複数作業者を想定したアノテーションツールの作成と機能の検討

星島洸明
京都大学大学院情報学研究科
hoshijima.komei.72x@st.kyoto-u.ac.jp

西村太一
京都大学大学院情報学研究科
nishimura.taichi.43x@st.kyoto-u.ac.jp

亀甲博貴
京都大学 学術情報メディアセンター
kameko@i.kyoto-u.ac.jp

森信介
京都大学 学術情報メディアセンター
forest@i.kyoto-u.ac.jp

1 はじめに

近年，市民科学 (Citizen Science) 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行うことを指す．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．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）」 [1] が挙げられる．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，多くの研究分野において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いる．そのため，タスクに応じたラベルをつけるアノテーションツールの需要がある．アノテーションツールを使用するのが複数人であると想定すると，作業する人それぞれが対象の専門的知識をどの程度有しているかは不明である．また CUI に慣れていない人も含まれるためツールはマウスのクリックと簡単なキーボードからの入力のみで扱えることが好ましい．

本研究では，非専門家が複数人いる状態で効率的にアノテーションすることを目指して，固有表現を題材にしたアノテーションツールを試作した．実験ではアノテーションツールを使用して固有表現のアノテーションコーパスを作成した．実験の結果，固有表現認識モデルの支援がある状態のほうが，支援がない場合よりラベル付けされたアノテーションコーパスの品質は向上した．作成したそれぞれのコーパスで学習し比較した結果，支援がある場合で作成したアノテーションコーパスで学習したほうが F 値が向上した．

2 関連研究

2.1 アノテータが複数いる状態のアノテーション

複数のアノテータが同一の文書に対してアノテーション作業を行う際，複数のアノテーションデータから最も優れているラベルデータを選出し，作成されたアノテーションデータで学習したモデルでまだアノテーションされてないテキストに対してラベル候補を提示する研究がある [2].

クラウドソーシングでラベル付きのデータを収集するとアノテータがラベル付けしたタスクに対して専門的知識をどの程度有しているかはアノテータそれぞれで異なる．そのため，アノテーションの品質管理はクラウドソーシングにおいて重要な課題である．アノテータつけたラベルが誤りであるかを推定する方法として同じデータを複数のアノテータにラベル付けしてもらい，その結果の多数決をとることでラベルの統合を行う方法がある．しかしこの方法は各アノテータの誤り率が同等であることを前提とするため必ずしも効率的な方法ではない．アノテータの誤り率を推定するため専門的知識の有無などを考慮に入れた方法が提案されている [3].

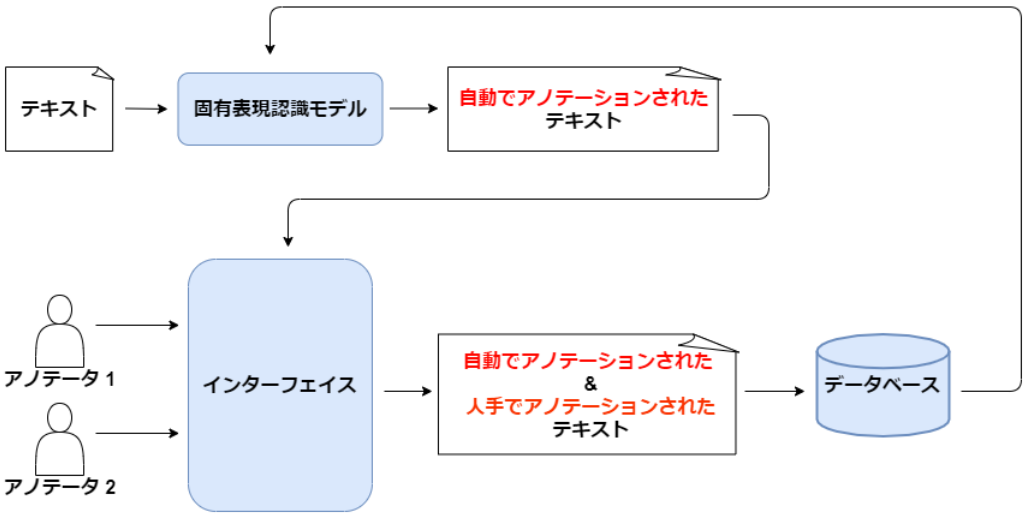


図 1　試作したツールの全体図.

2.2　非専門家がいる状態のアノテーション

Snow らの研究 [4] ではクラウドソーシングによって専門家，非専門家それぞれがアノテーションコーパスを作成した．非専門家のアノテータが作成したデータの品質の偏りがあるため，データの品質からアノテータの評価を行い管理した．その結果，設定した 5 つのタスクにおいて非専門家が作成したアノテーションコーパスは専門家が作成したアノテーションコーパスの 4 倍用意すると識別機の性能が同じになった．

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生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た研究がある [5].

2.3　既存のアノテーションツール

系列ラベリングで使用される既存のテキストアノテーションツールはショートカットキーの設定 [6]，能動学習を使用可能 [7] といった機能があり，効率的なアノテーションコーパスの作成ができる．本研究では，こうしたツールを参考に，(1) ラベル候補の提示機能と (2) ショートカットキーの設定機能が実装されており，効率よくアノテーションを行うことができる．

図 2　ツールのインターフェイス

3　アノテーションツール

複数人が作業することを想定した，ラベル付けをするためのツールを開発した．本ツールの概要図を図 1 に示す．ツールのラベル付けするときのインターフェイスを図 2 に示す．本ツールはアノテータが複数人であることを想定しており，各アノテータはアカウントを作成し各々が作業するプロジェクトを作成する．次にプロジェクト毎にテキストに付与したいラベルを設定する．作業するテキストには事前に学習した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案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ラベルを削除する．ツールによって付与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それらの時間をデータベースに格納する．各アノテータによってラベル付けされたデータを元の学習データに加えて，新たにモデルを学習す

表 1　レシピ用語のタグ一覧.

タグ	意味
Ac	調理者の動作
Af	食材の変化
D	継続時間
F	食材
Q	分量
Sf	食材の様態
St	道具の様態
T	道具

ることで自動ラベリングモデルの精度をより良くしてアノテータの負担を削減する.

4　実験

訓練データから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観測する実験を行う．本実験では上記のように新たにモデルの学習はしていない．アノテータによってラベル付けされたデータを正解データと比較して評価する.

4.1　実験設定

本実験では，アノテーション課題としてレシピ分野を対象とした固有表現認識を行う．使用するデータセットは，笹田らの基準 [8](表 1) に基づき定義されたレシピ用語を，IOB 形式でレシピ分野に精通した人がラベル付けしており，これを正解データとして活用した．支援のための固有表現認識モデルにはこのデータセットの学習データ 2,358 文使用した．また，テストデータには 387 文使用した．固有表現認識モデルには BiLSTM-CRF [9] を用いた．アノテータは 3 名で，日本語母語話者でありレシピ分野に関して専門的知識は有していない．各アノテータそれぞれは同じ 200 文にラベル付けをした．200 文のうち 100 文は支援あり，100 文は支援なしでラベル付けを行った．なお，支援ありの 100 文はアノテータ毎に異なる.

以下に実験で使用したパラメータを説明する．単語埋め込み層の入力は 1,640, 出力は 512 で，LSTM 層の出力は 1,024，最終的な出力は 17，バッチサイズは 10 で実施した．アノテータ支援のための固有表現認識モデルの適合率，再現率，F 値はそれぞれ 0.8164, 0.8425, 0.8292 であった.

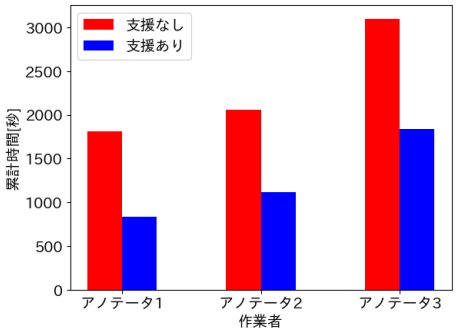


図 3　作業にかかった時間.

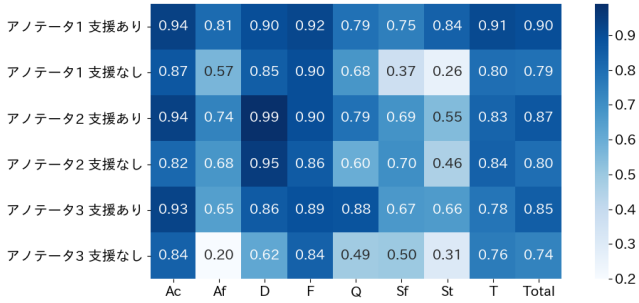


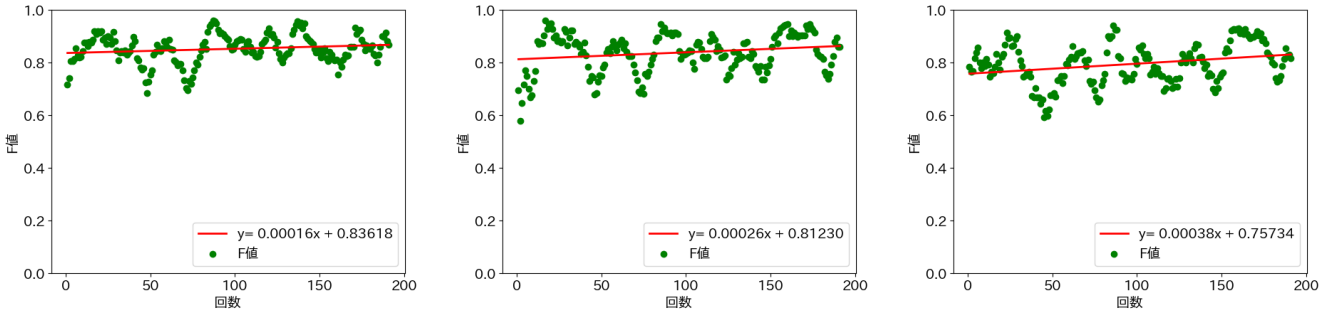
図 4　ラベル付けされたデータの品質.

4.2　実験手順

本ツールではアノテータが各自アカウントを作成しそれぞれのアノテータ情報とアノテーションデータを紐付けられている．そのため，アノテーション作業にかかる時間をツールで測定してツールによる支援がある場合と支援がない場合でかかる時間の差を見られるようになっている．アノテータが作業する際，ツールによる支援がある場合には，事前に学習した固有表現認識モデルを使用して，ツールの画面に表示されるテキストの中から固有表現を認識して画面上に表示する．ツールによる支援はそれぞれのアノテータが作業する 200 文のうちランダムに選ばれた 100 文に対して行われた．支援がある文と支援がない文はランダムに選ばれ，アノテータは作業を行う文を選択するまで支援があるかないか知ることができない.

作成されたアノテーションコーパスを正解データと比較して適合率，再現率，F 値を計算した.

4.3　結果



アノテータ 1　　アノテータ 2　　アノテータ 3

図 5　ラベル付けされたデータの品質の移動平均.

表 2　ラベル付けされたデータの品質.

	適合率	再現率	F 値
支援なし	0.8218	0.7497	0.7839
支援あり	**0.8955**	**0.8596**	**0.8770**

表 3　ラベルの認識性能 (学習データ:100 文).

	適合率	再現率	F 値
支援なし	0.6000	0.5856	0.5923
支援あり	**0.6427**	**0.6337**	**0.6370**

4.3.1　作業時間

作業にかかった時間を図 3 に示す．今回実験に参加したいずれのアノテータでも，支援がある場合の方が支援がない場合よりも少ない時間で作業を終了させることができた．

4.3.2　データ品質

ラベル付けされたデータの品質を図 4 に示す．値は F 値を表す．ラベル「Af」,「St」に関してアノテータ 2 の支援なしの場合は他のアノテータより比較的品質が高かった．そのためアノテータ 2 はこれらのラベルを付けることが比較的得意である可能性がある．

アノテータ 3 名のラベル付けされたデータの品質を表 2 に示す．表中の値は固有表現認識モデルによる支援がない場合と支援がある場合それぞれにおける適合率，再現率，F 値を平均で示している．支援なしの場合よりも支援ありの場合のほうがラベル付けされたデータの品質は向上した．支援のための固有表現認識モデルによる自動ラベリングの品質の F 値は，支援がある場合よりは低く支援がない場合より高い結果になった．

アノテータが作成したデータの品質が作業の経過とともに変化するか調べるため，各アノテータのデータ品質の移動平均を計算した．各アノテータのデータの品質の移動平均を 10 文毎にとった結果を図 5 に示す．いずれのアノテータも作業を進めるにしたがってラベル付けしたデータの品質は向上していった．向上した理由がアノテータが固有表現認識モデルの出力から学習したためか次第に作業に慣れていったためかはさらなる調査が必要である．

4.3.3　固有表現認識性能

各アノテータがラベル付けしたデータで学習した固有表現認識モデルの性能を表 3 に示す．いずれのアノテータでも支援がある場合の方が支援がない場合よりラベル付けされたデータの品質が向上した．

5　おわりに

本研究では，非専門家が専門分野のアノテーションコーパスを効率的に作成するためのツールを試作した．試作したツールは事前に学習したモデルによるラベルの提案機能を持つ．支援があると効率よくアノテーション作業がすすめられ，ラベル付けデータの品質は向上した．また，アノテータがそれぞれ提出したアノテーションコーパスで学習した固有表現認識モデルは支援がある場合のほうが固有表現認識性能が高くなった．

アノテータ毎のラベル付けされたデータの品質から，(1) 特定のアノテータは特定のラベルの種類に関してラベル付け作業が得意である場合があること，(2) 特定のラベルに比較的詳しいと思われるアノテータとそうではないアノテータがそれぞれ作成したアノテーションコーパスではデータの品質が異なることが分かった．そのためアノテータ毎に比較的信用できるラベルに重みを付けてアノテータテータコーパスを作成していくと固有表現認識モデルを効率的に学習できる可能性がある．

今後は，本ツールにアノテータのラベル付けの誤り率からアノテータの品質を考慮し管理する機能を追加する予定である．また，能動学習などを活用して固有表現認識モデルを逐次更新をする機能を追加することも検討する．

参考文献

[1] 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 2021-1-13).

[2] 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.

[3] Peter Welinder, Steve Branson, Pietro Perona, and Serge Belongie. The multidimensional wisdom of crowds. Advances in neural information processing systems, Vol. 23, pp. 2424–2432, 2010.

[4] Rion Snow, Brendan O'connor, Dan Jurafsky, and Andrew Y Ng. Cheap and fast–but is it good? evaluating non-expert annotations for natural language tasks. In Proceedings of the 2008 conference on empirical methods in natural language processing, pp. 254–263, 2008.

[5] 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.

[6] doccano. https://github.com/doccano/doccano. (参照 2021-1-13).

[7] Bill Yuchen Lin, Dong-Ho Lee, Frank F. Xu, Ouyu Lan, and Xiang Ren. AlpacaTag: An active learning-based crowd annotation framework for sequence tagging. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 58–63, Florence, Italy, July 2019. Association for Computational Linguistics.

[8] 笹田鉄郎, 森信介, 山肩洋子, 前田浩邦, 河原達也. レシピ用語の定義とその自動認識のためのタグ付与コーパスの構築. 自然言語処理, Vol. 22, No. 2, pp. 107–131, 2015.

[9] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 260–270, 2016.