ルールベースの深層格定義および自動付与とその LLM 統合による含意関係認識における効果検証

荒沢 康平 [1] 狩野 芳伸 [1]
[1] 静岡大学
{karasawa, kano}@kanolab.net

概要

LLM（Large Language Model、大規模言語モデル）の発展に伴い、多くの言語処理タスクで LLM が高い性能を発揮している。しかし、深層格で表現されるような複雑な事物の関係を正確に捉えられているかは不明である。我々は独自に定義した深層格とその付与ルール群に基づき、ルールベースの深層格自動付与システムを構築した。その出力と LLM を組み合わせることで、言語処理タスクの性能が一般に向上しうると期待する。本研究では性能検証として、JNLI データセットを用いて含意関係認識を行った。その結果、GPT-4o との組み合わせで正解率が 2.67 ポイント向上し 84.75 の SoTA 評価値を達成し、我々の自動深層格付与システムの有効性を示した。

1 はじめに

昨今の LLM（Large Language Model、大規模言語モデル）の発展に伴い、多くの言語処理タスクで LLM が高い性能を発揮している。しかし、深層格で表現されるような複雑な事物の関係を正確に捉えられているかは不明である。我々は独自に定義した深層格とその付与ルール群に基づき、ルールベースの深層格自動付与システムを構築した。その出力と LLM を組み合わせることで、言語処理タスクの性能が一般に向上しうると期待する。

河崎ら [1] は、言語学におけるいわゆる深層格は「述語に係る名詞の述語に対する意味的関係である」との解釈を示している。国立国語研究所 [2] は深層格について「人によっては述語に対する名詞の意味役割などと称し、平々の言語に依存しない普遍的なものであるか、あるいは個々の言語の特質を包含する最小公倍数的なものであることを前提とする。」とした一方で、「深層格についてまだ定説と言えるものがない」との見解を示した。本研究においては河崎らの深層格に対する解釈に準拠する。

渋木ら [3] は構文解析済みの単文に対して深層格を推測する手法を提案した。タグ付きコーパスから得られる「深層格選好」を利用し、タグなしコーパスから多様な言語表現に対応する規則を学習する。深層格選好を手掛かりとすることで言語依存性を低減し、日本語と英語の EDR コーパスを用いた実験では、クローズドデータで日本語 81.2 %・英語 78.5 %、オープンデータで日本語 69.5 %・英語 73.5 %の精度が報告されている。しかし、推測対象を「一つの動詞と二つの名詞で構成される単文」としており、この手法の適用範囲は限定的である。

河崎ら [1] は、入力層に名詞と動詞の分散表現、助詞の one-hot 表現を用いた 3 層のニューラルネットワークを構築し、それを用いて名詞 1 単語の深層格推定をする手法を提案し、日本語の深層格推定において 9 割以上の適合率を達成した。

深層格推定に関するこれらの先行研究は深層格推定対象が「名詞 1 単語」であることや「一つの動詞と二つの名詞で構成される単文」であるなど手法の適用範囲が限定的であることから、汎用的で網羅的な深層格自動付与システムが存在しないことが課題となっている。

国立国語研究所の報告 [2] では各種表層格ごとの深層格とその成立条件を定義している。例えばガ格の場合、「動作主格」「経験者格」「無意志主体格」「対象格」「属性格」の 5 種類の深層格を定義している。しかし、深層格の成立条件を定義しているだけのため、文に対して深層格を付与するには人手の判断が必要である。この課題を解決するべく本研究では成立条件をプログラムに落とし込み、1 文に対して自動で深層格を付与するルールベースのプログラムを開発した[1)]。

深層格にはフィルモア [4] が提唱した「一文一格

1） オープンソースで公開予定

の原理」という制約がある。これは「同じ格が、複合的な場合を除いて、単文（基本文）に一回しか起こらないとする。」という制約である。深層格自動付与プログラムを作成するにあたり、この制約に違反することがないようルールに調整を加えた。

本研究では性能検証として、含意関係認識を行った。その結果、実験したどの LLM でも正解率が向上し、自動深層格付与システムの効果が示された。

2 関連研究

2.1 含意関係認識データセット

日本語の含意関係認識データセットには、JSNLI[5]、JSICK[6]、JNLI[7] 等がある。それぞれ事前学習済みモデル [8] をファインチューンした Accuracy の報告値をカッコ内に表記して紹介する。

JSNLI[5] は英語の大規模な含意関係認識データセットである SNLI[9] を日本語に機械翻訳したものである（0.929）。JSICK[6] は様々な言語現象を含む英語の含意関係認識データセットである SICK[10] を人手で日本語に翻訳したものである（0.84）。JNLI[7] は、日本語理解ベンチマーク JGLUE[7] に含まれる含意関係認識データセットで、翻訳を介さずに作成された（0.906）。JNLI はクラウドソーシングで構築され、画像の内容を文章で表現させることで、含意と中立のラベルを持つ文ペアを作成している。表現した文章に対して矛盾する文ペアは、クラウドソーシングによって構築された。

2.2 LLM を用いた含意関係認識

OpenAI 社の GPT-4o[11] は、さまざまなベンチマークにおいて人間に匹敵しうるパフォーマンスを示している。執筆時点では、GPT-4o を高性能な代表モデルのひとつとして性能検証することは妥当と考えられる。Nejumi LLM リーダーボード Neo[12] で公開されている、 gpt-4o-2024-05-13 で JNLI を評価した結果では、Accuracy 0.85 を達成している。

高性能なオープンモデルの日本語 LLM として、LLM-JP が挙げられる。Nejumi LLM リーダーボード Neo に掲載されている JNLI の性能上位 2 件[2)]の LLM は、それぞれ Accuracy が 0.91, 0.89 である。ただし、GPT-4o よりも後者二つの LLM の性能が高いのは、後者二つのインストラクション訓練データに JNLI が含まれているためと考えられ、JNLI の性能比較としては適切ではないため対象外とした。

2） llm-jp-13b-instruct-full-jaster-v1.0 (https://huggingface.co/llm-jp/llm-jp-13b-instruct-full-jaster-v1.0) および llm-jp-13b-instruct-lora-jaster-v1.0 (https://huggingface.co/llm-jp/llm-jp-13b-instruct-lora-jaster-v1.0)

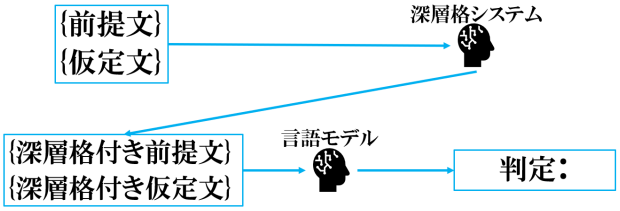

図 1 提案手法の概要

3 提案手法

本研究では独自のルールに基づく深層格自動付与システムを提案し、その性能検証として含意関係認識を実験する。テストデータとしては 1 から日本語で構築された JNLI[7] を使用する。まず、深層格自動付与システムを用い、JNLI の前提文と仮定文それぞれに深層格を付与する。次に、深層格が付与された前提文と仮定文を LLM に入力し、ゼロショットで推論する (図 1)(以下、この提案手法を「深層格あり手法」とする)。

ベースライン手法としては、LLM に対して原文のままの JNLI の前提文と仮定文を入力し、ゼロショットで推論する (以下、このベースライン手法を「深層格なし手法」とする)。

さらに、「深層格あり手法」と「深層格なし手法」をアンサンブルした手法も提案する (以下、この提案手法を「アンサンブル手法」とする)。これは、前提文と仮定文に出現する深層格が完全一致する場合は「深層格あり手法」の推論結果を、深層格が非一致の場合は「深層格なし手法」の推論結果をそれぞれ採用する手法である。「深層格が完全一致する」とは、前提文と仮定文に出現する深層格の種類・数が一致する場合を示す。具体的には、図 2 のような形となる。「深層格が非一致」とは、「深層格が完全一致しない」状態を指す。

前提 箱の中に < 対象格 > パンが < 主格 > 並んでいます。< 述語動詞 >
仮定 箱の中に < 対象格 > ドーナツが < 主格 > 並んでいます。< 述語動詞 >

図 2 深層格が完全一致する場合の例

3.1 深層格自動付与システム

深層格自動付与システムは、国語研が定義した深層格とその成立条件を基礎として我々が独自に開発

したものである。国語研が定義した深層格は種類が非常に多く、役割の重複が見られたため、自動処理に耐えうるようそれらを整理・統一する作業を行った。また、実際の文例を検討していくと、国語研の深層格だけでは十分にカバーできないケースがあった。そうしたケースに対応すべく、新たな深層格を独自に定義した。これらの修正と追加の結果、本研究において定義する深層格は表 1 の通りとなった。

この定義を用いた深層格自動付与システムの処理を説明する。まず入力文に対して KNP 4.19 [13][14][3] を用いて形態素解析を、Juman++ 2.0.0-rc3[15] [4] を用いて構文解析を行う。解析結果から動詞に直接係る名詞のかたまりを抽出し、それぞれの表層格と素性を獲得する。表層格と素性の情報と作成した深層格ルールを照らし合わせて、名詞のかたまりそれぞれに深層格を付与する。たとえば先の図 2 では深層格付与の結果、＜対象格＞＜主格＞＜述語動詞＞のタグが挿入されている。

深層格名	特徴
主格	ある動作や状態変化を引き起こす主体。
対象格	ある動作や状態変化の影響を受けるもの。
受け手格	ある動作や状態変化の影響を受けるもの。 主にヲ格を持つ動詞の二格に付与する。
数量格	数や量を表す。
時間格	時間を表す。
場所格	場所を表す。
部位格	場所を表す名詞の中で 特に動物の体の一部を表す。
起点格	物事の起点を表す。 主にカラ格に付与する。
終状態格	変化の結果を表す。
手段・道具格	有形の道具の他に抽象的な手段も含む。
相手格	ある動作の相手を表す。 主にト格に付与する。

表 1　独自に定義した深層格の一覧

3.2　プロンプト

提案手法では、Nejumi LLM リーダーボード Neo の評価指標の一つである llm-jp-eval で使用されている JNLI 専用プロンプト[5] を微調整し、利用する（図 5）。深層格あり手法では「#前提文#」と「#仮定文#」には深層格自動付与システムで深層格が付与された文を使用するほか、元のプロンプトに「注意：」の項目を加える。深層格なし手法では、図 5 から「注意：」の項目を削除し、「#前提文#」と「#仮定文#」に深層格が付与されていない原文を入れたプロンプトを使用する。

3）https://nlp.ist.i.kyoto-u.ac.jp/index.php?KNP
4）https://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN%2B%2B
5）https://github.com/llm-jp/llm-jp-eval/blob/dev/src/llm_jp_eval/jaster/jnli.py

4　実験設定

商用モデルとして gpt-4o-2024-11-20(以下、GPT-4o とする)、オープンモデルとして cyberagent/Llama-3.1-70B-Japanese-Instruct-2407[6] [16](以下 Llama3.1 とする)、日本語モデルとして llm-jp-3-13b-instruct[7] を用いた。出力の一貫性をできるだけ保つため、これらのモデルではすべて temperature を 0.1 に設定した。

本研究では訓練を行わないため、評価データのみを必要とする。評価では JNLI の valid-v1.1.json を用いて、矛盾・含意・中立の 3 値分類を行い、Accuracy, F1 値、precision, recall を算出した。

5　実験結果

表 2 に実験の評価結果を示す。GPT-4o では、「アンサンブル手法」が「深層格なし手法」の正解率を 2.67 ポイント上回り、Llama3.1 では、「深層格あり手法」が「深層格なし手法」の正解率を 1.47 ポイント上回った。

また、インストラクション訓練データに JNLI が含まれている llm-jp-13b-instruct-full-jaster-v1.0 を除くと、GPT-4o を用いた提案手法の「アンサンブル手法」が正解率 84.75 で最高性能となった。これは知る限り執筆時点での世界最高性能である。

6　考察

6.1　深層格付与により正答できた事例

GPT-4o を対象とした「深層格あり手法」「深層格なし手法」「アンサンブル手法」それぞれにおける、正解ラベルと推論ラベルの混合行列を、表 3 に示す。

正解が矛盾であった問題における正解数は「深層格あり手法」が 643 件で最も多かった。また、正解が矛盾の問題を中立と誤答した数は「深層格なし手法」の 199 件から「深層格あり手法」の 83 件に顕著に減少している。これらのことから、矛盾と中立の区別が難しい問題に深層格が効果を発揮した可能性があると言える。

具体例を図 3 に示す。＜主格＞＜述語動詞＞タグの部分は一致しており、＜場所格＞タグのみが差分

6）4bit 量子化を施したモデルで推論した。
7）https://huggingface.co/llm-jp/llm-jp-3-13b-instruct

モデル名	深層格	Acc	矛盾 (734 件)				含意 (353 件)				中立 (1347 件)			
			正解	P	R	F1	正解	P	R	F1	正解	P	R	F1
GPT-4o	あり	82.99	**643**	87.12	**87.60**	87.35	302	60.40	85.55	70.80	1075	**89.88**	79.80	84.54
GPT-4o	なし	82.08	524	92.74	71.38	80.67	**321**	65.51	**90.93**	**76.15**	1153	83.61	85.59	84.63
GPT-4o	アン	**84.75**	612	92.44	83.37	**87.67**	316	63.83	89.51	74.51	1135	88.88	84.26	**86.50**
llama3.1	あり	74.97	394	97.76	53.67	69.29	261	58.91	73.93	65.57	1170	73.67	86.85	79.71
llama3.1	なし	73.50	378	96.92	51.49	67.25	277	56.07	78.47	65.40	1134	73.16	84.18	78.28
llama3.1	アン	74.36	409	96.69	55.72	70.69	273	55.37	77.33	64.53	1128	74.30	83.74	78.73
llm-jp-3	あり	72.92	536	78.82	73.02	75.80	37	**84.90**	10.48	18.65	1202	70.29	89.23	78.63
llm-jp-3	なし	75.96	356	**97.80**	48.50	64.84	224	74.41	63.45	68.49	**1269**	71.73	**94.20**	81.44
llm-jp-3	アン	77.64	473	92.20	64.44	75.86	164	73.21	46.45	56.83	1253	73.83	93.02	82.32

表 2 実験結果 (アン: アンサンブル, Acc: Accuracy(%), P: Precision(%), R: Recall(%), F1: F1-Score(%), llm-jp-3: llm-jp-3-13b-instruct)

true\pred	矛盾 (件)			含意 (件)			中立 (件)		
手法の深層格使用	あり	なし	アンサンブル	あり	なし	アンサンブル	あり	なし	アンサンブル
矛盾 (734 件)	**643**	**524**	**612**	8	11	10	83	199	112
含意 (353 件)	13	5	7	**302**	**321**	**316**	38	27	30
中立 (1347 件)	82	36	43	190	158	169	**1075**	**1153**	**1135**

表 3 GPT-4o を用いた深層格あり、深層格なし、アンサンブルの 3 手法に対する混合行列

となっている。このように深層格付与によって差分が明確になり、正解が矛盾の問題の精度向上につながったと考えられる。

前提 台所の冷蔵庫の中に < 場所格 > 人が < 主格 > います。< 述語動詞 >
仮定 台所の冷蔵庫やレンジの前の調理台の傍に < 場所格 > 人が < 主格 > います。< 述語動詞 >

図 3 正解：矛盾、深層格なしの出力：中立、深層格ありの出力：矛盾 の例

6.2 深層格付与でも不正解であった事例

次に、「深層格あり手法」が誤答した場合に着目すると、正解が中立の問題を「深層格あり手法」が含意と誤答した数は 190 件と顕著に多い。

具体例を図 4 に示す。深層格付与システムは KNP の係り受け解析結果に依存しており、述語動詞に対して直接係り受け関係がある名詞句に対して深層格を付与する。KNP の解析結果では「おもちゃのハサミで」が「しています。」に対して直接係り受け関係がある名詞句ではなかったため、前提文の該当名詞句に対して < 手段・道具格 > が付与されず、前提文と仮定文で深層格が非一致となっている。< 手段・道具格 > の差分が明確でないため、< 手段・道具格 > に関する含意関係が正しく捉えられなかった可能性がある。

ほかにも、深層格が非一致である場合や < 述語動詞 > が異なる場合に不正解となる傾向があった。付与ルールの改善により深層格付与の間違いを減らす、省略された深層格を補完する、深層格定義セット自体の妥当性検証などが、今後の課題である。

前提 迷彩柄の服を着た女性がおもちゃのハサミで男性のネクタイを切るまねを < 対象格 > しています。< 述語動詞 >
仮定 軍服の女性が巨大なハサミで < 手段・道具格 > 男性のネクタイを < 対象格 > 鋏んでいます。。< 述語動詞 >

図 4 正解：中立、深層格ありの出力：含意 の例

7 結論

本研究では、国語研究所の深層格定義を整理した独自の深層格セットと付与ルールを整備し、ルールベースの深層格自動付与システムを構築した。このシステムを用いて深層格を自動付与したうえで、各種 LLM を用いた JNLI に対する評価実験を行った。

結果、深層格を付与することで前提文と仮定文の差分が明確化され、矛盾ラベルに対する推論性能の向上が見られた。「深層格なし手法」に対し提案手法である「アンサンブル手法」の正解率が GPT-4o で 2.67、llama3.1 で 0.86、llm-jp-3 で 1.68 ポイント向上した。GPT-4o の「アンサンブル手法」では正解率 84.75 で最高性能を示した。これは知る限り執筆時点で State-of-the-Art の性能である。

今後の研究課題としては、付与ルールの改善、省略された深層格の補完、深層格定義セット自体の妥当性検証などが挙げられる。また、自動深層格付与と LLM の組み合わせはタスクを問わず適用可能であり、ほかのタスクでも有効性を検証したい。

謝辞

本研究はJSPS科研費（JP22H00804）、JSTさきがけ（JPMJPR2461）、JST AIP加速課題（JPMJCR22U4）、およびセコム科学技術財団特定領域研究助成の支援をうけた。

参考文献

[1] 河崎工, 木村昌臣. 単語の分散表現を利用した深層格推定手法の提案. 日本知能情報ファジィ学会 ファジィ システム シンポジウム 講演論文集, Vol. 34, pp. 100–103, 2018.

[2] 国立国語研究所. 日本語における表層格と深層格の対応関係. 国立国語研究所, 1997.

[3] 渋木英潔, 荒木健治, 桃内佳雄, 栃内香次. 単語概念の深層格選好に基づく深層格推測手法. 電子情報通信学会論文誌, Vol. J89–D, No. 6, pp. 1413–1428, 2006.

[4] Filmore C.J. **Some problems for case grammar**. Monograph Series on LaRguages and Linguistics 24（Washington D．C．：Georgetown University Press, 1971.

[5] 吉越卓見, 河原大輔, 黒橋禎夫. 機械翻訳を用いた自然言語推論データセットの多言語化. 情報処理学会 第 244 回自然言語処理研究会, Vol. 2020-NL-244, No. 6, pp. 1–8, 7 2020.

[6] 谷中瞳, 峯島宏次. Jsick: 日本語構成的推論・類似度データセットの構築. 人工知能学会全国大会論文集, Vol. JSAI2021, pp. 4J3GS6f02–4J3GS6f02, 2021.

[7] 栗原健太郎, 河原大輔, 柴田知秀. Jglue: 日本語言語理解ベンチマーク. 自然言語処理, Vol. 30, No. 1, pp. 63–87, 2023.

[8] 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.

[9] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In Lluís Màrquez, Chris Callison-Burch, and Jian Su, editors, **Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing**, pp. 632–642, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[10] Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. A SICK cure for the evaluation of compositional distributional semantic models. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, **Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)**, pp. 216–223, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA).

[11] OpenAI : Aaron Hurst and et al. Gpt-4o system card, 2024. 2410.21276 https://arxiv.org/abs/2410.21276.

[12] llm-jp-eval リーダーボード Neo,（2024-12 閲覧）. https://wandb.ai/wandb-japan/llm-leaderboard/reports/Nejumi-LLM-Neo--Vmlldzo2MTkyMTU0.

[13] 河原大輔, 黒橋禎夫. 自動構築した大規模格フレームに基づく構文・格解析の統合的確率モデル. 自然言語処理, Vol. 14, No. 4, pp. 67–81, 2007.

[14] 遼平笹野, 禎夫黒橋, Sasano Ryohei, Kurohashi Sadao. 大規模格フレームを用いた識別モデルに基づく日本語ゼロ照応解析. 情報処理学会論文誌, Vol. 52, No. 12, pp. 3328–3337, 12 2011.

[15] 一森田, 禎夫黒橋. Rnn 言語モデルを用いた日本語形態素解析の実用化. 第 78 回全国大会講演論文集, Vol. 2016, No. 1, pp. 13–14, 03 2016.

[16] Ryosuke Ishigami. cyberagent/llama-3.1-70b-japanese-instruct-2407, 2024.

A 付録

以下は、タスクを説明する指示と、文脈のある入力の組み合わせです。要求を適切に満たす応答を書きなさい。

指示
前提と仮説の関係を entailment、contradiction、neutral の中から回答してください。それ以外には何も含めないことを厳守してください。

制約：
- 前提から仮説が、論理的知識や常識的知識を用いて導出可能である場合は entailment と出力
- 前提と仮説が両立しえない場合は contradiction と出力
- そのいずれでもない場合は neutral と出力

注意：
- 前提と仮定の双方に <> で囲われたタグがあるが、述語と名詞の意味的な関係を表す文法上のカテゴリーである深層格を示している。この深層格も根拠として判定せよ。

入力:
前提：何種類ものケーキが < 主格 > テーブルに < 対象格 > 並んでいます。< 述語動詞 >
仮定：テーブルの上にたくさんのケーキが < 主格 > あります。< 述語動詞 >

応答:

図 5 提案手法のプロンプトテンプレート（入力部分を変更して利用）