日本語における医療用語の難易度辞書の半自動構築

杉原 壮一郎[1] 梶原 智之[1] 二宮 崇[1] 若宮 翔子[2] 荒牧 英治[2]
[1] 愛媛大学 [2] 奈良先端科学技術大学院大学
{sugihara@ai., kajiwara@, ninomiya@}cs.ehime-u.ac.jp {wakamiya, aramaki}@is.naist.jp

概要

本研究では，日本語における医療用語の難易度辞書を構築する．医療従事者と患者のコミュニケーションを円滑化するために，難解な専門用語を平易に言い換える医療テキスト平易化の研究が英語を中心に進められている．しかし，日本語では本タスクのための言語資源が充分に整備されていない．そこで我々はまず，20 代から 50 代までのアノテータ 40 人を対象に，日本語の医療用語 1 万語の難易度を調査した．そして，アノテーション結果に基づき医療用語の難易度推定器を訓練し，万病辞書に含まれる全 37 万語の難易度を自動推定した．本稿では，難易度調査および難易度推定の結果を報告する．

1 はじめに

近年，医師を中心とする医療従事者と患者の間のコミュニケーションが重要視されている [1]．医療従事者と患者のコミュニケーション不足は，診断の伝達ミスや治療方針の行き違いにつながり，医療訴訟などのトラブル [2] を引き起こしうる．医療現場におけるコミュニケーションを難しくする要因のひとつに，医療従事者と患者の間の専門知識の差がある．特に，医療用語には難解なものが多く [3]，そのままでは患者が理解することは難しいため，医療従事者は平易な表現への言い換えによって患者に分かりやすく伝える工夫が重要である．

このような課題を解決するために，英語を中心に医療テキスト平易化の研究 [4–6] が進められている．しかし，日本語では医療テキスト平易化のための辞書やコーパスが整備されていない．本研究では，日本語における医療テキスト平易化に取り組むための第一歩として，医療用語の難易度辞書を構築する．

我々はまず，クラウドソーシングによって 40 人の被験者を募集し，非医療従事者に対して医療用語の難易度を調査した．1 万語の医療用語への難易度調査の結果を分析したところ，年齢を重ねるごとに未知の医療用語が減少することや，男性が妊娠や出産に関する医療用語を知らない傾向にあるなど，被験者の属性ごとの特徴が観察された．さらに，医療用語の文字種や頻度，単語分散表現などの素性を用いた機械学習によって医療用語の難易度推定器を訓練し，既存の単語難易度推定器よりも高い性能を達成した．最終的に，表 1 に例示するように，万病辞書[1] [7] に含まれる 37 万種類の日本語の医療用語（病名・症状表現）の難易度を推定し，難易度辞書[2] を公開した．

表 1 医療用語の難易度推定の例

難易度	医療用語
1（平易）	衰弱, 偏食, めまい, ニコチン依存症
2	植物状態, 感電死, 色覚異常, 胃癌検診
3	急性胃腸炎症状, 脳障害, 若年性脱毛症
4	術後食道裂孔ヘルニア, 後天性てんかん
5（難解）	掌蹠膿疱症性骨関節炎, HCAHC

2 関連研究

単語難易度辞書は，テキスト平易化 [8–10] の研究に応用されている．英語では，Pavlick and Nenkova [11] は，平易なコーパスと通常のコーパスの単語出現確率の対数比を用いて単語難易度を推定した．また，Pavlick and Callison-Burch [12] および Maddela and Xu [13] は，語彙の一部に対して人手で単語難易度を付与し，そのアノテーション結果に基づき単語難易度を推定する機械学習モデルを訓練した．日本語の特に医療ドメインにおいては，平易に書かれた大規模コーパスを使用できないため，前者のアプローチは採用できない．そこで本研究では，後者のアプローチで単語難易度辞書を構築する．

日本語の単語難易度辞書としては，言語教育の分野で，日本語能力試験出題基準語彙表[3] や日本語教

1） https://sociocom.naist.jp/manbyou-dic/
2） https://github.com/EhimeNLP/J-MeDic-Complexity
3） https://www7a.biglobe.ne.jp/nifongo/data/

育語彙表 [14] などが開発されている．自然言語処理の分野では，梶原ら [15] が多クラス分類問題として単語難易度の推定に取り組んでおり，日本語教育語彙表を教師データとして，品詞，文字種，頻度，単語分散表現 [16] を素性とする Support Vector Machine (SVM) を訓練している．医療用語の難易度推定に取り組んだ山本ら [17] は，非医療従事者が医療用語を難解に感じる要因として複合語を挙げており，文字種や頻度に加えて文字種ごとの文字数や構成素数の素性を用いて難易度推定の性能を改善した．本研究でも，これらの先行研究と同様に，機械学習ベースの単語難易度推定器を訓練する．

3　日本語の医療用語の難易度調査

3.1　クラウドソーシングによる難易度調査

本研究では，医療用語の難易度辞書を構築するために，非医療従事者を対象に難易度調査を実施した．医療用語は，日本語の大規模な病名辞書である万病辞書[1] [7] のうち，信頼度レベルの高い先頭 3 万語の中から 1 万語を無作為抽出した．

多様な被験者を対象とするために，年代（20 代，30 代，40 代，50 代）と性別（男性，女性）の組み合わせで 8 グループからそれぞれ 5 人ずつ，合計 40 人の被験者をクラウドソーシングサービスのランサーズ[4]で募集した．なお，被験者には 1 語あたり 1 円の報酬（時給 1,000 円の想定）を支払った．

非医療従事者である被験者に対して，1 万語のそれぞれに以下の 5 段階の難易度の付与を依頼した．

1. 日常会話の中で使う
2. 使ったことがある
3. 意味が分かる
4. 見た／聞いたことはあるが，意味は分からない
5. 見た／聞いたことがなく，意味が分からない

高品質な調査のために，被験者に対して 2 種類のフィルタリングを実施した．まず，300 件の小規模なアノテーションを依頼し，その回答を著者が確認し，問題のない被験者にのみ残り 9,700 件を継続依頼した．また，1 万件のアノテーション終了後に，年代と性別のグループごとにアノテーションの一致率を求めた．Quadratic Weighted Kappa (QWK) [18] を計算し，グループ内の誰かとの一致率が 0.3 を下回る被験者は除外し，新たに被験者を募集した．

4）　https://www.lancers.jp

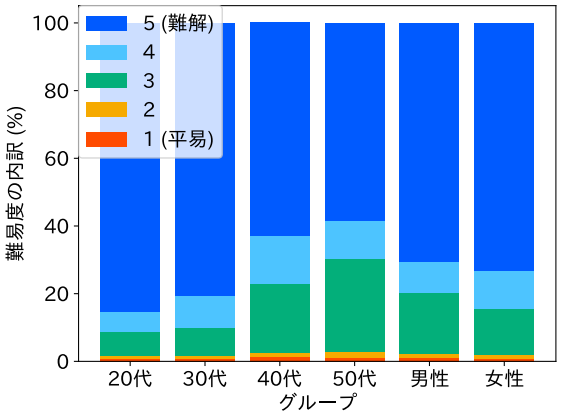

図 1　年代や性別ごとの難易度の分布

3.2　難易度調査の結果の分析

前節における医療用語の難易度調査の結果から，年代や性別ごとの特徴を分析する．まず，年代や性別ごとの難易度ラベル（1 が最も平易で 5 が最も難解）の分布を図 1 に示す．20 代や 30 代では 1 割程度の医療用語しか理解できないが，年齢を重ねることに知らない医療用語が減っていくことがわかる．しかし，50 代でも 7 割以上の医療用語の意味が分からない（4 の水色または 5 の青色）ことが明らかになった．なお，性別による医療用語の難易度の差はあまりなかった．

次に，特定の年代以上で意味が分かる医療用語の事例を観察する．ある年代以上の全員が 1〜3 の平易な難易度ラベルを付け，それ未満の年代において 1 人以上が 4 または 5 の難解な難易度ラベルを付けた医療用語の一部を表 2 に示す．日常会話で使うことの多い「しゃっくり」や「かぜ」などは全員が知っており，年齢を重ねるにつれて患者が増加する傾向にある「食道ポリープ」や「大腿骨骨折」などは 40 代以上または 50 代以上の被験者のみが知っている．これらの結果から，難易度調査に年代ごとの特徴が反映されていると示唆される．

最後に，特定のグループには意味が分からない医療用語の事例を観察する．あるグループ内の全員が 4 または 5 の難解な難易度ラベルを付け，その他のグループでは 1 人以上が 1〜3 の平易な難易度ラベルを付けた医療用語の一部を表 3 に示す．20〜30 代の若い男性は，「異常胎位」や「早発卵巣不全」などの妊娠や出産に関する医療用語の一部を知らないことがわかった．これらの結果から，難易度調査に性別ごとの特徴が反映されていると示唆される．

表 2　特定の年代以上で意味が分かる医療用語の例

グループ	医療用語
全員	息切れ, しゃっくり, かぜ, 口内炎
30 代以上	細菌性結膜炎, 水銀中毒, 腎尿管結石
40 代以上	肺移植ドナー, 食道ポリープ, てんかん発作
50 代以上	大腿骨骨折, 痴呆症, 脳死状態, 近視

表 3　特定のグループには意味が分からない医療用語の例

グループ	医療用語
20〜30 代の男性	双胎妊娠新生児, 異常胎位, 口蓋切創, 早発卵巣不全, ウェルシュ菌食中毒
20〜30 代の女性	胃穿孔, 単純ヘルペス, 慢性前立腺炎, 慢性アルコール性脳症候群, 嗄声

4　日本語の医療用語の難易度推定

本研究では，医療用語の難易度推定に関する先行研究 [17] で用いられた 3 種類の素性（基本素性）に加えて，新たに 3 種類の提案素性を用いて，機械学習ベースの難易度推定器を訓練する．機械学習には，先行研究 [15, 17] と同様に SVM を用いる．

素性として，先行研究 [17] では Twitter から得た単語頻度を用いているが，Twitter の API 制限の変更にともないデータの取得が難しくなったため，本研究では使用しない．また，現代日本語書き言葉均衡コーパス（BCCWJ）からの単語頻度も，先行研究 [17] において効果が見られなかったとの報告に従い，本研究では使用しない．

4.1　基本素性

文字種に関する素性　これは，医療用語を構成する文字の種類（ひらがな，カタカナ，漢字，数字，英字）を表現する素性である．各文字種の有無を表現する 0/1 バイナリ素性（5 次元），文字種ごとの文字数を表現する整数素性（5 次元），各文字種が最大何文字連続しているかを表現する整数素性（5 次元）の合計 15 次元で構成される．

構成素に関する素性　これは，医療用語を構成する形態素数を表現する素性である．用語を MeCab[5] (IPADIC) [19] で分かち書きした際の形態素数を表現する 1 次元の整数素性で構成される．

頻度に関する素性　これは，医療用語を構成する文字や形態素のコーパス中での出現頻度を表現する素性である．医療用語に含まれる形態素の頻度の合計，平均，最大，最小，先頭の形態素の頻度，末尾の形態素の頻度の 6 種類の頻度情報を素性として用いる．コーパスには日本語 Wikipedia を用い，形態素解析器には MeCab を用いた．頻度は対数化して使用し，先行研究 [17] と同様に，頻度が 0 の場合は $\log 0$ の代わりに 0 とした．同様の素性抽出を文字単位でも実施し，合計 12 次元の実数素性を用いる．

5）https://taku910.github.io/mecab/

4.2　提案素性

提案素性 1：Web コーパス上での頻度　大規模 Web コーパスの CC-100[6] [20] を用いて，基本素性と同様に頻度を数え，12 次元の実数素性を用いる．複数のコーパスを用いて頻度を数えることは単語難易度の推定に貢献することが知られている [21] が，先述のとおり本研究では先行研究 [17] で用いられた Twitter や BCCWJ のコーパスを使用しないため，代わりに大規模 Web コーパスを採用する．

提案素性 2：医療用語の単語単位の頻度　先行研究 [17] では文字や形態素の単位で頻度を数えているが，本研究では医療用語を分割せずに単語単位でも頻度を数え，素性に追加する．これは，MeCab での形態素解析の際に，万病辞書 [7] の MeCab 用辞書[7] を使用することで実現する．Wikipedia と CC-100 の両コーパスで医療用語の単語単位の頻度を数え，それぞれ対数化して 2 次元の実数素性を用いる．

提案素性 3：単語分散表現　先行研究 [15] でも用いられている単語分散表現を本研究でも用いる．訓練済みの fastText[8] [22] を採用し，医療用語が複数の構成素から成る場合はそれらのベクトルを平均し，300 次元の実数素性を用いる．

5　評価実験

1 万語の医療用語に対する難易度調査のアノテーションを用いて，難易度推定の評価実験を行う．

5.1　実験設定

データ　3.1 節で収集した 40 人分の難易度ラベルを平均し，整数になるように四捨五入し，1 万語の医療用語に対する正解の難易度（1 から 5 までの 5 ラベル）を定義する．ラベルが順序尺度であるため，評価指標には正解率とともに QWK [18] を用いた．表 4 に示すように，訓練データと評価データを 9:1 の割合で無作為に分割して実験した．本データ

6）https://data.statmt.org/cc-100/
7）https://sociocom.naist.jp/j-meddic-for-mecab/
8）https://fasttext.cc/docs/en/crawl-vectors.html

表 4 データセットの統計

ラベル	1	2	3	4	5	合計
訓練	33	100	341	2,650	5,876	9,000
評価	4	11	38	294	653	1,000
合計	37	111	379	2,944	6,529	10,000

セットは，最も平易なラベル 1 が 0.4%で最少，最も難解なラベル 5 が 65%で最多という不均衡データであるため，層化分割[9]を用いて訓練データと評価データのラベル比率が等しくなるように調整した．

モデル　難易度推定モデルには，scikit-learn（1.3.2）の SVM[10]（RBF カーネル）[23] を用い，多クラス分類モデルを実装した．ハイパーパラメータの C は {1, 5, 10, 50, 100} の中から，gamma は {0.0001, 0.0005, 0.001, 0.05, 0.1} の中から，5 分割交差検証のグリッドサーチ[11]によって QWK が最高となる組み合わせを採用した．素性は標準化[12]して使用した．

比較手法　提案手法の有効性を検証するために，2 種類のベースラインと性能を比較する．ひとつは，最頻クラスである 5 のラベルを常に出力する単純なベースラインである．もうひとつは，先行研究 [17] を再現した 4.1 節の基本素性のみを用いるベースラインである．提案手法には，4.1 節の基本素性に加えて 4.2 節の提案素性も用いる．

5.2　実験結果

実験結果を表 5 に示す．基本素性のみを用いるベースラインは，最頻クラスと同等の正解率であり，充分な性能が得られていない．提案手法は，基本素性ベースラインと比べて正解率で 14 ポイント，QWK で 28 ポイントと，大きく性能を改善できた．

それぞれの提案素性の有効性を明らかにするために，提案手法から提案素性をひとつずつ除外するアブレーション分析を実施した．どの素性を除外した場合にも，正解率と QWK の両方が低下することから，提案素性はいずれも有用であることがわかる．なお，提案素性 3 を除外した際に大きく性能が低下することから，単語分散表現が特に重要であると言える．また，提案素性を単独で使用して難易度推定を試みたところ，提案素性 1 のみで比較手法を上回る性能を達成できたことから，大規模 Web コーパスから得られる単語頻度の情報が医療用語の難易度推定に有用であることが明らかになった．

表 5 医療用語の難易度推定の実験結果

	正解率	QWK
最頻クラス	0.653	-
基本素性のみ	0.653	0.456
提案手法	**0.793**	**0.732**
提案手法 w/o 提案素性 1	0.782	0.729
提案手法 w/o 提案素性 2	0.785	0.695
提案手法 w/o 提案素性 3	0.718	0.612
提案素性 1 のみ	0.658	0.483
提案素性 2 のみ	0.612	0.444
提案素性 3 のみ	0.768	0.660

6　おわりに

本研究では，非医療従事者 40 人を対象に 1 万語の日本語医療用語の難易度調査を実施し，難易度推定器を構築した．難易度調査からは，年齢を重ねるごとに未知の医療用語が減るとはいえ，50 代でも 7 割以上の医療用語は難解であることが明らかになった．難易度推定の実験からは，大規模 Web コーパスから得られる単語頻度や単語分散表現の素性が有用であることが明らかになった．最終的に，約 8 割の正解率で 5 段階の難易度を分類できる日本語医療用語の難易度推定器を構築し，万病辞書の用語を対象とする難易度辞書を公開した．医療用語は，本研究で扱った病名・症状名だけでも 37 万語と膨大であり，また近年の「COVID-19」の出現など，日々更新されている．そのため，患者への調査を経ずに迅速に単語難易度を推定できる本研究の貢献は大きい．

今後の課題として，患者数や罹患した際の症状の深刻度など，医療分野に特有の素性を追加することで，難易度推定の性能を改善したい．また，病名・症状名に加えて，医薬品や人体部位，検査用語など，より多様な医療用語の難易度推定に展開していきたい．なお，本研究で扱った難易度は患者が判断する難易度であり，患者自身は平易だと思っていても，医学的には誤解されている用語 [3] については検出できない．これも今後の課題として取り組みたい．

9）https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html
10）https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html
11）https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html
12）https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

謝辞

本研究は，内閣府総合科学技術・イノベーション会議の戦略的イノベーション創造プログラム（SIP）第 3 期「統合型ヘルスケアシステムの構築」（研究推進法人：JST）の助成を受け実施された．

参考文献

[1] Jennifer Fong Ha and Nancy Longnecker. Doctor-patient Communication: A Review. **Ochsner Journal**, Vol. 10, No. 1, pp. 38–43, 2010.

[2] 小坂義弘. インフォームドコンセントの意味を考え直す. 日本臨床麻酔学会誌, Vol. 20, No. 10, pp. 595–597, 2000.

[3] 田中牧郎. 病院の言葉をわかりやすく―国語研究所の取り組みを通じて―. 日本内科学会雑誌, Vol. 101, No. 4, pp. 1145–1149, 2012.

[4] Gondy Leroy and James E. Endicott. Combining NLP with Evidence-Based Methods to Find Text Metrics Related to Perceived and Actual Text Difficulty. In **Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium**, pp. 749–754, 2012.

[5] Sebastian Joseph, Kathryn Kazanas, Keziah Reina, Vishnesh Ramanathan, Wei Xu, Byron Wallace, and Junyi Jessy Li. Multilingual Simplification of Medical Texts. In **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pp. 16662–16692, 2023.

[6] Ziyu Yang, Santhosh Cherian, and Slobodan Vucetic. Data Augmentation for Radiology Report Simplification. In **Findings of the Association for Computational Linguistics: EACL 2023**, pp. 1922–1932, 2023.

[7] Kaoru Ito, Hiroyuki Nagai, Taro Okahisa, Shoko Wakamiya, Tomohide Iwao, and Eiji Aramaki. J-MeDic: A Japanese Disease Name Dictionary Based on Real Clinical Usage. In **Proceedings of the Eleventh International Conference on Language Resources and Evaluation**, pp. 2365–2369, 2018.

[8] Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. Optimizing Statistical Machine Translation for Text Simplification. **TACL**, Vol. 4, pp. 401–415, 2016.

[9] Reno Kriz, Eleni Miltsakaki, Marianna Apidianaki, and Chris Callison-Burch. Simplification Using Paraphrases and Context-Based Lexical Substitution. In **Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 207–217, 2018.

[10] Sanqiang Zhao, Rui Meng, Daqing He, Andi Saptono, and Bambang Parmanto. Integrating Transformer and Paraphrase Rules for Sentence Simplification. In **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing**, pp. 3164–3173, 2018.

[11] Ellie Pavlick and Ani Nenkova. Inducing Lexical Style Properties for Paraphrase and Genre Differentiation. In **Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 218–224, 2015.

[12] 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**, pp. 143–148, 2016.

[13] Mounica Maddela and Wei Xu. A Word-Complexity Lexicon and A Neural Readability Ranking Model for Lexical Simplification. In **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing**, pp. 3749–3760, 2018.

[14] Yuriko Sunakawa, Jae ho Lee, and Mari Takahara. The Construction of a Database to Support the Compilation of Japanese Learners' Dictionaries. **Acta Linguistica Asiatica**, Vol. 2, No. 2, pp. 97–115, 2012.

[15] 梶原智之, 西原大貴, 小平知範, 小町守. 日本語の語彙平易化のための言語資源の整備. 自然言語処理, Vol. 27, No. 4, pp. 189–210, 2020.

[16] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed Representations of Words and Phrases and their Compositionality. In **Advances in Neural Information Processing Systems**, pp. 3111–3119, 2013.

[17] 山本英弥, 伊藤薫, 荒牧英治. 複合語の構成素情報を考慮した病名難易度の推定. 言語処理学会第 25 回年次大会, pp. 1495–1498, 2019.

[18] Jacob Cohen. Weighted Kappa: Nominal Scale Agreement Provision for Scaled Disagreement or Partial Credit. **Psychological Bulletin**, Vol. 70, No. 4, pp. 213–220, 1968.

[19] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying Conditional Random Fields to Japanese Morphological Analysis. In **Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing**, pp. 230–237, 2004.

[20] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised Cross-lingual Representation Learning at Scale. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 8440–8451, 2020.

[21] Tomoyuki Kajiwara and Mamoru Komachi. Complex Word Identification Based on Frequency in a Learner Corpus. In **Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications**, pp. 195–199, 2018.

[22] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching Word Vectors with Subword Information. **TACL**, Vol. 5, pp. 135–146, 2017.

[23] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine Learning in Python. **JMLR**, Vol. 12, No. 85, pp. 2825–2830, 2011.