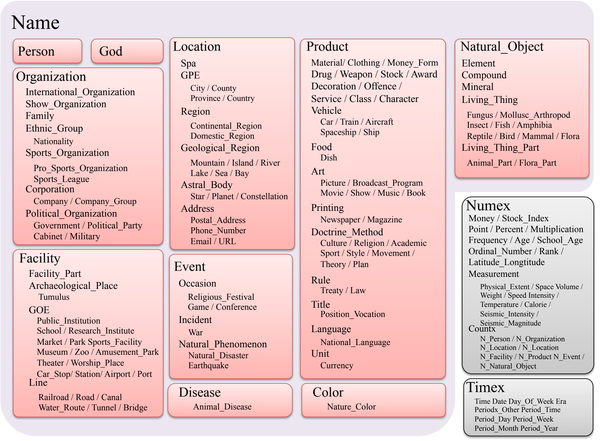

拡張固有表現定義
（https://sites.google.com/site/extendednamedentity711）

属性名	説明
SID	本データにおけるID
wikipedia_ID	WikipediaのID
Entry	項目名
Clean_entry	標準化された項目名
Page_property	ページの種類
Redirect_to	リダイレクト先
Redirect_from	リダイレクト元
Link_from-N	被リンク数
Link_anchor	リンク元のアンカー文字列
Category_info	Wikipediaのカテゴリ情報
First_sentence	説明文中の最初の文（自動抽出）
Listed_in	リストされている一覧ページ

Wikipedia 構造化データ


```
{
 "SID": 161224,
 "wikipedia_ID": "259974",
 "entry": "東京都立新宿高等学校",
 "clean_entry": "東京都立新宿高等学校",
 "page_property": "Normal",
 "redirect_to": "",
 “redirect_from”: [“新宿高校”, “新宿高等学校”, “都立新宿高等学校”, “東京都立新宿高校", "東京府立第六中学校"],
 "link_from_N": 276,
 "link_anchor": [{"count": 1, "anchor": "都立六中"}, {"count": 1, "anchor": "都立新宿高"}, {"count": 11, "anchor": "東京府立第六中学校"}, {"count": 2, "anchor": "新宿高等学校"}, {"count": 16, "anchor": "新宿"}, {"count": 1, "anchor": "東京都立第六高等学校"}, {"count": 4, "anchor": "東京府立六中"}, {"count": 1, "anchor": "旧制東京都立第六中学校"}, {"count": 2, "anchor": "新宿高"}, {"count": 4, "anchor": "府立六中"}, {"count": 8, "anchor": "都立新宿高校"}, {"count": 8, "anchor": "新宿高校"}, {"count": 216, "anchor": "東京都立新宿高等学校"}, {"count": 1, "anchor": "東京都立新宿高校"}],
 “category_info”: [“東京都区部の公立高等学校|しんしゆく”, “新宿区の学校|しんしゆくこう”, “学校記事”],
 “first_sentence”: “東京都立新宿高等学校(とうきょうとりつ しんじゅくこうとうがっこう)は、東京都新宿区内藤町に所在する都立高等学校。",
 “listed_in”: [“旧制中等学校・新制高校のナンバースクール一覧”, “東京都立新宿高等学校の人物一覧", "東京都高等学校一覧", "旧制中等教育学校の一覧 (東京都)"]
}
```

関根聡、安藤まや　（ランゲージ・クラフト）
松田耕史、鈴木正敏、乾健太郎　（東北大学）

質問応答、情報抽出、対話、機械翻訳等に必要な名前に関する知識。
Wikipediaの全項目に拡張固有表現をタグ付したデータです。
約2万項目は人手で作成、残りは機械学習でタグ付けをしています。
人手チェックを継続するため、一部を有償で公開します。

属性名	説明
SID	本データにおけるID
ENE	拡張固有表現情報
annotation_flag	アノテーション情報

ENE+W データ


```
{
 "SID": 161224,
 "ENE": ["学校名"],
 "annotation_flag": "HAND.LC_annotator_201511”
}
{
 "SID": 161225,
 "ENE": ["美術博物館名"],
 "annotation_flag": "AUTO.TOHOKU_201601”
}
{
 "SID": 161226,
 "ENE": ["CONCEPT"],
 "annotation_flag": "AUTO.TOHOKU_201601”
}
{
 "SID": 161227,
 "ENE": ["地名_その他"],
 "annotation_flag": "AUTO.TOHOKU_201601”
}
{
 "SID": 161228,
 "ENE": ["郡名"],
 "annotation_flag": "AUTO.TOHOKU_201601”
}
```


データの統計情報

説明	データ数
付与データ数	22,677
IGNORED	621
CONCEPT	2,660

説明		データ数
付与された拡張固有表現の種類数	1	21,624
	2	850
	3	187
	4	14
	6	2

高頻度の拡張固有表現定義

人名	**4,041**	**競技会名**	**625**
番組名	2,395	プロ競技組織名	484
企業名	1,701	地位職業名	462
市区町村名	975	映画名	438
製品名__その他	964	公園組織名	363
日付表現	916	学校名	326
文学表現	909	主義方式名__その他	288

低頻度の拡張固有表現定義

頻度	種類数	例
0	55	URL、人数、温度、絵画名
1	8	船名、恒星名、時刻表現
2-5	16	運河名、取引所名、橋名、罪名
6-10	23	地震名、条件名、学齢、湾名
11-20	23	公共機関名、例祭名、国籍名

判断の材料

項目の分類は文脈情報がないので曖昧性解消が難しい。複数の分類を許すが、説明文の最初の文に現れているものだけに限定する。そうしない場合には、非常にマイナーな分類が可能になってしまう。

Wikipediaにおける偏在

一般の人が自由に項目を立てられるという特徴から、番組名、ゲーム、ブランド、プログラム言語などの製品名__その他、体制、方式、企画などの主義方式名__その他の項目が極めて多かった。

システマティック・ポリセミー

文学名であると同時に、映画名、番組名であるようなもの。魚名であると同時に食べ物名であるような例が非常に多い。エンティティーに固有のものではなく、カテゴリに共有して見られる曖昧性が存在する。

被リンク数のバイアス

人手でのタグ付与する項目の基準として、被リンク数を用いたが。しかし、リンク構造には大きな偏りがあり、リンク数が興味の度合いを示しているとは限らない。ページ閲覧数などを使うということが考えられる。

一般表現

拡張固有表現と判断できない一般的表現は「CONCEPT」として明示的に分類する。

拡張固有表現に不足しているもの

拡張固有表現に不足しているものとして、王朝名、ダンス、音楽、ゲームの種類名、恐竜名などが見つかった。