Unsupervised Classification of Student Dialogue Acts With Query-Likelihood Clustering

Aysu Ezen-Can
Department of Computer Science
North Carolina State University
Raleigh, North Carolina 27695
aezen@ncsu.edu

Kristy Elizabeth Boyer
Department of Computer Science
North Carolina State University
Raleigh, North Carolina 27695
keboyer@ncsu.edu

ABSTRACT

Dialogue acts model the intent underlying dialogue moves. In natural language tutorial dialogue, student dialogue moves hold important information about knowledge and goals, and are therefore an integral part of providing adaptive tutoring. Automatically classifying these dialogue acts is a challenging task, traditionally addressed with supervised classification techniques requiring substantial manual time and effort. There is growing interest in unsupervised dialogue act classification to address this limitation. This paper presents a novel unsupervised framework, query-likelihood clustering, for classifying student dialogue acts. This framework combines automated natural language processing with clustering and a novel adaptation of an information retrieval technique. Evaluation against manually labeled dialogue acts on a tutorial dialogue corpus in the domain of introductory computer science demonstrates that the proposed technique outperforms existing approaches. The results indicate that this technique holds promise for automatically understanding corpora of tutorial dialogue and for building adaptive dialogue systems.

Keywords

Tutorial dialogue, dialogue act modeling, unsupervised machine learning

1. INTRODUCTION

Tutorial dialogue systems are highly effective at supporting student learning [1, 8, 9, 11, 13, 14, 20]. However, these systems are time-consuming to build because of the substantial engineering effort required within their various components. For example, understanding and responding to the rich variety of student natural language input has been the focus of great attention, addressed by a variety of techniques including latent semantic analysis [15], enriching natural language input with spoken language capabilities [21], linear regression for assessing correlation of dialogue acts with learning [8] and integration of multiple dialogue policies [12]. However, a highly promising approach is to automatically mine models of user utterances from corpora of dialogue using machine learning techniques [16, 24].

A task of particular importance in modeling student utterances is determining the *dialogue act* of each utterance [25, 28]. The premise of dialogue act modeling is that it captures the communicative goal or action underlying each utterance, an idea that emerged within linguistic theory and has been leveraged with great success by dialogue systems researchers [2][27]. Dialogue act modeling, in practice, is based on creating taxonomies to use in dialogue act classification. Within tutorial dialogue systems, first the dialogue act for a student utterance is inferred, and this label serves as the basis for selecting the next tutorial strategy.

There are two approaches for learning dialogue act models from a corpus: supervised and unsupervised. Supervised models require a manually labeled corpus on which to train, while unsupervised models employ machine learning techniques that rely solely on the structure of the data and not on manual labels. A rich literature on supervised modeling of dialogue acts has shown success in this task by leveraging a variety of lexical, prosodic, and structural features [29, 30]. However, supervised models face two significant limitations. First, manual annotation is a time-consuming and expensive process, a problem that is compounded by the fact that many annotation schemes are domain-specific and must be re-engineered for new corpora. Second, although there are standard methods to assess agreement of different human annotators when applying a tagging scheme, developing the tagging scheme in the first place is often an ill-defined process. In contrast, unsupervised approaches do not rely on manual tags, and construct a partitioning of the corpus that suggests a fully data-driven taxonomy. Unsupervised approaches have only just begun to be explored for dialogue act classification, but early results from the computational linguistics literature suggest that they hold promise [10, 24], and a very recent finding in the educational data mining literature has begun to explore these techniques for learning-centered speech [25].

This paper presents a novel approach toward unsupervised dialogue act classification: *query-likelihood clustering*. This approach adapts an information retrieval (IR) technique based on query likelihood to first identify utterances that are similar to a target utterance. These results are then clustered to identify dialogue acts within a corpus in a fully unsupervised fashion. We evaluate the proposed technique on a corpus of task-oriented tutorial dialogue collected through a textual, computer-mediated dialogue study. How best to evaluate unsupervised techniques is an open research question since there is no "perfect" model that the results can be compared to. We therefore examine two complementary evaluation criteria that have been used in prior work: quantitative evaluation with respect to manual labels [10, 25], and detailed qualitative inspection of the clustering to determine whether it learned "natural" groupings of utterances [24]. The results demonstrate that query-likelihood clustering performs significantly better than majority baseline chance compared to manual labels. In addition, the proposed algorithm outperforms a recently reported unsupervised approach for speech act classification within a learning-centered corpus. Finally, qualitative analysis suggests that the clustering does group together many categories of utterances in an intuitive way, even

highlighting in a fully data-driven fashion some ways in which the original hand-authored dialogue act taxonomy could be revised and improved in the future.

2. RELATED WORK

Dialogue act classification aims to model the intent underlying each utterance. Supervised dialogue act modeling has been well studied in the computational linguistics literature, applying techniques such as Hidden Markov Models [30] and Maximum Entropy classifiers [4][29]. For tutorial dialogue, promising approaches have included an extension of latent semantic analysis [28], a syntactic parser model [22], and vector-based classifiers [6].

Compared to the rich body of work on supervised dialogue act modeling, a much smaller body of work has focused on unsupervised approaches. A recent non-parametric Bayesian approach used Dirichlet Process Mixture Models [10], which attempt to identify the number of clusters non-parametrically. Another recent work on unsupervised classification of dialogue acts modeled a corpus of Twitter conversations using Hidden Markov Models combined with a topic model built using Latent Dirichlet Allocation [24]. This corpus was composed of small dialogues about many general subjects discussed on Twitter. In order for the dialogue act model not to be distracted by different topics, they separated content words from dialogue act cues with the help of the topic model. In our tutoring corpus, however, the content words reveal important information about dialogue acts. For example, the word "help" is generally found in utterances that are requesting a hint. Therefore, our model retains content words.

Rus et al. utilize clustering to classify dialogue acts within an educational corpus [25], forming vectors of utterances using the leading tokens (words and punctuation marks), and using string comparison as the similarity metric. As they mention, this string comparison may not be sufficient to generalize word types used within the same context. For example, 'hello' and 'hi' are different according to string comparison; however, they are part of the same dialogue act, in that they both serve as a greeting. Our clustering approach uses query likelihood to group similar words that can be used for the same intention, and we use a blended part-of-speech tag and word feature set which overcomes the challenge introduced by string comparisons. The results suggest that these extensions improve upon existing clustering techniques.

3. TUTORING CORPUS

The corpus consists of dialogues collected between pairs of tutors and students collaborating on the task of solving a programming problem as part of the JavaTutor project during spring 2007. The tutor and student interacted remotely with textual dialogue through computer interfaces. There were forty-three dialogues in total, with 1,525 student utterances (averaging 7.54 words per utterance) and 3,332 tutor utterances (averaging 9.04 words per utterance). This paper focuses on classifying the dialogue acts of student utterances only. Within an automated tutoring system, tutor utterances are system-generated and their dialogue acts are therefore known. The corpus was manually segmented and annotated with dialogue acts, one dialogue act per utterance, during prior research that focused on supervised dialogue act annotation and dialogue structure modeling [7]. While the manual dialogue act labels are not used in model training, they are used to evaluate the unsupervised clustering. Table 1 shows manually labeled tags and their frequencies. The Kappa for agreement on these manual tags was 0.76. An excerpt from the corpus is presented in Table 2.

Table 1: Dialogue act tags with examples and student frequencies from corpus

Tag	Act	Description	Freq
Q	Question	*A general question which is not specific to the task*	276
EQ	Evaluation Question	*A question about the task*	416
S	Statement	*A statement of fact*	211
G	Grounding	*Acknowledgement of previous utterance*	192
EX	Extra-Domain	*Any utterance that is not related to the task*	133
PF	Positive Feedback	*Positive assessment of knowledge or task*	116
NF	Negative Feedback	*Negative assessment of knowledge or task*	92
LF	Lukewarm Feedback	*Assessment having both positive and negative assessments*	32
GRE	Greeting	*Greeting words*	57

Table 2: Excerpt from the corpus (typographical errors originated in corpus)

	Utterance	Tag
Student:	so obviously here im going to read into the array list and pull what we have in the list so i can do my calculations	*S*
Tutor:	something like that, yes	*LF*
Tutor:	by the way, an array list (or ArrayList) is something different in Java. this is just an array.	*S*
Student:	ok	*G*
Student:	im sorry i just refer to it as a list because thats what it reminds me it does	*S*
Student:	stores values inside a listbox(invisible)	*S*
Tutor:	that's fine	*EX*
Tutor:	ok, so what are we doing here?	*EQ*
Student:	im not sure how to read into the array	*NF*

4. QUERY-LIKELIHOOD CLUSTERING

This section describes our novel approach of adapting information retrieval (IR) techniques combined with clustering to the task of unsupervised dialogue act classification. IR is the process of searching available resources to retrieve results that are similar to the query [3]. IR techniques are mostly used in search engines to retrieve results that are similar to given queries. In the proposed approach, the target utterance that is to be classified is used as a query and its similar utterances are gathered using query likelihood. Then, the query likelihood results are provided to the clustering technique.

4.1 Natural Language Preprocessing

At its core, query-likelihood information retrieval operates at the token, or word, level. In order to prepare the corpus for this application, several preliminary experiments were conducted to determine the appropriate type of preprocessing. It was observed that preprocessing is a crucial step in order to increase the discriminating cues extracted from the corpus.

Part-of-speech (POS) tagging is a technique for labeling each word according to its grammatical part of speech such as noun, verb, and adjective. This procedure allows us to generalize words to their functionalities in sentences. For example, the pronouns 'you' and 'it' are grouped in the same POS tag: PRP standing for personal pronoun. The generalization provided by this part-of-speech backoff can be useful in dialogue act classification [5, 6, 28]. We experimented on querying with both actual words and with full part-of-speech backoff. The best results were produced by a combination of words and POS tags. This hybrid approach replaces function words such as determiners ('the', 'a'), conjunctions ('and', 'but'), and prepositions ('in', 'after') with their POS tags. Content words were retained but stemmed (e.g., 'parameter' becomes 'paramet', 'completely' becomes 'complet') to reduce the number of distinct words in the vocabulary of the corpus under consideration. This choice was motivated by the observation that in this task-oriented domain, important information about the dialogue act resides in content words. For instance, the word 'confused' reveals important information about the state in which the student is and it is likely that the student might be requesting a hint.

It was noted that in this domain of computer science tutoring, the natural language contains special characters that indicate a semantically important entity related to the domain, such as short bits of programming code. Although they are important with regard to the tutoring task, they require additional preprocessing in order to be handled appropriately by automated natural language processing techniques. Therefore, code segments in the corpus were replaced with meaningful tags representing them. For instance, segments about array indexing, which may originally have appeared as 'x[i]' and been mishandled, were replaced with the text 'ARRAY_INDEXING'. If-statements, loops and arithmetic operations were all replaced in the corpus using similar conventions. All procedures in natural language processing is automated using parser therefore, the human-intervention was in deciding on the procedure to use (retain content words, replace function words) not in its implementation.

4.2 Query-Likelihood Representation

The query likelihood model treats each utterance as a document. The target utterance whose dialogue act is to be predicted becomes the query, and we apply information retrieval by querying the target utterance in the corpus. This query produces a ranked list of documents, from most likely to least likely, and this list is used to identify those utterances that are most similar to the target. The query likelihood implementation from the Lemur Project was used in this work [31].

The ordering of words contains important information about the structure of utterances. For example, the word pair 'I am' is mostly found in statements whereas if we regard them separately, 'I' can belong to a question such as 'What should I do next?' or 'am' can be part of a evaluating question 'Am I doing this correct?' (After preprocessing 'I am' becomes 'PRP (personal pronoun) VBP (present tense singular verb, non third person)' however, the same issue applies to the POS tags as well.) Because of the importance of word ordering on inferring the structure of utterances, we utilized a modified query approach that considers bigrams (pairs of adjacent words occurring in each utterance) rather than unigrams (individual words).

Table 3 displays several original utterances, their modified forms after POS backoff and stemming, and the query combinations that were submitted to the algorithm. The POS tag VBZ represents third person present tense singular verbs, DT represents determiners, TO is the same as the word 'to', VBD stands for past tense verbs, WDT and WRB are interrogatives, and MD represents modal verbs. Figure 1 shows an example query with a question and its similar utterances retrieved.

Table 3: Original utterances, their processed versions and query combinations

Utterance	POS+ stemming	Query combination
I'm reading it right now	VB read PRP right now	(VB read) (read PRP) (PRP right) (right now)
what is the basic structure to begin an array?	WDT VBZ DT basic structur TO begin DT array?	(WDT VBZ) (VBZ DT) (DT basic) (basic structur) (structur TO) (TO begin) (begin DT) (DT array) (array ?)
that was correct	WDT VBD correct	(WDT VBD) (VBD correct)
how do you think you should start it?	WRB do PRP think PRP MD start PRP?	(WRB do) (do PRP) (PRP think) (think PRP) (PRP MD) (PRP start) (start PRP) (PRP ?)

Query:

How can I solve this problem?

Query Likelihood results:

How can I do addition?

What would be the results?

Which should go first?

Figure 1: Sample query and its results

4.3 Clustering

The similarity results from querying were used as the distance metric in a *k*-means clustering algorithm. The implementation of this idea relies on creating binary vectors for similar utterances and then grouping those vectors. Each utterance that is present in the similarity list is represented as a 1, while the others are represented with a value of 0. In this way, each target utterance in the corpus is represented by a vector indicating the utterances that are similar to it. The entire unsupervised dialogue act classification algorithm is summarized in Figure 2.

Let D be a corpus of utterances $D = \{u_1, u_2 \dots u_n\}$
Then the goal is:

$\forall u_i \in D$, identify l_i = dialogue act label of u_i

Procedure:

For each u_i

1. Set target utterance $q_i = u_i$
2. Let the query likelihood result of utterance u_i be $R = (u_t, u_{t+1} \dots u_z)$ such that R is the result of *queryLikelihood*(q_i)
3. Create vector of query results indicator variables $V_i = (v_1, v_2 \dots v_j \dots v_n)$ such that

   $v_j = 1$ if $u_i \in R$

   else $v_j = 0$

Let the total vector be $V_T = (V_1, V_2, \dots, V_n)$

Return clusters $C = \{c_1, c_2 \dots c_k\}$ such that C is the result of *k-means*(V_T)

Figure 2: Query-likelihood clustering algorithm.

5. EXPERIMENTS

The goal of the experiments is to apply the novel unsupervised technique of query-likelihood clustering to discover student dialogue act clusters within the corpus of tutorial dialogue. We utilize a two-pronged evaluation consisting of quantitative comparison in terms of accuracy on manual labels, as well as qualitative examination of the resulting clusters. This section first presents the model-learning process, including parameter tuning on a development set, and then presents quantitative and qualitative evaluations on the remainder of the corpus. We also compare performance of the proposed approach to a state-of-the-art unsupervised technique for speech act labeling in a learning-centered corpus.

5.1 Parameter Tuning

In order to train the unsupervised model, three parameters had to be set. Two of these parameters are used within the query phase and the last one applies to the clustering phase. The two parameters to be determined in the query phase are *b*, the blind relevance feedback threshold, and *n,* the number of top query likelihood results to be used while creating vectors for clustering. The parameter related to the clustering phase is the distance metric. In order to tune these parameters, a 25% validation set, constituting 10 randomly selected sessions, was drawn from the corpus. The parameter tuning was conducted in a sequential manner that allows the latter steps to use already fixed parameters.

Token weighting. Prior to tuning the parameters, an additional optional set of token weights was tuned for use during querying. This optional parameter was desirable based on observations that some POS tags should be weighted more than others for identifying similar results to a target utterance. For example, interrogatives (question words such as *what*, *where*, *when*, *how*, and *why*) and question marks are highly discriminating for question dialogue moves. Weights were learned for this subset of tokens using an incremental approach as shown in Table 4. First, the mean average precision values of query likelihood results without any weights were given. Then, weights for WDT (POS tag for *what* and *which*) were utilized within the experiments. Having determined the proper weight for WDT, different weights for WRB (POS tag for *why, where, how, when*) were tried. Finally, the weight for question marks was set.

Table 4: Mean average precision (MAP) results for weighting interrogative parts of speech and punctuation

Weights	MAP
no weight	0.1239
WDT = 10	**0.2359**
WDT = 100	0.2326
WRB = 10	**0.2358**
WRB = 100	0.2339
'?' = 10	0.2457
'?' = 100	**0.2567**

Relevance feedback threshold (*b*). In a typical query likelihood scenario, relevance feedback on the retrieved results is provided by human users and used to improve the model performance. However, in a fully unsupervised scenario, human relevance feedback is not available. Our unsupervised approach utilizes pseudo-relevance feedback (blind relevance feedback), which assumes that the top *b* documents retrieved are relevant to the query [26]. Taking the top *b* documents into account, the algorithm automatically finds words that are important for those documents and therefore may be useful for the query. In order to find the important words, relevance models for each retrieved document in the ranked list are computed, where a relevance model is the probability of features used in the query given the document. In our experiments, the features are composed of bigrams, which are pairs of adjacent words within an utterance. Therefore, the relevance model of a document is the probability of its bigrams given the whole utterance. Then, relevance models of the top *b* documents are sorted and the top terms are determined, which are used to expand the original query. Having chosen those words, the algorithm expands the initial query by appending the newly found terms and running the query again. This procedure of enriching the query with top relevant results is done in order to avoid sparse ranked lists. Like the other parameters described in this section, *b* was tuned on a development set. Table 5 shows the resulting best *b*-value of 30.

Table 5: *b*-value MAP results

b	**5**	**10**	**15**	**20**	**25**	**30**	**35**
MAP	0.343	0.34	0.342	0.346	0.351	**0.352**	0.351

Top query threshold (*n*). Given the *b*-value, we moved on to tuning *n*-values, which represent the number of top query likelihood results to be used in forming the vectors for subsequent clustering. A higher *n* value will treat a larger number of utterances retrieved during querying as "similar" to the target. A minimum of *n*=5 was chosen to sufficiently populate the vectors since the non-zero entries determine clusters, and we explored whether larger *n* values performed better; however, the optimal value was *n*=5 as shown in Table 6.

Table 6: *n*-value MAP results with set *b*-value

n-value	**5**	**10**	**15**	**20**	**30**	**100**
MAP	**0.517**	0.432	0.409	0.398	0.395	0.307

Distance metric. We experimented with multiple distance metrics to determine which metric performed best within this novel context. The candidate distance metrics included the widely used Euclidean and Manhattan distances. Also considered was another widely used similarity metric in text mining, cosine similarity, which increases as the two vectors have non-zero entries in the same positions. This approach is particularly effective for sparse vectors. The final candidate was Borda count [18], a metric that weights non-zero entries according to their ranks in query likelihood. In our scenario, the first utterance retrieved by query likelihood gets the highest rank. This approach reforms vectors in a weighted manner, after which Euclidean distance is applied.

Within the development set, cosine distance performed with highest accuracy at 43.43% compared to manual labels. Borda count achieved 42.63%, Euclidean distance 41.64%, and Manhattan distance 41.82%. Since cosine similarity is computed by the dot product of two vectors divided by the product of their magnitudes, it takes the size of vectors into account and provides a ratio of matching 1 values in the same position in both vectors. In this way, intersecting non-zero entries are valued with respect to the size of the vectors.

5.2 Accuracy in Identifying Manually Labeled Dialogue Acts

Having tuned the parameter values using the development set, we trained a model on the remainder of the corpus. This unsupervised model did not take manual labels into account during training, but we evaluate its performance with respect to manual labels. Due to the nature of unsupervised approach, the optimal number of clusters was not known. Therefore, we explore varying numbers of clusters using *k*-Means, a standard clustering algorithm that aims to cluster observations into *k* clusters so that each element has the highest similarity to every other element in the cluster [19]. In addition to *k*-means clustering, we experimented with a non-parametric Bayesian approach used in recent unsupervised dialogue act classification work in a non-tutoring domain [10].

We calculate the accuracy of the approach for classification by comparing to manual labels. Figure 3 presents the accuracy of the proposed approach using the *k*-means clustering algorithm in Weka [17]. To label a target utterance, the query likelihood results were retrieved for that utterance, and then its vector was provided to the clustering algorithm. The resulting cluster in which the target utterance resides was interpreted as the dialogue act label of the target utterance. The majority label of the cluster was taken as its dialogue act label. For comparison, the accuracy results are compared against the majority class baseline of 25.87%, Evaluation Question (EQ). This is the most commonly occurring student dialogue act in the corpus and therefore represents the expected accuracy of a model that performs equal to chance. Additionally, we compare our implementation against that of the recent approach of Rus et al. [25], which clusters utterances via word similarity using a specified number of leading tokens of each utterance.

The highest accuracy achieved by query-likelihood clustering was 41.84% with k=8. This accuracy therefore constitutes a 61.9% improvement over baseline chance. We experimented with the Rus et al. leading-tokens clustering using two to five leading tokens as they suggest. Five leading tokens performed best, yielding 34.90% accuracy at k=7.

In order to provide a comparison across corpora, we consider the results from Rus et al. on their corpora of educational games. Their highest accuracy was 37.9%, compared with a majority class baseline of 28.5% (statements). This is a 32.98% improvement over the majority baseline chance. Another algorithm tried by Rus et al. was Expectation Maximization, which achieved 37.9% accuracy on their combined corpus of educational games. This algorithm achieved its highest accuracy of 30.47% with four leading tokens on our corpus. We also experimented with Dirichlet process clustering as used by Crook et al. [10]. Dirichlet process clustering gave substantially lower results on our corpus after natural language preprocessing, with an accuracy of 24.48% compared to the 41.84% of query-likelihood clustering.[1]

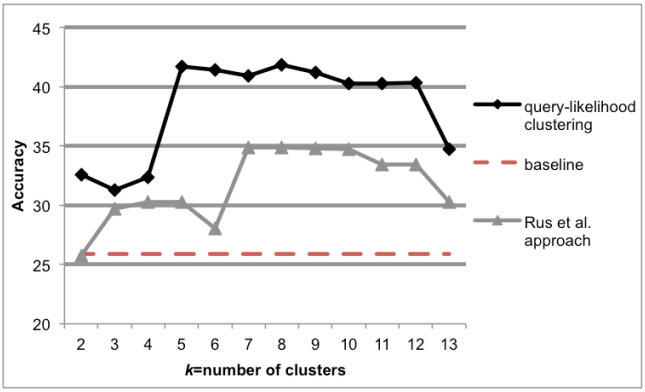

Figure 3: Accuracy results (%) for query-likelihood clustering, Rus et al. approach with 5 leading tokens and majority baseline.

Finally, in order to evaluate the relative contribution of the query-likelihood clustering components of automatic natural language preprocessing (POS tagging, stemming) and vector enhancement with information retrieval, we performed two experiments omitting each of these components from the procedure. Parameters were re-tuned for each experiment utilizing the same development set split used earlier (25% of the corpus). With the best-performing model size of k=8 clusters, omitting the natural language preprocessing step produced accuracy of 37.59% and omitting the information retrieval step produced accuracy of 35.68%, each of which is substantially lower than the query-likelihood clustering approach of 41.84%. However, these simpler approaches resulted in a smaller number of clusters for their best-fit models, achieving their highest accuracies of 41.06% and 40.45%, respectively, when k=4. This accuracy is only modestly lower than the query-likelihood clustering accuracy of 41.84%; however, a significant drawback is that the number of clusters is much smaller than are typically distinguished by dialogue act classification schemes, and would likely not result in sufficiently fine-grained distinctions to support dialogue management.

5.3 Analysis of Clusters

As shown in Table 7, the approach was particularly successful in clustering negative feedback utterances (NF), evaluation questions (EQ) and groundings (G). 64.06% of all NF utterances are grouped in one cluster (*Cluster 2*), while 64.09% of EQ utterances are in one cluster *(Cluster 5)* and 94.74% of all G utterances are in another cluster *(Cluster 1)*. The Q and EQ tagged utterances,

[1] The Dirichlet clustering implementation from mahout.apache.org was used for this analysis.

which are structurally very similar in that they both are questions, were grouped into one cluster (*Cluster 5)*, which constitutes 58.54% of all Q and EQ tagged utterances. In another cluster, 14.26% of all Q and EQ utterances were grouped together *(Cluster 6)*.

Table 7: Student utterance distributions over clusters using manual tags (majority dialogue act in bold for each cluster)

Clusters/ Tags	*C1*	*C2*	*C3*	*C4*	*C5*	*C6*	*C7*	*C8*
GRE	31	1	0	4	8	0	0	2
EX	22	21	5	16	8	0	1	21
Q	1	13	11	20	121	33	0	36
PF	57	16	6	7	0	1	0	5
S	11	32	**21**	**29**	5	0	0	**51**
G	**144**	0	0	1	0	0	**6**	1
EQ	1	18	17	14	**191**	**43**	0	14
NF	3	**41**	9	4	0	0	0	7
LF	1	15	2	3	0	0	0	1

In order to examine the structure of the clusters in more detail, Table 8 provides sample utterances from four of the eight clusters. The manual labels of the utterances are also given in parenthesis. Some clusters perform intuitively; for example, *Cluster 1* is dominated by grounding, with many positive feedback utterances as well. These positive feedback and grounding utterances have similar surface forms such as "yeah," "right," and "oh," and distinguishing them further will require modeling the broader dialogue structure (for example, a notion of dialogue history) in future work.

Cluster 2 is dominated by negative feedback (NF) with a large number of statements (S). This cluster captures the vast majority of negative feedback moves, indicating that these moves are highly distinguishable. Most of the statements were in negative tone or expressing confusion. For example, the utterance "Im not sure if it is asking if the PARAMETER is how ever far you are away from NUMBER or the actual number you are away from NUMBER," which is labeled as statement, shows confusion although it is a statement in terms of its intention.

Cluster 3 and *Cluster 4* are highly impure. Closer inspection reveals that *Cluster 3* mostly has broken utterances such as "correction digit", "is in", "digit" and some statements. *Cluster 4* contains statements and extra domain utterances that are primarily statements. Moreover, there are some implicit questions such as "thats another thing I was going to ask am I just storing the values in METHOD_NAME and sending them to PARAMETER" and "So I have to call upon the PARAMETER class and use a method in there right?". *Cluster 7* is highly sparse, containing only seven utterances. These tend to be highly similar in structure, such as "oh ok," and "oh dear." The very low distance between these utterances pulled them tightly into one cluster that was judged as distinct from the others.

Cluster 5 and *Cluster 6* are dominated by evaluation questions (EQ), with a substantial number of general questions (Q) as well. The dominance of EQ in both clusters may relate to its high frequency within the corpus. However, these two question acts tend to exhibit close structural similarity although their roles are different (asking for feedback versus asking a general conceptual question). For example, the utterance 'do i have to set it to a PARAMETER?' is a question; however, a deeper analysis shows that it is more specifically an evaluation question that requests feedback on the task.

Table 8: Utterances from selected clusters with manual tags

Cluster 1

- right (G)
- ahh (G)
- ok (G)
- yeah (G)
- yes (PF)
- heheh yeah that would work (PF)
- I see that (PF)
- gotcha (EX)
- Yes Giving me definitions to various commands and such (EX)
- Ohh yes substantially (EX)

Cluster 2

- not really yet (NF)
- im not completely sure about how to do this (NF)
- the parsing im not sure about (NF)
- to be honest im not even sure what an array is (NF)
- im not sure how to read into the array (NF)
- I don't know how to do this (NF)
- and I know there is more to this line but I cannot remember the command (NF)
- Not yet (NF)
- im not so good at ARRAY just yet (NF)
- but I'm not exactly sure how to do that (NF)
- Im not sure if it is asking if the PARAMETER is how ever far you are away from NUMBER or the actual number you are away from NUMBER (S)
- I am asking how to do whatever drawing I need to do in the METHOD_CALL method (S)

Cluster 5

- So what's wrong with this? (Q)
- Can't manually turn an integer into a string? (Q)
- Then how would I incorporate that with the METHOD_CALL? I think it's asking me to use that in some why but it's not supplying arguments to do so (Q)
- are we done extracting digits? (Q)
- how do i sum the digits? (Q)
- do i have to set it to a PARAMETER? (EQ)
- thats another thing I was going to ask am I just storing the values in METHOD_NAME and sending them to PARAMETER? (EQ)
- why is what I just highlighted underlined in red doesn't that mean its wrong? (EQ)
- does extracting have to do with METHOD_NAME? or anything (EQ)

Cluster 6

- What is the next step? (Q)
- What do I write in it? (Q)
- what do i do first? (Q)
- so what can i do to fix what i was doing? (Q)
- does that look ok? (EQ)
- is this correct? (EQ)
- is this what i need to do? (EQ)

The presence of impure and sparse clusters prompted an experiment to explore whether other models with similar but slightly lower overall accuracy would yield a more desirable clustering. Therefore, we explored using an information criterion, Hartigan's rule of thumb, that utilizes the number of parameters in the model. This metric identified ten clusters as optimal, with a slightly lower accuracy (40.28%) compared to eight clusters, and the sparse clusters remained. We also experimented with the X-Means algorithm that utilizes Bayesian Information Criterion for splitting clusters [23], which resulted in four clusters with 36.72% accuracy.

6. DISCUSSION

A strength of unsupervised approaches is that because they do not rely on any manually engineered tagging schemes, they reflect the structure of the corpus in a fully data-driven way. In our case, the results highlight challenges of utilizing pedagogically driven manual dialogue act classification taxonomies within automated approaches. For example, a cross-cutting issue with the clustering presented here is that EX dialogue acts are distributed almost equally across several clusters. In the manual tagging, EX is a catch-all tag for conversation that was not directly related to tutoring. This tag was applied at the structural level, so if a question such as 'Should I close the door?' was not task-related it would have been tagged EX, as would its answer, 'Yes.' This distinction was desirable from a pedagogical perspective, but from a linguistic perspective it conflates dialogue act with topic. Future work will explore combining unsupervised dialogue act modeling with unsupervised topic modeling in order to address this type of modeling challenge. From a dialogue act research perspective, it is important to consider the issue of conflating act with topic when devising manual tagging schemes that may become the target of automated approaches in later work.

While the proposed algorithm is promising in that it outperforms current unsupervised approaches for dialogue act modeling, it has several notable limitations. One limitation is algorithmic complexity, which is quadratic over the size of the corpus. This complexity is inherent in the binary representation of each utterance as a vector with similarity to other utterances. Another limitation of the proposed approach arises with clustering algorithms in general, which is that a significant amount of human intelligence is often required to decide on the number of suitable clusters for the corpus. Nonparametric approaches to automatically identifying the number of clusters performed worse than parametric approaches in the current analyses; however, nonparametric approaches in general are an important area for future study. Finally, the query-likelihood clustering approach does not consider higher-level dialogue structure; it clusters one utterance at a time. This limitation leads to trouble disambiguating utterances with similar surface features. A highly promising direction to address this limitation is to enhance the algorithm with structural features such as dialogue history.

7. CONCLUSION AND FUTURE WORK

We have presented a novel student dialogue act classification model, query-likelihood clustering, which classifies dialogue acts in an unsupervised fashion by adapting techniques from information retrieval with a clustering approach. Although the technique did not utilize manual labels for model training, it performed substantially better than baseline chance at classifying utterances when compared to manually applied dialogue act tags. Moreover, query-likelihood clustering outperformed several currently reported approaches in the recent computational linguistics and educational data mining literature. It discovered a best-fit model with eight clusters, a close correspondence to the nine dialogue acts based on the handcrafted dialogue act taxonomy.

This novel approach holds great promise for dialogue act classification. Several directions are particularly important for future work. Multimodal features, such as posture, gesture, or facial expression of students, may hold great potential for providing dialogue act cues. Additionally, future work should focus on modeling higher-level dialogue structure such as adjacency pairs and discourse structure within unsupervised frameworks, and on devising suitable novel techniques for joint dialogue act and topic modeling within task-oriented tutorial dialogue. Finally, evaluating unsupervised techniques is an open research question. Future work will focus on further developing research techniques for evaluating unsupervised dialogue act classification. Together these research directions will allow unsupervised classification of dialogue acts within large corpora.

8. ACKNOWLEDGMENTS

This work is supported in part by the North Carolina State University Department of Computer Science and 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.

9. REFERENCES

[1] Aleven, V. and Koedinger, K.R. (2002). An effective metacognitive strategy: learning by doing and explaining with a computer-based Cognitive Tutor. *Cognitive Science*. 26, 2, 147–179.

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

[3] Baeza-Yates, R. and Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. ACM Press.

[4] Bangalore, S., Fabbrizio, G. Di 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.

[5] Becker, L., Basu, S. and Vanderwende, L. (2012). Mind the Gap : Learning to Choose Gaps for Question Generation. *North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 742–751.

[6] 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. *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue-SIGDIAL'10*, 297–305.

[7] Boyer, K.E., Phillips, R., Ingram, A., Ha, E.Y., Wallis, M., Vouk, M. and Lester, J. (2011). Investigating the Relationship Between Dialogue Structure and Tutoring Effectiveness: A Hidden Markov Modeling Approach. *The International Journal of Artificial Intelligence in Education (IJAIED)*, 65–81.

[8] Chen, L., Eugenio, B. Di, Fossati, D., Ohlsson, S. and Cosejo, D. (2011). Exploring Effective Dialogue Act Sequences in One-on-one Computer Science Tutoring Dialogues. *Proceedings of the 6th Workshop on*

Innovative Use of NLP for Building Educational Applications., 65–75.

[9] Chi, M., VanLehn, K. and Litman, D. (2010). Do micro-level tutorial decisions matter: applying reinforcement learning to induce pedagogical tutorial tactics. *International Conference on Intelligent Tutoring Systems*, 224–234.

[10] Crook, N., Granell, R. and Pulman, S. (2009). Unsupervised classification of dialogue acts using a Dirichlet process mixture model. *Proceedings of the 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue - SIGDIAL'09*, 341–348.

[11] Dzikovska, M., Steinhauser, N., Moore, J.D., Campbell, G.E., Harrison, K.M. and Taylor, L.S. (2010). Content, Social, and Metacognitive Statements: An Empirical Study Comparing Human-human and Human-computer Tutorial Dialogue. *Proceedings of the European Conference on Technology Enhanced Learning*, 93–108.

[12] Dzikovska, M., Moore, J.D., Steinhauser, N., Campbell, G., Farrow, E. and Callaway, C.B. (2010). BEETLE II : a system for tutoring and computational linguistics experimentation. *Proceedings of the ACL 2010 System Demonstrations*, 13–18.

[13] D'Mello, S.K., Lehman, B. and Graesser, A.C. (2011). A Motivationally Supportive Affect-Sensitive AutoTutor. *R. A. Calvo and S. K. D'Mello (Eds.), New Perspectives on Affect and Learning Technologies*, pp. 113–126.

[14] Forbes-Riley, K. and Litman, D. (2009). Adapting to Student Uncertainty Improves Tutoring Dialogues. *Proceedings of the International Conference on Artificial Intelligence in Education*, 33–40.

[15] 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.

[16] 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 the 13th Annual SIGDIAL Meeting on Discourse and Dialogue, Seoul, Republic of Korea*, 247–256.

[17] Hall, M., National, H., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I.H. (2009). The WEKA Data Mining Software: An Update. *ACM SIGKDD Explorations Newsletter*. 11, 1, 10–18.

[18] Jain, A., Nandakumar, K. and Ross, A. (2005). Score normalization in multimodal biometric systems. *Pattern Recognition*. 38, 12, 2270–2285.

[19] Kanungo, T., Member, S., Mount, D.M., Netanyahu, N.S., Piatko, C.D., Silverman, R. and Wu, A.Y. (2002). An Efficient *k*-Means Clustering Algorithm: Analysis and Implementation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 24, 7, 881–892.

[20] Kumar, R., Ai, H., Beuth, J. and Rosé, C.P. (2010). Socially Capable Conversational Tutors Can Be Effective in Collaborative Learning Situations. *Proceedings of the International Conference on Intelligent Tutoring Systems*, 156–164.

[21] Litman, D. and Silliman, S. (2004). ITSPOKE : An Intelligent Tutoring Spoken Dialogue System. *Demonstration Papers at NAACL HLT '10 Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

[22] 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 the Intelligent Tutoring Systems Conference*, 65–71.

[23] Pelleg, D. and Moore, A. (2000). X-means: Extending K-means with Efficient Estimation of the Number of Clusters. *Proceedings of the Seventeenth International Conference on Machine Learning*, 727-734.

[24] Ritter, A., Cherry, C. and Dolan, B. (2010). Unsupervised Modeling of Twitter Conversations. *Proceedings of NAACL HLT '10 Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 172–180.

[25] Rus, V., Moldovan, C., Niraula, N. and Graesser, A.C. (2012). Automated Discovery of Speech Act Categories in Educational Games. *Proceedings of International Conference on Educational Data Mining, 25-32*,

[26] Salton, G. and Buckley, C. (1990). Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science*. 41, 4, 288–297.

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

[28] Serafin, R. and Di Eugenio, B. (2010). Dialogue Act Classification, Higher Order Dialogue Structure, and Instance-Based Learning. *Journal of Dialogue & Discourse. 1, 2, 81-104*.

[29] 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.

[30] 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. *Association for Computational Linguistics*. 26, 3, 339–373.

[31] Strohman, T., Metzler, D., Turtle, H. and Croft, W.B. (2005). Indri : A language-model based search engine for complex queries. *Proceedings of the International Conference on Intelligent Analysis*.