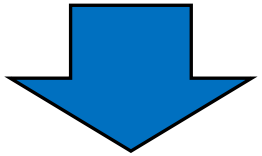
Wikipedia構造化プロジェクト
構造化手法について

2018/10/18

豊橋技術科学大学　応用数理ネットワーク研究室

研究背景

色々な手法で作成したデータがある方が良い

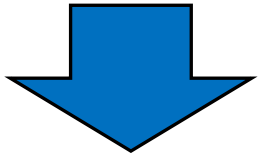


ルールベースで出来る範囲で頑張る！

属性値抽出の流れ

- 使用データ : Wikipediaのダンプデータ(JSON)を使用

データの整形



各属性ごとに，ルールベースでの属性値抽出

データの整形

ページ番号と記事のデータをまとめる

整形前（配布された Wikipedia のダンプデータ）

```
{"index":{"_type":"page","_id":"3240437"}}
{"template":["Template:各年の文学ヘッダ", ...}
```

整形後

```
{"index":{"_type":"page","_id":"3240437"},
 "template":["Template:各年の文学ヘッダ", ...}
```

不要な文字列の除去

- ファイルに埋め込まれているBOMやノーブレークスペース等の不要な文字列

→　sed等を用いて除去

Infoboxの取り出し

source_text

```
{{基礎情報 会社\n|社名=ウォルマート\n|英文社名=Wal-MartStores,Inc.\n|ロゴ=[[ファイル:Walmart logo.svg|270px]]\n|種類=[[公開会社]]\n|市場情報={{上場 情 報|NYSE|WMT}}\n|国籍={{USA}}\n|本社郵便番号=\n|本社所在地=[[アーカンソー州]][[ベントンビル (アーカンソー州)|ベン
                    ⋮
```

⇒

JSON 形式に変換

```
"infobox":{
  "社名":"ウォルマート",
  "英文社名": "Wal-Mart...",
  "ロゴ": "[[ファイル:Wal...",
  "種類": "[[公開会社]]",
  "市場情報":"{{ 上場情...",
  "国籍": "USA",
  "本店所在地":"[[デラウ...",
  "設立": "[[1969 年]][[10...",
  "業種": "小売業",
                    ⋮
}
```

Wiki記法等の除去

- Wiki記法等の除去に用いた正規表現

```
.*<em>(.+)</em>.*
¥{¥{(nobold|[Ss]mall(er)?|my)¥|([^|}]+)(¥}¥})?
¥{¥{[Ll]ang¥|[^|]+¥|(.+)(¥}¥})?
<!--.+-->
¥{¥{仮リンク¥|([^|]+)¥|[a-z]{2}¥|([^|]+)¥}¥}
¥{¥{[Rr]¥|[a-zA-Z0-9¥|]+¥}¥}
(:[a-z]+(-[a-z]+)?:)([A-Za-z  ]+)
¥(¥(.+?¥)¥)
<ref>.+</ref>
</?[A-Za-z]+>
^:?[A-Z]?[a-z]+(¥||:)(.+)$
^.+¥|
```

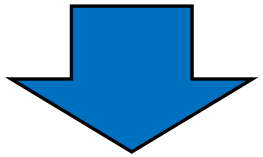
属性値の抽出：市区町村カテゴリ

ふりがな

- “opening_text”がある場合
  →最初に出現した“()”の中を抽出

    例 : “安土町(あづちちょう)は、滋賀県の東部…”
    → [“あづちちょう”]

- “opening_text”がない場合
  →Infoboxの“よみがな”, “正式名称”等のキーから値を抽出



- ひらがな・カタカナ以外の文字を含む場合は候補から外す

人口密度

- Infoboxに“^人口密度.*”にマッチするキーが
  - ある場合　→取り出した文字列に単位をつけて解答とした

    例：{“人口密度（平方キロ当たり）”：“87”}
    → [“87人/㎢”]

  - ない場合　→本文から以下の正規表現にマッチする文字列を抽出

“(¥d[¥d,.]+*人*/((km|mi)[2²]|㎢)”

ただし，複数存在する場合は最初の2つを採用

座標・緯度

- Infoboxの “緯度”, “latitude”キーから抽出

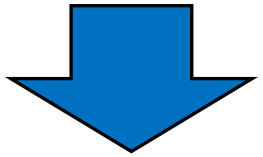
  例 : {“緯度” : “北緯34度46分6秒”}

  → [“北緯34度46分6秒”]

- 本文から以下の正規表現にマッチする文字列を抽出

  “(北|南)緯[¥d.度分秒]+)”

  “(¥d+度¥d+分¥d+秒～¥d+度¥d+分¥d+秒)”



- 解答候補の最初3つを解答とした

属性値の抽出：人名カテゴリ

ふりがな

- 以下に示すInfoboxキーの内, かな文字の連続を解答とする
  - ふりがな
  - カタカナ表記
  - 和名
  - 発音

例 : グッチ裕三
→“ふりがな” : [“ぐっち ゆうぞう”]
→[“ぐっち ゆうぞう”]

- ただし, 人名の区切りを示す記号[a]も認めている
- 見出し語と同じ場合は無視する

[a] “=”や“=”, “・”

両親

- 両親は義理の関係も存在するため，それらも取得
- Infobox値の中に”別名”も併せて記載されている場合はそれらも同時に抽出
  - 父母
  - (養｜義)?(父｜母)親?

例：後鳥羽天皇
→ {“父親”, [“[[高倉天皇]]”],
“母親”, [“[[坊門殖子]]（七条院）”]}
→ [“高倉天皇”, “七条院” , “坊門殖子”]

別名

- リングネームなど職業独特のものに全て対応するわけにはいかないので，”～ネーム”など別名を表す属性の要素も抽出

  - 渾名
  - ニックネーム
  - 字
  - リングネーム
  - etc…

例：トマス・コクラン
→ “nickname” : [“海の狼”]
→ [“海の狼”]

生年月日

- 紀元前や閏，年号などを考慮しつつ，日付表現を抽出
- “生年月日と年齢”テンプレート[a] からも抽出

例 : スティーヴ・ライヒ

→ “Born” : [“{{生年月日と年齢|1936|10|3}}…”]

→ [“1936年10月3日”]

- 誕生日
- 生年月日
- 生誕
- 出生
- birth_date
- etc…

[a] {{生年月日と年齢 | 1993 | 2 | 4}}

属性値の抽出 : 空港名カテゴリ

IATA, ICAO　両空港コード

- おおよそInfoboxに値として存在しているため，それを取得
  - IATA　→ 3文字
  - ICAO　→ 4文字

例 : 隠岐空港
{ “IATA” : [“OKI”], ICAO : [“RJNO”] }
→ [“OKI”], [“RJNO”]

国

- Infoboxのキー“国”を抽出
- Wikipediaで使用される国名コードから国名への変換表を用意し，それを利用して変換を行う
- 国名コードは正規表現

  “¥{¥{([A-Z¥-]+[0-9]*)¥}¥}”

  で示される文字列である必要がある

  例 : 隠岐空港

  → “国” : [“{{JPN}}”] → [“日本”]

座標・緯度

- Infobox内部では素直に記載されていない場合が多い
- 分割して記載されていることがある
  →これらに対応して抽出
- 表記法が複数あるが,示されているものは全て抽出
  - 緯度
  - (緯 | 経)度 (度| 分 | 秒)

例 : ピサ空港
→ {“緯度度”: [“43”], “緯度分”: [“41”],
“緯度秒” : [“02”], “緯度NS” : [“N”]}
→[“北緯43度41分02秒”]

年間(発着回数｜利用者数)データの年

- Infoboxのキーにどの年かが記載されていないため，
  どちらも以下のデータから採用している

  - 統計年
  - stat-year

例 : キャンベラ国際空港

→ “統計年” : [“2011年”]

→ [“2011年”]

その他

- 手を出せていない属性が多数ある
  - 名称由来
  - 名称由来人物の地位職業名
  - 旧称
  - 滑走路数
  - 近隣空港

属性値の抽出 : 企業名カテゴリ

売上高(単体，連結)

- Infoboxの“売上高”, “revenue”キーから抽出
- 金額に加えて，その金額が”単体”か”連結”かを判別する

例 : “単体 : 1000億円
連結 : 2000億円”

- (1)HTMLタグの“
”や読点を区切り文字とし，テキストを分割

“単体 : 1000億円
 連結: 2000億円”

→ “単体 : 1000億円”,
“連結 : 2000億円”

売上高(単体，連結)

- (2)単体か連結かを判断する手がかり語を探す

“単体：1000億円”
“連結：2000億円”

手がかり語

単体・・・“単体”, “単独”
連結・・・“連結”

- (3)金額を表す以下の正規表現にマッチした部分を抽出

“¥d.*円”

“単体：1000億円”
“連結：2000億円”
→
売上高(単体)：[“1000億円”]
売上高(連結)：[“2000億円”]

従業員数(単体，連結)

- Infoboxの“従業員数”キーから抽出
- 数字” + “名” or “人”の部分を抽出
- 単体か連結かを判断する方法は，売上高と同様に
  - (1)テキストを分割
  - (2)手がかり語を探す

例 :
{“従業員数” : “単体 : 1000人
連結 : 2000人”}
→ 従業員数(単体) : [“1000人”]
従業員数(連結) : [“2000人”]

設立年

- Infoboxの“設立”, “設立日”キー等から抽出
- 正解データの属性値は
  - YYYY年MM月DD日
  - YYYY年(元号YY年)MM月DD日

の形式が多い

例 : {“設立” : [“1937年8月28日”]}
→ [“1937年8月23日”]

この形式にマッチングするような正規表現を用いる

“¥d{0,4} 年 ((明治|大正|昭和|平成)¥{0,2} 年)*¥d{1,2} 月*¥d{1,2} 日*”

○○データの年

- 売上高，従業員数，資本金など
- 日付の表現＋“期”で書かれる
  - YYYY年MM月期
  - YYYY年(元号YY年)MM月期

この形式にマッチングするような正規表現を用いる

“¥d{0,4} 年 ((明治|大正|昭和|平成)¥{0,2} 年）)*¥d{1,2} 月*¥d{0,2} 日*期”

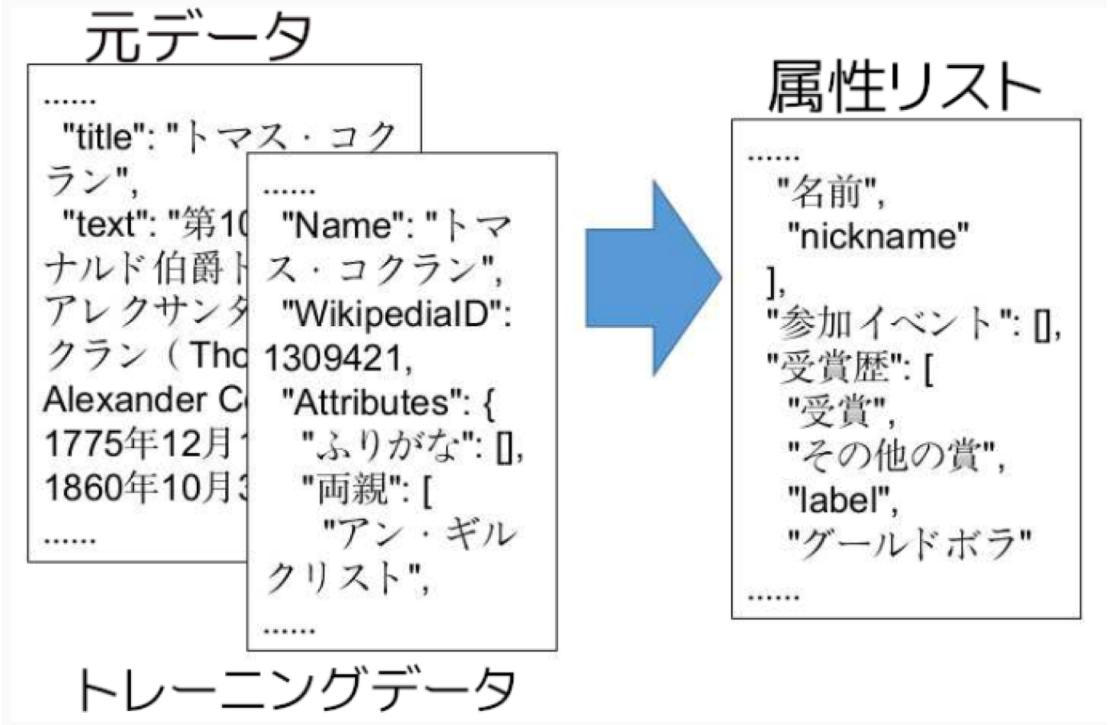
例：“資本金”：“1000万円(2016年3月期)”
→ [“2016年3月期”]

その他

- Infoboxのみでは抽出できないような属性がいくつかあった
  - 起源
  - 取扱商品
  - 業界内地位・規模
  - etc…
- これらに取り組んでいるグループもいらっしゃるので,
  手法が気になります！

前提

- 提供された解答データ(トレーニングデータ)
  → Infobox属性との対応



前提

- 人名と同様にInfoboxの属性値を確認
- 年月表現などは分割して属性値となる場合があるので注意
- 数値の単位は記述されていないことがあるので補完

例 : “標高 m = 89.4”
→ [“89.4m”]

属性名の逆引き

- ダンプデータの属性名と正解データの属性名が違う問題

例：　正解データ　　→ “子会社・合弁会社”
　　ダンプデータ　→ “主要子会社”

- 正解データの属性値から，ダンプデータのInfoboxの属性名を逆引きし，正解データの属性値を含む属性を属性名とした.

正解データ

{…, “子会社・合弁会社” : [“A社”], …}
{…, “子会社・合弁会社” : [“B社”], …}

ダンプデータ

{…, “主要子会社” : [“A社”, “C社”], …}
{…, “子会社” : [“B社”], …}