Using Speech Act Profiling for Deception Detection

Douglas P. Twitchell, Jay F. Nunamaker Jr., and Judee K. Burgoon

Center for the Management of Information,
114 McClelland Hall, Tucson AZ 85721, USA
{dtwitchell,jnunamaker,jburgoon}@cmi.arizona.edu

Abstract. The rising use of synchronous text-based computer-mediated communication (CMC) such as chat rooms and instant messaging in government agencies and the business world presents a potential risk to these organizations. There are no current methods for visualizing or analyzing these persistent conversations to detect deception. Speech act profiling is a method for analyzing and visualizing online conversations, and this paper shows its use for distinguishing online conversations that express uncertainty, an indicator of deception.

1 Introduction

Whether it's governments protecting their citizens from terrorists or corporations protecting their assets from fraud, many organizations are interested in finding, exposing, and ridding themselves of deception. With the increasing use of computer-mediated communication (CMC) tools such as chat, instant messaging, and e-mail, persistent conversations are becoming more common and are increasingly used as methods for transmitting deception. However, automated tools for studying human behavior in online conversations are rare. Even more rare are tools for aiding deception detection in these conversations.

The need for tools to aid in searching, analyzing, and visualizing text-based online conversations for deception is evident. The National Association of Securities Dealers (NASD) recently issued a Notice to Members that clarifies the responsibility of securities firms to store all instant messaging conversations for three years [1]. As the use of instant messaging in the finance industry becomes more commonplace, there is a need for software to filter through the large amount of resultant data. Furthermore, both management and government regulators are likely to be interested in deceptive messages that occur in these large repositories of conversations. Software to make the search for and the visualization of deceptive conversations easier than reading through all of the conversations or using keyword search would be useful.

1.1 Deception

Deception is defined as the active transmission of messages and information to create a false conclusion [2]. Messages that are unknowingly sent by a sender

H. Chen et al. (Eds.): ISI 2004, LNCS 3073, pp. 403–410, 2004.

are not considered deceptive, as there is no intention to deceive. The problem is that most people are poor at detecting deception even when presented with all of the verbal and non-verbal information conveyed in a face-to-face discussion. The problem becomes even more acute when the deception is conveyed in text such as written legal depositions, everyday email, or instant messaging, and it is nearly impossible when there are large amounts of text to sift through. Furthermore, deception strategies may change from situation to situation as the deceiver attempts to fool possible detectors.

Deception in face-to-face conversations has been studied extensively for many years. Several meta-analyses exist that attempt to summarize the large body of studies in lying and deception. Zuckerman et. al. [3] found in their meta-analysis that negative statements, verbal immediacy and discrepancies in the narrative were the most powerful indicators of deception. DePaulo's latest meta analysis [4] revealed several correlates of deception including internal consistency and logical structure of a story, how "engaged" the participant is, the number of details in the participants message, and how uncertain the participant seems.

Interpersonal Deception Theory (IDT) presents deception as a strategic interaction among participants. According to IDT, before a conversation, both the sender and receiver of deception bring to an interaction their expectations, goals, familiarity, etc. During the interaction or conversation, the sender will begin his or her deceit with certain strategies but will modify those strategies throughout the interaction based on perceived deception success. The receiver, on the other hand, begins with some level of suspicion (even if that level is zero), which is modified throughout the interaction based on credibility judgments. Both parties will likely leak nonstrategic behaviors indicating their psychological state. In the end, both sender and receiver will be able to evaluate their success at deceiving and detecting deceit, respectively. Speech acts are the vehicle by which both strategic and nonstrategic behaviors are transmitted.

1.2 Deception Detection

To advance toward the goal of automated deception detection in text-based CMC, Zhou et. al. [5, 6] introduce a number of automatable linguistic cues to deception. They show that some are not only significant but are also in line with theoretical constructs used in face-to-face deception theories. For example, their study of task-based email communications found that the diversity of language of deceivers was significantly less than that of truthful participants. They extend their idea further in a more recent paper [7] by feeding the automatically detected cues into several statistical and machine learning techniques to classify messages as truthful or deceptive. This method turns out to have between 60% and 70% accuracy–better than chance.

Despite its accuracy, the cues used in this method have some shortcomings. Many of the cues are best suited for asynchronous messages like email. Because of its asynchronous nature, email messages tend to be longer than synchronous CMC such as IM and chat. The average length of the messages in the Desert Survival study used by Zhou et al. [5] is 133 words. Messages, or in this case individual utterances, in the two corpora utilized by this paper rarely exceed 20 words. The lack of message length renders many of the cues found by Zhou et. al. [5] incalculable. For example, lexical diversity is calculated as the number of unique words in a message divided by the total number of words. This potential deception cue becomes useless in messages that are so short that not a single word is repeated (the Lexical Diversity is 1). Similarly, cues such as the number of sentences, ratio of passive voice verbs to total verbs, and the amount of positive affect expressed by a message become useless when the messages only have a few words. For synchronous CMC such as IM and chat another method is needed.

1.3 Speech Act Profiling

Speech act profiling is a method of automatically analyzing and visualizing synchronous CMC such as chat and instant messaging with an aim to make searching for deception in large amounts of conversational data easier than searching by keywords and reading through whole conversations. Also presented is a visualization scheme for conversations that emphasizes the nature of each participant's conversation as a whole and allows the analyst to get an overall view of the interaction between the participants. The method, called speech act profiling, is based on the work of Stolcke et. al [8] on dialog act modeling. Their method utilizes n-gram language modeling and hidden Markov models to classify conversational utterances into 42 dialog act categories. Speech act profiling takes the probabilities (not the classifications) created by the combination of the language model and the hidden Markov model, sums them for the entire conversation. The resulting probabilities are an estimate of the number of each of the dialog acts uttered by the participants. The probabilities for each participant can be separated and displayed on a radial graph. The graph is further organized according to Searle's [9] taxonomy of speech acts, which gives the analyst an overall view of the conversation. The resulting conversation profiles could be useful in a number of situations, including visualizing multiple simultaneous CMC conversations, testing hypotheses about the conversation's participants, and, of course, the post-hoc analysis of persistent conversations for deception. A full introduction to speech act profiling, including a guide to the speech acts and abbreviations, can be found in [10].

2 Methodology

The speech act profiles used in this paper come from the StrikeCom corpus, which originated in a study where subjects played a three-player military decision-making game named StrikeCom. The game required teams to search a grid-based game board for enemy camps. Each player had two assets with which to search the board. During each of five turns, the players searched the board and submitted their search. At the end of each turn, the game returned one of three results: likely nothing found, uncertain, or likely something found. After the end of the five searching turns, the teams used the information to place bombs for destroying the enemy camps. During some of the games, one of the players was given the actual locations of the enemy camps and told to deceptively steer the group away from bombing those locations. The game was designed to foster communication and allow experimenters to assign different players to various treatments such as deceptive and truthful. All communication was conducted through text-based chat, which was captured and now comprises the StrikeCom corpus. The conversations in the StrikeCom corpus are different from the telephone transcriptions used to train the speech act profiling model; however, as some of the examples will show, the profiling method is robust enough to be useful despite these differences.

Of a total of 32 games played, 16 included a player who was told that he or she was actually an agent for the enemy and that they should use deception attempt to steer the other players off course. The other 16 games were played without any mention of deception. In all instances of the deception condition, the participant playing the user "Space" was instructed to be the deceiver. All other players were unaware of the possible deception. The 16 deceptive games include 2,706 utterances containing a total of 13,243 words.

Figure 1 is a speech act profile created from all of the utterances from a single game. In this particular game the profile indicates that the participant playing Space1 is uncertain compared to the other participants, Air1 and Intel1, as indicated by the greater number of Maybe/Accept-parts (maybe) and opinions (sv) and fewer statements (sd). An example of this uncertain language is shown in the excerpt in Table 1. Early in the game, Space1 hedges the comment "i got a stike on c2" with the comment "but it says that it can be wrong...". Later Space1 qualifies his advocacy of grid space e3 with "i have a feeling". In reality there was no target at e3, and Space 1 was likely attempting deceive the others as instructed. In Depaulo et. al.'s meta-analysis of deception [4], vocal and verbal impressions of uncertainty by a listener were significantly correlated with deception (d = .30). That is, when deception is present, the receiver of the deceptive message often notices uncertainty in the speaker's voice or words. Since the voice channel isn't available in CMC, any uncertainty would have to be transmitted and detected using only the words. The uncertainty is transimitted in the words is picked up by the profile in Figure 1 in the form of a high proportion (relative to the other players) of Maybe/Accept-parts (maybe) and opinions (sv) and a low proportion of statements (sd).

Table 1. Excerpt of conversation represented by the speech act profile in Figure 1

Speaker	Utterance
Space1	i got a stike on c2.
Space1	but it says that it can be wrong...
...	...
Space1	i have a feeling theres on at e3... also , on the next turn we need to check columns one and two.

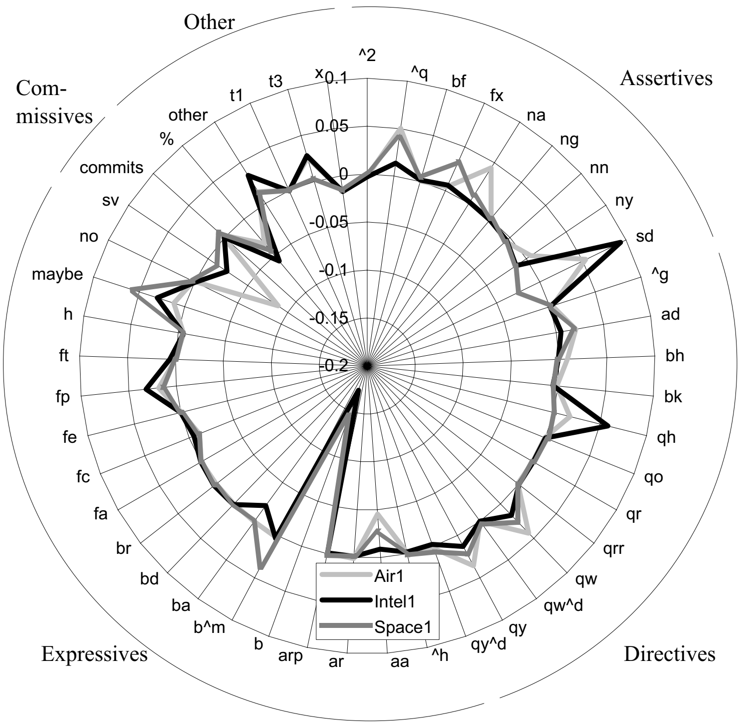

Fig. 1. Sample speech act profile from the StrikeCom corpus showing submissive and uncertain behavior by the deceiver

2.1 Preliminary Statistical Analysis

The example shown above is useful for envisioning how an investigator might use speech act profiling; however, the probabilities produced by speech act profiling can also be used for statistical comparison. More specifically, the average probabilities for each speaker obtained from the speech act profile represent the probable proportion of utterances that were of a given speech act. These probable proportions can be compared across experimental treatments when attempting to support hypotheses.

In addition to the profile shown in Figure 1, another way to detect the possible uncertainty is to obtain the proportion of all speech acts that express uncertainty. For example, Hedge and Maybe/Accept-part are two speech acts that express uncertainty. A Hedge is used specifically by speakers to introduce uncertainty into their statement (for example: *I'm not quite sure, but I think we should probably do it*). Maybe/Accept-part, also indicates uncertainty as in the phrase *It might be.* [11]. A set of speech acts that often express uncertainty are shown in Table 2. These uncertain speech acts can be combined by summing their probable proportions. The result is the probable proportion of speech acts that express uncertainty.

Table 2. Speech acts that often express uncertainty

Opinion (sv)	Tag-Question (^g)
Hedge (h)	Abandoned, Turn-Exit or Uninterpretable (%)
Dispreferred Answers(arp)	Acknowledge(Backchannel) (b)
Appreciation (ba)	Downplayer (bd)
Backchannel in question form (bh)	Response Acknowledgment (bk)
Signal-non-understanding(br)	Negative-non-no answers (ng)
Other answers (no)	Yes-No-question (qy)
Wh-Question (qw)	Open-Question (qo)
Or-Question (qr)	Or-Clause (qrr)
Declarative Yes-No-Question (qy^d)	Declarative Wh-Question (qw^d)

Given uncertainty's correlation with deception, the uncertain speech acts, and the StrikeCom corpus that contains both deceptive and truthful participants, the following hypotheses can be constructed.

Hypothesis 1: *Deceptive participants in a conversation will have a higher proportion of uncertain speech acts than truthful participants in other conversations.*

Hypothesis 2: *Deceptive participants in a conversation will have a higher proportion of uncertain speech acts than their truthful partners.*

The results of a simple t-test do not support Hypothesis 1, but the results of a paired t-test do lend support to Hypothesis 2. As expected, deceptive participants express more uncertainty in their language than do their truthful partners. The specific results are shown in Table 3.

Table 3. Means (and p-values)[†] for uncertainty

Deceptive Participants	Partners of Deceptive Participants	Unrelated Participant
0.32	0.27(0.03*)	0.29(0.20)

[†] p-values indicate the probability that the mean is different than the mean uncertainty of deceptive participants

[*] Significant at $\alpha = .05$ (one-tailed)

Hypothesis 2's significance is likely due to the tendency for participants in a conversation to sometimes compensate for their partner's behavior [12]. So, if one participant in a conversation expresses what seems to the other participants to be overly uncertain language, the other participants will likely compensate by using more certain language. Therefore, the difference in uncertainty between a deceptive participant and his or her partner's language will likely be more than the difference between a deceptive participant and another unrelated, truthful participant.

The results of this statistical test serve a two-fold purpose: (1) they lend support to previously supported hypotheses, namely that uncertain language accompanies deception; (2) they give statistical support to the validity of using speech act profiling to aid in deception detection in persistent conversations. Obviously, there are several potential hypotheses one could test on deception using the StrikeCom corpus and speech act profiling, but the focus of this paper is the use of speech act profiling with deception, not deception itself. Therefore, the statistical support of the validity of speech act profiling is of more import in this paper than replicating previous findings in the deception field.

3 Conclusions

Speech act profiling creates a new way to visualize and analyze online conversations for deception. The example profile and accompanying statistical analysis illustrate how insight into conversations and their participants can be gained and how that insight can be used to improve deception detection. Speech act profiling as it is presented here is in its first stages of development, but still represents a potentially valuable automated tool for detecting deception.

References

1. NASD: Instant messaging: Clarification for members regarding supervisory obligations and recordkeeping requirements for instant messaging. http://www.nasdr.com/pdf-text/0333ntm.pdf (2003)
2. Buller, D.B., Burgoon, J.K.: Interpersonal deception theory. Communication Theory **6** (1996) 203–242
3. Zuckerman, M., Driver, R.E.: Telling lies: Verbal and nonverbal correlates of deception. In Siegman, A.W., Feldstein, S., eds.: Multichannel Integrations of Nonverbal Behavior. Lawrence Erlbaum Associates, Hillsdale, New Jersey (1985)
4. DePaulo, B.M., Malone, B.E., Lindsay, J.J., Muhlenbruck, L., Charlton, K., Cooper, H.: Cues to deception. (under review) (2000)
5. Zhou, L., Twitchell, D.P., Qin, T., Burgoon, J.K., Nunamaker Jr., J.F.: An exploratory study into deception detection in text-based computer-mediated communication. In: Thirty-Sixth Annual Hawaii International Conference on System Sciences (CD/ROM), Big Island, Hawaii, Computer Society Press (2003)
6. Zhou, L., Burgoon, J.K., Nunamaker, J.F.J., Twitchell, D.P.: Automated linguistics based cues for detecting deception in text-based asynchronous computer-mediated communication: An emperical investigation. Group Decision and Negotiation (In press) **13** (2004)
7. Zhou, L., Twitchell, D.P., Qin, T., Burgoon, J.K., Nunamaker Jr., J.F.: Toward the automatic prediction of deception - an empirical comparison of classification methods. Journal of Management Information Systems (In Press) (2004)
8. Stolcke, A., Reis, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., Taylor, P., Van Ess-Dykema, C., Martin, R., Meteer, M.: Dialogue act modeling for automatic tagging and recognition of conversational speech. Computational Linguistics **26** (2000) 339–373

9. Searle, J.R.: A taxonomy of illocutionary acts. In: Expression and Meaning: Studies in the Theory of Speech Acts. Cambridge University Press, Cambridge, UK (1979) 1–29
10. Twitchell, D.P., Nunamaker, J.F.J.: Speech act profiling: A probabilistic method for analyzing persistent conversations and their participants. In: Thirty-Seventh Annual Hawaii International Conference on System Sciences (CD/ROM), Big Island, Hawaii, IEEE Computer Society Press (2004)
11. Jurafsky, D., Shriberg, E., Biasca, D.: Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual, Draft 13 (1997)
12. Burgoon, J.K., Stern, L.A., Dillman, L.: Interpersonal Adaptation: Dyadic Interaction Patterns. Cambridge University Press, Cambridge, UK (1995)