Linguistic Cues to Social Information
Description of Proposed Research

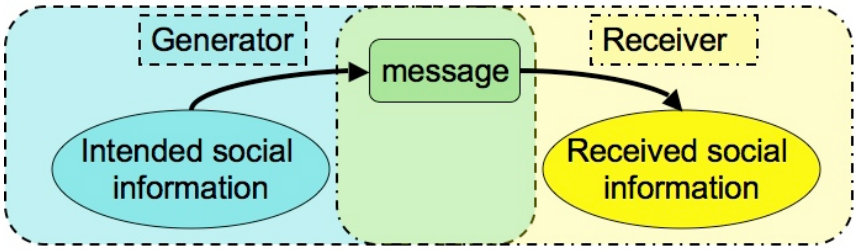
Principal Investigator:
Professor Lisa Pearl
Department of Cognitive Sciences
University of California, Irvine, CA 92697-3435
Email: lpearl@uci.edu

Co-Investigator:
Professor Mark Steyvers
Department of Cognitive Sciences
Email: msteyver@uci.edu

Co-Investigator:
Professor Padhraic Smyth
Department of Computer Science
Email: smyth@ics.uci.edu

1. Introduction

Blogs, emails, text messages and voice communications provide a rich data source to study the patterns of communication between people. While many information extraction techniques have been developed to automatically analyze text, the current state of technology is constrained to analyze the content and form of messages at the surface level. A major gap in current technology is the ability to reliably identify social information from text. By social information, we refer to subtle information inherent in text such as a speaker's motives (e.g. intent to persuade, intent to deceive), emotional states (e.g. anger or embarrassment from talking about a taboo topic), attitudes (e.g. disbelief, neutrality), and position in a social network (e.g. higher or lower social status). Interestingly, while humans can often effortlessly identify this information from language, currently no information extraction systems are available that are capable of doing the same. The figure to the left demonstrates the process of social information transmission from a generator (e.g. the speaker) to a receiver (e.g. someone listening to the speaker) via a message consisting of language text. The receiver must, in effect, reconstruct the generator's mental state in order to infer what social information was intended by the generator's message. The clues the receiver uses to do this are the linguistic cues in the message itself and the social information these cues are typically associated with, based on the receiver's experience.



The critical technical barrier for the development of social information extraction systems is a lack of a data – there are no large text databases annotated with social information. Because of this lack of data, standard machine learning techniques cannot be used to identify useful linguistic features. Thus, the lack of data leads to an accompanying lack of knowledge about the relevant features for identifying this social information. We propose to address this gap in current technology with a new approach in computational social science, integrating social science theory, linguistics, and machine learning techniques in order to create both the necessary large-scale database and the computational techniques to automatically identify the social information available in text. Our proposed research can be divided into two main tasks:

- **Task 1:** Creation of a large-scale database of text that is labeled with the social information being expressed, including messages that are easily identified as reflecting specific social information and messages likely to be confused as reflecting different social information than intended. Using game environments, the social information in messages will be identified through the collective effort of a large number of humans. In this *human computation* approach, we rely on the social skills of a large number of individuals for the creation as well as interpretation of text messages related to various social information.

- **Task 2:** Application of current machine learning techniques to learn what linguistic cues can be used to identify social information in text. Because the database developed in task 1 has a known ground truth (messages labeled definitively with specific social information), the predictive accuracy of the classifiers can be assessed in a cross-validation procedure, such as training on part of the text with observed social labels and testing on the remaining text. In addition, these classifiers can be used to identify social information in real-world data sets, such as open-source document collections like blogs, online discussion forums, and Twitter.

The research results will be documented in conference/journal papers, and the database made available to the scientific community. There are several impacts for the proposed research. First, this research will lead to the creation of the first large-scale text database with a known ground truth of relevant social information, which should jumpstart research in the area of social information extraction. Secondly, the human computation research can also yield examples of messages that are likely to be confused as indicating social information other than what was intended, which provides cues that lead to communication failure in humans. Thirdly, the application of machine learning techniques can lead to computational models that automatically identify social information from text, which are useful for automatically assessing the social information inherent in a message or conversation. All three of these results will be of great interest to extra-mural funding sources such as DARPA, IARPA, AFRL/AFOSR, and NSF. This proposal therefore serves as a vital stepping stone for future proposals for funding agencies such as these.

2. Creating a large database of text annotated with social information

Background. In machine learning research, reliable databases are generally required in order to develop reliable algorithms. Resources such as the Linguistic Data Consortium (LDC) at the University of Pennsylvania are dedicated to providing researchers with large quantities of natural language data on which to base their research. Unfortunately, very few databases annotated with social information exist, and the few that do are small in size. A recent addition to the LDC demonstrates this: the Language Understanding Annotation Corpus (LUAC) (Diab *et al.* 2009) includes text annotated with "committed belief", which "distinguishes between statements which assert belief or opinion, those which contain speculation, and statements which convey fact or otherwise do not convey belief." This is meant to aid in determining which beliefs can be ascribed to a communicator and how strongly the communicator holds those beliefs. Nonetheless, this is still a small sample of the possible social information contained in text, which includes a communicator's motives, attitudes, emotional states, and position in a social network. Moreover, the LUAC contains only about 9000 words across two languages (6949 English, 2183 Arabic), which is small compared to the corpora generally available for natural language processing (e.g. the English Gigaword corpus (Graff, 2003) contains 1756504 words).

Another tack taken by researchers has been to use open-source data that are likely to demonstrate certain social information by happenstance, e.g. online gaming forums with games that happen to involve the intent to deceive (e.g. Zhou, 2008: Mafia game forums). While these data sets are larger in size, they do not have the breadth of coverage in terms of what social information they can capture because, by nature, the games only explicitly involve one kind of social information (e.g. deception); other social information cannot reliably be attributed to the text. So, the relation between other types of social information (e.g. confidence, frustration, disbelief) and language text is unknown.

Database creation. A main innovation of this project is to use human computation for creating the database, where we leverage the knowledge contained in a population. People can often detect social information that is conveyed through text. For example, consider the following message: "*Clearly, sir, we must increase recruitment*." Given only the text itself, we can infer (a) the speaker intends to persuade the listener, and (b) the speaker is likely speaking to someone who has higher social status. Our data collection procedure utilizes this ability and is similar to the idea of *games with a purpose* (GWAPs) (von Ahn, 2006). GWAPs are currently being used to accumulate information about many things that

humans find easy to identify, such as objects in images, the musical style of songs, and common sense relationships between concepts (von Ahn, 2006). We intend to develop a web-based GWAP that creates a database of messages annotated with social information. In this game, unpaid participants provide knowledge about the social information in text. Because the data collection procedure occurs in the context of a game, large quantities of data can be accrued from a variety of sources.

The GWAP will encourage participants to both generate messages that reflect specific social information and to label messages created by other participants as reflecting specific social information. For example, one participant might generate the message "*Clearly, sir, we must increase recruitment*" for the social information of "persuading"; a different participant would see this message and label it as an example of "persuading". With enough game players, many messages will be created that clearly reflect different social information. Without any of the participants necessarily having expert knowledge or training, we expect that the cumulative knowledge will be quite reliable. For example, the same text can be evaluated by many different people to reduce the effect of idiosyncratic responses from a few individuals.

A great advantage of this database will be that many different kinds of social information can be labeled all at once. We can gauge how clearly a message reflects social information by how often it is labeled by others as reflecting that social information. In addition, by the very nature of the GWAP, we can also assess which social information is easily confused by humans, e.g. politeness with embarrassment, or confidence with deception. This will aid the development of models that extract social information and can also provide helpful information to analysts who wish to identify social information from text, e.g. by sending alerts that certain messages are likely to be ambiguous. Moreover, these confusing messages provide insight into cues that humans rely on (in these cases, erroneously) to assess social information from language text. Thus, by identifying confusing messages, we can both design computational models capable of avoiding the mistakes humans make and assess the social background knowledge humans bring to language interpretation.

Initially, the GWAP will run as a separate application that participants play on a computer in a social science laboratory (the offline version) or which they navigate to on the web (the online version). For web-based participation, participants will be encouraged to play via social networking sites such as facebook, livejournal, and myspace. Later, the game can be integrated into the social networking sites that have game capability, such as facebook, to increase participation.

Pilot study. We have already collected pilot data with the offline version of the GWAP that involves eight types of social information indicative of several social aspects: politeness (indicates social status), rudeness (indicates attitude, taboo topics), embarrassment (indicates taboo topics), formality (indicates social status), persuading (indicates intent to persuade), deception (indicates intent to deceive), confidence (indicates attitude), and disbelief (indicates attitude). So far, participants have created 290 messages and annotated 270 of them. Some sample messages that are correctly and incorrectly identified are shown below:

Social Dimension		*Message*
Generated	*Received*	
deception	deception	"Oh yeah...your hair looks really great like that...yup, I love it...it, uh, really suits you..."
embarrassment	embarrassment	"Oh... we're not dating. I would never date him... he's like a brother to me.."
rudeness	persuading	"James, Bree doesn't like you. She never did and never will!"
deception	persuading	"Little Sara, its okay. Don't listen to what those older boys told you. Santa Clause does exist."

In addition to identifying messages that are good examples of expressing particular social information, we can also identify types of messages that are likely to be confused. Below we see a confusion matrix of social information. The table shows the likelihood that a message will be received as expressing specific social information (in the columns), given that it has been generated with specific social information in mind (in the rows). In other words, we show the probability distribution *p*(*received* | *generated*). What we see from the pilot data is that people are more likely to correctly identify a message expressing rudeness (p = .72) and confidence (p = .71) and less likely to correctly identify a message expressing deception (p = .38) or formality (p = .36). Also, we can see that a deceptive message can be mistaken for a persuading message (p= 0.25), a message expressing disbelief can be mistaken for a message expressing deception (p = .14) or confidence (p = .18), and a persuading message can be mistaken for a deceptive message (p = .13) or confidence (p = .08), among other observations. Some of these may be expected, e.g. confidence with persuading since someone who is trying to persuade will likely be confident about the topic. Others may be unexpected a priori, such as mistaking disbelief for deception.

	deception	politeness	rudeness	embarrassment	confidence	disbelief	formality	persuading
deception	.38	.08	.13	.03	.03	.08	.05	.25
politeness	.03	.58	.03	.06	.03	.00	.17	.11
rudeness	.10	.00	.72	.00	.00	.14	.00	.03
embarrassment	.03	.10	.13	.60	.00	.10	.03	.00
confidence	.00	.06	.00	.00	.71	.03	.00	.20
disbelief	.14	.04	.07	.00	.18	.57	.00	.00
formality	.00	.29	.07	.04	.04	.00	.36	.21
persuading	.13	.03	.05	.03	.08	.05	.05	.59

The goal is to gather as much data as possible, primarily using the web. Based on von Ahn (2006), we might expect (using rough estimates) 1000 people to play every month. We might also expect a person to create 10 messages and interpret 100 messages in a month, leading to about 10 annotations per message. This gives an average of 10,000 annotated messages per month that the GWAP runs. Messages that are reliably interpreted will make up some proportion of the dataset, between 30 and 80% of the messages annotated, according to our pilot results. Thus, we estimate between 3000 and 8000 annotated messages for the database per month that the GWAP runs. Messages reliably confused as expressing different social information than intended will comprise a smaller portion of the annotated messages.

3. Developing and deploying machine-learning classifiers for social information

Identifying linguistic cues. Once a database of sufficient size has been created, we can identify linguistic cues that signal social information. Linguistic cues for identifying information in text have often been at the word-level in prior research. For example, positive and negative affect words (e.g. *lovely* vs. *stupid*) have been used in sentiment analysis to summarize whether a document is positive or negative. In deception detection research, informative word-level cues include counting first and third person pronoun usage (e.g. *me* vs. *them*) (Anolli *et al.* 2002), and noting the number of "exception words" (e.g. *but, except, without*) (Gupta & Skillicorn 2006). In addition, informative shallow text properties have also been identified (Zhou *et al.* 2004), such as (a) number of verbs, words, noun phrases, and sentences, (b) average sentence and word length, and (c) word type to word token ratio.

As we recognize the potential diagnostic power of both word-level and shallow text properties, we propose to explore both these kinds of linguistic cues as well as cues corresponding to deeper levels of linguistic structure. Some examples of potential deeper-level cues include whether optional arguments are given for verbs (e.g. *lunch* is an optional argument for the verb *eat*) and the use of subordinate clauses (e.g. *if I did it* is a subordinate clause in *If I did it, he did it, too*).

The design of the GWAP itself can help generate messages likely to exhibit deeper linguistic cues by controlling what kinds of messages participants are allowed to express. For example, initially participants have very few restrictions on what words they can use to express a particular social goal. However, later stages of the GWAP can prohibit the use of obvious words in order to force participants to rely on more subtle properties of the language to communicate specific social information. For example, for the social

information of "politeness", a participant might initially use words like *please* and *thank you* in a successful message (e.g. "*I'd like these, please"*); if these words are later prohibited, a participant might rely more on subtler properties such as using subordinate clauses in order to express the same social information successfully (e.g. "*If you could be so kind as to get these for me, I'd appreciate it."*). In this way, we can generate example messages for the database that use less obvious linguistic cues, and then determine which deeper-level cues are likely to indicate which social information.

Building classifiers. We can utilize a number of supervised machine learning techniques to build classifiers for the purpose of predicting the original social information associated with a message. This will help in building systems that can automatically identify social information from text. We can also build classifiers that predict how the message might actually be perceived (and potentially misperceived) by a human. This second application is useful because it will identify messages that might lead to consistent misinterpretation by humans.

The input to these classifiers consists of a large number of linguistic features such as presence/absence of specific words, syntactic features (e.g. argument structure, subordinate clauses), semantic features (e.g. WordNet features like semantically similar words), as well stylistic features (e.g. constructions like *if you please*). A number of classifiers can be compared such as sparse logistic regression (which is able to handle a large number input features), support vector machines, and baseline models such as naïve bayes.

Deploying classifiers. Once we have developed classifiers using the social information database, we can deploy them on unannotated data sets, such as blogs, online discussion forums, and Twitter. The classifiers will automatically indicate what social information a message is trying to convey, and indicate messages that are likely to be confused by people. Using the classifier, we will be able to generate a list of the top-ranked messages for particular social information (e.g. "intent to persuade"). We can assess whether these top-ranked messages accord with human intuition by having trained annotators decide what social information a top-ranked message is expressing and how well that message expresses it. This allows us to improve the performance of the classifier as well as add new annotated messages to the social information database.

4. References to Supporting Material

Anolli, L., Balconi, M., and Ciceri, R. (2002). Deceptive Miscommunication Theory (DeMiT): A New Model for the Analysis of Deceptive Communication. In Anolli, L., Ciceri, R. and Rivs, G. (eds)., *Say not to say: new perspectives on miscommunication,* 73-100. IOS Press.

Diab, M., Dorr, B., Levin, L., Mitamura, T., Passonneau, R., Rambow, O., and Ramshaw, L. (2009). Language Understanding Annotation Corpus. *LDC*, Philadelphia.

Graff, D. (2003). English Gigaword. *Linguistic Data Consortium*, Philadelphia.

Gupta, S. & Skillicorn, D. (2006). Improving a Textual Deception Detection Model, *Proc. of the 2006 conf. of the Center for Advanced Studies on Collaborative research.* Toronto, Canada.

Von Ahn, L., Kedia, M. and Blum, M. (2006). Verbosity: A Game for Collecting Common-Sense Facts, *In Proceedings of the SIGCHI conference on Human Factors in computing systems*, Montréal, Québec, Canada.

Zhou, L., Burgoon, J., Nunamaker, J., and Twitchell, D. (2004). Automating linguistics-based cues for detecting deception in text-based asynchronous computer-mediated communication. *Group Decision and Negotiation. 13*, 81-106.

Zhou, L., & Sung, Y. (2008). Cues to deception in online Chinese groups. *Proceedings of the 41st Annual Hawaii international Conference on System Sciences, 146*. Washington, DC: IEEE Computer Society.