ファインチューニング技術を用いた名寄せタスクの実装検証

張 龍傑 [1]
[1] 株式会社レゾナック
zhang.longjie.xljrh@resonac.com

概要

本研究は、大規模顧客名辞書を対象とした名寄せ（エンティティマッチング）タスクにおいて、BERT系モデルのファインチューニング手法を比較し、性能と計算コストのバランスを検証したものである。従来広く用いられてきた多クラス分類モデル（BertForSequenceClassification）は、分類クラス数が非常に多い場合に精度低下が生じることが知られている。そこで本研究では、分類モデルによる直接予測に加え、Sentence-BERT を用いたベクトル類似度ベースの名寄せ手法を検討した。

顧客名辞書から同一ラベルのペア（Positive Pair）と、ランダムまたは類似度上位から抽出した異なるラベルのペア（Negative Pair）を作成し、複数のファインチューニング方式を比較した。特に、類似度が高いがラベルが異なる Hard Negative Pair を用いた Sentence-BERT は、全手法の中で最高精度を達成した。一方、分類モデルを用いたベクトル類似度方式は精度はやや劣るものの、学習コストを大幅に削減できることが確認された。

1 はじめに

名寄せタスク（エンティティマッチング）は、商品情報や会社名などのデータに含まれるレコード同士が同一の実体かどうかを判定するタスクである。従来はルールベースの手法が多く研究されていました [1] が、近年では Transformer ベースのモデルを用いたエンティティマッチングが盛んに行われている。本タスクに関する研究の多くは Sentence-BERT を利用しており、参考文献 [2]、[3]、[4] では、意味が類似するデータペア（Positive Pair）と意味が異なるデータペア（Negative Pair）を教師信号として Sentence-BERT を学習させ、類似関係を検出している。しかし、データ数が増えるとペアの総数は入力データ数の二乗に近いオーダー $O(n^2)$ となり、学習コストが非常に大きくなる。例えば、レゾナックでは月に一度、顧客名の名寄せ作業を行い、各顧客名に所属するグループ会社名を付与している。これまでの作業で得られた、所属グループ名が付与された顧客名の辞書（表 1）には、顧客名が 113,000 件、グループ名が 46,808 件含まれている。もしすべての同一ラベルの組を Positive Pair、異なるラベルの組を Negative Pair として総当たりで作成すると、教師データの件数はおおよそ $\frac{n(n-1)}{2}$（ここで n=113,000）となり、約 6.4×10^9 件に達する。したがって、全組合せを用いた学習は現実的ではなく、別の方法を検討する必要がある。

表 1 顧客名–グループ名辞書

顧客名	グループ名
株）レゾナック山崎事業所	レゾナック
昭和電工ガスプロダクツ株式会社	レゾナック
Resonac Corporation	レゾナック
日産自動車	日産自動車
トヨタ自動車	トヨタ自動車
⋮	⋮

本研究では、すべての Positive Pair と一部の Negative Pair を選択して教師データを作成し、Sentence-BERT で学習する。さらに、一部の Negative Pair だけでも教師データの量は膨大になるため、もう一つ低コストな学習方法を提案する。具体的には、BertForSequenceClassification のファインチューニング機能を用いることで、明示的に Positive Pair と Negative Pair を大量に作成せずに学習可能であり、性能はやや低下するものの、計算コストを大幅に削減できることを示した。

新しい顧客名に所属グループ名を付与する作業は、46,808 クラスへの分類問題（multi-label classification）に相当する。AI を用いて自動分類する際に考えられる手法は、主に次の二つである。

- 手法 1：辞書を用いて分類モデルを学習し、そ

のモデルで新しい顧客名に所属グループ名を付与する方法。
- 手法 2：エンコーダーモデルで名寄せ対象の顧客名をベクトル化し、辞書中で最も類似する顧客名を探索し、その顧客名の所属グループ名を付与する方法。

手法 1 に関して、分類モデル BertForSequenceClassification[1] を用いて事前学習済みの BERT モデルをファインチューニングして分類を行う手法は、多くの研究で利用されている。しかし、分類クラス数が多い場合には性能が低下することが指摘されている。手法 2 に関して、本研究で使用するエンコーダーモデルは、東北大学が開発した BERT[2] である。なお、ファインチューニングの方法により複数のモデルを用意する。

2 名寄せシステム

本節では、まず前節に紹介した名寄せ手法 2 について詳しく紹介し、次に名寄せ用モデルのファインチューニングについて述べ、最後に、教師データの作成方法について述べる。

2.1 名寄せ方法

手法 2 では、ファインチューニングした BERT モデルに顧客名を入力し、そのベクトル表現を取得する。本研究では、得られたベクトルに対して CLS Pooling を適用し、入力トークン列の先頭に挿入された特殊トークン [CLS] の出力を採用する。その後、辞書に登録されている顧客名の中から、名寄せ対象の顧客名と最も類似度が高い顧客名を探索し、その顧客名に対応するグループ名を付与する。

2.2 事前学習済みモデルのファインチューニング

本研究では、以下の 5 種類のファインチューニング手法を提案する。各手法は、Sentence-BERT または BERT による分類モデル（BertForSequenceClassification）を基盤とし、教師データとして Positive Pair と Negative Pair を用いる。Negative Pair は、ランダムに選択したもの（Random Negative Pair）と、類似度が高いもの（Hard Negative Pair）の 2 種類を用意する。

Sentence-BERT 用教師データの構成

- **Positive Pair**
  顧客名辞書に含まれる同一ラベルの顧客名ペア。総数は 1,223,266 件。
- **Random Negative Pair**
  異なるラベルの顧客名ペアをランダムに抽出（重複なし）し、5,000,000 件生成[3]。
- **Hard Negative Pair**
  異なるラベルの顧客名ペアのうち、事前学習済み BERT によるベクトル類似度が高いものを抽出。抽出方法：顧客名 113,000 件をバッチ化し、各バッチ内で類似度上位 100 件を選択。累計 5,000,000 件に達した時点で抽出を終了[4]。このようなペアは、既存研究 [5], [6], [7] において Hard Negative Pair と呼ばれる。

2.2.1 Classification

顧客名辞書に含まれる顧客名 113,000 件とグループ名（ラベル）46,808 件を用いて、BertForSequenceClassification をファインチューニングする。

2.2.2 Sentence-BERT + Random Negative Pair

Positive Pair と Random Negative Pair を用い、顧客名ペアの CLS ベクトル間類似度に基づく損失関数で Sentence-BERT をファインチューニングする。

2.2.3 Sentence-BERT + Hard Negative Pair

Positive Pair と Hard Negative Pair を用い、類似度が高いがラベルが異なるペアに対して、類似度を低くするよう Sentence-BERT をファインチューニングする。

2.2.4 One Sentence + Random Negative Pair

Positive Pair と Random Negative Pair を用いた教師データを、2 つの顧客名を [SEP] トークンで結合し 1 つの文として入力する。Positive Pair にはラベル 1、Negative Pair にはラベル 0 を付与し、BertForSequenceClassification をファインチューニングする。

1） https://huggingface.co/docs/transformers/model_doc/bert
2） https://huggingface.co/tohoku-nlp/bert-base-japanese-whole-word-masking
3） 件数は、Azure の GPU VM（Standard NC8as T4 v3）を用い、2 エポックで約 40 時間の学習を見込んで決定した
4） Negative Pair の総数は非常に多いため、すべてのペアの類似度を計算することは不可能でこの方法を採用した

表 2　評価用データ

部署名	顧客名件数 (重複なし)
部署 A	446
部署 B	336
部署 C	369
部署 D	60
部署 E	161

2.2.5　One Sentence + Hard Negative Pair

Positive Pair と Hard Negative Pair を用いた教師データを、同様に [SEP] トークンで結合し、ラベル 1（Positive）、ラベル 0（Negative）を付与する。BertForSequenceClassification をファインチューニングする。

ファインチューニング方法 2.2.1 は、学習コストが最も低い学習方法であり、また、下記の例から、Sentence-BERT のように Positive Pair や Negative Pair を作成しなくても、同一ラベルを持つ顧客名のベクトルは互いに近づき、異なるラベルを持つ顧客名のベクトルは遠ざかる効果が得られる[5)]。

2.3　評価用データと評価指標

本名寄せタスクの評価用データには、レゾナック社内の 5 つの部署の実績データを利用する。評価用データの顧客名には、日本語・英語のほか、中国語、タイ語、韓国語も含まれる。データの構成は表 2 に示す。正解ラベルは担当者が付与する。評価指標には、多クラス分類タスクでよく用いられる Accuracy を使用する。

3　試験の手順

3.1　前処理

顧客名の表記ゆれを減らすため、以下の簡単な前処理を行う。

1. 全角の英数字を半角に変換する。
2. 半角カタカナを全角カタカナに変換する。
3. 発音が近いカタカナを統一する（例："ヴァ"→"バ"、"ヴィ"→"ビ"、"ウ゛"→"ブ"、"ヴぇ"→"べ"、"ヴォ"→"ボ"）。
4. 各種記号の表記を統一する。

情報の損失を最小限に抑えるため、上記以外の処理（例：数字の削除など）は行わない。

5)　詳細は付録 A に参照する

3.2　モデル学習

本研究では、2.2 節で述べた 5 種類のファインチューニング手法（Sentence-BERT + Random Negative Pair、Sentence-BERT + Hard Negative Pair、One Sentence + Random Negative Pair、One Sentence + Hard Negative Pair、Classification）を用いて BERT 系モデルの学習を行う。学習は Azure の GPU VM（Standard NC8as T4 v3）上で実施した。

各試験の詳細は以下の通りである。

- Sentence-BERT + Random Negative Pair および Sentence-BERT + Hard Negative Pair

  Positive Pair（1,223,266 件）と Negative Pair（Random または Hard、5,000,000 件）を合わせた計 6,223,266 件のペアを用いて Sentence-BERT をファインチューニングした。学習は 2 エポックで行い、合計の学習時間は約 43 時間であった。
- One Sentence + Random Negative Pair および One Sentence + Hard Negative Pair

  上記と同じ Positive Pair と Negative Pair のそれぞれのペアを、[SEP] トークンで結合して「1 つの文」として入力する形式に変換した教師データ（計 6,223,266 件）を用い、BertForSequenceClassification をファインチューニングした。学習は 1 エポックで行い、学習時間は約 32 時間であった。
- Classification

  顧客名辞書に含まれる顧客名 113,000 件を入力、グループ名（クラス）46,808 件を出力ラベルとする多クラス分類タスクとして BertForSequenceClassification をファインチューニングした。学習は 100 エポックで実施し、合計の学習時間は約 25 時間であった。

3.3　予測

手法 1 では、ファインチューニング方法「Classification」により学習した分類モデルを用いて、顧客名のラベルを直接予測する。

手法 2 では、ファインチューニングなしの BERT モデルおよび 2.2 節で述べた 5 種類のファインチューニング手法で学習した BERT モデルに顧客名を入力し、そのベクトル表現を取得する。取得したベクトルに基づき、辞書に登録されている顧客名の中から、名寄せ対象の顧客名と最も類似度が高い顧

表 3　性能検証

手法	ファインチューニング方法	学習時間	部署 A	部署 B	部署 C	部署 D	部署 E
手法 1: 顧客名を分類モデルに入力しラベルを予測する	Classification	25 hour/100 epoch	0.161	0.083	0.136	0.167	0.130
手法 2: 名寄せ対象の顧客名と最も類似度が高い顧客名のグループ名を付与する	Classification	25 hour/100 epoch	0.825	0.830	0.694	0.883	0.938
	Sentence-BERT + Random	43 hour/2 epoch	0.879	0.887	0.713	0.900	0.925
	Sentence-BERT + Hard	43 hour/2 epoch	**0.888**	**0.899**	**0.726**	**0.917**	**0.957**
	One Sentence + Random	32 hour/1 epoch	0.861	0.866	0.675	0.867	0.894
	One Sentence + Hard	32 hour/1 epoch	0.787	0.809	0.648	0.800	0.814
	ファインチューニングなし		0.805	0.821	0.656	0.783	0.820

客名を探索し、その顧客名に対応するグループ名を付与する。

4　性能評価

前節で紹介したレゾナック社内の 5 つの部署のデータに対して、前述の 7 つの手法の性能評価を行った。

結果を表 3 に示す。5 つの部署すべてにおいて、Sentence-BERT でファインチューニングした BERT モデルを用い、手法 2 で名寄せを行う方法が最も高い性能を示した。特に、教師データに Hard Negative Pair を用いた場合、全手法の中で最高の性能を達成した。

一方、手法 1 では「Classification」により学習した分類モデルを用いて顧客名のラベルを直接予測したが、すべての部署で性能が低かった。しかし、同じモデルを用いても、手法 2 のようにベクトル類似度に基づく名寄せを行うと、比較的良好な性能が得られた。

One Sentence 方式の教師データに関しては、Random Negative Pair を用いた場合は 4 番目に高い性能を示したが、Hard Negative Pair を用いた場合は性能が低下した。このことから、One Sentence 方式で学習した BERT モデルをベクトル類似度による名寄せに用いるのは適切でない可能性がある。

これは、付録に説明した BertForSequenceClassification が「同一ラベルの文を近づけ、異なるラベルの文を遠ざける」ように学習する一方で、Negative Pair の 2 つの顧客名を [SEP] トークンで結合して 1 つの文として入力すると、モデルがそれらを近い表現として扱ってしまうため、性能が低下したと考えられる。

なお、別の予測方法として、予測対象の顧客名と辞書内の顧客名を [SEP] トークンで結合し、モデルに入力して最も高い確率を出した顧客名のグループ名を付与する方法も考えられる。しかし、辞書には約 11 万件の顧客名が登録されており、1 件の予測に対して全件との比較が必要となるため、計算コストが非常に高くなるという課題がある[6)]。

5　おわりに

本研究では、複数の方法でファインチューニングした BERT モデルを利用し、ベクトル類似度に基づくレゾナック社内の顧客名の名寄せタスクにおける性能検証を行った。顧客名辞書に登録されている顧客名の数が多いため、Sentence-BERT による学習時に全てのペアを用いることは現実的ではない。そこで、一部の Negative Pair を選定してファインチューニングを行った結果、良好な性能が得られた。特に、Negative Pair の中から類似度が高い Hard Negative Pair を用いた場合、最も高い性能を示した。

一方、BertForSequenceClassification でファインチューニングしたモデルを用いてベクトル類似度に基づく名寄せを行った場合、Sentence-BERT と比較すると精度はやや低下するものの、学習コストが低いという利点がある。

6)　本研究で使用している Azure の GPU VM（Standard NC8as T4 v3）では、1 件の予測に 20 分以上を要する。

謝辞

本研究に際し、多くの貴重なご助言を賜りました高仁子氏、川原悠氏、辰巳大祐氏、岡田拓明氏に、心より感謝申し上げます。また、東京科学大学の岡崎直観先生には、終始ご指導とご助言を賜り、厚く御礼申し上げます。

参考文献

[1] L. Li, J. Li, and H. Gao. Rule-based method for entity resolution. **IEEE Transactions on Knowledge and Data Engineering**, Vol. 27, No. 1, pp. 250–263, 2014.

[2] 石井佑樹, 佐々木稔. Sentence-bert と語義定義文を利用した語義間の類義判定手法. 言語処理学会 第 30 回年次大会 発表論文集 (2024 年 3 月), pp. 1–5, 2024.

[3] 加納渉, 竹内孔一. Sentence-bert を利用した faq 検索におけるデータ拡張手法. 言語処理学会 第 28 回年次大会 発表論文集 (2022 年 3 月), pp. 1–5, 2022.

[4] Y. Li, A. Doan Y. Suhara, and W. Tan. Deep entity matching with pre-trained language models. **https://arxiv.org/pdf/2004.00584**, pp. 1–15.

[5] J. Robinson, C. Chuang, S. Sra, and S. Jegelka. Contrastive learning with hard negative samples. **https://arxiv.org/pdf/2010.04592**, pp. 1–28, 2021.

[6] R. Jiang, T. Nguyen, P. Ishwar, and S. Aeron. Supervised contrastive learning with hard negative samples. **https://arxiv.org/pdf/2209.00078**, pp. 1–8, 2024.

[7] W. Liu, Z. Yang, C. Li, Z. Hong, J. Ma, Z. Liu, L. Zhang, and F. Huang. Contrastive learning with hard negatives for sentence embeddings. **Applied Soft Computing**, Vol. 184, pp. 1–14, 2025.

A　分類モデル学習による埋め込みベクトルの改善効果

事前学習済み BERT によるベクトル化と、2.2.1 で示した手法でファインチューニングした後の BERT によるベクトル化の変化を比較した。検証対象は、トヨタ公式サイトおよび「かいとビジネス」に掲載された日立製作所、旭化成、日産自動車、レゾナックのグループ会社名、計 166 件である[7)]。会社名の全体情報を表す [CLS] トークンのベクトルを取り出し、UMAP で二次元に写像してプロットした結果を図 1 および図 2 に示す。

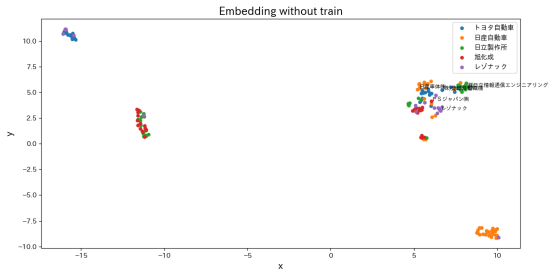

図 1　学習前の会社名

図 1 は、事前学習済みの BERT モデルで取得したベクトルをプロットした結果である。全部の会社名は 4 つのクラスタに分かれていて、それぞれのクラスタには違うグループに属する会社名が混在してしまっている。

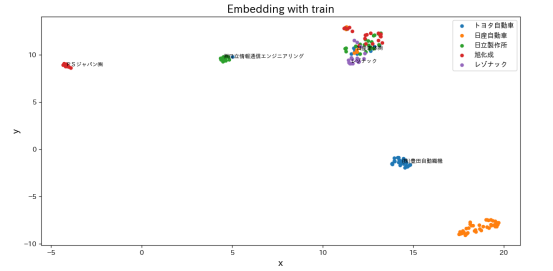

図 2　学習後の会社名

図 2 は 2.2.1 で示した手法でファインチューニングした BERT モデルによるベクトルをプロットした結果である。5 つのクラスタのうち 1 つはまだクラスタが混合した状態にあるが、残りの 4 つのクラスタには同じグループに所属する会社名がきれいにまとまっている。

7)　https://global.toyota/jp/company/profile/toyota-group/, https://kaito-business.com/archives/1594, https://kaito-business.com/archives/1609, https://kaito-business.com/archives/1999