Email is a Stage: Discovering People Roles from Email Archives

Anton Leuski
Institute for Creative Technologies
13274 Fiji Way, Marina del Rey, CA 90292
leuski@ict.usc.edu

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: Content Analysis and Indexing; H.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Experimentation, Human Factors

1. INTRODUCTION

Imagine a social scientist looking at the National Security Council emails for background information on how a policy decision was made; imagine a biographer accessing an email archive of a prominent scientist to find her role in a seminal discovery; or imagine an individual pondering over his own personal email collection to remember who was responsible for what part of the project he participated in three years ago. In all these scenarios the person making the analysis wants to determine who were the key actors, what roles did they play, what actions did they take, and what was the outcome of those actions.

We are interested in developing automatic techniques that would allows us to extract such information from email collections. In this paper we focus on one aspect of this problem: detecting the people roles. A single person can "play" different "personas" almost at the same time: e.g., she is a graduate student and a research assistant and a friend and so on. These roles are reflected in the content of all her communications with the outside world and thus in her email messages as well. We want to detect those roles so we can determine the relationship between the actors, – e.g., who started a project and who was responsible in bringing it to completion. We also want to treat emails created by the same person in different roles separately, – e.g., we would like to separate her personal emails from the professional ones.

We view a collection of emails as a multi-party dialog where each message carries one or several social actions or "speech acts" [4]. We believe that an individual's role can be detected by analyzing the patterns of the speech acts in the her incoming and outgoing emails. This paper consists of two parts. We start by describing an automatic classifier that allows us to detect the speech acts in email messages. We then present a study on how these speech acts can characterize a person's role.

SIGIR'04, July 25–29, 2004, Sheffield, South Yorkshire, UK.
ACM 1-58113-881-4/04/0007.

2. SPEECH ACTS

Electronic mail is an ubiquitous communication medium that carries an enormous amount of information [3]. Despite the wide spread of email, obtaining a useful collection of email proved to be a very difficult task due to significant privacy concerns about the use of such a collection. People are very reluctant to part with their emails. Presently we know of only one publicly available collection of organizational email [1].

Our internal efforts to collect emails resulted in approximately 500 email messages from 5 people in our research group. These messages are either sent or received by one of those people to or from somebody either inside or outside of the research group. An incoming message can be directly addressed to the person or be a copy of the message broadcasted to a group of people. The messages vary in size from a one word confirmation note to couple pages of a research plan description.

We hand-tagged the messages with 8 speech acts as defined in Table 1. Note that one message can be assigned multiple speech acts. For example, if someone reports on a completed task and asks what to do next, we tagged the message with both "provide information" and "request advice".

Table 1: Speech act statistics.

speech act	example	count
plan	We are going to do ...	10
request advice	What should I do next?	11
request meeting	Let meet and discuss this.	29
request action	Please reserve a room	96
request info	Do you have the url?	127
provide info	Here is the url you wanted	334

We processed the collection to remove the headers, signatures, and all quoted text from every email. The resulting message texts were stemmed and stopped. We extracted all unigrams, bigrams, and trigrams that ap-

peared more than twice in the collection and used them as features to create a feature vector for every message. The features were weighted using the standard $tf \times idf$ schema.

We trained a single Support Vector Machine (SVM) classifier for every speech act class using the SVM^{Light} package [2]. We used 10-fold cross validation to test the performance of the classifier. Table 2 shows the precision, recall, and accuracy numbers for the classifiers created for the four largest speech act classes.

Table 2: Accuracy numbers for four the most frequent speech acts.

speech act	Precision	Recall	Accuracy
request meeting	0.87	0.96	0.99
request action	0.97	0.74	0.92
request info	0.67	0.72	0.84
provide info	1.00	0.86	0.88
average	0.87	0.82	0.91

The results in Table 2 indicate that we can automatically detect speech acts in email messages with a high accuracy. The next step is to see if we can use the speech act information to determine the people roles.

3. ROLES

The positions (or roles) of the five people that shared their emails with us are: "professor, head of the research group", "graduate student", "secretary", "researcher", and "programmer". Assuming that the speech act classes are independent, we computed the normalized email activity per person per speech act: for every speech act we took the number of emails with the speech act sent or received by the person and divided it by the total number of emails of that person we had in the collection. Averaging this normalized email activity across all people gives us the expected likelihood of observing a particular speech act in a person's mailbox and the baseline for our analysis. The standard deviation of the sample serves as the comparison scale. If the actual number of emails with the speech act differs from the average by more than one standard deviation, we consider that an important feature of the person's role.

We collected all the instances of high and low speech act occurrences in people mailboxes in Table 3. There "+" indicates a significantly high amount of the particular speech act class in either incoming or outgoing email. Conversely, "–" indicates a significantly low amount.

We showed a version of the table that had no names attached to the columns to three people in our research group, described the experiment to them, told the names of the people from whom we collected the emails, and asked them to assign each person to a column in the table. Our judges were quick to solve the problem with an average 91.7% agreement among them and the ground truth. It gives us a good indication that pattern analysis of incoming and outgoing emails with different speech acts may allow us to detect and define people roles. We believe that given enough training data we can develop an automatic classifier that will handle the role detection analysis.

Table 3: Unusual email activity for five people with different roles arranged by speech act.

people	1	2	3	4	5
incoming email					
plan					–
request advice	+				–
request meeting	+			+	–
request action			+		
request information				–	+
provide information			–		
outgoing email					
plan	+				
request advice		+			
request meeting		+			
request action	+			–	
request information				–	+
provide information				+	

4. FUTURE WORK

Our present goal is to get more email data. We are searching for a sufficiently large collection of organizational email. The collection of the National Security Council email from 1985-1987 [1] is an example of a collection that was made public through an official process. It consists of approximately 300 messages. We are talking to the National Security Archives to see if they have a larger collection available.

Another possibility is to continue collecting emails from the people in our research group. While it seems like the easiest course of action, it is not: our experience suggests that people are very reluctant to share their emails even for confidential research purposes. It also suffers from a very limited use of the final collection: the privacy concerns will never allow for such a collection to be made public and therefore it will be impossible to repeat the experiments.

5. REFERENCES

[1] T. Blanton, editor. *White House E-Mail: the top secret computer messages the Reagan-Bush White House tried to destroy*. New Press, New York, 1995.

[2] T. Joachims. Transductive inference for text classification using support vector machines. In *Proceedings of International Conference on Machine Learning (ICML)*, 1999.

[3] P. Lyman and H. R. Varian. How much information, 2000. Retrieved on 11/10/02 from `http://www.sims.berkeley.edu/research/projects/how-much-info/internet/emaildetails.html`.

[4] J. Searle. *Speech acts: An essay in the philosophy of language*. Cambridge University, Cambridge, England, 1969.