クラウドソーシングにおける成果物の品質維持のための ダミー問題出題手法の検討

清水 伸幸，山下 達雄，塚本 浩司，颯々野 学(ヤフー株式会社)
{nobushim, tayamash, kotsukam, msassano}@yahoo-corp.jp

１．まえがき

近年、インターネットを通じて、不特定多数の人々により一定のタスクを成し遂げるクラウドソーシングと呼ばれるサービスの利用が急速に拡大している。計算機科学分野においても、人間の知的活動を計算機で補佐するため、特に米国においては画像の注釈付けや Web コンテンツの分類などの教師データの収集や、自動要約、機械翻訳など自然言語処理システムの評価などへのクラウドソーシングの利用が始まっている。クラウドソーシングの中でも、特にマイクロタスクと呼ばれる細分化されたタスクを行うタイプのサービスは、Yahoo!クラウドソーシングをはじめとするサービスプラットフォームによって、日本でも手軽に利用できるようになり、多くの人間の知見を集約することが可能となった。

こうして、新しいツールとして認識されつつあるクラウドソーシングであるが、品質の高い成果物を得るためには従来のように専門家にタスクの依頼を行うケースとは異なるタスクの設定手法が必要であることがわかってきた。従来と異なる点の一つは、タスクの意図を正確に伝えるため、タスクの表現に注意を払う必要がある点である。例えば、固有名詞抽出の技術を評価するため、抽出したある名詞 X について、クラウドソーシングで「X は飲食店の名前ですか？」と、タスクを設定したとする。単純な質問のようだが、結果は少なくとも以下の３つの解釈が混じったものとなる。（1）店としての存在の有無にかかわらず、文字列として飲食店らしいかどうかで判断すればよい、という解釈。（2）検索などで、店として現実に開店していることを確認して答えを出す、という解釈。（3）よくある店の知名度調査だと解釈し、店の名前を以前に聞いたことがあるかないかで返答すれば良い、という解釈。このように、一見単純そうな質問でも、タスク依頼者の意図とタスクを行うユーザーの解釈が分かれる事態は頻繁に起こる。従来と異なる二つ目の点は、手軽に報酬を得ることを目的として低品質の成果物を納める不誠実なユーザーや、設問の説明文を理解していないとみられるユーザーが一部存在していることである。成果物の品質を管理するにあたっては、上記の解釈の問題と、不誠実、あるいは設問文に対する理解の低いユーザーの存在の問題とを区別することが非常に重要である。

そこで、本研究では、現実のタスクを Yahoo!クラウドソーシングに入稿し、 その過程で得られたクラウドソーシング利用の注意点に関する知見を報告する。タスクは「検索連動型広告において入札すべき、広告の内容に合致した適切な検索語を提案すること」とした。検索連動型広告とは、検索サイトにおいて、利用者の入力した検索語に基づいて検索結果のページに表示されるテキスト広告である。この形式の広告の特徴は、検索サイトの利用者が実際に入力した検索語に関連する広告が表示されるため、他のウェブ広告と比べても利用者の興味に合致した広告を提示できることである。検索連動型広告では、広告出稿者は任意の検索語を指定してオークション形式で入札し、入札の結果、検索語が一致し、かつ指定した料金の範囲である、などの条件を満たせば広告が表示される。この広告方法の課題としては、広告に適した検索語が何であるか常に自明であるとは言えない点である。例えば、屋形船の船主が「花火」という検索語を入札し、屋形船の売り上げが大きく伸びたケースが存在する等、広告の内容と、それに有効な検索語とが表層的に異なる内容であることも多い。タスク選択の理由は、このタスクで得られるリソースが広範囲に様々な広告主に必要とされ、ウェブ広告の効果改善に重要であり、かつ、自然言語処理での単語間の関係抽出という重要なタスクと共通する性質を持っているためである。

本研究ではこういった広告と、それに有効な検索語のペアを入手するため、統計的手法で候補となる検索語を生成し、クラウドソーシングを利用して広告に対して適切な検索語であるかどうかを判定する実験を行う。次節ではクラウドソーシングでの成果物の品質を管理するため、一般的な手法を解説する。実験ではこれらの手法を活用し、それぞれの利点を調査する。３節において、実験結果を説明した後、４節において、クラウドソーシング利用のベストプラクティスにつながる注意点をまとめる。

２．クラウドソーシングにおける成果物の品質管理手法

既に述べたように、クラウドソーシングにおける課題の

一つは成果物の品質の管理である。Amazon Mechanical Turk など､商用プラットフォームの一部ではタスク依頼者が成果物を確認した上で､低品質な成果物には報酬の支払いを行わないオプションが用意されているものもある。しかしながらこの手法は､そもそも大量の作業結果全てに対してエキスパートによるチェックを行うことが前提であり､クラウドソーシングの利点を損なうため､現実的な解とはいえない。

一般的にクラウドソーシングでの品質管理に用いられる手法には２つのアプローチが存在する。一つは冗長化である。これは､同じタスクを複数のユーザーに依頼し多数決を取るなど､成果物を統計的に統合することにより､質の高い成果物を獲得するやり方である (Dawid and Skene 1979, Whitehill et al 2009)。冗長化では品質問題は改善されるが､何らかの理由で多数派となるユーザーがタスク依頼者の意図と異なる行動をとった際には機能しないため､頑健ではない。もう一つのアプローチは、正解のわかっているタスクを依頼することで､ユーザーの真面目さやタスクの理解度を測る方法である。こちらのアプローチでは、エキスパートが事前に作成した品質の高い正解付きのデータ(Gold Standard Data)が必要となるため、コストが増加する。特に、依頼者が必要としている本来のタスクと見分けがつかないダミー問題と呼ばれる問題を設定する手法では､依頼者側での準備が必要となり大きなコストがかかってしまう。

本稿では、これらの品質管理手法に加え、汎用的に用いることが可能な人間と機械を判別するチューリングテストの効用を調査する。このテストは一般に CAPTCHA と呼ばれ、近年問題になっているボット(自動プログラム)を用いたサービスの不正利用に対抗するために考案されたものである。CAPTCHA の基本的な形式は、歪曲やノイズが付加された文字列画像を WEB ページに提示し、閲覧者がその文字を判読できるか否かを試すものであるが､ここではより一般なチューリングテスト問題を利用することとする。

チューリングテスト問題は､上記のタスクと見分けの付かない Gold Standard Data を用いたダミー問題と比較すると、汎用性があり全てのタスクで利用が可能である一方、本来のタスクとの違いが一瞥できてしまうため､チューリングテスト問題だけをきちんと解く不誠実なユーザーを判別できない可能性がある。

３．実験

実際の広告は著作権上の問題で実験での利用が困難であるため､広告の代わりに以下の様な擬似的な特集記事の見出しを自動生成し､特集記事の見出しが与えられた検索語に対して適切であるかどうかを判断するタスクを設定した。

> 【花火】という単語でヤフー検索をしたユーザーに、
> 下記のリンクを表示して、興味を持つと思いますか？
> 「Yahoo!【屋形船】特集！」

実際のタスクでは、【】内の単語は自動的に抽出した単語ペアが入ることになる。ユーザーには、はい・いいえの二択が選択肢として与えられ､同時にコメント欄を設けて自由記述でのフィードバックも可能とした。１タスクは５問のセットとし､そのうちの１問はチューリングテスト問題、もう１問は Gold Standard Data を用いたダミー問題とした。各ユーザーに Gold Standard Data が重複して現れないよう、一人最大でタスクを４回まで実施可能とし、Yahoo!クラウドソーシングのタスクとして設定した。

チューリングテスト問題としては、“あなたは人間ですか？”または、“あなたはロボットですか？”という質問を使用した。一方の Gold Standard Data を用いたダミー問題では、以下の単語ペアを利用し、そのときのユーザーの正解率は以下の様になった。

【毛沢東】【メリーズお尻拭き】	96%
【箸置き】【通学路】	95%
【ソビエト連邦】【葛根湯】	94%
【グルコサミン】【アラビア人】	94%
【有言実行】【おむつ】	94%
【ヘモグロビン】【違約金】	89%
【サンボ】【まかない】	84%

これらのダミー問題は、すべて、答えが「いいえ」と選択されるよう単語ペアを作っている。この理由は、ダミー問題以外の単語ペアは､コーパスから統計的に自動生成しているため、全体的に「はい」が正解であることが多く、注意力が減少したユーザーがすべての設問に同じ「はい」の答えを選択しないよう、定期的に「いいえ」の答えがあるダミー問題を出現させることが望ましいからである。

ここで、【サンボ】【まかない】のペアの正解率は８４%で、他のダミー問題より低い。回答者コメントを調査したところ、タスク依頼者としては格闘技の「サンボ」を想定していたが､「サンボ」という名前の牛丼屋も存在しており、お店を検索したのであれば、まかないにも興味があっておかしくないことが判明した。このように Gold Standard Data を用いたダミー問題であっても､クラウドソーシング実施の際には十分な注意が必要である。

実験の結果、３７７人のユーザーがタスクを行い、チューリングテスト問題に１９人が間違った回答をした。さらに残りのユーザーのうち、正解率の低い、【ヘモグロビン】

【違約金】と【サンボ】【まかない】の２ペアを除外したダミー問題に間違ったユーザーは３５人となり、この手法でフィルターした結果、８４．３%のユーザーが残ることとなった。タスクでは、１２０１個の自動的に抽出した単語ペアを、単語ペア毎に最大５人までのユーザーに提示しており、５人全員が検索語に対して適切な特集ページであると判断した単語ペアは１６５個獲得できた。以下がその例となる。

【マンション】【住まい】
【スポーツクラブ】【フィットネス】
【ギフト】【快気祝い】
【痩せる】【ダイエット】
【自動車学校】【免許】

同様に５人中４人が適切であると判断した単語ペアは３０９個あり、合わせて４７４個の正例となる単語ペアをクラウドソーシングの結果として得ることができた。

一方、５人が適切でないと判断した単語ペアは８９個存在した。以下に例を挙げる。

【多汗症】【豊胸】
【トランペット】【フルート】
【金運】【フィーリング】
【卵】【辺見えみり】
【犬】【猫　販売】

これらの負例は、本来、検索連動型広告のキーワードとしては必要なものではないが、将来のクラウドソーシング再利用のため、Gold Standard として、ダミー問題として利用できないか追加調査を行った。しかしながら、どのペアも人手で作ったダミー問題と比べ、関連性が高いと答えるユーザーが存在したため、負例のダミー問題には適さないことが判明した。適切な負例が抽出できなかった原因としては、元々のデータを統計的な関連度の高さに基づいて生成したため、利用したデータに関連度が十分に低い単語ペアがなかったことが考えられる。

本稿の例や、自動同義語抽出など、正例のみを優先的に生成する言語処理の手法は一般だが、このように生成されたデータをクラウドソーシングで単純なデータソースとして利用すると、十分な負例（同義語の例では、絶対に同義語でない単語ペア）が混ざらない、という点に注意されたい。

更に、ユーザーから得られたコメントを精査したところ、今回のタスク設定では、単語ペア間の関連性は無いとはいえないが、検索者がどのような人物であるかによって興味の度合いが変化する単語ペアが与えられた場合に、判断が揺れるケースがあることが判明した。例えば、

【過払い】【債務整理】
金融関連の働き手は、興味を持つと思う。

【肌着】【ベビー肌着】
赤ちゃんを持つ親なら興味はあると思います。

【ごみ屋敷】【大田区　粗大ごみ】
地域を特定しすぎ、検索者が大田区に住んでいれば。

【人探し】【浮気　妻】
人探しは条件が広く、浮気妻まではいかないと思う。

といったコメントが得られた。こういった検索者の属性が検索語と広告の適合度に与える影響については、タスク設定の際に考慮できていなかった。

ここから得られる知見は、自然言語処理でクラウドソーシングを利用するプロセスは言語処理におけるコーパスの作成(Hovy et al. 2006)に似ているということである。アノテーション・ガイドラインを決めてコーパスを作る過程では、当初から予想していない例外的な用例を見つけた時、ガイドラインを更新してコーパスを作り直し、アノテーションの揺れを最小化するといったプロセスが行われる。クラウドソーシングでも、例外的で判断がわかれるケースが逐次発見されるため、タスクの説明文を都度更新してインクリメンタルにデータを整備することが質の高い成果物を得る重要なポイントとなる。

４．結論

本研究での実験から、一度クラウドソーシングに出してフィードバックを得なければ、Gold Standard Data と当初思われていたものであっても、ダミー問題として適切か判断できないケースがあることがわかる。また、チューリングテスト問題は有効だが、それだけを真面目に回答する人がいる以上、設問との区別がつかないダミー問題は必要である。自然言語処理においてクラウドソーシングを利用する際、本稿の知見をベースに、ベストプラクティスが作られて行くことを願っている。以下に本文中で述べられなかったベストプラクティスへ向けた提言を記しておく。

1　実データの利用方法が想像できる質問の仕方をすることで、タスク製作者が思ってもみなかったユーザーの解釈や誤解が減らせる。
2　設問の説明には具体的な例を入れ、検索が必要なら検索へのリンクを貼る。
3　新しいタスクをクラウドソーシングに入稿する際に

は、ダミー問題も含めて小規模なデータでパイロットテストを行い、何度か試行錯誤をした上で大規模なデータを投入する。

4　自由記述のコメント欄をつけて、ユーザーが判断に迷う際にはどうして迷ったのかフィードバックを得て、次のイテレーションでのタスクの改善につなげる。

5　２つの質問を組み合わせて一つの答えを求める等、認知言語学的に処理が難しい聞き方をしない。例：「以下のブログの内容は飲食店についてのものですか？また、このお店に行ってみたいと思いますか？はい・いいえ」

謝　辞

弊社インターンシップ期間中、特許庁の石川雄太郎氏に本発表の調査の実施、データの解析にご協力いただきました。有難うございました。

参考文献

- Dawid, A. P. and Skene, A. M. (1979) “Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm”, Journal of the Royal Statistical Society. Series C (Applied Statics), Vol. 28, No. 1, pp. 20–28
- Hovy, E., Marcus, M., Palmer, M., Ramshaw, L. and Weischedel, R. (2006) "OntoNotes: The 90% Solution", Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers, pp. 57-60
- Whitehill, J., Ruvolo, P., Wu, T., Bergsma, J., and Movellan, J. (2009) “Whose Vote Should Count More: Optimal Integration of Labels from Labelers of Unknown Expertise”, in Advances in Neural Information Processing Systems 22

略歴

著者 1 氏名：清水　伸幸　（ヤフー株式会社 Yahoo! JAPAN 研究所 主任研究員）

2006 年ニューヨーク州立大学オルバニー校にて博士課程修了。2007 年より東京大学情報基盤センター特任助教。2010 年より同センター特任講師。2011 年より Yahoo! JAPAN 研究所勤務。クラウドソーシング企画、自然言語処理と機械学習の研究開発に従事。博士（情報工学）。

著者 2 氏名：山下　達雄 (ヤフー株式会社 Yahoo!JAPAN 研究所 上席研究員)

2000 年奈良先端科学技術大学院大学情報科学研究科単位取得退学。2000 年より株式会社富士通研究所研究員。2005 年よりヤフー株式会社勤務。自然言語処理技術の研究と Web サービスへの適用に従事。博士(工学)。

著者 3 氏名：塚本　浩司 (ヤフー株式会社 Yahoo!JAPAN 研究所 上席研究員)

1998 年東京大学工学系研究科修了。1998 年より株式会社富士通研究所研究員。2004 年より 1 年間スタンフォード大学言語情報研究センター客員研究員。2009 年より Yahoo!JAPAN 研究所勤務。オンライン広告技術に関するプロジェクトマネージャーや、自然言語処理、機械学習、情報検索、広告配信技術の研究開発等に従事。

著者 4 氏名：颯々野　学 (ヤフー株式会社 Yahoo!JAPAN 研究所 言語処理・機械学習チームリーダー)

1991 年京都大学工学部電気工学第二学科卒業。同年より富士通研究所研究員。1999 年より 1 年間、米国ジョンズ・ホプキンス大学客員研究員。2006 年よりヤフー株式会社勤務。自然言語処理の研究に従事。2008 年京都大学大学院情報学研究科知能情報学専攻博士後期課程修了。博士（情報学）。