12-3-2014

Automated Speech Act Classification in Tutorial Dialogue

Borhan Samei

AUTOMATED SPEECH ACT CLASSIFICATION IN TUTORIAL DIALOGUE

by

Borhan Samei

A Thesis

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Master of Science

Major: Computer Science

The University of Memphis

December 2014

This work is dedicated to my father, Jafar Samei, who supported me through my life and taught me the meaning of love. He will always be remembered and in my heart through every step I had have taken and will be taking towards the end of my life. Rest in peace dad.

Borhan
December 2014

Abstract

Samei, Borhan. MS. The University of Memphis. December 2014. Automated Speech Act Classification in Tutorial Dialogue. Major Professor: Dr. Vasile Rus.

Speech act classification is the task of detecting speakers’ intentions in discourse. Speech acts are based on the language as action theory according to which when we say something we do something. Speech act classification has various application in natural language processing and dialogue-based intelligent systems. In this thesis, we propose machine learning models for speech act classification that account for both content of the current utterance and context (previous utterances) of dialogue and we present this work on two domains: human-human tutoring sessions and multi-party chat based intelligent tutoring systems. The proposed speech act classification models were trained and tested on chat utterances extracted from the tutoring sessions and based on the domain specific properties of the datasets were designed to work with hierarchical and granular speech act taxonomies.

Table of Contents

List of Tables

1 Introduction

In natural language interactions, a sentence or utterance usually represents an indirect intention or function. For example, the utterance “What is your name?” is about asking someone’s name; the speech act and its function is asking a question while the direct semantic meaning is specifically asking what someone’s name is. Speech act has theoretical roots in linguistics and philosophy of language.

For years, philosophers assumed that a ‘statement’, for example, was only to ‘describe’ some state of affairs or facts while grammarians had pointed out that besides statement, one may express a command or wish (Austin, 1962). Philosophers, presented the theoretical roots of speech acts (language as action theory) as the intentions behind an utterance (e.g. sentence) which in turn determines the function of a discourse unit in dialog (Austin, 1962; Searle, 1969).

Austin (1962) proposed three levels of language: locutionary, illocutionary, and prelocutionary. The locutionary act is the performance of an utterance, more precisely, the locutionary act of an utterance is the actual utterance and its ostensible meaning, comprising phonetic, phatic and rhetic acts corresponding to the verbal, syntactic and semantic aspects.

An illocutionary act is the actual meaning and intent of the utterance and in some cases an utterance also has a perlocutionary act, i.e., the effect of the utterance on its audience. For example, the utterance “Do not push the red button!” is a locutionary act with certain semantic and phonetic features as well as the illocutionary act of warning someone not to push the red button and the prelocutionary act of persuading someone not to push the red button.

Speech act (Searle, 1969) or dialog act (Austin, 1962) is equivalent to a conversational game move (Power, 1979) and has also been referred to as adjacency pair part

(Schegloff, 1967). Speech acts play an important role in language and discourse processing specifically in applied natural language processing where the general goal is either to understand the intentions of natural language utterances or to generate language by artificially intelligent agents and modeling the dialog interactions.

Examples of such systems are Intelligent Tutoring Systems (ITS; Rus, D'Mello, Hu, & Graesser, 2013) where an intelligent agent is playing the role of student, tutor, or both with the usual goal of maximizing learning gains. These systems require an interaction mean between human and intelligent agents and one form interaction is dialogue. Within the framework of dialogue-based ITS (Rus et al., 2013), the intelligent agent is required to respond to humans which in turn implies the need for natural language processing. For example, the intelligent agent needs to determine if it is being asked a question or given a command in order to take proper action, e.g. give a more informative hint or just answer a question or say *Hello!* in case the student was greeting.

Based on several application of speech acts, researchers have presented machine learning and statistical modeling approaches to automatically identify the speech acts to help design and improve natural language processing models in artificially intelligent systems. In our work, we applied machine learning techniques to data sets of dialog-based tutoring sessions which were extracted from an intelligent tutoring system and an online human-human tutoring service with the ultimate goal of modeling the dialog in tutoring sessions.

1.1 Taxonomy

In order to represent the speech act of an utterance, a set of speech act categories is defined. The set of categories is also known as the speech act taxonomy. Based on the application and domain, different taxonomies may be defined to represent the most important acts that one is interested in identifying. Also, as the cost of tagging speech acts by human experts increases with large data sets, researchers have become more interested in approaches to automatically detect the speech acts using machine learning techniques, i.e., speech act classification, which emphasize the importance of having a proper taxonomy to be able to capture the acts that are more important in a particular domain.

As mentioned, speech act classification is based on the "language as action" theory which defines speech acts based on the illocutionary force of an utterance. Austin (1962) proposed a taxonomy of five speech act categories which are based on performative verbs (Table 1).

Table 1. Speech Act Taxonomy of Austin.

Category	Definition	Example
Vertictives	The delivering of a finding upon evidence.	describe, estimate
Exercitives	The giving of a decision.	order, command
Commissives	To commit the speaker to a certain action.	nominate, declare
Expositives	The expounding of views, arguments, and usages and references.	illustrate, accept
Behabitives	The notion of reaction to other people's behavior.	apologize, thank

Austin's taxonomy suggests using verbs to identify speech acts. It is important to note that not all the verbs are illocutionary verbs and particular verbs may overlap in some categories which leads to confusion between verbs and acts and implies the need to explore more sensible taxonomies (Searle, 1969, 1976).

Searle (1969, 1976) proposed a taxonomy on the basis of illocutionary and grammatical indicators which included five categories: representatives, directives, commissives, expressives, and declarations. While the first three categories of Searle's taxonomy are equivalent to Austin's, they added the "expressives" category and distinguished declaration from other categories which expanded the taxonomy beyond verbs (Table 2).

Table 2. Speech Act Taxonomy of Searle.

Category	Definition	Example
Representatives	Committing the speaker to something's being the case	Boast, conclude
Directives	To get the hearer to do something	Command, order
Commissives	To commit the speaker to some future action	This has been suggested to me by Julian Boyd.
Expressives	To express the psychological state	I congratulate you on winning the race.
Declarations	Successful performance of action	I declare: my position is hereby terminated.

D'Andrade and Wish (1985) proposed seven categories of speech act: assertions, questions, requests and directives, reactions, expressive evaluations, commitment, and declaration. They extended Austin and Searle's taxonomies by defining categories such as requests and questions as separate speech acts (D'Andrade & Wish, 1985).

Domain specific taxonomies are used in certain application. In an intelligent tutoring system, for instance, the speech act classification can be used to understand the student utterances which in turn is needed to produce a proper response from the intelligent agents. One of the main categories of speech acts in tutoring systems is 'question' which needs to be answered either by the student or tutor AI agents.

Graesser and Person (1994) proposed 18 categories for questions asked in tutoring and their question categories were applied in intelligent tutoring systems (Graesser & Person, 1994).

Forsyth and Martell (2007) developed a speech taxonomy for online chat corpus, including statement, system, greet, emotion, wh-question, yes/no question, continuer, accept, reject, bye, yes answer, no answer, emphasis, other, and clarify. Researchers have proposed several taxonomies to clearly articulate different situations and most of the taxonomy sets are based on a domain and application of the speech act classification (Forsyth & Martell, 2007).

Along with theoretically designed taxonomies, researchers have also proposed unsupervised approaches to detect the speech act categories based on data-driven techniques. Rus, Moldovan, Niraula, and Graesser (2012) applied unsupervised clustering methods using the leading tokens of each utterance as features to find the speech acts using the data sets from online chat interactions. The resulting clusters of utterances were labeled by experts and formed seven main speech act categories: Expressive Evaluation, Greeting, Metastatement, Question, Reaction, Request, and Statement (Rus, et al., 2012).

Table 3. Data Driven Speech Act Taxonomy.

Speech act category	Example from dataset
Expressive Evaluation	Your stakeholders will be grateful!
Greeting	Hello!
MetaStatements	Oh yeah, last thing.
Statement	A physical representation of data.
Question	What should we do?
Reaction	Thank you
Request	Please check your inbox

The previous work on speech act classification were mostly based on general speech act categories which in turn improved the performance of classification models. In our work, we first use a general taxonomy on a data set of multi-party chat dialogues (Chapter 3) and next we define and use a granular and hierarchical taxonomy which we use on data set of human on-to-one tutoring sessions (Chapter 4).

1.2 Machine Learning

In machine learning field, classification is the task of assigning an input instance with a tag or category from a predefined set of categories. Detecting the speech act of an utterance falls into the classification problems where the goal is to tag an utterance with its speech act and the speech act categories are defined in the taxonomy. Within this framework designing a model for speech act classification often implies detecting features that are predictive for speech acts and choosing a proper machine learning algorithm, e.g., Decision Trees, Naïve Bayes, etc. to learn from a set of annotated instances (i.e., training set) (Moldovan, Rus, & Graesser, 2011; Olney et al., 2003) or in an unsupervised manner (Ezen-Can & Boyer, 2014). Researchers have also applied statistical approaches for modeling dialog and tagging speech acts (Stolke et al., 2000).

Classification is equivalent to function estimation in which the function's values are nominal. The learning process of a classification results in a classifier which is built given a set of training data. The most common approach to learning a classifier is supervised learning. Supervised machine learning implies learning from labeled training data. In supervised classification the training instances are usually tagged by human experts or any other means to generate a set of gold standard data. Other machine learning approaches are semi-supervised and unsupervised. In semi-supervised approach the model is initially learned from a set of tagged data but is then tested and/or retrained with untagged instances. Unsupervised learning is when no tagged data is provided and the model is learned based only on the observed features of the data. In computational linguistics, the cost of having experts tag training instances to form a gold standard set is usually high which is why researchers have been interested in applying semi-supervised (e.g., co-training and self-training) and unsupervised approaches to particular problems, e.g., part of speech tagging (Abney, 2007).

In order to be able to learn from a set of training instances, the instances are represented by features (attributes). Reducing an instance to its features (i.e., feature extraction) is an important step in machine learning.

A feature may be a nominal value, a real number, or any other type of attribute which represents a property of the data or observations. The number and types of features and the learning algorithm directly affect the complexity of learning process.

A particularly simple supervised learning algorithm which is common in classification is Naïve Bayes. Naïve Bayes is a probabilistic classifier and it based on the Bayes theorem. Naïve Bayes was first introduced by the text retrieval community (Russel, 2003). The Naïve Bayes model is represented by a set of conditional probabilities which outputs the probability of seeing a certain instance belonging to a category. Other popular classifier learning algorithms are decision trees and logistic regression. Based on the nature of the training instances and features used, one algorithm may be more appropriate for a certain problem; however, researchers usually compare the performance of different algorithms to come up with the final best model.

In our work, the goal is to design a classifier to tag speech acts in tutoring session utterances. We applied supervised machine learning techniques to two datasets: Multi-party chat-based ITS (Chapter 3) and transcripts from online tutoring services (Chapter 4). The next chapter is an overview of the related work in the field of speech act classification and it is followed by a description and results of our work on two datasets (Chapters 4 and 5) and future work (Chapter 6).

2 Related Work

Researchers have proposed several approaches to speech act classification, mostly supervised machine learning techniques. There has been a variety of work on building models based on features ranging from positional information of the turns in dialogue (Freschke, Gurevych, & Chebotar, 2012), lexical or syntactical features (Bangalore, Di Fabbrizio, & Stent, 2008; Stolcke et al., 2000), etc.

Freschke et al. (2012) developed an annotation schema based on the Wikipedia Talk pages where a conversation was divided into turns and each turn could consist of multiple speech acts. Their taxonomy included 17 categories and they trained binary classifiers for each category using Naïve Bayes, J48 decision trees and Support Vector Machines. Their features included uni-, bi-, and trigrams, the time distance of turns (in seconds) and the length of current, previous, and next turn, the position of turn within the discussion thread and a binary feature to represent weather a turn references or is referenced by other turns. They achieved an F-score of 0.82. The main property of their approach is training a separate classifier for each category and building a classification pipeline (Freschke et al., 2012).

A common application of speech act classification is in combination with speech recognition to model dialogue structure in spoken conversations, e.g., phone conversations. Stolcke et al. (2000) proposed a statistical approach to predict dialogue acts based on lexical, collocational, and prosodic cues. Their data set included 1155 human-human phone conversations and they achieved an accuracy of 0.65 on recognized speech and 0.71 on the word transcripts (Stolcke et al., 2000).

Kim, Cavedon, and Baldwin (2010) proposed speech act classifiers based on features such as bag of words along with information about dialogue structure such as the author of utterances and tested different models on online chat utterances extracted from online-shopping customer feedback data. They found that adding the information about the structure of dialogue improved their models' performance while sequential models (CRF) showed the best accuracy (Kim et al., 2010). In other work, Tavafi, Mehdad, Joty, Carenini, and Ng (2013) used the speaker of utterances as features and showed the effectiveness of SVM-hmm models on speech act classification. (Tavafi et al., 2013)

Ashok, Borodin, Stoyanchev, and Ramakrishnan (2014) presented a model based on several feature sets such as unigrams, syntactic, context-related, task-related, and presence of words. They tested the performance of multiple classifiers such as Support Vector Machine (SVM), J48 Decision Tree and Random Forest. The best performance (0.9 precision) was achieved by Random Forest classifier (Ashok et al., 2014).

In addition to the mentioned approaches, researchers have also proposed models for speech act classification using acoustic as well as lexical information (Jurafsky, Shriberg, Fox, & Curl, 1998; Rangarajan Sridhar, Bangalore, & Narayanan, 2009) or non-verbal features such as body posture (Ha, Grafsgaard, Mitchell, Boyer, & Lester, 2012). Rangarajan Sridhar et al. (2009) proposed modeling the sequence of acoustic-prosodic values as n-gram features and using maximum entropy model for speech act classification. Their model used context in the form of lexical, syntactic, and prosodic cues from preceding utterances which yielded an accuracy of 0.72 (Rangarajan Sridhar et al., 2009).

Speech act classifiers are often based on supervised machine learning, however there has been several works on building classifiers in an unsupervised setting. Unsupervised clustering of utterances is an example of such work (Ritter, Cherry, & Dolan, 2010; Rus et al., 2012). Joty, Carenini, and Lin (2011) proposed unsupervised algorithms based on Hidden Markov Modeling (HMM) to classify speech acts in email and forum conversations (Joty et al., 2011). Crook, Granell, and Pulman (2009) investigated applying Dirichlet Process Mixture Model for unsupervised clustering of dialogue utterances (Crook, Granell, & Pulman, 2009).

Researchers have proposed unsupervised classification and clustering techniques (Ezen-Can & Boyer 2014) as well as supervised algorithms (Moldovan et al., 2011; Onley et al., 2003; Rasor, Olney, & D'Mello, 2011; Samei, Li, Keshtkar, Rus, & Graesser, 2013) to model discourse within the framework of intelligent tutoring systems where human interacts with artificially intelligent agents.

In this thesis, we applied supervised machine learning and trained and tested classifiers in transcripts from online tutoring services sessions (Tutor.com) as well as multiparty chat conversations from an intelligent tutoring system (Landscience). The models investigated in this thesis build upon on Rus and colleagues' (2012) work which showed that general speech act categories can be predicted by using the first few tokens of utterances as features. We extended the feature sets, in particular we added features that capture the context of the previous dialogue, and tested multiple learning algorithms.

3 Multi-party dialogue (Landscience)

In the first step towards examining speech act classification models, we designed models on a data set of multi-party dialogue from student-mentor chat sessions in an online tutoring system (Landscience). Land Science is an epistemic game-like ITS designed to simulate an urban and regional planning internship experience for students (Shaffer & Gee, 2007). During the game students make land use decisions with the guidance of the mentor in order to meet the desires of virtual stakeholders. The students communicate with their team members and the mentor through a text-based chatting interface.

Currently, the human mentor guides the students to play the game while an artificially intelligent agent (AutoMentor) is being developed based on analysis of conversations between students and human mentor,. To find out the best prediction or conditions of the AutoMentor's conversational patterns, analyses have been conducted using the features of the speech act, a state transition network between adjacent speech acts, an epistemic network analysis, the newness and relevance of the chat contributions in the discourse space, along with the game elements, i.e., time parameters, topics, and activities at the different game stages. It is beyond the scope of this thesis, however, to discuss the various components and mechanism of AutoMentor. Instead, the immediate goal is to examine different models for speech act classification in mentor-student interactions.

3.1 Dataset

Our training data was extracted from a dataset of mentor-student chat utterances from seven Land Science full sessions. A hundred high- and middle-school students participated in the game in three conditions: during vacation, in school, and remote with one on-site meeting before they started the game. A total number of 26,148 chat utterances were generated by the players and the mentor. About 55% utterances were posted by the 100 players. We randomly extracted chat utterances to form our training data and adjusted the training data to include an even distribution of 30 instances per speech act category.

3.2 Approach

Our approach to speech act classification is a supervised machine learning approach. In this approach, models of the tasks are proposed as sets of features. Parameters of these models are learned/trained from annotated data and the performance of the learned models is then assessed on new, test data. The parameters of the proposed models are learned using several machine learning algorithms, i.e., decision trees and naïve Bayes.

The feature set was designed based on two principles: first, it is intuitively inferred and tested that human identified the speech act of an utterance as soon as they heard the first few words (Rus et al., 2012), namely, the first leading tokens. However, the context of an utterance is assumed to improve accuracy, e.g. it is more likely that after a *question* an *answer* follows as opposed to a *greeting*. Thus, the second feature set included the contextual information, e.g. speech act category of the more recent few utterances.

Briefly, our feature set consisted of content (non-contextual) features of the current utterance and contextual features (speech acts of previous utterances). The content, non-contextual features include the first two tokens and the last token which were represented as the actual string of characters (tokens) and the length of the utterance by word. The rest of the features captured contextual information with the five prior utterances (the speech acts and authors of these utterances). In this chapter, we use a general taxonomy consisting of a set of seven categories which was proposed by Rus et al. (2012) (Table 3).

3.3 Human Annotation

In order to examine the performance of our models, a set of mentor-student chat utterances were extracted randomly among different groups and different stages in the Land Science game. This data set was annotated by one human expert within the context of the chats. The human expert had access to the whole dialogue and context of the conversation. This annotated data set is deemed as the reference annotation and includes 30 utterances per speech act category. In order to examine the impact of the limited contextual information defined in our automated models (speech acts of previous five utterances), this data set was further annotated by another human judge in two forms, respectively.

First, the utterances were randomly ordered and the rater annotated them without considering the limited context. Second, each utterance was accompanied by the speech act category of five prior utterances and rater annotated the data considering the contents of the current utterance and prior context. Table 4 presents inter-reliability data for the two sets of ratings (with and without context) and comparisons with the reference annotation.

As shown in Table 4, the inter-rater reliability between human judges improved significantly by adding contextual information. The agreement on Metastatement, for example, improves significantly by adding context. Metastatements are the utterances which are generated when the players try to fix a communication break or confusion and knowing the context of a Metastatement utterance (prior utterances) helps identifying them since they often tend to refer to prior utterances to address an issue or a misunderstanding.

Table 4. Human Experts‘ Agreements.

Categories	Kappa		
	With Context-Without Context	Without Context-Reference annotation	With Context-Reference annotation
Greeting	0.87	0.85	0.90
Metastatement	0.55	0.53	0.92
Question	0.89	0.79	0.86
Reaction	0.18	0.27	0.44
Request	0.78	0.72	0.86
Statement	0.32	0.18	0.58
Expressive Evaluation	0.56	0.51	0.75

Besides adding context to our feature set, we refine the taxonomy to a hierarchical structure with three macro-categories: Initiative, Responsive, and Other. Each of new categories refer to a set of sub categories from the original taxonomy: Initiative (Question, Request, (fact) Statement), Responsive (Expressive Evaluation, Reaction), and Other (Metastatement, Greeting).

The Initiative category included speech acts corresponding to a speaker initiating a new dialogue segment. A fact statement, for example, can be uttered by a speaker who starts a conversation by simply asserting a fact. Similarly, Questions and Requests are typically initiated by a speaker and do not reflect reactions to the conversational partner(s)' moves. Responsive categories represented a response to a prior (initiative) speech act. Other category could occur in both situations. For instance, a greeting can be initiated or can be in response to another greeting. This new categorization is a bit challenging because of this dual behavior of the speech acts in the Other category. It violates the typical assumption in speech act classification that the three categories are disjoint or that an utterance may belong to only one category. Actually, some utterances received multiple labels from human annotators during the annotation process, e.g. statement and reaction were often confused. We had only used one of the labels in our experiments presented later.

This new structure allowed us to use different feature sets and models to maximize the accuracy for different speech act categories based on their nature and application. Using the annotated data set, we apply Decision Trees and Naïve Bayes machine learning models to create the automated speech act classifier. The performance of our models is presented in next section.

3.4 Results and Discussion

As it is shown in table 4, having contextual information improves the accuracy of human judgments. In fact, the more we know about context the better we can make decisions. Moreover, having contextual information dramatically changes the decision made by expert for certain speech act categories such as Reaction.

For the automated classification, our features set consist of two types of features: a set of 10 features which represent the context of the utterance by looking at the speech act category and speaker of five prior utterances, and 4 features representing the semantic information of the individual utterances including the first two tokens, last token, and the length of the utterance. The performance of proposed models is tested with two feature sets:

- **Semantic**: the feature set contains only the semantic information of the current utterance, i.e., three leading tokens, last token and the length of the utterance.
- **Contextual & Semantic**: the feature set contains both contextual and semantic information.

Besides the feature set, as mentioned earlier, the speech act taxonomy also has a vital impact on the performance of machine learning models. Moreover, because some categories are used more frequently, a high accuracy in predicting them is crucial to the performance of the conversational tutor, i.e., AutoMentor in our case. This requires a detailed evaluation of the models.

The hierarchical structure of the taxonomy enables the use of different models for classifying the different levels of categories. First, we need to classify the utterances into the three main categories: Initiative, Responsive, and Other. Next, for each such category a secondary classifier is trained to classify utterances inside each category.

For the first level of classification, we train the model on the reference annotations with two feature sets: semantic features only, and both semantic and contextual features. Tables 5 and 6 show the performances of different machine learning models on predicting the first level categories. The Naïve Bayes approach has a better performance for this step since the categories are generalized and using the same feature set the nodes in the decision tree overlap and may lead to confusion.

Table 5. Precision (P) and Recall (R) of First Level Models.

	Naïve Bayes				*Decision Tree(J48)*			
	Cont. & Sem.		**Semantic**		**Cont. & Sem.**		**Semantic**	
Category	***P***	***R***	***P***	***R***	***P***	***R***	***P***	***R***
Initiative	0.75	0.75	0.70	0.86	0.60	0.78	0.59	0.77
Responsive	0.64	0.63	0.78	0.55	0.73	0.45	0.73	0.45
Other	0.73	0.75	0.72	0.70	0.61	0.56	0.63	0.58

Table 6. Accuracy (A) and Kappa (K) of First Level Models.

	Naïve Bayes				*Decision Tree(J48)*			
	Cont. & Sem.		**Semantic**		**Cont. & Sem.**		**Semantic**	
Category	***A***	***K***	***A***	***K***	***A***	***K***	***A***	***K***
Initiative	0.75	0.57	0.86	0.58	0.78	0.38	0.77	0.36

Respon-sive	0.63	0.49	0.55	0.53	0.45	0.43	0.45	0.43
Other	0.75	0.64	0.70	0.59	0.56	0.43	0.58	0.46

Table 5 shows that adding context to the feature set improves the precision for Initiative category however for the Responsive category the precision is lower with context information. Overall context does not improve the performance of first level categorization except for slight changes in accuracy for Responsive and Other categories, since the first level categories are few and very broad. It is also learned that Naïve Bayes algorithm is more suitable than decision tree in first level categorization. Next, we focused on assessing the accuracy of speech act classification models with single and two layer taxonomy. Using the reference annotations as our training data, we created decision trees and Naïve Bayes learning models using WEKA (Hall et al., 2009) and we tested our models with 10-fold cross validation. The overall performance of models is evaluated with the two feature sets (semantic, context) and we also measure the performance of models with the 2-layer taxonomy vs. single layer flat taxonomy.

Table 7. Accuracy (A) and Kappa (K) with 1-layer and 2-layer Taxonomy.

Model	***Naïve Bayes***				***Decision Tree***			
Taxonomy	***1-layer***		***2-layer***		***1-layer***		***2-layer***	
Feature set	*A*	*K*	*A*	*K*	*A*	*K*	*A*	*K*
Semantic	0.53	0.29	0.74	0.70	0.55	0.48	0.50	0.47
Sem. & Cont.	0.54	0.47	0.75	0.71	0.56	0.48	0.64	0.58

As seen in table 7, Naïve Bayes algorithm is more suitable than Decision Tree and the maximum accuracy is achieved when the 2-layer taxonomy is used. The two feature sets however show a lower impact on the overall performance. It seems that the positive impact of the contextual features is low. Considering the significant impact of the 2-layer taxonomy on the overall performance of the models, in the next steps we use this multi-level taxonomy structure and evaluate the performance of the models on each category to examine the impact of contextual features in more detail.

Table 8. Accuracy and Kappa of Naive Bayes Models with Different Feature Sets on Second Level. (I=Initiative R=Responsive)

	Accuracy			***Kappa***		
Feature set	*I*	*R*	*Other*	*I*	*R*	*Other*
Semantic	0.71	0.78	0.76	0.56	0.56	0.53
Contextual & Semantic	0.72	0.66	0.88	0.58	0.33	0.76

Table 8 shows the performance of Naïve Bayes model for second level categorization by first level categories (groups of second level speech acts) to examine the general impact of context on predicting these groups of speech acts. The accuracy and kappa of 'Other' speech acts improves by adding context to our feature set, however the performance on 'Responsive' category becomes lower. It also inferred that context has a low positive impact on 'Initiative' categories. The different impact of context features on first level categories suggests that different feature sets improves the performance of the model on predicting certain categories while the performance decrease for some categories. We extend this analysis to the next level and examine the performance on individual second level categories (table 9).

Table 9. The Performance of Naive Bayes Models on Second Level categories with and without Context Features. (**P** = Precision, **R** = Recall, **A** = Accuracy, **K** = Kappa)

	Semantic	Contextual & Semantic

Categories	**P**	**R**	**A**	**K**	**P**	**R**	**A**	**K**
Expressive Evaluation	0.73	0.9	0.90	0.56	0.61	0.86	0.86	0.33
Greeting	0.78	0.73	0.73	0.53	0.96	0.8	0.80	0.76
Metastatement	0.75	0.8	0.96	0.76	0.82	0.96	0.96	0.76
Question	0.89	0.56	0.56	0.58	0.8	0.53	0.53	0.5
Reaction	0.87	0.66	0.66	0.56	0.77	0.46	0.46	0.33
Request	0.75	0.73	0.73	0.62	0.82	0.8	0.80	0.72
Statement	0.59	0.83	0.83	0.5	0.61	0.83	0.83	0.51

Based on the evaluations presented in Table 9, the context features have a positive impact on Greeting, Metastatement, Request, and Statement. The rest of the categories, on the other hand, have a better performance without contextual features. The results of the automated classification process are mixed but overall they suggest the use of the 2-layer taxonomy allowing us to apply different features and models for classification of different categories. Unlike human experts, the machine learning models' accuracy decreased on certain speech acts such as Responsive categories when we add context to the feature set.

3.5 Conclusion

Speech act classification has various applications in intelligent systems. Intelligent Tutoring Systems as an example can use speech act classification to identify student's inquiries in natural language interactions and determine an appropriate response from the intelligent tutor. In this chapter, we examined the role of context and taxonomy structure in the performance of automated speech act classification.

Based on the nature of speech act categories, we divided our taxonomy in three groups: Initiative, Responsive, and Other. This breaks the classification in two levels. In the first level classification, contextual features seem to not have a significant impact on the performance of models; however for the second level classification depending on the first level class, adding context improves the performance on certain categories such as Request, Greeting, and Metastatement while the performance of the model on some categories such as Question and Expressive evaluation is lower when we add context to the feature set.

The results presented in previous sections show that having some sort of contextual information has a positive impact on the accuracy of speech act classification for both human and computer. The nature of speech act categories used as the taxonomy set, on the other hand, is another main factor in the performance of certain models. In next chapter, we extend the scope of our approach to a larger data set from one-on-one tutoring sessions, use granular hierarchical taxonomy, and design models with similar and more features.

4 Online Tutoring Services (Tutor.com)

In previous chapter, we worked on a relatively smaller data set from multi-party chat dialogue sessions and used a general taxonomy with several machine learning models to train and test speech act classifiers. We used two kinds of feature sets and the results showed that adding contextual information to our feature set has a mixed impact on the performance. We also applied a hierarchical structure to the taxonomy which enabled using different kinds of features and models for different speech act categories.

In this chapter we extend our analysis and examine different models on a larger data set which is extracted from one-on-one tutoring sessions as opposed to multi-party chat. The taxonomy used in this chapter is more granular and based on a hierarchical structure, i.e., each speech act has a set of sub-categories (subacts). The feature sets are extended and more algorithms are tested in this chapter.

As mentioned earlier, in this work we used a data set of one-on-one human tutoring sessions which were extracted from Tutor.com algebra and physics tutoring sessions. The data set consisted of 1,438 sessions which included 95,526 utterances generated by tutors and students. The first step was to develop a taxonomy and have human experts tag the data set with speech acts.

The taxonomy used in this work was developed with the assistance of 20 subject matter experts (SMEs), all experienced tutors and tutor mentors working for Tutor.com which resulted in a fine-grained hierarchical taxonomy including 15 main categories where each main dialog act category consists of different sub-categories which resulted in 133 distinct dialog acts. Table 10 shows a list of main dialog acts with example.

Table 10. Top-level Speech Acts' Definition and Examples.

Act	Description	Example
Answer	A statement made in response to a question	Any non-zero integer.
Assertion	A free-standing statement (no prior question)	We have to keep the equation balanced.
Clarification	An statement serving to clarify a prior statement.	I mean both forces acting together.
Confirmation	A statement serving to confirm the truth or accuracy of a prior statement.	Right.
Correction	An statement serving to correct a prior statement.	Actually, -3.
Directive	An utterance in the form of an imperative.	Now draw the graph.
Explanation	An utterance in the form of an explanation.	Because there are no horizontal forces acting on it.
Expressive	An utterance in the form of an expressive.	Oh!
Hint	An utterance designed elicit another utterance by providing partial information.	Aren't you forgetting something?
Promise	An utterance that commits the speaker to a future action.	I will help you understand this.
Prompt	A utterance designed to elicit another utterance.	And then….
Question	An utterance in the form of a question	What are you having trouble understanding?
Reminder	An utterance in the form of a reminder.	Remember you need to subtract from both sides….
Request	An utterance in the form of request.	Could you help me?
Suggestion	An utterance in the form of a suggestion.	How about squaring both terms?

The dialog acts were defined and refined to minimize the overlap and maximize the coverage of distinct acts. The resulting taxonomy was described with examples and guidelines which were used by human annotators to tag the tutoring sessions to be used as training data in our models. The full taxonomy (dialog acts and their sub-acts) can be found in the appendix. A group of 20 experienced tutoring experts were consulted in the development of taxonomy and trained to tag the data set. The human tagging process included 4 major phases: development of taxonomy, 1st round tagging, reliability check, 2nd round tagging, reliability check, and final tagging phase.

The experts were divided into two groups: Taggers and Verifiers. In the first 2 tagging phases, each tagger was given a session transcript and asked to annotate the utterances. The resulting tagged session was then assigned to a verifier who went through the annotations, reviewed the tags and made necessary changes. In the reliability check steps, experts tagged each transcripts independently and in the final tagging phase, 5 experts went through all the tagged sessions and repeated the verification process to form a solid training set with the best quality of human tags.

Table 11. Agreements of Taggers (T) and Verifiers (V) on Top-level Speech Act (Act) and Sub-categories (SubAct).

		% Agreement		Kappa	
Phase	# Sessions	Act	SubAct	Act	SubAct
1st round	738	98	88	0.91	0.87
Reliability	39	79	63	0.75	0.62
2nd round	700	94	90	0.93	0.90
Reliability	36	81	66	0.77	0.64

Table 11 shows the agreement of experts on annotations prior to the final tagging round. The agreement of Taggers and Verifiers was approximately 90% with a slightly higher agreement on the second round which shows to what extent the verifiers made changes to the initial annotations. This results suggest a high agreement between experts' annotation while it is important to note that most of the disagreement was in the second level categories (subacts). Table 12 shows the distribution of speech act categories in the training data. It is observed that the highest frequent categories are Assertion and Expressive while other categories such as Hint and Promise have lowest frequency.

Table 12. Distribution of Speech Acts in the Training Data.

Act	Count	%
Answer	1130	1.2
Assertion	29890	32.3
Clarification	609	0.6
Confirmation	6620	7.1
Correction	2065	2.2
Directive	2006	2.1
Explanation	1941	2.0
Expressive	22198	24.0
Hint	341	0.3
Promise	303	0.3
Prompt	6186	6.6
Question	2553	2.7
Reminder	337	0.3
Request	14243	15.4
Suggestion	2028	2.1

4.1 Models

In supervised machine learning, the features used to represent the data play an important role in learning and the performance of the learned models. In previous studies, different kinds of features have been used ranging from the content, context, and domain specific properties of the dialogue sessions. In order to build the speech act classifier, we applied the following 3 kinds of featuresets.

- Simple features

Based on previous research (Rus et al., 2012), 3 leading tokens of an utterance were shown to be good predictors for speech act. Thus, we extracted the following features of each utterance: 1st token, 2nd token, 3rd token, last token, and length of utterance (i.e., number of tokens).

- Extended features

Using the Correlation Feature Selection (CFS) measure, we found that 1st and last token are the most predictive features and in order to add contextual information (features of prior utterances) we extended the simple features by adding the 1st and last token of three previous utterances to our feature set. CFS evaluates subsets of features with the assumption that the feature subsets which contain features that are highly correlated with the human classification are better than others.

- Conditional Random Fields features

One of learning algorithms that we applied was Conditional Random Fields (CRF) models on training data. Based on the nature of the CRF algorithm, we attempted to use more tokens as features and tried to cover the full content of utterances. In addition to the above features, to further investigate the CRF models we extracted the distribution of utterances length and found out the average length of the utterance is 12 tokens. Hence, another feature set that we used specifically for CRF models is the 20 first tokens of the utterance plus a context window of three previous utterances. Note that this size of 20 tokens covers most of the utterances (i.e., utterances that had 20 or less tokens) in full content while some had more tokens. If an utterance has less than 20 tokens then the respective features are set to a default value which represents blank tokens. This was an attempt to improve CRF models and since it increases our feature space dramatically it could not be efficiently applied in other algorithms.

The mentioned feature sets were used to create different models with multiple learning algorithms. One common property of the above feature sets is that they all represent the content of utterances by tokens. This enabled the models to capture specific properties of “chat” interactions where tokens are representative of certain functions whereas in spoken language it’s the words and acoustic features that play this role.

4.2 Algorithms

In order to learn the classification models from the training data, several learning algorithms can be applied. We applied supervised machine learning, i.e., the algorithms to learn a classifier from a set of tagged data. Four learning algorithms were used and evaluated: Naïve Bayes, Bayes Net, Logistic Regression, and Conditional Random Fields (CRF). Each of the algorithms has certain properties that take into account different characteristics of data.

- Naïve Bayes is the simplest learning algorithm for classification. It is based on the 'naïve' assumption that the features are independent. Based on the Bayes theorem, Naïve Bayes classifiers are learned by calculating a set of conditional probabilities which represent the probability of a class given a feature (e.g. $P(C_1 |f_1)$, $P(C_1 |f_2)$, etc.). The output of Naïve Bayes classifiers given the input 'I' is the probability of 'I' belonging to each category (in our case speech act category). The class with the highest probability is picked as the prediction. The predictions probabilities are also known as confidence. Naïve Bayes is a popular baseline classification algorithm because it can be efficiently learned in a supervised setting on small data sets, due to the features independence assumption.

- Bayes Net (Bayesian Network) is a probabilistic graphical model which represents the dependencies between features by a directed acyclic graph (DAG). In a Bayesian Networks nodes represent random variables (features) and edges represent conditional dependencies. Each node is associated with a probability function which given particular a set of values for the node's parent variables produces the probability of the variable represented by that node. Bayesian networks are also based on Bayes theorem but the difference between a Bayes Net and a Naïve Bayes model is that Bayes Nets represent the dependencies between features where as Naïve Bayes assumes that features are independent.

- Logistic Regression is another probabilistic model that is used in classification. A logistic regression model, uses logistic functions to model the prediction based on features. In the basic case, logistic regression is used to predict a binary class, however it can be modified to predict a multi-nominal variable by creating a regression function for each category and picking the prediction (category) based on the highest probability among all the functions' values. The fact that the logistic regression creates a separate model for each category makes it less efficient in cases with more number of categories and features. In our case, the size of taxonomy is large and the training data is also big which makes it less efficient to apply Logistic Regression, however in order to examine its performance, we applied this algorithm to our top-level classification model.

- Conditional Random Fields (CRF) is a probabilistic model designed mainly for tagging sequential data (Lafferty, McCallum, & Pereira, 2001). While the ordinary classification algorithms often label a single data point without taking the context into account, CRF models are appropriate for structural prediction, i.e., they use the neighbors (context) to make predictions and maintain the structural and sequential relationship between the labels. Since based on definition, several speech act categories (e.g. Answer) rely on the prior context, we applied CRF to investigate its performance along with the other common classification algorithms.

4.3 Results and Discussion

In order to examine the performance of different learning algorithms on our data set, three kinds of features (Simple, Extended, and CRF) were extracted from the full set of training data consisting of 95,526 utterances extracted from 1,438 annotated tutoring sessions and models were trained and tested using each feature set.

We used WEKA toolkit which is a Java package containing implementation of popular learning algorithms (Hall et al., 2009) and the CRF++ package to train and test the models. The size of our training data is larger than the data presented in most of the previous work on speech act classification and our data is within the domain of human one-on-one tutoring sessions which enables further analysis of the dialogue models to investigate the impact of dialogue moves on learning.

Extracting the simple features (leading tokens of utterance) we found approximately 2,500 distinct values for each of the features (out of ~95,000 utterances). The running time complexity of our algorithms is directly related to the number of features and data points. We tested out models using 10-fold cross validation.

In 10-fold cross validation, the available annotated data is split in 10 “folds;” 9 of the folds are used for training and one for testing. This process is repeated 10 times, once for each of the folds. The average performance measures across all iterations are reported.

As our taxonomy represents a hierarchical structure between the speech acts, we divided the speech acts into two categories: top-level speech acts and subcategories. This structure allows us to build models on different levels of speech act classification and ultimately design a hierarchical classifier which first tags an utterance with the top-level speech act and based on the top-level predict the appropriate sub-category. Based on the division of taxonomy in top-level and subcategories, we first trained and tested the models to predict the top-level speech act. Table 13 shows the results of 10-fold cross validation on the top-level classification models.

Table 13. 10-fold Cross Validation of Algorithms with Different Features for Top-level Speech Act Classification (P = precision, R = recall, F = F-measure)

Algorithm	FeatureSet	%Accuracy	Kappa	P	R	F
Naïve Bayes	Simple	72.5	0.65	0.71	0.72	0.70
Naïve Bayes	Extended	72.3	0.64	0.70	0.72	0.68
Bayes Net	Simple	72.6	0.65	0.71	0.72	0.71
Bayes Net	Extended	72.5	0.65	0.70	0.72	0.70
Logistic Regression	Simple	76.6	0.70	0.72	0.76	0.73
Logistic Regression	**Extended**	**77.4**	**0.71**	**0.70**	**0.77**	**0.73**
CRF	Simple	72.7	0.45	0.60	0.42	0.50
CRF	Extended	71.9	0.44	0.57	0.42	0.48
CRF	CRF	76.6	0.47	0.60	0.45	0.51

As seen in table 13, the best performance on top-level classification is achieved by Logistic Regression algorithm, however all the algorithms yield and accuracy of more than 70% which is the baseline accuracy.

The kappa values represent the extent to which each algorithm is performing better than chance and the logistic regression with extended feature set has the highest kappa which signifies the agreement of this algorithm with the expert tags.

It is interesting to note that the extended feature set does not improve the algorithms significantly which implies that adding the contextual information, i.e., prior utterances, is either not useful or not sufficiently representing the context. This behavior of contextual features have been previously shown in speech act classification models on a multi-party chat based tutoring system in chapter 3 and in Samei and colleagues (2013).

The top-level classification models provide reasonably accurate performance as the inter-rater agreement on top-level in the independent annotations yielded ~%70 which is comparable to the best models' accuracy. We further trained and tested models to classify utterances in the second level of speech act categories. The baseline models for the full classification (top-level and subcategories) is created by appending the subcategory to the top-level categories in our taxonomy. This results in a flat set of 133 categories.

The size of this flat taxonomy immediately limits the performance of classification models. As the taxonomy size increases more training data is needed to capture different characteristics of each category; however, the performance of models on this taxonomy illustrates a baseline accuracy which other models can be compared to.

The performance of models on the flat taxonomy are shown in the table 14. As it seen the accuracy of models is lower than the top-level classifiers. The drop in performance may be attributed to at least two factors: (1) the nearly ten-fold increase in the number of features (from 15 main speech acts to 133 subacts), leading to sparser data; and (2) the greater likelihood that the subtypes will be confused by human annotators. Human experts' agreements were also lower in the subcategory level which suggests the likelihood of confusion on that level.

Bayesian based algorithms (i.e., Naïve Bayes and Bayes Net) were applied to the flat taxonomy and the best performance was achieved by Bayes Net algorithm which is an extension to Naïve Bayes taking the features dependencies into account. This implies the importance of representing the dependencies between features to be able to make distinction between subcategories. The simple vs. extended features again doesn't show a significant impact on the performance.

Table 14. Results for Combinations of Models and Feature Sets (Dialogue Act Subtypes) (P=Precision R=Recall)

Algorithm	FeatureSet	%Accuracy	Kappa	P.	R.	F-measure
Naïve Bayes	Simple	51.9	0.49	0.52	0.51	0.47
Naïve Bayes	Extended	48.5	0.45	0.49	0.48	0.42
Bayes Net	**Simple**	**53.1**	**0.50**	**0.52**	**0.53**	**0.49**
Bayes Net	Extended	51.2	0.48	0.51	0.51	0.46

Next, we attempted to create a classifier for each set of subcategories. In other words, for each speech act a classifier was trained to predict its corresponding subcategories. In our data set, a set of utterances tagged with each speech act category formed the training data for learning its subcategories (subacts). Table 15 shows the performance of these classifiers which were trained on 70% and tested on 30% of the dataset.

Table 15. Performance of Subact Classifiers for each Speech Act Category.

Model	Accuracy	Kappa	Precision	Recall	F-measure
Answer	52.8	0.43	0.23	0.21	0.22
Assertion	57.67	0.42	0.47	0.35	0.4
Clarification	40.44	0.17	0.35	0.23	0.28
Confirmation	92.65	0.77	0.48	0.41	0.44
Correction	62.36	0.43	0.28	0.27	0.28
Directive	61.79	0.52	0.43	0.39	0.41
Explanation	54.47	0.25	0.21	0.19	0.2
Expressive	76.81	0.74	0.62	0.55	0.59
Hint	67.65	0.34	0.41	0.37	0.39
Promise	95.6	0	0.24	0.25	0.24
Prompt	64.28	0.3	0.51	0.42	0.46
Question	60.7	0.49	0.33	0.28	0.3
Reminder	47.79	0.25	0.42	0.39	0.4
Request	56.21	0.49	0.4	0.35	0.37
Suggestion	70.23	0.43	0.23	0.22	0.23

As shown in table 15, the subact classifiers yield an average accuracy of approximately %65 and kappa of 0.4. Particular subact categories such as the subcategories of Expressive are predicted with better accuracy and kappa (76% and 0.74), however one notable result is the lower precision and recall on the subact classifiers. This is due to the fact that some subcategories are too rare and there is not sufficient instances associated with them to let the models learn effectively.

Table 16. Subacts Predicted with Best Precision and Recall (precision > = 0.75)

Subact	**Precision**	**Recall**
AssertionURL	0.98	0.96
PromiseProcess	0.96	1
ExpressiveGreeting	0.95	0.98
ConfirmationPositive	0.95	0.97
ExpressiveLineCheck	0.88	0.77
QuestionAffect	0.86	0.93
ExpressiveNeutral	0.86	0.9
ConfirmationNegative	0.85	0.82
ExpressiveConfusion	0.84	0.58
DirectiveDialogControl	0.82	0.8
ExpressiveFarewell	0.8	0.81
ReminderProcess	0.8	0.62
ExpressivePraise	0.79	0.67
AnswerAffectPostive	0.77	0.95
RequestConfirmationUnderstanding	0.77	0.73
ExpressiveThanks	0.77	0.79
SuggestionProcess	0.76	0.83
CorrectionTypo	0.75	0.82

Table 17. Performance of Multi-layer Classifications.

Accuracy (%)	39.36
Kappa	0.36
Precision	0.19
Recall	0.15

To take a closer look at these classifiers, table 16 shows the subacts that were predicted with best precision and recall. Finally, we tested the final models, by first applying the speech act (top-level) classifier to the test set and based on the predicted speech act applied one of the subcategory models (presented in table 15) to predict the subact. Table 17 shows the results of this multi-layer classification.

As seen in table 17, the overall performance of the multi-layer classification models is significantly low. This is due to the fact the errors from top-level (speech act) are cascaded in the second level (subacts) besides the size of the taxonomy and lack of instances in some subact categories. The results imply the need to further investigate the design of the taxonomy as well as learning algorithms. However the performance of the models presented in this chapter are reasonable when applied separately and not in a multi-level approach.

4.4 Conclusion

The results of the different models and algorithms showed that the top-level speech acts can be predicted with a reasonable accuracy, however to be able to tag utterances with both top-level and subcategories a multi-level classification needs to be applied. In this work, we applied common classification algorithm such as, Naïve Bayes, Bayesian Networks, Logistic Regression, and Conditional Random Fields (CRF). Each of the mentioned algorithms has their particular properties that make them suitable for certain problems.

Naïve Bayes is the simplest classifier which is based on the Bayes theorem and works with conditional probabilities to learn the relation between each feature and the prediction. The naïve assumption in the Naïve Bayes algorithm is the independence of features. The Bayesian Network algorithm extends Naïve Bayes by taking into account the dependencies between features and while both Naïve Bayes and Bayesian

Networks are based on probabilistic modeling and Bayes theorem, the difference in the features dependency assumption makes them suitable for different problems.

We also applied CRF algorithm which is suitable for sequential classification. Adding the contextual information to the models didn't show a significant impact thus by applying an algorithm which is designed for sequential context-based classification (CRF) we further investigated the context-based design of speech act classification, however CRF models did not perform better than other algorithms such as Logistic Regression.

Another algorithm that we applied was Logistic Regression. Logistic Regression when adopted for multi-nominal classification creates a separate classifier for each class (category). In our experiments, Logistic Regression achieved the best performance on the top-level classification. This implies that the models will perform better if each category of speech act is learned separately which means to some extent resolves the confusion between different categories. This finding supports the hierarchical structure of the taxonomy which in a way forces the classification model to learn each class separately.

The hierarchical and granular taxonomy enables modeling the dialogue in a more precise way, i.e., identifying patterns and strategies used in tutoring sessions. Creating speech act classifiers becomes handy when the cost of human annotation is high. In our case, the data set that was annotated by human experts, and represents a small portion of a larger data set of available transcripts of tutoring sessions. The ultimate goal of this work is to build a model to be applied to a set of not-seen and untagged data and use the speech acts as means of modeling the discourse.

Learning a model on a set of tagged data and then applying it to a larger set of untagged instances is a known problem often approached as semi-supervised learning.

The semi-supervised framework enables the models to be updated and retrained as more untagged data get tagged. The proposed models in this chapter can be used as initial models for a semi-supervised classifier to which ultimately will identify speech acts in real time.

5 Future Work

In the thesis, we investigated speech act classification models on two datasets and with different taxonomies. The certain properties of a dataset and taxonomy have an impact on the performance of the models. While the baseline performance of the proposed models is reasonable, there are several directions for future work to improve the models.

All in all, the future directions to the proposed approach is to design a framework specifically appropriate for speech act classification models from the taxonomy structure to the learning process that can be tested in separate settings. The algorithms and models presented in this thesis are general classification algorithms which are designed to work in the generic framework of classification problems. In chapter 4 we used a hierarchical taxonomy and investigated models to predict the top-level speech acts as well as subacts while in chapter 3 we also modified the taxonomy to multi layers.

In order to develop classifiers with the hierarchical structure of taxonomy, we will investigate modifications to the algorithms to learn in a hierarchical way while avoiding cascaded errors in the second level. One approach can be Classifier Ensemble in which a set of classifiers will be applied to an instance and the set of predictions will be generated based on which we can examine an algorithm to identify the final prediction.

Another direction for future work is to examine the taxonomies and take a bottom-up data-driven approach to find the best structure for the taxonomy. We can apply unsupervised clustering techniques to form a set of speech acts and within each set find subcategories.

Moving the analysis in both top-down and bottom-up approach to find a common ground where the models do their best is another direction to take for future work.

Since the cost of human tagging is high, we will investigate more models with unsupervised and semi-supervised techniques such as self-training and co-training. In a semi-supervised approach we initially train classifiers on tagged data and then apply the learned models to a set of untagged instances and the predictions with high confidence are added to the training set and update the models. The models proposed in previous chapters can be used as initial classifiers and updated with more data in a semi-supervised approach.

The work presented in this thesis will be used as baseline approach to future analysis. With different general approaches to classification problems we will attempt to create a specific classification algorithm to take advantage of taxonomy structure and nature of speech act classification which is a base component of intelligent systems with natural language processing and language understanding.

References

Abney, S. (2007). *Semisupervised Learning for Computational Linguistics* (1st ed.). Chapman & Hall/CRC.

Ashok, V., Borodin, Y., Stoyanchev, S., & Ramakrishnan, I. (2014). Dialogue Act Modeling for Non-Visual Web Access. *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)* (pp. 123-132). Philadelphia, PA, U.S.A.: Association for Computational Linguistics.

Austin, J. L. (1962). *How to do things with words.* Oxford.

Bangalore, S., Di Fabbrizio, G., & Stent, A. (2006). Learning the Structure of Task-driven Human-human Dialogs. *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics* (pp. 201-208). Stroudsburg, PA, USA: Association for Computational Linguistics.

Crook, N., Granell, R., & Pulman, S. (2009). Unsupervised Classification of Dialogue Acts using a {D}irichlet Process Mixture Model. *Proceedings of the SIGDIAL 2009 Conference* (p. 341-348). London, UK: Association for Computational Linguistics.

D'Andrade, R. G., & Wish, M. (1985). Speech act theory in quantitative research on interpersonal behavior? *Discourse Processes, 8*(2), 229-259.

Ezen-Can, A., & Boyer, K. (2014). Combining Task and Dialogue Streams in Unsupervised Dialogue Act Models. *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)* (pp. 113-122). Philadelphia, PA, U.S.A.: Association for Computational Linguistics

Ferschke, O., Gurevych, I., & Chebotar, Y. (2012). Behind the Article: Recognizing Dialog Acts in Wikipedia Talk Pages. *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 777-786). Stroudsburg, PA, USA: Association for Computational Linguistics.

Forsyth, E. N., & Martell, C. H. (2007). Lexical and discourse analysis of online chat dialog. *Semantic Computing, 2007. ICSC 2007. International Conference on*, (pp. 19-26).

Graesser, A. C., & Person, N. K. (1994). Question asking during tutoring. *American educational research journal, 31*(1), 104-137.

Ha, E. Y., Grafsgaard, J. F., Mitchell, C. M., Boyer, K. E., & Lester, J. C. (2012). Combining Verbal and Nonverbal Features to Overcome the 'Information Gap' in Task-oriented Dialogue. *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue* (pp. 247-256). Stroudsburg, PA, USA: Association for Computational Linguistics.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA Data Mining Software: An Update. *SIGKDD Explor. Newsl., 11*(1), 10-18.

Joty, S., Carenini, G., & Lin, C.-Y. (2011). Unsupervised Modeling of Dialog Acts in Asynchronous Conversations. *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Three* (pp. 1807-1813). AAAI Press.

Jurafsky, D., Shriberg, E., Fox, B., & Curl, T. (1998). Lexical, Prosodic, and Syntactic Cues for Dialog Acts. *Discourse Relations and Discourse Markers,* (pp. 114-120).

Kim, S. N., Cavedon, L., & Baldwin, T. (2010). Classifying Dialogue Acts in One-on-one Live Chats. *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing* (pp. 862-871). Stroudsburg, PA, USA: Association for Computational Linguistics.

Lafferty, J. D., McCallum, A., & Pereira, F. C. (2001). Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. *Proceedings of the Eighteenth International Conference on Machine Learning* (pp. 282-289). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Moldovan, C., Rus, V., & Graesser, A. C. (2011). Automated Speech Act Classification For Online Chat. *MAICS,* (pp. 23-29).

Olney, A., Louwerse, M., Matthews, E., Marineau, J., Hite-Mitchell, H., & Graesser, A. (2003). Utterance classification in AutoTutor. *Proceedings of the HLT-NAACL 03 workshop on Building educational applications using natural language processing-Volume 2*, (pp. 1-8).

Power, R. (1979). The organization of purposeful dialogs. *Linguistics, 17,* 105-152.

Rangarajan Sridhar, V. K., Bangalore, S., & Narayanan, S. (2009). Combining Lexical, Syntactic and Prosodic Cues for Improved Online Dialog Act Tagging. *Comput. Speech Lang., 23*(4), 407-422.

Rasor, T., Olney, A., & D'Mello, S. K. (2011). Student Speech Act Classification Using Machine Learning. *FLAIRS Conference.*

Ritter, A., Cherry, C., & Dolan, B. (2010). Unsupervised Modeling of Twitter Conversations. *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 172-180). Stroudsburg, PA, USA: Association for Computational Linguistics.

Rus, V., D’Mello, S., Hu, X., & Graesser, A. (2013). Recent Advances in Conversational Intelligent Tutoring Systems. *AI Magazine, 34*(3), 42-54.

Rus, V., Moldovan, C., Niraula, N., & Graesser, A. C. (2012). Automated Discovery of Speech Act Categories in Educational Games. *International Educational Data Mining Society*.

Russell, S. J., & Norvig, P. (2003). *Artificial Intelligence: A Modern Approach* (2 ed.). Pearson Education.

Samei, B., Li, H., Keshtkar, F., Rus, V., & Graesser, A. C. (2014). Context-Based Speech Act Classification in Intelligent Tutoring Systems. *Intelligent Tutoring Systems - 12th International Conference, {ITS} 2014, Honolulu, HI, USA, June 5-9, 2014. Proceedings*, (pp. 236-241).

Schegloff, E. A. (1968). Sequencing in Conversational Openings. *American Anthropologist, 70,* 1075-1095.

Searle, J. R. (1969). *Speech acts: An essay in the philosophy of language* (Vol. 626). Cambridge university press.

Searle, J. R. (1976). A classification of illocutionary acts. *Language in society, 5*(01), 1-23.

Shaffer, D. W., & Gee, J. P. (2007). Epistemic games as education for innovation. In *BJEP Monograph Series II, Number 5-Learning through Digital Technologies* (Vol. 71, pp. 71-82). British Psychological Society.

Stolcke, A., Coccaro, N., Bates, R., Taylor, P., Van Ess-Dykema, C., Ries, K., . . . Meteer, M. (2000). Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech. *Computational Linguistics, 26*(3), 339-373.

Tavafi, M., Mehdad, Y., Joty, S., Carenini, G., & Ng, R. (2013). Dialogue Act Recognition in Synchronous and Asynchronous Conversations. *Proceedings of the SIGDIAL 2013 Conference* (pp. 117-121). Metz, France: Association for Computational Linguistics.

Appendix: Speech Act Taxonomy

Top-level	Sub-Category
Answer	
	Approach
	Affect:Postive
	Affect:Negative
	Affect:Neutral
	Calculation
	Concept
	InstructionalContext
	Metacognition
	Process
	PriorKnowledge
	PriorKnowledge:Negative
	PriorKnowledge:Positive
	ProblemStatement
	Understanding
	Understanding:Negative
	Understanding:Positive
	Unspecified
Assertion	
	Affect
	Approach
	Calculation
	Concept
	InstructionalContext
	Metacognition
	Process
	ProblemStatement
	Prior Knowledge:Positive
	Prior Knowledge:Negative
	Understanding:Positive
	Understanding:Negative
	URL
	Unspecified
Correction	
	Approach
	Calculation
	Concept
	ProblemStatement
	Process
	Typo
	Unspecified

Clarification	
	Approach
	Calculation
	Concept
	InstructionalContext
	Metacognition
	Process
	ProblemStatement
	Unspecified
Confirmation	
	Positive
	Neutral
	Negative
	Unspecified
Continuation	
	Continuation
Directive	
	Approach
	Attention
	Calculation
	Concept
	DialogControl
	InstructionalContext
	Metacognition
	Process
	Unspecified
Expressive	
	Acknowledgment
	Apology
	Confirmation:Positive
	Confirmation:Negative
	Confusion
	Celebration
	Farewell
	Greeting
	InstructionalContext
	Laugh/Smile
	LineCheck
	Metacognition
	Mistake
	Negative
	Neutral
	Praise
	Positive
	Thanks

	Understanding
	Unspecified
Explanation	
	Approach
	Calculation
	Concept
	InstructionalContext
	Metacognition
	Process
	ProblemStatement
	Unspecified
Hint	
	Approach
	Calculation
	Concept
	Unspecified
Promise	
	Calculation
	InstructionalContext
	Metacognition
	Process
	Unspecified
Prompt	
	Approach
	Calculation
	Concept
	Process
	Unspecified
Question	
	Affect
	Approach
	Calculation
	Concept
	InstructionalContext
	Metacognition
	Process
	Prior Knowledge
	ProblemStatement
	Understanding
	Unspecified
Reminder	
	Approach
	Calculation
	Concept
	InstructionalContext

	Metacognition
	Process
	Unspecified
Request	
	Confirmation
	Confirmation:PriorKnowledge
	Confirmation:Approach
	Confirmation:Calculation
	Confirmation:Process
	Confirmation:ProblemStatement
	Confirmation:InstructionalContext
	Confirmation:Metacognition
	Confirmation:Concept
	Confirmation:Understanding
	Clarification
	DialogControl
	Explanation
	Process
	Unspecified
Suggestion	
	Approach
	Attention
	Calculation
	Concept
	InstructionalContext
	Metacognition
	Process
	Unspecified
Unspecified	