# いくつかの意味論的なタスクにおける単一事例規範と対照学習規範の併用

小早川 健[1*] 塩田 雄大[2] 望月 貴裕[1]
[1]NHK 放送技術研究所, [2]NHK 放送文化研究所
*kobayakawa.t-ko@nhk.or.jp

## 概要

いくつかの意味論的なタスクをとりあげ，BERT 型モデルに対して対照学習規範を用いた学習の効果を検証する．対照学習規範を BERT に適用した先行研究は，基となる BERT の埋込層に追加された Pooling 層のみを対照学習規範で学習するものが多いのに対して，この報告では，BERT の埋込層も含めた学習を行う．実験によると，対照学習規範は，単一事例規範と併用することによって，性能改善が達成できることがわかった．

## 1 はじめに

従来はかなり困難とされてきた自然言語処理タスクにおいて，深層学習の進歩によって一定の性能が達成されるようになってきている．ニューラルネットワークの構造やパラメータ最適化の手法もさることながら，正解ラベルを必要としない一般の文を事前に学習したモデルを基として，タスク固有の層をネットワーク構造に追加して適応化学習 (fine-tuning) をする手法はデファクトスタンダードになっている．実在する多量の文に言語モデル規範を適用して学習された事前学習モデルは広く一般に配布されている．それは，事前学習では，入力単語に対する埋込表現が十分な精度で獲得され，それが故に，タスク固有の層は比較的少量の fine-tuning でもよい精度が達成できるということが前提としてあると思われる．

そのような BERT 型モデルの成功にも拘らず，事前学習によって獲得された埋込表現の精度が十分かどうかは，議論の余地がある．例えば，埋込表現を特徴量に用いたクラスタリングを実施すると，素の BERT 型モデルのものよりも，対照学習を用いたモデルのほうがよい結果が得られる．これは，対照学習がクラスタリングと相性がよいというだけではなく，獲得された埋込表現の精度そのものがよい可能性が考えられる．深層距離学習の分野では埋込表現の精度そのものが着目され，さまざまな対照学習損失が提案されている一方で，一般に収束が遅いことや，性能改善に繋がらないことがあることも指摘されている．

以上に鑑みて，この研究では，クラスタリング以外の複数の意味論的なタスクを選定し，対照学習することの効果を検証したい．具体的には，1) スパムメールの判定, 2) SNS 発信が意見性を持つかどうかの判定, 3) 語用調査のための述語生成の 3 つを取り上げる．1) のスパムメールの判定は，スパムメールかどうかを 2 値判定するタスクで，公開データを用いる．2) は，SNS 発信に含まれる意見性を帯びた文の中から，意見の対象が具体的に記述されているかどうかを 2 値判定するタスクである．NHK の番組に関して意見が述べられている発信について，具体的なシーンを意見対象として述べているかどうかを判定する．3) は，文章校閲を視野に入れた，語用調査に AI を活用するための基礎調査である．過去の語用調査で検討項目に上がった言い回しに着目し，適切な述語を生成できるかどうか，単語穴埋め問題として解かせるタスクである．

## 2 関連研究

大規模な事前学習と 11 種の自然言語処理タスクに適応化学習を適用した BERT [1] は現在のデファクトスタンダードになっている．対照学習規範を用いた研究は，古くは署名検証 [2], word2vec の Continuous BOW モデルの埋込表現に適用したもの [3], BERT に適用した sentence-transformers(原論文では Sentence-Bert)[4] がある．sentence-transformers は，BERT の embedding 層に mean pooling を施し，対となる学習事例を教師とした cos 損失などを用いて学習するものである．また，深層距離学習として [5]

がある．

対照学習に用いられる損失関数は，様々なものが提案されているが，この研究では，[6, 7, 8] を用いる．

語用調査に AI を活用する事例には，プロンプトを用いた先行研究 [9, 10] がある．NHK 放送文化研究所では，ことばの研究として，日本語共通語の背景と現状分析のための調査研究による分析を行っている [11]．実験のうちのひとつは，こちらから題材を選択した．

# 3　モデル

対照学習済みのモデルとして配布されているものは，多段階の事前学習ステップのひとつとして，対照学習規範が単独で用いられることが多い．ここでは，タスク固有のレイヤーを学習の段階で併用し，タスクに適用した性能を監視しながら学習を進めることのできるモデルを提案する．

そのために，この研究では，2 種類の異なる損失関数の和を全体的な損失関数とする．深層学習型のニューラルネットワークで構築されたモデルの概要を図 1 に示す．一般に配布されている BERT 型のモデルを基にして，学習された embedding 層を下層に持つ．上層に，タスク固有層と一般的な sentence-transformers 型のプーリング層を並列に持ち，それぞれの損失関数を通じて，学習データから損失が計算される．図の左側は，タスク固有層で構築されるネットワークであり，1 入力テキストと 1 正解事例が対応付けから損失が計算され，このような規範を単一事例規範と呼ぶことにする．図の右側は，一般的な sentence-transformer モデルで採用される mean pooling 層であり，複数の事例とその関係性を表す正解事例から計算される損失が計算され，このような規範を対照学習規範と呼ぶことにする．

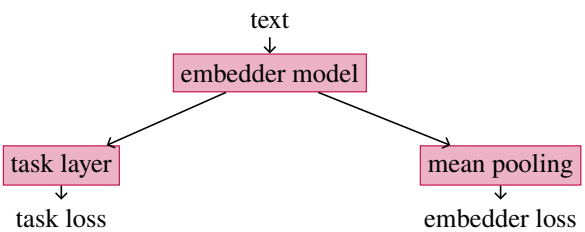


**図 1**　モデル

タスク固有層の中身は，ニューラルネットワークの一般的な構築法に従って，2 値分類問題では 1 層と sigmoid 活性関数を通じて binary cross entropy が，トークン生成問題では，語彙の数だけ並列する全結合層と softmax 関数を通じて cross entropy がタスク固有の損失関数 (task loss) となる．対照学習にはさまざまな損失関数が提案されてるが，そのうちのいくつかを実験にて検証する．対照学習は，オンライン学習のミニバッチに含まれる事例の中から，同じラベルの組を正例，異なるラベルの組を負例とする．

## 3.1　2 値分類の損失関数

2 値分類のための対照学習損失 (embedder loss) には，TripletLoss[7] を用いた．基準となる事例と同じカテゴリや異なるカテゴリに属する事例との距離の関係を学習するもので，ミニバッチの中から選ばれた 1 つの事例を基準とし，基準となる事例と同じカテゴリに属する事例を正例とし，その距離を $d_{\mathrm{ap}}$ とする．また，基準となる事例と異なるカテゴリに属する事例を負例とし，その距離を $d_{\mathrm{an}}$ とする．そして，一定のマージン $m$ 以上の差が開くような損失関数である．

$$\sum \max\left(d_{\mathrm{ap}} + m - d_{\mathrm{an}}, 0\right) \tag{1}$$

マージン $m$ は，実験を通じて既定値 0.5 に固定した．

## 3.2　述語生成の損失関数

述語生成のための対照学習損失 (embedder loss) には，Contrastive Loss[6] と AnglELoss[8] の和を用いた．

Contrastive Loss は，2 つの事例のユークリッド距離を $D$ として，損失関数

$$\begin{cases} \frac{1}{2}D^2 & (\text{正例のとき}) \\ \frac{1}{2}\max(0, m - D)^2 & (\text{負例のとき}) \end{cases} \tag{2}$$

で与えられる．AnglELoss は，複素数の埋込を出力する層の上に，正例どうしがなす角 $\Delta\theta_{ij}$ と負例どうしがなす角 $\Delta\theta_{mn}$ の差が広がるような損失関数

$$\log\left[1 + \sum \exp\left(\frac{\Delta\theta_{ij} - \Delta\theta_{mn}}{\tau}\right)\right] \tag{3}$$

で与えられる．マージン $m$ と温度 $\tau$ は，実験を通して，それぞれ既定値 0.5 と 1/20 に固定した．

# 4　実験

スパムメールの判定，意見性有無の判定，語用調査のための述語生成の 3 種類の実験を行った．それぞれの実験に別々のラベル付きの正解を用意した．

いずれの実験でも，数量が学習 8: 評価 2 の割合となるように分割し，学習データを用いて学習しながら評価データの損失関数を監視し，損失関数の減少が一定回数見られなくなるところで学習を停止した．

## 4.1 スパムメールの判定

代表的な 2 値分類の問題として，スパムメール判定のデータセットを用いる．表 1 にデータセットの詳細、表 2 に実験設定、および表 3 に実験結果を示す。単一事例のみを用いて学習した場合と比較して，対照学習規範である TripletLoss を併用した場合に性能向上が見られることがわかる．

| データセット名 | メッセージ数 |
|---|---|
| FredZhang7/all-scam-spam[1)] | 42,619 |

**表 1** スパムメール判定のデータ

| モデル | sentence-transformers/all-MiniLM-L6-v2 [2)] |
|---|---|
| 単一事例の損失関数 | binary cross entropy |
| 対照学習の損失関数 | TripletLoss |
| ミニバッチの事例数 | 1,024 |

**表 2** スパムメール判定の実験設定

| model type | 適合率 | 再現率 | f1 |
|---|---|---|---|
| binary CE only | 0.492 | 0.771 | 0.618 |
| +Triplet Loss | 0.897 | 0.949 | 0.922 |

**表 3** スパムメール判定の評価評価指標

## 4.2 意見性有無の判定

評判分析はいくつかの要素がある分析技術であり，そのひとつに，意見性の有無を判定するタスクがある．ここでは，番組に関する意見を分析するために，番組の具体的なシーンを対象として意見が述べられているかどうかを 2 値で判定するものであり，想定される利用法は，具体的なシーンを含まない発信をモデルによって自動的に除去することである．高い信頼性を持って除去できること，つまり高い precision が望ましい．表 4 に記すテレビ番組を対象として，放送時間帯になされた発信が，具体的なシーンに言及したものかどうかを判定する．表 5 に実験設定を示す．

1) https://huggingface.co/datasets/FredZhang7/all-scam-spam

2) https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

実験結果を表 6 に示す．評価データで f1 値の最も高い場所での precision と recall を合わせて示す．単一事例のみを用いて学習した場合と比較して，対照学習規範である TripletLoss を併用した場合に性能向上が見られることがわかる．

| 番組名 | 放送日 | 発信数 |
|---|---|---|
| らんまん | 2023.4.3〜4 | 6,964 |
| 虎に翼 | 2024.4.1〜3 | 15,239 |

**表 4** 意見性有無判定のデータ

| モデル | tohoku-nlp/bert-base-japanese-v3 [3)] |
|---|---|
| 単一事例の損失関数 | binary cross entropy |
| 対照学習の損失関数 | TripletLoss |
| ミニバッチの事例数 | 1,024 |

**表 5** 意見性有無判定の実験設定

| model type | 適合率 | 再現率 | f1 |
|---|---|---|---|
| binary CE only | 0.620 | 1.000 | 0.765 |
| +Triplet Loss | 0.847 | 0.912 | 0.879 |

**表 6** 意見性有無判定の性能評価指標

## 4.3 語用調査のための述語生成

ことばの研究として取り上げられたことのある例文を基にして，述語を空欄にした穴埋め問題を作成し，適切な候補を生成できるかどうか検証する．「犬にエサをやる。」という例文に着目し，述語の「やる」を空欄にして，単語を生成させる．

着目する例文: 犬にエサをやる。

単語穴埋め問題

犬にエサを [MASK]。

空欄の候補を生成し，生成された候補について，次の採点基準によって採点を行った．

**1** 文法，意味，語用ともに適切
**2** 文法，意味は適切だが，語用が不適切
**3** 文法は適切だが，意味，語用が不適切
**4** 文法，意味，語用が不適切

個々の候補のスコアを $\text{score}^{(\text{bare})}$ としたときに，生成候補上位 $N$ 個の総合評価指標は，生成された候補

3) https://huggingface.co/tohoku-nlp/bert-base-japanese-v3

の順番を加味した平均 (weighted evaluation)

$$\sum_{i=1}^{N} \frac{1}{i} \text{score}^{(\text{bare})} \Big/ \sum_{i=1}^{N} \frac{1}{i} \tag{4}$$

を用いた．これは，ラベルが正誤の 2 つの場合には，Mean Reciprocal Rank(MRR) に相当する．

学習と評価に用いるデータは，穴埋めされる単語を 100 個生成し，人手によって採点した．

単一事例規範として，マスクされた単語を復元する際に，生成された単語のクロスエントロピーに対して，採点基準に従って重み付けしたものを損失とした．対照学習に用いるデータは，学習データから 2 つの事例を抽出するすべての組み合わせに対して，同じカテゴリに含まれる場合は 1 を，異なるカテゴリに含まれる場合は 0 をラベルとして与えた．

表 7 にある設定を用いて，学習する様子を図 2 に示した．採点基準の選択肢にあるように，性能指標は値が小さいほどよく，学習が進むにつれて，性能が改善していることがわかる．学習が進んだ状態での性能を表 8 に示す．

| モデル | tohoku-nlp/bert-base-japanese-v3 |
|---|---|
| 生成する候補数 | 100 |
| 単一事例の損失関数 | Masked LM Loss |
| 対照学習の損失関数 | ContrastiveLoss, AnglELoss |
| ミニバッチの事例数 | 8 |

**表 7**　実験設定

| model type | weighted evaluation |
|---|---|
| masked LM loss only | 2.384 |
| +contrastive & angular loss | 1.986 |

**表 8**　モデル別の総合性能指標

## 5　おわりに

単一事例規範を単独で用いる学習と，対照学習規範を併用する学習について，比較実験による検証を行った．スパムメールの判定，SNS 発信が意見性を持つかどうかの判定, 語用調査のための述語生成の 3 種類の実験設定に置いて，それぞれ性能を比較したところ，いずれの場合に置いても，単一事例規範を単独で用いるよりも，対照学習規範を併用する学習のほうが性能が高いことが確かめられた．

改善の度合いには差があり，2 値分類の性能改善に比べて，述語生成の性能改善の度合いが少なかった．今後は，その原因と対策を検討していきたい．

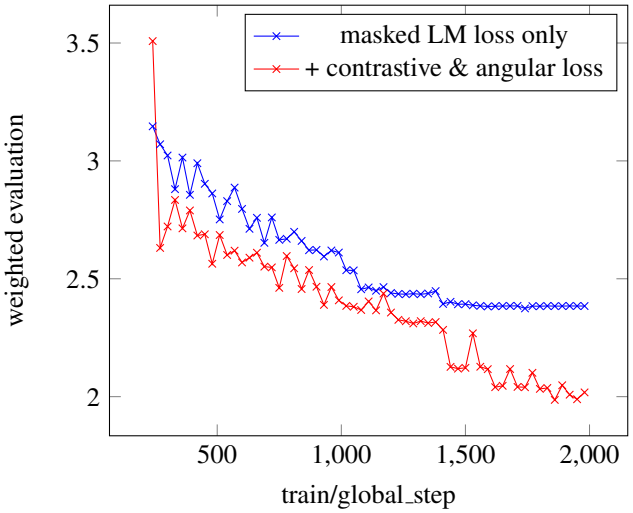


**図 2**　学習曲線

## 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 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.

[2] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. Signature verification using a "siamese" time delay neural network. In J. Cowan, G. Tesauro, and J. Alspector, editors, **Advances in Neural Information Processing Systems**, Vol. 6. Morgan-Kaufmann, 1993.

[3] Tom Kenter, Alexey Borisov, and Maarten de Rijke. Siamese CBOW: Optimizing word embeddings for sentence representations. In Katrin Erk and Noah A. Smith, editors, **Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 941–951, Berlin, Germany, August 2016. Association for Computational Linguistics.

[4] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, **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, Hong Kong, China, November 2019. Association for Computational Linguistics.

[5] Kevin Musgrave, Serge J. Belongie, and Ser-Nam Lim. Pytorch metric learning. **ArXiv**, Vol. abs/2008.09164, , 2020.

[6] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In **2006 IEEE**

**Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)**, Vol. 2, pp. 1735–1742, 2006.

[7] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**, June 2019.

[8] Xianming Li and Jing Li. AoE: Angle-optimized embeddings for semantic textual similarity. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 1825–1839, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[9] Daniel Fried, Nicholas Tomlin, Jennifer Hu, Roma Patel, and Aida Nematzadeh. Pragmatics in language grounding: Phenomena, tasks, and modeling approaches. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, **Findings of the Association for Computational Linguistics: EMNLP 2023**, pp. 12619–12640, Singapore, December 2023. Association for Computational Linguistics.

[10] Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. A fine-grained comparison of pragmatic language understanding in humans and language models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 4194–4213, Toronto, Canada, July 2023. Association for Computational Linguistics.

[11] ことばの研究. https://www.nhk.or.jp/bunken/research/kotoba/index.html.