市民科学でのアノテーション作業支援と作業者の能力向上支援

星島洸明[1]　西村太一[1]　亀甲博貴[2]　森信介[2]
[1] 京都大学大学院情報学研究科　[2] 京都大学学術情報メディアセンター
{hoshijima.komei.72x,nishimura.taichi.43x}@st.kyoto-u.ac.jp
{kameko,forest}@i.kyoto-u.ac.jp

概要

近年，市民科学を活用した研究プロジェクトが注目されている．このようなプロジェクトは多数の非専門家が作業をすることで少数の専門家のみのときより効率的にプロジェクトを進行させられる特徴がある．報酬を作業の動機とするクラウドソーシングと異なり，市民科学では作業者は科学や学問への貢献や知的好奇心を動機をしている．そのため作業者は作業対象への知識の向上を需要として持っており，本論文では成果物の分析をして，作業を通じて作業対象への知識が増えたり作業能力が向上していたりといった情報をフィードバックとして与えることを最終的な目標とする．また，非専門家による成果物の品質にばらつきが生じる場合でも学習データとして有効に活用する．

1　はじめに

近年，市民科学を活用した研究プロジェクトが注目されている．市民科学は非専門家が科学的な調査や研究活動をすることを指し，市民科学の特徴は少数の専門家のみで作業すると時間と費用が膨大にかかるプロジェクトでも多数の非専門家がプロジェクトに参加すると効率的に作業を進められることである．多数の人が参加することで効率的に作業を進める方法であるクラウドソーシングでは，インターネットを通じて不特定多数の人にタスクを依頼し，作業量に応じた報酬を受け取ることが作業に取り組む動機となっている．クラウドソーシングサービスの代表例として Amazon Mechanical Turk が知らており，画像データへのアノテーション作業や画像の内容を描写する文章作成など [1, 2] 高い専門的知識や技能を必要としない作業が主に依頼されている．市民科学では作業者の目的は報酬を得ることではなく，科学や学問への貢献，知的好奇心が動機である [3] ため，ボランティア型のクラウドソーシングといえる．クラウドソーシングを利用してデータを収集することは多くの研究でみられるが，対して市民科学が関わるデータ収集の研究はそれほど多くはない [4]．市民科学のプロジェクトの参加者は動機が知的好奇心にあるため，作業者が作業対象への知識を向上させる需要があると考えられる．よって作業者の作業によって作成された成果物の分析をして，作業を通じて知識が実際に増えているかのフィードバックを与えることが有用であると考えられる．

市民科学で作業対象が専門的知識を必要とし，作業者の専門的知識の有無を理由に作業への参加を拒まない場合，作業の結果の成果物の品質は一定ではなくなる．作業対象に対して専門的知識を多く有する人は高品質な成果物を作成しやすく，専門的知識をあまり有しない人は高品質な成果物を作成しにくい傾向があると考えられる．市民科学では作業対象に興味を持っているが専門的知識の少ない人が作業に関わっていくことで，作業対象の知識が増える可能性がある．

市民科学のプロジェクトの例として京都大学古地震研究会が 2017 年に公開した市民参加型史料翻刻プロジェクト「みんなで翻刻（地震史料）」 [3, 5] があげられる．翻刻とはくずし字で書かれている古文書などの歴史史料を一文字ずつ活字に書き起こしていく作業のことを指す．「みんなで翻刻」は web 上で歴史史料を翻刻するためのアプリケーションであり，翻刻作業には研究者だけでなく一般の人々も参加している．

このように，市民科学のプロジェクトでデータに対して人手でラベルを付与することが求められており，タスクに応じた付加情報をつけるアノテーションツールの需要がある．本論文ではモデルの精度向

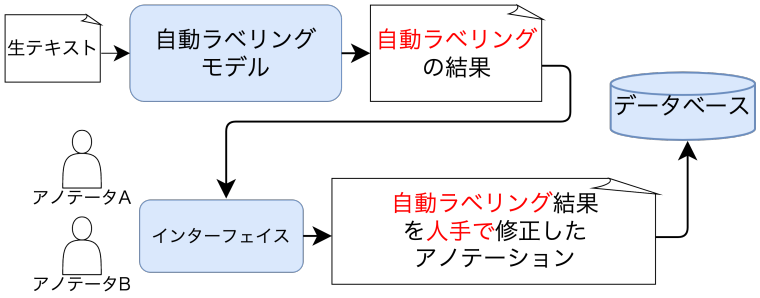


図 1　作成したアノテーションツールの概要図.

上につながる学習データを作成するために，市民科学のプロジェクトで使用されることを想定したアノテーションツールを作成する．

本論文では市民科学のプロジェクトで非専門家で作業する際に，第一に作業者の能力向上を作業者にフィードバックとして与え，第二に非専門家の成果物からモデル性能を向上させる学習データを作成することを目的とする．

2　関連研究

2.1　複数アノテータによるアノテーション

アノテータが複数いる状態で同一のテキストに対して作業をすると，アノテータの数だけのデータが作成される．各データでラベル付けされる内容は異なり，学習データとして扱うためには各データを一つに集約する必要がある．各データを一つに集約する際にトークン単位で多数決をする手法 [6] が知られている．集約したデータを学習データに加えてモデルを学習し，未作業のテキストにモデルによる自動ラベリングをすると，モデル精度を向上させるデータが作成できることが知られている [7].

2.2　非専門家を含む場合のアノテーション

専門分野に関する知識を有しない非専門家と専門家が同じテキストに対してラベル付けをする場合，まず非専門家がテキストのラベル付けをしてその生成物を専門家が修正することで，アノテーションコーパスを作成するコストを削減しアノテータ間の一致率を向上させることができる [8].

自動ラベリングモデルによる出力結果から人手で修正を加えるアノテーション方法は作業時間を短縮させることができる [9] が，誤った出力結果に影響されてデータが作成される可能性がある [10]．市民科学のプロジェクトに限らず非専門家によって大規模データセットが作成される場合，非専門家への作業支援は必要であるが，この点を考慮に入れる必要がある．

3　課題設定

「みんなで翻刻」のような市民科学のプロジェクトは 2 つの課題を抱えている．1 つ目は市民科学の参加者が作業対象の知識を向上させているかをフィードバックとして与えると有用であると考えられるいう点である．クラウドソーシングのワーカーが報酬を動機としているのに対し，市民科学の参加者は科学や学問への貢献や科学的好奇心を動機としている．そのため作業対象の知識を増やしたいという作業者の需要があると考えられる．作業を通じて実際に作業者の知識が増えているかをフィードバックすることで，市民科学に参加する動機をより強固にして，プロジェクト成功や発展につながる可能性がある．2 つ目は専門的知識が必要とされる分野で作業する際に，作業者の専門的知識に幅がある状態では，作成された成果物の品質にばらつきが生じるという点である．作業者の能力にばらつきがあることを前提に，自動ラベリング結果を表示したり [11, 12]，作業者の習熟度とタスクの難易度を考慮にいれてタスクを各作業者に適切に分配したり [13] して，作業を支援することができる．以上のような課題がある中で，市民科学において非専門家が作業する場合，モデルの性能を向上させるような高品質なデータを作成し，作業者の作業能力が向上しているかをフィードバックとして与えることを本論文の目的とする．

4　アノテーションツール

テキスト中の固有表現をアノテーションするツールを作成した．本ツールの概要図を図 1 に示し，ツールでラベル付けするときの作業画面を図 2 に示

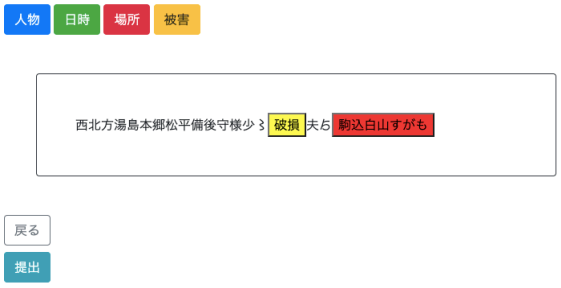


図 2　アノテーションツールの作業画面.

す．作業するテキストには学習した自動ラベリングモデルを使用してラベル候補をアノテータに提示する．アノテータはツールによって提示されたラベルの範囲と種類が正確であると判断した場合はそのままにしておき，誤っていると判断した場合は修正する．ツールによって提示されたラベル以外にラベルを付ける必要があるとアノテータが判断すれば新たにテキストにラベルを追加する．ツールはアノテータが作業開始から作業終了までの時間を測定しており，ラベル付けされたデータとともに測定時間をデータベースに格納する．

5　実験

学習したモデルを使用して，ラベル付けする候補をアノテータに提示して，人手でそれを修正することでアノテーション作業の効率化が見られるか実験する．またアノテータによって作成されたデータを元の学習データに追加してモデルを学習し，データ追加前のモデルと性能を比較して評価する．

5.1　実験設定

本実験では，固有表現をアノテーション対象とし，対象データセットは料理分野のレシピテキストと歴史資料テキストの 2 つである．歴史資料テキストは江戸時代に発生した安政江戸地震についてのテキストであり，固有表現タグを付与したデータセットである．料理分野のデータセットを作成するとき文を単語分割し，単語ごとに固有表現タグを付与する．歴史資料は現代語のテキストと異なる文法，語彙が含まれ，一文ごとの区切りも曖昧であるため，歴史資料のデータセットを作成するときは文を文字単位で分割した後に文字ごとに固有表現タグを付与する．作業支援のための自動ラベリングモデルの学習に使用したデータセットを表 1 に示す．表中の数字は文数を表す．なお自動ラベリングを学習するために使用した学習データを以降では元データと呼ぶ．

表 1　作業支援をする固有表現認識モデルの学習に使用したデータ.

対象データ	学習データ	開発データ	テストデータ
レシピテキスト	2,358	372	387
歴史資料テキスト	2,295	286	286

表 2　作業支援をする固有表現認識モデルの性能.

学習データ	適合率	再現率	F 値
レシピテキスト	0.860	0.902	0.880
歴史資料テキスト	0.618	0.623	0.621

アノテータは 3 名で，日本語母語話者であり，料理分野，歴史資料分野に関して専門的知識は有していない．以後アノテータ 3 名をそれぞれ A,B,C とする．各アノテータそれぞれが同じ 200 文にラベル付けをし，200 文のうち 100 文は作業支援あり，100 文は支援なしでラベル付けを行った．なお，レシピテキストでは支援ありの 100 文はアノテータ毎に異なり，歴史資料テキストでは全アノテータで同じ支援ありの 100 文に作業を行った．

固有表現認識モデルは Flair[14] を使用した．固有表現認識モデルの学習で使用する分散表現は，Flair が提供する文字ベースの日本語の事前学習済み言語モデル [15] を使用して構成した．以下にモデル学習時の各パラメータを説明する．隠れ層のサイズは 256，学習率は 0.1，バッチサイズは 32，最大エポック数は 150 とした．作業支援をするレシピテキストでの固有表現認識モデルの性能と歴史史料テキストでの固有表現認識モデルの性能を表 2 に示す．

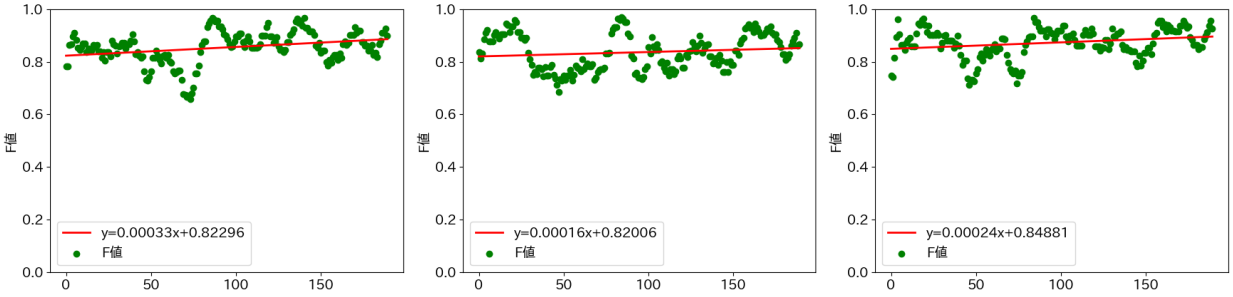
5.2　評価指標

データ品質の評価には各アノテータの成果物を元の学習データに追加して学習したモデルの性能の測定を行った．データ作成の効率性の評価には作業時間の比較を行った．

5.3　実験結果

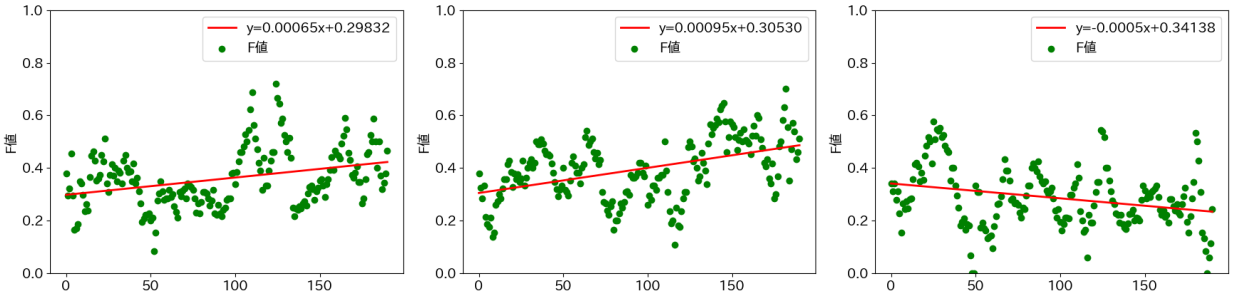
5.3.1　データ作成の効率性

各作業設定で作業にかかった時間の平均を表 3 に示す．レシピテキスト，歴史資料テキストの両データセットにおいて，生テキストから人手アノテーションをする場合より自動ラベリング結果から人手アノテーションをする場合の方が作業時間は短くなった．



アノテータ A　　アノテータ B　　アノテータ C

図 3　レシピテキストでの成果物の F 値の移動平均.



アノテータ A　　アノテータ B　　アノテータ C

図 4　歴史史料テキストでの成果物の F 値の移動平均.

表 3　作業時間の比較.

対象データ	作業設定	作業時間 (秒)
レシピテキスト	生テキストから	2,231
	自動ラベリング結果から	1,261
歴史資料テキスト	生テキストから	1,741
	自動ラベリング結果から	1,362

5.3.2　成果物の品質推移

アノテータが作成した成果物の品質が作業の経過とともに変化するか調べるため，各アノテータの成果物の F 値の移動平均を計算した．各アノテータの成果物を提出順に並べ，F 値の移動平均を 10 文毎にとった結果を図 3，4 に示す．レシピテキストではいずれのアノテータも作業を進めるにしたがって，成果物の F 値の移動平均は向上していった．歴史資料テキストではアノテータ A,B では移動平均が向上したが，アノテータ C は減少していった．この結果から，歴史資料テキストではモデルの自動ラベリング性能がレシピテキストほど高くないため，モデルの出力結果を利用した作業者の能力向上が起こりにくかった可能性が考えられる．

6　おわりに

本論文では市民科学でアノテーション作業が行われる場面を想定して，自動ラベリングモデルによる作業支援を行った．成果物の品質推移を測定することで作業能力の変化をフィードバックとして与え，モデル性能を向上させるデータ作成を目的とし，実験を行った．今後は自動ラベリング結果がラベル毎に異なる信頼度を持つことを考慮して，ラベル毎の信頼度を視覚的に表示するシステムを構築する．

参考文献

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In **2009 IEEE conference on computer vision and pattern recognition**, pp. 248–255. Ieee, 2009.

[2] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In **European conference on computer vision**, pp. 740–755. Springer, 2014.

[3] 橋本雄太, 加納靖之, 一方井祐子, 小野英理ほか. 『みんなで翻刻』 の運用成果と参加動向の報告. じんもんこん 2020 論文集, Vol. 2020, pp. 39–46, 2020.

[4] Linda See, Peter Mooney, Giles Foody, Lucy Bastin, Alexis Comber, Jacinto Estima, Steffen Fritz, Norman Kerle, Bin Jiang, Mari Laakso, et al. Crowdsourcing, citizen science or volunteered geographic information? the current state of crowdsourced geographic information. **ISPRS International Journal of Geo-Information**, Vol. 5, No. 5, p. 55, 2016.

[5] みんなで翻刻｜歴史資料の参加型翻刻プラット

フォーム. https://honkoku.org. (参照 2021-12-28).
[6] Filipe Rodrigues, Francisco Pereira, and Bernardete Ribeiro. Sequence labeling with multiple annotators. **Machine learning**, Vol. 95, No. 2, pp. 165–181, 2014.
[7] An T Nguyen, Byron C Wallace, Junyi Jessy Li, Ani Nenkova, and Matthew Lease. Aggregating and predicting sequence labels from crowd annotations. In **Proceedings of the conference. Association for Computational Linguistics. Meeting**, Vol. 2017, p. 299. NIH Public Access, 2017.
[8] Mary Martin, Cecilia Mauceri, Martha Palmer, and Christoffer Heckman. Leveraging non-specialists for accurate and time efficient AMR annotation. In **Proceedings of the LREC 2020 Workshop on "Citizen Linguistics in Language Resource Development"**, pp. 35–39, Marseille, France, May 2020. European Language Resources Association.
[9] Claudia Schulz, Christian M Meyer, Jan Kiesewetter, Michael Sailer, Elisabeth Bauer, Martin R Fischer, Frank Fischer, and Iryna Gurevych. Analysis of automatic annotation suggestions for hard discourse-level tasks in expert domains. **arXiv preprint arXiv:1906.02564**, 2019.
[10] Karën Fort and Benoît Sagot. Influence of pre-annotation on pos-tagged corpus development. In **The fourth ACL linguistic annotation workshop**, pp. 56–63, 2010.
[11] Seid Muhie Yimam, Chris Biemann, Richard Eckart de Castilho, and Iryna Gurevych. Automatic annotation suggestions and custom annotation layers in webanno. In **Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations**, pp. 91–96, 2014.
[12] Todd Lingren, Louise Deleger, Katalin Molnar, Haijun Zhai, Jareen Meinzen-Derr, Megan Kaiser, Laura Stoutenborough, Qi Li, and Imre Solti. Evaluating the impact of pre-annotation on annotation speed and potential bias: natural language processing gold standard development for clinical named entity recognition in clinical trial announcements. **Journal of the American Medical Informatics Association**, Vol. 21, No. 3, pp. 406–413, 2014.
[13] Yinfei Yang, Oshin Agarwal, Chris Tar, Byron C Wallace, and Ani Nenkova. Predicting annotation difficulty to improve task routing and model performance for biomedical information extraction. **arXiv preprint arXiv:1905.07791**, 2019.
[14] Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. FLAIR: An easy-to-use framework for state-of-the-art NLP. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)**, pp. 54–59, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.
[15] Alan Akbik, Duncan Blythe, and Roland Vollgraf. Contextual string embeddings for sequence labeling. In **Proceedings of the 27th International Conference on Computational Linguistics**, pp. 1638–1649, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics.
[16] Tetsuro Sasada, Shinsuke Mori, Tatsuya Kawahara, and Yoko Yamakata. Named entity recognizer trainable from partially annotated data. In **Conference of the Pacific Association for Computational Linguistics**, pp. 148–160. Springer, 2015.

表 4 レシピテキストで部分的アノテーションを使用した固有表現認識モデルの性能.

学習データ	適合率	再現率	F 値
元データ	0.825	0.875	0.849
元データ+フルアノテーション	**0.832**	0.877	**0.854**
元データ+部分的アノテーション	0.827	0.877	0.851

表 5 歴史資料テキストで部分的アノテーションを使用した固有表現認識モデルの性能.

学習データ	適合率	再現率	F 値
元データ	0.705	0.678	0.691
元データ+フルアノテーション	0.706	0.673	0.689
元データ+部分的アノテーション	**0.711**	**0.690**	**0.701**

A 付録

A.1 部分的アノテーションを使用した固有表現認識性能

学習に部分的アノテーションを使用するために，まず複数人で収集したアノテーションを統合する．各アノテーションを使用して，トークン単位で多数決を行い，過半数の場合ラベルを採用して統合したアノテーションデータを作成する．

次に以下の手順で部分的アノテーションを作成し，モデルを学習する．第一に，アノテーションデータを学習データ，開発データ，テストデータに分割する．第二に，各固有表現が部分的アノテーションに含まれるかの全組み合わせを作成する．第三に，学習データを使用して，固有表現の組み合わせにもとづいた部分的アノテーションを作成する．レシピテキストの固有表現は 8 個, 歴史資料テキストの固有表現は 4 個あるので，上記の手順でそれぞれ 254 個，14 個の部分的アノテーションを作成した．第四に，作成した各部分的アノテーションを元データに追加し，それぞれモデルを学習する．各モデルを開発データで評価し，最も性能が高いモデルを 1 つ選択する．第五に，選択したモデルをテストデータで評価する．

実験では部分的アノテーションを学習可能な固有表現認識モデルである PWNER[16] を使用した．元データに部分的アノテーションを追加して学習したときの固有表現認識モデルの性能を表 4,5 に示す．レシピテキストでは元データにフルアノテーションを加えたとき最も精度が高く，歴史資料テキストでは元データに部分的アノテーションを加えたとき最も精度が高くなった．この結果から，部分的アノテーションを使用することでフルアノテーションをそのまま使用するよりもモデルの精度が高くなる場合があることが示された．