公的文書平易化における出力制御の効果

丸山 拓海　　山本 和英
長岡技術科学大学
{maruyama, yamamoto}@jnlp.org

1 はじめに

テキスト平易化とは、ある文書を読者にとって理解しやすい文書に書き換えるタスクである。この技術によって、子供や非母語話者、障がい者などをはじめとする幅広い読者に多くのテキストリソースを提供することが可能となる。 読者ごとに読解力や理解語彙は異なるため、その読者自身にとって理解しやすい文を出力することがテキスト平易化の目的である。 すなわち、同一の入力文であったとしても、読者のニーズに合わせて出力を制御する必要がある。 本研究では、学校や病院・市役所等の公共施設で配布されているような、重要性の高い公的文書を対象に、出力文長、編集の度合い・平易化レベルを制御可能なモデルを利用し、その効果を検証する。

2 関連研究

出力を制御する方法は、入力文に情報を付与する手法 [1, 2] と出力側に何らかの制約を与える手法 [3] の 2 種類に大別される。 本研究では、前者のラベルを付与する方法により出力制御を試みる。 ラベルを挿入することで出力制御を行う研究は様々なタスクで行われている。 例えば、翻訳タスクでは、目的言語の表すラベルを入力言語の文頭に挿入し、単一モデルによる多言語翻訳を実現している [4]。 また、出力言語を制御するだけでなく、挿入するラベルを置き換えることにより、翻訳出力の難易度やスタイル、出力文長などの制御も行うことが可能である。 ここでは、平易化レベルや編集の度合い、出力文長などに注目した制御を試みる。

3 公的文書書き換えコーパス

本研究では、Moku らが利用した平易化コーパス [5] と同じコーパスを利用する。これは「やさしい日本語」のプロジェクトで作成されたものであり、約 40 名の日本語教師が、市役所や病院、学校等の公共施設で配布される公的文書を「やさしい日本語」に書き換えたものである。 このコーパスは原文である公的文書約 1,100 文書と共にその逐語訳、意訳、要約という 3 段階の平易文を含む対訳コーパスである。 それぞれの翻訳の位置付けは下記の通りである。

- **逐語訳 (literal translation)**: 日本語文の難解な語彙や句をやさしい表現に書き換えたもの。
- **意訳 (free translation)**: 文意等を損なわないように可能な限り、やさしい表現に書き換えたもの。
- **要約 (summary)**: 可能な限り文を平易化したもの。

これらは、一定の文法基準 [1] と旧日本語能力試験 2 級レベルの語彙のみに可能な限り制限されるよう、書き換え作業を行なっている。 コーパスにおける「やさしい」の基準は各日本語教師の主観である。

4 手法

入力文に対して、出力文を制御しうるラベルを文頭または文末に付与したデータで学習を行う。 ここでいうラベルとは、具体的には次の 2 つのことを指す。

- **ドメインラベル**: どの平易化レベルに変換するのかを表すラベル。例えば、逐語訳に変換するのであれば [2literal]、意訳に変換するのであれば、[2free] などのラベルを挿入する。
- **編集操作ラベル**: どのような操作をどの程度行うかを表すラベル。 たとえば、文長を入力文の 80%程度の文を出力したいのであれば、$[cr_{0.8}]$ といったラベルを挿入する。

[1] http://human.cc.hirosaki-u.ac.jp/kokugo/EJyasashisa-kijyun.html

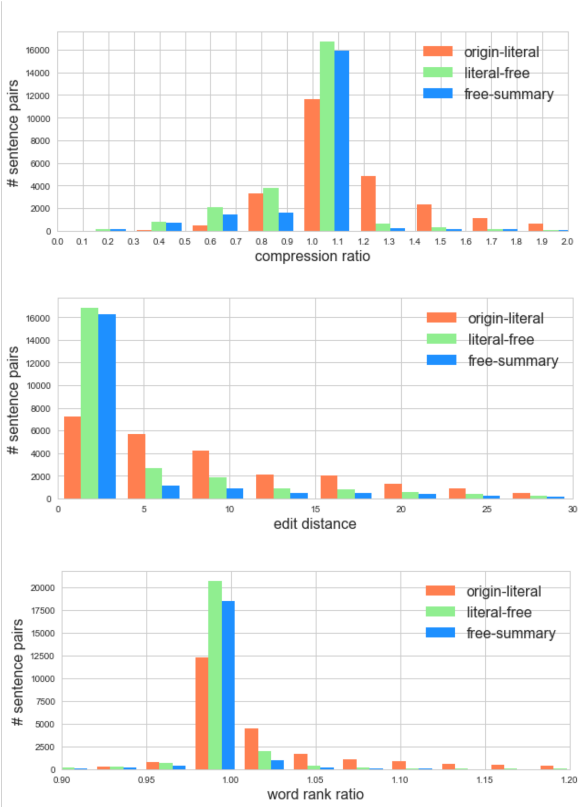

図 1: 学習データにおける各特徴量の分布

編集操作ラベルにおいて、着目した特徴量は次の 3 つである。 各特徴量ごとに 10 段階のラベルを利用する。

- **圧縮率 (compression ratio)**: (出力文の単語数) / (入力文の単語数)。 0.0 から 0.2 ごとに 1 つのラベルを割り当て。
- **編集距離 (edit distance)**: レーベンシュタイン距離。 0.0 から 3.0 ごとに 1 つのラベルを割り当て。
- **Word rank ratio**: 原文及び平易文に含まれる単語の頻度における第三四分位数の比。 単語の頻度は BCCWJ より獲得する。 0.9 から 0.02 ごとに 1 つのラベルを割り当て。

図 1 に、学習データにおける各特徴量の分布を示す。圧縮率の分布 (図 1 上段) に注目すると、原文-逐語訳においては文が長くなっていることが分かる。 これは、原文に含まれる難解な表現を平易な言葉で説明しているためであると考えられる。 一方で、逐語訳-意訳, 意訳-要約の変換では、文長を短くする変換が多く、各平易化タスクにおいて異なった編集を行なっていることが分かる。

また、編集距離や word rank の分布 (図 1 中段・下段) に注目すると、原文-逐語訳変換では編集距離や word rank ratio が大きいペアが多く含まれているのに対し、逐語訳-意訳変換や意訳-要約変換では少ない。 これらより、主要な換言操作は逐語訳への変換時に、高度な要約処理や文圧縮操作は意訳・要約への変換時に、というように段階的に平易化処理を行なっていることが分かる。

5 実験

5.1 実験設定

学習データには、公的文書書き換えコーパスに含まれる全平易化データ 134,523 文ペアを利用した。 4 単語未満または 128 単語を超える文を含むペアは除外した。 テスト時における編集操作ラベルは、入力文と参照文を比較して得られたラベルを付与する。 しかしながら、この設定は理想的なラベルを付与した状態であり、実際利用する場合には、不可能な設定である。 そこで、全ての入力文に対して、同一のラベルを付与した結果についても評価を行う。

平易化を行う系列変換モデルとして、16 個の self-attention heads を備えた 6 層の Transformer を使用した。 word embedding layer の次元は 512 とし、feed forward network の次元は 2048 とした。 最適化には、Adam を利用した。 学習率は 10^{-4} とし、100,000 iteration で学習を終了させた。

評価には、BLEU とテキスト平易化で広く利用されている SARI[6] を用いる。 BLEU は文分割を含まない平易化タスクにおいては流暢性と妥当性に対して正の相関があると報告されている。 また、SARI は平易化において広く用いられている評価尺度であり、文の平易さと正の相関があると報告されている。

5.2 テキスト平易化に関する評価

表 1 に評価結果を示す。 ドメインラベル・編集ラベルどちらにおいても、文末にラベルを挿入した方が文頭に挿入した場合の結果よりも全体的にスコアが高かったため、表 1 には文末挿入の結果のみを示している。

ここで、“Single” はテストデータとドメインが一致するトレーニングデータのみで学習したモデルである。例えば、原文-逐語訳変換を評価する際には、原文-逐語訳ペアのみを用いて学習を行う。 一方、“Joint” は、全

表 1: ラベルの文末挿入による可読性制御の結果

挿入ラベル	原文-逐語訳		原文-意訳		原文-要約		逐語訳-意訳		逐語訳-要約		意訳-要約	
	BLEU	SARI	BLEU	SARI	BLEU	SARI	BLEU	SARI	BLEU	SARI	BLEU	SARI
ラベルなし												
Single	39.97	49.13	28.42	44.21	24.53	42.82	66.70	34.12	52.89	32.86	**73.82**	32.83
Joint	39.20	48.45	37.47	49.16	35.96	48.45	59.94	32.40	56.51	33.07	63.90	30.17
ドメイン	39.81	49.24	36.15	48.78	33.20	46.89	57.76	32.19	52.76	32.23	60.32	29.54
編集操作 (best)												
圧縮率	41.39	48.98	37.47	48.58	35.22	47.36	64.99	32.59	60.22	32.87	70.18	31.21
編集距離	39.99	47.22	35.35	45.91	33.54	44.68	68.24	32.67	**63.10**	32.64	73.64	31.85
word rank	38.50	46.81	36.05	46.86	34.16	45.98	61.57	32.59	58.57	33.82	67.13	31.45
編集操作 (gold)												
圧縮率	**44.52**	**51.06**	**41.61**	**51.03**	**39.61**	49.88	65.18	34.38	60.77	35.21	70.35	32.56
編集距離	44.36	50.91	41.01	50.94	38.98	**50.00**	**67.03**	**35.15**	61.85	**35.42**	72.96	**33.11**
word rank	40.27	48.08	38.32	49.72	36.88	49.19	58.50	32.57	56.12	33.78	63.20	30.37

て平易化ペアを利用して学習したモデルである。 これらには、ラベルの付与は行なっていない。 編集操作ラベルの結果においては、各特徴量それぞれ 10 段階のラベルを付与して出力させた結果のうち、全平易化タスクにおける BLEU・SARI の平均が最も高かったラベルの結果を “編集操作 (best)” とし、参照文を考慮した上で各文において適切なラベルを付与した結果を “編集操作 (gold)” とした。

ドメインラベルの結果に注目すると、全体的に改善はみられなかった。 今回のドメインラベルは 3 種類と非常に荒い粒度のラベルであり、モデルに与える情報としては少なく、ノイズとなっている可能性がある。

編集操作ラベル (gold) の結果では、圧縮率ラベル、編集距離ラベルにより、“Joint” の結果に比べ大きく性能を改善している。 圧縮率ラベルにおいては、原文から逐語訳・意訳・要約への変換それぞれにおいて高いスコアを有しており、同一の入力に対して目的の平易化度合いに応じて出力を制御できていることが分かる。また、編集距離ラベルでは、逐語訳-意訳変換をはじめとする編集の少ない平易化ペアにおいて良い結果を残している。 これは、書き換えを必要としない変換の場合に、編集距離ラベルにより、モデル側にコピー操作を行うよう適切に指示できているからであると考えられる。

5.3 出力制御に関する分析

ここでは、各特徴量を 3 段階のラベルによって制御した際の出力への影響を分析する。 各編集ラベルにおける出力への影響を図 2 に示す。

圧縮率制御の結果に注目すると、付与するラベルを大きな圧縮率を表すものに変化させるにしたがって、圧縮率の分布も全体的に右へとシフトしていくことから、ラベルの目的とした制御が適切に行えていることが分かる (図 2 上段左)。

次に、編集距離制御の結果に注目すると、圧縮率ラベルと同様に、ラベルの目的とした制御が適切に行えていることが分かる (図 2 中段中)。 また、編集距離の制御に伴って、圧縮率及び word rank も変化している。 圧縮率では、編集距離ラベルを大きくするにつれて文が長くなる傾向にあり (図 2 中段左)、word rank ratio では、編集距離ラベルを大きくするにつれて、より頻度の高い単語を出力するようになることが分かる (図 2 中段右)。 このような傾向は、多くの編集を行うようモデルに指示した結果、頻度の高い平易な単語を繰り返して出力しているためであると考えられる。

6 まとめ

公的文書書き換えコーパスの性質を利用したドメインラベルと、平易化における編集操作に着目した編集

図 2: 各編集ラベルにおける出力への影響

操作ラベルの 2 種類を利用しテキスト平易化における出力制御を行った。 ドメインラベルでは性能の改善には至らなかったが、編集操作ラベルでは平易化性能を改善しつつ、圧縮率や編集距離に対して適切に出力を制御できることを示した。 今回、文レベルの特徴量をラベルとして導入したが、今後はより細かい粒度での特徴量の導入を検討したい。

謝辞

本研究は、平成 29-31 年学術研究助成基金助成金 挑戦的研究 (萌芽) 課題番号 17K18481 の助成を受けています。

参考文献

[1] Carolina Scarton and Lucia Specia. Learning simplifications for specific target audiences. In *ACL*, Vol. 2, pp. 712–718, Melbourne, Australia, July 2018.

[2] Louis Martin, Benoît Sagot, Éric de la Clergerie, and Antoine Bordes. Controllable sentence simplification. *arXiv preprint arXiv:1910.02677*, 2019.

[3] Daiki Nishihara, Tomoyuki Kajiwara, and Yuki Arase. Controllable text simplification with lexical constraint loss. In *ACL: Student Research Workshop*, pp. 260–266, Florence, Italy, July 2019.

[4] Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, 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.

[5] Manami Moku, Kazuhide Yamamoto, and Ai Makabi. Automatic easy japanese translation for information accessibility of foreigners. In *Proceedings of the Workshop on Speech and Language Processing Tools in Education*, pp. 85–90, 12 2012.

[6] 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.