Hierarchical Dialogue Act Classification in Online Tutoring Sessions

Borhan Samei Vasile Rus Benjamin Nye Donald M. Morrison
Institute for Intelligent Systems
University of Memphis
bsamei@memphis.edu

ABSTRACT
As the corpora of online tutoring sessions grow by orders of magnitude, dialogue act classification can be used to capture increasingly fine-grained details about events during tutoring. In this paper, we apply machine learning to build models that can classify 133 (126 defined acts plus 7 to represent unknown and undefined acts) possible dialogue acts in tutorial dialog from online tutoring services. We use a data set of approximately 95000 annotated utterances to train and test our models. Each model was trained to predict top level Dialogue Acts using several learning algorithms. The best learning algorithm from top level Dialogue Acts was then applied to learn subcategories which was then applied in multi-level classification.

Keywords
Dialogue Act, Tutoring dialog, Machine Learning, Classification

1. INTRODUCTION
A speech or dialogue act is a construct in linguistics and the philosophy of language that refers to the way natural language performs actions in human-to-human language interactions, such as dialogues. [1] [6] In order to represent the Dialogue Act of an utterance, a set of Dialogue Act categories is defined. The set of categories is also known as the Dialogue Act taxonomy.

In this paper we examine different models on a relatively large data set which is extracted from one-on-one online tutoring sessions. The taxonomy used in our work is based on a hierarchical structure, i.e., each Dialogue Act has a set of sub-categories (subacts). The size of our training data is larger than the data presented in most of the previous work on Dialogue Act classification, which helps support this more fine-grained structure. We used WEKA toolkit [2] and the CRF++ package to train and test the models and Mallet [3] java library was used to train and test Logistic Regression models. Since our data is within the domain of human one-on-one tutoring sessions, this work enables further analysis of models to investigate the impact of dialog moves on learning. The feature sets used to train these models include the leading tokens of an utterance in addition to contextual information (i.e., features of previous utterances).

2. METHOD
The taxonomy used in this work was developed with the assistance of 20 subject matter experts (SMEs), all experienced tutors and tutor mentors. The resulting hierarchical taxonomy includes 15 main categories where each main dialog act category consists of different sub-categories which resulted in 133 distinct dialog acts out of which 7 categories were defined to represent unknown and undefined cases.

Once the taxonomy was available, a set of 1,438 sessions were manually tagged. The human tagging process included 4 major phases: development of taxonomy, 1st round tagging, reliability check, 2nd round tagging, reliability check, and final tagging phase.

The experts were divided into two groups: Taggers and Verifiers. In the first 2 tagging phases, each tagger was given a session transcript and asked to annotate the utterances. The resulting tagged session was then assigned to a verifier who went through the annotations, reviewed the tags and made necessary changes. In the reliability check steps, experts tagged each transcript independently.

Since the Verifiers were modifying tags already established by the Taggers in the 1st and 2nd round cases, the agreement was expected to be high. The agreement of Taggers and Verifiers was approximately 90%, with a slightly higher agreement on the second round. This shows to what extent the verifiers made changes to the initial annotations (about 10% of tags changed). The reliability checks involved completely independent tagging, in which human experts yielded an agreement of approximately 80% on top level and 60% on subact level. The final annotations were used as training data for our machine learning models. In order to build the Dialogue Act classifier, we applied the following 3 kinds of feature sets.

- Simple features: Based on previous research, 3 leading tokens of an utterance were shown to be good predictors for Dialogue Act [4]. Thus, we extracted the following features of each utterance: 1st token, 2nd token, 3rd token, last token, and length of utterance (i.e., number of tokens).

- Extended features: Using the Correlation Feature Selection (CFS) measure, we found that 1st and last token are the most predictive features and in order to add contextual information (features of prior utterances) we extended the simple features by adding the 1st and last token of three previous utterances to our feature set.

The above feature sets were used to create different models with multiple learning algorithms. Four learning algorithms were used and evaluated: Naïve Bayes, Bayesian Networks, Logistic Regression, and Conditional Random Fields (CRF). Each of the algorithms has certain properties that take into account different characteristics of data.

3. RESULTS & DISCUSSION
Based on the division of taxonomy in top-level and subcategories, we first trained and tested the models to predict the top-level Dialogue Act. Table 1 shows the results of 10-fold cross validation on the top-level classification models.

Table 1. 10-fold Cross Validation of Algorithms with Different Features for Top-level Dialogue Act Classification.

Algorithm	FeatureSet	Accuracy%	Kappa

Naïve Bayes	Simple	72.5	0.65
Naïve Bayes	Extended	72.3	0.64
Bayes Net	Simple	72.6	0.65
Bayes Net	Extended	72.5	0.65
Logistic Regression	Simple	76.6	0.70
Logistic Regression	**Extended**	**77.4**	**0.71**
CRF	Simple	72.7	0.45
CRF	Extended	71.9	0.44

As seen in table 1, the best performance on top-level classification is achieved by the Logistic Regression algorithm; however, all the algorithms yield and accuracy of more than 70%. It is interesting to note that the extended feature set does not improve the algorithms significantly which implies that adding the contextual information, i.e., prior utterances, is either not useful or not sufficiently representing the context. The diminished role of contextual features is not surprising. It has been previously indicated that they do not play a significant role in Dialogue Act classification models on a multi-party chat based tutoring system [5].

We further trained and tested models to classify utterances in the second level of Dialogue Act categories. For each Dialogue Act a classifier was trained to predict its corresponding subcategories. Table 2 shows the performance of these classifiers which were trained on 70% and tested on 30% of the dataset. A 10-fold cross-validation was not possible in this case due to too few instances for some subcategories.

Table 2. Performance of Subact Classifiers within each Dialogue Act Category using Logistic Regression algorithm.

Model	N	Accuracy%	Kappa
Answer	1130	52.8	0.43
Assertion	29890	57.6	0.42
Clarification	609	40.4	0.17
Confirmation	6620	92.6	0.77
Correction	2065	62.3	0.43
Directive	2006	61.7	0.52
Explanation	1941	54.4	0.25
Expressive	22198	76.8	0.74
Hint	341	67.6	0.34
Promise	303	95.6	0.00
Prompt	6186	64.2	0.30
Question	2553	60.7	0.49
Reminder	337	47.7	0.25
Request	14243	56.2	0.49
Suggestion	2028	70.2	0.43

As shown in Table 2, the subact classifiers yield an average accuracy of approximately 65% and kappa of 0.4. Next we created a single model to classify Dialogue Act and Subact. By combining the top-level dialogue acts with their subacts, this produced a flat taxonomy with 133 categories. Table 3 shows the performance of our models with flat taxonomy using 10-fold cross validation.

Table 3. Performance of models with flat taxonomy.

Algorithm	FeatureSet	Accuracy	Kappa
Naïve Bayes	Simple	51%	0.49
Naïve Bayes	Extended	48%	0.45
Bayes Net	**Simple**	**53%**	**0.50**
Bayes Net	Extended	51%	0.48
Logistic Regression	Extended	44%	0.42
Logistic Regression	Simple	43%	0.41

Table 3 shows that the flat taxonomy classification improved the accuracy of our model significantly when compared to the the multi-level classification. It is worth noting that these results approach the agreement of human experts when they annotated independently, which was 66%.

4. CONCLUSION

The results of the different models and algorithms showed that the top-level Dialogue Acts can be predicted with a reasonable accuracy. However to be able to tag utterances with both top-level and subcategories a combined classification needed to be applied, rather than a hierarchical approach. Multiple classification algorithm were effective, such as Naïve Bayes, Bayesian Networks, Logistic Regression, and Conditional Random Fields (CRF).

The ultimate goal of this work is to build a model to be applied to a set of not-seen and untagged data and use the Dialogue Acts as means of modeling the discourse. The proposed models in this paper can be used as initial models for a semi-supervised classifier which will ultimately identify Dialogue Acts in real time.

5. REFERENCES

[1] Austin, J. L. 1962. *How to do things with words*: Oxford.

[2] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. 2009. The WEKA Data Mining Software: An Update: *SIGKDD Explor. Newsl., 11*(1), 10-18.

[3] McCallum, Andrew Kachites. 2002. "MALLET: A Machine Learning for Language Toolkit: http://mallet.cs.umass.edu.

[4] Rus, V., Moldovan, C., Niraula, N., & Graesser, A. C. 2012. Automated Discovery of Dialogue Act Categories in Educational Games: *International Educational Data Mining Society*.

[5] Samei, B., Li, H., Keshtkar, F., Rus, V., & Graesser, A. C. 2014. Context-Based Dialogue Act Classification in Intelligent Tutoring Systems: *Intelligent Tutoring Systems - 12th International Conference, {ITS} 2014, Honolulu, HI, USA, June 5-9, 2014. Proceedings*, 236-241.

[6] Searle, J. R. 1969. *Dialogue Acts: An essay in the philosophy of language*: Cambridge university press.