HR ドメイン特化の疎ベクトル検索モデルの構築

Dat P.T. Nguyen[1]　渡會 恭平[1]　加藤 遼[1]
[1] 株式会社ビズリーチ
{dat.nguyen, kyohei.watarai, ryo.kato}@bizreach.co.jp

概要

本研究では，人材採用プラットフォームにおけるレジュメ検索[1)]の高精度・低レイテンシかつ高い解釈性の実現を目的とし，疎ベクトル検索モデル SPLADE[1] の導入と，HR (Human Resources) ドメイン特有の課題に対応した学習手法を提案する．従来のキーワード検索では語彙の不一致により，ユーザーである企業の採用担当者の検索意図を十分に反映できない課題があった．これに対し，我々は高い解釈性と検索精度を両立する SPLADE を採用し，さらに HR データ特有の多対多の関係性や企業の採用担当者の主観的判断によるノイズを考慮した学習データ構築手法を開発した．実データを用いた評価実験の結果，提案手法は汎用 SPLADE モデルの再現率を上回り，かつ既存の一致検索システムに肉薄する性能を達成し，HR 領域におけるセマンティック検索の有効性を実証した．

1　はじめに

人材採用プラットフォームにおいて，膨大なデータベースから適切な求職者を効率的に発見する検索技術の重要性は高まっている．現在，ビズリーチにおけるレジュメ検索は，職種や年収などの属性をもとにしたハードフィルタと，フリーワードによるソフトフィルタの組み合わせによって実現されている．しかし，キーワード検索は文字列の完全一致に基づいているため，検索クエリとレジュメテキスト間で語彙が一致しない場合に対応できないという課題がある．例えば，企業の採用担当者が特定のスキルや経験を意図して検索を行っても，求職者が異なる同義語や表現を用いていた場合，その求職者は検索結果から除外されてしまう．加えて，キーワード検索は文脈や多義性の扱いに限界があり，企業の採用担当者が持つ潜在的な要件を汲み取ることが難しい．そのため，意図した検索結果を得るには，ユーザー自身による複雑な条件設定が不可欠となっている．

これらの課題を解決するため，本研究では，セマンティック検索の実現に向けた疎ベクトル検索モデルの導入と，HR ドメイン特有の課題に対応した学習手法を提案する．疎ベクトル検索モデルとして，SPLADE [1] を採用する．SPLADE は BERT [2] 等の Transformer モデルを基盤としつつ，学習時に FLOPS 正則化項を用いることで出力ベクトルをスパースにする手法である．これにより，入力テキストの意味的に関連する拡張語に重みを付与できるため，密ベクトル検索と同等の精度と，キーワード検索同様の高速性・解釈性を両立することが可能となる．

次に，HR ドメインへの適応について述べる．MS MARCO [3] などの汎用データセットで学習されたモデルは，「RPO」や「ダイレクトリクルーティング」といった HR 特有の専門用語や文脈を十分に獲得できていない．そこで本研究では，独自データセットを用いたファインチューニングを行うことで，この語彙ギャップを解消する．検証においては，ドメイン適応を施さない汎用モデルと，適応後の特化型モデルを比較し，HR 領域の検索タスクにおけるドメイン適応の有効性を明らかにする．

2　関連研究

SPLADE におけるドメイン適応は，特定領域での検索性能を高める上で極めて有効である．

Coudurier ら (2024) [4] は，第 1 段階検索におけるゼロショット性能の限界に対し，SPLADE に対するドメイン適応の有効性を実証した．彼らはターゲットドメインでの事前学習により，汎化性能を維持しつつ，ドメイン特有の語彙や文脈への適合性を大幅に向上できることを示した．これは，疎ベクトルモデルにおいても，ドメイン適応が検索精度の向上に貢献することを示している．

ドメイン適応の成功事例として法務分野が挙げ

1) 「レジュメ」とはビズリーチ上の「職務経歴書」を指します

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

られる．法律文書は極めて長い文長や条文番号などの厳密なエンティティ参照を含むため，汎用モデルでの対応が難しい．SPLADE Legal French [5] は，フランス語の法令データを用いた学習により，汎用的な CamemBERT ベースのモデルと比較して大幅な精度向上を達成した．これは専門性の高い領域において，ドメイン特化型の学習が必要であることを示した．

このように他分野での有効性が確認されている一方で，HR 領域における SPLADE の研究は依然として少ない．特に日本語の HR ドメインでは，求人票や職務経歴書といった特有の文書構造や，業界用語とスキル要件の複雑な関係性が存在するにもかかわらず，適応事例は未開拓である．そこで本研究では，日本語 HR ドメインに焦点を当て，最適なモデル構成および学習手法を検討することで，同領域の情報検索における性能向上を目指す．

3 手法

本章では，本研究における HR ドメインへの適用，および特有のデータ課題への対処法について詳述する．SPLADE モデルの概要については，付録 A に詳細を記載している．

3.1 HR ドメイン特化型 SPLADE モデル開発

HR 領域の検索システムにおいては，スカウト候補となるレジュメを漏れなく検索結果の上位に提示する高い再現率が極めて重要である．これを実現するための学習データとして，以下の定義を採用する．

- POSITIVE ドキュメント: 企業の採用担当者が特定のクエリで検索した結果，実際にスカウトを送信したレジュメ
- NEGATIVE ドキュメント: POSITIVE ドキュメントを除くレジュメ群からフィルタリングおよびランダムサンプリングしたレジュメ

3.2 HR データの複雑性とノイズへの対応

一般的な QA タスクのデータセットの特徴である 1 クエリ対 1 ドキュメントとは異なり，HR データは多対多の複雑な構造を持つ．すなわち，1 つのクエリに対して複数のレジュメがスカウトされ，逆に 1 つのレジュメは複数の異なるクエリによって発見される．また，スカウト判断は企業の採用担当者の主観に依存するため，検索行動ログには本質的にノイズが含まれる．一般的な密ベクトル検索の学習手法（In-Batch Negatives 等）を適用すると，この複雑性とノイズにより学習が収束しない問題が発生した．そこで，本研究では以下のデータ構築手法を提案する．

1. POSITIVE ドキュメントの増強: 学習の安定性を高めるため，あるクエリに対する POSITIVE データを，そのクエリと類似性の高い別のクエリでスカウトされたレジュメによって増強する．これにより，疎なデータセットにおいても十分な正例を確保し，モデルのロバスト性を向上させる．
2. クラスタリングによるノイズ除去: 曖昧なスカウト判断を含む類似クエリ群をクラスタリングし，その中から代表的なクエリのみを選択して学習データとする．これにより，企業の採用担当者の主観による矛盾を低減し，学習データの質を向上させる．

3.3 学習パイプライン

本研究で構築した HR ドメイン特化型 SPLADE モデル（以下，HR-SPLADE）は，「ドメイン適応事前学習（Domain Adaptive Pretraining）」によって基礎的な言語能力を高め、さらに SPLADE++[6] の学習手法に従い「知識蒸留（Knowledge Distillation）」によって高い検索精度を実現する構成となっている．HR-SPLADE を構築したパイプラインは、図 1 のように大きく 3 つのフェーズに分けて構成される．

- フェーズ 1: HR-Domain Adaptive Pretraining：独自語彙の定義と、HR ドメインに最適化したベースモデルの事前学習
- フェーズ 2: Knowledge Distillation: 検索ログと Cross-Encoder を用いた教師データ作成
- フェーズ 3: Fine-tuning: HR-SPLADE モデルの学習と最適化

4 実験

本実験では，求職者検索フェーズにおけるモデルの性能を評価する．具体的には，検索クエリに対してモデルが出力した上位 N 件の求職者リストの中に，過去に実際にスカウトが送信された求職者が含まれる割合を測定し，その再現率である $Recall@N$ を評価指標とする．本実験において比較・検証を行

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

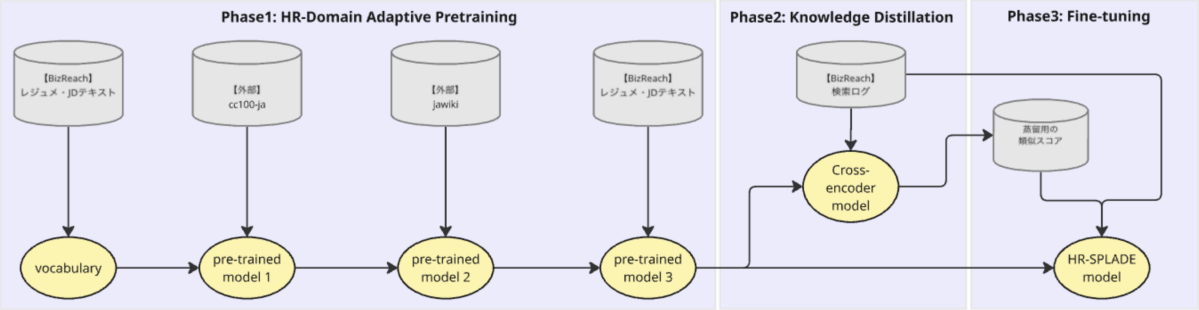

図 1　HR-SPLADE の学習パイプライン

うモデルは以下の通りである．

4.1　比較対象モデル

以下の 3 つのモデルを比較対象とする．

- BM25：現状のキーワード検索モデル．ログから抽出された，現行の ElasticSearch の検索結果をシミュレートしたもの．
- japanese-splade-base-v1[2)]：大規模な汎用コーパスで学習された汎用的な SPLADE モデル．
- HR-SPLADE（提案モデル）：本研究で提案する HR ドメインデータで学習された SPLADE モデル．2.2 節のデータロバスト性向上手法を適用したもの．

4.2　学習データと評価データ作成

ビズリーチの過去の検索ログから，検索条件，検索結果，および実際のスカウト履歴を抽出し，検索日時に基づいて時系列で学習用と評価用に分割した．データの件数を表 1 に示す．

項目	学習データ	評価データ
検索条件総数	23,190	10,349
検索結果の impression 総数[3)]	9,335,809	4,121,664
スカウト総数	393,386	36,083

表 1　データ統計情報

4.3　実験結果

システム要件に基づき，リトリーバルステージの性能を評価するため，再現率（Recall）指標を重視する．Rec@K は，POSITIVE ドキュメントが検索結果の Top K 件の中に含まれていた割合を示す．定量評価結果を表 2 に示す．

評価データには過去の実際の検索ログを利用しているため，全てのスカウト送信済みレジュメが検索結果の上位 1,000 件に含まれている場合，実際の検索エンジンで採用されている BM25 モデルの Rec@1000 は 1.0 となる．このため，再現率においては BM25 が理想的なモデルに近い値を示し，他のモデルがこれを上回ることはできない．

5　考察

表 1 の結果より，以下の点が確認された．

- BM25 は，検索キーワードがレジュメにそのまま含まれるケースにおいては極めて高い再現率を示すが，セマンティック検索への移行により，より多くの求職者を発見できる可能性がある．
- japanese-splade-base-v1 は，セマンティックな検索能力を有しながらも，HR ドメイン特有の語彙や文脈を十分に学習できていないため，ベースラインと比較して大幅に低い再現率に留まった．
- 本研究が提案した HR-SPLADE は，汎用 SPLADE モデルと比較して全ての K において再現率が向上した．特に Rec@100 では約 7 ポイントの改善が見られ，実用上重要な上位候補への適切なレジュメの提示に大きく貢献する．
- Rec@1000 において 95% 以上の POSITIVE ドキュメントが網羅されていることから，リトリーバルステージの上位 1,000 件をリランキングステージへの入力候補とすることで，適合文書が最終的なランキング上位に含まれる可能性が高まると考えられる．
- この性能向上は，3.2 節で提案した HR ドメイン特有の複雑なスカウトログに対応するためのデータロバスト性向上手法が効果的に機能し，モデルがドメインに特化した適切な疎ベクトル表現を獲得できたことを示唆している．
- HR-SPLADE モデルが小規模（約 3000 万パラ

2）　https://huggingface.co/hotchpotch/japanese-splade-base-v1
3）　検索を実行したときに検索結果一覧に表示されたレジュメの総合件数

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

モデル	Param 数	Rec@10	Rec@100	Rec@500	Rec@1000
BM25	-	0.3831	0.9103	0.9910	1.0000
japanese-splade-base-v1	111M	0.1685	0.5562	0.8375	0.9167
HR-SPLADE	30M	0.2032	0.6287	0.8977	0.9583

表 2　定量評価結果

メータ）であるにもかかわらず，汎用モデル（約 1 億 1000 万パラメータ）を上回る性能を示したことは，ドメイン特化型モデルの有効性を強調するとともに，検索のレイテンシーを削減する上でも有益である．

6　結論

本研究は，HR ドメインにおけるセマンティック検索の実現を目的とし，疎ベクトル検索モデル SPLADE を HR ドメインのスカウトログデータでファインチューニングする手法を提案した．HR ドメインデータ特有の多対多の複雑性に対応するため，ポジティブデータの増強とノイズとなるクエリの除外による学習のロバスト性向上策を導入した．定量評価の結果，提案モデルは汎用 SPLADE モデルと比較して，Rec@K において一貫して高い性能を示し，HR 領域におけるセマンティック検索の有効性を証明した．

今後は，本モデルで抽出された求職者一覧に対し，ランキング学習モデルを連携させることで，最終的な検索順位付けの精度向上を目指す．

参考文献

[1] Thibault Formal, Carlos Lassance, Benjamin Piwowarski, and Stéphane Clinchant. Splade v2: Sparse lexical and expansion model for information retrieval, 2021.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[3] Luiz Bonifacio, Vitor Jeronymo, Hugo Queiroz Abonizio, Israel Campiotti, Marzieh Fadaee, Roberto Lotufo, and Rodrigo Nogueira. mmarco: A multilingual version of the ms marco passage ranking dataset, 2022.

[4] Mathias Vast, Yuxuan Zong, Benjamin Piwowarski, and Laure Soulier. **Simple Domain Adaptation for Sparse Retrievers**, p. 403–412. Springer Nature Switzerland, 2024.

[5] Antoine Louis, Gijs van Dijck, and Gerasimos Spanakis. Know when to fuse: Investigating non-english hybrid retrieval in the legal domain. **CoRR**, Vol. abs/2409.01357, , 2024.

[6] Thibault Formal, Carlos Lassance, Benjamin Piwowarski, and Stéphane Clinchant. From distillation to hard negative sampling: Making sparse neural ir models more effective. In **Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval**, SIGIR '22, p. 2353–2359, New York, NY, USA, 2022. Association for Computing Machinery.

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

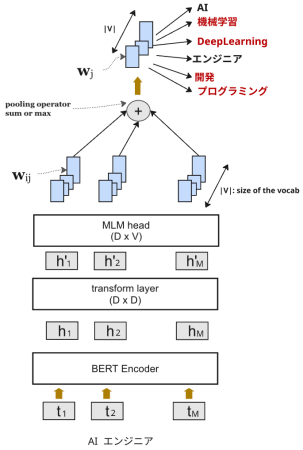

図 2 SPLADE モデルのアーキテクチャと疎ベクトル生成メカニズム

A 参考情報

A.1 SPLADE モデルの概要

SPLADE は, BERT の最終出力から語彙全体に対応する疎ベクトルを生成する機構である．SPLADE[1] をベースとしたアーキテクチャを図 2 に示す．

BERT による埋め込みと語彙への射影

入力テキストのトークン $t = (t_1, t_2, \ldots, t_M)$ (M はトークン数) を BERT エンコーダーに入力し，隠れベクトル $(h_1, h_2, \ldots, h_M)$ を取得する．次に，それぞれの隠れベクトル h_i を transform layer で変換し，BERT の語彙数 $|V|$ に対応する次元へ射影し，入力トークン t_i が語彙内の各トークン j に与える重要度 w_{ij} を以下の式で計算する．

$$w_{ij} = \text{transform}(h_i)^T E_j + b_j \tag{1}$$

ここで，E_j は語彙内のトークン j に対応する埋め込みベクトル，b_j はバイアスパラメータである．その後，SPLADE は Pooling 層を通して，入力全体に対する各語彙のスコア w_j を集約する．

疎ベクトル生成

最終的な疎ベクトル w は，各語彙トークン j のスコア w_j を以下の式で算出することで得られる．

$$w_j = \max_{i \in t} \log(1 + \text{ReLU}(w_{ij})) \tag{2}$$

ここで Max Pooling は，入力全体における各語彙の最大重要度を抽出する．これによりトークンごとのスコアが統合され，入力テキストのスパース表現が生成される．

スパース性を確保する正則化と損失関数

SPLADE を効果的に学習させるには，適合性を高めるだけでなく，ベクトルが過度に密になることを防ぎ，スパース性を保つための工夫が必要である．

1. 損失関数

適合性を学習するための損失関数として，SPLADE では In-Batch Negatives を用いた以下の損失関数を採用する．

$$\mathcal{L}_{\text{rank-IBN}} = -\log \frac{e^{\text{sim}(q_i, d_i^+)}}{e^{\text{sim}(q_i, d_i^+)} + e^{\text{sim}(q_i, d_i^-)} + \sum_j e^{\text{sim}(q_i, d_{i,j}^-)}} \tag{3}$$

ここで，q_i はクエリのベクトル，d_i^+ と d_i^- はそれぞれ q_i に対する正例および負例ドキュメントのベクトルを表す．また，$d_{i,j}^-$ はバッチ内の i 以外のクエリに対する正例ドキュメント（クエリ i に対する負例として扱われる）のベクトルであり，$\text{sim}(q, d)$ はクエリベクトル q とドキュメントベクトル d の類似度を計算する関数である．

2. スパース正則化

ベクトル内の非ゼロ要素数を抑制するため，FLOPS 正則化を導入する．この手法は各語彙トークンの重みの二乗和を最小化することでベクトル全体の活性化を抑え，結果としてスパース性を高める．

$$\mathcal{L}_{\text{FLOPS}} = \sum_{j \in V} \left(\frac{1}{B} \sum_{i=1}^{B} w(d_i)_j \right)^2 \tag{4}$$

ここで，B はバッチサイズ，$w(d_i)_j$ はドキュメント d_i における語彙 j の重み（非負の値）である．この正則化項を損失関数に加えることで，モデルは必要な単語のみを拡張するように学習する．

3. 全体の損失関数

最終的な損失関数は，クエリとドキュメントの両方に対して正則化項を加えたものとなる．

$$\mathcal{L} = \mathcal{L}_{\text{rank-IBN}} + \lambda_q \mathcal{L}_{reg}^q + \lambda_d \mathcal{L}_{reg}^d \tag{5}$$

$\mathcal{L}_{reg}^q$ と $\mathcal{L}_{reg}^d$ はそれぞれクエリとドキュメントの FLOPS 正則化項であり，ハイパーパラメータ λ_q, λ_d によってクエリとドキュメントのベクトルのスパース性を調整する．

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).