Vol.2021-NL-249 No.2
2021/7/27

大規模言語モデルの語彙的関係知識推定における日英間の比較調査

阿部 香央莉[1,a)]　北山 晃太郎[1]　松田 耕史[2,1]　吉川 将司[1,2]　乾 健太郎[1,2]

概要：近年各種タスクで最高性能を発揮している大規模言語モデル (LM) に対して，その内部にどのような語彙知識（上位下位・対義など）が蓄えられているかを調査する研究が盛んに行われている．しかし，これらの研究では基本的に英語を対象にした結果が報告されていることが多い．そこで，本研究では英語・日本語 LM 双方における語彙的関係知識推定の結果を比較し，既存研究で報告されている現象が英語と類型的特徴の大きく異なる日本語において観察されるのか検証する．

1. はじめに

近年，BERT[1] や RoBERTa[2] などの大規模事前学習済み言語モデル (pre-trained LM) が各種タスク（品詞タグ付け，固有表現抽出，文書分類など）において台頭しているが，その LM の中にどのような知識が蓄えられているかは自明ではない．そのため，LM がどのような知識を内部に保存しているかを調査する知識推定 (probing) の研究が盛んとなっている．中でも，Masked Language Model (MLM) タスクを解くように訓練された LM の穴埋め能力を利用して，テンプレートを用いた穴埋めタスク（cloze test, fill-in-the-blank task 等と称される）の精度によって LM の能力を測る研究がいくつかある．たとえば，“Dante was born in [MASK] in 1295.” などのようなテンプレートに対して，[MASK] トークンの位置にどのような単語が入るかを LM に予測させ，その予測が Wikipedia 等の知識ベースに書かれている事実に沿っているかを調査する事実知識に関する研究（factual probing）[3], [4] や，常識的知識や意味役割・イベント知識などの推定 [5] を行う研究などがあげられる．同様に，WordNet[6] のエントリとテンプレートを組み合わせることで，上位語 (hypernym)・対義語 (antonym)・共通の上位語を持つ同位語 (co-hyponym)・訛り (corruption) に関する語彙知識を推定する研究 [7] もある．しかし，これらの研究成果の多くは英語で報告されていることが多く，異なる言語で学習された LM においても，既存の研究で述べられていることが通用するのかは解明されていない．特に，日常的に扱われる文字種が複数あり，英語と言語類型が遠く離れている日本語において，既存研究の知見がそのまま反映されるかは自明ではない．たとえば，語彙知識推定の場合，「父」を示す単語には「おとうさん」「お父さん」「おとっつあん」「父親」「パパ」など，同じ概念を表すだけでもひらがな・カタカナ・漢字の三種類の文字を使い，ここで挙げた以上に表現のバリエーションがある．これら全ての表現が同一の概念を表すことや，それぞれの表現に応じて適切な対義語が存在することなどは，学習データの分布や単語分割の仕様にその精度が大きく依存する LM において捉えられていない可能性が高い．

そこで，本研究では，上位語・対義語・同位語を対象とする語彙的関係知識推定に着目し，同様の実験を日本語で行った際に，現状の LM の知識推定の研究において蓄えられている知見が英語と言語特性の大きく異なる日本語においても通用するのかどうか検証する．また，今回の実験では先行研究 [7] の手法に基づき，語彙的関係知識推定のためのデータセットを Open Multilingual WordNet に含まれる日本語のエントリから抽出して自動的に作成する．

結果として，英語での先行研究で述べられていた「(高頻度語では比較的推定できる一方で）低頻度語においては語彙知識の推定ができない」というような頻度による推定性能の傾向は日本語では現れず，一つの言語に対しての実験のみで LM の傾向を結論づけることへの懸念が裏付けられた．また，日本語における推定結果をもとに現状の LM の性能の分析を行なった結果，対義語については全く推定ができておらず，上位語に関しては予測が簡単だと思われる例については高い精度で推定できるものの，そうでない例に関しては著しく性能が低下することが判明した．

[1] 東北大学
Tohoku Univeristy, Sendai, Miyagi 980–8579, Japan
[2] 理化学研究所
RIKEN, Nihonbashi, Tokyo 103–0027, Japan
[a)] abe-k@tohoku.ac.jp

表 1 単語エントリの例（英語）[7].

関係 r	キーワード w_{key}	正解候補の集合 T
上位語	basketball	{game, ball, sport...})
対義語	new	old
同位語	samosa	{pizza, sandwich, salad, ...})

表 2 単語エントリの例（日本語）.

関係 r	キーワード w_{key}	正解候補の集合 T
上位語	トランペット	{ ブラス, 金管楽器, 真鍮, ...})
対義語	男	{ 女，女性 })
同位語	ピンク	{ ブロンド, ブルー, ...})

表 3 LM の語彙的関係知識予測に用いる英語のテンプレート例 [7].

関係 r	テンプレート p
上位語	(a) w_{key} is a [MASK]. "w_{key}" refers to a [MASK]. (a) w_{key} is a kind of [MASK].
対義語	w_{key} is the opposite of [MASK]. someone who is w_{key} is not [MASK]. something that is w_{key} is not [MASK]. w_{key} is not [MASK].
同位語	w_{key} and [MASK]. "w_{key}" and "[MASK]".

2. LM を用いた語彙的関係知識推定

2.1 英語での設定

語彙的関係知識推定には，WordNet から得られる単語エントリ（表 1），[MASK] トークンを含むテンプレート（表 3）の二つを用意する必要がある．先行研究では，四種類の関係（上位語，対義語，同位語，訛り[*1]）を扱い，あるキーワードに対しこれらの関係にある単語を予測できそうなテンプレートを複数作成し，それらのテンプレートを LM に入力した時に得られる [MASK] の埋め込み表現と類似度の高い単語候補（top-k）を列挙して，その候補中に正解となるような単語が含まれているかどうかを MRR で評価する．

単語エントリ x は（関係 r，キーワード w_{key}，正解候補の集合 T）の三つ組で表される．関係 r が対義語の場合，基本的に正解候補は一つの単語となる．関係 r が上位語や同位語の場合は，複数の正解候補を持つことが考えられる．また，各関係 r に対し，キーワードからその関係に対応する単語を予測できるようなテンプレート集合を $P(r)$ とし，テンプレート $p \in P(r)$ について，単語 w_{key} を用いて具体的な文にしたものを $p[w_{key}]$ とする．この $p[w_{key}]$ において，LM に [MASK] に対応する top-k を予測させた時に何番目に正解単語 $w_t \in T$ が予測できたか，をそのテンプレートにおける順位 $\mathrm{rank}'(p[w_{key}], w_t)$ とみなす．たとえば，関係 r = 上位語について，キーワード w_{key} を basketball とし，"basketball is a [MASK]." という文における [MASK] トークンに当てはまる語 top-3 を LM に予測させた時，尤度が高い順番に baseball, ball, sport という結果が得られた場合，正解候補の一つである ball が 2 番目に出力されているため，$\mathrm{rank}'(p[w_{key}], w_t) = 2$ となる．今回は，複数のテンプレートを用いるため，各エントリ x に対する $\mathrm{rank}(x)$ は複数テンプレートを試したうち最も rank′ が高かったものを用いる．こうして得られた各エントリの順位に対し，MRR (Mean Reciprocal Rank) を計算する．$\mathrm{rank}(x)$ および MRR は以下の式で表される．

$$\mathrm{rank}(x) = \min_{p \in P(r), w_t \in T} \mathrm{rank}'(p[w_{key}], w_t)$$
$$MRR = \frac{1}{|X|} \sum_{x \in X} \frac{1}{\mathrm{rank}(x)} \quad (1)$$

英語による先行研究 [7] では，各単語エントリのキーワードに対して Wikipedia データから得られる頻度情報を紐付け，エントリを高・中・低頻度の 3 種類のサブセットに分けて[*2]，頻度別での穴埋め予測の精度の比較を行い，その結果として「低頻度語においては各種語彙的関係にある語がうまく予測できない」という現象が報告されている．本研究では，同様の設定で実験を行い，この現象が日本語においても確かめられるかを検証する．英語側の実験では，先行研究において提供されているコード[*3]を利用し，再実験を行った結果を示す．先行研究に従い，語彙知識推定を行う LM には BERT-base (uncased)[*4] を採用した．このとき，LM の予測候補数は $k = 5$ とした．

表 4 LM の語彙的関係知識予測に用いる日本語のテンプレート例.

関係 r	テンプレート p
上位語	w_{key} とは [MASK] である。 「w_{key}」とは [MASK] の一種である。 「w_{key}」とは [MASK] のことを指す。
対義語	w_{key} と [MASK] は対の関係にある。 w_{key} と [MASK] は逆の関係にある。 w_{key} と [MASK] は反対の関係にある。 w_{key} の反対は [MASK] である。 w_{key} は [MASK] ではない。
同位語	w_{key} と [MASK]。 「w_{key}」と「[MASK]」。 w_{key} と [MASK] の違い。 w_{key} と [MASK] の違いについて。

2.2 日本語での設定

日本語においても英語と同様の実験を行うために，日本

[*1] ここでは，正規形から文字が追加・消失した単語を訛りと呼ぶことにする．

[*2] 頻度の基準は，Wikipedia における出現回数が 10 回より少ない単語を低頻度，10 回以上 100 回未満の単語を中頻度，100 回以上の単語を高頻度としている．

[*3] https://github.com/timoschick/am-for-bert

[*4] https://huggingface.co/bert-base-uncased

語の単語エントリ（表 2）とテンプレート（表 4）を作成する．このとき，英語で検証された語彙的関係のうち訛り (corruption) に関しては，送り仮名の有無など複雑な問題が混じってくることが考えられたため，訛りを除いた上位語，対義語，同位語の三種類の語彙的関係について調査を行うこととした．実験で扱う LM として，日本語 BERT の bert-japanese[*5]を採用した．英語と同様に，LM の予測候補数は $k = 5$ とした．

2.2.1 WordNet からの単語エントリ取得

Open Multilingual WordNet から上位語および同位語を取得する際，簡易化のため，今回は間接的な上位語は扱わず直接上位と定義されているもののみを使用した．

また，Open Multilingual WordNet から対義語エントリを取得する際，英語の場合には各見出し語 (lemma) に対し対義語となる見出し語が定義されているが，日本語の場合，その対義語の整備が行き届いていない[*6]．そのため，今回の実験では一度英語の見出し語を経由することで，日本語の対義語を取得することにした．例として，「男」に対応する対義語を手に入れる際は，一度英語の “man” に対応する synset を経由して “woman” に対応する synset を取得し，その synset から日本語の見出し語を手に入れる，という処理を行う．本来はある語に対する正解の対義語は一つとなるのが理想的だが，この手順では必ずしも対応する一つの対義語を手に入れることができないため，日本語の実験設定では複数の正解候補を許容することとした（表 2 の対義語を参照）．

また同位語に関して，一部の単語（色や動物を表す概念など）に関しては，単純に直接の上位 synset から他の下位 synset を取得してきても，正解候補が膨大になり，正解を判断するのに計算コストが高くなる恐れがある．したがって，今回の実験設定では同位語の候補を五つの synset に絞り，その五つから得られる見出し語集合から正解を判定することとした．

上記の手順で取得された各関係の単語エントリに対し，bert-japanese を事前学習する際に用いられた 2020 年 8 月 31 日時点の Wikipedia ダンプデータから頻度情報を算出し，英語における実験設定と同様の基準で単語エントリを高・中・低頻度のサブセットに分けた．

2.2.2 テンプレートの作成

テンプレートによる予測を LM に行わせるにあたり，上位語などの単語を適切に取り出せるようなテンプレートを用意する必要がある．このとき，英語の LM 用に作られたテンプレートが必ずしも他の言語において適切とは限らないという懸念が生じる．たとえば，同位語のテンプレートにおいて，「w_{key} and [MASK].」の直訳である「w_{key} と [MASK]。」というテンプレートの穴埋めを日本語 BERT に行わせた場合，[MASK] トークンの位置に「は」「で」などの助詞に該当する単語が候補として予測されることが多く見られた．これは，一般的に英語の “and” が前のフレーズと後のフレーズを等位に接続する用途でしか使われないことに対し，日本語の「と」には “and” 以外の用法も存在するためであると考えられる．これは，英語の直訳によるテンプレートでは [MASK] に同位語が来ることを暗に制限できないという問題を示唆している．また，英語のテンプレートの差分は，冠詞（a）の有無やクォーテーション（“”）の有無で作られているものが複数見受けられる．こういったテンプレートでは，冠詞をキーワードや [MASK] トークンの直前に付加することで，[MASK] トークンに必ず名詞が現れるように制御する効果が生まれている可能性があるが，日本語の場合，冠詞に該当する概念がないためこのような工夫は困難である．そのため，日本語テンプレートにおいては，冠詞による差分は扱わないものとし，クォーテーションによる差分に相当するものとして鍵括弧（「」）による差分でのテンプレートを作成した．

結果として，今回の実験では，先行研究における英語テンプレート（表 3）を参考に，概ねその翻訳となるようなテンプレートと，新たに日本語において適切な単語を予測できそうなテンプレートを人手で作成した．

[*5] https://huggingface.co/cl-tohoku/bert-base-japanese-v2
[*6] 2021 年 7 月現在，ドイツ語やポルトガル語などの別の言語においても，同様に対義語は整備されていなかった．

表 5 頻度別語彙的関係知識データセットの統計．表中の数値は単語エントリの数に相当する．

	英語 [7]			日本語		
	高頻度	中頻度	低頻度	高頻度	中頻度	低頻度
上位語	4,750	1,785	1,191	20,151	7,487	7,472
対義語	266	58	41	1,055	301	400
同位語	6,126	2,740	1,960	20,376	6,766	7,007

表 6 英語・日本語両言語における語彙的関係知識推定の MRR．英語の値については再実験を行った結果を記載．

	英語			日本語		
	高頻度	中頻度	低頻度	高頻度	中頻度	低頻度
上位語	0.391	0.298	0.252	0.184	0.230	0.200
対義語	0.368	0.090	0.119	0.042	0.072	0.076
同位語	0.270	0.156	0.124	0.095	0.090	0.080

3. 実験

2 節で述べたデータセットを使用して，英語・日本語両言語における語彙的関係知識の推定を行う．頻度別の語彙的関係知識推定用データセットの内訳は表 5 に示す通りである．このデータセットを用いて，語彙知識推定を行った結果を表 6 に示す．

表 6 より，英語においては，先行研究で述べられている

表 7 日本語の上位語推定における実際の例．top-5 nearest neighbor は，「w_{key} は [MASK] の一種である．」をテンプレートとして使用した時の予測結果を示す．

エントリ	頻度	WordNet による正解	top-5 nearest neighbor
フィッシュ・アンド・チップス	高	料理	寿司，以下，フィッシュ，**料理**，カクテル
ホッキョクグマ	高	クマ，熊	カニ，カエル，卵，魚，鳥類
排他的論理和ゲート	低	ゲート	**ゲート**，以下，これ，次，それ
レモンメレンゲパイ	低	パイ	**パイ**，以下，パン，菓子，ケーキ

通り低頻度になればなるほど語彙的関係知識推定の精度が下がっていく傾向が見られた．しかし，日本語においては，頻度による大きな傾向は見られないという結果となった．具体的に，同位語に関しては，頻度問わず MRR は横ばいで全て 0.1 以下というかなり低い値が得られ，また上位語や対義語においては，高頻度の MRR が中頻度や低頻度のものよりも低くなるという結果となった．この結果は，英語における実験結果が必ずしも他言語で同じ傾向を示すわけではないということを裏付けている．

しかし，MRR の値を観測するだけでは，実際に日本語における語彙的関係知識推定でどのようなエラーが生じているのか自明ではない．したがって，以降の 4〜6 節では，日本語での語彙的関係知識推定におけるエラーについて，何がエラーの原因となっているかを解明するため，実際の予測例を交えてより詳細な分析を行う．

4. 分析：上位語について

4.1 定性評価：実際の予測例

表 7 に，日本語での上位語推定における実際の例を示す．

高頻度語において上位語を一つも生成できなかった例（「ホッキョクグマ」など）に関しては，カタカナで表された種族名などが多く見られた．これらは，各エントリのキーワードに対してサブワード分割処理を行なった際に，キーワードが細切れに分割されてしまうことで，そのキーワードが持つ本来の意味を LM 内でうまく類推できていないということが考えられる．例えば，文字数の多いキーワードでありながらも上位語推定に成功した「フィッシュ・アンド．チップス」は「フィッシュ/・/アンド/・/チップ/##ス」と分割されたのに対し，「ホッキョクグマ」は「ホ/##ッキ/##ョ/##ク/グ/##マ」と分割されていた（「/」は各トークンの境目を表す）．そこで，キーワードのサブワード分割数の長さと LM の知識推定の正否の関係について，4.2 節でさらに詳細な分析を行う．

また，低頻度語において上位語を推定できた例（「排他的論理和ゲート」「レモンメレンゲパイ」など）は、キーワード中に上位語の概念を表す意味が含まれていることが多い傾向が見られた．これらの例に関しては，キーワードの一部を抜き出して答えれば正解となるようなものが多く，予測が非常に簡単な例であると見られる．このような，キーワード中に正解単語が含まれる例に関しても，4.3 節でさらに詳細な分析を行う．

4.2 定量評価：キーワードのサブワード分割数と精度の分析

LM にテンプレート中の [MASK] トークンに近い単語 top-5 を予測させ，その top-5 中で正解単語を一つでも予測できたエントリの総数を，キーワードのサブワード分割数ごとに示したものが表 8 となる[*7]．

表 8 より，top-5 中で正解単語を予測できたか否かで結果を見てみると，高頻度語は比較的どの分割数においても正解率（%）が一定より高いことが読み取れる．この結果と表 6 の結果を照らし合わせると，日本語 BERT は正解となる単語を予測することには成功しているものの，高い順位で出力することができていない，ということになる．

実際の例を見てみると，サブワード分割数が 8 以上のエントリの中には，他言語の WordNet エントリをそのまま日本語訳したような，日本語の単語エントリとして不自然な例が複数見られた（「オーブ/##ン/で/焼/##か/れ/た/食品」など）．また，サブワード分割数が 2 で予測に失敗した単語は，人間にとっても難しい例（「徒路」「千鳥草」）や，送り仮名・変換の揺れによるもの（「有り/##明け」「郵便/うけ」など）が多く見られた．送り仮名や変換の揺れは，先行研究で扱われていた英語の訛り (corruption) に近いものと考えられる．そのため，これらの例は本来調査したい上位語の知識推定に加え，英語では訛りに相当するような別の問題が混在してしまっていると考えられる．

また，不正解となったエントリの中には，LM の文字語彙セット中に入っていない旧字体を単語中に含んでいることによって，トークナイズ時に未知語 ([UNK]) トークンになってしまっている例も複数見受けられた．特に低頻度語セットにおいて，分割後に未知語トークンが出現するエントリは 198/7,472 件存在した（比較として，日本語高頻度セットでは 23/20,151 件，英語は低頻度セットでも未知語トークンの出現数は 0 件となっていた）．

[*7] 低頻度語かつ分割数 1 の 138 件のエントリのうち，キーワードが完全に “[UNK]” になっているものを除くと 17 件になる．この場合，正解数は 1/17 となり，正解率は 5.9%となる．また中頻度かつ 分割数 1 の 97 件のうち，同様の例を除くと 2 件になる．この場合，正解数は 0/2 で正解率は 0%となる．言い換えると，中・低頻度語でキーワード分割数 1 のエントリの中には，“[UNK]” をキーワードとしているのにもかかわらず正解を予測できてしまっている例が 2 件ずつある，ということになる．

表 8 各頻度セット中で，予測された top-5 が一つでも正解単語と一致した上位語エントリの数（#）および割合（%）

	分割数	10 以上	9	8	7	6	5	4	3	2	1
高頻度	#	-	-	-	-	1/5	15/29	99/345	729/2,312	3,525/11,545	1,149/5,915
	%	-	-	-	-	20.0	51.7	28.7	31.5	30.5	19.4
中頻度	#	-	-	1/1	1/2	10/33	40/126	182/569	757/2228	1308/4431	2/97
	%	-	-	100.0	50.0	30.3	31.7	32.0	34.0	29.5	2.1
低頻度	#	0/3	1/3	2/3	2/11	17/78	69/272	253/984	782/2729	893/3251	3/138
	%	0.0	33.3	66.7	18.2	21.8	25.4	25.7	28.7	27.5	2.2

表 9 高頻度キーワードを持つ上位語エントリに対する RR の分布.

	0	0.2	0.25	0.333	0.5	1
全体	**72.6**	2.0	2.9	3.5	5.6	13.3
キーワードに正解単語が含まれる（全体の 16.3%）	14.3	2.7	3.4	5.1	14.5	**60.1**
キーワードに正解単語が含まれない （全体の 83.7%）	**83.7**	1.9	2.8	3.2	4.0	4.4

これらの例から，今回用いた語彙知識推定用データセットの構築法は，WordNet という共通基盤から単語エントリを得ることで異言語間でもある程度同一条件での実験が可能ではあるものの，特定の言語での語彙的関係知識の推定を目的とした際，その単語エントリの妥当性を検証する必要があることが明らかになった．特に日本語においては，送り仮名や変換の揺れ，文字単位での未知語の出現などを考慮する必要があることがわかった．

4.3 定量評価：キーワード中に正解候補を含む・含まないエントリにおける推定精度

今回日本語での実験において使用したエントリのうち，正解単語がキーワード内に含まれているエントリ（「排他的論理和ゲート」など）は，高頻度・中頻度・低頻度においてそれぞれ 3,276 件，1,658 件，1,493 件 （全エントリ中の 16.3%, 22.1%，20.0%）見られた．このようなエントリを一つのサブセットとして見た時の MRR はそれぞれ $0.704, 0.764, 0.736$ となり，正解がキーワード中に含まれている例においては顕著に予測が当たりやすい傾向が見られた．比較として，正解がキーワードに含まれていないエントリにおける MRR はそれぞれ $0.085, 0.077, 0.066$ となり，キーワード中に正解単語が含まれていない場合，日本語 BERT は顕著に推定に失敗することもこの数値から読み取れる．

また，高頻度語の各エントリに対する RR の分布を表 9 に示す．RR は各サブセット全体で平均を取る前の値であり，各エントリにおける順位 $\mathrm{rank}(x)$ の逆数である．今回の実験では，BERT に top-5 を予測させているため，RR の値は $\{0.0, 0.2, 0.25, 0.333..., 0.5, 1.0\}$ の 6 種類に限られる．たとえば，複数あるテンプレートのうちどれか一つでも top-1 に正解単語を予測できていたら 1.0，どのテンプレートにおいても正解単語を予測できなかった場合は 0.0 となる．

表 9 から，正解がキーワード中に含まれるエントリに関しては，予測が成功しやすい上に top-1 に出力できている傾向も高い（6 割のエントリで top-1 に出力できいる）ことが読み取れる．反対に，正解がキーワード中に含まれないエントリに関しては，およそ 8 割程度のエントリが正解単語を一つも予測できていないことも読み取れる．これらのことから，現状の日本語 BERT の語彙知識推定で正解できているものは，キーワードの一部を抜き出すだけで正解となる比較的簡単と見られるものが多く，日本語 BERT が本当に上位下位関係に関する知識を蓄えられているかは定かではない，ということがわかった．

5. 分析：対義語について

表 10 に，日本語での対義語推定における実際の例を示す．表 10 に示した通り，対義語推定における日本語 BERT の性能は非常に悪く，実際の予測結果を見てみると，そもそも [MASK] トークンの予測においてひらがな一文字や「反対」などの意味を持つ単語などが出現し，全体的にほぼどのエントリにおいてもまともな単語を予測できていない傾向が見られた．この現象は，「男」や「右」など高頻度かつ対義語が自明である単語に関しても同様に見られており，少なくとも現状の日本語 BERT は，穴埋め形式の対義語推定においてまともに機能していないことが判明した．

6. 分析：同位語について

表 11 に，日本語での同位語推定における実際の例を示す．2 節でも述べた通り，今回の実験においては計算量の問題を考慮し，同位語の候補を 5synset までに絞っている．そのため，表中の「ピンク」の例のように，日本語 BERT が同位語として適切な単語（「赤」「オレンジ」など）を予測できているにもかかわらず，正解候補にその単語がないため，正解と判定されない，もしくは過少評価されてしまうという問題が見られた．

表 10 日本語の対義語推定における実際の例．top-5 nearest neighbor は，「w_{key} と [MASK] は対の関係にある。」「w_{key} の反対は [MASK] である。」をテンプレートとして使用した時の予測結果を示している．

エントリ	頻度	WordNet による正解	top-5 nearest neighbor
男	高	女，女性	[と，の，し，お，ん]，[反対．自由，逆，自然，曖昧]
右	高	レフト，左，左側，左手	[と，し，の，ん ##ち]，[反対，自由，悪，真，逆]
お子様	高	親	[ママ，母，子供，子，女房]，[反対，**親**，逆，父，悪]
おとっつあん	低	マザー，実母，母，母親	[と，で，プリン，は，び]，[反対，賛成，逆，同じ．これ]

表 11 日本語の同位語推定における実際の例．top-5 nearest neighbor は，『「w_{key}」と「[MASK]」。』をテンプレートとして使用した時の予測結果を示す．また，WordNet による正解のうち，[] に括られている単語は同じ synset に属す見出し語であることを示す．

エントリ	頻度	WordNet による正解	top-5 nearest neighbor
ピンク	高	[ブロンド]，[ブルー，...]，[ブラウン，...]，[グリーン，...]	ピンク，赤，オレンジ，白，**ブルー**
犬	高	雌，[フォックス，狐，稲荷]，[ウルフ，オオカミ，狼]	猫，犬，人，ネコ，ウサギ
センチリットル	低	[l，l，L，ミル，cc，[UNK]]	リットル，センチ，センチメートル，キログラム，ミリ

ただし，「犬」の例に関しては，「猫」や「ウサギ」などの単語が出力できているものの，WordNet の直接上位語が「イヌ科動物」であることにより，同位語 synset を五つに制限しなかったとしても，正解と判定できない可能性がある．また，「センチリットル」の例に関しては，「リットル」に相当する synset が正解として選ばれているものの，正解候補の中に「リットル」そのものがなく，リットルと出力されているのに正解と判定できていない．このように，同位語に関してはその正解候補となる概念が広いこともあるためか，WordNet から得られるエントリの質が適切と断言できない例が多く，今回の実験結果から決定的な結論を出すのは早計である．今後の展開として，今回実験に使用したエントリが適切かどうかをクラウドソーシングなどを用い人手でフィルタリングするなどの方針を考えている．

7. 議論

7.1 テンプレートを用いた知識推定について

2 節において述べた通り，テンプレートによる知識推定にはその言語において適切に推定したい知識を掘り出すためのテンプレートを考案する必要があり，テンプレートが妥当かどうかを判断するにはその言語についての母語話者レベルの知識が必要になる．また，テンプレートを母語話者が新しく考案したところで，そのテンプレートが必ずしも LM に上位語などの目的の単語を出力させるような働きをするとは限らない．これは，自然言語で書かれたテンプレートの穴埋め方式による枠組み自体が持つ懸念である．このような人手作成されたテンプレートの恣意性を排除するために，知識推定に用いるテンプレート自体を LM によって予測生成させる，という研究もいくつか現れている [8], [9]．しかし，これらの手法によって得られたテンプレートは人間が理解できるような系列にはなっておらず，解釈性が悪いという難点がある．テンプレートを用いた穴埋め形式による知識推定の是非については，いまだに議論の余地があり，より適切な LM の知識推定手法についても引き続き研究を続けていく必要がある．

7.2 WordNet による評価について

現状，日本語側のデータは完全自動で取得した後，特にフィルタリングなどの処理は行っていないため，4 節や 6 で示したように，ノイジーなエントリを複数含むデータとなっている．今回の実験では，上位語（および同位語）の正解として直接関係する上位語のみを採用しているが，直接関係する上位語でも人間の目から見て直感的でない例（「犬」の上位語として「イヌ科動物」など）がある．しかし，WordNet の木を遡って間接的な上位語も全て許容すると，膨大な正解候補が生まれてしまい，結果的に正解データとしての質が落ちてしまうという難点がある．これらのことから，現状完全自動の手法で，適切な語彙的関係知識データセットを得るのは難しいと考えられる．したがって，今後の方針として，クラウドソーシングによって人間が見ても妥当と判断できる語彙的関係データセットを構築することが挙げられる．

また，そもそも WordNet 自体が整備・維持コストの問題を抱えているため，WordNet のような語彙資源自体を LM で作成するという新たな試みも行われている [10]．この研究では複数の言語（10 言語）での実験が行われているが，日本語は含まれていない．この試みを日本語で行うのも今後の課題の一つとして考えている．

8. おわりに

LM の語彙的関係知識推定について，先行研究において英語で行われた実験を日本語に変えて調査を行い，先行研究で示されている傾向が英語と言語特性の大きく異なる日本語でも見られるのかどうかを検証した．結果として，同様の実験設定において日本語 BERT の知識推定を行ったものの，日本語 BERT では先行研究で述べられていた頻度による精度の傾向は出ず，全体的にどの語彙的関係およびどの頻度においても英語ほどの性能に達さなかった．この結果は，英語のみによる実験で LM の知識推定の性能について結論づけるのは早計である，という主張を裏付ける一つの根拠となる．

さらに，日本語での推定結果において，詳細なエラー分析を行い，正解候補となる単語をキーワード中に含むエントリにおいては，頻度に関係なく顕著に性能が高くなり，反対に含まないエントリにおいては性能が低くなることがわかった．また，特に低頻度語においては送り仮名・変換揺れ，未知語トークンの出現により推定がより困難になっている可能性があることも判明した．

今回の実験では語彙的関係知識推定に用いるエントリを完全自動で取得していることにより，データのノイジーさについての議論が生じる．したがって，今後の方針としては，クラウドソーシングで日本語の語彙的関係知識推定用のデータを綺麗にすること，また，穴埋め形式による知識推定という方法論の妥当性についても検証しながら，引き続き現状の LM の問題点をさらに詳細に探りたい．

謝辞 本研究は JSPS 科研費 JP20J21694 の助成を受けたものです．

参考文献

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

[2] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. and Stoyanov, V.: RoBERTa: A Robustly Optimized BERT Pretraining Approach, *arXiv preprint arXiv:1907.11692* (2019).

[3] Petroni, F., Rocktäschel, T., Riedel, S., Lewis, P., Bakhtin, A., Wu, Y. and Miller, A.: Language Models as Knowledge Bases?, *Proceedings of EMNLP-IJCNLP*, Hong Kong, China, Association for Computational Linguistics, pp. 2463–2473 (online), DOI: 10.18653/v1/D19-1250 (2019).

[4] Hosseini, A., Reddy, S., Bahdanau, D., Hjelm, R. D., Sordoni, A. and Courville, A.: Understanding by Understanding Not: Modeling Negation in Language Models, *Proceedings of NAACL*, Online, Association for Computational Linguistics, pp. 1301–1312 (online), DOI: 10.18653/v1/2021.naacl-main.102 (2021).

[5] Ettinger, A.: What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models, *CoRR*, Vol. abs/1907.13528 (online), available from ⟨http://arxiv.org/abs/1907.13528⟩ (2019).

[6] Miller, G. A.: WordNet: A Lexical Database for English, *Commun. ACM*, Vol. 38, No. 11, p. 39–41 (online), DOI: 10.1145/219717.219748 (1995).

[7] Schick, T. and Schütze, H.: Rare Words: A Major Problem for Contextualized Embeddings and How to Fix it by Attentive Mimicking, *Proceedings of AAAI*, Vol. 34, No. 05, pp. 8766–8774 (online), DOI: 10.1609/aaai.v34i05.6403 (2020).

[8] Shin, T., Razeghi, Y., Logan IV, R. L., Wallace, E. and Singh, S.: AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts, *Proceedings of EMNLP*, Online, Association for Computational Linguistics, pp. 4222–4235 (online), DOI: 10.18653/v1/2020.emnlp-main.346 (2020).

[9] Zhong, Z., Friedman, D. and Chen, D.: Factual Probing Is [MASK]: Learning vs. Learning to Recall, *Proceedings of NAACL*, Online, Association for Computational Linguistics, pp. 5017–5033 (online), DOI: 10.18653/v1/2021.naacl-main.398 (2021).

[10] Chen, C., Lin, K. and Klein, D.: Constructing Taxonomies from Pretrained Language Models, *Proceedings of NAACL*, Online, Association for Computational Linguistics, pp. 4687–4700 (online), DOI: 10.18653/v1/2021.naacl-main.373 (2021).