ニューステキストの要約及び平易化

菅井 内音　　西川 仁　　徳永 健伸
東京工業大学 情報理工学院
sugai.n.ab@m.titech.ac.jp {hitoshi,take}@c.titech.ac.jp

1 はじめに

近年, 自動要約及び平易化といった, テキストの読解を支援する技術への需要が高まっている. 長いテキストから重要箇所を抽出し短くまとめる「要約」は読み手の迅速な内容把握を可能にする一方, 専門用語などの難解な表現に対し削除及び易しい表現への置換を行う「平易化」は外国人や子供など語彙知識が不足している読み手の読解を補助する.

本研究では, 単一ニューステキストに対し自動要約及び平易化の両方を行い, より多くの読み手にとって理解しやすい, 短く平易なテキストの生成を目指す. 対象テキストは NHK NEWS WEB[1](以下, NNW) の記事と, その記事に対し構成変更による情報削減と平易な書き換えを行った NHK NEWS WEB EASY[2](以下, NNWE) の記事の対とする. 我々はこの目的を達成するための過程を, 記事全体に対する処理 (以下, 記事処理) と各文に対する処理 (以下, 文処理) の 2 つに分け, それぞれ別々にモデルを実装する. 前者は記事から重要文を抽出する「要約」側のタスクであるが, 後者は文分割や短縮, 書き換えといった「要約」及び「平易化」の両方に属するタスクである. ただし, 本研究では文処理に関する諸操作を「平易な記事には不要である, 難解な表現に対する操作」と位置付け, 平易化タスクにおいて広く用いられている統計的機械翻訳により文処理モデルを実装する. モデルの学習には, NNW-NNWE の記事対から記事処理及び文処理のそれぞれ用に構築した擬似パラレルコーパスを用いる. コーパス構築の際は, 単語分散表現のアライメントに基づき文間の類似度を計算することによる, NNW 記事中の文-NNWE 記事中の文の自動対応づけの結果を利用する. また, 生成実験の際は, 各モデルの適用順序を変えた場合や, 文処理モデルの学習に擬似パラレルコーパスだけでなく既存の大規模平易化コーパスを組み合わせて用いた場合の差異についても検証する.

[1] https://www3.nhk.or.jp/news/
[2] https://www3.nhk.or.jp/news/easy/

2 関連研究

2.1 NHK NEWS WEB EASY

NNWE は, NNW に掲載されている記事を「やさしい日本語[3]」の考え方により書き直した, 小・中学生や外国人向けの WEB ニュースサイトである. NNWE は平日に 4-5 記事程度ニュースを掲載しており, 各記事の下部には書き換え元となった NNW の記事へのリンクがある. 田中ら [12] によれば, NNW から NNWE への書き換えは, 日本語教師 (難解な表現の平易化) と記者 (記事の構成変更・要約) による人手での共同作業により行われている.

熊野ら [10] は, NNWE の記事制作過程で作成された日本語教師による平易な書き換え事例をパラレルコーパスとして利用し, 統計的機械翻訳によるやさしい日本語への自動変換を行っている. ただし, この書き換え事例のコーパスは非公開で利用できないため, 本研究では NNW-NNWE の記事対から記事処理及び文処理用のパラレルコーパスを擬似的に構築する.

2.2 やさしい日本語コーパス

前述したように, NNWE は「やさしい日本語」の考え方により NNW の記事を書き換えて作成される. 以下では「やさしい日本語」の観点で構築された言語資源について述べる.

Maruyama and Yamamoto [3] は, 田中コーパス [6] から作られた日英機械翻訳の対訳コーパス[4] 全 50000 文の日本語部分を原テキストとし, それを人手によりやさしい日本語に書き換えることで「やさしい日本語コーパス」[5]を構築した. また, Katsuta and Yamamoto [1] は, クラウドソーシングにより田中コーパスから新たに 35000 文をやさしい日本語へ書き換えることで「やさしい日本語拡張コーパス」[6]を構築した.

[3] http://human.cc.hirosaki-u.ac.jp/kokugo/EJ1a.htm
[4] https://github.com/odashi/small_parallel_enja
[5] http://www.jnlp.org/SNOW/T15
[6] http://www.jnlp.org/SNOW/T23

2.3 平易化用擬似パラレルコーパスの構築

梶原ら [9] は, word2vec の単語分散表現のアライメントに基づく文間の類似度推定手法として Song and Roth [5] の提案した Maximum Alignment を用い類似文同士を対応づけることで, English Wikipedia から平易化用の擬似パラレルコーパスを構築した.

本研究でも NNW-NNWE の記事対中で Maximum Alignment により文同士の対応づけを行うが, 対応づけの結果を記事処理及び文処理用の両方の擬似パラレルコーパス構築に利用する.

2.4 読み手に合わせた要約生成

永塚ら [8] は, 英語で書かれた複数のニュース記事を単一記事へ要約し, 複数の段階で語彙を平易化してそれぞれ公開している Web サイト「Breaking News English」[7]から複数文書要約コーパス「Breaking News English Corpus」を構築した. このコーパスにより, 「読み手に合わせて平易化された要約生成」の実現が期待されるが, 実際にこのコーパスを用いた実験の例は未だ報告されていない.

本研究では日本語の単一文書に対して要約及び平易化の両方を試みる. その過程で, 記事処理及び文処理用で別々に擬似パラレルコーパスを構築し, また実際に適用することでその効果を検証する.

3 擬似パラレルコーパスの構築

まず, Web クローリングにより, 2018 年 7 月 14 日から 2019 年 10 月 31 日までに発行された NNWE の記事と, 各記事の書き換え元となっている NNW の記事の本文を 1461 対収集した[8]. 収集後, NNW の記事中に存在するサブタイトルを削除した. 次に, 収集した各記事を文分割し, MeCab[9]により各文を単語分割した. その際, 全角数字は全て半角に修正した. その後, 2.3 節で述べた Maximum Alignment により, NNWE の記事中の各文に対して書き換え元の NNW の記事中の各文との類似度を計算し, 最も類似度の高い文同士を対応づけた.

対応づけの結果から, NNW の記事中のある 1 文が NNWE の記事中の複数の文と対応する場合のみ, 対応している NNWE の複数の文を先頭から順に結合することで 1 文とみなし, NNW の通常文-NNWE の平易な文の文対からなる文処理用の擬似パラレルコーパスを構築した. また, NNW の記事から, NNWE の記事中のどの文とも対応づいていない文を不要文とみなし削除したものを擬似要約とすることで, NNW の記事-擬似要約の記事対からなる記事処理用の擬似パラレルコーパスを構築した.

最後に, 記事単位で訓練データとテスト用データの比がおよそ 95:5 となるように分割を行った. 記事処理用の擬似パラレルコーパスについては 1461 記事対中 1388 記事対を訓練データ, 73 記事対をテストデータとした. 文処理用の擬似パラレルコーパス (以下, N コーパスも同義) については, 7321 文対中入出力の単語長が 100 以上である 141 文対を削除した上で, 7180 文対中 6808 文対を訓練データ (うち 334 文対はチューニング用), その他 372 文対をテストデータとした.

対応づけを自動で行ったため, N コーパスには意味が一致しない文対が存在する. また, 書き換えモデルを学習させるにはデータ量がやや少ないこともあり, そのままモデル学習に用いる場合主に文法性の保持に関する懸念がある. そこで, 2.2 節で述べた「やさしい日本語コーパス」と「やさしい日本語拡張コーパス」を統合したものを Y コーパスとし, N コーパスと組み合わせることでその問題を解決することをねらう. Y コーパスは合計 85000 文対と十分なデータ量を保有し, また人手で構築されているため文対間で意味は基本的に一致している. 以下, Y コーパスと N コーパスを組み合わせたものを Y+N コーパスと呼ぶ.

4 要約及び平易化の実験

4.1 実験方法

3 章で構築した擬似パラレルコーパスのテストデータである NNW の 73 記事に対し, 記事処理及び文処理それぞれのモデルを適用することで, 短く平易な記事を生成する実験を行った.

記事処理については, 西川ら [11] の提案した隠れ半マルコフモデルに基づく抽出型要約の学習器を実装し, 記事処理用の擬似パラレルコーパスの訓練データにより目標記事長を 300 としてパラメータを学習させたモデルを用いた. 文処理については, フレーズベースの統計的機械翻訳ツールである Moses ver.4.0[10]を使用し, 言語モデルと書き換えモデルを N コーパス及び Y+N コーパスの訓練データで学習させた 2 種類のモデルを用いた. また, 各モデルの適用順序について, 1) 記事

[7] https://breakingnewsenglish.com
[8] NNW のリンク取得に成功したもののみ収集した.
[9] https://github.com/neologd/mecab-ipadic-neologd
[10] http://www.statmt.org/moses/

処理モデルで重要文を抽出してから, 文処理モデルで各文の難解な表現に対し分割・短縮・書き換えを行う方法, 2)1 とは逆に, 文処理モデル → 記事処理モデルの順に適用する方法の 2 通りを行った.

文処理モデルの学習に用いるコーパスと, モデルの適用順序を変えることによる計 4 種類の生成記事に加え, ベースラインとして記事処理及び文処理の一方のみを行った記事と, 元記事 (NNW) のそれぞれについて, 実際の NNWE の記事を参照記事とした評価を行った. 評価指標には BLEU-1 [4], ROUGE-1 [2], SARI [7] を用いた. これらはそれぞれ機械翻訳, 要約, 平易化のタスクにおいて広く用いられている指標であるが, 本実験では BLEU 及び ROUGE で参照記事に対する生成記事の精度 (precision) 及び再現性 (recall) を計り, SARI で平易化の総合的な評価を行う.

4.2 実験結果

生成された各記事について, 上記の評価値及び記事長の中央値を算出した結果を表 1 に示す. 「文処理」列は, 文処理モデルの学習に用いたコーパスを示す. 「順序」列について, 例えば「記事 → 文」は記事処理モデル → 文処理モデルの順にモデルを適用したことを示す. 評価に平均値ではなく中央値を用いるのは, 文処理モデル → 記事処理モデルの順にモデルを適用する手法において記事長が目標の 50%以下となった記事が複数存在し, 各評価指標の平均値に大きな負の影響を与えたためである[11].

記事長については, 記事処理モデルのみを適用した記事よりも, 記事処理モデル → 文処理モデルの順にモデルを適用した記事の方が短くなっている. このことは, 文処理モデルの適用により想定していた文短縮が実際に行われたことを示している.

各評価値を比較すると, まず BLEU については, 記事処理及び文処理モデルの両方を適用した記事の全てが, 一方のみを適用した記事及び元記事を上回っている. また, ROUGE についても, 記事処理及び文処理モデルの両方を適用した記事の全てが記事処理モデルのみを適用した記事を上回っている. さらに, SARI についても記事処理及び文処理モデルの両方を適用した記事の全てが記事処理モデルのみを適用した記事及び元記事を上回っている. これらの事実は, 記事処理及び文処理モデルの両方を適用することで, 元記事や一方のモデルのみを適用した記事より参照記事 (NNWE) に近い, 短く平易な記事を生成できたことを示している.

記事処理及び文処理モデルの適用順序が同じ記事間で比較すると, 文処理モデルの学習に N コーパスを用いた記事の評価値が, Y+N コーパスを用いた記事を全体的に上回っている. この結果の主な原因として, Y コーパスと N コーパスの性質の違いが挙げられる. Y コーパスは短い会話文を中心として構成されており, 平易な書き換えの手法も単純なものが多いが, NNW-NNWE の記事対から擬似的に構築された N コーパスには長文も多く含まれ, 文対間での書き換えも表現の削除・具体化など複雑になる傾向がある. よって, Y コーパスを N コーパスと組み合わせることは, NNW-NNWE 間の平易な書き換えを反映することにはあまり寄与できなかった可能性が考えられる.

文処理	順序	記事長	BLEU	ROUGE	SARI
N	記事 → 文	262	0.509	0.487	0.515
Y+N	記事 → 文	278	0.497	0.469	0.479
N	文 → 記事	297	0.533	0.484	0.516
Y+N	文 → 記事	295	0.508	0.461	0.473
記事処理のみ		296	0.456	0.400	0.404
文処理 (N) のみ		498	0.428	0.672	0.600
文処理 (Y+N) のみ		529	0.397	0.667	0.552
元記事 (NNW)		572	0.356	0.552	0.130
参照 (NNWE)		301	1.000	1.000	1.000

表 1: 要約及び平易化の定量評価結果

4.3 生成例と考察

図 1 に, ズワイガニ販売のイベントについての NNW, NNWE の記事及び NNW に対し記事処理及び文処理モデルの両方を適用し生成した 3 種類の記事を示す. 生成されたどの記事においても, 「催し」→「イベント」 などの適切な書き換えが見られる一方, 書き換えミスも複数見られる.

先に記事処理モデルを適用した 2 記事 (2 段 1, 2 列目) を比較すると, N コーパスを文処理モデルの学習に用いた記事では, 「入ってを」など文が成り立たなくなってしまう致命的なミスが Y+N コーパスを用いた記事よりも多く見られる. Y+N コーパスで学習させた文処理モデルは, Y コーパスの文対で基本的な平易化事例を多く学習している分, N コーパスで懸念される文法性の問題をある程度解消できる可能性がこの例から示唆される.

次に, 文処理モデルの学習に Y+N コーパスを用いるが, モデルの適用順序が違う 2 記事 (2 段 2, 3 列目) を比較すると, 両者の内容はかなりの部分で一致しているが, 先に文処理モデルを適用した記事のみ, NNWE

[11] このような記事が生成されたこと自体が, 文処理モデル → 記事処理モデルの順に適用する手法の脆弱性を示唆しているとも考えられる.

元記事(NNW)		参照記事(NNWE)
...この催しは、地元の漁協などが毎年、開いているもので、石川県輪島市の港の近くの特設会場では、水産会社などがブースを並べてカニを販売しています。会場には、ズワイガニのオス約5000匹とメス約１万匹が並べられ、このうち、地元の漁協が出した店では、オスが１匹2000円から、メスが１匹600円で売られていました。 会場では、多くの観光客などが長い列を作って、お目当てのカニを買い求めていました。また、メスのカニが1匹丸ごと入った特製のみそ汁を味わったり、炭火焼きのコーナーで買ったばかりのカニを焼いたりして、冬の味覚を堪能していました。金沢市から訪れた小学2年生の男の子は...		...石川県輪島市の港で、たくさんの店が集まってズワイガニを売るイベントがありました。会場には、ズワイガニの雄が５０００匹と雌が１万匹ぐらい並んでいました。漁協の店では、雄は１匹２０００円以上、雌は１匹６００円で売っていました。店の前にはたくさんの人が並んで、どのカニがいいか選んでいました。カニが入ったみそ汁を飲んだり、買ったカニを焼いて食べたりする場所もあって、多くの人が楽しんでいました。金沢市から来た小学校２年生の男の子は...
記事処理→文処理(Nコーパス)	記事処理→文処理(Y+Nコーパス)	文処理(Y+Nコーパス)→記事処理
...このイベントは、の漁協などが、毎年開いていて、石川県輪島市の港の近くの特設会場では、会社の説明などが入ってを並べカニを売っています。会場には、ズワイガニ5000匹ぐらいの雄と 雌ぐらい1万匹並べました。しかし、この中の漁協が出した店では、雄が1匹2000円から、雌が1匹600円で売っていました。金沢市から来た小学生の男の子は...	...このイベントは、漁協などが毎年開いているもので、石川県輪島市の港の近くの特設会場では、今月の会社などが机を置いてカニを売っています。会場には、ズワイガニの約5000匹雄 と雌の約1万匹を並べました。このうちの漁協が出した店では、雄が1匹2000円から、雌が1匹600円で売られていました。金沢市から 来た小学生の男の子は...	...このイベントは、漁協などが毎年開いているもので、石川県輪島市の港の近くの特設会場では、今月の会社などが机を置いてカニを売っています。会場には、ズワイガニの約5000匹雄 と雌の約1万匹を並べました。このうちの漁協が出した店では、雄が1匹2000円から、雌が1匹600円で売られていました。 会場では、多くの人などが長い列を作って、た人のカニを食べていました。金沢市から来た小学生の男の子は...

図 1: 本実験で生成された記事の例 (一部. 赤字は適切な書き換え, 青字は誤った書き換え)

の記事でも描写されている会場での観光客の様子を述べる文が存在する (下線部, 変換ミスを含む). 文処理モデルで文短縮が行われるため, 記事処理モデルを適用する際の目標長が同じ場合は, 先に文処理モデルで各文の長さを短くすることで, より多くの重要文を目標長内に収められる可能性がこの例から示唆される [12].

5 おわりに

本研究では, NNW の記事に対し要約及び平易化の両方を行うことを試みた. その過程を記事全体に対する処理と各文に対する処理に分けてモデルを実装し, それぞれのモデルの学習には, NNW-NNWE の記事対から構築した擬似パラレルコーパスを用いた. 生成された記事に対する定量評価の結果, 記事処理及び文処理モデルの両方を適用することで, 元記事 (NNW) や一方のモデルのみを適用する場合に比べ BLEU, ROUGE, SARI の値が向上し, より参照記事 (NNWE) に近い, 短く平易な記事を生成できることが示された. また, 実際の生成例から, 文処理モデルの学習の際 N コーパスと Y コーパスを組み合わせることで文法的なミスを多少抑えられる可能性や, 文処理モデルを先に適用することで, 記事処理モデルを適用する際により多くの重要文を目標長内に収められる可能性が示唆された.

今後の課題としては, 記事処理及び文処理モデルの学習の際に, 単語の汎化や間違った対応づけの除去など改善の余地があり, 有効な処理方法について検証する必要がある. また, より質の高い擬似パラレルコーパスを構築するために, 文単位でなく文節単位で自動対応づけを行う手法も検討していく.

[12] 先に記事処理モデルを適用する場合は目標長を長めに設定することも考慮する必要があるが, 本研究では適切な長さを求めることができず同じ長さに設定した.

参考文献

[1] Akihiro Katsuta and Kazuhide Yamamoto. Crowdsourced Corpus of Sentence Simplification with Core Vocabulary. In *Proceedings of the 11th International Conference on Language Resources and Evaluation*, 2018.

[2] Chin-Yew Lin. ROUGE: A Package for Automatic Evaluation of Summaries. In *Proceedings of the Workshop on Text Summarization Branches Out*, pp. 74–81, 2004.

[3] Takumi Maruyama and Kazuhide Yamamoto. Simplified Corpus with Core Vocabulary. In *Proceedings of the 11th International Conference on Language Resources and Evaluation*, 2018.

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

[5] Yangqiu Song and Dan Roth. Unsupervised Sparse Vector Densification for Short Text Similarity. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1275–1280, 2015.

[6] Yasuhito Tanaka. Compilation of A Multilingual Parallel Corpus. In *Proceedings of Pacific Association for Computational Linguistics*, pp. 265–268, 2001.

[7] Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, Vol. 4, pp. 401–415, 2016.

[8] 永塚光一, 渥美雅保. Breaking News English Corpus: マルチレベルテキスト平易化に特化した複数文書要約コーパスの提案. 言語処理学会第 25 回年次大会, pp. 351–354, 2019.

[9] 梶原智之, 小町守. 平易なコーパスを用いないテキスト平易化. 自然言語処理, Vol. 25, No. 2, pp. 223–249, 2018.

[10] 熊野正, 後藤功雄, 田中英輝. 統計機械翻訳を用いたニュース文のやさしい日本語への自動変換. 2015 年映像情報メディア学会年次大会, 32D-2, 2015.

[11] 西川仁, 有田一穂, 田中克巳, 平尾努, 牧野俊朗, 松尾義博. 識別的半マルコフモデルによるテキスト結束性を考慮した単一文書要約. 情報処理学会論文誌, Vol. 57, No. 2, pp. 769–782, 2016.

[12] 田中英輝, 熊野正, 後藤功雄, 美野秀弥. やさしい日本語ニュースの制作支援システム. 自然言語処理, Vol. 25, No. 1, pp. 81–117, 2018.