



# 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)

- ① 三菱総合研究所に所属しています。
- ② 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。
- ③ 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。

## 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

トップページの情報に対し検索・スコアリング条件を指定できます。結果はスコア順に表示されます。トップページ同様本サイトの運営者 (Satoshi Takahashi) は本サイト (すべての情報・翻訳含む) の品質を保証せず、本サイト (すべての情報・翻訳含む) 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。

**検索・スコアリング条件**

メタデータ全体を含めた検索を行う。cs.CLなどカテゴリ指定も有効になります。

フィルタ設定 (key1, key2, ...)

検索ワードを入れてください

スコアリング設定 (key1score1, key2score2, ...)

検索ワード+加重スコアの形式で入力してください

表示回数 - 1 +

検索 (結果: 134件)

**条件の保存**

検索条件を保存 【保存チェックなし】

保存した条件を削除

検索条件はLocalStorageに保存されます。

<b>FreeSOLO: Learning to Segment Objects without Annotations (191.8)</b> 我々は、単純なインスタンスセグメンテーションメタデータSOLO上に構築された自己教師型インスタンスセグメンテーションフレームワークを提案する。このフレームワークは、従来のSOLOよりも、より少ないメタデータで、より高い精度を実現する。	<b>CAISE: Conversational Agent for Image Search and Editing (109.6)</b> 画像検索・編集のための自動会話エージェント(CAISE)のデータセットを提案する。私たちの知る限り、これは対話型画像検索とアプリケーションの提案を提供する最初のデータセットです。アシスタントアプリケーションが	<b>NoisyTune: A Little Noise Can Help You Finetune Pretrained Language Models Better (98.6)</b> 訓練済み言語モデル(PLM)の微調整は、下流タスクの成功に不可欠である。PLMは、事前訓練の番号に過度に適合する危険性を伴う。本稿では、事前訓練の	<b>Auto-scaling Vision Transformers without Training (84.3)</b> 本研究では、視覚変換器(ViT)の自動スケールアップフレームワークAs-ViTを提案する。As-ViTは、ViTを効率的かつ原則的に自動的に発見し、スケールアップする。As-ViTは統合されたフレームワークとして、分類と検出に
---	---	--	--

<b>(7) Learning Multi-Object Dynamics with Compositional Neural Radiance Fields [63.4]</b> 本稿では、暗黙的オブジェクトエンコーダ、ニューラルレージアンサーフィールド(NeRF)、グラフニューラルネットワークに基づく画像観測から構成するモデルを学習する手法を提案する。NeRFは3D以前の強みから、シーンを表現するための一般的な選択肢となっている。提案手法では、学習した潜在空間にRTを応用し、そのモデルと暗黙的オブジェクトエンコーダを用いて、暗黙的空間を効率的かつ効果的にサンプリングする。	<b>(8) FreeSOLO: Learning to Segment Objects without Annotations [191.8]</b> 我々は、単純なインスタンスセグメンテーションメタデータSOLO上に構築された自己教師型インスタンスセグメンテーションフレームワークであるFreeSOLOを紹介する。また、本手法では、複雑なシーンからオブジェクトを分離して検出する、新たなローカライズ対応事前学習フレームワークを提案する。	<b>(9) Pretraining with Words [50.1]</b> 本稿では、単語ではなく1で構文を表現する。その結果、テキストや機械読解の大半は、中国語でWord
<b>(15) First is Better Than Last for Training Data Influence [44.9]</b> 既存の手法はモデルパラメータによる影響の度合いに基づいている。そこで我々は、最後の層ではなく、単語埋め込み層上で動作させるTraIn-WEという手法を提案する。また、TraIn-WEは、3つの言語分類タスクにおけるケース削除評価において、最終層に適用される他のデータ駆動手法よりも4倍に向上させることを示した。	<b>(16) On Learning Mixture Models with Sparse Parameters [44.3]</b> 本研究では、高次元入力/出力メタベクトルの混合について検討し、これらのベクトルの両方と混合成分について考察する。混合空間の次元に相対的にサンプル複雑性が依存する回帰変換のための効率的なアルゴリズムを提供する。	<b>(17) Sample Efficient</b> まず、データ拡張整合性に効率的であることが構成するDAC解析の
<b>(3) Cognitive Semantic Communication Systems Driven by Knowledge Graph [33.3]</b> 知識グラフを利用した認知意味コミュニケーションフレームワークを提案する。意味情報抽出のシンプルで汎用的で解釈可能なソリューションを開発した。提案システムは、データ伝送率の信頼性の観点から、他のベンチマークシステムよりも優れている。	<b>(24) No-Regret Learning in Games is Turing Complete [33]</b>	

## 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

```
9 output_string = StringIO()
10 with open(pdf_file, 'rb') as in_file:
11     parser = PDFParser(in_file)
12     doc = PDFDocument(parser)
13     rsrcmgr = PDFResourceManager()
14     device = TextConverter(rsrcmgr, output_string, laparams=LAParams(boxes_flow=0.3, line_margin=1.0))
15     interpreter = PDFPageInterpreter(rsrcmgr, device)
16     for idx, page in enumerate(PDFPage.create_pages(doc)):
17         interpreter.process_page(page)
18 print(output_string.getvalue().replace('.', '\n'))
19
```

TrustworthyAI:FromPrinciplestoPracticesBOLI,JDAIRResearch,ChinaPENGQI,JDAIRResearch,USABOLI,USASHUAI, However,manycurrentAI systems were found vulnerable to imperceptible attacks, biased against underrepresented groups, lacking in ,whichnotonlydegradesuserexperiencebuterodesthesociety's trust in AI systems  
Inthisreview,westrivetoprovideAIpractitionersacomprehensiveguidetowardsbuildingtrustworthyAI systems  
WefirstintroducetheoreticalframeworkofimportantaspectsofAI trustworthiness, including robustness, generalization, exp  
Wethensurveyleadingapproachesintheseaspectsintheindustry  
TounifythecurrentfragmentedapproachestowardstrustworthyAI, weproposeasystematicapproachthatconsiders theentirelifecyc  
Inthisframework, weoff

TounifythecurrentfragmentedapproachestowardstrustworthyAI, wepr  
oposeasystematicapproachthatconsiders theentirelifecycleofAI system  
s, ranging from data acquisition to model development, to development an  
d deployment, finally to continuous monitoring and governance

### 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>

ライセンス: Creative Commons — Attribution 4.0 International — CC BY 4.0

※当該PDFからpdfminerによるテキスト抽出を実施

## 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

## 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

### ① Genalog

- <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

### ② NL-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

### ③ nlpaug

- <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

## 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

②データ拡張

- 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

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、ライセンスが調べきれなかったため

## 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」

- ① 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**
- ② モデルファイル: [https://fugumt.com/pdf\\_ocr\\_model.zip](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。

## PDF/OCR用モデルは妥当な翻訳が可能

### 英語テキスト

To unify the current fragmented approaches toward trustworthy AI, we propose a systematic approach that considers the entire lifecycle of AI systems, ranging from data acquisition to model development, to development and deployment, finally to continuous monitoring and governance.

### 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

we propose a systematic approach、data acquisition to model Modeling、Development and Deployment、finally to continuous monitoring and governanceに現在のフラグメンテーションされたapproaches to unify。

### 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>

ライセンス: [Creative Commons — Attribution 4.0 International — CC BY 4.0](#)

※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

## 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