

[S1-P27]

商品検索のための追加事前学習としての言い換えに基づく対照学習

杉山 誠治, 近藤 里咲, 梶原 智之, 二宮 崇 (愛媛大学)

1. 背景：マスク言語モデリングに基づく事前学習済みモデルのファインチューニング

事前学習済みモデルをファインチューニングして、タスクに特化したモデルを作成

マスク言語モデリングの事前学習では入力文の一部をマスクし、マスクされたトークンを予測^[1]

[1] Devlin et al. (2019) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Qクエリ
ヘッドホン

商品A

商品B

商品C

→ マスク言語モデル →

埋め込み計算

商品A

クエリ

商品B

商品C

→ ランキング

1位

2位

3位

2. 課題：事前学習では、類似する文の埋め込みを必ずしも近づけていない^[2]

✓ 類似する文同士を近づけることで、文対モデリングタスクをさらに改善できる

✓ 自然言語推論（NLI）データセットを用いた対照学習が有効^[3-5]だが、英語以外の言語には大規模なNLIデータセットは存在しない

→ **NLIデータセットを用いずに、類似する文の埋め込みを近づけたい**

●: クエリと類似する文 ●: クエリと類似していない文

ヘッドホン

夏におすすめな帽子

2024の最新イヤホン

近づける

ヘッドホン

夏におすすめな帽子

2024の最新イヤホン

[2] Li et al. (2020) On the Sentence Embeddings from Pre-trained Language Models

[3] Gao et al. (2021) SimCSE: Simple Contrastive Learning of Sentence Embeddings

[4] Jiang et al. (2022) PromptBERT: Improving BERT Sentence Embeddings with Prompts

[5] Chuang et al. (2022) DiffCSE: Difference-based Contrastive Learning for Sentence Embeddings

3. 提案手法：言い換え辞書を用いて正例ペアを作成し、事前学習済みモデルを対照学習

生コーパスから多言語に対応する言い換え辞書^[6]を用いて言い換え文を作成 → 正例ペアとして最も相応しい文を選択

事前学習済みモデルに追加事前学習として**対照学習**を適用 → 類似する文の埋め込みを近づける

事前学習 → 類似する文を近づける → 追加事前学習 → タスクに特化させる → ファインチューニング

言い換え元の文
ミカンの味がする。

言い換え辞書

言い換え元	言い換え先	言い換え確率
ミカンの	の 柑橘類 の	0.07
ミカンの	柑橘 系	0.10
ミカンの	柑橘 系 の	0.12
ミカンの	柑橘類 の	0.63

言い換え文

の 柑橘類 の 味がする。
柑橘 系 味がする。
柑橘 系 の 味がする。
柑橘類 の 味がする。

Perplexity

55.0
71.1
38.8
50.7

近づける

言い換え元の文

選択した言い換え文
柑橘系の味がする。

無関係な文
鳥が鳴いている。

無関係な文
魚が泳いでいる。

無関係な文
子供が走っている。

文の流暢性を測るPerplexityを計算

Perplexityの値が最も低い1文を正例ペアとする

[6] Ganitkevitch and Callison-Burch (2014) The Multilingual Paraphrase Database

4. 評価実験：商品検索タスクをはじめとした文対モデリングタスクで性能が改善

4.1 実験設定

Wiki40Bコーパスから言い換え元の文を抽出

言い換え辞書：EhiMerPPDB^[7]（日），PPDB 2.0^[8]（英）

• 言い換え確率が {0.1, 0.2, 0.3, 0.4, 0.5} 以上からなる

• 言い換え辞書を使用して、辞書サイズを変更する

• 対照学習のデータ数を {1万, 2万, 4万, 8万, 16万} にする

→ 検証用データで最適な組合せを選択し、評価用データで評価

モデル：mBERT^[9], 日本語 BERT^[10], 英語 BERT^[11]

タスク：商品検索（分類・リランキング）, 類似度推定, 含意関係認識, 言い換え認識

日本語データセット

	訓練用	検証用	評価用
Shopping Queries ^[12]	294,874	32,272	118,907
JSTS	11,205	1,246	1,457
JNLI	18,065	2,008	2,434
JSICK	4,500	500	4,927
PAWS_X	49,401	2,000	2,000

英語データセット

	訓練用	検証用	評価用
Shopping Queries ^[12]	1,254,438	138,625	425,762
STS-B	5,749	1,500	1,379
SNLI	550,152	10,000	10,000
SICK	4,439	495	4,906
PAWS	49,401	8,000	8,000

4.2 実験結果

タスク	商品検索（分類）	商品検索（リランキング）	類似度推定		含意関係認識		言い換え認識
評価指標	Micro-F1	nDCG	Pearson		Accuracy		Accuracy
日本語データセット	Shopping Queries		JSTS	JSICK	JNLI	JSICK	PAWS_X
mBERT	0.559	0.813	0.863	0.898	0.862	0.873	0.786
mBERT + 提案手法	0.586	0.832	0.870	0.902	0.875	0.875	0.795
日本語 BERT	0.590	0.827	0.916	0.919	0.893	0.891	0.798
日本語 BERT + 提案手法	0.594	0.841	0.916	0.922	0.890	0.890	0.822
英語データセット	Shopping Queries		STS-B	SICK	SNLI	SICK	PAWS
mBERT	0.651	0.845	0.847	0.871	0.873	0.850	0.913
mBERT + 提案手法	0.655	0.845	0.843	0.869	0.877	0.855	0.929
英語 BERT	0.654	0.844	0.838	0.870	0.888	0.862	0.914
英語 BERT + 提案手法	0.655	0.845	0.853	0.893	0.888	0.871	0.919

追加事前学習として類似する文の埋め込みを近づけることで、**ファインチューニングの性能を改善**

[7] <https://github.com/EhimeNLP/EhiMerPPDB>

[8] <http://paraphrase.org/#/download>

[9] <https://huggingface.co/google-bert/bert-base-multilingual-uncased>

[10] <https://huggingface.co/tohoku-nlp/bert-base-japanese-v3>

[11] <https://huggingface.co/google-bert/bert-base-uncased>

[12] <https://github.com/amazon-science/esci-data>

謝辞：本研究は 株式会社メルカリ R4D の支援を受けて実施した