人間と AI の協働による作業品質制御の研究

研究代表者　　小　山　　聡　　北海道大学大学院情報科学研究院　准教授

1 はじめに

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に仕事を依頼できる仕組みであり、マイクロタスクと呼ばれる、一つの画像の分類のような仕事量も作業時間も小さなタスクから、プロジェクトと呼ばれる、作業量も多く期間も長期に渡るタスクまで、幅広い仕事が扱われている。近年、クラウドソーシングが、機械学習に関わる様々な場面で用いられるようになってきている[鹿島他, 2016]。たとえば、機械学習に必要な正解付きの訓練データは従来は専門家が作成してきたが、クラウドソーシングで比較的容易に作成することができるようになり、機械学習の適用範囲が広がった。しかし、専門的知識のない作業者の作業結果には品質が保証できないという問題があり、クラウドソーシングにおいては作業の品質制御が重要な研究課題となっている。

従来の品質制御は依頼者の側が、作業結果をチェックして品質の低い作業を却下したり、過去の作業履歴に基づいて作業者をフィルタリングしたりといった形で行われ、作業者の側は積極的には関与してこなかった。これは、作業者を入力（タスク）を受け取って結果を出力する単純な入出力関数とみなす考え方である。実際には作業者は高度な知能を備えた人間であり、タスクの難易度や必要な時間、作業結果に対する自信[Oyama et al., 2013]といった、作業品質に関わる高度な知識を提供することが可能である。また、作業結果を利用する際も、単純な多数決ではなく、機械学習などの AI 技術を用いた、より高度な品質制御アルゴリズムが用いられるようになってきており[Atarashi et al., 2018]、人間と AI がうまく協調して作業を行う必要性が高まっている。

本研究調査では、人間と AI との能力を高い次元で組み合わせた協働を実現するための基礎研究と、感情分析への応用研究を行った。具体的には、作業者に品質に関する多様な情報をフィードバックしてもらうためのタスク設計、クラウドソーシングによる作業結果を集約し、専門家と同様な結果を出力するための機械学習モデルの構築、複数の作業者が関与するプロジェクト型のクラウドソーシングで報酬を分配するためのインセンティブ設計、および品質制御モデルの感情分析への応用といった研究を行った。

2 品質制御のタスク設計

クラウドソーシングにおいては、作業者の経験や能力にばらつきがあるため、作業者の能力に応じた難易度のタスクを提示することが、作業品質の向上に重要である。これは、オンライン教育において、学習者の習熟度に応じた難易度の問題を提示する課題とも共通する。タスクの難易度を推定することは簡単ではないため、クラウドソーシングの作業者からタスクの難易度に関するフィードバックを得て利用することを考えた。魔法陣パズルを例題として、作業者からフィードバックを収集するためのタスク画面を設計し、実際のクラウドソーシングサービスを用いて、作業者から得た主観的な難易度に関するフィードバックと、客観的な難易度とを比較し検証を行った。ここで、n次魔方陣とは、$n \times n$の正方形のマス目において、全ての縦・横・対角線の列の合計が全て同じ値となるように 1 からn^2までの数を一つずつ配置したものであり，行、列、対角線の和は全て$\frac{n(n^2+1)}{2}$となる。魔法陣パズルは、このうちいくつかのマス目が空欄となっており、そこに入る数を決めて魔法陣を完成させるというものである。魔法陣パズルを例題として扱ったのは、空欄の数の制御によって様々な難易度の問題を生成されることや、コンピューターに解かせることで客観的な難易度が計算できることなどが理由である。タスクとして用いたのは 4 次の魔法陣であり、図 1 に実際にクラウドソーシングで実行したタスク画面を示す。

公益財団法人電気通信普及財団
研究調査助成報告書　第35号　2020年度

図 1 タスク画面の例

このタスクでは、本来の目的である問題への解答の他に、作業者の解答に対する主観的な「自信」、問題の「難易度」、解答にかかった「時間」も入力させている。従来研究で自信を正解確率として入力させた際、二択問題にもかかわらず 50%以下の確率を入力するといった問題が生じた。そこで今回は、作業者の負荷低減のため、「自信」と「難易度」は数値ではなく、それぞれ「簡単、やや簡単、普通、やや難しい、難しい」「できなかった、多分できなかった、どちらとも言えない、多分できた、できた」の 5 段階から選択させて入力させることとした。クラウドソーシングサービス Lancers（https://www.lancers.jp/）を用いて、330 個の魔方陣パズルについて、418 人の作業者から合計 9,059 件のデータを収集した。図 2 は作業者のフィードバックから得たタスクの難易度に関する指標と、客観的な指標を比較したものである。客観的指標のうち、「コンピューター」は計算機で問題を解くのにかかる時間であり、「正解率」は複数の被験者に同じ問題を解かせた結果の平均である。「項目反応理論」は試験問題の設計などに用いられる手法であり[豊田, 2005]、単純な正解率でなく、解答者の能力も考慮して、問題の難易度を推定する。クラウドソーシングにおいては、全てのタスクをすべての作業者が行うとは限らないため、不完全なデータにも適用可能な EM タイプ IRT[作村他, 2014]と呼ばれるモデルを用いた。これらの客観的な指標に対し、主観的な指標はいずれも高い相関を示している。「コンピューター」による難易度に対する相関が他のものよりも小さくなっているのは、計算論的な難易度と、人間が感じる難易度が必ずしも一致していないことを表している。この実験で簡単な入力でも作業品質に関する充分な情報を収集できることを確認した。これにより、同じ問題を多くの人に解いてもらわなくても、主観的なフィードバックにより、品質制御に利用可能な情報を入手可能であることが分かった。

	解答時間	解答への自信	主観的難易度	空欄数
コンピューター	0.25	-0.24	0.2	0.22
正解率	0.87	**-0.98**	0.95	**0.9**
項目反応理論	**0.88**	-0.97	**0.96**	**0.9**

図 2 難易度に関する客観的な指標と主観的な指標の相関

3 品質制御の機械学習モデル構築

クラウドソーシングにおいては、同じタスクに対する複数の作業者の結果を集約することで、最終的な品

質を向上させることが行われる。結果を集約する際には、全ての作業者の結果を対等に扱うのではなく、作業の信頼性に応じて重み付けをすることが必要である。たとえば、製品等のアイテムのレビューにおいては、信頼性のある専門家に大量の製品の評価を依頼することは、作業量の面でも人件費のコストの面でも現実的でない場合が多い。一方、不特定多数の一般の人の評価は、大量のアイテムに対しても入手可能であるが、信頼性にばらつきがある。そこで、専門家の回答を正解例として、教師付き学習で集約の際の作業者の重みを決定するモデルを構築した。ここでは、作業者の回答を集約した結果が、なるべく専門家の回答と近くなるように重みを学習する。具体的には、アイテムiへの作業者jのスコアをx_{ij}、専門家のスコアをy_iとすると以下のような線形モデルで集約を行う。

$$y_i = w_1\ x_{i1} + w_2\, x_{i2} + \cdots + w_J\, x_{IJ}$$

回答が実数値の場合、この問題は回帰問題として定式化できるが、同じような回答をする作業者が複数存在すると多重共線性と呼ばれる状況が生じ、係数の推定が不安定になる。たとえば、作業者jと作業者kが複数のアイテムに全く同じスコアを付けていた場合、二人にどのように重みを振り分けるかは任意であり、重みが作業者の信頼性を正しく反映しないという問題が生じる。そこで、各作業者の重みを適切に設定するため、ゲーム理論におけるシャプレイ値の考え方を用いるシャプレイ値回帰[Lipovetsky & Conklin, 2001]を導入した。シャプレイ値[Shapley, 1953]とは、協力ゲームで提携における参加者の貢献を測る指標である。回帰モデルにおける各説明変数の重要度を測る単純な方法としては、その変数を除いたモデルと、含んだモデルとの間で誤差を比較するというものがある。しかしクラウドソーシングにおいては、作業者の回答は独立でないので、この解答は他にどのような回答者がいるかに大きく依存し、公平に作業者の貢献を測ることができない。シャプレイ値は提携に参加する順番の全ての順列での平均を用いるため、この問題が生じない。クラウドソーシングにおいては、作業者の部分集合が提携に対応し、各作業者が参加することで、正解に対する誤差がどれだけ減少するかに基づいて重みを計算する。

シャプレイ値回帰では、作業者の全ての部分集合において回帰誤差を計算する必要があり、作業者数が多い場合に計算量が大きくなる。クラウドソーシングでは、一般に全てのタスクをすべての作業者が行うわけではないため、同じタスクを行ったことがない作業者のグループが存在する。このような場合、作業者をノードとし、同じタスクを実行したことがある作業者をエッジで結んだ作業者協働グラフを考えると、エッジの比率が少ない疎なグラフとなる。ここで、シャプレイ値を計算するには、全ての部分グラフを考慮する必要はなく、連結成分にだけ回帰を実行すればよいことを示した。グラフの連結成分を効率よく列挙できる方法として、フロンティア法[Kawahara et al., 2017]が提案されている。

Web 検索において、2,665 件のクエリと URL の組に対してクラウドソーシングで 5 段階の適合性判定を行った Web Search Relevance Judging Dataset[Zhou et al., 2012]を用いて実験を行った。作業者の総数 177 人であったが、その中から 10 人、15 人、20 人をランダムに抽出して作業者協働グラフを作成した。図 3 に作業者数と必要な計算量の関係について示す。単純な方法では、可能な提携の数だけ回帰分析を実行する必要があるが、連結成分だけを考慮することで、その数を 10%以下に削減できることを確認した。

Num. of workers	Num. of connected components	Max. Num. of coalitions	Reduction rates
10	86.6	1013	8.5%
15	3072.6	32752	9.3%
20	96585.8	1048555	9.2%

図 3 作業者数と計算量の関係

4 プロジェクトにおけるインセンティブ設計

クラウドソーシングにおいて、作業者に参加へのインセンティブを与えるためには、適切な報酬を支払う必要がある。作業が一人で完結するマイクロタスクではなく、複数の作業者の協力が必要なプロジェクト型のクラウドソーシングの場合、作業者が納得する報酬の分配方法を考案する必要がある。このような状況を解析する理論的な枠組みとして、協力ゲーム理論がある。協力ゲームにおいては、提携と呼ばれる参加者の

組合せに応じて、異なる報酬が与えられる。このとき、参加者間でどのような提携が形成され、どのように報酬が配分されるかを解析するのが、協力ゲーム理論の目的であり、これまで多くの理論的な研究成果が蓄積されてきた。一方で、理論的に導出された最適な提携や報酬の支払い額と、実際に人間が望ましいと思う提携や報酬が、必ずしも一致しているとは限らない。そこで、協力ゲームにおいて、報酬の分配方法を参加者間の交渉によって定めさせる実験が行われた[Nash et al., 2012]。これらの実験は、複数の被験者を研究室に集めて実際に交渉をさせるため、手間やコストの制約から大規模化は困難であるが、もしこのようなデータが十分にあれば、適切な報酬の分配方法を機械学習によって定めることも可能となる。そこで、クラウドソーシングを用いて、作業者に適切な報酬配分案を提案させる方法を考案した。図 4 にクラウドソーシングのタスク画面を示す。このタスクにおいては、3 社での共同プロジェクトを例として、提携に応じて異なる報酬が得られるとき、どのような提携の組み方と報酬の分配方法を提案するかを尋ねている。

クラウドソーシングで得られた結果を分析したところ、参加者は均等な報酬の配分を好む傾向があるなど、実際の被験者実験と類似した結果が得られており、クラウドソーシングを用いる方法は有望であると考えられる。

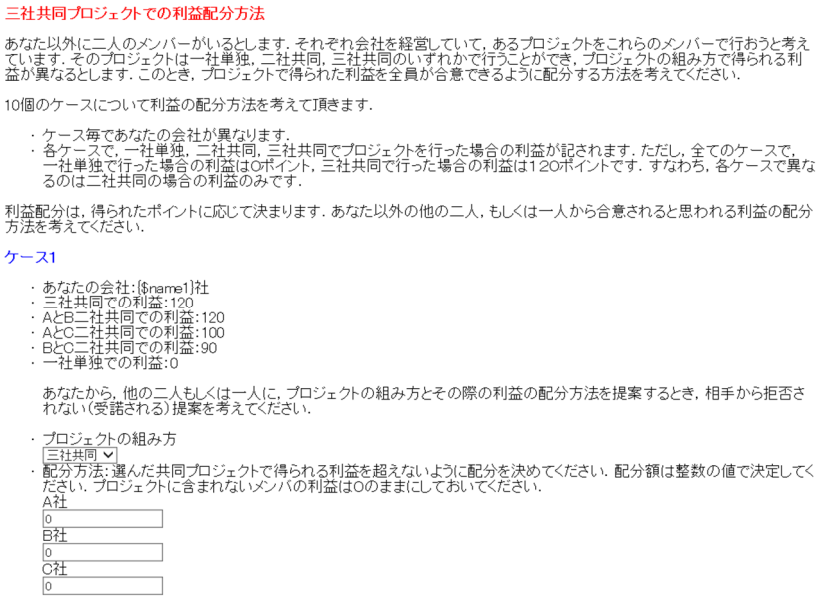

図 4 共同プロジェクトでの利益配分に関するタスク

5 品質制御モデルの感情分析への応用

近年、ソーシャルネットワーク等に投稿された文章が予想外の反響や「炎上」といった現象を引き起こすことが頻繁に起こっており、文書に人々がどのように感情的に反応するかを予測する、感情分析に対するニーズが高まっている。文章に対する感情推定は、物語文などを対象に長い研究の歴史があり[Ptaszynski et al, 2013]、最近では機械学習を用いた手法も研究されている。機械学習を用いる際には、テキストに対して、「喜び」や「怒り」とった感情ラベルを付与した正解データが必要であるが、クラウドソーシングを用いて正解データを収集し、感情ラベル間の依存関係を考慮して精度よく正解ラベルを推定する研究も行われている[Duan et al, 2014]。従来の文章に対する感情分析手法は、正解としての感情が人によらずに定まるという前提に立ち、一つの文章に対してただ一通りの出力を正解として持つことを仮定することが多い。しかし実際には、同じ内容を表した文章であっても、それぞれの読み手ごとに抱く感情は異なる場合が多い。たとえば、「阪神が巨人に勝った」というニュース記事に対して、阪神ファンはポジティブな感情を抱くが、巨人ファンはネガティブな感情を抱くであろう。より詳細な感情分析を行うには、読み手の個性による感情応答の違いを考慮した、感情推定モデルを構築する必要がある。これにより、読み手に応じてポジティブな感情をもたらす文章を提示したり，逆にネガティブな感情をもたらす恐れのある文書の提示を避けたり

することが可能になると期待できる。また、ある文書を人々の集団に提示した場合，何割ぐらいの人がどのような感情を持つか，感情の分布を推定することも可能になる。

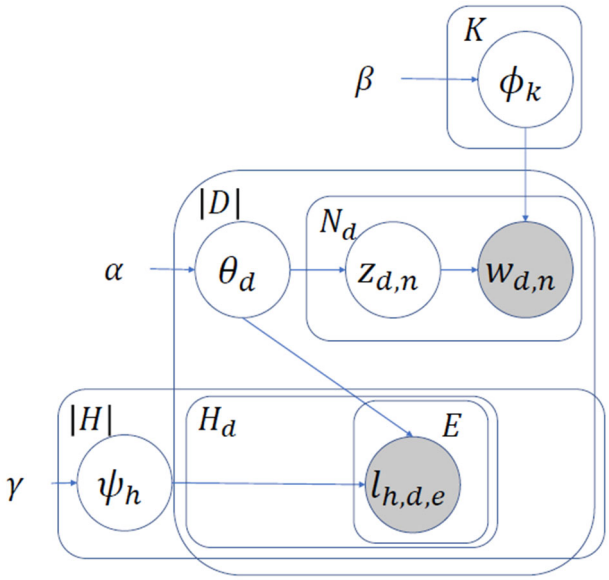

図 5 読み手に依存した感情応答の確率モデル

そこで、文章と読み手それぞれに潜在的な特徴を仮定し、それらの組合せによって感情応答が定まるというモデルを提案した。図 5 に提案したモデルのグラフィカルモデルを示す。ここでは、同じ記事に同時に複数の感情ラベルが付与される、マルチラベル問題として定式化する。これにより、一つの記事が「怒り」と「悲しみ」という感情応答を引き起こす効果をモデル化することができる。ここでHは読み手の集合であり、H_dはd番目の記事を読んだ読み手の集合、Eは感情応答の集合である。Kは潜在的なトピックの数であり、$\boldsymbol{\theta}_d$はトピックの分布で表される記事の潜在的な特徴、$\boldsymbol{\psi}_h$は読み手の潜在的な特徴である。感情文章からの潜在特徴の抽出は、Latent Dirichlet Allocation[Blei et al., 2003]をベースにしている。感情応答$l_{h,d,e}$は記事の潜在特徴$\boldsymbol{\theta}_d$と読み手の潜在特徴$\boldsymbol{\psi}_h$から、ベルヌーイ分布に従い

$$l_{h,d,e} \sim \mathrm{Ber}(\langle \boldsymbol{\theta}_d, \boldsymbol{\psi}_h \rangle)$$

で生成されると仮定している。

モデル	他の作業者のラベルの利用	ラベルなしデータに対する予測	ラベルなしデータの訓練での利用
DSモデル	○	×	×
PCモデル	○	○	×
IPCモデル	×	○	×
提案モデル	○	○	○

図 6 既存モデルと提案モデルの比較

従来のクラウドソーシングにおける品質制御モデルと提案モデルとの比較を図 6 に示す。DS モデル[Dawid & Skene, 1979]は EM 法を用いた古典的なラベル統合の方法であり、データへのラベルのみを考慮して、データの内容は考慮しないため、ラベルが全く付与されていないデータに対する予測を行うことはできない。PC モデルは各作業者が個別の分類器を持つモデルであり、ラベルなしデータへの予想が可能である[Kajino et al., 2012]。ここで、各作業者の分類器のパラメータは、ベースとなる分類器のパラメータと、各個人の特性を表す部分との和になっており、ベースとなる分類器をとおして他の作業者のラベルも利用することができる。IPC モデルは比較実験のために PC モデルを単純化したモデルであり、ベースとなる分類器を持たずに、各作業者のモデルが独立に学習される。これらのモデルは、ラベルなしのデータを訓練で利用することはで

きない。これに対して提案モデルは、他の作業者のラベルを利用することができ、ラベルなしデータを訓練において利用し、ラベルなしデータに対する予測を行うことも可能であるという特徴を持つ。

感情ラベル	怒り	悲しみ	喜び	嫌悪	驚き	恐怖
出現頻度	1,738	1742	1,861	1,248	1,067	1,836

図 7 感情ラベルの出現頻度

livedoor ニュースコーパス（https://www.rondhuit.com/download.html#ldcc）から抽出した 220 件の記事に対してクラウドソーシングサービス Lancers で、各記事に感情ラベルを付与した。各記事に対して 30 名の作業者にラベル付けを依頼し、異なる作業者の数は 95 名であった。感情ラベルとしては、「怒り」「悲しみ」「喜び」「嫌悪」「驚き」「恐怖」の 6 種類を用い、1 つの記事に少なくとも 1 つ以上のラベルを付与するマルチラベルの設定を採用した。各感情ラベルの出現頻度を図 7 に示す。220 件のラベル付きデータのうち、200 件を訓練、20 件をテスト用として用いた。また、これらとは別に 550 件のラベルなしデータも実験において用いた。

提案手法と既存手法による各感情ラベルに対する ROC-AUC の値と平均を図 8 に示す。ROC-AUC は、横軸を擬陽性率、縦軸を感度として分類器の閾値を変えてプロットした ROC 曲線の下の面積であり、正例と負例の割合が不均衡な場合にも分類器の能力の比較に用いることができる。提案手法は基本的なもの(文書無)の他に、一度もラベル付けをされていないラベルなしデータを訓練に用いる設定（UL）と訓練時に予測対象とする記事を参照するトランスダクティブ学習の設定（文書有）も用いた。感情ラベル間で差がみられるが、平均的には提案手法が既存手法を上回っていることが分かる。今回はラベルなしデータを用いた効果は確認できなかったが、その理由としては、ラベルなしデータの数が比較的少なかったことが考えられる。

	怒り	悲しみ	喜び	嫌悪	驚き	恐怖	平均
PC	0.569	0.539	0.635	0.641	0.563	0.599	0.591
IPC	0.556	0.578	0.565	0.604	0.584	0.554	0.573
提案（文書有）	0.655	0.669	0.564	0.581	0.669	0.665	0.634
提案ＵＬ（文書有）	0.644	0.688	0.602	0.577	0.634	0.676	0.637
提案（文書無）	0.650	0.685	0.570	0.597	0.675	0.661	0.640
提案ＵＬ（文書無）	0.613	0.687	0.559	0.567	0.643	0.683	0.625

図 8 感情ラベル推定の ROC-AUC

【参考文献】

[鹿島他, 2016] 鹿島久嗣, 小山聡, 馬場雪乃: ヒューマンコンピュテーションとクラウドソーシング（機械学習プロフェッショナルシリーズ), 講談社, 2016.

[作村他, 2014] 作村建紀, 徳永正和, 廣瀬英雄: EM タイプ IRT による不完全マトリクスの完全化とその応用, 情報処理学会論文誌数理モデル化とその応用, Vol.7, No.2, pp.17-26, 2014.

[豊田, 2005] 豊田秀樹: 項目反応理論・理論編—テストの数理（統計ライブラリー), 朝倉書店, 2005.

[Atarashi et al., 2018] Kyohei Atarashi, Satoshi Oyama, and Masahito Kurihara: Semi-supervised Learning from Crowds Using Deep Generative Models, In Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI 2018), pp.1555-1562, 2018.

[Blei et al., 2003] David M Blei, Andrew Y Ng, and Michael I Jordan: Latent Dirichlet Allocation. Journal of Machine Learning Research, 3(Jan):993–1022, 2003.

[Dawid & Skene, 1979] A. P. Dawid and A. M. Skene: Maximum Likelihood Estimation of Observer Error-rates Using the EM Algorithm. Journal of the Royal Statistical Society. Series C (Applied Statics), 28(1):20–28, 1979.
[Duan et al., 2014] Lei Duan, Satoshi Oyama, Haruhiko Sato, and Masahito Kurihara. Separate or Joint? Estimation of Multiple Labels from Crowdsourced Annotations. Expert Systems with Applications, 41(13):5723–5732, 2014.
[Kajino et al., 2012] Hiroshi Kajino, Yuta Tsuboi, and Hisashi Kashima. A Convex Formulation for Learning from Crowds. In Proceedings of the 26th AAAI Conference on Artificial Intelligence (AAAI 2012), pp. 73–79, 2012.
[Kawahara et al., 2017] Jun Kawahara, Takeru Inoue, Hiroaki Iwashita, and Shin-ichi Minato: Frontier-Based Search for Enumerating All Constrained Subgraphs with Compressed Representation, IEICE Transactions on Fundamentals of Electronics Communications and Computer Sciences E100.A(9):1773-1784, 2017.
[Lipovetsky & Conklin, 2001] Stan Lipovetsky and Michael Conklin: Analysis of Regression in Game Theory Approach, Applied Stochastic Models in Business and Industry 17(4):319 – 330, 2001.
[Nash et al., 2012] John F. Nash Jr., Rosemarie Nagel, Axel Ockenfels, and Reinhard Selten: The Agencies Method for Coalition Formation in Experimental Games, Proceedings of the National Academy of Sciences, 109(50):20358-20363, 2012.
[Oyama et al., 2013] Satoshi Oyama, Yukino Baba, Yuko Sakurai, and Hisashi Kashima: Accurate Integration of Crowdsourced Labels Using Workers' Self-Reported Confidence Scores, In Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI 2013), pp.2554-2560, 2013.
[Ptaszynski et al, 2013] Michal Ptaszynski, Hiroaki Dokoshi, Satoshi Oyama, Rafal Rzepka, Masahito Kurihara, Kenji Araki, and Yoshio Momouchi. Affect Analysis in Context of Characters in Narratives. Expert Systems with Applications, 40(1):168–176, 2013.
[Shapley, 1953] Lloyd S. Shapley: A value for N-person Games, Contributions to the Theory of Games, 2 (28), 307-317, 1953.
[Zhou et al., 2012] Dengyong Zhou, Sumit Basu, Yi Mao and John C. Platt: Learning from the Wisdom of Crowds by Minimax Entropy, In Advances in Neural Information Processing Systems 25 (NIPS 2012), 2012.

〈発　表　資　料〉

題　　名	掲載誌・学会名等	発表年月
項目反応理論に基づく魔方陣パズルのユーザー適応的自動生成	第 11 回データ工学と情報マネジメントに関するフォーラム(DEIM Forum 2019)論文集	2019 年 3 月
Aggregating Crowd Opinions Using Shapley Value Regression	Proceedings of the 12th Multi-Disciplinary International Conference on Artificial Intelligence (MIWAI 2018)	2018 年 11 月
Analysis of Coalition Formation in Cooperative Games Using Crowdsourcing and Machine Learning	Proceedings of the 32nd Australasian Joint Conference on Artificial Intelligence (AI 2019)	2019 年 12 月
A Personalized Affect Response Model for Online News Articles	Proceedings of the Fifth Linguistic and Cognitive Approaches To Dialog Agents Workshop (LaCATODA 2019)	20019 年 8 月