CF-006

試験問題画像からの XML メタタグ検出
XML meta tag detection in exam questions on images

松本 涼† 遠藤 聡志‡
Ryo Matsumoto Satoshi Endo

1 はじめに

国立情報学研究所のプロジェクト「ロボットは東大に入れるか」(以下「東ロボ」) は、センター試験を計算機に自動で解かせる取り組みを行っており、2016 年時点で、全教科合計の偏差値 57.1 を達成した[*1]。

しかしながら、東ロボくんの入力は画像ではなく、独自規格の XML（以下「東ロボ XML」）データである。そして、東ロボ XML データの作成は人手で行われている [1]。このため、「End-to-End ではない」や「人件費」などの問題がある。また、そもそも XML への変換の自動化は難しく、自動化を試みた先行研究 [1] でも、全ての情報を取得できていない。

そこで本研究は、センター試験 XML データへの自動生成モデルの開発を目標として、東ロボ XML を分析して整理する。その後、東ロボの XML タグと領域を、Object Detection, Intance Segmentation モデルを用いて抽出する手法を検討する。また、実際に教師データを作成し、XML タグと対応する実データ領域がどの程度抽出できるかを検証する。

2 研究目的

2.1 東ロボくん プロジェクト

東ロボくん プロジェクトとは、「ロボットは東大に入れるか」という試みである。大学入試センター試験および二次試験形式模試を計算機に実際に解かせており、2015 年時点のセンター試験模試の合計点は、学生の平均点をすでに超えていた。2019 年のセンター試験 英語では、185 ／ 200 点 を記録している。ただし、その目的は「試験問題に正解する」ことであるため、入力データは画像形式ではなく、より扱いやすく問題のみに集中できる XML 形式であることが多い [2][3][1]。

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 データ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、独自に策定した XML 仕様（以下、「東ロボ XML 形式」）を元に記述されている。なお、センター試験 XML データの作成は人手で行われた [1]。

2.3 東ロボ XML 形式

東ロボ XML は XML 規格を元とするため、「メタデータと実データ」に分けられる。またメタデータは、図 1 のように「要素と属性」を持つ。センター試験 XML データにおける、要素の例を図 2 の赤四角に、属性の例を図 3 の青四角にそれぞれ示す。

```
<要素1 @属性1 @属性2 ...>
  <要素2 @属性3 @属性4 ...>
    実データ
  </要素2 >
</要素1 >
```

図 1 東ロボ XML における要素と属性

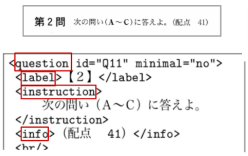


図 2 東ロボ XML における要素の例

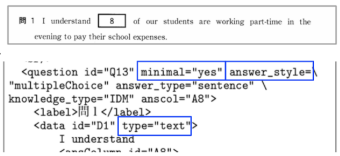


図 3 東ロボ XML における属性の例

東ロボ XML において、実データとは試験問題内の文章や画像情報であり、メタデータとは実データの試験問題における役割やデータ形式（メタ情報）、および複数の実データ間の関係性（メタ構造）を表す情報である。

以上を考慮して、東ロボ XML を、以下の 5 種類の情報の集合として定義する。

- メタデータ
  - 要素（メタデータの名前）
  - 属性（要素の補足情報）
  - 構造（並列、包含、...）
- 実データ
  - 文章情報
  - 画像情報（図表、グラフ、写真、...）

なお、メタデータはタグと呼ばれる <>,< /> 記号で表現され、タグは、図 1 のような実データを囲む形で配置される。従って、メタデータに対応する実データ領域が、試験問題画像上に必ず存在する。また、要素は、図 2 のように <> の先頭（左端）に 1 つだけ指定する。属性は、要素の補足説明であり、図 3 のように要素の右側に必要なだけ追加できる。

2.4 メタデータの要素

本研究の目的は、センター試験 XML の自動生成モデルを開発することである。そのためには、上述の 5 種類の情報全てをセンター試験画像のみから抽出し、構造情報を用いて東ロボ XML 形式に構造化する必要がある。本稿では、その第 1 段階として、メタデータの要素の抽出を行う。

東ロボ XML の要素は、全部で 24 種類ある。要素の一覧を表 1 に示す。東ロボ XML の仕様書である問題構造アノテーション仕様書[*3]を網羅しているが、東ロボ XML にはページの区別がないため、関連する要素を一

† 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://21robot.org/progress.html
*2 https://21robot.org/dataset.html
*3 https://21robot.org/dataset.html

部変更した。具体的には、page_num や ignore を追加し、question は question_max, question_min に分解した。

要素名	要素の説明
title	タイトル（教科名など）
info	配点などの情報、「全問必答」などの指示。
question_max	問題領域（ページ内最大）
question_min	問題領域（解答番号ごとに max を分割）
instruction	問題文
data	参考情報（文, 図表, グラフ, 写真...）
label	ラベル（「表 1」「図 1」「A」「ア」など）
caption	（主に表や図などに対する）説明文
note	注。
img	画像データ（写真, 地図, 概念図, グラフ...）
tbl	表（全体）
row	表中の行
cell	表中のセル
ansColumn	解答番号（ 3 など）
choices	選択肢（全体）
choice	選択肢（個々）
cNum	選択肢の番号
formula	数式
uText	下線付きテキスト
lText	ラベル付きテキスト（下線なし）
blank	空欄
ref	参照（data, uText, lTxt, note, ...）
page_num	ページ番号
ignore	ページ右下の数字 など

表 1　メタデータの要素（全 25 種類）の一覧

3　先行研究

センター試験 XML データの自動生成の先行研究として、磯崎ら [1] の研究がある。磯崎ら [1] の提案システムを、図 4 に示す。

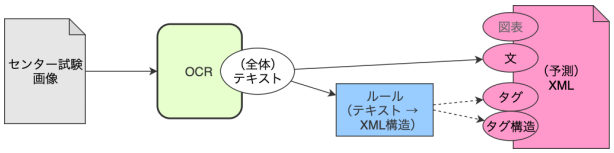


図 4　先行研究のモデル図

まず、市販の OCR ソフトによる文字認識（図 4 左側）で文章情報を抽出し、文章情報からルールベースで XML タグを付与する。

この提案システムにより、以下の成果が得られている。まず、いくつかの XML タグ（要素）について、抽出および構造化が可能であることを示した。次に、記号文字（② や 3 など）がセンター試験には多く存在し、それらが OCR ミス を起こしやすいことを示した。なお、記号文字については、前処理である程度対処出来ると主張している。具体的には、画像内の記号文字を「画像のテンプレートマッチング」で検出し、OCR ソフトで読める画像に置き換えている。

しかしながら、磯崎ら [1] の提案システムには未解決の課題も多い。例えば、

1. 英語教科の文章問題のみにしか対応できない。
2. 文章情報のみから XML タグを抽出するため、画像情報に関する XML タグは抽出できない。また、抽象的なタグ（data, tbl, question, ref など）にも対応できない。
3. 文章情報のみから XML タグを抽出するため、OCR ミスすると XML 化も失敗する可能性が高い。（文字情報に対する依存度が高い）
4. 記号文字などの OCR ミスが起きやすい文字を探すのに人手が必要になる。また、縦横のサイズの違いなど、やや高度な手作業が必要になる可能性もある。

以上の課題を緩和または解決する手法として、次節において、Object Detection のアプローチの適用を検討する。

4　アプローチ

本研究は、センター試験 XML データの自動生成を End-to-End で行うモデルの開発を目標とする。その第 1 段階としてメタデータの要素の抽出を行うが、本稿では、これを Object Detection タスクとして解くことを提案する。

4.1　Object Detection

Object Detection とは、以下の 2 つのタスクの複合タスクである [4]。

与えられた画像に対して、

- （人間・車・自転車・犬・猫などの）所定のクラスの物体インスタンスが、画像内に存在するか判定する。
- 存在する場合は、各物体インスタンスの「クラス」と「領域（一般には bounding box（外接矩形））」を返す。

東ロボ XML においては、Object Detection モデルの入力は「センター試験問題画像」であり、出力は「メタデータの要素」と「要素の実データ領域」である。

4.2　Object Detection の利点

まず、Object Detection は画像からクラスを予測するため、理論上は、画像情報に関する XML タグ全てを扱うことができる。また、文章も画像上に存在するため、文章情報に関する XML タグも全て扱うことができる。

より具体的には、

1. 全教科の全要素を扱うことができる。
2. 画像情報に関する東ロボ XML 要素を抽出できる。また、抽象的な要素（data, tbl, question, ref など）にも対応できる。
3. 文章情報からの XML タグ予測を組み合わせることで、安定的な要素の抽出ができる可能性が高い。
4. アノテーションという単純作業だけで、要素の抽出の自動化が可能となる。

などの利点がある。

5　実験

5.1　データセットの作成

2011 年度センター試験の英語、数学ⅡB、地理B（3 教科 1 年分）の画像に対して、アノテーションを行ない、データセットを作成した。アノテーションの例を図 5 に示す。

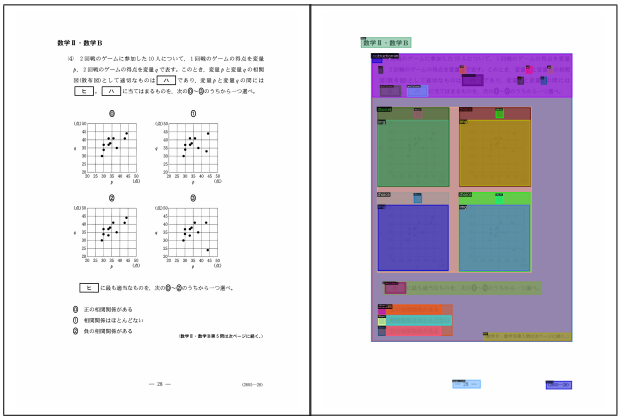


図 5　アノテーションの例

また、データセットとして英語、数学ⅡB、地理Bの 3 教科を選定した理由を、以下に示す。

- 英語
  先行研究 [1] と条件を揃えるため。
- 数学ⅡB
  数式や図形などが含まれるため。
- 地理B
  （地）図、表、写真、グラフ など、多くの画像情報が含まれるため。

データ数は 79 ページ分で、Object Detection タスクで一般的に使われる MS COCO[5] などのデータセットと比較すると、非常に少ない。ただし、センター試験はレイアウトやデザインがある程度統一されており、少数のデータでもある程度は学習可能であると予測する。

5.2　Detectron2 による学習

本稿では、Object Detectoin モデルとして Detectron2[*4] を使用する。Detectron2 は ResNeSt[6] を Backbone とし、Classification や、(Panoptic, Instance, Semantic) Segmentation, Object Detection などを学習可能なモデルである。Detectron2 は Detection と Segmentation を同時に学習、予測するため、出力は「矩形領域（の左上と右下の座標）(Detection)」および「スコア (%)」[7] と、「矩形領域 (Detection) 内の Segmentation マスク (2 値)」である。今回は Instance Segmentation の学習済みモデルを使用して、転移学習を行った。

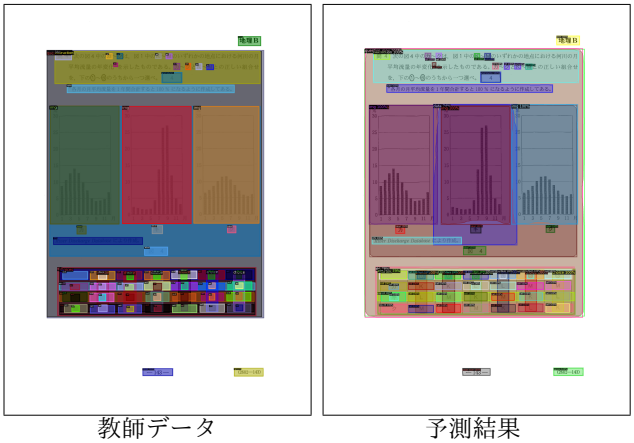
評価指標には、MS-COCO[5] Object Detection データセットの評価指標[*5]である平均適合率 (Average Precision, AP)（以下、「COCO AP」）を用いた。

*4　https://github.com/zhanghang1989/detectron2-ResNeSt
*5　http://cocodataset.org/#detection-eval

6　結果と考察

6.1　予測結果の例と考察

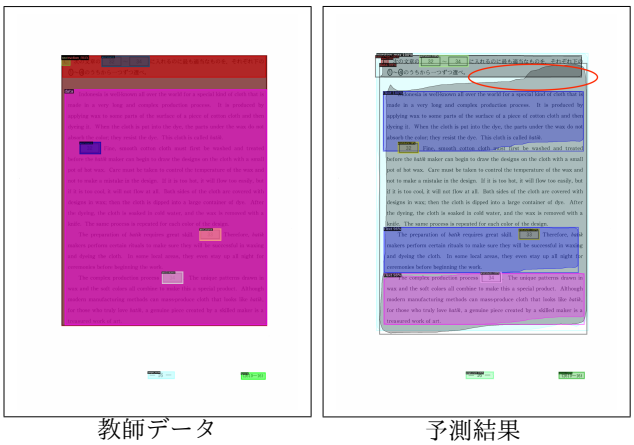
テストデータに対する予測結果の例を、図 6, 7 に示す。各図の左側が教師データの可視化、右側が予測結果の可視化である。



教師データ　　予測結果

図 6　予測結果の例 1（地理B）

図 6 は、上から「問題文」、「グラフ」、「(表形式の）選択肢」という構成のページである。右側の予測結果について、「問題文」周辺は、note まで含めてほぼ完璧に抽出できている。「グラフ」周辺は、中央の青矩形（data）が大きく間違えているが、スコアが 75%であるため、閾値を設ければ回避可能である。「(表形式の）選択肢」については、row が 2 行目のみ検出できていないが、それ以外については抽出できている。以上のことから、グラフや表などの画像情報が含まれるページでも、メタデータの要素が抽出可能であることが確認できる。



教師データ　　予測結果

図 7　予測結果の例 2（英語）

図 7 は、英語の文章読解のページである。右側の予測結果について、英語文章（data 99%）の Segmentation の際に、赤丸で示した余白部分も含んでしまっており、結果として、Detection 領域が上側に大幅にはみ出している。しかしながら、文章領域において重要なのは、Segmentation 領域に「必要十分な文字情報が含まれていること」である。従って、Segmentation 領域が上側の問題文を含んでいない限りにおいて、大きな問題ではないと考える。しかし、IText として 3 箇所が誤認識されて

いることは問題である。lText はラベル付きテキストを表すが、明らかに間違えている上に、スコアが 95%以上となっている。段落の行数がほぼ同じで、俯瞰すると 4 つの例文が並んでいるようにも見えるため、そのように見間違えた可能性が高い。このような検出ミスへの対応は、今後の課題である。

6.2 実験結果と考察

テストデータに対する Object Detection (box) および Instance Segmentation (mask) の平均適合率 (COCO AP) を、表 2 に示す。なお、AP_{75} は IoU 値 が 0.75 以上の予測領域を正答としたときの平均適合率である。

	AP	AP_{75}
box	83.405	94.366
mask	81.463	93.749

表 2 Detectorn2 の予測結果の平均適合率

表 2 より、AP は box, mask ともに 80 以上であり、高い精度であると言える。さらに、AP_{75} は 90 を超えており、非常に良い結果である。

本稿の目的は、センター試験 XML データの自動生成のための、東ロボ XML タグの要素の抽出である。この要素の役割は 2 つ存在する。1 つ目は、他のメタデータ（属性、構造）の予測に使うためであり、Detection 出力のクラスが該当する。2 つ目は、実データ（「文章」「画像」情報）の OCR および詳細解析に使うためであり、Detection 出力のクラスと Segmentation 出力の領域が該当する。よって、平均適合率および IoU 値は、出来るだけ高い精度である方が良い。

ただし、要素のクラスと領域は、他のメタデータや実データからも修正することが可能である。例えば、文の解析結果からのその領域の要素の訂正や、文が問題文であれば、ページ内の他の要素の推定が可能である。よって、Detectron2 のみで全ての要素が抽出できる必要はなく、その前提において、今回の結果は十分な精度であると考えている。

7 まとめ

本研究では、まずセンター試験画像から東ロボ XML データへの変換の自動化を目標とし、東ロボ XML を 5 つの情報の集合として定義した。次に、メタデータの要素について、Detectron2 (Object Detection, Instance Segmentation) モデルによる抽出手法を提案した。また、実際に教師データを作成し、要素と対応する実データ領域がどの程度抽出できるかを検証した。実験の結果、COCO AP は 80 以上、AP_{75} であれば 90 以上の精度が得られたため、Detectron2 モデルによって、XML タグおよび領域を（第 1 段階としては十分な精度で）抽出できることが示された。

8 今後の課題と展開

データセットの拡充と、データセットの規模と精度の関係性の調査が今後の課題である。特にデータセットの拡充については、センター試験の全教科、他年度のデータに加えて、漢・数・英検、TOEIC、その他資格試験についてもアノテーションを行い、東ロボ XML の仕様自体も見直していく予定である。

また、センター試験 XML データの自動生成の次の段階として、実データ（文章、画像情報）の抽出の自動化と、他のメタデータ（属性、構造）のうちの構造の予測を行う予定である。

まず、文章情報には OCR を行う必要があるが、本稿の成果である、要素が同一である領域に対して OCR を行うことができる。すなわち、OCR ミスの原因となる画像情報や特殊文字について、それらを除いた領域で OCR を行う。また、画像情報についても同様に、解析しやすい領域に限定して解析を行うことができる。さらに、今回は図 8 のような、図表内の label（「A」「ア」などの記号文字）についてもアノテーションを行なっているため、label 以外の情報の抽出に専念でき、構造化の骨格となることが期待される。

図 8 画像情報内の label 領域の予測結果

次に、他のメタデータ（属性、構造）について、構造は XML 構造（包含、順番など）を予測するタスクとなるが、Detectron2 の Detection 結果から得られる要素のクラスと領域の情報から、十分に予測できると考えている。最後に、属性については、要素や構造と比べて多種多様な上に、実データを解析しないと抽出不可能な概念も存在するため、予測および抽出は最も難しいと考えられる。よって、属性の抽出については、実データの取得が自動化できてから取り組む予定である。

参考文献

[1] 磯崎 秀樹, 佐藤 文香, and 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.

[2] 狩野 芳伸 and 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.

[3] 宮尾 祐介 and 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.

[4] Li Liu, Wanli Ouyang, Xiaogang Wang, Paul Fieguth, Jie Chen, Xinwang Liu, and Matti Pietikäinen. Deep learning for generic object detection: A survey. *International journal of computer vision*, 128(2):261–318, 2020.

[5] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[6] Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Zhi Zhang, Haibin Lin, Yue Sun, Tong He, Jonas Muller, R. Manmatha, Mu Li, and Alexander Smola. Resnest: Split-attention networks. *arXiv preprint arXiv:2004.08955*, 2020.

[7] 直人 井上, 諒佑 古田, 俊彦 山崎, and 清晴 相澤. 類似シーン画像の統計情報に基づく物体検出のフィルタリング. In 第 *78* 回全国大会講演論文集, volume 2016, pages 215–216, mar 2016.