# 単語分散表現による語義の近似を用いた語彙平易化手法

Lexical Simplification Using Word Embedding to Approximate Word Sense

高田祥平<sup>\*1</sup> 荒瀬由紀<sup>\*1</sup> 内田諭<sup>\*2</sup> Shohei Takada Yuki Arase Satoru Uchida

\*1大阪大学大学院情報科学研究科

Graduate School of Information Science and Technology, Osaka University

\*<sup>2</sup>九州大学大学院言語文化研究院 Faculty of Languages and Cultures, Kyushu University

Authentic English passages are not always appropriate for learners due to their vocabulary level; hence teachers sometimes have to modify the text by making sentences simpler or replacing difficult words with easier ones. This process, however, takes time and could be a burden for teachers. The present study aims to build an automatic lexical simplification system that can assist teachers in preparing materials for classes and examinations. The proposed system first selects target words based on CEFR levels and then lists candidates from a thesaurus. Then, the paraphrasablity of each candidate is examined using a word embedding method. The results show that the proposed method can provide correct candidates for more cases than the baseline and existing methods and is robust even when the target is a polysemous word.

# 1. はじめに

英語の教育現場では,教育者は学習者の英語の習熟度に合わせて教材として用いる英文の難易度の調整を行う.例えば, 学習者の理解の妨げになる難しい語について,平易な語への言い換えや注釈をつけるなどの処理が必要となる.これは教育者 自身の経験に基づいて行われ,多くの時間と労力を要する作業 であるため負担が大きい.そこで本研究では,英語教育者の教 材作成支援を目的とし,入力された英文中の難単語を自動で 抽出し,言い換え可能で平易な語を提示する手法を提案する. 英語習熟度の高い教育者は,提示された候補語から言い換えに 適切な候補を選定できると期待できる.一方で,不要に多くの 候補を提示することは,返って教育者の負担となってしまう. そこで本研究では,多くの難単語に対してできるだけ精度の高い候補語のリストを出力するシステムの実現を目指す.

本研究では語彙平易化を以下の3つのステップで実現する. まず,言い換え対象となる難しい単語 (TARGET と呼ぶ) を 特定する.次に、TARGET の言い換え先候補となる平易な語 (CANDIDATE と呼ぶ)を選定する.最後に、入力文の文脈 に応じた CANDIDATE を判定して出力する. TARGET と 言い換え可能な CANDIDATE は、TARGET が表れる文脈 において、TARGET と同一の意味を表現し、かつ文法的に も整合性を保たなければならない. そこで, この条件を満た す CANDIDATE を判定するための指標として、本研究では 語義とコロケーションを用い、教育支援に適した言い換え手 法を提案する.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 Language (CEFR) [Alderson 07] に準拠した難易度辞書を使用し て TARGET を特定する.次に,類義語辞書から難易度辞書 を用いて CANDIDATE の選定を行う. 最後に, コロケーショ ンの評価指標を用いて有意に共起する語を特定し、単語ベク トルの学習を行う. これにより TARGET を言い換え可能な CANDIDATE を機械学習により判別する.

連絡先: 荒瀬由紀, 大阪大学大学院情報科学研究科, 〒 565-0871 大阪府吹田市山田丘1−5, arase@ist.osaka-u.ac.jp



図 1: 手法の概要図

言い換え精度の評価実験の結果,既存の単語ベクトルを用 いるベースラインや品詞情報を付与した単語ベクトルを用いる 既存手法と比較して,提案手法はより多くの TARGET に対 して正解 CANDIDATE の提示が可能であり,教育支援とし て望ましい性質を持つことが明らかとなった.

## 2. 提案手法

提案する語彙平易化手法の概要を図1に示す.提案手法で は、まず入力文から難易度の高い語を特定し、TARGETとす る.次に、類義語辞書からTARGETの類義語を取得し、平 易なものをCANDIDATEとする.得られたCANDIDATE と入力文から各種特徴量を抽出し、機械学習により言い換えが 可能と判定したCANDIDATEのリストを出力する.各手順 における処理について、詳細を以下に記述する.

### 2.1 TARGET の特定

Leroy ら [Leroy 13] は大規模コーパスにおける単語の出現 頻度を難易度の指標として TARGET を特定している.一般 的に,難易度の高い語は出現頻度が低く,平易な語は出現頻度 が高い傾向にある [Leroy 11] が,出現頻度の閾値を適切に設 定することは困難である.

本研究では、単語難易度の付与のため、CEFR-J Wordlist Version 1.3<sup>\*1</sup> を用いる. このリストは CEFR に基づいて単 語を簡単なものから順に A1, A2, B1, B2 の4段階のレベ ルで示したものである. これに English Vocaburaly Profile<sup>\*2</sup> のリストを結合することで、さらに上位の C1, C2 レベルの 単語を追加し、合計で9,277 個の見出し語を含む6段階の単 語難易度辞書を作成した. この辞書では単語の原形と品詞ご とに難易度が関連付けられている.入力文中の語に難易度を 対応させるには各単語の原形と品詞の情報が必要であるため、 Stanford CoreNLP [Manning 14] を用いた解析を行う. 難易 度の上位3段階である B2, C1, C2 レベルが付与された名詞、 動詞、形容詞、副詞となる語を TARGET とする.

## 2.2 CANDIDATE の絞り込み

CANDIDATE の絞り込みにおいて, Glavašら [Glavaš 15] は単語ベクトルの類似度を用いている. TARGET とのベクト ルのコサイン類似度の高い語を CANDIDATE としているが, 対義語が含まれる割合も高く, CANDIDATE リストに多くの 不正解候補が含まれるため,後の CANDIDATE 判別に高い 精度が要求されることになる.

本研究では、CANDIDATE リストの品質を担保するため、 CANDIDATE を Thesaurus.com<sup>\*3</sup> から作成した類義語辞書 から抽出する. CANDIDATE の選定手法について、まず入 力文から抽出した各 TARGET について類義語辞書から、同 一の見出し語と品詞に関連付けられた類義語を取り出す. 得 られた類義語に対して難易度付与を行い、TARGET 難易度 が下位 3 段階である A1, A2, B1 レベルが付与された語を CANDIDATE とする.

## 2.3 CANDIDATE の選定

本研究では、言い換え可能な CANDIDATE の判別におい て考慮すべき要件を TARGET との語義が一致すること、入 力文における TARGET と周辺の語との文法的、文脈的整合 性を維持することと定義する. これらを考慮する特徴量を抽出 し、入力文において言い換えが可能な CANDIDATE を機械 学習によって分類する.

Paetzold ら [Paetzold 16] は品詞情報を付与した TARGET のベクトルと CANDIDATE のベクトルの類似度, 2-gram お よび 3-gram の言語モデルスコア, CANDIDATE の品詞別の 出現頻度比を特徴量として組み合わせている. しかし, この 特徴量セットでは同じスペル, 同じ品詞を持つが語義が異なる ケースに対応できないと考えられる. 語義曖昧性解消の既存研 究では, Navigli ら [Navigli 09] は 25% 以上の単語について 正確に語義を判別することができておらず, 語義の高精度な判 定は未だ困難である.

既存の単語分散表現では、周辺に出現する単語によって意味を近似するが、一単語につき一つの分散表現となるため、複数の語義が混在した表現となってしまう.本研究では、ある単語と有意に共起する単語が語義を限定する手がかりとなると仮定する.そして、このような共起語を制約として単語分散表現に加えることで、語義を近似する単語分散表現を生成する. 具体的には、TARGET と CANDIDATE の組み合わせについて、入力文中に TARGET と CANDIDATE の両方と有意 に共起する語 (PairWord) があると仮定し, PairWord を付与 した単語ベクトルの類似度を用いることで語義推定を模倣し た特徴量を加える.生成する単語ベクトルは単語と PairWord の2つ組から生成するため,学習データがスパースになると予 想される.そこで提案手法ではカバレッジを向上させるため, PairWord を付与した単語ベクトル類似度に通常の単語ベクト ルの類似度,言語モデルスコアを加えた3種類の特徴量を用い る.そして Support Vector Machine (SVM) [Cortes 95] を 用いて CANDIDATE の言い換えの可否を分類する.以下で は PairWord を付与したベクトル類似度を取得する手順につ いて詳細を述べる.

#### 2.3.1 PairWord の特定

提案手法では、入力文において TARGET と最も共起関係 の強い語を PairWord として付与する. 共起関係の指標とし て Mutual Information (MI スコア) [Barnbrook 96] を用い る. 2 単語 a, bの MI スコアは以下の式で表され、f(a) お よび f(b) はそれぞれ単語 a と b のコーパス内での出現頻度、 f(a,b) は a と b のコーパス内での共起頻度、N はコーパスの 総単語数を示す. MI スコアは 2 単語の共起頻度が大きいほど 高い値を示す. 一方で、高頻度語については MI スコアは値 が小さくなる. MI スコアは事前に Wikipedia の dump デー タ\*4 を用いて計算したものを使用する.入力文中の全単語か ら TARGET との MI スコアが最も高い語を PairWord とし て抽出する.

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

#### 2.3.2 単語ベクトル類似度の計算

PairWord の情報を付与した単語ベクトルの生成のため,事前 に Wikipedia の dump データを学習データとし,word2vec の continuous bag-of-words [Mikolov 13a, Mikolov 13b] に よりベクトルの生成を行う.学習データの各単語にアンダー バーで PairWord を結合することで,PairWord の情報を各 トークンに付与することができる.得られたベクトルを用いて, TARGET と PairWord を付与したトークン,CANDIDATE と PairWord を付与したトークンのベクトルの類似度を求め, 特徴量として用いる.

## 3. 評価実験

本章では,提案手法による CANDIDATE の言い換え性能 を評価する.

### **3.1** データセット

評価用データとして、CANDIDATE の言い換えの可否をネ イティブ話者がアノテートした Gold-standard データセット を用いる.このデータは類義語辞書と CEFR レベルに基づい た TARGET と CANDIDATE について、文法的整合性、語 義的な類似度、入力文におけるコロケーションの3項目につ いて判定することで言い換え可能な CANDIDATE の判別を 行う.アノテーションの対象テキストは Rice 大学の公開教科 書データ\*<sup>5</sup> と九州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を用いる.Rice 大学の教科書データ を基に作成したデータは CEFR-LS [Uchida 18]\*<sup>6</sup> として公開 している.

<sup>\*1</sup> 東京外国語大学投野由紀夫研究室. (http://www.cefr-j.org/ download.html より 2018 年 1 月ダウンロード)

<sup>\*2</sup> http://www.cefr-j.org/download.html

<sup>\*3</sup> http://www.thesaurus.com/

<sup>\*4</sup> https://dumps.wikimedia.org/enwiki/

<sup>\*5</sup> http://cnx.org/

<sup>\*6</sup> http://www-bigdata.ist.osaka-u.ac.jp/arase/pj/ lex-simplification.zip

データの種類	TARGET 数	CANDIDATE 数	
学習データ	450	2,123	
開発データ	72	811	
テストデータ	270	3,140	

表 1: 実験データの内訳

アノテーションの結果,言い換え可能な CANDIDATE を 一つ以上持つ TARGET 792 個,その CANDIDATE 9531 個 (このうち,言い換え可能な正解 CANDIDATE は 1954 個) を 評価用データとして用いる.実験には,データセットを表 1 の ように分割して使用する.テストデータについてはドメイン の偏りを避けるため,Rice 大学の教科書データ 9 種類それぞ れから TARGET 30 個ずつをランダムに抽出したものを使用 する.学習データについては,正例と負例の偏りを小さくす るため,各 TARGET について正解 CANDIDATE と不正解 CANDIDATE の数が同数になるようにサンプリングする.

#### 3.2 評価指標

英語習熟度の高い教育者は候補リストから言い換え可能な CANDIDATE を選定できると期待できる.そのため,教育者 の教材作成支援においては、より多くの TARGET に対して 正確に正解 CANDIDATE を少なくとも一つは出力できるこ とが望ましい.この評価指標として、以下の式で示す $T_{\text{Eval}}$ を 用いる.*T*は TARGET の集合,f(t)は TARGET のtに対 して正解 CANDIDATE を一つ以上出力すれば 1,それ以外 は 0 をとる関数を示す.

$$T_{\text{Eval}} = \frac{1}{|T|} \sum_{t \in T} f(t)$$

一方で、言い換え候補として示すリスト中の不正解 CANDI-DATE は少ない方が、教育者の判断の負担を減らすため望まし い、そのため、各 CANDIDATE リストに含まれる言い換えが 可能な正解 CANDIDATE の割合が高いリストを出力すること も重要である。そこで、以下に示す  $C_{\text{Precision}}$  を用いた評価を行 う. T' は正解 CANDIDATE を一つ以上出力した TARGET の集合、 $c_{\text{correct}}$  は TARGET の t が持つ正解 CANDIDATE 数、 $c_{\text{output}}$  は t に対して出力した CANDIDATE 数を示す。

$$C_{\text{Precision}} = \frac{1}{|T'|} \sum_{t \in T} \frac{c_{\text{correct}}}{c_{\text{output}}}$$

ー般的に、出力 CANDIDATE 数が増えると、多くの TAR-GET に対して正解 CANDIDATE を出力できるようになる 一方で、各出力リストに含まれる不正解 CANDIDATE の割 合も増加する。そのため、 $T_{\text{Eval}} \geq C_{\text{Precision}}$ はトレードオフ の関係にある。

## 3.3 実験設定

比較手法として, Paetzold ら [Paetzold 16] の特徴量セット を用いる.また, PairWord を付与した単語ベクトル類似度の 有効性を評価するため,通常の単語ベクトル類似度,言語モデ ルスコアのみを組み合わせたものをベースラインとし,比較手 法とする.

SVM の実装としては Libsvm<sup>\*7</sup> を用いる. RBF カーネル を使用し、ハイパーパラメータは  $c \ge \gamma$  の 2 種類を開発デー タを用いてチューニングする.

特徴量セット	$C_{\mathrm{Precision}}$	$T_{\rm Eval}$	$T_{\rm out}$	$C_{\rm out}$
提案手法	0.43	$0.81^{*}$	255	<b>1</b> , <b>245</b>
既存手法	0.43	0.76	250	1,173
ベースライン	$0.50^*$	0.67	228	782

表 2: 特徴量セット毎の言い換え精度と出力 TARGET および CANDIDATE 数. \* は他の手法と統計的有意差を確認したことを示す.また, $T_{out}$  と  $C_{out}$  はそれぞれ出力 TARGET と 出力 CANDIDATE の合計数を表す.

特徴量セット	$C_{\mathrm{Precision}}$	$T_{\rm Eval}$	$T_{\rm out}$	$C_{\rm out}$
提案手法	0.39	$0.80^{*}$	116	622
既存手法	0.37	0.71	113	588
ベースライン	$0.42^*$	0.55	99	343

表 3: 多義語のみを TARGET とした場合の言い換え精度と 出力数.\* は他の手法と統計的有意差を確認したことを示す.

He took Psyche to his palace and <u>showered</u> her with		
gifts, yet she could never see his face.		
TARGET: shower		
正解 CANDIDATE: give		
提案手法の出力: give (PairWord: gift)		
既存手法の出力: pour		
ベースラインの出力: storm		

表 4: 言い換え成功例

#### 3.4 実験結果

評価実験の結果を表2に示す、表2より、ベースラインや 既存手法と比較して、提案手法の方が $T_{\text{Eval}}$ が高い、これは、 提案手法がより多くの対象語に対して正解候補語を出力できる ことを示している、一方で、 $C_{\text{Precision}}$ は $T_{\text{Eval}}$ とトレードオ フの関係となるが、 $C_{\text{Precision}}$ は最も高いベースラインに比べ 7%の減少に抑えられている、このことから、提案手法は教育 者支援のための語彙平易化において望ましい性能を持つことが 分かる、また、表2の右側より、他の手法と比較して提案手法 がより多くの TARGET について CANDIDATE を出力でき ていることが分かる。

Thesaurus.com において複数タブを持つ語を多義語とし, テストデータから多義語 (122 個) のみを TARGET とした場 合の評価結果を表 3 示す.全ての TARGET を使用したとき と比較し,  $C_{\text{Precision}}$  と  $T_{\text{Eval}}$  は全体的に減少しており,多義 語の言い換えが困難であることが分かる.しかし,提案手法の  $T_{\text{Eval}}$  は全ての TARGET を対象とした場合とほぼ同等の値を 示しており,他の手法と比較して多義語に頑健な手法であると いえる.

言い換えの成功例を表4に示す.動詞の shower は多くの 場合「雨が降る」や「浴びせる」という語義で用いられるが, 表4の入力文においては「与える」という語義で用いられて いる.通常の単語ベクトルや品詞情報のみを付与した単語ベク トルでは,出現頻度が高いと考えられる前者の語義に近い語の 類似度が高い.一方で,提案手法では PairWord として gift を抽出しており,正解 CANDIDATE である give を出力でき ている.

言い換えの失敗例を表5に示す. 表5では, pioneer は「開

<sup>\*7</sup> https://www.csie.ntu.edu.tw/~cjlin/libsvm/

Think back to <u>pioneer</u> days, when individuals knew how to do so much more than we do today, from building their homes, to growing their crops, to hunting for food, to repairing their equipment.

TARGET: pioneer 正解 CANDIDATE: settler 提案手法の出力: innovator (PairWord: equipment)

表 5: 言い換え失敗例 (1)

Some disciplines such as biophysics and biochemistry build on both life and physical sciences and are interdisciplinary.

TARGET: discipline 正解 CANDIDATE: area 提案手法の出力: curriculum (PairWord: biochemistry)

表 6: 言い換え失敗例 (2)

拓者」という語義で用いられているが、PairWord として不適 切な equipment を抽出しており、不正解となった. この例で は、day を PairWord として抽出できれば正解である settler を出力できると考えられる.そこで、入力文の構文解析を行い、 TARGET との依存関係を持つ語を優先的に PairWord の検 索対象とすることで、言い換え精度を高めることができると考 えられる.このような依存関係を利用して PairWord を特定 するため、対象単語との依存関係を持つ頻度を用いた MI ス コアデータの拡充が今後の課題である.

また、PairWord は適切であるが、不正解となった例を表 6 に示す。表 6 では discipline は「分野」という語義で用いられ ている。PairWord としてはその分野名である biochemistry が抽出されており、適当であると考えられるが、正解である area の出力はできなかった。この原因として、提案手法では各 単語と Pairword を連結した単語ベクトルの学習を行うため、 学習データのスパース性が顕著となったためと考えられる。こ の例における biochemisry と area は学習データにおいて 2 回 しか Pairword の関係を持っておらず、ベクトルの学習が困難 であったと推察できる。これは TARGET や PairWord が専 門用語などの低頻度語である場合において特に問題となる。そ のため、目的のドメインに応じたコーパスを学習データとして 用いることが必要であることが分かる。

# 4. まとめ

本研究では、英語教育支援を目的とした語彙平易化手法を提 案した.提案手法では、CEFR レベルに基づいた TARGET の特定、類義語辞書を用いた CANDIDATE の選定、共起語 を付与した単語ベクトルを用いた CANDIDATE の判別の 3 つのステップで語彙平易化を実現した.言い換え精度の評価実 験の結果、提案手法では既存手法よりも多くの TARGET に 対して正解 CANDIDATE を出力することができた.出力結 果を解析したところ、構文解析を利用した PairWord の特定 や、目的のドメインに適したコーパスを用いたベクトルモデル の学習により精度の向上が見込まれる.

今後は,依存構造を考慮した MI スコアデータの拡充,フ レーズを対象とした平易化への手法の拡張を行う予定である.

## 謝辞

本研究は公益財団法人 KDDI 財団の助成を受けたものです. また, Gold-standard データセットの作成にあたり,アノテー ションを行って頂きました九州大学大学院言語文化研究院の Christopher G. Haswell 准教授に感謝の意を表します.

# 参考文献

- [Alderson 07] Alderson, C.: The CEFR and the Need for More Research, The Modern Language Journal, Vol. 91, pp. 659–663 (2007)
- [Barnbrook 96] Barnbrook, G.: Language and Computers: A Practical Introduction to the Computer Analysis of Language, Edinburgh University Press (1996)
- [Cortes 95] Cortes, C. and Vapnik, V. N.: Support-vector Networks, Machine Learning, Vol. 20, No. 3, pp. 273– 297 (1995)
- [Glavaš 15] Glavaš, G. and Štajner, S.: Simplifying Lexical Simplification: Do We Need Simplified Corpora?, in Proc. of ACL-IJCNLP, pp. 63–68 (2015)
- [Leroy 11] Leroy, G. and Endicott, J. E.: Term Familiarity to indicate Perceived and Actual Difficulty of Text in Medical Digital Libraries, Digital Libraries: For Cultural Heritage, Knowledge Dissemination, and Future Creation, pp. 307–310 (2011)
- [Leroy 13] Leroy, G., Endicott, J. E., Kauchak, D., Mouradi, O., and Just, M.: User Evaluation of the Effects of a Text Simplification Algorithm Using Term Familiar- ity on Perception, Understanding, Learning, and Infor- mation Retention, Journal of medical Internet research, Vol. 15, No. 7 (2013)
- [Manning 14] Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D.: The Stanford CoreNLP Natural Language Processing Toolkit, in Proc. of ACL, pp. 55–60 (2014)
- [Mikolov 13a] Mikolov, T., Chen, K., Corrado, G., and Dean, J.: Efficient Estimation of Word Representations in Vector Space, in Proc. of ICLR (2013)
- [Mikolov 13b] Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J.: Distributed Representations of Words and Phrases and their Compositionality, in Proc. of NIPS, pp. 3111–3119 (2013)
- [Navigli 09] Navigli, R.: Word Sense Disambiguation: A Survey, ACM Computing Surveys, Vol. 41, No. 2, p. 10 (2009)
- [Paetzold 16] Paetzold, G. H. and Specia, L.: Unsupervised Lexical Simplification for Non-Native Speakers, in Proc. of AAAI, pp. 3761–3767 (2016)
- [Uchida 18] Uchida, S., Takada, S., and Arase, Y.: CEFRbased Lexical Simplification Dataset, in Proc. of LREC (2018, to appear)