「構造化知識を使った言語処理応用」ワークショップ

属性値抽出タスク
ベースラインシステム紹介

2023/01/18

森羅2022 実行委員
三浦 明波

属性値抽出タスクの概要

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

属性値抽出をSQuAD形式で定式化

- 森羅2019,2020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]

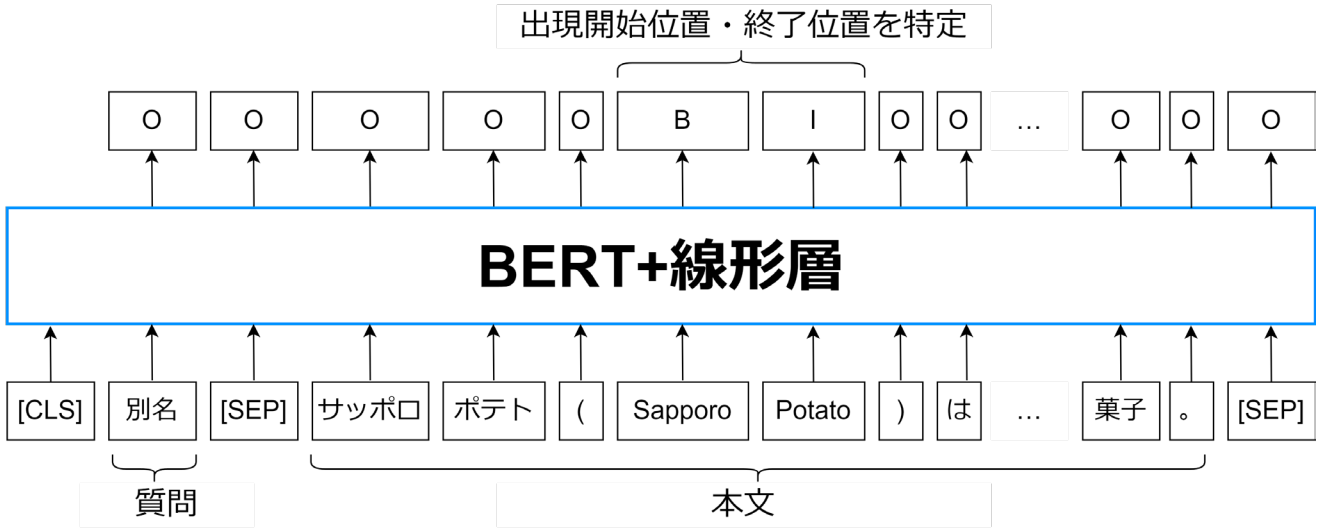
QAペアを抽出

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



訓練条件

利用データ

- 森羅2022 カテゴリ分類タスク推論結果 (今回のタスクでは正解ラベルは与えられていない)
- 森羅2022 属性値抽出タスク用訓練データ (Wikipedia2019とWikipedia2017の混合アノテーション)

利用モデル

- NICT BERT 日本語 Pre-trained モデル (ICT_BERT-base_JapaneseWikipedia_32K_BPE)

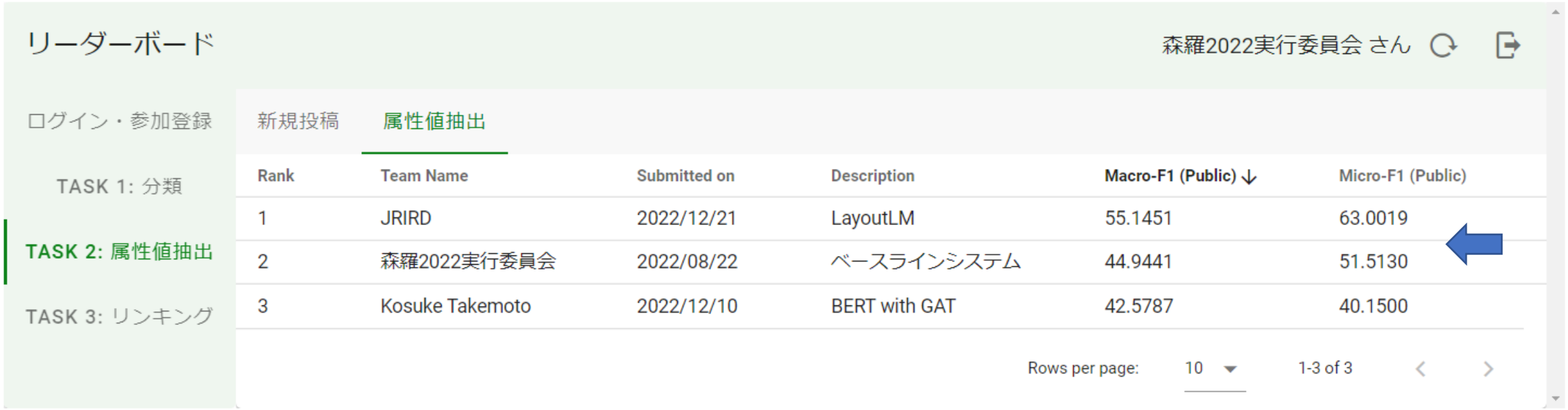
その他の設定

- グルーピング中間学習なし（グルーピング版は検証・訓練・推論間に合わず未提出)

実装公開URL

- https://github.com/akivajp/shinra-attribution-extraction

評価結果



(2023年1月18日14時現在)

- 過去タスクでのスコアから70弱程度のスコアを期待したが、マクロ平均スコアなどは低い
  - 前段のカテゴリ分類タスクでページに対する正解カテゴリは不明、ノイズが乗る
  - カテゴリ数が非常に多く、特定カテゴリ・属性の抽出が困難？

カテゴリ別のスコア

126カテゴリ

16カテゴリ

ENE ID	ENE Ja	Precision	Recall	F1 (↓ソート)	訓練サンプル数	正解サンプル数	提出抽出数	正解数
1.5.3.6	海洋名	**0.77778**	**0.84**	**0.80769**	4588	25	27	21
1.5.3.7	湾名	**0.66667**	**1**	**0.8**	3484	10	15	10
1.5.3.4	河川名	**0.75**	**0.83333**	**0.78947**	7066	18	20	15
1.8.3.2	選挙名	**0.7**	**0.90323**	**0.78873**	9524	31	40	28
1.6.4.2	城名	**0.8125**	**0.72222**	**0.76471**	2904	18	16	13
…	…	**…**	**…**	**…**	…	…	…	…
1.7.19.8	制度名	**0.25**	**0.04348**	**0.07407**	2094	23	4	1
1.5.1.0	ＧＰＥ＿その他	**0.08333**	**0.04545**	**0.05882**	6194	22	12	1
1.6.4.4	公共機関名	**0.11111**	**0.03846**	**0.05714**	3889	26	9	1
1.7.7	情報機器名	**0.0678**	**0.04651**	**0.05517**	2587	86	59	4
1.7.14.3	書物名	**0.25**	**0.01724**	**0.03226**	6788	58	4	1
1.7.19.13	学問名	**0**	**0**	**0**	1851	**2**	**0**	**0**
…	…	**…**	**…**	**…**	…	**…**	**…**	**…**
1.7.19.14	理論名	**0**	**0**	**0**	1215	**2**	**0**	**0**

正解サンプル数が少なく、スコア0のカテゴリが16個も存在

カテゴリ+属性別のスコア

直感: 訓練データが多いほどスコアが良い？→必ずしもそうでもなさそう

976行

102行（スコア0ばっかり）

ENE Ja	属性名	F1	訓練サンプル数（↓ソート）	ページ数	平均サンプル数
競技リーグ名	**歴代所属チーム**	**0.361702**	**23888**	173	**138.0809249**
家系名	所属する人物	0.734848	12129	196	61.88265306
公演組織名	上演作品	0.85	11687	181	64.56906077
人名	作品	0.674556	6397	75	85.29333333
路線名__その他	停車場	0.387097	5963	92	64.81521739
…	…	…	…	…	…
宿泊施設名	設計者・組織	1	51	21	2.428571429
橋名	設計者・組織	0.666667	50	26	1.923076923
恒星名	発見方法	0	3	3	1
医薬品名	属する医薬品	0	1	1	1
車名	販売国・ＧＰＥ	0	0	0	
…	…	…	…	…	…
自然物名__その他	下位の自然物	0	0	0	

競技リーグ名・歴代所属チームの例

なぜこんなにサンプル数が多い？

プリメーラ・ディビシオン

88の言語版

ページ　ノート　閲覧　編集　履歴表示

出典: フリー百科事典『ウィキペディア（Wikipedia）』

この項目では、スペインのトップリーグについて説明しています。その他の用法については「プリメーラ・ディビシオン (曖昧さ回避)」をご覧ください。

プリメーラ・ディビシオン（西: Primera División）または、**ラ・リーガ**（LaLiga）は、リーガ・デ・フトボル・プロフェシオナル（LFP、スペインプロリーグ機構）が運営するスペインのプロサッカーリーグ。正式名称は、**ラ・リーガ・サンタンデール**（LaLiga Santander）と称している。

日本では**リーガ・エスパニョーラ**（西: Liga Española、「スペインリーグ」の意味）の通称で知られていたが、リーグ側は2016年より**ラ・リーガ**（西: LaLiga）という名称を用いている[1][2][3]。

イングランドのプレミアリーグ、イタリアのセリエA、ドイツのブンデスリーガ、フランスのリーグ・アンとともにヨーロッパの五大プロサッカーリーグを形成している[4]。

この記事の項目名には以下のような表記揺れがあります。
ラ・リーガ

ラ・リーガ サンタンデール
LaLiga Santander

加盟国	スペイン
大陸連盟	UEFA
創立	1929年
参加クラブ	20
リーグレベル	第1部
下位リーグ	セグンダ・ディビシオン

概要 [編集]

スペインのプリメーラ・ディビシオンは、イングランドのプレミアリーグ、イタリアのセリエA、ドイツのブンデスリーガ、フランスのリーグ・アンとともにヨーロッパの五大プロサッカーリーグを形成している[4]。

プリメーラ・ディビシオンでは、創設以来計62のチームが参加している。9つのチームがチャンピオンを獲得しており、レアル・マドリードCFが35回、FCバルセロナが26回獲得している。バルセロナは1929年に初優勝し、アスレティック・ビルバオがリーグ初期

表からチーム名を
全部抽出！

2020年代 [編集]

所属クラブ（2021-22シーズン） [編集]

チーム名	監督	自治州	ホームタウン	スタジアム	収容人数	前年度成績
アトレティコ・マドリード	ディエゴ・シメオネ	マドリード州	マドリード	ワンダ・メトロポリターノ	68,000人	優勝
レアル・マドリード	カルロ・アンチェロッティ			サンティアゴ・ベルナベウ	81,044人	2位
ラージョ・バジェカーノ	アンドニ・イラオラ			デ・バジェカス	14,708人	6位☆
ヘタフェCF	ミチェル		ヘタフェ	コリセウム・アルフォンソ・ペレス	17,000人	15位
アスレティック・ビルバオ	マルセリーノ・ガルシア・トラル	バスク州	ビルバオ	サン・マメス・バリア	53,332人	10位
デポルティーボ・アラベス	ハビエル・カジェハ		ビトリア=ガステイス	メンディソロツァ	19,840人	16位
レアル・ソシエダ	イマノル・アルグアシル		サン・セバスティアン	レアレ・セグロス・スタジアム	32,076人	5位
カディスCF	アルバロ・セルベラ	アンダルシア州	セビリア	ヌエボ・ミランディージャ	25,033人	12位
セビージャFC	フレン・ロペテギ			ラモン・サンチェス・ピスファン	45,500人	4位
レアル・ベティス	マヌエル・ペレグリーニ			ベニート・ビジャマリン	60,720人	6位
グラナダCF	ディエゴ・マルティネス		グラナダ	ヌエボ・ロス・カルメネス	22,094	9位
RCDエスパニョール	ビセンテ・モレノ	カタルーニャ州	バルセロナ	RCDEスタジアム	40,500人	優勝☆
FCバルセロナ	シャビ・エルナンデス			カンプ・ノウ	99,354人	3位
バレンシアCF	ホセ・ボルダラス	バレンシア州	バレンシア	エスタディオ・デ・メスタージャ	49,677人	13位
レバンテUD	パコ・ロペス			シウダ・デ・バレンシア	25,354人	14位
ビジャレアルCF	ウナイ・エメリ		ヴィラ=レアル	エスタディオ・デ・ラ・セラミカ	24,500人	7位
エルチェCF	フラン・エスクリバ		エルチェ	マルティネス・バレーロ	33,732	17位
セルタ・デ・ビーゴ	エドゥアルド・コウデ	ガリシア州	ビーゴ	バライードス	29,000人	8位
CAオサスナ	ハゴバ・アラサテ	ナバラ州	パンプローナ	エル・サダール	15,326人	11位
RCDマジョルカ	ハビエル・アギーレ	バレアレス諸島州	パルマ・デ・マヨルカ	ソン・モイシュ	23,142人	2位☆

- **BERTでも、表から適切なカタカナ文字列を大量抽出するのは難しい？**
- **負例をどう適切に学習させるかも課題の一つ**

まとめ

- 属性値抽出タスクをSQuADタスク形式に変換し、
  スパニング問題としてBERTのfine-tuningで対応
  （過去タスクで最高精度の手法）

- 前段タスクでのノイズやカテゴリ数・属性数の多さから、
  性能に大幅に課題あり

- 主な課題:
  - 訓練データ不足への対応 (多段階学習など?)
  - 表形式データへの対応 (ルールベースとのハイブリッド手法など?)
  - 負例をどう学習に取り入れるか
    （誤った推論結果をサンプリングして学習に利用?）

補遺

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 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]**

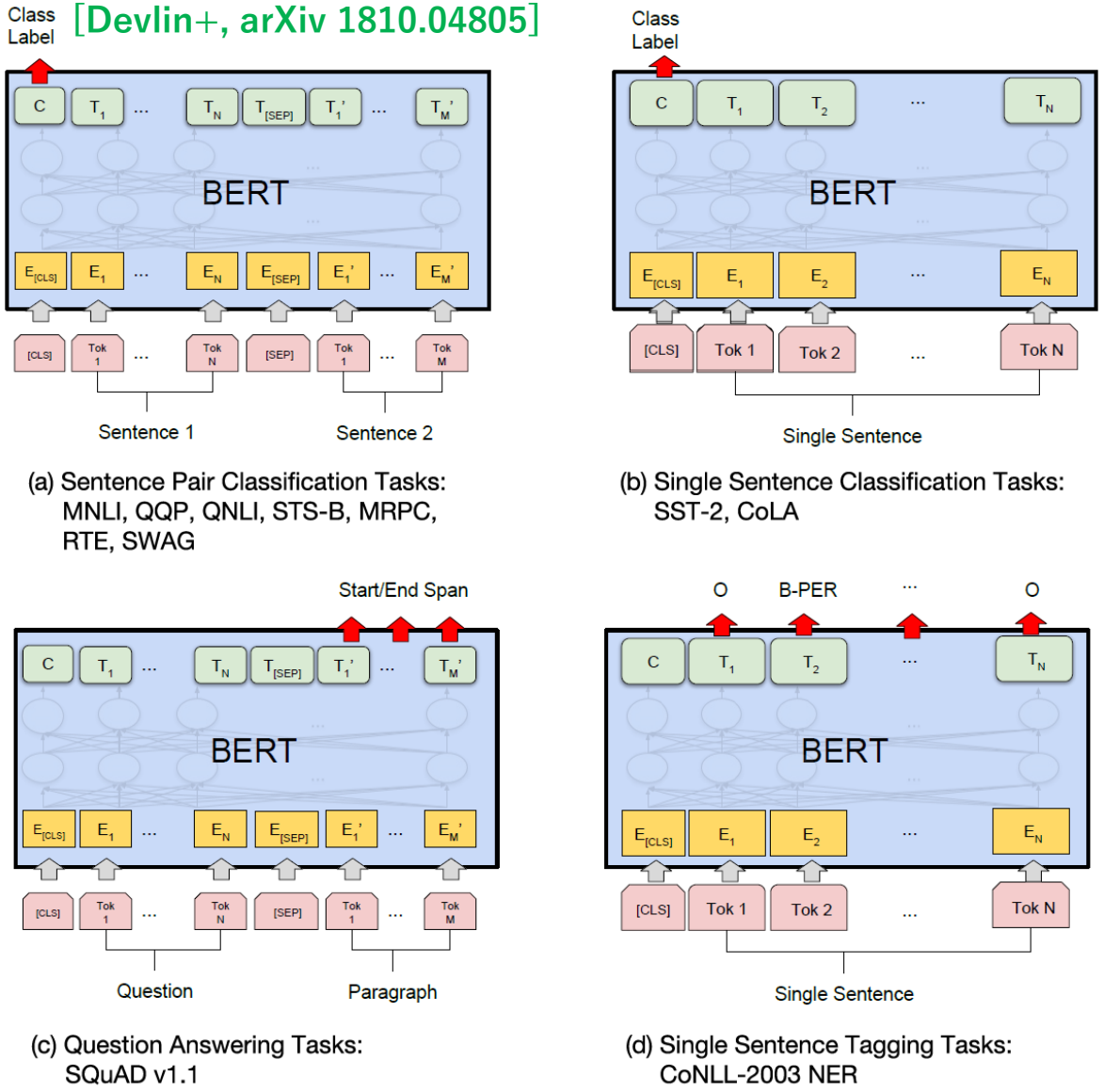
BERTを用いた4つのタスクアプローチ

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)

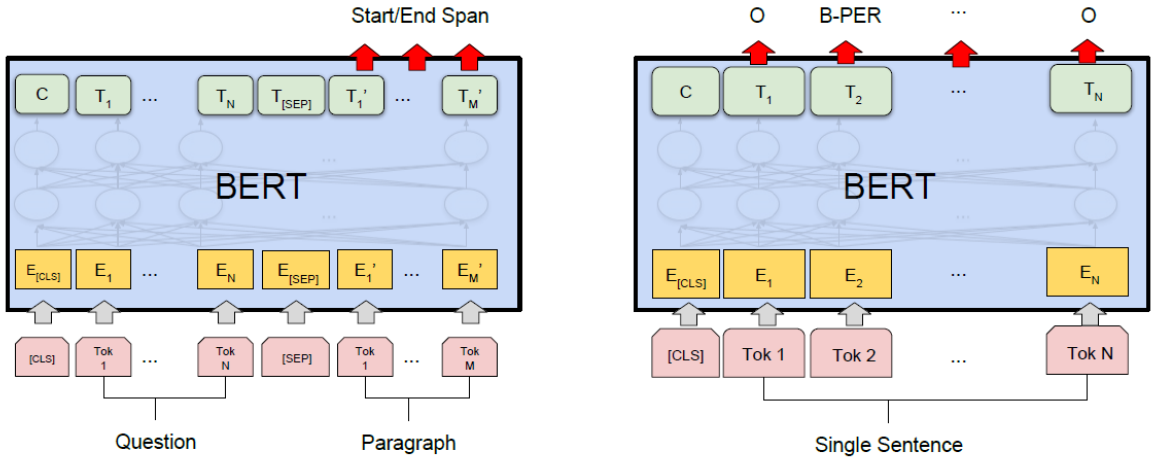
今回は(c)と(d)に注目

[Devlin+, arXiv 1810.04805]



系列タギングとスパニング

[Devlin+, arXiv 1810.04805]

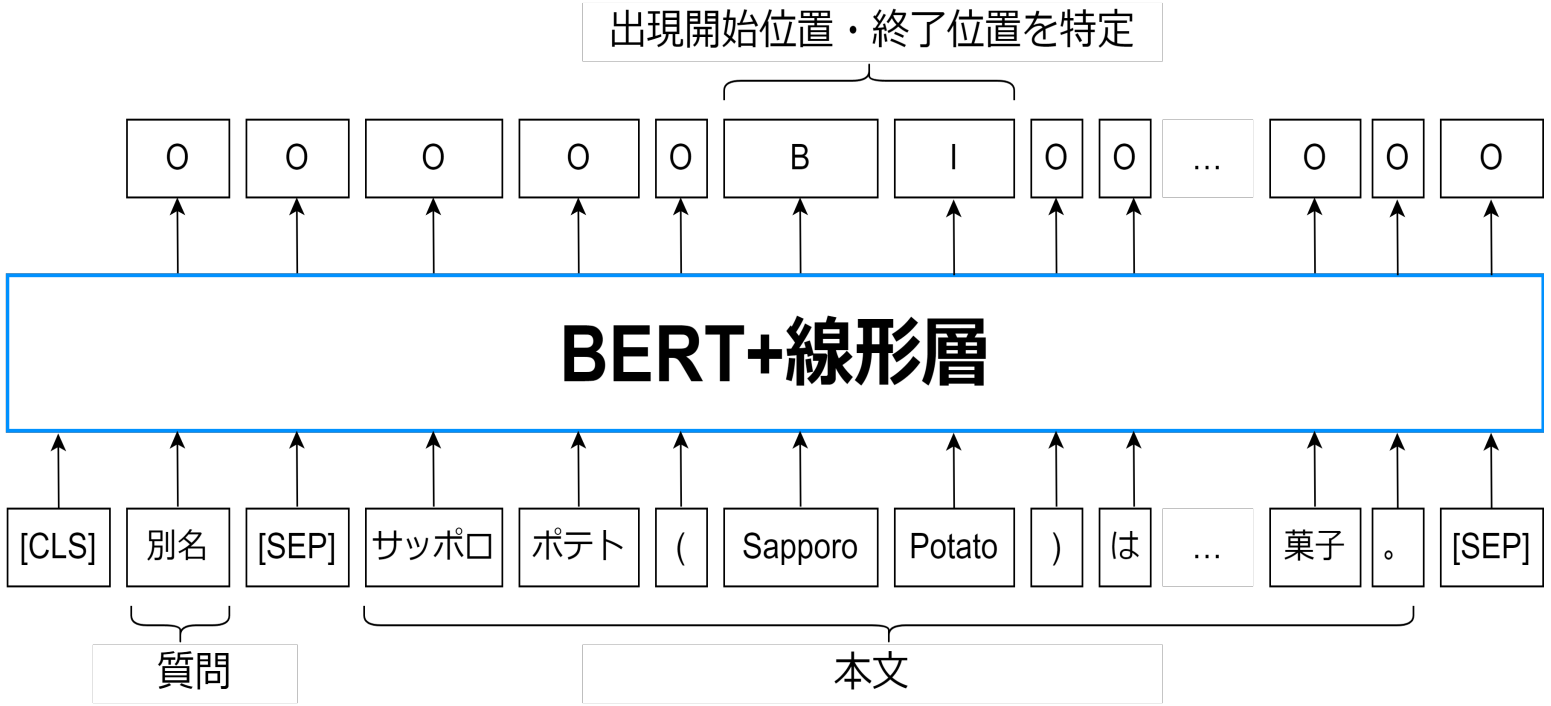


(c) Question Answering Tasks: SQuAD v1.1

(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

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)