# 対話型検索に向けた Semantic Parsing のデータ生成フレームワークの提案

丸山拓海　椎橋怜史　谷山徹　神谷慶　嶋村昌義

株式会社 LIFULL
{MaruyamaTakumi, ShiibashiSatoshi, TaniyamaToru, KamiyaKei, ShimamuraMasayoshi}@lifull.com

## 1 はじめに

弊社では「LIFULL HOME'S 住まいの窓口[1)]」という，家探しや家づくりに関することを専属アドバイザに無料で相談出来るサービスを運営している．本サービスは実際の店舗だけでなく，ビデオ通話や LINE 経由の相談窓口も用意されており，アドバイザとの対話を経て，より好みに合った物件探しをすることができる．一般的に，対話業務を効率化するために自動応答チャットボットが活用される傾向があるため，今回我々は対話業務の自動化を想定し，住まい探しに関するユーザ発話の言語解析を試みた．具体的には，ユーザ発話を適切な論理形式へと変換する Semantic Parsing とそのデータ構築に取り組んだ．

Semantic Parsing とは，自然言語文を特定のアプリケーションのための論理形式へと変換するタスクである． ここでいう論理形式とは，ラムダ式 [1] や SQL 文 [2, 3, 4]，プログラミング言語 [5] などの形式を指す．また，独自定義のツリー構造 [6, 7] を利用している研究も存在する． 自然言語をこのような特定の形式に変換することで，ユーザ発話を入力とした情報検索や AI アシスタントの制御が可能となる． しかしながら，論理形式をアノテーションするには一定レベルの専門知識が必要とされるため，データセットの構築は容易ではない．

そこで，我々は Wang ら [8] の英語を対象としたデータ構築手法を参考に，日本語に合わせた文法規則の定義と対話文脈への拡張を行い，論理形式アノテーションを言い換え作業に変換する新たなデータ構築フレームワークの作成を試みた．本論文の貢献を以下に示す．

- 日本語 Semantic Parsing データセットを効率的に開発するフレームワークの提案
- 対話履歴を利用した実験により，提案手法で構築したデータの効果を検証

1） https://www.homes.co.jp/counter/

## 2 関連研究

Semantic Parsing は知識データベースに対する質問応答や AI アシスタントの自然言語による制御など幅広い分野への応用が期待されるタスクである．

Semantic Parsing を行う方法として，組合せ範疇文法に基づく手法 [1, 9] や，質問応答ペアから学習する方法 [10] など様々な手法が提案されている． 近年では，深層学習をベースとした系列変換モデルによる手法も数多く提案されている [11, 12, 13, 14]．その一方で，データセットの構築に専門的知識が必要とされることから，コストが非常に高く，新規ドメインへの適用の難しさが課題となっている．

そこで，Wang ら [8] は，ドメインに依存しない文法規則とドメイン固有の語彙規則の組み合わせから擬似的な「自然言語文–論理形式ペア」を生成したのち，自然言語文側を人手で言い換えることで，より流暢かつ多様性のあるデータ構築を目指した．

## 3 データ生成手法

本論文で提案するデータ生成手法の手順を図 1 に示す．大まかな流れは次の通りである: (1) ドメイン固有の語彙規則 (3.2 節) を定義したのち，それらを (2) 文法規則 (3.3 節) によって組み合わせることで擬似的な文と論理形式のペアを生成する．その後，(3) 擬似生成文を人手で言い換えることで，(4) 言い換え文と論理形式のペアを獲得する (3.4 節)． このように，擬似生成文を経由することで，専門的知識が必要とされる論理形式のアノテーション作業が言い換え作業に置き換わり，より容易にデータ構築することが可能となる． ポイントは擬似データ生成部


**図 1** 擬似データ生成フロー

分である．このステップでは，対象となるドメインごとに構築する語彙規則を，フレームワーク側で提供するドメイン非依存の文法規則により適切に組み合わせ，自然言語文と Lambda-DCS(3.1 節) と呼ばれる論理形式のペアを生成する．詳細は，3.2 節および 3.3 節で述べる．

## 3.1 Lambda-DCS

Lambda-DCS[15] とは，Semanitc Parsing における知識グラフを背景とした論理形式の 1 つである．具体的には，名詞や数値表現 (e.g. Apartment, 1LDK) などをノードとし，その関係性 (e.g. floorPlan) をエッジとするグラフで表現される．このグラフ上のノードを辿っていくことで，任意の検索対象を表現する (3.3 節)．例えば，「間取りが 1LDK」を表現する場合，Lambda-DCS では，floorPlan.1LDK のように記述する．このように，ノード (1LDK) とエッジ (floorPlan) を結合する際に "." を用いて表現する．また，論理積 ∧ や論理和 ∨，否定論理 ¬ を利用し，条件を組み合わせていくことで更に複雑な表現を実現する． より詳細には Liang の文献 [15] を参照されたい．

**表 1** 語彙規則の例

| 住まい探しドメイン | | | |
|---|---|---|---|
| (1) | 賃貸アパート | → | TYPENP[Apartment] |
| (2) | 1LDK | → | NUM[1LDK] |
| (3) | 間取り | → | NP/ga[floorPlan] |
| (4) | 賃料 | → | NP/ga[rent] |
| (5) | ペット入居可 | → | NP/no[pet_friendly] |

**表 2** 文法規則の例

| 文法規則:基本 | | | |
|---|---|---|---|
| (1) | NP/no[$r$] の TYPENP[$x$] | → | NP[type . $x \wedge r$] |
| (2) | NP/ga[$r$] が NUM[$n$] | → | NP/no[$r.n$] |
| (3) | AP[$r$] NP[$x$] | → | NP[$x \wedge r$] |
| (4) | NP/no[$r$] ではない | → | AP[$\neg(r)$] |
| (5) | NP[x] か NP[y] | → | NP[$(x \vee y)$] |
| 文法規則:比較級・最上級 | | | |
| (6) | NP/ga[$r$] が NUM[$n$] 以上 | → | NP/no[$r. \geq .n$] |
| (7) | NP/ga[$r$] が NUM[$n$] 以下 | → | NP/no[$r. \leq .n$] |
| (8) | NP[$x$] で最も AP[$r$] | → | NP[argmax($r, x$)] |
| 文法規則:合計・平均 | | | |
| (9) | NP[$x$] の NP/ga[$r$] の合計 | → | NP[sum($r, x$)] |
| (10) | NP[$x$] の NP/ga[$r$] の平均 | → | NP[avg($r, x$)] |
| 文法規則:条件追加・削除・置換 | | | |
| (11) | NP[$x$] を追加 | → | NP[append($x$)] |
| (12) | NP[$x$] を削除 | → | NP[delete($x$)] |
| (13) | NP[x] を NP[y] に置換 | → | NP[replace($x, y$)] |

## 3.2 語彙規則

語彙規則では，3.1 節で述べたグラフのノードとエッジに，対応する自然言語と論理形式を定義する．具体的には，「自然言語 → 文法ラベル [論理形式]」のような形式を用いて定義する．ここで「住まい探し」ドメインを例に考えると (表 1)，ノードとして (1) 賃貸アパートや (2)1LDK，エッジとして (3) 間取りや (4) 賃料などが考えられる．これらを，文法ラベル (e.g. TYPENP, NUM, NP/no) と文法規則を用いて適切に組み合わせることで，擬似データを生成する．

## 3.3 文法規則

文法規則では，文法ラベルの組み合わせによる変換規則を記述する (表 2)． 生成する自然言語は日本

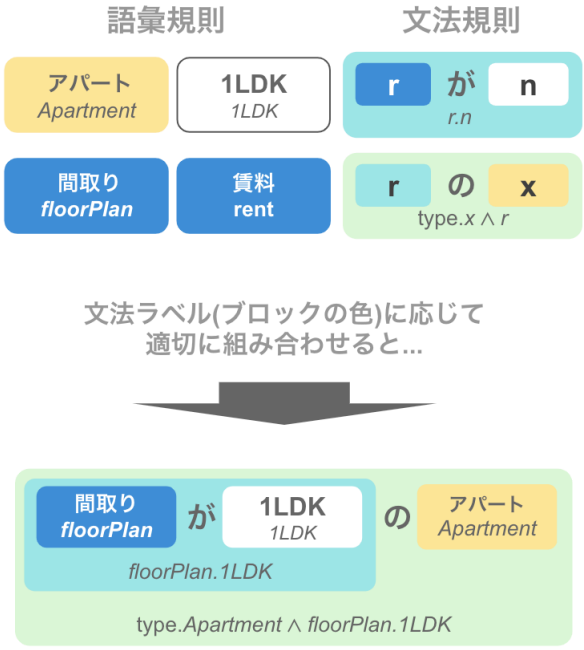


**図 2**　擬似データ生成例

語を対象としているため，Wang ら [8] の規則とは大きく異なる規則を作成した．ただし，Wang らの規則と同様に比較級や最上級，平均などの基本的な表現は網羅している．

近年の Semantic Parsing では，対話履歴に基づいて論理形式をアノテーションしたデータセット [16] や対話履歴を入力として扱うモデル [13] が提案されているが，今回は問題を簡略化するため 1 発話単位での解析に焦点を当てている．そのうえで，前後の対話文脈の内容に対して，条件の追加や修正ができるように，表 2 の (11)，(12)，(13) のような追加・削除・置換の規則を加えている．

これらの文法規則と語彙規則で定義した文法ラベルに基づいて，語句と論理形式を適切に組み合わせることで擬似的な自然言語文と論理形式のペアを生成する．例えば，語彙規則 (表 1) の (1), (2), (3) と文法規則 (表 2) の (1), (2) を文法ラベルに従って適切に組み合わせると，図 2 に示すように，“NP[NP/no[NP/ga[間取り] が NUM[1LDK]] の TYPENP[賃貸アパート]] → type.Apartment ∧ floorPlan.1LDK” というペアが得られる．

しかし，文法ラベルだけでは不自然な組み合わせが生成される恐れがある．例えば語彙規則 (表 1) の (2), (4) と文法規則 (表 2) の (2) の組み合わせから，“NP/ga[賃料] が NUM[1LDK] → rent.1LDK” などが生成されてしまう．そこで，語彙規則を定義する際には，それぞれに意味的な型を持たせて，その型制約により不自然な組み合わせの生成を抑制する．例えば，次のような生成途中の状態を考える: “NP/ga[間取り] が NUM[$x$] → floorPlan.$x$”．ここで，既に “floorPlan” が確定していることから，$x$ の値を決定する際には間取りを表す具体的な数値のみ e.g.) ワンルーム，1K, 1LDK, 2K, ... しか受け付けないといった制約を与える．

**表 3**　生成データ例

| | |
|---|---|
| 擬似生成 | 書斎が付いている 築年数 が 古い という条件を追加 |
| 言い換え | 書斎が必要です．築年数は古くても構いません． |
| 論理形式 | append( ageOfHouse.Old ∧ has_study) |
| 擬似生成 | 既にリフォーム・リノベーションされている 賃貸戸建て の中で最も 賃料 が 安い もの |
| 言い換え | リノベーション済みの賃貸戸建ての中で一番安いところだとどんな感じですか？ |
| 論理形式 | argmax(rent.Cheap, type.RentalHouse ∧ is_renovated) |

### 3.4　擬似生成文の言い換え

獲得した擬似生成文と論理形式のペアのうち，擬似生成文側を人手で言い換えることで，より流暢かつ多様性のあるデータの構築を目指す． 言い換え作業は，本論文の著者 5 人で行った．構築したデータ例を表 3 に示す。

また，5 人の間で，共通 20 文の言い換え作業を行い，作業者間の言い換えにおける類似性を BLEU によって評価した (図 3)．その結果，BLEU:0.1〜0.5 辺りに集中していることから，言い換え元 (擬似生成文) が共通していても作業者によって表現が異なり，多様な言い換え表現を得られていることが分かった．

## 4　実験

提案手法により生成したデータを Transformer[17] に学習させ，実際に自然言語文から論理形式への変換することで，データ生成手法の効果を検証する．評価には，擬似生成文の言い換えによって得られた評価セット 125 ペアおよび住まいの窓口の対話履歴を参考に構築した評価セット 173 ペアの 2 つを利用した．学習には言い換えデータ 1,229 ペアと，

表 4　実験結果

| | 言い換え評価セット | | | 対話履歴評価セット | | |
|---|---|---|---|---|---|---|
| | BLEU | Partial [%] | Exact [%] | BLEU | Partial [%] | Exact [%] |
| (a) 言い換え文のみ | 78.21 | 66.40 | 32.00 | 66.20 | 60.12 | 29.48 |
| (b) + 擬似生成文 | 77.51 | 66.40 | 32.80 | 70.91 | 60.69 | 38.15 |
| (c) + BERT | 84.88 | 70.40 | 46.40 | 67.78 | 58.38 | 39.84 |
| (d) 擬似生成文のみ | 56.50 | 12.80 | 4.80 | 45.85 | 6.35 | 0.00 |

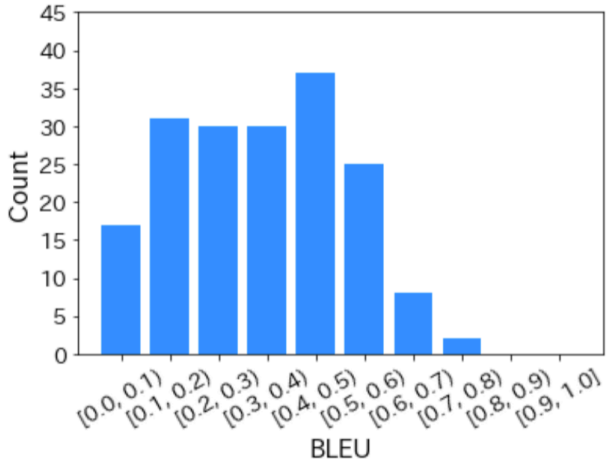


図 3　作業者間の言い換えの類似性

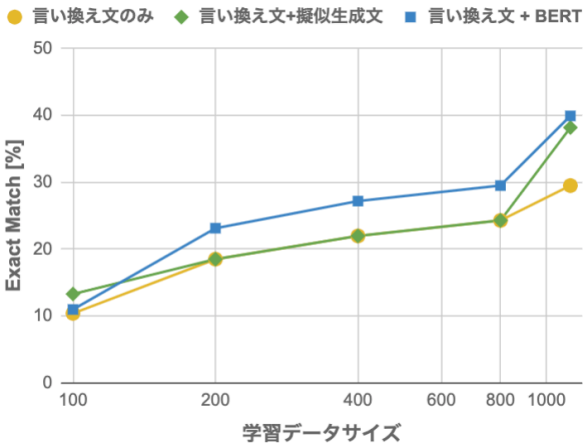


図 4　データサイズとモデルの性能

事前学習用に擬似生成データ 50,496 ペアを用いた．また，Encoder に学習済み BERT[2] を利用し [18]，事前学習の効果を比較した．Transformer の Encoder および Decoder の各パラメータは BERT[19] を参考にした．

### 4.1　実験結果

実験では学習方法および学習データの異なる次の 4 つの手法を比較した: (a) 言い換え文–論理形式ペアのみを学習データとしたモデル，(b) 擬似生成文–論理形式ペアを事前学習したのちに言い換え文–論理形式ペアで Fine-tuning したモデル，(c) 事前学習済み BERT を Encoder として与えたのちに言い換え文–論理形式ペアで Fine-tuning したモデル，(d) 擬似生成文–論理形式ペアのみを学習データとしたモデル (表 4)．評価には，BLEU および正解の論理形式と完全一致 (Exact)・1 つ以上の項が一致 (Partial) する出力をした割合を指標に用いる．

言い換え文評価セットと対話履歴評価セットの結果を比較すると傾向に大きな違いはなく，提案した生成手法が同じ方法で構築された言い換え文評価セットだけでなく目的の対話履歴の解析にも有効であることが分かる．さらに，(a) 言い換えデータのみの学習と (d) 擬似データのみの学習結果の差から，人手による言い換え作業の効果が見て取れる．

対話履歴評価セットにおける Partial および Exact に注目すると，まだ実応用できる精度とは言い難いが，図 4 に注目すると，さらにデータサイズを拡張することで，性能改善の見込みがあることが分かる．今後は，モデルの性能改善に寄与する学習データを効率的に収集する方法を検討していきたい．

## 5　おわりに

本論文では，語彙規則と文法規則の組み合わせから自然言語文–論理形式ペアを生成する，新たなデータ構築手法を提案した．自動生成された擬似データを経由することで，専門知識が必要とされる論理形式のアノテーションが擬似生成文の言い換え作業に置き換えられ，より容易にデータセットを作ることが可能となった．生成データを学習した変換モデルにより，対話履歴を論理形式に変換する実験を行い，提案したデータ生成手法の有効性を確認した．

2)　https://github.com/cl-tohoku/bert-japanese

# 参考文献


[1] Qingqing Cai and Alexander Yates. Large-scale semantic parsing via schema matching and lexicon extension. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 423–433, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

[2] Victor Zhong, Caiming Xiong, and Richard Socher. Seq2sql: Generating structured queries from natural language using reinforcement learning. *CoRR*, Vol. abs/1709.00103, , 2017.

[3] Catherine Finegan-Dollak, Jonathan K. Kummerfeld, Li Zhang, Karthik Ramanathan, Sesh Sadasivam, Rui Zhang, and Dragomir Radev. Improving text-to-SQL evaluation methodology. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 351–360, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[4] Tao Yu, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, Zilin Zhang, and Dragomir Radev. Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-SQL task. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 3911–3921, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[5] Y. Oda, H. Fudaba, G. Neubig, H. Hata, S. Sakti, T. Toda, and S. Nakamura. Learning to generate pseudo-code from source code using statistical machine translation. In *2015 30th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, pp. 574–584, 2015.

[6] Chris Quirk, Raymond Mooney, and Michel Galley. Language to code: Learning semantic parsers for if-this-then-that recipes. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 878–888, Beijing, China, July 2015. Association for Computational Linguistics.

[7] Kavya Srinet, Yacine Jernite, Jonathan Gray, and Arthur Szlam. CraftAssist instruction parsing: Semantic parsing for a voxel-world assistant. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4693–4714, Online, July 2020. Association for Computational Linguistics.

[8] Yushi Wang, Jonathan Berant, and Percy Liang. Building a semantic parser overnight. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1332–1342, Beijing, China, July 2015. Association for Computational Linguistics.

[9] Chitta Baral, Juraj Dzifcak, Marcos Alvarez Gonzalez, and Jiayu Zhou. Using inverse lambda and generalization to translate English to formal languages. In *Proceedings of the Ninth International Conference on Computational Semantics (IWCS 2011)*, 2011.

[10] Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. Semantic parsing on Freebase from question-answer pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1533–1544, Seattle, Washington, USA, October 2013. Association for Computational Linguistics.

[11] 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, Berlin, Germany, August 2016. Association for Computational Linguistics.

[12] Jayant Krishnamurthy, Pradeep Dasigi, and Matt Gardner. Neural semantic parsing with type constraints for semi-structured tables. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 1516–1526, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[13] Alane Suhr, Srinivasan Iyer, and Yoav Artzi. Learning to map context-dependent sentences to executable formal queries. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 2238–2249, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[14] Zhichu Lu, Forough Arabshahi, Igor Labutov, and Tom Mitchell. Look-up and adapt: A one-shot semantic parser. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 1129–1139, Hong Kong, China, November 2019. Association for Computational Linguistics.

[15] Percy Liang. Lambda dependency-based compositional semantics. *CoRR*, Vol. abs/1309.4408, , 2013.

[16] Shashank Srivastava, Amos Azaria, and Tom Mitchell. Parsing natural language conversations using contextual cues. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pp. 4089–4095, 2017.

[17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

[18] Sascha Rothe, Shashi Narayan, and Aliaksei Severyn. Leveraging pre-trained checkpoints for sequence generation tasks. *arXiv*, pp. arXiv–1907, 2019.

[19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *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.