

### 3.13 ロボットは東大に入れるか

—代々木ゼミナールセンター模擬試験による中間評価—

松崎 拓也 (名古屋大学/国立情報学研究所)

横野 光 (国立情報学研究所)

宮尾 祐介 (国立情報学研究所)

#### それなに？

「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研究プロジェクトである。「東ロボ」では、大学入試試験問題を自動的に解くソフトウェアの開発を行っている。その狙いは、言語処理および人工知能関連技術を総ざらえし、問題テキストの解析から分野特有の演繹、そして解答の生成までを含めた統合的な知的処理課題をベンチマークとして、今後10年の言語処理・AI関連技術の伸びしろと限界を見極めることにある。プロジェクトの目標は2016年度にセンター試験で高得点を獲得し、2021年度に東大合格レベルに到達することである。

#### どうやるの？

試験問題には図・表・イラストなども含まれるが、その中心は自然言語テキストである。このため、各科目とも言語処理が自動解答の基盤となっている。その一方で、問題に解答するためには、言語理解に加えて推論、数値解析、教科書など知識源の検索、画像理解といった種々の技術が必要となる。「東ロボ」の技術的目標の1つは、これらの知的情報処理技術と言語処理の接合を通じて、計算機による課題解決の基盤として頑健に動作する言語処理技術を確立するとともに、言語理解との接合によって種々の知的情報処理の間口を格段に広げることにある。

#### かだいは？

「東ロボ」では代々木ゼミナール(代ゼミ)主催のセンター模擬試験を用いた中間評価を2013年・2014年に行った。2014年の点数と受験者中での偏差値を表-1に示す。表のように、各科目とも受験

者平均に近い得点を獲得している。ただし数学および物理の結果は言語処理の一部を手で代替して得たものであり、これら理数系2科目では言語処理部分の自動化が直近の課題となる。英語・国語では、長文読解問題に対する解答手法の改良が大きな課題となっている。日本史・世界史では言語知識・分野知識など異種の知識を統合的に利用するシステムの実現が課題である。以降では、物理および日本史・世界史について、代ゼミ模試の結果に対する分析の一端を紹介する。

#### ぶんせき？

##### 物理のエラー分析

物理の解答システムでは、問題文の意味解析によって得られた形式表現をもとに、物理シミュレーションで状況を再現し、その結果を用いて解答するという手法を採用した。

エラー分析では、特に問題で与えられた状況が形式表現で記述できるか否かに注目した。模試5回分の小問104問のうち状況の記述ができなかったものは25問あり、その主な原因としてシミュレーションのモデルの不足が挙げられる。本手法ではシミュレータで実行可能なモデルとして問題を抽象化しており、このモデルの仕様が形式表現に対応している。したがって、状況を再現できるモデルがなければその問題の形式記述は得られない。現時点では「床の上の物体の運動」のような基本的な状況を中心にモデルを構築しているが、試験問題には「水中に伏せたコップを沈めると、コップの中に水が入り込んだ(図-1)」のようにモデルとしての抽象化が困難な状況や、「質量2kgのおもりをつるすと切れる糸」

### 3.13 ロボットは東大に入れるか

科目	得点/満点	偏差値
英語	95/200	50.5
国語	現代文	49/100
	古文	20/50
数学 IA	40/100	46.9
数学 IIB	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の結果

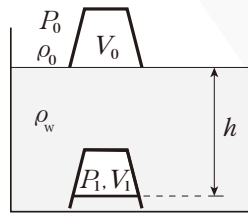


図-1 シミュレーションが困難な状況の例

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、次の内から一つ選べ。

1. 中王国は、ヒクソスを撃退してエジプトを再統一した。
2. 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-2 センター世界史の問題例

のように構築済みの「糸」モデルの拡張が必要な要素が含まれていた。モデルの追加は今後の課題であるが、少数のモデルで多くの問題が被覆できるだけでなく、問題文からの自動生成が実現可能であるようなモデルの設計を考える必要がある。

#### 日本史・世界史のエラー分析

日本史、世界史のセンター試験問題は、歴史的事実に関する選択肢から正しいものを選ばせる問題が多い。図-2に具体例を示す（説明の都合上、選択肢は2つのみ示す）。

これらの問題は「富士山の高さは？」のような質問を典型とする従来の質問応答よりやや複雑である。従来の質問応答では、基本的に事実を述べた文（たとえば「富士山の高さは3776m」）を検索することが必要となる。一方、試験問題の場合は事実と異なる選択肢が存在し、しかもそのような選択肢は表層的には事実と非常に近いテキストである。よって、正しい文と誤った文を識別するためには、「富士山」「高さ」といったキーワードの出現をもとに事実を検索する従来の手法に加え、「非事実らしさ」を評価する何らかの仕組みが必要となる。

狩野<sup>1)</sup>は、教科書などの知識源から、問題文と選択肢に含まれるキーワードによって文書を検索し、キーワードが検索結果に含まれない場合にはペナルティを与えることで、正しい文と誤った文を識別する手法を提案した。表-1の結果は、この手法によるものである。

エラー分析によると、現在のシステムの誤りの多くは、各キーワードが問題を解くために重要かどうかの分別ができないことに起因する。上記の例では、「ノモス」というキーワードがあるが、「ノモス」と「中王国」「新王国」が存在した時期には500～1000年ほどの隔りがあり、問われている事柄の正誤を判

定する上で「ノモス」は無関係である。しかし、解答システムはこのキーワードも含めてスコア付けを行ったため正誤判定を誤った。このように、キーワードとして重要かどうかは問題内容や文脈に依存して決まる場合が多い。たとえば、「君」「編集」といった一般語がカギとなる場合もあれば、「アジア系」といういかにも専門用語らしき単語が無関係である場合もある。つまり、正誤判定におけるキーワードの重要性は、歴史の知識を前提としており、ここに現在のシステムの限界がある。

これ以外には、同義フレーズや反義語などの言語知識や、時間や地理に関する一般知識を組み合わせることで正誤判定を行う必要がある場合などが見られた。これらは1つのアイデアですべてが解決できるような問題ではなく、個別の技術やリソースを地道に積み重ね、全体として1つのシステムとして動作させることが必要である。

#### 展望

「東ロボ」では、言語処理に限らず関連領域の研究者のさらなる参加を求めつつ、今後も模試や実際の試験問題を用いた評価および結果の分析を継続的に行っていき、それぞれの技術の伸びしろと限界を見極めていきたい。

#### 参考文献

1) 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（2014）。

（2015年10月6日受付）

松崎 拓也（正会員） matuzaki@nuee.nagoya-u.ac.jp  
 専門は言語処理。東大助教，国立情報学研究所特任准教授を経て現在，名古屋大学准教授。

横野 光（正会員） yokono@nii.ac.jp  
 専門は言語処理。東工大精密工学研究所研究員を経て現在，国立情報学研究所特任助教。

宮尾 祐介（正会員） yusuke@nii.ac.jp  
 専門は言語処理。東大助教を経て現在，国立情報学研究所准教授。