森羅プロジェクト交流会
取り組み共有

森羅2022 BERTワークショップ2
2日目 (2022/10/27)

横浜市立大学（学生） 武本孝輔

ちょっとだけ自己紹介

twitter@de_suke6

- 武本 孝輔（たけもと こうすけ）
- 社会人学生
- 近畿大学経済学部を卒業後、文系SEとして数年働く
- データサイエンスに出会い学びなおしを決意
- 現在、横浜市立大学データサイエンス学部３年
  (修士まではいきます)
- 自然言語を含む非構造化データと深層学習の研究室に入ったところ
- 本格的に機械学習やるのはこれから

本題

今、森羅プロジェクト2022で取り組んでいること

- HTML構造を補助情報として利用した属性抽出
  （まだまだちゃんとできたとは言えません）

HTMLを使う動機

は、オウム目インコ科に属する鳥類でオーストラリアの固有種。

分布 [編集]

クイーンズランド州中部からニューサウスウェールズ州東部、ビクトリア州、南オーストラリア州のアデレードのあたりまでと、タスマニア島。

形態 [編集]

全長20-23cm、体重約70g[1]。

全身はほぼ黄緑色で、耳羽、前頭部、くちばしは赤色。頭頂は青みのある緑色。後頭部から背面にかけてはオリーブ色。尾羽の下面は基部が赤色で、先端に行くにしたがって黄色みを

保全状況評価

LEAST CONCERN

(IUCN Red List Ver.3.1 (2001))

絶滅 | 絶滅危惧 | Least 低危険種

EX EW CR EN VU NT LC

分類

界：動物界 Animalia
門：脊索動物門 Chordata
亜門：脊椎動物亜門 Vertebrata
綱：鳥綱 Aves
目：オウム目 Psittaciformes
科：インコ科 Psittacidae
亜科：ヒインコ亜科 Psittacinae

- HTMLだと

```
<table>
<tbody><tr>
<td><a>界</a>
</td>
<td>:
</td>
<td><a>動物界</a> <span><a>Animalia</a></span>
</td></tr>

<tr>
<td><a>門</a>
</td>
<td>:
</td>
<td><a>脊索動物門</a> <span><a>Chordata</a></span>
</td></tr>
```

HTMLを使う動機

は、オウム目インコ科に属する鳥類でオーストラリアの固有種。

分布 [編集]

クイーンズランド州中部からニューサウスウェールズ州東部、ビクトリア州、南オーストラリア州のアデレードのあたりまでと、タスマニア島。

形態 [編集]

全長20-23cm、体重約70g[1]。

全身はほぼ黄緑色で、耳羽、前頭部、くちばしは赤色。頭頂は青みのある緑色。後頭部から背面にかけてはオリーブ色。尾羽の下面は基部が赤色で、先端に行くにしたがって黄色みを

保全状況評価

LEAST CONCERN

(IUCN Red List Ver.3.1 (2001))

絶滅 | 絶滅危惧 | Least 低危険種

EX EW CR EN VU NT LC

分類

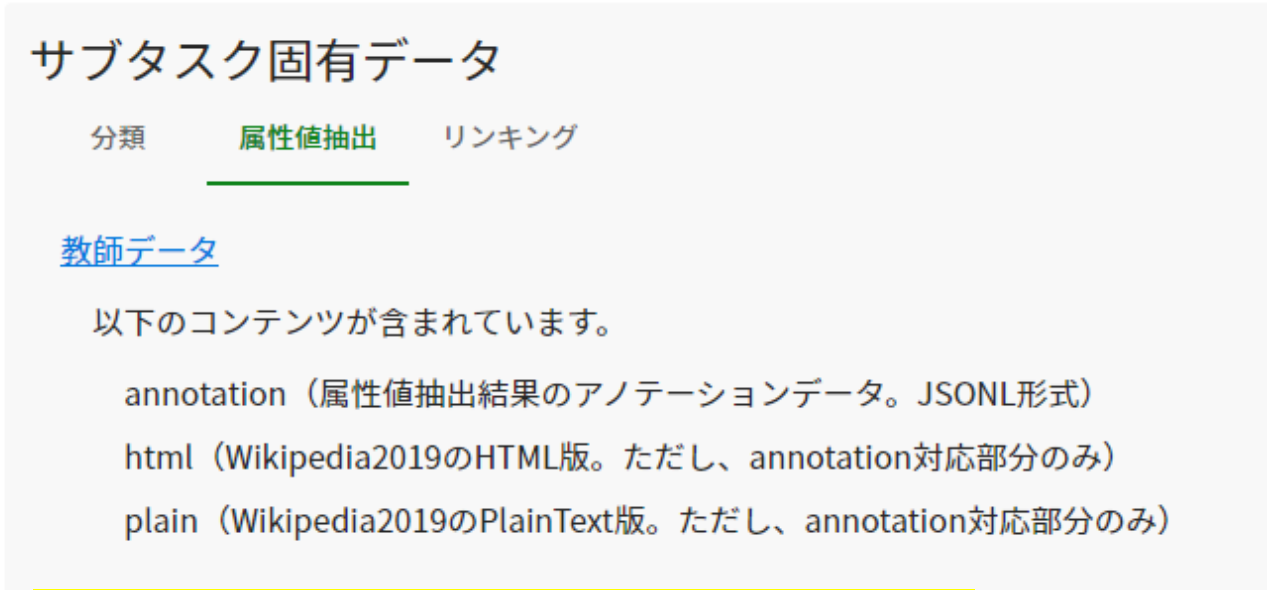
界：動物界 Animalia
門：脊索動物門 Chordata
亜門：脊椎動物亜門 Vertebrata
綱：鳥綱 Aves
目：オウム目 Psittaciformes
科：インコ科 Psittacidae
亜科：ヒインコ亜科 Psittacinae

- テキストだと

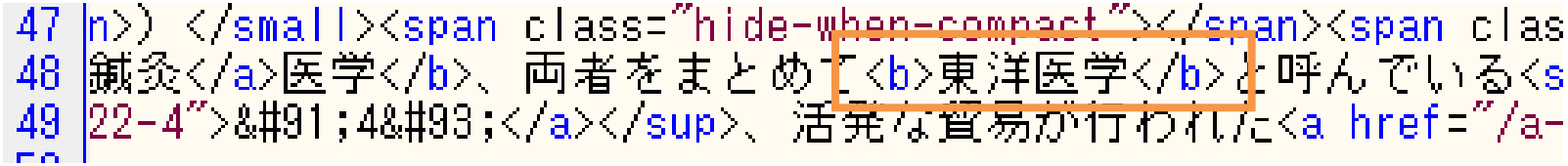
```
界
:
動物界 Animalia


門
:
脊索動物門 Chordata
亜門
```

属性抽出の教師データ

https://2022.shinra-project.info/data-download



19806ページ、914006個のアノテーション



一旦、HTML構造を素朴に利用

```
<p>フラミンゴ（Flamingo）という名前は<a href="/a-sumida/wiki2019_1/index.php/%E3%83%A9%E3%83%86%E3%83%B3%E8%AA%9E" title="ラテン語">ラテン語</a>で「<a href="/a-sumida/wiki2019_1/index.php/%E7%82%8E" title="炎">炎</a>」を意味するflammaに由来している。</p>
```

```
<p>フラミンゴ（Flamingo）という名前は<a>ラテン語</a>で「<a>炎</a>」を意味するflammaに由来している。</p>
```

- HTMLの属性（attribute）は使わず、上記のようにHTMLを簡略化してタグの存在と種類のみ用いる
- HTMLタグは特殊トークンとして扱う（損失関数からは除外）
- BERT(中山さんのRoBERTa)でファインチューニング
- NERワークショップ1日目で配布されたShinra2022NERWorkshop.ipynb を流用

実現のためのタスク

- HTMLから属性（attribute）を削除、処理対象を<body>以下に
  - それに伴い“html_offset”を調整
- TokenizerがHTMLタグを一つのトークンとして出力
  - japanese_roberta_tokenizerを改変
  - IPAdicにユーザ辞書を追加
  - HTMLタグをBPEの例外に
  - HTMLタグにtoken idを割り当て
- HTMLタグをBERTのEmbeddingsに追加
  - 標準正規分布からランダムに
- アウトプットをアノテーションの形式に
  - BERTの出力を線形変換したIOB2タグから”html_offset”または”text_offset”を作成

ハイパーパラメータ

MAX_LENGTH	128（colab上でやってるのでこれぐらいが限界…）
BATCH_SIZE	96
EPOCH	5
認識するHTMLタグの数	104（教師データにて50回以上現れたもの）
属性の種類	231 （教師データに現れるすべての属性名）

評価

- テストデータ（教師ありデータの10%）だと
  - Recall: 42.077%
  - Precision: 53.102%
  - F1: 46.951%
- リーダーボードだと

リーダーボード

Takemoto Kosuke2 さん

ログイン・参加登録

TASK 1: 分類

TASK 2: 属性値抽出

新規投稿　属性値抽出

Rank	Team Name	Submitted on	Description	Macro-F1 (Public) ↓	Micro-F1 (Public)
1	森羅2022実行委員会	2022/08/22	ベースラインシステム	44.9441	51.5130
2	Kosuke Takemoto	2022/10/15	BERT with html	35.0061	35.3854

拡張固有表現（カテゴリ）ごとにモデル構築

クイーンズランド州中部からニューサウスウェールズ州東部、ビクトリア州、南オーストラリア州のアデレードのあたりまでと、タスマニア島。

形態 [編集]

全長20-23cm、体重約70g[1]。

全身はほぼ黄緑色で、耳羽、前頭部、くちばしは赤色。頭頂は青みのある緑色。後頭部から背面にかけてはオリーブ色。尾羽の下

保全状況評価

LEAST CONCERN

(IUCN Red List Ver.3.1 (2001))

絶滅 / 絶滅危惧 / Least 低危険種

EX EW CR EN VU NT LC

分類

界：動物界 Animalia

門：脊索動物門 Chordata

亜門：脊椎動物亜門 Vertebrata

綱：鳥綱 Aves

目：オウム目 Psittaciformes

就航路線 [編集]

太字は同空港をハブ空港にしている航空会社。

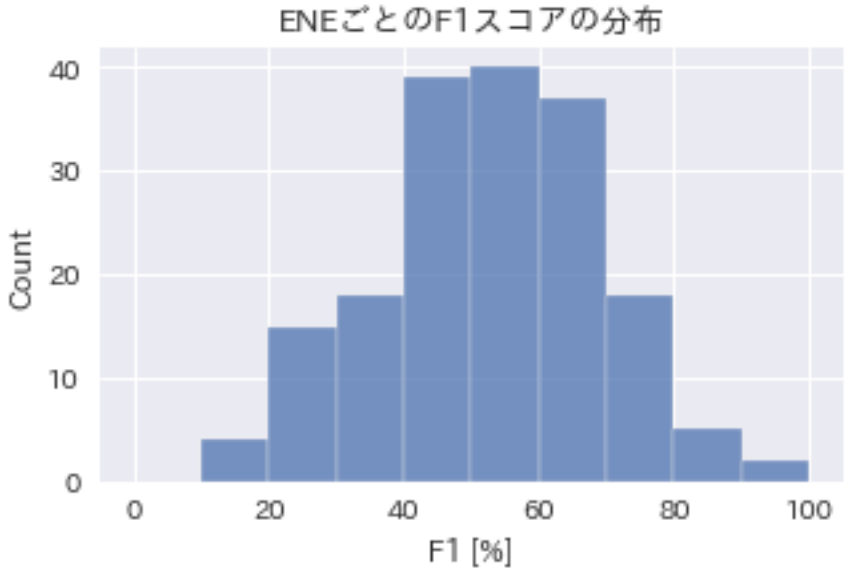
北ターミナル [編集]

航空会社	
日本航空 (JAL)[128]	**（北海道）** 女満別（夏期限定運航）、旭川（繁忙 **（東北）** 青森、三沢、秋田、花巻、仙台、山形 **（関東）** 東京/羽田、東京/成田 **（信越）** 松本（夏期限定運航）、新潟 **（近畿）** 但馬 **（中国・四国）** 出雲、隠岐、松山 **（九州・沖縄）** 福岡、長崎、大分、熊本、宮崎、
天草エアライン (AHX)[129]	熊本

今までのモデルをベースとして、拡張固有表現ごとにモデルを追加学習
150個のモデルを作成（EPOCH = 2）
リーダーボード評価用データを拡張固有表現ごとに対応するモデルで出力
教師データにないカテゴリを評価する際は、ベースのモデルで評価

評価

- テストデータ
  - Macro-F1: 53.02%（↑6%）
  - F1スコアとページ数の相関係数 0.103
    ページ数が少ないからF1スコアが低いとは限らない



- リーダーボード

Takemoto Kosuke2 さん

Description	Macro-F1 (Public) ↓	Micro-F1 (Public)
ベースラインシステム	44.9441	51.5130
BERT with html	35.0061	35.3854

Takemoto Kosuke2 さん

Description	Macro-F1 (Public) ↓	Micro-F1 (Public)
ベースラインシステム	44.9441	51.5130
BERT with html	40.5548	44.6001

今の実装の問題（要件への対応）

“html_offset”には”text”にHTMLタグが含まれているものがある

- 京都市東山区清閑寺霊山町
- 京都市東山区…
- 訓練データの914006個のアノテーション中、87320個が該当

単語が複数の属性を持つ場合に非対応

- 訓練データの914006個の注釈中、19475個は重複している注釈
- 今回のモデルだと複数の属性を持つ単語に非対応なので、
  一つの単語に対して、最大一つの注釈しか出力しない

今後とりくみたいこと

問題への対応	教師データの”text”にタグが含まれている場合への対応 複数属性への対応 HTMLタグの選別
GNN+BERT	DOMをグラフとして、GNNでテキストノードの埋め込みベクトルを得る。 BERTのトークンベクトルに得られたノードのベクトルを加える[1]

[1] 植畳, 數見拓朗, 小泉和之. HTML 構造を補助情報として利用する日本語ブログ記事からの固有表現抽出. 言語処理学会第 28 回年次大会発表論文集, 2022.

参考文献

[1] 植塁, 數見拓朗, 小泉和之. HTML 構造を補助情報として利用する日本語ブログ記事からの固有表現抽出. 言語処理学会第 28 回年次大会発表論文集, 2022.

以上です
ご清聴ありがとうございました

今の実装の問題（モデル）

補足資料

HTMLタグによるMAX_LENGTH(=128)の圧迫

- 一度にBERTに入力できるトークン長には限界が
- HTMLタグが多いと、それだけ自然言語の部分（文脈）が短くなってしまう
- 自然言語よりもトークン化するHTMLタグが有用な情報をもたないとダメ
- 対応策：対象とするHTMLタグの選別？