Automated Discovery of Speech Act Categories in Educational Games

Vasile Rus
Department of Computer Science
The University of Memphis
Memphis, TN 38152
vrus@memphis.edu

Cristian Moldovan
Department of Computer Science
The University of Memphis
Memphis, TN 38152
cmldovan@memphis.edu

Nobal Niraula
Department of Computer Science
The University of Memphis
vrus@memphis.edu
nbnraula@memphis.edu

Arthur C. Graesser
Institute for Intelligent Systems
The University of Memphis
365 Innovation Drive, Memphis, TN 38152
graesser@memphis.edu

ABSTRACT

In this paper we address the important task of automated discovery of speech act categories in dialogue-based, multi-party educational games. Speech acts are important in dialogue-based educational systems because they help infer the student speaker's intentions (the task of speech act classification) which in turn is crucial to providing adequate feedback and scaffolding. A key step in the speech act classification task is defining the speech act categories in an underlying speech act taxonomy. Most research to date has relied on taxonomies which are guided by experts' intuitions, which we refer to as an extrinsic design of the speech act taxonomies. A pure data-driven approach would discover the natural groupings of dialogue utterances and therefore reveal the intrinsic speech act categories. To this end, this paper presents a fully-automated data-driven method to discover speech act taxonomies based on utterance clustering. Experiments were conducted on three datasets from three online educational games. This work is a step towards building speech act taxonomies based on both extrinsic (expert-driven) and intrinsic aspects (data-driven) of the target domain.

Keywords

Speech act discovery, dialogue systems, educational games.

1. INTRODUCTION

An important task in dialogue-based educational systems is the detection of students' intentions from their natural language input, which we refer to as utterances. Speakers' intentions are modeled using elements from the speech act theory (Austin, 1962; Searle, 1969). Speech act theory was developed based on the "language as action" assumption as explained later. The automated detection of speaker's intentions in dialogues is known as the task of speech act classification.

Examples of speech acts are Questions, Statements, or Greetings. For instance, the hearer infers from the following utterance *How did you do that?* that the speaker is asking a Question, which informs the hearer to prepare an answer. Sometimes the speaker just states something as in the Statement, *The situation is getting worse every day.*, or greets someone as in *Hello!*.

Our work is conducted in the context of multi-party epistemic games in which chat rooms play an important role. For instance, in an Urban Science game, 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. The players interacted with the mentor through a chat facility provided in the game. All chat among players and mentors was logged.

If the mentor role is to be automated, in a tutoring system, we need to automatically manage the dialogue which involves identifying student-players' intentions (speech act classification) based on their utterances as well as to select the best speech acts the auto-mentor system needs to produce (speech act prediction) for feedback and scaffolding. The details of the games from which we collected data are presented in the Experiments and Results section.

The task of speech act classification has been extensively addressed by the intelligent tutoring systems (ITS; [1,2]) and natural language processing (NLP; [3,4,5]) communities. The related task of speech act prediction, which is about deciding what next speech act the automated dialogue system should generate, has also been investigated to some extent [6,7,8].

The NLP and ITS communities have addressed mainly the task of speech act classification and usually in simpler setups than ours: one-to-one dialogues, e.g. between an intelligent tutor and a student user or between a ticket-booking system and a human traveler. In contrast, the present study addresses multi-party dialogues in which more than two dialogue partners are involved. This has implications on the adopted solution to classify or discover the speech acts. Some predictive features that are easy to extract in dialogues between two partners become more challenging in speech act classification or discovery for multi-party dialogues. For example, the previous speech act feature which is useful to predict the current speech act in dialogues between two partners, e.g. after a Question by one speaker an Answer by the other speaker follows, becomes more challenging in multi-party dialogues because the previous speech act is not always directly linked to the current speech act, as in the case of a third partner joining the discussion suddenly.

Furthermore, the solutions to the task of speech act classification proposed by the ITS and NLP researchers are not fully automated because the important step of specifying the speech act taxonomy is manually handled by experts [9]. The expert-generated

taxonomies are specified extrinsically as experts generate them in an ad-hoc manner without an exhaustive analysis of the available data. Indeed, Andernach, Poel, and Salomons [10] indicates that experts define taxonomies based on their intuitions with minimum information from actual data which makes it hard to define a set of rules that different human annotators (or machines) could consistently apply to data in order to derive the same speech acts for similar utterances. In general, experts define a wishful taxonomy and then the hope is the automated algorithms could learn automatically the patterns to detect the speech acts in the taxonomy. There are other lingering issues with the expert-defined taxonomies as Traum [9] pointed out. Among these issues, Traum mentions the "significant challenges for creating a taxonomy of dialogue acts that can be understood and used by researchers other than the taxonomy designers."[1] We believe that a data-driven approach to discover or at least inform the creation of speech act taxonomies could be extremely useful. This work is a step in this direction of creating taxonomies based on both extrinsic and intrinsic processes.

We propose a data-driven approach that infers the intrinsic speech act categories from the data based on the similarities of the dialogue utterances according to some model, e.g. using lexical and positional information from the utterances. The method is based on clustering algorithms, both parametric (K-Means) and non-parametric (Expectation-Maximization), to group dialogue utterances into homogeneous groups which are then used to define the speech act categories. An automated method to discover the speech categories could complement and also be used as a validation tool for expert-defined taxonomies. The natural language community has largely ignored the task of automated discovery of speech act taxonomies; there has been only one early attempt nearly two decades ago [10]. To the best of our knowledge, no previous work proposed such an automated method for speech act discovery in the area of dialogue-based intelligent tutoring systems and the larger ITS community.

Our effort fits within the grander goal of building data-driven dialogue managers [11, 12]. The closest work to our own effort in the area of educational systems is by Kristy Boyer and colleagues ([12, 13]). They automatically derived 'dialogue modes' from sequences of dialogue acts (a modern definition of speech acts), instead of asking experts to define the dialogue modes. The best number of dialogue modes is found intrinsically by selecting inferred sets of dialogue modes that maximize a log-likelihood fit function. We follow a similar idea but instead of inferring sets of dialogue modes we infer categories of speech acts and rely on clustering algorithms instead of Hidden Markov Models as Boyer and colleagues did. Hidden Markov Models are best suited for inferring hidden variables from sequences of events. In our case, we were interested in the discovery of hidden similarity patterns among individual utterances and thus clustering was a natural choice. We chose K-Means and Expectation Maximization (EM) as the clustering algorithms. The former requires a priori specification of the number of clusters expected while EM can automatically infer the number of clusters through cross validation. The appealing of K-Means is its simplicity and the ease of interpretation, e.g. a centroid vector for each cluster is provided which can be used to interpret the cluster. In the case of K-Means we experimented with several pre-specified numbers of clusters. By default, the results thus obtained are compared with the expert-defined number of clusters, i.e. the expert speech act categories.

The rest of the paper is organized as in the followings. The next section provides an overview of speech act theory and speech act taxonomy work. We then provide the conceptual framework behind our basic idea to cluster dialogue utterances. The Experiments and Results section describes our experimental setup and the results obtained. We conclude with Conclusions and Future Work.

2. RELATED WORK

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 construct in linguistics and the philosophy of language that refers 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 [14]. According to Searle [15], 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. Second, there is the illocutionary act, which is the real intended meaning of the utterance, its semantic force. Third, there is the perlocutionary act which is the practical effect of the utterance, such as scaring, persuading, and encouraging.

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.

[1] Dialogue acts, speech acts, communicative acts, conversational acts, conversational moves, or dialogue moves are terms used by different researchers to refer to the same general concept [9].

Speech Act Category	Example	Count
Statement	I'll be your planning consultant.	605
Request	Click that and click "New Staff Page"	343
Reaction	Ah, I see.	642
MetaStatement	i didn't understand what maya wanted	176
Greeting	Hello!	103
ExpressiveEvaluation	good!!!!!!!!!!	166
Question	why am i getting notes from people not in my group?	646
Other	same thing what	87

Table 1. Our flat Speech Act Taxonomy with examples for each speech act category.

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” [14]. 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.”

The present study assumes there is one speech act per utterance and the set of speech acts used are all at the same level of depth thereby forming a flat hierarchy. These simplification assumptions are appropriate for a first attempt at automating the speech act discovery process.

2.1 Speech Act Taxonomies

As already mentioned, the tasks of speech act classification and prediction requires the existence of a predefined set of speech act categories or speech act taxonomy.

Researchers agree that defining a taxonomy in general and a speech act taxonomy in particular implies a balancing act between power and simplicity ([9, 16]). That is, defining a taxonomy implies interactions between the experts’ conceptual view of the target domain with an emphasis on power, i.e. capturing fine distinctions that would maximize reaching the goal the taxonomy will serve such as effective tutoring dialogue in our case, and the need for reliable annotation and predictions, i.e. maximizing the reliability with which human annotators can tag the speech acts in which case a few, well-defined categories are better than many, sophisticated categories.

Less emphasis has been paid to the relation between the taxonomy and the actual method to automatically recognize the speech acts in the taxonomy. In other words, taxonomies were refined by observing how reliably human annotators can use them to annotate data D’Andrade and Wish [17]. The degree to which the human annotators’ process may be replicated through an automated method or the intrinsic similarities among dialogue utterances within the constraints of a chosen model, e.g. leading tokens utterances [18], has been left as an afterthought. Our work is a step towards building taxonomies based on both expert and data-driven approaches which we believe could lead to a needed trade-off between power and accuracy. That is, while expert-defined taxonomies could lead to best outcomes conceptually but may sometimes be hard to detect, the data-driven approaches would lead to taxonomies that are derived from patterns in the data and would therefore result in good speech act classification performance. A mixed approach could provide the necessary trade-off between desirable speech act categories and classification performance. It should be noted that experts do consult data, in a limited way, when deriving their taxonomies [17]. However, an automated method for grouping dialogue utterances as proposed here would infer speech act categories from the entire available data in a systematic way.

We analyzed the speech act taxonomies proposed by researchers over the years. Some are flat while others are multi-layered. The layers in the multi-layered taxonomies can be viewed as levels, in which higher level speech acts are composed of lower level speech acts, or ranks, in which layers represent different phenomena [9]. We present next a summary of the most important ones as judged from a history and relevance to our own work.

The classic categorization of Austin [14] postulates five major speech act classes based on five categories of performative verbs: Expositives - verbs asserting or expounding views, classifying usages and references; Exercitives – verbs issuing a decision that something is to be so, as distinct from a judgement that it is so; Verdictives - verbs delivering a finding, official or unofficial, upon evidence or reason as to value or fact; Commissives - verbs commiting the speaker to some course of action; and Behabitives - verbs involving the attitudinal reaction of the speaker to someone’s conduct or fortunes [17].

The taxonomy proposed by Searle [15] consists of six major classes: Representatives - committing the speaker to something’s being the case; Directives - attempt by speaker to get the hearer to do something; Commissives – committing the speaker to some course of action; Expressives - expressing the psychological state specified; Declarations - bringing into existence the state

described in the proposition and Representative; and Declarations - giving an authoritative decision about some fact.

The category scheme proposed by D'Andrade and Wish [17] treats most utterances as conveying more than one speech act and does not attempt to establish a hierarchical order among multiple speech acts. The primary motivation for the speech act coding system was a desire to investigate correspondences between speech acts and adjectival "dimensions" descriptive of interpersonal behavior. In order for a classifying system to be useful for measuring interpersonal communication, the distinctions reflected by the coding scheme should be relevant to native speakers' perceptions and evaluations of interaction. Their classes are: Assertions (Expositives), Questions (Interrogatives), Requests and Directives (Exercitives), Reactions, Expressive Evaluations (Behabitives), Commitments (Commissives) and Declarations (Verdictives, Operatives).

While there seems to be some consensus on the existence of some speech acts, like greetings, questions, answers, etc., the efficiency of a particular taxonomy for solving a particular problem ultimately rests on the task at hand. For instance, Olney and colleagues [19] used a taxonomy that divided questions into 16 subcategories and had only 3 classes for the rest of the utterances, which was suitable for a particular intelligent tutoring environment. The 16 subclasses of Questions were: Verification, Disjunctive, Concept Completion, Feature Specification, Quantification, Definition, Example, Comparison, Interpretation, Causal Antecedent, Causal Consequence, Goal Orientation, Instrumental/Procedural, Enablement, Expectational and Judgmental.

In the case of Verbmobil, a research project aiming to develop a system that can recognize, translate and produce natural utterances, the taxonomy used takes into consideration in which of the five dialogue phases the actual speech acts occur. The main classes of their taxonomy tree are: Request, Suggest, Convention, Inform and Feedback which all yield subclasses. For instance, the Convention class is composed of the following subclasses: Thank, Deliberate, Introduce, Politeness Formula and Greeting [20].

In our work, we will use the set of speech act categories, shown in Table 1. The speech act categories are based on theoretical schemes that also can be reliably coded by trained judges [14, 15, 17, 19]. We use this reference taxonomy as a benchmark for comparison purposes with the automatically derived set of speech act categories.

3. THE APPROACH

Our approach to the automatic identification of speech acts classes is achieved using clustering algorithms.

Clustering is the unsupervised classification of data points (usually represented as vectors in a multidimensional space) into groups (clusters) based on similarity. A cluster is therefore a collection of objects which are similar to each other in the same cluster and are dissimilar to objects belonging to other clusters. The clustering problem has been addressed in many contexts and by researchers in many disciplines. This reflects the broad appeal of clustering and its usefulness as one of the steps in exploratory data analysis. In our case, we use clustering to discover intrinsic speech acts in dialogues from online educational games.

Table 2 offers examples of utterances belonging to three different speech act categories as defined by experts. In our method, the clustering algorithm would be fed a set of utterances of this type (see Table 2) and produce clusters in which similar utterances, i.e. utterances encoding the same speech act, belong to the same cluster. A quick post-hoc analysis by a human interpreter of the clusters thus obtained would allow the labeling of each cluster with a speech act label. For instance, by analyzing the utterances in the first column in Table 2, we immediately realize that they are all greetings and therefore a good label for such a cluster would be Greetings corresponding to the speech act category of Greetings. In this paper, however, we use the expert-labeled speech act categories to evaluate the obtained clusters.

An important step in clustering a set of data points, including dialogue data, is how to represent the data. In general, clustering algorithms require a vector representation. The dimensionality of the vector space is a choice the experimenter makes. In our case of clustering dialogue utterances, we rely on the hypothesis that good speakers in collaborative (as opposed to competitive or deceitful) dialogues make their intentions clear early on in their utterances allowing hearers to detect the speakers' intentions. Intuitively, the first few words of a dialog utterance are very informative of that utterances speech act. We could even show that some categories follow certain patterns. For instance, Questions usually begin with a wh- word while speech acts such as Greetings use a relatively small bag of words and expressions, i.e. Greetings are closed-class of utterances similar to function words such as prepositions or determiners.

In the case of other classes, distinguishing the speech act after just the first few words is not trivial, but possible. It should be noted that in typed dialogue, which is a variation of spoken dialogue, some information is lost. For instance, humans use spoken indicators such as the intonation to identify the speech act of a spoken utterance. We must also recognize that the indicators allowing humans to classify speech acts also include the expectations created by previous speech acts, which are discourse patterns learned naturally. For instance, after a first greeting another greeting that replies to the first one is more likely. In multi-party dialogue the previous speech act is more complex so consecutive utterances may or may not be directly related. We ignored such intonation and contextual clues so far in our work in order to explore the potential of classifying speech acts based on words alone. We do plan to incorporate contextual clues in future experiments.

One other argument in favor of this leading words assumption is the evidence that hearers start responding immediately (within milliseconds) or sometimes before speakers finish their utterances ([21] - pp.814). Further evidence of the leading words or tokens hypothesis has been provided by Moldovan, Rus, and Graesser [18] who showed that using "leading tokens" in an utterance leads to impressive speech act classification performance.

Therefore, we adopted a model in which each utterance is represented by its leading tokens (words and punctuation). This model includes the tokens themselves as well as their positions thus relying on lexical, punctuation, and positional information. Punctuation is useful in chat rooms as one of its functions is to encode intonational information which is lost in typed dialogues.

Greetings	Questions	Expressive Evaluation
Bye	what do i say ?	nice work , Player112 .
Bye Player102 !	hahah what ??	this chat thing is sooooo cool
bye guys	yep what now ?	nice work everyone , check your inbox
Bye	what do you like to do , etc .	That 's great .
Bye	What sort of background qualifies you for this internship?	Player109 great .
Bye !	what was in your notes ?	thanks for your help , laura

Table 2. Example of dialogue utterances that belong to the same speech act category as identified by experts.

4. EXPERIMENTAL SETUP AND RESULTS

We present in this section the experiments we conducted and the results obtained by automatically clustering dialogue utterances in order to discover the intrinsic speech act categories in the data.

The results are reported in terms of accuracy with respect to the expert-labeled speech act categories. After clustering the utterances, the expert-assigned label of the majority of the instances in a cluster is assigned as the predicted label of the cluster and thus all the instances in that cluster are given this label. Accuracy is then computed as the percentage of correctly predicted instances.

There are two major categories of clustering algorithms. Hierarchical clustering algorithms produce a nested series of partitions based on a criterion for merging or splitting clusters based on similarity. Partition based clustering algorithms identify the partition that optimizes (usually locally) a clustering criterion. Example algorithms from each category are hierarchical agglomerative (HAC) and K-means, respectively. HAC produces a hierarchical structure of clusters while K-means leads to a flat, direct clustering. In HAC, each data point is initially regarded as an individual cluster and then the task is to iteratively combine two smaller clusters into a larger one based on the distance between their data points. In the K-means algorithm, we specify a priori the number of clusters (K) we would like to have in the end. The algorithm usually starts with K seed data points which are considered as individual clusters. In subsequent iterations, the remaining data points are added to some cluster based on the distance to the centroid of each cluster. The centroid is an abstract data point of an existing cluster that is found by averaging over all the other points in the cluster. A distance metric must be defined for clustering algorithms. In our experiments, we used Euclidian and Manhattan distances. The reported results are with the Euclidian distance which produced results similar to the Manhattan distance. To perform clustering, we needed to set a couple parameters: number of clusters, which informs the clustering algorithm how many clusters to generate, and seed. The seed value is used in generating a random number which is, in turn, used for making the initial assignment of instances to clusters. In general, K-means is quite sensitive to how clusters are initially assigned and thus it is often necessary to try different values and evaluate the results. We have explored seed values between10 and 100 with an increment of 10. The best obtained results are reported, which correspond to seed values of 10 and 20. We used EM and K-means implementation from WEKA [22].

We collected dialogue utterances from three online educational games. While in general a dialogue utterance or turn may contain one or more sentences, in our context an utterance usually contains one sentence, with few exceptions. Therefore, the sentence was chosen as the unit of analysis. This choice can also be justified by the fact that it is closer to the ideal situation in which one and only one speech act is performed per unit of speech, i.e. an utterance.

A first data set used for this analysis came from a study run using an epistemic game, Urban Science. Urban Science is an educational game in which players, using iPlan, a custom-designed Geographic Information System, work as urban planners to change the look and feel of Madison, Wisconsin. They listen to people's concerns, redesign the city, and present their findings to family, friends, and planning experts. Urban Science explores how innovative technology-based learning environments modeled on the professional practices of urban planners inform students' understanding of ecology. The main goal of the game is to help players learn about ecology, develop self-confidence and presentation skills, and start to see the world through the eyes of a problem-solving urban planner.

The Urban Science chat data was collected from a November 2008 game run in Milwaukee and consists of online chat posts by the students and mentors exchanging information about the game rules, content, questions, advice, suggestions, according to the game plan. The posts, collected by the game log, were further preprocessed first by splitting them into sentences, and secondly by manually labeling each sentence with a category of the 8-class taxonomy (Statements, Requests, Reactions, Meta Statements, Greetings, Expressive Evaluations, Questions and Others).

The resulting 2768 sentenced were manually classified separately by two trained annotators. Most of the speech act categories had high levels of reliability (kappas greater than 0.7) among the human coders, but two of the categories (Meta Statement and Other) had moderate kappa scores of 0.546 to .587. The overall mean kappa score across all 8 speech act categories was 0.797.

The class distribution is shown in Table 1. If one were to randomly assign a speech act category according to these distributions, the likelihood of selecting the correct speech act category by chance would be .18. The average number of tokens per sentence is 7.57, with a Standard Deviation of 6.40.

The clustering results for the Urban Science data using the Expectation-Maximization algorithm are shown in Table 3, second column. The first column represents the number of leading tokens used for a particular trial. For instance, the third row from the top corresponds to the model in which three leading tokens

Leading Tokens	#Clusters/Urban Science	#Clusters/Land Science	#Clusters/Nephrotex	#Clusters/Combined
2 Tokens	5/34.4%	4/38.7%	3/38.9%	7/34.7%
3 Tokens	**6/40.3%**	**5/42.5%**	**4/38.2%**	**6/37.8%**
4 Tokens	4/36.2%	5/34.2%	4/38.7%	6/35.4%
5 Tokens	5/39.9%	5/36.4%	5/36.4%	6/37.9%

Table 3. Results with Expectation-Maximization clustering algorithm.

N	Urban Science	Land Science	Nephrotex	Combined
6 clusters	29.6%	36.4%	28.6%	35.2
7 clusters	27.2%	31.1%	26.9%	31.5
8 clusters	29.1%	30.3%	26.3%	27.8
9 clusters	31.3%	28.9%	26.1%	28.0
10 clusters	27.6%	26.2%	25.7%	27.0

Table 4. Results with K-Means clustering algorithm.

were used. The results show that the three leading tokens provide the best results and yields six discovered clusters. When evaluated against expert-assigned labels, the accuracy was 40.3% for the leading three tokens model. A random guess would uniformly assign a dialogue utterance to each of the eight speech acts for an accuracy of 12.5%. Compared to the expert-defined speech act categories, the EM algorithm does not identify Greetings and Other speech acts. Greetings are mostly clustered in the predicted Reactions cluster.

Results for K-Means are shown in Table 4. The results are all for leading three tokens which was the best model when using the non-parametric EM algorithm. Remember that we do not have to specify a priori how many clusters we should expect when using the EM algorithm which is the reason we first used EM to find the best model to use for the discovery of intrinsic speech act categories in the data. The first column in Table 4 indicates the number of clusters used. We tried values around the expert-defined number of clusters, which was eight clusters.

Land Science is another computer-based educational game, in which players become interns at the office of a fictitious urban and regional planning firm. The players have to weigh the trade-offs of land use decisions in ecologically-sensitive areas, interact with virtual stakeholders, and develop land use plans for local and national sites. It is a 10 hour game played in schools or out-of-school enrichment programs.

The Land Science data was collected from the log of a game run in 2010 at Massachusetts Audubon Society. The online chat posts were split into 4131 sentences which were than manually labeled independently by two humans. The inter-judge reliability scores ranged from 0.501 for the category Other to 0.918 for the category Question, with a mean of 0.755.

The class distribution is as follows: 2.3% Others, 2.3% Expressive Evaluations, 2.7% Greetings, 7.8% Requests, 8.4% Meta Statements, 19.0% Questions, 28.2% Statements and 28.9% Reactions, which means that the chance of the corpus is .21. The average number of tokens per sentence is 6.85, with a Standard Deviation of 6.69.

The results on the Land Science data set are shown in the third column of Tables 3 and 4. The best results are again for a model in which the three leading words were used. However, in this case the number of intrinsic speech act categories, i.e. clusters, is five. MetaStatements, Greetings, and Other are not identified as clusters by the three leading tokens model and the non-parametric EM algorithm.

Nephrotex is an educational game in which undergraduate engineering students role-play as professional engineers-in-training in order to develop the skills, knowledge, identity and values of engineers. In Nephrotex, students are welcomed as early career hires into the fictitious company Nephrotex, whose core technology is the ultrafiltration unit, or dialyzer, of a hemodialysis machine. The students' assigned task is to design a next-generation dialyzer that incorporates carbon nanotubes and chemical surfactants into the hollow fibers of the dialyzer unit.

Online chat posts were collected from a game run in 2011 and subsequently split into 1000 sentences which were later manually classified by two humans. The kappa scores for each of the eight categories when comparing the two trained judges ranged from .41 for class Other to .94 for class Question with an average of .68

The class distribution shows the following hierarchy: 1.1% Others, 1.4% Greetings, 2.4% Expressive Evaluations, 4.0% Meta Statements, 5.6% Requests, 17.3% Questions, 20.2% Reactions and 48.0% Statements, which indicates that the corpus' chance is .30 The average number of tokens per sentence was 9.01, with a Standard Deviation of 6.38.

The large corpus obtained by combining the previous three corpora, consists of a number of 7899 sentences, each labeled with one of the eight speech act categories. The distribution is as follows: 2.4% Others, 2.9% Greetings, 3.6% Expressive Evaluations, 7.1% Meta Statements, 9.1% Requests, 20.3% Questions, 25.8% Reactions and 28.5% Statements, resulting in a chance of .20. The average number of tokens per sentence is 7.37, with a Standard Deviation of 6.59.

For the Nephrotex corpus, the best results are obtained using the two leading tokens. However, the results obtained with the three leading tokens are comparable in terms of accuracy but not in the number of clusters discovered, three versus four. Because the three leading tokens model has been best in the other datasets, we incline to declare it a winner in this case too.

Finally, we also experimented with a combined dataset. Results are presented in the last column of Tables 3 and 4.

4.1 Balanced Data Set

Because the three datasets collected were dominated by certain categories, e.g. Questions, Reactions, and Statements, we wondered about the ability of the clustering algorithms to discover the intrinsic speech act categories when the data would be uniformly distributed.

To achieve this goal, we ran experiments on a balanced dataset of speech acts by extracting from the combined data set an equal number of utterances for each speech act. In the process, we dropped the *Other* category as too few utterances were available. In the end, we obtained a balanced data set of seven speech act categories, each category containing 230 utterances each.

N	#Clusters/Accuracy
2	4/29.8%
3	6/28.3%
4	5/31.7%
5	6/31.1%

Table 5. Accuracy and number of clusters obtain with EM algorithm on the balanced data set.

From the results in Table 5, we can see that the accuracy is quite similar for all values of N, i.e. the number of leading words used as predicting features in clustering. The leading three words generate six clusters (out of seven in the gold standard). MetaStatements were mostly labeled as Greetings, Statements, and Expressive Evaluations. For instance, the MetaStatement, "Yay!" expressing an emotion is similar to a Greeting because of its short length and exclamation mark. For short utterances which are shorter than the number of tokens used in a given model we introduce dummy values for missing tokens, e.g. NONE. So, "Yay!" and "Hi!" have similar representations except for the first tokens which explains why they are clustered. Given that ideally we would like to have a trade-off between the complexity of the model used, in our case defined by how many tokens are employed (the more tokens the more complex the model), discrimination power (number of distinguishable clusters), and performance, we conclude from the results in Table 5 that using the three leading words is best.

5. CONCLUSIONS AND FUTURE WORK

We proposed in this paper a fully automated method to speech act discovery. As we already mentioned, this work is a step towards a process of defining the speech act taxonomy using both extrinsic and intrinsic aspects of the target domain. The extrinsic aspects comprise of the goals of the system that needs the speech act taxonomy and the experts' knowledge and biases. The intrinsic aspects relate to the actual similarities of the actual data. A trade-off between the extrinsic and intrinsic forces could lead to a robust speech act taxonomy that is both informed by experts' views and by the actual data.

We presented results on the original dataset as well as on balanced datasets in which the gold standard (i.e., the speech act categories are validated by experts) had same numbers of utterances for each speech act. The balanced datasets offer a more fair comparison of the clustering method of the utterances in our epistemic games. However, sometimes domains such as educational systems may be biased towards particular speech acts in which case the original datasets offers us a view at the "real" world and how the proposed methods work in real settings.

A drawback of the proposed model for representing dialogue utterances, i.e. the N leading tokens, is that the distance between two dialogue utterances is based on string operations rather than lexico-semantic distances which would be more meaningful for natural language dialogues. That is, two utterances that contain the words 'hi' and 'hi' would be close in a string-based representation while 'hi' and 'hello' or 'hi' and 'bye' would not. While for the former example of 'hi' and 'bye' one could argue for the creation of a different cluster, or speech act category, for the former they should definitely be in the same cluster. One solution is to modify the clustering library in WEKA [22] to include a lexico-semantic distance based on word-to-word similarity measures, e.g. using the WordNet similarity library [23]. We do plan to explore this line of research in the future.

As one last conclusion, our work showed that there is close relationship between the model used, e.g. the number of leading tokens, and the number of intrinsic clusters found in the data. This result should inform the developers of speech act classifier who used a particular model about the power of that model to discover the intrinsic, extrinsic, or intrinsic-extrinsic speech act categories adopted.

6. ACKNOWLEDGMENTS

This research was supported in part by Institute for Education Sciences under awards R305A100875 and by the National Science Foundation awards #0904909 and NSF#0918409. Any opinions, findings, and conclusions or recommendations expressed in this material are solely the authors' and do not necessarily reflect the views of the sponsoring agencies.

7. REFERENCES

[1] Marineau, J., Wiemer-Hastings, P., Harter, D., Olde, B., Chipman, P., Karnavat, A., Pomeroy, V., Graesser, A., and the TRG. (2000). Classification of speech acts in tutorial dialog. In Proceedings of the workshop on modeling human teaching tactics and strategies at the Intelligent Tutoring Systems 2000 conference, pp. 65–71, 2000.

[2] Serafin, R. and Di Eugenio, B. (2004). FLSA: Extending Latent Semantic Analysis with features for dialogue act classification. ACL04, 42nd Annual Meeting of the Association for Computational Linguistics, Barcelona, Spain.

[3] Reithinger, N. and Maier. E. (1995). Utilizing statistical dialogue act processing in Verbmobil. In ACL95, Proceedings of the 33rd AnnualMeeting of the Association for Computational Linguistics.

[4] Ries, K. (1999). HMM and Neural Network Based Speech Act Detection. In Proceedings of ICASSP 99.

[5] Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., Taylor, P., Martin, R., Van Ess-Dykema, C., and Meteer, M. (2000). Dialogue act modeling for automatic tagging and recognition of conversational speech. Computational Linguistics, 26(3):339–373.

[6] Nagata, M. and Morimoto, T. (1993). An experimental statistical dialogue model to predict the Speech Act Type of

the next utterance. In Proceedings of the International Symposium on Spoken Dialogue (ISSD-93), pages 83--86, Waseda University, Tokyo, Japan.

[7] Reithinger, N. (1995). Some Experiments in Speech Act Prediction. In AAAI 95 Spring Symposium on Empirical Methods in Discourse Interpretation and Generation, Stanford University.

[8] Bangalore, S., and Stent, A. (2009). Incremental parsing models for dialog task structure. In Proceedings of the Meeting of the European Chapter of the Association for Computational Linguistics.

[9] Traum, D. R. (2000). 20 Questions for Dialogue Act Taxonomies, in Journal of Semantics, 17(1):7--30, 2000.

[10] Andernach, T., Poel, M. and Salomons, E. (1997). Finding classes of dialogue utterances with Kohonen networks. In: Daelemans, W., van den Bosch, A. and Weijters, A., editors, Workshop Notes of the ECML/MLnet Workshop on Empirical Learning of Natural Language Processing Tasks, pp. 85-94, Prague, Czech Republic.

[11] Bangalore, S., Di Fabbrizio, G., & Stent, A. (2008). Learning the structure of task-driven human-human dialogs. IEEE Transactions on Audio, Speech, and Language Processing, 16(7), 1249-1259.

[12] Boyer, K.E., Ha, E.Y., Phillips, R., Wallis, M. D., Vouk, M.A. and Lester, J.C. (2010). Dialogue Act Modeling in a Complex Task-Oriented Domain. In Proceedings of the 11th Annual SIGDIAL Meeting on Discourse and Dialogue, Tokyo, Japan, 2010, 297-305.

[13] Boyer, K.E., Ha, E.Y., Wallis, M.D., Phillips, R., Vouk, M.A., and Lester, J.C. (2009). Discovering Tutorial Dialogue Strategies with Hidden Markov Models. (2009). In Proceedings of the 14th International Conference on Artificial Intelligence in Education (AIED '09), Brighton, U.K., 2009, 141-148.

[14] Austin, J.L. 1962. How to do Things with Words. Oxford University Press, 1962.

[15] Searle, J.R. 1969. Speech Acts. Cambridge University Press, GB, 1969.

[16] Nielsen, R.D., Buckingham, J., Knoll, G., Marsh, B., and Palen, L. (2008). A Taxonomy of questions for question generation. In Vasile Rus and Art Graesser (Eds.): Proceedings of the Workshop on the Question Generation Shared Task and Evaluation Challenge, Arlington, Virginia, September 25-26, 2008.

[17] D'Andrade, R.G.; and Wish, M. (1985). Speech Act Theory in Quantitative Research on Interpersonal Behavior. Discourse Processes 8:2:229-258, 1985.

[18] 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.

[19] Olney, A.; Louwerse, M.; Mathews, E.; Marineau, J.; Hite-Mitchell, H.; and Graesser, A. (2003). Utterance Classification in AutoTutor. Building Educational Applications using Natural Language Processing: Proceedings of the Human Language Technology, Philadelphia, PA.

[20] Alexandersson, J.; Buschbeck-Wolf, B.; Fujinami, T.; Maier, E.; Reithinger, N.; Schmitz, B.; Siegel, M. (1998). Dialogue Acts in VerbMobil-2. volume 226, Verbmobil Report, German Research Center for Artificial Intelligence (DFKI), Saarbrcken, 1998.

[21] Jurafsky, D. and Martin, J.H.(2009). Speech and Language Processing. Prentice Hall, 2009.

[22] Witten, I. H.; and Frank, E. (2005). Data Mining: Practical machine learning tools and techniques, 2nd Edition. Morgan Kaufmann, San Francisco.

[23] Pedersen, T., Patwardhan, S., and Michelizzi, J. (2004). WordNet::Similarity - Measuring the Relatedness of Concepts, In the Proceedings of the Nineteenth Na-tional Conference on Artificial Intelligence (AAAI-04), pp. 1024-1025, July 25-29, 2004, San Jose, CA (Systems Demo).

.