Vol.2016-NL-229 No.13
2016/12/21

平易なコーパスを用いない テキスト平易化のための単言語パラレルコーパスの構築

梶原 智之[1,a)]　小町 守[1,b)]

概要：統計的機械翻訳の枠組みを用いたテキスト平易化が近年活発に研究されているが、その学習に必要な単言語パラレルコーパスを人手で構築することはコストが高い。そのため、公開されておりテキスト平易化のために自由に利用できるのは、English Wikipedia と Simple English Wikipedia のコンパラブルコーパスから単言語アライメントによって自動的に構築された英語のパラレルコーパスのみであるが、Simple English Wikipedia のように平易に書かれた大規模なコーパスは英語以外の多くの言語では利用できない。そこで、我々は日本語をはじめとする任意の言語でのテキスト平易化を実現することを目指し、生コーパスのみからテキスト平易化のための単言語パラレルコーパスを自動構築する手法を提案する。我々はまず文のリーダビリティを計算し、生コーパスを難解な文からなるコーパスと平易な文からなるコーパスに分解する。そして、単語分散表現を用いて計算される単語アライメントに基づく文間類似度によって、難解な文と平易な文の文アライメントを求める。我々の提案手法は、ラベル付きデータや辞書などの外部知識を必要とせず、生コーパスのみを用いてテキスト平易化のための単言語パラレルコーパスを自動構築するので、任意の言語に適用できる。フレーズベース統計的機械翻訳を用いたテキスト平易化の実験の結果、提案手法は平易なコーパスを用いずに入力文よりも平易な同義文を生成することができた。

1. はじめに

難解なテキストの意味を保持したまま平易に書き換えるテキスト平易化は、言語学習者や子どもをはじめとする多くの読者の文章読解を支援する。近年、テキスト平易化を同一言語内の翻訳問題と考え、統計的機械翻訳の枠組みで入力文から平易な同義文を生成する研究 [1–8] が盛んである。しかし、異言語間の機械翻訳モデルの学習に必要な異言語パラレルコーパスとは異なり、テキスト平易化モデルの学習に必要な単言語パラレルコーパスの構築はコストが高い。これは、日々の生活の中で対訳（異言語パラレル）データが大量に生産および蓄積されるのとは異なり、難解なテキストを平易に書き換えることは自然には行われないためである。そのため、公開されておりテキスト平易化のために自由に利用できるのは、English Wikipedia [*1] と Simple English Wikipedia [*2] から自動的に構築された英語のパラレルコーパス [2,3,9,10] のみであるが、Simple English Wikipedia のように平易に書かれた大規模なコーパスは英語以外の多くの言語では利用できない。そこで本研究では、日本語をはじめとする任意の言語でのテキスト平易化を実現することを目指し、生コーパスのみからテキスト平易化のための単言語パラレルコーパスを自動構築する手法を提案する。

本研究の概要を図 1 に示す。我々は、まず文の可読性を表すリーダビリティを計算し、生コーパスを難解な文からなるコーパスと平易な文からなるコーパスに分解する。次に、難解な文と平易な文の全ての組に対して、単語分散表現を用いて計算される単語アライメントに基づく文間類似度を求める。そして、任意の閾値以上の類似度を持つ難解な文と平易な文の組を抽出し、テキスト平易化のための単言語パラレルコーパスを構築する。このような単言語パラレルコーパスを用いてフレーズベースの統計的機械翻訳モデルを学習することで、統計的機械翻訳の枠組みで入力文から平易な同義文を生成するテキスト平易化を実現できる。

本研究で生コーパスを分割して得られる難解なコーパスと平易なコーパスは、English Wikipedia と Simple English Wikipedia のようなコンパラブルコーパスではないため、得られるのは雑音の多い文対である。しかし、フレーズベース統計的機械翻訳が必要とするのはフレーズ単位の変換対であり、これは部分的な対応しか持たない雑音の多いパラレルコーパスからも獲得することができる。そして、最終的には言語モデルによるリランキングを行うため、雑

[1] 首都大学東京 システムデザイン研究科
[a)] kajiwara-tomoyuki@ed.tmu.ac.jp
[b)] komachi@tmu.ac.jp
[*1] http://en.wikipedia.org
[*2] http://simple.wikipedia.org

ⓒ 2016 Information Processing Society of Japan

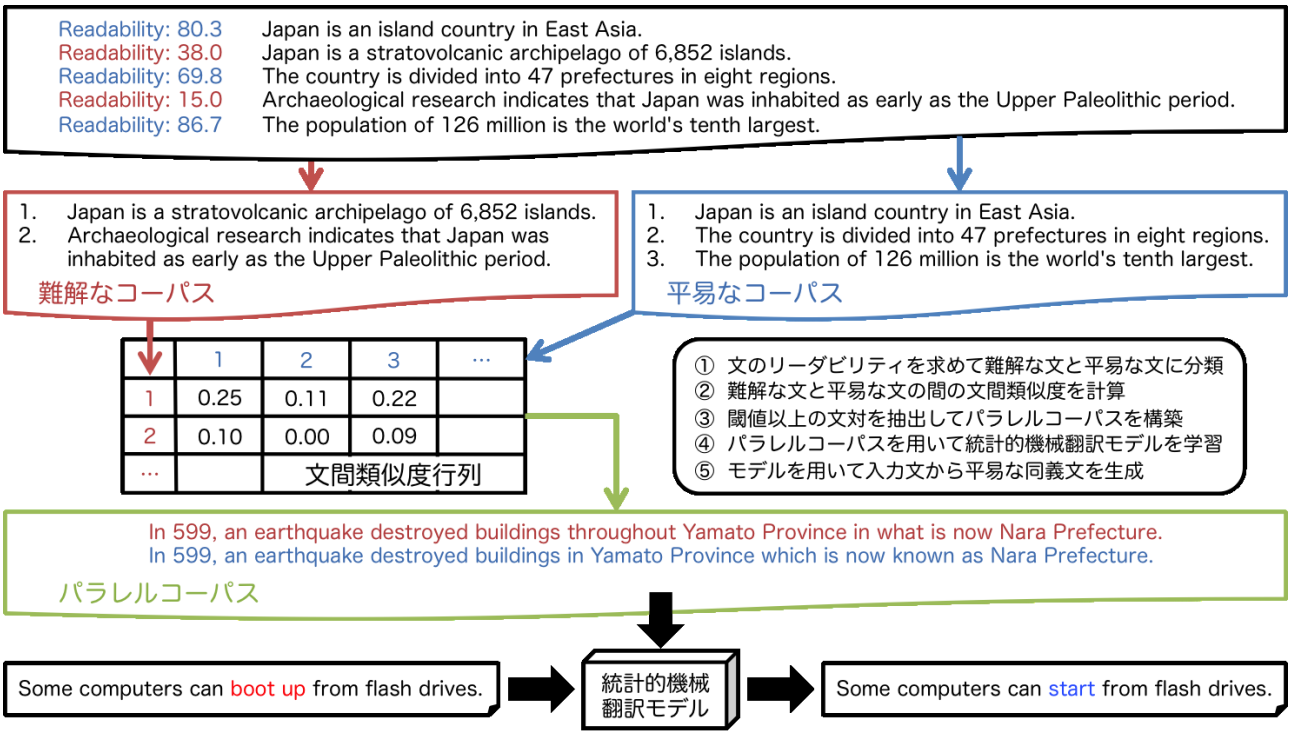

図 1 単言語パラレルコーパスの構築および統計的機械翻訳の枠組みでのテキスト平易化

音の多い変換対からでも、その中に適切なフレーズペアが得られていれば平易な同義文を生成することができる。本研究では、これを実験的に示す。

Xu ら [11] によって構築されたテキスト平易化のためのマルチリファレンスのデータセット [*3] を用いた評価の結果、我々のコーパスを用いて学習したモデルは、BLEU および SARI の自動評価尺度において先行研究のコーパスを用いて学習したモデルと同等の性能が得られることがわかった。BLEU は意味や文法、SARI は難易度に関して人手評価との相関が高いテキスト平易化のための自動評価尺度である [11]。つまり、Simple English Wikipedia のような平易に書かれた大規模なコーパスを利用することなく、入力文を平易に書き換えることができた。

本研究の貢献は、外部知識を利用することなくテキスト平易化を実現したことである。これまでは、人手で構築された難解な文と平易な文のパラレルコーパス [12]、平易に書かれた大規模なコーパス（Simple English Wikipedia）、文間類似度のラベル付きデータ [13–17]、言い換え知識 [18–20] などの言語資源が豊富に存在する英語を中心にテキスト平易化の研究が進められてきたが、本研究ではこれらの外部知識を利用することなく生コーパスのみからテキスト平易化のための単言語パラレルコーパスを自動構築し、統計的機械翻訳の枠組みでテキスト平易化を実現した。生コーパスは多くの言語で大規模に利用できるので、今後は本研究の成果をもとに、文のリーダビリティを測定可能な任意の言語でテキスト平易化を実現できるだろう。

[*3] https://github.com/cocoxu/simplification

2. 関連研究

2010 年以降、統計的機械翻訳の枠組みでのテキスト平易化の研究が盛んである。特に英語では、English Wikipedia と Simple English Wikipedia をコンパラブルコーパスと考え、ここから抽出された単言語パラレルコーパス [2,3,9,10] を用いた統計的機械翻訳の枠組みでのテキスト平易化 [2–6, 10] が盛んに研究されている。Coster and Kauchak [3] や Kajiwara and Komachi [10] は、標準的なフレーズベースの統計的機械翻訳ツール Moses [21] を用いて、統計的機械翻訳の枠組みでのテキスト平易化と自動評価尺度 BLEU [22] による評価を行った。BLEU ではリファレンスとの比較によって出力文の意味や文法の正しさを評価するが、テキスト平易化では入力文よりも平易な文を出力したいため、リーダビリティや SARI を用いて入力文と出力文を難易度の観点でも比較することが望ましい。一方で、フレーズの削除などのテキスト平易化に特化した翻訳モデル [2,4,5] も提案され、リーダビリティや BLEU の改善が示された。英語以外の言語では、ポルトガル語 [1]、スペイン語 [7]、日本語 [8] で統計的機械翻訳の枠組みでのテキスト平易化が行われている。本研究では我々は任意の言語でのテキスト平易化を実現することを目指し、Simple English Wikipedia を用いることなく English Wikipedia のみからテキスト平易化のための単言語パラレルコーパスを構築する。そして、Moses を用いた標準的な統計的機械翻訳の枠組みでのテキスト平易化を行ったときの性能をリーダビリティ（Flesch Reading Ease）と BLEU および SARI で評価する。

ⓒ 2016 Information Processing Society of Japan

90 ∼ 100	Very Easy
80 ∼ 89	Easy
70 ∼ 79	Fairly Easy
60 ∼ 69	Standard
50 ∼ 59	Fairly Difficult
30 ∼ 49	Difficult
0 ∼ 29	Very Difficult

表 1 リーダビリティとその解釈

文間類似度計算手法	*G vs. O*		*G+GP vs. O*	
	MaxF1	AUC	MaxF1	AUC
Zhu et al.	0.550	0.509	0.431	0.391
Coster and Kauchak	0.564	0.495	0.415	0.387
Hwang et al.	0.712	0.694	0.607	0.529
Kajiwara and Komachi	**0.717**	**0.730**	**0.638**	**0.618**

表 2 パラレルデータとノンパラレルデータの 2 値分類

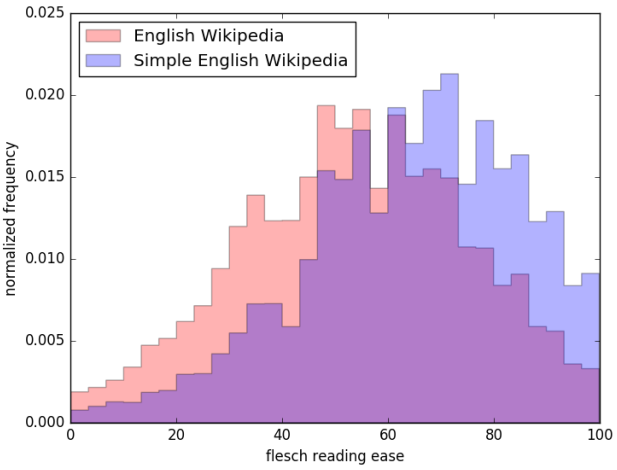

図 2 リーダビリティの分布

これまでに 4 種類の英語のテキスト平易化のための単言語パラレルコーパスが、English Wikipedia と Simple English Wikipedia を用いて構築されている。Zhu ら [2] は、文を TF-IDF ベクトルとして表現し、そのベクトル間のコサイン類似度を用いて初めてテキスト平易化のための単言語パラレルコーパス *4 を構築した。Coster and Kauchak [3] は、TF-IDF ベクトル間のコサイン類似度に加えて文の出現順序を考慮することで、より高精度にテキスト平易化コーパス *5 を構築した。しかし、Zhu らや Coster and Kauchak の手法では、異なる単語間の類似度を考慮していない。難解な表現から平易な表現への書き換えが頻繁に行われるテキスト平易化タスクにおいては、異なる単語間の類似度も適切に測定したい。Hwang ら [9] は、国語辞典の見出し語と定義文中の単語の共起を用いて、異なる単語間の類似度も考慮してテキスト平易化コーパス *6 を構築した。Kajiwara and Komachi [10] は、単語分散表現のアライメントに基づく文間類似度を用いて、辞書などの外部知識を使わずに異なる単語間の類似度を考慮してテキスト平易化コーパス *7 を構築した。本研究では最も性能の高い Kajiwara and Komachi の手法を用いて難解な文と平易な文の文アライメントを求める。

テキストの可読性を評価するリーダビリティ尺度としては、Flesch Reading Ease Formula [23] や Flesch-Kincaid Grade Level [24] がよく知られている。これらはいずれも、単語数と音節数を用いてリーダビリティを計算する。リーダビリティ尺度は言語ごとに開発されており、例えば日本語では李ら [25]、佐藤 [26]、藤田ら [27] の研究がある。本研究では 0 から 100 までのスコアに基づく Flesch Reading Ease Formula を用いて英文のリーダビリティを求める。

3. テキスト平易化コーパスの構築

本研究では English Wikipedia *8 からテキスト平易化のための単言語パラレルコーパスを自動構築する。Wikipedia からの本文抽出には WikiExtractor *9、テキストのトークナイズには NLTK 3.2.1 *10 をそれぞれ使用し、10 単語以上の文 (6,283,703 文) のみを対象とする。まず 3.1 節では、Flesch Reading Ease Formula を用いてリーダビリティを計算し、生コーパスを難解な文からなるコーパスと平易な文からなるコーパスに分解する。次に 3.2 節では、単語分散表現のアライメントに基づく文間類似度を用いて難解な文と平易な文の文アライメントを求め、テキスト平易化のための単言語パラレルコーパスを構築する。

3.1 リーダビリティに基づく難解な文と平易な文の分類

Flesch Reading Ease Formula では、α を単語数、β を 1 単語あたりの平均音節数として、文のリーダビリティを次のように定義する。

$$FleschReadingEase = 206.835 - 1.015\alpha - 84.6\beta \quad (1)$$

Flesch Reading Ease Formula を用いて計算された文のリーダビリティはスコアが大きいほど平易であることを意味し、各スコアは表 1 のように解釈できる。

図 2 に、English Wikipedia と Simple English Wikipedia の文のリーダビリティの分布を示す。English Wikipedia には、平易な文書である Simple English Wikipedia に比べてリーダビリティの低い難解な文が多いことがわかる。しかし、全ての文が難解なわけではないので、大規模なコー

*4 https://www.ukp.tu-darmstadt.de/data/sentence-simplification/simple-complex-sentence-pairs/
*5 http://www.cs.pomona.edu/~dkauchak/simplification/
*6 http://ssli.ee.washington.edu/tial/projects/simplification/
*7 https://github.com/tmu-nlp/sscorpus
*8 https://dumps.wikimedia.org/enwiki/20160501/
*9 https://github.com/attardi/wikiextractor/
*10 http://www.nltk.org/

ⓒ 2016 Information Processing Society of Japan

パスの中から平易な文を抽出することで、平易な文書に頼ることなく平易なコーパスを得ることができる。そして、平易な文を除いた難解なコーパスと、抽出された平易な文からなる平易なコーパスを利用することで、これまでEnglish Wikipedia と Simple English Wikipedia に対して使われてきた手法 [10] を適用して同様のテキスト平易化のための単言語パラレルコーパスを構築できる。

あるコーパスを分割して得られる難解なコーパスと平易なコーパスの組は、English Wikipedia と Simple English Wikipedia の組とは異なりコンパラブルコーパスではない。そのため、同義や含意の関係にある文対は少量しか得られないと考えられる。しかし、本研究ではフレーズベースの統計的機械翻訳を用いてテキスト平易化を行うため、以下の 3 つの理由でこの問題の影響は少なく、雑音の多い文対からでも重要な知識を獲得することができる。

- テキスト平易化は同一言語内の翻訳問題であるため、入力文に含まれる多くの単語をそのまま出力することができる（変換しないことが正解である）。そのため、異言語間の翻訳問題とは異なり、適切な変換対が少量しか得られないことが致命的な問題にはならない。
- フレーズベースの統計的機械翻訳では、フレーズ単位の変換対を学習する。難解なフレーズとその言い換えである平易なフレーズの組は、同義や含意の関係にある文対からだけではなく、類義の関係にある文対からも得ることができる。
- フレーズベースの統計的機械翻訳では、最終的に言語モデルによるリランキングを行うため、雑音の多いフレーズペアを獲得していても、平易な言い換えとして適切なフレーズペアをその中に含むことができれば適切な平易文が得られる。

図 2 のリーダビリティの分布から、English Wikipedia は 60 未満のリーダビリティを持つ文の割合が高く、Simple English Wikipedia は 60 以上のリーダビリティを持つ文の割合が高いことがわかる。そこで本研究では、English Wikipedia から抽出した 6,283,703 文を、60 未満のリーダビリティ（Very Difficult ～ Fairly Difficult）を持つ 3,689,227 文からなる難解なコーパスと 60 以上のリーダビリティ（Standard ～ Very Easy）を持つ 2,358,921 文からなる平易なコーパスに分割する。なお、0 未満または 100 を越えるリーダビリティを持つ 235,555 文（数百単語の長文や箇条書きなど）はリーダビリティを測定不能として除外する。

3.2 文間類似度に基づく文アライメント

我々は Kajiwara and Komachi [10] と同じく、Song and Roth [28] によって提案された単語分散表現のアライメントに基づく文間類似度（Maximum Alignment）を用いて難解な文と平易な文の文アライメントを得る。Maximum Alignment では、文 x に含まれる各単語 x_i に対して最も類似度が高い文 y 中の単語 y_j を選択し、それらの $|x|$ 個の単語の組み合わせについて計算した単語間類似度 $\phi(x_i, y_j)$ を平均して $\mathrm{STS}_{\mathrm{asym}}(x, y)$ を求める。$\mathrm{STS}_{\mathrm{asym}}(x, y)$ は非対称なスコアであるため、$\mathrm{STS}_{\mathrm{asym}}(x, y)$ と $\mathrm{STS}_{\mathrm{asym}}(y, x)$ の平均値を用いて対称な文間類似度 $\mathrm{STS}_{\mathrm{sym}}(x, y)$ を計算する。ここで、$\phi(x_i, y_j)$ は単語 x_i と単語 y_j の間の単語間類似度を表し、本研究ではコサイン類似度を用いる。

$$\mathrm{STS}_{\mathrm{asym}}(x, y) = \frac{1}{|x|} \sum_{i=1}^{|x|} \max_j \phi(x_i, y_j) \quad (2)$$

$$\mathrm{STS}_{\mathrm{sym}}(x, y) = \frac{1}{2}(\mathrm{STS}_{\mathrm{asym}}(x, y) + \mathrm{STS}_{\mathrm{asym}}(y, x)) \quad (3)$$

Hwang ら [9] は、English Wikipedia と Simple English Wikipedia から抽出された文対に対して *Good*（2 文間の意味が等しい）、*Good Partial*（一方の文が他方の文を含意する）、*Partial*（部分的に関連する）、*Bad*（無関係）の 4 種類のラベルを人手で付与した 67,853 文対（277 *Good*、281 *Good Partial*、117 *Partial*、67,178 *Bad*）のデータ [*6] を公開している。Kajiwara and Komachi [10] はこのデータセットを用いて、Maximum Alignment の文間類似度によってパラレルデータとノンパラレルデータの 2 値分類を行った。*Good* のラベル付きデータのみをパラレルデータとする場合（*G vs. O*）と *Good* および *Good Partial* の 2 つのラベル付きデータをパラレルデータとする場合（*G+GP vs. O*）の 2 つの設定で 2 値分類を行い、F1 の最大値（MaxF1）と Area Under the Curve（AUC）の 2 つの尺度で評価した結果が表 2 である。先行研究の文間類似度と比較して、Maximum Alignment が難解な文と平易な文の文アライメントのために有効な手法であることがわかる。

我々は公開されている学習済みの単語分散表現 [*11] を使用し、全ての難解な文と平易な文の組み合わせに対して Maximum Alignment を用いて文間類似度を計算した。ノイズを軽減するために単語間類似度が 0.5 以上の単語対のみを単語アライメントに使用し、文間類似度が 0.5 以上である 2,072,572 文対を抽出してテキスト平易化のための単言語パラレルコーパスを構築した。

3.3 テキスト平易化コーパスの概要

表 3 に、本研究で構築したテキスト平易化のための単言語パラレルコーパスの文間類似度ごとの例を示す。0.8 を越える類似度の高い文対には、難解な語句から平易な語句への言い換え（precipitation → rainfall）が見られたり、難解な文が平易な文の意味を含意するような例が見られる。0.6 から 0.8 の中程度の類似度を持つ文対は、同義関係や含意関係ではないが、関連する内容について書かれている。0.6 未満の低い類似度を持つ文対は、共通する語句や関連す

[*11] https://code.google.com/archive/p/word2vec/

文間類似度	難解な文	平易な文
0.99	Climate in this area has mild differences between highs and lows, and there is adequate **precipitation** year round.	Climate in this area has mild differences between highs and lows, and there is adequate **rainfall** year round.
0.88	The new German Empire included 25 states (three of them, Hanseatic cities) **and the imperial territory of Alsace-Lorraine**.	The new German Empire included 25 states, three of them Hanseatic cities.
0.77	In 1996, she received the Primetime Emmy Award for Outstanding Supporting Actress in a Comedy Series, an award she was nominated for on seven occasions.	In 2006 and 2008, she received Emmy nominations for Outstanding Supporting Actress in a Drama Series.
0.66	The album reached number two in the UK Albums Chart and was certified double platinum by the British Phonographic Industry (BPI).	The single reached number one in the UK and has been certified platinum by the BPI, selling 600,000 copies.
0.55	Bombed as a target of the Oil Campaign of World War II, Erfurt suffered only limited damage and was captured on 12 April 1945, by the US 80th Infantry Division.	During World War II the city suffered only some damage and was liberated by the British 8th army on 20 June 1944.

表 3 我々が構築したテキスト平易化コーパスの文間類似度ごとの例

る語句を含んではいるが、文全体としては強い関連がない。

本研究では English Wikipedia を用いてテキスト平易化のための単言語パラレルコーパスを構築した。しかし、Wikipedia は冒頭に概要を書き、本文で詳しく説明するという独特の構造を持つため、難解な文と平易な文のアライメントを求める本手法で実際には概要と本文のアライメントが得られているという可能性がある。概要と本文のアライメントが得られている場合、これは Wikipedia の文書構造に依存するため、任意の生コーパスに本手法が適用できるという主張が成立しない。そこで我々は、本研究で構築したコーパスからランダムに 5,000 文対を抽出し、それぞれの文対がどこから得られているのかを調査した。まず、文間類似度が $(0.9, 1.0]$ の 1,000 文対については、全ての文対が English Wikipedia 中の異なる記事から獲得されていた。次に、文間類似度が $(0.8, 0.9]$ の 1,000 文対については、1 文対のみが同一記事内の概要と本文のペアであり、残りの 999 文対は異なる記事から獲得されていた。続いて、文間類似度が $(0.7, 0.8]$ の 1,000 文対については、5 文対が同一記事から得られた文対であり、残りの 995 文対は異なる記事から獲得されていた。なお、同一記事から得られた文対のうち、3 文対は本文中の連続する 2 文からなる文対であり、1 文対は本文中の連続しない 2 文からなる文対であり、1 文対が概要と本文のペアであった。同様に、文間類似度が $(0.6, 0.7]$ の 1,000 文対については 945 文対、$(0.5, 0.6]$ の 1,000 文対については 930 文対が異なる記事から得られた文対であった。ほとんどの文対が異なる記事の組から得られていることから、提案手法が Wikipedia の文書構造に依存しない手法であることがわかる。

図 3 に、本研究で構築したコーパスの文間類似度とコーパスサイズの関係を示す。文間類似度が 0.94 以上で 10 万文、同様に 0.79 以上で 50 万文、0.64 以上で 100 万文、0.55 以上で 150 万文、0.51 以上で 200 万文となっている。

4. 統計的機械翻訳を用いたテキスト平易化

我々は生コーパスのみから構築したテキスト平易化のための単言語パラレルコーパスの有効性を調査するために、統計的機械翻訳の枠組みでテキスト平易化モデルを学習し、Simple English Wikipedia を使って構築された既存のテキスト平易化のための単言語パラレルコーパスを用いて学習したモデルとの比較を行う。

4.1 実験設定

我々はテキスト平易化を難解な文から平易な文への翻訳問題と考え、対数線形モデルを用いてモデル化する。

$$\begin{aligned}\hat{s} &= \underset{simple}{\operatorname{argmax}} P(simple|complex) \\ &= \underset{simple}{\operatorname{argmax}} P(complex|simple)P(simple) \\ &= \underset{simple}{\operatorname{argmax}} \sum_{m=1}^{M} \lambda_m h_m(simple, complex)\end{aligned} \tag{4}$$

対数線形モデルでは M 個の素性関数 $h_m(simple, complex)$ および各素性に対する重み λ_m を考え、翻訳確率 $P(simple|complex)$ をモデル化する。テキスト平易化の場合は、入力の難解な文に対して素性関数の重み付き線形和を最大化する平易な文 $\hat{s}$ を探索する問題を考える。素性関数としては、フレーズの平易化モデル $\log P(complex|simple)$ や言語モデル $\log P(simple)$ などを用いる。

我々はフレーズベースの統計的機械翻訳ツールである Moses 2.1 [21] を使用し、パラレルコーパスからの単語アライメントの獲得には GIZA++ [29] を用いた。言語モデルには、KenLM [30] を用いてパラレルコーパスの平易な文（比較手法では Simple English Wikipedia から抽出した文、提案手法では English Wikipedia のうち Flesch Reading Ease Formula で計算されるリーダビリティが 60 以上の文）か

ⓒ 2016 Information Processing Society of Japan

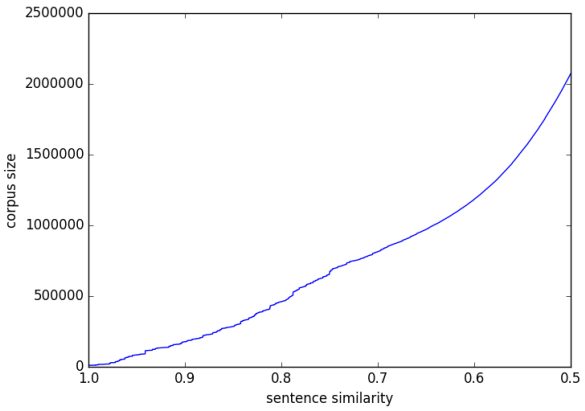

図 3 文間類似度とコーパスサイズ

	文対数	FRE	BLEU	SARI
Baseline (none)	0	54.5	99.4	25.9
Zhu Corpus	108,016	59.7	84.7	34.7
Coster Corpus	137,362	59.8	86.4	34.1
Hwang Corpus	284,738	61.0	81.3	34.5
Kajiwara Corpus	492,993	61.7	78.4	34.9
Xu (PPDB) [11]	(tuning) 2,000	67.9	72.4	37.9
Ours	100,000	54.9	94.9	29.1
Ours	500,000	55.3	92.7	31.1
Ours	1,000,000	56.9	88.0	33.7
Ours	1,500,000	58.2	83.2	34.4
Ours	2,000,000	59.2	79.1	34.1
Ours (all)	2,072,572	58.9	78.0	34.0

表 4 統計的機械翻訳の枠組みでのテキスト平易化の実験結果

ら 5-gram 言語モデルを構築した。テストデータには、Xu ら [11] によって公開されているマルチリファレンスのパラレルコーパス [*3] を使用した。これは、English Wikipedia から抽出された難解な 350 文に対して、それぞれ 8 人が平易な同義文を付与したものである。本研究では、このマルチリファレンスのパラレルコーパスを用いて Flesch Reading Ease（FRE）、BLEU および SARI [11] による自動評価を行った。なお、トレーニングデータからはテストデータに含まれる English Wikipedia の文を除外した。

4.2 実験結果および考察

表 4 に統計的機械翻訳の枠組みでのテキスト平易化の実験結果を示す。Baseline (none) は、書き換えを行わず入力文をそのまま出力するベースラインである。Zhu Corpus [2]、Coster Corpus [3]、Hwang Corpus [9] および Kajiwara Corpus [10] は、我々の実験設定と同じくフレーズベースの統計的機械翻訳ツール Moses によってテキスト平易化を行うが、難解なコーパス（English Wikipedia）と平易なコーパス（Simple English Wikipedia）の両方を用いて各研究で構築されたパラレルコーパスをトレーニングに使用している。Xu (PPDB) 法 [11] は、パラレルコーパスから単語アライメントによって平易化規則を獲得するのではなく、言い換え辞書である PPDB [18] のフレーズペアと言い換え確率を統計的機械翻訳のフレーズテーブルの代わりに使用している。また、Xu (PPDB) 法は SARI に最適化するために 2,000 文のマルチリファレンスデータでチューニングしている。Ours は我々の提案手法であり、文間類似度の高い順に 10 万文、50 万文、100 万文、150 万文、200 万文をそれぞれ抽出して比較している。

表 4 の FRE の結果から、我々の提案手法では入力文よりもリーダビリティの高い文を出力できていることがわかる。特に、文間類似度の上位 200 万文を用いてトレーニングした場合には、Simple English Wikipedia を使ってトレーニングした場合と同等のリーダビリティを持つ文を出力することができた。

FRE が出力文のみを用いてリーダビリティのみを評価するのに対して、BLEU は出力文とリファレンスの両方を、SARI は入力文と出力文とリファレンスの全てを用いて文法や意味も評価する。Xu ら [11] は、BLEU が Grammar や Meaning の観点で人手評価との相関が高く、SARI が Simplicity の観点で人手評価との相関が高いことを報告している。我々の提案手法では、文間類似度の上位 150 万文を用いてトレーニングした場合に SARI が最大となった。このとき、BLEU は Simple English Wikipedia を使ってトレーニングした場合や大規模な言い換え知識を用いてチューニングした場合と同等であり、入力文の文法や意味を保持した文を出力することができた。SARI でも、我々の提案手法は Simple English Wikipedia を用いる比較手法と同等の性能を発揮した。

これらの結果から、生コーパスのみを用いて構築したパラレルコーパスでトレーニングする提案手法が、フレーズベースの統計的機械翻訳を用いるテキスト平易化において、比較手法と同等の性能を発揮することがわかった。この手法は、平易に書かれた大規模なコーパスあるいは大規模な言い換え知識が利用できない言語でのテキスト平易化においても効果を発揮することが期待できる。

表 5 に、統計的機械翻訳の枠組みでのテキスト平易化の例を示す。我々の提案手法は、Reference1 と同様に不要な表現 “both” を省略し、Reference2 や 3 と同様に難解な表現 “numerous” を平易な表現 “many” に言い換えた。比較手法の中には、難解な表現 “extremely” から平易な表現 “very” への言い換えも見られる一方で、“world” や “speaking” など必要以上の省略も見られる。

5. おわりに

本研究では、平易に書かれた大規模なコーパスや言い換え知識などの外部知識に頼らず、生コーパスのみを用いてリーダビリティと単語分散表現のアライメントに基づく文

ⓒ 2016 Information Processing Society of Japan

Input	Offenbach's numerous operettas, such as *Orpheus in the Underworld*, and *La belle Hélène*, were extremely popular in both France and the English-speaking world during the 1850s and 1860s.
Reference1	Offenbach's ~~numerous~~ operettas, such as *Orpheus in the Underworld*~~,~~ and *La belle Hélène*, were ~~extremely~~ **very** popular in ~~both~~ France and the English-speaking world during the 1850**'**s and 1860**'**s.
Reference2	Offenbach's ~~numerous~~ **many** operettas, such as *Orpheus in the Underworld*, and *La belle Hélène*, were ~~extremely~~ **very** popular in ~~both~~ France and **in** the English-speaking world during the 1850s and 1860s.
Reference3	Offenbach's ~~numerous~~ **many** operattas, ~~such as~~ **including** *Orpheus in the Underworld*~~,~~ and *La belle ~~Hélène~~* **Helene**, were ~~extremely~~ **very** popular in ~~both~~ France and the English-speaking world ~~during~~ **in** the 1850s and 1860s.
Reference4	During the 1850s and 1860s, *Orpheus in the Underworld*~~,~~ and *La belle Hélène*~~,~~ were ~~extremely popular~~ **some of the famous** Offenbach's ~~numerous~~ operettas~~,~~ in both France and the English-speaking world.
Reference5	Offenbach's numerous operettas, such as *Orpheus in the Underworld*, and *La belle Hélène*, were ~~extremely~~ **very** popular in both France and the English-speaking world during the 1850s and 1860s.
Reference6	Offenbach's numerous operettas, such as *Orpheus in the Underworld*, and *La belle Hélène*, were ~~extremely~~ **very great** popular in both France and the English-speaking world during the 1850s and 1860s.
Reference7	Offenbach's ~~numerous~~ **a great number of** operettas, such as *Orpheus in the Underworld*, and *La ~~belle Hélène~~* **beautiful woman Helene**, were ~~extremely popular~~ **greatly pleasing to all** in both France and the English-~~speaking world~~**talking earth** during the 1850s and 1860s.
Reference8	Offenbach's ~~numerous~~ **a great number of** operettas, such as *Orpheus in the Underworld*, and *La ~~belle Hélène~~* **beautiful woman Helene**, were ~~extremely popular~~ **greatly pleasing to all** in both France and the English-~~speaking world~~**talking earth** during the 1850s and 1860s~~.~~
Zhu Corpus	Offenbach's ~~numerous~~ **many** operettas, such as *Orpheus in the Underworld*, and *La belle Hélène*, were extremely popular ~~in~~ both France and the English-speaking ~~world during~~ **in** the 1850s and 1860s.
Coster Corpus	Offenbach's ~~numerous~~ **many** operettas, such as *Orpheus in the Underworld*, and *La belle Hélène*, were extremely popular ~~in~~ both **in** France and the English-speaking ~~world during~~ **in** the 1850s and 1860s.
Hwang Corpus	Offenbach's ~~numerous~~ **many** operettas, such as *Orpheus in the Underworld*, and *La belle Hélène*, were extremely popular in both France and the English-speaking world during the 1850s and 1860s.
Kajiwara Corpus	Offenbach's ~~numerous~~ **many** operettas, such as *Orpheus in the Underworld*, and *La belle Hélène*, were ~~extremely~~ **very** popular in both France and the English-speaking world during the 1850s and 1860s.
Xu (PPDB)	Offenbach's ~~numerous~~ **many** operettas, such as *Orpheus in the Underworld*, and ~~*La*~~ **The** *belle Hélène*, were ~~extremely~~ **very** popular in both France and the English-~~speaking~~ world ~~during~~ **in** the 1850s and 1860s.
Ours	Offenbach's ~~numerous~~ **many** operettas, such as *Orpheus in the Underworld*, and *La belle Hélène*, were extremely popular in ~~both~~ France and the English-speaking world during the 1850s and 1860s.

表 5 統計的機械翻訳の枠組みでのテキスト平易化の例

間類似度によってテキスト平易化のための単言語パラレルコーパスを自動構築した。また、統計的機械翻訳の枠組みでのテキスト平易化の実験によって、平易な大規模コーパスを用いてトレーニングする先行研究と同様に、入力文をより平易な同義文に変換できることを確認した。

これまでは、豊富な言語資源が存在する英語を中心にテキスト平易化の研究が進められてきたが、生コーパスは英語以外の多くの言語でも大規模に利用できるので、今後は本研究の成果をもとに、文のリーダビリティを測定可能な任意の言語でテキスト平易化が実現できるだろう。

参考文献

[1] Specia, L.: Translating from Complex to Simplified Sentences, *Lecture Notes in Computer Science*, Vol. 6001, pp. 30–39 (2010).
[2] Zhu, Z., Bernhard, D. and Gurevych, I.: A Monolingual Tree-based Translation Model for Sentence Simplification, *Proceedings of the 23rd International Conference on Computational Linguistics*, Beijing, China, pp. 1353–1361 (2010).
[3] Coster, W. and Kauchak, D.: Simple English Wikipedia: A New Text Simplification Task, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Portland, Oregon, USA, pp. 665–669 (2011).
[4] Coster, W. and Kauchak, D.: Learning to Simplify Sentences Using Wikipedia, *Proceedings of the Workshop on Monolingual Text-To-Text Generation*, Portland, Oregon, USA, pp. 1–9 (2011).
[5] Wubben, S., van den Bosch, A. and Krahmer, E.: Sentence Simplification by Monolingual Machine Translation, *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, Jeju Island, Korea, pp. 1015–1024 (2012).
[6] Štajner, S., Bechara, H. and Saggion, H.: A Deeper Exploration of the Standard PB-SMT Approach to Text Simplification and its Evaluation, *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, Beijing, China, pp. 823–828 (2015).

ⓒ 2016 Information Processing Society of Japan

[7] Štajner, S., Calixto, I. and Saggion, H.: Automatic Text Simplification for Spanish: Comparative Evaluation of Various Simplification Strategies, *Proceedings of the International Conference Recent Advances in Natural Language Processing*, Hissar, Bulgaria, pp. 618–626 (2015).

[8] Goto, I., Tanaka, H. and Kumano, T.: Japanese News Simplification: Task Design, Data Set Construction, and Analysis of Simplified Text, *Proceedings of MT Summit XV*, Miami, Florida, USA, pp. 17–31 (2015).

[9] Hwang, W., Hajishirzi, H., Ostendorf, M. and Wu, W.: Aligning Sentences from Standard Wikipedia to Simple Wikipedia, *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Denver, Colorado, USA, pp. 211–217 (2015).

[10] Kajiwara, T. and Komachi, M.: Building a Monolingual Parallel Corpus for Text Simplification Using Sentence Similarity Based on Alignment between Word Embeddings, *Proceedings of the 26th International Conference on Computational Linguistics*, Osaka, Japan, to appear (2016).

[11] Xu, W., Napoles, C., Pavlick, E., Chen, Q. and Callison-Burch, C.: Optimizing Statistical Machine Translation for Text Simplification, *Transactions of the Association for Computational Linguistics*, Vol. 4, pp. 401–415 (2016).

[12] Xu, W., Callison-Burch, C. and Napoles, C.: Problems in Current Text Simplification Research: New Data Can Help, *Transactions of the Association for Computational Linguistics*, Vol. 3, pp. 283–297 (2015).

[13] Agirre, E., Cer, D., Diab, M. and Gonzalez-Agirre, A.: SemEval-2012 Task 6: A Pilot on Semantic Textual Similarity, **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, Montréal, Canada, pp. 385–393 (2012).

[14] Agirre, E., Cer, D., Diab, M., Gonzalez-Agirre, A. and Guo, W.: *SEM 2013 shared task: Semantic Textual Similarity, *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, Atlanta, Georgia, USA, pp. 32–43 (2013).

[15] Agirre, E., Banea, C., Cardie, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Guo, W., Mihalcea, R., Rigau, G. and Wiebe, J.: SemEval-2014 Task 10: Multilingual Semantic Textual Similarity, *Proceedings of the 8th International Workshop on Semantic Evaluation*, Dublin, Ireland, pp. 81–91 (2014).

[16] Agirre, E., Banea, C., Cardie, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Guo, W., Lopez-Gazpio, I., Maritxalar, M., Mihalcea, R., Rigau, G., Uria, L. and Wiebe, J.: SemEval-2015 Task 2: Semantic Textual Similarity, English, Spanish and Pilot on Interpretability, *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, Denver, Colorado, USA, pp. 252–263 (2015).

[17] Agirre, E., Banea, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Mihalcea, R., Rigau, G. and Wiebe, J.: SemEval-2016 Task 1: Semantic Textual Similarity, Monolingual and Cross-Lingual Evaluation, *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, San Diego, California, pp. 497–511 (2016).

[18] Ganitkevitch, J., Van Durme, B. and Callison-Burch, C.: PPDB: The Paraphrase Database, *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Atlanta, Georgia, USA, pp. 758–764 (2013).

[19] Pavlick, E., Rastogi, P., Ganitkevitch, J., Van Durme, B. and Callison-Burch, C.: PPDB 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification, *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, Beijing, China, pp. 425–430 (2015).

[20] Pavlick, E. and Callison-Burch, C.: Simple PPDB: A Paraphrase Database for Simplification, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Berlin, Germany, pp. 143–148 (2016).

[21] Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A. and Herbst, E.: Moses: Open Source Toolkit for Statistical Machine Translation, *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, Prague, Czech Republic, pp. 177–180 (2007).

[22] Papineni, K., Roukos, S., Ward, T. and Zhu, W.-J.: Bleu: a Method for Automatic Evaluation of Machine Translation, *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, Pennsylvania, USA, pp. 311–318 (2002).

[23] Flesch, R.: A new readability yardstick, *Journal of Applied Psychology*, Vol. 32, pp. 221–233 (1948).

[24] Kincaid, J. P., Fishburne Jr., R. P., Rogers, R. L. and Chissom, B. S.: Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel, *Technical report, Defence Technical Information Center (DTIC) Document*, pp. 8–75 (1975).

[25] 李在鎬，長谷部陽一郎，柴崎秀子：読解教育支援のためのリーダビリティー測定ツールについて，言語処理学会第 15 回年次大会発表論文集，pp. 713–716 (2009).

[26] 佐藤理史：均衡コーパスを規範とするテキスト難易度判定，情報処理学会論文誌，Vol. 52, No. 4, pp. 1777–1789 (2011).

[27] 藤田早苗，藤野昭典，小林哲生：教科書を基準とする難易度推定，人工知能学会第 29 回全国大会, 2N3-5，pp. 1–4 (2015).

[28] Song, Y. and Roth, D.: Unsupervised Sparse Vector Densification for Short Text Similarity, *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Denver, Colorado, USA, pp. 1275–1280 (2015).

[29] Och, F. J. and Ney, H.: A Systematic Comparison of Various Statistical Alignment Models, *Computational Linguistics*, Vol. 29, No. 1, pp. 19–51 (2003).

[30] Heafield, K.: KenLM: Faster and Smaller Language Model Queries, *Proceedings of the Sixth Workshop on Statistical Machine Translation*, Edinburgh, Scotland, pp. 187–197 (2011).

ⓒ 2016 Information Processing Society of Japan