書籍の著者名曖昧性解消における評価コーパスの自動構築

中山 祐輝　　Sudha Bhingradive　　村上 浩司
楽天株式会社 楽天技術研究所
{yuki.b.nakayama, sudha.bhingardive, koji.murakami}@rakuten.com

1 はじめに

インターネットの普及によって、ある人物に関する所望の文書をウェブ上で検索することが日常的に行われている。その際には、30%のクエリが人名を含む [2]。しかし、クエリを人名とした検索結果の中には、利用者の情報要求とは異なる同姓同名の人物による文書が混在している場合があり、これらの文書は、適合文書を探し出す上で妨げとなる。例えば、Google で「田中実」をクエリとして検索したとき、「俳優」の田中実さんが上位を占めており、もし他の田中実さんに関する文書を獲得したい場合、利用者は適合文書が存在するかもわからず、検索リストを下位に向かって見続けなければならない。本論文では字面上同一の人名が、複数の異なる人物を持つ性質を人名の曖昧性という。

ウェブ検索だけでなく、大規模な論文数や書籍数を保持するデジタルライブラリやオンラインショッピングでの検索においても人名の曖昧性は高い。学術論文の代表的なデジタルライブラリである DBLP[1]では、アルファベット順に整列された著者名から、ある著者名の論文一覧を閲覧できる。しかし、「C. Chen」という著者は、「Chao Chen」など名が省略されない異表記名が 179 種類も DBLP 内に存在する [8]。そのため、どの論文が興味のある「C. Chen」に対する論文かを発見するには時間を要する。また、国立国会図書館[2](NDL: National Diet Library) などの書籍におけるデジタルライブラリでは著者名や書籍タイトルをキーとした詳細検索を行える。しかし、同名の異なる著者ごとに書籍を網羅的に区別するまでには至っていない。このように、人名の曖昧性はメディアと言語に依存せず、普遍的にまんえんする問題である。異なる著者ごとに文書をグループ化できれば、ファセットサーチなどで異なる人物ごとに検索結果を提示することができ、適合文書へのアクセスが容易となる。また、推薦システムなど種々の応用が期待できる。

以上の背景によって、人名の曖昧性を解消するためにウェブ検索とデジタルライブラリ内の学術論文を対象に、様々な手法が提案されてきた [7, 3, 4, 8, 11, 6, 12, 9]。しかし、書籍における人名の曖昧性解消の研究はない。そこで本論文では、書籍を対象とした人名曖昧性解消用のための評価用コーパスの構築を目的とする。評価コーパスの構築は、手法の客観的な評価を行うために重要な言語資源であり、人名の曖昧性解消の発展に貢献する。

書籍における著者名曖昧性解消の重要性を確かめるために、代表的な書籍 EC サイトである楽天ブックス[3]を対象に、著者情報を持つ 2,402,680 冊の書籍のうち、どれくらいの書籍が曖昧性のある著者によって書かれたかを NDL の API[4]を用いて推定した。まず、楽天ブックスに登録されている 666,186 個の字面上の異なり著者名を API の入力とする。API の出力は、NDL に登録されている書籍内での同姓同名の異なり著者一覧であり、図 1 のような形式で出力する。666,186 個の著者名の中で、444,469 個の著者名がマッチした。マッチした著者名の中で、曖昧性のある著者名の数は 37,383 個 (8.4%) だった。また、8.4%の著者名による書籍数は 373,370 冊あり、少なくとも 15%の書籍が曖昧性のある著者によって書かれていた。

田中, 実, 1907-1978	田中, 実, 1927-
田中, 実, 1918-	田中, 実, 1946-
田中, 実, 1921-1993	田中, 実, 1910-
田中, 実, 1947-	田中, 実, 1923-
田中, 実, 1949-	田中, 実, 1944-
田中, 実, 1935-	田中, 実, 1946- 理科教育学
田中, 実, 1950-	田中, 実, 1925-
田中, 実, pub. 2018	田中, 実, コンサルタント
田中, 実, pub. 2018.4	田中, 実, 1950- 医師
	.....

図 1: 国立国会図書館の API の出力結果例

2 関連研究

人名の曖昧性解消と最も類似する研究タスクに、著者名の正規化とエンティティリンキングがある。Dumont ら [5] は、表記揺れ (例：A.C. Doyle v.s. Arthur Conan Doyle) と綴り誤り (例：Arthur Konan Doyle) が頻繁に発生する著者情報を整備するために、書籍の著者名を正規化する手法を提案した。我々の研究と同様に、書籍の著者に焦点を当てている。しかし、我々のタスクは同姓同名の著者を個人ごとに区別することであり、彼らとは目的が異なる。Zhu ら [14] は、テキスト中のエンティティを Wikipedia などの知識ベースに登録されているエンティティのいずれかに紐づけるエンティティリンキ

[1] https://dblp.uni-trier.de
[2] https://www.ndl.go.jp
[3] https://books.rakuten.co.jp
[4] https://id.ndl.go.jp/information/sparql/

ング手法を提案した。このタスクは、既知のいずれかのエンティティに振り分けるカテゴリ分類であるのに対して、人名の曖昧性解消はカテゴリが用意されておらず、著者のグループ化が目的である。

ウェブ検索とデジタルライブラリの学術論文を対象として、人物の曖昧性解消における様々な手法が提案されてきた。ウェブ検索における人名の曖昧性解消 [7, 3, 4, 6] は、基本的にウェブページやスニペットの内容を手掛かりに、クラスタリングを行う。また、学術論文を対象とした多くの手法では、著者の所属、論文のキーワードとアブストラクト、会議名もしくはジャーナル名、参考文献を手掛りとする [8, 11, 12, 9]。しかし、書籍を対象とした時、本文の収集は困難であり、かつ上記のようなメタデータは一般的に流通していないため、書籍を対象とした人名の曖昧性解消手法を新たに考案する必要がある。

ウェブ検索における人名曖昧性解消のための代表的なコーパスとして、WePS コーパス [1] と MC4WEPS コーパス [10] がある。これらは、49 個もしくは 110 個の人名それぞれにおいて、人物名をクエリとしてウェブ検索を行う。そして、最大で検索結果上位 100 程度の文書を人手でグループ分けして評価コーパスを構築した。学術論文検索でも様々な評価コーパスが構築されている [11]。しかし、書籍を対象とした評価コーパスは存在しない。

3 評価コーパスの構築

本研究は、著者名の集合 $A \ni a$ と、著者名 a よって書かれた書籍の集合 $B_a \ni b$ が与えられた時、B_a を個々が書いた書籍に分割されるようなクラスタ集合 $C_{a,b}$ を獲得する問題を想定する。構築したコーパスの一部を表 1 に示す。評価コーパスは、各著者名ごとに表 1(a) のような正解クラスタと、表 1(b) のようなクラスタリングに用いるメタデータからなる。(a) は、著者 ID と著者名を先頭行にもつ。二行目からは、クラスタ名とそのクラスタに属する書籍からなり、書籍は ISBN で区別する。メタデータは、正解クラスタに紐づけるための ISBN、書籍タイトル、著者名、出版社名、ジャンル ID からなる。著者名は、最大で 2 名とする。ジャンル ID は、先頭の 3 桁をルートジャンル、後続の 3 桁ずつを中間ジャンルもしくはリーフジャンルとした木構造を模している。ジャンルは、楽天ブックスのジャンル体系にしたがう。表 2 にコーパスの基礎情報を示す。著者名の総数は 2,013 個であり、クラスタ数が増加すると著者名あたりの書籍数が増加する傾向にある。

表 1(a) のコーパスを構築するために、楽天ブックスの著者情報を有効利用する。楽天ブックスの特長として、同姓同名の異なる著者同士を区別するために、「田中実（英語学）」や「田中実（1949-）」など括弧内に肩書き、もしくは生没年がごくわずかな著者名に付加されている。また、同じ肩書きの異なる著者には、「弁護士 1」や「弁護士 2」のように通し番号が付与されている。この付加情報が異なれば、異なる著者による書籍と仮定する。

上記の仮定に基づく自動構築方法を以下で説明する。まず、付加情報を持つ著者名を抽出する。次に抽出された著者名ごとに以下の処理を行う。付加情報が異なれば、異なるクラスタに書籍を割り当てる。クラスタ名は、付加情報からの肩書きもしくは生没年とする。しかし、付加情報を調査したところ、以下の事例があった。

(1) **肩書きに表記の揺れが生じている**
「デンマーク語」と「ﾃﾞﾝﾏｰｸ語」のように、半角と全角の違いによって、異なる肩書きとみなされる付加情報があった。また、「医学」と「医学博士」のような類似する肩書きを持つ著者の書籍同士を比較したところ、両者の書籍は同じ著者によって書かれている事例があった。

(2) **肩書きが付与されている著者と生没年が付与されている著者は同一でない可能性がある**
「英語学」と「1949-」が付与されている著者の書籍同士を比較したところ、同一の著者によって書かれている場合があった。

(3) **肩書きと生没年とは関連のない付加情報がある**
「赤塚不二夫」は、著者名の曖昧性を持たない。しかし、赤塚不二夫（「おそ松くん」）と赤塚不二夫（『おそ松くん』）のような著者を区別すべきでない付加情報が含まれることで、誤って曖昧性を持つと特定される著者名があった。

これらに対処するために、以下のフィルタリングを行なう。(1) については、肩書きと生没年を半角に統一する。次に、肩書きが同じで通し番号を含むペアを除き、肩書きの類似度を Jaccard 係数で測定する。そして、この値が 0.5 より大きければ、どちらか一つの肩書きを選択し、その著者の書籍を評価コーパスから除く。(2) については、生没年が付与されている著者の書籍をコーパスから除く。(3) については、1 節で述べた方法で同姓同名の異なり著者一覧を取得し、異なり著者数が 1 もしくはマッチしなかった著者名を評価データから除く。具体的には、API 用の URI[5]に、以下の sparql クエリを URI の末尾に結合することで著者一覧を取得する。

```
PREFIX rdfs:
  <http://www.w3.org/2000/01/rdf-schema#>
PREFIX skos:
  <http://www.w3.org/2004/02/skos/core#>
PREFIX foaf:
  <http://xmlns.com/foaf/0.1/>
SELECT ?name WHERE{
  ?auth foaf:primaryTopic ?entity .
  ?auth rdfs:label ?name .
  ?entity foaf:name “田中実”
```

[5]http://id.ndl.go.jp/auth/ndla

表 1: 書籍における曖昧性解消の評価コーパス

(a) 正解クラスタ

name324:田中実
民法学:9784326450312 9784313314795 9784326400447 9784808902285 9784641037342 9784882611745 9784808902278 9784766414004 9784326400003 9784766409307 9784766409291 9784842020099...
英語学:9784896840193 9784410363269 9784410362262 9784410111068 2100011154575 9784410362255 9784327440954 9784410363245 9784344405196 9784410111051 9784569572031 9784410363252...
不動産投資:9784828407722 9784931421905 9784898001639
...

(b) 書籍のメタデータ

isbn	title	author1	author2	publisher	pubdate	genreid
9784832974111	札幌の自然を歩く 道央地域の地質あんない	宮坂省吾	田中実	北海道大学出版会	20110600	001012005
9784896840179	英語のニュアンス Q & A 洋販新書	田中実		IBC パブリッシング	19970700	001020007
9784882611745	ケースで学ぶ借地・借家法 ケースで学ぶシリーズ	田中実	藤井輝久	信山社 大学図書	19910200	001008009001
9784641180239	医療の法律紛争 医師と患者の信頼回復のために 有斐閣選書	田中実	藤井輝久	有斐閣	19861000	001010010001

表 2: 評価コーパスの基礎情報

異なり著者数	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	19
著者名の数	1,091	559	263	112	71	42	38	16	7	10	5	3	3	4	2	2	2
書籍数/著者名	17.1	18.1	24.0	27.9	30.2	36.0	41.9	41.2	26.9	33.6	39.2	32.7	45.3	89.8	50.0	32.5	46.0

NDL は、同姓同名の著者個人を日本目録規則に基づいた方法で区別しており[6]、信頼性の高い書誌情報である。

単に NDL の著者情報をエクスポートするだけで、正解クラスタリングを構築することも考えられる。例えば、NDL が提供する ISBN 検索は、特定の URI[7] に ISBN を追加して問い合わせることで、「田中, 実, 1921-1993」のような上述の規則にしたがった付加情報を出力する。しかし、構築した評価コーパス中の 2,013 個の著者名による全書籍に対して ISBN 検索を行なったところ、わずか 646 個の著者名にのみ、80%以上の書籍に付加情報が存在した。よって、一著者あたりの被覆率が低いため、NDL のみから正解クラスタリングを構築すべきでない。

4 評価実験

まず、構築した評価コーパスにおける正解クラスタの妥当性を評価した。評価方法は、3 節の終わりで述べた被覆率が比較的高い 646 個の著者名を対象に、NDL で集計されたクラスタと、構築したコーパスにおけるクラスタとの一致度を測定した。一致度の尺度には、クラスタリング結果の評価に用いられる純度と逆純度の F 値を用いた。測定の結果、0.995 という高い F 値を達成したため、品質の高い正解クラスタと考える。

次に、基本的な曖昧性解消手法を実装し、評価コーパスを用いてベースライン手法との比較評価を行った。実装した手法は、2 値分類ステップとクラスタリングステップからなる。2 値分類ステップは、同姓同名の著者によって出版された書籍のペアが与えられたとき、二つの書籍が同一の著者によって書かれたか否かを分類する。このステップでは、図 2 の Tran[13] らのフレームワークを利用した。彼らは、論文メタデータから素性を設計し、複数の多層パーセプトロンを用いたアンサンブルによって分類を行なった。本研究で用いた 2 値分類のための素性を表 3 に示す。クラスタリングステップは、新規の書籍が与えられたとき、すでにグループ化された書籍とペアを作り、図 2 のモデルを適用する。そして、$p(y = \text{同じ著者} \mid x)$ が最大となる書籍が属するクラスタに新規の書籍を割り当てる。もし、いずれのクラスタにおいても $p(y = \text{同じ著者} \mid x) > \theta$ を満たさない場合は、新たなクラスタに書籍を割り当てる。ベースライン手法として、Artiles ら [1] も用いた ALL-IN-ONE というすべての書籍を同じクラスタに割り当てる手法を採用した。

表 4 は、評価コーパスを用いて著者名による 10 分割交差検定で学習とテストを実施した結果である。特に、NDL の所蔵数に対するコーパス中における書籍数の割合が高い著者名を掲載した。基本手法は、ベースラインより優れた F 値を達成した。

5 おわりに

これまで、ウェブ検索とデジタルライブラリの学術論文を対象とした人名の曖昧性解消手法が提案されてきた。しかし、書籍を対象とした人名曖昧性解消の研究はない。本論文では、楽天ブックスの著者情報に付加され

[6] https://www.ndl.go.jp/jp/data/faq/author.html
[7] http://iss.ndl.go.jp/api/opensearch?isbn=

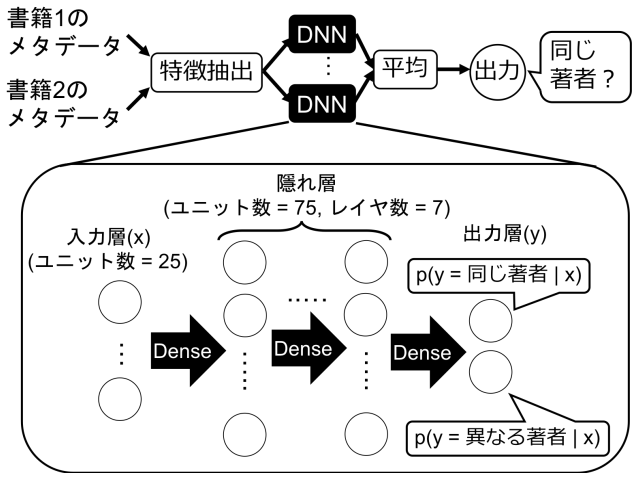

図 2: Tran ら [13] の曖昧性解消手法

表 3: 2 値分類に用いる素性の一覧

カテゴリ	素性	ユニット数
タイトル	6 尺度による類似度	6
	6 尺度による共著者の類似度	6
	同じ共著者？	1
著者	書籍 1 の著者数	1
	書籍 2 の著者数	1
	著者数の差	1
出版社	6 尺度による類似度	6
ジャンル	リーフジャンルが同じ？	1
	ジャンルノードの重複数	1
出版日	出版日の差	1

表 4: 10 個の著者名における実験結果

著者名	所蔵冊数		被覆率	異なり著者数	F 値	
	NDL(A)	評価コーパス (B)	(B)/(A)		ベースライン	本手法
井上正治	152	106	70%	3	0.596	**0.832**
浅井隆	245	224	91%	4	**0.907**	0.692
伊藤典子	74	63	85%	5	0.633	**0.942**
佐藤裕之	71	66	93%	4	0.718	**0.985**
山本祐司	102	79	77%	4	0.838	**0.924**
矢上裕	61	103	169%	2	0.821	**0.990**
山下純一	58	50	88%	2	0.684	**0.945**
原田治	74	66	89%	3	**0.706**	0.696
浜田翔子	79	104	132%	2	0.948	**1.000**
増田豊	55	41	75%	4	0.794	**0.909**
平均	97.1	90.2	96.8%	3.3	0.764	**0.891**

る肩書きと生没年を有効利用し、著者情報のフィルタリングを行うことで、書籍における著者名曖昧性解消のための評価コーパスを構築した。構築したコーパスの正解クラスタと、国立国会図書館に登録されている書誌情報の一致度を評価し、評価コーパスの品質を評価した。今後は、書籍の著者名曖昧性解消の手法を提案する。構築したコーパスは、情報学研究データレポジトリ[8]（IDR）にて公開する予定である。

参考文献

[1] Javier Artiles, Julio Gonzalo, and Satoshi Sekine. WePS 2 evaluation campaign: overview of the web people search clustering task. In *Proceedings of the WWW Web People Search Evaluation Workshop*, pp. 64–69, 01 2009.

[2] Javier Artiles, Julio Gonzalo, and Felisa Verdejo. A testbed for people searching strategies in the www. In *Proceedings of SIGIR*, p. 569–570, 2005.

[3] Liwei Chen, Yansong Feng, Lei Zou, and Dongyan Zhao. Explore person specific evidence in web person name disambiguation. In *Proceedings of EMNLP-CoNLL*, pp. 832–842, Jeju Island, Korea, 2012.

[4] Agustín D. Delgado, Raquel Martínez, Víctor Fresno, and Soto Montalvo. A data driven approach for person name disambiguation in web search results. In *Proceedings of COLING*, pp. 301–310, 2014.

[5] Béranger Dumont, Simona Maggio, Ghiles Sidi Said, and Quoc-Tien Au. Who wrote this book? a challenge for e-commerce. In *Proceedings of the 5th Workshop on Noisy User-generated Text*, pp. 121–125, 2019.

[6] Hojjat Emami. A graph-based approach to person name disambiguation in web. *ACM Transction on Management Information Systems*, Vol. 10, No. 2, 2019.

[7] Xianpei Han and Jun Zhao. Structural semantic relatedness: A knowledge-based method to named entity disambiguation. In *Proceedings of ACL*, p. 50–59, 2010.

[8] Ijaz Hussain and Sohail Asghar. A survey of author name disambiguation techniques: 2010–2016. *The Knowledge Engineering Review*, Vol. 32, , 12 2017.

[9] Kunho Kim, Shaurya Rohatgi, and C. Lee Giles. Hybrid deep pairwise classification for author name disambiguation. In *Proceedings of CIKM*, p. 2369–2372, 2019.

[10] Soto Montalvo, Raquel Martínez, Leonardo Campillos, Agustín D. Delgado, Víctor Fresno, and Felisa Verdejo. MC4WEPS: a multilingual corpus for web people search disambiguation. *Language Resources and Evaluation*, Vol. 51, No. 3, pp. 805–832, Sep 2017.

[11] Mark-Christoph Müller. Semantic author name disambiguation with word embeddings. In *Research and Advanced Technology for Digital Libraries*, pp. 300–311, 2017.

[12] L. Peng, S. Shen, J. Xu, Y. Fu, D. Li, and A. L. Jia. Diting: An author disambiguation method based on network representation learning. *IEEE Access*, Vol. 7, pp. 135539–135555, 2019.

[13] Hung Nghiep Tran, Tin Huynh, and Tien Do. Author name disambiguation by using deep neural network. *Intelligent Information and Database Systems*, pp. 123–132, 2014.

[14] Ganggao Zhu and Carlos A. Iglesias. Exploiting semantic similarity for named entity disambiguation in knowledge graphs. *Expert Systems with Applications*, Vol. 101, pp. 8–24, 2018.

[8] https://www.nii.ac.jp/dsc/idr/