クラウドソーシングでの固有表現アノテーションにおける不良回答の検出

福光 嘉伸[1,2,a)]　松田 裕貴[1,2,b)]　諏訪 博彦[1]　安本 慶一[1]

概要：データのアノテーション作業をクラウドソーシングで行うことで，低コストで機械学習のための学習データを収集できる．しかし，コストと引き換えに得られるデータの品質に大きなばらつきがあり，特に対価として報酬を付与すると可能な限り速く回答を行おうとする行動により不良回答が発生する問題がある．そこで，本研究では固有表現アノテーションを対象として不良回答をリアルタイムで検出することを目的とし，作業中の画面操作から得られるカーソル移動量や操作時間などの特徴量を用いた検出手法を提案する．本稿では，クラウドワーカーを対象に検証実験を行うとともに，分類精度の改善を目的に個人差を反映できる特徴量を追加する．機械学習モデルによる分類では，学内学生とクラウドワーカーのデータ両方を用いたデータセットにおいて 0.747 の Accuracy が得られ，ラベル付与数に関する特徴量が分類において重要であることが分かった．

キーワード：アノテーション，不良回答検出，機械学習

1. はじめに

マイクロタスク型クラウドソーシングは，インターネット上で不特定多数の人々に短時間で遂行可能な業務を水平分散的に委託することで，低コストで大規模なデータ収集・分析を実現する手法であり，様々な用途での活用が進んでいる．機械学習の分野においては，モデル構築・精度向上のために大量の学習データが必要となるため，データのアノテーション作業をクラウドソーシングで行うことで低コストで学習データを収集することが重要となる．しかし，アノテーションを行うユーザのスキルや意欲に差があるため，コストと引き換えに得られるデータの品質に大きなばらつきが生じるという課題がある [1]．特に，回答の対価として報酬を付与する場合に可能な限り速く回答を行おうとする行為が発生する（努力の最小限化）と考えられる．人間は思考に時間を使うことで様々な誤りを減らすため [2]，努力の最小限化によって思考時間が減少することで誤ったラベリングを行い，不良回答が発生するという問題がある．また，クラウドソーシングでは金銭を報酬とする場合に，アノテーションの質が低下するとの報告がある [3]．誤ったラベルが多量に含まれる学習データを用いた場合，機械学習モデルの精度が低下する可能性がある．そのため，そうしたノイズとなりうるデータの発生を検知・防止することが求められている．

社会心理学といった質問紙調査（アンケート調査）を多く取り扱ってきた分野では，より正確な回答結果を得るために努力の最小限化の傾向を検出する手法が考案されている．評価尺度を質問紙に取り入れて努力の最小限化の傾向を示す個人を検出する方法として，Instructional Manipulation Check（IMC）[4] や Directed Questions Scale（DQS）[5] がある．しかし，これらの回答者を疑うような質問を回答者自身が認識可能な形で提示する方法は，回答者に心理的負担を与え，内発的動機を損なう可能性がある．それにより，その質問自体が努力の最小限化の傾向を引き起こす可能性がある．そこで，後上ら [6], [7] らは，評価尺度を取り入れずに努力の最小限化の傾向を検出することを目的として，スマートフォンの画面操作記録を特徴量とする機械学習による検出手法を提案している．しかし，この手法ではアンケートの回答が全て終わってからしか最小限化の傾向の検出を行うことができないことが課題点として挙げられる．同一人物が一度しか実施することができないタスクや，データの母集団が限られている場合では，これまでの手法では不良回答の発生により最終的に得られるデータが不足してしまう恐れがある．この観点はアノテーションについ

[1] 奈良先端科学技術大学院大学
Nara Institute of Science and Technology
[2] 共同第一著者，Co-first author
[a)] fukumitsu.yoshinobu.ft5@is.naist.jp
[b)] yukimat@is.naist.jp

ても同様に問題となり，母集団が限定されるアノテーションタスクで対象者が不良回答を行った場合にはそのデータを利用できなくなったり，最終的に構築するモデルの精度に悪影響を及ぼしたりすることが懸念される．このことから，アノテーション作業中にリアルタイムに不良回答の検知を行いデータ品質を向上する重要性が高いと考える．

そこで，我々はデータのアノテーションタスクを対象として不良回答をリアルタイムに検知する手法を提案した [8]．提案手法では，アノテーション作業中のクリックやマウスカーソルの移動等の画面操作をリアルタイムに記録し，得られる特徴量を用いた不良回答検出を行う．先行研究 [9] では，不良回答の分類モデルを作成するために，固有表現ラベリングのタスクを対象として奈良先端科学技術大学院大学の学生を対象に学習データの取得実験を実施し，適切回答・不良回答の分類を行うモデルを構築した．本稿では，クラウドワーカーを対象にデータ取得及び検証実験を行うとともに，分類精度の改善を目的に個人差を反映できる特徴量である例題の差分特徴量を追加する．また，本実験では新たに取得したクラウドワーカーのデータに，先行研究の学内学生のデータを合わせたデータセットを構築することで，データ増加による分類精度改善を図る．データセットの構築後，各モデルの評価を行い，得られた結果についての分析結果について報告する．クラウドワーカーのデータのみを用いたデータセットでは，先行研究の学内学生のデータを用いたデータセットと同等もしくは少し高い精度を取得した．また，学内学生とクラウドワーカーのデータ両方を用いたデータセットにおいて 0.747 の Accuracy が得られた．Permutation importance によって各特徴量のモデルへの寄与度を算出した結果，ラベル付与数に関する特徴量が分類において重要であることが分かった．

2. 関連研究

従来の不良回答を検出する研究として，リッカート式やテキストベースの回答を対象としたものが一般的であった．後上ら [6], [7] は，一回のスクロール操作による画面移動量やスクロール速度などの画面操作を記録するシステムを作成し，画面操作を特徴量として用いることで，85.9%の検出率を達成しているが，回答完了後にしか検出できないためリアルタイム性がなく，回答者へのフィードバックや改善に時間がかかってしまう．また，不良なクラウドワーカーへの対応策を講じておらず，取得したデータの一部を不良データとして削除しなければならない可能性がある．

一方，アノテーションに関して客観的な品質を測る指標についても研究されている．Otani ら [10] は，物体検出精度の品質を測る指標として，一般的な mAP（Mean Average Precision）に加え，誤差を含む検出結果を正解に修正するためのコスト OC-cost（Optimal Correction Cost）を提案している．また，アノテーションやマイクロタスクに対するユーザの作業に対するモチベーションの向上に重点が置かれており，出力結果の品質向上には課題が残っている．Sihang ら [11] は，クラウドソーシングでマイクロタスクを作業として与える際に，従来の Web インターフェースの代わりに会話型インターフェースを用いることで，ユーザのやる気を向上させる手法を提案している．Jeffrey ら [12] は，長時間の作業中に適度な休憩を与えることで，ユーザの定着率を大幅に改善し，作業に対する関与を高めることを明らかにした．しかし，これらの手法はデータの品質向上に直接結びついているわけではない．

白砂ら [13] は，膀胱鏡画像が異常か正常か判断するタスクにおいて作業前に 1 秒待ち時間を設定することでアノテーション品質を向上できることを明らかにしたが，数%の改善にとどまっており，長い思考時間が必要となるタスクでは効果が小さいと考えられる．これらの課題を踏まえ，リアルタイムに不良回答の検出を行い，不良回答を改善することのできる方法を検討する重要性は高いと考えられる．

先行研究 [8], [9] では，固有表現アノテーションタスクを対象に，不良回答検出を行う手法を提案し，奈良先端科学技術大学院大学の学生を対象にデータ取得実験を行った．取得したデータを基にモデルを構築し分析をした結果，0.738 の Accuracy が得られ，分類失敗のデータに関してはタスク間の難易度差・作業速度や本人の能力などの個人差が生じていることが原因であると考察した．そのため，本稿では作業速度・作業時間に依存しない個人差を反映するような特徴量を追加し，クラウドワーカーを対象に検証実験を行う．

3. 提案手法

本稿では，対象とするマイクロタスクとして，自然言語アノテーションの一つである「固有表現（Named Entity）アノテーション」に着目し，タスク中に発生する不良回答をリアルタイムに検出することを目的としている．その実現のために，クリックやカーソル移動といった「画面操作ログ」の収集方法，および収集されたログデータに基づく不良回答検出モデルを提案する．

3.1 想定シナリオと提案手法の概要

本稿で想定するアノテーションタスク遂行の想定シナリオおよび，提案手法の概要を**図 1** に示すとともに，以下に各項目について詳細を述べる．

(1) アノテーションタスク依頼者がクラウドソーシングサービス等でアノテータを募集する．アノテータは，自然言語の専門家ではなく，アノテーション対象となる言語を母国語とする一般の市民（クラウドワーカー）を想定する．
(2) アノテータが依頼されたアノテーションタスクを遂行

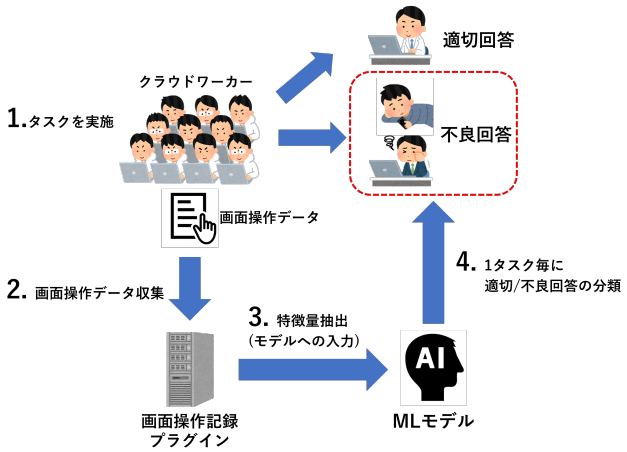

図 1　提案手法の概要

図 2　ラベル付与の様子

する．アノテーションタスクはそれぞれアノテータが保有する PC によって遂行されることを想定する．

(3) この際，アノテーションシステムに搭載されたプラグイン（後述）がアノテータの作業時の画面操作ログを定常的に収集する．

(4) 得られたログから特徴量を抽出し，抽出された特徴量を入力とする機械学習モデルにより，不良回答を検出する．

3.2　固有表現アノテーションの作業手順

本研究では，アノテーション作業として特に「固有表現（Named Entity）アノテーション」に焦点を当てている．アノテーション作業は，オープンソースソフトウェアのアノテーションツールである LabelStudio [14] 上で行うことを想定する．アノテータは与えられる文章内に含まれる文字列に対して固有表現のラベル付けをする．固有表現の例としては，「地名」・「施設名」・「乗り物名」・「路線名」などが挙げられ，それぞれについて「固有名詞」・「一般名詞」などの区別がある．LabelStudio 上でのアノテーション作業の詳細な手順を以下に示す．

(1) タスク一覧から順にタスクを選択すると，アノテーション対象の文章が与えられる．

(2) 文章上部に並んでいるラベル一覧から，付与したいラベルをクリックし選択する．

(3) アノテーション対象の文章内の任意の文字列をドラッグすることで範囲選択する．

表 1　抽出する特徴量

特徴量	単位	差分特徴量
回答時間	s	
非操作時間	s	
	回	
カーソル移動量（x 軸, y 軸, 合計）	px	
クリック回数	回	
カーソル n 秒以上停止回数 {n=1,3,5,10,30}	回	
クリック間隔 平均時間	s	○
クリック間隔 標準偏差	s	○
クリック間隔（最大，最小，第 1・3 四分位数）	s	○
クリック間隔の最大値，最小値の差分	s	○
クリック間隔の第 1・3 四分位数の差分	s	○
文字選択回数	回	
ラベル選択回数	回	
付与ラベル数（重複除く，重複含む）	個	
平均カーソル移動量（1 秒毎，1 ラベル付与毎）	px	○
平均クリック回数（1 秒毎，1 ラベル付与毎）	回	○
平均文字選択回数（1 秒毎，1 ラベル付与毎）	回	○
1 秒間に n ピクセル以上移動していない回数 {n=100,300,500,1000}	回	
n 秒毎のカーソル移動速度（最大，最小第 1・3 四分位数）{n=0.5,1,3,5,10}	px/s	
n 秒毎のカーソル移動速度の最大値，最小値の差分 {n=0.5,1,3,5,10}	px/s	
n 秒毎のカーソル移動速度の第 1・3 四分位数の差分 {n=0.5,1,3,5,10}	px/s	
n 秒毎のカーソル移動速度の標準偏差 {n=0.5,1,3,5,10}	px/s	

(4) 範囲選択が確定すると手順（2）で選択したラベルが範囲選択した文字に対して付与される．なお，誤ってラベルを付与した場合は**図 2** の左側に示すラベル欄から削除する必要がある．

(5) 文章中の全ての固有表現にラベル付けが完了したら，Submit ボタンを押し，タスク内容を保存する．

3.3　画面操作記録プラグインと抽出する特徴量

アノテーション作業を遂行する際の「画面操作」を記録する LabelStudio のプラグインについて述べる．LabelStudio はブラウザで利用可能な Web ベースのアノテーションツールであることから，アノテータがシステム上で行う操作を記録する機能をフロントエンドの Javascript プラグインとして実装した．プラグインは，ウィンドウ・タブのアクティブ表示時間などの情報に加えて，マウスイベントやクリックイベントなどのイベント駆動の情報を逐次収集しデータベースへと格納する．

次に，得られた画面操作ログデータに基づいて抽出される特徴量について述べる．本研究では，アノテーションタスクの操作内容に合わせた特徴量，かつリアルタイムなデータ抽出に対応可能な特徴量を抽出することとした．本稿で使用する基本となる特徴量（以降，基本特徴量）を**表 1** に示す．先行研究 [9] の結果として，作業速度・作業時間に依存しない個人差を反映するような特徴量を増やすことに

よって分類精度が改善できる可能性が示唆されたため，例題（ベースライン）と実際のアノテーションタスクへの回答の特徴量の差（差分特徴量）を追加した．この差分特徴量については，表 1 の「差分特徴量」の欄に “〇” で示す．

固有表現アノテーションは，文章中から固有名詞（人名や書籍名など）や日時表現など固有表現を抽出し，ラベルを付与するタスクであるため，マウスカーソルを移動させることによる文字選択（該当単語をドラッグ操作で囲う）や，選択した領域へのラベル付与（指定されたリストから対応するラベルを選択する）といった画面操作が含まれる．タスクの実施時には，文章を見てどの単語が固有表現であり，どのカテゴリに属するかを判断する認知処理が必要とされる．不良回答では，速く回答を行うために十分な数の単語にラベルを付与していないことや，文章中の固有表現に適切なラベルが付与されていないこと，回答に迷ってマウスカーソルの速度にばらつきがあることが考えられる．このことから，適切/不良回答間で回答時間，文字選択回数やカーソル移動量・速度等に違いが出ると考えられる．また，付与されたラベルには，ユーザのラベル削除忘れ等の理由で同じ文字列に対して複数付与されている場合がある．そのため，付与ラベル数では同一文字列に対して重複してラベル付けがされていた際に，加算する場合としない場合どちらも抽出する．

3.4 不良回答検知モデルの構築

最後に，前述の画面操作記録プラグインによって収集された画面操作ログデータによって導出される特徴量を入力とする不良回答検知モデルを構築する．機械学習のアルゴリズムとしては，後上らの報告 [7] で最も優れた性能を示した LightGBM（LGBM）[15] を用いることとする．LightGBM は実行時間が短くリアルタイムでの向いており，一般的に少ないサンプルデータでも高い精度が出やすい，かつ決定木ベースのアルゴリズムであるため不要な特徴量が含まれていても精度が低下しにくいという特徴がある．データセットのサンプルサイズが不均衡の場合，学習が上手くできない可能性があるため SMOTE [16] を用いてオーバーサンプリングを行う．また，ハイパーパラメータは Optuna を用いて最適化する．

4. 実験

4.1 データ収集実験

3 章で述べた画面操作記録プラグインを用いて実施した学習データ取得実験および取得データセットについて説明する．本実験は，奈良先端科学技術大学院大学の学生に加え，クラウドソーシングサービスであるクラウドワークス[*1]とランサーズ[*2]のワーカーを対象として募集を行い，

[*1] https://crowdworks.jp/
[*2] https://www.lancers.jp/

自宅を出てのぞみに乗り山を越え新大阪駅に到着しました．新大阪駅からJR 京都線とバスを乗り継ぎ，奈良駅に着きました．猿沢池までの道は歩きました．

凡例
地名（固有名詞）・地名（一般名詞）
施設名（固有名詞）・施設名（一般名詞）
乗り物名（固有名詞）・乗り物名（一般名詞）
路線名（固有名詞）・路線名（一般名詞）

図 3 地球の歩き方旅行記データセットを用いた固有表現アノテーションの実施例

被験者数は学内学生 61 人，クラウドワークス 76 人，ランサーズ 23 人の計 160 人であった．被験者の内訳は，10 代から 60 代の男性 79 名，女性 81 名である．学内学生には報酬として 1000 円相当のギフトカードを付与し，クラウドワーカーには契約金額として 1000 円を支払うことで，アノテーション作業を実施してもらいその際の画面操作を記録することで学習データを取得する．なお，本研究は人を対象とする研究に関する倫理審査委員会の承認を受けて実施した（承認番号：2020-I-2）．

以降では，実験で用いたアノテーションタスクおよび得られたデータセットについて述べる．

4.1.1 アノテーションタスクの概要

本実験におけるアノテーション作業としては，「地球の歩き方旅行記データセット[*3] [17], [18]」を用いた固有表現ラベリングを対象とした．この文章中の固有表現の内，地名・施設名・乗り物名・路線名や道・橋等の経路に対して固有名詞/一般名詞を区別してラベルを付与する．この 4 種類に該当しない表現や単語はラベル付与の対象外であり，各タスクには 2〜9 個の付与対象の固有表現が含まれる．アノテーションの手順は 3.2 節に示す内容に準ずる．具体的なアノテーションの実施例を**図 3** に示す．

本実験はデータセットのうち日本語文章の 24 文を選択した．全ての文へのアノテーションの終了，もしくは作業実施から 30 分を経過した時点でのタスクへのアノテーションの終了をもって，本実験の作業終了とした．

4.1.2 データセット

地球の歩き方旅行記データセットのグラウンドトゥルースと，実際に被験者がラベリングを行った各タスクの結果を比較し，適切回答であるか不良回答であるかを判別する．実験参加者 160 人の内，指示通りに作業を実施した，かつデータの使用許可が得られた 146 人のデータを分析対象とする．前処理として，タスクの回答時間が 10 秒未満，5 分以上のデータに関しては異常値として取り除く．また，ラベル付与対象が 2 つおよび 3 つの文章が 2 文のみだったた

[*3] https://www.nii.ac.jp/dsc/idr/arukikata/

表 2 データセットに含まれるデータ数（クラウドワーカー実験）

	データセット Crowd $D_{\pm0}$	データセット Crowd $D_{\pm50}$
適切回答	750	924
不良回答	876	702

表 3 データセットに含まれるデータ数（2 実験合計）

	データセット All $D_{\pm0}$	データセット All $D_{\pm50}$
適切回答	1191	1463
不良回答	1435	1163

め，これらを外れ値として扱い分析対象から取り除く．ラベル範囲・ラベル種類の両方がグランドトゥルースと一致したもののみ正解とし，そのときの F1 score（F1 値）を評価値とする．以下の式 1，2，3 より F1 値を算出する．

$$Precision = \frac{\text{文中の正解ラベル数}}{\text{文中の付与ラベル数}} \quad (1)$$

$$Recall = \frac{\text{文中の正解ラベル数}}{\text{グランドトゥルースのラベル数}} \quad (2)$$

$$F1score = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3)$$

不良回答とみなす F1 値の閾値について，付与対象が最も少ない 4 つの場合に，ラベル付与不足・付与過剰を 1 つずつまで許容するように設定する．そのため，F1 値が 0.7 以上のラベル済みデータを適切回答，0.7 未満を不良回答として扱いデータセット Crowd $D_{\pm0}$ を作成する．

自然言語領域において正解とするラベル範囲に関しては，文字数の ±N% までは正解とみなすとして部分一致を許容する場合もある．明確に許容範囲の値が決められているわけではなく，対象とするアノテーション作業にあわせて設定する必要がある．本アノテーション作業でラベル付与の範囲不足が発生する原因に，複合名詞に対して片方の名詞にのみラベルを付与する場合が挙げられる．そのため，文字数の ±50% までの部分一致を正解とみなして，F1 値を算出し，データセット Crowd $D_{\pm50}$ を作成する．また，データセット Crowd $D_{\pm0}$，Crowd $D_{\pm50}$ はクラウドワーカーを対象に取得したデータのみで構築されており，学内学生実験と同等の精度が出るかや差分特徴量追加による影響を調査する．作成した各データセットに含まれる適切回答・不良回答のデータ数を**表 2** にそれぞれに示す．

次に，データ増加による影響を調査するため学内学生実験のデータとクラウドワーカー実験のデータを組み合わせてデータセットを構築する．同一の F1 値の基準とラベル範囲の基準を用いて，データセット All $D_{\pm0}$ とデータセット All $D_{\pm50}$ を作成し，適切回答・不良回答のデータ数を**表 3** にそれぞれ示す．

4.2 モデルの評価方法

データセット Crowd $D_{\pm0}$，Crowd $D_{\pm50}$，All $D_{\pm0}$，All $D_{\pm50}$ それぞれを用いて機械学習モデルを構築し，モデル性能を評価する．この際，Crowd $D_{\pm0}$，Crowd $D_{\pm50}$ ではベースラインとの差分特徴量を含める場合と，含めない場合の精度を算出し，差分特徴量の影響を評価する．本稿では，汎化性能を考慮した上で分類精度を算出するために，ある一人の被験者のデータをテストデータとし，それ以外の被験者データで構築したモデルを評価する工程を，全ての被験者について繰り返し行う Leave-One-Participant-Out 交差検証（LOPOCV）を用いることとした．分類の評価指標として Precision，Recall，Accuracy を用い，モデルの評価を行う．また，同様の特徴量を数種類抽出している場合，特徴量重要度にバイアスが生じる場合がある [19]．そのため，クラウドワーカー実験のモデルに関して，Permutation importance [20] による特徴量のモデルへの寄与度を算出し，特徴量の重要度を調査する．

表 4 データセット Crowd $D_{\pm0}$ の分類結果（差分特徴量あり）

		予測	
		適切回答	不良回答
正解	適切回答	505	245
	不良回答	253	623

Precision: 0.673　Recall: 0.666　Accuracy: 0.694

表 5 データセット Crowd $D_{\pm50}$ の分類結果（差分特徴量あり）

		予測	
		適切回答	不良回答
正解	適切回答	789	135
	不良回答	281	421

Precision: 0.854　Recall: 0.737　Accuracy: 0.744

5. 実験結果・考察

表 4，**表 5** に画面操作記録から抽出した基本特徴量と差分特徴量を合わせて学習したモデルの分類結果を示す．表中に正解データの適切回答/不良回答それぞれに対して予測した結果をデータ数で示し，表下部に分類モデルの各評価指標の値を示す．データセット Crowd $D_{\pm0}$ は 0.694，データセット Crowd $D_{\pm50}$ は 0.744 の Accuracy であった．結果から，先行研究での完全一致のみ正解の場合で 0.690，部分一致を許容する場合で 0.738 の Accuracy を示した結果と比較して，同等もしくは少し高い分類精度を取得することができたといえる．

次に，クラウドワーカー実験のデータを用いたデータセットで，ベースラインとの差分特徴量を除いて学習したモデルの分類結果を**表 6**，**表 7** に示す．差分特徴量なしの場合に，データセット Crowd $D_{\pm0}$ は 0.692，データセット Crowd $D_{\pm50}$ は 0.732 の Accuracy を得た．この結果から差分特徴量は，データセット Crowd $D_{\pm0}$ の精度向上への影響は小さいが，データセット Crowd $D_{\pm50}$ では精度向上に寄与しているといえる．これは，完全一致のみを正解とするデータセット Crowd $D_{\pm0}$ では不良回答とみなされているが，部分一致を許容するデータセット Crowd $D_{\pm50}$ の際に適切回答とみなされる境界にあたるようなデータに対

表 6 データセット Crowd $D_{\pm 0}$ の分類結果（差分特徴量なし）

		予測	
		適切回答	不良回答
正解	適切回答	519	231
	不良回答	270	606

Precision: 0.692　Recall: 0.658　Accuracy: 0.692

表 7 データセット Crowd $D_{\pm 50}$ の分類結果（差分特徴量なし）

		予測	
		適切回答	不良回答
正解	適切回答	749	175
	不良回答	260	442

Precision: 0.811　Recall: 0.742　Accuracy: 0.732

表 8 データセット All $D_{\pm 0}$ の分類結果

		予測	
		適切回答	不良回答
正解	適切回答	831	360
	不良回答	428	1007

Precision: 0.698　Recall: 0.660　Accuracy: 0.700

表 9 データセット All $D_{\pm 50}$ の分類結果

		予測	
		適切回答	不良回答
正解	適切回答	1259	186
	不良回答	478	702

Precision: 0.871　Recall: 0.725　Accuracy: 0.747

して個人差を表す特徴量が有用であることを示唆する．

次に，学内学生実験とクラウドワーカー実験のデータ，どちらも含むデータセットを用いて学習したモデルの分類結果を**表 8**，**表 9** に示す．データセット All $D_{\pm 0}$ は 0.700，データセット All $D_{\pm 50}$ は 0.747 の Accuracy であった．結果から，先行研究での学内学生実験のみのデータセットと比較して，完全一致のみを正解とする場合と部分一致を許容する場合どちらにおいてもデータ数を増やすことで 0.01 程度，Accuracy の向上をすることができたといえる．これは被験者数が増え，様々な特徴の適切回答・不良回答のデータを学習することができたからであると考えられる．

本研究では，マウス速度やクリックに関して同様の特徴量を複数種類抽出しているため，特徴量重要度にバイアスが生じている可能性がある．そのため，Permutation importance を用いて特徴量のモデルへの寄与度の分析を行うこととする．**表 10**，**表 11**，**表 12**，**表 13** にそれぞれのデータセットで Permutation Importance による特徴量の寄与度が高かった上位 10 個を示す．

表の左欄は，寄与度上位 10 個の特徴量名，右欄には特徴量の寄与度を示す．各表からどのデータセットにおいても，ラベル付与数（label_num）の寄与度が高いことが分かる．また，学内学生実験・クラウドワーカー実験のデータを合せたデータセット 2 つで文字選択回数に関する特徴量（seleck_num/label_type，select_num/label 等）の寄与度が

表 10 データセット Crowd $D_{\pm 0}$ における寄与度が高い特徴量

特徴量名	寄与度
label_num	0.1890 ± 0.0344
mousespeed_y_boxdiff_0.5s	0.0221 ± 0.0098
click_interval_maxmindiff	0.0202 ± 0.0158
click_interval_min	0.0147 ± 0.0090
mousespeed_y_min10s	0.0141 ± 0.0132
mousespeed_75%_10s	0.0135 ± 0.0107
mousespeed_x_boxdiff_0.5s	0.0129 ± 0.0131
mousespeed_boxdiff_0.5s	0.0095 ± 0.0099
select_num/label_base	0.0104 ± 0.0100
label_type_num	0.0104 ± 0.0063

表 11 データセット Crowd $D_{\pm 50}$ における寄与度が高い特徴量

特徴量名	寄与度
label_num	0.2209 ± 0.0556
mousespeed_x_std10s	0.0374 ± 0.0237
label_type_num	0.0233 ± 0.0143
mousespeed_x_min10s	0.0129 ± 0.0157
mousespeed_min5s	0.0123 ± 0.0039
mousespeed_x_25%_1s	0.0117 ± 0.0072
select_num/time	0.0110 ± 0.0049
mousespeed_max3s	0.0110 ± 0.0092
y_move/time	0.0104 ± 0.0063
mousespeed_25%_0.5s	0.0104 ± 0.0092

表 12 データセット All $D_{\pm 0}$ における寄与度が高い特徴量

特徴量名	寄与度
label_num	0.1551 ± 0.0176
select_num/label_type	0.0129 ± 0.0088
mousespeed_x_maxmindiff3s	0.0129 ± 0.0103
select_num/label	0.0125 ± 0.0117
count_3s	0.0118 ± 0.0061
mousespeed_x_25%_1s	0.0110 ± 0.0181
click_num/label	0.0095 ± 0.0168
mousespeed_max3s	0.0095 ± 0.0099
mousespeed_x_25%_3s	0.0068 ± 0.0046
mousespeed_y_boxdiff0.5s	0.0068 ± 0.0089

高い．これは，不良回答では迷いや集中切れ等が原因で文字選択の画面操作が多くなるからであると考えられ，先行研究 [9] の結果を支持する．

最後に，分類精度がラベル付与数（label_num，label_type_num）に依存している可能性を考慮し，各データセットにおいてそれら 2 つの特徴量を学習させずに評価指標を算出し，Accuracy の評価を行う．各データセットと，それに対応するラベル付与数の特徴量を学習しなかった場合とした場合それぞれの Accuracy を**表 14** に示す．

表 14 各データセットにおけるラベル付与数の分類精度への影響

データセット名	ラベル付与数なし時 Accuracy	ラベル付与数あり時 Accuracy
Crowd $D_{\pm 0}$	0.674	0.694
Crowd $D_{\pm 50}$	0.725	0.744
All $D_{\pm 0}$	0.681	0.700
All $D_{\pm 50}$	0.722	0.747

表 13 データセット All $D_{\pm 50}$ における寄与度が高い特徴量

特徴量名	寄与度
label_num	0.1890 ± 0.0092
select_task/label	0.0213 ± 0.0232
select_num/label	0.0190 ± 0.0110
label_type_num	0.0183 ± 0.0124
x_move/label	0.0125 ± 0.0104
mousespeed_y_min0.5s	0.0114 ± 0.0130
mousespeed_min10s	0.0091 ± 0.0061
select_task/label_type	0.0087 ± 0.0137
mousespeed_75%_0.5s	0.0076 ± 0.0068
select_num/time	0.0076 ± 0.0083

結果から各データセットでラベル付与数を学習しなかった場合に 0.02 程度 Accuracy が低下していることが分かる．このことから，ラベル付与数の特徴量の重要度は高いものの，マウス速度や文字選択等の画面操作記録の特徴量のみでも分類が可能であるといえる．そのため，各モデルではラベル付与数と画面操作記録の特徴量を組み合わせることにより分類を行っていると考えられる．

6. おわりに

本研究はクラウドソーシングにおけるマイクロタスク，特に固有表現アノテーションを対象とし，ユーザが取る不良回答をリアルタイムに検知する手法を提案し，機械学習モデルを構築したのち評価を行った．提案手法はアノテーション作業中の画面操作をリアルタイムに記録し，特徴量を抽出することで作業実施中の不良回答検出を実現する．機械学習モデルによる分類では，クラウドワーカーを対象にデータ取得及び検証実験を行った結果，学内学生を対象とした先行研究と同等もしくは少し高い分類精度を取得することができ，特徴量の重要度からラベル付与数が分類に重要であることが分かった．また，学内学生とクラウドワーカーのデータを合わせたモデルでは，部分一致を許容するデータセットで 0.747 の Accuracy で分類することができ，データ数を増やすことで分類精度の改善をすることができた．

今後は，実用に向けて特徴量やデータの追加を行うことなどで分類精度の向上を図る．本稿では，各タスク実施時に得られる特徴量のみを抽出しているが，累積の付与ラベル数や直近タスクでの付与ラベル数などの時系列的な特徴量を扱うことで不良回答傾向を検出できる可能性がある．また，連続で不良回答と分類された場合に，不良回答とみなし介入を行うなどの手法を取ることで精度の改善をできると考えられる．

謝辞 本研究の一部は，JST さきがけ（JPMJPR2039）の助成を受けたものである．

参考文献

[1] Daniel, F., Kucherbaev, P., Cappiello, C., Benatallah, B. and Allahbakhsh, M.: Quality Control in Crowdsourcing: A Survey of Quality Attributes, Assessment Techniques, and Assurance Actions, *ACM Comput. Surv.*, Vol. 51, No. 1 (online), DOI: 10.1145/3148148 (2018).

[2] Kolvoort, I. R. and van Maanen, L.: Causal reasoning under time pressure: testing theories of systematic non-normative reasoning patterns, *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 43, No. 43 (2021).

[3] Maddalena, E., Ibáñez, L.-D., Reeves, N. and Simperl, E.: Qrowdsmith: Enhancing Paid Microtask Crowdsourcing with Gamification and Furtherance Incentives, *ACM Trans. Intell. Syst. Technol.*, Vol. 14, No. 5 (online), DOI: 10.1145/3604940 (2023).

[4] Oppenheimer, D. M., Meyvis, T. and Davidenko, N.: Instructional manipulation checks: Detecting satisficing to increase statistical power, *Journal of Experimental Social Psychology*, Vol. 45, No. 4, pp. 867–872 (online), DOI: https://doi.org/10.1016/j.jesp.2009.03.009 (2009).

[5] Maniaci, M. R. and Rogge, R. D.: Caring about carelessness: Participant inattention and its effects on research, *Journal of Research in Personality*, Vol. 48, pp. 61–83 (online), DOI: https://doi.org/10.1016/j.jrp.2013.09.008 (2014).

[6] 後上正樹，松田裕貴，荒川豊，安本慶一：オンラインアンケート回答時のスマートフォン画面操作状況に基づく不適切回答検出，第 25 回一般社団法人情報処理学会シンポジウム・インタラクション 2021，pp. 11–20 (2021).

[7] Gogami, M., Matsuda, Y., Arakawa, Y. and Yasumoto, K.: Detection of Careless Responses in Online Surveys Using Answering Behavior on Smartphone, *IEEE Access*, Vol. 9, pp. 53205–53218 (online), DOI: 10.1109/ACCESS.2021.3069049 (2021).

[8] 福光嘉伸，松田裕貴，諏訪博彦，安本慶一：クラウドソーシングを用いたアノテーションにおける不良回答の検出手法，研究報告ユビキタスコンピューティングシステム（UBI），Vol. 2023-UBI-79, No. 18, pp. 1–6 (2023).

[9] 福光嘉伸，松田裕貴，諏訪博彦，安本慶一：固有表現アノテーションにおける画面操作記録を用いた不良回答検出，研究報告ヒューマンコンピュータインタラクション（HCI），Vol. 2024-HCI-206, No. 40, pp. 1–7 (2024).

[10] Otani, M., Togashi, R., Nakashima, Y., Rahtu, E., Heikkilä, J. and Satoh, S.: Optimal Correction Cost for Object Detection Evaluation, *The IEEE/CVF Computer Vision and Pattern Recognition Conference*, CVPR'22 (2022).

[11] Qiu, S., Gadiraju, U. and Bozzon, A.: Improving Worker Engagement Through Conversational Microtask Crowdsourcing, *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI'20, pp. 1–12 (online), DOI: 10.1145/3313831.3376403 (2020).

[12] Rzeszotarski, J. M., Chi, E., Paritosh, P. and Dai, P.: Inserting micro-breaks into crowdsourcing workflows, *The First AAAI Conference on Human Computation and Crowdsourcing*, HCOMP'13, pp. 62–63 (2013).

[13] Masaru, S., Kagawa, R. and Hidehito, H.: A one-second wait improves judgment accuracy: A mouse tracking reveals cognitive processes during choice behaviors, *Proceedings of the 45th Annual Conference of the Cognitive Science Society* (2023).

[14] HumanSignal, Inc.: Label Studio, `https://github.com/heartexlabs/label-studio` (2019). (Accessed on 2023-09-01).

[15] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. and Liu, T.-Y.: LightGBM: A Highly Efficient Gradient Boosting Decision Tree, *Advances in Neural Information Processing Systems*, NIPS'17, Vol. 30 (2017).

[16] Chawla, N. V., Bowyer, K. W., Hall, L. O. and Kegelmeyer, W. P.: SMOTE: synthetic minority oversampling technique, *J. Artif. Int. Res.*, Vol. 16, No. 1, p. 321–357 (2002).

[17] Ouchi, H., Shindo, H., Wakamiya, S., Matsuda, Y., Inoue, N., Higashiyama, S., Nakamura, S. and Watanabe, T.: Arukikata Travelogue Dataset, *arXiv*, No. 2305.11444, pp. 1–6 (online), DOI: 10.48550/arXiv.2305.11444 (2023).

[18] Higashiyama, S., Ouchi, H., Teranishi, H., Otomo, H., Ide, Y., Yamamoto, A., Shindo, H., Matsuda, Y., Wakamiya, S., Inoue, N., Yamada, I. and Watanabe, T.: Arukikata Travelogue Dataset with Geographic Entity Mention, Coreference, and Link Annotation, *arXiv*, No. 2305.13844, pp. 1–11 (online), DOI: 10.48550/arXiv.2305.13844 (2023).

[19] Strobl, C., Boulesteix, A.-L., Zeileis, A. and Hothorn, T.: Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution., *BMC bioinformatics*, Vol. 8, p. 25 (online), DOI: 10.1186/1471-2105-8-25 (2007).

[20] Fisher, A., Rudin, C. and Dominici, F.: All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously (2019).