PDF/OCR用の機械翻訳モデルを作ってみた

高橋 怜士（@voleneko）

2022年3月4日

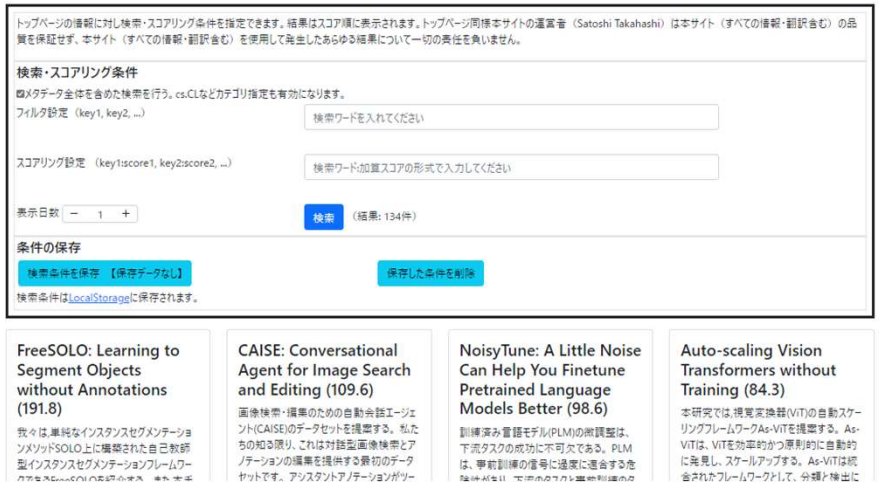
自己紹介: 高橋 怜士（たかはし さとし）

- 三菱総合研究所に所属しています。
- 趣味で開発・実験した内容を紹介します。
  品質よりも個人の楽しさを重視した開発であり
  技術選定は合理的とは限りません。
- 記載内容は個人の見解であり、会社・所属機関の意見を代表するものではありません。

背景① fugumt.comの紹介

arXivの新着情報を翻訳するサイトを運営

- arXivからAI関連のデータを日々100件程度取得・翻訳
- CC-0, CC-BYなどライセンスが許せば**PDF全体を翻訳**
- 独自モデルを使って翻訳（Marian-NMTを利用）
- 検索・スコアリング、VRモードなど多機能



背景② fugumt.comの問題

テキスト抽出時、単語認識に失敗することがある

処理の流れ

1. pdfminer によるテキスト抽出 → **2. nltk/pySBD による文分割** → **3. Marian-NMT(FuguMT) による英語→日本語翻訳**

```
output_string = StringIO()
with open(pdf_file, 'rb') as in_file:
    parser = PDFParser(in_file)
    doc = PDFDocument(parser)
    rsrcmgr = PDFResourceManager()
    device = TextConverter(rsrcmgr, output_string, laparams=LAParams(boxes_flow=0.3, line_margin=1.0))
    interpreter = PDFPageInterpreter(rsrcmgr, device)
    for idx, page in enumerate(PDFPage.create_pages(doc)):
        interpreter.process_page(page)
print(output_string.getvalue().replace('.', '¥n'))
```

```
TrustworthyAI:FromPrinciplestoPracticesBOLI,JDAIResearch,ChinaPENGQI,JDAIResearch,USABOLIU,JDAIResearch,USASHUAIDI,
However,manycurrentAIsystemswerefoundvulnerabletoimperceptibleattacks,biasedagainstunderrepresentedgroups,lackingin
,whichnotonlydegradesuserexperiencebuterodesthesociety' strustinallAIsystems
Inthisreview,westrivetoprovideAIpractitionersacomprehensiveguidetowardsbuildingtrustworthyAIsystems
WefirstintroducethetheoreticalframeworkofimportantaspectsofAItrustworthiness,includingrobustness,generalization,exp
Wethensurveyleadingapproachesintheseaspectsintheindustry
TounifythecurrentfragmentedapproachestowardstrustworthyAI,weproposeasystematicapproachthatconsiderstheentirelifecyc
Inthisframework,weoffe
```

TounifythecurrentfragmentedapproachestowardstrustworthyAI,weproposeasystematicapproachthatconsiderstheentirelifecycleofAIsystems,rangingfromdataacquisitiontomodeldevelopment,todevelopmentanddeployment,finallytocontinuousmonitoringandgovernance

下記PDFを利用しています

Bo Li, Peng Qi, Bo Liu, Shuai Di, Jingen Liu, Jiquan Pei, Jinfeng Yi, & Bowen Zhou. (2021). Trustworthy AI: From Principles to Practices.

https://arxiv.org/abs/2110.01167

※当該PDFからpdfminerによるテキスト抽出を実施

課題解決のアプローチ

失敗に対応可能な機械翻訳モデルを構築する

①パラメータ調整など

- pdfminerのパラメータを調整して正しく単語抽出ができるようにする
- ライブラリを変更する

②テキストの修正

- 文分割前にテキスト（英語）を修正する
- 「Input:スペースを除去したテキスト」→「Output: スペース除去前のテキスト」となるモデルを構築

③翻訳エンジンの改善

- スペースがなくても動作する翻訳エンジンを構築する
- データセットを上記前提で拡張・合成する

何を入れても対応できる機械翻訳は便利そうという趣味的理由で選択

1. pdfminer によるテキスト抽出 → **2. nltk/pySBD による文分割** → **3. Marian-NMT(FuguMT) による英語→日本語翻訳**

合成データ作成方法

OCRエラーも考慮して拡張（合成）データを構築
処理時間の観点からnlpaugを使用

①Genalog

- https://github.com/microsoft/genalog
- テキストをドキュメント形式で画像化しAzureOCRを実行、テキスト対応を取ることが可能

②NL-Augmenter

- https://github.com/GEM-benchmark/NL-Augmenter
- 多様なpluginを通してNLPのデータ拡張が可能なフレームワーク
- テキスト→画像化→変形→ Tesseract OCRというデータ拡張pluginが存在

③nlpaug

- https://github.com/makcedward/nlpaug
- 多数のデータ拡張手法に対応したライブラリ
- OCRエラーをシミュレーションする形のデータ拡張が可能

計算時間の観点で実際にOCRを行うアプローチは選択できなかった

翻訳エンジン構築の流れ

一般的な手順でニューラル機械翻訳モデルを構築

通常の構築フロー（ベースライン）　　**本件の構築フロー（PDF/OCR用）**

①データ作成

- Creative Commonsライセンスのデータ＋独自収集データ 15M対訳ペア
  ※データ元はhttps://github.com/s-taka/fugumtをご参照
  ※データクリーニングはhttps://staka.jp/wordpress/?p=571の通り

- スペース除去とOCRエラーを再現したテキスト（英語）を追加

②データ拡張

③事前学習

- 英→日、日→英データを混ぜた事前学習（検証するとBLEUで約+1ptの効果）
  ※Liang Ding, Di Wu, & Dacheng Tao. (2021). Improving Neural Machine Translation by Bidirectional Training.

④モデル構築

- Marian-NMT (transformer) + sentence pieceによるモデル構築
  ※計算資源の問題でBack translationは行わない

⑤評価

- 独自テストセットに対してBLEUで評価
  ※上記テストデータが学習に含まれない（記号除去・小文字化して一致する文がない）事は確認済み
  ※WMT21 news taskなどを使いたいが、ライセンスが調べきれなかったため

結果

両データでPDF/OCR用モデルが優れた性能

→ 複数処理を1モデルにする対応は有効（？）

ベースライン	VS	PDF/OCR用
BLEU=17.8	オリジナルデータ	BLEU=18.6
BLEU=　3.2	スペース除去データ	BLEU=18.5

※BLEU計測はdetokenize後の訳文と正解に対して「sacrebleu --tokenize ja-mecab –b」

- デモサイト: https://devneko.jp/demo/
  - No.1:fugumt.comモデル　No.2:大文字小文字を区別しないモデル
  - **No.3:ベースライン　　No.4:PDF/OCR用モデル**
- モデルファイル: https://fugumt.com/pdf_ocr_model.zip
  - CC BY-SA 4.0、研究用を目的に公開

作者は本モデルの動作を保証しません。
本モデルを使用して発生したあらゆる結果について一切の責任を負いません。

翻訳例

PDF/OCR用モデルは妥当な翻訳が可能

英語テキスト

TounifythecurrentfragmentedapproachestowardstrustworthyAI,weproposeasystematicapproachthatconsiderstheentirelifecycleofAIsystems,rangingfromdataacquisitiontomodeldevelopment,todevelopmentanddeployment,finallytocontinuousmonitoringandgovernance.

ベースライン

信頼に値するAI、AIシステムの再ライフサイクルを考察するweproposeasystematicapproach、dataacquisitiontomodel Modeling、Development and Deployment、finallytocontinuousmonitoring andgovernanceに現在のフラグメンテーションされたapproachestounify。

PDF/OCR用

信頼できるAIに対する現在の断片化されたアプローチを統一するために、データ取得からモデル開発、開発とデプロイ、さらには継続的な監視とガバナンスまで、AIシステムのライフサイクル全体を考慮した体系的なアプローチを提案する。

下記PDFのテキストを利用しています

Bo Li, Peng Qi, Bo Liu, Shuai Di, Jingen Liu, Jiquan Pei, Jinfeng Yi, & Bowen Zhou. (2021). Trustworthy AI: From Principles to Practices.

https://arxiv.org/abs/2110.01167

※当該PDFからpdfminerによるテキスト抽出を実施

※抽出したテキストを機械翻訳

今後の取り組み

機械翻訳モデル、データ拡張・合成について特殊な用途に対応したモデルを構築（予定）

機械翻訳モデル

SQuADなど位置情報が重要なデータに対しその情報を壊さない機械翻訳モデル

データ拡張・合成

PDF翻訳に採用予定のLayout-Parser + Tesseract OCRに特化したOCRエラー再現データ拡張・合成ライブラリ