# コピー機構と長さ正規化を用いたData-to-text生成

川添 正太郎　　西川 仁　　徳永 健伸
東京工業大学 情報理工学院
{kawazoe.s.aa@m, hitoshi@c, take@c}.titech.ac.jp

## 1 はじめに

Data-to-text 生成は非言語的なデータを入力としてテキストを生成する課題である [6]．本研究は属性と値の組からなる表を入力とし，表中の重要な情報をテキストとして生成する課題を扱う．表からのテキスト生成は質問応答システムや対話エージェントを構築するための基礎技術として期待されている [2]．近年，テキスト生成の手法として seq2seq [1] が広く使われているが，seq2seq は以下の二つの問題をもっている．

**(1) 未知語問題**
seq2seq の辞書は，計算コストを抑えるために一定数の高頻度語から構築する．このため，異なり数では多くを占める低頻度語は辞書に含まれず，生成時には辞書に含まれない語を特別な記号 UNK で置換してテキストを生成する．テキスト中のこのような特殊記号はテキストの質を低下させる．

**(2) 短文生成問題**
seq2seq は短いテキストを生成する傾向がある．この傾向は，ニューラル機械翻訳の分野で議論されている [4, 10]．正解テキストより短いテキストには言及されるべき重要な情報が欠落している可能性が高い．さらに，テキスト生成の評価指標として使われる BLEU [5] には短いテキストの評価を下げる brevity penalty が組み込まれているため，短いテキストの BLEU スコアは小さくなる傾向がある．この傾向は機械翻訳の分野で指摘されているが [10]，Data-to-text 生成においては指摘されてこなかった．

本研究では，未知語問題を緩和するために，コピー機構 (copy mechanism) [7] を用いて入力から出力へ語をコピーし，辞書に含まれない語の出力を可能にする．また，短文生成問題を緩和するために，長さ正規化 (length normalization) [9] を用いてデコーディング時に短いテキストが生成されることを抑制する．提案手法を評価するために WikiBio [3] データセットを使用する．WikiBio は人物に関する Wikipedia の表と人物紹介文の組からなる (表 1)．

本研究の貢献は以下のとおりである:

- 自動評価と人手評価の両方を行い，コピー機構と長さ正規化が生成されるテキストの質を改善することを示す．提案手法は最先端のモデルを BLEU スコアで 1.04 ポイント上回る．
- brevity penalty が，機械翻訳の分野だけでなく，WikiBio においても BLEU スコアに大きな影響を与えることを示す．

表

| | |
|---|---|
| **Name** | Kava Huihahau |
| **Fullname** | Kava Huihahau |
| **Birth Date** | 08 August 1982 |
| **Birth Place** | Tonga |
| **Position** | Defender |
| **National Years** | 2003 – 2007 |
| **National Team** | Tonga |
| **National Caps** | 10 |
| **National Goals** | 0 |

人物紹介文
Kava Huihahau (born 8 August 1982) is a Tongan former international footballer who played as a defender.

表 1: サッカー選手に関する Wikipedia の表，およびその内容を説明する人物紹介文．表の左列が属性，右列が値である．

## 2 手法

### 2.1 定式化

入力の表は，属性と値の組からなり，それぞれの値は一つ以上の語から構成されている (表 1)．それぞれの語に一つの属性を対応付けることができるため，入力の表は属性と語の組の系列とみなせる．形式的には，入力の表は $X = \{(f_i, w_i)\}_{i=1}^{|X|}$ と表される．ただし，$f_i$ は属性，$w_i$ は語，$|X|$ は $X$ に含まれる語の総数である．例えば，表 1 の表は，形式的には $\{(\text{Name}, \text{Kava}), (\text{Name}, \text{Huihahau}), \ldots\}$ と表される．

入力の表 $X$ が与えられたとき，モデルは出力される語の系列，すなわち文 $Y = \{y_t\}_{t=1}^{|Y|}$ を生成する．ただし，$y_t$ は出力される語，$|Y|$ は出力される文の長さである．

### 2.2 位置の入力系列への付与

先行研究にならい [3]，入力系列に位置 $p_i^+$ および $p_i^-$ を組み込む．$p_i^+$ は，$w_i$ を含む値の先頭から数えた $w_i$ の位置である．$p_i^-$ は，$w_i$ を含む値の末尾から後ろ向きに数えた $w_i$ の位置である．$p_i^+$ と $p_i^-$ を組み込むことにより，表 1 の表は，表 2 のように，$w_i$，$f_i$，$p_i^+$，

| $i$ | $f_i$ | $w_i$ | $p_i^+$ | $p_i^-$ |
|---|---|---|---|---|
| 1 | Name | Kava | 1 | 2 |
| 2 | Name | Huihahau | 2 | 1 |
| 3 | Fullname | Kava | 1 | 2 |
| 4 | Fullname | Huihahau | 2 | 1 |
| 5 | Birth Date | 08 | 1 | 3 |
| 6 | Birth Date | August | 2 | 2 |
| 7 | Birth Date | 1982 | 3 | 1 |
| | | ⋮ | | |

表 2: 表 1 の表の，入力系列としての表現．

および $p_i^-$ の組の系列として表される．これらの埋め込みベクトル $\boldsymbol{w}_i$，$\boldsymbol{f}_i$，$\boldsymbol{p}_i^+$，および $\boldsymbol{p}_i^-$ を結合したベクトル $[\boldsymbol{w}_i; \boldsymbol{f}_i; \boldsymbol{p}_i^+; \boldsymbol{p}_i^-]$ がエンコーダへの入力となる．ただし，$[;]$ はベクトルの結合を表す．

### 2.3 コピー機構

seq2seq では，辞書に含まれるすべての語 $w$ に対して，$w$ が出力される確率 $P_{\text{vocab}}(w)$ を計算する．デコーディングの時刻 $t$ において，入力の表 $X$ およびこれまでに出力された語の系列 $y_{<t} = \{y_1, ..., y_{t-1}\}$ が与えられたとき，語 $y_t$ が出力される確率は以下で定式化される:

$$P(y_t|y_{<t}, X) = P_{\text{vocab}}(y_t). \tag{1}$$

ここで，$y_t$ が辞書に含まれない場合，$P(y_t|y_{<t}, X) = P_{\text{vocab}}(y_t) = 0$ となる．

式 (1) により，seq2seq は辞書に含まれる語しか出力できないため，未知語問題を引き起こす．この問題を緩和するために，式 (1) にコピー機構を組み込んだ pointer generator [7] を用いる．これは以下で定式化される:

$$P(y_t|y_{<t}, X) = p_{\text{gen}} P_{\text{vocab}}(y_t) + (1 - p_{\text{gen}}) \sum_{i:w_i=y_t} a_i^t. \tag{2}$$

ただし，$p_{\text{gen}} \in [0, 1]$ は各 $t$ に対して決まる数，$\{a_i^t\}_{i=1}^{|X|}$ は注意機構の重みである．式 (2) で，pointer generator は注意機構の重みを，入力から出力へと語がコピーされる確率として再利用している．これにより，$y_t$ が辞書に含まれていなくても，$\{w_i\}_{i=1}^{|X|}$ に含まれていれば，確率 $P(y_t|y_{<t}, X)$ を割り当てることができる．

### 2.4 デコーディング

入力の表 $X$ が与えられたとき， デコーダはスコア関数を最大にするような出力系列 $Y$ を探索する．標準的なスコア関数は以下で計算される:

$$s(X, Y) = \sum_{t=1}^{|Y|} \log P(y_t|y_{<t}, X). \tag{3}$$

$\log P(y_t|y_{<t}, X)$ は常にゼロ以下の値をとるため，$s(X, Y)$ は系列長 $|Y|$ が大きくなるにつれて単調に減少する．このスコア関数の性質が短文生成問題を引き起こす一因となる．短いテキストは，言及されるべき重要な情報が欠落している可能性があるだけでなく，BLEU スコアを低下させる原因にもなる．

### 2.5 なぜ短文生成問題が BLEU を減少させるのか

BLEU スコア [5] は $n$-gram 適合率に基づく評価指標である．$n$-gram 適合率は，短いテキストで大きくなる傾向がある．この傾向を補正するため，Papineni ら [5] は BLEU スコアに，短いテキストの評価を下げる brevity penalty を導入した．brevity penalty は長さ比 (length ratio) の関数である．これは $lr = c/r$ で計算される．ただし，$c$ はコーパス内で生成されたすべてのテキストの長さの合計，$r$ はコーパス内のすべての正解テキストの長さの合計である．次に，brevity penalty は以下で計算される:

$$bp = \begin{cases} e^{1-1/lr} & (lr \leq 1) \\ 1 & (lr > 1) \end{cases}. \tag{4}$$

コーパス全体に対する BLEU (BLEU-4) スコアは以下で計算される:

$$\text{BLEU} = bp \cdot \exp\left(\frac{1}{4}\sum_{n=1}^{4} \log p_n\right). \tag{5}$$

ただし，$p_n$ は修正 $n$-gram 適合率 (modified $n$-gram precision) [5] である．式 (4) および (5) から，brevity penalty は長さ比が 1 より小さいときに BLEU スコアを減少させる．この結果，短文生成問題は BLEU スコアを低下させる原因となる．

### 2.6 長さ正規化

短文生成問題を緩和するために，式 (3) を，長さ正規化 [9] を導入することで修正する．長さ正規化を導入したスコア関数 $s_{\text{LN}}(X, Y)$ は以下で計算される:

$$lp(Y) = \frac{(5 + |Y|)^\alpha}{(5 + 1)^\alpha}, \tag{6}$$

$$s_{\text{LN}}(X, Y) = s(X, Y)/lp(Y). \tag{7}$$

ただし，$\alpha$ はハイパーパラメータである．$\alpha$ を増加させることで生成されるテキストは長くなる．このため，$\alpha$ をチューニングすることで，brevity penalty の影響を最も緩和するような長さ比を探索することができる．

## 3 実験

### 3.1 データセット

本研究は，WikiBio [3] データセットを用いる．WikiBio は人物に関する Wikipedia の表と人物紹介文の組からなり，全部で 728,321 個の組を含む．人物紹介文は，Wikipedia の記事の最初の一文から抽出さ

| 手法 | BLEU |
|---|---|
| KN [3] | 2.21 |
| NLM [3] | 4.17 |
| Template KN [3] | 19.80 |
| Table NLM [3] | 34.70 |
| Copy mechanism+EMA [8] | 46.76 |
| S2S | 38.35 |
| PG | 42.95 |
| PG+LN | **47.80** |

表 3: 自動評価の結果.

れたものである．Lebret ら [3] に従い，データセットを訓練集合 (80%)，検証集合 (10%)，およびテスト集合 (10%) に分割する．

### 3.2 ベースライン

提案手法を，以下の従来手法と比較する．(1) Kneser-Ney 言語モデル (KN) [3]，(2) ニューラル言語モデル (NLM) [3]，(3) テンプレートを用いた Kneser-Ney 言語モデル (Template KN) [3]，(4) 表に条件付けられた条件付きニューラル言語モデル (Table NLM) [3]，(5) コピー機構と指数移動平均を用いたモデル (Copy mechanism+EMA) [8].

本研究では，seq2seq (S2S)，pointer generator (PG)，および pointer generator と長さ正規化を組み合わせたモデル (PG+LN) を実装し，これらを従来手法と比較する．

### 3.3 評価指標

#### 自動評価

Lebret ら [3] に従い，BLEU-4 スコアを自動評価指標として用い，その計算には `mteval-v13a.pl` スクリプトを用いた．

#### 人手評価

流暢さ (fluency) と情報性 (informativeness) を人手評価指標として用いた．Amazon Mechanical Turk を用い，流暢さと情報性について個別に評価タスクを設計した．各評価タスクについて，正解文 (REF)，S2S，PG，および PG+LN の計四文の組を無作為に 200 個選び，各組に 10 名のワーカー (評価者) を割り当てた．評価対象となる文の順序を無作為に並び替えてワーカーに提示し，5 点のリッカート尺度で評価させた．

流暢さの評価タスクでは，REF，S2S，PG，および PG+LN の計四文を，1 (最も流暢でない) から 5 (最も流暢である) で評価した．情報性の評価タスクでは，S2S，PG，および PG+LN の計三文を，REF の情報をどれだけ網羅しているかに基づき，1 (最も情報が豊富でない) から 5 (最も情報が豊富である) で評価した．

| $P$ | #W | #R | S2S | PG | PG+LN | REF |
|---|---|---|---|---|---|---|
| 流暢さ | | | | | | |
| 0 | 119 | 2000 | 3.55 | 3.68 | 3.67 | 3.61 |
| 5 | 107 | 1710 | 3.61 | 3.75 | 3.74 | 3.68 |
| 10 | 95 | 1335 | 3.58 | 3.77 | 3.75 | 3.68 |
| 15 | 83 | 991 | 3.68 | 3.86 | 3.79 | 3.71 |
| 20 | 71 | 743 | 3.53 | 3.77 | 3.66 | 3.57 |
| 25 | 59 | 500 | 3.41 | 3.75 | 3.59 | 3.47 |
| 情報性 | | | | | | |
| 0 | 131 | 2000 | 3.30 | 3.44 | 3.47 | - |
| 5 | 117 | 1741 | 3.34 | 3.48 | 3.51 | - |
| 10 | 105 | 1476 | 3.36 | 3.50 | 3.55 | - |
| 15 | 91 | 1142 | 3.34 | 3.53 | 3.57 | - |
| 20 | 79 | 962 | 3.38 | 3.55 | 3.62 | - |
| 25 | 65 | 689 | 3.44 | 3.62 | 3.76 | - |

表 4: 人手評価の結果．$P$: カットオフ閾値，#W: ワーカー，#R: 評価，REF: 正解文．

| #U | #S | BLEU | #R | F | I |
|---|---|---|---|---|---|
| 0 | 20,929 | 44.39 | 770 | 3.70 | 3.35 |
| 1 | 31,176 | 38.53 | 560 | 3.56 | 3.32 |
| 2 | 13,776 | 34.09 | 400 | 3.47 | 3.27 |
| 3 | 4,379 | 30.13 | 180 | 3.36 | 3.27 |
| 4 | 1,610 | 26.52 | 70 | 3.21 | 3.14 |

表 5: UNK 記号の数の S2S の性能への影響．#U: UNK 記号，#S: 文，#R: 評価，F: 流暢さ，I: 情報性．

## 4 結果

### 4.1 自動評価

表 3 に自動評価の結果を示す．S2S に比べて，PG は BLEU を 4.60 改善している．さらに，PG に比べて，PG+LN は BLEU を 4.85 改善している．従来研究と比較すると，PG+LN は最先端のモデル [8] を BLEU で 1.04 上回っている．

### 4.2 人手評価

一部のワーカーの評価時間が極端に短い，あるいは長い場合があった．これらの評価は信頼性に欠けると判断したため，カットオフ閾値 $P$ を用いて，評価時間の平均が上位 $P$%または下位 $P$%となるワーカーを除去した．ただし，真の $P$ を決める手段はないため，$P = 0, 5, 10, 15, 20, 25$ でのすべての結果を示す．

表 4 に人手評価の結果を示す．$P$ に関わらず，スコアの順位は，流暢さで PG>PG+LN>REF>S2S，情報性で PG+LN>PG>S2S である．REF (正解文) が流暢さで三位となった理由として，1) WikiBio は自動で構築されているため [3] 人物紹介文に非文が含まれることがある，2) アラビア文字のような英語でない文字が含まれることなどが考えられる．

| 手法 | UNK% |
|---|---|
| S2S | 6.01% (85,182/1,416,411) |
| PG | 0.71% (10,042/1,406,827) |
| PG+LN | 1.41% (26,939/1,906,169) |

表 6: UNK 記号の割合.

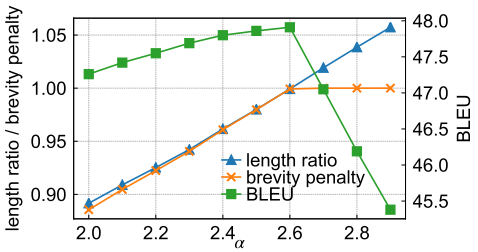


図 1: 長さ比，brevity penalty，および BLEU スコアに対する $\alpha$ の影響.

## 5 議論

### 5.1 コピー機構の効果

表 5 は，UNK 記号の数が S2S の性能にどれだけ影響するかを示している．UNK 記号が増えるにつれ，すべての指標が減少している．この結果は，未知語問題が，生成されたテキストの質を下げることを示唆している．

表 6 はテスト集合において出力されたすべての語に対する UNK 記号の割合を示している．コピー機構により，UNK 記号の割合は 6.01% (S2S) から 0.71% (PG) へと減少している．この減少は，コピー機構が未知語問題を緩和し，テキストの質の改善につながったことを示唆している．

### 5.2 長さ正規化の効果

図 1 は，式 (6) の $\alpha$ の，長さ比，brevity penalty，および BLEU スコア (それぞれ検証集合における PG+LN での値) への効果を示している．この図から，以下のことが観測できる．1) $\alpha$ が増加するほど長さ比も増加する．2) 長さ比が約 1.0 となり brevity penalty の増加が止まったとき，BLEU スコアが最大になる．これらの観測は 2.6 節における説明と一貫している．また，テスト集合では，PG の長さ比は 0.74，PG+LN では 1.00 である．この長さ比の増加は，長さ正規化が brevity penalty の影響を緩和することによって BLEU スコアを改善したことを示唆している．

### 5.3 生成例

表 7 に生成例を示す．S2S は，辞書に含まれない正解文の語 “schumnig” を出力できず，これを UNK 記号に置き換えている．反対に，PG および PG+LN は “schumnig” を出力できた．これらの例は，コピー機構が未知語問題を緩和したことを示唆している．さらに，PG+LN は “who currently plays for ec kac of the austrian hockey league ( ebel )” を生成し，これは正解文に含まれる “who is currently playing with ec kac of the austrian hockey league ( ebel )” とほとんど一致している．この一致は，長さ正規化が短文生成問題を緩和したことを示唆している．

| | |
|---|---|
| **REF** | martin *schumnig* ( born july 28 , 1989 ) is an austrian professional ice hockey defenceman *who is currently playing with ec kac of the austrian hockey league ( ebel )* . |
| **S2S** | martin UNK ( born july 28 , 1989 ) is an austrian ice hockey defenceman . |
| **PG** | martin schumnig ( born july 28 , 1989 ) is an austrian professional ice hockey defenceman . |
| **PG+LN** | martin schumnig ( born july 28 , 1989 ) is an austrian professional ice hockey defenceman who currently plays for ec kac of the austrian hockey league ( ebel ) . |

表 7: 生成されたテキストの例.

## 6 おわりに

本研究は Data-to-text 生成に取り組み，seq2seq がもつ二つの問題，未知語問題と短文生成問題を指摘した．本研究は，これらの問題をコピー機構と長さ正規化を用いて緩和した．WikiBio データセットにおける結果は，提案手法が最先端のモデルを BLEU で 1.04 上回ることを示した．また，人手評価の結果は，流暢さと情報性の観点で提案手法の有効性を示した．

## 参考文献


[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *ICLR*, 2015.

[2] Jun-Wei Bao, Duyu Tang, Nan Duan, Zhao Yan, Yuanhua Lv, Ming Zhou, and Tiejun Zhao. Table-to-text: Describing table region with natural language. In *AAAI*, 2018.

[3] Rémi Lebret, David Grangier, and Michael Auli. Neural text generation from structured data with application to the biography domain. In *EMNLP*, 2016.

[4] Kenton Murray and David Chiang. Correcting length bias in neural machine translation. In *WMT*, 2018.

[5] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2002.

[6] Ehud Reiter and Robert Dale. *Building natural language generation systems*. Cambridge university press, 2000.

[7] Abigail See, Peter J. Liu, and Christopher D. Manning. Get to the point: Summarization with pointer-generator networks. In *ACL*, 2017.

[8] Hamidreza Shahidi, Ming Li, and Jimmy Lin. Two birds, one stone: A simple, unified model for text generation from structured and unstructured data. *CoRR*, 2019.

[9] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, 2016.

[10] Yilin Yang, Liang Huang, and Mingbo Ma. Breaking the beam search curse: A study of (re-)scoring methods and stopping criteria for neural machine translation. In *EMNLP*, 2018.