

コモンセンス知識を利用した物語中の登場人物の感情推定

田辺 ひかり

小川 哲司

小林 哲則

林 良彦

早稲田大学理工学術院

tanabe@pcl.cs.waseda.ac.jp

1 はじめに

物語における登場人物の感情推定においては、直接的な感情表現を手がかりとするだけではなく、常識的知識を援用して「行間を読む」ことが必要となる。本研究では、物語を構成する各文における登場人物の感情を推定するというタスクにおいて、事態に関するコモンセンス知識を導入し、さらに、対象文の先行文脈を考慮することで、感情推定の精度を向上させることを試みる。

本研究で対象とする Story commonsense データセット [1] は、5文からなる短い物語の各文に対し、登場人物の感情カテゴリが、Plutchik の基本感情 [2] を利用したマルチラベルの形態で付与されている。このため、対象タスクはマルチラベルの分類問題として定式化される。本研究では、外部のコモンセンス知識として、出来事と出来事、出来事と人物の心的状態の関係に関する事態知識を収集した知識グラフ ATOMIC [3] を使用し、文脈情報としては、対象文の先行文を用いた。

評価実験において、対象文、文脈、外部知識それぞれからの情報の統合方法に関して複数の設定を比較した結果、外部知識、文脈の双方を用いることが推定精度向上に有効であるが、その際は、双方の情報を別々に処理した後にアンサンブル的に統合することが有効であること、および、用いる外部知識はタスクに適合するように事前に選別しておくことが有用であることなどが確認された。

2 関連研究

Rashkin ら [1] は物語中の人物の心的状態を推定するためのデータセットとして Story commonsense データセットを構築した。物語理解のデータセットでは、物語の中で起きたことや物語の前後を推論する QA 形式のものが多く [4][5][6]、人物の心的状態に焦点を当てたデータセットは珍しい。

Story commonsense データセットには基本的欲求と人物の感情の二つの側面でラベル付与がなされている。Paul と Frank [7] は、ConceptNet [8] を利用した基本的欲求ラベルの推定を行っている。物語と欲求カテゴリを関係付ける可能性のあるパスを ConceptNet から取り出し、PageRank などのグラフ理論的指標により算出したパスの重要度が高い上位 k 本のパスを推定に利用することで、外部知識を用いないモデルを上回る精度を達成した。しかし、本研究で扱う感情推定は行っていない。

Story commonsense データセットにおける感情推定を行った研究として、[9, 10, 11] がある。対象タスクがマルチラベル分類問題であるため、カテゴリ間の依存・相関関係の扱いが課題となる。

田辺ら [9][11] は Plutchik の環の構造、学習データにおける感情ラベルの共起関係を用い、感情極性の二値分類とのマルチタスク学習を行った。さらに推定対象文の先行文脈を取り込むことで推定精度を向上させた。また、モデルから得られる文の分散表現の主成分分析から Plutchik の環の構造が概ね妥当であることを示した。

Gaonkar ら [10] は、Label-Embedding Attentive Network (LEAM) [12] を利用することにより、感情カテゴリ間関係を考慮し、さらに、BERT を用いた label embedding の獲得やラベル同士の共起関係の利用、ラベルなしデータを用いた半教師あり学習を行うことにより、当該タスクの SOTA を達成した。

これらの研究はカテゴリ間関係の利用に焦点を当てており、Paul と Frank の研究のような外部知識は利用していない。Paul と Frank の研究では ConceptNet のようなエンティティを単位とした知識グラフを利用しているものの、ATOMIC で提供されるような出来事を単位とするコモンセンス知識は利用しておらず、知る限りでは、本研究は Story commonsense データセットにおける感情推定に対するコモンセンス知識利用の初めての試みである。

3 データセットと知識リソース

3.1 Story commonsense データセット

Rashkin らは物語文中の登場人物の内面状態を推定するために、Story commonsense データセット [1] を作成した。このデータセットは5文からなる短い物語を集めた Story cloze test [13] データセットの一部に対して、一文ごと・登場人物ごとに心的状態をアノートしている。心的状態は基本的欲求と感情の二つの観点でラベル付けされており、それぞれ複数のラベルの付与が許容されている。これにより、多面的な心的状態が表されるが、タスクはマルチラベル分類問題となる。なお、本研究の対象である感情ラベルには Plutchik [2] の8感情 (joy, trust, fear, surprise, sadness, disgust, anger, anticipation) が用いられている。

3.2 コモンセンス知識グラフ: ATOMIC

ATOMIC [3] は、事態¹⁾間の前後・因果関係などに関する常識的知識(コモンセンス)を if-then 関係として整理した知識グラフである。特に、出来事と関係する人物の感情・意図などの心的状態の関係知識を収集している点が特徴的であり、物語文からの感情推定に有用であることが期待できる。

以下に ATOMIC における互いに関連する三つ組群の例を示す。一行目は各三つ組に共通する head ノードであり、IfEvent を表す。二行目以降の各行は、三つ組における relation エッジ, tail ノードを表しており、これらにより ThenEvent を表す。これには、状態や性質に関する記述も含まれる。

```
[PersonX repels PersonY's attack]
- [xAttr] -> [PersonX is brave]
- [xEffect] -> [PersonX's heart races]
- [xEffect] -> [PersonX gains an enemy]
-[ oEffect] -> [PersonY gets hurt]
```

表 1 に ATOMIC で使用される三つ組・エッジのタイプ・定義などを表す。最左列は、三つ組が (1) 人物の内面状態を表す IfEvent-MentalState, (2) 人物の態度や人柄を表す IfEvent-Persona, (3) IfEvent と関連する出来事を表す IfEvent-Event の3つのタイプに大分類されることを示す。感情推定に関しては、特に (1) と (2) のタイプの三つ組が有効であることが期待できる。第二列に示すエッジタイプでは、head ノー

1) いわゆる出来事(event)だけでなく、性質や状態なども含む。

ドの出来事を行う主体(PersonX)とそれ以外の客体(others)を区別している。

4 感情推定モデル

図 1 に提案する感情推定モデルを示す。対象文とそれに関連するコモンセンス知識を用いることにより感情推定を行う知識利用モデルからの尤度(logits_p)と、対象文と先行文脈から感情推定を行う文脈利用モデルの尤度(logits_s)を次式の重み付き和により統合し、最終的な推定結果を得る。

$$\text{logits} = w \text{logits}_s + (1 - w) \text{logits}_p \quad (1)$$

なお、それぞれのモデルにおけるエンコーダーは事前に別個に学習されたものを使用する。

4.1 知識利用モデル

外部知識の抽出: 感情推定に有用そうな情報を ATOMIC から抽出しモデルに入力する。抽出は物語文と三つ組の head ノードの単語マッチングによって行う。まず、物語文の単語から品詞が 'PART', 'PUNCT', 'CCONJ', 'DET' のいずれかであるものをストップワードとして削除し、残された単語と ATOMIC 中の IfEvent との重複単語数が閾値以上の三つ組を一定個数取得する。今回は、三つ組の量と物語との合致度に関する定性分析から閾値を3とした。実験においては、モデルへ入力する知識を表 1 に示した ATOMIC の三つ組タイプが IfEvent-Persona, IfEvent-MentalState のものに限定する場合と全てを用いる場合を比較する。

対象文のエンコーディング: 以下の入力形式により感情推定の対象文を BERT [14] によりエンコードし、[CLS] ベクトルを分散表現として利用する。

```
[CLS]{人物}#{対象文}[SEP]
```

外部知識の導入: 対象文と同様に BERT を用いたエンコーディングを行うため、抽出した三つ組をエッジを表 1 最右列に示す疑似自然文に置換し、三つ組の各要素 head, relation, tail は別個にエンコードする。得られた分散表現と物語文の分散表現の source-target attention を行い、物語文との関連度で重み付けされた三つ組分散表現に更新する。Attention で重み付けされた三つ組の分散表現と対象文の分散表現を連結したのち、全結合層と softmax 関数に通し、8つの各感情クラスの有無を示す尤度を求める。この尤度による識別結果と正例との損失を binary cross-entropy loss によって計算し、パラメータ更新

表1 ATOMICのエッジと自然文への書き換え表現一覧

三つ組タイプ	エッジタイプ (ラベル)	三つ組の定義・説明	自然文変換時の表現
IfEvent -MentalState	X intent (xIntent)	主体が IfEvent を起こした意図はなにか？	Because X wanted
	X reaction (xReact)	IfEvent の後、主体はどう感じるか？	Then PersonX feels
	Other reaction (oReact)	IfEvent の後、客体はどう感じるか？	Then others feel
IfEvent -Persona	X attribute (xAttr)	主体はどのような態度に見られるか？	PersonX is seen as
IfEvent -Event	X need (xNeed)	主体が IfEvent を起こす前に必要な出来事はなにか？	Before, PersonX needed
	Effect on X (xEffect)	IfEvent は主体にどのような影響を与えるか？	Has an effect on PersonX
	X want (xWant)	IfEvent の後、主体は何をしたくなるか？	As a result, PersonX wants
	Other want (oWant)	IfEvent の後、客体は何をしたくなるか？	As a result, others want
	Effect on other (oEffect)	IfEvent は客体にどのような影響を与えるか？	Has an effect on others

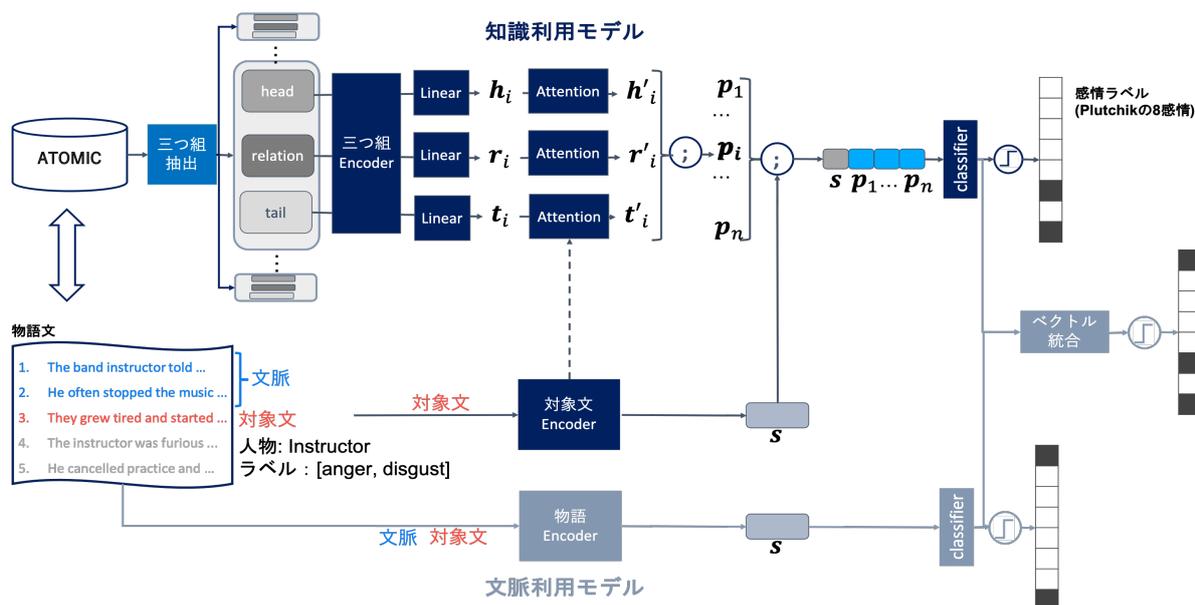


図1 提案する感情推定モデル

を行う。

4.2 文脈利用モデル

登場人物の感情は、物語中の一文のみだけでなく複数文に渡る出来事や状況の連鎖によって生じうる。このような場合、推定モデルへ入力する対象文一文のみでは正しく推定することは難しい。そこで、対象文より前に何が起きたか、そもそもどのような状況にあるのかといった文脈を入力に加える。

Rashkin ら [1] は、先行文脈において推定対象の登場人物が現れる文に限定したが、本研究では複数人物間のインタラクションが、対象の人物が陽に現れない文に記述される可能性を考慮し、対象文に先行する文を文脈として利用する [11]。

BERT への入力形式は、以下のように対象文と文脈を記号#により区分する形式とした。

[CLS]{人物}#{文脈}#{対象文}[SEP]

なお、後述の実験においては、比較のために、導入する文脈を空として文脈モデルを利用する設定での評価(表2の#7)も行っている。

5 実験

5.1 実験設定

Story commonsense データセットの train セットには正例ラベルが付与されていないため、dev セットを学習データとして使用し、test セットを用いて評価を行った。データ量(物語文・登場人物対)は、学習データ 13,004 対、テストデータ 11,859 対である。

物語と三つ組のエンコードには、事前学習済みの BERT のモデル bert-base uncased を使用し、学習データで fine-tune した。物語文の最大入力長は 20

表2 マルチラベル感情分類への外部知識と文脈の利用の効果

#	モデルモニック	知識利用モデル	文脈利用モデル	P	R	F1
1	Gaonkar-Semi_supervision			57.94	76.35	65.88
2	Gaonkar-Learned_correlations			56.50	71.47	63.11
3	Ensemble (提案手法)	onlyEmo	✓	64.40	56.36	60.11
4	Sent+ATOMIC(Emo)	onlyEmo		61.79	54.47	57.90
5	Sent+ATOMIC(All)	All		59.23	53.47	56.20
6	Sent+Cxt		✓	62.51	56.64	59.43
7	Sent		†(文脈不使用)	60.21	54.59	57.26

単語とし、先行文による文脈を用いる場合は80単語に制限した。BERTから得る[CLS]分散表現の次元サイズは765とした。学習における最適化にはAdam [15]を用い、エポック数は32、バッチサイズは32とした。ATOMICから抽出する三つ組の最大個数は40個とした。式1において、知識利用モデルと文脈利用モデルの影響力を調節する重み w は、予備実験から $w = 0.7$ とした。評価指標にはマイクロ平均による精度 (Precision), 再現率 (Recall), F1 値 (F1)を用いた。

5.2 実験結果

表2に各設定における推定精度を示す。知識利用モデルのカラムにおけるonlyEmoは、ATOMICの三つ組タイプをIfEvent-Persona, IfEvent-MentalStateに限定した場合、Allは限定しない場合を表す。参考のため、Gaonkarらの実験結果 [10]を#1, #2に引用している。

#1, #2と#3の比較から、提案手法は現在のSOTAを達成しているモデルよりもPが上回っているが、Rでは劣っている。これは、Gaonkarらの手法 [10]がラベル間の共起関係を強く意識した手法であることによる。

#4と#5の比較からは、利用する外部知識のタイプをあらかじめ対象タスクに応じて限定できるなら、そうすべきことが分かる。このことは、#7に対する#5の比較では外部知識を限定なしで利用するとかえって精度が下がるのに対し、#7と#4に比較では精度が向上していることから裏付けられる。

#3と#4の比較からは、外部知識だけでなく、文脈を併用することが有効であることが分かる。事前実験として知識利用モデル側で文脈を利用する設定も試したが、精度が低下したことから、外部知識と文脈は別に扱い、後にアンサンブル的に統合するのが良いと考えられる。

#3と#4の違いは、文脈利用モデルのアンサンブルの有無であり、後者ではF1が60.11ptから57.90ptへ2.21pt低下している。これに対し、#3と#6の差異は、知識利用モデル (onlyEmo条件) 使用の有無であるが、後者におけるF1の低下は、0.68ptにとどまっている。この差から、現状のモデルにおいては、文脈利用による精度向上への寄与の方が commonsense外部知識より大きい可能性が示唆される。現在の文脈の扱いは非常に単純化したものであるにも関わらずこのような傾向が見られたことから、外部知識の利用に関しては、まだ多くの改善の余地が残っているものと考えられる。

6 むすび

本研究では、物語を構成する各文における登場人物の感情を推定するというタスクにおいて、事態に関する commonsense知識 (ATOMIC) を導入することにより、感情推定の精度を向上させることを試みた。また、その際に物語テキストの文脈を利用することの効果について実験的に調査した。

Story commonsense データセットを用いた実験結果からは、上記のアンサンブルモデルが有効であること、用いる外部知識は、可能ならばタスクに適合するように事前に限定しておくことが有用であること、外部知識を用いる効用は文脈から得られる後よりも小さく、外部知識の利用法にはまだ多くの課題が残っていることなどが分かった。

今後の課題としては、外部知識からの三つ組の抽出法を改善し、利用する知識情報を増やす一方で、これにより増大するノイズを低減することが挙げられる。後者に関しては、Story commonsense データセットに収録されている基本欲求を利用し、それに基づく行動・出来事と引き起こされる登場人物の感情との相互作用や、物語の進行と感情の推移の連関関係などを扱っていく必要があると考えている。

参考文献

- [1] Hannah Rashkin, Antoine Bosselut, Maarten Sap, Kevin Knight, and Yejin Choi. Modeling naive psychology of characters in simple commonsense stories. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2289–2299, Melbourne, Australia, July 2018. Association for Computational Linguistics.
- [2] Robert Plutchik. A general psychoevolutionary theory of emotion. In Robert Plutchik and Henry Kellerman, editors, *Theories of Emotion*, pp. 3 – 33. Academic Press, 1980.
- [3] Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33, pp. 3027–3035, 2019.
- [4] Matthew Richardson, Christopher J.C. Burges, and Erin Renshaw. MCTest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 193–203, Seattle, Washington, USA, October 2013. Association for Computational Linguistics.
- [5] Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 839–849, San Diego, California, June 2016. Association for Computational Linguistics.
- [6] Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *AAAI spring symposium: logical formalizations of commonsense reasoning*, pp. 90–95, 2011.
- [7] Debjit Paul and Anette Frank. Ranking and selecting multi-hop knowledge paths to better predict human needs. In *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. 3671–3681, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.
- [8] Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 31, 2017.
- [9] 田辺ひかり, 小川哲司, 小林哲則, 林良彦. 感情推定における感情カテゴリに関する先験的知識の利用. 言語処理学会第 26 回年次大会発表論文集, pp. 1404–1407, 2020.
- [10] Radhika Gaonkar, Heeyoung Kwon, Mohaddeseh Bastan, Niranjana Balasubramanian, and Nathanael Chambers. Modeling label semantics for predicting emotional reactions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4687–4692, Online, July 2020. Association for Computational Linguistics.
- [11] Hikari Tanabe, Tetsuji Ogawa, Tetsunori Kobayashi, and Yoshihiko Hayashi. Exploiting narrative context and a priori knowledge of categories in textual emotion classification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 5535–5540, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.
- [12] Guoyin Wang, Chunyuan Li, Wenlin Wang, Yizhe Zhang, Dinghan Shen, Xinyuan Zhang, Ricardo Henao, and Lawrence Carin. Joint embedding of words and labels for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2321–2331, Melbourne, Australia, July 2018. Association for Computational Linguistics.
- [13] Nasrin Mostafazadeh, Michael Roth, Annie Louis, Nathanael Chambers, and James Allen. LSDSem 2017 shared task: The story cloze test. In *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*, pp. 46–51, Valencia, Spain, April 2017. Association for Computational Linguistics.
- [14] 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, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.
- [15] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.