敵対的生成ネットワークを用いた記号的知識蒸留

日浦隆博 [1,2]　河野誠也 [2,1]　Angel Garcia Contreras[2]　吉野幸一郎 [2,1]
[1] 奈良先端科学技術大学院大学　[2] 理化学研究所 GRP
hiura.takahiro.hu6@is.naist.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能力を持つものの、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、LLM が持つカバレージを維持しつつ知識推論の性能を上げる方法が模索されている。その中でも記号的知識蒸留は、大規模言語モデルの出力を知識グラフとして転移しつつ知識推論モデルの学習に用いるもので、非常に膨大な知識グラフの獲得が期待できる。しかし、LLM の出力にはしばしば意味的・文法的に誤ったデータが多数含まれ、それらを除去するフィルタが必要となることが知られている。本研究では、敵対的な学習方法を導入し、知識の生成（選択）と蒸留を同時に学習することで、追加のアノテーションを必要とせずにこうしたフィルタを実現する手法を提案する。

1　はじめに

近年，大規模言語モデル（Large Language Mode; LLM）が持つ推論能力により，様々な自然言語処理タスクにおける精度向上が報告されている [1, 2]．しかし LLM を知識推論システムとして用いようとした場合，それなりの割合で誤りを含む推論結果を出力することも知られており，知識グラフを用いた推論モデルの研究が依然として必要である [3]．

知識グラフは，事実や概念などの知識をノード（エンティティ），それらの関係性をエッジとして扱う表現方法であり，豊富な知識を体系的に保存することができる．これを用いた推論モデルでは，多段的な推論が可能であり，LLM による推論よりも高性能であることが示されている [4]．しかしこうした推論モデルの構築には大規模な知識グラフが必要である．

記号的知識蒸留 [4] は LLM を用いた知識グラフの自動構築手法である．具体的には，あらかじめ欲しいドメインのデータを用意しておき（ゴールドデータ），それらを few-shot プロンプティングを用いて LLM に与えることで，同じドメインのデータ（シルバーデータ）を LLM に出力させる．LLM が出力したシルバーデータには意味的・文法的に誤ったデータが多数含まれていることが知られており，シルバーデータの正誤を判定するフィルタを適用する．従来の記号的知識蒸留手法では，シルバーデータの一部に正誤のアノテーションを実施し，教師あり学習によりデータの蒸留を行なっていた．

しかしこの方法では，ドメインデータに対してフィルタ学習用のデータを準備する手間が掛かる．そこで本研究では，敵対的生成ネットワーク（Generative Adversarial Network; GAN）のような敵対的な学習手法により，LLM の推論候補に対する自動的な記号的知識蒸留を行う手法を提案する．実験では，京都大学の黒橋研究室が公開している事態間関係知識のデータセット [5] で実験・検証を行った．

2　関連研究

LLM を用いた知識グラフの自動構築手法は大きく分けて二つに分けられる．一つは生のテキストデータから知識関係を抽出する手法，もう一つは LLM が持つ知識を蒸留する手法である．

前者は，入力センテンスからエンティティとエッジを特定し，対応する知識グラフを抽出する手法である．一番単純な方法としては，知識グラフをトリプル (entity1-edge1-entity2 のような形式) の順列で表現することで，sequence-to-sequence の問題設定に落とし込み，その問題設定において LLM を活用する方法が挙げられる．Melnyk ら [6] は知識グラフ構築をエンティティを特定するステップと，エッジを特定するステップに分割して学習する手法を提案した．また Liu ら [7] は，同様の 2 ステップでの処理を教師なし学習で行う手法を提案した．

生のテキストから知識グラフを抽出する方法は，

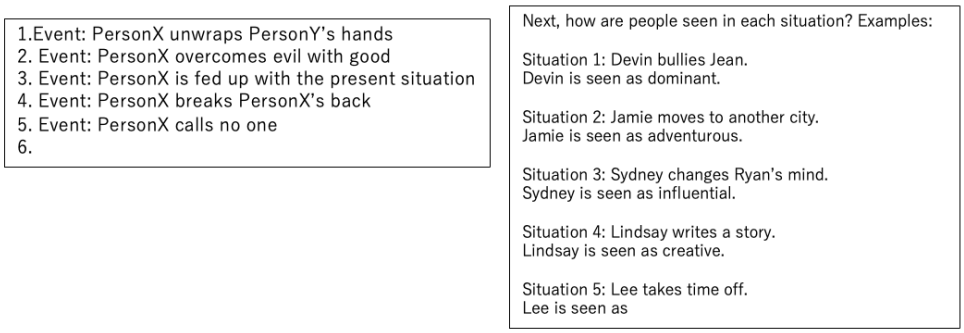

図 1　Symbolic Knowledge Distillation の例

センテンスと知識グラフの対応関係が明確であるという利点があるが，最終的な知識グラフの規模が用意したテキストのデータ数に依存するという問題点がある．また，実際には生テキストに対する自動抽出の精度には限界がある．それに比べて LLM が持つ知識を蒸留する手法は，LLM が事前学習を通じて得た膨大な知識を知識グラフの形式で抽出することができ，最終的に獲得できる知識グラフが大規模になるという利点がある．West ら [4] は few-shot プロンプティングを用いて，事前に用意した知識グラフのデータ（ゴールドデータ）を参考に，同じドメインの知識グラフデータ（シルバーデータ）を LLM から蒸留する手法を提案した（記号的知識蒸留）．図 1 に記号的知識蒸留の例を示す．左側がエンティティを LLM に生成させるためのプロンプト例である．右側は，head エンティティとエッジ関係が与えられた時の，head エンティティに対応する tail エンティティを LLM に生成させるプロンプト例である．ゴールドデータを利用してプロンプト内で例示することにより，LLM はその関係やドメインを考慮したシルバーデータを生成することができる．この head エンティティを生成するプロンプトと，tail エンティティを生成するプロンプトを組み合わせることで LLM から知識グラフを生成することができる．また，この手法により獲得したシルバーデータには，意味的・文法的に誤ったデータも多数含まれているため，それらを蒸留するためのフィルタが必要になる．West らは，シルバーデータの一部を人の手でアノテーションし，教師あり学習で蒸留フィルタの学習を行なった．

3　提案手法

本研究では記号的知識蒸留の手法を発展させ，LLM による知識の生成とフィルタによる知識の蒸留を敵対的に行う知識グラフの自動構築手法を提案する．具体的には，記号的知識蒸留におけるシルバーデータに対するフィルタを知識生成結果の選択に対して敵対的に学習することにより，LLM の学習結果から追加のアノテーションを必要とせずに記号的知識蒸留を行う．用意するモデルは，Selector と Discriminator の二つである．Selector の学習では LLM から生成された複数のシルバーデータを与え，最もゴールドデータに近そうなものを選択する．Discriminator は Selector によって選択されたシルバーデータと同じ head エンティティおよびエッジ関係を持つゴールドデータを与えられ，この識別を行う．Selector の学習の流れを図 2 に示す．敵対的学習を通して，Selector は Discriminator を騙すような選択方針を学習し，Discriminator はゴールド/シルバーデータに対する分類能力を向上させる．通常敵対的生成ネットワークでは Discriminator と Generator のパラメータを更新するが，Generator を Selector に置き換えることで LLM のパラメータを更新しない軽量な敵対的学習を実現する．

3.1　Discriminator

Discriminator は，入力データに対してゴールド/シルバーデータの分類を行う．ゴールドデータ集合からランダムにサンプリングしたデータを正例，50 個のシルバーデータを Selector に入力し，最もスコアの高かったデータを負例として学習を行う．学習ではクロスエントロピー誤差を最小化する．

$$L_D(x,\ y) = -y\log(D(x)) - (1-y)\log(1-D(x)) \quad (1)$$

x はデータ，y は対応するラベルである．x がゴールドデータの場合 $y = 1$，シルバーデータの場合 $y = 0$ である．また $D(x)$ はデータ x を Discriminator に入力した時の出力スコアである．

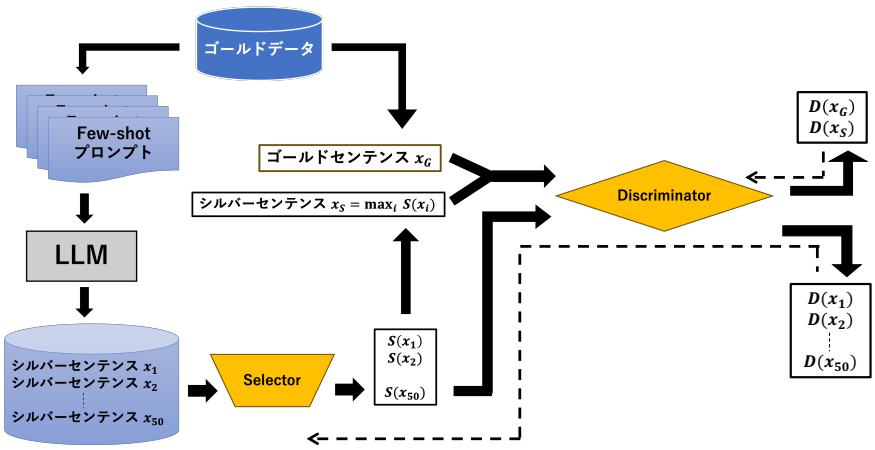

図 2　Selector の学習方法

3.2　Selector

Selector の学習では，Discriminator がゴールドデータだと判断するデータに対して高いスコアを，シルバーデータだと判断するデータに対して低いスコアを出力することを目指す．Selector の学習に以下のマージンベースのペアワイズ目的関数を用いることで，データの順序関係を学習することができる．

$$L_{pairwise}(x_{high}, x_{low}, m) = max(0, (m-(s_{high}-s_{low})) \tag{2}$$

s_{high}，s_{low} はデータ x_{high}，x_{low} に対する Selector の予測スコアであり，データ x_{high} はデータ x_{low} よりもランクが高いものとする．つまり，$D(x_{low}) < D(x_{high})$ の関係にあるデータペアである．式 (2) は，s_{high} が s_{low} よりも m 以上大きい場合 0，小さい場合は小さいほど大きな値になる．m はハイパーパラメータである．

今回は Selector への入力データを Discriminator の出力スコアにより三つのカテゴリに分類した．(a)$D(x) \leq 0.35$，(b)$0.35 < D(x) < 0.5$，(c)$0.5 \leq D(x)$ の三つである．(a) のデータ集合 X_a，(b) のデータ集合 X_b，(c) のデータ集合 X_c から一つずつ取り出したデータの組み (x_a, x_b, x_c) に対する目的関数を以下のように定義する．

$$L_S(x_a \in X_a, x_b \in X_b, x_c \in X_c) = L_{pairwise}(x_b, x_a, 8) + L_{pairwise}(x_c, x_b, 1) \tag{3}$$

Discriminator の出力を段階的に扱う理由は，学習が進むと，ゴールドデータに近いデータは 0.5 付近に収束するため，それらと $D(x) \leq 0.35$ の明らかなシルバーデータを区別して扱うためである．よって (a) と (b) の間の距離は大きく，(b) と (c) の間の距離は小さくなるように，m の値をそれぞれ 10 と 1 に設定した．また学習が進むと (a) に含まれていたデータに対する discriminator の出力が徐々に増加し (b) に含まれてしまい．学習が不安定になる問題があるため，4 iterations に 1 回の割合で discriminator の学習に使用するシルバーデータを (a) の中からサンプリングする．

4　実験と結果

実際に提案した敵対的ネットワークによる記号的知識蒸留の評価を行う．今回は京都大学の黒橋研究室が公開している事態間関係知識データセット [5] に含まれる，「原因・理由」を表す関係セットを用いて実験を行った．ただしこれらは，コーパスから自動抽出したもので誤りを含む場合があるので，そこから手作業で 739 個を抜粋したものをゴールドデータとして使用した．

Selector と Discriminator は東北大学が huggingface で公開している”cl-tohoku/bert-large-japanese-v2”[1] をベースに学習をした．蒸留候補のシルバーデータを生成する LLM モデルは，Elyza が huggingface で公開している”elyza/ELYZA-japanese-Llama-2-7b”[8, 9][2] を使用した．シルバーデータ生成の際の LLM の設定は，top_p=0.9，top_k=0 であり，各プロンプトに対して 10 個のデータを生成する．また，今回用いたプロンプトの例を図 3 に示す．また temperature は 0.9 に設定する．ただしこの状態だと学習初期の，Selector がデータを無作為に選択するような場合でも，Discriminator が分類するのが難しいデータが選

1）https://huggingface.co/cl-tohoku/bert-large-japanese-v2
2）https://huggingface.co/elyza/ELYZA-japanese-Llama-2-7b

ばれてしまう可能性が高く，学習が進まなくなる．そこで temperature=1.8 で生成したシルバーデータを用いて Discriminator のみを，30 iterations だけ事前学習した．そして，Discriminator を 1 回，Selector を 1 回学習するステップを 1 iteration とし，これを 70 iterations 繰り返すことで学習を行った．

```
事態1を原因・理由として発生する事態2を生成してください

事態1:値段が張る
事態2:迷う

事態1:ポイントが使える
事態2:購入

事態1:電球が切れる
事態2:交換

事態1:送金
事態2:報告

事態1:土が良い
事態2:育つ

事態1:肉体労働
事態2:疲れる

事態1:紛失
事態2:再購入

事態1:知見を得る
事態2:報告

事態1:効果性が抜群だ
事態2:試す

事態1:料金が変動
事態2:問い合わせる

事態1:インフルエンザが流行
事態2:
```

図 3　シルバーデータ生成のプロンプト例
head エンティティ「インフルエンザが流行」と「原因・理由」の関係にある tail エンティティを生成

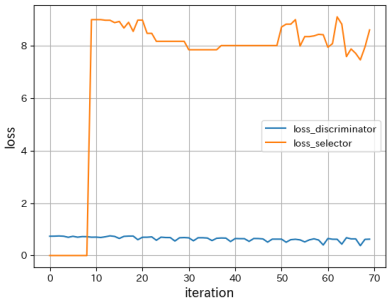

図 4　損失の推移

損失の推移を図 4 に示す．また，1，30，70 iteration 目での評価データに対する Selector の予測結果を図 5 に示す．図 5 より，学習が進むにつれて，スコアが相対的に変化し，より head エンティティ「映画館に行く」と原因・理由の関係にありそうな「泣く」「映画を見る」の tail エンティティ候補が高く評価されるようになっていることがわかる．

5　まとめ

本研究では記号的知識蒸留に対して敵対的生成ネットワークを適用し，生成結果に対する追加アノテーションなしで適切な関係知識を蒸留する手法を提案しその評価を行った．その結果，評価データにおいて学習が進むにつれてランキングの変動が見られた．しかし学習の収束が安定しないという問題点があり改善が必要である．また人手評価を通して，Selector の精度を調べる必要もある．

また今後の展望として，特に特許を対象とした知識グラフ [10] を用いた推論モデルの構築を行う予定である．

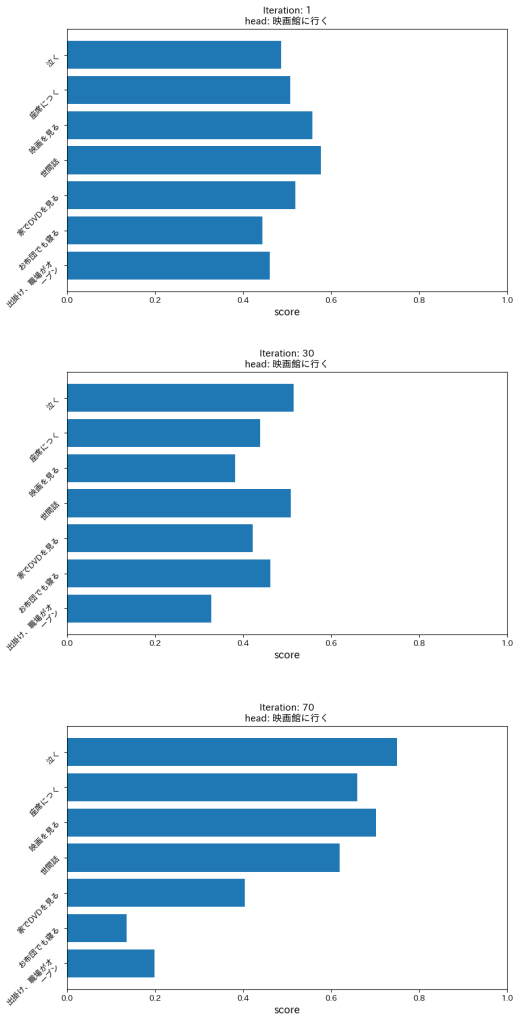

図 5　評価データに対するスコアの推移

謝辞

本研究は，JST ムーンショット型研究開発事業 JPMJMS2236，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] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. **Advances in neural information processing systems**, Vol. 33, pp. 1877–1901, 2020.

[2] 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.

[3] Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 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.

[4] Peter West, Chandra Bhagavatula, Jack Hessel, Jena Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. 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.

[5] Tomohide Shibata, Shotaro Kohama, and Sadao Kurohashi. A large scale database of strongly-related events in japanese. In **LREC**, pp. 3283–3288, 2014.

[6] Igor Melnyk, Pierre Dognin, and Payel Das. Knowledge graph generation from text. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, **Findings of the Association for Computational Linguistics: EMNLP 2022**, pp. 1610–1622, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[7] Lixian Liu, Amin Omidvar, Zongyang Ma, Ameeta Agrawal, and Aijun An. Unsupervised knowledge graph generation using semantic similarity matching. In Colin Cherry, Angela Fan, George Foster, Gholamreza (Reza) Haffari, Shahram Khadivi, Nanyun (Violet) Peng, Xiang Ren, Ehsan Shareghi, and Swabha Swayamdipta, editors, **Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing**, pp. 169–179, Hybrid, July 2022. Association for Computational Linguistics.

[8] Akira Sasaki, Masato Hirakawa, Shintaro Horie, and Tomoaki Nakamura. Elyza-japanese-llama-2-7b, 2023.

[9] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023.

[10] 日浦隆博, 吉田奈央, 松井陽子, 河野誠也, 野中尋史, 吉野幸一郎. 科学知識発見を目的とした特許のアノテーション. 言語処理学会第 30 回年次大会論文集, 2024.