Vol.2023-NL-256 No.9
2023/5/28

系列変換モデルにおける語彙制約を用いた複数出力候補の統合

宮野 稜大[1,a)]　梶原 智之[2,b)]　荒瀬 由紀[1,c)]

概要：機械翻訳では生成文を後編集することで品質を改善する Automatic Post-Editing (APE) が用いられるが，構築コストの高い人手タグ付きコーパスを要するため，他の言語生成タスクに適用するハードルは高い．本研究では APE に着想を得つつ，人手によるコーパスを必要としない手法を提案する．具体的には，系列変換モデルの N ベスト出力を用いて正誤予測を行い，誤りと予測された語を負の制約，正しいと予測された語を正の制約として，語彙制約を適用したデコードを再び行う．これにより，N ベスト出力に含まれていた正しい語を含みつつ，誤りを避けた文を生成する．言い換え生成および要約タスクにおける提案手法の有効性を評価する実験を行った．その結果，いずれのタスクにおいても提案手法はビームサーチによる文生成を上回る性能を達成することを確認した．

1. はじめに

Bidirectional Auto-Regressive Transformer (BART) [1] や Text-To-Text Transfer Transformer (T5) [2] に代表されるように，大規模コーパス上で事前訓練された系列変換モデルを所望のタスクのコーパスで fine-tuning するアプローチにより，言語生成の品質は飛躍的に向上した．言語生成タスクの一種である機械翻訳では，出力文の品質をさらに改善するアプローチとして，翻訳結果を自動的に修正する Automatic Post-Editing (APE) が研究されている [3–5]．APE モデルは，入力文，機械翻訳モデルの出力文，人間が修正した出力文の三つ組からなる Post-Edit コーパスを用いる教師あり学習で訓練される．大規模な Post-Edit コーパスを用いることで，強力な機械翻訳モデルによる出力文の品質を顕著に改善できる [5] が，Post-Edit コーパスは構築コストが高いため，既にコーパスが存在する機械翻訳以外の言語生成タスクへの APE 適用は困難である．

本研究は APE に着想を得つつ，Post-Edit コーパスに依存しない生成文の品質向上手法を提案する．出力文に含まれる誤りおよび正解の「検出」のみであれば，系列変換モデルの fine-tuning に用いるパラレルコーパスにより訓練できることに着目し，それらを語彙制約 [6] として文を再生成するアプローチを取る．さらに N ベストの出力文を用いることで，出力すべき語・出力すべきでない語を幅広く獲得し，語彙制約に適用する．具体的には，まず系列変換モデルにビームサーチを適用することで N ベストの出力文を生成した後，N ベスト出力に含まれる各トークンの正誤予測を行う．そして，入力文を再度系列変換モデルに入力し，語彙制約を適用したデコードを行うことで，予測された誤りを含まず，正解と期待されるトークンを含んだ出力文を得る．提案手法はパラレルコーパスが存在するあらゆる言語生成タスクに適用でき，高い汎用性を持つ．

提案手法の有効性を検証するために，言い換え生成タスク [7] および要約タスク [8–10] という 2 つの言語生成タスクにおける評価実験を実施した．その結果，両タスクにおいて，提案手法に基づく語彙制約を用いた手法がベースラインを上回る性能を達成した．提案手法における語彙制約の品質が生成文に与える影響について分析し，正の制約の再現率向上，偽陽性率低減が重要であることを示した．

2. 関連研究

本章では提案手法と関連が深い技術として，APE，モデルアンサンブル，語彙制約について議論する．

2.1 機械翻訳における Automatic Post-Editing

APE [3,4] で用いる Post-Edit コーパスとして，100K 以上の三つ組から構成される大規模なコーパスを用いる場合，

[1] 大阪大学大学院情報科学研究科
Graduate School of Information Science and Technology, Osaka University
[2] 愛媛大学大学院理工学研究科
Graduate School of Science and Engineering, Ehime University
[a)] miyano.ryota@ist.osaka-u.ac.jp
[b)] kajiwara@cs.ehime-u.ac.jp
[c)] arase@ist.osaka-u.ac.jp

最先端の APE モデルが強力な機械翻訳システムの出力を大幅に改善することが報告されている [5]．一方で，小規模な Post-Edit コーパスを用いる場合，APE モデルは機械翻訳システムの出力を顕著に改善できないことが報告されている [11, 12]．APE を用いた系列変換モデル出力の改善には，構築コストの高い大規模な Post-Edit コーパスが必要であり，これが存在しないタスクには APE の適用が困難である．一方，提案手法は Post-Edit コーパスに依存しないため，幅広い言語生成タスクへの適用が可能である．

2.2 言語生成タスクにおけるモデルアンサンブル

提案手法は N ベスト出力の正誤予測に基づく語彙制約により新たな出力を生成するが，これは N ベスト出力を正誤予測に基づき組み合わせているとも捉えられる．この観点では，言語生成におけるモデルアンサンブル手法とも関連が深い．言語生成タスクにおけるモデルアンサンブルでは，複数の系列変換モデルを用いることで，生成文の品質向上を目指している [13, 14]．複数の訓練済み系列変換モデルを用意し，各デコードステップにおいて全モデルのトークン予測確率の平均を計算した後，出力トークンを決定する．アンサンブルするモデルの数に比例して訓練コストが増大するため，多数のモデルを用いることは現実的に困難である．一方，提案手法では，単一の系列変換モデルの複数の出力候補を統合することにより，生成文の品質向上を目指している．このため，訓練コストは統合する出力候補数に依存しない．

2.3 語彙制約による生成文の品質改善

提案手法で用いる語彙制約は，タスクに関連する人間の知識を基に制約を生成し，デコーダを制御することで出力文の品質を向上する手法として用いられてきた．語彙制約を適用した言語生成として，Chatterjee ら [15] および Hokamp ら [16] は，専門用語の対訳辞書を制約として機械翻訳タスクに語彙制約手法を適用した．Lu ら [6] は，所与のキーワードを含む文を生成する Table-to-Text タスクや質問文生成タスクに語彙制約を用いている．Dehghan ら [17] および Zetsu ら [18] は，トークンの難易度に基づいて制約を生成してテキスト平易化に，Kajiwara [19] はスタイルに関連する語彙を制約としてスタイル変換に応用している．これら既存研究では，語彙制約を生成するためのタスクに固有の知識が明確であることを前提としている．一方提案手法はタスクに依らず利用可能な系列変換モデル出力を基に語彙制約を作成するため，幅広い言語生成タスクへの適用が可能である．

3. 事前知識：NEUROLOGIC★ の語彙制約

提案手法では，state-of-the-art の語彙制約手法である NEUROLOGIC★ [6] を用いる．NEUROLOGIC★ は語彙制約を高速に実現するデコーディング手法であり，系列変換モデルの構造を変更することなく，訓練済みのモデルに直接適用できる．

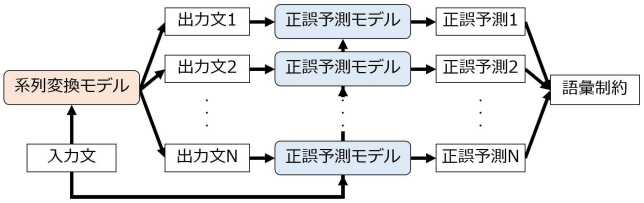

図 1 提案手法の構成

NEUROLOGIC★ は与えられた語彙制約を基に，以下の手順で各時刻のトークンを生成する．

(1) 出力候補のトークンについて，そのトークンの後に生成されるトークン列 (Lookahead) を予測する．
(2) 予測した Lookahead を基に，将来の語彙制約充足率を計算する．
(3) 出力候補トークンの生成確率と将来の語彙制約充足率を基に，出力候補の枝刈りを行う．
(4) 語彙制約充足の種類に基づき，残された出力候補のグルーピングを行う．
(5) 出力候補を生成確率および語彙制約充足率を基にスコア付けし，各グループの最良候補の中から出力トークンを選択する．これにより，出力空間の幅広い探索を行う．

以上の手順により，生成確率と将来の語彙制約充足率が高い候補を高速に探索し出力できる．

4. 提案手法

提案手法の構成を図 1 に示す．提案手法では，ある入力文について系列変換モデルによる出力を複数生成し，全ての出力について正誤予測を行う（4.1 節）．それらを統合し，正と予測されたトークンを出力文に加えるべき正の語彙制約，誤と予測されたトークンを出力文から除外すべき負の語彙制約とすることで語彙制約を生成する（4.2 節）．そして再度入力文を系列変換モデルに入力し，デコーダに NEUROLOGIC★ [6] による語彙制約を適用して最終的な出力文を得る．

4.1 系列変換モデル出力の正誤予測

系列変換モデル出力の正誤予測では，図 2 に示す通り，予測モデルに RoBERTa [20] を用い，各トークンに対する正，誤の二値分類を行う．各トークンの正誤予測においては，出力文だけでなく系列変換モデルへの入力が手がかりになると考えられる．そこで，図 2 に示す通り，予測モデルへの入力は “<INPUT> 系列変換モデル入力 <QUERY> 系列変換モデル出力” とする．[*1]

[*1] <INPUT> および <QUERY> は特殊トークンとして RoBERTa の語彙に追加する．

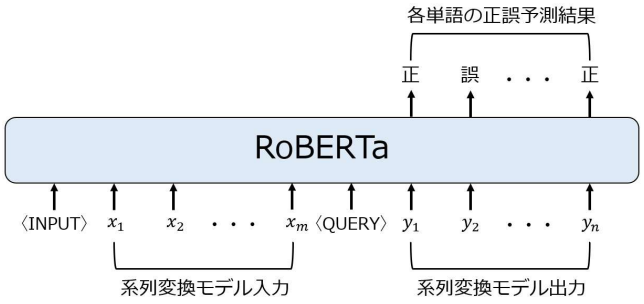

図 2 RoBERTa による系列変換モデル出力の正誤予測

正誤予測モデルの訓練には，系列変換モデルの訓練コーパスと同じものを用いる．正解ラベルは系列変換モデルの出力文と参照文を比較することで作成する．訓練コーパスの入力文を系列変換モデルに入力し，ビームサーチにより N ベストの出力文を得る．各出力文について，参照文に含まれるトークンに正のラベルを，含まれないトークンに誤のラベルを付与する．以上のようにして生成した正解ラベルを用いて，RoBERTa を fine-tuning する．

4.2 正誤予測を統合した語彙制約の作成

提案手法では，ある入力文について得られた N ベストの出力文について正誤予測を行う．正誤予測は文脈に依存するため，同じトークンが出力文によっては異なるラベルを予測されることがある．そこで，正誤ラベルの予測回数の多数決を取ることで最終的なラベルを決定し，語彙制約に追加する．正のラベルを持つトークンは正の語彙制約として出力を促進し，誤のラベルを持つトークンは負の語彙制約として出力を抑制する．なお，正誤ラベルの予測回数が等しい場合，そのトークンは語彙制約に含めない．

5. 実験設定

言語生成タスクにおける提案手法の有効性を検証するため，言い換え生成および要約による評価実験を実施する．本章では提案手法の実装について述べ，評価実験における共通の設定を説明する．

5.1 正誤予測モデル

正誤予測モデルは HuggingFace Transformers [21] ライブラリを用いて実装し，事前学習済みモデルとして “roberta-base”[*2] [20] を用いた．4.1 節で述べた通り，fine-tuning 済みの系列変換モデルとその訓練データを用いて，正誤予測モデル用の正解ラベルを作成する．RoBERTa の訓練では 1 エポックごとに検証データで F1 スコアを計算し，3 回改善が見られなくなったところで fine-tuning を終了する．

出力数 N，すなわち系列変換モデルのデコーダにおけるビーム幅の候補を，$1, 5, 10, 20, 30 \ldots, 100$ から探索する．それぞれのビーム幅で正誤予測モデルを訓練し，検証データの F1 スコアが最も高いモデルを最終的な正誤予測モデルとして採用する．

[*2] https://huggingface.co/roberta-base

表 1 評価実験におけるデータセットの文章数

データセット	訓練用	検証用	評価用
DIRECT	64,126	–	7,372
CNN/Daily Mail (Version 3.0.0)	287,113	13,368	11,490
XSum	204,045	11,332	11,334

また評価実験では正誤予測モデルが理想的な予測をできたと仮定した場合のオラクルの語彙制約による提案手法の性能も検証する．出力文と参照文双方に存在する語を正の制約に，出力文と参照文を比較し，出力文にのみ存在する語を負の制約とすることでオラクルの語彙制約を作成し，NEUROLOGIC* の制約として適用する．

5.2 デコーダの設定

NEUROLOGIC* の実装は Lu ら [6] の実装[*3]に従う．ただし，Lu らの公開プログラムでは負の語彙制約が実装されていないため，負の語彙制約を適用できるようコードの修正を行った．NEUROLOGIC* を用いて語彙制約を考慮したデコードを行う際のビーム幅は，ベースラインとする純粋なビームサーチと同じ設定を用いるものとする．具体的には，ベースとする系列変換モデルの各タスクにおけるビームサーチのデフォルト値を用いた．

6. 言い換え生成タスクにおける評価

言い換え生成タスクのひとつである，間接的・直接的応答間の言い換えタスク [7] における提案手法の有効性を検証する．間接的応答とは，発話者の要求や意図を直接的に言及せず，言外に含んだ発話であり，本タスクでは間接的応答とそれに対応する直接的応答を相互に言い換える．

6.1 コーパスと評価指標

データセットには，表 1 に示す DIRECT (Direct and Indirect REsponses in Conversational Text) [7] を用いる．DIRECT コーパスは間接・直接発話言い換えコーパスであり，71,498 の間接的応答と直接的応答の対からなる．

DIRECT コーパスは既存のマルチドメイン・マルチターンのタスク指向対話コーパス MultiWOZ 2.1 (Multi-Domain Wizard-of-Oz 2.1) [22,23] を拡張したものであり，MultiWOZ における対話履歴と元の応答，元の応答をより間接的に言い換えた発話，元の応答をより直接的に言い換えた発話が利用できる．本実験では，間接的応答を直接的応答に言い換える Indirect-to-Direct タスク，直接的応答を間接的応答に言い換える Direct-to-Indirect タスクの両方に取り組む．なお，両タスクについて，対話履歴を考慮する設定と考慮しない設定の 2 種類の設定を用いる．評価指標は

[*3] https://github.com/GXimingLu/a_star_neurologic

表 2 間接直接発話変換タスクにおける性能評価 (BLEU). † はベースラインとの有意差が存在するスコアを表す.

	Indirect-to-Direct		Direct-to-Indirect	
	w/ history	w/o history	w/ history	w/o history
beam-search	35.57	34.38	26.92	26.63
NEUROLOGIC★ (正負)	36.43†	35.42†	30.21†	30.57†
NEUROLOGIC★ (正)	**36.95**†	**35.94**†	**30.89**†	**31.33**†
NEUROLOGIC★ (負)	35.84†	34.82†	29.97†	30.12†
NEUROLOGIC★ (正負，オラクル)	**65.55**†	**65.31**†	**60.23**†	**60.71**†
NEUROLOGIC★ (正，オラクル)	57.85†	57.38†	49.42†	50.15†
NEUROLOGIC★ (負，オラクル)	51.60†	50.98†	45.24†	45.54†

DIRECT の設定に従って BLEU [24] を用い，ブートストラップ法 [25] に基づき有意水準 5% で有意差検定を行う．

6.2 系列変換モデルと比較手法

系列変換モデルとして，DIRECT コーパスにて fine-tuning した BART [1] を用いる．モデルの実装には HuggingFace Transformers ライブラリを使用し，事前学習済みモデルとして “facebook/bart-base”[*4]を用いた．BART の入力フォーマットおよび訓練設定は高山ら [7] に従う．

実験では，語彙制約を用いず fine-tuning した BART においてビームサーチを用いて出力を生成するベースライン（“beam-search” と表記）と比較する．ビーム幅は “facebook/bart-base” のデフォルトに従い 4 とした．また，正負それぞれの語彙制約の効果を検証するため，正の語彙制約のみを用いる提案手法（“NEUROLOGIC★ (正)” と表記）および負の語彙制約のみを用いる提案手法（“NEUROLOGIC★ (負) ” と表記）の性能も評価する．

6.3 実験結果

表 2 に実験結果を示す．ここで “w/ history” は対話履歴を考慮する設定を表し，“w/o history” は対話履歴を考慮しない設定を表す．上段には，提案手法に基づく語彙制約を用いた場合の性能を示す．全てのタスクで提案手法 (NEUROLOGIC★ (正負)) はベースラインよりも有意に高い BLEU スコアとなった．また，語彙制約の種類による性能変化に注目すると，正の語彙制約のみを用いる場合（NEUROLOGIC★ (正)）が最も高い性能を達成している．

下段には，参照文を用いて生成したオラクルの語彙制約を用いた場合の性能を示す．オラクルの制約を用いる場合，正負両方の語彙制約を用いる提案手法が正のみ，負のみの制約を用いる場合に比べ顕著に高い性能となった．予測した語彙制約を用いる場合と比べて，Indirect-to-Direct タスクでは最大 29.39 ポイント，Direct-to-Indirect タスクでは最大 30.14 ポイント BLUE スコアが向上している．

以上の結果から，正誤予測モデルの性能を改善することで間接直接発話変換タスクの性能をさらに向上できると考えられる．語彙制約と生成品質に関する考察は 8 章で行う．

7. 要約タスクにおける評価

本章では要約における提案手法の有効性を評価する．

7.1 コーパスと評価指標

データセットには, 表 1 に示す CNN/Daily Mail [8,9][*5]および XSum (The Extreme Summarization) [10] を用いる．CNN/Daily Mail データセットは，CNN および Daily Mail の記事とその記事の要約であるハイライトを収集したデータセットであり，約 310K のニュース記事とハイライトの対からなる．CNN/Daily Mail の文数の平均は，記事が 30.7，ハイライトが 3.8 である．

XSum データセットは，BBC の記事とその要約を収集したものであり，約 230K の記事と要約の対からなる．XSum の文数の平均は，記事が 19.8，要約が 1.0 である．XSum データセットは，CNN/Daily Mail データセットよりも要約の文数が少なく，抽象的な要約を行う必要がある．

要約タスクにおいては，評価指標として ROUGE [26] を用い，近似的ランダム化検定 [27] に基づき，R=1000，有意水準 5% で有意差検定を行う．

7.2 系列変換モデルと比較手法

系列変換モデルとして Lewis ら [1] により fine-tuning された BART である “facebook/bart-large-cnn”[*6]および “facebook/bart-large-xsum”[*7]を用いる．なお，正誤予測モデルの入力の最大長は 512 であるため，入力長が 512 を超える場合，要約記事を前半と後半に分割し，それぞれを BART の出力要約と繋げて正誤予測モデルの入力とする．

また 6 章同様，語彙制約を用いず純粋なビームサーチのみ用いる BART をベースライン (“beam-search”) とする．CNN/Daily Mail では “facebook/bart-large-cnn” のデフォ

[*4] https://huggingface.co/facebook/bart-base

[*5] バージョン 3.0.0 を利用.

[*6] https://huggingface.co/facebook/bart-large-cnn

[*7] https://huggingface.co/facebook/bart-large-xsum

表 3　要約タスクにおける性能評価 († はベースラインとの有意差が存在するスコアを表す)

	CNN/Daily Mail				XSum			
	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-LSUM	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-LSUM
beam-search	44.04	21.08	30.63	40.99	45.46	22.35	37.23	37.21
NEUROLOGIC★ (正負)	**44.99**†	21.64†	31.58†	**41.99**†	45.69†	22.37	**37.39**†	**37.39**†
NEUROLOGIC★ (正)	44.76†	21.33†	30.66	41.72†	**45.87**†	22.19†	37.16	37.16
NEUROLOGIC★ (負)	44.72†	**21.72**†	**31.61**†	41.76†	45.18†	22.25	37.23	37.24
NEUROLOGIC★ (正負，オラクル)	**61.74**†	**33.84**†	**38.73**†	**54.75**†	**67.23**†	**42.72**†	**54.67**†	**54.69**†
NEUROLOGIC★ (正，オラクル)	56.05†	30.58†	36.98†	52.05†	61.22†	35.68†	47.54†	47.52†
NEUROLOGIC★ (負，オラクル)	53.30†	29.89†	38.40†	50.09†	55.33†	32.88†	47.10†	47.14†

ルトのビーム幅に従い，デコードにはビーム幅 4 のビームサーチを用いる．Xsum では “facebook/bart-large-xsum” のデフォルトのビーム幅に従い，デコードにはビーム幅 6 のビームサーチを用いる．加えて，正の語彙制約のみ用いる提案手法 (“NEUROLOGIC★ (正)”)，負の語彙制約のみ用いる提案手法 (“NEUROLOGIC★ (負)”) も評価する．

7.3　実験結果

表 3 に要約タスクの実験結果を示す．上段には，提案手法に基づく語彙制約を用いた場合の性能を示す．CNN/Daily Mail，XSum 双方において，正負の語彙制約を用いる提案手法 (NEUROLOGIC★ (正負)) がベースラインを有意に上回り，人手評価と高い相関を持つ ROUGE-L [26] において最も高いスコアを達成している．

下段には，参照文を用いて生成したオラクルの語彙制約を用いた場合の性能を示す．いずれも正負の語彙制約を用いる提案手法 (NEUROLOGIC★ (正負)) が正のみ，負のみの制約を用いるよりも顕著に高い性能となった．正誤予測を用いる場合と比べて，オラクルの制約を用いることで，CNN/Daily Mail では ROUGE-1 は最大 16.75 ポイント，ROUGE-2 は最大 12.20 ポイント，ROUGE-L は最大 6.85 ポイント，ROUGE-LSUM は最大 12.76 ポイント向上している．Xsum では ROUGE-1 は最大 21.54 ポイント，ROUGE-2 は最大 20.35 ポイント，ROUGE-L は最大 17.28 ポイント，ROUGE-LSUM は最大 17.30 ポイント向上している．

以上の結果から，要約タスクにおいても，正誤予測モデルの性能が提案手法による生成文の品質改善において重要なことが分かる．本実験では要約タスクにおける記事長が 512 トークンを超える場合，半分に分割してそれぞれ入力するが，それによって正誤予測性能が低下したと考えられる．要約においては正誤予測において長い文章を扱えるよう，改善が必要である．語彙制約と生成文の品質に関する考察を 8 章で行う．

8.　語彙制約と生成文品質に関する考察

本章では語彙制約品質が最終的な生成文に与える影響について，定量的 (8.1 節) および定性的 (8.2 節) に考察する．

8.1　定量的分析

正の制約は指定した語の生成を促進するため，生成文に直接影響を与える．また正の語彙制約に誤りである語を指定してしまうと「本来出力すべきでない語の出力を促進」する．一方，負の語彙制約は指定した語の出力は抑制するが，「出力すべきでない語」は膨大に存在するため，生成文への影響は間接的と言える．以上より語彙制約が満たすべき性質として，(1) 正の制約の再現率が高いこと，(2) 正の制約に出力すべきでない語を含む偽陽性率が低いこと，(3) 負の語彙制約の再現率が高いこと，が考えらえる．これらを検証するため，正の語彙制約を C_p，負の語彙制約を C_n，参照文中の語の集合を W_{ref}，N ベスト中の語の集合を W_N として，以下の指標を計算する．参照文に対する正の制約の再現率を $R_{p,\text{oracle}} = \frac{|C_p \cap W_{\text{ref}}|}{|W_{\text{ref}}|}$，参照文に対する偽陽性率を $F_{p,\text{oracle}} = \frac{|C_n \cap W_{\text{ref}}|}{|W_{\text{ref}}|}$ とする．N ベスト出力に対する正の制約の再現率を $R_{p,N} = \frac{|C_p \cap W_N \cap W_{\text{ref}}|}{|W_N \cap W_{\text{ref}}|}$，負の制約の再現率を $R_{n,N} = \frac{|C_n \cap \{W_N \setminus W_{\text{ref}}\}|}{|W_N \setminus W_{\text{ref}}|}$ とする．

結果を表 4 に，正負の語彙制約を用いた提案手法のベースライン (ビームサーチ) に対するスコアの差分とともに示す．いずれのタスクでも，N ベスト出力に対する正の制約の再現率は 49% から 58% にとどまっている．また偽陽性率が 10% から 16% となっており，提案手法に悪影響を及ぼしていると考えられる．正誤予測における正の制約の再現率向上，偽陽性率低減が今後の課題である．

間接直接発話変換タスクでは，N ベスト出力に対する負の制約の再現率が要約タスクに比べて低い．これらのタスクでは正の語彙制約のみ用いる提案手法が高い性能を持つことが示されたが，今後負の制約の再現率を改善することが効果的と考えられる．

これらの指標の絶対値とベースラインに対する性能向上幅には明確な相関関係はみられない．これは正負の語彙制約による効果はベースとなる系列変換モデルのもつ言語生成能力にも依存するためと考えられる．

8.2　モデルの出力例

Direct-to-Indirect タスクにおける入力文，参照文，提案手法に基づき作成された語彙制約，提案手法に基づくモデル出力 (NEUROLOGIC★ (正負))，ビームサーチによる

表 4 各タスクにおける正誤予測モデルの性能評価

タスク	$R_{p,\mathrm{oracle}}(\uparrow)$	$F_{p,\mathrm{oracle}}(\downarrow)$	$R_{p,N}(\uparrow)$	$R_{n,N}(\uparrow)$	BLEU	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-LSUM
Indirect-to-Direct w/ history	0.36	0.10	0.53	0.07	+0.86	–	–	–	–
Indirect-to-Direct w/o history	0.34	0.11	0.51	0.08	+1.04	–	–	–	–
Direct-to-Indirect w/ history	0.34	0.16	0.53	0.11	+3.29	–	–	–	–
Direct-to-Indirect w/o history	0.32	0.13	0.49	0.14	+3.94	–	–	–	–
CNN/Daily Mail	0.32	0.15	0.54	0.18	–	+0.95	+0.56	+0.95	+1.00
XSum	0.33	0.12	0.58	0.13	–	+0.23	+0.02	+0.16	+0.18

表 5 提案手法に基づくモデルの出力例 (間接直接発話変換タスク)

正の語彙制約	**one**, **better**, **than**, No, Thanks, you, than
負の語彙制約	**more**, **services**, **expected**, **from**, so, will, further, do, needed, is, for
入力文	That is what I want from you. Thanks a lot.
参照文	No **one** can serve me **better than** you. Thanks.
beam-search	No **more services expected from** you. Thanks.
NEUROLOGIC* (正負)	No **one** can serve me **better than** you. Thanks.
正の語彙制約	*in*, *the*, East, area
負の語彙制約	*my*, *preferred*, stay, to, would, for, any, comfortable, go, it, there, choice, option
入力文	I need the one in the East area
参照文	East area is *my preferred* location
beam-search	East area is *my preferred* location
NEUROLOGIC* (正負)	Is there any one available *in the* East area?

生成文 (beam-search) の例を表 5 に示す．

1 つ目の例では，ビームサーチでは参照文にある “one”，“better” および “than” を含む文を生成できておらず，また出力された “more”，“services”，“expected” および “from” は提案手法では負の語彙制約としている．提案手法ではこれらをそれぞれ正の語彙制約，負の語彙制約とすることで，ベースラインを改善している．

2 つ目の例では，ビームサーチでは参照文にある “my” と “preferred” を正しく出力できている．一方提案手法では，これらを負の語彙制約としており，また参照文に存在しない “in” と “the” を誤って正の語彙制約としたことで，ビームサーチより悪化している．これらの例からも，正誤予測モデルの性能改善が最終的な生成文の品質向上には重要であることが分かる．

9. おわりに

本研究では，系列変換モデルの N ベスト出力に対する正誤予測に基づき作成した語彙制約をデコーダに適用し，生成文の品質を改善した．評価実験により，言い換え生成および要約タスクにおける提案手法の有効性を示した．

今後は，語彙制約の作成に用いる正誤予測モデルの性能を向上させることで，生成文の品質のさらなる改善を目指す．また，機械翻訳など他の言語生成タスクにも提案手法を適用し，生成文の評価および分析を行うことも今後の課題である．さらに，提案手法とモデルアンサンブルの融合も考えられる．提案手法を用いて複数モデルの出力文を効果的に組み合わせることで，少数のモデルを用いたモデルアンサンブルにおいても，生成文の品質を改善できると期待される．

謝辞

本研究は JSPS 科研費 JP21H03564 の助成を受けたものです．

参考文献

[1] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V. and Zettlemoyer, L.: BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension, *in Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 7871–7880 (2020).

[2] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W. and Liu, P. J.: Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer, *Journal of Machine Learning Research*, Vol. 21, No. 140, pp. 1–67 (2020).

[3] Chatterjee, R., Freitag, M., Negri, M. and Turchi, M.: Findings of the WMT 2020 Shared Task on Automatic Post-Editing, *in Proceedings of the Conference on Machine Translation (WMT)*, pp. 646–659 (2020).

[4] Correia, G. M. and Martins, A. F. T.: A Simple and Effective Approach to Automatic Post-Editing with Transfer Learning, *in Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 3050–3056 (2019).

[5] Chollampatt, S., Susanto, R. H., Tan, L. and Szymanska, E.: Can Automatic Post-Editing Improve NMT?, *in Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2736–2746 (2020).

[6] Lu, X., Welleck, S., West, P., Jiang, L., Kasai, J., Khashabi, D., Le Bras, R., Qin, L., Yu, Y., Zellers, R., Smith, N. A. and Choi, Y.: NeuroLogic A*esque Decoding: Constrained Text Generation with Lookahead Heuristics, *in Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 780–799 (2022).

[7] 高山隼矢，梶原智之，荒瀬由紀：対話における間接的応答と直接的応答からなる言い換えコーパスの構築と分析，自然言語処理，Vol. 29, No. 1, pp. 84–111 (2022).

[8] See, A., Liu, P. J. and Manning, C. D.: Get To The Point: Summarization with Pointer-Generator Networks, *in Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1073–1083 (2017).

[9] Hermann, K. M., Kociský, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M. and Blunsom, P.: Teaching Machines to Read and Comprehend, *in Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, pp. 1693–1701 (2015).

[10] Narayan, S., Cohen, S. B. and Lapata, M.: Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization, *in Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1797–1807 (2018).

[11] Ive, J., Specia, L., Szoc, S., Vanallemeersch, T., Van den Bogaert, J., Farah, E., Maroti, C., Ventura, A. and Khalilov, M.: A Post-Editing Dataset in the Legal Domain: Do we Underestimate Neural Machine Translation Quality?, *in Proceedings of the Language Resources and Evaluation Conference (LREC)*, pp. 3692–3697 (2020).

[12] Chatterjee, R., Federmann, C., Negri, M. and Turchi, M.: Findings of the WMT 2019 Shared Task on Automatic Post-Editing, *in Proceedings of the Conference on Machine Translation (WMT)*, pp. 11–28 (2019).

[13] Garmash, E. and Monz, C.: Ensemble Learning for Multi-Source Neural Machine Translation, *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pp. 1409–1418 (2016).

[14] Imamura, K. and Sumita, E.: Ensemble and Reranking: Using Multiple Models in the NICT-2 Neural Machine Translation System at WAT2017, *Proceedings of the 4th Workshop on Asian Translation (WAT2017)*, pp. 127–134 (2017).

[15] Chatterjee, R., Negri, M., Turchi, M., Federico, M., Specia, L. and Blain, F.: Guiding Neural Machine Translation Decoding with External Knowledge, *in Proceedings of the Conference on Machine Translation (WMT)*, pp. 157–168 (2017).

[16] Hokamp, C. and Liu, Q.: Lexically Constrained Decoding for Sequence Generation Using Grid Beam Search, *in Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1535–1546 (2017).

[17] Dehghan, M., Kumar, D. and Golab, L.: GRS: Combining Generation and Revision in Unsupervised Sentence Simplification, *in Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 949–960 (2022).

[18] Zetsu, T., Kajiwara, T. and Arase, Y.: Lexically Constrained Decoding with Edit Operation Prediction for Controllable Text Simplification, *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pp. 147–153 (2022).

[19] Kajiwara, T.: Negative Lexically Constrained Decoding for Paraphrase Generation, *in Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 6047–6052 (2019).

[20] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. and Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach, *arXiv preprint arXiv:1907.11692* (2019).

[21] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q. and Rush, A.: Transformers: State-of-the-Art Natural Language Processing, *in Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 38–45 (2020).

[22] Budzianowski, P., Wen, T.-H., Tseng, B.-H., Casanueva, I., Ultes, S., Ramadan, O. and Gašić, M.: MultiWOZ - A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling, *in Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5016–5026 (2018).

[23] Eric, M., Goel, R., Paul, S., Sethi, A., Agarwal, S., Gao, S., Kumar, A., Goyal, A., Ku, P. and Hakkani-Tur, D.: MultiWOZ 2.1: A Consolidated Multi-Domain Dialogue Dataset with State Corrections and State Tracking Baselines, *in Proceedings of the Language Resources and Evaluation Conference (LREC)*, pp. 422–428 (2020).

[24] Papineni, K., Roukos, S., Ward, T. and Zhu, W.-J.: Bleu: a method for automatic evaluation of machine translation, *in Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 311–318 (2002).

[25] Koehn, P.: Statistical Significance Tests for Machine Translation Evaluation, *in Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 388–395 (2004).

[26] Lin, C.-Y.: ROUGE: A Package for Automatic Evaluation of Summaries, *Text Summarization Branches Out*, pp. 74–81 (2004).

[27] Riezler, S. and Maxwell, J. T.: On Some Pitfalls in Automatic Evaluation and Significance Testing for MT, *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pp. 57–64 (2005).