卒業論文

クラウドソーシングを用いた
対訳辞書作成のためのタスク割り当て手法

指導教官　　村上 陽平 准教授

立命館大学情報理工学部
情報コミュニケーション学科 4 回生
2600160250-0

地田 大樹

2019 年度 秋学期 卒業研究 3(2Q)
令和 2 年 1 月 31 日

クラウドソーシングを用いた
対訳辞書作成のためのタスク割り当て手法

地田 大樹

内容梗概

インドネシアは 700 以上の言語がある多言語国家であるが，国策により公用語のインドネシア語での教育が徹底されたことで，多くの地方語の話者が減少している．特に，消滅の危機に瀕している危機言語は 147 言語にものぼる．このような地方語の話者を増やすために，地方語間でのコミュニケーションを支援するための対訳辞書が必要となる．一方で，クラウドソーシングによって対訳辞書などの言語資源を作成することが主流になりつつある．不特定多数の作業者に作業を依頼するクラウドソーシングでは，作業者の能力にはばらつきがあり，実行結果の品質を保証することが困難であるため，ゴールドタスク（正解既知問題）を用いることにより，能力が高いと推測される作業者のみにタスクを割り当てる方法が用いられる．しかしながら，ゴールドタスクを生成するのは非常に難しく，コストがかかることが知られている．さらに，インドネシアの地方語は話者が少なく，危機言語を含む複数の地方語を話すことができる人はさらに限られるため，能力の高い作業者はあまり見込めない．このような作業者の能力の平均が低い集団では，同じタスクを複数の作業者に割り当て，多数決を用いる方法では，誤った回答を採用する可能性が高いため，品質管理を行うことがうまくできない．

そこで，本研究では，作業結果による作業者の動的な信頼値評価手法を導入する．具体的には，信頼値に応じてタスクの割り当てられやすさを調整し，さらに信頼値がある一定の値よりも高い作業者（一定以上の能力があると推測される作業者）のみが対訳評価タスクを実行可能にする．これにより，ゴールドタスクを用いることなく信頼値の低い作業者（能力が低いと推測される作業者）を徐々に排除していき，高い品質の実行結果を得ることを目指す．本手法の実現にあたり，取り組むべき課題は以下の 2 点である．

作業者とタスクのモデル化

信頼値を求めるための多様なタスク割り当て手法を設計し，それぞれの有用性の検証をするために，ワークフローを構成するタスクと，そのタスク

を実行する作業者のモデルを作成する必要がある．

信頼値の算出

作業者の作業結果から効果的に信頼値を計算する方法を設計する必要がある．加えて，信頼値を用いた，信頼できる作業者を判別するための基準となる閾値の設定と，タスクの割り当てられやすさの調整を行う必要がある．

一つ目の課題に対しては，作業者の能力値をベータ分布を用いて 0 から 0.7 の間で設定した．作業者の能力が高いほど，正しく作業を行うとする．能力値の上限を 1 ではなく 0.7 としたのは，能力の高い作業者は限られているという制約があるためである．そして，自由回答型タスクである対訳作成タスクと，真偽型タスクである対訳評価タスクからなるワークフローとし，タスクの実行結果は作業者の能力によって確率的に決定されるとする．対訳作成タスクによって作成されたある対訳一つあたりに複数回の対訳評価タスクを行い，それらの結果の多数決をとることで，その対訳に対する最終的な評価を決定する．二つ目の課題に対しては，単純化のために，信頼値を正答数と誤答数の差とした．そして，信頼値が 1 以上である作業者を信頼できる作業者とみなし，対訳評価タスクは信頼できる作業者のみに割り当てるようにする．加えて，各作業者の信頼値をもとに，タスク割り当て確率における重みを計算し，これを用いて，タスクの割り当てられやすさの調整を行う．本研究の貢献は以下の通りである．

作業者とタスクのモデル化

作業者の能力値を事前に設定し，タスクの実行結果は作業者の能力値により決定されるようにモデル化を行なった．これにより，様々な手法でのシミュレーションを同じ条件で行うことが可能となった．

信頼値の算出

作業結果に基づく信頼値の計算と，各作業者の信頼値に基づいたタスク割り当て手法の定式化を行った．その結果，信頼値を用いる提案手法では，既存手法の半分以下のゴールドタスク数で，同程度の正確性を得ることに成功した．

Task Assignment for Crowdsourced Bilingual Dictionary Creation

Hiroki CHIDA

Abstract

Indonesia is a multilingual country with more than 700 local languages. However, the number of local language speakers has been decreasing due to the thorough education of Indonesian, the official language of Indonesia, by national policy. In particular, as many as 147 languages are on the verge of extinction. Therefore, the bilingual dictionaries to support communication between these local languages are required to increase the number of speakers. In natural language processing domain, crowdsourcing is becoming mainstream to create language resources including bilingual dictionaries. To assure the quality of crowdsourced output with the various abilities of workers, gold tasks (tasks the answers of which are known) are used to measure a worker's abilities and to allocate tasks to good-quality workers only. However, generating gold tasks is very difficult and costly. In addition, it is difficult to find local language speakers, especially multilingual speakers, which means that there are few high-ability workers. In this situation, controlling quality by assigning the same tasks to multiple workers and using majority vote to select the final result does not work well.

In this paper, we propose a method to dynamically evaluate the reliability of workers based on their work results. Specifically, we decide the probability of task assignment based on the reliability, and only the workers whose reliability are higher than a threshold (the workers whose abilities are expected to be high) can evaluate the bilingual text generated by the other workers. In this way, we aim to gradually eliminate the workers who have low reliability (the workers who are expected low-quality workers) and to get good quality results without using gold tasks. To this end, we address the following two problems.

Modeling workers and tasks

In order to design various task assignment methods for evaluating reliability and to verify the usefulness of each workflow, it is necessary to model the tasks that compose the workflow and the workers who execute the tasks.

Evaluation of reliability

It is necessary to design the method for effectively evaluating the reliability based on the work result. In addition, it is necessary to set a threshold to detect the reliable workers using the reliability and to decide the probability of task assignment based on the reliability

For the first problem, we set the worker's ability value between 0 and 0.7 using beta distribution. The higher the worker's ability, the more correct the work. The upper limit of the ability value was set to 0.7 instead of 1 because of the constraint that the number of workers with high ability is limited. The workflow consists of one translating task, which is an open-ended task, and several evaluating tasks, which are true-false tasks. The results of the task execution are probabilistically determined by the worker's ability. The final evaluation of the bilingual text is determined by allocating multiple evaluation tasks for each bilingual text created by the translation task and taking a majority vote of the results. For the second problem, we define the reliability as the difference between the number of correct answers and the number of incorrect answers for each worker. Then, we set the threshold as 1, so the workers whose reliability are 1 or greater are regarded as reliable workers. We allocate evaluation tasks only to reliable workers. In addition, we calculate the weight in task assignment based on the reliability of each worker, and we adjust the probability of task assignment using the weight.

The contributions of this paper are as follows:

Modeling workers and tasks

We set the workers' abilities beforehand and the result of a task relies on the ability of the worker who executes the task. By using this model, it is possible to simulate a variety of methods under the same conditions.

Evaluation of reliability

We designed the method for evaluating the reliability of workers based on their work results and formulated the task assignment methods based on the reliability. As a result, the proposal method provides the same level of accuracy with less than half number of gold tasks compared to the previous methods.

クラウドソーシングを用いた
対訳辞書作成のためのタスク割り当て手法

目次

第1章　はじめに

インドネシア周辺には，147 もの地方語が消滅の危機に瀕しており，これらの地方語の保護支援，および地方語間のコミュニケーションの支援を行うための対訳辞書が必要である．これらの言語の保護支援のための対訳辞書などの言語資源の作成に，クラウドソーシングが用いられている．

クラウドソーシングとは，インターネットを通じて，不特定多数の人に仕事を依頼する仕組みのことであり，人手が必要な大量の作業を発注することができる．特に，計算機では比較的困難だが，人間にはそれほど難しくないタスクを発注するのに用いられる．

クラウドソーシングでは，不特定多数の作業者にタスクを発注するため，作業者の能力にばらつきがある．そして，タスクの実行結果は作業者の能力に依存するため，実行結果の品質を保証することは困難である．そのため，クラウドソーシング における品質管理はとても重要な課題である．品質の良い作業結果を用いるために最も重要なことは，能力の高い作業者にタスクを割り当ててもらうことである．しかし，作業者の能力は，タスクを実行してもらうまではわからない．そこで，ゴールドタスク（正解既知問題）を用いて，能力の高い作業者を判別する方法が用いられている．ゴールドタスクを用いて能力が高いと推測された作業者のみにタスクを割り当てることによって．品質の高い作業結果を得ることができる，しかし，ゴールドタスクを生成するのは非常に難しく，コストがかかる．加えて，人間が作業を行うため，間違える可能性を完璧に排除することはできない．そのため，ある単一の作業者の作業結果をそのまま利用することは困難である．そこで，複数の作業者に同じタスクを割り当て，多数決を取る方法が用いられる．しかし，インドネシアの地方語は話者が少なく，危機言語を含む複数の地方語を話すことのできる人はさらに限られるため，能力の高い作業者はあまり見込めない．このような作業者の能力の平均が低い集団では，多数決を用いて正しい答えを導く方法はあまり有効ではない．

本研究では，これらの問題を解決するために，作業結果による作業者の動的な信頼値評価手法を導入すことにより，動的に信頼できる作業者を判別するというアプローチをとった．信頼値に基づいて各作業者のタスクの割り当てられやすさの調整を行い，一部タスクにおいては信頼値が一定の値よりも高い，信頼できる作業者のみに割り当てる方法を提案する．このアプローチを実現する

にあたって，以下の課題に取り組む必要がある．

作業者とタスクのモデル化

信頼値を求めるための多様なタスク割り当て手法を設計し，それぞれの有用性の検証をするために，ワークフローを構成するタスクと，そのタスクを実行する作業者のモデルを作成する必要がある．

信頼値の算出

作業者の作業結果から効果的に信頼値を計算する方法を設計する必要がある．加えて，信頼値を用いた，信頼できる作業者を判別するための基準となる閾値の設定と，タスクの割り当てられやすさの調整を行う必要がある．

本稿の残りは以下のような構成となっている．第 2 章でクラウドソーシングにおける品質管理方法に関する関連研究を紹介し，第 3 章で作業者とタスクをどのようにモデル化したのかについて説明する．第 4 章で提案手法である作業結果による作業者の動的な信頼値評価手法について，信頼値の計算方法，および信頼値を用いたタスクの割り当て方法について具体的に説明する．その後に，第 5 章でこの手法の評価を行い，第 6 章で本稿をまとめる．

第2章　クラウドソーシングにおける品質管理方法

この章では，まずクラウドソーシングの概要について説明する．その後，クラウドソーシングで重要視されている品質管理方法として最も重要である，能力による作業者の選択方法について説明する．

2.1　クラウドソーシング

クラウドソーシングとは，インターネットを用いて不特定多数の人に仕事を依頼すること，もしくはその仕組みのことを指す．一般的に，画像のラベリングや文章の翻訳などのような，数秒から数分で実行でき，それほど高い専門知識を必要としないタスクが主に取り扱われている．Amazon Mechanical Turk[1)]（AMT）などの，クラウドソーシングの巨大なプラットフォームが存在するため，インターネットを通じて大勢の作業者を容易に確保することができる．そのため，特にコンピュータのみでは実行することが困難だが，人間の持つ能力を用いればそれほど難しくはないタスクを実施するのに適している．

クラウドソーシングを用いた言語資源の作成も盛んに行われており，AMT を用いて，英語とスペイン語間の用例対訳を作成する手法 [1] などが提案されている．他にも，用例対訳の収集，共有を目的とした多言語用例対訳共有システム TackPad の開発が行われており，主に医療の分野に特化した多言語の用例対訳の収集を行なっている [2]．従来は，専門家に言語資源の作成を依頼するのが主流であり，コストがかかるものであることが知られている，しかし，クラウドソーシングを用いることで，比較的安価に作成できるようになったため，言語資源に関する様々な研究が行われている．その中でも，クラウドソーシングを用いることで，正確性を保つことが難しくなったため，品質管理方法が議論されている [3].

クラウドソーシングにおいて重要視されている研究内容として，クラウドソーシングにおける品質管理方法がある．人間が作業を行うため，必ずしも正しい作業結果を得られるとは限らない．加えて，不特定多数の人に作業を依頼するため，能力の低い作業者や，意図的に品質の低い作業を行う作業者（スパムワーカ）が作業を行う可能性を完璧に排除することは難しい．そのため，単一作業者の作業結果を採用することは困難である．そこで，品質管理手法の研究では，

[1)] Amazon Mechanical Turk (https://www.mturk.com)

主に 2 つのアプローチから品質管理を行う研究がされている．

- 作業結果を集約して全体の品質を向上させる方法
- 個々の作業結果の品質を向上させる方法

前者は主に，作業結果から誤りを取り除くことで高品質な結果を得ることを試みるアプローチである．例として，同じタスクを複数の作業者に割り当て，冗長性を持たせた上で多数決を取る方法が用いられている．しかしながら，多数決を用いる方法では，作業者の能力が高い場合は正しい答えを導くことができる一方で，作業者の能力が低い（２値選択型タスクの場合は正解率 50%以下である）場合には，正解を導き出すことは困難である [4]．このように，クラウドソーシング上の作業者は能力にばらつきがあることが多いため，能力の低い作業者と能力の高い作業者の回答を同等に扱い，多数決をとることは賢明ではない．そこで，作業者によって回答の重みを変える，重み付き多数決が用いられている [5]．

後者は，タスクを作業者に依頼する前に報酬やタスクの設計，作業者の選択を行うことで，作業者によるタスクの実行結果そのものの向上を試みるアプローチである．例として，報酬分配を各作業者の評判情報を用いて行う手法 [6] や，あるタスクをより小さなタスクに分解するための手法 [7] 等が提案されている．なかでも，能力が高いと推測される作業者を事前に抽出し，抜き出した作業者にタスクを割り当てる手法は，タスク実行前に，能力の低い作業者やスパムワーカを排除することができ，能力が高いと推定される作業者のみが実際のタスクを行うため，特に作業結果の品質の向上が期待される．このような，タスク割り当て手法による品質管理法について，次章で詳しく述べる．

2.2 タスク割り当て

タスク割り当ては，これから依頼するタスクに対して，高い品質の作業結果を返すことが期待できる作業者を抜き出す手法であり，事前に作業者の能力を推定する必要がある．しかし，クラウドソーシング上に存在する作業者の能力は千差万別であり，作業者の能力を事前に知ることは困難である．そこで，作業者自身にタスクの回答の確信度を申告させる手法 [8] や，作業者に他の信頼できる作業者を紹介させる手法-[9] が研究されている．しかし，作業者の申告情報は不確かであるため，それ以外の情報を利用するのが賢明である．そこで，過去のタスクの多数決の結果をそのタスクの正解とし，各作業者の能力を推定

する手法が提案されている [10]．しかし，この手法は多数決の結果を用いるため，作業者の能力の平均が低い場合に適応するのが困難だと予想される．他にも，作業者の振る舞いから作業者の能力を推定する研究が行われている．具体的には，タスクフィンガプリントと呼ばれる，マウスやキーボード操作のタイミングや回数，処理時間などを作業者の画面操作ログとして抜き出し，作業者の振る舞いとして用いる手法 [11] や，ページのスクロールや，ラジオボタンの間隔から，作業者がどの問題を回答しているのかを予測し，各問題ごとの作業時間をその作業者の振る舞いとして用いる手法 [12] などが考案されている．このような作業者の振る舞いを特徴として機械学習を行い，作業結果の品質の推定が行われている [11]．

なかでも，最も正確に作業者の能力を推定する方法として，あらかじめ正解のわかっているタスク（ゴールドタスク）が用いられている．例として，ゴールドタスクを事前に割り当て，作業者の回答を評価することで作業者のフィルタリングを行う方法や，通常のタスクにゴールドタスクを紛れ込ませ，作業者の能力を測定し，選別する方法がその例である [13]．これらの方法により作業者の能力が低いと判定できた場合は，それ以降その作業者にはタスクの割り当てを行わない，一部のタスクに制限を設ける，もしくは，その作業者の作業結果を使用しないといった対策を講じることが可能である．この方法は作業者の能力の平均が高くない場合において，もっとも効果的な作業者の能力推定方法だと考えられる．しかし，実際のタスクにゴールドタスクを紛れ込ませる場合，答えのわかっているゴールドタスクへの回答に対しても報酬を支払わなくてはならない．事前にゴールドタスクを作業者に割り当て，作業者の能力を測定する場合，能力を推定したい作業者全員に対してゴールドタスクを割り当てる必要があるため，単純な作業量効率が悪くなってしまうという欠点がある．さらに，ゴールドタスクを生成をするのは大変困難であり，コストがかかることが知られているため，ある時点までに収集されたデータをもとに自動的にゴールドタスクを生成する方法が提案されている [14]．

本研究では，危機言語を対象とする，クラウドソーシングを用いた対訳辞書作成を想定している．そのため，危機言語を含む複数言語を話すことができる作業者は少なく，作業者の能力の平均が高くないことが容易に想像できる．本研究では，このような作業者の能力の平均が低い群衆から，答えのわかっているタスクをあまり使うことなく能力が高い作業者を抽出し，タスクを割り当て

ることを目的としている．これは，ゴールドタスクを用いることなく，作業者の能力を正確に評価する必要があることを意味する．

第3章　モデリング

危機言語を対象とした対訳辞書の作成を行うことを想定し，特性を理解した上で，クラウドソーシング作業者，タスク，そしてワークフローのモデル化を行う．

3.1　作業者

作業者の能力が高いほどタスクの実行結果の品質は高くなる．ここでは，作業者の能力を多言語における語彙力とし，能力が高くなればなるほど，その作業者が認識している語彙が多くなるとする．そして単純化のために，タスクの実行結果の品質は，作業者の能力によって確率的に決まると仮定する．ただし実際には，作業者がタスクを行い，その結果がわかるまでは能力が判明しないため，作業者の能力をベータ分布を用いて表す．その確率密度関数 $f(x|a,v)$ は式（1）で表す [15]．

$$f(x|a,v) = \mathrm{Beta}(\frac{a}{\min(a,0.7-a)v}, \frac{0.7-a}{\min(a,0.7-a)v}) \times 0.7 \qquad (1)$$

ここで，$a \in (0,0.7)$ は作業者の能力の平均値を正規化した値であり，$v \in (0,1)$ は作業者の能力の分散を決定するパラメータである．v は 0 に近づくほど分散が 0 に近づき，v が 1 に近づくと平均が a のベータ分布の中で最も分散が大きくなる．上記の作業者のモデルは， [15] で採用されたものを基に，危機言語を対象としたクラウドソーシングによる対訳辞書作成というフィールドの特性に合わせて作業者の能力の上限を 0.7 になるように改良したものである．

3.2　タスク

作業者に割り当てるタスクは，自由入力タスクである対訳作成タスクと，複数の 2 値選択型タスクである対訳評価タスクの 2 種類であるとする．

対訳作成タスク

与えられた単語や文章の対訳をに自由に作成する自由入力型タスク

対訳評価タスク

対訳作成タスクによって作成された対訳が“正しい”か“間違っている”かを評価する 2 値選択型タスク

対訳作成タスクの実行結果は，作業者が与えられた単語の対訳を知っている

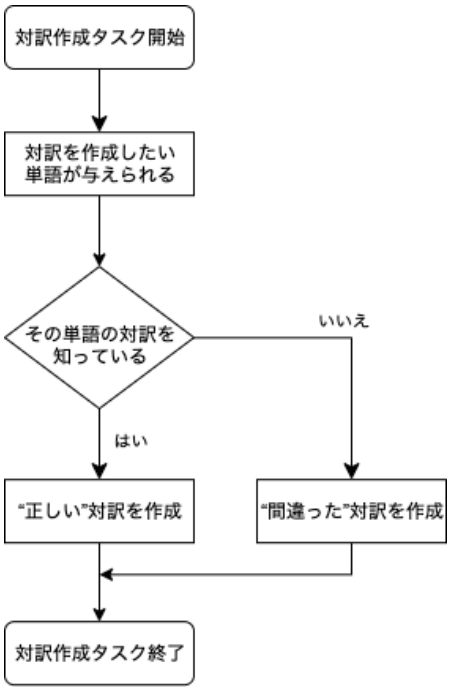

図 1: 対訳作成タスクのモデル

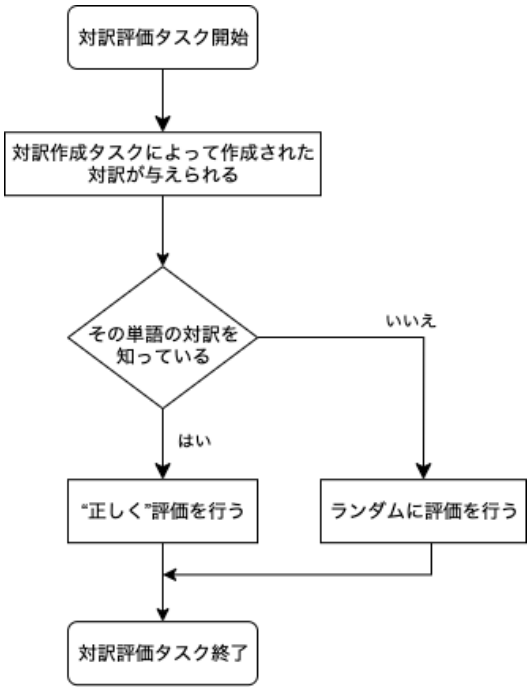

図 2: 対訳評価タスクのモデル

場合は”正しい”対訳を作成し，もしもその単語の対訳を知らない場合は”正しい”対訳を作成できず，”間違った”対訳を作成するとする．そのため，完璧に作業者の能力に依存し，正しい対訳が作成されるか，そうでないかが明確に決定されるとする（図 1）．しかし，対訳評価タスクは 2 値選択型タスクであるため，作業者が与えられて単語の正しい対訳を知っている場合は”正しい”評価を行うが，その単語の対訳を知らなかった場合は，”正しい”か”間違っている”かの 2 値から無作為に選択するとする（図 2）．そのため，対訳評価タスクにおいては，作業者の能力がどれだけ低くても，50%以上の確率で”正しい”評価を行うことは保証されている．

3.3 ワークフロー

対訳作成タスクと，複数の対訳評価タスクで構成されるワークフローを考える（図 3）．対訳作成タスク 1 回に対して複数回の対訳評価タスクを行うことで，冗長性を確保する．つまり，対訳作成タスクによって作成された対訳の最終的な評価は，対訳評価タスクの実行結果の多数決で決定されるとする．最終的に得られる対訳の品質については（表 1）の通りである．何も対訳が得られなかった場合は，対訳作成タスクから再度行い，全ての単語について対訳を得ること

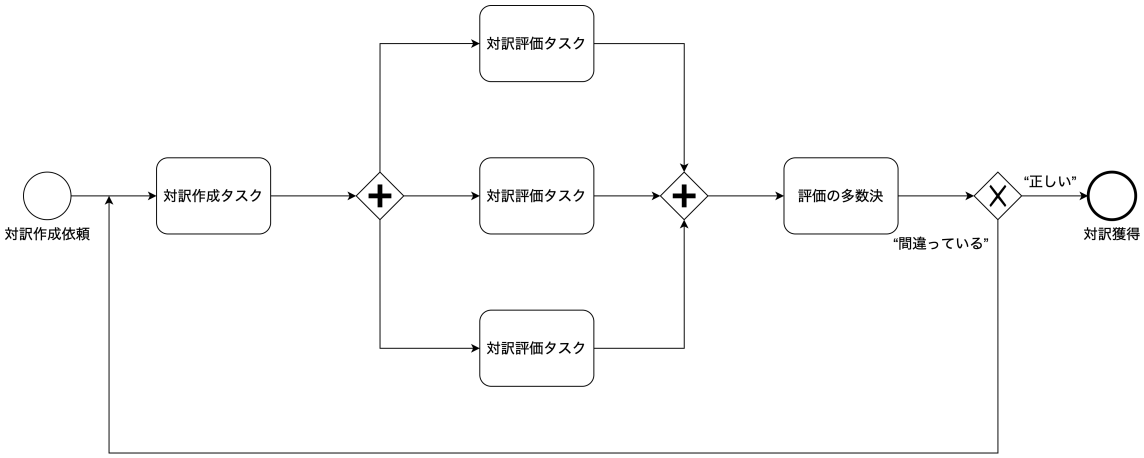

図 3: 対訳辞書作成におけるワークフロー

表 1: 最終的に得られる対訳の品質

		対訳評価タスク	
		“正しい”評価	“間違った”評価
対訳作成タスク	“正しい”作成	“正しい”対訳獲得	対訳なし
	“間違った”作成	対訳なし	“間違った”対訳獲得

ができるまで繰り返す.

第4章　作業者の動的な信頼値評価手法

本研究では，クラウドソーシングにおける作業者の実行結果を用いて，作業者の信頼値を計算することで能力が高いと推測される作業者を判別することを目指す．そのための具体的な方法についてこの章で説明する．

4.1　信頼値の計算方法

対訳評価タスクは2値選択型タスクであるため，作業者の能力を測るのには適していない．そのため，作業者に行ってもらうタスクのうち，対訳作成タスクの実行結果のみを用いて信頼値の計算を行う．各作業者に信頼値というパラメータを設定し，初期値を0とする．

- ある対訳作成タスクににによって作成された対訳が，対訳評価タスクの実行結果の多数決によりにより“正しい”と判断された場合，その対訳の作成者の信頼値を $+1$ する
- ある対訳作成タスクににによって作成された対訳が，対訳評価タスクの実行結果の多数決によりにより“間違っている”と判断された場合，その対訳の作成者の信頼値を -1 する

この計算を1つの単語の対訳の評価が完了するたびに行う．

4.2　信頼値を用いたタスクの割り当て方法

各作業者の信頼値を用いることにより，２種類のタスク割り当てにおける手法を考案した．

- 閾値を用いた対訳評価タスクの割り当て
- 重み付き確率を用いたタスク割り当て

前者に関しては，対訳評価タスクに制限を設けた．対訳作成タスクについては作業者全体から無作為に選んだ作業者に行ってもらうこととし，対訳評価タスクについては，信頼値が1以上の作業者を信頼できる作業者とみなし，信頼できる作業者のみが行うことができるとする．これにより，対訳評価の間違いが少なくなることが期待される．

後者に関しては，対訳作成タスクと対訳評価タスクの両方について，タスクの割り当てられやすさの調整を各作業者の信頼値を用いた重み付けに基づいて行った．あるタスクを実行可能な総作業者数が n である時，i 番目の作業者の重

み w_i は式（2）のように計算する．

$$w_i = 1 + r_i - r_{min} \tag{2}$$

r_i は i 番目の作業者の信頼値を表しており，r_{min} はそのタスクを実行可能な全作業者の信頼値のうち，最も小さい値を表している．式（2）のように計算することで，信頼値が最も小さい作業者の重みが 0 になる（タスクが割り当てられる確率が 0 になる）のを避けることができる．そして，作業が進むにつれて作業者間での信頼値の差が大きくなることで，重みの差も拡大していく．

ある作業者にタスクが割り当てられる確率 p_i は重みを用いることで，式（3）のように計算できる．

$$p_i = \frac{w_i}{w_1 + w_2 + w_3 + \cdots + w_i + \cdots + w_n} \tag{3}$$

これらの計算を，タスクを割り当てる度に行うことで，信頼値の高い作業者（能力が高いと推測される作業者）にはタスクが割り当てられやすくし，信頼度の低い作業者（能力が低いと推測される作業者）にはタスクが割り当てられづらくすることで，自動的に能力が低いと推測される作業者を排除していくことができる．

第5章　評価

第4章で説明した作業結果を用いた動的な信頼値の評価手法の有用性を検証するために，第3章で作成したモデルを用いてシミュレーションを行い，既存の品質管理手法の効果と比較する．

5.1　評価方法

提案するタスク割り当て手法を含む手法に対して，作成された対訳の正確性，作成された正しい対訳1つあたりの作業量，そして，使用されたゴールドタスクの数の観点からの評価を行う．

1. 提案手法1（Reliable1）
   これまでの章で説明してきた，作業結果による作業者の動的な信頼値の評価を行い，信頼値による実行可能タスクの制限を設ける手法．
   最初の信頼できる作業者についてはゴールドタスクを用いて，対訳評価に必要な最少人数のみ発見することとする．そして，途中で信頼できる作業者の人数が対訳評価に必要な最少人数に満たなくなった場合，追加でゴールドタスクを作業者に割り当て，信頼できる作業者の人数を増やすこととする．
2. 提案手法2（Reliable2）
   提案手法1と同じく，作業結果による作業者の動的な信頼値の評価を行う方法であるが，こちらは信頼値による実行可能タスクの制限に加えて，重み付き確率を用いたタスク割り当ても行う手法である．
   最初の信頼できる作業者についてはゴールドタスクを用いて，対訳評価に必要な最少人数のみ発見することとする．そして，途中で信頼できる作業者の人数が対訳評価に必要な最少人数に満たなくなった場合，追加でゴールドタスクを作業者に割り当て，信頼できる作業者の人数を増やすこととする．これにより，作業者の中から無作為にタスクを割り当てる場合と，タスクの割り当てられやすさを信頼値に基づいて調節した場合の比較を行うことができる．
3. 比較手法1（GoldTaskOnlyEvaluation）
   提案手法と同様に，信頼できる作業者のみが対訳評価タスクを行うが，ゴールドタスクを用いて信頼できる作業者を判別する手法．

実際の作業を行う前に，事前評価としてゴールドタスクを作業者全員に割り当て，正解者を信頼できる作業者とみなす．これにより，答えがわかっているタスクを使って信頼できる人を判別した場合と，作業結果から計算した信頼値をもとに信頼できる作業者を判別した場合との比較を行うことができる．

4. 比較手法 2（GoldTask）
   ゴールドタスクを用いて信頼できる作業者を判別し，信頼できる作業者のみが全てのタスクを行う手法．
   実際の作業を行う前に，事前評価としてゴールドタスクを作業者全員に割り当て，正解者を信頼できる作業者とみなす．これにより，対訳評価タスクだけでなく，対訳作成タスクも信頼できる人のみが行った場合との比較を行うことができる．
5. 比較手法 3（Random）
   全てのタスクにおいて，作業者全体から無作為に選び，割り当てる方法．これにより，作業者の能力の測定を全く行わなかった場合との比較を行うことができる．

上記で説明した手法それぞれに対して，以下に示す指標を用いて作成された対訳の正確性，作成された正しい対訳 1 つあたりの作業量，そして，使用されたゴールドタスクの数の測定を行う．

1. 作成された対訳の正確性
   それぞれの手法によって作成された対訳の正確性は以下のように計算する．

$$\text{正確性} = \frac{\text{作成された対訳のうち，正しい対訳数}}{\text{作成された対訳の総数}} \tag{4}$$

   これにより，各手法での単純な成果物の品質を比較することができる．
2. 作成された正しい対訳 1 つあたりの作業量
   それぞれの手法によって作成された正しい対訳 1 つあたりの作業量は以下のように計算する．

$$\text{正しい対訳 1 つあたりの作業量} = \frac{\text{作業量}}{\text{作成された対訳のうち，正しい対訳数}} \tag{5}$$

   作業量は，割り当てられたゴールドタスク，対訳作成タスク，対訳評価タスクの総合計である．これにより，各手法での作業効率を比較することができる．

3. 使用されたゴールドタスクの数
   提案手法と，ゴールドタスクを用いる比較手法 1 および比較手法 2 において使用されたゴールドタスクのと比べることで，使用されたゴールドタスクの数にどの程度差異があるのかを知ることができる．

上記で説明した観点についての評価を行うために，各手法でシミュレーションを行なった．その際の条件は以下の通りである．

- 対訳を作成する単語の個数：1000 個
- 作業者の人数：100 人
- 作業者の能力：
  3.1 のモデルに基づいてに決定し，作業者の能力の平均は 0.2 から 0.6 の間で 0.1 刻みに変化させて比較を行う．分散は 0.5 とする．
- ゴールドタスク：
  - 事前評価 1 回につき，ゴールドタスクを 1 回割り当て．その正解者を信頼できる作業者とみなす
  - 事前評価 1 回につき，ゴールドタスクを 3 回割り当て，3 回中 2 回正解した作業者を信頼できる作業者とみなす

  これらの 2 種類の場合を比較する．

乱数による偏りを排除するために，各状態ごとにシミュレーションを 100 回行った結果の平均値を用いた．

5.2 結果

5.2.1 事前評価 1 回につきゴールドタスク 1 回の場合

正確性は提案手法である Reliable2 が一番良い結果になり，それよりも少し正確性が低い，Reliable1 と GoldTask がほぼ同じぐらいの正確性になった（図 4）．

正しい対訳 1 つ当たりの作業量は，Reliable2 と GoldTask がほぼ同率でもっとも少ない結果となった．他の 3 つの手法に関してはほとんど差はなかった（図 5）．

使用されたゴールドタスクの数は，GoldTask と GoldTaskOnlyEvaluation でのでの使用数が等しく，一番多かった．提案手法である Reliable1 と Reliable2 はそれの半分から 1/5 程度の使用数となった（図 6）．

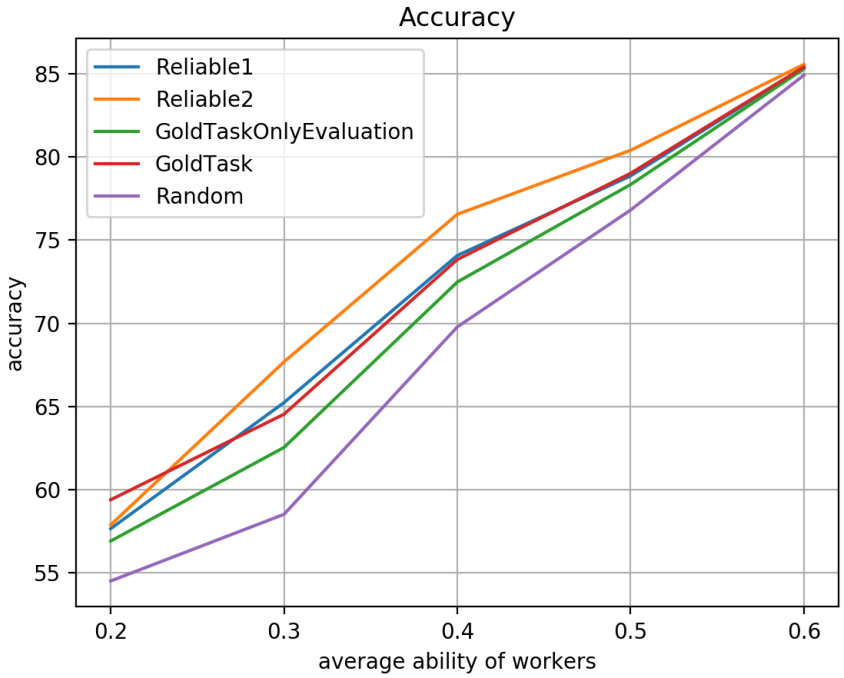

図 4: 作成された対訳の正確性（事前評価 1 回につきゴールドタスク 1 回）

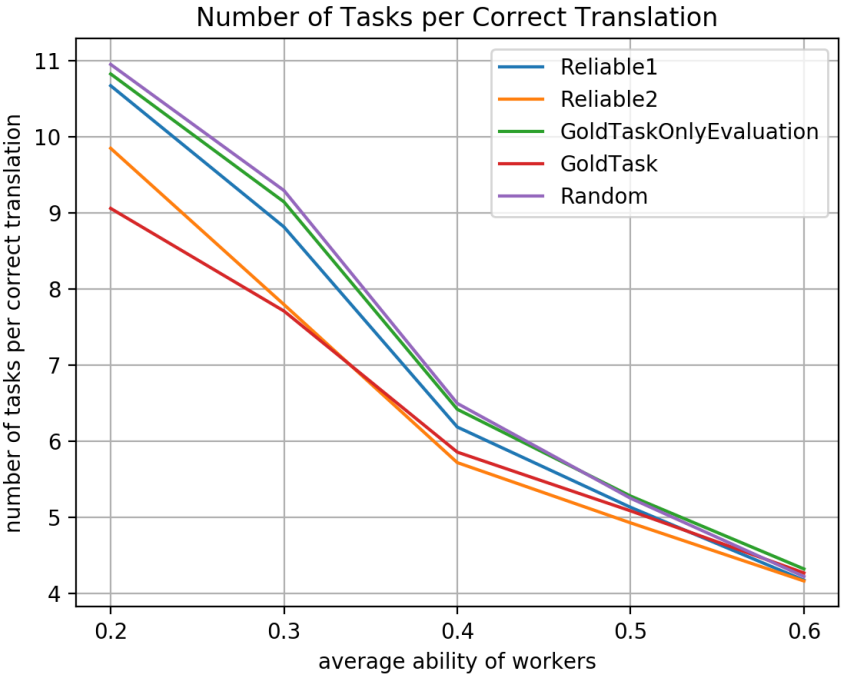

図 5: 正しい対訳 1 つあたりの作業量（事前評価 1 回につきゴールドタスク 1 回）

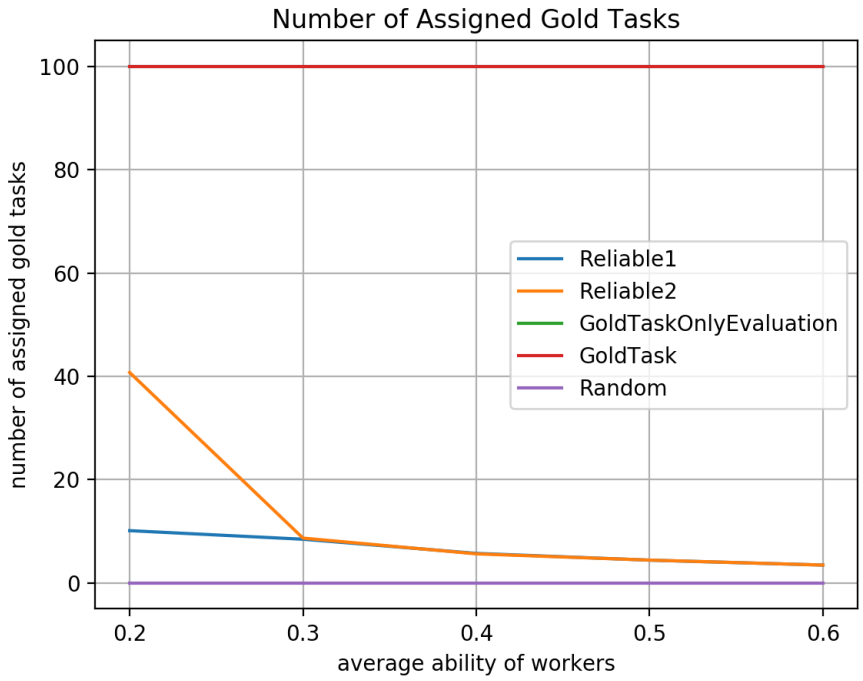

図 6: 使用されたゴールドタスクの数（事前評価 1 回につきゴールドタスク 1 回）

5.2.2　事前評価 1 回につきゴールドタスク 3 回の場合

正確性は，GoldTass と Reliable2 とが一番良い結果になり，それよりも少し正確性が低い，Reliable1 と GoldTaskOnlyEvaluation がほぼ同じぐらいの正確性になった（図 7）．

正しい対訳 1 つ当たりの作業量は，Reliable2 と GoldTask がほぼ同率でもっとも少ない結果となった．他の 3 つの手法に関してはほとんど差はなかった（図 8）．

使用されたゴールドタスクの数は，GoldTask と GoldTaskOnlyEvaluation ででの使用数が等しく，一番多かった．提案手法である Reliable1 と Reliable2 はそれの 1/3 から 1/6 程度の使用数となった（図 9）．

5.3　考察

5.3.1　正確性

事前評価 1 回あたりに用いるゴールドタスクの回数を変えることによって，ゴールドタスクでの作業者の能力推定の精度を上げることができる．そのため，事前評価 1 回につきゴールドタスクを 3 回割り当てる場合のシミュレーション

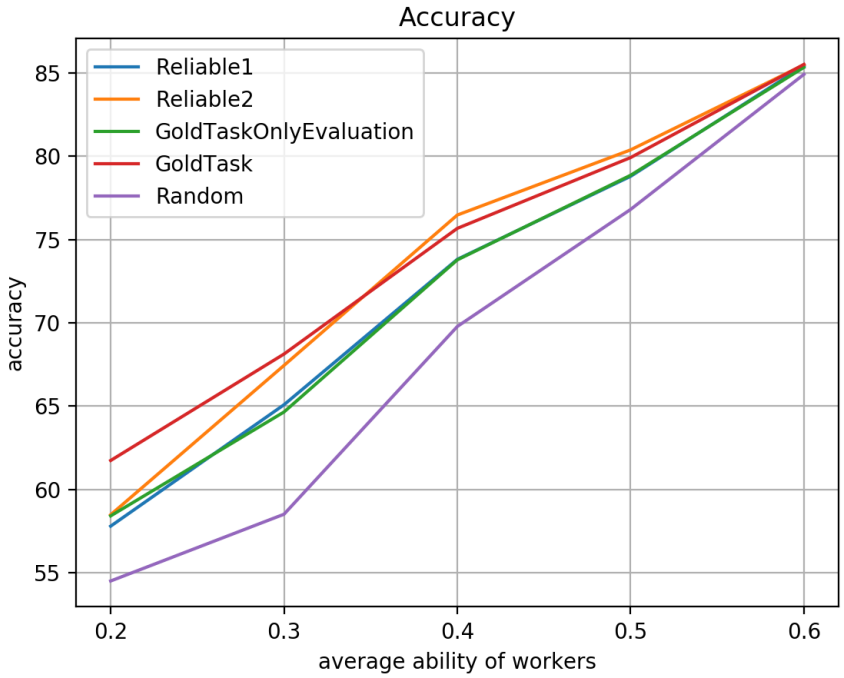

図 7: 作成された対訳の正確性（事前評価 1 回につきゴールドタスク 3 回）

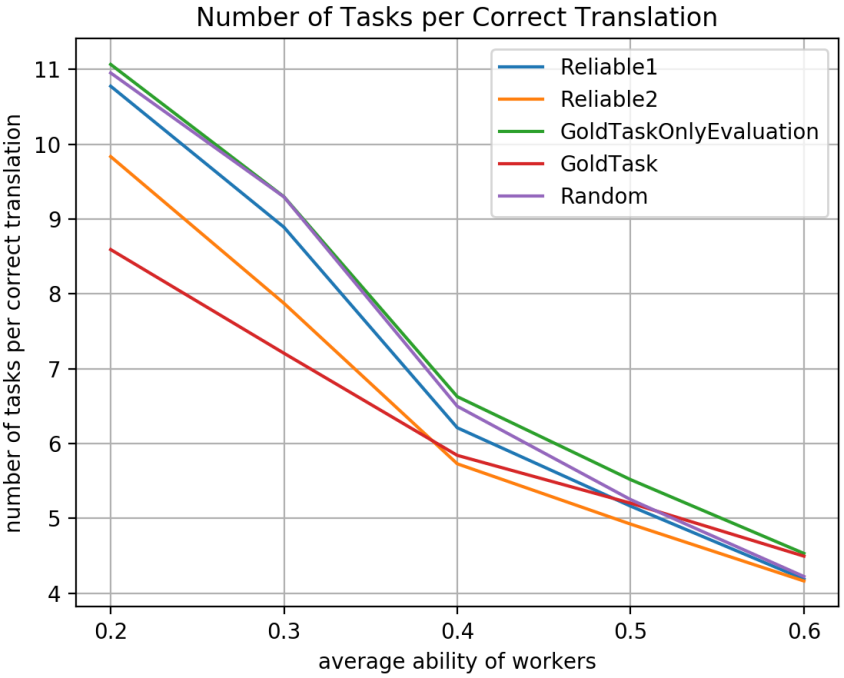

図 8: 正しい対訳 1 つあたりの作業量（事前評価 1 回につきゴールドタスク 3 回）

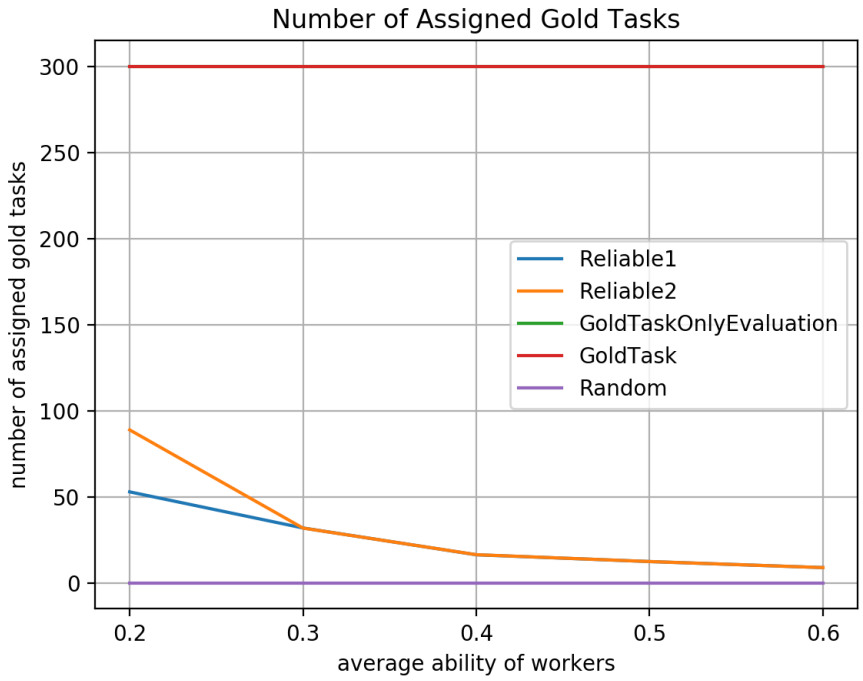

図 9: 使用されたゴールドタスクの数（事前評価 1 回につきゴールドタスク 3 回）

では，GoldTask と GoldTaskOnlyEvaluation の方が正確に作業者の能力を測れる．そのため，提案手法よりも正確性が高いと予想していたが，実際は，提案手法と大差はない結果になった．

Reliable1 と GoldTaskOnlyEvaluation では，信頼できる作業者の判別方法が異なるだけで，対訳作成タスクは作業者の中からランダムに割り当て，対訳評価タスクは信頼できる作業者の中からランダムに割り当てるため，タスクの割り当て方法は全く同じである．これらの手法を比較すると，事前評価 1 回につき，ゴールドタスクを 1 回使用した場合では，Reliable1 の方が正確性が高かった．そして，事前評価 1 回につき，ゴールドタスクを 3 回使用した場合では，両者に差はほとんど見られなかった．

さらに，使用されたゴールドタスク数を比べても，提案手法は，従来の手法の半分以下しか使用していない．このことから，信頼できる作業者による他の作業者の能力の評価手法は，ゴールドタスクを用いた作業者の能力の評価手法と同等か，より有効であると考えられる．

5.3.2 効率

ゴールドタスクの使用回数が多いのにもかかわらず，正しい対訳 1 つあたりの作業量は，GoldTask がもっとも少ない結果になった．これは，ゴールドタスクの正解者が，対訳の作成を含む全ての作業を行うためであると考えられる．提案手法では，対訳作成タスクを作業者全体から選択して割り当てる一方で，GoldTask では，ゴールドタスクの正解者のみが，対訳作成タスクと対訳評価タスクの両方を行う．そのため，対訳作成タスクにより間違った対訳が作成される可能性がそもそも低いため，効率が良いと考えられる．つまり，効率を上げるためには，対訳評価タスクだけでなく，対訳作成タスクにも能力の高い作業者を割り当てることが有効であることがわかる．Reliable2 では各作業者の信頼値から重みを計算し，これを用いてタスクの割り当てられやすさの調整を行なった．その結果，正しい対訳 1 つあたりの作業量は GoldTask と同程度であった．このことから，信頼値による重み付き確率を用いたタスク割り当て方法は，ゴールドタスクを用いた，能力の低い作業者の排除とおおよそ同じ精度であると考えられる．

5.3.3 ゴールドタスク数

作業者の能力の平均値が 0.2 の時に，Reliable2 のゴールドタスク使用数が Reliable1 に比べて多かった．これは，途中で信頼できる作業者の人数が対訳評価に必要な最少人数に満たなくなり，その度にゴールドタスクの割り当てを行い，追加で信頼できる作業者の探索を行なっているためであると考えられる．Reliable1 でも同様に，信頼できる作業者の人数不足は発生するが，Reliable2 では，より信頼できる作業者の人数が足りなくなりやすいと考えられる．なぜなら，Reliable2 では信頼値による重みづきタスクの割り当て手法を用いているためである．これにより，信頼できる作業者の中に混じった，能力があまり高くないが信頼値が高い作業者に対訳作成タスクが割り当てられやすくなる．そして，そのような作業者が作成した対訳が間違っていると評価されることで信頼値が減少し，閾値を下回ることで，信頼できる作業者からの排除されるためであると推測される．

GoldTask と GoldTaskOnlyEvaluation では，信頼できる作業者を抽出するために，事前評価としてゴールドタスクを全作業者に割り当てなければならない．それに比べて，提案手法では，対訳評価タスクを行うのに必要な最少人数の信頼できる作業者を抽出することができれば良いため，ゴールドタスクの使用回

数を従来の半分以下に抑えることが可能であると考えられる.

第6章　おわりに

クラウドソーシングに限らず，人に何か作業を依頼する場合には，故意であろうとなかろうと，作業結果に誤りが含まれる可能性がある．そのため，誤りを除去する，または誤りがそもそも起こらないように工夫し，作業結果の品質を管理する手法は必要不可欠である．

本研究では，作業結果による作業者の動的な信頼値評価によるタスク割り当て手法を導入することで，少ないゴールドタスク数でも，能力の高い作業者を推定できることを示した．そして，信頼値による実行可能タスクの制限や，信頼値による重み付き確率を用いたタスク割り当てにより，能力が低いと推測される作業者を排除することで，作業結果の品質の向上を行なった．

本研究の貢献は以下の2点である．

作業者とタスクのモデル化

作業者の能力値を事前に設定し，タスクの実行結果は作業者の能力値により決定されるようにモデル化を行なった．これにより，様々な手法でのシミュレーションを同じ条件で行うことが可能となった．

信頼値の算出

作業結果に基づく信頼値の計算と，各作業者の信頼値に基づいたタスク割り当て手法の定式化を行った．その結果，信頼値を用いる提案手法では，既存手法の半分以下のゴールドタスク数で，同程度の正確性を得ることに成功した．

また，本稿では未解決の問題も存在する．今回は，作業者の能力値のモデル化しか行なっていない．そのため，タスクの実行結果は作業者の能力値のみを用いてを決定し，各タスクの難易度は全く考慮していない．そのため，各タスクの難易度を設定し，これを考慮したシミュレーションを行う必要がある．

さらに，信頼値についても，単純化のために，正答数と誤答数の差とした．本来であれば，正解した評価者の信頼値を集約し，対訳作成に付与する信頼値を計算を行うべきであるが，本研究ではそこまで至らなかった．対訳評価タスクにおいても，信頼値に基づいた回答の重み付けを用いることで，多数決の精度を向上させるべきであった．

加えて，作業者が故意に間違った回答や，でたらめな回答を行うケースを考慮していない．そのため，迷惑行為を行う作業者（スパムワーカ）のモデル化

を行い，スパムワーカが存在する群集に対しても，提案手法が有効であるのかを検証する必要がある．

謝辞

本研究を行うにあたり，熱心なご指導，ご助言を賜りました村上陽平准教授に深謝申しあげます．また，普段からお世話になっている社会知能研究室の皆様に心より感謝いたします．

参考文献

[1] Negri, M. and Mehdad, Y.: Creating a Bi-lingual Entailment Corpus through Translations with Mechanical Turk: $100 for a 10-day Rush, *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pp. 212–216 (2010).

[2] 福島拓, 吉野孝, 重野亜久里: 正確な情報共有のための多言語用例対訳共有システム, 情報処理学会論文誌コンシューマ・デバイス＆システム, Vol. 2, No. 3, pp. 23–33 (2012).

[3] Gilles Adda, Benoît Sagot, K. F. and Mariani, J.: Crowdsourcing for Language Resource Development: Critical Analysis of Amazon Mechanical Turk Overpowering Use, *5th Language and Technology Conference* (2011).

[4] Victor S. Sheng, F. P. 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*, pp. 614–622 (2008).

[5] Rion Snow, Brendan O'connor, D. J., Ng, A. Y.: Cheap and fast–but is it good? evaluating non-expert annotations for natural language tasks, *Proceedings of the 2008 conference on empirical methods in natural language processing*, pp. 254–263 (2008).

[6] Zhang, Y. and van der Schaar, M.: Reputation-based Incentive Protocols in Crowdsourcing Applications, *2012 Proceedings IEEE INFOCOM*, pp. 2140–2148 (2012).

[7] Anand Kulkarni, M. C. and Hartmann, B.: Collaboratively crowdsourcing workflows with turkomatic, *Proceedings of the acm 2012 conference on computer supported cooperative work*, pp. 1003–1012 (2012).

[8] 小山聡, 馬場雪乃, 櫻井祐子, 鹿島久嗣: クラウドソーシングにおけるワーカーの確信度を用いた高精度なラベル統合, 人工知能学会全国大会論文集 第 27 回全国大会 (2013).

[9] 西智樹, 小出智士, 大野宏司, 長屋隆之: ソーシャルネットワークを用いたクラウドソーシングの品質向上, 人工知能学会全国大会論文集 第 27 回全国大会 (2013).

[10] Pinar Donmez, J. G. G. and Schneider, J.: Efficiently Learning the Accuracy of Labeling Sources for Selective Sampling, *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 259–267 (2009).

[11] Rzeszotarski, J. M. and Kittur, A.: Instrumenting the crowd: using implicit behavioral measures to predict task performance, *Proceedings of the 24th annual ACM symposium on User interface software and technology*, pp. 13–22 (2011).

[12] Matthias Hirth, Sven Scheuring, T. H. C. S. and Tran-Gia, P.: Predicting result quality in crowdsourcing using application layer monitoring, *2014 IEEE Fifth International Conference on Communications and Electronics (ICCE)*.

[13] Gabriella Kazai, Jaap Kamps, M. K. and Milic-Frayling, N.: Crowdsourcing for book search evaluation: impact of hit design on comparative system ranking, *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pp. 205–214 (2011).

[14] David Oleson, Alexander Sorokin, G. L. V. H. J. L. and Biewald, L.: Programmatic Gold: Targeted and Scalable Quality Assurance in Crowdsourcing, *Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence*, pp. 43–48 (2011).

[15] Shinsuke Goto, T. I. and Lin, D.: Understanding Crowdsourcing Workflow: Modeling and Optimizing Iterative and Parallel Processes, *Fourth AAAI Conference on Human Computation and Crowdsourcing*, pp. 52–58 (2016).