A Dialog System to Detect Deception

Yuiko Tsunomori, Graham Neubig, Takuya Hiraoka, Masahiro Mizukami, Sakriani Sakti, Tomoki Toda, Satoshi Nakamura

Abstract When humans attempt to detect deception, they perform two actions: looking for telltale signs of deception, and asking questions to attempt to unveil a deceptive conversational partner. There is a significant amount of prior work on automatic deception detection, which focuses on the former. On the other hand, we focus on the latter, constructing a dialog system that asks questions to attempt to catch a potentially deceptive conversation partner. We propose several dialog strategies for this task, and measure the deception detection accuracy of each, finding that a more intelligent dialog strategy is slightly more effective at detecting deception.

1 Introduction

Dishonesty is a fundamental part of human life, and thus there is a significant interest in figuring out whether a particular conversational partner is telling the truth or not. Because it is known that it is not easy to detect deception during dialog, skilled interrogators use many techniques to detect deception [4], which include both looking for telltale signs and asking questions so that the features that give away a liar are more easily exposed [11].

In recent years, research on detecting deception automatically using machine learning techniques has achieved some degree of success [6, 8]. For example, Hirschberg et al. [6] performed deception detection experiments over dialogs in which an interviewer questions an interviewee, and used acoustic/lexical information to achieve an accuracy of 66.4%, higher than the chance rate of 60.2%. It should be noted that this previous research deals with only detecting deception in a particular, already performed dialog. In the analogy to human interrogators, this is equivalent to "looking for the telltale signs of deception," which, while important, is only half of the interrogators job. The other half, asking questions to cause deception

Yuiko Tsunomori, Graham Neubig, Takuya Hiraoka, Masahiro Mizukami, Sakriani Sakti, Tomoki Toda, Satoshi Nakamura
Nara Institute of Science and Technology, e-mail: neubig@is.naist.jp

features to be exposed, has not been covered in previous work, with one exception being our previous research in which [10] we focused on dialog in which an interviewer is questioning an interviewee, and examined which types of questions can cause a deceiver to show signs of deception (explained in Section 2).

In this paper, based on this analysis, we design a deception detecting dialog system. The goal of the system is, like that of a human interrogator, to perform utterances that make the interlocutor reveal acoustic or linguistic signs of deception. To do so, we first perform data-driven dialog modeling on a corpus of actual interviews (Section 3), then propose two dialog strategies based on these models and our previously performed analysis (Section 4). Pilot experiments in which users converse with dialog systems using various strategies find that the choice of a dialog strategy has a small but noticeable effect on deception detection accuracy (Section 5).

2 Dialog Scenario and Analysis

Interviews are one of the most common situations in which deception may occur (because the interviewee wants to be seen in a favorable light), but should ideally be detected (because the interviewer wants to assess the interviewee appropriately). The English CSC corpus [3] and the Japanese JDC corpus [10] are recorded using an interview-based scenario as follows:

1. The experimenter tells subjects that the experiment seeks to identify individuals who fit a "target profile" for 6 areas (politics, music, geography, food, interactive, and survival).
2. Subjects take a written test in the 6 areas before starting the interview.
3. Test scores are manipulated so that all subjects score too high to fit the profile in 2 areas, too low in 2, and correctly in 2. Subjects are told these scores.
4. Subjects are told that the experiment is actually about identifying people who can convince others, and those who convince the interviewer that they fit the target profile in all areas can get a prize.
5. Subjects attempt to convince the interviewer that their scores in each of the 6 areas match the target profile, while the interviewers attempt to uncover the truth.

Thanks to studies on deception detection using this data [6, 5], and it is now possible to detect, to some extent, whether a dialog participant is lying from acoustic or prosodic cues. On the other hand, in our previous work [10], we performed a preliminary examination of what can be done for more proactive deception detection; not *"how can we detect the clues of deception"*, but *"how can the interviewer elicit clues of deception?"* In this examination, we found two traits of interviewer questions that were able to effectively improve the accuracy of deception detection. First, "check" questions, which attempt to confirm priorly stated information were effective, likely because it may make a liar more nervous to have the interviewer confirm whether what they said previously is correct. Second, short questions were also found to be effective, presumably because the interviewee has less time to think

Table 1 Example categories of utterance contents

Category	Topic	Test	Score
About	Whether the interviewee matches the target profile e.g. "How did you do on the music section?"	Contents of the test e.g. "What kind of problem question was given?"	Score of the test e.g. "What score do you think you got on the test?"

about their answer. In the following sections, we incorporate these insights into the design of our deception detecting dialogue system.

3 Modeling Interview Dialog

To construct the skeleton of our dialog manager, we first define dialog acts, then build a data driven model for interview dialog.

To define dialog acts for interview dialog, we start with general purpose functions (GPFs; [2]), a general framework for defining dialog acts in conversation. Particularly, due to the features of interview dialogue, we can assume that the interviewer will be in control of the conversation and also will mainly be asking questions to the interviewee, we solely model the interviewer's question dialog acts in four categories: yes/no propositional questions (ProQ), questions requiring a set of information as an answer (SetQ), questions requiring a choice (ChoiceQ), and questions confirming provided information (CheckQ). These dialog acts are still too coarse for our purposes, so we sub-divide the dialog acts to the point where they can be used to drive the proposed interviewer dialog system. Specifically, we sub-categorize the 4 question GPFs with 11 categories expressing the utterance contents, some examples of which are shown in (Tab. 1). We hand-annotated 7 dialogs in the JDC corpus with these fine-grained tags by hand.

Based on these fine-grained dialog acts, we train hidden Markov Models (HMMs) to learn the dialog structure of interviews. We assume that each fine-grained dialog act is an observation, and train HMMs to model the dialog sequences using the EM algorithm. We change the number of the states between 1 and 10, and perform 100 HMM training runs until convergence for each number of states, so that the total number of trained HMMs is 1,000. Finally, we choose the most suitable of the 1,000 HMMs using the the Minimum Description Length (MDL) criterion [9].

4 Rule-based Dialog Strategies

Based on the insights gained from the model trained in the previous section, as well as the analysis from a previous study [10], we next devise two dialog strategies that have the potential to elicit signs of deception. Because of the relatively small training data, instead of taking an entirely data-driven approach, we instead opt to use the model as guidance, but create the actual dialog strategy by hand.

Static Strategy: In this system, we create a dialog strategy that proceeds through the dialog asking the same questions regardless of the user response. The order of the questions is structured manually to generally follow the structure learned by the HMM model. For example, for each topic the interviewer generally asks whether or not the interviewee fits the target profile, then asks the reason why the user believed so. To ensure that the dialog is not too monotonous, we slightly vary the pattern from topic to topic.

CheckQ Strategy: One of the insights obtained from our previous research was that interviewers tend to use check questions CheckQ when they are suspicious of the interviewee's previous utterance, and these CheckQ questions also proved the most effective in eliciting signs of deception. To attempt to take advantage of this fact, we create an additional strategy that treats CheckQ differently from other questions. Specifically, when training the HMM model to decide the dialog structure, we first remove all the CheckQ questions from the corpus, then create a fixed dialog strategy that is the same as the above dialogue strategy with the exception of not including any CheckQ actions. Next, in real-time, the system calculates the probability of a lie for each user utterance using the deception detection module, the automatic parallel of guessing whether the user's utterance is "suspicious" or not. Finally, if the probability of a lie is over 0.5, the system performs a CheckQ action, attempting to catch the user on any suspicious utterances. After this CheckQ question, the system continues where it left off with the original fixed strategy.

5 Experimental Evaluation

In our experimental evaluation, we examine the effect of the choice of dialog strategy on deception detection accuracy by having users perform dialog with systems using different strategies, and measure the accuracy of the deception detection model over dialogs performed using each strategy. We perform the experiment using a WoZ setting, in which the wizard performs spoken language understanding (SLU) and natural language generation (NLG). In NLG, the system has a question template for each dialog act, and the Wizard generates reply sentences based on this template. The templates are intentionally chosen to be short, 1-10 words, as short questions were shown be effective in our previous analysis. In addition, to prevent entrainment to unnatural speech synthesis, or other differences resulting from talking to systems from affecting results, we have the wizard read the sentences directly to the user. For deception detection, we use the acoustic/prosodic features [10], which include fundamental frequency F_0, power, and phoneme duration, and use Bagging of decision trees [1] as a classifier.

The theme of the dialog is chosen among the 6 areas in the tests in JDC. Regardless of the result of the test, the subjects attempt to convince the wizard that their scores matched the target profile. In preliminary experiments, we found that it was not possible to obtain sufficient deception detection accuracy with no training data for the target speaker, so we recruited 8 users that participated in the collection of JDC, and used the JDC data for each interviewee as deception detector training data.

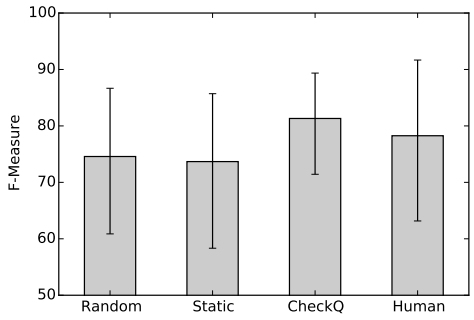

Fig. 1 F-measure of deception detection for each system

Table 2 Example dialogs translated from Japanese (SP: Speaker / W: Wizard, S: Subject)

System	SP	Dialogue	System	SP	Dialogue
Random	W	What score do you think that you got on the test?	CheckQ	W	How did you do on civics?
	S	I think about 80.		S	I matched the profile.
	W	How did you do on civics?		W	Is that true?
	S	I matched the profile there.		S	Of course.
All	W	How did you do on music?	Human	W	How did you do on civics?
	S	I matched the profile there.		S	My score matched the profile.
	W	Why do you think so?		W	Is that true?
	S	I've played piano since I was young.		S	Yes.

The subject talks once to each of the systems, in random order by subject to prevent conditioning effects. Following JDC, the veracity of subjects' utterances is annotated by having all subjects push a "truth" or "lie" button during the experiment for each utterance. We consider the users' utterances as test data and detect deception every utterance and evaluate the deception detection accuracy for each system.

Based on this protocol, we evaluated the following systems.

Random: Randomly choose the system dialog acts with equal probability.
Static: Static strategy of Section 4.
CheckQ: CheckQ strategy of Section 4.
Human: A system where the wizard directly chooses which dialog act to use. In this case, the lie probability is still calculated by the deception detector, and the wizard is allowed to reference this probability when choosing then next act.

In Fig. 1, we show the F-measure of deception detection for utterances elicited by each system, and in Tab. 2, we show an example of dialogs for each system. Error bars represent the 95% confidence interval according to bootstrap resampling [7]. From these results, we can see that, while the error region is large due to the relatively small number of utterances per system, the CheckQ rule system appears to have slightly higher deception detection accuracy than the other two automated systems, and at a similar level to Human. On the other hand, there was no obvious difference between the Random policy and the Static policy trained on the interview

corpus, indicating that the utilization of deception detection during dialogue and appropriately timed CheckQs is likely the reason for the difference.

Taking a look at the example dialog (which is relatively representative of the dialogs as a whole), we can see that these results make sense intuitively as well. The dialog of CheckQ and the Human system are close, being both natural, and taking advantage of CheckQ questions to attempt to confirm previous information, potentially catching the user off guard. Interestingly, the percentage of lies for the CheckQ system (68%) was also most similar to Human (65%), compared to Static (71%) which elicited more lies, and Random (53%), which elicited fewer lies.

6 Conclusion and Future Work

In this paper, we described construction of a deception detecting dialog system, including modeling of interviewer's questions in human-human interaction, and construction of dialog strategies based on these models and other insights. In pilot experiments, we found that a system utilizing questions that confirm previous information achieves slightly higher accuracy than two other baselines.

It should be noted that these pilot experiments are on a somewhat small scale, with only eight participants, and thus it is necessary to perform large-scale experiments to confirm these tendencies with more statistical power. Once these experiments have been confirmed, we aim to create a fully automatic dialog system that is not reliant on a Wizard of Oz setting, and test this system "in the wild" with more broad-ranging deceptive tasks.

References

1. Breiman, L.: Bagging predictors. Machine learning **24**(2), 123–140 (1996)
2. Bunt, H., et al.: ISO 24617-2: a semantically-based standard for dialogue annotation. In: Proc. LREC (2012)
3. DePaulo, B.M., et al.: Cues to deception. Psychological bulletin **129**(1), 74 (2003)
4. Ekman, P.: Telling Lies. W. W. Norton & Company (1985)
5. Enos, F.: Detecting deception in speech. Ph.D. thesis, Columbia University (2009)
6. Hirschberg, J., et al.: Distinguishing deceptive from non-deceptive speech. In: Proc. Eurospeech (2005)
7. Koehn, P.: Statistical significance tests for machine translation evaluation. In: Proc. EMNLP (2004)
8. Pérez-Rosas, V., Mihalcea, R.: Cross-cultural deception detection. In: Proc. ACL (2014)
9. Stolcke, A., Omohundro, S.: Hidden Markov model induction by bayesian model merging. Advances in neural information processing systems (1993)
10. Tsunomori, Y., et al.: An analysis towards dialogue-based deception detection. In: Proc. IWSDS (2015)
11. Vrij, A., et al.: Outsmarting the liars: Toward a cognitive lie detection approach. Current Directions in Psychological Science **20**(1), 28–32 (2011)