CF-007

試験問題画像における XML 構造と属性の推定
Estimation of Attributes and Structured from Exam Questions

松本 涼† 遠藤 聡志‡
Ryo Matsumoto Satoshi Endo

1 はじめに

人工知能技術で大学入試問題を解くことに挑戦する「ロボットは東大に入れるか」（以下、「東ロボ」）プロジェクト [1] は、大学入試センター試験（以下、「センター試験」）の模擬試験において、2016 年時点で全教科合計の偏差値 57.1 を達成した。しかしながら、その入力は人手で電子化、アノテーションされた XML データであり [2]、「XML 化にコストがかかる」、「人が XML データを作成しないと解答動作を行えない」などの課題がある。従って、センター試験の画像から、End-to-End で XML データを生成するシステムが必要となるが、そのためには、抽象的な複数の情報の抽出、推定および統合を行う必要がある。

磯崎ら [4] は、センター試験英語の画像データに文字認識（OCR）を行い、得られた文字列にルールベースで XML タグを付与する手法を提案しており、いくつかの情報について抽出・推定、および XML 化を実現した。しかしながら、全ての情報の抽出、推定は実現できていない。また、XML データの精度評価も行われていない。

そこで本研究では、XML の仕様を参考に、XML データを要素、属性、構造、内容の 4 種類の情報に分けて抽出、推定し、精度評価を行うことを提案する。著者ら [5] はこれまでに、要素の抽出と精度評価を行なった。よって本稿では、属性および構造の推定と精度評価を行う。また、End-to-End での XML データ生成について、考察を行う。

2 研究背景

2.1 東ロボくん プロジェクト

東ロボプロジェクト [1] は、大学入試センター試験および二次試験問題を解く人工知能の開発を目的としている。2015 年時点で、センター試験模試の合計点が学生の平均点を超えており、2019 年センター試験の英語筆記本試験において、185 点（偏差値 64.1）を達成した*1。しかしながら、その目的は「試験問題に正解する」ことであり、問題を（解く前の）読み取る動作において、人の手作業を許容している。すなわち、入力データは画像形式ではなく、計算機にとって扱いやすい XML 形式を前提としており [1][2][3]、XML データの作成は、人が行う必要がある [4]。

2.2 センター試験 XML データ

センター試験 XML データは、1990〜2017 年度の大学入試センター試験問題を、XML 形式で電子化したデータセットである。東ロボくんプロジェクトが公開*2しており、プロジェクトが独自に策定した XML ベースのアノテーション仕様（以下、「東ロボ XML 仕様」）に従って記述されている。

3 End-to-End 自動 XML 化システム

XML(eXtensible Markup Language) *3 は，あらゆる文書を電子化する目的で策定されたマークアップ言語である。本研究は、XML を要素、属性、構造、内容に分解し、それぞれについて情報の抽出・推定を行い、得られた情報を XML データとして統合する End-to-End モデル（以下、「E2E XML 生成システム」）を提案することを目標としている。

3.1 XML タグ、内容について

XML において、デザインや構造、意味などのメタ情報は、図 1 のように、文や図表などを XML タグでマークアップすることで表現される。

<XMLタグ1>
<XMLタグ2>
単語、文、文章、図表、地図、...
</XMLタグ2 >
</XMLタグ1 >

図 1 XML タグと内容の位置関係

XML タグには、start-tag と end-tag がある。一般的に、XML タグは start-tag と end-tag で囲まれた情報（以下、「内容情報」）にメタ情報を付与する。

内容情報には、文章情報や画像情報、音声、動画情報、XML データなどがある。センター試験 XML データには、問題文や図、表の各セルの値、図表のキャプション文、数式の記号などがある。

3.2 要素、属性について

XML タグは、要素（element）と属性（attribute）情報を持つ。センター試験 XML データの例を、図 2 に示す。図 2 は、2011 年の地理 B 科目、第 1 問の問 2（一部省略）であり、左側が写真、右側が XML データである。XML データに、要素を赤四角で、属性を青四角でそれぞれ示す。

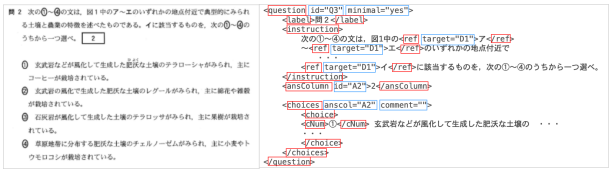

図 2 要素、属性の例

† 琉球大学大学院理工学研究科情報工学専攻,Graduate School of Engineering and Science, University of the Ryukyus
‡ 琉球大学工学部工学科知能情報コース,Computer Science and Intelligent Systems, University of the Ryukyus
*1 https://www.nii.ac.jp/news/release/2019/1118.html
*2 https://21robot.org/dataset.html
*3 Extensible Markup Language (XML) 1.1 (Second Edition) W3C Recommendation https://www.w3.org/TR/xml11/

3.2.1 要素の抽出

要素は XML タグの先頭（左端）に 1 つだけ指定でき、XML タグの名前を表す。なお、著者らはこれまでに、センター試験画像からの要素情報の抽出に物体検出モデルを適用する実験を行い、全ての要素が、平均適合率 mAP=0.8 以上の精度で抽出可能であることを示した [5]。要素情報の抽出に物体検出を適用する手法は、論文 PDF[6] や、古典・近代書籍 [7][8]、などにも適用され、高精度で抽出できることが確かめられている。

3.2.2 属性の推定

属性は、要素の補足説明であり、要素の右側に必要なだけ記述できる。また、属性は name と value からなり、name=value の形で表記される。例えば、文書内の写真に、要素名が data である XML タグを付与し、その type が “写真” であることを表現したい場合、XML データは図 3 のようになる。

```
<data  type="写真"  size="640x320" ... >
   写真データ
</data>
```

図 3　属性の例

なお、本研究では属性を @name と表記する。例えば図 3 は、「(要素 data) の @type は “写真” である」となる。表 1 に、東ロボ XML 仕様における属性（一部抜粋）を示す。

表 1　東ロボ XML 仕様における属性（一部抜粋）

要素	属性	意味
exam	@year	年度
	@subject	科目名
	@range_of_options	選択問題の範囲
	@num_of_options	選択する選択問題の数
question	@minimal	それ以上小問を含まない？
	@answer_style	回答方式（選択, 記述, ...）
	@answer_type	答えのタイプ
	@knowledge_type	必要となる知識
	@anscol	対応する解答欄の @id の値
data	@type	text, image, table, complex
cell	@rowspan	（表内に、横方向の セルの連結がある場合） 結合されているセルの数
	@colspan	（表内に、縦方向の セルの連結がある場合） 結合されているセルの数
ref	@target	参照記号の参照先の 要素の@id の値
blank	@digits	空欄に指定された桁数

表 1 の他に、@id（ref の @target などの参照のための、XML タグの識別番号）や @comment（コメント（任意））、exam における @source（問題の出題機関 (大学入試センター)）、@srcTxtURL（問題の URL（任意））、誤字脱字や欠落に対応するための @correction（文字化け箇所の正しい文字）、@string（テキスト欠落部分の正しい文字）がある。

3.3 構造について

要素が複数ある場合、要素の順序は木構造（特に内包、並列関係）を表現する。例えば、図 3 の写真が、センター試験の選択問題における選択肢であるとき、XML タグは図 4 のように表現される。

```
<choices>
   <choice> <data  type="写真">
      写真データ 1
   </data> </choice>
   <choice> <data  type="写真">
      写真データ 2
      ...
   </data></choice>
</choices>
```

図 4　構造の例

図 4 において、choices と choice は親子関係にあるため、順序を入れ替えられない。choice と data は、start-tag と end-tag の位置関係が同じため、入れ替えることは可能である。ただし、センター試験においては、choice と data の間に cNum（選択肢番号。①, ②, ...）が入ることがあり、その場合は、choice は cNum と data の親要素である必要がある。このように、XML タグの順序は重要な意味（以下、「構造情報」）を持つ。

3.4 E2E XML 生成システム

XML データは、図 5 のような配置で、要素（赤）、属性（青）、構造（タグの順序）、内容（緑）の 4 種類の情報が記述されたデータである。

```
<要素1  @属性1  @属性2  ...>
   <要素2  @属性3  @属性4  ...>
      内容
   </要素2 >
</要素1 >
```

図 5　要素、属性、構造、内容の記述方法

センター試験の画像から End-to-End で XML を自動生成するシステムの実現には、これら 4 つの情報を抽出、推定し、XML に統合する必要がある。本稿では、そのうちの属性と構造の推定実験を行う。また、E2E XML 生成システムの実装の課題について、考察を行う。

4 実験

センター試験 XML データの全情報の抽出、推定という目標のために、属性および構造情報の推定実験を行った。本実験の目的は、以下の 2 つである。

- 属性と構造の推定において利用可能な情報を、要素、属性、構造、内容の視点で整理する。
- 一部の属性と構造の情報について、機械学習アルゴリズムで推定実験を行い、推定精度のベースラインを示す。

4.1 属性、構造の推定に利用可能な情報

本実験で利用可能な情報は、「推定対象の情報が含まれるページの画像データ（以下、「ページ画像」）」と「XML データ」である。

「XML データ」のうち、属性、構造の推定どちらにも利用可能な情報を、「ページ画像」、「要素（名前、領域）」、「内容（文章）」とする。また、属性の推定には、「構造」および「推定対象ではない属性」の情報を、構

造の推定には、「属性」の情報を、それぞれ利用できるものとする。すなわち、属性または構造を推定する際、その他の情報は完全に取得できた（理想的な）状態を前提として、実験を行う。

4.2 属性推定

4.2.1 推定対象の情報

本実験では、属性 @answer_type と @knowledge_type の属性値の推定を行った。表 2 に、それぞれの属性値の例を示す。@answer_type は、問題の答えのタイプを 14 クラス（複数指定可能）で分類し、@knowledge_type は、問題の解答に必要となる知識を、10 以上のクラス（複数指定可能）で分類している。(”DM”のみ、科目ごとに異なる子クラスを持つ。英語科目には、辞書的知識（PRN, DIC_O,...）や文章読解（R_QA, R_SUM,...）、グラフ・図表の読み取り（IC_G, IC_P, ...）など、計 21 クラスがある）

表 2 属性値の例

属性名	属性値の例
answer_type	sentence, term(person, location, ...), formula,image(graph, map, ...), referenceSymbol, orthography, symbol-symbol-symbol, ... など
knowledge_type	KS(外部の知識源), RT(資料テキストからの判断), IC（グラフ、図表読み取り）, GK(一般的知識), DM(ドメイン依存モデル), ’DIC_O,GK’, ’R_QA,IC_O’ ... など

@answer_type と @knowledge_type どちらも複数ラベルを指定可能であるため、マルチラベル分類問題であるが、本実験では、XML データセットに存在するクラスのみの多クラス分類問題として推定した。

4.2.2 データセット

1987 〜 2017 年度の英語（筆記）科目の XML データからデータセットを作成した。@answer_type、@knowledge_type は、 要素 question（小問）の属性であるため、入力データは question に関係する情報が望ましいと考えられる。よって本実験では、「直前の大問の問題文（instruction）」（内容）と、「question タグで囲まれた XML データ」（要素（名前）、属性、構造、内容）を入力データとした。入力データの例を、図 6 に示す。

Unnamed: 0	answer_type	<instruction/>	contents
0	(symbol-sentence)*2	次の問い(問１・問２)において，下線部(a)・(b)の単語のアクセント(強勢)の位置が正しい...	<label>問１</label> <ansColumn id="A1">1</ansC...
1	(symbol-sentence)*2	次の問い(問１・問２)において，下線部(a)・(b)の単語のアクセント(強勢)の位置が正しい...	<label>問２</label> <ansColumn id="A2">2</ansC...
2	sentence	次の会話の下線部(1)〜(4)について，それぞれ以下の問い(問１〜４)に示された①〜④のうち...	<label>問１</label> <ansColumn id="A3">3</ansC...

図 6 属性推定における入力データの例

4.2.3 モデル

@answer_type、@knowledge_type 属性それぞれについて、LSTM 文章分類モデルで学習を行った。Tokenizer は、入力が XML データであることから、Sentence Piece[9] を使用した。Tokenize 前後のデータの例を、図 7 に示す。

```
<label>問14</label>
<data id="D20" type="text">
  <ansColumn id="A21">21</ansColumn>
</data>
```

```
['_<', 'label', '>', '問', '14</', 'label', '>',
 '_<', 'data', '_id', '=', '"', 'D', '20', '"', '_type', '=', '"', 'text', '"', '>',
  '_<', 'ansColumn', '_id', '=', '"', 'A', '2', '1', '"', '>', '2', '1</', 'ansColumn', '>',
 '</', 'data', '>', ]
```

図 7 Sentence Piece による Tokenize 前後の入力データ

4.2.4 結果と考察

@answer_type、@knowledge_type の属性値の、LSTM 文章分類モデルによる推定精度 (正解率) を表 3 に示す。

表 3 属性値の推定精度

属性名	正解率
@answer_type	0.93
@knowledge_type	0.42

この結果から、@answer_type より @knowledge_type の方が推定が難しい情報であると言える。原因はいくつか考えられるが、まず、@answer_type は 29 クラスの分類問題であるのに対し、@knowledge_type は 138 クラスあることが挙げられる。また、@answer_type は「選択肢のデータ形式」から予測できるが、@knowledge_type は「問題を解くにはどこを検索すれば良いか」という情報であり、問題文をより詳細に読む必要があることも大きく影響していると考えられる。

なお、３つ目の原因として、入力データの XML タグが問題文を詳細に読む際のノイズになる可能性もあるため、XML タグを除いての実験も行なった。結果は、正解率 が 0.45 であり、多少の精度向上は見られたが、本質的な改善には至らなかった。@knowledge_type の推定精度向上は今後の課題であるが、解決方針として、モデルの変更（マルチラベル分類モデルや、Bidirectional LSTM モデルなどの利用）を考えている。

4.3 構造推定

本実験では、構造情報（要素の順序）の推定を行った。E2E XML 生成システムにおいて、構造推定までの処理の過程は、図 8 になる予定である。

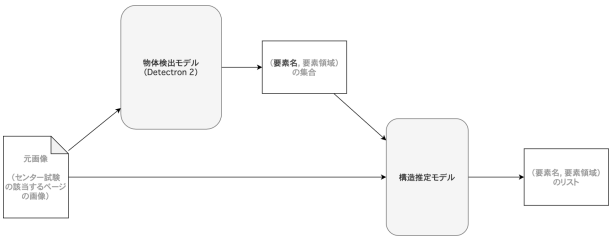

図 8 構造推定までの処理の流れ

構造推定は、要素の抽出（物体検出）モデルから得られた「要素（名前, 領域）の集合」と「元画像」を入力として、正しい順序にソートされた「要素のリスト」を出

力するタスクである。なお、本実験では要素領域 と 元画像 の情報は使用せず、要素名のみで推定を行った。

4.3.1 データセット

2011 年度の数学 IIB 科目の XML データからデータセットを作成した。教師データは、E2E XML 生成システムの入力がページ単位であることから、「ページごとの要素のリスト」とした。その際、センター試験 XML データにはページの概念がないため、ページの境界の意味で、page_num（ページ番号）タグを手作業で追加した。学習データの例を図 9 に、元画像を図 10 に示す。


```
["question", "label", "/label", "info", "/info", "info", "/info",

"instruction",
"formula", "/formula", "formula", "/formula", "formula", "/formula",
"formula", "/formula", "formula", "/formula", "formula", "/formula",
"formula", "/formula", "formula", "/formula", "formula", "/formula",
"formula", "/formula", "formula", "/formula", "formula", "/formula",
"formula", "/formula", "formula", "/formula", "formula", "/formula",
"formula", "/formula", "formula", "/formula", "formula", "/formula",
"formula", "/formula", "formula", "/formula", "formula", "ansColumn",
"ansColumn", "/formula", "formula", "/formula", "formula", "/formula",
"formula", "/formula", "formula", "ansColumn", "/formula", "formula",
"ansColumn", "ansColumn", "/formula", "formula", "ansColumn", "/formula",
"formula", "ansColumn", "ansColumn", "ansColumn", "ansColumn", "/formula",
"ref", "/ref", "ref", "/ref",
"/instruction",

"choices",
"choice", "cNum", "/cNum", "formula", "/formula",
"/choice", "choice", "cNum", "/cNum", "formula", "/formula",
"/choice", "choice", "cNum", "/cNum", "formula", "/formula",
"/choice", "choice", "cNum", "/cNum", "formula", "/formula",
"/choice", "/choices",

"page_num", "/page_num"]
```


図 9　構造推定における、教師データの例

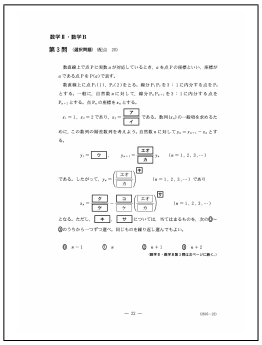

図 10　元画像（2011 年 数学 IIB, p7）

なお、数学科目において、数式は MathML 仕様（数式を表現する、東ロボ XML 仕様とは異なる XML 規格）で記述されているが、本実験では formula タグに省略した。E2E XML 生成システムにおいて要素 formula は、MathML 仕様を学習したもう一つの E2E XML 生成システムによって、MathML 形式に変換される予定である。

4.3.2 モデル

Seq2Seq モデルを用いて、要素名のソートを行った。エンコーダーおよびデコーダーは LSTM である。入力は、「シャッフルされた要素名の集合」とし、出力に対する教師データは、「正しい順序に並べられた要素名のリスト」とした。また、評価指標には BLEU を使用した。

なお、ページ内の要素の総数は最大で 80 程度であったため、Seq2Seq モデルの最大入力系列長は 100 とした。

4.3.3 結果と考察

構造（要素の順序）情報の、Seq2Seq モデルによる推定の結果を、表 4 に示す。

表 4　構造の推定精度

	精度（BLEU）
構造（要素名のみ）	0.80

本実験では、入力として 要素領域 と 元画像 の情報は使用していないが、BLEU 値 0.8 が得られた。このことから、センター試験のほとんどの情報は、一定のルールで配置（レイアウト）され、パターン化されていると考えられる。学生にとっては、情報の検索の高速化（欲しい情報がありそうな位置を推測しやすい）というメリットになるが、その具体的な効果は未知数である。

構造推定モデルの応用例としては、ある情報の集合が与えられたとき、「センター試験風な情報の配置」に並べ替える などが考えられる。

5　End-to-End モデル実現に向けて

5.1　属性、構造の推定における入出力

属性は、図 3 のように、要素の右側に name=value の形で記述されるため、属性名（name）と 属性値（value）の推定が必要になる。属性名は、東ロボ XML 仕様によって属性と要素が多対多で定義されているため、要素によって自動的に定まる。属性値は、例えば表 1 など様々な種類があるが、要素と属性名によって、属性値の推定タスク（分類、検索など）が定まるため、属性値の推定のための「モデル」と「入力データ」を決定できる。まとめると、E2E XML 生成システムにおける属性推定は、要素の抽出 → 属性名の取得 → 推定タスクの決定 → モデルと入力データの決定 → 属性値の推定という流れで実現できると考えられる。

構造は、構造の推定実験から、要素名のみで BLEU 値 0.8 の精度で推定可能であると分かった。構造の推定までの流れは、図 8 を想定している。

5.2　今後の課題

E2E XML 生成システムの開発には、少なくとも以下の 3 つの課題がある。

1 つ目は「東ロボ XML 仕様の網羅」である。特に属性について、@answer_type、@knowledge_type 以外の属性の推定は、まだ実現できていない。少なくとも、表 1 の属性全てについて、推定モデル（アルゴリズム）と入力データを用意する必要がある。また、数式に対応するため、MathML の学習も必要になる。

2 つ目は「抽出、推定、統合を制御する機能の開発」である。「属性値の推定モデルと入力データを、要素と属性名から決定する機能」や、「要素、属性、構造、内容情報の抽出・推定タスクの切り替えを制御する機能」、「抽出・推定した情報を、XML データにまとめる機能」などを開発する必要がある。

3 つ目は「センター試験の仕様への対応」である。英語、数学以外の教科への適用や、教科ごとの特徴を利用した推定モデルの調整と精度向上などがある。

参考文献

[1] 新井紀子, 松崎拓也. ロボットは東大に入れるか. **人工知能**, 27(5):463–469, 2012.

[2] 宮尾 祐介, 川添 愛. 「大学入試問題を解く」ことから見える言語, 知識, 世界理解に関する研究課題. **人工知能学会誌**, 27(5):470–478, 2012.

[3] 狩野 芳伸, 川添 愛. 「ロボットは東大に入れるか」 歴史科目の自動解答. **人工知能学会誌**, 31(6):813–819, 2016.

[4] 磯崎 秀樹, 佐藤 文香, 木野内 友梨. センター試験英語入試問題の自動 xml 化について. **言語処理学会 第 *22* 回年次大会 発表論文集**, 2016.

[5] 松本涼, 遠藤聡志. 試験問題画像からの xml メタタグ検出. **第 *19* 回情報科学技術フォーラム *(FIT)***, 2020.

[6] Minghao Li, Yiheng Xu, et al. Docbank: A benchmark dataset for document layout analysis, 2020.

[7] 青池亨, 木下貴文, 里見航, 川島隆徳. 国立国会図書館のデジタル化資料の検索・提供方法改善を目的とした技術開発とデータセット構築の取組について. **人工知能学会全国大会論文集**, JSAI2020:3Rin472–3Rin472, 2020.

[8] Zejiang Shen, Kaixuan Zhang, and Melissa Dell. A large dataset of historical japanese documents with complex layouts, 2020.

[9] Taku Kudo and John Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*, 2018.