日本語 Text2Data のためのデータセットの検証および実社会応用

仲田将斗 [1]　亀甲博貴 [2]　森信介 [2]
[1] 京都大学大学院 情報学研究科　[2] 京都大学 学術情報メディアセンター
nakata.masato.26m@st.kyoto-u.ac.jp　{kameko,forest}@i.kyoto-u.ac.jp

概要

テキストアナリティクスにおける実験設定の記述や再現は，非専門家の研究者にとって困難な課題である．本研究では，構造化データとテキストを相互変換する Data2Text および Text2Data タスクの基盤強化を目的とし，日本語 Wikipedia から抽出した 123M 文規模の新たなデータセット Japinax を構築した．実験では，事前学習済み言語モデルの fine-tune におけるデータサイズと変換精度の相関を分析し，実用的な精度を得るための定量的な指標を明らかにした．さらに，実社会応用として Web アプリケーション LiTA を開発し，研究支援機能を提供するとともに，ユーザからのフィードバックを通じて継続的にデータセットを拡充するシステムを実現した．

1　はじめに

人文学や社会科学において，研究対象とするテキスト集合から統計的な知見を得るために，情報学の活用が広がっている．**テキストアナリティクス**とは，単語の出現頻度分析や共起ネットワーク，トピックモデルなどの手法を用いて，テキストを定量的に分析する分野である．しかし，これらの分析手法を適切に選択し，各種パラメータを調整・理解するには，統計学や情報学に関する高度な専門知識が要求される．そのため，非専門家である研究者にとって，実験の設定内容を論文中に必要十分に記述することや，他者の論文から実験設定を正しく再現することは，依然として高い障壁となっている [1]．

こうした課題に対し，森田ら [1] は，テキストアナリティクスの操作ログ（構造化データ）と論文中の説明文（自然言語）の間の双方向変換モデルを提案した．この変換は，構造化データからテキストを生成する Data2Text と，テキストから構造を抽出する Text2Data の両面を併せ持つタスクである．しかし，先行研究で構築された日本語データセット Texylon は，人手によるアノテーションに基づいているため，データサイズが制限されていた．

本研究では，まず日本語における Data2Text および Text2Data の基盤となる大規模なデータセットの構築に取り組む．日本語 Wikipedia のダンプデータから，表（‘‘data’’）とテキストを紐付けた新たなデータセット Japinax を構築した．本データセットは 123M 文から構成され，日本語におけるデータ・テキスト変換研究の新たなベンチマークとなるものである．

その上で，事前学習済み言語モデルをベースとし，追加学習に用いるデータセットのサイズが，変換精度にどのような影響を与えるかを分析した．実験を通じて，具体的な量的目標を示した上で，それを補うための仕組みを提案する．すなわち，実社会応用として Web アプリケーション LiTA を開発し，ユーザからのフィードバックを通じて継続的にデータセットを拡充するシステムを実現した．

2　関連研究

2.1　Data2Text

Data2Text とは広く ‘‘構造化データから，それを記述する自然言語への変換’’ を扱うタスクである．よく扱われる構造化データには，MR（meaning representation; 意味表現），グラフ構造，表形式データなどが挙げられる．

Data2Text 用データセットとしては，WebNLG [2]，WIKITABLETEXT [3] などが存在する．WebNLG はグラフ構造，とくに知識グラフに対する Data2Text のためのデータセットであり，後者は表形式データに対する Data2Text を扱う．

たとえば Kale らは，transformer encoder-decoder である T5 [4] を用いたモデルを提案し，従来手法を上

回っている [5]．彼らは大規模な英語コーパスで事前学習された T5 モデルを，WebNLG 等の Data2Text データセットで fine-tune することで，質の高いテキスト出力を得たと報告している．その際，生の構造化データを単一の文字列へ‘‘線形化（linearization）’’し，言語モデルへの入力としている．

2.2 Text2Data

IE（information extraction; 情報抽出）の一分野をなすタスクとして，Data2Text の逆変換 Text2Data がある．前述の Kale らのモデルのように，線形化手法を用いた Text2Data 用 T5 が Wu らによって提案された [6]．これは BERT [7] による NER（named entity recognition; 固有表現認識）を上回る精度を記録し，線形化手法の優位性を示した．

また，近年の LLM（large language model; 大規模言語モデル）の台頭により，Text2Data は劇的に精度が向上した．TKGT [8] は LLM を用いた Text2Data モデルの代表例である．これは線形化手法をも大きく上回り，とくに評価指標の一つ header F1 [6] においては 100% を達成し，Wu らの 86.02 ~ 88.02% を凌いだ．ただし LLM はパラメータ数が非常に多く，TKGT も合計で 76 B ほどのパラメータを有している．この事実は前節で述べたような ‘‘実システムでの応用’’ をするにあたって，大きな障害となる．すなわち，RAM 等のハードウェア資源を多く消費するため，普及に必要なコストが高くなる．そのため我々は，LLM に頼らずに Text2Data の手法を模索する必要がある．

2.3 テキストアナリティクス分野

テキストアナリティクスは，これら Data2Text, Text2Data の応用が期待される分野である．テキストアナリティクスとはテキスト集合の性質を調べるための統計的・確率的手法全般を指す．テキストアナリティクスを行うためのツールはいくつか知られているが，その利用には専門性を要することが問題となっている [1]．

そのような当該分野の研究者の負担を軽減するため，森田らはテキストアナリティクス論文執筆・再現に際して役立つデータセット Texylon およびモデルを公開した [1]．Texylon は日本語で書かれたテキストアナリティクス実験の論文から収集した Data2Text データセットであり，その ‘‘data’’ としては実験設定を表す MR が収録されている．ただし，Texylon は人手でアノテーションしたデータセットであり，サイズの小ささが問題視されている．

表 1 評価に使用したデータセット

データセット	言語	ドメイン	文数
WIKITABLETEXT	英	マルチジャンル	13,320
Texylon	日	テキスト分析	423
Japinax	日	マルチジャンル	123M

3 データセット

本研究では，Text2Data におけるモデルの性能を評価するため，言語やドメインの異なる三種類のデータセットを用いる．

WIKITABLETEXT は，Wikipedia から抽出された多様なジャンルの表と，それに対応する短い説明文のペアとから構成される英語のデータセットである [3]．本データセットは，特定のドメインに特化しない汎用的な Data2Text/Text2Data に対する標準的なベンチマークとして広く利用されている．各データは，一つのテーブルに含まれる属性名と値のペアの集合と，その内容を要約または説明する一つの英文によって構成される．

Texylon は，日本語のテキストアナリティクス分野の学術論文から収集された，専門的なデータセットである．本データセットは，実際の学術論文内に掲載されている表と，その表の論文中での言説の対とを人手でアノテーションすることで構成された．

Japinax [1)] は，日本語の Data2Text/Text2Data のために，新たに Wikipedia から構築した日本語データセットである．

構築手順は以下の通りである．まず，日本語 Wikipedia の最新のダンプデータ[2)]から，記事内のセクション（section）構造を保持したまま，テキストとテーブルを抽出する．次に，各記事をセクション単位で分割し，各セクション内に含まれる本文テキストと，そのセクション内に配置されている表を一つのサンプルとして紐付ける．一般にセクションは複数の表を含むため，テキストと表は 1:N に対応する．本データセットは合計 122,577,476 文から構成される．

1） /jaˈpɪnəks/
2） 2025 年 11 月 26 日現在

4 実験

4.1 モデル

本研究では，Text2Data タスクを解くにあたり，次のアーキテクチャを採用する．いずれのモデルにおいても，大規模コーパスで事前学習された言語モデルをベースとし，各データセットを用いて fine-tune を行うことで，構造化データへの変換能力を学習させる．

Linearized Text2Data モデル（*Lin*） データ全体の構造を一度に生成することを目的とした，Seq2Seq としての定式化である．本手法では，入力として自然言語テキスト `[TEXT]` を与え，出力として特殊なタグを用いた線形化（linearization）済み構造化データを生成する [5]．具体的には，Kale らに倣い，次の形式の擬似的なマークアップテキストを用いる．

```
<table>
    <cell>
        [VALUE]
        <header> [KEY] </header>
    </cell>
</table>
```

本論文では，このアーキテクチャを *Lin* と呼ぶことにする．

4.2 評価指標

本研究では，推定された表データ（key-value 対）の各 `[KEY]` に対して，生成された `[VALUE]` が正しく推定できているかを定量的に測定する．評価には以下の三種類の指標を用いる．

- **Exact match (F1):** 生成された値が正解ラベルと完全に一致するかどうかを判定する．
- **BLEU:** 生成された値と正解ラベルの間の N-gram の重なりを測定する [9]．
- **Semantic similarity** 表層的な一致でなく，意味的な近接性を評価する．事前学習済み sentence-transformers モデルである `all-MiniLM-L6-v2`[3] を用いて 384 次元の密ベクトルへと埋め込み，その平均コサイン類似度を計算する．

3） https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2 — 2026/01/08 アクセス．

表 2 モデルサイズとデータセットサイズの関係．モデルは fine-tune に使用したデータセットの割合 [%] とパラメータ数（small/base）を表す

モデル	Exact	BLEU	Semantic
2% small	0.000	00.15	0.293
4% small	0.000	00.17	0.288
8% small	0.000	00.01	0.198
20% small	0.000	76.87	0.908
40% small	0.000	78.74	0.916
60% small	**0.001**	**81.42**	**0.923**
2% base	0.000	63.40	0.874
4% base	0.000	75.39	0.914
8% base	0.001	81.12	0.924
20% base	0.003	83.58	0.930
40% base	0.007	84.28	0.932
60% base	**0.011**	**84.66**	**0.934**

4.3 実験 1

まずモデルのパラメータ数と学習データサイズが Text2Data の変換精度に与える影響を調査するため，次の設定で実験を行った．ベースとなる言語モデルには，CodeContests [10] 等のデータセットで事前学習された FLAN-T5 [11] を採用し，異なるパラメータサイズ（60.5M *vs.* 0.2B）の間で比較を行った．さらに訓練データサイズを段階的に変化させた WIKITABLETEXT を用いて，*Lin* アーキテクチャで fine-tune を実施した．

4.4 結果 1

実験 1 の結果を表 2 に示す．この表から分かるように，パラメータの少ない（60.5M）モデルでは，訓練データが 20%（2,000 文）を境に著しく精度向上し，そこから頭打ちとなる．その評価値は，パラメータの多い（0.2B）モデルを 4~8% の訓練データ（400~800 文）で学習させたものに及ぶ．このことから，パラメータ数を減らすと，およそそれに反比例して多くの訓練データを用意する必要があると言える．

4.5 実験 2

続いて，本研究で構築した Japinax（§3）を中心に，日本語における Text2Data の性能とドメイン適応能力を評価した．ベースモデルには，OSCAR [12, 13] 等の日本語コーパスで事前学習された日本語 T5 (0.2B) を使用した．実験 1 と同じようにデータサイ

表 3 Japinax の N [%] で学習したモデルの精度評価（unseen 設定）

モデル	Exact	BLEU	Semantic
2%	0.000	00.02	0.085
4%	0.000	00.02	0.089
8%	0.000	00.02	0.080
20%	0.000	00.02	0.085
40%	0.000	00.02	0.090
60%	0.000	00.02	0.083
80%	0.000	00.02	0.084
100%	0.000	00.01	0.08

表 4 Japinax の N [%] で学習したモデルの精度評価（fine-tuned）

モデル	Exact	BLEU	Semantic
2%	0.142	00.00	0.503
4%	0.148	00.00	0.489
8%	0.006	55.47	0.789
20%	0.150	00.00	0.579
40%	0.206	00.00	0.644
60%	0.235	70.36	0.592
80%	0.174	00.00	0.583
100%	0.164	00.00	0.561

ズを変化させながら Japinax で fine-tune し，その上で以下の二つのシナリオを検証した．

第一に，未知の（unseen）ドメインに対するゼロショット能力を測るため，学習に用いていない Texylon データセットに対して直接テストを実施した．第二に，少数の Texylon データを用いて追加の fine-tune を行い，その後同じテストデータにおける性能を測定した．これらの実験において，fine-tune に用いる Japinax のサイズを変化させることで，大規模な一般ドメインデータでの学習が，専門的な特定ドメイン（Texylon）への適応にどのような影響を与えるかを分析した．

4.6 結果 2

Unseen 設定での結果を表 3 に，Texylon で fine-tune したものの結果を表 4 に示す．前者においてはほとんどのケースで推定に失敗しているのに対して，fine-tune してドメイン適用することで，精度が大幅に改善されている．

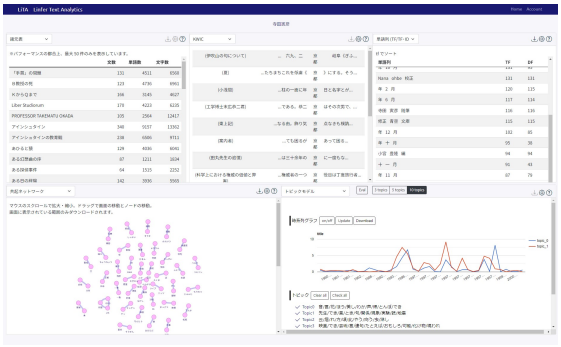

図 1 テキストアナリティクスツール LiTA の概観

5 アプリケーションとして

これまでの実験（§4）により，およそ数千ほどのデータがあれば，Text2Data モデルを十分に学習できることが分かった．そこで，このようなデータを収集しつつ，実社会への応用として使えるツールとして，Web アプリケーション LiTA[4] を開発した（図 1）．

Text2Data タスクの応用先の一つに，論文テキストから実験設定への自動変換による，研究者支援がある．その中でも，専門的な情報学的計算に不慣れな人文系の研究者を対象に，**テキストアナリティクス分野**における研究者支援が進められてきた [1]．

本研究はその一環として，1) 論文再現の AI アシスタント，2) テキストアナリティクスの実験，3) ユーザのフィードバックによる，さらなるデータセットの拡充の三点が可能な Web アプリケーションを開発した．

6 結論

我々はまず，日本語の Text2Data データセットの不足を補うため，Wikipedia から 123M 文を収集し，新たなデータセット Japinax を構築した．さらに，この Japinax や既存の英語データセットを用いて，データセットのサイズと推定精度との間の詳細な傾向を示した．ドメイン適用に必要な訓練データ量を明確にした上で，特定ドメインにおけるデータセットのさらなる収集と，実社会での応用とに焦点を当てた Web アプリケーションの公開へと至った．

今後は，このアプリケーションの安定した稼働および普及に努めつつ，より精度の高い Text2Data モデルの研究を進めなければならず，そのためのデータセット収集は急務である．

4） © 2020 Linfer Inc.

参考文献

[1] Masato Nakata, Kosuke Morita, Hirotaka Kameko, and Shinsuke Mori. Texylon: Dataset of log-to-description and description-to-log generation for text analytics tools. In Toyotaro Suzumura and Mayumi Bono, editors, **New Frontiers in Artificial Intelligence**, pp. 269–283, Singapore, 2024. Springer Nature Singapore.

[2] Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. Creating training corpora for NLG micro-planners. In Regina Barzilay and Min-Yen Kan, editors, **Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 179–188, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[3] Junwei Bao, Duyu Tang, Nan Duan, Zhao Yan, Yuanhua Lv, Ming Zhou, Tiejun Zhao, and Min-Yen Kan. Table-to-text: Describing table region with natural language. In **Proceedings of the AAAI Conference on Artificial Intelligence, 32 Volume**, pp. 5020–5027. Association for the Advancement of Artificial Intelligence, February 2018.

[4] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. **Journal of Machine Learning Research**, Vol. 21, No. 140, pp. 1–67, 2020.

[5] Mihir Kale and Abhinav Rastogi. Text-to-text pre-training for data-to-text tasks. In **Proceedings of the 13th International Conference on Natural Language Generation**, pp. 97–102, Dublin, Ireland, December 2020. Association for Computational Linguistics.

[6] Xueqing Wu, Jiacheng Zhang, and Hang Li. Text-to-table: A new way of information extraction. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 2518–2533, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[7] 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.

[8] Peiwen Jiang, Xinbo Lin, Zibo Zhao, Ruhui Ma, Yvonne Jie Chen, and Jinhua Cheng. TKGT: Redefinition and a new way of text-to-table tasks based on real world demands and knowledge graphs augmented LLMs. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing**, pp. 16112–16126, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[9] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In **Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics**, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[10] Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Rémi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, Thomas Hubert, Peter Choy, Cyprien de Masson d'Autume, Igor Babuschkin, Xinyun Chen, Po-Sen Huang, Johannes Welbl, Sven Gowal, Alexey Cherepanov, James Molloy, Daniel Mankowitz, Esme Sutherland Robson, Pushmeet Kohli, Nando de Freitas, Koray Kavukcuoglu, and Oriol Vinyals. Competition-level code generation with alphacode. **arXiv preprint arXiv:2203.07814**, 2022.

[11] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models, 2022.

[12] Pedro Javier Ortiz Su'arez, Laurent Romary, and Benoit Sagot. A monolingual approach to contextualized word embeddings for mid-resource languages. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 1703–1714, Online, July 2020. Association for Computational Linguistics.

[13] Pedro Javier Ortiz Su'arez, Benoit Sagot, and Laurent Romary. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019, pp. 9 – 16, Mannheim, 2019. Leibniz-Institut f'ur Deutsche Sprache.