# Using Topic Models to Bridge Coding Schemes of Differing Granularity

Whitney L. Cade and Andrew Olney
{wlcade, aolney}@memphis.edu
Institute for Intelligent Systems, University of Memphis

Abstract. While Intelligent Tutoring Systems (ITSs) are often informed by the data extracted from tutoring corpora, coding schemes can be time consuming to implement. Therefore, an automatic classifier may make for quicker classifications. Dialogue from expert tutoring sessions were analyzed using a topic model to investigate how topics mapped on to pre-existing coding schemes of different granularities. These topics were then used to predict the classification of words into moves and modes. Ultimately, it was found that a decision tree algorithm outperformed several other algorithms in this classification task. Improvements to the classifier are discussed.

## 1 Introduction

While expert human-to-human tutoring is considered to be the most effective form of tutoring [2], human tutors are costly and in short supply. Therefore, researchers strive to understand their pedagogical techniques and implement them in an Intelligent Tutoring System (ITS). To understand what these techniques are and how they are implemented, corpus analysis is often used to study tutors. These data are noisy and complex by nature, but coding schemes are one method of understanding the data in a corpus. However, coding schemes take time and manpower to implement, and are not always cost-effective. Automatic tools are quicker and can also provide some information about a corpus. One automatic tool is the Latent Dirichlet allocation (LDA) model [1], or a topic model. This method is unsupervised and easily interpretable, and the output can be used to "tag" words as belonging to a certain semantic category. This makes the topic a possible feature that could be used in either a larger classifier or a manual coding scheme. In this study, we examine the possibility of using a word's topic as derived from a topic model to predict its label in two coding schemes of differing grain sizes.

## 2 Methods

We used a previously collected corpus of tutoring sessions conducted by expert tutors (see [3]). 40 tutoring sessions were recorded and transcribed, then coded according to two coding schemes. The move coding scheme (with 43 components) is a fine grained coding scheme, usually taking less than 1 conversational turn. It tends to capture small pedagogical and motivational phrases. The mode coding scheme (with 8 components) is a coarse-grained coding scheme, usually taking 10 or more turns, and it captures the overall structure of the tutoring session.

The transcripts were cleaned of common high-frequency words, i.e. "stop words", so that only content words remained. Each conversational turn was made into a single document. These transcripts were input to topic modeling software of our own design, with the

number of topics set to 100, the prior for topics appearing in a document (α) set to 1, and the prior for words appearing in a topic (β) set to 0.01. The topic model assigned each word in the corpus to a topic which was then paired with the move and mode category associated with each word. Preliminary explorations of the data revealed that three dialogue moves were highly gregarious; therefore, words from these categories were not used in either the move or the mode analyses. After every word's topic was paired with its corresponding move or mode, it was then formatted for the Weka machine learning toolkit. Weka allows comparison of several machine learning algorithms on a variety of dimensions such as percent classified correctly. We used Weka to answer the following question: given the topic, can the mode or move be predicted?

## 3 Results & Discussion

Five machine learning algorithms were chosen to classify the data, and one algorithm was selected to serve as the baseline for comparison (for further information on each algorithm, see [4]). They are: ZeroR (baseline algorithm, chooses the majority class), J48 (decision tree algorithm), IBk (k=10; nearest-neighbor learner), LogitBoost (boosting algorithm) and SMO (support vector machine algorithm). Ultimately, all algorithms performed better than the *ZeroR* baseline algorithm for both the move (baseline: 10.05% correct) and mode (baseline 50.15% accurate) coding scheme ($p < .05$). Although many algorithms post similar results, J48 has the advantage that its associated decision tree is easily interpretable. Its accuracy is 19.19% for the move coding scheme, and 52.30% for the mode coding scheme.

In conclusion, it seems that topics alone are a viable predictor for modes and moves, but they do not provide a full picture by themselves. A confusion matrix of each coding scheme's results reveals that moves and modes with high entropy (where the content relies heavily on the domain rather than a formulaic saying) are harder to predict than low entropy categories. Additionally, it seems apparent that modes are harder to predict than moves, which may be due to their context-dependent nature. These results may be useful first steps in building an online classifier for an ITS.

## References

[1] Blei. D., Ng, A., Jordan, M. Latent Dirichlet Allocation. *Journal of Machine Learning Research,* 2003, 3, p. 993–1022.

[2] Bloom, B. S. The 2 sigma problem: The Search for Methods of Group Instruction as Effective as One-to-One Tutoring. *Educational Researcher,* 1984, 13, p. 4 - 16.

[3] Person, N. K., Lehman, B., Ozbun, R. Pedagogical and Motivational Dialogue Moves Used by Expert Tutors. Presented at the *17th Annual Meeting of the Society for Text and Discourse,* 2007.

[4] Witten, I.H., Frank, E., *Data Mining: Practical Machine Learning Tools and Techniques with Java implementations,* 1998. San Francisco, CA: Morgan Kaufmann.