森羅プロジェクト2023 キックオフミーティング

公開データ・評価に関するご説明

2023/05/18

森羅2023 実行委員
三浦 明波 (株式会社アティード)

公開データの種類について

公開データの種類について

- 森羅2023特設ページから全てDL可能: https://2023.shinra-project.info/
- 全タスク共通データ (2019年版/2021年版Wikipedia全記事データ, 4つのフォーマット)

	Cirrus Search Dump (JSON Lines)	Wiki Dump (XML)	HTML	Plain Text
W2019 (訓練用)	全記事	全記事	全記事	一部のみ (属性値抽出・リンキングの訓練データ対象記事のみ)
W2021 (推論対象)	全記事	全記事	全記事	全記事

- カテゴリ・属性・タスク定義データ (JSON Lines形式で1つのファイルに集約)
- サブタスク固有データ (各データは JSON Lines 形式、昨年タスクから細かい変更あり)

	カテゴリ分類	属性値抽出	リンキング	End-to-End
W2019 訓練データ	W2019全記事 ラベル付き	W2017+W2019一部記事 ラベル付き	W2017+W2019一部記事 ラベル付き	←の各タスクデータの活用を想定
W2021 リーダーボード	W2021一部記事 ラベル無し	W2021一部記事 カテゴリ分類ラベル付き 属性値ラベル無し	W2021一部記事 カテゴリ分類・属性ラベル付き リンキングラベル無し	W2021一部記事 ラベル無し

※ 個別サブタスクにも取り組みやすいよう森羅2021年以前の形式でデータ公開

公開データ: Wikipedia全記事データ

- 日本語Wikipediaの2019年版 (W2019) と2021年版 (W2021)について、以下の4種類のファイルを公開

1. MediaWikiダンプファイル (XML形式)

   Wikipediaの全記事をXML形式で出力したもの。
   WikiExtractorなどのOSSツールでデータ取り出し可能。

2. 全記事HTMLファイル ({ページID}.html)

   1.のMediaWikiダンプファイルを元にWikipediaを再現・クロールして取得したHTML

3. プレーンテキストファイル ({ページID}.txt)

   2.のHTMLファイルを元にHTMLタグを取り除きテキスト情報のみを取得したファイル
   (W2019については対象カテゴリの記事のみ、W2021については全記事を公開)

4. Cirrus Searchダンプファイル (JSON形式)

   構造化の対象となる標準名前空間内の全記事をJSON形式で出力したもの。
   リダイレクトページや、テンプレートなどの特殊ページは除外されている。

カテゴリ・属性・タスク定義データ

●各タスクのデータ作成（アノテーション）
および訓練・推論を行う上で基準となるタスク定義データ

- ENE9.0定義
  - ENE ID、日本語カテゴリ名、英語カテゴリ名、上位カテゴリ、下位カテゴリ、カテゴリ説明（日本語・英語）を規定
- 属性・タスク定義
  - ENE9.0定義の各末端カテゴリ(子を持たないカテゴリ)について、
    属性名（日本語）、属性説明（日本語）、
    属性値抽出の対象となるか、属性値リンキングの対象となるか（bool値)を規定

定義データ内の各行の例 (JSON Line)


```
{
 "ENE_id": "1.1",
 "name": { "en": "Person", "ja": "人名" },
 "definition": { "en": "A name of …", "ja": "人の名前。¥n人名の…" },
 "children_category": [], "parent_category": "1",
 "attributes": [{
  "name": "異表記", "description": "見出し語の異表記。…", "extraction_task": true, "linking_task": false
 }, …]
}
```

タスク固有データ①

カテゴリ分類タスク

分類タスク: 公開データ

- 訓練用ラベル付きデータ:
  - W2019全記事 (920,044ページ) に対する
    ENE9.0ラベル付きデータ (**公開中**)

- 開発データ:
  - W2021から選出されたページIDと正解ENE IDの対応データ (近日公開)

- 個別タスク評価用データ:
  - W2021から選出されたリーダーボード評価用データ (近日公開)
  - W2021から選出された本評価用データ (近日公開)

- タスク内容
  - W2021の記事のページIDとタイトルの一覧の入力を元に、
    各ページIDに対するENE9.0ラベルを付与して出力

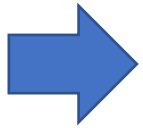
分類タスク: 入出力フォーマット

訓練データ例:

```
{
   "page_id": "72942", // ページID
   "title": "バックス (ローマ神話)",  // ページタイトル
   "ENEs":
      {
         "HAND.AIP.202304": [
               {
                  "ENE": "1.2", // 正解カテゴリID
                  "prob": 1, // 常に1
               }
         ]
      }
}
```

評価用データ例:

```
{
   "page_id": "401755",
   "title": "覚信尼",
   "ENEs":
      { } // 空の状態で配布
}
```



提出データ例:

```
{
   "page_id": "401755",
   "title": "覚信尼",
   "ENEs":
      {
         // "AUTO." 以降に任意のシステム名を入力
         "AUTO.{YOUR_SYSTEM}.202306": [
               {
                  "ENE": "1.1", // 予測カテゴリID
                  "prob": 0.85, // 予測確率スコア
               }
         ]
      }
}
```

- 評価のポイント
  - 評価方法: page_id, title, ENEの全てが一致する場合のみ正解とみなし、マイクロ平均F-1スコアで評価
  - ENEフィールドにはカテゴリ名ではなくカテゴリID ("1.2"など) を記載
  - ページIDのフィールド名は "pageid" ではなく "page_id" であり、整数値ではなく文字列値を取ること

タスク固有データ②

属性値抽出タスク

属性値抽出タスク: 公開データ

- 訓練用ラベル付きデータ:
  - W2017一部記事(8,867ページ), W2019一部記事 (18,942ページ)
    に対してENE9.0ラベルと全属性値のラベルが付与されたデータ(**公開中)**
  - 合計27,809ページ, 約130万件の属性値
  - 各ページIDに対応するHTML, プレーンテキストは訓練用Zipファイルに同梱

- 個別タスク評価用データ:
  - W2021から選出されたリーダーボード評価用データ (近日公開)
  - W2021から選出された本評価用データ (近日公開)

- タスク内容
  - W2021の記事のページID・タイトル・ENE9.0ラベルの一覧入力を元に、
    対象となる178カテゴリに該当する全ページから
    抽出対象属性の全属性値（属性名と出現位置)を列挙

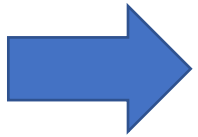
属性値抽出タスク: 入出力フォーマット

訓練データ例:

```
{
   "page_id": "239553",
   "title": "人体解剖学",
   "ENE": "1.7.19.13",
   "attribute": "別名",
   "text_offset": {
        "start": { "line_id": 27, "offset": 88 },
         "end": { "line_id": 27, "offset": 101},
         "text": "human anatomy"
   },
   "html_offset": {
        "start": { "line_id": 27, "offset": 592 },
        "end": { "line_id": 27, "offset": 605 },
        "text": "human anatomy"
   },
}
```

評価データ例:

```
{
   "page_id": "1101676",
   "title": "数値流体力学",
   "ENE": "1.7.19.13",
}
```



提出データ例:

```
{
   "page_id": "1101676",
   "title": "数値流体力学",
   "ENE": "1.7.19.13",
   "attribute": "読み",
   // text_offset か html_offset の
   // どちらか一方のみでOK
   "text_offset": {
        "start": { "line_id": 28, "offset": 7 },
         "end": { "line_id": 28, "offset": 19 },
         "text": "すうちりゅうたいりきがく"
   },
   "html_offset": {
        "start": { "line_id": 28, "offset": 17 },
        "end": { "line_id": 28, "offset": 29 },
        "text": "すうちりゅうたいりきがく"
   },
}
```

- 評価のポイント:
  - 複数の属性値が存在する場合、その数だけJSON Lineを出力
  - page_id, title, ENE, offset (text_offset または html_offset) の全てが一致した場合に正解とみなす
  - text_offset と html_offset の両方が提出された場合は text_offset のみが評価に利用される
  - line_id, offsetは0オリジンの整数値

タスク固有データ③

属性値リンキングタスク

リンキングタスク: 公開データ

- 訓練用ラベル付きデータ:
  - W2017一部記事(1,208ページ), W2019一部記事 (481ページ)に対して ENE9.0ラベル、全属性値とそのリンク先のラベルが付与されたデータ (**公開中**) (抽出属性値数: 59,666、有効リンク数: 42,606)
  - 全ページ数: 1,689, リンキング対象の属性値数: 68,956, 有効リンク数: 50,565
  - 各ページIDに対応するHTML, プレーンテキストは訓練用Zipファイルに同梱
- 個別タスク評価用データ:
  - W2021から選出されたリーダーボード評価用データ (近日公開)
  - W2021から選出された本評価用データ (近日公開)
- タスク内容
  - W2021の記事のページID・タイトル・ENE9.0ラベル・属性値の一覧入力を元に、適切なリンク先が存在する場合にはリンク先としてページIDを付与、存在しない場合には null 値を付与

リンキングタスク: 入出力フォーマット

訓練データ例:

```
{
    "page_id": "2725947",
    "title": "フォッカー F.XXXVI",
    "ENE": "1.7.18.3",
    "attribute": "種類",
    "text_offset": {
        … // 属性値抽出と同様
        "text": "旅客機"
    },
    "html_offset": {
        … // 属性値抽出と同様
    },
    "link_page_id": "80372",  // リンク対象ページID
    // 以下は過去タスクからの慣習で存在するが今回は不要
    "link_type": {
      "later_name": false, "part_of": false, "derivation_of": false
    }
}
```

評価データ例:

```
{
    "page_id": "2240470",
    "title": "ユーロコプター EC 155",
    "ENE": "1.7.18.3",
    "attribute": "バリエーション",
    "text_offset": {
        … // 属性値抽出と同様
        "text": "EC 155B1"
    },
    "html_offset": {
        … // 属性値抽出と同様
    },
}
```

提出データ例:

```
{
    "page_id": "2240470",
    "title": "ユーロコプター EC 155",
    "ENE": "1.7.18.3",
    "attribute": "バリエーション",
    "text_offset": {
        … // 属性値抽出と同様
        "text": "EC 155B1"
    },
    "html_offset": {
        … // 属性値抽出と同様
    },
    // 推定されたリンク対象ページID
    // 存在しないと判断した場合は
    // "link_page_id": null を明示的に指定
    "link_page_id": "2240470"
}
```

- 評価ポイント:
  - page_id, title, ENE, attribute, offset, link_page_idの全てが一致した場合に正解とみなす (対象ページが存在しないことを正しく推定した場合も正解にカウント)
  - 対象ページが存在しないと判断した場合は明示的に "link_page_id": null を指定して出力

まとめ

- 森羅2023特設ページから全てDL可能: https://2023.shinra-project.info/
- 全タスク共通データ (2019年版/2021年版Wikipedia全記事データ, 4つのフォーマット)

	Cirrus Search Dump (JSON Lines)	Wiki Dump (XML)	HTML	Plain Text
W2019 (訓練用)	全記事	全記事	全記事	一部のみ (属性値抽出・リンキングの訓練データ対象記事のみ)
W2021 (推論対象)	全記事	全記事	全記事	全記事

- カテゴリ・属性・タスク定義データ (JSON Lines形式で1つのファイルに集約)
- サブタスク固有データ (各データは JSON Lines 形式、昨年タスクから細かい変更あり)

	カテゴリ分類	属性値抽出	リンキング	End-to-End
W2019 訓練データ	W2019全記事 ラベル付き	W2017+W2019一部記事 ラベル付き	W2017+W2019一部記事 ラベル付き	←の各タスクデータの活用を想定
W2021 リーダーボード	W2021一部記事 ラベル無し	W2021一部記事 カテゴリ分類ラベル付き 属性値ラベル無し	W2021一部記事 カテゴリ分類・属性ラベル付き リンキングラベル無し	W2021一部記事 ラベル無し

※ 昨年と異なり属性値抽出・リンキングでEnd-to-End前提の評価は撤廃