学生・若手研究者のためのBERTワークショップ2
(固有表現タスク　第2回目)

# 属性値抽出タスク
# BERTを用いたアプローチ例

2022/10/27

森羅2022 実行委員
三浦 明波

# 属性値抽出タスクのご紹介

カテゴリ分類
(文書分類) タスク

**サッポロポテト**

出典: フリー百科事典『ウィキペディア (Wikipedia)』

**サッポロポテト** (Sapporo Potato) はカルビーが製造販売するジャガイモをベースとするスナック菓子[1][2][3]。

**歴史**

カルビー創業者・松尾孝が1967年に「かっぱえびせん」をアメリカニューヨークの国際菓子博覧会に出典したとき[1][2][4][5]、衝撃を受けたのは会場内に山のように積まれたポテトチップスだった[1][6]。当時のアメリカでは

(…省略…)

**ページをカテゴリに分類**
**カテゴリ: 食べ物名**

※ 森羅2022では階層化された294カテゴリに分類

**分類されたカテゴリ (食べ物名) に基づいて属性値を抽出**

**属性値抽出タスク**

| タイトル: サッポロポテト, ページID: 956852, カテゴリ: 食べ物名__その他 | | |
|---|---|---|
| 別名 | Sapporo Potato | 1行目10文字目～1行目23文字目 |
| 生産者・組織 | カルビー | 1行目26文字目～1行目29文字目 |
| 販売者・組織 | カルビー | 1行目26文字目～1行目29文字目 |
| 材料 | ジャガイモ | 1行目37文字目～1行目41文字目 |
| 材料 | 小麦粉 | 40行目1文字目～40行目3文字目 |
| … | … | … |
| 種類 | スナック菓子 | 1行目49文字目～1行目54文字目 |
| 属性名 | 属性値 | 出現位置 |

- Wikipediaの記事から、該当するカテゴリに応じて属性名・属性値・出現位置を特定・列挙するタスク
- 属性名を固有表現クラス、属性値を固有表現と置き換えれば固有表現抽出タスクに類似
  ただし…　固有表現抽出タスクの一般的手法だけでは難しい課題もあり、工夫が求められる

# 分類カテゴリの定義について

Extended Named Entity –Ver 9.0.0- (拡張固有表現, ENE9.0.0)は階層構造で定義される

1 名前 (Name)

| ENE | | | 正例 | 負例 | |
|---|---|---|---|---|---|
| 1.0 名前__その他 | | | | | |
| 1.1 人名 | | | 福沢諭吉, エドガー・アラン・ポー, 春日局, R・ゼーリック | 田中(→ ＣＯＮＣＥＰＴ), ポパイ(→ キャラクター名), 浦島太郎(→ キャラクター名), 寅さん(→ キャラクター名) | 属性 |
| 1.2 神名 | | | アテネ, インドラ, ゼウス, ヘラクレス | 守り神(→ ＣＯＮＣＥＰＴ), 女神(→ ＣＯＮＣＥＰＴ), 現人神(→ ＣＯＮＣＥＰＴ), 八百万の神(→ ＣＯＮＣＥＰＴ) | 属性 |
| 1.3 生物呼称名 | 1.3.0 生物呼称名__その他 | | | | |
| | 1.3.1 動物呼称名 | 1.3.1.0 動物呼称名__その他 | たま, ポチ, トントン | ゴールデンレトリバー(→ 哺乳類名), カクレクマノミ(→ 魚類名), ネッシー(→ 架空生物名) | 属性 |
| | | 1.3.1.1 競走馬名 | オグリキャップ, ディープインパクト, トウカイテイオー, ナスルーラ系(競走馬の血統) | 競走馬(→ ＣＯＮＣＥＰＴ ) | 属性 |
| | 1.3.2 植物呼称名 | | 富田の一本松, 練馬白山神社の大ケヤキ, ロイヤル・オーク | バラ(→ 植物名) | 属性 |
| | 1.4.0 組織名__その他 | | 孔門の十哲, 向田ファミリー, 精華町町内会, 警視庁・神奈川県警合同捜査本部 | アイユーブ朝（→ 政治的組織名）, 全国黒人向上協会（→非営利団体名） | 属性 |
| | 1.4.1 国際組織名 | | 国連, ユニセフ, 北大西洋条約機構, 世界保健機関 | 国際陸上競技連盟(→ 競技連盟名) | 属性 |

←
各カテゴリの[属性]リンクから、抽出対象の属性名の定義と抽出された属性値の例が確認できる

https://liat-aip.sakura.ne.jp/ene/ene9/definition_jp/html/enetree.html

# 属性値抽出タスクの処理の流れ

本文(HTMLかプレーンテキストから選択可)の他に、分類カテゴリなども入力情報として利用

タイトル: サッポロポテト
ページID: 956852
分類カテゴリ: 食べ物名__その他
本文:
サッポロポテト (Sapporo Potato) はカルビーが製造販売するジャガイモをベースとするスナック菓子。
(後略)

属性値抽出

サッポロポテト (**Sapporo Potato**) は**カルビー**が製造販売する**ジャガイモ**をベースとする**スナック菓子。**
[別名] [生産者・組織] [材料] [種類]
[販売者・組織] ※ 同一の部分文字列に複数属性名が該当

各属性名の候補に対して、対応する属性値となる部分文字列・出現箇所を全て列挙

# 属性値抽出と固有表現抽出の違い

- 抽出対象の属性がカテゴリや文脈に応じて変化する
  例)
  同じ「カルビー」という部分文字列でも、
  「サッポロポテト」（食べ物名)の記事では
  「生産者・組織」「販売者・組織」などの属性に該当し、
  「カルビーポテト」（企業名）の記事では
  「主要株主」などの属性に該当する

- 抽出する属性の種類 (属性名)の数:
  - 1000種類以上(森羅2022タスクの定義では1722の異なり数)の属性名があり、
    固有表現(10前後)や拡張固有表現(数百)で扱うクラス数とは桁違いに多い

➢固有表現抽出で用いられる一般的なアプローチ
(各トークンをクラス毎のIOBタグに分類)の場合、対処に難しい場合がある
  - 1000以上の属性名に対するIOBタグに分類?
  - 記事のカテゴリ毎に異なるモデルを作成して学習?

# 別の類似タスク: SQuADの紹介

- SQuADとは?
  - データセット・タスクの名前で、**S**tanford **Qu**estion **A**nswering **D**atasetの略 **[Rajpurkar+, 2016]**
  - 以下のようなテキスト(段落)と質問文のペアの入力から、
    質問文を元に「テキストの中から正解部分を出力」する抜き出し問題

例)

| | |
|---|---|
| テキスト | Beyoncé Giselle Knowles-Carter (/biːˈjɒnseɪ/ bee-YON-say) (born September 4, 1981) is an American singer, songwriter, record producer and actress. Born and raised in **Houston, Texas**, she performed in various singing and dancing competitions as a child, and rose to fame in the late 1990s as lead singer of R&B girl-group Destiny's Child. (後略) |
| 質問文 | In what city and state did Beyonce  grow up? |
| 回答 | テキスト: Houston, Texas<br>開始位置: 166文字目 |

- 高難易度なタスクとされてきたが、BERTの登場で高い精度で推論可能となった **[Devlin+, 2018]**

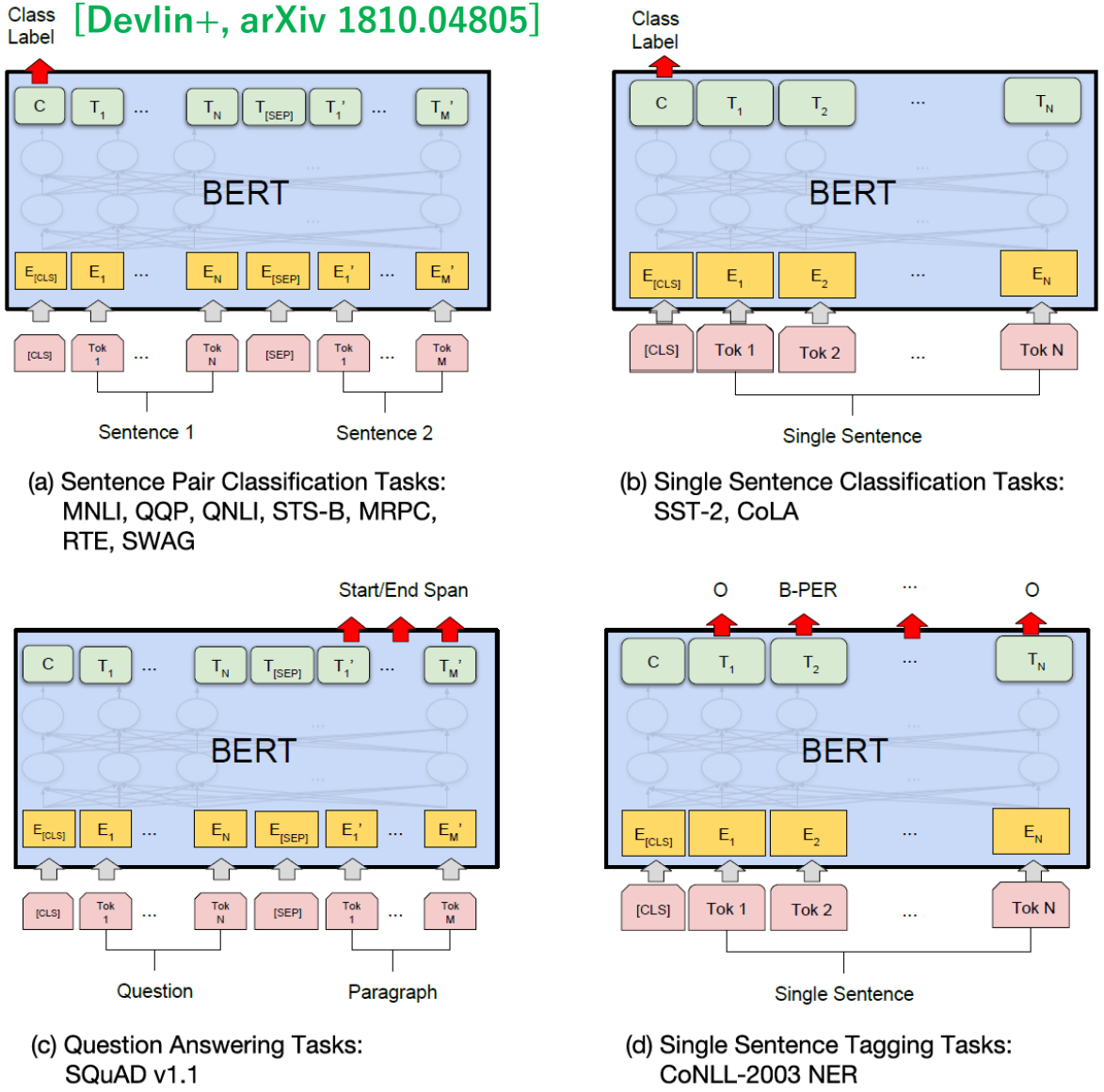
# BERTを用いた4つのタスクアプローチ

BERTが得意とする4つのタスク類型

(a) マッチング (連結した2文間の関係を分類)
(b) 分類 (入力された1文の内容を元に分類)
**(c) スパニング**
**(連結したクエリと本文から、**
**本文の各トークンに対して分類して範囲推定)**
**(d) 系列タギング**
**(入力された文の各トークンに対して分類)**

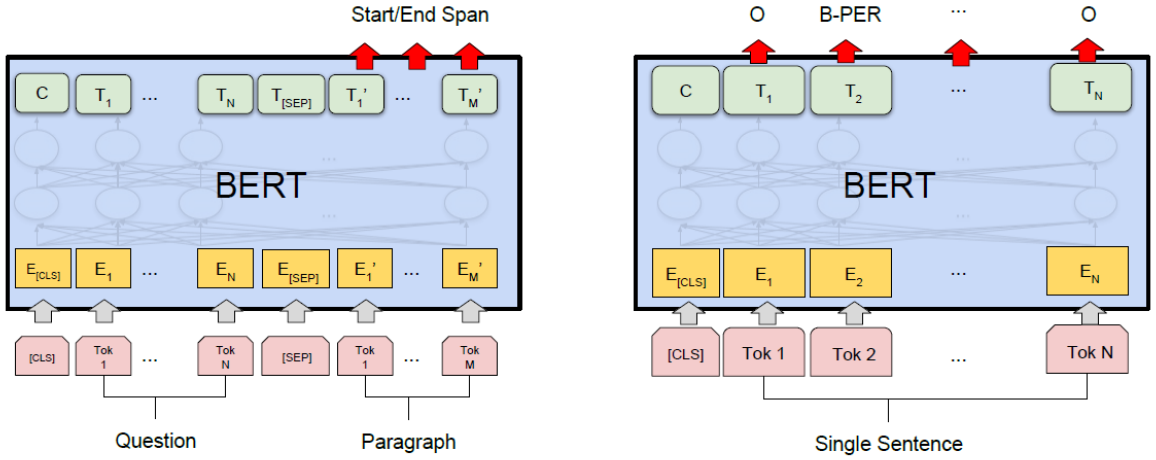
今回は(c)と(d)に注目

[Devlin+, arXiv 1810.04805]



# 系列タギングとスパニング

[Devlin+, arXiv 1810.04805]




(c) Question Answering Tasks: SQuAD v1.1

(d) Single Sentence Tagging Tasks: CoNLL-2003 NER

- 固有表現抽出では(d)の系列タギングが有効
- SQuADでは(c)のスパニングで高い性能
- 属性値抽出では(d)も使えなくはないけど、SQuAD同様の定式化をすれば(c)で推論できる！
  - ✓ 単一モデルでも、複数の属性名・属性値の抽出ができる（同一部分文字列に対しても複数の属性を抽出できる）
  - ✓ 単一モデルでも、複数カテゴリの記事からの抽出に対応できる
  - × 系列タギングのように一括で抽出できないので、抽出対象の数だけ比例して遅くなる

# 属性値抽出をSQuAD形式で定式化

- 森羅2020の属性値抽出で最高精度となった手法であり **[石井, 2021]**、今年度タスクにおけるベースラインに採用
- 属性名を質問、該当する属性値の出現箇所1つ1つを回答とみなして、QAペアを抽出

サッポロポテト（**Sapporo Potato**）は**カルビー**が製造販売する**ジャガイモ**をベースとする**スナック菓子。**
[別名] [生産者・組織] [販売者・組織] [材料] [種類]

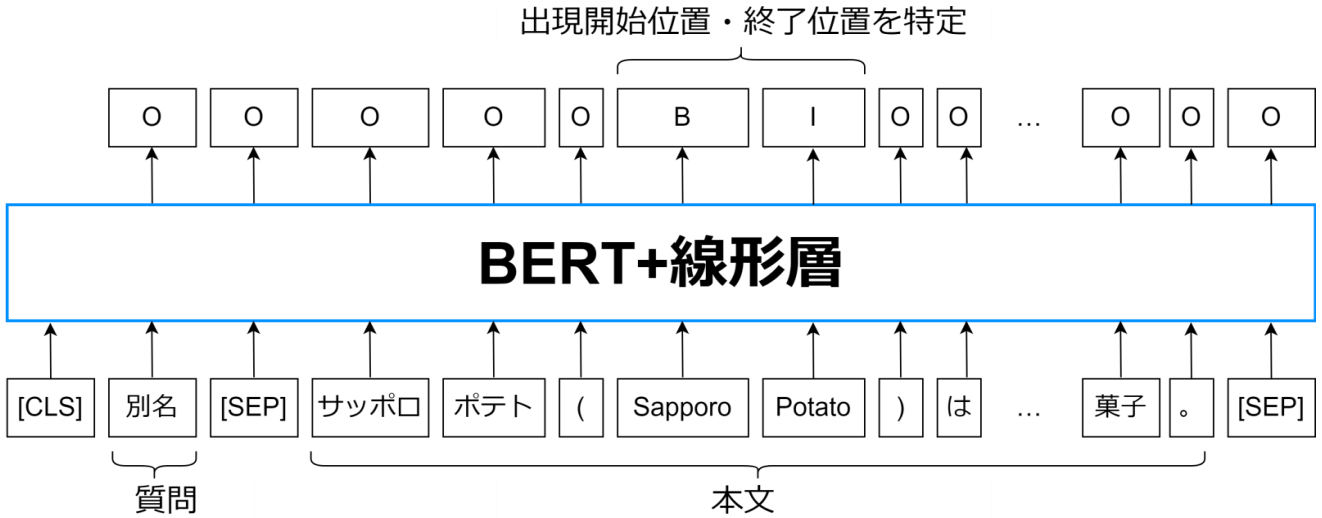
QAペアを抽出

質問: 別名
回答: Sapporo Potato
(10文字目から開始)

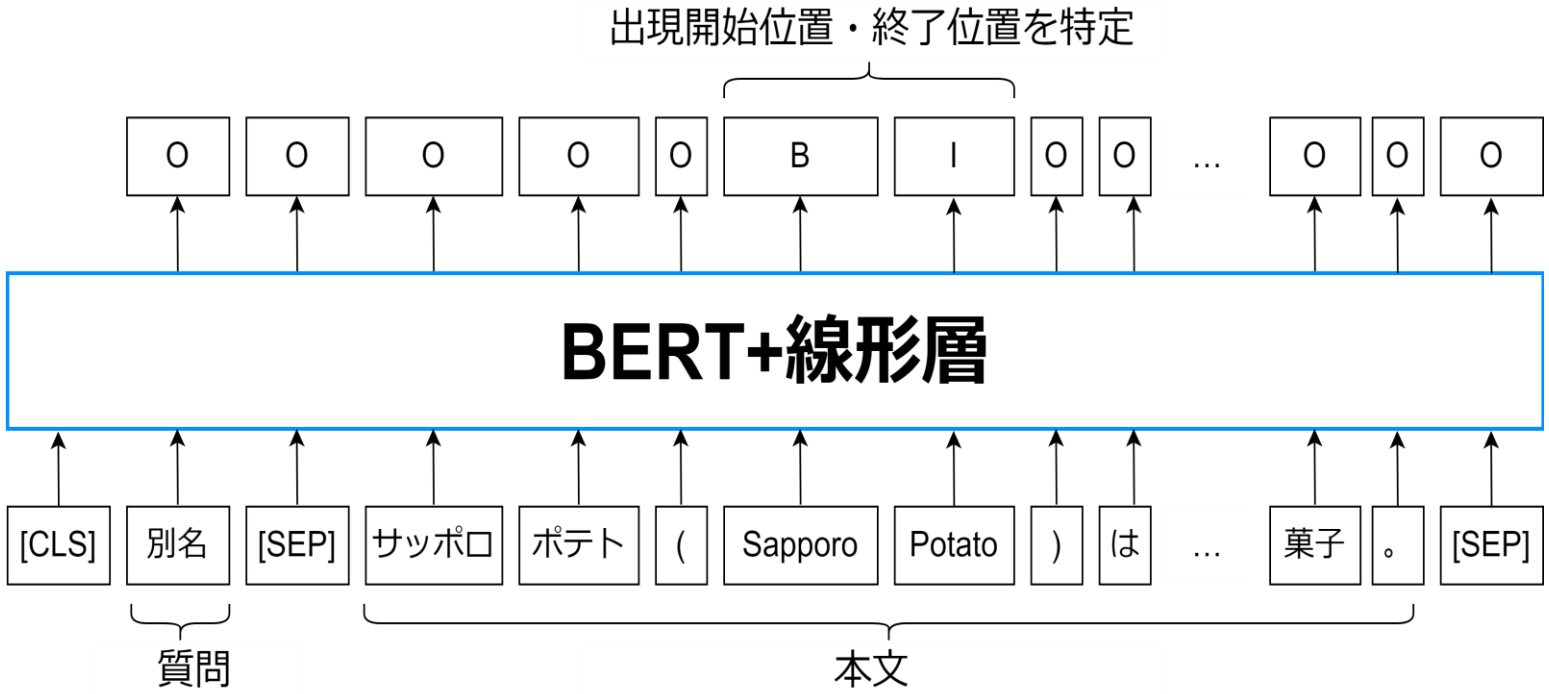
質問: 生産者・組織
回答: カルビー
(26文字目から開始)

…

日本語BERTで
1ペアずつ処理



# 実装面での扱い




運営のベースラインシステム実装:
https://github.com/akivajp/shinra-attribution-extraction

- 属性名を質問文とみなし1文目に、本文を2文目として結合して入力
- BERTの出力層では、入力トークンに対する属性値の開始位置にBタグ、継続位置にIタグ、それ以外の部分にはOタグを分類して出力する
(SQuADでは出現位置は1箇所のみだが、属性値抽出では複数出現する場合は、それに応じた数だけBタグ・Iタグを同時出力)

# さらなる改善のために

- 入力文をHTMLで扱う場合、タグの扱いをどうするか
  - Aタグなどは重要なヒントになるはず
- BERTは最大512トークンまでといった系列長の制約があり、
  全文一括ではなく段落毎に属性値の有無や出現位置の推定が必要
- カテゴリ分類との複合タスクとして取り組む場合、
  前段の分類精度の影響も大きくなる
  - n-bestなどを利用して改善の余地あり
- RoBERTaなどの後発のモデルに変更したり、
  出力層側での追加処理など

# まとめ

- 属性値抽出タスクのご紹介
  - 固有表現抽出と比べてチャレンジングなポイント
- SQuADタスクのご紹介と、
  BERTの主要なタスクアプローチについてご紹介
- SQuADタスクを転用してBERTで
  属性値抽出タスクを行うアプローチ例(ベースライン)のご紹介