ツリーバンクの言語学的妥当性の自動評価

富田 朝 [1]　谷中 瞳 [2]　戸次 大介 [1]
[1] お茶の水女子大学 [2] 東京大学
{tomita.asa, bekki}@is.ocha.ac.jp
hyanaka@is.s.u-tokyo.ac.jp

概要

自然言語推論において、理論言語学に基づく解析手法では、推論の精度は統語解析の妥当性に依存する。統語解析の妥当性を保証するには、パーザの学習・評価に用いるデータが言語学的に妥当であることが重要であるが、統語構造を提供するツリーバンクの妥当性評価は十分に行われておらず、従来の手法では言語学的妥当性を適切に捉えることができない。そこで本研究では、理論言語学と型理論を基盤とする新たな評価手法を提案し、この評価手法を用いてツリーバンクの妥当性評価を行う。

1　はじめに

自然言語処理のさまざまなタスクの中でも、推論は重要な役割を果たしている。近年、大規模言語モデル (LLM) による推論の精度は目覚ましい向上を見せている [1, 2] 一方で、言語学的アプローチに基づく自動推論の研究も進展してきている [3, 4, 5]。とくに、計算可能な自然言語の意味論として依存型意味論 [6] が登場し、依存型の自動証明器である wani [7] の開発に成功するなど、理論言語学に基づいた言語処理の応用可能性が広がっている。このような理論言語学に基づく自動推論では、統語論や意味論の分析を推論プロセスに統合することで、推論結果だけでなく、その過程を証明図として明確に提示することが可能である。ここで出力される証明図は、推論結果の妥当性を保証できるというだけでなく、仮に推論の結果が誤っていた場合に、誤りの原因を検知することができるため、推論システムのさらなる改善へとつなげられるという利点もある。

推論の精度は、推論を支える前処理である統語解析や意味解析の出力に大きく影響を受ける。とくに、誤りを含む統語解析や意味解析は、誤った推論結果へとつながるため、統語・意味解析を通して言語学的に妥当な統語構造と意味表示を得ることは、推論精度の改善に不可欠な要素である。しかし、統語構造や意味表示の正しさは、統語解析器や意味解析器が高い精度を示すこと自体によっては保証されない。現に、現在の統語解析器 [8, 9] の精度は、評価用データセットに対しては高い精度を実現しているものの、誤りが含まれたデータセットで学習・評価を行っているために、受身・使役をはじめとした複雑な構文に対しては、出力に誤りが含まれることが知られている [10]。

一般に、統語解析器の学習・評価データセットとして、統語ラベル付きコーパスであるツリーバンクが用いられる。さまざまな言語や形式文法のツリーバンク [11, 12, 13, 14, 15, 16] が開発されるとともに、ツリーバンクに含まれる語彙の網羅性や言語処理タスクでの評価実験が行われている。一方で、ツリーバンクのデータが言語学的にどれほど妥当であるかという観点からの評価が十分に行われていない。

本研究では、ツリーバンクの言語学的妥当性を評価手法として、統語論に基づいたツリーバンクの評価手法と、意味論に基づいたツリーバンクの評価を提案する。また、これらの評価手法を用いて、Tomita et al. [17] で構築された CCG ツリーバンクの言語学的妥当性の評価実験を行い、結果を議論する。

2　CCG ツリーバンクの構築

2.1　組合せ範疇文法 (CCG)

理論言語学に基づいた自動推論の統語解析では、自然言語を形式文法に基づいた統語構造へと変換する。ここで用いられる形式文法として、文脈自由文法 (Context Free Grammar; CFG) や範疇文法 (Categorial Grammar) などがあげられる。中でも、範疇文法の一種である組合せ範疇文法（Combinatory Categorial Grammar; CCG）[18, 19] は、チョムスキー階層の 1.5 型文法である弱文脈依存文法の中で最も

弱い生成能力を持つという特徴があり、自然言語の必要十分な記述に適している。CCG に基づいたツリーバンクとしては、英語の CCGbank [12] や、日本語の日本語 CCGbank [20] などがあげられる。これらは、文脈自由文法のツリーバンク [11] や係り受け構造のツリーバンクからの自動変換によって構築されており、既存の統語解析器 [21, 22, 23, 9, 24, 25] の学習・評価用データとして活用されている。

2.2 言語学的に妥当なツリーバンク

日本語 CCGbank には、受身や使役などの格交替が起こる文の分析に誤りがある [10]。また、用言の活用形をはじめとした統語情報が不足しているという課題もある。そこで、言語学的に妥当な日本語 CCG ツリーバンクの構築手法として、日本語統語解析器 lightblue [26] を用いた Reforging（2.2.2 に後述）を提案し、新たに言語学的な妥当性に着目した日本語 CCG ツリーバンク lightblue CCGbank を構築した [17]。

2.2.1 日本語統語解析器 lightblue

lightblue [26] は CCG に基づく統語構造と依存型意味論に基づく意味表示を並行して計算することのできる解析器である。CCG ツリーバンクを学習データに用いたニューラル統語解析器 [9] とは異なり、あらかじめ形態素解析器 Juman [27] の辞書データから変換した語彙を含む CCG の辞書と、CCG の組合せ規則を用いて統語・意味解析を行うため、学習データとしてツリーバンクを必要としないという特徴がある。さらに、lightblue の辞書には、用言の活用形や活用の種類、テンスをはじめとした、詳細な統語素性の情報を含むこと、CCG 統語構造と並行して、依存型意味論に基づく意味表示を出力できること、などの利点もある。

2.2.2 Reforging の概要

lightblue は上記の利点がある一方で、Juman から変換した辞書に用言の項構造の誤りが含まれるという課題を残している。そこで、先行研究 [17] では、lightblue を用いた解析時に、他の言語資源 [16, 28] から用言の項構造を抽出し、lightblue の語彙項目を削除・追加する項構造合成モジュールを作成した。この項構造合成モジュールと、解析器 lightblue を用いたツリーバンクの構築手法を Reforging [17] と呼び、それによって言語学的に妥当な日本語 CCG ツリーバンク lightblue CCGbank の構築を試みた。lightblue CCGbank は、ABC ツリーバンク [16] から抽出した 13,653 文に、CCG 統語構造および依存型意味論に基づく意味表示を付与したデータセットである。残る課題は、lightblue ツリーバンクの言語学的妥当性をどのように評価するかである。

3 言語学的妥当性の評価手法

3.1 従来のツリーバンクの評価指標

一般に、ツリーバンクの評価には、辞書のエントリ数、辞書被覆率、パーザの解析精度などの指標が用いられる。この章では、これらの従来の評価指標はいずれも完全なものではない、ということを指摘する。

3.1.1 辞書のエントリ数・被覆率

ツリーバンクの葉ノードの語から辞書を構築し、そのエントリ数や辞書被覆率を測定することで、ツリーバンクの網羅性を評価できる。辞書被覆率とは、未知の形態素が正しく辞書に登録されている割合のことである。

しかし、被覆率が高いことは、データセットの妥当性を保証するものではないということに注意したい。被覆率は、辞書に登録された形態素がどれだけ未知語に対して適切に対応しているかを示す指標であり、ツリーバンクのデータ数が十分であることを示すものに過ぎず、ツリーバンクのデータが妥当であるかどうかについては、被覆率では判断できない。したがって、被覆率が高いからといって、データの品質や妥当性は保証されない。

3.1.2 パーザーの解析精度

パーザーを用いた解析精度での評価では、パーザーに訓練データとしてツリーバンクを学習させ、与えられた文に対してパーザーがどれだけ正確に統語構造を解析できるかの評価を行う。Evalb[1)] などのソフトウェアを用いて precision, recall, F 値, タグ付けの accuracy などを測定する。

しかし、パーザの解析精度が高いことと、データセットが妥当であることは異なる概念である。解析精度は、学習データを用いて訓練されたパーザが、評価データに対してどれだけ正解データに近い出力を生成できるかを示す指標である。しかし、学

1） https://nlp.cs.nyu.edu/evalb/

習データや評価データに誤りが含まれている場合、パーザはその誤った情報に基づいて高精度な解析を行うことができるが、その解析結果が実際のデータの妥当な構造や意味を反映しているとは限らない。したがって、パーザの解析精度が高い場合でも、それがデータセットの妥当性を保証するものではない。

3.2 提案手法

3.1 で見てきたように、従来のツリーバンク評価手法では、言語学的な妥当性の定量評価を行う手法としては不十分である。

また、CCG の統語構造と依存型意味論の意味表示の評価には理論言語学の高度な知識が必要になるため、人手での評価のコストは高く、大規模なツリーバンクの評価方法としては不適切である。そこで、本研究では、データ数の多いツリーバンクの言語学的妥当性を自動で評価する手法を提案する。lightblue CCGbank では、1 つの文に対して CCG 統語構造と依存型意味論の意味表示が付与されているため、統語論の側面からの統語構造の評価指標と、意味論の側面からの意味表示の評価指標を新たに考案し、この 2 つの評価値を組み合わせることで、多角的な評価を可能にする。

3.2.1 統語論に基づいた評価

lightblue CCGbank の全ての文は、ABC ツリーバンクから抽出されているため、lightblue CCGbank のそれぞれの統語構造に、対応する ABC ツリーバンクの統語構造が存在する。そこで、言語学者によるアノテーションによって構築された ABC ツリーバンクの統語構造が妥当であると仮定し、lightblue CCGbank の統語構造と ABC ツリーバンクの統語構造の一致度をスコア化することで、統語構造の信頼性を評価する。

ABC ツリーバンクで用いられる ABC 文法は、関数適用規則に関数合成規則を加えた範疇文法である。ABC 文法と CCG では、統語範疇の定義や unary 規則[2)]の定義が異なるため、そのまま比較することはできない。そこで、ABC ツリーバンクに含まれる統語範疇を CCG て使われる統語範疇へ変換し、lightblue CCGbank との一致度をスコア化することで評価を行う。スコアは以下の手順で計算する。

2) ABC ツリーバンクでは、名詞の名詞句化を $N \Rightarrow NP$ という単項規則で表現する一方、lightblue では、空範疇である存在量化詞 $\exists \vdash T/(T\backslash NP)/N$ と関数適用規則で表現する

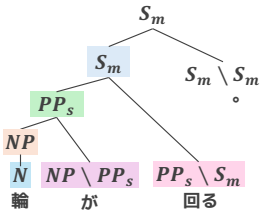

図 1 変換前の統語構造

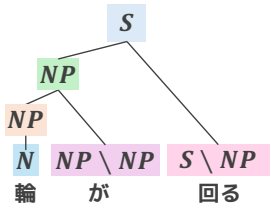

図 2 CCG へ変換した構造

図 3 ABC ツリーバンクの統語構造の変換

```
[(N, 輪),(NP, 輪),(NP\NP, が),
(NP, 輪が),(S\NP, 回る),(S, 輪が回る)]
```

図 4 図 2 のカテゴリと表層形のペアのリスト

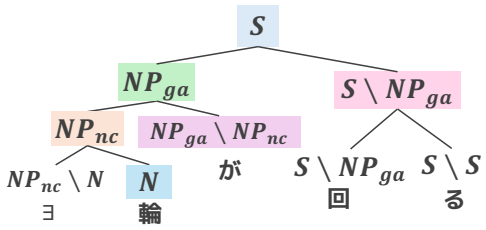

図 5 lightblue CCGbank の統語構造

```
[(NP\N, ヨ),(N, 輪),(NP, 輪),
(NP\NP, が),(NP, 輪が),(S\NP, 回),
(S\S, る),(S, 輪が回る)]
```

図 6 図 5 の表層形とカテゴリのペアのリスト

1. ABC 文法を CCG へ変換する (図 3)
2. 1 の統語構造と lightblue CCGbank の統語構造のそれぞれについて、統語構造の統語範疇と表層形のペアのリストを作成する (図 4, 図 6)
3. ABC ツリーバンクのリスト (図 4) の要素のうち、lightblue CCGbank のリスト (図 6) に含まれる要素の割合をスコアとする

この評価手法の利点として、CCG の空範疇と ABC ツリーバンクの unary 規則の比較ができること、用言の分析における差異[3)]にも対応できること、などがあげられる。一方で、ABC ツリーバンクの全ての統語構造が正しいとは限らない中で ABC ツリーバンクを正解データとして仮定する必要があることや、ABC ツリーバンクに含まれない統語素性の評価ができない点が課題として残る。

3) 「走る」という用言の分析を、CCG では、語幹「走」＋活用語尾「る」とする一方、ABC ツリーバンクでは、「走る」で一つの動詞として分析する

3.2.2 意味論に基づいた評価

lightblue CCGbank に含まれるすべての統語構造には、依存型意味論に基づく意味表示が付与されている。この意味表示を活用し、型理論の観点からその妥当性を評価する手法を提案する。

まず、lightblue を用いて型検査を行い、型検査に通過する意味表示の割合を計算する。型検査とは、依存型意味論の意味表示が整合な型を持つかを判定する手続きで、意味表示が type という型を持つことを証明できれば、型検査は成功する。付録の図 8 は型検査が成功した場合の証明図の例である。

型検査は、意味表示が不整合である場合に失敗するが、CCG と依存型意味論を統語論、意味論の体系として使用している場合、意味表示が不整合になることは起こり得ないことが証明されている。つまり、型検査の失敗は、統語構造の実装に誤りがあるということを示唆する。この性質を利用することで、意味表示を合成的に導出できるかという観点で統語構造の正しさを評価することができる。また、統語構造に対して、意味表示のレベルで型理論に基づいた評価を実施できる点も本手法の強みである。

一方で、型検査を通過した意味表示と対になる統語構造が必ずしも妥当であるとは限らないため、この評価指標のみで言語的妥当性を完全に保証することは難しい。より厳密な評価を行うには、3.2.1 の評価手法をはじめ、他の妥当性の評価指標と組み合わせることが望ましい。

4 評価実験

4.1 実験設定

lightblue CCGbank に含まれる各ジャンルからランダムにサンプリングを行い、合計で ？文 を対象に評価を実施した。以下の指標に基づき、統語構造を総合的に評価した。

統語構造のスコア平均 3.2.1 で提案した手法に基づき、統語構造を 100 点満点で評価し、ジャンルごとの平均スコアを算出する。

型検査の通過率 3.2.2 で示した型理論に基づく手法を用い、型検査を通過した割合を計測する。

総合的な評価 統語構造のスコアが 50 点以上かつ型検査に通過したデータの割合を計算する。

4.2 実験結果

実験の結果を表 1 に示す。

表 1 評価結果

ジャンル	データ数	スコア平均	型検査通過率	総合
青空文庫	75	43.1	69.3	21.1
聖書	24	37.8	58.3	17.2
書籍	6	43.0	66.7	25.0
辞書	60	53.0	48.3	25.9
会議録	21	32.9	90.5	16.0
フィクション	18	49.2	77.8	33.3
法律	6	12.0	66.7	0.0
その他	30	49.5	73.3	30.2
ニュース	30	40.7	76.7	11.8
ノンフィクション	6	40.0	100.0	25.0
話し言葉	30	31.8	73.3	16.7
テッドトーク	15	52.9	73.3	28.6
教科書	120	48.1	55.0	23.1
wikipedia	15	37.7	66.7	16.7
Total	456	40.8	64.9	22.3

全体の平均スコアは 40.8 点であり、評価対象の文 456 文中 296 文が型検査に通過し、通過率は 64.9% であった。

スコア平均が最も高かったジャンルは辞書で、53.0%の統語構造が他コーパスと一致した。また、ノンフィクションは 6 文中全ての文が型検査を通過し、14 ジャンルで最も高い通過率となった。総合的な評価が最も高かったフィクションでは、3 割のデータで妥当性が保証された一方、法律、ニュース、聖書などのジャンルでは総合値が伸び悩んだ。この結果は、lightblue がドメイン固有の表現が含まれる文の解析に課題があるということを示唆しており、lightblue の辞書の拡張によって妥当性のさらなる向上が見込まれる。

5 おわりに

本研究では、統語論および意味論に基づいたツリーバンク評価指標を提案し、多角的な観点からツリーバンクの妥当性を自動で評価する枠組みを確立した。これにより、従来の評価指標では捉えきれなかった統語構造の妥当性をより詳細に分析することが可能となった。

妥当性を保証するツリーバンクの整備は、推論精度の向上のみならず、エラー検知や説明可能性の向上といった、今後の言語処理システムに求められる要件を満たすための重要なステップとなる。

今後は、ツリーバンクの構築過程へのフィードバック機構の導入することで lightblue CCGbank のさらなる妥当性の向上を目指す。また、スコア算出方法の改善をはじめとした評価手法自体の改善方法についても検討していく予定である。

謝辞

本研究の一部は、JST CREST JPMJCR20D2、JSPS 科研費学術変革領域研究（B）「ナラティブ意識学」JP24H00809 の支援を受けたものである。

参考文献

[1] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. **CoRR**, Vol. abs/2110.14168, , 2021.

[2] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. **CoRR**, Vol. abs/2201.11903, , 2022.

[3] Koji Mineshima, Pascual Martínez-Gómez, Yusuke Miyao, and Daisuke Bekki. Higher-order logical inference with compositional semantics. In **Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing**, pp. 2055–2061, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[4] Lasha Abzianidze and Johan Bos. Towards universal semantic tagging. In Claire Gardent and Christian Retoré, editors, **Proceedings of the 12th International Conference on Computational Semantics (IWCS) — Short papers**, 2017.

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

[7] 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)**, 2020.

[8] Hiroshi Noji and Yusuke Miyao. Jigg: A framework for an easy natural language processing pipeline. In **Proceedings of ACL-2016 System Demonstrations**, pp. 103–108, Berlin, Germany, August 2016. Association for Computational Linguistics.

[9] Masashi Yoshikawa, Hiroshi Noji, and Yuji Matsumoto. A* CCG parsing with a supertag and dependency factored model. In **Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 277–287, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[10] Daisuke Bekki and Hitomi Yanaka. Is Japanese CCGBank empirically correct? A case study of passive and causative constructions. In **Proceedings of the 21st International Workshop on Treebanks and Linguistic Theories (TLT, GURT/SyntaxFest 2023)**, pp. 32–36, Washington, D.C., March 2023. Association for Computational Linguistics.

[11] Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large annotated corpus of English: The Penn Treebank. **Computational Linguistics**, Vol. 19, No. 2, pp. 313–330, 1993.

[12] Julia Hockenmaier and Mark Steedman. CCGbank: A corpus of CCG derivations and dependency structures extracted from the Penn Treebank. **Computational Linguistics**, Vol. 33, No. 3, pp. 355–396, 2007.

[13] Johan Bos, Cristina Bosco, and Alessandro Mazzei. Converting a dependency treebank to a categorial grammar treebank for Italian. In **Proceedings of the eighth international workshop on treebanks and linguistictheories (TLT8)**, pp. 27–38, Italy, Milan, 2009.

[14] Stephen A. Boxwell and Chris Brew. A pilot Arabic CCGbank. In **Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)**, Valletta, Malta, May 2010. European Language Resources Association (ELRA).

[15] Julia Hockenmaier. Creating a CCGbank and a wide-coverage CCG lexicon for German. In **Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics**, pp. 505–512, Sydney, Australia, July 2006. Association for Computational Linguistics.

[16] Yusuke Kubota, Koji Mineshima, Noritsugu Hayashi, and Shinya Okano. Development of a general-purpose categorial grammar treebank. In **Proceedings of the Twelfth Language Resources and Evaluation Conference**, pp. 5195–5201, Marseille, France, May 2020. European Language Resources Association.

[17] Asa Tomita, Hitomi Yanaka, and Daisuke Bekki. Reforging : A method for constructing a linguistically valid Japanese CCG treebank. In Neele Falk, Sara Papi, and Mike Zhang, editors, **Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop**, pp. 196–207, St. Julian's, Malta, March 2024. Association for Computational Linguistics.

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

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

[20] Sumire Uematsu, Takuya Matsuzaki, Hiroki Hanaoka, Yusuke Miyao, and Hideki Mima. Integrating multiple dependency corpora for inducing wide-coverage Japanese CCG resources. In **Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 1042–1051, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

[21] Stephen Clark and James R. Curran. Wide-coverage efficient statistical parsing with CCG and log-linear models. **Computational Linguistics**, Vol. 33, No. 4, pp. 493–552, 2007.

[22] Mike Lewis and Mark Steedman. A* CCG parsing with a supertag-factored model. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, **Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 990–1000, Doha, Qatar, October 2014. Association for Computational Linguistics.

[23] Mike Lewis, Kenton Lee, and Luke Zettlemoyer. LSTM CCG parsing. In Kevin Knight, Ani Nenkova, and Owen Rambow, editors, **Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 221–231, San Diego, California, June 2016. Association for Computational Linguistics.

[24] Miloš Stanojević and Mark Steedman. Max-margin incremental CCG parsing. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 4111–4122, Online, July 2020. Association for Computational Linguistics.

[25] Yuanhe Tian, Yan Song, and Fei Xia. Supertagging Combinatory Categorial Grammar with attentive graph convolutional networks. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 6037–6044, Online, November 2020. Association for Computational Linguistics.

[26] 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.

[27] Daisuke Kawahara and Sadao Kurohashi. Case frame compilation from the web using high-performance computing. In **Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)**, pp. 1344–1347, Genoa, Italy, May 2006. European Language Resources Association (ELRA).

[28] Nobuhiro Ueda, Kazumasa Omura, Takashi Kodama, Hirokazu Kiyomaru, Yugo Murawaki, Daisuke Kawahara, and Sadao Kurohashi. Kwja: A unified japanese analyzer based on foundation models. In **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics: System Demonstrations**, pp. 538–548, Toronto, Canada, 2023.

A　付録

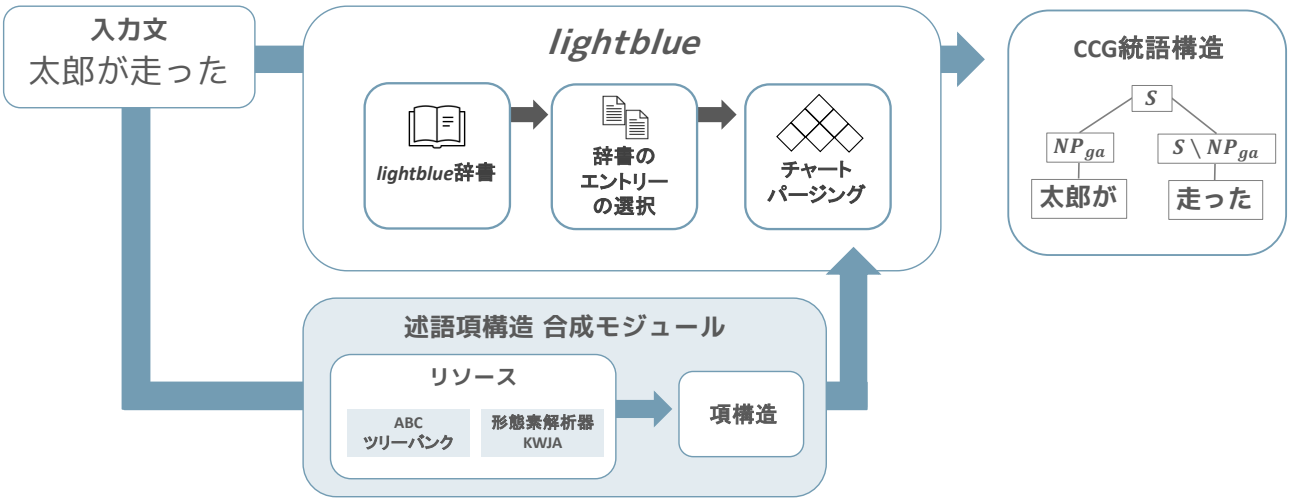

図 7　リフォージングの全体像

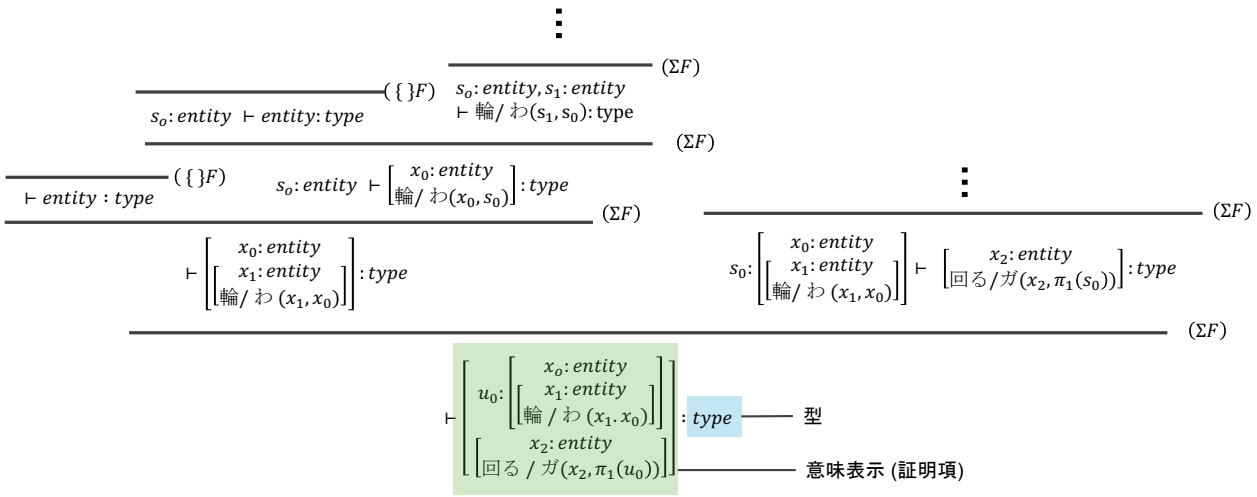

図 8　依存型意味論の型検査図の一部