Wikidataによる文章アノテーションシステム

大阪電気通信大学・情報通信工学部・情報工学科
古崎研・瀬野匡史

アプリケーションリンク

Wikidataによる文章アノテーションシステムとは？

- WikidataやDBpediaなどのLODを利用した実用的なシステムを作成したいというところから探し始めDBpedia Spotlightを見つけ、これのWikidata版を作成することが出来ないかというところから開発を始めました。
- 文章を入力すると、名詞にWikidataのハイパーリンクテキストを付与して文章を出力します。
- オプションとしてカテゴリーの選択、ラベル検索設定、名詞のフィルタリングなどを用意しています。

実際にWikipediaの日本のトップ記事をアノテーションしてみた

Wikidataによる文章アノテーションシステム

使い方(githubレポジトリ)

カテゴリー:

- [ ] 活動
- [ ] 解剖学的構造
- [ ] 疾患
- [ ] 薬
- [x] 組織
- [ ] 人物
- [x] 場所
- [ ] 化学物質
- [ ] 生物
- [ ] 仕事

例文1 例文2 例文3

▸ 詳細設定

日本国（にほんこく、にっぽんこく、英: Japan）、または日本（にほん、にっぽん）は、東アジアに位置する民主制国家[1]。首都は東京都[注 3][2][3]。全長3500キロメートル以上にわたる国土は、主に日本列島[注 7]および千島列島・南西諸島・伊豆諸島・小笠原諸島などの弧状列島により構成され[3][4]、大部分が温帯に属するが、北部や島嶼部では亜寒帯や熱帯の地域がある[5][6]。地形は起伏に富み、火山地・丘陵を含む山地の面積は国土の約75 % を占め[6]、沿岸の平野部に人口が集中している。国内には行政区分として47の都道府県があり、日本人（大和民族・琉球民族・アイヌ民族[注 8]・外国系の人々）と外国人が居住し、日本語を通用する[2][3]。明治維新後の1889年（明治22年）に大日本帝国憲法を制定し立憲国家となる。太平洋戦争後の連合国による占領期の1947年（昭和22年）には現行の日本国憲法を施行し民主制へ移行。1952年（昭和27年）、

アノテーション実行

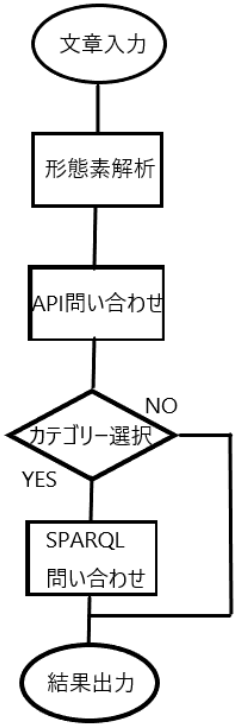
アノテーション結果

click

日本国（にほんこく、にっぽんこく、英: Japan）、または日本（にほん、にっぽん）は、東アジアに位置する民主制国家[1]。首都は東京都[注 3][2][3]。全長3500キロメートル以上にわたる国土は、主に日本列島[注 7]および千島列島・南西諸島・伊豆諸島・小笠原諸島などの弧状列島により構成され[3][4]、大部分が温帯に属するが、北部や島嶼部では亜寒帯や熱帯の地域がある[5][6]。地形は起伏に富み、火山地・丘陵を含む山地の面積は国土の約75 % を占め[6]、沿岸の平野部に人口が集中している。国内には行政区分として47の都道府県があり、日本人（大和民族・琉球民族・アイヌ民族[注 8]・外国系の人々）と外国人が居住し、日本語を通用する[2][3]。明治維新後の1889年（明治22年）に大日本帝国憲法を制定し立憲国家となる。太平洋戦争後の連合国による占領期の1947年（昭和22年）には現行の日本国憲法を施行し民主制へ移行。1952年（昭和27年）、サンフランシスコ平和条約により主権を回復した[3]。

アノテーションシステムのフローチャート

- まず初めにテキストを入力したら、形態素解析を行い、アノテーション対象の名詞を抽出します。
- 次に抽出した名詞をMediaWikiのAPI[1][2]に問い合わせて、WikidataのIDやURLなどの情報を取得します。
- そして、カテゴリーが選択されていたのなら名詞にそのカテゴリーに該当しているものがないかSPARQLで問い合わせます。
- 最後にテキストにハイパーリンクを付与、文章を組み立て出力します。



評価

- 評価方法としてWikipediaの記事を使用し、トップ記事を完全一致検索でアノテーションして再現率・適合率を求めました。アノテーションするWikipediaの記事はトップ記事の文字数が150文字以上あるものをランダムで選出しました。
- 再現率はWikipediaがアノテーションしているものをアノテーションシステムが同じものをどれだけアノテーション出来ているかというものである。
- 適合率はアノテーションシステムがアノテーションしたものをWikipediaがどれだけ同じものをアノテーションしているかというものである

Wikipedia記事との比較

日本国（にほんこく、にっぽんこく、英: Japan）、または**日本**（にほん、にっぽん）は、東アジアに位置する民主制国家[1]。首都は東京都[注 2][2][3]。

全長3500キロメートル以上にわたる国土は、主に日本列島[注 6]および千島列島・南西諸島・伊豆諸島・小笠原諸島などの弧状列島により構成され[3][4]、大部分が温帯に属するが、北部や島嶼部では亜寒帯や熱帯の地域がある[5][6]。地形は起伏に富み、火山地・丘陵を含む山地の面積は国土の約75%を占め[6]、沿岸の平野部に人口が集中している。国内には行政区分として47の都道府県があり、日本人（大和民族・琉球民族・アイヌ民族[注 7]・外国系の人々）と外国人が居住し、日本語を通用する[2][3]。

上:Wikipedia
下:アノテーションシステム

:厳密に一致

:リダイレクトや厳密に一致しないものの正解と言えるもの

日本国 （にほんこく、**にっぽん** こく、**英** : **Japan** ）、または**日本** （にほん、**にっぽん** ）は、**東アジア** に**位置** する**民主制** **国家** [1]。**首都** は**東京都** [注 2][2][3]。 **全長** 3500**キロメートル** 以上にわたる**国土** は、主に**日本列島**[注 6]および**千島列島** ・**南西諸島** ・**伊豆諸島** ・**小笠原諸島** などの**弧状列島** により**構成** され[3][4]、大**部分** が**温帯** に属するが、**北部** や**島嶼** **部** では**亜寒帯** や**熱帯** の**地域** がある[5][6]。**地形** は起伏に富み、**火山** **地** ・**丘陵** を含む**山地** の**面積** は**国土** の約75 **%** を占め[6]、**沿岸** の**平野** **部** に**人口** が集中している。国内には**行政区分** として47の**都道府県** があり、**日本人** （**大和民族** ・**琉球民族** ・**アイヌ民族**[注 7]・**外国** **系** の**人々** ）と**外国人** が**居住** し、**日本語** を通用する[2][3]。

- 評価数は少ないですが、この結果の再現率は 既存研究結果[3]のエンティティタイプの予測結果が64％となっているところを鑑みるとあまり悪くないのではと思います。
- 適合率はあまりいい結果が出ませんでした。基本的にアノテーションシステムでアノテーションされる単語はWikipediaよりかなり多いのでこの評価方法では低くなってしまいます。

	ページ名	URL	再現率(%)	適合率(%)		再現率平均(%)	適合率平均(%)
2	コネティカットのひょこひょこおじさん	https://ja.	40	19		46.3	22.2
3	馬王堆漢墓	https://ja.	40	40			
4	プレミアリーグ (バレーボール)	https://ja.	66	16			
5	太刀川正三郎	https://ja.	71	17			
6	ワカオライデン	https://ja.	47	41			
7	ローデシアン・リッジバック	https://ja.	25	7			
8	モンギ・スリム	https://ja.	46	38			
9	フェアチャイルドセミコンダクター	https://ja.	75	16			
10	ジャッキー吉川とブルー・コメッツ	https://ja.	28	7			
11	カミラ (イギリス王妃)	https://ja.	25	21			

まとめ

- 再現率はそこそこいい結果が得られたと考えています。
- 適応率が低いので適応率を上げるために、Wikipediaでは同じ単語は一度しかアノテーションされていないので一度アノテーションした単語は再度アノテーションしないよう変更する。そしてアノテーションする名詞を抽出する設定はどれが適切なのか検討したいと思っています。
- そして今回適応率を評価するときにテキストと意味が合っているアノテーションが出来ていたとしてもWikipediaに載っていなければ不正解となっているため、その部分を人の手で評価出来たらと考えています。

参考文献

- [1] MediaWiki API help action=wbsearchentities
  https://www.wikidata.org/w/api.php?action=help&modules=wbsearchentities
- [2] MediaWiki API help list=search (sr)
  https://www.mediawiki.org/w/api.php?action=help&modules=query%2Bsearch
- [3] Knowledge Graph Entity Type Prediction with Relational Aggregation Graph Attention Network
  https://2022.eswc-conferences.org/wp-content/uploads/2022/05/paper_45_Zou_et_al.pdf