A Preliminary Investigation of Learner Characteristics for Unsupervised Dialogue Act Classification

Aysu Ezen-Can
Department of Computer Science
North Carolina State University
aezen@ncsu.edu

Kristy Elizabeth Boyer
Department of Computer Science
North Carolina State University
keboyer@ncsu.edu

ABSTRACT

For tutorial dialogue systems, classifying the dialogue act (such as questions, requests for feedback, or statements) of student natural language utterances is a central challenge. Recently, momentum is building for the use of unsupervised machine learning approaches to address this problem because they reduce the manual tagging required to build dialogue act models from corpora. However, unsupervised models still do not perform as well as supervised models in terms of accuracy. This paper presents an unsupervised dialogue act modeling approach that leverages the influence of learner characteristics, particularly students' perceptions of their own skill, on their language use. The experimental findings show that leveraging skill perception within dialogue act classification improves performance of the models, producing better accuracy. This line of investigation will inform the design of next-generation tutorial dialogue systems, which leverage machine-learned models to adapt to their users.

Keywords

Tutorial dialogue, learner characteristics, dialogue act classification, unsupervised machine learning.

1. INTRODUCTION

Tutorial dialogue is a highly effective form of instruction, and much of its benefit is thought to be gained from the rich natural language dialogue exchanged between tutor and student [2]. In order to model tutorial dialogue for the purposes of building tutorial systems or for studying human tutoring, *dialogue acts* provide a valuable level of representation. Dialogue acts represent the underlying intention of utterances (for example, to ask a question, agree or disagree, or to give a command) [1]. For tutorial dialogue systems, dialogue act classification is crucial to understanding students' utterances and developing tutorial strategies [6].

Today's tutorial dialogue systems utilize a variety of dialogue act classification strategies. Historically when machine learning has been used to devise tutorial dialogue classifiers, these have been *supervised* classifiers, which require training on a manually labeled corpus. However, supervised techniques face substantial limitations in that they are labor-intensive due to the manual annotation and handcrafted dialogue act taxonomies that are usually domain-specific. To overcome these challenges, unsupervised dialogue act modeling techniques have been investigated in recent years.

Despite this growing focus on developing unsupervised dialogue act classifiers, these models still underperform compared to supervised approaches in their accuracy for classifying according to manual tags. However, while unsupervised models to date have considered such things as lexical features (the words found in the utterance) and syntactic features (the structure of the sentence), they have not considered learner characteristics, such as skill perception, which are believed to influence the structure of tutorial dialogue [3]. Learner characteristics also play an influential role in learning in web-based courses [5].

This paper investigates whether the performance of an unsupervised dialogue act classifier can be improved by taking a specific learner characteristic into account. We utilize *skill perception*, a student's ranking of her own skill as she perceives it compared to others. Specifically, we train unsupervised dialogue act models that are tailored to students of a specific skill perception level, and we compare those models to ones trained without restricting by that learner characteristic. This unsupervised training is conducted entirely without the use of manual tags. We then test the models on held-out test sets within leave-one-student-out cross validation, and compare the resulting classification accuracy according to their previously applied manual tags. The results can inform the way that next-generation tutorial dialogue systems conduct their real-time dialogue act classification.

2. DIALOGUE ACT MODELING

The corpus used in this study consists of computer-mediated student-tutor interactions during an introductory computer science programming task [4]. Throughout the data collection, students and tutors communicated through a textual dialogue-based learning environment while working on Java programming. Students were given a pre-survey that included items on computer science. The pre-survey included an item that asked students to rate how skilled they are in the domain compared to others. We refer to this response as skill perception. Students (n=42) were divided into groups (high and low skill perception) based on the median score.

The corpus containing 1,640 student utterances was manually annotated with dialogue act tags in a previous work [4]. There are seven student dialogue acts in total (*Answer, Acknowledgement, Statement, Question, Request for Feedback, Clarification and Other*) where the majority class baseline chance is 39.95%. As required by unsupervised modeling, these dialogue act tags are not available during model training, but we use them for evaluation purposes to calculate accuracy on a held-out testing set.

We hypothesize that dialogue act models built using unsupervised machine learning will perform substantially better when customized to specific learner group skill perception. The corpus is partitioned by skill perception and we examine whether an unsupervised dialogue act classifier trained only on students with high skill perception performs better on a test set of dialogue acts

from high skill perception students, compared to a classifier trained on a mixture of high and low skill perception students.

In order to gather accuracy data across these characteristics, we conduct leave-one-student-out training and testing folds. The testing set for each of the *n* folds consists of all of a single student's dialogue utterances and the model is trained on the remaining *n-1* students. We compute the average test set performance of the model across all folds for each learner characteristic partition. The performance metric utilized in this study is *accuracy* compared to the manually labeled dialogue acts where accuracy is the number of utterances in the test set that were classified the same as their manual label, divided by the total number of utterances in the test set. Our unsupervised dialogue act classifier leverages the *k*-medoids clustering technique.

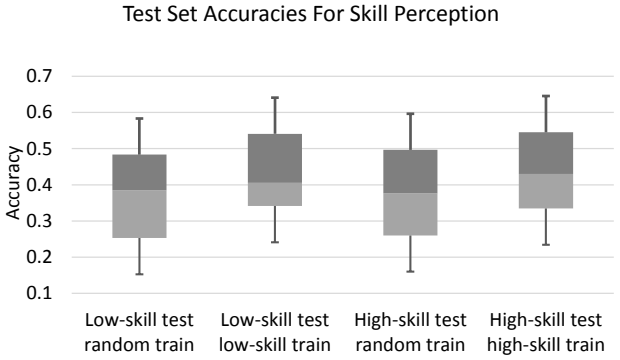

Figure 1: Leave-one-student-out test set accuracies for models by skill perception

For students with low skill perception ($n_{lowSkill}$=26) the average performance of the dialogue act classification model trained on utterances of randomly selected students is 0.39 (σ=0.17) whereas the accuracy rises to 0.43 (σ=0.17) for a tailored model trained only on students with low skill perception (Figure 1). This is not a statistically significant difference after Bonferroni correction. For these students, 11 out of 26 cases improved their performance by utilizing the learner characteristic (five of them above 5% and five of them above 15%), six of them were affected negatively (four of them below 5% decrease) and nine of them achieved the same performance.

The same pattern is visible for students with high skill perception ($n_{highSkill}$=16). For these students, the average test set accuracy increases from 0.38 (σ=0.14) to 0.42 (σ=0.13) gained by using utterances of students with high skill perception rather than learning from utterances selected randomly, again not statistically significant after Bonferroni correction. Six of the cases out of sixteen improve test set accuracy (four of them above 15%), three of them degrades (two of them below 5%) and seven cases perform equally.

Although the differences in model performance were not statistically reliable for students in different skill perception groups, we observed some interesting patterns within these groups (Table 1). Students with low skill perception tended to use more utterances such as, "ok I am getting it," which may be a type of affective or face-saving dialogue move. Students in the high skill perception group seem to exhibit more social, relaxed utterances, reflected by examples such as, "cool cool" and "yeah haha."

	Low Skill Perception	High Skill Perception
Acknowledgments	- oh - ok I am getting it - ok I get it! - interesting - oh ok	-cool cool -yeah haha - yep lol -yep! exciting stuff! - sure
Questions	- what do i do now - can you explain more about the scanner line - so what is this doing exactly? -why is not it prompting me to enter my name?	-comments are just a way to write notes to others to help them understand right? - out of curiosity would not it make sense to switch those last two lines of code?

Table 1: Selected utterances from clusters tailored to skill perception

3. CONCLUSION

Understanding student natural language within intelligent tutoring systems is a critical line of investigation for tutorial dialogue systems researchers. For dialogue act classification in particular, the field has only begun to explore unsupervised approaches and to investigate the range of features that are beneficial within this paradigm. We have presented a first attempt to leverage learners' perception of their own skill within a dialogue act classification model. It is hoped that the research community can continue to build richer models of natural language understanding for students of all learner characteristics in order to enhance learning.

ACKNOWLEDGMENTS

Thanks to the members of the JAVATUTOR project and the LearnDialogue group at NC State University. This work is supported in part by the National Science Foundation through Grant DRL-1007962 and the STARS Alliance, CNS-1042468. Any opinions, findings, conclusions, or recommendations expressed in this report are those of the participants, and do not necessarily represent the official views, opinions, or policy of the National Science Foundation.

REFERENCES

[1] Austin, J.L. 1962. *How To Do Things With Words*. Oxford University Press.

[2] Bloom, B.S. 1984. The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-one Tutoring. *Educational Researcher*. 4–16.

[3] Boyer, K.E., Vouk, M.A. and Lester, J.C. 2007. The Influence of Learner Characteristics on Task-Oriented Tutorial Dialogue. *Proceedings of AIED*, 365–372.

[4] Ha, E.Y., Grafsgaard, J.F., Mitchell, C.M., Boyer, K.E. and Lester, J.C. 2012. Combining Verbal and Nonverbal Features to Overcome the "Information Gap" in Task-Oriented Dialogue. *Proceedings of thel SIGDIAL Meeting on Discourse and Dialogue,* 247–256.

[5] Hershkovitz, A. and Nachmias, R. 2011. Online Persistence In Higher Education Web-Supported Courses. *The Internet and Higher Education*. 14, 2.,98–106.

[6] Marineau, J., Wiemer-Hastings, P., Harter, D., Olde, B., Chipman, P., Karnavat, A., Pomeroy, V., Rajan, S. and Graesser, A. 2000. Classification of Speech Acts in Tutorial Dialog. *Proceedings of the Workshop On Modeling Human Teaching Tactics And Strategies at ITS*, 65–71.