Vol.2016-NL-227 No.12
2016/7/30

# 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た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、テキスト平易化のための単言語パラレルコーパスは、英語や独語など限られた言語でしか整備されていない。そこで本稿では、単語の分散表現に基づいて計算される文間類似度を用いて、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手法を提案する。我々は難解な文と平易な文からなる任意の文対に対して、一方の文中の各単語に対して最も類似度の高い他方の文中の単語を割り当てる多対一の単語アライメントを考え、それらの単語間類似度の平均値によって文間類似度を定義する。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しないため、任意の言語に適用できる。実験の結果、我々の提案手法は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の自動構築タスクにおいて state-of-the-art を更新した。また、統計的機械翻訳の枠組みを用いたテキスト平易化の実験結果も、我々の提案手法によって構築されたコーパスが既存のテキスト平易化のためのコーパスよりも優れていることを示した。本研究ではテキスト平易化を対象にしたが、単言語パラレルコーパスは言い換えや文圧縮などの分野でも有用な言語資源である。

## 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、テキスト平易化のための単言語パラレルコーパスは、英語 [2,3,9,10]、ポルトガル語 [11]、スペイン語 [12]、デンマーク語 [13]、ドイツ語 [14]、イタリア語 [15]、日本語 [8] の 7 言語でしか整備されていない。なお、このうち公開されているコーパスは英語のもののみである。そこで本研究では、ラベル（文間類似度などのスコアや同義・含意などのラベル）付きデータや辞書などの外部知識を必要とせず低コストに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 [*1] を提案する。

本研究では、難解なテキストと平易なテキストからなるコンパラブルコーパスが与えられたときに、それぞれの難解な文と平易な文の組に対して単語分散表現のアライメントに基づく文間類似度を計算する。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することで、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。本研究の概要を図 1 に示す。このような単言語パラレルコーパスを用いて統計的機械翻訳モデルを学習することによって、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。

我々は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の自動構築タスクにおける評価用データセット [*2] を用いて、提案手法の内的評価と外的評価を行った。この評価用データセットは、難解な文と平易な文の組に人手でパラレル（同義である）かノンパラレル（同義でない）かのラベルが付与されたものである。内的評価として、我々の提案手法で計算した文間類似度によって評価用データセット中の各文対をパラレルデータとノンパラレルデータに 2 値分類したところ、F1 スコアの向上を確認できた。また外的評価として、我々が構築したコーパスと先行研究で

[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] https://github.com/tmu-nlp/sscorpus
[*2] http://ssli.ee.washington.edu/tial/projects/simplification/

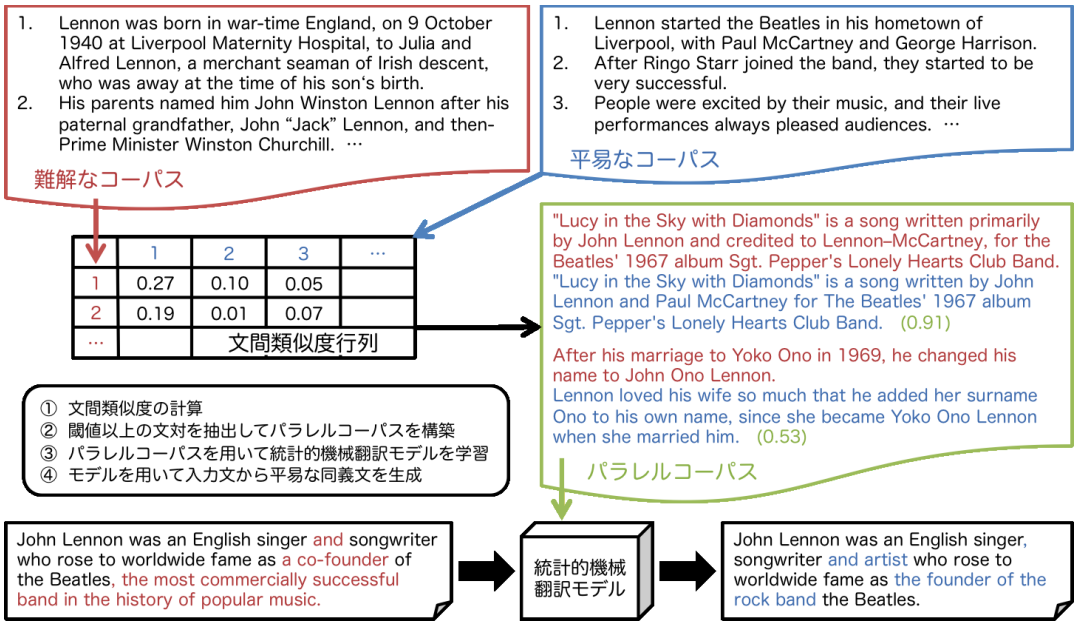


**図 1** 単言語パラレルコーパスの構築および統計的機械翻訳の枠組みでのテキスト平易化

公開されているコーパスを用いてそれぞれフレーズベースの統計的機械翻訳モデルを学習したところ、評価用データセットのパラレルデータに対するテキスト平易化においてBLEU スコアの向上を確認できた。

本研究の貢献は以下の 3 つである。

- 提案手法はパラレルデータとノンパラレルデータの分類タスクにおいて先行研究よりも F1 で 3.1 ポイントの改善（0.607 → 0.638）を行い、高精度に単言語パラレルコーパスを構築することができる。
- 本研究で構築したコーパスを用いて学習したテキスト平易化モデルは、既存のコーパスで学習したモデルよりも BLEU で 3.2 ポイントの性能改善（44.3 → 47.5）を行うことができる。
- 提案手法は文間類似度の計算にラベル付きデータや辞書などの外部知識を必要としないため、低コストにテキスト平易化コーパスを自動構築できる。

## 2. 関連研究

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 [*3] と Simple English Wikipedia [*4] 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] 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] が盛んに研究されている。Coster and Kauchak [3] は、標準的なフレーズベースの統計的機械翻訳ツール Moses [16] と自動評価尺度 BLEU [17] を用いて、統計的機械翻訳の枠組みでのテキスト平易化を行った。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。Štajner ら [6] は、Moses と BLEU を用いた統計的機械翻訳の枠組みでのテキスト平易化について、学習に使用する単言語パラレルコーパスの量や質を変化させて考察し、適度な文間類似度（0.5 から 0.6）を持った単言語パラレルコーパスを用いた学習がテキスト平易化タスクにおいて効果的であることを示した。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、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において、BLEU で評価される性能を改善する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。

これまでに 3 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 [*5] 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 [*6] 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することができない。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

[*3] http://en.wikipedia.org
[*4] http://simple.wikipedia.org
[*5] https://www.ukp.tu-darmstadt.de/data/sentence-simplification/simple-complex-sentence-pairs/
[*6] http://www.cs.pomona.edu/~dkauchak/simplification/

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 [*2] を構築した。本研究では我々は、単語分散表現を用いることで辞書などの外部知識に頼らず異なる単語間の類似度を考慮する低コストなテキスト平易化コーパスの自動構築手法を提案する。

文間の意味的類似度を計算する Semantic Textual Similarity (STS) タスク [18] では、word2vec [19] などの単語の分散表現の成功を受け、異なる単語間の類似度を考慮する手法が提案されている。SemEval-2015 の STS タスク [20] では、word2vec の単語分散表現や PPDB [21] の言い換えを用いた単語アライメントに基づく教師あり学習の手法 [22] が最高精度を達成している。同じく word2vec の単語分散表現に基づく教師なしの文間類似度計算手法 [23–25] も提案されている。文間類似度のラベル付きデータを必要としないこれらの教師なし手法は、テキスト平易化のための単言語パラレルコーパスの自動構築にも応用できる。

## 3. 単語分散表現のアライメントに基づく文間類似度の計算

我々はテキスト平易化のための単言語パラレルコーパスの自動構築のために、単語分散表現のアライメントに基づく 4 種類の文間類似度の計算手法を提案する。 3.1 節から 3.3 節で説明する手法は、Song and Roth [24] 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の計算手法を本タスクに応用するものである。 3.4 節の Word Mover's Distance [25] も、単語分散表現のアライメントに基づく文間類似度の計算に用いることができる。

### 3.1 Average Alignment

文 $x$ と文 $y$ の間の全ての単語の組み合わせについて単語間類似度を計算し、それらの $|x||y|$ 個の単語間類似度を平均して文間類似度 $\mathrm{STS}_{\mathrm{ave}}(x,y)$ を求める。

$$\mathrm{STS}_{\mathrm{ave}}(x,y) = \frac{1}{|x||y|}\sum_{i=1}^{|x|}\sum_{j=1}^{|y|}\phi(x_i,y_j) \qquad (1)$$

ここで、$x_i$ および $y_j$ は、それぞれ文 $x$ および文 $y$ に含まれる単語を表す。また、$\phi(x_i,y_j)$ 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。

### 3.2 Maximum Alignment

3.1 節で説明した Average Alignment は単語分散表現に基づく文間類似度の計算方法として直感的であるが、同義の文対 $(x,y)$ を考えても全ての単語の組み合わせについて単語間類似度 $\phi(x_i,y_j)$ が高くなるとは考えにくく、多くの単語間類似度は 0 に近い値を取るノイズになると考えられる。そこで文 $x$ 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のみ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 $\mathrm{STS}_{\mathrm{max}}(x,y)$ を計算する。

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

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

### 3.3 Hungarian Alignment

Average Alignment および Maximum Alignment は、それぞれ多対多および多対一の単語アライメントに基づく文間類似度の計算方法と考えることができる。本節では $x$ および $y$ の 2 文を単語をノードとする 2 部グラフとして考え、一対一の単語アライメントに基づく文間類似度の計算方法を定義する。この 2 部グラフは、単語間類似度 $\phi(x_i,y_j)$ を重みとする重み付きの辺を持つ重み付き完全 2 部グラフである。この完全 2 部グラフの最大マッチングを求めることで、単語間類似度の総和を最大化する一対一の単語アライメントを得ることができる。2 部グラフの最大マッチング問題は、Hungarian 法 [26] を用いて解くことができる。そこで文 $x$ に含まれる各単語 $x_i$ に対して Hungarian 法によって文 $y$ 中の単語 $h(x_i)$ を選択し、それらの $\min(|x|,|y|)$ 個の単語の組み合わせについて計算した単語間類似度を平均して文間類似度 $\mathrm{STS}_{\mathrm{hun}}(x,y)$ を求める。

$$\mathrm{STS}_{\mathrm{hun}}(x,y) = \frac{1}{\min(|x|,|y|)}\sum_{i=1}^{\min(|x|,|y|)}\phi(x_i,h(x_i)) \quad (4)$$

### 3.4 Word Mover's Distance

Word Mover's Distance [25] も、単語の分散表現を用いた多対多の単語アライメントに基づく文間類似度の計算に用いることができる。Word Mover's Distance は、文 $x$ から文 $y$ へと単語を輸送する輸送問題を解く Earth Mover's Distance [27] の特殊な場合に相当する。

$$\mathrm{STS}_{\mathrm{wmd}}(x,y) = 1 - \mathrm{WMD}(x,y) \qquad (5)$$

$$\mathrm{WMD}(x,y) = \min\sum_{i=1}^{n}\sum_{j=1}^{n}\mathcal{A}_{ij}\psi(x_i,y_j) \qquad (6)$$

$$\text{subject to : } \sum_{j=1}^{n}\mathcal{A}_{ij} = \frac{1}{|x|}freq(x_i)$$

$$\sum_{i=1}^{n}\mathcal{A}_{ij} = \frac{1}{|y|}freq(y_j)$$

ここで、$\psi(x_i,y_j)$ は単語 $x_i$ と単語 $y_j$ の間の単語間非類似度（距離）を表し、本研究ではユークリッド距離を用いる。また、$\mathcal{A}_{ij}$ は文 $x$ 中の単語 $x_i$ から文 $y$ 中の単語 $y_j$ への輸送量を表す行列であり、$n$ は語彙数、$freq(x_i)$ は文 $x$ 中での単語 $x_i$ の出現頻度である。

## 4. 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

我々は文間類似度を用いた文アライメント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構築し、単語分散表現のアライメントに基づく文間類似度計算手法の有効性を検証する。まず 4.1 節では、English Wikipedia と Simple English Wikipedia から抽出された文対に対して人手でラベル付けされたデータを用いてパラレルデータとノンパラレルデータの 2 値分類を行い、単語分散表現のアライメントに基づく文間類似度の有効性を示す。次に 4.2 節では、提案手法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構築し、定性的な評価を行う。そして 4.3 節では、我々の構築したコーパスと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それぞれテキスト平易化モデルを学習し、統計的機械翻訳の枠組みでの評価によって提案手法の有効性を示す。

### 4.1 文間類似度によるパラレルデータとノンパラレルデータの 2 値分類

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 段階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 [*2] を公開している。我々はこ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の自動構築タスクにおける評価用データセットを用いて、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 値分類を行う。*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 つの尺度で評価する。このパラレルデータとノンパラレルデータの 2 値分類の性能が高いことは、その文間類似度の計算手法がテキスト平易化のための単言語パラレルコーパスの自動構築タスクにとって有用であるということを意味する。

パラレルデータとノンパラレルデータの 2 値分類の結果を表 1 に示す。上段の 3 つの手法はテキスト平易化のための単言語パラレルコーパス構築の先行研究であり、下段の 5 つの手法は単語分散表現に基づく文間類似度計算手法である。Additive Embeddings は、単語アライメントを使用しない比較手法 [23] であり、単語の分散表現を足し合わせることによって文の分散表現を構成し、コサイン類似度によって文間類似度を計算した。本実験では、単語分散表現に基づく文間類似度計算のために、公開されている学習済みの単語分散表現 [*7] を用いた。

| 文間類似度計算手法 | *G vs. O* | | *G+GP vs. O* | |
|---|---|---|---|---|
| | MaxF1 | AUC | MaxF1 | AUC |
| Zhu et al. [9] | 0.550 | 0.509 | 0.431 | 0.391 |
| Coster and Kauchak [9] | 0.564 | 0.495 | 0.415 | 0.387 |
| Hwang et al. [9] | 0.712 | 0.694 | 0.607 | 0.529 |
| Additive Embeddings | 0.691 | 0.695 | 0.518 | 0.487 |
| Average Alignment | 0.419 | 0.312 | 0.391 | 0.297 |
| Maximum Alignment | 0.717 | 0.730 | **0.638** | **0.618** |
| Hungarian Alignment | 0.524 | 0.414 | 0.354 | 0.275 |
| Word Mover's Distance | **0.724** | **0.738** | 0.531 | 0.499 |

**表 1** パラレルデータとノンパラレルデータの 2 値分類

Average Alignment、Maximum Alignment および Hungarian Alignment については、単語アライメントのノイズ除去を行った。 3.2 節でも述べたが、同義の文対 $(x, y)$ を考えても全ての単語対について単語間類似度が高くなるとは考えにくく、どの単語アライメントの手法を用いても単語間類似度が低いにも関わらず対応付けられてしまう単語対が存在する。このようなノイズとなる単語対の影響を抑えるため、我々は $\phi(x_i, y_j) > \theta$ の単語間類似度を持つ単語対 $(x_i, y_j)$ のみを用いて単語アライメントを行った。この閾値 $\theta$ は MaxF1 を最大化するように選択し、Average Alignment については *G vs. O* の分類時に 0.89、*G+GP vs. O* の分類時に 0.95、Maximum Alignment については *G vs. O* の分類時に 0.28、*G+GP vs. O* の分類時に 0.49、Hungarian Alignment については *G vs. O* の分類時に 0.98、*G+GP vs. O* の分類時に 0.98 を採用した。

表 1 の実験結果から、*Good* とその他の 2 値分類においては多対多の単語アライメントに基づく提案手法である Word Mover's Distance が最も高い性能を示した。また、*Good+Good Partial* とその他の 2 値分類においては多対一の単語アライメントに基づく提案手法である Maximum Alignment が最も高い性能を示した。なお、Maximum Alignment は *Good* とその他の 2 値分類においてもテキスト平易化のための単言語パラレルコーパス構築の先行研究よりも高い性能を示した。

図 2 および図 3 に、パラレルデータとノンパラレルデータの 2 値分類における Precision-Recall 曲線を示す。図 3 の *Good+Good Partial* とその他の 2 値分類において、赤で示す Maximum Alignment が他の単語分散表現に基づく文間類似度計算手法よりも高い性能を示すことがわかる。

テキスト平易化では、難解な表現から平易な表現への言い換えだけではなく、文中の重要ではない難解な表現を省略することによって読みやすい短文を生成することも多い [10]。そこで、テキスト平易化のための単言語パラレルコーパスには、対応する難解な文と平易な文が同義である *Good* の文対だけでなく、難解な文が平易な文の意味を含意する *Good Partial* の文対も含めることが重要である。そ

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

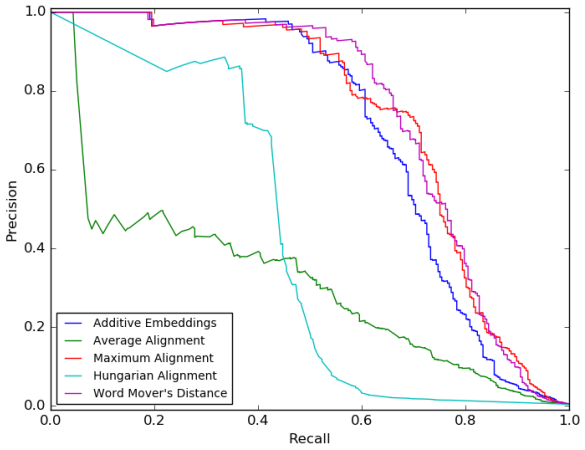


**図 2** *Good* とその他の 2 値分類における PR 曲線

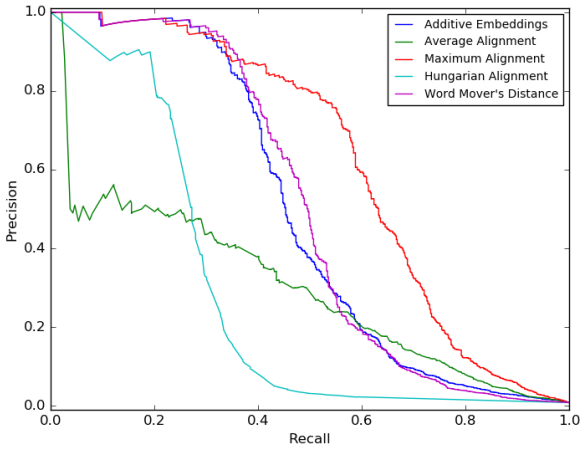


**図 3** *Good+Good Partial* とその他の 2 値分類における PR 曲線

のため、*Good+Good Partial* とその他の 2 値分類において最も高い性能を示す Maximum Alignment が、テキスト平易化のための単言語パラレルコーパス構築に最も適した文間類似度の計算手法であると言える。

### 4.2 英語のテキスト平易化コーパスの構築

我々は 4.1 節で最も高い性能を示した Maximum Alignment を用いて、English Wikipedia（normal）[*8] と Simple English Wikipedia（simple）[*9] 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した。まずタイトルが完全一致する normal と simple の記事を集めて、126,725 の文書対を得た。これらの文書対に対して WikiExtractor [*10] を用いた本文抽出と NLTK 3.2.1 [*11] を用いたトークナイズを行ったところ、normal と simple の平均文長はそれぞれ 25.1 語および 16.9 語であった。

これらの文書対ごとに、全ての normal 文と simple 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。表 1 の実験結果（MaxF1）から単語間類似度および文間類似度の閾値を設定し、単語間類似度が 0.49 以上である場合のみ単語アライメントを行い、文間類似度が 0.53 以上である場合のみ文アライメントを行った。こうして、126,725 文書対から 492,993 文対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した。

表 4 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。0.9 以上の類似度を持つ文対には、同義表現の言い換え（purchased → bought）が見られる。0.7 以上の類似度を持つ文対には、重要ではない表現の削除（such as ...）が見られる。0.7 未満の類似度を持つ文対には、表層としては数単語しか一致していないような言い換え・含意・類義文が見られる。

[*8] https://dumps.wikimedia.org/enwiki/20160501/
[*9] https://dumps.wikimedia.org/simplewiki/20160501/
[*10] https://github.com/attardi/wikiextractor/
[*11] http://www.nltk.org/

### 4.3 既存のテキスト平易化コーパスとの比較

我々は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、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。本研究では、Zhu ら [2]、Coster and Kauchak [3]、Hwang ら [9] の、English Wikipedia と Simple English Wikipedia から構築された既存の 3 つのテキスト平易化コーパスと我々のコーパスを比較する。

我々はテキスト平易化を normal 文から simple 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。

$$\begin{aligned}\hat{s} &= \underset{simple}{\operatorname{argmax}} P(simple|normal) \\ &= \underset{simple}{\operatorname{argmax}} P(normal|simple)P(simple) \\ &= \underset{simple}{\operatorname{argmax}} \sum_{m=1}^{M} \lambda_m h_m(simple, normal)\end{aligned} \tag{7}$$

対数線形モデルでは $M$ 個の素性関数 $h_m(simple, normal)$ および各素性に対する重み $\lambda_m$ を考え、翻訳確率 $P(simple|normal)$ をモデル化する。テキスト平易化の場合は、入力の normal 文に対して素性関数の重み付き線形和を最大化する simple 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 $\log P(normal|simple)$ や言語モデル $\log P(simple)$ などを用いる。それぞれのテキスト平易化のための単言語パラレルコーパスのうち、無作為抽出された 500 文対は MERT [28] によるチューニングのために使用し、残りの全ての文対を平易化モデルのトレーニングのために使用した。デコーダにはフレーズベースの統計的機械翻訳ツールである Moses [16]、トレーニングデータ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 Simple English Wikipedia [*9] 全体から 5-gram 言語モデルを構築した。テストデータには Hwang らによって公開されている人

| テキスト平易化コーパス | 文対数 | 語彙数 (normal) | 語彙数 (simple) | 平均文長 (normal) | 平均文長 (normal) | BLEU |
|---|---|---|---|---|---|---|
| Zhu et al. | 100,000 | 173,463 | 143,030 | 21.2 | 17.4 | 41.8 |
| Zhu et al. (All) | 107,516 | 181,459 | 149,643 | 21.2 | 17.4 | 42.0 |
| Coster and Kauchak | 100,000 | 112,744 | 102,418 | 23.7 | 21.1 | 43.8 |
| Coster and Kauchak (All) | 136,862 | 132,567 | 120,620 | 23.6 | 21.1 | 44.3 |
| Hwang et al. | 100,000 | 117,474 | 103,427 | 25.3 | 21.2 | 42.9 |
| Hwang et al. (Good) | 154,305 | 152,419 | 133,825 | 25.2 | 21.2 | 42.9 |
| Hwang et al. | 200,000 | 175,416 | 145,773 | 25.6 | 20.5 | 43.1 |
| Hwang et al. (Good+Partial) | 284,238 | 212,138 | 164,979 | 26.0 | 19.8 | 43.9 |
| Hwang et al. | 300,000 | 217,699 | 167,945 | 26.1 | 19.7 | 42.9 |
| Hwang et al. (All) | 391,116 | 248,510 | 184,521 | 26.5 | 19.4 | 43.1 |
| Ours | 100,000 | 122,390 | 112,670 | 23.9 | 21.8 | 43.2 |
| Ours | 200,000 | 180,776 | 151,815 | 24.7 | 20.1 | 45.7 |
| Ours | 300,000 | 219,628 | 174,576 | 25.2 | 19.0 | 46.4 |
| Ours (All) | 492,493 | 274,775 | 198,043 | 25.3 | 17.9 | **47.5** |

**表 2** 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との比較

| | |
|---|---|
| Input | Mozart's Clarinet Concerto and Clarinet Quintet are both in A major, and generally Mozart was more likely to use clarinets in A major than in any other key besides E-flat major. |
| Reference | Mozart used clarinets in A major often. |
| Zhu et al. | Mozart's Clarinet Concerto and Clarinet Quintet are both in A major, and generally Mozart **which he** more likely to use clarinets in A major than in any other key besides E-flat major. |
| Coster and Kauchak | **Mozart was** Clarinet Concerto and Clarinet Quintet are both in A major, and **Mozart used clarinets in A major often**. |
| Hwang et al. | Mozart's Clarinet Concerto and Clarinet Quintet are both in A major, and generally Mozart was more likely to use clarinets in A major than in any other key besides E-flat major. |
| Ours | Mozart's Clarinet Concerto and Clarinet Quintet are both in A major, and **Mozart used clarinets in A major often**. |

**表 3** 統計的機械翻訳の枠組みでのテキスト平易化の例

手のラベル付きデータ [*2] のうち、2 文間の意味が等しいという *Good* のラベルが付いた 277 文対を使用し、BLEU [17] による評価を行った。

表 2 に各コーパスの文対数、語彙数、平均文長、BLEU を示す。Hwang らのコーパスでは各文対に文間類似度が付与されており、コーパス全体が文間類似度によって *Good* (0.67 以上)、*Partial* (0.53 以上)、*Remaining* (0.45 以上) の 3 つに分割されているので、それぞれと比較した。まず BLEU に注目すると、我々の構築したコーパスで学習したテキスト平易化モデルが、統計的機械翻訳の枠組みでのテキスト平易化において既存の他のコーパスで学習したモデルよりも高い性能を示した。また表 2 には、Hwang らのコーパスと我々の構築したコーパスにおいて、それぞれの手法で計算された文間類似度の高い順に 10 万文対、20 万文対、30 万文対を抽出してコーパスサイズを揃えたときの BLEU も示した。コーパスサイズに関わらず我々のコーパスで学習したモデルの BLEU が高いことから、性能の差がコーパスの量のみに起因するものではないことがわかる。なお、Zhu らと Coster and Kauchak のコーパスからも 10 万文対を抽出して BLEU を求めたが、これらのコーパスでは各文対に文間類似度が付与されていないため無作為に 10 万文を抽出しており、これらは参考の値である。

我々の構築したコーパスの平均文長は、既存の他のコーパスよりも normal 文と simple 文の差が大きく、English Wikipedia と Simple English Wikipedia の全体の平均文長 (25.1 および 16.9) に近いことがわかる。これは、Maximum Alignment が文長に関わらず適切に文間類似度を計算できていることを意味する。

表 3 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。我々のコーパスで学習したモデルは、入力文を適切に平易化し、Reference を含意する文を出力できた。Coster and Kauchak のコーパスで学習したモデルは、適切な平易化も行っているが、誤変換も行い非文を出力した。Hwang らのコーパスで学習したモデルは、無難な出力を行い、入力文を書き換えなかった。Zhu らのコーパスで学習したモデルは、誤変換のみを行い、非文を出力した。

## 5. おわりに

本研究では、単語の分散表現に基づいて計算される文間類似度を用いて、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手法を提案した。我々は、単語分散表現のアライメントに基づく 4 種類の文間類似度計算手法を提案し、一方の文中の各単語に対して最も類似度の高い他方の文中の単語を割り当てる多対一の単語アライメ

| 文間類似度 | normal | simple |
|---|---|---|
| 0.99 | Woody Bay Station was **purchased** by the Lynton and Barnstaple Railway Company in 1995 and, after much effort, a short section of railway reopened to passengers in 2004. | Woody Bay Station was **bought** by the Lynton and Barnstaple Railway Company in 1995 and, after much effort, a short section of railway reopened to passengers in 2004. |
| 0.90 | The fort was used by the Office of the Commissioners of Crown Lands, **who** had been evacuated from their central London offices **during World War II**. | **During World War II,** the Fort was used by the Office of the Commissioners of Crown Lands, **which** had been evacuated from their central London offices. |
| 0.80 | This work continued with the 1947 paper "Types of polyploids: their classification and significance", which **detailed a system for the classification of polyploids and** described Stebbins' ideas about the role of paleopolyploidy in angiosperm evolution**, where he argued that chromosome number may be a useful tool for the construction of phylogenies**. | This work continued with the 1947 paper "Types of polyploids: their classification and significance", which described Stebbins' ideas about the role of paleopolyploidy in angiosperm evolution. |
| 0.70 | Mir **has been** a significant influence on late 20th-century art, in particular the American abstract expressionist artists **such as Motherwell, Calder, Gorky, Pollock, Matta and Rothko, while his lyrical abstractions and color field paintings were precursors of that style by artists such as Frankenthaler, Olitski and Louis and others**. | Mir **was** a significant influence on late 20th-century art, in particular the American abstract expressionist artists. |
| 0.63 | **The couple** has **four children:** | **She** has **two daughters and two sons.** |
| 0.53 | Ithaca is **in the rural Finger Lakes region about northwest of New York City; the nearest larger cities, Binghamton and Syracuse, are an hour's drive away by car, Rochester and Scranton are two hours, Buffalo and Albany are three.** | Ithaca is **a city in upstate New York, America.** |

**表 4** 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

ントを利用する文間類似度計算手法の有効性を実験的に示した。我々の提案手法は、English Wikipedia と Simple English Wikipedia から抽出された文対に対して、パラレルデータとノンパラレルデータの 2 値分類を行う文間類似度の内的評価において最高性能を達成した。また、我々の提案手法によって構築されたテキスト平易化コーパスは、統計的機械翻訳の枠組みでのテキスト平易化を行う文間類似度の外的評価においても最高性能を達成した。

本研究では English Wikipedia と Simple English Wikipedia という難易度の異なるコンパラブルコーパス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した。しかし、このような難易度の異なるコンパラブルコーパスを大規模に入手することは多くの言語では難しい。今後は文の難易度推定手法と組み合わせることによって、大規模な単言語コーパスからの単言語パラレルコーパス構築に本手法を拡張し、多言語でテキスト平易化コーパスを構築したい。

また、単言語パラレルコーパスは、言い換えや文圧縮などの同一言語内のテキストからのテキスト生成タスクにおいても有用な資源である。今後は本研究で提案した単語分散表現に基づく文間類似度を用いた単言語パラレルコーパス構築手法の、他のタスクへの適用可能性を検討したい。

## 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).

[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] 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).

[11] Caseli, H. M., Pereira, T. F., Specia, L., Pardo, T. A., Gasperin, C. and Aluísio, S. M.: Building a Brazilian Portuguese Parallel Corpus of Original and Simplified Texts, *Advances in Computational Linguistics, Research in Computer Science*, Mexico City, Mexico, pp. 59–70 (2009).

[12] Bott, S. and Saggion, H.: An Unsupervised Alignment Algorithm for Text Simplification Corpus Construction, *Proceedings of the Workshop on Monolingual Text-To-Text Generation*, Portland, Oregon, USA, pp. 20–26 (2011).

[13] Klerke, S. and Søgaard, A.: DSim, a Danish Parallel Corpus for Text Simplification, *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, Istanbul, Turkey, pp. 4015–4018 (2012).

[14] Klaper, D., Ebling, S. and Volk, M.: Building a German/Simple German Parallel Corpus for Automatic Text Simplification, *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations*, Sofia, Bulgaria, pp. 11–19 (2013).

[15] Brunato, D., Dell'Orletta, F., Venturi, G. and Montemagni, S.: Design and Annotation of the First Italian Corpus for Text Simplification, *Proceedings of The 9th Linguistic Annotation Workshop*, Denver, Colorado, USA, pp. 31–41 (2015).

[16] 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).

[17] 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).

[18] 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).

[19] Mikolov, T., Chen, K., Corrado, G. S. and Dean, J.: Efficient Estimation of Word Representations in Vector Space, *Proceedings of Workshop at the International Conference on Learning Representations*, Scottsdale, Arizona, USA (2013).

[20] 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).

[21] 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).

[22] Sultan, M. A., Bethard, S. and Sumner, T.: DLS@CU: Sentence Similarity from Word Alignment and Semantic Vector Composition, *Proceedings of the 9th International Workshop on Semantic Evaluation*, Denver, Colorado, USA, pp. 148–153 (2015).

[23] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. and Dean, J.: Distributed Representations of Words and Phrases and Their Compositionality, *Advances in Neural Information Processing Systems*, Lake Tahoe, Nevada, USA, pp. 3111–3119 (2013).

[24] 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).

[25] Kusner, M., Sun, Y., Kolkin, N. and Weinberger, K.: From Word Embeddings To Document Distances, *Proceedings of The 32nd International Conference on Machine Learning*, Lille, France, pp. 957–966 (2015).

[26] Kuhn, H. W.: The Hungarian Method for the assignment problem, *Naval Research Logistics Quarterly*, Vol. 2, pp. 83–97 (1955).

[27] Rubner, Y., Tomasi, C. and Guibas, L. J.: A Metric for Distributions with Applications to Image Databases, *Proceedings of the Sixth International Conference on Computer Vision*, Washington, DC, USA, pp. 59–66 (1998).

[28] Och, F. J.: Minimum Error Rate Training in Statistical Machine Translation, *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, Sapporo, Japan, pp. 160–167 (2003).

[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).