Analyzing Microtext: Papers from the 2011 AAAI Workshop (WS-11-05)

What Are Tweeters Doing: Recognizing Speech Acts in Twitter

Renxian Zhang
The Hong Kong Polytechnic University
csrzhang@comp.polyu.edu.hk

Dehong Gao
The Hong Kong Polytechnic University
csdgao@comp.polyu.edu.hk

Wenjie Li
The Hong Kong Polytechnic University
cswjli@comp.polyu.edu.hk

Abstract

Speech acts provide good insights into the communicative behavior of tweeters on Twitter. This paper is mainly concerned with speech act recognition in Twitter as a multi-class classification problem, for which we propose a set of word-based and character-based features. Inexpensive, robust and efficient, our method achieves an average F1 score of nearly 0.7 with the existence of much noise in our annotated Twitter data. In view of the deficiency of training data for the task, we experimented extensively with different configurations of training and test data, leading to empirical findings that may provide valuable reference for building benchmark datasets for sustained research on speech act recognition in Twitter.

Introduction

As an information exchange platform, the microblogging service of Twitter has fostered a generation of tweeters and continued to expand this generation. For newcomers eager to join, they will have to understand what the tweeters are doing by posting tweets under a topic.

Essentially, twittering is a communicative act due to its social networking nature. When communicating with tweets, people may share information, ask questions, make suggestions, express sentiments, etc. which are all instances of "speech acts" (Austin, 1962). In this work, we are concerned with recognizing such speech acts in tweets as a tentative step towards tweet content understanding.

The automatic recognition of speech acts has significant impacts on Twitter and tweeters. 1) For **Twitter itself**, speech act recognition helps to reveal how a topic is constituted in terms of tweeters' speech acts and whether there is any topic shift. For example, a news topic is normally expected to be mainly composed of fact statements, and if the actual speech act distribution deviates from our expectation (e.g., a considerable proportion of making suggestions under a news topic), we can suspect topic shift or spamming. 2) For **tweet posters**, speech act recognition helps us to understand their behavior as a community defined by their common interest in a certain topic, as well as their behavior as individuals susceptible to the others in the same community. For example, observing a fashion group's speech acts and individuals' speech act variation along a timeline may shed interesting light on how a common interest in a certain aspect (e.g., criticizing or asking for price) has evolved. 3) For **tweet readers**, the distribution of speech act types under a specific topic enables them to follow a topic with mostly desirable speech acts (e.g., fact statements for a news topic) or follow only tweets with desirable speech acts (e.g., comments only for a celebrity topic), thus helping them to become efficient readers in a sea of tweets.

To recognize speech acts in tweets, we basically follow the popular taxonomy of Searle (1975): assertives (asserting something true), commissives (committing to future action), directives (getting the listener to do something), declaratives (bringing out a different state of world), and expressives (expressing psychological state). However, not all of them are readily applicable to Twitter speech acts because commissives are rare and declaratives are even rarer according to our observation of the experimental data. We also observe that there are two distinct types of directives: asking questions and making suggestions. Table 1 lists all the 5 speech act types we use, alongside the corresponding Searle's types.

Table 1: Types of Speech Acts

Our Types	Searle's (1975) Types
Statement	Assertive
Question	Directive
Suggestion	
Comment	Expressive
Miscellaneous	Commissive
	Declarative

Twitter speech act recognition is thus a multi-class classification task, and our primary goal in this work is to find a set of robust features appropriate for it. As we are not aware of any sizable speech act-annotated Twitter data at the time of writing, we will also explore on what level training data should be prepared – topic-level, category-level, or the whole Twitter space – and whether one model trained on one topic/category can adapt to a different topic/category. The empirical results will help to define the scope of the task and inform training data preparation for future development of the task. Those are our contributions in the current work, which will benefit any future research pursuing this direction.

In the rest of the paper, after surveying some related work, we discuss the various features to be used. After that, we report our experiments on 6 annotated Twitter datasets and discuss the results. Finally, the major finding and future work are summarized.

Related Work

The notion of speech act was proposed nearly half a century ago by the philosopher Austin (1962) and the speech act theory has since established itself in pragmatics (Levinson, 1983), characterized by sustained interest in its taxonomical (Searle, 1975) and logical aspects (Searle and Vanderveken, 1985).

In computational linguistics, speech act is also known as dialogue act and the main interest is in their ability to automatically recognize and model conversation (Stolcke et al., 2000), which relies on annotated corpora such as Switchboard-DAMSL (Jurafsky et al., 1997) and Meeting Recorder Dialog Act (Dhillon et al., 2004).

Such corpora use telephone or face-to-face conversation data, which cannot directly facilitate work on the massive data in electronic forms: email, instant messaging, and microblog. To circumvent the problem, semi-supervised classification (Jeong et al., 2009) or unsupervised models (Ritter et al., 2010) have been proposed. However, no speech act-annotated corpus on Twitter has been released.

We are not aware of published work on Twitter speech act recognition as a classification task. Although Jeong et al. (2009)'s work is close, the extremely noisy nature of Twitter data makes some of their fine-grained features (e.g., subtree features) inapplicable and leaves an open question whether intensive Twitter text normalization (Kaufmann and Kalita, 2010) is necessary for speech act recognition.

Features for Twitter Speech Act Recognition

A major challenge for any work on Twitter data is to deal with its notorious noisiness: spelling mistakes, grammatical errors, mixture of letter and non-letter characters, etc. For our task, we have two options to tackle the problem: 1) converting Twitter text to "normalized" text as was done by Kaufmann and Kalita (2010) and then extracting features, which is nontrivial; 2) extracting robust features directly from the noisy text. We prefer the second option partly because we aim at simplicity and robustness. More importantly, noise in Twitter may hold telltale features for speech act recognition, such as the repeated use of *?* or *!* to predict "question" or "comment".

In the following, we describe the features we use for recognizing the 5 speech acts (Table 1) as a classification task, including word-based and character-based features.

Word-based Features

We have 535 word-based features composed of two major types. All of them are binary-valued, indicating their presence or not.

Cue Words and Phrases

Some speech acts are typically signaled by some cue words or phrases, such as *whether* for "question" and *could you please* for "suggestion". There are some manually compiled lexicons for speech act cues (e.g., Wierzbicka, 1987), but we refrain from using them for two reasons. First, the cue lexicons are very limited, consisting mostly of verbs. But words of other parts of speech (including closed-class words) and phrases may be equally predictive. Second, such lexicons only serve standard English, not Twitter English, which is rife with non-standard spellings, acronyms, and abbreviations.

Therefore, we semi-automatically compiled a cue lexicon of Twitter English from our annotated datasets with a total of 8613 tweets (see details in Experiments). First, high-frequency unigrams, bigrams, and trigrams are collected for "statement", "question", "suggestion", and "comment". Then we manually checked them and come up with a list of 531 features. Table 2 shows some examples.

Table 2: Examples of Cue Words and Phrases

	Examples	Total
Unigrams	*know, hurray, omg, pls, why …*	268
Bigrams	*do it, i bet, ima need, you can …*	164
Trigrams	*?!?, heart goes out, rt if you …*	99

Non-cue Words

Some special words, though not intuitively cuing speech acts, may indirectly signal speech acts. We use four types of such non-cue words explained in the following.

Abbreviations and Acronyms: one feature indicating whether such shortened word forms appear. We collected the lexicon from online [1] and published (Crystal, 2006) resources, a total of 1153 words. Examples are *4ever* for "forever" and *tq* for "thank you".

We then restore the shortened words to their original forms before extracting the next two features: opinion words and vulgar words, both relying on full spellings.

Opinion Words: one feature indicating whether opinion words appear. We used the SentiWordNet (Baccianella et al., 2010) and Wilson Lexicon (Wilson and Wiebe, 2003), which are well known for the task of sentiment analysis.

[1] http://www.chatslang.com

As we are only interested in strong opinion words, we build a lexicon by intersecting highly opinionated words (positive score + negative score ≥ 0.5) from SentiWordNet with the "strong" words from the Wilson Lexicon, resulting in a total of 2460 words, like *shallow*, *vague*, *scary*, etc.

Vulgar Words: one feature indicating whether vulgar words appear. We used the API from an online resource[2] and collected 341 words like *c**t* and *f**k*[3].

Emoticons: one feature indicating whether emoticons appear. We collected 276 emoticons from an online resource[4], such as O:) and *-*.

Character-based Features

We have 8 character-based features composed of two types, which indicate the frequency and position of special characters and are either binary- or ternary- valued.

Twitter-specific Symbols

We concentrate on the 3 symbols specific to Twitter: #, @, and *RT*. # is a hashtag marker often used in a mention of something to be stated about or commented on; @ is a prefix to a tweeter account, which tends to be associated with questions or suggestions; *RT* stands for retweeting and its presence, especially in the initial position, strongly indicates a statement. Repeated use of them are even stronger indicators of possible speech acts.

Each of those symbols is associated with 2 features: 1 binary-valued feature indicating whether the symbol is in the initial position of a tweet and 1 ternary-valued feature indicating whether the symbol does not appear (0), appear 1 or 2 times (1), or appear more than 2 times (2).

Indicative Punctuations

We single out 2 punctuations: *?* and *!* as the former often indicates a question and the latter is likely to indicate a comment or suggestion. Repeated use of them increases the likelihood. Each of them is associated with 1 ternary-valued feature indicating zero appearance (0), 1 or 2 appearances (1), or 3 or more appearances (2).

Experiments

We ran experiments on 6 sets of Twitter data with hand-annotated speech act labels by focusing on three issues: 1) the effects of different features on speech act classification; 2) the effect of data granularity (topic-level, category-level, or Twitter space-level) on classification performance; 3) the adaptability of a classification model trained on one topic/category to test data on another topic/category. The details are provided below.

Annotated Data

Using the Twitter API, we collected tweets of 6 randomly chosen trending topics determined by Twitter.com from March 1, 2011 to March 31, 2011. The topics fall into three categories – *News*, *Entity*, *Long-standing Topic* (*LST*) – that correspond to the three "topic types" defined by Zhao and Jiang (2011). We manually annotated all the 8613 tweets as one of *Sta* (statement), *Que* (question), *Sug* (suggestion), *Com* (comment), or *Mis* (Miscellaneous)[5]. The topics and tweet numbers are shown in Table 3[6].

Table 3: Topics and Numbers of Tweets

Category	Topic	# Tweets
News	Japan Earthquake (JE)	1742
	Libya Releases (LR)	1408
Entity	Dallas Lovato (DL)	677
	Nikki Taylor (NT)	786
LST	#100factsaboutme (FM)	2000
	#sincewebeinghonest (SH)	2000

Different categories/topics of tweets have different speech act distributions. Due to space limitations, we illustrate the speech act distributions in three topics, one for each category, in Figures 1 to 3.

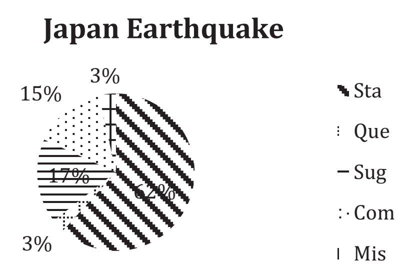

Figure 1: Speech Act Distribution for News: JE

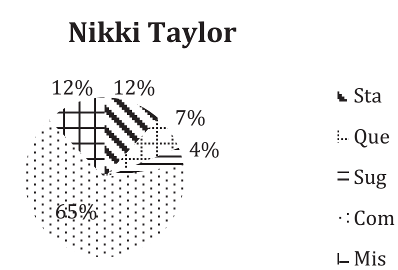

Figure 2: Speech Act Distribution for Entity: NT

[2] http://www.noswearing.com/dictionary
[3] For ethical concerns, we mask part of the words here and deliberately avoid using them in other examples.
[4] http://www.sharpened.net/emoticons/

[5] We will make the data and all our used resources publicly available.
[6] The two LST topics consist of 2000 tweets randomly sampled from the original collections, which are too large to be comparable to the other topics.

#sincewebeinghonest

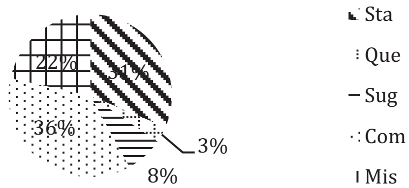

Figure 3: Speech Act Distribution for LST: SH

Our collected datasets demonstrate marked speech act distribution patterns. For example, the news topic "Japan Earthquake" is predictably dominated by statements but also contains many suggestions (e.g., about what people can do to help victims) that might interest some readers.

Experimental Setup

The raw text data were lightly preprocessed and the features were extracted by regular expression patterns. The only "normalization" is to restore the abbreviations and acronyms to find opinion words and vulgar words.

We did two sets of experiments. In the first set, we classified tweets in each topic using different features. In the second, we tried using different training and test data configurations on the topic, category, and whole dataset level and obtained interesting results. For all classifications, we split the annotated data into 90% training data + 10% test data and use the state-of-the-art SVM classifier with a linear kernel[7]. The multi-class problem is solved by using the one-against-the-rest paradigm.

Results and Discussions

Now we report the F1 (the harmonic mean of precision and recall) scores in various settings. Unless otherwise specified, all the results are from ten-fold cross validation.

Using Different Features

To find out what features are useful for speech act recognition, we experimented with cue words, non-cue words, symbols (character-based features), and all features. Table 4 lists the F1 scores on each speech act with different features, as weighted averages of the 6 topics according to tweet numbers. The "Avg" is the weighted average according to the number of each speech act type.

Table 4: F1 Scores for Different Features

Feature	Sta	Que	Sug	Com	Mis	Avg
Cue	0.788	0.455	0.554	0.623	0.422	0.668
Non-cue	0.671	0.088	0.068	0.355	0.074	0.447
Symbols	0.681	0.473	0.039	0.412	0.097	0.483
All	0.798	0.597	0.564	0.670	0.446	0.695

Treated individually, cue words and phrases are the best features overall, outstripping the second-best symbols by 38%. On each speech act, it defeats non-cue words and symbols with the only exception of "questions" because the punctuation *?* is a more reliable indicator of questions than question cue words. This proves the value of cue words for speech act recognition. With regard to the fact that our cue lexicon is only of a moderate size (531 unigrams, bigrams, and trigrams) based on limited data, better performance is anticipated with a larger and better cue lexicon, a joint product of statistical support from large Twitter corpora, well-designed Information Extraction patterns, and well-concerted human examination.

Character-based features (symbols) outperform non-cue features in almost all columns (with the only exception of "suggestion") and occasionally defeat cue features for reasons explained. Considering the fact that the non-cue lexicon contains more than 4000 tokens whereas we only have 5 symbols, the superiority of the latter becomes more noticeable. Although the non-cue words bear the characteristics of cyber English, most of them do not fit the Twitter text. On the other hand, the handful of symbols can capture speech act regularities better than words. Such evidence shows that the Twitter text has a distinct style and not all purported "noises" are noisy (e.g., *?!?).*

Without exception, using all features achieves better performance than using any one type of feature alone. Clearly all our proposed features contribute to the recognition of speech acts in Twitter. The weighted average F1 over all speech acts is nearly 0.7, achieved by our inexpensive method and little Twitter text normalization – even with no spelling correction or hashtag keyword splitting.

Recognition of statements and comments is conspicuously better than that of questions and suggestions. Whether this suggests that the former have more identifiable features than the latter remains to be tested against more data. Unsurprisingly, the recognition of "miscellaneous" is the worst because it is a heterogeneous group that may contain very different speech act types and non-speech acts. In future extensions of the work, we suggest breaking it down to more homogeneous subsets.

Using Different Training/Test Configurations

The various categories and topics of Twitter may be a barrier to speech act recognition and its application. If different categories/topics of Twitter data share little common ground in speech acts, we will have to sample data from each category/topic to do successful speech act recognition. Otherwise it is possible to select data regardless of Twitter category/topic or adapt a model trained on one category/topic to a different category/topic.

Drawing on the empirical conclusion from the first set of experiments, we classified speech acts using all the

[7] We used a Python wrapper of the LIBSVM tool: http://scikit-learn.sourceforge.net/modules/svm.html

proposed features but different training/test data configurations. The **homogenous** configurations use training and test data of the same kind on three levels: topic, category, and the whole dataset. The **heterogeneous** configurations use training and test data of different kinds on two levels: topic and category.

The last row in Table 4 is the average result of the homogenous topic-level configuration, which is now expanded with F1 scores on individual topics into Table 5. In this and the following tables, the "Dataset 1 : Dataset 2" notation means Dataset 1 is used for training and Dataset 2 is used for testing. If they are the same, a 9:1 training/test split is implied; otherwise the whole Dataset 1 and Dataset 2 are used for training and test, respectively.

Table 5: Homogenous Topic-level F1 Scores

	Sta	**Que**	**Sug**	**Com**	**Mis**	**Avg**
JE : JE	0.861	0.419	0.708	0.546	0.082	0.749
LR : LR	0.947	0.394	0.000	0.826	0.424	0.891
DL : DL	0.774	0.821	0.697	0.826	0.327	0.779
NT : NT	0.222	0.571	0.167	0.819	0.351	0.649
FM : FM	0.766	0.220	0.163	0.504	0.591	0.664
SH : SH	0.583	0.475	0.434	0.593	0.400	0.533
All	0.798	0.597	0.564	0.670	0.446	0.695

Our first observation is that the performance of speech act recognition is not proportionate to the amount of training data. Overall, speech act recognition on the news topics is better than that on the entity topics (with the least tweets), which in turn is better than that on the long-standing topics (with the most tweets). This can be attributed to the fact that news and entities (usually about celebrities or trendy concepts) are shaped by tweeters' collective behavior that can be mostly captured by our shallow features. In contrast, long-standing topics are typically composed of casual chatters that demonstrate varied individual behaviors that are hard to be captured by our shallow features. Moreover, tweets in such topics display more interpersonal speech acts, such as "promises" and "threats", which have been relegated to "miscellaneous" in our current scheme. This also explains why the "Mis" scores for FM and SH are relatively high among the topics.

After mixing the tweets from the same category (News = JE + LR, Entity = DL + NT, LST = FM + SH), we obtain the homogenous category-level F1 scores in Table 6.

Table 6: Homogenous Category-level F1 Scores

	Sta	**Que**	**Sug**	**Com**	**Mis**	**Avg**
News : News	0.901	0.508	0.715	0.692	0.097	0.810
Entity : Entity	0.651	0.780	0.590	0.812	0.246	0.716
LST : LST	0.401	0.698	0.358	0.473	0.387	0.549
All	0.673	0.705	0.581	0.629	0.335	0.673

The weighted averages of news and entity are very close to those based on individual topics (0.812 for news and 0.709 for entity), but the LST average is 10% lower than the topic-level average (0.599). This suggests that the news and entity topics share more common ground in speech acts than the long-lasting topics. The total category-level average, however, is comparable to the topic-level average.

The result of mixing all topics regardless of their categories is shown in Table 7.

Table 7: Homogenous Whole Dataset-level F1 Scores

	Sta	**Que**	**Sug**	**Com**	**Mis**	**Avg**
All : All	0.770	0.636	0.577	0.612	0.209	0.639

To our delight, the total average score is not much worse than the topic-level weighted average as the degradation is 8%. Moreover, the scores on "statements", "questions" and "suggestions" are even higher than the topic-level or category-level averages. Serious degradation is only observed for "miscellaneous". As mentioned before, this is not a clearly defined type and when data from different categories are mixed, it is even harder to determine miscellaneous speech acts. As this is the main performance bottleneck for the homogenous whole dataset-level configuration, we propose using a finer taxonomy of speech acts that fit the Twitter data in the future.

As we are also interested in how well a model trained on one topic/category can fit another topic/category, we first experimented on the topic-level training/test configurations, i.e., using data from different topics within the same category for training and test. The results are shown in Table 8. In this table and Table 9, the F1 scores are not the results from 10-fold cross validation because cross validation is not applicable to heterogeneous training/test configurations.

Table 8: Heterogeneous Topic-level F1 Scores

	Sta	**Que**	**Sug**	**Com**	**Mis**	**Avg**
LR : JE	0.798	0.427	0.032	0.435	0.046	0.575
JE : LR	0.905	0.235	0.316	0.538	0.000	0.792
NT : DL	0.096	0.688	0.036	0.627	0.059	0.414
DL : NT	0.346	0.490	0.197	0.524	0.000	0.425
SH : FM	0.692	0.290	0.250	0.465	0.470	0.592
FM : SH	0.529	0.104	0.112	0.362	0.353	0.386

Compared with Table 5 row by row, the least degradations are on SH/FM : SH (–10.8%) and JE/LR : JE (–11.1%) and the most are on DL/NT : DL (–46.8%) and NT/DL : NT (–34.4%). It seems that news and long-standing topics are more adaptable than entity topics in terms of speech act classification.

Finally, we tried different training/test configurations on the category level and report the result in Table 9.

Compared with Table 6, the least degradations are on News/Entity: News (–26.6%) and Entity/News : Entity (–29.2%) and the worst are on News/LST : News (–79.3%) and LST/Entity : LST (–52.6%). Therefore the best adaptability is observed between news and entity tweets, as is also reflected in the absolute average scores. The Long-standing category seems rather incompatible with news or entity, as the lowest average scores all include LST as training or test data.

Table 9: Heterogeneous Category-level F1 Scores

	Sta	Que	Sug	Com	Mis	Avg
Entity : News	0.709	0.338	0.427	0.368	0.038	0.594
LST : News	0.126	0.374	0.228	0.313	0.082	0.167
News : Entity	0.425	0.516	0.158	0.649	0.016	0.507
LST : Entity	0.275	0.554	0.404	0.480	0.185	0.420
News : LST	0.432	0.302	0.144	0.472	0.024	0.344
Entity : LST	0.246	0.258	0.222	0.446	0.046	0.260

Based on Tables 8 and 9, we find that training adaptability is generally better on the topic level than on the category level. Such empirical findings lead to the observation that in the absence of training data, a heterogeneous training/test configuration is feasible, with the annotation effort/classification performance trade-off at stake. The adaptability of a classification model is generally better if both training and test data are from the same category, especially for news and long-standing topics. If training and test data from different categories are to be used, they are preferably from news and entity. It is ill-advised to use long-standing data in a heterogeneous category-level training/test configuration. Rather, a topic-level training/test configuration within the long-standing category is more advisable.

Conclusion and Future Work

Speech acts are a window to tweeters' communicative behavior collectively or individually. Our preliminary work puts forth a new promising topic in content-oriented research of microtext such as Twitter text.

We address speech act recognition in Twitter as a multi-class classification problem by proposing a set of word-based and character-based features that can be easily harvested from raw data or free resources. Our inexpensive and robust method results in an average F1 of nearly 0.7.

Among the features, speech act cue words and phrases are most valuable. The easily extractable character-based symbols are also useful as they capture the unique textual feature of Twitter.

We have explored in-depth the effect of configuration of training and test data on classification performance. The empirical findings suggest that mixed training/test is nearly as successful as per-category or per-topic training/test, and that a heterogeneous training/test configuration is also feasible between different topics within one category or between different categories at some acceptable annotation effort/classification performance tradeoff.

In the future, we will construct a more complete set of typical speech acts in Twitter by breaking up the "miscellaneous" type. Research in a semi-supervised approach is also underway to overcome the problem of annotated data insufficiency. It is not clear whether alternative non-learning methods of finding speech acts as topics (e.g., LDA) will work, which we will also explore.

References

Austin, J. 1962. *How to Do Things with Words*. Oxford: Oxford University Press.

Baccianella, S., Esuli, A., and Sebastiani, F. 2010. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. *Proceedings of the Seventh conference on International Language Resources and Evaluation*

Crystal, D. 2006. *Language and the Internet, 2nd edition*. Cambridge, UK: Cambridge University Press

Dhillon, R., Bhagat, S., Carvey, H., and Shriberg, E. 2004. *Meeting Recorder Project: Dialog Act Labeling Guide*. Technical report, International Computer Science Institute.

Jeong, M., Lin, C-Y., and Lee, G. 2009. Semi-supervised Speech Act Recognition in Emails and Forums. In *Proceedings of EMNLP*, pages 1250–1259.

Jurafsky, D., Shriberg, E., and Biasca, D. 1997. *Switchboard SWBD-DAMSL Labeling Project Coder's Manual, Draft 13*. Technical report, University of Colorado Institute of Cognitive Science.

Kaufmann, M., and Kalita, J. 2010. Syntactic Normalization of Twitter Messages. In *Proceedings of ICON-2010: 8th International Conference on Natural Language Processing*.

Levinson, S. 1983. *Pragmatics*. Cambridge: Cambridge University Press.

Ritter, A., Cherry, C, and Dolan, B. 2010. Unsupervised Modeling of Twitter Conversations. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL*, pages 172–180.

Searle, J. 1975. Indirect speech acts. In P. Cole and J. Morgan (eds.), *Syntax and semantics, vol. iii: Speech acts* (pp. 59–82). New York: Academic Press.

Searle, J., and Vanderveken, D. 1985. *Foundations of Illocutionary Logic*. Cambridge: Cambridge University Press.

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

Wierzbicka, A. 1987. *English Speech Act Verbs: A Semantic Dictionary*. Orlando: Academic Press.

Wilson, T., and Wiebe, J. 2003. Identifying Opinionated Sentences. In *Proceeding of NAACL 03*, 33–34.

Zhao, X., and Jiang, J. 2011. *An Empirical Comparison of Topics in Twitter and Traditional Media*. Technical Report, Singapore Management University School of Information Systems.