Vol.2011-DBS-153 No.8
2011/11/3

日本語の単語難易度推定による VOD 講義の難易度推定

中西聖明[†1]　木藤善信[†1]　木村祐介[†1]
椎名広光[†2]　北川文夫[†2]

日本人教員が留学生に講義を行う際に日本語の使用レベルが理解できていないために，留学生は非常に困難な状況に置かれていると考えられる．そこで講義での日本語使用レベルを把握するために，資料が残っている VOD 講義で使われているスライドと発話に対して日本語単語の難易度判定を行い，それを利用して講義の難易度を評価する．日本語の難易度は，日本語能力試験の 4 区分を用いて行い，日本語能力試験の問題で使用されている漢字や単語の級を難易度として用いる．漢字や単語がこれまで日本語能力試験に出現していない場合は，難易度を推定によって求める．未区分の漢字や単語の難易度推定法については，日本語の辞書データを利用して漢字や単語の難易度をサポートベクタマシン (SVM) による学習し，未区分である物に対して Web 上の文から学習パラメータを求めてサポートベクタマシンによる難易度判定を行う．

Estimation method of difficluty on VOD lecture by rating of Japanese words

In recent years, for many teachers have insufficient comprehension of international students, many international students have difficult problem in learning. Then, for comprehension of Japanese on the lecture, we estimate difficulty of lecture by ratio of japans words in VOD lecture. We classify Japanese word difficulty into four classes as to JLPT (Japanese Language Proficiency Test), and we deal with kanji and word class, which are used in JLPT, as difficulty. Kanji and word have not appeared in JLPT, they estimate difficulty. Regarding kanji and words is not belonging in classes, they are estimated by SVM (Support Vector Machine) using Japanese dictionary data. System learns kanji and word difficulty by using Japanese dictionary data with SVM (Support Vector Master). On the other hand, parameters of unknown word in Japanese dictionary data are estimated by using online data as Web search engine.

1. はじめに

日本語を母国語としない留学生にとっては，日本語がある程度上達しなければ日本人向けの講義を受講することは容易ではなく，ほとんど理解ができていないと考えられる．一方，講義を行う日本人教員も留学生に講義を理解してもらいたいと考えているが，講義の現状を理解できているわけではない．そこで講義での日本語使用レベルを把握するために，講義の発話や資料がすべて残っているインターネット環境を利用して講義を行う VOD 講義を対象に，VOD 講義で使われている Microsoft PowerPoint(以下，PPT) と PPT 毎の発話に対して日本語の難易度を評価法を提案する．特に，本研究では評価手法として VOD 講義で使われている PPT と PPT 毎の発話の日本語の漢字や単語を日本語能力試験[22] の試験区分を利用して，講義の日本語難易度として評価する．また，日本語能力試験や徳弘による試験区分データ (以下，試験区分データ)[13] に，これまで現れていない漢字や単語に対しては，難易度を推定によって求める．難易度が不明の漢字や単語の難易度推定法については，日本語の辞書データを利用して漢字や単語の学習パラメータを多クラスサポートベクタマシン (以下, SVM)[11),19)] による学習を行いし，難易度を判定されていない未区分である漢字や単語に対して Web 上の文から学習パラメータを求めて SVM による難易度判定を行っている．

2. VOD システムによる e-Learning 講義システム

本研究で作成しているシステムは，岡山理科大学を含む関連 6 大学で構成している教育コンソーシアムにおける単位互換制度を利用した VOD による e-Learning 講義のシステム[7] を利用している (図 1)．特に本研究では，講義名「データベース」の PPT 情報とその発話内容を用いて，解析を行っている．

3. VOD 講義の難易度の評価 (未区分データの推定なし)

本研究では VOD 講義の難易度の評価として，日本語の難易度を利用して講義の難易度をはかろうとしている．日本語の難易度については，財団法人　日本国際教育支援協会[23]

†1 岡山理科大学大学院　総合情報研究科
Graduate School of Informatics, Okayama University of Science
†2 岡山理科大学 総合情報学部
Faculty of Infomatics, Okayama University of Science

Vol.2011-DBS-153 No.8
2011/11/3

図 1 Tablet PC 上での VOD 実行画面

表 1 PPT と講義の発話の日本語難易度 (未区分データなし)

PPT (頁)	表示時間	PPT の難易度		発話の難易度	
		漢字	単語	漢字	単語
1	00:00〜01:29	該当なし	該当なし	2.8	2.8
2	01:29〜09:12	2.5	2	2.8	2.8
3	09:12〜12:40	3.5	2.5	2.9	2.5
4	12:40〜14:26	該当なし	該当なし	3.3	3.2
5	14:26〜17:04	2.1	2.1	2.6	2.4
6	17:04〜19:10	3.5	4	2.8	2.5
7	19:10〜19:22	該当なし	該当なし	2.0	4.0
8	19:22〜21:55	2.4	2.5	2.7	2.6
9	21:55〜22:04	該当なし	該当なし	2.7	3.0
10	22:04〜24:14	2.4	1.5	2.8	2.7
11	24:14〜24:55	2.5	2	2.5	2.0

と独立行政法人　国際交流基金国際交流基金[24] による日本語能力試験[22] で使われる試験の区分 (旧試験の 1 級から 4 級を利用, 現在は N1 から N5 に区分) を利用[13] して，各試験区分で現れたことがある，または現れる候補となる単語や漢字をその難易度とし利用している[13]．表 1 に，講義名「データベース」の 14 回目第 1 セクションを PPT 頁毎に PPT と発話の出現単語と漢字の難易度の平均を示す．

4. SVM による学習と未区分単語・漢字の推定

4.1 SVM による学習と VOD 講義難易度の処理概要

前章の **3** では，講義の PPT や発話に対する単語や漢字の難易度は，試験区分データ[13] の単語の日本語能力試験で既出であるものとそれに相当する試験区分の級が判明しているものだけで，PPT 毎に難易度を判定している．しかし，単語の試験区分が判明していないものも多く存在している．それら未区分である単語や漢字の試験区分を推定する必要がある．そこで，SVM の学習パラメータの算出と学習および難易度判定を次の手順で難易度を推定する．

(1) 試験区分データ[13] の単語に対して大きく分けてつぎの 2 つから単語の学習パラメータを学習する (図 2).

●国語辞典[25] の見出し語と意味 (説明文) から単語の学習パラメータ.
●漢字の難易度から合成した単語の学習パラメータ.

作成した学習パラメータは，RBF カーネルを用いた多クラス SVM で機械学習を行う．上記の学習パラメータについては，**4.2** で示す.

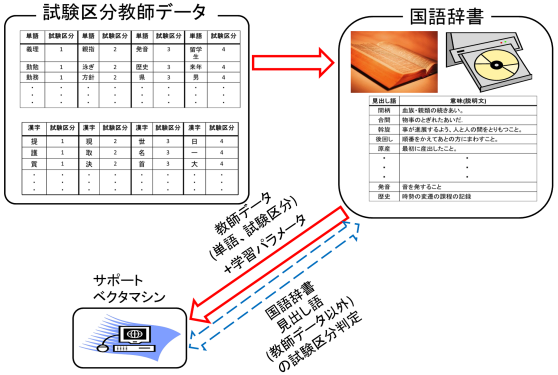

図 2 難易度の学習の概要

(2) 国語辞書の見出し語すべてを (1) で学習した多クラス SVM によって試験区分を作成する (図 2).

(3)VOD 講義の PPT の文字と発話の単語や漢字を取り出し試験区分を判定する (図 3).

(3-1) 試験区分データ[13] にある単語はその試験区分を利用する.

(3-2) 国語辞書の見出し語にある単語は (2) の結果を利用する.

(3-3)(3-1),(3-2) に含まれない未区分の単語については Web 上の検索エンジン (Google[26]) の検索結果で表れる検索結果 1 位の Web ページにある未区分単語のある文を取り出し，その文から学習パラメータを算出し，SVM により試験区分を計算する (図 3).

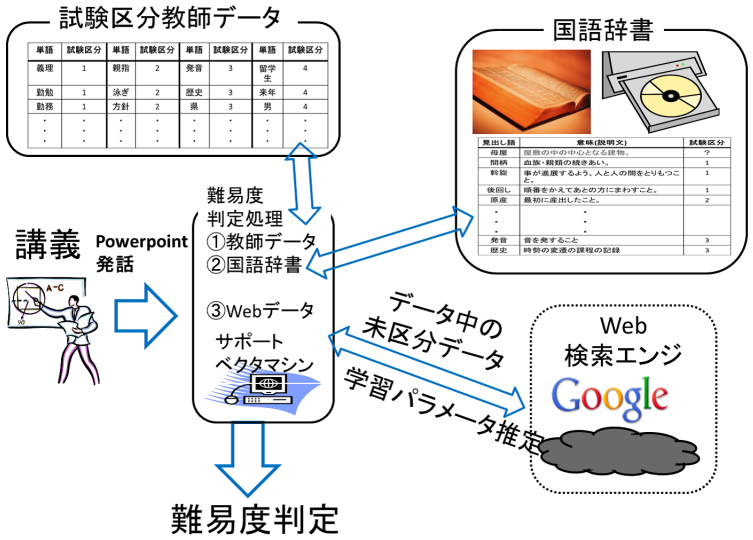

図 3 難易度の処理過程の概要

4.2 SVM に対する単語の学習パラメータ

本研究で作成する学習パラメータ (共起語級別比，係り受け級別比，漢字単語難易度，漢字意味難易度) の例を表 2 に示す．各学習パラメータについては，以下の節で述べる．

表 2 国語辞書単語難易度の学習パラメータ

語	級	C_w	D_w	共起語級別比 CR_w^L				係り受け級別比 DR_w^L				E_1^w	E_2^w
				1 級	2 級	3 級	4 級	1 級	2 級	3 級	4 級		
亜	1	24	0	0.21	0.54	0.08	0.17	0.00	0.00	0.00	0.00	1	2.1
秋	4	207	12	0.10	0.48	0.14	0.27	0.50	0.25	0.25	0.00	3	2.0
朝	4	82	2	0.05	0.22	0.09	0.05	0.00	0.00	0.50	0.50	3	1.9
味わい	1	3	0	0.33	0.67	0.00	0.00	0.00	0.00	0.00	0.00	3	2.5
意図	1	21	4	0.05	0.24	0.09	0.62	0.25	0.50	0.25	0.00	3	1.9
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
入る	4	99	4	0.13	0.52	0.11	0.23	0.25	0.75	0.00	0.00	4	2.2

4.2.1 国語辞書からのパラメータ生成

日本語単語の試験区分の判定を行うために SVM に対する学習パラメータについて説明する．学習パラメータは， 学習データを試験区分データ[13] の単語に対して，教師データの試験区分のほかに国語辞書[25] の見出し語と意味 (説明文) から共起語級別比 CR_w^L と係り受け級別比 DR_w^L の 2 種類のパラメータを作成する.

(1) 試験区分: 試験区分データから日本語能力試験で表れる試験区分 (教師データ).

(2) 共起語級別比 CR_w^L: 国語辞書中の意味 (説明文) 中の単語 w が共起する他の単語の試験区分 L の級ごとの比率.
共起する単語の試験区分の頻度を C_w^L とするとき，級ごとの比率 CR_w^L は,

$$CR_w^L = \frac{C_w^L}{C_w}, \quad C_w = \sum_{i=1}^{4} C_w^i$$

で表す.

国語辞典の見出し語「歴史」の例では， 意味 (説明文) 中の単語「時勢」に対して，意味 (説明文) に試験区分 1 級の「変遷」， 試験区分 2 級の「過程」， 「記録」があるので， 共

Vol.2011-DBS-153 No.8
2011/11/3

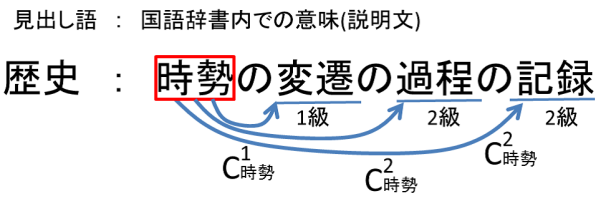

図 4 級別共起頻度

起する単語の試験区分の頻度 $C^1_{\text{時勢}}$ に 1, $C^2_{\text{時勢}}$ に 2 を加算し，国語辞書全体から頻度を計算したのち，共起語級別比 $CR^L_{\text{時勢}}$ を求める． 「時勢」以外にも意味 (説明文) にある単語「変遷」， 「過程」， 「記録」も同様に計算する (図 4).

(3) 係り受け級別比 DR^L_w: 国語辞書中の意味 (説明文) ごとに単語 w と直接係受けする単語の試験区分 L の級の比率.
係り受けする単語の級の頻度を D^L_w とするとき，級ごとの比率 DR^L_w は,

$$DR^L_w = \frac{D^L_w}{D_w}, \quad D_w = \sum_{i=1}^{4} D^i_w$$

で表す.

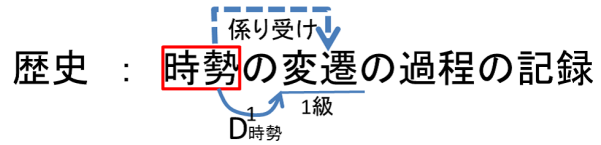

図 5 級別相関度

国語辞典の見出し語「歴史」の例では， 意味 (説明文) 中の単語「時勢」と直接係受け関係にある単語「変遷」の試験区分 1 級を利用して $D^1_{\text{時勢}}$ に 1 を加え，国語辞書全体から頻度を計算したのち係り受け級別比 $DR^L_{\text{時勢}}$ を求める (図 5).

4.2.2 漢字難易度からのパラメータ生成

(1) 漢字単語難易度 E^w_1:単語 w を構成する漢字の中で最も高い試験区分.
漢字単語難易度 E^w_1 は単語 $w = c_1c_2\cdots c_i\cdots$, c_i は i 番目の文字, 各文字 c_i の漢字難易度 C_i とするとき,

$$E^w_1 = max(C_1, C_2, \cdots, C_i, \cdots).$$

なお， 試験区分データ[13] の中で漢字難易度が記述されていないものは 1 級とする.

国語辞典の見出し語「歴史」の例では， 各漢字の試験区分が図 6 のようになっていると， 歴史の「歴」が 2 級,「史」が 2 級から最大値の 2 級が「歴史」の漢字単語難易度 $E^{\text{歴史}}_1 = 2$ となる.

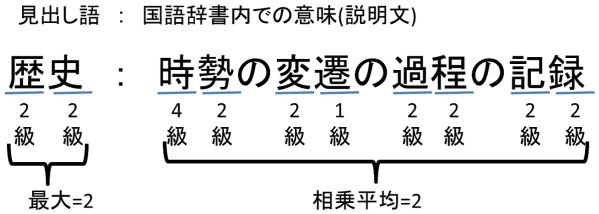

図 6 漢字の級判定

(2) 漢字意味難易度 E^w_2:意味 (説明) で表れる漢字の難易度.
単語 w の意味 (説明文) 中に現れる漢字の難易度の相乗平均で求める．試験区分データと国語辞書の見出しにない漢字については Web 上の Google 検索[26] を利用して， 第一候補の Web ページの内の対象の単語を含む文を意味として扱い，その中にも漢字がなかった場合は， 4 級とする.
単語 w の漢字別推定単語難易度を E^w_2 とすると,

$$E^w_2 = \sqrt[N_c]{\prod_{i=1}^{N_c} C_i}.$$

国語辞典の見出し語「歴史」の例では， 意味 (説明文) の漢字「時」:4 級， 「勢」:2 級,「変」:2 級， 「遷」:1， 「過」:2 級,「程」:2 級， 「記」:2 級， 「録」:2 級であるので,

Vol.2011-DBS-153 No.8
2011/11/3

$$E_2^{\text{歴史}} = \sqrt[8]{4 \cdot 2 \cdot 2 \cdot 1 \cdot 2 \cdot 2 \cdot 2 \cdot 2} = 2.0$$

5. SVM による日本語の難易度推定

学習パラメータの種類を組み合わせて 4 種類の RBF カーネルを用いた多クラス SVM[16] によって学習を行い，結果比較を行う．

5.1 SVM による日本語単語の難易度推定

初めに単語の難易度を推定するための学習として単語をベースとして算出される学習パラメータである共起語級別比, 係り受け級別比を用いて SVM の学習を行い試験区分データの再評価を行った結果ををを表 3 に示す．

また，講義データベースの試験区分データにない国語辞書の見出し語を評価した結果を表 4 に，国語辞書の見出しにない未区分単語を Web から抽出して評価した結果を表 5 に示す．

表 3 日本語単語難易度よる試験区分データの難易度推定

		正しい級別			
		1 級	2 級	3 級	4 級
判	1 級	41.00	35.71	24.57	21.14
定	2 級	16.00	9.42	21.28	16.00
結	3 級	13.42	16.86	16.86	13.43
果	4 級	29.57	38.57	37.14	49.42

表 4 日本語単語難易度よる国語辞書の見出し語の単語

1 級	2 級	3 級	4 級
28.37	10.39	12.20	31.71

表 5 日本語単語難易度よる未区分単語の Web データによる難易度推定

1 級	2 級	3 級	4 級
47.15	6.98	11.30	34.57

表 3 では， 試験区分が 1 級と 4 級に判定される傾向が強く，両端の難易度の級に押し付けてしまう傾向があるため，これを利用すると表 5 の未区分単語のようにどちらかに振り分けてしまう結果が出たと考えられる． また全体として難しく判定する傾向が考えられる．

5.2 漢字単語難易度推定から単語の難易度推定

前節の単語をベースとした学習パラメータに対して，文字をベースにして構成する単語の難易度を推定を行った．ここでは， 漢字単語難易度 E_1^w のみで SVM の学習を行っている．試験区分データの再評価を行った結果ををを表 6 に示す．

また，講義データベースの試験区分データにない国語辞書の見出し語を評価した結果を表 7 に，国語辞書の見出しにない未区分単語を Web から抽出して評価した結果を表 8 に示す．

表 6 単語の構成漢字による試験区分データの難易度推定

		正しい級別			
判	1 級	33.66	22.71	13.85	8.17
定	2 級	48.97	54.78	54.00	32.87
結	3 級	11.69	13.69	24.16	28.52
果	4 級	5.69	8.82	7.99	30.43

表 7 単語の構成漢字による国語辞書の見出し語の難易度推定

1 級	2 級	3 級	4 級
12.98	52.50	6.58	27.94

表 8 単語の構成漢字による未区分単語の Web データによる難易度推定

1 級	2 級	3 級	4 級
7.44	42.68	13.10	36.78

試験区分データ[13] は，表 6 から試験区分が 2 級に判定されているものが多く，2,3 級の中央に難易度を判定する傾向が考えらえる．よって，表 7 の国語辞書の見出し語のように 1 級ではなく 2 級が多いとと判定され，同様に表 8 の未区分単語でも 2 級が多いとなっている．

5.3 漢字単語難易度と漢字意味難易度による単語難易度推定

5.2 では， 辞書の見出し語から算出した難易度のパラメータ E_1^w のみを学習に利用したのに対して， 辞書の意味 (説明文) から算出したパラメータ E_2^w も含めた 2 つのパラメータで SVM の学習を行った．試験区分データの再評価を行った結果を表 9 に示す．

また，講義データベースの試験区分データにない国語辞書の見出し語を評価した結果を表 10 に，国語辞書の見出しにない未区分単語を Web から抽出して評価した結果を表 11 に示す．

表 9 単語, 意味の構成漢字による試験区分データ難易度推定

		正しい級別			
		1 級	2 級	3 級	4 級
判	1 級	35.14	20.28	8.00	5.14
定	2 級	48.57	58.28	51.42	24.28
結	3 級	1.14	20.28	3.42	1.42
果	4 級	15.14	1.142	37.16	69.14

表 10 単語, 意味の構成漢字による国語辞書の見出し語の難易度推定

1 級	2 級	3 級	4 級
10.26	50.22	9.01	30.51

表 11 単語, 意味の構成漢字による未区分単語の Web データによる難易度推定

1 級	2 級	3 級	4 級
8.00	44.21	7.98	39.81

試験区分データ[13] は，表 9 から試験区分が 2 級に集まる傾向がありつつ，4 級である物を 4 級に判定するように見られる．よって，表 7 の国語辞書の見出し語や表 8 の未区分単語でも 2 級と 4 級が多く，**5.2** よりも 4 級が多くなるところが特徴的である．これは辞書の意味 (説明文) から算出した E_w^2 の計算方法に依存して結果が引きずられていると考えられる．

5.4 単語難易度と漢字難易度の全てから推定

学習パラメータとして， 共起語級別比 CR_w^L， 係り受け級別比 DR_w^L，漢字単語難易度 E_1^w，漢字意味難易度 E_2^w を利用して SVM の学習を行った．試験区分データの再評価を行った結果を表 12 に示す．

また，講義データベースの試験区分データにない国語辞書の見出し語を評価した結果を表 13 に，国語辞書の見出しにない未区分単語を Web から抽出して評価した結果を表 14 に示す．

表 12 全てのパラメータによる試験区分データの難易度推定

		正しい級別			
		1 級	2 級	3 級	4 級
判	1 級	49.14	31.14	26.57	22.85
定	2 級	7.14	10.85	10.57	8.00
結	3 級	18.28	25.42	23.42	17.42
果	4 級	25.42	32.57	39.42	51.71

表 13 全てのパラメータによる国語辞書の見出し語の難易度推定

1 級	2 級	3 級	4 級
41.17	17.64	8.82	32.35

表 14 全てのパラメータによる未区分単語の Web データによる難易度推定

1 級	2 級	3 級	4 級
42.45	8.49	16.98	32.08

全ての学習パラメータを利用した方法は，共起語級別比 CR_w^L と係り受け級別比 DR_w^L から学習される結果よりも，漢字単語難易度 E_1^w と漢字意味難易度 E_2^w によっていくぶん緩和された学習結果になっているのではないかと考えられる．

6. VOD 講義の難易度の評価 (未区分データの推定あり)

先の **3**(VOD 講義の難易度の評価 (未区分データの推定なし)) では，試験区分データにのっている単語だけでそれ以外の単語は試験区分を推定せずに VOD 講義の難易度を評価した． それに対して SVM で未区分データを含めて評価した場合の VOD 講義の難易度を表 15 に示す．

未区分データの推定は， 試験区分が高い方向に推定されるので，各 PPT 毎における難易度の評価も高くなっている．また， 全てのパラメータを学習するよりも漢字から算出した E_1^W と E_2^W を使うと難易度評価が低くなっている．学習パラメータの算出方法の性質から難易度評価が低くなる傾向は類推できる．しかしながら学習パラメータの算出時間と SVM

表 15 PPT と講義の発話の日本語難易度 (未区分データ含む)

PPT (頁)	表示時間	PPT の難易度															発話の難易度													
		教師データ				辞書				Web				平均	E_1	E_1&E_2	教師データ				辞書				Web				平均	E_1	E_1&E_2
		1	2	3	4	1	2	3	4	1	2	3	4				1	2	3	4	1	2	3	4	1	2	3	4			
1	00:00〜01:29	該当なし												該当なし	該当なし	該当なし	0	6	2	4	0	0	1	0	3	2	1	2	2.6	2.8	2.1
2	01:29〜09:12	0	1	0	0	0	0	0	0	1	0	0	0	1.5	2.0	2.5	7	7	6	16	3	2	1	2	9	3	5	11	2.9	2.8	2.3
3	09:12〜12:40	1	0	0	1	0	0	0	0	1	0	0	0	2.0	2.5	1.9	4	7	3	5	0	1	1	0	1	1	3	4	2.8	2.5	2.7
4	12:40〜14:26	該当なし												該当なし	該当なし	該当なし	0	5	1	6	0	0	0	0	4	0	1	2	2.7	3.2	2.4
5	14:26〜17:04	2	4	1	1	0	1	0	1	2	0	1	1	2.3	2.1	2.4	3	3	1	4	1	0	0	4	3	1	1	1	2.5	2.4	2.6
6	17:04〜19:10	0	0	0	0	0	0	0	0	3	0	1	2	2.3	4.0	2.5	0	14	3	4	1	0	0	1	6	2	1	5	2.6	2.5	2.3
7	19:10〜19:22	該当なし												該当なし	該当なし	該当なし	0	0	0	1	0	0	0	0	0	0	0	0	4.0	4.0	2.0
8	19:22〜21:55	0	3	0	1	1	1	0	0	0	0	0	1	2.4	2.5	3.0	3	7	2	5	3	0	0	0	5	0	3	3	2.8	2.6	2.3
9	21:55〜22:04	該当なし												該当なし	該当なし	該当なし	0	1	1	0	0	0	0	0	0	1	0	0	2.3	3.0	1.9
10	22:04〜24:14	2	2	0	0	2	0	0	1	0	0	0	0	1.9	1.5	1.8	1	4	1	3	3	1	0	2	3	0	1	2	2.4	2.7	2.5
11	24:14〜24:55	0	1	0	0	0	0	0	0	3	0	0	0	1.3	2.0	2.5	0	1	1	0	0	0	0	0	1	0	0	0	2.0	2.0	2.6

の学習時間も考えると少ないパラメータで済ませるのも一つの手と考えられる．

7. ま と め

本研究では， 日本語の難易度推定を利用して VOD 講義の難易度について評価を行った．日本語の難易度としては， 日本語能力試験の試験区分を利用しており，難易度が未区分であるデータについては， SVM による結果で推定を行っている． 未区分であるデータは， 利用頻度が少ないため難しいと判定されているケースが多いと考えられるため，SVM の学習結果もそれに引きずられているのではないかと考えられる．

SVM の結果が試験区分を難しい方に判定してしまう問題点は，原因の 1 つに名詞が連接する単語が，その構成する単語が簡単なものであっても国語辞書には載っていることが少なく Web 上の検索結果を利用して難易度を判定するためと考えられる．また，2 つ目の原因として学習用に利用した国語辞書や未区分の単語を Web から検索するときの検索エンジンの結果の特徴を考慮しなければならない． 例えば， 検索エンジン[26] の場合は， 第一候補がよく利用されている Web ページであろうとの予測のもとに第一候補の Web ページを利用することにしたが，用語的な単語の多くは，それを説明する Wikipedia[27] から単語や漢字を含んだ文を取り出すことが多く，日本語としては難しい表現が多いと思われる． よって，学習データも適切なものを選ぶ必要もあるが，テストするときのデータを取り出す先も考慮しなければならいと考えられる．今後は， 学習する辞書データの変更した場合の振る舞いの違いや，優しい表現を取り出す手法に本研究を発展させたいと考えている．

参 考 文 献

1) A.P.Dempster, N.M.Laird, and D.B.Rubin.: Maximum likelihood form incomlete data via the EM algorithm. Journal of the Royal Statistical Society series B, Vol. 39, No.1, pp.1-38, 1977.
2) 三上, 増山, 中川: ニュース番組における字幕生成のための文内短縮による要約, 言語処理学会論文誌「自然言語処理」, Vol.6, No.6, pp.65-81, 1999.
3) 伊藤, 藤井, 石川: 音声文書検索を用いたオンデマンド講義システム, 電子情報通信学会技術研究報告 SP 音声, Vol.101, No.523, pp.55-60, 2001.
4) 北，津田，獅々子: 情報検索アルゴリズム, 2002.
5) Haruo Yokota, Takashi Kobayashi, Taichi Muraki, and Satoshi Naoi.: UPRISE: Unified Presentation Slide Retrieval by Impression Search Engine. IEICE Trans. on Info. and Syst., Vol. E87-D, No. 2, pp. 397-406, 2 2004.
6) 森本, 室田, 清水: 教育用動画像検索システムと時間情報同期方法の開発. 電子情報通信学会論文誌 D-I, Vol. J88-D-I, No. 10, pp. 1515-1524, 2005.
7) 北川，大西: 対面講義と e-learning(LMS + VOD) とを併用した講義形式の実践と分析, 日本教育情報学会学会誌 Vol.22 No.3 pp.57-66, 2007.
8) 金森, 竹之内, 村田: R で学ぶデータサイエンス 5 パターン認識, 共立出版, 2009.
9) 小林，椎名，北川: 字幕データを用いた VOD 教材検索システムの提案, pp416-417, 教育情報システム学会第 31 回全国大会, 2009.

Vol.2011-DBS-153 No.8
2011/11/3

10) 小林，椎名，北川: 字幕データ付き VOD 講義の単語頻度に対する混合正規分布モデルによる映像区間の推定, pp.306-307, 日本教育情報学会第 26 回年会, 2010.
11) Vladimir N. Vapnik, "Statistical Learning Theory", John Wiley & Sons, 1998.
12) 小山，小林，椎名，北川, 字幕データ付き VOD 講義の単語頻度に対するカーネル密度推定による映像区間の推定, pp.267-268, 教育情報システム学会第 32 回年会, 2010.
13) 徳弘: " 日本語学習のためのよく使う順漢字 2100 ", 三省堂, 2008.
14) 豊田, データマイニング入門, 東京図書, 2008.
15) E.L. Allwein, R.E. Schapire and Y. Singer. Reducing Multiclass to Binary: A Unifying Approach for Margin Classifiers, Journal of Machine Learning Research, 2000.
16) F. Aiolli, A. Sperduti Multiclass Classification with Multi-Prototype Support Vector Machines, Journal of Machine Learning Research,2005 pp.817-850.
17) J. Weston, C. Watkins. Multi-class Support Vector Machines, Technical Report CSD-TR-98-04,1998.
18) M. Gonen, A. Gonul and E. Alpaydm. Multiclass Posterior Probability Support Vector Machines, Proc. of IEEE Transaction on Neural Networks,2008.
19) J.Shawe-Taylor, N.Cristianini:カーネル法によるパターン解析, 共立出版, 2010.
20) Kobayashi, N., Koyama, N., Shiina, H., Kitagawa, F., "Estimation of movie segments by Gaussian mixture models on VOD lecture with Japanese Subtitle", Proceeding of Pacling 2011, #47, pp1-4, 2011.
21) Cabocha, http://chasen.org/~taku/software/cabocha/
22) 日本語能力試験公式ウェブサイト:http://www.jlp t.jp/
23) 日本国際教育支援協会:http://www.jees.or.jp/
24) 国際交流基金国際交流基金: http://www.jpf.go.jp/j/
25) EDR 電子化辞書: http://www2.nict.go.jp/r/r312/EDR/J_index.html
26) Google: http://www.google.co.jp/
27) ウィキペディア: http://ja.wikipedia.org/wiki/メインページ