クラウドソーシング用マイクロタスク設計支援のための ユーザフィードバックの収集手法

林　亮太† 清水　伸幸†† 森嶋　厚行†††

† 筑波大学大学院 図書館情報メディア研究科　〒 305–8550 茨城県つくば市春日 1–2
†† ヤフー株式会社　〒 107–6211 東京都港区赤坂 9-7-1 ミッドタウン・タワー
††† 筑波大学 図書館情報メディア系/知的コミュニティ基盤研究センター　〒 305–8550 茨城県つくば市春日 1–2
E-mail: †ryota.hayashi.2014b@mlab.info, ††nobushim@yahoo-corp.jp, †††mori@slis.tsukuba.ac.jp

あらまし　クラウドソーシング用マイクロタスクの設計において質問文が曖昧であると，ワーカがリクエスタの意図しない解釈をしてしまう場合がある．ワーカがリクエスタの意図しない解釈のままタスクに回答することは，タスク結果の品質低下につながる．そこで本論文では，マイクロタスク設計において質問文をよりよいものにすることを目的として，ワーカのフィードバックを利用する手法を提案する．提案手法では，リクエスタが作成した原案となる質問文に対するワーカの様々な解釈を，クラウドソーシングによって収集する．その後，質問文改善に有益と思われるものを抽出し，これをリクエスタに提示することによって，質問文改善の材料を提供する．

キーワード　クラウドソーシング , タスク設計 , コメント解析

1. はじめに

近年，群衆の知や力を利用して，計算機だけでは解決が困難な問題に取り組むクラウドソーシングが注目を集めている．クラウドソーシングとは，リクエスタと呼ばれる問題解決を望む人が，ワーカと呼ばれる不特定多数の人々にタスクの処理を委託することである．クラウドソーシングのうち，委託するものが短時間で処理可能なタスク（マイクロタスク）であるものをマイクロタスク型クラウドソーシングと呼ぶ．

マイクロタスク型クラウドソーシングにおける重要な課題として，どのようにしてマイクロタスク内の質問文の曖昧性を除去するかということが挙げられる．マイクロタスク内の質問文が曖昧であると，質問文が同じタスクでもそれぞれのワーカが様々な解釈をしてしまい，タスク結果の品質が下がってしまう．このような曖昧性を除去するには，多くのワーカたちがとりうる解釈をリクエスタが把握していなければならない．しかしながら，熟練のリクエスタでさえもそのような多様な解釈を把握することは困難である．したがって，ワーカたちの様々な解釈をフィードバックとして提供することは有用であると考えられる．

このような背景の下，我々はマイクロタスクの質問文の解釈を収集する事を提案してきた．本論文では，(1) 我々が構築したコーパス [12] に対するインターアノテータアグリーメントを求め，(2) 集めた解釈の中で有益なものを見つける基準を決めるための主観評価実験を行い，さらに，(3) 機械学習による自動分類を試みた結果を示す．

提案手法全体のポイントは次の 3 つである．

ワーカの解釈を取り扱うための枠組みの提案とコーパスの作成. 本手法では，まず，質問文に対する回答とその理由をワーカに訊ねることによってワーカの解釈を収集する．これは，質問文を改善したいマイクロタスクの最後に，質問文に対する判断の理由を記述する欄を設け，それをクラウドソーシングすることによって行う．例えば，ワーカに対して，生まれたての赤ちゃんの写真を見せて,「この写真には不快な内容が含まれますか?」というタスクがあったとすれば，それに対してワーカは質問そのものに対する回答「いいえ，不快な内容は含まれません」と，その理由「赤ちゃんは可愛く，赤ちゃんを好きでない人はいないから」を答えることになる．本研究では，ワーカの質問文に対する解釈を，質問文への回答とその回答の理由の組として表現し，この組を viewpoint と呼ぶ．先の例では，viewpoint は組 (「いいえ，不快な内容は含まれません」,「赤ちゃんは可愛く，赤ちゃんを好きでない人はいないから」) となる．

我々は，論文 [12] において，viewpoint に関するアノテーションガイドラインを提案し，アノテーションされた viewpoint のコーパスを作成した．本論文ではこのガイドラインに基づいて構築したコーパスに関して，インターアノテータアグリーメントを求めた．

有益な viewpoint の抽出. 収集した viewpoint の数が多くなると，その中からより有益な viewpoint を抽出することは重要な課題となる．本手法では，有益な viewpoint かどうかを判断する基準として汎用性を用いる．汎用性とは，質問文に回答する際に，その viewpoint がより多くのデータに対して有用であるという性質である．クラウドソーシングを用いた主観評価実験により，我々は viewpoint 内の論理構造がこの汎用性を決める重要な要因であることを発見した．この結果は，我々が提案したアノテーションスキーマが，汎用性の高い viewpoint を抽出するのに役立つということを示している．

例えば,「この写真には不快な内容が含まれますか？」という質問に対して「いいえ」と回答し，その回答の理由が，それぞれ「野球のバッドは不快ではないから」「日没の写真はた

いてい不快でないから」「これは果物であり，不快ではないから」である viewpoint があるとする．これらの viewpoint は，「不快でない」ものの名前を単に挙げているだけである．これらの viewpoint から，ワーカが質問文に対してどのように解釈したのかを把握するには，ワーカによって記述された「不快でない」ものを列挙する必要がある．しかしながら，一般に「不快でない」ものというのは非常に多く存在するため，それらをすべて列挙するのは不可能である．それゆえに，ワーカが質問文に対してどのように解釈したのかを把握するには，これらの viewpoint は限定的すぎると考えられる．これに対して，より有益な viewpoint は「ヌードや暴力的な表現が含まれていないから」といったようにワーカの判断基準をはっきり述べているようなものであると考えられる．

5. 節で示すように，汎用性の低い viewpoint 全てが有益でないとはいえないが，汎用性が高い viewpoint はその質問文に対するワーカの解釈を直接表しているといえるため，汎用性はその viewpoint がリクエスタにとって有益かどうかを判断する重要な基準となりえる．

Viewpoint の分類実験．汎用性が高い viewpoint を抽出するための viewpoint に対するラベル付けを，機械学習によって行えることを示す．本手法では，極大部分文字列を特徴量として，アノテーションされたコーパスを学習データとしたサポートベクターマシンを使用する．実験結果は，分類器が 88.3% の精度で viewpoint のラベルを推定できることを示した．インターアノテータアグリーメントのスコアから，この精度は，まだ改善の余地があると考えられる．

本論文は以下のようにして構成される．2. 節では，関連研究について述べる．3. 節では，本論文における viewpoint と，その型の定義を行う．また，コーパス作成のための viewpoint の収集方法と，それらに対するアノテーションについて説明する．4. 節以降が本論文の中心である．4. 節では，構築したコーパスのインターアノテータアグリーメントについて検証を行った結果を報告する．5. 節では，どの viewpoint の型が有益であるかを決定する．6. 節では，viewpoint の型を推定するための分類器や特徴量について説明する．また，分類器の性能についても報告する．最後に，7. 節では，本研究の要点についてまとめ，今後の課題を述べる．

2. 関 連 研 究

マイクロタスク型クラウドソーシングにおける，リクエスタにとっての有益な viewpoint の抽出に関する研究は，我々の知る限りでは存在しない．しかし，関連する研究として次のようなものが挙げられる．

(1) 自然言語処理における文章推測の研究．自然言語処理の分野では，[9], [11] において，質問文内の暗黙の前提の発見を行う研究がなされている．これらの研究は，ある文章から推測できる別の文章を発見する文章推測の研究 [10] の一部である．文章推測の研究では，記述されたテキストの意味や，それらに関連する事実に焦点を当てている．これに対し本研究では，ある文章に対するコメントの読み手（リクエスタ）にとっての有用性の観点から，コメントを分類することに焦点を当てている．

(2) タスク結果の品質向上を目的とした研究．タスクの委託前にタスク設計を改善したり，より良いタスク結果をもたらすワーカを抽出する手法 [2], [4] や，タスク完了後に適切でない回答をフィルタリングする手法 [1], [7] に関する研究が存在する．これらの手法は，我々の提案する手法を用いて抽出する viewpoint を用いることによって，さらにタスク結果の品質を向上させることができると考えられる．

3. Viewpoint へのアノテーション

本節では，論文 [12] で構築したコーパス作成について説明する．このコーパスは，後ほど他のマイクロタスクの viewpoint に対する自動的なラベリングのために用いられる．

3.1 Viewpoint の収集

マイクロタスクの作成．まず，viewpoint を収集するための，元となるマイクロタスクの生成を行う．データには，日本の電子商取引サイトである ASKUL(注1) で販売されている商品の画像データを用いる．ASKUL では，多くの商品が階層構造を持つカテゴリに分類されている．元となるマイクロタスクは，この各データが，特定のカテゴリに属しているかを問うタスクである．ここで取り扱う特定のカテゴリの選出にあたって，以下の要件がある．

- マイクロタスクの数が膨大になることを防ぐため，特定のカテゴリ内の商品数も膨大でない
- 多様な viewpoint 収集のため，特定のカテゴリ内の商品も多様である
- 質問文を簡潔にするため，特定のカテゴリの名前は簡潔である

したがって，以下の条件を満たすカテゴリを「食品 / 飲料」カテゴリから選出した．

(1) 食品カテゴリから 2 レベル以上下位に位置する．

(2) 少なくとも 3 つのサブカテゴリを含む．

(3) カテゴリ名が他のカテゴリ名を 2 つ以上含むものでない（例 :「コーヒー / お茶」）

選出されたカテゴリは 7 つであり，このカテゴリから 28 個の商品画像データが得られた．図 1 に，これらのカテゴリを問う質問文を示す．この 7 つの質問を，28 個の商品画像データに対して行うこととし，元となるマイクロタスクを $7 \times 28 = 196$ 個作成した．

次に，元となるマイクロタスクに，ワーカがその回答を行った理由を問う質問文とその入力フォームを追加する．これらを追加した結果出来上がったマイクロタスクの一例を図 2 に示す．ワーカは「これはお茶ですか？」といった 1 つ目の質問に対して「はい」「いいえ」「わからない」のどれかで回答し，なぜその回答を選んだのかの理由を 2 つ目の質問で答える．このタスク結果によって，質問に対する回答とその回答を選んだ理由の組，すなわち viewpoint を得ることができる．

(注1) : http://www.askul.co.jp/

これはコーヒーですか？
これは炭酸飲料ですか？
これは緑茶ですか？
これは紅茶ですか？
これはお茶ですか？
これは調味料ですか？
これはインスタント食品ですか？

図 1　コーパス作成に用いる元の質問

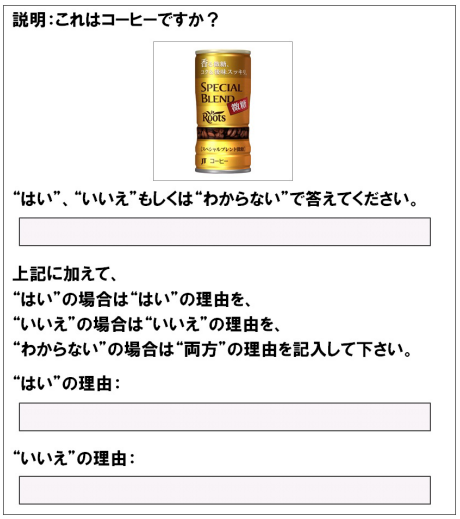

図 2　viewpoint 収集用タスク

マイクロタスクのクラウドソーシング. Yahoo!クラウドソーシングを通して，作成されたマイクロタスク 1 つにつき，20 人のワーカに問い合わせ，20 個の結果を得る．ワーカの必要人数の削減のため，重複のない 5 つのマイクロタスクを 1 セットとし，それぞれのワーカに 1 セット単位でタスクの委託を行った．このセットを作成するためにダミーのマイクロタスクを 4 個作成し，マイクロタスク数が 5 で割り切れる数（$196 + 4 = 200$）に設定した．また，ワーカに対する 1 セットあたりの報酬は 2 円に設定し，ワーカは多くとも 1 セットのタスクしか処理できないよう設定した．すなわち，得られる最大の viewpoint の数は $200 \times 20 = 4000$ であるが，理由を問う 2 つ目の質問に対する回答が任意であるため，結果として得られた viewpoint の数は 1413 であった．また，このタスクを処理したワーカの人数は，387 人であった．

3.2　本論文で用いる表記法

本論文では，viewpoint の収集を行うタスク結果の表記法として，次を用いる．W をワーカの集合，Q を質問文の集合，$A = \{$ はい, いいえ $\}$ を取り得る判断の集合，$C = A \cup \{\text{Unsure}\}$ を取り得る選択肢の集合とする．このとき，タスク結果の各インスタンスを (w, q, d, a, r) という組で表す．ここで，r は，ワーカ $w \in W$ が提示された画像データ d を対象とする質問文 $q \in Q$ に対して判断 $a \in A$ を行った理由である．ワーカが q に対して「わからない」と回答した場合，そのワーカには「はい」「いいえ」両方の判断の理由を入力してもらう．その場合，我々は 2 つのインスタンス $(w, q, d, \text{はい}, r_1), (w, q, d, \text{いいえ}, r_2)$ を手に入れる．

また，viewpoint は理由 r と結論 c からなると考えられ，これを論理式 $r \to c$ で置き換える．例えば，「これはカップに入った黒い液体．したがって，これはコーヒーである.」という viewpoint があったときに，r は「これはカップに入った黒い液体」，c は「これはコーヒーである」を表す．

ここで，viewpoint $r \to c$ はタスク結果のインスタンス (w, q, d, a, r) から入手することができる．なぜなら，もし a が「はい」（もしくは「いいえ」）であれば，c は肯定形（あるいは否定形）の記述に定まるからである．例えば，先ほどコーヒーの例における c は,「これはコーヒーですか？」という質問 q に対する「はい」という判断 a によって,「これはコーヒーである」という結論 c が導き出せる．したがって本論文では，必要であれば $r \to (q, a)$ という記法で viewpoint を表すこととする．

3.3　viewpoint の型

論理構造の観点から，viewpoint の型を定義する．viewpoint を表す論理式 $r \to c$ において，判断理由 r あるいは c が肯定的な意味を持つ場合それを “P”，否定的な意味を持つ場合はそれを “N”で表すこととする．まず，N と P の組み合わせによって，次の単純な 4 つの型を定義する．

PP $(r \to c)$: （例）ラベルにコーヒーと書いてあるので，これはコーヒーである．

NP $(\neg r \to c)$: （例）コーヒーでない証拠がないので，これはコーヒーである．

PN $(r \to \neg c)$: （例）これはカレーである．したがって，コーヒーではない．

NN $(\neg r \to \neg c)$: （例）コーヒー豆から作られていないので，コーヒーではない．

これらに加えて，次の 2 つの複合型を定義する．ここで複合型とは，判断理由に肯定形と否定形の両方が含まれているものであり，判断理由 $r_1, \neg r_2, \ldots$ 間の論理演算子（「かつ」「または」など）は区別しない．

PNP $(r_1, \neg r_2 \to c)$: （例）ラベルにコーヒーと書いてあり，かつコーヒーでないという証拠がないため，コーヒーである．

PNN $(r_1, \neg r_2 \to \neg c)$: （例）液体の色が黒でなく，透明であるので，コーヒーでない．

また,「コーヒーだからコーヒーである」というような viewpoint をトートロジーと呼ぶ．

定義 1 (トートロジー) *$PP\ (r \to c)$ or $NN\ (\neg r \to \neg c)$ のうち，$r = c$ となるものを, トートロジーと呼ぶ.*

トートロジーかそうでないかを区別するために，トートロジーとなっている viewpoint を表す 2 つの型を定義する．

P $(c \to c)$: （例）コーヒーであるから，コーヒーである．

N $(\neg c \to \neg c)$: （例）これはコーヒーでない．したがって，コーヒーではない．

トートロジーにおいては，前件 r が実用的な内容を提供しないことから，前件 r を表す “P”, “N”を取り除くことによってトートロジーを表現する．また，PNN において，その一部分を構成する NN がトートロジーとなっている場合，PN に分類する．

3.4　アノテーションガイドライン

viewpoint の収集を行った後，各 viewpoint に対して，型を

表すラベルを手動で付与する．ここで，viewpoint $r \rightarrow (q, a)$ は，タスク結果のインスタンス (w, q, d, a, r) から得られる．a が「はい」であったときには，c が肯定形（P）となるため，PP, NP, PNP のどれかを型を表すラベルとして付与する．また，a が「いいえ」であったときには，c が否定形（N）となるため，PN, NN, PNN のどれかを型を表すラベルとして付与することとなる．ただし，例外として次のようなケースには，“wrong” ラベルを付与する．

- 判断 a が「はい」（あるいは「いいえ」）であるのにも関わらず,「いいえ」（あるいは「はい」）と回答した理由を入力するためのフォームにその回答理由を記述している．
- フォームに記述された内容が，理由を表す文章となっていない（「これはカレーですか？」,「あああああ」など）
- フォームに記述された内容が，文脈に沿ったものでない（質問「これはお茶ですか？」に対して「いいえ」と回答しているにも関わらず，入力された理由が「これは茶葉を使用して作られたものなので，お茶です」というように,「はい」の回答理由を記述している）

表 1 に，アノテーション例を示す．各 viewpoint について，アノテーションの根拠を次から示す．

PP1. 「粉末緑茶という単語がパッケージに書かれている」が判断理由 r であり,「お茶である」が結論 c であるため，PP となる．

PP2. 「サイダーと記載してある」が判断理由 r であり，判断 a が「はい」であるため，結論 c が「炭酸飲料である」となり，PP となる．ここで，結論 c は判断 a によって決まるため，理由を入力するためのフォームに記述された文章には，結論 c が書かれていなくともよい．

NP1. 判断理由 r が「茶葉をポットに入れて作る，といった本格的な作り方の食品ではない」という否定形であり，判断 a が「はい」であるため，NP となる．

PN1. 一見すると NN に見える．しかし，文章中の「コーヒーでない」は判断理由 r を述べたものではなく，結論 c を述べたものである．判断理由 r を述べている部分は「（これは）ドレッシング」であるので，PN となる．このようなパターンの viewpoint は非常に多かった．

PN2. 判断 a が「いいえ」であり，理由を入力するためのフォームに記述された文章は「これはコーヒーである．したがって，お茶ではない」と解釈できるため，PN である．

NN1. 判断理由 r が「ミルクティーに緑茶の成分は入っていない」で，否定形であり，判断 a が「いいえ」である．したがって，NN となる．

NN2. これは複雑なケースである．理由を入力するためのフォームに記述された文章は「調味料は，料理に味をつけるもの」であり，否定形の表現は含まれない．しかし，以下の理由で NN とする．まず，判断 a が「いいえ」であるので，結論 c が「調味料でない」となり，候補は NN か PN，PNN に絞られる．つぎに，入力された文章「調味料は，料理に味をつけるもの」は「調味料であれば，料理に味をつけるものである」ということを意味する．これの対偶をとると「料理に味をつけるものでないならば，調味料ではない」という文章になる．判断 a から導かれた結論 c が「調味料ではない」であるので，この文章が厳密に viewpoint を表す文章となっていると考えられる．したがって，判断理由 r は「料理に味をつけるものでない」，結論 c が「調味料ではない」であるので，NN となる．一般に，その viewpoint に関して (1) 質問が「これは X ですか？」である．(2) 記述された文章に，X であるための必要条件が含まれる．(3) 判断 a が「いいえ」である．の 3 つを全て満たすとき，その viewpoint を NN と結論づける．

PNN1. 記述された文章には,「これはハーブ茶であり」「コーヒーは原料に含まれない」という 2 つの判断理由 $r_1, \neg r_2$ と,「コーヒーではない」という結論 c が含まれる．r_1 が肯定形，$\neg r_2$ が否定形であるので，PNN とする．このように，判断理由に肯定形と否定形の両方を含み，結論が否定形であるものが PNN の典型的なパターンであった．

PNN2. 記述された文章は,「インスタント食品とは手軽に簡単に利用できるもの」「これは豆を煎ってから使用するもの」という理由を表す部分と,「いいえ，になりうる.」という結論 c を表す部分からなる．理由を表す部分の 1 つ目に関しては，NN2 で示したパターンと同じで，(1) 質問が「これはインスタント食品ですか？」である．(2) 文章内に「インスタント食品とは手軽に利用できるもの」というインスタント食品であるための必要条件が含まれている．(3) 判断 a が「いいえ」である．という 3 つ条件が成り立つため，NN である．また，理由を表す部分の 2 つ目については，判断理由 r が肯定形，判断 a が否定形である典型的な PN である．したがって，全体としてみると PNN となる．ここで説明したように，文章内に 2 つ以上の判断理由を含む場合，まずそれぞれの理由に関してどの型に当てはまるかを考えた上で，それらを結合することによって,（複合型の）viewpoint の型を決定する．

4. インターアノテータアグリーメントとコーパスの統計

本節では，インターアノテータアグリーメントと，アノテーションしたコーパスの統計について報告する．アノテーションガイドラインに従い，収集した viewpoint に対して 2 人がアノテーションを行った．その結果，2 人のアノテータ間で異なるアノテーションがされたのはわずか 28 個であり，アグリーメントを表す一般的な尺度である κ 統計量は，0.968 という非常に高い値 [3] となり，アノテーションガイドラインが明確であることが示された．

また，表 2 に，各 viewpoint の型の数と割合を示す．W は wrong の略記である．この統計から，PN の数が多い一方，PNP,NP がコーパス内に 1 つしか存在しないことがわかる．PNP，NP が少ないのは，次のような理由によるものと考えられる．NP である viewpoint $\neg r \rightarrow c$ の対偶をとると，$\neg c \rightarrow r$ となる．ここで，多くのケースでは，c を満たさないものは非常に多く存在する．そして，判断理由 r となりうるような，それらに共通する属性を表すシンプルなフレーズを提供することは難しいと考えられる．例えば，結論 c が「コーヒーである」

型	ID	質問文	判断	理由を入力するためのフォームに記述された文章
PP	PP1	これはお茶ですか？	はい	粉末緑茶という単語がパッケージに書かれているので，お茶である．
	PP2	これは炭酸飲料ですか？	はい	サイダーと記載してあるため．
NP	NP1	これはインスタント食品ですか？	はい	茶葉をポットに入れて作る，といった本格的な作り方の食品ではないので，はい，となりうる．
PN	PN1	これはコーヒーですか？	いいえ	ドレッシングはコーヒーではないので
	PN2	これはお茶ですか？	いいえ	コーヒー
NN	NN1	これは緑茶ですか？	いいえ	ミルクティーに緑茶の成分は入っていないから
	NN2	これは調味料ですか？	いいえ	調味料は，料理に味をつけるもの．
PNN	PNN1	これはコーヒーですか？	いいえ	これはハーブ茶であり，コーヒーは原料に含まれないので，コーヒーではない．
	PNN2	これはインスタント食品ですか？	いいえ	インスタント食品とは手軽に簡単に利用できるものだが，これは豆を煎ってから使用するものなので，いいえ，になりうる．

表 1　アノテーションの例

PP	(P)NP	PN	NN	PNN	P	N	W	Total
247	2	824	110	95	23	4	108	1,413
17.5%	0.1%	58.3%	7.8%	6.7%	1.6%	0.3%	7.6%	100%

表 2　各 viewpoint の型の数と割合

であったとき，コーヒーでないものの共通の属性を言い表すことは簡単ではない．したがって，NP の PNP においては否定形の判断理由が作りにくく，その結果数が少なかったのだと考えられる．

5.　主観評価タスク

論理構造に着目したアノテーションが，汎用性の高い viewpoint の抽出に役立つことを示すため，各 viewpoint が様々なデータに対する判断に役立つかどうかをクラウドソーシングによって検証する．本論文では，これを主観評価タスクと呼ぶ．

5.1　実 験 設 定

タスク． 3.1 節で述べたように，viewpoint $r \rightarrow (q, a)$ を得るために，ワーカにある画像データ d を対象とした質問 q に対して判断 a を下した理由 r を尋ねた．主観評価タスクは，このようにして得られた各 viewpoint が，他の画像データ $d'(\neq d)$ を対象とした質問 q に対して判断をするときに役立つかを尋ねるものである．主観評価タスクを作成するため，収集した viewpoint と，質問に対する正しい解答を含むデータ（ゴールドスタンダードデータ）を用いる．具体的には，タスクは以下のようにして作成する．

ステップ 1. (w, q, d, a, r) から得られた viewpoint$r \rightarrow (q, a)$ それぞれに対して，正しい判断の解答が a であるデータ $d'(\neq d)$ をランダムに選ぶ．これは，質問 q に対する判断を，d と d' で同じにするためである．ここで，型が wrong である viewpoint に対しては，タスクを生成しない．その理由は，wrong 型の viewpoint は意味をなさないため，汎用性が低いことは明らかだからである．

ステップ 2. ステップ 1 によって選ばれた d' それぞれに対して，図 3 に示すような主観評価タスクを生成する．主観評価タスクには 2 つの質問が含まれる．1 つ目の質問は，そのviewpoint を収集したときに行った質問 q と同じものである．この質問は，提示する d' に対して，ワーカが正しい判断 a ができるかを確認するために行う．2 つ目の質問は，d' を対象とする質問 q に回答するときに，提示する viewpoint が役立つかどうか問うものである．

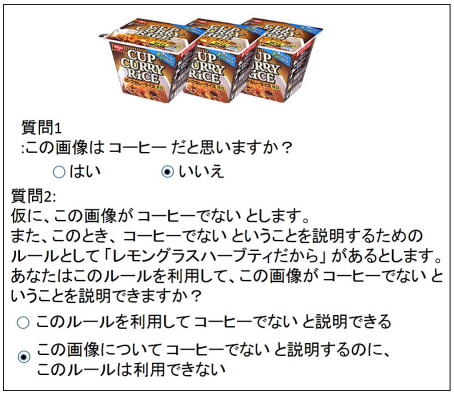

図 3　主観評価タスク

このタスクについて複数人に問い合わせたときに，2 つ目の質問に対して「役立つ」と答えた人が多ければ，その viewpoint はより汎用性が高いものであるといえる．言い換えれば，「役立たない」と答えた人が多ければ，その viewpoint は特定のデータ d にしか当てはまらない，汎用性の低い viewpoint ということになる．

ワーカ． 多数決をとるため，生成した各主観評価タスクに対して，重複のない 9 人のワーカに問い合わせた．結果，主観評価タスクを行ったワーカの人数は，実質 1058 人であった．

5.2　実 験 結 果

表 3 に，主観評価タスクの結果を示す．各行は，viewpoint の型ごとに結果を集計したものとなっており，各行 2 列目は提示する d' に対して，少なくとも 1 人のワーカが正しい判断 a をした viewpoint の数である．また 3 列目と 4 列目は，2 列目で示した人数のうち，過半数が d' に対して「役立つ」と回答された viewpoint について，それぞれ数と割合で示している．

結果から，他のデータに対して「役立つ」と答えた人の割合が，各型で明らかに異なることがわかる．重要なのは以下の点である．

type	1 人以上が正しい判断をした数	過半数が「役立つ」と判断した数	割合
PP	247	96	38.9%
NP	1	0	0.0%
PNP	1	1	100.0%
PN	824	115	14.0%
NN	110	48	43.6%
PNN	95	22	23.1%
P	23	22	95.7%
N	4	3	75.0%

表 3　クラウドが「役立つ」と判断した viewpoint の数と割合

- P と N は，他のデータに「役立つ」とされる viewpoint の割合が明らかに高い．これは，P と N が「インスタント食品であるから，インスタント食品である」というようなトートロジーであり (節 3.3)，どのデータに対しても当てはまるものであったからであると考えられる．
- PN は，他のデータに「役立つ」とされる viewpoint の割合が，PP と NN に比べて低い．これは，PN は「これはお茶であるので，炭酸飲料ではない」など特定のデータ（ここでは，「お茶」）に対する判断にしか役立たないものが多かったからであると考えられる．言い換えれば，PN の viewpoint $r \to c$ において，r が表現するデータの集合が小さいということであり，多くのケースで d' がその集合の中に含まれなかったということである．一方で，PP や NN における r は,「これは食べ物ではないから（お茶ではない）」など，特定のデータだけに依拠したものではなく，比較的汎用性が高いものが多いことが示されている．
- PNN については，若干 PN よりも，他のデータに対して「役立つ」と回答されている割合が高い．これは，PNN の viewpoint は「これはお茶であり，食品ではない．したがって，インスタント食品ではない.」というような，汎用性の低い PN と比較的汎用性の高い NN の複合型であり，何人かのワーカが汎用性の低い PN の部分（ここでは,「これはお茶であり」）を無視したからであると考えられる．

「役立つ」と回答される割合は高ければ，r と c が表す集合が似ているということがいえる．したがって，実験結果で割合が高かった PP，NN が c の特徴をうまく表現していると考えられ，それぞれのワーカの c に対する解釈を表しているといえる．ここで，P と N の viewpoint はトートロジーであるため，割合が高くとも提案手法には役立たないことに注意したい．

これらから，論理構造の観点から提案したアノテーションガイドラインが，収集した多くの viewpoint の中から，有益なものを抽出する際の重要な手がかりとなることを示した．

6. 分類実験

viewpoint の数が増えるにつれ，それを分析するリクエスタの負担も増えてしまう．したがって，分類器による viewpoint のフィルタリングがリクエスタにとって必要となってくる．

viewpoint の型を推定するため，単純な線形分類器を用いる．ライブラリとして libliner [5] を用い，L2-regularized logstic regression を指定したうえで，その他のパラメータはデフォルトのものを使用した．文書分類と同じように，viewpoint i を入力ベクトル $x_i \in R^m$ で表現し，出力ラベル y_i が viewpoint の型を表すとする．入力ベクトルは，以下の 4 ステップで作成する．(1) まず，各 viewpoint $r \to (q, a)$ が生成される元となった質問文 q を分析し，質問対象物を特定する．例えば,「これはインスタント食品ですか？」という質問では,「インスタント食品」が質問対象物となる．その後，viewpoint の判断理由 r のもととなる，理由を入力するフォームに記述された文字列の中から，質問対象物を特別なシンボル T で置き換える．(2)(1) で生成した文字列の中から，名詞を特定し，すべて特別なシンボル O で置き換える．(3)(2) で各 viewpoint に対して生成された文字列の終端に特殊なシンボルを付与した後，それらを鎖状につなぐ．このようにしてできた一つの文字列の中から，極大部分文字列 [6], [8] を抽出する．ここで，極大部分文字列は，本質的に分類器の重み学習に必要な部分文字列そのものである [8]．極大部分文字列についてのより詳しい説明は, [8] を参照されたい．この極大部分文字列を，入力ベクトルの次元として用いる．ここで入力ベクトル x_i の j 次元目の値 n は，対応する極大部分文字列がその viewpoint i 内に出現した回数であるとする．(4) さらに，入力ベクトルの次元に値がバイナリとなる 5 つの次元を追加する．そのうち 3 つの次元は,「はい」「いいえ」「わからない」がそれぞれ選択されたかどうかを表すものであり，残り 2 つの次元は「はい」の理由を入力するフォーム,「いいえ」の理由を入力するフォームに入力したかどうかを表すものである．この次元の追加は，wrong をうまく検出するために行う．

訓練データとして，次の飲料に関する質問文から得られた viewpoint の集合を用いた．

（1） これはコーヒーですか？

（2） これは炭酸飲料ですか？

（3） これはお茶ですか？

（4） これは紅茶ですか？

（5） これは緑茶ですか？

検証データとして最後の質問文である「これは緑茶ですか？」から得られた viewpoint の集合を用いた．また，テストデータとして，以下の 2 つの質問文から得られた viewpoint の集合を用いた．

（1） これは調味料ですか？

（2） これはインスタント食品ですか？

このように，訓練データとテストデータで使用されている語彙が異なるように意図的に設定した．その理由は，学習された分類器が特定のドメインに依存していないことを示すためである．ここで，5. 節で述べたように，有益であるのは PP,NN であり，それ以外の型の viewpoint は有益でないとしている．分類器の測定方法を以下に示す．

評価に用いる尺度は，精度 (正しくラベル付けされた viewpoint の割合)，適合率 P（分類器が「有益である（型が PP あるいは NN である）」と判断した viewpoint のうち，実際に有

益であった viewpoint の割合），再現率 R(有益な viewpoint のうち，分類器が「有益である」と判断した viewpoint の割合)，そして P と R の調和をとる F 値 F_1（$F_1 = 2PR/(P+R)$)）を用いる．テストの結果，精度は 88.3% であった，また，F 値について，最も良かった結果は適合率 $P = 72.6\%$，再現率 $R = 72.6\%$ が 72.6% となったときの $F_1 = 72.6\%$ であった．この結果は，現在のコーパスの大きさでは自動的なアノテーションは難しいことを示している．インターアノテータアグリーメントの値が高いことから，さらなる改善の余地があると考えられる．

7. おわりに

タスク内の曖昧な記述を減らそうとしているマイクロタスク設計者を助けるために，タスクの結果に影響を及ぼすワーカの viewpoint について研究した．まず，質問への解答とその理由といった観点から viewpoint を定義した．そして viewpoint に対するアノテーションガイドラインを考案し，クラウドソーシングを用いたユーザ調査の結果から，PP と NN と呼ばれる 2 つの viewpoint がとりわけ有益であることを決めた．さらに，他の有益でない viewpoint から PP,NN を区別できる分類器を作成した．

今後の課題として挙げられるのは，本手法によって実際にリクエスタに記述を書き換えてもらい，曖昧性が下がることを示すことである．また，自然言語処理の標準に従うと，収集した viewpoint の数が十分とはいえないため，データを拡大することを計画している．

謝　辞

本研究の一部は科研費 (#25240012) による．

文　献

[1] Srikanth Jagabathula andLakshminarayanan Subramanian andAshwin Venkataraman. Reputation-based worker filtering in crowdsourcing. In *Advances in Neural Information Processing Systems 27:Annual Conferenceon Neural Information Processing Systems 2014, December 8-13 2014,Montreal, Quebec, Canada*, pp. 2492–2500, 2014.

[2] Ailbhe Finnerty andPavel Kucherbaev andStefano Tranquillini andGregorio Convertino. Keep it simple: reward and task design in crowdsourcing. In *Biannual conference of the Italian chapter of SIGCHI, CHItaly '13,Trento, Italy - September 16 - 20, 2013*, pp. 14:1–14:4, 2013.

[3] Ron Artstein and Massimo Poesio. Inter-coder agreement for computational linguistics. *Comput. Linguist.*, Vol. 34, No. 4, pp. 555–596, dec 2008.

[4] Djellel Eddine Difallah, Gianluca Demartini, and Philippe Cudré-Mauroux. Pick-a-crowd: tell me what you like, and i'll tell youwhat to do. In *22nd International World Wide Web Conference, WWW'13, Rio de Janeiro, Brazil, May 13-17, 2013*, pp. 367–374, 2013.

[5] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. Liblinear: A library for large linear classification. *J. Mach. Learn. Res.*, Vol. 9, pp. 1871–1874, jun 2008.

[6] Matthias Gallé. The bag-of-repeats representation of documents. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, pp. 1053–1056, New York, NY, USA, 2013. ACM.

[7] Nguyen Quoc Viet Hung, Nguyen Thanh Tam, Ngoc Tran Lam, and Karl Aberer. An evaluation of aggregation techniques in crowdsourcing. In *Web Information Systems Engineering - WISE 2013 -14th International Conference, Nanjing, China, October 13-15, 2013,Proceedings, Part II*, pp. 1–15, 2013.

[8] D. Okanohara and J. Tsujii. Text categorization with all substring features. In *SIAM International Conference on Data Mining (SDM)*, pp. 838–846, 2009.

[9] Galina Tremper. Weakly supervised learning of presupposition relations between verbs. In *Proceedings of the ACL 2010 Student Research Workshop*, ACLstudent '10, pp. 97–102, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[10] Hila Weisman, Jonathan Berant, Idan Szpektor, and Ido Dagan. Learning verb inference rules from linguistically-motivated evidence. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pp. 194–204, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[11] Katja Wiemer-Hastings and Peter Wiemer-Hastings. Dp: A detector for presuppositions in survey questions. In *Proceedings of the Sixth Conference on Applied Natural Language Processing*, ANLC '00, pp. 90–96, Stroudsburg, PA, USA, 2000. Association for Computational Linguistics.

[12] 丹治寛佳, 清水伸幸, 森嶋厚行, 北川博之. マイクロタスク型クラウドソーシングにおける質問文改善の支援手法. In *DEIM2015 C6-1*, 2015.