Exploiting Discourse Structure for Spoken Dialogue Performance Analysis

Mihai Rotaru
University of Pittsburgh
Pittsburgh, USA
mrotaru@cs.pitt.edu

Diane J. Litman
University of Pittsburgh
Pittsburgh, USA
litman@cs.pitt.edu

Abstract

In this paper we study the utility of discourse structure for spoken dialogue performance modeling. We experiment with various ways of exploiting the discourse structure: in isolation, as context information for other factors (correctness and certainty) and through trajectories in the discourse structure hierarchy. Our correlation and PARADISE results show that, while the discourse structure is not useful in isolation, using the discourse structure as context information for other factors or via trajectories produces highly predictive parameters for performance analysis.

1 Introduction

Predictive models of spoken dialogue system (**SDS**) performance are an important tool for researchers and practitioners in the SDS domain. These models offer insights on what factors are important for the success of a SDS and allow researchers to assess the performance of future system improvements without running additional costly user experiments.

One of the most popular models of performance is the PARADISE framework proposed by (Walker et al., 2000). In PARADISE, a set of *interaction parameters* are measured in a SDS corpus, and then used in a multivariate linear regression to predict the target performance metric. A critical ingredient in this approach is the relevance of the interaction parameters for the SDS success. A number of parameters that measure the dialogue efficiency (e.g. number of system/user turns, task duration) and the dialogue quality (e.g. recognition accuracy, rejections, helps) have been shown to be successful in (Walker et al., 2000). An extensive set of parameters can be found in (Möller, 2005a).

In this paper we study the utility of *discourse structure* as an information source for SDS performance analysis. The discourse structure hierarchy has been shown to be useful for other tasks: understanding specific lexical and prosodic phenomena (Hirschberg and Nakatani, 1996; Levow, 2004), natural language generation (Hovy, 1993), predictive/generative models of postural shifts (Cassell et al., 2001), and essay scoring (Higgins et al., 2004).

We perform our analysis on a corpus of speech-based tutoring dialogues. A tutoring SDS (Litman and Silliman, 2004; Pon-Barry et al., 2004) has to discuss concepts, laws and relationships and to engage in complex subdialogues to correct student misconceptions. As a result, dialogues with such systems have a rich discourse structure.

We perform three experiments to measure three ways of exploiting the discourse structure. In our first experiment, we test the predictive utility of the discourse structure in itself. For example, we look at whether the number of pop-up transitions in the discourse structure hierarchy predicts performance in our system.

The second experiment measures the utility of the discourse structure as contextual information for two types of *student states*: correctness and certainty. The intuition behind this experiment is that interaction events should be treated differently based on their position in the discourse structure hierarchy. For example, we test if the number of incorrect answers after a pop-up transition has a higher predictive utility than the total number of incorrect student answers. In contrast, the majority of the previous work either ignores this contextual information (Möller, 2005a; Walker et al., 2000) or makes limited use of the

Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006), pages 85–93,
Sydney, July 2006. ©2006 Association for Computational Linguistics

discourse structure hierarchy by flattening it (Walker et al., 2001) (Section 5).

As another way to exploit the discourse structure, in our third experiment we look at whether specific trajectories in the discourse structure are indicative of performance. For example, we test if two consecutive pushes in the discourse structure are correlated with higher learning.

To measure the predictive utility of our interaction parameters, we focus primarily on *correlations* with our performance metric (Section 4). There are two reasons for this. First, a significant correlation between an interaction parameter and the performance metric is a good indicator of the parameter's relevance for PARADISE modeling. Second, correlations between factors and the performance metric are commonly used in tutoring research to analyze the tutoring/learning process (Chi et al., 2001).

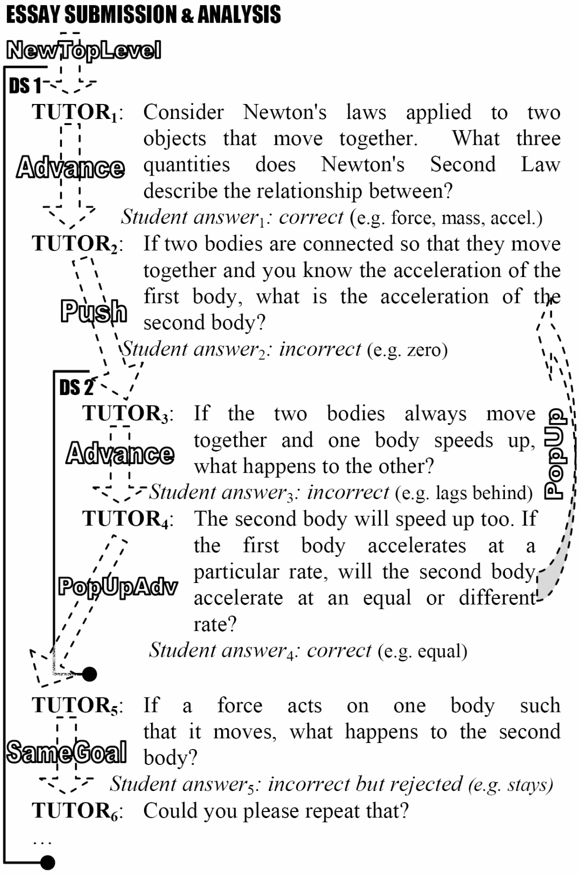
Our correlation and PARADISE results show that, while the discourse structure is not useful in isolation, using the discourse structure as context information for other factors or via trajectories produces highly predictive parameters for performance analysis.

2 Annotation

Our annotation for discourse structure and student state has been performed on a corpus of 95 experimentally obtained spoken tutoring dialogues between 20 students and our system **ITSPOKE** (Litman and Silliman, 2004). ITSPOKE is a speech-enabled version of the text-based Why2-Atlas conceptual physics tutoring system (VanLehn et al., 2002). When interacting with ITSPOKE, students first type an essay answering a qualitative physics problem using a graphical user interface. ITSPOKE then engages the student in spoken dialogue (using head-mounted microphone input and speech output) to correct misconceptions and elicit more complete explanations, after which the student revises the essay, thereby ending the tutoring or causing another round of tutoring/essay revision. Each student went through the same procedure: 1) read a short introductory material, 2) took a pretest to measure the initial physics knowledge, 3) work through a set of 5 problems with ITSPOKE, and 4) took a posttest similar to the pretest. The resulting corpus had 2334 student turns and a comparable number of system turns.

2.1 Discourse structure

We base our annotation of discourse structure on the Grosz & Sidner theory of discourse structure (Grosz and Sidner, 1986). A critical ingredient of this theory is the intentional structure. According to the theory, each discourse has a discourse purpose/intention. Satisfying the main discourse purpose is achieved by satisfying several smaller purposes/intentions organized in a hierarchical structure. As a result, the discourse is segmented in discourse segments each with an associated discourse segment purpose/intention. This theory has inspired several generic dialogue managers for spoken dialogue systems (Bohus and Rudnicky, 2003).



TUTOR$_1$: Consider Newton's laws applied to two objects that move together. What three quantities does Newton's Second Law describe the relationship between?

Student answer$_1$: correct (e.g. force, mass, accel.)

TUTOR$_2$: If two bodies are connected so that they move together and you know the acceleration of the first body, what is the acceleration of the second body?

Student answer$_2$: incorrect (e.g. zero)

TUTOR$_3$: If the two bodies always move together and one body speeds up, what happens to the other?

Student answer$_3$: incorrect (e.g. lags behind)

TUTOR$_4$: The second body will speed up too. If the first body accelerates at a particular rate, will the second body accelerate at an equal or different rate?

Student answer$_4$: correct (e.g. equal)

TUTOR$_5$: If a force acts on one body such that it moves, what happens to the second body?

Student answer$_5$: incorrect but rejected (e.g. stays)

TUTOR$_6$: Could you please repeat that?

…

Figure 1. The discourse structure and transition annotation

We automate our annotation of the discourse structure by taking advantage of the structure of the tutored information. A dialogue with ITSPOKE follows a question-answer format (i.e. system initiative): ITSPOKE asks a question, the student provides the answer and then the process is repeated. Deciding what question to ask, in what order and when to stop is hand-authored beforehand in a hierarchical structure that resembles the discourse segment structure (see Figure 1). Tutor questions are grouped in segments which correspond roughly to the discourse segments. Similarly to the discourse segment purpose, each question segment has an associated tutoring goal or purpose. For example, in

ITSPOKE there are question segments discussing about forces acting on the objects, others discussing about objects' acceleration, etc.

In Figure 1 we illustrate ITSPOKE's behavior and our discourse structure annotation. First, based on the analysis of the student essay, ITSPOKE selects a question segment to correct misconceptions or to elicit more complete explanations. This question segment will correspond to the top level discourse segment (e.g. DS1). Next, ITSPOKE asks the student each question in DS1. If the student answer is correct, the system moves on to the next question (e.g. $Tutor_1 \rightarrow Tutor_2$). If the student answer is incorrect, there are two alternatives. For simple questions, the system will simply give out the correct answer and move on to the next question (e.g. $Tutor_3 \rightarrow Tutor_4$). For complex questions (e.g. applying physics laws), ITSPOKE will engage into a *remediation subdialogue* that attempts to remediate the student's lack of knowledge or skills. The remediation subdialogue is specified in another question segment and corresponds to a new discourse segment (e.g DS2). The new discourse segment is dominated by the current discourse segment (e.g. DS2 dominated by DS1). $Tutor_2$ system turn is a typical example; if the student answers it incorrectly, ITSPOKE will enter discourse segment DS2 and go through its questions ($Tutor_3$ and $Tutor_4$). Once all the questions in DS2 have been answered, a heuristic determines whether ITSPOKE should ask the original question again ($Tutor_2$) or simply move on to the next question ($Tutor_5$).

To compute interaction parameters from the discourse structure, we focus on the transitions in the discourse structure hierarchy. For each system turn we define a **transition** feature. This feature captures the position in the discourse structure of the current system turn relative to the previous system turn. We define six labels (see Table 1). **NewTopLevel** label is used for the first question after an essay submission (e.g. $Tutor_1$). If the previous question is at the same level with the current question we label the current question as **Advance** (e.g. $Tutor_{2,4}$). The first question in a remediation subdialogue is labeled as **Push** (e.g. $Tutor_3$). After a remediation subdialogue is completed, ITSPOKE will pop up and it will either ask the original question again or move on to the next question. In the first case, we label the system turn as **PopUp**. Please note that $Tutor_2$ will not be labeled with PopUp because, in such cases, an extra system turn will be created between $Tutor_4$ and $Tutor_5$ with the same content as $Tutor_2$. In addition, variations of "Ok, back to the original question" are also included in the new system turn to mark the discourse segment boundary transition. If the system moves on to the next question after finishing the remediation subdialogue, we label the system turn as **PopUpAdv** (e.g. $Tutor_5$). Note that while the sum of PopUp and PopUpAdv should be equal with Push, it is smaller in our corpus because in some cases ITSPOKE popped up more than one level in the discourse structure hierarchy. In case of rejections, the system question is repeated using variations of "Could you please repeat that?". We label such cases as **SameGoal** (e.g. $Tutor_6$).

Discourse structure transitions	
Advance	53.4%
NewTopLevel	13.5%
PopUp	9.2%
PopUpAdv	3.5%
Push	14.5%
SameGoal	5.9%
Certainty	
Certain	41.3%
Uncertain	19.1%
Mixed	2.4%
Neutral	37.3%
Correctness	
Correct	63.3%
Incorrect	23.3%
Partially Correct	6.2%
Unable to Answer	7.1%

Table 1: Transition and student state distribution.

Please note that each student dialogue has a specific discourse structure based on the dialogue that dynamically emerges based on the correctness of her answers. For this reason, the same system question in terms of content may get a different transition label for different students. For example, in Figure 1, if the student would have answered $Tutor_2$ correctly, the next tutor turn would have had the same content as $Tutor_5$ but the Advance label. Also, while a human annotation of the discourse structure will be more complex but more time consuming (Hirschberg and Nakatani, 1996; Levow, 2004), its advantages are outweighed by the automatic nature of our discourse structure annotation.

We would like to highlight that our transition annotation is *domain independent* and *automatic*. Our transition labels capture behavior like starting a new dialogue (NewTopLevel), crossing discourse segment boundaries (Push, PopUp, PopUpAdv) and local phenomena inside a discourse segment (Advance, SameGoal). If the discourse structure information is available, the transition information can be automatically computed using the procedure described above.

2.2 Student state

Because for our tutoring system student learning is the relevant performance metric, we hypothesize that information about student state in each student turn, in terms of correctness and certainty, will be an important indicator. For example, a student being more correct and certain during her interaction with ITSPOKE might be indicative of a higher learning gain. Also, previous studies have shown that tutoring specific parameters can improve the quality of SDS performance models that model the learning gain (Forbes-Riley and Litman, 2006).

In our corpus, each student turn was manually labeled for *correctness* and *certainty* (Table 1). While our system assigns a correctness label to each student turn to plan its next move, we choose to use a manual annotation of correctness to eliminate the noise introduced by the automatic speech recognition component and the natural language understanding component. A human annotator used the human transcripts and his physics knowledge to label each student turn for various degrees of correctness: correct, partially correct, incorrect and unable to answer. “Unable to Answer” label was used for turns where the student did not answer the system question or used variants of “I don’t know”.

Previous work has shown that certainty plays an important role in the learning and tutoring process (Pon-Barry et al., 2006; VanLehn et al., 2003). A human annotator listened to the dialogues between students and ITSPOKE and labeled each student turn for its perceived degree of certainness. Four labels were used: certain, uncertain, neutral and mixed (both certain and uncertain). To date, one annotator has labeled all student turns in our corpus[1].

3 Interaction parameters

For each user, interaction parameters measure specific aspects of the dialogue with the system. We use our transition and student state annotation to create two types of interaction parameters: **unigrams** and **bigrams**. The difference between the two types of parameters is whether the discourse structure context is used or not. For each of our 12 labels (4 for correctness, 4 for certainty and 6 for discourse structure), we derive two unigram parameters per student over the 5 dialogues for that student: a *total* parameter and a *percentage* parameter. For example, for the ‘Incorrect’ unigram we compute, for each student, the total number of student turns labeled with ‘Incorrect’ (parameter Incorrect) and the percentage of such student turns out of all student turns (parameter Incorrect%). For example, if we consider only the dialogue in Figure 1, Incorrect = 3 (Student$_{2,3,5}$) and Incorrect% = 60% (3 out of 5).

Bigram parameters exploit the discourse structure context. We create two classes of bigram parameters by looking at *transition–student state* bigrams and *transition–transition* bigrams. The transition–student state bigrams combine the information about the student state with the transition information of the previous system turn. Going back to Figure 1, the three incorrect answers will be distributed to three bigrams: Advance–Incorrect (Tutor$_2$–Student$_2$), Push–Incorrect (Tutor$_3$–Student$_3$) and PopUpAdv–Incorrect (Tutor$_5$–Student$_5$). The transition–transition bigram looks at the transition labels of two consecutive system turns. For example, the Tutor$_4$–Tutor$_5$ pair will be counted as an Advance–PopUpAdv bigram.

Similar to the unigrams, we compute a total parameter and a percentage parameter for each bigram. The percentage denominator is number of student turns for the transition–student state bigrams and the number of system turns minus one for the transition–transition bigram. In addition, for each bigram we compute a *relative percentage* parameter (bigram followed by %rel) by computing the percentage relative to the total number of times the transition unigram appears for that student. For example, we will compute the Advance–Incorrect %rel parameter by dividing the number of Advance–Incorrect bigrams with the number of Advance unigrams (1 divided by 2 in Figure 1); this value will capture the percentage of times an Advance transition is followed by an incorrect student answer.

4 Results

We use student learning as our evaluation metric because it is the primary metric for evaluating the performance of tutoring systems. Previous work (Forbes-Riley and Litman, 2006) has suc-

[1] The agreement between the manual correctness annotation and the correctness assigned by ITSPOKE is 90% (kappa of 0.79). In a preliminary agreement study, a second annotator labeled our corpus for a binary version of certainty (uncertainty versus other), resulting in a 90% inter-annotator agreement and a kappa of 0.68.

cessfully used student learning as the performance metric in the PARADISE framework. Two quantities are used to measure student learning: the pretest score and the posttest score. Both tests consist of 40 multiple-choice questions; the test's score is computed as the percentage of correctly answered questions. The average score and standard deviation for each test are: pretest 0.47 (0.17) and posttest 0.68 (0.17).

We focus primarily on correlations between our interaction parameters and student learning. Because in our data the pretest score is significantly correlated with the posttest score, we study *partial* Pearson's correlations between our parameters and the posttest score that account for the pretest score. This correlation methodology is commonly used in the tutoring research (Chi et al., 2001). For each trend or significant correlation we report the unigram/bigram, its average and standard deviation over all students, the Pearson's Correlation Coefficient (R) and the statistical significance of R (p).

First we report significant correlations for unigrams to test our first hypothesis. Next, for our second and third experiment, we report correlations for transition–student state and transition–transition parameters. Finally, we report our preliminary results on PARADISE modeling.

4.1 Unigram correlations

In our first proposed experiment, we want to test the predictive utility of discourse structure in isolation. We compute correlations between our transition unigram parameters and learning. We find no trends or significant correlations. This result suggests that discourse structure in isolation has no predictive utility.

Here we also report all trends and significant correlations for student state unigrams as the baseline for contextual correlations to be presented in Section 4.2. We find only one significant correlation (Table 2): students with a higher percentage of neutral turns (in terms of certainty) are negatively correlated with learning. We hypothesize that this correlation captures the student involvement in the tutoring process: more involved students will try harder thus expressing more certainty or uncertainty. In contrast, less involved students will have fewer certain/uncertain/mixed turns and, in consequence, more neutral turns. Surprisingly, student correctness does not significantly correlate with learning.

Parameter	Mean (SD)	R.	p
Neutral %	37% (8%)	-.47	.04

Table 2: Trend and significant unigram correlations

4.2 Transition–student state correlations

For our second experiment, we need to determine the predictive utility of transition–student state bigram parameters. We find a large number of correlations for both transition–correctness bigrams and transition–certainty bigrams.

Transition–correctness bigrams

This type of bigram informs us whether accounting for the discourse structure transition when looking at student correctness has any predictive value. We find several interesting trends and significant correlations (Table 3).

The student behavior, in terms of correctness, after a PopUp or a PopUpAdv transition is very informative about the student learning process. In both situations, the student has just finished a remediation subdialogue and the system is popping up either by reasking the original question again (PopUp) or by moving on to the next question (PopUpAdv). We find that after PopUp, the number of correct student answers is positively correlated with learning. In contrast, the number, the percentage and the relative percentage of incorrect student answers are negatively correlated with learning. We hypothesize that this correlation indicates whether the student took advantage of the additional learning opportunities offered by the remediation subdialogue. By answering correctly the original system question (PopUp–Correct), the student demonstrates that she has absorbed the information from the remediation dialogue. This bigram is an indication of a successful learning event. In contrast, answering the original system question incorrectly (PopUp–Incorrect) is an indication of a missed learning opportunity; the more events like this happen the less the student learns.

Parameter	Mean (SD)	R.	p
PopUp–Correct	7 (3.3)	.45	.05
PopUp–Incorrect	2 (1.8)	-.42	.07
PopUp–Incorrect %	1.6% (1.2%)	-.46	.05
PopUp–Incorrect %rel	17% (13%)	-.39	.10
PopUpAdv–Correct	2.5 (2)	.43	.06
PopUpAdv–Correct %	2% (1.3%)	.52	.02
NewTopLevel–Incorrect	2.3 (1.8)	.56	.01
NewTopLevel–Incorrect %	1.9% (1.4%)	.49	.03
NewTopLevel–Incorrect %rel	15% (12%)	.51	.02
Advance–Correct	40.5 (9.8)	.45	.05

Table 3: Trend and significant transition–correctness bigram correlations

Similarly, being able to correctly answer the tutor question after popping up from a remediation subdialogue (PopUpAdv–Correct) is positively correlated with learning. Since in many cases, these system questions will make use of the knowledge taught in the remediation subdialogues, we hypothesize that this correlation also captures successful learning opportunities.

Another set of interesting correlations is produced by the NewTopLevel–Incorrect bigram. We find that the number, the percentage and the relative percentage of times ITSPOKE starts a new essay revision dialogue that results in an incorrect student answer is positively correlated with learning. The content of the essay revision dialogue is determined based on ITSPOKE's analysis of the student essay. We hypothesize that an incorrect answer to the first tutor question is indicative of the system's picking of a topic that is problematic for the student. Thus, we see more learning in students for which more knowledge gaps are discovered and addressed by ITSPOKE.

Finally, we find the number of times the student answers correctly after an advance transition is positively correlated with learning (the Advance–Correct bigram). We hypothesize that this correlation captures the relationship between students that advance without having major problems and a higher learning gains.

Transition–certainty bigrams

Next we look at the combination between the transition in the dialogue structure and the student certainty (Table 4). These correlations offer more insight on the negative correlation between the Neutral % unigram parameter and student learning. We find that out of all neutral student answers, those that follow an Advance transitions are negatively correlated with learning. Similar to the Neutral % correlation, we hypothesize that Advance–Neutral correlations capture the lack of involvement of the student in the tutoring process. This might be also due to ITSPOKE engaging in teaching concepts that the student is already familiar with.

Parameter	Mean (SD)	R.	p
Advance–Neutral	27 (8.3)	-.40	.08
Advance–Neutral %	21% (6%)	-.62	.00
Advance–Neutral %rel	38% (10%)	-.73	.00
SameGoal–Neutral %rel	35% (31%)	.46	.05

Table 4: Trend and significant transition–certainty bigram correlations

In contrast, staying neutral in terms of certainty after a system rejection is positively correlated with learning. These correlations show that based on their position in the discourse structure, neutral student answers will be correlated either negatively or positively with learning.

Unlike student state unigram parameters which produce only one significant correlation, transition–student state bigram parameters produce a large number of trend and significant correlations (14). This result suggests that exploiting the discourse structure as a contextual information source can be beneficial for performance modeling.

4.3 Transition–transition bigrams

For our third experiment, we are looking at the transition–transition bigram correlations (Table 5). These bigrams help us find trajectories of length two in the discourse structure that are associated with better student learning. Because our student state is domain dependent, translating the transition–student state bigrams to a new domain will require finding a new set of relevant factors to replace the student state. In contrast, because our transition information is domain independent, transition–transition bigrams can be easily implemented in a new domain.

The Advance–Advance bigram covers situations where the student is covering tutoring material without major knowledge gaps. This is because an Advance transition happens when the student either answers correctly or his incorrect answer can be corrected without going into a remediation subdialogue. Just like with the Advance–Correct correlation (recall Table 3), we hypothesize that these correlations links higher learning gains to students that cover a lot of material without many knowledge gap.

Parameter	Mean (SD)	R.	p
Advance–Advance	35 (9.1)	.47	.04
Push–Push	2.2 (1.7)	.50	.03
Push–Push %	1.8% (1.3%)	.52	.02
Push–Push %rel	11% (7%)	.52	.02
SameGoal–Push %rel	18% (23%)	.49	.03

Table 5: Trend and significant transition–transition bigram correlations

The Push–Push bigrams capture another interesting behavior. In these cases, the student incorrectly answers a question, entering a remediation subdialogue; she also incorrectly answers the first question in the remediation dialogue entering an even deeper remediation subdialogue. We hypothesize that these situations are indicative of big student knowledge gaps. In our corpus, we find that the more such big knowledge gaps are discovered and addressed by the system the higher the learning gain.

The SameGoal–Push bigram captures another type of behavior after system rejections that is positively correlated with learning (recall the SameGoal–Neutral bigram, Table 4). In our previous work (Rotaru and Litman, 2006), we performed an analysis of the rejected student turns and studied how rejections affect the student state. The results of our analysis suggested a new strategy for handling rejections in the tutoring domain: instead of rejecting student answers, a tutoring SDS should make use of the available information. Since the recognition hypothesis for a rejected student turn would be interpreted most likely as an incorrect answer thus activating a remediation subdialogue, the positive correlation between SameGoal–Push and learning suggests that the new strategy will not impact learning.

Similar to the second experiment, the results of our third experiment are also positive: in contrast to transition unigrams, our domain independent trajectories can produce parameters with a high predictive utility.

4.4 PARADISE modeling

Here we present our preliminary results on applying the PARADISE framework to model ITSPOKE performance. A stepwise multivariate linear regression procedure (Walker et al., 2000) is used to automatically select the parameters to be included in the model. Similar to (Forbes-Riley and Litman, 2006), in order to model the learning gain, we use posttest as the dependent variable and force the inclusion of the pretest score as the first variable in the model.

For the first experiment, we feed the model all transition unigrams. As expected due to lack of correlations, the stepwise procedure does not select any transition unigram parameter. The only variable in the model is pretest resulting in a model with a R^2 of .22.

For the second and third experiment, we first build a baseline model using only unigram parameters. The resulting model achieves an R^2 of .39 by including the only significantly correlated unigram parameter: Neutral %. Next, we build a model using all unigram parameters and all significantly correlated bigram parameters. The new model almost doubles the R^2 to 0.75. Besides the pretest, the parameters included in the resulting model are (ordered by the degree of contribution from highest to lowest): Advance–Neutral %rel, and PopUp–Incorrect %. These results strengthen our correlation conclusions: discourse structure used as context information or as trajectories information is useful for performance modeling. Also, note that the inclusion of student certainty in the final PARADISE model provides additional support to a hypothesis that has gained a lot of attention lately: detecting and responding to student emotions has the potential to improve learning (Craig et al., 2004; Forbes-Riley and Litman, 2005; Pon-Barry et al., 2006).

The performance of our best model is comparable or higher than training performances reported in previous work (Forbes-Riley and Litman, 2006; Möller, 2005b; Walker et al., 2001). Since our training data is relatively small (20 data points) and overfitting might be involved here, in the future we plan to do a more in-depth evaluation by testing if our model generalizes on a larger ITSPOKE corpus we are currently annotating.

5 Related work

Previous work has proposed a large number of interaction parameters for SDS performance modeling (Möller, 2005a; Walker et al., 2000; Walker et al., 2001). Several information sources are being tapped to devise parameters classified by (Möller, 2005a) in several categories: dialogue and communication parameters (e.g. dialogue duration, number of system/user turns), speech input parameters (e.g. word error rate, recognition/concept accuracy) and meta-communication parameters (e.g. number of help request, cancel requests, corrections).

But most of these parameters do not take into account the discourse structure information. A notable exception is the DATE dialogue act annotation from (Walker et al., 2001). The DATE annotation captures information on three dimensions: speech acts (e.g. acknowledge, confirm), conversation domain (e.g. conversation- versus task-related) and the task model (e.g. subtasks like getting the date, time, origin, and destination). All these parameters can be linked to the discourse structure but flatten the discourse structure. Moreover, the most informative of these parameters (the task model parameters) are domain dependent. Similar approximations of the discourse structure are also common for other SDS tasks like predictive models of speech recognition problems (Gabsdil and Lemon, 2004).

We extend over previous work in several areas. First, we exploit in more detail the hierarchical information in the discourse structure. We quantify this information by recording the discourse structure transitions. Second, in contrast to previous work, our usage of discourse structure is domain independent (the transitions). Third, we exploit the discourse structure as a contextual information source. To our knowledge, previous work has not employed parameters similar with our transition–student state bigram parameters. Forth, via the transition–transition bigram parameters, we exploit trajectories in the discourse structure as another domain independent source of information for performance modeling. Finally, similar to (Forbes-Riley and Litman, 2006), we are tackling a more problematic performance metric: the student learning gain. While the requirements for a successful information access SDS are easier to spell out, the same can not be said about tutoring SDS due to the current limited understanding of the human learning process.

6 Conclusion

In this paper we highlight the role of discourse structure for SDS performance modeling. We experiment with various ways of using the discourse structure: in isolation, as context information for other factors (correctness and certainty) and through trajectories in the discourse structure hierarchy. Our correlation and PARADISE results show that, while the discourse structure is not useful in isolation, using the discourse structure as context information for other factors or via trajectories produces highly predictive parameters for performance analysis. Moreover, the PARADISE framework selects in the final model only discourse-based parameters ignoring parameters that do not use the discourse structure (certainty and correctness unigrams are ignored).

Our significant correlations also suggest ways we should modify our system. For example, the PopUp–Incorrect negative correlations suggest that after a failed learning opportunity the system should not give out the correct answer but engage in a secondary remediation subdialogue specially tailored for these situations.

In the future, we plan to test the generality of our PARADISE model on other corpora and to compare models built using our interaction parameters against models based on parameters commonly used in previous work (Möller, 2005a). Testing if our results generalize to a human annotation of the discourse structure and automated models of certainty and correctness is also of importance. We also want to see if our results hold for performance metrics based on user satisfaction questionnaires; in the new ITSPOKE corpus we are currently annotating, each student also completed a user satisfaction survey (Forbes-Riley and Litman, 2006) similar to the one used in the DARPA Communicator multi-site evaluation (Walker et al., 2002).

Our work contributes to both the computational linguistics domain and the tutoring domain. For the computational linguistics research community, we show that discourse structure is an important information source for SDS performance modeling. Our analysis can be extended easily to other SDS. First, a similar automatic annotation of the discourse structure can be performed in SDS that rely on dialogue managers inspired by the Grosz & Sidner theory of discourse (Bohus and Rudnicky, 2003). Second, the transition–transition bigram parameters are domain independent. Finally, for the other successful usage of discourse structure (transition–student state bigrams) researchers have only to identify relevant factors and then combine them with the discourse structure information. In our case, we show that instead of looking at the user state in isolation (Forbes-Riley and Litman, 2006), combining it with the discourse structure transition can generate informative interaction parameters.

For the tutoring research community, we show that discourse structure, an important concept in computational linguistics theory, can provide useful insights regarding the learning process. The correlations we observe in our corpus have intuitive interpretations (successful/failed learning opportunities, discovery of deep student knowledge gaps, providing relevant tutoring).

Acknowledgements

This work is supported by NSF Grant No. 0328431. We would like to thank Kate Forbes-Riley, Joel Tetreault and our anonymous reviewers for their helpful comments.

References

D. Bohus and A. Rudnicky. 2003. *RavenClaw: Dialog Management Using Hierarchical Task Decomposition and an Expectation Agenda.* In Proc. of Eurospeech.

J. Cassell, Y. I. Nakano, T. W. Bickmore, C. L. Sidner and C. Rich. 2001. *Non-Verbal Cues for Discourse Structure.* In Proc. of ACL.

M. T. H. Chi, S. A. Siler, H. Jeong, T. Yamauchi and R. G. Hausmann. 2001. Learning from human tutoring. *Cognitive Science, 25.*

S. D. Craig, A. C. Graesser, J. Sullins and B. Gholson. 2004. Affect and learning: an exploratory look into the role affect in learning with AutoTutor. *Journal of Educational Media, 29.*

K. Forbes-Riley and D. Litman. 2005. *Using Bigrams to Identify Relationships Between Student Certainness States and Tutor Responses in a Spoken Dialogue Corpus.* In Proc. of SIGdial.

K. Forbes-Riley and D. Litman. 2006. *Modelling User Satisfaction and Student Learning in a Spoken Dialogue Tutoring System with Generic, Tutoring, and User Affect Parameters.* In Proc. of HLT/NAACL.

M. Gabsdil and O. Lemon. 2004. *Combining Acoustic and Pragmatic Features to Predict Recognition Performance in Spoken Dialogue Systems.* In Proc. of ACL.

B. Grosz and C. L. Sidner. 1986. Attentions, intentions and the structure of discourse. *Computational Lingustics, 12*(3).

D. Higgins, J. Burstein, D. Marcu and C. Gentile. 2004. *Evaluating Multiple Aspects of Coherence in Student Essays.* In Proc. of HLT-NAACL.

J. Hirschberg and C. Nakatani. 1996. *A prosodic analysis of discourse segments in direction-giving monologues.* In Proc. of ACL.

E. Hovy. 1993. Automated discourse generation using discourse structure relations. *Articial Intelligence, 63*(Special Issue on NLP).

G.-A. Levow. 2004. *Prosodic Cues to Discourse Segment Boundaries in Human-Computer Dialogue.* In Proc. of SIGdial.

D. Litman and S. Silliman. 2004. *ITSPOKE: An intelligent tutoring spoken dialogue system.* In Proc. of HLT/NAACL.

S. Möller. 2005a. *Parameters for Quantifying the Interaction with Spoken Dialogue Telephone Services.* In Proc. of SIGDial.

S. Möller. 2005b. *Towards Generic Quality Prediction Models for Spoken Dialogue Systems - A Case Study.* In Proc. of Interspeech.

H. Pon-Barry, B. Clark, E. O. Bratt, K. Schultz and S. Peters. 2004. *Evaluating the effectiveness of Scot:a spoken conversational tutor.* In Proc. of ITS Workshop on Dialogue-based Intelligent Tutoring Systems.

H. Pon-Barry, K. Schultz, E. O. Bratt, B. Clark and S. Peters. 2006. Responding to Student Uncertainty in Spoken Tutorial Dialogue Systems. *International Journal of Artificial Intelligence in Education, 16*.

M. Rotaru and D. Litman. 2006. *Dependencies between Student State and Speech Recognition Problems in Spoken Tutoring Dialogues.* In Proc. of ACL.

K. VanLehn, P. W. Jordan, C. P. Rosé, D. Bhembe, M. Böttner, A. Gaydos, M. Makatchev, U. Pappuswamy, M. Ringenberg, A. Roque, S. Siler and R. Srivastava. 2002. *The Architecture of Why2-Atlas: A Coach for Qualitative Physics Essay Writing.* In Proc. of Intelligent Tutoring Systems (ITS).

K. VanLehn, S. Siler, C. Murray, T. Yamauchi and W. B. Baggett. 2003. Why do only some events cause learning during human tutoring? *Cognition and Instruction, 21*(3).

M. Walker, D. Litman, C. Kamm and A. Abella. 2000. Towards Developing General Models of Usability with PARADISE. *Natural Language Engineering*.

M. Walker, R. Passonneau and J. Boland. 2001. *Quantitative and Qualitative Evaluation of Darpa Communicator Spoken Dialogue Systems.* In Proc. of ACL.

M. Walker, A. Rudnicky, R. Prasad, J. Aberdeen, E. Bratt, J. Garofolo, H. Hastie, A. Le, B. Pellom, A. Potamianos, R. Passonneau, S. Roukos, G. Sanders, S. Seneff and D. Stallard. 2002. *DARPA Communicator: Cross-System Results for the 2001 Evaluation.* In Proc. of ICSLP.