k 近傍事例に基づく埋め込み表現のドメイン適応と検索への応用

五藤巧 [1]　堤田恭太 [2]　村瀬文彦 [2]　三谷陽 [2]　渡辺太郎 [1]
[1] 奈良先端科学技術大学院大学　[2] 株式会社デンソー
{goto.takumi.gv7, taro}@is.naist.jp
{kyota.tsutsumida.j7m, fumihiko.murase.j6b, akira.mitani.j5g}@jp.denso.com

概要

特定ドメインを対象とする検索タスクにおいて埋め込みモデルのドメイン適応は重要であるが，対象とするドメインの種類数が膨大な場合には追加学習による方法を適用するのは難しい．本研究では，ニューラルモデルに基づく密ベクトル検索において，モデルを追加学習することなくドメイン適応させる方法を提案する．具体的には，ドメイン適応先の単言語コーパスからクエリの k 近傍事例を計算して，近傍事例の埋め込み表現を統合した表現によってクエリの表現を調整する．実験では，独自に収集するデータを用いた故障事例検索タスクにおいて，提案法により検索性能が向上すること，また検索結果が解釈性の向上につながることを示す．

1　はじめに

密ベクトルによる文書検索では，BERT [1] などの埋め込みモデルによって計算される埋め込み表現によって文書間の類似度を計算し，類似度が上位の文書を提示する．対象とするテキストが特定のドメインを対象とする場合には，そのドメインに適応した埋め込みモデルを使うことで検索性能が向上すると考えられる．単純なドメイン適応方法には追加学習を考えることができ，例えば，適応先ドメインのテキストを用いて SimCSE [2] の学習手法を適用することが考えられる．しかし，ドメインの種類数が 100 種類以上など多数存在する場合には，ドメインごとの追加学習は難しい．また，対象とするドメインのテキストに秘匿性がある場合，それらのテキストをモデルの学習に用いるのは望ましくない．

このような課題を解決するため，本研究では，k 近傍事例に基づいて埋め込み表現をドメイン適応する方法を提案する．具体的には，適応先ドメインの単言語コーパスからクエリの k 近傍事例を検索し，検索事例の埋め込み表現を距離で重み付けした表現を獲得する．この表現とクエリの元の埋め込み表現との線形補完を取ることで，ドメイン適応させた表現を獲得する．後段の検索タスクでは，検索される事例の表現も同様にドメイン適応させることで，ドメイン知識を考慮した検索を行う．したがって，図 1 の概念図に示すように，推論時にはドメイン適応した表現を獲得するための検索と，実際の検索タスクを解くための検索のパイプラインとなる．提案法が必要とするのは対象ドメインの単言語コーパスのみであるため，ラベルのアノテーションやモデルの追加学習が必要ないことが利点である．

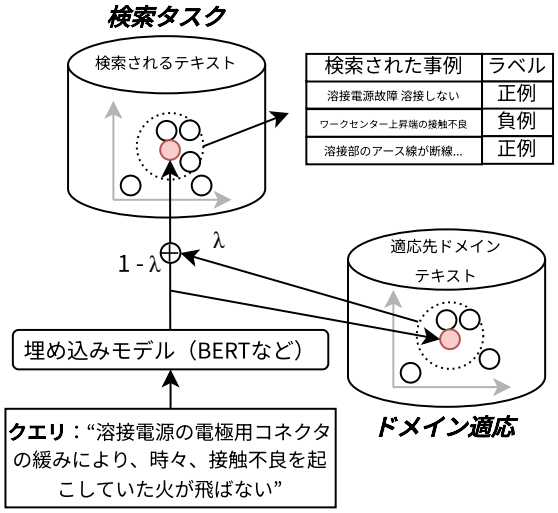

図 1　提案法の概念図．

実験では，内部で開発されたデータに基づく故障事例検索タスクを対象に，データ収集元の事業部や機器をドメインとみなした実験を行う．結果から，提案法は適応元と適応先のドメインの乖離がある程度小さい場合に有効であることを示す．また，ドメイン適応のための検索におけるハイパーパラメタと性能の関係や，検索結果の提示による解釈性の利点

について議論する．

2 手法

提案法は既存モデルの埋め込み表現を後処理として調整することによって，モデルを追加学習することなくドメイン適応させる．適応先ドメインの文書集合を使用し，既存モデルによって計算された入力テキストの埋め込み表現と，文書集合から検索された k 近傍事例の埋め込み表現を統合することで調整する．提案法に必要なのは，埋め込みモデル Enc とドメイン適応先の文書集合 $\mathcal{D} = \{\boldsymbol{x}_i\}_{i=1}^{|\mathcal{D}|}$，および入力のクエリ $\boldsymbol{q}$ である．なお，埋め込みモデルは適当なテキスト $\boldsymbol{x}$ を入力として d 次元の埋め込み表現 $\mathrm{Enc}(\boldsymbol{x}) \in \mathbb{R}^d$ を計算するモデルとする．

まず，入力テキストの埋め込み表現 $\mathrm{Enc}(\boldsymbol{q})$，および文書集合の各事例の埋め込み表現 $\{\mathrm{Enc}(\boldsymbol{x})|x \in \mathcal{D}\}$ を計算する．次に，$\mathrm{Enc}(\boldsymbol{q})$ をクエリとして k 近傍事例 $\mathcal{K} \subseteq \mathcal{D}$ を検索し，次式に従い k 近傍事例の表現を統合する．

$$\boldsymbol{h}_{k\mathrm{NN}} \propto \sum_{\boldsymbol{x} \in \mathcal{K}} \mathrm{Enc}(\boldsymbol{x}) \exp\left(\frac{\mathrm{Sim}(\mathrm{Enc}(\boldsymbol{x}), \mathrm{Enc}(\boldsymbol{q}))}{\tau}\right) \quad (1)$$

式 1 は，それぞれの近傍事例の埋め込み表現 $\mathrm{Enc}(\boldsymbol{x})$ の重み付き和を，exp() の項を重みとして計算することを示している．Sim() は 2 つの埋め込み表現の類似度を定量化する尺度で，負のユークリッド距離や余弦類似度を使用できる．τ は温度パラメータで，この値が大きいほどそれぞれの近傍事例にかかる重みが均一化する．この定式化は，クエリに関連するドメイン知識に問い合わせながら表現を獲得するために，入力クエリと単言語コーパスとの Cross Attention を k 近傍事例のみに限定して計算しているとも解釈できる．最後に，この表現と元の表現の線形補完を取ることでドメイン適応した表現を得る．

$$\boldsymbol{h}'_{\boldsymbol{q}} = (1 - \lambda)\mathrm{Enc}(\boldsymbol{q}) + \lambda \boldsymbol{h}_{k\mathrm{NN}} \quad (2)$$

λ は重み付けのための係数であり，値が大きいほど近傍事例から計算される表現を重要視する．

3 実験

3.1 故障事例検索タスク

本研究では，内部で開発されたデータに基づく故障事例検索タスクを用いて提案法のドメイン適応能力を明らかにする．本タスクの目的は，ユーザが入力する機器の故障に関する説明文を入力として，故障の処置内容を提示することである．この目的のため，本研究で独自に収集する保全履歴データを活用する．保全履歴データは，表 1 に示すように，故障事例を記述した文書とその処置内容を記述した文書のペアからなるデータである．検索では処置内容を直接検索するのではなく，故障事例の記述のみの類似度に基づいて事例を検索し，検索された事例に紐つく処置内容を提示することとする．保全履歴データは様々な事業部や機器を対象として収集されているため，これらをそれぞれドメインとみなすことができる．本研究では特定の事業部が対象とする溶接機に関する機器を適応先ドメインとし，提案法のドメイン適応性能を明らかにする．

表 1 保全履歴データの例．クエリに対する正例と負例を一件ずつ示した．クエリの一例は “溶接電源の電極用コネクタの緩みにより、時々、接触不良を起こしていた 火が飛ばない”.

故障事例の記述	処置内容の記述
正例	
溶接の位置ズレ 溶接形状が悪い	溶接ポイント変更実施（P105、P106）
負例	
ボルトの疲労 lm ガイドボルト折損	取り換え&心出し

表 2 データセットの統計量.

データセット	文書数	備考
クエリ	10	
ラベル付きデータ	995	正例 218 件，負例 777 件
ラベルなしデータ	244,895	

使用するデータセットの統計量を表 2 に示す．クエリは，ユーザが入力すると仮定する故障事例の記述である．ラベル付きデータは，検索される故障事例の記述に対して，検索されるべきかどうかを人手でラベル付けしたもので，いずれのクエリについても共通である．すなわち，どのクエリが入力されても正例のいずれかが検索されるべきである．また，ラベルなしデータは適応先ドメインについて対象とする事業部は同じであるが機器は異なる．

3.2 実験設定

ラベルなしデータ 224,895 件を適応先ドメインの単言語コーパス $\mathcal{D}$ として用いて，ドメイン適応した表現を検索タスクに利用する．ラベル付きデータは後段の検索タスクで検索されるテキスト集合であり，それぞれの埋め込み表現は提案法でドメイン適応される．埋め込みモデルには東北大日本語

表 3 故障事例検索タスクにおける実験結果．ID はドメイン内設定，OOD はドメイン外設定で追加事前学習することに対応する．提案法ありのグループでは，提案法なしから改善したものに下線を引いた．

	性能			
k	8	16	32	64
BM25	86.25	85.62	83.75	69.84
提案法なし				
東北大 BERT	86.25	81.25	74.37	64.06
+追加学習 (OOD)	85.00	82.50	75.62	67.96
+追加学習 (ID)	95.00	88.12	81.87	72.65
提案法あり				
東北大 BERT	87.50	79.38	71.85	62.19
+追加学習 (OOD)	87.50	84.38	78.12	71.56
+追加学習 (ID)	90.00	87.50	80.62	71.72

BERT（tohoku-nlp/bert-base-japanese-v2）を用いる．実験では共通して，トークンレベルの表現を平均プーリングすることで文書単位の表現を計算することとし，埋め込み表現の類似度の尺度 Sim(·) には余弦類似度を用いる．

比較手法　埋め込みモデルを実際に追加学習したモデルと比較することで，提案法が追加学習と同等のドメイン適応性能を有するかを確認する．この目的のため，事業部の観点からドメイン外設定とドメイン内設定の 2 種類について，東北大日本語 BERT をマスクトークン復元タスクで追加事前学習した．ドメイン内設定では，目的のドメインと同じ事業部の故障事例を，またドメイン外設定では異なる事業部の保全履歴データを用いた．学習データの事例数はどちらも 1 万件である．ドメイン内設定は理想的な設定であるものの，ドメインの数だけモデルを学習・運用する必要があるため，これを避けることが本研究の目的である．ドメイン外設定は，事業部や機器で共通するドメイン知識は獲得しているが，ある特定の機器についての専門用語にはドメイン適応させる余地があると考えることができ，実用的な設定である [1]．また疎ベクトル検索による手法である BM25 とも比較する [2]．

評価方法　ある一つのクエリについて上位 N 件の検索結果を計算し，その中に含まれる正例の件数で評価する．評価データにクエリは 10 件あるため，各クエリの正例の割合を平均したものを最終的な評価値とする．また，検索結果の上位数件のみを提示したり，下位の事例も含めてより多くの件数を提示するなどの多様な状況における性能を評価するため，$N = \{8, 16, 32, 64\}$ について評価する．なお，提案法では式 1 の $|\mathcal{H}|$，同式の温度パラメータ τ，式 2 の重み λ を決定する必要があるが，実験では評価におけるそれぞれの N について，オラクルの設定を使用して結果を報告する [3]．オラクル設定とした理由は，評価データのクエリ数が 10 件と限られており，開発データが捻出できないためである [4]．

3.3 実験結果

実験結果を表 3 に示す[5]．表は 3 つのグループに分かれており，“提案法なし” はモデルが計算する埋め込み表現をそのまま用いて検索を行い，“提案法あり” は提案法により埋め込み表現をドメイン適応させてから検索する．“追加学習” は，追加事前学習したモデルに対応し，ID はドメイン内設定，OOD はドメイン外設定にそれぞれ対応する．

まず，提案法なしのグループの中で結果を比較することにより，埋め込みモデルをドメイン適応させることで検索性能が向上することが分かる．ドメイン内設定では特に性能が向上することから，対象とするドメインに近いデータで追加学習するほど検索性能が向上する．次に，“提案法なし” と “提案法あり” との比較結果から，ドメイン外設定（OOD）で提案法が有効であるが，その他のモデルではほとんど有効性が確認できなかった．東北大 BERT を対象とする場合，適応元から適応先へのドメインは大きく乖離しており，提案法では適応させるのが難しい．一方，ドメイン外設定では事業部共通の知識は獲得しているためドメインの乖離は小さく，そのような場合に提案法が有用であると考えられる．

BM25 との比較では，$N = 64$ のように下位の検索事例も含めて提示する場合には提案法が有用であるが，$N \leq 32$ の場合には BM25 が上回る．タスクの傾向として，溶接機に関する故障事例では “溶接” などの特定の単語が文書に含まれやすく，BM25 でも比較的検索が容易であると考えられる．

1） 例えば，保全履歴データをあらゆる事業部や機器について混ぜたデータで，一度のみモデルを追加事前学習することに対応する．
2） 実装には BM25s [3] を用いた．
3） 付録 A に示す候補を試行する．
4） 10 件のデータを分割して交差検証のように実験することもできるが，数件のデータのみから決定された検索設定の信頼性は低いと考えられ，妥当な実験設定とは言い難い．
5） 使用した検索設定を付録 B に示す．

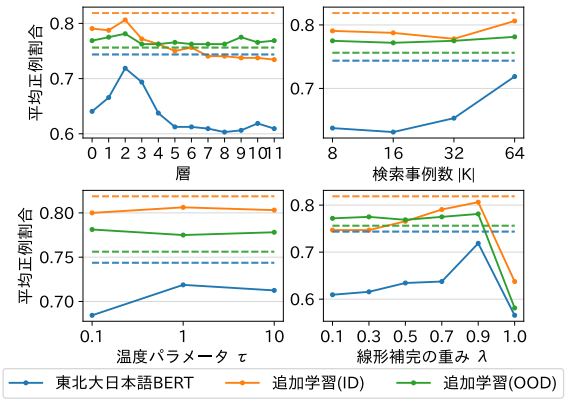

図 2　$N = 32$ において検索設定を変化させた時の最高性能の変化．実線は提案法の性能，破線は提案法を用いない場合の性能である．

4　分析

4.1　検索設定による性能変化

3.2 節で述べた検索設定の各項目が，ドメイン適応性能にどのように影響するか分析する．図 2 は，$N = 32$ において，それぞれの検索設定を x 軸の値に固定したまま他の設定を変更した時の，最も高い評価値をプロットしたものである．例えば，温度パラメータ τ における $x = 0.1$ のプロットは，$\tau = 0.1$ に固定したまま，その他の $|\mathcal{K}|$ や λ などの設定を全ての候補試行する．破線は提案法を用いない場合の性能であり，表 3 で報告した値と等しい．

図 2 から，2 層付近の浅い層，より大きな検索事例数，高い λ がより高いドメイン適応性能につながることが分かる．浅い層が有効なのは，文全体の意味よりも単語単位の表現を維持する表現が重要であることを示唆する．また，λ についての傾向は近傍事例から計算された表現が重要であることを示唆するが，$\lambda = 1.0$ では性能が急激に低下する．この理由については，式 2 において $\boldsymbol{h}_{k\mathrm{NN}}$ の方がノルムが小さい傾向にあり，$\lambda = 0.9$ のように極端な重みを割り当てることでノルムの差を解消していると考えられる．以上のような傾向は見られるものの，追加事前学習を行なった場合には検索設定の変化に対して性能はあまり変化しない．一方，東北大日本語 BERT では性能変化が顕著であり，適応するべきドメインの乖離が大きな場合には，検索設定がもつ性能への影響がより顕著に現れると考えられる．

表 4　提案法のドメイン適応のための検索における検索結果．ラベル付きデータに含まれる事例 “原因不明 レクチ端子台の焦げが発生する” をクエリとして，ラベルなしデータから検索された事例上位 3 件を示す．

原因不明 クランク曲げの線ガイドがコイルに干渉する
真因不明、コイルチャックがゆるい？インシュズレが普段より多く発生する
原因不明。 エキセンヘッド回転軸の位置ズレが発生。

4.2　近傍事例の提示による解釈性

提案法は 1 段階目の検索事例を提示することによって解釈性とすることができ，モデルの信頼性が向上したりエラー分析が容易になったりする．エラー分析の例を表 4 に示す．この例では，“原因不明” というフレーズに検索が依存しており，処置内容を知るためにより重要だと思われる “レクチ端子台”，“焦げ” などの単語が無視されている．このような近傍事例はドメイン適応において適切ではなく，得られた表現は後段の検索タスクではノイズになる可能性があり，改善の余地がある．

5　関連研究

k 近傍事例に基づいて BERT モデルを特定のタスクに適応させる方法は，単語穴埋め形式の質問応答 [4] や，の文分類タスク [5] について応用例がある．これらの先行研究では，k 近傍事例に紐つくラベルの情報を用いることで確率分布について線形補完するが，提案法では検索タスクを対象とするため埋め込み表現について線形補完する点が異なる．また，SimCSE [2] などの文書集合から埋め込みモデルを学習する方法との比較では，提案法は基本的に学習不要でドメイン適応可能な利点がある．

6　おわりに

本研究では，検索タスクにおいて追加学習なしで埋め込み表現のドメイン適応を行うために，適応先ドメインの文書集合からクエリに対応する k 近傍事例を検索し，元の表現と近傍事例の表現との線形補完を取ることでドメイン適応させる方法を提案した．独自に収集するデータに基づく故障事例検索タスクによる実験結果から，提案法がドメイン適応性能を有すること，検索結果の提示が解釈性を高めることを述べた．今後は，BEIR [6] などの公開データを用いて提案法の有効性を幅広く調査する予定である．

参考文献

[1] 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.

[2] Tianyu Gao, Xingcheng Yao, and Danqi Chen. SimCSE: Simple contrastive learning of sentence embeddings. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 6894–6910, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[3] Xing Han Lù. Bm25s: Orders of magnitude faster lexical search via eager sparse scoring, 2024.

[4] Nora Kassner and Hinrich Schütze. BERT-kNN: Adding a kNN search component to pretrained language models for better QA. In Trevor Cohn, Yulan He, and Yang Liu, editors, **Findings of the Association for Computational Linguistics: EMNLP 2020**, pp. 3424–3430, Online, November 2020. Association for Computational Linguistics.

[5] Linyang Li, Demin Song, Ruotian Ma, Xipeng Qiu, and Xuanjing Huang. Knn-bert: Fine-tuning pre-trained models with knn classifier, 2021.

[6] Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In **Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)**, 2021.

A　試行した検索設定

1 段階目の検索の設定について，次の候補を試行した．層については 0 層（埋め込み層）から 12 層までの層を試行した．式 1 の検索事例数 $|\mathcal{K}|$ は $|\mathcal{K}| = \{8, 16, 32, 64\}$，同式の温度パラメータ τ は $\tau = \{0.1, 1.0, 10.0\}$，式 2 の線形補完の重み λ $\lambda = \{0.1, 0.3, 0.5, 0.7, 0.9, 1.0\}$ を候補とした．

B　最適な検索設定

3.3 節で報告した結果におけるドメイン適応のための検索についてのハイパーパラメータを表 5 に示す．各モデルおよび各 N についてオラクルの設定が存在することに注意されたい．

表 5　それぞれのモデルおよび N における最適なハイパーパラメータ．

モデルと N	層	$\lvert\mathcal{K}\rvert$	τ	λ
東北大 BERT				
$N = 8$	2	64	1	0.9
$N = 16$	2	64	1	0.9
$N = 32$	2	64	1	0.9
$N = 64$	2	64	1	0.9
追加学習 (OOD)				
$N = 8$	5	64	10	0.9
$N = 16$	3	64	0.1	0.9
$N = 32$	2	64	0.1	0.9
$N = 64$	2	64	0.1	0.9
追加学習 (ID)				
$N = 8$	1	8	0.1	0.7
$N = 16$	2	64	0.1	0.9
$N = 32$	2	64	1	0.9
$N = 64$	1	16	0.1	0.7