イベント構造に基づく物語生成モデルの改良に向けて

古川　好　鶴岡 慶雅
東京大学大学院 情報理工学系研究科
{furukawa, tsuruoka}@logos.t.u-tokyo.ac.jp

1　はじめに

ニューラルネットワーク技術の発達や、容易に入手することができる大規模なテキストデータの増加などに後押しされ、自然言語処理の分野は目覚ましい勢いで発展している。その中でも文章生成の分野は、従来の注意機構を発展させた Transformer [1] や、更にその Transformer をベースとして作られた言語モデルである GPT-2 [2] の考案など、著しい発展を遂げている。

文章生成分野の中でも、小説や物語のようなものの生成を目標とするストーリー生成 (Story Generation) と呼ばれる分野が存在しており、近年は翻訳・要約などのタスクで使われるようなモデルを利用し、小説や物語を生成しようとする試みが盛んに行われるようになっている。

しかし従来のモデルには、一般的な文章、つまり文法的、意味的には間違いはないが人間が読んでいて面白いと感じることはないような文章を生成しやすい、長期の依存関係を捉えにくく一貫性を持った文章を生成することが難しい、文生成時に一トークンずつ生成を行うため長文の生成に時間がかかるなどの課題が存在している。そのため小説や物語のように、一貫したテーマを持ち、意味的に整合性が取れており、実際に人間が読んで新鮮味や面白さを感じることができるような文章を生成することは、依然として困難な課題となっている。

本研究では、これまでのストーリー生成研究及びその他の文生成タスクにおける研究を参考とし、簡潔なモデルを用いて整合性･面白さ･一貫性を確保した文章を生成することを目指す。そのために、Matrin ら [3] によるイベント構造を用いたストーリーの階層的な生成を目指した研究を下地とし、より整合性・面白さ・一貫性を確保した長めの文章を生成できるような拡張を行い、既存の Transformer をベースとした seq2seq モデルによる生成文との比較を行った。

2　関連研究

コンピュータによりストーリーを生成しようとする試みは古くから行われており、例えば 1970 年代には Meehan ら [4] がルールベースでストーリーを生成する TALE-SPIN というシステムを開発しており、2000 年代には McIntyle ら [5] らによって、大量のデータから関連性のグラフを作り、それを用いてストーリーを生成するという研究も行われている。近年ではニューラルネットワークを用いた文章生成技術の発達により、ニューラルネットワークを利用したストーリー生成の研究も盛んに行われるようになってきている。

多くの機械学習タスクと同じく、ニューラルネットワークを利用したストーリー生成も End-to-End に行われることが多いが、近年急激に盛んになった研究分野であるためか、入力や出力が研究によって多様であるという特徴がある。例えば、 Fan ら [6] による研究では、入力にはタイトルを出力するように訓練された言語モデルの出力を用いて、そのタイトルと関係のある長文ストーリーを丸ごと生成することを試みている。その一方で、 Guan ら [7] や Ippolito ら [8] は、入力としてストーリーの一部を用い、その残りの部分を出力することを試みている。

ストーリー生成において他の文章生成系タスクと比べて特に重要視されるのが、生成された文章がもっともらしい (plausible) か、面白い (interesting) か、一貫性がある (coherent) か、というものである。もっともらしさというのは、その文章が意味的、あるいは文法的に整合性が取れているかということを表している。文法自体は正しかったとしても、何らかの理由がない限り、例えば「夜に太陽が上っている」などの描写があったとすればそれは意味的には整合性が取れているとは言えないだろう。面白さというのは、その文章を人間が読んで面白いと思うかどうかということを表している。翻訳や要約などの文章生成では、出力文は正解文の内容とどれだけ類似しているかが評価され、実

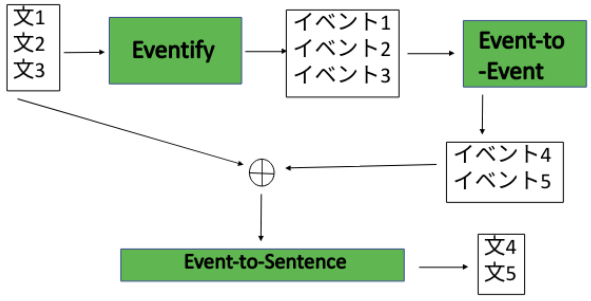

図 1: 提案モデル

際に人が読んで面白さを感じるかというものは評価されない。しかしストーリー生成では、最終目標は人間に読んでもらうことであるので、面白さというものが非常に重要となる。一貫性というのは、文章がある主題から逸れずに書かれているかどうかということを表している。もっともらしさや面白さというものがあっても、ある程度同じ軸に沿って書かれた文でなければストーリーが破綻していると考えられてしまう。

ストーリー生成研究全般に通じる特徴であるが、既存のモデルにはストーリーを段階的に生成するものが多い。一つの seq2seq モデルを利用して入力文から直接ストーリーを出力しようとするのではなく、一度入力文を何らかの中間表現に変換してから、それを用いてストーリーを出力する。例えば、Martin ら [3] は入力文を主語 (s)、述語 (v)、目的語 (o)、補助情報 (m) の 4 情報からなる抽象的な event に変換し、seq2seq を用いてその続きとなる event に変換。それを別の seq2seq を用いて文章に戻すことで続きの文章を生成することに成功している。

3 提案手法

本研究における提案モデルを図 1 に記す。

この提案モデルでは、三つの文からなる文章を入力とし、訓練時にはその続きとなる二つの文からなる文章を出力する。 Martin ら [3] の研究のように、従来の研究においては、ストーリーの続きを出力させるような場合は入力、出力ともに一文である場合がほとんどであるが、本研究においては複数文を出力することを目指す。 複数文の出力はより応用的なタスクであると考えられる。また、従来のモデルと同じように階層的な生成を行うことで、一貫性を保ちつつ生成されたストーリーのクオリティを向上させることを目指している。

以下では、提案モデルによる文章の生成過程について説明する。まず、入力文となる文 1, 2, 3 から (s, v, o, m) からなる イベント 1, 2, 3 を生成する。s, v, o, m は先述したように主語、述語、目的語、補助情報を指し、それぞれベクトルで表現される。これを行うのが図 1 における Eventify モジュールである。Eventify モジュールは stanfordNLP によるパースとルールベースによる変換プログラムから構成されている。

こうして作成されたイベント 1, 2, 3 を入力として、その続きとなるイベント 4, 5 を出力するのが Event-to-Event モジュールである。このモジュールは Transformer によって構成されているが、デコーダにおけるトークンのサンプリングには出力文の多様性を増すために、一般に使われる Beam Search ではなく Top-p Sampling [9] を採用している。

こうして得られたイベント 4, 5 と文 1, 2, 3 を結合したものを入力とし、所望の文である文 4, 5 を出力するのが Event-to-Sentence モジュールである。このモジュールは LSTM ベースの Pointer-Generetor Networks [10] をベースとした seq2seq により構成されている。 LSTM ベースの seq2seq は、 CNN ベースの seq2seq や Self-Attention ベースの Transformer と比べると計算時間などの点で不利ではあるが、長期の依存関係を捉えやすいとされており [11]、ストーリー生成のような場合では有利に働くと考えられる。また、Pointer-Generator Networks を使うことで、入力文から一部をコピーし、それによって生成されるストーリーの一貫性を確保することを目標としている。この Pointer-Generator Networks もデコーダ部分では Beam Search ではなく Top-p Sampling を使用しており、生成分の多様性を高めようとしている。

4 実験

図 1 で表される提案モデルと、通常の Transformer モデルを用いて文の生成を行った。98161 対のデータからなる ROCStories データセット [12] を使用した。これらの内 8 割の 78529 対を訓練データ、1 割の 9816 対をバリデーションデータ、1 割の 9816 対をテストデータとし、 80 エポック訓練を行った。

4.1 結果

出力結果の例として、得られた出力文の一例を表 1 に示す。

表 1: 出力結果例

入力文
Jeff invited his friends over to play board games on Saturday night. They arrived at his house early that evening. The six of them sat around a big table.
出力文（Transformer）
So we do friends. Man took around loved me.
出力文（提案モデル）
Jeff actually beat kids. They Watched Jeff lucky guy.
参照文
They took turns deciding which game to play. They spent six hours playing different board games.

4.2 考察

ストーリー生成タスクにおいては出力文は、入力文に対しての整合性や一貫性、面白さを保ってさえいれば正しい出力として許容される。そのため、表 1 における提案モデルの出力文は参照文とは大きく異なっているが、それを理由として不適切な出力とすることはできないと考えられる。5 節で後述するが、翻訳などのタスクでよく使われる BLEU スコアは、定性的には参照文とどれだけ n-gram ベースで一致しているかを測っているため、ストーリー生成において BLEU などの評価基準を用いることは不適切であると考えられる。

生成された文章の全体的な傾向として、ランダムサンプリングの一種である Top-p Sampling を用いているため、生成された文章が文法的に誤っている場合がよく見られるという問題点があった。表 1 の出力文 (Transformer) などもその一例である。この傾向は特に生成された文が長いほど顕著に見られ、ストーリー生成タスクにおいても必ずしも Beam Search よりランダムサンプリング手法が有利になるわけではないことを表していると考えられる。

5 今後の課題

今後の課題としては、以下の三点が挙げられる。

一つは、デコードにおいて確率分布から一つの単語を決定する際の方法を変更するというものである。従来のモデルでよく使われる beam search では、入力文 $\boldsymbol{x}$ に対し出力文 $\boldsymbol{y}$ を、式 1 の近似解となるように決定する。

$$\boldsymbol{y} = \arg\max_{\boldsymbol{y'}} P(\boldsymbol{y'}|\boldsymbol{x}) \tag{1}$$

しかし、Holtzman [9] らによると、実際に人間の書く文章に言語モデルの確率を割り当てて調査した場合、Beam Search により生成された文章のように確率が高いわけではなく、低めの確率が割り当てられることが多い。そのため、真に人間の書いた文章を再現するためには他のデコード方法を使わなければならない。本研究で用いた Top-p Sampling はこの問題を解決するために考案された手法であるが、人間の書いた文章を再現するには不十分であると考えられる。

そのため、新たなデコード手法を提案する必要がある。考えられる手法としては、出力された文章が意味的・構造的に正当性があると判断される範囲で最も確率の低いトークンを選ぶ、などというものがある。このような考えに基づくデコード手法は著者の知る限り提案されておらず、今後はこの手法について模索していきたいと考えている。

二点目は、生成文の評価を正しく下すというものである。文生成タスクでしばしば使われる BLEU や ROUGE のような機械評価軸は、 n-gram をベースとしており生成分と正解文が一致しているほど高スコアであると判断するというものである。ストーリー生成のように明確な正解がなく、整合性・面白さ・一貫性さえ保たれていればどのような文章が生成されても問題がないというようなタスクにおいては、このような評価基準を使うことは適切でないと考えられる。そのため、人手評価や新たな機械評価を用いてモデルを評価しなければならない。HUSE [13] のように、人手評価と機械評価を融合させ、かつ生成された文章が訓練時に利用した文からの盗用ではなく、オリジナリティのある文章を出力できているかというところまで含めて判断するような評価手法を用いて判断することが有効であると考えられる。

もう一つは、依然として扱う文章が短いというものである。本研究では入力文として三つの文からなる文章を使い、それが二つの文からなる文章を出力することを目指したが、ストーリーと聞いて我々が想像するような文章はより長いものである。Fan ら [14] の研究のように、長めのストーリー全体を生成することを目指した研究や、WritingPrompts データセット [6] といった長文生成に適したデータセットも存在している。

これらを利用して本研究のような入力文章の続きを出力することを目指した研究は存在しておらず、より長文の含まれるデータセットを利用した場合の方が、提案手法である Pointer-Generato Networks の使用などが有効に働くことが多いのではないかと考えられる。

6 結論

本研究では、ニューラルネットワークを用いたストーリーの生成を試みた。ストーリーの生成において重要となるのは、生成された文章自体の面白さやデコーダにおけるサンプリングの方法、評価基準など、一般的な文生成タスクではあまり重要視されないような箇所であることが多い。今後はこういった部分についての手法を考案しつつ、生成される文章の傾向を制御できるような方向に研究を進めていきたいと考えている。

参考文献

[1] Ashish Vaswani, et al. Attention is all you need. In *NIPS*. 2017.

[2] Alec Radford, et al. Language models are unsupervised multitask learners. 2018.

[3] Lara J. Martin, et al. Event representations for automated story generation with deep neural nets. In *AAAI*, 2017.

[4] James R. Meehan. Tale-spin, an interactive program that writes stories. In *IJCAI*, 1977.

[5] Neil McIntyre and Mirella Lapata. Learning to tell tales: A data-driven approach to story generation. In *ACL*, August 2009.

[6] Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. In *ACL*, July 2018.

[7] Jian Guan, Yansen Wang, and Minlie Huang. Story ending generation with incremental encoding and commonsense knowledge. In *AAAI*, 2019.

[8] Daphne Ippolito, et al. Unsupervised hierarchical story infilling. In *Proceedings of the First Workshop on Narrative Understanding*, June 2019.

[9] Ari Holtzman, et al. The curious case of neural text degeneration. *ArXiv*, Vol. abs/1904.09751, , 2019.

[10] Abigail See, Peter J. Liu, and Christopher D. Manning. Get to the point: Summarization with pointer-generator networks. In *ACL*, pp. 1073–1083, July 2017.

[11] Ke Tran, et al. The importance of being recurrent for modeling hierarchical structure. In *EMNLP*, October-November 2018.

[12] Nasrin Mostafazadeh, et al. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *NAACL: Human Language Technologies*, pp. 839–849, San Diego, California, June 2016.

[13] Tatsunori Hashimoto, Hugh Zhang, and Percy Liang. Unifying human and statistical evaluation for natural language generation. pp. 1689–1701, 01 2019.

[14] Angela Fan, Mike Lewis, and Yann Dauphin. Strategies for structuring story generation. In *ACL*, July 2019.