

# 日本語を母国語としない人向けの表現簡易化変換システム

金沢工業大学 中沢研究室 4EP1-05 上田祐輝

## 研究の背景

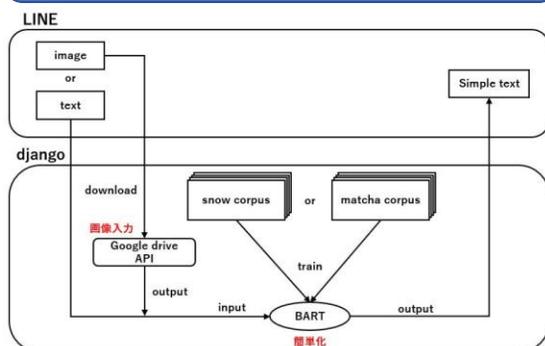
### 現状

外国からの移住者が増加する中で「やさしい日本語」が注目されており、ニーズも確認できるが、実際に使用される場合はまだ少ない。

### 問題点

難解文をやさしい日本語に変換するサービスは少なく、テキスト入力のみに対応しているため、利用者が難解な文を書き起こす必要がある。

## システム概要



### 1. Lineからテキストを取得

→画像での入力を受けた際は、OCR機能を利用し、画像内のテキストを取得。

### 2. やさしい日本語化

→強化学習を行ったBARTを用いて入力文をやさしい日本語に変換

**入力** 土足厳禁

**出力** 靴を脱いでください

### 3. やさしい日本語を返信

→文法的に問題のない出力を選定

## 実装方法

### botの作成

フレームワークであるdjangoを用いてOCRと推論機能を実装する。

### OCRの実装

Google drive APIを用いてGoogle driveのOCR機能を用いる。

### BARTの強化学習

難解な日本語文と平易な日本語文の文対を有する以下の2つのコーパスのどちらか、または両方を用いて強化学習を行う。

#### 1. snowコーパス

83k文対を有する。大学生と有志によって作成されており、文対量は多いが品質は保証されていない。

#### 2. matchaコーパス

15k文対を有する。日本語教育能力検定試験に合格し、技能実習生が対象の日本語教師を経験した専門家によって作成されており、品質が保証されている。

## 評価方法

・出力文がやさしい(平易な)日本語であるか。

上記をSARIで評価する。

[SARI = 1/3(追加F1 + 削除F1 + 保持F1)]

## 今後の展望

追加事前学習を行うことで同様のデータセットでもBARTの平易化タスクでの性能が向上するという論文の手法を試す。文頭の出力が安定しない問題解決法を模索する。