Vol.2017-SLP-117 No.6
2017/7/28

fastTextを用いた対話からの欺瞞検知と分析

細見 直希[1,a)]　Sakriani Sakti[1]　吉野 幸一郎[1,2]　中村 哲[1]

概要：欺瞞は人間同士のコミュニケーションにおいてしばしば起こるものであるが，それを検知することは人間にとって容易ではない．これに対し，機械ならば人間では捉えることが難しい特徴を利用することが可能であり，人間よりも高精度に欺瞞を検知することができる可能性が示唆されている．特に教師あり学習による機械学習では，事例ごとの正負（欺瞞・非欺瞞）を識別器に与えることで，どのような特徴が欺瞞検知に有効であるかを推定することができる．この機械学習の手法及び特徴空間の構築については近年盛んに研究されており，特に言語特徴においてはfastTextと呼ばれる分散表現を利用する識別器が文章からの評判分析や文に対するタグ予測の課題で高い精度を挙げることが報告されている．本研究では，このfastTextを用いて各発話に欺瞞ラベルが付与されたインタビュー形式の対話を学習させることで，人間よりも有意に高精度な欺瞞検知器が構築可能であることを示した．またそれに加え，音響特徴を用いる識別器との性能比較や人間による欺瞞検知能力についての検証分析を行った．

1. はじめに

欺瞞は人間同士の対話においてしばしば行われるものであり，時として対話相手の利益を損ねたり，発話者が不当な利益を享受するために行われる．欺瞞の検知を行うことができれば，こうした悪意のある情報のやりとりを防ぐことが可能となる．

しかし，人間にとって欺瞞の検知は困難な課題であることが過去の研究から言われている．Bondらは過去の200以上の様々な欺瞞検知に関する研究のメタ分析を行い，特別な訓練を受けていない人間による欺瞞検知の平均正解率が54%程度だったことを報告している[1]．また，Levineらは真実と欺瞞のメッセージを収録したビデオをそれぞれセグメントに分け，テストとして呈示するメッセージに含まれる真実の割合を変化させる実験を行うことで，人間による欺瞞検知の正解率が呈示される真実の割合に依存することを示している[2]．言い換えれば，人間が欺瞞検知を行おうとしても，一般にその精度がチャンスレベル（まぐれ当たりの確率）を大きく上回らないことが示されている．人間による欺瞞検知がうまくいかない要因については様々な議論がなされているが，その原因の1つに人間がバイアスを有していることが挙げられる．例えば，人間は相手の発言を実際の真偽に関わらず真実だと判断する傾向（真実バイアス）[3]や，相手を嘘つきだと思って見ているとそのように見えてしまう確証バイアス[4]を有していることが知られている[5]．また，欺瞞時に特徴的な現象を人間が把握できないことも原因の1つだと考えられる．そこで本研究では，人間が持つようなバイアスを持たない計算機を利用し，機械学習（教師あり学習）を用いて欺瞞検知を行うことを考える．教師あり学習では，欺瞞発話の事例を学習し，検知精度を最大化するような特徴量の選択を行う．そのため，人間が欺瞞検知に有効であると把握できない特徴を用いて検知を行うことができる可能性がある．今回は言語特徴を利用するfastTextや音響特徴を利用する識別器を用いて欺瞞検知器を構築し，人間よりも高精度な検知性能を目指す．また，検知に有効な特徴についての議論を行う．

2. 欺瞞と欺瞞検知

欺瞞については様々な研究者が定義をしている．例えばKraussは「欺瞞者が虚偽であるとみなす信念や理解を他者に抱かせようとする行為」[6]，Vrijは「メッセージの伝達者が真実でないとみなす信念を他者に形成しようとする，事前の通告の無い成功する可能性も失敗する可能性もある意図的試み」と定義した[7]．本研究でもこれらの定義にならう．ここで重要な点は欺瞞が意図的な行為であるということであり，虚偽記憶や無知，誤りなどについてはここでは欺瞞として扱わないということである．また，「欺瞞」の関連語である「嘘」は厳密には意味が異なるが，互換性のある語として用いているVrij[7]にならって，以降は文脈に応じてそれらを適宜用いる．

[1] 奈良先端科学技術大学院大学情報科学研究科
8916-5, Takayama-cho, Ikoma, Nara, 630-0192, Japan
[2] 科学技術振興機構
4-1-8, Honmachi, Kawaguchi, Saitama, 332-0012, Japan
[a)] hosomi.naoki.hg6@is.naist.jp

欺瞞検知の方法については，ポリグラフや fMRI などを用いて人間の生理的反応を計測することで高い検知性能が得られるという報告がある [7]．しかし，検知の対象者に専用の計測機器を繋いだり，特別な質問手順により反応を引き出したりする方法では，自由なコミュニケーション中の欺瞞を検知することが難しい．そこで本研究では，そのような特別な手続きを必要としない，言語や音声の特徴を用いた欺瞞検知について議論する．人間同士の対話からの機械学習を用いた欺瞞検知については，いくつか研究がされている．例えば，Hirschberg らは標準アメリカ英語を母国語とする英語話者による欺瞞発話を含むインタビューを収録したコーパスを作成し，それに対して Ripper rule-induction classifier を用いて語彙，音響，韻律及び話者依存の特徴量による欺瞞識別器を構築することで正解率がチャンスレベルに対して約 6%改善したことを報告している [8]．また，近年では Levitan らが標準英語及び北京語を母国語とする英語話者を対象としたコーパスを作成し，RandomForest を用いて音響，韻律特徴や話者の性格などの特徴として利用する識別器を構築し，チャンスレベルに対して約 10%高い精度を実現している [9]．

3. CSC Deceptive Speech

嘘を含むようなインタビューの場面が収録されたコーパスとして CSC Deceptive Speech がある [8]．これは標準アメリカ英語を話す男性 16 人，女性 16 人の合計 32 人の実験協力者を対象に各 25 分から 50 分間のインタビューを実施し，合計約 7 時間分の実験協力者の発話を収録したコーパスである．コーパスには録音された音声とそれを書き起こしたテキスト，回答の真偽についてのラベル情報が含まれている．コーパスに含まれる発話の例は表 1 の通りである．また，このコーパスの収録の条件については，以下の通りである [10]．

(1) 実験協力者のパフォーマンスがある目標プロフィールに適合するかどうか調査することを告げられる．
(2) 実験協力者は 6 項目（music, interactive, survival skills, food and wine knowledge, NYC geography, civics）について事前の調査を受ける．
(3) この 6 項目の回答について，2 項目が目標プロフィールに適合，4 項目が不適合となるよう実験者が調査結果を操作し，その結果を実験協力者に伝える．
(4) 実験協力者には本人のインタビュー結果（4/6 項目が不適合）が告げられ，次のインタビューでインタビュアーに対して彼ら自身が目標プロフィールに適合していることを納得させられれば，賞金がもらえることを告げられる．
(5) 次のインタビューで，実験協力者はインタビュアーに対して全項目の調査の結果において目標プロフィールに適合していることを主張する．実験協力者はインタビュアーからの各質問に対する自身の回答ついて，真実と欺瞞を示す 2 つのペダルのうちの 1 つを押すことによって，回答が真実であるか欺瞞であるかを示す．

表 1 ：CSC Deceptive Speech に含まれる発話例

発話	ラベル
well, yeah, there's a chance.	真実
uh actually, I did well. excellent.	欺瞞

コーパス中の発話に対する真偽のラベルは，実験協力者が押した真実，欺瞞のペダル情報に基づいて正解が与えられている．本研究では，句読点や休止によって自動的に分割されたものを発話単位として以降の実験に利用する．

4. 提案手法

本節ではまず，言語特徴を用いた教師あり学習による欺瞞検知器について説明する．また，それに加えて音響特徴を用いる識別器，それぞれの識別器を組み合わせて 1 つの識別器としたものをそれぞれ構築する．

4.1 言語特徴による識別器

言語特徴による識別器では Joulin らによって提案された文分類器である fastText のアーキテクチャを利用する．fastText は図 1 のように Word2Vec のアーキテクチャである CBOW モデル [11] とよく似た構造をもち，文の bag of n-grams を入力として埋め込み層を通して分散表現 e_i を構築し，それらのベクトルを平均化した m を識別器の入力として用いる．識別器はこれを入力としてラベルを予測する全結合ニューラルネットワークである．本研究ではこれを真実と欺瞞の 2 クラス分類に用い，識別器の出力である y と正解ラベルを用いて交差エントロピーを最小化するように学習を行う．fastText では文のラベル情報を利用して分散表現を構築し，同時に識別器の重みを学習するため，識別に最適化された分散表現を学習することが期待できる．実際に，文章からの評判分析や文に対するタグ予測の課題において高い性能を発揮することが報告されている [12]．

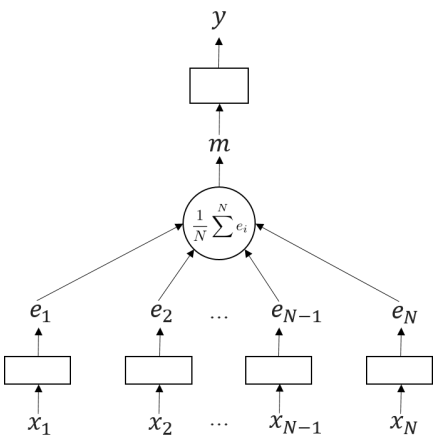

図 1 ：fastText のアーキテクチャ

4.2 音響特徴による識別器

音響特徴による識別器では，中間層 2 層の全結合ニューラルネットワークを用いた．入力としては，F0 や RMS Energy などを含む 384 次元の特徴量 [13] を用いた．音響特徴は発話音声から OpenSMILE[14] を利用して抽出した．入力及び中間層は特徴量の次元数と同数のユニット，出力層は 1 ユニットとし，活性化関数にはシグモイド関数を用いた．

4.3 2 つの識別器を組み合わせた識別器

言語特徴と音響特徴を用いた識別器は，それぞれ異なる種類の欺瞞の識別を学習していることが期待される．そこで 4.1 と 4.2 の識別器を組み合わせて新たな 1 つの識別器を構築する．提案する識別器は図 2 の通りである．y_1，y_2 はそれぞれの識別器の出力であり，w_1，w_2 を重みとして，その重み付き線形和を出力する．

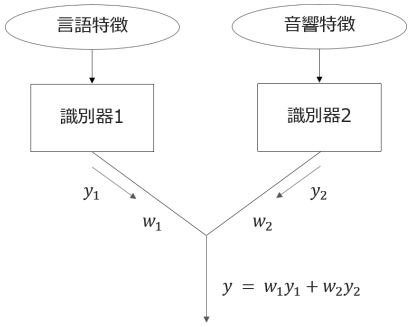

図 2 :提案手法の識別器

5. 実験

先行研究 [8] では CSC Deceptive Speech に含まれる発話について，その長短に関わらず識別器の学習，テストセットとして利用していた．しかし，コーパスには極めて短くそのラベルを予測することが困難な発話が含まれている．そこで本実験では句読点を含めて 5 単語より長い発話のうち，ランダムに抽出された 100 発話（真実 50 発話，欺瞞 50 発話）をテストセットとし，テストセットとは別に 4000 発話（真実 2000 発話，欺瞞 2000 発話）の学習セットを用意した．本実験ではこのデータを用いて，提案する識別器の欺瞞検知性能を検証する．加えて，同様の課題を人間を対象に実施することで，人間の欺瞞検知能力との比較を行う．人間による欺瞞検知実験では，一般財団法人国際ビジネスコミュニケーション協会が公開する PROFICIENCY SCALE を英語能力の参考として，TOEIC®スコア 730 以上（Mean=861.7，SD=117.7）の英語を母国語としない大学院生（日本人男性 3 名，外国人男性 3 名）によって実験を行った．実験協力者は欺瞞及び真実の発話の参考として学習セットを自由に利用できるものとし，テストセットに含まれる発話の書き起こし文の読解，音声の聴解，及び読解と聴解の 3 つの形式でテストセットのラベル予測を行った．

表 3 ：人間の検知結果

	Accuracy	Precison	Recall	F-measure
音声	0.515	0.524	0.370	0.414
テキスト	0.510	0.515	0.387	0.425
音声+テキスト	0.512	0.498	0.360	0.405

表 4 ：提案手法の検知結果

	Accuracy	Precison	Recall	F-measure
音響特徴	0.580	0.577	0.600	0.588
言語特徴	0.620	0.630	0.580	0.604
音響特徴+言語特徴	0.640	0.667	0.560	0.609

識別器による欺瞞検知実験では，4 章のそれぞれの識別器について学習セットのうち 90%（真実 1800 発話，嘘 1800 発話）で学習，10%（真実 200 発話，嘘 200 発話）でパラメータのチューニングを行い，テストセットのラベルの予測を行った．なお，本実験では対話の文脈情報は用いず，当該発話のみから判定を行った．また，4.1,4.2,4.3 のいずれの識別器についても実装には Theano（バージョン 0.9.0）バックエンドの Keras(バージョン 1.2.1) を利用した．4.1 の言語特徴による識別器では bag of bigrams を入力とし，30 次元の分散表現を構築した．4.2 の音響特徴による識別器では中間層の各層でユニットの 10%をランダムにドロップアウトして学習を行った．2 つの識別器を組み合わせた 4.3 の重み w_1，w_2 については，バリゼーションセットに対して Accuracy が最も高くなった $w_1 = 0.6$，$w_2 = 0.4$ とした．いずれのモデルの学習においてもバッチサイズ 10，エポック数 10 とし，最適化アルゴリズムとして Adam を用いた．パラメータは Adam の提案論文 [15] に準じた．

結果については，欺瞞である発話を正例，真実である発話を負例とし，表 2 及び以下の式 (1) から (4) を指標として用いて評価を行った．

表 2 ：予測結果の分類

		予測ラベル	
		正例	負例
真のラベル	正例	TP	FN
	負例	FP	TN

$$Accuracy = \frac{TP + TN}{FP + FN + TP + TN} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{FN + TP} \quad (3)$$

$$F - measure = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \quad (4)$$

6. 結果と考察

実験結果について，人間による欺瞞検知の結果を表 3，提案する識別器による検知結果を表 4 に示す．人間による

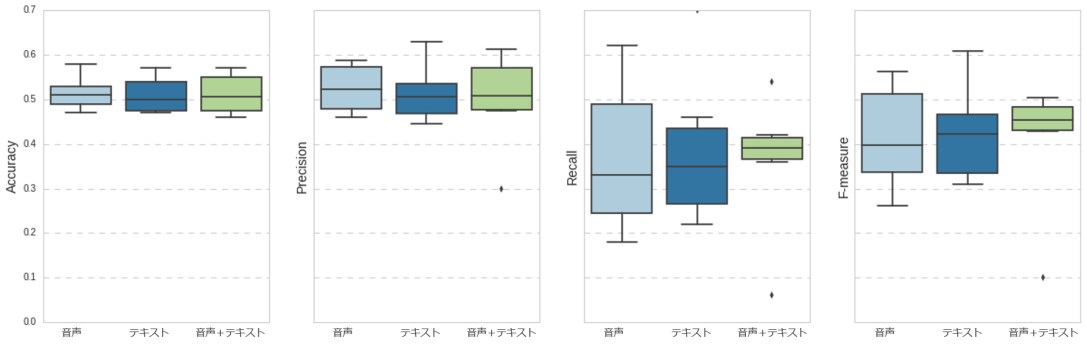

図 3 ：テスト形式ごとの人間による検知結果の分布

欺瞞検知の実験について，実験協力者間の発話に対する予測ラベルの一致度を調べるために Fleiss の Kappa 係数を算出した．一般に Kappa 係数は 0.6 以上だと予測の一致度が高いと見なされるが，本実験では 0.16 となり，実験協力者間で予測したラベルにばらつきが見られる結果となった．表 3 について，本実験のチャンスレベル 50%という課題に対して人間の平均 Accuracy は，いずれのテスト形式でも 51%程度となった．この Accuracy について片側二項検定を行ったところ，チャンスレベルと比較して有意差は見られなかった（$p > .05$）．さらに，図 3 からわかるように，欺瞞検知能力は実験協力者間でのばらつきが大きい．特に Recall のばらつきが大きい原因として，一方のラベルばかりを予測する実験協力者がいたことが考えられる．

また，実験終了後に実験協力者に対して行った欺瞞検知において音声，テキスト，音声とテキスト両方のどの形式が一番検知に役立ったかという質問に対しては，両方との回答が最も多かった．しかし，その他の形式のテスト結果と比較して，両方を利用した場合の指標に有意な差は見られなかった．また，発話が真実か欺瞞かを判断する際に何を基準としたか自由回答のアンケートを実施した結果，言語情報については”um”や”uh”などのフィラー，強調語の副詞，躊躇いを意味する語，音声情報については吃音，イントネーション，声の大小や高低を判断基準にしたという回答が得られた．しかし，本実験の結果からそれらの判断基準のほとんどが効果的ではないことがわかった．ただし，本実験では英語を母国語としない実験協力者による英語対話からの欺瞞検知だったため，実験協力者が正確に欺瞞を検知できなかった可能性が残されている．

表 4 に示す提案手法の教師あり学習による識別器を用いる欺瞞検知では，どの評価指標においても人間による欺瞞検知結果よりも高い結果となった．Accuracy については，言語特徴を利用した 4.1 の識別器と，言語特徴と音声特徴を利用した 4.3 の識別器を組み合わせた場合でチャンスレベルよりも 10%以上高い．この結果と人間による検知で最も Accuracy が高かった音声の聞き取りのみによる検知結果についてそれぞれ片側二項検定を行ったところ，提案手法では人間よりも有意に高精度な欺瞞検知が可能であることが確認できた（$p < .05$）．また，提案手法で得られた F-measure について，人間と比較して大幅に高くなっていることが確認できた．

7. おわりに

本研究では，欺瞞を含むインタビュー形式の対話を対象とした教師あり学習による欺瞞検知と，人間が検知を行った場合との性能比較を行った．実験結果から，言語特徴を利用する識別器，言語特徴を利用する識別器と音響特徴を利用する識別器を組み合わせた識別器では，チャンスレベル及び人間による検知能力と比較して有意に高い Accuracy を得られることが確認できた．また，2 つの識別器を組み合わせた場合では，言語特徴のみを利用する識別器よりも高い Accuracy が得られた．今後は個々の特徴量及び識別器の改良行い，より高精度な識別器の構築を目指す．また，実際に欺瞞と判定できたもので利用された特徴量についての分析，規模の大きなテストセットを用いた英語を母国語とする協力者による欺瞞検知実験を行う予定である．

参考文献

[1] Bond Jr, C. F. and DePaulo, B. M.: Accuracy of deception judgments, *Personality and social psychology Review*, Vol. 10, No. 3, pp. 214–234 (2006).
[2] Levine, T. R., Kim, R. K., Sun Park, H. and Hughes, M.: Deception detection accuracy is a predictable linear function of message veracity base-rate: A formal test of Park and Levine's probability model, *Communication Monographs*, Vol. 73, No. 3, pp. 243–260 (2006).
[3] McCornack, S. A. and Parks, M. R.: Deception detection and relationship development: The other side of trust, *Annals of the International Communication Association*, Vol. 9, No. 1, pp. 377–389 (1986).
[4] Nickerson, R. S.: Confirmation bias: A ubiquitous phenomenon in many guises., *Review of general psychology*, Vol. 2, No. 2, p. 175 (1998).
[5] 村井潤一郎：社会心理学における嘘研究: 現状と展望, 嘘に対する雑感 (特集うそ・ウソ・嘘)，心理学ワールド，No. 71, pp. 5–8 (2015).

[6] Krauss, R. M.: Impression formation, impression management, and nonverbal behaviors, *Social cognition: The Ontario Symposium*, Vol. 1, Erlbaum Hillsdale, NJ, pp. 323–341 (1981).
[7] Vrij, A.: *Detecting lies and deceit: Pitfalls and opportunities*, John Wiley & Sons (2008).
[8] Hirschberg, J., Benus, S., Brenier, J. M., Enos, F., Friedman, S., Gilman, S., Girand, C., Graciarena, M., Kathol, A., Michaelis, L. et al.: Distinguishing deceptive from non-deceptive speech., *Interspeech*, pp. 1833–1836 (2005).
[9] Levitan, S. I., An, G., Wang, M., Mendels, G., Hirschberg, J., Levine, M. and Rosenberg, A.: Cross-cultural production and detection of deception from speech, *Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection*, ACM, pp. 1–8 (2015).
[10] Enos, F.: *Detecting deception in speech*, Columbia University (2009).
[11] Mikolov, T., Chen, K., Corrado, G. and Dean, J.: Efficient estimation of word representations in vector space, *arXiv preprint arXiv:1301.3781* (2013).
[12] Joulin, A., Grave, E., Bojanowski, P. and Mikolov, T.: Bag of Tricks for Efficient Text Classification, *arXiv preprint arXiv:1607.01759* (2016).
[13] Schuller, B., Steidl, S. and Batliner, A.: The interspeech 2009 emotion challenge, *Tenth Annual Conference of the International Speech Communication Association* (2009).
[14] Eyben, F., Weninger, F., Gross, F. and Schuller, B.: Recent developments in opensmile, the munich open-source multimedia feature extractor, *21st ACM international conference on Multimedia*, ACM, pp. 835–838 (2013).
[15] Kingma, D. and Ba, J.: Adam: A method for stochastic optimization, *arXiv preprint arXiv:1412.6980* (2014).