AN OVERVIEW OF DIALOGUE AND SEMANTIC PROCESSING IN EDUCATIONAL TECHNOLOGIES

Vasile RUS, Nobal NIRAULA, Mihai LINTEAN, Arthur GRAESSER

The University of Memphis
Memphis, TN 38152
E-mail: vrus@memphis.edu

Abstract: Dialogue-based educational applications such as Intelligent Tutoring Systems interact with the student user mainly through conversational dialogue that mimics student-human tutor interactions.

There are two majors tasks that such systems need to handle: dialogue processing and assessing the correctness of student utterances. We present in this chapter an overview of dialogue and semantic processing in dialogue-based Intelligent Tutoring Systems with an emphasis on two state-of-the-art systems, AutoTutor and DeepTutor.

Key words: dialogue processing, semantic processing, and educational technologies.

1. INTRODUCTION

1.1. Dialogue and Discourse Processing in Educational Technologies

Dialogue and Discourse are areas of Natural Language Processing/Computational Linguistics that deal with manipulating larger chunks of human language, i.e. more than a sentence/utterance. Dialogue processing focuses on aspects of language that are specific to conversations between two or more partners. A key characteristic of dialogue is the taking of turns by the speakers. Each turn contains both content and dialogue management segments. The dialogue management segment is necessary to coordinate and optimize the joint action which a conversation or dialogue is. For instance, after a speaker says something new it is important for the hearer to acknowledge the understanding (or misunderstanding) of the new information before moving the dialogue forward. Furthermore, dialogue can be viewed as a set of actions, or speech acts, that together form a plan meant to achieve a certain communicative goal such as teaching Newton's 3rd law or buying an airline ticket. Based on this view of dialogue, we will focus in more depth on the issues of speech act classification and dialogue management in dialogue-based educational technologies. Furthermore, we will address the issue of utterance understanding, which is related to the content part of dialogue. In particular, we will exemplify these issues and corresponding solutions we have developed part of the AutoTutor (Graesser, Rus, et al., 2008) and DeepTutor (Rus et al., 2012) projects, two intelligent tutoring systems with natural language dialogue.

Discourse processing deals with aspects of language that are specific to long chunks of texts that form a whole such as paragraphs and documents. Two important issues in discourse processing are cohesion and coherence. Cohesion is the explicit link of sentences through mechanisms such as

pronouns and lexical chains. Coherence is about the logical connections of ideas in a discourse. A discourse could be coherent but not cohesive. However, cohesion is a desirable feature of discourse when ease of understanding is the goal, e.g. news reading. Low discourse cohesion could help in context such as learning where the goal is to make students think, i.e. poor cohesion forces students to discover themselves the logical but not explicit connections between ideas in a discourse. It should be noted that poor cohesion helps high knowledgeable students. Students will low knowledge levels benefit from high cohesion discourse.

We will focus in this chapter only on dialogue and semantic processing aspects of dialogue-based intelligent tutoring systems.

2. DIALOGUE PROCESSING IN INTELLIGENT TUTORING

2.1 Intelligent Tutoring Systems

It is easy to justify the use of tutors from the standpoint of learning gains. Students in human tutoring conditions show learning gains of 0.4-0.9 (non-expert tutors such as paraprofessionals, cross-aged, or peer tutors) to 0.8-2.3 SD (expert tutors) compared to students in traditional classroom instruction and other suitable controls (Bloom, 1984; Cohen, Kulik, & Kulik, 1982; Chi, Roy, & Hausmann, in press; Person, Lehman, & Ozbun, 2007; VanLehn et al., 2007).

Encouraged by the effectiveness of one-on-one human tutoring, computer tutors that mimic human tutors have been successfully built with the hope that a computer tutor could be afforded for every child with access to a computer.

In this chapter, we focus on some dialogue and semantic processing challenges in dialogue-based Intelligent Tutoring Systems (ITSs). Two examples of state-of-the-art dialogue ITSs are AutoTutor and DeepTutor. AutoTutor is an intelligent tutoring system that helps students learn Newtonian physics, computer literacy, and critical thinking topics through tutorial dialogue in natural language (Graesser, Rus, et al., 2008). DeepTutor is the first intelligent tutoring system based on Learning Progressions (LPs; Stevens, Delgado, & Krajcik, 2009; Rus et al., 2012). Learning progressions are a recently proposed framework by the science education research community as a way forward in science education.

As we already mentioned, computer tutors with natural language interaction try to mimic the conversation between a human tutor and a student. We present next an overview of analyses of the structure of dialogue in two human-to-human instructional interactions: classroom versus tutoring.

2.2 The Structure of Dialogue in Classroom Instruction versus Tutoring

Classrooms typically consist of the teacher presenting didactic lessons aligned with a curriculum, of presenting problems and worked out solutions, and of frequently interrogating the students with Initiate-Respond-Evaluate (IRE) sequences. The IRE sequence in a classroom consists of the teacher initiating a question, the student giving a short-answer response, and the teacher giving a positive or negative evaluation of the response (Mehan, 1979; Sinclair & Coulthart, 1975). The analogue in the tutoring session would be the exchange below on the subject matter of Newtonian physics.

(1) TUTOR: According to Newton's second law, force equals mass times what?
(2) STUDENT: acceleration
(3) TUTOR: Right, mass times acceleration.

Or

(2) STUDENT: velocity

(3) TUTOR: Wrong, it's not velocity, it is acceleration.

Obviously, there are more innovative classroom environments that deviate from this simple sketch, but this does depict most classrooms.

The discourse and pedagogical structure of a tutoring session is somewhat different from the typical classroom. Although there is a tendency for poor tutors to simply lecture like a teacher, most tutors spend considerable time presenting problems or asking difficult questions that are answered collaboratively by the tutor and tutee (Chi et al., 2001; Graesser et al., 1995; Person & Graesser, 1999). According to Graesser and Person (1994), tutors frequently implement the following 5-step tutoring frame:

(1) TUTOR asks a difficult question or presents a problem.
(2) STUDENT gives an initial answer.
(3) TUTOR gives short feedback on the quality of the answer.
(4) TUTOR and STUDENT have a multi-turn dialogue to improve the answer.
(5) TUTOR assesses whether the student understands the correct answer.

It is quite apparent that this 5-step tutoring frame involves collaborative discussion, joint action, and pressure for the tutee to construct knowledge rather than merely receiving knowledge. The role of the tutor shifts from being a knowledge-teller to a guide for collaborative knowledge construction. The relevant metaphor shifts from a classroom to collaborative work. The major dialogue structure in both AutoTutor and DeepTutor follows this 5-step dialogue structure presented above and elaborated further next. The structure differs from one system to another depending on what other instructional strategies are implemented. For instance, in DeepTutor we use a pedagogical strategy of first identifying the major concepts and principles, then applying them to the current problems, and finally putting everything together in a coherent solution to the target problem. AutoTutor implemented a pedagogical strategy based on the zone of proximal development where the next step in solving a problem is chosen best on its closeness to students' proven knowledge.

2.3 The Structure of Dialogue in AutoTutor and DeepTutor

The structure of the dialogue in both AutoTutor and DeepTutor as well as in human tutoring (Chi et al., 2001, 2004; Graesser, & Hu, & McNamara, 2005; Shah et al., 2002) follows an Expectation and Misconception Tailored (EMT) dialogue. EMT dialogue is the primary pedagogical method of scaffolding good student answers. Both the two ITSs and human tutors typically have a list of expectations (anticipated good answers) and a list of anticipated misconceptions associated with each main question. For example, expectations E1 and E2 and misconceptions M1 and M2 are relevant to the example physics problem below.

PHYSICS QUESTION: If a lightweight car and a massive truck have a head-on collision, upon which vehicle is the impact force greater? Which vehicle undergoes the greater change in its motion, and why?

E1. The magnitudes of the forces exerted by A and B on each other are equal.
E2. If A exerts a force on B, then B exerts a force on A in the opposite direction.
M1: A lighter/smaller object exerts no force on a heavier/larger object.
M2: Heavier objects accelerate faster for the same force than lighter objects.

AutoTutor and DeepTutor guide the student in articulating the expectations through a number of dialogue moves: pumps (what else?), hints, and prompts for the student to fill in missing words. Hints and prompts are carefully selected by the system to produce content in the answers that fill in missing content words, phrases, and propositions. For example, a hint to get the student to articulate expectation E1 might be "What about the forces exerted by the vehicles on each other?"; this hint would ideally elicit the answer "The magnitudes of the forces are equal." A prompt to get the student to say "equal" would be "What are the magnitudes of the forces of the two vehicles on each other?" As the learner expresses information over many turns, the list of expectations is eventually covered and the main question is scored as answered.

Dialogue moves and the problems AutoTutor and DeepTutor can tutor on are stored in a *curriculum script*. The curriculum script is a knowledge structure employed by novice tutors that largely determines the content and flow of a tutoring session (Graesser et al., 1995; McArthur et al., 1990; Putnam, 1987). The continuum of information provided by the tutor in different types of moves is worth noting. Moves at the beginning of the tutorial interaction (i.e., pumps and hints) provide less information to the student than moves towards the end of a lengthy interaction (i.e., prompts and assertions). AutoTutor and DeepTutor promote active construction of knowledge by giving more information only when the learner is floundering (Graesser, Person, & Magliano, 1995; Chi et al., 2001). Results from AutoTutor experiments show pumps and hints are correlated with greater student prior knowledge than prompts and assertions (Jackson, Person, & Graesser, 2004). Therefore, AutoTutor is sensitive and adaptive to the knowledge states of the learner. DeepTutor is still in development, we do not have concrete results with respect to its impact on student learning.

Human tutors, AutoTutor, and DeepTutor are dynamically adaptive to the learner in ways other than coaching them to articulate expectations. There is the conversational goal of correcting misconceptions that arise in the student's responses. When the student articulates a misconception, the tutor acknowledges the error and corrects it. There is another conversational goal of giving feedback to the student on their contributions. For example, the tutor gives short feedback on the quality of student contributions. The tutor accommodates a mixed-initiative dialogue by attempting to answer the student's questions when the student is sufficiently inquisitive to ask questions. The tutor asks counter-clarification questions (e.g., I don't understand your question, so could you ask it in another way?) when the tutor does not understand the student's question. Tutors are considered more adaptive to the student to the extent that they correct student misconceptions, give correct feedback, answer student questions, and ask clarification questions to insure the grounding of content. AutoTutor, DeepTutor, and other dialogue-based intelligent tutoring systems implement these features of conversational responsiveness.

AutoTutor relies on curriculum scripts which primarily list the expectations and misconceptions associated with a problem without modelling the natural progression of students' understanding of the underlying concepts associated with that problem. DeepTutor relies on Learning Progressions that are built around the idea of model-based reasoning. LPs represent ecologically valid models of understanding a topic in the form of sequences or hierarchies of increasingly sophisticated models. LPs will bring tutoring systems closer to the ideal of full adaptivity by assessing the student's level in the LP for a topic and thus guiding the feedback of tutoring based on that information.

2.4 Dialogue Processing in DeepTutor

A main component in the intelligent tutoring system DeepTutor is the dialogue management component that implements the framework and structure of tutorial dialogue that we presented earlier. An important function of the dialogue management component is deciding what DeepTutor should say next, the ideal sequence of tutor moves that will accomplish the tutorial dialogue goals and optimize student learning.

The dialogue manager in DeepTutor is implemented as a production rule system. Production rules are appropriate for mixed-initiative dialogue systems in which both the system and the user can take control of the conversation at any moment (Jurafsky & Martin, 2009) and have been successfully used in AutoTutor (Graesser, Rus, et al., 2008). There are several major changes that we made to a standard production rule dialogue manager. First, we added dialogue strategies for establishing common ground (Clark, 1996). While we implemented conversational goals used in other dialogue-based tutoring systems, e.g. AutoTutor, that include coaching students to articulate expectations, correcting students' misconceptions, and attempting to answer the student's questions when the student is sufficiently inquisitive to ask questions, we have a set of new conversational goals in DeepTutor. The additional conversational goals are perfect grounding at each turn, providing accurate feedback on students' contributions, error-handling (for handling cases when the system cannot accurately interpret what the student is saying), naturalness of dialogue, and optimizing knowledge transfer. Components that explicitly handle dialogue moves associated with these goals were designed and added to the core dialogue management module. For instance, we built a module to detect the need for establishing common ground and another to initiate and handle the dialogue moves to establish common ground. Consider the scenario in which the tutoring system presents a rare or unseen word X to a student and the student replies with "*What is X?*" That is an indication of request for grounding. By the same token, we can imagine a student replying *Yes* to a comprehension-gauging question on behalf of the tutoring system. The system would then skeptically challenge the student with a statement to double-check the student's understanding. If the student stumbles then that is an indication of knowledge transfer failure and thus the system must activate a component to optimize the transfer of knowledge. One important note here is that the form of the verification statement can make a huge difference on the successful implementation of this knowledge transfer verification strategy. Isaacs and Clark (1987) show that for an expert and novice to understand each other they must adapt their references such that the experts provide and novices acquire specialized knowledge. In other words, at the beginning of the tutor-tutee interaction there must be a phase in which the two conversational partners need to agree on the terminology used for referring to various entities and relations. As Isaacs and Clark (1987) have shown, novices (i.e., tutees in our case) can quickly adopt the experts' terminology.

The DeepTutor dialogue management module also includes advanced strategies for error-handling and naturalness of dialogue. Rapid re-prompting is the preferred strategy for first-level error prompting (Cohen et al., 2004). In rapid re-prompting, the system first rejects a user utterance by saying *I'm sorry*. In case of a second rejection, the system then applies progressive prompting, which implies the system is giving more hints about the form of an easy-to-understand utterance. Natural conversations can be achieved by simulating what human speakers do when engaged in dialogues. For instance, Herbert and Meredyth (2004) showed that speakers monitor their speech and make repairs when needed as in "The truck ... I mean, the lighter truck." We simulate in DeepTutor making speech errors and then correcting them as real human speaker, and therefore tutors, do. The question rises whether students have different expectations from a computer tutor than a human tutor. We have plans to conduct an experiment that tests for such differences in expectations.

3. SPEECH ACT CLASSIFICATION

As part of dialogue management, an intelligent tutoring system must detect students' intentions in dialogue: is the student providing more information contributing towards a full answer to the task at hand?, is the student asking a question?, is the student greeting?. Identifying students' intentions is very important in planning the next dialogue move by the system. For instance, after a

greeting such *Hi!* a polite system would respond with a greeting first before guiding the student towards the learning goal.

In DeepTutor as well as AutoTutor, speakers' intentions are modelled using elements from the speech act theory (Austin, 1962; Searle, 1969). Speech act theory has been developed based on the language as action assumption which states that when people say something they do something. Speech act is a term in linguistics and the philosophy of language referring to the way natural language performs actions in human-to-human language interactions, such as dialogues. Its contemporary use goes back to John L. Austin's theory of locutionary, illocutionary and perlocutionary acts (Austin 1962). According to Searle (Searle 1969), there are three levels of action carried by language in parallel: first, there is the locutionary act which consists of the actual utterance and its exterior meaning; then, there is the illocutionary act, which is the real intended meaning of the utterance, its semantic force; finally, there is the perlocutionary act which is the actual effect of the utterance, such as scaring, persuading, encouraging, etc. It is interesting to notice that the locutionary act is a feature of any kind of language, not only natural ones, and that it does not depend on the existence of any actor. In contrast, an illocutionary act needs the existence of an environment outside language and an actor that possesses intentions, in other words an entity that uses language for acting in the outside environment. Finally, a perlocutionary act needs the belief of the first agent in the existence of a second entity and the possibility of a successful communication attempt: the effect of language on the second entity, whether the intended one or not, is taking place in the environment outside language, for which language exists as a communication medium. As opposed to the locutionary act, the illocutionary and perlocutionary acts do not exist in purely descriptive languages (like chemical formulas), nor in languages built mainly for functional purposes (like programming languages). They are an indispensable feature of natural language but they are also present in languages built for communication purposes, like the languages of signs or the conventions of warning signals. In a few words, the locutionary act is the act of saying something, the illocutionary act is an act performed in saying something, and the perlocutionary act is an act performed by saying something. For example, the phrase "Don't go into the water" might be interpreted at the three act levels in the following way: the locutionary level is the utterance itself, the morphologically and syntactically correct usage of a sequence of words; the illocutionary level is the act of warning about the possible dangers of going into the water; finally, the perlocutionary level is the actual persuasion, if any, performed on the hearers of the message, to not go into the water. In a similar way, the utterance "By the way, I have a peanut butter sandwich with me; would you like to have a bite?" can be decomposed into the three act levels. The locutionary act is the actual expressing of the utterance, the illocutionary act is the offer implied by the phrase, while the perlocutionary act, namely the intended effect on the interlocutor, might be impressing with own selflessness, creating a gesture of friendliness, or encouraging an activity, in this case eating.

The notion of speech act is closely linked to the illocutionary level of language. The idea of an illocutionary act can be best captured by emphasizing that "by saying something, we do something" (Austin 1962). Usual illocutionary acts are: greeting ("Hello, John!"), describing ("It's snowing."), asking questions ("Is it snowing?"), making requests ("Could you pass the salt?"), giving an order ("Drop your weapon!"), making a warning ("The floor is wet!"), or making a promise ("I'll return it on time."). The illocutionary force is not always obvious and could also be composed of different components. As an example, the phrase "It's cold in this room!" might be interpreted as having the intention of simply describing the room, or criticizing someone for not keeping the room warm, or requesting someone to close the window, or a combination of the above. A speech act could be described as the sum of the illocutionary forces carried by an utterance. It is worth mentioning that within one utterance, speech acts can be hierarchical, hence the existence of

a division between direct and indirect speech acts, the latter being those by which one says more than what is literally said, in other words, the deeper level of intentional meaning.

In the phrase ”Would you mind passing me the salt?”, the direct speech act is the request best described by ”Are you willing to do that for me?” while the indirect speech act is the request ”I need you to give me the salt.” In a similar way, in the phrase ”Bill and Wendy lost a lot of weight with a diet and daily exercise.” the direct speech act is the actual statement of what happened ”They did this by doing that.”, while the indirect speech act could be the encouraging ”If you do the same, you could lose a lot of weight too.”

A major approach we adopted to the task of automated speech act classification is based on the idea that the leading tokens in an utterance are indicative of the speaker’s intention, i.e. speech act. For instance, a question most likely starts with a wh-word, such as How, followed by an auxiliary verb. In contrast, a statement starts with noun or pronoun followed by a verb.

We assumed there is one speech act per utterance and the set of speech acts used are all at the same level of depthness forming a flat hierarchy. These simplification assumptions are appropriate for a first attempt at automating the speech act classification process and testing our leading tokens model. In one set of experiments, we used the following set of speech acts: Statement, Request, Reaction, MetaStatement, Greeting, Expressive Evaluation, Question, and Other. On a data set that came from a study run using an epistemic game, *Urban Science*, in which players take on the role of an intern for an Urban Planning company and are provided guidance from a mentor on the proper steps to be taken in redesigning a city, we obtained a best accuracy of 68.30% (kappa=60.51%). Accuracy measures how well the predicted speech act categories match the correct categories, which were annotated by human experts. Kappa statistics measure the performance of a method while accounting for chance (a value of kappa>60% is very good, meaning the method does much better than chance). The data set we used was a random sample of 750 mentor contributions and 750 player contributions from a chat among players and mentors that included 1,956 mentor contributions and 2,175 player contributions. The 750 mentor and 750 student contributions were further split into speech acts, using end of sentence punctuation marks (i.e., periods, question marks, and exclamation marks) as delimiters. This process resulted in 901 mentor speech acts and 765 player speech acts. Each of these speech acts were hand annotated by one trained annotator. Prior to this annotation, two assistants, trained on the speech act categories, independently annotated 1,500 speech acts. The average inter-rater reliability (across all categories) was Kappa = 0.87. Each of the 1,666 mentor and player speech acts was annotated using only one of the categories from the speech act taxonomy.

We experimented with n=2..8 leading tokens to make predictions about the speech act categories of the utterances (the average contribution has 7.26 tokens) with Naïve Bayes and Decision Trees as the pattern learning algorithms. The results confirmed our hypothesis that using just the 3 leading tokens in each utterance is as good as or better than using the leading 8 tokens. Using just one or two leading tokens was no better than 3 leading tokens. More details about this work can be found in Moldovan, Rus, & Graesser (2011).

Our basic model has its own limitations, which explains the very good but not perfect performance results. We plan to extend the basic model we proposed here with more contextual clues, which we believe will lead to further improvements in performance. Contextual clues will exploit discourse sequential patterns that humans most likely take advantage of, such as the fact that after a greeting another greeting follows as a response to the first one.

4. SEMANTIC PROCESSING IN AUTOTUTOR AND DEEPTUTOR

Besides the dialogue management issues discussed above, an ITS must address the content part of dialogue. Indeed, one fundamental problem in dialogue-based ITSs is to evaluate students' answers to a given problem. There are two types of such assessments: global versus local. In the global assessment, the overall student answer to the target problem is evaluated. In local evaluation, a particular student answer, for instance in response to tutor questions such as What does Newton's second law says?, is being evaluated. We address next the local evaluation of such responses that occurs in the middle of the dialogue between the student and the system.

During the dialogue that takes place between the system and the student, the system checks whether the student can articulate each of a set of ideal steps, called expectations, of the ideal answer. Furthermore, the system also compares the student response to possible misconception stored in the curriculum script. The following expectation and student answer is reproduced from a previous experiment with AutoTutor (Graesser et al. 2004).

Expectation: *The person and the object cover the same horizontal distance.*
Student Answer: *The two objects will cover the same distance.*

The challenge is to decide whether the expectation (E) is semantically similar to the student answer (A). This is a typical text-to-text similarity problem. A number of text-to-text semantic similarity tasks have been identified in natural language processing, including paraphrase identification (Dolan et al., 2004), recognizing textual entailment (Dagan et al., 2005), and elaboration detection (McCarthy & McNamara, 2008). These fundamental tasks are in turn important to a myriad of real world applications such as providing evidence for the correctness of answers in Question Answering (Ibrahim et al., 2003), increase diversity of generated text in Natural Language Generation (Iordanskaja et al., 1991), and detecting students' mental models in educational systems (Lintean, Rus, & Azevedo, 2012).

Given its importance, the semantic similarity problem has been addressed by many research groups. The majority of the explored solutions measure the degree of similarity at word and structural level between the two texts that are being evaluated and then use that information in conjunction with a simple learning approach of finding an optimal threshold that separates paraphrases from non-paraphrases, or some more complicated machine learning methods (i.e. meta-classifiers, or voting schemes between several classifiers). Similarity solutions range from simple word overlap to weighted word overlap to greedy methods that incorporate also syntactic or phrase overlap (Corley & Mihalcea, 2005; Lintean & Rus, 2011; Kozareva & Montoyo, 2006; Madnani, Tetreault & Chodorow, 2012; Socker et al., 2011; Qiu et al., 2006).

We proposed several solutions over the years to this problem. We highlight next a greedy method (Rus & Graesser, 2006; Lintean & Rus, 2011) and a more recently proposed solution (Rus & Lintean, 2011) that solves the problem of semantic similarity using an optimization algorithm based on the job assignment problem (Kuhn, 1955; Munkres, 1957).

4.1 The Greedy Approach

In the greedy approach words from one sentence (usually the shorter sentence) are greedily matched, one by one, starting from the beginning of the sentence, with the most similar word from the other sentence. In case of duplicates, the order of the words in the two sentences was important such that the first occurrence matches with the first occurrence and so on. To be consistent across all methods presented here and for fairness of comparison across these methods, we require that words must be part of at most one pair. It should be noted that others, e.g. Corley and Mihalcea

(2005), did not impose such a requirement and therefore some words could be selected to be part of more than one pair.

The greedy method has the advantage, over the other methods, of being simple and fast, while also effectively using the natural order of words within the sentence, which partially encodes the syntactic information between them. The obvious drawback of the greedy method is that it does not aim for a global maximum similarity score. The optimal methods described next solve this issue.

4.2 Optimal Word-to-Word Matching

The optimal method aims at finding the best overall word-to-word match, based only on the similarities between words. This is a well-known combinatorial optimization problem. The assignment problem is one of the fundamental combinatorial optimization problems and consists of finding a maximum weight matching in a weighted bipartite graph. Given a complete bipartite graph, *G = (S, T, E)*, with n worker vertices (*S*), n job vertices (*T*), and each edge $e_{s\in S,t\in T} \in E$ having a non-negative weight $w(s,t)$ indicating how qualified a worker is for a certain job, the task is to find a matching *M* from *S* to *T* with maximum weight. In case of different numbers of workers or jobs, dummy vertices could be used.

The assignment problem can be formulated as finding a permutation π for which $S_{OPT} = \sum_{i=1}^{n} w(s_i, t_{\pi(i)})$ is maximum. Such an assignment is called optimum assignment. An algorithm, the Kuhn-Munkres method (Kuhn, 1955), has been proposed that can find a solution to in polynomial time (see Dawes, 2011 for a complete formal description of the algorithm).

In our case, we modeled the semantic similarity problem as finding the optimum assignment between words in one text, T_1, and words in another text, T_2, where the fitness between words belonging in opposite texts can be measured by any word-to-word semantic similarity function. That is, we are after a permutation π for which $S_{OPT} = \sum_{i=1}^{n} word\text{-}sim(v_i, w_{\pi(i)})$ is maximum where *word-sim* can be any word-to-word similarity measure, and *v* and *w* are words from the texts T_1 and T_2, respectively.

4.3 Summary of Results

We have evaluated the above methods on various corpora among which we mention the Microsoft Research Paraphrase (MSRP) Corpus (Dolan, Quirk, and Brockett 2004). MRSP is a standard dataset for evaluating approaches to paraphrase identification. The corpus contains a total of 4076 training instances (of which 2753 are TRUE paraphrases) and 1725 testing instances (of which 1147 are TRUE paraphrases). Although the corpus has its issues and limitations, confirmed by other researchers as well (Madnani, Tetreault & Chodorow, 2012), it has been so far the largest publicly available annotated paraphrase corpus and has been used in most of the recent studies that addressed the problem of paraphrase identification and semantic similarity assessment. The results varied based on the exact word-to-word similarity measures used yielding best results for JCN (Jiang & Conrath, 1997) for an accuracy of 73.20% for the greedy method and 74.20% for the optimal matching method.

5. CONCLUDING REMARKS

Intelligent Tutoring Systems with natural language interaction are complex systems whose main form of interaction with the student is through dialogue. We presented in this chapter an overview of some of the most challenging issues related to dialogue and semantic processing in

such systems. While the current solutions to these issues are good enough to allow ITS researchers to focus on core learning issues such as student assessment, feedback, and tutorial and pedagogical strategies, there is much more to be done in terms of perfecting the dialogue and semantic processing aspects of dialogue-based ITSs. For instance, because the semantic assessment component is not perfect a most frustrating experience for students happens when they provide a correct answer but the system labels it as incorrect and therefore responds with negative feedback. To avoid such experiences with possible negative impact learning and motivation, there is need to further the performance of the natural language processing aspects of dialogue-based ITSs.

6. REFERENCES

1. AUSTIN, J.L. 1962. How to do Things with Words. Oxford University Press, 1962.
2. BLOOM, B. S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. Educational Researcher, 13, 4-16.
3. CHI, M. T. H., SILER, S., JEONG, H., & HAUSMANN, R. (2001). Learning from human tutoring. Cognitive Science, 25, 471-534.
4. CHI, M. T. H., SILER, S. A., & JEONG, H. (2004). Can tutors monitor students' understanding accurately? Cognition and Instruction, 22(3): 363-387.
5. CHI, M. T. H., ROY, M., & HAUSMANN, R.G.M. (2008). Observing tutorial dialogues collaboratively: Insights about human tutoring effectiveness from vicarious learning. Cognitive Science, 32, 301-341.
6. CLARK, H. H. (1996). Using language. Cambridge: Cambridge University Press.
7. COHEN, P. A., KULIK, J. A., & KULIK, C. C. (1982). Educational outcomes of tutoring: A meta-analysis of findings. American Educational Research Journal, 19, 237-248.
8. COHEN, M. H., GIANGOLA, J. P., & BALOGH, J. (2004). Voice User Interface Design, Addison-Wesley Professional, ISBN: 0321185765.
9. CORLEY, C. and MIHALCEA, R. (2005). Measures of Text Semantic Similarity, in Proceedings of the ACL workshop on Empirical Modeling of Semantic Equivalence, Ann Arbor, MI, June 2005.
10. DAGAN, I., GLICKMAN, O., and MAGNINI, B. (2005). The PASCAL recognizing textual entailment challenge. In Proceedings of the PASCAL Workshop.
11. DOLAN, W.B., QUIRK, C., and BROCKETT, C. (2004). Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In Proceedings of the 20th International Conference on Computational Linguistics, Geneva, Switzerland.
12. GRAESSER, A. C., PERSON, N. K., & MAGLIANO, J. P. (1995). Collaborative dialogue patterns in naturalistic one-to-one tutoring. Applied Cognitive Psychology, 9, 495–522.
13. GRAESSER, A. C., LU, S., JACKSON, G. T., MITCHELL, H., VENTURA, M., OLNEY, A., & LOUWERSE, M.M. (2004). AutoTutor: A tutor with dialogue in natural language. Behavioral Research Methods, Instruments, and Computers, 36, 180-193.
14. GRAESSER, A. C., HU, X., & MCNAMARA, D.S. (2005). Computerized learning environments that incorporate research in discourse psychology, cognitive science, and computational linguistics. In A.F. Healy (Ed.), Experimental Cognitive Psychology and its Applications: Festschrift in Honor of Lyle Bourne, Walter Kintsch, and Thomas Landauer. Washington, DC: American Psychological Association.
15. GRAESSER, A.C., RUS, V., D'MELLO, S. K., and JACKSON, G. T. (2008). AutoTutor: Learning through natural language dialogue that adapts to the cognitive and affective states of the learner. In D. H. Robinson & G. Schraw (Eds.), Current perspectives on cognition, learning and instruction: Recent innovations in educational technology that facilitate student learning (pp. 95–125). Information Age Publishing.
16. HERBERT, H. C., & MEREDYTH, A. K. (2004). Speaking while monitoring addressees for understanding. Journal of Memory and Language, 50(1), 62-81.
17. IBRAHIM, A., KATZ, B., and LIN, J. (2003). Extracting structural paraphrases from aligned monolingual corpora. In Proceeding of the Second International Workshop on Paraphrasing, (ACL 2003).

18. IORDANSKAJA, L., KITTREDGE, R., and POLGERE, A. (1991). Natural Language Generation in Artificial Intelligence and Computational Linguistics. Lexical selection and paraphrase in a meaning-text generation model, Kluwer Academic.
19. ISAACS, E. A., & CLARK, H. H. (1987). References in conversations between experts and novices. Journal of Experimental Psychology: General, 116, 26-37.
20. LINTEAN, M., RUS, V., & AZEVEDO, R. (2012). Automatic Detection of Student Mental Models during Prior Knowledge Activation in MetaTutor, International Journal of Artificial Intelligence in Education.
21. LINTEAN, M., & RUS, V. (2011). Dissimilarity Kernels for Paraphrase Identification, Proceedings of the 24th International Florida Artificial Intelligence Research Society Conference, Palm Beach, FL, May, 2011.
22. JACKSON, G.T., PERSON, N.K., and GRAESSER, A.C. (2004) Adaptive Tutorial Dialogue in AutoTutor. Proceedings of the workshop on Dialog-based Intelligent Tutoring Systems at the 7th International conference on Intelligent Tutoring Systems (pp. 368-372). Universidade Federal de Alagoas, Brazil, 9-13.
23. JIANG, J.J. & CONRATH, D.W. (1997). Semantic Similarity based on corpus statistics and lexical taxonomy. In Proceedings of the International Conference on Research in Computational Linguistics.
24. JURAFSKY, D., & MARTIN, J. (2009). Speech and Language Processing, Prentice-Hall, 2009, Second Edition.
25. KOZAREVA, Z., and MONTOYO, A. (2006). Lecture Notes in Artificial Intelligence: Proceedings of the 5th International Conference on Natural Language Processing (Fin-TAL 2006). chapter Paraphrase Identification on the basis of Supervised Machine Learning Techniques.
26. KUHN, H. W. (1955). The Hungarian Method for the assignment problem. Naval Research Logistics Quarterly, volume 2, pages 83-97.
27. MADNANI, N., TETREAULT, J. and CHODOROW, M. (2012). Re-examining Machine Translation Metrics for Paraphrase Identification. Proceedings of the NAACL-HTL Conference. Montreal, Canada. Association for Computational Linguistics.
28. MCCARTHY, P.M. and MCNAMARA, D.S. (2008). User-Language Paraphrase Corpus Challenge, online, 2008.
29. MCARTHUR, D., STASZ, C., & ZMUIDZINAS, M. (1990). Tutoring techniques in algebra. Cognition and Instruction, 7, 197-244.
30. MEHAN, H. (1979). Learning lessons. Cambridge, MA: Harvard.
31. 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.
32. MUNKRES, J. (1957). Algorithms for the assignment and transportation problems. Journal of the Society for Industrial and Applied Mathematics, volume 5(1), pages 32-38. Society for Industrial and Applied Mathematics.
33. PERSON, N., LEHMAN, B., & OZBUN, R. (2007). Pedagogical and motivational dialogue moves used by expert tutors. Presented at the 17th Annual Meeting of the Society for Text and Discourse. Glasgow, Scotland.
34. PERSON, N. K., & GRAESSER, A. C. (1999). Evolution of discourse in cross-age tutoring. In A. M. O'Donnell & A. King (Eds.), Cognitive perspectives on peer learning (pp. 69-86). Mahwah, NJ: Erlbaum.
35. PUTNAM, R. T. (1987). Structuring and adjusting content for students: A study of live and simulated tutoring of addition. American Educational Research Journal, 24(1), 13–48.
36. QIU, L., KAN, M. Y., CHUA, T. S. (2006). Paraphrase Recognition via Dissimilarity Significance Classification. Proceedings of the EMNLP, pages 18-26.
37. RUS, V. & GRAESSER, A.C. (2006). Deeper Natural Language Processing for Evaluating Student Answers in Intelligent Tutoring Systems, Proceedings of the Twenty-First National Conference on Artificial Intelligence (AAAI-06).

38. RUS, V., GRAESSER, A.C., CONLEY, M., GIRE, E., FRANCESCHETTI, D., LINTEAN, M., NIRAULA, N., BAGGETT, W., & BARGAGGLIOTTI, A. (2012). DeepTutor: Promoting Deep Learning, 5th International Conference on Educational Data Mining, Chania, Crete, Grece, 19-21 June, 2012.
39. SEARLE, J.R. 1969. Speech Acts. Cambridge University Press, GB, 1969.
40. SHAH, F., EVENS, M.W., MICHAEL, J., & ROVICK, A. (2002). Classifying student initiatives and tutor responses in human keyboard-to keyboard tutoring sessions. Discourse Processes, 33, 23-52.
41. SINCLAIR, J. M., & COULTHARD, R. M. (1975). Towards an analysis of discourse: The English used by teachers and pupils. London: Oxford University Press.
42. SOCKER, R., HUANG, E.H., PENNINGTON, J., NG, A.Y., and MANNING, C.D. (2011). Dynamic Pooling and Unfolding Recursive Autoencoders for Paraphrase Detection. Advances in Neural Information Processing Systems, volume 24. NIPS.
43. STEVENS, S. E., DELGADO, C., & KRAJCIK, J. S. (2009). Developing a hypothetical multi-dimensional learning progression for the nature of matter. Journal of Research in Science Teaching.
44. VANLEHN, K., GRAESSER, A. C., JACKSON, G. T., JORDAN, P., OLNEY, A., & ROSE, C. P. (2007). When are tutorial dialogues more effective than reading? Cognitive Science, 31, 3-62.