Analysis of Intelligent Tutoring Systems

Yash Sinha[1], Geetanjli Chugh[2] and Rongrong Wang[3]

Abstract— **This document fulfills the course requirement of CIS6930 - DIALG SYS NAT LANG IF. The document gives an introduction to the three research papers reviewed by us and follows with individual summaries and critique of the papers. The last section of the document summarizes the approach we took for data collection, data extraction and finally discusses the result of the process for the second question of the assignment.**

I. INTRODUCTION

Intelligent tutoring systems are clearly one of the successful enterprises in Artificial Intelligence. The three papers we analyzed are based on the new generation of intelligent tutoring systems namely AutoTutor[1] and ITSPOKE[3]. These tutoring systems pose challenging problems to the students and take feedback. They assist in learning gains and clarify the misconceptions of the students using interactive human tutor simulation. The first two papers also talk about the recent advances in this domain wherein interactive 3-D simulations can help the students in easier learning. Another improvement is that of recognizing and responding to students' emotion and thus making the experience more interactive.
The third paper 'Toward Detecting Emotions in Spoken Dialogs'[2] proposed a negative and non-negative emotion-recognition method, using thecombination of acoustic, lexical anddiscourse information to greatly improve the emotional classification performance.

II. AutoTutor: An Intelligent Tutoring System With Mixed-Initiative Dialogue

A. Summary of the paper

The research paper AutoTutor: An Intelligent Tutoring System With Mixed-Initiative Dialogue [1] tries to give a brief overview about AutoTutor which is an intelligent tutoring system that simulates a human tutor. This is an application of dialogue system wherein the conversation between the simulator and the student is done in natural language. The author states how the AutoTutor system assists the students by helping them construct improved answers through marginal learning and also helps them clarify their misconceptions. The paper introduces the teaching strategies of AutoTutor. It tries to evaluate students contribution to the conversation through Latent Semantic Analysis wherein it matches the students answers to either expectation or misconception. The dialogue pattern used by AutoTutor is Expectation and Misconception Tailored dialogues (EMT). This dialogue pattern is flexible in the sense that the trajectory of the conversation isnt fixed. The conversation is carried forward based on the students answers. The goals of AutoTutor are to assist the students in covering the list of expectations and also clarifying their misconceptions. The AutoTutor tries to achieve these by giving feedback to the students answer in the form of hints, prompts, assertions and other dialogue moves.

The author states four methods that were used to evaluate the performance of AutoTutor. The dataset involves the students in controlled laboratory experiments as well as classrooms. The paper discusses two of the performance metrics:

- **Evaluation of Dialogue Quality**
  The author talks about how the dialogue quality of AutoTutor was tested based on the students discretion on whether the dialogue presented to them was that generated by the AutoTutor or a human tutor. The set of dialogues were generated half by a human tutor and half by AutoTutor. The results were surprising as the students were unable to discriminate whether the dialogues were human tutor-generated or by AutoTutor. Thus, the paper claims that AutoTutor is a good simulator as it allows for a smooth conversation with the students.
- **Evaluation of Pedagogical Quality**
  The author states that in this method the students take a pre-test and then interact with the AutoTutor for a few hours. Then a post-test is completed by the students. The paper states that in ten experiments involving over 1000 students, AutoTutor enhanced the learning with gains ranging between .2 and 1.5 sigma with a mean of .8. The paper draws a comparison with the learning gain through a human tutor wherein the enhanced learning was with an effect of .4 sigma.

The paper also discusses the architecture of AutoTutor and briefs about its different components. The process of taking the input from the student it, analyzing the statement and responding back with an answer is also stated in the paper. The paper then talks about the new advances in AutoTutor including interactive 3-D simulation and responding based on students emotions. In conclusion the paper talks about the limits of the natural language systems and states that AutoTutor is suitable for conversations that involve low-to-medium user knowledge of the topic as in case the shared knowledge is high, the expectation of precision from the tutor

[1] Yash Sinha is in the Department of Computer Science University of Florida Gainesville Florida, USA
[2] Geetanjli Chugh is in the Department of Computer Science University of Florida Gainesville Florida, USA
[3] Rongrong Wang is in the Department of Computer Science University of Florida Gainesville Florida, USA

and student will be high and this leads to high risk of failure to meet those expectations.

B. Strengths

There are multiple strengths of the paper. The paper tries to state the working of the dialogue system and its advantages over a human tutor by giving concrete examples and results. The tests are illustrative and give a clear picture of the dataset which is important in understanding the testing scenarios. The paper gives statistical measures of the results which help in straightforward comparison with a human tutor.

C. Weaknesses

There are a few weaknesses in the paper. While the strategies and working of AutoTutor are well explained, the order of the topics are a little distracting. While the reader is focused on the comparison of the dialogue system with a human tutor, the author introduces the architecture of the system. That seems a bit abrupt as the more than 3/4th of the paper didnt talk about the architecture rather the practical application of AutoTutor. The author then again dives into the recent advances which are a bit incoherent with the previous topic of the architecture. And finally, in the topic of recent advances of AutoTutor, the author tries to explain the interface of itself which is incoherent with the topic itself. The conclusion is not in line with the entire length of the paper where the author never discusses the disadvantages of the language system when the results are presented. These disadvantages are presented in conclusion and thus instead of an expected summary of the paper, the reader is presented with a new topic altogether.

III. ITSPOKE: An Intelligent Tutoring Spoken Dialogue System

A. Summary of the paper

This research paper 'ITSPOKE: An Intelligent Tutoring Spoken Dialogue System' [3] discusses ITSPOKE which is a human-computer tutoring system that uses an interactive pedagogical agent to engage students in a spoken dialogue. It also gets feedback from them, corrects misconceptions and elicits more complete explanations. The main idea behind this research is to analyze the utility of adding spoken language capability to dialogue tutoring system. ITSPOKE platform is used to examine whether speech and acoustic-prosodic information can improve the performance of dialogue tutoring system. The author discusses the architecture of the system. The ITSPOKE is developed using Python wrapper which acts as a proxy server between Why-2-Atlas (text based tutoring system) server and client. For performance analysis, students with no prior knowledge of Physics were assessed based on a pre-test and a post-test. The post-test was done after the students interacted with ITSPOKE for 4 hours and worked through 5 Physics problems. When the result of tests was analyzed a Word Error Rate of 31.2% was found.

The paper is concluded by the stating that the work on ITSPOKE is still in progress. It tries to identify the improvement areas of ITSPOKE as compared to text with respect to a variety of evaluation metrics.

B. Strengths

Main strength of this paper are its historical analysis and structured flow. This paper shows the co-relation with the older versions (text-based tutoring systems like Why-2-Atlas) of tutoring system which helps us to understand the basis of research and its future scope. The author also explained the architecture of ITSPOKE and illustratively shows that this is a better version. The paper also explains the working of the tutoring system using conversational dialogue examples and how it interacts with the students. The paper talks about the possible future advancements of enhancing these systems by adding responses by recognizing students emotions.

C. Weaknesses

It fails to clarify results of performance analysis with concrete data comparisons. The results shown are not summarized properly and the paper is unable to conclude the analysis of adding spoken language capabilities in dialogue systems. The paper has ambiguity in its application description. I misinterpreted a few sentences while reading. A better clarity in the process description could have helped in better understanding of the system's working.

IV. Toward detecting emotions in spoken dialogs

A. Summary of the paper

In this paper [2], the authors combined acoustic, lexical and discourse signal sources to detect negative and non-negative emotions from human speech, with the data obtained from a realistic human-machine interaction call center. They provided an in-depth investigation by mining the data from a large data pool first, then used a systematic acoustic feature selection method to generate 5 best features, which have better emotion-recognizing performance. After that, the paper used the emotional salience theory for automatic word-emotion recognition. It also used 5 actions from speech as the emotional classifiers to study the discourse information in human-computer dialogues. At last, they combined the output getting from acoustic, lexical, and discourse classifiers together, and achieved a good recognition performance, especially from the combination of acoustic and lexical information.

For more specified details in the paper, I want to say the authors emphasized the difficulties that exist in emotion recognition at first, for there is no accurate definition,nor the categories,of emotions, and thats why only negative emotion and non-negative ones were studied. The paper also used objective and subjective measures (which used 4 human taggers to classify the emotion categories) to narrow down the data that would be more meaningful for this study. As for selecting the best acoustic correlates, the paper used a forward selection method to select top 10 or 15 acoustic features in male and female respectively. And the results were confirmed effective in reducing the dimension by using a principal component analysis method. When the authors tended to combine the three resources from the simple average output of each results, they found the best performance were achieved mostly by the combination of lexical and acoustic information, left out of discourse information. The reason was found that the lexical and discourse information classifiers were dependent to each other, which will not help to improve the recognition performance.

B. Strengths

As for my critiques of this paper, I think the strength that stood out the most is the idea to add two more sources-lexical and discourse source- into the acoustic one to achieve better performance, thus making machines to recognize emotions more appropriately. The second impressive thing is that the authors divided emotions into two categories (negative and non-negative ones) to solve the problem of unable to define and recognize all kinds of emotions that exist, besides, the detection of negative emotion can also be used to provide better services for the application being investigated. The last thing I think is desirable is the paper processed the data systematically. For it mined the data using both objective and subjective approaches. Where in the subjective approach, there were 4 listeners randomly choosing the utterances to tag their corresponding emotions, and a kappa static was used to measure the degree they agree with each others' judgment. This can greatly ensure high accuracy of the emotion classifying and help to make sure the data selected were potentially more useful in this research.

C. Weaknesses

However, there are three aspects I think could be further improved. Firstly the data used could be obtained from a bigger source, rather than a small call center. As it is well known that the larger the database is, the more accurate the results could be. The second weakness is it took only one utterance each turn based on the situation of MOST dialogs when processing the data, but the problem is there might be more utterances in every turn in a dialog, neglecting this rare occurrence might negatively influence the final result. The last defection is when labeling the users response to its speech-act categories, the paper did the task by using only one person, and hence the results of the effects in discourse information could be biased. It will be better if there are more taggers to do the job, to guarantee the accuracy of the classifying process.

V. Basic processing on textual natural language data

A. Data collection process

The requirement for data was at least 100 utterances. Such number of utterances with two-party dialogues are hard to find in an interview excerpt or a movie scene. We tried to find scripts from different plays and found publicly available script from here: **http://www.one-act-plays.com/comedies/open_door.html**

The script is taken from a play where two persons converse with each other for a total of 178 utterances. This script consists of a total of 4,065 words and 18,489 characters if no white spaces are considered. Below is an excerpt from the play.

LADY TORMINSTER: Oh!

SIR GEOFFREY: [Rising.] Hullo! Don't be afraid–it's only I!

LADY TORMINSTER: What a start you gave me Why haven't you gone to bed?

SIR GEOFFREY: I'm tired of going to bed. One always has to get up again, and it becomes monotonous. Why haven't you gone to sleep?

LADY TORMINSTER: I don't know–it's too hot, or something. I've come for a book.

SIR GEOFFREY: Let me choose one for you.

[He goes to the table.]

LADY TORMINSTER: Why were you sitting in the dark?

SIR GEOFFREY: Because the light annoyed me. What sort of book will you have? A red one or a green one?

LADY TORMINSTER: Is there a virtue in the colour of the binding?

SIR GEOFFREY: Why not? They're all the same inside. There are three hundred ways, they say, of cooking a potato–there are as many of dressing up a lie, and calling it a novel. But it's always the same old lie. Here take this. [He hands her a book.] Popular Astronomy. That will send you to sleep.

LADY TORMINSTER: The stars frighten me. But I'll try it. Good-night.

SIR GEOFFREY: Good-night.

LADY TORMINSTER: And you really had better go to bed.

SIR GEOFFREY: I move as an amendment that you sit down and talk.

B. Data Manipulations

We have manipulated the data to suit a better approximation.

- We have removed the occurrences of '–' in the script as there are few pair of words that would have been counted as a single word. eg. of cooking a potato–there are as many
- We also handled the scene descriptions included inside [..] as these were not part of any dialogues
- We removed the punctuation because they would unnecessarily contribute to the length of the words.

C. Problem Approach

We divided the utterances of both the persons in two different sets and then count the total number of words in both the sets. The number of utterances by each speaker is obtained by calculating the occurrences of the identifier 'Speaker Name:'. After that the solutions are trivial using the following formulas:

- Number of words = Total number of strings in each lists

- Average word per turn = (Total number of words by the speaker)/(Total Dialogue turns of the speaker)
- Average length of words = (Total length of lists as string array removing the white spaces)/(Total number of words spoken by the speaker)

The code has been uploaded here: **goo.gl/JbZ59r**

D. Results

These were the results that were obtained:

- Person1 makes 89 dialogue turns
- Person2 makes 89 dialogue turns
- Person1 says 1502 words
- Person2 says 1907 words
- Person1 says 16.88 words on an average per turn
- Person2 says 21.43 words on an average per turn
- Average length of words spoken by Person1 is: 3.82
- Average length of words spoken by Person2 is: 4.04

We can infer from the results that Person1 speaks less number of words and that is apparent from the fact that she speaks less words on an average per turn and also the average length of words spoken by her are also less given the fact that both the speakers take an equal number of dialogue turns.

REFERENCES

[1] AC Graesser, P Chipman, and BC Haynes. AutoTutor: An intelligent tutoring system with mixed-initiative dialogue. *IEEE Transactions on*, 2005.

[2] CM Lee and SS Narayanan. Toward detecting emotions in spoken dialogs. *IEEE transactions on speech and*, 2005.

[3] DJ Litman and S Silliman. ITSPOKE: An intelligent tutoring spoken dialogue system. *Demonstration papers at HLT-NAACL 2004*, 2004.