修士論文

クラウドソーシングを用いた
対訳辞書作成のための品質管理手法

指導教官　　村上 陽平 准教授

立命館大学大学院情報理工学研究科
修士課程情報理工学専攻

地田 大樹

令和 4 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品質管理手法

地田 大樹

内容梗概

現在，対訳辞書などの言語資源をクラウドソーシングで作成することが主流になりつつある．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．

しかしながら，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．特に低資源言語間の対訳辞書の場合，複数の低資源言語を話すことができる人は限られ，作業者の能力の平均が低い．そのため，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く，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，超問題（複数のタスクをまとめてひとつのタスクとみなしたもの）を用いた回答統合手法を用いることで，高信頼な作業者の少ない環境での品質向上を目指す．具体的には，能力の高い作業者は超問題への回答において一致しやすいため，超問題を用いることで能力の高い作業者が多数派になる可能性を高める．

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．

高信頼な評価者の選択

超問題では，少数の高品質作業者が作業者集団の中に含まれることを前提としている．したがって，不特定多数からなるクラウドソーシングに超問題を適用するには，不特定多数の作業者の中から高品質な作業者を少なくとも二人は選択できる必要がある．

作業時間の短縮

作業者が作成した対訳の正誤を正しく評価できたとしても，間違った対訳が作成されると，作成のやり直しが発生してしまい，作業量が増えてしまう可能性がある．そのため，能力の高い作業者にタスクを積極的に割り当てる必要がある．また，超問題では，高信頼な作業者が評価者に少ないと

超問題の多数決が不調となり，評価のやり直しが生じ得る．評価のやり直しはすでに行った評価作業を無効とするため，作業時間を短縮するためには，評価結果の再利用が必要である．

一つ目の課題に対しては，作業結果に基づく作業者の動的な信頼値評価を行い，能力が高いと推定される評価者を選択した．

二つ目の課題に対しては，信頼値に基づく作成者の選出を行った．また，超問題の多数決において多数決解が決まらず，個々の問題に分解した後の多数決でもうまくいかないものがあった場合，分解後の多数決でうまく答えを得ることに成功した問題に正解した評価者だけで再度多数決を行うようにした．

シミュレーションと実データを用いて，作成された辞書の正確性と作業量について評価を行い，提案手法の有効性を検証した．

本研究の貢献は以下の通りである．

高信頼な評価者の選択

各作業者の作業結果より算出された信頼値に基づき評価者を選出することで，シミュレーションによる評価では，10〜17%ほど正確性を向上させることに成功した．また，実データによる評価では，3%ほど正確性を向上させることに成功した．

作業時間の短縮

信頼値の高い作業者に積極的に対訳作成タスクを割り当て，超問題の多数決が不調だった場合に評価の再利用を行うことで，シミュレーションによる評価において，一定数以上能力値が高い作業者が含まれる場合には，作業量を 2000〜2500 ユニットほど削減することに成功した．また，実データによる評価では，600 ユニットほど削減することに成功した．

Quality Control for Crowdsourced Bilingual Dictionary Creation

Hiroki CHIDA

Abstract

Recently, crowdsourcing is becoming mainstream to create language resources including bilingual dictionaries. Crowdsourcing is a scheme for requesting work from a large and open group of people via the Internet, and it can be used to order a large amount of works that requires human labor. Crowdsourcing is especially used to request tasks that are relatively difficult for computers, but not so difficult for humans. However, in crowdsourcing, where the tasks are executed by an unspecified number of workers the abilities of whom vary, it is difficult to guarantee the quality of the execution results. Especially in the case of bilingual dictionaries creation between low-resource languages, the number of people who can speak multiple low-resource languages is limited, and the average ability of workers is low. This results in the method of assigning the same task to multiple workers and using majority voting has a high possibility of obtaining wrong answers, and quality control cannot be performed well.

Therefore, we aim to improve quality in an environment with a small number of highly reliable workers by using an answer aggregation method on hyper questions (multiple tasks that are considered together as one task). Since workers with high ability tend to agree on the answers to hyper questions, the method increases the possibility that workers with high ability will be in the majority. To this end, we address the following two problems.

Selecting highly reliable evaluators

In the answer aggregation method on hyper questions, it is assumed that a small number of high quality workers are involved. Therefore, it is necessary to select highly reliable evaluators from a crowd.

Reducing the number of tasks

Even if a worker is able to correctly evaluate whether a bilingual text is correct or not, if an incorrect bilingual text is produced, the worker may have to redo the translation, which may increase the number of tasks. Therefore, it is necessary to proactively assign tasks to workers with high

ability.

For the first problem, we dynamically evaluate the reliability of workers based on their work results and selected evaluators who were estimated to be highly competent.

For the second problem, we selected the translator based on the reliability. In addition, when the majority vote for a hyper question fail and there are no majority answers for some of decoded tasks, the another round of majority voting is taken by the evaluators who voted majority answers for the rest of succeeded tasks.

We used simulations and an actual data to verify the effectiveness of the proposed method by evaluating the accuracy and the work quantity of the generated dictionary.

The contributions of this paper are as follows:

Selecting highly reliable evaluators

By selecting evaluators based on the reliability calculated from the results of each worker's work, we succeeded in improving the accuracy by 10 to 17% in the simulation evaluation. In the evaluation using actual data, we succeeded in improving the accuracy by 3%.

Reducing the number of tasks

By proactively assigning translation tasks to workers with high reliability, and by reusing the evaluations when the majority vote on hyper questions fail, we were able to reduce the work quantity by 2000-2500 units when the number of workers with high ability was more than a certain number in the simulation evaluation. In the evaluation using actual data, we succeeded in reducing the work quantity by about 600 units.

クラウドソーシングを用いた
対訳辞書作成のための品質管理手法

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ない．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，低資源言語間の対訳辞書の場合，複数の低資源言語を話すことができる人は限られるため，作業者の能力の平均が低いと予想される．このような能力の平均が低い群衆では，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，超問題（複数のタスクをまとめて一つのタスクとみなしたもの）を用いた回答統合手法を用いることで，高信頼な作業者の少ない環境での多数決の品質を向上させるというアプローチを取った．能力の高い作業者は超問題への回答において一致しやすいため，超問題の多数決を取ることで，能力の高い作業者が多数派になる可能性が高くなる．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．

高信頼な評価者の選択

超問題では，少数の高品質作業者が作業者集団の中に含まれることを前提

としている．したがって，不特定多数からなるクラウドソーシングに超問題を適用するには，不特定多数の作業者の中から高品質な作業者を少なくとも二人は選択できる必要がある．

作業時間の短縮

作業者が作成した対訳の正誤を正しく評価できたとしても，間違った対訳が作成されると，作成のやり直しが発生してしまい，作業量が増えてしまう可能性がある．そのため，能力の高い作業者にタスクを積極的に割り当てる必要がある．また，超問題では，高信頼な作業者が評価者に少ないと超問題の多数決が不調となり，評価のやり直しが生じ得る．評価のやり直しはすでに行った評価作業を無効とするため，作業時間を短縮するためには，評価結果の再利用が必要である．

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 章で対訳辞書作成のためのワークフローについて説明する．第 4 章と第 5 章でそれぞれ超問題を用いた回答統合手法と，信頼値に基づくタスク割り当て手法について具体的に説明する．その後に，第 6 章でシミュレーションモデルによりこれらの手法の評価を行い，第 7 章では実データを用いたモデルにより各手法の評価を行う．さらに．第 8 章では，実データをもとに回帰モデルを作成し，これにより各手法の評価を行う．そして，第 9 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].

2.2　品質管理手法

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，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

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

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 [3]．

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 [4] 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 [5] 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抽出した後に，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るため，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．

2.3　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，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 [6]．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．

そこで，あらかじめ正解のわかっているタスク（ゴールドタスク）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．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 [7]．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，この方法では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 [8]．

本研究では，危機言語を含む，低資源言語を対象とする，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，これらの言語を複数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にも有効な回答統合手法と，各作業者の作業結果より算出された信頼値に基づくタスクを割り当て手法を組み合わせることで，作成される対訳の品質向上を目指す．

第3章　対訳辞書作成のためのワークフロー

3.1　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1）．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．“正しい”対訳が作成され，それが“正しく”評価された場合は，“正しい”対訳を獲得する．“間違った”対訳が作成され，それが“間違った”評価をされた場合は“間違った”対訳を獲得する．それ以外の場合は，対訳は獲得されないとする．対訳が獲得されなかった場合は，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ができるまで繰り返す．

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

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”対訳を作成	“正しい”対訳獲得	獲得される対訳なし
	“間違った対訳を”作成	獲得される対訳なし	“間違った”対訳獲得

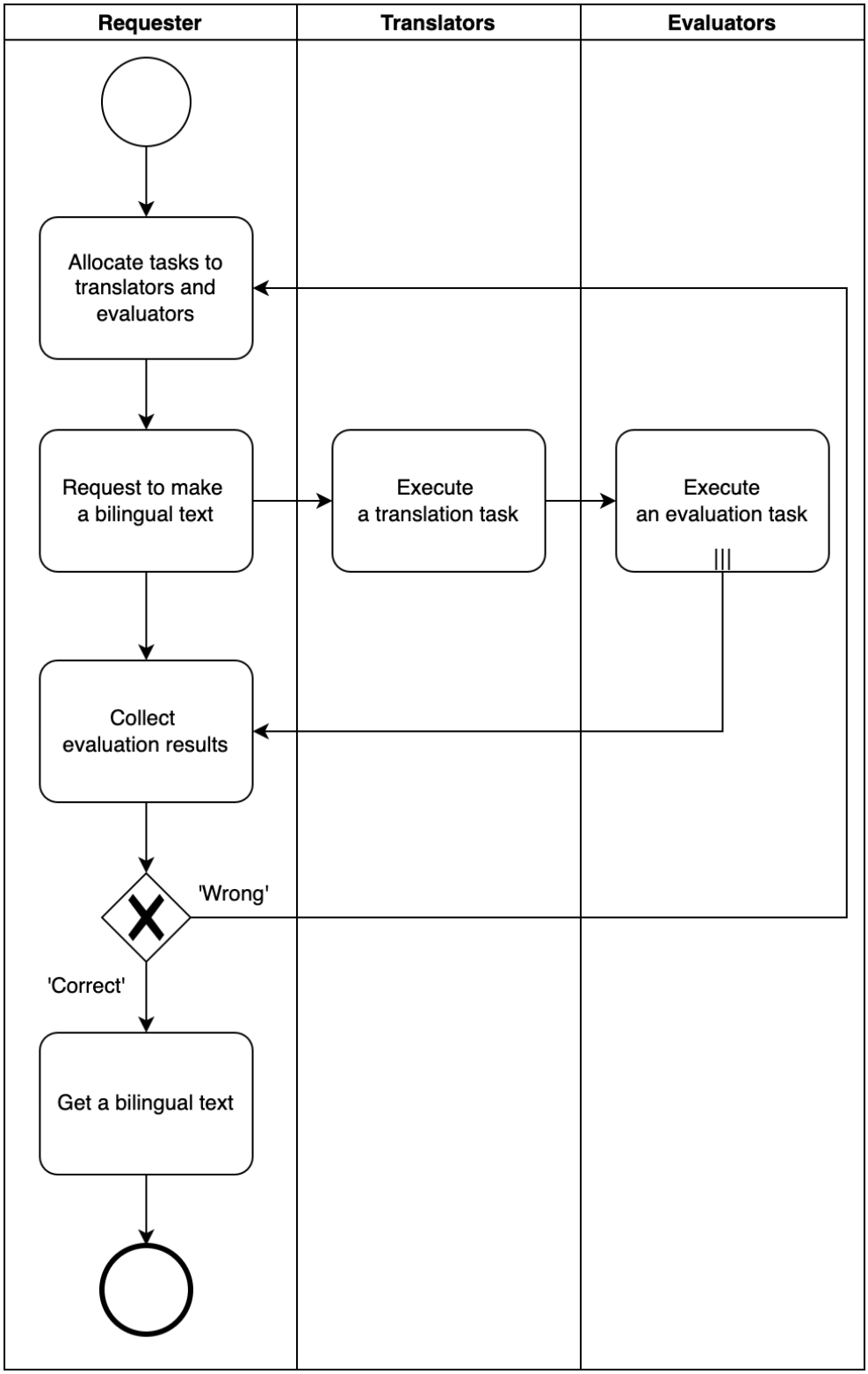

図 1: 対訳辞書作成におけるワークフロー

第4章　超問題を用いた回答統合手法

4.1　超問題

従来の回答統合手法は，多数派の意見を強調する性質があるために，多数派が誤答するような問題では失敗することが多かった．そこで，このような正しく回答できる少数の専門家の意見を強調し，正しい回答統合結果を得るために．「超問題」とそれに基づく回答統合手法が提案されている [9].

超問題は問題集合 Q の部分集合である．特に要素数 k の超問題を「k-超問題」と表記する．例えば，問題集合として $[1,2,3,4]$ が与えられたときに，その 3-超問題は，$[1,2,3]$, $[,2,4]$, $[1,3,4]$, $[2,3,4]$ である．また，超問題に対する回答を，超問題を構成する回答の連結で表す．ある作業者が問題 1 に ‘○’, 問題 2 に ‘×’, 問題 3 に ‘○’ と答えている場合，超問題 $[1,2,3]$ に対する回答は ‘○×○’ となる．

4.2　超問題に基づく多数決

超問題を用いた具体的な回答統合手法として，超問題に対する回答で多数決を取ることで専門家の意見を強調する．今回は，対訳評価タスクの回答統合に超問題を用いた多数決を用いる．

具体的には，対訳作成タスクで作成されたいくつかの対訳に対する対訳評価タスクをまとめて問題集合 Q を作成し，その中から k-超問題を構築する．そして各超問題に対する作業者の回答の多数決を取り，最終的にその結果を単一問題に分解し，その単一問題に対して再度多数決を取ることで，各単一問題に対する最終的な答えを得る．

以上の手続きを，図 2 を用いて説明する．この例では，5 人の評価者が 4 つの作成された対訳（日本語 - 英語）について “○（正しい）” か “×（間違っている）” かを評価している．まず，k-超問題を構築し，単一問題に対する回答を超問題に対する回答として変換する．この例では $k=3$ としており，$[1,2,3]$, $[,2,4]$, $[1,3,4]$, $[2,3,4]$ の 4 件の超問題を構築する．また，例えば評価者 C の超問題 $[1,2,3]$ に対する回答は ‘○×○ ’ となる．次のステップでは超問題上での多数決を行う．4 件の超問題全てで ‘○○○’ が多数決解となる．3 番目のステップでは，超問題の多数決解から，各単一問題の多数決解への投票を得る．例えば，超問題 $[1,2,3]$ の多数決解 ‘○○○’ から，作成された対訳 1,2,3 のそれぞれに対する ‘○’ という投票を得る．最後のステップでは，この投票に対して多数

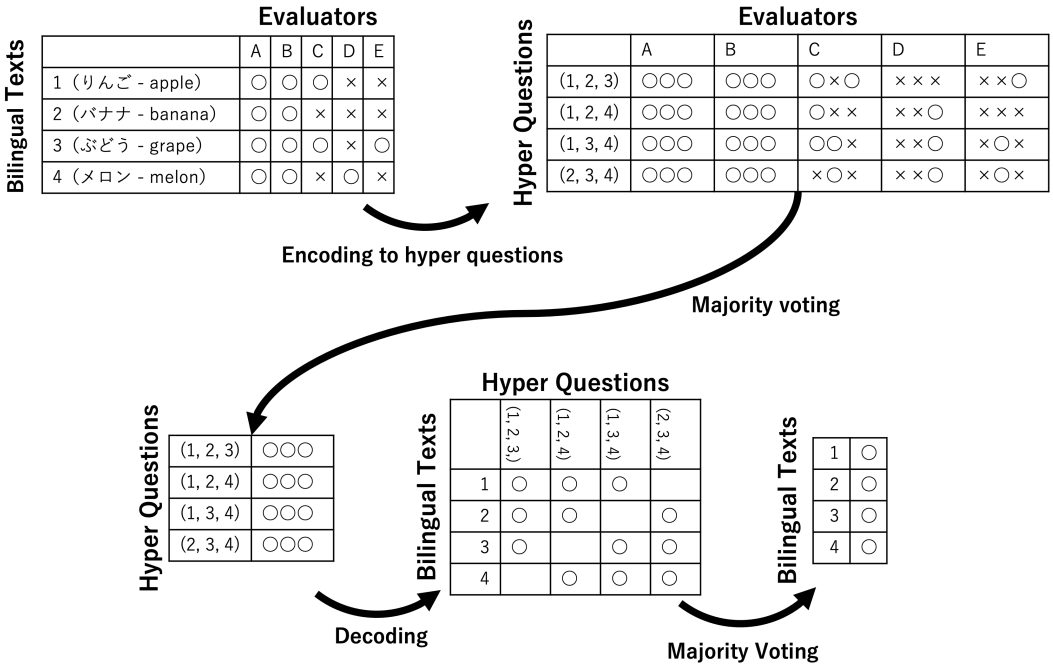

図 2: 超問題を用いた多数決の例

決を行い，各単一問題に対する最終的な答えを得る．この例では，全ての作成された対訳に対する評価の正答が‘○’であり，5 人の評価者のうち 2 名が専門家であり，常に正しい回答を行なっている．単純な多数決は作成された対訳 2 の評価で失敗してしまうが，超問題を用いた多数決を用いることで作成された対訳全てに対して正しい評価を得ることができる．

4.3 評価の再利用

図 3 のように超問題の多数決において多数決解が決まらない場合がある．この例では，個々に分解した後の多数決でもうまくいかず，対訳 1，3 に対する評価が決まらない．このような場合には，分解後の多数決でうまく答えを得ることに成功した問題に全て正解した作業者だけで再度多数決を行うものとする．この例では，対訳 2，4 両方の評価に正解した作業者 A，B，D の対訳 1，3 それぞれに対する評価の多数決を取る．すると，対訳 1，3 に対する評価は‘○’が二つと‘×’が一つとなるため，対訳 1, 3 は正しいと評価が決まる．これにより，作業のやり直しを防ぎ，作業量を削減することができる．

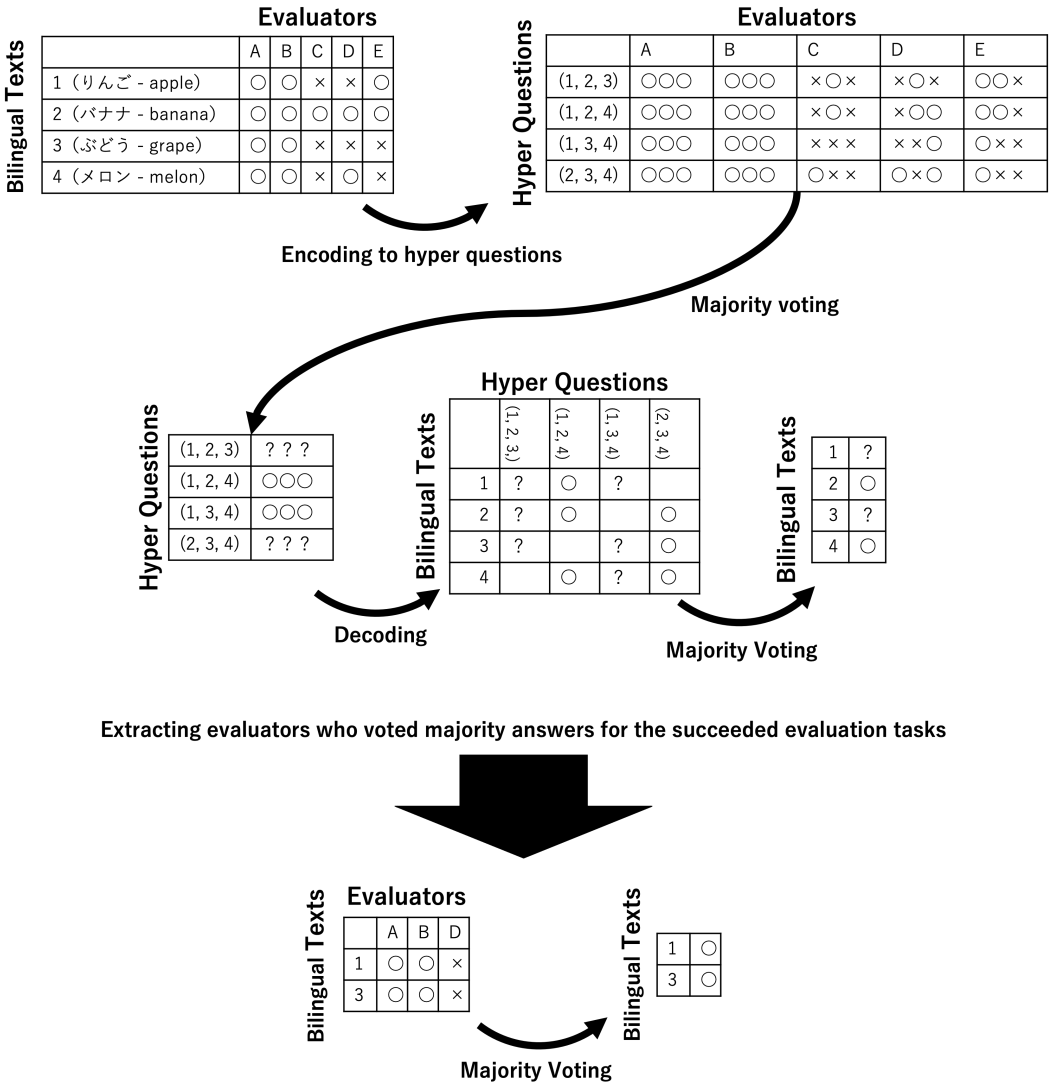

図 3: 超問題を用いた多数決がうまくいかない場合の評価の再利用の例

第5章　信頼値に基づくタスク割り当て手法

本研究では，作業結果を用いて能力が高いと推測される作業者を判別し，彼らに積極的にタスクを割り当てることで，品質の向上とコストの削減を目指す．そのために，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．

5.1　単純な多数決における信頼値評価手法

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．各作業者の対訳作成タスクと対訳評価タスクの作業結果から以下のように信頼値を付与する．

- 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
- 対訳評価タスクにおいて，多数派の評価をした評価者の信頼値を $+1$ する
- 対訳評価タスクにおいて，少数派の評価をした評価者の信頼値を -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．

5.2　超問題に基づく多数決における信頼値評価手法

5.1 章と同様に，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．各作業者の対訳作成タスクと対訳評価タスクの作業結果から以下のように信頼値を付与する．

- ある対訳作成タスクによって作成された対訳が，対訳評価タスクの回答統合の結果“正しい”と評価された場合，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よって作成された対訳が，対訳評価タスクの回答統合の結果“正しい”と評価された場合，その対訳の作成者の信頼値を -1 する

- ある作業者が，ある問題集合 Q に含まれる作成された対訳全てに対してに行なった評価が，対訳評価タスクの回答統合で得られた最終的な評価と過半数以上一致していた場合，その評価者の信頼値を $+1$ する
- ある作業者が，ある問題集合 Q に含まれる作成された対訳全てに対してに行なった評価が，対訳評価タスクの回答統合で得られた最終的な評価と過半数以上異なっている場合，その評価者の信頼値を -1 する

この計算を 1 つ問題集合 Q に含まれる作成された対訳全てに対する評価が完了するたびに行う．

5.3 信頼値に基づくタスク割り当て

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（1）のように計算する．

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

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（1）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（2）

のように計算できる．

$$p_i = \frac{w_i}{w_1 + w_2 + w_3 + \cdots + w_i + \cdots + w_n} \tag{2}$$

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．これによって，品質の向上とともに，コストの削減が期待できる．

第6章　シミュレーションによる評価

6.1　モデリング

シミュレーションによる評価を行うにあたって，低資源言語を対象とした対訳辞書の作成を行うことを想定し，クラウドソーシングの作業者，タスクのモデル化を行う．

6.1.1　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，$x(0 \leq x \leq 1)$ で表す．x が 1 に近づくほど，その作業者が認識している語彙が多くなり，タスクを正しく行うことが可能になる．一方，x が 0 に近づくほど，作業者の認識している語彙が少なくなり，タスクを誤る可能性が高まる．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．本研究では，先行研究に従い，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（3）で表す [10]．

$$f(x|a,v) = \mathrm{Beta}(\frac{a}{\min(a,1-a)v}, \frac{1-a}{\min(a,1-a)v}) \tag{3}$$

ここで，$a \in (0,1)$ 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，[10] で採用されたものである．

6.1.2　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場合は“正しい”対訳を作成し，その単語の対訳を知らない場合は“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

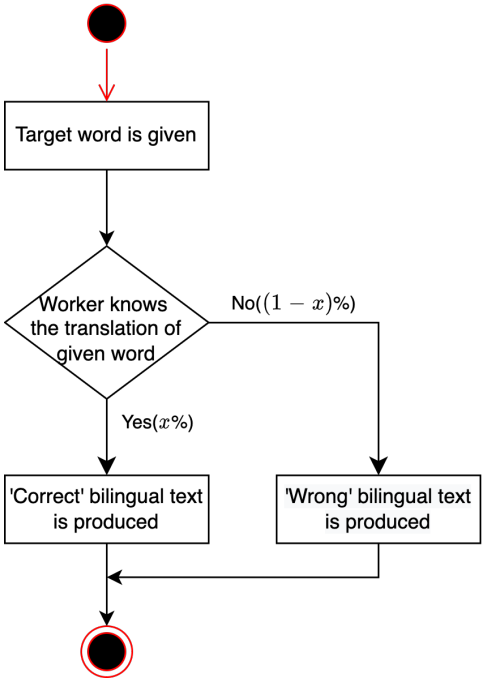

図 4: 対訳作成タスクのモデル

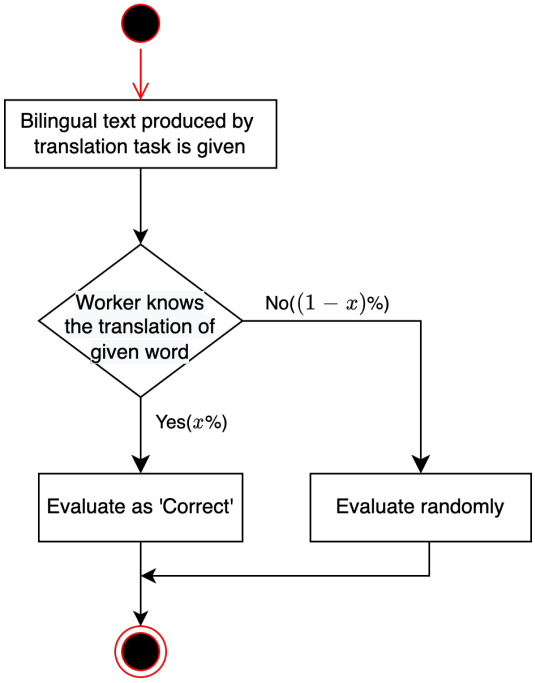

図 5: 対訳評価タスクのモデル

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 4）．しかし，対訳評価タスクは 2 値選択型タスクであるため，作業者が与えられた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 5）．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．

6.2 超問題の最適化

シミュレーションによる評価を行うにあたって．低資源言語を対象とした対訳作成を行うことを想定し，超問題に基づく多数決を対訳評価タスクの解答統合に最適化させる必要がある．

6.2.1 パラメータ

1 つの問題集合 Q に含まれるタスクの数 q と，Q の部分集合である超問題の要素数 k，そして，評価者の人数 n の三つが考えられる．

6.2.2 実験設定

6.2.1 で説明した 3 つのパラメータを最適化するために q を $[3, 4, 5, 6, 7, 8, 9]$，k を $[2, 3, 4, 5, 6, 7, 8]$ に変化させて実験を行った．10 人の作業者が 1000 語の作

成された対訳に対して“正しい”か“間違っている”かを評価をしている場合を想定し，正解率と．作業回数を算出した．まずは，評価者の人数 n を 3 に固定して，専門家の数が専門家の人数が $[2,3,4]$ の場合についての実験を行い，専門家の人数が与える影響を調べた．その後に，専門家の人数を 3 に固定して，評価者の人数が $[3,4,5]$ の場合についての実験を行い，評価者の人数が与える影響を調べた．最後に，q と k をがどのような影響を及ぼすのかを調べた．非専門家はランダムに評価するとし，専門家の正解率は 0.85 とした（能力値 0.7 の場合の対訳評価タスクの正解率）．

6.2.3 結果

表 2, 3, 4 は，評価者の人数が 3 人，専門家の数を $[2,3,4]$ の時の正確性と作業回数の表である．これらの表から，専門家の数が増えれば，正解率が高くなる．作業回数に関しては，k が 3 以下の時は専門家の数を増やしても変化は少なかった．しかし，k が 4 以上の時は，専門家の数が増えれば作業回数が少なくなる傾向にあった．

表 3, 5, 6 は，専門家の人数が 3 人，評価者の人数 $[3,4,5]$ の時の正確性と作業回数の表である．これらの表から，評価者の人数が増えても正確性にはあまり影響がないことがわかった．しかし，評価者の人数が増えると，作業回数が少なくなることがわかった．

全体的に，k を大きくすれば，正解率が上がるが，作業回数が多くなる．q に関しては正解数と作業結果のどちらにもあまり影響がないことがわかった．しかし，k が 4 以上の時，k が 3 以下の時と比べて作業回数が大幅に増えてしまう．そのため，今回は $q=4, k=3, n=5$ を採用する

表 2: $n = 3$, 専門家の数 $= 2$ の時の正確性と作業回数

	k=2	k=3	k=4	k=5	k=6	k=7	k=8
q=3	65.97 6586.8	-	-	-	-	-	-
q=4	62.47 4770.6	66.72 11422.2	-	-	-	-	-
q=5	63.36 6006	68.14 14030.1	66.88 28731	-	-	-	-
q=6	62.0 4669.2	65.36 9327	70.84 40680	69.7 53581.8	-	-	-
q=7	64.64 5688.3	67.21 10817.1	71.66 35593.8	70.36 79057.8	71.15 95809.5	-	-
q=8	62.62 4643.1	68.46 11977.2	67.89 23438.4	74.84 91704.6	74.27 144061.8	73.47 166977.6	-
q=9	64.12 5519.4	66.96 9721.5	70.99 29232	75.19 90760.8	77.77 267559.8	79.62 296641.2	79.74 297824.3

表 3: $n = 3$, 専門家の数 $= 3$ の時の正確性と作業回数

	k=2	k=3	k=4	k=5	k=6	k=7	k=8
q=3	73.23 6307.5	-	-	-	-	-	-
q=4	69.83 4542.6	75.64 10258.8	-	-	-	-	-
q=5	72.21 5628.6	78.22 12382.2	76.94 22837.5	-	-	-	-
q=6	69.54 4557.6	74.26 8541.0	80.45 31653.6	79.87 39588.6	-	-	-
q=7	71.9 5360.1	75.92 9425.7	82.37 26271.3	82.28 53030.7		-	-
q=8	82.57 63101.7	77.85 11079.9	78.15 19518.9	84.75 60927.6	86.73 88812.0	85.45 103048.8	-
q=9	71.71 5285.7	76.6 8687.7	82.13 22168.5	87.12 59193.6	88.83 137662.2	89.01 159159.9	89.53 160093.3

表 4: $n = 3$, 専門家の数 $= 4$ の時の正確性と作業回数

	k=2	k=3	k=4	k=5	k=6	k=7	k=8
q=3	78.65 5920.2	-	-	-	-	-	-
q=4	75.76 4349.4	82.45 9096.6	-	-	-	-	-
q=5	77.27 5274.0	84.31 10426.2	84.42 18045.0	-	-	-	-
q=6	74.98 4318.5	79.99 7335.6	86.32 22834.8	86.25 29190.0	-	-	-
q=7	77.73 4955.7	82.88 8030.4	90.04 20025.3	88.64 35941.2	88.41 43559.7	-	-
q=8	75.95 4234.2	85.03 9154.8	86.49 14637.9	92.02 40358.4	91.86 55188.0	91.83 66043.2	-
q=9	77.28 4822.8	82.36 7350.3	88.28 16596.6	92.85 38304.6	92.31 80604.6	92.45 93725.4	93.5 94562.5

表 5: $n = 4$, 専門家の数 $= 3$ の時の正確性と作業回数

	k=2	k=3	k=4	k=5	k=6	k=7	k=8
q=3	71.77 6623.6	-	-	-	-	-	-
q=4	71.95 5476.0	74.95 6976.8	-	-	-	-	-
q=5	71.15 797.6	74.63 7862.4	76.74 16358.0	-	-	-	-
q=6	72.54 5791.6	74.24 7304.0	78.99 16770.0	79.81 26000.8	-	-	-
q=7	72.29 5807.6	75.02 6962.0	77.86 13402.4	84.78 30671.2	83.12 43532.4	-	-
q=8	71.33 5619.2	75.74 6979.2	77.47 12798.8	86.72 31314.8	86.85 4376.4	86.07 72339.2	-
q=9	71.37 5906.8	73.56 6919.6	80.19 13630.8	83.57 28017.6	90.25 83781.2	90.16 106211.2	89.57 107150.0

表 6: $n = 5$, 専門家の数 $= 3$ の時の正確性と作業回数

	k=2	k=3	k=4	k=5	k=6	k=7	k=8
q=3	74.83 10732.0	-	-	-	-	-	-
q=4	72.06 6871.0	72.47 7355.5	-	-	-	-	-
q=5	75.03 8649.0	73.96 8315.5	77.59 12415.5	-	-	-	-
q=6	72.33 6561.0	74.0 7592.5	78.93 12106.0	81.08 20125.0	-	-	-
q=7	73.99 7663.5	75.27 7079.0	79.3 10902.0	83.21 21666.5	82.68 34568.5	-	-
q=8	72.48 6329.5	74.64 6891.0	77.93 10497.0	84.33 20835.5	87.63 37445.0	86.02 52688.0	-
q=9	73.54 7041.5	74.6 7076.0	80.95 10993.5	83.18 20269.0	88.45 51442.5	88.03 76268.5	88.34 78381.5

6.3 評価方法

提案手法を含む手法に対して，作成された対訳の正確性，全ての対訳が得られるまでにかかった総作業量の観点から評価を行う．

1. 提案手法 1（Reliable_hyper_reuse）
   超問題を用いた回答統合手法と，信頼値に基づくタスク割り当て手法を組み合わせたモデル．超問題の多数決が失敗した場合には評価の再利用を行う．
2. 提案手法 2（Reliable_hyper）
   超問題を用いた回答統合手法と，信頼値に基づくタスク割り当て手法を組み合わせたモデル．
3. 比較手法 1（Random_hyper）
   超問題を用いた回答統合手法を用いたモデル．タスクの割り当ては，全て作業者全体から無作為に行う．
4. 比較手法 2（Reliable）
   信頼値に基づくタスク割り当て手法を用いたモデル．対訳評価タスクの回答統合には単純な多数決を用いる．
5. 比較手法 3（Random）
   作業者全体から無作為にタスク割り当てを行い．対訳評価タスクの回答統合には単純な多数決を用いるモデル．

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，全ての対訳が得られるまでに必要な作業量の測定を行う．

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. 全ての対訳が得られるまでに必要な作業量
   作業量は、すべての対訳が得られるまで実行された対訳作成タスクと対訳評価タスクにかかる合計ユニットタイムであるとする．ユニットタイムは各タスクの実行にかかる推定時間から算出され，対訳作成タスクは対訳評価タスクに比べて難易度が高いため，対訳作成タスクは 3 ユニット，対訳評価タスクは 1 ユニットの時間がかかると定義した．このモデルは [11] で採用されたものである．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．

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：20 人
- 作業者の能力：6.1.1 のモデルに基づいてに決定し，作業者の能力の平均 a は 0.2 から 0.7 の間で変化させて比較を行う．分散 v は 0.5 とする．作業者の能力値の分布は図 6 の通りである．
- 1 つの問題集合 Q に含まれるタスク数 q：4 個
- 超問題の要素数 k：3 個
- 評価者の人数 n：5 人

また，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．

6.4 結果

6.4.1 正確性

提案手法である Reliable_hyper_reuse と Reliable_hyper の正確性が高く，Reliable, Random_hyper, Random と続いた．Reliable_hyper_reuse と Reliable_hyper の正確性はほとんど同じだったが，Reliable_hyper の方が若干低かった．正確性が一番高かった Reliable_hyper_reuse, Reliable_hyper と，3 番目の Reliable の間には 5〜15%ほどの差があった（図 7）．

6.4.2 作業量

超問題を用いた回答統合手法を使用する, Reliable_hyper_reuse，Reliable_hyper，Random_hyper の作業量が多くなる傾向にあった．しかし，Reliable_hyper_reuse と Reliable_hyper に関しては作業者の能力値の平均が 0.5 以上の場合，単純な多数決を用いるモデルよりも少なくなった（図 8）．

6.4.3 信頼値

信頼値に基づくタスク割り当て手法を用いる Reliable, Reliable_hyper, Reliable_hyper_reuse については，作業者の能力値と信頼値の関係についての調査を行った（図 9, 10, 11）．どの手法に関しても能力の高い作業者の信頼値は高くなる傾向にあった．しかし，超問題を用いた回答統合手法を使用する，Reliable_hyper と Reliable_hyper_reuse に関しては，作業者の能力値の平均 a が 0.3 以下の場合，

精度があまり高くなかった．そして，Reliable_hyper と Reliable_hyper_reuse において a が 0.4 以上の場合，能力値が 0.6 以上になると能力値と信頼値の相関が見られたが，能力値が 0,6 以下の作業者の信頼値にほとんど差はなかった．

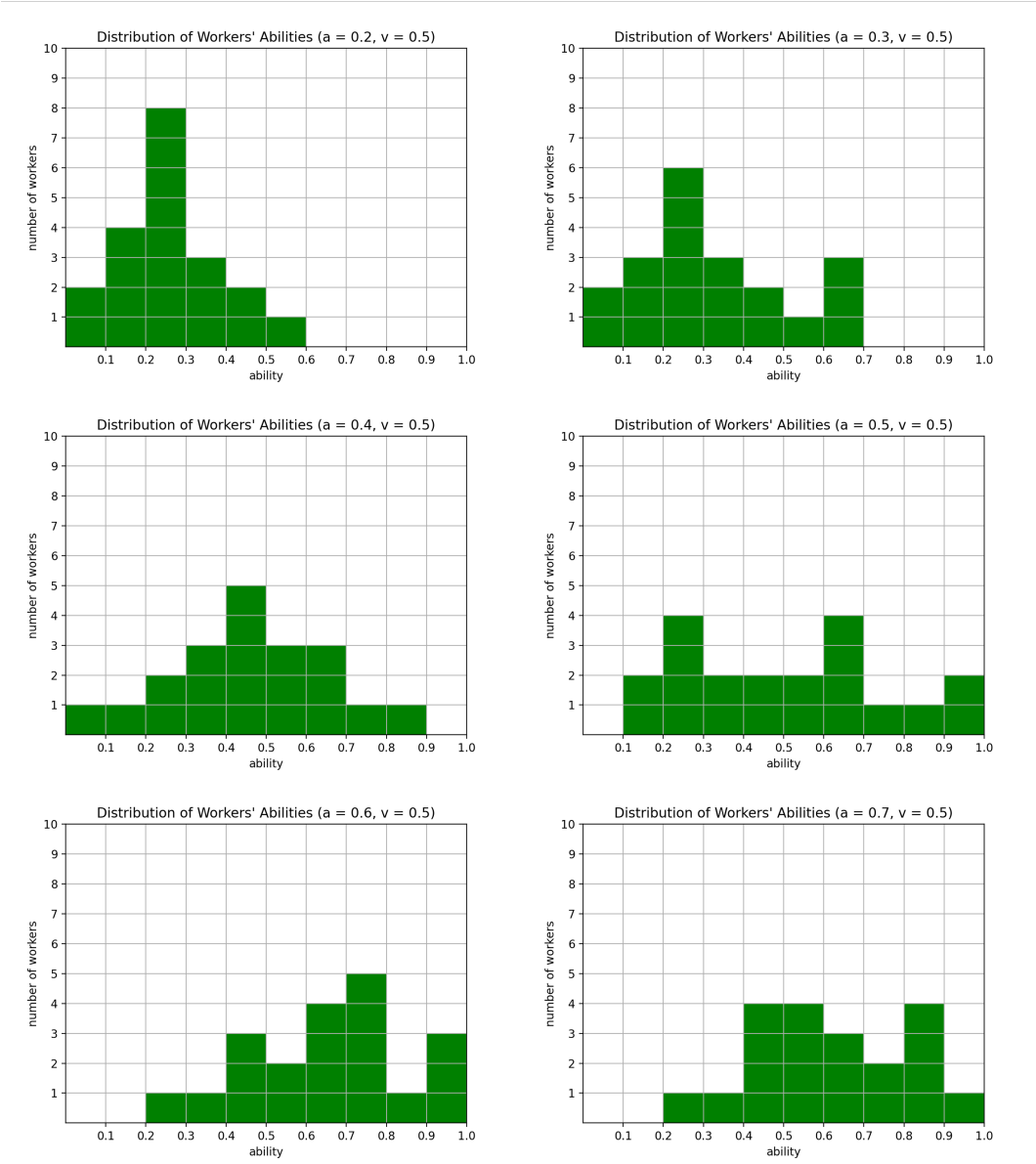

図 6: 作業者の能力値の分布

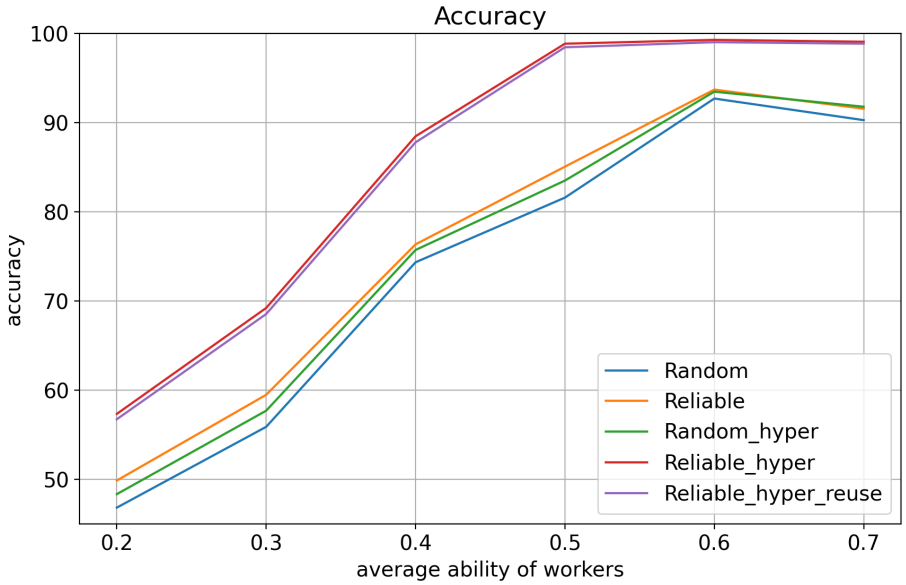
Accuracy
accuracy
average ability of workers
Random
Reliable
Random_hyper
Reliable_hyper
Reliable_hyper_reuse

図 7: 正確性

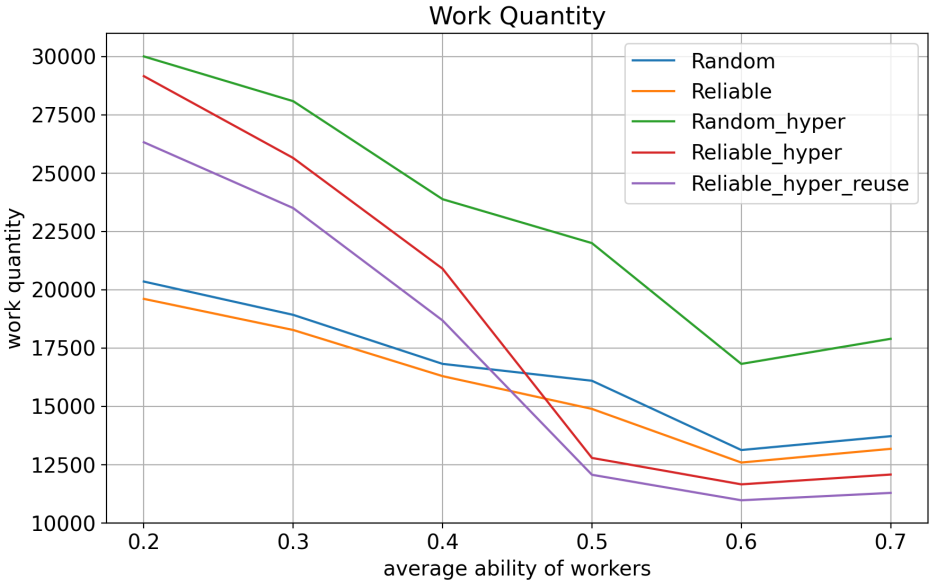
Work Quantity
work quantity
average ability of workers
Random
Reliable
Random_hyper
Reliable_hyper
Reliable_hyper_reuse

図 8: 作業量

6.5 考察

6.5.1 正確性

Reliable と Random_hyper はどちらも Random より正確性が高かったことから，信頼値に基づくタスク割り当て手法と，超問題を用いた回答統合手法が有効であることがわかった．そして，Reliable と Random_hyper を比べると，Reliable の方が正確性が高かったことから，回答統合の精度を高めるよりも，能力の高い作業者にタスクを割り当てる方が効果的であることがわかった．さらに，信頼値に基づくタスク割り当て手法と，超問題を用いた回答統合手法を組み合わせた Reliable_hyper_reuse と Reliable_hyper の正確性が特に高かったことから，これらの手法を個々に用いるよりも，組み合わせた場合に特に有効であることがわかった．そして，Reliable_hyper_reuse と Reliable_hyper を比べると，Reliable_hyper の方が若干正確性が低かったことから，評価の再利用をすると，少し回答の信頼性が下がることがわかった．

6.5.2 作業量

超問題を用いた回答統合手法を使用する，Reliable_hyper_reuse，Reliable_hyper，Random_hyper の作業量が多くなる傾向にあったことから，やり直しが数多く発生していることが容易に想像できる．これは，超問題での多数決は，通常の多数決と比べて決まりづらく，対訳評価タスクの回答統合で評価が決まらない場合があることが原因だと考えられる．しかし，評価の再利用を行う Reliable_hyper_reuse の作業量が Reliable_hyper，Random_hyper と比べて少ないことから，評価の再利用がやり直しの削減に効果があることがわかる．そして，作業者の能力値の平均が 0.5 以上の場合，Reliable_hyper_reuse，Reliable_hyper の作業量が単純な多数決を用いるモデルよりも少なくなった．このことから，能力の高い作業者が一定数以上存在する群衆の中から，高品質な作業者に対訳評価タスクを割り当てることができれば，寧ろ超問題での多数決は一致しやすく，超問題の多数決で評価が決まらない場合のやり直しが発生しづらいと考えられる．さらに，Reliable_hyper_reuse，Reliable_hyper では，対訳作成タスクも信頼値の高い作業者に優先的に割り当てられるため，そもそも間違った対訳が作成されることが少ないことも作業量が削減された理由の一つだと考えられる．高品質な作業者（能力値が 0.7 以上の作業者）の人数については，図 6 の $a = 0.4$ と $a = 0.5$ のグラフが示すように，作業者の能力値が 0.4 の時は 2 人，0.5 の時

は 4 人であった．このことから，高品質な作業者が 2 人では少なすぎると考えられる．なぜなら，2 人の高品質な作業者は対訳作成タスクに割り当てられることが多く，対訳評価タスクはあまり割り当てられないため，対訳作成タスクにより正しい対訳が作成されたとしても，超問題の多数決がうまく機能していないからであると考えられる．

6.5.3 信頼値

Reliable, Reliable_hyper, Reliable_hyper_reuse のいずれに関しても，能力の高い作業者の信頼値は高くなる傾向にあったことから，信頼値を計算することでうまく作業者の能力を推定することができたと考えられる．しかし，超問題を用いた回答統合手法を使用する，Reliable_hyper と Reliable_hyper_reuse に関しては，作業者の能力値の平均 a が 0.3 以下の場合，精度があまり高くなく，信頼値がマイナスになっていることから，能力の低い作業者同士は超問題の多数決において回答が一致しづらいため，信頼値が下がり続けていると考えられる．これは，単純な多数決を用いる Reliable においては，a が 0.3 以下の場合でもある程度能力値と信頼値に相関が見られることからも言える．そして，Reliable_hyper と Reliable_hyper_reuse において，a が 0.4 以上の場合，能力値が 0.6 以上になると能力値と信頼値の相関が見られたが，能力値が 0,6 以下の作業者の信頼値にほとんど差はなかったのに対して，Reliable においては，a に関係なく，ある作業者集団の中で相対的に能力値が高い作業者は信頼値が高くなる傾向にあった．このことから，超問題を用いた回答統合手法と信頼値に基づくタスク割り当て手法を組み合わせることで，絶対的に能力の高い作業者のみを抽出することができると考えられる．

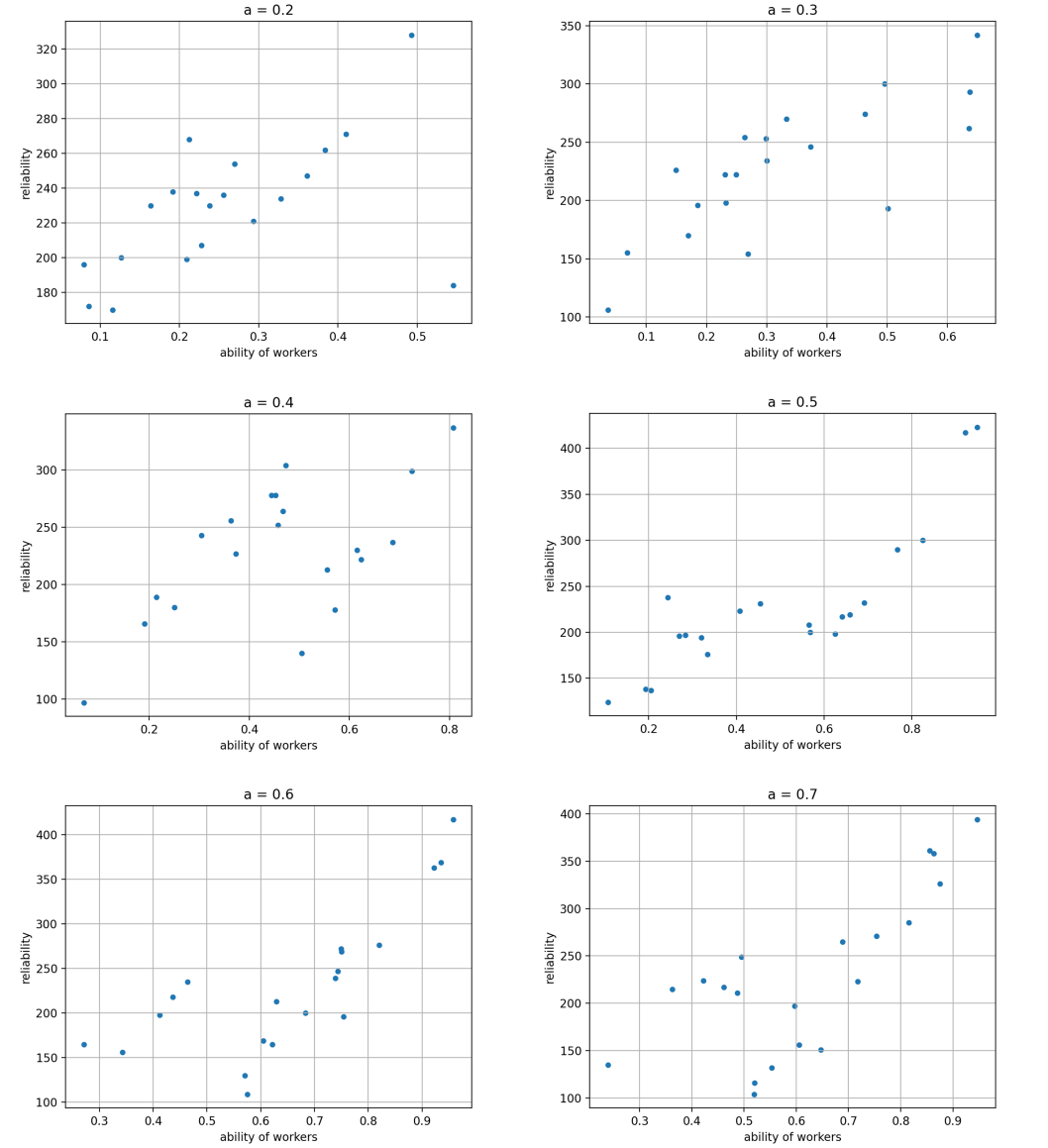

図 9: Reliable を用いたときの作業者の能力値と信頼値の分布

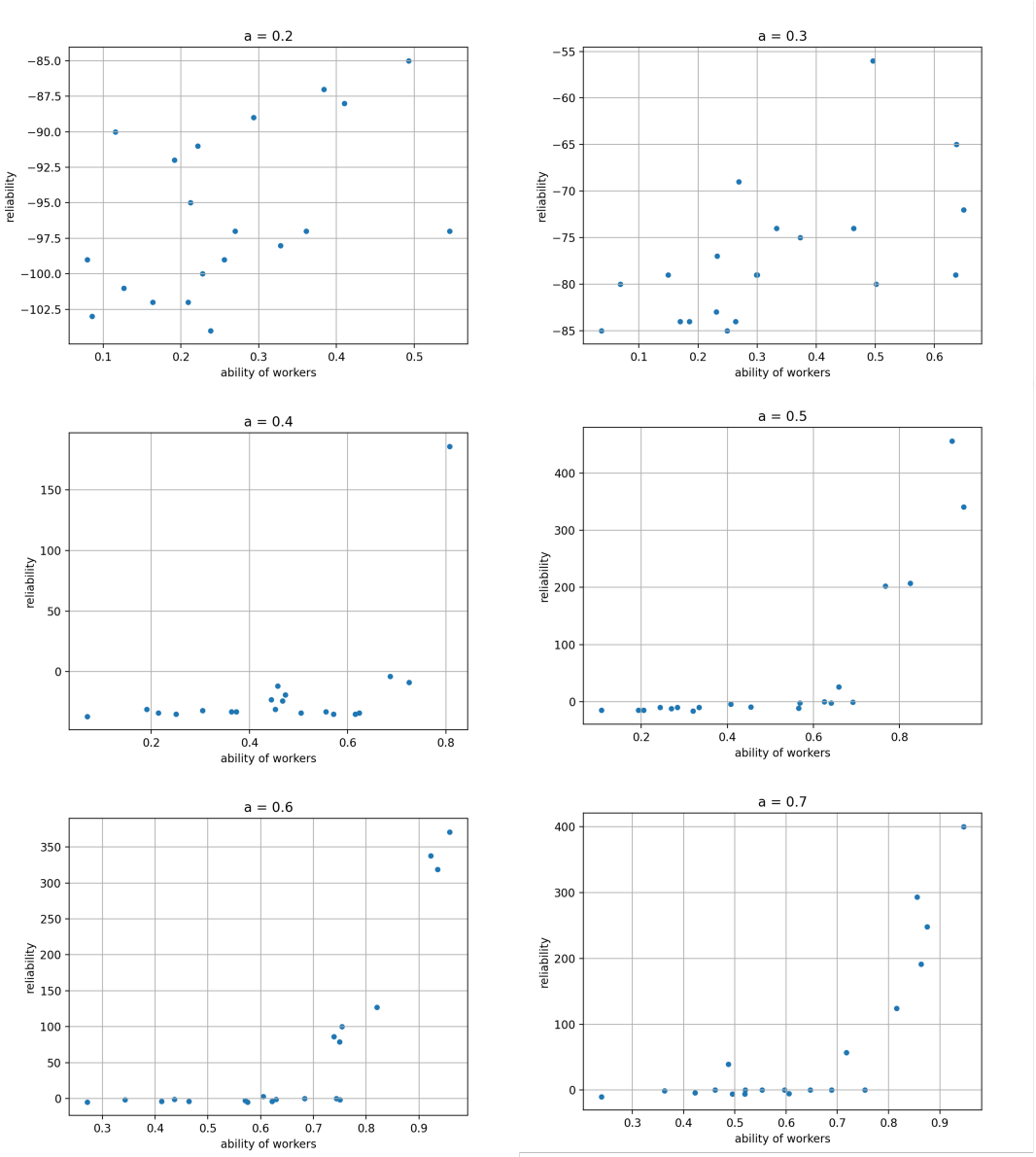

図 10: Reliable_hyper を用いたときの作業者の能力値と信頼値の分布

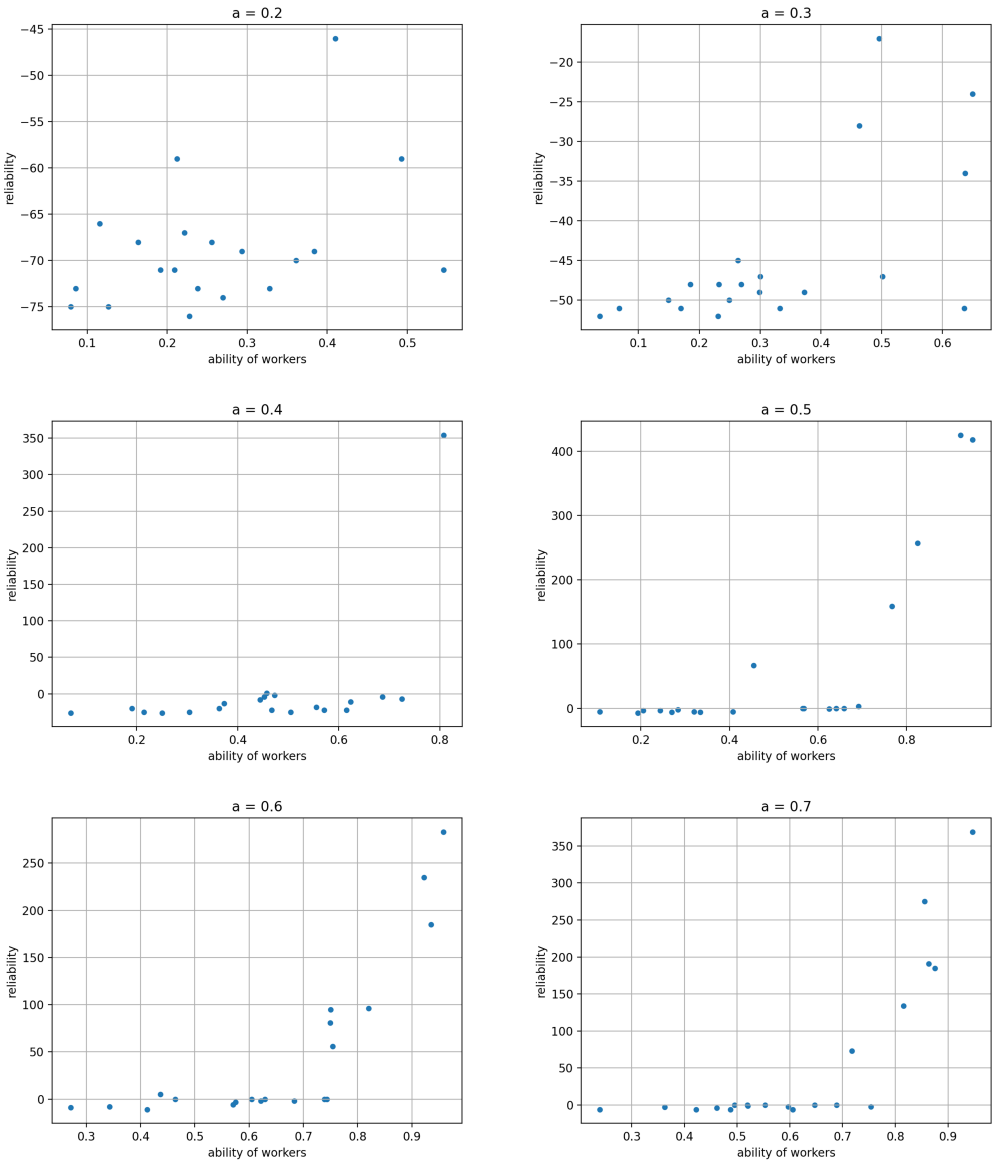

図 11: Reliable_hyper_reuse を用いたときの作業者の能力値と信頼値の分布

第7章　実データによる評価

7.1　実データ

実際の作業者20名にクラウドソーシングで1000語に対する対訳を作成してもらった．それぞれの作業者に1000語全ての対訳を作成してもらい，作成された対訳の中から重複を除いたものに対してさらに“正しい”のか“間違っている”のか評価を行ってもらった．

実際の作業者の対訳作成タスクの正解率と対訳評価タスクの正解率は表7の通りである．作業者の対訳作成タスクの正解率の平均は約66.3%，対訳評価タスクの正解率の平均は約68.6%であった．作業者の対訳評価タスクの正解率の分散は約0.27であった．

7.2　評価方法

6.3章で説明した5つの手法に対して，7.1章で説明した実データを用いてシミュレーションを行なった場合の正確性と作業量の評価を行った．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：20人
- タスクの実行結果：実データにおける各タスクの実行結果を参照
- 1つの問題集合Qに含まれるタスク数q：4個
- 超問題の要素数k：3個
- 評価者の人数n：5人

7.3　結果

実データにおけるそれぞれの手法の正確性と作業量は表8の通りである．

7.3.1　正確性

超問題を用いた回答統合手法を使用する, Reliable_hyper_reuse，Reliable_hyper，Random_hyperの正確性が高く，ほぼ同等だった．そして，単純な多数決を用いるReliable, とRandomは少し正確性が低かった．

7.3.2　作業量

作業量の少ない方から，Reliable_hyper_reuse，Reliable, Random, Reliable_hyper，Random_hypeという順番になった．

表 7: 実際の作業者の対訳作成タスクの正解率と対訳評価タスクの正解率

作業者	対訳作成タスクの正解率	対訳評価タスクの正解率
1	82.5	72.34875445
2	82	90
3	79.8	69.0747331
4	78.9	70.88967972
5	78.6	69.34163701
6	78.1	75.65836299
7	77.3	82.68683274
8	77.1	59.03914591
9	75.5	63.77224199
10	73.7	72.66903915
11	72.5	69.39501779
12	71.9	70.42704626
13	66.7	66.19217082
14	66.1	59.0569395
15	61.4	63.32740214
16	59.5	68.02491103
17	43.5	65.44483986
18	39.5	52.54448399
19	31.8	62.27758007
20	29.6	70.21352313

表 8: 実データにおけるそれぞれの手法の正確性と作業量

手法	正確性	作業量
Random	80.37	11448
Reliable	80.77	11285
Random_hyper	82.03	14419
Reliable_hyper	82.97	12312
Reliable_hyper_reuse	82.6	10813

7.3.3 信頼値

信頼値に基づくタスク割り当て手法を用いる Reliable, Reliable_hyper, Reliable_hyper_reuse については，作業者の能力値（対訳作成タスクの正解率）と信頼値の関係についての調査を行った（図 12, 13, 14）．どの手法に関しても能力の高い作業者の信頼値は高くなる傾向にあった．しかし，超問題を用いた回答統合手法を使用する，Reliable_hyper と Reliable_hyper_reuse に関しては，能力が高いのにも関わらず，信頼値があまり高くない作業者が特に目立った．

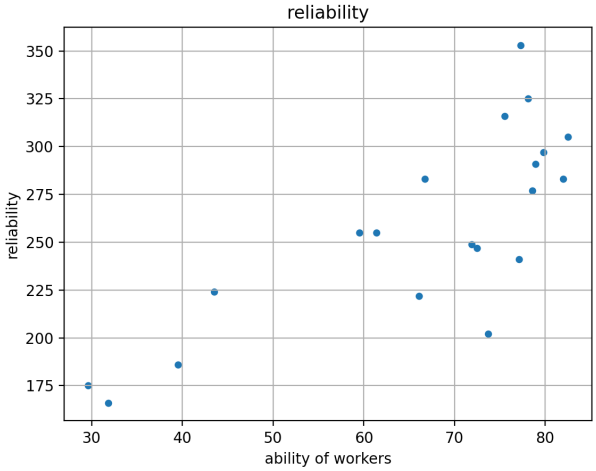

図 12: Reliable を用いたときの作業者の能力値と信頼値の分布

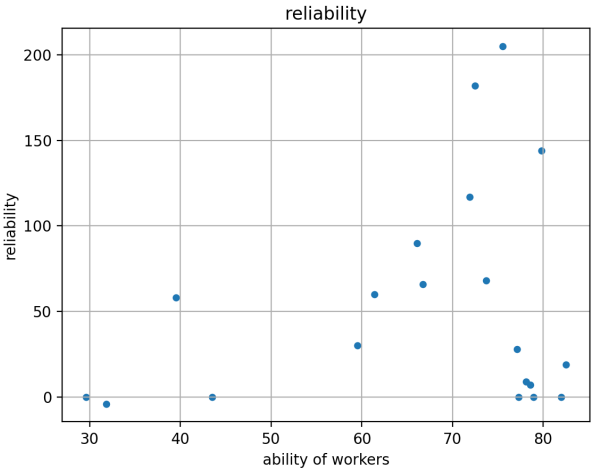

図 13: Reliable_hyper を用いたときの作業者の能力値と信頼値の分布

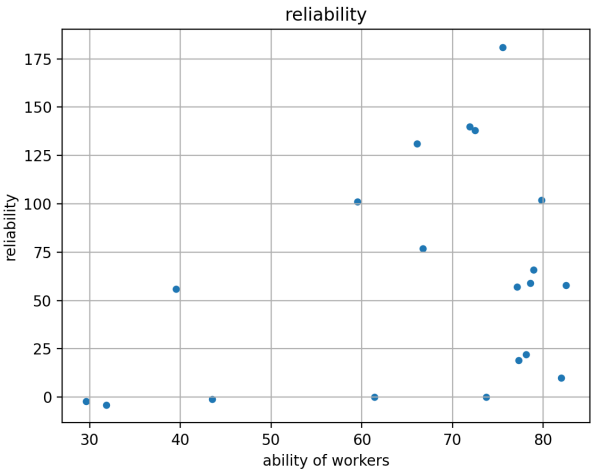

図 14: Reliable_hyper_reuse を用いたときの作業者の能力値と信頼値の分布

7.4 考察

7.4.1 正確性

Reliable_hyper_reuse，Reliable_hyper，Random_hyper の正確性が高かったことから，超問題を用いた回答統合手法が有効であることがわかった．しかし，Reliable_hyper_reuse と Reliable_hyper を比べると，Reliable_hyper_reuse の方が若干正確性が低かったことから，評価の再利用をすると，少し回答の信頼性が下がることがわかった．そして，Random と Reliable の正確性がほとんど変わらず，Reliable_hyper と Random_hyper に関しても同様であるため，信頼値に基づくタスク割り当て手法は正確性を向上させるのにあまり有効ではないことがわかった．これは，作業者間の対訳評価タスクにおける正解率の差があまり大きくないため，信頼値の高い作業者を選出しても，多数決の結果にあまり差が見られなかったことが原因であると考えられる．

7.4.2 作業量

Reliable_hyper，Random_hyper の作業量が多くなる傾向にあったことから，超問題での多数決は，通常の多数決と比べて決まりづらく，対訳評価タスクの回答統合で評価が決まらない場合にやり直しが発生していることが原因だと考えられる．そして，Random と Reliable の作業量がほとんど変わらないにもかかわらず，Reliable_hyper と Random_hyper の作業量を比べると Reliable_hyper の方が少ないことから，超問題を用いた回答統合手法と信頼値に基づくタスク割り当て手法を組み合わせると，作業量の削減には有効であることがわかった．これは，対訳作成タスクを信頼値の高い作業者に優先的に割り当てることで，そもそも間違った対訳が作成されることが少ないことが理由だと考えられるさらに，Reliable_hyper_reuse の作業量が一番少ないことから，評価の再利用が有効であることもわかる．

7.4.3 信頼値

Reliable, Reliable_hyper, Reliable_hyper_reuse のいずれに関しても，能力の高い作業者の信頼値は高くなる傾向にあったことから，信頼値を計算することで，うまく作業者の能力を推定することができたと考えられる．しかし，超問題を用いた回答統合手法を使用する，Reliable_hyper と Reliable_hyper_reuse に関しては，能力が高いのにも関わらず，信頼値があまり高くない作業者が目立った．そこで，各作業者の能力値と割り当てられたタスク数の関係を追加で調査

した（図 15, 16, 17）．すると，Reliable_hyper と Reliable_hyper_reuse に関しては，能力が高いのにも関わらずほとんどタスクが割り当てられていない作業者や，中には一度もタスクが割り当てられていない作業者も存在した．これは，全体的に対訳評価タスクの正解率がそこまで高くないため，対訳評価タスクの回答統合の精度があまり良くないことが原因だと考えられる．超問題の多数決でうまく評価が決まらず，やり直しが発生した結果，信頼値の更新が滞る．すると，一定数の既に信頼値を獲得している作業者にタスクが割り当てられ続ける状態が続くため，一度もタスクを割り当てられていない作業者や，信頼値が低い作業者にタスクが割り当てられる確率が低くなる．そのため，一度もタスクが割り当てられないままの作業者が存在するのだと考えられる．

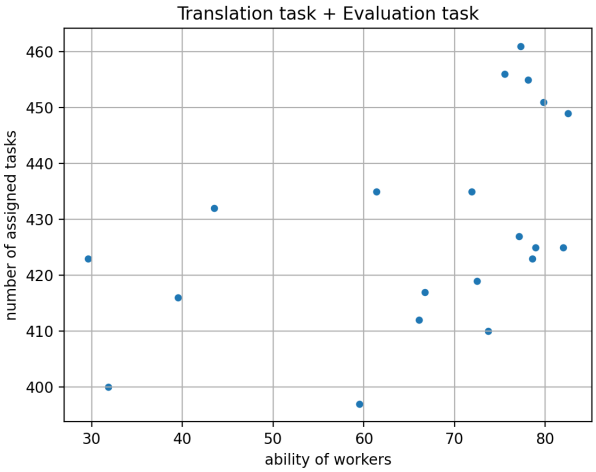

図 15: Reliable を用いたときの作業者の能力値と割り当てられたタスク数の分布

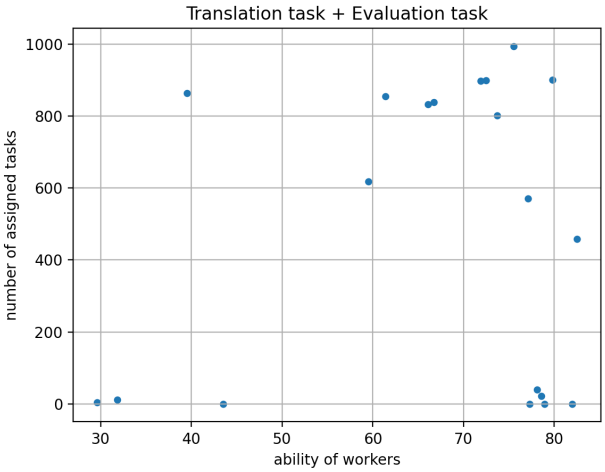

図 16: Reliable_hyper を用いたときの作業者の能力値と割り当てられたタスク数の分布

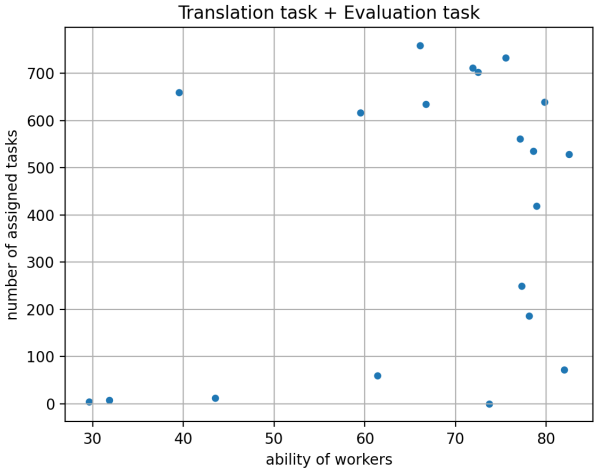

図 17: Reliable_hyper_reuse を用いたときの作業者の能力値と割り当てられたタスク数の分布

第 8 章　回帰モデルによる評価

シミュレーションによる評価では，各作業者の能力値に基づき各タスクの実行結果がに決定するとした．このモデルでは，能力値は作業者が与えられた単語の対訳を知っている確率とし，対訳作成タスクの正解率よりも必ず対訳評価タスクの正解が高くなるように設定した（6.1.2 章）．しかし，実データにおける作業者の対訳作成タスクの正解率と対訳評価タスクの正解率を比べた場合，必ずしも対訳評価タスクの正解率が対訳作成タスクの正解率よりも高くなるということはなかった（表 7）．そこで，回帰分析を用いて実データの作業者の各タスクの正解率の関係を調べた．そして，回帰モデルを用いることで，より実データに近いシミュレーションを行うことができる．

8.1　回帰モデル

今回はサポートベクター回帰を用いた．実データにおける各作業者の対訳作成タスクの正解率と対訳評価タスクの正解率（表 7）の組み合わせ 20 個ののうち 15 個を訓練データとし，残りの 5 個をテストデータとした．そして，2 種類のカーネル関数（linear, rbf）を用いて学習を行い，精度指標として決定係数 R^2 を比較した．その結果，linear カーネルの方が精度が高いという結果になった（図 18）．そのため，linear カーネルを用いた回帰モデルを採用する．

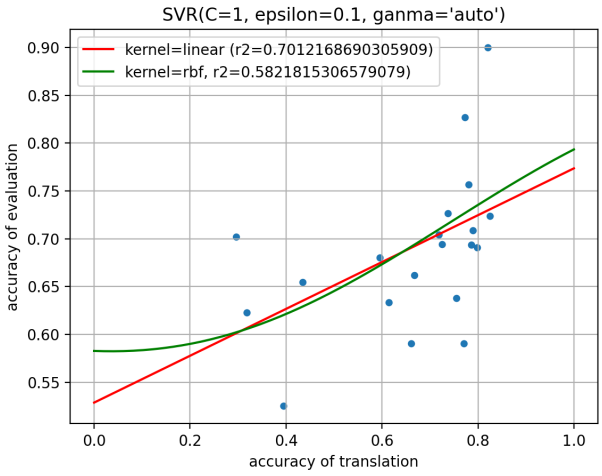

図 18: 実データにおける作業者の各タスクの正解率の分布と回帰モデルによる予測結果

8.2 評価方法

6.3 章で説明した 5 つの手法に対して，回帰モデルを用いて作業者の能力を設定した場合の正確性と作業量の評価を行った．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：20 人
- 作業者の能力：回帰モデルによって決定し，作業者の能力の平均 a は 0.2 から 0.7 の間で変化させて比較を行う．分散 v は 0.27 とする．作業者の能力値の分布は図 19 の通りである．
- タスクの実行結果：実データにおける各タスクの実行結果を参照
- 1 つの問題集合 Q に含まれるタスク数 q：4 個
- 超問題の要素数 k：3 個
- 評価者の人数 n：5 人

8.3 結果

8.3.1 正確性

提案手法である Reliable_hyper_reuse と Reliable_hyper の正確性が高く，Reliable, Random_hyper, Random と続いた．Reliable_hyper_reuse と Reliable_hyper の正確性はほとんど同じだった（図 20）．

8.3.2 作業量

超問題を用いた回答統合手法を使用する, Reliable_hyper_reuse，Reliable_hyper，Random_hyper の作業量が多くなる傾向にあった．Random_hyper の作業量が最も多く，Reliable_hyper, Reliable_hyper_reuse と続いた（図 21）．

8.3.3 信頼値

信頼値に基づくタスク割り当て手法を用いる Reliable, Reliable_hyper, Reliable_hyper_reuse については，作業者の能力値（対訳作成タスクの正解率）と信頼値の関係についての調査を行った（図 22, 23, 24）．どの手法に関しても能力の高い作業者の信頼値は高くなる傾向にあった．

8.4 考察

8.4.1 正確性

シミュレーションによる評価と同じような結果になったが，全体的に正確性が低くなった．これは，回帰モデルではシミュレーションモデルと比べて対訳評価タスクの正解率が低いためだと考えられる．

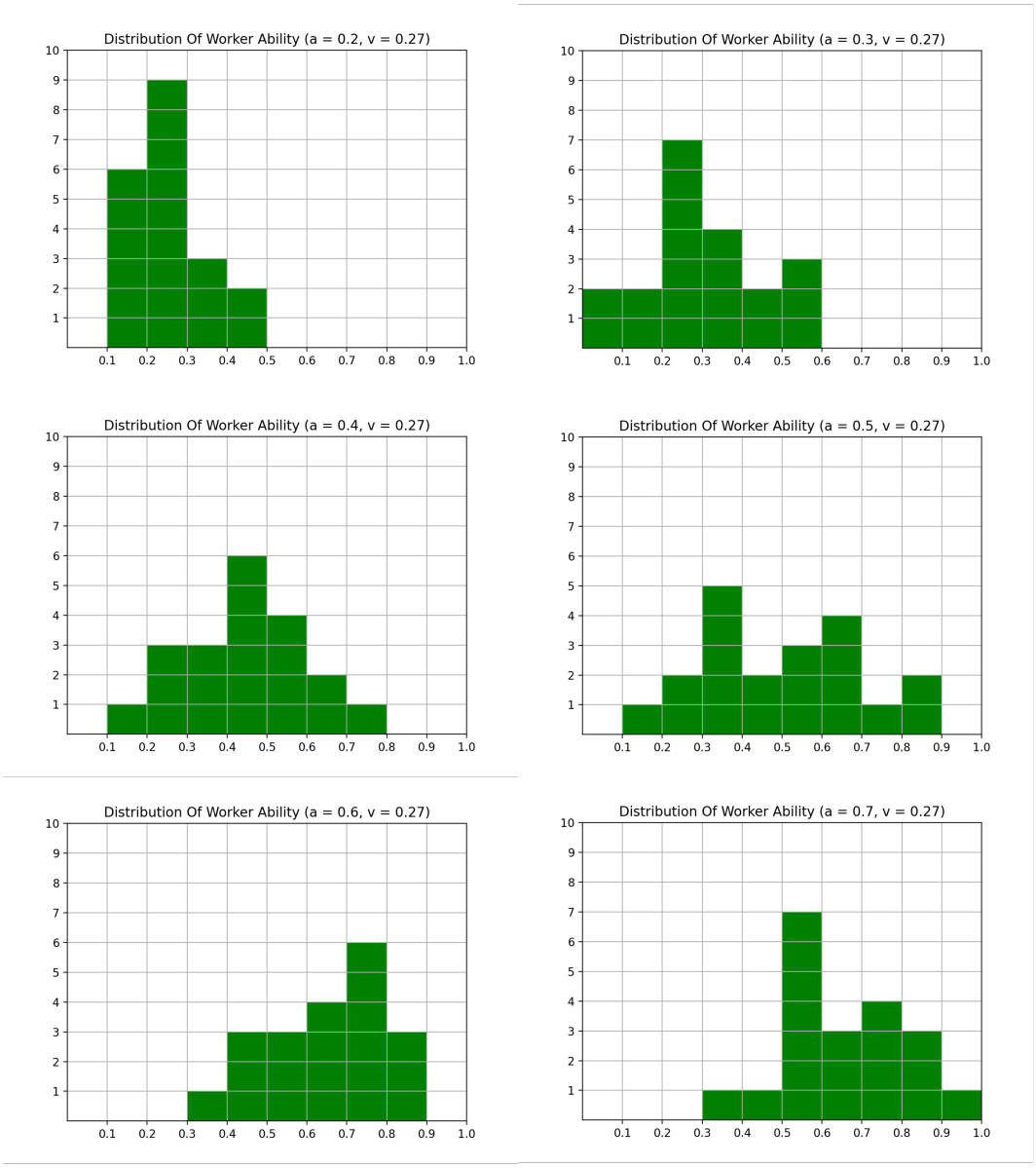

図 19: 作業者の能力値の分布

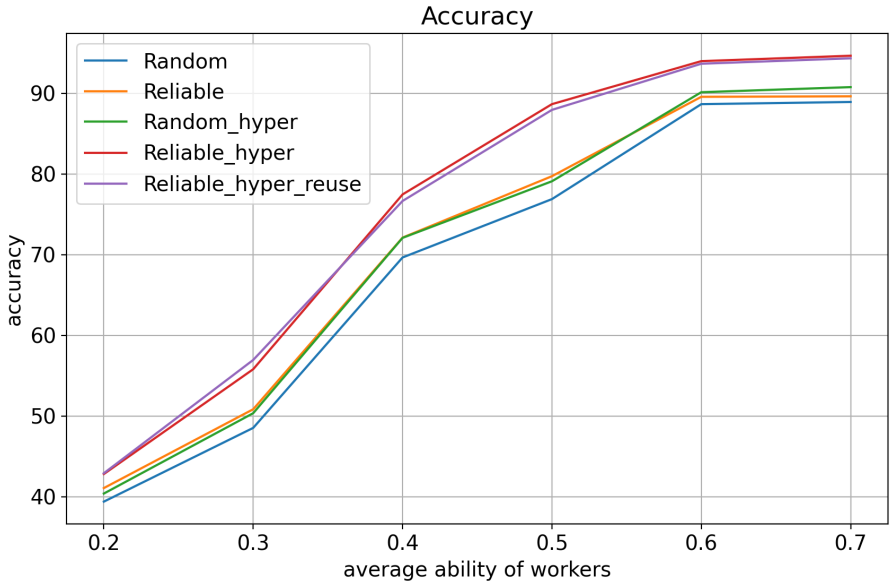

図 20: 正確性

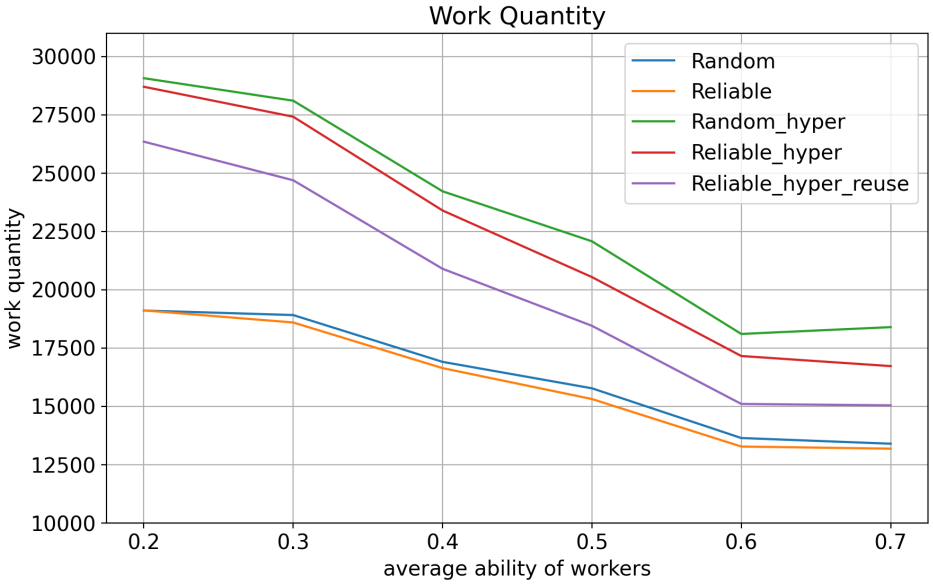

図 21: 作業量

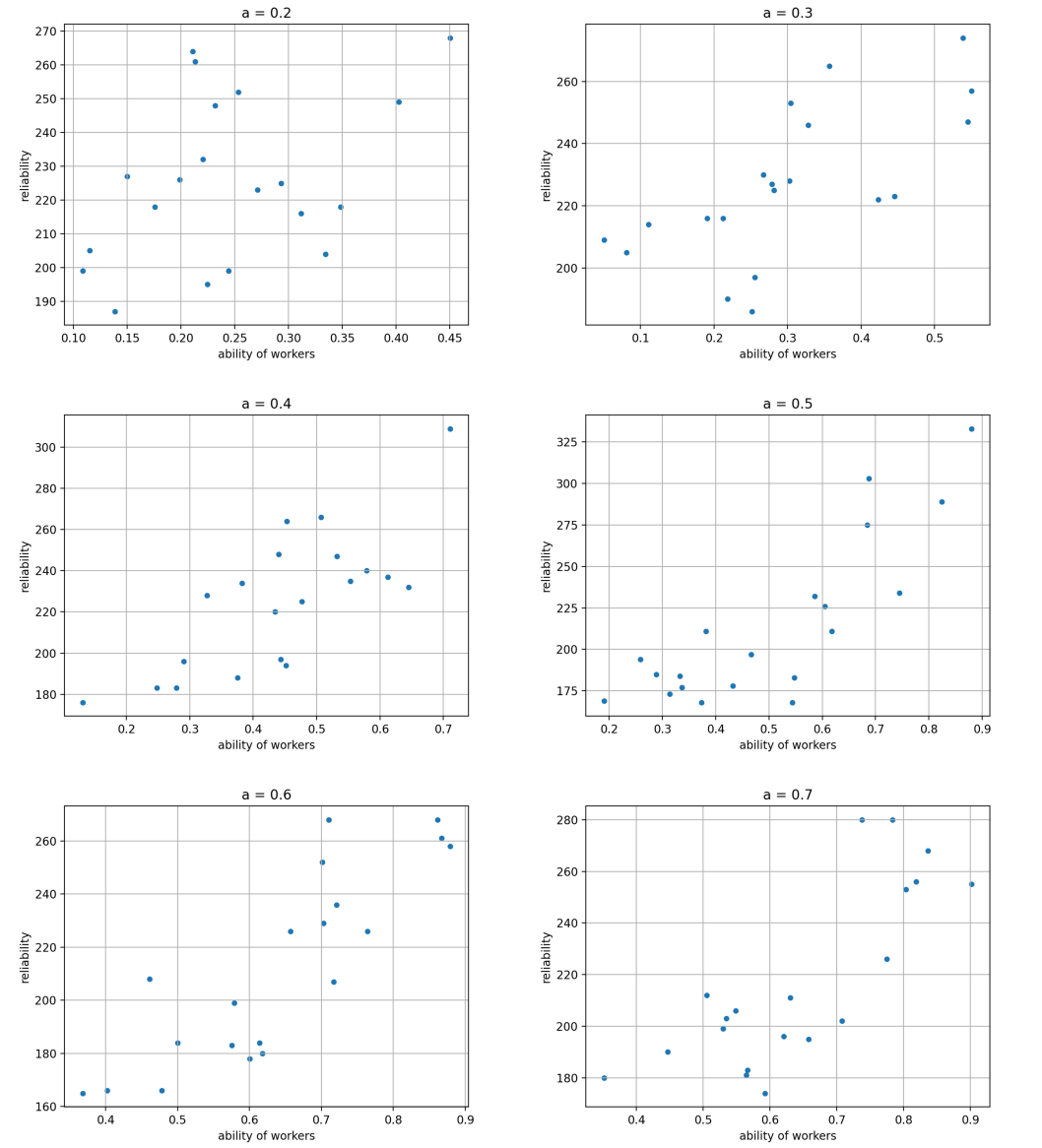

図 22: Reliable を用いたときの作業者の能力値と信頼値の分布

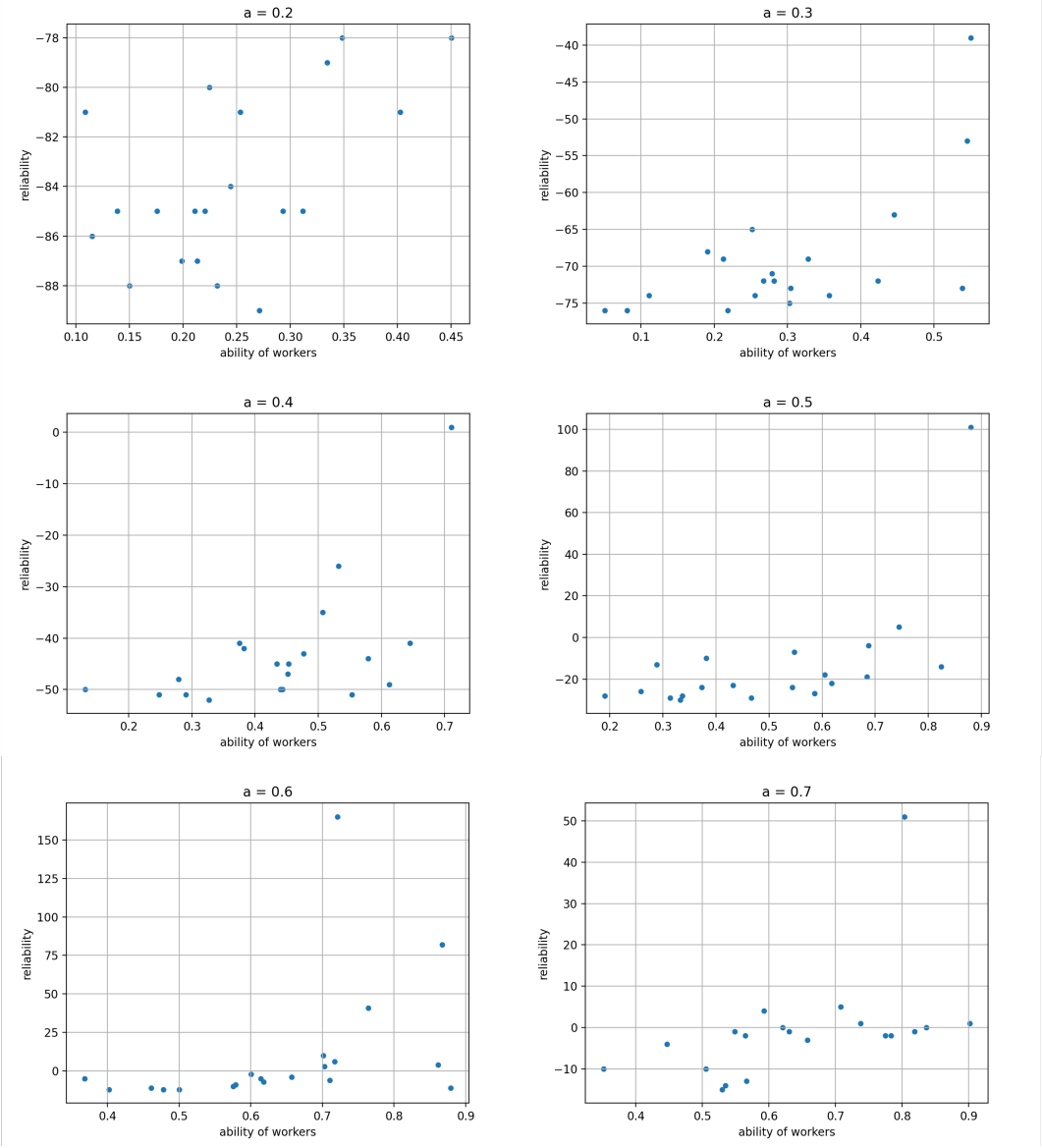

図 23: Reliable_hyper を用いたときの作業者の能力値と信頼値の分布

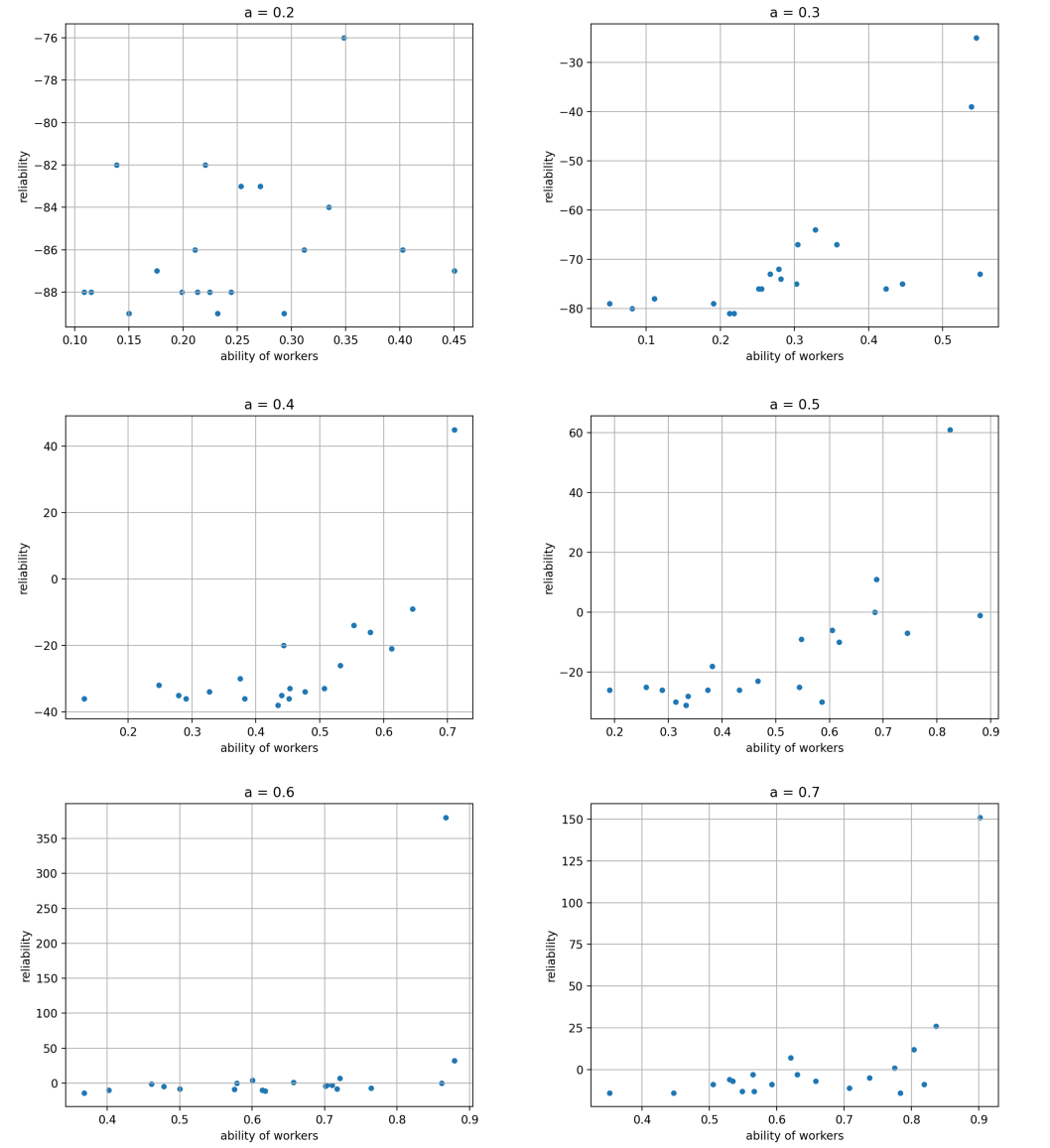

図 24: Reliable_hyper_reuse を用いたときの作業者の能力値と信頼値の分布

8.4.2 作業量

シミュレーションによる評価と同じような結果になったが，作業者の能力値の平均が 0.5 以上の場合でも，Reliable_hyper_reuse，Reliable_hyper の作業量が単純な多数決を用いるモデルよりも少なくなることはなかった．これは，回帰モデルではシミュレーションモデルと比べて対訳評価タスクの正解率が低いため，評価のやり直しが発生することが多いことが原因であると考えられる．

8.4.3 信頼値

シミュレーションによる評価と同じような結果になったが，全体的に信頼値の精度が下がったように感じられる．これは，回帰モデルではシミュレーションモデルと比べて対訳評価タスクの正解率が低いためだと考えられる．そして，実データによる評価と同じように，能力が高いのにも関わらず，信頼値があまり高くない作業者が目立った．さらに，シミュレーションモデルの分布と比べて，より極端に信頼値が高い作業者とそうでない作業者の信頼値の差が激しかった．そこで，各作業者の能力値と割り当てられたタスク数の関係を追加で調査した（図 25, 26, 27）．すると，信頼値の格差ほど割り当てられたタスクの偏りはないことがわかった．このことから，能力の高い作業者でも，対訳評価タスクの正解率がそこまで高くないため，対訳評価タスクに失敗して信頼値を失う場合が多いと考えられる．もしも実際の作業者の対訳評価タスクの正解率が回帰モデルと同様にそこまで高くないとするのならば，対訳評価タスクの結果に基づき信頼値を計算するのは有効でないかもしれない．

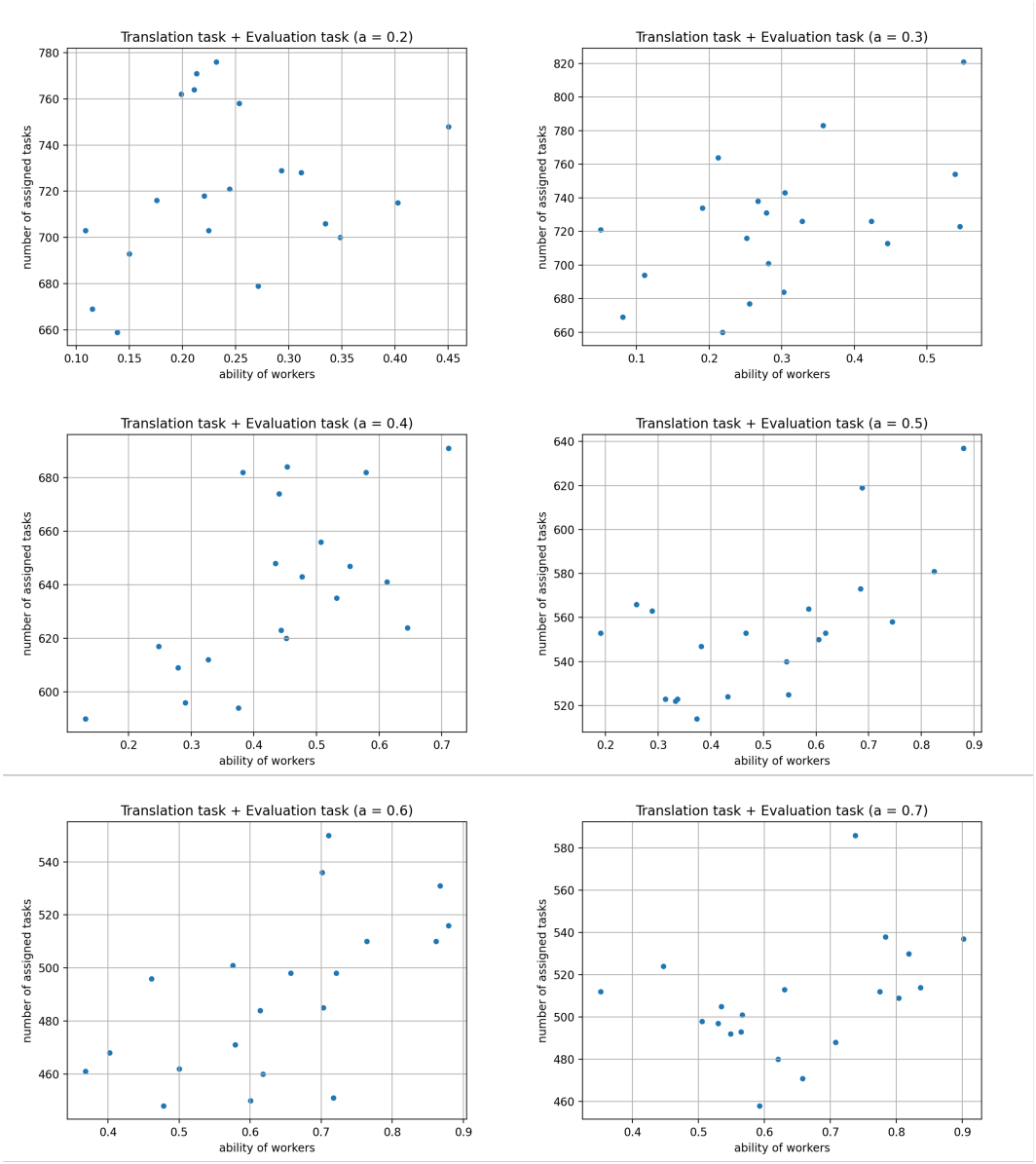

図 25: Reliable を用いたときの作業者の能力値と割り当てられたタスク数の分布

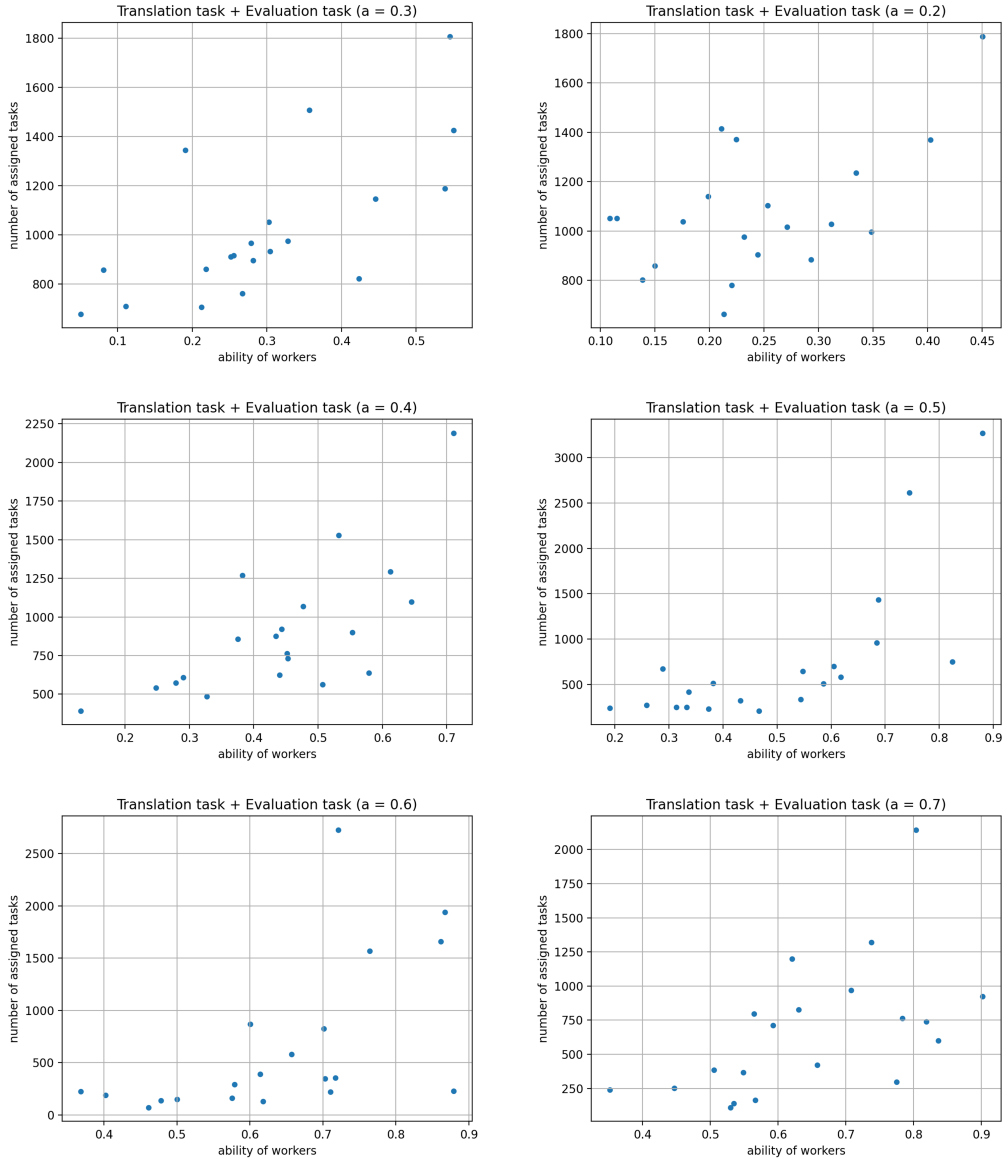

図 26: Reliable_hyper を用いたときの作業者の能力値と割り当てられたタスク数の分布

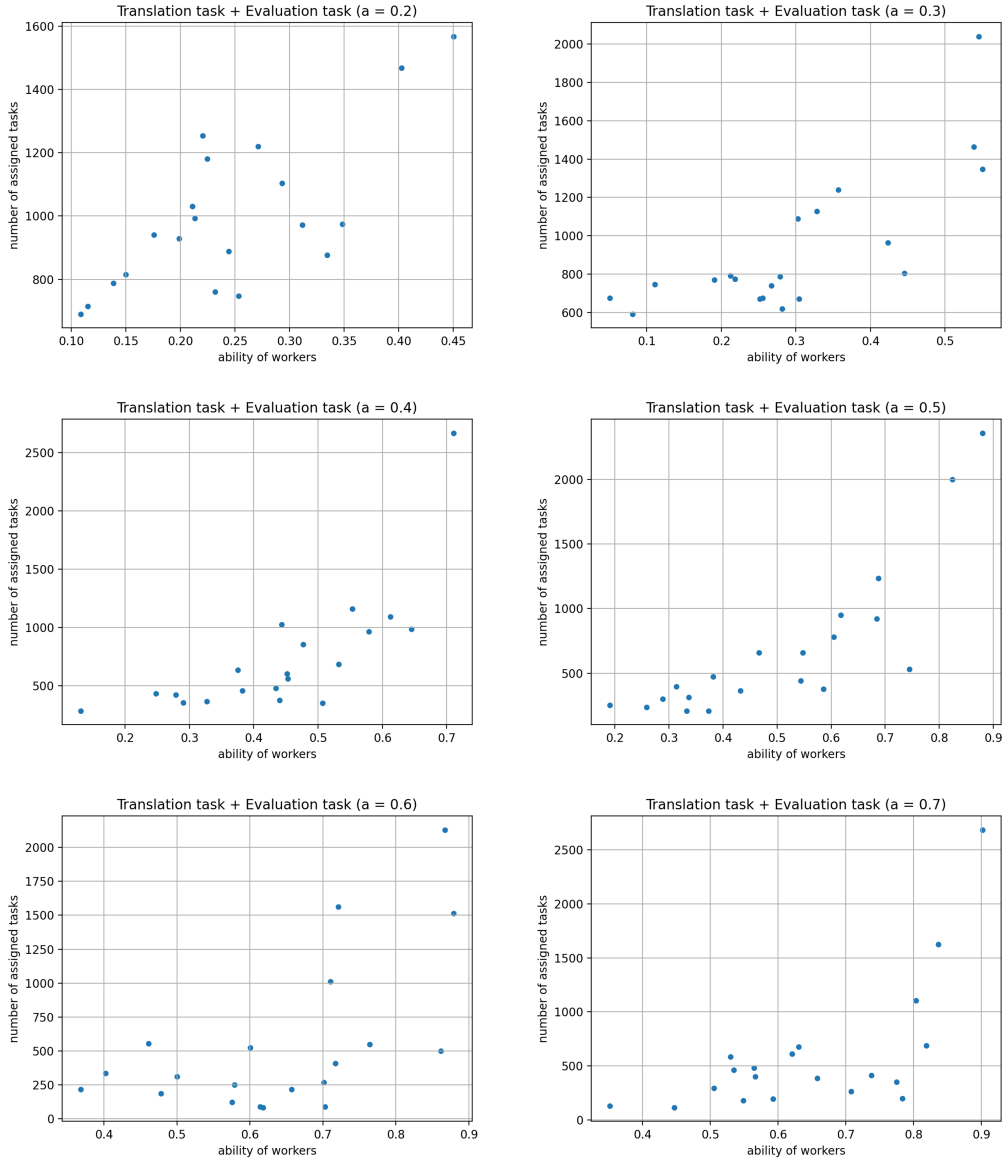

図 27: Reliable_hyper_reuse を用いたときの作業者の能力値と割り当てられたタスク数の分布

第9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，超問題を用いた回答統合手法を用いることで，高信頼な作業者の少ない環境でも正確性を向上させることに成功した．そして，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，能力の高い作業者を推定し，積極的にタスクを割り当てることで正確性の向上と作業量の削減に成功した．また，超問題を用いた回答統合手法と信頼値に基づくタスク割り当て手法を組み合わせることで，それぞれの手法を単体で用いるよりもより効果があることを示した．

本研究の貢献は以下の2点である．

高信頼な評価者の選択

各作業者の作業結果より算出された信頼値に基づき評価者を選出することで，シミュレーションによる評価では，10〜17%ほど正確性を向上させることに成功した．また，実データによる評価では，3%ほど正確性を向上させることに成功した．

作業時間の短縮

信頼値の高い作業者に積極的に対訳作成タスクを割り当て，超問題の多数決が不調だった場合に評価の再利用を行うことで，シミュレーションによる評価において，一定数以上能力値が高い作業者が含まれる場合には，作業量を2000〜2500ユニットほど削減することに成功した．また，実データによる評価では，600ユニットほど削減することに成功した．

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，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ことや，対訳評価タスクに信頼値に基づいた回答の重み付けを用いるなど，改善の余地は大いある．

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，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，Mondheera PITUXCOOSUVARN 助教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*, Association for Computational Linguistics, pp. 212–216 (2010).

[2] 福島拓, 吉野孝, 重野亜久里ほか: 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 (CDS), Vol. 2, No. 3, pp. 23–33 (2012).

[3] 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*, pp. 614–622 (2008).

[4] Zhang, Y. and Van der Schaar, M.: Reputation-based incentive protocols in crowdsourcing applications, *2012 Proceedings IEEE INFOCOM*, IEEE, pp. 2140–2148 (2012).

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

[6] Donmez, P., Carbonell, J. 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. 259–268 (2009).

[7] Kazai, G., Kamps, J., Koolen, M. 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).

[8] Oleson, D., Sorokin, A., Laughlin, G., Hester, V., Le, J. 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).

[9] Li, J., Baba, Y. and Kashima, H.: Hyper questions: Unsupervised targeting of a few experts in crowdsourcing, *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pp. 1069–1078 (2017).

[10] Goto, S., Ishida, T. 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).

[11] Nasution, A. H., Murakami, Y. and Ishida, T.: Plan Optimization to Bilingual Dictionary Induction for Low-resource Language Families, *Transactions on Asian and Low-Resource Language Information Processing*, Vol. 20, No. 2, pp. 1–28 (2021).

[12] Adda, G., Sagot, B., Fort, K. and Mariani, J.: Crowdsourcing for language resource development: Critical analysis of amazon mechanical turk overpowering use, *5th Language and Technology Conference* (2011).

[13] Snow, R., BrendanO ' connor, Jurafsky, D., 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).

[14] 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), 一般社団法人人工知能学会, pp. 2M5OS07b2–2M5OS07b2 (2013).

[15] 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), 一般社団法人人工知能学会, pp. 3M3OS07d4–3M3OS07d4 (2013).

[16] 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).

[17] Hirth, M., Scheuring, S., Hoßfeld, T., Schwartz, C. and Tran-Gia, P.: Predicting result quality in crowdsourcing using application layer monitoring,

2014 IEEE Fifth International Conference on Communications and Electronics (ICCE), IEEE, pp. 510–515 (2014).

[18] Murakami, Y.: Indonesia language sphere: An ecosystem for dictionary development for low-resource languages, *Journal of Physics: Conference Series*, Vol. 1192, No. 1, IOP Publishing, p. 012001 (2019).

[19] Nasution, A. H., Murakami, Y. and Ishida, T.: A generalized constraint approach to bilingual dictionary induction for low-resource language families, *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, Vol. 17, No. 2, pp. 1–29 (2017).