修士論文

複合語の構成素情報を考慮した医療用語難易度の推定

山本 英弥

2019 年 3 月 8 日

奈良先端科学技術大学院大学
情報科学研究科 情報科学専攻

本論文は奈良先端科学技術大学院大学情報科学研究科に
修士（工学）授与の要件として提出した修士論文である．

山本 英弥

審査委員：

松本 裕司 教授	（主指導教員）
中村 哲 教授	（副指導教官）
田中 宏季 助教	（副指導教官）
荒牧 英治 特任准教授	（副指導教官）

複合語の構成素情報を考慮した医療用語難易度の推定*

山本 英弥

内容梗概

近年，インフォームド・コンセントという語が注目を浴びている．この語は，患者などの非医療者が治療などに関して説明を受け，内容を理解した上で，自らの意志で医療者の方針に合意するという意味である．この語が示すように，質の高い医療を実現するために，医療者と非医療者のコミュニケーションの重要性が高まっている．それに伴い，医療者はより正確で分かりやすい説明を求められるようになってきた．しかし，限られた時間の中で，説明の正確さとわかりやすさを両立するには，難解な医療用語の使用を避けられない場合があり，全ての語を噛み砕いて説明することができないと考えられる．また，医療者は医療用語に慣れているために，非医療者にとって難解な医療用語が理解しづらく，噛み砕いて説明するべき語がわからない可能性がある．したがって，両者のコミュニケーションの質をより高くするためにまず必要なことは，多くの非医療者にとって難解な医療用語を，お互いが把握することだと考えられる．このような背景のもと，国立国語研究所は，アンケートを用いて非医療者にとって難解な医療用語を取り上げ，それらがどのような要因で難解に感じられるかを調査した「『病院の言葉』をわかりやすくする提案」を公開した．しかし対象語彙は約 60 語と小規模であり，実用にはより大規模なものが求められる．そこで本研究では，大規模な「難解な医療用語リスト」を作成するために，機械学習を用いた医療用語の難易度推定手法は確立されていないため，一般的な日本語の難易度を推定する既存の手法に基づいて新たに提案する．クラウドソーシングを用いて医療用語が難解に感じる要因を調査した結果，複合語が難解であるということがわかった．これを構成素数などの素性として表現し，既存の手法に追加して医療用語の難易度推定をした結果，既存の手法と比較して決定係数で 27 ポイントの向上を実現した．これにより自動での「大規模な難解な医療用語リ

*奈良先端科学技術大学院大学 情報科学研究科 情報科学専攻 修士論文, 2019 年 3 月 8 日.

スト」作成を実現した.

キーワード

自然言語処理，医療情報，クラウドソーシング，語彙難易度，医療用語

Estimation Methods for Medical Term's Difficulty Utilizing Information on Constituents*

Hideya Yamamoto

Abstract

As we often hear the term 'informed consent' these days, good communication between the medical staffs and the non-medicals (e.g.patients) is paid much attention in medical context for the improvement of quality medical care quality. To communicate with patients effectively, medical staffs are required to satisfy both of precision and understandabilty of explanations. Though medical staffs sometimes cannot avoid to use difficult medical terms for precise explanation, these terms also can be obstacles for non-medicals to understand. In such cases, medical staffs have to add more explanations on these terms so that patients can understand. Moreover, medical staffs are so familiar with medical terms that they cannot instantly distinguish which terms are difficult for non-medicals. For the first step to solve these problems, it is important to understand which words are difficult for many non-medicals and which words need additional explanations for effective communication. As the first attempt, National Institute of Japanese Language and Linguistics (NINJAL) published 'list of difficult medical terms', using questionnaire to investigate what factors are related to the difficulty of 60 target medical terms. However, 60 terms is not large enough for practical use, so this study propose automated method of estimating medical terms' difficulty which enables to get larger one, using machine learning. Investigation via crowdsourcing showed that complex terms are difficult for non-medicals. The suggested method reflects this nature of medical

*Master's Thesis, Department of Information Science, Graduate School of Information Science, Nara Institute of Science and Technology, March 8, 2019.

terms by adding information on constituents of a noun compound as features to an existing method. In the experiment of estimating medical terms' difficulty, the new features increased the R^2 score by 27 points compared with an existing method, which was proposed to estimate common words' difficulty.

Keywords:

Natural Language Processing, medical informatics, crowdsourcing, vocabulary levels, medical terms

目次

図目次

第 1 章　はじめに

1.1　背景

近年，インフォームド・コンセントという語が注目を浴びている．これは，患者などの非医療者が治療などに関してよく説明を受け，十分に内容を理解した上で，自らの意志で医療者の方針に合意するということである．このように，質の高い医療の実現のために，医療者と非医療者のコミュニケーションの重要性が高まっている．それに伴い，医療者はより詳しい説明や分かりやすい説明を求められるようになってきた．

しかし，実際の臨床の現場では，医療者と非医療者のコミュニケーションが完全には成立していない場合があると考えられる．例えば，医療者から以下のような処置の説明を口頭で受ける場面を考える．

> 今回行う処置の狙いは，炎症を抑えることです．今回は ERCP を使った手術です．ERCP というのは内視鏡検査の一つです．嚢胞があると思われるところへ体外から造影剤を入れるルートが確保できないためです．処置そのものは外科的な手術になります．管を肋骨の下あたりから刺して，ERCP を見ながら嚢胞がありそうな位置まで入れます．これで上手くいけば，炎症の原因と思われる嚢胞が体外へ出ていくルートが確保できるので，炎症も下がると思われます．ただし今回行う処置に伴うリスクとしては，出血が多くなる可能性が挙げられます．血液検査の結果を見てると，血小板が少ないようなので，先に血小板の輸血をして，それから行います．

これらの説明は，インタラクションなく一方的にされることも多く，患者の理解の度合いには個人差があると考えられる．例えば上の例において，「炎症」，「内視鏡検査」，「造影剤」，「ルート」，「嚢胞」，「血小板」，「輸血」といった医療用語が出現するが，それらの意味をそれぞれ理解できるかどうかは各々の持っている知識次第であると考えられる．

ここで難しい点は，限られた時間で正確な説明を求められる医療者にとって，難解な医療用語の使用を避けることが不可能な場合があるということである．また，上述の「炎症」などの医療用語を全て噛み砕いて説明すると，説明全体がわかりにくくなると考える可能性がある．そのため，噛み砕いて説明する語は最低限に抑える，例えば上述の語のうち「ルート」と「嚢胞」だけは説明を加える，などの工夫が必要である．しかし，非医療者にとってどういった語が難解であるのかを医療者が理解できていないため，噛み砕いて説明するべき語の判断ができない．

以上を踏まえると，コミュニケーションの質をより高くするためにまず必要なことは，どういった医療用語が多くの非医療者にとって難解なのかを，医療者が理解しておくことだと考えられる．それによって，最低限それらの語については非医療者が理解できるよう，医療者が特に注意深く説明することができると思われる．

こういった背景の下，国立国語研究所は注意すべき医療用語をアンケートなどを用いて約 60 語を取り上げた「『病院の言葉』をわかりやすくする提案」を公開した*．これらの語は，「もう少し説明を加えたほうがいい語」や「噛み砕いて説明したほうがいい語」などとそれぞれに特徴がつけられ類型化されて公開された．しかしながら，この提案は 60 語程度と小規模であり，これらの他にも注意すべき語は多数存在すると考えられ，実用面を考えるとより大規模なものが必要である．

1.2　本研究の概要と貢献

前節の背景を踏まえて，国立国語研究所の提案のような手作業では大規模なリストの提示が難しいため，機械学習を用いることで各語に対する難易度の決定を自動化し，より高精度な医療用語の難易度推定を行い，大規模なリストの提示を可能とする手法を提案する．しかし医療用語の難易度推定手法は確立されていない．そのため本研究は，まず（A）従来の日本語難易度推定手法を用いたベースライン手法の作成および検討を行い，その結果を踏まえて（B）医療用語難易度推定のために用意した素性を用いた難易度推定手法の検討を行う．

前者（A）に関して，コーパスでの頻度情報が語の難易度推定に有効であるという先行研究の主張や，ユーザの大半が非医療者である SNS（ソーシャル・ネット

*http://www2.ninjal.ac.jp/byoin/

ワーキング・サービス）での出現頻度情報が医療用語の難易度推定にも有用であろうという予測のもと，各語の Twitter での出現頻度（以降，Twitter 頻度）を素性に加えて一般語の難易度推定を行った．その結果，既存の素性と比較して精度が向上した．この結果を考慮し，既存の素性に Twitter 頻度を追加したものをベースライン手法とし，医療用語の難易度推定を行った．その結果，同音異義語の頻度情報が難易度推定結果に影響を与えていると考えられる例など，素性が不十分であることが考えられる結果が得られた．

後者（B）に関しては，前者の結果を考慮して，医療用語の難易度推定に有用な素性を追加するために，まずクラウドソーシングを用いて医療用語が難解に感じる要因の調査を行った．調査結果として，「複合語が難しい」などの結果が得られたため，各語の構成素に関する情報などを素性として加え，医療用語の難易度推定を行った．その結果，構成素数などの複合語の特徴を考慮した素性を加えることで，医療用語難易度推定の精度が大幅に向上した．

本研究の貢献は大きく以下の二つである．（1）一つ目は，新たな素性を追加することで医療用語の難易度推定精度を向上させたという手法面での改良である．先行研究は外国人教育用である日本語教育語彙表の難易度を対象にしていたのに対し，本研究は日本人向けの医療用語難易度を対象にしており，医療用語以外にも，特有の表現を多く用いる専門用語に対しても有用である可能性がある．（2）二つ目は，機械学習を用いた手法であるため，手作業に比べると精度は劣るものの，国立国語研究所の提案に比べて大規模な医療用語の難易度に関するリストが得られたという結果面での貢献である．

1.3　本論文の構成

本論文の構成は，以下の通りである．2 章で日本語難易度推定と医療用語の難易度に関する関連研究について述べる．1.2 節の（A）に関しては，3 章で一般語の難易度推定実験でベースライン手法を決定し，4 章でそのベースライン手法を用いた医療用語の難易度推定を行い，ベースライン手法での結果を検討する．1.2 節の（B）に関しては，5 章で，ベースライン手法での課題に対応するため，医療用語が難解に感じる要因の調査をし，その結果を反映した医療用語難易度推定について述

べる．最後に 6 章で結論を述べる．

第 2 章　日本語難易度推定と医療用語の難易度に関する関連研究

本章では本研究に関連する研究について紹介する．本研究では医療用語の難易度推定を主な目的としているが，確立された手法はない．そのため，既存の一般的な日本語の難易度推定手法に基づいて医療用語の難易度推定を試みる．2.1 節で医療用語の難易度に関する関連研究を紹介し，2.2 節で日本語の難易度に関する関連研究について述べ，2.3 節で関連研究のまとめと本研究の方針を述べる．

2.1　医療用語の難易度に関する関連研究

医療用語の難易度に関する研究は，我々の知る限りほとんど存在しない．国立国語研究所が公開した「『病院のことば』をわかりやすくする提案」*が唯一の例といえる．この提案は，医療者と非医療者のコミュニケーションを円滑に行うために，両者間で使用される単語の意味が異なっていたり，非医療者が理解できていなかったりすることによって，医師から意味が伝わりにくい医療用語を，アンケートなどを用いてピックアップしたものである．この提案の優れている点は，人手での調査を行ったために信頼性が高いことと，それぞれの医療用語の性質を類型化したことだ．それぞれの医療用語の意味が伝わらない原因を，1．患者に言葉が知られていない，2．患者の理解が不確か，3．患者に心理的負担があるの三つに分類し，それぞれ分かりやすく伝える工夫を提示している（図 2.1）．さらに選定した 57 語については，具体的にそれらの類型に当てはめ，非医療者が最低限知っておくべき点や注意すべき点などをまとめて公開している（図 2.2）．しかしながらこの提案は，提示した語数が 57 語と小規模である．煩雑な手続きプロセスを経て作られたものであるため，あまねく全ての医療用語に対して同様に公開することは困難である．

*http://www2.ninjal.ac.jp/byoin/

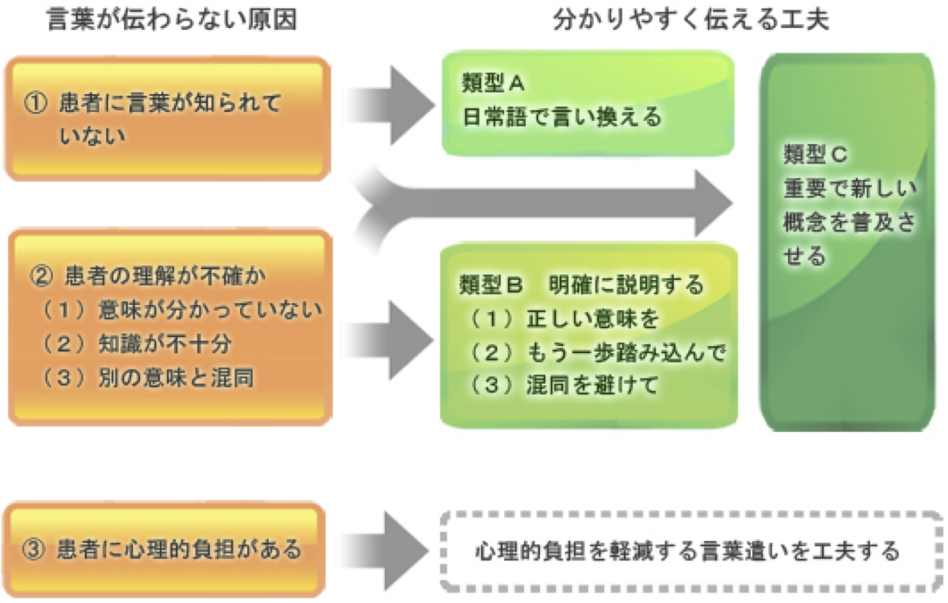


図 2.1　医療用語の意味が伝わらない原因とそれぞれに対する類型（http://www2.ninjal.ac.jp/byoin/より抜粋）

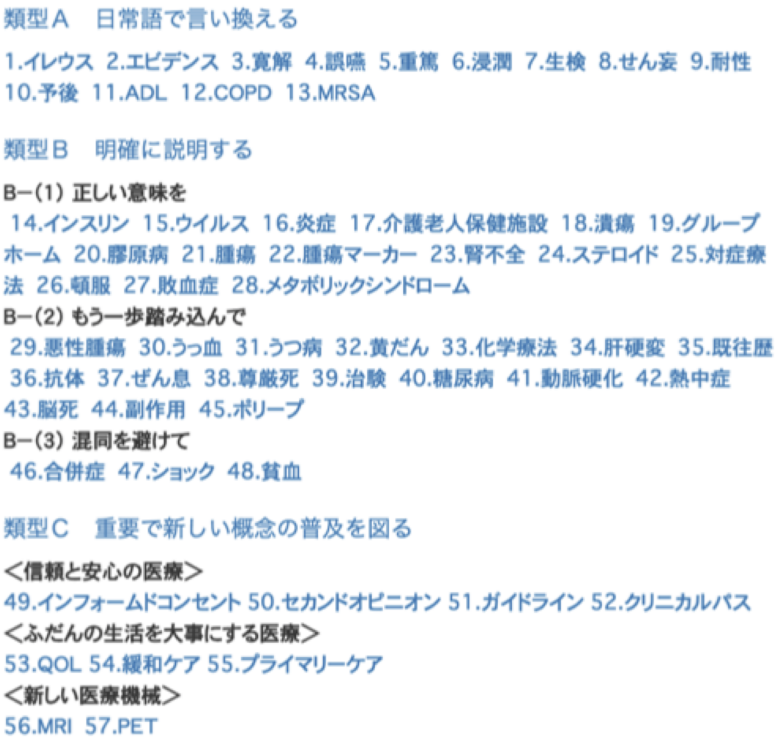


図 2.2　各医療用語の類型化例
（http://www2.ninjal.ac.jp/byoin/より抜粋）

2.2　日本語の難易度推定に関する関連研究

日本語の難易度推定に関する研究について述べる．2.2.1 節で難易度推定に関する研究の概観を述べ，2.2.2 節で本研究の基本となるサポートベクターマシン（SVM）を用いた日本語の難易度推定に関する先行研究の詳細を述べる．

2.2.1 難易度推定に関する研究の概観

英語においては，語や文の平易化タスクという形で，難易度推定に関する研究が行われてきた．その理由として，英語には言語資源が豊富にあるということが挙げられるといわれている [1]．言語資源の例としては，平易に書かれた大規模コーパスである Simple English Wikipedia [2]，難解な文と平易な文のパラレルコーパス [3, 4, 5]，大規模な言い換え辞書である PPDB (The Paraphrase Database) [6, 7] を元に，難解な語句から平易な語句への言い換え辞書である Simple PPDB [8] などが挙げられる．

その一方で，日本語の難易度推定に関する研究はあまり行われていない．これは，英語のような言語資源がないことが挙げられる．一般的な語彙の難易度は概念的な難易度のことを指すが，テキストの可読性を表すリーダビリティの観点からの難易度もあり，いくつかの評価軸に対して定義できると考えられる．リーダビリティの標準的な評価に関しても，英語には Flesch Reading Ease (FRE)[†]など確立された評価指標が存在するのに対し，日本語には「帯」[9] や「jReadability」[‡]といったシステムが開発されているが，確立はされていない．ここでは本研究での難易度の意味に近い概念的な難易度に関する研究について述べる．

難易度について最もよく使用される指標は出現頻度である．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り [10]，それを利用した語彙の平易化を行った研究もある [11, 12]．日本語の難易度推定の研究では，日本語教育語彙表[§]という外国人の語彙教育を目的に 17,920 語の難易度を 6 段階で表したリストがデータとして使われている．梶原らは日本語教育語彙表の難易度を元に日本語単語の難易度推定器を作成し，簡易な語への変換を行うパラフレーズ辞書（Simple PPDB: Japanese）を構築した [1]．素性には， Wikipedia 本文での頻度情報に加えて，リーダビリティの先行研究で有効な素性であるとされた単語長や文字種，Wikipedia 本文をもとにした word2vec[13] による単語分散表現を用いた．これらを用いて SVM で日本語教育語彙表の難易度を推定するタスクを行い，単語分散表現を用いることで難易度推定精度が大幅に向上したという結果を報告した．

[†]Flesch, R. (1948). FLESCH READING EASE READABILITY FORMULA
[‡]http://jreadability.net
[§]http://jhlee.sakura.ne.jp/JEV.html

しかし，word2vec による単語分散表現は特性上不安定であり，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．高田らはこの word2vec の不安定さを実験で示し，単語分散表現を用いる代わりに現代日本語書き言葉均衡コーパス（BCCWJ ;The Balanced Corpus of Contemporary Written Japanese） [14] 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ている [15].

2.2.2 SVM を用いた難易度推定手法

本研究は，日本語教育語彙表を対象とした日本語難易度推定を行った高田らの先行研究 [15] に基づいてベースライン手法を設定するため，その詳細について述べる．以下で，モデルとして用いた SVM ，彼らが使用した語彙リソースである日本語教育語彙表，使用したモデルと素性，その結果について述べる．

SVM

高田らの研究では，難易度推定のモデルとして多クラス分類モデルが使用された．多クラス分類モデルは，目的変数が各サンプルの所属する群を表現するラベルである離散値であり，その目的変数を予測するために使用される．高田らの研究では，目的変数が日本語教育語彙表に掲載されている難易度であるため，多クラス分類モデルとして SVM による分類（以下 SVC と称する）を用いた．本章では， SVM について説明する．

SVM のモデルには，クラス分類問題に使われる SVC と，回帰問題使われる 回帰（以下 SVR と称する）の 2 種類が存在する．

まず，SVC について説明する．SVM は，「マージン最大化」を基本アイデアとし，SVC ではこのアイデアの下で，訓練データの集合からクラスを分類する境界を学習する．ここで登場するマージンとは図 2.3 に示すように，分類する境界を挟んだ両クラスがどれだけ離れているかを表す度合いであり，SVC ではこれを最大化することが目的となる．SVC では，マージン以上の領域に属するデータの損失は 0 とみなされる．

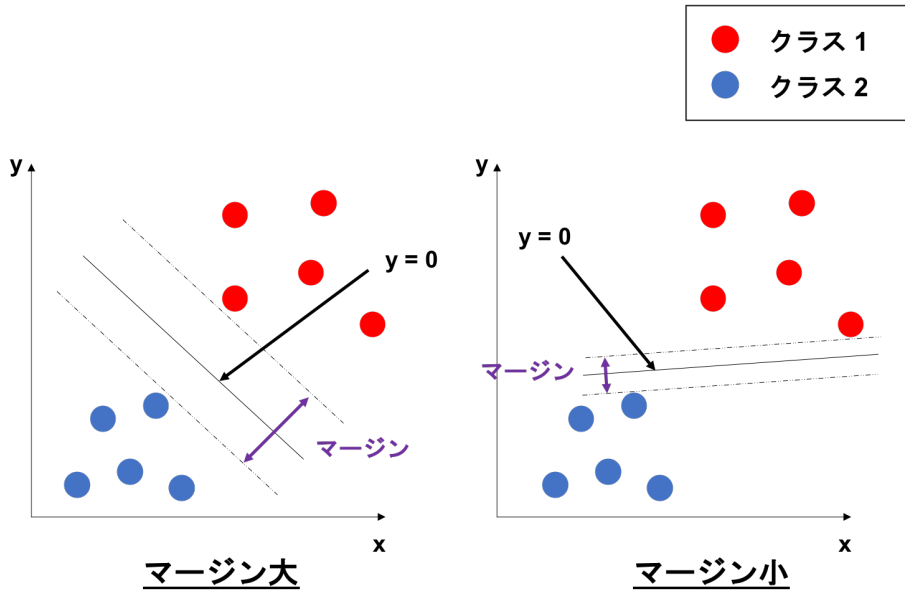


図 2.3　SVM の分類境界とマージンの例

SVR は SVC と同様にマージン最大化の考え方の下で回帰を行っている．単純な二乗誤差を考慮する通常の線形回帰と異なり，回帰直線から残差の絶対値が ε 以下の場合は残差によるペナルティが 0 になるようになっている．入力する特徴量ベクトル $\mathbf{x_i}$ に対して，正解の値が y_i ，回帰式が式（2.2.1）でそれぞれ表される時に，その効果を与えるのが式（2.2.2）で表されるスラック関数 ξ_i であり，式（2.2.3）の損失関数を最小化するように重み $\mathbf{w}$，$\mathbf{w_0}$ 学習が進められる．式（2.2.3）の C はコストパラメータと呼ばれ，この値が大きいほど誤分類を許容しないように，小さいほど許容するように学習が進む．

$$f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + \mathbf{w_0} \tag{2.2.1}$$

$$\xi_i = max(0, |y_i - \mathbf{w}^T\mathbf{x}| - \varepsilon) \tag{2.2.2}$$

$$L = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i \xi_i \tag{2.2.3}$$

以上に加えて今回はカーネル法を用いた非線形な分類と回帰を行うためそれについても説明する．カーネル法は，データをそのまま直線や平面で分類できない時に特徴量ベクトルを高次元空間に写像し，その空間で最適化計算を行うものである．

そのために必要なカーネル関数としては，式（2.2.4）で定義される多項式カーネルや式（2.2.5）で定義される動径基底関数カーネル（RBF カーネル）が代表的である．

$$K(\mathbf{x_i}, \mathbf{x_j}) = (\alpha \mathbf{x_i} + \mathbf{x_j})^p \tag{2.2.4}$$
$$K(\mathbf{x_i}, \mathbf{x_j}) = exp(-\gamma \|\mathbf{x_i} - \mathbf{x_j}\|^2) \tag{2.2.5}$$

難易度推定対象

難易度推定対象の日本語教育語彙表は，17,920 語の日本語教育用語彙の読み，難易度，品詞などが収録されたノンネイティヴの日本語教育を目的に作られたものである（表 2.1）．難易度は 6 段階（初級前半，初級後半，中級前半，中級後半，上級前半，上級後半）に分けられており，3 段階（初級・中級・上級のそれぞれの前後半をまとめたもの）としても利用可能である．高田らの研究では，これらの語彙とそれぞれの難易度をデータとした．難易度には上記の 3 段階と 6 段階の両方が用いられた．

表 2.1　日本語教育語彙表の各項目の説明と例

項目名	説明	例
No.	項目の ID 番号.	102
標準的な表記	新常用漢字表に従った標準的な表記．漢字かな交じり表記	赤ちゃん
読み	カタカナ表記での読みの情報.	アカチャン
語彙の難易度	日本語教育上の語彙レベル.	2. 初級後半
品詞	UniDic に基づいた品詞（品詞 2（詳細））と簡易版の品詞（品詞 1）．複合語の場合は and で全品詞を登録.	名詞-普通名詞-一般
語種	UniDic に基づいた語種.	和語

モデルと素性

難易度推定モデルは SVC を用い，カーネルには RBF カーネルを用いた．難易度推定に使用した素性は，各語のコーパスでの頻度情報と表層的な情報である．以下に詳細を示す．

A. **Wiki 頻度（1 次元）**：MeCab[16] と MeCab に追加する単語分かち書き辞書の mecab-ipadic-NEologd[¶]によって日本語 Wikipedia を単語分割した際の各単語の対数頻度．

B. **文字単位の Wiki 頻度（4 次元）**：単語に含まれる各文字単位の日本語 Wikipedia 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．

C. **単語長（1 次元）**：単語の文字列としての長さ．

D. **文字種（3 次元）**：文字種（ひらがな，カタカナ，漢字）に関する素性．

   文字種の有無　単語に含まれる文字種の有無（1, 0）
   文字種の文字数　単語に含まれる文字種の文字数

E. **BCCWJ 頻度（14 次元）**：BCCWJ 中の単語単位の対数頻度．BCCWJ に含まれる 13 種類のジャンルごとの頻度と，全体における頻度の対数．これらの情報をまとめた BCCWJ 語彙表から抽出．

なお，頻度が 0 のものは，対数頻度を 0 ($\log 0 = 0$) とした．頻度情報は，SemEval-2012 の English Lexical Simplification タスクの Baseline で優れたスコアを示したこと， Wikipedia 中の単語単位の頻度を素性として利用している研究がある [8, 1] ことなどから，有用な素性になると考えられる．単語長は，Flesch ら [17] が用いていることなどから，リーダビリティを測る有用な素性になると考えられる．

[¶]https://github.com/neologd/mecab-ipadic-neologd

文字種は，簡易な語にはひらがなやカタカナが多く，漢字が少ないという仮定の下，使われている．高田らは，これらの素性の組み合わせを変えて，難易度推定実験を行った．

結果と考察

難易度推定実験として，3 段階難易度の予測と 6 段階難易度の予測の 2 つのタスクが行われた．その結果，Wiki 頻度，文字単位の Wiki 頻度，単語長，文字種の文字数，BCCWJ 頻度を用いたモデルが両タスクにおいて最高の精度を出した．特に文字単位の Wiki 頻度と BCCWJ 頻度は，素性から除くと難易度推定精度が大きく低下することから，難易度推定のための重要な素性であることが示唆された．

我々の考察を加えると，文字単位の Wiki 頻度に関しては，平易な単語に用いられやすいひらがなやカタカナは頻度が大きくなり，難解語に用いられると考えられる漢字の頻度は低くなると考えられ，日本語の難易度推定に有用であることが考えられる．また，BCCWJ 頻度に関しては，難易度推定の対象となった日本語教育語彙表そのものが BCCWJ などを元に作られたものであり，そこでの頻度が難易度に関連している可能性が考えられる．

2.3 まとめ

本研究は，国立国語研究所の提案よりも大規模な医療用語の難易度リストを提示するために，手作業で大量の医療用語の難易度を調査することは困難なため，機械学習を用いて医療用語の難易度を推定することを目的としている．しかし，医療用語の難易度を推定する手法は確立されていないため，既存の一般的な日本語の難易度推定手法に基づいて，医療用語の難易度推定を試みる．そのため，日本語の難易度推定に関する関連研究を紹介した．

日本語の難易度推定において，単語分散表現を素性として用いると難易度推定精度が向上するという報告がある [1] が，単語分散表現は不安定な性質があり，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．高田らは，この性質を実験で示し，BCCWJ

での頻度情報などを用いて，難易度推定精度は単語分散表現を用いた場合と比較して劣るものの，再現性の高い日本語の難易度推定手法を提案した [15].

再現性の高さが本研究では重要であると考え，本研究では高田ら [15] の手法に基づいて医療用語の難易度推定を試みる．ただし，いずれの手法も外国人教育を目的とした日本語難易度リストを対象に行っているため，用いられた素性が医療用語の難易度推定においても精度向上に寄与するとは限らないため，この手法が医療用語の難易度推定に適用可能かを検討し，改善点を模索する．

第 3 章　一般語の難易度推定実験によるベースライン手法の検討

本研究では，まず一般語の難易度推定実験を行い，その結果を元に医療用語難易度推定のベースライン手法を設定するため，2 章では，医療用語の難易度や日本語の難易度推定に関する先行研究を紹介し，ベースライン手法の基本となる日本語の難易度推定を行った研究の詳細について述べた．本章では，難易度推定に有用であると思われる web コーパス頻度に関する 2 種類の素性を新たに追加し，先行研究と同様に日本語教育語彙表を対象に難易度推定実験を行い，ベースライン手法を検討する．3.1 節で方法を述べ，3.2 節で結果を述べる．

3.1　方法

本節では実験に用いるモデルと素性，対象とするデータ，評価方法について述べる．

3.1.1　モデルと素性

モデルは先行研究と同様に SVC を使用し，カーネルはそれぞれのタスクに対して，線形カーネルと RBF カーネルの 2 種類を用いる．用いる素性は，2.2.2 節で紹介した先行研究の素性に，Twitter 頻度と文字単位の Twitter 頻度を追加した．以降で用いる Twitter のツイートデータは 2012 年の一年間のツイート約 1,580 万件を，Wikipedia の記事は Wikipedia 日本語版のダンプサイト*から 2018 年 6 月時点で最新のものをそれぞれ用いた．

a. **Twitter 頻度（1 次元）** ：MeCab[16] と MeCab に追加する単語分かち書き辞書の mecab-ipadic-NEologd†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．

*https://dumps.wikimedia.org/jawiki
†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.

Twitter 頻度は Wiki 頻度と同様で，Wiki 頻度の対象とするコーパスを Wikipedia の記事から Twitter のツイートデータに変更した素性である．文字単位の Twitter 頻度に関しても同様で，文字単位の Wiki 頻度の対象とするコーパスを Wikipedia の記事から Twitter のツイートデータに変更した素性である．これらを追加した理由は．非医療者が主なユーザであると考えられる Twitter における頻度情報が，一般の日本語のみならず，医療用語の難易度推定にも有用であると考えたためである．素性の組み合わせは，全素性や関連研究のと同じ条件のものなど，数パターンを用意した.

3.1.2 対象データ

関連研究と同様に日本語教育語彙表の難易度を対象とした．ただし今回は，同一表記で品詞が異なる語は頻度情報の区別ができないため，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（以下実験ボキャブラリと称する）を対象とした（表 3.1）.

表 3.1 実験ボキャブラリの難易度の度数分布表

難易度	1	2	3	4	5	6
語数（語）	178	415	1,399	4,576	4,768	1,072

3.1.3 評価方法

日本語教育語彙表の 3 段階難易度の推定と 6 段階難易度の推定の二種類のタスクを行う．対象のデータをトレーニングデータ（80%）とテストデータ（20%）に分け，トレーニングデータで 5 分割のグリッドサーチ交差検証を行い，そこでの平均

スコアが最も高かったパラメータを選び，テストデータでのスコアを実験でのスコアとする．スコアはミクロ平均の accuracy を採用する．accuracy は以下の数式で表される．

$$accuracy = \frac{\sum_i H(i)}{n} \tag{3.1.1}$$

$$H(i) = \begin{cases} 1 & (y_{true,i} = y_{pred,i}) \\ 0 & (otherwise) \end{cases} \tag{3.1.2}$$

accuracy は，正解数をサンプル数で割った値である正答率であるため，この値が大きいほど性能が高いことを示す．

3.2 結果と考察

線形カーネルを用いた結果を表 3.2 に，RBF カーネルを用いた結果を表 3.3 にそれぞれ示す．表内の数字は accuracy の値を示す．素性として 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 頻度を用いたモデルを基本素性と呼ぶ．表内の w/o は後続の素性を基本素性から除いたことを示す．この二つのタスクにおいて両カーネルの結果に大きな差はないが，RBF カーネルを用いた方がおおよそ全ての素性で高い精度が出た．素性に関しては，両カーネルにおいて BCCWJ 頻度を除くと，3 段階難易度推定タスクにおいては 3 ポイント以上，6 段階難易度推定タスクにおいては 2.5 ポイント以上，それぞれ下がることから，関連研究と同様に BCCWJ 頻度の重要性が確認された．また，RBF カーネルにおいては，関連研究手法に Twitter 頻度を素性として追加することで両タスクにおいて精度が向上し，Twitter 頻度も難易度推定に有用であることが示唆された．

表 3.2　予備実験の結果（線形カーネル）

使用する素性	3 段階	6 段階
+ 文字単位の Twitter 頻度	0.701	0.542
基本素性	0.693	0.520
w/o Twitter 頻度 （関連研究手法）	0.702	0.535
w/o BCCWJ 頻度	0.659	0.494
w/o 文字単位の Wiki 頻度	0.698	0.524
w/o 文字種の文字数	0.695	0.518
w/o 単語長	0.709	0.540
w/o Wiki 頻度	0.689	0.510

表 3.3　予備実験の結果（RBF カーネル）

使用する素性	3 段階	6 段階
+ 文字単位の Twitter 頻度	0.711	0.554
基本素性	0.719	0.553
w/o Twitter 頻度（関連研究手法）	0.715	0.535
w/o BCCWJ 頻度	0.669	0.508
w/o 文字単位の Wiki 頻度	0.701	0.540
w/o 文字種の文字数	0.701	0.531
w/o 単語長	0.709	0.538
w/o Wiki 頻度	0.699	0.535

第 4 章　ベースライン手法による医療用語難易度推定

3 章では，日本語教育語彙表を対象に予備実験を行い，医療用語の難易度推定のベースライン手法を決定した．本章では，ベースライン手法を用いて医療用語難易度推定を行い，その結果を評価をする．以下で，難易度推定方法とその評価実験について述べる．

4.1　難易度推定方法

ベースライン手法を用いた医療用語の難易度推定方法について述べる．以下で，使用するモデル，素性について述べる．

モデル

モデルには SVC を使用し，予備実験の結果を考慮し，カーネルは非線形カーネルである RBF カーネルを用いる．モデルは全て sklearn (0.18.1) の SVM *を用いる．また，多クラス分類への拡張は one-versus-rest 法で行う．

素性

予備実験において 3 段階と 6 段階の両タスクで先行研究の素性を上回った基本素性を，本章で医療用語難易度推定に用いる素性とする．

4.2　評価実験

4.1 節の方法で医療用語の難易度を推定した結果の評価実験を行う．以下で実験の方法と結果および考察を述べる．

*https://scikit-learn.org/0.18/modules/classes.html#module-sklearn.svm

4.2.1 実験方法

本節では，前述した方法での医療用語難易度推定結果の評価方法について述べる．大人数に対して人手での主観評価を行うため，非医療者の募集をしやすいクラウドソーシングを用いて評価を行った．評価対象の語彙とクラウドソーシングのタスクに関する詳細，学習データを以下に示す．

評価対象語彙

この実験では，医療用語である基準を万病辞書[†]に掲載されている語とした．万病辞書とは，電子カルテや退院サマリといった医療文書から症状や病名に関連する語を幅広く抽出したデータベースであり，58 万語近くを収録している（2018 年 6 月時点）．3 名の医療従事者がそれぞれの医療用語に ICD10 コード を付与し，その一致度合いに応じて信頼度が付与されている．万病辞書の構成を表 4.1 に示す．本研究で対象とするのは，この万病辞書の「出現形」とする．

今回の評価は人手を利用するため，全ての語を評価対象にすることはできない．評価対象の語を限定するために万病辞書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 語を対象とした．

表 4.1　万病辞書の各項目の説明と例

項目名	説明	例
出現形	電子カルテや退院サマリから抽出された症状・病名	転移性新生物
ICD10 コード	出現形に対応する，ICD10 対応標準病名マスター‡に記載されている ICD10 コード．ただし，次の場合には -1 を付与： a. 4 つ以上のコードが存在する場合（3 つまでは全て付与） b. 断片的な情報のみで判断が困難な場合 c. コードが存在していない場合	C80
標準病名	出現形に対応する，ICD10 対応標準病名マスターに記載されている ICD10 対応標準病名	悪性奇形腫
MedDRA コード	ICH 国際医薬用語集（Medical Dictionary for Regulatory Activities; MedDRA）の日本語版 MedDRA/J の基本語（Preferred Term; PT）	転移性新生物
信頼度 LEVEL	それぞれ以下の基準でつけられている． S: ICD10 対応標準病名マスターに記載されている症状・病名 A: 2 名以上の医療従事者が同じコードを付与した症状・病名 B: 2 名以上の医療従事者が相談してコードを付与した症状・病名 C: 1 名の医療従事者がコードを付与した症状・病名 D: 計算機が自動的に割り当てた症状・病名	C

タスク設定・難易度決定

Yahoo!クラウドソーシング[§]を用いて，医療関係の業務に携わったことのない 100 人をリクルートした．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」という質問をし，前述の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 件法で回答してもらう．タスク概要の画面を図 4.1 に，タスクの画面の例を図 4.2 に示す．表 4.2 に示すように，それぞれの回答に対して点数を割り当て，各語の難易度を 100 人の平均点として定め，これを以降 crowd difficulty level (CDL) と呼ぶことにする．また，タスクの参加者には，回答の送信をもって以下の内容に同意したものとみなすことを伝えている．

- 回答内容は研究利用の範囲を超えて使用することはないということ
- 個人の特定などはされないということ

本タスクは 2018 年 9 月 5 日に実施した．ここで得られた各語の CDL と，6 段階難易度推定のタスクで学習した分類器を用いて難易度を推定した結果を比較する．

学習データ

分類器を構築するための学習データは，日本語教育語彙表に掲載されている実験ボキャブラリの語およびそれぞれに対応する 6 段階難易度難易度とする．

[‡]http://www2.medis.or.jp/stdcd/byomei/index.html
[§]https://crowdsourcing.yahoo.co.jp/
[§]http://sociocom.jp/ |data/2018-manbyo/index.html

表 4.2　クラウドソーシングでの回答結果と点数の対応

回答	点数
難しいと思う	4
やや難しいと思う	3
やや簡単だと思う	2
簡単だと思う	1

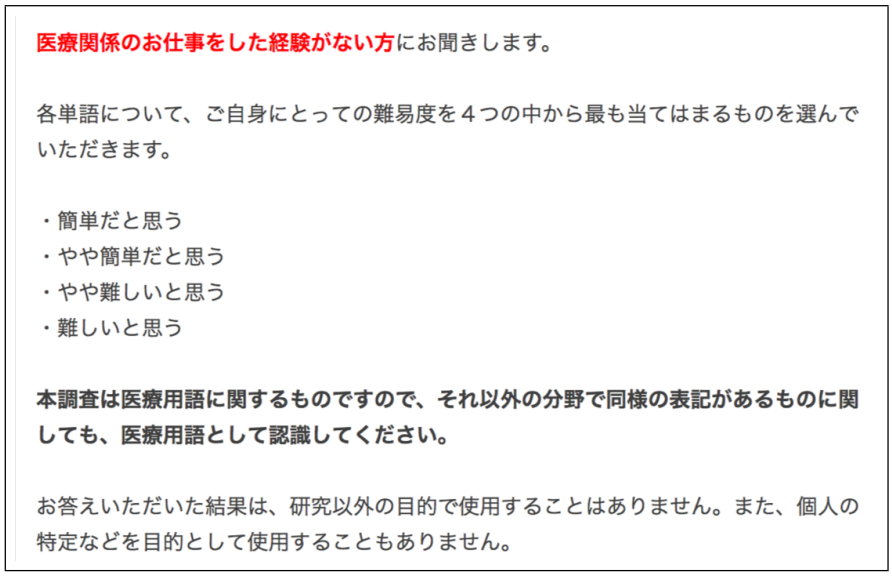


図 4.1　クラウドソーシングのタスク概要の画面

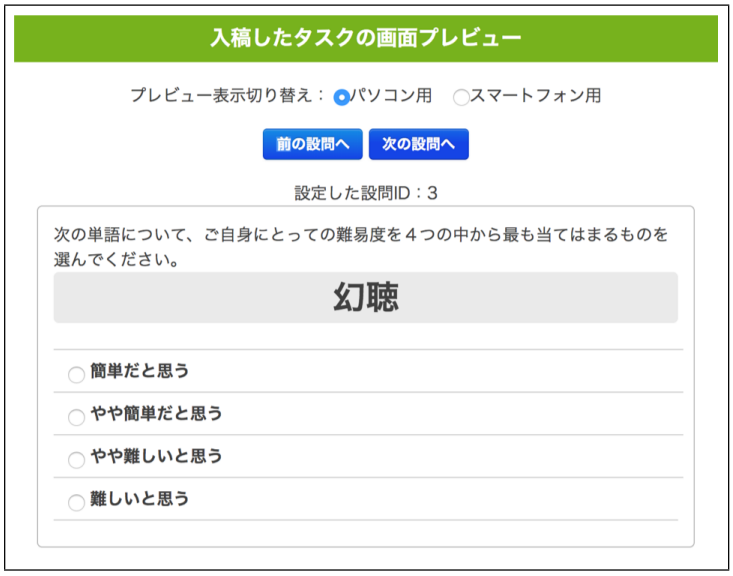


図 4.2　クラウドソーシングのタスク画面の例

4.2.2　実験結果と考察

評価結果を図 4.3 に示す．推定難易度毎に箱ひげ図にしており，縦軸が CDL，横軸が推定難易度である．推定難易度 1 から 3 に含まれている語は表 4.3 に示す 10 語しかなかった．推定難易度 4 から 6 に関しては，推定難易度の中央値で見ると CDL と正の相関となった．

推定難易度 1 から 3 の語については，今回の訓練対象である日本語教育語彙表の難易度 1 から 3 は日常用語の基本レベルであるため，医療用語が少数しか分類されないのは妥当と考えられる．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という点も妥当と考えられる．しかし，「よう」「IC」の二語は例外と考えられる．医療用語としては，「よう」は皮膚などにできる腫れ物，「IC」はう蝕という歯の病気をそれぞれ表現する語である．これらは CDL が大きい，すなわち非医療者にとっての難易度が高く，また知名度も低いと思われるにもかかわらず，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．これらについては，「よう」は助動詞の「よう」，「IC」は集積回路を表す「IC」の意味も語であるため，同音異義語の頻

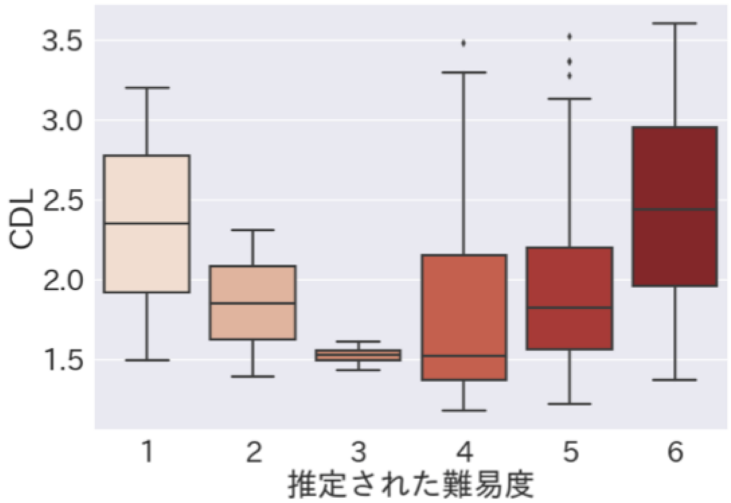


図 4.3　難易度推定結果

表 4.3　推定難易度が 3 以下の語

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

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 から 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である．しかしながら個々の事例を見ると，身近な語だと考えられる「耳垢」や「歯垢」が推定難易度 6 となっているなど課題も存在する．

以上のような長所と短所が挙げられるが，ノンネイティヴの日本語教育を目的に作られた日本語教育語彙表の難易度を学習データとしていることによるものであるのか，使用した素性によるものなのか，そういった点について検討が必要であると思われる．医療用語難易度の教師データを作り，それらを用いて医療用語の難易度推定に適した素性の調査を行う．

第 5 章　医療用語が難解に感じる要因の調査と新規素性の検討

4 章ではベースライン手法を用いた医療用語難易度推定を行った．その結果，頻度情報などによる用いる素性を改良することで，難易度推定精度が向上する可能性が示唆された．本章では，新たに追加する素性を検討するために，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，医療用語難易度推定に適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．

難易度は主観的な尺度であり，定量的な評価が困難である．日本語教育語彙表における難易度は，BCCWJ 頻度や文字種の文字数などを用いることで上手く表現することができた．しかし，医療用語の難易度は日本語教育語彙表のそれとは性質が異なる可能性があり，どういった要因が難易度と関係しているのかを推測するのが難しいという問題がある．

そこで本章では，医療用語の難易度と関連する要因をクラウドソーシングを用いて調査し，結果を難易度推定のための素性として数値化し，それらの素性を加えて難易度推定精度を向上させる．また，4 章で作成した医療用語難易度の教師データが 260 語と機械学習に用いるデータとしては小規模であるため，1000 語分の教師データを作成する．以下で，医療用語の難易度と関連する要因の調査，調査結果を反映した医療用語難易度の推定方法，評価実験について述べる．

5.1　医療用語の難易度と関連する要因の調査

難易度と関連する要因の調査をクラウドソーシングを用いて行う．最初にクラウドソーシングを用いた仮説収集についての関連研究を紹介し，その後，調査方法と結果について述べる．

5.1.1　クラウドソーシングを用いた仮説収集に関する関連研究

主に医療系の研究において，クラウドソーシングを用いた仮説収集を行った研究がいくつか報告されている．Bevelander ら [18] は，成人の肥満に関する仮説をク

ラウドソーシングで収集し，新しい有用な仮説が得られることを示した．米良ら [19] は，仮説収集のみならず，その仮説も尤もらしさの検証も自動で行える手法を提案した．同様に荒牧ら [20] は，アレルギーの原因となるリスク因子の仮説をクラウドソーシングで収集及び検証をし，その有用性を示した．

これらの研究が示すように，未知の仮説を収集するためにクラウドソーシングを用いることは有用な方法であると考えられる．これに従って，医療用語が難解に感じる要因に関してクラウドソーシングを用いて調査および検討をする．

5.1.2 調査方法

医療用語が難解に感じる要因の調査および検討の方法について述べる．調査は，Yahoo!クラウドソーシングを用いて 500 人のクラウドワーカーに対して行う．タスクの画面を 5.1 に示す．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

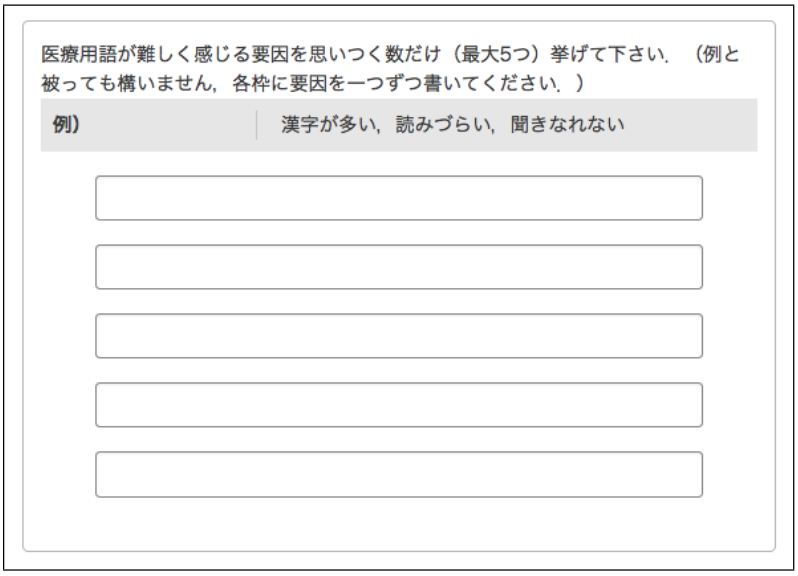


図 5.1　タスク画面

また，タスクの参加者には，回答の送信をもって以下の内容に同意したものとみなすことを伝えている．

- 回答内容は研究利用の範囲を超えて使用することはないということ
- 個人の特定などはされないということ

このタスクの回答結果を人手で整理し，素性化できるかなどを検討する．

5.1.3　調査結果

前述のタスクを 2018 年 11 月 20 日から 11 月 21 日に実施した．調査結果を表 5.1 に示す．およそ回答内容は表中の「回答内容」のいずれかに分類でき，さらに，既存の素性で表現できる特徴，新たに素性を作成することで表現できる特徴，素性として表現できない特徴に分類した．「漢字が多い」「アルファベットなどの略称」「カタカナの意味がわからない」は文字種の文字数や文字種の有無で，「聞きなれな

い言葉の羅列」は各種頻度情報で，「長い」は単語長で，「難しい漢字が使われる」は文字単位の頻度情報によって，それぞれ既存の素性で表現できる特徴であると考えられる．また，「複合的な言葉」は各語を形態素解析器によって構成素に分け，構成素数をカウントすることで，「漢字が連なっている場合に区切り位置がわからない」は漢字の連続数をカウントすることで，それぞれ新たな素性で特徴を表現できると考えられる．「専門用語が入っていると分からなくなる」「怖さを感じる」「部位の表現が難しい」「似た用語が多い」「聞き取れない」「字面だけでは意味を想像できない」などは素性として表現できない特徴であると考えられる．

表 5.1　回答内容の分類

分類	回答内容
既存の素性で 表現できる特徴	漢字が多い，アルファベットなどの略称， カタカナの意味がわからない， 聞きなれない言葉の羅列，長い， 難しい漢字が使われる
新たに素性を作成することで 表現できる特徴	複合的な言葉， 漢字が連なっている場合に 区切り位置がわからない
素性として 表現できない特徴	専門用語が入っていると分からなくなる， 怖さを感じる，部位の表現が難しい， 似た用語が多い， 聞き取れない， 字面だけで意味を想像できない，

5.2　調査結果を考慮した語の難易度推定方法

5.1 節の調査結果から素性として表現できる回答を素性として加えて，医療用語および一般の日本語の難易度推定方法について述べる．以下で用いるモデルと素性について述べる．

モデル

本研究では，難易度の教師データが連続値である医療用語の難易度推定には SVR を，難易度の教師データが離散値である一般語の難易度推定には SVC をそれぞれ使用した．また，カーネルには非線形カーネルである RBF カーネルを用いた．モデルは全て sklearn (0.18.1) の SVM を用いた．また，多クラス分類への拡張は one-versus-rest 法で行った．

素性

難易度推定に使用する素性について説明する．

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*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）**：単語に含まれる各文字単位の日本語 Wikipedia および 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．

*http://sociocom.jp/ data/2018-manbyo/data/MANBYO_201806_Dic-utf8.dic

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

I. **単語長（1 次元）**：単語の文字列としての長さ

J. **BCCWJ 頻度（14 次元）**：2 章の BCCWJ 頻度と同様

K. **頻度情報（2 次元）**：MeCab と mecab-ipadic-NEologd によって日本語 Wikipedia および Twitter を単語分割した際の各単語のそれぞれの対数頻度

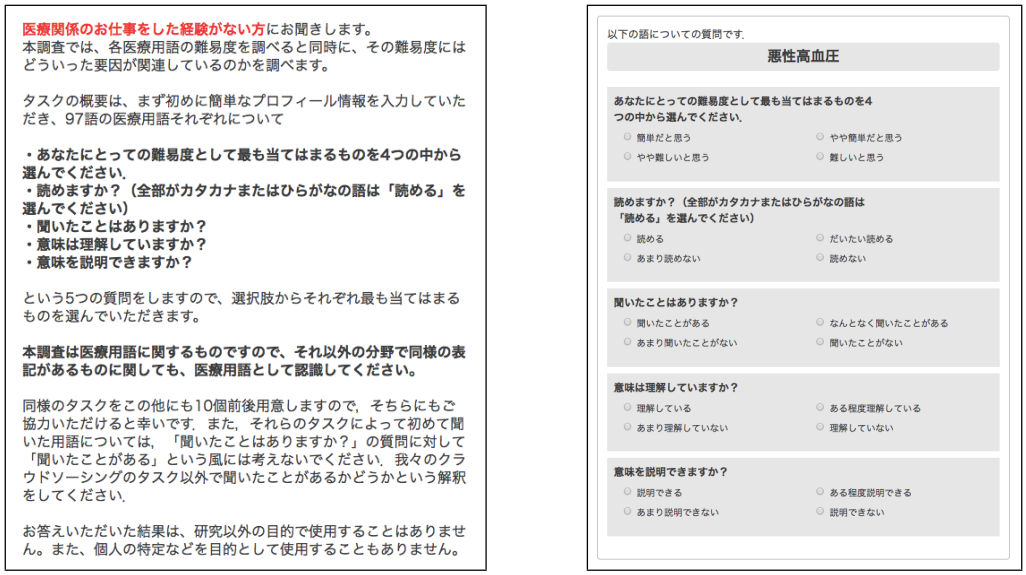
以降で使用する「全提案素性」は上記の全ての素性を用いた 60 次元の素性，「基本素性」は 3.2 節で定義した 24 次元のベースライン素性を表す．

5.3 評価実験

節の方法で医療用語の難易度を推定した結果の評価実験を行う．しかし，実験を行うために十分な量の医療用語難易度の教師データがないため，まずデータの作成をする．その後，それらを用いて評価実験を行う．以下で，データの作成方法，実験の方法，結果および考察を述べる．

5.3.1 医療用語難易度の教師データの作成

医療用語難易度の教師データの作成については，Yahoo!クラウドソーシングを用いて 1,000 語を対象に行った．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のないクラウドワーカーのみをリクルートした．タスクの画面の例を図 5.2 に示す．



タスク概要の画面　　各語に対する回答画面例

図 5.2　クラウドソーシングのタスク画面の例

各語に対する質問は以下の①から⑤の 5 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？

難易度そのものを問うために質問①を用意し，その他の要因との関連も見るために②から⑤の質問を用意した．それぞれに対する回答は 4 件法で行う．選択肢はそれぞれ以下の通りである．

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

これを各語に対して 500 人を対象に行う．表 5.2 に示すように，それぞれの質問に対するそれぞれの回答に点数を割り当て，各語に対して質問別に点数を算出す

る．また，タスクの参加者には，回答の送信をもって以下の内容に同意したものとみなすことを伝えている．

- 回答内容は研究利用の範囲を超えて使用することはないということ
- 個人の特定などはされないということ

本タスクは 2018 年 11 月 29 日から 12 月 3 日に実施した．全語の点数の平均と分散を表 5.3 に示す．可読性についての質問②の平均点が他より低くなった．また，質問同士の相関係数を表 5.4 に示す．この表からも，可読性以外は質問①と強い相関を示していることがわかるため，質問①，③，④，⑤に関しては同様の内容を示していると考え，これ以降，各語の難易度は質問①の 500 人の平均点とする．

表 5.2　クラウドソーシングでの回答結果と点数の対応

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

表 5.3　全語の質問毎の平均と分散

	①	②	③	④	⑤
平均	2.99	1.68	3.03	3.15	3.23
分散	1.12	1.05	1.17	1.08	1.05

表 5.4　質問同士の相関係数

	①	②	③	④	⑤
①	1.00	0.361	0.692	0.746	0.733
②	0.361	1.00	0.358	0.341	0.313
③	0.692	0.358	1.00	0.833	0.798
④	0.746	0.341	0.833	1.00	0.929
⑤	0.733	0.313	0.798	0.929	1.00

評価者間の一致度を検討するために，Fleiss の κ 係数 [21] を用いる．Fleiss の κ 係数は式（5.3.1）で表される．

$$\begin{aligned} \kappa &= \frac{P_o - P_e}{1 - P_e} \\ P_o &= \frac{\sum_i \sum_j {X_{ij}}^2 - nm}{nm(m-1)} \\ P_e &= \sum_j {P_j}^2 \\ p_j &= \frac{1}{nm} \sum_i X_{ij} \end{aligned} \tag{5.3.1}$$

ここで m はワーカの人数，n はサンプル数，X_{ij} は，カテゴリ j に分類されたサンプル i の評価数であり，今回は $m = 500, n = 1000$ である．計算の結果，κ 係数は 0.0983 であり，Z 統計量が 1,799 となったため，標準正規分布表と照合すると $\mathrm{p} < 0.01$ となるため，評価者間で一定の合意があると考えられる．

5.3.2　実験方法

本節では，5.2 節の方法で医療用語と一般の日本語の難易度推定を行った結果の評価方法について述べる．以下で対象データ，評価指標，試行回数について述べる．

対象データ

難易度推定する対象は医療用語と一般語の二種類で，医療用語は前節で作成

した教師データ，一般語は日本語教育語彙表とする．日本語教育語彙表の難易度推定は 3 段階と 6 段階の両方を対象とした．

評価指標

医療用語の難易度推定実験は，連続値である難易度を推定する回帰問題であるため，評価指標として，R^2，neg_MAE，neg_MSE の三種類を用いた．それぞれの指標は以下の式で表される．

$$R^2 = 1 - \frac{\sum_i (y_{true,i} - y_{pred,i})^2}{\sum_i (y_{true,i} - \overline{y})} \tag{5.3.2}$$

$$neg\_MAE = -\frac{\sum_i |y_{true,i} - y_{pred,i}|}{n} \tag{5.3.3}$$

$$neg\_MSE = -\frac{\sum_i |y_{true,i} - y_{pred,i}|^2}{n} \tag{5.3.4}$$

ここで，$y_{true,i}, y_{pred,i}$ を語 i の正解難易度と予測難易度の対として，$y_{true,i}$ は各サンプルの正解難易度，$y_{pred,i}$ は各サンプルの予測難易度，$\overline{y}$ は正解難易度の平均値，n はサンプル数をそれぞれ表す．三つの各指標は，値が大きければ大きいほど性能が高いことを示す．

一般語の難易度推定実験の評価指標としては，3.1.3 節と同様にミクロ平均の accuracy を用いた．accuracy は，正解数をサンプル数で割った値である正答率であるため，大きければ大きいほど性能が高いことを示す．3 段階難易度と 6 段階難易度の両方を対象に実験したため，それぞれで accuracy を算出する．

評価内容

「全提案素性」から素性を一種類ずつ除いて，10 分割の交差検証を行い，各指標の値を 10 回の平均とし，比較する．また，モデルのパラメータの組み合わせは，平均スコアが最も高い組み合わせを選択する．その際のスコアは，

医療用語の難易度推定においては決定係数（R^2 スコア）を，一般語の難易度推定においてはミクロ平均の accuracy をそれぞれ適用する．

5.3.3 実験結果

医療用語と一般語の難易度を推定した実験結果を以下に示す．

医療用語難易度推定結果

医療用語の難易度を推定した結果を表 5.5 に示す．表中の各指標の値は 10 回の試行の平均値を示す．結果は R^2 の値が高い順に並べている．また，表内の w/o は後続の素性を「全提案素性」から除いたことを示す．ある素性を除いて難易度推定をした結果，精度が下がるということは，除いた素性が精度向上に寄与したと考えることができる．したがって，R^2 を基準とすると，「構成素頻度情報」が最も精度向上に寄与した素性と捉えることができる．

一般語難易度推定結果

一般語の難易度を推定した結果を表 5.6 に示す．表中の値は 10 回の試行の平均値を示す．結果は 3 段階難易度推定タスクでの accuracy の値が高い順に並べている．また，表 5.5 と同様に，表の下に出現する素性が精度向上に寄与した素性と考えることができる．したがって，3 段階難易度と 6 段階難易度の両方で，BCCWJ 頻度 が最も重要な素性であると捉えることができる．しかし，どの素性にも大きな差はなく，基本素性を用いた場合と大きな差は確認されなかった．

5.3.4 考察

医療用語難易度推定に関して，表 5.5 の通り，基本素性に加えて今回新たに追加した新規素性を加えることで，難易度推定の精度が向上した．これには二つの要因が考えられる．一つは，日本語教育語彙表の難易度と性質が異なることだ．日本語教育語彙表の難易度は，BCCWJ 頻度を中心に人手で作られたものであるため，BCCWJ 頻度を中心とした少数の素性で難易度推定することが可能だったが，医療

用語の難易度はそういった頻度情報は使わずに人間の主観によって作られたものであるため，そういった単純な素性だけでは特徴を捉えきれないということが考えられる．もう一つは，医療用語には複合語が多いことだ．調査結果にもあるように，複合語の切れ目がわからないということが難しく感じる要因である可能性があり，今回新たに追加した素性によってそういった特徴を捉えることができ，難易度推定精度の向上に寄与したことが考えられる．

一般語難易度推定に関して，表 5.6 の通り，新規素性を加えることで基本素性の精度を上回る組み合わせもあったが，大きく向上することはなかった．これは，日本語教育語彙表には複合語が少ないこと（対象語の平均構成素数：1.02）が原因と考えられる．また上述の通り，日本語教育語彙表の難易度は主に BCCWJ 頻度などを中心に作られたものであり，基本素性で対象語の難易度に関する特徴を捉えられていると考えられる．

表 5.5　新規素性を加えた医療用語難易度推定結果
（括弧内の数字は全提案素性との差）

素性	R^2	neg_MAE	neg_MSE
全提案素性	**0.647**	**-0.244**	**-0.101**
w/o 先頭の文字と最後の文字の頻度情報	0.641 (-0.006)	**-0.244** (±0)	**-0.101** (±0)
w/o BCCWJ 頻度	0.635 (-0.012)	-0.246 (-0.002)	-0.103 (-0.002)
w/o 単語長	0.629 (-0.018)	-0.249 (-0.005)	-0.105 (-0.004)
w/o 文字種の文字数	0.626 (-0.021)	-0.249 (-0.005)	-0.106 (-0.005)
w/o 文字種の有無	0.625 (-0.022)	-0.247 (-0.003)	-0.104 (-0.003)
w/o 文字種の連続数	0.623 (-0.024)	-0.251 (-0.007)	-0.106 (-0.005)
w/o 素数	0.621 (-0.026)	-0.250 (-0.008)	-0.108 (-0.007)
w/o 頻度情報	0.618 (-0.029)	-0.252 (-0.010)	-0.109 (-0.008)
w/o 先頭の構成素と最後の構成素の頻度情報	0.617 (-0.030)	-0.248 (-0.004)	-0.107 (-0.006)
w/o 文字単位の頻度	0.610 (-0.037)	-0.252 (-0.008)	-0.109 (-0.008)
w/o 構成素頻度情報	0.578 (-0.069)	-0.262 (-0.018)	-0.119 (-0.018)
基本素性	0.370 (-0.277)	-0.320 (-0.076)	-0.178 (-0.077)

表 5.6　新規素性を加えた一般語難易度推定結果
（括弧内の数字は全提案素性との差）

素性	3 段階難易度推定	6 段階難易度推定
w/o 構成素頻度情報	**0.716** (+0.003)	0.547
w/o 文字種の有無	0.714 (+0.001)	0.544 (-0.006)
w/o 文字種の文字数	0.713 (±0)	0.546 (-0.004)
w/o 先頭の構成素と最後の構成素の頻度情報	0.713 (±0)	0.548 (-0.002)
全提案素性	0.713	**0.550**
w/o 文字種の連続数	0.712 (-0.001)	0.546 (-0.004)
w/o 構成素数	0.712 (-0.001)	0.549 (-0.001)
w/o 単語長	0.712 (-0.001)	0.547 (-0.003)
w/o 先頭の文字と最後の文字の頻度情報	0.712 (-0.001)	0.543 (-0.007)
w/o 文字単位の頻度	0.711 (-0.002)	0.546 (-0.004)
w/o 頻度情報	0.709 (-0.004)	0.546 (-0.004)
基本素性	0.700 (-0.013)	0.538 (-0.012)
w/o BCCWJ 頻度	0.694 (-0.019)	0.523 (-0.027)

第 6 章　おわりに

本研究では，非医療者にとって難解な医療用語の存在を，医療者と非医療者のお互いが知ることが，質の高い医療の実現のために重要であると考え，大規模な医療用語難易度リスト獲得のための医療用語難易度推定手法を提案した．まず，一般的な日本語に対する既存の日本語難易度推定手法を用いて予備実験を行った．既存の難易度推定手法の素性は，コーパスでの頻度情報，語に含まれる各文字の頻度情報，単語長，文字種の文字数の 23 次元で，これらの素性だけを用いた手法ではコーパスでの頻度情報が主な情報となり，日本語教育語彙表の難易度を推定するためには十分であったが，医療用語の難易度を決め得る特徴を捉えるには不十分な部分があった．捉えるべき医療用語の特徴を調査するため，非医療者が医療用語を難解に感じる要因をクラウドソーシングを用いて質問した．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，「構成素数」など複合語の特徴を捉えられるような素性を加えた．また，医療用語の難易度に関する教師データが存在しなかったため，クラウドソーシングを用いて作成した．これらを用いて再度実験を行ったところ，既存の素性に比べて大幅に難易度推定精度が向上した．

医療用語には複合語が多いことなどを考慮すると，今回追加した複合語の特徴を捉えられるような素性が，医療用語の難易度推定精度向上に寄与する可能性が示唆された．また，一般語の難易度推定においては，これらの素性によって精度が大きく向上しなかったが，日本語教育語彙表の語には複合語が少なかったことが要因であると考えられ，このタスク特有の現象と考えることもできる．そのため，法律用語など複合語が多いと思われるジャンルの専門用語に関しても，医療用語と同様にこれらの素性を用いることで，より正確に語の難易度を推定できる可能性がある．

複合語が多いという医療用語の特徴を本研究で提案した手法で捉えることができ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ことが期待できる．本研究の成果が，より質の高い医療の実現に役立つことを願う．

謝辞

本研究は，多くの方々からご指導・ご協力をいただいたことで推し進めることができました．本研究を行うにあたり，副指導教官である荒牧英治特任准教授には丁寧なご指導及びご助言を賜りましたこと，不自由なく日頃の研究室活動を送ることができる環境を提供していただきましたことに深く御礼申し上げます．また，主指導教官である松本裕治教授，副指導教官である中村哲教授，田中宏季助教には，ご多忙にも関わらず本研究の論文の審査を引き受けてくださったこと，そして各ゼミナール・コロキアムにおいて的確なご指導・ご助言を賜りましたことに深く御礼申し上げます．私とは異なった視点からご質問・ご指導いただけたことによって，新たな問題点や活路を見出すことができました．そして，伊藤薫博士研究員には研究テーマ決定からお世話になり，実験設計や軌道修正など様々なご相談をお受けいただきました．その都度適切なご助言を賜りましたことを深く御礼申し上げます．また，事務処理をはじめ，私たちが研究を快適に行うための環境を整えてくださったソーシャル・コンピューティング研究室のスタッフの皆様に心より御礼申し上げます．また，私が研究を行う上で不足していた知識を丁寧にご享受いただいたソーシャル・コンピューティング研究室の先輩方には心より御礼申し上げます．最後になりましたが，二年間お世話になりました同期及び後輩の学生の皆様に心より御礼申し上げます．

参考文献

[1] 梶原智之, 小町守. Simple ppdb: Japanese. 言語処理学会第 23 回年次大会発表論文集, pp. 529–532, 2017.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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 classifi-

cation. 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.

[8] 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)*, Vol. 2, pp. 143–148, 2016.

[9] 佐藤理史ほか. 均衡コーパスを規範とするテキスト難易度測定. 情報処理学会論文誌, Vol. 52, No. 4, pp. 1777–1789, 2011.

[10] Gondy Leroy and James E Endicott. Term familiarity to indicate perceived and actual difficulty of text in medical digital libraries. In *International Conference on Asian Digital Libraries*, pp. 307–310. Springer, 2011.

[11] Gondy Leroy, James E Endicott, David Kauchak, Obay Mouradi, and Melissa Just. User evaluation of the effects of a text simplification algorithm using term familiarity on perception, understanding, learning, and information retention. *Journal of medical Internet research*, Vol. 15, No. 7, 2013.

[12] 高田祥平, 荒瀬由紀, 内田諭. 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. 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. 3G205–3G205. 一般社団法人 人工知能学会, 2018.

[13] 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.

[14] 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.

[15] 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.

[16] 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*, 2004.

[17] Rudolph Flesch. A new readability yardstick. *Journal of applied psychology*, Vol. 32, No. 3, p. 221, 1948.

[18] Kirsten E Bevelander, Kirsikka Kaipainen, Robert Swain, Simone Dohle, Josh C Bongard, Paul DH Hines, and Brian Wansink. Crowdsourcing novel childhood predictors of adult obesity. *PloS one*, Vol. 9, No. 2, p. e87756, 2014.

[19] 米良俊輝, 若宮翔子, 荒牧英治, 森嶋厚行. クラウドソーシングのみによる因果関係発見の試み. 人工知能学会全国大会論文集 2017 年度人工知能学会全国大会 (第 31 回) 論文集, pp. 4O1OS17a2–4O1OS17a2. 一般社団法人 人工知能学会, 2017.

[20] Eiji Aramaki, Shuko Shikata, Satsuki Ayaya, and Shin-Ichiro Kumagaya. Crowdsourced identification of possible allergy-associated factors: automated hypothesis generation and validation using crowdsourcing services. *JMIR research protocols*, Vol. 6, No. 5, 2017.

[21] Joseph L Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, Vol. 76, No. 5, p. 378, 1971.

発表リスト

[1] 山本英弥，伊藤薫，矢野憲，若宮翔子，荒牧英治，検査表現と潜在的解釈の相互変換方法とその評価法の設計，第 37 回医療情報学連合大会.
[2] Hideya Yamamoto, Kaoru Ito, Chihiro Honda, Eiji Aramaki, Does Digital Dementia Exist?, In Proceedings of AAAI 2018 Spring Symposium.
[3]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.

付録

万病辞書に掲載されいている信頼度が S の語の中から 5,000 語を対象に難易度推定し，推定難易度が低い 50 語と高い 50 語を以下に示す．学習データには，5.3.1 節で作成した 1,000 語のデータ，素性には全提案素性を用いた．

推定難易度が低い 50 語

語	推定難易度
脳腫瘍	2.276
心筋梗塞	2.299
胃潰瘍	2.313
白血病	2.323
損傷	2.327
肺炎	2.354
不整脈	2.368
妊娠	2.372
骨折	2.373
心不全	2.386
精神病	2.388
癌	2.392
狂犬病	2.395
不眠症	2.402
自閉症	2.403
狭心症	2.409
幻覚	2.415
脱臼	2.415
腱鞘炎	2.416
老衰	2.419
衰弱	2.424
脳出血	2.433
神経症	2.437
結核	2.448
骨粗鬆症	2.454

語	推定難易度
貧血	2.455
溺死	2.460
即死	2.460
食中毒	2.462
昏睡	2.463
緑内障	2.465
肺結核	2.472
敗血症	2.473
窒息	2.475
気管支炎	2.476
悪夢	2.477
捻挫	2.480
浮腫	2.480
白内障	2.480
赤痢	2.483
不器用	2.484
糖尿	2.484
発熱	2.486
脳卒中	2.487
肝炎	2.489
帯状疱疹	2.489
外傷	2.490
肺癌	2.491
神経質	2.493
結膜炎	2.494

推定難易度が高い 50 語

語	推定難易度	語	推定難易度
肺サルコイドーシス	3.388	進行性骨化性線維異形成症	3.406
筋強直性ジストロフィー	3.389	気腫性腎盂腎炎	3.407
髄膜炎菌性心内膜炎	3.391	アルコール性多発ニューロパチー	3.415
四日熱マラリア	3.391	会陰裂傷第 1 度	3.411
進行性多巣性白質脳症	3.394	筋サルコイドーシス	3.415
ソマトスタチノーマ	3.394	二次性甲状腺機能亢進症	3.415
交代性上斜位	3.394	低トロンビン血症	3.416
会陰裂傷第 4 度	3.395	線状強皮症	3.416
ハシトキシコーシス	3.395	癌性ニューロパチー	3.419
水晶体原性虹彩毛様体炎	3.396	近位尿細管性アシドーシス	3.419
三日熱マラリア	3.397	全身性エリテマトーデス	3.424
高グルカゴン血症	3.397	二次性肺高血圧症	3.426
手掌筋膜拘縮	3.398	白質ジストロフィー	3.427
高ガストリン血症	3.398	心因性多飲症	3.432
会陰裂傷第 3 度	3.399	反応性穿孔性膠原線維症	3.433
高メチオニン血症	3.400	ラ音	3.433
肺アミロイドーシス	3.400	異染性白質ジストロフィー	3.436
高プロリン血症	3.401	異型乳管過形成	3.438
好酸球性膿疱性毛包炎	3.401	多巣性運動ニューロパチー	3.440
低レニン性低アルドステロン症	3.401	心アミロイドーシス	3.442
遠位尿細管性アシドーシス	3.402	心サルコイドーシス	3.445
肝サルコイドーシス	3.404	両大血管右室起始症	3.446
会陰裂傷第 2 度	3.405	手指振戦	3.462
眼サルコイドーシス	3.405	円板状エリテマトーデス	3.463
腎アミロイドーシス	3.405	二次性高脂血症	3.464