JLR2024 日本語言語資源の構築と利用性の向上

学術論文のPDF文書からのテキスト抽出における課題

福田健人 (放送大学教養学部情報コース)

2024-03-15

概要

- モチベーション
  - 日本語LLMのための高品質な学習データ源として PDF文書を使いたい
  - ファーストステップとして学術論文を対象としたい
- 結論
  - 日本語の横書き2カラムレイアウトの論文に対しては、ndl_layout（レイアウト要素の検出）とPyMuPDF（各要素からのテキスト抽出）を組み合わせた手法が有力
- 未解決の課題
  - 図表、数式からの安定した情報抽出
- 長期的な課題
  - 学術論文以外にも安定して適用できる高速な手法の開発

LLMの学習データセット

- 書籍のスキャンデータ - The Pileなど
- Webサイトのクロールデータ - Common Crawlなど
  - **HTML**から本文を抽出して利用
  - （Wikipediaの記事データ: 質が高く、ライセンスが明確で、収集が容易）

PDFは？

- ほとんどの場合、フィルタリングによって破棄している
- なぜか？
  - 本文抽出が困難
    - マークアップ言語であるHTMLと比較して、組版記述ファイルである PDFでは本文抽出が難しい
  - 総数が少ない
    - CommonCrawlではtext/htmlが98%程度に対してapplication/pdfは1%程度
- 本当に破棄していいのか？
  - 公刊資料などが多く、概して良質なデータであることが期待できる
  - 元のテキスト量は1/100でも、フィルタリングの歩留まりが 100倍だとしたら？
  - 参考: SwallowコーパスではCommonCrawlの日本語データを6%程度までフィルタリングしている

なぜ学術論文？

- 解析が容易である
  - レイアウトが単純で共通性が高い
- メタデータが充実している
  - 文書の分野、発表年などのメタデータが整備されている
  - 「直近5年の有機化学に関する文書だけを集中的に利用したい」といったニーズに応えられる
- （許可さえ得られれば）クローリングが容易
  - 論文リポジトリは一覧ページなどが使いやすい形で提供されている
  - ※クローリングを原則禁止しているリポジトリもある

抽出処理のデモ

短期的な課題

- 図表と数式の扱い
  - 現状では適切なパース処理が実装できていないため、すべて無視している
  - 特に図表はバリエーションが多く、情報抽出は難しい
- 処理速度の向上
  - Google Colab環境のV100インスタンスでは1ページの処理に5秒程度を要する

長期的な課題

- 学術論文以外にも安定して適用できるかどうかの検証が必要
  - 直近6年間にJ-Stageに収録された論文は 2万本あまり
  - トークン数にすると約 0.1B程度
  - LLMの学習データとしては十分でない
  - ➔ 論文以外の公刊文書についてもテキスト抽出ができるか検証したい