自然言語推論への応用を志向した セマンティックパージングの性能評価

船蔵颯 [1,2,3]　峯島宏次 [1]
[1] 慶應義塾大学　[2] 京都大学大学院
[3] 株式会社キカガク
funakura.hayate.28p@st.kyoto-u.ac.jp　minesima@abelard.flet.keio.ac.jp

概要

本研究は、セマンティックパージングで広く採用されている、Smatch をはじめとするグラフマッチングによる評価指標が、セマンティックパージングの主たる下流タスクのひとつである自然言語推論における性能を必ずしも保証しないということを示すものである。上記を示すために、ファインチューニングおよび In-context learning に基づくパーザを構築し、グラフマッチングによる評価と自然言語推論を志向した評価をそれぞれのパーザに適用した。結果として、グラフマッチングによる評価と自然言語推論への応用可能性との間にはギャップがあるということを報告する。

1　はじめに

セマンティックパージング (semantic parsing) は、自然言語表現をデータベースクエリや論理式などの構造化された表現に変換するタスクである。解析結果の活用先は文書分類 [1] や質問応答 [2] など多岐にわたり、自然言語推論もそのひとつである [3, 4, 5]。

セマンティックパージングも他の自然言語処理タスクと同様、ニューラルネットワークを用いた手法が台頭している [6, 7]。これらセマンティックパージングの手法の評価指標として、Smatch [8] が広く採用されている。Smatch は二つのグラフがどの程度一致しているかを測る指標であり、主に Abstruct Meaning Representation(AMR)[9] に適用される。AMR は $R(t_1, t_2)$ の形式をとる原子命題の連言とみなすことができる。ある二つの AMR を g_1, g_2 としたとき、g_1, g_2 の Smatch は以下で計算される。

$$2 \cdot \frac{P \cdot R}{P + R}$$

ただし、T_1, T_2 を g_1, g_2 それぞれに含まれる原子命題の集合とし、$M = T_1 \cap T_2$ としたとき、$P = M/|T_2|, R = M/|T_1|$ である。このように、Smatch は二つのグラフの一致度を、各グラフに含まれる原子命題の F-score によって表す。本稿では、Smatch およびその派生系も含めた、グラフ間の一致度による評価を**グラフマッチングによる評価**と呼ぶこととする。

本研究で主眼とするのは、自然言語推論への応用を志向したセマンティックパージングの性能評価である。自然言語推論とは、自然言語で表現されたいくつかの仮定 $\mathrm{P}_1, \ldots, \mathrm{P}_n$ と、仮説 H の論理的な関係（含意、矛盾、中立）を予測するタスクである。$\mathrm{P}_1, \ldots, \mathrm{P}_n$ および H を、それぞれの意味を表す形式言語表現（意味表示）に変換し、得られた意味表示に対して定理証明器を適用することで推論を行う、という手法が検討されてきている [3, 4, 5]。このような手法のコンポーネントとしてのセマンティックパーザには、後段の定理証明において正しい推論結果が導かれるような意味表示の出力が要請される。つまり、ある文 S について、正解の意味表示 $\mathrm{SR_g(S)}$ があるとしたとき、パーザの出力 $\mathrm{SR_p(S)}$ は $\mathrm{SR_g(S)}$ と論理的に等価であるのが理想である。グラフマッチングによる評価は、グラフの構成要素がどの程度一致しているかを表す指標であるから、予測結果と正解の論理的な関係を表すものではない。

そこで本研究は、**グラフマッチングによる評価で高い性能を達成したパーザが、自然言語推論を志向した評価でも高い性能を達成するとは必ずしも言えない**という仮説を検証するための実験を行った。実験では、英語の文を一階述語論理の意味表示に変換するパーザを、グラフマッチングによる評価と自然言語推論を志向した評価の両方で評価した。セマンティックパージングの手法としては、Transformer 型の事前学習済みニューラルネットワー

クをセマンティックパージング用に再学習するファインチューニングと、GPT-4o/4o mini にいくつかの解析例を提示し、その上で未知の文を解析させる In-context learning (ICL) の二つを採用した。実験の結果として、グラフマッチングによる評価と自然言語推論への応用可能性との間にはギャップがあるということを報告する。

2 背景と関連研究

自然言語表現を形式的な意味表現に変換する手法は、記号論理と形式意味論の分野で広く研究されてきた [10]。従来の述語項構造に加えて、否定や量化などのスコープ関係や、テンス、モダリティや談話関係などの談話レベルの豊かな情報を組み込んだ意味表現へ変換する open-domain の解析手法は、CCG などの豊かな構文情報をもつリソースの登場と構文解析技術の発展を経て大きな進展を見せた [11]。さらに、PMB [12] やなどの意味表現が直接付与されたコーパスの開発が進むにつれて、ニューラルネットネットワークに基づく手法、特に系列変換に基づく手法が広く研究され、コード生成 [13]、質問応答 [14]、文生成 [6] などに広く応用されている。

このように当初、セマンティックパージングの研究は、翻訳や要約、質問応答や言い換えなど意味処理を要する様々な下流タスクを実現・改善するのに有効だと考えられており、グラフマッチングに基づく柔軟な評価はこうした幅広い応用を念頭においたものとなっていると言える。一方で、事前学習モデルの登場により、形式的な意味表現がこれら多様な下流タスクに応用される機会は従来よりも減りつつあり、セマンティックパージングの有効性についても議論が続いている [15]。

形式的な意味表現が重要となるタスクの一つに論理的な含意関係の判定がある。従来、記号論理的なアプローチ [10] のもとで研究されてきた論理表現の多くは、自動定理証明 (Automated Theorem Proving) [16, 17] と組み合わせて、正確な論理推論（含意判定や整合性判定）を行うことを念頭において設計されたものである。このパラダイムを直接、含意関係認識に適用した初期の試みとしては [18] がある。また、CCG 構文解析に基づくセマンティックパージングを自動定理証明に基づいて多様な自然言語推論に適用した研究としては、[19] や [3] が挙げられる。一方、ニューラルセマンティックパージングと定理証明を組み合わせた手法が自然言語の推論にどの程度有効であるのかは、われわれの知る限り、まだ十分に解明されていない。

正確な論理推論が必要とされる問題の一つは、数学の定理証明である。自然言語で書かれた証明から定理証明器、及び、Coq や Lean などの証明支援系で処理可能な形式的な表現に変換する手法は、Autoformalization として広く研究されている [20]。数学の証明を典型とする複雑な問題解決に必要な「正確な論理推論」を言語モデルのみで実現できるかどうかは依然として明確ではなく、現状では論理的な推論能力を実現するために、構造的な意味表現に基づく記号処理と統計的な言語モデルを組み合わせた仕組み（Neuro-Symbolic AI）の模索が続いている [21]。

以上の背景をもとに、この研究では、現状のニューラルセマンティックパージングの手法（系列変換、および、LLM の In-context learning）が定理証明と組み合わせて自然言語推論にどの程度有効であるのかを検証することを目的とし、より正確かつ頑健な自然言語推論モデルの構築に寄与することを目指すものである。

3 実験設定

本節では、本研究で実施したファインチューニングおよび ICL による実験それぞれの設定について述べる。

3.1 データセット

本研究では、 SICK [22] をベースとして作成されたデータセットを使用した。SICK は文間類似度および自然言語推論のデータセットであり、量化や否定など論理的な表現を伴う推論を含む点に特徴がある。各サンプルは文ペアおよび文間の類似度 ($\in \{1, 2, 3, 4, 5\}$)、推論関係ラベル ($\in \{\texttt{entailment}, \texttt{contradiction}, \texttt{neutral}\}$) がアノテーションされている。

Haruta ら [4, 23] では、動詞の Event Semantics と形容詞や比較表現の Degree Semantics を組み合わせた 1 階述語論理の拡張系 (Typed First-Order Logic) のもとで、文の意味表示を計算し、自動定理証明に基づいて自然言語推論を行うシステムが実装されている。このシステムは SICK も対象としており、本研究ではこのシステムによる解析結果を正解として使用した。

使用したのは SICK のテストデータであり、上述

のシステムによってセマンティックパージングおよび自動定理証明を行い、推論関係ラベルが正解と一致したもののみを採用した。各サンプルは文とそれに対応する意味表示からなり、全体で 8,140 サンプルである。以下はサンプルの例であり、各 a が自然言語文、b がそれに対応する論理式である。

(1) a. Someone is typing.
b. $\exists x_1 \exists e_2[\text{type}(e_2) \wedge (\text{Subj}(e_2) = x_1)]$

(2) a. There is no clown singing.
b. $\neg\exists x_1(\text{sing}(x_1) \wedge \text{clown}(x_1))$

(3) a. Two young women are sparring in a kickboxing fight.
b. $\exists x_1(\exists d_2 \exists x_3(\text{woman}(x_1) \wedge (d_2 = \text{th}(\text{woman})) \wedge \text{young}(x_1, d_2)) \wedge \text{many}(x_1, 2) \wedge \exists e_4(\text{spar}(e_4) \wedge (\text{Subj}(e_4) = x_1) \wedge \exists x_5(\text{fight}(x_5) \wedge \text{kickboxing}(x_5) \wedge \text{in}(e_4, x_5) \wedge (\text{Subj}(e_4) = x_5))))$

得られたデータセットを $\mathcal{D}$ で表す。

3.2 ファインチューニング

ファインチューニングによる実験では、$\mathcal{D}$ をランダムシャッフルしたのち 8:2 に分割し、前者を学習データ、後者をテストデータとして使用した。学習データを $\mathcal{D}_{\text{train}}$、テストデータを $\mathcal{D}_{\text{test}}$ で表す。

モデルは Transformer 型モデルである Byt5 [24] の Small モデルを使用した。エポックは 20、バッチサイズは 16 に設定し、学習率は 1×10^{-5}、重み減衰は 0.01 とした。それ以外の値は Hugging Face Transformers 4.46.2 のデフォルト値を採用した。

3.3 In-context Learning

ICL による実験では、GPT-4o および GPT-4o mini を使用した。セマンティックパージングを行うよう指示した英文と、$\mathcal{D}_{\text{train}}$ からランダムに取得した 3 つの文・意味表示ペアを Few-shot 事例としてプロンプトに含めた。プロンプトの全文は付録に記載する。

プロンプトの末尾に $\mathcal{D}_{\text{test}}$ の各文を挿入し、セマンティックパージングを行わせた。再現性を担保するため、ランダムシードを 0 に設定した。また、温度パラメータは 0.7 に設定した。

表 1 実験結果

手法	F-score	Acc	Fwd Acc	Bwd Acc	Nonwff
SFC	0.7775	0.3286	0.4369	0.3589	0.1522
4o	0.7958	0.2073	0.3236	0.2401	0.0588
4o mini	0.7780	0.2085	0.3045	0.2438	0.1448

3.4 評価指標

各手法によるセマンティックパージング結果について、グラフマッチングによる評価と自動定理証明による評価を行った。

グラフマッチングによる評価には、Counter [25] を使用した。[1)] Counter は、否定や量化などのスコープに意味があるようなグラフ構造用に Smatch を改変したものである。Counter が対応しているのは談話表示構造 (DRS) であるため、パーザの予測結果を DRS に変換し、正解の DRS と予測結果の DRS との間での F-Score を Counter で算出した。変換の例を (4) に示す。ここでの DRS は変数間の関係を表す `b1 REF x1` などの表現の集まりとして表記され、この表記は clausal form と呼ばれる。

(4) a. $\exists x_1(\text{girl}(x_1) \wedge \exists x_2(\text{horse}(x_2) \wedge \exists e_3(\text{ride}(e_3) \wedge (\text{Subj}(e_3) = x_1) \wedge (\text{Acc}(e_3) = x_2))))$
b.
```
b1 REF x1
b1 REF x2
b1 REF e3
b1 girl x1
b1 horse x2
b1 ride e3
```

自動定理証明による評価には、一階述語論理の定理証明器である Vampire [26] を使用した。[2)] パーザの予測結果が正解と論理的に等価であるか、あるいは過不足があるかを調査するため、予測結果から正解への論理的含意および正解から予測結果への論理的含意が成立するかどうかを調べた。いずれの含意も成立する場合、両者は論理的に等価であり、前者のみが成立する場合は不要な式が予測結果に含まれていること、後者のみが成立する場合は必要な式が予測結果に含まれていないことをそれぞれ意味する。

4 結果と議論

実験の結果を表 1 に示す。ファインチューニングを SFT、GPT-4o を 4o、GPT-4o mini を 4o mini とそれぞれ表記する。最左の値はグラフ間の F-score であり、それ以降は自動定理証明による評価結果である。Acc は予測結果と正解の間で双方向の導出が成立した文の割合、Fwd Acc は正解から予測結果への含意のみが成立する文の割合、Bwd Acc は予測結果から正解への含意のみが成立する文の割合、Nonwff は予測結果が Well-formed formula でない文の割合を表す。

いずれの手法もグラフマッチングによる F-score は 0.8 付近の値である一方で、正解と論理的に等価な意味表示を出力できたのは SFT で全体の 3 割程度、ICL では全体の 2 割程度であった。また、4o は F-score が最も高い値であるにもかかわらず、Acc は最も低い値をとっている。このことは、**グラフマッチングによる評価指標のもとでより高い性能を示す手法が、論理推論においても高い性能を示すわけでは必ずしもない**ということを示唆する。

5 おわりに

本研究では、自然言語を一階述語論理に基づく意味表示へと変換するパーザを構築し、グラフマッチングによる評価と定理証明器による評価を行った。構築した三つのパーザいずれについても、グラフマッチングに基づく F-score は 0.8 程度である一方で、正解と論理的に等価な意味表示を出力できたサンプルはテストデータ全体の半数にも満たないことが示された。この結果は、グラフマッチングによる評価指標のもとでより高い性能を示す手法が、論理推論においても高い性能を示すわけでは必ずしもないということを示唆する。今後は、定理証明器による推論結果を最適化するようなニューラルセマンティックパージングおよび、形式統語論・形式意味論に基づくルール主体のセマンティックパージングの両者を検討する必要がある。

謝辞

本研究は JST CREST、JP-MJCR2114 の支援を受けたものです。

1） https://github.com/RikVN/DRS_parsing?tab=readme-ov-file

2） https://github.com/vprover/vampire

参考文献

[1] Li Dong, Furu Wei, Shujie Liu, Ming Zhou, and Ke Xu. A statistical parsing framework for sentiment classification. **Computational Linguistics**, Vol. 41, No. 2, pp. 293–336, 2015.

[2] Wen-tau Yih, Xiaodong He, and Christopher Meek. Semantic parsing for single-relation question answering. In **Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)**, pp. 643–648, 2014.

[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, 2015.

[4] Izumi Haruta, Koji Mineshima, and Daisuke Bekki. Combining event semantics and degree semantics for natural language inference. In **Proceedings of the 28th International Conference on Computational Linguistics**, pp. 1758–1764. International Committee on Computational Linguistics, 2020.

[5] Hayate Funakura and Koji Mineshima. Computational semantics and evaluation benchmark for interrogative sentences via Combinatory Categorial Grammar. In **Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation**, pp. 764–774, 2023.

[6] Ioannis Konstas, Srinivasan Iyer, Mark Yatskar, Yejin Choi, and Luke Zettlemoyer. Neural AMR: Sequence-to-sequence models for parsing and generation. In **Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 146–157, 2017.

[7] Xuefeng Bai, Yulong Chen, and Yue Zhang. Graph pre-training for AMR parsing and generation. **arXiv preprint arXiv:2203.07836**, 2022.

[8] Shu Cai and Kevin Knight. Smatch: an evaluation metric for semantic feature structures. In **Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)**, pp. 748–752, 2013.

[9] Irene Langkilde and Kevin Knight. Generation that exploits corpus-based statistical knowledge. In **COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics**, 1998.

[10] Patrick Blackburn and Johan Bos. **Representation and Inference for Natural Language: A First Course in Computational Semantics**. CSLI, 2005.

[11] Johan Bos, Stephen Clark, Mark Steedman, James R Curran, and Julia Hockenmaier. Wide-coverage semantic representations from a CCG parser. In **Proceedings of the 20th international conference on Computational Linguistics**, pp. 1240–1246, 2004.

[12] Lasha Abzianidze, Johannes Bjerva, Kilian Evang, Hessel Haagsma, Rik van Noord, Pierre Ludmann, Duc-Duy Nguyen, and Johan Bos. The Parallel Meaning Bank: Towards a multilingual corpus of translations annotated with

compositional meaning representations. In **Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers**, pp. 242–247, 2017.

[13] Wang Ling, Phil Blunsom, Edward Grefenstette, Karl Moritz Hermann, Tomáš Kočiský, Fumin Wang, and Andrew Senior. Latent predictor networks for code generation. In **Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 599–609, 2016.

[14] Li Dong and Mirella Lapata. Language to logical form with neural attention. In **Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 33–43, August 2016.

[15] Rik van Noord, Antonio Toral, and Johan Bos. Character-level representations improve DRS-based semantic parsing even in the age of BERT. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 4587–4603, 2020.

[16] Melvin Fitting. **First-Order Logic and Automated Theorem Proving**. Springer, 1996.

[17] Alan Robinson and Andrei Voronkov. **Handbook of Automated Reasoning**, Vol. 1. Elsevier, 2001.

[18] Johan Bos and Katja Markert. Recognising textual entailment with logical inference. In **Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing**, pp. 628–635, 2005.

[19] Lasha Abzianidze. A tableau prover for natural logic and language. In **Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing**, pp. 2492–2502, Lisbon, Portugal, 2015.

[20] Yuhuai Wu, Albert Qiaochu Jiang, Wenda Li, Markus Rabe, Charles Staats, Mateja Jamnik, and Christian Szegedy. Autoformalization with large language models. **Advances in Neural Information Processing Systems**, Vol. 35, pp. 32353–32368, 2022.

[21] Henry A Kautz. The third AI summer: AAAI Robert S. Engelmore Memorial Lecture. **AI Magazine**, Vol. 43, No. 1, 2022.

[22] 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 **Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)**, pp. 216–223, 2014.

[23] Izumi Haruta, Koji Mineshima, and Daisuke Bekki. Implementing natural language inference for comparatives. **Journal of Language Modelling**, Vol. 10, No. 1, p. 139â191, Nov. 2022.

[24] Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. Byt5: Towards a token-free future with pre-trained byte-to-byte models. **Transactions of the Association for Computational Linguistics**, Vol. 10, pp. 291–306, 2022.

[25] Rik van Noord, Lasha Abzianidze, Hessel Haagsma, and Johan Bos. Evaluating scoped meaning representations. In **Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)**. European Language Resources Association (ELRA), 2018.

[26] Laura Kovács and Andrei Voronkov. First-order theorem proving and vampire. In **International Conference on Computer Aided Verification**, pp. 1–35. Springer, 2013.

付録

以下に、3.3 節で述べた実験において使用したプロンプトを記載する。$\mathcal{D}_{\text{test}}$ の各文を{text}と置換することでプロンプトを作成し、セマンティックパージングを行った。

```
Refer to the example below and perform a semantic analysis of the sentence at
    the bottom. Only output the formula.

Text: There is no boy playing outdoors and there is no man smiling.
Formula: (-exists x1.(_boy(x1) & exists x2.(_outdoors(x2) & exists e3.(_play(e3)
     & (Subj(e3) = x1) & (Acc(e3) = x2)))) & -exists x4.(_man(x4) & exists e5.(
    _smile(e5) & (Subj(e5) = x4))))

Text: Four children are doing backbends in the park.
Formula: exists x1.(_child(x1) & _many(x1,_4) & exists x2.(_backbend(x2) &
    exists e3.(_do(e3) & (Subj(e3) = x1) & (Acc(e3) = x2) & exists x4.(_park(x4)
    & _in(e3,x4) & (Subj(e3) = x4)))))

Text: A black dog and a small white and black dog are looking up at a kitchen
    countertop.
Formula: (exists x1.(_dog(x1) & _black(x1) & exists e2.(_look(e2) & (Subj(e2) =
    x1) & exists x3.(_countertop(x3) & _kitchen(x3) & _at(e2,x3)) & _up(e2))) &
    exists x4.(exists d5 x6.(_dog(x4) & _white(x4) & _black(x4) & (d5 = _th(_dog
    )) & _small(x4,d5)) & exists e7.(_look(e7) & (Subj(e7) = x4) & exists x8.(
    _countertop(x8) & _kitchen(x8) & _at(e7,x8)) & _up(e7))))

Text: {text}
Formula:
```