自動生成した NLI データを用いた教師なし文埋め込みの改良

佐藤 蒼馬 [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] 名古屋大学大学院情報学研究科
{sato.soma.y7,tsukagoshi.hayato.r2}@s.mail.nagoya-u.ac.jp
{sasano,takedasu}@i.nagoya-u.ac.jp

概要

デコーダ系の大規模言語モデル (LLM) は自然言語処理の多くのタスクにおいて高い性能を示しており、文埋め込み生成においても PromptEOL [1] というデコーダ系モデルが Semantic Textual Similarity (STS) タスクにおいて最高性能を達成している。しかし、PromptEOL が高い性能を示すのは、人手で構築された自然言語推論 (Natural Language Inference: NLI) データセットを用いて fine-tuning した場合であり、人手で構築されたデータを利用しない場合の STS の性能は 6 ポイント程度低い値となっている。本研究では LLM を用いて NLI データセットを自動生成し、PromptEOL の fine-tuing に利用することで、教師なし設定における文埋め込み生成の高性能化を目指す。STS タスクで評価した結果、人手で構築された大規模なデータセットを利用しない設定において 82.21 という既存手法を上回る性能を達成した。

1 はじめに

文埋め込みは検索や含意関係認識など多くのタスクに利用できることから、広く研究されている。特に、事前学習済み言語モデルを fine-tuning することで文埋め込みを生成する手法が多く提案されている。例えば、エンコーダ系のモデルを用いたものとして、NLI 分類で BERT [2] を fine-tuning する SentenceBERT [3]、定義文を用いて BERT を fine-tuning する DefSent [4]、対照学習により BERT を fine-tuning する SimCSE [5]、プロンプトベースの文埋め込み手法により静的なトークン埋め込みのバイアスを取り除く PromptBERT [6]、エンコーダ・デコーダ系のモデルを用いたものとして、NLI データセットおよび QA データセットを用いて T5 [7] を fine-tuning する Sentence-T5 [8] 等が挙げられる。

また近年、デコーダ系の LLM に基づく手法が種々のベンチマークタスクで高い性能を示している。文埋め込み生成についても、デコーダ系の LLM をバイ・エンコーダとして使用し、埋め込みを生成する SGPT [9] や、一つの単語のみに焦点を当てる制限を設けたプロンプトベースの手法で文埋め込みを生成する PromptEOL [1] 等が提案されており、PromptEOL は人手で構築されたデータを利用する設定において、現時点でもっとも高い STS の性能を達成している。しかし、人手で構築された NLI データセットを利用しない場合、それより大きく低い性能にとどまっている。

本研究では LLM を利用して NLI データセットを自動構築し、構築した NLI データセットを PromptEOL の fine-tuning に利用することで、人手で構築したデータを利用しない設定においても高い性能を持つ文埋め込みモデルの構築を目指す。本研究で提案するフレームワークの概要を図 1 に示す。まず、Wikipedia から抽出した文に対し、それが含意する文、それと矛盾する文を生成するようなプロンプトをそれぞれ適用し、NLI データセットを自動生成する。続いて、自動生成された NLI データセットを用いて PromptEOL の fine-tuning を行い、fine-tuning 済みのモデルを用いて文埋め込みを生成する。

2 PromptEOL

PromptEOL はデコーダ系の大規模言語モデルに入力するプロンプトを工夫することで質の高い文埋め込みを生成する手法である。図 1 下部の埋め込み生成で示すように、「`This sentence: "[text]" means in one word: "`」の [text] を埋め込みを生成したい文に置換し、「`in one word: "`」の直後の埋め込みを、入力文の埋め込みとして用いる。デコーダ系の大規模言語モデルは次単語予測タスクで事前訓練されたモデルであるため、入力文を 1 単語で言い換えるような単語埋め込みを出力させるように入力を工夫することで、文埋め込みを生成することができる。

SimCSE の教師あり学習同様、NLI データセット

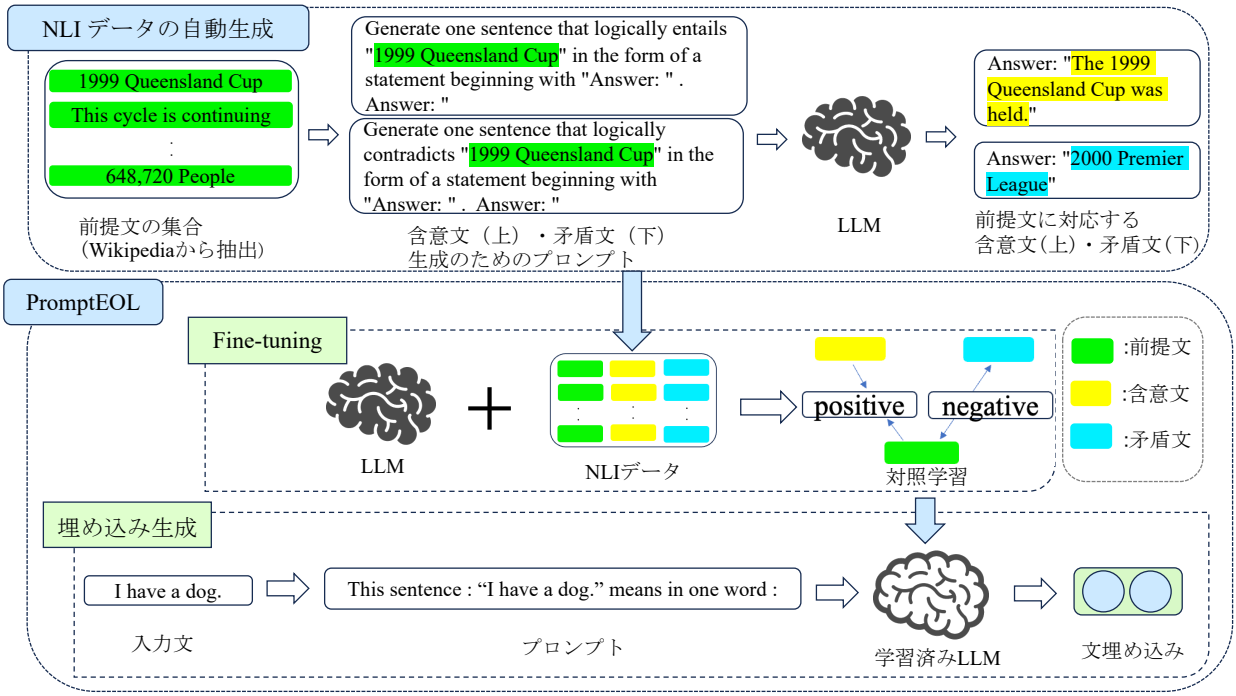

図 1　自動生成した NLI データを用いた教師なし文埋め込みの改良の概要

で対照学習した LLM を用いることで、より高性能な文埋め込みモデルを得ることが可能となる。図 1 の下部に示すように、NLI データセットを用いる場合、PromptEOL は含意関係にある文の埋め込みが近づき、矛盾関係にある文の埋め込みが遠ざかるように LLM を fine-tuning した上で、fine-tuning 後の LLM を用いて文埋め込みを生成する。NLI データセットを用いて fine-tuning を行うことで、STS タスクにおいて 6 ポイント程度高いスコアを達成することが報告されている [1]。

3　NLI データセットの自動生成

本研究では、LLM を用いて自動生成した NLI データを用いて PromptEOL を fine-tuning することで、人手で構築された大規模なデータセットを利用せずに、高性能な文埋め込み生成モデルを構築することを目的とする。本節では本研究における NLI データの生成手順について説明する。

3.1　既存の NLI データセット

NLI データセットは、前提文 (premise) と仮説文 (hypothesis) からなる文のペアに対し、含意 (entailment)、中立 (neutral)、矛盾 (contradiction) のいずれかのラベルが付与されたデータセットである。人手で構築された代表的な NLI データセットとして Stanford NLI (SNLI) [10] コーパスや、Multi-Genre NLI (MNLI) コーパス [11]、Cross-Lingual NLI (XNLI) コーパス [12] があり、それぞれ 579,000、433,000、112,500 個の文ペアで構成されている。SimCSE や PromptEOL が SNLI コーパスおよび MNLI コーパスを用いているなど、NLI データセットは多くの文埋め込み生成モデルにおいて利用されている。本研究では、SNLI コーパスと MNLI コーパスを統合したデータを利用することとし、以降ではこのデータのことを人手 NLI データセットと呼ぶ。

3.2　NLI データの自動生成

前提文となる文を、以下のプロンプトの {premise} の部分に置換し LLM に入力することで、それぞれ前提文に対応する含意文、矛盾文を生成する。この際、「Answer: "」に続いて出力された、次の「"」までの文字列を生成文とみなす。

含意文生成用プロンプト

```
Generate one sentence that logically entails "{premise}" in the form of a statement beginning with "Answer:". Answer: "
```

矛盾文生成用プロンプト

```
Generate one sentence that logically contradicts "{premise}" in the form of a statement beginning with "Answer:". Answer: "
```

few-shot 学習を採用する場合は、人手 NLI データセットからいくつかの文ペアを取り出し、上述のプロンプトの前に例として加え、含意文、矛盾文を生成する。例えば、1-shot の前提文、含意文のペアとして、“Fun for adults and children.”、“Fun for both adults and children.” を使用する場合のプロンプトは以下のようになる。

含意文生成用プロンプト (1-shot)

```
Generate one sentence that logically entails "Fun for adults and children." in the form of a statement beginning with "Answer:". Answer: "Fun for both adults and children."
Generate one sentence that logically entails "premise" in the form of a statement beginning with "Answer:". Answer: "
```

4 実験

自動生成した NLI データセット、および、それを用いて生成した文埋め込みの評価を行った。

4.1 NLI データセットの評価

評価方法 DeBERTa V2 XXLarge モデル [13] に MNLI タスクを学習させた deberta-v2-xxlarge-mnli[1] を用いて、自動生成した NLI データセットの質を評価した。具体的には、自動生成した NLI データセット中の各事例について、deberta-v2-xxlarge-mnli を用いた含意、中立、矛盾の三値分類を行い、分類結果と NLI データセットに付与されたラベルの一致度を算出することでデータセットの質を評価した。

実験設定 0-shot および人手 NLI データセットについては、含意および矛盾の文ペアをそれぞれランダムに 3000 ペアずつ、計 6000 ペア取り出し、それらの文ペアに対し自動分類結果とのラベルの一致度を算出した。few-shot 学習では、含意 10 個、矛盾 10 個の計 20 個の異なるプロンプトに対し、それぞれ 1000 ペアずつ、計 20000 個の文ペアを生成し、それらの文ペアに対し自動分類結果とのラベルの一致度を算出した。また、LLM には llama-2-7b-chat [14] を用いた。生成元となる前提文の集合として、SimCSE [5] の教師なし学習で用いられた、Wikipedia からランダムに抽出された 100 万文を用いた。その際、人手 NLI データセットの分布に近くなるようにトークン数が 4 以上 32 以下の文のみ使用した。few-shot 学習は 5-shot、10-shot、20-shot の 3 パターンで実施した。

1) https://huggingface.co/microsoft/deberta-v2-xxlarge-mnli

表 1 NLI データセットのラベルと自動分類の一致度

データセット	含意	矛盾
自動生成 (0-shot)	0.160	0.888
自動生成 (5-shot)	0.867	0.930
自動生成 (10-shot)	0.909	0.935
自動生成 (20-shot)	0.933	0.940
人手 NLI データセット	0.929	0.941

実験結果 一致度の算出結果を表 1 に示す。shot 数が増えるほど一致度が向上し、20-shot の場合、人手 NLI データセットと同等の一致度となった。特に含意ペアについては、few-shot 学習により大幅に一致度が向上しており、few-shot 学習の効果は非常に大きいと言える。以上の結果から、10-shot や 20-shot を用いて自動生成した NLI データセットはある程度、高品質なものとなっている可能性が高いと考えられる。付録 A に 0-shot、20-shot の設定で自動生成された NLI データの例を示す。

4.2 文埋め込みの評価

評価方法 NLI データセットを利用して生成された文埋め込みの評価を行った。文埋め込みの評価は主に Semantic Textual Similarity (STS) タスクで行った[2]。STS タスクは、文ペアが与えられた時に、モデルを用いて文ペアの意味的な類似度を計算し、それがどの程度人間による評価に近いかを検証することで、モデルが文の意味的類似性を正しく推定できるかを評価するタスクである。本研究では多くの先行研究と同様に、文ペアの文埋め込みの余弦類似度と人手評価による類似度とのスピアマンの順位相関係数により文埋め込みの質を評価した。

実験設定 STS データセットとして、先行研究と同様 STS 2012–2016 [15, 16, 17, 18, 19]、STS-B [20]、SICK-R [21] の 7 つのデータセットを用いた。LLM には llama-2-7b [14] を用い、64,000 事例の NLI データセットを用いて fine-tuning を行った。学習時の `batch_size` は 256、`warm_up` に用いたステップ数は全体の 10%、学習率は 5e-4 とした。学習の際は、5 ステップごとに STS-B の開発セットに対するスピアマンの順位相関係数を計算し、最もスコアが高い時点のモデルを最終的な評価に用いた。

2) SentEval の後段タスクを用いた評価も実施したが、先行研究 [1] で報告されている通り、NLI データセットを用いて fine-tuning を行うことの有効性は確認できなかった。後段タスクを用いた評価実験の結果は付録 B に記載する。

表 2 文埋め込みの余弦類似度と人手評価とのスピアマンの順位相関係数 (表内の値は全て 100 をかけたもの)

	STS12	STS13	STS14	STS15	STS16	STS-B	SICK-R	Avg.
Without fine-tuning								
PromptEOL-Llama-2-7b	59.91	78.86	68.74	75.71	73.39	73.48	71.26	71.62
Fine-tuning on unsupervised datasets								
unsupervised-SimCSE-RoBERTa	72.86	83.99	75.62	84.77	81.80	81.98	71.26	78.90
PromptRoBERTa	73.94	84.74	77.28	84.99	81.74	81.88	69.50	79.15
Fine-tuning on automatically generated datasets with n-shot								
PromptEOL+CSE-Llama-2-7b (n=0)	69.88	85.80	78.08	81.18	81.61	82.01	72.13	78.67
PromptEOL+CSE-Llama-2-7b (n=5)	73.25	87.60	81.58	85.45	83.67	84.52	74.96	81.58
PromptEOL+CSE-Llama-2-7b (n=10)	74.32	87.66	81.88	85.79	84.04	85.47	76.38	82.21
PromptEOL+CSE-Llama-2-7b (n=20)	74.12	87.74	82.14	85.25	83.99	85.51	76.05	82.11
Fine-tuning on manually annotated datasets								
supervised-SimCSE-RoBERTa	77.46	87.27	82.36	86.66	83.93	86.70	81.95	83.76
PromptEOL+CSE-Llama-2-7b	78.21	89.63	84.74	88.70	85.80	88.49	82.29	85.41

乱数による実験結果への影響を低減するため、0-shot、および、人手データセットを用いた実験では、NLI データセットの並び順を変えて 3 回実験し、それらの平均スコアを最終的なスコアとした。few-shot 学習では、異なる例文を用いて生成した 10 個の NLI データセットに対しそれぞれ実験し、それらの平均スコアを最終的なスコアとした。

また、比較のため、fine-tuning を行わない llama-2-7b ベースの PromptEOL を用いた評価も行った。さらに NLI データセットを使用しない unsupervised-SimCSE-RoBERTa、人手 NLI データセットで学習した supervised-SimCSERoBERTa、教師なし学習において最高性能を達成している PromptRoBERTa についても先行研究のスコアを引用し比較を行った。

実験結果 STS タスクの実験結果を表 2 に示す．0-shot で fine-tuning した PromptEOL モデルと fine-tuning を行わない PromptEOL モデルのスコアを比較すると、平均スコアがおよそ 7 ポイントほど上昇しており、生成した NLI データセットを用いた fine-tuning が有効であることが確認できる。さらに、0-shot と few-shot の結果を比較すると、few-shot 学習により生成した NLI データセットを用いたモデルの方が 3 ポイント程度高くなっており、few-shot 学習の有効性が確認できる。特に 10-shot に対する平均スコアは 82.21 と、unsupervised-SimCSE-RoBERTa や PromptRoBERTa の性能を上回っており、人手で構築された大規模なデータセットを利用しない設定において最高性能を達成した。

教師あり学習と比較すると、教師あり学習した PromptEOL の平均スコアは 85.41 と、10-shot に対する平均スコア 82.21 より、3.2 ポイント高いスコアであった。fine-tuning を行わない場合の平均スコアは 71.62 であるので、few-shot 学習により自動生成した NLI データセットの効果は非常に大きいと言える。

5 おわりに

本研究ではデコーダ系の大規模言語モデルを用いて NLI データセットを自動生成し、PromptEOL の fine-tuning に利用することで、教師なし設定における文埋め込み生成の高性能化に取り組んだ。STS タスクを用いた評価実験の結果、few-shot 学習により自動生成した NLI データセットを用いることで、人手で構築された大規模なデータセットを利用しない設定において、既存手法を大きく上回る性能を達成した。今後の課題としては以下の 2 つが挙げられる。まず、提案した枠組みは人手で構築された大規模なデータセットを必要としないことから多くの言語に応用可能であると考えられるが、本研究での実験は英語のみを対象としている。手法の言語横断的な有用性を示すためには、英語以外の言語も対象とした実験を行う必要がある。また、後段タスクを対象とした実験では、そもそも NLI データセットを用いた fine-tuning が有効でないという結果となったが、提案した枠組みは NLI データ以外にも応用可能であることから、後段タスクごとに適したデータセットを検証し、各タスクに適した文埋め込みの生成を目指すことが考えられる。

謝辞

本研究は JSPS 科研費 JP21H04901 の助成を受けたものです。

参考文献

[1] Ting Jiang, Shaohan Huang, Zhongzhi Luan, Deqing Wang, and Fuzhen Zhuang. Scaling Sentence Embeddings with Large Language Models. **arXiv:2307.16645**, 2023.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In **Proceedings of NACCL'19**, pp. 4171–4186, 2019.

[3] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In **Proceedings of EMNLP-IJCNLP'19**, pp. 3982–3992, 2019.

[4] Hayato Tsukagoshi, Ryohei Sasano, and Koichi Takeda. DefSent: Sentence Embeddings using Definition Sentences. **Proceedings of ACL'21**, pp. 411–418, 2021.

[5] Tianyu Gao, Xingcheng Yao, and Danqi Chen. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In **Proceedings of EMNLP'21**, pp. 6894–6910, 2021.

[6] Ting Jiang, Jian Jiao, Shaohan Huang, Zihan Zhang, Deqing Wang, Fuzhen Zhuang, Furu Wei, Haizhen Huang, Denvy Deng, and Qi Zhang. PromptBERT: Improving BERT Sentence Embeddings with Prompts. In **Proceedings of EMNLP'22**, pp. 8826–8837, 2022.

[7] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. **arXiv:1910.10683**, 2019.

[8] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Sentence-T5: Scalable Sentence Encoders from Pre-trained Text-to-Text Models. In **Findings of ACL'22**, pp. 1864–1874, 2022.

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

[10] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In **Proceedings of EMNLP'15**, pp. 632–642, 2015.

[11] Adina Williams, Nikita Nangia, and Samuel Bowman. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In **Proceedings of NACCL'18**, pp. 1112–1122, 2018.

[12] Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. XNLI: Evaluating Cross-lingual Sentence Representations. In **Proceedings of EMNLP'18**, pp. 2475–2485, 2018.

[13] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. In **Proceedings of ICRL'21**, 2021.

[14] Thomas Scialom Hugo Touvron and et al. Llama 2: Open Foundation and Fine-Tuned Chat Models. **arXiv:2307.09288**, 2023.

[15] Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. SemEval-2012 Task 6: A Pilot on Semantic Textual Similarity. In **Proceedings of SemEval'12**, pp. 385–393, 2012.

[16] Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. *SEM 2013 shared task: Semantic Textual Similarity. In **Proceedings of *SEM'13**, pp. 32–43, 2013.

[17] Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. SemEval-2014 Task 10: Multilingual Semantic Textual Similarity. In **Proceedings of SemEval'14**, pp. 81–91, 2014.

[18] Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Iñigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, German Rigau, Larraitz Uria, and Janyce Wiebe. SemEval-2015 Task 2: Semantic Textual Similarity, English, Spanish and Pilot on Interpretability. In **Proceedings of SemEval'15**, pp. 252–263, 2015.

[19] Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. SemEval-2016 Task 1: Semantic Textual Similarity, Monolingual and Cross-Lingual Evaluation. In **Proceedings of SemEval'16**, pp. 497–511, 2016.

[20] Eneko Agirre Iñigo Lopez-Gazpio Lucia Specia Daniel Cer, Mona Diab. SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation. In **Proceedings of SemEval'17**, pp. 1–14, 2017.

[21] 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 LREC'14**, pp. 216–223, 2014.

表 3　0-shot 学習で自動生成された NLI データの例

前提文	自動生成された含意/矛盾文	自動推定
It is a colorless liquid.	含意文：It has no color.	**含意**
	矛盾文：It is a colored liquid.	**矛盾**
He spent several months in prison.	含意文：He was convicted of a crime and sentenced to several months in prison.	中立
	矛盾文：He was released from prison yesterday.	中立
Her last public performance was in 1954.	含意文：She retired from the music industry after that year.	中立
	矛盾文：She is still actively touring and performing today.	**矛盾**
This later became simply the Amps.	含意文：The band's early name, 'The Amps,' was later simplified to just 'The Amps.'	中立
	矛盾文：The Amps never became simply this.	**矛盾**

表 4　20-shot 学習で自動生成された NLI データの例

前提文	自動生成された含意/矛盾文	自動推定
It is a colorless liquid.	含意文：Liquid is colorless.	**含意**
	矛盾文：It is a solid.	**矛盾**
He spent several months in prison.	含意文：He was in prison.	**含意**
	矛盾文：He was never in prison.	**矛盾**
Her last public performance was in 1954.	含意文：She performed in 1954.	**含意**
	矛盾文：She has been in hiding for the past 50 years.	中立
This later became simply the Amps.	含意文：A group of musicians formed a band.	中立
	矛盾文：This later became the Bamps.	**矛盾**

表 5　SentEval の後段タスクにおける正解率 (%)

	MR	CR	SUBJ	MPQA	SST	TREC	MRPC	Avg.
Without fine-tuning								
PromptEOL-Llama-2-7b	90.53	92.45	96.22	91.24	95.39	96.20	74.96	91.00
Fine-tuning on unsupervised datasets								
unsupervised-SimCSE-RoBERTa-large	82.74	87.87	93.66	88.22	88.58	92.00	69.68	86.11
PromptRoBERTa	83.82	88.72	93.19	90.36	88.08	90.60	76.75	87.36
Fine-tuning on automatically generated datasets with n-shot								
PromptEOL+CSE-Llama-2-7b (n=0)	89.92	92.45	95.47	90.32	93.83	94.00	72.40	89.77
PromptEOL+CSE-Llama-2-7b (n=5)	89.44	92.69	94.89	90.74	93.65	94.36	72.35	89.73
PromptEOL+CSE-Llama-2-7b (n=10)	89.21	92.80	94.80	90.87	93.42	93.42	72.69	89.60
PromptEOL+CSE-Llama-2-7b (n=20)	89.07	92.83	94.73	90.77	93.05	94.36	73.76	89.80
Fine-tuning on manually annotated datasets								
supervised-SimCSE-RoBERTa-large	77.46	87.27	82.36	86.66	83.93	86.70	81.95	83.76
PromptEOL+CSE-Llama-2-7b	89.77	93.28	95.99	90.66	95.20	96.20	74.86	90.85

A　自動生成された NLI データの例

0-shot 学習、20-shot 学習で自動生成された NLI データの例を表 3、表 4 にそれぞれ示す。表 3、表 4 で前提文は同一である。

0-shot 学習により生成された含意文の多くは中立と推定されているものの、含意はしていなくとも類似した意味の文を生成できていることがわかる。一方、矛盾文については never といった否定語や colored という colorless とは逆の意味の単語を用いることで、多様な文が生成できていることがわかる。20-shot 学習により生成された文は前提文を強く参照している傾向があり、精度の高い含意文、矛盾文を生成できていることが確認できる。また、0-shot 学習と比べ、生成される文は短くなる傾向があった。

B　後段タスクを用いた評価

生成された文埋め込みの、後段タスクにおける有用性を検証するため、SentEval の後段タスクを用いた評価を実施した。SentEval の後段タスクでは文埋め込みを入力として使用し、その上で線形分類器の訓練を行う。具体的には、各文から生成された文埋め込みを特徴として使用し、ロジスティック回帰などの線形分類器を訓練する。訓練された分類器を用いて分類タスクのパフォーマンスを評価し、正解率や他の関連メトリックスを測定することで、文表現の効果性を定量的に評価する。

結果を表 5 に示す。先行研究 [1] で報告されているのと同様に、fine-tuning を行うことによる後段タスクの性能向上は確認できなかったが、0-shot、few-shot ともに人手 NLI データセットで学習した場合と同等のスコアとなった。