4Y-06

CEFR難易度別単語リストとZipf scaleの対応付けによる教師なし難解語検出

伊藤和樹† メッサー真秀† 島本大輔† 撫中達司†

東海大学 情報通信学部†

概要

教師ありデータを必要としない難解語検出(Complex Word Identification :CWI)のための新しい教師なし学習モデルを提案する．CWI は，難しい文章をより平易な文章に変換する文章平易化タスクに活用する基礎技術として実用化が期待される.しかし，扱う単語の種類が膨大で，かつモデルの性能が教師ありデータ付きコーパスに強く依存する性質があるため，ラベル付きコーパスを用いた教師あり学習による CWI は実用化に至っていない.

本研究の目的は，言語能力を初心者の A1 からネイティブに近いレベルの C2 までの 6 段階で評価する国際指標である CEFR(Common European_Framework_of_Reference,Languages: learning,teaching,assessment)と日常会話における単語頻度を計測する Zipf scale を対応付けることにより，個人の英語学習レベルに基づいた教師なし難解語検出を行うことである

1. はじめに

文章平易化(Text Simplification)[1]とは，文章の意味を保持しつつ，読みやすさと理解力を向上させるために，難解な表現を平易な表現に変換するタスクである．これは，識字能力の低い人，自閉症などの認知障害，失語症や失読症を持つ人の読書評価ツール,として利用することができる[2]．難解語検出は文章平易化における主要タスクである．これは，文章中の語が読者にとって難しい語か否かを識別するタスクである．近年，読者の読解力レベルに合わせた文章平易化の実現を目指す研究が行われている[3]．その主な用途としては，第二言語学習者への読解支援や，教師が教材の難易度を調整するための支援がある．それに伴い，読者の読解力レベルに合わせた難解語検出の必要性が生じている.

Unsupervised Complex Word Identification Based on Mapping CEFR Difficulty Word Lists to Zipf Scale

† Kazuki Itoh Messer Matthew
Daisuke Shimamoto Tatsuji Munaka
† School of Information and Telecommunication Engineering, Tokai University

2. 先行研究

教師なし難解語検出の研究としてはいくつか提案されているが[4]，これらの先行研究では，大学の講義(e.g.,accounting,research)などのドメインテキスト固有の特徴に基づくものであり，汎用的な教師なし難解語検出は実現されていない．本研究では，CEFR 難易度別単語リストと Zipf scale の対応付けによる，ドメインテキストに依存しない，教師なし難解語検出を提案する.

3. 提案手法

3.1 提案手法の概要

本研究では，CEFR 難易度別単語リストと Zipf scale の対応付けによる難解語検出を提案する．CEFR は欧州評議会(Council of Europe)が作成した学習者の外国語能力を測るための共通の指標である．このリストでは，表 1 のように学習者のレベルごとに単語がランク付けされている[5][6]．Zipf scale は日常における単語頻度を計測する指標[7]である.

表 1.CEFR 難易度別単語リスト

	説明
A1	簡単なコミュニケーションや情報交換をするための基本的な能力．例)メニューに関する簡単な質問をし，簡単な答えを理解できる.
A2	シンプルでわかりやすい情報を扱い，身近な文脈で自己表現を始める能力.例)予測可能な簡単な話題について,日常的な会話に参加できる.
B1	身近な場面では限定的に自己表現し,非定型的な情報には一般的に対処する能力.例)銀行で口座開設を依頼できる（簡単な手続きの場合）.
B2	ほとんどの目的を達成し，様々なテーマで自分を表現することができる能力．例)訪問者を案内し,その場所について詳しい説明ができる.
C1	適切さ,繊細さ,不慣れな話題への対応力など,いかにうまく伝えるかに重点を置いたコミュニケーション能力.例)敵対的な質問にも自信をもって対応できる.自分の発言の順番を確保し,守れる.
C2	学術的または認知的に要求の多い教材を扱い,ある面では平均的なネイティブスピーカーよりも高度なレベルで言語を効果的に使用する能力.例)文章中から素早く,関連する情報を探し出し，主要なトピックを把握できる.

3.2 閾値決定方法

CEFR のレベルと Zipf scale により計算される日常における単語の頻度を表す値を対応付ける．具体的には，CEFR 難易度別単語リストの全単語(9937 語)の Zipf scale を計測し，難易度(A1～C2)ごとに正規分布として扱う．

今回は，A1 と B1，B1 と C1 の正規分布曲線の交点を求めることにより，この交点を難易度間の境界点と解釈し，これを閾値とした．

境界点の抽出においては，修飾関係のある品詞ごとにグルーピングをする．これは，全ての単語を同じグループとして扱うのではなく，品詞ごとの特性を活かして分類することにより，境界値をより明確に抽出できると考えたことによる．具体的には，名詞・形容詞リスト，動詞・前置詞リスト，副詞リストの 3 つのリストに分けた．リストごとに単語頻度の計測を行い，正規分布として扱い(図 2)，A1・B1 閾値と B1・C1 閾値を算出した(表 3)．なお，これらの閾値は少数第 4 位で四捨五入したものである．結果として，図 2，表 3 に示す様に，品詞ごとの特性を考慮し，正規分布曲線の交点を用いることで，閾値を算出することを可能にした．

図 2. 各リストの分布図

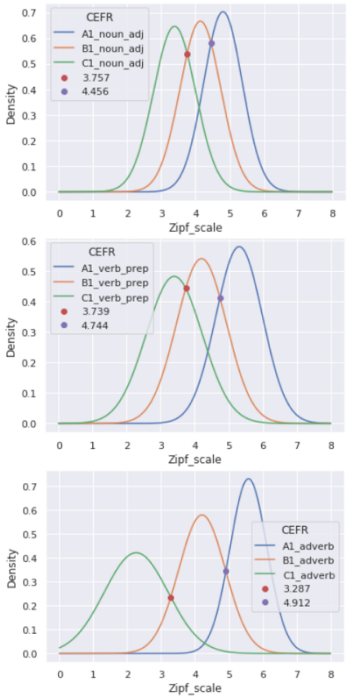

表 3. 各リストの閾値

リスト	A1・B1 閾値	B1・C1 閾値
名詞・形容詞	4.456	3.757
動詞・前置詞	4.744	3.739
副詞	4.912	3.287

表 4. 例

単語	Zipf scale	分類結果
write	5.04	A1 以上 B1 未満
publish	4.10	B1 以上 C1 未満
compose	3.46	C1 以上

4. 考察

表 4 に，本研究の教師なし難解語検出の例を示す．難解語「compose」は動詞であり，表 4 の動詞・前置詞リストによると，「compose」の Zipf scale は 3.46 であることから，C1 以上の難易度の単語と分類される．より平易な同義語として，publish(B1～C1),write(A1～B1)を挙げることができる．

Weblio によると，この難解語「compose」は英検 2 級以上とされているため，C1 以上という分類結果は相当であると考える．なお，本研究の正規分布曲線による閾値算出では，リスト間の分布が σ +-1 をもって明確に区別できることを期待したがこれは達成できなかった．

5. おわりに

本稿では，教師ありデータを必要としない難解語検出のための新しい教師なし学習モデルを提案した．CEFR 難易度別単語リストと Zipf scale による単語頻度の値を対応付けることにより，CEFR による個人の英語学習レベルに基づいた教師なし難解語検出を可能にした．これにより，ある単語の難易度を算出したい際には，その単語の Zipf scale を計測することにより，その単語がどの難易度レベル(A1,B1,C1)に属する単語であるのかを分類することが可能となる．

今後の展望としては，本研究での CEFR レベルに基づく分類結果が,難解語検出という目的に叶うものかどうかを評価する手法について検討したい．

参考文献

[1] Automated Text Simplification: A Survey
[2] Controllable Text Simplification with Explicit Paraphrasing
[3] Controllable Sentence Simplification with a Unified Text-to-Text Transfer Transformer
[4] Vicomtech at ALexS 2020: Unsupervised Complex Word Identification Based on Domain Frequency
[5] CEFR Levels https://www.examenglish.com/CEFR/cefr.php (cited 2023-January-11)
[6] Open Language Profiles English datasets https://github.com/openlanguageprofiles/olp-en-cefrj (cited 2023-January-11)
[7] wordfreq3.0.3 https://pypi.org/project/wordfreq/ (cited 2023-January-11)