人と AI の能力向上を共に目指したマイクロタスク割り当て手法

中山　拓海†　　松原　正樹††　　森嶋　厚行††

† 筑波大学情報学群情報メディア創成学類　〒 305-8550 茨城県つくば市春日 1-2
†† 筑波大学 図書館情報メディア系　〒 305-8550 茨城県つくば市春日 1-2
E-mail: †takumi.nakayama.2020b@mlab.info, ††{masaki,mori}@slis.tsukuba.ac.jp

あらまし　データ駆動型 AI は人間が提供するデータを参考に学習して構築される. 近年は, AI が人間の能力を超えることもあり, 人間の能力向上のために AI の振る舞いを参考にする事例が現れている. クラウドソーシングにおいても, 人間の能力向上のために AI を利用するという研究がある. しかし, 人間と AI の能力を向上させ合うサイクルを実現する方法はまだ明らかではない. 本研究では, クラウドソーシングにおいて人間と AI がそれぞれの能力を向上させるタスク割り当て手法を提案する. 提案手法では, まず人間と AI の能力を Dawid と Skene のモデルと能動学習を応用して推定する. その後, AI が得意なタスクは AI の予測を参考回答としてそのタスクが苦手な人間に提示し, AI が苦手なタスクはそのタスクが得意な人間に割り当てる. 本論文では, 提案手法を利用した実験を行った結果を報告する.

キーワード　機械学習，クラウドソーシング，インタラクションデザイン，タスク割り当て

1　はじめに

近年, AI 技術の発展により AI が人間の能力を超えることもあり, 人間の能力向上のために AI の振る舞いを参考にする事例が現れている. 例えば, 将棋界では 2017 年に名人に勝利する将棋 AI が構築 [1] されたが, 将棋 AI の出力する指し手を学習して棋力を向上させているプロ棋士などが登場している [2]. つまり, 従来は AI は人間が提供するデータを参考に学習して構築されていたが, 人間の能力向上のために AI の振る舞いを参考にする事例が現れている.

クラウドソーシングにおいても, 人間の能力向上のために AI の出力を利用するという研究がある [1] [8] [12]. 一方, データ駆動型 AI はクラウドソーシングを通じて人間が提供する大量のデータを参考に学習して構築される（例えば ImageNet [3] など）. このように, 一方向的に人間が AI から学習する・AI が人間から学習という研究の前例はあるものの, 双方向的な繰り返しによって人間と AI の能力を向上させ合うサイクルを実現する方法はまだ明らかではない.

そこで本研究では, クラウドソーシングにおいて人間と AI がそれぞれの能力を向上させるタスク割り当て手法を提案する. 提案手法では, まず人間と AI の能力を Dawid と Skene のモデル [2] と, 能動学習の方法 [7] を組み合わせ推定する. その後, AI が得意なタスクは AI の予測を参考回答としてそのタスクが苦手な人間に提示し, AI が苦手なタスクはそのタスクが得意な人間に割り当てる. このプロセスを繰り返すことにより両方の能力を向上させることを目指す.

本論文では, クラウドソーシングで依頼した不特定多数のワーカに対し, 提案手法を利用した実験を行った結果を報告する. 方言の分類問題を題材に提案手法に基づいて人間と AI へのタスク割り当てを 4 サイクル行った. その結果, 人間と AI, 両方の能力が向上することが確認された.

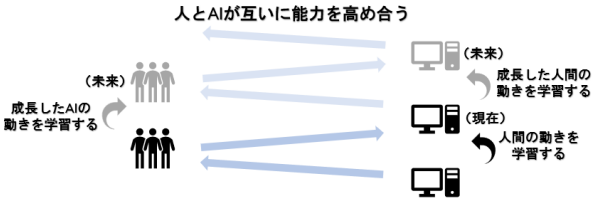

図 1　人間と AI の学習サイクル

2　関連研究

2.1　Dawid と Skene のアルゴリズム

本研究では, タスクはいくつかの分野にグループ分けでき, 各分野ごとのワーカの能力を計測する方法として Dawid と Skene のモデル [2] を用いる. Dawid と Skene のモデルでは, EM アルゴリズムを用いて, 分類対象タスクの正解と人間ワーカの能力を交互に推定する. EM アルゴリズムは, タスクの正解を推定する E ステップと, ワーカの能力を推定する M ステップで構成されており, E ステップで推定したタスクの正解を元に, M ステップでワーカの能力を推定し, M ステップで推定したワーカの能力を考慮して, E ステップでタスクの正解を推定するという動作を繰り返す. このアルゴリズムでは, 各ラベルに対してワーカの能力を算出することができ, 本研究ではワーカの各方言に対する能力を推定するために利用している.

2.2　Self Correction

Shah らはクラウドソーシングのタスクにおいて, 回答を選んだ後, 他者の回答を参考情報として提示し, 提示された回答を見

1：公益社団法人, 日本将棋連盟, 「第 2 期電王戦二番勝負, PONANZA の 2 連勝で幕を下ろす」(2017 年 05 月 22 日), https://www.shogi.or.jp/news/2017/05/2ponanza2.html, (参照 2020-12-24)

2：Yahoo Japan ニュース, 「羽生善治「将棋の世界は, 人間が AI から学んでいく時代に入っている」新井紀子×羽生善治×重松清〈前編〉」(2020 年 07 月 10 日), https://news.yahoo.co.jp/articles/1e48d73e51ba9e09a5cd28593b75dbcc6b4bdb27, (参照 2020-12-24)

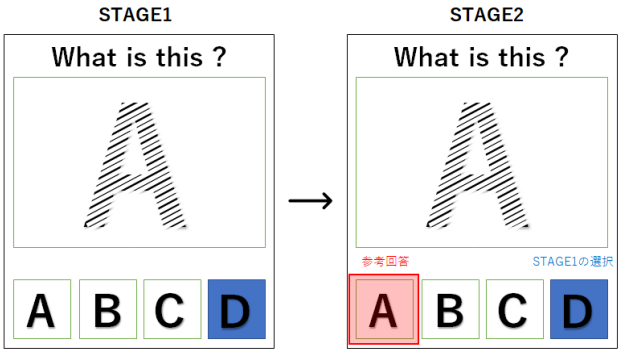

図 2　SelfCorrection の実施例: 2 つの STAGE からなり, STAGE1 ではワーカが自分の考えで選択し, STAGE2 では STAGE1 で選んだ回答と, 参考回答として他のワーカの回答が表示される

て自身の回答を変更すること許可する Self Correction [10] という方法を提案した（図 2）. 具体的には Self Correction は 2 つのステージからなる. Stage1 では, ワーカが自由に回答を選び, その後の Stage2 では, Stage1 で選んだ選択肢と, 他のワーカが選んでいる選択肢を表示する. Stage2 では他のワーカの回答を見た上で, 自身の回答を変更することができる.

これによりタスク結果の品質向上がシミュレーションにより示されている.

また Kobayashi らは, SelfCorrection の繰り返しによりタスク結果の品質向上だけでなくワーカの能力向上の効果があることを示した [8] [6].

2.3　AI の能力を向上させる研究

データ駆動型 AI はクラウドソーシングで大量にデータを集めて構築することが通常である. その中で, AI を精度向上させる方法として能動学習がある. Lewis らは, 最も不確定な要素を人間のワーカにラベル付けしてもらう Uncertainly Sampling における, Margin Sampling を提案している [7]. Margin Sampling というのは, （「1 番目に確率の高いラベル」の確率 − 「2 番目に確率の高いラベル」の確率）の値が小さいほど AI の確信度が低いとする方法である.

本研究では, タスクに対する AI の確信度を計算し, AI が得意とするタスク, 苦手とするタスクを判断するために利用している. これによって苦手なタスク（確信度の低いタスク）を, その分野が得意な人間に回答してもらうことができる. 確信度の低いタスクを, 能力の高い人間に回答してもらう理由としては, 能動学習で AI の精度を向上させるためには, 確信度の低いタスクに対して, 正解データをラベル付けする必要があるからだ.

2.4　AI を利用して人間の能力を向上させる研究

特に本研究に類似する研究として, Abad らが行った, クラウドソーシングの品質保証のために, 指導用データとして相応しいデータを, ワーカの回答から作成するという研究がある [1]. この研究では, ワーカの回答した問題に対して, AI の予測に基づいて順位をつけ, その順位が最も高いものを, タスクを説明するための例題とし, その次に順位が高いいくつかの問題を, ワーカが練習として解く問題とする. その後, 選ばれなかった問題に対しては, 指導用データを用いたクラウドソーシングにて, ラベル付けを行ってもらう. その後, ラベル付けした回答も含めて AI による計算を行い, 指導用データを更新し, 再度クラウドソーシングを行っていくという研究である. 実験の結果, トレーニング無しでクラウドソーシングを行うワーカより, 関連研究の手法を用いたワーカの正解率が高く, 正解データを必要とせずとも, ワーカの能力を向上させることが出来る, ということが確認されている. しかし, この研究では, AI の成長に関しては研究を行っておらず, 本研究の, 人と AI を共に成長させるという目的とは異なる.

また, 2.2 の SelfCorrection では, STAGE2 で他のワーカの回答を見せる代わりに, AI の予測回答を参考回答として見せることでも, ワーカの能力が向上することが確認されている [8] [6].

その他にも, クラウドソーシングにおいてワーカーの信念や先行概念を推論し, パフォーマンスに応じたボーナスを与える手法をとる研究や [11] や, 学習者に合った教育フレームワークを選択するという研究 [5], 多クラス分類タスクの精度を向上させることを目標とした研究 [4] などもあるが, どれも AI の能力向上を目指してはいない.

3　提 案 手 法

提案手法の全体像を述べ, 詳細のアルゴリズムについて記述する.

3.1　全　体　像

提案手法は, AI の能力が高い分野に関しては, その分野が苦手な人間が, AI の回答を参考に学習し, AI の能力が低い分野に関しては, その分野が得意な人間が, AI の学習に必要なデータを提供する, という戦略をとる.

具体的な流れについて, 下記の通りである（図 3）. まず, 依頼者から受け取った全タスクデータに対して, タスク割り当てプログラムを実行し, クラウドソーシングで依頼するタスクと, そうでないタスクに分ける. その後, 依頼するタスクに関しては, クラウドソーシングプラットフォームを用いて, クラウドワーカーに依頼する. この時, クラウドワーカーには SelfCorrection を用いたタスクを行ってもらうが, その具体的な内容に関しては 4 章の実験で説明する.

次に, クラウドソーシングの結果に対し, Dawid と Skene のモデルを使用するプログラムと, 機械学習を行うプログラムを実行する. Dawid と Skene のモデルを使用したプログラムでは, EM アルゴリズムにより, ワーカの能力推定と, 正解のラベル推定を交互に行うが, 提案手法では, ワーカの能力に対する推定結果のみを使用する.

機械学習のプログラムでは, 2 つのフェーズに分けてその動作を行う. 学習フェーズでは, クラウドソーシングの結果を機械学習し, 方言を分類する AI を生成する. 機械学習の方法は, ナイーブベイズ分類を使用し, 文章中の単語の出現率を計算することで, 推定を行う.

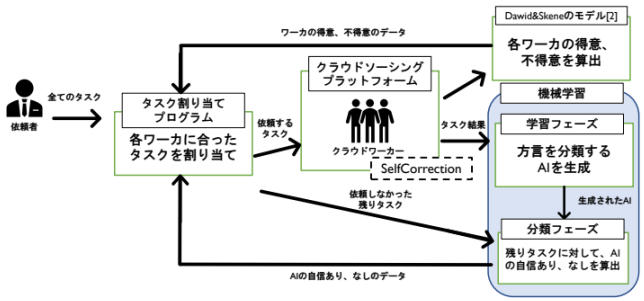

図 3　提案手法の全体像

また, 今回実施した実験では, Oyama の研究 [9] を参考に, ワーカの回答への自信を考慮しており, タスクの際に自信の有無を聞き, 自信があるとされた回答のみを学習する方法をとっている. これは, 誤った回答が広がることを防ぐ目的で実装している. 提案手法で利用している, 人間の能力を推定する Dawid と Skene のモデルは重み付け多数決であり, 多数の人間が回答する選択肢を正解とする傾向があり, それによってワーカの能力も推定される. 以前の研究室における被験者実験の際に, 参加者の多くが同じ誤った選択肢を選び, それがサイクルを重ねるごとに, 誤った選択肢が正解としてワーカ全体に広まるという結果になった. それを踏まえ, 今回実施したクラウドソーシングでの被験者実験に関しては, 誤った回答が広がることを防ぐ目的として, 自信度を採用している.

分類フェーズでは, 学習フェーズによって生成された AI が, タスク割り当てプログラムで依頼しないとした, 残りのタスクに対して分類を行う. その後, 分類結果に対して, 能動学習における Margine Sampling を適用し, AI のタスクに対する確信度を出力する.

こうして得られた, ワーカの能力を推定したデータと, AI のタスクに対する確信度を出力したデータを, タスク割り当てプログラムに送り, 2 つのデータを元に各ワーカに対してタスク割り当てを行う. 具体的なタスク割り当ての方法に関しては次の 3.2 節において説明する.

3.2　タスク割り当てアルゴリズム

変数, 関数の定義

タスク割り当てアルゴリズムで使用する変数, 関数の定義については表 1 で示す.

アルゴリズムの説明

本研究で取り扱う問題は, 方言データ U, ワーカ集合 W とその能力 $C(w_i)$ が与えられたときに, タスク割り当て T を出力することである. 以下に問題を解決するアルゴリズムを示す.

1 行目で集合 U の上から x 個を U_x, 下から y 個を U_y とする新たな集合を作る. U_x には AI の確信度の高いタスク, U_y には AI の確信度の低いタスクが格納されている.

2 行目では, 1 行目で作成した, 確信度の高いタスクの集合 U_x と, 確信度の低いタスクの集合 U_y をクラスごとに分類し, 新たな集合 U_{xi}, U_{yi} に格納する. 例えば, U_{x1} には U_x のうち, クラス 1 に属するタスクが格納されている.

3〜5 行目では, 方言ごとに x と y に割り当てる人数を計算する. 例えば a_{x1} は, クラス 1 の x に割り当てる人数で, クラス 1 において能力の高いとされた人の数を表している.

表 1　変数, 関数の定義

変数	変数の説明
U	方言データ
W	ワーカ集合.$(w_1 \dots w_n)$
$C(w_i)$	ワーカ w の能力
n	クラス数
x	確信度の高いタスクの数
y	確信度の低いタスクの数
T	タスク割り当て表
U_x, U_y	割り当てに使用する方言データ
U_{xi}, U_{yi}	方言ごとの確信度順に並んでいる方言データ
a_{xi}, a_{yi}	各方言に割り当てる人数
W_i	各方言ごとのワーカの集合.
W_{xi}, W_{yi}	方言ごとの確信度順に並んでいるワーカデータ

6, 7 行目は, クラスの数だけ新しくワーカの集合を作り, それぞれを能力順にソートする.

8 行目では, 4, 5 行目で作成した W_i を a_{xi}, a_{yi} に従って分割する. 例えば, $|W_1| = 10, a_{x1} = 6, a_{y1} = 4$ の場合, W_{x1}は W_1 のうち下から 6 人, W_{y1}は W_1のうち上から 4 人を格納した集合となる.

9 行目では, 8 行目までに作成した, 各クラスごとのワーカとタスクの集合に従って組み合わせを作り, タスクを割り当てる. 結果はタスク割り当て表 T に格納する.

Algorithm 1 Algorithm for Assignment

Input: U, W, w, n, x, y

Output: T

方言ごとのタスク集合を作る

1: $U_x, U_y \leftarrow$ 方言データ U の上から x, 下から y 個を集合とする

2: $U_{xi}, U_{yi} \leftarrow U_x, U_y$を各方言ごとに分類

方言ごとに割り当てる人数を出す

3: $z_i = |U_{xi}| + |U_{yi}|$

4: $a_{xi} = (|U_{xi}|/z_i)|W|$.

5: $a_{yi} = |W| - a_{xi}$

方言ごとのワーカ集合を作る

6: $Wi \leftarrow$　ワーカ W から方言ごとの集合を作る

7: Wi をソートして能力順に並べる

8: $W_{xi}, W_{yi} \leftarrow$ ワーカ W_iの上から a_{xi}人, 下から a_{yi}人をそれぞれ集合とする

タスク割り当て表を作る

9: $T \leftarrow$ 各方言の, ワーカ (W_{xi}, W_{yi}) とタスク (U_{xi}, U_{yi}) に従って組み合わせを作る

4　実　　験

本実験では, 提案手法を適用したタスクを, 数サイクルに分けてクラウドワーカに実施してもらい, 人間と AI のタスクに対する正解率が向上するのか検証した. 実験は Yahoo クラウドソーシングを用いて不特定多数のワーカに依頼し, オンライン上で実施した. 評価方法としては, 各サイクルごとの人間と AI

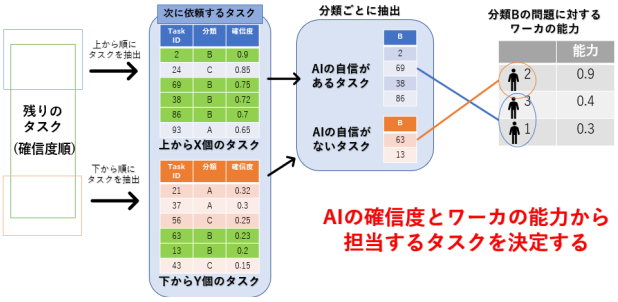

図 4　タスク割り当てのアルゴリズム

それぞれの正解率を算出し, その正解率の推移を見ることで判断した. なお, 本実験は筑波大学研究倫理審査委員会により承認（第 20-106 号）を受けて実施している.

4.1　設　　定

タスク

タスクの内容は, 図 5 のような方言が含まれる文章に対して, どの地域の方言かを問う問題である. 回答の際には, その回答に対する自信の有無を聞き, 自信ありと選択された回答のみを機械学習の対象とした. 対象とする地域は, 北海道, 東北, 中部, 関西, 九州で, 問題に答える際には, 5 つの選択肢の中から選んで回答する. 問題文は, Yahoo クラウドソーシングにおいて, 不特定多数のワーカに方言の文章の入力を依頼し, そこで収集されたデータから, 抜粋して作成している. 各地域に対して 20 個ずつ問題文を作成した合計 100 問を, 参加人数分複製した. タスクには, SelfCorrection を用いており, 図 5 のように 1 回目に AI の予測がない状態で回答した後, 自身の回答と AI の回答を提示し, 変更をするかを被験者に問う. 本実験では, 後者の結果をタスク結果として使用している. また, 1 回目の問題に関しては, AI がまだ生成されていないため, ランダムに問題文を選び, 全員に同じタスクを割り当てたものをタスクデータとしている.

クラウドソーシングプラットフォーム

この実験では, Yahoo クラウドソーシングと Crowd4U を利用した. Yahoo クラウドソーシングはワーカを募集し, 参加したワーカに Crowd4U へのリンクを渡すために使用した. Crowd4U では, 参加者に実際にタスクを行ってもらう. 各サイクルごとのタスクを予め用意し, 実験開始の段階で, タスクのデータファイルをアップロードすることで, 即座にタスクを行うことができる状態になっている. アップロードするタスクのデータは, 被験者の ID, タスクの ID, 問題文, AI の予測から構成されている. ただし, 一回目のサイクルに関しては, AI の生成を行えないため, タスクデータに AI の予測は含まれいない. また, タスク終了後はタスク結果を出力したファイルをダウンロードする. そのファイルのデータは AI を生成するプログラムと, Dawid と Skene のモデルのプログラムに渡す.

人間

出身や能力は未知のものとし, 11 人に依頼した. ワーカには事前に実験開始の時刻を伝え, 指定時刻になり次第, 事前に渡したタスクへのリンクから実験に参加してもらう方法を取った.

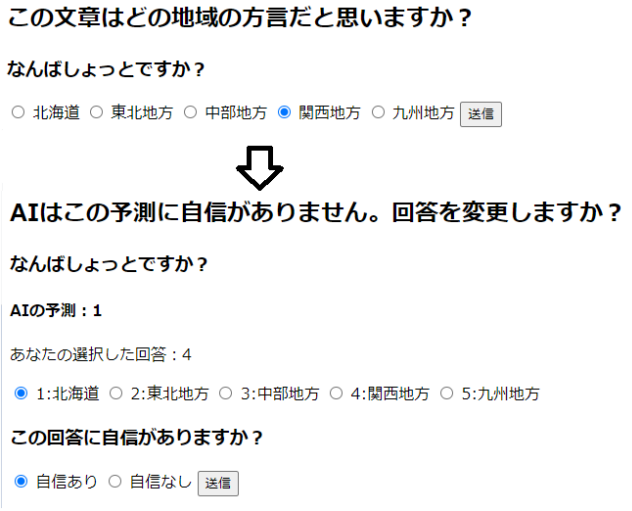

図 5　実際のタスク画面

AI

AI は各サイクルごとに, ワーカ全員がタスクを終了した後, その都度タスク結果を学習して生成した. 生成された AI は, 依頼しなかったタスクに対して, ラベルの予測と確信度の計算を行う. 計算後は, そのデータをタスク割り当てプログラムに送る. 機械学習の方法は, ナイーブベイズ分類を使用し, 文章中の単語の出現率を計算することで, 推定を行った. 具体的には, 学習データにおける, 各方言の文書数を総文書数で割った値（方言出現率）と, 各方言内の文書出現率 を計算し, その 2 つを掛け合わせたものを文章内の方言出現率としている. また, その出現率を 2.3 節の Margine Sampling [2] に適用し, AI の確信度の計算を行っている.

タスク割り当て

Dawid と Skene のモデルによる, 人間の各方言の能力を出力したデータと, 生成された AI による, 各タスクに対する予測ラベルと確信度を出力したデータを受け取り, タスク割り当てを行う. 割り当て後は, 次に依頼するタスクデータと, 依頼しない残りのタスクデータの 2 つのファイルが出力される. 後者に関しては, 次のサイクルで AI を生成した後, その AI によって分類が行われる.

実施方法

実験はオンラインで実施した. 実験中, タスク結果を学習し, 新たにタスク割り当てを行う際には, ワーカには待機してもらい, 新しいタスクデータをアップロード後, 次のサイクルの時刻になり次第, タスクを実施してもらった.

実験は以下の流れに従い実施

（1） タスクデータを Crowd4U にアップロードする

（2） ワーカがタスクに回答する

（3） すべての回答が終了後, 図 3 の流れに従い, 新たなタスク割り当て表と, 割り当て表に従ったタスクデータを作成する

（4） 1〜3 を設定したサイクル数回繰り返す

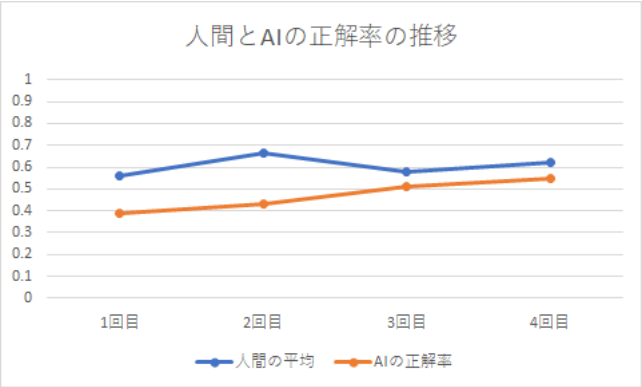

図 6　被験者と AI の正解率の推移. 赤線が AI の平均正解率, 青線が被験者の正解率を表している.

4.2　実験の結果

図 6 は, 各回ごとの被験者の平均正解率と, 各回までに得られたデータを機械学習し, 生成された AI の正解率の推移を表している. 縦軸は, 正解率の値を 0〜1 までの値で表し, 横軸はサイクル数を表している. 各線はそれぞれの正解率を表しており, 赤線が AI の正解率, 青線が被験者の平均正解率を示している.AI の正解率に関しては, 学習後生成された AI が全問題に対して予測を行い, その正解率を記載している.

図 7 はすべての回答における被験者の平均正解率と, 自信ありとした回答における被験者の平均正解率を比較したものである. 表の値は正解率を 0〜1 までの値で表し, 実際の値を小数点第 3 位を四捨五入したものを記載している.

	1回目	2回目	3回目	4回目
すべて	0.56	0.67	0.58	0.62
自信ありのみ	0.76	0.79	0.73	0.75

図 7　すべての回答における被験者の平均正解率と, 自信ありとした回答における被験者の平均正解率の比較

図 8 と図 9 では, 方言ごとのデータについてまとめている.

図 8, 9 は人間と AI それぞれの方言ごとの正解率の推移を表しており, 横軸と縦軸は, 図 6 と同じ表現を用いている.

図 10 は関西地方に関する成績が最も低下したワーカと, 最も向上したワーカそれぞれの個人成績を表しており, 表の値は正解率を 0〜1 までの値で表し, 実際の値を小数点第 3 位を四捨五入したものを記載している.

5　考　　察

図 6 から 1 回目と 4 回目を比較して, 人間の平均正解率は 0.56 から 0.62 へと増加しており, 増加率は高くはないが, 人間のタスクに対する能力が上昇傾向にあるといえる. また, AI の正解率に関しては, 0.39 から 0.55 へと増加しており, AI の精度はサイクルを重ねるごとに上昇している. 人間の能力の変化が, AI の能力の変化と比較して小さい理由として, AI の能力が人間の能力より低く, AI の予測が人間の学習に及ぼす影響が小

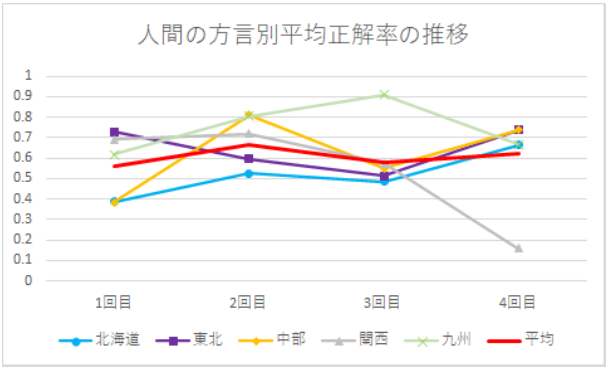

図 8　方言ごとの人間の平均正解率の推移

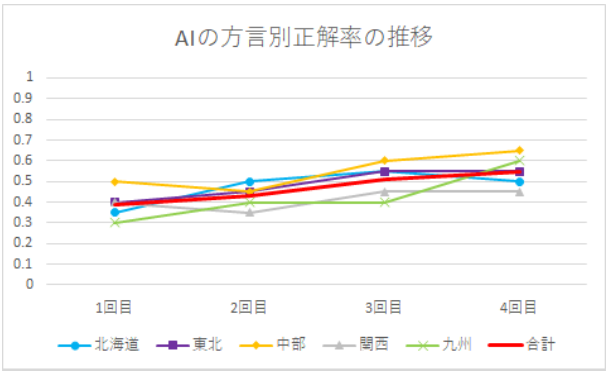

図 9　AI の方言ごとの正解率の推移

最も関西が低下したワーカ

ID:14	北海道	東北	中部	関西	九州
1回目	0	0.4	0.2	0.8	0
2回目	0.17	0.4	0.67	0.75	0.8
3回目	0	0.25	0.25	0.57	1
4回目	0.5	0.33	0.89	0	0.25

最も関西が向上したワーカ

ID:16	北海道	東北	中部	関西	九州
1回目	0.8	0.6	0	0.2	0.6
2回目	0.63	1	0.33	0.5	0.71
3回目	0.5	0.5	0.4	0.6	1
4回目	1	0.5	0.5	0.5	1

図 10　関西地方における成績が, 最も向上したワーカと, 最も低下したワーカそれぞれの個人成績

さかったのではないかと推察される. また, 1, 2 回目と 3, 4 回目を比べると人間と AI の正解率の差が小さくなっている. このことから, 更に学習を継続することで, AI の精度がより上昇することが予測される. そのため, 今後サイクル数を変更した実験を行い, 正解率の変化について更に検証を行う必要がある.

図 7 を見ると, すべてのサイクルにおいて, 自信ありと回答した際の平均正解率は, 全ての回答に対する平均正解率より高く, AI が学習するデータとして, より相応しいものを選ぶことが出来ていると言える.

図 8 と図 9 を見ると, 方言ごとの条件の違いによって, 正解

率の変化が異なっている. 図 8 における人間の平均正解率に関しては, 北海道と中部との 2 つの地域に関しては, 1 回目と 4 回目を比べると数値が上昇し, 東北と九州に関しては, 2, 3 回目において正解率の変化はあるものの, 1 回目と 4 回目を比べると正解率がほぼ変化していない. また, 関西に関しては, 1 回目から 4 回目にかけて正解率が下降傾向にある. 図 9 における AI の正解率に関しては変化量は異なるが, どの方言においても 1 回目から 4 回目にかけて正解率が上昇している.

図 8 と図 9 を比較すると, 人間の正解率が上昇した北海道と中部に関しては, AI の正解率が他の地域と比較して高いが, 人間の正解率が下降した関西に関しては, AI の正解率も低いという結果になっている. また, 1 回目と 4 回目を比較して正解率が変化しなかった, もしくは低下した東北, 関西, 九州に関しては, 1 回目の人間の正解率が高く AI の正解率と差がある一方で, 正解率の上昇した中部と九州は, 1 回目の正解率が低く, AI の正解率とも差が小さい. これらのことから, 人間の能力の成長には, AI の正解率と人間の正解率の関係によって変化し, 特に,AI の正解率が人間の正解率と同程度, もしくはそれ以上の場合, 人間の能力が向上するのではないかと考察できる.

また, 人間全体の正解率が低下した関西地方について, 図 10 において個人ごとの成績の考察を行う. 図 10 では, 方言全体の正解率が低下している一方で, 関西地方の成績が上昇したワーカも存在するが, 元々正解率の高かったワーカの能力がかなり低下していることがわかる. このことから, 今回関西地方全体の成績が低下した理由としては, 元々正解率の高かった多くのワーカが, 精度の低い AI の回答に影響を受け, 正解率が低下してしまったことにあると考えられる. 特に今回の実験では, 図 5 のように,AI の回答が自信あるか, ないかをタスク上部に表記しており,AI の回答が自信ありの際に, その回答が間違っていた場合, ワーカに対して悪影響が出てしまったのではないかと考えられる. そのため今後の実験において,AI の回答の見せ方について変更を行うことなど, 精度の低い AI による影響の与え方を改善していく必要がある.

今回の実験では, 元々の人間の正解率が AI の正解率より高いという条件の下で, 比較的正解率の低かった AI の精度がより向上するという結果になった. 今後の課題として, 予め学習したＡＩを使用し, 人間よりもＡＩが正解率が高いという条件で実験を行うことや, サイクル数を増やして今後の正解率の推移を見ることなど, 条件を変更して更に実験を行い, どういった条件で人間とＡＩが成長するかより深く検証し, それに伴い提案手法を改善していくことで, 有効性を向上させていく必要がある.

6 結　　論

本論文では, SelfCorrection と Dawid と Skene のモデル, 能動学習の 3 つの技術を組み合わせ, 人間と AI の能力を成長させる方法を提案した. 具体的には, AI の能力が高い分野に関しては, その分野が苦手な人間が, AI の回答を参考に学習し, AI の能力が低い分野に関しては, その分野が得意な人間が, AI の学習に必要なデータを提供する, という戦略をとり, それをクラウドソーシングタスクに組み込んだ. このクラウドソーシングタスクを, 不特定多数のクラウドワーカに依頼し, その結果を検証したところ, 人間と AI が教えあうことにより, それぞれの能力が成長することが確認された.

今後の課題として, クラウドソーシング実験を更に行い, 能力が上昇する条件をより明確にし, その上で提案手法を改善し, この手法の有効性を向上させる必要がある.

謝　　辞

本研究の一部は, JST CREST(JPMJCR16E3) および AIP チャレンジの支援を受けたものである. ここに謝意を示す.

文　　献

[1] Azad Abad, Moin Nabi, and Alessandro Moschitti. Autonomous crowdsourcing through human-machine collaborative learning. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 873–876, 2017.

[2] Alexander Philip Dawid and Allan M Skene. Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, Vol. 28, No. 1, pp. 20–28, 1979.

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

[4] Xiaoni Duan and Keishi Tajima. Improving multiclass classification in crowdsourcing by using hierarchical schemes. In *The World Wide Web Conference*, pp. 2694–2700, 2019.

[5] Liu Jiacheng, Hou Xiaofeng, and Tang Feilong. Fine-grained machine teaching with attention modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34, pp. 2585–2592, 2020.

[6] Masaki Kobayashi, Hiromi Morita, Masaki Matsubara, Nobuyuki Shimizu, and Atsuyuki Morishima. Empirical study on effects of self-correction in crowdsourced image classification tasks. In *2020 Human Computation Journal*, 2020.

[7] David D Lewis and Jason Catlett. Heterogeneous uncertainty sampling for supervised learning. In *Machine learning proceedings 1994*, pp. 148–156. Elsevier, 1994.

[8] Masaki Matsubara, Masaki Kobayashi, and Atsuyuki Morishima. A learning effect by presenting machine prediction as a reference answer in self-correction. In *2018 IEEE International Conference on Big Data (Big Data)*, pp. 3522–3528. IEEE, 2018.

[9] Satoshi Oyama, Yukino Baba, Yuko Sakurai, and Hisashi Kashima. Accurate integration of crowdsourced labels using workers' self-reported confidence scores. In *Twenty-Third International Joint Conference on Artificial Intelligence*, 2013.

[10] Adish Singla, Ilija Bogunovic, Gábor Bartók, Amin Karbasi, and Andreas Krause. Near-optimally teaching the crowd to classify. In *ICML*, Vol. 1, p. 3, 2014.

[11] Runzhe Yang, Yexiang Xue, and Carla Gomes. Pedagogical value-aligned crowdsourcing: Inspiring the wisdom of crowds via interactive teaching. 2018.

[12] Jing Zhang, Huihui Wang, Shunmei Meng, and Victor S Sheng. Interactive learning with proactive cognition enhancement for crowd workers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34, pp. 540–547, 2020.