# 言語処理技術の統合的評価基盤としての大学入試問題

宮尾祐介
国立情報学研究所
yusuke@nii.ac.jp

川添愛
津田塾大学/国立情報学研究所
zoeai@tsuda.ac.jp

松崎拓也
東京大学
matuzaki@is.s.u-tokyo.ac.jp

横野光
国立情報学研究所
yokono@nii.ac.jp

## 1　はじめに

本稿では，大学入試問題を言語処理研究のベンチマークとして利用する意義について議論する．まず，実際の試験問題の分析を通して，自然言語処理の観点から主な研究課題を概観する．特に，自然言語で記述された問題への回答に必要な技術と，現在の言語処理技術との乖離に着目する．また，現在整備を進めているリソースを紹介し，用語認識や含意関係認識などの基盤技術から，質問応答や要約などの応用まで幅広い言語処理技術の研究・評価に利用可能であることを示す．

大学入試問題は人間の知的能力を客観的に測定するために設計され，問題の多くは自然言語テキストとして与えられることから，知的システムを目指す自然言語処理研究にとって興味深い題材である．一方で，必要とされる知識はごく限られ，必ず答えが存在し，合理的な解法で答えが導けることが保証されている．このように人工的に制限されつつ多様な知的処理が必要なタスクを対象とすることで，個別化した言語処理技術を統合的に評価することができ，また一般社会に対して分かりやすく自然言語処理の成果を示すことができる．さらに，試験問題を解くプロセスを言語処理タスクとして見ることで，試験問題が何を測定しているのか，人間の知的能力とどのような関係にあるのか，といった疑問に対する手がかりが得られると期待される．

小学生の国語や算数の問題を対象とした研究では，世界知識や常識の必要性が問題点として挙げられている [8, 4, 5]．大学入試は必要な知識が多い反面，一般常識に依存する部分が少なく，解法が合理的・論理的に説明できる場合が多い．したがって言語処理タスクとしてはより取り組みやすいと言える．本研究は，制限された世界の中で必要とされる暗黙の知識や意味理解の仕組みを明らかにすることを目指している．

## 2　タスク設定

本研究は問題を理解し答えを導く部分に焦点を当てるため，人間が実際に試験を受ける際の様々な難事は捨象する．まず，入出力は試験問題の画像や筆記ではなく，XML 形式のテキストデータとする．図 1 に実際のデータの一部を示す．大問・小問 (question)，指示文 (instruction)，文章や写真などの資料 (data)，回答欄 (ansColumn)，選択肢 (choice) などの文書構造や，下線 (uText) やラベル (label) などの記号が XML で構造化されている．また，自然言語テキスト中以外の

```
<exam subject="SekaishiA(main exam)" year="2009">
 <title>2009 年度 本試験 世界史A</title>
 <question id="Q1" minimal="no">
  <label>【1】</label>
   <instruction>モニュメントや歴史的建造物について述べた次の文章A-
   Cを読み，以下の問い (問 1-11) に答えよ．(配点 33)</instruction>
   <data id="D0" type="text">
    <label>A</label>
    現在，アテネの中心部の丘にその偉容を誇る<uText id="U1"><label>
    (1)</label>パルテノン神殿</uText>は，...
   </data>
   <data id="D9" type="image">
    <img src="K000W11-001.png"/>
   </data>
   <question id="Q2" minimal="yes">
    <label>問 1</label>
    <instruction>
     前の<ref target="D9">写真</ref>の左手前のアーチを多用した建
     物はローマ時代の劇場であるが，古代ローマを代表する建造物として
     正しいものを，次の 1-4 のうちから一つ選べ．
    </instruction>
    <ansColumn id="A1">1</ansColumn>
    <choices>
     <choice><cNum>1</cNum>ピサ大聖堂</choice>
     <choice><cNum>2</cNum>コロッセウム</choice>
...
```

図 1: 2009 年度センター試験世界史 A の XML データ

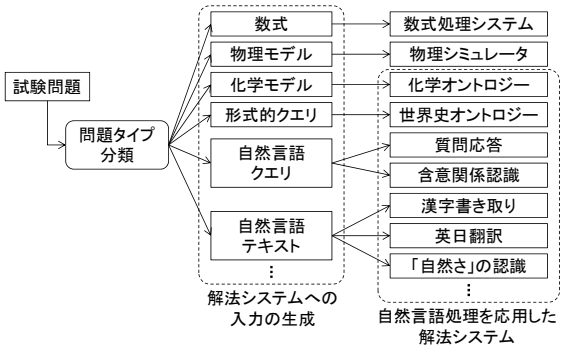


図 2: 試験問題回答システムの基本設計

参照関係も XML でマークアップする (ref)．

さらに，各設問に対して問題タイプを示すラベルを人手で付与する．現実の試験問題は多様な自然言語処理タスクを含んでおり，全てを一度に研究対象とすることは難しい．問題タイプを予め付与することで，3 節に示すような各言語処理タスクを個別に抽出して研究対象とすることができる．

## 3　問題分析

本研究では，図 2 に示す基本設計を想定する．まず問題タイプにより分岐し，それぞれの問題タイプについて専門の解法システムを用意する．例えば，数学の計算問題であれば数式処理システムが利用でき，世界史

の知識を問う問題であれば，後述するように質問応答や含意関係認識に帰着することができる[1]．これにより，試験問題全体ではなく，特定の自然言語処理タスクに焦点を絞ることができる．現在は人手で付与した問題タイプラベルにより分岐するが，将来的には自動分類を適用することを想定する．

この基本設計を前提とすると，自然言語処理は主に2つの場面で必要となる．一つは，解法システムへの入力を生成する場面である．例えば，数学の問題を解くためには数式を立てる必要があるが，問題は自然言語テキストとして与えられる．よって，テキストを入力として数式処理システムへの入力を生成するタスクとしてとらえられる（3.1節）．もう一つは，解法システムの中で言語処理技術を応用する場面である．漢字書き取りや翻訳はほぼ自明であるが，後述するように様々な自然言語処理技術が応用できる可能性がある（3.2節–3.7節）．また，多くの問題で共通して必要となる要素技術（3.8節）や，問題タイプ分類を前提としない問題回答システムについても議論する（3.9節）．

## 3.1 形式表現への写像

数学や物理の問題の多くは，問題文から数式を立てるタスクと考えられる [6] [2]．これは，自然言語テキストを形式表現に変換するタスクととらえられ，情報抽出や論理表現への写像手法の延長線上にある．しかし，手がかり表現や固有表現・専門用語が明確でなく，テキストが記述する世界の抽象的構造を認識する必要がある場合もあるため，興味深い研究課題である．

## 3.2 知識を問う問題

知識を問う問題とは，教科書や参考書に書いてあることをどれだけ覚えているかが評価される問題である．例えば，ある出来事が起きた年代を問う問題があり，典型的な質問応答タスクと見られる．ただし，実際の設問は必ずしも質問文の形式ではなく，問題文を質問応答システムへの入力に変換する必要があるが，これは3.1節で挙げたタスクと同様の難しさがある．

また，大学入試センター試験のような選択式試験では，文として与えられた言明の真偽を判定する問題がよく出題される．これは質問応答ではなく教科書などの知識を前提とした含意関係認識とみなすことができる．例えば図3上に示した問題は，真偽判断の根拠となる文章が教科書やWikipediaに存在するため（図3下），根拠となる文章から各選択肢が含意できるかどうかを判断する問題に帰着できる [3, 9]．

## 3.3 オントロジーに基づく推論

図3の問題は原理的には含意関係認識に帰着できるが，様々な推論が複雑に絡み合っており，現在一般的な含意関係認識手法で解くのは困難である．一方で，試験問題で問われる範囲は限定されているため，範囲内の用語や概念を形式的に記述し，オントロジーに基づく推論に帰着する手法も考えられる．さらに，図4のような定性的推論が必要な問題は，数式に帰着することはできず，含意関係認識に帰着することも難しいため，このアプローチが第一候補となる．

既存研究において，化学知識をオントロジーで記述し，問題に回答するシステムが開発されている [1]．このシステムは自然言語を直接入力とするわけではないが，このようなオントロジーを構築できれば，3.1節の技術と結合することで上記のような問題を解くことができる．しかし，大学入試が対象とする概念は化学物質のような定義が明らかなものに限られない．例えば地理においては「トウモロコシ」「農地」などは重要な概念であるが，いわゆる固有表現や専門用語と違い形式的定義が自明でない．抽象概念が多い分野におけるオントロジーに基づく知識の構造化・形式化や，それに基づく用語・概念認識は重要な研究テーマである．

安全保障理事会 (安保理) についての記述として誤っているものを，次のうちから一つ選べ．
1. 安保理の常任理事国は，手続事項以外の事項について，拒否権をもっている．
2. 安保理は，国際社会の平和と安全の維持または回復に必要な軍事的措置を決定する場合には，あらかじめ総会の承認を得なければならない．
（2009年度センター試験　政治・経済）

……1945年2月に開催されたヤルタ会談において，大国の拒否権は手続事項に適用されないこと，紛争の平和的解決が試みられている間は当事国は表決に加わらないとの妥協が成立した．…すなわち，常任理事国の1か国でも反対投票を投じれば決議は否決されるため，常任理事国は拒否権を有していることになる．
…武力による威嚇・武力の行使が禁止され，平和破壊行為・侵略行為が国連安全保障理事会で認定されれば，必要に応じて非軍事的措置，非軍事的措置が行き詰れば軍事的措置でこれを排除する体制が整備された．

図 3: 言明の真偽を判断する問題（上）と根拠となるWikipediaの文章（下）

脳下垂体からは様々なホルモンが分泌される．それらの作用を調べるために，ラットを麻酔し，苦痛のない状態で脳下垂体の摘出手術を行い，その後の様子を観察した．脳下垂体を摘出した後，ラットに起こる変化として最も適当なものを選べ．
1. 尿量が増加する．
2. 代謝が盛んになる．（2009年度センター試験 生物I）

図 4: 定性的推論が必要な問題

## 3.4 概念化・事例化

3.2，3.3節とは異なるタイプの問題として，抽象概念と事例の関係を判断する問題がある．図5の例では，選択肢で与えられた事例が抽象概念（自我同一性を見失っている状態）と一致しているかどうかを認識する必要がある．抽象概念と事例の関係は含意・推論関係ではなく，また事例を予め列挙することは不可能である．概念の定義と事例が持つ抽象的構造との一致を認識する手法が求められる．

[1] 本稿における試験問題に対する解釈や解き方は筆者らの自然言語処理研究者としての解釈であり，公式見解ではない．
[2] 正しい数式が立てられれば，数式処理システムや物理シミュレータで答えが導けると仮定している．

自我同一性を見失っている心理状態の例として最も適当なものを一つ選べ．
1. 定年で，仕事を辞め，空虚さを感じていた時もありました．今，これまでの人生を振り返って自分史を書き始めています．思いのほか，たくさんの人にお世話になってきた自分を改めて感じています．
2. 子どもも大学に入り，家を出ていきました．心の中にぽっかりと穴が空いた感じが続いています．自分の人生っていったい何だったのだろうか，自分の存在意義って何なのだろうか，いろいろと思い悩んでいます．
（2009 年度センター試験　倫理）

図 5: 抽象概念と事例の一致を判断する問題

秀吉による全国統一には，鎌倉幕府以来の武士社会における結合の原理に基づく面がある．秀吉はどのようにして諸大名を従えたか．2 行以内で述べなさい．
（2009 年度東京大学前期試験　日本史）

図 6: 出題意図・要求に沿った知識の要約が必要な問題

### 3.5　読解問題

国語に限らず，倫理や現代社会などの社会系科目にも読解問題は多く見られる．読解問題については，含意関係認識に帰着する手法 [2, 7] や，問題タイプごとに専用システムで解く手法 [8] が提案されている．特に，論文 [8] で指摘されているように，世界知識や常識的推論が必要とされる場合が問題である．例えば国語の読解問題では，本文から推測できることとできないことを厳密に区別しつつ，明示的に述べられていない感情（行動や人間関係から読み取れる心情）や，論旨の流れから見た自然な結論の認識が要求される．参照解析などの要素技術の高度化に加え，意見ホルダーや感情の経験者の同定，常識的推論の妥当性の判定など，これまでにあまり取り組まれていない技術が必要である．

### 3.6　知識・文章の要約

国語や社会系科目の問題では，持っている知識あるいは問題中で与えられた文章を要約する問題が多く見られる．しかし，単純な要約ではなく，出題意図に沿った要約や，一定の条件を満たす要約が要求されることが多い．要約の条件としては，「指定された語句を使って」といった比較的単純なものもあるが，図 6 の例のように，具体的な事例（秀吉による個々の大名の制圧）についての知識を，明示的・暗黙的に指定された観点（鎌倉幕府以来の結合の原理）から，個々の事例を抽象化した形でまとめることが要求される場合もある．特定の観点に従った要約は既に先行研究があるが，上記の例のように要求される観点を把握すること自体が難しい場合も多く，また，抽象的な観点に従った要約を生成するためには，3.2, 3.3, 3.4 節で議論した様々な意味解析技術を統合的に利用する必要がある．

### 3.7　文章の自然さの認識

英語の会話・文章穴埋め問題では，文章としての一貫性・自然さを認識する必要がある．文章の自然さのレベルには，単語の選択制限や熟語の知識から判断できる比較的容易なものに加え，談話構造の認識が必要なもの，および世界知識・理解が必要なものがある（図 7）．談話構造の認識が必要な問題としては，接続詞や助動詞の時制の選択があり，談話中の命題が成立する時間の特定，および因果関係や話者の知識状態の認識が必要である．世界知識が必要な問題では，「乳歯が抜けた後に永久歯が生える」のような一般知識に加え，「クローゼットの中では，服は畳まれるのではなくハンガーに掛けられる」のような，必然とは言い難い慣習や傾向が求められることもある．単語の共起関係やスクリプト知識で近似できる可能性もあるが，本質的な解決は極めて難しい．

次の会話の空欄に入れるのにもっとも適当なものを選べ．
Zack: もう 10 時だ．ボブがトイレから帰ってきたら店を出た方がいいね．割り勘にする？
Koji: いや．僕は君たちよりたくさん食べて飲んだから，多く払うべきだと思う．
Zack: （空欄）
Koji: それは公平だね．
1. 雑誌のクーポンを持ってくればよかった．
2. 別々に払うようにする？
（2009 年度センター試験　英語）

図 7: 文章の自然さの認識が必要な問題（日本語訳）

### 3.8　要素技術

試験問題は誤解が生じないように分かりやすく記述されているが，自然言語処理の観点では必ずしも解析は容易ではない．特に，ほぼ全科目を通じて広い意味での参照解析が必須である．ここでいう参照解析とは，既存の照応解析や共参照解析が対象とするものだけでなく，テキストの表層に明示的に現れない先行詞や，表層・構文構造に現れない参照表現（トピックなど）の認識を含む [6]．また，問題文中には一般的でない固有名や記号・数式，さらに穴埋め問題用の空白など，各教科に特有な表現が多数含まれる．頑健な形態素・構文解析手法の開発も重要な研究課題である．

### 3.9　問題文の理解

2 節で述べたように，当面の研究課題は，典型的な問題タイプについて個別の解法システムを開発し，その基盤となる言語処理技術を向上させることである．しかし，試験問題として問われうる全ての問題タイプを網羅することは不可能であるため，将来的な課題として，想定外の問題に対しても問題文を理解し，蓄積した解法システム・基盤技術を適切に組み合わることで解法システムを自動生成する手法を開発することが考えられる．これを達成するためには，上述の様々な解法システム・基盤技術の相互運用性を高め，言語処理技術とその入出力に関するオントロジーを整備することで，大学入試タスク全体を分析・形式化しなければならない．もしこれが実現されれば，3.1 節や 3.3 節の技術を，問題文の要求を言語処理ワークフローへと翻訳するという一段階抽象的な処理に適用することが可能になる．

## 4 リソース

3 節で述べたように，大学入試問題を解くためには多様な自然言語処理技術が必要なため，全てを一度に研究対象とすることは難しい．異なる研究グループが全体システムへの寄与を評価しながら並行して研究を進めるために，以下のようなリソースおよび基盤システムを開発し，公開する予定である．

**文書構造アノテーション**　過去 22 年分のセンター試験問題全科目，および各科目の教科書データに対し，図 1 のように主要な文書構造をタグ付けした．

**問題タイプアノテーション**　試験問題の各小問ごとに問題タイプを表すラベルを付与した．分類基準として，回答方式（記述式，選択式等），答えの表層的な形式（文，用語，記号，画像等），回答に必要な知識・技術の種類（教科書的知識，一般的知識，図表の読み取り，テキストの理解等）の三つを設けている．

**用語アノテーション**　社会系・理科系の試験問題と教科書テキストに現れる重要な用語をタグ付けした．各科目に共通あるいは固有の重要概念の分類カテゴリを 50 程度設け，その下位クラスに相当する概念およびインスタンスの両方をアノテーション対象としている．

**含意関係認識評価データ**　センター試験の正誤を問う設問を利用し，含意関係認識評価データを構築した．センター試験の選択肢と，その正誤を判定する根拠となる Wikipedia の記述をペアとしたデータであり，NTCIR-9 RITE タスクにおいて提供された [3, 9][3]．

その他形態素や係り受け構造など基本的な言語構造のアノテーションと，3.8 節で議論した参照関係のアノテーションコーパスを開発・公開する予定である．

また，問題タイプに基づき対象とすべき問題のみを抽出して評価することができる基盤システムを開発している．これを利用することにより，例えば質問応答で答えられる設問のみを対象として，独自の質問応答システムを評価し，また全体システムへの寄与を評価することができる．それぞれの問題タイプについては簡単なベースラインシステムをあわせて公開する予定である．質問応答システムなど，それ自体で複合的なシステムについては，それぞれのシステムをさらにモジュール化する．したがって，構文解析や参照解析などの要素技術を改良・差し替え，全体システムへの寄与を評価することができる．

## 5 おわりに

本稿では，大学入試問題の分析に基づき，問題に回答するために必要となる自然言語処理技術を概観した．既存技術の延長線上に位置づけられるものも多いが，現状の技術では実際の問題に回答するには不十分な場合が多く，構文解析・参照解析などの要素技術の高度化と併せて研究すべきテーマは多い．これらは大学入試問題に特化した技術ではなく，様々な応用にも密接に関係している．例えば，3.2，3.4 節のタスクは高度な検索技術と考えられ，3.3 節は抽象概念を扱う分野（情報科学や法学など）の知識の構造化にも共通の研究課題である．3.5，3.6 節はテキスト情報の高度な編集・編纂タスクであり，3.7 節の技術は対話システムや文章生成にそのまま応用可能なものである．また，3.9 節は自然言語の入力に反応して行動するという本質的な自然言語理解への一つのアプローチである．これらのタスクはそれぞれ独立ではなく，深い意味理解の様々な現れ方と考えられる．

本稿で挙げた自然言語処理タスクの多くは，試験問題が本来測定しようとしている能力とは異なる場合が多いことに注意すべきである．これは，人間が無意識に行っている意味理解が現在の自然言語処理のボトルネックになっているからであり，画像理解や物理世界の理解においても同様の問題がある [6]．大学入試という制限された世界の中で，どこまで意味理解の仕組みを明らかにできるかが問われていると言える．

[3] 本データは NTCIR より公開予定である．

## 参考文献


[1] J. Angele, E. Moench, H. Oppermann, S. Staab, and D. Wenke. Ontology-based query and answering in chemistry: OntoNova @ Project Halo. In *Proc. 2nd ISWC*, 2003.

[2] I. Dagan, O. Glickman, and B. Magnini. The PASCAL recognising textual entailment challenge. In *Machine Learning Challenges*, Vol. 3944 of *LNCS*. 2006.

[3] H. Shima, H. Kanayama, C.-W. Lee, C.-J. Lin, T. Mitamura, Y. Miyao, S. Shi, and K. Takeda. Overview of NTCIR-9 RITE: Recognizing inference in text. In *Proc. NTCIR-9*, 2011.

[4] 安西祐一郎, 上里譲, 田村淳. 算数の文章題を解くシステムにおける問題の内部表現について. 知識工学と人工知能, 1985.

[5] 稲葉栄美子, 渡部広一, 河岡司. 常識知識を用いた算数問題の意味理解. 第 159 回自然言語研究会, 2004.

[6] 横野光, 稲邑哲也. テキストからの物理モデル生成に向けて. 言語処理学会第 18 回年次大会, 2012.

[7] 笠原要, 平博順, 永田昌明, 柴田知秀, 黒橋禎夫. 複数文からなる文章読解タスクへのテキスト含意認識の適用. 言語処理学会第 17 回年次大会, 2011.

[8] 関根聡, 齋藤真実, 岡田美江, 井佐原均. 小学 2 年生の問題を解く—電脳優子 2 年生国語・概要. 言語処理学会第 11 回年次大会, 2005.

[9] 宮尾祐介, 嶋英樹, 金山博, 三田村照子. 大学入試センター試験を題材とした含意関係認識技術の評価. 言語処理学会第 18 回年次大会, 2012.