

# On a Text-processing Approach to facilitating Autonomous Deception Detection

Therani Madhusudan,  
MIS Dept., Univ. of Arizona  
Tucson, AZ 85750

**Abstract**—Current techniques towards information security have limited capabilities to detect and counter attacks that involve different kinds of masquerade and spread of misinformation executed over long time periods to achieve malicious goals. Detection of such deceptive information obtained during online interactions (emails, chat room conversations) is the first step before counter strategies can be developed. With the large-scale use of information technologies as a general communication medium, facilitating deception detection is a key enabler to utilizing information systems to their fullest potential. This article presents a framework for Computer-Aided Deception Detection building on the Interpersonal Deception Theory (IDT) of human interpersonal communication research and text-processing techniques for facilitating deception analysis of text-oriented communication. A state-transition diagram based framework is proposed to model the dynamic evolution of an interpersonal conversation between a sender and receiver based on the IDT-based process schemata. The framework is then utilized to develop a deception detection agent to process textual information. Deception detection is defined as a process of model verification. Architecture of a prototype under development and open problems for further research in this area are outlined.

## I. INTRODUCTION

The widespread use of information obtained via electronic media to support decision-making permeates all aspects of everyday life for the average human user. Information obtained via email or a website may implicitly be trusted by the average user and who may act upon the same. Consider for example an erroneous press release regarding Emulex Corp. by Internet Wire (a press release distribution service), based on a forged email sent to Internet Wire regarding company affairs, which triggered a stock drop of 61% in August'2000. Internet Wire did not bother to verify the origin or contents of the email. Further, users of the information (stock holders, brokers etc.) acted implicitly trusting that information from InternetWire had to be true. These kind of attacks target the way humans assign meaning to the electronic content received or obtained and act upon the same (and referred

to in some literature as “semantic attacks”). Electronic information thus obtained is a major vulnerability and an open target for attack by malefactors. Actions based on misinformation could lead to potential benefits to the perpetrator in the long or short run. Few people take the time during routine, normal activities to corroborate the veracity of the information they base their decisions on by examining the source credentials, finding alternate opinions or checking the content with a third party considered an authority on the nature of the content[16], [20].

Though it is possible to diligently manually verify the semantic content of information obtained electronically, it takes an inordinate amount of time and resources. Such manual verification may be near impossible under conditions of duress when information changes rapidly and exhibits a wide range of content (which sometimes may be new), and the time available for judicious rational decision-making is a limiting factor. Such conditions occur in many contexts such as 1) on a daily basis in contexts such as electronic markets, real-time business transactions and airtraffic control 2) under battlefield conditions 3) in life and death situations such as in hospitals, industrial and automobile accidents. In some cases, when a manual operator detects an error, it may be possible to check (for example, an air traffic controller can override a malfunctioning radar based on visual detection) but it may not always be possible leading to potentially disastrous actions. With the recent focus on increasing cyber security, improving risk management and disaster prevention, it is essential to build information management tools that provide reliable information to aid the human decision-makers.

Computer-aided tools to assist in the information content and source verification process would be highly beneficial both in extreme conditions and also for normal usage in mission-critical information technology-based applications. In extreme conditions, computer-aided deception detection(CADD) assistance may help sort through all the available electronic information and filter the implausible ones, reducing the load on final manual verification be-

fore decisions are made and ensuing actions taken. Under benign conditions, such CADD tools may be embedded in modern information access tools and validate the information content or source and flag the user if any potential aberration is detected. Further, these embedded tools may also be used to instill the discipline of corroborating information obtained and train the manual users to a new paradigm of accessing and using electronic information.

In this article, an approach to facilitating deception detection during electronically mediated conversations by modeling the deception detection process as a sequence of computational model verification steps is discussed. Two basic models are outlined, namely, 1) A communication model illustrating the structure of the communication interaction. It consists of submodels of the message and information content, the information source(sender) and the information sink(receiver) and 2) The deception detection process model which operates on the communication model and illustrates the dynamics of how meaning is assigned to a message and the meaning is trusted by the information user. The models are developed, based on the Interpersonal Deception Theory (IDT) proposed in [7]. IDT models are based on social science studies into human communication. These models outline the interactive communication processes and describe the various acts of the participants during a deceptive interpersonal communication interaction. Computational models that facilitate automated identification of deception in both face-to-face and electronically mediated conversations and mimic the underlying reasoning of the participants are not currently available *a priori* and need to be developed via experimentation and the synthesis of our current knowledge from the communication and psychological research literature.

The focus of our current work is on identifying the possibility of deception in a text-oriented communication interaction such as email conversations or chatroom dialogues. This article describes how the communication and process models briefly discussed above can be instantiated using textprocessing techniques and thus facilitate deception analysis of textual electronic conversations. Further, an exploratory computational architecture to facilitate incremental development of the various structural and process models that constitute the phenomena of deception is outlined. The key feature of the proposed approach is the synthesis of recent advances in natural language text processing techniques with the behavioral models of communication to facilitate deception analysis. The approach provides for incremental model development and experimentation as the various aspects of deception are identified and analysed.

The rest of the article is organized as follows: Section

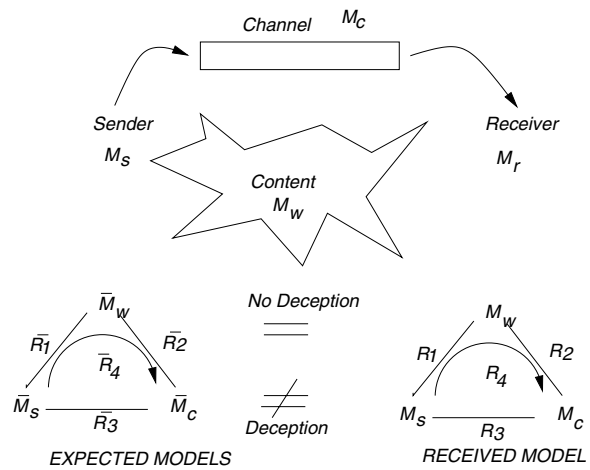


Fig. 1. A Communication model for Deception detection

2 describes computational models for deception detection building on basic ideas of IDT. Section 3 details the requirements for a computational framework and the role of textprocessing techniques in the proposed framework. Section 4 provides a brief survey of the relevant background literature and concluding remarks.

## II. PROBLEM DESCRIPTION

Firstly, we describe the structural aspects of a communication interaction followed by the computational process model for deception. For purposes of the initial development of our framework, we consider as a basic building block, the communication interaction between a single information receiver and single sender. Section 2.2 further elucidates this model based on IDT. A dynamic computational process model is developed and the process of deception detection is defined in this framework.

### A. Definition of Deception

For purposes of this article, we assume that information is transmitted from an information sender to a receiver. Further, this information transfer process can be 1) synchronous or asynchronous and 2) a one-shot communication or a dialogue. The process of information transfer from sender to receiver is illustrated in the top of Figure 1.

There is a sender (modelled as  $M_s$ ) who encodes the world information (or content) modelled as  $M_w$  and puts the message ( $m$ ) on the information channel modelled as  $M_c$ . The channel is then accessed by the receiver (modelled by  $M_r$ ) either autonomously or on cue and the message is decoded before initiating an ensuing action. The ensuing action may initiate a dialogue as more information is requested by the receiver or the receiver may act based on the message. For purposes of our discussion, we

assume that the overall semantics of the message  $m$ , as inferred by the receiver is based on some relevant composition of the following key submodels:

- Model of the message content which models real world information provided by  $M_w$ . This constitutes all models that correspond to the phenomena or event in the real world under consideration. The content model may even be executable code that mimics some phenomena.
- Model of the sender which models properties of the sender provided by  $M_s$ . The model captures the senders characteristics, his psychological makeup, familiarity with the content and channel, access to the information, reliability, knowledge etc.
- Model of the information channel and media,  $M_c$ , from which properties of the nature of channel can be inferred. This model provides the general characteristics regarding the media used to convey the message such as text, images and video. Encoding information as text is much different from encoding the same information as images. Further different channels have varying capabilities in transmitting the message.
- Model of the receiver  $M_r$ , that provides characteristics of the receiver when the message is being received. For example, it could model the receivers familiarity with either the content or the media. For example, an illiterate receiver could misinterpret a text message.

For purposes of this article, the receiver is considered to be a rational entity (a computational agent) without any bias of any kind. In reality, a receiver model may be necessary based on the kind of human user who may be aided by such an computational agent. Thus a rational receiver deals with the composition of the three basic submodels as shown at the bottom of Figure 1.

The triad (called `Received models`) at the bottom of the figure illustrates that the totality of the final message  $m$  received is based on the following seven basic constituent entities for a given  $M_r$ :

- Model of content as it maps to the world,  $M_w$
- Model of the sender - properties of the entity that encodes the message,  $M_s$
- Model of the information channel - properties of the media such as text, audio, video provided by  $M_c$
- Relationships between the world model and sender model collectively called,  $R_1$ .
- Relationships between the sender model and the information channel model collectively called,  $R_3$ .
- Relationships between the world model and information channel model collectively called,  $R_2$

- Relationships between all the three basic models collectively called,  $R_4$ .

The tuple (the set of submodels)  $(\overline{M}_w, \overline{M}_c, \overline{M}_s, \overline{R}_1, \overline{R}_2, \overline{R}_3, \overline{R}_4)$  denotes the **actual or true** set of models and denoted as `Expected models` in the figure for a given receiver. Different receivers may have different expected models based on the goals of the conversational interaction. A receiver enters into a conversation either voluntarily upon request or to meet a personal information-seeking goal and participates in the conversation with cogent expectations of how the conversation is to evolve, what information is to be obtained or provided etc. These expectations constitute the expected model. Developing a general set of plausible expected models for different kinds of conversational goals is difficult but for specific tasks such as domain-oriented information retrieval or domain-oriented problem solving such as playing games it may be possible. Distinct variations between the expected model and the reality (the true models obtained during conversation) may indicate a plausibility of deception. A simple baseline (from a computational perspective) is that when the expected and true models are equivalent for a given receiver  $M_r$ , there is no deception. If there is a discernable well-defined difference with reasonable accuracy, we may detect deception. *Deception detection is the process of identifying these inconsistencies via verifying the received model against the expected model.*

Deception can be introduced by manipulation of any combination of the above seven basic constituents of a received message by a receiver. In the introductory example (of the forged press release), since  $M_s$  nor  $R_1$  was verified (source credentials) nor  $M_w$  (message corroborated),  $M_c$ ,  $R_3$ ,  $R_2$  nor  $R_4$  did not raise any cues (Email was forged, such messages are possibly sent via email by relevant people in the know), a hoax occurred. Activities such as falsifying information or introducing trojan horses (manipulating  $M_w$ ), identity deception (manipulating  $M_s$ ), tampering with editing text or images (manipulating  $M_c$ ), forgery (manipulating both  $M_s$  and  $M_w$ ), digital counterfeiting (manipulating all three ( $M_c$ ,  $M_w$ ,  $M_s$ ) models) can all lead to deception. It is also important to note that any inconsistency between these three models i.e. a violation of their relationships can possibly identify deception for a receiver. Since these three models need to be mutually consistent, deceivers go to great lengths, to meet the criteria for all three expected models to be consistent when indulging in a deceptive act for a receiver. Deceivers usually are successful when the basic verification steps of checking equivalence between the expected and received are not conducted or cursorily skipped by the receiver or if the re-

ceiver model is biased or incomplete with respect to the message. It is also important to note that the more complex the message, the more difficult it is to verify and thus enables deception. Research in human communication has been shown that deception and suspected deception arise in nearly a quarter of all conversations. However, detecting deception with reliable accuracy is extremely difficult for humans. Facilitating the same using computational means is a much harder problem. Further because of the tacit and implicit aspects of human communication and interaction, explicating the various cues of deception, formally and explicitly representing the same, to facilitate detection is a complex task. Detection deception in our framework is equivalent to detecting inconsistencies within the tuple  $(M_c, M_w, M_s, R_1, R_2, R_3, R_4)$  (by comparing expected and received) for a given receiver,  $M_r$ . It is also possible that for a given message  $m$ , one receiver may be weaker (i.e. more deceivable) compared to another depending on the relative strengths of their expected models. Inconsistencies have to be detected across multiple receiver models of increasing strength, before deception can be identified in a consistent manner for an autonomous detection system. From an automated deception detection perspective, we need the strongest receiver model which may have to be developed incrementally. This concludes the description of the structural aspects of the communication model. Further discussion on the encoding of this model in the context of textual electronic conversations is discussed in Section 3.

### B. Relationship to IDT

IDT is a macroscopic theory of deception and builds on a merger of interpersonal communication and deception principles designed to better account for deception in interactive contexts[9], [13]. Readers are referred to the above articles for a better background to IDT. The focus of the rest of this section is on discussing the relationship between the structural model tuple  $(M_c, M_w, M_s, M_r, R_1, R_2, R_3, R_4)$  (both expected and received) with respect to IDT and how IDT may provide insights into developing a computational model for deception detection. Relevant assumptions and insights from the theory are discussed in the following paragraphs.

Firstly, the process of interactive, interpersonal deception is goal-oriented and intentional in nature involving strategic and non-strategic behaviors. The strategic actions of a sender (which may not be discerned by the receiver) are accompanied by non-strategic actions (consisting of perceptive, cognitive and emotional processes) which may provide a clue to the receiver about

the situation. Identification of these clues, called *leakages* in the deception literature[17] is a key step in detection. Goal identification along with accompanying non-strategic actions may provide templates of things to scan and search for in the course of a conversation during the deception detection process. Thus based on a sequence of messages  $(m_1, m_2, \dots)$ , it may be possible to identify overall goals and also elicit models and relationships from the same. Thus identification of the tuple  $(M_c, M_w, M_s, M_r, R_1, R_2, R_3, R_4)$  may be possible based on the goals, context and sequence of messages in the conversation. However, these models evolve during an interaction and accounting for the same is a key consideration. This model evolution history needs to be considered during deception detection.

The complexity of detection is largely due to its dynamic nature. The communication process involves active participation by both sender and receiver. Thus both sender and receiver models  $(M_s, M_r)$  may be modulated during course of the interaction. The behavioral patterns (the relationships  $R_i$  adapt to feedback between the participants, evolution of the context and changes in topics. Messages (from senders) entail communicative acts that enable management of these behavioral patterns, steer the course of conversation, manage impressions and emotions and influence the receiver. Further, nominally receivers implicitly trust communication (unless experience suggests otherwise) as a social norm. Deceivers exploit such trust by controlling information via different kinds of encoding that alter veracity, completeness, directness, clarity and personalization. Deceivers create desired impressions, process rapidly changing streams of information, make sense of incongruent and ambiguous information, control effects of leakage cues and keep the interaction oriented towards the goal. Similarly, an alert receiver is conducting an heightened surveillance of the interaction, managing his responses and emotions and executing a safe conversational strategy to detect duplicity if it exists or conduct a normal interaction otherwise. The process as illustrated above (and studied in [8], [6], [11], [12], [10]) illustrates the evolution of the tuple  $(M_c, M_w, M_s, M_r, R_1, R_2, R_3, R_4)$  for both sender and receiver as the conversation evolves and messages are exchanged between the participants.

The process of deceptive interpersonal interaction is also spatio-temporal in nature and is illustrated in IDT as shown in Figure 2. (reproduced from [7]). The process of deception begins in a well-defined context and relationship between a sender and receiver. Within the context are the cognitive and behavioral factors that define sender/receiver interaction and set the stage for evo-

lution of the interaction. The cognitive aspects consist of goals, behaviors and background knowledge accompanied by specific skills and patterns of skill usage (repertoire) for both participants. The timeline of interaction is divided into preinteraction, a variety of intermediate stages and finally a post interaction stage. The preinteraction stage defines the initial behaviors (during time period  $T1$ ) when the interaction begins. As the interaction evolves, these in turn affect detection accuracy, credibility judgements and suspicion displays (time periods  $T2$  and  $T3$ ). As the information is processed, receiver cognitions and behaviors influence sender cognitions and behaviors (time periods  $T4$  and  $T5$ ) followed by feedback adjustments (time period  $T6$ ) and so forth. These stages continually iterate and evolve through time till the interaction terminates and postinteraction processes are initiated. It is important to note that during a time-period multiple messages may be exchanged between the sender and the receiver and the above schematic identifies key transition stages during the conversation. The macroscopic process outlined here provides a highlevel staged approach to understanding the process of deception. The process schema is generic enough to handle to both synchronous and asynchronous as well as various kinds of face-to-face versus spatially separated interactions.

IDT thus provides a basis to understand how the tuple  $(M_c, M_w, M_s, M_r, R_1, R_2, R_3, R_4)$  (both expected and received) from the previous section evolve through time during the course of an interaction. The theory provides a framework to define the details of the constituent models and the nature of the relationships between them. Further, as illustrated in IDT, these models constantly evolve in a complex manner intermingled with deception detection steps (conducted consciously or unconsciously by the receiver). Using the IDT process as a basis, from a computational perspective, a deceptive conversational interaction can be viewed as a timed sequence of models,  $\sum_1, \sum_2, \sum_3, \dots, \sum_t, t = 1, 2, 3, \dots, n$  and each  $\sum = (M_c, M_w, M_s, M_r, R_1, R_2, R_3, R_4)$ . From our perspective, a transition between  $\sum_i \rightarrow \sum_j, i \neq j, j \geq i, i \in t$  occurs when a message is exchanged either from a sender to receiver or vice-versa. Each  $\sum_i$ , henceforth called a conversational state, captures the current state of the interaction between the sender and receiver. Given the conversational state model, the deception conversation can be modelled as a state-transition diagram. Further, within each conversational state, either participant has the choice to continue with the nominal thread of conversation or initiate queries towards detection/suspicion processes. Further, consistency checking for deception detection could be triggered at the end of every transition or

at some predefined critical states at the receiver whereas at the sender, goal consistency checking could be initiated to ensure that the conversation is moving towards a goal.

An illustration of a conversational state diagram (to mimic IDT) is shown in Figure 3. The preinteraction phase is the *START* state, followed by transitions (shown by filled arrowheads) to states  $1, 2, 3, \dots, 9$ , and then the *END* state. Each transition is assumed to be caused by a message exchange, leading to possibly a new state or a transition back to the same state (a loop) as shown in state 5. In each time-period  $T_i$ , there could possibly be many states with transitions depending on the nature of the conversation. Two outgoing transitions are shown from state 2 namely,  $2 \rightarrow 3$  and  $2 \rightarrow 3'$ . The state sequence  $(2, 3', 4', 5', 6', 7', 8')$  (shown with unfilled arrowheads), illustrates a corrective message sequence that could be initiated by the sender when it is detected that the conversation is moving away from the goal, possibly in state 2. Similarly, the receiver could initiate transitions (i.e. a conversation sequence) to elicit more information and confirm the suspicion of detection.

The conversational state diagram of Figure 3 is one obtained by a third-party playing the role of an observer and describes the *joint interaction* between the communicating parties. In realtime situations, each participant will have their own local view of the conversational state diagram as each process the information in their messages and initiate actions accordingly as shown in Figure 4. Further, the local view of each agent may not even overlap either in the state descriptions or the transitions. The sender may use a local state diagram to steer the conversation towards a set of goals whereas the receiver may use another local state diagram to guide the conversation to confirm his/her suspicion and detect deception. A generic interpersonal interaction thus defined has no bounds either on the number of states, the number of transitions, or the patterns amongst the transitions unless explicit communication protocols are defined. Many other properties of this framework are to be defined and further investigated.

Though the state-based model provided here is simplistic in nature, it provides a basis to develop a detection agent that tracks the evolution of the conversation (represented by the state diagram). Developing a library of standard and abnormal conversational patterns may prove beneficial for detecting anomalies in any given observed conversation. Further, the model is generic in that a human detection agent may also be introduced at any state to identify deception given the current history, defined by the sequence of states traversed thus far, and the accompanying messages. In a similar manner, instead of deception-detection agents, it is also possible to envision, deception-

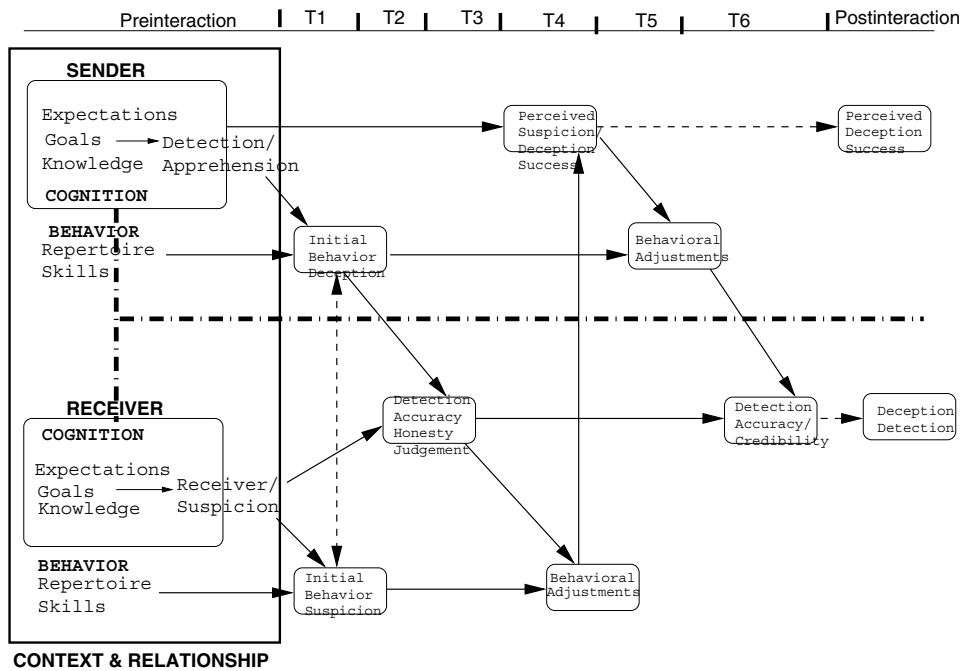


Fig. 2. Schematic Deception process model in IDT

generating agents, that introduce deception into a conversation.

The state diagram enables modeling various types of conversations, namely,

- Two active participants, synchronous or asynchronous, since we have abstracted time in terms of states. States need not be traversed in temporal order.
- Only receiver, one-way conversations, with a passive sender. For example, a website is modelled as a passive sender, responding in a predefined manner to every message from receiver
- Face-to-face or electronically mediated or even by a third-party may be modelled by this framework.

The above model is inspired by the philosophy of language as action theory from AI and computational linguistics and theory of speech acts[21]. Further details of the role of natural language processing in facilitating deception detection is discussed in the ensuing section.

We describe the utility of the framework for deception detection in the following section for a one-way conversation.

### III. PROPOSED COMPUTATIONAL FRAMEWORK

The evolution of the joint state transition diagram models an interpersonal conversation as discussed above. In the following paragraphs, I describe the design of a deception detection agent that builds a receivers local view of the state diagram and coordinates its internal and external (message response or query) actions accordingly based on a text-oriented conversation.

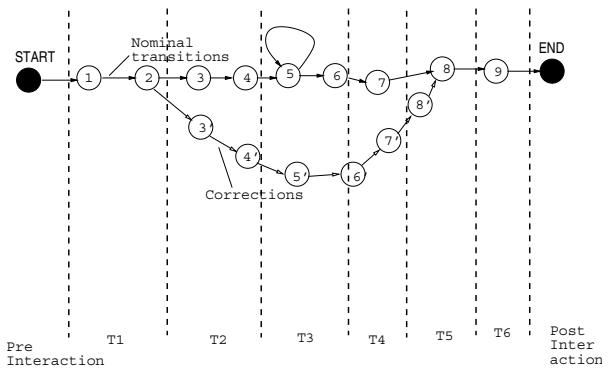


Fig. 3. State Transition diagram for a deceptive interpersonal conversation

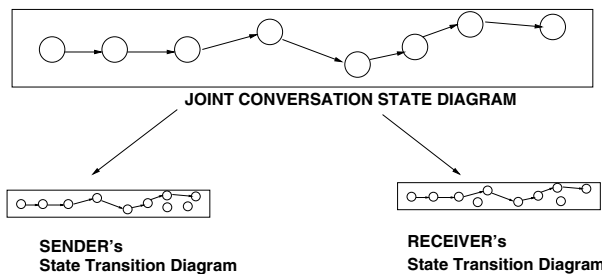


Fig. 4. Multiple views of a Conversation

#### A. Design of the Receiver Agent

For illustrative purposes, (without loss of generality), we focus on a two-state model (consisting of adjacent states), wherein, the receiver agent (also called the Deception Detection(DD) agent) begins from a given state, receives a message  $m$  and reaches the ensuing state. The

generic tasks of a receiver agent, when it enters a new state are as follows (in order):

- Receive message and extract the content to update the local state  $\Sigma$ .
- Update the current internal representation of the local state diagram  $\Sigma$  and identify new current state.
- Evaluate current state and choose a course of action. There are two possible classes of choices: 1) choose nominal interaction or 2) Suspect detection and initiate additional information gathering actions to confirm or deny.
- Generate appropriate response by encoding the message according to chosen action and transition to next state.

The key steps in the above procedure are extraction of content from the message, identifying leakages etc. and evaluation of the newly generated state after all updates. The performance and behavior of the receiver agent is dependent on the intelligence and ability of the agent based on the knowledge embedded in the same. Similar processing steps possibly characterize the ability of a human agent as to how the message is interpreted in context and evaluated.

### B. Message Extraction and Deception Detection Processes

Message extraction occurs at two levels, syntactic and semantic. At the syntactic level, the message can be encoded in various formats, namely, speech sounds, textual interaction (such as email), pictures (sketches, images and diagrams) and video. Further, we assume that message content can be obtained by the receiver (by decoding any encrypted message if necessary). Encryption techniques protect the message content from tampering when the messages are in the channel. In this article, our primary focus is on deception detection of textual messages (such as email or press releases from websites etc). Speech dialogs can be transcribed into a conversation and presented textually and will not be further discussed here.

We assume that agents involved in a conversation speak the same language i.e. share the same syntax including grammar and vocabulary. By parsing the messages (using relevant natural language processing techniques), information may be extracted. The information to be extracted from the messages depends on the representation of the models ( $M_c, M_w, M_s, M_r, R_1, R_2, R_3, R_4$ ) and is discussed below.

1) *Representation of Models:* Representation of the models is based on conventional knowledge representation techniques from the field of AI[19]. The models

$M_i$  are encoded using declarative representations with attached procedures. The relations  $R_i$  which encode consistency relationships are designed in terms of if-then rules. In our current prototype, the agents use a first-order predicate logic formalism to encode the various aspects of each model. Each of the models are currently under development and we illustrate the same. The semantics and syntax of the representation is under development.

Senders, receivers and the channels are modelled in terms of frame-based representation (a variant of predicate logic). Frames consist of slots, wherein each slot is a property either in declarative form or an attached procedure that can be executed to infer relationships or constraints. For example, a partial channel representation is shown below (the text in the parenthesis are possible values for the attribute):

```
Channel_type:(email-text, source-text,
speech-transcribed-text)
Conversation_type:(Synchronous,Asynchronous)
Interaction_location:(Online,face-to-face)
Channel_related_cues:(misspellings, repetition
overuse_of_exclamations)
Channel_language:(English, Latin, Esperanto)
Number_of_participants:()
Channel_usage_characteristics:()
```

Attached procedures for such channel descriptions could be those that update related values of the slots when one of them changes. For example, changing languages may enforce a completely different set of textual cues specific to that language. Similarly senders and receivers are also modelled. Example attributes are sender location, sender history, senders language abilities, senders channel use characteristics etc. Relationships are codified as constraints using if-then rules. Many of these rules are bidirectional and enforce different kinds of consistency checks. Examples of some such rules are: A sender from this geographical location should know about a certain local tourist attraction. A person working in this department should not have access to this information. We are currently developing generic model representations and rules.

The state transition diagram and the overall conversation are explicitly stored in persistent storage, called a conversation repository. A conversational state is defined by attributes such as the message received, states of the component models, time when the state was entered, exited, duration spent in that state and number of times the state was visited. Storing the state diagram provides a means to analyze conversations in a mechanistic manner. Further, to facilitate reasoning about the evolution of states, at each state, we define a module called the

expected state generator. This expected state generator is developed based on experience across human interactions and plausible courses of interpersonal behavior. The DD agent uses this state generator to posit a new expected state from the current state whenever it receives/sends a message. This expected state is then compared to the state after message reception. Details of design of the expected state generator are currently under development and is based on a rule-based finite state machine.

Information extraction from the textual messages modifies the current models of the sender, receivers or channels or triggers a variety of rules to be fired. One of the key issues is updating the knowledge with respect to new information. The DD agent may possibly recognize all the information relevant to its knowledge base but miss knowledge beyond its current scope. Those predicates currently not in scope are assumed to be false. This assumption, called closed-world, assumption is a limiting factor and may be extended based on addition techniques of non-monotonic and default reasoning. In our planned initial studies, we aim to have a human-aided deception detection agent, to minimize such problems.

2) *Deception detection:* Deception detection transpires during the evaluation of a new state after message updates have been performed. The evaluation process has to verify that each of the constituent models and relationships are really true for any message. Inconsistencies in any of these seven constituents (between the received and the expected) is a potential indicator of deception. We believe, human detectors potentially verify all these seven constituents simultaneously based on various cues and how this is done is an open area for further study.

In our proposed approach, we postulate a sequential deception detection process. The evaluation process executes each of the following seven basic verification steps and their potential interactions before converging to a decision that the message is potentially deceptive or not and selecting a course of action. Each verification step is a complex activity involving multiple submodels. A classification of possible submodels and computational techniques for verification is currently under research. The steps of the evaluation process include 1) Channel model verification: Does the received message exhibit the necessary properties that could be expected on this channel? such as is the syntax correct? Is the grammar plausible? 2) Sender model verification: Is the sender authentic and exhibit the expected characteristics for such a message? Is this consistent with sender history ? 3) Relationship between Channel and Sender: Is it possible for the sender to be capable enough to encode the message in this channel? 4) Content model verification: Is the message con-

tent plausible? (independent of the sender) 5) Relationship between Content and Sender: Can the sender be familiar with such content? 6) Relationship between Content and Channel: Is it possible to send such content via such media? 7) Relationship between Content, Sender and Channel: Is it possible for all of these to occur together without any inconsistencies? and 8) The questions above are examples of those that could be possibly asked during verification of each model. Additional possible questions may depend on the goal of the interaction and conversational domain.

As mentioned earlier, each verification step consists of comparing the received model and the expected model. Many kinds of comparative techniques may be used - qualitative, quantitative, logical or an hybrid. The above model representation facilitates a logical framework for comparison. Once a message is received and information is extracted, a new state  $\sum_{new}$  is obtained. The new state entails the received model. The DD agent during the state update steps also creates an expected state  $\sum_{exp}$ , using the expected state generator. The key comparison step is to infer that  $\sum_{new} \subseteq \sum_{exp}$  i.e. the new state is a subset of all that is implied by the expected state.

A variety of mechanistic logical inference techniques are available such as resolution theorem proving and induction. For the logic-based representation for each model component, we compare the expected versus the real state by trying to infer clauses in the expected state models based on the received model. Different inference techniques may be need to be used based on the nature of the underlying representation and is currently a topic of research in our framework. Verifying one relationship may involve validating a variety of possible concurrent relationships before the received message can be trusted.

Current research is focused on developing the model representations for some basic test domains. In the following paragraphs, a brief discussion of the deception detection agent along with the role of text-processing techniques is provided.

### C. DD Agent Architecture

The DD Agent process can be envisioned as a sequence of verification steps wherein each task deals with verification of a particular constituent model under the guidance of the generic process of Section 3.1. Verifying a particular received model may spawn numerous concurrent activities. With this perspective, a DD Agent can be envisioned as a coordination agent which initiates these model verification steps and then coordinates the inferring process based on the results received. The coordination rules will be developed as we understand the

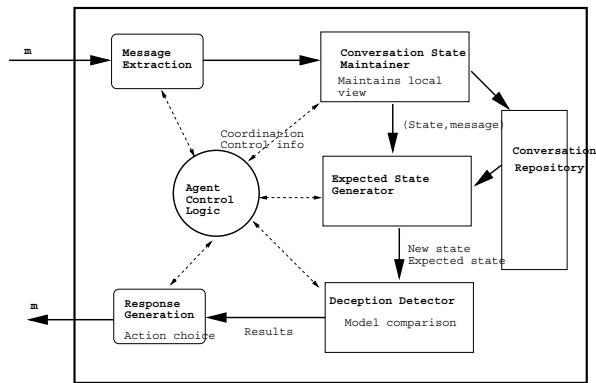


Fig. 5. Deception Detection Agent Architecture

interactions between the constituent models better based on studies of human deception processes. From an experimental perspective, we need a computational framework with the following key requirement, namely Compositionality - computational models and humans need to be combined and controlled in any aspect of the deception detection process. The models acquired may encompass both structural knowledge (domain (content) independent such as generic sender/receiver behaviors, generic channel models and common-sense knowledge) and domain-dependent semantic knowledge. This allows for setting up different experimental and configurational scenarios to elucidate the models. For example, model comparisons between expected and received can be conducted manually, and differences tracked via an user-interface. Such data can be analysed to build better models of the overall conversation. Additionally criteria required are extensibility, modularity and distributability of the setup for experimental purposes.

A generic architecture of the DD agent is shown in Figure 5. The agent control module coordinates the agents tasks. The key deception related modules are the conversation state maintainer, the state generator and the evaluator. Messages are received, semantic information extracted (via natural language processing), deception evaluation is performed and an appropriate response is generated. The conversational repository stores models, rules etc, and facilitates functioning of the deception agent. Each module embeds its own internal processing engine to facilitate appropriate reasoning and representation techniques. For example, the evaluator embeds a theorem-prover to make inferences. Additional internal structure is currently under development and is described in [18]

We have thus far developed a generic process execution engine with a discrete task model that is currently being adapted to execution the detection workflow presented above. The prototype engine has been built using a component framework and is being currently being tested for

verification of textual media. In the following section, a discussion of the role of text processing is techniques is presented.

#### D. Role of Text-processing techniques

Text-processing algorithms form the core of the message extraction and response generation modules of the Deception Detection Agent. Message extraction from free unstructured text has been studied in the information extraction literature and readers are referred to [1], [15], [2]. Basic information extraction has been studied in the context of factual event extraction and named entity extraction. Our work is currently exploring the utility of these techniques in the context of textual analysis. Further, in the proposed framework, we are developing standardized "cue" libraries of conversational textual features such as collections of noun phrases, verbs, adjectives, repeat utterances of word combinations etc. which indicate different types of deception-related cues. Extracting such cues from free conversational text and comparing with those in the library is a key indicator. Response generation for facilitating different kinds of conversations has also been studied in the natural language processing literature. Utilization of such techniques for managing conversations is a key research area. Discourse and dialogue analysis is another key research area in natural language processing and many enable the analysis of the evolution of the joint conversational state diagram outlined earlier. Rules for managing deceptive conversations and its interactions with messages and sender/receiver models need to be studied further. Our current work is embedding basic text processing techniques for the different phases of detection and developing generic computational representations is extremely complex. A notable point is that detecting deception even within textual conversations is a complex task. Consideration of realtime interactions along with speech and other modalities only illustrates the complexity of the problem.

#### IV. DISCUSSION AND CONCLUDING REMARKS

The proposed framework builds on ideas from two key perspectives, namely 1) Communications research and 2) The Natural Language processing literature in Artificial Intelligence. A background on the latest psychological literature on deception is summarized in [22]. IDT is one amongst multiple, interdisciplinary theories of deception such as Uncertainty Reduction Theory[3], Behavioral Cue Theory[5], Expectancy violations theory[4] and Channel Expansion theory[14].

A major question while using the above theories to build a computational tool for detection is what aspects of

these theories are amenable to computation and wherein is manual involvement required to utilize the theory. Our approach has been process-driven and hence with IDT providing a larger process framework, we intend to situate each of the above theories as subprocesses within the same. Our approach has been top-down and the strategy is to capture necessary knowledge for deception incrementally. Textual message encoding and decoding is an important step in our framework and recent advances in natural language processing techniques may benefit deception detection.

The key characteristic of the proposed approach is the synthesis of empirical behavioral research with the concurrent development of a computational and logical framework to facilitate decision-making regarding the nature of the message in its totality. Adapting theories of human interaction and behavior to online interaction and covering both syntactic and semantic issues is extremely complex. A prototype implementation is in progress to evaluate the potential of this approach in various contexts. The current framework is by no means comprehensive considering the scale of the socio-technical problem and has raised more research questions for the future, some of which have been presented in earlier sections. Our future work is focused on enabling detection techniques such as Criteria-based conditional assessment(CBCA) and Reality monitoring(RM)[22] in the context of this framework.

## REFERENCES

- [1] James Allen. *Natural Language Understanding*. Benjamin Cummings, 1994.
- [2] D Appelt. IJCAI tutorial on information extraction. Technical report, SRI, 1999.
- [3] C.R. Berger. Communication under uncertainty. In M.E. Roloff and G.R. Miller, editors, *Interpersonal processes: New directions in Communications Research*. Sage, Newbury, CA, 1987.
- [4] C.F. Bond, A. Omar, U. Pitre, B.R. Lashley, L. Skaggs, and C.T. Kirk. Fishy-looking liars: deception judgement from expectancy violation. *Journal of Personality and Social Psychology*, 63:969–977, 1992.
- [5] C.F. Jr Bond and A.O. Atoum. Deception judgements: Cue theory and beyond. In *Annual meeting of the Society of Experimental Social Psychology*, October 1999.
- [6] D.B. Buller, A. Burgoon, J.K. and Buslig, and J. Roiger. Testing interpersonal deception theory: the language of interpersonal deception. *Communication Theory*, 6:268–289, 1996.
- [7] D.B. Buller and J.K. Burgoon. Interpersonal deception theory. *Communication Theory*, 6:203–242, 1996.
- [8] D.B. Buller, J.K. Burgoon, A. Buslig, and J. Roiger. Interpersonal deception:viii, nonverbal and verbal correlates of equivocation from the bayelas et al. (1990) research. *Journal of Language and Social Psychology*, 13:396–417, 1994.
- [9] D.B. Buller, J.K. Burgoon, C. White, and A.S. Ebesu. Interpersonal deception:vii, behavioral profiles of falsification, concealment and equivocation. *Journal of Language and Social Psychology*, 13:366–395, 1994.
- [10] J.K. Burgoon, A. Buller, D.B. and Ebesu, and P. Rockwell. Interpersonal deception:v, accuracy in deception detection. *Communication Monographs*, 61:303–325, 1994.
- [11] J.K. Burgoon and D.B. Buller. Interpersonal deception:iii, effects of deceit on perceived communication and nonverbal behavior dynamics. *Journal of Nonverbal Behavior*, 18:155–184, 1994.
- [12] J.K. Burgoon and D.B. Buller. Interpersonal deception:iv, effects of suspicion on perceived communication and nonverbal behavior dynamics. *Human Communication Research*, 22:163–196, 1994.
- [13] J.K. Burgoon, D.B. Buller, K. Floyd, and J. Grandpre. Deceptive realities: sender, receiver and observer perspectives in deceptive conversations. *Communications Research*, 23:724–748, 1996.
- [14] J. Carlson and R. Zmud. Channel expansion theory and the experiential nature of media richness perceptions. *Academy of Management Journal*, 42(2):153–170, 1999.
- [15] R. Cole, J. Mariani, H. Uszkoreit, A. Zaenen, and V. Zue. *Survey of the State of the Art in Human Language Technology*. Cambridge University Press, 1999.
- [16] Dorothy Denning. *Information Warfare and Security*. Addison-Wesley, 1999.
- [17] P. Ekman and W.V. Friesen. Nonverbal leakage and clues to deception. *Psychiatry*, 32:88–105, 1969.
- [18] Therani Madhusudan. Autonomous deception detection agents. Technical report, University of Arizona, 2002. Unpublished manuscript.
- [19] Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Prentice-Hall, 1995.
- [20] Bruce Schneier. *Secrets and Lies: Digital Security in a Networked World*. John Wiley, 2000.
- [21] J. R. Searle. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press: Cambridge, England, 1969.
- [22] Aldert Vrij. *Detecting Lies and Deceit: The Psychology of Lying and Implications for Professional Practice*. John Wiley, 2000.