「ロボットは東大に入れるか」プロジェクト 代ゼミセンター模試タスクにおけるエラーの分析

松崎拓也[1]，横野光[2]，宮尾祐介[2]，川添愛[2]，狩野芳伸[3]，加納隼人[1]，佐藤理史[1]
東中竜一郎[4]，杉山弘晃[4]，磯崎秀樹[5]，菊井玄一郎[5]，堂坂浩二[6]，平博順[7]，南泰浩[8]
[1]名古屋大学大学院，[2]国立情報学研究所，[3]静岡大学，
[4]NTT コミュニケーション科学基礎研究所，[5]岡山県立大学，[6]秋田県立大学
[7]大阪工業大学，[8]電気通信大学

1 はじめに

「ロボットは東大に入れるか」（以下，「東ロボ」）は国立情報学研究所を中心とする長期プロジェクトであり，AI 技術の総合的ベンチマークとして大学入試試験問題に挑戦することを通じ，自然言語処理を含む種々の知的情報処理技術の再統合および新たな課題の発見と解決を目指している．プロジェクトの公式目標は 2016 年度に大学入試センター試験において高得点を挙げ，2021 年度に東大 2 次試験合格レベルに達することである．プロジェクトでは，2016 年度のセンター試験「受験」に至るまでの中間評価の一つとして，2013 年度，2014 年度の 2 回に渡り代々木ゼミナール主催の全国センター模試 (以下，代ゼミセンター模試) を用いた各科目の解答システムの評価を行い，その結果を公表した．表 1 に 2014 年度の各科目の得点と偏差値を示す[1]．2013 年度の結果については文献 [22] を参照されたい．

2014 年度の評価では，英語・国語・世界史 B で受験者平均を上回る得点を獲得するなど，大きな成果があった一方で，その得点に端的に現れているように，残された問題も大きい．本稿では，代ゼミセンター模試およびその過去問を主たる評価データとして各科目の解答システムのエラーを分析し，言語処理・知能処理の多様な側面に対する現在の NLP/AI 諸技術の達成度に関してひとつの見取り図を与えるとともに，今後の課題を明らかにする．以下では，まず知的情報処理課題としてのセンター模試タスクの概要をまとめたのち，英語，国語現代文（評論），古文，数学，物理，日本史・世界史の各科目について分析結果を述べる．

[1]数学・物理に関しては他の科目と異なる付加情報を含む入力に対する結果である．詳細はそれぞれに関する節を参照のこと．国語は，未着手の漢文を除いた現代文・古文の計 150 点に関する偏差値を示す．

科目		得点/満点	偏差値
英語		95/200	50.5
国語	現代文	49/100	54.2
	古文	20/ 50	
数学 I・数学 A		40/100	46.9
数学 II・数学 B		55/100	51.9
物理		31/100	49.0
日本史 B		44/100	48.2
世界史 B		52/100	56.1

表 1: 2014 年度 代ゼミセンター模試 (第 1 回) に対する得点と偏差値

2 センター試験タスクの概要

表 2，表 3 に，2014 年度代ゼミセンター模試 (第 1 回) の世界史 B・日本史 B・数学 (I+A，II+B の合計)・物理，国語・英語を対象とした問題分類の結果を示す．表内の各数字は，各カテゴリに分類された問題数およびその割合（カッコ内）である．ここでは，一つの問題が複数のカテゴリに属する場合も許している．これらの分類は解答タイプ（解答形式および解答内容の意味的カテゴリ）と解答に必要となる知識のタイプに関するアノテーション [13] から得られたものであるが，読みやすくするために，表中では各カテゴリにそれらのアノテーションを要約・再解釈したラベルを与えている．

表 2 に示されるように，社会科目ではほとんどの問題が教科書内の知識を正しく記憶しているかどうかを問う問題であり，形式は真偽判定型と factoid 質問型が多い．問題中で与えられた資料文に関する読解問題や一般常識の関わる問題の割合は低いことから，大多数の問題に対しては外部の知識源を適切に参照し，要

	世界史 B	日本史 B	数学	物理
真偽判定	13 (36%)	15 (42%)	0 (0%)	1 (4%)
factoid 質問	10 (28%)	7 (19%)	0 (0%)	2 (7%)
読解	0 (0%)	2 (6%)	0 (0%)	0 (0%)
教科書の暗記	36 (100%)	36 (100%)	0 (0%)	0 (0%)
一般常識	0 (0%)	0 (0%)	0 (0%)	0 (0%)
画像・図表理解	5 (14%)	5 (14%)	6 (6%)	20 (100%)
ドメイン固有推論	0 (0%)	0 (0%)	96 (100%)	29 (100%)
数式理解	0 (0%)	0 (0%)	96 (100%)	25 (89%)
問題総数	36	36	96	29

表 2: 問題分類（社会科目・理数系科目）

	国語	英語
語彙知識	10 (29%)	21 (45%)
文法知識	1 (3%)	20 (43%)
読解	14 (40%)	20 (43%)
一般常識	0 (0%)	6 (13%)
状況理解	0 (0%)	8 (17%)
修辞構造理解	4 (11%)	0 (0%)
画像理解	0 (0%)	10 (21%)
問題総数	35	47

表 3: 問題分類（国語・英語）

求される解答形式に合わせた出力へ加工することで解答できる可能性が示唆される．すなわち，現行の質問応答および検索をベースとした方法によって解ける可能性がある．他方，数学・物理に関しては，問題のほぼすべてが「ドメイン固有推論」に分類されている．すなわち，単に知識源を参照するだけでは解答できず，数理的演繹やオントロジーに基づく推論などが必要となることが示唆される．特に，数学・物理の問題のほとんどが数値ないし数式を答える問題であるため，数値計算ないし数式処理は必須である．言語処理と数値・数式処理の統合は，境界領域の研究として興味深い．数学・物理の間の違いとして，画像・図表の理解を必要とする問題の割合の差が見て取れる．数学では数表および箱ひげ図の理解を有する大問が 1 題あったが，平面幾何やベクトルの問題で与えられた図に関しては必要な情報が全て問題文で与えられており，解答する上で図を理解する必要はない．

数学は全ての問題が数式交じりの日本語テキスト中の空欄を数式や記号で埋める形式をとっている．調査対象とした 2014 年度代ゼミセンター模試（第 1 回）では，箱ひげ図および数表の理解を要する問題が一問あったが，これを含め全てが表 2 の「ドメイン固有推論」および「数式理解」に属する問題となっている．また，平面幾何およびベクトルの問題で図が計 3 つ含まれていたが，必要な情報は全て問題文に与えられており，これらの図に関しては解答する上で理解する必要はない．

英語と国語の問題分類は，他科目とは大きく異なっている．英語に関する節で述べるように，語彙知識，文法的知識を問う問題は，現在の言語処理技術の射程内のものが多数ある．しかし，英語・国語で大きな割合を占める読解問題は，これを研究課題とする取り組みが近年開始されたものの [15, 16]，言語処理・知的情報処理課題としての定式化を含め，未解決の部分が多いタイプの問題である．さらに，英語問題には一般常識を問う問題，状況理解を問う問題（旅行者と案内人の会話として適切なものを選択する問題など），画像理解（絵の説明として適切なものを選ぶ問題など）が含まれるが，これらは一部に研究課題として非常に難しいものを含んでいる．この点で，少なくとも現時点では，英語で満点に近い高得点を得ることは困難であると考えられる．

3 英語問題のエラー分析

3.1 はじめに

本稿では，東ロボプロジェクト [22] において英語チームが英語問題を解いたときのエラーを分析した結果について述べる．特に，代ゼミセンター模試の 6 回分（2012 第 1 回，2013 第 1 回〜第 4 回，2014 第 1 回）を中心に分析を行った．

今回の分析は現状一定の精度で解けている短文問題（すなわち，大問 1 から大問 3）のみについて行っている．また，短文問題の中で文脈に合わない文を選ぶという問題（3B）については未着手のため触れていない．また，意見要旨把握問題については，会話文完成問題と同じ解き方で解いているため，会話文完成問題の分析をもって，この問題の分析とする．なお，2014 年度の代ゼミセンター模試の英語問題を解いた手法については，文献 [23] に詳述されているので参照されたい．

3.2 発音・アクセント問題

ここ数年の発音・アクセント問題は発音箇所が異なる・同じ箇所やアクセント位置が異なる・同じ箇所を選択する問題であり，音声認識用の辞書を用いることですべて解くことが出来ている．しかし，1987 年から 2009 年までのセンター試験の発音アクセント問題は 28/85（約 32%）しか解くことができていない．これらの問題は，強勢の位置を問うものが多く，文脈を理解しないと解くことができない．

図 1 は強勢の問題の例である．下線部の単語のうち，強勢が置かれるものをそれぞれ選択する．(1) の下線部では，worse が正解となるが，worse を強く読むかどうかは文脈に依存する．

1999 年までの発音・アクセント問題で解けていない問題を分析したものを表 4 に示す．辞書やプログラムの整備などにより対応できるものを短期に対応できる問題（18 問），それ以外を長期間必要な問題（33 問）あった．発音・アクセント問題の出現する箇所に発音以外の問題があり，これらは 6 問あった．強勢の問題は近年コーパスベースの手法で取り組んでいる文献 [19] もあるが，まだ取り組みが少ないのが現状である．

3.3 文法・語法・語彙問題

文法・語法・語彙問題とは，文中の空欄に最もふさわしい語句を 4 つの候補の中から選ぶ問題である．代ゼミセンター模試の過去 6 回分には，この問題が合わせて 60 問出題されている．英語チームでは，単語 N-gram を用いて，最も確率が高くなる候補を選ぶ方法を用いた．本手法では，47 問解くことができた．解くことが出来なかった問題の要因は以下の通りであった．

問題 ID	要因
2012-1-A10	反実仮想
2012-1-A16	複数文
2013-1-A12	複数文
2013-2-A9	遠い依存関係
2013-3-A8	成句
2013-3-A9	遠い依存関係
2013-3-A13	反実仮想
2013-4-A10	局所的な高い確率に引きずられる
2013-4-A11	成句
2013-4-A17	関係代名詞（of which の用法）
2014-1-A11	遠い依存関係（much less)
2014-1-A13	成句 (never ... without)
2014-1-A14	局所的な高い確率に引きずられる

反実仮想のように，条件文に呼応する場合はそれを踏まえる必要があるが，N-gram ではそれが捉えられていなかった．また，複数文で前半部分を受けて後半の単語を選ぶ問題についても同様に答えられていない．遠い依存関係は N-gram によって捉えにくいものであるが，今回の代々木模試 2014-1 の 2 問については，Dependency Language Model [3] に基づく手法で答えられることを確認した．成句については受験においてよく出現するものであり，新聞記事の出力分布からずれるために答えられていないと思われる．関係代名詞の用法については，解くためには文法的な観点が必要と思われる．

今回の分析対象である 60 問で人間（受験生）とシステムの正答傾向に違いがあるかを分析した．ここで，人間の回答として，受験生の選択した割合が最も高かったものを用いている．なお，人間は 48 問（80%）正解している．クロス表を作成したところ以下の様になった．

	システムが正解	システムが不正解	合計
人間が正解	37	11	48
人間が不正解	10	2	12
合計	47	13	60

システムと人間の両方が解けるものはある程度共通しているものの，それぞれ得意・不得意があることも分かる．システムが正解することと，人間が正解することが独立であるか，本クロス表についてカイ二乗検定を行ったところ $\chi^2 = 0.07$（$p = 0.94$）であり有意差は見られなかった．人間とシステムの解き方は独立であることが示唆される．

60 問の各設問について人間とシステムの選択肢の順序を求め，代表的な順位相関係数である Spearman の ρ および Kendall の τ の平均値を求めた．ここで，人間の選択肢の順位とは選択した割合によるものであり，システムの順序とは，N-gram 確率によって得ら

Maya : Here comes our train. It's not too crowded.
Jeff : Do the trains (1)get any worse than this?
Maya : Oh, yes. During the morning rush hour (2)they're twice as bad.
Jeff : I can't imagine a train being more crowded than this. Where I'm from, (3)we can always get a seat.
Maya : You were lucky, but you'll have to get used to the crowds here. How do you get to school? Do you take a train?
Jeff : No, (4)I walk to school.

図 1: 強勢問題の例

	分類	1987	1991	1993	1995	1997	1999	2001	2003	2005	2007	2009	集計
		5/7	0/7	8/10	2/8	2/9	1/9	2/8	0/6	0/6	3/8	5/7	
短期に対応できる問題	問題分類の誤り		5										5
	辞書にない可能性						1						1
	文章中の単語の音素の発音（組み合わせ選択）				2		2						4
	文中の単語中音素の強勢（組み合わせ選択）					2			2	2			6
	強勢（組み合わせ選択）										2		2
長期間必要な問題	文章（文）中の語の強勢（単語選択）	2	2		2	2	2	4	4	4			22
	文章中の語の強勢（仲間はずれ選択）			2				2					4
	文中の単語中音素の強勢（組み合わせ選択）				2								2
	強勢による意図の選択										3	2	5
発音・強勢以外の問題（対話）						3	3						6

表 4: 発音・アクセント問題の分類

れる確率値の大きい順に並べたものである．その結果，ρ と τ の平均はそれぞれ 0.07，0.06 となり，ほぼ無相関であった．ここからも，人間とシステムは異なった解き方で問題を解いていることが示唆される．

システムが正解し人間が不正解であった問題は 10 問であり，この内訳は以下の通りである．人間は英語の典型的用法を知らないことで不正解になっているケースがほとんどであった．これらはシステムがデータにより正解できるものである．また前置詞の用法も英語に慣れていないと難しく，受験生には解けなかったようである．

問題 ID	問題内容
2012-1-A11	英語的用法 (in demand)
2012-1-A12	前置詞選択
2012-1-A13	難しい単語 (equipped)
2012-1-A17	難しい文法 (little から始まる構文)
2013-1-A15	英語的用法（difficulty finding)
2013-2-A8	英語的用法（errors/mistakes の使い分け）
2013-2-A11	前置詞選択 (of the station)
2013-2-A12	英語的用法 (not working)
2013-3-A10	英語的用法 (for the time being)
2013-3-A11	英語的用法 (to my taste)

人間が正解しシステムが不正解であった問題は 11 問および，どちらも不正解だった 2 問 (2013-1-A12, 2014-1-A14) は，システムが解けなかった問題として前掲した 13 問である．これらは，ほとんど人間（受験生）であれば間違えることのない問題のようである．

会話の流れの自然さの推定方法	アノテーション	
	無し	有り
発話意図と感情極性	8/18	6/18
発話意図のみ	7/18	7/18
感情極性のみ	8/18	6/18

表 5: アノテーションの有無による会話文完成問題の正解率の変化

3.4 会話文完成問題

会話文完成問題は，二人の話者の会話の空所に適切な文を 4 つの選択肢から選び，会話文を完成させる問題である．この問題を解くため，4 つの選択肢の各場合について会話文の流れの自然さを推定し，最も自然な流れとなる選択肢を選ぶという方法を用いた．会話文の流れの自然さは (a) 発話意図（表明，評価など）の流れの自然さと (b) 感情極性（ポジティブかネガティブ）の流れの自然さから成る．(a) は Switchboard Dialog Act Corpus[7] から発話意図列の識別モデルを CRF によって学習し，発話意図列の生起確率に基づいてスコアを計算した．(b) は感情極性コーパス [14] から SVM により識別モデルを学習し，感情極性がポジティブあるいはネガティブである確率に基づいてスコアを計算した．それぞれのスコアの重み付き和を最終的なスコアとした．

エラー分析のため，代ゼミセンター 6 回分の問題について，会話中のすべての発話および選択肢に対し，1 名の評価者がアノテーションを行い，発話意図のラベルと感情極性の度合を付与した．アノテーションに基づき，(a), (b) のスコアを計算した．(a) は付与された発話意図列の N-gram 確率をコーパスから計算したものをスコアとした．(b) は付与された感情極性の度合に基づいてスコアを計算した．コーパスから学習したモデルに基づいてスコアを算出する場合（アノテーション無し）とアノテーションに基づいてスコアを算出する場合（アノテーション有り）を比較し，正解率がどう変わるかを検証した．その結果を表 5 に示す．

表において，発話意図のスコアと感情極性のスコアの両方を使う場合は，正解率が最大となるように重みを調整した．表から分かるように，感情極性に関して，アノテーション無しの方がアノテーション有りの場合よりも正解率が若干高い．アノテーション無しの場合は，感情極性コーパスを使うことにより，ポジティブ／ネガティブな文に現れる単語の出現確率を考慮してスコアを計算していることに対して，アノテーション有りの場合は，そのような単語の出現確率を精密に考慮できないことが性能低下につながった可能性がある．本質的にアノテーション無しの方が性能が良いかどうかはより多くのデータを使って判断することが必要である．

本手法は発話意図のスコアと感情極性のスコアの重み付き和で最終的なスコアを計算しているが，どちらのスコアを優先すべきかは問題による．実際，発話意図のスコアと感情極性のスコアのいずれかが最大となる選択肢を選ぶことができたすると，アノテーション無しでは 18 問中 13 問，アノテーション有りでは 18 問中 10 問が正解となる．発話意図と感情極性のスコアのいずれを使って問題を解くべきかを適切に判断することは今後の課題の一つである．

3.5 語句整序完成問題

語句整序完成問題とは，空所を含む文に対して，与えられた数個の単語列を適切に並べ替えて，文法・意味的に正しい文を完成させる問題である．我々は，文法・語法・語彙問題と同様に言語モデルを用いてこの問題に取り組んだ．具体的には，単語列のすべての並びを列挙し，もっとも文としての確率が高いものを選ぶ手法を用いた．

過去の代ゼミセンター模試では 18 問あり，このうち，15 問（83%）システムは正答することができた．正解できなかった 3 問についてはエラーの要因は以下であった．

2013-1-A21-22　関係代名詞 (a product that will sell)
2013-1-A23-24　遠い依存関係
2013-3-A21-22　局所的な高い頻度の語句

ここでの要因は，文法・語法・語彙問題とほぼ同様である．システムの正解率もほぼ同じであることから，N-gram によって解くことのできる問題はおおよそ 80%であることが確認できる．

今回の分析対象である 18 問について，人間とシステムの正答傾向に違いがあるかも分析した．以下はそのクロス表である．

	システムが正解	システムが不正解	合計
人間が正解	14	1	15
人間が不正解	1	2	3
合計	15	3	18

本クロス表についてフィッシャーの直接確率検定を行ったところ p 値は 0.06 であり有意傾向にあった．こ

問題	未知語（句） 正解の選択肢	エラーのタイプ＆分析
2013-1-2	take a rain check (accept your offer later) 雨天順延券の意味	• イディオム辞書（Wiktionary）の不備 • OED・英辞郎には記載されている • To ask that an arrangement be postponed or an offer taken up at a later date. • 現状，go with you, weather permitting を選んでしまう.
2013-3-2	cognate (related)	• cognate と unfamiliar は，共起しやすく誤って選択してしまう.
2013-4-1	put the shoe on the right foot (criticize the person who is to blame)	• イディオム辞書（Wiktionary）の不備 • OED・英辞郎に記載されている. • To put the blame on the real offender. • 現状，justify what the person has done を選んでしまう.

表 6: 未知語（句）語彙推測問題のエラー内訳

れは，文法・語法・語彙問題と異なるところであり，システムと人間はより近い解き方をしているのではないかと考察される．

人間が不正解でありシステムが正解したものは 1 問（2012-1-A25-26）だった．“all I could think about” という構文が受験生に取っては難しいながら，英語の典型的な用法であり，システムにとっては N-gram で解ける問題だったことによる．

3.6 未知語（句）語彙推測問題

この問題は，出現頻度が低く一般にはあまり知られていないような文章中の単語またはフレーズについて語義を推定し，与えられた選択肢の中から最も意味の近い語義を選択する問題である．今回，word2vec [12] を用い，未知の語句と選択肢のベクトルをそれぞれ求め，コサイン類似度の高いものを選択する手法を用いた．なお，未知の単語が慣用句の場合は，イディオム辞書によって事前に語釈文に置き換えた上でベクトルを算出している．

過去 5 回の代ゼミセンター模試の全 12 問について，9 問（75%）解くことができた．正解できなかった 3 つの問題の内訳を表 6 に示す．二つはイディオム辞書の不備に依る．今回は，Wiktionary から作成したイディオム辞書を用いたが，そのカバレッジが低かった．これらはよりカバレッジの大きい Oxford English Dictionary を用いることで解決できることが分かった．もう一つは単語 “cognate” であるが，単語であっても，辞書の語釈文によって置き換えてベクトルを算出することでこちらも解けることが分かった．すなわち，単語，イディオムについて，置き換える・置き換えないという操作が正しくできれば，本問題については解くことができると言える．

本文:
... ヨーロッパ流の芸術観では、芸術とは自然を素材にして、それに人工を加えることで完成に達せしめられた永遠的存在なのだから、A 造型し構成し変容せしめよう という意志がきわめて強い。それが芸術家の自負するに足る創造であって、それによって象徴的に、彼等自身が永生への望みを達するのである。...

問２　傍線部A「造形し構成し変容せしめよう」とあるが、それはどういうことか。その説明として最も適当なものを、次の１～５のうちから一つ選べ。

1 変化し続ける自然を作品として凍結することにより、一瞬の生命の示現を可能にさせようとすること。
2 時間とともに変化する自然に手を加え、永遠不変の完結した形をそなえた作品を作り出そうとすること。

図 2: 評論傍線部問題の例 (2007 年本試験 第 1 問の問 2)

3.7 英語：まとめと今後の課題

本稿では，東ロボプロジェクトにおいて英語チームが英語問題を解いたときのエラーを分析した結果について述べた．長文読解問題はまだチャンスレベルの正答率であるため，今回は分析対象としなかったが，今後解答できるようになっていくにつれ，エラーを分析していく予定である．なお，長文において，特に問題だと考えている課題は 3 つある．意味を反転させるような表現の扱い，共参照解析，メタ言語（文章自体への言及）である．また，過去の代ゼミセンター模試の長文（特に大問 6 の論述に関する問題）を分析したところ，選択肢に関連のある一文を長文から抽出できれば解ける問題が 25 問中 11 問あったが，その他は複数の文の統合が必要なものであった．要約技術の適用や文の統合といった技術が必要になってくると思われる．

4 国語 評論問題のエラー分析

4.1 センター試験『国語』評論傍線部問題

本節では，大学入試センター試験『国語』評論の**傍線部問題**と呼ばれる問題を取り扱う．傍線部問題の具体例を図 2 に示す．この図に示すように，傍線部問題は，何らかの評論から抜き出された文章 (本文) を読んだ上で設問文を読み，5 つの選択肢のうちから正解の選択肢を 1 つ選ぶという選択式の問題である (紙面の都合上，図 2 には 2 つしか選択肢を記載していない)．傍線部問題は，センター試験『国語』評論の配点の約 2/3 を占めている．

4.2 傍線部問題の解法

東ロボ国語チームは，傍線部問題の自動解法として，これまでに**本文照合法** [17]，およびその一部を拡張した**節境界法** [8] を提案，実装した．本節ではこれらの解法について概説する．

4.3 本文照合法

本文照合法は，

- 正解選択肢を選ぶ根拠は，本文中に存在する [2, 6]
- 意味的に似ているテキストは，表層的にも似ていることが多い

という考え方 (仮説) に基づく解法である．具体的には，次のような方法で傍線部問題を解く．

1. **入力:** 本文，設問，選択肢集合を入力する．
2. **照合領域の決定:** 選択肢と照合する本文の一部 (照合領域) を定める．照合領域は，本文中の傍線部を中心とした連続領域とする．
3. **選択肢の事前選抜:** 考慮の対象外とする選択肢を除外する．具体的には，ある選択肢について，自分以外の選択肢との文字の一致率の平均値が最も小さい選択肢を除外する．
4. **照合:** 考慮の対象とする選択肢をそれぞれ照合領域と比較し，照合スコアを求める．照合スコアには，照合領域とその選択肢との間の共通する要素の割合 (オーバーラップ率 [4]) を用いる．
5. **出力:** 照合スコアの最も高い選択肢を解答として出力する．

この本文照合法には，以下の 3 つのパラメータが存在する．

- 照合領域として本文のどの範囲を選ぶか
- 照合スコアをどのような単位で計算するか (何のオーバーラップ率をスコアとするか)
- 選択肢の事前選抜を行うか

これらのパラメータは，以降で述べる節境界法にも共通する．

4.4 節境界法

節境界法は，長い文を複数のまとまりに区切るという戦略に基づき，本文照合法の一部を拡張した解法である．具体的には，本文照合法の照合ステップにおいて，照合領域と選択肢に節境界検出に基づいた節分割を行い，その結果を照合スコアの計算に利用する．節は「述語を中心としたまとまり」[9] と定義される文法単位であり，おおよそ述語項構造に対応する．

節境界検出には，節境界検出プログラム Rainbow [5] を用いる．Rainbow は，文の節境界の位置を検出し，節の種類のラベル (節ラベル) を付与するプログラムである．Rainbow によって付与された節境界で区切られた部分を節とみなして，節分割を行う[2]．

節境界法では，照合スコアを以下のような方法で計算する．

Step1 照合領域 t と選択肢 x に節境界検出を行い，それぞれ節の集合 T, X に変換する．

Step2 T と X を用いて選択肢 x の照合スコアを計算する．具体的には，X 内の各節 $c_x \in X$ のスコアの平均値を，選択肢 x のスコアとする．節 c_x のスコアは，c_x と，T 内の各節 $c_t \in T$ との類似度の最大値とする．

節同士の類似度は，節同士の共通する要素の割合 (オーバーラップ率 [4]) と，2 つの節の節ラベルが一致する場合のボーナスの和と定義する．

4.5 評価実験

センター試験の過去問および代々木ゼミナールのセンター模擬試験過去問 (以下，代ゼミ模試とよぶ) を用いて，本文照合法および節境界法の評価を行った．センター過去問は 10 回分，代ゼミ模試は 5 回分の試験データを使用した．傍線部問題の総数は，センター過去問が 40 問，代ゼミ模試が 20 問である．

4.5.1 実験結果

本文照合法ソルバーと節境界法ソルバーを，センター過去問，および代ゼミ模試に適用した結果 (正解数) を表 7 に示す．この表の P-m-n は，照合領域 (本文の傍線部の前後何段落を照合領域とするか) を表し，C^1 や L などは，オーバーラップ率として何の一致率を用いるかの単位を表す (たとえば C^1 は文字 unigram

[2] 厳密には本来の節の定義からは外れる場合がある．

	C^1		C^2		L		W	
	non	ps	non	ps	non	ps	non	ps
P-0-0	14/11	17/16	12/11	16/13	9/16	14/19	12/15	15/19
	10/2	**11**/3	6/2	7/3	8/3	**10**/4	8/6	9/6
P-b-0	19/15	**20**/19	14/17	16/18	14/**22**	17/**22**	13/**23**	15/**25**
	8/5	8/4	6/4	9/4	7/3	8/3	9/3	**10**/3
P-1-1	15/19	16/**21**	14/16	15/17	18/**20**	19/**23**	15/**25**	17/**28**
	5/3	5/4	4/3	4/3	5/3	6/3	4/4	7/3
P-1-0	16/15	18/19	14/16	16/17	14/18	18/**20**	12/19	17/**23**
	6/6	7/6	5/3	7/4	6/5	8/4	9/6	**10**/5
P-a-0	**20**/16	**20**/18	15/15	17/16	13/**21**	15/**22**	14/**20**	15/**22**
	10/6	9/5	9/7	**11**/7	5/4	8/4	7/5	8/6
P-b-c	14/15	16/18	13/13	14/16	18/17	19/**22**	13/**21**	15/**24**
	10/5	**10**/4	8/5	8/4	7/6	9/5	7/5	7/5
P-b-1	16/16	16/19	12/17	13/18	17/**20**	18/**23**	14/**23**	15/**26**
	7/4	7/3	7/5	7/4	5/5	6/4	6/2	8/2

表 7: センター過去問と代ゼミ模試に対する正解数 (本文照合法/節境界法，上段がセンター 40 問，下段が代ゼミ模試 20 問に対する結果)

		R@1	R@2	R@3
センター	(節) P-1-1, W, ps	28	29	34
(40 問)	(節) P-b-1, W, ps	26	32	33
	(節) P-a-0, C^2, ps	16	28	36
代ゼミ	(本) P-0-0, C^1, ps	11	12	18
(20 問)	(本) P-a-0, C^2, ps	11	14	17
	(本) P-b-0, W, non	9	15	19

表 8: ソルバー出力の上位に正解が含まれる設問数

(本文の解答根拠部分)
... しかしながら **映画を見るという行為は、一瞬たりとも休むことのない時間の速度にとらわれ、その奴隷と化する**ことでもあった。 ... だが**映画は一方通行的に早い速度で流れる時間に圧倒されて、ついにはひとつの意味しか見出せない**危険な表現であり、...

(正解選択肢)
映画は、限られた時間のなかで壮大な時空間を描き出すようなことを可能にしたが、**映画に見入っている時間をきびしく制限しようとすることで、観客の眼差しを抑圧してしまう**ことになった。

図 3: タイプ C の難問の例 (2005 年本試験 第 1 問の問 4)

を用いることを表す)．また，選択肢の事前選抜を行う場合を ps，行わない場合を non で表す．これらのパラメータの組み合わせ 56 通りについて，正解数を調査した．

表 7 では，本文照合法ソルバー，節境界法ソルバーの正解数を，この順に斜線で区切って示している．また，上段にはセンター過去問の正解数，下段には代ゼミ模試の正解数を示している．半数以上の問題に正解した場合の正解数は，ボールド体で示している．

表 7 を見ると，センター試験と代ゼミ模試の問題は，性質が異なるということがわかる．センター過去問に関しては，多くのパラメータ (45/56) において，節境界法の正解数が本文照合法の正解数以上となったのに対し，代ゼミ模試に関しては，56 通りすべてのパラメータにおいて，本文照合法の正解数が節境界法の正解数以上となった．また，本文照合法では 2 つの問題データ間で正解率があまり変わらないのに対し，節境界法では全体的にセンター過去問よりも代ゼミ模試の正解率の方が低い．

ソルバーは，解答を出力する際，照合スコアの高い順に選択肢番号を出力するが，このとき，スコア上位に正解が含まれた設問数を表 8 に示す．パラメータは，センター過去問または代ゼミ模試で，比較的成績のよいものを 3 つ選んだ．R@n は，スコア順位で n 位までに正解が含まれたことを表す．(節)，(本) はそれぞれ節境界法，本文照合法を表す．

表 8 を見ると，ほとんどの問題で正解選択肢が選択肢 5 つのうちの上位 3 位までには入ることがわかる．スコア上位の選択肢に対して，本文と合致しない部分の検出ができれば，より正解数が向上することが期待できる．

4.5.2 典型的な難問例

本文照合法，および節境界法は，いずれも文字列の表層的類似度を照合スコアに用いているため，本文の解答根拠部分と選択肢との間で表層的に全く異なる言

	R@1	R@2	R@3
(節) P-1-1, W, ps	4	9	15
(節) P-b-1, W, ps	1	7	14
(節) P-a-0, C^2, ps	6	10	15
(本) P-0-0, C^1, ps	10	14	18
(本) P-a-0, C^2, ps	12	14	16
(本) P-b-0, W, non	10	16	19

表 9: 受験生の選んだ選択肢上位にソルバー出力が含まれる設問数

	$SA = HA$	$SA \neq HA$	
		ソルバー	受験生
正解	9	2	5
不正解	1	8	5

表 10: ソルバーと受験生のマーク率 1 位の解答が一致したときの正解数

い回しが用いられているような問題には正解できない. センター過去問の 40 問の傍線部問題を調査したところ，そのような問題は多く存在した．その中でも，以下の 3 つのタイプの問題は，ソルバーにとって特に難問であると考えられる.

A 本文で抽象的に述べている内容を具体的に述べた選択肢を選ぶ設問 (40 問中 2 問)

B 本文で具体的に述べている内容を抽象的に述べた選択肢を選ぶ設問 (40 問中 4 問)

C 本文と選択肢の抽象度は同じだが，選択肢が本文の内容を，句以上の大きな単位で全面的に言い換えている設問 (40 問中 16 問)

タイプ C の設問の例を図 3 に示す.

4.5.3 人間の解答との比較

代ゼミから提供されたデータを用いて，ソルバーの解答傾向が人間 (受験生) のそれと似ているかの比較を行った．代ゼミ模試 20 問において，ソルバーの解答結果と，受験生の解答番号別マーク率を比較した. 受験生の選んだ選択肢 n 位までにソルバーの選んだ選択肢が含まれる設問数を表 9 に示す．この表の R@n は，受験生のマーク率順位の n 位までにソルバー出力が含まれたことを表す．また，ソルバー (本文照合法 P-0-0, C^1, ps) と受験生のマーク率 1 位の解答が一致したときの正解数を表 10 に示す．この表の SA はソルバー解答，HA は受験生の解答マーク率 1 位の選択肢を表す.

表 9 を見ると，節境界法に比べて，本文照合法の解答傾向の方が受験生と似ている．代ゼミ模試において節境界法より本文照合法の方が好成績であったことを考慮すると，代ゼミ模試においては，受験生と解答傾向が似ているソルバーの方が，正解率が高くなると考えられる.

表 10 を見ると，ソルバーは，平均的な受験生と同じ番号を出力したときはおおむね (9/10) 正解し，異なる番号を出力したときはおおむね (8/10) 不正解であるということがわかる．すなわち,「人間が解けず，ソルバーが解ける」という問題は少なく (2/10)，ソルバーが解ける問題は受験生も解けることが多い (9/11) ということがいえる.

4.6 国語：まとめと今後の課題

本節では，東ロボ国語チームが提案，実装した評論傍線部問題の自動解法とその成績，および解答結果の分析について述べた．実装した本文照合法，節境界法は，いずれも文字列の表層的類似度を用いる解法であり，本質的に正解できない難問もあるものの，ソルバーは，適切なパラメータさえ選べば，多くの問題に対してスコア順位で上位に正解選択肢を出力できた.

現在のソルバーは，全ての傍線部問題に対して同じパラメータ，同じ解法で解答するが，今後は，問題を換言型，理由型などいくつかの型に分類し，より適したパラメータ，特徴を用いて解く必要があると考えられる．たとえば，傍線部の理由を問う理由型の問題の場合，本文傍線部周辺の比較的狭い領域の，因果関係を表す表現などが手がかりとなるであろう．また，評論には例示や引用がしばしば用いられるため，本文および選択肢を，本質的に重要な部分とそうでない部分に分け，重要な部分のみで照合を行うようなアプローチも有用であると考えられる.

5 国語 古文問題のエラー分析

センター試験『国語』の古文問題で出題される問題の種類には以下のものがある.

- 傍線部現代語訳
- 文法問題
- 内容理解問題 (心情把握，理由説明など)

- 和歌 (内容理解，技法など)
- 文章表現技法

このうち文法問題に対しては形態素解析器 MeCab と中古和文 UniDic によって得られた古文本文の形態素解析結果を用いて解答し，傍線部現代語訳と内容理解問題に対しては統計的機械翻訳によって古文本文を現代語文に翻訳し，それと選択肢との類似度を計算し，最も類似度が高い選択肢を解として出力するという手法で解答を行う [21]．和歌問題と文章表現技法に関してはそれぞれに対応した解答器を作成せず，内容理解問題と同じ手法を用いている．

2013，2014 年度の代々木ゼミナールのセンター模試 5 回分に対して行った評価での解答器の正答数を表 11 に示す．

	現代語訳	文法問題	内容理解	その他	計
2013-1	1/3	1/1	2/2	0/2	4/8
2013-2	2/3	1/1	1/2	2/2	6/8
2013-3	1/3	0/1	1/2	1/2	3/8
2013-4	2/3	0/1	0/2	0/2	2/8
2014-4	1/3	0/1	1/2	1/2	3/8

表 11: 評価結果 (センター模試 5 回分)

以下，文法問題と内容理解問題についての誤りについて述べる．

5.1 誤り分析

5.1.1 文法問題解答

文法問題では傍線部の表現の構成要素，その品詞，助動詞の場合は用法が問われることが多い．この種の問題に対する解答器の誤りの原因は，解答器が用いている中古和文 UniDic の品詞体系と高校で教えられる古典文法の品詞体系の差異と，助動詞の用法の統語的・意味的違いが識別できていなかったということであった．

前者は具体的には高校教育の古典文法に存在する “形容動詞” が中古和文 UniDic に存在しないというものである．例えば “心細げなり” は一語の形容動詞であるが，中古和文 Unidic では “心細 (形容詞)-げ (接尾辞)-なり (助動詞)” となり，この違いが誤答の原因となった．これに対しては形容動詞が UniDic の文法でどのように表現されているかを規則として記述しておけば解決できると考えられる．

また，文法問題では助動詞の用法が正しく推定できるかが重要であるとされ，誤答の選択肢の中には正答のものと助動詞の用法に関する項目だけが異なるようなものが存在することがある．このような選択肢集合から正答を導くためには助動詞の用法推定は必須となる．現時点の解答器では複数の用法を持つ助動詞に対して文脈に即した用法の推定を行っておらず，これが原因で誤答になる問題があった．

複数の用法を持つ助動詞に対して文章中での用法を推定するという問題は一種の語義曖昧性解消であり，例えば助動詞毎に用法推定モデルを学習しそれを適用することでこの問題に対応することが考えられる．

5.1.2 傍線部現代語訳，内容理解問題解答

内容理解問題に対して，解答器は統計的機械翻訳によって古文本文を現代語訳し，それと選択肢との類似度によって解を決定する．翻訳モデルの実装には moses を利用し，学習コーパスには小学館の新編日本古典文学全集の電子化されている 31 巻分から抽出した 86,684 文対を用いた．

構築した翻訳モデルの性能評価のため，試験問題の古文本文に翻訳モデルを適用して得られた現代語訳の BLEU 値を求めた．結果を表 12，実際に得られた翻訳結果の例を図 4 に示す．

年度-回	2013-1	2013-2	2013-3	2013-4	2014-1
BLEU	23.61	19.56	24.49	32.32	15.72

表 12: 問題本文に対する BLEU 値

古文	世の人いかに言い、沙汰しけんとこそ推し量らるれ。
システム訳	世間の人はどうしたことか言葉、噂しけんと想像されることである。
参照訳	世間の人はどのように言い、噂をしただろうかと想像される。

図 4: 翻訳結果の例

BLEU 値は平均で 23.14 であり，十分な性能であるとは言い難い．生成された翻訳は助動詞などの機能表現は比較的正しく訳されているが，内容語に関しては誤っていたり，古語がそのまま表れている場合が多いという傾向があった．これは学習に用いた対訳コーパスが少量であったことに起因すると考えられる．

解答器は得られた翻訳文を選択肢との比較に用いるため，翻訳モデルの性能が解答器の性能に影響を与え

問題 ID	正解	default	full	raw
2013-2-A25	1	3,1,4,5,2	3,1,4,5,2	3,1,4,5,2
2013-3-A25	2	3,5,2,1,4	3,5,2,1,4	3,5,2,1,4
2013-4-A25	2	5,2,3,1,4	5,2,3,1,4	5,2,3,1,4
2013-4-A26	1	4,1,3,5,2	4,1,3,5,2	4,1,3,5,2
2014-1-A26	4	5,4,2,1,3	5,4,2,1,3	5,4,2,1,3

表 13: 異なる現代語訳による評価結果

る．そこで人手による参照訳を現代語訳として用いた場合に解答器の性能がどう変化するかを検証した．参照訳では古文本文に表れていない主語などの要素や理解に必要な補助的な情報が追加されていることがあるため，その有無の影響もあわせて調査した．各現代語訳毎の本文との類似度に基づいて選択肢を並び替えたものを表 13 に示す．“default” は翻訳モデルで得られた現代語訳，“full” は参照訳，“raw” は参照訳から本文には表れていない情報を削除した訳をそれぞれ表す．

解答器は選択肢と本文との類似度を単純なコサイン類似度で計算しているため，意味的には同じ表現であっても表層が異なっているものを正しく扱うことができていない．選択肢には本文の内容の言い換えや要約によって参照訳の表現がそのまま用いられていることが少ないため，単純な類似度計算では現代語訳による性能に違いが表れなかったと考えられる．

センター試験古文問題は古典文法で記述された文章が理解できるかが重要とされているため，内容理解問題に関しては例えば，小説の登場人物の心情をその行動から推測するといったような現代文の問題と比べて，表層に書かれてあることがそのまま解答の手がかりとなることが多い．従って，正しい現代語訳には解答に必要な情報が含まれていると考えられるため，それと選択肢との内容の類似度を計算する手法を改善する必要がある．

実際の問題解答においては，全ての古典の単語が現代語に訳せることは必須とされておらず，例えば “をかし” のようないわゆる重要単語が文脈に応じて正しく訳せるかどうかが鍵となっていることが多い．このことから，内容の類似度計算を問題解答のみに限定すると，重要単語のような正確に翻訳すべき箇所を推定し，その対応に焦点を当てた計算手法を考慮することが有効であると考えられる．

5.1.3 受験者の解答との相関

受験者が最も多くマークしたものを人間の解答と見なし，システムの解答との関係を調査した．結果を表 14 に示す．また，表 15 に問題分類ごとの正答数を示す．

	システムが正解	システムが不正解	合計
人間が正解	16	18	34
人間が不正解	2	4	6
合計	18	22	40

表 14: 人間の解答とシステムの出力の関係

分類	問題数	人間	システム
現代語訳	15	13	7
文法問題	5	5	2
内容理解	10	9	5
和歌解釈	5	3	2
文章表現	5	4	2

表 15: 問題分類ごとの正答数

評価で利用した問題の実際の受験者が最も多くマークした選択肢とシステムの出力が一致した割合は 0.45(18/40) であった．受験者が最も多くマークしたにも関わらずそれが正解でなかった問題は，そうでない問題に比べて難易度が高いと考えられる．今回の評価で利用した問題のうちそのような問題は 6 問あり，システムが正答したものは文章表現に関するものと内容理解に関するものの 2 問であった．人間の解答において正答率が低かったものは和歌解釈問題であり，5 問中 3 問が正答であった．システムの正答も 2 問と少なく，和歌に関する問題は両者にとって難しいことが分かる．

5.2 未対応の問題

現時点の解答器では和歌の問題と表現技法に関する問題が未対応である．和歌の問題は和歌の表現技法に関する問題と，内容理解に関する問題に分類できる．内容理解に関しては本文の内容理解と同様の手法での解答が考えられるが，和歌には枕詞や縁語のような特有の表現技法や比喩などが用いられる傾向が強く，その現代語訳は本文のものとは異なっているため，本文に対する翻訳モデルをそのまま適用できない．また，現時点で対訳コーパスとして利用できる歌集は古今和歌集のみであるため，これのみから和歌用の翻訳モデルを新たに構築することも困難である．

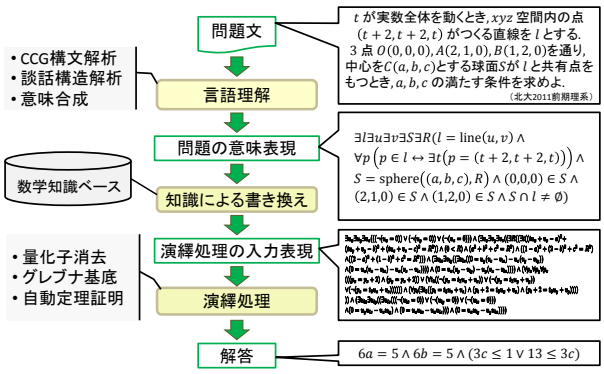

図 5: 数学解答システムの構成

古典文法で記述されたテキストが新たに生成されることは基本的にはないため，現存する全ての古典テキストに対して電子化を行い，対訳を作成すれば現代語訳に関わる問題は解決する．しかし，コストがかかるため現実的とは言えない．このことから古文問題解答では少量の対訳コーパスが利用できるという状況で，どのように効率的に翻訳精度を向上させるか，そのためにはどのような知識を構築すれば良いかが一つの焦点となる．

5.3 まとめと今後の課題

現時点の古文問題解答器は，文法問題に関しては中古和文 UniDic を用いた形態素解析結果と選択肢との比較，内容理解問題に関しては機械翻訳による現代語訳と選択肢との比較により問題に解答する．その正答率は 0.45(18/40) であった．解答器で用いている技術は非常に基本的なものであり，特に古文問題解答に特化した処理は行っていない．今後はこれまでに述べたように，解答性能の向上に向けて，助動詞の用法推定，現代語翻訳の性能向上，重要単語に焦点を当てた内容の類似度計算手法の改良を行う．

また，言語資源として今回のタスクでは古文-現代文対訳コーパス，中古和文 UniDic を用いたが，今後，コーパスの規模の拡大を望むことは困難である．そこで，受験者が学習に利用する古語辞典などから解答に必要な知識を抽出し，それを用いた手法にも取り組む予定である．

6 数学解答システムの意味合成処理における「理論的エラー」の分析

数学解答システムは，図 5 のような構成になっており，大きく分けて，自然言語で記述された数学問題を論理表示へと翻訳する言語理解処理と，形式化された表現のみを操作対象とする書き換えおよび演繹処理の二つの部分に分けられる [11, 10]．前半の言語理解処理の中心は組合せ範疇文法（Combinatory Categorial Grammar, CCG）[18] の枠組みを用いた意味合成であるが，現在，構文解析および照応・文脈解析などが開発中であり，言語理解部分の自動化については道半ばの段階である．そこで，これまでのシステム評価は，係り受け解析・照応解決の結果や評価時点の辞書に存在しない語の用法を人手でアノテートした問題文を入力とし，これら付加された情報も用いて CCG による意味合成結果を半自動的に得る形で行ってきた．このため，本節では通常の意味での言語処理のエラー，すなわち曖昧性解消における誤りや辞書の被覆率の不足に関して分析するのではなく，CCG の枠組みによる我々の言語分析で現在カバーできていない現象の分析，言い換えれば言語の形式的分析の不足に起因する「理論的なエラー」の分析を行う．以下では，2014 年度第一回代ゼミセンター模試の数学 IA・数学 IIB のデータを対象に，上記の意味のエラーのうち 2 例を挙げ解説する．

6.1 型違いの並列

形式意味論の分野では述語論理やその変種が意味表現のための体系として用いられ，句・単語などの「部分の意味」から「文（章）全体の意味」を合成するための形式的手段としては，真理値の型 t および個体 (モノ) の型 e の 2 種類の基底型に基づく型付きラムダ計算を用いるのが一般的である．本研究でも，意味表現と合成のための体系として型付きラムダ計算を用いているが，個体については単一の型 e ではなく，例えば「実数」「2 次元平面上の点」「整数のリスト」といった多数の型を区別する体系となっている．

このように個体に関し様々な型を付けた体系を用いることで，述語の項などに関する選択制限を型の整合性として形式的に扱うことができる．これによって，統語的曖昧性のために多数発生する意味的に不整合な解釈の大多数を構文解析の途中で排除できると期待される．

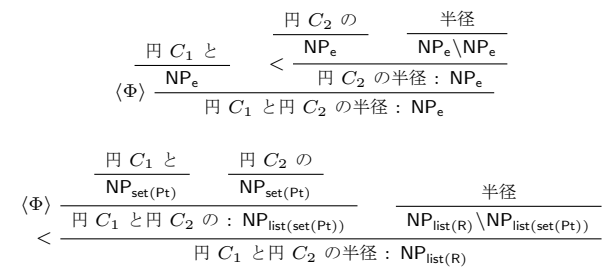

図 6: 並列句の誤った解析（上）と正しい解析（下）

型によって制約された言語解析が有効に働くであろう場面の代表的なものは並列句の解析である．図 6 は並列を含む名詞句「円 C_1 と円 C_2 の半径」について，単一の個体型 e のみを持つ体系での導出のひとつ（上段）と，多数の個体型をもつ我々の体系における導出（下段）を示している．上の導出で「円 C_1」と「円 C_2 の半径」が誤って並列されているように，個体の型として単一の型 e のみを考える体系では意味的な不整合性を型によって検出することはできない．これに対し，我々の意味表現体系では実数の型 R，図形（点集合）の有限列の型 list(set(Pt))，実数の有限列の型 list(R) などを区別し，さらに並列句を構成する文法規則 ⟨Φ⟩ で 2 つの娘句が同一の意味的型を持つことを保証することで，図 6 下段に示す「円 C_1」と「円 C_2」が並列される解析のみが許される．

このように，細かく型付けされた意味表示体系は曖昧性解消の手段としての効果が期待できるが，その副作用として，型が異なる句どうしの並列句が扱えなくなるという問題がある．そのような「型違いの並列」は頻度は少ないものの，

> ... を満たす ΔABC について，その重心 G の座標と面積 S を求めよ．

といった表現としてしばしば実際の数学問題に現れる．この例では平面上の点の型 Pt をもつ「その重心 G の座標」と実数の型 R を持つ「面積 S」が並列されており，我々の現在の意味表現体系では扱うことができない．実際に，2014 年度の代ゼミセンター模試でも

> $\overrightarrow{CP}$ を $\vec{a},\vec{b},t$ を用いて表すと... となる．

という表現を含む問題 (数学 II・B 第 4 問) が出題され，ここでは $\vec{a}$ と $\vec{b}$ が 2 次元ベクトルの型 Vec2D，t は実数の型 R をもつために解析ができなかった．

この「型違いの並列」の問題を解決し，かつ型の区別による曖昧性解消を行うための方法の一つは，並列名詞句を現在の枠組みのように型 α の要素のみからなる有限列（list(α) 型）として扱うのではなく，α 型のモノと β 型のモノからなる 2 つ組の型 (α,β)，α，β，γ 型のモノからなる 3 つ組の型 (α,β,γ)，... を適当な n つ組まで考え，異なる型の要素からなる k つ組として分析することである．この方式では，上記の模試問題の「$\vec{a},\vec{b},t$」の部分は型 (Vec2D, Vec2D, R) を持つことになる．その上で，このような型違いの並列句を項として取りうる「(を用いて) 表せ」「(を) 求めよ」のような動詞に関しては，引数として 2 つ組，3 つ組，...，n つ組を取る用法に対応した n 個の異なる辞書定義を与えればよい．

当然ながら，この方式では，型は異なるものの実質的には非常に類似した意味表現を持つ多数の辞書定義によって辞書が肥大化し，解析の際の計算コストが増大する恐れがある．しかし，これまでの観察によれば「型違いの並列句」を項とし得る述語はごく限られている．よって，それら少数の述語に対してのみ n つ組までを項に持つ複数の辞書定義を与え，通常の型が揃った並列句を項として取る述語についてはこれまで通り list(α) 型の引数を取る辞書定義を与えることで，上記のデメリットをある程度押さえることができる．

しかし，この体系では有限列の型 list(α) と同種のものの 2 つ組，3 つ組，... の型 (α,α)，(α,α,α)，... がともに存在するため，型 α をもつ名詞句どうしの並列に対しては，解析の途中段階で常にリスト型と k つ組の型をもつ複数の解釈が発生するという複雑さは依然として避けられない．

6.2 行為結果の表現

現在の我々の意味表現体系で扱えない別の例として，行為や操作の結果を表す表現を取り上げる．2014 年度センター模試数学 I・A では

> 104 を素因数分解すると ア^3・イウ である．

という文を含む出題があったが，現在の我々の文法体系ではこの文に対する意味合成ができない．同様の「X を V すると Y となる」という構造を持つ文 (以下,「行為結果文」と呼ぶ) は他にも

- n を 2 乗すると 4 の倍数となる．
- 放物線 C を y 軸方向に 1 だけ平行移動すると放物線 D となる．
- 円の半径を 2 倍にすると面積は 4 倍になる．

など種々あり，数学テキストでは比較的よく現れるタイプの文である．

動詞「なる」および接続助詞「と」の通常の用法も考慮すると，行為結果文「X を V すると Y となる」の意味表現としてもっとも表層構造に忠実なのは以下のようなものだろう：

1. 行為 V の前の世界 W_1 と行為後の世界 W_2 には，ともにモノ X が存在する．
2. 行為 V の結果モノ X の性質は変化し，行為後の世界 W_2 ではモノ X とモノ Y は一致する，あるいはモノ X は W_2 では性質 Y を満たす．

ここでは行為 V の前・後における世界の変化を捉えるために，ある種の時間の概念（ないし複数世界間の推移）が意味表示の体系に持ち込まれている．

しかし，上記のような一連の行為結果文から問題を解くために読み取る必要がある意味内容は通常の述語論理の枠組みで十分表現可能である（例えば，上の箇条書きの最初に例に対しては「n^2 は 4 で割り切れる」）．また，明示的に時間の推移を表す「点 P は速度 v で動き，時刻 t に点 Q に到達する」といった表現を含む問題は比較的少数であることも考えあわせると，現在の開発段階で意味表現に時間の概念を持ち込む利得は意味表現・言語解析および推論の複雑化に見合わないと考える．

幸い，これまでに観察された行為結果文は定型的なものが多く，時間を含まない現在の枠組みで，必要な意味表現を合成することは多くの場合に可能であると思われる．特に「X を V すると Y となる」という形の文については，少なくとも下記の 2 つの方針が考えられる：

方針 1 主節「Y となる」はガ格のゼロ代名詞を持ち，そのゼロ代名詞は間接照応で「X を Y した結果」を指すと考える．

方針 2 句「V すると」は右にガ格を欠いた一項述語，左にヲ格名詞句を項として取ると考える（即ち，「V すると」は範疇 $\mathsf{S\backslash NP_o/(S\backslash NP_{ga})}$ を持つ）．

方針 1 のゼロ照応の解決は，行為結果文の定型性を利用することで比較的容易に実現できると予想されるが，方針 2 は，利点として CCG による統語・意味解析の枠組み内で全ての意味合成処理が行えることに加え，例えば「2 乗すると 10 を超える奇数」のような連体修飾の形も上記の範疇を持つ「V すると」の語彙項目によって同時に扱える点が挙げられる：

$$
\dfrac{\dfrac{\dfrac{\text{2 乗すると}}{\mathsf{S\backslash NP_o/(S\backslash NP_{ga})} : \lambda P.\lambda x.P(x^2)} \quad \dfrac{\text{10 を超える}}{\mathsf{S\backslash NP_{ga}} : \lambda x.x > 10}}{\mathsf{S\backslash NP_o} : \lambda x.x^2 > 10} > \mathbf{rel}}{\mathsf{N/N} : \lambda N.\lambda x.Nx \wedge x^2 > 10} \quad \dfrac{\text{奇数}}{\mathsf{N} : \lambda x.\mathrm{odd}(x)} >
$$
$$
\mathsf{N} : \lambda x.\mathrm{odd}(x) \wedge x^2 > 10
$$

6.3 数学：まとめと今後の課題

本節では数学自動解答システムにおける意味合成の理論的部分に関して，現在のシステムの限界の一端をセンター模試に現れた例を通して解説した．一つ目の「型違いの並列」の例は，型付きの関数型プログラミング言語あるいはそのモデルである型付きラムダ計算における技術を言語分析に直接応用しようとする際に生じる問題であり，二つ目の「行為結果表現」の例は，統語構造となるべく並行的で，一般性のある意味分析を目指す形式意味論的なアプローチと，適切な複雑さの範囲で最も点を取れるシステムを目指す工学的要請のギャップに起因する問題であったと言えるだろう．今後は，ここで挙げた 2 例を含む意味合成おける現実的な問題に関し，重要な現象から分析・実装を行うとともに，言語処理部分の自動化を進め，曖昧性解消の段階でのエラーについても観察と解決を進める．

7 物理問題のエラー分析

大学入試における物理の問題の多くは，問題に記述された状況において，ある物理現象が起きたときの物理量についてのもの (e.g. “物体が停止した時間”) や，物理現象が起きるための条件となる物理量についてのもの (e.g. “棒がすべり出さないための静止摩擦力”) である．本研究ではこの種の問題解答に向けて，物理シミュレーションによって問題に書かれている状況を再現し，得られた結果を用いた手法で取り組んでいる [20]．提案手法による問題解答の流れを図 7 に示す．

解答器は自然言語文で記述された問題を入力として受け取り，まず意味解析を行い，状況の記述と解答形式の記述からなる形式表現を生成する．次に形式表現を元に物理シミュレーションを行い，得られた結果から問題に記述されている物理現象が起きた時刻における物理量を特定し，解答形式にあわせて出力することで問題に解答する．センター試験のような選択式の場合は，得られた物理量と選択肢とを比較し，その差が最も小さくなるものを解答する．

2014 年度のセンター模試による評価では形式表現からシミュレーション結果の取得に焦点を当て，人手で

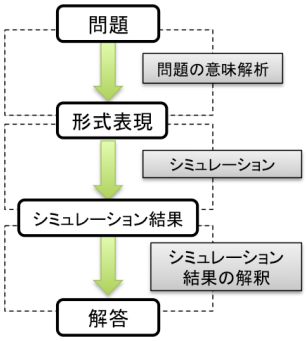

図 7: 物理解答システムの概要

記述した問題の形式表現を入力とし，そこから得られたシミュレーション結果から解答が導けるかどうかを人手で判断することで解答を行うという設定とした．

人手で記述した形式表現には基本的には誤りは含まれないと考えられるため，正しくシミュレーションできれば正答に到達できると考えられる．しかし，全ての問題に対してシミュレーション結果が得られるというわけではない．その主な理由としては以下の 3 つがある．

(i) 形式表現の記述が困難

(ii) シミュレータで実行できない

(iii) 問題がシミュレーションが困難な物理現象を含む

(ii) に関してはモジュールのバグなどによるもの，(iii) は電磁誘導のような現象が対象の問題である．本稿では (i) に焦点を当て，その詳細について述べる．

7.1 形式表現

本手法で用いている形式表現は一階述語論理の形式で記述している．定義している述語は物体，物理量，物体に対する操作，物理現象を表す 4 種類のものである．このうち物体に対する操作と物理現象を表す述語に関しては，事象が起きた時間関係を明示するためにイベント変数を導入している．

現時点における形式表現の定義でどの程度の問題が記述できるかを代々木ゼミナールの 2013，2014 年度のセンター模試 5 回分を用いて評価した．結果を表 16 に示す．状況記述の項は実際に形式表現で記述できた小問の数，括弧はその割合を示している．また，+の値は新たに述語を定義すれば形式表現として記述可能となる小問の数を示している．

形式表現に用いる述語は過去のセンター試験問題を対象とした調査結果を基に人手で定義したものである．

大分類	小分類	小問数	状況記述
力学	物体の運動	14	11(0.79)+3
	力のつり合い	24	15(0.63)+3
	円運動	5	3(0.6)
	圧力	6	3(0.6)
エネルギー	エネルギー	6	2(0.33)+1
	仕事	1	1(1.00)
波	波の性質	17	6(0.35)+9
	音	14	6(0.43)+4
	光	9	3(0.33)+3
電気と磁気	電気回路	8	5(0.63)
合計		103	55(0.53)

表 16: 形式記述の分析 (試験 5 回分)

図 8: 複雑な形状の例

表 16 の状況記述の項の “+” の値は新しく述語を定義することで状況の記述が可能となった問題の数を示す．

高校物理の分野は力学，エネルギー，波，電磁気の 4 つからなる．本研究では力学から着手しているため，力学においては状況記述が可能な問題の割合が大きい．

状況の記述ができないと判断された問題の原因を表 17 に示す．

原因	問題数
シミュレータでの再現が困難な表現を含む	12
現状の定義では記述できない表現を含む	10
その他	2

表 17: 形式表現での記述が不可能であった原因

以下上位 2 つの原因について詳細を述べる．

7.1.1 シミュレータでの再現が困難な表現を含む

問題には “平らな床の上に置かれた物体” のように基本的な要素で構成された状況だけでなく，図 8 左のように画像によって与えられた複雑な形状の要素が出現するものもある．また，図 8 右では “細い金属でできた棒を直角に折り曲げ……” という物体に対する操作によってまっすぐな状態から変形された棒の形状が示されている．このような状況を形式的に表現する手段にはそれぞれの物体を直接指すような (例えば，“穴の空いた円板 (x)” のような) 述語を新たに定義すると

いうものが考えられる．同様に “棒を曲げる” といった操作も “曲げる (x,e) ∧ 棒 (x)” という形で記述することが可能である．しかし，本手法では形式表現はシミュレータの入力に用いられるため，定義された述語に対応するシミュレータ上の機能が存在している必要がある．例えば，前述の “曲げる” のような物体に対する変形操作は物体の挙動のシミュレーションでは対象とされていない．この点において，上図のような状況は現時点では実行可能な形式表現として記述することができないと判断する．

7.1.2　現状の定義では記述できない表現を含む

現時点では，例えば，“質量 $2m$ のおもりを鉛直に吊すと切れてしまう糸” のように，その問題においてのみ特殊な性質を有しているような物体の記述ができないという問題がある．

質量や位置などの物理量と同様にこの性質を物体の属性として記述することは可能であるが，その場合，どのような性質があり得るかをあらかじめ列挙できるかどうかが問題となる．

7.2　残っている課題

現時点では数値データとして出力されるシミュレーション結果を人手で解釈して解答しているが，最終的にはこの部分も自動化する必要がある．物理の問題では単純に物理量を問われるようなものや，例えば “止まった時の時刻” のような基本的なものから，“動いている人にはサイレンはどのように聞こえるか” といったような自然文として記述される物理現象の定性的な挙動まで様々な問いがなされる．解答に必要な情報はシミュレーション結果から得ることができるが，それをどのように解釈して解答に用いれば良いかはこの問われていることから判断しなければならない．前者の物理量を答えるような基本的な問題であればその記述は容易であるが，後者の場合，そもそもその解答形式をどのように形式的に記述し得るかを考える必要がある．

問題文からの形式表現の生成も残されている課題であるが，物理の問題には問題文とともに状況を示した図が添付されていることが多い．必要な情報が全てテキストで与えられている問題もあるが，7.1.1 節で挙げたように図でしか与えられない情報が存在する場合もある．従って，問題を正確に理解するためには画像の解釈を行う必要がある．

物理の問題には，シミュレーションで解けないような問題も含まれている．例えば，虹や夕焼けのように日常で遭遇する現象がどういう物理現象なのかが問われるような，物理に関する一般的な理解を問うような問題である．この種の問題の解答には，例えば，光の色によって屈折のされ方が変わるといった物理現象の性質に関する知識を保有しておく必要がある．

また，エネルギーに関する問題は基本的に各種エネルギーの公式から得られるエネルギーの値をエネルギー保存則に基づいて関係づけて計算することで解答するという形式のものであり，用いられるエネルギーの公式では具体的にどのようにエネルギーが生じるかは重要ではなく，例えば，水力発電所で生じるエネルギーも，手回し発電機で生じるエネルギーも同じ公式で得ることができる．解答に必要な物理量が与えられれば，それがどのような形で与えられるかは問題を解くという点では関係が無く，力学の問題などに比べて自由に状況を設定することができ，表現のバリエーションが他の分野に比べて多くなっている．そのため，この種の問題を一般化して形式表現にするのが困難であり，加えて，電熱器や発電機など様々な要素が出てくるためシミュレーションで解くのは困難である．

この種の問題に対しては，人間が解く時と同様に，エネルギー保存則を用いて，解答に必要な数式を導出することで解答する別の手法が必要となる．

7.3　まとめと今後の課題

物理問題の多くは問題で与えられた状況に対して起きた物理現象について，その時の物理量やその物理現象が成立するための条件を問うものである．このような問題に対して，我々は問題に書かれている状況を認識し，その状況を物理シミュレータによってシミュレーションし，得られた物理量をもとに解答するというアプローチで取り組んでいる．これまで，十分な範囲の問題を記述することができ，かつ，その情報から物理シミュレーションが可能となるような形式表現の定義を行ってきた．表 16 に示すように，まだ記述できない問題は残っているため，今後も定義を改良する必要があるが，同時に自然文として記述されたテキストからこの形式表現への変換にも取り組む予定である．

また，本節の最初で述べたように物理の問題の全てがシミュレーションで解答できるというわけではない．例えば，電磁誘導に関する問題では，ある操作を行ったときに物理的挙動がどのように変化するか，といった定性的な内容が問われることが多い．物理的な挙動

の定性的な推論に関しては，定性推論と呼ばれる領域で研究が行われており (e.g. [1])，これらを踏まえて物理問題解答のための定性推論についても取り組む予定である．

8　世界史・日本史のエラー分析

本節では，2014 年度の代ゼミ模試に対し，狩野 [24] のシステムが出力した解答のエラー分析について報告する．本システムは，山川出版社の世界史または日本史の用語集を知識源とし，設問から抽出したキーワードが知識源の中でどのように分布しているかをスコアとして算出し，解答を選択する．具体的には，設問および知識源に対して以下の各処理を行い，解答の選択を行う．[3]

1. 問題文解析 : 問題文のテキストを前処理し，キーワード抽出を行う対象テキストを切り出す．一般に，設問は背景説明のテキストや導入文，実際に正誤判定の対象となる文など，複数のテキストから構成される．そこで，これらのテキストから後段の処理で必要となるテキスト箇所を抽出する必要がある．

2. キーワード抽出 : 前処理した問題文テキストから，スコア付けに用いるキーワードを抽出する．キーワードリストとして，Wikipedia の見出し語から自動抽出した語句を人手でクリーニングしたものを用い，単純なマッチングでキーワード抽出を行った．

3. 知識源検索 : 抽出したキーワードで知識源を検索し，キーワードに合致するテキストを得る．

4. スコア付け : キーワードと検索結果テキストとの一致度をスコア付けする．後述するように，センター試験では文の正誤を判定する問題が多い．誤りを含む文では，知識源のまとまった範囲内にキーワードが出現せず，別の場所に出現すると考えられる．したがって，検索結果テキストにキーワードが含まれない場合は，ペナルティとして負のスコアを与える．

5. 解答選択 : 文の正誤を判定するタイプの問題に対しては，スコアが大きいものを正しい文として解答を選択する．語句を解答するタイプの問題（いわゆるファクトイド型質問応答に相当）に対しては，選択肢に挙げられた語句を問題文テキストに埋め込み，文の正誤判定問題に帰着して解答を行う．年代を解答する問題については，検索結果テキスト中の年代表現を抽出することで解答を行う．

[3]本システムは図表の処理は行っておらず，図表に対して人手でアノテーションされたテキストを利用して解答を行う．

> 下線部 1（ノモスという多くの小国家）に関連して，ノモスは王国へ発展するが，古代エジプトの王国について述べた文として正しいものを，次の 1〜4 のうちから一つ選べ．
> 2. 中王国は，ヒクソスを撃退してエジプトを再統一した．
> 4. ヒッタイトは史上はじめて鉄製武器を使用し，新王国と争った．

図 9: 問題文解析の誤りの例

表 18 に世界史，表 19 に日本史の問題タイプとエラー分析結果を示す．センター試験の世界史・日本史では，選択肢として与えられた文に対して正誤を判定するタイプの問題が大きな割合を占める（例えば図 9）．語句や年代を解答するタイプの問題（例えば図 11）はいわゆるファクトイド型質問応答に見えるが，知識源中の解答に関連する記述は多くの場合一つしか無く，大規模テキストを利用した解答の aggregation といった技術は利用できない．したがって，語句・年代と問題文との組合せの正誤を判定するタスクに帰着される．このように，知識源を的確に参照しつつ，文の正誤を判定するという処理は，上記のようにテキストの前処理，キーワード抽出，検索，スコア付け等，複合的な処理が必要であり，また各処理で高い精度が要求される．各処理は当然不完全なものであり，必ずしも排他的な関係にあるわけでもない．よって，最終的に誤答が出力された要因を単一の原因に帰着することは難しいため，表 18，表 19 では，複数の要因は別個にカウントしてエラーの分類を行った．

「問題文解析」は，問題に解答するための情報が書かれた問題文テキストを切り出す処理に起因するエラーである．図 9 に例を示す．[4]この問題では，問題文中に「ノモス」「王国」「古代エジプト」といったキーワードが現れるが，実はこれらの情報は選択肢の正誤判定には無関係である．つまり，選択肢の文のみを用いて正誤の判定を行うことができる．一方，問題によっては問題文中のキーワードが正誤判定に必要な場合や，さらに背景説明のテキストも参照する必要があることもある．次のエラー要因とも関連するが，どこまでの

[4]問題例を挙げる際には，紙面の都合上，選択肢の一部のみ抜粋する．

	文の正誤判定	語句の解答	年代の解答	図を参照	計
問題文解析	2	2	0	0	4
キーワード抽出	4	0	0	0	4
一般知識	0	1	1	0	2
言語知識	2	0	0	0	2
言語構造	1	1	0	0	2
その他	2	1	0	0	3
誤答数	8	4	1	0	13
問題数	22	8	3	3	36

表 18: 世界史の問題タイプと誤答の要因

	文の正誤判定	語句の解答	年代の解答	図を参照	計
問題文解析	1	3	0	0	4
キーワード抽出	4	1	0	0	5
一般知識	8	1	0	0	9
言語知識	1	0	0	0	1
言語構造	2	1	0	0	3
その他	0	0	0	1	1
誤答数	13	4	0	1	18
問題数	23	7	2	4	36

表 19: 日本史の問題タイプと誤答の要因

下線部 c（大和地方を中心とした政治連合であるヤマト政権が誕生し）に関連して，ヤマト政権の支配体制に関して述べた文として正しいものを，次の 1〜4 のうちから一つ選べ.
2. 君（公）や直の姓の有力豪族が国政を担当し，ヤマト政権の中枢をになった.
3. 国造に任命され，地方の支配権をヤマト政権から保証された地方豪族もいた.

図 10: キーワード抽出の誤りの例

背景説明：戦後すぐの 1949 年に行われた群馬県岩宿遺跡における発掘調査以降，各地で　ア　の地層から打製石器が発見され，日本における旧石器時代の文化の存在が明らかになった.
設問：空欄アイに入る語句の組合せとして正しいものを，次の 1〜4 のうちから一つ選べ.
2. ア 更新世　イ ひすい（硬玉）
4. ア 完新世　イ ひすい（硬玉）

図 11: データベース・オントロジー的知識が利用できる例

テキストをキーワード抽出の対象とすべきかは単純には決定できない.

「キーワード抽出」は，当該問題を解くのに必要・不必要なキーワードを分別できていないことに起因するエラーである．図 10 に例を示す．この例では，2 は誤った文であるが,「君」「直」などがキーワードとして認識されず，これらの語が知識源に現れなかったにも関わらずペナルティがかからなかったため，正しい文と判定されてしまった．これ以外にも，例えば「法制」「編集」といった一般語がその問題文中では重要なキーワードとなっているような場合や，逆に「アジア系」のような専門用語らしい語が知識源には明示的に書かれていないため，ペナルティがかかってしまった例がある．世界史・日本史の知識がある程度ある人間が読めば，重要なキーワードと重要ではない（知識源に明示的に書かれていなくても正誤判定には影響しない）キーワードがある程度区別できるが，これを実現するのは容易ではない.

「一般知識」は，解答のために必要な知識が明示的に知識源に記述されていないことに起因するエラーである．世界や日本の地理・時代に関する知識，一般常

下線部 c（紀元前 4 世紀には西日本，ついで東日本の東北地方へも広がった）に関して述べた次の文 X・Y について，その正誤の組合せとして正しいものを，下の 1〜4 のうちから一つ選べ．
X 北海道に水稲耕作は及ばず，採集・狩猟・漁労中心の続縄文文化が続いた．
Y 沖縄・南西諸島には南方から伝わった農耕が広まり，独特な貝塚文化が始まった．
知識源：擦文文化　7〜13 世紀頃に北海道に広く展開した鉄器文化．名称は続縄文土器と土師器の影響を受けて誕生した櫛の歯のような文様を持つ擦文土器に由来する．農耕もあるが，主生業は狩猟・漁労．北海道式古墳も築造．

図 12: 一般知識が必要な例

設問：下線部 d（水稲耕作を基礎とする弥生文化）に関連して，弥生時代の農耕について述べた文として正しいものを，次の 1〜4 のうちから一つ選べ．
2. 籾を田に直接まく直播が行われ，田植えは始まっていなかった．
4. 収穫した稲の脱穀は，木製農具の臼と竪杵を用いて行った．
知識源：木製農具　耕作具として鍬・鋤，水田面を平均にならすえぶり，精穀具としての木臼・竪杵などがあり，イチイガシ・栗などの堅い木材でつくられた．

図 13: 言語知識が必要な例

識に照らした判断，等が必要とされる．単純な例としては，図 11 のように，「岩宿遺跡」がどの時代の遺跡か，という知識を予め用意しておけば解答できるような問題もある．このような知識は必ずしも教科書・用語集に明示されているわけではないが，データベースなどの形式で整理しておくことは可能である．より困難な例を図 12 に示す．この場合，知識源の文章を読めば，「北海道に水稲耕作は及ばず」が妥当であることが分かるが，この判断のためには農耕，狩猟，水稲といった概念の知識と，それらを対比して判定を行う処理が必要である．このように知識源の記述と設問の記述が直接一致しないケースは特に日本史の問題に多い．

「言語知識」および「言語構造」は，自然言語処理技術の利用・高精度化により解決できる可能性のあるエラーである．前者は，例えば「解読」と「未解読」が反義語であるといった語彙知識や，「収穫した稲の脱

背景説明：…『宋書』倭国伝には 5 人の倭王が朝貢したことの記述があるが，そのうち倭王　ウ　は雄略天皇に比定されている．
設問：空欄ウエに入る語句の組合せとして正しいものを，次の〜のうちから一つ選べ．
2. ウ 武　エ 須恵器
4. ウ 興　エ 須恵器
知識源：倭の五王　421〜502 年に五人の倭王 (讃・珍〈彌〉・済・興・武) が宋などの南朝と 13 回通交した記事が『宋書』などにある．讃は仁徳か応神・履中，珍は反正か仁徳，済は允恭，興は安康，武は雄略の諸天皇に比定されている．

図 14: 言語構造が必要な例

穀」と「精穀具」のパラフレーズ関係など，言語知識を利用することで正答が得られる可能性があるものである．「解読」「未解読」のような例であれば，言語リソースの整備により解決できる可能性が高い．しかし，図 13 に示すような例はパラフレーズ認識あるいはテキスト間含意関係認識に相当するものもあり，必ずしも容易に解決できるものではない．後者は，係り受け解析，述語項構造解析，否定の解析などによって，文の意味の違いを認識することが必要とされるものである．典型的には，文章中に複数の命題が記述されている場合がある．図 14 に示す例では，正解は 2 であるが，4 のキーワードも同一文章中に含まれているため，スコアが同率となり，最終的に誤った解答を選択してしまっている．この例は係り受けあるいは述語項構造が正確に得られれば正しい解答が得られると期待される．また，図 13 の例では，システムは 2 を選択したが，これは知識源には「田植えをした可能性も高まった．」と記述されており，否定やモダリティの正確な解析によって正答が得られる可能性がある．ただし，このような言語知識・言語解析は新たなエラー要因を持ち込むため，単純にこれらの技術を導入することでは全体の正答率が下がる可能性が高い．必要な場面で適切かつ正確に言語処理技術を利用する必要がある．

8.1　世界史・日本史：まとめと今後の課題

本節では，世界史・日本史の試験問題を対象に，狩野 [24] のシステムのエラー分析を行った．本システムは構文解析，意味解析等の自然言語処理を行わず，設問と知識源とのキーワードの一致をスコア付けする方式をとっている．自然言語処理の立場からは，より深

い言語処理技術を利用することで正答率を上げるというアプローチが考えられるが，エラー分析の結果からは，それにより正答できる問題はそれほど多くなく，また精度が不十分な言語リソース・言語解析を導入することによる副作用も懸念される．一方，問題文の前処理やキーワード抽出に起因するエラーはまだ一定数残っており，これらは改善の余地があると考えられる．また，特に日本史では一般知識・常識や，知識源に直接記述されていない知識を統合的に利用する必要がある問題が見られる．これを解決することは容易ではないが，自然言語理解の興味深い未解決問題の一つと見ることもできる．

9 おわりに

本稿では大学入試センター試験形式の模試問題データを主たる対象として，英語・国語（現代文評論，古文)・数学・物理・日本史・世界史の各科目に対する解答システムのエラーを分析した．言語系科目（英語・国語)，数理系科目（数学・物理)，人文系科目（日本史・世界史）に渡る多様な知的課題に並行的に取り組み，現時点での課題を通覧することで現在の NLP/AI 諸技術の限界と達成に関しひとつの見取り図を提出できたと考える．

また，上記の科目グルーピングの内部においても，第 2 言語能力の測定を主眼とする英語の読解問題と，母語による「読解能力」を評価する国語の読解問題の違いや，言語理解・演繹の両フェーズにおいて時間の推移を本質的に扱う必要のある物理と，オブジェクト間の静的な関係についての理解が主要な課題となる数学の違いなど，言語処理課題として興味ある差が存在すると考えられる．さらに，本稿では触れなかったが，これまでの評価結果から，解答システムの構成によらず日本史は世界史に比べ一般に難しいらしい，との観察が得られており，類似の問題形式をもち，ともに検索・質問応答・含意関係認識などの複合課題と見える日本史・世界史の間にも，言語処理タスクとしての性質の違いが存在することが示唆される．

これらの言語処理課題としての性質の違い，また，言語理解のフェーズに後続する演繹的処理から言語理解部分への異なる要請（エラー許容度・表示形式の違い等々）は，東ロボプロジェクト全体としての問題設定の内部で興味深いスペクトラムをなしており，各科目の解答システム開発を通じた今後の技術的成果およびエラー分析結果を基礎として,「言語を理解して問題を解く」とは全体としてどういうことなのか，また，どこまでが，なぜ機械化可能なのかを見極めることが今後の課題である．

謝辞

本研究を推進するにあたって，大学入試センター試験問題のデータをご提供下さった独立行政法人大学入試センターおよび株式会社ジェイシー教育研究所に感謝いたします．また，模擬試験データおよび解答分布データをご提供下さった学校法人高宮学園に感謝いたします．また，日本史および世界史用語集の電子データをご提供くださった山川出版社に感謝いたします．

参考文献

[1] Kenneth D. Forbus. Qualitative process theory. *Artificial Intelligence*, Vol. 24, No. 1-3, pp. 85–168, 1984.

[2] 船口明. きめる！センター国語現代文. 学研教育出版, 1997.

[3] Joseph Gubbins and Andreas Vlachos. Dependency language models for sentence completion. In *Proc. EMNLP*, pp. 1405–1410, 2013.

[4] 服部昇平, 佐藤理史. 多段階戦略に基づくテキストの意味関係認識: RITE2 タスクへの適用. 情報処理学会研究報告 2013-NL-211 No.4/2013-SLP-96 No.4, 情報処理学会, 2013.

[5] 加納隼人, 佐藤理史. 日本語節境界検出プログラム Rainbow の作成と評価. FIT2014 講演論文集 第 2 分冊 pp.215-216, 2014.

[6] 板野博行. ゴロゴ板野のセンター現代文解法パターン集. 星雲社, 2010.

[7] Daniel Jurafsky, Elizabeth Shriberg, and Debra Biasca. Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual, draft 13. Technical Report Technical Report 97-02, University of Colorado, Boulder. Institute of Cognitive Science, 1997.

[8] 加納隼人, 佐藤理史, 松崎拓也. 節境界検出を用いたセンター試験『国語』評論傍線部問題ソルバー. 情報処理学会研究報告 2015-NLP-220, 情報処理学会, 2015.

[9] 益岡隆志, 田窪行則. 基礎日本語文法 -改訂版-. くろしお出版, 1992.

[10] Takuya Matsuzaki, Hidenao Iwane, Hirokazu Anai, and Noriko Arai. The complexity of math problems – linguistic, or computational? In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pp. 73–81, Nagoya, Japan, October 2013. Asian Federation of Natural Language Processing.

[11] Takuya Matsuzaki, Hidenao Iwane, Hirokazu Anai, and Noriko H. Arai. The most uncreative examinee: A first step toward wide coverage natural language math problem solving. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, pp. 1098–1104, 2014.

[12] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Proc. NIPS*, pp. 3111–3119, 2013.

[13] Yusuke Miyao and Ai Kawazoe. University entrance examinations as a benchmark resource for nlp-based problem solving. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pp. 1357–1365. Asian Federation of Natural Language Processing, 2013.

[14] Bo Pang and Lillian Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proc. ACL*, pp. 115–124, 2005.

[15] Anselmo Penas, Eduard Hovy, Pamela Forner, Alvaro Rodrigo, Richard Sutcliffe, Corina Forascu, and Caroline Sporleder. Overview of QA4MRE at CLEF 2011: Question answering for machine reading evaluation. In *CLEF 2011 Labs and Workshop Notebook Papers*, pp. 19–22, 2011.

[16] Anselmo Penas, Eduard Hovy, Pamela Forner, Alvaro Rodrigo, Richard Sutcliffe, Caroline Sporleder, Corina Forascu, Yassine Benajiba, , and Petya Osenova. Overview of QA4MRE at CLEF 2012: Question answering for machine reading evaluation. In *CLEF 2011 Labs and Workshop Notebook Papers*, pp. 303–320, 2011.

[17] 佐藤理史, 加納隼人, 西村翔平, 駒谷和範. 表層類似度に基づくセンター試験『国語』現代文傍線部問題ソルバー. 自然言語処理 vol.21 No.3 pp.465-483, 言語処理学会, 2014.

[18] Mark Steedman. *The Syntactic Process*. Bradford Books. Mit Press, 2001.

[19] Xiao Zang, Zhiyong Wu, Helen Meng, Jia Jia, and Lianhong Cai. Using conditional random fields to predict focus word pair in spontaneous spoken english. In *Proc. INTERSPEECH*, pp. 756—760, 2014.

[20] 横野光, 稲邑哲也. 論理演算と物理シミュレーションの結合による物理問題解答. 2014 年度人工知能学会全国大会, 2014.

[21] 横野光, 星野翔. 統計的現代語訳モデルを用いたセンター試験古文問題解答. 第 5 回コーパス日本語学ワークショップ, 2014.

[22] 新井紀子. ロボットは東大に入れるか. イースト・プレス, 2014.

[23] 東中竜一郎, 杉山弘晃, 磯崎秀樹, 菊井玄一郎, 堂坂浩二, 平博順, 南泰浩. センター試験における英語問題の回答手法. 言語処理学会第 21 回年次大会（NLP2015）, 2015.

[24] 狩野芳伸. 大学入試センター試験歴史科目の自動解答. 2014 年度人工知能学会全国大会（第 28 回）, 2014.