An Analysis Towards Dialogue-based Deception Detection

Yuiko Tsunomori, Graham Neubig, 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 has been significant prior work on automatic deception detection that attempts to learn signs of deception. On the other hand, we focus on the second action, envisioning a dialogue system that asks questions to attempt to catch a potential liar. In this paper, we describe the results of an initial analysis towards this goal, attempting to make clear which questions make the features of deception more salient. In order to do so, we collect a deceptive corpus in Japanese, our target language, perform an analysis of this corpus comparing with a similar English corpus, and perform an analysis of what kinds of questions result in a higher deception detection accuracy.

1 Introduction

Dishonesty is a fundamental part of human life, and as a result 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 a number of techniques to detect deception [3], which include both looking for telltale signs and asking questions so that the features that give away a liar are more easily exposed [13].

In recent years, there has been research on detecting deception automatically using machine learning techniques, and these works have achieved some degree of success. For example, Hirschberg et al. [4] performed deception detection experiments on an English corpus including deception (the CSC corpus) using acoustic/prosodic and lexical features, and achieved an accuracy (66.4%) higher than the

Y. Tsunomori, G. Neubig, S. Sakti, T. Toda, S. Nakamura,
Nara Institute of Science and Technology (Japan)
e-mail: {tsunomori.yuiko.tq1, neubig, ssakti, tomoki, s-nakamura}@is.naist.jp

chance rate (60.2%). In addition, Pérez-Rosas and Mihalcea [9] reported that there are difference in the lexical characteristics of deception between cultures or languages, although they make no mention of acoustic/prosodic features.

It should be noted that this previous research deals with only detecting deception in a particular, already performed dialogue. 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 features to be exposed, has not been covered in previous work. In our work, we envision a deception detecting dialogue system that can detect deception by not only looking for the telltale signs, but also by asking questions to cause features of deception to be exposed. In this paper, we take a first step towards this goal by identifying not only which features can be used to create a deception detecting classifier, but also which types of questions can cause a deceiver to show signs of deception. If these questions are made clear, in future work it will be possible to create a dialogue system that focuses on these questions, and thus may be more effective at eliciting signs of deception.

In this research, we make two contributions:

- The first is that, as our target language is Japanese, we collect a Japanese corpus modeled after similar English corpora of detective speech. We perform deception detection experiments using these corpora and comparing features, both lexical and acoustic/prosodic, that can be used to detect deception effectively in Japanese and English.
- The second, and main, novel contribution is that we analyze which types of questions made by the interrogator make it possible to detect deception effectively on this corpus. Specifically, we examine the dialog act of questions and lengths of questions that elicit utterances that are easy or difficult to classify.

2 Collection and Annotation of the Japanese Deception Corpus

Before performing research on data-driven deception detection, it is necessary to have a corpus, and a number of resources have been created in previous works. The CSC corpus [4] recorded interviews where subjects were encouraged to lie to an interviewer, and were motivated by financial incentive to deceive successfully. Interviews were performed in English, with a total of 22 interviews ranging from 25 to 50 minutes. Furthermore, there is the Idiap Wolf Corpus [5], an audio-visual corpus containing natural conversational data of volunteers who took part in a competitive role-playing game in a group conversational scenario. Four groups of 8-12 people were recorded in English.

However, while excellent resources exist for English, there are fewer resources for other languages. In Japanese, there is the Indian Poker corpus [8], an audio-visual corpus containing natural conversational data of 18 subjects who took part in 3-person games of Indian poker. However this resource is not publicly available, and because we assume a one-on-one dialogue system, a corpus recorded with three participants is not suitable for our research. Thus, as a first step in our research, we collect a corpus of deceptive utterances with two goals: first to allow comparative studies of deception detection in speech across languages and cultures, and also to provide further resources for our work on deception detecting dialogue systems, which will use Japanese as its target language. To do so, we collect dialogs, make transcriptions, and annotate "lie" labels under the same conditions as the CSC corpus [4].

2.1 Corpus Collection

In order to collect our corpus of deceptive speech, we follow the recording paradigm of the CSC corpus, as we describe below. As an example of scenes in which deception regularly occurs, the dialog recording assumes a self-presentational dialogue [2] between an interviewer and interviewee. The recording process is 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. The subjects take a written test in the 6 areas before starting the interview.
3. The 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. The experimenter tells the subjects the score.
4. The subjects are told that the experiment is actually about identifying people who can convince others that they fit the target profile in all areas. They are told that those who succeeded at deceiving the interviewer into believing that they fit the target profile in all areas can get a prize.
5. The subjects attempt to convince the interviewer that their scores in each of the 6 areas matched the target profile. The interviewers' task is determining how subjects had actually performed, and the interviewer is allowed to ask any questions other than those that were actually part of the tasks the interviewee had performed.

2 people were recruited as interviewers, and 10 people were recruited as subjects. The total number of dialogs is 10, and the total time is about 150 minutes. The total number of utterances is 1069 and total number of sentence-like units (SUs) is 1671, where a SU is a unit that divides utterances by punctuation marks and a stops. We have named this corpus the "Japanese Deception Corpus (JDC)", and make it available for research purposes.[1] We show part of the JDC corpus in Table 1.

[1] http://ahclab.naist.jp/resource/ja-deception/

Table 1 Example dialog (I/：Interviewer, P/：Subject)

Speaker	Transcription	Label
I	音楽に関して, あなたはマッチしていましたか？ How did you do on the music section?	
P	はい, マッチしていました. I matched the desired profile on that section.	Lie
I	それはなぜだと思いますか？ Why do you think so?	
P	えーと, そこそこ答えれたからです. Uh, I was able to answer so-so.	Truth
P	小さい頃からずっとピアノをやっていたので. I have played piano since I was a child.	Lie

2.2 *Annotation*

In order to label the veracity of subjects' SUs, we asked all subjects to push a "truth" or "lie" button during the interview for each SU. SUs including a lie in any parts are defined as a lie. Labels for lies were obtained automatically from button-push data and hand-corrected for alignment. The number of SUs labeled "truth" was 1401 and the number labeled "lie" was 270.

3 Features for Deception Detection

In order to perform deception detection experiments, it is necessary to define features that may be indicative of deception. Based on previous research [4], we extract lexical and acoustic/prosodic features which may characterize deceptive speech. The extracted features are reported in Table 2.

- **Acoustic/prosodic features**
  As acoustic/prosodic features, we use fundamental frequency F_0, power, and phoneme duration. F_0 is obtained using the Snack Sound Toolkit [11], and phoneme duration is obtained using Kaldi [10].

- **Lexical features**
  To extract lexical features, we first perform word segmentation and POS tagging of the Japanese sentences using MeCab [6] and then use this information to calculate features. Of the listed features, topic indicates the test area under consideration, and noise indicates the presence of a cough or a sound resulting from the subject contacting the microphone. The frequency of positive-emotion words is extracted using Semantic Orientations of Words [12]. In addition to the previously proposed features for English, we add "the Japanese particles at the end of the sentence" which takes advantage of the fact that sentence final particles indicate when the speaker has confidence in their utterance (the "yo" particle),

Table 2 Acoustic/prosodic, lexical, and subject-dependent features

Category	Description
Lexical	Topic, Laugh, Noise, Disfluency, Third person pronoun, Denial, Yes/No, End of sentence, Verb base form, Cue phrase, Question, Positive words, Agree, Filled pause
F_0	Median, Percentage of median in SU
Phoneme duration	Vowel, Average, Max
Power	Average, First and Last frame of SU
Subject-dependent	Gender, Frequency of filled pause and cue phrase

is attempting to seek the agreement of the listener (the “ne” particle), or other similar factors.

- **Subject-dependent features**
  We also extract features related to the characteristics of the subject. We use the gender, the frequency of cue phrases (e.g. well, actually, basically), and the frequency of filled pauses.

4 Deception Detection Experiments

Based on the data and features described in the previous sections, we first perform experiments on binary classification between deceptive and non-deceptive utterances. To solve this classification, we use Bagging of decision trees [1] as implemented in the Weka toolkit, which gave the best performance of the methods that we tested. The evaluation of the experiments is performed by leave-one-out cross-validation which uses 1670 SUs for training and 1 SU for testing.

4.1 Discussion

Table 3 shows the classification results. “Japanese” is the classification rate using the JDC corpus that we described in section 2, and “English” is the accuracy the same classifier (minus the Japanese-specific features) using a part of the CSC corpus. “Human” indicates the accuracy of manual classification, where utterances are classified by a different person from the subjects. He classified each SU without considering the context.

In Japanese, the accuracy of the classification using acoustic/prosodic and subject-dependent features is the highest, higher than the chance rate by about 7%. Similarly in English, the accuracy using acoustic/prosodic and subject-dependent features is also highest, higher than the chance rate by about 17%. The accuracy of utterances

Table 3 Classification accuracy and deception detection F-measure for acoustic/prosodic (AP), lexical (L) and subject-dependent (S) features

	Japanese		English	
Features	Rate (%)	F-measure (%)	Rate (%)	F-measure (%)
Chance rate	83.8	0.0	71.4	0.0
AP	90.5	60.2	86.8	74.5
L	84.2	7.6	71.4	14.7
AP+S	90.7	61.4	88.1	77.7
L+S	85.2	31.5	76.8	52.9
AP+L	89.9	56.9	86.8	74.6
AP+L+S	90.2	58.1	87.8	77.2
Human	83.0	28.4		

Table 4 Accuracy between subjects (AP+L+S)

Subject	A	B	C	D	E	F	G	H	I	J
Chance rate (%)	93.5	89.4	92.4	76.9	78.7	75.9	64.9	82.9	73.1	83.9
Accuracy (%)	93.2	89.4	93.2	80.2	86.5	88.6	84.7	87.4	73.1	93.7

Table 5 Example dialogs（G/：Higher accuracy，A/：Lower accuracy）

Subject	Transcription
G	SN はー，そうですね，まぁMP 七八割は答えれたかなっていう位ですね. SN Ah, so uh MP, may be I was able to answer for 70% of the test.
A	たぶん大丈夫だと思います. I think that it's probably OK.

classified by humans is mostly the same as the chance rate, demonstrating the difficulty of deception detection for humans. As well, the rate using lexical features alone in Japanese and English is almost equal to the chance rate. Because the accuracy of classification improved after adding the frequency of cue phrases and filled pauses, we can see that subject-dependent features are effective to detect deception in both English and Japanese. Finally, we measured statistical significance between the results using Fisher's exact test, and found significant differences between the chance rate and systems using acoustic/prosodic features, acoustic/prosodic + subject-dependent, acoustic/prosodic + lexical, and acoustic/prosodic + lexical + subject-dependent ($p < 0.01$).

Table 4 shows the deception detection accuracy between subjects. Additionally, Table 5 shows the example dialogs between subjects in which deception detection is easy and difficult. In this figure SN indicates a noise and MP indicates a disfluency. The examples are subjects' replies to the interviewers' questions about the result of the test, and subject G is a speaker with a high deception detection accuracy and A is a speaker who has a low deception detection accuracy. It can be seen that A has many SN and MP, with an unsteady voice. On the other hand, G doesn't have many distinguishing differences from true utterances.

Table 6 Effective Features

Category	English	Japanese
Lexical	Noise, Third person pronoun, YesNo	Verb base
Subject-dependent	Frequency of cue phrase	
F_0	Median	Median
Phoneme duration	Average, Vowel	Vowel
Power	Average, First and Last frame of SU	Last frame of SU

4.2 Cross-lingual Comparison of Effective Features

In this section we compare the difference in features that are effective in deception detection across the two languages. Using best-first search, we did feature selection maximizing the rate of classification on the training data. Table 6 shows the resulting selected features. As acoustic/prosodic features, the median of F_0, average of vowel phoneme duration, and the last frame of power were found effective for both Japanese and English. Potential reasons why these features were effective for both Japanese and English are as follows:

- Last frame of power
  Generally, the change of feelings (such as uncertainty) tend to appear at the end of an utterance.
- Median of F_0
  It is said that people often change voice tone when they tell a lie [2].
- Vowel duration
  It is possible that people tend to speak at different speeds when lying or telling the truth.

Regarding lexical features, the results were greatly different between Japanese and English. In English, noise, third person pronoun, and containing “Yes” or “No” were effective. On the other hand, for Japanese the lexical features used in this research were largely ineffective, with only containing a verb base form proving effective.

5 Analysis of Types of Questions that Detect Deception Effectively

As our final goal is to build a dialogue system that can perform or aid deception detection, in this section we summarize our main results, analyzing what kind of questions this system can perform to make it easier to detect deception. Assume that we have question of the interviewer q and its corresponding response r. In order to perform this analysis, we separate all responses r into classes based on some feature of the corresponding q, then measure the accuracy of deception detection for each class. If a particular class has higher deception detection accuracy, it can be said that q of this type are effective at drawing out features of deception that can be easily detected automatically, and are thus questions that a deception detecting dialogue system should be focusing on.

5.1 Analysis of Question Dialogue Acts

First, we hypothesize that the variety (dialogue act) of the interviewers' utterance has an effect on the ease of detecting deception. In order to test this hypothesis, we use the dialogue act of q as the class into which we divide the responses r. Each utterance of the interviewer is annotated with a general-purpose function (GPF) defined by the ISO international standard for dialog act annotation (ISO24617-2, 2010). In this paper, annotators assign GPF manually. For approximately 10% of the corpus, 2 annotators annotate GPFs and the mean rate of agreement is 80%. Of these, we focus on situations where the annotator performs one of the following dialogue acts. The definitions of each dialogue act are quoted from the standard:

- **CheckQ**
  Communicative function of a dialogue act performed by the sender, S, in order to know whether a given proposition is true, about which S holds an uncertain belief that it is true. S assumes that addressee A knows whether the proposition is true or not, and puts pressure on A to provide this information
- **ChoiceQ**
  Communicative function of a dialogue act performed by the sender, S, in order to know which one from a given list of alternative propositions is true; S believes that exactly one element of that list is true; S assumes that the addressee, A, knows which of the alternative propositions is true, and S puts pressure on A to provide this information.
- **ProQ**
  Communicative function of a dialogue act performed by the sender, S, in order to know whether a given proposition is true. S assumes that A knows whether the proposition is true or not, and puts pressure on A to provide this information.[2]
- **SetQ**
  Communicative function of a dialogue act performed by the sender, S, in order to to know which elements of a certain set have a named property. S puts pressure on the addressee, A, to provide this information. S believes that at least one element of the set has the named property, and S assumes that A knows which are the elements of the set that have the property.

In Table 7, we show classification results for the subjects' SUs corresponding to each type of labeled GPF in the interviewers' utterance. In this case, we are most interested in the case where lies are correctly classified as lies (lie recall), as these indicate the possibility that the system can detect when the conversational partner

[2] This is a superset of checkQ, so ProQ in this work indicates all ProQ that are not CheckQ.

is lying. In Fig. 1, we show lie recall. Confidence intervals for $p<0.05$ is calculated by the Clopper-Pearson method.

From these results, we can see that the category with the highest rate of lies that are correctly classified as lies is for SUs corresponding to CheckQ. In responses to CheckQ questions, subjects tend to talk about the previous speech again when the interviewer asks them to confirm previous information. This is interesting in that Meyer [7] reported that interviewers often let liars talk about same speech to detect deception. The result that CheckQ is most effective to detect deception is in concert with this observation. On the other hand, the lowest rate of lie classified as lies is SUs corresponding to ProQ, which conceivably put less pressure on the interviewee, as they only need to answer yes or no.

In addition, in Table 8 we show the number of words per SU in the interviewee response to each question. From this table, we can see that the length of utterances corresponding to CheckQ is the shortest. Again, this is in concert with the observation of Meyer [7], that lies are more easily exposed from the extremely short utterances.

5.2 Analysis of Question Length

One potential reason why CheckQ was found to be effective was because it allows us to ask about details of the story, which has the potential to shake the confidence of the potential lier and cause deception features to be exposed [7]. Particularly, for a question that causes the speaker to review previously stated information such as CheckQ, subjects occasionally cannot answer about small points of the made-up story. In this case, it is also conceivable that the question length is important, because subjects think about a made-up story while listening to the question. Thus, we hypothesized that deception features are exposed more easily when interviewer asks shorter questions, and that the subjects can lie more skillfully when interviewer asks longer question.

To assess this hypothesis, we calculated the detection rate corresponding to each question length. In Fig. 2, we show lie recall corresponding to each question length. Like before, confidence intervals $p<0.05$ are calculated by the Clopper-Pearson method. For example, $1 \sim 10$ is the lie recall that classified for subjects' SUs corresponding to $1 \sim 10$ words question.

From these results, we can see that lie recall corresponding to questions of $1 \sim 10$ words is the highest. Because of this, we can say that, as expected, the short questions are effective to detect deception. Based on these results, and combined from the results in the previous section, it is likely that an interviewer, and by proxy, a dialogue system, that asks short questions that confirm previous information will be more effective at detecting deception.

Table 7 Detail of the classification results

	Recall(%)		Rate(%)	
Class	True	Lie	Accuracy	Chance rate
CheckQ	99.4	83.3	95.3	77.7
ChoiceQ	100.0	80.0	96.4	78.6
ProQ	100.0	69.1	91.5	66.7
SetQ	99.5	73.7	94.0	75.8

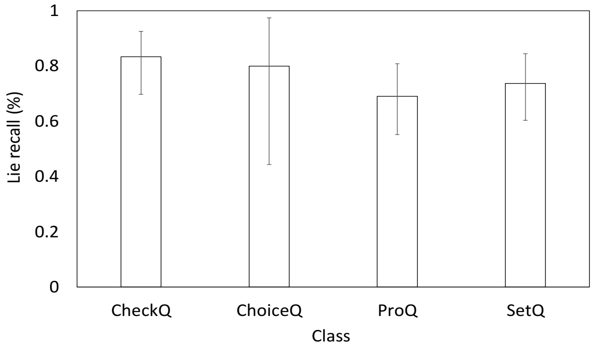

Fig. 1 Lie recall corresponding to each question

Table 8 Mean length of SUs corresponding to each question

CheckQ	ChoiceQ	ProQ	SetQ	Average
6.4	12.2	11.8	18.1	12.3

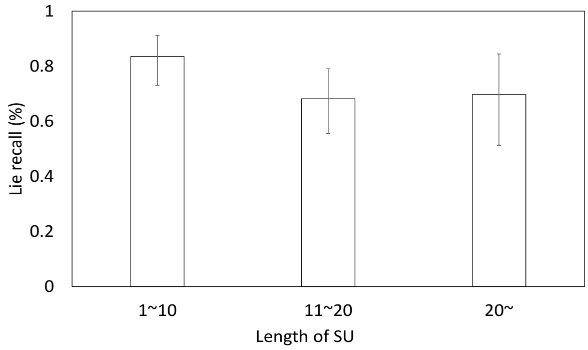

Fig. 2 Lie recall corresponding to each question length

6 Conclusion and Future Work

In this paper, we described the collection of a Japanese deception corpus and experiments in detecting deception. We performed classification using features that were shown be effective in English by previous research [4], and confirmed that these features were also effective to some extent for Japanese. We then performed on analysis of the relationship between types of questions an interviewer makes and the ease of detecting deception. We confirmed that Check questions were the most effective variety to elicit utterances that make it easier to detect deception, and that features of deception are exposed easily by asking short questions.

In future research, we plan to further analyze other aspects of questions that may influence the accuracy of deception detection. We will also perform the actual implementation of a deception detecting dialogue system based on our analysis of these effective questions.

References

1. L. Breiman. Bagging predictors. *Machine learning*, Vol. 24, No. 2, pp. 123–140, 1996.
2. B. M. DePaulo, J. J. Lindsay, B. E. Malone, L. Muhlenbruck, K. Charlton, and H. Cooper. Cues to deception. *Psychological bulletin*, Vol. 129, No. 1, p. 74, 2003.
3. P. Ekman. *TELLING LIES*. W. W. Norton & Company, 1985.
4. J. B. Hirschberg, S. Benus, J. M. Brenier, F. Enos, S. Friedman, S. Gilman, C. Girand, M. Graciarena, A. Kathol, L. Michaelis, et al. Distinguishing deceptive from non-deceptive speech. *Proc. Eurospeech*, 2005.
5. H. Hung and G. Chittaranjan. The IDIAP wolf corpus: exploring group behaviour in a competitive role-playing game. In *Proc. The international conference on Multimedia*, pp. 879–882. ACM, 2010.
6. T. Kudo, K. Yamamoto, and Y. Matsumoto. Applying conditional random fields to japanese morphological analysis. *Proc. EMNLP*, Vol. 4, pp. 230–237, 2004.
7. P. Meyer. *LIE SPOTTING*. Griffin, 2011.
8. Y. Ohmoto, K. Ueda, and T. Ohno. A method to detect lies in free communication using diverse nonverbal information: Towards an attentive agent. In *Active Media Technology*, pp. 42–53. Springer, 2009.
9. V. Pérez-Rosas and R. Mihalcea. Cross-cultural deception detection. *Proc. ACL*, pp. 440–445, 2014.
10. D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely. The kaldi speech recognition toolkit. *Proc. ASRU*, 2011.
11. K. Sjolander. Tcl/tk snack toolkit. 2004. http://www.speech.kth.se/snack/.
12. H. Takamura, T. Inui, and M. Okumura. Extracting semantic orientations of words using spin model. *Proc. ACL*, pp. 133–140, 2005.
13. A. Vrij, P. A. Granhag, S. Mann, and S. Leal. Outsmarting the liars: Toward a cognitive lie detection approach. *Current Directions in Psychological Science*, Vol. 20, No. 1, pp. 28–32, 2011.