議論の流れを利用した皮肉を含むコメントの検出

武石 健吾
東京工業大学 情報理工学院
takeishi.k.aa@m.titech.ac.jp

徳永 健伸
東京工業大学 情報理工学院
take@c.titech.ac.jp

1 背景と目的

皮肉は複雑な言語現象であり，発話の真の意味がその文字通りの意味とは異なる．皮肉の理解はソーシャルメディアを対象としたオピニオンマイニングや感情分析において難しい課題となる．皮肉検出の先行研究では，主に文中の語彙的な手がかりを利用している [1]．最近の研究では皮肉検出における文脈の重要性を指摘しているものが多い．ソーシャルメディア・テキストのコメントは，多くの場合，先行するコメントや世界で起きている出来事に関連して行われる．これは文脈が皮肉の検出に重要な役割を果たすことを示唆している．

本研究は議論の流れを用いて皮肉を含んだコメントを検出する手法を提案する．議論を構成する各コメントはその議論のトピックに対して直接言及しているものと別のコメントに対して言及しているものに分けられる．後者のコメントにさらに別のコメントが付けられることで議論が構成される．本研究で用いる議論の流れとは検出対象とするコメントからトピックまで親コメントを辿るパス中のコメントの系列を指す．

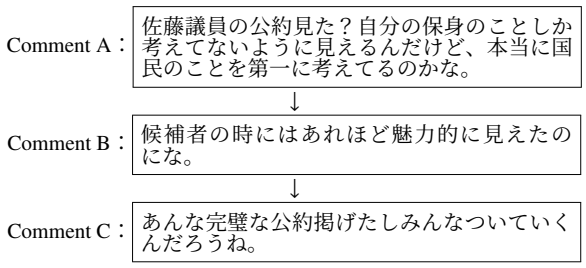

図 1 不自然な議論の流れ

議論の流れを皮肉の検出に利用する動機は議論の流れからして不自然である皮肉を含んだコメントに注目することである．図 1 は，政治カテゴリに投稿された佐藤という議員が掲げた公約に関するトピックへのコメントである．Comment A に対して Comment B が投稿され，さらに Comment B に対して Comment C が投稿されている．Comment A と Comment B を参照すると佐藤議員の公約への批判の流れが発生している．この次に投稿されるコメントが公約を擁護するものである場合，逆説の意味を持つフレーズが入ることが予想される．しかし，続くコメントは逆説フレーズがないにも関わらず公約に賛成するような文意を持っている．このことから Comment C は皮肉を含んだコメントであることがわかる．このように議論の流れを参照することで，皮肉の検出精度が向上することが期待される．

2 関連研究

本研究はオンラインディスカッションフォーラム Reddit[1)] に投稿されたコメントを検出対象とする．ある著者があるトピック（ニュース記事へのリンクなど）を Reddit に投稿すると，著者同士でこのトピックについて Web サイトで議論することができる．これにより Reddit は図 2 のような各コメントに親コメントが含まれるツリー構造の会話構造が確立されている．

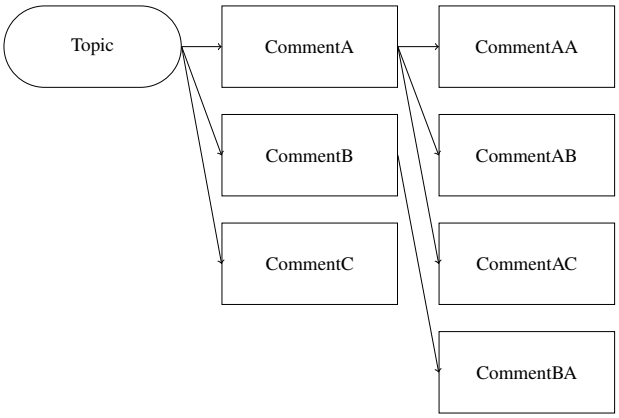

図 2 Reddit の構造

Reddit に投稿されたコメントの皮肉検出の研究の動向について以下で紹介する．文献 [2] ではオンラインソーシャルメディアディスカッションでの皮肉検出のために，内容駆動型と文脈駆動型の両方のモデリングのハイブリッドアプローチを採用し

1） http://reddit.com

た CASCADE というモデルを提案している．文脈情報の 1 つ目は著者の行動特性を表現する著者ベクトルである．著者の過去の投稿を利用して，ライティングスタイルとパーソナリティインジケーターをモデル化し，その後，正準相関分析（CCA）を使用して包括的な著者ベクトルに融合する．文脈情報の 2 つ目は談話ベクトルである．これは，同じフォーラムに属するこれらの統合されたコメントのドキュメントモデリングによって行われる．これらの談話の特徴は，皮肉を検出するために必要なジャンルの情報とともに，重要な文脈情報，背景の手がかりを与える．内容のモデリングには CNN (Convolutional Neural Network) を使用して構文上の特徴を抽出する．本研究では CASCADE をベースラインとし，不自然な議論の流れを特徴量として加えたモデルを提案する．文献 [3] では特定のコーパスで皮肉を含むコメントを検出するために，マルチヘッドアテンションベースの双方向ロングショートメモリ（MHA-BiLSTM）ネットワークを提案している．また，感情や句読点などの様々な特徴量を検討している．

3 提案モデル：議論の流れの導入

本研究は文献 [2] で提案されているモデル CASCADE をベースラインとし，CNN，Sentence-BERT [4]，MT-DNN[5] をそれぞれベースとしたモデルを 3 つ提案する．

3.1 CASCADE

CASCADE はコメントの内容および文脈の情報を活用する． 内容モデリングでは，CNN を使用してコメントのベクトル表現 $\vec{c}_{ij}$ を生成している．このベクトル $\vec{c}_{ij}$ は，構文情報と意味情報の両方を取得する．文脈モデリングは，著者ごとの特徴量とトピックのカテゴリを表す’’ Subreddits’’ ごとの特徴量をそれぞれ学習し，著者ベクトル $\vec{u}_i$ と談話ベクトル $\vec{t}_j$ を生成する．その後，これら 3 つのベクトル $\vec{c}_{ij}$，$\vec{u}_i$，および $\vec{t}_j$ を連結し皮肉または非皮肉に分類する．

3.2 議論の流れを表すテキスト

本研究では議論の流れを表現するために「話題テキスト」と「議論テキスト」を利用する．話題テキストは議論の始まりに位置するトピックを形成しているテキストである．図 2 のように全てのコメントは一つのトピックを根ノードとする木のノードである．話題テキストは議論の先頭に位置しており議論の方向性を決定すると考えられる．一方，議論テキストは議論の流れを形成している各コメントのテキストを連結して生成する．連結の際にはコメントの境目を明確にするために区切り文字 “<END>” を各コメントから取得したテキストの間に差し込む．議論テキストは議論の流れ自体を表している．これら 2 つのテキストと検出対象のコメントのテキストである「対象テキスト」との関係をモデリングすることで不自然な議論の流れを利用して皮肉を含むコメントを検出する．

3.3 CNN をベースとしたモデル

CNN でのテキストのバッチモデリングでは，テキストの長さを統一するために制限またはパディングが行われる．先行研究ではテキストの単語数を制限する際は設定値に達するまで先頭から順に単語を取得しており，本研究でも対象テキスト及び話題テキストは同様の方法で長さの制限を行う．一方で議論テキストの場合，議論の先頭に近いコメントほど検出対象から遠ざかり検出に有用な情報が少ないことが予想されるため，各コメントのテキストから同数の単語を取得した後に区切り文字 “<END>” を各コメントから取得したテキストの間に差し込むことで議論テキストを生成する．この時，区切り文字の数も含めてテキストの長さが統一されるように単語の取得数を調整する．その後，話題テキスト，議論テキスト，対象テキストを区切り文字 “<END_C>” を介して遠結する．

次に連結したテキストを CNN に入力として加えベクトル表現を生成する．この表現は両テキストの構文情報と意味情報に加えテキスト間の関係を含む．その後 CASCADE と同様に CNN を用いて生成したベクトルに著者ベクトルと談話ベクトルを連結し皮肉または非皮肉に分類する．話題テキスト，議論テキスト及び対象テキストの長さは先行研究と等しくなるようそれぞれ 100 に設定する．よって結合したテキストの長さは区切り文字も含めて 302 となる．それに合わせて関係ベクトルのサイズは 302 に設定する．

3.4 Sentence-BERT をベースとしたモデル

本研究では皮肉検出タスクを文関係の推論タスクと捉える．そこで Sentence-BERT [4] をベースとし

たモデルを検討する．Sentence-BERT の文埋め込みは InferSent [6] や Universal Sentence Encoder [7] などの他の最先端の文埋め込み方法よりも優れていることが知られている．また，Fine-tuning 用のデータセットとして SNLI [8] と Multi-Genre NLI [9] を用いており，NLI (Natural Language Inference) タスクに強みを持つ．

Sentence-BERT を使用して関係ベクトルを生成し，CASCADE のベクトル結合層に加えたモデルを検討する．文献 [4] の実験で最高性能であったベクトルの連結方法を参考にし，テキスト A とテキスト B の関係ベクトル v_A と v_B を利用した $(v_A, v_B, |v_A - v_B|)$ を採用する．すなわち，結合層に含まれるベクトルは CNN で生成した対象テキストのベクトル表現 $\vec{C}_{ij}$，Sentence-BERT で生成した関係ベクトル $(v_A, v_B, |v_A - v_B|)$，著者ベクトル $\vec{u}_i$，談話ベクトル $\vec{t}_j$ である．本研究では，Sentence-BERT を用いて話題テキスト，議論テキスト，対象テキストをそれぞれ $v_{\text{topic}}, v_{\text{chain}}, v_{\text{target}}$ にベクトル化し，$v_{\text{target}}, v_{\text{chain}}, v_{\text{topic}}, |v_{\text{target}} - v_{\text{chain}}|, |v_{\text{target}} - v_{\text{topic}}|$ を連結して関係ベクトルとして使用する．モデルに使用する Sentence-BERT のパラメータは文献 [4] に従う．

3.5 MT-DNN をベースとしたモデル

文関係の推論タスクで他のモデルを上回る性能を示した MT-DNN [5] を皮肉検出タスクで利用することを検討する．MT-DNN は BERT をマルチタスク学習に拡張したモデルであり，一般的な言語理解評価ベンチマーク 9 タスク中 8 つで最高の結果が出ている．また SNLI，SciTail といった，文関係の推論タスクの精度が素の BERT に比べて大幅に改善されている．本研究では皮肉検出で有用と考えられ CASCADE にもコンテキストモデリングとして導入されている，著者ごとの特徴量と Subreddits ごとの特徴量をそれぞれ学習し生成された著者ベクトルと談話ベクトルを MT-DNN に導入する．

皮肉検出タスクは MT-DNN が適応する 4 つのタスクのうち，文関係の分類タスクに該当する．このタスクは，長さ m の Premise：$P = (p_1, \ldots, p_m)$ と長さ n の Hypothesis：$H = (h_1, \ldots, h_n)$ を入力として加え，P と H 間の論理関係 R を見つける．出力モジュールの設計は，ニューラル NLI モデルである確率的回答ネットワーク (SAN) [10] の回答モジュールを使用している．その中で t ステップ目の状態 s_t とその前の状態 s_{t-1} および Premise のワーキングメモリ M_p から計算される x_t を用いて各ステップ $t \in 0, 1, \ldots, T-1$ の関係の確率分布 ${P_t}^r$ は式 (1) で計算される．

$$ {P_t}^r = \text{softmax}(\theta_4[s_t; x_t; |s_t - x_t|; s_t \cdot x_t]) \qquad (1) $$

式 (1) の 4 つのベクトルが連結されている部分に著者ベクトルと談話ベクトルも連結することで素性としてモデルに加える．モデルに使用する MT-DNN のパラメータは文献 [5] に従う．

表 1　各データセット中のコメント数

	訓練データ		テストデータ	
	均衡	非均衡	均衡	非均衡
皮肉	179,700	179,700	44,637	44,637
非皮肉	179,700	539,100	44,637	133,911

4 評価実験

4.1 データセット

評価のために自己注釈付き Reddit コーパス SARC [11] を用いる．このデータセットには Reddit に投稿された皮肉を含む/皮肉を含まないコメントの 553M の例が含まれている．各コメントには，コメントのテキスト，著者，所属する SubReddit，および親コメントなどの情報が含まれている．公開されている不均衡データ[2)]の訓練データとテストデータからランダムサンプリングし，皮肉を含むコメント数と皮肉を含まないコメント数の割合が 1:1 の均衡データセットと両カテゴリのコメント数の割合が 1:3 の不均衡データセットの 2 つを作成した．モデルの学習には訓練データの 10%を開発データとして使用する．データセットの統計を表 1 に示す．

これらのデータセットは大規模であるため同じモデルでの結果のバラつきは小さいと判断し，モデルごとにシード 3 回を変更した結果の平均値を計算する．評価指標には Precision (P), Recall (R), Accuacy (A), F 値 (F) を用いる．再現率と適合率のどちらも重要であるという考えられているため，この中で Accuacy と F 値を重要視してモデル性能の比較を行う．

4.2 結果と考察

実験結果を表 2 と表 3 に示す．議論の流れを導入した 3 つのモデルの性能が CASCADE の性能を上

2）https://nlp.cs.princeton.edu/SARC/2.0/main/

表 2　均衡データセットの結果

	P	R	A	F
CASCADE	0.836	0.864	0.847	0.850
CNN	0.849	0.875	0.859	0.861
SBERT	0.854	0.883	0.866	0.868
MT-DNN	0.869	0.904	**0.883**	**0.886**

表 3　不均衡データセットの結果

	P	R	A	F
CASCADE	0.767	0.726	0.876	0.746
CNN	0.774	0.779	0.888	0.776
SBERT	0.797	0.753	0.890	0.774
MT-DNN	0.841	0.768	**0.906**	**0.803**

回っていることがわかる．また，3 つのモデルの中でも MT-DNN をベースにしたモデルの性能が他を上回っている．これを更に調査するために，不均衡データセットについて議論の流れ (議論テキスト) を利用できるコメントと利用できない (議論の先頭に位置する) コメントに分けて各指標を算出した結果を表 4 と表 5 に示す．

表 4　議論の流れを利用できるコメントの結果

	P	R	A	F
CASCADE	0.967	0.743	0.757	0.840
CNN	0.920	0.967	0.899	0.943
SBERT	0.930	0.949	0.895	0.939
MT-DNN	0.931	0.971	**0.913**	**0.951**

まず，CASCADE と CNN 及び SBERT ベースとのモデルの性能の違いに着目する．議論の流れを利用できないコメントでは Accuracy 及び F 値に目立った差は見られない．一方，議論の流れを利用できるコメントでは議論の流れを導入した 2 つのモデルの性能が優っている．つまり，テキストの構文情報と意味情報の利用における性能の違いはなく，議論の流れの導入によって性能が向上したと結論づけられる．次に CNN 及び SBERT ベースのモデルと MT-DNN ベースのモデルの性能の違いに着目する．Accuracy 及び F 値は議論の流れの有無に関わらず性能は向上している．この場合，議論の流れが有用に働いたか判断することができない．そこで，話題テキストと議論テキストを使わず MT-DNN の全ての Premise テキストを空白テキストにしたモデルの性能を検証する．各データセットでの実験結果を表 6 と表 7 に示す．

CASCADE と MT-DNN (Premise なし) では後者のモデルの方が性能が良いことから MT-DNN の内容

表 5　議論の流れを利用できないコメントの結果

	P	R	A	F
CASCADE	0.709	0.719	0.887	0.714
CNN	0.715	0.707	0.887	0.711
SBERT	0.740	0.679	0.890	0.708
MT-DNN	0.800	0.690	**0.905**	**0.741**

表 6　均衡データセットにおける Premise の効果

	P	R	A	F
CASCADE	0.836	0.864	0.847	0.850
MT-DNN (Premise なし)	0.849	0.895	0.868	0.871
MT-DNN (Premise あり)	0.869	0.904	0.883	0.886

ベースのモデリングの方が性能が高いと考えられる．MT-DNN では Premise がある方が性能が高いことから MT-DNN が議論の流れをモデリングしたことが皮肉検出に有効に働いていると考えられる．

5　結論

本研究ではオンラインディスカッションフォーラム内の皮肉を含むコメントの検出に議論の流れを利用することを提案した．議論の流れを導入したモデルとして CNN ベースのモデル，SBERT ベースのモデル，MT-DNN ベースのモデルを検討し，性能の比較を行なった．その結果，議論の流れが皮肉検出に有用に働くことを確認した．今後の課題は議論を構成するコメントのテキストからの単語の取得方法の更なるバリエーションである．本研究では各コメントのテキストからの単語の取得は先頭から行ったが，同じテキストでも単語を取得する部分を工夫する余地がある．

参考文献

[1] Roger Kreuz and Gina Caucci. Lexical influences on the perception of sarcasm. In *Proceedings of the Workshop on computational approaches to Figurative Language*, pp. 1–4, 2007.

[2] Devamanyu Hazarika, Soujanya Poria, Sruthi Gorantla, Erik Cambria, Roger Zimmermann, and Rada Mihalcea. Cascade: Contextual sarcasm detection in online discussion forums. *arXiv preprint arXiv:1805.06413*, 2018.

[3] A. Kumar, V. T. Narapareddy, V. Aditya Srikanth, A. Malapati, and L. B. M. Neti. Sarcasm detection using multi-head

表 7　不均衡データセットでの各評価指標

	P	R	A	F
CASCADE	0.767	0.726	0.876	0.746
MT-DNN (Premise なし)	0.797	0.745	0.889	0.770
MT-DNN (Premise あり)	0.841	0.768	0.906	0.803

attention based bidirectional lstm. *IEEE Access*, Vol. 8, pp. 6388–6397, 2020.

[4] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

[5] Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. Multi-task deep neural networks for natural language understanding. *arXiv preprint arXiv:1901.11504*, 2019.

[6] Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. Supervised learning of universal sentence representations from natural language inference data. *arXiv preprint arXiv:1705.02364*, 2017.

[7] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*, 2018.

[8] Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*, 2015.

[9] Adina Williams, Nikita Nangia, and Samuel R Bowman. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*, 2017.

[10] Xiaodong Liu, Kevin Duh, and Jianfeng Gao. Stochastic answer networks for natural language inference. *CoRR*, Vol. abs/1804.07888, , 2018.

[11] Mikhail Khodak, Nikunj Saunshi, and Kiran Vodrahalli. A large self-annotated corpus for sarcasm. *arXiv preprint arXiv:1704.05579*, 2017.