複合語の構成素情報を考慮した病名難易度の推定

山本 英弥　　伊藤 薫　　荒牧 英治
奈良先端科学技術大学院大学 情報科学研究科
{yamamoto.hideya.xx7, kito, aramaki}@is.naist.jp

1 はじめに

臨床において医療者と非医療者のコミュニケーションの重要性が指摘されている．しかしながら，実際の臨床の現場ではいくつかの困難が存在する．その一つとして難解な医療用語の存在がある．正確な説明を求められる医療者にとって，難解な医療用語を使用せざるを得ないこともあるが，その場合も最大限わかりやすく説明を行う必要がある．したがって，難解な医療用語を使わざるを得ない場合は，使用者がその用語が難しい用語であることを把握し，慎重に使う必要がある．しかし，医療者はあまりに医療用語に慣れてしまっており，一般の患者（以降，非医療者）がどのような用語を難しいと感じるかを把握しにくい．このため，国立国語研究所は「難解な語リスト」を作成しているが，60 語程度しか収録されておらず，より大規模なものが求められる．そこで本研究では，既存の一般的な日本語の難易度推定手法を医療用語向けに改良し，大規模な「難解な医療用語リスト」の作成が可能な機械学習を用いた医療用語の難易度推定手法を提案する．

2 関連研究

一般的な日本語の難易度推定は，英語のそれらに比べて困難なタスクであると言われている．その理由として，英語には言語資源が豊富にある [1–7] のに対して，日本語にはあまりないことが挙げられる．特に，日本語の難易度に関する正解データが確立されておらず，日本語学習者向けに作成され，17,920 語の難易度を 6 段階で示した日本語教育語彙表[1]が用いられることが多い．

この日本語教育語彙表を用いた研究はいくつかある．梶原らは，web 上での頻度情報や単語長や文字種などに加えて， Wikipedia 本文を基にした Word2vec による単語分散表現を用いることで，難易度推定精度が向上すると報告した [8]．しかしながら，単語分散表現は不安定な特性があり，同じコーパス・同じパラメータを用いた場合でも各単語ベクトルは初期値に依存し，エポックごとに異なるという問題がある．髙田らはこの Word2vec の不安定さを実験で示し，単語分散表現の代わりに現代日本語書き言葉均衡コーパス (BCCWJ) [9] での頻度情報や Wikipedia での文字単位の頻度情報を用いる，再現が容易な難易度判定手法を提案した [10]．

[1] http://jhlee.sakura.ne.jp/JEV.html

3 予備実験

本章では，従来の日本語難易度推定手法をベースラインとして評価するため，髙田ら [10] の素性に性質の似た素性を追加し，日本語教育語彙表で難易度推定器を構築し，医療用語の難易度推定を行う．

3.1 方法

実験に用いる素性および医療用語の正解データについて述べる．なお，学習データは，BCCWJ 語彙表と日本語教育語彙表のいずれかに同一の標準的な表記が複数あるものを除外した 12,408 語（以下学習ボキャブラリと称する）を対象とし，6 段階難易度を用いる．単語の難易度推定モデルには髙田ら [10] と同じ RBF カーネルによる SVM を用いる．学習データに用いる日本語教育語彙表の難易度は離散値であるため，scikit-learn(0.18.1)[2]の SVC を用いる．パラメータ C と gamma は，グリッドサーチの結果最もスコアの高かった組み合わせを採用する．

3.1.1 素性

本研究では，髙田らの用いていた素性に以下の二つを追加したものを用いる．

a. **Twitter 頻度（1 次元）** ：MeCab[3]と mecab-ipadic-NEologd[4]によって日本語 Twitter を単語分割した際の各単語の対数頻度．

[2] https://scikit-learn.org/stable/
[3] http://taku910.github.io/mecab/
[4] https://github.com/neologd/mecab-ipadic-neologd

b. **文字単位の Twitter 頻度（4 次元）**：単語に含まれる各文字単位の日本語 Twitter 中の頻度. 文字単位の頻度の総和，平均，最大，最小それぞれに対して対数をとった値.

なお，log0 = 0 とする．予め日本語教育語彙表の 6 段階難易度で予備実験を行い，最もスコアの良かった素性の組み合わせ（Wiki 頻度 + 文字単位の Wiki 頻度 + 単語長 + 文字種の文字数 + BCCWJ 頻度 + Twitter 頻度，以降では基本素性と呼ぶ）を用いて分類器を構築し，医療用語の難易度推定に用いる．

3.1.2 データの作成

医療用語の難易度のリストは存在しないため，新たに作成する．作成に当たっては，大人数の非医療者を簡易にリクルーティングできるためクラウドソーシングを用いる．

対象語彙

今回の評価は人手で行うため，全ての語を評価対象にすることは現実的でない．従って，評価対象は医療用語リソースである万病辞書[5]に掲載されている語のうち，本研究で使用している Twitter のツイートデータと BCCWJ 語彙表の両方に登場する 260 語に限定する．

タスク設定・難易度決定

Yahoo!クラウドソーシング[6]を用いて，医療関係の業務に携わったことのない 100 人をリクルートする．その 100 人に対して，「次の単語について，ご自身にとっての難易度を 4 つの中から最も当てはまるものを選んでください」と質問し，対象語彙の難易度をそれぞれ「難しい」「やや難しい」「やや簡単」「簡単」の 4 件法で回答させた．これらに対し前から順に 1-4 点を割り当て，100 人の平均点を各語の難易度として定め，その値を crowd difficulty level (CDL) と呼ぶことにする．

3.2 実験結果・考察

医療用語の評価結果を図 1 に示す．縦軸が CDL，横軸が推定難易度である．推定難易度が 1-3 の語は表 3.2 に示す 10 語のみであった．

医療用語が少数しか推定難易度 1-3 に分類されないことについては，今回の訓練対象である日本語教育語彙表の難易度 1-3 が日常用語の基本レベルであるため妥当と言える．また，全体的に知名度が高いと思われる語が多く，CDL が小さい，すなわち非医療者にとっての難易度が低いものが多いのも妥当と言える．ただし，皮膚などにできる腫れ物を指す「よう」，う蝕という歯の病気を指す「IC」は CDL が大きいにもかかわらず，推定難易度がそれぞれ 1 と 2 であり例外的な振る舞いをしている．これらは，「よう」は助動詞の「よう」，「IC」は集積回路を表す「IC」の意味もある多義語であるため，同音異義語の頻度情報が難易度推定に影響した可能性が考えられる．推定難易度 4-6 の語については，推定難易度の群間で比較すると望ましい結果に見えるが，「耳垢」「歯垢」が推定難易度 6 となっているなど，個々の事例については課題も見られた．

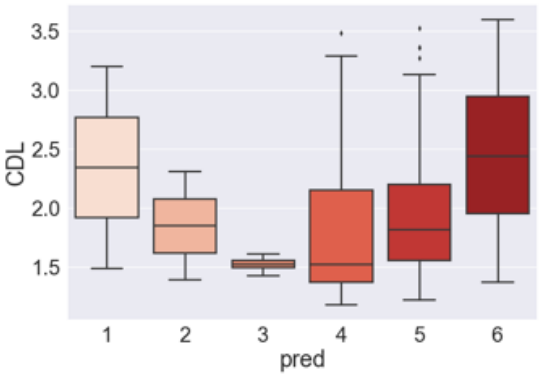

図 1: 予備実験結果

表 1: 推定難易度が 3 以下の語

推定難易度	語 (CDL)
3	自殺 (1.51)，不安 (1.56)，ストレス (1.43)，ショック (1.61)，インフルエンザ (1.49)，アレルギー (1.54)
2	IC(2.31)，心配 (1.39)
1	風邪 (1.49)，よう (3.2)

4 提案手法

前章での考察を踏まえて本章では，そもそも我々非医療者が医療用語を難解に感じる要因について調査し，それを反映した素性を追加することでより正確な難易度推定を目指す．

4.1 方法

非医療者が医療用語を難解に感じる要因の調査と，それを反映した素性および医療用語の正解データの作成について述べる．単語の難易度推定モデルには前章と同様に RBF カーネルによる SVM を用いる．学習データに用いる医療用語の難易度は連続値であるため，回帰問題として扱うために sklearn(0.18.1) の SVR[7]を

[5] http://sociocom.jp/~data/2018-manbyo/data/MANBYO_201806.zip
[6] https://crowdsourcing.yahoo.co.jp/
[7] https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html

用いる．パラメータ C と gamma は，グリッドサーチの結果最もスコアの高かった組み合わせを採用する．難易度推定結果の評価指標には，R_2，neg_MAE，neg_MSE の三種類を用いた．いずれの指標も，値が大きい方が性能が良いことを表す．

4.1.1 非医療者が医療用語を難解に感じる要因の調査

医療用語が難解に感じる要因の調査は，Yahoo!クラウドソーシングを用いて 500 人に対して行う．質問文及び回答の形式は以下の通りである．

> 医療用語が難しく感じる要因を思いつく数だけ（1〜5 個）挙げて下さい．（例と被っても構いません，各枠に要因を一つずつ書いてください．）
> 例）漢字が多い，読みづらい，聞きなれない

回答結果を検討すると，「複合的な言葉」「漢字が連続する場合に区切り位置がわからない」のような原因が挙げられており，前者は各用語をさらに形態素解析器により構成素数をカウントすることで，後者は漢字の連続数をカウントすることで新たな素性として組み込めると考えられる．その他，既存の素性で表現できそうな特徴や，素性として表現するのが困難な特徴も見られた．

4.1.2 素性

提案手法では調査結果を踏まえ，前章の基本素性に変更あるいは追加を行い，以下の素性を使用する．

A. **文字種の文字数（5 次元）**：各文字種（ひらがな，カタカナ，漢字，数字，アルファベット）の文字数．

B. **文字種の有無（5 次元）**：各文字種（ひらがな，カタカナ，漢字，数字，アルファベット）の有無

C. **文字種の連続数（3 次元）**：各文字種（カタカナ，漢字，アルファベット）の連続数

D. **構成素数（2 次元）**：語を構成する構成素数に関する情報（MeCab での分割数，MeCab 対応万病辞書[8]を追加した MeCab での分割数 - MeCab での分割数）

E. **構成素頻度情報（12 次元）**：語の構成素毎の Wikipedia, BCCWJ, Twitter それぞれにおける頻度情報

F. **文字単位の頻度（8 次元）**：語に含まれる文字毎の Wikipedia および Twitter における頻度情報

G. **先頭の文字と最後の文字の頻度情報（4 次元）**：各語の先頭の文字と最後の文字の Wikipedia, Twitter における頻度情報

H. **先頭の構成素と最後の構成素の頻度情報（4 次元）**：各語の先頭の構成素と最後の構成素の Wikipedia, Twitter における頻度情報

4.1.3 実験対象データの作成

Yahoo!クラウドソーシングを用いて，万病辞書の中から無作為に 1,000 語を選択し調査する．非医療者にとっての難易度を調査するため，医療関係の業務に携わったことない人をリクルートする．各語に対する質問は以下の五つである．

> ①： あなたにとっての難易度として最も当てはまるものを 4 つの中から選んでください．
> ②： 読めますか？（全部がカタカナまたはひらがなの語は「読める」を選んでください）
> ③： 聞いたことはありますか？
> ④： 意味は理解していますか？
> ⑤： 意味を説明できますか？

また，それぞれに対する回答は以下を選択肢とした 4 件法で行う．

> ①： 簡単だと思う・やや簡単だと思う・やや難しいと思う・難しいと思う
> ②： 読める・だいたい読める・あまり読めない・読めない
> ③： 聞いたことがある・なんとなく聞いたことがある・あまり聞いたことがない・聞いたことがない
> ④： 理解している・ある程度理解している・あまり理解していない・理解していない
> ⑤： 説明できる・ある程度説明できる・あまり説明できない・説明できない

各語に対して上記の質問に 500 人が回答した．各選択肢にはそれぞれ前から順に 1-4 点を割り当て，質問①に対する 500 人の平均点を最終的な各語の難易度として定めた．

4.2 結果と考察

医療用語難易度推定結果の一部を表 4.2 に示す．表の素性は，全提案素性はここまで紹介した全ての素性（60 次元），基本素性は前章の基本素性（23 次元），-（マイナス）は全提案素性からその素性を除いたものを表す．

[8] http://sociocom.jp/~data/2018-manbyo/data/MANBYO_201806_Dic-utf8.dic

表 2: 新規素性を加えた医療用語難易度推定結果

素性	R_2	neg_MAE	neg_MSE
全提案素性	**0.647**	**-0.244**	**-0.101**
-BCCWJ 頻度	0.635	-0.246	-0.103
-構成素頻度情報	0.578	-0.262	-0.119
基本素性	0.370	-0.320	-0.178

全提案素性を用いた場合，基本素性に比べて R_2 スコアで 27 ポイント以上高く，医療用語難易度推定に関しては大幅に精度が向上したと言える．また，全提案素性から BCCWJ 頻度を除いた場合よりも構成素頻度情報を除いた場合の方が精度が 5.7 ポイント低下した．また，BCCWJ 頻度を除いた場合でも全提案素性を用いた場合と大差なかった．

これらをまとめると，医療用語難易度推定においては，日本語教育語彙表の難易度推定で有用と言われた BCCWJ 頻度はあまり有用ではなく，むしろ構成素頻度情報のような複合語などの特徴を捉えられる素性が有用であると考えられる．また，全提案素性で日本語教育語彙表の難易度推定を行った結果，基本素性と精度に大差なかったが，これは日本語教育語彙表に含まれる語彙に複合語が少ないことが原因と考えられる．

5 おわりに

本研究では，医療者が難解な医療用語の存在について知ることが，円滑な診療において重要であると考え，より正確な医療用語難易度推定を目指した．そして，一般的な日本語に対する既存の難易度推定方法を用いた予備実験結果を踏まえ，非医療者が医療用語を難解に感じる要因を調査した．その結果，「複合的な言葉」などが難解に感じるという意見が多かったため，それらを素性として加え再度実験を行ったところ，大幅に難易度推定精度が向上した．

複合語が多いという医療用語の特徴を捉えたことが精度向上の要因であると考えられるが，より正確な難易度推定を目指すためには，今回の結果を個々に確認し，より細かに人手でチェックを行いフィードバックすることが必要であると思われる．さらに必要な素性やモデルを見つけることで，より正確で大規模な医療用語難易度リストが得られるだろう．

参考文献

[1] Tomoyuki Kajiwara and Mamoru Komachi. Building a monolingual parallel corpus for text simplification using sentence similarity based on alignment between word embeddings. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pp. 1147–1158, 2016.

[2] Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych. A monolingual tree-based translation model for sentence simplification. In *Proceedings of the 23rd international conference on computational linguistics*, pp. 1353–1361. Association for Computational Linguistics, 2010.

[3] William Coster and David Kauchak. Simple english wikipedia: a new text simplification task. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pp. 665–669. Association for Computational Linguistics, 2011.

[4] William Hwang, Hannaneh Hajishirzi, Mari Ostendorf, and Wei Wu. Aligning sentences from standard wikipedia to simple wikipedia. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 211–217, 2015.

[5] Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. Ppdb: The paraphrase database. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 758–764, 2013.

[6] Ellie Pavlick, Pushpendre Rastogi, Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. Ppdb 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, Vol. 2, pp. 425–430, 2015.

[7] Ellie Pavlick and Chris Callison-Burch. Simple ppdb: A paraphrase database for simplification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 143–148. Association for Computational Linguistics, 2016.

[8] 梶原智之, 小町守. 平易なコーパスを用いないテキスト平易化. 自然言語処理, Vol. 25, No. 2, pp. 223–249, 2018.

[9] Kikuo Maekawa, Makoto Yamazaki, Takehiko Maruyama, Masaya Yamaguchi, Hideki Ogura, Wakako Kashino, Toshinobu Ogiso, Hanae Koiso, and Yasuharu Den. Design, compilation, and preliminary analyses of balanced corpus of contemporary written japanese. In *LREC*, 2010.

[10] 髙田理功, 梶原智之, 奥村紀之. 再現が容易な単語の平易化判定手法. 人工知能学会全国大会論文集 2018 年度人工知能学会全国大会 (第 32 回) 論文集, pp. 2L403–2L403. 一般社団法人 人工知能学会, 2018.