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

Real-time Detection and Intervention to Careless Responses in Micro-task Crowdsourcing

福光 嘉伸 † 松田 裕貴 †,‡ 諏訪 博彦 †,‡ 安本 慶一 †,‡
Yoshinobu Fukumitsu Yuki Matsuda Hirohiko Suwa Keiichi Yasumoto

1. はじめに

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では，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品質に大きなばらつきがあり，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ない．例えば，回答の対価として報酬を付与する場合，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，結果として不適切な回答が発生するという問題がある．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 [1] 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．しかしながら，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 [2, 3] 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制約がある．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適切な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．このことから，リアルタイムに不適切な回答の検知を行い，介入することで行動変容を行う重要性が高いと考える．

本稿では，データのアノテーションタスクを対象として不適切な回答をリアルタイムに検知する手法と，行動変容のための介入方法を提案する．アノテーションにおいて不適切な回答が発生するユーザの状況として，作業内容を見ず意図的に速く回答を行う場合と，注意散漫や疲労により集中が切れた場合が考えられる．そこで，検討する提案手法では，作業中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適切回答検出を行う（適切な回答・意図的な不適切回答・偶発的な不適切回答の 3 分類問題として取り扱う）．さらに，不適切回答の検出結果を基に不適切な回答の発生状況に合わせた介入を行うことで，行動変容を実現する．

2. 関連研究

不適切な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．尾崎ら [4] はアンケートの回答時間や連続同一回答数を特徴量として抽出し，機械学習によって，努力の最小限化の傾向を検出する方法を提案している．また，後上ら [2, 3] 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 [5] 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．

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 [6] 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．

† 奈良先端科学技術大学院大学，
Nara Institute of Science and Technology
‡ 理化学研究所革新知能統合研究センター（AIP），
RIKEN AIP

アノテーションやマイクロタスクに対するユーザのやる気や作業の質を向上させるための研究や，行動変容を促すことを目的とした介入方法に関する研究は多数行われている．Sihang ら [7] 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 [8] 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．Peng ら [9] は，クラウドソーシングのマイクロタスクにおいて，少量の娯楽をタスクの間に提供することで作業の品質を維持しながら，ユーザの定着率を大幅に改善できることを明らかにした．また，大山ら [10] は，参加型センシングを対象として，ボタンのタップと携帯電話を振る行為の 2 種類で貢献の意思表示を行うことで速く回答を行う不適当な行為を抑制する方法を提案した．Zhang ら [11] は，ユーザに対する情報提示によりユーザの歩数の増加を促す行動変容において，提示する情報の対話スタイル（情報の粒度や婉曲表現の度合い）によって行動変容の効果が変化することを明らかにしている．

このように，不適切な回答を検出する既存研究 [2, 3] 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適切な回答の検出を行うことができない制約がある．また，不適切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適切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 [7, 8, 9] 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．これらのことから，リアルタイムに不適切な回答の検知を行い，不適切な回答を改善し，出力結果の質を向上させるように行動変容を行うことのできる介入方法を検討する重要性は高いと考えられる．

3. 提案手法

本研究の目的は，クラウドソーシング上でマイクロタスクを実施する際に，ユーザの不適切回答傾向をリアルタイムに検知するとともに，適した介入方法によってユーザに働きかけることにより，タスクの遂行品質を向上させることである．

3.1 提案手法の概要

提案手法の概要および動作の流れについて図 1 に示すとともに，以下に詳細を述べる．

1. クラウドソーシング依頼者がマイクロタスクを設定し，ユーザを募集する．

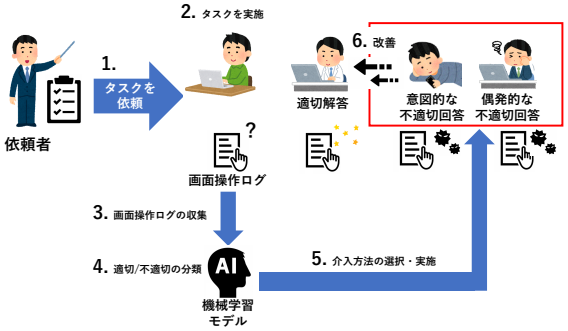

図 1: 提案手法の概要

2. ユーザが提示されたマイクロマスクを実施する．

3. この際，マイクロタスク遂行システム上では，ユーザの作業時の画面操作ログをシステムが定常的に収集する．

4. 得られたログを入力とし，機械学習モデルによって不適切回答傾向を検出する．

5. 不適切回答傾向が検出された場合は，その傾向の種類に応じたユーザへの介入方法を選択・実施することによって，ユーザの行動変容を促す．

6. 最終的に，不適切回答を行っていたユーザはシステムからの介入により，行動を改善し適切な回答を行うようになる．

なお本研究では，不適切回答傾向は 2 つのパターンがあると想定する．1 つ目は，インセンティブを目的とし，一貫して不誠実な態度の回答を行う「意図的な不適切回答」である．2 つ目は，ユーザは適切な回答をしたいと考えているが，注意散漫や疲れなどで集中が切れた状態に起因して発生する「偶発的な不適切回答」である．そこで本研究では，不適切回答傾向の問題を，「適切」・「不適切」の二値分類ではなく，「適切回答」・「意図的な不適切回答」・「偶発的な不適切回答」の 3 クラスの分類として取り扱う．

以降では，想定するマイクロタスクについて整理し，その結果を基にしたシステム設計について述べる．

3.2 想定するマイクロタスク

本研究では，クラウドソーシング上で行うマイクロタスクとして，機械学習における学習データのアノテーションを想定する．検討中のアノテーションタスクとしてバウンディングボックス・ラベルの付与，文書の文字起こし，固有表現ラベリング，感情アノテーションがある．以降では，各アノテーションタスクの作業内容の説明を行う．

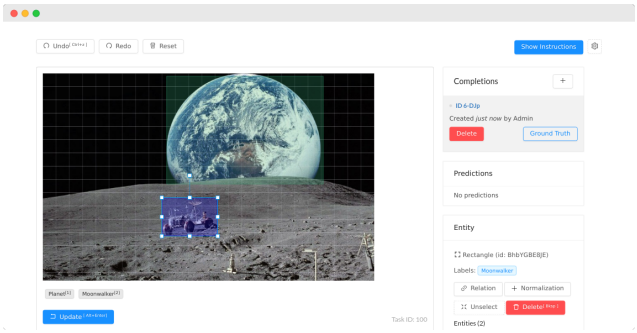

図 2: バウンディングボックス・ラベル付与を行う画面例．Label Studio の Github リポジトリ（`https://github.com/heartexlabs/label-studio`）より引用

画像アノテーション　画像・映像に対してラベルを付与する形態のタスクである．

バウンディングボックス・ラベルの付与：画像解析分野において物体検出・認識モデルを構築するために，画像データに対してバウンディングボックスおよびラベルが付与されたデータセットが必要となる．図 2 に示すように，物体が映っている画像の領域（バウンディングボックス）を選択し，ラベルを付与するタスクである．タスクには，マウスカーソルを移動させることによる領域選択（例：矩形領域をドラッグ操作で描く）や，選択した領域へのラベル付与（例：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，画像を見て何がどこに映っているかを判断する認知処理が必要とされる．不適切な回答では，速く回答を行うために十分な数の対象にラベルを付与していないことや，画像から得られる情報をもとに適切な領域ラベルを付与していないことが考えられる．そのため，適切/不適切回答間で回答時間やカーソル移動量（PC の場合）に違いが出ると考えられる．

文書の文字起こし：文字認識分野において手書き文書や近代文書に対する認識モデルを構築するために，文書の画像データに対してラベルが付与されたデータセットが必要となる．パブリックドメイン OCR 学習用データセット [12] や近代雑誌データセット [13] に含まれるデータのように，文字が映っている画像の領域を選択し，記載の文字をラベルとして付与するタスクである．タスクに必要な処理や，不適切回答が起こる状況はバウンディングボックスタスクと同様であると考えられる．

自然言語アノテーション　文章など，自然言語に対してラベルを付与する形態のタスクである．

固有表現ラベリング：自然言語処理の分野における固有表現抽出モデルを構築するためにデータセットが必要となる．図 3 に示すように，文章中から固有名詞（人名や

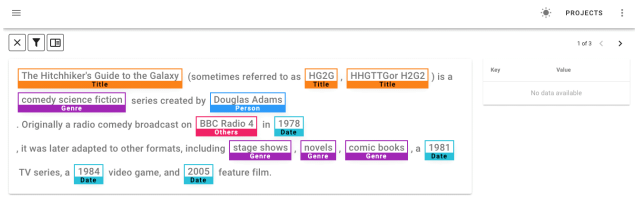

図 3:　固有表現ラベリングの画面例．doccano の Web サイト（`https://doccano.github.io/doccano/tutorial/`）より引用

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．タスクには，マウスカーソルを移動させることによる領域選択（例：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（例：指定されたリストから対応するラベルを選択する）やスクロール操作によって文章を送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．不適切な回答では，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が考えられる．そのため，適切/不適切回答間で回答時間やスクロール速度に違いが出ると考えられる．

感情アノテーション：感情分析モデルを構築するために，文章や単語に対して感情データがラベルとして付与されたデータセットが必要となる．感情分類の代表的なデータセットに chABSA-dataset [14] があり，このデータセットでは各文の感情分類だけでなく，文中のどの単語がネガティブであるかポジティブであるかを表す情報を含んでいる．タスクには，マウスカーソルを移動させて選択肢を選択する画面操作が含まれる．不適切な回答では，文章及び単語を読まずに同一回答を繰り返すことや，文章を理解出来ずに適切なラベルが付与されていないことが考えられる．そのため，適切/不適切回答間で連続同一回答数や回答時間，スクロール速度に違いが出ると考えられる．

3.3　システム設計

提案システムの構成を図 4 に示す．提案システムは，アノテーションタスク時の画面操作ログを取得するシステム，不適切回答を検出（分類）するシステム，分類結果に合わせた介入を行うシステムで構成される．また，本研究ではアノテーションタスクを扱うため，ユーザの端末をスマートフォンに限定せず，PC での作業も対象とする．以降では，タスク実施環境として後上ら [2, 3] でも用いられているオープンソースの調査システム LimeSurvey を用いると仮定し，各システムの詳細について述べる．

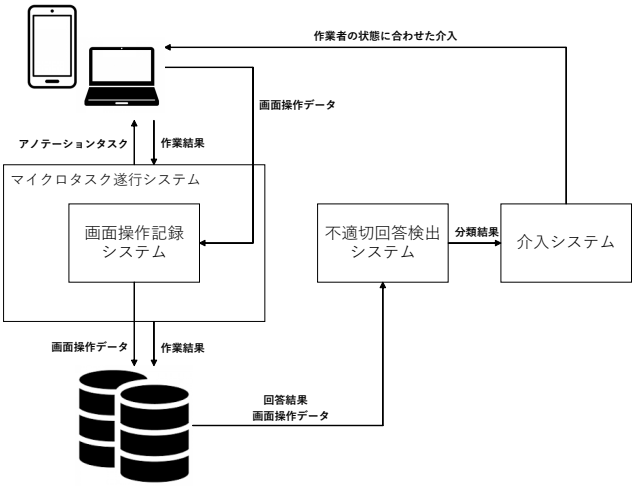

図 4: システム設計図

3.3.1 画面操作記録システム

画面操作が記録可能なシステムの詳細について述べる．画面操作ログ取得には，後上ら [2, 3] が作成した OperationLogger を拡張し用いる．

次に，画面操作データから抽出する特徴量について述べる．本研究では，後上ら [2, 3] が用いている特徴量を拡張し，アノテーションタスクに対応でき，リアルタイムに抽出することのできる特徴量を追加する．さらに，各タスクの実施をトリガーとし特徴量を抽出する．

表 1 に用いる特徴量とその単位を示す．カーソル移動量は，ユーザの使用端末が PC の場合にのみ抽出する特徴量とする．また，一般的なクラウドソーシングツールで抽出できるものは “–”，新たに追加する特徴量のうち後上らの提案したものは “◯”，新たに考案し追加するものは “●” で独自追加の列に示す．新たに追加した特徴量は，対象とするアノテーションタスクの操作内容に合わせて考案されている．作業内容を見ず意図的に速く回答を行う場合では，同じ回答を内容を見ずに繰り返したり速く回答することがあるため，現状の連続同一回答数や問題間の回答時間などの特徴量が有用であると考えられる．注意散漫や疲労により集中が切れた場合では，前問から作業速度が下がることや作業が止まることがあるため，問題間の回答時間の変化や非操作時間などの特徴量が有用であると考えられる．

3.3.2 不適切回答検出システム

アノテーション作業時の不適切回答の検出を行うシステムについて述べる．機械学習による分類を各タスク終了時に行うことで，不適切な回答をリアルタイムで検出する．画面操作データから特徴量を学習させることで分類モデルを作成し，分類を行う．学習データにおいて 3 クラス間でデータ数に違いがある場合は，ダウンサンプリングもしくは機械学習のアルゴリズムを工夫することにより考慮を行う．機械学習のアルゴリズムは，実験を行い取得した特徴量の傾向を基に最適なものを選択する．また，不適切回答をする状態になる前/なった後の特徴量を比較することで，分類に有用な特徴量を選択する．

なお，学習に用いるデータセット構築にあたっては，画面操作データにラベルを付与する必要がある．各ブロックの末尾に「疲れているか」を問う質問を入れることで，疲労状態でのタスク実施（偶発的な不適切回答）であるかのラベル付けを行う.「疲れているか」の質問に対して「いいえ」と答えたかつ，アノテーションタスクで正しい回答を行っていない場合は，意図的に行った不適切回答としてラベル付けを行う．タスクでは常に一定の基準でラベル付けを行うために，一問当たりの認知負荷や与えるタスクの重さが調整可能なアノテーションタスクを選択する．

表 1: 抽出する特徴量

特徴量	単位	独自追加
回答時間	s	–
前問までとの回答時間の変化	%	●
非操作時間	s	●
非操作時間が長すぎる回数	回	◯
テキストの削除回数	回	◯
スクロール長	px	◯
スクロール時間	s	◯
スクロール速度	px/s	◯
前問までとのスクロール長の変化	%	●
前問までとのスクロール時間の変化	%	●
前問までとのスクロール速度の変化	%	●
逆スクロール回数	回	◯
現状の連続同一回答数	問	–
文字数	文字	–
カーソル移動量	px	●
クリック回数	回	●
クリック間隔	s	●
前問までとのカーソル移動量の変化	%	●
前問までとのクリック回数の変化	%	●
前問までとのクリック間隔の変化	%	●

–： 一般的な特徴量，◯： 後上ら [2, 3] による特徴量，
●： 新規に追加する特徴量

表 2: 分類結果に対応する介入方法

分類結果	介入方法
適切回答	-
意図的な不適切回答	ポップアップ表示
	アイコン表示
偶発的な不適切回答	短時間の休憩
	少量の娯楽

3.3.3 行動変容のための介入システム

分類モデルの結果を基にタスク実施中のユーザに行動変容を促すシステムについて述べる．行動変容の方法として画面表示や作業提示による介入を行う．不適切回答が起こる状態それぞれに対応した介入を行うことで，適切な回答をするよう行動変容を促すことが可能であると考えられる．表 2 にそれぞれの分類結果に対応する介入方法を示す．

ポップアップ表示では，注意喚起や大山ら [10] の貢献の意思表示を作業途中で行い，速く回答を行う不適当な行為を抑制することを検討している．アイコン表示では，目のアイコンを表示することでユーザに監視されていることを明示せずに知らせることで行動変容を促す．また，短時間の休憩や少量の娯楽がユーザの定着率を大幅に改善でき，作業に対する関与を高めることは既存研究 [8, 9] で明らかになっているため，タスク間にこれらをシステムが提示し，ユーザに与えることで行動変容を促す．

4. 実験概要・計画

本章では，実際に行う実験の概要や計画について示す．本稿の手法は，リアルタイムの介入を行うことでユーザのタスク中の行動変容を目的として提案される．以下に，目的を達成するための実験概要を示す．

(1) 提案手法の不適切回答の検出・分類を行う分類モデルを作成するために，アノテーションタスクの各設問を作成し，学習データを取得する実験を行う．タスクは複数ブロックに分割された形式で実施し，全て答えが既知の問題を扱う．また，ブロック毎に問題数を変更することで，問題数による疲労や作業内容を見ず意図的に速く回答を行う行為への影響も調査する．

- 自然言語処理向けのオープンソースのアノテーションツールには，doccano [15] がある．doccano が対象とするタスクには，感情分析等のテキスト分類，系列ラベリング，要約や翻訳などの系列変換があるため，システムを組み込むことで提案手法をツール上のタスクに適用できると考えられる．
- バウンディングボックス・ラベルの付与を行うことのできるオープンソースのアノテーションツールには，Label Studio [16] がある．doccano 同様，システムを組み込むことで提案手法をツール上のアノテーションタスクに適用できると考えられる．

(2) 3.2 節に示したアノテーションタスクをクラウドソーシング上でユーザに与え，ラベル付けされた画面操作データを 3.3.1 節のシステムによりリアルタイムに取得し，特徴量を抽出する．このタスクの実施により，適切回答・意図的な不適切回答・偶発的な不適切回答の 3 クラスそれぞれの作業結果とその際の画面操作記録のデータを取得することができると考えられる．また，アノテーションタスクはブロック毎に問題数を変更することで，問題数による疲労や意図的に速く回答を行う行為への影響も調査する．

(3) 3 クラス間の特徴量の傾向の違いを分析することで，学習に用いる特徴量とデータの特性に適切なアルゴリズムを選択し，分類モデルの作成を行う．また，アノテーションタスクの種類による特徴量の違いを分析し，必要であれば特徴量の補正を行う．分類モデルの精度評価には，Accuracy，Precision，Recall，F1 Score を用い，汎化性能の検証は，10-fold cross validation によって行う．不適切回答の検出方法としては，1 つタスクが終わるごとに分類を行うことで逐次的な検知を行う．分類にかかる実行時間として，次のタスクが終わるまでに検出が出来ることを要件とする．

(4) 分類モデル作成後に，介入によってどの程度の確率で不適切回答が改善されるかの実験を行い，行動変容を促す介入の手法について，効果を測定する．

5. まとめ

本研究はクラウドソーシングにおけるマイクロタスク，特に機械学習のためのアノテーションタスクを対象とし，ユーザが取る不適切な回答をリアルアイムに検知する手法と，介入によって行動変容を促す方法を組み合わせることで回答の品質を向上するためのシステムを提案した．提案手法で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に不適切な回答を検出できるよう検討を行った．また，アノテーションタスクにおける不適切回答として，意図的な不適切回答・偶発的な不適切回答の 2 種類があると定義し，それらについての介入方法について整理した．

今後の予定としては，各システムおよびタスクの作成を行い実際のクラウドワーカーを対象とした実験を実施し，提案手法の有効性を検証することが挙げられる．

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] 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.

[2] 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.

[3] 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.

[4] 尾崎幸謙, 鈴木貴士. 機械学習による不適切回答者の予測. 行動計量学, Vol. 46, No. 2, pp. 39–52, 2019.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] Peng Dai, Jeffrey M Rzeszotarski, Praveen Paritosh, and Ed H Chi. And Now for Something Completely Different: Improving Crowdsourcing Workflows with Micro-Diversions. In *Proceeding of The 18th ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW'15, pp. 628–638, 2015.

[10] Kohei Oyama, Yuki Matsuda, Rio Yoshikawa, Yugo Nakamura, Hirohiko Suwa, and Keiichi Yasumoto. A Method for Expressing Intention for Suppressing Careless Responses in Participatory Sensing. In *18th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*, MobiQuitous'21.

[11] Zhihua Zhang, Juliana Miehle, Yuki Matsuda, Manato Fujimoto, Yutaka Arakawa, Keiichi Yasumoto, and Wolfgang Minker. Exploring the Impacts of Elaborateness and Indirectness in a Behavior Change Support System. *IEEE Access*, Vol. 9, pp. 74778–74788, 2021.

[12] 国立国会図書館. パブリックドメイン ocr 学習用データセット（令和 3 年度 ocr テキスト化事業分）, 2022. https://github.com/ndl-lab/pdmocrdataset-part1.

[13] 人文学オープンデータ共同利用センター. 近代雑誌データセット, 2017. http://codh.rois.ac.jp/modern-magazine/.

[14] Takahiro Kubo and Hiroki Nakayama. chABSA-dataset, 2018. https://github.com/chakki-works/chABSA-dataset.

[15] Hiroki Nakayama. doccano, 2018. https://github.com/doccano/doccano.

[16] Hartex. Label studio, 2019. https://github.com/heartexlabs/label-studio.