Preliminary Results on Dialogue Act Classification in Chat-based Online Tutorial Dialogues

Vasile Rus, Rajendra Banjade
Department of Computer Science
The University of Memphis
Memphis, TN 38152
{vrus,rbanjade}@memphis.edu

Nabin Maharjan, Donald Morrison
The University of Memphis
Memphis, TN 38152
{nmharjan}@memphis.edu

Steve Ritter, Michael Yudelson
Carnegie Learning, Inc.
437 Grant Street, Suite 918
Pittsburgh, PA 15219, USA
{sritter}@carnegielearning.com

ABSTRACT

We present in this paper preliminary results with dialogue act classification in human-to-human tutorial dialogues. Dialogue acts are ways to characterize the intentions and actions of the speakers in dialogues based on the language-as-action theory. This work serves our larger goal of identifying patterns of tutors' actions, in the form of dialogue act and subact sequences, that relate to various aspects of learning. The preliminary results we obtained for dialogue act classification using a supervised machine learning approach are promising.

Keywords

dialogue acts, intelligent tutoring systems, instructional strategies.

1. INTRODUCTION

A key research question in intelligent tutoring systems and in the broader instructional research community is understanding what expert tutors do. A typical operationalization of this goal of understanding what expert tutors do is to define the behavior of tutors based on their actions.

In our case, because the focus is tutorial dialogues, we model the actions of tutors using dialogue acts inspired from the *language-as-action* theory [1, 7]. According to the language-as-action theory, *when we say something we do something*. Therefore, we map all utterances in a tutorial dialogue onto corresponding dialogue acts using a predefined dialogue act taxonomy, which is described later. It should be noted that automatically discovered dialogue act taxonomies are currently being built [6]. However, we chose to work with an expert-defined taxonomy of dialogue acts, developed by experts based on dialogue and pedagogical theories [5], because it better serves our larger research goals of testing such theories.

2. THE APPROACH

We adopted a supervised machine learning method to automate the process of dialogue act classification. This implies the design of a feature set which can then be used together with various supervised machine learning algorithms such as Naive Bayes, Decision Trees, and Bayes Nets. For automated dialogue act classification, researchers have considered rich feature sets that include the actual words (possibly lemmatized or stemmed) and n-grams (sequences of consecutive words). Besides the computational challenges posed by such feature-rich methods, it is not clear whether there is need for so many features to solve the problem of dialogue act classification.

Our approach is based on the observation that humans infer speakers' intention after hearing only a few of the leading words of an utterance [4]. One argument in favor of this assumption is the evidence that hearers start responding immediately (within milliseconds) or sometimes before speakers finish their utterances ([5] - pp.814).

Intuitively, the first few words of a dialog utterance are very informative of that utterance's dialogue act. We could even show that some categories follow certain patterns. For instance, Questions usually begin with a *Wh*-word while dialogue acts such as Greetings use a relatively small bag of frozen words and expressions.

In the case of other dialogue act categories, distinguishing the dialogue act after just the first few words is not trivial, but possible. It should be noted that in typed dialogue, which is a variation of spoken dialogue, some information is not directly available. For instance, humans use spoken indicators such as the intonation to identify the dialogue act of a spoken utterance. We must also recognize that the indicators allowing humans to classify dialogue acts also include the expectations created by previous dialogue acts, which are discourse patterns learned naturally. For instance, after a first Greeting another Greeting that replies to the first one is more likely. We used intonational clues in our work to the extent that such information is indirectly available to us, in the form of punctuation marks, in typed/chat-based dialogues. We did incorporate contextual clues in our preliminary experiments, e.g. we used as a feature the dialogue act of the previous utterance, but the results did not improve significantly. It is important to note that the present study assumes there is one direct speech act per utterance.

3. THE TAXONOMY

The current coding taxonomy builds on an earlier taxonomy that sought to identify patterns of language use in a large corpus of online tutoring sessions conducted by human tutors in the domains of Algebra and Physics [5]. The taxonomy is considerably more granular than previous schemes such as the one used by Boyer and colleagues [2].

The most recent version of the taxonomy employs two levels of description. At the top level, it identifies 16 standard dialogue categories including Questions, Answers, Assertions, Clarifications, Confirmations, Corrections, Directives, Explanations, Promises, Suggestions, and so forth. It also includes two categories, Prompts and Hints, that have particular pedagogical purposes. Within each of these major dialogue act categories we identify between 4 and 22 subcategories or subacts.

4. EXPERIMENTS AND RESULTS

We have used in our experiments 288 tutorial sessions (containing about 17,537 utterances) between professional human tutors and actual college-level, adult students. These sessions are a subset of of a larger sample of 500 sessions randomly selected from a corpus of 17,711 sessions we obtained from an organization that offers online human tutoring services. Students taking two college-level developmental mathematics courses (pre-Algebra and Algebra) were offered these online human tutoring services at no cost. The same students had access to computer-based tutoring sessions through Adaptive Math Practice, a variant of Carnegie Learning' Cognitive Tutor. It should be noted that students may or may not initiate a tutorial dialogue with a human tutor while attending those courses. This is important to note as there could be a self-selection bias in those tutorial dialogues that we used.

Expert Annotation Process

The 288 sessions we used here were manually labelled by a team of 6 trained annotators, all of whom were experienced classroom math teachers. Each session was first manually tagged by two independent annotators, i.e. they did not see each other's tags. Then, the tags of the two independent annotators were double-checked by a verifier, who also happens to be the designer of the taxonomy. The verifier had full access to the tags assigned by the independent taggers. The role of the verifier was to resolve discrepancies. The inter-annotator agreement for the two independent annotators was Cohen's kappa=0.72 for dialogue acts and kappa=0.60 for dialogue acts and subacts combined.

The agreement was best for Expressives (0.88), Assertions (0.81), Requests (0.78) and worst for Hints (0.2), Clarifications (0.33), and Explanations (0.42).

Results

For space reasons, we summarize the results of our supervised machine learning approach in terms of accuracy and Cohen's kappa relative to the final tag adjudicated by the verifier using a 10-fold cross-validation approach. We only provide results on dialogue act classification (no subacts) for the same space reasons.

The model

Our model for predicting dialogue acts consists of the following five features/predictors: the leading three tokens in an utterance, the last token such as a question mark ('?') at the end of a question, and the length of the utterance. We experimented with other features such as the speaker (student vs. tutor), the position of the utterance in the dialogue, e.g. an utterance at the beginning of a session is more likely a Greeting, the previous dialogue act, but we have not noticed any significant impact on performance relative to the five-feature model mentioned above. More powerful models that do account explicitly for sequential observations are needed, e.g. Conditional Random Fields.

We experimented with our 5-feature model in combination with a number of machine learning algorithms including Naïve Bayes, Decision Trees, and Bayes Nets. We also experimented with sequential models based on Conditional Random Fields but the results, again, were not better. The best results, obtained with BayesNets, are summarized below.

D-Act classification Results

Using all features leads to 67.27% accuracy and Cohen's kappa of 0.58. The speaker does not seem to have an impact as the results accuracy is 66.74%. The same for position, if removed the resulting accuracy is 66.77%. The remaining features are indeed important as if another is removed the accuracy drops significantly below 60.00%.

Our plan next is to annotate more sessions up to 500 and retrain our models. Once the accuracy is at acceptable level, we will use the classifiers to automatically tag tens of thousands of sessions with dialogue acts and subacts. Once the sequences of actions and subactions are available, we will identify patterns of tutor and student actions that related to learning and affect and which could then be used in the development of automated intelligent tutoring systems or in a hybrid system where both human and intelligent tutors co-exist.

Acknowledgments. This research was sponsored by a subcontract to The University of Memphis from Carnegie Learning, Inc., under award W911QY-15-C-0070 by Department of Defense. Any opinions, findings, and conclusions or recommendations expressed in this material are solely the authors'.

5. REFERENCES

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

[2] Boyer, K.E., Phillips, R., Ingram, A., Ha, E.Y., Wallis, M.D., Vouk, M.A., & Lester, J.C. (2011). Investigating the relationship between dialogue structure and tutoring effectiveness: a hidden Markov modeling approach, The *International Journal of Artificial Intelligence in Education (IJAIED)*, Vol. 21 No. 1, 2011, 65-81.

[3] Jurafsky, Dan.; and Martin, J.H. (2009). Speech and Language Processing. Prentice Hall, 2009.

[4] Moldovan, C., Rus, V., & Graesser, A.C. (2011). *Automated Speech Act Classification for Online Chat*, The 22nd Midwest Artificial Intelligence and Cognitive Science Conference, Cincinatti, OH, April 2011 (Best Student Paper Award - Honorary Mention).

[5] Morrison, D. M., Nye, B., Samei, B., Datla, V. V., Kelly, C., & Rus, V. (2014). *Building an Intelligent PAL from the Tutor.com Session Database-Phase 1: Data Mining*. The 7th International Conference on Educational Data Mining, 335-336.

[6] Rus, V., Graesser, A., Moldovan, C., & Niraula, N. (2012). *Automatic Discovery of Speech Act Categories in Educational Games*, 5th International Conference on Educational Data Mining (EDM12), June 19-21, Chania, Greece.

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