CopyNet with SeqGAN による自動校閲システムの構築

永井 涼雅† 前田 亮‡

†‡立命館大学情報理工学部 〒525-8577 滋賀県草津市野路東 1-1-1

E-mail: †is0367fs@ed.ritsumei.ac.jp, ‡amaeda@is.ritsumei.ac.jp

あらまし 新聞記事は，記者によって原稿が作成された後に，校閲などの作業を通して発行される．校閲作業には，誤字脱字や名詞・動詞などの変換ミスの修正，助詞の誤用の修正などが含まれる．現状では，これらの作業は人手により行われているが，校閲作業には自動化できる部分が多いと考えられる．

新聞記事の校閲では，校閲前後での文章の差異が比較的少ない．そのため，本論文で提案する深層学習による自動校閲では，入力文からコピーすることを学習できる CopyNet モデルを用いる．加えて，精度の向上に寄与できると考え，画像処理分野で注目されている深層学習モデルの一つである GAN を自然言語処理に適用した SeqGAN を用いる．また，校閲前の新聞記事の入手が困難であることから，校閲ルールの逆適用により疑似的に校閲前記事を生成する．評価実験の結果，提案手法が従来手法よりも高い精度が得られた．

キーワード 深層学習，自然言語処理

1. はじめに

新聞記事は，記者によって原稿が作成された後に，校閲などの作業を通して発行される．新聞記事は世の中の出来事を正しく世間に伝えるためにある．その中で，校閲は，新聞記事において，その内容を正確かつ明瞭に修正する役割を持つ．もし校閲作業が杜撰であれば，誤った情報を世間に流布し，最悪の場合，名誉棄損など法に触れてしまう可能性もある．校閲作業には，誤字脱字や名詞・動詞などの変換ミスの修正，助詞の誤用の修正などが含まれる．現状では，これらの作業は人手で行われているが，校閲作業には自動化できる部分が多いと考えられる．

校閲作業の自動化には，深層学習の適用などが考えられる．しかし，自然言語に深層学習を適用する研究は，機械翻訳に比べて自動校閲では少ない．校閲作業は，機械翻訳と異なり，入力言語と出力言語が同じであることが特徴である．そのため，機械翻訳で用いられる通常の seq2seq（sequence to sequence）ではなく文章校閲分野における手法が有効となってくる．また，校閲の特徴として，校閲前後の記事の差異が比較的少ないことがあげられる．このことから，校閲前の記事の内容を校閲後の記事にコピーするという処理が精度向上に効果的であると考えられる．本研究では，入力文からコピーすることを学習する深層学習モデルと，敵対的な識別器の適用により自動校閲の精度向上を目指す．

2. 関連研究

機械翻訳や文法校正など様々な分野で深層学習が適用されている．RNN（Recurrent Neural Network）は深層学習モデルの一つであり，時系列処理を目的としたモデルである．RNN は隠れ層が次の隠れ層に繋がっており，ある時点での状態を次の状態に渡すことができる．しかし，RNN では長期的な依存関係を記憶することが困難である．そこで RNN に三つのゲート（入力ゲート，出力ゲート，忘却ゲート）を追加し，長期的な依存関係にも対応した LSTM（Long Short-Term Memory）が研究された．LSTM は，RNN と比較して精度に優れているが，ゲートを三つ追加していることで処理時間に劣る．Cho[1]らは LSTM のゲートを統合する GRU（Gated Recurrent Unit）を開発した．RNN などの深層学習モデルを利用して文章を生成する手法に Sutskever[2]らの seq2seq がある．Seq2seq は入力データを一つの潜在ベクトルに変換する Encoder と，潜在ベクトルを Encoder から受け取り，各時系列で最適な単語を予想し出力する Decoder に分けられる．Seq2seq では文章を受け取り，文章を出力することから，翻訳タスクによく利用されている．翻訳タスクにおいて，翻訳元と翻訳後の単語の対応関係は重要である．Seq2seq では Encoder が入力文の情報を特定のサイズのベクトルに変換するため，系列長が長くればなるほど元の文章の情報が損なわれるという欠点があった．Bahdanau[3]らは Decoder が局所的にフォーカスするモデルである Attention を開発した．通常，文章校正，校閲では入力文章と出力文章に大きな差異がない．よって，入力文章から単語をコピーして，そのまま出力することが精度に寄与する．Gu[4]らの CopyNet は，そのアイデアを基にしたモデルである．本研究ではこのモデルを用いて校閲文章の出力を行う．

乾[5]らは，深層学習を用いて文章生成タスクと校閲

編集予測タスクを同時に学習させることで，校閲自動化タスクの精度を向上させた．しかし，乾らの研究と同じように，校閲前の新聞記事を入手することは困難である．また乾らのモデルは校閲前後で文章の差異が比較的少ないことに注目していない．

深層学習を用いた自動生成は，画像処理においても盛んである．自動生成の精度を向上させる手法としてGoodfellow[6]らのGAN（Generative Adversarial Networks）がある．GANは2つのモデルを競わせることにより，精度を向上させる．GANは生成器と識別器に分けられる．生成器は乱数の入力から画像を生成する．一方で，識別器は生成器が生成した画像と実際の画像を識別できるように学習する．この2つを交互に学習させることにより，生成器はより本物に近いデータを生成するようになる．しかし，GANの仕組みをそのまま自然言語処理の分野に転用することができない．GANで用いられるデータは連続的であるが，自然言語処理では入力値は辞書ID（離散値）であるためである．自然言語処理にGANを適用した例として，Yu[7]らのseqGANがある．識別器に強化学習の手法を取り入れたことで，入力が文章であったとしてもGANが適用可能となった．しかし，この手法は学習の精度を上げる代わりに学習に多大な時間を要する．そこでseqGANの中でも，時間を最も要するロールアウトの操作を削除し，単一の報酬のみを逆伝播させることで時間を短縮している．白井ら[8]により，この手法が機械翻訳タスクにおいて精度の向上に寄与できると報告されている．

3. 提案手法

本研究における提案手法はデータ生成部と深層学習部の2つの部分に分けられる．データ生成部では，実際の新聞記事のデータに対して置換，削除，挿入の操作を確率的に行うことで擬似的に誤り文書を生成する．深層学習部では，GRUをベースにした双方向seq2seq，そしてCopyNetを敵対的な識別器の有無も交えて実装する．

3.1 データ生成部

データ生成部は全てのデータセットに対して適用する．データ生成部は置換，削除，挿入からなる．データ生成の際は形態素解析器であるmecab-ipadic-neologd[9]を用いて単語単位に分割する．名詞に対して5%の確率でそれと類似した単語と置換を行う．類似判定は事前に学習したword2vecにより行う．word2vecは入力に使う新聞記事コーパスを用いて学習させる．例えば，word2vecは「天下人」に対して「信長」と置換を行う．また同様に名詞に対して5%の確率で同音異義語との入れ替えを行う．同音異義語は同音異義語辞書[10]から検索し，検索に一致したものからランダムに置換を行う．また，全ての単語に対して5%の確率で，削除処理を行う．全ての助詞に対して5%の確率で，その直後にランダムな助詞を挿入する．

3.2 深層学習部

提案モデルは生成器と識別器に分かれる．生成器は，正解データと擬似校閲前データから学習し，校閲後データを出力する．識別器は，生成器が出力したデータと正解データを学習し識別する．生成器と識別器は敵対的であり，生成器は識別器を騙すためにより実際の文章に近いものを出力し，識別器はそれを見分けようとする（図1）．

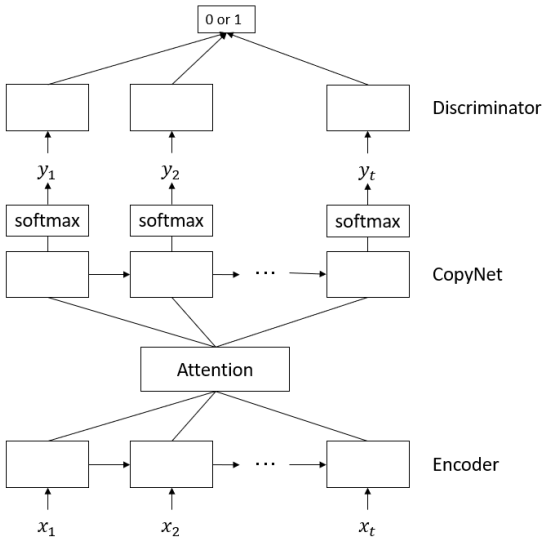

図 1 提案モデル

ここでx_tは $t \in \{1, \ldots, n\}$ における入力データであり，同様にy_tは$t \in \{1, \ldots, n\}$ におけるDecoderの出力である．

3.2.1 GRUからなるEncoder-Decoderを用いた文章生成

本研究では，校閲後文章の生成にはGRUからなるEncoder-Decoderモデルを使用する．同様にEncoder部分には双方向GRUモデルを使用する．GRUは各タイムステップ$s \in \{1, \ldots, m\}$でタイムステップ$s-1$もしくは$s+1$の状態と，入力単語であるx_sで式（1），式（2）のように計算を行う．

$$\left[\overrightarrow{h_s}\right] = \overrightarrow{GRU}\left(x_s, \overrightarrow{h_{s-1}}\right) \qquad (1)$$

$$\left[\overleftarrow{h_s}\right] = \overleftarrow{GRU}\left(x_s, \overleftarrow{h_{s+1}}\right) \qquad (2)$$

ここでh_sはタイムステップ s における隠れベクトルを表す．出力は式（1），式（2）を連結し，一つの状態とする．

$$h_s = \left[\overrightarrow{h_s}; \overleftarrow{h_s}\right] \qquad (3)$$

Decoder では各タイムステップ$t \in \{1, \ldots, n\}$でタイムステップ$t-1$における隠れベクトルと，予測された単語x_tで式（4）のように計算を行う．

$$dh_t = \overrightarrow{GRU}(w_t, dh_{t-1}) \quad (4)$$

ここでh_1は Encoder の出力により定義される．$t > 1$において，入力w_tは$t-1$における Decoder 出力である．Decoder の出力であるy_tは，Decoder の GRU の出力であるdh_tを softmax 層に通すことで各語彙に対する確率分布が出力される（式 5），（式 6）．

$$y_t = softmax(dh_t) \quad (5)$$

$$softmax(dh_t) = \frac{e^{dh_t}}{\sum_{i=1}^{n} e^{dh_i}} \quad (6)$$

3.2.2 CopyNet による文章生成

機械翻訳と比べ，文章校正，校閲は入力文章と正解文章の差異が小さい．そのことから，入力文章を出力文章にコピーすることが有効であると考えられる．CopyNet は，Attention をベースに入力文章に出現する単語がどれだけ出力文章に現れるかに焦点を当てて研究された．Attention は式（4）の GRU 入力に対し，Encoder の位置情報を含んだコンテキストc_tを追加する．Encoder の隠れベクトルであるh_sを用いてc_tは式（7）のように計算される．

$$c_t = \sum_{s=1}^{T_s} \alpha_{ts} h_s \quad (7)$$

ここでα_{ts}はh_jのアノテーションであり，式（8），式（9）のように計算される．

$$\alpha_{ts} = \frac{exp(e_{ts})}{\sum_{k=1}^{T_x} exp(e_{tk})} \quad (8)$$

$$e_{ts} = a(dh_{t-1}, h_s) \quad (9)$$

GRU の一つ前の状態を表すdh_{t-1}と Encoder の各ステップの隠れベクトルであるh_sを用いて，それぞれの対応関係のスコアをバッチごとの積により計算したものがe_{ts}である．

CopyNet の Encoder は実験の際にベースラインとして用いる双方向 seq2seq と全く同じである．Decoder は，Generate-Mode と Copy-Mode の二つのモードからなる複合的な確率モデルから，出力単語を選ぶ．出力単語は式（10）より算出される．

$$\tilde{y_t} = y_t + copy\_score \quad (10)$$

y_tは Generate-Mode で算出され，Decoder の出力単語の予測分布である．copy_score は Copy-Mode で算出され，入力文中の単語がコピーされる確率である．

一般的な RNN ベースの seq2seq でのタイムステップ t における Decoder の入力は，一つ前の出力のy_{t-1}である．CopyNet でのタイムステップ t における Decoder の入力は$e(y_{t-1}) + \zeta(y_{t-1})$である．$e(y_{t-1})$は$y_{t-1}$の word embedding である．$\zeta(y_{t-1})$は$y_{t-1}$が入力文にある場合，それに対応する隠れ層のベクトルを返し，なければ 0 を返す．

3.2.3 敵対的モデル

本研究では，seq2seq に対して敵対的な識別器を適用する．識別器は，生成器である seq2seq から生成された系列と正解データの系列を正しく識別することを目的とする．あらかじめ訓練された生成器から出力された文章と正解データを用いて識別器を訓練する．識別器は 2 層の双方向の LSTM から構成される．訓練された識別器に時系列を入力し，それが seq2seq から出力されたものなら 0 を，正解データならば 1 を出力する．seq2seq は，この識別器が騙されるような文章を出力することを目的とする．

4. 評価実験

4.1 使用したデータセット

本研究ではデータセットとして毎日新聞記事の 1995 年，2011 年のコーパスを用いる．1995 年の記事数は 111,501 である．2011 年の記事数は 96,563 である．また，コーパスは句点ですべて分割し，それぞれを別のデータとした．分割したデータのうち 15 文字以上の文章，808,591 件を本研究で用いるデータとした．

通常，文章を入出力として扱う際は単語ごとに分割し，それぞれに辞書番号を割り振る．すべての語彙に辞書番号を割り振る場合，コーパスが大きくなればなるほど必要とするメモリや処理時間が大きくなる．そのため，低頻度語を未知語として統一する必要がある．しかし，未知語への置き換えは扱えない単語を生み，深層学習の精度が下げる．そこで本研究では sentencepiece [11]を用いる．sentencepiece は，与えられたコーパスを指定された数で語彙分割する．そのため，本研究では未知語は存在しない．入出力の語彙数はともに 8,002 である．

4.2 実験設定

Encoder には 3 層の GRU を用いる．一方で Decoder では 2 層の双方向 GRU をベースとして用いる．ミニバッチサイズは 256 とした．また word embedding による密次元と隠れ層の次元は 256 とした．デコーディングの際，単語は貪欲法により最も確率が高いとされた単語を出力する．また各層におけるドロップアウトの確率は 0.5 とする．

4.3 結果

本研究における提案モデルの精度は BLEU[12]を用いて測定する．BLEU は，機械翻訳の分野において一般的な評価尺度であり，n-gram の適合率によって算出される．従来手法である双方向 seq2seq と提案手法の実験結果を表 1 に示す．

実験結果から，識別器なし CopyNet の精度が，日本語新聞記事において，ベースラインである seq2seq よりもおよそ 15%上回った．また seq2seq では，敵対的な識別器を適用した場合の精度が，適用しない場合よりもおよそ 3%上回った．CopyNet では，敵対的な識別器を適用した場合の精度が,適用しない場合よりも 2%上回った．

表 1 実験結果

BLEU	seq2seq	CopyNet
識別器なし	49.70	64.38
識別器あり	53.08	66.38

5. 考察

表 1 の結果より CopyNet が校閲タスクにおいて有効であることがわかった．このことから日本語文章の校閲タスクにおいても，入力文章を出力文章へコピーするという処理がモデルの精度に大きく寄与できるということがわかる．またベースラインに敵対的な識別器を適用した場合も精度の向上があることがわかった．このことは，単一の報酬のみを用いる seqGAN を日本語校閲タスクに用いた場合でも有効であることを示していると考えられる．

6. まとめと今後の展望

本研究では校閲前後の新聞記事に差異が比較的少ないことに着目し,校閲タスクに CopyNet を適用した．また，校閲タスクが生成タスクであることから seqGAN を適用した．結果から，提案手法が精度を向上させることが確認された．

提案手法では，校閲前後の差異が比較的少ないことに着目しているが，校閲作業独自の編集操作や同義表現の置換などはまったく考慮していない．今後の展望としては，校閲前後のパターンを調査し，上記の問題により焦点を当てたシステムを開発していきたいと考えている．

参 考 文 献

[1] Kyunghyun Cho，Fethi Bougares，Holger Schwenk，Dzmitry Bahdanau，Yoshua Bengio，“Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation”, Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing，pp.1724–1734, 2014.

[2] Ilya Sutskever, Oriol Vinyals, Quoc V. Le,“Sequence to sequence learning with neural networks”, arXiv:1409.3215, 2014.

[3] Dzmitry Bahdanau, KyungHyun Cho, Yoshua Bengio, “Neural machine translation by jointly learning to align and translate”, arXiv:1409.0473, 2015.

[4] Jiatao Gu，Zhengdong Lu，Hang Li，Victor O.K. Li，“Incorporating copying mechanism in sequence-to-sequence learning”, Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pp.1631–1640, 2016.

[5] Yuta Hitomi，Hideki Tamamori，Naoaki Okazaki，Kentaro Inui，“Proofread Sentence Generation as Multi-Task Learning with Editing Operation Prediction”, Proceedings of the Eighth International Joint Conference on Natural Language Processing, pp.436–441, 2017.

[6] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio, “Generative adversarial nets”, arXiv:1406.2661, 2014.

[7] Lantao Yu，Weinan Zhang，Jun Wang，Yong Yu, “SeqGan: sequence generative adversarial nets with policy gradient”, arXiv:1609.05473, 2017.

[8] 白井 圭祐，二宮 崇，森 信介，“敵対性学習を用いたニューラル機械翻訳”, 言語処理学会 第 23 回年次大会 発表論文集, pp.1105-1108, 2017.

[9] ksasao/homonym – GitHub, https://github.com/ksasao/homonym（参照 2020 年 1 月 2 日）

[10] mecab-ipadic-NEologd : Neologism dictionary for MeCab – GitHub, https://github.com/neologd/mecab-ipadic-neologd（参照 2020 年 1 月 2 日）

[11] sentencepiece – GitHub, https://github.com/google/sentencepiece（参照 2020 年 1 月 2 日）

[12] Kishore Papineni，Salim Roukos，Todd Ward，Wei-Jing Zhu, “BLEU: a Method for Automatic Evaluation of Machine Translation”, Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp.311–318, 2002.