エンティティ解決手法を応用したデータクリーニングのための不整合検出

大森　弘樹†　　清水　敏之††　　吉川　正俊††

† 京都大学大学院情報学研究科　〒 606–8501 京都府京都市左京区吉田本町
†† 京都大学大学院情報学研究科　〒 606–8501 京都府京都市左京区吉田本町
E-mail: †hiroki@db.soc.i.kyoto-u.ac.jp, ††{tshimizu,yoshikawa}@i.kyoto-u.ac.jp

あらまし　専門性の高いデータの分析や検索などを行うためには，クリーンなデータを必要とする場合が多い．単純な誤字や脱字は既存の手法により高い精度で検出可能であるが，単語が同じエンティティを指しているかどうかを判断するには，クリーニング対象のデータについての，辞書や専門家の知見といった外部知識の利用が必要である．
本研究は，関係データベースに格納された科学データのメタデータや論文の書誌情報といった専門性の高いデータを対象とし，エンティティ解決に用いられる技術を応用することで，関係データ内の同一のエンティティを指しているが表現が異なる，いわゆる表記ゆれなどの不整合な値の候補を検出する手法を提案する．本手法により検出された不整合な値の候補は，データ管理者に提示され，データ管理者の確認を踏まえたインタラクティブなデータクリーニングのために利用することを想定している．

キーワード　関係データ，データクリーニング，エンティティ解決，機械学習

1 はじめに

現在，ビッグデータの利用は官民共に盛んである．そうしたデータの利用分野の一つとして機械学習があり，機械学習の利活用もまた盛んである．機械学習やデータの分析を行うためには，空値の補填や値に一貫性をもたせるなどのデータ整備が重要となる一方で，データ整備は多大な労力を必要とする [1, 2].

本研究では，科学メタデータや論文の書誌情報といった専門性の高いデータを対象とし，エンティティ解決に用いられる技術を応用することで，関係データ内に存在している，同一のエンティティを指しているが表現が異なる，いわゆる表記ゆれなどの不整合な値の候補を検出する手法を提案する．本手法により検出された不整合な値の候補は，データ管理者に提示され，データ管理者の確認を踏まえたインタラクティブなデータクリーニングのために利用することを想定している [3–5].

本研究が主にクリーニングの対象としている値は，単純な誤字脱字などの誤りではなく，同義語や語順が異なる語などの単純な文字列のパターンでは捉えることの難しい不整合な値である．そのような不整合な値は，単純には辞書などの外部知識を利用することでクリーニングが可能である．しかし，辞書を作成するには，クリーニング対象とするデータの専門性が高い場合において，専門家の多大な労力が必要である可能性が高く，利用可能な外部知識を作成するために必要とされるコストが多大であることが多い．本研究はそうしたコストを軽減するために，データの値の中から不整合な値である可能性があるものを候補として発見し，人間，特にデータ管理者をユーザとして想定し，ユーザに不整合な値の候補の提示を行うことで，効率よく不整合な値の修正を行うことを目的とする．

修正に専門家を必要とするような不整合を含むデータの例を図 1 に示す．この例では 1 つの組が 1 つのデータに対応しており，それぞれの組に t_1 から t_4 の ID を割り当てている．カテゴリの列はそのデータの属するカテゴリを，提供者の列はそのデータの提供者の名前を，作成機関の列はデータを作成した機関の名前を，収録年はそのデータが収録された年を表している．図 1 の例では，提供者の欄について，t_1 とそれ以外の組とで姓名の順番が異なっており，不整合が起きている．こうした姓名の順番の判断などは辞書を用いるなどといった機械的な判断が難しい．また，機関の略称として “KU”が用いられているが，その略称に対応する機関名は “Kyoto univ.”と “Kobe univ.” の 2 種類が表の中に存在しており，これもまた機械的な判断が難しい．本研究では，こうした機械的に判断することが難しい不整合な値の候補を検出する手法を提案する．検出された不整合な値の候補は，ユーザに提示され，ユーザの確認を踏まえたインタラクティブなデータクリーニングのために利用されることを想定している．

	カテゴリ	提供者	作成機関	収録年
t_1:	ocean	Omori Hiroki	KU	2011
t_2:	forest	Hiroki OMORI	Kyoto univ.	2016
t_3:	soils	Taro YAMADA	Kobe univ.	2017
t_4:	ocean	Hiroki OMORI	Kobe univ.	2018

図 1: 不整合なメタデータの例

本研究は，データクリーニングの分野における既存手法を適用しづらいデータを対象に，データクリーニングを行うことを今後の目標の一つとしている．データクリーニングの分野における既存研究としては，属性間の従属性などを用いたルールベース手法が存在している [6]．しかし，ルールベース手法には recall が低くなるという欠点が存在している．その欠点を解決するため，機械学習を用いた手法が提案され，高い precision と recall でエラーを検出することが可能となっている [7]．機

械学習手法はラベル付けのコストがあり，クリーニング対象のデータの全体ではなく一部にラベル付けを行い，訓練データとして用いる．しかし，訓練データ外に未知のパターンの不整合が存在しているような状況も考えられ，そうした不整合は機械学習を用いた手法による検出が難しくなる．誤字脱字といった単純なエラーは，少ない訓練データでもデータ全体のエラーの傾向を掴むことが可能だが，表記ゆれなどの不整合な値は，訓練データに含まれていない場合，機械学習を用いた手法による対処が困難であると我々は考えた．

こうした訓練データ外に未知のパターンの不整合が存在してしまう状況では，データの一部を訓練データとする機械学習より，ルールベースの手法のほうが適しているが，先述の通り，ルールベースの手法には recall が低くなるという課題がある．

そのため，本研究はルールベース手法の欠点を克服するように，単語の分散表現と機械学習を用いることによって，通常のルールベース手法より更に柔軟で緩和された，recall が高くなるかわりに precision が低くなるようなルールを用いて不整合な値の候補を検出することを目的としている．低くなった precision は，検出された不整合な値の候補を，ユーザの確認を踏まえたインタラクティブなデータクリーニングに適用することで補うことができると考えている．

本論文の構成を以下に示す．第 2 節では本論文と関係している研究を紹介する．第 3 節では提案するデータクリーニング方式がどのようなものであるかを説明する．第 4 節では実際のデータに対して本手法を適用した実験の内容とその結果に加えて，今後の実験計画について示す．第 5 節では本論文の提案する手法の性質や能力についての議論を行う．第 6 節では本論文のまとめを述べ，今後の課題について議論する．

2 単語の分散表現を用いたエンティティ解決手法

本研究のエンティティ解決に関する手法の大部分は Muhammad らの研究 [8] を基にしている．この研究はタプルを分散表現することによって，タプルを対象にエンティティ解決することを目的としている．つまり，あるタプルともう一つのタプルが同一のエンティティを指しているか否かを判定することを目的としている．そのためにこの研究ではいくつかの手法が示されており，中でも本研究が参考にしている手法の概要を以下に示す．

この研究は，タプルを分散表現するために，まず，単語の分散表現を用いている．この研究は，GloVe [9] という単語をベクトルとして分散表現する手法により，ウェブアーカイブなどの一般的なコーパスから学習された学習済みの単語ベクトル集合を利用している．しかし，そうした学習済みのデータは出現頻度の少ない単語や造語，組織名や個人名は含まれていない場合が多い．そこで，既存の単語ベクトル集合に含まれていない未知の単語にも単語ベクトルを与えるため，クリーニング対象のテーブル上でその未知の単語が属する属性とタプルにおいて，共起頻度の高い単語を意味の近しい単語と考え，共起頻度の高い単語上位 k 件の単語ベクトルの単純加算平均をとったベクトルを，その未知の単語の初期単語ベクトルとして設定する．

そこから Retrofitting という手法を適用し，属性における共起単語同士，タプルにおける共起単語同士の距離が近づくよう単語ベクトル全てを修正する [10]．

これで単語ごとに単語ベクトルが設定されたが，一つの属性に複数の単語が存在する場合がある．その場合は属性に対して一つのベクトル表現となるよう，値を空白で区切り，それぞれの単語ベクトルを単純加算平均してできたベクトルを，属性について分散表現しているベクトルとする．その結果，一つのタプルに対して属性数分の分散表現が存在するようになる．属性数を m，単語ベクトルの次元数を d とすると $m * d$ 次元の行列ができることになる．

この行列を用いてタプル間の類似度を表すベクトルを作成する．タプルごとに各属性に一つずつ分散表現ベクトルが存在しているため，各属性ごとのベクトル同士でコサイン類似度を計算する．その結果，m 次元の，属性の値の類似度を各要素に持つ，類似度ベクトルが計算される．

こうして計算されたタプルのペア間の類似度ベクトルに対し，ニューラルネットを用いて，それぞれのタプルに対応しているエンティティが同一であるか否かの二値分類学習を行う．

本研究とこの関連研究との差異は主に 2 点存在している．1 点目はこの関連研究はタプルを対象にしてエンティティ解決することを目的としているが，本研究は着目属性の値を対象にしてエンティティ解決し，更にユーザに結果を提示することでデータクリーニングすることを目的としている点である．2 点目は本研究ではエンティティ解決を行って，その上で誤分類された結果を利用している点である．この関連研究がラベルを正しいものとしてエンティティ解決していることに対して，本研究はラベルに誤りが含まれている前提の下でエンティティ解決をしている．

3 提 案 手 法

3.1 入　　力

本研究は広く利用されていることから関係データをクリーニング対象とする．まず，関係データ形式のクリーニング対象であるテーブルを一つ入力に与える．また，不整合な値を発見したいとユーザが考えている着目属性を入力に与える．この着目属性の値に対してエンティティ解決を行っていく．さらに，クリーニング対象とするテーブルの単語を含む学習済みの単語ベクトル集合を与える．この単語ベクトル集合は，対象とするテーブルの全単語を含んでいる必要はなく，テーブルの単語の一部を含んでいるだけで良い．本研究は GloVe [9] という単語をベクトルとして分散表現する手法により，一般的な Web ページなどのコーパスから学習された学習済みの単語ベクトル集合を利用する．

3.2 エンティティ解決手法の応用

本研究はユーザが与えた着目属性の値をなんらかのエンティティに対応するラベルだと捉えてエンティティ解決を行う．つ

まり，あるタプルの着目属性の値ともう一つのタプルの着目属性の値が同一のエンティティを指しているか否かを判定する．2 節で述べた研究 [8] では，タプルの指すエンティティに着目してエンティティ解決を行っていたが，本研究ではタプルの着目属性の値に着目してエンティティ解決を行う．

また，着目属性以外の属性は，着目属性の値に対応するエンティティの属性だと捉えることとする．本研究はタプルのペアの着目属性以外の属性の値から，タプルのペアの着目属性の値が一致しているかどうか判定できるようにニューラルネットを訓練することで，エンティティ解決を行う．データベースのタプルを訓練データとテストデータに分割し，それぞれのデータセットの中のタプルから作られる全てのペアを学習や予測に利用する．

3.3 共起関係を用いた単語ベクトルの調整

本研究は 2 節で述べた研究 [8] と同様に，着目属性以外の属性の値を表現する手段として，単語の分散表現を利用する．GloVe によってウェブアーカイブから学習された学習済みの単語ベクトル集合を用いる．そこに含まれない未知の単語が属する属性とタプルにおいて，共起頻度の高い単語上位 10 件の単語ベクトルの単純加算平均をとったベクトルを，その未知の単語の初期単語ベクトルとして設定する．そこから Retrofitting という手法を適用し，属性における共起単語同士，タプルにおける共起単語同士の距離が近づくよう単語ベクトル全てを修正する [10].

3.4 タプル間の類似度

2 節で述べた研究 [8] と同様に，1 つの属性に複数の単語が存在する場合は，属性に対して一つのベクトル表現となるよう，値を空白で区切り，それぞれの単語ベクトルを単純加算平均してできたベクトルを属性を分散表現するベクトルとする．その結果，一つのタプルに対して着目属性を除いた属性数分のベクトルができる．属性数を m，単語ベクトルの次元数を d とすると $m-1*d$ 次元の行列ができることになる．

2 節で述べた研究では，こうしてできた行列を用いてタプル間の類似度を測ったが，本研究では着目属性の値が指すエンティティに対応している各属性間での類似度を測ることになる．つまり，着目属性以外の各属性の値に対しての類似度を測ることになる．タプルごとに着目属性を除いた各属性に 1 つずつ分散表現ベクトルが存在しているため，着目属性を除いた各属性ごとのベクトル同士でコサイン類似度を計算する．その結果，タプルのペアに対して，$m-1$ 次元の，属性の値の類似度を各要素に持つ，類似度ベクトルが計算される [8].

3.5 二値分類学習

こうして計算された類似度ベクトルに対し，ニューラルネットを用いて二値分類学習を行う．2 節で述べた研究 [8] では，タプルが指すエンティティが一致しているかどうかの学習を行ったが，本研究では着目属性の値が一致するかどうかの学習を行う．本研究での分類先の二値クラスはポジティブとネガティブで，入力として与えられたタプルのペアの着目属性の値が一致するならポジティブ，そうでないならネガティブとして分類する．ニューラルネットを用いて学習することにより，単純に値の類似度の比較を行うより柔軟に学習できることを期待している．

クリーニング対象とするテーブル内のタプルを訓練データとテストデータとに分割し，訓練データの中のタプルから作られる全てのペアを対象に学習を行う．また，訓練データの中のタプルから作られる全てのペアと，テストデータの中のタプルから作られる全てのペアを対象に，ラベルの予測を行う．本論文で実施した実験では学習の際のコストと精度の釣り合いが取れているため，全タプルからランダムに 20%取得したタプルを訓練データとして使用している．訓練データとして利用するタプルの割合が大きいほど予測の精度は向上するが，学習に要する時間は増大するため，大きなデータを対象とするには全てを訓練データとして利用できない状況もあると考えている．

3.6 ユーザへの提示

本研究はニューラルネットが学習した結果，訓練データとテストデータそれぞれに対して行う二値分類予測の誤分類について着目する．誤りの中でも特にポジティブと誤分類された値について着目する．

ポジティブとして誤分類されたタプルのペアは，ニューラルネットが着目属性以外の属性の類似度を見た結果，着目属性の値が一致していると判断したが，実際には着目属性には異なる値が入っていたタプルのペアである．エンティティ解決の観点から見ると，着目属性の値をエンティティに対応するラベル，その他の属性の値をエンティティに属する属性と見なせる．そのため，ポジティブとして誤分類されたタプルのペアの着目属性の値は，属性が類似していることから対応しているエンティティが同一であることが疑われるが，エンティエィに対応するラベルは異なっている状態と見なせる．よって，誤って同一のエンティティに対して異なるラベルが付与されている可能性があると考えることができる．従って，ポジティブに誤分類されたタプルのペアは，着目属性の値は同じエンティティを指しているのに，実際には誤って異なる表記の値が入力されている可能性がある．こういったポジティブに誤分類されたタプルのペアの着目属性の値を，表記ゆれなどの不整合な値の候補としてユーザに提示することにより，効率よく不整合な値を修正していけることを本手法は期待している．

4 実　　験

4.1 定性的評価

本研究は定性的な評価を行うために，データ統合・解析システム DIAS (Data Integration and Analysis System)[1] に提供されたデータに付属しているメタデータの一部を対象に提案手法の実施を行った．DIAS のメタデータは本来 XML 形式であるが，それを簡易的に 12 属性の関係データへと変換し提案する方式を適用した．一つの行が一つのデータセットに対応し

1 : https://www.diasjp.net

ており，今回は 426 データセットのメタデータに対して実験を行った．

このメタデータは，カテゴリ，制作された日時，提供されたデータセットの作成者，その所属機関，メタデータの著者やその所属機関を属性として持つ．DIAS は観測によって得られた地球各地での多様な観測データを収集しており，それに付随しているメタデータも多様であるため，科学データを対象にした実験用データとして有用である．

上記データの属性のうち，そのデータがどのようなカテゴリのデータであるかを示しているカテゴリの属性，そのデータのドキュメントを書いた著者名を表す著者名の属性，その著者の所属機関名を表す所属機関の属性，そのデータセット自体を作成した人の名前を表す作成者名の属性，その作成者の所属機関名を表す作成機関の属性，そして，データについて問い合わせを行う連絡先の機関名を表す連絡機関の属性に着目して提案手法を実施した．

入力として与える学習済みの単語ベクトル集合として，GloVe を用いて Common Crawl のウェブアーカイブから学習された学習済みの単語ベクトル集合 [2]を使用した．最初にユーザが着目属性とする属性を指定する必要性があるが，事前にデータを観察した際，不整合な値と疑われる値が多かった連絡機関の属性を着目属性として指定した．

全体の 20% のタプル学習用の訓練データに，全体の 80% のタプルをテストデータに用いた．また，バリデーションデータの比率を 20% に，学習率を 0.02 に設定し，セル数が 500 で最終的に二値分類を行う 2 層のニューラルネットを用いて 20 エポックの学習を行った．タプルの全てのペアに対するポジティブとネガティブの二値分類を学習し，着目属性である連絡機関の値が一致する場合はポジティブ，そうでなければネガティブに分類するように学習を行う．

訓練データに含まれるポジティブなタプルのペアは 1467 件，ネガティブなタプルのペアは 16813 件であった．テストデータに含まれるポジティブなタプルのペアは 2938 件，ネガティブなタプルのペアは 24557 件であった．訓練データに対する分類の正答率は約 99%，テストデータに対する分類の正答率は約 97% となった．訓練データに対し，誤ってポジティブと判断されたタプルのペアは 99 件，誤ってネガティブと判断されたタプルのペアは 155 件であり，テストデータに対し，誤ってポジティブと判断されたタプルのペアは 169 件，誤ってネガティブと判断されたタプルのペアは 533 件であった．

誤ってポジティブとして判断された，つまり，誤って同一の値だとニューラルネットに判断された着目属性の値のペアの一部を表 1 に示す．これらの値のペアは，本研究が不整合な値としてユーザに対して提示したいと考えている値である．この表にある確信度とは，二値分類に用いるニューラルネットの出力の値を指しており，1 に近づくほどポジティブとして確信しており，0 に近づくほどネガティブと確信していると解釈することが可能である．

また，組織名が null である場合，同じ値が入っているものとして学習を行っているが，今回は．不整合な値の候補を検出することが目的であるため，表 1 から null を含む値の不整合の候補を削除している．null の扱いについては，第 5 節にて議論を行う．null を含む候補を省いた 70 件程度のリストの中から，不整合な値ではないかと疑われる値がいくつか観察できた．学習の段階では全てを小文字にしカンマを削除しているが，表記が見づらくなるため，本来の大文字混じりの表記に戻した例を次に示す．

- “JAMSTEC/DrC” と “DrC/JAMSTEC”
- “Independent Administrative Institution Japan Agency for Marine-Earth Science and Technology, Institute of Observational Research for Global Change” と “Research Institute for Global Change, JAMSTEC”
- “Center for Environmental Measurement, National Institute for Environmental Studies” と “Center for Global Environmental Research, National Institute for Environmental Studies”
- “Forestry and Forest Products Research Institute” と “Cold Regions Environment Conservation Research Group, Hokkaido Research Center, Forestry and Forest Products Research Institute”

しかし，これらの値が真に同じエンティティを指しているかを判断するには，組織名全体に対して専門的な知見を持っている人間が必要であり，precision や recall などの定量的な評価は困難である．そのため，DIAS とは別のデータを対象にして定量的な評価を行う実験を現在計画している．

4.2 定量的評価実験の計画

不整合な値が存在していない整備されたデータに対してノイズを付与し，不整合な値を人工的に起こしたデータを用いて定量的に評価する実験を計画している．dblp[3]が提供している論文の書誌情報が格納された XML ファイルを，関係データ形式に格納したテーブルを対象に実験を行う予定である．データベースに関するトップカンファレンスや論文誌に採用された論文のデータを収集し，著者名に対して，ノイズを付与し人工的に不整合な値を発生させ，ユーザに提示する不審な値にどれくらい不整合な値が含まれているかについてや，ユーザに提示する不審な値が全体の不整合な値の中のどれくらいの割合をカバーできているかについての評価を行うことを計画している．

5 議　　論

本研究の手法を同じ値が極端に少ない属性を着目属性として実施すると，タプルのペアの着目属性の値が同じ値かそうでないかに対する正解率を上げるように二値分類の学習をさせているため，全てのペアに対して，着目属性の値が同じではないと判定するような望ましくない学習を起こす傾向にある．今後，どのようなデータに対して本研究の手法が有効であるかの評価

2 : https://github.com/stanfordnlp/GloVe

3 : https://dblp.uni-trier.de

表 1: 出力結果の一部

値 A	値 B	最大確信度	平均確信度
“jamstec/drc”	“jamstec/rigc”	0.979	0.979
“cptec/inpe”	“jamstec/drc”	0.945	0.945
“cptec/inpe”	“jamstec/rigc”	0.940	0.914
“jamstec/drc”	“drc/jamstec”	0.892	0.892
“institute of meteorology and hydrology”	“department of meteorology and hydrology”	0.853	0.853
“independent administrative institution japan agency for marine-earth science and technology institute of observational research for global change”	“research institute for global change jamstec”	0.836	0.836
“japan agency for marine-earth science and technology”	“tokyo metropolitan university”	0.827	0.823
“center for environmental measurement national institute for environmental studies”	“center for global environmental research national institute for environmental studies”	0.941	0.793

を行っていく必要性がある．また，本研究が既存手法に対してどのような差別化が図れているかについての評価を行うことも今後の課題である．ルールベースを用いた既存手法と機械学習を用いた既存手法の両方と比較する必要があると考えている．

本研究の手法の応用事例として，本論文では誤ってポジティブと分類したタプルのペアの着目属性の値をユーザに提示しているが，誤ってネガティブと分類したタプルのペアの着目属性の値に着目することも考えられる．誤ってネガティブと分類したタプルのペアの着目属性の値は，実際は同じ値であるのにニューラルネットは異なる値であると判断された値である．そのため，それらの値は詳細化して別の表現となっているべき値が，同一の表現にまとめられてしまっている値であるとみなすことが可能である．具体的な状況としては，同一の組織でありながら全く異なる分野のデータを収集しているいくつかの部署が同一のデータベースにデータを提供しており，提供しているデータに付属しているメタデータの提供機関の属性にはそれぞれの部署名が入力されていることが望ましいといった状況が考えられる．このような状況で，提供機関に組織名までしか入っておらず，部署名が記入されていない場合，それぞれの部署名を入力することによって表記の詳細化を行うように，ユーザに促すことが可能であると考えている．

また，null を用いた応用事例も考えられる．誤ってポジティブとして分類したタプルのペアの着目属性が，一方には値が入っていてもう一方は null であった場合，ニューラルネットはその値と null が同一であると判断しているため，null にその値が入るべきではないかとユーザに提案することも可能であると考えている．しかし，本来 null には異なる値が入っているが，同じ値であるとして学習を行うため，ニューラルネットの学習に悪影響を与えている可能性も存在している．

6 おわりに

本研究は単純な誤字脱字のようなエラーではなく，同じエンティティを指しているが値の表記が異なるような，いわゆる表記ゆれなどの不整合な値の候補をユーザに提示することで，効率よく不整合な値を修正していくことを目的とするデータクリーニング手法を提案した．また，実際のデータに対して提案手法を適用し，データクリーニングの対象としている不整合な値ではないかと疑われる候補が抽出できていることを確認した．

今後の課題としては，まず人工的に不整合な値を混入したデータに対して実験を行うことで定量的な評価を実施することや，本研究が有効に働くデータはどのようなものかを調査することが挙げられる．

謝　辞

本研究の一部は JSPS 科研費 JP17H06099, JP18H04093, JP18K11315 の助成を受けたものです．

文　献

[1] Sean Kandel, Jeffrey Heer, Catherine Plaisant, Jessie Kennedy, Frank van Ham, Nathalie Henry Riche, Chris Weaver, Bongshin Lee, Dominique Brodbeck, and Paolo Buono. Research directions in data wrangling: Visualizations and transformations for usable and credible data. *Information Visualization*, 10(4):271–288, 2011.

[2] Tamraparni Dasu and Theodore Johnson. *Exploratory data mining and data cleaning*, volume 479. John Wiley & Sons, 2003.

[3] Maksims Volkovs, Fei Chiang, Jaroslaw Szlichta, and Renée J Miller. Continuous data cleaning. In *Data Engineering (ICDE), 2014 IEEE 30th International Conference on*, pages 244–255. IEEE, 2014.

[4] Jian He, Enzo Veltri, Donatello Santoro, Guoliang Li, Giansalvatore Mecca, Paolo Papotti, and Nan Tang. Interactive and deterministic data cleaning. In *Proceedings of the 2016 International Conference on Management of Data*, SIGMOD ’16, page 893907, New York, NY, USA, 2016. Association for Computing Machinery.

[5] Protiva Rahman, Courtney Hebert, and Arnab Nandi. Icarus: minimizing human effort in iterative data completion. *Proceedings of the VLDB Endowment*, 11(13):2263–2276, 2018.

[6] Theodoros Rekatsinas, Xu Chu, Ihab F Ilyas, and Christopher Ré. Holoclean: Holistic data repairs with probabilistic inference. *Proceedings of the VLDB Endowment*, 10(11):1190–1201, 2017.

[7] Alireza Heidari, Joshua McGrath, Ihab F. Ilyas, and Theodoros Rekatsinas. Holodetect: Few-shot learning for error detection. In *Proceedings of the 2019 International*

Conference on Management of Data, SIGMOD '19, page 829846, New York, NY, USA, 2019. Association for Computing Machinery.

[8] Muhammad Ebraheem, Saravanan Thirumuruganathan, Shafiq Joty, Mourad Ouzzani, and Nan Tang. Distributed representations of tuples for entity resolution. *Proceedings of the VLDB Endowment*, 11(11):1454–1467, 2018.

[9] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics.

[10] Manaal Faruqui, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard Hovy, and Noah A. Smith. Retrofitting word vectors to semantic lexicons. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1606–1615, Denver, Colorado, May–June 2015. Association for Computational Linguistics.