Tatara: 支援機能を持ったアノテーションツールの構築

山村 崇 [1,3]　嶋田 和孝 [1,3]　吉川 和 [2,3]　岩倉 友哉 [2,3]
[1] 九州工業大学　[2] 株式会社富士通研究所　[3] 理研 AIP-富士通連携センター

{t_yamamura, shimada}@pluto.ai.kyutech.ac.jp
{y.hiyori, iwakura.tomoya}@fujitsu.com

1　はじめに

昨今の情報システムにおいて，機械学習は欠かせない存在となっている．機械学習を用いたシステムの精度向上のため，一般には機械学習のアルゴリズムの提案や改良に焦点が当たりがちであるが，実際には適用するデータの量や質が何よりも重要であることが多い．少量の学習データしかない場合は，できるだけノイズを含まない良質なデータが必要であるし，量を拡充したいという思いは実用面では必ずつきまとう．一方で，良質のデータを用意することは容易ではない．さらに大量のデータを人間が手作業のみで準備することは現実的ではない．したがって，効率的に学習のためのデータを収集し，適切なラベル付けなどの作業（アノテーション）が必要である．

以上のように，良質なデータをできるだけ大量に用意することは，何らかのシステムを作るために不可欠である．そのために，できるだけ容易にデータ作成を行えるアノテーション技術の確立が急務である．本研究では，この問題を解消するための効率的なアノテーション技術や環境の整備をすることを目的としている．図 1に全体像を示す．最終的なゴールは，ある分野の専門知識が十分ではないようなアノテータ（アノテーション作業をする人物）が，一定の専門性や分野特有の知識が必要なデータに対して効率的にアノテーションを行える枠組みを構築することである．そのために，類似データから構築された分類モデルの情報や少量の専門家から得られた知識を基に構築された予測モデル/フィードバックモデルを構築し，アノテータに気づきを与え，効率的で質の高いアノテーションが行えるようなツール（Tatara）を構築する．

本稿では，その中でもスパンベースの系列のラベリングを対象とする．具体的には化合物に関するラベリングタスクを対象とした分析を行う．システムの背後で動作する予測モデルを擬似的に作成し，その予測モデルがどの程度アノテーションに影響するかを調査する．

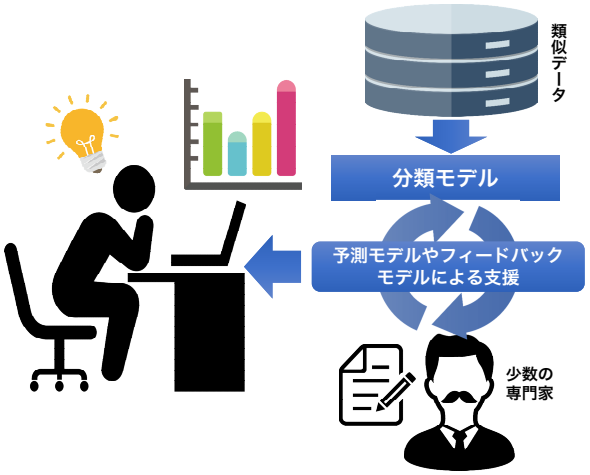

図 1: システムの全体像.

2　関連研究

Amazon Mechanical Turk などの出現により，近年ヒューマンコンピュテーションとクラウドソーシングに関する研究が活発に行われている [1]．Qing ら [5] は，アノテーション統合手法（単純な多数決，Social choice theory に基づく手法，生成モデルに基づく手法）を NLP の分類タスクで比較した．分析の結果，事例ごとのアノテータ数が多いほど精度が上がるといった示唆が得られている．Shah and Zhou [6] は self-correction（自己訂正）という考え方に基づき，アノテーションの品質を向上させる手法について提案している．本研究で提案するツールも機械学習や少数の一定の質を保ったアノテーションデータ，アノテータの知見を有効利用し，支援しようとする点は類似している．

実際に，Web 上やスタンドアローンで動作するアノテーションツールも多く存在する．BRAT はスパンアノテーションをするための最も有名なツールの一つである [7]．YEDDA のようにマルチプラットフォームに対応し，軽量なアノテーションツールも存在する

[9]．doccano[1]はスパンアノテーションのみならず，文書のラベル付けなど，さまざまなアノテーション作業が可能である．本研究で開発しているアノテーションツールも同様にスパンアノテーションだけではなく，文のラベル付けが可能である [8]．このように，近年，様々なアノテーションツールやシステムが提案・開発されており，これはアノテーション手法や作業環境に関する研究の重要性の高さを示唆している．

3　アノテーションツール

Tatara は，Python と Javascript によって開発されており，Web ブラウザ上で動作する．システムの外観を図 2に示す．現段階ではスパンアノテーション（図 2(a)）と文単位のラベル付け（図 2(b)）が可能である．doccano と同様にショートカットキーによるラベル付けが可能であり，加えて，プルダウンメニューによる要素の選択，直前のアクションの継続や取り消しも可能である．さらに YEDDA などのように，管理者権限でログインすれば，どのタグがどのぐらい選ばれているかなど，アノテーション結果の分析画面（図 2(c)）を閲覧することができる．

4　実験

本節では，Tatara が提供する支援機能の有効性検証を行う．図 1に示したように，本来であれば，別のデータなどから学習された分類モデルや少量の専門家による知見を利用することが求められるが，本稿では，この予測/フィードバックモデルを既知のデータから擬似的に生成することで，その擬似的な支援機能がどの程度アノテーションの正確さなどに影響するかなどを検証する．

4.1　検証対象

本稿では，図 2(a) のスパンアノテーションを対象とする．データは ChemdNER コーパス [2] であり[2]，科学論文の抄録中の要素（エンティティ）について，エンティティの範囲（開始文字位置・終了文字位置）と 7 つのラベル（TRIVIAL, SYSTEMATIC, ABBREVIATION, FORMULA, FAMILY, IDENTIFIER, MULTIPLE）に対して一つのラベルを割り振る問題である．

4.2　擬似的な支援機能の設定

今回設定する支援機能は，タグ付けすべき箇所を事前に候補として列挙するものとする．ChemdNER のデータに対しての Name Entity Recognition タスクの精度は 90%強であることが報告されている [4]．現状で，この最高精度相当の NER モデルが利用できると仮定し，タグ付けすべき箇所の候補を例示する支援機能を擬似的に作成する．

本稿での支援機能のポリシーとして，支援箇所は多い方がアノテータの作業は楽になるが，一定の正確さを保たなければならない，ということを念頭に置くことにする．現状の精度が 90%程度であるという前提から，Recall を 50%に抑えれば，一定の正確性が保てるであろうというナイーブな仮定を置き，その場合に全く正しくない箇所を候補として列挙する割合を 10%程度だと仮定する．そして，ChemdNER coprus にすでにアノテートされている結果（正解データ）を利用し，上記の割合に沿って，支援箇所を決定する[3]．さらに，選択範囲については多少の誤りや誤差があるもの

(a) スパンアノテーション

(b) 文単位のアノテーション

(c) 管理者画面による結果の分析

図 2: システムの外観．

[1] https://github.com/chakki-works/doccano

[2] ただし，実験で利用したガイドラインは ChemdNER-patent[3] であることに注意．加えて，現段階 (2019 年 1 月時点) ではコーパス，ガイドラインともリンク切れである．

[3] つまり，正解データの 50%程度をタグ付け箇所の候補として利用する．

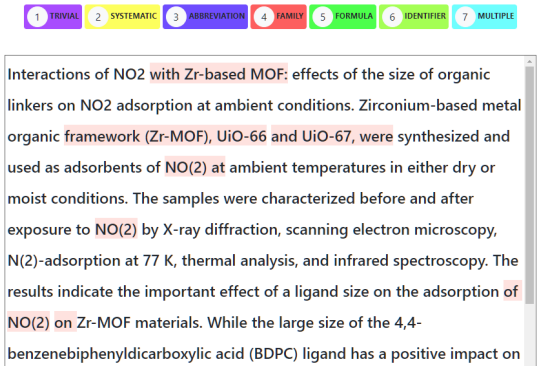

図 3: タグ付け箇所の候補を可視化.

と仮定し，前後の単語を 25%程度の確率で誤って含むようにする．これを擬似的な支援機能の出力とする．支援箇所はアノテートしたいデータがツールで開かれた場合に，薄いピンク色で提示される[4]．具体的な例を図 3に示す．図中の「NO(2)」などが，すでにコーパスでアノテートされている正解データであり，「with Zr-based MOF」の「with」は前後誤差の条件によって色づけがされている．

4.3 実験設定

支援機能によってどの程度アノテーションの正確さに差が出るかなどを評価した．実験では 15 人の専門家に 10 個の文書 (doc1〜doc10) のアノテーションを依頼した．これら 10 文書は，ChemdNER corpus から，1 文書中に 10 個以上のエンティティと 5 種類以上のラベルがアノテーションされており，かつ 50%以上が異なるエンティティであるという条件を満たすものからランダムに選んだ．

アノテータは，doc1 から順番にアノテーションを行うこととした．アノテータは過去に ChemdNER patent に関するアノテーションの経験があり，アノテーションそのものについては一定の知識を持っているが，本ツールを利用するのは初めてである．従って，アノテーション実験中に生じる何らかの問題が支援機能によるものではなく，ツールに不慣れなことから生じるものの場合もある．そこで，実験では最初の 3 つのデータを練習用と位置づけ（ただし，被験者をそのことは知らない），doc4〜doc10 の 7 文書を分析に利用した．支援機能の有無によるアノテーション結果を分析するために，まずアノテータをほぼ同数になるように A 群と B 群に分けた．その上で，各文書において，支援機能のあり/なしが A 群と B 群で異なるようにアノテーションタスクを割り振った[5]．

[4] 今回の支援機能では，サポートするのは位置だけであり，ラベルの種類までは考慮していない．

候補箇所			候補なし	合計
正解	前後誤差	間違い		
21	53	14	88	176

表 1: 正解データのエンティティ数と候補箇所の内訳.

	支援機能なし			支援機能あり		
	精度	再現率	F 値	精度	再現率	F 値
平均	0.467	0.474	0.455	**0.538**	**0.506**	**0.499**

表 2: 正解データとの一致率（精度，再現率，F 値）.

表 1に，支援機能として候補する箇所の内訳と各文書の正解データのエンティティ数を示す．4.2節で述べたように，正解データのエンティティ数に対して候補する箇所の割合は 50%程度とした．また，候補する箇所の 90%は正解のエンティティを含むようにし，10%は間違った箇所を候補とするように設定した．

支援機能の有無によるアノテーションの正確さを分析するために，各文書に対して正解データとアノテータが付与したエンティティとの一致率を計算した．正解データのエンティティとアノテータが付与したエンティティの開始文字位置・終了文字位置・ラベルの 3 つが一致していれば正解とした．各文書に対して，精度，再現率，F 値を算出し，それぞれの平均値を算出した．

また，アノテーションを開始した時刻（アノテーション画面の起動時）から，終了した時刻（アノテーションの提出処理時）の差をアノテーションの所要時間として計測した．

4.4 実験結果と分析

表 2に，支援機能なしと支援機能ありの場合におけるアノテーション結果を示す．実験データ全体では，支援機能を活用したアノテーションが，支援機能なしと比較して，すべての評価で上回っていることを示している．傾向として精度が大きく向上しており，候補箇所を利用したアノテーションを行うことで，アノテータがエンティティをより正確にアノテーションできるようになったと考えられる．

より具体的な一致率の向上の要因を分析するために，候補箇所の種類ごとによる再現率を分析した．表 3に，正しい候補箇所（正解）と前後に誤差を含んだ候補箇所（前後誤差），および候補していないエンティティごとにおける再現率を示す．支援機能なしのアノテー

[5] つまり，ある文書を A 群が支援機能ありでアノテートした場合，B 群は支援機能なしでアノテートする．

	支援機能なし			支援機能あり		
	正解	前後誤差	候補なし	正解	前後誤差	候補なし
平均	0.426	**0.597**	0.416	**0.474**	0.582	**0.473**

表 3: エンティティの種類ごとの再現率.

	支援機能なし	支援機能あり
平均	18.508	**16.788**

表 4: アノテーションの所要時間（分）.

ションと比較して，支援機能ありのアノテーションでは，正しい候補箇所の再現率は向上している一方で，前後に誤差を含んだ候補箇所のエンティティに対しては再現率が低下していることが確認された．本実験で扱ったような化合物名のエンティティでは複合語が多いこともあり，前後にノイズを含んだような箇所を候補として列挙するとアノテータは正しい判断が難しくなると考えられる．そのため，本当に正しいと考えられる箇所のみを候補とするような方策を立てることで，より正確なアノテーションが期待できる．また，候補されていない箇所のエンティティに対する再現率も向上していることから，アノテータは候補されていない箇所に対しても適切なアノテーションができていることがわかる．加えて，候補なしの箇所でも再現率が上がっているというこの実験結果は，部分的な支援機能がアノテータのアノテーション作業への集中力や注意力の向上（たとえば，候補が示されることで候補提示箇所へのアノテーションが素早くできるようになり，それに比例して，提示された候補箇所以外の場所を注視できる時間が増えたなど）につながった可能性を示しており，支援機能を持ったアノテーションツールの潜在的な有効性を示していると考えられる．この点についてはさらに詳細な分析が必要である．

アノテーションの所要時間[6]を表 4に示す．支援機能なしと比較して，支援機能ありのアノテーションの所要時間が少ないことから，本実験の支援機能を活用することで，効率的かつより正確なアノテーションを支援できることがわかる．

5 おわりに

本論文では，アノテータに対する支援機能を持ったアノテーションツール Tatara を構築し，化合物に対するスパンアノテーションタスクを対象として実験を行なった．本実験の支援機能として，化合物名の候補だと考えられる箇所に対してアノテータに事前に提示した．実験では，支援機能を用いたアノテーションが，支援機能がない場合と比べて，アノテーションの所要時間を削減しつつ，アノテーションの正確さを支援できることを確認した．

本稿では化合物名に関するスパンアノテーションを扱ったが，3節で述べたように，本ツールは他のアノテーションタスクもサポートしている．実際に，我々が別の研究で開発している対話コーパス（Kyutech コーパス）のアノテーションにも利用できることを確認している [8]．今後も様々なタスクについてその有効性を検証する予定である．現在我々はアノテータの専門性を考慮した割り当て法についても研究を進めている [10]．このようなアプローチとの連携も今後の課題の一つである．なお，本アノテーションツール Tatara は，オープンソースとして公開予定[7]である．

[6] 5 分以上の操作がない時刻は，タグ付けを行う時刻の間隔の平均値で置き換えた．10 分・30 分・60 分以上の場合でも確認したが，いずれも支援機能ありの場合が所要時間が短くなった．

参考文献

[1] 鹿島久嗣, 小山聡, 馬場雪乃. ヒューマンコンピュテーションとクラウドソーシング. 講談社, 2016.

[2] Martin Krallinger et al. The CHEMDNER corpus of chemicals and drugs and its annotation principles. *Journal of Cheminformatics*, Vol. 7, No. 1, p. S2, Jan 2015.

[3] Martin Krallinger et al. Overview of the CHEMDNER patents task. In *Proceedings of the 5th BioCreative Challenge Evaluation Workshop*, pp. 63–75, 2015.

[4] Ling Luo, Zhihao Yang, Pei Yang, Yin Zhang, Lei Wang, Hongfei Lin, and Jian Wang. An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition. *Bioinformatics*, Vol. 34, No. 8, pp. 1381–1388, 2018.

[5] Ciyang Qing, Ulle Endriss, Raquel Fernandez, and Justin Kruger. Empirical analysis of aggregation methods for collective annotation. In *Proceedings of the 25th International Conference on Computational Linguistics, COLING 2014*, pp. 1533–1542, 2014.

[6] Nihar Shah and Dengyong Zhou. No oops, you won't do it again: Mechanisms for self-correction in crowdsourcing. In *Proceedings of The 33rd International Conference on Machine Learning*, pp. 1–10, 2016.

[7] Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. brat: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 102–107, 2012.

[8] 山村崇, 嶋田和孝. Kyutech コーパスにおけるアノテーションツールの試作. 情報処理学会 九州支部 若手の会セミナー, pp. 31–35, 2018.

[9] Jie Yang, Yue Zhang, Linwei Li, and Xingxuan Li. YEDDA: A lightweight collaborative text span annotation tool. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pp. 31–36, 2018.

[10] 吉川和, 他. コーパス作成における専門性を考慮した作業者割当ての提案と化学分野での評価. 言語処理学会第 25 回年次大会, 2019.

[7] http://www.pluto.ai.kyutech.ac.jp/~shimada/resources.html