文書単位の日本語テキスト平易化コーパスの構築に向けて

長井 慶成　岡 照晃　小町 守
東京都立大学
nagai-yoshinari@ed.tmu.ac.jp　{teruaki-oka, komachi}@tmu.ac.jp

概要

日本語の平易化の研究は，英語など第二言語学習者の多い言語に比べ言語資源に乏しく，研究も盛んでない現状である．その打開に向け，我々は新たな文書単位の日本語テキスト平易化コーパスの構築を目指している．本稿ではまず毎日新聞と毎日小学生新聞を対象に，同じ内容を扱った記事同士を比較する予備調査を行った．その結果から，5 つの平易化操作を定義し，平易化前後の文アライメント評価データを作成した．さらにテキスト平易化に向けて作られた既存の文アライナーでの評価分析に取り組んだ．

1　はじめに

テキスト平易化とは，理解が困難な文書である「難解文書」の持つ意味を保ちながら，語彙や文法的な複雑さの抑えられた解釈のしやすい「平易文書」へと換言するタスクである．平易化の利点として，テキストの読解レベルを下げることで子供や第二言語学習者などの非ネイティブスピーカーの読解支援 [1] が可能な点がある．

英語には「Wikipedia と Simple English Wikipedia」や，「ニュース記事とそれを平易に書き換えたもの」など大規模なコンパラブルテキストが既に存在しており，これらを対応付けたテキスト平易化コーパスは英語の平易化の研究促進に貢献している．一方，日本語の平易化のための言語資源は，英語に比べて乏しいのが現状である．既存の日本語の平易化コーパスの多くは，語彙平易化や文法平易化といった特定の平易化に焦点を当て，また文単位の平易化のみを考慮したものとなっている．英語のような大規模なテキスト平易化コーパスが存在しないため，日本語では文書単位の平易化がどのような操作により実現されているかの調査は十分に行われておらず，未だ不透明な部分が多い．

このような課題に対処すべく，我々は日本語の文書単位で，かつ平易化の操作を網羅した新たな日本語テキスト平易化コーパスの構築に取り組んでいる．本稿ではその前段階としてコンパラブルテキストである毎日新聞と毎日小学生新聞の記事データの同じ内容を扱った記事同士を比較する予備調査を行う．調査の結果から日本語の文書単位の平易化に必要な平易化操作を定義し，それに基づきアノテーションを行なって「文アライメント評価データ」を作成した．最後に作成したデータを用いて，既存の自動文アライナーの評価を行なった．

本稿の主な貢献は，以下の 3 点である．

- 予備調査から，日本語の文書単位の平易化における 5 つの平易化操作を新たに定義した．
- 定義した平易化操作や文の対応が付与された「文アライメント評価データ」を作成した．
- 文アライナーでの評価分析により，平易化操作や文の対応の自動付与の可能性を示した．

2　関連研究

2.1　英語の平易化コーパス

英語の代表的な平易化コーパスには，Wikipedia とニュース記事をドメインとした 2 種類が存在する．

Wikipedia をドメインとした平易化コーパス [2][3] は，Simple English Wikipedia と English Wikipedia を対応付けることで構築されている．Simple English Wikipedia[1] は，一部の English Wikipedia[2] の記事を平易化したテキストである．平易化はユーザーが行っており，基本的な語彙や文法に限定した書き換え，詳細追加などで記事を作成している．

ニュースドメインの平易化コーパスとして，Newsela[3] コーパス [4] がある．この平易化コーパスは，ニュース記事を人手により 4 段階のレベルで書き換えることで構築されたコーパスである．

1）https://simple.wikipedia.org/
2）http://en.wikipedia.org/
3）https://newsela.com/data/

日本語の Wikipedia には Simple English Wikipedia にあたるような平易化された記事は存在しない．また Newsela のような複数の難易度を考慮したコーパスも存在しないため，日本語の平易化のための言語資源は乏しい現状にある．

2.2 日本語の平易化コーパス

『SNOW T15: やさしい日本語コーパス』[5] は，語彙平易化に焦点を当てたコーパスである．日本語の基本的な語として 2,000 語を選定し，難解な文をその語彙のみを用いた表現に書き換えて構築された．

他にこのコーパスを参考にした『SNOW T23: やさしい日本語拡張コーパス』[6] や文法平易化に焦点を当てた『日本語文法平易化コーパス』[7] などもあるが，いずれも特定の平易化に特化し，また文単位の平易化のみを考慮したものである．

我々は日本語の文書単位での平易化を見据えている．それに向けて，以上のような文単位の日本語平易化コーパスとは異なり，特定の平易化に限定せず，平易化の操作を網羅した文書単位の日本語テキスト平易化コーパスの構築に取り組んでいる．

3 予備調査

この章では，文書単位での平易化に必要となる平易化操作を新たに定義するための予備調査を行う．次の 4 章で，定義した平易化操作や対応する平易文書の文 ID を難解文書の各文にアノテーションして文アライメント評価データを作成する．

3.1 データセット

難解文書と平易文書の関係があると期待される日本語のコンパラブルテキストとして，毎日新聞と毎日小学生新聞の記事がある．毎日小学生新聞は小学生向けの日刊紙であり，公式からの言及はないが，毎日新聞の記事を小学生が理解できるような平易な表現に書き換えたと思われる記事が散見される．そのため本稿では毎日新聞記事を「難解文書」，毎日小学生新聞記事を「平易文書」として扱う．

予備調査では，まず毎日新聞と毎日小学生新聞の記事集合に対して記事アライメントを行い，同じ内容を扱った記事ペアを抽出する．次に目視で記事ペアの文の対応付けを行い，どのような操作により平易な表現への書き換えが行われているかを調査する．その結果から日本語の文書単位の平易化に必要な平易化操作を新たに定義する．

表 1 平易化操作の種類と操作

種類	操作
編集	難解記事の文が平易側の文と 1 対 1 で対応
スプリット	難解記事に存在する文が平易版で複数存在
マージ	難解記事の複数文が，平易版の 1 文と対応
文の削除	難解記事のみに存在し，平易版では未記載
文の挿入	補足説明を加えるなど，平易版のみに存在

予備調査には毎日新聞，および毎日小学生新聞の一年分の掲載記事を収録した 2014 年版の『CD-毎日新聞データ集』，『CD-毎日小学生新聞 記事データ集』を使用する．

3.2 記事の対応付け

毎日新聞と毎日小学生新聞の記事集合から同じ内容を扱った記事ペアを抽出する．毎日小学生新聞の特徴として，漢字の後に括弧でふりがなを振る傾向がある．これを取り除くため，両方のデータセットで前処理として丸括弧とその中の文字列は削除する．前処理を加えた記事に対し，ユーザー辞書に「令和」を登録した mecab ipadic-NEologd で形態素解析を行い，2 文字以上の動詞，形容詞，名詞を抽出する．ただし数字，ストップワード（一部の非自立語動詞），基本形が「*」のものなどは除く．毎日小学生新聞の掲載日ごとの記事を形態素解析して抽出した単語の頻度ベクトルを作成，tf-idf で重み付けした．掲載日の前後 7 日の毎日新聞記事と余弦類似度をとることで記事を自動抽出した．

bag-of-words での記事アライメントでは，掲載時間のずれにより詳細まで完全に一致したものを得る事はできない（例：被害状況が更新されるなど）．また毎日小学生新聞への書き換えでは不要な句や節を削除する平易化の操作が行われており，記事の類似度が高いからといって必ずしも良い対応の取れた記事のペアとは限らない．そのため目視で確認し，内容が一致した類似度が 0.70 以上の記事ペアから類似度 0.05 刻みに 3 件ずつ計 18 件を抽出した．

3.3 結果

抽出した各記事を鍵括弧外の「。」，「？」，「！」で文分割し，文の対応を目視で確認した．対応が付いていると判断した文同士を比較して平易化のために行われている操作を調査した．操作の種類ごとに分類した結果，表 1 に示す計 5 つの平易化操作を新たに定義した．各平易化操作の例は，付録 A 参照．

表 2　毎日新聞記事へのアノテーション（一部抜粋）

ラベル	対応 ID	文 ID	文
文の挿入	3	0	囲碁：１０歳・仲邑初段、悔しい船出 【大阪】
スプリット	1,2	1	先月１０歳になったばかりの囲碁の史上最年少棋士、仲邑菫初段が２２日、大阪市の日本棋院関西総本部で公式戦初対局に臨んだが、大森らん初段に敗れ、ほろ苦いプロデビューとなった。
編集	2	2	早碁のテレビ棋戦「竜星戦」予選。

表 3　対応する毎日小学生新聞記事（一部抜粋）

文 ID	文
1	４月１日付で囲碁の史上最年少棋士になった仲邑菫初段が２２日、大阪市で公式戦の初めての対局に臨みました。
2	第２９期竜星戦予選Ｂ１回戦で、同期の大森らん初段に敗れ、ほろ苦いプロデビューとなりました。
3	序盤は互角の戦いだったものの、大森初段が中盤で優勢を築きました。

表 4　平易化操作ラベルの割合と平均文長

平易化操作	総数	割合 (%)	平均文長 (文字)
編集	213	41.1	49.9
スプリット	32	6.2	72.8
マージ	32	6.2	48.3
文の削除	227	43.8	47.3
文の挿入	14	2.7	⟨33.2⟩

4　評価データの作成

4.1　データセット

文アライメント評価データの作成には，2019 年版の『CD-毎日新聞データ集』，『CD-毎日小学生新聞 記事データ集』のデータセットを使用する．記事が対応付いていると判断する閾値を類似度 0.75 以上として，3.2 節と同様に記事アライメントを行う．その結果，毎日新聞 66,858 件と毎日小学生新聞 3,529 件から 442 件の記事ペアを抽出した．そのうちの 287 件がすべて文が「。」，「？」，「！」のいずれかで終わる記事であった．見出しなどではなく文に加えられる平易化操作に関心があるので，この 287 件の中から目視で対応がとれていることを確認した 50 件の記事ペアを抽出した．

4.2　アノテーション

抽出した記事ペアの毎日新聞と毎日小学生新聞の記事を読み比べ，毎日新聞記事の各文に対して表 1 の平易化操作，毎日小学生新聞記事の対応する文の ID を付与するようにアノテーションを 1 名の作業者に依頼した．依頼時，毎日小学生新聞で挿入された文 ID を記録するため文 ID が 0 の行を用意し，文のカラムには記事の見出しを入れた．毎日新聞記事への人手アノテーションの結果の例と，その記事に対応付く毎日小学生新聞記事を表 2 と表 3 に示す．

5　分析

5.1　平易化操作ラベルの割合

毎日新聞と毎日小学生新聞の間で行われる平易化操作の傾向をつかむために，毎日新聞記事の各文に対して付与された表 1 の 5 つの平易化操作の総数と割合, 付与された文の平均文長を調査した．その結果を表 4 に示す．ただし，付録 B のように毎日新聞と毎日小学生新聞には文書の性質に違いがある．そのため毎日小学生新聞の文に付与される「文の挿入」は，毎日新聞の文に付与する他の平易化操作の平均文長とは単純な比較はできない．

最も多く行われた平易化操作は「文の削除」であった．これは難解な文を平易な文へと変換することを前提とした「文単位の平易化」にはない「文書単位の平易化」特有の操作である．すべての難解な文を平易なものにするのではなく，ときには削除して文書全体を要約する形で毎日新聞記事から毎日小学生新聞記事へ書き換えていることが伺える．

スプリットのラベルの付いた文の対応 ID をみると文の対応は 1:n となっており，n:m のような複数文が対応するような複雑な平易化操作は抽出した 50 件ではみられなかった．スプリットで連続する文に分割されるのは 32 回のうち 30 回であり，また付与された文の平均文長は他の操作よりも長いことから，文書の構造を変化させるのではなく，文長を短くすることで理解しやすいものに書き換える傾向があるといえる．

マージはスプリットと同数だが，マージは n:1 の関係をとる平易化操作であり，毎日新聞の各文にアノテーションを行っているため実際の操作回数よりも多くなっている．連続するマージを 1 回とカウントした場合の総数は 24 回であり，平易化操作としてはスプリットの方が多用されている．

表 5　平易化操作ラベルの条件

平易化操作	条件 1	条件 2	条件 3	条件 4
編集	○	×	×	×
スプリット	○	○	×	×
マージ	○	×	○	×
文の削除	×	×	×	×
文の挿入	×	×	×	○

5.2　アライナーによる文の対応付けと分類

既存の文アライナーで作成した評価データの文アライメントを行い，その結果をもとに「平易化操作」と「対応 ID」を自動付与する．これを評価データの人手アノテーションと比較して一致率を調査する．

5.2.1　アライナー

CATS[4)][8][9] は，段落や文を log tf-idf で重み付けした文字 3-gram で表現し，文書間で最大類似度のものと対応付けるアライナーである．本稿では文アライナーとして使用し，対応元文書の各文の「対応 ID（対応する文の ID）」と「余弦類似度」を求める．1:n の対応をとるスプリットやマージの検出のために「毎日新聞-毎日小学生新聞」の両方向で行う．

5.2.2　平易化操作ラベルの自動付与

作成した評価データの 50 記事から抽出した 15 記事をもとに以下の 4 つの条件を設定し，表 5 のように満たすかで平易化操作を自動付与する．ラベルを付与する毎日新聞の文を「S_{mai}」，CATS での毎日新聞から毎日小学生新聞への対応付けで S_{mai} と最大類似度をとる毎日小学生新聞の文を「S_{sho}」とする．

条件 1　毎日小学生新聞から毎日新聞への対応付けで S_{sho} の最大類似文は S_{mai} で，類似度は 0.20 以上

条件 2　毎日小学生新聞の文に S_{sho} 以外に S_{mai} が最大類似文，かつ類似度が 0.20 以上の文がある

条件 3　毎日新聞の文に S_{mai} 以外に S_{sho} が最大類似文，かつ類似度が 0.20 以上の文がある

条件 4　毎日小学生新聞の文で，毎日新聞のすべての文との類似度が 0.13 以下

5.2.3　評価

CATS の出力と表 5 をもとに自動で付与された各平易化操作ラベルの総数と，それが人手アノテーションしたラベルと対応 ID のどちらとも一致するものの総数と割合を表 6 に示す．「なし」は自動付与では文の挿入は行われていないとされたが，人手アノテーションではあると判断されたものを表す．

表 6　平易化操作ラベルの一致率

平易化操作	総数	一致数	一致率 (%)
編集	158	137	86.7
スプリット	18	16	88.9
マージ	31	9	29.0
文の削除	157	154	98.1
文の挿入	10	8	80.0
なし	2	0	0.0
全体	376	316	84.0

表 7　自動付与と人手によるアノテーションの比較

人手	自動	毎日新聞
D	M	「Ｎ７００Ｓ」は現行の「Ｎ７００Ａ」の後継として開発。
E	M	２０２０年７月２４日に開幕する東京五輪直前のデビューを計画し、米テキサス州の高速鉄道や台湾新幹線などへの売り込みを狙う。

マージを除く平易化操作は人手アノテーションと高い一致率であった．マージの一致率が低い要因として 2 点考えられる．1 点目として文字 3-gram での対応付けにより内容は異なるが出現単語が類似した文とも高い類似度となり，誤った対応付けをすることがある．2 点目としてマージの判断は人手でも困難なことが挙げられる．毎日小学生新聞の『「Ｎ７００Ｓ」は来年の東京オリンピック直前にデビューする予定で，アメリカの高速鉄道や台湾新幹線などへの売り込みを目指しています。』という文に対応する毎日新聞の文に人手と自動では表 7 のように異なるラベルを付与した．この場合は自動付与の方が適切なラベルといえる．文アライナーの活用することで人手付与の誤りを検知でき，より高い精度で平易化操作ラベルを付与することが期待される．

6　おわりに

本稿では，毎日新聞と毎日小学生新聞の記事ペアから平易化の予備調査を行った．「調査により新たに設計した 5 つの平易化操作」と「対応する文 ID」が付与された文アライメントの評価データを作成した．また，文アライナーにより自動での平易化操作を分類した結果，人手アノテーションと高い一致率であった．このことから，文アライナーを活用して自動で平易化操作ラベルや対応する文 ID を付与することで，十分なコンパラブルテキストさえ用意すれば，大規模な文書単位の平易化コーパスを構築できる可能性を示した．

4)　https://github.com/neosyon/SimpTextAlign

謝辞

本研究成果は，国立研究開発法人情報通信研究機構（NICT）の委託研究「自動翻訳の精度向上のためのマルチモーダル情報の外部制御可能なモデリングの研究開発」により得られたものである．

参考文献

[1] Jan De Belder and Marie-Francine Moens. Text simplification for children. In **SIGIR Workshop on Accessible Search Systems**, pp. 19–26, 2010.

[2] Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych. A monolingual tree-based translation model for sentence simplification. In **International Conference on Computational Linguistics**, 2010.

[3] William Coster and David Kauchak. Simple English Wikipedia: A new text simplification task. In **Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies**, pp. 665–669, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

[4] Wei Xu, Chris Callison-Burch, and Courtney Napoles. Problems in current text simplification research: New data can help. **Transactions of the Association for Computational Linguistics**, Vol. 3, pp. 283–297, 2015.

[5] Takumi Maruyama and Kazuhide Yamamoto. Simplified Corpus with Core Vocabulary. In Nicoletta Calzolari (Conference chair), Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, **Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)**, Miyazaki, Japan, May 7-12 2018. European Language Resources Association (ELRA).

[6] Akihiro Katsuta and Kazuhide Yamamoto. Crowdsourced Corpus of Sentence Simplification with Core Vocabulary. In Nicoletta Calzolari (Conference chair), Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, **Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)**, Miyazaki, Japan, May 7-12 2018. European Language Resources Association (ELRA).

[7] 稲岡 夢人, 山本 和英. 日本語文法平易化コーパスの構築. 言語処理学会第 25 回年次大会, pp.375-378, March 2019.

[8] Sanja Stajner, Marc Franco-Salvador, Simone Paolo Ponzetto, Paolo Rosso, and Heiner Stuckenschmidt. Sentence alignment methods for improving text simplification systems. In **Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 2: Short Papers**, pp. 97–102, 2017.

[9] Sanja Štajner, Marc Franco-Salvador, Paolo Rosso, and Simone Paolo Ponzetto. CATS: A tool for customized alignment of text simplification corpora. In **Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)**, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA).

A　平易化操作ごとの例

表 8　平易化操作ごとの例

平易化操作	毎日新聞	毎日小学生新聞
編集	参院によると、介助者が同席する重度障害者の質問は初めて。	参議院によると、介助者がつきそう重度障害者の質問は初めてです。
スプリット	対イラン圧力強化の一環で、米連邦法に基づき外国の政府機関そのものをテロ組織に指定するのは初めて。	イランへの圧力強化の一つです。 アメリカ連邦法に基づいて外国の政府機関をテロ組織に指定するのは初めてです。
マージ	船名の表示はなかったが、漁船の登録番号から所有者が判明した。 漁船は、釜石市唐丹町花露辺の漁師、佐々木清文さんが所有していた「清昭丸」。	船の名前の表示はありませんでしたが、漁船の登録番号から釜石市の漁師、佐々木清文さんが持っていた「清昭丸」だと分かりました。
文の削除	ＡＩの基礎研究から健康医療や社会インフラ、工場などで活用する応用分野まで幅広い研究を行う。	-
文の挿入	-	ＮＥＣは試作機を飛ばしてデータを集め、２０年代半ばには人の移動の実現につなげたいと考えています。

B　作成した文アライメントの評価データの統計量

表 9　評価データの統計量

	平均文字数	平均文数	平均文長 (文字)
毎日新聞	505.0	10.1	50.1
毎日小学生新聞	274.9	6.1	44.9

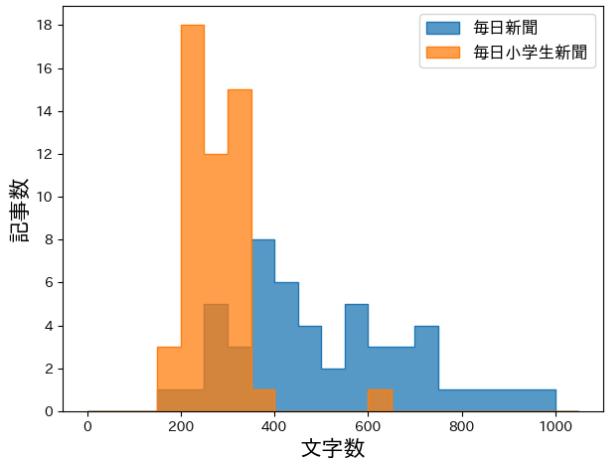

図 1　評価データの文字数のヒストグラム

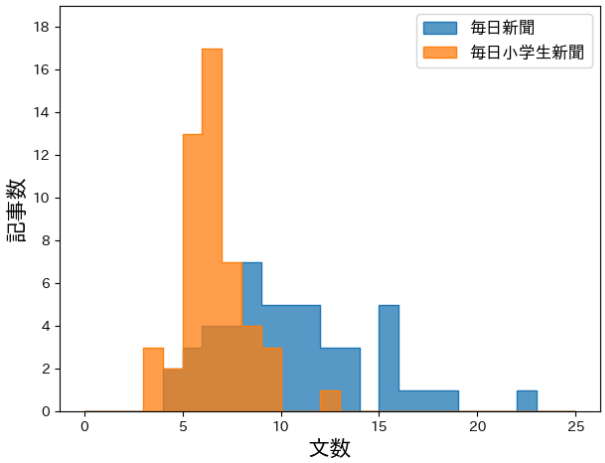

図 2　評価データの文数のヒストグラム