修士論文

クラウドソーシングベースのアノテーションタスクにおける回答の質低下の動的検知

福光 嘉伸
奈良先端科学技術大学院大学
先端科学技術研究科
情報理工学プログラム

主指導教員: 安本 慶一
ユビキタスコンピューティングシステム 研究室（情報科学領域）

令和 6 年 3 月 15 日提出

本論文は奈良先端科学技術大学院大学先端科学技術研究科に
修士 (工学) 授与の要件として提出した修士論文である。

福光 嘉伸

審査委員：

主査　安本 慶一　（情報科学領域 教授）
荒牧 英治　（情報科学領域 教授）
諏訪 博彦　（情報科学領域 准教授）
松田 裕貴　（情報科学領域 助教）

クラウドソーシングベースのアノテーションタスクにおける回答の質低下の動的検知*

福光 嘉伸

内容梗概

アノテーション作業をクラウドソーシングで行うことにより，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，得られるデータの品質に大きなばらつきがあり，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って不良回答が発生する問題がある．誤ったラベルが含まれる学習データを用いると，機械学習モデルの精度が低下する恐れがあるため，不良回答の発生を検知・防止することが重要となる．先行研究では，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が，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検出を行うことができない．そのため，データの母集団が限られている場合，不良回答の影響によってデータが不足する恐れがある．そこで，本研究ではアノテーションタスク，特に固有表現（Named Entity）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，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から得られる特徴量を用いた検出手法を提案する．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 を用いて画面操作を記録するプラグインを開発し，学内学生およびクラウドワーカーを対象として画面操作データを収集して機械学習による不良回答検出を行った．LightGBM（LGBM）を分類アルゴリズムとして用い，Leave-One-Participant-Out 交差検証による評価を行った結果，学内学生を対象とした実験では 0.738 の Accuracy が得られた．また，不

*奈良先端科学技術大学院大学 先端科学技術研究科 修士論文, 令和 6 年 3 月 15 日.

良回答ではラベル付与毎の平均カーソル移動量・クリック回数・文字選択回数が多くなり，分類において重要であることが分かった．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，学内実験と同等もしくは少し高い分類精度を示し，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において 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を改善できた．これらのモデルでは，特徴量の重要度より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で分類を行っていると考察した．

キーワード

アノテーション，クラウドソーシング，不良回答検出，機械学習

Dynamic Detection of Response Quality Deterioration for Annotation Tasks in Crowdsourcing*

Yoshinobu Fukumitsu

Abstract

Annotation tasks can be performed via crowdsourcing to collect training data for machine learning at a low cost. However, the quality of the data obtained can vary, and there is an issue with careless responses due to workers trying to answer as quickly as possible to earn rewards. It is important to detect and prevent the occurrence of careless responses, as training data containing incorrect labels may reduce the accuracy of the machine learning model. Prior research has proposed machine learning-based detection methods using smartphone screen operation logs as features to detect ”Satisficing” without using evaluation scales. However, this method can only be used to detect careless responses after all of the survey responses have been completed. Therefore, if the data population is limited, the data may be insufficient due to the influence of careless responses. Therefore, this study proposes a real-time method for detecting careless responses in annotation tasks, especially named entity annotation. This method utilizes features such as cursor movement and response time, obtained from screen operations during the annotation task. We have developed a plugin for Label-Studio, which records screen operations carried out during annotation tasks. Screen operation logs were collected from graduate students at Nara Institute of Science and Technology and cloudworkers, and machine learning was used to detect careless responses. We used LightGBM (LGBM) as the classification algorithm and performed

*Master’s Thesis, Graduate School of Science and Technology, Nara Institute of Science and Technology, March 15, 2024.

evaluation using Leave-One-Participant-Out cross-validation. As a result of the experiment with graduate students, an accuracy of 0.738 was obtained. The average cursor movement, the number of clicks, and the number of text selections per label assignment increased for careless responses, which was important in the classification. As a result of data collection and validation experiments on crowdworkers, it is shown that the classification accuracy is the same or slightly higher than in student experiments. From the feature importance, we found that the number of assigned labels was important for classification. We obtained an accuracy of 0.747 in the dataset that combining data from students and crowdworkers, and the classification accuracy was improved by increasing the number of data. Based on the feature importance, it is considered that in these models, classification was performed by combining the number of assigned labels and screen operation features.

Keywords:

Annotation, Crowdsourcing, Detection of Careless Responses, Machine Learning

目 次

1. 序論 **1**

2. 関連研究 **4**

3. 画面操作記録を用いた不良回答検出法 **6**

3.1 想定シナリオと提案手法の概要 6

3.2 固有表現アノテーションの作業手順 8

3.3 画面操作記録プラグインと抽出する特徴量 8

3.4 不良回答検知モデルの構築 11

4. 学内実験 **12**

4.1 データ収集実験 12

4.1.1 アノテーションタスクの概要 12

4.1.2 データセット 13

4.2 モデルの評価方法 14

4.3 実験結果・考察 15

5. クラウドソーシング実験 **20**

5.1 データ収集実験 20

5.1.1 アノテーションタスクの概要 20

5.1.2 データセット 20

5.2 モデルの評価方法 21

5.3 実験結果・考察 22

6. 結論 **32**

謝辞 **33**

参考文献 **34**

研究業績 **37**

図 目 次

1 提案手法の概要 6
2 ラベル付与を行う画面例 7
3 ラベル付与の様子 7
4 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 13
5 データセット Stu $D_{\pm 0}$ の特徴量の重要度 16
6 データセット Stu $D_{\pm 50}$ の特徴量の重要度 17
7 ラベル付与毎の平均カーソル移動量 18
8 ラベル付与毎の平均クリック回数 18
9 ラベル付与毎の平均文字選択回数 19
10 データセット Crowd $D_{\pm 0}$ の特徴量の重要度 25
11 データセット Crowd $D_{\pm 50}$ の特徴量の重要度 26
12 データセット All $D_{\pm 0}$ の特徴量の重要度 27
13 データセット All $D_{\pm 50}$ の特徴量の重要度 28

表 目 次

1 抽出する特徴量 10
2 各データセットに含まれるデータ数（学内実験） 14
3 データセット Stu $D_{\pm 0}$ の分類結果 15
4 データセット Stu $D_{\pm 50}$ の分類結果 15
5 各データセットに含まれるデータ数（クラウドソーシング実験） . 21
6 各データセットに含まれるデータ数（2 実験合計） 21
7 データセット Crowd $D_{\pm 0}$ の分類結果（差分特徴量あり） 22
8 データセット Crowd $D_{\pm 50}$ の分類結果（差分特徴量あり） 22
9 データセット Crowd $D_{\pm 0}$ の分類結果（差分特徴量なし） 23
10 データセット Crowd $D_{\pm 50}$ の分類結果（差分特徴量なし） 23
11 データセット All $D_{\pm 0}$ の分類結果 24

12　データセット All $D_{\pm 50}$ の分類結果 24
13　Permutation Importance による寄与度（データセット Crowd $D_{\pm 0}$） 29
14　Permutation Importance による寄与度（データセット Crowd $D_{\pm 50}$） 29
15　Permutation Importance による寄与度（データセット All $D_{\pm 0}$） . . 30
16　Permutation Importance による寄与度（データセット All $D_{\pm 50}$） . . 30
17　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 31

1. 序論

マイクロタスク型クラウドソーシングとは，インターネット上で不特定多数の群衆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，低コストで大規模のデータを取得および解析することを可能とする方法であり，様々な用途での活用が進んでいる．特に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ことから，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，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，品質管理が課題となっている [1]．クラウドソーシングでは，アノテーションを行うユーザが必ずしも正確に回答するとは限らず，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ように行動すること（努力の最小限化）が考えられる．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により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恐れがあるため，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．三浦ら [4] は The Attentive Responding Scale（ARS）という矛盾を問う評価尺度を質問紙に取り入れることで，努力の最小限化の傾向を示す個人を検出する方法を提案している．また，同様に評価尺度を質問紙に取り入れることで，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 [5] や Directed Questions Scale [6] がある．しかし，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，回答者の内発的動機が損なわれることで，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 [7, 8] 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検出を行うことができないことが問題となりうる．同一人物が一度しか実施することができないタスク（調査内容を知る前と知った後で回答が変化する性質があるもの）など，データの母集団が限られている場合，これまでの方法では不良回答の影響によって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ても同様に問題となりうる．本人しかアノテーションができないようなタスクの場合，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う重要性が高いと考える．

本研究ではアノテーションタスク，特に固有表現（Named Entity）アノテーション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する．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．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 を用いて画面操作を記録するプラグインを開発し，学内学生およびクラウドワーカーを対象として学習データの取得実験を実施する．実験により得られた画面操作記録から特徴量を抽出し，適切回答・不良回答の分類を行うモデルを構築し，不良回答検出を行う．また，クラウドワーカーを対象とした実験では，学内実験の検出結果を踏まえて新たに特徴量を追加する．

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 交差検証による評価方法を用いた．その結果，学内学生を対象とした場合では 0.738 の Accuracy が得られ，ラベル付与数と，ラベル付与毎の平均カーソル移動量・クリック回数が，分類において重要であることが分かった．また，分類に失敗したデータについては適切回答/不良回答間で特徴の違いが小さいもしくは混在していると考えられる．クラウドソーシング上でクラウドワーカーを対象とした場合では，例題作

業時の特徴量値をベースラインとして設定し，差分特徴量を算出し特徴量として扱うことにより，0.747 の Accuracy が得られた．分析の結果，ラベル付与数の特徴量の重要度が高く，不良回答・適切回答の境界にあたるようなデータに対してベースラインとの差分特徴量が有用であることが分かった．

本論文の構成は次の通りである．2 章で関連研究について紹介する．3 章で本研究で提案する画面操作記録を用いた不良回答検出法について述べる．4 章では，学内学生を対象としたデータ収集での実験設定について説明し，得られたデータを用いた機械学習による不良回答検出の結果および考察を述べる．5 章では，実際にクラウドワーカーを対象としデータ収集および検証実験を行った結果と，機械学習での評価結果およびそれらの考察を述べる．最後に，6 章で本研究のまとめと今後の展望を述べる．

2. 関連研究

不良回答を検出する既存研究としては，リッカート式やテキストベースのアンケートを対象としたものが一般的である．後上ら [7, 8] 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．加えて，中川ら [9] は，アンケート回答中の迷いが反映されたタッチ操作ログを取得するべく，スライドバーや拡大鏡を活用した 2 種類の回答 UI を提案している．また，Mara ら [10] は，実際の職場環境におけるストレスを，マウス動作やキーボードのタイピング動作，心拍変動から特徴量を抽出することによって検出する手法を提案している．

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 [11] 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モチベーションや作業及びデータの質を向上させるための研究は多数行われている．Sihang ら [12] 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 [13] 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．また，白砂ら [14] 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．

このように，不良回答を検出する既存研究 [7, 8] では，リアルタイム性がなく，回答が全て終わってからしか不良回答の検出を行うことができないという課題を残している．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．アノテーションやマイクロタスクに対する既存研究 [12, 13] では，やる気や定着率を向上させる点に留まっており，出力結果（データセット）の質を向上させること

は難しい．待ち時間を作業前に設ける研究 [14] では，数%の改善にとどまっている，かつ長い思考時間が必要となるタスクではその効果は小さいと考えられる．これらのことから，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．

3. 画面操作記録を用いた不良回答検出法

本研究では，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．

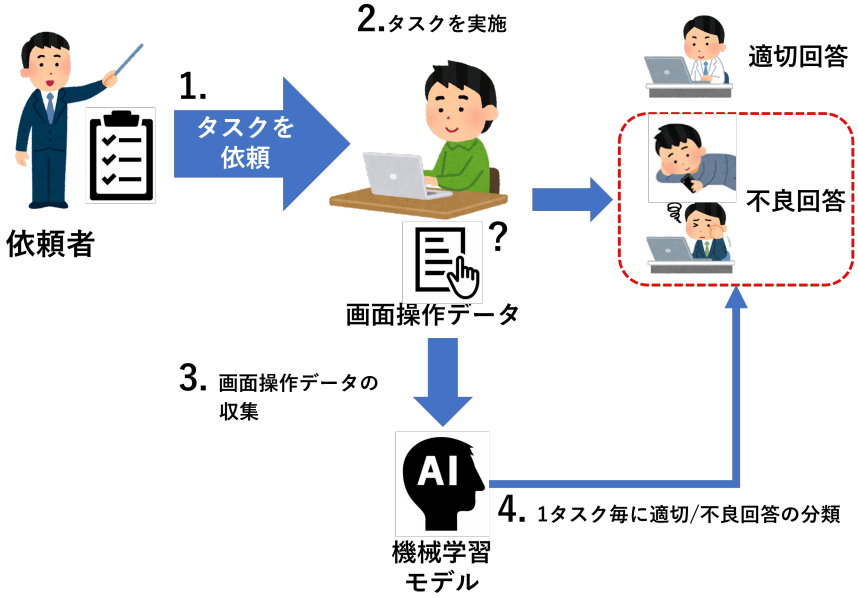

図 1 提案手法の概要

3.1 想定シナリオと提案手法の概要

本研究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. 得られたログを入力とする機械学習モデルによって，不良回答を検出する．

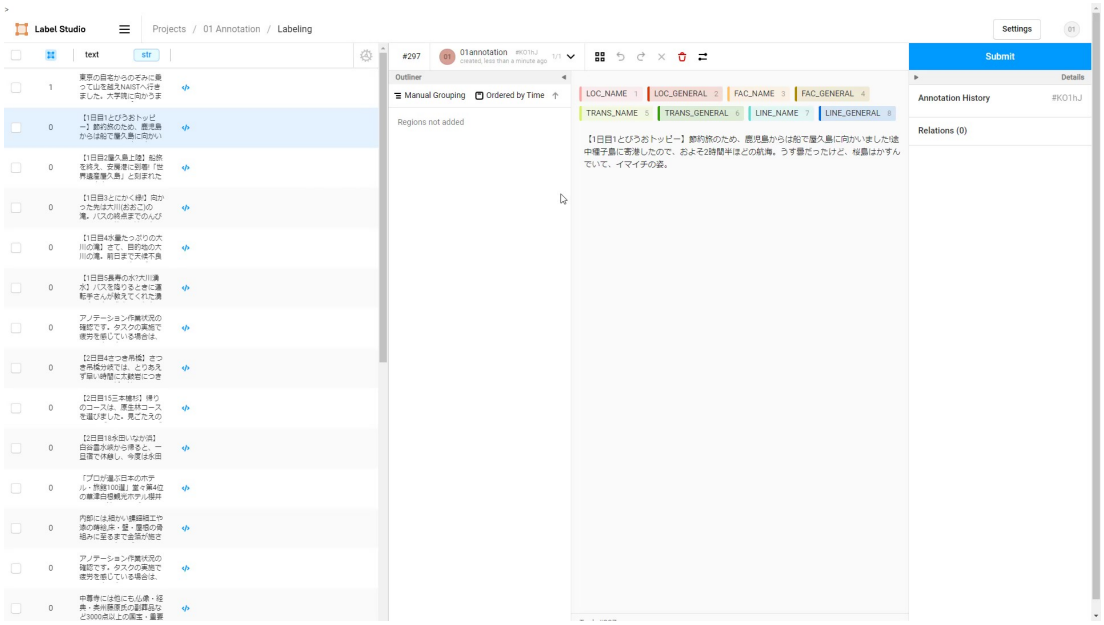

図 2 ラベル付与を行う画面例

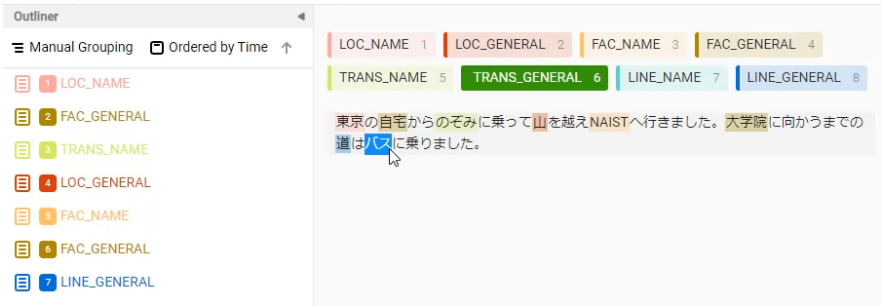

図 3 ラベル付与の様子

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 [15] 上で行うことを想定する（図 2）．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様子を図 3 に示し，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.
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 3 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．表 1 に用いる特徴量とその単位とを示す．本研究では，後上らの手法 [8] で用いられた特徴量に加えて，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新たに採り入れることとした．

本研究で使用する基本となる特徴量（以降，基本特徴量）を表 1 に “〇” で示す．後述の 4 章では，この基本特徴量を用いた学内の学生を対象とした実験を行っている．その結果として，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．なお，差分特徴量を追加した場合のモデルの性能評価については 5 章にて追実験した結果を報告する．

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，ラベル付与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．

表 1 抽出する特徴量

特徴量	単位	基本特徴量	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	○	
ラベル付与数（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	○	

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 [8] で最も優れた性能を示した LightGBM（LGBM）[16] 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 [17] 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. 学内実験

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を対象として募集を行い，被験者数は 61 人であった．被験者の内訳は，20 代の男性 39 名，女性 22 名で情報工学を専門としない学生も含む．被験者に，実際に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．被験者には，報酬として 1000 円相当のギフトカードを付与した．なお，本研究は奈良先端科学技術大学院大学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[1] [18, 19]」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 4 に示す．

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

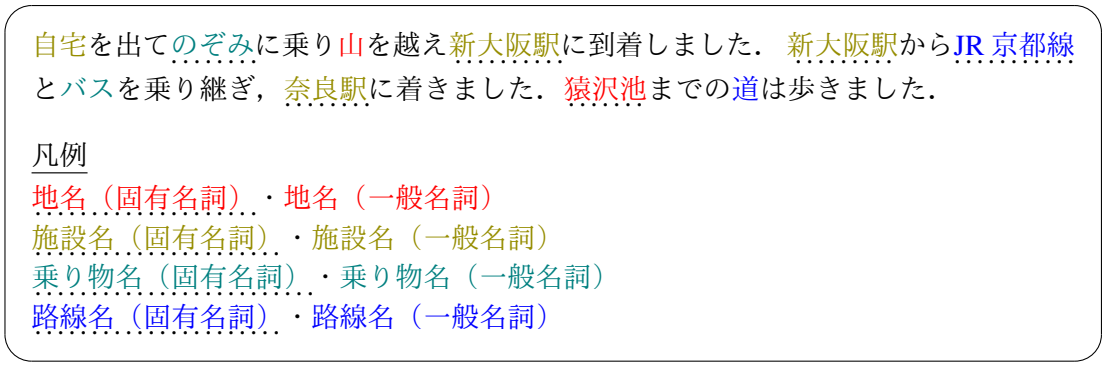

図 4 地球の歩き方旅行記データセットを用いた固有表現アノテーションの実施例

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（1 文毎）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．被験者 61 人のうち，指示通りに作業を実施した 56 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）を各タスクに対しての評価値とする．以下の式 1，2，3 より各タスク（1 文毎）において F1 値を算出する．

$$Precision = \frac{\text{文中のラベル付与正答数}}{\text{文中のラベル付与数}} \tag{1}$$

$$Recall = \frac{\text{文中のラベル付与正答数}}{\text{グランドトゥルースのラベル数}} \tag{2}$$

$$F1score = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

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 Stu $D_{\pm 0}$ を作成する．

また，自然言語領域において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 Stu $D_{\pm 50}$ を作成する．

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．

表2 各データセットに含まれるデータ数（学内実験）

	データセット Stu $D_{\pm 0}$	データセット Stu $D_{\pm 50}$
適切回答	421	521
不良回答	578	478

4.2 モデルの評価方法

データセット Stu $D_{\pm 0}$，Stu $D_{\pm 50}$ それぞれを用いて機械学習モデルを構築し，提案手法の性能を評価する．本研究では，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，モデルの推定に寄与している特徴量の分析のため，ジニ係数による重要度を算出する．

4.3 実験結果・考察

表 3，表 4 に画面操作記録から抽出した基本特徴量を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 Stu $D_{\pm 0}$ は 0.690，データセット Stu $D_{\pm 50}$ は 0.738 の Accuracy であった．結果から，固有表現アノテーションでは，部分一致を許容するタスクの場合に，不良回答がより検知がしやすいといえる．これは完全一致のみ正解ラベルとする場合は適切回答の基準が厳しく，正常に作業を行っている人も不適切回答に分類されるからだと考えられる．また，分類に失敗したデータについては適切回答/不良回答間で特徴の違いが小さいもしくは混在していると考えられる．主な原因として，アノテーション作業の実施において作業速度や本人の能力などの個人差や，文章構造によって生じるタスク間の難易度差が挙げられる．

表 3 データセット Stu $D_{\pm 0}$ の分類結果

		予測	
		適切回答	不良回答
正解	適切回答	287	176
	不良回答	134	402

Precision: 0.612　Recall: 0.682　Accuracy: 0.690

表 4 データセット Stu $D_{\pm 50}$ の分類結果

		予測	
		適切回答	不良回答
正解	適切回答	413	154
	不良回答	108	324

Precision: 0.728　Recall: 0.793　Accuracy: 0.738

図 5，図 6 にそれぞれのデータセットで特徴量の重要度が高かった上位 20 個

を示す．図の縦軸は各特徴量を示し，横軸の値が大きいほどその特徴量が分類精度の向上に重要であることを示す．データセット Stu $D_{\pm 0}$，Stu $D_{\pm 50}$ のどちらにおいても重要度の上位 20 個に入っていた特徴量はその特徴量名の左に「*印」を付している．図からラベル付与数（label_num）や，ラベル付与毎のカーソル移動量・クリック回数・文字選択回数に関する特徴量（cursor_move/label, x_move/label, y_move/label, click_num/label, select_num/label）が分類に重要であることが分かる．また，マウス速度（mousespeed_y_min1s）やクリック間隔（click_interval_min, click_interval_avg）も同様に分類に重要であるといえる．

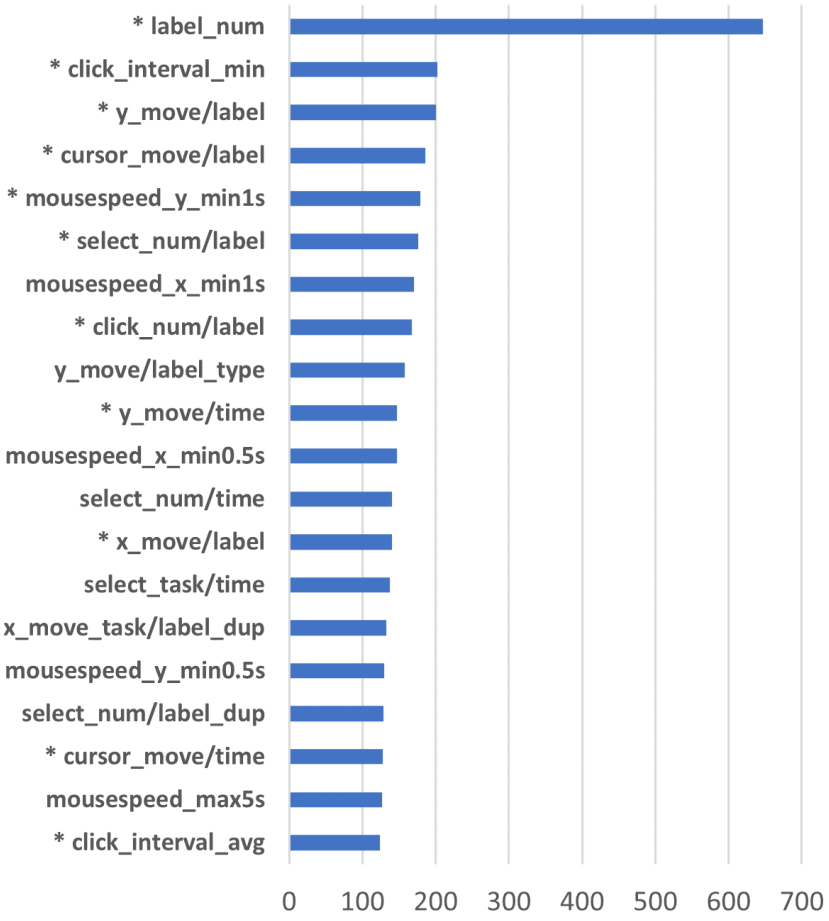

図 5 データセット Stu $D_{\pm 0}$ の特徴量の重要度

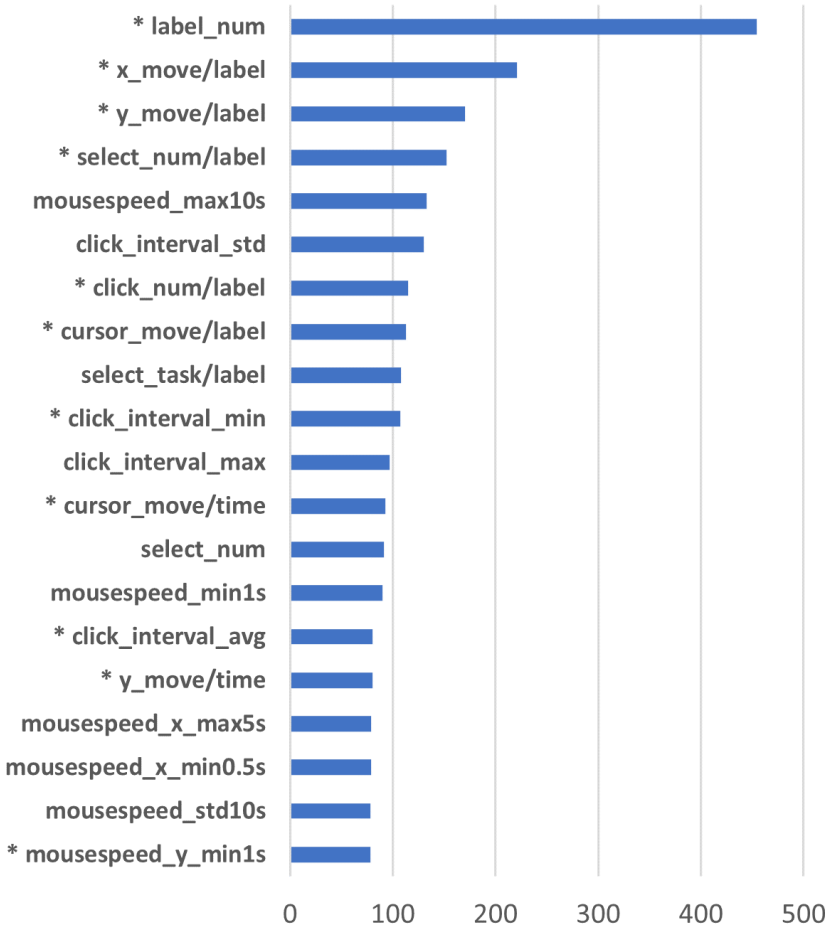

図 6 データセット Stu $D_{\pm 50}$ の特徴量の重要度

次に，どちらのデータセットでも特徴量の重要度分析で上位にあった特徴量の中から，ラベル付与毎のカーソル移動量（合計）・クリック回数・文字選択回数について，適切回答群と不良回答群との比較を箱ひげ図で図 7，図 8，図 9 に示す．図より，データセット間で大きな傾向の違いはないといえるが，適切回答群に比べて不良回答群は，ラベル付与毎のカーソル移動量・クリック回数・文字選択回数については，不良回答群では中央値が大きいかつ，中央値から最大値にかけての値のばらつきが大きいことが分かる．このことから，不良回答では迷いや集中切れ等が原因でカーソル移動やクリック，文字選択などの画面操作が多くなる特

徴があると考えられる．そのため，カーソル速度の低速移動の回数・秒数や，マウスカーソルの直線/曲線動作回数等の迷いや集中切れを表す特徴量を抽出することで精度が改善できる可能性がある．また，分類失敗の原因として挙げた作業者の個人差に関して，作業速度や作業時間に依存しないような特徴量を増やすことで同様に分類精度が改善できると考えられる．

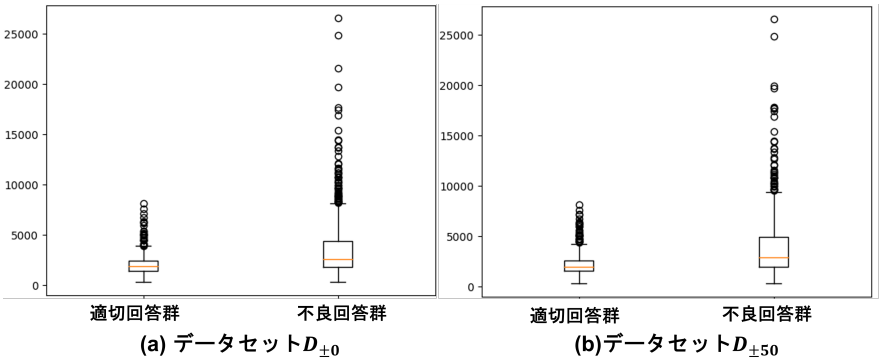

図 7 ラベル付与毎の平均カーソル移動量

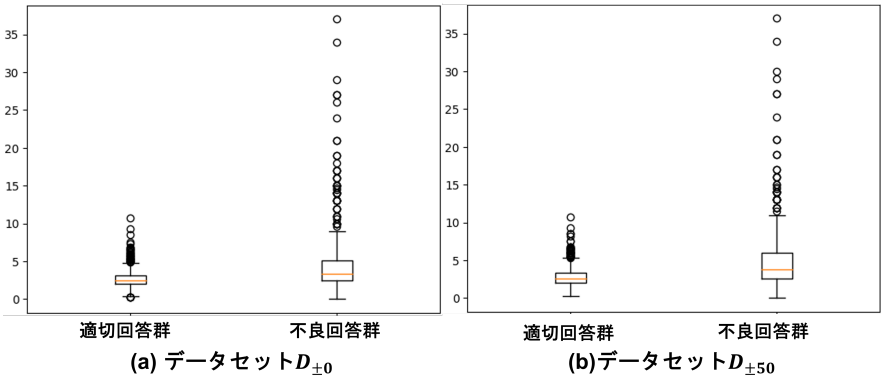

図 8 ラベル付与毎の平均クリック回数

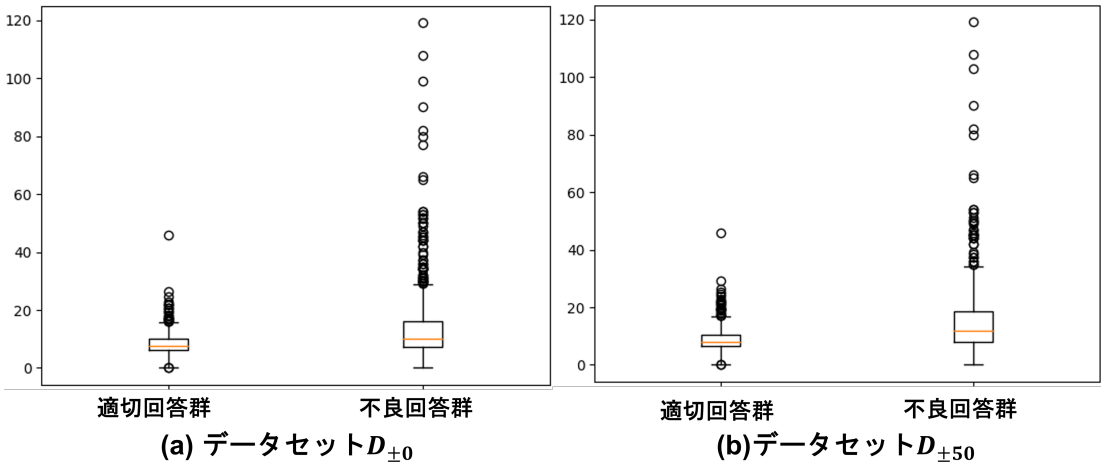

図 9 ラベル付与毎の平均文字選択回数

5. クラウドソーシング実験

5.1 データ収集実験

学内実験と同様に 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クラウドソーシングでの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，クラウドソーシングサービスであるクラウドワークス[2]とランサーズ[3]のワーカーを対象として募集を行い，被験者数はクラウドワークス 76 人，ランサーズ 23 人の計 99 人であった．被験者の内訳は，10 代から 60 代の男性 40 名，女性 59 名である．被験者に，契約金額として 1000 円を支払い，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．

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．

5.1.1 アノテーションタスクの概要

本実験は 4.1.1 項と同じタスク設定を用いて，日本語文章の 24 文を選択した．また，3.2 項に示した個人差を反映した差分特徴量を抽出するために，2 文の例題をタスクの前に追加で設定した．操作に慣れていない等の影響を考慮し，1 文目は差分特徴量算出の対象とせず，2 文目を算出の対象として抽出することとした．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．

5.1.2 データセット

4.1.2 項と同様に，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・不良回答を判別しデータセットを構築する．実験参加者 99 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 90 人のデータを分析対象とする．F1 値が 0.7 以上のラベ

[2] https://crowdworks.jp/
[3] https://www.lancers.jp/

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 $D_{\pm 0}$ を作成するとともに，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 $D_{\pm 50}$ を作成する．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 5 にそれぞれに示す．

次に，学内実験のデータとクラウドソーシング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 $D_{\pm 0}$ とデータセット All $D_{\pm 50}$ を作成し，適切回答・不良回答のデータ数を表 6 にそれぞれ示す．

表 5 各データセットに含まれるデータ数（クラウドソーシング実験）

	データセット Crowd $D_{\pm 0}$	データセット Crowd $D_{\pm 50}$
適切回答	750	924
不良回答	876	702

表 6 各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 $D_{\pm 0}$	データセット All $D_{\pm 50}$
適切回答	1191	1463
不良回答	1435	1163

5.2 モデルの評価方法

データセット Crowd $D_{\pm 0}$，Crowd $D_{\pm 50}$，All $D_{\pm 0}$，All $D_{\pm 50}$ 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 $D_{\pm 0}$，Crowd $D_{\pm 50}$ 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．4.2 節と同じく LOPOCV により交差検証を行い，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 [20]．そのため，クラウドソーシング実験のモデルに関して，ジニ係数 [21] による重要度を算出するとともに，Permutation importance [22] による特徴量のモデルへの寄与度も算出することとする．

5.3 実験結果・考察

表 7，表 8 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 $D_{\pm 0}$ は 0.694，データセット Crowd $D_{\pm 50}$ は 0.744 の Accuracy であった．結果から，4.3 節の学内実験の結果と比較して，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．また，データセット Crowd $D_{\pm 0}$ に比べ，データセット Crowd $D_{\pm 50}$ で不良回答を誤って適切回答と分類してしまうデータが増加する要因として，部分一致を許容することで適切回答のデータ割合が多くなり特徴が多様になったことが考えられる．

表 7 データセット Crowd $D_{\pm 0}$ 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 8 データセット Crowd $D_{\pm 50}$ 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

次に，クラウドソーシング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 9，表 10 に示す．差分特徴量なしの場合に，データセット Crowd $D_{\pm 0}$ は 0.692，データセット Crowd $D_{\pm 50}$ 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 $D_{\pm 0}$

の精度向上への影響は小さいが，データセット Crowd $D_{\pm 50}$ 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 $D_{\pm 0}$ 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 $D_{\pm 50}$ 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して個人差を表す特徴量が有用であることを示唆する．

表 9 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 10 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

次に，学内実験とクラウドソーシング実験のデータ，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 11，表 12 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，学内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．

図 10，図 11，図 12，図 13 にそれぞれのデータセットで特徴量の重要度が高かった上位 20 個を示す．図の縦軸は各特徴量を示し，横軸の値が大きいほどその特徴

量が分類精度の向上に重要であることを示す．図からラベル付与数（label_num）どのデータセットでも分類に重要であることが分かる．また，差分特徴量を含むデータセットの，データセット Crowd $D_{\pm 0}$ ・データセット Crowd $D_{\pm 50}$ においては，一部の差分特徴量 (y_move/time_base，click_interval_25%_base) がどちらのデータセットの特徴量の重要度の上位に入っていることが分かる．そのため，差分特徴量は，個人差を考慮した特徴量を扱うことができるため一部のデータの分類に対して有用であると考えられる．

表 11 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 12 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

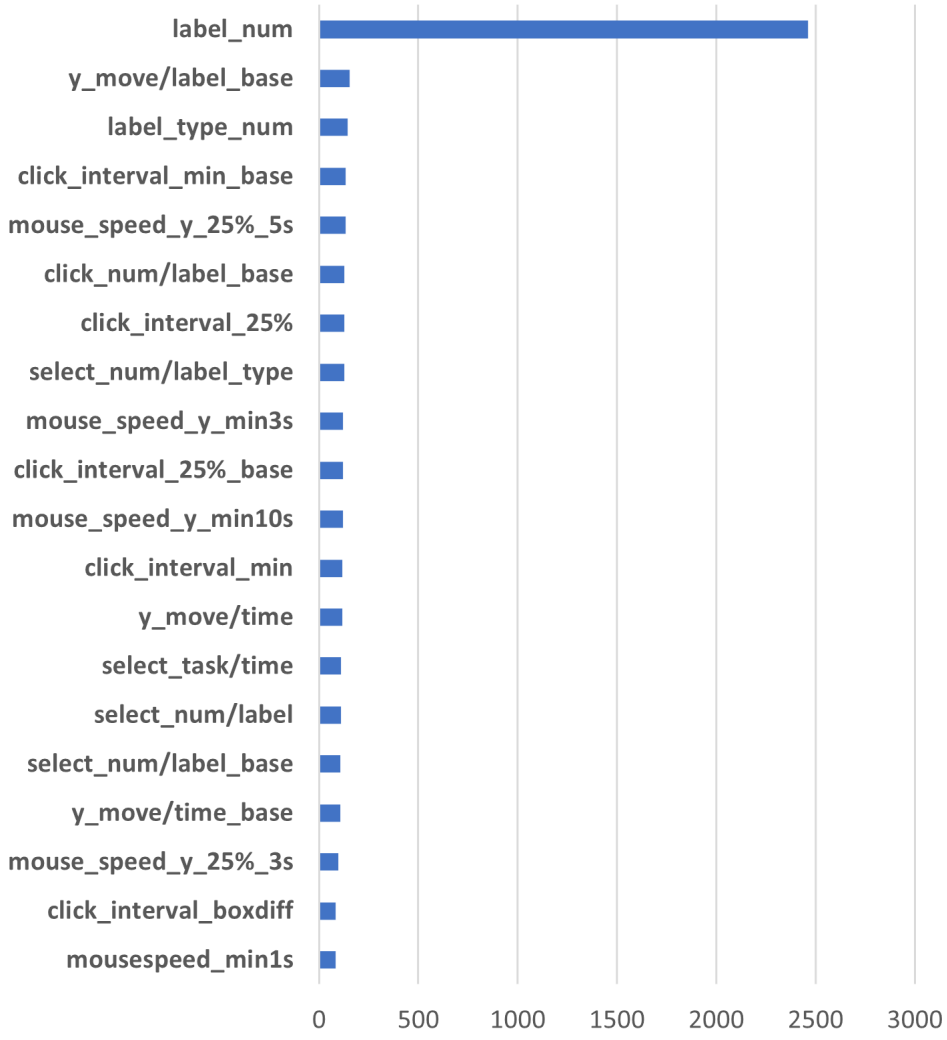

図 10 データセット Crowd $D_{\pm 0}$ の特徴量の重要度

図 11 データセット Crowd $D_{\pm 50}$ の特徴量の重要度

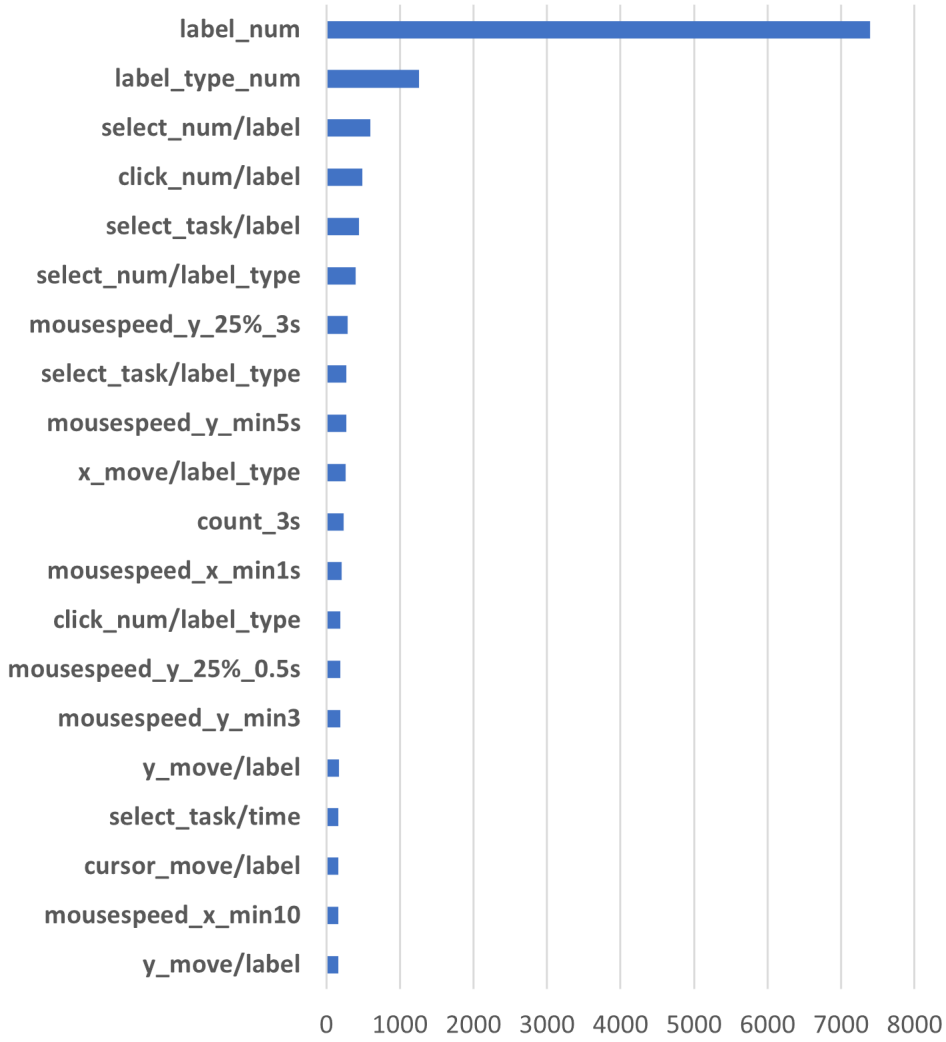

図 12 データセット All $D_{\pm 0}$ の特徴量の重要度

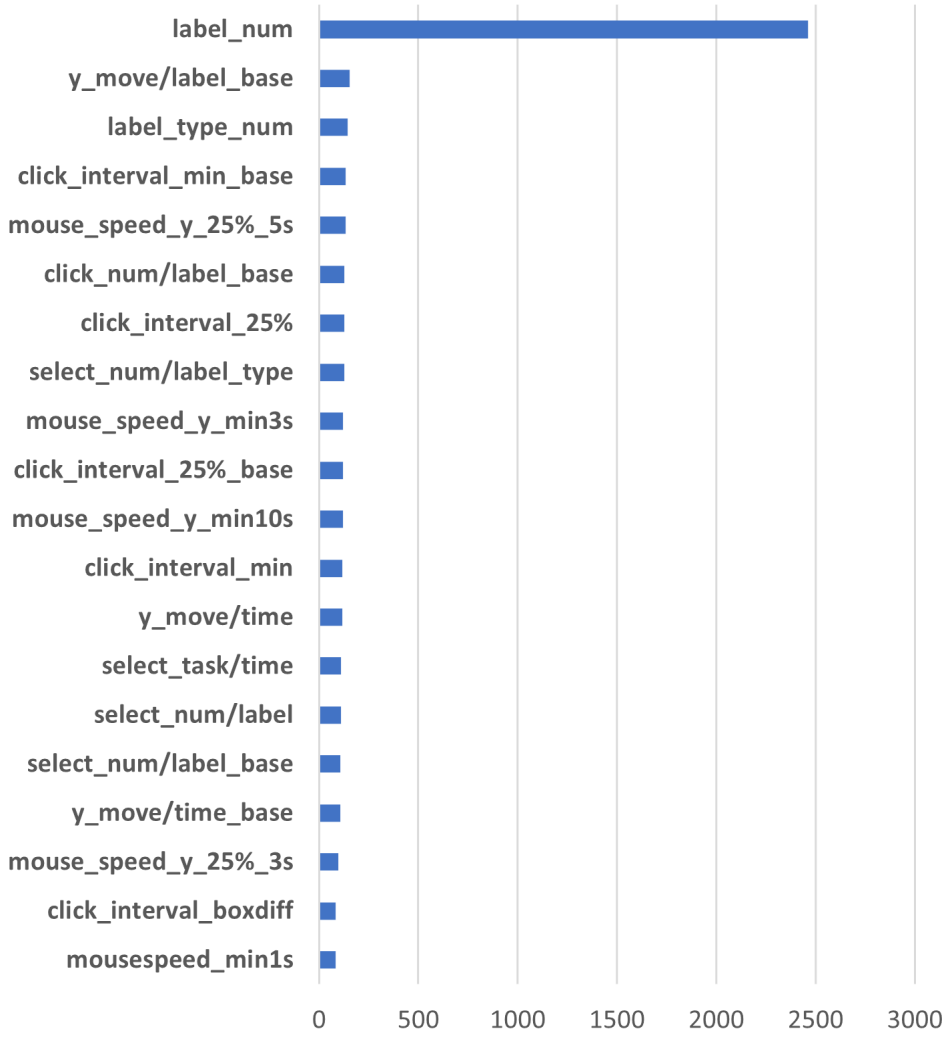

図 13 データセット All $D_{\pm 50}$ の特徴量の重要度

いずれのデータセットでもラベル付与数 (label_num) の重要度が特に高い．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 13，表 14，表 15，表 16 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 13 Permutation Importance による寄与度（データセット Crowd $D_{\pm 0}$）

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 14 Permutation Importance による寄与度（データセット Crowd $D_{\pm 50}$）

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 15 Permutation Importance による寄与度（データセット All $D_{\pm 0}$）

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

表 16 Permutation Importance による寄与度（データセット All $D_{\pm 50}$）

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

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．各表からジニ係数による特徴量の重要度と同じく，ラベル付与数（label_num）の寄与度が高いことが分かる．また，学内実験・クラウドソーシング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）の寄与度が高い．これは，4.3 節に示した通り，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 17 に示す．

表 17 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
データセット Crowd $D_{\pm 0}$	0.674	0.694
データセット Crowd $D_{\pm 50}$	0.725	0.744
データセット All $D_{\pm 0}$	0.681	0.700
データセット All $D_{\pm 50}$	0.722	0.747

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. 結論

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，学内実験により得られたデータを用いたモデルでは部分一致を許容するデータセットにおいて 0.738 の Accuracy を示し，不良回答ではラベル付与毎の平均カーソル移動量・クリック回数・文字選択回数が多くなり，分類において重要であることが分かった．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，学内実験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，学内実験とクラウドソーシング実験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．

しかし，実運用をする上で十分な精度は得られていないため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する必要がある．本研究では，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，累積のラベル付与数や直近タスクでのラベル付与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．本研究では，適切回答・不良回答の 2 クラスの分類を行ったが，不良回答の中でも意図的な場合と偶発的な場合を区別して分類し，それぞれに対応した介入方法を考案することも実用に向けて重要な点である．

謝辞

本研究を進めるにあたり，安本慶一教授には，本研究のテーマを通して先端技術の研究に取り組むきっかけを与えていただき，研究全般に関し論文添削や提案手法について多くのご指導・ご助言を賜りました．また，充実した研究環境の提供など，研究活動をご支援いただきました. 感謝の意を表すとともに，心より厚く御礼申しあげます.

荒牧英治教授には，ご多忙の中，論文審査委員を引き受けていただいた上で，副指導教官として本研究に様々なご意見及びご助言をいただきました．感謝の意を表すとともに，心より厚く御礼申し上げます.

諏訪博彦准教授には，本研究を進めるにあたり，実験結果に対するご意見や実験方法へのご助言など，気付けなかった点について多くのご指摘・ご指導をしていただきました．個別での研究に関する相談にも，時間を割いていただき丁寧に回答してくださいました. 感謝の意を表すとともに，心より厚く御礼申しあげます.

松田裕貴助教には，実験用のプログラムを作成する際の技術的な支援，及び実験手段などについて直接相談させていただき多くのご助言を賜りました．感謝の意を表すとともに，心より厚く御礼申しあげます.

金岡恵事務補佐員，山内奈緒事務補佐員には，研究活動において学会や出張に関する事務処理を始め，研究生活の様々な場面でご支援いただきましたこと，謹んで感謝申し上げます.

最後に，本研究を含む大学院での活動において，大学院生活を共にしたユビキタスコンピューティングシステムの先輩，同輩及び後輩の皆様，そして，今日までの学生生活を支えてくれた家族に心より感謝申し上げます.

参考文献

[1] Florian Daniel, Pavel Kucherbaev, Cinzia Cappiello, Boualem Benatallah, and Mohammad Allahbakhsh. Quality control in crowdsourcing: A survey of quality attributes, assessment techniques, and assurance actions. *ACM Comput. Surv.*, Vol. 51, No. 1, 2018.

[2] Ivar R Kolvoort and Leendert van Maanen. 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] Eddy Maddalena, Luis-Daniel Ibáñez, Neal Reeves, and Elena Simperl. Qrowdsmith: Enhancing paid microtask crowdsourcing with gamification and furtherance incentives. *ACM Trans. Intell. Syst. Technol.*, Vol. 14, No. 5, sep 2023.

[4] 三浦麻子, 小林哲郎. オンライン調査における努力の最小限化を検出する技法. 社会心理学研究, Vol. 32, No. 2, 2016.

[5] Daniel M. Oppenheimer, Tom Meyvis, and Nicolas Davidenko. Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, Vol. 45, No. 4, pp. 867–872, 2009.

[6] Michael R. Maniaci and Ronald D. Rogge. Caring about carelessness: Participant inattention and its effects on research. *Journal of Research in Personality*, Vol. 48, pp. 61–83, 2014.

[7] 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.

[8] Masaki Gogami, Yuki Matsuda, Yutaka Arakawa, and Keiichi Yasumoto. Detection of careless responses in online surveys using answering behavior on smartphone. *IEEE Access*, Vol. 9, pp. 53205–53218, 2021.

[9] Takaaki Nakagawa, Yutaka Arakawa, and Yugo Nakamura. Augmented Web Survey with enhanced response UI for Touch-based Psychological State Estimation. In *2022 IEEE 4th Global Conference on Life Sciences and Technologies*, LifeTech, pp. 91–95, 2022.

[10] Mara Naegelin, Raphael P. Weibel, Jasmine I. Kerr, Victor R. Schinazi, Roberto La Marca, Florian von Wangenheim, Christoph Hoelscher, and Andrea Ferrario. An interpretable machine learning approach to multimodal stress detection in a simulated office environment. *J. of Biomedical Informatics*, Vol. 139, No. C, 2023.

[11] Mayu Otani, Riku Togashi, Yuta Nakashima, Esa Rahtu, Janne Heikkilä, and Shin'ichi Satoh. Optimal Correction Cost for Object Detection Evaluation. In *The IEEE/CVF Computer Vision and Pattern Recognition Conference*, CVPR'22, 2022.

[12] Sihang Qiu, Ujwal Gadiraju, and Alessandro Bozzon. Improving Worker Engagement Through Conversational Microtask Crowdsourcing. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI'20, pp. 1–12, 2020.

[13] Jeffrey M Rzeszotarski, Ed Chi, Praveen Paritosh, and Peng Dai. Inserting micro-breaks into crowdsourcing workflows. In *The First AAAI Conference on Human Computation and Crowdsourcing*, HCOMP'13, pp. 62–63, 2013.

[14] Shirasuna Masaru, Rina Kagawa, and Honda Hidehito. 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*, 7 2023.

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

[16] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Advances in Neural Information Processing Systems*, Vol. 30 of *NIPS'17*, 2017.

[17] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *J. Artif. Int. Res.*, Vol. 16, No. 1, p. 321–357, jun 2002.

[18] Hiroki Ouchi, Hiroyuki Shindo, Shoko Wakamiya, Yuki Matsuda, Naoya Inoue, Shohei Higashiyama, Satoshi Nakamura, and Taro Watanabe. Arukikata travelogue dataset. *arXiv*, No. 2305.11444, pp. 1–6, 2023.

[19] Shohei Higashiyama, Hiroki Ouchi, Hiroki Teranishi, Hiroyuki Otomo, Yusuke Ide, Aitaro Yamamoto, Hiroyuki Shindo, Yuki Matsuda, Shoko Wakamiya, Naoya Inoue, Ikuya Yamada, and Taro Watanabe. Arukikata travelogue dataset with geographic entity mention, coreference, and link annotation. *arXiv*, No. 2305.13844, pp. 1–11, 2023.

[20] Carolin Strobl, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC bioinformatics*, Vol. 8, p. 25, 02 2007.

[21] Leo Breiman. Random forests. *Mach. Learn.*, Vol. 45, No. 1, pp. 5–32, October 2001.

[22] Aaron Fisher, Cynthia Rudin, and Francesca Dominici. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously, 2019.

研究業績

本論文に関係のある国内会議

1. 福光嘉伸，松田裕貴，諏訪博彦，安本慶一，マイクロタスク型クラウドソーシングにおける不適切回答のリアルタイム検出・介入手法の検討，2022 年度 情報処理学会関西支部 支部大会 講演論文集，pp.1-6，2022

2. 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），2023-UBI-79，No.18，pp.1-6，2023

3. 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），2024-HCI-206，No.40，pp.1-7，2024

4. 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での固有表現アノテーションにおける不良回答の検出，研究報告行動変容と社会システム（BCSS），2024-BCSS-19，pp.1-8，2024

5. Yoshinobu Fukumitsu，Yuki Matsuda，Hirohiko Suwa，Keiichi Yasumoto，Detecting Careless Responses in Dataset Annotation using Screen Operation Logs，*2024 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom '24 TrustSense)*，2024

本論文に関係のない国内会議

1. 平良繁幸，松田裕貴，福光嘉伸，諏訪博彦，安本慶一会話型インタフェースを用いたオンラインアンケートの回答態度改善手法の検討，2023 年度 情報処理学会関西支部 支部大会 講演論文集，pp.1-4，2023，

受賞

1. 2022 年度 情報処理学会関西支部 支部大会 支部大会奨励賞： 福光 嘉伸：マイクロタスク型クラウドソーシングにおける不適切回答のリアルタイム検出・介入手法の検討