Context Features in Email Archives

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

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 organizing "that workshop" three years ago. In all these scenarios the information about the people who send and received the emails is a vital part of the information seeking process. A model of writers and readers of the messages as well as the links between them is an important context for understating and interpreting the content of the emails.

The following list is an enumeration of some of the properties for that model:

- **Role** A single person "plays" different roles or "personas" at different moments of her life, often these roles exist almost at the same time: e.g., she is a graduate student and a research assistant and a music lover and a cooking expert and a friend and so on. Knowing the role of the sender or the addressee may have a profound effect on interpreting the content of the message. For example, we might want to treat emails created by the same person in different roles separately and separate her personal emails from the professional ones.

- **Relationship** A relationship is a link between a pair of roles: e.g., if A is a research adviser and B is a student, then knowing that A supervises B is an important connection between the two personas. For example, if we want to trace how a particular project was started we might want to look at messages that went from A to B. If, on the other hand we want to find out the results of the project we should look at the message from B to A. Note that the roles and relationships are relative: person A might be a supervisor of person B but at the same time she could be subordinate to person C.

- **Expertise** Understanding how well a person knows a particular topic may affect our belief in the information she provides. We are likely to trust the message content more if we know that the sender is an expert on the subject of the email.

- **Trust** Knowing that somebody is an expert is not sufficient unless we also trust her expertise.

The preceding list is by no means exhaustive. It serves as a way of focusing our attention on that email is first of all is a communication media and it is important to understand the people who are sending the messages, their intention, knowledge and expertise. To model these aspects of interaction we turn to the speech act theory.

Speech act theory is built on the foundation laid by Wittgenstein and Austin [1]. Ludwig Wittgenstein began a line of thought called "ordinary language philosophy." He taught that the meaning of language depends on its actual use. Language, as used in ordinary life, is a language game because it consists of rules. In other words, people follow rules to do things with the language.

According to Searle [6], to understand language one must understand the speakers intention. Since language is intentional behavior, it should be treated like a form of action. Thus Searle refers to statements as *speech acts*. The speech act is the basic unit of language used to express meaning, an utterance that expresses an intention. Normally, the speech act is a sentence, but it can be a word or phrase as long as it follows the rules necessary to accomplish the intention. When one speaks, one performs an act. Speech is not just used to designate something, it actually does something. Speech act stresses the intent of the act as a whole. According to Searle, understanding the speakers intention is essential to capture the meaning. Without the speakers intention, it is impossible to understand the words as a speech act.

There exist several taxonomies of speech acts. For example, Searle defined four types of speech acts: utterance acts, propositional acts (referring is a type of

propositional act), illocutionary acts (promises, questions and commands) and perlocutionary acts (to elicit some behavioral response from the listener). Bach and Harnish [2] created a hierarchy of 38 different speech acts that fall into four general categories: constatives, directives, commissives and acknowledgments (e.g., statements, requests, promises and apologies accordingly).

We believe that a person's role, her relationship to other people, her expertise and trust is defined by her actions and therefore reflected in her language. We expect that we can model these properties by analyzing patterns of speech acts in her communications, i.e., in her incoming and outgoing emails. Our intention is to use statistical classification techniques to detect and assign the speech acts to individual email messages by analyzing the content of the email. Then we plan to determine persistent patterns of speech acts among multiple mailboxes to define the roles.

Using the same speech acts we can model the other context properties described at the beginning of this paper. For example, to define a relationship between two people, we need to analyze their roles and the patterns of speech acts in the emails they exchange. We can determine a person's expertise by combining topic-based clustering with the speech act analysis: e.g., we count how often a person sends out *statements* on a particular topic. Consequently, trust can be defined as a proportion of all requests on a particular topic that end up in the person's inbox.

In the rest of the paper I summarize some experiments on detecting speech acts and persons' roles. These experiments were done on a small email corpus of approximately 500 messages collected from five members of our research group. I have presented the results of that study at the last year SIGIR [5]. During this year we have extended the study to the publicly available Enron collection [4] of approximately 300,000 messages. We used a more complex taxonomy of speech acts adapted from the work by Bach and Harnish [2]. We observed a similar performance in detecting the individual speech acts. Presently we are working on a system that would use speech act patterns to determine a person's role.

2. EXPERIMENTS

We hand-tagged the messages with 6 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".

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 appeared 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 [3]. We used 10-fold cross validation to test the performance of the classifier. We observed 87% precision and 82% recall on average across 4 largest speech act classes.

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

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 2. There "+" indicates a significantly high amount of the particular speech act class in either incoming or outgoing email. Conversely, "–" indicates a significantly low amount.

Table 2: 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				+	

The patterns in table 2 can be interpreted the following way: The first person (person "1") is getting asked for

advice quite a lot and he often sends out requests for action and plan kind of messages, which looks like what a supervisor would do. He is the professor and the head of the research group. The third person ("3") receives a lot of requests for action and other people do not ask her for information – she is the secretary. The second person ("2") often needs advice and wants to meet with other people – he is the student. Interpreting the other two patterns is left as an exercise for the reader.

3. REFERENCES

[1] J. L. Austin. *How to do Things with Words.* Oxford, 1962.

[2] K. Bach and R. M. Harnish. *Linguistic Commuication and Speech Acts.* MIT Press, Cambridge, Mass., 1979.

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

[4] B. Klimt and Y. Yang. Introducing the enron corpus. In *Proceedings of the First Conference on Email and Anti-Spam (CEAS)*, Mountain View, CA, 2004.

[5] A. Leuski. Email is a stage: discovering people roles from email archives. In *Proceedings of 27th annual international ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'04)*, pages 502–503, Sheffield, United Kingdom, 2004. ACM Press.

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