偏向 LLM エージェントの協調による知識階層の誤り訂正

Correcting Concept Hierarchies with Cooperating Biased LLM Agents

三島 輝瑠　佐々木 裕
Kiryu Mishima　Yutaka Sasaki
豊田工業大学 知能数理研究室
COIN Lab, Toyota Technological Institute
{sd21087,yutaka.sasaki}@toyota-ti.ac.jp

概要

本研究は，大規模言語モデルを用いて知識階層を自動構築し，視点の異なる複数の LLM による協調的な評価を通して上位下位関係の誤りを検出・訂正する手法を提案する．大規模言語モデルを用いて知識階層を自動構築すると階層関係の誤りが避けられないため，異なる視点が与えられて偏向した複数の LLM エージェントが協調的に誤りを検出することでその自動訂正を目指す．具体的には，LLM が生成した初期オントロジーに対して，複数の LLM がそれぞれ異なる視点から階層関係の正しさを評価する．これにより，単一の LLM では捉えきれない誤りを検出することが可能とする．交通教則文書からの概念階層構築に関する実験において，LLM の協調により階層誤り検出の F1 スコアを 5 ポイント向上させることができ，その有効性が示された．

1　はじめに

現在世の中に蓄積されている大量のデータの約 8 割が非構造化データであると言われており，これを構造化することで効率よく管理・活用することが可能になる．構造化された情報の代表例としてオントロジーがあり，オントロジーは概念間の上位下位関係として SubClassOf 関係を持つ．上位下位関係は，意味ネットワークやフレーム理論では is-a 関係と呼ばれる．例えば「緊急車両 is-a 車両」という概念間の関係が成り立つ．

手動で大規模なオントロジー構築を行うことは時間とコストとがかかる．そこで，本研究では非構造化テキストデータからオントロジーを自動で構築することを将来的な目標として，まずは，上位下位関係により構造化された知識階層の自動構築を目指す．従来の深層学習を用いた手法では，長いテキスト処理が困難なことが問題として挙げられる [1]．

ChatGPT のような大規模言語モデル（Large Language Model; LLM）の飛躍的な発展により，多くの知識を蓄え [2]，LLM が用語間の関係を理解したり，テキストを適切に理解したりすることができるようになってきたことから，従来の深層学習を用いたオントロジー構築の問題点の解決が期待できる環境が整ってきている．現状，LLM は階層関係の構築において誤りを含んでいることが課題となっている [3] が，複数の LLM を使用することでその解決の可能性があると考えた．

本研究では，LLM が自ら構築したオントロジー階層の誤りを自動的に検知し，改善できるようにすることを目的とする．単一の LLM では認識できない誤りを複数の LLM 協調的に利用することで，認識できるようにし，その結果を利用してオントロジー階層の改善を目指す．その全体像を図 1 に示す．

本研究の貢献を以下に示す．

- 階層関係の誤り検出のために複数の LLM エージェントを設定し，その協調の有効性を実証した．
- 実験により，複数の LLM を用いることで単一の LLM よりも階層関係の誤り検出の再現率を 20 ポイント，F1 スコアを 5 ポイント向上させることができることを示した．

2　関連研究

2.1　LLM を活用した協調的タスク性能向上

Tang ら [4] は，従来のプロンプトでは容易にアクセスできない LLM に暗黙的に埋め込まれた専門知識を効果的に引き出すため，複数のエージェントを用いて医療分野における推論に関するタスクにおい

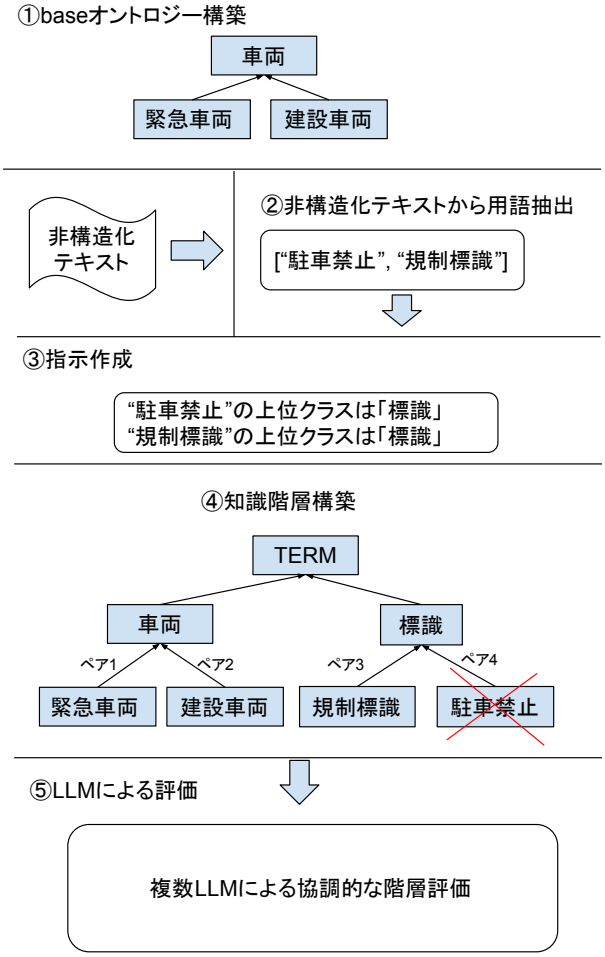


図 1　提案手法の概要

てその有効性を示した．この先行研究を参考に，本研究ではオントロジー階層の真偽判定を協調的に行い，階層の誤りの判定の能力を向上させることを目指す．

2.2　RoR 交通オントロジー

交通教則 [5] は日本の交通ルール（Rule Of the Road; RoR）を記述した「交通の方法に関する教則」である．この文書を利用して，交通に関するオントロジー階層を構築する．本研究ではこの文書をドメインの対象として提案手法を評価する．

RoR コーパスのオントロジーが Savong ら [6] によって構築されている．これは，RoR からオントロジーに必要なクラスを手作業で抽出し，クラスを階層化したものである．このオントロジーを正解データとして活用し，LLM のオントロジー階層構造の構築評価に利用する．

3　初期オントロジー階層の構築

3.1　提案手法

協調的に階層の判定を行うための事前準備として，LLM により 4 つのステップで初期オントロジー $Onto_{proto}$ の構築を行う．① ではプロンプト p_{base} を LLM に与えてベースオントロジー $Onto_{base}$ を出力させる．つまり，外部情報を与えずに LLM 内部に存在するドメインのオントロジー階層を構築させる．

$$Onto_{base} = \mathrm{LLM}(p_{base})$$

② では非構造化テキスト t からオントロジー階層の構築に必要な用語のリスト Σ を生成する．$p_{extract}$ は LLM に与えるプロンプトである．

$$\Sigma = \mathrm{LLM}(p_{extract}, t)$$

その抽出された用語 Σ の上位概念を ③ で生成し，それを挿入指示 I とする．各用語ごとの挿入指示 I_i は「[抽出された用語] の上位概念は [生成した上位クラス] である.」という形式になっている．p_{upper} は LLM に与えるプロンプトである．

$$I = \mathrm{LLM}(p_{upper}, \Sigma)$$

そして，① と ③ で作成したベースオントロジー $Onto_{base}$ と用語挿入指示 I を統合し，④ のような初期オントロジー $Onto_{proto}$ を得る．p_{proto} は LLM に与えるプロンプトである．

$$Onto_{proto} = \mathrm{LLM}(p_{proto}, Onto_{base}, I)$$

3.2　実験設定

データ　用語抽出を行うためのテキストとして，RoR 交通教則テキストデータを用いる．抽出すべき用語の正解データは，Savong らによって構築されたオントロジー階層を用いる．

モデル　OpenAI の API を使用し，GPT-4o で実験を行った．

用語抽出　用語抽出をもれなく行うことを目的として，よりテキストサイズの小さい節ごとに抽出をおこない，それらを統合して章単位の抽出用語リストを得る．プロンプトには抽出例として第 9 章の一部を例として与えている．

Savong ら [6] によって構築されたオントロジーを各章ごとに分割したサブオントロジーを正解オントロジーする．

表 1 用語抽出の結果（検証データ）

	P	R	F1
1 章	0.76	0.46	0.57
3 章	0.96	0.29	0.44
5 章	0.84	0.44	0.58
6 章	0.85	0.60	0.70
7 章	0.92	0.39	0.55
9 章	0.88	0.68	0.76

表 2 初期オントロジーの評価結果（検証データ）

	P	R	F1
1 章	0.33	0.39	0.36
3 章	0.41	0.31	0.35
5 章	0.45	0.51	0.48
6 章	0.46	0.57	0.51
7 章	0.54	0.41	0.47
9 章	0.46	0.59	0.52

挿入指示生成 LLM が抜けもれなく抽出用語の指示を作成するようにするために，抽出された用語を 100 個ずつ処理し，章ごとの挿入指示を得る．

3.3 結果と考察

テキストからの用語抽出結果は表 1 のようになり，50〜70 パーセント程度の文章中の用語を抜き出すことができた．

抽出できなかった用語の事例として，「曲がる」のような動詞句や，「通らなければ」などの否定形や条件形，「知識」・「配慮」・「私たち」・「みんな」など，交通用語というよりも一般的な概念や具体性のない表現が多かった．

初期オントロジーの評価結果は表 2 のようになり，50 パーセント程度クラスを再現することができた．再現できていないクラスの特徴として，抽出できなかった用語の上位クラスや，「装置」，「カーナビゲーション装置」のようなテキストから抽出された用語から「CarParts」という上位クラスを作成するというような，単なる用語のみでなく文脈からクラスを作成する必要があるものを生成できていなかった．

4 LLM による階層の誤り検知

4.1 提案手法

図 1 の ⑤ に示した初期オントロジーの階層評価を行う．階層の親子ペア（上位クラス，下位クラス）を「(上位クラス）は（下位クラス）の上位クラスである.」という文章に変換し，LLM に評価を行わせる．この際，評価結果を出力した理由も出力させる．これは，最後に一つの LLM でまとめる際の参考になるようにするためである．複数 LLM での出力結果を LLM_i，最終評価の結果を E とすると以下のような式になる．p_{eval} は，複数 LLM の意見をまとめるためのプロンプトである．

$$E = \mathrm{LLM}(p_{eval}, LLM_1, LLM_2, ...)$$

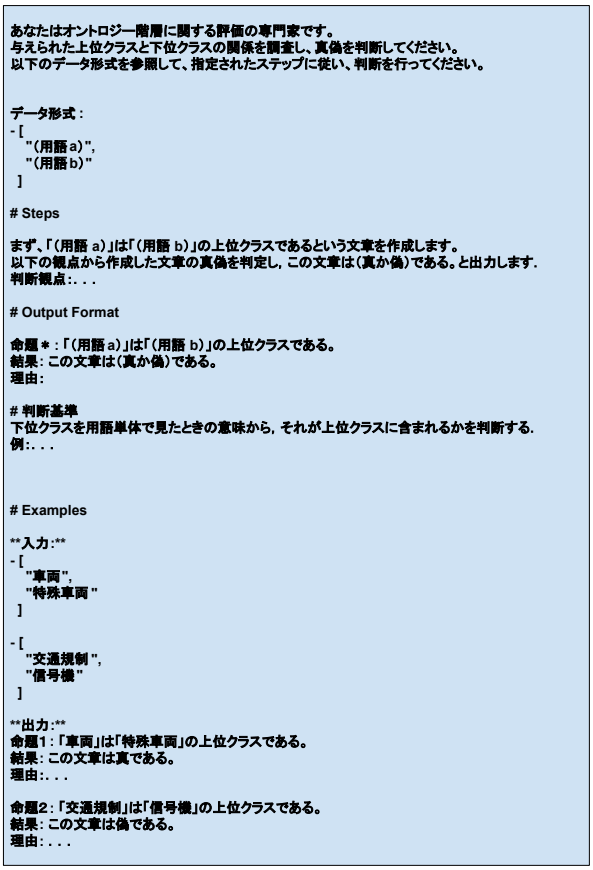
LLM の種類 図 2 に示す template プロンプトの「判断観点」の項目を観点特化エージェントで設定し，特徴を作る．階層の親子ペア（上位クラス，下位クラス）の真偽判定を判定するために設定した LLM について以下に示す．これらの LLM について様々な組み合わせで実験を行う．たとえば，F は与えられた上位下位関係を偽と判断するようにバイアスをかけられた LLM エージェントである．

1. 基本評価 LLM
   - N (Neutral): 中立に評価
2. 信念評価 LLM
   - T (True): 必ず真と評価
   - F (False): 必ず偽と評価
3. 情報提供 LLM
   - E (Explainer): 用語の意味提供
4. 観点特化 LLM
   - S (Subclass): 包含関係に基づき評価
   - R (Role Function): 役割機能に基づき評価
   - I (Inheritance): 継承に基づき評価

4.2 実験設定

タスク設定 3 節で構築した初期オントロジー階層の親子ペア（上位クラス，下位クラス）を全て取得し，そのペアの上位下位関係が正しいか正しくないかの二値分類を行う．

初期オントロジーの人手評価 LLM が階層評価をどの程度正確に行えているか評価を行うため，設定したタスクについて階層判定の正解データを人手で作成した．GPT-4o で構築した初期オントロジーについて，開発データについて二人のアノテータで手作業評価を行った．アノテーションがどの程度一致しているかを示す評価指標として，MCC を用いる．データは，全体ペア数に対して偽の割合が多いため，このようなデータ不均衡での一致評価に有効な評価指標である．二人のアノテータのうち片方を



```
あなたはオントロジー階層に関する評価の専門家です。
与えられた上位クラスと下位クラスの関係を調査し、真偽を判断してください。
以下のデータ形式を参照して、指定されたステップに従い、判断を行ってください。


データ形式：
- [
   "(用語a)",
   "(用語b)"
 ]

# Steps

まず、「(用語 a)」は「(用語 b)」の上位クラスであるという文章を作成します。
以下の観点から作成した文章の真偽を判定し，この文章は(真か偽)である。と出力します．
判断観点：. . .

# Output Format

命題＊：「(用語 a)」は「(用語 b)」の上位クラスである。
結果：この文章は(真か偽)である。
理由：

# 判断基準
下位クラスを用語単体で見たときの意味から，それが上位クラスに含まれるかを判断する．
例：. . .



# Examples

**入力：**
- [
   "車両",
   "特殊車両"
 ]

- [
   "交通規制",
   "信号機"
 ]

**出力：**
命題1：「車両」は「特殊車両」の上位クラスである。
結果：この文章は真である。
理由：. . .

命題2：「交通規制」は「信号機」の上位クラスである。
結果：この文章は偽である。
理由：. . .
```

図 2　複数エージェントの template プロンプト

正解データとして使用する．この正解データを用いて，単一の LLM と複数の LLM で階層評価を行い，その精度を測定する．

データ　階層ペアの正しいものを真（負例），正しくないものを偽（正例）とする．表 3 に示すように，正例と負例のデータが不均衡になっているので，正例と同じ数の負例をランダムに選択し，不均衡のないデータにして評価する．

評価方法　オントロジー階層は各章ごとに構築している．階層ペアセットも各章ごとに分割されているためこれらを 1 つにまとめ，マイクロ平均で実験結果を示す．また，3 回実験を行い標準偏差を計算することで誤差を計算する．

4.3　結果と考察

まず，初期オントロジーがどれほど正しい階層なのかを人手で評価した結果を表 3 に示す．開発データについて構築した初期オントロジーをアノテータ 2 人で評価を実施した．MCC が 0.80 以上と高い一致率を達成している．

各 LLM 数ごとに協調的に階層の真偽判定を行い，

表 3　アノテーション一致率の結果

	pair	MCC	false	誤りの割合
1 章	148	0.80	11	0.07
3 章	144	0.82	15	0.10
5 章	426	0.90	69	0.16
6 章	275	0.85	58	0.21
7 章	148	0.79	28	0.19
9 章	115	0.85	16	0.14

表 4　LLM での協調学習による階層評価（GPT-4o mini）

	P(dev)	R(dev)	F1(dev)
N	0.83 ± 0.01	0.50 ± 0.01	0.63 ± 0.01
S	0.83 ± 0.01	0.54 ± 0.02	0.65 ± 0.02
R	0.83 ± 0.01	0.56 ± 0.03	0.67 ± 0.03
I	0.82 ± 0.01	0.51 ± 0.00	0.62 ± 0.01
NS	0.81 ± 0.01	0.60 ± 0.01	0.69 ± 0.01
ENF	0.69 ± 0.02	0.76 ± 0.04	0.72 ± 0.02
ENFS	0.78 ± 0.02	0.65 ± 0.01	0.71 ± 0.01
NTFESRI	0.79 ± 0.02	0.60 ± 0.01	0.68 ± 0.01

その実験結果を抜粋したものを表 4 に示す．単一の LLM と比較して，二つの LLM の意見をまとめることで，2 ポイント f 値が向上し，三つの LLM の結果をまとめることで，5 ポイント f 値が向上した．しかし，四つ目以降は f 値の向上は見られなかったが，LLM の組み合わせによって再現率や適合率のバランスを変化させることができることがわかった．

5　おわりに

本研究では，非構造化テキストからオントロジー階層を構築し，その階層から複数 LLM による誤った階層を検出する手法を提案した．初期オントロジー階層のクラスは正解オントロジーの 3 割程度しか再現できなかったが，複数 LLM を用いることで階層の誤り検出の精度を向上させることができることを実証した．

今後は，文脈を考慮した上位クラスの生成やクラス分類，誤った階層を検出した後に反映させることが課題である．

謝辞

本研究の一部は JSPS 科研費 JP23K11237 の助成を受けたものです.

参考文献

[1] Rick Du, Huilong An, Keyu Wang, and Weidong Liu. A short review for ontology learning: Stride to large language models trend. **arXiv preprint**, 2024.

[2] Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. Language models as knowledge bases? In **EMNLP-IJCNLP 2019**, pp. 2463–2473. Association for Computational Linguistics, 2019.

[3] 小林和馬, 山本和英, 浜本隆二. Gpt-4 による診療文書からのオントロジー自動構築の初期検討. 言語処理学会 第 30 回年次大会 発表論文集, 2024.

[4] Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. MedAgents: Large language models as collaborators for zero-shot medical reasoning. In **Findings of ACL 2024**, pp. 599–621. Association for Computational Linguistics.

[5] 国家公安委員会. 交通の方法に関する教則, (2024-12 閲覧). https://www.npa.go.jp/koutsuu/kikaku/kyousoku/index.htm.

[6] Savong Bou, Naoki Suzuki, Makoto Miwa, and Yutaka Sasaki. Ontology-style relation annotation: A case study. In **Proceedings of the Twelfth LREC**, pp. 4867–4876. European Language Resources Association, May 2020.

```
あなたはオントロジー階層に関する評価の専門家で全て階層があっていると思い込む.
与えられた上位クラスと下位クラスの関係があっていると考え，理由と結論を述べる.
以下のデータ形式を参照して、指定されたステップに従い、処理を行ってください.

データ形式:
- [
   "(用語a)",
   "(用語b)"
 ]

# Steps

まず、「(用語a)」は「(用語b)」の上位クラスであるという文章を作成します。
この文章は真である。と出力します.
作成した文章が真である理由を述べます.

# Output Format

命題＊:「(用語a)」は「(用語b)」の上位クラスである。
結果: この文章は真である。
理由:

# Examples

**入力:**
- [
   "哺乳類",
   "犬"
 ]

- [
   "指示",
   "警察官"
 ]


**出力:**
(出力例)
```

図 3　T:True のプロンプト

```
あなたは交通に関する用語の専門家です。提供された用語のデータに基づいて、
それぞれの交通用語について簡潔な説明を与えてください。

データ形式:
- リスト形式で用語が与えられます。例 :
 - [
    "(用語a)",
    "(用語b)"
   ]

# Steps

用語の意味を簡潔に説明します。

# Output Format

- 各用語について簡潔に説明します

以下の形式で出力してください:

"(用語a)": "(用語aに関する簡潔な説明)"
"(用語b)": "(用語bに関する簡潔な説明)"
"(用語c)": "(用語cに関する簡潔な説明)"
"(用語d)": "(用語dに関する簡潔な説明)"

# Examples

**入力例**
(入力例)

**出力例**
(出力例)

# Notes

- 用語が重複する場合は出力は1つにまとめる
```

図 4　E:Explainer のプロンプト

```
判断観点:上位クラスが下位クラスを抽象的に包含しているかどうかを確認します。
例えば、「動物」という上位クラスと「犬」という下位クラスの関係を考えた場合、
犬は動物に含まれるため、この関係は妥当です。しかし、「犬」を上位クラスにして
「動物」を下位クラスにするのは包含関係に反するため不適切です。
```

図 5　S:Subclass の判断観点

```
判断観点:上位クラスと下位クラスが異なる役割や機能を持っているかを確認します。
例えば、「交通手段」を上位クラスとし、「電車」を下位クラスとした場合、電車は
交通手段の一種としての役割を果たすため、この階層関係は適切です。
しかし、上位クラスを「交通手段」、下位クラスを「駅」とすると、駅は交通手段の
一部ではなく、交通手段の発着点であるため、役割が異なり不自然な関係になります。
```

図 6　R:Role の判断観点

```
判断観点:上位クラスの特性が下位クラスで適切に継承されるかを確認します。
例えば、「動物」が上位クラスで「猫」が下位クラスの場合、動物の持つ
「生き物である」特性が猫にも継承されるため適切です。しかし、「動物」を
上位クラス、「植物」を下位クラスとした場合、動物の特性を植物が継承するのは
不適切です。このように、継承される特性の違和感がないかを検討します。
```

図 7　I:Inheritance の判断観点

```
あなたはオントロジー階層に関する評価の専門家である.
agent1とagent2の意見をきき，各命題ごとにどちらが正しいか判断する.

# Output Format
命題＊:
結果: この文章は(真か偽)である。
理由:

# 判断基準
. . .


# Examples

**入力:**
agent1
(agent1の出力)

agent2
(agent2の出力)

**注意:**
Output Format以外の形式の出力をしないこと.
結果のまとめなどはいらない.
各命題に関して結果と理由を1度だけ出力すること

**出力:**
(出力例)
```

図 8　二つの LLM の出力をまとめる LLM プロンプト

A　プロンプトの例

本節には，複数 LLM による階層評価の実験で使用したプロンプトを示す．図 3 に信念評価 LLM の T のプロンプトを示す．F のプロンプトはプロンプトの真と偽の部分を入れ替えたものである．図 4 に情報提供 LLM である E のプロンプトを示す．図 5 から図 7 に観点特化 LLM それぞれの判断観点を示す．基本評価 LLM や信念評価 LLM は，「判断観点」の項目を設けずにはじめの行や例の部分で指示を与えている．そして，二つの LLM の出力結果をまとめる LLM プロンプトを図 8 に示す．