単語難易度推定による中日単語学習システム

藏　培　慶[1]　　小林 伸行[2]　　椎名 広光[3]

[1] 岡山理科大学大学院 総合情報研究科 情報科学専攻

[2] 山陽学園大学 総合人間学部 生活心理学科

[3] 岡山理科大学 総合情報学部 情報科学科

zangpq@gmail.com[1]，koba_nob@sguc.ac.jp[2]，shiina@mis.ous.ac.jp[3]

1　はじめに

現在，日本の大学に多くの留学生が在学している．日本語を母語としない留学生にとって，日本語の講義を受けるには，日本語の学習を行うことが前提になっている．実際，多くの留学生は短い期間の日本語学習，日本語の理解は不十分なまま留学してきている．そのため講義で使われるすべての単語を理解できていないと考えられる．そこで講義前や後に，講義で使用された日本語単語の確認や学習するシステムが必要と考えられる．一方，外国語の語彙学習には，すでに数多くのソフトウェアが開発されている．それらのうち語彙選択問題のアプリケーションでは，そのほとんどは外国語に対する母国語の選択問題によって構成されている．また，外国語が対応の母国語及び発音によって構成されている．しかし，両方を連動作させているアプリケーションは少ない．そこで，本研究では，日本語の語彙学習と中国語の語彙学習のシステムを連動させる学習システムを開発した．

2　語彙学習システムの概要

語彙学習システム開発に Java 言語をデータベースに SQLite を用いて Android Tablet 端末に実装した．このシステムは，学習モードでは，音声による学習モードと語彙選択学習モードと混合クイズ学習モードを用意している．テストモードでは，通常テストモードと混合テストモードを用意している．システムの特長的な項目として，学習項目の設定や選択問題の誤選択で利用する単語の難易度の推定を行った．学習の終わりに，テスト及び解答時間を測り，また，再度誤選択した単語の学習及び学習効果にコメントすることができるようにしている．

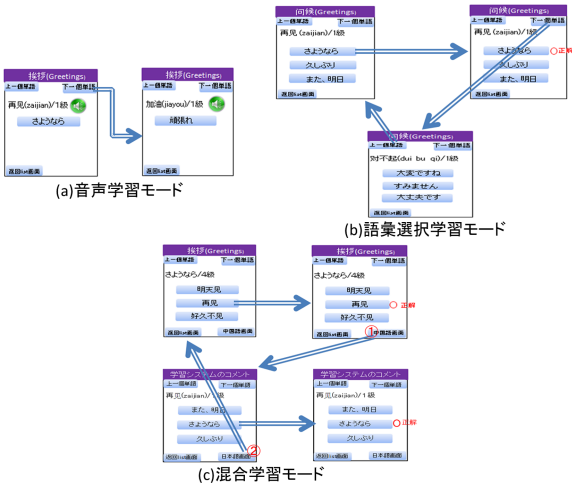

図 1: 語彙学習システム

2.1　学習モードについて

語彙学習システムでは，単語を学習，語彙選択で学習，語彙選択でご選択した単語の再学習が行えるようになっている．学習モードとしては，つぎの 3 種類を用意している．

(1) 音声による学習モード (図 1(a))：各言語の単語の発音を続けて学習する．このモードでは，ユーザーが単語リストから学習したい単語を選択して単語の発音を聞く．また，この単語に対応する日本語 (あるいは中国語) 及び難易度を表示する．

(2) 語彙選択学習モード (図 1(b))：言語ごとに連続して選択問題で学習する．学習時に，誤選択した単語は再度学習ができる．

(3) 混合クイズ学習モード (図 1(c)):選択問題で中国語と日本語を混合して出題する．手順としては，正解した選択肢を問題として，反対側の言語の単語を学習する．図 3 の例では日本語単語「さようなら」を学習する場合，最初にこの単語の学習画面では三つ選択肢から正しい選択肢と考えるものを選択する．選択

後， 結果画面に移る．また， 結果画面では選択したものの正誤を表示する．もし結果画面（あるいは学習画面）では「中国語画面」ボタン ① の選択後， 日本語単語「さようなら」が対応の中国語単語「再見」の学習画面に移る．同様に， 中国語単語「再見」の学習画面（あるいは結果画面）では「日本語画面」ボタンの選択後， 日本語単語「さようなら」の学習画面に移る．

2.2 テストモードについて

単語の学習終了後， 学習効果を測るため， 語彙テストを行うことができる．テストモードでは， 学習した単語のテスト及びテストにかかる時間を計算することができ， 次の 2 種類のモードを用意している．

(1) 通常テストモード：同じ言語， 同じカテゴリーの単語が一緒にテストを行う．テストの終わりに， 正解率と時間を表示する．

(2) 混合テストモード：言語を区別しないで， ただし，単語のカテゴリーが同じであれば， 一緒にテストを行う．また (1) と同様に正解と時間を表示する．

3 単語難易度の推定

3.1 単語難易度推定方法

学習単語の難易度については， 中国語については中国語検定 (HSK[1]) の試験難易度を，日本語については日本語能力試験 (JPLT)[2] の試験区別を利用した．単語の難易度については，あらかじめ試験区別の難易度が判明している単語から， 単語から判明していない単語の難易度を機械学習のサポートベクタマシン（SVM[6]）を利用して推定し， 推定された単語を再び利用して難易度を利用するブートストラップ法を利用している．単語の難易度推定の過程を以下に示す．

Step1:単語の難易度が判明している初期データの辞書また意味記述文を取得する．

Step2: SVM よる繰り返し学習

Step2-1:見出し語の難易度が初期データにない場合は，意味記述中の単語の難易度分布を用いて難易度を推定する．

Step2-2:意味記述中の単語の難易度がすべて決定していない場合は， 難易度が判明している単語から頻度分布を作り， 見出し語の難易度を推定する．

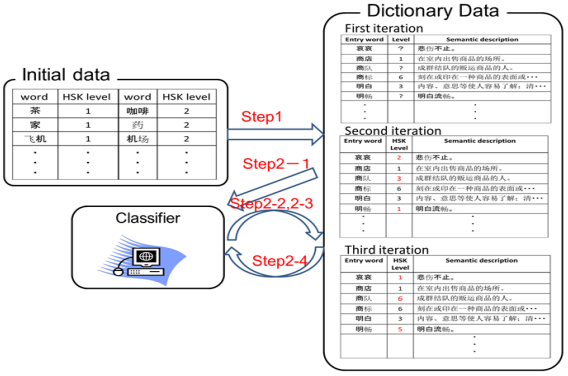

図 2: 難易度推定過程

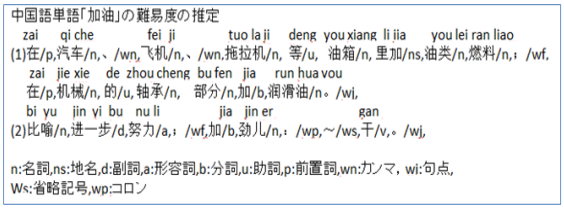

図 3: 詞素解析例

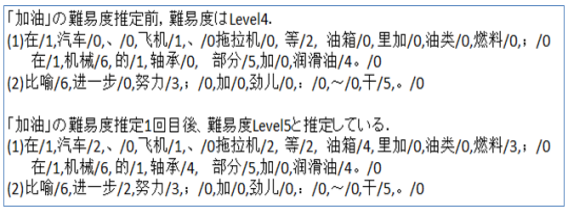

図 4: 中国語難易度推定過程

Step2-3:意味記述中の単語の難易度を更新して， 初期学習データの見出し語の難易度を組み合わせたパラメータを用いて SVM による分類学習を実行する．

Step2-4:Step2-3 を繰り返す．辞書の難易度推定結果の変更が収束すれば終了する．

3.2 中国語単語難易度の推定

中国検定 (HSK) の難易度は， Level1 から Level6 の 6 段階に分かれており， Level1 が最も易しく，Level6 が最も難しい．推定では， 意味記述を品詞に分解するが， 例えば「加油」(jia you) の詞素解析結果を図 3 に， 難易度推定の過程を図 4 に示す．図 4 では，あらかじめ難易度が判明している単語にのみ難易度がつけられている難易度推定 0 回目で，難易度分布 (4/11,1/11,1/11,1/11,2/11,2/11) から SVM を利用して難易度 Level4 と推定している．0 回目の状態から辞書 [3] の見出し語の難易度を推定し，その難易度を利用して単語の難易度をつけたのが難易度推定 1 回目

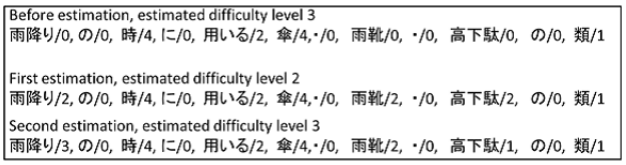

図 5: 日本語難易度推定過程

で, 難易度分布 (4/11,4/11,2/11, 3/11,2/11,2/11) から SVM を利用して難易度 Level5 と推定している． なお， 図 4 は難易度が判明しない単語や記号については， Level を 0 としている．難易度の推定過程の意味記述に現れる「汽」(qi che) は教師データには含まれていないため， 1 回目の難易度推定では Level2 として， 「加油」の推定に使用される．

3.3 日本語単語難易度の推定

日本語の難易度については，日本語能力試験 (JLPT) の旧試験の区分を難易度として利用している．難易度については， Level4 が最も易しく， Level 1 が最も難しい.「雨具」(amagu) の意味記述を用いた難易度推定の過程を図 5 に示す．図 5 では， あらかじめ難易度が判明している単語にのみ難易度がつけられている難易度推定 0 回目で, 難易度分布 (1/4,1/4,0/4,2/4) から SVM を利用して難易度 Level3 と推定している．0 回目の状態から辞書の見出し語の難易度を推定し，その難易度を利用して単語の難易度をつけたのが難易度推定 1 回目で, 難易度分布 (1/7,4/7,0/7,2/7) から SVM を利用して難易度 Level2 と推定している． 難易度推定 2 回目は， 1 回目の難易度を利用して見出し語の難易度を推定したものを再度割り付けており，難易度分布 (1/7,3/7,1/7,2/7) から SVM を利用して難易度 Level3 と推定している． なお， 図 5 は難易度が判明しない単語や記号については， 0 としている．また， 意味記述に現れる「雨靴」(amagutsu) は教師データには含まれていないため，0 回目の難易度推定では Levlel0,1 回目の難易度推定では Level2, 2 回目の難易度推定でも Level2 と推定して， 「雨靴」の推定に使用される．

4 学習単語の選出

4.1 語彙カテゴリーの選択

語彙カテゴリーの選択は， New Standard Japanese for Sino-Japan Communication Primary[5] から抽出し， 日本で生活するうえでの必要と思われる順に並べ替えている．学習カテゴリーは (1) 挨拶 (2) 交通機関 (3) 食べ物 (4) 飲み物 (5) 気象 (6) 職業 (7) 色 (8) 週 (9) 数詞 (10) 衣服 (11) 人物 (12) 岡山理科大学のデータベースの講義の 12 に分けられている．

中国語					
Category	Problem Word	HSK Level	Correct Word	Incorrect Word1	Incorrect Word2
Greetings 问候	你好 Hello	1	こんにちは /4	おはよう /4	こんばんは /4
	再见 Goodbye	1	さようなら /4	ひさしぶり /4	また明日 /4
	早上好 Good morning	1	おはよう /4	こんばんは /4	こんにちは /4
	明天见 See you Tomorrow	1	また明日 /4	お世話になりました /3	ひさしぶり /4
Transportation 交通工具	火车 Bus	1	列車 /2	トラック /2	地下鉄 /4
	公共汽车 Bus	2	バス /1	地下鉄 /4	タクシー /2
	地铁 Subway	3	地下鉄 /4	列車 /2	トラック /2
	出租汽车 Taxi	6	タクシー /2	バス /1	列車 /2
Food 食物	曲奇 Cookies	2	クッキー /4	パン /4	ビスケット /1
	面包 Bread	3	パン /4	ビスケット /1	クッキー /4
	蛋糕 Cake	3	ケーキ /2	クッキー /2	ゆで卵 /1
	饼干 Biscuit	4	ビスケット /1	パン /4	ベーコン /2

図 6: 中国語単語リストと選択単語例

4.2 学習単語と誤選択単語の選出

出題する単語については， 前節の語彙カテゴリーのなかから各言語の難易度を利用して， 中国語の場合は HSK の試験区分の易しい順に， 日本語の場合はＪＬＰＴの易しい順としている．また， 正解でない誤選択の単語選出手順は， 次の通りである．

(1) 同一カテゴリーの中から誤選択の単語を決める．

(2) 誤選択単語は， 正解の単語の試験区分に近いものを選択する．

ただし， 実際の語彙学習システムの開発では， 誤選択単語の選出は， コンピュータで自動選出したもの中から人手チェックして選んでいる．単語の意味を考慮していないためである．図 6 と図 7 に， 3 カテゴリーと出題単語の 4 つの例と選択子の単語を示す．また， 中国語の単語の後ろには HSK の試験区分を付

日本語					
Category	Problem Word	JLPT Level	Correct Word	Incorrect Word1	Incorrect Word2
Greetings 挨拶	こんにちは	4	你好 /1	早上好 /1	晚上好 /1
	おはよう	4	早上好 /1	晚上好 /1	你好 /1
	さようなら	4	再见 /1	明天见 /1	好久不见了 /1
	また明日	4	明天见 /1	好久不见了 /1	不用谢 /1
Transportation 交通機関	地下鉄	4	地铁 /3	无轨电车 /2	直达快车 /2
	列車	2	火车 /1	无轨电车 /2	特别快车 /2
	タクシー	2	出租汽车 /6	地下鉄 /4	バス /1
	バス	1	公共汽车 /2	火车 /1	地铁 /3
Food 食べ物	パン	4	面包 /3	饼干 /4	蛋糕 /4
	クッキー	4	曲奇 /2	蛋糕 /4	蜂蜜 /2
	ケーキ	2	蛋糕 /3	曲奇 /2	面包 /3
	ビスケット	1	饼干 /4	面包 /3	蛋糕 /4

図 7: 日本語単語リストと選択単語例

し， 日本語単語の後ろに JLPT の試験区分を付している．

5 学習効果の評価

学習モードの語彙選択問題には， 中国語と日本語をそれぞれ順に学習する順学習モードと， 混合して学習する混合モードがある．各方式による学習効果について， 被験者による評価実験を行った． 評価実験の実施を 5 回行い， 1,2,3 回目と 4, 5 日は連続した日にちで，3 回目と 4 回目の間には 2 カ月の期間を置いた．各モードごとの平均正解率と回答に使用した時間にかかった平均時間 (平均回答時間) を図 8 と図 9 に示す．平均正解率では、順学習モードは初めから高い正解率を持っているのに対して、混合モードは当初の正解率は 70%程度であるが評価回数ごとに正解率が上がっている。また、平均回答時間においては、混合モードの操作性のため時間はかかっているが、評価回数が多くなるについて順学習モードと同じ時間で回答が済むようになっている。

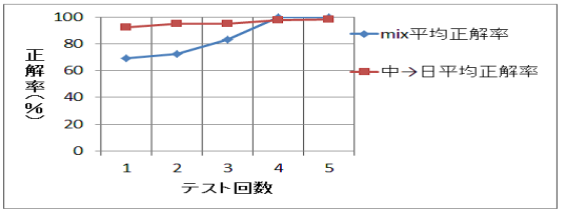

図 8: 正解率の変化

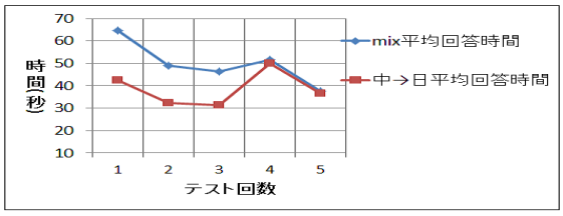

図 9: 学習時間の変化

6 まとめ

本研究において， 単語難易度の推定する場合， 推定した単語の難易度のレベルの誤差が存在する．できるだけ誤差を縮小するように， もっと高精度な難易度の推定方法の使用は必要である．また， 現在まで学習単語の数は 230 個程度であるので， 単語の数の追加も必要である．

参考文献

[1] Officical site for Chinese Language Test, http://www.chinesetest.cn/index.do

[2] Japanese-Language Proficiency Test, http://www.jlpt.jp

[3] Hanyu Da Cidian, Publishing House of the Chinese Dictionary, 1998.

[4] Y. Tokuhiro, Kanji2100 Listed according to Frequency and Familiarity, Sanseido, 2008.(in Japanese)

[5] New Standard Japanese for Sino-Japan Communication Primary, Peoples Education Press,2005.

[6] V.Vapnik, Statiscal Learning Theory, Springer, 1988.