# Text Simplificationのための文難易度の２値分類手法の検討

高田祥平 †, 水嶋海都 ‡, 荒瀬由紀 ‡,
† 大阪大学工学部電子情報工学科, ‡ 大阪大学大学院情報科学研究科マルチメディア工学専攻
{takada.syouhei, mizushima.kaito, arase}@ist.osaka-u.ac.jp

## 1 はじめに

インターネットや情報処理技術の発達により多くの電子テキストが利用できるようになっている。しかし、難解な情報を持つテキストについては誰もがその内容を理解できるわけではなく、テキストの読みやすさの向上が求められている。その需要から、与えられたテキストを万人に理解しやすいテキストへと変換するtext simplificationというタスクが研究されるようになった。

Text simplificationの目的は大きく2つに分けられる。1点目は年齢が低い、または十分な教育を受けられていないなどの理由から言語能力が低い人やノンネイティブ話者のためにテキストを読みやすいものに変換することである。2点目は機械翻訳や要約、言語生成といった自然言語処理で扱いが容易なテキストに変換することである。

Text simplificationのプロセスとしてはテキストの中の余分なフレーズを取り除くこと、複雑なフレーズや文構造を置換すること、テキストの意味をわかりやすくするために新たなフレーズを挿入することの3点に分けられる。このプロセスは機械翻訳や要約といった他の言語処理のプロセスと類似しており、text simplificationの実装にはそれらの言語処理の手法が利用されている。しかし既存研究では、そもそもtext simplificationが必要かどうかの判定はされておらず、充分simpleと言える文に対しては、システムを適用することで可読性を下げる要因となる。またtext simplification 手法の学習においても、コーパスに含まれるtext simplificationが不要な文が学習効果を下げてしまう。

そこで本稿では、入力テキストに対してtext simplificationが必要なnormal文か、simplificationが不要なsimple文かを2値分類により判別する。分類の手法として、テキストからNapolesら[4]が提案した特徴量に、構文木の特性、単語難易度、パラフレーズを用いた特徴量を追加し、Support Vector Machine (SVM)[5] を用いて分類する。また、使用した特徴量セットごとの分類精度から、分類に有効な特徴についての考察も行う。

## 2 関連研究

Costerら[2]は英語のtext simplificationタスクを英語テキストから簡単な英語テキストへの翻訳ととらえ、統計的機械翻訳システムであるMoses[9]を用いてtext simplificationを実現している。Text simplificationが行われた前後のテキストデータを対訳文としてMosesに学習させ、フレーズの言い換えだけでなく削除についての規則を加えることで、text simplificationの品質を改善している。

またtext simplificationの限定的な問題として、テキスト中の語彙を簡単なものに変換するタスクであるlexical simplificationも研究されている。Hornら[3]はtext simplificationが行われた前後のテキストデータから単語の言い換え規則を抜き出し、単語の出現頻度や複数の言語モデルを使用してランク付けを行い、lexical simplificationを実現している。

一方で、NapolesらはSimple Wikipediaがtext simplificationのコーパスとして利用できるか検証している。WikipediaとSimple Wikipediaの記事単位および文単位を入力として、どちらから抽出されたか2値分類を行っている。記事単位の分類は99%を超える高い精度で行うことができたが、文単位での分類についてはSVMを用いたもので77%という精度となった。目的は異なるが、入力がsimpleなテキストかどうかの判定をしているという点で、本研究と関連が深い。本稿ではNapolesらが使用した特徴量の中のbag-of-wordsを使用したモデルをベースラインとし、構文情

表 1: Wikipedia コーパスの例

| | |
|---|---|
| normal | It has the highest elevation of any market town in England. |
| simple | It is the highest market town in England. |

表 2: 実験データの詳細

| | normal | simple |
|---|---|---|
| 文数 | 117,023 | 166,121 |
| 単語数 | 3,045,293 | 3,554,303 |
| 単語のタイプ数 | 105,705 | 108,525 |

報や単語難易度など、多様な特徴量を用いることで分類性能を向上できるか検証する。

## 3 コーパス

本稿の実験では Coster ら [1] が Wikipedia[*1] と Simple Wikipedia[*2] 中の記事に含まれる文の対応を取ることで作成したコーパス (Wikipedia コーパス）[*3] を使用した。以下では text simplification を行う前のテキストを simple とし、行った後のテキストを normal とする。表 1 にコーパスの例を示す。

Wikipedia コーパスはおよそ 137,000 組のテキストの対からなっており、その中には複数の文同士がアライメントされたもの、同一の文がアライメントされたものが含まれている。同一の文がアライメントされているものは、text simplification の必要がないデータとみなして simple 文として用いた。また、Wikipedia コーパスは自動生成されたものであるためノイズとなり得るデータが含まれている。そこで構文解析器 Enju[*4][8] での解析でエラーが発生したテキスト、URL のみのテキスト、単語の平均文字数が 2 以下となるテキストをノイズとして取り除いた。最終的に得られた実験データの詳細を表 2 に示す。

[*1]https://www.wikipedia.org/
[*2]https://simple.wikipedia.org/
[*3]http://www.cs.pomona.edu/ dkauchak/simplification/
[*4]http://www.nactem.ac.uk/enju/index.ja.html

表 3: 特徴量表

| 特徴量セット名 | 内容 |
|---|---|
| bag-of-words | テキストに出現する単語の頻度 |
| 品詞別の単語の個数 | 品詞ごとの単語の個数 |
| 表層から得られる特徴 | テキストの文字数、単語数、単語の文字数、簡単な単語の含まれている割合 |
| 構文解析から得られる特徴 | 最大の句の単語数、句中の単語数の割合 |
| 単語の難易度 | 難易度別の単語の個数 |
| パラフレーズ知識 | PPDB より抽出した、simple/normal フレーズの個数 |
| 構文の複雑さ | 構文木の深さ、関係詞句の個数 |

## 4 特徴量

表 3 に本研究で用いる特徴量の概要を示す。表 3 の上から 4 種類が Napoles らの実験で使用された特徴量となっている。なお、各要素は最大値が 1 で最小値が 0 となるように正規化を行った。

### 4.1 Napoles らによる特徴量

**bag-of-words ベクトル**　最も基本的な特徴として、コーパス内の単語の出現頻度から bag-of-words ベクトルを生成した。ベクトルの要素としては頻度が 30 以上の語でストップワードを除いた 7469 語を使用した。

**品詞別の単語の個数**　Enju の解析結果から得られた単語の全品詞タグの出現頻度と 2 つの品詞の共起頻度を特徴量とした。また名詞の単数形と複数形や固有名詞のタグをまとめて名詞として扱うなど、名詞、動詞、形容詞、副詞、限定詞、関係詞の 6 種類は関連するタグをまとめてタグの種類を減らした頻度の抽出も行った。

**表層から得られる特徴**　複雑なテキストは簡単なものと比較して文や単語が長いという仮説から、各テキストの文字数と単語の個数、単語の平均の文字数を特徴量として抽出した。また、Simple Wikipedia は簡単な単語を用いて記事を書くよ

うに推奨されている。Simple Wikipedia のガイドラインで示されている Basic English 850 (BE850) list[7] を使用し、各テキストに含まれる単語の割合も抽出した。

**構文木における特徴**　テキストの構造的な複雑さを表す特徴量として、Enju による解析木を用いる。各テキストの構文木から名詞句、動詞句、前置詞句、関係詞句の 4 種類の句の最大の単語数をそれぞれ抽出した。また関係詞句の中の名詞句の単語数の割合など、句中の単語数の割合についても特徴量として加えた。

### 4.2　提案する特徴量

**単語の難易度**　各単語について BE850 に含まれているかどうかだけでなく、さらに細かい難易度を考慮するため、投野ら [6] によって作成された CEFR-J Wordlist を用いる。このリストは Common European Framework of Reference (CEFR) 基準に単語を分類したリストで、7821 語をその見出し語と品詞をもとに A1、A2、B1、B2 の 4 段階に分けたものである。CEFR-J Wordlist にリストアップされた単語とその難易度から、テキストに含まれる 4 種の難易度別の単語数を特徴量とした。

**パラフレーズ知識**　テキストのフレーズについて、より簡単なフレーズに言い換えられるものが含まれているならば、そのテキストは text simplification が可能であると考えられる。フレーズの言い換え表現を収集した Paraphrase Database (PPDB)[*5][10] の Lexical と Phrasal のデータを利用する。Wikipedia コーパスにおける言い換え表現の頻度を取り、normal と simple に含まれる頻度が異なるフレーズの対について、normal の頻度が高いものを normal フレーズ、simple の頻度が高いものを simple フレーズとして抽出した。以下に例を示す。

$$\text{is accountable for} \rightarrow \text{is in charge of}$$

得られたフレーズの対の総数は Lexical が 127,159 組、Phrasal が 186,852 組となった。この言い換えリストの simple/normal フレーズが各テキストに含まれている個数をそれぞれ抽出し、特徴量とした。

[*5] http://www.cis.upenn.edu/ ccb/ppdb/

表 4: 特徴量セット別のパラメータ値とテストセットにおける分類精度

| 特徴量セットの名前 | C | 分類精度 (%) |
|---|---|---|
| ベースライン | 0.5 | 66.19* |
| ＋品詞別の単語の個数 | 0.3 | 67.30 |
| ＋表層から得られる特徴 | 0.0 | 66.50* |
| ＋構文解析から得られる特徴 | 0.4 | 66.25* |
| ＋単語の難易度 | 0.4 | 66.40* |
| ＋パラフレーズ知識 | 0.2 | 66.10* |
| ＋構文の複雑さ | 0.4 | 66.46* |
| 特徴量全て | 0.6 | 67.52 |
| 「パラフレーズ知識」以外全て使用したモデル | 0.3 | 67.63 |

**構文の複雑さ**　構文の複雑さについて考慮するため、Enju による解析結果からテキストの構文木の深さと関係詞句の部分木の個数を特徴量として追加した。複数の文を含むテキストについては、深さが最大のものを使用する。

## 5　評価実験

### 5.1　実験設定

実験データについて、トレーニングデータを 226,518 文、ディベロップメントデータを 28,316 文、テストセットを 28,317 文とした。分類には SVMLight[*6] を使用した。線形カーネルを使用し、ディベロップメントデータを用いてハイパーパラメータを設定した。Bag-of-words による分類をベースラインとして、bag-of-words ベクトルに他の特徴量を一つ追加したモデルおよび特徴量を組み合わせたモデルによる分類の精度と比較し、各特徴量について評価する。

### 5.2　実験結果

各モデルのパラメータと分類精度を表 5.2 に示す。表 5.2 では、符号検定により有意差が認められた分類結果を * で示す。この結果から、bag-of-words と品詞に関する素性を組み合わせたモデルで分類精度が頭打ちになっており、他の特徴量が精度に貢献しなかったことが示される。

[*6] http://svmlight.joachims.org/

実験データを検証したところ、3語以下の単語変換のみからなる simplification の割合がおよそ 46%であることから、多くの text simplification の変換はテキストの一部しか行われていない。そのため、テキスト全体を見る指標である構造的な特徴量は効果が限定されていたと考えられる。このことから、text simplification は翻訳とは性質が異なる問題であり、困難な課題であることが分かる。

パラフレーズ知識を用いた特徴量については、追加することで精度が低下することが分かった。パラフレーズペアのうち、どちらが simple かを判断するには今回用いたような簡潔な手法では不十分であることがうかがえる。単語難易度も加味する、simple フレーズの決定に閾値を導入するなど、今後改善の余地がある。

また、simple 文、normal 文について特徴量の分布が 2 つのクラスで非常に近いことが分かった。例えば、テキストの単語数について、simple 文と normal 文の平均値はそれぞれ 21.4±0.1 語と 26.0±0.1 語、標準偏差はそれぞれ 11.3 と 13.8 となっていた。このことから、テキストのみから抽出する特徴量では分類精度を向上するのは難しいことが分かる。今後、固有名詞抽出や専門用語辞書など、外部知識を利用するアプローチも検討する必要がある。

## 6 まとめ

本稿では、テキストが simple であるかどうかを判断することを目的として、Napoles らの特徴量に加え、単語難易度や構文木の複雑さ、またパラフレーズを考慮する特徴量を用いてテキストの 2 値分類を行った。結果として、bag-of-words と品詞に関する特徴量の組み合わせを向上する特徴量は得られず、normal/simple 文の判別は困難な課題であることが分かる。

今後は、専門用語辞書などの外部リソースの利用や、他の分類器を用いた性能評価を実施する予定である。

## 参考文献


[1] W.Coster and D.Kauchak. “Simple English Wikipedia : A New Text Simplification Task,” Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, pages 665–669, (June 2011).

[2] W.Coster and D.Kauchak. “Learning to Simplify Sentences Using Wikipedia,” Proceedings of the workshop on monolingual text-to-text generation, pages 1–9, (June 2011).

[3] C.Horn, C.Manduca, and D.Kauchak. “Learning a Lexical Simplifer Using Wikipedia,” Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pages 458–463, (June, 2014).

[4] C.Napoles and M.Dredze. “Learning Simple Wikipedia : A Cogitation in Ascertaining Abecedarian Language,” Proceedings of the Workshop on Computational Linguistics and Writing, pages 42–50, (June 2010).

[5] C.Corinna and V.Vapnik. “Support-Vector Networks,” Machine learning 20.3,pages 273-297, (1995).

[6] CEFR-J Wordlist Version 1 (2013) 東京外国語大学投野由紀夫研究室. http://www.cefr-j.org/download.html.

[7] C.K.Ogden. “Basic English: A General Introduction with Rules and Grammar,” London: Paul Treber & Co., Ltd. (1930).

[8] Y.Miyao and J.Tsujii. “Feature Forest Models for Probabilistic HPSG Parsing,” Computational Linguistics. 34.1. pages 35–80, MIT Press, (2008).

[9] P,Koehn, H.Hoang, A.Birch, C.Callison-Burch, M.Federico, N.Bertoldi, B.Cowan, W.Shen, C.Moran, R.Zens, C.Dyer, O.Bojar, A.Constantin, and E.Herbst. “Moses: Open Source Toolkit for Statistical Machine Translation,” Proceedings of the Annual Meeting of the Association for Computational Linguistics, pages 177–180, (June 2007).

[10] G.Juri, B.V.Durme, and C.Callison-Burch. “PPDB: The Paraphrase Database,” Proceedings of the North American Chapter of the Association for Computational Linguistics Human Language Technologies, pages 758–764, (June 2013).