

Text Extraction From PDF

Problem Statement

Accurate metadata extraction from born-digital academic PDFs depends on first obtaining a complete, correctly ordered textual representation of each document’s contents. This task is complicated by:

- **Heterogeneous publisher templates:** (e.g., IEEE, Springer) that differ in font usage, heading conventions, and artifact placement.
- **Variable page layouts:** single-column, multi-column, or hybrid—that break naive top-to-bottom reading order.
- **Parser-specific limitations:** in handling encoded characters, ligatures, figures, footnotes, and incremental updates inside the PDF file structure.

The immediate goal is therefore to identify which open-source parsing library - **PyMuPDF**, **pypdfium2**, **pdfminer.six**, or **PyPDF2** delivers the highest fidelity text extraction on *first pages* of arXiv papers representing diverse layouts and styles.

“Highest fidelity” will be quantified against a manually verified ground-truth transcript using four automated text-similarity measures:

1. **Character Error Rate (CER)** – percentage of character insertions, deletions, and substitutions.
2. **Word Error Rate (WER)** – percentage of word-level insertions, deletions, and substitutions.
3. **BLEU Score** – n-gram precision-based overlap between extracted text and reference.
4. **ROUGE-L** – longest-common-subsequence recall and precision, capturing sentence-level ordering.

Data

Aspect	Detail
Source dataset	DocBank —token-level annotations (bounding box, font, text) for arXiv papers published 2014-2018.
Rationale	<i>Born-digital academic PDFs</i> with ground-truth tokens; diverse layouts and publisher styles.
Selection method	<div>1. Queried arXiv API for 5 disciplines (CS, Stats, Math, EESS, Econ).</div> <div>2. Cross-referenced results with DocBank to keep only papers with existing annotations.</div> <div>3. Downloaded PDFs + annotations for 101 papers (distribution: CS 31, Stat 22, Math 18, EESS 18, Econ 12).</div>
Reading-order ground truth	<div>1. YOLOv12 segments each page into logical blocks.</div> <div>2. Group DocBank tokens by detected segments.</div> <div>3. Feed segments + unassigned tokens into LayoutReader (LayoutLM-based seq2seq) to predict inter-segment reading order.</div> <div>4. Concatenate ordered tokens to obtain full reference text per page.</div>

token | boundingbox ((x0, y0), (x1, y1)) -> (x0, y0, x1, y1) | color (R, G, B) | font | label

Inverse	201	111	275	134	0	0	0	RQRKXE+NimbusRomNo9L-Medi	title	
Reinforcement			280	111	428	134	0	0	RQRKXE+NimbusRomNo9L-Medi	title

Page—header 0.43
Inverse Reinforcement Learning via Deep Gaussian Processes

Text 0.82

Ming Jin* Andreas Damianou¹
EECS, UC Berkeley, USA Amazon.com, Cambridge, UK
jinming@eecs.berkeley.edu andreas.damianou@amazon.com

Text 0.74

Pieter Abbeel
EECS, UC Berkeley, USA
pabbeel@berkeley.edu

Text 0.77

Costas Spanos
EECS, UC Berkeley, USA
spanos@berkeley.edu

Page-header 0.76

Section-head Text 0.96
Abstract observing its demonstration

Text 0.95

header 0.40

observing its demonstrations or trajectories in the task. It has been successfully applied in scientific inquiries, e.g., animal and human behavior modeling (Ng et al. 2000), as well as practical challenges, e.g., navigation (Ratliff et al. 2006; Abbeel et al. 2008; Ziebart et al. 2008) and intelligent building controls (Barrett and Linder 2015). By learning the reward function, which provides the most succinct and transferable definition of a task, IRL has enabled advancing the state of the art in the robotic domains (Silver et al. 2007).

Text 0.96

er et al. [2007].

Previous IRL algorithms treat the underlying reward as a linear (Abbeel and Ng [2004], Ratliff et al. [2006], Ziebart et al. [2008], Syed and Schapire [2007], Ratliff et al. [2009]) or non-parametric function (Levine et al. [2010], [2011]) of the state features. Main formulations within the linearity category include maximum margin (Ratliff et al. [2006]) which presupposes that the optimal reward function leads to maximal difference of expected reward between the demonstrated and random strategies, and feature expectation matching (Abbeel and Ng [2004], Syed et al. [2008]) based on the observation that it suffices to match the feature expectation of a policy to the expert in order to guarantee similar performances. The reward function can be also regarded as the parameters for the policy class, such that the likelihood of observing the demonstrations is maximized with the true reward function, e.g., the maximum likelihood (Ziebart et al. [2008]).

Text 0.97

ebart et al. [2008].

As the representation power is limited by the linearity assumption, nonlinear formulations (Levine et al. [2010]) are proposed to learn a set of composite features based on logical conjunctions. Non-parametric methods, pioneered by (Levine et al. [2011]) based on Gaussian Processes (GPs) (Rasmussen [2006]), greatly enlarge the function space of latent reward to allow for non-linearity, and have been shown to achieve the state of the art performance on benchmark tests, e.g., object world and simulated highway driving (Abbeel and Ng [2004], Syed and Schapire [2007], Levine et al. [2010], [2011]). Nevertheless, the heavy reliance on predefined or handcrafted features becomes a

Section—header 0.82

1 INTRODUCTION

Text 0.88

The problem of inverse reinforcement learning (IRL) is to

Text 0.71

on that the agent subsumes by

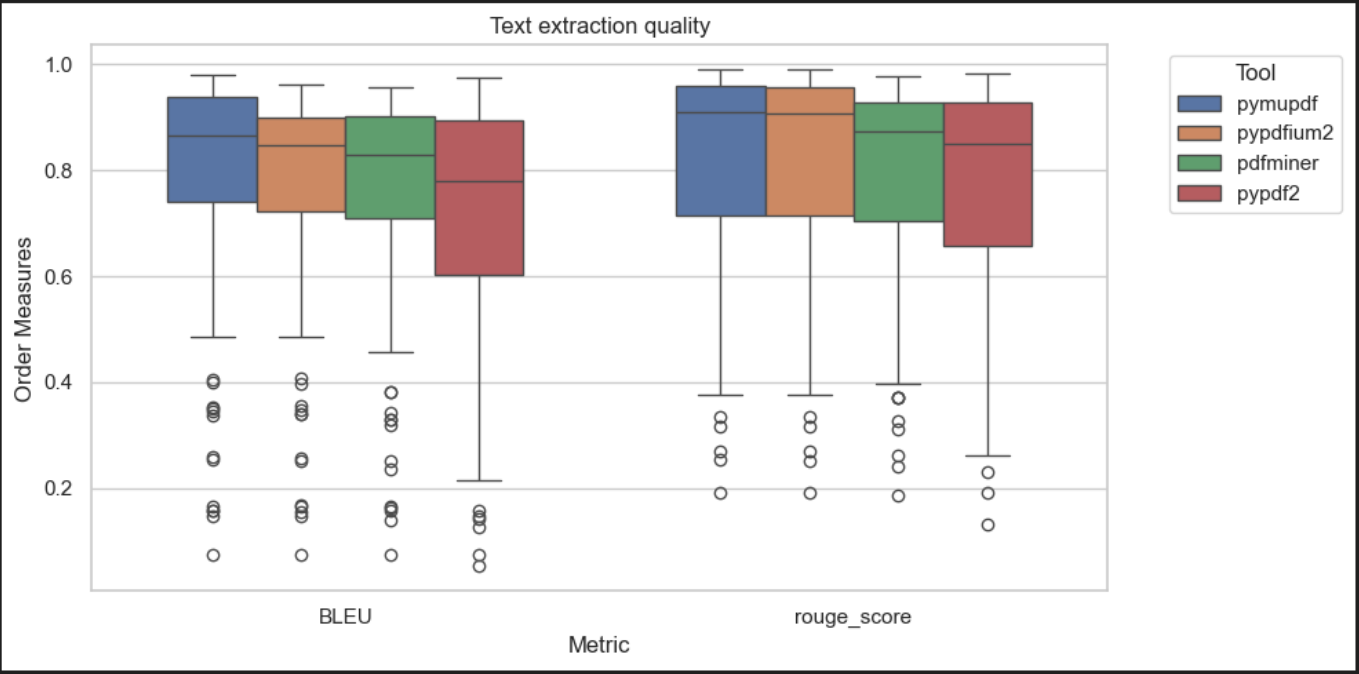
This research is funded by the Republic of Singapore's National Research Foundation through a grant to the Berkeley Education Alliance for Research in Singapore (BEARS) for the Singapore-Berkeley Building Efficiency and Sustainability in the Tropics (SinBerBEST) Program. BEARS has been established by the University of California, Berkeley as a center for research and education in Singapore.

Text 0.46

Results

PyMuPdf

- avg_BLEU=0.7817
- avg_rouge_score=0.8086



PyMuPdf

- avg_CER=0.2805
- avg_WER=0.3544

