マルコフ連鎖モデルを用いた文章校正のためのデータ拡張

永井 涼雅† 前田 亮‡

†立命館大学情報理工学研究科 〒525-8577 滋賀県草津市野路東 1-1-1
‡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 自動的に文章の校正を行う Grammatical Error Correction（GEC）では，一般に校正前のコーパスと校正後のコーパスを必要とする．一般的に，これらのコーパスは入手が困難である．本研究では，サイズの小さなコーパスからマルコフ連鎖モデルを用いて疑似的に校正文章の生成を行う．また，そうして作成した疑似的な校正後文章に対して従来手法である逆翻訳やルールベースの拡張を行い疑似的な校閲前文章を生成する．これらの拡張文章対を用いることにより，文書校正モデルの精度向上を目指す.

キーワード 深層学習，自然言語処理

1. はじめに

世界的な国際化の進展に伴い，日本にも国際化の波が来ている．日本政府観光局によると 2019 年の訪日外国人観光客数は 31,882,049 人にも上る[1]．2009 年の訪日外国人観光客数が 6,789,658 人であることから，この 10 年で日本の社会はより国際化したということがわかる．しかし，訪日外国人観光客にとって日本には言語という大きな障壁がある．観光庁が 2018 年から 2019 年に行った「日本の旅行中に困ったこと」についてのアンケートによると「施設等のスタッフとのコミュニケーションがとれない」が最も多く，全体の 20.6%を占めた[2]．このように，外国人観光客にとって言語の壁は大きく，その改善が国際化社会において重要であることがわかる．特に国際的なコミュニケーションの増加に伴い,留学生などの増加も見込まれる．日本語的な誤りや読みづらさを含んだ文章は，外国人はもちろんのこと，日本人ですら理解できないことがある． その中で，文章校正は特別な役割を持つ．

現在，日本語初学者のみならず，新聞記者のようなプロフェッショナルですら何気なく Word などに搭載されている自動校正機能を利用していることが考えられる．これらは，従来は巨大なコーパスから編集のパターンを保存しルールベースで校正するものが多かったが，現在では深層学習を活用したものも現れている[3].

深層学習は教師あり学習であるため，その学習には多量のコーパスを必要とする．特に文章校正タスクでは，校正前の文章と校正後の文章を用意する必要があるため，その構築には高いコストがかかる．本研究では，マルコフ連鎖モデルを用いて，校正後の文章を疑似的に拡張する．深層学習による文章分類によって，疑似校正後文章の中から，明らかにマルコフ連鎖モデルにより生成されたものを発見し,除外する.さらに，ルールベースや深層学習による逆翻訳など複数の手法を用いて，疑似校正後文章と本物の校正後文章から校正前の文章を疑似的に拡張する．それによって，コーパス全体の総量を増やし，従来の校正モデルの精度の向上を目指す.

2. 関連研究

2.1 マルコフ連鎖モデル

マルコフ性（Markov property）とは，ある現在の状態から，次の状態を予測する際に，それ以前の過去の状態は無関係であるという性質である[4]．例えば，状態が $x_1 \dots x_i$ $\{1, 2, \dots i-1, i\}$という時系列であるとき，x_{i+1}は$P(x_i|x_{i+1})$で決定される.マルコフ連鎖モデルの一種に，n 階マルコフ連鎖モデルがある．通常のマルコフ連鎖モデルがx_{i+1}を予測する際にx_iしか用いないのに対して，n 階マルコフ連鎖モデルでは$x_{i-n} \dots x_i$を用いてx_{i+1}を予測する．n は任意に指定する．このとき，n が大きければ大きいほど意味が通っているがオリジナリティのない文章であると言われている．マルコフ連鎖モデルは，コーパスの拡張を行う際に用いられることがある．機械学習用のデータとしてマルコフ連鎖モデルで拡張した文章は，コーパスを再利用するため，過学習を引き起こす可能性がある.

2.2 深層学習による自然言語処理

2.2.1 概要

自然言語処理で扱うデータは文章であることから，深層学習モデルは時系列処理に特化したモデルであることが想定される．RNN（Recurrent Neural Network）は深層学習モデルの一つである．ある時点の状態を次の状態に渡すことで，時系列のあるデータを処理できる.RNN には長期的な単語の依存関係を参照できないという欠点がある．そのため，LSTM（Long Short-Term Memory）や GRU（Gated Recurrent Unit）などの長期的な依存関係を記憶，または忘却するためのユニットを拡張したモデルも開発されている.

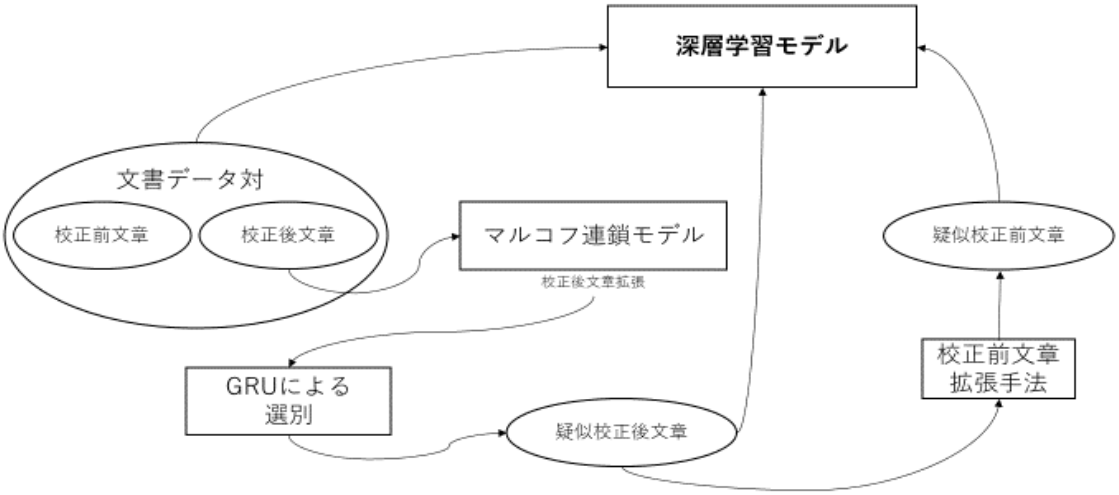

図 1：提案手法全体の流れ

本研究で扱う深層学習タスクは文章分類と文章生成である．文章分類は，入力した文章を複数のカテゴリに自動的に識別する．文章生成では，入力した時系列データから新たなデータを出力する．翻訳タスクや要約タスクなどで広く使われている．本研究では，文章校正を，誤り文から正しい文を生成する翻訳タスクとして扱う．

2.2.2 文章分類

深層学習を用いた文章分類タスクでは，入力した文章を複数のカテゴリのうちの一つに分類することを学習する．文書内の単語を，順序関係を考慮せずにひとまとめにする Bag of Words など，その手法は幅広い．近年では，深層学習を用いた文書分類が主流となっている．本研究では，双方向の GRU モデルをマルコフ連鎖から生成した疑似校正後文章の一部と実際の文章で学習する．学習した GRU モデルを用いて，疑似校正後文章のうち，「マルコフ連鎖で生成されたような文章」と削除する．

2.2.3 文章生成

従来の深層学習を用いた文章校正タスクの多くに時系列処理モデルが採用されている．Hitomi [5]らは，校閲前文章から校閲後文章を生成する校正タスクと校閲前文章のどこをどのように編集するか予測するタスクを LSTM ベースの Seq2seq と並行して学習させることで従来手法よりも高い精度を記録した．このように，Seq2seq が校正タスクに有効であることが一般的に知られている．Seq2seq による深層学習は，教師あり学習である．そのため，その精度はコーパスのサイズに左右される．校正タスクの精度を向上させるために，校正前のコーパスを拡張する手法はいくつか存在する．中島[6]らは，校正後の文章にルールベースによりノイズを混ぜ，校正前の文章を生成している．

3. 提案手法

3.1 概要

本研究では，本来手に入れることが難しい校正前の文章だけでなく，校正後の文章も疑似的に生成することによりコーパスのサイズを拡張する．まず，n 階マルコフ連鎖モデルにより，校正後の文章を疑似的に生成する．マルコフ連鎖モデルにより生成された文章には，文法的に誤りを含むものも多い．そこで，マルコフ連鎖モデルで疑似的に生成したコーパスと実際のものを用いて学習した分類器により文法的に筋の通ったものとそうでないものを分類し，前者のみを校正後の文章として利用する．疑似的に生成したものを含む校正後コーパスを対象に，ルールベースや逆翻訳などの複数の手法により，校正前文章を疑似的に生成する．そうして生成したコーパスを用いた場合とオリジナルのコーパスだけを用いた場合で Seq2seq による校正タスクの精度がどれほど変化するかを測定する．この提案手法の全体の流れを図 1 に示す．精度指標としては，BLEU を用いて，結果を考察する．

3.2 マルコフ連鎖モデルによるデータ拡張

n 階マルコフ連鎖モデルを用いて，校正後文章を拡張する．n 階マルコフ連鎖モデルでは n の値を大きく設定するとオリジナリティが失われ，小さく設定すると文脈的な正しさを失う．本研究では，GRU による文章の選別を行うため，生成文章のオリジナリティを重視し n=2 と設定する．マルコフ連鎖モデルにより生成された文章の例を表 1 に示す．

表 1：マルコフ連鎖モデルによる出力例

1. できるだけ早くそれをした。
2. あなたは何になりますか。
3. ボブは会に参加しなかったなかった。
4. 彼が試験の結果はまだ生きている。

表 1 を見ると，1 と 2 は意味的に正しく，3 と 4 は誤っている．本研究では，コーパスと同程度の文書数をマルコフ連鎖により生成する．

3.3 拡張データの選別

マルコフ連鎖モデルは，表 1 の 3 と 4 のように意味的に誤っている文章を生成することがある．そのような文章が，コーパス内に「校正後の文章」として存在してしまうと，それがノイズとなり，モデルの精度を下げてしまう可能性がある．そのため，そのような文章をあらかじめ排除する必要がある．本研究では，山本[7]らの「SNOW T15:やさしい日本語コーパス」5 万件のうち，やさしい日本語文章 1 万件とそれを用いて生成した疑似校正後文章 1 万件を用いて分類器を学習する．通常のやさしい日本語を「文法的に正しい文章」，マルコフ連鎖により疑似的に生成したものを「文法的に誤った文章」と定義し，文書分類タスクとして学習する．学習の結果，GRU は 68.4%の精度を記録した．この学習済み分類器を用いて，4 万件のやさしい日本語文章から生成した疑似校正後文章 4 万件を分類する．本来，すべてが偽物であると分類されるはずだが，本物であると分類されるものもある．本研究では，本物であると分類された文章を新しいコーパスとして，4 万件の中に加え入れる．

3.4 疑似校正前文章の生成

拡張した疑似校正後文章には対となる校正前文章が存在しない．そこで，本研究では分散表現による名詞と形容詞の置換，ルールベースによる置換や削除処理，逆翻訳の 3 種類の手法によって疑似的に校正前文章を生成する．

3.4.1 分散表現による生成

本研究では，[8]にて公開されている学習済みの Word2vec モデルを利用する．疑似校正後文章に対して，mecab-ipadic-NEologd により形態素解析を行い，品詞が名詞または形容詞であったものに対して，最も類似する単語で置換する．

3.4.2 ルールベースによる生成

ルールベースにより，機械的かつ確率的に疑似校正後文章を編集し，疑似的に校正前文章を生成する．本研究では，ルールを表 2 のように設定する．各ルールの確率はそれぞれ独立しており，処理は文単位で行う．

表 2：疑似校正前文章生成のルール

1. 33%の確率で一様乱数により決定された二つの単語の位置を入れ替える，33%の確率でこの操作を二回行う，34%でなにもしない
2. 各単語に対して，5%の確率で削除を行う
3. 各単語に対して，10%の確率で同じ単語を 1 つ後ろに挿入する

3.4.3 逆翻訳による生成

文章校正タスクでは，ニューラルネットワークに校正前文章から校正後文章を生成することを学習させるのが一般的である．文章拡張では，その逆の入出力を行うことで，疑似的な校正前文章の生成を行う．このタスクを逆翻訳という．小川[9]らは，逆翻訳によるデータ拡張が深層学習モデルの精度向上に寄与することを示した．Gu[10]らは，文章校正タスクの入出力の差異が少ないことに注目し，入力から出力へ単語をコピーすることを学習するモデルを発表した．本研究では，逆翻訳モデルとして，Gu らのモデルを採用する．「SNOW T15:やさしい日本語コーパス」のむずかしい日本語を入力とし，やさしい日本語を出力として学習する．学習したモデルに疑似校正後文章を入力し，疑似的に校正前文章を生成する．

3.5 ベースモデル

本研究の最終的な目標は，ベースラインである Seq2seq の校正精度を向上させることである．Seq2seq は GRU から構築する．拡張データセットが，入力コーパスから生成されることから過学習を起こすことが想定される．そのため，Seq2seq に word-dropout を適用し，実験を行う．Word-dropout は，入力単語のベクトルを一定確率で 0 にする．

4. 評価実験

4.1 使用したデータセット

本研究の全体を通して「SNOW T15:やさしい日本語コーパス」を用いる．本コーパスは，外国人にとっても理解しやすいやさしい日本語，理解しづらいむずかしい日本語，その英訳の 3 対を 5 万件収録している．本研究では，そのうちやさしい日本語とむずかしい日本語のみを利用する．やさしい日本語を正しい文章，むずかしい日本語を誤っている文章として扱う．

4.2 実験設定

マルコフ連鎖や 3 種類の手法により拡張したコーパスを，分類器を学習する際に用いた 1 万件を除いた 4 万件に加え，Seq2seq を学習する．ベースラインとする Seq2seq は 1 層の双方向 GRU から構成されている．また，各層には Dropout を適用し，それらの確率は 0.2 とする．出力層では Softmax を用いて，単語の出現確率を計算する．入力に用いる文章データは，

表 3：実験の結果

拡張なし	拡張+逆翻訳	拡張+ルールベース	拡張+word2vec
11.92	12.42	12.97	13.56

Sentencepiece[11]を用いて形態素解析を行う．Sentencepiece は，教師なし学習により，コーパスを指定された単語数に抑えられるように辞書を作成し，形態素解析ができる．これにより，コーパス内の低頻度語を out-of-vocabulary として処理する必要がなくなる．

疑似校正後文章の選別を行う分類器は双方向，隠れ層のサイズは 128，各層に Dropout を適用し，その確率は 0.2 とする．また，活性化関数は sigmoid とする．また，実験の際に使用する Seq2seq モデルは活性化関数を Softmax とし，それ以外は分類器と同様とする．Seq2seq の学習は，50 エポック，バッチサイズは 128，最適化関数は Adam，損失関数は Pytorch の NLLoss を用いる．

4.3 実験結果

本研究では，マルコフ連鎖により 4 万件の文章を生成した．そのうち，元のコーパスと全単語が重複したものは存在しない．また，分類器によって 4 万件のうち，マルコフ連鎖により生成されたように見える文章と判断され，削除されたものは 9,039 件である．また，残った疑似校正後文章に対して 3.4 節の 3 つの手法をそれぞれ用いて疑似的に 3 種類の校正前文章を生成する．

分類器の学習に用いなかった「SNOW T15:やさしい日本語コーパス」4 万件のうち，1 万件を Seq2seq のテストデータとし，残りのデータを 8:2 の比率で学習データと検証データにわける．その学習データに，マルコフ連鎖により生成された疑似校正後文章と疑似校正前文章を追加し，学習を行う．テストデータの校正前文章を入力とし，学習した Seq2seq の精度を測定する．結果は表 2 となる．精度は BLEU によって計測する．BLEU は n-gram の一致度により精度を測る．結果は 0～1 で表示され，その数字に 100 を掛けることで 100 分率として表記することができる．本研究では，BLEU の結果に 100 を掛け，少数第 3 位を四捨五入し表示する．

5. 結果と考察

マルコフ連鎖モデルと 3 種類の手法によって，校正タスクにおける Seq2seq モデルの BLEU スコアがわずかであるが上昇した．この手法は，コーパスの件数を校正前後問わず増やせることから，極端にコーパスが少ない場合などに有用である．特に文章校正タスクは，校正前後のコーパスを手に入れることが難しいため，有効である．

実験結果のうち，もっとも精度が高いものでさえ，拡張なしのものと比較して 1.64 しか変わらない．この原因は二つ考えられる．一つ目は，過学習である．マルコフ連鎖モデルは，文章の単語順序確率を学習し文章を生成することから，元のコーパスに出現する単語と語順以外を出力することができない．そのため，全文一致ではなくとも部分的に一致する文章が大量に生成されることになる．その結果，モデルの学習を阻害されることが考えられる．今後の研究で，校正後文章の単語や語順を，文法的正しさを崩さないように，全く新しい疑似校正後文章を生成する手法を考案する必要がある．二つ目は，マルコフ連鎖モデルにより生成された文章の意味的正しさである．現状では，マルコフ連鎖モデルで生成した文章を，GRU を用いて選別している．GRU 自体の精度は期待できるほど高くなく，選別された文章の意味的正しさは期待できない．今後，分類器をより精度が期待できるモデルに変更する必要がある．

参　考　文　献

[1] 日本政府観光局：月別・年別統計データ（訪日外国人・出国日本人），https://www.jnto.go.jp/jpn/statistics/visitor_trends/（参照 2020 年 12 月 16 日）

[2] 国土交通省観光庁：訪日外国人が旅行中に困ったこと、受入環境整備の課題が明らかになりました～受入環境について訪日外国人旅行者にアンケート調査を実施～，https://www.mlit.go.jp/kankocho/news08_000267.html（参照 2020 年 12 月 16 日）

[3] 杉本昭彦：リクルートの校閲 AI が驚異的な効果 検出率は人を超え数秒で完了，日経 XTREND，2018.

[4] 太田博三：文章自動生成手法の一考察 -文と文のつながりを課題として-，情報処理学会研究報告，Vol.2017-IFAT-127 No.3，2017 年 7 月 22 日．

[5] Yuta Hitomi，Hideaki Tamori，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，436–441，2017.

[6] 中島寛人，山田剛：誤り文の自動生成による校正エンジンの学習，言語処理学会 第 24 回年次大会，2018 年 3 月．

[7] 山本和英，丸山拓海，角張竜晴，稲岡夢人，小川耀一朗，勝田哲弘，髙橋寛治：やさしい日本語対訳コーパスの構築，言語処理学会第 23 回年次大会，pp.763-766，2017 年 3 月．

[8] Pre-trained word vectors of 30+ languages– GitHub，https://github.com/Kyubyong/wordvectors

（参照 2020 年 12 月 16 日）

[9] 小川耀一朗，山本和英：日本語誤り訂正における擬似誤り生成による訓練データ拡張，言語処理学会第 26 回年次大会，pp.505-508 ，2020 年 3 月.

[10] 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.

[11] SentencePiece ‐ GitHub，
https://github.com/google/sentencepiece
（参照 2020 年 12 月 16 日）