Vol.2021-IFAT-143 No.1
2021/7/30

Next Sentence Predictionを応用した学術文献著者同定

金沢 輝一[1]　蔵川 圭[2]　安達 淳[1]

概要：学術論文の著者同定は，学術情報サービスにおけるコンテンツへの到達性を改善するだけでなく，研究力分析を支える基礎情報の整備においても重要な役割を担っている．本稿は論文と研究者の研究トピックの類似性を推定し，これに基づいて著者を同定する手法を提案する．従来，著者の所属機関情報のない書誌情報や単著論文では著者同定の精度を安定して得ることは困難とされてきた．本稿では，5 万例以上の論文著者を提案手法で同定する評価実験を実施して，著者氏名と論文のタイトルのみを用いて和英いずれにおいても F 値 0.75〜0.76 の同定精度が得られることを検証した．

Academic Article Author Identification using the Next Sentence Prediction

KANAZAWA Teruhito[1]　KURAKAWA Kei[2]　ADACHI Jun[1]

Abstract: Identifying authors of academic articles is an important process in developing research intelligence as well as in improving the accessibility of academic information services. In this paper, we propose a method of identifying authors based on the similarity of the subject between the article and the researches of candidate researchers. It had been considered difficult to obtain stable accuracy in identifying single authors and authors without affiliation information. However, we concluded that the proposed method using only author names and article titles achieved F-measure values of 0.75 to 0.76 in our experiment to identify more than 50,000 authors, regardless of Japanese or English the titles are written in.

1. はじめに

学術コンテンツサービスにおいて，研究者や論文著者の情報同定の重要性が増している．ORCID[*1]等の研究者識別子の活用が促進されているものの，自動処理あるいは人の手によって人物同定が行われる割合は依然として大きい．日本語では人物同定を「名寄せ」とも表現するが，氏名は識別子としての一意性を持たないので，実際の人物同定は他の手がかりも参照しながら行われる．例えば論文著者の同定では所属機関，共著者等が用いられる．

しかしこれらの情報は常に参照可能とは限らない．著者の所属機関の情報が書誌情報から取得できない場合や単著論文の場合は，氏名だけを手掛かりとして同定を行うため低精度となりがちである．

我々は，論文の書誌情報としてほぼ必ず取得可能な著者氏名と論文のタイトルのみを用いて，実用的な同定精度を得る手法の実現を試みた．本稿では BERT の次文推定を応用して論文と研究者の研究トピックの類似性を推定し，これに基づいて著者を同定する手法を提案する．本稿のリサーチクエスチョンは以下の通りである．

RQ1: 最小限の情報のみを用いて，学術論文の著者同定においてどの程度の精度を得られるか．

RQ2: 言語によらず安定した同定精度が得られるか．

これらに対する答えを導き出すため，和英の論文の著者を同定するタスクを設定して，提案手法の評価を行う．

[1] 国立情報学研究所
National Instutite of Informatics, Chiyoda-ku, Tokyo 101–8430 Japan
[2] 2021 年 3 月まで [1]．4 月より大学改革支援・学位授与機構
National Institution for Academic Degrees and Quality Enhancement of Higher Education, Kodaira-shi, Tokyo 187–8587 Japan
[*1] https://orcid.org/

2. 関連研究

2.1 人物の同定

論文の著者を特定する人物同定の手法は，候補者集合や判断に用いる情報源等の特性に応じたものが各種提案されている．

既知の著作物を持つ候補者の中から，判定対象の文章自身を判断材料として著者を特定する形式の人物同定は著者推定 (Authorship Attribution) と呼ばれている．PAN[*2] という組織が 2011 年から継続的に著者推定の評価タスクを整備しており [1]，現時点で最も新しい 2019 年のタスクに挑戦した中では tf-idf によってテキストを特徴量ベクトルに変換し，SVM で同定を行う手法が一般的で，特に SVD による次元圧縮を取り入れた Muttenthaler らの手法 [2] が最も高い F 値を示した．

一方で，文章の内容以外の手がかりから人物を同定する場合もある．本稿が想定するのはこちらの形式である．山田らは氏名，共著者，所属機関名，メールアドレス，研究内容情報，引用・被引用関係情報等を総合的に用いた論文著者等の同定処理手法を提案している [3]．単著論文の場合や，異動によって所属機関と共著者の両方が変わってしまう場合等において適切に同定できないケースが存在すると報告されている．この課題の克服には，論文の書誌情報としてほぼ必ず取得可能な著者氏名と論文のタイトルのみで安定した精度を得る手法の検討が効果を持つ．

2.2 話題の類似性判断

藤野らは論文の概要をトピックモデルで分析し，論文の著者が特定の組織の研究者であることを判断する手法を提案している [4]．この手法では判定対象の組織に応じて最適なトピック数を事前学習に基づいて設定する必要があり，また多クラス分類問題に最適化されていないため，同様の仕組みを著者個人の同定に単純には適用できない．

桂井らは論文のタイトルを構成する語彙から研究分野を推定して，同名の研究者の中から著者を同定する手法を提案している [5]．論文のタイトルだけを手掛かりとして，所属，共著者および発行年を組み合わせた同定手法より高い同定精度を得られることが評価実験によって示されている．ただし候補者の中に必ず正解が含まれていることを仮定した評価であり，候補者に正解を含まない状況での同定精度は未検証である．

3. 提案手法

3.1 提案手法の概要

本研究では，論文の著者に対して研究者リストに収録されている研究者のいずれであるか，あるいは収録されていない者であるかを推定する著者同定タスクを対象とする．研究者リストに多重登録者は存在しないものとみなす．よって 1 人の著者に対する正解となる研究者は最大 1 人である．研究者リストは各研究者について氏名のほか，携わった研究課題のタイトル一覧を記載しているものとする．

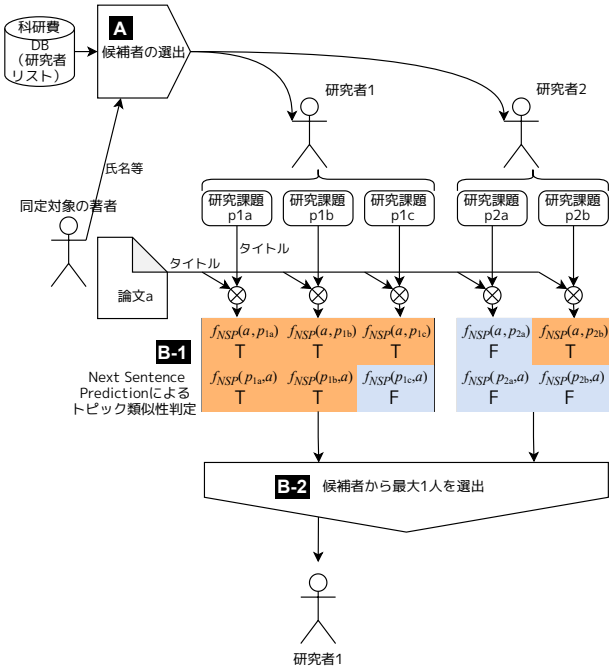

図 1 提案手法のデータフロー

提案手法は (1) 研究者リストから候補者を選出する工程（**図 1** の A）と (2) 候補者から最大 1 人を選出する工程（同 B-1，B-2）で構成する．以下に各工程の詳細を述べる．

3.2 工程 1: 候補研究者の選出

前半の工程では，氏名の同一性または類似性に基づいて以下の基準のいずれかに該当する研究者を候補者として選出する．

(1) 論文著者と漢字表記のフルネームが同一の研究者がいれば，それらを候補者として抽出する．
(2) 前項の基準に該当者がない場合，論文著者とローマ字表記のフルネームが同一の研究者がいれば，それらを候補者として抽出する．
(3) 前項までの基準に該当者がない場合，論文著者名がファーストネームをイニシャルにした表記であり，フルネームを同様の表記に変換して一致する研究者がいれば，それらを候補者として抽出する．
(4) 前項までの基準に該当者がない場合，論文著者名が姓のみの表記であれば，姓が一致する研究者を候補者として抽出する．

3.3 工程 2: 候補者からの選出

後半の工程では，次文推定 (Next Sentence Prediction)

[*2] https://pan.webis.de/

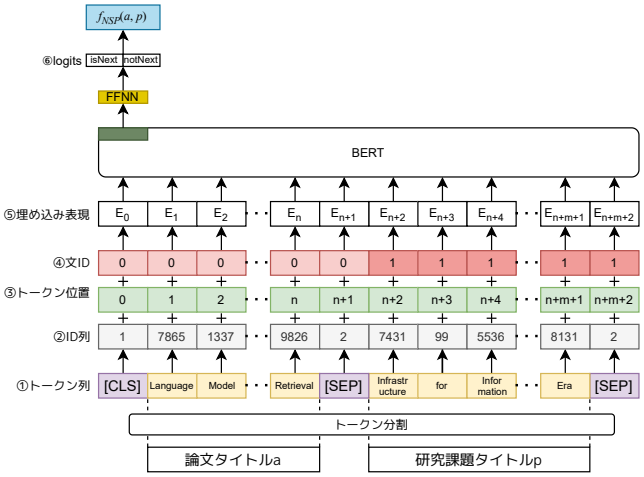

図 2 BERT による次文推定

を用いて研究課題と論文との類似性を判定し，候補者の中から最大 1 人を選出する．ここでは，類似のトピックを共有する研究課題と論文のタイトルならば，連続する 2 文として解釈できるという仮定を置いている．

提案手法では，次文推定が事前学習のタスクに組み込まれている BERT[6] によって次文推定を行うこととした．BERT の次文推定は**図 2** の流れで行われる．① 推定対象の 2 文がそれぞれトークン分割された後，② 先頭と文間，そして末尾に区切りトークンを加えて ID 列に変換される．このトークン ID と，③ 列中での位置と，④ どちらの文に属するかを示す識別子の 3 要素からなるベクトルを構成し，これが ⑤ 埋め込み表現に変換されて BERT の Transformer 構造の入力となる．最終隠れ層の出力から順伝播型ニューラルネットワーク (Feedforward NN; FFNN) によって ⑥ 研究課題のタイトルが論文のタイトルの次文でありうる尤度 isNext と，そうでない尤度 notNext の 2 つの logit が得られる．文 y が文 x の直後に出現する文でありうるかどうかの真偽値 $f_{\mathrm{NSP}}(x,y)$ は 2 つの logit によって以下のように定義される．

$$f_{\mathrm{NSP}}(x,y) = \begin{cases} \text{true} & \text{isNext} > \text{notNext} \\ \text{false} & \text{isNext} \leq \text{notNext} \end{cases}$$

この次文推定を用いて，候補者が N 件の研究課題に参画している場合に，当該候補者が参画した研究課題群と論文との類似性を示す $2N$ 次元の可変長ベクトル，

$$(f_{\mathrm{NSP}}(a,p_1), f_{\mathrm{NSP}}(a,p_2), \ldots, f_{\mathrm{NSP}}(a,p_N),$$
$$f_{\mathrm{NSP}}(p_1,a), f_{\mathrm{NSP}}(p_2,a), \ldots, f_{\mathrm{NSP}}(p_N,a))$$

が得られる．ただし a は論文のタイトル，p は候補者の研究課題のタイトルである．次にこの類似性ベクトルを，論文 a のトピックが研究課題群全体と類似しているといえるかどうかの単独の論理値に変換し，最終的に候補者の中から最大 1 人を選出する．変換と選出の手順は，実データを用いた評価の結果に基づいて設定することとする．

表 1 評価に用いたデータセット

判定基準	著者数（延べ）	異なり研究者数	同一研究者の最多著者カウント
(1)	49,879	22,537	50
(2)	362	230	8
(3)	1,206	785	10
合計	51,447	23,166	50

4. 実験

4.1 データセット

提案手法による論文著者の同定には，研究者の研究トピックを抽出するための情報源と，同定対象の論文のトピックを抽出するための情報源が必要である．また，精度を測定するためには，論文著者と研究者の正しい同定情報も必要である．ここでは，学術情報データベース CiNii Articles[*3]に収録されている記事の著者の科研費研究者番号を以下の基準で推定したデータセットを構築して用いた．科研費の研究課題に関する情報は KAKEN[*4]から取得した．

(1) DOI[*5]が紐づいている CiNii Articles の論文を抽出する．当該 DOI が紐づいた成果物を有する科研費研究課題があれば，その参画研究者と CiNii Articles 論文の著者をフルネームで比較する．一致する参画研究者が 1 人だけならば著者とみなす．複数いる場合は絞り込みは行わずデータセットに含めない．

(2) 前項とほぼ同様の条件で，著者名がイニシャル表記で記載されている場合は，科研費研究課題の参画研究者名をイニシャル表記に変換したものと比較する．ただし同じ論文の別の著者として既に前項の条件に適合した研究者は除外する．一致する参画研究者が 1 人だけならば著者とみなす．

(3) 前項までとほぼ同様の条件で，姓のみを比較する．ただし同じ論文の別の著者として既に 1, 2 の条件に適合した研究者は除外する．一致する参画研究者が 1 人だけならば著者とみなす．

これらの基準で科研費研究者番号が推定できた著者数を**表 1** に示す．

4.2 論文と研究課題のトピック類似性判断基準

4.2.1 トピック類似性判断基準の評価

本節では，工程 2 で候補者から 1 人を選出する際に用いる，類似性ベクトルを単独の真偽値に変換する手順を決定する．**表 2** に示した 4 つの条件は，変換手順の候補である．この中から，変換後の真偽値が著者本人であるか否かと強い相関を持つものを選ぶこととする．

*3 https://ci.nii.ac.jp/
*4 https://kaken.nii.ac.jp/
*5 https://www.doi.org/

表 2 トピック類似性判定条件の性能

番号	定義（連続文として成り立つ条件）	数式表現	R(再現率)	P(適合率)
1	論文タイトルと全ての研究課題名がどちらを先頭にしても成立	$\forall p \in \mathrm{Project}(r)\ f_{\mathrm{NSP}}(a,p) \wedge f_{\mathrm{NSP}}(p,a)$	.1192	.9989
2	論文タイトルと全ての研究課題名がいずれかの順序で成立	$\forall p \in \mathrm{Project}(r)\ f_{\mathrm{NSP}}(a,p) \vee f_{\mathrm{NSP}}(p,a)$	.2367	.9989
3	論文タイトルと 1 つ以上の研究課題名がどちらを先頭にしても成立	$\exists p \in \mathrm{Project}(r)\ f_{\mathrm{NSP}}(a,p) \wedge f_{\mathrm{NSP}}(p,a)$	.3461	**.9991**
4	論文タイトルと 1 つ以上の研究課題名がいずれかの順序で成立	$\exists p \in \mathrm{Project}(r)\ f_{\mathrm{NSP}}(a,p) \vee f_{\mathrm{NSP}}(p,a)$	**.3644**	.9989

精度評価には，変換後の真偽値（推測値）と著者本人か否かの真偽値（実値）との混同行列から得られる以下の 2 つの指標を用いる．

$$(\text{再現率 } R) = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \tag{1}$$

$$(\text{適合率 } P) = \frac{\mathrm{TP}}{\mathrm{TP+FP}} \tag{2}$$

ただし TP は著者本人が推測真値を得た人数，FP は他研究者が推測真値を得た人数，FN は著者本人が推測偽値を得た人数である．

4.2.2 評価結果と判断基準の決定

4.1 節で構築したデータセットから 3.2 節の手順で選出した各候補者について求めた単独の真偽値が，著者本人に対して真値であった場合を真陽性として，表 2 の 4 つの条件を評価した結果を**表 2** に示す．

今回評価した 4 つの中には，再現率と適合率がともに最高となる単一の条件は存在しなかったが，いずれも同程度に高い水準の適合率であった．そこで候補者から最大 1 人を選出する判断基準は，再現率の大きい条件 3 と 4 を組み合わせて次のように定義する．これによって正しい同定結果が多く得られると期待できる．

(1) 論文タイトルと 1 つ以上の研究課題名がいずれかの順序で連続文として成り立つ（条件 4 を満たす）候補が 1 人だけいれば，当該研究者が著者だと推定する．
(2) 1 の該当者がなく，論文タイトルと 1 つ以上の研究課題名がどちらを先頭にしても連続文として成り立つ（条件 3 を満たす）候補が 1 人だけいれば，当該研究者が著者だと推定する．

4.3 著者の同定

4.3.1 言語モデル

BERT の言語モデルとして日本語には CiNiiBERT[7] を，英語には SciBERT[8] を使用した．CiNiiBERT は CiNii Articles の日本語で記述された記事概要から言語モデルを学習したものである．4.1 節で構築したデータセットと使用した記事レコードが一部重複しているが，CiNiiBERT と本稿の同定アルゴリズムでは次文推定の対象が異なり，前者は記事概要同士，後者は記事のタイトルと科研費研究課題のタイトルであるため，不適切な学習には該当しないと判断した．

4.3.2 評価指標

評価方法は，4.1 節で構築したデータセットの各論文著者について科研費研究者番号を推定する精度を計測するというものである．評価対象のシステムは各論文著者に対して最大 1 つの研究者番号を回答することができ，確信を持った回答が困難な場合には無回答を許容する．

同定精度はデータセットの全論文著者の回答の正誤に基づいて正答数を TP，誤答数を FP，無回答数を FN とした場合の式 (1), (2) からそれぞれ再現率，適合率を，また以下の式から F 値を求めた．

$$(\text{F 値}) = \frac{2RP}{R+P}$$

4.3.3 比較対象

評価実験では以下の手法を比較対象とした．

(1) **候補者からランダムに 1 人選んで回答する．**提案手法と同じ基準で選出した候補者からランダムに 1 人に絞り込む．科研費研究者と同名の別人の論文で誤判定が生じて適合率が低下すると予想される．正解の研究者を含んでいるテストケースで正解の研究者が選出される確率の期待値は候補者数を $|C|$ とすると $\frac{1}{|C|}$ であることを利用して，実際の試行はせず期待値を計算した．
(2) **候補者の唯一性に基づく同定．**$|C| = 1$ の場合だけ回答し，$|C| > 1$ は全て無回答とする．この判断基準ではランダムに選んだ場合の適合率低下を回避できる一方で，複数の候補者がいる場合は全て無回答となるため再現率は低下すると予想される．
(3) **CiNii Articles の著者同定．**CiNii Articles は，相澤の手法 [9] を応用して，氏名に加えて論文の共著関係や所属等の情報を用いたクラスタリングと，科研費研究者番号に基づくクラスタリング矛盾の解消によって著者同定を行っている [10, 11]．ただし CiNii Articles の公開メタデータから抽出した同定結果は，適切にチューニングされた状態か疑わしい結果であったため，参考情報とした．

[3] では著者をユニークな研究者 ID と同定するという，本稿と同種の課題設定の下に評価を行って，再現率 0.9079，適合率 0.9863 という結果を得ている．しかし評価データは 1 人当たり平均 430 件の論文に紐づいている 12 名の研究者とのことであり，本稿で実施した評価とは条件が大きく異なっている．精度指標を直接比較しても適切な考察が

表 3 評価結果

同定アルゴリズム	タイトルの記述言語	FN(無回答数)	TP(正答数)	FP(誤答数)	R(再現率)	P(適合率)	F 値
提案手法	日本語	17,967	32,435	1,045	.6305	.9688	**.7638**
	英語	18,847	31,715	885	.6165	**.9729**	.7547
候補者からランダム選択	–	**11,647**	**33,745**	6,055	**.6559**	.8155	.7271
候補者の唯一性	–	20,155	30,297	995	.5889	.9682	.7324
CiNii Articles （参考）	–	49,878	1,519	**50**	.0295	.9681	.0573

表 4 提案手法の工程ごとの性能

提案手法内の工程	タイトルの記述言語	候補数・回答数	TP(正答数)	FP(誤答数)	R(再現率)	P(適合率)	F 値
工程 1: 候補者選出	–	71,720	41,377	30,343	.8043	(.5769)	(.6719)
工程 2:	日本語	33,480(.4668)	32,435(.7839)	1,045(.0344)	.6305	.9688	.7638
候補者から 1 人を選出	英語	32,600(.4545)	31,715(.7665)	885(.0292)	.6165	.9729	.7547

導かれないため，比較対象には加えないこととした．

4.3.4 評価結果

評価結果を**表 3** に示す．

提案手法はトピックの類似判定を日本語で行った場合に再現率，適合率，F 値がそれぞれ 0.6305, 0.9688, 0.7638 で，英語の場合それぞれ 0.6165, 0.9729, 0.7547 であった．

候補者からランダムに選択した場合は，予想通り提案手法よりも適合率が 15 パーセントポイント以上劣っていた．

唯一の候補者のみを回答した場合は，予想通り提案手法よりも再現率が約 2〜4 パーセントポイント劣っていた．

CiNii Articles の適合率は提案手法と同程度だが，再現率は 0.0295 と非常に低く，提案手法の 20 分の 1 以下にとどまっている．

提案手法の工程ごとの評価指標を**表 4** にまとめた．工程（1）はできる限り正解研究者を候補から漏らさないよう，1.0 に近い高い再現率であることが望ましいが，0.8 程度であった．これは評価用のデータセットを作成するための 4.1 節の手順で同定できた研究者が，3.2 節の手順による候補者の選出では漏れたことを意味する．例えば論文の書誌情報に「鈴木一『朗』」と記載されている著者が研究者リストでは「鈴木一『郎』」だった場合に，当該論文を成果物とする科研費研究課題に鈴木姓の研究者が鈴木一郎しか参画していなければデータセット作成手順では同一人物とみなされるが，提案手法の手順では同姓の研究者を候補に含めていないので漏れてしまう．候補に含めていないのは，全ての鈴木姓の研究者からトピックの類似性のみで 1 人に絞り込むことになれば精度維持が困難なためである．

次に，言語による性能の違いを考察するためにトピックの類似性判定に使用したタイトルの言語ごとに性能指標を算出する．候補者数 $|C|$ で分類した部分集合ごとの適合率を日本語と英語それぞれで算出した結果を**図 3** に，同じく再現率を**図 4** に示す．$|C|$ が大きくなるにつれてテストケースが減少して統計的誤差が大きくなっており，$|C| \geq 11$ は一つの部分集合にまとめている．日本語では，$|C|$ によらず適合率がほぼ一定の性能を維持しつつ，再現率は $|C| = 1$ の場合に 0.9269 という高い値だが $|C| = 2$ では 0.4097 に低下し，それ以降もほぼ単調に低下していることが分かる．これは判断基準を設定した際の期待通りの挙動である．一方英語では候補者数の増加と共に適合率が低下してしまっている．再現率の低下が日本語よりも著しく，$|C| = 6$ の時に 0.0114 である（日本語では 0.1966）．$|C|$ が大きくなるにつれて，正解の研究者が候補に含まれているのに無回答とするテストケースが増え，相対的に不正解の割合が高まっていることが図 4 から読み取れる．

5. 考察

RQ1:「最小限の情報のみを用いて，学術論文の著者同定においてどの程度の精度を得られるか」については，適合率で約 0.97 を達成した．[6] によると BERT の次文推定精度は 97〜98%とされており，類似タスクであるトピックの類似性判定において同程度の精度が得られたことは興味深い．今回の評価では他データで学習した言語モデルをそのまま用いて一定の性能を得たが，ファインチューニングによって精度を改善できる可能性がある．また [12, 13] 等のモデルでは，次文推定タスクによるパラメータ習得は言語モデルとしての性能向上への寄与が小さいのではないかとの疑念から，次文推定を用いないか，あるいは類似タスクで置き換えることを提案している．次文推定以外にトピックの類似性の判定が可能な機能があれば比較評価の対象となるだろう．

$|C| = 1$ に限定すればトピックの類似性で判断せずに全ての候補者を回答した方（候補者の唯一性による同定）が，ほぼ同じ適合率（それぞれ 0.9699, 0.9682）で優れた再現率を得られた（それぞれ 0.5638, 0.5889）が，これは研究者リストの研究者と紐づく著者だけで評価していることによるもので，科研費研究者番号を持たない著者が多数いる CiNii Articles 全体では同様の結果にならない．研究者リストに含まれていない著者の抽出は自動化が困難なため本稿の評価では扱わなかった．同姓同名の著者のうち一部だけが科研費研究者である場合には，候補者の唯一性に基づ

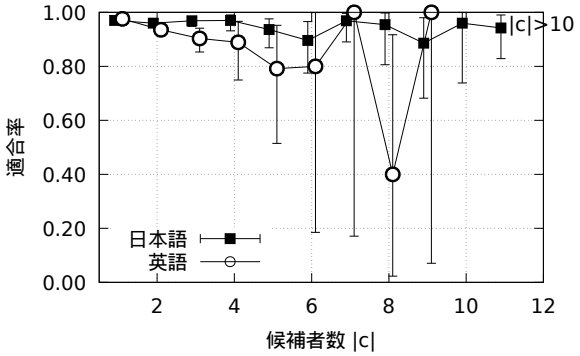

図 3 候補者数と適合率
エラーバーは符号検定有意水準 99%の範囲．各言語の結果が重なるのを避けるため横軸を ±0.1 ずらして描画している．

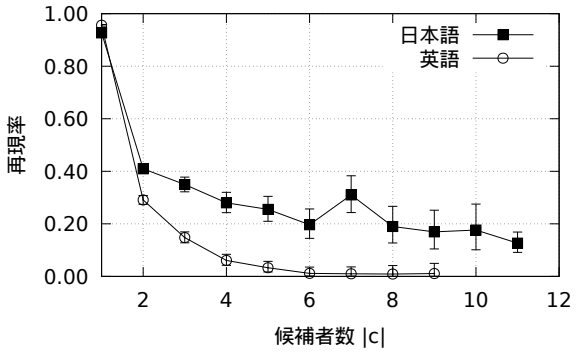

図 4 候補者数と再現率
エラーバーは符号検定有意水準 99%の範囲．

く手法では全て科研費研究者が著者という誤った同定が行われるが，提案手法はトピックの類似性に基づいて分別を行うことで誤同定を抑止できると予想される．

RQ2:「言語によらず安定した精度が得られるか」については，日本語と英語のいずれを用いても最終的には同程度の精度が得られた．しかし図 3, 4 からは特性の差異が読み取れる．$|C| = 1$ では英語のトピック類似度による判定の方が適合率・再現率ともに良好だが，候補者の増加に伴って急速に性能を悪化させている．にもかかわらず提案手法が高い精度を達成したことに関しては，候補者の選出工程で漢字かな表記とローマ字表記の両方を利用して，平均 1.4 人程度（約 5 万人の著者に対して選出された候補者の総数は約 7 万人）に絞り込めた工程（1）の寄与が大きいと言える．図 4 が示すように工程（2）の再現率は，候補者が 2 人以上になると極端に低下してしまうため，氏名の情報がローマ字表記でしか得られないデータベースでは再現率の低下が予想される．この点は例えば ORCID の研究者情報等での評価で検証が可能と思われる．

6. おわりに

本稿では，学術論文の著者同定処理を，氏名と論文および研究課題のタイトルという最小限の情報のみで行う手法を提案し，実用的な精度が得られることを評価実験で確認した．

提案手法は BERT の次文推定を応用して論文と研究者の研究トピックの類似性を推定するもので，日本語と英語で同程度の精度が得られた．また，候補者を平均 2 人以下に絞り込めていることの寄与が大きいことが確認された．

提案手法は所属機関や共著者等による同定処理と相補的に同定精度を改善できると考えられる．その場合，複数の判定器の出力の統合方法について検討していく必要がある．

高精度の著者同定は，学術情報サービスには利便性を，例えば関連論文を同著者のものとその他に分類して表示する等の形でもたらし，研究力分析には精緻さを，例えば研究業績の網羅的な把握等によってもたらす．これらの実現に向けて本研究を発展させていきたい．

参考文献

[1] Kestemont, M. et al.: Overview of the Cross-domain Authorship Attribution Task at PAN 2019, *CLEF,* (2019).
[2] Muttenthaler, L., Lucas, G., and Amann, J.: Authorship Attribution in Fan-fictional Texts given Variable Length Character and Word n-grams, *CLEF,* (2019).
[3] 山田 智之, 西 信能, 佐藤 友思, 棚橋 佳子, 渡辺 麻子, 松邑 勝治, 黒沢 努, 矢口 学: 高精度研究者人名名寄せによる効率的な研究成果情報の集積方法, 情報プロフェッショナルシンポジウム予稿集 2010, pp. 117–122, DOI: 10.11514/infopro.2010.0.117.0, (2010).
[4] 藤野 友和, 濱田 ひろか: 学術文献 DB における著者識別のためのトピックモデリングの利用とその性能比較, 特集 Institutional Research と統計科学, 統計数理, vol. 68, No. 2, pp. 209–218, (2020).
[5] Katsurai, M. and Ohmukai, I.: Author Matching across Different Academic Databases: Aggregating Simple Feature-Based Rankings, *Proceedings of the 18th Joint Conference on Digital Libraries,* JCDL '19, pp. 279–282, DOI: 10.1109/JCDL.2019.00046, (2019).
[6] Devlin, J. et al.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, arXiv: 1810.04805 [cs.CL], (2019).
[7] 壹岐 太一, 金沢 輝一, 相澤 彰子: 学術分野に特化した事前学習済み日本語言語モデルの構築, 情処研報, vol. 2020-IFAT-139, No. 4, pp. 1–6, (2020).
[8] Beltagy, I., Lo, K., and Cohan, A.: SciBERT: Pretrained Language Model for Scientific Text, *EMNLP,* arXiv: 1903.10676, (2019).
[9] 相澤 彰子: データベースとウェブの連携による情報の獲得と利用に関する研究, 『言語処理技術の深化と理論・応用の新展開』2009 年度 科研・合同シンポジウム, http://www.forest.eis.ynu.ac.jp/NLPsympo2009/files/AizawaA.pdf, (2009).
[10] 国立情報学研究所: CiNii 著者検索について, https://cinii-blog.tumblr.com/post/486298233/cinii-author-search, (2010).
[11] 相澤 彰子: 科研費研究課題「データベースとウェブの連携による情報の獲得と利用に関する研究」成果報告書, https://kaken.nii.ac.jp/file/KAKENHI-PROJECT-21300058/21300058seika.pdf, (2012).
[12] Liu, Y. et al.: RoBERTa: A Robustly Optimized BERT Pretraining Approach, arXiv: 1907.11692 [cs.CL], (2019).
[13] Lan, Z. et al.: ALBERT: A Lite BERT for Self-supervised Learning of Language Representations, *International Conference on Learning Representations,* arXiv: 1909.11942 [cs.CL], (2020).