Automated Speech Act Classification For Online Chat

Cristian Moldovan and **Vasile Rus**
Department of Computer Science
Institute for Intelligent Systems
The University of Memphis
Memphis, TN 38152, USA
cmldovan|vrus@memphis.edu

Arthur C. Graesser
Department of Psychology
Institute for Intelligent Systems
Institute for Intelligent Systems
The University of Memphis
art.graesser@gmail.com

Abstract

In this paper, we present our investigation on using supervised machine learning methods to automatically classify online chat posts into speech act categories, which are semantic categories indicating speakers' intentions. Supervised machine learning methods presuppose the existence of annotated training data based on which machine learning algorithms can be used to learn the parameters of some model that was proposed to solve the task at hand. In our case, we used the annotated Linguistic Data Consortium chat corpus to tune our model which is based on the assumption that the first few tokens/words in each chat post are very predictive of the post's speech act category. We present results for predicting the speech act category of chat posts that were obtained using two machine learning algorithms, Naïve Bayes and Decision Trees, in conjunction with several variants of the basic model that include the first 2 to 6 words and their part-of-speech tags as features. The results support the validity of our initial assumption that the first words in an utterance can be used to predict its speech act category with very good accuracy.

Introduction

The task of speech act classification involves classifying a discourse contribution, e.g. an utterance, into a speech act category selected from a set of predefined categories that fulfill particular social discourse functions. Examples of speech act categories 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 following Statement, *The situation is getting worse every day.* or greets someone as in *Hello!* .

In this paper, we propose an automated method to classify online chat posts into speech act categories. The proposed automated method relies on a model that emphasizes the use of the first tokens or words in an utterance to decide their speech act category. For instance, a Question can be distinguished from a Statement based on the first words because usually a Question starts with question words such as *How* which is followed by an auxiliary verb such as *did*. Our model is based on the assumption that humans do infer speakers' intentions early on when they hear the first few words of an utterance. To automate the process, we framed our problem as a supervised machine learning problem in which we map the previously described model into a set of features and then use machine learning algorithms to learn the parameters of the model from annotated training data. We test in this paper this hypothesis and report how well the first 2 to 6 words of an utterance can diagnose its speech act. The tuned models are then evaluated on separate test data sets. In particular, we work with online chat conversations in which participants in online chatrooms converse with each other via computer networks. Each online chatroom participant can see everyone else's dialogue turns, or chat posts, and respond.

The rest of the paper is organized as in the followings. The next section presents theoretical background on speech acts as well as an overview of various speech act taxonomies. The *Approach* section offers the details of our approach. The following section describes related work addressing the task of speech act classification in similar contexts, e.g. online chats. The *Experiments and Results* section provides a summary of the experiments and results. The *Conclusions* section ends the paper.

Language As Action - Speech Acts

Speech act theory has been developed based on the language as action assumption which states that when people say something they do something. Speech act is a term in linguistics and the philosophy of language referring to the way natural language performs actions in human-to-human language interactions, such as dialogues. Its contemporary use goes back to John L. Austin's theory of *locutionary*, *illocutionary* and *perlocutionary acts* (Austin 1962). According to Searle (Searle 1969), there are three levels of action carried by language in parallel: first, there is the locutionary act which consists of the actual utterance and its exterior meaning; then, there is the illocutionary act, which is the real intended meaning of the utterance, its semantic force; finally, there is the perlocutionary act which is the actual effect of the utterance, such as scaring, persuading, encouraging, etc.

It is interesting to notice that the locutionary act is a feature of any kind of language, not only natural ones, and that it does not depend on the existence of any actor. In contrast, an illocutionary act needs the existence of an environment outside language and an actor that possesses intentions, in other words an entity that uses language for acting in the outside environment. Finally, a perlocutionary act needs the belief of the first agent in the existence of a second entity and the possibility of a successful communication attempt: the effect of language on the second entity, whether the intended one or not, is taking place in the environment outside language, for which language exists as a communication medium. As opposed to the locutionary act, the illocutionary and perlocutionary acts do not exist in purely descriptive languages (like chemical formulas), nor in languages built mainly for functional purposes (like programming languages). They are an indispensable feature of natural language but they are also present in languages built for communication purposes, like the languages of signs or the conventions of warning signals.

In a few words, the locutionary act is the act *of* saying something, the illocutionary act is an act performed *in* saying something, and the perlocutionary act is an act performed *by* saying something. For example, the phrase *"Don't go into the water"* might be interpreted at the three act levels in the following way: the locutionary level is the utterance itself, the morphologically and syntactically correct usage of a sequence of words; the illocutionary level is the act of warning about the possible dangers of going into the water; finally, the perlocutionary level is the actual persuasion, if any, performed on the hearers of the message, to not go into the water. In a similar way, the utterance *"By the way, I have a peanut butter sandwich with me; would you like to have a bite?"* can be decomposed into the three act levels. The locutionary act is the actual expressing of the utterance, the illocutionary act is the offer implied by the phrase, while the perlocutionary act, namely the intended effect on the interlocutor, might be impressing with own selflessness, creating a gesture of friendliness, or encouraging an activity, in this case eating.

The notion of speech act is closely linked to the illocutionary level of language. The idea of an illocutionary act can be best captured by emphasizing that "by saying something, we do something" (Austin 1962). Usual illocutionary acts are: greeting (*"Hello, John!"*), describing (*"It's snowing."*), asking questions (*"Is it snowing?"*), making requests (*"Could you pass the salt?"*), giving an order (*"Drop your weapon!"*), making a warning (*"The floor is wet!"*), or making a promise (*"I'll return it on time."*). The illocutionary force is not always obvious and could also be composed of different components. As an example, the phrase *"It's cold in this room!"* might be interpreted as having the intention of simply describing the room, or criticizing someone for not keeping the room warm, or requesting someone to close the window, or a combination of the above. A speech act could be described as the sum of the illocutionary forces carried by an utterance. It is worth mentioning that within one utterance, speech acts can be hierarchical, hence the existence of a division between direct and indirect speech acts, the latter being those by which one says more than what is literally said, in other words, the deeper level of intentional meaning. In the phrase *"Would you mind passing me the salt?"*, the direct speech act is the request best described by *"Are you willing to do that for me?"* while the indirect speech act is the request *"I need you to give me the salt."* In a similar way, in the phrase *"Bill and Wendy lost a lot of weight with a diet and daily exercise."* the direct speech act is the actual statement of what happened *"They did this by doing that."*, while the indirect speech act could be the encouraging *"If you do the same, you could lose a lot of weight too."*

In our work presented here, we assume there is one speech act per utterance and the set of speech acts used are all at the same level of depthness forming a flat hierarchy. These simplification assumptions are appropriate for a first attempt at automating the speech act classification process and testing our leading tokens model. Furthermore, the LDC data set imposed further constraints on our experiments as the LDC corpus does assume only one speech act per chat posts and also uses a flat set of speech act categories.

Speech Act Taxonomies

The task of speech act classification, the focus of our paper, requires the existence of a predefined set of speech act categories or speech act taxonomy. Researchers have proposed various speech act taxonomies over the years. We present next a summary of the most important important ones as judged from historical and relevance perspectives.

The classic categorization of (Austin 1962) postulates five major speech act classes based on five categories of performative verbs: Expositives - verbs asserting or expounding views, clasifying 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 (D'Andrade and Wish 1985).

The taxonomy proposed by (Searle 1969) 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 Declarations - giving an authoritative decision about some fact.

The category scheme proposed by (D'Andrade and Wish 1985) treats most utterances as conveying more than a 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 particu-

Table 1: Literature Speech Act Taxonomies

Name	Main Classes
Austin	Expositives, Exercitives, Verdictives, Commissives, Behabitives
Searle	Representatives, Directives, Commissives, Expressives, Declarations, Representative Declarations
D'Andrade and Wish	Expositives, Interrogatives, Exercitives, Reactions, Verdictives, Commissives, Behabitives
VerbMobil	Request, Suggest, Convention, Inform, Feedback

lar problem ultimately rests on the task at hand. For instance (Olney, et al. 2003) uses a taxonomy that divides questions into 16 subcategories and has only 3 classes for the rest of the utterances, which is suitable for an Intelligent Tutoring environment. The 16 subclasses of Questions are: 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 longterm interdisciplinary Language Technology research project with the aim 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 ramify into subclasses. For instance, the Convention class is composed of the following subclasses: Thank, Deliberate, Introduce, Politeness Formula and Greeting. (Alexandersson, et al. 1997)

A summary of the theoretical speech act taxonomies and the Verbmobil taxonomy mentioned above are presented in Table 1. In our work, we will use the LDC set of speech act categories, which are described later.

The Approach

As we already mentioned, we adopted a supervised machine learning method to automate the process of speech act classification. Machine learning methods imply the design of a feature set which can then be used together with various machine learning algorithms. We used two such algorithms, Naïve Bayes and Decision Trees, to learn the parameters of the basic model and induce classifiers that can categorize new utterances into speech act categories. Naïve Bayes are statistical classifiers that make the naïve assumption of feature independence. While this assumption means models that are too simplistic at times, it helps with better estimating the parameters of the model which in turn leads to good classifiers in general. Decision Trees are based on the idea of organizing the features in a hierarchical decision tree based on information gain. More informative features are always higher in the tree.

In the automated speech act classification literature, researchers have considered rich feature sets that include the actual words (possibly lemmatized or stemmed) and n-grams (sequences of consecutive words). In almost every such case, researchers apply feature selection methods because considering all the words might lead to overfitting and, in the case of n-grams, to data sparseness problems because of the exponential increase in the number of features. Besides the computational challenges posed by such feature-rich methods, it is not clear whether there is need for so many features to solve the problem of speech act classification.

We believe that humans infer speakers' intention after hearing only few of the leading words of an utterance. One argument in favor of this assumption is the evidence that hearers start responding immediately (within milliseconds) or sometimes before speakers finish their utterances ((Jurafsky and Martin 2009) - pp.814). This paper is a first step towards exploring the validity of such a hypothesis within the context of automated speech act classification of online chat posts.

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 Answers, Accepting, or Rejecting, contain a semantic equivalent of *yes* or *no* among the first words, and Greetings use a relatively small bag of words and expressions. 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. We ignored such intonational and contextual clues so far in our work in order to explor the potential of classifying speech acts based on words alone. We do plan to incorporate contextual clues in future experiments.

A key decision when developing methods to classify speech acts is choosing the speech act taxonomy. In our work presented in this paper, we adopted the taxonomy proposed by the developers of the LDC chat corpus (Forsyth and Martell 2007). The taxonomy is presented in Table 2. We will use chat posts and their speech acts from the LDC corpus to illustrate the basic idea of our leading tokens approach. We picked examples of posts labeled as Yes/No Questions from the corpus. Selecting the first few words as features seems to be a good approach after seeing the following 12 randomly selected instances of the Yes/No Questions class: *"is 10-19-20sUser68 back yet"*, *"Any women from*

Table 2: Speech act taxonomy and frequencies in the LDC online chat corpus

Classification	Percent	Example
Statement	34.50%	10-19-40sUser11...some people have a lot of blank pages
System	17.02%	JOIN
Greet	13.40%	Hey You
Emotion	11.52%	lmao
Wh-Question	5.33%	where from@11-09-adultsUser12
Yes/No Question	5.22%	wisconsin?
Continuer	3.48%	but i didnt chance it
Accept	2.45%	ok
Reject	2.14%	I can't do newspaper.. I can't throw that far and stairs give me problems
Bye	1.57%	goodnite
Yes Answer	1.17%	yeah
No Answer	0.94%	nope 11-09-adultsUser27
Emphasis	0.48%	Ok I'm gonna put it up ONE MORE TIME 10-19-30sUser37
Other	0.43%	0
Clarify	0.34%	i mean the pepper steak lol

Nashville in here?", *"are you a male?"*, *"hey any guys with cams wanna play?"*, *"any guyz wanna chat"*, *"any single white females?"*, *"r u serious"*, *"can't sleep huh?"*, *"really?"*, *"any girls wanna chat with 24/m"*, *"22/m/wa any ladies want to chat"*, *"can i talk to him!!"*. The word *"any"* seems to appear often and so are the forms of the auxiliary verb *"to be"* and modal verbs. It would also seem very useful to use a lemmatizer or stemmer, that map morphological variations of the same word to a canonical form, adapted to the specific environment of online chat. For instance, we would like to automatically decide that *"guyz"* is the same as *"guys"* and that the words *"r"* and *"u"* may in fact be an abbreviation for *"are you"*. Without this additional knowledge many resemblences would be lost. Also, the post *"really?"* has no common feature with the others, except for the question mark.

Some other speech act classes are even more suitable to this approach. As an example, we will provide 12 randomly selected instances labeled as Yes Answer in the same corpus: *"yes 10-19-20sUser30"*, *"sure 10-19-20sUser126"*, *"yes 10-19-20sUser115!!!!"*, *"yes"*, *"yep"*, *"yes...."*, *"yes i sleep"*, *"yeah..."*, *"U are Yes"*, *"Yes i would 10-19-30sUser12"*, *"yep....cool...kool..."*, *"yep..."*. The word *"yes"*, usually on the first position in the post, is a powerful common feature, as well as the relatively short length of the posts. A common feature is also the usage of pronouns, especially *"I"*. However, without knowing that *"yes"*, *"yep"* and *"yeah"* are variants of the same word, any automated classification method would lose a significant amount of accuracy.

A previous attempt by (Marineau, et al. 2000) explored classification using the first three words of each utterance. We extended the range and used from the first two words up to the first six words of each post. Using more words does provide more information and thus an easier way to differentiate between classes. However, due to the nature of the corpus we used, sometimes considering too many words is a disadvantage, because many posts are only one or two words long and labeling the missing positions with a *none* tag means encouraging a classifier to find common features between all short utterances, regardless of their different words. We must introduce artificial values such as *none* for missing positions in short posts in order to generate values for all the six features, for instance, in models where we use the first 6 words in chat posts to predict the speech acts. We only used the first 6 leading words as the average length in our LDC corpus was 4.67 words meaning models with 6 words should use up all the words in the posts, on average, to make predictions.

Related Work

Forsyth and Martell (Forsyth and Martell 2007) developed a speech act classifier on the LDC corpus, using the taxonomy of (Wu, Khan, Fisher, Shuler and Pottenger 2005). The corpus consisted of online chat sessions in English between speakers of different ages. Their prediction model relied on a set of 22 features that include: the number of chat posts ago the user last posted something, number of chat posts in the future that contain a yes/no pattern, total number of users currently logged on, the number of posts ago that a post was a JOIN (System message), total number of tokens in post, first token in post contains *"hello"* or variants, first token in post contains conjunctions such as *"and"*, *"but"*, *"or"*, etc., number of tokens in the post containing one or more *"?"* and number of tokens in the post in all caps. The values for all the features were normalized. The first 9 features were based on the distance of the post to specific posts around it, while the rest of the features were based on the density of some key words in the post or in the first token of the post belonging to a specific speech act category. The machine learning algorithms they used were Backpropagation Neural Network and Naïve Bayes, with the former performing better. Neither method seemed to make a reasonable classification unless the frequency of the class was higher than 3%.

Obviously, in the classification system of Forsyth and

Table 3: 10-fold cross-validation on LDC online chat corpus

n	Naïve Bayes Accuracy	Kappa	Precision	Recall	F-Measure	Decision Trees Accuracy	Kappa	Precision	Recall	F-Measure
2	74.14	.676	.719	.741	.714	78.33	.727	.772	.783	.772
3	73.05	.662	.698	.731	.697	78.35	.727	.772	.784	.772
4	72.57	.656	.690	.726	.690	77.27	.711	.755	.773	.746
5	72.17	.651	.671	.722	.683	77.27	.711	.755	.773	.746
6	71.70	.645	.662	.717	.677	77.32	.711	.755	.773	.746

Table 4: 10-fold cross-validation on LDC online chat corpus without "System" posts

n	Naïve Bayes Accuracy	Kappa	Precision	Recall	F-Measure	Decision Trees Accuracy	Kappa	Precision	Recall	F-Measure
2	66.64	.558	.646	.666	.634	71.80	.622	.702	.718	.702
3	65.28	.543	.627	.653	.615	71.87	.623	.704	.719	.703
4	64.46	.533	.618	.645	.604	71.82	.622	.702	.718	.703
5	64.03	.527	.605	.640	.598	71.77	.621	.701	.718	.702
6	63.51	.520	.585	.635	.591	71.82	.622	.702	.718	.703

Martell, the order of posts in the chat and automatic system messages (like JOIN or PART) played a major role. As far as syntactical information is concerned, they started from the assumption that the first word of a post is very important for determining the speech act of the post, especially in the case of Wh-Question, Yes/No Question, Continuer, Yes Answer and No Answer. Also, the question mark and the exclamation mark were considered indicative.

In order to automatically classify speech acts, (Samuel, Carberry and Vijay-Shanker 1998) applied a Transformation-Based Learning machine learning algorithm on Reithinger and Klessen's training set (143 dialogues, 2701 utterances) and on a disjoint testing set (20 dialogues, 328 utterances) (Reithinger and Klesen 1997). The features investigated were punctuation marks, speaker direction (provided by the corpus), number of words in utterances, speech acts of previous and following utterances, and a feature called dialogue act cues. The latter is finding the n-grams for n = 1,2,3 that minimize the entropy of the distribution of speech acts in a training corpus. Other processing steps they used included filtering out irrelevant dialogue act cues and clustering semantically-related words. The results showed a comparison between features: manually selected cue phrases, word n-grams, and entropy-minimization cues, all combined with the additional processing steps. The best results were obtained using entropy minimization with filtering and clustering.

Experiments and Results

The LDC online chat corpus is a product of the Naval Postgraduate School (Lin 2007). It contains 10,567 posts from different online chat rooms in English. All the posts had to go through a sanitizing process in order to protect user privacy, so that the corpus could be made available to the larger research community. All the user screen names were replaced with a mask, for example *killerBlonde51* was replaced by *101930sUser112*.

The original motivation for the development of the corpus was an attempt to automatically determine the age and gender of the poster based on their chat style, using features like average number of words per post, vocabulary breadth, use of emoticons and punctuation. Subsequently, the corpus was manually annotated with part of speech labels for each word and a speech act category per post. An automatic speech act classifier was used for this purpose and then each post was manually verified (Forsyth and Martell 2007). We take advantage in our experiments of the part of speech information available in the LDC corpus by incorporating this information in our basic model. We report results with and without part of speech information, which was included in the basic model in the form of part of speech tags for each word considered in a particular instance of the model.

The part of speech (POS) tagging used the Penn Treebank tagset with some additions specific to the problems related to a chat corpus. Abbreviations such as *"lol"* and emoticons such as *":)"* are frequently encountered and since they all convey emotion they were treated as individual tokens and tagged as interjections (*"UH"*). Also, some words that would normally be considered misspelled and were practically standard online were treated as correctly spelled words and tagged according to the closest corresponding word class. For example, the word *"wont"* if treated as a misspelling would normally be tagged as *"^MD^RB"*, the character ^ referring to a misspelling. The same word would be tagged as *"MD"* and *"RB"* when referring to *"modal"* and *"adverb"*, respectively. However, since it was highly frequent in the chat domain, *"wont"* was tagged as *"MD"*. In contrast, words that were just plain misspelled and did not

Table 5: 10-fold cross-validation on LDC online chat corpus without "System" posts and without POS tags

	Naïve Bayes					**Decision Trees**				
n	Accuracy	Kappa	Precision	Recall	F-Measure	Accuracy	Kappa	Precision	Recall	F-Measure
2	69.40	.574	.641	.694	.641	71.79	.622	.701	.718	.703
3	66.88	.546	.632	.669	.613	71.82	.622	.703	.718	.703
4	65.74	.532	.618	.657	.598	71.85	.623	.703	.719	.703
5	64.57	.517	.594	.646	.584	71.80	.623	.702	.718	.703
6	63.89	.507	.587	.639	.576	71.78	.622	.701	.718	.702

appear frequently were tagged with the misspelled version of the tag, for example the word *"intersting"* was tagged as *"^JJ"* (Forsyth and Martell 2007).

In the LDC corpus, each post was assigned a single speech act category from the 15 categories of the chat taxonomy proposed by (Wu, Khan, Fisher, Shuler and Pottenger 2005). Those categories along with examples and their frequencies in the corpus are represented in Table 2.

In order to implement the machine learning approach, we extracted for each of the 10,567 posts the first n tokens ($n = 2..6$) and their part of speech (POS) tags. Furthermore, we recorded the annotated speech act category as the correct class of the post, which is needed during training. We then use Naïve Bayes and Decision Trees (J48) from WEKA (Witten and Frank 2005) to induce classifiers based on the leading n tokens and their POS tags. We experimented with several variants of the basic model by generating an instance for each $n = 2..6$. The accuracy of the induced classifiers was measured using 10-fold cross-validation. In the cases in which the post had less than n tokens, we replaced the empty feature slots with a dummy token and a dummy POS tag. A summary of results is shown in Table 3.

The System class clearly increases the performance of the classifier due to the large number of instances and their simplicity. Practically, the System posts are "PART", "JOIN", and just a few other variants. In a second round of experiments, we wanted to investigate the validity of our approach on real posts only, i.e. we did not take into account the System posts. As already mentioned, the actual System messages are too specific to a particular chat system and they are not natural language. As before, we extracted for each post the first n tokens and the speech act category. On the remaining 7,935 posts, we applied the same Naïve Bayes and Decision Trees (J48) classifiers with 10-fold cross-validation. The results are presented in Table 4. A significant drop in accuracy can be noticed. Still, the performance of the proposed approach is very good on the natural, non-System posts. We also observe that there is no major difference among the various models when the number of leading words is vare. This may be explain in the case of chat posts by the relative short nature of these posts.

In Table 5, we provide the results obtained by the induced classifiers without System posts and without parts of speech tags for the words. These results reveal the power of our basic model alone (without part of speech information) on natural posts (System posts are not included). An interesting finding is the fact that best results using Naïve Bayes are obtained when using only the first two leading words in a chat post instead of more. When using Decision Trees, results obtained with the first two leading words are as good as when using even 6 words. Thus, we can conclude that our hypothesis that the first few leading words of an utterance are very diagnostic of that utterances' speech act is true, at least for online chat posts, the focus of our experiments.

Conclusions

Our results acknowledge the fact that the first few tokens of a chat post are indicative of the speech act of the post. It is worth mentioning the chat language could be considered more challenging than natural conversation language. Indeed, online chat being an environment that apparently encourages extreme creativity and exhibits a very high tolerance to misspellings and breaking language rules. For instance, "*hey, heya, heyheyhey, heys, heyy, heyyy, heyyyy, heyyyyy, heyyyyyy, heyyyyyyy, heyyyyyyyy, heyyyyyyyyy, heyyyyyyyyyy, heyyyyyyyyyyyyyy*" are in the LDC corpus chat-specific variants of the same greeting word. To this point, we did not use a lemmatizer-like tool that could reduce the high number of variants for the same token, especially in the case of speech acts with a high emotional content, such as rejections, accepting, emotion and emphasis, and also in the case of greetings and yes/no questions/answers, which in a literary corpus would usually be represented by a relatively small number of expressions, but which in the chat environment are especially targeted by language creativity.

One future extension we plan to do is utilizing word n-grams for detecting speech acts. N-grams could better capture word order and thus better differentiate between patterns such as "*do you*" (a bigram), which most likely indicates a question, and "*you do*", which indicates a Command. Furthermore, we plan on using the dialog act cues proposed by (Samuel, Carberry and Vijay-Shanker 1998) for detecting n-grams that minimize the entropy of the distribution of speech acts in our training corpus and apply them as features for speech act classification. We also plan to test our leading words hypothesis on dialogue data that is naturally longer than chat posts. We hope to understand better from such data whether only two or three leading words are enough as opposed to six.

References

Alexandersson, J.; Buschbeck-Wolf, B.; Fujinami, T.; Maier, E.; Reithinger, N.; Schmitz, B.; Siegel, M. 1997. *Dialogue Acts in VerbMobil-2.* volume 226, Verbmobil Report, German Research Center for Artificial Intelligence (DFKI), Saarbrcken, 1998

Austin, J.L. 1962. *How to do Things with Words.* Oxford University Press, 1962.

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

Forsyth, E.N.; and Martell, C.H. 2007. *Lexical and Discourse Analysis of Online Chat Dialog.* International Conference on Semantic Computing (ICSC 2007), pp. 19-26.

Jurafsky, Dan.; and Martin, J.H. 2009. *Speech and Language Processing.* Prentice Hall, 2009.

Lin, J. 2007. *Automatic Author profiling of Online Chat Logs.* M.S.Thesis, Naval Postgraduate School, Monterey.

Marineau, J.; Wiemer-Hastings, P.; Harter, D.; Olde, B.; Chipman, P.; Karnavat, A.; Pomeroy, V.; and Graesser, A. 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 conference.

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.

Reithinger, N.; and Klesen, M. 1997. *Dialogue act classification using language models.* In Proceedings of EuroSpeech-97, pages 2235-2238.

Samuel, K.; Carberry, S.; and Vijay-Shanker, K. 1998. *Dialogue Act Tagging with Transformation-Based Learning.* In Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics, pages 1150-1156, Montreal, 1998.

Searle, J.R. 1969. *Speech Acts.* Cambridge University Press, GB, 1969.

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

Wu, T.; Khan, F.M.; Fisher, T.A.; Shuler, L.A.; and Pottenger, W.M. 2002. *Posting Act Tagging using Transformation-Based Learning.* In Proceedings of the Workshop on Foundations of Data Mining and Discovery, IEEE International Conference on Data Mining (ICDM'02), December 2002.