

S16 米国特許のFタームへの自動分類

中央大学 理工学部 経営システム工学科 蜂須賀 笙太、難波 英嗣

背景

日本と米国の分類体系の違い

米国特許
IPC | 約6万個

日本特許
Fターム | 約40万個

↓

日本特許の分類体系は
米国特許よりも細分化されている

目的

米国特許にFタームを付与する

↓

米国特許の詳細かつ的確な
検索が可能になる

利用データ

翻訳機

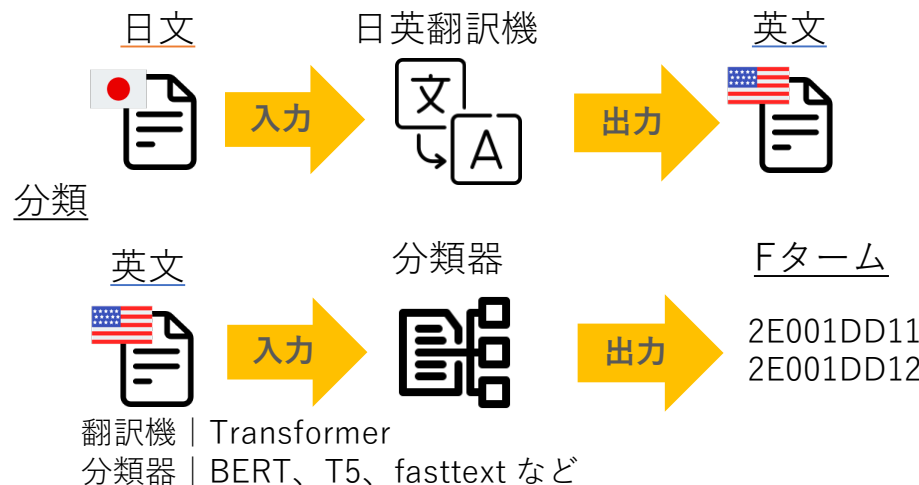
- NTCIR-10 特許機械翻訳テストコレクション

分類器

- J-PlatPat 2019~2021年 日本特許
- USPTO Bulk Data 2019~2021年 米国特許
(日本特許の Patent Family)

分類の手法

翻訳 | 学習時のみ



教師データとしてFタームが必要になるので、日本特許とパテントファミリー^{*1}を利用する

学習データ | 日本特許(約30万個)

テストデータ | 米国特許(約2万個)

(日本特許の Patent Family)

入力となるテキストには要約文を利用し、**Fターム**は日本特許の利用データで500以上出現する**2220個**を用いる

^{*1} Patent Familyとは同じ発明で複数の国に出願している特許の集まりのこと

分類結果

表1 BERT 分類結果

	precision	recall	F1
micro	0.538	0.315	0.397
macro	0.306	0.219	0.224

BERTのbert-base-uncasedを使用

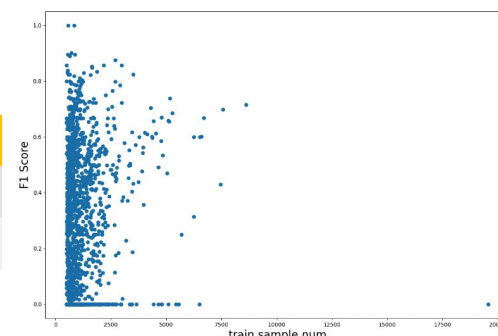


図1 各Fタームの学習データ数とF値の散布図

recallが小さいため、付与されたFタームの数が実際より少ない

各FタームのF値は学習データ数に依存せず、各Fタームごとに難易度が異なる

今後の課題

- さまざまなモデルの実装
T5、fasttextなど
- データセットの調整
データの追加など
- recallの値の改善
- 各Fタームの分類結果の優劣の原因分析