Classifying Sentences as Speech Acts in Message Board Posts

Ashequl Qadir and Ellen Riloff
School of Computing
University of Utah
Salt Lake City, UT 84112
{asheq,riloff}@cs.utah.edu

Abstract

This research studies the text genre of message board forums, which contain a mixture of expository sentences that present factual information and conversational sentences that include communicative acts between the writer and readers. Our goal is to create sentence classifiers that can identify whether a sentence contains a speech act, and can recognize sentences containing four different speech act classes: *Commissives*, *Directives*, *Expressives*, and *Representatives*. We conduct experiments using a wide variety of features, including lexical and syntactic features, speech act word lists from external resources, and domain-specific semantic class features. We evaluate our results on a collection of message board posts in the domain of veterinary medicine.

1 Introduction

In the 1990's, the natural language processing community shifted much of its attention to corpus-based learning techniques. Since then, most of the text corpora that have been annotated and studied are collections of *expository text* (e.g., news articles, scientific literature, etc.). The intent of expository text is to present or explain information to the reader. In recent years, there has been a growing interest in text genres that originate from Web sources, such as weblogs and social media sites (e.g., tweets). These text genres offer new challenges for NLP, such as the need to handle informal and loosely grammatical text, but they also pose new opportunities to study discourse and pragmatic phenomena that are fundamentally different in these genres.

Message boards are common on the WWW as a forum where people ask questions and post comments to members of a community. They are typically devoted to a specific topic or domain, such as finance, genealogy, or Alzheimer's disease. Some message boards offer the opportunity to pose questions to domain experts, while other communities are open to anyone who has an interest in the topic.

From a natural language processing perspective, message board posts are an interesting hybrid text genre because they consist of both expository text and conversational text. Most obviously, the conversations appear as a thread, where different people respond to each other's questions in a sequence of posts. Studying the conversational threads, however, is not the focus of this paper. Our research addresses the issue of conversational pragmatics within individual message board posts.

Most message board posts contain both expository sentences as well as speech acts. The person posting a message (*the "writer"*) often engages in speech acts with the readers. The writer may explicitly greet the readers (*"Hi everyone!"*), request help from the readers (*"Anyone have a suggestion?"*), or commit to a future action (*"I promise I will report back soon."*). But most posts contain factual information as well, such as general knowledge or personal history describing a situation, experience, or predicament.

Our research goals are twofold: (1) to distinguish between expository sentences and speech act sentences in message board posts, and (2) to clas-

Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pages 748–758,
Edinburgh, Scotland, UK, July 27–31, 2011.

sify speech act sentences into four types: *Commissives*, *Directives*, *Expressives*, and *Representatives*, following Searle's original taxonomy (Searle, 1976). Speech act classification could be useful for many applications. Information extraction systems could benefit from filtering speech act sentences (e.g., promises and questions) so that facts are only extracted from the expository text. Identifying *Directive* sentences could be used to summarize the questions being asked in a forum over a period of time. *Representative* sentences could be extracted to highlight the conclusions and beliefs of domain experts in response to a question.

In this paper, we present sentence classifiers that can identify speech act sentences and classify them as *Commissive*, *Directive*, *Expressive*, and *Representative*. First, we explain how each speech act class is manifested in message board posts, which can be different from how they occur in spoken dialogue. Second, we train classifiers to identify speech act sentences using a variety of lexical, syntactic, and semantic features. Finally, we evaluate our system on a collection of message board posts in the domain of veterinary medicine.

2 Related Work

There has been relatively little work on applying speech act theory to written text genres, and most of the previous work has focused on email classification. Cohen et al. (2004) introduced the notion of "email speech acts" defined as specific verb-noun pairs following a pre-designed ontology. They approached the problem as a document classification task. Goldstein and Sabin (2006) adopted this notion of email acts (Cohen et al., 2004) but focused on verb lexicons to classify them. Carvalho and Cohen (2005) presented a classification scheme using a dependency network, capturing the sequential correlations with the context emails using transition probabilities from or to a target email. Carvalho and Cohen (2006) later employed N-gram sequence features to determine which N-grams are meaningfully related to different email speech acts with a goal towards improving their earlier email classification based on the writer's intention.

Lampert et al. (2006) performed speech act classification in email messages following a verbal response modes (VRM) speech act taxonomy. They also provided a comparison of VRM taxonomy with Searle's taxonomy (Searle, 1976) of speech act classes. They evaluated several machine learning algorithms using syntactic, morphological, and lexical features. Mildinhall and Noyes (2008) presented a stochastic speech act model based on verbal response modes (VRM) to classify email intentions.

Some research has considered speech act classes in other means of online conversations. Twitchell and Jr. (2004) and Twitchell et al. (2004) employed speech act profiling by plotting potential dialogue categories in a radar graph to classify conversations in instant messages and chat rooms. Nastri et al. (2006) performed an empirical analysis of speech acts in the away messages of instant messenger services to achieve a better understanding of the communication goals of such services. Ravi and Kim (2007) employed speech act profiling in online threaded discussions to determine message roles and to identify threads with questions, answers, and unanswered questions. They designed their own speech act categories based on their analysis of student interactions in discussion threads.

The work most closely related to ours is the research of Jeong et al. (2009) on semi-supervised speech act recognition in both emails and forums. Like our work, their research also classifies individual sentences, as opposed to entire documents. However, they trained their classifier on spoken telephone (SWBD-DAMSL corpus) and meeting (MRDA corpus) conversations and mapped the labelled dialog act classes of these corpora to 12 dialog act classes that they found suitable for email and forum text genres. These dialog act classes (addressed as speech acts by them) are somewhat different from Searle's original speech act classes. They also used substantially different types of features than we do, focusing primarily on syntactic subtree structures.

3 Classifying Speech Acts in Message Board Posts

3.1 Speech Act Class Definitions

Searle's (Searle, 1976) early research on *speech acts* was seminal work in natural language processing that opened up a new way of thinking about con-

versational dialogue and communication. Our goal was to try and use Searle's original speech act definitions and categories as the basis for our work to the greatest extent possible, allowing for some interpretation as warranted by the WWW message board text genre.

For the purposes of defining and evaluating our work, we created detailed annotation guidelines for four of Searle's speech act classes that commonly occur in message board posts: *Commissives*, *Directives*, *Expressives*, and *Representatives*. We omitted the fifth of Searle's original speech act classes, *Declarations*, because we virtually never saw declarative speech acts in our data set.[1] The data set used in our study is a collection of message board posts in the domain of veterinary medicine. We designed our definitions and guidelines to reflect language use in the text genre of message board posts, trying to be as domain-independent as possible so that these definitions should also apply to message board texts representing other topics. However, we give examples from the veterinary domain to illustrate how these speech act classes are manifested in our data set.

Commissives: A *Commissive speech act* occurs when the speaker commits to a future course of action. In conversation, common Commissive speech acts are promises and threats. In message boards, these types of Commissives are relatively rare. However, we found many statements where the main purpose was to confirm to the readers that the writer would perform some action in the future. For example, a doctor may write *"I plan to do surgery on this patient tomorrow"* or *"I will post the test results when I get them later today"*. We viewed such statements as implicit commitments to the reader about intended actions. We also considered decisions not to take an action as Commissive speech acts (e.g., *"I will not do surgery on this cat because it would be too risky."*). However, statements indicating that an action will not occur because of circumstances beyond the writer's control were considered to be factual statements and not speech acts (e.g., *"I cannot do an ultrasound because my machine is broken."*).

Directives: A *Directive speech act* occurs when the speaker expects the listener to do something as a response. For example, the speaker may ask a question, make a request, or issue an invitation. Directive speech acts are common in message board posts, especially in the initial post of each thread when the writer explicitly requests help or advice regarding a specific topic. Many Directive sentences are posed as questions, so they are easy to identify by the presence of a question mark. However, the language in message board forums is informal and often ungrammatical, so many Directives are posed as a question but do not end in a question mark (e.g., *"What do you think."*). Furthermore, many Directive speech acts are not stated as a question but as a request for assistance. For example, a doctor may write *"I need your opinion on what drug to give this patient."* Finally, some sentences that end in question marks are rhetorical in nature and do not represent a Directive speech act, such as *"Can you believe that?"*.

Expressives: An *Expressive speech act* occurs in conversation when a speaker expresses his or her psychological state to the listener. Typical cases are when the speaker thanks, apologizes, or welcomes the listener. Expressive speech acts are common in message boards because writers often greet readers at the beginning of a post (*"Hi everyone!"*) or express gratitude for help from the readers (*"I really appreciate the suggestions."*). We also found Expressive speech acts in a variety of other contexts, such as apologies.

Representatives: According to Searle, a *Representative speech act* commits the speaker to the truth of an expressed proposition. It represents the speaker's belief of something that can be evaluated to be true or false. These types of speech acts were less common in our data set, but some cases did exist. In the veterinary domain, we considered sentences to be a Representative speech act when a doctor explicitly confirmed a diagnosis or expressed their suspicion or hypothesis about the presence (or absence) of a disease or symptom. For example, if a doctor writes that *"I suspect the patient has pancreatitis."* then this represents the doctor's own proposition/belief about what the disease might be.

Many sentences in our data set are stated as fact but could be reasonably inferred to be speech acts. For example, suppose a doctor writes *"The cat has*

[1] Searle defines Declarative speech acts as statements that bring about a change in status or condition to an object by virtue of the statement itself. For example, a statement declaring war or a statement that someone is fired.

pancreatitis.". It would be reasonable to infer that the doctor writing the post diagnosed the cat with pancreatitis. And in many cases, that is true. However, we saw many posts where that inference would have been wrong. For example, the following sentence might say *"The cat was diagnosed by a previous vet but brought to me due to new complications"* or *"The cat was diagnosed with it 8 years ago as a kitten in the animal shelter"*. Consequently, we were very conservative in labelling sentences as Representative speech acts. Any sentence presented as fact was not considered to be a speech act. A sentence was only labelled as a Representative speech act if the writer explicitly expressed his belief.

3.2 Features for Speech Act Classification

To create speech act classifiers, we designed a variety of lexical, syntactic, and semantic features. We tried to capture linguistic properties associated with speech act expressions as well as discourse properties associated with individual sentences and the message board post as a whole. We also incorporated speech act word lists that were acquired from external resources, and used two types of semantic features to represent semantic entities associated with the veterinary domain. Except for the semantic features, all of our features are domain-independent so should be able to recognize speech act sentences across different domains. We experimented with domain-specific semantic features to test our hypothesis that Commissive speech acts can be associated with domain-specific semantic entities.

For the purposes of analysis, we partition the feature set into three groups: *Lexical and Syntactic (LexSyn) Features*, *Speech Act Clue Features*, and *Semantic Features*. Unless otherwise noted, all of the features had binary values indicating the presence or absence of that feature.

3.2.1 Lexical and Syntactic Features

We designed a variety of features to capture lexical and syntactic properties of words and sentences. We described the feature set below, with the features categorized based on the type of information that they capture.

Unigrams: We created bag-of-word features representing each unigram in the training set. Numbers were replaced with a special # token.

Personal Pronouns: We defined three features to look for the presence of a 1st person pronoun, 2nd person pronoun, and 3rd person pronoun. We included the subjective, objective, and possessive form of each pronoun (e.g., *he*, *him*, and *his*).

Tense: Speech acts such as Commissives can be related to tense. We created three features to identify verb phrases that occur in the *past*, *present*, or *future* tense. To recognize tense, we followed the rules defined by Allen (1995).

Tense + Person: We created four features that require the presence of a first person subjective pronoun (I, we) within a two word window on the left of a verb phrase matching one of four tense representations: *past*, *present*, *future*, and *present progressive* (a subset of the more general *present* tense representation).

Modals: One feature indicates whether the sentence contains a modal (*may*, *must*, *shall*, *will*, *might*, *should*, *would*, *could*).

Infinitive VP: One feature looks for an infinitive verb phrase ('to' followed by a verb) that is preceded by a first person pronoun (I, we) within a three word window on the left. This feature tries to capture common Commissive expressions (e.g., *"I definitely plan to do the test tomorrow."*).

Plan Phrases: Commissives are often expressed as a plan, so we created a feature that recognizes four types of plan expressions: *"I am going to"*, *"I am planning to"*, *"I plan to"*, and *"My plan is to"*.

Sentence contains Early Punctuation: One feature checks for the following punctuation marks within the first three tokens of the sentence: , : ! This feature was designed to recognize greetings, such as: *"Hi," ,* or *"Hiya everyone !"*.

Sentence begins with Modal/Verb: One feature checks if a sentence begins with a modal or verb. The intuition is to capture interrogative and imperative sentences, since they are likely to be Directives.

Sentence begins with WH Question: One feature checks if a sentence begins with a WH question word (Who, When, Where, What, Which, What, How).

Neighboring Question: One feature checks whether the following sentence contains a question mark '?'. We observed that in message boards, *Directives* often occur in clusters.

Sentence Position: Four binary features represent the relative position of the sentence in the post. One feature indicates whether it is the first sentence, one feature indicates whether it is the last sentence, one feature indicates whether it is the second to last sentence, and one feature indicates whether the sentence occurs in the bottom 25% of the message. The motivation for these features is that Expressives often occur at the beginning and end of the post, and Directives tend to occur toward the end.

Number of Verbs: One feature represents the number of verbs in the sentence using four possible values: 0, 1, 2, >2. Some speech acts classes (e.g., Expressives) may occur with no verbs, and rarely occur in long, complex sentences.

3.2.2 Speech Act Word Clues

We collected speech act word lists (mostly verbs) from two external sources. In Searle's original paper (Searle, 1976), he listed words that he considered to be indicative of speech acts. We discarded a few that we considered to be overly general, and we added a few additional words. We also collected a list of speech act verbs published in (Wierzbicka, 1987). The details for these *speech act clue lists* are given below. Our system recognized all derivations of these words.

Searle Keywords: We created one feature for each speech act class. The Representative keywords were: (*hypothesize*, *insist*, *boast*, *complain*, *conclude*, *deduce*, *diagnose*, and *claim*). We discarded 3 words from Searle's list (*suggest*, *call*, *believe*) and added 2 new words, *assume* and *suspect*. The Directive keywords were: (*ask*, *order*, *command*, *request*, *beg*, *plead*, *pray*, *entreat*, *invite*, *permit*, *advise*, *dare*, *defy*, *challenge*). We added the word *please*. The Expressives keywords were: (*thank*, *apologize*, *congratulate*, *condole*, *deplore*, *welcome*). We added the words *appreciate* and *sorry*. Searle did not provide any hint on possible indicator words for Commissives, so we manually defined five likely Commissive keywords: (*plan*, *commit*, *promise*, *tomorrow*, *later*).

Wierzbicka Verbs: We created one feature that included 228 speech act verbs listed in the book *"English speech act verbs: a semantic dictionary"* (Wierzbicka, 1987)[2].

3.2.3 Semantic Features

All of the previous features are domain-independent and should be useful for identifying speech acts sentences across many domains. However, we hypothesized that semantic entities may correlate with speech acts within a particular domain. For example, consider medical domains. Representative speech acts may involve diagnoses and hypotheses regarding diseases and symptoms. Similarly, Commissive speech acts may reveal a doctor's plan or intention regarding the administration of drugs or tests. Thus, it may be beneficial for a classifier to know whether a sentence contains certain semantic entities. We experimented with two different sources of semantic information.

Semantic Lexicon: Basilisk (Thelen and Riloff, 2002) is a bootstrapping algorithm that has been used to induce semantic lexicons for terrorist events (Thelen and Riloff, 2002), biomedical concepts (McIntosh, 2010), and subjective/objective nouns for opinion analysis (Riloff et al., 2003). We ran Basilisk over our collection of 15,383 veterinary message board posts to create a semantic lexicon for veterinary medicine. As input, Basilisk requires seed words for each semantic category. To obtain seeds, we parsed the corpus using a noun phrase chunker, sorted the head nouns by frequency, and manually identified the 20 most frequent nouns belonging to four semantic categories: DISEASE/SYMPTOM, DRUG, TEST, and TREATMENT.

However, the induced TREATMENT lexicon was of relatively poor quality so we did not use it. The DISEASE/SYMPTOM lexicon appeared to be of good quality, but it did not improve the performance of our speech act classifiers. We suspect that this is due to the fact that diseases were not distinguised from symptoms in our lexicon.[3] Representative speech acts are typically associated with disease diagnoses

[2] openlibrary.org/b/OL2413134M/English_speech_act_verbs

[3] We induced a single lexicon for diseases and symptoms because it is difficult to draw a clear line between them semantically. A veterinary consultant explained to us that the same term (e.g., diabetes) may be considered a symptom in one context if it is secondary to another condition (e.g., pancreatitis) but a disease in a different context if it is the primary diagnosis.

and hypotheses, rather than individual symptoms.

In the end, we only used the DRUG and TEST semantic lexicon in our classifiers. We used all 1000 terms in the DRUG lexicon, but only used the top 200 TEST words because the quality of the lexicon seemed questionable after that point.

Semantic Tags: We also used bootstrapped contextual semantic taggers (Huang and Riloff, 2010) that had been previously trained for the domain of veterinary medicine. These taggers assign semantic class labels to noun phrase instances based on the surrounding context in a sentence. The taggers were trained on 4,629 veterinary message board posts using 10 seed words for each semantic category (see (Huang and Riloff, 2010) for details). To ensure good precision, only tags that have a confidence value ≥ 1.0 were used. Our speech act classifiers used the tags associated with two semantic categories: DRUG and TEST.

3.3 Classification

To create our classifiers, we used the Weka (Hall et al., 2009) machine learning toolkit. We used Support Vector Machines (SVMs) with a polynomial kernel and the default settings supplied by Weka. Because a sentence can include multiple speech acts, we created a set of binary classifiers, one for each of the four speech act classes. All four classifiers were applied to each sentence, so a sentence could be assigned multiple speech act classes.

4 Evaluation

4.1 Data Set

Our data set consists of message board posts from the Veterinary Information Network (VIN), which is a web site (www.vin.com) for professionals in veterinary medicine. Among other things, VIN hosts message board forums where veterinarians and other veterinary professionals can discuss issues and pose questions to each other. Over half of the small animal veterinarians in the U.S. and Canada use the VIN web site.

We obtained 15,383 VIN message board threads representing three topics: cardiology, endocrinology, and feline internal medicine. We did basic cleaning, removing html tags and tokenizing numbers. We then applied the Stanford part-of-speech tagger (Toutanova et al., 2003) to each sentence to obtain part-of-speech tags for the words. For our experiments, we randomly selected 150 message board threads from this collection. Since the goal of our work was to study speech acts in sentences, and not the conversational dialogue between different writers, we used only the initial post of each thread. These 150 message board posts contained a total of 1,956 sentences, with an average of 13.04 sentences per post. In the next section, we explain how we manually annotated each sentence in our data set to create gold standard speech act labels.

4.2 Gold Standard Annotations

To create training and evaluation data for our research, we asked two human annotators to manually label sentences in our message board posts. Identifying speech acts is not always obvious, even to people, so we gave them detailed annotation guidelines describing the four speech act classes discussed in Section 3.1. Then we gave them the same set of 50 message board posts from our collection to annotate independently. Each annotator was told to assign one or more speech act classes to each sentence (COM, DIR, EXP, REP), or to label the sentence as having no speech acts (NONE). The vast majority of sentences had either no speech acts or at most one speech act, but a small number of sentences contained multiple types of speech acts.

We measured the inter-annotator agreement of the two human judges using the kappa (κ) score (Carletta, 1996). However, kappa agreement scores are only applicable to labelling schemes where each instance receives a single label. Therefore we computed kappa agreement in two different ways to look at the results from two different perspectives. In the first scheme, we discarded the small number of sentences that had multiple speech act labels and computed kappa on the rest.[4] This produced a kappa score of .95, suggesting extremely high agreement. However, over 70% of the sentences in our data set have no speech act at all, so NONE was by far the most common label. Consequently, this agreement score does not necessarily reflect how consistently the judges agreed on the four speech act classes.

[4]Of the 594 sentences in these 50 posts, only 22 sentences contained multiple speech act classes.

In the second scheme, we computed kappa for each speech act category independently. For each category C, the judges were considered to be in agreement if both of them assigned category C to the sentence or if neither of the judges assigned category C to the sentence. Table 1 shows the κ agreement scores using this approach.

Speech Act	Kappa (κ) score
Expressive	.97
Directive	.94
Commissive	.81
Representative	.77

Table 1: Inter-annotator (κ) agreement

Inter-annotator agreement was very high for both the Expressive and Directive classes. Agreement was lower for the Commissive and Representative classes, but still relatively good so we felt comfortable that we had high-quality annotations.

To create our final data set, the two judges adjudicated their disagreements on this set of 50 posts. We then asked each annotator to label an additional (different) set of 50 posts each. All together, this gave us a gold standard data set consisting of 150 annotated message board posts. Table 2 shows the distribution of speech act labels in our data set. 71% of the sentences did not include any speech acts. These were usually expository sentences containing factual information. 29% of the sentences included one or more speech acts, so nearly $\frac{1}{3}$ of the sentences were conversational in nature. Directive and Expressive speech acts are by far the most common, with nearly 26% of all sentences containing one of these speech acts. Commissive and Representative speech acts are less common, each occurring in less than 3% of the sentences.[5]

4.3 Experimental Results

4.3.1 Speech Act Filtering

For our first experiment, we created a *speech act filtering classifier* to distinguish sentences that contain one or more speech acts from sentences that do not contain any speech acts. Sentences labelled as having one or more speech acts were positive instances, and sentences labelled as NONE were negative instances. Speech act filtering could be useful for many applications, such as information extraction systems that only seek to extract facts. For example, information may be posed as a question (in a Directive) rather than a fact, information may be mentioned as part of a future plan (in a Commissive) that has not actually happened yet, or information may be stated as a hypothesis or suspicion (in a Representative) rather than as a fact.

[5]These numbers do not add up to 100% because some sentences contain multiple speech acts.

Speech Act	# sentences	distribution
None	1397	71.42%
Directive	311	15.90%
Expressive	194	9.92%
Representative	57	2.91%
Commissive	51	2.61%

Table 2: Speech act class distribution in our data set.

We performed 10-fold cross validation on our set of 150 annotated message board posts. Initially, we used all of the features defined in Section 3.2. However, during the course of our research we discovered that only a small subset of the lexical and syntactic features seemed to be useful, and that removing the unnecessary features improved performance. So we created a subset of *minimal lexsyn features*, which will be described in Section 4.3.2. For speech act filtering, we used the *minimal lexsyn features* plus the speech act clues and semantic features.[6]

Class	P	R	F
Speech Act	.86	.83	.84
No Speech Act	.93	.95	.94

Table 3: Precision, Recall, F-measure for speech act filtering.

Table 3 shows the performance for speech act filtering with respect to Precision (P), Recall (R), and F-measure score (F).[7] The classifier performed well, recognizing 83% of the speech act sentences with 86% precision, and 95% of the expository (no

[6]This is the same feature set used to produce the results for row E of Table 4.

[7]We computed an F1 score with equal weighting of precision and recall.

	Features	Commissives			Directives			Expressives			Representatives		
		P	R	F	P	R	F	P	R	F	P	R	F
	Baselines												
	Com baseline	.45	.08	.14	-	-	-	-	-	-	-	-	-
	Dir baseline	-	-	-	.97	.73	.83	-	-	-	-	-	-
	Exp baseline 1	-	-	-	-	-	-	.58	.18	.28	-	-	-
	Exp baseline 2	-	-	-	-	-	-	.97	.86	.91	-	-	-
	Rep baseline	-	-	-	-	-	-	-	-	-	1.0	.05	.10
	Classifiers												
U	Unigram	.45	.20	.27	.87	.84	.85	.97	.88	.92	.32	.12	.18
A	U+all lexsyn	.52	.33	.40	.87	.84	**.86**	.98	.88	.92	.30	.14	.19
B	U+minimal lexsyn	.59	.33	.42	.87	.85	**.86**	.98	.88	.92	.32	.14	.20
C	B+speechActClues	.57	.31	.41	.86	.84	.85	.97	.91	**.94**	.33	.16	**.21**
D	C+semTest	.64	.35	.46	.87	.84	.85	.97	.91	**.94**	.33	.16	**.21**
E	D+semDrug	.63	.39	**.48**	.86	.84	.85	.97	.91	**.94**	.32	.16	**.21**

Table 4: Precision, Recall, F-measure for four speech act classes. The highest F score for each category appears in boldface.

speech act) sentences with 93% precision.

4.3.2 Speech Act Categorization

BASELINES

Our next set of experiments focused on labelling sentences with the four specific speech act classes: *Commissive*, *Directive*. *Expressive*, and *Representative*. To assess the difficulty of identifying each speech act category, we created several simple baselines using our intuitions about each category.

For Commissives, we created a heuristic to capture the most obvious cases of future tense (because Commissive speech acts represent a writer's commitment toward a future course of action). For example, the presence of the phrases 'I will' and 'I shall' were hypothesized by Cohen et al. (2004) to be useful bigram clues for Commissives. This baseline looks for future tense verb phrases with a 1st person pronoun within one or two words preceding the verb phrase. The **Com** baseline row of Table 4 shows the results for this heuristic, which obtained 8% recall with 45% precision. The heuristic applied to only 9 sentences in our test set, 4 of which contained a Commissive speech act.

Directive speech acts are often questions, so we created a baseline system that labels all sentences containing a question mark as a Directive. The **Dir** baseline row of Table 4 shows that 97% of sentences with a question mark were indeed Directives.[8] But only 73% of the Directive sentences contained a question mark. The remaining 27% of Directives did not contain a question mark and generally fell into two categories. Some sentences asked a question but the writer ended the sentence with a period (e.g., *"Has anyone seen this before."*). And many directives were expressed as requests rather than questions (e.g., *"Let me know if anyone has a suggestion."*).

For Expressives, we implemented two baselines. **Exp** baseline 1 simply looks for an exclamation mark, but this heuristic did not work well (18% recall with 58% precision) because exclamation marks were often used for general emphasis (e.g., *"The owner is frustrated with cleaning up urine!"*). **Exp** baseline 2 looks for the presence of four common expressive words (*appreciate*, *hi*, *hello*, *thank*), including morphological variations of *appreciate* and *thank*. This baseline produced very good results, 86% recall with 97% precision. Obviously a small set of common expressions account for most of the Expressive speech acts in our corpus. However, the word "hi" did produce some false hits because it was used as a shorthand for "high", usually when reporting test results (e.g., *"hi calcium"*).

[8] 235 sentences contained a question mark, and 227 of them were Directives.

Finally, as a baseline for the Representative class we simply looked for the words *diagnose(d)* and *suspect(ed)*. The **Rep** baseline row of Table 4 shows that this heuristic was 100% accurate, but only produced 5% recall (matching 3 of the 57 Representative sentences in our test set).

CLASSIFIER RESULTS

The bottom portion of Table 4 shows the results for our classifiers. As we explained in Section 3.3, we created one classifier for each speech act category, and all four classifiers were applied to each sentence. So a sentence could receive anywhere from 0-4 speech act labels indicating how many different types of speech acts appeared in the sentence. We trained and evaluated each classifier using 10-fold cross-validation on our gold standard data set.

The *Unigram (U)* row shows the performance of classifiers that use only unigram features. For Directives, we see a 2% F-score improvement over the baseline, which reflects a recall gain of 11% but a corresponding precision loss of 10%. The unigrams are clearly helpful in identifying many Directive sentences that do not end in a question mark, but at some cost to accuracy. For Expressives, the unigram classifier achieves an F score of 92%, identifying slightly more Expressive sentences than the baseline with the same level of precision. For Commissives and Representatives, the unigram classifiers performed susbtantially better than their corresponding baseline systems, but performance is still relatively weak.

Row A (*U+ all lexsyn*) in Table 4 shows the results using unigram features plus all of the lexical and syntactic features described in Section 3.2.1. The lexical and syntactic features dramatically improve performance on Commissives, increasing F score from 27% to 40%, and they produce a 2% recall gain for Representatives but with a corresponding loss of precision.

However, we observed that only a few of the lexical and syntactic features had much impact on performance. We experimented with different subsets of the features and obtained even better performance when using just 10 of them, which we will refer to as the *minimal lexsyn* features. The *minimal lexsyn* feature set consists of the 4 Tense+Person features, the Early Punctuation feature, the Sentence begins with Modal/Verb feature, and the 4 Sentence Position features. Row B shows the results using unigram features plus only these *minimal lexsyn* features. Precision improves for Commissives by an additional 7% and Representatives by 2% when using only these lexical and syntactic features. Consequently, we use the *minimal lexsyn* features for the rest of our experiments.

Row C shows the results of adding the speech act clue words (see Section 3.2.2) to the feature set used in Row B. The speech act clue words produced an additional recall gain of 3% for Expressives and 2% for Representatives, although performance on Commissives dropped 2% in both recall and precision.

Rows D and E show the results of adding the semantic features. We added one semantic category at a time to measure the impact of them separately. Row D adds two semantic features for the TEST category, one from the Basilisk lexicon and one from the semantic tagger. The TEST semantic features produced an F-score gain of 5% for Commissives, improving recall by 4% and precision by 7%. Row E adds two semantic features for the DRUG category. The DRUG features produced an additional F-score gain of 2% for Commissives, improving recall by 4% with a slight drop in precision.

4.4 Analysis

Together, the TEST and DRUG semantic features dramatically improved the classifier's ability to recognize Commissive speech acts, increasing its F score from 41% $\rightarrow$ 48%. This result demonstrates that in the domain of veterinary medicine, some types of semantic entities are associated with speech acts. Our intuition behind this result is that commitments are usually related to future actions. In veterinary medicine, TESTS and DRUGS are associated with actions performed by doctors. Doctors help their patients by prescribing or administering drugs and by conducting tests. So these semantic entities may serve as a proxy to implicitly represent actions that the doctor has done or may do. In future work, explicitly recognizing actions and events many be a worthwhile avenue to further improve results.

We achieved good success at identifying both Directives and Expressives, although simple heuristics also perform well on these categories. We showed that training a Directive classifier can help to iden-

tify Directive sentences that do not end with a question mark, although at the cost of some precision.

The Commissive speech act class benefitted the most from the rich feature set. Unigrams are clearly not sufficient to identify Commissive sentences. Many different types of clues seem to be important for recognizing these sentences. The improvements obtained from adding semantic features also suggests that domain-specific semantics can be useful for recognizing some speech acts. However, there is still ample room for improvement, illustrating that speech act classification is a challenging problem.

Representative speech acts were by far the most difficult to recognize. We believe that there are several reasons for their low performance. First, Representatives were sparse in the data set, occurring in only 2.91% of the sentences. Consequently, the classifier had relatively few positive training instances. Second, Representatives had the lowest inter-annotator agreement, indicating that human judges had difficulty recognizing these speech acts too. The judges often disagreed about whether a hypothesis or suspicion was the writer's own belief or whether it was stated as a fact reflecting general medical knowledge. The message board text genre is especially challenging in this regard because the writer is often presumed to be expressing his/her beliefs even when the writer does not explicitly say so. Finally, our semantic features could not distinguish between diseases and symptoms. Access to a resource that can reliably identify disease terms could potentially improve performance in this domain.

5 Conclusions

Our goal was to identify speech act sentences in message board posts and to classify the sentences with respect to four categories in Searle's (1976) speech act taxonomy. We achieved good results for speech act filtering and the identification of Directive and Expressive speech act sentences. We found that Representative and Commissive speech acts are much more difficult to identify, although the performance of our Commissive classifier substantially improved with the addition of lexical, syntactic, and semantic features. Except for the semantic class information, our feature set is domain-independent and could be used to recognize speech act sentences in message boards for any domain. Furthermore, our features only rely on part-of-speech tags and do not require parsing, which is of practical importance for text genres such as message boards that are littered with ungrammatical text, typos, and shorthand notations.

In future work, we believe that segmenting sentences into clauses may help to train classifiers more precisely. Ultimately, we would like to identify the speech act expressions themselves because some sentences contain speech acts as well as factual information. Extracting the speech act expressions and clauses from message boards and similar text genres could provide better tracking of questions and answers in web forums and be used for summarization.

6 Acknowledgments

We gratefully acknowledge that this research was supported in part by the National Science Foundation under grant IIS-1018314. Any opinions, findings, and conclusion or recommendations expressed in this material are those of the authors and do not necessarily reflect the view of the U.S. government.

References

James Allen. 1995. *Natural language understanding (2nd ed.)*. Benjamin-Cummings Publishing Co., Inc., Redwood City, CA, USA.

Jean Carletta. 1996. Assessing agreement on classification tasks: the kappa statistic. *Comput. Linguist.*, 22:249–254, June.

Vitor R. Carvalho and William W. Cohen. 2005. On the collective classification of email "speech acts". In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 345–352, New York, NY, USA. ACM Press.

Vitor R. Carvalho and William W. Cohen. 2006. Improving "email speech acts" analysis via n-gram selection. In *Proceedings of the HLT-NAACL 2006 Workshop on Analyzing Conversations in Text and Speech*, ACTS '09, pages 35–41, Stroudsburg, PA, USA. Association for Computational Linguistics.

William W. Cohen, Vitor R. Carvalho, and Tom M. Mitchell. 2004. Learning to classify email into "speech acts". In *EMNLP*, pages 309–316. ACL.

Jade Goldstein and Roberta Evans Sabin. 2006. Using speech acts to categorize email and identify email gen-

res. In *Proceedings of the 39th Annual Hawaii International Conference on System Sciences - Volume 03*, pages 50.2–, Washington, DC, USA. IEEE Computer Society.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The weka data mining software: an update. *SIGKDD Explor. Newsl.*, 11:10–18, November.

Ruihong Huang and Ellen Riloff. 2010. Inducing domain-specific semantic class taggers from (almost) nothing. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 275–285, Stroudsburg, PA, USA. Association for Computational Linguistics.

Minwoo Jeong, Chin-Yew Lin, and Gary Geunbae Lee. 2009. Semi-supervised speech act recognition in emails and forums. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3 - Volume 3*, EMNLP '09, pages 1250–1259, Stroudsburg, PA, USA. Association for Computational Linguistics.

Andrew Lampert, Robert Dale, and Cecile Paris. 2006. Classifying speech acts using verbal response modes. In *Proceedings of the 2006 Australasian Language Technology Workshop (ALTW2006)*, pages 34–41. Sydney Australia : ALTA.

Tara McIntosh. 2010. Unsupervised discovery of negative categories in lexicon bootstrapping. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 356–365, Stroudsburg, PA, USA. Association for Computational Linguistics.

John Mildinhall and Jan Noyes. 2008. Toward a stochastic speech act model of email behavior. In *CEAS*.

Jacqueline Nastri, Jorge Pena, and Jeffrey T. Hancock. 2006. The construction of away messages: A speech act analysis. *J. Computer-Mediated Communication*, pages 1025–1045.

Sujith Ravi and Jihie Kim. 2007. Profiling student interactions in threaded discussions with speech act classifiers. In *Proceeding of the 2007 conference on Artificial Intelligence in Education: Building Technology Rich Learning Contexts That Work*, pages 357–364, Amsterdam, The Netherlands, The Netherlands. IOS Press.

Ellen Riloff, Janyce Wiebe, and Theresa Wilson. 2003. Learning subjective nouns using extraction pattern bootstrapping. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, pages 25–32, Stroudsburg, PA, USA. Association for Computational Linguistics.

John R. Searle. 1976. A classification of illocutionary acts. *Language in Society*, 5(1):pp. 1–23.

Michael Thelen and Ellen Riloff. 2002. A bootstrapping method for learning semantic lexicons using extraction pattern contexts. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10*, EMNLP '02, pages 214–221, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 173–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

Douglas P. Twitchell and Jay F. Nunamaker Jr. 2004. Speech act profiling: a probabilistic method for analyzing persistent conversations and their participants. In *System Sciences, 2004. Proceedings of the 37th Annual Hawaii International Conference on*, pages 1–10, January.

Douglas P. Twitchell, Mark Adkins, Jay F. Nunamaker Jr., and Judee K. Burgoon. 2004. Using speech act theory to model conversations for automated classification and retrieval. In *Proceedings of the International Working Conference Language Action Perspective Communication Modelling (LAP 2004)*, pages 121–130.

A. Wierzbicka. 1987. *English speech act verbs: a semantic dictionary*. Academic Press, Sydney, Orlando.