Classification of Speech Acts in Tutorial Dialog

Johanna Marineau[1], Peter Wiemer-Hastings[2], Derek Harter[1], Brent Olde[1], Patrick Chipman[1],
Ashish Karnavat[1], Victoria Pomeroy[1], Sonya Rajan[1], Art Graesser[1], and the Tutoring Research Group

[1]Department of Psychology, University of Memphis, Memphis, TN 38152-6400
jmarinea@memphis.edu or a-graesser@memphis.edu
[2]ICCS/Division of Informatics, 2 Buccleuch Place, Edinburgh, EH89LW, Scotland
peterwh@cogsci.ed.ac.uk

Abstract Computer models of tutorial dialog need to segment and classify the learner's contributions into sequences of speech acts. The responses of the computer tutor we have developed (called AutoTutor) need to be sensitive to the speech acts of the learner's contribution during the previous turn. We developed and tested some models that classified the speech acts in naturalistic tutorial dialog into four categories: Assertions, Yes/No Questions, WH-Questions, and Frozen Expressions. The three models consist of (1) a Brill part of speech tagger, a syntactic parser, and a symbolic post-processor, (2) a feed-forward neural network, and (3) a model based on surface linguistic features. The parsing model had the best performance; classifying speech acts with 79% accuracy. Based on the data from this study, further research will be directed toward the usefulness of a hybrid model that uses a parser to extract the main clause, which can then be classified using a neural network.

1 Introduction

With every turn in a conversation, humans analyze the content and the speaker's intent behind each speech act. Theoretical frameworks for categorizing these speech acts have been developed in philosophy, linguistics, and semantics for nearly 40 years [4, 25, 26]. These theoretical frameworks have been the basis for empirical classification systems that operate on specific types of corpora [8, 9, 17, 24]. Each of these systems contained from 4 to 10 categories of speech acts. Four of the frequent categories in these systems are the four categories below.

1. Assertions. These are claims about the world or statements of fact.
2. Questions. These are frequently subdivided into yes/no versus wh-questions.
3. Directives. These are requests for an action or information.
4. Responses. These are usually short reactions to the previous speaker's contribution.

Of course, there are other categories of speech acts, such as promises, declarations, and expressive evaluations. However declarations and promises are not very frequent in the context of tutoring. The goal of our research is to develop a speech act classification system that can operate as part of a robust computerized dialogue system in tutoring.

As an example, consider the speech act classifier in the AutoTutor system [11, 15]. AutoTutor's primary goal is to model human-human tutorial dialogues. In such dialogues, the tutor normally takes control of the dialogue by presenting problems to the student, getting answers and other contributions back, and providing help when needed. Sometimes, however, the student takes the initiative, as in the case of student questions. For example, the student might ask for the definition of a technical term (e.g., "What does ROM mean?") or a confirmation that information is correct ("Isn't ROM Read Only Memory?"). A student might ask the tutor to repeat something ("What?", "Could you say that again?"). For the computerized tutor to be able to appropriately respond in such mixed-initiative dialogue, the system must understand not only the information content of the speech act, but also the speaker's intentions. The intentions are to some extent made clear by the categorization of the speech act.

When the student types his questions or contributions into the AutoTutor system, a series of analyses are initiated that respond with an appropriate dialogue moves. A Dialogue Advancer Network (DAN) is designed to select an appropriate discourse marker (for example "Well", or "Okay") as needed and the next dialogue move. Dialogue moves are selected from curriculum scripts that contain the questions, hints, prompts, elaborations, and summaries written for each subtopic [22, 23]. The student's previous speech act must be correctly classified for the DAN to select a move that will be conversationally smooth and pedagogically effective. While exploring the different speech act classification systems, it was necessary to consider the special constraints of a human-computer tutorial dialogue. For example, D'Andrade and Wish [8] used a corpus that included conversations among members of a family. They had access to the cues that humans have in face-to-face conversation, such as intonation, gestures, and facial expressions. These cues are not available in a human-computer interaction where students enter the content of their turns by keyboard. As another example, Samuel, Carberry and Vijay-Shanker [24] developed a machine learning approach to speech act classification, which classified speech acts with 76% accuracy. Unfortunately, this model was trained and tested on a written corpus that used cues from subsequent utterances in the dialog (i.e., information that occurred after the speech act being classified). This type of information would be unavailable in an ongoing tutorial dialog in which the speech acts are classified as they occur. Other statistical models of speech act classification have been developed to facilitate computer-human dialogue [18, 19]. Nagata & Morimoto's [18] model utilizes syntactic categories, which correspond to the speech act categories used in our analysis, and intention categories, which are specific to their corpus of scheduling dialogues. Reithinger & Klesen [19] developed a model that classified 18 types of dialogue acts with 67% accuracy for German text and 75% for English text, using a corpus of scheduling dialogues from VERBMOBIL [27].

Neural networks have been developed by Kipp [16] for classifying speech acts . Using Elman's simple recurrent network (SRN) [10], Kipp's network contained 18 modular networks, with one module for each speech act category. A selector network then received the weighted outputs from the modular networks and determined the correct speech act category. When evaluating a corpus of dialogue acts taken from VERBMOBIL[27], the network's performance was 60%. This modular network was developed after initial experiments using a feedforward network, which yielded performance of only 45%.

When we analyzed naturalistic human-to-human tutoring [12, 13] and tutorial dialogues with AutoTutor [15], there were a number of speech act categories that frequently occurred and that were functionally significant. The categories included (1) Assertions, (2) WH-questions, (3) YES/NO questions, and (4) Directives. Examples of these speech acts categories will be described later, when the tutoring corpus is discussed. There are also a number of other categories that periodically occurred, such as metacognitive comments ("I don't understand", "This makes sense") and short responses ("Uh-huh", "Okay"). Metacognitive comments and short responses are normally expressed as frozen expressions that distinctively signal their conversational function. The present study focused on Assertions, WH-questions, YES/NO-questions, and Directives because there was some flexibility in their composition in addition to their being relevant to tutorial dialog and their being adopted by many other researchers who investigate speech act classification.

This paper compares the performance of three different speech act classification systems that have been implemented on computer. Each of these computational models categorize a speech act on the basis of a variety of textual cues, including punctuation, number of words in the utterance, parts of speech, and the order of subconstituents. Each model was tested on the same corpus of 100 speech acts. The three architectures used by the classifiers that we have analyzed are:

(a) A feed-forward neural network model
(b) A classical Natural Language Processing (NLP) system that tags and parses the texts and then runs the results through a post-processing module
(c) A simple production rule model that uses word-based regular expressions and surface cues

This paper first describes the corpus that was used to train and test the different models. Then it describes the architecture of each model and its performance. Finally, we discuss our results and describe the future directions of the project.

2 Corpus

A corpus of 426 speech acts was created by sampling the transcripts of human-to-human tutorial dialogs collected by Graesser and Person [12]. These transcripts recorded the interactions between a human tutor and human student discussing research methods in psychology. The speech acts were randomly selected from the student turns in 32 different sessions and compiled into a database of 128 Assertions, 128 Questions, 128 Responses, and 42 Directives. We attempted to get equivalent numbers of speech acts in each category, but the frequency of Directives was low because student Directives are relatively rare in practice. For the purpose of the present analysis, Responses were removed, and Questions were subcategorized into WH-Questions and Yes/No Questions. The resulting database contained 128 Contributions, 41 WH-Questions, 87 Yes/No Questions, and 42 Directives (298 speech acts in total). The division of the Questions into two categories creates unequal numbers of questions in each question category, but we were limited by the number of available questions in the tutoring corpus. Each model of speech act classification was subsequently tested on 100 speech acts, 25 from each category, which were randomly selected from the database. The database items were marked to include a part of speech tag for each word, the word's position in the speech act, and the category of the tutor's and the student's last speech act. Examples of the items from each category are given below:

(1) Assertions.
The second independent variable would be outside distractions.
Well, you would have to define what you mean by it.

(2) WH-questions

What I mean, what exactly was a scatterplot?
How did you change that?

(3) Yes/No questions
Should I do another one?
Do you want me to give an example?

(4) Directives
Be more specific, please.
Wait a minute!

3 Models

As previously mentioned, we evaluated three different speech act classification models: a connectionist model, a classical NLP model, and a simple production rule model. This section describes the architecture of each model.

3.1 Neural Network Model

Previous analyses of speech acts in tutorial dialogues have suggested that a neural network model could be effective in classifying speech acts. In particular, the first three words of a speech act were believed to be a significant cue for classification [15], and the previous two turns in a dialogue are known to be moderately predictive of the next speech act category [14]. This prior research suggests the likely features that could be very effective input units in a neural network system. Furthermore, because some speech acts are ambiguous from the standpoint of speech act classification, a neural network might outperform a symbolic system by weighing conflicting information.

A feed-forward neural network model was developed with an architecture that had input units, one hidden layer, and output units. There were 57 total input units. The tutor's previous speech act was represented by four units, one for each of the possible categories. The student's previous speech act was represented in the same way. Sentence-ending punctuation was given by four input units, one for a period, one for a question mark, one for an exclamation point, and one for “other.” The length of the speech act was encoded by three units, one for 1-5 words, another for 6-10 words, and the third for more than 10 words. Next, for each of the first three words of the utterance, there were 17 units, which corresponded to the 17 part of speech categories in the system's tag set. The activation on each node was proportional to the frequency of the word with that part of speech tag. Fifteen hidden units were used and the learning rate was .01. The network was trained on the entire training set of 198 speech acts and then tested on the test set of 100 items.

3.2 Parsing Model

The classical NLP speech act classifier was created by combining a shareware part of speech (POS) tagger, a shareware parser, and a symbolic post-processor module. The shareware systems that we used were Brill's rule-based tagger [6] and Abney's CASS/SCOL parser [1].

Brill's tagger is the *de facto* state-of-the-art tagging system, which is used as an input filter for many different NLP systems. Its tags are taken from the Penn Treebank's Wall Street Journal corpus. The part of speech tags used by the Penn Treebank corpus were grouped to correspond to the 17 part of speech tags used our model. The tagger first gives the most frequent tag for each word in the input sentence. Then it applies a series of rules, which modify the tags based on the surrounding context. The tagger has been shown to give correct tags in over 90% of cases in several different tests on written corpora.

Abney's SCOL system uses the CASS cascading grammar parser. This parser takes a tagged text as input, and then tries to apply grammar rules to the input. The parser differs from a traditional parser in that the grammar rules are separated into levels, starting with the simplest components and moving up to complex combinations. The parser attempts to apply the rules one level at a time. Once it moves to a higher level, it will no longer attempt to apply lower-level rules. In this manner, the parser avoids many of the pitfalls caused by ambiguity in natural language; at each level, it creates the components that it can be sure of, leaving difficult choices for later processing, making the parser run quite quickly. It does not always produce a complete parse for an input sentence, but the partial parses that it produces are highly likely to be correct.

We processed 198 speech acts (the full set of 298 items minus the 100 test items) with the Brill tagger and SCOL parser. Then we analyzed the resulting syntactic structures, and manually extracted regularities to make a simple decision tree [5, 20,21]. This decision tree contained seven decisions, for example: "Is there an imperative construction in the parse tree?" Because the training set had a severe lack of examples for some speech act categories, the decision tree was extended after the first round of testing to handle additional examples. The final tree contained 10 decisions.

3.3 Simple Rule-based Classifier

As a control, the final model tested was the speech act classifier currently being used in the AutoTutor system [15]. This model also uses a decision tree, but with word-based cues. Due to the requirements of the tutoring system, this model classifies student input into a slightly different set of categories: Yes/No Questions, Definitional Questions (What is X?), Frozen Expressions (canonical phrases like, "What did you say?"), and Assertion. For the purpose of this analysis, Definitional Questions were labeled as WH-Questions and the Frozen Expressions were labeled as Directives. The Frozen Expressions used by this speech act classifier were compiled by identifying Directives, short Responses, and WH-questions like the one shown above. Analyses revealed that 36% of the Frozen Expressions were Directives. However, it should be noted once again that the speech acts in the sample of 100 were equally divided among Assertions, WH-Questions, YES-NO questions, and Directives.

4 Results

4.1 Neural network

The neural network model correctly classified 71% of the Assertions, 66% of the WH-questions, 63% of the Directives, and 58% of the Yes/No questions. Overall performance was 65%. Table 1 shows the classifications made by the network for each category. The categories on the left column of the table are the categories of human experts. The categories on the top are the ones assigned by the neural network. The cells on the diagonal show the percentage of each speech act that were correctly classified. A sensitivity analysis of the network indicates it substantially relied on the punctuation and the length of the speech act as primary cues in deciding the speech act category. This accounts for the majority of the incorrect classifications that the network made: identifying Assertions as Directives (24%), WH-questions as Yes/No questions (28%), Directives as Assertions (33%), and Yes/No questions as WH-questions (27%).

Table 1. Neural Network Classifier Results (%)

	Assertion	WH-Question	Directive	Yes/No Question
Assertion	71	5	24	0
WH-Question	6	66	0	28
Directive	33	0	63	4
Yes/No Question	15	27	0	58

4.2 Parsing Model

The model using the tagger, parser, and post-processor had the best performance of the three systems. Table 2 shows that this model correctly classified 68% of the Assertions, 85% of the WH-questions, 80% of the Directives, and 83% of the Yes/No questions, for an overall performance of 79%. The results of this model may be slightly inflated since, for this first set of experiments, modifications were made to the decision tree after the first testing. Errors made by this system were similar to the errors made by the neural network. That is, there was a blur between the two categories of questions and between Assertions and Directives.

Table 2. Parsing Model Results (%)

	Assertion	WH-Question	Directive	Yes/No Question
Assertion	68	8	24	0
WH-Question	0	85	15	0
Directive	12	0	80	8
Yes/No Question	0	13	4	83

4.3 Simple Rule-based Classifier

The overall performance of the AutoTutor speech act classifier was 53%. As shown in Table 3, this model was correctly able to identify Assertions and YES/NO questions (100% for each category), but performance was very poor for WH-questions and Directives (8% and 4%, respectively). This model's poor performance on these two categories may be related to the differences between the test set and the types of input it was designed to classify. For example, not all WH-questions are definitional questions ("What is an X?").

Table 3. Results of Simple Rule-based Classifier (%)

	Assertion	WH-Question	Directive	Yes/No Question
Assertion	100	0	0	0
WH-Question	8	8	0	84
Directive	92	0	04	4
Yes/No Question	0	0	0	100

5 Conclusions

Although each model performed above chance (25%), none of them is ideal for use in a dialogue system like AutoTutor. The tutor's next dialogue move depends on the correct classification of the student's previous speech act, so an enhanced performance is needed. Some obvious measures can be taken to improve the performance of both the neural network and the symbolic parser.

The neural network was trained on 66% of the corpus (198 speech acts) so that each model could be tested on the same set of 100 speech acts. By removing 25 speech acts from each category, the training set of WH-questions and directives was reduced to 16 and 17 speech acts, respectively. These two categories also had the lowest performance scores. Simply increasing the corpus size, to produce a larger and more evenly distributed training set, may improve network performance. This network's architecture includes input units for the part of speech of the first three words of each speech act. Unfortunately, speech acts that begin with an introductory clause may cause incorrect classifications ("After you turn on the computer, does ROM get used?"). The first three words used in making the classification are not in the main clause of the speech act, so the intention of the question is missed. These preposed subordinate clauses are very frequent in tutoring sessions because the student needs to set up the context (via the preposed clause) before the focal question is asked.

Our next research goal is to evaluate a speech act classifier that combines a parsing system and a neural network. Speech acts would be assigned part of speech tags, and then parsed to identify the main clause, as opposed to subordinate clauses. The main clause can then be evaluated by the neural network for classification.

We are currently assuming a transparent relationship between the surface speech act of an utterance and the speaker's intention. This is often, but not always, the case. To use a classical example, if I want you to open the window, depending on my politeness and the social rank between us, there are several ways in which I could communicate my intent. I could assert that it is warm in the room, ask you if you mind opening the window, ask you directly to open the window, or command you to open the window. In the tutoring domain, the relationship between the tutor and student is relatively well specified, so we can safely assume that when a student asks a question, she is expecting some kind of answer. In some cases, however, the student presents an assertion that looks like a question, for example, "Could it be RAM?" In such instances, the student uses the question format only to indicate that she is not sure of her answer. We want the system to be aware of this, but also to evaluate the response as an assertion. We will focus on this issue in future research.

References

1. Abney, S.: Partial parsing via finite-state cascades. In Proceedings of the ESSLLI '96 Robust Parsing Workshop (1996)
2. Abney, S.: Methods and statistical linguistics. In J. Klavans, P. Resnik (eds.), The Balancing_Act. Cambridge, MA: MIT Press (1996)
3. Allen, J.: Natural language understanding. Redwood City, CA: Benjamin Cummings (1995)
4. Austin, J. L.: How to do things with words. Cambridge, MA: Harvard University Press (1962)
5. Breiman, L., Friedman, J., Olshen, R., Stone, C.: Classification and regression trees. Wadsworth and Brooks (1984)
6. Brill, E.: Some advances in rule based part of speech tagging. In Proceedings, Third International Workshop on Parsing Technologies,

7. Tilburg, The Netherlands (1994)
8. D'Andrade, R., Wish, M.: Speech act theory in quantitative research on interpersonal behavior. Discourse Processes **8** (1985) 229-259
9. Dore, J.: Children's illocutionary acts. In R. Freedle (ed.), Discourse Comprehension and Production. Hillsdale, NJ: Erlbaum (1977)
10. Elman.J.L.: Distributed representations, simple recurrent networks, and grammatical structure. Machine Learning 7 (1991) 195-225
11. Graesser, A. C., Franklin, S., Weimer-Hastings, P., TRG: Simulating smooth tutorial dialogue with pedagogical value. Proceedings of the American Association for Artificial Intelligence Menlo Park, CA: AAAI Press (1998) 163-167
12. Graesser, A.C., Person, N. K.: Question asking during tutoring. American Educational Research Journal **31** (1994) 104-137
13. Graesser, A.C., Person, N.K., Magliano, J.: Collaborative dialog patterns in naturalistic one-on-one tutoring. Applied Cognitive Psychology **9** (1995) 359-387
14. Graesser, A. C., Swamer, S., Hu, X.: Quantitative discourse psychology. Discourse Processes **23** (1997) 229-263
15. Graesser, A. C., Wiemer-Hastings, K., Wiemer-Hastings, P., Kreuz, R., and the TRG: AutoTutor: A simulation of a human tutor. Journal of Cognitive Systems Research **1** (1999) 35-51
16. Kipp, M.: The neural pathway to dialogue acts. Proceedings of the 13th European Conference on Artificial Intelligence (1998) 175-179
17. Labov, W., & Fanshel, D. Therapeutic discourse: Psychotherapy as conversation. New York: Academic Press (1977)
18. Nagata, M., Morimoto, T.: First steps toward statistical modeling of dialogue to predict the speech act type of the nest utterance. Speech communication, 15 (1994) 193-203
19. Reithinger, N., Klesen, M.: Dialogue act classification using language models, in Proceedings of EuroSpeech-97, Rhodes (1997) 2235-2238
20. Quinlan, J.R.: Induction of decision trees. Machine Learning **1** (1986) 81-106
21. Quinlan, J.R., Rivest, R.: Inferring decision trees using the minimum description length principle. Information and Computation **80** (1989)
22. Person, N.K., Bautista, L., Kreuz, R.J., Graesser, A.C., and the Tutoring Research Group: The dialogue advancer network: a conversation manager in autotutor. To appear in the proceedings of the workshop on modeling human teaching tactics and strategies. ITS, Montreal, Canada (in press)
23. Person, N.K., Graesser, A.C., the TRG: Designing AutoTutor to be an effective conversational partner. To appear in the Proceedings for the 4th International Conference of the Learning Sciences Ann Arbor, MI (in press)
24. Samuel, K., Carberry, S., Vijay-Shanker, K.: Dialogue act tagging with transformation –based learning. In Proceedings of the 17th International Conference on Computational Linguistics & the 36th Annual Meeting of the Association for Computational Linguistics., Montreal, Canada (1998) 1150-1156
25. Searle, J. R.: A taxonomy of illocutionary acts. In K. Gunderson (ed.), Language, mind, and knowledge. Minneapolis: University of Minnesota Press (1975)
26. Vendler, Z.: Res cogitans: An essay in rational psychology. Ithaca, NY: Cornell University Press (1972)
27. Wahlster, W.: Verbmobil- translation of face-to-face dialogues, Technical report, German Research Centre for Artificial Intelligence (DFKI), (1993). In Proceedings of MT Summit IV, Kobe, Japan, 1993