人工知能学会研究会資料
SIG-SLUD-B502-04

嘘を発見する対話システム

A Dialog System to Detect Deception

角森 唯子[1]　Graham Neubig　Sakriani Sakti　平岡 拓也
水上 雅博　戸田 智基[2]　中村 哲*
Yuiko Tsunomori[1]　Graham Neubig　Sakriani Sakti　Takuya Hiraoka
Masahiro Mizukami　Tomoki Toda[2]　Satoshi Nakamura

奈良先端科学技術大学院大学 情報科学研究科
Nara Institute of Science and Technology, Graduate School of Information Science

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 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 for an interview task that acts as an interviewer asking questions to attempt to catch a potentially deceptive interviewee. We propose several dialog strategies for this system, and measure the utterance-level deception detection accuracy of each, finding that a more intelligent dialog strategy results in slightly better deception detection accuracy.

1　はじめに

日常生活の中で、嘘をつく、もしくはつかれる場面が非常に多い。このため、対話相手の発言が真実であるか、偽りであるかを見分ける技術が人間にとって非常に重要である。しかし、このように嘘を見分けるために、高度な技術が必要であり、例えば検事などは高度な技術を身につけている [4]。この技術は大きく分けて (1) 嘘の特徴を見分けることと、(2) 嘘の特徴が露呈しやすくなるように質問を行うことに分類できる [13]。

近年では、機械学習により嘘を検出する技術に関する研究がなされており、ある程度の実績を収めている [7, 10]。例えば、Hirschberg ら [7] は面接対話において、音響的・言語的特徴を用いた嘘検出を行い、チャンスレート (60.2%) を有意に上回る分類精度 (66.4%) を実現している。これら研究は、上述の「嘘の特徴を見分ける」技術をシステム上に実現することに目的的としている。

一方、「嘘の特徴が露呈しやすくなるように質問を行う」行為をシステム上で実現することを目的とした先行研究はほとんど存在しない。著者らが知る限り、我々が以前行った、対話中の嘘が露呈しやすくなる質問の分析 [12] のみであり、これに関して 2 節で詳しく説明する。この分析では、質問の様々な種類の中で、発話時間の短い質問、もしくは以前の発話から得られた情報を確認する質問は効果的に見分けやすい嘘を誘導するという結果が得られた。

本稿では、この分析に基づいて、**嘘を発見する対話システム**を提案する。このシステムの目標は、人間が嘘を発見しようとするときと同じように、嘘の音響的・言語的特徴が現れやすくなるような発話を行うことである。このようなシステムを構築するために、まず質問者が対象者に質問を行う面接対話のコーパスに対して、対話の流れをモデル化する (3 節)。そして、この分析を基に、嘘が露呈しやすくなるような対話戦略を提案する (4 節)。このモデルの対話戦略の有効性を検証するために、実際のユーザーが様々な対話戦略を用いたシステムと対話を行った実験を通して、対話戦略を工夫することにより対話中の嘘の検出率が向上することを確認した (5 節)。

2　対話シナリオと分析

嘘が頻繁に発生する場面の 1 つとして、面接が挙げられる。この場合、面接を受けている**対象者**は好意を持たれるために、嘘をついたり、脚色したりすることがある。そのため、面接を行っている**質問者**は対象者を正しく評価するために、対象者が嘘をついている場合、その嘘を暴くことが重要となる。

*連絡先： 奈良先端科学技術大学院大学 情報科学研究科
奈良県生駒市高山町 8916-5
E-mail: neubig@is.naist.jp
[1] 現在、NTT ドコモ勤務
[2] 現在、名古屋大学勤務

このような面接に着目して収集されたコーパスとして、英語では CSC コーパス [3]、日本語では我々が構築した日本語偽言コーパス (Japanese Deception Corpus; JDC [12]) が存在する。収録の手順を以下に示す。

1. 対象者が 6 項目 (政治、音楽、地理、食べ物、インタラクティブ、サバイバル) に対する、ある「目標プロフィール」に適合するかどうかを調べるための実験であると対象者に伝える。
2. 対象者に 6 項目のテストを収録前に受けてもらう。
3. 実験者は、2 項目が適合、4 項目が不適合となるように結果を操作し、その結果を対象者に伝える。
4. この実験の本当の目的は「目標プロフィールに適合している」と主張し、質問者を納得させることができる人物を探すことであり、上手く納得させれば賞金があると対象者に伝える。
5. 対象者は、全項目のテスト結果において「目標プロフィールに適合している」と、面接で質問者に主張する。質問者は、目標プロフィールやテスト内容についての知識はなく、いかなる質問をしても構わないとする。

このような手続きに従って収集されたデータに基づいて、対話中の嘘を検出する研究は多くなされており [7, 5]、音響的・言語的特徴を用いて、対話の参加者は嘘をついているかどうかはある程度判別可能となっている。また、我々は以前このデータに基づいて、「嘘の特徴をいかに見分けるか」だけではなく、「嘘が露呈しやくなるようにどのような質問をすれば良いか」を調査している [12]。この調査では、嘘検出の精度向上につながる質問に共通する特徴を 2 つ発見した。まず、今までの発話内容の信憑性を確かめる**確認質問**は特に嘘を検出する上で有効であった。これは、発言内容が確認されることにより、質問される対象者の不安が募らされるためであると考えられる。また、**発話時間の短い質問**も有効性が確認された。これは、発話時間が短ければ、質問対象の話者は十分な答えを練ることが出来ないためであると考えられる。

次節以降、分析から得られたこれらの知見を対話戦略に反映させる手法を提案する。

3 面接対話のモデル化

対話管理部の大枠を構築するために、まず対話行為を定義し、質問者と対象者の間の対話をデータに基づいてモデル化する。

表 1: 発話内容のカテゴリの例

カテゴリ	定義
項目	対象者が目標プロフィールに適合するかどうかについて 例:「音楽の部分はどうでしたか？」
テスト	テストの内容について 例:「どのような問題がありました？」
得点	テストの点数について 例:「どのような点数になったと思いますか？」

3.1 対話行為の設計

質問者の対話行為は、一般目的機能 (General purpose functions; GPF [2]) を基に定義される。GPF は一般的な対話行為を定義するための枠組みであるが、これを面接で発生するような発話に合わせて工夫する。具体的には、面接の性質から、質問者が対話の流れを決め、主に対象者への質問を行うことを考慮して、以下の 4 種類の質問に関する対話行為に基づいて質問者の対話行為を設計する：

CheckQ 聞き手によって与えられた命題の真偽について、話し手が確信が持てていない場合に使用。話し手が命題の真偽を確認するため、聞き手に対して情報提供を促す。

ChoiceQ 話し手が、聞き手が知っていると仮定する命題のリストの中から、真の要素がどれかを知るために、聞き手に対して情報提供を促す。

ProQ 話し手が、聞き手が知っていると仮定する命題の真偽を得るために、聞き手に対して情報提供を促す。

SetQ 話し手が、聞き手が知っていると仮定する集合の中から、どの要素が固有の特性を持つかを知るために、聞き手に対して情報提供を促す。

これらの対話行為はまだ対話制御を行う上で粒度が荒いため、更に各対話行為を細分化する。具体的には、4 種類の質問 GPF タグと、表 1 で示すような対話内容を表現する 11 種類のカテゴリに関するタグを付与し、計 44 種類の細分化された対話行為を作成する。JDC コーパスに含まれる全対話の各発話にこの対話行為タグを人手で付与する。

3.2 HMM に基づく対話の流れのモデル化

これらの対話行為の情報を用いて、面接対話を隠れマルコフモデル (Hidden Markov Model; HMM) でモデル化する。HMM では、各細分化された対話行為の系列を観測データとし、パラメータを EM アルゴリズムで学習する。HMM の状態数を 1 から 10 へと変動させ、各状態数においてランダムにパラメータを初期化してから学習を行う手続きを 100 回繰り返し、1,000 個の HMM が学習される。この 1,000 個の中から、最小記述長 (Minimum Description Length; MDL [11]) 基準に基づいて最良のパラメータを持つモデルを選択する。

$$MDL = -\ln L + \frac{k \ln n}{2}$$

ここで、L は学習データにおける HMM の尤度で n はそのデータの数であり、k は 0 でない遷移と生成確率の数である。

図 1 に、最良として選択された HMM モデルを示す。ここで、四角は状態を示し、四角の中の情報は観測され得る対話行為とその生成確率を示す。状態の間の矢印は遷移確率を示す。なお、図には 0.1 を超える確率のみを表示している。

4 嘘を発見する対話戦略

前節のデータに基づくモデルと、我々の以前の研究 [12] から得られた知見に基づき、嘘の特徴を引き出すような対話戦略を 2 つ設計する。学習データが比較的小規模であるため、前節のモデルをそのまま用いずに、これを基に対話戦略を人手で構築することにする。

4.1 決定的な対話戦略

まず、ユーザーが何を回答したとしても、同じ流れで対話を進めていく**決定的な対話戦略**を構築する。対話の流れは前節のデータに基づくモデルを参考にして、決定的に対話の流れを記述するルールを人手で作成する。例えば、システムがある項目に関する対話を開始する際、まずユーザーが目標プロフィールに適合するかどうかを聞いてから、なぜ適合したと思うのかに関する質問をする。対話が単調になり、ユーザーが飽きてしまうことを防ぐために、項目ごとに質問のパターンを前節のモデルの許容範囲内で変動させる。

4.2 嘘検出と確認質問に基づく対話戦略

我々が今までに行った調査 [12] により、人間の質問者が対象者の発言が怪しいと思った時に、その質問内容を確認する確認質問 (CheckQ) を行うという知見が得られている。また、様々な種類の中で、確認質問は嘘の特徴を引き出すのに最も効果的な種類であることも確認している。この 2 つの知見に基づいて、**嘘検出と確認質問に基づく対話戦略**も提案する。

具体的には、まず確認質問を他の質問種と区別して扱うために、対話行為系列からすべての確認質問を取り除き、確認質問なしの質問系列に基づき HMM モデルを 3.2 節の手続きに基づき再学習した。実際に対話を行う際に、この HMM に基づいて作成された対話を進めていきながら、ユーザーの各発話に対して、その発話が嘘である確率を計算する。そして、嘘の確率が 0.5 より高かった発話（つまり「怪しい」発話）に対してのみ、確認質問を行い、ユーザーの嘘を暴こうとする。確認質問を行ってから、通常のモデルに戻り、対話を中断した時点から再開始する。

5 実験的評価

5.1 実験設定

実験的評価では、対話戦略を工夫することにより、嘘の検出率の向上が図れるかを明らかにするために実験を設計する。人間が対話システムの機能の一部を補う WoZ 法 [6] を用いて、対話制御をシステムに任せ、言語生成と言語理解を人間の WoZ に任せる。言語生成では、各対話行為に対する発話テンプレートが用意され、このテンプレートに基づき WoZ が発話を生成する。以前の調査で得られた、短い質問が最も効果的であるという知見に基づき、テンプレートをすべて 1-10 単語の比較的短いものとする。また、不自然な合成音声へのエントレインメント [9] 等、システムと直接対話することによる副作用をなるべく減らすために、WoZ 役の人間が直接発話を読んで発話する形で対話を進める。

嘘の検出自体には、基本周波数、パワー、音素継続長などを含む音響・韻律特徴量 [12] を用いる。分類器として、試した様々な分類器の中で最も高い精度を実現した決定木のバギング [1] を用いる。

対話の項目は JDC の 6 つの項目の中で選択する。ユーザーが事前にテストを受け、テストの点数に関わらず、WoZ に点数が目標プロフィールに適合したと主張する。予備実験で、ユーザー自身の学習データがなければ十分な嘘検出精度が得られないことが分かったため、JDC の収録に参加したユーザー 8 名を実験の対象者とし、各ユーザーの JDC データを嘘検出器の学習データとする。各ユーザーは各システムと 1 回ずつ対話を行い、システム提示順による実験への影響を避けるため、システムの提示順をランダムにする。JDC と同じ手順に従い、ユーザーが対話を進めると同時に、

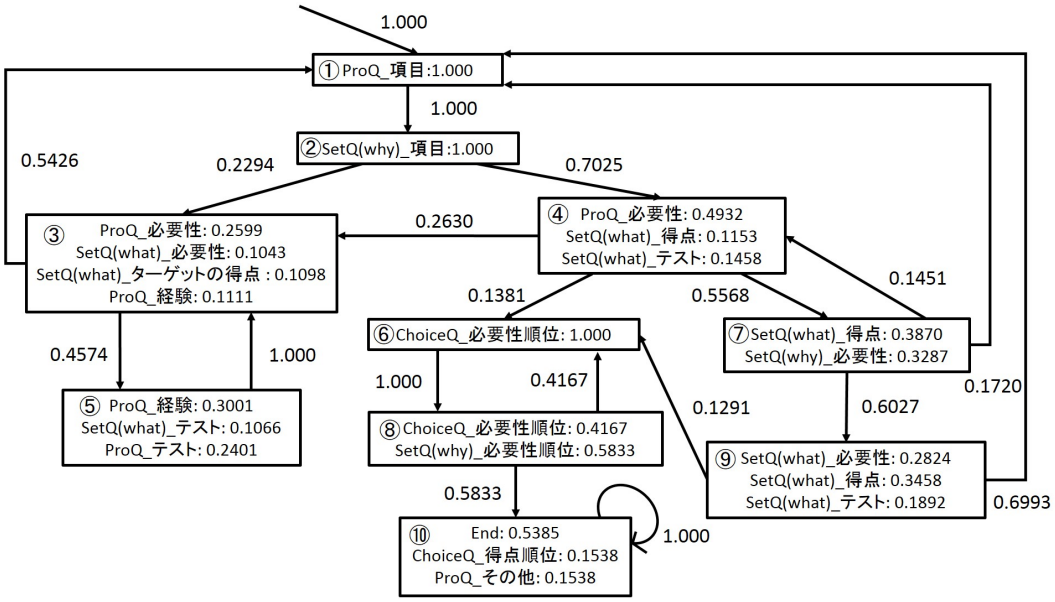

図 1: 面接対話の HMM モデル

各発話を行っている際に「真実」か「嘘」に対応するボタンを押してもらう。このラベルを正解として、嘘検出モデルを用いて検出の F 値を計算する。

5.2 評価用システム

評価には、下記の 4 種類の条件を調査する：

Random: 等確率でランダムに対話行為を選択するシステム。

Static: 4.1 節の通り、JDC コーパスの HMM モデルを参考に人手規則を記述した決定的なシステム。

CheckQ: 4.2 節の通り、JDC コーパスの HMM モデルを参考にした人手規則に加えて、嘘の確率の高い発話に対してのみ確認質問を行うシステム。

Human: 人間の WoZ が直接利用する対話行為を選択するシステム。選択の際、CheckQ システムが利用している嘘確率が WoZ に提示され、対話行為選択の判断材料にして良いとする。

5.3 実験結果と考察

図 2 に各システムにおける嘘検出の F 値を示し、表 2 に各システムの対話例を示す。エラーバーはブートストラップ法 [8] により得られた 95%の信頼区間を示す。

この結果から、各システムにおける総発話数が少ない関係で信頼区間が広いが、CheckQ システムは他の自動システムを若干上回り、Human と同程度の嘘検出率を実現していることが分かる。一方、Random と Static

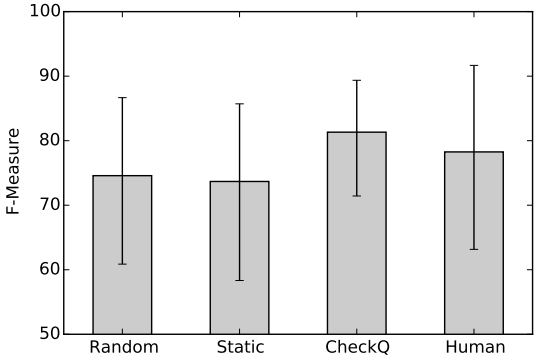

図 2: 各システムの嘘検出 F 値

の間に差がほとんどないことから、JDC に基づくモデル化より、嘘の検出とその結果に基づく確認質問の実施が嘘検出率の向上につながっている可能性が高いことが分かる。

なお、対話例に目を向けると、数量評価の差の理由が直感的に分かる。CheckQ と Human ではいずれも確認質問を混ぜながら、自然に対話を進めている。これと比較して、Static は自然に対話を進めているが、適宜に確認質問を行っておらず、嘘を発見するチャンスを逃していることが分かる。

6 おわりに

本稿では、嘘を発見する対話システムを構築するために、人間同士の面接対話のモデル化と、このモデル

表 2: 各システムにおける対話例（SP:話者、W:WoZ、U:ユーザー）

システム	SP	発話
Random	W	点数はどれぐらいだったでしょう？
	U	80 点ぐらいだと思います。
	W	政治の項目についてはどうですか？
	U	プロフィールに適合しました。
Static	W	音楽の項目はどうでしたか？
	U	プロフィールに適合しました。
	W	なぜそう思ったのですか？
	U	子どもの頃からピアノを弾いているからです。
CheckQ	W	政治の項目はどうでしたか？
	U	プロフィールに適合しました。
	W	本当ですか？
	U	もちろん。
Human	W	政治の項目はどうでしたか？
	U	点数がプロフィールに適合しました。
	W	本当ですか？
	U	はい。

や確認質問に関する知見に基づく対話戦略を提案した。実験では、嘘を検出し、嘘の確率が高い発話に対して確認質問を行う対話戦略が効果的であることが示唆された。

今後の課題として、本稿で行ったユーザー 8 名の比較的小規模な実験を、より大規模に拡張した場合の傾向を確認することが挙げられる。この傾向を確認してから、全自動で嘘を発見する対話システムを構築し、人間の WoZ に頼らない実験で効果を確かめていきたい。

参考文献

[1] L. Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

[2] H. Bunt, J. Alexandersson, J.-W. Choe, A. C. Fang, K. Hasida, V. Petukhova, A. Popescu-Belis, and D. R. Traum. Iso 24617-2: A semantically-based standard for dialogue annotation. In *Proceedings of LREC*, pages 430–437. Citeseer, 2012.

[3] B. M. DePaulo, J. J. Lindsay, B. E. Malone, L. Muhlenbruck, K. Charlton, and H. Cooper. Cues to deception. *Psychological bulletin*, 129(1):74, 2003.

[4] P. Ekman. *TELLING LIES*. W. W. Norton & Company, 1985.

[5] F. Enos. *Detecting deception in speech*. PhD thesis, Columbia University, 2009.

[6] N. M. Fraser and G. N. Gilbert. Simulating speech systems. *Computer Speech & Language*, 5(1):81–99, 1991.

[7] 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.

[8] P. Koehn. Statistical significance tests for machine translation evaluation. In *Proceedings of EMNLP*, 2004.

[9] R. Levitan and J. B. Hirschberg. Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. *Proc. Interspeech*, 2011.

[10] V. Pérez-Rosas and R. Mihalcea. Cross-cultural deception detection. *Proc. ACL*, pages 440–445, 2014.

[11] A. Stolcke and S. Omohundro. Hidden Markov model induction by bayesian model merging. *Advances in neural information processing systems*, 1993.

[12] Y. Tsunomori, G. Neubig, S. Sakti, T. Toda, and S. Nakamura. An analysis towards dialogue-based deception detection. In *Proceedings of IWSDS*, Busan, Korea, January 2015.

[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*, 20(1):28–32, 2011.