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

吉川 和[1,2]　金子 貴美[1,2]　岩倉 友哉[1,2]　吉田 宏章[1]　熊野 康孝[3]
嶋田 和孝[2,4]　ジェプカ ラファウ[2,5]　シフィエチコフスカ パトリツィア[5]

[1] 株式会社富士通研究所　[2] 理研 AIP-富士通連携センター　[3] 株式会社ジー・サーチ
[4] 九州工業大学大学院 情報工学研究院　[5] 北海道大学大学院 情報科学研究科

{y.hiyori, kaneko.kimi, iwakura.tomoya, yoshida.hiro-15,
kumano.yasutaka}@fujitsu.com,
shimada@pluto.ai.kyutech.ac.jp, {rzepka, swieczkowska}@ist.hokudai.ac.jp

1　はじめに

近年の計算機環境の進歩により，機械学習による自然言語処理が広がりを見せている．機械学習のうち，教師あり学習手法では，人手でアノテーションした大規模なコーパスを用いることで，精度改善が期待される．しかしながら，コーパスの作成においては人手作業に頼る部分が大きく，大規模かつ高品質なコーパスの作成が課題となる．コーパス構築を効率化するための一つの方法として，クラウドソーシングの利用があげられる．クラウドソーシングを用いることで，人名や組織名などを対象とした固有表現抽出 [2, 6]，含意関係や語義曖昧性解消 [9] といった一般的な知識で判断可能なタスクおよび，文法的な知識を要する品詞タグ付け [3] などで，低コストでアノテーションを行なえることが示されている．

このように従来のクラウドソーシングの多くが一般的な知識で解けるタスクを対象としてきたのに対し，医学や化学といった専門知識が必要な分野では，専門性を持った作業者を確保する必要がある．Nye ら [7] は，非専門家の作業者と専門家の作業者を組合せて，医療分野の文書のアノテーションを実施している．

Nye らの方法では専門家の作業量が軽減できる可能性が示唆されたが，大規模な専門分野コーパス構築においては，詳細化された専門分野に対する問題が残る．たとえば，化学分野で無機化学の専門家が創薬の文書にアノテーションをする場合や，人工知能分野で自然言語処理の専門家が画像処理の文書にアノテーションをする場合などは，専門外の分野を対象とすることになる．特に，化学や医療といった専門分野のアノテーションでは，細分化された分野全てに対し十分な作業者の確保が保証できないという課題がある．

本論文では，作業者の専門性を考慮したアノテーション対象文書の割当てに基づくコーパス作成手法を提案する．従来のクラウドソーシングを対象とした研究では，複数のアノテーション結果の統合方法 [6]，複数のアノテーションを考慮した学習方法 [8]，専門家の発見 [4] といった研究が行なわれてきた．これらに対し，本論文は，専門性が一致する作業者が確保できない際を想定した割当手法に関するものであり，専門と異なっていても，作業者に近い分野の文献を割当てることができれば，比較的良い品質の作業が期待されるという考えに基づく．本論文では，作業者が読んだ論文から見積もった専門性を基に，アノテーション対象を割当てを行なう．

本手法の評価のため，化学を専攻する学生を中心に作業者を集めて実験を行なった．日英の化学文献へのアノテーションの結果から，提案手法による割当てが，質の良いコーパス構築に貢献することが示唆された．

2　アノテーション対象

今回対象とするタスクは，化学文献への化学用語のアノテーションである．以下に示す日英の化学文献各約 500 件のタイトルおよびアブストラクトにアノテーションを行った．

2.1　英語

英語文献として，CHEMDNER コーパス [1] を利用した．CHEMDNER は化学分野の文献のアブストラクト 10,000 件に対し，7 種類の化学用語のクラスを付与したものである．今回のアノテーション対象として

表 1: 専門分野とアノテーション対象のカテゴリ.

カテゴリ	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
英語	✓		✓(※)	✓(※)	✓	✓	✓	✓	✓	✓	✓
日本語	✓	✓	✓	✓	✓	✓	✓	✓			

カテゴリ: (1) 物理化学, (2) 分析化学, (3) 無機化学, (4) 錯体化学, (5) 有機化学, (6) 高分子化学, (7) 医療, (8) 創薬, (9) 生化学, (10) 応用化学, (11) 毒性学

(※) 英語の (3) (4) はジャーナル名による区別ができなかったため, 統合して 1 カテゴリとして扱う.

は, CHEMDNER のテストセット 3000 件から 504 件を選択した. 論文の選択においては, まず各論文に, ジャーナル名をもとに表 1 に記載の 9 カテゴリを重複を許して付与し, 各カテゴリの件数がおよそ均等になるように選択を行なった.

今回のアノテーションの定義も, CHEMDNER に従った.[1] 評価用の gold standard data としては, CHEMDNER のテストセットのアノテーションを利用した.

2.2 日本語

日本語文献として, JDREAM III Ⅲ[2] から論文を取得して利用した. 表 1 に記載の 8 カテゴリから, 2017 年 9 月時点での最新各 2,500 件, 計 20,000 件を取得した. その後, 簡単なフィルタリングにより, アブストラクトに化合物名の含まれていそうな文献に絞り, カテゴリの内訳が均等になるように, 計 520 件を選択した.

日本語のアノテーションの定義は, 筆者らが独自に定義したものを利用した. 付与するタグは 21 種類で, 化合物名のほか, 医薬品名や物性を表す表現にもタグを付与する.

評価用の gold standard data としては, 専門家に依頼し作成したものを用いる.[3]

3 アノテーション作業者

作業者として, 化学を専攻する学生 49 名に作業を依頼した. このうち, 英語と日本語のアノテーションを行ったのはそれぞれ 40 名, 20 名である.[4] 各作業者には, 専門性を見積もるために, 事前に下記の情報を申告してもらった.

- 専門分野（表 1 のカテゴリから選択. 複数選択）
- 最近読んだ論文 5 件のタイトル・アブストラクト

申告された専門性の分布を表 2 に示す.

また, 評価データ構築のために, 化学文献へのアノテーション経験のある専門家を 17 名雇用した.

4 作業者割当て手法

表 1 に示されるように, 化学には細分化された専門分野があり, 取り扱う化合物の種類も分野により異なる. このような細分化した専門性がある領域におけるアノテーションのための作業割当て手法を提案する.

4.1 専門性の見積り

アノテーション対象の割当てにおいては, 専門分野と提出論文から見積もられる専門性を用いる.

専門分野 表 2 にある作業者から申告された専門分野は, 作業対象の論文（以下, タスク論文）と直接紐づけることができる有用な情報である. 今回試行する全ての割当て手法において, 作業者 i のタスク論文 j に対する専門性 $I_{\mathrm{cat}}(i,j)$ を考慮する. $I_{\mathrm{cat}}(i,j)$ は作業者 i の申告した専門分野とタスク論文 j のカテゴリに重複があれば 1, 無ければ 0 とする.

提出論文 専門分野の情報は有用だが, 化学のように分野が細分化されていると, 各分野に対し専門性を持つ作業者が十分に確保できない場合がある. そこで, 分野をまたぐ作業者の適性を見積もるため, 作業者から提出された論文（以下, 提出論文）と作業対象の論文の意味的類似度を利用する.

[1] 実際には, CHEMDNER ガイドラインが入手できなかったため, その後開催された特許向けタスク CHEMDNER patent のガイドラインを代替として利用した. こちらも CHEMDNER ガイドラインに基づいてるものの, 一部改変が行なわれている. しかしながら, 予備調査の結果, それらの改変が今回の評価には大きな影響がないと判断して進めた.

[2] https://jdream3.com/

[3] 評価用 gold standard data は, まず, 各アブストラクトにつき 2 名でアノテーションを行なった後, 二つの結果を基に, 最終的なアノテーションを決定する形で作成した.

[4] 日英両方をアノテーションした作業者もいるため, 作業者数の合計と作業依頼者総数は異なる.

表 2: 各カテゴリに専門性を持つ作業者数.

カテゴリ	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	その他
英語作業者 (40 名)	9	7	6	1	14	6	4	4	10	7	4	14
日本語作業者 (20 名)	2	2	3	2	2	3	3	5	6	2	0	3

カテゴリ: (1) 物理化学, (2) 分析化学, (3) 無機化学, (4) 錯体化学, (5) 有機化学, (6) 高分子化学, (7) 医療, (8) 創薬, (9) 生化学, (10) 応用化学, (11) 毒性学. ” その他” は (1)〜(11) 以外が専門の場合.

まず，提出論文およびタスク論文に対応する分散表現を計算する．分散表現としては，論文のタイトルとアブストラクトに出現する内容語の単語分散表現（200 次元）をユークリッドノルムが 1 になるように正規化し，平均をとったものを用いた．英語・日本語の単語分散表現は，`word2vec` [5] を用いて学習した．学習データとして，英語は CHEMDNER 全件および MEDLINE 2017 年版 [5] から CHEMDNER で使用されているジャーナルの論文を抽出したものを，日本語は 2 章に記載の，JDREAM III から取得した論文計 20,000 件を用いた.

4.2 作業対象に対する適性スコアの計算

作業者 i のタスク論文 j への適性スコア $s_{i,j}$ として，ベースラインの申告された専門分野のみを考慮する「cat」に加えて, 提案手法である提出論文とタスク論文の類似度に基づく「catsim-avg」と「catsim-nearest」を用いる.

(`cat`) $I_{\mathrm{cat}}(i,j)$.

(`catsim-avg`) $I_{\mathrm{cat}}(i,j) + \mathrm{sim}(\frac{1}{L}\sum_{l=1}^{L} d_{il}^{(\mathrm{s})}, d_j^{(\mathrm{t})})$

(`catsim-nearest`) $I_{\mathrm{cat}}(i,j) + \max_l \mathrm{sim}(d_{il}^{(\mathrm{s})}, d_j^{(\mathrm{t})})$

$d_{il}^{(s)} (l = 1, \ldots, L)$ および $d_j^{(t)}$ はそれぞれ, 提出論文とタスク論文に対応する分散表現を表す．今回は各作業者から 5 件ずつ論文が提出されたので, $L = 5$ である．$\mathrm{sim}(\cdot,\cdot)$ はベクトルのコサイン類似度を表す．すなわち, (`catsim-avg`) では提出論文の分散表現の平均をとったものとタスク論文の類似度を, (`catsim-nearest`) では提出論文とタスク論文の最大類似度を計算する.

4.3 割当ての作成

各作業者の専門性を見積った後に，タスク論文への作業者の割当てを行なう．割当ては整数線形計画問題（ILP）として以下のように定式化する:

$$
\begin{aligned}
\text{Maximize} \quad & \sum_{i\in I, j\in J} s_{i,j}\cdot x_{i,j} \\
\text{s.t.} \quad & k \le \sum_{j\in J} x_{i,j} \le k+1 \quad \forall i \in I, \\
& \sum_{i\in I} x_{i,j} = 1 \quad \forall j \in J, \\
& x_{i,j} \in \{0,1\} \quad \forall i \in I, \forall j \in J.
\end{aligned}
$$

ここで「Maximize $\sum_{i\in I, j\in J} s_{i,j}\cdot x_{i,j}$」が目的関数であり，$x_{i,j}$ について最適化することで, 専門性を考慮した適正スコアの合計値が最大になるような割当を見つけるという意味となる．$s_{i,j}$ は, 4.1 節に示される作業者 i のタスク j に対する適性スコアである．$x_{i,j}$ は割当結果を示す変数で，作業者 i がタスク j に割当てられた場合には 1, そうでなければ 0 である．I は作業者の集合, J はタスク論文の集合である．k は作業者 1 人あたりに割当てるべきタスク数を示し，$k = \lfloor |J|/|I| \rfloor$ である．全ての作業者に均等にタスクを割当てるためにこのように設定した．英語では，$|J| = 504$, $|I| = 40$, 日本語では，$|J| = 520$, $|I| = 20$ となる.

「$k \le \sum_{j\in J} x_{i,j} \le k+1 \quad \forall i \in I$」が一作業者あたりの論文割当件数の制約にあたり，「$\sum_{i\in I} x_{i,j} = 1 \quad \forall j \in J$」が，一つの論文につき，一作業者を割当てるための制約となる.

ILP の求解は PuLP [6] で COIN CBC ソルバを用いて行い，いずれも最適解を得た.

5 評価実験

5.1 作業概要

本実験の目的は，割当て手法によりコーパスの質に違いが生じるかを検証することである.

作業に不慣れなことに起因する作業品質の変化を軽減するため, 作業者には, タスク論文に対するアノテーション作業に入る前に，アノテーションの定義の理解

[5] https://www.nlm.nih.gov/bsd/medline.html

[6] https://github.com/coin-or/pulp

表 3: 実験結果．gold standard data を基にした精度．

英語			
割当て手法	Recall	Precision	F_1
cat	0.525	0.481	0.502
catsim-avg	0.550	0.496	0.522
catsim-nearest	**0.569**	**0.507**	**0.536**
日本語			
cat	0.455	**0.510**	0.481
catsim-avg	0.458	0.502	0.479
catsim-nearest	**0.465**	**0.510**	**0.486**

の確認を兼ねて 5 件の例題について作業を行ってもらった．

その後，各作業者には，割当てられた論文について作業を行ってもらった．実験では，比較対象の 3 種類の手法を用いた 3 種類の割当てをそれぞれ全作業者を対象に作成し，これらの割当て結果をマージしたものを各作業者に割当てた．割当て結果間には重複があるため，作業者によって総作業量にはばらつきがある．作業順はランダムにし，各タスク論文がどの割当てに起因するものかは作業者には知らせていない．

アノテーション作業には BRAT[10] を用いた．

5.2 実験結果

割当て後に一部作業キャンセルが生じたため，全ての割当て手法について作業結果が出そろったのは英語 323 件，日本語 375 件となった．以下ではこれらの文献を対象に，作業品質の確認を行った．

アノテーション品質の評価には，gold standard data に対する recall, precision, F-measure (F_1) を用いる．アノテーションの開始位置・終了位置・タグ種別全てが一致した場合のみ正解と判断し，タグ種別違いやアノテーション範囲のずれは全て誤りとみなした．

表 3 に実験結果を載せる．ベースライン (`cat`) と比較し，`catsim-nearset` が日英ともに高い F_1 値を示している．`catsim-avg` については，英語では改善がみられるが，日本語では precision が悪化しており改善がみられなかった．

ベースラインと提案手法によるアノテーションの差を確認するため，文字ベースの BIO タグに基づく McNemar 検定を行ったところ，日英ともに，いずれの提案手法もベースラインと有意差（$p < 0.01$）がみられた．

6 まとめ

本論文では，作業者の専門性を考慮したタスク割当てに基づくコーパス作成手法を提案した．日英の化学文献を用いた評価から，本提案手法による専門性の考慮により，アノテーションの精度改善が行なえる可能性が示唆された．

今後は，誤りの傾向と割当て手法の関係をより具体的に検証し，高品質な専門分野コーパスの構築に向けてさらなる検討をすすめる．

参考文献

[1] Martin Krallinger et al. The chemdner corpus of chemicals and drugs and its annotation principles. *Journal of Cheminformatics*, Vol. 7, p. S2, 2015.

[2] Tim Finin, Will Murnane, Anand Karandikar, Nicholas Keller, Justin Martineau, and Mark Dredze. Annotating named entities in twitter data with crowdsourcing. In *Proc. of NAACL HLT 2010 Workshop on CSLDAMT '10*, pp. 80–88, 2010.

[3] Dirk Hovy, Barbara Plank, and Anders Søgaard. Experiments with crowdsourced re-annotation of a pos tagging data set. In *Proc. of ACL'14 (Volume 2: Short Papers)*, pp. 377–382, 2014.

[4] Hongwei Li, Bo Zhao, and Ariel Fuxman. The wisdom of minority: Discovering and targeting the right group of workers for crowdsourcing. In *Proc. of WWW'14*, pp. 165–176, 2014.

[5] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *CoRR*, Vol. abs/1301.3781, , 2013.

[6] An Thanh Nguyen, Byron Wallace, Junyi Jessy Li, Ani Nenkova, and Matthew Lease. Aggregating and predicting sequence labels from crowd annotations. In *Proc. of ACL'17 (Volume 1: Long Papers)*, pp. 299–309, 2017.

[7] Benjamin Nye, Junyi Jessy Li, Roma Patel, Yinfei Yang, Iain Marshall, Ani Nenkova, and Byron Wallace. A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. In *Proc. of ACL'18 (Volume 1: Long Papers)*, pp. 197–207, 2018.

[8] Filipe Rodrigues, Francisco Pereira, and Bernardete Ribeiro. Sequence labeling with multiple annotators. *Machine Learning*, Vol. 95, No. 2, pp. 165–181, May 2014.

[9] 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 *Proc. of EMNLP*, pp. 254–263, 2008.

[10] 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 *Proc. of EACL*, pp. 102–107, 2012.