辞書の意味を利用した日本語単語と英語単語の難易度推定

中西　聖明 [1]　木藤　善信 [2]　木村　祐介 [3]　椎名　広光 [4]　北川　文夫 [5]
NakanishiOus@gmail.co.jp[1],sigel4280@live.com [2], CmbdlDpg2f2@gmail.com[3],
shiina@mis.ous.ac.jp[4], kitagawa@mis.ous.ac.jp[5]
岡山理科大学大学院　総合情報研究科 [1,2,3]
岡山理科大学　総合情報学部 [4,5]

1　まえがき

日本語を母国語としない留学生などの日本語学習者が日本語を学習する際に，学習すべき同一レベルの単語を提示できると，その必要性に答えらえるのではないかと期待される．しかしながら，日本語を母国語とするものは，単語レベルを表す指標を持っているわけではなく，漢字の難易度に頼っている印象があり，日本語学習者とは差異があるように感じられる．

そこで本研究では，日本人からみた日本語単語の難易度と日本語を母国語としないものからみた日本語単語の難易度を評価し，同様に英語の単語に対しても母国語とするものとそうでないものからみた評価を行うのが目標である．手法としては，日本人向けの日本語試験の漢字の日本漢字能力検定 [5] や外国人向けの日本語能力試験 [2, 3] で出題された単語の級を初期値として，国語辞書の意味で使われている単語の難易度分布 (級別頻度) を多クラスサポートベクタマシン [1](以降,SVM) を用いて，見出し語の日本語難易度の推定を行う．そして推定された日本語難易度を再び意味に適用してすべての見出し語の難易度を全て計算し終わるまで繰り返している．同様に英英辞書 [6] に対して英語による意味を利用して難易度を推定を行い日本語と英語の単語の難易度の特徴について述べる．

また，国語辞書の記述は編纂者により統一化された静的なデータであるのに現在一般的に利用されている記述である動的なデータを考えることも必要である．なぜなら他国語の学習を行う際にも学習言語での Web 検索を行ったり日常的に他国語ページを閲覧することなど Web ページを利用した多くの学習方法が考えられためである．そこで辞書データを Web 検索データに置き換えた難易度推定についても実施した．

表 1: 初期学習データの組み合わせ

学習対象	初期学習データ	
	母国語者用	非母国語者用
(日) 国語辞書	① 漢字合成	② 日本語能力試験
(英)WordNet	③ Spelling Bee	④ 英検
(日)Web データ	⑤ 漢字合成	⑥ 日本語能力試験
(英)Web データ	⑦ Spelling Bee	⑧ 英検

2　学習対象と SVM への初期学習データの組み合わせ

本研究では，単語難易度を推定に利用している SVM の学習パラメータを，日本語と英語の 2 種類，辞書データと Web 検索データの 2 種類，難易度の初期データを母国語の利用者と母国語としない利用者の 2 種類の組み合わせ 8 種類に分けて行っている．難易度の初期データは，組み合わせによって相違し，本研究では表 1 のように難易度データを初期学習データとしている．

3　辞書データを対象とした難易度評価

3.1　SVM への辞書データによる学習パラメータの構築

辞書からの学習パラメータの作成は，初期学習データから見出し語の難易度を教師データ，見出し語の意味内の単語の難易度頻度分布を正規化して学習パラメータとする．初期データにないために難易度が決まっていない見出し語に対しては，SVM による意味内の単語の難易度が推定できてからその難易度頻度分布を用いて SVM によって推定する．最終的には，初期データ以外の見出し語は意味中の単語の難易度頻度

Copyright(C) 2012 The Association for Natural Language Processing.
All Rights Reserved

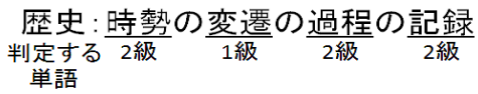

図 1: 辞書の難易度パラメータ

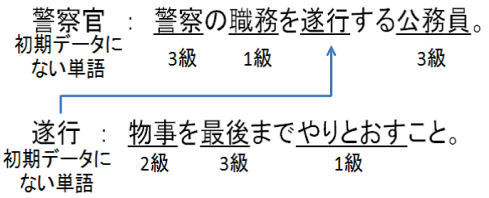

図 2: 見出し語の難易度推定

分布を学習パラメータとして推定を行う．これらについて以下に示す．

(1) 辞書データによる初期学習パラメータ

難易度が L 級区分あるとしたときに，見出し語 w の意味内に現れる単語の難易度 (級) ごとの頻度を $D(l,w)$,$l = 1 \ldots L$, とすると，正規化した難易度頻度分布 $DR(l,w)$ は，

$$DR(l,w) = \frac{D(l,w)}{\sum_{i=1 \ldots L} D(i,w)}$$

で求められる．

国語辞書の見出し語「歴史」の図 1 の例では，意味の記述が「時勢の変遷の過程の記録」となっており，その時の各単語難易度ごとの頻度から難易度分布 $DR(l,\mathrm{w})$ が求められ，$DR(1,\text{歴史}) = 0.25$，$DR(2,\text{歴史}) = 0.75$，$DR(3,\text{歴史}) = 0.00$，$DR(4,\text{歴史}) = 0.00$ となる．

(2) 辞書データによる SVM の繰り返し処理による見出し語の難易度推定

初期データにないために難易度が決まっていない見出し語の難易度 (級) は，見出し語の難易度 (級) を推定するため学習パラメータが決定してから，言い換えると意味中に現れる単語の難易度 (級) が決定してから SVM によって推定を行う．図 2 の例では見出し語「警察官」の意味中には難易度 (級) 未推定単語「遂行」が含まれる．そこで「遂行」の難易度 (級) を SVM により推定した後に「遂行」の推定結果を利用し見出し語「警察官」パラメータ作成を行う．

3.2 日本語単語の難易度データ

日本語単語の難易度については，単語の難易度よりも漢検等の漢字難易度の方が知られている．そこで日本語単語の難易度は，単語を構成する漢字難易度から合成する．漢字難易度には漢字能力検定の難易度を利用し，単語を構成する漢字のうち最も難易度の高い級別の漢字を単語の難易度とする．図 3 の単語「難易度」の例では「難易度」を構成する漢字のうち最も難しい級別の漢字「難」の級別の 5 級を単語「難易度」の難易度とする

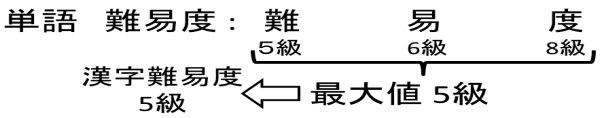

図 3: 日本語単語の難易度設定

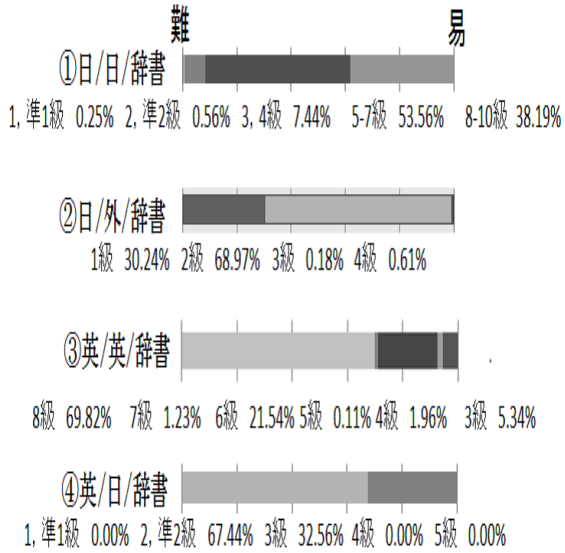

図 4: 辞書データに対する SVM による単語の難易度推定結果

3.3 辞書データによる難易度推定の結果

国語辞書を対象とし，初期データを第 3.2 節で述べた漢字の難易度から生成されたことで日本語単語の難易度があらかじめ判明しているデータをテストセットとした結果を図 4-①に示す．また初期データを日本語の学習者を対象とした日本語能力試験の試験区分に用いた場合の結果を図 4-②に示す．

一方，英語辞書を対象とし，英語を母国語とする人を対象とした Spelling Bee の試験区分に用いた時の Spelling Bee での試験区分があらかじめ判明しているデータをテストセットとした結果を図 4-③に示す．また，初期データに英語の学習者を対象とした英検の難易度を用いた場合の結果を図 4-④に示す．

Copyright(C) 2012 The Association for Natural Language Processing.
All Rights Reserved

4 Web 検索データを対象とした難易度評価

4.1 SVM への Web 検索データからの学習パラメータの構築

Web 検索データからの学習パラメータの作成には，難易度の初期データを用いて難易度を推定するが，初期データをテストセットとした結果が収束するまで検索を繰り返すことで難易度が未決定の単語の評価を行う．最終的にはテストの結果が収束するまでに難易度を推定した全ての単語の検索データの難易度分布をパラメータとして学習する．

4.2 Web 検索データからのパラメータ作成

Web 検索データからのパラメータ作成は，推定単語の Web 検索を行い最も上位の検索結果文中の単語の難易度を利用する．初期学習データから推定単語の難易度を教師データ，検索結果文内の推定単語を除いた単語の難易度分布を正規化したものを学習パラメータとする．初期データにないために難易度が決まっていない単語に対しては，SVM により検索結果文内の単語難易度を決め難易度分布から SVM によって推定する．これらについて以下に示す．

(1)Web 検索データによる初期学習パラメータ

難易度が L 級区分あるとしたときに，単語 w の検索結果文内に現れる単語の難易度ごとの頻度を $S(l,w)$,$l = 1 \ldots L$, とすると，正規化した難易度分布 $SR(l,w)$ は，

$$SR(l,w) = \frac{S(l,w)}{\sum S(l,w)}$$

で求められる．

推定単語「男性」の図 5 の例では，Web 検索結果文が「男性は，女性と対比される人間の性別のこと」となっており，この時の各単語難易度の頻度から難易度分布 $SR(l,\mathrm{w})$ が求められ，$SR(1,$男性$) = 0.25$, $SR(2,$男性$) = 0.50$, $SR(3,$男性$) = 0.25$, $SR(4,$男性$) = 0.00$ となる．

(2)Web 検索データによる SVM による繰り返し学習

初期データにないために難易度 (級) が決まっていない単語の難易度 (級) は，単語の難易度を推定するため学習パラメータが決定してから SVM によって推定を行う．図 6 の例では単語「視力」の検索結果文内には難易度 (級) 未推定単語単語「識別」が出現するため，

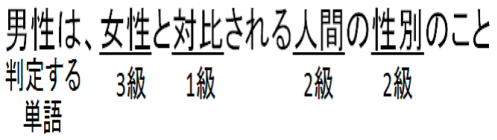

図 5: 検索結果文中出現単語級別比

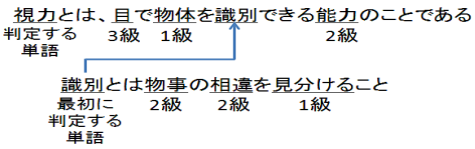

図 6: Web 検索による難易度推定

単語「識別」の難易度 (級) 推定を行った後に単語「視力」のパラメータ作成を行う．また，単語のパラメータ設定のために繰り返し検索を行う場合は以下の収束条件を満たすとき検索を打ち切り，それまでに難易度評価が行われた単語のみを単語として扱いパラメータの作成を行うこととする．初期データの難易度 (級) を d_1 から d_2 とし，難易度 (級) 推定のための検索を行っていない単語全体に検索を行った回数を t とする．初期データをテストセットとした結果，難易度 i 級の単語が j 級と推定された割合を $x^t_{i,j}(0 \leq x^t_{i,j} \leq 1)$ とすると収束条件を

$$\sum_{i=d_1}^{d_2} \sum_{j=d_1}^{d_2} \frac{(x^t_{i,j} - x^{t+1}_{i,j})^2}{(d_2 - d_1 + 1)^2} < 0.0001$$

としている．

4.3 Web 検索データによる難易度推定の結果

日本語での Web 検索データを対象とし，初期データを第 3.2 節で述べた漢字の難易度から生成された日本語単語の難易度があらかじめ判明しているデータをテストセットとした結果を 7-⑤に示す．また初期データを日本語の学習者を対象とした日本語能力試験の試験区分を用いた場合の結果を 7-⑥に示す．

一方，英語での Web 検索データを対象とし，英語を母国語とする人を対象とした Spelling Bee の試験区分に用いた時の Spelling Bee での試験区分があらかじめ判明しているデータをテストセットとした結果を 7-⑦に示す．また，初期データに英語の学習者を対象とした英検の難易度を用いた場合のテストの結果を 7-⑧に示す．

Copyright(C) 2012 The Association for Natural Language Processing.
All Rights Reserved

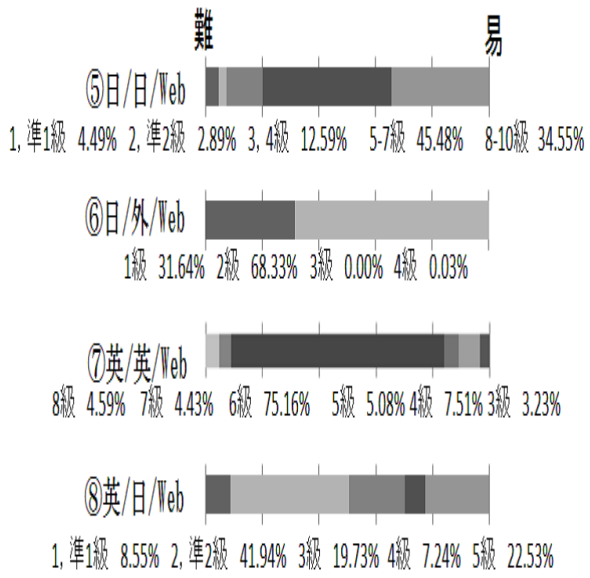

図 7: Web 検索データに対する SVM による単語の難易度推定結果

5 評価

国語辞書についての評価 (図 4-①,②) の結果から辞書に記述されている日本語は外国人にとって難易度が高く，日本人にとっては比較的容易に記述していると評価していることが読み取れる．また，英語辞書についての評価 (図 4-③,④) では母国語者であるアメリカ人と英語を母国語としない日本人の双方にとって難しく評価される結果となっている．これらのことより日本語の辞書の記述単語は日本人と外国人にとって大きく難易度差があるが，英英辞書においてはアメリカ人と日本人との難易度差が小さいと推測される．

日本語の Web 検索データの評価 (図 7-⑤,⑥) では日本人，外国人にとっての難易度評価がともに国語辞書と同程度で記述されていることが読み取れる．日本語単語の Web 検索を行った結果では Web 上の辞書，百科事典の文を利用する場合が多く，これらの Web ページが難解な単語で記述されていることが同程度の結果となった要因として推測される．英語においては英語を母国語とするアメリカ人にとっての評価 (図 7-⑦) を辞書に比べ低い難易度に評価している．英単語の検索結果では企業の Web ページが上位に表示されることが多く，企業の Web ページ上の文はアメリカ人にとって容易に理解できるよう馴染みのある単語で記述されているからだと考えられる．英語の学習者である日本人にとっての Web 検索結果文の評価 (図 7-⑧) から Web 検索結果文は辞書の評価に比べ難易度が上下に分かれる傾向があることが読み取れる．企業の Web ページ上の文のアメリカ人にとっては容易に理解できる馴染みのある単語が日本人にとっては馴染みのあるものではなく難易度が上がりうる場合もあったと推測される．

6 まとめ

本研究では辞書データ，Web 検索データを利用して日本語単語と英単語の難易度評価を行った．また，初期学習データとしては各言語の学習者とその言語を母国語とする人にとっての難易度を利用し，SVM による単語の難易度評価を行った．結果として日本語においては辞書の意味と Web 検索結果文の難易度評価の差異が小さく，日本人向けの基準と外国人向けの基準の難易度差が大きくなった．一方，英語においては辞書の意味と Web 検索結果文の難易度評価の差異が大きく，アメリカ人向けの基準と日本人向けの基準の難易度差が小さくなった．また，現状では使用する辞書の性質が難易度に大きく影響しており，このことが各評価間の差異に繋がった一因と考えられる．単語評価においても初期学習データの対象年齢，学歴等に差異があるためこれらの差異を考慮しなければならない．今後は， 学習する辞書データの差異を吸収し，SVM の分離平面の適切な選択の手法に本研究を発展させたいと考えている．

参考文献

[1] V. Vapnik, Statistical Learning Theory, Springer, 1998.

[2] 日本語能力試験公式ウェブサイト, http://www.jlpt.jp

[3] 徳弘康代, 日本語学習のためのよく使う順漢字 2100, 三省堂,2008.

[4] 北原保雄, 明鏡国語辞典 第二版, 大修館書店, 2010.

[5] 日本漢字能力検定協会, http://www.kanken.or.jp/index.php

[6] WORDNET, http://wordnet.princeton.edu/

[7] 実用英語技能検定, http://www.eiken.or.jp/

[8] Scripps National Spelling Bee, http://www.spellingBee.com/

[9] Yahoo Japan デベロッパーネットワーク, http://developer.yahoo.co.jp/

Copyright(C) 2012 The Association for Natural Language Processing.
All Rights Reserved