E-050

文脈依存語に対する対照学習に基づく皮肉検出
Sarcasm Detection based on Contrastive Learning for Context-Dependent Words

テサウンエー‡ 綱川 隆司‡ 西田 昌史‡
Thet Saung Aye Takashi Tsunakawa Masafumi Nishida

1. はじめに

皮肉表現の検出は、テキストの感情を正確に分析するために重要である。たとえば、"What a lovely reason to cancel my weekend plans."という表現は、表面的には肯定的な意味を持つ単語を含むが、実際には否定的な皮肉表現として用いられている場合が多い。このような表現を適切に検出できなければ、誤った感情判定につながるおそれがある。

皮肉表現の検出には、単語やフレーズの表面的な意味だけでなく、文脈に依存した解釈を正しく捉えることが不可欠である。たとえば、前出の文における "lovely" は、"What a lovely afternoon to sit outside and read." における "lovely" とは、皮肉としての解釈が異なる。Ghosh ら[1]は発話と対象語が与えられた場合に、その対象語の意味が文字通りのものか皮肉的なものかを判別する手法を提案した。

本研究では、文脈に応じて意味が変化する語に対して、皮肉的な用法と文字通りの用法を区別可能な埋め込み表現を獲得することを目的とする。文脈に応じて文字通りにも皮肉的にも解釈され得る語（以下、文脈依存語と呼ぶ）に着目し、対照学習を用いてその埋め込み表現の獲得を試みる。学習された埋め込み表現の評価として、その表現を用いた線形分類器を皮肉検出タスクで学習し、皮肉検出用データセット SARC を用いて評価する。特定の語に関して、対照学習により得られた埋め込み表現が、従来のファインチューニング手法と同等の性能を達成することを示す。

2. 関連研究

2.1 皮肉検出

皮肉検出は、文章が皮肉を含むか否かを識別する二値分類問題として扱われてきた。近年では、事前学習済みの言語モデルに対し、皮肉を含む文章とそのラベル情報を用いて再学習を行うことで、性能を最適化する手法が広く採用されている。それに加えて、発話に関連する文脈的・感情的情報の活用も提案されている[2-4]。

発話全体に対して皮肉性の有無を予測する手法とは異なり、Ghosh ら[1]は、与えられた発話および注目語に基づき、その語が文字通りの意味で用いられているのか、皮肉的に使われているのかを識別する手法を提案している。本研究は、語義における皮肉的曖昧性を解消するという目的で文脈依存語に着目している点において、Ghosh らの手法と共通している。一方で本研究では、文脈間の意味の対比に基づいて埋め込み表現を学習するという、異なるアプローチを採用しており、この点で相違がある。

2.2 対照学習

対照学習は、類似したサンプル同士の表現を近づけ、非類似なサンプル同士の表現を遠ざけることで、効果的な特徴表現を獲得することを目的とする手法である（Hadsell ら[5]）。本研究では Gao ら[6]が提案した対照学習の目的関数を採用した。文字通りに使われている 2 つの異なる文 x_i, x_i^+、同じ語が皮肉的に使われている文 x_i^- に対し、それぞれの文埋め込み表現が h_i, h_i^+, h_i^-とする。N 文からなるバッチに対して、目的関数は以下のように定義される。

$$-\log \frac{\exp(\mathrm{sim}(h_i,h_i^+)/\gamma)}{\sum_{j=1}^{N}\left(exp\left(\mathrm{sim}\left(h_i,h_j^+\right)/\gamma\right)+\alpha^{\mathbb{I}_i^j} exp\left(\mathrm{sim}\left(h_i,h_j^-\right)/\gamma\right)\right)} \quad (1)$$

γ は温度のハイパーパラメータであり、$\mathrm{sim}(h_1, h_2)$ はコサイン類似度を表す。$\mathbb{I}_i^j \in \{0,1\}$ は i=j のとき、かつそのときに限り 1 となる指示関数である。本研究では、この目的関数を用いて事前学習モデルの学習を行った。

表 1　文脈依存語と学習用トリプレット数

important(4519), interesting(1304), wonderful(1240), confused(795), lovely(532), horror(518), pathetic(450), creators(340), arrogant(205)

3. 提案手法

本研究では、まず皮肉的に用いられる語を抽出する既存手法を用いて文脈依存語を抽出し、それらを含む文を基に、皮肉文と文字通り文の対照文ペアを構築する。次に、これらの文対を用いて対照学習を行い、皮肉表現を効果的に捉える文埋め込み表現の獲得を目指す。最後に、得られた文埋め込みを固定し、単純な線形分類器によって皮肉検出タスクの性能を評価し、従来のファインチューニング手法との比較を行う。

3.1 学習データの構築

文脈に応じて文字通りにも皮肉的にも解釈され得る語を抽出するために、既存研究を参照した。Aye ら[7]は皮肉検出の教師あり学習とコサイン類似度を用いて、皮肉的に用いられる単語を抽出する手法を提案した。本研究ではその単語集合を文脈依存語として採用し、これの文脈を基に対照学習を行う。

対照学習で用いる正例ペアと負例ペアの作成には、すでに皮肉に関する正解ラベルが付与されている教師あり学習用データセットを使用した。まず、文脈依存語を含む表現を抽出した。データのバランスを保つため、皮肉文と文字通りの文は同数となるように調整し、（h_i, h_i^+），（h_i, h_i^-）ペアを生成した。表 1 には、文脈依存語と、それに対する学習に用いたトリプレット（h_i, h_i^+, h_i^-）の数を示す。

対照学習およびベースライン手法の評価、ならびにベースライン手法の学習に必要な皮肉検出用データは、文脈依存語を含む文と、対応する元の教師あり学習データセットの正解ラベルを用いて構成した。

‡ 静岡大学　Shizuoka University

3.2 文埋め込み表現の対照学習と評価

3.1 節で構築した対照文ペアを用いて、式 1 を最適化した。本節では、文埋め込み表現を得るためのプーリング手法および対比損失における重み係数について説明する。

一般に、事前学習済み言語モデルの出力から文の埋め込み表現を得る際には、文頭の特殊トークン（例：BERT における[CLS]）に対応する隠れ状態を利用する方法が広く用いられている。一方で、意味的類似度の計算などを目的とするタスクでは、特に最初と最終の隠れ層のすべてのトークン出力を平均化して文埋め込みとする手法の方が効果的であることが報告されている（Reimers ら, 2019[8]）。本研究では複数のプーリング手法を比較した結果、[CLS]トークンによるプーリングでは F1 スコアが 0.702 と、平均プーリングの 0.738 と比較して性能が低下した。

式 1 における重み係数 α は、文脈依存語を含む文に対し、その語が文字通りに使用される場合と皮肉的に使用される場合との対比が損失関数に与える影響度を調整するパラメータである。直感的には、α を高く設定することで、皮肉との違いがより強調されると考えられる。複数の値を設定して比較実験を行った結果、α=20 のときに最も良好な性能を示したため、以下の実験ではその値を用いた。

対照学習によって得られた文埋め込み表現の有効性を確認するため、これらの埋め込みを固定した状態で、単純な線形分類モデルを皮肉検出タスクで学習させ、性能を評価した。

3.3 ベースラインモデル

ベースラインとする皮肉検出のファインチューニング手法には、従来の研究において高い精度を示した手法を採用した [2-3]。これらの手法では、事前学習済みモデルに対し、発話とその皮肉ラベルを用いて再学習を行う。本研究においても同様に、文脈依存語を含む文とその皮肉ラベルを用いてモデルの学習を行った。

4. 実験設定

実験には、Reddit の投稿を収集して構築された皮肉検出用データセット SARC 2.0（Khodak ら [9]）を使用した。文脈依存語 9 語（表 1）に対して、データセットから合計 42,465 件の投稿を抽出した。データは、学習用 80%、検証用 10%、テスト用 10%に分割した。表 1 に示すように、対照学習用トリプレットの数は、「important」に対する 4519 組から、「arrogant」に対する 205 組までの範囲である。

本研究では、BERT（Devlin ら[10]）の事前学習済み言語モデルをベースモデルとして採用した。モデルの事前学習済みチェックポイントを初期値として用い、SARC 2.0 の学習データに対して再学習を行った。

学習率、バッチサイズ、エポック数などをグリッドサーチにより最適化した。その結果、対照学習は、学習率 5×10^{-5}、バッチサイズ 32、エポック数 20 で学習を行った。ベースライン手法のファインチューニングには、学習率 2×10^{-5}、バッチサイズ 32、エポック数 4 を用いた。全ての手法において、クロスエントロピー損失を用いて学習した。

5. 結果と考察

従来のファインチューニング手法と、提案する対照学習手法による皮肉検出性能を比較した。加えて、事前学習モデルから得られる文埋め込み表現について、[CLS]プーリング手法と、隠れ層の平均による文埋め込みを固定し、線形変換モデルで皮肉検出性能を比較した。簡潔さのため、各文脈依存語の皮肉判定に対する適合率、再現率、および F1 スコアの平均とその標準偏差、さらに各手法における最高および最低性能を示した単語を報告する。

表２　皮肉検出の性能評価

Model	Precision	Recall	F1	Max F1 (word)	Min F1 (word)
Avg. BERT embeddings	0.807 ± 11	0.578 ± 8	0.664 ± 4	0.734 (important)	0.621 (confused)
BERT-[CLS] embeddings	0.693 ± 10	0.550 ± 7	0.608 ± 5	0.703 (important)	0.555 (horror)
Finetune-BERT	0.765 ± 6	0.756 ± 7	0.760 ± 6	0.845 (creators)	0.698 (confused)
Contrastive Learning-BERT(Proposed)	0.705 ± 8	0.777 ± 7	0.738 ± 7	0.821 (creators)	0.642 (pathetic)

表２は、皮肉検出タスクの性能評価を示している。提案手法である対照学習は、従来のファインチューニング手法と比べて、平均 F1 スコアの差は 0.03 未満で、ほぼ同等の性能を示した。また、提案手法で学習したモデルは、事前学習モデルから得た文埋め込みを用いた場合と比較して、F1 スコアが 0.664 から 0.738 まで向上した。対照学習により、文脈依存語の皮肉と文字通りの用法を区別する埋め込み表現が得られることが示唆された。

さらに、否定的な語を含む皮肉的文脈の認識は困難であることが示唆された。「confused」「pathetic」「horror」は、いずれの手法においても最低の性能を示していた。

6. おわりに

本研究より対照学習が、文脈依存語に対して、皮肉的な用法と文字通りの用法を区別可能な埋め込み表現を獲得することができたことを示唆した。

対照学習は、ラベルが類似した文をクラスター化するように学習するため、皮肉検出のような分類タスクにおいては、線形変換モデルの学習時に目的の不一致が生じる可能性がある。今後は、対照学習に分類目的を組み込むような手法の検討が必要である。

参考文献

[1] Ghosh et al. Sarcastic or Not: Word Embeddings to Predict the Literal or Sarcastic Meaning of Words. EMNLP. 2015.
[2] Potamias et al. A transformer-based approach to irony and sarcasm detection. Neural Comput & Applic 32. 2020.
[3] Babanejad et al. Affective and Contextual Embedding for Sarcasm Detection. COLING 28. 2020.
[4] He et al. Sarcasm Detection Base on Adaptive Incongruity Extraction Network and Incongruity Cross-Attention. Appl. Sci. 2023.
[5] Hadsell et al. Dimensionality Reduction by Learning an Invariant Mapping. CVPR'06. 2006.
[6] Gao et al. SimCSE: Simple Contrastive Learning of Sentence Embeddings. EMNLP. 2021.
[7] Aye et al. 皮肉検出における BERT の単語埋め込みベクトルの分析. WiNF. 2024.
[8] Reimers et al. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. EMNLP-IJCNLP. 2019.
[9] Khodak et al. A Large Self-Annotated Corpus for Sarcasm. LREC. 2018.
[10] Devlin et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL. 2019.