Vol.2024-NL-259 No.20
2024/3/11

多様なプロンプトを用いた言語モデルの多角的な知識評価

趙 信[1,a)]　吉永 直樹[2,b)]　大葉 大輔[2,c)]

概要：言語モデルが学習データから獲得した関係知識の量を評価をするために，関係知識を表現する穴埋め文（プロンプト）を用いた評価手法が広く用いられている．しかしながら，この評価手法では，評価に用いるプロンプトの言語表現によって評価結果が大きく変動することが知られており，異なる言語モデルの有する関係知識の量を適切に評価することが難しい．そこで本稿では，言語モデルの関係知識をより公平かつ多角的な観点から評価するため，多様なプロンプトに基づく評価データセット MyriadLAMA と，このデータセットを活用した関係知識評価手法 BELIEF を提案する．MyriadLAMA は，既存の関係知識評価データセットである LAMA-UHN の各プロンプトに対し，構文的・意味的に多様なプロンプトを人手と大規模言語モデル (GPT-4) を用いて半自動で大量に生成したものである．BELIEF は，この多数のプロンプトを用いることで，関係知識評価におけるプロンプトバイアスの影響を低減すると共に，知識の一貫性と信頼性の観点も含めた多角的な関係知識の評価を実現する．実験では，複数の事前学習済みモデルの関係知識評価を通して，提案手法の有効性を確認する．

1. はじめに

大規模テキストから学習した事前学習済み言語モデルは，学習過程でテキストに含まれる関係知識を，暗黙のうちに獲得，保持することから，知識ベースとしての活用が期待されており，モデルが有する関係知識を評価する研究が行われるようになっている．言語モデルの保持する関係知識を評価する手法としては，関係知識を表現する穴埋め文（プロンプト，例: John Lennon の出身国は [MASK] である）の空欄（[MASK]）のエンティティを言語モデルで予測する LAMA probe [1] が用いられる．LAMA probe では，言語モデルの予測精度をもとに，モデルが有する知識の量を評価する．

LAMA probe は言語モデルの有する知識を評価する手段として有用ではあるが, 一方で, 単一のプロンプトのみで関係知識の有無を評価すると，その結果はプロンプトの言語表現の些細な違いに強い影響を受けてしまう [2], [3], [4]. 実際に言語モデルを知識ベースとして活用する際には，多様なユーザのクエリに対して知識を返す必要があり，単一のプロンプトに基づく予測精度で知識の量を評価することにはリスクがある．実際，各関係知識ごとに複数のプロンプトを用意した評価データセットを用いた評価により，予測精度が大きく変動することが報告されており [5], [6]，言語モデルの知識ベースとしての有用性をより適切に評価する手法を確立することが期待されている．

本研究では，単一プロンプトに依存した関係知識評価では言語モデルが有する関係知識を適切に評価することが困難であることを踏まえ，多種多様なプロンプトを用いた言語モデルの関係知識評価手法 BELIEF と，その評価のためのデータセット MyriadLAMA を提案する．まず，提案する評価手法のために，意味的にも構文的にも多様なプロンプトを用いて既存の LAMA-UHN データセット [7] のプロンプトを拡張した MyriadLAMA を構築する．MyriadLAMA は，関係知識のタイプごとに人手で用意した少数の構文的・意味的に多様なプロンプトをもとに，大規模言語モデル (LLM, 具体的には GPT-4) を用いて大量のプロンプトを自動生成することで，半自動で多種多様なプロンプトを構築する．この MyriadLAMA を活用し，BELIEF では多様なプロンプトに対するモデルの出力分布を統合して，個別のプロンプトにおける結果の揺らぎを均一化した予測精度の評価を行い，さらにはモデルが有する関係知識の信頼性と頑健性の評価を行う．

実験では，異なるテキスト集合，また損失関数を用いて学習された複数の事前学習済みモデル BERT に提案手法 BELIEF を適用し，その関係知識評価を通して BELIEF の有効性を評価した．その結果，多数のプロンプトを用いることで，単一プロンプトにおける評価と比較して予測精度の評価を適切に行えることを確認した．また，言語モデ

1　東京大学大学院 情報理工学系研究科
2　東京大学 生産技術研究所
a)　xzhao@tkl.iis.u-tokyo.ac.jp
b)　ynaga@iis.u-tokyo.ac.jp
c)　oba@tkl.iis.u-tokyo.ac.jp

ルが有する知識の頑健性，信頼性の評価を行い，予測精度が示唆する知識の量とは異なる観点でモデルの知識を評価することが可能であることも確認した．

2. MyriadLAMA の構築

本節では，本研究で構築した多様なプロンプトに基づく関係知識評価データセット MyriadLAMA について説明する．本研究で提案する言語モデルの関係知識評価手法では，構文的・意味的に多様なプロンプトを用いてモデルが保持する関係知識の評価を行い，その結果を統合することで，個々のプロンプトのバイアスの影響を抑えて知識の量（精度），頑健性，信頼性を評価する．これまでもプロンプトを多様化したデータセットは複数提案されているものの，それらは関係の言語表現の言い換えに焦点を当てたもので，関係知識に含まれるエンティティの表現の多様性が捉えられていない．さらに，関係の言語表現についても，提案手法で用いるには多様性が不十分であるため，新しい評価データセットを構築することとした．

本研究では，具体的に既存の関係知識評価データセット LAMA-UHN [7] を拡張して MyriadLAMA を構築した．LAMA-UHN は，Wikipedia に含まれる個別の関係知識に対応する単一プロンプト[*1]から構成されており，各関係知識は ⟨ 主体, 関係, 対象 ⟩（例: ⟨ 東京, 首都, 日本 ⟩）の 3 つ組（以下，**知識トリプル**）で構成されている．“関係” ごとに単一のテンプレート表現（以下，**関係テンプレート**，例: [X] は [Y] の首都である）が用意されている．LAMA-UHN を用いた関係知識評価の基本手順は，まず評価対象の知識トリプルを用いてテンプレートを埋め，[Y] を [MASK] トークンに置き換えて**マスクプロンプト**（これ以降，**プロンプト**）を生成する．次に，評価対象の言語モデルにプロンプトを入力して “対象”を正しく推測できるかを確認する．

MyriadLAMA では，“関係” に対して多様なテンプレート表現を提供するのみならず，“主体” や “対象” の言語表現も拡張する．具体的には，表層的な言語表現の揺らぎを排して互いに独立な関係知識を考え，この関係知識に含まれるエンティティ（“主体” と “対象”）と関係の言語表現を言い換えにより多様化することで，多様なプロンプトを生成する．このとき，表層の揺らぎを縮退した知識トリプルを**固有トリプル**，各固有トリプルに含まれるエンティティと関係表現を具体化した知識トリプルを**派生トリプル**と呼び，区別することとする．例えば，固有トリプル ⟨E_{John Lennon}, R_{born-in}, E_{United Kingdom}⟩ は，エンティティと関係に具体的な言語表現を与えることで，複数の派生トリプル (⟨John Lennon, born in, UK⟩, ⟨John Lennon, birthplace, United Kingdom⟩) などに対応する．派生トリプルを使用して一つプロンプトを生成できる（例: John Lennon was born in [MASK]）．以下で，拡張方法の概要を説明する．詳しい拡張方法は付録 A.1.1 を参照されたい．

表 1: LAMA-UHN と MyriadLAMA の統計情報．

	LAMA-UHM	MyriadLAMA
関係テンプレート数	41	4100
固有トリプル数	27,106	34,048
派生トリプル数	27,106	21,140,500
主体-関係ペア数	24,643	24,643
プロンプト数	24,643	6,492,800

エンティティの拡張． LAMA-UHN に含まれる知識トリプルは Wikipedia に基づく知識ベース T-REx [8] のサブセットであり，その “主体-関係” ペアの対象の種類は限定されている．これに対して，MyriadLAMA では，“主体-関係” をキーとして T-REx の知識ベースを検索し，他の許容される “対象” を包含するように “対象” を拡張した．例えば，John Lennon が演奏できる楽器について，E_{guitar} のみを LAMA-UHN に含まれている場合，E_{piano}も含むように固有トリプルを拡張する．また，Wikidata[*2]に含まれるエイリアスを使用して，主体および対象の表現も拡張した．例えば，E_{United Kingdom}は複数の表現（United Kingdom, UK, Britain）として表せる．

関係テンプレートの言い換え． 既存研究では，関係テンプレートと意味的に等価な言い換え表現を人手 [6] や関係抽出手法を利用 [5] して少数生成しているが，本研究では，含意を伴う表現や異なる構文（例：陳述，質問-回答）を含む，より多様的な関係テンプレートを人手で生成した上で，さらに得られた言語表現の言い換えを LLM を用いて自動生成した．具体的な手順は以下の通りである．まず各 “関係” に対して 5 つの意味的・構文的に異なる関係テンプレートを作成し，次に GPT-4 API[*3]を使用して各テンプレートの言い換えをそれぞれ 20 個生成した．全てのテンプレートの妥当性を人手で確認し，結果として，41 の関係に対して計 4100 件のテンプレートを作成した．

既存研究 [9] において，言語モデルの関係知識評価の性能が予測する MASK トークン数に大きく依存するという報告があることを考慮し，本研究では，単一トークンの予測に焦点を当てる．本研究では，実際に評価で利用した BERT [10] とその亜種が使用する WordPiece トークナイザにより，単一トークンとなる “対象” を予測対象とした派

[*1] LAMA-UHN は LAMA データセット [1] から対象が主体に文字列として包含されるプロンプト（例: Apple Watch is a product of [MASK].）や人名の主体から母国語を対象として予測するプロンプト（例: The native language of Jean Marais is [MASK].）のような知識が必ずしも必要ないプロンプトを削除したもので，関係知識の評価を行う上で LAMA より適切なデータセットとなっている．

[*2] https://www.wikidata.org/wiki/Wikidata:Data_access

[*3] OpenAI: gpt-4-1106-preview

生トリプルのみをプロンプトの生成に用いた．

表 1 に，LAMA-UHN と MyriadLAMA について，“主体-関係” ペア数，プロンプト数，関係テンプレート数と各種トリプル数をそれぞれ示す．まず，T-REx を用いて “対象” エンティティを拡張したことで，固有トリプルの数は 27,106 から 34,048 まで増加した．さらに，“主体”, “対象” エンティティの同義表現と “関係” に対して半自動生成した多様な関係テンプレートを組み合わせることで，派生トリプルの数は LAMA-UHN の 27,106 から 21,140,500 と，約 778 倍増加した．プロンプトは派生トリプルから生成するであるものの，“対象” の言語表現を含まないため，プロンプトの数は派生トリプルより少なくなるが，プロンプトの数は 27,106 から 6,492,800 と 239 倍に増加した．拡張したプロンプトの質の評価については，付録 A.1.2 に記載した関係知識評価の結果を参照されたい．

3. BELIEF: 多様なプロンプトに基づく言語モデルの多角的な関係知識評価

本節では，PLM が有する知識をより適切に評価するため，多様なプロンプトを用いた関係知識評価手法 BELIEF を提案する．従来の単一プロンプトに基づく関係知識評価はプロンプトの言語表現に強い影響を受ける [2], [3], [4]．そこで，BELIEF では，MyriadLAMA（2 節）の多様なプロンプトを使用することで，個別のプロンプトのバイアスの影響を低減した関係知識評価を行うとともに，言語モデルが記憶する知識の量（精度）に加えて，モデルが提示する知識の一貫性と信頼性を考慮した評価を実現する．以下では，まず評価の前提となる情報を述べたのち（3.1 節），評価に用いる各指標の説明を行う（3.2-3.4 節）．

3.1 前提

MyriadLAMA は一対多となる関係知識や，同じ “対象” を指す多様な言語表現を含むため，同じプロンプトに対して正解となる “対象” トークンが複数存在しうる．例えば，主体 E_{John Lennon}と関係 R_{born-in}に対して，正解となるトークンには “UK” や “Britain” などが該当する．モデルが最も確からしいと出力したトークンの精度により関係知識を評価する場合は，これらのありうる正解うち，どの表現を予測した場合でも，モデルは知識を有すると評価される[*4]ことが望ましい．そこで，MyriadLAMA に含まれる “主体-関係” ペアの集合を T，特定の “主体-関係” ペア $t \in T$ に対応するプロンプトの集合を P_t，t に対応する正解の “対象” トークンの集合を C_t としたときに，i 番目のプロンプト $p_t^i \in P_t$ の推定結果の正解とする “対象” として，評価対象の PLM が最も高い出力確率を付与したトークン $a_t^i \in C_t^i$ を求め，これを評価に利用する．また，p_t^i に対する評価対象の言語モデルの予測結果は，[MASK] に対応する出力分布を $\mathcal{O}_i^t = \{(w_j, o_j) | \sum_j o_j = 1\}$ とした時，トークン $\hat{w}_i^t = \mathrm{argmax}_{w_j,(w_j,o_j)\in\mathcal{O}_i^t} o_j$ と定義する．

[*4] 一対多となる関係知識の評価をどのように行うべきかについては議論 [11] があるが，本研究では，一対一の関係知識で “対象” のエイリアス（例: United Kingdom, UK, Britain）を考慮するなど，一対多ではない関係知識も含めて統一的に扱うため，このような評価を行った．

3.2 精度とその揺らぎ

BELIEF では，与えられた “主体-関係” ペアに対応する “対象” の予測精度を評価する際，単一プロンプトではなく，多様なプロンプト (節 2) の結果を集約することで，個別のプロンプトのバイアスに強く依存しない精度を求めることができるだけでなく，精度の揺らぎを評価することもできる．なお，[MASK] トークンに対応する出力確率の Top-1 の評価だけではモデルの出力分布の部分的な側面しか評価できないため，Top-K トークンを考慮した柔軟な評価指標を提案する．

精度: 評価データセット (MyriadLAMA) に含まれるプロンプトのうち，出力確率 Top-K に正解トークンが含まれるプロンプトの割合 (Acc@K)，および Top-K トークンにおける正解トークンの平均逆順位 (MRR) を精度の評価指標として用いる:

$$\mathrm{Acc@}K = \frac{\sum_{t\in T}\sum_{i}^{|P_t|} \mathbb{1}[\mathrm{rank}(a_t^i, \mathcal{O}_i^t) \leq K]}{\sum_{t\in T}|P_t|} \quad (1)$$

$$\mathrm{MRR} = \frac{1}{\sum_{t\in T}|P_t|}\sum_{t\in T}\sum_{i}^{|P_t|}\frac{1}{\mathrm{rank}(a_t^i, \mathcal{O}_i^t)} \quad (2)$$

ここで，$\mathrm{rank}(a_t^i, \mathcal{O}_i^t)$ は正解の “対象”a_t^i の出力確率分布 $\mathcal{O}_i^t$ における順位を表し，$\mathbb{1}[x]$ は x が True なら 1 を返し，False なら 0 を返す．

精度の揺らぎ: 次に，異なるプロンプトに対する PLM の精度の揺らぎを評価する．各 “主体-関係” ペア t について，それぞれランダムに一つ対応するプロンプトを選択して精度を計算する（式 1 および式 2 において $|P_t| = 1, \forall t \in T$）．これを N 回繰り返して得られた精度 v_i の集合を $V_{\mathrm{Acc@}K}$，V_{MRR} とする（実験では $N = 5000$）．V_* に対し，範囲（最大値と最小値の差）と標準偏差を求めることで，精度の揺らぎを評価する:

$$\mathrm{range} = \max(V_*) - \min(V_*) \quad (3)$$

$$\mathrm{stdev} = \sqrt{\frac{1}{N}\sum_{v_i\in V_*}^{N}(v_i - \frac{1}{N}\sum_{v_i\in V_*}^{N} v_i)^2} \quad (4)$$

ここで V_* は $V_{\mathrm{Acc@}K}$ または V_{MRR} である．

3.3 一貫性

“主体-関係” ペア t に対し，多様なプロンプト P_t を言語モデルに入力して予測された “対象” の一貫性を評価する．

表 2: 多様なプロンプトに基づく関係知識評価手法 BELIEF による BERT とその亜種の評価結果.

PLMs	精度 (Acc@1/Acc@10/MRR) ↑		精度の揺らぎ (Acc@1/Acc@10/MRR) ↓		一貫性 ↑	信頼性 ↓
	LAMA-UHN	MyriadLAMA	range	stdev	Consist@1	Overconf@K (k=1,10)
$\text{BERT}_{\text{base}}$	.2403/.5377/.1767	.1103/.3053/.1766	.1508/.2815/.1946	.0221/.0404/.0269	.167	.220/.288
$\text{BERT}_{\text{large}}$	**.2454**/**.5509**/**.3456**	.1185/.3173/.1861	.1525/.2871/.1922	.0231/.0400/.0274	**.180**	.218/.290
BERT_{wwm}	.2448/.5248/.3380	**.1453**/**.3638**/**.2188**	**.1502**/**.2728**/**.1681**	**.0218**/**.0374**/**.0258**	.084	**.116**/**.164**

具体的には，ある “主体-関係” ペア t に対応するプロンプト p_t^i に対する評価対象の言語モデルの予測結果 $\hat{w}_i^t$ が他のプロンプト $p_t^j \in P_t (j \neq i)$ に対する予測結果 $\hat{w}_t^j$ とどれだけ一致しているかを，全ての “主体-関係” ペア $\in T$ に対して算出し平均する [5], [12]:

$$\text{Consist@1} = \frac{1}{|T|} \sum_{t \in T} \frac{\sum_{i,j:i \neq j, i,j \leq |P_t|} \mathbb{1}[\hat{w}_i^t = \hat{w}_j^t]}{\frac{1}{2}|P_t|(|P_t|-1)} \quad (5)$$

3.4 信頼性

[MASK] トークンに対して予測した “対象” の出力確率の最大値と前述の精度を比較することで，言語モデルが自身の予測をどれだけ過信しているか評価できる．ここでは，拡張した多様なプロンプト集合 (2 節) を用いることで，言語モデルの過信度をプロンプトバイアスの影響を抑えて評価する．過信度の計算は，深層学習モデルのキャリブレーション [13] に着想を得たものである．まず，各プロンプトに対して [MASK] の最大出力確率（以後，確信度とする）を算出する．次に，プロンプトを確信度降順でソートし，これを M 個のビン $(P^{(1)}, P^{(2)}, \ldots, P^{(M)})$ に分割して，各ビン i に対して，精度の平均 $\overline{\text{Acc@}K}^{(i)}$ および確信度の平均 $\overline{o_{max}}^{(i)}$ をそれぞれ求める．最後に，これらの差分を全てのビンに渡って平均することで言語モデルが “対象” を予測する際の過信度を評価する:

$$\text{Overconf@}K = \sum_{i=1}^{M} \frac{|P^{(i)}|}{M} (\overline{o_{max}}^{(i)} - \overline{\text{Acc@}K}^{(i)}) \quad (6)$$

Overconf@K の値が小さければ小さいほど，言い換えると，言語モデルの確信度と精度が近いほど，言語モデルは自身の予測を過信していないと評価でき，予測の確信度を信頼することができる．

4. 実験

本節では，MyriadLAMA を用いて言語モデルの関係知識を評価する手法 BELIEF を複数の言語モデルの評価に適用し，プロンプトが持つバイアスの影響を再確認する．また，BELIEF を用いることで，言語モデルが保持する関係知識について，その知識の量（精度）に加えて一貫性や信頼性の観点でも，よりプロンプトバイアスの影響の少ない評価が行えることを確認する．

4.1 設定

本稿では，評価対象の言語モデルとして事前学習済みモデル BERT [10] とその亜種である $\text{BERT}_{\text{base}}$ (bert-base-uncased[*5])，$\text{BERT}_{\text{large}}$ (bert-large-uncased[*6])，および BERT_{wwm} (bert-large-uncased-whole-word-masking[*7][*8]) を使用する．$\text{BERT}_{\text{large}}$ と BERT_{wwm} は，110M パラメタを持つ $\text{BERT}_{\text{base}}$ の約 3 倍の 340M パラメタを持つ．BERT_{wwm} は，$\text{BERT}_{\text{large}}$ と同数のパラメタを持つが，事前学習におけるマスキングのアプローチが異なる．具体的には，BERT_{wwm} は 1 つの単語に対応する全てのトークンを同時にマスキングするのに対し，$\text{BERT}_{\text{large}}$ および $\text{BERT}_{\text{base}}$ は一部トークンのマスクを許す．

精度の揺らぎ（節 3.2）を計算するには，各 “主体-関係” ペアごとにプロンプトを一つサンプルして精度計算を複数回（N 回）実行する必要がある．N 回の各試行では，各 “主体-関係” ペア t に対応するプロンプトをランダムに一つ選ぶ．ここで，“主体-関係” ペア t に対してプロンプトをランダムに一つ選ぶことは，言い換えると，関係に対応するテンプレートをランダムに一つ選択することである．本実験では，各関係に対するテンプレートのランダムサンプルを事前に $N = 5000$ 回[*9]試行し，一つずつ精度計算に使用した．この時，同じ関係に対しては同じサンプリングが行われることになることに注意されたい．

4.2 結果: 単一プロンプトに基づく関係知識評価の脆弱性

BELIEF を用いた言語モデルの関係知識の評価結果を表 2 に示す．表から，今回評価した BERT およびその亜種は，多様なプロンプトに対して，関係知識の予測精度は大きく揺らぎ，予測結果の “一貫性” が低く，さらに自身の性能を過信する傾向があることが確認された．以下では，

*5 https://huggingface.co/bert-base-uncased
*6 https://huggingface.co/bert-large-uncased
*7 https://huggingface.co/bert-large-uncased-whole-word-masking
*8 https://github.com/google-research/bert
*9 なお，検定におけるブートストラップ法と同様に，この試行では同じプロンプトに関する予測は複数回行う必要はなく，1 回のみ評価してその結果を用いれば良いことに注意されたい．

$\text{BERT}_{\text{large}}$ を例に，関係知識評価の傾向を分析する．

まず，精度については，単一プロンプトに基づく LAMA-UHN を基にした評価では Acc @ 1 が 0.2454 であるのに対し，多様なプロンプトを用いた BELIEF による評価では Acc @ 1 は 0.1184 となっており，低い．対応する精度の揺らぎの大きさから，単一プロンプトを用いた関係知識評価によって，知識の予測精度が高く見積もられていた可能性を示唆している．

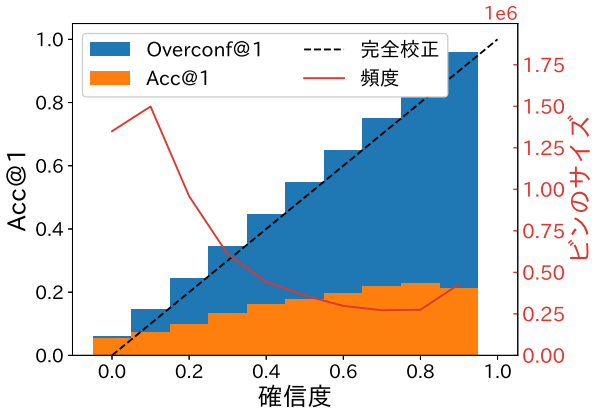
また，表 2 より，評価に用いるプロンプトによって，精度に大きな揺らぎが出うることも分かる．平均精度 (Acc@1) が 0.1184 と低い値であるのにもかかわらず，精度の最大値は 0.2072，最小値は 0.0469 と，大きく乖離していた．これは，従来の単一プロンプトに基づく評価では，選択されたプロンプトが正にも負にも評価結果に大きな影響を及ぼしていることを強調している．さらに，精度の高い標準偏差 (stdev) や一貫性 (Consist@1) の低さから，使用するプロンプトによって多様な評価結果が得られることを示しており，ここでも単一プロンプトに基づく評価の問題点を浮き彫りにしている．LAMA-UHN を用いた場合，$\text{BERT}_{\text{large}}$ は BERT_{wwm} より優れた予測精度を示しているが，BELIEF による評価ではこれが逆転している．Remiers [14] は，学習が初期値に大きく依存する深層学習モデルの比較を単一の評価値を用いて計算することのリスクを指摘し，スコアの分布に基づく比較を提案しているが，これと同様に，単一のプロンプトを用いた関係知識評価は実際のモデルの性能において，不適切な結論を導く可能性があることが確認された．

最後に，図 1a に $\text{BERT}_{\text{large}}$ の確信度 o_{max} と精度 Acc@1 の対応関係を示す．図より，$\text{BERT}_{\text{large}}$ は確信度が高いプロンプトに対しても精度は改善せず，自信の予測に対して過信していることが示された．さらに，表 2 から，K を大きくすると，より過信度 Overconf@K は悪化することも分かった．

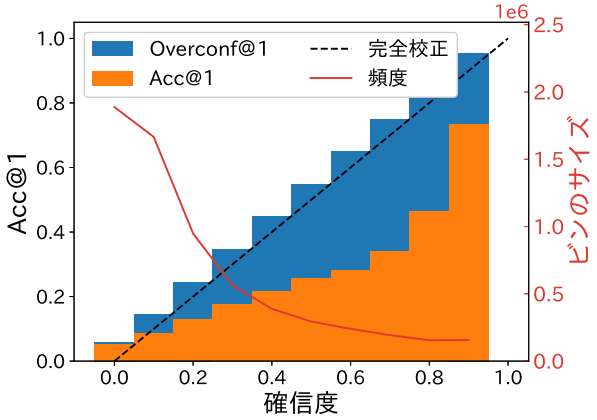
4.3 結果: BELIEF を用いた言語モデルの関係知識評価

ここでは，BELIEF の適用により得られた，バイアスの影響が少ない関係知識評価結果を異なる BERT モデル間で比較することで，BERT とその亜種が保持する関係知識を深く分析する．

表 2 より，$\text{BERT}_{\text{base}}$ に比べて $\text{BERT}_{\text{large}}$ の方が精度，一貫性と信頼性の指標において上回った．加えて，BERT_{wwm} は，一貫性以外の指標において，更なる性能改善を示した．これは，パラメタサイズの増強だけではなく，学習戦略（この場合，マスキング方法）を最適化することが事前学習における知識獲得において重要な要素であることを示唆している．ここで，平均的に最も良い性能を発揮する BERT_{wwm} は，多様なプロンプトを用いた際に，予測精度の揺らぎが小さい一方で，予測の一貫に課題が残ることが分かった．これは，予測の精度の一貫性が必ずしも両立しないことを示唆している．



(a) $\text{BERT}_{\text{large}}$



(b) BERT_{wwm}

図 1: 精度と確信度の関係．

BERT_{wwm} は，精度およびその頑健性だけでなく，信頼度の観点においても優れた能力を示した．図 1b に BERT_{wwm} の確信度 (o_{max}) と実際の精度 (Acc@1) の対応を示す．図から，$\text{BERT}_{\text{large}}$ とは対照的に，BERT_{wwm} は確信度と実際の精度との間に強い相関があることがわかる．

5. 関連研究

本節では，まず既存の言語モデルの関係知識評価手法について概観した後，それらの関係知識評価手法で用いられるプロンプトのバイアスの影響を指摘した研究について説明する．次に，本研究とも関係するプロンプトバイアスの影響を低減する関係知識評価手法について議論する．

5.1 プロンプトに基づく言語モデルの関係知識評価

Petroni らは大規模テキストから学習された言語モデルを知識ベースとして活用することを念頭に，穴埋め文（プロンプト）を用いて関係知識の評価を行う手法 LAMA probe [1] を提案した．Petroni らの研究を受けて，言語モ

デルが有する知識をより効果的に引き出すべく，プロンプトに基づく関係知識評価の精度をプロンプトの最適化により改善する試みが行われた．具体的に，Shin らは [15] 言語モデルによる知識の想起を補助するトークンをモデルの学習データから検索してプロンプトを拡張することで，関係知識の予測精度を改善する手法を提案している．また，関係知識の予測に対して最適化された，埋め込みに基づくソフトプロンプトを用いて知識の予測精度を改善する研究 [16], [17], [18] も複数提案されている．これらの研究では，主として言語モデルの知識ベースとしての利用を促進する目的で，関係知識の予測精度を高めるプロンプトを提案しているが，言語モデルが有する知識を評価するという観点では，言語モデルに最適化されたプロンプトを用いることが必ずしも言語モデルが有する知識を正しく評価することに繋がるとは限らない．

Jiang ら [6] や Elazar ら [5] は，LAMA データセットに含まれるプロンプトの言い換えを人手で作成して言語モデルの関係知識評価を行い，関係知識の予測精度がプロンプトによって大きく変動することを指摘した．Kassner ら [2] は否定表現など特定の言語表現を含むプロンプトを，また Petroni ら [3] や Ravichander らは [4] は語彙や構文を変更したプロンプトを言語モデルが適切に扱うことができない点をそれぞれ指摘している．また，プロンプトに含まれる補助的な文脈 [7] や，その順序 [19] が関係知識の予測精度に影響を与えることも報告されている．これらの研究は，言語モデルの関係知識評価を適切に行うことの難しさを示すと共に，個別の関係知識に対して単一のプロンプトを用いた言語モデルの関係知識評価の限界を示すものである．

5.2 関係知識評価におけるプロンプトバイアスの影響

プロンプトを用いた言語モデルの関係知識評価では，“主体” の部分文字列である “対象” を予測するプロンプトや，母国語を人名から予測するプロンプト，また出力分布に大きな偏りがある “関係” について，出力に対してプロンプトが有するバイアスを利用してモデルが関係知識の予測を行うことが指摘されている [20]．また，関係知識の予測精度を改善するためにプロンプトを最適化する過程で，学習・評価データ間のドメインの重なりによる出力分布の漏洩や，関係知識評価データセット自体のバイアスによって，言語モデルが有する知識が過度に見積もられるリスクも指摘されている [16], [21], [22], [23], [24].

関係知識評価において異なるプロンプトを用いた場合の出力の一貫性を評価するために, LPAQA [5] や ParaRel [6] など複数の評価データセットが構築されている．これらのデータセットでは，本研究と同様に，同一の関係知識に対して用意した複数のプロンプトに対する言語モデルの出力の揺らぎを問題とし，単一プロンプトを用いた関係知識評価の脆弱性を指摘している [5], [6], [12]．LPAQA や ParaRel が提供するプロンプトは数が少なく，表現できる多様性に限界があり，本研究で提案する MyriadLAMA のように，安定した関係知識評価を行う用途で用いることは難しい．

5.3 バイアスを除去したプロンプトに基づく関係知識評価

言語モデルが有する知識を正しく評価するために，モデルの出力に含まれるプロンプト由来のバイアスを除去する手法 [25], [26], [27] が検討されている．これらの手法では，関係知識評価時のモデルの出力分布に含まれるプロンプト由来のバイアスを検出し，分布を補正することによってバイアスを除去する．また，Yoshikawa ら [28] は言語モデルが行う予測の確信度を手がかりに，関係知識の評価を行う上でより適切なプロンプトを選別する手法を提案している．Newman ら [29] はアダプタを用いてプロンプトを埋め込みに写像することで，プロンプトに含まれる具体的な言語表現が関係知識評価に及ぼす影響を低減している．これらの手法は関係知識評価におけるプロンプトバイアスを軽減することで，言語モデルが有する知識をより正確に評価するものであり，プロンプトを多様化することでプロンプトバイアスの影響を低減する提案手法とは直交するアプローチとなっており，提案手法と組み合わせて用いることが可能である．

また，プロンプトの言い換えを入力して，プロンプトの埋め込み空間に摂動に加えて得られた複数の出力分布をアンサンブルすることで，個別のプロンプトのバイアスの影響を抑えた関係知識の評価手法 [6], [17], [30] も提案されている．これらの手法が，複数のプロンプトを併用してバイアスを軽減した予測を行うのに対し，本研究ではプロンプトごとに出力を得ることで，精度のみならず一貫性や信頼性などの多角的な評価を可能とする点で異なる．

6. 終わりに

本稿では，事前学習済み言語モデルが有する関係知識をより頑健に評価することを目的として，多様なプロンプトを有する評価データセット MyriadLAMA を構築し，このデータセットに基づく多角的な関係知識評価手法 BELIEF を提案した．我々はまず，既存の評価データセット LAMA-UHN をもとに，大規模言語モデルを併用して，評価対象の関係知識に含まれる関係とエンティティの表現を多様化した評価データセットを MyriadLAMA を半自動構築した．この MyriadLAMA を活用して，言語モデルの有する関係知識の量（精度），一貫性，信頼性を多角的に評価する手法 BELIEF を設計した．BELIEF では，多様なプロンプトを用いることで，個別のプロンプトのバイアスに強く影響されず，事前学習済みモデルが有する関係知識

の評価や比較を行うことが可能となっている．実験では，BELIEF を BERT とその亜種の評価に適用することで単一プロンプトに基づく関係知識評価の脆弱性を示しただけでなく，LAMA-UHN では見られなかった大規模言語モデル間の性能差を確認した．これにより，大規模言語モデルの有する関係知識を評価する手法として BELIEF が有効であることを示した．

MyriadLAMA は膨大なプロンプトを含むため，評価のコストが大きいという問題点が残されている．今後は，今回半自動構築した MyriadLAMA から，多様性を確保したまま頑健な関係知識評価が行えるサブセットを抽出することを検討している．また，MyriadLAMA については，準備が整い次第公開する予定である．

謝辞 本研究は東京大学生産技術研究所特別研究経費，JSPS 科研費 JP21H03494，JP21H03445，および JST，CREST，JPMJCR19A4 の支援を受けたものである．

参考文献

[1] Petroni, F., Rocktäschel, T., Riedel, S., Lewis, P., Bakhtin, A., Wu, Y. and Miller, A.: Language Models as Knowledge Bases?, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (Inui, K., Jiang, J., Ng, V. and Wan, X., eds.), Hong Kong, China, Association for Computational Linguistics, pp. 2463–2473 (online), DOI: 10.18653/v1/D19-1250 (2019).

[2] Kassner, N. and Schütze, H.: Negated and Misprimed Probes for Pretrained Language Models: Birds Can Talk, But Cannot Fly, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (Jurafsky, D., Chai, J., Schluter, N. and Tetreault, J., eds.), Online, Association for Computational Linguistics, pp. 7811–7818 (online), DOI: 10.18653/v1/2020.acl-main.698 (2020).

[3] Misra, K., Ettinger, A. and Rayz, J.: Exploring BERT's Sensitivity to Lexical Cues using Tests from Semantic Priming, *Findings of the Association for Computational Linguistics: EMNLP 2020* (Cohn, T., He, Y. and Liu, Y., eds.), Online, Association for Computational Linguistics, pp. 4625–4635 (online), DOI: 10.18653/v1/2020.findings-emnlp.415 (2020).

[4] Ravichander, A., Hovy, E., Suleman, K., Trischler, A. and Cheung, J. C. K.: On the Systematicity of Probing Contextualized Word Representations: The Case of Hypernymy in BERT, *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics* (Gurevych, I., Apidianaki, M. and Faruqui, M., eds.), Barcelona, Spain (Online), Association for Computational Linguistics, pp. 88–102 (online), available from ⟨https://aclanthology.org/2020.starsem-1.10⟩ (2020).

[5] Elazar, Y., Kassner, N., Ravfogel, S., Ravichander, A., Hovy, E., Schütze, H. and Goldberg, Y.: Measuring and Improving Consistency in Pretrained Language Models, *Transactions of the Association for Computational Linguistics*, Vol. 9, pp. 1012–1031 (online), DOI: 10.1162/tacl_a_00410 (2021).

[6] Jiang, Z., Xu, F. F., Araki, J. and Neubig, G.: How Can We Know What Language Models Know?, *Transactions of the Association for Computational Linguistics*, Vol. 8, pp. 423–438 (online), DOI: 10.1162/tacl_a_00324 (2020).

[7] Petroni, F., Lewis, P., Piktus, A., Rocktäschel, T., Wu, Y., Miller, A. H. and Riedel, S.: How Context Affects Language Models' Factual Predictions, *ArXiv*, Vol. abs/2005.04611 (online), available from ⟨https://api.semanticscholar.org/CorpusID:212411919⟩ (2020).

[8] Elsahar, H., Vougiouklis, P., Remaci, A., Gravier, C., Hare, J., Laforest, F. and Simperl, E.: T-REx: A Large Scale Alignment of Natural Language with Knowledge Base Triples, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, European Language Resources Association (ELRA), (online), available from ⟨https://aclanthology.org/L18-1544⟩ (2018).

[9] Zhao, X., Yoshinaga, N. and Oba, D.: Tracing the Roots of Facts in Multilingual Language Models: Independent, Shared, and Transferred Knowledge.

[10] Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *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)* (Burstein, J., Doran, C. and Solorio, T., eds.), Minneapolis, Minnesota, Association for Computational Linguistics, pp. 4171–4186 (online), DOI: 10.18653/v1/N19-1423 (2019).

[11] Nagasawa, H., Heinzerling, B., Kokuta, K. and Inui, K.: Can LMs Store and Retrieve 1-to-N Relational Knowledge?, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)* (Padmakumar, V., Vallejo, G. and Fu, Y., eds.), Toronto, Canada, Association for Computational Linguistics, pp. 130–138 (online), DOI: 10.18653/v1/2023.acl-srw.22 (2023).

[12] Fierro, C. and Søgaard, A.: Factual Consistency of Multilingual Pretrained Language Models, *Findings of the Association for Computational Linguistics: ACL 2022* (Muresan, S., Nakov, P. and Villavicencio, A., eds.), Dublin, Ireland, Association for Computational Linguistics, pp. 3046–3052 (online), DOI: 10.18653/v1/2022.findings-acl.240 (2022).

[13] Desai, S. and Durrett, G.: Calibration of Pretrained Transformers, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Webber, B., Cohn, T., He, Y. and Liu, Y., eds.), Online, Association for Computational Linguistics, pp. 295–302 (online), DOI: 10.18653/v1/2020.emnlp-main.21 (2020).

[14] Reimers, N. and Gurevych, I.: Reporting Score Distributions Makes a Difference: Performance Study of LSTM-networks for Sequence Tagging, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (Palmer, M., Hwa, R. and Riedel, S., eds.), Copenhagen, Denmark, Association for Computational Linguistics, pp. 338–348 (online), DOI: 10.18653/v1/D17-1035 (2017).

[15] Shin, T., Razeghi, Y., Logan IV, R. L., Wallace, E. and Singh, S.: AutoPrompt: Eliciting Knowledge from Language Models with Automatically Gen-

erated Prompts, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Webber, B., Cohn, T., He, Y. and Liu, Y., eds.), Online, Association for Computational Linguistics, pp. 4222–4235 (online), DOI: 10.18653/v1/2020.emnlp-main.346 (2020).

[16] Zhong, Z., Friedman, D. and Chen, D.: Factual Probing Is [MASK]: Learning vs. Learning to Recall, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T. and Zhou, Y., eds.), Online, Association for Computational Linguistics, pp. 5017–5033 (online), DOI: 10.18653/v1/2021.naacl-main.398 (2021).

[17] Qin, G. and Eisner, J.: Learning How to Ask: Querying LMs with Mixtures of Soft Prompts, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T. and Zhou, Y., eds.), Online, Association for Computational Linguistics, pp. 5203–5212 (online), DOI: 10.18653/v1/2021.naacl-main.410 (2021).

[18] Li, Y., Che, T., Wang, Y., Jiang, Z., Xiong, C. and Chaturvedi, S.: SPE: Symmetrical Prompt Enhancement for Fact Probing, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (Goldberg, Y., Kozareva, Z. and Zhang, Y., eds.), Abu Dhabi, United Arab Emirates, Association for Computational Linguistics, pp. 11689–11698 (online), DOI: 10.18653/v1/2022.emnlp-main.803 (2022).

[19] Lu, Y., Bartolo, M., Moore, A., Riedel, S. and Stenetorp, P.: Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Muresan, S., Nakov, P. and Villavicencio, A., eds.), Dublin, Ireland, Association for Computational Linguistics, pp. 8086–8098 (online), DOI: 10.18653/v1/2022.acl-long.556 (2022).

[20] Poerner, N., Waltinger, U. and Schütze, H.: E-BERT: Efficient-Yet-Effective Entity Embeddings for BERT, *Findings of the Association for Computational Linguistics: EMNLP 2020* (Cohn, T., He, Y. and Liu, Y., eds.), Online, Association for Computational Linguistics, pp. 803–818 (online), DOI: 10.18653/v1/2020.findings-emnlp.71 (2020).

[21] Cao, B., Lin, H., Han, X., Sun, L., Yan, L., Liao, M., Xue, T. and Xu, J.: Knowledgeable or Educated Guess? Revisiting Language Models as Knowledge Bases, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (Zong, C., Xia, F., Li, W. and Navigli, R., eds.), Online, Association for Computational Linguistics, pp. 1860–1874 (online), DOI: 10.18653/v1/2021.acl-long.146 (2021).

[22] Youssef, P., Koraş, O., Li, M., Schlötterer, J. and Seifert, C.: Give Me the Facts! A Survey on Factual Knowledge Probing in Pre-trained Language Models, *Findings of the Association for Computational Linguistics: EMNLP 2023* (Bouamor, H., Pino, J. and Bali, K., eds.), Singapore, Association for Computational Linguistics, pp. 15588–15605 (online), DOI: 10.18653/v1/2023.findings-emnlp.1043 (2023).

[23] Li, S., Li, X., Shang, L., Dong, Z., Sun, C., Liu, B., Ji, Z., Jiang, X. and Liu, Q.: How Pre-trained Language Models Capture Factual Knowledge? A Causal-Inspired Analysis, *Findings of the Association for Computational Linguistics: ACL 2022* (Muresan, S., Nakov, P. and Villavicencio, A., eds.), Dublin, Ireland, Association for Computational Linguistics, pp. 1720–1732 (online), DOI: 10.18653/v1/2022.findings-acl.136 (2022).

[24] Cao, B., Lin, H., Han, X., Liu, F. and Sun, L.: Can Prompt Probe Pretrained Language Models? Understanding the Invisible Risks from a Causal View, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Muresan, S., Nakov, P. and Villavicencio, A., eds.), Dublin, Ireland, Association for Computational Linguistics, pp. 5796–5808 (online), DOI: 10.18653/v1/2022.acl-long.398 (2022).

[25] Zhao, T., Wallace, E., Feng, S., Klein, D. and Singh, S.: Calibrate Before Use: Improving Few-Shot Performance of Language Models, *International Conference on Machine Learning*, (online), available from ⟨https://api.semanticscholar.org/CorpusID:231979430⟩ (2021).

[26] Dong, Q., Dai, D., Song, Y., Xu, J., Sui, Z. and Li, L.: Calibrating Factual Knowledge in Pretrained Language Models, *Findings of the Association for Computational Linguistics: EMNLP 2022* (Goldberg, Y., Kozareva, Z. and Zhang, Y., eds.), Abu Dhabi, United Arab Emirates, Association for Computational Linguistics, pp. 5937–5947 (online), DOI: 10.18653/v1/2022.findings-emnlp.438 (2022).

[27] Wang, Y., Lu, D., Kong, C. and Sang, J.: Towards Alleviating the Object Bias in Prompt Tuning-based Factual Knowledge Extraction, *Findings of the Association for Computational Linguistics: ACL 2023* (Rogers, A., Boyd-Graber, J. and Okazaki, N., eds.), Toronto, Canada, Association for Computational Linguistics, pp. 4420–4432 (online), DOI: 10.18653/v1/2023.findings-acl.270 (2023).

[28] Yoshikawa, H. and Okazaki, N.: Selective-LAMA: Selective Prediction for Confidence-Aware Evaluation of Language Models, *Findings of the Association for Computational Linguistics: EACL 2023* (Vlachos, A. and Augenstein, I., eds.), Dubrovnik, Croatia, Association for Computational Linguistics, pp. 2017–2028 (online), DOI: 10.18653/v1/2023.findings-eacl.150 (2023).

[29] Newman, B., Choubey, P. K. and Rajani, N.: P-Adapters: Robustly Extracting Factual Information from Language Models with Diverse Prompts, *ArXiv*, Vol. abs/2110.07280 (online), available from ⟨https://api.semanticscholar.org/CorpusID:238856973⟩ (2021).

[30] Kamoda, G., Heinzerling, B., Sakaguchi, K. and Inui, K.: Test-time Augmentation for Factual Probing, *Findings of the Association for Computational Linguistics: EMNLP 2023* (Bouamor, H., Pino, J. and Bali, K., eds.), Singapore, Association for Computational Linguistics, pp. 3650–3661 (online), DOI: 10.18653/v1/2023.findings-emnlp.236 (2023).

付　　録

A.1 MyriadLAMA の構築と評価

A.1.1 MyriadLAMA におけるトリプルの拡張方法

本付録では，MyriadLAMA において固有トリプルから派生トリプルを得る詳細な手順を説明する．2 節で述べたように，本研究ではまず・LAMA-UHN [7] に含まれる固有トリプルについて，“対象” を T-REx [5] を用いて拡張する．次に，得られた固有トリプルに対し，エンティティ (“主体” と “対象”) に紐づく具体的な言語表現と，人手と LLM を併用して多様化した関係テンプレートを組み合わせて派生トリプルを生成した．以下で，その手順を説明する．

A.1.1.1 エンティティの拡張

T-REx の検索による固有トリプルの拡張: LAMA-UHN は，T-REx [8] から派生した LAMA データセットのサブセットであり，T-REx は，Wikidata の知識トリプルと Wikipedia テキスト（記事概要）の間の大規模なアラインメントから成る，1100 万の知識トリプルを含む知識ベースである．T-REx は NoSub，AllEnt，SPO の 3 つの異なる手法を用いて知識トリプルと記事概要の間のアライメントを行っており，高い精度（それぞれ 0.98，0.96，0.88）で知識トリプルと概要の対応づけを行っている．LAMA では最も高精度の NoSub によって概要と対応づけされた知識トリプルのみを利用しているが，関係知識評価データセットの構築においては知識トリプルの正確性のみが問題となるため，MyriadLAMA ではより広範囲の知識トリプルを利用してデータセットを拡張した．具体的には，T-REx に存在する NoSub，AllEnt，SPO によって概要に対応付された全て知識トリプルを検索候補として，LAMA-UHN に存在する “主体-関係” ペアをクエリとする検索を行い，より多様な “対象” エンティティを獲得した．これにより，固有トリプル数を 27,106 から 34,048 まで増補した（表 1）．

エイリアス（同義表現）を用いたエンティティの拡張: 次に，Wikidata から得られるエンティティのエイリアスを利用して，固有トリプルの “主体” と “対象” の言語表現（とその言い換え）を獲得した．具体的には，エンティティの Wikidata 識別子[*10] と Wikidata API[*11] を使用して，エンティティの（英語の）エイリアス表現を取得した．“主体” と “対象” のエイリアスを後述する関係テンプレートと組み合わせることで，新しい派生トリプルを生成できる．N 個の “主体” と M 個の “対象” が与えられた場合，一つの関係テンプレートから生成できる派生トリプルの数は $N \times M$ である．

[*10] https://www.wikidata.org/wiki/Wikidata:Identifiers
[*11] https://www.wikidata.org/wiki/Special:EntityData/<entity_identifier>.json

A.1.1.2 関係テンプレートの多様化

関係テンプレートを多様化するに際しては，単純にその種類数を増やすだけでなく，言語表現の多様性を高めることに留意し，LAMA-UHN に含まれる関係テンプレートをもとに，人手と大規模言語モデル (LLM) を用いて 2 ステップで書き換えるアプローチを採用した．まず，元の LAMA-UHN に含まれる各関係テンプレートから人手で 5 つの関係テンプレートを作成した．次に，LLM (GPT-4) を利用し，各関係テンプレートから 19 の追加テンプレートを自動生成した．これにより，各関係につき合計 100 の関係テンプレートが得られた．以下で，それぞれのステップを説明する．

ステップ 1: 人手による関係テンプレートの書き換え. 人手による関係テンプレートの書き換えは，本論文の第一著者が実施した．関係テンプレートの書き換えでは，単純な語彙の置換に留まらず，“主体” と “対象” 間の関係を維持した上で，構文的・意味的な多様な言語表現に書き換えることが望ましい．そこで，Wikidata が提供する関係の説明[*12]を参照しつつ，意味的に等価な言い換えだけでなく，含意関係として成立する表現への書き換えも行なった．例えば，“[X] died in [Y].”[*13]という関係テンプレートについては，文構造を変更し，対象に関する情報を付加して，“[X] resided in [Y] until death.” のような関係テンプレートに書き換えている．また，語彙についても “dead/death” のような単純かつ直接的な言い換えは後述の LLM に任せる想定で，“[X] spent the last years of life in [Y].” のような含意関係にある関係テンプレートへと書き換えた．また，文型の多様性を担保すべく，“質問？応答” 形式への書き換えも行った．この書き換えでは，質問は “主体” と “関係” に関する情報を包含し，応答は該当する “対象” を指す．

これらの作業を進める過程で，LAMA-UHN の一部テンプレートが Wikidata に定義された関係の本来の意味を部分的にしか表現できておらず，特定の知識トリプルに対して不適切となっていることが判明した．例えば，“対象が主体が属する創作物や芸術作品のジャンル” を定義する関係 P136[*14]では，対象エンティティのカテゴリとして音楽，映画，文学など多様なカテゴリがあり得るが，LAMA-UHN における P136 の初期テンプレートでは対象を “[X] plays [Y] music.” と音楽に限定しており，音楽以外のカテゴリに関する正確な情報収集が困難であった．このような問題を有する関係テンプレートについては，元のテンプレートを削除し，新たに 5 つの関係テンプレートを作成した．

ステップ 2: GPT-4 による関係テンプレートの言い換え. 元の関係テンプレートと人手で書き換えた関係テンプレー

[*12] https://www.wikidata.org/wiki/Property:<relation_identifier>
[*13] https://www.wikidata.org/wiki/Property:P20
[*14] https://www.wikidata.org/wiki/Property:P136

トをもとに，OpenAPI によって提供される GPT4-API (gpt-4-1106-preview[*15]) を使用して，言語表現の言い換えを自動生成した．このとき，使用した指示プロンプトは以下の通りである：

> *You are a professional tool that can paraphrase sentences into natural sentences that can correctly represent the relationship between [X] and [Y], without repetition. Make the paraphrase as diverse as possible using simple words. Please paraphrase the given sentence 19 times.*

重複する文が生成された場合，重複する文を排除した上で，指示に従い新しい関係テンプレートを生成する作業を，異なる 19 個のテンプレートが得られるまで繰り返し実施する．また，特定の関係において，意味的に不適格な関係テンプレートが生成されるケースが確認された．そこで，そのような場合には，“主体” と “対象” の関係を捉えるために，エンティティのカテゴリ情報を加えた指示プロンプトを代わりに用いた．例えば，“[X] used to work in [Y].”[*16]という関係テンプレートを言い換える際は，初期の指示プロンプトに以下のプロンプトを追加入力した:

> *Be aware that [Y] is the geographic location but NOT company or organization, where persons or organizations were actively participating in employment, business or other work.*

この結果として，“[X] used to work in [Y].” からは以下のような関係テンプレートが自動生成された:

- “[X] was formerly employed in [Y].”
- “[X] once worked at [Y].”
- “[Y] was the place where [X] used to be engaged in work.”

A.1.2 MyriadLAMA の評価

我々が提案する知識プロービング手法 BELIEF では，関係知識ごとに多様性と品質を兼ね備えた膨大なプロンプトが必要となる．品質はプロンプトが対象分布の一部を正確に捉えることを保証し，一方で多様性は複数のプロンプトが真の知識分布の異なる側面を捉えることを保証する．本節では，これらの要素を精度，精度の揺らぎと一貫性の 3 つの側面から確認する．

A.1.2.1 プロンプト品質の評価

まず，MyriadLAMA の関係テンプレートの品質を，派生プロンプトを用いて BERT の亜種から得られる知識の精度を通して評価した．まず，各関係ごとに，関係テンプレートの精度 (Acc@1) で評価して，最小値，最大値，および平均値を求め，それらを 41 関係に対して平均して Acc@1 を計算した．表 A·1 に示されているように，MyriadLAMA のプロンプト品質は大きく変動しているが，高品質のプロンプトは LAMA-UHN のものよりも顕著に優れている．MyriadLAMA の平均精度は LAMA-UHN よりも低いが，LAMA-UHN が厳選したエンティティとテンプレートを使用しているのに対し，MyriadLAMA は半自動で書き換えた関係テンプレートを用いているためと考えられる．

表 A·1: MyriadLAMA と LAMA-UHN の Acc@1

	Min	Max	Avg	LAMA-UHN
$\text{BERT}_{\text{base}}$	.0000	.3534	.1103	.2403
$\text{BERT}_{\text{large}}$	.0007	.3728	.1185	.2454
BERT_{wwm}	.0015	.3695	.1453	.2448

表 A·2: 主体と関係の多様性評価 (%)

	一貫性		Acc@1 の範囲 (min/max)	
	主体	関係	主体	関係
$\text{BERT}_{\text{base}}$	57.45	15.04	6.73/14.41	0.00/35.34
$\text{BERT}_{\text{large}}$	54.97	15.48	7.14/15.54	0.07/37.28
BERT_{wwm}	50.05	10.57	8.31/18.84	0.15/36.95

A.1.2.2 プロンプト多様性の評価

次に，MyriadLAMA の派生プロンプトの多様性を評価するため，書き換えられた関係テンプレートと拡張された主体を一貫性と精度の範囲の観点から評価する．具体的には，全プロンプト集合を複数の部分集合に分割し，各部分集合内の “主体” または “関係” には一種類の表現のみが含まれるように設定して得られる派生プロンプトの Acc@1 を BERT の亜種を用いて評価した．

表 A·2 に結果を示す．この結果からは，“主体” のエイリアスによる一貫性と精度の揺らぎは関係テンプレートの一貫性と精度の揺らぎほど強いではないことが分かるが，関係知識評価への影響は無視できない．これらの結果は，既存研究で確認された単一プロンプトごとの関係知識評価の揺らぎを再確認するものであり，同時に，MyriadLAMA におけるプロンプトの多様性が関係知識評価にもたらす評価の頑健性を示唆する．

[*15] https://platform.openai.com/docs/models/gpt-4-and-gpt-4-turbo
[*16] https://www.wikidata.org/wiki/Property:P937