 of the 48 features results in substantial amount of reduction in computing power, yet utilized most of the essential information captured from acoustic features.

The initial 40 mfcc features are the result of applying a set of 40 standard filters, which are available as part of VoiceBox Matlab Toolbox [10]. These mfccs reflects the distribution of energy level. The 4 amplitude features are: the value of the amplitude of the overall segment, mean, median, and standard deviation value of the 1 second windows in a given speech segment. Similarly, the 4 pitch features are: pitch of the overall segment, mean, median, and standard deviation of pitch over 1 second windows in a given segment. The pitch features were extracted according to the YIN

algorithm [12]. The amplitude features reflect the intensity and energy level of speech. Therefore, the mean value of amplitude could show the volume of the speaker and the standard deviation of amplitude could be used to show the amount of variation of intensity in the speaker's speech segment. The pitch features also reflect intensity of speech. All three types of features represent structural aspect of the speech signal rather than the content.

The third and final stage involved predicting whether it is possible to use machine learning to automatically label segments of speech as a "reasoning" or "non reasoning" contribution with high enough accuracy using the set of features just described. We used a structural support vector machine (SVM) learning algorithm [7]. We ran two sets of different experiments. Note that our data consists of conversation from two different meetings, each lasting 30 minutes in length. We will refer to each meeting as meeting 1 and meeting 2 from here on. Although the task given for the two meetings were identical in that three participants build an egg holder with given materials, the nature of the conversation was different because different participants were involved in the meeting. For our purpose, the amount of reasoning differed for each meeting. 208 out of 619 segments (30% of data) were coded as "reasoning" contributions in meeting 1, where as 59 out of 721 segments (10% of data) were coded as "reasoning" contributions in meeting 2.

The goal of the first set of experiment was to test whether the addition of acoustic features would improve the prediction. Using the same training (meeting 1 data) and test (meeting 2 data) sets, we ran four experiments, where we varied the number of features. For the baseline case, only speaker id and speech duration features were used. Then for the subsequent 3 experiments we added additional features, which reflects the structural aspects of speech, namely 1 pca, 2 pca, and 3 pca. 1 pca is the top feature that we narrowed down from the 48 acoustic features (40 mfcc, 4 amplitude, 4 pitch) that we collected using the principle component analysis (PCA). 2 pca are the top two features, and 3 pca are the top three features.

The first set of experiment would not only show the usefulness of acoustic features, but also how "good" our model is when tested on a new set of data that is collected from different participants. However, we expected that the performance of the model would not be as good as it could be given that it was build using data from only one meeting. Therefore, for the second set of experiment, we combined data from both meeting 1 and 2 to create the training and the test set. The training set used was 80% of randomly selected data from each of meeting 1 and 2. The remaining 20% from both meetings were used as the test set. The details and results of these experiments are presented in section 4.2

4.2 Results

The results of the two sets of machine learning experiments are shown in table 2. For the first set of experiments, as we added additional acoustic features, the f-score improved as expected. Although adding 1pca did not improve the f-score, additional 2 or 3 top acoustic features improved the f-score from about 42% to 47%. The improvement was mostly due to improvement in the recall, rather than precision.

The second set of experiments showed similar results in that addition of acoustic features improved the f-scores. However, with the introduction of a more varied data set where we combined meeting 1 and meeting 2 data, the best f-score is 62%. This shows promise in that with the addition of more data, our score could be improved.

Table 2. Machine learning experiment results

exp#	training set	test set	features used	Recall (%)	Precision (%)	F-score (%)
1	100% of meeting1	100% of meeting2	id, duration	40.68	43.64	42.11
			id, duration, 1 pca	38.99	46.93	42.59
			id, duration, 2 pca	49.15	46.03	47.54
			id, duration, 3 pca	47.46	45.90	46.67
2	80% of (meeting1 + meeting2)	remaining 20% of (meeting1 + meeting2)	id, duration	35.29	92.31	51.06
			id, duration, 1 pca	44.12	83.33	57.69
			id, duration, 2 pca	52.94	75	62.07
			id, duration, 3 pca	52.94	75	62.07

For both sets of experiments, duration of segment was the top indicator for determining whether a contribution contained reasoning or not. This result matches the heuristic that if a contribution contains reasoning, it is longer duration because the speaker needs time to express his thoughts.

5 Conclusions and Current Directions

In this paper, we presented our work towards automatic detection of reasoning displays in speech data. Our work shows promise in that 1) humans can distinguish reasoning and non-reasoning statements with reasonable reliability, and 2) using machine learning, classification of a statement as reasoning/ non-reasoning is feasible, even with limited training data. However, our results should be considered preliminary since the amount of data was limited to two meeting sessions. We are currently collecting and annotating audio data from additional meetings to validate our result further as well as testing its generality across a wider variety of student groups. In terms of technological improvements, our next steps include incorporating sequential information explicitly within the feature space rather than incorporating it as part of the structural svm [7] machine learning algorithm. We expect this to improve performance since it would simplify the complexity of the learning algorithm. In addition, we would like to find other features that would reflect the coding process used by human annotators or the structure of the language data by incorporating content related features. One initial step towards incorporating such content related feature would be using speech recognition to spot key words that indicate cause and effect relationships such as “because” or “for”.

Acknowledgments. This research was supported in part by xxx.

References

1. Ang, et. al. Prosody based automatic detection of annoyance and frustration. In Proc. *International conference spoken language processing*, (2002)
2. Baker, R.S., Corbett, A.T., Koedinger, K.R. Detecting Student Misuse of Intelligent Tutoring Systems. Proceedings of the 7th International Conference on Intelligent Tutoring Systems, 531-540 (2004)
3. Chi, M.T.H., Bassok, M., Lewis, M.W., Reinman, P., & Glaser, R. Self-explanations: how students study and use examples in learning to solve problems. Cognitive Science, 13, 145-182 (1989)
4. De Lisi, R., Goldbeck, SL. Implications of the Piagetian Theory for peer learning. In O'Donnell A. M., Kig A. Cognitive perspectives on peer learning. Ner Jersey, Laurence Erlbaum Associated Inc. pp.3-37. (1999)
5. Feng, M., Beck, J., & Heffernan, N. Using Learning Decomposition and Bootstrapping with Randomization to Compare the Impact of Different Educational Interventions on Learning. In Barnes, Desmarais, Romero & Ventura (Eds) Proc. of the 2nd International Conference on Educational Data Mining. pp. 51-60 (2009)
6. Kim, T. Chang, A., Pentland, A. Meeting Mediator: Enhancing Group Collaboration with Sociometric Feedback, In Proceedings of Conference on Computer Supported Collaborative Work, San Diego, CA, pp.457-466 (2008)
7. Structural Support Vector Machine. http://svmlight.joachims.org/svm_struct.html
8. Ranganath, R., Jurafsky, D., & McFarland, D. It's Not You, it's Me: Detecting Flirting and its Misperception in Speed-Dates. Proceedings of EMNLP (2009)
9. Teasley, S. D. Talking about reasoning: How important is the peer in peer collaboration? In L. B. Resnick, R. Säljö, C. Pontecorvo & B. Burge (Eds.), Discourse, tools and reasoning: Essays on situated cognition, pp. 361-384 (1997)
10. Voicebox. http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html
11. Weinberger A., Fischer F. A framework to analyze argumentative knowledge construction in computer supported collaborative learning. Computers & Education; vol 46, pp.71 - 95 (2006)
12. YIN: fundamental frequency estimator. http://labrosa.ee.columbia.edu/doc/yin.html