A Tutorial Markov Analysis of Effective Human Tutorial Sessions

Nabin Maharjan and Vasile Rus
Department of Computer Science / Institute for Intelligent Systems
The University of Memphis
Memphis, TN, USA
{nmharjan,vrus}@memphis.edu

Abstract

This paper investigates what differentiates effective tutorial sessions from less effective sessions. Towards this end, we characterize and explore human tutors' actions in tutorial dialogue sessions by mapping the tutor-tutee interactions, which are streams of dialogue utterances, into streams of actions, based on the language-as-action theory. Next, we use human expert judgment measures, *evidence of learning* (EL) and *evidence of soundness* (ES), to identify effective and ineffective sessions. We perform sub-sequence pattern mining to identify sub-sequences of dialogue modes that discriminate good sessions from bad sessions. We finally use the results of sub-sequence analysis method to generate a tutorial Markov process for effective tutorial sessions.

1 Introduction

Identifying effective instructional strategies, i.e., strategies that induce learning gains, has been a key research question in the Intelligent Tutoring Systems (ITSs) community. There are two common approaches used to address this research question: (1) hypothesize and validate through experimentation strategies guided by sound pedagogical theory (Aleven et al., 2001; Rus et al., 2017a) and (2) discover strategies employed by expert tutors (Boyer et al., 2011; Rus et al., 2015; Ohlsson et al., 2007). In our work, we adopt the latter approach which typically consists of mining patterns associated with successful tutorial sessions in large collections of recorded human tutoring sessions. (Boyer et al., 2011; Cade et al., 2008; Rus et al., 2015; Ohlsson et al., 2007).

It is important to note that discovering effective tutoring strategies by studying the strategies used by expert tutors is challenging because what characterizes tutoring expertise is still an open question to some degree (Rus et al., 2015). A tutor who employs sound strategies may appear less expert when working with students having low abilities or lacking in motivation. On the other hand, an average tutor may seem expert if s/he only works with high ability and highly motivated students. We lack student ability and prior knowledge information in our data and therefore focus on *effective tutoring* rather than *expert tutoring*. *Effective tutoring* refers to tutoring that yields learning gains. In sum, we study in this paper strategies of effective tutors as reflected in effective tutorial sessions.

In this paper, we worked with an annotated corpus of human tutoring sessions from which we identified effective sessions based on human expert judgments (see Section 5). We mapped tutorial sessions onto sequences of dialogue acts and dialogue modes (Cade et al., 2008), explained later, using a predefined coding taxonomy (see Section 3). We then conducted sub-sequence pattern mining to identify sub-sequences of dialogue modes that occur in effective tutoring sessions but not in ineffective tutoring sessions. We used these distinctive sub-sequences of modes to build a Markov process for effective tutorial sessions. Finally, using the tutorial Markov process, we analyzed and searched for dialogue mode patterns associated with effective tutoring sessions.

2 Related Work

Discovering the structure of tutorial dialogues and tutors' strategies has been a main goal of the intelligent tutoring research community from the very beginning because such tutorial session structures and strategies must be understood in order to be replicated in the development of effective

Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications, pages 30–34
Melbourne, Australia, July 19, 2018.

ITSs. Graesser et al. (1995) proposed a five-step general structure of collaborative problem solving during tutoring. Cade et al. (2008) examined likely sequences of dialogue modes in expert tutoring. Boyer and colleagues (2009; 2010) applied hidden Markov models to discover effective dialogue modes inherent in the tutoring sessions. Rus et al. (2017b) used a supervised machine learning method to automatically map tutorial sessions into dialogue acts, sub-acts and modes and then analyzed human tutoring sessions using profile comparison and sequence logos to discover effective tutorial strategies in terms of dialogue acts and modes. Our work further contributes to this area of research by characterizing effective tutorial sessions in terms of dialogue mode sequential patterns and tutorial Markov processes.

3 Coding Taxonomy

The role of the coding taxonomy is to help us map tutors and tutees' utterances in tutorial dialogues onto actions, i.e., dialogue acts, based on the language-as-action theory (Austin, 1975; Searle, 1969) according to which *when we say something we do something*. For example, the utterance: *"There is an useful idea called 'conservation of energy"'* is categorized as an *Assertion* dialogue act, i.e., the utterance is making an assertion. Because the assertion is about *"conservation of energy"*, a *Concept*, we consider this as a specialized assertion about a concept, i.e., an Assertion-Concept dialogue act-subact combination.

We group sequences of dialogue acts and sub acts into higher level constructs, i.e., dialogue modes. Dialogue modes represent contiguous sequences of dialogue acts-subacts that together serve particular pedagogical purposes, e.g., a sequence of hints in the form of questions may reflect a scaffolding instructional strategy in which the tutee works mostly by herself on the current instructional task while the tutor offers help, when needed, through such hints.

The dialogue act and mode taxonomies are adapted to our context from a set of earlier taxonomies which were created to analyze a large corpus of online tutoring sessions conducted by human tutors in the domains of Algebra and Physics (Morrison et al., 2014). There are 17 top level expert-defined dialogue act categories. Each dialogue act category may have 4 to 22 subcategories or sub-acts. For example, we distinguish *Assertions* that reference aspects of the tutorial process itself *(Assertion:Process)*; domain concepts (*Assertion:Concept*), or the the use of lower-level mathematical calculations (*Assertion:Calculation*). Further, we have a set of 17 dialogue modes: *Assessment*, *Closing*, *Fading*, *ITSupport*, *Metacognition*, *MethodID*, *Modeling*, *OffTopic*, *Opening*, *ProblemID*, *ProcessNegotiation*,*RapportBuilding*, *RoadMap*, *SenseMaking*, *Scaffolding*, *SessionSummary* and *Telling*. A detailed description of the dialogue modes is described by Morrison and colleagues (Morrison et al., 2014).

Annotation	Agreement (%)	Kappa
Act	77	0.72
Act-subact	62	0.60
Mode	44	0.37
Mode*	53.8	0.48
Mode**	64.3	0.60

Table 1: Average Inter Annotator Agreement Between Two Independent Annotators. Mode* and Mode** represent dialogue mode agreement between verifier and first annotator and, verifier and second annotator respectively.

4 Data

A large corpus of about 19K tutorial sessions between professional human tutors and actual college-level, adult students on Algebra domain was collected via an online human tutoring service. 500 tutorial sessions containing 31,299 utterances were randomly selected for annotation. The sessions were manually labeled by a team of 6 subject matter experts (SMEs). They were trained on the taxonomy of dialogue acts, sub-acts, and modes. Each session was manually tagged by two independent annotators without looking at each other's tags to eliminate any labeling bias problems. The tags of the two independent annotators were double-checked by a verifier who resolved discrepancies in tags, if any. It should be noted that though the average inter-annotator agreement is apparently low, the final agreement of annotators with the verifier is higher (see Table 1). The verifier also happened to be the designer of our dialogue taxonomy. In this paper, dialogue mode should be interpreted as dialogue mode-switch.

Mode sub-sequences (p-value in bracket)
Fading-Closing (0.0002)
Scaffolding-Scaffolding (0.0008)
Fading (0.0009)
Scaffolding-Fading (0.0055)
Fading-Fading (0.01)
ProblemID-Fading (0.02)
ProblemID-Scaffolding-Scaffolding(0.0362)
Closing (0.05)
Fading-RapportBuilding (0.05)
Fading-ProcessNegotiation (0.06)
Fading-Scaffolding (0.07)
Fading-ProcessNegotiation-Closing (0.07)
Fading-Fading-Scaffolding (0.08)
ProblemID-Fading-Scaffolding (0.10)
Fading-Scaffolding-Fading (0.14)
Scaffolding-Closing (0.15)
Opening-ProblemID-Fading (0.18)
RapportBuilding (0.18)
Fading-Scaffolding-Closing (0.18)
Scaffolding-Fading-Scaffolding (0.20)
Scaffolding-RapportBuilding (0.23)
Fading-Scaffolding-ProcessNegotiation (0.26)
RapportBuilding-Closing(0.37)

Table 2: Discriminant mode sub-sequences.

5 Markov Analysis of Tutorial Sessions

In order to identify effective sessions, the SMEs also rated each tutorial session using a 1-5 scale (5 being best score) along two dimensions: evidence of learning (EL) and evidence of soundness (ES). The ES score is supposed to measure the degree to which the tutor applied pedagogically sound tactics. On the other hand, the EL score indicates whether there is strong evidence that the student learned from the tutoring session. The ES and EL distinction was designed in order to separate confounding factors such as learners' engagement in the session. For instance, a tutor may apply sound pedagogy but the student may not learn as all they might be interested is to find a quick answer to their (homework) problem. It should be noted that most of the sessions we had access to were in the context of homework help. That is, students start a session by asking for help with a particular problem. While EL and ES were supposed to capture different things, the EL and ES scores were found to be highly correlated (Pearson co-efficient of 0.7).

We used both EL and ES to capture overall quality of tutoring sessions. We categorized all sessions having ES and EL scores ≤ 2 as ineffective, and all sessions rated with ES = 5 and EL ≥ 4 as effective.

We conducted discriminant mode sub-sequence analysis using *Traminer* package in R. It should be noted that a sub-sequence is not necessarily a contiguous sequence of observations, however, the order of the observations is preserved. For example, (*Fading*)-(*Closing*) is a valid sub-sequence of dialogue modes formed from the (*Fading*)-(*ProcessNegotiation*)-(*Closing*) contiguous session fragment. We generated sub-sequences up to length 7 from all annotated tutorial sessions.

The *Traminer* algorithm first finds the most frequent sub-sequences by counting their distinct occurrences and then applies a Chi-squared test (Bonferroni-adjusted) to identify sub-sequences that are statistically more (or less) frequent in each group. We used a p-value < 0.4 to generate a sufficient number of likely distinctive sub-sequences of modes (Table 2). Once the significant sub-sequences were identified, we generated a state transition matrix, explained next.

5.1 State Transition Matrix

We created a state transition matrix with modes as the states. We ignored sub-sequences of unit length as they don't indicate an observed transition. For sub-sequences spanning more than two states, we split them into multiple bigram sub-sequences. For example, we obtained bigram sub-sequences *Opening-ProblemID* and *ProblemID-Fading* from the *Opening-ProblemID-Fading* sub-sequence. We discarded self-transition paths since modes are actually mode switches in our case. Therefore, we discarded transition path *Scaffolding-Scaffolding* from the *ProblemID-Scaffolding-Scaffolding* sub-sequence.

We used confident scores of discriminant sub-sequences to compute transition probabilities. The confidence score of a discriminant sub-sequence is the probability that the sub-sequence is coming from an effective session, i.e., 1 - p-value. We computed a confidence score of a path as the confidence score of the discriminant sub-sequence the path belongs to. For example, for *Opening-ProblemID-Fading* (0.18), the confidence score of paths *Opening-ProblemID* and *ProblemID-Fading* is 1-0.18 = 0.82. We weighted an edge as the cumulative sum of its confi-

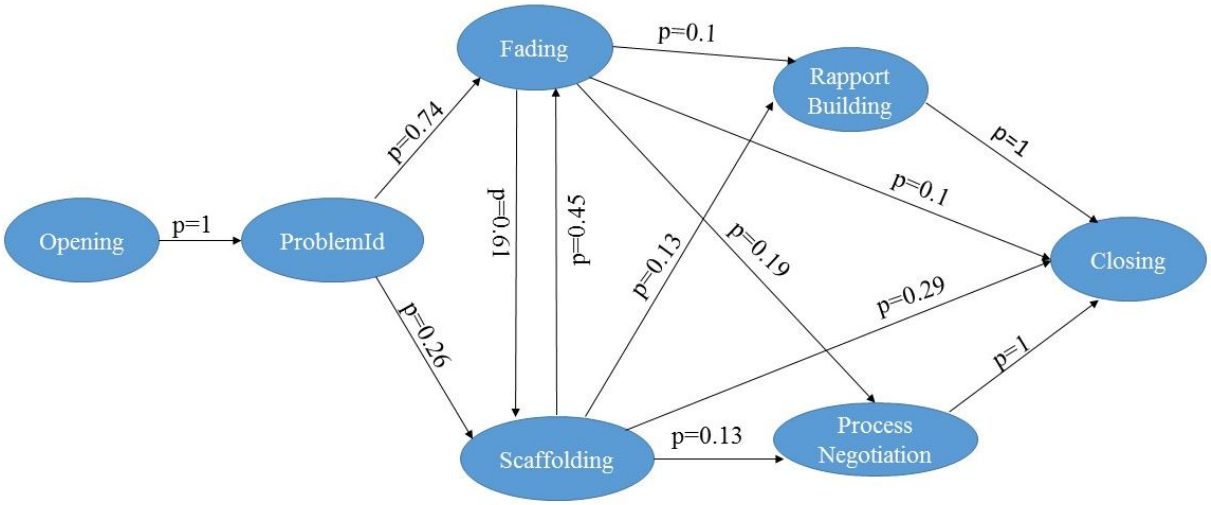

Figure 1: Tutorial Markov Process for effective tutorial sessions.

dence scores from all the sub-sequences where the path is present. For example, path *ProblemID-Fading* is present in sub-sequences *ProblemID-Fading* (0.02) and *Opening-ProblemID-Fading* (0.18). So, the weight of the path *ProblemID-Fading* is: 0.98 + 0.82 = 1.8. Finally, we normalized the weight of each path A-B to represent transition probability by dividing it by the sum of the weights of all possible transitions from A. For example, the weight of the path *ProblemID-Fading* is normalized by dividing it by the sum total of weights of all the transitions from *ProblemID* state.

5.2 Tutorial Markov Interpretation

Figure 1 shows the state transition graph of the underlying Markov process corresponding to the above state transition matrix. In the figure, the states are dialogue modes whereas transitions are generated using only the discriminant sub-sequences of modes. Each path has been labeled with the corresponding transition probabilities.

The Markov process reveals that any sequence of modes it can generate starts with an *Opening* and ends with a *Closing* state and is likely to have a large number of *Scaffolding - Fading* switches/transitions. This result partly supports theoretical expert tutoring models based on the *modeling-scaffolding-fading* paradigm (Rogoff and Lave, 1984). The high occurrences of these modes provide evidence that effective tutors monitor and engage students more and provide help only when needed. Cade et al. (2008) also found that *Scaffolding* was a highly occurring mode in expert tutoring. They found a relatively low occurrence of the *Fading* mode, which they suggested might be explained by time constraints, i.e. the tutoring session prevented tutors from spending too much time in the *Fading* mode.

The Markov process also resembles to some degree Graesser's (Graesser et al., 1995) 5-step dialogue framework, which captures the tutorial phases prevalent in collaborative problem-solving tutoring: *i) Tutor asks question, ii) Student answers question, iii) Tutor gives short feedback, iv)Tutor and student collaboratively improve the quality of the answer, v) Tutor assesses student's understanding*. One probable effective tutorial path from the Markov process, which might be comparable to Graesser's framework, is *Opening - ProblemId - Fading - Scaffolding - Fading - ProcessNegotiation - Closing*. Indeed, the sub-path *ProblemId - Fading - Scaffolding - Fading - ProcessNegotiation* resembles Graesser's 5-step framework.

The first 3 phases in Graesser's framework don't align with the initial modes of the suggested learning path. This might be because of the difference in the tutoring environment. Graesser assumed tutor-driven sessions, which with a tutor first asking a question or presenting a problem for the learner to solve, followed by a student answer, etc. In our case, it is the students who are seeking help from tutors on specific problems. Initially, in our case, the tutor works together with the student to understand the problem (*ProblemId*). Then, the tutor fades, allowing the student to work on the problem by herself (*Fading*). The tutor may switch

between *Scaffolding* and *Fading* to provide help (*Scaffolding*), only when needed. In this sense, the last two elements in Graesser's framework can be considered to be aligned with the *Scaffolding - Fading* pattern.

The additional benefit of this Markov process representation is that it suggests multiple possible paths or meta-strategies that can lead to learning gains.

6 Conclusion

We used human expert judgment scores to identify effective and ineffective tutoring sessions. We conducted discriminant mode sub-sequence analysis based on which we generated a Markov process for effective tutorial sessions. We found that sequences of dialogue modes derived from the Markov process are most likely to have many *Scaffolding* and *Fading* modes. Furthermore, the inferred Markov process suggests a new model for tutoring when students ask for help as opposed to tutor-driven sessions, which was modeled in the past. Our future work is to expand our understanding of the effective strategies in effective tutorial sessions while accounting for other factors such as students' prior knowledge.

Acknowledgments

This work was partially supported by The University of Memphis and a contract from the Advanced Distributed Learning Initiative of the United States Department of Defense.

References

Vincent Aleven, Octav Popescu, and Kenneth R Koedinger. 2001. Towards tutorial dialog to support self-explanation: Adding natural language understanding to a cognitive tutor. In *Proceedings of Artificial Intelligence in Education*, pages 246–255.

John Langshaw Austin. 1975. *How to do things with words*. Oxford university press.

Kristy Elizabeth Boyer, Eunyoung Ha, Michael D Wallis, Robert Phillips, Mladen A Vouk, and James C Lester. 2009. Discovering tutorial dialogue strategies with hidden markov models. In *AIED*, pages 141–148.

Kristy Elizabeth Boyer, Robert Phillips, Amy Ingram, Eun Young Ha, Michael Wallis, Mladen Vouk, and James Lester. 2010. Characterizing the effectiveness of tutorial dialogue with hidden markov models. In *International Conference on Intelligent Tutoring Systems*, pages 55–64. Springer.

Kristy Elizabeth Boyer, Robert Phillips, Amy Ingram, Eun Young Ha, Michael Wallis, Mladen Vouk, and James Lester. 2011. Investigating the relationship between dialogue structure and tutoring effectiveness: a hidden markov modeling approach. *International Journal of Artificial Intelligence in Education*, 21(1-2):65–81.

Whitney Cade, Jessica Copeland, Natalie Person, and Sidney DMello. 2008. Dialogue modes in expert tutoring. In *Intelligent tutoring systems*, pages 470–479. Springer.

Arthur C Graesser, Natalie K Person, and Joseph P Magliano. 1995. Collaborative dialogue patterns in naturalistic one-to-one tutoring. *Applied cognitive psychology*, 9(6):495–522.

Donald Morrison, Benjamin Nye, Borhan Samei, Vivek Varma Datla, Craig Kelly, and Vasile Rus. 2014. Building an intelligent pal from the tutor. com session database phase 1: Data mining. In *Educational Data Mining 2014*.

Stellan Ohlsson, Barbara Di Eugenio, Bettina Chow, Davide Fossati, Xin Lu, and Trina C Kershaw. 2007. Beyond the code-and-count analysis of tutoring dialogues. *Artificial intelligence in education: Building technology rich learning contexts that work*, 158:349.

Barbara Ed Rogoff and Jean Ed Lave. 1984. *Everyday cognition: Its development in social context.* Harvard University Press.

Vasile Rus, Rajendra Banjade, Nobal Niraula, Elizabeth Gire, and Donald Franceschetti. 2017a. A study on two hint-level policies in conversational intelligent tutoring systems. In *Innovations in Smart Learning*, pages 171–181. Springer.

Vasile Rus, Nabin Maharjan, and Rajendra Banjade. 2015. Unsupervised discovery of tutorial dialogue modes in human-to-human tutorial data. In *Proceedings of the Third Annual GIFT Users Symposium*, pages 63–80.

Vasile Rus, Nabin Maharjan, Tamang Lasang, Michael Yudelson, Susan Berman, Fancsali Stephen, and Steve Ritter. 2017b. An analysis of human tutors actions in tutorial dialogues. In *Proceedings of the Thirtieth International Florida Artificial Intelligence Research Society Conference*.

John R Searle. 1969. *Speech acts: An essay in the philosophy of language*, volume 626. Cambridge university press.