Toward Adaptive Unsupervised Dialogue Act Classification in Tutoring by Gender and Self-Efficacy

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, and statements) of student natural language utterances is a central challenge. Recently, unsupervised machine learning approaches are showing great promise; however, these models still have much room for improvement in terms of accuracy. To address this challenge, this paper presents a new unsupervised dialogue act modeling approach that leverages non-cognitive factors of gender and self-efficacy to better model students' utterances during tutorial dialogue. The experimental findings show that for females, leveraging learner characteristics within dialogue act classification significantly 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 with the help of non-cognitive factors.

Keywords

Tutorial dialogue, learner characteristics, dialogue act classification, unsupervised machine learning, adaptive 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 [7, 17, 36]. In order to model tutorial dialogue for the purposes of building tutorial systems or for studying human tutoring, *dialogue acts*, which capture both cognitive and non-cognitive aspects of dialogue utterances, 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) [3, 32]. Within the computational linguistics and dialogue systems literature, automatically classifying dialogue acts has been a focus of research for several decades [6, 14, 35]. For tutorial dialogue systems, dialogue act classification is crucial to understanding students' utterances and developing tutorial strategies [8, 24].

Today's tutorial dialogue systems utilize a variety of dialogue act classification strategies, some rule-based and some statistical [13]. 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. The same is true within the broader dialogue systems research community: dialogue act classifiers have historically either been handcrafted and rule-based, or learned with supervised machine learning techniques [11, 14, 22, 29]. 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 including hidden Markov models [20, 21, 30], Dirichlet Process clustering [12, 23], *k*-means clustering [31], and query-likelihood clustering [15] 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 non-cognitive factors, such as gender and self-efficacy, which are believed to influence the structure of tutorial dialogue [10]. Cognitive factors such as skill mastery has been widely studied in learning environments. However, there is a smaller body of work on adaptive learning environments using non-cognitive factors. A variety of learner characteristics, including non-cognitive factors, play an influential role in learning, not only in tutoring but in classroom settings [1], and in web-based courses [19]. Prior work on learner characteristics has focused on building adaptive systems based on different user groups [16], tutorial feedback selection [9] and identifying students that need remedial support [27]. Identifying clusters of student characteristics is also an active area of research [4, 25–27].

This paper investigates whether the performance of an unsupervised dialogue act classifier can be improved by taking these factors into account. Because non-cognitive factors are shown to affect language, we believe that training dialogue act classifiers tailored to specific learner characteristics can help tutorial dialogue systems to understand students better. We utilize two learner characteristics: gender, as self-reported by students on a survey and domain-specific self-efficacy, as measured by a validated instrument for determining a student's confidence in her own abilities. Specifically, we train unsupervised dialogue act models that are tailored to students of specific gender and self-efficacy level, and we compare those models to corresponding ones trained without restricting by that learner characteristic. This unsupervised training is conducted entirely without the use of manual tags. We then test all of 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 show that for female students, utilizing learner characteristics statistically significantly improves dialogue act classification models. For self-efficacy groups, improvement is observed but not at a statistically reliable level. This paper constitutes the first research toward incorporating non-cognitive factors into unsupervised dialogue act classifiers for

tutorial dialogue with the overarching goal of providing personalized learning for students. We first administered a survey to collect these characteristics via self-report, and then learned a dialogue act classifier tailored to those characteristics. These results can inform the way that next-generation tutorial dialogue systems conduct their real-time dialogue act classification and language adaptation.

2. RELATED WORK

Dialogue act modeling is an important level of representation within dialogue systems. Following theories proposed several decades ago within philosophy and linguistics [3, 32], dialogue act classification aims to capture the intention of an utterance; for example, in tutoring some dialogue acts involve asking questions or giving or requesting feedback. While a long-standing line of investigation has focused on handcrafted or supervised machine learning techniques for dialogue act classification [11, 14, 22, 29], only recently is a body of work emerging on unsupervised approaches to this problem. Most of this work has been done outside of educational domains, with a proposed hidden Markov model in the domains of Twitter posts [30] and emails [21], Dirichlet Process Mixture Models for a train fare dialogue domain [12] and for navigating buildings [23], and a Chinese Restaurant Process approach for spoken Japanese [20].

Another important difference between the current work and prior research is in the features used, namely the non-cognitive characteristics of gender and self-efficacy. Prior work has used a variety of features for performing supervised dialogue act classification, including prosodic and acoustic features which involve the profile of the sound signal itself [35], lexical features such as words and sequences of words [34], syntactic features including part-of-speech tags [6, 24], dialogue structure features such as taking the initiative and the previous dialogue act [33] as well as task/subtask features in tutorial dialogue [8, 18]. Within unsupervised dialogue act classification a subset of these features have also been used such as words [12], state transition probabilities in Markov models [23], topic words [30], function words [15], a smaller subset of words containing beginning portions of utterances [31], part-of-speech tags and dependency trees [21]. While a variety of experiments have demonstrated the utility of these features in several domains, no prior work has reported on an attempt to include the factors considered here, in order to improve the performance of an unsupervised dialogue act classifier. To investigate this, we build dialogue act classifiers that learn from utterances of specific learner groups and predict dialogue acts of students according to their learner characteristics.

3. CORPUS

The corpus used in this study consists of student-tutor interactions in an introductory computer science programming task [18]. Throughout the data collection, freshman engineering students and tutors communicated through a textual dialogue-based learning environment while working on Java programming. The ethnicity of students participated in this study is distributed as follows: 26 white, 9 Asian, 3 Latino, 2 African American, 1 Middle Eastern and 1 Asian American. An excerpt from the corpus is shown in Table 1.

Students were given a pre-survey that included survey items on computer science self-efficacy, such as 'I am sure I can learn programming'. This self-efficacy scale was adapted directly from the Domain-specific Self-Efficacy Scale [5], with five items measured on a Likert scale from 1-5 (1 being lowest self-efficacy, 5 being highest). Students also completed a demographic questionnaire from which gender was obtained. For self-efficacy, students were divided into classes based on the median score across all students on that scale. Along with gender, this produces two partitions of the 42 students: females (12) and males (30), low (24) and high self-efficacy students (18).

Table 1: Excerpt of dialogue with a *male* student in the *low self-efficacy* group

Role	Utterance	Dialogue Act
Tutor	You'll need to end every Java statement with a semi colon	*S*
Student	Got it!	*ACK*
Tutor	This is to let Java know where each statement ends	*S*
Tutor	Ah no prompt!	*S*
Tutor	Why do you think that is?	*Q*
Student	I wish I knew...	*A*
Student	I don't think I spelled anything wrong	*S*
Tutor	Ah it's actually pretty easy	*S*
Tutor	The order of the lines matters	*S*

The corpus containing 1640 student utterances was manually annotated with dialogue act tags in previous work [18] (Table 2). 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.

Table 2: Student dialogue acts and distributions

Student Dialogue Act	Example	Distribution
A (answer)	*yeah I'm ready!*	39.95%
ACK (acknowledgement)	*Alright*	21.31%
S (statement)	*i am taking basic fortran right now never seen literal before*	21.20%
Q (question)	*what does that mean?*	15.15%
RF (request feedback)	*better?*	0.98%
C (clarification)	**html messing*	0.79%
O (other)	*haha*	0.61%

4. DIALOGUE ACT MODELING BASED ON LEARNER CHARACTERISTICS

We hypothesize that dialogue act models built using unsupervised machine learning will perform substantially better when customized to specific learner groups. Specifically, we investigate whether by training a model only on students of a particular learner characteristic, that model would perform significantly better at predicting the dialogue acts of unseen students with the same learner characteristic compared to a model that was trained on students of all learner characteristics.

We note that because the same corpus is being partitioned in two different ways, the same student will occur in one of the gender groups and in one of the self-efficacy groups. This choice to partition in 2-way splits rather than *2n*-way splits where *n* is the number of learner characteristics is because of issues that arise with sparsity. This interdependence between partitions is a limitation to note; however, as discussed in Section 5, this

interdependence can be taken into account for making decisions within a tutorial dialogue system by employing a suite of classifiers within a voting scheme.

4.1 Experimental Design

For gender and self-efficacy, we will test whether an unsupervised dialogue act classifier trained only on students with that characteristic outperforms a classifier that is not specialized by this characteristic. 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 (where *n* varies depending on which learner group is being considered) consists of all of a single student's dialogue utterances and the model is trained on the remaining *n-1* students. The average number of utterances per student in the corpus is 36.8 (σ=12.07; min=16; max=64). These are therefore the average, minimum, and maximum number of utterances across the leave-one-student-out test sets.

We compute the average test set performance of the model across all folds for each non-cognitive characteristic partition. The performance metric utilized in this study is *accuracy* compared to the manually labeled dialogue acts described in the previous section, where accuracy is computed as the number of utterances in the test set that were classified according to their manual label, divided by the number of utterances total in the test set. As described in 4.2, the process of labeling via unsupervised classification involves taking the majority vote within each cluster.

For constructing the folds, we take an approach to balance the sample size available to model training. This balancing approach is needed to ensure that each model is trained on a similar size of data. Consider, for example, the partition of gender. Without a balanced sampling approach the leave-one-student-out testing folds for the un-specialized classifier for female students would include n_{female}=12 test folds but the available data for each training fold would be n_{total}-1 = 41. In contrast, the specialized classifier trained only on female students would still include n_{female}=12 test points but the available data for each training fold would be n_{female}-1 = 11. Therefore, each un-specialized classifier was trained on a randomly selected subset of the corpus. In the case of females, each of the 12 testing folds will utilize a model trained on 11 data points. The specialized classifier will use 11 female data points, and the un-specialized classifier will use 11 randomly selected data points. In this way, we investigate how well a model predicts dialogue acts of a student with and without utilizing learner characteristic information.

4.2 Unsupervised Dialogue Act Models

Our unsupervised dialogue act classification approach leverages the *k*-medoids clustering technique [28]. This approach groups similar utterances together, and is similar to the more familiar *k*-means algorithm except that in *k*-medoids, the centroid of each cluster must be an actual data point within the corpus rather than a potentially artificial data point computed as the mean of distances. Our experiments with *k*-medoids have demonstrated that it outperforms a variety of other unsupervised machine learning approaches for the task of dialogue act classification in tutorial dialogue, although the results of such experiments are beyond the scope of this paper since our goal is to investigate the *differential benefit* of adding learner characteristic features to the model, not to compare different unsupervised approaches.

The *k*-medoids algorithm requires seeding clusters at the beginning of each training fold and then proceeds by distributing data points to clusters according to their closest centroids until convergence upon the model. In the standard *k*-medoids algorithm, the seeds are randomly selected. However, we employ a greedy seed selection approach intended to mitigate the effects of the unbalanced distribution of dialogue acts in the corpus [2]. Within this greedy seed selection, an initial seed is randomly selected and then each of the subsequent seeds are selected by choosing the point that maximizes its distance from the already-selected seeds. The goal in using this approach is to select the seeds from diverse utterances so the algorithm produces better clusters, and our initial experiments indicated that it substantially improves the model.

In addition to its seeding approach, the *k*-medoids approach requires the number of clusters *k* to be set prior to model training. To discover the number of clusters, we experimented with *X*-Means and Expectation Maximization clustering, both of which attempt to identify the optimal number of clusters. Both of these algorithms converged at four clusters as the optimal choice, so we proceed with *k*=4. However, perhaps in part due to the benefit of the greedy seed selection made possible by *k*-medoids, these models performed with substantially worse overall accuracy than *k*-medoids.

The utterances were represented as vectors with each column matching a token (punctuation and words) in the corpus and each row matching an utterance. There were a total of 877 distinct tokens.

With these parameters in place, first the clusters were formed using each training set, and then for each utterance of the student held out within the leave-one-student-out fold, we computed the closest cluster to that utterance as indicated by average cosine distance to each point in the cluster. The closest cluster was selected as the cluster to which the test utterance belongs, and the majority vote of the cluster was assigned to the test utterance as its dialogue act label. For each leave-one-student-out testing fold, the accuracy was computed by comparing these cluster-assigned labels to the manual dialogue act tags.

4.3 Experimental Results

This section presents experimental results for unsupervised dialogue act classification based on learner characteristics. We compare each model built separately by gender and self-efficacy level to the models that are built using utterances from randomly selected students, *i.e.* not utilizing learner characteristic information. Each comparison in this section is conducted with a one-tailed *t*-test with a post-hoc Bonferroni correction. The threshold for statistical reliability after the correction has been taken as α=0.05.

Gender. As shown in Figure 1, the average leave-one-student-out cross-validation accuracy for the model built using female students' utterances (n_{female}=12) is higher than the model built on randomly selected students. In each test run, all of one female's utterances were left out to be used as the test set, and the dialogue act model was built on the remaining eleven female students' utterances. This process was repeated for each female student. Note that for each of the eleven students, all utterances from that student were considered. Average test set accuracy for the model with randomly selected students was 0.41 (σ=0.2), whereas the average test set accuracy for the dialogue act classification model that was built utilizing female students' utterances only was 0.56 (σ=0.19). After a Bonferroni correction this difference was statistically significant (p_{Bonf}<0.05).

For male students (n_{male}=30), the average accuracy is only slightly higher with the models tailored to males 0.43 (σ =0.13) than the models learned for randomly selected students 0.40 (σ =0.12), and this difference is not statistically significant (Figure 1). Looking more closely at the results, we find that for eight of the thirty males within the corpus, a tailored model outperformed the random model (with five of these seeing more than 10% increase in accuracy), while twenty-two of the cases saw no difference in accuracy between the random and tailored conditions. Two of the males saw a decrease in accuracy for the tailored condition.

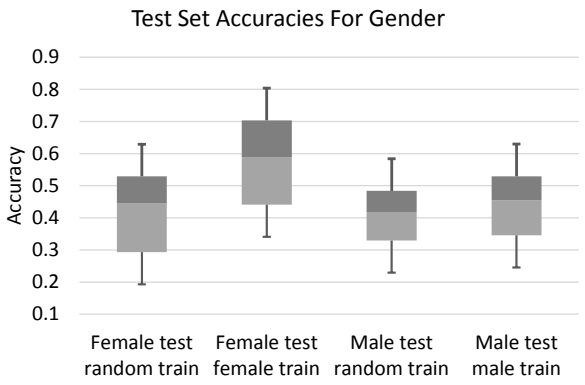

Figure 1: Leave-one-student-out test set accuracies for models by gender

Self-efficacy. Models built using the self-efficacy learner characteristic predict the unseen utterances' dialogue acts marginally more successfully than models that do not use this information, though these differences are not statistically reliable. For students with low self-efficacy (n_{lowEff}=24) the average test set accuracy for dialogue act models that selected students randomly is 0.38 (σ=0.16) and it increases to 0.43 (σ=0.17) with dialogue act models that learn only from low-self-efficacy students' utterances (Figure 2). In fifteen out of twenty-four cases the dialogue act models tailored to low self-efficacy groups outperform models that are trained on randomly selected students (eight of the cases with more than a 10% increase), while in seven of the cases the performance is decreased by utilizing the learner characteristic (five of them by more than a 5%) and in two of the cases the accuracy remains the same.

The improvement obtained by utilizing learner characteristics in dialogue act classification task is also marginal for high-self-efficacy students, where $n_{highEff}$=18. The average performance for the random model is 0.41 (σ=0.14) whereas the model achieves 0.47 (σ=0.11) accuracy when trained only on utterances of high-self-efficacy students. This improvement was statistically significant before Bonferroni correction but not afterward. In seven out of eighteen cases, models trained on utterances of high self-efficacy students improved test set accuracy (five of them above 15% improvement) and in two of the cases the learner characteristic decreases the performance (both of them below 5% decrease). Nine of the cases remained unaffected in their dialogue act classification accuracy.

The average accuracies over the leave-one-student-out cross-validation folds can be found in Table 3. Models tailored to learner groups uniformly outperform their counterpart, and the improvement is statistically significant for females.

Figure 2: Leave-one-student-out test set accuracies for models by self-efficacy

Table 3: Average test set accuracies for each learner characteristic (p<0.05 after Bonferroni correction)**

Learner characteristic group	Model restricted by learner characteristic	Model built on randomly selected students
Females	**0.56****	0.41
Males	**0.43**	0.40
Low self-efficacy	**0.43**	0.38
High self-efficacy	**0.47**	0.41

5. DISCUSSION

Dialogue act classification is a central task for tutorial dialogue systems. Without accurate dialogue act classification, systems cannot adapt and respond appropriately. Unsupervised machine learning approaches to dialogue act classification are a highly promising new area of study, and we have presented the first unsupervised dialogue act classifier tailored to learner characteristics. The experimental results demonstrated that dialogue act classifiers that leverage the non-cognitive factors of gender and self-efficacy outperform those that do not, and in the case of female students the improvement was statistically significant. This section presents some examples of the learned dialogue act clusters and discusses the implications of this work for tutorial dialogue systems.

First, we examine clusters from the gender-tailored unsupervised dialogue act classifier. Table 4 displays a selection of utterances that were clustered together during the unsupervised training of the model, and afterward the clusters were labeled for testing purposes using the manual tags that comprise the majority of each cluster. For those in Table 4 the clusters were labeled as Acknowledgments and Questions. By examining the structure of these clusters we gain some intuition as to the types of regularities that help the tailored models to perform significantly better. We see females in this study tended to use acknowledgment phrases such as, "oh I see" and "makes sense," while males tended to use the phrasing, "got it" more frequently. Within the cluster labeled as questions, we observe that females tended to request more feedback, an observation that also emerged in prior work within a different corpus in the same domain collected approximately six years earlier [10]. On the other hand, male students tended to ask more general questions.

In addition, we observe some example clusters from the models based on self-efficacy in Table 5. Students with high self-efficacy tend to use more confident utterances such as "absolutely" compared to "ok" used by low-self efficacy students. We note that questions in the low self-efficacy group often make an implicit

request for reassurance within their task-based questions, such as, "and that is it?". In contrast, students in the high self-efficacy group more often ask contentful questions.

Table 4: Selected utterances from clusters tailored to gender

	Females	Males
Acknowledgements	- oh I see - make sense - yup - aha! -hahaha its ok	- got it - ok i got it - alright i got it - gotcha alrigth - cool - sure thing
Questions	-is this right? -does that work? -should I run it? -was i supposed to put that before something? -so for line number could i have typed system out println monopoly instead of println x if i wanted to?	-so will testing always be related to running the program -so it is kinda like saying x number or something in algebra? -why does not it stop on the next line in this case

Table 5: Selected utterances from clusters tailored to self-efficacy

	Low Self-Efficacy	High Self-Efficacy
Acknowledgements	- ok - yes there were a lot of things i felt like i had to switch around - that makes sense now	-cool! -oh ok that works - yep got that - absolutely
Questions	-so what exactly am i supposed to be doing? - is there something specific i need to call my game - i finished reading should i click compile again? -and that is it?	-what is the best way to do that? - ok so tell me if this makes sense string declares the variable and then line number tells me what that variable is value is?

Limitations. The present work has several notable limitations. First, as mentioned previously, the partitions of the corpus are not independent; that is, the same student, and associated utterances, are present within one gender group and one self-efficacy group. Because these partitions are not independent, care must be taken when interpreting the findings. Furthermore, it is possible that the self-efficacy of students can change in the course of tutoring, which would not be handled by a classifier built using a one-time self-report. However, we believe that the current approach holds great promise for real-time tutorial dialogue classification. By building separate classifiers by learner characteristic, a suite of classifiers (each smaller and faster than one built on the entire corpus) can be run in parallel and can vote for the classification of a given students' utterance. However, as is the case with the work presented here, splitting the corpus results in a substantially reduced sample size on which to train, which partially explains the lack of statistically reliable results observed here. Our work has begun to explore the use of intrinsic metrics for accuracy (rather than relying on manual tags), which has the potential to dramatically increase the available data to any dialogue act classifier and mitigate issues of sparsity that arise when splitting by learner characteristics.

6. CONCLUSION AND FUTURE WORK

More accurately understanding student natural language within intelligent tutoring systems is a critical line of investigation for tutorial dialogue systems researchers. 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 non-cognitive factors within such a dialogue act classification model, achieving statistically significant improvements in dialogue act modeling for female students, and increasing the models' performance by small margins for the self-efficacy groups.

Building upon these first steps, there are several promising future directions. First, while sample size prohibited exploring some other learner characteristics here, other characteristics are likely highly influential and should be investigated. These may include ethnicity, personality, and other non-cognitive factors. Additionally, while the current work focused on analyzing dialogue, another aspect of the tutorial interaction that presents challenges in understanding is the task model. Models that aim to understand students' problem-solving activities and infer their goals or plans may benefit substantially from leveraging learner characteristics. 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 improve the student experience and enhance learning by adaptation.

ACKNOWLEDGMENTS

The authors wish to thank the members of the Center for Educational Informatics at North Carolina State University for their helpful input. 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] Ames, C. and Archer, J. 1988. Achievement Goals in the Classroom: Students' Learning Strategies and Motivation Processes. *Journal of Educational Psychology*. 80, 3, 260–267.

[2] Arthur, D. and Vassilvitskii, S. 2007. k-means++: The Advantages of Careful Seeding. In *Proceedings Of The Eighteenth Annual ACM-SIAM Symposium On Discrete Algorithms*. 1027–1035.

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

[4] Azarnoush, B., Bekki, J.M. and Bernstein, B.L. 2013. Toward a Framework for Learner Segmentation. *JEDM*. 5, 2, 102–126.

[5] Bandura, A. 2006. Guide for Constructing Self-Efficacy Scales. *Self-efficacy Beliefs Of Adolescents*. 5, 307–337.

[6] Bangalore, S., Di Fabbrizio, G. and Stent, A. 2008. Learning the Structure of Task-Driven Human-Human Dialogs. *IEEE Transactions on Audio, Speech and Language Processing*. 16, 7, 1249–1259.

[7] Bloom, B.S. 1984. Sigma of Problem: The Methods Instruction One-to-One Tutoring. *Educational Researcher*. 4–16.

[8] Boyer, K.E., Ha, E.Y., Phillips, R., Wallis, M.D., Vouk, M.A. and Lester, J.C. 2010. Dialogue Act Modeling in a Complex Task-Oriented Domain. In *Proceedings of SIGDIAL*. 297–305.

[9] Boyer, K.E., Phillips, R., Wallis, M., Vouk, M. and Lester, J. 2008. Balancing Cognitive and Motivational Scaffolding in Tutorial Dialogue. In *Proceedings of ITS*, 239–249.

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

[11] Buckley, M. and Wolska, M. 2008. A Classification of Dialogue Actions in Tutorial Dialogue. In *Proceedings of the 22nd International Conference on Computational Linguistics*. 1, 73–80.

[12] Crook, N., Granell, R. and Pulman, S. 2009. Unsupervised Classification of Dialogue Acts Using a Dirichlet Process Mixture Model. In *Proceedings of SIGDIAL*. 341–348.

[13] Dzikovska, M.O., Farrow, E. and Moore, J.D. 2013. Combining Semantic Interpretation and Statistical Classification for Improved Explanation Processing in a Tutorial Dialogue System. In *Proceedings of AIED*. 279–288.

[14] Eugenio, B. Di, Xie, Z. and Serafin, R. 2010. Dialogue Act Classification , Higher Order Dialogue Structure, and Instance-Based Learning. *Dialogue & Discourse*. 1, 2, 1–24.

[15] Ezen-Can, A. and Boyer, K.E. 2013. Unsupervised Classification of Student Dialogue Acts With Query-likelihood Clustering. In *Proceedings of EDM*, 20–27.

[16] Forbes-Riley, K. and Litman, D.J. 2009. A User Modeling-Based Performance Analysis Of A Wizarded Uncertainty-Adaptive Dialogue System Corpus. In *Proceedings of INTERSPEECH*, 2467–2470.

[17] Graesser, A.C., Wiemer-Hastings, K., Wiemer-Hastings, P. and Kreuz, R. 1999. AutoTutor: A Simulation Of A Human Tutor. *Cognitive Systems Research*. 1, 1, 35–51.

[18] 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. In *Proceedings of SIGDIAL*, 247–256.

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

[20] Higashinaka, R., Kawamae, N., Sadamitsu, K., Minami, Y., Meguro, T., Dohsaka, K. and Inagaki, H. 2011. Unsupervised Clustering of Utterances Using Non-Parametric Bayesian Methods. In *Proceedings of INTERSPEECH*, 2081–2084.

[21] Joty, S., Carenini, G. and Lin, C.-Y. 2011. Unsupervised Modeling Of Dialog Acts In Asynchronous Conversations. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence*. 1807–1813.

[22] Keizer, S., Akker, R. and Nijholt, A. 2002. Dialogue Act Recognition with Bayesian Networks for Dutch Dialogues. In *Proceedings of the SIGDIAL Workshop*, 88–94.

[23] Lee, D., Jeong, M., Kim, K., Ryu, S. and Geunbae, G. 2013. Unsupervised Spoken Language Understanding for a Multi-Domain Dialog System. *IEEE Transactions On Audio, Speech, and Language Processing*. 21, 11, 2451–2464.

[24] 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. In *Proceedings of the Workshop On Modeling Human Teaching Tactics And Strategies at ITS*. 65–71.

[25] Meece, J.L. and Holt, K. 1993. A Pattern Analysis Of Students' Achievement Goals. *Journal Of Educational Psychology*. 85, 4, 582–590.

[26] Merceron, A. and Yacef, K. 2003. A Web-Based Tutoring Tool With Mining Facilities to Improve Learning and Teaching. In *Proceedings of AIED*, 201–208.

[27] Merceron, A. and Yacef, K. 2005. Clustering Students To Help Evaluate Learning. *Technology Enhanced Learning*. 171, 31–42.

[28] Ng, R.T. and Han, J. 1994. Efficient and Effective Clustering Methods for Spatial Data Mining. In *Proceedings of the 20th International Conference on Very Large Data Bases*, 144–155.

[29] Reithinger, N. and Klesen 1997. Dialogue Act Classification Using Language Models. In *Proceedings of EuroSpeech*, 2235–2238.

[30] Ritter, A., Cherry, C. and Dolan, B. 2010. Unsupervised Modeling of Twitter Conversations. In *Proceedings of the Association for Computational Linguistics*, 172–180.

[31] Rus, V., Moldovan, C., Niraula, N. and Graesser, A.C. 2012. Automated Discovery of Speech Act Categories in Educational Games. In *Proceedings of EDM*, 25–32.

[32] Searle, J.R. 1969. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press.

[33] Serafin, R. and Di Eugenio, B. 2004. FLSA: Extending Latent Semantic Analysis With Features For Dialogue Act Classification. In *Proceedings of the Association for Computational Linguistics*, 692–699.

[34] Sridhar, V.K.R., Bangalore, S. and Narayanan, S.S. 2009. Combining Lexical, Syntactic and Prosodic Cues For Improved Online Dialog Act Tagging. *Computer Speech & Language*. 23, 4, 407–422.

[35] Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., Taylor, P., Martin, R., Ess-Dykema, C. Van and Meteer, M. 2000. Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech. *Computational Linguistics*. 26, 3, 339–373.

[36] VanLehn, K., Jordan, P.W., Rosé, C.P., Bhembe, D., Bottner, M., Gaydos, A., Makatchev, M., Pappuswamy, U., Ringenberg, M., Roque, A., Siler, S. and Srivastava, R. 2002. The Architecture Of Why2-Atlas: A Coach For Qualitative Physics Essay Writing. In *Proceedings of ITS*, 158–167.