負の語彙制約に基づくニューラル言い換え生成

梶原 智之
大阪大学データビリティフロンティア機構
kajiwara@ids.osaka-u.ac.jp

1 はじめに

言い換え生成とは、入力文と意味的に等価な文を生成するタスクの総称である。これらの技術を用いることで、我々はテキストにおける意味以外の情報を自動的に制御することができる。代表的な言い換え生成としては、テキストの難易度を制御するテキスト平易化やスタイルを制御するスタイル変換などのサブタスクがある。これらの研究は、人々のコミュニケーションや言語学習を支援するだけでなく、他の自然言語処理アプリケーションの性能改善 [1] にも貢献する。

言い換え生成は、同一言語内の機械翻訳の問題として解くことができる。そのため、難易度 [2,3] やスタイル [4,5] の異なる言い換え文対が収集され、各サブタスクに特化したパラレルコーパスが構築されている。初期にはフレーズベース [2,4] や構文ベース [6,7] の統計的機械翻訳に基づく手法が考えられたが、ニューラル機械翻訳 [8] の成功を受け、近年では注意機構に基づく seq2seq モデル [9,10] が提案されている。

機械翻訳では、入力文に出現する全ての単語を目的言語の単語に書き換える。しかし、言い換え生成では全ての単語を書き換える必要はない。ある基準が与えられた場合に、入力文中の基準を満たさない表現を検出し、それらを書き換える。例えばテキスト平易化では、基準はテキストの難易度であり、難解な表現を平易な同義表現に書き換える。入力文の一部のみを書き換えれば良いというタスクの特徴のために、機械翻訳のアプローチを用いる言い換え生成モデルは、しばしば保守的に振る舞い、書き換えるべき表現をそのまま出力文にコピーしてしまう [9,10]。この問題を解決するために、我々はまず入力文に対して言い換え箇所の検出を行い、負の語彙制約によってそれらの表現をそのまま出力することを避ける言い換え生成を行う。

本研究では、英語のテキスト平易化 [3] およびスタイル変換 [5] の 2 つのタスクで実験を行った。実験の結果、提案手法は積極的な書き換えを促進し、両方のタスクで言い換え生成の品質を改善することができた。

	BLEU(S,T)	BLEU(T,R)	SARI
Source	100.0	21.37	2.75
Reference	18.30	100.0	70.25
PBMT-R	75.60	18.19	15.77
Hybrid	25.64	14.46	30.00
EncDecA	52.91	21.70	24.12
Dress	39.96	23.26	27.37
Dress-LS	43.00	24.33	26.63

表 1: テキスト平易化における先行研究の性能評価

	BLEU on E&M		BLEU on F&R	
	(S,T)	(T,R)	(S,T)	(T,R)
Source	100.0	49.09	100.0	51.03
Reference	27.08	100.0	28.37	100.0
RBMT	71.11	61.78	71.43	67.25
PBMT	51.62	68.22	51.56	72.94
NMT-Comb	54.16	68.41	54.66	74.22
BiFT-Ens	55.86	71.36	59.48	74.49
MultiTask	54.55	72.13	58.11	75.37

表 2: スタイル変換における先行研究の性能評価

2 先行研究

表 1 に、テキスト平易化の先行研究とその性能を示す。PBMT-R [11] は、フレーズベースの統計的機械翻訳に基づく手法であり、n-best 出力を入力文との非類似度によってリランキングする。Hybrid [12] は、文分割モデルと言い換えモデルを統合した手法であり、まず入力文に文分割と削除を施し、続いてフレーズベースの統計的機械翻訳を用いて平易な表現へ言い換える。EncDecA [9] は、注意機構に基づくニューラル機械翻訳を用いる手法であり、Dress および Dress-LS のベースとなる。Dress [9] は、強化学習に基づく手法

であり、EncDecA による事前学習に続いて、テキスト平易化のための評価尺度である SARI [7] などを用いてモデルを最適化する。Dress-LS [9] は、Dress において過去のモデル出力を考慮しない手法であり、単に流暢な出力ではなく、平易な言い換えを促進する。

表 2 に、スタイル変換の先行研究とその性能を示す。RBMT [5] は、ルールベースの機械翻訳に基づく手法である。PBMT [5] は、フレーズベースの統計的機械翻訳に基づく手法であり、パラレルコーパスとともに入力側のスタイルのコーパスを用いて自己訓練を行う。NMT-Comb [5] は、ニューラル機械翻訳に基づく手法であり、パラレルコーパスとともに入出力の各スタイルのコーパスから PBMT によって生成された疑似パラレルコーパスを用いて訓練を行う。BiFT-Ens [10] は、マルチリンガルのニューラル機械翻訳 [13] に基づくアンサンブル手法であり、formal と informal のスタイル間の両方向の言い換えを単一モデルで訓練する。MultiTask [10] は、スタイル変換と機械翻訳のマルチタスク学習に基づく手法であり、仏語から formal な英語および informal な英語への翻訳タスクを用いる。

表 1 および 2 に、英語のニュース記事を英語母語話者の子ども向けに平易化するための Newsela データセット [3] および informal な英語を formal に言い換えるための GYAFC データセット [5] において、入力文と出力文の間の BLEU [14] および出力文と参照文の間の BLEU を示す。GYAFC データセットには、Entertainment&Music (E&M) と Family&Relationships (F&R) の 2 つのドメインが含まれる。どのタスクにおいても、モデル出力は参照文よりも BLEU(S,T) が高い。これは、先行研究の各モデルが入力文の大部分を出力文にコピーする保守的な言い換えを行っていることを意味する。この問題を解決するために、本研究では入力文から言い換え箇所を検出し、それらの表現を出力文にコピーしない積極的な言い換えを行う。

3 提案手法

我々の提案手法では、まず所与の入力文に対して言い換え箇所の検出（3.1 節）を行う。次に、学習済みの言い換え生成モデルを用いて入力文を言い換えるが、ビームサーチに負の語彙制約（3.2 節）を加え、検出された表現を含まない言い換え文を選択する。本手法はビームサーチのみを変更するため、任意の言い換え生成モデルに適用でき、モデルの再訓練も必要ない。

	Train	Dev	Test
Newsela	94,208	1,129	1,077
GYAFC-E&M	52,595	2,877	1,416
GYAFC-F&R	51,967	2,788	1,332

表 3: データセットの文対数

3.1 言い換え箇所の検出

訓練用パラレルコーパスを用いて自己相互情報量によって言い換え箇所を検出するための辞書を構築する。

$$\mathrm{PMI}(w,s) = \log \frac{p(w,s)}{p(w)p(s)} = \log \frac{p(w|s)}{p(w)} \quad (1)$$

言い換え時には、所与の入力文に対して、入力側のスタイル s（難解／informal）との自己相互情報量が閾値 θ を超える単語 w を言い換え箇所として検出する。

3.2 負の語彙制約に基づく言い換え生成

語彙制約に基づくテキスト生成 [15–17] は、ビームサーチに制約を加えることで、出力テキストに所望の単語を出現させることを可能にする。この技術は、機械翻訳の後編集 [15] や所与の画像タグを用いる画像キャプション生成 [16] において有効性が示されている。

言い換え生成においては、出力文に出現させるべき単語が事前に与えられるという状況は想定しにくい。そのため、機械翻訳の後編集や画像キャプション生成で用いられた正の語彙制約をそのまま本タスクに適用することはできない。一方で、特定の単語を出力文に出現させないという負の語彙制約は、言い換え生成のために有望である。なぜなら、例えばテキスト平易化は、入力文に出現するような難解な表現を使わずに言い換え文を生成するタスクだからである。

本研究では、最も高速に動作する語彙制約アルゴリズムである動的ビーム割当 [17] を用いて、ビームサーチに負の語彙制約を導入する。負の語彙制約では、指定された単語を含む候補をビームサーチの際に除外する。これによって、生成される言い換え文には、前節で検出された単語が出現しなくなる。

4 評価実験

表 3 に示すデータセットを用いて、テキスト平易化およびスタイル変換の評価実験を行う。

	Newsela					GYAFC-E&M				GYAFC-F&R			
	Add	Keep	Del	BLEU	SARI	Add	Keep	Del	BLEU	Add	Keep	Del	BLEU
Source	0.0	60.3	0.0	21.4	2.8	0.0	85.4	0.0	49.1	0.0	85.8	0.0	51.0
Reference	100	100	100	100	70.3	57.2	82.9	61.2	100	56.5	82.7	60.6	100
Dress-LS	2.4	60.7	44.9	24.3	**26.6**								
MultiTask						33.0	90.0	59.5	**72.1**	32.9	90.3	61.4	75.4
RNN-Base	1.8	60.8	22.3	24.1	17.4	31.9	90.0	57.5	71.2	32.9	90.5	61.1	74.7
RNN-PMI	2.8	61.1	36.5	**24.7**	22.8	33.5	90.0	59.9	71.7	34.3	90.9	63.1	75.9
RNN-Oracle	10.4	82.9	89.9	36.4	40.0	34.8	92.7	72.4	75.2	35.7	93.2	74.6	79.3
CNN-Base	2.2	60.0	39.6	23.5	24.5	33.5	89.9	59.9	71.0	33.6	91.0	63.2	75.2
CNN-PMI	2.3	59.9	44.7	23.6	26.3	33.9	89.6	60.6	70.7	34.0	90.9	63.6	75.5
CNN-Oracle	8.0	77.9	87.0	30.6	39.3	35.0	91.5	70.9	73.5	35.4	92.6	73.5	78.3
SAN-Base	1.8	60.9	23.8	24.0	17.8	34.4	90.0	59.9	71.8	34.5	91.1	63.2	76.7
SAN-PMI	2.5	61.3	38.0	24.6	23.3	35.2	90.0	61.2	**72.1**	35.3	91.1	64.0	**77.0**
SAN-Oracle	10.1	82.0	89.4	35.9	39.9	36.6	92.4	71.4	75.1	36.6	92.9	73.7	79.8

表 4: テキスト平易化（難解→平易）およびスタイル変換（informal → formal）における言い換え生成の評価

4.1 実験設定

テキスト平易化のために、先行研究 [9] と同じ設定で分割および tokenize した Newsela データセット [3] を用いる。また、スタイル変換のために、Moses ツールキット[1]によって normalize および tokenize した GYAFC データセット [5] を用いる。いずれのタスクにおいても、Byte Pair Encoding [18] を用いてトークン数を 16,000 に制限する。GYAFC では、informal なスタイルから formal なスタイルへの言い換え時にのみ、人手評価と自動評価に相関がある [5] ことが報告されているため、本研究ではこの設定でのみ実験を行う。

語彙制約のために、3.1 節で説明した自己相互情報量を用いて閾値 θ を超える単語を言い換え箇所として検出する。この閾値 θ は、開発データにおける出力文と参照文の間の BLEU を最大化する値を選択する。

言い換え生成モデルとして、我々は Sockeye ツールキット[2]を用いて Recurrent Neural Network (RNN)、Convolutional Neural Network (CNN) および Self-Attention Network (SAN) の各モデルを構築する。RNN モデルは、符号化器と複合化器の両方で 512 次元の 1 層 LSTM を利用し、MLP に基づく 512 次元の注意機構を用いる。CNN モデルは、符号化器と復号化器の両方で 512 次元の 8 層モデルを利用し、カーネルサイズは 3 とする。SAN モデルは、符号化器と復号化器の両方で 512 次元の 6 層モデルを利用し、単一の注意ヘッドを用いる。埋め込み層は全てのモデルで 512 次元とし、符号化器と復号化器および出力層で重みを共有する。正則化には埋め込み層および隠れ層にて確率 0.2 で dropout を適用し、さらに layer-normalization および label-smoothing を使用する。最適化には adam を利用し、バッチサイズを 4,096 トークンとして 1,000 更新ごとに開発データで perplexity を評価し、5 回改善が見られなくなったところで訓練を終了する。

評価尺度には主に BLEU [14] を使用し、テキスト平易化の実験においては SARI [7] も併用する。また、モデルの詳細な比較のために、入力文に対して追加された単語 (Add)、入出力間で保持された単語 (Keep)、入力文から削除された単語 (Del) の F 値[3]を評価する。

比較手法には、上述の詳細な評価のために、テストデータにおけるモデル出力が公開されている先行研究を用いる。特に、表 1 および 2 に示した BLEU が最も高い、テキスト平易化における Dress-LS [9] およびスタイル変換における MultiTask [10] を我々の提案手法と比較する。比較手法に合わせて、スタイル変換においては我々の提案手法も両方向のドメイン混合アンサンブルモデルを採用する。つまり、各ドメインの訓練データを結合し、文頭に {to-formal} または {to-informal} のラベルを追加した上で、シードの異なる 4 つのモデルをアンサンブルする。

言い換え箇所を理想的に検出できた場合の提案手法の有効性を調査するため、オラクル検出も実験する。オラクル検出では、入力文に出現する単語のうち参照文に出現しない全ての単語を語彙制約として利用する。

[1]https://github.com/moses-smt/mosesdecoder
[2]https://github.com/awslabs/sockeye

[3]GYAFC データセットのテストデータはマルチリファレンスのため、Add、Keep、Del の各 F 値は 100.0 にならない。

4.2 実験結果

表 4 に、実験結果を示す。3 つのモデル構造について 3 つのデータセットで評価したところ、GYAFC-E&M における CNN を除く 9 つ中 8 つの設定では、提案手法である PMI が BLEU および SARI で語彙制約を利用しない Base を改善した。さらに、Newsela および GYAFC-F&R では、比較手法を凌ぐ性能を達成できた。Dress-LS が強化学習を用いて評価尺度を直接最適化していることや、MultiTask が大規模な外部データに頼っていることを考慮すると、シンプルな seq2seq モデルに基づく提案手法の有効性は高い。

モデル出力の詳細な分析の結果、全てのモデル構造において、PMI は Add および Del の観点で Base を常に改善している。これは、負の語彙制約の導入によって、モデルが入力文中の難解な表現や informal な表現を積極的に書き換えるようになったことを意味する。

一方で、Oracle と比較すると、PMI は Del の観点で改善の余地が大きい。これは、難解な表現や informal な表現の検出性能が十分でないことを意味する。本研究では訓練用パラレルコーパスを用いて言い換え箇所を検出したが、今後はパラレルコーパスに限らず、より大規模なデータを用いて言い換え箇所を検出したい。

また、Oracle も含めて、Add の観点ではモデル出力と参照文の間に大きな隔たりがある。これは、平易な表現や formal な表現など、目的のスタイルに関する情報をモデルが十分に獲得できていないことを意味する。やはり今後は、パラレルコーパスに限らず目的のスタイルに属するより大規模なコーパスや辞書を活用し、言い換え生成モデルの性能を改善したい。

5 おわりに

本研究では、言い換え生成モデルの保守的な振る舞いを改善するために、言い換え箇所の検出およびビームサーチへの負の語彙制約の導入を提案した。英語のテキスト平易化およびスタイル変換における実験結果は、提案手法が RNN ／ CNN ／ SAN の全てのモデル構造について、多くの場合に言い換え生成の品質を改善できることを示した。我々の提案手法は、入力文に出現する難解／ informal な表現を削除し、言い換え文へ平易／ formal な表現を追加することを促進する。

謝辞

本研究は JST（ACT-I、課題番号: JPMJPR18UB）の支援を受けたものである。

参考文献

[1] Sanja Štajner and Maja Popović. Can Text Simplification Help Machine Translation? *Baltic Journal of Modern Computing*, Vol. 4, No. 2, pp. 230–242, 2016.

[2] William Coster and David Kauchak. Simple English Wikipedia: A New Text Simplification Task. In *Proc. of ACL*, pp. 665–669, 2011.

[3] Wei Xu, Chris Callison-Burch, and Courtney Napoles. Problems in Current Text Simplification Research: New Data Can Help. *TACL*, Vol. 3, pp. 283–297, 2015.

[4] Wei Xu, Alan Ritter, William B. Dolan, Ralph Grishman, and Colin Cherry. Paraphrasing for Style. In *Proc. of COLING*, pp. 2899–2914, 2012.

[5] Sudha Rao and Joel Tetreault. Dear Sir or Madam, May I Introduce the GYAFC Dataset: Corpus, Benchmarks and Metrics for Formality Style Transfer. In *Proc. of NAACL*, pp. 129–140, 2018.

[6] Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych. A Monolingual Tree-based Translation Model for Sentence Simplification. In *Proc. of COLING*, pp. 1353–1361, 2010.

[7] Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. Optimizing Statistical Machine Translation for Text Simplification. *TACL*, Vol. 4, pp. 401–415, 2016.

[8] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural Machine Translation by Jointly Learning to Align and Translate. In *Proc. of ICLR*, pp. 1–15, 2015.

[9] Xingxing Zhang and Mirella Lapata. Sentence Simplification with Deep Reinforcement Learning. In *Proc. of EMNLP*, pp. 584–594, 2017.

[10] Xing Niu, Sudha Rao, and Marine Carpuat. Multi-Task Neural Models for Translating Between Styles Within and Across Languages. In *Proc. of COLING*, pp. 1008–1021, 2018.

[11] Sander Wubben, Antal van den Bosch, and Emiel Krahmer. Sentence Simplification by Monolingual Machine Translation. In *Proc. of ACL*, pp. 1015–1024, 2012.

[12] Shashi Narayan and Claire Gardent. Hybrid Simplification using Deep Semantics and Machine Translation. In *Proc. of ACL*, pp. 435–445, 2014.

[13] Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Vigas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. *TACL*, Vol. 5, pp. 339–351, 2017.

[14] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proc. of ACL*, pp. 311–318, 2002.

[15] Chris Hokamp and Qun Liu. Lexically Constrained Decoding for Sequence Generation Using Grid Beam Search. In *Proc. of ACL*, pp. 1535–1546, 2017.

[16] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Guided Open Vocabulary Image Captioning with Constrained Beam Search. In *Proc. of EMNLP*, pp. 936–945, 2017.

[17] Matt Post and David Vilar. Fast Lexically Constrained Decoding with Dynamic Beam Allocation for Neural Machine Translation. In *Proc. of NAACL*, pp. 1314–1324, 2018.

[18] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural Machine Translation of Rare Words with Subword Units. In *Proc. of ACL*, pp. 1715–1725, 2016.