新聞記事データを用いたテキスト平易化

Text simplification using newspaper articles

小藤　直紀　Naoki KOTO
難波　英嗣　Hidetsugu NANBA
竹澤　寿幸　Toshiyuki TAKEZAWA

広島市立大学　情報科学部
School of Information Sciences, Hiroshima City University

Automatic text simplification attempts to automatically transform complex sentences into their simpler variants without significantly changing the original meaning. Several researches on automatic text simplification have conducted based on a large-scale monolingual parallel corpus. However, it is costly to manually construct a parallel corpus for text simplification. Therefore, we investigate automatic construction of a large-scale simplified corpus for Japanese from newspaper database corpora. In this paper, we examined several methods for sentence alignment of texts with different complexity levels. Using the best of them, we sentence-align the Mainichi newspaper and Mainichi newspaper for elementary students, thus providing large training materials for automatic text simplification systems.

1. はじめに

現在，日本には 200 万人以上の在留外国人や，旅行客など多種多様な人種，国籍の外国人が日本に滞在している．近年，そのような人々に向けて「やさしい日本語」(http://human.cc.hirosaki-u.ac.jp/kokugo/EJ1a.htm)というものが考案されている．「やさしい日本語」とは，外国人にもわかりやすい日本語であり，地震などの災害が発生した際の情報伝達や，行政窓口の外国人対応など様々な状況で扱われている．難解な日本語を平易化し，情報伝達を容易にする「テキスト平易化」は，今後さらに重要なタスクになると考えられる．そこで本研究では，日本語テキストを平易化する手法を提案する．

近年，テキスト平易化を同一言語内の翻訳問題と考え，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が盛んに行われている．その代表的な手法に，難解なテキストと平易なテキストからなる大規模な単言語パラレルコーパスを用いたテキスト平易化が挙げられる．しかし，その多くは英語で作成された大規模パラレルコーパスを用いた手法である．そこで，本研究では毎日新聞と毎日小学生新聞の記事データをそれぞれ難解なテキスト，平易なテキストとして対応付けを行い，大規模パラレルコーパスを作成する．それを用いて文の平易化を行うことにより，円滑な情報提供の手助けをすることが本研究の目的である．

論文の構成は以下のとおりである，2 章では関連研究を紹介する．3 章ではテキスト平易化システムの構築について述べる．4 章では評価実験とその結果，考察を述べる．5 章で本論文をまとめる．

2. 関連研究

本研究の関連研究として，2.1 節では，やさしい日本語について，2.2節では，英語のパラレルコーパスによるテキスト平易化について，2.3 節では，平易なコーパスを用いないテキスト平易化について，2.4節では，SCDVを用いた文書分類について，それぞれ述べる

2.1 やさしい日本語

近年，外国人にも理解が容易な「やさしい日本語」を用いる取り組みが盛んであり，災害時の緊急連絡や，マニュアルなどに用いることで被害の減少が見込める．原文からやさしい日本語へ変換する技術に関して，例えば梶原らが作成した語彙平易化システム[梶原 17]などが挙げられる．

Maruyama らは日本語を対象にした自動平易化に用いる言語資源が少ないことを挙げ，「やさしい日本語」という観点で日本語に関する言語資源構築を行った[Maruyama 18]．原テキストには small parallel enja: 50k En/Ja Parallel Corpus for Testing SMT Methods が用いられている．Maruyamaらはこのコーパス内にある 5 万文すべてに対して，2,000 語の平易な語で構成される語彙集合「やさしい語彙」を用いて文単位の書き換え，やさしい語彙の追加，削除を行うことでやさしい日本語コーパスを作成した．本研究では新聞記事コーパスを用いることにより，より精度の高いパラレルコーパスの構築を目指す．

2.2 英語のパラレルコーパスによるテキスト平易化

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を用いたテキスト平易化手法が提案されている．Zhu らは English Wikipedia と Simple English Wikipedia から得られた大規模な並列データセットを使用した統計的機械翻訳による手法をベースとした，ツリーベースの翻訳モデル(TSM)によるテキスト平易化モデルを提案している[Zhu 10]．Zhu らは Wikipedia の言語リンクをたどり，English Wikipedia と Simple English Wikipedia から 65,133 対の対応付けを行った．その後，JWPL を用いて特定のタグを削除，両テキストからプレーンテキストを抽出した． 本研究では毎日新聞および毎日小学生新聞を用いる．対応付けには記事番号や「ある」，「いる」などの特徴を得られない動詞などを除いたベクトルを作成し，コサイン類似度を用いて対応付けを行う．

2.3 平易なコーパスを用いないテキスト平易化

梶原[梶原 18]らは生コーパスのみからテキスト平易化のための疑似パラレルコーパスを自動構築するフレームワークを提案している．生コーパスから無作為に抽出した2つの文に対して

連絡先：
小藤 直紀，難波 英嗣，竹澤 寿幸
広島市立大学 情報科学部
〒731-3194 広島市安佐南区大塚東三丁目 4 番 1 号
{koto, nanba, takezawa}@ls.info.hiroshima-cu.ac.jp

タスクに応じた品質推定を行い，一定以上の尤度を持つ文対を疑似パラレルコーパスとして抽出する

梶原らが，リーダビリティ推定で得た難解なサブコーパスと平易なサブコーパスを用いて文間類似度を計算し，英語のテキスト平易化のための疑似パラレルコーパスを構築した．本研究では，毎日新聞と毎日小学生新聞をそれぞれ難解なテキスト，平易なテキストとしてパラレルコーパスを構築する．毎日新聞と毎日小学生新聞で対応するテキストの難易ははっきりとしているので，リーダビリティを推定することなく，日本語の平易化のためのパラレルコーパスを作成することができる．

2.4 SCDV を用いた文書分類

Word2Vec のベクトル空間に対して GMM(Gaussian Mixture Models)でクラスタリングを行い，各単語がどのトピックに属しているのか考慮したベクトル空間に修正することで，単語ベクトルから文書ベクトルを作成する際の精度向上を目的とする手法 SCDV[Mekala 17]が他クラス文書分類問題において，他の手法と比べて優れていることが示されている．本研究でも SCDV が有効であると考え，記事間の類似度を計算する際の，特徴付きベクトルの作成に用いる．

3. テキスト平易化システムの構築

3.1 節ではテキストの平易化手順について述べる．3.2 節では，パラレルコーパスの構築手順である記事単位のアライメント取得手法について，3.3 節では，文単位のアライメント取得手法について，3,4節では，記事間の対応付けについて，3,5 節では，テキスト平易化ついて述べる

3.1 平易化手順

テキスト平易化手順として，毎日新聞および毎日小学生新聞の記事を難解なテキストおよび平易なテキストとして対応付けてパラレルコーパスを構築する．新聞記事同士を対応付けするにあたり，記事単位の対応付けをおこなった後，対応している記事内で文単位の対応付けを行い，アライメントをとる．その後，翻訳作業を行う．

3.2 記事単位の類似度判定

本研究ではコサイン類似度を用いて記事間の類似度を計算した．コサイン類似度計算に用いる各記事ベクトルは TF-IDF により取得する．また，単語の意味をベクトルとして表現する word2vec をクラスタリングし，文書の特徴を得る SCDV という手法が近年盛んに用いられる．本研究の平易化テキストに対しても，文字の一致率だけでは無く，文字の意味的特徴を考慮することが必要であると考え．SCDV を用いる．

3.3 文単位の類似度判定

難解な文と平易な文の同義性を評価し，対応付けするために本研究では，単語分散表現のアライメントに基づく，3 つの文間類似度の計算手法[Song 15]と Word Mover's Distance[Rubner 98]を本タスクに用いる．以下にそれぞれの計算手法を記載する．

(1) Average Alignment

全ての単語の組み合わせについて単語間類似度を計算，平均して文間類似度$S_{ave}(x,y)$を求める．$\Phi(x_i,y_j)$はコサイン類似度である．

$$S_{ave}(x,y)=\frac{1}{|x||y|}\sum_{i=1}^{|x|}\sum_{j=1}^{|y|}\Phi(x_i,y_j)$$

(2) Maximum Alignment

文に含まれる各単語に対して最も類似度が高い単語を選択し，それらの単語のみ計算した単語間類似度$\Phi(x_i,y_j)$を平均して$S_{asym}(x,y)$を求める．$S_{asym}(x,y)$は非対称なスコアなので，それらの平均値を用いて対象な文間類似度$S_{max}(x,y)$を計算する．

$$S_{asym}(x,y)=\frac{1}{|x|}\sum_{i=1}^{|x|}max_j\,\Phi(x_i,y_j)$$

$$S_{max}(x,y)=\frac{1}{2}\left(S_{asym}(x,y)+S_{asym}(y,x)\right)$$

(3) Hungarian Alignment

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を定義する．文xに含まれる各単語x_iに対して Hungarian 法[Kuhn 95]によって文y中の単語$h(x_i)$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を平均した文間類似度$S_{hun}(x,y)$を求める．

$$S_{hun}(x,y)=\frac{1}{min(|x|,|y|)}\sum_{i=1}^{|x|}\Phi\left(x_i,h(x_i)\right)$$

(4) Word Mover's Distance

$\psi(x_u,y_v)$は単語x_uと単語y_v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A_{uv}は文x中の単語x_uから文y中の単語y_vへの輸送量を表す行列であり，nは語彙数，$freq(x_u)$は文x中での単語x_uの出現頻度である．

$$S_{wmd}(x,y)=1-WMD(x,y)$$

$$WMD(x,y)=\min\sum_{u=1}^{n}\sum_{v=1}^{n}A_{uv}\psi(x_u,y_v)$$

$$\sum_{v=1}^{n}A_{uv}=\frac{1}{|x|}freq(x_u)$$

$$\sum_{v=1}^{n}A_{uv}=\frac{1}{|y|}freq(y_v)$$

3.4 記事間の対応付け

3.2 節で述べた記事単位の類似度を計測する際に，同じ内容の毎日新聞記事と毎日小学生新聞記事の日付に何日のずれがあるか調査した．調査の結果，前後1日より 2 日，前後 2 日より 3 日と対応している記事間の日付は近いほど対応数は増加していた．また，ある日付の毎日小学生新聞が，その日付の毎日新聞および，その日付より新しい毎日新聞と対応していることは極めて少なかった．人手による対応データは，毎日新聞から 1 日遅れで毎日小学生新聞の記事と対応している記事が記載されることが多い結果となっていた．

3.5 テキスト平易化

3.2 節および 3.3 節で提案した手法を毎日新聞および毎日小学生新聞に適用し，日本語テキスト平易化パラレルコーパスを自動構築する．構築したパラレスコーパスは，形態素解析器を用いて単語区切りを入れた後，機械翻訳システムを適用することで，日本語テキストの平易化システムを構築することができる．従来のテキスト平易化研究では，機械翻訳システムとして統計翻訳器が利用されることが多かったが，本研究でも，統計翻訳パッケージのひとつである cicada(http://att-astrec.nict.go.jp/product/cicada/)を用いる．この他，近年，機械翻訳分野で活発に研究されている seq2seq モデルをはじめとしたニューラル機械翻訳も利用して平易化システムを構築する．なお，本研究では，テキスト平易化パラレルコーパスの構築に主眼を置いているため，評価は記事単位の類似度判定および文単位の類似度判定についてのみ行う．

4. 実験

本章では，本研究で行った実験条件，実験結果，またその考察について，4.1 節では記事単位の類似度判定実験，4.2 節では文単位の類似度判定実験について述べる．

4.1 記事単位の類似度判定実験

(1) 実験条件

実験データおよび評価

毎日新聞，毎日小学生新聞ともに 2016 年の 1 年間分を使用して，人手により記事の対応付けを行い，記事単位の類似度判定実験に用いた．人手により作成された 856 対の記事単位の対応は A 評価:顕著に対応が取れる B 評価:対応を取れる という観点で対応付けが行われている． A 評価のものだけを用いて評価した場合と，A 評価と B 評価の両方を用いた場合の 2 通りで実験を行う．なお，評価には平均精度(average precision)を用いる．

比較手法

本研究では，以下の比較手法に基づき記事単位の類似度判定実験を行う．

- TF-IDF: 記事中の各単語の重み付けに TF-IDF 手法を用い，毎日小学生新聞の 1 記事に対して，その記事の日付当日および，その日付から前後 4 日分の毎日新聞の記事との類似度を計算
- SCDV: SCDV を用いてその記事の日付当日および，その日付から前後 3 日分までの毎日新聞記事との類似度を計算

(2) 実験結果

評価結果を表 1 に示す．表 1 より，対象記事に対して以前の記事だけでなく，TF-IDF を用いたアライメントの評価と SCDV を用いたアライメントの評価を比較したとき，TF-IDF を用いたアライメントのほうが優れていることがわかる．

表 1 記事単位の対応付け(平均精度)

手法	A 評価	A+B 評価
TF-IDF	0.9206	0.9042
SCDV	0.8734	0.8520

(3) 考察

本節では 4.1.2 節の実験結果に対する考察を述べる．結果より，SCDV よりも TF-IDF を用いて作成した記事ベクトルがより優れていた．正解データと比較した際に，TF-IDF は正しいアライメント結果を示し，SCDV では誤ったアライメント結果を示した例を以下に示す．正解データの毎日小学生新聞記事を図 1，SCDV のアライメント結果である毎日新聞記事を図 2 に示す．毎日小学生新聞での読み仮名は省略している．

＜毎日小学生新聞 2016 年 1 月 5 日，記事番号 7 番＞
第92回東京箱根間往復大学駅伝競走が2、3日、神奈川県箱根町―東京・大手町の往復10区間217・1キロメートルで行われ、青山学院大学が往路と復路を制して、2年連続2度目の総合優勝を果たしました。総合記録は10時間53分25秒。

図 1 毎日小学生新聞記事本文(一部抜粋)

<毎日新聞記事(2016 年 1 月 3 日，記事番号 77 番>
前回大会に続いて、トヨタ自動車が後半に実力を発揮した。群馬県の上州路を舞台に1日に行われた「ニューイヤー駅伝第60回全日本実業団対抗駅伝競走大会」。トヨタ自動車が2連覇を達成して通算優勝回数も5位タイの「3」とした。

図 2 毎日新聞記事本文(一部抜粋)

図 1 と図 2 は記事内容が異なり，チーム名などの名詞は違うが，順位の表記や記事の構成，その他の用語は正解データである毎日新聞記事(2016 年 1 月 4 日，記事番号 46 番)と非常に似ていた．SCDV の特徴および利点はテキストに対してクラスタリングを行い，各単語がどのトピックに属するかを加味したベクトル空間を作成できる点であるが，本研究で用いている新聞記事は元々各記事の内容に統一性があり，似た形式の記事が見られるため，SCDV を用いることによる利益が，充分に得られなかったことが原因と考えられる．

4.2 文単位の類似度判定

本節では，3.2 節で挙げた文単位の類似度判定実験に関する実験条件，実験結果，またその考察について述べる．

(1) 実験条件

実験データおよび評価

2016 年度の毎日新聞および毎日小学生新聞の対応している記事内で，文を人手により対応付けたものを文単位の類似度判定実験に用いた．文は句点で改行されており，先頭に記事番号と行番号が与えられ整理されている，この作業も人手により行った．人手により作成された 2,813 対の文単位の対応は A 評価:顕著に対応が取れる B 評価:対応を取れる，という観点で対応付けが行われている． A 評価のものだけを用いて評価した場合と，A 評価と B 評価の両方を用いた場合の 2 通りで実験を行う．なお，評価には平均精度(average precision)を用いる．

実験手法

本実験において，文単位のアライメントは 3.3 節で述べた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，Average Alignment，Maximum Alignment，Hungarian Alignment，Word Mover's Distance を用いて文間の類似度を判定する．A+B 評価の正解データを用いる．

(2) 実験結果

4 種類のアライメント手法に対して．A 評価の正解データのみを用いた場合と，A 評価と B 評価の正解データ両方を用いた場合の平均精度による評価結果を以下の表 2 に示す．今回の実験では Maximum Alignment が A 評価の正解データのみを用いた場合，A 評価と B 評価の正解データ両方を用いた場合のいずれでも最も優れた精度を示している．

表 2 文単位のアライメント評価(平均精度)

手法	平均精度	
	A 評価	A＋B 評価
Average Alignment	0.8483	0.8407
Maximum Alignment	**0.9802**	**0.9766**
Hungarian Alignment	0.9658	0.9617
Word Move's Distance	0.6109	0.5912

(3) 考察

本節では 4.2.2 節の実験結果に対する考察を述べる．評価の結果より，Maximum Alignment が最も優れた平均精度を示している．以下に正解データと Maximum Alignment の結果が異なっている毎日新聞および日小学生新聞の，記事本文の抜粋を記載する．id は年月日，記事番号，行を表す．

＜毎日新聞＞

<id=160105178-1>東京都中央卸売市場「築地市場」（中央区）で5日、新春恒例の初競りが行われ、青森県大間産の200キロのクロマグロが、昨年の3倍超の1400万円（1キロ当たり7万円）で競り落とされた。

＜毎日小学生新聞＞

<id=S160106002-1>今年11月に移転する築地市場（東京都中央区）で5日、「最後の初競り」が行われました。

＜毎日小学生新聞＞

<id=S160106002-2>青森県大間産のクロマグロ200キログラムに、昨年の約3倍にあたる1400万円（1キログラム当たり7万円）の値段がつきました。

人手による正解データでは，<id=160105178-1>に対して<id=S160106002-2>が対応している．これは，互いに含まれている単語と文脈を考慮して対応付けたと考えられる．一方，Maximum Alignment のアライメント結果では <id=160105178-1>に<id=S160106002-1>が対応付けられている，Maximum Alignment が分割された文のどちらが十分に文脈を汲み取れているかの判断を誤ったことが原因と考えられる．しかし全体的には非常に高い精度を示しているのも，また事実である．

5. おわりに

本研究では，毎日新聞と毎日小学生新聞から構築されたパラレルコーパスによるテキスト平易化手法を提案した．パラレルコーパスの構築手順である記事単位の類似度判定実験では TF-IDF による評価結果が SCDV を用いた評価結果より優れた結果となり，文単位の類似度判定実験では．4 つのアライメント手法のうち Maximum Alignment による評価結果が最も優れたか結果を示した．以上のアライメント手法により 30,940 文対からなる平易化のパラレルコーパスを構築した．

参考文献

[Kuhn 95] Harold William Kuhn: The Hungarian Method for the Assignment Problem, Naval Research Logistics Quarterly,Vol.2, pp. 83-97, 1995.

[Maruyama 17] Takumi Maruyama and Kazuhide Yamamoto: Simplified Corpus with Core Vocabular, Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC), 2018.

[Mekala 17] Dheeraj Mekala, Vivek Gupta, Bhargavi Paranjape, and Harish Karnick: SCDV : Sparse Composite Document Vectors Using Soft Clustering over Distributional , Representations, Proceedings of EMNLP 2017, pp. 659-669, 2017.

[Rubner 98] Yossi Rubner, Carlo Tomasi, and Leonidas Guibas: A Metric for Distributions with Applications to Image Databases, Proceedings of the 6th International Conference on Computer Vision, pp. 59-66, 1998.

[Song 15] Yangqiu Song and Dan Roth: 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, pp. 1275–1280, 2015.

[Zhu 10] Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych: A Monolingual Tree-based Translation Model for Sentence Simplification, Proceedings of the 23rd International Conference on Computational Linguistics, pp.1353–1361, 2010.

[梶原 17] 梶原 智之, 小町 守: Simple PPDB: Japanese, 言語処理学会第 23 回年次大会, pp. 529-532, 2017.

[梶原 18] 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.