

# 対象に固有な特徴を捉えたキャッチコピー生成

三沢 翔太郎 西埜 徹 谷口 友紀 三浦 康秀 佐藤 政寛 大熊 智子  
富士ゼロックス株式会社

{misawa.shotaro, nishino.toru, taniguchi.tomoki, yasuhide.miura,  
sato.masahiro, ohkuma.tomoko}@fujixerox.co.jp

## 1 はじめに

ニュース記事や通販サイトの商品ページには本文とともにタイトルやキャッチコピーが付与されている。このようなキャッチコピーはユーザーの関心が得られる表現にすることで、閲覧数や商品購買数の向上に繋がると考えられる。そのためキャッチコピーは単純な文書要約とは異なり、対象が持つ固有の強みや性質を正確に表すこと、全ての情報を網羅するわけではなく着目すべき点のみを含むこと、魅力的な表現であることなどの性質を持つことが望ましい。本研究では、このようなキャッチコピーの自動生成を行う。

ルールやテンプレートなしに柔軟なキャッチコピーを生成する方法には、機械翻訳や要約などで広く用いられているニューラルネットワークベースの Attention 機構付き Encoder Decoder モデル (EncDec) がある [1]。しかし、EncDec は頻出する単語やフレーズを生成しやすく、特徴のないキャッチコピーを生成してしまう傾向にある [2]。そのため、EncDec でキャッチコピーを生成すると、対象が持つ固有の性質がキャッチコピーに反映されにくくなってしまう。

そこで本研究では、より具体的に対象の特徴を表すキャッチコピーの生成を目指す。入力文の特徴を復元できるキャッチコピーは具体的な内容を含むと仮定し、生成した文から入力文を復元する機構を同時に学習する。ここで、キャッチコピーは入力文の内容を網羅する必要がないため、入力文をそのまま生成するのではなく圧縮した情報のみ復元する。これにより対象の入力文を特定できるような独自の文が生成されやすくなり、より特徴を表したキャッチコピーとなる。

本研究は以下の3点を目的とする。(1) 生成文から入力文を復元する機構を備えることで特徴性を考慮したキャッチコピーを生成するモデルを提案する。(2) 就職・転職のポータルサイト Wantedly<sup>1</sup>のデータを利用し、手法の有効性を確認する。(3) 適切性と特徴性という二つの人手評価の方法を提示し、提案手法の有効性を確認する。

## 2 関連研究

**キャッチコピー自動作成:** キャッチコピーの自動作成に関する研究は主に関連文書を収集しルールなどで書き換える方法が主流である [3]。これに対してルールや関連文書収集などなしに、大量の入力文とキャッチコピー

<sup>1</sup>Wantedly はウォンテッドリー株式会社の商標です。

のペアを用いて学習し、入力文からキャッチコピーを生成する方法が提案されている。これらは EncDec をベースとしており、出力文長を制御する機構 [4] や強化学習を組み合わせた方法などが提案されている [5]。しかし、これまでキャッチコピーの特徴性に着目した研究は取り組まれていない。

**ダイバーシティーの拡張:** EncDec で一般的な文が生成されるという問題に対して、出力のダイバーシティーを拡張する方法が存在する。近年注目されている方法は、Variational Auto Encoder を応用し、ランダムな要因を加えることで特定の表現ばかりを出力しないようにする方法である [2]。この方法は単純に表現の幅を広げることを目的としているが、本研究は入力文独自の情報を含ませることを目的としているためアプローチが異なる。この他に、入力文の様々な単語に着目しながら文を生成する方法が提案されており、Attention 機構を用いた Coverage 機構 [6] や入出力の対応関係を同時に学習し利用する方法などがある [7]。これらの方法は入力文の内容を網羅するという考えであるが、キャッチコピー生成においては必ずしもすべての情報を均等に含む必要はないため、本課題に対する効果は限定的であると考えられる。

**入力文を復元する機構:** 本研究と同様に入力文を復元することで性能向上を図る手法が提案されている。目的の文を生成するモデルに加えて Decoder の中間表現から元文を再現するモデルを同時に学習する方法 [8], [9] や、ビームサーチで複数の候補文を生成したのちに復元のしやすさも評価基準として生成文を選択する方法などがある [10]。これらは、機械翻訳など入力情報と出力情報がほぼ同じ情報量を持っている場合を対象としている。しかし、キャッチコピー生成は出力情報が入力情報よりも極端に情報量が少ない問題設定であり、これらの方法をそのまま適用して性能が改善するとは限らない。

## 3 入力文復元機構を備えたキャッチコピー生成モデル

### 3.1 概要

提案手法は Attention 機構付き Encoder Decoder モデルをベースとして、生成されたキャッチコピーから入力文に含まれている単語を復元する機構を備える。この復元機構は学習時にのみ利用し、復元機構で生じた誤差は生成モデルまで伝搬させる。これにより、生

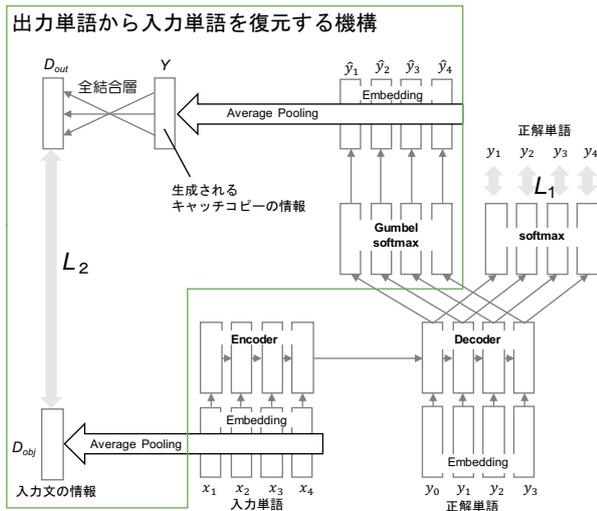


図 1: 提案手法の概要

成したキャッチコピーから入力文の復元を可能にするために、生成モデルは入力文に応じて出力を変化させ、対象の入力文に固有の内容を表すような単語を特に生成しやすくなる。図 1 に提案手法の概要を示す。

### 3.2 Attention 機構付き Encoder Decoder Model

EncDec は Encoder で入力文に対する情報を獲得し、その情報を元に Decoder は生成した単語を考慮しながら 1 単語ずつキャッチコピーを生成する。また、Attention 機構では Decoder で各単語を生成する際に、入力文の単語に重み付けしながら Encoder から獲得した中間表現を扱う。Encoder, Decoder と Attention は同時に学習され、生成する際はビームサーチを用いて事後確率の高い系列を獲得する。実験では Encoder と Decoder の内部構造には Bidirectional Gated Recurrent Unit (GRU) を用いる。

### 3.3 生成結果から入力単語を復元する機構

この機構は、Decoder から生成された単語から入力文に含まれる単語の情報を復元する。これは学習時に EncDec の性能を改善するためのみに利用し、以下の流れで処理を行う。

**生成されるキャッチコピーの情報獲得** 一般に Decoder において各単語を生成する際は、GRU からの出力値を元に全結合層と Softmax 層を用いることで各単語の生成確率を計算する。この機構においては Softmax 層を Gumbel Softmax 層 [11] に置換することで、確率値ではなく式 (1) に示すように一意の単語を獲得する。さらに獲得された単語の Embedding の平均値を式 (2) のように算出し、これを生成されるキャッチコピーの情報として扱う。

$$\hat{y}_t = \text{gumbel}(d_t) \quad (1)$$

$$Y = \sum_{t=1}^T \text{Emb}(\hat{y}_t) / T \quad (2)$$

$\text{gumbel}(\cdot)$ ,  $\text{Emb}(\cdot)$  はそれぞれ gumbel softmax 層, Embedding 層からの出力を、 $d_t$  は  $t$  単語目の Decoder からの出力を、 $T$  は生成文の単語数を表す。

**全結合層** 上記で獲得した生成されるキャッチコピーの情報を入力として全結合層  $f$  で式 (3) で示すように線形変換する。この層の導入により、入力文とキャッチコピーで使われる単語の傾向や情報量に違いがある場合でも両者の距離を近づけやすくなる。

$$D_{out} = f(Y) \quad (3)$$

**入力文の情報獲得** この機構でキャッチコピーから復元する対象である入力文に関する情報は、式 (4) で示すように入力文に含まれる単語の Embedding の平均値を用いる。

$$D_{obj} = \sum_{i=1}^I \text{Emb}(x_i) / I \quad (4)$$

$I$  は入力文の単語数を、 $x_i$  は  $i$  番目の入力単語を表す。

### 3.4 目的関数

提案手法では、Decoder からの生成された単語と正解の単語から計算されるクロスエントロピー損失に加えて、生成文から入力文を復元する機構において生じた誤りを元に算出した損失を導入する。後者のロス  $D_{out}$  と  $D_{obj}$  の距離で定義し、実験では正規化したコサイン距離を用いた。学習時には二つの損失を同時に学習するため、以下の式 (5) の損失を最適化する。

$$L = L_1 + \lambda \times L_2(D_{obj}, D_{out}) \quad (5)$$

$L_1$  は生成単語に関するクロスエントロピー損失で、 $L_2$  は二つのベクトルの距離から算出した損失であり、 $\lambda$  は二つの損失の学習度合いを制御するハイパーパラメータである。

## 4 実験

### 4.1 実験設定

**実験データ:** 就職・転職ポータルサイトである Wantedly に掲載されている情報を用いて、募集要項からその募集要項に付与されたキャッチコピーを生成するタスクで評価実験を行なった。このデータはヘッドラインとキャッチコピーの性質を併せ持つ。実験に用いたデータの例は図 2(a) に示す。また、データは同一企業が同一の集合に属さないようにランダムに分割し、その統計的特徴は表 1 に示す。

**実験条件:** Embedding の次元と GRU のユニットはともに 200 とし、形態素解析は mecab ipadic-neologd を利用し、Embedding はランダム初期化した。また、エポック数は最大 15、バッチサイズは 5、最適化は学習率 0.001 の Adam で行い、ビームサーチのサイズは 5 と

<p><b>【募集要項】</b>  Wantedlyは海外に進出しています。現在では、海外ユーザー向けに Rails アプリと iOS/Android アプリを展開しています。Wantedly 海外版の開発は、ただ多言語化するだけでなく、国ごとのマーケットの違い、ユーザーの課題、サービスのフェーズを理解して、海外にスケールさせていく仕事です。つくる過程では、現地で活動しているメンバーと密にコミュニケーションをしながら開発を進めていくことが重要となってきます。海外のユーザーにも使われるプロダクトをチーム開発してみたい人、海外進出することで Wantedly のサービスを育てていきたい人、ぜひ一緒に多くの人に使われるプロダクトをつくっていきましょう！</p> <p><b>【正解のキャッチコピー】</b>  Wantedly のサービスを海外にスケールさせたいエンジニア Wanted !</p>
--

(a) 募集要項とキャッチコピーの例

<p><b>【EncDec出力】</b>  海外インターン！0→1を創る  iOSエンジニアを募集！</p>	<p><b>【提案手法出力】</b>  海外向けアプリでグローバルに活躍  したいiOSアプリエンジニアを募集！</p>
--	--

(b) EncDec と提案手法が生成したキャッチコピー

図 2: データの例とモデルが生成したキャッチコピー

表 1: データ概要

	募集数	企業数	平均出力長	平均入力長
学習	58,461	5,736	15.59	347.14
開発	6,290	717	15.38	314.75
テスト	7,573	721	15.57	362.67

した。また、入力は募集要項の最初の 100 単語を用いた。提案手法の  $\lambda$  は  $\{0.25, 0.5, 0.75, 1.0, 2.0, 3.0, 4.0, 5.0\}$  から開発データに対する ROUGE-L が最良になる 4.0 を用いる。さらに比較のために、EncDec に Coverage 機構を付与したモデル (Coverage), 提案手法の復元機構で用いるキャッチコピーの情報として Embedding ではなく BoW の平均を利用したモデル (bow), さらに Gumbel Softmax を利用せず Softmax の出力を利用したモデル (softmax) を比較する。

## 4.2 評価指標

自動評価と人手評価を実施した。自動評価では ROUGE-1,L を計算し、10 回の実験の平均値を用いた。

提案手法では入力に対して適切なキャッチコピーを生成し、その上で対象のより具体的な特徴を捉えることを目指す。そこで人手評価は、適切性と特徴性という 2 つの独立した観点で、日本語学を学んだアノテーションの専門家が行った。

**適切性評価:** 適切性は募集要項とキャッチコピーを提示し、キャッチコピーが募集要項に合致しているかどうかを判断した。100 件に対してアノテーションを実施し、適切なキャッチコピーの割合を算出した。

**特徴性評価:** 適切性と評価を切り分けるために募集要項は提示せず、両方の手法で生成したキャッチコピーのみを提示し、どちらのキャッチコピーがより具体的な求人を表しているかを判断した。ただし、判断不可能も許容した。1000 件のキャッチコピーに対して評価を行い、より具体的であると評価されたキャッチコピーの割合を算出した。

表 2: 実験結果

	ROUGE-1	ROUGE-L
EncDec	20.15	16.77
Coverage	20.40	16.88
提案手法 (bow)	**21.06	**17.47
提案手法 (softmax)	**20.97	**17.44
提案手法	** <b>21.28</b>	** <b>17.75</b>

ただし、\*\*は EncDec と比較して 1% 有意であることを、太字は各指標で最も性能が高いものを表す。

表 3: 人手評価の結果

	適切性	特徴性
EncDec	0.18	0.46
提案手法	<b>0.31</b>	<b>0.50</b>

## 4.3 実験結果

表 2 に自動評価を用いた実験結果を示す。この結果から、EncDec と比較して、ROUGE-L が 0.98pt 向上しており提案手法が有効といえる。また、提案手法の全ての設定において EncDec よりも性能が高いため、復元することによる効果があることがわかる。さらに、各設定を比較すると Gumbel Softmax や Embedding を使う効果がそれぞれあることがわかる。対して、Coverage の性能向上幅は非常に限定的であるため、キャッチコピーにおいては網羅性が重要でないという仮説が正しかったと考えられる。これらの性能に関する傾向は ROUGE-L だけではなく ROUGE-1 もほぼ同様であることから、提案手法は文脈を考慮せずに特徴的な単語を生成することで復元を容易にしているわけではないと考えられる。

表 3 に人手評価による実験結果を示す。この結果から適切性と特徴性ともに提案手法で向上していることがわかる。特に適切性は大幅に向上しており、入力文を復元する機構を付与したことで、入力文と異なる内容の文を生成する確率が減少したと考えられる。また特徴性も向上していることから、入力文の内容に合致する文を生成することで、より具体的に見えるキャッチコピーが生成されることが確認できた。

## 5 考察

**適切性の例:** 図 2(a) にある募集要項から EncDec と提案手法で生成したキャッチコピーを図 2(b) に示す。どちらのキャッチコピーも「海外」と「アプリエンジニア」は述べられており、募集要項に即した内容が生成されている。しかし、EncDec は「海外インターン」や、新規事業を思わせる「0 → 1 を創る」という募集要項に合致しない表現が生成されている。対して、提案手法は「海外向けアプリ」と募集要項の最も重要な点を抑えることができているといえる。

**特徴性の例:** 表 4 に同一の募集要項に対して、EncDec と提案手法が生成した文を示す。1 と 2 の生成結果に関しては提案手法がより特徴的であり、3 番は EncDec が特徴的であると評価されたものである。提案手法の 1 と 2 のキャッチコピーは「エンタメ集団」や「自動車

表 4: EncDec と提案手法が生成したキャッチコピーに関する特徴性比較の例

	EncDec	提案手法
1	一緒に会社を創り上げていく仲間を募集	<社長直下>急成長のエンタメ集団を牽引する人事採用担当者を募集!
2	世界を舞台にイケてるものを生み出したいデザイナー募集	世界最大級の自動車業界にイノベーションを起こすデザイナー募集!
3	海外で暮らしをサポートするハウスマネージャー募集	フリーター歓迎! 国際色豊かなメンバーとして加わりませんか?

業界」など業界を限定する単語が入っており、より具体的な内容が記載されていることがわかる。このことより、提案手法では入力文を復元するために特徴的な単語を生成していることが考えられる。しかし、3のキャッチコピーは「フリーター歓迎」や「国際色豊か」などの特徴的な表現は含まれているものの、EncDecのキャッチコピーには職種を限定する「ハウスマネージャー」という表現が含まれるため EncDec がより特徴的であると判断された。これは、提案手法が職種や業種などの重要な語を区別して扱えないため生じたと考えられる。

**復元機構の性能:** 提案手法の復元機構を学習することで、入力文を推定できるような単語が生成されているかを確認した。ここでは、EncDec で生成した文から入力文を推定する性能を擬似的に評価するために、復元機構の損失を全結合層のみに伝搬させて EncDec には影響を与えない方法を用意した。なお、評価には生成文から算出した中間表現と入力文を構成する単語の平均 Embedding との間のコサイン類似度を用いた。その結果、提案手法は 0.996 で、全結合層のみ学習したモデルは 0.888 であった。このことから、復元機構を追加し学習に利用することで、入力文をより復元できる文が生成されていることが確認できた。

**生成文の特徴性:** 募集要項にバリエーションがあり、提案手法が特徴的なキャッチコピーを生成できると仮定すると、募集要項ごとに類似しないキャッチコピーを生成すると考えられる。そこで、EncDec と提案手法それぞれが生成したキャッチコピー集合に対して、式 (6) の  $most\_similar_i$  の平均を算出した。

$$most\_similar_i = \max(\sum_{j \in C \cap j \neq i} ROUGE-L_i^j) \quad (6)$$

$C$  はキャッチコピー集合を、 $ROUGE-L_i^j$  は  $i$  番目と  $j$  番目のデータに対するキャッチコピーを比較した際の ROUGE-L を表す。

その結果、EncDec は 75.27 であり提案手法は 71.77 であった。正解のキャッチコピー集合でこの値を算出すると 55.96 であるため改善の余地はあるが、提案手法の方が入力文に対して特徴的なキャッチコピーを生成できていることがわかる。

さらに提案手法は、正解のキャッチコピーが特徴的なデータに対して特に性能が高いと考えられる。そこで式 (6) を各テストデータの正解キャッチコピーに対して算出し、テストデータをその値で分類し、それぞれに対する提案手法と EncDec の ROUGE-L の差を比較した。この結果、 $most\_similar_i$  が 0.3 以下のテストデータでは 1.90pt, 0.4 以下では 1.35pt, 0.5 以下では 0.74pt 提案手法の性能が高かった。このことから提案手法は特徴性の高いテストデータに対して特

に性能が高く、求める効果が獲得できている。

**入力文とキャッチコピーの単語一致:** 提案手法が生成したキャッチコピーの方が入力文の内容をより表しているかを確認する。ここでは生成されたキャッチコピーの単語のうち、募集要項の中でモデルの入力に用いていない 100 単語目以降の単語と一致する割合を計算した。正解のキャッチコピーは 47.26% であるのに対して、従来手法は 42.96% で提案手法は 44.49% であり、提案手法の方が入力文の内容をより忠実に再現していることが確認できた。

## 6 おわりに

本研究ではキャッチコピー生成において、入力情報を生成した単語情報から推定する機構を追加し、生成モデルと同時に学習する方法を提案した。就職・転職ポータルサイトのデータを利用して、自動評価において従来の Attention 機構付き Encoder Decoder モデルよりも高い性能であることを確認した。さらに、人手評価を実施し、提案手法の方がより特徴的な表現が生成されることが確認された。今後は、入力文を復元する機構を備えた他の手法との比較を行う。さらに、復元すべき単語と復元しなくても良い単語を自動で切り分けられるように復元部分にも Attention を導入することで性能を向上させる。

## 参考文献

- [1] Dzmitry Bahdanau, Kyunghyun Cho and Yoshua Bengio. Neural Machine Translation by Jointly Learning to Align and Translate. arXiv preprint arXiv:1409.0473, 2014.
- [2] Tiancheng Zhao, Ran Zhao and Maxine Eskenazi. Learning Discourse-level Diversity for Neural Dialog Models using Conditional Variational Autoencoders. In *ACL*, 2017.
- [3] 山根宏彰, 荻原将文. Web 情報を用いたキャッチコピー自動生成システム. ファジシステムシンポジウム講演論文集 28, 2012.
- [4] 牧野拓哉, 岩倉友哉, 高村大也, 奥村学. Minimum Risk Training に基づく要約モデルの出力長制御. 言語処理学会第 24 回年次大会, 2018.
- [5] Yen-Chun Chen and Mohit Bansal. Fast Abstractive Summarization with Reinforce-Selected Sentence Rewriting. In *ACL*, 2018.
- [6] Abigail See, Peter J. Liu and Christopher D. Manning. Get To The Point: Summarization with Pointer-Generator Networks. In *ACL*, 2017.
- [7] 清野舜, 高瀬翔, 鈴木潤, 岡崎直観, 乾健太郎, 永田昌明. ニューラルヘッドライン生成における誤生成問題の改善. 言語処理学会第 24 回年次大会, 2018.
- [8] Zhaopeng Tu, Yang Liu, Lifeng Shang, Xiaohua Liu and Hang Li. Neural Machine Translation with Reconstruction. In *AAAI*, 2017.
- [9] Longyue Wang, Zhaopeng Tu, Andy Way and Qun Liu. Learning to Jointly Translate and Predict Dropped Pronouns with a Shared Reconstruction Mechanism. In *EMNLP*, 2018.
- [10] Isabelle Augenstein and Anders Sogaard. Multi-Task Learning of Keyphrase Boundary Classification. In *ACL*, 2017.
- [11] Eric Jang, Shixiang Gu and Ben Poole. Categorical Reparameterization with gumbel-softmax. arXiv preprint arXiv:1611.01144, 2016.