

# 位置エンコーディングを用いた出力長制御

高瀬 翔      岡崎 直観  
東京工業大学

{sho.takase@nlp., okazaki@c.titech.ac.jp}

## 1 はじめに

ニューラルネットを用いたエンコーダ・デコーダモデルは機械翻訳、要約、キャプション生成など、様々な自然言語生成タスクで高性能を達成している [9, 11, 15]. ただし、要約生成の実用では出力の長さ制限が課せられることがある. 多くのエンコーダ・デコーダモデルは出力の長さを考慮しないため、図1に示すように、正解とエンコーダ・デコーダの出力との長さに大きな乖離が生じる. そこで、エンコーダ・デコーダモデルに長さ制御の機構を組み込む方法が提案されている.

Fan ら [1] は出力長を制御するため、エンコーダ・デコーダモデルに出力長を表現する埋め込みを組み込んだ. この手法では、各長さの埋め込み表現は独立であるため、訓練データにおいて稀な長さの出力文を生成することが原理的に難しい. これに対し、出力長を離散的ではなく連続的に表現して、エンコーダ・デコーダモデルに組み込む手法も提案されている.

Kikuchi ら [2] と Liu ら [6] はそれぞれ、LSTM と CNN の内部ベクトルを出力長に応じて定数倍する手法を提案した. これらの手法では、デコーダの初期ステップに所望の長さを与えるが、以降の生成ステップにおいて、長さに関する信号が与えられることはない. これらに対し、本研究では生成時のトークンを出力するごとに、終端までの残り長さを明に与える手法を提案する. 具体的には、機械翻訳において高い性能を達成している Transformer [14] の位置エンコーディングに出力長の情報を組み込む. 提案手法は出力長を連続的に表現するため、訓練データに出現していない長さの出力も可能である. 要約生成タスクの一つである、見出し文生成において長さ制御の実験を行い、提案手法が出力の長さを制御できるだけでなく、ROUGE スコアも向上させることを示す.

## 2 位置エンコーディング

Transformer [14] では、トークンの埋め込み表現にそのトークンの位置に対応した位置エンコーディング (positional encoding) を足し合わせ、エンコーダおよびデコーダに入力する. トークンの位置を  $pos$ , 埋め込み表現の次元数を  $d$  とすると、位置エンコーディングの  $i$  番目の次元の値  $PE_{(pos,i)}$  は次式で得られる.

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{\frac{2i}{d}}}\right) \quad (1)$$

$$PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{\frac{2i}{d}}}\right) \quad (2)$$

つまり、位置エンコーディングの各次元は周期  $10000^{2i/d} \times 2\pi$  の正弦波 (と余弦波) で表現される. 位置エンコーディングは

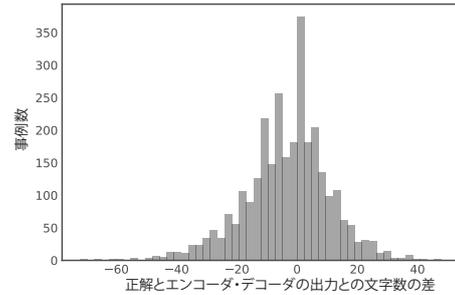


図1: Annotated English Gigaword のテストセット [9] における正解の見出し文と LSTM エンコーダ・デコーダ [7] で生成した見出し文との文字数の差

位置のみに依存した値である、すなわち、トークンの絶対位置を表現している.

本研究では、式 (1) と (2) を与えられた出力長までの距離に依存した式に拡張する. デコーダ側の位置エンコーディングを提案した式に置き換えることで、文生成を行う際に所望の出力長までの残りを明に考慮する. 本稿では、終端までの比率に応じた位置エンコーディング (length-ratio positional encoding,  $LRPE$ ) と、終端までの差に応じた位置エンコーディング (length-difference positional encoding,  $LDPE$ ) の2種類を提案する.  $LRPE$  と  $LDPE$  は、それぞれ以下の式で計算する.

$$LRPE_{(pos,len,2i)} = \sin\left(\frac{pos}{len^{\frac{2i}{d}}}\right) \quad (3)$$

$$LRPE_{(pos,len,2i+1)} = \cos\left(\frac{pos}{len^{\frac{2i}{d}}}\right) \quad (4)$$

$$LDPE_{(pos,len,2i)} = \sin\left(\frac{len - pos}{10000^{\frac{2i}{d}}}\right) \quad (5)$$

$$LDPE_{(pos,len,2i+1)} = \cos\left(\frac{len - pos}{10000^{\frac{2i}{d}}}\right) \quad (6)$$

ここで、 $len$  は所望の出力長を表す.  $LRPE$  は終端に対する、残り長さの比率が同程度の位置に対して、似た値を返す. 最も単純な例として、 $d$  番目の次元、すなわち、 $2i = d$  (または  $2i + 1 = d$ ) について考えると、 $LRPE$  は  $\sin(pos/len)$  (あるいは  $\cos(pos/len)$ ) となるため、 $pos = 5$  かつ  $len = 10$  や、 $pos = 10$  かつ  $len = 20$  のように、残り長さの比率が同じとなる位置については同じ値が得られる.  $LDPE$  は終端までの残り長さが同じとなる位置に対して、常に同じ値を返す.

$LRPE$  と  $LDPE$  はデコーダ側において、ベースとなる Transformer [14] と同様、入力トークンの埋め込み表現と足し

合わせる。訓練時には、正解出力の長さを所望の出力長  $len$  として学習し、テスト時には、 $len$  に所望の値を代入することで出力長の制御を行う。

### 3 実験

#### 3.1 データセット

提案手法が指定された長さで質の良い見出し文を生成できるか検証するため、日本語と英語の見出し文生成タスクで実験を行った。日本語の見出し生成には人見らの構築したデータセットを用いる [17]。訓練には新聞記事の先頭 3 文と見出し文の約 160 万対を用いた。評価時には、新聞記事 1 つに対して 10 ( $len = 10$ ), 13 ( $len = 13$ ), 26 ( $len = 26$ ) 文字以内という文字数制限で作成された 3 種類の見出し 1,181 件を用いた。この評価データは新聞社で実際に使用された見出しを収録しているため、実応用を想定した性能評価に近い。

英語の見出し生成では、Annotated English Gigaword に Rush ら [9] の前処理スクリプト\*1を適用し、得られた訓練データ（記事の先頭の文と見出し文の約 380 万対）を学習に用いた。一方、評価について、日本語とは異なり、ある記事に対して複数の長さの見出し文を収録したデータは存在しない。そこで、Rush らの前処理スクリプトで得られたテストセットから、特定の長さの見出し文を、それぞれの長さについて 3,000 件ずつランダムに抽出し、これを評価データと見なすことで、文字数制限付きの評価を模擬した。長さの条件としては、文字数 1 から 30 ( $len = 30$ ), 30 から 50 ( $len = 50$ ), 50 から 75 ( $len = 75$ ) を採用した。さらに、既存研究で報告されている評価結果と比較するため、DUC-2004 [8] での評価も行った。

なお、日本語、英語のいずれの言語においても、ひとつの記事に対して異なる長さの見出し文を大量に含んだデータはないため、ひとつの記事に対してひとつの見出しが付与された訓練データを用いている。

本実験では、出力文字数の制御が可能かを検証するため、デコーダ側は文字単位の語彙を用いた。これに対し、エンコーダ側はサブワード [3, 10] を用いて語彙を構成した。サブワード語彙を構成する際のハイパーパラメータは、サイズが概ね 16,000 となるように設定した。

#### 3.2 ベースライン

既存研究で提案された、2 種の出力長制御手法と Transformer をベースラインとして用いる。

**LenInit** Kikuchi ら [2] はデコーダ側の LSTM の記憶セルの初期値によって、出力長を制御する手法 (LenInit) を提案した。

$$m = len \times b, \quad (7)$$

ここで、 $b$  は訓練対象となるベクトルである。本実験では、広く使われている、Luong らの LSTM エンコーダ・デコーダ [7]\*2 に LenInit を組み込んだ。公平な比較のため、見出し文生成において、シンプルな LSTM エンコーダ・デコーダで最高性能を達成しているハイパーパラメータ [13] を用いて学習を行った。

\*1ただし、実応用に近い状況とするため、以下のスクリプトから低頻度語と数字の置換処理を除去したものを用いた。  
<https://github.com/facebookarchive/NAMAS>

\*2実装としては <https://github.com/mlpnl/mlpnl-nmt> を用いた。

**Length Control (LC)** Liu ら [6] は CNN エンコーダ・デコーダにおいて、入力トークンの埋め込み表現に、所望の出力長を掛け合わせることで、出力長を制御する手法 (Length Control, LC) を提案した。本実験では、Liu らと同様のハイパーパラメータを用いて学習を行った。

**Transformer** 提案手法は Transformer [14] を元としているため、オリジナルの Transformer との比較も行う\*3。ハイパーパラメータは Vaswani らのベースモデルと同様の値を用いた。

#### 3.3 結果

表1は日本語テストデータにおける、各モデルでの ROUGE-1 (R-1), 2 (R-2), L (R-L) を示している。Transformer と提案手法の組み合わせ (Transformer+LRPE, Transformer+LDPE) は、各長さ ( $len = 10, 13, 26$ ) において、全てのベースラインよりも高い ROUGE 値を達成している。加えて、LRPE では、絶対位置を示す位置エンコーディング (+PE) を導入した際に性能が向上している。LRPE については、出力長を制御しつつ、生成結果の質を向上させるために、絶対位置を示す位置エンコーディングも併用することが有効であると考えられる。

表2は English Gigaword から構築したテストデータにおける ROUGE 値を示している。この表において、提案手法 (LRPE, LDPE) は  $len = 75$  において、ベースラインよりもスコアが大きく向上している。また、LRPE について、表1と同様、絶対位置を導入した際にスコアが向上しており、提案手法では最も高いスコアを達成している。一方で、 $len = 30, 50$  において、提案手法の性能はベースの Transformer より低いスコアとなっている。この結果は、訓練データ中に 50 文字以下の見出し文が多数を占め、ベースの Transformer も 30 文字から 50 文字の出力が多いためであると考えられる。しかしながら、節3.4で述べるように、ベースの Transformer の出力のほとんどは、所望の長さとは大きく異なっている。

表3は DUC-2004 テストデータにおける ROUGE 値を示している。この表においては、提案手法として、英語テストデータで性能の高かった LRPE のスコアを示している。既存の評価手順 [8] に従い、本研究でも生成結果の先頭 75 byte で評価を行った。提案手法である LRPE はベースの Transformer からスコアを大きく向上させ、LenInit の R-2 を除き、ベースラインよりも高いスコアを達成している。さらに、PE を加えると、全てのベースラインよりも高いスコアを達成した。この表には既存研究で報告されているスコアも記した。提案手法は既存の出力長制御モデルよりも高いスコアを達成し、最高性能のモデル [12] と同程度のスコアとなっていることが分かる。

本研究では文字単位のデコーダを用いているため、元記事とは無関係な単語を生成してしまう場合がある。これに対処するため、上位  $m$  件の生成結果について\*4、元記事の単語を含んでいる数でリランキングを行った。提案手法とリランキングの組み合わせ (表3における +リランキング) は最も高い ROUGE 値を達成した。

\*3 <https://github.com/pytorch/fairseq> で公開されている実装を用いた。

\*4 本実験では  $m = 20$  とした。

モデル	len = 10			len = 13			len = 26		
	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
ベースライン									
LenInit	37.03	16.83	35.76	40.97	18.93	38.47	46.11	21.41	37.54
LC	36.56	16.34	35.41	40.52	19.10	38.52	42.60	20.04	36.12
Transformer	32.71	14.67	31.31	42.06	20.70	39.12	43.74	21.04	36.18
提案手法									
Transformer+LRPE	41.45	20.47	40.10	<b>46.14</b>	22.72	<b>43.56</b>	50.20	24.38	40.48
+PE	42.20	<b>21.05</b>	<b>40.95</b>	45.71	<b>23.04</b>	43.16	<b>50.94</b>	24.43	41.17
Transformer+LDPE	<b>42.32</b>	20.07	40.83	45.97	22.02	43.30	50.76	<b>24.56</b>	<b>41.21</b>
+PE	41.85	19.77	40.59	45.06	21.33	42.22	50.81	24.34	41.02

表1: 日本語テストデータにおいて、各長さの見出し文を生成した際の ROUGE 値。

モデル	len = 30			len = 50			len = 75		
	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
ベースライン									
LenInit	45.38	26.46	44.05	48.04	25.27	45.20	51.14	25.95	46.81
LC	45.13	26.88	44.01	46.57	24.53	44.16	48.41	24.54	44.92
Transformer	47.53	<b>29.74</b>	<b>46.38</b>	49.62	<b>27.64</b>	46.86	45.55	23.49	42.22
提案手法									
Transformer+LRPE	46.88	27.37	45.38	49.71	25.70	46.62	53.19	27.23	<b>48.49</b>
+PE	<b>47.77</b>	28.07	46.29	<b>50.18</b>	26.02	<b>46.97</b>	<b>53.32</b>	<b>27.28</b>	48.43
Transformer+LDPE	47.49	27.17	45.99	49.64	25.74	46.47	53.10	27.06	48.30
+PE	47.12	26.97	45.64	49.45	25.45	46.38	52.81	26.90	48.11

表2: Annotated English Gigaword から抽出した各長さに対するテストデータにおける ROUGE 値。

モデル	R-1	R-2	R-L
ベースライン			
LenInit	29.56	11.07	26.26
LC	28.87	10.73	25.68
Transformer	25.30	8.89	22.56
提案手法			
Transformer+LRPE	30.87	10.72	26.71
+PE	31.33	11.12	27.24
+リランキング	<b>32.32</b>	<b>11.43</b>	<b>28.04</b>
出力長制御を行う既存手法			
Kikuchi ら [2]	26.73	8.39	23.88
Fan ら [1]	30.00	10.27	26.43
上記以外の既存研究			
Rush ら [9]	28.18	8.49	23.81
Suzuki ら [12]	32.28	10.54	27.80
Zhou ら [16]	29.21	9.56	25.51
Li ら [5]	31.79	10.75	27.48
Li ら [4]	29.33	10.24	25.24

表3: DUC-2004 における ROUGE 値。

### 3.4 出力長の正確性

Lin ら [6] に従い、生成した要約と所望の長さとの分散を用いて、出力長の制御がどの程度正確にできているか検証する。生成した  $n$  個の要約文に対し、分散 ( $var$ ) を以下のように計算する。

$$var = 0.001 \times \frac{1}{n} \sum_{i=1}^n |l_i - len|^2 \quad (8)$$

ここで、 $len$  は所望の長さ、 $l_i$  は生成した見出し文の長さである。

表4に、各手法の各所望の長さに対する分散を示した。この表から、提案手法である LRPE, LDPE は見出し文の出力長を正確に制御できており、既存手法である LenInit や LC と比較し、より所望の長さに近い見出し文を生成できていることが分かる。この結果は、提案手法が生成中のトークンを出力するごとに所望の長さまでの残りを考慮するため、所望の長さで生

成を正確に終わることができることを示唆している。また、終端までの差を考慮する LDPE の方が、LRPE よりも正確に出力長を制御できる。

提案手法は、原理的に、訓練データに出現していない長さの見出し文を生成することができる。このことを検証するため、訓練データから所望の長さに該当する見出し文を除去し、提案モデルの訓練を行った。結果は表4の下段に示している。この表から、所望の長さの訓練事例を除いたデータで学習した場合でも、分散は訓練事例全体を用いた場合と同程度であることが分かる。すなわち、提案手法は、所望の長さが訓練データに出現していない場合でも、見出し文の長さを正確に制御できる。

日本語テストデータにおいて、ランダムに選択した元記事に対する、提案手法 (Transformer+LRPE+PE) で生成した各長さの見出し文と正解の見出し文を図2に示す。提案手法はいくつかの例において、人手と遜色のない見出し文を生成している。しかしながら、図中の下線の例のように、主語と述語の関係が元記事と矛盾する (3行目,  $len = 10, 13$ ), 単語の途中で出力が切れる (4行目,  $len = 26$ ), 修飾先の単語がなく文として適切でない (5行目,  $len = 10$ ) などの問題もある。これらの改善は今後の課題である。

## 4 おわりに

本研究では、要約生成において出力長の制御を行うため、所望の長さに依存する位置エンコーディングを提案した。実験により、提案手法は所望の長さが訓練データに含まれていない場合でも、所望の長さで出力を行えることが分かった。日本語の評価データでは、出力長の制御を正確に行いつつ、ROUGE 値の向上を確認した。英語の評価データにおいても、出力長の制御が正確に行えることを確認し、DUC-2004 において最高スコアを達成した。

	分散					
	日本語テストデータ			英語テストデータ		
モデル	$len = 10$	$len = 13$	$len = 26$	$len = 30$	$len = 50$	$len = 75$
ベースライン						
LenInit	0.005	0.002	0.002	0.014	0.009	0.006
LC	0.006	0.002	0.004	0.057	0.033	0.039
Transformer	8.989	5.581	2.510	10.824	10.412	27.366
提案手法						
Transformer+LRPE	0.002	0.002	0.001	0.004	0.002	0.003
+PE	0.004	0.002	0.002	0.001	0.003	0.003
Transformer+LDPE	<b>0.000</b>	<b>0.000</b>	<b>0.000</b>	0.002	<b>0.001</b>	<b>0.001</b>
+PE	<b>0.000</b>	<b>0.000</b>	<b>0.000</b>	0.003	<b>0.001</b>	<b>0.001</b>
所望の出力長を除いた訓練データで学習した提案手法						
Transformer+LRPE	0.001	0.002	0.002	<b>0.000</b>	<b>0.001</b>	<b>0.001</b>
+PE	0.004	0.002	0.001	0.002	<b>0.001</b>	0.003
Transformer+LDPE	<b>0.000</b>	<b>0.000</b>	<b>0.000</b>	0.003	<b>0.001</b>	<b>0.001</b>
+PE	<b>0.000</b>	<b>0.000</b>	<b>0.000</b>	0.004	0.002	<b>0.001</b>

表4: 各手法で出力した見出し文の所望の長さとの分散。

元記事	$len = 10$	$len = 13$	$len = 26$
国際通貨基金(IMF)は22日、最新の世界経済見通しを発表した。2018年の世界全体の成長率見通しは3.9%として、昨年10月発表の前回予想から0.2ポイント上方修正した。米トランプ政権が実現した税制改革による影響で、米国内見通しを引き上げたことなどが要因だ。	成長見通しを引き上げ	世界経済成長見通し引き上げ	世界経済の成長見通し引き上げ_トランプ減税の効果で
月の地下に、長さ約50キロに及ぶ長大な空洞があることが、日本の月探査機「かぐや」の観測データから判明した。過去の火山活動で生じたとみられる。宇宙航空研究開発機構(JAXA)が18日、発表した。	世界成長率、上方修正	世界成長予想、IMF上方修正	世界成長率、3.9%に上方修正_IMF、18年見通し
トランプ米大統領が表明した鉄鋼やアルミ製品への新たな関税をめぐり、米政府高官は「どの国も例外ではない」と明言した。最大の標的の中国だけでなく、日本を始めとする同盟国も影響は避けられない見通しだ。欧州などは報復措置を準備しつつある。	月の地下に巨大な空洞	月の地下に50キロの空洞	月の地下に長さ50キロの空洞_探査機「かぐや」で判明
ノルウェーのノーベル委員会は6日、2017年のノーベル平和賞を、核兵器の非合法化と廃絶を目指す国際NGOで、今年の核兵器禁止条約成立に貢献した「核兵器廃絶国際キャンペーン」(ICAN)に与えると発表した。授賞式は12月10日にオスロである。ライスアンデセン委員長は授賞理由について、「核兵器の使用がもたらす破壊的な人道面での結末を人々に気づかせ、条約に基づく核兵器禁止の実現へ画期的な努力を重ねてきた」と説明。	月の地下に長大な空洞	月地下に長大な空洞_月探査	月の地下に長さ50キロの空洞_月探査機「かぐや」観測
血液中のビタミンDの濃度が低いとがんになるリスクが高くなることがわかった。国立がん研究センターのチームが8日、英医学誌ブリティッシュ・メディカル・ジャーナルに論文を発表した。ビタミンDはがん細胞の増殖を抑える働きがあるためとみられる。	トランプ関税に報復も	EU「トランプ関税」報復も	「トランプ関税」EUは報復も_日本、米政権に配慮要望
	トランプ氏、報復準備	トランプ米大統領、報復準備	「どの国も例外ではない」_トランプ新関税、米高官明言
	平和賞に国際NGO	ノーベル平和賞にICAN	国際NGO「核兵器廃絶国際キャンペーン」(ICAN)
	ノーベル平和賞決まる	ノーベル平和賞に核兵器廃絶	ノーベル平和賞に「核兵器廃絶国際キャンペーン」_ノル
	がんとビタミンD関係	ビタミンD低いとがんリスク	血中ビタミンD低いと、がんリスク上昇_肝臓がんで顕著
	ビタミンD濃度低いと	ビタミンD濃度低いとがんに	ビタミンD濃度低いとがんリスク_国立がん研究センター

図2: 日本語テストデータにおける、元記事に対する、各長さの人手の見出し文(上段)と提案手法で生成した見出し文(下段)。

**謝辞** 本研究成果は国立研究開発法人情報通信研究機構(NICT)の委託研究「多言語音声翻訳高度化のためのディープラーニング技術の研究開発」により得られたものです。また、朝日新聞社の人見雄太氏、田森秀明氏に実験についての助言をいただきました。

## 参考文献

- Angela Fan, David Grangier, and Michael Auli. “Controllable Abstractive Summarization”. In: *WMT*. 2018, pp. 45–54.
- Yuta Kikuchi et al. “Controlling Output Length in Neural Encoder-Decoders”. In: *EMNLP*. 2016, pp. 1328–1338.
- Taku Kudo. “Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates”. In: *ACL*. 2018, pp. 66–75.
- Haoran Li et al. “Ensure the Correctness of the Summary: Incorporate Entailment Knowledge into Abstractive Sentence Summarization”. In: *COLING*. 2018, pp. 1430–1441.
- Piji Li et al. “Deep Recurrent Generative Decoder for Abstractive Text Summarization”. In: *EMNLP*. 2017, pp. 2091–2100.
- Yizhu Liu, Zhiyi Luo, and Kenny Zhu. “Controlling Length in Abstractive Summarization Using a Convolutional Neural Network”. In: *EMNLP*. 2018, pp. 4110–4119.
- Thang Luong, Hieu Pham, and Christopher D. Manning. “Effective Approaches to Attention-based Neural Machine Translation”. In: *EMNLP*. 2015, pp. 1412–1421.
- Paul Over, Hoa Dang, and Donna Harman. “DUC in Context”. In: *Information Processing & Management* 43.6 (2007), pp. 1506–1520.
- Alexander M. Rush, Sumit Chopra, and Jason Weston. “A Neural Attention Model for Abstractive Sentence Summarization”. In: *EMNLP*. 2015, pp. 379–389.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. “Neural Machine Translation of Rare Words with Subword Units”. In: *ACL*. 2016, pp. 1715–1725.
- Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. “Sequence to Sequence Learning with Neural Networks”. In: *NIPS*. 2014, pp. 3104–3112.
- Jun Suzuki and Masaaki Nagata. “Cutting-off Redundant Repeating Generations for Neural Abstractive Summarization”. In: *EACL*. 2017, pp. 291–297.
- Sho Takase, Jun Suzuki, and Masaaki Nagata. “Direct Output Connection for a High-Rank Language Model”. In: *EMNLP*. 2018, pp. 4599–4609.
- Ashish Vaswani et al. “Attention is All you Need”. In: *NIPS*. 2017, pp. 5998–6008.
- Oriol Vinyals et al. “Show and Tell: A Neural Image Caption Generator”. In: *CVPR*. 2015, pp. 3156–3164.
- Qingyu Zhou et al. “Selective Encoding for Abstractive Sentence Summarization”. In: *ACL*. 2017, pp. 1095–1104.
- 人見 雄太 et al. “出力長制御を考慮した見出し生成モデルのための大規模コーパス”. In: 言語処理学会年次大会. 2019, (to appear).