Improving "Email Speech Acts" Analysis via N-gram Selection

Vitor R. Carvalho
Language Technologies Institute
Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh PA
vitor@cs.cmu.edu

William W. Cohen
Machine Learning Department
Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh PA
wcohen@cs.cmu.edu

Abstract

In email conversational analysis, it is often useful to trace the the intents behind each message exchange. In this paper, we consider classification of email messages as to whether or not they contain certain intents or email-acts, such as "propose a meeting" or "commit to a task". We demonstrate that exploiting the contextual information in the messages can noticeably improve email-act classification. More specifically, we describe a combination of n-gram sequence features with careful message preprocessing that is highly effective for this task. Compared to a previous study (Cohen et al., 2004), this representation reduces the classification error rates by 26.4% on average. Finally, we introduce Ciranda: a new open source toolkit for email speech act prediction.

1 Introduction

One important use of work-related email is negotiating and delegating shared tasks and subtasks. To provide intelligent email automated assistance, it is desirable to be able to automatically detect the *intent* of an email message—for example, to determine if the email contains a request, a commitment by the sender to perform some task, or an amendment to an earlier proposal. Successfully adding such a semantic layer to email communication is still a challenge to current email clients.

In a previous work, Cohen et al. (2004) used text classification methods to detect "email speech acts". Based on the ideas from Speech Act Theory (Searle, 1975) and guided by analysis of several email corpora, they defined a set of "email acts" (e.g., *Request*, *Deliver*, *Propose*, *Commit*) and then classified emails as containing or not a specific act. Cohen et al. (2004) showed that machine learning algorithms can learn the proposed email-act categories reasonably well. It was also shown that there is an acceptable level of human agreement over the categories.

A method for accurate classification of email into such categories would have many potential applications. For instance, it could be used to help users track the status of ongoing joint activities, improving task delegation and coordination. Email speech acts could also be used to iteratively learn user's tasks in a desktop environment (Khoussainov and Kushmerick, 2005). Email acts classification could also be applied to predict hierarchy positions in structured organizations or email-centered teams (Leusky, 2004); predicting leadership positions can be useful to analyze behavior in teams without an explicitly assigned leader.

By using only single words as features, Cohen et al. (2004) disregarded a very important linguistic aspect of the speech act inference task: the textual context. For instance, the specific sequence of tokens "Can you give me" can be more informative to detect a *Request* act than the words "can", "you", "give" and "me" separately. Similarly, the word sequence "I will call you" may be a much stronger indication of a *Commit* act than the four words separately. More generally, because so many specific

sequence of words (or n-grams) are inherently associated with the intent of an email message, one would expect that exploiting this linguistic aspect of the messages would improve email-act classification.

In the current work we exploit the linguistic aspects of the problem by a careful combination of n-gram feature extraction and message preprocessing. After preprocessing the messages to detect entities, punctuation, pronouns, dates and times, we generate a new feature set by extracting all possible term sequences with a length of 1, 2, 3, 4 or 5 tokens.

Using this n-gram based representation in classification experiments, we obtained a relative average drop of 26.4% in error rate when compared to the original Cohen et al. (2004) paper. Also, ranking the most "meaningful" n-grams based on Information Gain score (Yang and Pedersen, 1997) revealed an impressive agreement with the linguistic intuition behind the email speech acts.

We finalize this work introducing *Ciranda*: an open source package for Email Speech Act prediction. Among other features, Ciranda provides an easy interface for feature extraction and feature selection, outputs the prediction confidence, and allows retraining using several learning algorithms.

2 "Email-Acts" Taxonomy and Applications

A taxonomy of speech acts applied to email communication (email-acts) is described and motivated in (Cohen et al., 2004). The taxonomy was divided into *verbs* and *nouns*, and each email message is represented by one or more verb-noun pairs. For example, an email proposing a meeting and also requesting a project report would have the labels *Propose-Meeting* and *Request-Data*.

The relevant part of the taxonomy is shown in Figure 1. Very briefly, a *Request* asks the recipient to perform some activity; a *Propose* message proposes a joint activity (i.e., asks the recipient to perform some activity and commits the sender); a *Commit* message commits the sender to some future course of action; *Data* is information, or a pointer to information, delivered to the recipient; and a *Meeting* is a joint activity that is constrained in time and (usually) space.

Several possible verbs/nouns were not considered here (such as *Refuse*, *Greet*, and *Remind*), either because they occurred very infrequently in the corpus, or because they did not appear to be important for task-tracking. The most common verbs found in the labeled datasets were *Deliver*, *Request*, *Commit*, and *Propose*, and the most common nouns were *Meeting* and *deliveredData* (abbreviated as *dData* henceforth).

In our modeling, a single email message may have multiple *verbs-nouns* pairs.

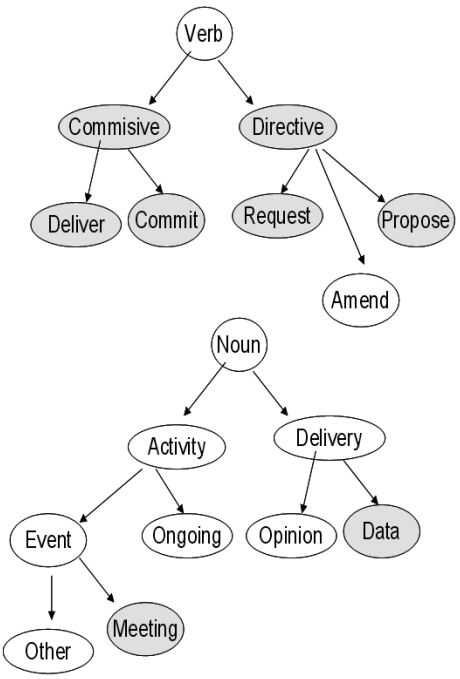

Figure 1: Taxonomy of email-acts used in experiments. Shaded nodes are the ones for which a classifier was constructed.

Cohen et al. (2004) showed that machine learning algorithms can learn the proposed email-act categories reasonably well. It was also shown that there is an acceptable level of human agreement over the categories. In experiments using different human annotators, Kappa values between 0.72 and 0.85 were obtained. The Kappa statistic (Carletta, 1996) is typically used to measure the human inter-rater agreement. Its values ranges from -1 (complete disagreement) to +1 (perfect agreement) and it is defined as (A-R)/(1-R), where A is the empirical probability of agreement on a category, and R is the probability of agreement for two annotators that

label documents at random (with the empirically observed frequency of each label).

3 The Corpus

The *CSpace* email corpus used in this paper contains approximately 15,000 email messages collected from a management course at Carnegie Mellon University. This corpus originated from working groups who signed agreements to make certain parts of their email accessible to researchers. In this course, 277 MBA students, organized in approximately 50 teams of four to six members, ran simulated companies in different market scenarios over a 14-week period (Kraut et al.,). The email tends to be very task-oriented, with many instances of task delegation and negotiation.

Messages were mostly exchanged with members of the same team. Accordingly, we partitioned the corpus into subsets according to the teams. The 1F3 team dataset has 351 messages total, while the 2F2, 3F2, 4F4 and 11F1 teams have, respectively, 341, 443, 403 and 176 messages. All 1716 messages were labeled according to the taxonomy in Figure 1.

4 N-gram Features

In this section we detail the preprocessing step and the feature selection applied to all email acts.

4.1 Preprocessing

Before extracting the n-grams features, a sequence of preprocessing steps was applied to all email messages in order to emphasize the linguistic aspects of the problem. Unless otherwise mentioned, all preprocessing procedures were applied to all acts.

Initially, forwarded messages quoted inside email messages were deleted. Also, signature files and quoted text from previous messages were removed from all messages using a technique described elsewhere (Carvalho and Cohen, 2004). A similar cleaning procedure was executed by Cohen et al. (2004).

Some types of punctuation marks (“,;:.)(][”) were removed, as were extra spaces and extra page breaks. We then perform some basic substitutions such as: from “*’m*” to “ *am*”, from “*’re*” to “ *are*”, from “*’ll*” to “ *will*”, from “*won’t*” to “*will not*”, from “*doesn’t*” to “*does not*” and from “*’d*” to “ *would*”.

Any sequence of one or more numbers was replaced by the symbol “[number]”. The pattern “[number]:[number]” was replaced with “[hour]”. The expressions “*pm* or *am*” were replaced by “[pm]”. “[wwhh]” denoted the words “*why*, *where*, *who*, *what* or *when*”. The words “*I*, *we*, *you*, *he*, *she* or *they*” were replaced by “[person]”. Days of the week (“*Monday*, *Tuesday*, ..., *Sunday*”) and their short versions (i.e., “*Mon*, *Tue*, *Wed*, ..., *Sun*”) were replaced by “[day]”. The words “*after*, *before* or *during*” were replaced by “[aaafter]”. The pronouns “*me*, *her*, *him*, *us* or *them*” were substituted by “[me]”. The typical filename types “*.doc*, *.xls*, *.txt*, *.pdf*, *.rtf* and *.ppt*” were replaced by “.[filetype]”. A list with some of these substitutions is illustrated in Table 1.

Symbol	Pattern
[number]	any sequence of numbers
[hour]	[number]:[number]
[wwhh]	“why, where, who, what, or when”
[day]	the strings “Monday, Tuesday, ..., or Sunday”
[day]	the strings “Mon, Tue, Wed, ..., or Sun”
[pm]	the strings “P.M., PM, A.M. or AM”
[me]	the pronouns “me, her, him, us or them”
[person]	the pronouns “I, we, you, he, she or they”
[aaafter]	the strings “after, before or during”
[filetype]	the strings “.doc, .pdf, .ppt, .txt, or .xls”

Table 1: Some PreProcessing Substitution Patterns

For the *Commit* act only, references to the first person were removed from the symbol [person] — i.e., [person] was used to replace “he, she or they”. The rationale is that n-grams containing the pronoun “I” are typically among the most meaningful for this act (as shall be detailed in Section 4.2).

4.2 Most Meaningful N-grams

After preprocessing the 1716 email messages, n-gram sequence features were extracted. In this paper, n-gram features are all possible sequences of length 1 (unigrams or 1-gram), 2 (bigram or 2-gram), 3 (trigram or 3-gram), 4 (4-gram) and 5 (5-gram) terms. After extracting all n-grams, the new dataset had more than 347500 different features. It would be interesting to know which of these n-grams are the “most meaningful” for each one of email speech acts.

1-gram	2-gram	3-gram	4-gram	5-gram
?	do [person]	[person] need to	[wwhh] do [person] think	[wwhh] do [person] think ?
please	? [person]	[wwhh] do [person]	do [person] need to	let [me] know [wwhh] [person]
[wwhh]	could [person]	let [me] know	and let [me] know	a call [number]-[number]
could	[person] please	would [person]	call [number]-[number]	give [me] a call [number]
do	? thanks	do [person] think	would be able to	please give give [me] a call
can	are [person]	are [person] meeting	[person] think [person] need	[person] would be able to
of	can [person]	could [person] please	let [me] know [wwhh]	take a look at it
[me]	need to	do [person] need	do [person] think ?	[person] think [person] need to

Table 2: Request Act:Top eight N-grams Selected by Information Gain.

One possible way to accomplish this is using some feature selection method. By computing the Information Gain score (Forman, 2003; Yang and Pedersen, 1997) of each feature, we were able to rank the most "meaningful" n-gram sequence for each speech act. The final rankings are illustrated in Tables 2 and 3.

Table 2 shows the most meaningful n-grams for the *Request* act. The top features clearly agree with the linguistic intuition behind the idea of a *Request* email act. This agreement is present not only in the frequent 1g features, but also in the 2-grams, 3-grams, 4-grams and 5-grams. For instance, sentences such as "What do you think ?" or "let me know what you ..." can be instantiations of the top two 5-grams, and are typically used indicating a request in email communication.

Table 3 illustrates the top fifteen 4-grams for all email speech acts selected by Information Gain. The *Commit* act reflects the general idea of agreeing to do some task, or to participate in some meeting. As we can see, the list with the top 4-grams reflects the intuition of commitment very well. When accepting or committing to a task, it is usual to write emails using "Tomorrow is good for me" or "I will put the document under your door" or "I think I can finish this task by 7" or even "I will try to bring this tomorrow". The list even has some other interesting 4-grams that can be easily associated to very specific commitment situations, such as "I will bring copies" and "I will be there".

Another act in Table 3 that visibly agrees with its linguistic intuition is *Meeting*. The 4-grams listed are usual constructions associated with either negotiating a meeting time/location ("[day] at [hour][pm]"), agreeing to meet ("is good for [me]") or describing the goals of the meeting ("to go over the").

The top features associated with the *dData* act in Table 3 are also closely related to its general intuition. Here the idea is delivering or requesting some data: a table inside the message, an attachment, a document, a report, a link to a file, a url, etc. And indeed, it seems to be exactly the case in Table 3: some of the top 4-grams indicate the presence of an attachment (e.g., "forwarded message begins here"), some features suggest the address or link where a file can be found (e.g., "in my public directory" or "in the etc directory"), some features request an action to access/read the data (e.g., "please take a look") and some features indicate the presence of data inside the email message, possibly formatted as a table (e.g., "[date] [hour] [number] [number]" or "[date] [day] [number] [day]").

From Table 3, the *Propose* act seems closely related to the *Meeting* act. In fact, by checking the labeled dataset, most of the *Proposal*s were associated with *Meeting*s. Some of the features that are not necessarily associated with *Meeting* are " [person] would like to", "please let me know" and "was hoping [person] could".

The *Deliver* email speech act is associated with two large sets of actions: delivery of data and delivery of information in general. Because of this generality, is not straightforward to list the most meaningful n-grams associated with this act. Table 3 shows a variety of features that can be associated with a *Deliver* act. As we shall see in Section 5, the *Deliver* act has the highest error rate in the classification task.

In summary, selecting the top n-gram features via Information Gain revealed an impressive agreement with the linguistic intuition behind the different email speech acts.

Request	Commit	Meeting
[wwhh] do [person] think	is good for [me]	[day] at [hour] [pm]
do [person] need to	is fine with [me]	on [day] at [hour]
and let [me] know	i will see [person]	[person] can meet at
call [number]-[number]	i think i can	[person] meet at [hour]
would be able to	i will put the	will be in the
[person] think [person] need	i will try to	is good for [me]
let [me] know [wwhh]	i will be there	to meet at [hour]
do [person] think ?	will look for [person]	at [hour] in the
[person] need to get	$[number] per person	[person] will see [person]
? [person] need to	am done with the	meet at [hour] in
a copy of our	at [hour] i will	[number] at [hour] [pm]
do [person] have any	[day] is fine with	to go over the
[person] get a chance	each of us will	[person] will be in
[me] know [wwhh]	i will bring copies	let's plan to meet
that would be great	i will do the	meet at [hour] [pm]
dData	**Propose**	**Deliver**
– forwarded message begins	[person] would like to	forwarded message begins here
forwarded message begins here	would like to meet	[number] [number] [number] [number]
is in my public	please let [me] know	is good for [me]
in my public directory	to meet with [person]	if [person] have any
[person] have placed the	[person] meet at [hour]	if fine with me
please take a look	would [person] like to	in my public directory
[day] [hour] [number] [number]	[person] can meet tomorrow	[person] will try to
[number] [day] [number] [hour]	an hour or so	is in my public
[date] [day] [number] [day]	meet at [hour] in	will be able to
in our game directory	like to get together	just wanted to let
in the etc directory	[hour] [pm] in the	[pm] in the lobby
the file name is	[after] [hour] or [after]	[person] will be able
is in our game	[person] will be available	please take a look
fyi – forwarded message	think [person] can meet	can meet in the
just put the file	was hoping [person] could	[day] at [hour] is
my public directory under	do [person] want to	in the commons at

Table 3: Top 4-grams Selected by Information Gain

5 Experiments

Here we describe how the classification experiments on the email speech acts dataset were carried out. Using all n-gram features, we performed 5-fold crossvalidation tests over the 1716 email messages. Linear SVM[1] was used as classifier. Results are illustrated in Figure 2.

Figure 2 shows the test error rate of four different experiments (bars) for all email acts. The first bar denotes the error rate obtained by Cohen et al. (2004) in a 5-fold crossvalidation experiment, also using linear SVM. Their dataset had 1354 email messages, and only 1-gram features were extracted.

The second bar illustrates the error rate obtained using only 1-gram features with additional data. In this case, we used 1716 email messages. The third bar represents the the same as the second bar (1-gram features with 1716 messages), with the difference that the emails went through the preprocessing procedure previously described.

The fourth bar shows the error rate when all 1-gram, 2-gram and 3-gram features are used and the 1716 messages go through the preprocessing procedure. The last bar illustrates the error rate when all n-gram features (i.e., 1g+2g+3g+4g+5g) are used in addition to preprocessing in all 1716 messages.

In all acts, a consistent improvement in 1-gram performance is observed when more data is added, i.e., a drop in error rate from the first to the second bar. Therefore, we can conclude that Cohen et al. (2004) could have obtained better results if they had used more labeled data.

A comparison between the second and third bars reveals the extent to which preprocessing seems to help classification based on 1-grams only. As we can see, no significant performance difference can be observed: for most acts the relative difference is

[1]We used the LIBSVM implementation (Chang and Lin, 2001) with default parameters.

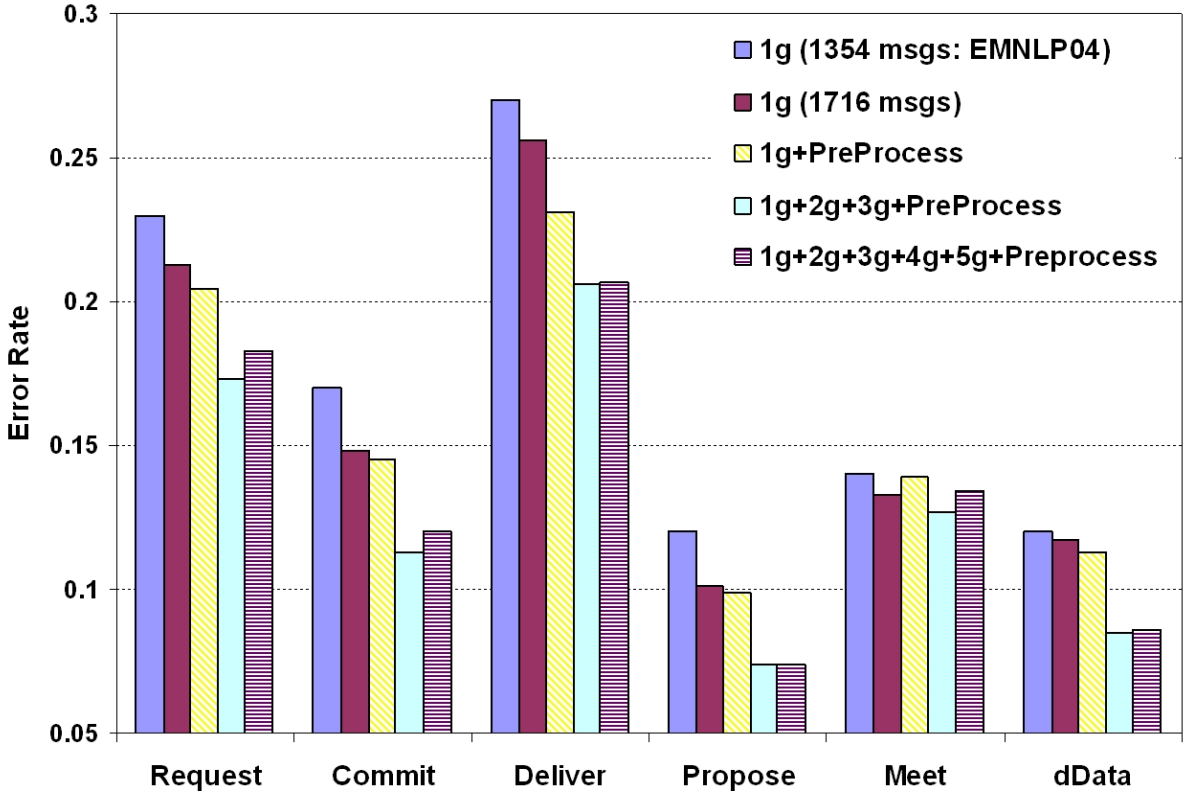

Figure 2: Error Rate 5-fold Crossvalidation Experiment

very small, and in one or maybe two acts some small improvement can be noticed.

A much larger performance improvement can be seen between the fourth and third bars. This reflects the power of the contextual features: using all 1-grams, 2-grams and 3-grams is considerably more powerful than using only 1-gram features. This significant difference can be observed in all acts. Compared to the original values from (Cohen et al., 2004), we observed a relative error rate drop of 24.7% in the *Request* act, 33.3% in the *Commit* act, 23.7% for the *Deliver* act, 38.3% for the *Propose* act, 9.2% for *Meeting* and 29.1% in the *dData* act. In average, a relative improvement of 26.4% in error rate.

We also considered adding the 4-gram and 5-gram features to the best system. As pictured in the last bar of Figure 2, this addition did not seem to improve the performance and, in some cases, even a small increase in error rate was observed. We believe this was caused by the insufficient amount of labeled data in these tests; and the 4-gram and 5-gram features are likely to improve the performance of this system if more labeled data becomes available.

Precision versus recall curves of the *Request* act classification task are illustrated in Figure 3. The curve on the top shows the *Request* act performance when the preprocessing step cues and n-grams proposed in Section 4 are applied. For the bottom curve, only 1g features were used. These two curves correspond to the second bar (bottom curve) and forth bar (top curve) in Figure 2. Figure 3 clearly shows that both recall and precision are improved by using the contextual features.

To summarize, these results confirm the intuition that contextual information (n-grams) can be very effective in the task of email speech act classification.

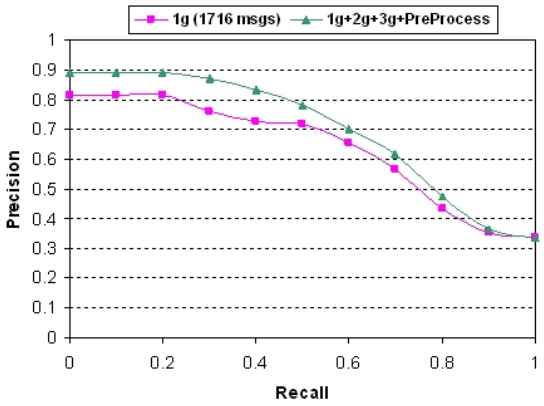

Figure 3: Precision versus Recall of the Request Act Classification

6 The Ciranda Package

Ciranda is an open source package for Email Speech Act prediction built on the top of the Minorthird package (Cohen, 2004). Among other features, Ciranda allows customized feature engineering, extraction and selection. Email Speech Act classifiers can be easily retrained using any learning algorithm from the Minorthird package. Ciranda is currently available from `http://www.cs.cmu.edu/~vitor`.

7 Conclusions

In this work we considered the problem of automatically detecting the intents behind email messages using a shallow semantic taxonomy called "email speech acts" (Cohen et al., 2004). We were interested in the task of classifying whether or not an email message contains acts such as "propose a meeting" or "deliver data".

By exploiting contextual information in emails such as n-gram sequences, we were able to noticeably improve the classification performance on this task. Compared to the original study (Cohen et al., 2004), this representation reduced the classification error rates by 26.4% on average. Improvements of more than 30% were observed for some acts (*Propose* and *Commit*).

We also showed that the selection of the top n-gram features via Information Gain revealed an impressive agreement with the linguistic intuition behind the different email speech acts.

References

[Carletta1996] Jean Carletta. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254.

[Carvalho and Cohen2004] Vitor R. Carvalho and William W. Cohen. 2004. Learning to extract signature and reply lines from email. In *Proceedings of the Conference on Email and Anti-Spam*, Palo Alto, CA.

[Chang and Lin2001] Chih-Chung Chang and Chih-Jen Lin, 2001. *LIBSVM: a library for support vector machines*. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[Cohen et al.2004] William W. Cohen, Vitor R. Carvalho, and Tom M. Mitchell. 2004. Learning to classify email into "speech acts". In *Proceedings of Empirical Methods in Natural Language Processing*, pages 309–316, Barcelona, Spain, July.

[Cohen2004] William W. Cohen, 2004. *Minorthird: Methods for Identifying Names and Ontological Relations in Text using Heuristics for Inducing Regularities from Data.* `http://minorthird.sourceforge.net`.

[Forman2003] George Forman. 2003. An extensive empirical study of feature selection metrics for text classification. *The Journal of Machine Learning Research*, 3:1289–1305.

[Khoussainov and Kushmerick2005] Rinat Khoussainov and Nicholas Kushmerick. 2005. Email task management: An iterative relational learning approach. In *Conference on Email and Anti-Spam (CEAS'2005)*.

[Kraut et al.] R.E. Kraut, S.R. Fussell, F.J. Lerch, and A. Espinosa. Coordination in teams: Evidence from a simulated management game. To appear in the Journal of Organizational Behavior.

[Leusky2004] Anton Leusky. 2004. Email is a stage: Discovering people roles from email archives. In *ACM Conference on Research and Development in Information Retrieval (SIGIR)*.

[Searle1975] J. R. Searle. 1975. A taxonomy of illocutionary acts. In *In K. Gunderson (Ed.), Language, Mind and Knowledge.*, pages 344–369, Minneapolis, MN. University of Minnesota Press.

[Yang and Pedersen1997] Yiming Yang and Jan O. Pedersen. 1997. A comparative study on feature selection in text categorization. In *Proceedings of ICML-97, 14th International Conference on Machine Learning*, pages 412–420.