# In-Context Evaluation of Unsupervised Dialogue Act Models for Tutorial Dialogue

**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

Unsupervised dialogue act modeling holds great promise for decreasing the development time to build dialogue systems. Work to date has utilized manual annotation or a synthetic task to evaluate unsupervised dialogue act models, but each of these evaluation approaches has substantial limitations. This paper presents an in-context evaluation framework for an unsupervised dialogue act model within tutorial dialogue. The clusters generated by the model are mapped to tutor responses by a handcrafted policy, which is applied to unseen test data and evaluated by human judges. The results suggest that in-context evaluation may better reflect the performance of a model than comparing against manual dialogue act labels.

## 1 Introduction

A central focus within the dialogue systems research community is developing techniques for rapidly constructing dialogue systems. One technique that has proven highly promising is to take a corpus-based approach to dialogue system authoring, for example by bootstrapping policy learning (Henderson, Lemon, & Georgila, 2008; Williams & Young, 2003), predicting what a human agent would do (Bangalore, Di Fabbrizio, & Stent, 2008), or learning supervised dialogue act models (Stolcke et al., 2000). Traditionally, these corpus-based approaches require some amount of manual annotation prior to learning the dialogue models. In many cases, this manual annotation is a problematic bottleneck for system development.

For tutorial dialogue systems, which aim to support students in acquiring skills or knowledge, heavy manual annotation is often required for learning models that classify student utterances with respect to dialogue acts (Forbes-Riley & Litman, 2005; Serafin & Di Eugenio, 2004), questioning strategies (Becker, Palmer, Vuuren, & Ward, 2012), or information sharing (Mayfield, Adamson, & Rosé, 2012)

For dialogue act modeling in particular, recent work has demonstrated the great promise of unsupervised approaches, which are learned without the use of manual labels (Crook, Granell, & Pulman, 2009; Ezen-Can & Boyer, 2013; Ritter, Cherry, & Dolan, 2010). However, because gold standard labels are not a part of model learning, how to best evaluate unsupervised models represents a significant open research question (Vlachos, 2011).

Most quantitative evaluations of unsupervised dialogue act models have relied on agreement with manual dialogue act annotations, though these annotations were not used in model learning (Crook et al., 2009; Rus, Moldovan, Niraula, & Graesser, 2012; Ezen-Can & Boyer, 2013). Relying on manually tagged dialogue act labels to evaluate an unsupervised model has two major drawbacks: it does not fully avoid the manual annotation bottleneck, and it imposes a hand-authored criterion onto a fully data-driven model, which may be unnecessarily limiting. Distinctions made by an unsupervised model may be useful within a dialogue system, even if these categories are different from the distinctions made within a hand-authored dialogue act tagset.

This paper presents a novel evaluation framework for unsupervised dialogue act classification of user utterances within tutorial dialogue. Instead of attempting to evaluate the model intrinsically, we evaluate its performance on



*Proceedings of the SIGDIAL 2013 Conference*, pages 324–328,
Metz, France, 22-24 August 2013. 

an external task: triggering an appropriate utterance via a simple dialogue policy. This evaluation, which does not require an end-to-end dialogue system, judges the model in the simulated context of the target task. The results demonstrate that this in-context evaluation may be equally useful as comparing against gold standard dialogue act labels, while substantially reducing the time required for human annotation.

## 2 Related Work

Perhaps the earliest unsupervised approach for dialogue act modeling investigated hidden Markov models with a bag-of-words approach in a meeting scheduling domain (Woszczyna & Waibel, 1994), using perplexity with respect to manual labels for evaluating the number of hidden states. Dirichlet process clustering has been investigated for dialogue act classification in the train fares and scheduling domain (Crook et al., 2009), evaluating on intra-cluster similarity and inter-cluster similarity along with error rates with respect to manual labels. Another Bayesian approach utilized hidden Markov models and topic modeling to classify Twitter posts (Ritter et al., 2010). Notably, Ritter et al. utilize an utterance ordering task, rather than manual labels, for quantitative evaluation. Most recently, standard *k*-means and EM clustering algorithms were used for dialogue act clustering on an educational corpus, and the model's accuracy was again evaluated with respect to manual labels (Rus et al., 2012). The current paper builds on these prior findings by applying a recently developed clustering framework and proposing a novel in-context evaluation scheme that can be used regardless of the unsupervised dialogue act modeling technique underlying it.

## 3 Dialogue Act Clustering

We consider an unsupervised dialogue act classification model on a corpus of human-human student and tutor dialogues centered on a computer programming task within a textual dialogue environment (Boyer et al., 2009). There are 1,525 student utterances and 3,332 tutor utterances in the corpus. This paper focuses on dialogue act classification for student utterances, since in a tutorial dialogue system the tutor dialogue acts are system-generated.

The corpus was manually labeled in prior work with nine dialogue acts tailored to capture phenomena of interest within tutorial dialogue: general *Question*, *Evaluation Question* (request specific feedback on the task), *Statement*, *Positive Feedback*, *Lukewarm Feedback*, *Negative Feedback*, *Grounding*, *Greeting,* and *Extra-domain* (utterances that are off topic). The Kappa for agreement on these manual tags was 0.76. These tags will be used within the present work to compare the in-context performance of the unsupervised policy with a manual-tag policy, but the tags are not used to learn or tune the unsupervised model.

The unsupervised dialogue act model evaluated here is based on a recently developed approach that adapts the query-likelihood technique from information retrieval to rank utterances similar to each target utterance (Ezen-Can & Boyer, 2013). Each utterance within the training set is queried against all other utterances within the training set using bigram features.

Vectors encode the resulting utterance similarity, and these vectors are provided to a *k*-means clustering algorithm to partition the utterances into dialogue acts. Our recent work (Ezen-Can & Boyer, 2013) evaluated query-likelihood dialogue act clustering against two other approaches with respect to classifying manual labels, and the query-likelihood approach outperformed *k*-means clustering using leading tokens (Rus et al., 2012) and Dirichlet process clustering (Crook et al., 2009). In the current work we add to the feature vectors the first level of the parse tree as provided by the Stanford parser (Klein & Manning, 2003).

The number of clusters was selected based on sum of squared errors (SSE). As with many parameterized models, model fit tends to increase with more parameters, but there are important tradeoffs in computation time and risk of overfitting. In experiments, *k*=number of clusters ranged from 2 to 24. 21 clusters were chosen, corresponding to the rightmost "knee" within the SSE graph (see Appendix).[1]

## 4 Evaluation Framework

Evaluating unsupervised dialogue act clusters presents numerous challenges. In prior evaluations of query-likelihood clustering, we computed accuracy with respect to the manually applied dialogue act tags described earlier, demonstrating 41.64% accuracy for a model with 8 clusters, compared to 34.90% accuracy for the Rus et al.

[1] Selecting the number of clusters is a subjective decision. Nonparametric techniques, such as variations on Dirichlet process clustering, hold promise for addressing this limitation in the future.

(2012) *k*-means approach and 24.48% accuracy for Dirichlet process clustering (Crook et al., 2009) on our corpus. However, the goal of the current work is to substantially reduce the human tagging required to evaluate the model. We also aim to test the hypothesis that comparing against manual labels under-represents the utility of the unsupervised model. That is, a dialogue policy built on the unsupervised model could perform better than the relatively low classification accuracy for manual tags would suggest. Our evaluation will explore this hypothesis.

In order to achieve these goals, we first trained an unsupervised dialogue act model on 75% of the corpus using the query-likelihood approach described in Section 3. The resulting model has 21 clusters. Then, we handcrafted a dialogue policy for tutor responses by qualitatively examining each cluster of training data and creating one tutor response for each cluster. Some clusters and their corresponding tutor utterances are depicted in Figure 1. This policy was applied by classifying unseen utterances from a held-out test set (25% of the corpus) using the learned model (Figure 2). The result of this process is that for each student utterance from the test set, a tutor response is generated based on the policy. This process resulted in 373 student utterances, one for each utterance in the 25% testing set, each paired with a corresponding tutor response generated by the hand-authored policy.

The evaluation goal is to determine whether the responses made by this policy are reasonable, which will represent the utility of the unsupervised dialogue act model for its intended use within a dialogue manager. We used human judges to rate the output of the policy. Thirty student utterances and tutor responses were randomly selected from the available utterances generated by the test set. An example set of utterances and policies can be seen in the Appendix. These items were placed in a survey that asked the reader to rate the extent to which each tutor response makes sense given the student utterance. (One item was inadvertently omitted from the survey, resulting in 29 items that were evaluated by the judges and that will be analyzed here.) To avoid bias introduced by the ordering of items, they were presented in a different randomized order for each of the seven judges who completed the survey. (29 items from a comparison condition using manual tags were also randomly interleaved into the survey, as described later in this section.) Judges used a rating scale from 1 to 4 (1=*makes no sense*, 2=*makes a little sense*, 3=*makes a lot of sense,* and 4=*makes perfect sense*). Since the models only used the current student utterance, the dialogue history was also not shown to the human raters.

Across the seven judges, the average rating of the tutor responses selected by the unsupervised policy was 2.35. We also collapsed the ratings into *positive* (≥2.5 average across seven judges) and *negative* (<2.5 average). With this binary categorization, 44.8% of the time tutor responses generated by the unsupervised policy were rated positively. It is important to note that no information other than dialogue act was considered for generating the tutor responses; the tutor utterances were relatively content-free and based only on the dialogue act categorization given by the unsupervised model.

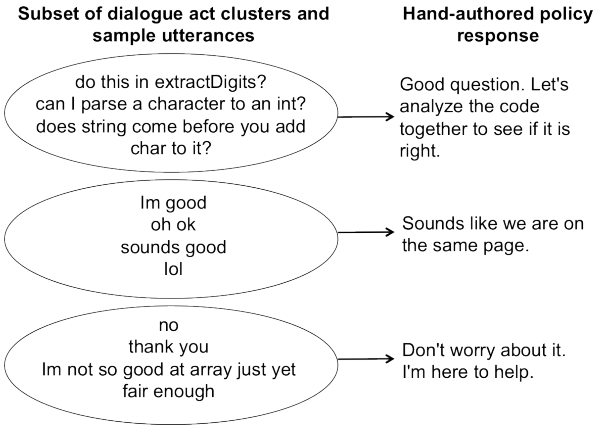


**Figure 1: Clusters from unsupervised dialogue act modeling and corresponding dialogue policy (typographical errors originated in corpus)**

For comparison, we also constructed a handcrafted dialogue policy using the manual dialogue act labels and applied this policy to the same utterances as were used to evaluate the unsupervised model. These pairs of student utterances and tutor responses were interleaved randomly on the same survey provided to seven human judges. The same tutor responses as in the unsupervised policy were used whenever possible for this manual-tag policy. The tutor responses generated from the manual-tag policy received an average score of 2.22, slightly lower than the average of 2.35 for tutor responses generated by the unsupervised policy. The binary positive-negative split for these ratings reveals that 31% were rated positively (≥2.5 average), compared to 44.8% for the unsupervised policy.

Direct comparisons between the unsupervised policy and the manual-tag policy must be interpreted with caution, in part because the unsupervised policy was more granular (based on 21

clusters) than the manual-tag policy (based on 9 tags) and also because it can be difficult to ensure that the two policies were of equal quality. On the other hand, the unsupervised policy utilized no manual labels and was applied to an unseen test set, while the manual-tag policy was based on reliable tags applied to the actual utterances from the testing set.

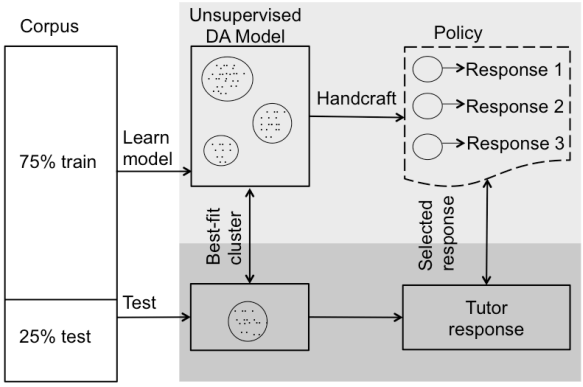


**Figure 2: Evaluation framework structure**

Finally, we evaluated the extent to which the 4-category rating scheme was reliable across judges. The weighted Kappa (Cohen, 1968), used for ordinal scales because it penalizes disagreements less if they are closer together, was 0.30 averaged across all pairs of judges, indicating *fair* agreement (Landis & Koch, 2013). For the collapsed binary ratings, average pairwise ordinary Kappa was 0.36.

## 5 Discussion

It was hypothesized that evaluating an unsupervised dialogue act model against manual labels may be an inappropriately strict metric, requiring the model to conform to the criteria used by humans to handcraft the manual tagset. Indeed, the accuracy of the unsupervised dialogue act model presented here with 21 clusters was 30.4% for identifying manual labels (arrived at by assigning the majority class tag to each unsupervised cluster after clustering was complete). The majority class baseline (most frequent student dialogue act tag) was *Evaluation Question* with a relative frequency of 25.87%, so on accuracy for identifying manual labels, the unsupervised model improved modestly over baseline. In contrast, when this unsupervised model was used to select a tutor response within a dialogue policy, the response was judged positively 44.8% of the time by human judges. Moreover, recall that the tutor responses were content free and took only the dialogue act label into account (no information state or topic). Therefore, it is meaningful to consider what percent of the time the responses were rated as making some sense (receiving a 2, 3, or 4 rating average across the human judges). By this criterion, 65.5% of tutor responses selected by the unsupervised policy were rated as sensible.

Finally, this evaluation approach demonstrates promise for alleviating the bottleneck of manual annotation for dialogue act models. Each item within the current evaluation survey required approximately 15 seconds to judge, using untrained human judges, for a total of approximately 1 hour of effort across *all seven* judges. The time required for handcrafting policies was relatively small, approximately 1 hour. In contrast, the dialogue act annotation scheme required approximately 35 seconds per utterance (amortizing substantial up-front training time for each annotator) when applied as part of previous work, for a total of approximately 50 hours per annotator.

## 6 Conclusion

Unsupervised dialogue act modeling holds great promise for decreasing development time of dialogue systems. We have presented an unsupervised dialogue act model and an evaluation framework to judge the utility of the unsupervised model within a dialogue management task. The results demonstrate that in-context evaluation of an unsupervised dialogue act model, rather than accuracy against manual labels, may better reflect the usefulness of the model for dialogue management. Furthermore, this evaluation technique may greatly reduce the time required by human judges to evaluate the model.

One of the most promising directions for future work involves devising unsupervised dialogue act models that leverage a richer representation in order to perform better. These rich features may include dialogue history, adjacency pair information, and topic modeling. Additionally, it is important for the community to evaluate unsupervised dialogue models in the full context of deployed systems.

**Acknowledgments**. This material is based upon work supported by the National Science Foundation under Grants DRL-1007962 and CNS-1042468. Any opinions, findings, conclusions, or recommendations expressed in this report are those of the authors and do not necessarily represent the views of the National Science Foundation.

## References


Bangalore, S., Di Fabbrizio, G., & 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., Palmer, M., Vuuren, S. Van, & Ward, W. (2012). Learning to Tutor Like a Tutor: Ranking Questions in Context. *Proceedings of the International Conference on Intelligent Tutoring Systems*, 368–378.

Boyer, K. E., Philips, R., Ha, E. Y., Wallis, M. D., Vouk, M. A., & Lester, J. C. (2009). Modeling Dialogue Structure with Adjacency Pair Analysis and Hidden Markov Models. *Proceedings of NAACL HLT*, 49–52.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, *70*(4), 213-220.

Crook, N., Granell, R., & Pulman, S. (2009). Unsupervised Classification of Dialogue Acts Using a Dirichlet Process Mixture Model. *Proceedings of SIGDIAL,* 341–348.

Ezen-Can, A., & Boyer, K. E. (2013). Unsupervised Classification of Student Dialogue Acts With Query-likelihood Clustering. *International Conference on Educational Data Mining*, 20–27.

Forbes-Riley, K., & Litman, D. J. (2005). Using Bigrams to Identify Relationships Between Student Certainness States and Tutor Responses in a Spoken Dialogue Corpus. *Proceedings of SIGDIAL*, 87–96.

Henderson, J., Lemon, O., & Georgila, K. (2008). Hybrid Reinforcement / Supervised Learning of Dialogue Policies from Fixed Data Sets. *Computational Linguistics, 34(4),* 487-511.

Klein, D., & Manning, C. D. (2003). Accurate Unlexicalized Parsing. *Proceedings of ACL*, 423–430.

Landis, J. R., & Koch, G. G. (1994). The Measurement of Observer Agreement for Categorical Data Data for Categorical of Observer Agreement The Measurement. *International Biometric Society*, *33*(1), 159–174.

Mayfield, E., Adamson, D., & Rosé, C. P. (2012). Hierarchical Conversation Structure Prediction in Multi-Party Chat. *Proceedings of SIGDIAL*, 60–69.

Ritter, A., Cherry, C., & Dolan, B. (2010). Unsupervised Modeling of Twitter Conversations. *Proceedings of NAACL HLT*, 172–180.

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

Serafin, R., & Di Eugenio, B. (2004). FLSA : Extending Latent Semantic Analysis with features for dialogue act classification. *Proceedings of ACL,* 692–699.

Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., Taylor, P., et al. (2000). Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech. *Computational Linguistics*, *26*(3), 339–373.

Vlachos, A. (2011). Evaluating unsupervised learning for natural language processing tasks. *Proceedings of EMNLP*, 35–42.

Williams, J. D., & Young, S. (2003). Using Wizard-of-Oz simulations to bootstrap Reinforcement-Learning- based dialog management systems. *Proceedings of SIGDIAL*, 135–139.

Woszczyna, M., & Waibel, A. (1994). Inferring linguistic structure in spoken language. *Proceedings of ICSLP,* 847-850.


## Appendix

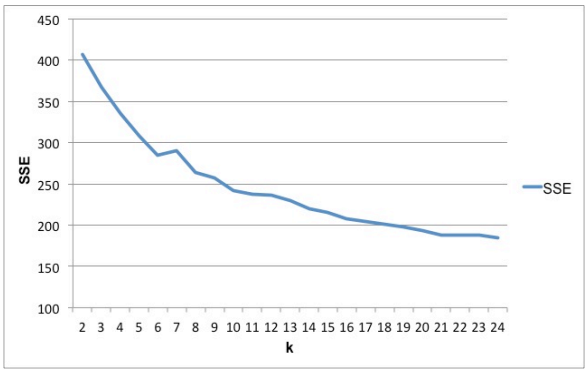


**Figure 3: Sum of squared errors graph**

**Table 1: Example student utterances and tutor responses**

| Student Utterance | Tutor Response (Unsupervised Policy) | Tutor Response (Manual-Tag Policy) |
|---|---|---|
| I'm trying to think, heh | Don't worry about it. I'm here to help. | That part is tricky. |
| how can I pull values out of an array or can I reference them with code like zipDigits[1]? | Great question. Let's look at the task decription together to see if it can help. | Good question. Let's analyze the code together to see if it is right. |
| thanks for the reminder | I'm here to help! | Great, seems like we agreed. |
| does that mean I should declare it at the top of the code? | Maybe you should try it out. | Good question. Let's analyze the code together to see if it is right. |