クラウドソーシングによる慣用句判定

町田 雄一郎　　柴田 知秀　　黒橋 禎夫
京都大学大学院 情報学研究科
{machida,shibata,kuro}@nlp.ist.i.kyoto-u.ac.jp

1　はじめに

近年，自然言語処理では大規模データを利用した機械学習による手法が成果を上げており，充実した言語リソースの構築や，評価セット，機械学習のための正解データ構築などが重要なものとなっている．このようなデータの作成を人手で行うには時間を要し，多くの場合専門家が行うため費用がかかることが問題である．そこで近年，インターネットを介してワーカーと呼ばれる世界中の不特定多数の人間に作業を安価に依頼することのできるクラウドソーシングの利用が増加している．クラウドソーシングを適切に用いることで，不特定多数の人間の判定を用いても，十分な精度のデータを高速かつコストを抑えて獲得できることが多くの研究により示されている [3].

本研究では，クラウドソーシングが言語リソース構築にどのように貢献できるかについて，慣用句を対象として分析する．現在存在している慣用句リソースには佐藤 [8] が作成した基本慣用句五種対照表[*1]があるが，佐藤も指摘しているように，基本的な慣用句でもこのリストからもれてしまっているものもある．そこで慣用句候補を自動的に収集し，クラウドソーシングによって不特定多数のワーカーの判定から新規に慣用句を獲得するタスクを行った．

2　先行研究

Snow らは様々な自然言語処理のタスクをクラウドソーシングで行い，専門家によるものと比較したところ複数のワーカーの判定を集約すれば高い精度の結果が獲得できることを示した [3]．これを発端に多くの NLP タスクをクラウドソーシングで行うという研究がなされている．例えば，推論規則の評価 [6] や，クラウドソーシングを利用して高い質の翻訳を獲得する [5] など，多くのタスクが行われており，自然言語処理分野でもクラウドソーシングは近年着目されている分野である．

また，クラウドソーシングで得られる判定の集約方法を工夫することで結果の品質を上げる研究も行われている．個々の能力に差があるワーカーの判定結果を集約する最も基本的な手法としては多数決が考えられるが，真の答えを潜在変数として判定を統計的に行う手法も研究されている．Dawid らは複数の医師の診断から適切な診断結果を得るという場面でこのような問題を考え，それぞれの医師の能力をパラメータとして考えることで，複数の判定から適切な判定を推定する手法を提案した [1]．近年ではこのような手法をクラウドソーシングに応用した手法が研究されている．[4]

3　慣用句判定タスク

3.1　慣用句の定義

日本語慣用句辞典 [9] ではおおよそ一般的な共通理解として「単語の 2 つ以上の連結体であって，その結びつきが比較的固く，全体で決まった意味を持つ言葉」であるとしているが，同時に「慣用句の定義はいまだに決定的なものがない」とも断っており，慣用句と他の表現との明確な区別は難しく，慣用句かどうか判別するにはかなり主観的な要因があることは否定出来ないと述べている．また基本慣用句五種対照表を作成した佐藤も，コロケーションと慣用句を明確に区別することは難しく，慣用句を判定する際は他の句と厳密に区別するのではなく，広い意味で「慣用句等の成句」として捉えるほうがよいと述べている．したがって，広い意味で慣用句を考えた時，多数の日本語使用者が持つ実際の感覚に根ざした判定は重要な基準となりうると考えられる．

[*1] http://kotoba.nuee.nagoya-u.ac.jp/jc2/kanyo/doc

3.2 慣用句候補の選定

ワーカーに提示する慣用句候補は，格フレーム [7] 中で直前格に含まれる名詞が 1 つしかないものを自動的に抽出したものである．これは慣用句には特殊な言い回しが多く，直前格に含まれる名詞が 1 つしかない場合が多いと考えたからである．例えば「焼く」の 6 番目のフレームにおいて直前格であるヲ格には「世話」の一つしかない．このような語句を慣用句候補として考えると全体で 4,677 句があった．また，結果の比較のため人手で作成された慣用句リストからも 559 句の慣用句を入れた．これは佐藤らの作成した基本慣用句五種対照表から，橋本ら [2] が選定したより基本的な慣用句のうち，述語項を含むものである．（今後，このリストを「人手慣用句リスト」と呼ぶ．）したがって計 5,236 句の慣用句候補についてワーカーからの判定を収集した．

3.3 慣用句判定タスクの概要

本タスクでは自動抽出した慣用句候補に対してクラウドソーシングを用いて慣用句かどうか判定した．慣用句判定タスクの概要を図 1 に示す．今回はクラウドソーシングサービスとして，Yahoo!クラウドソーシング[*2]を利用した．

本論文では混乱を防ぐためにに次のように語句を定める．

ワーカー　インターネットを通して判定を行う作業者

問題　ワーカーに対して提示される質問

タスク　いくつかの問題で構成される作業の単位

このサービスでは，ワーカーは報酬を得るために決められた問題数で構成されたタスクを行わなければならない．今回は 1 タスクを 5 問に設定した．問題は以下の様な形式である．

例：「骨を折る」は慣用句ですか．
(はい/いいえ/わからない)

回答の選択肢は「はい」「いいえ」「わからない」の 3 値である．ここで注釈として，「骨を折る」のように慣用句の意味とそのままの意味の両方で使われる場合は「はい」を選択して下さい，という説明をつけた．ワーカーに対しては図 2 のような画面が表示される．今回は 1 つの慣用句候補に対して 10 人のワーカーが判定し，1 人のワーカーは最大で 10 タスク行うことができるように設定した．また 1 タスク中，つまり 5 問中 1 問はワーカーの信頼性を図るための問題を入れた．このようにワーカーの信頼性を図る手法をクオリティ・コントロールと呼ぶことにする．

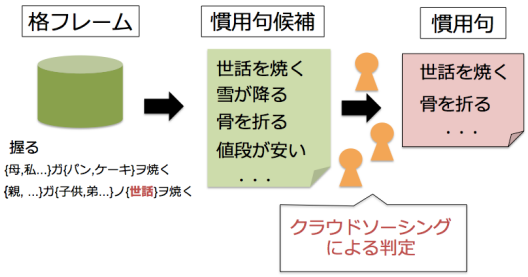

図 1　慣用句判定タスク概要

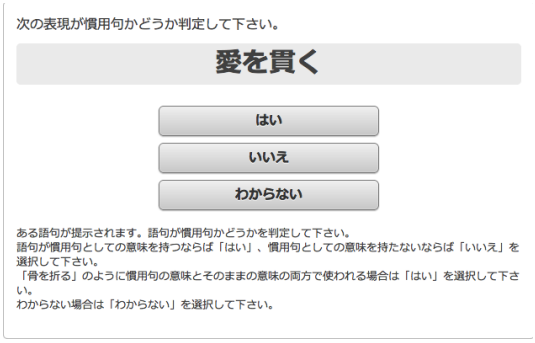

図 2　ワーカーに提示される画面

3.4 クオリティ・コントロール

クラウドソーシングでは不特定多数のワーカーに作業を依頼するため，中には全く真面目に作業をしないワーカーも存在する．このようなワーカーはフィルタリングする必要がある．今回はフィルタリングとして，事前に正解が分かっている問題を入れた．この問題は一般的な日本人ならば誰にでもわかるような問題であり，この問題の正解率を見ることでワーカーが真面目に仕事をしたかどうかを判定する．

例：「コンビニに行く」は慣用句ですか．
(正解として「いいえ」を期待している)

本論文ではこのような問題をチェック問題と呼ぶことにする．今回実施したクオリティ・コントロールはワーカーがどれだけ慣用句を知っているかというような言語能力を図るものではなく，あくまで問題文を正確に理解し，真面目に取り組んでいるかどうかを知るためのものである．

4 タスク結果

4.1 基本的な情報

今回の慣用句判定タスクで実行されたタスクの総数は 13,090 タスクで，チェック問題についた判定を除く

[*2] http://crowdsourcing.Yahoo.co.jp/

と慣用句候補についた判定は 52,360 個である．また，ワーカーの総数は 1,363 人であった．ワーカーが行ったタスク数の平均は 9.6 であり，ほとんどのワーカーが最大の 10 タスク行っているということがわかる．また，タスクは Yahoo!クラウドソーシングに投稿されてから約 1 時間で終了したので 1 分あたり 1,090 判定が獲得できたことになり，短時間で非常に多くの判定を獲得できた．

4.2 判定の集約処理

次に複数の判定を集約して最終的な判定結果を下す手法について考える．今回は，チェック問題の正答数によってワーカーを選別して多数決をとる方法と，EM 法によってワーカーの能力と問題の難易度の両者を考慮して判定を集約する手法を比較した．またこれ以降は，ワーカーがある問題に下した判定を「判定」と呼び，ある問題に対して複数のワーカーの判定を集約した結果を「判定結果」と呼んで区別することにする．

4.2.1 多数決

多数決にはチェック問題の正答数でフィルタリングされた結果を用いる．チェック問題を全く間違えてなかったワーカーは 603 人で 44.2% であった．ここで，最も厳しい基準でクオリティ・コントロールを実行すればチェック問題を 1 問でも間違えたワーカーの回答は切り捨てることになるが，ワーカーは人間であるため 1 度は押し間違える可能性もある．そこでチェック問題のミスを 1 問も許さない場合と，1 問までチェック問題のミスを許した場合の両方の場合を 4.3 節で比較した．また，今回はチェック問題に期待される回答を全て「いいえ」として設定してしまったため，全ての判定を「いいえ」としたワーカーを切り捨てることができなかった．そのため更に全て同じ判定を下したワーカーも除外することにした．

4.2.2 ワーカーの能力と問題の難易度を考慮した判定処理

多数決ではフィルタリングとしてチェック問題の正答数のみを考慮しており，これにパスしたワーカーの能力は全て同等と考えて集約しているが，個々のワーカーの能力や問題の難易度は様々であることを考慮して集約する手法も Whitehill らによって提案されている [4]．Whitehill らは，ある問題の真の判定結果を潜在変数とし，ワーカーの回答内容からワーカーの能力と各問題の難易度をパラメーターとして EM 法により両者を推定するというモデルを提案している．以下にこの手法の概要を述べる．

表 1　人手慣用句リストに対する判定結果

手法	Accuracy
多数決 （チェック問題のミスなし）	0.81
多数決 （チェック問題のミスが 1 問以下）	0.83
Whitehill 2009	0.95

N 個の問題があった時，それぞれの問題に対する真の判定結果を $Z_j \in \{0,1\}$ とする $(j = 1,2,...,N)$．また，I 人のワーカーがいた時，ワーカー i が問題 j に付与する判定を L_{ij} とする $(i = 1,2,...,I)$．そして，ワーカー i の能力を $\alpha_i(-\infty,\infty)$，問題 j の難易度を $1/\beta_j[0,\infty)$ とし，ワーカー i が問題 j に正解する確率を

$$Pr(L_{ij} = Z_j) = \frac{1}{1+\exp(\alpha_i\beta_j)} \tag{1}$$

と定義する．ここで $\alpha_i \to +\infty$ のときは，そのワーカーは常に正しい判定をし，逆に $\alpha_i \to -\infty$ の時は常に間違った判定をする．また，$1/\beta_j = 0$ の時は，ワーカーの能力がどれだけ高くても正しい判定をする確率は 1/2 となり，逆に $1/\beta_j \to \infty$ の時は，全てのワーカーが正解できる問題となる．各ワーカーと各問題に対してこのようなパラメータを考え，EM アルゴリズムを用いて真の判定結果と，パラメータを推定することで，各問題がとる判定結果を確率的に考えることができる．今回は多数決と Whitehill らの手法を用いて判定を処理し比較した．

4.3 2 手法の比較

今回は比較のため，人手慣用句リストの 559 句の慣用句に対しても判定をつけていたので，これらの慣用句に対してそれぞれの手法で判定を集約した結果について Accuracy で比較した．多数決については「はい」が過半数を超えたものを慣用句として考える．もし，「はい」と「いいえ」が同数であった場合など判定結果が一意に決まらない場合は全て「わからない」と判定した．また，Whitehill らの手法では，慣用句である確率が過半数の 0.5 以上のものを慣用句として考えた．

多数決間の比較ではチェック問題のミスを 1 問だけ許した場合の方が若干上回っているものの，3 つの手法中では Whitehill らの手法の方が最も高かった．そのため今回は判定の集約法として Whitehill らの手法を用いることにした．

4.4 獲得された慣用句についての考察

Whitehill らの手法で集計すると慣用句は 970 句あった．結果を図 2 に示す．まず人手慣用句リストに載っ

表 2　判定結果
(#は基本慣用句五種対照表に含まれているもの)

慣用句候補	多数決の結果	はい	いいえ	わからない	慣用句確率
#世話が焼ける	はい	5	0	0	1.00
歯車が噛み合う	はい	6	1	0	1.00
脇を固める	はい	5	0	0	0.999
雲行きがあやしい	はい	7	1	0	0.999
病床に臥せる	わからない	2	1	2	0.969
語尾を濁す	わからない	2	2	0	0.576
#鼻をつつく	いいえ	1	5	0	0.037
#油を絞る	いいえ	1	3	0	0.017

ていたもので，慣用句と判定されたものは 531 句あり，28 句は慣用句とは判定されなかった．また，自動的に獲得した慣用句候補では，439 句が慣用句と判定されており，4,295 句が慣用句とは判定されなかった．この内，60 句は人手慣用句リストの基になった佐藤の基本慣用句五種対照表に含まれていたので，結局新規に獲得できた慣用句は 379 句である．

獲得された慣用句を見てみると人手慣用句リストに載っていた慣用句の多くは，高い確率で慣用句であるという判定結果になっており，多くの人にとって基本的な慣用句であると考えられる．新たに獲得された慣用句として，「歯車が噛み合う」，「脇を固める」，「雲行きがあやしい」といったものがあるが，これらも同様に高い確率で慣用句であるという判定結果が出ているため基本的な慣用句を新規に獲得できたといえるだろう．

また人手慣用句リストにあっても一般的には慣用句と思われていないものも存在することがわかった．例えば人手慣用句リストに含まれていた「油を絞る」や「鼻をつつく」は慣用句であるとは判定されなかった．これらは辞書的には基本慣用句として扱われているが，ワーカーの判定に基づけば基本的なものではないと考えられる．

一方慣用句という判定結果が出たものでも明らかに慣用句ではないものも若干含まれていた．これらの例について詳しく見てみると，判定に参加した 10 人のワーカーの全員の信頼度が低く，信頼できる判定が全く集まらなかったことが原因であった．このような問題を解決するためのタスクデザインを考えることは今後の重要な検討課題である．

5　まとめと今後の展望

本論文では自動的に獲得したデータから不特定多数のワーカーの判定を用いて，新規に慣用句を獲得することができた．これまで慣用句は慣用句であるかないかという 2 値としてしか捉えることができなかったが，多数のワーカーの判定を集約することで慣用句らしさを定量的に捉えることができた．

一方で不特定多数のワーカーの判定を利用するため，適切なフィルタリングを行わなければ判定結果に影響が出てしまうという問題も依然存在している．今後はより有効なクオリティ・コントロール手法や判定集約手法を利用して，効率的な言語リソース構築の手法を研究したい．

参考文献

[1] A. P. Dawid and A. M. Skene. *Applied Statistics*, No. 1, pp. 20–28.

[2] Chikara Hashimoto and Daisuke Kawahara. Construction of an idiom corpus and its application to idiom identification based on WSD incorporating idiom-specific features. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP 2008)*, pp. 992–1001, 2008.

[3] Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. Cheap and fast—but is it good?: Evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pp. 254–263, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics.

[4] Jacob Whitehill, Paul Ruvolo, Ting fan Wu, Jacob Bergsma, and Javier Movellan. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pp. 2035–2043. 2009.

[5] Omar F. Zaidan and Chris Callison-Burch. Crowdsourcing translation: Professional quality from non-professionals. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pp. 1220–1229, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[6] Naomi Zeichner, Jonathan Berant, and Ido Dagan. Crowdsourcing inference-rule evaluation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 156–160, Jeju Island, Korea, July 2012. Association for Computational Linguistics.

[7] 河原大輔, 黒橋禎夫. 格フレーム辞書の漸次的自動構築. 自然言語処理, Vol. 12, No. 2, pp. 109–131, 2005.

[8] 佐藤理史. 基本慣用句五種対照表の作成. 情報処理学会研究報告, 2007-NL-178, pp. 1–6, 2007.

[9] 米川明彦, 大谷伊都子. 日本語慣用句辞典. 東京堂出版, 2005.