文脈構造を利用した埋め込み表現学習の提案

原田慎太朗
株式会社東芝 研究開発センター
shintaro.hk4@toshiba.co.jp

概要

文や画像などの埋め込み表現は、検索をはじめとしたアプリケーションに利用されており、関連研究が盛んに進んでいる。先行研究では、文脈構造を明示的に利用せずに同一の文および画像を正例としてそれ以外を負例とする対照学習が主流である。しかし、埋め込み表現を構築する上で文脈構造を利用することは重要である。本稿では、文脈構造を考慮することで表現能力を向上させるための最適輸送を用いた教師なし埋め込み学習手法を提案する。結果として、STS-B および SICK-R における定量評価では、最高性能と同等以上の性能を達成し、定性評価では、文脈構造の利用により頑健かつ質の良い埋め込み表現が得られた。

1 はじめに

埋め込み表現とは、テキストや画像などを表現する密ベクトルであり、類似事例検索や検索拡張生成 (Retrieval Augmented Generation: RAG) などのアプリケーションに利用されている。近年、注目を集めている埋め込み表現学習として、同一のテキストおよび画像を拡張したものを正例とし、その他のテキストおよび画像を負例とする教師なし対照学習がある。これらは教師なし学習にも関わらず、質の高い埋め込み表現を獲得できることが自然言語処理 [1] および画像処理 [2] において確認されている。文の埋め込みに対する教師なし対照学習である SimCSE[1] は追加のモジュールが必要なくドロップアウト [3] による単純な正例の拡張から質の良い埋め込み表現を獲得できるフレームワークとして知られている。しかし、先行研究では、文の埋め込み表現だけに着目して対照損失を計算するため、理想的な文の埋め込みを組み立てるうえで重要な文脈構造をうまく捉え切れていないと考える。

文脈構造を理解する上で、語順を正しく理解することは重要であり、この並び変え問題は様々な語学教育で取り入れられている。本稿では、語順の並び変え問題を参考にして、文脈構造予測を反映した対照損失を提案する。具体的には、ある文に着目して、文を構成するトークンの語順を入れ替えることで正例を作成する。拡張した文から元の文へと輸送することで語彙の並び替えを学習する。つまり、シャッフルされたトークン表現が正しく文脈を捉えられているかの判定を最適輸送を介して対照損失に組み込む。これにより、文脈構造を考慮しつつ文の埋め込みを獲得することができる。また、追加のモジュールを必要としないため、追加コストが発生しない。実験の結果、定量評価において文脈構造の獲得がより質が高い文の埋め込み表現の獲得に寄与し、定性評価において文脈構造の獲得がノイズに頑健な類似事例検索に寄与していることを確認した。

2 最適輸送

最適輸送とは、ある分布と別の分布の距離を計算するためのアルゴリズムである。分布の間の距離および対応関係（アライメント）を得ることができるため、類似度計算 [4] や機械翻訳 [5] など様々な自然言語処理に利用されている。輸送問題を解くためには、入力として、輸送前の分布 μ_s、輸送後の分布 $\mu_{s'}$、分布間の輸送コストを定める距離尺度 d の 3 つを必要とする。

$$\mu_s = \left\{(x_i, m_i)\right\}_{i=1}^{n},\ \ \mu_{s'} = \left\{(x'_j, m'_j)\right\}_{j=1}^{n'} \tag{1}$$

ここで、μ は点 $x_i \in \mathbb{R}^d$ に質量 $m_i \in [0, 1]$ を対応させた確率分布である。なお、条件として $\sum_i m_i = 1$ を満たすものとする。

最適輸送は分布の間の輸送コストを最小化する最適化問題として下記のように記述される。

$$\min_{T \in u(\mu_s, \mu_{s'})} \sum_{i,j} T_{i,j} c(x_i, x'_j) \tag{2}$$

$$u(\mu_s, \mu_{s'}) = \{T \in \mathbb{R}^{n \times m}; T_{i,j} \geq 0, T\mathbb{1} = m, T^T\mathbb{1} = m'\} \tag{3}$$

ここで、$T \in R^{m \times n}$ は輸送計画行列であり、$c(x_i, x'_j)$ は x_i と x'_j の間の輸送コスト計算する関数である。なお、分布の点にはトークン埋め込み表現、分布の質量には一様な重み、輸送コストには、ユークリッド距離を用いる [6]。

輸送問題を解くことで、出力として、分布の間の最適輸送コストとそれぞれの分布に含まれる点の対応関係（輸送計画行列）を獲得できる。輸送計画行列に含まれる要素である輸送量は $T_{i,j}$ は x_i と x'_j が近ければ大きくなるため、x_i と x'_j の間の類似度として利用できる。

3 SimCSE：代表的な埋め込み学習

文の埋め込み表現学習に対照学習を用いた先行研究の 1 つである SimCSE は、教師なしに正例の組を作るために、ドロップアウトによるデータ拡張を採用している。具体的には、同じ埋め込み表現に対して異なるドロップアウトを適用することで、元の意味は同じだが異なる埋め込み表現を獲得する。そして、図 1 に示す通り、文の埋め込みである CLS トークンの潜在表現を対照損失を介して互いに近づくよう学習する。

ドロップアウト拡張　正例の拡張としてドロップアウトによる潜在空間でのトークンのマスキングを行う。文の長さが 10 の時、削除する割合を $p_{\text{dropout}} = 0.1$ とすると、潜在空間において 1 個のトークン表現がランダムにマスキングされることを示す。なお、ドロップアウト層は新たに追加するのではなく、モデルに含まれるドロップアウト層を利用する。

目的関数　ドロップアウト拡張により得られた正例の組を下記の目的関数を介して近づける。なお、教師なし対照学習では、同じ文から拡張された文のみを正例とし、バッチ内に含まれるその他の文を負例とする。そのため、似た意味の文にもかかわらず負例として扱う場合があり、これは問題である。そこで、損失関数に負例は利用しない。

$$L = -\log \frac{e^{\text{sim}(h_i^{z_i}, h_i^{\hat{z}_i})/\tau}}{\sum_{j=1}^{N} e^{\text{sim}(h_i^{z_i}, h_j^{\hat{z}_j})/\tau}} \tag{4}$$

ここで、N はバッチサイズ、h はモデルより得られた潜在表現、$\text{sim}(x, y) = \cos(x, y)$ は埋め込み表現である x と y の cosine 類似度を計算する関数、z と $\hat{z}$ は適用された異なるドロップアウト、τ は類似度に対する敏感さを制御する温度パラメータである。

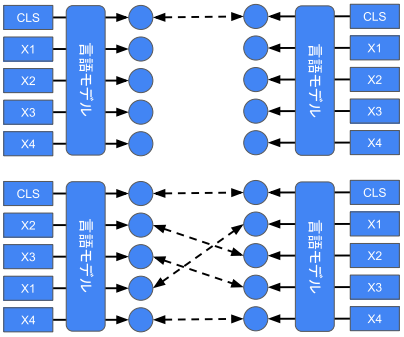

図 1　文の埋め込み表現における対照学習の概要。上が先行研究 [1, 7] で提案された CLS トークン間のみに着目する対応関係の例である。下が本稿で提案する全てのトークン間に着目する対応関係の例である。

4 最適輸送に基づく埋め込み学習

最適輸送を用いてシャッフルした文を正しい文に近づけることで並び替え問題を解く。並び替え問題として解くためには、ある文に対してシャッフルした文（問題）と何もしない文（正解）を用意して、ドロップアウト拡張により得られた正例の組を元の位置を参照して近づける。これにより、文脈構造を考慮した頑健な埋め込み表現が得られる。また、図 1 に示すように、ある 1 つの埋め込み表現に着目して対照学習を行う先行研究 [1, 7, 8] とは異なり、本研究は全ての埋め込み表現に着目して対照学習を行う。

並び替え拡張　正例の拡張としてシャッフルによるトークンの入れ替えを行う。文の長さが 10 の時、入れ替える割合を $p_{\text{shuffle}} = 0.1$ とすると、2 個のトークンがランダムに入れ替わることを示す。この時、シャッフルされる前の位置番号は保持して、対照損失の計算における正例の参照に利用する。なお、元の文とは関係がない特殊トークンの CLS や PAD 等は入れ替えの対象から外した。

目的関数　ドロップアウトおよび並び替え拡張により得られた正例の組を用いて、最適輸送で最適輸送行列を得る。そして、図 1 に示す通り、正例の位置番号を参照して最適輸送コストを類似度として下記の目的関数を介して近づける。なお、先行研究と同様に負例は利用しない。

$$L = \frac{1}{N} \sum_{k=1}^{N} -\log \frac{e^{\text{ot}(h_i^{z_i}, h_i^{\hat{z}_i}, d)/\tau}}{\sum_{j=1}^{T} e^{\text{ot}(h_i^{z_i}, h_j^{\hat{z}_j}, d)/\tau}} \tag{5}$$

ここで、N はバッチサイズ、T はトークン数、h はモデルより得られたトークンの埋め込み表現、

model	STS-B	SICK-R	Avg.
BERT	47.20	58.25	52.72
SimCSE	78.63	70.38	74.51
DiffCSE	80.25	**70.81**	75.53
Ours	**81.55**	70.66	**76.11**

表 1 STS-B と SICK-R における文の埋め込み表現の定量評価。最も高い値は太字で表す。

$\mathrm{ot}(x, y; d)$ は埋め込み表現である x と y の間の最適輸送量を計算する関数、z と $\hat{z}$ は適用された異なるドロップアウト、τ は類似度に対する敏感さを制御する温度パラメータである。

5 実験および評価

5.1 学習設定

訓練設定は、先行研究 [1] に従う。学習データは英語の Wikipedia データであり、テキストサイズはランダムにサンプリングされた 10^6 文である[1)]。モデルおよびトークナイザには事前学習済みの bert-base-uncased を用いた[2)]。学習に用いたハイパーパラメータは先行研究 [1] を参考にした。追加の最適輸送に伴うドロップアウト拡張と並び替え拡張のハイパーパラメータには、それぞれ $p_{\mathrm{dropout}} = 0.1$ と $p_{\mathrm{shuffle}} = 0.1$ を設定した。モデルの学習には NVIDIA RTX A6000 GPU 1 枚を用いた。比較対象として、同じ教師なしモデルである BERT[9]、SimCSE[1]、DiffCSE[7] を採用する。なお、モデルサイズは base で統一する。

5.2 評価結果

評価設定は、比較対象である先行研究 [1] に従う。評価ベンチマークには、STS-B と SICK-R を用いた。それぞれのベンチマークには、2 文の間の類似度 [0.0, 5.0] がアノテーションされている。評価指標には、モデルが予測したコサイン類似度と人手評価の類似スコアのスピアマンの相関係数を計算した。また、条件を整えるために、文の埋め込み表現には、CLS トークンの埋め込み表現ではなく、全てのトークンの平均埋め込み表現を用いた。

表 1 より、最高値を示す先行研究 [7] と比較して、STS-B では 1.3 ポイント向上、SICK-R では 0.15 ポイント低下、平均では 0.58 ポイント向上することが分かる。また、追加モジュールとして Generator と Dicriminator を必要とする DiffCSE と比較しても、追加モジュールなしに同等の結果を示すことが分かる。

1） https://github.com/princeton-nlp/SimCSE
2） https://huggingface.co/bert-base-uncased

6 考察

提案モデルが SimCSE よりも良い性能を示すことを説明するために、定性評価として類似事例検索を行う。STS-B のテストセットにおける全ての文（計 2,758 個）をデータベースとして利用して、クエリ文とデータベースに含まれる文をエンコードした埋め込み表現の間のコサイン類似度を計算することで、最近傍文を取得する。類似事例検索で得られた上位 3 例を表に示す。

1 つ目の検索クエリは、“unfortunately the answer to your question is we simply do not know.” である。両モデルともに正解を第 2 位の回答として取得している。しかし、第 1 位の回答として、SimCSE は、“my answer to your question is ‘probably not’.” を選択しており適切ではないが、提案モデルは “unfortunately, this question cannot be answered in its full generality.” と適切なものを選択している。これより、提案モデルは文脈の意味を正しく理解して文の埋め込み表現を構築できている。2 つ目の検索クエリは、“a man sleeps with a baby in his lap.” である。SimCSE は正解の回答を第 1 位の回答として取得しているが、提案モデルは第 2 位の回答として取得している。ただし、“赤ん坊を膝にのせて寝る” という情景は後者のほうが適切であり、提案モデルは文脈の意味を正しく文の表現に反映できている。

提案モデルが SimCSE よりも頑健に動作することを説明するために、定性評価として類似事例検索を行う。データベースは前述と同様の STS-B のテストセットであり、クエリ文とデータベースに含まれる文をエンコードした文の埋め込み表現の間のコサイン類似度を計算することで、最近傍文を取得する。類似事例検索で得られた上位 3 例を表に示す。

クエリは、前述した類似事例検索と同じであるが、なるべく意味が変わらないように語順を入れ替えた。1 つ目の検索結果の順位について、SimCSE では全ての順位に対して変動があるが、提案モデルでは順位の変動がない。2 つ目の検索の順位について、両モデルにおいて順位の変動を確認した。これより、提案モデルのほうが、文法的に正しくないクエリを与えても頑健に機能することが分かる。これは、類似事例検索として有用な要素である。

	SimCSE	Ours
Query: unfortunately the answer to your question is we simply do not know.		
#1	my answer to your question is ‘probably not’.	unfortunately, this question cannot be answered in its full generality.
#2	sorry, i don’t know the answer to your question.	sorry, i don’t know the answer to your question.
#3	i think that the short answer to your question is: no.	my answer to your question is ‘probably not’.
Query: a man sleeps with a baby in his lap.		
#1	a man asleep in a chair holding a baby.	a father napping in a chair with a baby on his lap.
#2	a father napping in a chair with a baby on his lap.	a man asleep in a chair holding a baby.
#3	man with pink shirt sleeping on chair with infant.	man with pink shirt sleeping on chair with infant.

表 2　文法的に正しい文に対する SimCSE と提案モデルを用いて検索された上位 3 事例。左の列は検索の順位を示す。

	SimCSE	Ours
Query: unfortunately we simply do not know the answer to your question is.		
#1	#2; sorry, i don’t know the answer to your question.	#1; unfortunately, this question cannot be answered in its full generality.
#2	#3; this doesn’t answer your question.	#2; sorry, i don’t know the answer to your question.
#3	#1; my answer to your question is ‘probably not’.	#3; this doesn’t answer your question.
Query: a baby in his lap a man sleeps.		
#1	#1; a man asleep in a chair holding a baby.	#1; a father napping in a chair with a baby on his lap.
#2	#3; man with pink shirt sleeping on chair with infant.	#3; man with pink shirt sleeping on chair with infant.
#3	#2; a father napping in a chair with a baby on his lap.	#2; a man asleep in a chair holding a baby.

表 3　文法的に正しくない文に対する SimCSE と提案モデルを用いて検索された上位 3 事例。左の列は検索の順位を示し、セミコロン前の番号は図 2 の類似事例検索での順位を示す。

7　おわりに

本稿では、文脈構造を捉え頑健な文の埋め込みを構築するために、語順の並び替え問題から着想を得た、最適輸送による対照学習手法を提案し検証した。結果として、追加のモジュールなしに先行研究と比較して同等以上の性能を発揮することを確認した。さらに、文の埋め込み表現の構築に文脈構造の利用が重要であることを確認した。ただし、正例を作るための並び替え拡張により入力の意味が変化する場合があり、これは今後の課題とする。

今後は、正例の拡張および損失関数におけるバリエーション（ハイパーパラメータの調整、潜在空間での正例拡張、負例の利用）の調査と日本語 [10] を含めた様々な言語の評価データセットにおいて検証に取り組みたい。また、提案手法は、画像パッチをトークンと見立てた画像処理モデルにも応用可能であり、画像埋め込み表現の質についても調査に取り組みたい。

参考文献

[1] Tianyu Gao, Xingcheng Yao, and Danqi Chen. SimCSE: Simple contrastive learning of sentence embeddings. In **Empirical Methods in Natural Language Processing (EMNLP)**, 2021.

[2] Kouta Nakata Yaling Tao, Kentaro Takagi. Clustering-friendly representation learning via instance discrimination and feature decorrelation. **Proceedings of ICLR 2021**, 2021.

[3] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. **Journal of Machine Learning Research**, Vol. 15, No. 56, pp. 1929–1958, 2014.

[4] Sho Yokoi, Ryo Takahashi, Reina Akama, Jun Suzuki, and Kentaro Inui. Word rotator’s distance. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 2944–2960, Online, November 2020. Association for Computational Linguistics.

[5] Jingjing Xu, Hao Zhou, Chun Gan, Zaixiang Zheng, and Lei Li. Vocabulary learning via optimal transport for neural machine translation. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, **Proceedings of the**

59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 7361–7373, Online, August 2021. Association for Computational Linguistics.

[6] Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. From word embeddings to document distances. In **International Conference on Machine Learning**, 2015.

[7] Yung-Sung Chuang, Rumen Dangovski, Hongyin Luo, Yang Zhang, Shiyu Chang, Marin Soljacic, Shang-Wen Li, Scott Yih, Yoon Kim, and James Glass. DiffCSE: Difference-based contrastive learning for sentence embeddings. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, **Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 4207–4218, Seattle, United States, July 2022. Association for Computational Linguistics.

[8] Yixuan Su, Fangyu Liu, Zaiqiao Meng, Tian Lan, Lei Shu, Ehsan Shareghi, and Nigel Collier. Tacl: Improving BERT pre-training with token-aware contrastive learning. **CoRR**, Vol. abs/2111.04198, , 2021.

[9] 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.

[10] Hayato Tsukagoshi, Ryohei Sasano, and Koichi Takeda. Japanese SimCSE Technical Report. **arXiv:2310.19349**, 2023.