Hierarchical Transformer によるストーリー生成

渥美和大 [1]　狩野芳伸 [2]

静岡大学

[1]katsumi@kanolab.net [2]kano@inf.shizuoka.ac.jp

1　はじめに

本稿で言うストーリー生成は、ストーリー内の１文目、あるいは１文目を含む複数文を入力として、それに続く文を生成することが目的のタスクである。本タスクは、ストーリー内に出現する登場人物、ストーリー内で起こるイベントの連鎖に一貫性を保ちながら文を生成する必要があるため難しい。

近年ニューラルネットワークをストーリーの自動生成に応用する研究が行われている[1、2、3、4]。本研究でも同様に、ニューラルネットワークを用いてストーリー生成を行う。ストーリー生成の手法として、1 つ目の文を入力として与えて、後に続く文を再帰的に生成する手法を用いる。モデルはHierarchical Recurrent Encoder Decoder(HRED)[5、6]のRNNをTransformer[7]に置き換えたHierarchical Transformer(HT)を扱う。

また、本研究ではDziriら[8]が、対話生成の研究において、Latent Dirichlet Allocation(LDA)[9]によって得られたトピックのトピックワードを用いて生成文の多様性を持たせた研究に着目し、複数のトピックからAttentionを取る手法を提案する。加えて、Takaseら[10]が見出し文生成の研究において、単語列に出力長までの相対的な位置を与える研究に着目し、文単位の相対的な位置を与える手法を提案する。実験では提案手法でHTを拡張したモデルでストーリーを生成し、HTと比較してBLEU、Perplexity、Distinct[11]の結果が改善されたこと示す。

2　関連研究

ニューラルネットワークによるストーリー生成は、sequence-to-sequence(seq2seq)[12]を基にする手法が主流である[1、2]。Martinら[1]は、ストーリー内のイベント表現からイベント表現を生成するモデルとイベント表現から自然文を生成するモデルを用いてストーリーを生成する手法を提案した。Guptaら[2]は、ストーリー内に現れる重要単語を抽出し、ストーリーの最終文を生成する手法を提案した。Guan[3]らは、HREDをベースにして、常識的知識を組み込みながらストーリーの最終文を生成する手法を提案した。また、本研究と同様に、ストーリー全体を生成する手法がある。Dauphinら[4]は、ConvS2S[13]をベースに用いて、特定のキーワード文からストーリー全体を生成した。これらの手法は入力と出力が１対１であるのに対し、本研究では文を再帰的に複数生成するため、ストーリーの文数を制御しながらストーリーを生成することができる。

3　背景

3.1　Transformer

はじめに、本研究で扱うHierarchical Transformerのベースとなる Transformerについて説明する。Transformerは、Attention、Feed Forward Network(FFN)のみを使うモデルである。EncoderはSelf-Attention、FFNの 2 つ、Decoderは、Self-Attention、Source-Target Attention、FFNの３つのサブレイヤーで構成され、それぞれのレイヤーは複数層スタックされる。また、Self-Attention、Source-Target Attentionでは、複数のHeadでAttentionを取るMulti-Head Attention(MultiHead)[7]が用いられ、(1)で表す。なお、QはQuery、KはKey、VはValueを表す。

$$MultiHead(Q,K,V) \quad (1)$$

Encoder、Decoder共にそれぞれの内部表現がQ、K、Vの入力になるMultiHeadをSelf-Attentionと呼び、DecoderにおいてQがDecoderの内部表現、K、VがEncoderの最終出力になるMultiHeadをSource-Target Attentionと呼ぶ。N層目の内部表現を$E^{(n)}$、N層目のSelf-Attentionの出力を$A^{(n)}$とすると、Encoderのサブレイヤーは(2)、(3)で表す。

$$A^{(n-1)} = MultiHead\left(E^{(n-1)}, E^{(n-1)}, E^{(n-1)}\right) \quad (2)$$

$$E^{(n)} = FFN\left(A^{(n-1)}\right) \quad (3)$$

N層目の内部表現を$D^{(n)}$、N層目のSelf-Attentionの出力を$F^{(n)}$、N層目のSource-Target Attentionの出力を$G^{(n)}$とすると、Decoderのサブレイヤーは(4)、(5)、(6)で表す。

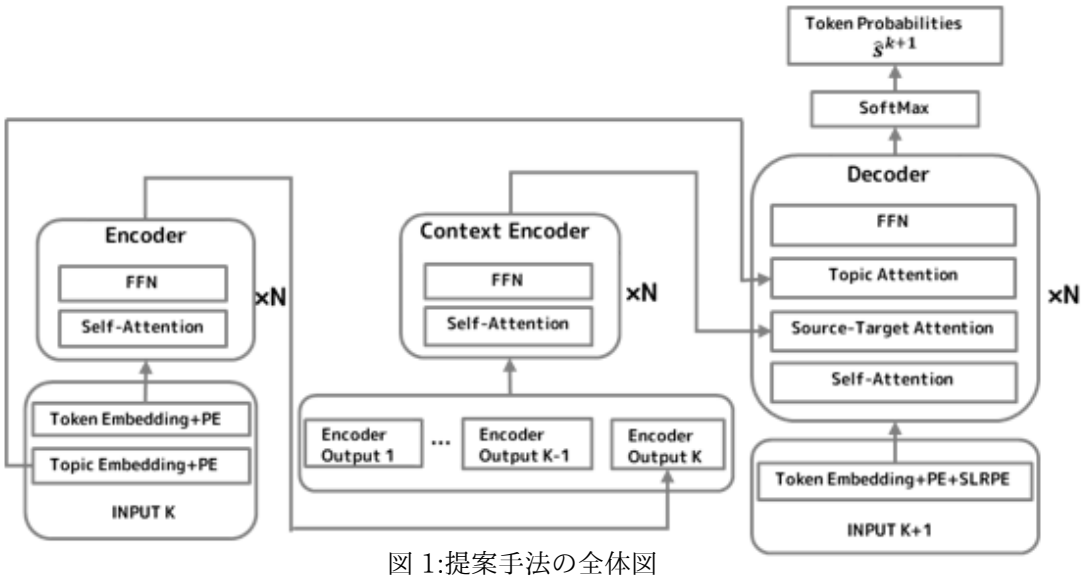

図 1:提案手法の全体図

$$F^{(n-1)} = MultiHead\left(D^{(n-1)}, D^{(n-1)}, D^{(n-1)}\right) \quad (4)$$

$$G^{(n-1)} = MultiHead\left(F^{(n-1)}, E^{(N)}, E^{(N)}\right) \quad (5)$$

$$D^{(n)} = FFN\left(G^{(n-1)}\right) \quad (6)$$

Transformerの入力は単語の埋め込み表現に加えて、単語の位置を表すPositional Encoding(PE)[7]が用いられる。PEは正弦波と余弦波によって単語の位置を表し、単語の埋め込み表現に加算される。単語の位置をpos、埋め込み表現の次元数をdとすると、PEは(7)、(8)で表す。

$$PE_{(pos,2i)} = \sin\left(pos/10000^{2i/d}\right) \quad (7)$$

$$PE_{(pos,2i+1)} = \cos\left(pos/10000^{2i/d}\right) \quad (8)$$

3.2 Hierarchical Transformer (HT)

HTはTransformerのEncoder、Decoderに加え文脈情報を扱うContext Encoderを追加する。Context Encoderは複数文を順番にEncoderに入力し、それぞれのEncoderの出力を加算したものを結合し、入力とする。HTは複数文の情報を扱うことで文脈情報を学習することができる。

Context EncoderのサブレイヤーはEncoderのサブレイヤーと同様のため説明は省く。DecoderではTransformerのサブレイヤーであるSource-Target AttentionのK、VをContext Encoderの出力に置き換える。

4 提案手法

本研究ではLDAから得られるトピックを用いたTopic Attention(4.2)、文の相対位置を表すSentence Level Relative Positional Encoding(4.3)を提案する。また、提案手法でHTを拡張し、ストーリーを再帰的に生成する。提案手法を含めたモデルの概要を図 1 に示す。

4.1 問題設定

K個の文からなる文章、I個の単語で構成されるK番目の文を(9)、(10)で表す。

$$S = s^{(1)}, \ldots\ s^{(k)}, \ldots s^{(K)} \quad (9)$$

$$s^{(k)} = s_1^{(k)}, \ldots, s_i^{(k)}, \ldots, s_I^{(k)} \quad (10)$$

本研究では、k番目までの文$S_{\leqq k}$を入力として$k+1$番目の文S_{k+1}を提案モデルによって生成することを目標とする。s_{k+1}を生成する確率を(11)で表す。

$$P(s^{k+1}|S_{\leqq k};\theta) = \Pi_{i=1}^{I} P\left(s_i^{k+1}|S_{\leqq k}, s_{<i}^{k+1};\theta\right) \quad (11)$$

4.2 Topic Attention (TA)

入力文のトピックを考慮するために、Decoderのサブレイヤーにk番目の文のトピックからのAttentionを取るTopic Attention(TA)を提案する。TAを追加したDecoderのサブレイヤーは(12)、(13)、(14)、(15)で表す。なお、$C^{(N)}$はContext Encoderの出力、Tはk番目の文をLDAによってトピック推定し、上位複数のトピックを単語同様に、埋め込み表現にPEを加算したものを表す。

$$F^{(n-1)} = MultiHead\left(D^{(n-1)}, D^{(n-1)}, D^{(n-1)}\right) \quad (12)$$

$$G^{(n-1)} = MultiHead\left(F^{(n-1)}, C^{(N)}, C^{(N)}\right) \quad (13)$$

$$R^{(n-1)} = MultiHead\left(G^{(n-1)}, T, T\right) \quad (14)$$

$$D^{(n)} = FFN\left(R^{(n-1)}\right) \quad (15)$$

	Perplexity	BLEU	Distinct-1	Distinct-2
HT	42.76	5.197	0.483	2.090
HT+TA	42.07	5.224	**0.701**	3.177
HT+SLRPE	43.70	5.162	0.451	1.762
HT+TA+SLRPE	**40.05**	**5.319**	0.662	**3.677**

表 1:評価結果

4.3 Sentence Level Relative Positional Encoding (SLRPE)

ストーリーの流れを表すために、生成中の文が相対的にストーリーのどの位置を生成しているかを表すSentence Level Relative Positional Encoding(SLRPE)を提案する。SLRPEはTakaseら[10]が提案したLength Ratio Positional Encoding(LRPE)で、単語に文章長に依存した単語の相対位置を与えていたのを、文数に依存した文の相対位置を与えるものである。SLRPEを(16)、(17)で示す。なお、iは次元の番号、posは現在の出力文が何文目かを表し、lenは生成する文数を表す。

$$SLRPE_{(pos,len,2i)} = \sin\left(pos/len^{2i/d}\right) \quad (16)$$

$$SLRPE_{(pos,len,2i)} = \cos\left(pos/len^{2i/d}\right) \quad (17)$$

SLRPEはDecoderでの単語埋め込み表現にPEと共に加算される。

5 実験

5.1 データセット

本研究では、ROCStories Corpus[14]をストーリー生成に用いる。このデータセットは、日常の出来事の因果関係と流れを表す連続した文から構成される。各ストーリーは独立しており、最終文でストーリーの結末を表す。各文の平均単語数は 9.8 個で、ユニークな単語は 42253 語存在する。全体のデータ数 101903 件のうち、93903 件を訓練データ、4000 件をテストデータ、4000 件を検証データとした。

5.2 比較モデル

比較モデルとして、下記の 4 つのモデルで学習、評価を行う。

HT:HREDのRNNをTransformerに置き換えたモデル
HT+TA:TA(4.2)でHTを拡張したモデル
HT+SLRPE:SLRPE(4.3)でHTを拡張したモデル
HT+TA+SLRPE:TA(4.2)、SLRPE(4.3)でHTを拡張したモデル

5.3 実験設定

全てのモデルでEncoder、Context Encoder、Decoderはレイヤーを 6 層スタックし、MultiHeadのHead数は 8、隠れ層の次元は 512 とした。学習時は語彙制限を行わず、Label Smoothingを行い、OptimizerとしてAdamを用いた。Adamの学習率は 0.001、その他のパラメータは推奨値で設定した。TAを用いる際はデータセット内の名詞、形容詞、動詞から、出現頻度が高い単語と低い単語を除いてLDAを学習し、LDAのトピック数は 50、扱うトピックは上位 5 つとした。また、SLRPEを用いる際は生成する文は 5 つとした。生成時にはBeam Search(Beam size:8)を行った。

5.4 評価

評価尺度として、テストデータに対するPerplexity、BLEU、Distinctを用いて提案手法の有用性を評価する。

5.5 評価結果

評価結果を表 1、各モデルの生成結果を表 2 に示す。HT+TAは、BLEU、Perplexityにおいて、HTに比べて上回る結果となった。生成結果を比較するとHTは入力文と生成文の間の関係があまり考慮されていないが、HT+TAは”pet”という単語に対して適切に続く文を生成している。この結果からTAによって入力文のトピックが考慮され、適切な文が生成されていることが分かる。

HT+SLRPEは、BLEU、Perplexity共に下回る結果となった。生成結果を比較すると、生成文は異なるがSLRPEを追加したことによる特徴の変化は確認できなかった。

HT+TA+SLRPE は BLEU 、 Perplexity において、HT+TA、HT+SLRPEに比べて上回る結果となった。HT+TAと生成結果を比較すると、HT+TA+SLRPEはHT+TA同様に”pet”という単語に対して適切に続く文を生成している。また、HT+TAでは、最終文の 4 文目で次文が存在するような結末として不適切な文を生成しているのに対し、HT+TA+SLRPEでは結末として適切な文を生成している。この結果からSLRPEによって文の相対

Input	Dan wanted a pet for Christmas.
HT	He decided to try out for the team. One day, he decided to buy a lottery ticket. He was very happy with his choice. He was so happy that he almost cried!
HT+TA	One day, He went to the animal shelter. He saw a lot of animals. He had a lot of fun with his kitten. He went to the store and bought all the ingredients he needed.
HT+SLRPE	One day, he decided to take a nap. He went to the store and bought all of the ingredients. They had a lot of fun. I was very happy with my choice.
HT+TA+SLRPE	One day, he went to the pet store. He decided to keep the puppy as a pet. He took him home and put him in bed. He had a great time.

表 2:生成結果

位置が与えられ、生成文の位置情報を考慮して文を生成できていることが分かる。

今回実験で扱った全てのモデルで、生成される単語の多様性が少ない問題があった。Distinctで評価したところ、Distinct-1 ではHT+TA、Distinct-2 はHT+TA+SLRPEが最も良い結果になった。この結果からTAが生成される単語の多用性を生成するのに貢献していることが分かる。

6 おわりに

本研究では、入力文のトピックからのAttentionを取るTAと文の相対位置を表すSLRPEを提案し、HTに拡張してストーリーを生成した。実験では、TA、TA+SLRPEにおいてBLEU、Perplexity、Distinctでの結果が改善された。今後は定性評価として、生成されたストーリーの人手による評価を行いたい。

参考文献

[1] Prithviraj Ammanabrolu, Ethan Tien, Wesley Cheung, Zhaochen Luo, William Ma, Lara Martin, and Mark Riedl, 2019. Guided neural language generation for automated storytelling. In Proceedings of the Second Workshop on Storytelling, pages 46–55.

[2] Prakhar Gupta, Vinayshekhar Bannihatti Kumar. Mukul Bhutani. and Alan W. Black, 2019. Writerforcing: Generating more interesting story endings. In Proceedings of the Second Workshop on Storytelling. pages 117-126.

[3] Jian Guan Yansen Wang and Minlie Huang, 2018. Story Ending ge-neration with incremental encoding and commonsense knowledge. arXiv preprint arXiv:1808.10113.

[4] Angela Fan Mike Lewis and Yann Dauphin, 2018. Hierarchical neural story generation. In ACL, pages 889-898.

[5] lulian V. Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau, 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In AAAI, pages 3776-3784.

[6] Alessandro Sordoni, Yoshua Bengio, Hossein Vahabi,Christina Lioma, Jakob G, and Jian-Yun Nie, 2015. A hierarchical recurrent encoder-decoder for generative context-aware query suggestion. In CIKM, pages 553-562.

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Lion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, 2017. Attention is all you need. In NIPS, pages 5998-6008.

[8] Nouha Dziri, Ehsan Kamallo, Kory W. Mathewson, K. and Osmar Zaiane, 2018. Augmenting neural response generation with context aware topical attention. arXiv preprint arXiv:1811.01063.

[9] David M. Blei, Andrew Y. Ng, and Michael. I. Jordan, 2003. Latent dirichlet allocation. In Journal of Machine Learning Research, pages 993–1022.

[10] Sho Takase, Naoaki Okazaki, 2019. Positional encoding to control output sequence length. arXiv preprint arXiv:1904.0418.

[11] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan, 2016. A diversity-promoting objective function for neural conversation models. In Proceedings of NAACL-HLT, pages 110– 119.

[12] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le, 2014. Sequence to sequence learning with neural networks. In NIPS, pages 2440-2448.

[13] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, Yann N. Dauphin, 2017. Convolutional sequence to sequence learning. arXiv preprint arXiv:1705.03122.

[14] Nasrin Mostafazadeh, Michael Roth, Annie Louis, Nathanael Chambers, and James Allen, 2017. LSDSem 2017 shared task: The story cloze test. In Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics, pages 46–51.