記号的知識蒸留における敵対的学習の利用とその評価

日浦隆博 [1,2]　河野誠也 [2,1]　Angel Garcia Contreras[2]　吉野幸一郎 [3,2,1]

[1] 奈良先端科学技術大学院大学　[2] 理化学研究所ガーディアンロボットプロジェクト
[3] 東京科学大学
hiura.takahiro.hu6@naist.ac.jp
{seiya.kawano，angel.garciacontreras，koichiro.yoshino}@riken.jp

概要

大規模言語モデル（LLM）は多くの自然言語処理タスクで顕著な性能を実現しているが，知識推論のようなタスクでは依然として性能に課題があり，知識トリプルを用いた適応により知識間の関係性を明示的に学習させるアプローチが注目されている．大規模・高精度な学習用トリプルの獲得に有効なのが記号的知識蒸留である．これは LLM の持つ膨大な知識をトリプルとして出力し，知識推論モデルの学習に使用する手法である．この手法ではトリプルの一部に人手評価を行い，フィルタモデルの学習を行っている．本研究では，敵対的な学習を導入することにより，人手評価を必要としない，自動的な記号的知識蒸留を実現する手法を提案し評価する．

1　はじめに

近年，大規模言語モデル（LLM）の登場により，様々なタスクにおける精度向上が報告されている [1, 2, 3]．しかし LLM をそのまま知識推論システムとして用いるには依然として課題があり，トリプルを用いた学習が注目されている [4]．トリプルとは，二つの知識とその間の関係性のセットであり，知識グラフのエッジと関連ノードに対応する．LLM は事前学習を通して知識間の関係性を暗黙的に学習しているが，トリプルを用いた学習によりそれを明示的に学習することができ，知識推論の性能向上につながる．ここで重要なのは，より大規模で高精度な学習用トリプルの獲得であり，それを実現する手段として記号的知識蒸留 [5] が有効である．

記号的知識蒸留の全体像を図 1 に示す．記号的知識蒸留は，LLM が事前学習を通じて獲得した膨大な知識をトリプルとして抽出し，推論モデルの学習に使用する手法である．記号的知識蒸留では既存のトリプルデータセット（ゴールドデータ）を LLM に与え，同じドメイン・形式の擬似的なトリプルデータ（シルバーデータ）を出力している．しかし LLM の出力には，意味的・文法的に誤ったデータが多く含まれており，学習データの低品質化に繋がるため，それに対するフィルタモデルが必要である．従来はシルバーデータの一部に正誤のラベル付けを人手で行い，教師あり学習によりフィルタを獲得していたため，アノテーションコストがかかるという問題があった．そこで本研究では，シルバーデータから意味的・文法的に正しいデータを選択する Selector モデル（フィルタの役割）と，Selector から選択されたデータを評価する Discriminatoir モデルを敵対的に学習することにより，シルバーデータに対する自動的なフィルタリングを行う手法を提案する．

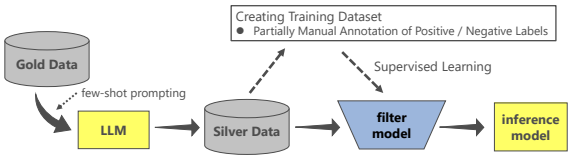

図 1　記号的知識蒸留の流れ

2　提案手法

本研究では敵対的学習を導入し，シルバーデータに対する人手でのアノテーションを必要としない，自動的な記号的知識蒸留の手法を提案する．

図 2 に敵対的学習の概要を示す．用意するモデルは，Selector と Discriminator である．Selector はシルバーデータから高品質なデータを選択するモデル，Discriminator は入力データがゴールドデータかシルバーデータかを識別するモデルである．敵対的学習を通して，Selector は Discriminator を騙すような選択方針の学習を，Discriminator は Selector によって選択されたデータが偽物（シルバーデータ）であ

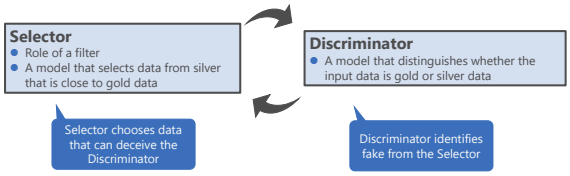

図 2　敵対的学習の概要

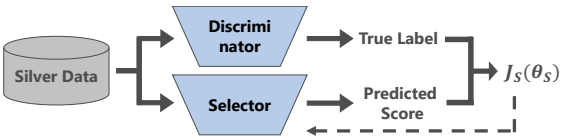

図 3　Selector の学習方法

ると識別する性能を向上させることで，最終的に Selector がより高品質なデータの選択を学習できるようになるイメージである．

また Selector と Discriminator は，BERT モデルに二層の線形層を連結することで構築した．

2.1　Selector

学習の流れを図 3 に示す．学習目標は Discriminator を騙せるデータの選択であるため，Discriminator の予測結果を正解ラベルとして，クロスエントロピー誤差 (CE) の最小化を通して学習を行う．

2.2　Discriminator

学習の流れを図 4 に示す．ゴールドデータを正例，シルバーデータを負例として学習を行う．学習に使用する負例は Selector が選択したデータである．敵対的な学習であるため，Discriminator の学習に使用する負例は，学習が進むにつれて徐々に正例に近づいていくことが望ましい．そこでサンプリング対象のデータを決定する閾値を用意し，Selector の出力がそれ以上のデータからサンプリングすることにし，その閾値を学習の経過に伴って増加させることにする．閾値の推移は図 5 の通りである．

Discriminator は式 1 の最小化で学習する．

$$J_D(\theta_D) = 1.0 \times \mathrm{CE}(\mathcal{D}_{\mathrm{train}}) + \lambda_{hinge} \times \mathrm{Hinge}(\mathcal{D}_{\mathrm{train}}) \quad (1)$$

$\mathcal{D}_{\mathrm{train}}$ は学習データセットを表している．CE はクロスエントロピー誤差，Hinge はヒンジロスであり，ヒンジロスは以下のように定義する．

$$\mathrm{Hinge}(\mathcal{D}_{\mathrm{train}}) = \frac{1}{|\mathcal{D}_{\mathrm{train}}|} \sum_{(x_i, y_i) \in \mathcal{D}_{\mathrm{train}}} \mathrm{hinge\_loss}(x_i, y_i) \quad (2)$$

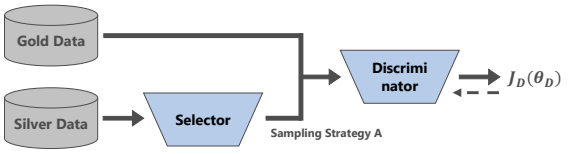

図 4　Discriminator の学習方法

$$\mathrm{hinge\_loss}(x_i, y_i) = \begin{cases} \max(0, 1 - D'(x_i)) & ,\text{if } y_i = 1 \\ \max(0, 1 + \min(D'(x_i), 0)) & ,\text{if } y_i = 0 \end{cases} \quad (3)$$

x_i はデータ，y_i はそれに対応する正解ラベルである．$D(x_i)$ はデータ x_i に対する Discriminator の出力，$D'(x_i)$ はロジットである．

通常の敵対的学習の場合，学習が進むと Discriminator の出力は 0.5 に収束するが，その場合，Selector の学習における正解ラベルが不安定になってしまう．そのためクロスエントロピーに加えてマージンを最大にするためのヒンジロスを導入している．

本研究の前提として，ゴールドデータは高精度で，シルバーデータには高品質・低品質なデータが混在していることがわかっている．これを踏まえて，正例・負例それぞれでヒンジロスを定義した．正例に対するヒンジロスは通常のヒンジロスである．負例に対するヒンジロスは，誤って正例と予測したデータのロスの影響を小さくするために，0 以上の入力に対しては定数 1 を返す関数で定義した．ヒンジロスに対する重み λ_{hinge} を 0.3 とした時の，各損失関数を図 6 に示す．また λ_{hinge} の学習経過による推移を図 7 に示す．

3　実験設定

3.1　使用データ

ATOMIC$^{20}_{20}$[6] をゴールドデータとして使用し，Llama2[3] を用いてシルバーデータを生成する（温度パラメータは 1.3）．評価用データも同様に生成する．評価用データは温度パラメータが 1.3，1.0，0.8 で，それぞれにおいて，同じ head が学習用シルバーデータに含まれているかどうかで 2 パターンずつ生成した．含まれているものを seen データ，含まれていないものを unseen データと呼ぶ．これは学習の過学習などを調べるためのものである．

ゴールドデータとシルバーデータのデータ件数は，それぞれ 16,461，148,936 であり，評価用データセットはそれぞれ 500 件ずつ生成した．各評価用データの正解ラベルの内訳は表 1 の通りである．

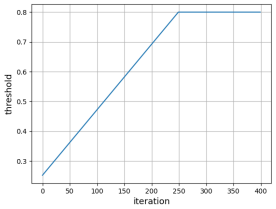

図 5 threshold

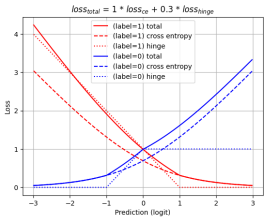

図 6 Discriminator の目的関数

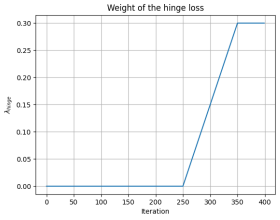

図 7 ヒンジロスの重み

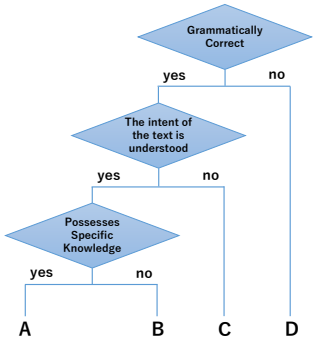

図 8 シルバーデータの負例に対する評価フロー

3.2 比較手法

Base モデル 敵対的学習モデルと比較するための Base モデルを学習する．これはシルバーデータ全体を負例，ゴールドデータを正例として，クロスエントロピー誤差により学習したモデルである．シルバーデータは高い温度パラメータで生成しているため，低品質なデータの割合が大きく，負例として機能することが期待できる．

Adv モデル Selector を Discriminator と敵対的に学習させたモデルを Adv モデルとする．また両モデルの初期状態は Base モデルとする．

4 実験結果

定量評価 図 9 に結果を示す．Adv は学習を 4 回行なった結果の平均である．accuracy は，Adv が全体的に Base を上回る結果になった．precision と recall では，precision は Adv が，recall は Base が大きくなる傾向になった．つまり Base は負例に対する識別の厳密さが低く，積極的に正例と予測する傾向であるのに対して，Adv は負例に対する識別の厳密さが増していると考えられる．これは敵対的学習における Discriminator の学習を通して，よりゴールドデータらしい事例のみを選択するように Selector が学習された結果ではないかと考える．これによる precision の増加と recall の減少はトレードオフの関係であるが，シルバーデータの生成は大規模に行えることを考えると，precision の向上は望ましい傾向である．またこれらの傾向は，seen と unseen で同じであることも確認できる．

表 1 各評価データのラベル内訳

Dataset		Label Proportion	
Data Type	Temperature	Positive	Negative
seen	0.8	224 (45.5%)	268 (54.5%)
seen	1.0	185 (37.5%)	308 (62.5%)
seen	1.3	81 (16.8%)	401 (83.2%)
unseen	0.8	231 (46.7%)	264 (53.3%)
unseen	1.0	164 (33.3%)	328 (66.7%)
unseen	1.3	64 (13.1%)	424 (86.9%)

学習されたフィルタの性質 負例に対する予測において，Base が正解したデータは，ほぼ Adv も正解するという結果になった．そこで Base が識別に失敗した負例に注目し，誤りの種類に基づいて図 8 のフローで評価することで，４つのカテゴリへの分類を行った．トリプルは head-relation-tail という構成であるが，head と tail を別々に評価し，片方が no になるとそれをそのデータ全体の評価とした．

D は文法的に誤ったデータであり，文法的誤り，文章の不成立などが含まれる．またゴールドデータと異なるフォーマットのデータに対しても D と判定した．C は意図が伝わらないデータであり，“PersonX sees gold flakes in your severely watered plants” のようなデータが挙げられる．B は具体的な知識を持たないデータであり，“PersonX practices” や “PersonX would never have to wait” といったデータがこれに該当する．A は head と tail に問題はないが，それらが relation の関係性になっていないデータである．Selector での識別を考えると，A に当てはまる負例の識別が最も難しく，A->B->C->D の順に識別が容易になると考えることができる．

評価結果を図 10 に示す．左が Adv も識別を誤った負例データ，右が Adv のみ識別に成功した負例データを表している．ここから，敵対的学習により，識別が容易な D のデータが最も多く識別できるようになったことがわかる．また識別が困難な A のデータも一定数，識別できるようになっていることがわかる．これはつまり，よりゴールドデータらし

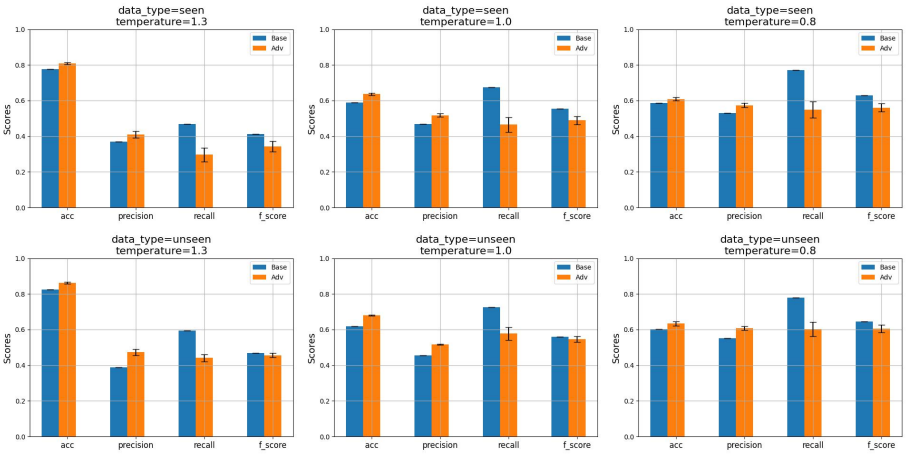

図 9　実験結果

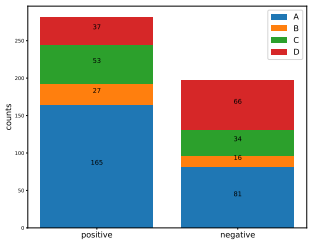

図 10　各カテゴリに含まれるデータ数

いデータの選択という観点から，ゴールドデータが持つフォーマットのような情報も考慮するフィルタが学習出来ていることが示されている．

5　関連研究

5.1　知識グラフの構築と拡張

コモンセンス知識グラフの有名な例として ConceptNet[7] や ATOMIC[8] がある．ConceptNet は単語やフレーズで表現された概念と，その関係性に関する知識グラフであり，一般常識や日常生活に関する関係性をカバーしている．ATOMIC は人間の行動などに焦点を当てた，if-then ルールに基づく知識グラフである．初期の ConceptNet[7] や ATOMIC は主にクラウドソーシングをもとに構築されたが，より大規模な知識グラフ獲得のために様々な方法が提案されてきた．例えば ConceptNet5.5[9] では，DBpedia[10] や WordNet[11] などとのリソース統合が行われた．ATOMIC$^{20}_{20}$[6] でも ConceptNet と ATOMIC との統合・拡張が行われた．TransOMCS[12] は生のテキストデータからの自動抽出による大規模知識グラフの獲得手法を提案した．近年では LLM を用いた拡張手法が注目されている．LLM は膨大な言語知識をパラメータ内部に保有していることが知られており [13, 14, 15]，この知識を活用することで，より大規模で多様な知識の獲得が期待されている．Yu らは電子商取引におけるユーザーの購入行動を LLM に与え，抽出した暗黙的な意図を知識グラフとして体系化するフレームワークを提案した [16]．west らは few-shot プロンプティングを用いて，既存の知識グラフを拡張する手法を提案した [5]．

5.2　LLM を用いた知識推論

LLM を知識ベースとして活用することは柔軟性などの点で有用であり，その可能性が模索されている [17, 18]．しかしコモンセンスなどの日常的な知識や一般的な理解はテキスト中で明言されることが少なく [19]，LLM の事前学習だけではその推論能力に限界がある．それに対して COMET[4] は，コモンセンス知識グラフのトリプルで適応された Transformer[20] ベースのモデルであり，比較的小さなモデルであっても高精度な推論が可能であることが示された．west らはトリプルを用いた適応において，学習用トリプルの規模と品質が推論精度の向上に重要であることを示した．

6　まとめ

本研究では敵対的学習を導入し，正例・負例のアノテーションなしで記号的知識蒸留を行う手法を提案した．その結果，シルバーデータ全体を負例として学習する Base と比較し，負例に対する識別が厳しくなり precision の向上が見られた．LLM を用いたシルバーデータの生成は，大規模に行うことが可能であるため，これは望ましい結果である．また負例に対する分析により，敵対的学習を通して，フォーマットに問題のあるデータなどをより適切に識別できるようになっており，また識別が難しいデータに対しても一定の割合で識別が可能になったことが分かった．この手法により，従来の記号的知識蒸留における問題点であったアノテーションコストの削減が期待できる．

謝辞

本研究は，JST ムーンショット型研究開発事業 JPMJMS2236 と JST 戦略的創造研究推進事業（ACT-X）JPMJAX22A4 の支援を受けたものです．

参考文献

[1] Jacob Devlin, et al. 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] Tom Brown, et al. Language models are few-shot learners. **Advances in neural information processing systems**, Vol. 33, pp. 1877–1901, 2020.

[3] Hugo Touvron, et al. Llama 2: Open foundation and fine-tuned chat models, 2023.

[4] Antoine Bosselut, et al. COMET: Commonsense transformers for automatic knowledge graph construction. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 4762–4779, Florence, Italy, July 2019. Association for Computational Linguistics.

[5] Peter West, et al. Symbolic knowledge distillation: from general language models to commonsense models. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, **Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 4602–4625, Seattle, United States, July 2022. Association for Computational Linguistics.

[6] Jena D Hwang, et al. (comet-) atomic 2020: On symbolic and neural commonsense knowledge graphs. In **Proceedings of the AAAI Conference on Artificial Intelligence**, Vol. 35, pp. 6384–6392, 2021.

[7] Hugo Liu and Push Singh. Conceptnet—a practical commonsense reasoning tool-kit. **BT technology journal**, Vol. 22, No. 4, pp. 211–226, 2004.

[8] Maarten Sap, et al. 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.

[9] Robyn Speer and otherse. Conceptnet 5.5: An open multilingual graph of general knowledge. In **Proceedings of the AAAI conference on artificial intelligence**, Vol. 31, 2017.

[10] Sören Auer, et al. Dbpedia: A nucleus for a web of open data. In **international semantic web conference**, pp. 722–735. Springer, 2007.

[11] Christiane Fellbaum. **WordNet: An electronic lexical database**. MIT press, 1998.

[12] Hongming Zhang, et al. Transomcs: From linguistic graphs to commonsense knowledge. **arXiv preprint arXiv:2005.00206**, 2020.

[13] Fabio Petroni, et al. Language models as knowledge bases? 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. 2463–2473, Hong Kong, China, November 2019. Association for Computational Linguistics.

[14] Zhengbao Jiang, et al. How can we know what language models know? **Transactions of the Association for Computational Linguistics**, Vol. 8, pp. 423–438, 2020.

[15] Damai Dai, et al. Knowledge neurons in pretrained transformers. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 8493–8502, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[16] Changlong Yu, et al. FolkScope: Intention knowledge graph construction for E-commerce commonsense discovery. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, **Findings of the Association for Computational Linguistics: ACL 2023**, pp. 1173–1191, Toronto, Canada, July 2023. Association for Computational Linguistics.

[17] Xiang Lorraine Li, et al. A systematic investigation of commonsense knowledge in large language models. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, **Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing**, pp. 11838–11855, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[18] Joe Davison, et al. Commonsense knowledge mining from pretrained models. 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. 1173–1178, Hong Kong, China, November 2019. Association for Computational Linguistics.

[19] Jonathan Gordon and Benjamin Van Durme. Reporting bias and knowledge acquisition. In **Proceedings of the 2013 workshop on Automated knowledge base construction**, pp. 25–30, 2013.

[20] Ashish Vaswani, et al. Attention is all you need. **Advances in neural information processing systems**, Vol. 30, , 2017.