共起関係を用いた単語の置き換えによる文章の可読性向上システム

栗田 雷也　　指導教員：山村 毅

1 はじめに

現代社会では, 我々の身の回りに多くの文が存在しており, 中には理解し難い文章（表現が硬い, 単語が難しい, 文節の順序が分かり難いなど）も少なくない. こういった文章を読むためには, 表現や単語, 文法などを調べたりといった作業を繰り返す必要があり, 多くの時間と労力がかかってしまう. そのため, 読むことや理解することが難しく, 読み手側の読む意欲を低下させてしまうことが考えられる. この問題に対して, 栗田ら [1] は文中の難単語を対象に，小学校で使われている簡単な類義語に置き換えることで文章の可読性向上手法の提案を行った.

本研究では, 文章の可読性に重きを置き，既存のシステムに加え，共起関係を用いた単語の置き換えを行うことで，文章の可読性向上させるシステムの考案を目的とする.

2 先行研究 [1]

2.1 システム概要

図 1 にシステム構成を示す. まず, システムは, 入力文に対して形態素解析を行い, 単語を取得する. 次に, 小学校単語データを参照し, 取得した単語の中から難単語を取得する. さらに, 難単語に対し, word embedding を用いて類義語抽出を行い, 類義語に対して, 小学校単語データを参照し, 置き換え候補リストを作成する. 最後に, 置き換え候補リストを元に単語の置き換えを行い, 文を出力する.

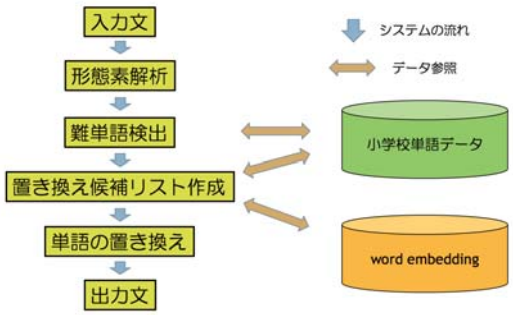

図 1　システム構成

2.2 難単語の検出

形態素解析の結果から取得した単語に対して, 条件 1~3 に従って単語を抽出する. これを難単語とする.

条件 1： ”名詞, 動詞, 副詞, 形容詞, 形容動詞, 連体詞”である
条件 2： ”数・助数詞・固有名詞・副詞可能”でない
条件 3： 小学校データを参照し, 小学校で使われていない

2.3 word embedding

word embedding とは，単語の概念を低次元の密なベクトル表現で表したものである. 対象の単語と似た概念を持つ単語は，ベクトル空間上で近くに存在するベクトルとして表現されることを利用し類義語抽出を行うことができる.

2.4 置き換え候補リストの作成

難単語を入力とし, word embedding を用いて類義語を出力する. 小学校単語データを参照し, 出力結果から小学校で使われている単語を抽出し, 類似度と学年情報を付与し類似度の高い順にリストに追加する. これを置き換え候補リストと呼ぶ.

2.5 単語の置き換え

2.4 節で述べた, 置き換え候補リストを参照し, 入力文に対して品詞情報の有無を用いた類似度の高い単語への置き換えを行う. また, 文章の自然さを維持する為, 必要に応じて, 置き換え前の単語に合わせて置き換え後の単語の活用を行う.

2.6 評価実験

word embedding の作成時のパラメーター調整が同義語検索に与える影響について評価実験を行った. パラメータの調整には，window size を 2~5 の 4 種類，学習モデルを日本語版 wikipedia と毎日新聞記事の 2 種類，学習モデルを CBOW と skip-gram の 2 種類に変化させ，組み合わせた計 16 種類の word embedding を作成した. 評価指標として,「設問 1：置き換えた文章と元の文章の意味が同じである.」,「設問 2：置き換えた文章が元の文章と比較してわかりやすい.」,「設問 3：置き換えた文章は自然である.」 の 3 項目を設定した. 表 1 に最も精度の良かった「日本語版 wikipedia / skip-gram / window size 2」の結果を示す. 置き換え後の文章のわかりやすさの精度が 30% に満たないことから，可読性の向上において精度が低いと言える. 精度の低い原因として，置き換える必要がない難単語を誤って置き換えていることが考えられる.

表 1　アンケート結果

	設問 1	設問 2	設問 3
精度	95 %	29 %	29 %

3 共起関係を用いた単語の置き換え

3.1 先行研究に対する分析

先行研究では，難単語として検出された単語を全て置き換えの対象としていたが，実際には検出された難単語の中で，どれくらいの難単語が単語の置き換えで対処できるのかをまずは分析する.

毎日新聞記事 200 文に対して解析を行ったところ，547 個の難単語が検出され，その内，145 個の難単語が置き換えで対処できることがわかった. さらに，置き換えで対処できる 145 個の難単語の内，置き換え候補リスト内に適切な類義語が存在する難単語は 57 個であった.

3.2 提案手法

3.1 節の分析結果より，本研究では，単語の置き換えによって対処できる 57 個の難単語を含んだ文を対象とする. 類義語抽出の際に word embedding の性質上，実際には関連した単語が抽出される. そのため，類似度のみで単語選択を行うと文章に対して適切でない単語が選択される. そこで，類義語を選択する際に，類似度だけでなく文章を構成する単語との結び付きの強さを考慮することで文章に対して適切な単語を選択できるよう

になると考えられる．単語の結び付きの強さを表す指標として共起関係を利用する．

3.3 共起関係

単語間の共起関係とは，単語間の結びつきの強さ，ある単語とある単語が文章中に同時に出現することを意味する．単語の類似性だけでなく，文章に適切な類義語を選択するために，単語の共起関係を利用する．

3.4 システムの追加

共起関係を用いた単語の置き換えを行うために，従来のシステムに”共起関係の作成”，”置き換え候補リストの更新”を新たな処理として加える．共起辞書は，置き換え候補リストの更新に利用し，難単語に対する各類義語の共起の出現頻度を獲得する．

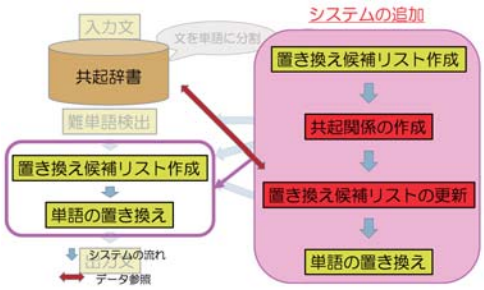

図 2　システムの追加

3.5 共起辞書

2010 年の毎日新聞記事 (63.8MB) の各文章に対して形態素解析を行い，2.2 節で述べた条件 1 と 2 に該当する単語を抽出する．抽出した各単語を 2 単語ずつ組み合わせ共起関係を作成し，各共起関係の出現頻度を計算しまとめたものを共起辞書とする．

3.6 共起関係の作成

入力文に対して形態素解析を行い，2.2 節で述べた条件 1 と条件 2 に該当する単語を抽出する．抽出した単語を難単語とその他の単語に分類し，難単語とその他の単語を 2 単語ずつ組み合わせ共起関係を作成する．作成した共起関係における難単語を類義語に置き換え，類義語と入力文を構成する各単語との共起関係を作成する．

3.7 置き換え候補リストの更新

3.6 節で述べた類義語と入力文を構成する各単語との共起関係に対して，共起辞書を参照し各共起関係における出現頻度を獲得する．獲得した各共起関係における出現頻度を合計し，合計値を各類義語の出現頻度とし，置き換え候補リストに各類義語の出現頻度を加え置き換え候補リストを更新する．

3.8 共起関係を用いた単語の置き換え

置き換え候補リスト内の最も高い類似度を”t”とし，条件 1〜3 に従って置き換え候補リストを分類し，各条件に合わせた方法で単語の置き換えを行う．ただし，選択する類義語は難単語と同じ品詞であることとし，必要であれば単語の活用を行う．

条件 1　t ≥0.85 の場合，類似度が 0.85 以上の類義語の中から類似度が最も高い類義語を選択

条件 2　0.85>t≥0.55 の場合，類似度が 0.55 以上の類義語の中から出現頻度が最も高い類義語を選択

条件 3　0.55>t の場合，類似度 × 出現頻度を総合値とし，総合値が最も高い類義語を選択

4 実験方法と結果

4.1 実験方法

2 章で述べた従来のシステムに対し，3 章で提案した処理を加えたシステム作成する．そして，それを用いて，3.1 節で述べた単語の置き換えで対処できる 57 個の難単語に対して，共起関係を用いた単語の置き換え行い，文章の可読性を向上させた文章を出力する．2.6 節で述べた「日本語版 wikipedia / skip-gram / window size 2」で学習を行った word embedding を用いて難単語に対する類義語を獲得する．出力した文章に対して weblio の類義語辞典を用いて作成した正解文と比較し考察を行う．

4.2 実験結果

本研究で提案した手法と先行研究で提案された手法を用いた単語の置き換え結果を表 2 に示す．結果として，本研究で提案した” 共起関係を用いた単語の置き換え方法”は，従来の手法より精度が約 12 ポイント向上した．これより，共起関係を用いた単語の置き換えは文章の可読性向上に有用であることがわかった．

表 2　正しく置き換えた単語数と精度

	単語数 (個)	正誤率 (%)
本研究で提案した手法	41（個）	72（%）
先行研究で提案された手法	34（個）	60（%）

4.3 検討・考察

単語の置き換えに共起関係を用いることで，従来の手法では正しく置き換えることができなかった 23 個の難単語の中から 10 個の難単語を正しく置き換えることができた．しかし，逆に 3 個の難単語の関しては，類似度順に単語を選択した場合には正しく置き換えることができていたが，共起関係を用いた場合は，誤って置き換えられていた．また，共起関係を用いた場合でも，正しく置き換えることができなかった 13 個の難単語において原因を調査すると，そのうち 11 個の難単語に関しては，置き換え候補リスト内の類義語の 60%〜88% の単語が類似していないことがわかった．さらに，そのうち 4 個の難単語に関しては，正解の単語が難単語と異なる品詞であり，本研究では，単語の選択の際に難単語と同じ品詞の単語を選択しているため選択の対象外となっていることがわかった．

5 まとめ

単語の置き換えの際に，共起関係を用いることで正しく単語の選択が行えるように改良を行った．実験を行った結果，従来の手法と比べ約 12 ポイント精度が向上しており，共起関係を用いた単語の置き換えは文章の可読性向上に有用であることがわかった．実験結果と考察より，今後の課題としては，word embeding による類義語の獲得の際に，関連性が高い単語ではなく類似している単語を獲得できるよう学習を行うこと，データ量を増やし類似度の閾値や総合値の計算式を学習によって決定できるようにすることなどが挙げられる．

参考文献

[1] 栗田雷也, 山村毅: ”word embedding 作成時の学習データの違いが同義語検索に与える影響について”, 第 15 回情報学ワークショップ (WiNF2017), PC-22, 中部大学, 2017.