LLM を用いた知識補完の統合による 自然言語推論システム lightblue の拡張

富田 朝 [1]　戸次 大介 [1]
[1] お茶の水女子大学
{tomita.asa, bekki}@is.ocha.ac.jp

概要

本研究では、形式意味論に基づく自然言語推論システムにおいて課題となる語彙知識の不足に対処するため、大規模言語モデル (LLM) を用いた公理自動生成手法を提案する。提案手法は、論理推論システム lightblue に LLM を統合し、推論過程で必要な語彙的知識のみを公理として動的に生成・補完することで、不要な公理の補完を抑制しつつ、人手を介さない自動的な知識補完を実現する。評価実験の結果、提案手法の有効性および論理推論における拡張性が示された。

1　はじめに

形式意味論に基づく自然言語推論 (Natural Language Inference; NLI) システムは、文の意味を論理式として明示的に表現することで、厳密で説明可能な推論を実現する枠組みである。中でも、組合せ範疇文法 (Combinatory Categorial Grammar; CCG) [1, 2] と高階論理に基づく推論システム [3, 4, 5] は、証明可能性に基づいて推論を行うことで、高い適合率 (Precision) を有する点に特徴がある。しかし、従来の論理推論システムは、語彙間の関係性に関する知識や世界知識に依存する推論を十分に扱えないという課題を残している。

論理推論システムにおいて、語彙的な推論や世界知識を扱うためには、推論過程で利用可能な公理を明示的に追加した上で推論を行う必要がある。しかし、推論システムへの公理の追加には、膨大な知識データセットの中から推論に必要な知識のみを適切に抽出することが困難であるという、いわゆるフレーム問題が存在する。特に、不要な公理の導入は探索空間の拡大を引き起こし、推論効率や正確性に影響を与える可能性があるため、知識補完の制御は実用的な論理推論システムにおける重要な課題である。

この課題に対して、これまで様々な公理生成手法が提案されてきた。例えば、WordNet [6] のような知識データベースを利用し、単語間の意味的関係性に基づいて公理をオンデマンドに生成する手法 [7] や、語や句の分散表現に基づく類似度を計算し、意味的に関連する表現間に公理を導入する手法などが存在する [8]。また別のアプローチとして、人間が推論過程に介入し、公理を逐次補完する対話的な自然言語推論システムも提案されている [9]。この手法では、論理推論システムを大規模な知識データベースに直接接続するのではなく、推論に必要な知識を人間が判断し、公理として追加することで、不要な知識の導入を抑制できるという利点がある。一方で、人間との対話を前提とするため、推論プロセスの完全な自動化が困難であるという制約がある。

そこで本研究では、人間の代替として大規模言語モデル (Large Language Models; LLM) を用い、論理推論システムに対して公理を自動的に追加する手法を提案する。本手法では、論理推論システムとして *lightblue* [10] を用い、定理証明器として *wani* [11] を採用する。LLM を推論ループに組み込むことで、推論に必要な語彙的な知識を状況に応じて補完しつつ、不要な公理の補完を抑制することを目指す。このような枠組みにより、知識補完の制御という課題に対して一つの実践的な解を与えるとともに、論理推論に基づく高い説明可能性を維持したまま、語彙的推論能力の向上を図ることが可能となる。本研究では、評価実験を通じて提案手法の有効性を検証する。

2　推論システム lightblue

本研究では論理推論システムとして lightblue を用いる。lightblue は、統語理論として組合せ範疇文法、意味の理論として依存型意味論 [12] を採用して

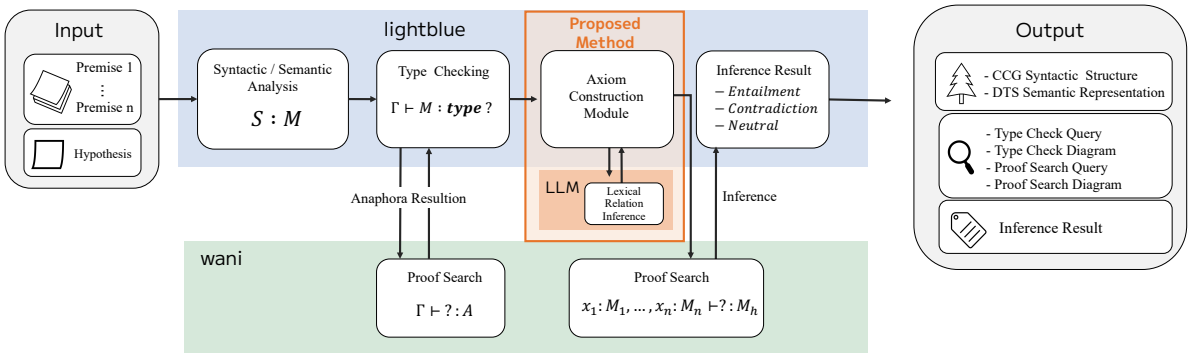

図 1　lightblue を用いた LLM-in-the-loop inference の構成図

おり、統語・意味解析器 [13] に依存型理論 [14] の定理証明器 wani を接続することで自然言語推論を行う。lightblue の特徴として、採用する意味理論である依存型意味論が証明論的意味論に基づいている点が挙げられる。これにより、照応解析や前提解決に加え、含意関係の判定そのものを証明探索という単一の操作に帰着できる。さらに、意味表示に対して型検査を行うことで、意味表示の不整合を検出し、システム全体の意味的整合性を保証することができる。

3　提案手法

システムの構成図を図 1 に示す。まず、前提文と仮説文を入力として受け取り、lightblue を用いて統語・意味解析を行う。続いて、得られた意味表示に対して型検査を行い、意味表示の整合性を検証する。この過程で、意味表示に未指定項が含まれる場合、すなわち照応解決・前提束縛が必要な場合には、定理証明器 wani を用いて証明探索を行うことで先行詞を同定する。従来の lightblue では、型検査が成功した後、前提文と仮説文の意味表示を定理証明器に渡すことで自然言語推論を行う。これに対し本研究では、この処理の中間に公理生成モジュールを組み込み、推論に必要な公理を動的に生成した上で、推論を行うようシステムを拡張する。

3.1　公理生成モジュール

公理生成の流れを図 2 に示す。以下、前提文に「太郎がりんごを食べた」、仮説文に「太郎がフルーツを食べた」が与えられた場合を例として取り上げ、この推論に必要となる公理の生成過程を説明する。本モジュールでは、lightblue による統語・意味解析および型検査の結果を用いて、推論に必要となる語彙間の意味関係を調べるための語彙の組（以下、Subgoal）を抽出し、公理を生成する。

3.1.1　Subgoal の特定

まず、前提文および仮説文の意味表示に含まれるすべての定項 (constant symbols) について、その型 (signature) と対応する統語的型 (syntactic type) を抽出する。次に、型と統語的型が一致する語の組を候補として列挙し、語彙間の意味関係を調べるための Subgoal を構成する。

この際、表層形が完全に一致する語の組は候補から除外する。これにより、前提文と仮説文の双方に出現する固有名詞など、包含関係が自明な語を除外できる。また、定項の型とその統語的型の一致に基づいて候補を選択することで、同一の統語的型を持つ語同士、すなわち品詞が一致する語の関係性のみに限定して語彙的な推論を行うことが可能となる。さらに、各語の組に対して組の要素を順序を反転させた組を追加することで、一方向の含意関係に限らず、その逆方向の関係性についても判定を行う。これにより、「りんご」と「フルーツ」の関係だけでなく、「フルーツ」から「りんご」への推論が必要となる場合にも対応できる。

3.1.2　知識の獲得と公理生成

続いて、各 Subgoal に対して大規模言語モデル (LLM) を用いて語彙的な包含関係の判定を行う。具体的には、語のペア (A, B) について、「A は B に包含されるか」という問いを与え、含意 (Yes)、矛盾 (No)、中立 (Unknown) のいずれかに分類させる。

LLM の判定結果に基づき、含意と判定された場合には含意公理を、矛盾と判定された場合には否定公理を生成する。中立と判定された場合には公理を生成しない。たとえば、LLM が「りんご」は「フルーツ」に意味的に包含されると判定した場合

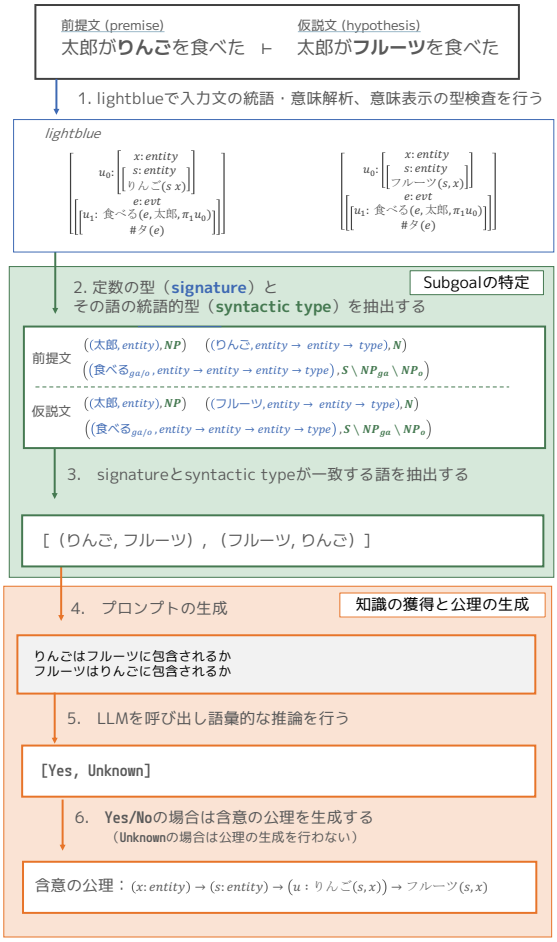

図 2　公理生成の流れと例

は、$(x : entity) \rightarrow (s : entity) \rightarrow (u : りんご(s,x)) \rightarrow フルーツ(s,x)$ という公理が追加される。もし、矛盾と判断された場合には、矛盾の公理として、$(x : entity) \rightarrow (s : entity) \rightarrow (u : りんご(s,x)) \rightarrow \neg フルーツ(s,x)$ が追加される。生成された公理群は、前提文・仮説文の意味表示とともに定理証明器に与えられ、自然言語推論に用いられる。

4　評価実験

提案手法の性能を (1) 大規模言語モデル (LLM) が語彙的な関係性を正しく判定できるか、(2) 生成された公理を用いることで論理推論システムが推論を正確に行えるか、という観点から評価する実験を行う。実験には GPT-4o を使用した。

4.1　データセット

本実験では、含意関係認識のデータセットである SICK [15] を専門家が日本語に翻訳したデータセット JSICK [16] を用いる。実験対象として、JSICK の train セットに含まれる約 5,000 件の文対のうち、前提文と仮説文の差分が名詞の置換のみであるデータ 310 件から、ランダムに 50 件をサンプリングした。さらに本研究では、語彙的推論の評価に焦点を当てるため、前提文および仮説文に含まれる普通名詞の一部を固有名詞に変換している。これは、普通名詞が存在量化を伴う解釈を持つ場合に生じる指示対象のずれを抑制し、語彙間の関係性に基づく推論のみを評価するためである。本研究で使用したデータセットに含まれる文対の例を表 1 に示す。

4.2　LLM の語彙的関係推論能力の評価

4.2.1　実験設定

提案手法に基づき抽出された Subgoal 191 組に対して、上位語・下位語関係に基づく 3 値の正解ラベル (Yes/No/Unknown) を人手で付与したデータを正解データとして使用し、LLM よる語彙間の関係性判定の正答率を評価した。

4.2.2　結果

結果を表 2 に示す。本実験の結果、LLM は語彙間の包含関係が明示的な場合には一定の精度で判断できる一方、矛盾と中立の判定において混同が生じやすいことが確認された。特に No（矛盾）クラスでは Precision および F1 が低く、矛盾関係の判定が不安定であることが示されている。

公理は Yes および No の判定結果に基づいて生成されるため、語彙間の関係性の判定誤りは誤った公理の生成につながる。実験結果より、Yes クラスは Recall が比較的高く、包含関係を広く検出できている一方で、Precision は限定的であり、必ずしもすべての含意公理が高い信頼性を持つとは言えないことが示唆される。一方で No クラスでは、中立の関係を矛盾と誤判定する例が存在しており、誤った矛盾公理が生成される可能性が高い。また、Unknown に分類されるべき語彙ペアに対して Yes または No の判定がなされる場合、公理が過剰に生成される要因となる。以上より、LLM を用いた語彙関係判定は含意公理の候補生成に一定の有効性を持つものの、その出力を無条件に公理として採用することは危険であり、特に矛盾公理の生成は制御する必要がある。

4.3　公理生成にともなう推論性能の評価

LLM が出力した語彙間の関係性の判定をもとに生成された公理を lightblue に組み込み、定理証明器

表 1 JSICK データセットから名詞の置換のみが行われている文対を抽出したデータの例。T から始まる文が前提文、H から始まる文が仮説文である。Subgoal の ✓ は含意、× は矛盾の関係を表している。

文のペア	ラベル	Subgoal	生成された公理
T: 花子が馬に乗っている **H:** 花子が動物に乗っている	Entailment	(馬, 動物) ✓ (動物, 馬)	$(x : entity) \rightarrow (s : entity) \rightarrow (u :$ 馬$(s,x)) \rightarrow$ 動物(s,x)
T: 太郎がジャガイモを切っている **H:** 太郎がトマトを切っている	Neutral	(ジャガイモ, トマト) (トマト, ジャガイモ)	
T: 太郎はインタビューを拒否している **H:** 太郎はインタビューを許可している	Contradiction	(許可, 拒否) × (拒否, 許可) ×	$(x : entity) \rightarrow (s : entity) \rightarrow (u :$ 許可$(s,x)) \rightarrow \neg$拒否(s,x) $(x : entity) \rightarrow (s : entity) \rightarrow (u :$ 拒否$(s,x)) \rightarrow \neg$許可(s,x)

表 2 LLM による語彙間の関係性判定の性能の結果

Label	Precision	Recall	F1
Yes	0.636	0.840	0.724
No	0.361	0.722	0.481
Unknown	0.939	0.764	0.843
Macro avg	0.645	0.776	0.683
Micro avg	0.770	0.770	0.770

を用いて推論を行うことで、提案システム全体としての推論性能を評価する。特に、公理の動的追加が語彙的推論を含む推論課題に与える効果を検証する。

4.3.1 実験設定

提案手法の有効性を測るため、以下の 2 つのベースラインとの比較を行う。

1. **Majority Baseline:** クラス分布の偏りによる影響を確認するため、学習データ中の最頻出ラベル (Neutral) を常に予測する単純なベースライン。
2. ***Vanilla lightblue*:** 公理生成モジュールを使用せず、デフォルトの lightblue のみで推論を行う設定。

評価には、前節で用いた JSICK 名詞置換データ 50 件を用いる。各文対に対して、まず lightblue により統語・意味解析および型検査を行い、提案手法においては必要に応じて公理生成モジュールを適用する。評価指標としては、正解ラベルに対する正解率 (Accuracy) に加え、適合率 (Precision)、再現率 (Recall) を用いる。

4.3.2 結果

結果を表 3 に示す。提案手法は Accuracy 0.72 を達成し、 Majority Baseline および Vanilla lightblue を上回る性能を示した。この結果から、LLM の判定に基づいて生成された公理を推論過程に組み込むことで、語彙的な推論を含む文対に対する全体的な推論性能が向上することが確認された。特に、公理生成を行わない Vanilla lightblue と比較して精度が向上している点は、LLM による知識補完が有効に機能したことを示唆している。前節（4.2 節）で確認された通り、LLM による語彙関係判定には誤りも含まれるものの、定理証明システムに統合することで、全体として語彙的な推論能力の向上に寄与することが確認された。

表 3 自然言語推論性能の比較結果

Method	Precision	Recall	Accuracy
Majority Baseline	–	–	0.580
Vanilla lightblue	0.898	0.262	0.600
提案手法	**0.919**	**0.338**	**0.720**

5 おわりに

本研究では、論理推論システムでの推論時に LLM を用いて公理を動的に生成・追加する手法を提案した。提案手法では、論理推論システム lightblue と定理証明器の推論ループに LLM を組み込むことで、語彙知識に基づく包含関係の公理を必要に応じて補完する枠組みを実現した。評価実験の結果、提案手法は公理生成を行わない Vanilla lightblue を上回る推論性能を示した。特に、LLM によって生成された語彙的公理が、論理推論における含意導出を効果的に補完していることが確認された。

本研究では、論理推論と LLM を対立的に捉えるのではなく、相補的に統合することで、語彙的推論を含むより実用的な自然言語推論が可能となることを示した。今後の課題としては、名詞間の包含関係にとどまらず、述語間の意味的関係性を扱えるように公理生成手法を拡張することが挙げられる。

謝辞

本研究は、JST 国家戦略分野の若手研究者及び博士後期課程学生の育成事業（博士後期課程学生支援）JPMJBS2404、JST 戦略的創造研究推進事業 CREST JPMJCR2565、および JSPS 科研費 JP23H03452 の支援を受けたものである．

参考文献

[1] Mark Steedman. **Surface Structure and Interpretation**. The MIT Press, Cambridge, 1996.

[2] Mark Steedman. **The Syntactic Process**. MIT Press, 2000.

[3] Pascual Martínez-Gómez, Koji Mineshima, Yusuke Miyao, and Daisuke Bekki. ccg2lambda: A compositional semantics system. In Sameer Pradhan and Marianna Apidianaki, editors, **Proceedings of ACL-2016 System Demonstrations**, pp. 85–90, Berlin, Germany, August 2016. Association for Computational Linguistics.

[4] Lasha Abzianidze. LangPro: Natural language theorem prover. In Lucia Specia, Matt Post, and Michael Paul, editors, **Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations**, pp. 115–120, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[5] Hai Hu, Qi Chen, Kyle Richardson, Atreyee Mukherjee, Lawrence S. Moss, and Sandra Kuebler. MonaLog: a lightweight system for natural language inference based on monotonicity. In Allyson Ettinger, Gaja Jarosz, and Joe Pater, editors, **Proceedings of the Society for Computation in Linguistics 2020**, pp. 334–344, New York, New York, January 2020. Association for Computational Linguistics.

[6] George A. Miller. WordNet: A lexical database for English. In **Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992**, 1992.

[7] Pascual Martínez-Gómez, Koji Mineshima, Yusuke Miyao, and Daisuke Bekki. On-demand injection of lexical knowledge for recognising textual entailment. In Mirella Lapata, Phil Blunsom, and Alexander Koller, editors, **Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers**, pp. 710–720, Valencia, Spain, April 2017. Association for Computational Linguistics.

[8] Hitomi Yanaka, Koji Mineshima, Pascual Martínez-Gómez, and Daisuke Bekki. Acquisition of phrase correspondences using natural deduction proofs. In Marilyn Walker, Heng Ji, and Amanda Stent, editors, **Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)**, pp. 756–766, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[9] Atsushi Sumita, Yusuke Miyao, and Koji Mineshima. Talking with the theorem prover to interactively solve natural language inference. In Kaibao Hu, Jong-Bok Kim, Chengqing Zong, and Emmanuele Chersoni, editors, **Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation**, pp. 411–420, Shanghai, China, 11 2021. Association for Computational Linguistics.

[10] Asa Tomita, Mai Matsubara, Hinari Daido, and Daisuke Bekki. Natural language inference with CCG parser and automated theorem prover for DTS. In Timothée Bernard and Timothee Mickus, editors, **Proceedings of the Second Workshop on the Bridges and Gaps between Formal and Computational Linguistics (BriGap-2)**, pp. 1–7, Düsseldorf, Germany, September 2025. Association for Computational Linguistics.

[11] Hinari Daido and Daisuke Bekki. Development of an automated theorem prover for the fragment of DTS. In **the 17th International Workshop on Logic and Engineering of Natural Language Semantics (LENLS17)**, 2017.

[12] Daisuke Bekki and Koji Mineshima. **Context-Passing and Underspecification in Dependent Type Semantics**, pp. 11–41. Springer International Publishing, Cham, 2017.

[13] Daisuke Bekki and Ai Kawazoe. Implementing variable vectors in a CCG parser. In **Logical Aspects of Computational Linguistics. Celebrating 20 Years of LACL (1996–2016)**, pp. 52–67, Berlin, Heidelberg, 12 2016. Springer Berlin Heidelberg.

[14] Per Martin-Löf. **Intuitionistic Type Theory Vol. 1**. Bibliopolis, 1984.

[15] 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).

[16] Hitomi Yanaka and Koji Mineshima. Compositional evaluation on Japanese textual entailment and similarity. **Transactions of the Association for Computational Linguistics**, Vol. 10, pp. 1266–1284, 2022.