機械学習を用いた学術分野の自動分類とWEBアプリによる実装

福島大学 地域未来デザインセンター　荻 多加之

導入

各研究の学術分野を分類することは、研究費申請や研究力分析・評価において重要である。しかし、専門外の人間が大量の研究課題を客観的に分類することは非常に困難である。一方で、機械学習が研究力分析・評価にも適用され始めている（Goh et al. 2020）。

そこで本研究では科研費の概要テキストのデータに基づき、機械学習によって各課題の学術分野をどの程度の精度で推定可能であるか検証した。

方法

1. モデルの訓練

科研費データベースから**2019～2021年度**開始の基盤BC（48,680件）をダウンロード

↓

ダウンロードしたデータから「**研究開始時の研究の概要**」と「**小区分**」を取り出す

↓

小区分を対応する11の**大区分に変換**し、訓練データを作成する

↓

事前学習済みの自然言語処理モデルをダウンロード

↓

訓練データを用い、事前学習済みのモデルの**追加学習（ファインチューニング）**を行う

2. 新規課題による分類テスト

科研費データベースから**2022年度**開始の基盤BC（12,952件）をダウンロード

↓

ダウンロードしたデータから「**研究開始時の研究の概要**」と「**小区分**」を取り出す

↓

小区分を対応する11の**大区分に変換**し、テストデータを作成する

↓

作成した追加学習済みモデルを用い、テストデータの「概要」文字列から**大区分を推定**

↓

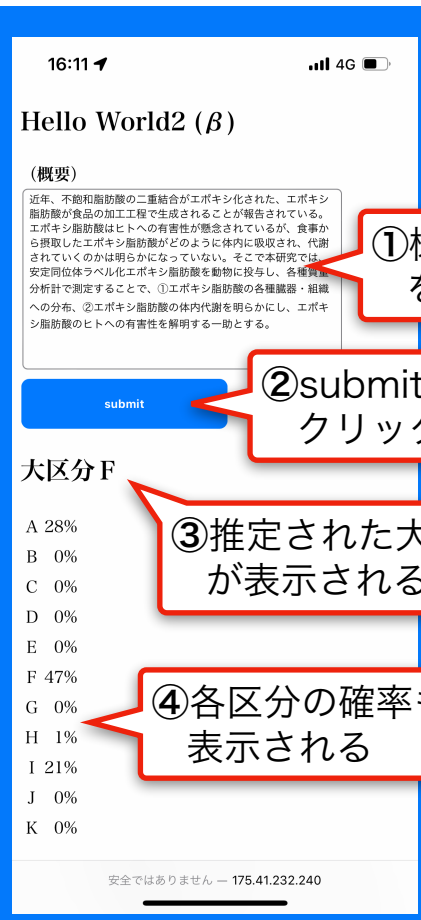
正解の大区分を参照し、**分類精度を検証**する

参考情報

- 2018年にGoogleから発表された自然言語処理モデル「BERT」を用いた。
- 東北大学 乾研究室で開発された訓練済み日本語BERTモデルを利用した。
- 我妻幸長氏によるオンライン講習のサンプルコードを参考にした。
- Goh et al., Evaluating human versus machine learning performance in classifying research abstracts. Scientometrics 125, 1197 (2020).
- 実装に用いた主なツール：Python, BERT, SageMaker Studio Lab, AWS EC2, Flask, etc.

審査区分推定ウェブサイト

http://175.41.232.240:5000



①概要テキストを貼り付け

②submitをクリック！

③推定された大区分が表示される

④各区分の確率も表示される

K 科研費データベースから**「研究開始時の研究の概要」**をコピー！

注）試験運用のため動作が不安定です。
繋がらない場合はスマホからアクセスしてみてください。

結果

2022年度の基盤BC課題の大区分推定

推定された大区分

正解の大区分	A	B	C	D	E	F	G	H	I	J	K	再現率	適合率	F値
A	**2962**	6	36	2	1	41	5	0	181	32	9	90%	92%	91%
B	3	**464**	23	35	13	4	3	1	2	26	9	80%	88%	84%
C	33	12	**630**	61	3	12	0	1	20	50	24	74%	75%	75%
D	0	26	65	**184**	32	8	0	2	8	1	9	55%	49%	52%
E	1	7	8	66	**209**	14	4	20	11	1	3	61%	64%	62%
F	25	0	19	4	16	**483**	53	21	62	3	15	69%	66%	67%
G	2	2	1	2	7	51	**268**	13	55	7	1	66%	61%	63%
H	0	0	0	1	40	32	52	**188**	322	1	2	29%	53%	38%
I	122	1	11	11	2	50	43	105	**4343**	17	5	92%	86%	89%
J	44	6	38	1	4	8	10	2	33	**283**	2	66%	67%	66%
K	15	2	13	7	2	30	2	1	15	0	**81**	48%	51%	49%

正解率	平均再現率	平均適合率	平均F値(定義A)	平均F値(定義B)
81%	**66%**	**68%**	**69%**	**67%**

大区分の分類では高い推定精度を示した。

2022年度の神経科学関連課題の小区分推定

推定された小区分

正解の小区分	一般	形態	機能	基盤脳	認知脳	病態	認知	再現率	適合率	F値
神経科学一般	**21**	1	13	2	1	4	0	50%	43%	46%
神経形態学	7	**6**	0	0	0	4	0	35%	67%	46%
神経機能学	4	0	**15**	4	5	4	0	47%	31%	37%
基盤脳科学	2	1	4	**2**	1	0	0	20%	15%	17%
認知脳科学	6	1	5	2	**8**	2	2	31%	47%	37%
病態神経科学	9	0	5	0	0	**46**	1	75%	77%	76%
認知科学	0	0	7	3	2	0	**14**	54%	82%	65%

正解率	平均再現率	平均適合率	平均F値(定義A)	平均F値(定義B)
52%	**45%**	**52%**	**46%**	**48%**

類似の小区分では精度が劣っていた。

考察

- 大区分レベルなど大まかな分類であれば、科研費以外の学術文書や助成金情報などにも応用可能と考えられる。
- 類似の小区分での分類精度が低かった。これは学習データが少ないこと、同じ単語が同じ文脈でどの区分でも使用されていることなどが原因として考えられる。
- 日本語Wikipediaをコーパスに用いて事前訓練したモデルであるため、専門用語の解釈で不十分な点が散見された。最新のモデル（RoBERTa等）を検討する必要がある。**（自然言語処理に詳しい共同研究者を募集中です！）**