# 定義文を用いた文埋め込み構成法

塚越 駿　　笹野 遼平　　武田 浩一
名古屋大学情報学部　　名古屋大学大学院情報学研究科
tsukagoshi.hayato@e.mbox.nagoya-u.ac.jp　　{sasano,takedasu}@i.nagoya-u.ac.jp

## 1 はじめに

近年，自然言語推論 (Natural Language Inference: NLI) データセットを用いて文埋め込みを構成する手法が成功を収めており，文類似度 (Semantic Textual Similarity: STS) タスクをはじめとする様々なタスクで活用されている [1, 2]．これらの手法では，NLI データセットの文ペアに付与されている「含意」「矛盾」「その他」のラベルを正しく分類するというタスクを通して文埋め込みを構成する．しかし，このような手法は，大規模な NLI データセットが整備されている言語でしか利用できないという問題がある．

本研究ではこの問題を解決するため，辞書に含まれる単語とその定義文が基本的に同一の意味内容を表すという関係に着目し，辞書の定義文を用いた文埋め込み構成法を提案する．辞書は NLI データセットと比べ，はるかに多くの言語において整備が行われており，定義文を用いた文埋め込み構成法は多くの言語に適用可能であると考えられる．

## 2 定義文を用いた文埋め込み構成法

本研究で提案する定義文による文埋め込み構成法は，NLI データセットを用いて文埋め込みを構成するモデルとして代表的な Sentece-BERT [1] と同様，BERT [3] や RoBERTa [4] などの事前学習済み言語モデルをベースとし，これらのモデルの出力から文埋め込みを構成する (図 1)．本節ではまず BERT と RoBERTa，および，Sentence-BERT を紹介し，続いて提案する文埋め込み構成法の説明を行う．

### 2.1 BERT と RoBERTa

BERT は複数層の Transformer [5] エンコーダで構成される事前学習済み言語モデルである．マスク穴埋め問題と次文予測によって，大規模なテキストデータで自己教師あり学習を行うことで言語知識を獲得し，文脈に従った単語埋め込みを出力する．マスク穴埋め問題は，入力文の中のトークンを一定

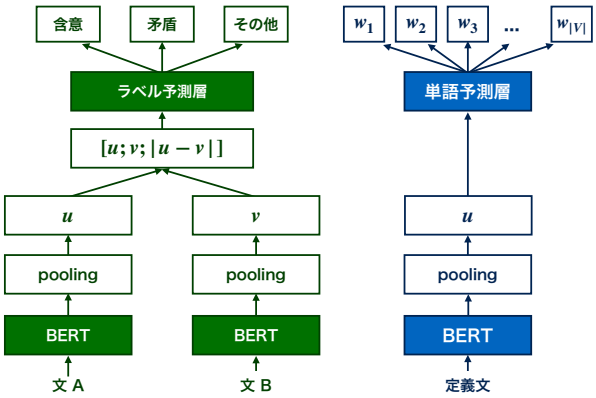


**図 1** Sentence-BERT (左) と提案手法 (右) の概要図

の割合で [MASK] という特殊トークンに置き換えてモデルへ入力し，[MASK] に対応する位置の最終層の埋め込みを用いて，置き換え前のトークンを予測するタスクである．次文予測は，文区切りトークン [SEP] で繋がれた 2 文を BERT に入力した際，文頭に付与された特殊トークン [CLS] の出力をもとにそれらが元のテキストデータで連続する 2 文であるかどうかを予測するタスクである．

一方，RoBERTa は BERT と同様の構造をしているが，BERT で事前学習時に行われていた次文予測を排除し，データサイズ，バッチサイズを大きくすることで BERT の改善を試みたモデルである．次項以降で説明する Sentence-BERT と提案手法はいずれもベースモデルとして BERT と RoBERTa が利用可能であるが，本稿におけるモデルの説明では基本的に BERT を用いた場合について説明する．

### 2.2 Sentence-BERT

NLI データセットを用いた文埋め込み構成法は Conneau ら [2] により提案された．Conneau らは，パラメータを共有した 2 つのモデルの学習を行う Siamese network 構造を用いて，類似した文が意味ベクトル空間上で近い位置に分布するように文埋め込みを構成する InferSent を提案した．Reimers ら [1] によって提案された Sentence-BERT は，InferSent と類

似した構造を採用しているが，文ベクトルの構成に fine-tuning した BERT を用いている．Sentence-BERT の概要を図 1 左に示す．Sentence-BERT では，まず文を BERT に入力し，その出力を pooling することで文ベクトルを構成する．pooling 手法には以下の 3 種類が用いられる．

**CLS** BERT の事前学習時に次文予測で用いられる [CLS] トークンの埋め込みを用いる．RoBERTa を用いる場合は [CLS] が存在しないため，代替として文頭トークン<s>の埋め込みを用いる．

**Mean** 文を構成するすべての語の文脈化単語埋め込みの平均を用いる．

**Max** 文を構成するすべての語の文脈化単語埋め込みの次元ごとの最大値を用いる．

pooling により得られる文ペアのそれぞれの文埋め込みを $\boldsymbol{u}, \boldsymbol{v}$ とする．それらを組み合わせたベクトル $[\boldsymbol{u}; \boldsymbol{v}; |\boldsymbol{u} - \boldsymbol{v}|]$ をクラス数と同じ数の出力次元を持つラベル予測層に入力し，文ペアに付与されている「含意」などのラベルを正しく予測できるように fine-tuning を行う．fine-tuning には，SNLI データセット [6]，および，Multi-Genre NLI データセット [7] を合わせた約 100 万文を用いる．

### 2.3 定義文を用いた文埋め込み

本研究では，単語とその定義文が同一の意味内容を表すことに着目し，定義文から単語予測を行うことで文埋め込みを構成する．提案手法の概要を図 1 右に示す．ここで，ある定義文 $X_k$ に対応する単語を $w_k$ と表し，BERT の事前学習時にマスク穴埋め問題で用いられる [MASK] の埋め込みから元のトークンを予測する層を単語予測層と呼ぶ．

まず，$X_k$ を BERT に入力し，出力を pooling することで $X_k$ の文埋め込み $\boldsymbol{u}_k$ を構成する．pooling 手法には Sentence-BERT と同様に CLS，Mean，Max の 3 種類を用いる．次に，構成した文埋め込み $\boldsymbol{u}_k$ を単語予測層に入力し，$X_k$ を入力としたときの $w_k$ の出現確率 $P(w_k|X_k)$ を得る．損失関数に交差エントロピー誤差を用いて，$P(w_k|X_k)$ を最大化するように fine-tuning を行う．

この際，単語予測層には事前学習時のパラメータを固定して用いる．これにより，追加の分類器などを新たに学習せず fine-tuning を行うことができる．また，事前学習時の単語予測層をそのまま使っているため，提案手法で得られる文埋め込みは，その文が表す意味内容に近い意味で使用されている単語が存在した場合，その文脈化単語埋め込みと類似するという性質が期待できる．

**表 1** データセットの統計値

| 全体 | 単語数 | 定義文数 | 平均文長 |
|---|---|---|---|
| 訓練データ | 29,413 | 97,759 | 9.921 |
| 開発データ | 3,677 | 12,127 | 9.874 |
| テストデータ | 3,677 | 12,433 | 9.846 |
| **BERT 語彙内** | **単語数** | **定義文数** | **平均文長** |
| 訓練データ | 7,732 | 54,142 | 9.531 |
| 開発データ | 936 | 6,544 | 9.512 |
| テストデータ | 979 | 6,930 | 9.551 |
| **RoBERTa 語彙内** | **単語数** | **定義文数** | **平均文長** |
| 訓練データ | 7,269 | 53,935 | 9.376 |
| 開発データ | 901 | 6,625 | 9.372 |
| テストデータ | 925 | 6,945 | 9.410 |

## 3 単語予測実験

提案手法により構成した文の埋め込みが，どの程度，文の意味を埋め込めているか評価するため，定義文の埋め込みを用いた単語予測実験を行った．

### 3.1 使用するデータセット

提案手法は単語と定義文のペアを必要とする．本研究では，石渡ら [8] が公開しているデータセットの中から，Oxford Dictionary の単語と定義文を利用した．各エントリは単語と定義文のペアからなり，一つの単語が複数の定義文を持ち得る．データは単語ごとに訓練データ/開発データ/テストデータに 8:1:1 の割合で分割した．また，提案手法は単語予測層に BERT または RoBERTa の事前学習時の層を用いるため，各モデルの語彙に含まれない単語に関する予測確率を単語予測層から得ることはできない．したがって，本実験ではデータセットの中から BERT，RoBERTa それぞれの語彙に含まれる単語とその定義文のみを用いる．利用したデータセットの統計値を表 1 に示す．

### 3.2 実験設定

事前学習済みモデルとして，Hugging Face[1] が公開しているライブラリである Transformers[2] から，BERT-base (bert-base-uncased)，BERT-large (bert-large-uncased)，RoBERTa-base (roberta-base)，RoBERTa-large (roberta-large) を利用した．fine-tuning の設定として，バッチサイズは 16，エポック数は 1，最適化手法に Adam を用いた．また，学習の開始時点では学

1) https://huggingface.co
2) https://github.com/huggingface/transformers

表 2　単語予測の性能

| Model | pooling | MRR | Top1 | Top3 | Top10 |
|---|---|---|---|---|---|
| BERT-base | CLS | .0009 | .0000 | .0000 | .0000 |
| (fine-tuning なし) | Mean | .0132 | .0001 | .0043 | .0242 |
| | Max | .0327 | .0157 | .0320 | .0626 |
| BERT-base | CLS | .3200 | .2079 | .3670 | .5418 |
| | Mean | .3091 | .1972 | .3524 | .5356 |
| | Max | .2939 | .1840 | .3350 | .5207 |
| BERT-large | CLS | .3587 | .2388 | .4139 | .6011 |
| | Mean | .3286 | .2091 | .3792 | .5723 |
| | Max | .2925 | .1814 | .3356 | .5194 |
| RoBERTa-base | CLS | .3436 | .2241 | .3983 | .5836 |
| | Mean | .3365 | .2170 | .3906 | .5783 |
| | Max | .3072 | .1941 | .3523 | .5386 |
| RoBERTa-large | CLS | .3863 | .2611 | .4460 | .6364 |
| | Mean | **.3995** | **.2699** | **.4634** | **.6599** |
| | Max | .3175 | .2015 | .3646 | .5543 |

習率を 0 とし，全学習ステップのうち 10%で，設定した値まで線形に学習率を増加させる warm-up を用いた．学習率は各モデル，pooling 手法ごとに $2^x \times 10^{-6}$, $x \in \{0, 0.5, 1, ..., 7\}$ の範囲で探索し，開発データでの平均逆順位 (Mean Reciprocal Rank; MRR) が最も高くなった学習率を使用した．異なるシード値で 10 回実験を行い，その平均を評価スコアとした．定義文を入力した際に出力される単語の予測確率から MRR と，1，3，10 位以内に正解が含まれる割合 (Top-$k$ accuracy) を算出し，評価に用いた．また，比較対象として BERT-base モデルを fine-tuning せずに使った場合の性能も算出した．

### 3.3　実験結果

実験の結果を表 2 に示す．fine-tuning しなかった場合の性能は pooling 手法として Max を選んだ場合が最も高かったものの，その Top-1 accuracy は 0.0157 と極めて低い値であり，高い性能を得るためには fine-tuning を行うことが必須であることが確認できる．提案手法では，base を用いたモデルより large を用いたモデル，BERT を用いたモデルより RoBERTa を用いたモデルの方が性能が高く，RoBERTa-large と Mean の組み合わせが最も高い性能を示した．また，RoBERTa-large 以外のモデルは，pooling 手法に CLS を用いたモデルが最も高い性能を示した．

## 4　応用タスクにおける性能評価

生成された文埋め込みの一般的な有用性を評価するため Semantic Textual Similarity (STS) タスク，および，文埋め込み評価のためのツールキットである SentEval を用いた評価を行った．

### 4.1　実験設定

本節では，Reimers [1] らにより報告されている各タスクの性能と，提案手法の性能との比較を行う．評価は，モデルごとに 3 節の実験で開発データの MRR が最も高かった pooling 手法を用いて行った[3)]．既存手法の性能は Reimers ら [1] の結果を引用した．

### 4.2　Semantic Textual Similarity タスク

Semantic Textual Similarity (STS) タスクは，文ペアが与えられた時に，その文ペアの意味的な類似度を推定するタスクである．教師なし設定の場合，STS データセットを用いた学習は行わず，事前に学習したモデルを用いて文ペアをそれぞれ文埋め込みに変換し，それらを用いて算出された文ペアの類似度と，人手評価との順位相関係数により評価する．本研究では，提案手法が一般の文に対して妥当な文埋め込みを構成できているか評価するため，STS12-16 [11–15]，STS benchmark [16]，SICK-Relatedness [17] の各データセットを用いた教師なし STS タスクにより評価を行った．これらのデータセットには文ペアとその類似度が含まれており，類似度は人手評価によって付与された 0 から 5 の実数である．文埋め込みの類似度には余弦類似度を用い，データセットに含まれる人手評価とのスピアマンの順位相関係数を算出した．異なるシード値で 10 回実験を行い，その平均を評価スコアとした．

実験の結果を表 3 に示す．提案手法で用いる学習データは，Sentence-BERT の学習に使用されているデータの 5% 程度の規模であるにも関わらず，提案手法の BERT-base，RoBERTa-base モデルは Sentence-BERT-base，Sentence-RoBERTa-base と遜色のない性能を示すことが確認できる．特に，STS12-16 の集約版である STS benchmark において，提案手法の RoBERTa モデルは高い性能を示した．

### 4.3　SentEval

SentEval [18] は，感情分類などを含む様々なタスクを集約したツールキットである．文埋め込みを入力とする分類器の学習を行い，その性能から文埋め込みがどのような情報を捉えているかを評価する．本研究では，Reimers ら [1] と同一のタスクについて実験を行った．各タスクの概要を表 4 に示す．評価には，前節の設定で fine-tuning を行ったモデルを使

3)　STS タスク，SentEval のそれぞれについて，全てのモデルと pooling 手法ごとの平均と標準偏差を付録に記載する．

**表 3**　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). STS-B は STS benchmark を，SICK-R は SICK-Relatedness を表す．各タスクごとに最良の結果を**太字**で示す.

| Model | STS12 | STS13 | STS14 | STS15 | STS16 | STS-B | SICK-R | Avg. |
|---|---|---|---|---|---|---|---|---|
| Avg. GloVe embeddings [9] | 55.14 | 70.66 | 59.73 | 68.25 | 63.66 | 58.02 | 53.76 | 61.32 |
| Avg. BERT embeddings | 38.78 | 57.98 | 57.98 | 63.15 | 61.06 | 46.35 | 58.40 | 54.81 |
| BERT CLS-vector | 20.16 | 30.01 | 20.09 | 36.88 | 38.08 | 16.50 | 42.63 | 29.19 |
| InferSent - Glove [2] | 52.86 | 66.75 | 62.15 | 72.77 | 66.87 | 68.03 | 65.65 | 65.01 |
| Universal Sentence Encoder [10] | 64.49 | 67.80 | 64.61 | 76.83 | 73.18 | 74.92 | **76.69** | 71.22 |
| Sentence-BERT-base (Mean) | 70.97 | 76.53 | 73.19 | 79.09 | 74.30 | 77.03 | 72.91 | 74.89 |
| Sentence-BERT-large (Mean) | 72.27 | 78.46 | **74.90** | **80.99** | 76.25 | **79.23** | 73.75 | 76.55 |
| Sentence-RoBERTa-base (Mean) | 71.54 | 72.49 | 70.80 | 78.74 | 73.69 | 77.77 | 74.46 | 74.21 |
| Sentence-RoBERTa-large (Mean) | **74.53** | 77.00 | 73.18 | 81.85 | 76.82 | 79.10 | 74.29 | **76.68** |
| 提案手法 (BERT-base, CLS) | 67.56 | 79.86 | 69.52 | 76.83 | 76.61 | 75.57 | 73.05 | 74.14 |
| 提案手法 (BERT-large, CLS) | 66.22 | **82.07** | 71.48 | 79.34 | 75.38 | 73.46 | 74.30 | 74.61 |
| 提案手法 (RoBERTa-base, CLS) | 65.55 | 80.84 | 71.87 | 78.77 | **79.29** | 78.13 | 74.92 | 75.62 |
| 提案手法 (RoBERTa-large, Mean) | 58.36 | 76.24 | 69.55 | 73.15 | 76.90 | 78.53 | 73.81 | 72.36 |

**表 5**　SentEval の各タスクにおける正解率 (%)．各タスクごとに最良の結果を**太字**で示す.

| Model | MR | CR | SUBJ | MPQA | SST-2 | TREC | MRPC | Avg. |
|---|---|---|---|---|---|---|---|---|
| Avg. GloVe embeddings | 77.25 | 78.30 | 91.17 | 87.85 | 80.18 | 83.00 | 72.87 | 81.52 |
| Avg. BERT embeddings | 78.66 | 86.25 | 94.37 | 88.66 | 84.40 | 92.80 | 69.45 | 84.94 |
| BERT CLS-vector | 78.68 | 84.85 | 94.21 | 88.23 | 84.13 | 91.40 | 71.13 | 84.66 |
| InferSent - GloVe | 81.57 | 86.54 | 92.50 | 90.38 | 84.18 | 88.20 | 75.77 | 85.59 |
| Universal Sentence Encoder | 80.09 | 85.19 | 93.98 | 86.70 | 86.38 | **93.20** | 70.14 | 85.10 |
| Sentence-BERT-base (Mean) | 83.64 | 89.43 | 94.39 | 89.86 | 88.96 | 89.60 | **76.00** | 87.41 |
| Sentence-BERT-large (Mean) | 84.88 | 90.07 | 94.52 | 90.33 | 90.66 | 87.40 | 75.94 | 87.69 |
| 提案手法 (BERT-base, CLS) | 80.94 | 87.57 | 94.59 | 89.98 | 85.78 | 89.73 | 73.82 | 86.06 |
| 提案手法 (BERT-large, CLS) | 85.79 | 90.54 | **95.58** | 90.15 | **91.17** | 90.47 | 73.74 | 88.20 |
| 提案手法 (RoBERTa-base, CLS) | 83.94 | 90.44 | 94.05 | 90.70 | 89.16 | 90.80 | 75.52 | 87.80 |
| 提案手法 (RoBERTa-large, Mean) | **86.47** | **91.53** | 95.02 | **91.15** | 90.77 | 92.33 | 73.91 | **88.74** |

**表 4**　SentEval の各タスクの説明

| タスク名 | 分類対象 | クラス数 |
|---|---|---|
| MR | 映画レビューの感情 | 2 |
| CR | 商品レビューの感情 | 2 |
| SUBJ | 映画/あらすじの主観性 | 2 |
| MPQA | 語句の極性 | 2 |
| SST-2 | 映画レビューの感情 | 2 |
| TREC | 質問の種別 | 6 |
| MRPC | 言い換えかどうか | 2 |

用した．SentEval 内の各タスクについて，提案手法により生成した文埋め込みを入力とする分類器を学習し，性能を評価した．文埋め込みを入力とする分類器の学習には Reimers ら [1] と同じ設定を用い，10 分割交差検証を行った．提案手法によるモデルの fine-tuning と性能評価は異なるシード値で 3 回行い，その平均を評価スコアとした．

実験の結果を表 5 に示す[4)]．提案手法の中では RoBERTa-large が最も高い性能を示している．また，base から large へモデルサイズを増加させると一貫して性能が向上した．提案手法の BERT-large，RoBERTa-base，RoBERTa-large における性能は Sentence-BERT-large を上回っていることから，提案手法による文埋め込みが様々なタスクに応用できる有用な情報を埋め込んでいることが確認できる．

4) Reimers ら [1] により Sentence-RoBERTa の性能は Sentence-BERT と同等と報告されているため記載を省略した．

## 5　おわりに

本研究では辞書の定義文を用いた文埋め込み構成法を提案した．また，単語予測実験，および，STS タスク，SentEval を用いた実験を通して，提案手法の有効性，および，大規模な NLI データセットを用いる既存手法と同等の性能を発揮することを示した．提案手法は多くの言語において整備されている辞書に基づく手法であり，他の言語に応用する場合であっても新規の言語資源を作成する必要がないという特長がある．また，得られる文埋め込みは，その文が表す意味内容に近い意味で使用されている単語が存在した場合，その文脈化単語埋め込みと類似するという性質を持つことが期待できる．

今後は，実際に英語以外の言語に応用した場合の性能評価や，文書分類タスクなどより広範な下流タスクへの応用の評価を行いたい．さらに，提案手法による定義文の埋め込みと文脈化単語埋め込みの意味ベクトル空間上での関係の解析を行いたい．

# 参考文献


[1] 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.

[2] Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 670–680, 2017.

[3] 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.

[4] Y. Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, M. Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692*, 2019.

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems (NIPS)*, pp. 5998–6008, 2017.

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

[7] Adina Williams, Nikita Nangia, and Samuel Bowman. 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)*, pp. 1112–1122, 2018.

[8] Shonosuke Ishiwatari, Hiroaki Hayashi, Naoki Yoshinaga, Graham Neubig, Shoetsu Sato, Masashi Toyoda, and Masaru Kitsuregawa. Learning to describe unknown phrases with local and global contexts. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pp. 3467–3476, 2019.

[9] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, 2014.

[10] Daniel Matthew Cer, Yinfei Yang, Sheng yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, C. Tar, Yun-Hsuan Sung, B. Strope, and R. Kurzweil. Universal Sentence Encoder. *arXiv:1803.11175*, 2018.

[11] Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. SemEval-2012 task 6: A pilot on semantic textual similarity. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Semantic Evaluation (SemEval)*, pp. 385–393, 2012.

[12] Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. *SEM 2013 shared task: Semantic textual similarity. In *Second Joint Conference on Lexical and Computational Semantics (*SEM)*, pp. 32–43, 2013.

[13] 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 the 8th International Workshop on Semantic Evaluation (SemEval)*, pp. 81–91, 2014.

[14] 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 the 9th International Workshop on Semantic Evaluation (SemEval)*, pp. 252–263, 2015.

[15] 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 the 10th International Workshop on Semantic Evaluation (SemEval)*, pp. 497–511, 2016.

[16] Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval)*, pp. 1–14, 2017.

[17] 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.

[18] Alexis Conneau and Douwe Kiela. SentEval: An evaluation toolkit for universal sentence representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC)*, pp. 1699–1704, 2018.

# A 付録

**表 5** 各モデル，pooling 手法ごとの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．表内の数値は，異なるシード値で 10 回評価を行った際の平均と標準偏差に 100 をかけたものである．

| Model | pooling | STS12 | STS13 | STS14 | STS15 | STS16 | STS-B | SICK-R | Avg. |
|---|---|---|---|---|---|---|---|---|---|
| BERT-base | CLS | 67.56 ±0.26 | 79.86 ±0.25 | 69.52 ±0.39 | 76.83 ±0.32 | 76.61 ±0.33 | 75.57 ±0.37 | 73.05 ±0.32 | 74.14 ±0.25 |
| | Mean | 67.30 ±0.44 | 81.96 ±0.24 | 71.92 ±0.28 | 77.68 ±0.47 | 76.71 ±0.48 | 76.90 ±0.40 | 73.28 ±0.30 | 75.11 ±0.21 |
| | Max | 64.61 ±0.87 | 82.06 ±0.21 | 72.43 ±0.31 | 76.56 ±0.74 | 75.61 ±0.43 | 76.61 ±0.52 | 72.15 ±0.46 | 74.29 ±0.33 |
| BERT-large | CLS | 66.22 ±0.79 | 82.07 ±0.39 | 71.48 ±0.33 | 79.34 ±0.44 | 75.38 ±0.60 | 73.46 ±0.45 | 74.30 ±0.50 | 74.61 ±0.41 |
| | Mean | 64.18 ±0.96 | 82.76 ±0.42 | 73.14 ±0.32 | 79.66 ±0.92 | 77.93 ±0.78 | 77.89 ±0.89 | 73.98 ±0.46 | 75.65 ±0.53 |
| | Max | 58.94 ±1.06 | 81.03 ±0.66 | 71.34 ±0.88 | 76.23 ±1.83 | 76.07 ±0.56 | 75.75 ±0.70 | 71.69 ±0.74 | 73.01 ±0.74 |
| RoBERTa-base | CLS | 65.55 ±0.89 | 80.84 ±0.26 | 71.87 ±0.39 | 78.77 ±0.70 | 79.29 ±0.27 | 78.13 ±0.61 | 74.92 ±0.18 | 75.62 ±0.38 |
| | Mean | 60.78 ±1.41 | 77.17 ±0.60 | 69.71 ±0.73 | 75.13 ±1.00 | 77.75 ±0.38 | 76.52 ±0.63 | 74.10 ±0.45 | 73.02 ±0.63 |
| | Max | 63.85 ±0.86 | 78.55 ±0.90 | 71.19 ±0.86 | 76.55 ±1.12 | 77.86 ±0.59 | 78.02 ±0.77 | 73.97 ±0.46 | 74.28 ±0.62 |
| RoBERTa-large | CLS | 63.84 ±1.34 | 77.33 ±2.53 | 68.64 ±1.34 | 72.86 ±1.96 | 77.13 ±1.32 | 78.32 ±1.08 | 74.14 ±1.31 | 73.18 ±1.20 |
| | Mean | 58.36 ±1.16 | 76.24 ±0.87 | 69.55 ±0.85 | 73.15 ±1.32 | 76.90 ±0.94 | 78.53 ±0.54 | 73.81 ±0.88 | 72.36 ±0.73 |
| | Max | 62.89 ±1.42 | 77.99 ±1.88 | 69.83 ±1.66 | 75.60 ±1.51 | 79.63 ±0.60 | 79.34 ±0.48 | 74.04 ±0.84 | 74.19 ±0.88 |

**表 6** 各モデル，pooling 手法ごとの SentEval の各タスクにおける正解率 (%)．表内の数値は，異なるシード値で 3 回評価を行った際の平均と標準偏差に 100 をかけたものである．

| Model | pooling | MR | CR | SUBJ | MPQA | SST-2 | TREC | MRPC | Avg. |
|---|---|---|---|---|---|---|---|---|---|
| BERT-base | CLS | 80.94 ±0.08 | 87.57 ±0.12 | 94.59 ±0.09 | 89.98 ±0.04 | 85.78 ±1.14 | 89.73 ±0.76 | 73.82 ±0.19 | 86.06 ±0.28 |
| | Mean | 81.84 ±0.17 | 88.20 ±0.04 | 94.82 ±0.12 | 89.94 ±0.12 | 86.49 ±0.20 | 89.73 ±0.31 | 75.32 ±0.78 | 86.62 ±0.18 |
| | Max | 80.74 ±0.16 | 88.00 ±0.09 | 94.32 ±0.07 | 89.92 ±0.25 | 85.03 ±0.09 | 89.13 ±0.50 | 74.11 ±0.49 | 85.89 ±0.02 |
| BERT-large | CLS | 85.79 ±0.19 | 90.54 ±0.26 | 95.58 ±0.14 | 90.15 ±0.04 | 91.17 ±0.06 | 90.47 ±0.95 | 73.74 ±0.61 | 88.20 ±0.07 |
| | Mean | 84.05 ±0.25 | 89.50 ±0.24 | 95.21 ±0.12 | 90.19 ±0.36 | 89.44 ±0.14 | 88.60 ±0.87 | 73.99 ±0.90 | 87.28 ±0.05 |
| | Max | 83.48 ±0.30 | 89.04 ±0.37 | 94.55 ±0.09 | 89.88 ±0.17 | 87.50 ±0.26 | 90.87 ±1.30 | 74.28 ±1.27 | 87.09 ±0.27 |
| RoBERTa-base | CLS | 83.94 ±0.30 | 90.44 ±0.49 | 94.05 ±0.06 | 90.70 ±0.17 | 89.16 ±0.22 | 90.80 ±0.35 | 75.52 ±0.42 | 87.80 ±0.20 |
| | Mean | 84.88 ±0.21 | 91.09 ±0.01 | 94.60 ±0.10 | 90.69 ±0.07 | 89.73 ±0.54 | 93.13 ±0.12 | 77.22 ±0.46 | 88.76 ±0.08 |
| | Max | 83.98 ±0.03 | 90.78 ±0.24 | 93.96 ±0.07 | 90.63 ±0.11 | 90.05 ±0.06 | 93.60 ±0.72 | 77.80 ±0.32 | 88.69 ±0.12 |
| RoBERTa-large | CLS | 85.63 ±0.27 | 90.74 ±0.15 | 94.53 ±0.14 | 91.20 ±0.11 | 90.08 ±0.59 | 93.53 ±0.76 | 72.66 ±1.73 | 88.34 ±0.28 |
| | Mean | 86.47 ±0.29 | 91.53 ±0.06 | 95.02 ±0.08 | 91.15 ±0.07 | 90.77 ±0.34 | 92.33 ±0.64 | 73.91 ±0.96 | 88.74 ±0.12 |
| | Max | 85.60 ±0.26 | 90.73 ±0.70 | 94.21 ±0.65 | 91.09 ±0.32 | 90.65 ±0.37 | 91.53 ±1.70 | 76.15 ±0.33 | 88.56 ±0.57 |