表の段階的な縮小による biography 生成

金子 知弘　　船越 孝太郎　　奥村 学
東京工業大学
{kaneko, funakoshi, oku}@lr.pi.titech.ac.jp

概要

本研究では，人物情報に関する表から説明文を生成する biography 生成における冗長性の改善と網羅性の向上に取り組む．一般に生成モデルでは，繰り返しなどの冗長な出力を行う問題があり，biography 生成においても表内の項目を複数回記述して冗長な出力を行ってしまうことがある．本研究では，biography 生成における冗長な生成を制御する方法の 1 つとして，表中から記述済みの項目を削除し，表を段階的に縮小する手法を提案する．実験では，SynthBio データセットを用い，自動評価において冗長性の改善，網羅性の向上を確認した．

1　はじめに

自然言語処理分野では，表 [1] やグラフ [2] など，構造化されたデータからのテキスト生成が重要な課題として位置付けられている．その中で Table-to-text タスク [3, 4] は，図 1 に示すように，表を入力し，表の内容を自然言語で記述するタスクである．生成されたテキストにより，表内の情報の関連性を理解しやすくなる利点がある．本研究では，その中でも，ある人物に関する情報をまとめた表を対象に biography を生成するタスクに取り組む．

一般に生成モデルでは，繰り返しなどの冗長な出力を行う問題があり，biography 生成においても表内の項目を複数回記述して冗長な出力を行うことがある．例えば，高い生成性能を持つ事前学習済みモデル T5 [5] を Wikiperson データセット [6] で追加学習し SynthBio データセット [7] で評価すると，図 1 のように，表内のセルが生成文で複数回記述され（赤色）冗長な生成が起きている．さらに，冗長な生成により，記述されないセルが発生する（青色）．

本研究は，このような内容記述の冗長性を減らすとともに網羅性を上げることを目的とし，表内から記述済みのセルを削除し，表を段階的に縮小する手法を提案する．この手法ではまずどのセルが記述されたかを判定し，その後記述されたと判定されたセルはクロスアテンションにおいてマスクを行う．これにより，人物情報に関する，より理解しやすい表の内容記述を生成することができる．

key	value
name	Sylvia Audrey Mabika
nationality	Zimbabwean
birth place	Harare, Zimbabwe
alma mater	Yale University
education	Masters in Literature
occupation	novelist, poet
notable works	The Reluctant Muse
language	English
partner	Kassim Abid

Sylvia Audrey Mabika is a Zimbabwean poet and novelist. Sylvia Audrey Mabika was born in Harare, Zimbabwe Zimbabwe. Sylvia Audrey Mabika was born in Harare, Zimbabwe Zimbabwe to a Zimbabwean mother and a Zimbabwean mother. Sylvia Audrey Mabika's first novel The Reluctant Muse was published in 2007. Sylvia Audrey Mabika's first novel The Reluctant Muse was published at Yale University. mabika's first novel The Reluctant Muse was published in English.

図 1　Table-to-text タスクの表と生成文の例．表内の赤色のセルは，生成文の赤色の単語列で複数回記述されている．一方，表内の青色のセルが記述されていない．

近年，人間のアノテータが介入した SynthBio データセットが提案され，biography 生成におけるモデル性能をより適切に評価できるようになった．そこで，本研究でも SynthBio データセットを用いて評価を行い，自動評価において冗長性の改善と網羅性の向上を確認した．

2　関連研究

人物情報の表から biography を生成するタスク [8, 9] は，文生成を目的とした WikiBio データセット [1] とテキスト生成を目的とした Wikiperson デー

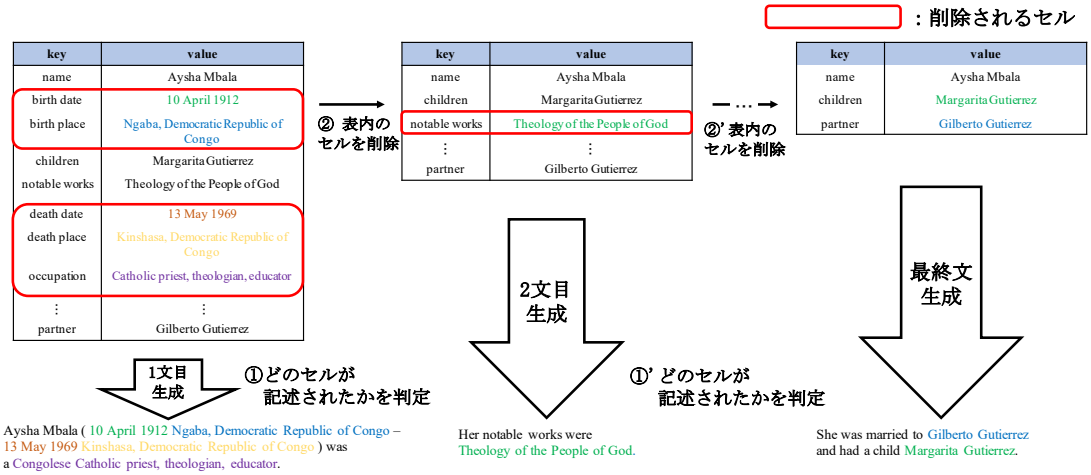

図 2　提案手法の流れ．色付きの文字は，記述されたと判定されたセルとそのセルに該当する出力文の単語列を表す．

タセットで研究されてきた．例えば，Liu ら [10] の入力にエンティティ情報を結合した学習方法，Wang ら [9] の表とテキストの類似性を制約に加算した生成フレームワーク，Tian ら [11] のデータからスコアを学習する変分ベイズ学習フレームワークなどがある．Wang ら [8] は，忠実度と網羅性の向上を目的とした 2 段階モデルを提案し，従来の最先端モデルのスコアを凌駕した．このモデルは，1 段階目で生成すべき単語を抽出し，2 段階目でその抽出した単語を元に，単語の削除と挿入を繰り返し行うことでテキスト生成を行っている．しかし，この手法ではモジュールが多くモデルが複雑化している．

3　提案手法

本節では，biography 生成における冗長な生成を制御する手法として，記述された表内のセルを削除し，表を段階的に縮小する手法を説明する．まず biography 生成では，表 T を入力とし，説明文 $Y = \{y_1, y_2, \cdots, y_m\}$ を生成する．表 T は $T = \{c_1, c_2, \cdots, c_n\}$ のようにセル c_i の集合として形式化され，セル c_i は $c_i = \langle k_i, v_i \rangle$ のようにキー値 k_i とバリュー値 v_i のペアで構成される．ただし，表 T をモデルに入力するときは，表の構造をモデルが理解しやすくするため，バリュー値の前後に 2 つの特別なトークン<SP1>・<SP2> を挿入する．例えば，<birth date, 10 April 1912>は，birth date <SP1> 10 April 1912 <SP2>とモデルに入力する．

具体的な提案手法の流れを図 2 に示す．まず，表から 1 文目が生成されたら，表内のどのセルが記述されたかの判定を行う（図 2 ①）．次に，記述されたと判定されたセルは表から削除する（図 2 ②）．そして，縮小された表を元に 2 文目を生成し，再び 2 文目において表内のどのセルが記述されたかを判定し（図 2 ①'），記述されたと判定されたセルは表から削除する（図 2 ②'）．この一連の流れを，表内のセルがなくなるまで繰り返し行うことで，冗長性を減らすと共に網羅性を上げる．

3.1　どのセルが記述されたかの判定

学習時，デコーダにどのセルが記述されたかを表すヘッドを構築することで判定を行う．具体的なヘッド構築の流れを図 3 に示す．まず，エンコーダの隠れ状態からセル表現を構築する（図 3 ①）．次に，出力文の各トークンからセル表現へのアテンションスコア（図 3 ②）を計算する．そして，事前に作成した教師データ（図 3 ③）とアテンションスコアからクロスエントロピー損失を計算する．最後に，クロスエントロピー損失を言語モデルの学習損失に加算する．

推論時は，構築したヘッドのアテンションスコアを元に，どのセルが記述されたかの判定を行う．具体的には，生成したトークンからあるセル表現へのアテンションスコアが，事前に設定した閾値を上回った場合，そのセルは「記述された」と判定する．

セル表現の構築　エンコーダの隠れ状態からセル表現をまず構築する．しかし，セルに含まれる全てのトークンの隠れ状態からセル表現を構築すると，セル同士でテキストが異なるため，隠れ状態の次元数も異なる．その結果，セル同士でセル表現の次元数も異なってしまう．そこで，モデル入力において各セルに付加した特別なトークン（<SP1>，<SP2>）の隠れ状態を結合したものをセル表現として利用す

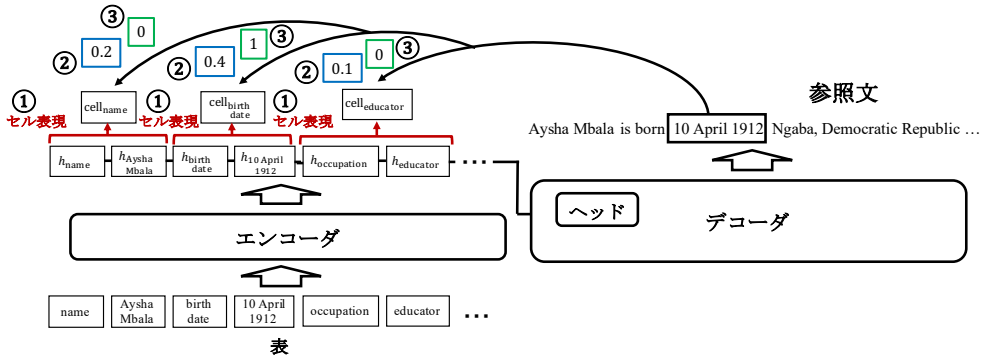

図 3　どのセルが記述されたかを表すヘッドの構築の流れ．①は，セルの隠れ状態からセル表現を構築する手順．例えば，<birth date,10 April 1912> はセル表現 $\mathrm{CR}_{\text{birth date}}$ となる．②は，出力文の各トークンから各セル表現へのアテンションスコア（青色）．③は，出力文の各トークンから各セル表現への教師データ（緑色）．例えば，出力文の 10 April 1912 というトークン列は，<birth date,10 April 1912> のセル表現 $\mathrm{CR}_{\text{birth date}}$ への教師データが 1 で，それ以外は 0 である．

る．具体的には，セル c_i に対してのセル表現 CR_i は，

$$\mathrm{CR}_i = \mathrm{concat}(h_{\mathrm{SP1}_i}, h_{\mathrm{SP2}_i}) \tag{1}$$

となる．ここで，h_{SP1_i} と h_{SP2_i} はそれぞれセル c_i の特別なトークン<SP1>と<SP2> の隠れ状態である．

アテンションスコアの計算　Transformer モデル [12] の出力文のトークンから入力文のトークンへのアテンションスコアの計算方法と同様に，出力文のトークン y_j からセル表現 CR_i へのアテンションスコア $A_{i,j}$ を計算する．

$$A_{i,j} = \underset{i}{\mathrm{softmax}}(K_{\mathrm{CR}_i} Q_j) \tag{2}$$

ここで，K_{CR_i} と Q_j は

$$K_{\mathrm{CR}_i} = W_{\mathrm{CR}} \times \mathrm{CR}_i, \tag{3}$$

$$Q_j = W_y \times y_j \tag{4}$$

であり，W_{CR} と W_y は学習可能な重み行列である．

クロスエントロピー損失の計算　出力文のトークン y_j からセル表現 CR_i へのアテンションスコア $A_{i,j}$ と教師データ $L^{n\times m}$ からクロスエントロピー損失を計算する．

$$\mathrm{loss}_{\mathrm{attn}} = \frac{1}{m} \sum_{1\le j\le m} \frac{1}{n} \sum_{1\le i\le n} L_{i,j} \log A_{i,j} \tag{5}$$

ここで，教師データ L は

$$L_{i,p:q} = 1,\ L_{l(\neq i),p:q} = 0 \qquad (y_{p:q} = {}^{\exists} v_i), \tag{6}$$

$$L_{l,p:q} = \frac{1}{n} \qquad (y_{p:q} \neq {}^{\forall} v_i) \tag{7}$$

である．式 (6) は，あるセル c_i のバリュー値 v_i と出力文の単語列 $y_{p:q}$ が完全一致[1]するとき，出力文の単語列 $y_{p:q}$ からセル表現 CR_i への教師データは 1 であり，それ以外のセル表現 $\mathrm{CR}_{l(\neq i)}$ への教師データは 0 であることを意味する．式 (7) は，出力文のトークン列 $y_{p:q}$ が全てのセルのバリュー値と完全一致しないとき[2]，出力文の単語列 $y_{p:q}$ から全てのセル表現への教師データは $\frac{1}{n}$ であることを意味する．

言語モデルの学習損失への加算　計算したクロスエントロピー損失 $\mathrm{loss}_{\mathrm{attn}}$ を言語モデルの学習損失 $\mathrm{loss}_{\mathrm{LM}}$ に加算する．

$$\mathrm{loss} = \mathrm{loss}_{\mathrm{LM}} + \mathrm{loss}_{\mathrm{attn}} \tag{8}$$

どのセルが記述されたかの判定を行う手法を説明してきたが，この手法では，生成文に対してどのセルも記述されていないと判定する場合がある．そのような場合には，記述されたと未だ判定されていないセルの中で，生成文の各トークンからのアテンションスコアが最大のセルをまずリストアップし，その中で最大のアテンションスコアであるセルを記述されたと判定する．

3.2　表内のセルの削除

表内のセルの削除は，Transformer モデル [12] のクロスアテンションをマスクすることで仮想的に行う．クロスアテンションとは，文生成時に入力のどこに注目するかを表す機構なので，クロスアテンションをマスクするということは，入力の表の特定のセルに注目させず記述させないことを意味する．

一般に biography 生成用のデータセットでは，表 1 に示すように，セルが参照文で 2 回以上言及される

1)　同じ表内で包含関係にあるバリュー値を持つ複数のセルが存在する場合（例えば，セル<birth place，Columbia> とセル<university，Columbia University> がある場合），出力文と最長で完全一致しているセルを優先する．

2)　これは，出力文の単語列 $y_{p:q}$ がどのセルも記述していないことを意味し，文章構成上必要な単語列などについてのものである．

表 1 Wikiperson でセルが参照文で言及される回数.

回数	学習データ	検証データ	評価データ
0 回	2879（0.18%）	394（0.20%）	344（0.18%）
1 回	1290936（81.8%）	155959（81.8%）	153092（81.7%）
2 回以上	284265（18.0%）	34207（18.0%）	33935（18.1%）

表 2 Wikiperson と SynthBio の統計量.

データセット	学習データ	検証データ	評価データ
Wikiperson	250,186	30,487	29,982
SynthBio	-	1,118	1,119

ことがある．このように参照文で 2 回以上言及されるセルは冗長であると考え，学習時はセルが最後に言及された時点でマスクし，推論時は 1 回目に記述された時点でマスクする．

4 実験

4.1 実験設定

学習データには Wikiperson データセット，検証・評価データには Wikiperson データセットと SynthBio データセットを用いた．これらのデータセットの統計量を表 2 に示す．

自動評価では，PARENT [13] と PARENT-T [9] を評価指標として用いた．PARENT は，表と参照文の両方との，n-gram 単位での生成文の一致度を測る．PARENT-T は，表のみを考慮した n-gram 単位での生成文の一致度を測る．テキスト長は，SentencePiece [3] を用いて文字列をトークン化し，トークン単位で生成テキストの長さを測ったものである．

先行研究 [8] は SynthBio データセットで評価しておらず，また，Wikiperson データセットでは T5 モデル [5] が先行研究 [8] よりも PARENT F1 が良いため，ベースラインとしては，T5 モデルと，繰り返し抑制の既存研究である，repetition penalty [14] を T5 上に実装したモデル（rep. penalty モデル）を用いた．提案モデルも T5 上で構築する．

より具体的には，本研究の実装には Huggingface [15] を用いた．モデルは T5-small を使用し，ヘッド数は 8，層数は 6 である．初期学習率は 5e-05 で，Adam optimizer [16] ($\beta_1 = 0.9$，$\beta_2 = 0.999$) によりパラメータを最適化した．学習は最大 60 エポック行い，最も検証データに対する損失の低いモデルを選択した．最大テキスト長は 512 で，語彙サイズは 32,103 である．学習バッチサイズは 64 である．検証・評価時のビームサイズは 4 とした．

セルが記述されたと判定する閾値は，検証データでチューニングした結果，0.50 とした．また，ベースラインモデルの repetition penalty 値は，検証データでチューニングした結果，1.90 とした．

3) https://github.com/google/sentencepiece

表 3 自動評価結果（Wikiperson）.

	PARENT			PARENT-T			テキスト長
	Pre.	Rec.	F1	Pre.	Rec.	F1	
T5	0.673	**0.567**	**0.602**	0.421	0.911	**0.568**	80.9
rep. penalty	0.658	0.560	0.592	0.408	**0.916**	0.557	76.1
提案モデル	**0.693**	0.533	0.589	**0.422**	0.877	0.563	67.3

表 4 自動評価結果（SynthBio）.

	PARENT			PARENT-T			テキスト長
	Pre.	Rec.	F1	Pre.	Rec.	F1	
T5	0.618	0.461	0.521	**0.508**	0.637	0.559	257.1
rep. penalty	**0.620**	0.488	0.540	0.491	0.684	0.565	212.3
提案モデル	0.611	**0.516**	**0.550**	0.482	**0.746**	**0.579**	243.7

4.2 実験結果

Wikiperson による評価結果を表 3 に，Synthbio による評価結果を表 4 に示す．Wikiperson による評価において，提案モデルはベースラインモデルよりも，PARENT F1，PARENT-T F1 が低い．提案モデルは推論時，セルが 1 回目で記述されたらマスクするため，1 つのセルは 1 回のみ記述される．Wikiperson は表 1 に示すように，セルの約 18.0%が 2 回以上言及されるため，Wikiperson において提案手法の有効性が示せなかったと考えられる．

それに対して，SynthBio による評価において，提案モデルの PARENT F1，PARENT-T F1 ともにベースラインモデルよりも高い．提案モデルはベースラインモデルと比較してテキスト長が短くなっており，しかも評価指標の F1 が向上していることから，冗長性が減った上で網羅性が向上したことがわかる．

5 結論

本研究では，冗長性の改善と網羅性の向上を目的として，記述された表内のセルを削除し，表を段階的に削除する手法を提案した．評価実験の結果，提案モデルは PARENT，PARENT-T F1 ともにベースラインモデルを上回った．特に，提案モデルはベースラインモデルと比較してテキスト長が短くなっており，評価指標の F1 が向上したことから，冗長性が改善した上で網羅性が向上したと言える．

謝辞

本研究を実施するにあたり，産業技術総合研究所 人工知能研究センター 高村大也氏，奈良先端科学技術大学院大学 上垣外英剛准教授，東京工業大学 小林尚輝氏に大変お世話になりました．深く感謝申し上げます．

参考文献

[1] Rémi Lebret, David Grangier, and Michael Auli. Neural text generation from structured data with application to the biography domain. In **Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing**, pages 1203–1213, Austin, Texas, November 2016. Association for Computational Linguistics.

[2] Soichiro Murakami, Akihiko Watanabe, Akira Miyazawa, Keiichi Goshima, Toshihiko Yanase, Hiroya Takamura, and Yusuke Miyao. Learning to generate market comments from stock prices. In **Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pages 1374–1384, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[3] Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. The E2E dataset: New challenges for end-to-end generation. In **Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue**, pages 201–206, Saarbrücken, Germany, August 2017. Association for Computational Linguistics.

[4] Lya Hulliyyatus Suadaa, Hidetaka Kamigaito, Kotaro Funakoshi, Manabu Okumura, and Hiroya Takamura. Towards table-to-text generation with numerical reasoning. In **Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)**, pages 1451–1465, Online, August 2021. Association for Computational Linguistics.

[5] 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**, 21(140):1–67, 2020.

[6] Qingyun Wang, Xiaoman Pan, Lifu Huang, Boliang Zhang, Zhiying Jiang, Heng Ji, and Kevin Knight. Describing a knowledge base. In **Proceedings of the 11th International Conference on Natural Language Generation**, pages 10–21, Tilburg University, The Netherlands, November 2018. Association for Computational Linguistics.

[7] Ann Yuan, Daphne Ippolito, Vitaly Nikolaev, Chris Callison-Burch, Andy Coenen, and Sebastian Gehrmann. Synthbio: A case study in faster curation of text datasets. In **Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)**, 2021.

[8] Peng Wang, Junyang Lin, An Yang, Chang Zhou, Yichang Zhang, Jingren Zhou, and Hongxia Yang. Sketch and refine: Towards faithful and informative table-to-text generation. In **Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021**, pages 4831–4843, Online, August 2021. Association for Computational Linguistics.

[9] Zhenyi Wang, Xiaoyang Wang, Bang An, Dong Yu, and Changyou Chen. Towards faithful neural table-to-text generation with content-matching constraints. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pages 1072–1086, Online, July 2020. Association for Computational Linguistics.

[10] Tianyu Liu, Xin Zheng, Baobao Chang, and Zhifang Sui. Towards faithfulness in open domain table-to-text generation from an entity-centric view. **Proceedings of the AAAI Conference on Artificial Intelligence**, 35(15):13415–13423, May 2021.

[11] Ran Tian, Shashi Narayan, Thibault Sellam, and Ankur P. Parikh. Sticking to the facts: Confident decoding for faithful data-to-text generation. **CoRR**, abs/1910.08684, 2019.

[12] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, L ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, **Advances in Neural Information Processing Systems**, volume 30. Curran Associates, Inc., 2017.

[13] Bhuwan Dhingra, Manaal Faruqui, Ankur Parikh, Ming-Wei Chang, Dipanjan Das, and William Cohen. Handling divergent reference texts when evaluating table-to-text generation. In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pages 4884–4895, Florence, Italy, July 2019. Association for Computational Linguistics.

[14] Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. CTRL: A conditional transformer language model for controllable generation. **CoRR**, abs/1909.05858, 2019.

[15] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations**, pages 38–45, Online, October 2020. Association for Computational Linguistics.

[16] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, **3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings**, 2015.