DEIM Forum 2021 B21-4

# 距離学習を用いたHuman-in-the-loop エンティティマッチングフレームワークの提案

大沢　直史† 　伊藤　寛祥†† 　福島　幸宏††† 　原田　隆史†††† 　森嶋　厚行††

† 筑波大学 知識情報・図書館学類　〒 305-0821 茨城県つくば市春日 1-2
†† 筑波大学 図書館情報メディア系　〒 305-0821 茨城県つくば市春日 1-2
††† 東京大学 情報学環・学際情報学府　〒 113-8654 東京都文京区本郷 7 丁目 3-1
†††† 同志社大学大学院総合政策科学研究科　〒 602-8580 京都府京都市上京区
E-mail: †naofumi.osawa.2020b@mlab.info, ††{ito,mori}@slis.tsukuba.ac.jp, †††fukusima-y@iii.u-tokyo.ac.jp, ††††ushi@slis.doshisha.ac.jp

**あらまし**　エンティティマッチングは情報統合における重要な問題であり，これまでも様々なアプローチが試みられてきた．本研究では，距離学習を用いた，二値分類モデルと人間による Human-in-the-loop によるエンティティマッチングのフレームワークを提案する．距離学習とは，レコード間の関係を考慮した埋め込み空間を学習する手法であり，意味的に類似したレコードは空間中で近い位置に埋め込まれるという特性をもつ．これを利用し，同一レコードの候補を埋め込み空間における近傍探索により発見・K 近傍クラフを構成し，グラフの縮約を繰り返すことでエンティティマッチングを実現する．ここで，二値分類モデルが一定の確信度を越えるレコードペアについては，分類モデルによるマッチングを行い，下回るペアについては人間が行うことで，精度の向上およびクラウドソーシングのコストの削減をねらう．実験では，国立国会図書館が所有する総合目録に対する書誌データに対し，本フレームワークを適用することで書誌統合を行った．実験結果より，提案フレームワークは，ナイーブな方法と比較して大幅な比較回数の削減を実現し，高精度かつ低コストにエンティティマッチングが行えることを示した．

**キーワード**　エンティティマッチング，Human-in-the-loop，距離学習

## 1 はじめに

エンティティマッチングとは，データベース中において同一の実体を参照しているレコードの集合を識別する問題である．エンティティマッチングは，データのクリーニングや複数のデータベースの統合において非常に重要であるが，エンティティ作成時の入力の揺れや欠損がしばしば生じ，完全にルールベースな手法ではすべてのエンティティに対して完全なマッチングを行うことはできないという問題がある．加えて，データベースがもつレコード数が大きい場合，すべてのレコードの組み合わせを参照し，それらが同じエンティティを参照しているか否かを判定する同定作業を行うためには膨大な計算時間がかかり，さらに同定にはその対象に合わせた正規化処理や同定キーの作成が必要になるといった問題がある．

本研究では Human-in-the-loop に基づくアプローチに距離学習モデルと二値分類を用いた手法を提案する．距離学習とは，データ間の関係を考慮し，同一のクラスに属するデータ同士は空間中で近い位置に埋め込み，異なるクラスに属するデータ同士は遠い位置に埋め込むようなマッピング関数を学習する手法である．適切な距離学習が行えれば，同一レコードの可能性がある候補のペア選択およびルールベースでは判定が難しいペアの選択について，アドホックでないアプローチを提供することが期待できる．

本論文で提案する手法は二値分類モデルとクラウドソーシングによる，Human-in-the-loop エンティティマッチングに距離学習を導入したものである．はじめにデータベースがもつデータを距離空間への埋め込みを行う．次に埋め込まれたデータをそれぞれノードとし，距離が最も近い上位 K 件のノードを接続する K 近傍グラフを構築する．そして各エッジに対し，はじめに二値分類モデルによる同定を行い，一定の閾値を越える確信度での予測値が得られた場合，レコードの統合・非統合を確定する．閾値を越える確信度が得られなかった場合，レコードペアに対してクラウドソーシングによる同定を行う．この閾値によって，クラウドソーシングによるタスク数を制御することができ，低い閾値を採用することでタスク数が減少し，高い閾値を採用することでタスク数が増加する．ここで，共有ノード数に基づくエッジのスコアリングを行い，このスコアが大きいものから同定を行うことで比較回数を削減する．

通常のデータ統合を考える場合，データの直積集合から全てのデータの組み合わせを参照し，それぞれが同一なデータであるか否かを判定する必要がある．その際，明らかにデータが同一ではない組み合わせに対しても同定判定を行う必要があり，比較回数に非常にコストがかかる．本研究では距離尺度を採用し，距離が近いデータのみを同定判定する距離学習によるブロッキング手法を用いた．これにより，明らかに同一ではな

いデータに対して判定を行う必要がなくなる．またスコアが高いノードペアは他のノードと共有するエッジを多く持つため．優先的に同定することで比較回数を大きく削減することができる．さらに，同定精度が高く，ヒューリスティックにタスクに対応することのできる人間ワーカを介入することで機械が同定判定を行えなかったデータに対しても漏れなく対応することができる．

実験では，国立国会図書館の総合目録に対する書誌同定の問題に対して提案手法を適用し，本手法の有効性を検証する．二値分類モデルとしては，レコードの各特徴量の類似度に基づくモデルを作成し，距離学習手法としては Siamese Network 構造を採用し，Contrastive Loss に基づく損失関数を最小化することで距離学習を行った．

実験の結果，二値分類モデルの学習においては 70%の精度が得られ，距離学習を用いたブロッキングにおいては，$K = 10$ の K 近傍グラフを構築することで，約 90%の再現率で同定候補をあげることができることを示した．また，本フレームワークにおいて二値分類モデルの確信度の閾値を上げて，積極的にクラウドワーカーをタスクに取り入れることにより，二値分類モデルのみを用いた場合よりも精度が向上することを示した．さらに，共有ノード数に基づくスコアリングを行うことで，ランダムにエッジを選択して統合する場合と比較して比較回数が減少することを示した．

本論文の構成は以下である．2 章で関連研究について説明を行い，3 章で Human-in-the-loop エンティティマッチングフレームワークを提案する．4 章では国立国会図書館の総合目録を用いたエンティティマッチング実験とその結果について示す．最後に，5 章において実験に対する考察を行い，6 章にて本論文をまとめる．

## 2 関 連 研 究

本章では，本研究の関連研究として，エンティティマッチングに対するブロッキング手法，クラウドソーシングを用いた書誌同定，Human-in-the-loop を利用したデータ統合手法，レコード同定問題についてそれぞれ述べる．

### 2.1 エンティティマッチングに対するブロッキング手法

エンティティマッチングにおいて，多くの場合，比較回数を削減する手法として「ブロッキング」と呼ばれる手法を用いることで，あらかじめ同定を行う候補を限定する．

standard blocking [1] では，キーとなる属性を定め，同一のキーを持つレコードをマッチングの候補集合とする．

キャノピークラスタリング [2] は教師無し学習手法の一つで，tf-idf を類似度のスコアとして用いる．tf-idf による距離がある一定以下であるようなエンティティをまとめ，キャノピーと呼ばれるクラスタを構成する．キャノピークラスタリングではクラスタを大まかに分類する数を見積もることができるため，k-means 法や凝集法に代表される一般的なクラスタリングアルゴリズムを行うための事前処理として用いられることが多い．

我々の研究ではブロッキングを使用せずに距離学習と K 近傍グラフの構築による比較回数の削減を試みる．

### 2.2 クラウドソーシングによる書誌同定

クラウドソーシングは，不特定多数の人々に仕事を依頼する仕組みを指す．原田ら [3] はこのクラウドソーシングを利用した書誌同定手法を提案した．ここでは，国立国会図書館総合目録と京都府立図書館が持つ書誌データ 700 万件の内，ISBN が空白となっている書誌 200 万件を機械的に突合し，類似度が極めて高い書誌と極めて低い書誌を除いた，機械的判別が難しい書誌データのペア 8000 件を算出し，クラウドソーシングによって同定を行った．同定の結果，書誌データが一致しているペアに対しては約 93 %，書誌データが一致していないペアに対しては 99 %の精度で同定を行うことが可能であった．

クラウドソーシングでは高い精度で同定を行うことが出来たが，計算機に比べて高速度に同定を行えないこと，金銭的なコストが生じること，同定対象のデータベースが非常に大きいことなどから全てのマッチングをクラウドソーシングで行うことは現実的ではないと考えられる．そこで本研究は，第一手順として高速度に処理の行える機械学習によるマッチングを行い，機械学習によって同定が行えなかった問題をクラウドソーシングのタスクとして人間が行う仕組みを採用する．

### 2.3 Human-in-the-loop を利用したデータ統合手法

近年，機械学習モデルと人間が相互補完しながら動作するシステムを指す「Human-in-the-loop」の研究が盛んに行われている．例として，Li [4] は Human-in-the-loop を用いたデータ統合手法の提案を行った．この手法では (1) 規則に基づいた候補ペアの生成 (2) クラウドに基づいた候補ペアの絞りこみの 2 つの段階を用いた手法を提案した．

規則に基づいた候補ペアの生成では，ある規則の集合を設定し，その規則を満たすレコードのペアを同一候補として挙げた．この場合，全てのペアが規則を満たすかどうかを確認する必要があるため，データの大きさによって計算量が膨大になってしまう．提案手法ではフィルタリングを用いた低コストな候補ペアの作成手法を提示した．

クラウドソーシングに基づいた候補ペアの絞りこみでは，選択，推論，絞り込みの 3 つの要素を用いた候補ペアの同定手法を提案した．クラウドソーシングではワーカにタスクを遂行してもらう際に報酬を支払う必要がある．そこで提案手法では，クラウドに提示するペアを削減できるような「有益な」ペアを優先的にクラウドに提示する．これによって同定する必要があるタスクを減らすことができ，クラウドに支払うコストが削減できる．

我々の研究では，候補ペアを作成する際には距離学習によって学習された書誌データをそのまま採用する．距離学習は類似しているデータは近く，非類似のデータは遠く埋め込まれるように学習が行われる．したがって我々の手法では知識や経験則，ルールに基づくブロッキングを持たず，距離空間において近いデータを同定候補集合として同定処理を行う．

Wang ら [5] はそれぞれのデータの全組み合わせに対し，ルールに基づく類似度のランキング付けを行い，ある閾値以上のペアに関してクラウドソーシングによる同定判定を行う．これにより高い精度での同定を可能としている．我々の研究では距離学習によってデータを適切な位置に埋め込み，K 近傍グラフによって隣接するノードを接続することで候補ペアの作成をしている点で異なる．

### 2.4 レコード同定問題

単一または異なる情報源の間で重複するエンティティを見出すエンティティマッチングの研究は古くから行われている [6]．エンティティマッチングはレコード同定や名寄せ等様々な名称で表現される．レコード同定は本来，ばらばらに記録がなされていた出生・死亡・婚姻・離婚などの記録を人物毎につなぎ合わせ，様々な統計解析処理を行うために用いられた．近年では多岐にわたるデータを対象としたデータベース分野での研究が行われている [7]．

桂井ら [8] の研究に，異なる複数の学術情報データベースから著者同定を行う研究がある．ここでは学術情報がもつメタデータに対して，文字列類似度やコサイン類似度，数値に関しては差分を用いたランキング付けを行い，その尺度を用いて同定を行うアプローチを行っている．

本研究では文字列類似度と編集距離の類似度指標を学習する二値分類モデルを構築した．適切な学習が行えた場合にマッチング処理の高速化が期待できる．

## 3 提 案 手 法

提案手法は Algorithm 1 として表される．提案手法では，はじめにデータベース $\mathcal{S}$ が持つ書誌データ $(\boldsymbol{a}, id)$ を距離空間上に埋め込む．距離空間への埋め込みは 4.2 節で述べる距離学習モデルを用いて行われる．次に埋め込まれたデータをそれぞれノードとし，距離が最も近い上位 K 件のノードを接続する K 近傍グラフを構築する．そして各ノードに対しスコアを計算し，スコアの高いノード同士を候補ペアとして同定を行い，同一データの集合を得る．ノード間がマッチした場合はノードが統合されることでグラフの縮約が行われ，一致しなかった場合はそのノード間のエッジを切断する．以上の操作をノードが独立するまで繰り返す．距離が近いノード間を接続することで，データの意味が似ているノードが接続されるため，データ統合におけるブロッキングの役割を果たす．さらにここでは，比較回数を削減するため，重複する隣接ノード数に基づくスコアリングを行い，この値が大きいものから判定を行う．表 1 に本論文で用いる記号の表記法を示す．

ノードペアは 4.2 節で述べる二値分類モデルによって構築される $AI$ を用いて同定を行う．$AI$ が $\alpha$ 以上の確信度をもって判定をした場合は $AI$ の結果を採用した同定，反対に確信度が $\alpha$ 以下の場合はクラウドソーシングのタスクとしてワーカが同定を行う．同定された場合はノードを統合し，グラフの縮約を行う．また同定されなかった場合にはノードペア間のエッジの切断を行う．以上の操作を繰り返すことで，距離空間上ではグラフの縮約が繰り返されエンティティが 1 つに定まったグラフが構築される．アルゴリズムの各構成要素については以降の節で説明する．

表 1　記号の定義

| 記号 | 定義 |
|---|---|
| $\mathcal{S} = \{S_i\}$ | データベース |
| $S_i = (\boldsymbol{a}, id)$ | 各データ ($\boldsymbol{a}$ = 特徴量，$id$ = 番号) |
| $\mathcal{D} = \{(S_i, S_j)\}$ | データが一致しているペアの集合 |
| $M_\Theta \colon \mathcal{S} \to \mathbb{R}^k$ | 距離学習 |
| $Nearest(S_i, K)$ | $i$ との距離が近い上位 $K$ 個のデータ |
| $\mathcal{G}_K = (\mathcal{V}, \mathcal{E})$ | データの K 近傍グラフ |
| $\mathcal{V} \subseteq \mathcal{P}(\mathcal{S})$ | K 近傍グラフにおけるノード集合 |
| $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ | K 近傍グラフにおけるエッジ集合 |
| $AI \colon \mathcal{V} \times \mathcal{V} \to [0, 1]$ | 機械学習による二値分類モデル |
| $Human \colon \mathcal{V} \times \mathcal{V} \to 0, 1$ | クラウドソーシングにおけるタスク |
| $\alpha$ | 二値分類モデルに対する閾値 |

**Algorithm 1** Proposed Method

**Input:** Database $S$, Pair $D$
**Output:** A set of nodes in the kNN-Graph $\mathcal{V}$

*Learn Metrics $M_\theta$ and AI*
Construct kNN-Graph $\mathcal{G}_K = (\mathcal{V}, \mathcal{E})$
**while** $\mathcal{E} \neq \varnothing$ **do**
  $(v_i, v_j) \leftarrow \underset{(v_i, v_j) \in \mathcal{E}}{\arg\max}\ \boldsymbol{Score}(v_i, v_j)$
  AI score $\leftarrow AI(v_i, v_j)$
  **if** AI score $\geqq \alpha$ **then**
    $D \leftarrow D \cup v_i \times v_j$
    $\mathcal{G} \leftarrow \mathcal{G} \setminus (v_i, v_j)$
  **else if** AI score $\leqq 1 - \alpha$ **then**
    $\mathcal{E} \leftarrow \mathcal{E} \setminus (v_i, v_j)$
  **else**
    Human score $\leftarrow Human(v_i, v_j)$
    **if** Human score == 1 **then**
      $D \leftarrow D \cup v_i \times v_j$
      $\mathcal{G} \leftarrow \mathcal{G} \setminus (v_i, v_j)$
    **else**
      $\mathcal{E} \leftarrow \mathcal{E} \setminus (v_i, v_j)$
    **end if**
  **end if**
**end while**

### 3.1 距 離 学 習

距離学習ではデータベースにおけるデータを埋め込み空間中にマップするための関数 $M_\Theta : \mathcal{S} \to \mathbb{R}^m$ を学習する．ここで $\Theta$ は距離学習のマッピング関数のパラメータである．距離学習の基本的なアイデアは，同一のクラスに属するデータは空間中で距離が近くなり，異なるクラスに属するデータは空間中で距離が遠くなるようにするというものである．データが一致しているペア $\mathcal{D}$ が与えられたとき，以下の最適化問題を解くことで距離学習のパラメータ $\Theta$ を学習する．

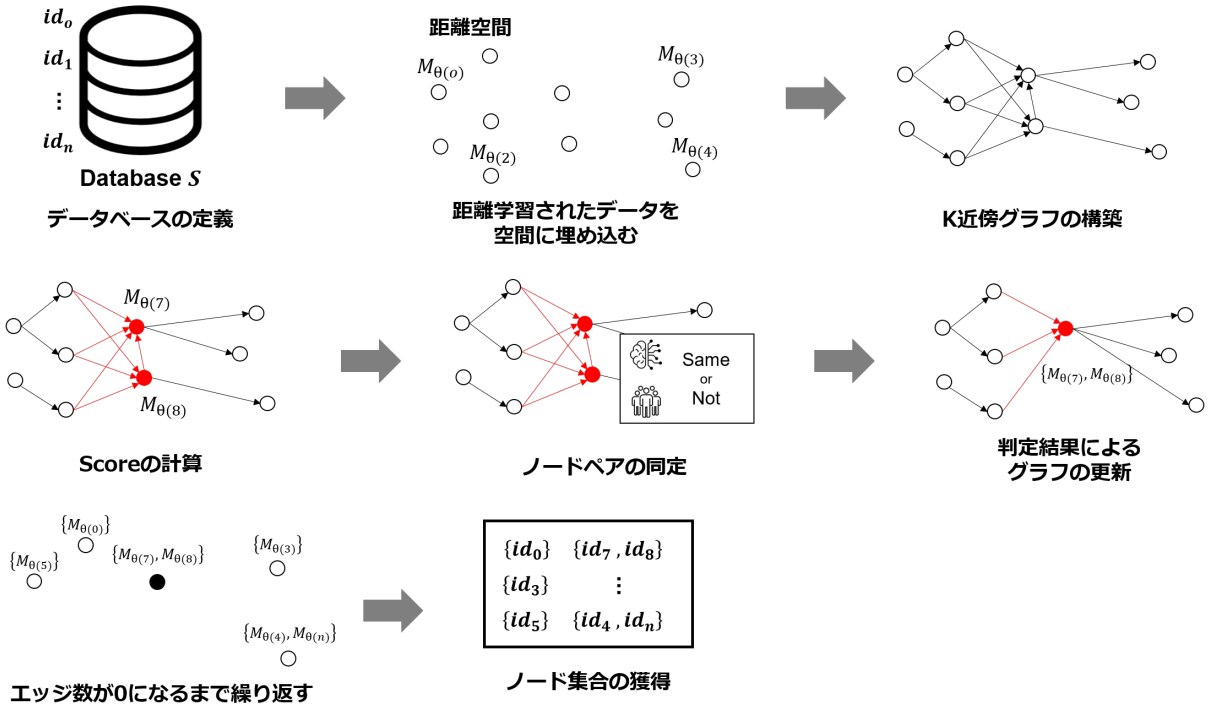



図 1　提案手法の概要

$$\Theta = \arg\min_{\Theta} \sum_{(S_i,S_j)\in\mathcal{D}} Dist_1(M_\Theta(S_i), M_\theta(S_j)) - \sum_{(S_i,S_j)\notin\mathcal{D}} Dist_2(M_\theta(S_i), M_\theta(S_j))$$

ここで，$Dist_1$,$Dist_2$ はそれぞれ $m$ 次元埋め込み空間における距離を表している．

### 3.2　距離学習に基づく K 近傍グラフの構成

本研究では距離学習で得られた埋め込み空間中でのデータ点に対して K 近傍グラフを構築する．K 近傍グラフを以下のように定義する．

$$\mathcal{G}_K = (\mathcal{V}, \mathcal{E})$$

ここで，$\mathcal{V} \subseteq \mathcal{P}(\mathcal{S})$ はノード集合を表し，各ノード $v \subseteq \mathcal{S}$ はデータの集合になる．K 近傍グラフを構成した初期状態においては，各ノード $v$ は各データ点ひとつのみからなる集合になる．すなわちエッジ集合 $\mathcal{E}$ は以下のように初期化される．

$$\mathcal{E} = \{(\{S_i\}, \{S_j\}) \mid S_j \in Nearest(S_i, K), S_i \in \mathcal{S}\}$$

ここで，エッジは有向エッジであり，$Nearest(S_i, K) \subseteq \mathcal{S}$ は埋め込み空間中における，データ $S_i$ と距離が近い上位 K 個のデータ集合を表す．提案フレームワークにおいて同一のデータと判定されたノードペアは K 近傍グラフ中で逐次縮約され，同一でないデータと判定されたノードペアのエッジはグラフから逐次削除されていく．本論文では，エッジ $(v_i, v_j)$ によるグラフ $\mathcal{G}$ の縮約を $\mathcal{G} \setminus (v_i, v_j)$ と表記する．

### 3.3　同定するノードの優先順位のスコアリング

本研究では，比較するペアの個数を最小化するため，同定するノード間に優先順位を計算する．ここでの基本的なアイデアは，K 近傍グラフにおいて，2 つのノードにおいて共有する隣接ノード数が多いほど，その 2 つのデータが統合されたときに縮約されるエッジ数が大きくなり，結果として比較が必要になるデータのペア数が少なくなるというものである．エッジ $(v_i, v_j)$ に対するスコアは以下のように計算する．

$$\boldsymbol{Score}(v_i, v_j) = |\mathcal{N}(v_i) \cap \mathcal{N}(v_j)|$$

ここで，$\mathcal{N}(v_i) \subseteq \mathcal{S}$ はノード $v_i$ に対するエッジをもつノード集合である．

## 4　実　　験

本章では，実データを用いた実験により，提案手法の有用性を確認する．実験では (1) 二値分類モデルによる書誌同定 (2) 距離学習による書誌データの表現 (3) エンティティマッチング実験の 3 種類の評価実験を行う．

### 4.1　使用データ

本論文では，国立国会図書館の総合目録によって作成された書誌データベース $\mathcal{S}$ 内の書誌データマッチング問題を考える．$\mathcal{S}$ は複数の図書館の書誌データベースが混在し，1 つのデータベースとなっている．これは共通のスキーマを持つ複数データベースと同義である．

各書誌データ $S_i$ は (1) タイトル (2) 巻号 (3) 著者 (4) 出版社 (5) 出版年 (6)ISBN(7) ページ数 (8) 外形の特徴量を持つ．

書籍出版物の書誌を特定することができる ISBN[1]は 1 冊の

1：https://isbn.jpo.or.jp/

書籍に対して固有の番号が付与される．そのため，本来ならばISBNが一致している書誌は同一の書誌であるといえるが，ISBNの取得には費用が掛かることなどから，市場に出ることのなくなった書籍のISBNを別な書籍に流用する，いわゆる使い回しの事例も存在する．また，ISBNが使用されるようになったのは1981年頃からであり，それ以前の書籍にはISBNが付与されていない．以上のように使い回しや付与されていない書誌が存在することから，実際の同定処理には用いず，4.1節で述べる機械学習モデルの学習の際に用いる．出版年に関して，表記方法の揺れが非常に多いために一律の正規化を行うことは非常に難しいと考え，本実験では使用しない．

### 4.2 二値分類モデルによる書誌同定

二値分類モデルによる書誌同定について述べる．これは提案手法の$AI(v_i, v_j)$にあたる．本実験では書誌データのペアを作成し，ペア間から算出される類似度を学習することで書誌データが同一なものであるか否かを判定する二値分類モデルを構築する．構築されたモデルに書誌データのペアを提示して，そのペアが一致しているか否かを判定する実験を行う．これにより，機械学習が書誌同定に有効であることを示す．二値分類モデルの構築にはKeras[2]を使用した．

#### 4.2.1 データセット

本実験は国立国会図書館が所有する総合目録の書誌データを使用し，ISBNと書誌データ間の類似度を用いて書誌データのペアを作成した．また,「書誌データが完全に一致していると判断できるデータ」と「書誌データが明らかに違うと判断できるデータ」に関しては扱わず，人間が判定を間違える可能性のある「極めて似ている書誌データ」を扱うため，ある一定の書誌データ類似度を持つデータのみに限定して学習を行った．以下の式を用いて類似度の閾値を決定し，書誌データのペアを作成した．

$$\mathcal{D} = \{(S_i, S_j) \mid 1.0 > 1.0 - Distance(S_i, S_j) \geqq P\}$$

本実験では$P = 0.6$とし， 書誌データの類似度が0.6以上1.0以下のデータを使用した．$Distance(S_i, S_j)$はPython標準ライブラリであるdifflibのSequenceMatcher[3]を使用し，書誌データ間の類似度を求めた．これにより，正解データ (類似度が高く，ISBNが一致している書誌のペア) が約6万ペア，不正解データ (類似度が高く，ISBNが一致していない書誌のペア) が約43万ペア作成された．ペアの一例を表2に示す．

次に，作成されたペアからニューラルネットワークに入力する特徴量を算出する．本実験ではISBNによる同定は行わず，結果の確認のみに使用することからISBNは使用しない．また，出版年は表記の揺れや作成者による入力の違い，書誌による入力の違いが激しく，一律にすべてを正規化した表現にすることが困難である．差分の大きい入力データを用いることはモデルの性能を著しく低下させると考え本実験では使用しない．よって4.1節で示した特徴量のうちISBNと出版年を除いた6つの特徴量に対し以下の類似度計算を行い，1冊の書誌データから36個の特徴ベクトルを算出した．

- Jaro-Winkler係数[4]
- Levenshtein係数[6]
- Dice係数 $= \frac{2\times|\boldsymbol{a}_{ik}\cap\boldsymbol{a}_{jk}|}{|\boldsymbol{a}_{ik}|+|\boldsymbol{a}_{jk}|}$
- Jaccard係数 $= \frac{|\boldsymbol{a}_{ik}\cap\boldsymbol{a}_{jk}|}{|\boldsymbol{a}_{ik}\cup\boldsymbol{a}_{jk}|}$
- Bigrams-Jaccard係数 $= \frac{|Bi(\boldsymbol{a}_{ik})\cap Bi(\boldsymbol{a}_{jk})|}{|Bi(\boldsymbol{a}_{ik})\cup Bi(\boldsymbol{a}_{jk})|}$
- Trigrams-Jaccard係数 $= \frac{|Tri(\boldsymbol{a}_{ik})\cap Tri(\boldsymbol{a}_{jk})|}{|Tri(\boldsymbol{a}_{ik})\cup Tri(\boldsymbol{a}_{jk})|}$

ここで，$\boldsymbol{a}_{ik}$は書誌$S_i$のk番目の特徴量を表す．編集距離であるJaro-Winkler係数とLeventhtein係数は，$\boldsymbol{a}_{ik}, \boldsymbol{a}_{jk}$のそれぞれが持つ特徴量の最大文字数を正規化した値を採用した．また，すべての特徴ベクトルを類似度に統一するため，Jaro-Winkler係数とLevenshtein係数に関しては1から算出された値を引き，類似度とした値を採用した．また，Bigram-Jaccard係数とTrigram-Jaccard係数はそれぞれ書誌データをBigram,Trigramによって分割された文字列を集合の要素としてJaccard係数によって計算を行った．

#### 4.2.2 学習とその結果

学習にはKeras[5]を使用し，二値分類を行うモデルを作成した．モデルは入力層，全結合層1，ドロップアウト層，全結合層2，ドロップアウト層，全結合層3，全結合層4，出力層で構成される．活性化関数は隠れ層はRelu関数を採用し，出力層はsigmoid関数とする．ドロップアウト層では前層の出力の10%のドロップアウト，損失関数にはクロスエントロピー誤差，最適化関数にはAdamを採用した．ペアデータ間から作成されたデータ約49万件のうち，34万件を学習データとして用い，15万件によって検証を行った結果を表3に示す．

表3 *AI*の予測による混同行列

| | | 予測したクラス | |
|---|---|---|---|
| | | 一致 | 不一致 |
| 実際のクラス | 一致 | 12519 | 5575 |
| | 不一致 | 113 | 127373 |

書誌割れに比べ，書誌誤同定は検索システムにおいて優先的に解決すべき問題であるといえる．横断検索システムにおいて，本来ならば検索結果に表示されるべき書誌が表示されていないという現象はユーザービリティに大きく影響するからである．結果より，書誌誤同定に関しては大きく貢献するモデルを構築することが可能であったが，書誌割れに対しては改善の余地が見つかった．本実験での特徴量算出は，文字列をベースに行った．表記に対する詳細な正規化や本実験で用いなかった出版年の採用等，モデルの性能を向上させることは可能であると考える．

2：https://keras.io/ja/
3：https://docs.python.org/ja/3/library/difflib.html
4：https://pypi.org/project/python-Levenshtein/
5：https://keras.io/ja/

表 2　書誌データのペア例

| 書誌 | タイトル | 巻号 | 著者 | 出版社 | Page | 外形 |
|---|---|---|---|---|---|---|
| $S_1$ | 若おかみは小学生! | | 亜沙美, 令丈ヒロ子 | 講談社 | 215p | 18cm |
| $S_2$ | 若おかみは小学生!：花の湯温泉ストーリー 1 | [PART1] | 亜沙美, 令丈ヒロ子 | 講談社 | 215p | 18cm |

### 4.3　距離学習による書誌データの表現

距離学習における書誌データの表現について述べる．これは提案手法の $M_\Theta$ にあたる．本実験では書誌データ間で発生する関係性を距離空間上に表現することを目的とする．

#### 4.3.1　データセット

4.1 節と同様の書誌データを用い，分散表現を獲得する．書誌データのベクトル化には Facebook 社が提供する fasttext[6] を採用した [9]．書誌データでは，特にタイトルにおいてしばしば造語，略語といった未知語が見受けられる．fasttext では Subword model と呼ばれる，単語を n-gram 集合として表現をしそれぞれの要素の分散表現の和を単語の分散表現とするモデルを採用しているため，レアな単語に対しても有効な分散表現を獲得することが可能である．分散表現は (1) タイトル (2) 巻号 (3) 著者 (4) 出版社 (5) ページ数 (6) 外形の 6 種のメタデータからそれぞれ 100 次元の分散表現を獲得し，600 次元の分散表現を得た．

#### 4.3.2　学　　習

距離学習はデータ間の類似度や距離といった計量を学習する手法である．意味が類似するデータは近く，非類似しているデータは遠くなるような特徴空間を学習していると言える．意味的な距離を考慮した特徴量空間の学習が可能となれば，未知のデータに対してもロバストに対応することが出来る．本研究では Siamese Network [10] を距離学習モデル $M_\Theta$ として採用し，最適な距離を学習するため，*Contrastive Loss* [11] と呼ばれる損失関数を使用する．*Contrastive Loss* は以下の式で表される．

$$\boldsymbol{Contrastive\ Loss} = \frac{1}{2}\left(\sum_{(S_i,S_j)\in\mathcal{D}} Euc(S_i,S_j)^2 - \sum_{(S_i,S_j)\notin\mathcal{D}} \max(M - Euc(S_i,S_j)),0)^2\right)$$

ただし，$M$ はマージンを表し，

$$Euc(S_i,S_j) = \|M_\theta(S_i) - M_\theta(S_j)\|_2$$

である．

学習には Keras[6] を使用した．以下に Siamese Network の構造を示す．

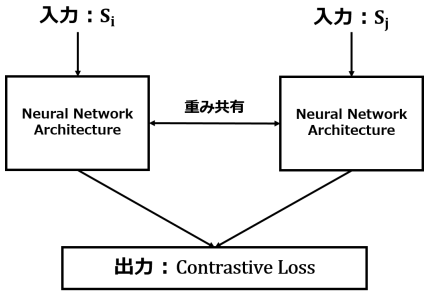



図 2　Siamese Network

また，Neural Network Architecture に使用したモデルは入力層，全結合層 1，ドロップアウト層，全結合層 2，ドロップアウト層，全結合層 3，ドロップアウト層，および出力層の順で構成される．全結合層と出力層は活性化関数に Relu 関数，損失関数には *ContrastiveLoss* を使用した．*ContrastiveLoss* が持つ距離 $D$ の算出には Neural Network Architecture の出力値である特徴ベクトルが使用されるため，間接的に Neural Network Architecture の重みを学習する．

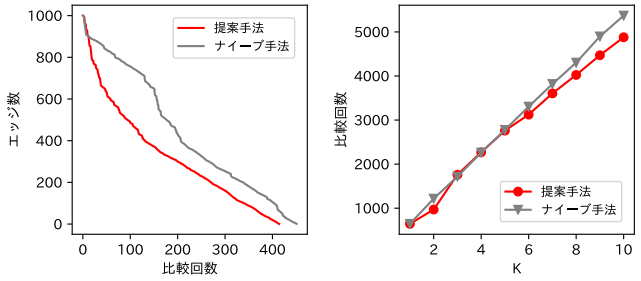
### 4.4　提案手法における評価実験

本章では，3 章で述べた提案手法の有効性を確認するための評価実験について述べる．実験は (1) 提案手法の比較回数の変化 (2) 提案手法によって獲得されるノード集合の再現率 (3) フレームワークの同定能力とタスク数についてそれぞれ確認を行う．同定能力の評価の指標は再現率と適合率の調和平均である F1 値を用いる．また比較対象は書誌データが持つノードを書誌 *id* の順に従って同定を行うナイーブな手法である．本実験では，*Human* は正しい判定をすることを仮定して行う．

#### 4.4.1　結　　果

図 3 に提案手法の比較回数の変化について示す．***Score*** の高いノードはエッジを多く共有していることから，優先的に同定を行うことで比較回数を削減することが可能である．また近傍探索数 K を大きくした場合にナイーブな手法に比べ比較回数の増加が小さいことを確認した．

6：https://fasttext.cc/




(a) エッジ数の変化　　(b) 比較回数の変化

図 3　**近傍探索数 $K$ とエッジ数に対する比較回数の変化**:(a) では $Score$ の高いノードからグラフを縮約することで，ランダムに選択したナイーブ手法と比較して比較回数を削減したことを示す．(b) 近傍探索数 K を大きくした場合，累計の比較回数に有意差が生じた．

本研究では K 近傍グラフのエッジに対して同定の判定を行う．このため，より高精度に同定を行うためには，マッチングすべきレコードのペアが K 近傍グラフのエッジにもれなく含んでいることが必要になる．図 4 は K 近傍グラフ構築時におけるマッチングペアの再現率である．再現率はデータの一致しているペアの集合である $\mathcal{D}$ に対し，K 近傍グラフにおけるエッジ集合をどれほど含むかを示すもので，以下のように表すことができる．

$$\mathcal{E}_{init} = \{(S_i, S_j) \mid S_j \in Nearest(S_i, K), S_i \in \mathcal{S}\}$$

$$Init\ Recall = \frac{|\mathcal{D} \cap \mathcal{E}_{init}|}{|\mathcal{D}|}$$

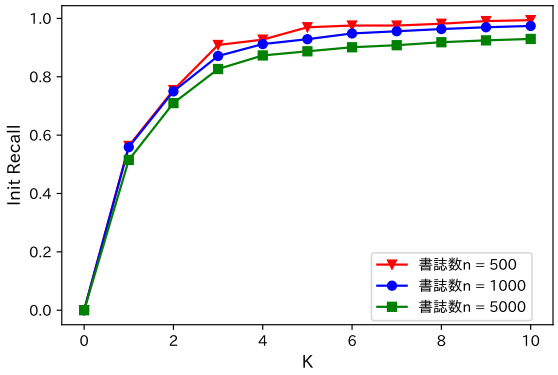



図 4　**K 近傍グラフ構築時の同定ペアの再現率**: K を大きくするほど，同定すべきペアの再現率が高くなる．

複数のデータ数で実験を行った結果，$K = 10$ で再現率は約 90%に達することを確認した．これより距離学習によって適切なデータの埋め込みを得られていること，K の値を適切に設定することで比較回数を削減することが可能である．

図 5 に $AI$ の閾値 $\alpha$ を変化させた場合の F1 値とワーカに割り当てたタスクの数を示す．クラウドソーシングによる同定シミュレーションの結果，$\alpha$ を大きくした場合に提案手法の F1 値が向上した．

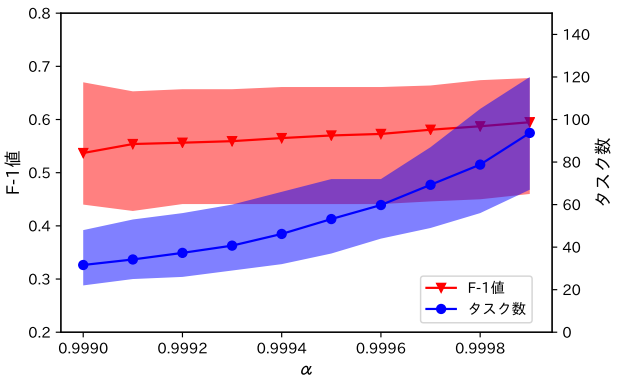



図 5　**$AI$ の閾値 $\alpha$ を変化させた場合の F 値とタスク数の変化**: $\alpha$ が大きいほどクラウドソーシングのタスク数が増加し，それに伴って F1 値も上昇する．

これより，$\alpha$ を大きくした場合にタスク数が増加し，フレームワークの性能を向上させることができる．一方で一般にタスク数が増えるほど金銭的なコストが大きくなるため，どのような閾値を用いるのが適切であるかは，フレームワークを適用する際の状況に依存する．

## 5　議　　論

提案手法の精度は，4.2 節で行った二値分類モデルによる書誌同定判定と，4.3 節で行った距離学習に依存する．本実験の結果より，(1) 二値分類モデルの精度 (2) 距離学習 (3) 提案フレームワークにおけるタスク割り当ての 3 つの課題が見つかった．

二値分類モデルの精度について，提案フレームワークは大きな割合で $AI$ の結果を採択するため，4.2 節で構築された二値分類モデルの同定精度に大きく依存する．本実験では，ほとんどの回答において高い確信度による判定が行われた．$AI$ が判定を間違えた結果を採択せずクラウドソースするためにはデータの細かな差分に対応可能な $AI$ モデルの構築が必要である．今後は様々な類似度のデータの学習や，各カラムに対する重みづけを考慮する必要がある．

エンティティマッチングを行う上で，全組み合わせを比較する場合，組み合わせ数は $\frac{n(n-1)}{2}$ 回 となり，データベースが大きくなるにつれて比較回数も大きく増加する．本実験では距離学習によりデータ間の関係を考慮した埋め込み空間を学習することで，図 4 に示すように $K = 10$ で再現率は 90%程度となった．これより 1 つのデータが比較する必要のあるデータ数を 10 程度に収めることができ，比較回数の大幅な削減を達成した．しかしながら上に示した $AI$ が判定を間違えたペアには距離が 0 である組み合わせが見受けられた．これは距離学習でメタデータが持つ意味を正確に表現することが出来ていないことが考えられる．例として，シリーズのある書誌が巻号の差分を表現できずに同じ位置に埋め込まれてしまうケースが存在した．今後は距離学習によって獲得されるデータの次元数を大きくし，どの次元数で適切にメタデータの意味を表現することが可能であるか確認する必要がある．またエンティティマッチン

グにおいて同定されうるデータの数は，エンティティの種類によって大きく差が生じる．提案手法では全てのデータに対して一律に同じ $K$ の値を採用したが，実際に行われている書誌同定や著者同定等のマッチングを考えるとあるエンティティには $K = x$ が十分である場合において，他のエンティティに対しては不十分であるということが考えられる．同定を行うデータ毎に適切な $K$ の値を割り当てることが可能であれば，更なる比較回数の削減が期待できる．

提案手法におけるタスク割り当てについて，本実験では $\alpha$ を限りなく大きくした場合の提案手法の F1 値は 0.6 程度であった．理由として *AI* が判定を誤ったペアデータに対して確信度の低い場合が極めて少なく，タスクが人間に割り当てられることが少なかったことが挙げられる．曖昧性の高いペアをクラウドソーシングタスクとして人間に割り当てるためには，二値分類モデルの性能向上に加え，予め曖昧性の高いデータを抽出するなど改善の余地が見られる．

## 6 まとめ

本論文では，共通のスキーマを持つ複数のデータベースを対象としたエンティティマッチング手法として，距離学習を用いた Human-in-the-loop エンティティマッチングフレームワークの提案を行った．提案手法の有効性を検討するため，国立国会図書館の総合目録を利用したいくつかの実験を行った．初めに，エンティティマッチングに対して二値分類がどの程度の性能を示すか検証した．次に，距離学習を用いて，同定候補集合の効率的な絞り込みに利用可能であるか検証を行った．最後に，提案フレームワークによるエンティティマッチングに取り組んだ．

提案手法は改善の余地はあるものの高い正解率を得た．二値分類モデルでは，エンティティマッチングには有効であるものの，タスク割り当て手法には改善の余地が見受けられた．距離学習の適用に関して，候補ペアの絞り込みに大きく貢献し比較回数を削減することができた．今後は次元数を変化させ，メタデータの持つ意味を適切に表現する次元数を獲得する必要がある．提案フレームワークを用いたエンティティマッチングでは，クラウドソーシングを組み合わせることで フレームワークのマッチング能力の向上を示すことが出来た．

今後の展望として，(1) *AI* と人間がどのようにタスクを分担するかというタスク割り当ての戦略の考案，(2) スパムワーカーを考慮したフレームワークの考案の 2 つを計画している．(1) については，エンティティマッチングにおいて，機械が処理を行う場合と人間が処理を行う場合とでは，マッチング精度と計算時間および金銭的コストがトレードオフの関係になる．これに対応して，マッチング精度とコストのバランスを最適にするためのタスク割り当ての戦略について考察することを計画している．(2) については，本実験においては，クラウドワーカーは常に正しい動作を行うと仮定したが，実際のクラウドソーシング環境を想定した場合，スパムワーカーの存在や，ワーカーの判定の間違いも考えられる．今後はこのようなクラウドワーカーの不確実な動作に対してロバストなフレームワークを構成することを計画している．

## 文　献


[1] Matthew A. Jaro. Advances in record-linkage methodology as applied to matching the 1985 census of tampa, florida. *Journal of the American Statistical Association*, Vol. 84, No. 406, pp. 414–420, 1989.

[2] William W. Cohen and Jacob Richman. Learning to match and cluster large high-dimensional data sets for data integration. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, p. 475–480, New York, NY, USA, 2002. Association for Computing Machinery.

[3] Harada Takashi, Fukushima Yukihiro, Sato Sho, Tsuruta Misato, Yoshimoto Ryuji, and Morishima Atsuyuki. Advancement of bibliographic identification using a crowdsourcing system. *Proceedings of the 9th Asia-Pacific Conference on Library & Information Education and Practice(A-LIEP 2019)*, pp. 71–82, 11 2019.

[4] Guoliang Li. Human-in-the-loop data integration. *Proc. VLDB Endow.*, Vol. 10, No. 12, p. 2006–2017, August 2017.

[5] Jiannan Wang, Tim Kraska, Michael J Franklin, and Jianhua Feng. Crowder: Crowdsourcing entity resolution. *arXiv preprint arXiv:1208.1927*, 2012.

[6] H. B. Newcombe, J. M. Kennedy, S. J. Axford, and A. P. James. Automatic linkage of vital records. *Science*, Vol. 130, No. 3381, pp. 954–959, 1959.

[7] 相澤彰子, 大山敬三, 高須淳宏, 安達淳. レコード同定問題に関する研究の課題と現状. 電子情報通信学会論文誌. D-I, 情報・システム, I-情報処理 = The transactions of the Institute of Electronics, Information and Communication Engineers. D-I, Vol. 88, No. 3, pp. 576–589, mar 2005.

[8] 桂井麻里衣, 大向一輝. 複数の異なる学術情報データベースを対象とした著者同定支援システムに関する検討. 第 10 回データ工学と情報マネジメントに関するフォーラム, 3 2018.

[9] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*, 2016.

[10] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. Signature verification using a "siamese" time delay neural network. In *Proceedings of the 6th International Conference on Neural Information Processing Systems*, NIPS'93, p. 737–744, San Francisco, CA, USA, 1993. Morgan Kaufmann Publishers Inc.

[11] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, Vol. 2, pp. 1735–1742, 2006.