ガウス埋め込みに基づく文表現生成

陽田 祥平 [1]　塚越 駿 [2]　笹野 遼平 [2]　武田 浩一 [2]
[1] 名古屋大学情報学部　[2] 名古屋大学情報学研究科
wt.50p.8613@s.thers.ac.jp tsukagoshi.hayato.r2@s.mail.nagoya-u.ac.jp
{sasano,takedasu}@i.nagoya-u.ac.jp

概要

近年、文の持つ情報を埋め込み空間上の点として表現する文ベクトルの研究が盛んである。しかし、点による文の表現は、文の持つ意味の広がりや包含関係などの文同士の非対称的な関係を表現できないなど、文が持つ多様な情報の一部しか表現できない。そこで本研究では、文をガウス分布として領域的に埋め込む手法、および、包含関係の認識のための類似度指標を提案する。実験を通し、自然言語推論タスクにおいて従来の点表現ベースの手法と同等の性能を達成できること、点表現では困難であった包含関係の向きの推定が可能であることを示す。

1　はじめに

文の持つ情報をベクトルで表現する文埋め込みは、文書分類や類似文検索、質問応答など様々な自然言語に関するタスクで用いられている。近年では、事前学習済み言語モデルを用いる機械学習ベースの文埋め込みの生成法が主流となっており、トピックや感情といった文が持つ情報を豊かに表現する文埋め込み生成の試みが広く行われて文埋め込み生成の主なモデルとしては、Siamese Network を用いて BERT [1] を微調整する Sentence-BERT [2] や、正例と負例の文ペアを構築して対照学習を行う SimCSE [3] などが挙げられる。しかし、これらは文をベクトルとして表現する手法であり、文同士の類似度の指標としては基本的にコサイン類似度のような対称的な指標が用いられることから、包含関係や階層構造など、2 文間の非対称的な関係を捉えることができない。

そこで本稿では、単語をガウス分布として表現するガウス埋め込みを基に、図 1 に示すように文をガウス分布として表現する手法を提案する。また、KL ダイバージェンスを用いて文同士の階層構造を表現可能な非対称的な類似度指標を提案する。自然言語推論 (Natural Language Inference; NLI) 分類タスク、および包含関係にある文ペアに対してどちらが包含する側であるかを推測するタスクを用いた実験から、提案手法が従来の文表現と意味表現において同等の性能を保ちつつ、文同士の包含関係を表現できることを示す。

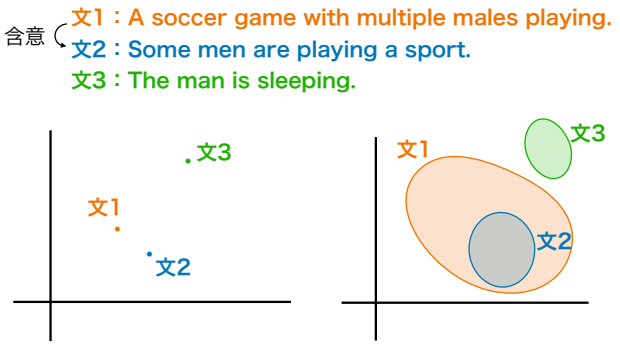

図 1　従来の点埋め込み (左図) と提案手法による埋め込み (右図) の概要図。点埋め込みでは包含関係にある文 1 と文 2 のうちどちらが他方を包含するかを表現できない一方、提案手法では包含関係を表現できる。

2　ガウス埋め込みに基づく文表現

本研究では、事前学習済み言語モデルに対照学習による微調整を適用して、包含関係を適切に捉えた文のガウス埋め込みを獲得する。本節ではまず、ガウス埋め込みの代表的な研究である Gaussian Embedding、および対照学習を用いた文埋め込みの獲得手法である SimCSE について述べ、続いてそれらを組み合わせた提案手法について述べる。

2.1　Gaussian Embedding

ガウス埋め込みの先行研究として、グラフの各ノード [4,5] やレビュー [6] をガウス分布で表現するものがあるが、その代表的な手法として、単語をガウス分布として表現する Gaussian Embedding [7] が挙げられる。Gaussian Embedding では単語 w_i を、平

均ベクトル μ_i と分散共分散行列 Σ_i を用いて次式のように表現する。

$$w_i = \mathcal{N}(x; \mu_i, \Sigma_i) \tag{1}$$

平均ベクトルが従来の点での表現に相当し、分散共分散行列が意味の広がりを表す。また 2 つの単語間の類似度として、次式で示す KL ダイバージェンスを用いている。

$$D_{KL}(N_i||N_j) = \int_{x \in \mathbb{R}^n} \mathcal{N}(x; \mu_i, \Sigma_i) \ln \frac{\mathcal{N}(x; \mu_i, \Sigma_i)}{\mathcal{N}(x; \mu_j, \Sigma_j)} \tag{2}$$

KL ダイバージェンスは左辺の引数を入れ替えることで値が変わる非対称的な指標であるため、階層構造といった埋め込み同士の非対称的な関係を捉えることができる。Vilnis ら [7] は KL ダイバージェンスを基に Skip-gram [8] に倣った手法を用いてモデルを学習させることで、単語の意味の広がりや含意関係を捉えた表現を構成できることを示した。

2.2 Supervised SimCSE

近年、文埋め込みの生成法についての研究が非常に盛んである [9–14]。その中でも代表的なものの一つに Supervised SimCSE [3] がある。Supervised SimCSE は、NLI データセットを用いた対照学習によって、文埋め込みモデルを学習する。NLI データセットは前提文と仮説文のペアで構成されており、各ペアには、前提文が仮説文を含意することを示す「含意」、前提文が仮説文と矛盾することを示す「矛盾」、含意でも矛盾でもないことを示す「中立」のいずれかのラベルが人手で付与されている。Supervised SimCSE では含意のラベルが付与されている文ペアを正例として埋め込みを近づけ、矛盾のラベルが付与されている文ペアおよびバッチ内で異なるペアに属する 2 文を負例として埋め込みを遠ざける対照学習を行うことで、2 文間の類似度を推定する Semantic Textual Similarity (STS) タスクにおいて高い性能を達成している。

2.3 提案手法

本研究では、文 s_k の意味をガウス分布 N_k として表現し、NLI データセットを用いて教師あり対照学習を行うことで、文のガウス埋め込みを獲得する。提案手法の概要を図 2 に示す。まず文 s_k を BERT に入力し、[CLS] トークンから文のベクトル表現 v_k を獲得する。続いて、v_k を 1 層の線形層からなるネットワーク 2 つに入力し、得られた出力をそれぞ

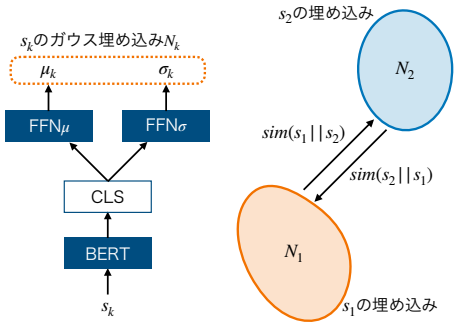

図 2 提案手法の概要図 (左図) と類似度 sim の概念図 (右図)。sim は非対称的であるため、$\text{sim}(s_1||s_2)$ と $\text{sim}(s_2||s_1)$ は異なる値をとる。

れガウス分布 N_k の平均ベクトル μ_k、分散共分散行列の対角成分 σ_k とする。なお本手法では計算の効率化のため、Gaussian Embedding と同様に分散共分散行列の対角成分のみを分散の表現として用いる。以降 σ_k は分散ベクトルと呼ぶ。

次に、文 s_j に対する文 s_i の類似度指標 $\text{sim}(s_i||s_j)$ を式 (3) により定義する。

$$\text{sim}(s_i||s_j) = \frac{1}{1 + D_{KL}(N_i||N_j)} \tag{3}$$

KL ダイバージェンスの値域は $[0, \infty)$ であることから、$\text{sim}(s_i||s_j)$ の値域は $(0, 1]$ となる。また、本指標は引数を入れ替えると値が変化する非対称な指標であり、N_i の分散が N_j の分散より大きい場合、$D_{KL}(N_i||N_j)$ の方が $D_{KL}(N_j||N_i)$ より大きくなる傾向があることから、$\text{sim}(s_j||s_i)$ の方が $\text{sim}(s_i||s_j)$ より大きくなる傾向がある。

包含関係の学習では Gaussian Embedding と同様に、より多様な文を包含する文ほどその埋め込みの分散が大きくなるように学習を行う。これを実現するため含意の文ペアを教師データとして用いて、包含する側である前提文の分散を大きく、包含される側である仮説文の分散を小さくなるように学習する。これは、上述した類似度指標の性質から、sim(仮説文 || 前提文) が sim(前提文 || 仮説文) より大きくなるように学習することで実現が可能である。また、含意関係にない、すなわち意味的に類似していない文ペアについては、sim(仮説文 || 前提文) が小さくなるように学習する。KL ダイバージェンスは分散に比べて平均の変化に大きく影響を受けるという性質があることから、この操作により 2 文の平均ベクトルの距離が大きくなることが期待される。

モデルの学習には、Supervised SimCSE と同様、NLI データセットを用いた対照学習を行う。対照学

習では、正例となる文同士の類似度が大きくなるように学習を行うと同時に、負例となる文同士の類似度が小さくなるように学習を行う。本研究では、正例と負例の選定法として以下の 3 種類を用いる。

含意集合　含意ラベルが付与された前提文と仮説文のペアの集合。意味的に類似した文同士が近づくよう、含意関係にある文同士を正例とする。

矛盾集合　矛盾ラベルが付与された前提文と仮説文のペアの集合。含意関係にない文同士が離れるよう、矛盾関係にある文同士を負例とする。

逆向き集合　含意集合の各ペアにおける前提文と仮説文を入れ替えたもの。包含する側である前提文の分散を大きく、包含される側である仮説文の分散を小さくなるように学習するため、逆向きの含意関係にある文同士を負例とする。

図 3 に、あるバッチにおける、上記 3 つの正例と負例の選定法の概要図を示す。

正例と負例の集合の各要素に対し sim(仮説文 || 前提文) を計算し、softmax 関数を適用したもののクロスエントロピー誤差を損失関数として定義する。データセット中の前提文、含意ラベルが付与された仮説文、矛盾ラベルが付与された仮説文のガウス埋め込みをそれぞれ N_i、N_i^+、N_i^- とし、損失関数のうち含意集合、矛盾集合、逆向き集合に対応する項をそれぞれ V_E、V_C、V_R としたとき、各バッチにおける損失関数は次式のように表せる。なお n はバッチサイズ、τ は sim の増幅の度合いを調整する温度係数である。

$$
\begin{aligned}
V_E &= \Sigma_{j=1}^{n} e^{\mathrm{sim}(N_j^+ || N_i)/\tau}, \\
V_C &= \Sigma_{j=1}^{n} e^{\mathrm{sim}(N_j^- || N_i)/\tau}, \\
V_R &= \Sigma_{j=1}^{n} e^{\mathrm{sim}(N_j || N_i^+)/\tau}, \\
\mathcal{L} &= \sum_{i=1}^{n} -\ln \frac{e^{\mathrm{sim}(N_i^+ || N_i)/\tau}}{V_E + V_C + V_R}
\end{aligned}
\tag{4}
$$

以上の手法により、意味的に近い文同士はその埋め込みの平均が近い値をとり、かつ包含関係にある 2 文では包含する側の分散が大きく、包含される側の分散が小さく学習されることが期待できる。

3　評価実験

NLI 分類と包含の向き推定という 2 つのタスクにより、提案手法により得られるガウス分布に基づく文表現の性能評価を行った。

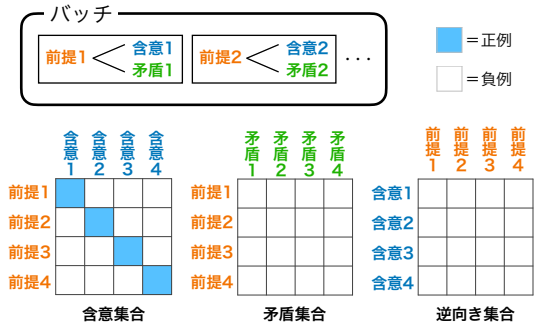

図 3　学習時における正例と負例の選定法の概要図。行列の各要素が sim(縦軸の文 || 横軸の文) に対応し、正例であれば 1 に、負例であれば 0 に近づける。

3.1　評価方法

NLI 分類　包含関係の認識においての既存の手法との性能を比較するため、NLI 分類を行なった。NLI 分類は、前提文と仮説文が提示され、前提文が仮説文を含意しているか、2 文の内容が矛盾しているか、関係がないかのいずれであるかを推測するタスクである。一般には上記の 3 値分類を行うが、本研究では NLI 分類を 2 値分類で行う。具体的には、sim(仮説文 || 前提文) の値が閾値以上であれば含意、そうでなければその他とする。

検証には Stanford NLI (SNLI) データセット [15] と SICK [16] データセットを使用し、分類の正解率と Precision-Recall (PR) 曲線の AUC を評価指標とした。分類の正解率は、開発セットにおいて閾値を 0 から 1 の間で 0.001 ずつ変化させた中で最も正解率が高くなる閾値を算出し、その閾値でテストセットの分類をした際の評価値を用いた。なおどちらのデータセットも各文ペアに含意、中立、矛盾の 3 種類のラベルが付与されているため、中立と矛盾を合わせてその他とした。またベースラインとして、学習済みの Supervised SimCSE モデル[1)]で同様の手法を用いた際の性能も評価した。

包含の向き推定　提案手法で獲得された文表現が包含関係を階層構造として捉えられるかを評価するために、含意関係にある事が分かっている 2 文 A、B に対し、どちらが包含する側の文であるかを予測するタスクを行う。包含の向きを決定する指標として、$\mathrm{sim}(A||B) < \mathrm{sim}(B||A)$ であれば A、そうでなければ B を包含する側だとする類似度ベースのものと、埋め込みの各次元の分散の積を $\det(\sigma_A)$、$\det(\sigma_B)$ とし、$\det(\sigma_A) > \det(\sigma_B)$ であれば A、そうで

1)　https://huggingface.co/princeton-nlp/sup-simcse-bert-base-uncased

なければ B を包含する側とする分散ベースのものを用いる。検証には SNLI と SICK のそれぞれのテストセットのうち、含意ラベルが付与されているもののみを用いた。

3.2 実験設定

事前学習済みモデルには HuggingFace[2] が公開しているライブラリである Transformers [17] から、BERT-base[3] を利用した。Supervised SimCSE と同様に、訓練データとして SNLI と Multi-Genre NLI (MNLI) データセット [18] を組み合わせたものを利用した。その他の詳細な実験設定は付録 A に示す。

損失関数については 2.3 節で述べた、含意集合、矛盾集合、逆向き集合の全てを用いる損失関数に加え、矛盾集合および逆向き集合が性能にもたらす影響を検証するため、含意集合のみ、含意集合と矛盾集合、含意集合と逆向き集合、含意集合と矛盾集合と逆向き集合の 4 種類の損失関数で実験を行った。

3.3 実験結果

NLI 分類タスクにおける実験結果を表 1 に示す。実験した 4 つの損失関数の中では、Supervised SimCSE での負例の組み合わせと同じ、含意+矛盾が最も高い性能を示した。含意と含意+矛盾、含意+逆向きと含意+矛盾+逆向きをそれぞれ比較すると、負例に矛盾集合を含むことで性能が向上し、SICK では Supervised SimCSE の性能を上回っていることが確認できる。一方、含意と含意+逆向き、含意+矛盾と含意+矛盾+逆向きに注目すると、負例に逆向き集合を含むことで性能が低下している。含意の文ペアは意味的に類似しているものが多く、それらを負例として用いる逆向き集合は NLI 分類においては性能に悪影響を与えると考えられる。

包含の向き推定の性能を表 2 に示す。包含の向き推定では、負例に逆向き集合を含むことで性能が大幅に上昇し、特に SNLI では約 97%という非常に高い割合で正しく包含の向きを推定する事ができた。また、SICK においては包含+逆向きが最も高い性能を示し、約 72%の割合で正しく包含の向きを推定できることがわかった。包含の向き推定では、SNLI と SICK の間で平均的な性能に大きな差が見られた。これは、SNLI と SICK で、文ペアの文長の比についての傾向が異なるためであると考えられる。SNLI は SICK と比較して、前提文が仮説文より長い傾向が強いため[4]、文が長いほど分散が大きくなるように学習を行うことで、包含の向き推定が比較的簡単にできるようになると考えられる。一方で SICK は、前提文と仮説文の長さの比が平均してほとんど同じであるため、文長の情報は包含の向き推定に寄与しない。したがって、このような文長によるバイアスを利用できないことが、SICK の性能が相対的に低いことの要因であると考えられる。これに対して、提案手法のうち、逆向き集合を含む設定は SNLI と SICK 双方の性能が相対的に高くなっている。このことから、逆向き集合を用いる学習手法が、文長のバイアスの影響を軽減し、意味的に妥当な文の包含関係を捉えるのに有用であることがわかった。

表 1 各手法の NLI 分類における正解率と PR 曲線の AUC。表中の Sup-SimCSE は Supervised SimCSE を表す。

	SNLI		SICK	
	正解率	AUC	正解率	AUC
Sup-SimCSE	74.96	66.76	**86.11**	**81.41**
含意	72.50	61.44	76.50	64.33
含意+矛盾	**78.33**	**72.50**	85.21	80.13
含意+逆向き	70.64	57.57	74.24	57.31
含意+矛盾+逆向き	77.51	69.33	84.37	78.92

表 2 包含の向き推定の実験結果。類似度ベースの手法の性能を「類似度」、分散ベースの手法の性能を「分散」の列に記載している。

	SNLI		SICK	
	類似度	分散	類似度	分散
含意	81.65	81.19	63.01	62.44
含意+矛盾	61.56	60.44	61.62	62.18
含意+逆向き	97.01	97.12	**71.23**	**71.93**
含意+矛盾+逆向き	**97.09**	**97.21**	69.22	70.13

4 おわりに

本研究では従来の文埋め込み手法が抱える、文の意味の広がりや包含関係を表現できないという問題を解決するため、事前学習済み言語モデルと対照学習を用いた、ガウス埋め込みに基づく文表現の生成手法を提案した。評価実験の結果、従来のベクトルによる文表現と同等の NLI 分類性能を達成したことに加え、非対称的な類似度を用いた学習により、従来では困難であった包含の向きの推定においても高い性能を達成した。今後はその他のタスクにおける性能や、平均・分散ベクトルの分布に着目した埋め込みそのものの解析も行なっていきたい。

2) https://huggingface.co
3) https://huggingface.co/bert-base-uncased
4) 実際の SNLI と SICK の文長の比を付録 B に示す。

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 **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)**, pp. 4171–4186, 2019.

[2] 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)**, pp. 3982–3992, 2019.

[3] Danqi Chen Tianyu Gao, Xingcheng Yao. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, p. 6894—6910, 2021.

[4] Shizhu He, Kang Liu, Guoliang Ji, and Jun Zhao. Learning to Represent Knowledge Graphs with Gaussian Embedding. **Proceedings of the 24th ACM International on Conference on Information and Knowledge Management**, 2015.

[5] Aleksandar Bojchevski and Stephan Günnemann. Deep Gaussian Embedding of Graphs: Unsupervised Inductive Learning via Ranking. In **6th International Conference on Learning Representations (ICLR)**, 2018.

[6] Danushka Bollegala Ichiro Sakata Masaru Isonuma, Junichiro Mori. Unsupervised Abstractive Opinion Summarization by Generating Sentences with Tree-Structured Topic Guidance. In **Transactions of the Association for Computational Linguistics (TACL)**, p. 945—961, 2021.

[7] Luke Vilnis and Andrew McCallum. Word Representations via Gaussian Embedding. **arXiv:1412.6623**, 2015.

[8] Kai Chen Greg Corrado Jeffrey Dean Tomas Mikolov, Ilya Sutskever. Distributed Representations of Words and Phrases and their Compositionality. **arXiv:1310.4546**, 2013.

[9] Zuozhu Liu Kwan Hui Lim Lidong Bing Yan Zhang, Ruidan He. An Unsupervised Sentence Embedding Method by Mutual Information Maximization. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 1601–1610, 2020.

[10] Shaohan Huang Zihan Zhang Deqing Wang Fuzhen Zhuang Furu Wei Haizhen Huang Denvy Deng Qi Zhang Ting Jiang, Jian Jiao. PromptBERT: Improving BERT Sentence Embeddings with Prompts. **arXiv.2201.04337**, 2021.

[11] Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. On the Sentence Embeddings from Pre-trained Language Models. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 9119–9130, 2020.

[12] Hongyin Luo Yang Zhang Shiyu Chang Marin Soljačić Shang-Wen Li Wen-tau Yih Yoon Kim James Glass Yung-Sung Chuang, Rumen Dangovski. DiffCSE: Difference-based Contrastive Learning for Sentence Embeddings. In **Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)**, pp. 4207–4218, 2022.

[13] Moin Nabi Tassilo Klein. SCD: Self-Contrastive Decorrelation of Sentence Embeddings. In **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)**, pp. 394–400, 2022.

[14] Niklas Muennighoff. SGPT: GPT Sentence Embeddings for Semantic Search. **arXiv.2202.08904**, 2022.

[15] Christopher Potts Christopher D. Manning Samuel R. Bowman, Gabor Angeli. A large annotated corpus for learning natural language inference. In **Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, p. 632—642, 2015.

[16] Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. A SICK cure for the evaluation of compositional distributional semantic models. In **Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)**, pp. 216–223, 2014.

[17] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-Art Natural Language Processing. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP): System Demonstrations**, pp. 38–45, 2020.

[18] Samuel Bowman Adina Williams, Nikita Nangia. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In **Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)**, p. 1112—1122, 2018.

[19] Frank Hutter Ilya Loshchilov. Decoupled Weight Decay Regularization. In **7th International Conference on Learning Representations (ICLR)**, 2019.

A　詳細な実験設定

提案手法によるモデルの学習では、SimCSE と同様、埋め込みの次元数を 768、エポック数を 3、温度係数を 0.05、最適化手法を AdamW [19] とした。バッチサイズ、学習率についてはそれぞれ {16, 32, 64, 128, 256}、{1e-5, 3e-5, 5e-5} の範囲で探索を行い、後述するモデルの学習中の評価において最も性能の高い組み合わせを使用した。また学習率は、学習の開始時点から線形に減衰するよう学習率スケジューリングを行った。各ハイパーパラメータの組み合わせごとの結果を表 3 に示す。

各実験では 100 step 毎に SNLI の開発セットを用いた NLI 分類タスクにおける PR 曲線の AUC を算出し、最も性能が高かった時点のモデルを最終的な性能評価に用いた。異なる乱数シード値で 5 回評価を行った際の平均を評価スコアとした。

表 3　各バッチサイズと学習率ごとのモデルに対する NLI 分類の PR 曲線の AUC。表内の数値には 100 をかけている。

		学習率		
		1e-5	3e-5	5e-5
バッチサイズ	16	64.39	66.02	**67.00**
	32	63.21	65.23	65.97
	64	60.79	64.13	65.27
	128	59.93	62.72	63.60
	256	58.86	61.50	62.52

B　前提文と仮説文の文長の比

SNLI および SICK の文ペアに対し、前提文と仮説文の長さの比の対数をとった値についてのヒストグラムを図 4 に示す。SICK は文長の差がないことを表す 0 付近に分布が集中しているが、SNLI は正の領域に分布が集中している。この事から、SNLI では前提文が仮説文より長い傾向にあることがわかる。

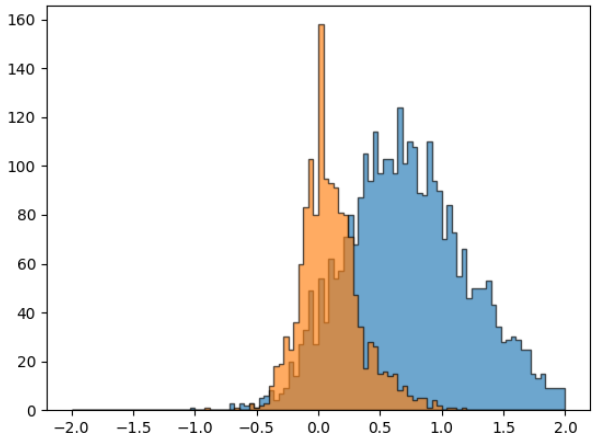

図 4　前提文と仮説文の長さの比のヒストグラム。横軸は前提文と仮説文の長さの比の自然対数、縦軸は文ペアの数を表し、青が SNLI、橙が SICK を表す。