

# 小学2年生の問題を解く - 電腦優子2年生国語・デモ

齋藤真実\*1 岡田美江\*1 関根聡\*2 井佐原均\*3

\*1長岡技術科学大学電気系

\*2株式会社ランゲージ・クラフト研究所 / ニューヨーク大学

\*3独立行政法人情報通信研究機構

## 1. はじめに

小学校2年生の国語の問題を解くシステム『電腦優子2年生国語』を作成している。本稿では、実際の問題の解き方を紹介し、評価結果について述べる。本年度大会の口頭発表で発表する別の稿(関根ら 2005)で、目的と概要、関連研究について述べ、使用した知識や解決できなかった問題に対して考察を行っている。

## 2. 概要

我々は問題を分析するために、小学2年生の国語の問題集を5冊購入し、問題を大まかに分類してみたところ「漢字」「言葉」「読解」「作文」の4種類に分類できた。さらに細かく分類した各問題に対して問題を解くプログラムを実装した。「作文」については本システムでは対応していない。本稿では、これらの問題に対して第3~5章の中で、それぞれの解き方とその評価結果について述べている。また、第6章ではシステム全体をまとめた評価を行っている。

表1~3には、「漢字」「言葉」「読解」のそれぞれの問題における、訓練データ中の割合と訓練データとテストデータでのテスト結果をまとめた。訓練データとは、システムを作るうえで参考にした1冊の問題集であり、評価用として、それとは別にテストデータを用意した。詳しくは本稿の第6章で述べる。また、表1と2には主に問題を解く上で重要となる知識やツールを示し、表3には使用した手法をまとめている。使用した知識やツールに関する詳細は別の稿(関根ら 2005)で述べている。

## 3. 漢字の問題

漢字の問題には大きく分けて「読み」の問題と「書き」の問題がある。「読み」の問題では形態素解析を利用し、「書き」の問題では大規模コーパスが、辞書から得た解答候補の中から答えを選ぶのに役に立った。詳しくは表1に示すとおりであるが、書き順を答える問題と漢字の部分を組み合わせる漢字を作る問題は電子化が難しいため、本システムでは対応していない。

実際に本システムで対応できる問題のうち、正解できなかったものは334問中7問であり、その内5問は形態素解析ミスによる「読み」の問題の誤りであった。

例1. 一分間。

形態素解析: いちぶんかん / 正解: いっぱんかん

特に「読み」の問題では、形態素解析以外に有効な知識を見つけられなかった。「書き」の問題では有効であったコーパスからは漢字の読み方が得られないので、「読み」の問題については何か別の知識を探す必要がある。

また、システムが正解できなかった残りの問題は「書き」の問題で、1問は同音異義語の問題(「弟」or「義弟」)、最後の1問はコーパス中に存在しなかったために誤った(「おじに あう。」は「会う」か「合う」か?)ものである。

表1. 漢字の問題の種類

問題の種類	訓練データ		訓練データ	テストデータ
	における	使用した	正解数	正解数(対象の
	問題の割合[%]	知識・ツール	(問題数)	問題数、全体の問題数)
読み	27	漢字辞書、形態素解析	96(100)	6(8, 8)
書き	61	単語辞書、 大規模コーパス	220(222)	63(66, 66)
書き順	6	-	0(20)	0(0, 2)
漢字の部分の 組み合わせ	3	-	0(10)	0(0, 0)
漢字の分類	3	単語辞書、シソーラス	11(12)	0(0, 0)
合計	100	-	327(364)	69(74, 76)

## 4. 言葉の問題

漢字の問題でも読解問題でもなく、語彙や文の構造に関する問題などを、ここでは「言葉の問題」と呼んでいる。表2に示す主な問題について簡単に説明する。

反対語を答える問題では、反義語辞書から複数候補を出し、大規模コーパスを用いて問題文に適切なものを選び、かなり解けている。

なかまの言葉の問題には、上位下位の関係を求める問題や違うなかまの言葉を選ぶ問題などの種類があるが、これらはシソーラスを利用することで大部分解ける。

助詞の問題は文章に合う助詞を選択するもので、例2に示すような問題である。

例2. [を/と/に]のうち、に合うことばを書きなさい。

(1) りんご みかん 買う。

出力: を を 正解: と を

選択肢が与えられているので、考えられる組み合わせから文を完成させ、大規模コーパスから尤もらしいものを選んだ。例2の問題では、どの組み合わせについてもコーパスに存在しなかったため、のある文節と動詞の文節を選び、「りんご 買う。」「みかん 買う。」などとしてコーパスで探した。そのため「りんごを 買う。」が最も多くコーパス中に存在したため不正解となった。

また、カタカナの問題はひらがなの言葉をカタカナに直すものであるが、「ー」は「あいうえお」で書かれているため、辞書を利用して正解を探した。

例3 . つぎの ことばを、かたかなで 書きなさい。  
 (1) てえぶる 正解：テーブル

表2を見ると、カタカナの問題の正解数が少ないが、これは、選択肢の言葉を分類してカタカナで書く問題がこの種に含まれているからである。例4の問題で、「外国から来たことば」などはシソーラスを用いても判断できないため、この種の問題は解くことができなかった。これらの問題は、カタカナの問題37問中16問存在した。

例4 . (1) (2)に かたかなで 書きなさい。  
 選択肢：ぼちゃん・のうと・がさがさ・ぶらし  
 (1) ものの音 (2) 外国から来たことば

ようすを表す言葉の問題には、例5に示すような問題や文章中に適切な言葉を選んで入れる問題などがある。例5の問題は問題文と選択肢のコーパス中での共起を使って解いた。(1)は正解したが(2)では「ゆっくり」と「のろのろ」の共起頻度が高く、正解の「ごろごろ」は得られなかった。

例5 . つぎの ようすを あらわす ことばを[ ]から えらんで、 で かこみなさい。  
 (1) 元気で そだつ ようす。  
 [ずんずん・すらすら・すくすく]  
 (2) 大きな ものが ゆっくり ころがる ようす。  
 [ころころ・ごろごろ・のろのろ]

また、言葉の問題は今回訓練データとして使用した問題集だけで、細かく分けて29種類あった。本システムでは、それぞれに対応したサブシステムを作成したが、他の問題集も含めると問題の種類は膨大であり、それらを逐一作成したのではコストがかかる。この問題をどのようにしてまとめて解くか、今後の課題である。

## 5 . 読解問題 (散文)

読解問題には説明文や物語文の読み取り問題である「散文」と詩の問題である「韻文」の二種類があるが、本システムでは散文のみを対象としている。表3で分類している各種の問題に対して利用した、いくつかの手法の中で、代表的な5つの手法について説明する。

表2 . 言葉の問題の種類

問題の種類	訓練データ		訓練データ	テストデータ
	における	使用した	正解数	正解数(対象の
	問題の割合[%]	知識・ツール	(問題数)	問題数, 全体の問題数)
反対の言葉	18	反義語辞書、 大規模コーパス	26(27)	12(15, 21)
なかまの言葉	11	シソーラス	14(17)	34(44, 83)
助詞の問題	19	大規模コーパス	25(28)	16(17, 17)
カタカナ	25	単語辞書、 形態素解析	18(37)	19(22, 52)
ようすを表す 言葉	19	大規模コーパス、 形態素解析	18(29)	16(20, 31)
文の構造	5	形態素解析	7(7)	20(22, 22)
仮名使い	2	-	0(3)	0(0, 19)
動詞の記述	2	-	0(3)	0(0, 0)
合計	100	-	108(151)	117(140, 245)

### a . パターンマッチング

特に穴埋め問題で利用し、コンテキストを使ったパターンマッチングで本文中から答えを探す。完全に一致しなかった場合はマッチングに利用する文字列を減らして繰り返した。( )ひとつに対して答えは一文節とした。

例6 . 穴埋め問題  
 本文(一部):「二、三日 たつと、その 花は しぼんで、だんだん 黒っぽい 色に かわって いきます。」  
 穴埋め文:「花は(1)(2)色に かわる。」  
 正解:(1)しぼんで (2)黒っぽい

本訓練データではこの手法は有効にはたらいだ。問題が高度になるにつれて、他の手法との連携が必要になってくるが、読解問題を解く上で重要な手法である。

### b . NE や直後にくる助詞の種類

「いつ」「どこ」のように、問題によって答えのNEタイプが限定できる場合には、本文中から時間表現、場所表現などを候補とする。また、問題文が「いつに」であれば、「時間表現+に」であるものを最有力候補とする。

例7 . いつのことを作文に書いていますか。  
 本文(一部):「日曜日に 家ぞくで 川口こに きました。みずうみでは、(後略)」  
 正解：日曜日

物語全体について聞いている「いつ」「どこ」の問題に関しては、それぞれのNEタイプの語の中で、本文中に最初に出現するものを答えている。形態素解析やNE抽出が上手くいかずに不正解となった問題があるが、この手法そのものは概ね有効である。

### c. 手がかり語を用いた部分マッチ

「何の A」や「どれくらいの B」のように、問題文中から A、B に相当するものを手がかり語として抽出し、それをもとに本文中から答えを探す。

例 8 . かえったばかりの ひなは、どれくらいの 大きさですか。  
 本文(一部):「かえったばかりの ひなは、子どもの ゆびの 先ぐらいの 大きさです。」  
 正解:子どもの ゆびの 先ぐらい

また、「なぜ・どうして」の理由を聞く問題では、問題文中から手がかり語を抽出するのではなく、独自に決めた手がかり語「それで」「だから」等を利用している。このように、手がかり語を用いた部分マッチには 2 種類あるが、どちらも本システムで有効にはたらいた。

### d. コーパス中での頻度の計算

接続詞の問題でのみ使用している。選択肢にある接続詞と直前の文の文末の一文節との共起の相互情報量をコーパスから求めて、それを比較している。

例 9 . 接続詞の選択問題  
 本文(一部):「(前略)たくさんの えいようを おくっているのです。 、花は、すっかり かれて、(後略)」  
 選択肢:ア.まず イ.やがて ウ.それとも  
 正解:イ

しかし、この手法はあまり有効ではなく、より広い素性を使った談話解析が必要である。

### e. 問題文中のキーワードの利用

この手法は主に a - d の手法によって、解答候補が複数見つかったときに利用している。本文中での各キーワードとの距離(形態素数の数)の総和が最も小さいものを解答とする。ここでいうキーワードは、問題文に含まれる内容語と未定義語であり、名詞の連続はひとつにまとめ、接辞も含めている。この手法は概ね有効であった。

表 3 の使用した手法の欄には、システムの中で優先度の高い順に示している。上述の 5 つの手法以外を用いて解いたものは f としている。このようにして問題を解いてみたが、このような手法では解けない問題もいくつか存在した。その中でも特徴的なものについて紹介する。

例 10 . 子だぬきは どうして「まきをせおって」いたのですか。  
 (1) を買う (2) の代わりにするため。  
 本文(一部):「すまんがたこを一つ売ってはくれんかな。 お金がないので、代わりにまきを集めてきたんだけど。」

穴埋め問題は本文と穴埋め文の間のパターンマッチングで解いているが、このように同じ形式で書かれていないものは正解を取り出せない。

例 11 . うめの花が、びっくりしたのは、 どうしてですか。  
 本文(一部):「あら。今ごろはまだ、みんないっしょにねむっているんじゃないの。」  
 正解:てんとう虫は、まだねむっているはずだから。

この問題ではうめの花がてんとう虫に対して言ったセリフから推測しなくてはならないが、これは非常に難しい問題である。しかし、小学 2 年生レベルの問題でも、このように登場人物のセリフから答えを推測する問題はいくつか見られた。

表 3 . 読解問題の種類

問題の種類	訓練データ		訓練データ	テストデータ	
	における	使用した		正解数	
	問題の割合[%]	手法	正解数(問題数)	(対象の問題数、全体の問題数)	
だれ	セリフを言った人物	5	b,a,f	5(6)	1(4, 4)
	それ以外	0	b,e,f		
何	何のように	22	c,b,e	17(26)	1(1, 6)
	何の		c,f,e		
	何を		c,a,f		
	何(が)を(に)		a,e		
	何と(言い)答え		b,a,f,e		
いつ	物語全体	4	b	3(5)	0(0, 0)
	物語の一部		-		
どこ	物語全体	4	b	3(5)	0(1, 1)
	物語の一部		b,f,c		
なぜ・どうして・なんのため		16	c,f	11(18)	0(1, 1)
どう		9	c	8(10)	0(0, 0)
どんな		0	-	0(0)	0(0, 1)
どのように		1	-	0(1)	0(0, 0)
どれくらい+形容名詞		2	b,c,f	1(2)	0(0, 0)
穴埋め		10	a	10(12)	4(9, 9)
疑問詞なし(気持ち・ようす)		4	-	0(5)	0(0, 0)
接続詞		2	d	1(2)	1(3, 3)
ストーリー経過順		10	f	8(11)	0(0, 0)
パラグラフ		10	f	7(12)	3(3, 3)
その他		1	-	0(1)	0(0, 6)
合計		100	-	74(116)	10(22, 34)

## 6. 評価

訓練データは公文出版の「学力チェックテスト国語 2 年生」を、評価用データは、同じ公文出版の「くもんの小学ドリル」シリーズの「漢字」「ことばと文章」「長文読解」の問題集を用いた。「漢字」と「長文読解」では、問題集の最後にあるまとめの問題を使用した。また「ことばと文章」ではまとめの問題には文章問題も含まれており、

問題集の全体の中から言葉の問題のみを取り上げた。したがってテストデータでは、言葉の問題が他と比べて多くなっている。

表1～3の右端の欄にテストデータでの各分類別の結果を記してあるが、それらをまとめたものを表4に示す。ただし、テストデータ中に訓練データには見られない構造をした問題が存在したため、実際にシステムで解くことのできるタイプのみを解いている。各表には実際に解いた問題の数(対象の問題数)とシステムは対応していないが、我々が分類したテストデータに含まれるすべての問題数(全体の問題数)を記してある。全体の問題に対する対象の問題の割合は、漢字で97.4%、言葉で57.1%、読解問題で64.7%であった。この値を見てみると、特に言葉の問題については半分近くの問題に対応できていないことがわかる。そのため、対象の問題での正解率は83.6%と高くても、全体としては47.8%という結果になっている。この原因は問題を細分化し、特定の問題形式に従ったサブシステムを多数作ることで対応しようとしたためであると考えられる。問題形式にあまりとらわれない柔軟な対応が必要である。また、読解問題で、我々のシステムが対応していなかったものには表の穴埋め問題などがあつた。

本システムが対象としている問題での正解率は全体で83.4%と高い値が得られた。特に漢字、言葉の問題では、テストデータでの正解率と訓練データでの正解率はほぼ同等で、問題が認識されれば安定して正解を出せることがわかる。しかし、読解問題の正解率は45.5%と低い。これは、読解のテストデータの量が少ないこともあるが、訓練データでは良好な成績であった「だれ」の問題と穴埋め問題が半分しか取れなかったことが大きい。特に、穴埋め問題では文節を一まとまりとして処理したことで上手くパターンにマッチしなかったことが原因となっている。

表4. テストデータでのテスト結果

	正解数	対象の	全体の	正解率	正解率	訓練データ全体の
		問題数	問題数	(対象)[%]	(全体)[%]	正解率[%]
漢字	69	74	76	93.2	90.8	89.8
言葉	117	140	245	83.6	47.8	71.5
読解	10	22	34	45.5	29.4	63.8
全体	196	235	355	83.4	55.2	80.3

### 参考文献

関根聡、齋藤真実、岡田美江、井佐原均: 2005. 「小学2年生の問題を解く - 電腦優子2年生・概要」第11回自然言語処理年次大会

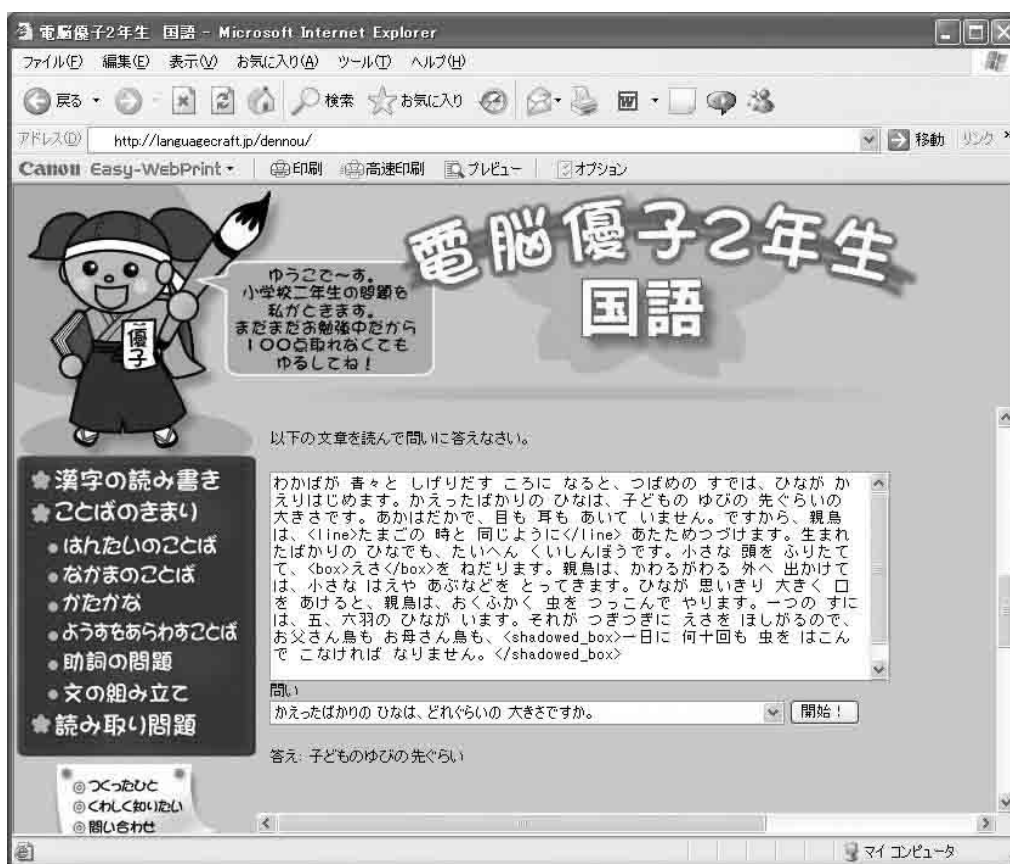


図1. 電腦優子2年生国語デモ画面