やさしい表現へのニュースの自動変換評価用データセットの構築

後藤 功雄　　熊野 正　　田中 英輝

NHK 放送技術研究所

1 はじめに

日本在住の外国人へニュースを分かりやすく伝えるために、ニュースをやさしい日本語で提供する研究を進めている。NHK ではこれまで、やさしい日本語ニュースへの書き換えの基準を作成し [2]、NHK のインターネットサービス「NEWSWEB EASY」でこの基準に沿って書き換えたニュースの配信を 2012 年 4 月から開始している。現在の書き換えは、支援システム [4] を利用して人手で行っており、一日あたり 5 記事のニュースを配信している。この支援システムに自動処理技術を導入することで作業を効率化して、より多くのやさしい日本語ニュースを配信することを目指している。

この自動処理技術の 1 つとして、統計的機械翻訳（SMT）の技術を用いて表現をやさしく自動変換することを考えている [3]。そこで、SMT による自動変換の実現と品質評価のために、これまでに蓄積されたやさしく書き換えられたニュースと書き換え前のニュースの記事対から対応する文対を抽出し、訓練データ、開発データ、テストデータからなる自動変換評価用データセットを構築した。

本稿は、構築した自動変換評価用データセットについて説明する。この説明の前に、まず、やさしい日本語ニュースへの書き換えと自動変換の対象とする書き換えについて述べる。

2 やさしい日本語ニュースへの書き換え作業と自動変換の対象

やさしい日本語ニュースは、普通の日本語ニュースを書き換えて表現をやさしくし、内容を簡潔にしたものである。やさしい日本語を学んだ日本語教師が主に表現をやさしくし、記者（経験者）が主に内容を簡潔にしている。日本語教師と記者がこれらの書き換えを交互に実施している。[*1]

これら 2 者の書き換えのうち、日本語教師が行っている書き換えを自動変換の対象とする。

3 自動変換評価用データセットの構築

本節では自動変換評価用データセットの構築方法について説明する。

3.1 正解文の設定

自動変換先の目標である正解文と入力のテスト文は以下のように設定した。日本語教師と記者が交互に実施する書き換えは必要に応じて複数回行われる場合がある。すなわち、1 つのニュースをやさしい日本語ニュースにする過程で日本語教師は複数回の書き換えを行う場合がある。正解文は、このうち最初の日本語教師の書き換え結果とする。そして、テスト文はその書き換え直前の文とする。これは次の理由による。2 回目以降の書き換えでは入力の多くの部分で既にやさしい表現になっていると考えられるので、書き換え直前の文をテスト文とした場合に書き換えが少ないためである。また、テスト文を最初に日本語教師が書き換える直前の文とし、正解文を 2 回目以降の日本語教師の書き換え結果にすると、日本語教師だけでなく、記者による書き

[*1] 日本語教師と記者のうち、先に日本語教師が元のニュースを書き換えた割合は次の通りである。NEWSWEB EASY が公開実験として開始した 2012 年 4 月は 58%、本運用を開始した 2013 年 5 月は 6%、2013 年 9 月は 1% である。サービス開始当初は、日本語教師と記者が同時に作業を開始していたため、約半分は日本語教師が元のニュースを書き換えていたが、最近ではほとんどの場合に先に記者が内容を簡潔にしている。これは、内容を簡潔にして量が少なくなってから表現をやさしくした方が日本語教師の作業量が減って効率が良いためである。

換えも含まれてしまい、表現をやさしくする部分に着目した評価ができなくなってしまうためである。

本稿では、日本語教師が最初に書き換える直前の文を「原文」、書き換えた文を「目的文」と呼ぶ。また、原文と対応する目的文との対をパラレル文対、書き換え前後の記事対をパラレル記事対と呼ぶ。

3.2 人手によるパラレル文対の抽出

評価の信頼性を確保するために、テスト文と正解文は人手により抽出して構築した。まず、パラレル記事対内の原文と目的文との文アラインメントを人手により付与し、パラレル文対を同定した。次に、テスト文を 1 文単位で自動変換した際の品質を評価することが目的であるため、以下の条件でパラレル文対からテスト文と正解文を抽出した。

- 原文の 1 文が 1 文以上の目的文に書き換えられたものであり、1 つの目的文が複数の原文にアラインメントされたものを除く。
- 原文の主要な内容が目的文に含まれていないものや、原文に含まれない内容が目的文に追加されているものを除く。原文のニュースの主要な内容を伝えるために必要な情報が含まれていれば、詳細な情報は省略されていてもよい。文脈に依存せず常に成り立つ情報が追加されていてもよい。

1 番目の条件を「文数制約」、2 番目の条件を「ノイズ制約」と呼ぶ。文脈に依存せず常に成り立つ情報の追加とは、例えば原文中の表現が「青森県」で、目的文中の表現が「東北地方の青森県」のような場合である。

テスト文と正解文を構築するために、490 記事対に対して人手で文アラインメントを付与した。この結果、上記の条件でテスト文と正解文を抽出できた記事数は 485 であった。490 記事対中の文数制約を満たす文アラインメントに対して、ノイズ制約を満たすもの（ノイズ無し）と満たさないもの（ノイズ有り）に分類した割合を表 1 に示す。

3.3 文アラインメント自動推定の要件

訓練データを構築するための文アラインメントは、コスト削減および今後も増加していくニュースデータを自動的に利用できるようにするために自動処理を用いる。まず、文アラインメント自動推定の要件を検討した。

表 1　原文 1 文–目的文 1 文以上のパラレル文対のノイズの有無

	文対数	割合
ノイズ無し	2,012	0.697
ノイズ有り	873	0.303

表 2　人手で付与した文アラインメントの種類毎の頻度

原文数–目的文数	頻度	割合
1–1	2201	0.698
1–2	575	0.182
1–3	90	0.029
1–0	71	0.023
0–1	67	0.021
2–2	52	0.016
2–1	46	0.015
2–3	17	0.005
1–4	16	0.005
その他	20	0.006

2 節で説明したとおり、やさしい日本語を学んだ日本語教師が主に表現をやさしくし、記者が主に内容を簡潔にしているが、この分担は明確に分かれたものではない。記者の書き換えで一部の表現をやさしく書き換えることはよく行われている。日本語教師の書き換えでは、表現をやさしくし、文を分割して文構造を簡単にするだけでなく、詳細な内容を省略したり、補足説明を追加したり、文の順番を変更する場合もある。日本語教師による書き換え前後の記事と文アラインメントの例を図 1 に示す。

人手で文アラインメントを付与した 490 記事対中の文アラインメントの種類毎の頻度と割合を表 2 に示す。原文が 1 文で目的文が 1 文以上のアラインメントが多いが、文の省略・追加や原文が複数文の場合もあることが分かる。

また文の順番の相関を示す Kendall の τ [*2]の平

[*2] この値は、次のように計算した。アラインメントされていない目的文は削除し、アラインメントされている原文の文番号を目的文に付与し、目的文に対して原文の文番号が小さい順に文番号を振り直した。この際に、原文の文番号が同じ場合の文番号は、目的文の文番号が小さい順とした。この振り直した文番号の並びと目的文の並び順に対して、Kendall の τ の値を計算した。

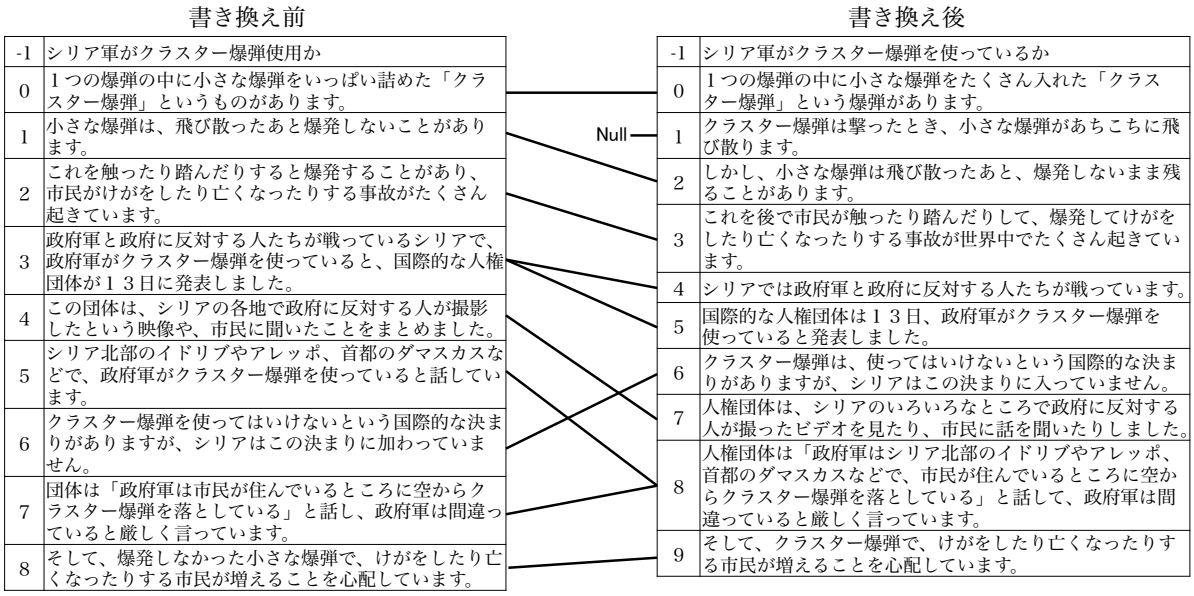

図1　書き換え前後の記事と文アラインメントの例（文番号-1 は記事タイトルを表す）

均値は 0.987 であった。値が 1 に近いため文の順番は多くの場合で一致していることが分かる。

これらのことから、原文と目的文の文アラインメント推定では、1 文対 1 文のほか複数文を含むアラインメント、文の省略、文の追加、文の順番を推定できることが必要である。

3.4　文アラインメント自動推定

原文と目的文は、一致する語が多く含まれており、多くの場合に文の順番も一致する。そのため、一致する語に基づいてダイナミックプログラミングを用いてアラインメントが交差しない範囲で文アラインメントを自動推定する手法を用いることにした。この手法の 1 つの実装である Champollion[1] を文アラインメント自動推定に用いた。この手法は、対応する文対の文の順番を同じに制限した上で[*3]、文の省略・追加、連続する複数文を含むアラインメントも推定することができる。訓練データ構築に用いた 1,559 記事対に対して、自動推定で獲得した文アラインメントの種類毎の数を表 3 に示す。

[*3] 順番が変わった場合はアラインメントできないが、順番の変更があった文について対応先無しと推定することができれば質の低いパラレル文対を抽出することを避けることができる。

表3　訓練データに対する自動アラインメント結果の数

原文数-目的文数	頻度	割合
1–1	7893	0.727
1–2	2336	0.215
1–3	383	0.035
2–1	111	0.010
2–2	44	0.004
1–4	39	0.004
0–1	32	0.003
1–0	19	0.002
3–1	5	0.000

表4　文アラインメントの Precision と Recall

	Precision	Recall
全体	0.872	0.885
1 原文–1 以上の目的文	0.881	0.942

3.5　文アラインメント推定の品質評価

人手で文アラインメントを付与した 490 記事対を用いて、文アラインメント自動推定結果の品質を評価した。アラインメントの単位を 1 つの対応関係（例えば 1 文対 1 文、1 文対 2 文、1 文対 0 文など）として、全体の文アラインメントおよび原文が 1 文で目的文が 1 文以上の文アラインメントに対する Precision と Recall を表 4 に示す。

なお、やさしい表現への自動変換では、変換でき

表 5　評価用データセットの文対数

	期間（括弧内は記事番号）	記事数	文対数
訓練	12/04/02(1)–14/02/26(3)	1,559	10,651
開発	14/02/26(4)–14/04/24(2)	170	723
テスト	14/04/24(3)–14/09/30(5)	485	2,012

ない表現はそのまま出力すれば入力に対して出力の品質が低下することはないが、変換に誤りが含まれると出力の品質が低下する。そのため、訓練データはデータ量が多いことよりノイズが少ないことの方が比較的重要であると考えられる。

4　自動変換評価用データセット

本節では、構築した自動変換評価用データセットの諸元を説明する。このデータセットは、2012 年 4 月から 2014 年 9 月までの間に蓄積したデータから構築した。この期間のデータのうち、普通のニュースからやさしい日本語ニュースの最終稿までの編集履歴の中に日本語教師の書き換えを含むものを利用した。

4.1　テストデータ

テストデータは最も新しい期間の記事を用いて構築した。3.2 節で説明した方法で人手によりテスト文と正解文となるパラレル文対を抽出した。テストデータは 2,012 文のテスト文とそれらの正解文から構成される。期間やデータ数を表 5 にまとめる。また、テスト文を抽出した記事データをテスト文の文脈データとしてデータセットに含めた。正解文の主な目的は、機械翻訳の自動評価手法などを用いて自動変換した文の品質評価に利用することである。

4.2　開発データ

開発データはテストデータより前の期間の記事を用いて構築した。テストデータと同じ方法で人手によりパラレル文対を抽出した。開発データは、723 文の原文とそれに対応する目的文からなる（表 5）。また、原文を抽出した記事データを開発データの文脈データとしてデータセットに含めた。開発データの目的は、SMT システムのパラメータのチューニングに利用することを想定している。

4.3　訓練データ

訓練データは開発データより前の期間の記事を用いて構築した。3.4 節で説明した方法で自動獲得した文アラインメントから、原文が 1 文で目的文が 1 文以上のパラレル文対を抽出した。訓練データのパラレル文対は、10,651 文の原文とそれに対応する目的文である（表 5）。

また、このパラレル文対を抽出したパラレル記事対もデータセットに含めた。そのため、他の文アラインメント手法を用いて新たにパラレル文対を抽出することも可能である。さらに、訓練データ期間の記事について普通のニュースからやさしい日本語ニュースの最終稿までの全ての編集過程の記事もデータセットに含めた。このデータは言語モデルの構築などに利用することを想定している。

5　おわりに

NEWSWEB EASY サービスのための作業で蓄積されたデータを用いて、やさしい表現へのニュースの自動変換評価用データセットを構築した。

今後はこのデータセットを用いて自動変換の評価・課題の調査・改善を進めていく予定である。また、現在はほとんどの場合に日本語教師による書き換えは記者による書き換えの後に行われているが、最初に表現をやさしく自動変換したほうが人間の作業効率が上がる可能性がある。そこで、この条件を前提にした評価データも構築したい。

参考文献

[1] Xiaoyi Ma. Champollion: A robust parallel text sentence aligner. In *Proceedings of LREC*, 2006.

[2] 田中英輝, 美野秀弥. やさしい日本語によるニュースの書き換え実験. 情報処理学会研究報告, Vol.2010-NL-199, No.11, 2010.

[3] 後藤功雄, 熊野正, 田中英輝. 一般のニュースからやさしい日本語ニュースへの書き換えの分析. 言語処理学会第 20 回年次大会, pp. 15–18, 2014.

[4] 美野秀弥, 田中英輝. ニュース原稿のやさしい日本語ニュースへの書き換え支援ツール―日本在住外国人のために―. 映像情報メディア学会年次大会, No.18-6, 2012.