Deception Detection in Human Reasoning

Deqing Li and Eugene Santos, Jr.
Thayer School of Engineering
Dartmouth College
Hanover, N.H., U.S.A
{ Deqing.Li, Eugene.Santos.Jr }@Dartmouth.edu

Abstract—**Even though people deal with deceptions throughout their whole lives, deception detection remains a challenging problem. The average detection rate for humans is only around chance, and detection skill is unlikely to be improved through training. Therefore, researchers have studied the features of deceptive behaviors that were largely ignored in human detection. For example, physiologists look at the physiological signals such as breathing rate and blood pressure, psychologists focus on non-verbal cues such as facial expressions and gestures, and computer scientists search for linguistic cues such as length of sentences. Although they all provide promising results, they seem to neglect a critical part in a person's communication: the reasoning behind the communicated content. In this paper, a method is proposed to detect deception by identifying inconsistencies, explaining the reasoning behind the inconsistencies, and measuring the likelihood of deception based on cues in reasoning. The initial experiment demonstrates the effectiveness of the approach in identifying and explaining communications containing inconsistencies. Reasoning cues that can best discriminate deception from truth are proposed, and aspects of the verification and measurement of such cues as possible future directions of work are discussed.**

Keywords- deception detection; reasoning; multi-agent system; bayesian network; probabilistic; modeling

I. INTRODUCTION

Deception, as Whaley [1] defines, is information designed to "manipulate the behavior of others by inducing them to accept a false or distorted presentation of their environment-physical, social, or political", and, as Burgoon and Buller [2] defines, is a "deliberate act perpetrated by a sender to engender in a receiver's beliefs contrary to what the sender believes is true to put the receiver at a disadvantage". Thus deceiving, in most cases, is a malicious act that can bring long-term and irreparable harm to receivers by updating their knowledge with false information. As such, detection of deception becomes very important because it brings awareness of the false information and the malicious intent of the sender, and undoes his manipulations as necessary. However on the other hand, detection of deception is extremely challenging. Humans perform badly in deception detection, and their performance has been shown to not be improvable through training [3]-[5].

To find mechanisms that help detect deception, researchers typically first focus on behavioral deceptions such as animal behavior [6], children behavior [7], military behavior [8], and internet behavior [9] because plenty of behavioral cues may be revealed subconsciously by people. Some researchers further attempt to study both behavior and communication content, which offer them more evidence for identifying potential conflicts or inconsistencies. However, we believe deception in communication alone possesses significant research potential as well as practical value nowadays since more and more communication content are documented by way of emails, tweeters, online messengers, blogs, sms, e-reports, social networks, and so on. With people relying more frequently on written communications, less and less non-verbal behaviors such as facial expressions will be accessible to detectors. That means many key observables are now missing from people's daily communications, which makes people more vulnerable to deceptions and lowers down their current chances of detection even more.

Without understanding the behavior of the deceiver, it is hard to detect deception, but this happens frequently in real life. For example, reports from intelligence analysts, analyses from financial consultants, and even diagnoses from doctors may contain deception to varying degrees. It can be extremely harmful for people to take the reliability of such reports for granted. To deal with only communication content, researchers for the past few decades have been using linguistic cues such as the length of the sentence, the number of verbs versus the number of nouns, and the use of some indicative words [10] [11]. Although these efforts provide strong theoretical bases as well as insight to the study of human behavior, there are several shortcomings in these approaches. Firstly, since linguistic cues are only leaked subconsciously, the detection is less effective when deceivers are asked to communicate in an official way, and in planned deception, cues can be readily studied by deceivers and words can be crafted in order to avoid the cues. Secondly, wording in answers can be structured by questions. For example, deceivers are generally observed to avoid the use of "I", but if being asked "did you kill the person?" most people will start by saying "Yes, I…" or "No, I…". Lastly, the directions of the cues are largely moderated by contextual factors [10]. Another method that can be used to mine the subjective information from communication content is sentiment analysis [12]. It is able to detect the attitude and emotion of authors at the document, sentence, and feature levels with good performance. However, to classify deceptive and non-deceptive information is a different matter. The application of sentiment analysis on deception detection is very limited. Researchers only succeeded in detecting duplicated opinions of online reviews using sentiment analysis [13].

978-1-4577-0653-0/11/$26.00 ©2011 IEEE

In this paper, we propose a method that not only can identify deception but also explains how the deception was formed given just the communication content. It differs from existing methods by capturing the reasoning process of senders, explaining their methods of manipulations, and searching for reasoning cues in their manipulations. In a reasoning process, people estimate consequences from causes or the other way around based on their prior knowledge. A sequence of such estimations about a specific topic forms the reasoning basis for the topic. By capturing the reasoning process, we make sure that deceivers who avoid linguistic-cue detection by changing their wording will be caught because the semantics of communication remains the same. By explaining the methods of manipulations, we can provide detailed evidence of a suspect to human detectors such that they can further review and/or confirm deception. Finally, by searching for cues that indicate the existence of deception, we also eliminate unexpected communications that are non-malicious. Our intuition is that there are some unique features in the deceivers' reasoning processes that they cannot avoid even though it is easy for them to avoid linguistic cues. These features are observable because i) memories from past experience differ from memories from imagination [14], and ii) there is a discrepancy between how people argue for what they truly believe, or are willing to embrace and how people argue for what they do not in terms of both the behavior and the reasoning [15] [16].

The paper is organized as follows: We first describe the four stages of Johnson et al.'s deception detection framework [17], and then explain how it is utilized in our model in Section II. In Section III, we introduce real life data that are used in our experiments and explain how to construct reasoning bases from the data in natural language for our experiments. We then describe our methodology and present our results in Section IV. Lastly, conclusions and future work are discussed in Section V.

II. Johnson et al.'s Deception Detection Framework

Johnson et al.'s deception detection framework [17] was originally used to automatically detect fraud in an accountant's reports. The authors examined the way auditors detect management's malicious manipulations of financial information so as to make the company appear more profitable than it actually is. They noticed that people can learn detection heuristics from past experience if a particular form of deception is frequent. However, deception detection is a low base-rate task, especially in domains where interactions and feedbacks are available. Thus, people's experience in detecting deception is fraught with failure. The framework [17] then was proposed to address this problem. There are four stages in the framework: Activation, Hypothesis generation, Hypothesis evaluation, and Global evaluation. In Activation, expectations and exact values are compared. The magnitude of the discrepancy between them determines whether inconsistencies are observed. In Hypothesis generation, they propose hypotheses to explain how deceptions are formed from the inconsistencies. Then the hypotheses are assessed on the basis of their materiality in the Hypothesis evaluation stage. Finally, all accepted hypotheses are aggregated and final judgment is produced in the Global evaluation stage. The performance of detection was shown to be better than human auditors.

We apply Johnson et al.'s deception detection framework but made modifications to it because deceptions in communication content are presented as natural language, which has a more flexible form than accountant reports and the verification of which is more subtle than confirming the functionality and the materiality of misrepresented numbers as in auditing accountant reports. The main modifications are illustrated below. In the Activation stage, we assume that there is more than one person giving his opinions on a topic over time and the group of the people shares some level of knowledge [18]. These assumptions are reasonable in professions such as security, medicine and finance. At the same time, these assumptions make sure that truth tellers are expected to be consistent with their past opinions, and because of this consistency, their opinions can be expected, and the discrepancy between the expectation and the exact opinion represents the inconsistencies in people's communication content during the activation. In the next stage "Hypothesis generation", the detection schema of auditing accountant reports is pre-defined by experts because the observed inconsistencies of each type of deception are fixed. However, since manipulations in communication is so flexible that there are an almost infinite number of ways to deceive through different combinations of inventing and hiding, we cannot simply categorize them into specific types, but we can explain how a deception was formed in an automatic fashion. To do that, we retrieve the deceiver's reasoning process on the inconsistent opinions, which indicates how the inconsistencies were structured to convey the conclusion. In the hypothesis verification stage, only assessing the materiality of a hypothesis in manipulating the environment as in auditing accountant reports cannot confirm a deception in communication. We need to evaluate the likelihood of a hypothesis being a true deception based on several reasoning cues that can best distinguish deceptive opinions from true opinions.

III. Construction of Reasoning Base from Natural Text

To conduct our experiment, large amounts of both true and false opinions from multiple subjects as well as their past opinions on the same topic are necessary. Besides, we require the precise classification of each sentence of all the false stories into "honest" and "deceptive" in order to serve as our ground truth. Since real life deception data over a long period are very rare and survey data would require massive human effort, we could not find available data that perfectly fit our experimental setting, and thus we chose survey data from Mihalcea and Strapparava [19] as part of our data, and simulate the additional data as necessary. In the survey data, one hundred test subjects were asked to imagine that they were taking part in a debate. To prepare for their speech, each of them needed to provide a story of at least 4 or 5 sentences to illustrate his true opinion on a topic. Next, they were also asked to provide stories to illustrate false opinions on the topic, that is, to lie about their opinions. For example, on the topic "abortion", subjects may argue about why it is good if they support abortion, and also argue about why it is bad as if they did not support it. They were asked to provide true arguments in the true stories, but were not required to provide all lies in the deceptive stories. It means that we only have ground truth for global lies, that are lies in the opinions on the topic, but not for local lies, that are lies in their individual arguments. We selected the data under the topic “abortion” because there are sufficient numbers of both positive and negative comments on it in both true and false stories, which provides us with complete information of how arguments from different sides influence the conclusion, or in other words, enable us to retrieve the reasoning behind the comments, and at the same time prevent us from being biased to one side of the story.

On the other hand, the simulated data were also based on the survey data. Firstly, a simulated human subject is necessary to provide us with history data and testing data that are consistent over time. A simulated human subject is called an agent in the experiment and the stories are assumed to be the presentations of their reasoning results given some evidence. Retrieval of the reasoning process or even retrieval of the semantic meaning from natural language is a major challenge for natural language processing. The difficulties in our work particularly include the selection of relevant arguments, the extraction of polarity on arguments, and the identification of the causal relationships between arguments. Although we note that sentiment analysis achieves encouraging results in capturing subjective information in communication, it is not fully applicable to our research since sentiment analysis is domain-specific and its capability to retrieve subjective information from objective expressions is yet immature. To capture the reasoning process that produces the stories, we use rule-based keyword mapping combined with some human effort. Specifically, we pick out the most frequent 30 arguments from all true stories by manually picking out the arguments containing the most frequent 30 words. For each of the arguments, we find its correspondence in each story by matching key words that can best distinguish the argument from others and obtain its polarity in the story using sentiment analysis techniques [12]. Then the polarities of all arguments in all the stories are encoded in a matrix in which the rows represent story ids and the columns represent argument ids. If some arguments are missing in a story, its polarity is marked as being in the middle of neutral and negative because by not mentioning some arguments it probably means that the authors did not think the arguments were supportive of their claims. Next, we use the correlations between arguments in the polarity matrix to generate a Bayesian network (BN) [20] in which each node is an argument. In the BN, we use one word to represent the id of each node (argument), and the argument “abort” is the conclusive argument for each story. The arguments and their ids are shown in Table 1. As such, the BN represents the reasoning process of a combination of all one hundred true stories, which can be regarded as a rational person who has both supporting and dissenting opinions on abortion from different perspectives. The impact of individual differences is reduced due to central tendency. In the experiment, this agent is called AgentZero. The algorithm we used to generate the BN is the PC algorithm [21]. PC algorithm is popularly used to generate BNs from correlations and partial-correlations of observations. The basic idea of PC algorithm is that partial correlation between random variables indicates d-separation. The advantage of using PC algorithm is to save computational and memory costs of searching a huge space of all possible BN structures while still being able to distinguish causality from correlation. During the generation of BNs, we varied the threshold of correlation under which the edge between two arguments should not be preserved. A high correlation threshold means that two arguments are allowed to be causally related only when they are strongly correlated. We measured the fit of the BN of each threshold with the true stories. The validation process is as follows: For each story, if the binary polarity of the conclusive node after belief updating corresponds with that in the story given that all non-conclusive nodes that are explicitly mentioned in the story are evidence, then the BN is validated to fit the story. The fit rates of three threshold values are shown in Table 2. It shows that 0.15 correlation threshold generates the BN that fits all the true stories best. 0.1 correlation threshold tends to connect arguments more than necessary and 0.2 correlation threshold tends to connect arguments less than necessary. The best generated BN is displayed in Fig. 1.

TABLE I. MOST FREQUENT 30 ARGUMENTS FROM TRUE STORIES AND CORRESPONDING IDS

Argument Id	Argument
Abort (conclusive argument)	I support Abortion.
Right	Women have the right to do whatever they want with their bodies.
Govern	Government should interfere with people’s decision on abortion.
Care	Unwanted children are put into dodgy care systems.
World	Unwanted children should be brought up in the world.
Life	Unwanted children’s lives are miserable.
Murder	Abortion is murder.
Health	Some pregnant women have health problems.
Option	Abortion is an option.
Time	The time to allow abortion should be fixed.
Early	Abortion should only be allowed at early time.
Population	Abortion can help birth control.
Adopt	Adoption is an option.
Carry	Women should be forced to carry babies.
Child	Children have right to life.
Couple	There are families and couples who want to adopt babies.
Educate	Education should be provided to prevent unwanted pregnancy.
Mistake	People use abortion to correct their mistakes.
Teenager	Some teenagers get pregnant.
Inconvenience	Pregnancy provides inconvenience.
Responsible	People should take responsibility.
Sex	People are forced to have sex.
Birth	Some pregnancies have birth defects.
Human	Unborn children are human.
Concept	Life starts from conception.
God	Religion plays an important role in the decision.
Circumstance	There are circumstances when people need abortion.
Want	People want abortion.
Legal	Abortion is legal.

TABLE II. FIT RATE OF BNS GENERATED WITH 0.2, 0.15 AND 0.1 CORRELATION THRESHOLDS

Threshold of correlation values	Fit rate
0.1	0.802083
0.15	0.822917
0.2	0.8125

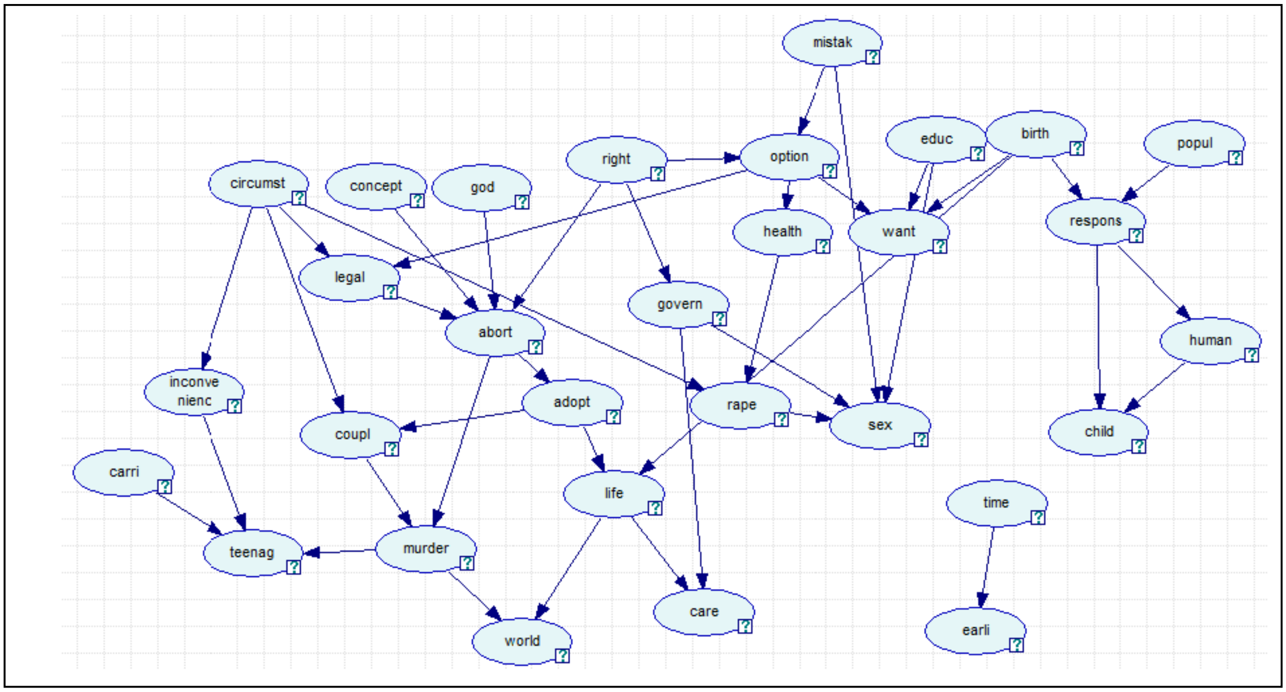

Figure 1. AgentZero generated from one hundred true stories with 0.15 correlation threshold

After AgentZero is generated, all other agents (for the group) are simulated by perturbing the conditional probability tables (CPTs) of AgentZero, which are called AgentOne, AgentTwo, and so on. The slight difference in their CPTs enables them to share similar but not exactly the same uncertainty in knowledge. We use a value from 0.1 to 0.4 to control the level of perturbation, e.g. 0.1 means that every conditional probability in the CPT is shifted within +/-0.1. Secondly, we simulated the normal behavior of all agents. We assume that normally, agents honestly present opinions following their reasoning results. Thus, a repeat of normal behavior is simulated by conducting belief updating on AgentZero, AgentOne, AgentTwo, and so on, given the same set of evidence that were randomly chosen. The posterior probabilities inferred from belief updating are then transformed into binary polarities by comparing which state has a higher probability.

IV. METHODOLOGY

Our system applies the main idea of the four stages in Johnson et al.'s deception detection framework [17]. In this section, we detail our methodology and analyze the experimental results in the first two stages and propose a methodology for the third stage.

A. *Activation stage*

The main purpose of activation is to find out the inconsistencies in the stories. According to the assumption we mentioned in *Section II* that people are consistent over time, even if an agent is different from others, he is consistently different. Thus we still can predict his opinion based on opinions from other people. If his exact opinion deviates too much from what we expected, his opinions are regarded as inconsistent. This is more realistic than to expect agents to agree with each other because distinctive opinions do not make a person deceptive but make him more valuable to the group. To implement this idea, we assume that the agents correlate with each other, that is to say, when Agent A agrees with Agent B in the past, Agent A is expected to agree with Agent B in the future. With this correlation, we can predict the future opinions of an agent given the future opinions of all other agents using GroupLens prediction method [22]. If the predicted opinion of an agent is significantly different from his exact opinion, then his exact opinions are regarded as inconsistent. A pilot study of the activation stage using simulated data can be found in [18] and [23]. In both works, we evaluated the detection rate of inconsistency while varying some parameters, namely, the number of agents, the repeat, the perturbation value, the number of evidence, and the time of standard deviations (stdev) that determines how much deviation we allow from exact opinion to expected opinion. To apply the framework to real life data, the methodology is modified from [18] and [23] because the data we use in this paper contain both binary values and probabilistic values. In practice, we first simulate the normal behavior of all agents and calculate the Pearson correlations of the binary polarities in the networks between each pair of agents using (1).

$$r_{AB} = \frac{Cov(A,B)}{\sigma_A \sigma_B} = \frac{\sum_i (A_i - \overline{A})(B_i - \overline{B})}{\sqrt{\sum_i (A_i - \overline{A})^2 \sum_i (B_i - \overline{B})^2}} \qquad (1)$$

where r_{AB} represents the Pearson correlation coefficient. For the i^{th} set of evidence, we define A_i as the polarity of expert A, and B_i as the polarity of expert B. $\bar{A}$ denotes the average of all polarities inferred from expert A's knowledge base given different sets of evidence, and likewise for B. In order to estimate what is a reasonable deviation from the prediction, we predict the normal behavior using GroupLens as in (2).

$$A_{X_{\Pr ediction}} = \bar{A} + \frac{\sum (B_{iX} - \bar{B}_i) r_{AB_i}}{\sum_i | r_{AB_i} |} \qquad (2)$$

where $A_{X_{predition}}$ denotes the predicted value of A. For the i^{th} agent, r_{ABi} denotes the Pearson correlation coefficient between A and B_i, and X denotes the random variable whose state is unknown by A, but is available by B_i. The error between the predicted value and the exact normal behavior is an estimation of reasonable deviation, that is, noise. In the test stage, we assume that only AgentZero exhibits deceptive behavior, and thus we use the polarity matrix of true stories again to represent true opinions of AgentZero in order to measure false alarm rate, and the deceptive behavior of AgentZero is represented by the polarity matrix of the false stories. We manually categorize the arguments into subjective arguments that demonstrate a strong attitude towards abortion and beliefs of some theories such as "unborn children are human" and objective arguments that serve as the premises on which abortion may be controversial such as "some pregnant women have health problems". Objective arguments are regarded as evidence if explicitly mentioned in the stories, and subjective arguments are regarded as reasoning results. Likewise, we predict the normal behavior for true stories using GroupLens with the evidence in true stories, and predict the normal behavior for false stories in the same way. If the error between the predicted value and the true opinion of AgentZero is beyond four stdev, that is, with 0.0004% of being noise, then the detection is a false alarm. If the error between the predicted value and the deceptive behavior is beyond four stdev, the detection is activated. Since the 100 true stories are provided by 100 different test subjects, their individual differences introduce noise to the inconsistency in true stories. Besides, the evidence combinations in the true stories are limited. To eliminate the individual differences and to cover a larger spectrum of possible evidence, we produce another set of true opinions artificially by sampling AgentZero with random evidence, but we note that some combinations of evidence may not occur in real world. To summarize, the process of activation can be simply described as in Table 3. Table 4 shows the performance of activation using our semi-simulated data.

It shows that around 8.6% of the arguments in false stories are classified as inconsistent, around 7.3% of the arguments in real true stories are classified as inconsistent and 1.8% of the arguments in artificial true stories are classified as inconsistent. Although the inconsistency rates in the false stories seem really low, it is reasonable since normally only part of the sentences in false stories are deceptive and we do not have the ground truth of that fidelity.

TABLE III. STEPS OF ACTIVATION STAGE

Step1	Build AgentZero from 100 true stories and simulate other agents.
Step2	Sample over the agents to simulate a history of opinions.
Step3	Obtain the correlations between agents, predict the history of opinions and measure the prediction errors.
Step4	Measure inconsistency in false story by taking the differences between false stories and predicted false stories
Step5	Measure inconsistency in true stories by taking the difference between real/artificial true stories and predicted true stories.

TABLE IV. DETECTION PERFORMANCE OF ACTIVATION STAGE WITH SEMI-SIMULATED DATA

Parameters	Agents = 10, Repeats = 100, Perturbation = 0.1, Evidence = 1-8, Times of std = 3		
	Inconsistency rate in False Story	***Inconsistency rate in Real True Story***	***Inconsistency rate in Artificial True story***
Max	0.61	0.36	0.0824
Min	0.0	0.0	0.0
Mean	0.0858	0.0729	0.0180
No. of stories with inconsistency	79	71	9

If we regard a story as detected if at least one of its arguments is detected, then the recall is 0.79, the precision with regard to real true stories is 0.5267, and the precision with regard to artificial true stories is 0.8977. The low precision with regard to real true stories shows that AgentZero seems to be unable to recover the original stories probably due to the individual differences between the test subjects who wrote the stories. We will show in the following stages that the inconsistencies from false stories and from true stories exhibit different features which can be used to distinguish their source of inconsistency.

B. Hypothesis Generation Stage

When deceiving, people usually manipulate one or several tokens in the reasoning chain, and let the rest of the tokens change accordingly. Finally the receivers would naturally infer the false conclusions by themselves. Manipulations come in various forms, but they have a common goal which is to convince the receiver of the false conclusion. Hypothesis generation provides a way to explain the flow of potential deceptions and shows how the false information was conveyed to him without being noticed. Besides, it also provides a skeleton, based on which we can retrieve the cues that indicate the existence of deception in hypothesis verification. Therefore, the main purpose of hypothesis generation is to identify the flow of the inconsistencies in the original BN structure. To do that, we calculate the correlations of the polarities between the inconsistent nodes within AgentZero using training data, and generate a BN structure using PC algorithm. Figure 2 shows a hypothesis generated from the following false story.

> *An unborn child is still a living being, who should be given the rights any other human has. This includes the right to life. No person, including the mother of an unborn child has the right to kill a living human being. If a woman finds herself with an unwanted pregnancy, she has the option to put the baby up for adoption. There are*

many loving couples who would love the opportunity to adopt a baby.

Figure 3 shows a hypothesis generated from the following true story.

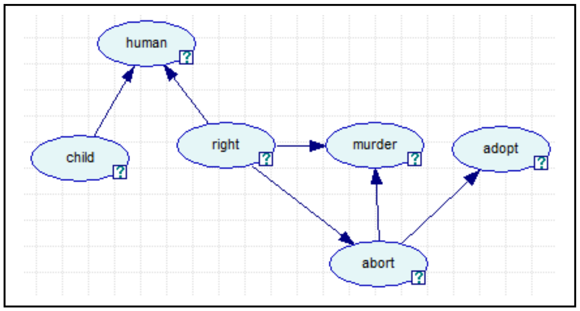

Figure 2. A hypothesis generated from a false story

Abortion is wrong and should never be resorted to, unless in very extreme cases where the life of the mother and/or the fetus is in danger. Even then it should be used sparingly. Abortion is not birth control. Birth control is something you do BEFORE a baby is conceived. Abortion is killing a baby that you have already conceived. And life begins at conception - a baby just a few weeks old has a beating heart. Some people would argue that abortion is a woman's choice, but what about the child's choice? The baby is being killed for her convenience. It is totally wrong. When an unwanted pregnancy occurs, the only humane option is adoption. Abortion is deliberately taking steps to end a life, to stop a beating heart. That is murder.

Note that consistent arguments in the stories are not included in the hypotheses. Although the exact meanings of the arguments vary slightly in the stories, both of the hypotheses propose reasonable connections between inconsistent arguments. For the true story, we notice that the hypothesis presents not only the arguments that were possibly fabricated by the author but also those (such as "rape") that were possibly hidden by the author.

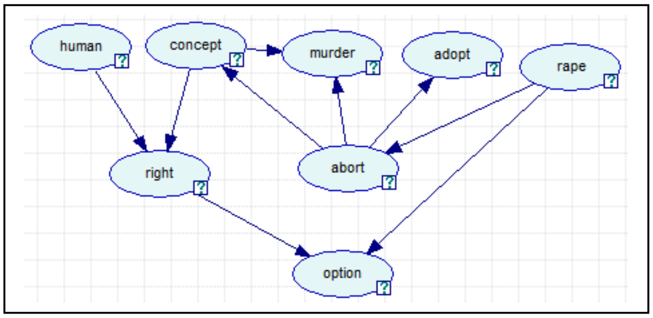

Figure 3. A hypothesis generated from a true story

C. Hypothesis Verification Stage

The activated inconsistencies detected from the activation stage may not necessarily be due to deception. It could be false alarm, opinion change or misinformation. False alarm is the honest words that are mistakenly classified as deception. Opinion change is the change in the reasoning base rather than the intentional change in the presentation. Technically, it is the change in conditional probabilities instead of the change in posterior probabilities. Misinformation differs from deception (also called disinformation in order to contrast misinformation) in that the speaker is not aware of the falsity in his presentation and does not present the false information on purpose. Since deceiver' intent is to convince the receiver of his attitude on a topic, all his arguments is derived from his conclusion. On the other hand, the truth teller intends to reach a conclusion from his arguments and evidence, and thus their arguments are driven by evidence. As such, we expect some observables in the properties of the deceivers' reasoning base due to the discrepancy between intentional manipulations and unintentional deviations from expectation.

There are three tasks we need to accomplish in this stage: (i) identify cues that can best discriminate deception from truth, (ii) automatically represent and measure cues, and (iii) build classification models to predict deception. Due to the lack of existing theories on the reasoning behavior of deceivers, we borrowed DePaulo et al.'s categories of linguistic cues [10], and propose the cue that *Hypothesis that intends to manipulate the final conclusion is more likely to be a deception (abbreviated as C1)*. It can be understood intuitively since if the conclusion is not manipulated all the other manipulations are fruitless. C1 is measured by the existence of the argument "abort". By classifying based on C1, we identified 61% of the hypotheses from false stories as deception, 30% of the hypotheses from real true stories as deception and 2% of the hypotheses from artificial true stories as deception. Although the detection on false stories does not seem very compelling, we believe that the performance will be improved if combinations of various cues were considered [10]. Following DePaulo et al.'s categories, we also propose the following categories from which the cues might be extracted.

1) (C2) Hypothesis in which critical arguments are manipulated less than noncritical arguments is more likely to be a deception.

2) (C3) Hypothesis that shows less diversity in arguments is more likely to be a deception.

3) (C4) Hypothesis in which all the nodes are functional to the conclusion, that is to say, increase the likelihood of the conclusion, is more likely to be a deception.

C2, C3 and C4 have not been studied yet in sufficient detail in this initial experiment. They will be explored and tested in future work. However, we believe these cues are sufficiently important to warrant a brief discussion. We note that since deceivers want to avoid arguments that can be easily disproved and that sometimes they have no idea what the best arguments could be if their conclusion were true, they tend to raise arguments that do not have a significant impact on the conclusion, e.g. when a person lies about his best friend he tends to illustrate all the nice characteristics of this person instead of how they interact with each other, which is not a very good reason for being a best friend, so this is why we propose C2. C2 also corresponds to the findings of Zuckerman et al. [24] that deceivers communicate in a less forthcoming way. C3 is true because deceivers display less diversity at the content level than did truth tellers [11], e.g. a truth teller who describes his best friend talks about all kinds of thing they have done together. Both C2 and C3 are due to the fact that truth tellers are backed by an accumulation of knowledge and experience that deceivers cannot imagine [25], and thus deceivers not only have less familiarity with the subject they are describing but also to allow fewer opportunity of being disproved [26]. We observe C4 because deceivers do not appreciate that memory is fallible and that stories always have two sides when they are defensive of their credibility [10]. Consequently, truth tellers are free to express uncertainties and arguments against their conclusions, while deceivers tend to include fewer imperfections in their stories. All of these cues are quite robust to the careful craft of wording and the change of contextual factors due to the nature of deceit and the deceivers' lack of true knowledge.

V. CONCLUSION AND FUTURE WORK

We proposed a model that identifies deception from communication content. It also can provide an explanation of how the deception was formed and what the deceiver intends to convey without any information about the intent of the deceiver. The model followed Johnson et al.'s detection framework, accomplished the first two stages of activating inconsistency and generating hypothesis of the flow of deception, and provided an initial solution for the third stage by defining reasoning cues to indicate the existence of deception. The contributions of this paper are as follows:

- We developed an approach to retrieve reasoning, represent reasoning and explain reasoning in terms of deception.
- Deceivers not only fabricate information but also hide information, which results in the incompleteness of relevant observables. Besides, the line between truth and deception is not always clear [10], so it is necessary to deal with the uncertainty in deception. Instead of just identifying the existence of deception, our approach is able to further pinpoint the information that is either being fabricated or being hidden by deceivers.
- We developed an approach to eliminate individual differences, which prevents the detection of distinctive but nondeceptive thoughts.
- We proposed cues in human reasoning that can discriminate deception from truth and other unexpected stories without malicious intent. These cues are relatively robust to "crafting" of words and changes of contextual factors.
- Our approach can potentially be used to train human detectors and serve as decision aids for them. It is promising especially where text is the only source of information.

Our future work is as follows: Since the performance of our approach depends very much on the extraction of the reasoning behind the data, we plan to improve the natural language processing technique applied to the data. Building models of the binary polarities on the arguments loses a lot of information such as the level of agreement. Besides, it is very time consuming to assign the polarities manually. Therefore, we will develop a method to measure the agreement of each story on each argument automatically. Moreover, the direction of the generated hypothesis is fairly arbitrary in the current experiment. We may propose a pre-defined hierarchy of the arguments to improve the directing job. Also, we will continue hypothesis verification by verifying the categories of cues, measuring the cues and classifying deception based on the combination of the cues. Lastly, Mihalcea and Strapparava's survey data was chosen for our preliminary study because deception data that is available to public is very rare and opinion-based lies are more difficult to catch than event/fact-based lies due to the fact that the difference between false opinion and changed opinion is ambiguous while false facts are definite and that deceivers are more likely to have complete knowledge to mimic true opinions than to mimic true events/facts. We will try to apply our method on other datasets such as the CSC (Columbia-SRI-Colorado) Deception Corpus [27], which records event-based lies and provides ground truth for both global and local lies.

ACKNOWLEDGMENT

This work was supported in parts AFOSR Grant Nos. FA9550-10-1-0499, FA9550-09-1-0716 and FA9550-07-1-0050, DHS Grant No. 2008-ST-061-BS0001-03, DTRA Grant No. HDTRA1-10-1-0096, and an ONR MURI.

REFERENCES

[1] B. Whaley, "Toward a general theory of deception," Military Deception and Strategic Surprise, J. Gooch and A. Perlmutter, Eds. London, U.K.: Frank Cass, 1982.

[2] J. Burgoon and D. Buller, "Interpersonal deception: III. Effects of deceit on perceived communication and nonverbal behavior dynamics," Journal of Nonverbal Behavior, vol. 18, no. 2, pp. 155-184, June 1994.

[3] M. G. Millar and K. U. Millar, "The effects of cognitive capacity and suspicion on truth bias," Communication Research, vol. 24, no. 5, pp. 556-570, Oct. 1997.

[4] C. V. Ford, Lies! Lies! Lies! The Psychology of Deceit. Washington. DC: American Psychiatric Press, 1996.

[5] P. E. Johnson, S. Grazioli, K. Jamal, and R. G. Berryman, "Detecting deception: adversarial problem solving in a low base-rate world," Cognitive Science, vol. 25, no. 3, pp. 355-392, June 2001.

[6] J. B. Bell and B. Whaley, Cheating and Deception. Transaction Publishers, 1991.

[7] B. Sodian, "The development of deception in young children," British Journal of Developmental Psychology, vol. 9, pp. 173-188, 1991.

[8] C. Cruickshank, Deception in World War Two. New York: Oxford University Press, 1979.

[9] S. Grazioli and S. L. Jarvenpaa, "Perils of internet fraud: an empirical investigation of deception and trust with experienced internet consumers," IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, vol. 30, no. 4, pp. 395-410, July 2000.

[10] B. M. DePaulo, J. J. Lindsay, B. E. Malone, L. Muhlenbruck, K. Charlton, & H. Cooper, "Cues to deception," Psychological Bulletin, vol. 129, pp. 74–112, 2003

[11] L. Zhou, J. K. Burgoon, D. P. Twitchell, T. Qin, and J. F. Nunamaker. "A Comparison of classification methods for predicting deception in computer-mediated communication," Journal of Management Information Systems, vol. 20, no. 4, pp. 139–165, 2004.

[12] B. Liu, "Sentiment Analysis and Subjectivity", Handbook of Natural Language Processing Issue, 1st ed., Taylor and Francis Group, Eds. CRC Press, 2010, pp. 1-38.

[13] N. Jindal and B. Liu, “Opinion spam and analysis”, Proceedings of the Conference on Web Search and Web Data Mining (WSDM), pp. 219-230, 2008.

[14] M. K. Johnson & C. L. Raye, “Reality monitoring”, Psychological Bulletin, vol. 88, pp. 67-85, 1981.

[15] A. Mehrabian, Nonverbal communication, Chicago: Aldine Atherton, 1972.

[16] M. Wiener & A. Mehrabian, Language within language: Immediacy, a channel in verbal communication, New York: Appleton-Century-Crofts, 1968.

[17] P. E. Johnson, S. Grazioli, K. Jamal, and R. G. Berryman, “Detecting deception: adversarial problem solving in a low base-rate world,” Cognitive Science, vol. 25, no. 3, pp. 355-392, June 2001.

[18] E. Santos, Jr. and D. Li, "Deception Detection in Multi-Agent Systems," IEEE Transactions on Systems, Man, and Cybernetics: Part A, vol. 40, no. 2, pp. 224-235, 2010.

[19] R. Mihalcea and C. Strapparava. “The lie detector: Explorations in the automatic recognition of deceptive language”, Proceedings of the ACLIJCNLP 2009 Conference Short Papers, 2009.

[20] J. Pearl, Probabilistic Reasoning in Intelligent Systems. San Francisco, CA: Morgan Kaufmann Publishers, 1988.

[21] P. Sprites, C. Glymour and R. Scheines, Causation, Prediction, and Search. New York: Springer-Verlag, 1993.

[22] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl, “GroupLens: An open architecture for collaborative filtering of Netnews,” Proceeding of the Conference on Computer Supported Cooperative Work, Chapel Hill, NC: ACM Press, 1994, pp. 175-186.

[23] E. Santos, Jr., D. Li, and X. Yuan, “On deception detection in multi-agent systems and deception intent”, Proceeding of SPIE, Orlando, FL: SPIE, March 2008, vol. 6965.

[24] M. Zuckerman, B. M. DePaulo & R. Rosenthal, “Verbal and nonverbal communication of deception,” Advances in experimental social psychology, L. Berkowitz, Eds. New York: Academic Press, 1981, vol. 14, pp. 1-59.

[25] H. Markus, “Self-schemata and processing information about the self”, Journal of Personality and Social Psychology, vol. 35, pp. 63-78, 1977.

[26] A. Vrij, Detecting lies and deceit. Chichester, England: Wiley, 2000.

[27] J. Hirschberg, S. Benus, J. M. Brenier, S. F. F. Enos, S. Gilman, C. Girand, M. Graciarena, L. M. A. Kathol, B. Pellom, E. Shriberg, and A. Stolcke, “Distinguishing deceptive from nondeceptive speech”, Proceedings of Eurospeech, Lisbon: ISCA, 2005.