英語教育支援のための Lexical Simplification: コロケーションスコアを用いたアプローチ

高田祥平 †, 荒瀬由紀 †, 内田諭 ‡
† 大阪大学大学院情報科学研究科, ‡ 九州大学大学院言語文化研究院
†{takada.syouhei, arase}@ist.osaka-u.ac.jp, ‡uchida@flc.kyushu-u.ac.jp

1 はじめに

国際化やインターネットの普及が進み，ノンネイティブ話者が英語のテキストを扱う機会は増加している．しかし，ノンネイティブ話者にとって，難解な情報を持つテキストの内容を理解することは困難であるため，テキストの読みやすさの向上が求められている．その需要に伴って，与えられたテキストを理解しやすいテキストへと変換するタスクである Text simplification について研究が進められている．Text simplification には文全体の書き換えを行うもの [5] もあるが，基礎となるアプローチとして入力テキスト中の理解が難しい単語をより簡単な単語に置き換える Lexical simplification[2][6] がある．Laufer[8] は適正な英文の理解を得るためには 95% 以上の単語が既知であることが望ましいと指摘しており，テキストの難易度について単語が寄与する部分は大きいと考えられる．英語の Lexical simplification を行うことで，英語教育者が学習者に示す英文の難易度を最適なものに変換することや，低頻度の難しい単語を置き換えることによって他の自然言語処理の精度を向上できると考えられる．

本稿では，英語教育者の教材準備支援を目的とした Lexical simplification 手法を提案する．英語教育者は自身の経験に頼って教材として用いる文章の難易度調整を行うがその負担は大きい．そこで提案手法では単語の置き換え候補を提示し，教育者が簡単に文章の難易度を調整できるようにする．Lexical Simplification のプロセスは，言い換えの対象となる単語 (TARGET と呼ぶ) の特定，単語の言い換え先候補 (CANDIDATE と呼ぶ) の抽出，文脈に合わせた CANDIDATE のランク付けの 3 つに分けられる．教材準備に用いること，また教育者の負担を軽減するための手法であることを考えると，高い精度で CANDIDATE を提示することが重要となる．提案手法では世界的な言語能力の評価指標である Common European Framework of Reference for Languages (CEFR) [3] に準拠した単語辞書を用いて TARGET を特定し，TARGET の類義語と周辺の語との関係に対して，共起頻度を基にしたスコア付けを行い，文脈に沿った CANDIDATE のランキングを出力する．

提案手法の性能を評価するため Gold-standard データを作成した．具体的には教科書の文章において CEFR レベルの高い難しい単語を TARGET とし，辞書から抽出した類義語を CANDIDATE 候補とし，ネイティブ話者によるアノテーションを行った．実験の結果，言語モデルスコアにより CANDIDATE をランキングするベースラインに比べ，提案手法は高い精度で CANDIDATE を決定できることが示された．

2 関連研究

Pavlick ら [10] は言い換え表現を収集した Paraphrase Database (PPDB)[*1][7] から，フレーズ単位の simplification のための変換規則を抽出し，Simple PPDB[*2] として公開している．Simple なフレーズの抽出には機械学習を用い，その特徴量としてフレーズの長さの他に単語のベクトル表現が利用されている．

また，Horn ら [6]，Biran ら [2] は共に Lexical simplification の変換規則抽出を Wikipedia[*3] と Simple Wikipedia[*4] から作成されたコーパス [4] を利用して行っている．入力文の文脈に対して適切な単語を CANDIDATE として選択するための基準として，Biran らは Wikipedia で CANDIDATE と共起した単語の頻度と入力テキストで CANDIDATE と共起した単語の頻度をそれぞれ特徴ベクトルとし，2 つのベクトルのコ

[*1] http://www.cis.upenn.edu/~ccb/ppdb/
[*2] http://www.seas.upenn.edu/~nlp/resources/simple-ppdb.tgz
[*3] https://www.wikipedia.org/
[*4] https://simple.wikipedia.org/

図 1: 提案手法の手順

サイン類似度を計算し，スコアとして用いている．一方で，Horn らは言語モデルや CANDIDATE の Simple Wikipedia における出現頻度， Google N-gram[*5] における出現頻度を利用したランク付けを行い，得られた CANDIDATE のランキングについて Amazon Mechanical Turk[*6] を用いた人手での評価を行っている．提案手法では英語教育者による教材作成支援を目的としており，CANDIDATE 生成の際の精度を重視した手法となっている点でこれらの既存研究と異なる．

3 提案手法

3.1 手法の概要

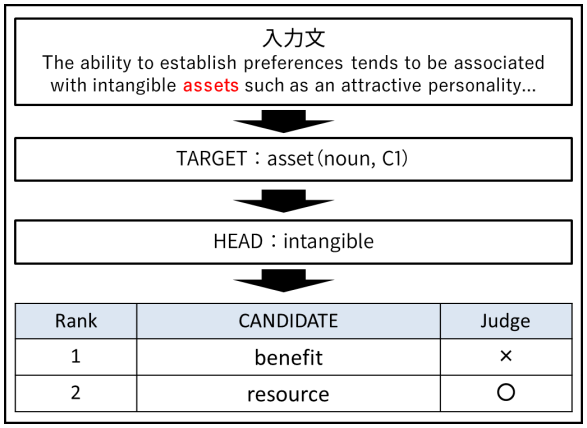
提案手法では CANDIDATE 選定の精度を重視し，TARGET のコロケーションスコアを用いる．図 1 に提案手法における単語の CANDIDATE 決定の手順を示す．TARGET と高いコロケーションスコアを持つ単語を HEAD とし，TARGET をある単語に置き換えたときにも HEAD とのコロケーション関係が成り立つものを CANDIDATE とする．各手順における処理について，以下に記述する．

TARGET の選定 入力文の各単語について，構文解析結果から得られる原形と品詞をもとに CEFR-J Wordlist Version 1[*7][12] を用いて難易度を付与する．このリストは CEFR 基準に単語を簡単なものから順に A1, A2, B1, B2 で示したもので，日本人学習者向けに特化したものである．これに English Vocaburaly Profile[*8] のリストを結合し，さらに上位の C1, C2 レベルの単語を追加した．難易度が B2 以上であるものを難単語とし，TARGET とする．本研究では構文解析器として Stanford CoreNLP[9] を利用した．

類義語の抽出 類義語辞書から TARGET の難易度が B1 以下である類義語を抽出する．

HEAD の探索 TARGET の前後 4 単語のウィンドウ内に存在する単語のうち，TARGET に対して高い MI スコアを持つ単語を TARGET とコロケーション関係を持つ HEAD とする．

CANDIDATE の決定 HEAD と TARGET の類義語との MI スコアを算出し，高いスコアを持つものを CANDIDATE として出力する．スコアを持つ類義語が複数存在する場合は，スコア順にランキングした類義語のリストを出力する．



Rank	CANDIDATE	Judge
1	benefit	×
2	resource	○

図 2: 出力の例 : “Judge” は評価実験によって得られたネイティブ話者によるアノテーションであり，この例では “resource” に置き換え可能なことを示している．実際のシステムでは “Judge” の項目は表示しない．

提案手法による Lexical simplification の例を図 2 に示す．

3.2 CANDIDATE のランク付け

提案手法では，単語の難易度の付与に CEFR-J Wordlist Version 1 を用いる．また，言い換え先の候補となる類義語辞書の生成には Thesaurus.com[*9] の

[*5] LDC カタログ : LDC2006T13
[*6] https://www.mturk.com/
[*7] http://www.cefr-j.org/download.html
[*8] http://www.englishprofile.org/wordlists
[*9] http://www.thesaurus.com/

データを使用する．2 単語の結びつきの強さを示す指標として Corpus of Contemporary American English (COCA) のデータから計算された MI スコア [1] を利用する．単語 A, 単語 B の MI スコアとは以下の式で表され，$f(A), f(B)$ はそれぞれ単語 A, B のコーパス内での出現頻度，$f(A,B)$ は A, B の共起頻度，n はコーパスの総語数，w は共起を見るウィンドウ幅を示す．COCA の MI スコアリストでは $w = 4$ に設定されている．共起頻度の低い語の MI スコアは信頼性が低いため，MI スコアリストから共起頻度が 6 以上のデータを用いる．MI スコアは 2 単語の共起強度が大きいほど高い値をとり，2 単語のコロケーション関係は深いといえる．

$$MI(A,B) = \log_2 \frac{nf(A,B)}{wf(A)f(B)}$$

4 評価実験

4.1 Gold-standard データの作成

提案手法の有効性を評価するため，Gold-standard データを作成した．実験対象のデータとして，大学の英語リーディングの授業で用いられる上級レベルのテキスト [11] を使用した．実験データから TARGET を含む文を抽出し，各 TARGET の類義語を CANDIDATE 候補として，ネイティブ話者に置き換え可能か文脈を考慮しながらアノテーションしてもらった．アノテータは 30 代後半のイギリス出身のネイティブ話者で，日本での英語教授歴が 15 年ある人物である．豊富な英語教育経験を有しており，アノテータとしての信頼性は高いと考える．

アノテータの負荷を軽減するため，TARGET の類義語数およびそれらの CEFR レベルによるアノテーション対象の選定を行った．具体的には CANDIDATE 候補は CEFR レベルが B1 以下のものとし，CEFR レベル B1 以下の類義語数が 30 個以下のものからランダムサンプリングしたものをアノテーション対象とした．その結果アノテーションを行うデータは，TARGET が 124 個であり，それぞれの類義語数の合計が 918 個となった．

アノテーションの結果，CEFR レベル B1 以下の類義語中には置き換え可能 (正解) となるものが存在しないケースがあることが判明した．124 個の TARGET のうち，48 個 (約 39%) においては置き換え可能な類義語が存在しなかった．本研究では正解となる類義語が存在する TARGET 76 個およびその CANDIDATE である 573 個の類義語を評価対象とする．

4.2 ベースライン

ベースラインとして，Horn ら [6] がランキング生成の特徴量の一つとして用いた言語モデルを利用したシステム (以降では LM システムと呼ぶ) と比較する．LM システムの構成は，TARGET の選定と類義語の抽出は提案手法と同様に行い，得られた類義語に言い換えた文に対して言語モデルによるスコア付けを行う．得られたスコアを用いてランキングしたリストを MI で出力したリストと比較する．使用する言語モデルのトレーニングデータには Google N-gram から出現頻度が 1,000 以上のものを利用した．

4.3 評価指標

システムの評価指標として，出力する CANDIDATE の適合率と再現率を用いる．ここで適合率とは各システムが出力する CANDIDATE の個数に対する正解 CANDIDATE 数の割合を示し，再現率はデータ全体の正解 CANDIDATE の個数に対するシステムが出力した正解 CANDIDATE 数の割合を示す．以降ではランキング上位 n 個の CANDIDATE を用いて算出した適合率，再現率をそれぞれ precision@n，recall@n と表記する．

また，TARGET に対して置き換え可能となる CANDIDATE を出力できるかを評価するため，カバレッジも評価する．評価対象の TARGET 数に対して，1 個以上の CANDIDATE を出力できた TARGET 数の割合をカバレッジとする．

4.4 実験結果

提案手法 (以降では MI システムと呼ぶ) と LM システムが出力する CANDIDATE ランキングに対し，precision@n，recall@n を評価したグラフを図 3 に示す．両システムにおける適合率を比較すると，MI システムでは precision@1 でも高い適合率を達成しており，出力数を増やしても高い適合率を維持している．これは適切な CANDIDATE を安定して提示できることを示しており，英語教育者支援という目的に対して望ましい特性である．

一方で，MI システムではランキング上位 5 個を CANDIDATE として出力しても，出力できる CANDIDATE 数は合計 51 個となり，LM システムに比べて再現率は低い．TARGET に対するカバレッジは評価対象である 76 個の TARGET に対して，MI システムが CANDIDATE を出力できたのは 32 個であり，カ

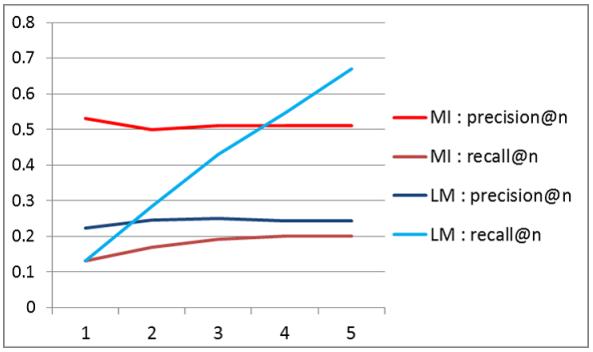


図 3: 適合率と再現率

バレッジは 42% にとどまった．この原因として，MI スコアリストの規模が小さく，TARGET および CANDIDATE と HEAD の組み合わせを充分にカバーしていないことが挙げられる．TARGET の前後 4 単語に対して MI スコアがリストにない場合は HEAD が決定できないため，CANDIDATE を出力できない．CANDIDATE を出力できなかった 44 個の TARGET のうち，24 個が HEAD を決定できなかったことが原因であった．

ただし，MI システムが 1 個以上の CANDIDATE を出力した TARGET に対して，正解となる CANDIDATE が少なくとも 1 個以上含まれる割合を求めたところ，71.9% となった．このことからも MI スコアを用いるアプローチは高い精度で CANDIDATE を決定できることがわかる．

5 まとめ

本稿では，英語教育者の支援を目的として MI スコアを用いた Lexical simplification 手法を提案した．実験の結果，言語モデルスコアにより CANDIDATE 候補のランキングを行うベースラインよりも高い精度で CANDIDATE を決定できることがわかった．一方で，CANDIDATE を推薦できる TARGET の割合がベースラインよりも小さく，提案手法のカバレッジを改善する必要があることがわかった．

今後は，Google N-gram を利用した MI スコアリストの拡充や，熟語やフレーズの simplification が出来るよう手法を拡張する予定である．

参考文献

[1] G. Barnbrook. “Language and Computers: A Practical Introduction to the Computer Analysis of Language,” Edinburgh University Press, (1996).

[2] O. Biran, S. Brody and N. Elhadad. “Putting it Simply: a Context-Aware Approach to Lexical Simplification,” Proc. of ACL, pages 496–501, (June 2011).

[3] A. J. Charles. “The CEFR and the Need for More Research,” The Modern Language Journal, Vol. 91, No. 4, pages 659–663, (2007).

[4] W. Coster and D. Kauchak. “Simple English Wikipedia : A New Text Simplification Task,” Proc. of ACL, pages 665–669, (June 2011).

[5] W. Coster and D. Kauchak. “Learning to Simplify Sentences Using Wikipedia,” Proc. of ACL, pages 1–9, (June 2011).

[6] C. Horn, C. Manduca, and D. Kauchak. “Learning a Lexical Simplifer Using Wikipedia,” Proc. of ACL, pages 458–463, (June 2014).

[7] G. Juri, B. V.Durme, and C. Callison-Burch. “PPDB: The Paraphrase Database,” Proc. of NAACL-HLT, pages 758–764, (June 2013).

[8] B. Laufer. “What Percentage of TextLexis is Essential for Comprehension,” Chapter 25 in Special language: From humans thinking to thinking machines, Multilingual Matters, pages 316–323, (1989).

[9] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard and D. McClosky. “The Stanford CoreNLP Natural Language Processing Toolkit,” Proc. of ACL, pages 55–60, (June 2014).

[10] E. Pavlick and C. Callison-Burch. “Simple PPDB: A Paraphrase Database for Simplification,” Proc. of ACL, pages 143–148, (August 2016).

[11] 九州大学大学院言語文化研究院 学術英語テキスト編集委員会. “Authentic Reader-A Gateway to Academic English,” 研究社, (July 2016).

[12] CEFR-J Wordlist Version 1 (2013) 東京外国語大学投野由紀夫研究室.