人材業界固有の表現を考慮した求人票のマルチラベル分類

原 龍昊 [1] 林 勝悟 [1]Dat P.T. Nguyen[1]
[1] 株式会社ビズリーチ
{longhao.yuan, shogo.hayashi, dat.nguyen}@bizreach.co.jp

概要

人材業界では，求人票や履歴書といった様々なテキストデータが存在し，その分類や予測が行われている．しかし，人材業界のテキストには専門用語などの固有の表現があるため，単純なテキスト分類モデルでは適切な分類が難しい．

本研究では，求人票の職種と業種のマルチラベル分類タスクに取り組む．まず，人材業界固有の表現に対応するために，求人票データを用いてドメイン特化 BERT モデルを訓練する．次に，無関係なテキストの影響を低減するために，文単位に分割した求人票のテキストと各カテゴリの埋め込みを BERT モデルを用いて生成し，各カテゴリとの最大類似度ベクトルに基づいてマルチラベル分類を行う．求人票データセットを用いた評価を行い，提案手法は比較手法よりも優れた結果を出し，分布シフトの状況下でも有効に働くことを確認した．

1 はじめに

人材の採用や管理を行う人材業界では，求人票や履歴書といった様々なデータが存在する．これらは主にテキストやカテゴリを持つ半構造化データとして表現され，システムで管理されている．例えば，求人票はポジション名や仕事内容といったテキストフィールドと，職種や業種といったカテゴリフィールドからなる．仕事内容を単一のカテゴリで表現することは難しいため，職種や業種はしばしば複数のカテゴリを持つマルチラベルとして表現される．

求人票のデータ形式はシステムごとに異なるため，あるシステムで管理されている求人票を異なるシステムで扱うためには，その形式に合わせて新たにデータを登録する必要がある．しかし，職種や業種はそれぞれ多数のカテゴリを持つため，システム間のカテゴリの対応関係を人間が正しく理解することは容易ではない．また，専門的な職種や業種には専門用語や固有の表現が存在するため，単純なテキスト分類手法では高い分類精度を達成することが難しい．例えば，求人票に「機械学習ライブラリの経験がある」という文がある場合，通常の手法は「ライブラリ」という単語を誤って書籍などの異なる文脈で意味を解釈してしまう可能性がある．

本研究では人材業界固有の表現を考慮した求人票のマルチラベル分類に取り組む．まず，人材業界固有の表現を学習するために，求人票データを用いて BERT モデルを訓練する．次に，無関係なテキストの影響を低減するために，文単位に分割した求人票のテキストとカテゴリテキストの埋め込みを訓練した BERT モデルを用いて生成する．それらの類似度行列を求め，各カテゴリとの最大類似度ベクトルに基づいて求人票のマルチラベル分類を行う．求人票データを用いた評価を行い，提案手法は比較手法よりも高い性能を発揮したことを確認した．

2 関連研究

テキスト分類は自然言語処理分野の最も基本的なタスクの一つである．基本的なテキスト分類手法はまずテキストをベクトルに変換し，そのベクトルを分類モデルを用いて分類する．テキストのベクトルへの変換には，TF-IDF や BM25，BERT[1] といった手法を用いることができる．分類器として，ロジスティック回帰や Random Forest，Neural Networks などを用いることができる [2].

機械学習や自然言語処理の技術を人材業界のデータに応用する研究は多数存在する．Luo らの研究 [3] では，T5 モデルを用いた質問応答タスクを通じて履歴書から学歴やスキルといった情報を抽出する試みが行われた．BERT モデルを用いた求人票からキーワード抽出 [4] や，求人票と履歴書のマッチング [5] といった研究も存在する．Decorte ら [6] は，BERT モデルを用いて履歴書の業界分類と求職者のキャリア予測を行った．日本語の求人テキストを用いた研究として，質問応答タスクを通じて求人情報を含むメール文書データから構造化データを抽出す

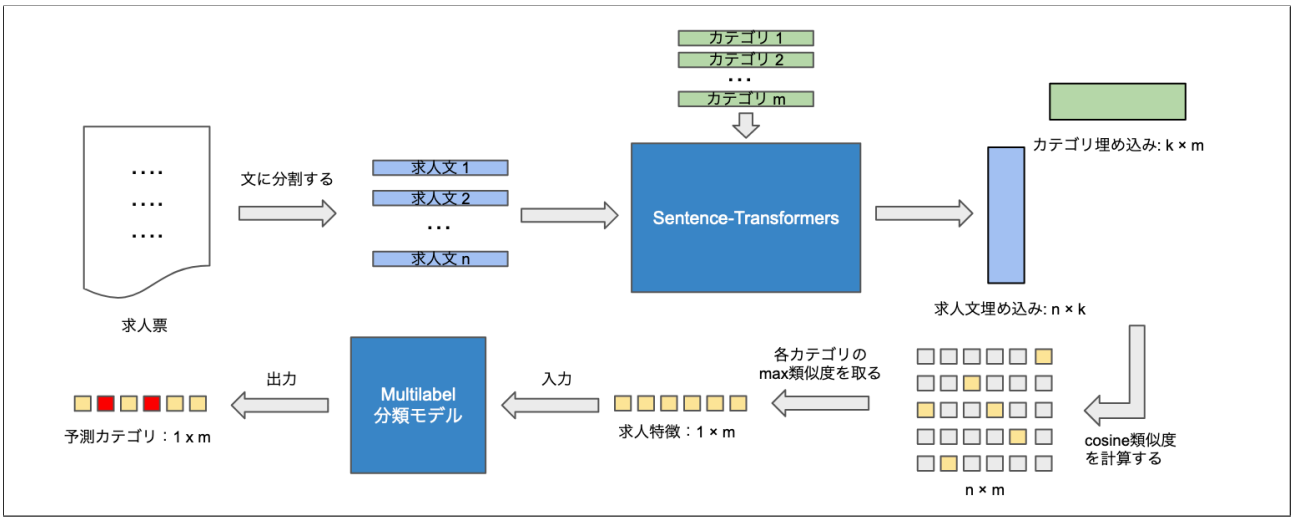

図 1　提案手法の流れ．

る研究も存在する [7].

近年，ChatGPT をはじめとする大規模言語モデルが様々なタスクで成功を収めている [8]．しかし，大規模言語モデルを大量の求人票データに適用するためには多大なコストが必要であり，API の利用にはデータのセキュリティに関する懸念も存在する．そのため，本研究では比較的低コストかつセキュリティが確保された環境で動作するモデルに焦点を当てる．

3　提案手法

提案手法の枠組みは，求人票データを用いたドメイン特化 BERT モデル (求人 BERT と呼ぶ) の訓練，求人票のテキストとカテゴリテキストの類似度ベクトルの生成，マルチラベル分類からなる．枠組みの全体の流れを図 1 に表す．

3.1　求人票データ

本研究では「ビズリーチ」[1)]に登録されている求人票データの職種と業種のマルチラベル分類タスクに取り組む．職種のカテゴリは 35 種類，業種のカテゴリは 18 種類存在する．各求人票の職種と業種はそれぞれ 1 つ以上のカテゴリを持つ．テキストフィールドとして，ポジション，募集条件，仕事内容，募集企業を持つ．職種と業種の分類に関係するポジション，募集条件，仕事内容のテキストを結合して 1 つのテキストとする．

1)　「ビズリーチ」は株式会社ビズリーチが運営するスカウト型転職サイトである．企業が求人票を添付したスカウトを求職者に送信し，求職者がそのスカウトに返信することで面談や面接などのマッチングが生じる．

3.2　求人 BERT の訓練

求人票の埋め込みモデルとして Sentence BERT [9] を用いる．まず，求人票データを使用して，マスク言語モデルにより BERT モデルの事前訓練を実施する．次に，類似するデータの組であるポジティブペア，類似しないデータの組であるネガティブペア，ポジティブペアとの識別が難しいネガティブペアであるハードネガティブペアを用いて BERT モデルの訓練を行う．経験的に，職種が同じ求人票は類似する傾向にある．また，同一求職者に送られるスカウトの求人票は類似する傾向にあり，また同一求職者が返信するスカウトの求人票も類似する傾向にある．そこで，以下の条件を持つ求人票の組をポジティブペア，ネガティブペア，ハードネガティブペアとする．

- **ポジティブペア**
  - 同じ職種を持つ
  - 同じ求職者にスカウト送信，返信されている
- **ネガティブペア**
  - 同じ職種と業種を持たない
  - 同じ求職者から返信されていない
  - 募集企業が異なる
- **ハードネガティブペア**
  - 同じ職種と業種を持つ
  - 同じ求職者にスカウト送信されていない
  - 異なる求職者にスカウト送信，返信されている
  - 募集企業が異なる

3.3 類似度ベクトルの生成

求人票のテキストには，例えば企業の所在地や沿革など，職種や業種に特別関係しない冗長な文章が含まれていることが多々ある．このような文章は，分類時のノイズとなるだけでなく，BERT モデルの入力トークンを消費してしまい，他の重要な文章が入力されなくなる可能性がある．そのため，求人票のテキストを文単位に分割してから求人票の埋め込み表現を求める．

提案手法は分類先のカテゴリテキストも活用するために，カテゴリテキストの埋め込みと求人票の埋め込みの類似度に基づいてマルチラベル分類を行う．求人票のテキスト d を文単位に分割して n の文集合 $\{s_1, s_2, \ldots, s_n\}$ を得る．BERT モデル E により，各文と m のカテゴリテキスト $\{c_1, c_2, \ldots, c_m\}$ の埋め込み $E(s_1), E(s_2), \ldots, E(s_n)$, $E(c_1), E(c_2), \ldots, E(c_m)$ を生成し，これらの類似度行列 $\mathbf{X}$ を求める．本研究では，文 s_i とカテゴリテキスト c_j の類似度として以下の cosine 類似度を用いる．

$$X_{ij} = \frac{E(s_i) \cdot E(c_j)}{\|E(s_i)\| \|E(c_j)\|}$$

各カテゴリごとに最大類似度を求め，d の特徴量として m 次元ベクトル $\mathbf{x}$ を得る．

3.4 マルチラベル分類

本研究では大量の求人票データを処理できるように，軽量なロジスティック回帰モデルをベースモデルとする一対他分類器を採用する．求人票とカテゴリの類似度ベクトル $\mathbf{x}$ に対して，関数 f_j $(j \in [m])$ を適用して，カテゴリ j が割り当てられる確率を以下のように求める．

$$f_j(\mathbf{x}) = \sigma(\mathbf{w}_j \cdot \mathbf{x} + b_j)$$

ここで，$\mathbf{w}_j$ は m 次元の重みベクトル，b_j はバイアス項，σ はシグモイド関数である．訓練データセットを用いてパラメータ $\mathbf{w}_j, b_j$ $(j \in [m])$ を最適化する．

「ビズリーチ」の求人票は 1 つ以上のカテゴリを持つが，一対他分類器は不均衡データの問題により，各カテゴリの分類確率が低くなる傾向がある．そこで，推論時に分類確率がハイパーパラメータ $\epsilon > 0$ より高いカテゴリを割り当てる．いずれのカテゴリの分類確率も ϵ より上回らない場合には，最も高い分類確率を持つカテゴリを割り当てる．以上をまとめると，x に割り当てられるカテゴリ集合 $\hat{Y}$ は以下のように表現される．

$$\hat{Y} = \{f_j(\mathbf{x}) \mid f_j(\mathbf{x}) > \epsilon, j \in [m]\} \cup \{\mathrm{argmax}_{j \in [m]} f_j(\mathbf{x})\}$$

実験では $\epsilon = 0.2$ に設定する．

4 実験

4.1 データセット

訓練データとして，「ビズリーチ」に登録されている求人票データ (ビズリーチデータと呼ぶ) を用いる．BERT モデルの訓練には，5 万組のポジティブペア，10 万組のネガティブペア，10 万組のハードネガティブペアを使用する．また，マルチラベル分類モデルの訓練には，5 万件の求人票データを使用する．

評価データとして，「ビズリーチ」の求人票データ 2 万件を用いる．また，インターネット上に公開されている求人票データ (外部データと呼ぶ)188 件を用いた評価も行う．ここでは，外部データに対してビズリーチデータの職種カテゴリと業種カテゴリを割り当てることが目的であり，正解カテゴリは人手で作成した．

4.2 BERT モデルの設定

BERT モデルとして，大量の求人票データを高速に処理できるよう，合計パラメータ数が 1,500 万程度の比較的に軽量なモデルを用いる．ネットワーク構造は 8 層の transformer 層と平均 pooling 層からなる．埋め込みベクトルの次元は 256 であり，最大入力トークン数は 2048 である．トークナイザーとして SudachiPy[10] を用いる．Transformers ライブラリ [11] で BERT モデルの実装を行った．

4.3 比較手法

本研究の目的は人材業界固有の表現を学習することにより，求人票データの分類精度を向上させることである．そこで，提案手法を以下の固有の表現を学習していないベクトル化手法である TF-IDF と一般的な公開コーパスで訓練された BERT モデル[2] (一般 BERT と呼ぶ) を比較手法として用いる．求人票のテキストを文単位に分割してからベクトル化する場合と，分割せずにベクトル化する場合の比較を行う．ベクトル化した後の処理は，提案手法と同じで

2) https://huggingface.co/pkshatech/simcse-ja-bert-base-clcmlp

手法	文単位分割	ビズリーチデータ			外部データ		
		Accuracy	Macro-F	Micro-F	Accuracy	Macro-F	Micro-F
TF-IDF	なし	0.5601	0.2637	0.6025	0.5353	0.1996	0.5859
TF-IDF	あり	0.5189	0.2476	0.5656	0.4456	0.2046	0.5351
一般 BERT	なし	0.5318	0.2560	0.5757	0.5310	0.1878	0.5768
一般 BERT	あり	0.4488	0.1581	0.5003	0.4489	0.1039	0.4893
ファインチューニング済み求人 BERT	なし	0.6069	0.2718	0.6587	0.5127	0.2169	0.5740
求人 BERT	なし	0.5405	0.2440	0.5811	0.5405	0.2440	0.5811
(提案手法) 求人 BERT	あり	**0.6207**	**0.4963**	**0.6682**	**0.5810**	**0.3150**	**0.6598**

表 1　職種予測の評価結果．太字は最も良い結果を示している．

手法	文単位分割	ビズリーチデータ			外部データ		
		Accuracy	Macro-F	Micro-F	Accuracy	Macro-F	Micro-F
TF-IDF	なし	0.5366	0.2680	0.5757	0.7042	0.1784	0.7534
TF-IDF	あり	0.4082	0.1500	0.4436	0.8244	0.2046	0.8586
一般 BERT	なし	0.5492	0.3359	0.5902	0.7753	0.1697	0.8139
一般 BERT	あり	0.4275	0.1468	0.4592	0.8406	0.2031	0.8674
ファインチューニング済み求人 BERT	なし	**0.6181**	**0.4024**	**0.6715**	0.8725	0.2812	0.8946
求人 BERT	なし	0.4227	0.1522	0.4534	0.4227	0.1522	0.4534
(提案手法) 求人 BERT	あり	0.5372	0.3915	0.5824	**0.9073**	**0.3991**	**0.9297**

表 2　業種予測の評価結果．太字は最も良い結果を示している．

ある．

また，文単位の分割なしのテキストを入力として，求人 BERT をマルチラベル分類タスクでファインチューニングしたモデル (ファインチューニング済み求人 BERT と呼ぶ) との比較も行う．

4.4　評価指標

評価指標として，以下で定義される Accuracy，Macro-F1，Micro-F1 を用いる．

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N}\frac{|Y_i \cap \hat{Y}_i|}{|Y_i \cup \hat{Y}_i|}$$

$$\text{Macro-F1} = \frac{1}{m}\sum_{j=1}^{m}\frac{2p_j r_j}{p_j + r_j}$$

$$\text{Micro-F1} = \frac{2pr}{p+r}$$

ここで，N はデータ数，$\hat{Y}_i$ と Y_i はそれぞれデータ i の予測カテゴリ集合と正解カテゴリ集合，p_j と r_j はそれぞれカテゴリ j に関する適合率と再現率，p と r はそれぞれデータとカテゴリ単位の適合率と再現率を表す．

4.5　結果

各手法の職種と業種予測の評価結果をそれぞれ表 1 と表 2 に示す．求人票のテキストの文単位の分割を行い，ドメイン特化の求人 BERT を用いる提案手法は，ビズリーチデータの業種予測問題を除いて，比較手法よりも優れた結果を出した．提案手法が高い性能を発揮できた理由は，求人票のテキストを文単位に分割することにより分類において無関係なテキストの影響を低減し，求人票のテキストを用いて訓練された BERT モデルにより人材業界固有の表現を獲得できたからだと思われる．

マルチラベル分類タスクで求人 BERT モデルのファインチューニングを行ったファインチューニング済み求人 BERT は，業種予測タスクにおいては提案手法よりも良い結果を出した．しかし，ファインチューニング済み求人 BERT はビズリーチデータに最適化されるように訓練されているため，求人票のテキストや職種，業種の分布がビズリーチデータと異なる外部データにおいて最も優れた結果を出すことができなかった．一方で，提案手法は人材業界固有の表現を学習し，求人票のテキストとカテゴリとの類似度に基づいて予測を行うことで，ビズリーチデータと外部データの分布が異なっていても高い性能を発揮することができたと思われる．

5　まとめ

本研究では，人材業界固有の表現を獲得するために，求人票データを用いてドメイン特化 BERT モデルの訓練を行った．また，無関係なテキストの影響を低減するために，求人票のテキストを文単位に分割してから，各カテゴリとの最大類似度を用いてマルチラベル分類を行う手法を提案した．職種と業種のマルチラベル分類タスクにおいて，提案手法は優れた性能を発揮し，分布シフトにも強いことを確認した．

参考文献

[1] Santiago González-Carvajal and Eduardo C Garrido-Merchán. Comparing bert against traditional machine learning text classification. **arXiv preprint arXiv:2005.13012**, 2020.

[2] Andrea Gasparetto, Matteo Marcuzzo, Alessandro Zangari, and Andrea Albarelli. A survey on text classification algorithms: From text to predictions. **Information**, Vol. 13, No. 2, p. 83, 2022.

[3] Yuxin Luo, Feng Lu, Vaishali Pal, and David Graus. Enhancing resume content extraction in question answering systems through t5 model variants. In **RECSYS in HR 2023 : the 3rd Workshop on Recommender Systems for Human Resources (RecSys in HR 2023), Proceedings**, Vol. 3490, 2023.

[4] Hussain Falih Mahdi, Rishit Dagli, Ali Mustufa, and Sameer Nanivadekar. Job descriptions keyword extraction using attention based deep learning models with bert. In **2021 3rd International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)**, pp. 1–6. IEEE, 2021.

[5] Changmao Li, Elaine Fisher, Rebecca Thomas, Steve Pittard, Vicki Hertzberg, and Jinho D Choi. Competence-level prediction and resume & job description matching using context-aware transformer models. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 8456–8466, 2020.

[6] Jens-Joris Decorte, Jeroen Van Hautte, Johannes Deleu, Chris Develder, and Thomas Demeester. Career path prediction using resume representation learning and skill-based matching. In **RECSYS in HR 2023 : the 3rd Workshop on Recommender Systems for Human Resources (RecSys in HR 2023), Proceedings**, Vol. 3490, 2023.

[7] 川崎拳人. 非構造化データの構造化における情報抽出. 研究報告音声言語情報処理 (SLP), Vol. 2020, No. 16, pp. 1–6, 2020.

[8] Tianyu Wu, Shizhu He, Jingping Liu, Siqi Sun, Kang Liu, Qing-Long Han, and Yang Tang. A brief overview of chatgpt: The history, status quo and potential future development. **IEEE/CAA Journal of Automatica Sinica**, Vol. 10, No. 5, pp. 1122–1136, 2023.

[9] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**. Association for Computational Linguistics, 2019.

[10] Kazuma Takaoka, Sorami Hisamoto, Noriko Kawahara, Miho Sakamoto, Yoshitaka Uchida, and Yuji Matsumoto. Sudachi: A japanese tokenizer for business. In **Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)**, 2018.

[11] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In **Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations**, pp. 38–45, 2020.