DEIM Forum 2021 I24-5

クラウドソーシングにおける作業依頼者の意図を適切に反映した作業指示の選定手法の提案

藤浦　礼奈† 　　鈴木　　優†

† 岐阜大学工学部電気電子・情報工学科　〒 501–1193 岐阜県岐阜市柳戸 1 番 1
E-mail: †w3033127@edu.gifu-u.ac.jp, ††ysuzuki@gifu-u.ac.jp

あらまし　我々は機械学習における学習データを作成するときに，手間を軽減するため，もしくは多数の意見を反映させるためにクラウドソーシングを利用する．作業を依頼するときに作業依頼者は曖昧な指示を提示すると，指示を理解できる作業者とできない作業者を発生させる可能性がある．これにより，データに対する作業者の考え方が一致しないことに繋がり，作業依頼者の望む結果が得られないことがある．そのため作業依頼者の望む結果を得るためには，曖昧でない適切な作業指示を加えることが効果的なのではと考えた．ところが，どの作業指示が適切であるかを選択することは難しい．そこで我々は適切な作業指示を判明させるために，作業者から得られるデータのうち，投票人数が割れているデータに着目した．適切だと思われる作業指示の候補をもとに作業を依頼し，投票人数が割れているデータ数を比較することにより，作業指示が適切に伝わっているかを判定する．これにより，適切な作業指示を決める．またこの手法は実際に作業を行う必要があるため，費用や手間などのコストがかかってしまう．そのため我々は，適切な作業指示を決定するために必要な作業結果を，作業者ごとと作業指示ごとに BERT を用いて予測する．これにより，コストをかけずに適切な作業指示を決定することを実現する．

キーワード　クラウドソーシング，機械学習

1　はじめに

クラウドソーシングは，目的のデータを収集する際に複数の作業者から容易にデータを収集することができるという利点から，様々な分野で活用されている．ところが，必ずしも作業依頼者の想定通りに作業がされるとは限らない．作業依頼者は作業を依頼する際に,「直感でコロナウィルスに関係あると思えば 1 を選択してください.」などの作業指示を提示する．また択一式の作業を行うとき，様々な意見を取り入れるために，複数の作業者に作業を依頼し得られたデータから，多数決により作業結果を決めることがある．多数決を取ることによって得られた結果は，作業者の意図した結果と意図していない結果の二つに分けることができる．作業者の意図した結果が得られる要因は，作業指示が作業者に適切に伝わったためである．また，作業者の意図していない結果は二つに分けることができる．一つ目は，作業者の大半が作業指示を誤解して作業を行った結果である．これは，作業指示の内容が誤っていたために起こる．二つ目は，作業指示を適切に理解した人と，誤解した人が混在して作業を行った結果である．これは，提示した作業指示が作業者にとって曖昧であったために起こる．そのため作業依頼者は望むデータを得るために，曖昧と判断されにくい適切な作業指示を明らかにしようとする．

しかし，適切な作業指示を明確にすることは簡単ではない．なぜなら作業依頼者は，作業者が作業について悩む要因を把握することが難しいためである．

我々はこれらの問題を解決するために，作業依頼者と作業者が相互に作用することが効果的であると考えた．これにはまず，作業者から得られたデータのうち投票人数が割れているデータから，作業指示の候補を決める必要がある．それぞれの作業指示で作業を行い，実際に作業者から得られた結果から投票人数が割れているデータ数が少ない方の指示を，適切な作業指示とする．これにより，作業者が指示に対して誤解した要因を考慮した作業指示の明確化が可能となる．

しかしこの手法は，実際に作業を行い適切な作業指示を決める必要があるため，費用や手間などの膨大なコストがかかる．そこで我々は，機械学習を用いることにより作業結果を予測することが，コスト削減に繋がるのではと考えた．これは，作業者ごとと作業指示の候補ごとに得られた作業結果を基にモデルを作成することにより実現する．これにより，人手で大量に作業を行わなくても，実際に人手で十分な作業量を行った場合と同等の結果を得ることができる．

評価実験では，以下の二つの項目について確かめた．一つ目は，作業指示を追加する前と比べて，投票人数が割れている数が減っているかどうかである．これは指示を追加することで，作業指示を理解している人と，誤解している人が混在するデータ数を減らすことができるのかを確かめるためである．二つ目は，得られた結果のラベルの訂正必要数が少ないかどうかである．これは，作業者の意図が正確に伝わったかどうかを明らかにするためである．また，機械学習において予測した結果も上記の 2 点を検証する．

以上の項目を確かめた結果，定めた適切な作業指示で作業を行うことで，投票人数が割れているデータ数を 100 件中 94 件減らすことができた．また，ラベルの訂正必要数は 22 件減ら

すことができた．

機械学習による作業結果の予測においても，実際に人手で作業を行った場合と同じ作業指示を選定することができた．

本研究の貢献は以下の二つである．

- 作業者に対する適切な指示は，作業者による投票によって決めることができることを明らかにしたこと．
- 機械学習を用いて作業者の付与するラベルを予測することにより，作業依頼のコストを抑えることができることを明らかにしたこと．

2 関 連 研 究

クラウドソーシングを行うにあたり詳細な作業指示を定めていないと，作業依頼者と作業者の間に誤解が生まれてしまうため，あまり良い結果を得ることができない [1]．丹治ら [2] は，作業者から得られた作業に対する判断基準をクラスタリングすることにより，作業指示決定のための適切な判断基準を収集する手法を提案している．より良い作業指示で作業を行うために，作業者の意図を反映する点は似ている．しかし本研究では，作業者の作業に対する判断基準を得ることはせずに，ラベルの間違いが一番多いデータの特徴から作業指示を導き出している点が異なる．

Aniket らの研究 [3] では高品質な作業結果を得るために，事前に作業者に対して作業に関する質問をすることで，作業結果の品質を向上させることができることが明らかになっている．この研究より，作業者に提示する指示内容は作業結果の品質に大きく関わっていると言える．

クラウドソーシングと機械学習を組み合わせた研究も存在する．Steve らの研究 [4] では，人手を機械学習の過程に組み込むことで，機械学習の精度を向上させることができると報告されている．これは，人間の感覚でしか判断することができない特徴と，機械学習がデータからパターンを見つけることができるという二つの特徴を組み合わせることによって，どちらかだけを行った場合では得られることのできない判断基準を得られるためである．

また，Amazon が提供しているクラウドソーシングサービスである Amazon Mechanical Turk[1]を使用した Vamsi らの研究 [5] では，クラウドソーシングを行う過程に機械学習を用いる研究を行っている．この研究では，クラウドソーシングをより効果的に行うために，作業者の能力や興味などの特徴から，その作業者に合う作業をモデルに基づいて推奨している．この研究により得られた作業結果の品質は，繰り返しのラベル付けから得られる結果の品質と近しい結果が得られることが判明している．

クラウドソーシングにおけるコストを削減するために，山下らの研究 [6] では，Dawid らの研究 [7] を応用した手法を採用している．Dawid らの手法は，医師の複数の診断結果を用いて，正しい診断行うために EM アルゴリズムを採用している．EM アルゴリズムは，正解と作業者の能力を交互に予測することができるが，より良い結果を得るためには十分な作業量が必要不

1：https://www.mturk.com/

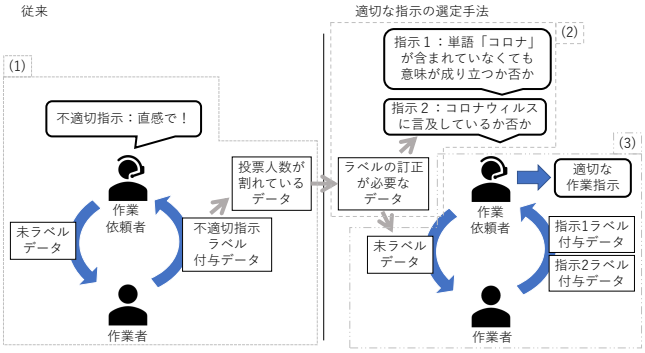

図 1　作業指示の選定手法の流れ

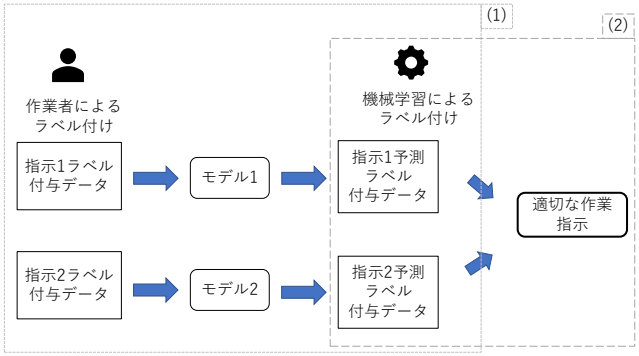

図 2　コスト削減手法の流れ

可欠である．そのため山下らは，クラウドソーシングにおけるコストを削減するために動的な EM アルゴリズムを採用し，作業者と同じように作業をする AI にも作業を行わせた．これにより，作業者と AI が判断しやすいタスクにおいて，高い正解率を導くためのコストの削減を実現している．

3 提 案 手 法

作業依頼者は，全ての作業者の意図を把握することは難しい．また，作業依頼者の望む結果を得るために，どの作業指示を加えたら作業者が迷うことなく作業を行うかを明らかにすることは容易ではない．

そこで我々は，作業者から得られたデータにより作業指示を決める．これにより作業者が作業に対して迷っている原因を明らかにしたうえで，作業の指示をすることができる．これを実現するために，図 1 と以下の手法 1 において適切な作業指示の選定手法の流れを示す．

[手法 1 – (1)]　クラウドソーシングにより得られたデータから，投票人数が割れているデータを取得する．3.1.1 節で説明する．

[手法 1 – (2)]　投票人数が割れているデータの中から，ラベルの訂正必要数とその特徴を明らかにする．また，再度作業を依頼したときにラベルが訂正されると思われる作業指示の候補を，明らかにした特徴をもとにいくつか定める．3.1.2 節で説明する．

[手法 1 – (3)] [手法 1 – (1)] と同じ条件のもとで集めた新たなデータを用いて，[手法 1 – (2)] で定めた作業指示の候補により作業を行う．作業結果より，投票人数が割れているデータ数が少ない方を適切な作業指示とする．3.1.3 節で説明する．

[手法 1 – (4)] [手法 1 – (3)] で明らかにした作業指示をもとに，再度 [手法 1 – (1)] のデータに対してクラウドソーシングを行う．得られた作業結果のうち投票人数が割れているデータ数とラベルの訂正必要数を明らかにし，[手法 1 – (2)] における値と比較する．3.1.4 節で説明する．

今回提案する手法 1 は，適切な作業指示を決めるために実際に作業を行う必要がある．そのため，費用や手間などの膨大なコストがかかってしまう．

そこで我々は，機械学習を用いて作業結果を予測することで，コストを抑えながら適切な作業指示を明らかにする．これを実現するために，図 2 と以下の手法 2 において機械学習を用いて適切な作業指示を決める際に，コストを削減するための手法の流れを示す．

[手法 2 – (1)] [手法 1 – (3)] で収集した作業結果を用いて，作業者と作業指示ごとに作業の傾向を学習させたモデルを作成する．ラベルが付与されていないデータに対して，作成したモデルによりラベルの予測をする．3.2.1 節で説明する．

[手法 2 – (2)] [手法 2 – (1)] でモデルにより予測したラベルから，投票人数が割れているデータ数が少ない作業指示を明らかにする．これにより，少ないコストで適切な作業指示を決定する．3.2.2 節で説明する．

3.1 適切な作業指示の選定手法

手法 1 では，作業者から得られたデータのうち投票人数が割れているデータをもとに作業指示を決定する．そのため，作業者が意図していない結果が得られる要因である作業指示の曖昧さを考慮して作業指示を決定することができる．

3.1.1 投票人数が割れているデータの収集

クラウドソーシングにより得られたデータのうち，投票人数が割れているデータをランダムに N_A 件取得する．これをデータ A とする．以下，取得するデータについては，データに付与されているラベルごとのデータ数が同数となるように取得する．つまり，取得したデータに付与されているラベルが二つの場合，各ラベル $N_A/2$ 件ずつ取得する．

本研究において投票人数が割れているかどうかは，(1) 式のように定義する．

$$v(T_i) = (m_1(T_i) - m_2(T_i)) - \frac{m_1(T_i)}{2} \tag{1}$$

T_i は i 番目のデータ ($i = 1, 2, \cdots, n$)，$v(T_i)$ は T_i における曖昧さ，$m_1(T_i)$ は T_i に投票可能なラベルのうち得票数が最も多いラベルの投票人数，$m_2(T_i)$ は T_i に投票可能なラベルのうち得票数が 2 番目に多いラベルの投票人数を表している．また，$v(T_i)$ における閾値を θ とし，$v(T_i)$ が θ 以下の値であれば投票人数が割れているデータ，θ より大きい値であれば，投票人数が割れていないデータと定義する．

つまり，3 種類のラベルが付与されているデータに対して，ラベル 1 に投票した人数が 6 人，ラベル 2 に投票した人数が 1 人，ラベル 3 に投票した人数が 0 人の場合，曖昧さの値は，$(6-1)-6/2=2$ となる．閾値を 0 とした場合，$v(T_i)$ は閾値 θ より大きい値のため，このデータは投票人数が割れていないデータと判定することができる．

3.1.2 作業指示の設定

データ A のうち，ラベルの訂正必要数とそのいくつかの特徴を明らかにする．以下本論文では，特徴という言葉は，言葉遊びや広告など，そのデータがどのようなものかを表す人間の感覚における特徴のことを指している．言葉遊びは「コロナでイチコロナ」，広告は「【緊急】マスク入荷しました！http...」などといったものである．ラベルの訂正が必要なデータの特徴の中から，データ数が最も多い特徴に着目する．その特徴をもとに，再度作業を行った際に正しいラベルが付与されるであろう適切な作業指示の候補を W 個定める．今回はどのような作業指示が効果的なのかを判明させるために，データ数が一番多い特徴のみに着目した．しかし，一番多い特徴のみに着目をして得られたデータが良い成果を得ることができたら，今回指定した作業指示を参考にして，今後は他の特徴に属するデータにも作業を行う．

例えば，データ A にラベルの訂正をしなければならないデータが 10 件あり，この 10 件のデータを，特徴 1 に属するデータに 3 件，特徴 2 に属するデータに 7 件分類できる場合，データ数の多い特徴 2 のみに着目をする．特徴 2 が「文章として成り立っていない」という特徴であったとする．特徴 2 に属するデータのラベルを訂正したい場合には,「コロナという単語が含まれていないと文章の意味が成り立たない場合は，関係あるのラベルを付与する.」など，着目する特徴に応じた作業指示の候補を定める．このとき，作業指示の候補の個数は 2 個か 3 個にすると良い．これは，作業指示を増やせば増やすほど 1 度にかかるコストが高くなってしまうことと，どの作業指示が効果的かを見逃してしまう可能性があるためである．また，一つの作業指示に対して，特徴に対する改善点を複数記入すると，作業者が作業指示を理解できない可能性が出てきてしまう．そのため，一つの作業指示につき，一つの改善点を指定することを推奨する．

3.1.3 作業指示の提示

3.1.3 節では，適切な作業指示を決定する際に AB テストを用いる．AB テストとはマーケティング戦略手法の一つであり，Web ページ上のオブジェクトの配置や表示の仕方が異なるパターン A とパターン B から，一定期間内においてどちらの方がより良い成果を得ることができるのかを測定するテストである．主に，Web ページや広告などに用いられることが多い．実

際に良い成果を得ることのできたパターンを Web ページに適用した結果，更なる利益を得ることができたという事例がある．そこで本研究では，適切な作業指示を指定するときに，AB テストに倣い，いくつかの指示パターンでの作業を実施する．作業結果より，一番良い成果が得られるパターンを導き，実行するという手法を提案する．

3.1.1 節でデータを収集したときと同じ条件のもと，新たに投票人数が割れているデータを N_B 件取得する．これをデータ B とする．なお，作業にかかるコストを考慮して，どの作業指示が良いかを判断できるだけのデータの件数が用意できれば，新たに収集するデータ数の N_B は必ずしも N_A と同じ数である必要は無い．作業者に対して W 個の作業指示とデータ B を提示し，各作業指示に応じて W 回データに対するラベル付けを依頼する．また作業依頼者はデータ B に対して，作業者に提示した作業指示を考慮せずに，主観で正しいと思うラベルを付与する．作業者と作業指示ごとに得られたデータ B のラベルを作業指示ごとにまとめ，投票人数が割れているデータが少ない方の作業指示を，適切な作業指示として定義する．

例えば定めた作業指示の個数が二つだった場合，データ B に対して，それぞれの作業指示に応じて 2 回ラベル付けを依頼する．このとき，作業者は同一のデータに対して 2 回作業を行なっているため，二つの作業指示を基準としてラベル付けが行われたデータが二つ得られる．二つの作業指示ごとに作業者の結果をまとめ,「コロナという単語が含まれていないと文章の意味が成り立たない場合は，関係あるのラベルを付与する.」という作業指示で得られた結果の方が投票人数が割れているデータが少ないとすると，この作業指示を適切な作業指示として定義する．

3.1.4 新たな作業指示の適用

データ A と 3.1.3 節で得られた適切な作業指示を作業者に提示し，再度ラベル付けを依頼する．このとき得られたデータをデータ A' とする．それぞれの作業者から得られたデータ A' のラベルをまとめ，投票人数が割れているデータ数が，作業指示を適切に決めていないときと比較して減っているかどうか確かめる．また，データ A' をまとめたラベルと 3.1.2 節で得られたラベルの訂正必要数を比較する．さらに，3.1.2 節にて作業指示の候補を定める際に着目した特徴に属するデータのラベルの訂正必要数が減っているかどうかを確認する．

例えば，データ A と 3.1.3 節の例で定めた適切な作業指示である「コロナという単語が含まれていないと文章の意味が成り立たない場合は，関係あるのラベルを付与する.」という作業指示を作業者に提示したとする．作業者から得られた結果 A' が，作業指示を適切に決めていないときと比較して投票人数が割れているデータ数が減っていたもしくは，ラベルの訂正必要数が減っていれば，作業指示を適切に伝えることができたと言える．

3.2 機械学習によるコスト削減の手法

我々は今回提案した適切な作業指示を選定するための手法において，実際に作業者から得られた結果をもとに定める手法を提案した．しかしこの選定手法は，実際に作業を行わなければならないため膨大なコストがかかってしまう．そのためコストを抑えつつ適切な作業指示を選定するためには，機械学習で作業結果を予測することがコスト削減に繋がるのではないかと考えた．

今回はモデル作成時に用いるデータ数に制限を設けていない．しかし人手でラベル付けを行うデータ数を極端に少なくすると，モデルの精度や汎用性が低くなるので，モデルを作成できるだけのデータ数は確保しなければならない．

3.2.1 モデルの作成とラベルの予測

3.1.3 節において得られたデータから，作業者と作業指示ごとに作業結果を収集する．収集したデータから，BERT を用いて作業者の作業の傾向を反映させたモデルを作成する．作成したモデルにより，ラベルが付与されていないデータに対して作業者が付与するであろうラベルを推定する．

今回の実験ではモデルを作成するにあたり，ファインチューニングが行えることと，自然言語処理のタスクにおいて優れた結果を得られるという利点より BERT を用いる．日本語でモデルを作成するにあたり，東北大学の乾・鈴木研究室が公開している訓練済み日本語 BERT モデル [2]の bert-base-japanese-whole-word-masking を使用した．

また，適切なエポック数を決定するために early stopping を用いた．今回は，10 エポック前の検証データの誤差がそれ以降増え続けていたら学習を終了させるという設定で行った．

検証データは，モデルに使用するデータのうち，データ数が少ないラベルに属するデータの約 10 分の 1 件をそれぞれのラベルに属するデータから抜き出し，作成する．訓練データは，モデルに使用するデータのうち，検証データに用いるデータの個数分削除したものから作成する．このとき，データ数が少ないラベルに属するデータは，データ数が多いラベルに属するデータと同数になるまで複製し，それぞれのデータを合算したものとする．これは，訓練用に用いるデータ数に偏りがあると，テストデータはデータ数の多い方のラベルに判定されやすくなってしまうためである．

例えば，作業者 A が作業指示 1 をもとに行った作業結果が 500 件得られたとする．これを，作業者と作業指示の個数分収集する．以下，作業者 A の作業指示 1 におけるデータには，関係あるというラベルに 100 件，関係ないというラベルに 400 件のデータが存在する場合を想定する．まず，作業者 A の作業指示 1 で取得した 500 件のデータのうち，学習時のエポック数を決めるための検証データを集める．データ数の少ないラベルである関係あるに属するデータを基準に，10 分の 1 である 10 件を取得する．関係ないのラベルのデータも同様に 10 件取得する．検証データとして用いていないデータのうち，データ数が少ないラベルである関係あるのデータ 90 件を，データ数が多いラベルである関係ないのデータ数の 390 件と同数となるように複製する．関係あると関係ないのデータ 390 件ずつ，合計 780 件を訓練データとする．訓練データから，BERT を用いてファインチューニングを行いモデルを作成する．ラベルが付与

2：https://github.com/cl-tohoku/bert-japanese

されていないデータに対して，作成したモデルによりラベルを予測する．これを作業者と作業指示の個数回分行う．

3.2.2 予測ラベルと実際のラベルの比較

3.2.1 節で作成したモデルにより予測されたラベルを多数決により決定する．多数決により決定したラベルのうち，投票人数が割れているデータ数が少ない方の作業指示を適切な作業指示とする．また，作業依頼者が正しいと思うラベルとの一致数を比較する．

例えば，作業指示 1 と作業指示 2 における予測ラベルの多数決の結果の投票人数が割れているデータ数が 50 件と 100 件であったとする．この場合，作業指示 1 の方が，投票人数が割れているツイート数が少ないため，作業指示 1 を適切な作業指示とする．

4 評価実験

4 章では，作業者から得られた結果を用いて，適切な作業指示を決定するための実験を行った．また，適切な作業指示を決める際にかかるコストを減らすために，機械学習を用いて作業傾向を学習させたモデルを作成した．これにより，作業者と作業指示ごとに作業結果を予測することで，実際に人手で作業した場合と同等の結果を得ることができるのかを検証した．

今回行った実験は，通常クラウドソーシングで行うような不特定多数の作業者に作業を依頼するのではなく，第一著者が所属している研究室のメンバーに作業を依頼した．そのため，作業の品質に悪影響を与える怠惰な作業者については考慮していない．

4.1 使用するデータセットの詳細

評価実験では，我々が公開している COVID-19 日本語 Twitter データセット [3]を用いる．このデータセットは，クラウドソーシングにて収集した．ツイートの収集期間は 2020 年 1 月 9 日から 2020 年 7 月 6 日までで,「COVID」または「コロナ」を含む合計 135,082 件のツイート ID とラベルとラベルに対応する投票人数が明記されている．なお，このデータセットを使用するには，TwitterAPI などを用いてツイート ID をツイートに変換する必要がある．

このデータセットを作成するにあたり，まずコロナウィルスに関係あるか，関係ないか，判定不可能かを問いている．関係あるを選択した場合，一般事実か，個人事実か，意見・感想か，これらのどれにも属さないかを問いている．一般事実は，ニュースなど一般的に公表されている情報を指す．個人事実は，ツイートの発信者の個人的な情報など一般的に公表されていない情報を指している．投票に参加した人数は，一つのツイートに対して 5 人から 10 人であり，過半数の投票数を得られたラベルがそのツイートのラベルとなる．また，曖昧さの定義である (1) 式を元に閾値 θ を 0 として得ることができたツイート数は，投票人数が割れていないツイートが 126,241 件，割れているツイートは 8,841 件であった．

3：http://www.db.info.gifu-u.ac.jp/data/Data_5f02db873363f976fce930d1

表 1　使用するデータセットのラベルの訂正必要数

		ラベル	
		関係ある	関係ない
投票人数	割れていない	9/50 (18%)	2/50 (4%)
	割れている	35/50 (70%)	1/50 (2%)

4.2 投票人数が割れているツイートと割れていないツイートについて

我々の公開しているデータセットを用いて，投票人数が割れているツイートと割れていないツイートに分ける．その中から，関係あるのラベルのツイートと関係ないのラベルのツイートをランダムで 50 件ずつ，合計 100 件取得する．その後，第一著者の主観によりそれぞれのツイートに対して正しいと思うラベルを付与し，データセットのラベルと比較する．

第一著者の正しいと思うラベルの基準は以下のように定めた．関係あるのラベルは，データセット作成時の基準に加え，コロナウィルスがまん延していたときに起きていた事象が分かるようなもの．関係ないのラベルはそれ以外のものとした．

第一著者が正しいと思うラベルと比較したときのラベルの訂正必要数は，それぞれ表 1 のようになった．表 1 より，投票人数が割れているツイートの方が，ラベルの訂正必要数が多いことが分かる．これは，データセット作成時に提示した作業指示が適切でなかったために，指示内容を理解できる人とできない人の結果が混在してしまったためである．また，投票人数が割れておらず関係あるのラベルが付与されているツイートのうち，ラベルの訂正必要数は 9 件あった．しかし 9 件のうち 5 件は，広告や診断メーカーなど，データセットから排除をすることが比較的容易なものであった．これらの結果より，1 節で述べたように，作業依頼者の望まないデータは主に投票人数が割れているデータに多く属していることが分かる．

4.3 実　験　1

実験 1 では，作業者から得られた作業結果から適切な作業指示を選定する実験を行った．

4.3.1 投票人数が割れているツイートの収集

我々の公開しているデータセット を用いて，投票人数が割れているツイートを収集する．(1) 式を元に閾値 θ を 0 とし，投票人数が割れているツイートの中から，コロナウィルスに関係あるツイートと関係ないツイートをランダムに 50 件ずつ，合計 100 件収集する．以下これをツイート群 A とする．

A に対して，第一著者の主観により正しいと思うラベルを付与する．A に元々付与されていたラベルと新たに第一著者が正しいと思うラベルを比較して，ラベルの訂正が必要なツイートの数を明らかにする．

4.3.2 作業指示の設定

A の中から，ラベルの訂正をする必要があるツイートの特徴を明らかにする．表 2 にラベルを訂正する必要があるツイートの特徴と件数とツイート例を示す．表 2 において，ツイート数が最も多い特徴であるコロナウィルスに言及していないに着目する．このツイートに対して，クラウドソーシングを行った際

表 2　投票人数が割れているツイートのうち，正しいラベルが付与されていないツイートの特徴

特徴	件数	ツイート例
コロナウィルスに言及していない	15	咳しまくってたら「コロナか !?」って心配された
言葉遊び	5	コロナでイチコロナつって wwww
質問箱など	4	みんなからの匿名質問を募集中！こんな質問に答えてるよ ····
意味が理解できない	4	コロナは止まってもいいけど、私はとまりませ〜ん。
タグのみ	3	#コロナに負けるな #がんばろう#covid-19
広告	2	【#マスク 入荷情報】マスクが入荷しました！今だけお得！#洗えるマスク
コロナ+感情のみ	1	コロナほんとうざいしね
自然言語処理で扱えない	1	おはようございます、今日は寒いですね。コロナも ··· https://

に正しいラベルが付与されるであろう二つの作業指示を以下のように定める．

作業指示 1　作業者の直感に加え,「コロナ」もしくは「COVID-19」という単語が無いと文章として意味が成り立たないものを関係ある．成り立つものを関係ないとする．

作業指示 2　作業者の直感に加え，コロナウィルスについて言及しているものを関係ある．言及していないものを関係ないとする．

4.3.3　作業指示の提示

A と同じ条件で，新たにコロナウィルスに関係あるツイートと関係ないツイートを 50 件ずつ，合計 100 件収集する．以下ツイート群 *B* とする．

4 名の作業者に対して定めた二つの作業指示と *B* を提示し，定めた作業指示の個数回分，つまり合計 2 回 *B* に対してラベル付けを依頼する．4 名の投票結果から多数決により決定したラベルの投票人数が割れているツイート数が少ない方を，適切な作業指示とする．

4.3.4　新たな作業指示の適用

4.3.3 節により判明した適切な作業指示と *A* を再度 4 名の作業者に提示し，*A* に対して新たにラベル付けを依頼する．得られた結果を *A′* とする．*A′* から多数決により決定したラベルのうち，*A* のときよりも投票人数が割れているツイート数が減っているかどうかと，ラベルの訂正必要数が減っているかどうかを調べる．

4.3.5　実験結果と考察

今回の実験は作業者が 4 名でラベルの選択項目が二つであったため，作業者全員の結果から多数決をとる際に各ラベルの投票人数が 2 名になっているツイートについては，ラベルを決定することができない．そのため第一著者が正しいと思うラベルと比較をする場合，投票人数が割れているツイートは無視をしている．また，今回は投票人数が割れているツイートの定義は (1) 式ではなく，ラベルを決定することができないツイートとした．

適切な作業指示を決定するために，作業指示 1 と作業指示 2 で作業を行った場合どちらが第一著者の望む結果を得ることができるかを明らかにした．結果を表 3 に示す．表 3 上段は作業者から得られた結果から多数決により決定したラベルのうち，

表 3　作業指示の決定

	作業指示 1	作業指示 2
投票人数が割れている数	7/100 (7%)	22/100 (22%)
ラベルの訂正必要数	16/93 (17%)	13/78 (17%)

表 4　ラベルの訂正必要数

		ラベル	
		関係ある	関係ない
投票人数	割れている	13/46 (28%)	1/48 (2%)

投票人数が割れているツイート数を表している．下段は第一著者が正しいと思うラベルと比較したときのラベルの訂正必要数を表している．

表 3 上段より，作業指示 1 の方が投票人数が割れているツイート数が少ないことが分かる．そのため作業指示 2 よりも作業指示 1 の方が，指示を理解できる作業者とできない作業者が混在しておらず，良い指示であると言える．また表 3 下段より，ラベルの訂正必要数は作業指示 1 の方が少々多かったものの，正しいラベルとの一致数は作業指示 1 の方が 12 件多かった．これにより，作業指示 1 の方が作業者の意図が適切に伝わる良い指示であるといえる．この結果より，作業指示 1 を適切な作業指示と定め，再度 *A* に対して作業指示 1 をもとにラベル付けを依頼した．

初めは投票人数が割れているツイート数は 100 件全てであった．しかし完全な比較にはならないが，新たに 4 名の作業者で作業を行った場合，*A′* から多数決により決定したラベルのうち投票人数が割れているツイート数は 6 件であった．これにより指示を加えていない状態と比較して，指示を誤解する作業者を減らすことができる作業指示を定めることができたといえる．

また，第一著者が正しいと思うラベルと比較したときのラベルの訂正必要数を表 4 に示す．表 1 の下段と表 4 のラベルの訂正必要数を比較すると，関係ないに属するツイートのラベルの訂正必要数に変わりはなかった．しかし，関係あるに属するツイートのラベルの訂正必要数は大幅に減らすことができた．また，作業指示 1 を設定する際に参考にしたコロナウィルスに言及していないという特徴に属するツイートのみに着目すると，ラベルの訂正必要数は 15 個から 0 個に減っていた．これにより，作業依頼者の意図を作業結果に適切に反映することができたと言える．

しかし今回の作業指示では，作業指示を決定する際に参考に

したツイート以外では，ラベルの訂正が必要なツイート数は 14 件も存在した．ラベルを訂正する必要があるツイートには，作業指示を絞ったために，作業指示として指定できなかったが違うラベルを付与してほしいツイートがいくつかあった．また，「ぶっコロナ」などのように，ツイート数が短すぎるためにツイートの意味を理解できないツイートがあった．

作業者から寄せられた意見の中にも文章量に関するものがあった．

- 「文章が短いものは判断しにくいので，消した方がいい.」
- 「作業をしているツイートには，文章量が短いものや意味不明と思われるツイートが多く見受けられるため，作業に意味があるのか疑問に感じた.」

そのため今後は，追加でツイートのラベルを訂正できるような作業指示の候補を新たに定めて再度作業を依頼することや，文章量が少なすぎるツイートの除外をすることにより，更に作業者に伝わりやすい作業指示を決定する必要がある．

4.4 実　験　2

実験 2 では，機械学習を用いて作業結果を予測することにより，適切な作業指示を決定する実験を行った．

4.4.1 モデルの作成とラベルの予測

4.3.3 節において得られた作業者と作業指示ごとのデータを用いる．今回はモデルの精度の計測も行う．そのため，作業者と作業指示ごとに得られた 100 件のデータのうち半分の 50 件をモデル作成用に，残りの 50 件をラベル推定用に用いる．

モデル作成用のデータを用いて，作業者と作業指示ごとに，作業結果を予測するためのモデルを作成する．モデル作成用のデータを見たところ，関係あるのラベルのツート数が 8 件しか無い作業者のデータが存在した．そのため検証データには，関係あると関係ないのラベルのツイートを 2 件ずつ，合計 4 件用いる．訓練データには，検証データで用いていないツイートを用いる．このとき，ツイート数が少ないラベルに属するツイートを，ツイート数が多いラベルに属するツイートと同数になるまで複製し，合算したものを用いる．

作業者と作業指示ごとに，訓練データと検証データを用いてモデルを作成する．BERT でファインチューニングを行い，作業者 4 名に対して作業指示二つ分の合計 8 個モデルを作成する．その後，ラベル推定用の 50 件のツイートに対して，作成したそれぞれのモデルにより作業者が付与するであろうラベルを予測する．

4.4.2 予測ラベルと実際のラベルの比較

4.4.1 節において予測した結果から多数決により決定したラベルのうち，投票人数が割れているツイート数を調べる．これにより，適切な作業指示を決定する．

また第一著者が正しいと思うラベルと比較し，作業依頼者の意図が反映された結果が得ることができたかを明らかにする．更に実際に作業者により付与されていたラベルとの誤り数を明らかにし，モデルの精度を測る．

4.4.3 実験結果と考察

機械学習により予測した作業結果から多数決により決定したラベルのうち，投票人数が割れているツイート数を表 5 に示す．機械学習により予測されたラベルと第一著者が正しいと思うラベルを比較したときのラベルの訂正必要数を表 6 に示す．予測されたラベルと作業者により実際に付与されたラベルとの誤り数を表 7 に示す．

表 5　予測されたラベルの投票人数が割れている数

	作業指示 1	作業指示 2
予測ラベルの多数決	3/50 (6%)	18/50 (36%)
作業者全員の多数決	1/50 (2%)	12/50 (24%)

表 6　予測されたラベルの訂正必要数

	作業指示 1	作業指示 2
作業者 A	13/50 (26%)	24/50 (48%)
作業者 B	10/50 (20%)	25/50 (50%)
作業者 C	13/50 (26%)	35/50 (70%)
作業者 D	16/50 (32%)	17/50 (34%)
多数決	11/47 (23%)	15/32 (47%)

表 7　予測されたラベルと作業者による実際のラベルとの誤り数

	作業指示 1	作業指示 2
作業者 A	7/50 (14%)	20/50 (40%)
作業者 B	5/50 (10%)	23/50 (46%)
作業者 C	13/50 (26%)	29/50 (58%)
作業者 D	15/50 (30%)	23/50 (46%)
多数決	9/50 (18%)	35/50 (70%)

表 5 の上段は，機械学習により予測されたラベルのうち投票人数が割れているツイート数を，下段は実際に作業者により付与されていたラベルの多数決結果のうち投票人数が割れているツイート数を表している．モデルにより予測されたラベルから多数決により決定したラベルのうち，投票人数が割れているツイート数は，作業指示 1 と作業指示 2 でそれぞれ 3 件と 18 件であった．ラベル推定用に用いたツイートに元々付与されていたラベルの多数決は，作業指示 1 と作業指示 2 で 1 件と 12 件であった．投票人数が割れているツイート数は，実際に作業者がラベルを付与した場合と比較して，予測ラベルの方が増えてしまった．しかし，予測ラベルは実際に作業者により得られた結果と同じように，作業指示 1 の方が投票人数が割れているツイート数が少ないという結果が得られた．

表 6 より，作業指示 1 において予測されたラベルの訂正必要数はどの作業者も 35%以下であった．また表 7 より，作業指示 1 において，実際に人手により付与されたラベルとの誤り数はどの作業者も 30%以下であった．これより，ある程度信頼できるモデルが作成できたといえる．

しかし問題点もいくつか存在する．それは，モデルの信頼性が必ずしも高いとは言えないことである. 表 6 より，作業者 B の作業指示 1 における予測ラベルの訂正必要数は 5 件であるため，一見すると非常に高い予測精度があると考えられる．しかし，実際に予測されたラベルには全て関係ないのラベルが付与されていた．そのため，もう一度作業者 B の作業指示 1 におけるモデルを作成したところ，半数以上に関係あるのラベルが付与された．これは，使用するデータ数が十分でないためにモデ

ルが作業者の作業傾向を把握することができなかったためだと考えられる．しかし，他の作業者のモデルを再度作成してみたところ，初めに作成したモデルの結果との差はあまり無かった．これより今回作成したモデルでは，データによってモデルの信用度にばらつきがあることが分かった．そのため，今後はモデルを作成するための必要最低限のデータ数を明らかにしたうえで，作業結果の予測をする必要がある．

表 6 と表 7 より，予測したラベルの訂正必要数と実際に作業者が付与したラベルとの誤り数は，どちらも作業指示 2 の方が多かった．モデルに使用したデータを見たところ，作業指示 1 より作業指示 2 の方が，作業者が付与するラベルにばらつきがあった．また同様に，ラベル予測用のデータに付与されていたラベルも，作業指示 2 の方がばらつきがあった．

これより，ラベルのばらつきが少ないデータでは，今回と同じような少ないデータ数でもある程度満足できる結果を得ることができることが明らかになった．しかしラベルのばらつきがあるデータを対象とする場合は，作業結果を正確に予測するだけの特徴を掴めていなかった．そのため少ないデータ数でモデルを作成する場合，ラベルのばらつきが少ないデータのみに適用するか，モデルを作成するためのデータ数を増やす必要がある．

今回の実験では，モデルに使用するデータ数が少ない事が以上の問題を引き起こした要因となった．そのため，今後は実際にクラウドソーシングを用いて，大量のデータからモデルを作成することを検討している．

5 おわりに

本稿では，クラウドソーシングにおける適切な作業指示は，作業者から得られる結果のうち投票人数が割れているデータにより決めることができることを明らかにした．また機械学習により作業結果を予測することで，コストを抑えることができることを明らかにした．その結果，作業依頼者が望まないデータに対して定めた適切な作業指示で作業を依頼することで，作業指示への誤解が少ないかつ正解ラベルとの一致率が高い結果を得ることができた．また一部のデータを除き，予測した作業結果は，作業者による実際の作業結果と近しい結果を得ることができたため，コストを削減するために機械学習を用いることの有用性を示すことができた．

しかし，作業の依頼をするデータのうちの一部の文章量が短すぎるために作業者を混乱させたり，扱うデータ数が少ないために必ずしも信頼できるモデルが作成できなかったなどの問題点がある．そのため今後は，文章量の少ないデータは除外し，データを十分に確保しても本稿と同じ結果が得られるかの検証を行いたい．

謝辞 本研究の一部は JSPS 科研費 18H03342，19H04221，19H04218，および大川情報通信基金の助成を受けたものです．

文　献

[1] Aniket Kittur, Jeffrey V. Nickerson, Michael Bernstein, Elizabeth Gerber, Aaron Shaw, John Zimmerman, Matt Lease, and John Horton. The future of crowd work. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, CSCW '13, p. 1301–1318, New York, NY, USA, 2013. Association for Computing Machinery.

[2] 丹治寛佳, 清水伸幸, 森嶋厚行, 北川博之. マイクロタスク型クラウドソーシングにおける質問文改善の支援手法. In *DEIM*, pp. online(C6–1), 2015.

[3] Aniket Kittur, Ed H. Chi, and Bongwon Suh. Crowdsourcing user studies with mechanical turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, p. 453–456, New York, NY, USA, 2008. Association for Computing Machinery.

[4] Steve Branson, Catherine Wah, Florian Schroff, Boris Babenko, Peter Welinder, Pietro Perona, and Serge Belongie. Visual recognition with humans in the loop. In *Proceedings of the 11th European Conference on Computer Vision: Part IV*, ECCV'10, p. 438–451, Berlin, Heidelberg, 2010. Springer-Verlag.

[5] Vamshi Ambati, Stephan Vogel, and Jaime Carbonell. Towards task recommendation in micro-task markets. In *Proceedings of the 11th AAAI Conference on Human Computation*, AAAIWS'11-11, p. 80–83. AAAI Press, 2011.

[6] 山下裕, 小林正樹, 若林啓, 森嶋厚行. クラウドソーシングにおける AI を利用したタスク削減手法. In *DEIM*, pp. 71,online(I5–3), 2020.

[7] A. P. Dawid and A. Skene. Maximum likelihood estimation of observer error‐rates using the em algorithm. *Journal of The Royal Statistical Society Series C-applied Statistics*, Vol. 28, pp. 20–28, 1979.