日本語単語の難易度推定の試み

水谷 勇介　　河原 大輔　　黒橋 禎夫
京都大学 大学院情報学研究科
{mizutani, dk, kuro}@nlp.ist.i.kyoto-u.ac.jp

1 はじめに

言語の学習において語彙の学習は非常に重要である。そこでは、易しい単語から徐々に難しい単語へという、語の難易度を踏まえた学習が必要である。これは、母語の学習においても、外国語の学習においてもそうであるし、今後、自然言語処理において意味の問題をより深く取り扱っていく上でも同様であると考える。

本論文では、さまざまな手がかりを用いて、語の難易度を 20 段階に分類する手法を検討する。本研究における単語の難易度は、単語が表す概念の難しさとする。つまり、読みや書きの難しさには左右されないものである。例えば「薔薇」という単語は漢字で書くのは難しいが、メジャーな花の名称であり、概念としての難易度は低い。

単語が表す概念の難しさといっても、誰にとっての難しさであるかによって順序は異なると考えられる。母語話者にとっての難しさと外国語としての学習者にとっての難しさは異なるであろうし、子供と大人、また個人差の問題もあるだろう。本研究では母語話者にとっての難しさを考えることとし、年代や個人差の問題はまずは考えないこととする。

分類を 20 段階としたことは、従来の研究・リソースにあるような一桁段階程度では学習の際の目安として十分でなく、逆に 50〜100 段階というような細粒度の分類は現実的ではないと考えたためである。単語の難易度は定義を含めて極めて難しい問題であることを覚悟の上で、今回はひとつの試行を行った。

2 関連研究

人手もしくは自動で単語に難易度を設定する研究はいくつか存在するが、それらは 3〜7 段階程度の粗い粒度の難易度を対象としている。

言語認知に関する研究において、単語親密度のデータベースが構築されている [1]。単語親密度とは、単語に対する被験者の主観的な馴染みの程度を表した尺度であり、複数の被験者が 1 から 7 までの 7 段階で評定したものを平均化したものである。被験者がよく理解している単語ほど親密度が高くなる傾向があるため、単語の難易度と相関があると考えられる。

難解な語句を平易な同義表現に変換するためのリソースとして、Simple PPDB: Japanese [2] が構築されている。この研究では、単語の難易度として、日本語教育語彙表 [1] の 3 段階 (初級、中級、上級) の難易度を採用し、SVM を用いた 3 クラス分類を行っている。SVM の素性としては、単語長、文字種、単語頻度、単語分散表現を用いている。学習データとして利用している日本語教育語彙表は、日本語を母語としない人々の日本語学習のために開発された辞書である。本研究では、単語頻度は手がかりとして用いる一方、文字種などは漢字の読み書きの難易度が考慮されてしまうので手がかりとして用いていない。

日本語を母語にしない人々の日本語学習を支援するための研究が行われている [3] [4]。しかし、これらの研究は、既に習得している母語と日本語の結びつけによる学習を想定しており、本研究で対象とする母語としての日本語学習を支援する試みとは目的が異なる。

3 単語の難易度推定

3.1 概要

本研究では、教科書コーパス語彙表 [5] を単語難易度の基準データとし、これを学習データとした機械学習によって単語難易度を推定する。機械学習には、Support Vector Regression (SVR) を用い、単語難易度と相関が高いと考えられる 3 種類の手がかりを素性として用いる。なお、本研究で扱う単語は、形態素解析器 JUMAN++[2] の単語辞書に掲載されている単語を基本とし、一文字漢字の単語の場合にはその直前の単語を付加したもの (「愛国/心」など) も単語とみなす。

[1] http://jhlee.sakura.ne.jp/JEL.html
[2] http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN++

教科書カテゴリ	語数
小学校前半	4,800
小学校後半	7,300
中学校	11,000
高校	21,000
合計	44,000

表 1: 教科書カテゴリごとの単語数

文書難易度	文書数
1	9,600
2	5,400
3	30,000
4	82,000
5	150,000

文書難易度	文書数
6	53,000
7	97,000
8	62,000
9	12,000
合計	500,000

表 2: 文書難易度ごとの文書数

3.2 リソース

以下では、まず教科書コーパス語彙表について説明し、次に単語難易度推定のための 3 種類の手がかりについて述べる。

3.2.1 教科書コーパス語彙表

教科書コーパス語彙表は、教科書コーパスから抽出した語彙リストである。教科書は国が学校教育を目的に検定したものであり、単語難易度の基準データとしてふさわしいと考え、単語難易度を機械学習するための学習データとして用いることにした。ここで注意すべきことは、単語難易度の正解データは存在せず、教科書コーパス語彙表をあくまで基準として用いていることである。

本語彙表には、小学校前半、小学校後半、中学校、高校の 4 つのカテゴリにおける単語出現頻度が掲載されている [3]。単語ごとに、出現した最も低い学年のカテゴリ (初出) を抽出し、表 1 に示すカテゴリごとの単語リストを得た。この初出のカテゴリを**教科書カテゴリ**と呼ぶ。

3.2.2 ウェブ出現頻度

単語の出現頻度が高くなるほど、単語の難易度は低くなる傾向があると考えられるため、単語の出現頻度を単語難易度推定の手がかりとして用いる。日本語 100 億文からなるウェブコーパスの形態素解析済みデータを用いて単語の出現頻度を計数した。

3.2.3 文書難易度初出・平均

文書の難易度が高くなるほど、そこに出現する単語の難易度も高くなる傾向があると考えられる。本研究では、コーパスを用意して各文書の難易度を推定し、それに基づく単語ごとの文書難易度を単語難易度推定の手がかりとして用いる。

まず、文書難易度を推定するために用いた「帯」[6] について説明する。このシステムは、「現代日本語書き言葉均衡コーパス」(BCCWJ) に 9 段階の難易度 (1-9) を付与し、難易度ごとに言語モデル (文字 bigram) を学習する。入力された文書に対して、各難易度の言語モデルで尤度を計算し、最大尤度をとる言語モデルに対応する難易度を出力する。BCCWJ への難易度付与は、次の四つのステップで行う。

1. シードとなる難易度つきコーパスを準備する
   本システムでは、教科書コーパス [7][4] という小中高大の各学年の教科書から抽出した 13 段階のコーパス (約 1,500 文書) が用いられている。
2. 難易度つきコーパスを用いて、文書難易度比較機を構築
   文字 bigram を素性とし、入力された文書ペアのどちらが難しいかを機械学習により判定している。
3. 文書難易度比較器を用いて BCCWJ の約 10,000 文書をソートし、正規分布に基づいて 9 段階の難易度を付与
4. 信頼度の高い難易度付与結果を抽出して新たな難易度つきコーパスとしてステップ 2 へ戻る

「帯」を 50 万文書からなるウェブコーパスに適用し、文書の難易度測定を 9 段階で行った。文書難易度ごとの文書数を表 2 に示す。本研究では、文書難易度初出と文書難易度平均を単語難易度推定の手がかりとして用いる。文書難易度初出とは、単語が出現した文書の中で最も低い文書難易度を指す。文書難易度平均とは、その単語の難易度ごとの出現頻度で重み付けした文書難易度平均値である。なお、難易度ごとの出現頻度は、各難易度の文書数 (表 2) で正規化している。

3.2.4 漢検難易度初出

日本漢字能力検定 (漢検) は、漢字能力を測定する技能検定である。10 級〜1 級の 12 段階の級からなり、10 級が小学校 1 年生修了程度で、1 級がもっとも難しい。本研究では、漢検の問題文中に使用された単語に

[3] 教科書コーパス語彙表は、UniDic に基づく単語単位を採用しており、約 50,000 語からなる。ここから、3.1 節で説明した我々の「単語」に合致するものを抽出し、約 44,000 語を得た。

[4] 3.2.1 節の教科書コーパスとは異なる別のものである。

級	語数	級	語数
1 級	6,700	5 級	1,600
準 1 級	5,100	6 級	1,900
2 級	2,200	7 級	2,400
準 2 級	2,300	8 級	2,200
3 級	2,600	9 級	1,200
4 級	3,200	10 級	1,700
		合計	33,000

表 3: 漢検問題文における初出の級ごとの単語数

ついて、出現した最も簡単な級 (初出) を単語難易度推定の手がかりとして用いる。平成 4 年から 28 年までに出題された漢検の問題文のうち、読み問題と書き問題 (選択式の問題は除外) に出現する単語を抽出し、表 3 に示す級ごとの単語リストを得た。

4 単語難易度の推定結果

4.1 実験設定

教科書コーパス語彙表から抽出した単語リストにおいて、すべての手がかりが存在する単語を抽出し、訓練・テストデータとして利用した。これは約 14,000 語からなる。教科書カテゴリの小学校前半から順に 1 から 4 の難易度を表す数値を与え、SVR を用いて 10 分割交差検定によって単語難易度を推定した。SVR には、scikit-learn (0.17.1)[5] の RBF カーネルを利用した。推定した実数値 (1〜4) を 0.15 ずつ 20 分割し、20 段階の難易度にマッピングした。これを**推定難易度レベル**と呼ぶ。20 がもっとも高い推定難易度レベルである。

4.2 結果と考察

表 4 に、推定難易度レベルと教科書カテゴリとの対応の例を表す。同じ教科書カテゴリに属する単語について、1 から 20 の推定難易度レベル順に見てみると、ある程度難しさの順番に並んでいるように感じられる。学習データとして用いた教科書コーパス語彙表のカテゴリは 4 段階しかないものの、それがうまく細粒度に分割できたと考えられる。

しかし、小学校前半の単語がレベル 19 までの高い推定難易度レベルにまで分布していることは改善の余地があり、小学校前半であれば高くてもレベル 10 程度までに分布している方が望ましいと考えられる。高校の単語がレベル 2 以外の広範囲のレベルに分布していることも同様である。

[5] http://scikit-learn.org/

個々の単語を見てみると、「落書き」や「ロッカー」など、教科書カテゴリでは高校に分類されているものの推定難易度レベルが低い単語や、「曙」や「編纂」など、小学校前半に分類されているものの推定難易度レベルが高い単語がある。これらは教科書カテゴリが適切ではなく、適正な難易度レベルが推定できていると考えられる。教科書カテゴリは、教科書コーパス語彙表において各カテゴリの出現頻度が 1 回でもあれば抽出しているため、データスパースネスの影響を受けていることが原因の一つと考えられる。

一方、低い推定難易度レベルに出現する「厳禁」や「打ち上げ」、高い推定難易度レベルに出現する「麦畑」や「ハンガー」などには違和感を感じる。「厳禁」に関しては、漢検の問題文中に「土足 げんきん」(「土足」の読み問題) という形で 10 級に、「打ち上げ」に関しても「打 ち上げ花火」(「打」の読み問題) という形で 8 級に出題されていたため、推定難易度レベルが低くなったと考えられる。

高い推定難易度レベルに「麦畑」や「ハンガー」のような簡単な単語が出現する原因としては、これらの単語は概念としては簡単であるものの、ウェブコーパス中に出現することが少ないため、ウェブ出現頻度などの手がかりが難易度の高い単語と似た傾向を示してしまったと考えられる。このように、我々が日常的に使っている単語の頻度がウェブコーパスには正確に反映されていないことが問題の一つとして挙げられる。

次に、難易度推定に用いた 4 つの手がかりのうち、どの手がかりが有効に働いたかを検証するために、4 つの手がかりから 1 つずつを除外したモデルを学習した。それぞれのモデルによる推定難易度レベルと教科書カテゴリの難易度値 (1〜4) との平均二乗誤差を計算した結果を表 5 に示す。結果としては、文書難易度初出と漢検難易度初出を除外した推定の結果がそれぞれ悪化した。従って、文書難易度初出と漢検難易度初出が難易度レベルの推定に重要であることがわかる。これは、単語を使用し始める境界である初出が、単語の概念としての知識を有しているか否かと関係していることを示している。一方、ウェブ出現頻度や文書難易度平均は、ウェブコーパスにおける単語出現頻度の影響があるため、本研究における単語難易度推定には有効に働かなかったと考えられる。

5 おわりに

本研究では、教科書コーパス語彙表を基準データとして、20 段階からなる単語難易度を推定した。手が

	小学校前半	小学校後半	中学校	高校
1	足跡, 電話, カード, 出来る	テーブル, 未知だ, 沖, 老いる	出口, 焼き芋, 財布, 一見	厳禁, 打ち上げ
2	地図, 砂漠, 郵便, 会社	空き地, 灰色, 貯金, ストロー	シチュー, ラケット, 中止, 本棚	
3	鞄, 鳴る, 揺れる, 苺	待ち合わせ, 平たい, 炭火, クリスマス	役員, 進学, 王子, 立ち寄る	落書き, 出店, 後回し, 刺さる
4	測定, 進める, 都合, 用紙	消毒, 土手, 区間, 通行	鳴り響く, 昼前, 競争, 蒸し焼き	引き戸, ロッカー
5	全国, 茹でる, 水玉, 電柱	逆上がり, ハムスター, 蛍, 見渡す	サーカス, 眠り, 苦笑い, 夜明け	玉手, 増産, 酒器, 満席
6	身近だ, 目立つ, 本文, 児童	周囲, 日頃, 野外, 種目	湧き出る, 新年, 成果, 上達	塩気, 切らす, 流れ星, 鼻歌
7	原稿, 等しい, 幸福だ, 主語	包帯, 通信, 目薬, 旅先	半袖, 休息, 始発, 会見	極力, 断然, 隈なく, 身元
8	確保, 箇所, 点検, 埋める	手ぬぐい, 冗談, 特色, 海底	判定, 愛らしい, 雑炊, 定食	商談, 出没, 自粛, 吸い物
9	茄子, 不意だ, 模型, 揺する	負担, 泡立てる, 雨雲, 船旅	家中, 講習, 踏切, 親友	標的, 切り札, 公認, 感想
10	ヒソヒソ, 日和, 順序, 学者	携わる, 輸出, 引き継ぐ, 総理	鏡, 作動, 禁物, 資質	夜通し, 検挙, 増築, 伝授
11	岸辺, 障子, 設備, 初夢	個体, 妻子, 謝罪, 隅々	無罪, 天下, 憎む, 系統	鏡台, 番茶, 応接, 急病
12	銅線, 至急, 暮れる, 若草	適する, 赤ん坊, 命令, 追放	操縦, 月刊, 面接, 相続	悲願, 兼用, 字幕, 離脱
13		叫び, 戦火, 捧げる, 近代	療法, 出現, 優勢だ, 工学	忠言, 耳たぶ, 全快, 絶壁
14	女将, 度胸, 作詞, 叫び声	丘陵, 商人, 茶の間, 皇后	競合, 滅亡, 昇進, 休眠	高慢だ, 明白だ, 筆跡, 猛威
15	ハンガー, ロンドン, 打ちのめす, 言い伝え	魚河岸, 書院, 漢語, 沼地	雨傘, 世俗, 濃紺, 彩色	印紙, 球菌, 忌引き, 粗暴だ
16	編纂, 水はけ	挿絵, 五色	一泡, 鼻緒, 軍備, 諛う	誘引, 地殻, 仏間, 年譜
17	揺りうごかす, 凸凹だ, 麦畑, 曙	仮名遣い, はにわ, 漆器, 渡来	地平, 猿人, 誹諧, 儒学	贈賄, 弔問, 兵糧, 藻屑
18	硯	武芸, 総画, 産湯, 参政	字句, 洋間, 世直し, 慣用	出仕, 精彩, 史家, 寵児
19	風見	一揆, 山伏, 旧居	現出, 知行, 遺構, 独善	接収, 傀儡, 居士, 軽重
20		起こり	紅花, 洛中, 篳篥, 保元	宝珠, 政教, 寛仁, 封ずる

表 4: 推定難易度レベルと教科書カテゴリの対応の例

使用する手がかり	平均二乗誤差
全ての手がかり	0.877
− ウェブ頻度	0.871
− 文書難易度初出	0.930
− 文書難易度平均	0.877
− 漢検難易度初出	0.943

表 5: 手がかりを一つずつ除外したときの平均二乗誤差

かりとしては、ウェブ出現頻度、文書難易度初出・平均、漢検難易度初出を用いたが、そのうち有効に働いた手がかりは文書難易度初出と漢検難易度初出であった。今後は、この難易度推定結果を用いた漢字の学習アプリを開発し、日本語学習を支援する研究を進めていく予定である。また、本研究で構築した単語難易度データベースは公開する予定である。

謝辞

本研究は京都大学と（公財）日本漢字能力検定協会の研究プロジェクト「人工知能（ＡＩ）による漢字・日本語学習研究」のもとで実施された。（公財）日本漢字能力検定協会からの研究助成に感謝致します。

参考文献

[1] 天野成昭, 近藤公久. 日本語の語彙特性, 第 1 巻. 三省堂, 2000.

[2] 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.

[3] 藏培慶, 小林伸行, 椎名広光. 単語難易度推定による中日単語学習システム. 言語処理学会第 20 回年次大会発表論文集, pp. 113–116, 2014.

[4] 中西聖明, 木藤善信, 木村祐介, 椎名広光, 北川文夫. 日本語の単語難易度推定による VOD 講義の難易度推定. Technical report, Information Processing Society of Japan, 2011.

[5] 田中牧郎, 相澤正夫, 斎藤達哉, 棚橋尚子, 近藤明日子, 河内昭浩, 鈴木一史, 平山允子. 言語政策に役立つ，コーパスを用いた語彙表・漢字表等の作成と活用, 2011. 特定領域研究「日本語コーパス」言語政策班.

[6] 佐藤理史. 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.

[7] 松吉俊, 近藤陽介, 橋口千尋, 佐藤理史. 全教科を収録対象とした日本語教育コーパスの構築. 言語処理学会第 14 回年次大会発表論文集, pp. 520–523, 2008.