Understanding Student Language: An Unsupervised Dialogue Act Classification Approach

Aysu Ezen-Can
North Carolina State University
aezen@ncsu.edu

Kristy Elizabeth Boyer
North Carolina State University
keboyer@ncsu.edu

Within the landscape of educational data, textual natural language is an increasingly vast source of learning-centered interactions. In natural language dialogue, student contributions hold important information about knowledge and goals. Automatically modeling the dialogue act of these student utterances is crucial for scaling natural language understanding of educational dialogues. Automatic dialogue act modeling has long been addressed with supervised classification techniques that require substantial manual time and effort. Recently, there is emerging interest in unsupervised dialogue act classification, which addresses the challenges related to manually labeling corpora. This paper builds on the growing body of work in unsupervised dialogue act classification and reports on the novel application of an information retrieval technique, the Markov Random Field, for the task of unsupervised dialogue act classification. 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 to education-centered unsupervised dialogue act classification. Unsupervised dialogue act classification techniques have broad application in educational data mining in areas such as collaborative learning, online message boards, classroom discourse, and intelligent tutoring systems.

1 INTRODUCTION

Natural language constitutes a vast portion of educational data. Recent years have seen a surge of educational data mining research aimed at modeling natural language data and leveraging those models to further student learning or to support effective teaching. Natural language has been mined within sources including lecture notes (Atapattu, Falkner, and Falkner, 2014), MOOC discussion forums (Wen, Yang, and Rosé, 2014), traditional class message boards (Yoo and Kim, 2014), computer-supported collaborative learning technologies (Kumar, Beuth, and Rosé, 2011), collaborative learning transcripts (D'Mello, Olney, and Person, 2010), tutorial dialogue systems (Graesser, VanLehn, Rosé, Jordan, and Harter, 2001), online discussion forums (Ferguson, Wei,

He, and Buckingham Shum, 2013) and student peer-reviews (Xiong and Litman, 2014). Models of natural language have been used to achieve goals including generating questions (Niraula, Rus, Stefanescu, and Graesser, 2014), assessing students' prior knowledge (Stefanescu, Rus, and Graesser, 2014), identifying social deliberative behavior (Xu, Murray, Woolf, and Smith, 2013), predicting task completion (González-Brenes, Mostow, and Duan, 2011), and detecting and predicting affect in educational games (Forsyth, Graesser, Pavlik Jr, Cai, Butler, Halpern, and Millis, 2013).

A primary focus of educational data mining of natural language interactions is to identify highly effective teaching strategies and implement them within educational systems that engage in dialogue (e.g., (Mostow, Beck, Cen, Cuneo, Gouvea, and Heiner, 2005)). Research suggests that dialogue-rich interactions may foster improved learning because of the adaptive collaboration mechanisms within dialogue (Graesser, Person, and Magliano, 1995), the available naturalness of expression (Litman, Rosé, Forbes-Riley, VanLehn, Bhembe, and Silliman, 2006), and by supporting students' self-explanation (Aleven, Popescu, and Koedinger, 2001) even for ill-defined domains (Weerasinghe, Mitrovic, and Martin, 2009).

Natural language dialogue can be modeled at many levels of granularity, and full natural language understanding involves a multi-step pipeline that links raw utterances to their semantics, intent, and context (Jurafsky and Martin, 2000). One of the most useful levels of dialogue modeling is *dialogue act classification* which identifies the communicative action or intent of each utterance, such as questions, hints, or statements (Allen, Schubert, Ferguson, Heeman, Hwang, Kato, Light, Martin, Miller, Poesio, et al., 1995; Core and Allen, 1997; Serafin and Di Eugenio, 2004; Traum, 1999; Stolcke, Ries, Coccaro, Shriberg, Bates, Jurafsky, Taylor, Martin, Van Ess-Dykema, and Meteer, 2000). Understanding student dialogue acts is a central challenge. For example, in a tutorial dialogue system, distinguishing whether the student is asking a question, requesting feedback, proposing a plan, or expressing affect is critical for subsequent tutorial move selections. Similarly, when modeling dialogues within message boards of MOOCs, student dialogue acts such as indirect questions, commitments, and social coordination may be of particular importance. In these and other contexts, dialogue act modeling provides key information to understanding student natural language contributions.

While dialogue act classification has been studied extensively for decades and reliable techniques exist for *supervised* dialogue act modeling, there are substantial challenges for scaling

these for educational data mining. First, supervised dialogue act models rely on handcrafted dialogue act tag sets which are often highly corpus-specific and require substantial consideration of the domain and of dialogue theory. Second, the manual effort required to label corpora so that supervised models can be trained is high. For these reasons, our work has focused for the past several years on unsupervised dialogue act modeling, which has only recently emerged as a research focus within the dialogue systems research community (Ezen-Can and Boyer, 2014b; Crook, Granell, and Pulman, 2009; Higashinaka, Kawamae, Sadamitsu, Minami, Meguro, Dohsaka, and Inagaki, 2011; Joty, Carenini, and Lin, 2011; Ritter, Cherry, and Dolan, 2010; Lee, Jeong, Kim, Ryu, and Lee, 2013) and to a very limited extent within educational data mining research (Ezen-Can and Boyer, 2013, 2014a; Rus, Moldovan, Niraula, and Graesser, 2012). The goal is to build well-formed models characterized by cohesive dialogue act groups that facilitate interpretation and subsequent use within systems to support teaching and learning.

This paper presents a novel approach to unsupervised natural language dialogue modeling for educational data mining, with application to tutorial dialogue. Leveraging highly effective techniques from the computational linguistics subfield of information retrieval, we propose a clustering approach based on a graph-based Markov Random Field (MRF) framework to group dialogue utterances with the same dialogue act in an unsupervised way, that is, without requiring the dialogue acts to be labeled manually ahead of time. We compare this new approach to prior unsupervised approaches in the literature and find that it outperforms all previously reported approaches, including our earlier work on query-likelihood clustering for dialogue act modeling (Ezen-Can and Boyer, 2013).

The organization of this article is as follows. We first give an overview of supervised and unsupervised classifiers in the Related Work section, followed in Section 3 by a description of the corpus used for modeling student utterances in this work. The steps undertaken for leveraging information retrieval techniques for dialogue act classification are described in Section 4. Section 5 presents experiments comparing MRF-based clustering with our prior query-likelihood approach and other unsupervised dialogue act modeling work in the educational data mining literature, detailing the results both quantitatively and qualitatively. Section 6 provides both quantitative and qualitative evaluation results with a held-out test set for the best performing MRF model. In Section 7 the performance of the dialogue act classifier is evaluated in terms of its capability for understanding students: we evaluate which dialogue acts are harder to distin-

guish by comparing the classifier's performance for specific dialogue acts. Finally in Section 8, we summarize the work presented in this article with some final remarks and future work.

2 Related Work

The idea that human conversation contains *speech acts* originated with sociolinguistic theorists (Austin, 1975; Searle, 1969). Dialogue act theory suggests that humans not only communicate factual information within natural language utterances, they often express underlying intended action (e.g., to ask a question, to give a command). The practical value of dialogue acts for computational linguistics has been well demonstrated, with an extensive literature on automated dialogue act classification approaches. Most of these approaches rely on supervised dialogue act classifiers. Hidden Markov models (Stolcke et al., 2000; Boyer, Ha, Phillips, Wallis, Vouk, and Lester, 2010), maximum entropy models (Rangarajan Sridhar, Bangalore, and Narayanan, 2009), conditional random fields (Quarteroni, Ivanov, and Riccardi, 2011), decision trees (Shriberg, Stolcke, Jurafsky, Coccaro, Meteer, Bates, Taylor, Ries, Martin, and Van Ess-Dykema, 1998) and support vector machines (Sadohara, Kojima, Narita, Nihei, Kamata, Onaka, Fujita, and Inoue, 2013) are some of the methods proposed by researchers for supervised dialogue act classification. For tutorial dialogue, promising approaches have included an extension of latent semantic analysis (Serafin and Di Eugenio, 2004), a syntactic parser model (Marineau, Wiemer-Hastings, Harter, Olde, Chipman, Karnavat, Pomeroy, Rajan, Graesser, Group, et al., 2000) and vector-based classifiers (Boyer et al., 2010), which typically achieve higher than 75% accuracy (Bangalore, Di Fabbrizio, and Stent, 2008; Forbes-Riley and Litman, 2005; Serafin and Di Eugenio, 2004).

Following this line of investigation, recent years have witnessed a growing body of work on *unsupervised* dialogue act modeling. Ritter et al. (2010) utilized Hidden Markov Models (HMMs) with topic information for modeling Twitter conversations. They separated content words (topic words) with the help of a Latent Dirichlet Allocation framework. Other research has followed this direction by proposing a variation of HMM for asynchronous conversations such as e-mails and forums, defining dialogue act emission distributions as mixture models and adding dialogue structure features (Joty et al., 2011). Dirichlet Process Mixture Models with a non-parametric Bayesian approach for train fares and timetables have also been explored (Crook

et al., 2009), and a subsequent improvement on that work used a hierarchical Dirichlet Process with Hidden Markov Models for extracting semantics from utterances (Lee et al., 2013). Lee and colleagues used a three-step approach: dialogue act, intent and slot entity recognition applied on spoken language. Similar to Ritter and colleagues, Lee et al. assume that each word is generated by one of three sources: words in the current dialogue act, general words and domain words. Another non-parametric Bayesian method, infinite HMM, has been explored within a Japanese discussion domain, and the results were compared with those obtained using the Chinese Restaurant Process showing that the infinite HMM performed better in terms of purity and F-measure (Higashinaka et al., 2011). There have also been attempts at clustering dialogue acts on educational corpora using k-means (Rus et al., 2012) as well as our prior work on combining query-likelihood with clustering (Ezen-Can and Boyer, 2013). Overall, the prior unsupervised approaches have substantially underperformed current supervised approaches. Additionally, with the exception of one hidden Markov Modeling approach (Joty et al., 2011), unsupervised models have not exploited word-ordering information, instead using a bag-of-words representation of utterances. The current work leverages word order for dialogue act clustering. We propose a dialogue act modeling framework that takes word-ordering information into account by representing the relationships between utterances as a Markov random field utilizing the probabilities obtained from the graph for clustering. This approach advances the state of the art for understanding students in terms of classifying dialogue acts.

3 Tutorial Dialogue Corpus

Our goal in this study is to understand students better by utilizing educational data mining for the task of dialogue act classification. Therefore, we mine a task-oriented tutorial dialogue corpus collected in 2007 for an introductory Java programming project. The corpus consists of student-tutor interactions in a computer-mediated environment while collaborating on computer programming problems (Boyer, Vouk, and Lester, 2007; Boyer et al., 2010; Boyer, Phillips, Ingram, Ha, Wallis, Vouk, and Lester, 2011). Students were allowed to ask questions and make fully unrestricted dialogue moves to their tutors via the textual communication channel in the course of solving programming tasks (see Figure 1).

The corpus consists of 1,525 student utterances from 43 distinct students (averaging 7.54

words per utterance) and 3,332 tutor utterances from two paid tutors (averaging 9.04 words per utterance). The corpus was manually annotated in prior work (with 0.80 Kappa) for both tutor and student dialogue acts that analyzed relationship between tutoring modes and student learning outcomes (Boyer et al., 2011). The tagging scheme consists of nine dialogue acts, the distribution of which is depicted in Table 1. The most frequently appearing dialogue act is EQ (evaluation questions) with 27.3%, which is the majority baseline chance constituting performance of a model that is equal to chance. Excerpts from the corpus can be seen in Table 2. For this article, the manual annotations are used only for evaluation purposes because unsupervised approaches assumes that manual labels are not accessible to the model building phase.

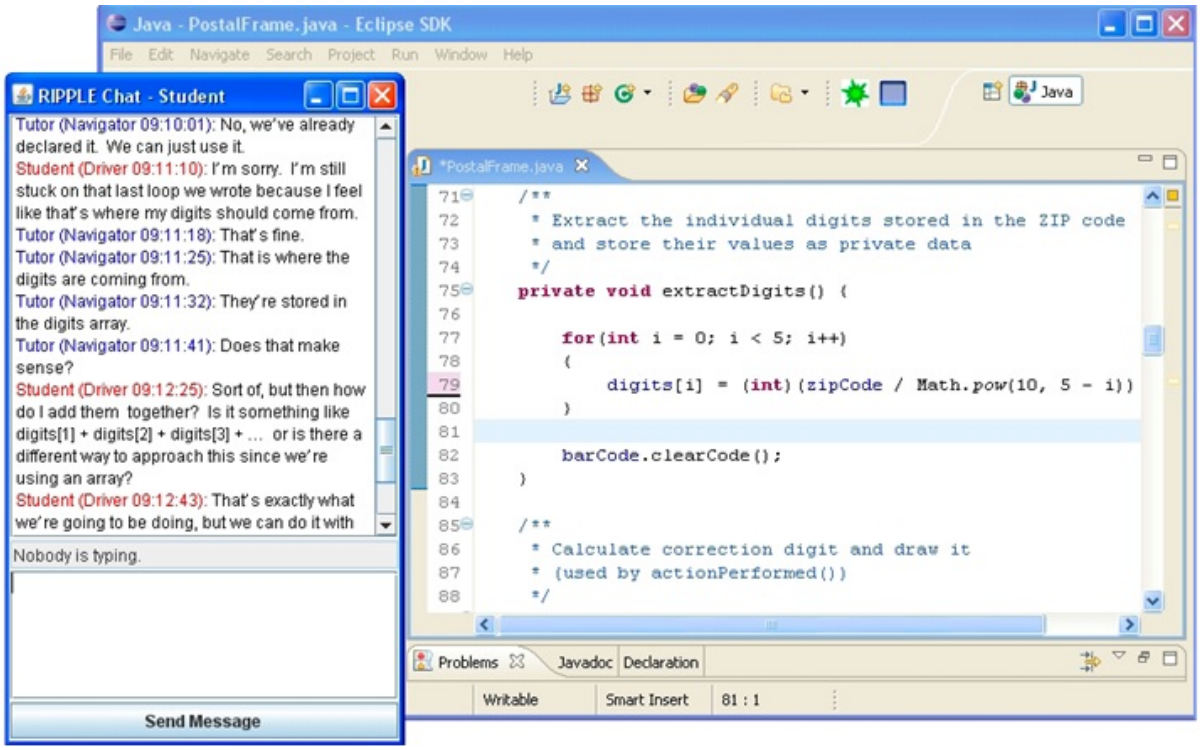

Figure 1: Screenshot from the human-human tutorial dialogue interface.

4 Unsupervised Dialogue Act Modeling

The goal of this work is to model student dialogue acts in an unsupervised manner. Our novel approach adapts information retrieval techniques combined with clustering for modeling student dialogue acts. In this section, the framework utilized for dialogue act classification is described.

Student Dialogue Act	Example	Percent of dialogue messages
Evaluation Question (EQ)	*can I parse a character to an int*	27.3
Question (Q)	*What is the next step?*	18.1
Statement (S)	*ohh let me try something else then*	13.8
Grounding (G)	*ok*	12.6
Extra-Domain (EX)	*Ok I have read over the first page*	8.7
Positive Feedback (PF)	*that works great*	7.6
Negative Feedback (NF)	*I'm having trouble figuring out how to pass the parameter to the other methods though*	6.0
Greeting (GRE)	*hi*	3.7
Lukewarm Feedback (LF)	*we've sort of learned them but im not exactly sure how to use them*	2.1

Table 1: Student dialogue act tags and their frequencies.

First we describe the natural language preprocessing, followed by the information retrieval techniques used to calculate utterance similarity and finally the clustering phase, which utilized calculated similarities to group utterances.

Information retrieval is the process of searching available resources to retrieve results that are similar to the query (Ricardo and Ribeiro-Neto, 1999). Probabilistic models such as language models are commonly used for this purpose (Manning, Raghavan, and Schütze, 2008). For example, in document search, documents that are expected to be relevant to a query are sorted according to their relevance to the query based on a similarity score. We adapt this information retrieval approach to compute similarity between utterances. This similarity information can then be used for clustering relevant utterances together, forming groups of utterances that we hypothesize represent the same dialogue act.

This section presents a novel approach to unsupervised dialogue act modeling: Markov Random Field based clustering. First we describe the technique that MRF-based clustering will be compared against, query-likelihood clustering, which is also based on information retrieval techniques. Query-likelihood clustering treats each student utterance as a query and retrieves other similar utterances from the corpus. MRF-based clustering builds on this approach with one key difference: MRF takes word ordering into account, unlike prior unsupervised dialogue act modeling approaches (Metzler and Croft, 2005). We hypothesize that MRF will perform better

Excerpt 1:
Student: will there be a loop in this? [*Q*] ***Tutor***: well, we need to add up all of the values in DigitArray [*S*] ***Tutor***: the easiest way to do that is with another loop [*S*] ***Tutor***: so yes [*S*] ***Student***: ok [*G*] ***Student***: do i declare a new in within this class or do i need to declare toward the beginning of the code? [*Q*] ***Tutor***: within this method is fine [*S*] ***Student***: cool [*G*]
Excerpt 2:
Student: what did i do wrong on this line [*Q*] ***Tutor***: Z isn't declared anywhere [*S*] ***Tutor***: so you want to declare Z as an array of int's [*S*] ***Tutor***: if you hit Save as you go those errors will update (Ctrl+S on the keyboard) [*EX*] ***Student***: how am i doing so far? [*EQ*] ***Tutor***: looks good [*PF*] ***Tutor***: keep going then we'll talk about how to shorten this up with a for-loop [*S*] ***Student***:how do i shorten this? [*Q*] ***Tutor***: well first let's try running it to make sure it's right [*S*] ***Tutor***: try putting a system.out.println in your method and run it [*S*] ***Student***: what is it meant to print out? [*Q*] ***Tutor***: if you call say System.out.println("5"); then that would print a 5 in the console when you ran it [*S*]

Table 2: Excerpts of dialogue from the corpus with the corresponding dialogue act tags in brackets.

than all prior techniques. The experiments described in the next section demonstrate this with empirical results.

In Section 4.1., we first explain the preprocessing required for modeling, and in Section 4.2., we detail the query-likelihood and MRF-based similarity calculations. Then, in Section 4.3. we describe the clustering.

4.1. Natural Language Preprocessing

When modeling natural language data, a series of natural language processing steps are often required prior to proceeding with further modeling. The types of features that are most useful to produce in the initial natural language processing steps are often not known in advance and is the subject of preliminary experimentation, as is the case in this work. In its raw form, natural

language dialogue utterances are a series of *tokens* which include words and punctuation. Some dialogue modeling approaches work best when given raw word-level tokens while other models benefit from abstracting the actual words in some way, for example to their parts of speech (POS). This work presents an example of this phenomenon. POS tagging labels each word according to their grammatical part of speech such as noun, verb, and adjective (Marcus, Santorini, and Marcinkiewicz, 1993). Because POS tagging represents words by their function in sentences, it provides a level of generalization that can be useful in dialogue modeling (Becker, Basu, and Vanderwende, 2012; Boyer et al., 2010; Di Eugenio, Xie, and Serafin, 2010). Because POS tagging is so useful, there are many available automatic POS taggers. In this work we utilize the Stanford Parser (Klein and Manning, 2003). We experimented with both actual words and with full part-of-speech backoff. While the best results for MRF-based clustering were achieved with raw features (actual words), query-likelihood clustering reached its best performance with a combination of raw features and POS tags. The hybrid approach utilized for query-likelihood clustering replaces function words such as determiners ("the", "a"), conjunctions ("and", "but"), and prepositions ("in", "after") with their POS tags.

A second important preprocessing choice is whether to use content words themselves or whether to stem them. Stemming is the process of generalizing words to their roots. For query-likelihood clustering, 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. We use the Snowball stemmer in this work[1].

Another consideration for preprocessing raw natural language dialogue data is how to represent special entities within the utterances. In our domain of computer science learning, the natural language contains special characters that indicate semantically important entities related to the domain, such as short bits of programming code. Students often incorporate some code such as function names or variable names into their text messages to tutors within the course of the dialogue. 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 manually 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*

[1] http://snowball.tartarus.org

statements, loops and arithmetic operations were all replaced in the corpus using similar conventions.[2]

4.2. Utterance Similarity Calculation

Having preprocessed the corpus, we now have a set of utterances that will be used for dialogue act classification. The next step is to calculate how similar these utterances are to each other so that we can cluster them. We report on the novel MRF-based model and compare it against query-likelihood modeling, both of which adapt information retrieval techniques as described in the following subsections.[3]

4.2.1. Query-Likelihood Model

We have previously reported on a technique to calculate similarities between utterances using query-likelihood (Ezen-Can and Boyer, 2013), and we describe this technique here for comparison. Query-likelihood is based on a language model which originates from the Bayesian assumption that each token is independent from every other token (Manning et al., 2008). This technique is widely used in search engines. Given a query, a list of documents that are expected to be relevant to the query are returned. We adapt this information retrieval goal to our purposes where we focus on finding similar utterances instead of documents. The process can be summarized as follows: *given a target utterance, find a list of utterances that are similar to the target.* The similar utterances are obtained from our corpus of student utterances. Table 3 presents two sample queries and top three most similar utterances retrieved by the query-likelihood model.

To achieve this, the query-likelihood model searches for words that are shared between the query and the utterances in the corpus. We use query-likelihood modeling to calculate proximity between utterances in a similarity space. For every target utterance whose dialogue act is to be classified, query-likelihood produces similarities to every other utterance in the corpus. In this way, we obtain a list of similar utterances for each utterance in the corpus. Because we would like to obtain groups of utterances that share the same dialogue act, having lists of similarity information is not sufficient. Therefore we need to use these lists for clustering. Using each

[2]Performing this annotation automatically is the focus of ongoing research. Differentiating programming code within natural language utterances is a challenging problem.

[3]The Lemur Project's information retrieval implementation (Indri) was used in this work (Strohman, Metzler, Turtle, and Croft, 2005).

Target utterance	Top three most similar student utterances
I am confused	- here's the part I am really confused on is this where I have to call up another class - and if so, then I guess I am sort of confused about how to retrieve the appropriate values from the table array - I'm confused on what I am going to put inside the loop
how can I solve	- how can I pull values out of an array or can I reference them with code like ARRAY_INDEXING - Is it like I have it on my screen And can I also set how long my array will be when I say private int PARAMETER -yea but i just can't remember to how to use the METHOD_CALL to get each individual number

Table 3: Sample queries and their top three query-likelihood results.

produced list, we create a vector representation that shows each utterance in the list as a 1 indicating presence of that utterance in the list of similar utterances and others as a 0 indicating absence of the utterance, to perform clustering. Figure 2 illustrates this process.

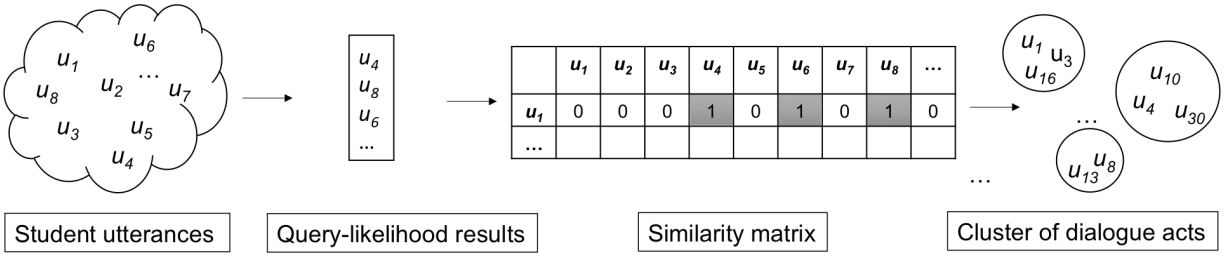

Figure 2: The query-likelihood clustering framework.

4.2.2. Markov Random Field Model

The previous section described the query-likelihood technique. The new approach presented here explicitly models the token ordering within a graphical representation, MRF graphs. In this way, while finding similarities between two utterances, we aim to consider the sequential order of each token rather than using a bag-of-words approach. For example, the utterances "I am correct" and "am I correct" have higher similarity scores to each other using query-likelihood clustering than MRF because all tokens are the same without considering the token ordering.

However, in this case a lower similarity is desirable as one of the utterances is a statement and the other is a question in terms of their communicative intentions. This distinction is the main motivation of MRF-based clustering. We hypothesize that if the similarities of utterances utilized in clustering technique are calculated more accurately by taking the token ordering into consideration, the clustering performance will improve as well. We therefore use a formulation of MRF that takes word ordering into account (Metzler and Croft, 2005).

MRF has been shown to be one of the best performing algorithms in the information retrieval community (Metzler and Croft, 2005). Therefore, proving its theoretical background is beyond the scope of this article. However, we would still like to provide motivation for using MRF for similarity calculation by discussing some differences with widely used metrics in the natural language processing literature. *Skip n-grams* are a generalization of *n-grams* where the words need not be consecutive, but there may be gaps within a window size. For example, if the window size is w, a bigram (n-gram with n=2) will consider a sequence of four tokens with the window of 2 tokens changing in between. Although this technique is widely used, skip n-grams require a window size to be set before computation whereas MRF does not, which is a motivating factor for using MRF. In addition, the use of n-grams gets more computationally expensive as n increases. While representing n-grams as feature vectors, the length of the vector is given by the whole set of n-grams in the corpus, making the vector large and the effect of the non-zero values lower, whereas in MRF likelihood summation, the non-zero values do not affect the similarity calculation. This property of MRF enables the nonzero values to be given more importance as desired. Further, MRF does smoothing for frequently occurring tokens, taking into account more important and representative words.

Recall that the purpose of using MRFs is to identify similar utterances. Another way to do this would be *longest common subsequence* to compute the similarity of not necessarily contiguous sequences of words. This technique is costly because all window sizes are computed, whereas it is possible to limit the window size (e.g., w=2) in MRF. In addition, it is possible to weight some values of w and n more than others in MRF by giving different weights to edges. The ordered tokens can be weighted more than single words ($n \geq 2$ more than $n = 1$) and phrases with one skipped token weighed more than multiple skipped tokens ($w = 1$ more than $w \geq 2$), whereas it is very difficult to do weighting with the longest common subsequence metric. It would require weighting some of the columns in the feature vector compared to a

better structured way provided by MRF. Motivated by the flexibility of MRF models, we apply this technique to calculate similarities between utterances.

First, MRF models are drawn and the probabilities obtained from the undirected graph serve as inputs to the clustering algorithm (i.e., k-medoids clustering) as shown in Figure 3. Suppose y_i is an utterance that we would like to calculate the similarities to every other utterance u_j in the corpus. We draw an MRF model with u_j being the root and $t_1...t_k...t_n$ the leaves where t_k are the tokens of the target utterance y_i. This graph is drawn for every utterance in the corpus for which we would like to calculate proximity to y_i. The edges represent how well the token t_k describes the utterance u_j. Using the cliques formed via edges between one utterance and each token, we create a similarity matrix. For instance, the first row of the similarity matrix shows the similarities of utterance u_1 to every utterance in the corpus and similarly the second row for utterance u_2. This produces a symmetric matrix, therefore it is sufficient to compute only the half above the diagonal. Then, these similarity values are input into the clustering technique that outputs groups of utterances that are hypothesized to share the same dialogue act.

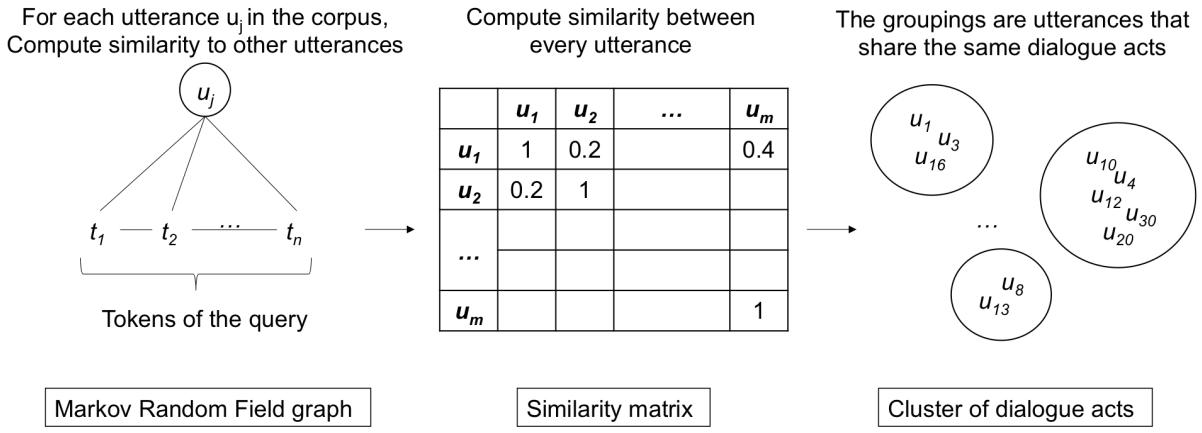

Figure 3: The MRF-based clustering framework.

To compute the needed similarities between every pair of utterances in the corpus, we consider 3-node cliques and 4-node cliques which are analogous to bigrams (n-grams with $n = 2$) and trigrams ($n = 3$) considering the root. Additionally, we consider all values of window size w, which is analogous to skip bigrams and trigrams. In the MRF, a 2-clique is formed when an edge exists between two nodes (a token t_k and utterance u_j). This represents a bag-of-words approach where we calculate how well token t_k describes utterance u_j. For calculating the edge

values of a 3-clique, we search for skip bigrams for all values of w limited by the length of u_j. We use the edge values obtained from the graphs as similarity measures between utterances. This approach is a generalization of all three techniques previously mentioned: bag-of-words, skip n-grams and longest common subsequence.

For the bag-of-words approach ($n = 1$, unigrams), we calculate a smoothed language model estimate $P(t_k|u_j)$ where the probability measures the likelihood of token t_k describing u_j. In other words, we wish to estimate the importance of this token in the utterance compared to the importance of this token in the entire corpus. This probability can be estimated as follows:

$$P(t_k|u_j) = [(1 - \alpha_d)\frac{freq(t_k, u_j)}{|u_j|} + \alpha_d\frac{freq(t_k, C)}{|C|}] \tag{1}$$

where $freq(t_k, u_j)$ is the frequency of token t_k in u_j, $|u_j|$ is the total number of tokens in u_j, $freq(t_k, C)$ is the frequency of token t_k in the corpus, and $|C|$ is the total number of tokens in the corpus. The smoothing is included to assign a non-zero probability to unseen words in utterance u_j that are present in the corpus, a common technique for natural language distributions where many words occur with low frequency (Zhai and Lafferty, 2001).

In addition to unigrams, we also wish to consider the more flexible skip n-grams with $w \geq 0$. To do this, we estimate the probability that tokens t_k and t_m, possibly separated by w other tokens, describe u_j. This probability $P(t_k, t_m|u_j)$ for 3-cliques is computed as follows:

$$P(t_k, t_m|u_j) = [(1 - \alpha_d)\frac{freq(t_k * t_m, u_j)}{|u_j|} + \alpha_d\frac{freq(t_k * t_m, C)}{|C|}] \tag{2}$$

where $freq(t_k * t_m, u_j)$ is the frequency of the phrases in utterance u_j that start with t_k and end with t_m with w tokens between them, and similarly $freq(t_k * t_m, C)$ is the count of such phrases in the entire corpus. By varying w, we identify all phrases in which the tokens occur in that order.

The process described above uses cliques to compute similarities among individual constituents t_k of each target utterance $y_i = t_1 t_2 ... t_n$ and all other utterances $u_j \in C$. We need an overall similarity between the target utterance y_i and all other utterances u_j. To compute this overall similarity, we sum over the 2-clique and 3-clique similarities using the following formula

adapted from information retrieval (Metzler and Croft, 2005):

$$M(u_j, y_i) = \sum_{c \in G} \lambda_i P(t_k|u_j) + \sum_{c \in G} \lambda_d P(t_k, t_m|u_j) \quad (3)$$

where λ_i is the weight of 2-cliques and λ_d is the weight of 3-cliques. These similarities between utterances are then placed into a matrix M.

4.3. CLUSTERING

The similarity results obtained as described above are used as the distance metrics for clustering dialogue acts. For query-likelihood clustering, each utterance that is present in the similarity list (above the defined similarity threshold) is represented as a 1, while the others are represented with a value of 0 (Ezen-Can and Boyer, 2013). For MRF-based clustering, each utterance is represented by a vector where similarities are obtained from probabilities in the Markov Random Field model, the matrix M described above. In this way, each target utterance in the corpus is represented by a vector indicating the utterances that are similar to it. Then the clustering takes the produced matrix as an input to group utterances that are similar to each other. We utilize a widely used clustering algorithm k-medoids (Ng and Han, 1994) for MRF-based clustering. The entire unsupervised dialogue act classification algorithm for MRF-based clustering is depicted in Table 4.

Let D be a corpus of utterances $D = u_1, u_2, ..., u_n$. Then the goal is:
$\forall\ u_j \in D$, identify l_j as dialogue act label of u_j
Procedure:
For each utterance u_j
1. Set target utterance to y_i
so that the similarities of y_i to every other utterance will be calculated
2. Build the Markov Random Field graphs G with the tokens of y_i
as the leaf nodes such that every other utterance in the corpus
is represented as roots of G
3. Create vector of similarity results indicator variables from G as
$V_j = (v_1, v_2, ..., v_t, ...v_n)$ such that v_t is obtained from cliques
formed by target utterance y_i and an utterance u_j from the corpus in G
Let the total vector be $V_T = (V_1, V_2, ..., V_j, ..., V_n)$
Return clusters $C = c_1, c_2, ..., c_k$ such that C is the result of $kmedoids(V_T)$

Table 4: Markov Random Field-based clustering algorithm.

5 EXPERIMENTS WITH COMPARISONS

The goal of the experiments is to determine whether MRF-based clustering outperforms query-likelihood clustering as hypothesized. Additionally, we compare our implementation against that of the recent approach of Rus et al. (2012), which clusters utterances in an educational corpus via word similarity with Euclidean distance, using a specified number of leading tokens of each utterance. We use accuracy to compare these three models. How to best measure accuracy is a non-trivial question for unsupervised models, but we follow the standard practice of comparing to manual labels, though those labels were not used during model training.

For evaluating the MRF-based dialogue act classification approach, we follow an isomorphic approach to the one used in our prior work on query-likelihood dialogue act classification (Ezen-Can and Boyer, 2013): we first retrieve the similarity results for each utterance, and then send the matrix M of similarities to the clustering algorithm. Then using majority voting, we label each utterance with the most frequent dialogue act tag in its cluster.

Training set accuracy evaluates the ability of the models to match the manual labels for the data on which they were trained. Following standard practice for unsupervised model evaluation (Higashinaka et al., 2011; Joty et al., 2011; Rus et al., 2012), we utilize training set accuracy for comparison of the models. The training set accuracy is computed as the number of utterances correctly classified divided by the total number of utterances in the training set.

We explore varying numbers of clusters and provide accuracies for each of them separately. Figure 4 depicts the training set accuracy results for MRF-based clustering compared to query-likelihood clustering, the Rus et al. approach with 5 leading tokens, as well as the random chance baseline. This random chance baseline is the most frequently occurring dialogue act, Evaluation Question (EQ), at 27.3%. We use this highest frequency class as the random chance baseline, rather than $1/9 = 11.1\%$ which would be the accuracy of random guesses among all tags disregarding their frequencies. Choosing the highest frequency tag is a more stringent baseline because a classifier that always guesses EQ will achieve 27.3% accuracy, higher than 11.1%. As shown in Figure 4, MRF-based clustering outperforms its counterparts substantially, confirming our hypothesis.

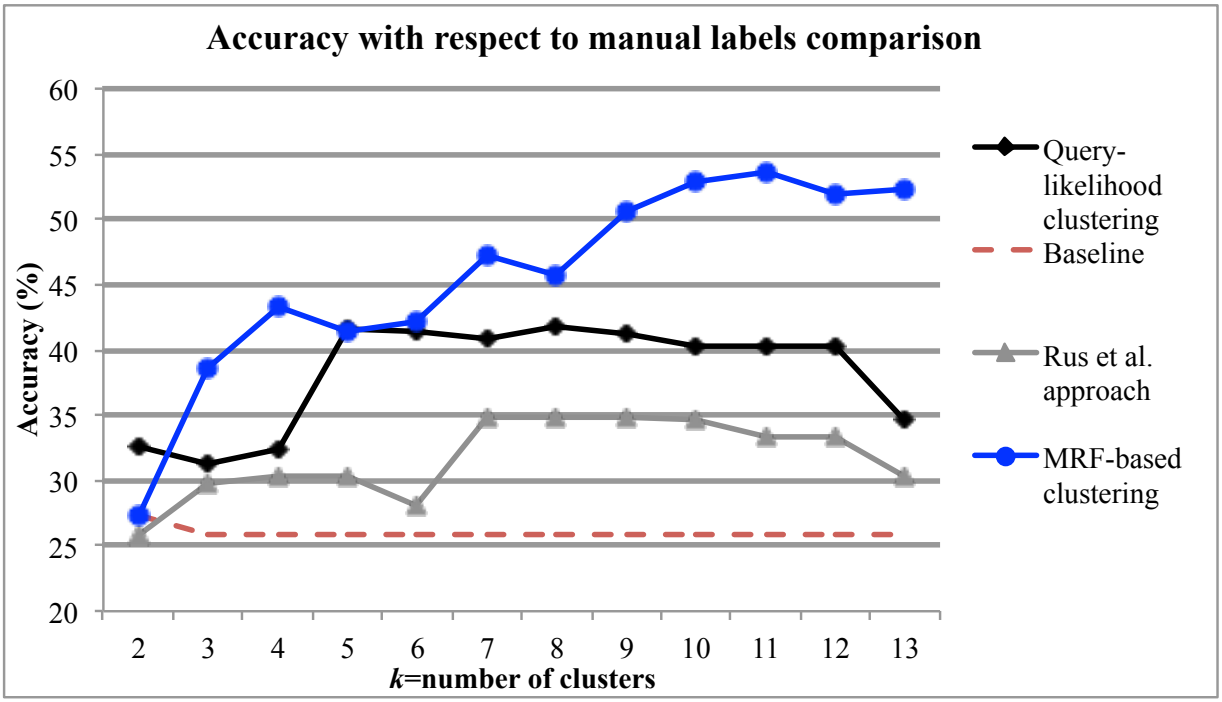

Figure 4: Comparison of two information retrieval inspired methods: query-likelihood clustering and MRF-based clustering as well as the only education-centered unsupervised dialogue act classification work by Rus et al.

6 Evaluation of MRF-Based Clustering

In addition to the training set accuracy used for comparison of different models, we are also interested in how well MRF-based clustering performs on unseen test utterances. To investigate this, this section reports on the selection of number of clusters k, followed by quantitative analyses of *test set accuracy, precision and recall*. Additionally, we examine the distribution of manual dialogue acts across the unsupervised clusters.

In order to determine the number of clusters for the MRF-based model, the Bayesian Information Criterion (BIC) is computed for each value of k. BIC penalizes the number of parameters, which is the number of clusters in our case. Lower BIC values represent a better fit to the data (Chen and Gopalakrishnan, 1998). In our corpus, a model with seven clusters achieved the lowest BIC value (see Figure 5). Note that, the lowest BIC value does not necessarily correspond to the model with highest accuracy because BIC does not consider manual labels, instead it only measures how coherent the clusters are considering distances between data points. The remainder of this section analyzes the seven-cluster model.

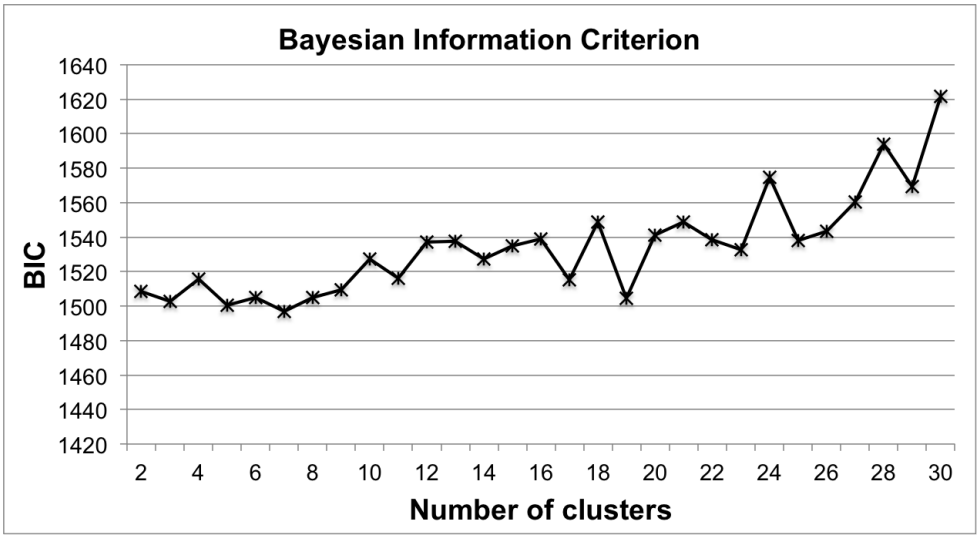

Figure 5: BIC values for varying number of clusters.

6.1. Test Set Accuracy, Precision and Recall

For analyzing the model's performance on unseen test data, we conduct leave-one-student-out cross-validation so that each student's utterances are included in the held-out test set once. In this way, we avoid providing our model with an unfair advantage because the utterances of the same student are not present both for training and testing. To label each held-out utterance, it is assigned to its closest cluster by taking the minimum average distance to all clusters. The majority label of the closest cluster is assigned as the dialogue act label of the test utterance. Then the test set accuracy is calculated as the number of utterances in the test set that are classified correctly, divided by the total number of utterances in the test set. For the MRF-based clustering, the overall accuracy is 36%. The F-measure for MRF-based clustering is 23.2%, with 24.5% precision and 24.0% recall. Because some dialogue acts have never become majority in any of the clusters, they were never predicted by the model (i.e., NF, LF and GRE). Considering only the dialogue act tags that were predicted by the model, the F-measure is 34.8%, with 36.8% precision and 36.1% recall. This performance is still well above baseline. Note that test set accuracy and F-measure were not reported for query-likelihood clustering nor by Rus et al. The confusion matrix in Figure 6 depicts the correct and incorrect predictions for the whole test set.

Similar to the training set evaluations, questions (Q and EQ) are among the most accurately predicted acts in the model. In contrast, there are some dialogue acts (NF, LF, GRE) that are not

True \ Predicted	Q	EQ	S	G	EX	PF	NF	LF	GRE
Q	55	195	13	13	0	0	0	0	0
EQ	107	263	23	23	0	0	0	0	0
S	66	83	49	11	1	1	0	0	0
G	6	13	17	130	12	14	0	0	0
EX	40	44	13	18	7	11	0	0	0
PF	24	25	7	12	5	43	0	0	0
NF	53	21	7	7	2	2	0	0	0
LF	14	9	3	2	2	2	0	0	0
GRE	18	15	2	2	11	8	0	0	0

Figure 6: Confusion matrix for leave-one-student-out cross validation.

predicted in the model. The reason for this is that these acts are so infrequent in the corpus that they were not assigned as the majority label of any cluster and therefore were never assigned to any test utterance.

To understand the performance seen above, we examine the distribution of dialogue acts among clusters for one of the folds of the MRF-based clustering (Figure 7). The majority dialogue act of each cluster is shown in bold.

Cluster No.	*Q*	*EQ*	*S*	*G*	*EX*	*PF*	*NF*	*LF*	*GRE*
1	16	16	53	52	**63**	36	18	7	43
2	**120**	114	77	2	31	25	54	12	5
3	10	15	**26**	0	6	0	2	1	0
4	7	14	4	**126**	10	4	4	1	1
5	18	**19**	11	7	11	11	5	2	0
6	1	0	1	0	5	**32**	0	2	0
7	98	**225**	37	1	4	1	2	2	6

Figure 7: Distribution of dialogue acts over clusters in one of the folds. Bold face indicates the majority dialogue act in each cluster.

The first cluster is mostly composed of extra-domain utterances (utterances that are off-topic)

as well as statements and groundings. Because the clustering approaches investigated here do not explicitly use topic information, they group on the surface-level features of utterances, which makes it challenging to distinguish utterances that are not related to the task. The second cluster and seventh cluster are mainly questions, both in the form of Q, a general question, and EQ, an "evaluation" question more specific to the task. In examining the second cluster we see that the model has difficulty differentiating these two question types. For the difficulty in differentiating questions and statements, we observed an interesting point about the corpus. Because the data collection was with novice students, it is very common for these students to use a declarative with a question mark attached such as "while declaring the loop I can use *i* again since it's local right?" and "no wait we need *if* for the conditions right". Given the features used within the models, these questions and statements appear similar in structure, which may explain the model's difficulty in differentiating them. The third cluster has mostly statements, and the model is successful in grouping grounding dialogue acts in the fourth cluster. The fifth cluster is highly impure and its interpretation is not straightforward. Finally, the sixth cluster groups positive feedback utterances together.

6.2. Qualitative Evaluation

In this section, we evaluate the resulting MRF-based clusters qualitatively by examining the utterances grouped in each cluster. Table 5 presents a sampling of utterances from each cluster of the model with seven clusters.

As shown by the examples of cluster 1, many extra-domain utterances and statements are grouped together. From surface-level features, it is difficult to distinguish dialogue acts that may be labeled differently by human coders; for example, "oh" was tagged as extra-domain and "oohhhh" as positive feedback. (Dialogue history is highly influential on different labels of these utterances and as discussed in Section 7, it is important to leverage within dialogue act models.) Likewise, the utterance "beginning now" which was labeled as off-topic is syntactically a statement.

Cluster 2 is mostly composed of questions and evaluation questions. The same finding we noted when examining the distribution of dialogue act tags within clusters is visible here, as both questions and evaluation questions are similar in structure.

The utterances in cluster 3 are generally in the form of statements regardless of their dialogue

act tags. For instance, the two utterances, "in a while loop" (which is tagged as question) and "so it would be like assignment" (which is tagged as evaluation question) are syntactically closer to statements than questions when the utterances are considered alone. Once again incorporating dialogue structure so that the model benefits more from the whole dialogue rather than the surface-level features alone is an important challenge moving forward, to address this issue.

It is clear that some clusters are influenced by words: cluster 4 sees many words like "ok" while cluster 6 sees many occurrences of "yes." In cluster 5, in addition to the questions, some utterances which seem like statements but which were manually tagged as questions appear. For example, "well this array is taking ints and we are putting in characters from a string" has a question tag from manual labeling.

7 Discussion and Limitations

Automatically understanding natural language that students exchange as they are learning is an increasingly prominent problem for educational data mining research. To achieve truly scalable models, unsupervised approaches hold great promise as they aim to address the labor-intensive nature of engineering taxonomies and manual labeling. We described an unsupervised model aimed at modeling student dialogue acts without requiring labor-intensive efforts for annotations and showed that 36% cross-validated test set accuracy is achievable by MRF-based clustering. Compared to a supervised model utilizing an extensive set of features including manually labeled task activities and hidden dialogue states learned by a Hidden Markov Model, which achieved 62.8% accuracy (Boyer et al., 2010), the results obtained by the proposed unsupervised model is promising. As promising as unsupervised models are, they pose important challenges. The dialogue act classifiers presented here have exemplified both the great promise and challenges of unsupervised dialogue modeling.

First, the results illustrate that while dialogue acts are a very useful distinction for educational dialogues, we also need other distinctions such as topic. For example, the dialogue act tag EX indicating off-topic utterances was not distinguished successfully by MRF-based clustering or the prior approaches used for comparison. We have begun to explore combinations of unsupervised dialogue act models with unsupervised topic models, which may address this challenge.

Cluster 1
-oh [*EX*]
-oohhhh [*PF*]
-let me fix this real fast [*S*]
-beginning now [*EX*]
-fair enough [*G*]
Cluster 2
-exactly how would i make an array an instance parameter [*Q*]
-do i set it equal to the conditions of parameter [*EQ*]
-could i just do that with a string [*EQ*]
-ok so how would i add up each digit [*Q*]
-all i remember how to do is convert strings into ints [*LF*]
Cluster 3
-in a while loop [*Q*]
-this should be a string [*EQ*]
-we want it to equal zero so should this be ten [*S*]
-go on [*Q*]
-so it would be like assignment [*EQ*]
Cluster 4
-ok [*G*]
-oh ok [*G*]
-ok like loop [*EQ*]
-ok however [*NF*]
Cluster 5
-is that already declared somewhere [*Q*]
-but since parameter is already given as an int [*Q*]
-is this different than my parameter thing [*Q*]
-and also if my parameter is correct [*EQ*]
-it looks like it is giong to come to that [*S*]
Cluster 6
-yes [*PF*]
-yes giving me definitions to various commands and such [*EX*]
-ohh yes [*EX*]
Cluster 7
-how can you get the sum of parameter [*EQ*]
-ok so it would be like assignment and so on until the last digit [*EQ*]
-what is the name of the postal code in the program [*Q*]
-can the chararray not be used outside because it is in a private method? [*EQ*]
-where are the drawing functions provided [*Q*]

Table 5: Example utterances from each cluster.

Second, the discrimination of dialogue act tags according to whether they referred specifically to the task (e.g., for questions and evaluation questions) was hard for the dialogue act classifier. For example, the data-driven groupings did not distinguish "exactly how would I make an array an instance parameter," which is manually annotated as a question, and "do i set it equal to the conditions of parameter," which is labeled as an evaluation question. In order to address this challenge, we have begun to explore both unsupervised and supervised semantic mapping from utterances to the learning task as a second stage to dialogue act classification. By doing this we may successfully group all questions together in one step, and then determine whether they refer to the task in a second step.

In addition, in this work we showed the promise of MRF-based clustering using only the current student utterance. Because each dialogue utterance is related to its predecessor, considering prior dialogue acts is an important way of representing dialogue history as shown in a supervised dialogue act classification task (Samei, Li, Keshtkar, Rus, and Graesser, 2014). Incorporating context-based features to the model presented in this work is a promising future direction.

8 Conclusion and Future Work

A tremendous amount of educational data is in the form of textual natural language. Whether in tutorial dialogue systems, textual collaborations, or MOOCs, these textual data capture intentions, goals, emotions, and other rich dimensions of human learning interactions. Dialogue act classification is an important step in understanding this natural language dialogue. This article proposed a new unsupervised dialogue act classifier inspired by information retrieval techniques, MRF-based clustering, and compared it to the previous best-performing unsupervised dialogue act classifiers for educational data. Experimental results showed that MRF-based clustering was more successful than its predecessor query-likelihood clustering, likely due to its ability to capture word ordering information.

Several future directions are promising. First, several features successfully utilized in supervised classification techniques remain unexplored for unsupervised models: these include multimodal features when available such as facial expression, posture, gesture, and speech signal. In addition, evaluation of unsupervised models constitutes an important research challenge. Most of the work to date has utilized manual labels as the gold standard. However, the un-

derlying assumption of accepting manual labels as gold-standard imposes unwanted restrictions on unsupervised models. Evaluation techniques that do not depend on manual labels should be investigated in the future; for example, our own future work is placing our best-performing unsupervised dialogue act models within a deployed tutorial dialogue system for end-to-end system evaluation. It is hoped that by moving the field of unsupervised dialogue act modeling forward we will enable better adaptive systems, and better automated understanding of human learning interactions.

9 Acknowledgments

The authors wish to thank to the members of the JavaTutor project, especially James Lester, Eric Wiebe, Bradford Mott, Eunyoung Ha, Christopher Mitchell, Joseph Grafsgaard, 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

Aleven, V., Popescu, O., and Koedinger, K. R. 2001. Towards tutorial dialog to support self- explanation : Adding natural language understanding to a Cognitive Tutor. *Proceedings of Artificial Intelligence in Education*, 246–255.

Allen, J. F., Schubert, L. K., Ferguson, G., Heeman, P., Hwang, C. H., Kato, T., Light, M., Martin, N., Miller, B., Poesio, M., et al. 1995. The TRAINS project: A case study in building a conversational planning agent. *Journal of Experimental & Theoretical Artificial Intelligence 7,* 1, 7–48.

Atapattu, T., Falkner, K., and Falkner, N. 2014. Acquisition of triples of knowledge from lecture notes: A natural language processing approach. In *Proceedings of the International Conference on Educational Data Mining*. 193–196.

Austin, J. L. 1975. *How to do things with words*. Vol. 1955. Oxford university press.

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.

Becker, L., Basu, S., and Vanderwende, L. 2012. Mind the gap: learning to choose gaps for question generation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 742–751.

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 the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue. Association for Computational Linguistics, 297–305.

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. *International Journal of Artificial Intelligence in Education 21,* 1, 65–81.

BOYER, K. E., VOUK, M. A., AND LESTER, J. C. 2007. The influence of learner characteristics on task-oriented tutorial dialogue. In *Proceedings of the 13th International Conference on Artificial Intelligence in Education (AIED)*. 365–372.

CHEN, S. S. AND GOPALAKRISHNAN, P. S. 1998. Clustering via the Bayesian information criterion with applications in speech recognition. In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing*. Vol. 2. 645–648.

CORE, M. G. AND ALLEN, J. 1997. Coding dialogs with the DAMSL annotation scheme. In *Proceedings of the AAAI Fall Symposium on Communicative Action in Humans and Machines*. 28–35.

CROOK, N., GRANELL, R., AND PULMAN, S. 2009. Unsupervised classification of dialogue acts using a Dirichlet process mixture model. In *Proceedings of the SIGDIAL 2009 Conference*. Association for Computational Linguistics, 341–348.

DI EUGENIO, B., XIE, Z., AND SERAFIN, R. 2010. Dialogue act classification, higher order dialogue structure, and instance-based learning. *Dialogue & Discourse 1,* 2, 1–24.

D'MELLO, S., OLNEY, A., AND PERSON, N. 2010. Mining collaborative patterns in tutorial dialogues. *Journal of Educational Data Mining 2,* 1, 2–37.

EZEN-CAN, A. AND BOYER, K. E. 2013. Unsupervised classification of student dialogue acts with query-likelihood clustering. In *Proceedings of the International Conference on Educational Data Mining*. 20–27.

EZEN-CAN, A. AND BOYER, K. E. 2014a. A Preliminary Investigation of Learner Characteristics for Unsupervised Dialogue Act Classification. In *Proceedings of the 7th International Conference on Educational Data Mining (EDM)*. 373–374.

EZEN-CAN, A. AND BOYER, K. E. 2014b. Combining task and dialogue streams in unsupervised dialogue act models. In *Proceedings of the 15th Annual SIGDIAL Meeting on Discourse and Dialogue*. 113–122.

FERGUSON, R., WEI, Z., HE, Y., AND BUCKINGHAM SHUM, S. 2013. An evaluation of learning analytics to identify exploratory dialogue in online discussions. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge*. ACM, 85–93.

FORBES-RILEY, K. AND LITMAN, D. J. 2005. Using bigrams to identify relationships between student certainness states and tutor responses in a spoken dialogue corpus. In *Proceedings of the 6th SIGDIAL Workshop on Discourse and Dialogue*. 87–96.

FORSYTH, C. M., GRAESSER, A. C., PAVLIK JR, P., CAI, Z., BUTLER, H., HALPERN, D., AND MILLIS, K. 2013. Operation aries!: Methods, mystery, and mixed models: Discourse features predict affect in a serious game. *Journal of Educational Data Mining 5,* 1, 147–189.

GONZÁLEZ-BRENES, J. P., MOSTOW, J., AND DUAN, W. 2011. How to classify tutorial dialogue? comparing feature vectors vs. sequences. In *Proceedings of the International Conference on Educational Data Mining*. 169–178.

GRAESSER, A., PERSON, N. K., AND MAGLIANO, J. P. 1995. Collaborative dialogue patterns in naturalistic one-to-one tutoring. *Applied cognitive psychology 9,* 6, 495–522.

GRAESSER, A. C., VANLEHN, K., ROSÉ, C. P., JORDAN, P. W., AND HARTER, D. 2001. Intelligent tutoring systems with conversational dialogue. *AI magazine 22,* 4, 39.

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 *INTERSPEECH*. 2081–2084.

JOTY, S., CARENINI, G., AND LIN, C.-Y. 2011. Unsupervised modeling of dialog acts in asynchronous conversations. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence*. 1807–1813.

JURAFSKY, D. AND MARTIN, J. H. 2000. *Speech & language processing*. Pearson Education.

KLEIN, D. AND MANNING, C. D. 2003. Accurate unlexicalized parsing. *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, 423–430.

KUMAR, R., BEUTH, J. L., AND ROSÉ, C. P. 2011. Conversational strategies that support idea generation productivity in groups. In *Proceedings of the Computer Supported Collaborative Learning (CSCL) Conference*. 398–405.

LEE, D., JEONG, M., KIM, K., RYU, S., AND LEE, G. 2013. Unsupervised spoken language understanding for a multi-domain dialog system. In *IEEE Transactions On Audio, Speech, and Language Processing*. Vol. 21. 2451–2464.

LITMAN, D. J., ROSÉ, C. P., FORBES-RILEY, K., VANLEHN, K., BHEMBE, D., AND SILLIMAN, S. 2006. Spoken versus typed human and computer dialogue tutoring. *International Journal of Artificial Intelligence in Education 16,* 2, 145–170.

MANNING, C. D., RAGHAVAN, P., AND SCHÜTZE, H. 2008. *Introduction to information retrieval*. Vol. 1. Cambridge University Press.

MARCUS, M. P., SANTORINI, B., AND MARCINKIEWICZ, M. A. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics 19,* 2, 313–330.

MARINEAU, J., WIEMER-HASTINGS, P., HARTER, D., OLDE, B., CHIPMAN, P., KARNAVAT, A., POMEROY, V., RAJAN, S., GRAESSER, A., GROUP, T. R., ET AL. 2000. Classification of speech acts in tutorial dialog. In *Proceedings of the Workshop on Modeling Human Teaching Tactics and Strategies at the Intelligent Tutoring Systems 2000 Conference*. 65–71.

METZLER, D. AND CROFT, W. B. 2005. A Markov random field model for term dependencies. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and development in information retrieval*. 472–479.

MOSTOW, J., BECK, J., CEN, H., CUNEO, A., GOUVEA, E., AND HEINER, C. 2005. An educational data mining tool to browse tutor-student interactions: Time will tell. In *Proceedings of the Workshop on Educational Data Mining, National Conference on Artificial Intelligence*. 15–22.

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.

NIRAULA, N. B., RUS, V., STEFANESCU, D., AND GRAESSER, A. C. 2014. Mining gap-fill questions from tutorial dialogues. In *Proceedings of the International Conference on Educational Data Mining*. 265–268.

QUARTERONI, S., IVANOV, A. V., AND RICCARDI, G. 2011. Simultaneous dialog act segmentation and classification from human-human spoken conversations. In *IEEE International*

Conference on Acoustics, Speech and Signal Processing. 5596–5599.

RANGARAJAN SRIDHAR, V. K., BANGALORE, S., AND NARAYANAN, S. 2009. Combining lexical, syntactic and prosodic cues for improved online dialog act tagging. *Computer Speech & Language 23,* 4, 407–422.

RICARDO, B.-Y. AND RIBEIRO-NETO, B. 1999. *Modern information retrieval*. Vol. 463. ACM press, New York.

RITTER, A., CHERRY, C., AND DOLAN, B. 2010. Unsupervised modeling of Twitter conversations. In *Proceedings of the Association for Computational Linguistics*. 172–180.

RUS, V., MOLDOVAN, C., NIRAULA, N., AND GRAESSER, A. C. 2012. Automated discovery of speech act categories in educational games. In *Proceedings of the International Educational Data Mining Society*. 25–32.

SADOHARA, K., KOJIMA, H., NARITA, T., NIHEI, M., KAMATA, M., ONAKA, S., FUJITA, Y., AND INOUE, T. 2013. Sub-lexical dialogue act classification in a spoken dialogue system support for the elderly with cognitive disabilities. In *Proceedings of the Fourth Workshop on Speech and Language Processing for Assistive Technologies*. 93–98.

SAMEI, B., LI, H., KESHTKAR, F., RUS, V., AND GRAESSER, A. C. 2014. Context-based speech act classification in intelligent tutoring systems. In *Proceedings of International Conference on Intelligent Tutoring Systems*. 236–241.

SEARLE, J. R. 1969. *Speech acts: An essay in the philosophy of language*. Cambridge university press.

SERAFIN, R. AND DI EUGENIO, B. 2004. FLSA: Extending latent semantic analysis with features for dialogue act classification. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*. 692–699.

SHRIBERG, E., STOLCKE, A., JURAFSKY, D., COCCARO, N., METEER, M., BATES, R., TAYLOR, P., RIES, K., MARTIN, R., AND VAN ESS-DYKEMA, C. 1998. Can prosody aid the automatic classification of dialog acts in conversational speech? *Language and speech 41,* 3-4, 443–492.

STEFANESCU, D., RUS, V., AND GRAESSER, A. C. 2014. Towards assessing students' prior knowledge from tutorial dialogues. In *Proceedings of the International Conference on Educational Data Mining*. 197–200.

STOLCKE, A., RIES, K., COCCARO, N., SHRIBERG, E., BATES, R., JURAFSKY, D., TAYLOR, P., MARTIN, R., VAN ESS-DYKEMA, C., AND METEER, M. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics 26,* 3, 339–373.

STROHMAN, T., METZLER, D., TURTLE, H., AND CROFT, W. B. 2005. Indri: A language model-based search engine for complex queries. In *Proceedings of the International Conference on Intelligent Analysis*. Vol. 2. 2–6.

TRAUM, D. R. 1999. Speech acts for dialogue agents. In *Foundations of Rational Agency*. Springer, 169–201.

WEERASINGHE, A., MITROVIC, A., AND MARTIN, B. 2009. Towards individualized dialogue support for ill-defined domains. *International Journal of Artificial Intelligence in Education 19,* 4, 357–379.

WEN, M., YANG, D., AND ROSÉ, C. P. 2014. Sentiment analysis in MOOC discussion forums: What does it tell us? In *Proceedings of the International Conference on Educational Data Mining*.

XIONG, W. AND LITMAN, D. 2014. Evaluating topic-word review analysis for understand-

ing student peer review performance. In *Proceedings of the International Conference on Educational Data Mining*.

XU, X., MURRAY, T., WOOLF, B. P., AND SMITH, D. 2013. Mining social deliberation in online communication–if you were me and I were you. In *Proceedings of the International Conference on Educational Data Mining*.

YOO, J. AND KIM, J. 2014. Capturing difficulty expressions in student online Q&A discussions. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*. 208–214.

ZHAI, C. AND LAFFERTY, J. 2001. A study of smoothing methods for language models applied to ad-hoc information retrieval. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 334–342.