デジタルプラクティス　Vol.9 No.4 (Oct. 2018)

特集招待論文

クラウドソーシングによる名刺データ化プロセスの実践

髙橋 寛治[1]　糟谷 勇児[1]　真鍋 友則[1]　中野 良則[1]　吉村 皐亮[1]　常樂 諭[1]

[1]Sansan（株）

Sansan（株）はクラウド名刺管理サービスを提供している．現在のデータ化精度は99.9%であり，ビジネスユースに耐え得る名刺読み取り精度を実現している．OCRのみではこの精度を実現できず，クラウドソーシングを活用することで高精度と低コストを実現している．本稿では，高精度かつ低コストなデータ化のためのクラウドソーシングの取り組み事例を紹介する．具体的には，（1）スパムワーカと疑われるワーカに対して，警告文を表示することで入力精度を89.4%から91.1%（入力ワーク時）に向上することができること，（2）報酬を2，3倍にした際の作業量の増加率が，それぞれ13.7%，31.8%（選択ワーク時）と必ずしも2，3倍にはならないことが分かったこと，（3）ワークの完了条件を2人のワーカの結果がマッチした時点と3人のワーカの結果がマッチした時点で変えた際に，入力精度に大きな差が見られなかったことなどを報告する．これらは既存の研究で報告された内容から逸脱するものではないが，実際の事業での応用において具体的な数値を元に報告したものとして有用な事例研究である．

1. はじめに

近年，人工知能技術は急速に成長しているが，コンピュータのみを用いた解決が困難な問題は多い．たとえば，レシートや郵便番号の読み取りといった高精度な文字起こしや構造化が求められる場面では，現状のOCR（Optical Character Recognition）や項目名推定の精度は追いついていない．決して精度が低いわけではないが，これらのビジネスを運営する上では1 文字の誤りでも許容し難い．この厳しい要件を解決する方法の1つに，クラウドソーシング[1]がある．

クラウドソーシングとは，コンピュータ技術では解決が困難な課題を人と組み合わせることで解決を狙うという考え方であるヒューマンコンピュテーションを，不特定多数の人にインターネット越しに仕事を依頼する仕組みのことである[2]．ヒューマンコンピュテーションを活用するために会社専属のワーカを雇う場合は，タスク供給の過不足の調整は困難な場合が多い．対して，クラウドソーシングはスケーラブルに対応可能である．このような特性から，データ化や機械学習用のアノテーションデータを作成するために活用されている[3]．

本稿では，当社におけるクラウドソーシングによる名刺データ化プロセスの実践について報告する．当社は第2章で述べるように名刺管理サービスを提供しており，高精度かつ高速なデータ化が事業の継続に必要不可欠である．創業以来，高精度に強みを持っており，人力による入力を活用している．大量に入力するために，自社で雇用するオペレータ以外に，クラウドソーシングを活用している．第3章ではクラウドソーシングのためのプラットフォームを示し，また，第4章ではクラウドソーシングで問題となる品質管理についての取り組み事例を紹介する．最後に，第5章では実運用の中で見えた課題について提起し検討を行う．クラウドソーシングの活用に本稿が役立てば幸いである．

2. 名刺管理サービスの概要

当社はビジネスの出会いを資産に変え，働き方を革新することをミッションとし，具体的には名刺サービスをクラウドアプリケーションとして提供する企業である．ユーザが名刺をカメラで撮影したり，スキャナで取り込んだりすると，画像が当社に送られ，複数の工程を経てデータ化される．データ化された名刺はクラウド上のサービスで閲覧，管理，活用することが可能となる．当社には，法人向けクラウド名刺管理サービスのSansan と個人向け名刺アプリのEightの2つの事業がある．SansanとEightのサービスはそれぞれのサービスを開発・運用するSansan事業部，Eight事業部によって運営されているが，名刺のデータ化からそのデータの分析・活用までのデータにまつわる処理をData Strategy ＆ Operation Center（以下DSOC）という部署が統括している．名刺のデータ化に求められるものとして，Sansanサービスはデータ化の精度とセキュリティを優先し，Eightサービスは低コストで大量に処理することを優先しているなど違いはあるが，どちらのサービスも大まかなフローは下記のようになる．

（1）画像を細切れのピースと呼ばれる単位の小画像に分割する
（2）ピースが名刺のどの項目か（会社名か，氏名か，住所かなど）を判定する（“選択”と呼ばれる工程）
（3）各ピースについて項目ごとのルールに基づいてテキスト化する（“入力”と呼ばれる工程）
（4）入力結果をマージし，1つの名刺として完成させる
（5）マージした結果をチェックし，必要なものについてはオペレータによって再入力される

この“入力”と“選択”において，ピースは画像処理や機械学習の技術を用いて自動で一次判定され，自動判定の信頼度が一定以下の場合にクラウドソーシングでそれぞれの工程を行う．信頼度は複数の判定エンジンを用いることで，その多数決によって算出する．以下，本稿では“入力”と“選択”工程をクラウドソーシングを用いて行うことについての知見について記述する．DSOCはこれまで数億枚の名刺をデータ化してきた実績があり，クラウドソーシングで処理したタスク数は数百億にのぼる．この規模でクラウドソーシングを事業に活用している例は国内では他に比肩する例はなく，その中で行われてきたさまざまな実験による知見は事例研究として高い価値があると考えている．

クラウド名刺サービスの価値を支える高速かつ高精度な大量のデータ化は，クラウドソーシングにより実現される．

3. クラウドソーシングによる名刺データ化

3.1 タスク設計における前提

名刺のデータ化は1文字の誤りが含まれると，サービスに適用できない．たとえば，E-mailアドレス中のアルファベットの1文字が間違っているとメールが届かなくなるからである．また，対象が名刺であるため，なるべく早くデータ化されないと記載されている連絡先情報を利用できない．すなわち，名刺管理サービスを提供する上では，高精度かつ高速なデータ化が必要不可欠である．

実サービスに耐え得る高精度かつ高速なデータ化を実現するべく，OCRや画像処理といった情報処理技術とワーカによる入力・選択を組み合わせることで，現在のデータ化システムの精度は99.9%に達し，月間数千万枚以上の名刺をデータ化している．我々は名刺データ化システムにおいて，以下の4つの軸を重要視している．これらの要素の頭文字を取り，名刺データ化システムをGEESと呼んでいる．

Global　より世界中のワーカに依頼できる
Elastic　繁忙／閑散期を柔軟に吸収できる伸縮自在な体制を組める
Efficient　より効果的な方法で役割配置できる
Scalable　よりスケーリングできる

以上のGEES の前提を踏まえた，名刺データ化フローとクラウドソーシングの活用について紹介する．

3.2　名刺データ化フロー概観

名刺のデータ化とは，撮影もしくはスキャンされた画像から，氏名や会社名といった項目ごとに文字列をデータベースに格納することである．名刺画像の例を**図1**に，対応する入力項目とその内容を**表1**に示す．図1の例は，スキャナで取り込まれた画像であるため，読み取りやすい例である．可能であれば，実際に名刺を見て入力し，データ化の難しさや手間を体感されてから，次に進まれたい．

図1　名刺（架空の名刺）

表1　図1の架空の名刺に対する項目と入力内容

項目	入力内容
氏名	田中 浩介
会社名・団体名	株式会社 C&S
部署	東京本社 企画制作部
役職	マネージャー
〒	162-0855
住所	東京都新宿区二十騎町 2-12-15
TEL	03-3409-3133
FAX	
携帯	
Email	swk_nep628@docomo.ne.jq
URL	www.c-and-s.com
上記項目に該当なし	

名刺の取り込みからデータ化のフローを**図2**に示す．スマートフォンで撮影もしくはスキャンされた名刺画像（1. 名刺取り込み）は，背景から名刺画像をトリミング（2. 背景分離）した後に，影の除去や文字を鮮明にさせる画像補正（3. 画像補正）が適用される．これは，スマートフォンで撮影された画像は，フラッシュによる白飛びや暗くて視認性が悪いものなど，ワーカの作業効率を妨げるものが含まれているからである．次に，深層学習を用いた手法により，名刺内テキストについて項目ごとの分割および項目名の分類（4. 項目分割）を行う．さらに項目を小さくする（5. セキュリティ項目細分割）ことで，ワーカが作業しやすくなる．細切れにされた項目の入力（6. 項目入力） をワーカに依頼し，その結果を各項目のデータと統合する（7. マージ）．最後に，機械学習やワーカによる目視を組み合わせることで入力項目のチェックや補正を行う（8. チェック&補正）．

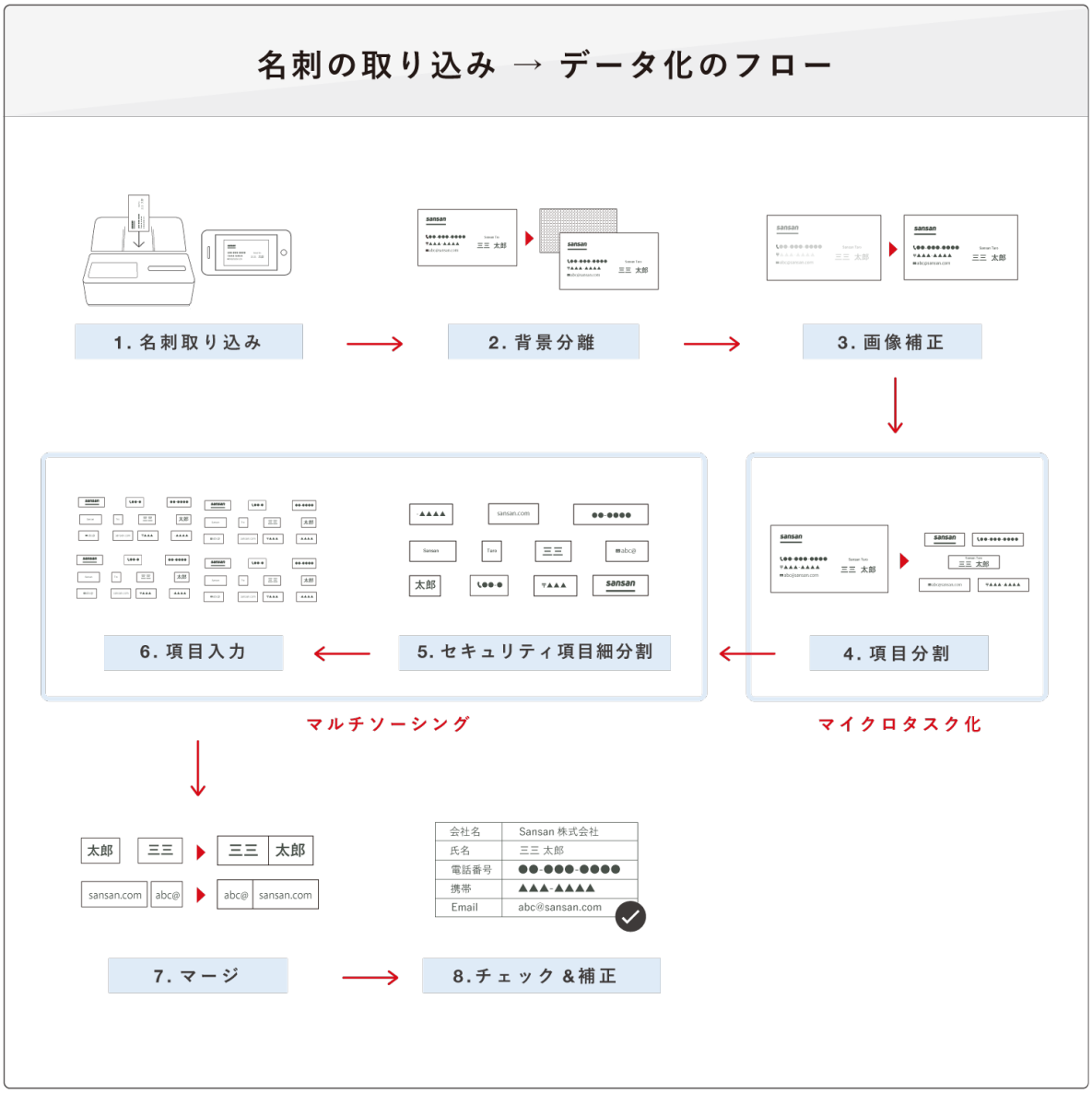

図2　名刺の取り込みからデータ化までのフロー

データ化フローにおいてクラウドソーシングを用いるのは，“入力”と“選択”工程である．入力とは，ワーカが画面に表示されるピース画像を見たまま入力するタスクである．選択とは，ワーカが画面に表示されるピース画像を見て，ピースが名刺のどの項目かを判定するタスクである．

3.3　マイクロタスク×マルチソーシングによるワーカの適材適所と個人情報の保護

名刺に含まれる情報は個人情報であるため，名刺をそのまま不特定多数のワーカに渡してデータ化を依頼することは不適切である．したがって，クラウドソーシングで依頼するタスクに個人情報が含まれる場合のタスク設計は，クラウドソーシングをビジネスに利用する際に避けては通れない問題である．こうしたタスク内の個人情報保護に着目した研究は，インスタンスプライバシ保護クラウドソーシング[4] と呼ばれ，当社でも画像処理を活かしたマイクロタスクとマルチソーシングの組合せにより，個人情報を保護しながら不特定多数のワーカへのタスク依頼を可能としている．

まず，マイクロタスク化について説明する．当社の名刺のデータ化（図2）の各処理工程の中でも，特に，（4. 項目分割）と（5. セキュリティ項目細分割）の2つは，個人情報保護と不特定多数のワーカへの作業依頼の両立の上で非常に重要な工程といえる．項目分割処理により，名刺

画像内で入力が必要な文字列が項目ごとに分割される．セキュリティ項目細分割処理により，氏名やE-mail，携帯電話番号などの個人情報を含む文字列は，情報として価値がない大きさに画像が分割される．すなわち，個人情報を保護しながらワーカへの作業の依頼が可能となる☆1．一般に，個人情報保護（プライバシ保護） 施策を強くすると，成果物の品質は低下すると考えられているが，当社が依頼しているタスク内容の場合には，読み取る対象にフォーカスされるため精度向上にも貢献していると考えられる．また，タスクが細分化されているため，一つひとつのタスクに対する作業量が減少しワーカの負担が軽減されるという利点がある．

このようにマイクロタスクとマルチソーシングを組み合わせることにより，全項目を入力していた当初と比べて個人情報を保護しながらコストを約1/7に抑えている．マイクロタスク化することの利点は，誰でも，いつでも，どこでも， 作業ができることでもである．たとえば，電話番号やE-mailはアルファベットと数値で構成されているため，日本語の読み書きができるワーカである必要がなくなる．また，一つひとつのタスクを行う際の入力文字数が少なくなるため，ワーカが少しの時間で取り組むことができる．

次に，マルチソーシングについて説明する．当社では，より多くのタスクを処理するためにワーカ数を増やすための，マルチソーシングと呼んでいる方法を名刺のデータ化プロセスの途中に採用している．ここでいうマルチソーシングとは，さまざまな媒体のワーカに作業依頼をすることである．当社の名刺データ化は，下記に示す3種類のワーカに支えられており，全世界で毎月数十万人のワーカが名刺のデータ化に日々取り組んでいる．

センターオペレータ
: 当社が雇用するワーカである．継続して教育できるため品質が高いがコストも高い．安定してタスクを処理することができる．

BPOオペレータ
: 当社と契約する海外を含めたビジネス・プロセス・アウトソーシングのワーカである☆2．教育が可能で，品質も良好である．

クラウドソーシングワーカ
: 当社が提供するプラットフォームで入力を行う不特定多数のワーカである．教育ができないため品質は少し下がるが，スケーラブルでコストが低いことが特徴である．

3.4　クラウドソーシングプラットフォームの紹介

“入力”と“選択”タスクのクラウドソーシングのための実際の作業プラットフォームを紹介する．

当社では，ワーカがいつでもどこでも作業できるように，PCとスマートフォンの両方で作業プラットフォームを提供している．ワーカはまず，“入力”か“選択”のどちらのタスクに取り組むかを決めて作業に取りかかるため， ワーカは選択した一方のタスクにのみ集中して取り組むことができる．タスクの終了については，ワーカが任意のタイミングで終了するか，タスクの枯渇が条件となる．

実際の作業プラットフォームのPC向け（**図3**，**図4**）とスマートフォン向け（**図5**）のスクリーンショットを例示する．図5に示すように， 一つひとつのタスクがマイクロタスク化されていることで， PCとスマートフォンのいずれにおいても，容易に作業できることが分かる．

図3　PC版の入力作業プラットフォームの画面

図4　PC版の選択作業プラットフォームの画面

図5　スマートフォン向けの作業プラットフォームの画面

PC版では，1日単位で取り組んだタスク数および，有効タスク数，順位を表示している．1つのタスクに対して複数回答を収集し，早い順で同じ回答が2つ得られた場合に確定としている（以降，この認定方法をマッチと呼ぶ）．このマッチしたタスクを有効タスクと呼ぶ．マッチした場合に報酬が支払われる．

3.5　運用上の工夫例

本節では，当社でのクラウドソーシングの運用上の工夫の一例を3つの文脈に分けて紹介する．

3.5.1　回答の精度向上のためのタスク設計

クラウドソーシングで得られるワーカの回答の品質を向上させるための1つの手段として，適切なタスク設定をすることが挙げられる．当社でも，ワーカの回答の精度向上のために，タスク設計に十分に気を配っており，その中の1つの方法として事前説明の徹底と，練習項目の実施をしている．具体的には，各ワーカが実際のタスクを行う前に必ず1日に1度事前説明を行い，練習項目に取り組んでもらっている．これにより，もしほかのクラウドソーシングタスクを兼業しているワーカがいたとしてもタスク内容の勘違いや，回答方法の勘違いが防げると考えられる．

実際，選択項目を追加した際に，事前説明と練習項目に当該項目を含めただけではワーカの誤りが散見された．追加した選択項目の練習問題をほかの練習問題と比べて3倍ほど導入した場合に，誤りが減少した．

3.5.2　ワーカのモチベーション担保

マイクロタスク型のクラウドソーシングタスクでは，基本的には単純作業を繰り返すことになるので，モチベーションを保つのが難しいという問題がある．ワーカのモチベーションが低下してくると，タスクに対する回答の精度が低下するどころか，最悪の場合ではワーカが作業をやめることに繋がってしまう．そこで，当社ではワーカのモチベーションを上げるための手段として，正確にタスクを遂行した回数に関するランキングを導入している．このランキングに基づいて，日々の上位者を決定し，上位者に対しては追加報酬が与えらえる．この方法により，真面目に取り組むことでより多くの報酬が貰えるというモチベーションが生まれると考えられる．

ランキングを導入した際に，ワーカから本制度の継続希望やランキングを週次や月次に変更できないかと，前向きな意見が寄せられた．意見を生むほどワーカのモチベーション維持に寄与していると考えられる．

3.5.3　ツール利用対策

当社では機械のみによるデータ化では十分な精度での名刺のデータ化は不可能と考えており，その機械では適切に処理できない部分を人手に頼るためにクラウドソーシングを利用している．しかしながら，ワーカの中には一部ではあるが，映し出された名刺の一部に対してOCR技術などを用いて入力タスクをこなすワーカも存在する．このようなワーカが増えてくると，当然クラウドソーシングを用いる意味がなくなってしまう．そこで，当社では，ツールの利用対策として，マッチ率が特定の値を下回るワーカに対してreCAPTCHA[5]を提示するという措置も施している．これにより，ツールの利用者を検知することができる．合わせて，OCRツールの利用自体を防ぐための方策として，提示する名刺の一部の画像に対して特定の処理を適用し，OCRツールによる読み取りが困難になるようにしている．

処理タスク数にある閾値を定めて，閾値以上のワーカに対してreCAPTCHAを適用し，効果を検証した．大量のタスクを処理しているワーカは，実際にタスクをこなしている場合とOCRツールの利用による機械的な処理の2つに大まかに分けられる．ここで，高い集中力を維持したワーカにreCAPTCHAが適用されないよう，いくつかの要因をもとに閾値を設定した．reCAPTCHAを適用した結果，大量にタスクを処理しているワーカの精度改善が見られたため，高精度を担保するという目的は達成できた．

4. パフォーマンス改善

本章では，名刺のデータ化を行う際の3つの課題についての説明を行い，各課題を解決するために行った実験とその結果を示すとともに，検証した内容について報告する．

4.1　グレーワーカ対策

「クラウドソーシング」では不正ユーザの存在が，精度を下げコストを上げる要因としてしばしば問題になる．本ケースにおいて想定される不正ユーザはOCR 技術を用いて自動処理を行うユーザや，不誠実な回答を繰り返すユーザであり，これらのユーザによるタスク処理は人手を介した精度向上に寄与しないため，このようなユーザをなるべく排除する必要がある．

排除するためにはまず，不正ユーザを検知する必要がある．我々は，次の数量が平均と乖離していた場合に不正ユーザ，すなわちスパムワーカの可能性が高いとし，「グレーワーカ」と定義してマークを行った．基準とした数量は，単位時間の処理件数の多さ，正答率の低さ，選択カテゴリの偏り，作業継続時間の長さなどである．処理件数が平均と乖離して過剰に多い場合は，自

動処理を行っていたり，正しく入力・選択をせずに先に進めていたりする可能性が強く，また正答率が圧倒的に低かったり常に同じ選択肢ばかりを選ぶユーザも，同種の不正を行っている可能性が高い．ただし，不正と不正でないユーザを既知としたデータが存在するわけではないため，不正と断定してユーザを排除するのは困難であったことから，これらグレーワーカに対し，警告文（**図6**）を出現させることで改善が見込まれるかどうかの検証を行った．

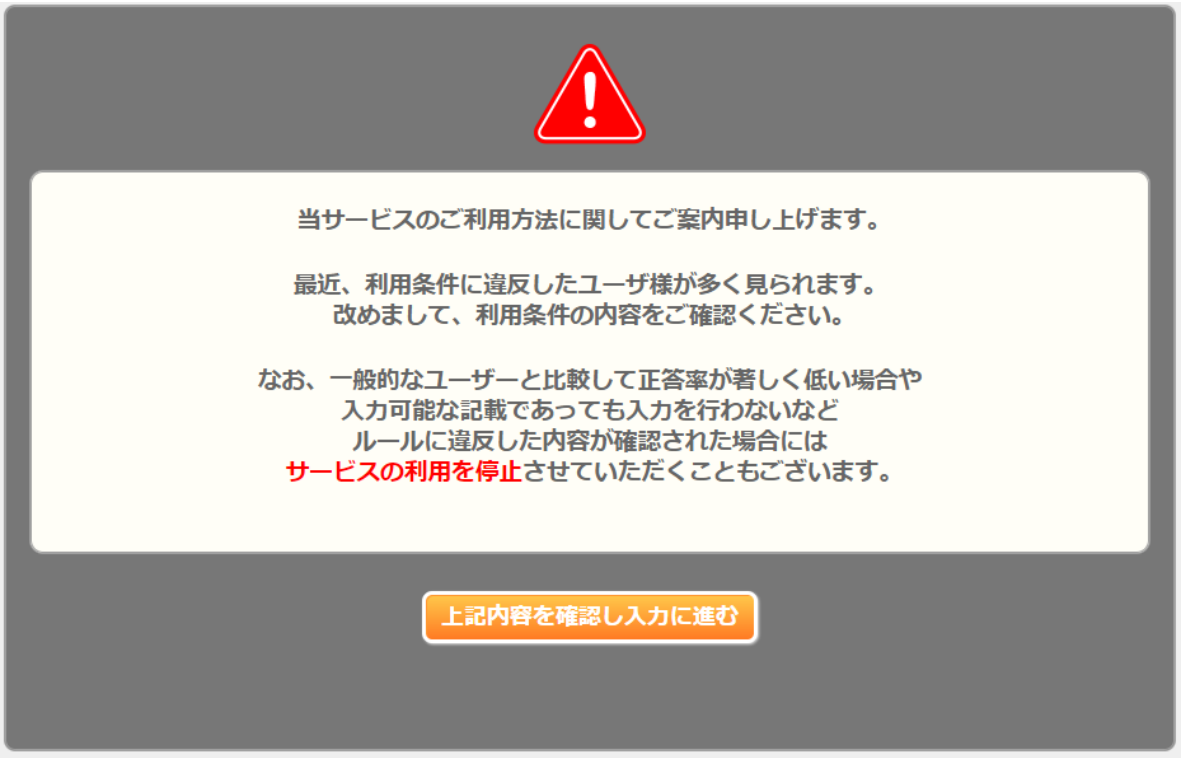

図6　グレーワーカに提示した警告文

警告文提示を受けたグレーワーカ2,585人のうち，579人の離脱が見られた．すなわちグレーワーカの23%程度が離脱し，77%が残存したが，警告文を提示する3カ月前と提示した3カ月後で比較すると，全体精度で1.7ポイントの精度向上があり，一定の効果が見られた．さらに警告文提示後も残存したグレーワーカから適当数をサンプリングし，それらの精度の変化について確認したところ，入力タスクで12.6ポイント選択タスクで8.3ポイントの精度向上が見られており，グレーワーカの態度改善に対しても，警告文の効果を確認することができた（**表2**）．

表2　警告文による精度変化

警告文	全体精度 (入力)	グレーワーカの精度 (入力)	グレーワーカの精度 (選択)
提示前	89.4 %	62.8 %	44.8 %
提示後	91.1 %	75.4 %	53.1 %

当初の想定ではスパムワーカは警告文による不正検知によって離脱のみを起こすものと考えており，継続および態度の改善はそれほど期待していなかったが，このような成果が見られたことから，警告文にスパムワーカの態度を正し，善良なワーカへ戻すための効果があることが分かった．また，想定よりも離脱者数が少なかったため，作業量への影響も小さく抑えることができた．

4.2　作業単価と処理量の関係

膨大な量の名刺をデータ化する際に発生する問題の1つとして，急激なタスクの増加に対して労働力の供給が追いつかなくなることが挙げられる．当社の名刺のデータ化プロセスの途中には人力による作業が必須となっているため，労働力の不足は名刺のデータ化の遅延に繋がることになる．この問題に対処するべく需要に対して十分な量の労働力の供給を得るためには，今まで以上の数のワーカが作業に取り組むか，既存ワーカの作業量を増やす必要がある．そこで当社は，労働力の供給を増やすため，ワーカの作業に対するモチベーション向上を意図して，作業単価を増加させた場合に労働力にどのような変化が見られるかを検証する実験を行った．

この実験では，実際の名刺のデータ化プロセスの中で，選択タスクについて，ある一定期間作業単価を増加させた場合のワーカ数とマッチ数の変化を調べた．具体的には，作業単価を元の単価の2倍，3倍に増加させて，ワーカ数とマッチ数の変化を確認した．その実験の結果を**表3**に示す．この実験の結果から，作業単価を高くしてもワーカの人数は増加しないという結果が得られた．これは，作業単価を上げたことの周知方法が適切ではなかった可能性がある．一方で，作業結果のマッチ数は大幅に増加していることが分かった．この理由として，次の2つの可能性があると考えられる．1つがワーカの作業量が増えた可能性，もう1つがワーカがタスクに対してそれ以前より真面目に取り組むようになり，正解率が上昇した可能性である．しかし，元々総作業数の9割程度がマッチすることを考えると，後者の可能性の影響は低いと考えられるため，ワーカの作業量が増えたと取るのが自然である．このことから，作業単価を高くすることによってワーカ数自体が増えることはないが，作業を行うワーカ一人ひとりの作業量は増加するといえる．しかしながら，作業単価を3倍にした場合でもマッチ数は高々30%程度しか増加していないため，費用対効果を考えると作業単価を高くして労働力を増やすことは実用には向かないと考えられる．

表3　選択タスクにおける作業単価の変化に対する1 日あたりのユーザ数とマッチ数の変化

作業単価	ワーカ数/日	ワーカ増加率	マッチ数/日	マッチ増加率
×1	7,256	-	2,278,634	-
×2	6,776	−6.615%	2,590,902	13.704%
×3	6,891	−5.030%	3,003,630	31.817%

本実験とおおむね同様の実験が，すでに2010年にHortonらによって行われている[6]．実験に用いられたタスクの内容は異なるが，賃金とワーカの作業との関係を明らかにする実験という意味では同じといえる．[6]によれば，賃金が低いほど早期に作業を辞めることが多くなるという主張がなされている．これは，裏を返せば賃金が高いほど作業時間が長くなるということにほかならない．今回の当社の実験結果からは，作業単価を上げた場合にマッチ数が増加していることからワーカの作業量が増加したと考えられ，この結果はHortonらが主張している内容と合致している．Hortonらによる実験の追試という意味でも本実験の社会的貢献は大きいといえる．

4.3　精度を担保するための冗長化

クラウドソーシングにおいてはクオリティ向上のために複数ワーカの回答を統合することが一般的であり，当社でも第3章で述べたように複数ワーカで回答が一致するまで冗長なタスクを依頼している．我々はサービスレベルとして99.9%という高い精度を掲げており，より多くのワー

カを用いてでもクオリティを改善したいため，クラウドソーシングのタスク完了のために何件のマッチを用いるのがよいか検討した．

実験は2週間にわたり「選択」タスクについて実施されており，前の1週では2ワーカのマッチが必要で，後の1週では3ワーカのマッチが必要な設定に切り替える．精度検証として，それぞれの期間にデータ化された項目に対するクオリティチェックを実施した．複数ワーカでマッチした項目と社員が付与した項目が異なる場合にミス判定となっている．各期間にデータ化された1,000ピースを抽出してチェックした結果を**表4**に示しているが，2マッチと3マッチでクオリティに大きな差異は見られなかった．

表4　必要マッチ数の精度への影響

必要マッチ数	チェック数	ミス件数	ミス割合
2 マッチ	1000	37	3.7%
3 マッチ	1000	35	3.5%

これは，1ワーカの回答の精度が高い場合に冗長性による精度改善幅が小さい[7] というよく知られた性質とも整合的な結果である．ここでは[7]にならって必要マッチ数と精度の関係について簡易な解析をする．1ワーカの正解率を p，必要マッチ数を $N+1$ とすると複数ワーカでマッチした項目の精度 q は以下の式で概算できる☆3．

$$q \sim \sum_{i=1}^{N} \binom{2N+1}{i} p^{2N+1-i}(1-p)^{i}$$

必要マッチ数と精度の関係を正解率ごとにプロットしたものが**図7**である．「選択」タスクについて1ワーカの正解率は90%強の水準であるから，1マッチから2マッチへの変更では7ポイントの精度改善が見込まれる一方，2マッチから3マッチへの変更では1ポイント程度の精度改善しか見込まれず，数理的にも実験結果を裏付けることができる．

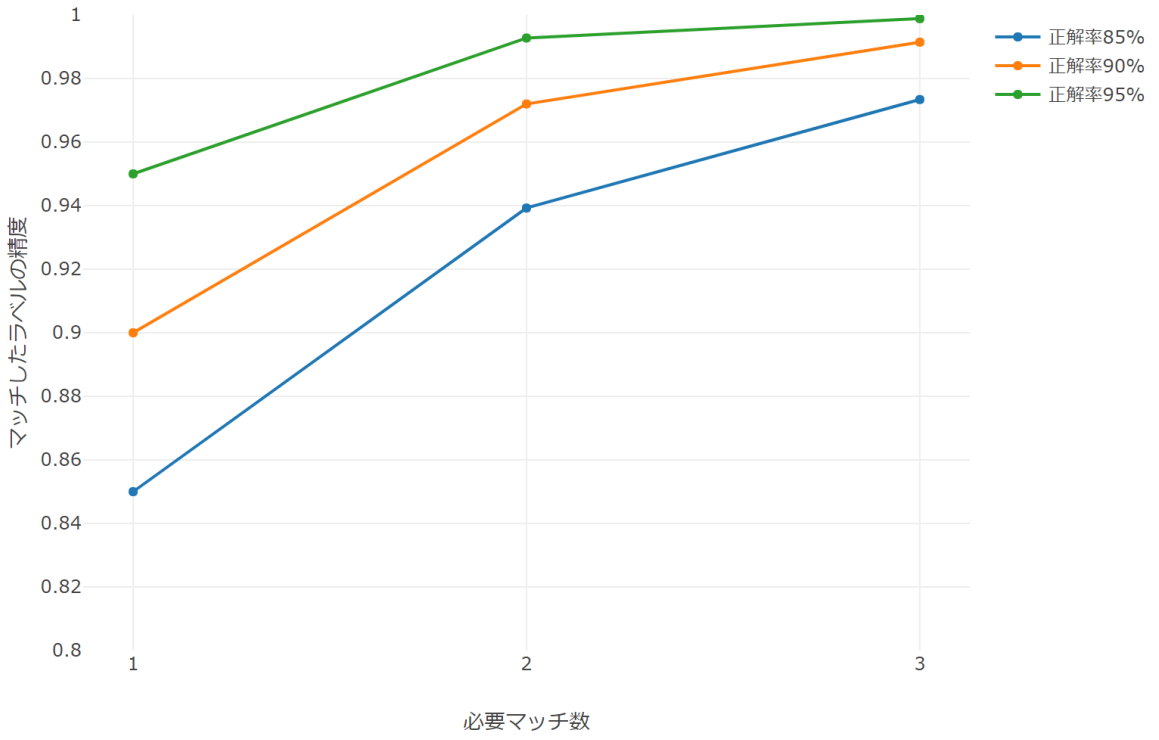

図7　必要マッチ数と理論精度

必要マッチ数を多くすることは，報酬の発生するワーカ増加によるコスト増加，冗長なタスク増加による処理量減少，さらに納品時間延長というデメリットがあるため，当社では2マッチを採用している．

5. 今後の課題

人工知能技術の業務課題への応用は昨今必須となりつつあるが，一般に業務工程をすべて人工知能技術で置換することは難しく，人手による作業と，工程を切り分けて運用する必要が生じる．工程を適切に切り分けるためには，人工知能と人の力の作業適性の違いを十分に把握している必要があり，最適に工程を分離することでコストパフォーマンスを最大化することができる．当社は画像認識技術と人力工程を分離・再統合するシステムを作成し，クオリティとコスト低減の両立を目標として改良を続けてきた．本稿で述べてきたように，人力工程にクラウドソーシングを利用することで，クオリティを担保するための人の特性を生かしつつ，かつコストを抑えるためのシステム作りを達成できたことは，大きな成果であったと考えている．今後，さらに精度を向上させつつコストを下げるためにクラウドソーシングを活用していく際の課題として，次のような事項が挙げられる．

クラウドソーシングでは，センターオペレータなどの専門職員と比べ，技術や学習の向上が見られにくい．これはワーカの継続性が低いことに加え，向上するためのモチベーションが欠けていることも要因と考えられる．単純な報酬アップなどでは継続や学習向上の施策としては限界があることが，本稿での調査を通して明らかになっているため，継続・学習を促すための施策の考案が必要であるとともに，一方で，そのようなクラウドソーシングの特性に合わせ，作業内容の方をより最適化することも，合わせて考える必要がある．すなわち，なるべく作業者の学習コストが少なくなるように，作業を簡便なものに切り分けていく工夫などである．

誰でも簡便にワークを開始できるということは同時に離脱もしやすいということであり，クラウドソーシングの，労働流動性が高いというこの特徴は労働供給の不安定さにつながっている．この不安定さを解消するためにも，継続ワーカを一定数確保するための施策は重要である．具体的には，継続しているワーカと離脱したワーカの特性や正解率，作業時間の違いなどを分析した上で，離脱ワーカを継続ワーカに近づけていくための工夫が必要になると考えている．

そのほかの課題として，現在当社がグローバル展開を進める中で，多言語入力の必要に迫られていることが挙げられる．そのためには海外へのクラウドソーシングをより積極的に展開していく必要があるが，ワーカの特性が国や文化に依存して異なるため，成功事例をそのまま持ち込んでもうまくいかないことが多いことに加え，昨今では特にアジア圏でPCを持っておらず，スマートフォンのみを保有しているようなワーカが多いため，タスクをスマートフォンでも簡便にできるように，画面設計等を一層工夫する必要がある．また，アジア圏ではインターネット回線が低速かつ不安定であるため，そのような環境下でも安定して機能するシステムを設計する必要性もある．

これらの課題を解決することで，より一層クラウドソーシングを活用できると考えている．

6. おわりに

セキュリティと利便性を両立させて，クラウドソーシングを活用しよう．

名刺管理サービスにおける高精度なデータ化を実現するためのクラウドソーシングの取り組みについて，事例紹介を行った．さらに，クラウドソーシングのパフォーマンス改善のための施策とその効果についての考察を紹介し，今後の課題を提示した．帳票や伝票など高精度な文字のデータ化が求められる場に，本稿で紹介した事例が参考になれば幸いである．

謝辞　本稿の記載内容は，Sansan（株）Data Strategy & Operation Centerの名刺データ化におけるクラウドソーシングの実践について活動成果をまとめたものであり，Data Strategy & Operation Centerの皆様を始めとして，名刺データ化にかかわる皆様に深謝いたします．

参考文献

1）Howe, J. : The Rise of Crowdsourcing, Vol.14, No.6, Wired Magazine (2006).
2）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，講談社 (2016).
3) Callison-Burch，C. and Dredze，M. : Creating Speech And Language Data with Amazon's Mechanical Turk， Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's MechanicalTurk, Association for Computational Linguistics, pp.1-12 (2010).
4 ） Kajino，H.，Baba ， Y. and Kashima，H. : Instance-Privacy Preserving Crowdsourcing，Second AAAI Conference on Human Computation and Crowdsourcing (2014).
5）Von Ahn, L., Blum, M. and Langford, J. : Telling Humans and Computers Apart Automatically, Commun. ACM, Vol.47，No.2, pp.56-60 (2004).
6）Horton, J. J. and Chilton, L. B. : The Labor Economics of Paid Crowdsourcing, Proceedings of the 11th ACM Con-ference on Electronic Commerce, EC'10, ACM, pp.209-218 (2010).
7）Sheng, V. S., Provost, F. and Ipeirotis, P. G. : Get Another Label? Improving Data

Quality and Data Mining Using Multiple, Noisy Labelers, Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD'08, ACM, pp.614-622 (2008).

脚注

☆1　たとえば，"E-mail:dummy-taro@hoge.com" は "dummy-taro"と"@hoge.com"に細分割されるため，それぞれのピースを見て作業するワーカはE-mail全体を知ることができない.
☆2　障がい者のデータ入力業務としても提供しており，隙間時間とは異なった新たな仕事を生み出している.
☆3　2マッチの精度は，3ワーカの内2ワーカ以上が正解する確率と同程度と見なせる.

髙橋 寛治（非会員）ka.takahashi@sansan.com
長岡技術科学大学大学院工学研究科電気電子情報工学専攻修了．2017年にSansan（株）に入社．自然言語処理に関連する研究開発に従事.

糟谷 勇児（非会員）kasuya@sansan.com
早稲田大学大学院情報ネットワーク専攻修了．2016年にSansan（株）に入社．判別エンジンやリコメンドエンジンの研究開発などに従事.

真鍋 友則（非会員）manabe@sansan.com
東北大学生命科学研究科修了，筑波大学大学院ビジネス科学研究科経営システム科学専攻修了（MBA）．同大学院博士後期課程在籍中．2017年にSansan（株）に入社．名刺データを用いた営業支援ツールの開発に従事.

中野 良則（非会員）nakano@sansan.com
大阪大学大学院経済学研究科博士前期課程修了．2017年にSansan（株）に入社し，リコメンドエンジン開発やデータ分析に従事.

吉村 皐亮（非会員）yoshimura@sansan.com
京都大学大学院情報学研究科知能情報学専攻修了．2018年にSansan（株）に入社．ワーカの性質に基づくタスク割り当てに関する研究開発などに従事.

常樂 諭（非会員）joraku@sansan.com
Sansan（株）の設立に合わせ，創業メンバとして参画．法人向け名刺管理サービス「Sansan」の開発部長・プロダクトマネージャを経て，現在は名刺のデータ化やデータの分析・活用を行うDSOCのセンター長を務める.

採録決定：2018年7月13日
編集担当：澤邉 知子（日本大学）