SHINRA2020-ML

Categorizing 30 language Wikipedia into Extended Named Entity categories

Overview of SHINRA2020-ML

Motivation

Explainable NLP applications need structured Knowledge Base

This Task

Categorize entities in 30 language Wikipedia into ENE (name ontology) using the categorized Japanese Wikipedia and the language links

Resource by Collaborative Contribution

The KB will be created using the outputs of the share-task

森羅 SHINRA
Structured Knowledge, built on Wikipedia and Extended Named Entities
Center for Advanced Intelligence Project, Riken, Japan

Structured KB

Toronto (CITY)	
country	Canada
…	
airport	Pearson Inter. Airport

Pearson Inter. Airport (AIRPORT)	
Location	Ontario
…	
Name Origin	Lester B. Pearson

Lester B Pearson (PERSON)	
birthday	23 April 1897
…	
Career	Ambassador to US, One of the founders of UN

DBpedia Freebase WIKIDATA yago BabelNet

Existing KB are very noisy

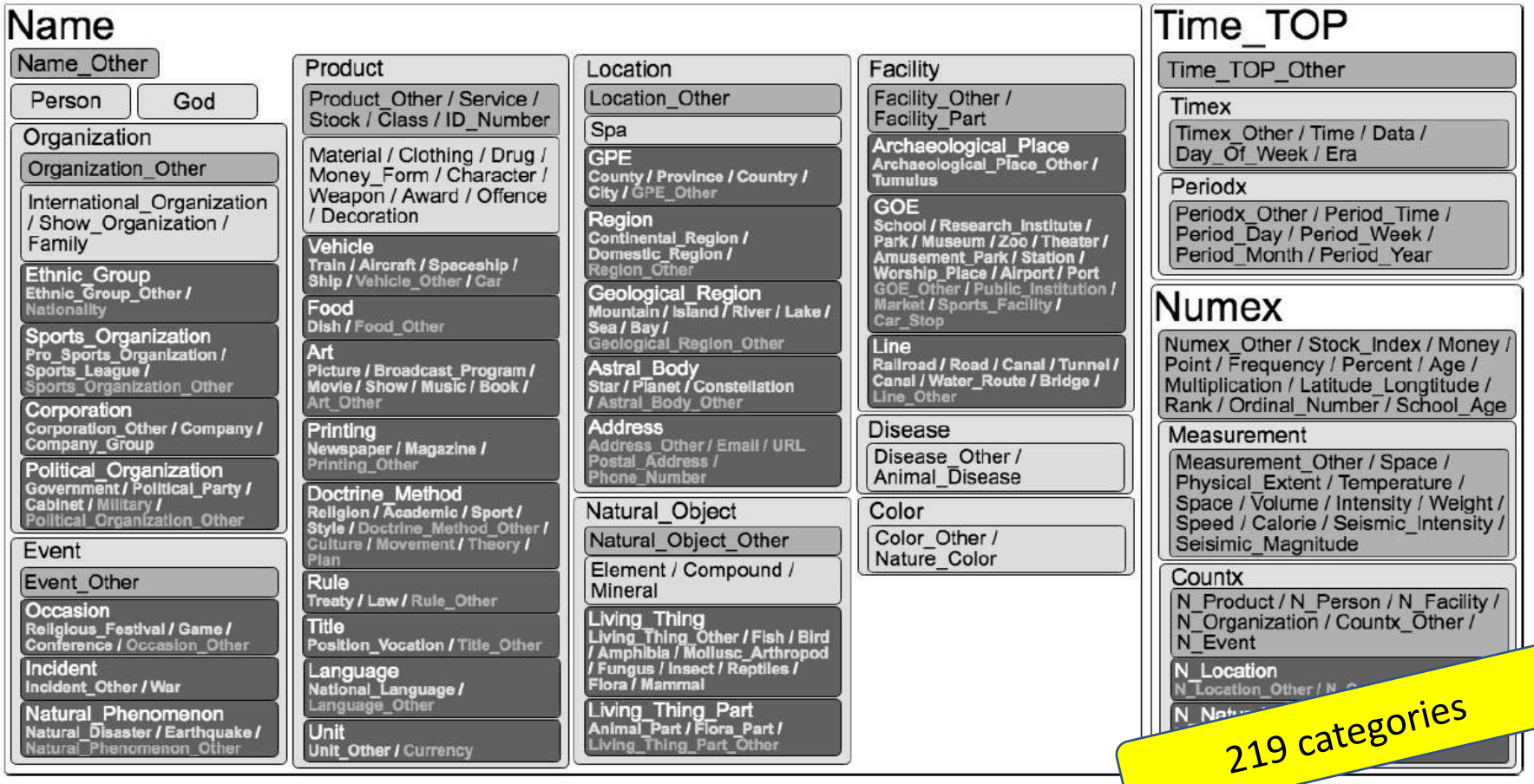
森羅
SHINRA
Structured Knowledge, built on Wikipedia and Extended Named Entities
Center for Advanced Intelligence Project, Riken, Japan

Design and Contents

- Top-down design
  - Use a name ontology well designed
- Bottom-up population
  - Populate by Crowdsourcing or by AI

森羅 SHINRA
Structured Knowledge, built on Wikipedia and Extended Named Entities
Center for Advanced Intelligence Project, Riken, Japan

Extended Named Entity (name ontology)



森羅
SHINRA
Structured Knowledge, built on Wikipedia and Extended Named Entities
Center for Advanced Intelligence Project, Riken, Japan

The TASK

Japanese Wikipedia

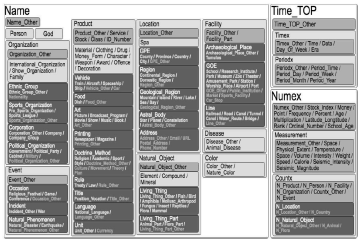
720K pages categorized

Family Name

City Name

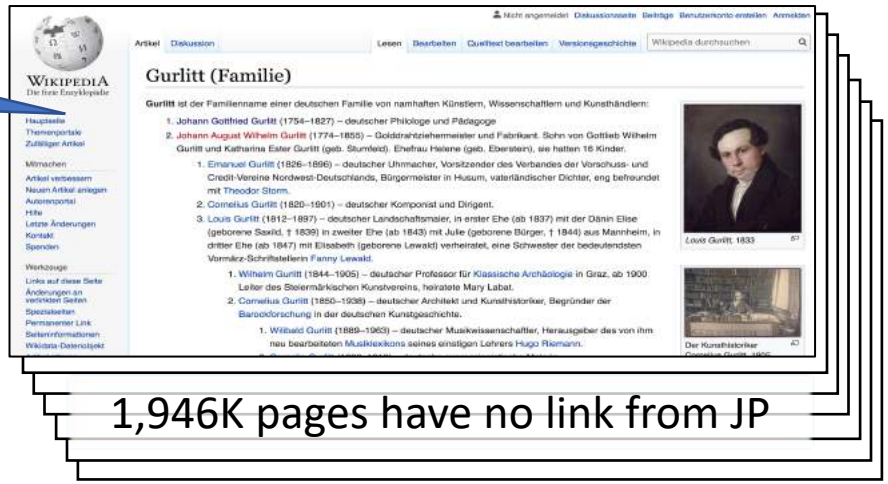
German Wikipedia

316K pages categorized by link



The task is to categorize all pages without link from JP

?



1,946K pages have no link from JP

森羅
SHINRA
Structured Knowledge, built on Wikipedia and Extended Named Entities
Center for Advanced Intelligence Project, Riken, Japan

The TASK

Japanese Wikipedia

720K pa

Family Name

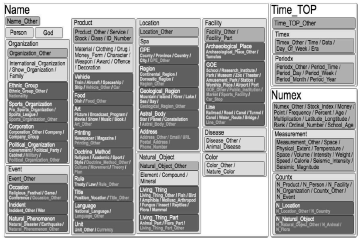
City Name

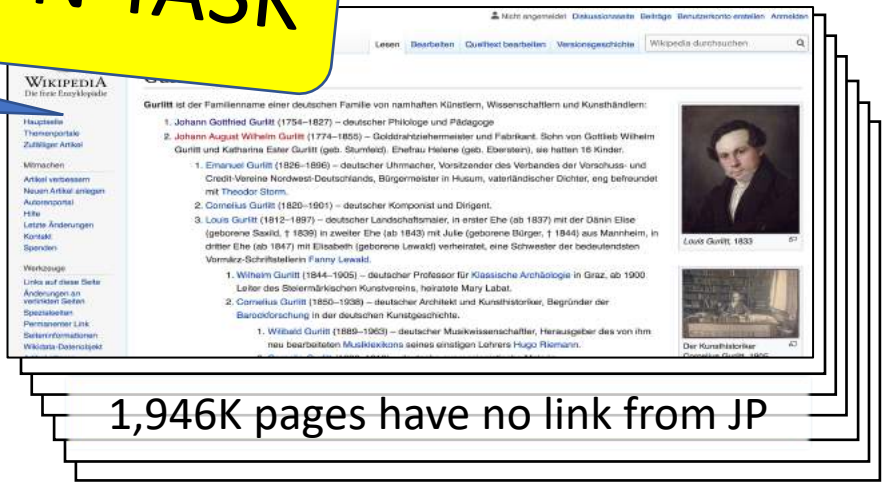
German Wikipedia

ages categorized by link

TEXT CATEGORIZATION TASK

The task is to categorize all pages without link from JP





1,946K pages have no link from JP

Target Languages

Language	Number of Users	Number of pages	Pages with links from jp	Link ratio
English	35,464,188	5,790,377	510,840	8.8
Spanish	5,289,422	1,500,013	283,539	18.9
French	3,334,739	2,074,648	359,783	17.3
German	3,101,292	2,262,582	316,652	14.0
Chinese	2,663,839	1,041,039	290,631	27.9
Russian	2,451,838	1,523,013	280,565	18.4
Portuguese	2,199,869	1,014,832	238,065	23.5
Italian	1,770,376	1,496,975	304,174	20.3
Arabic	1,611,381	661,205	75,773	11.5
Japanese	1,432,174	1,136,222	-	-
Indonesian	1,027,019	451,336	121,598	26.9
Turkish	1,021,218	321,937	118,107	36.7
Dutch	970,607	1,955,483	223,354	11.4
Polish	934,491	1,316,130	248,229	18.9
Persian	795,312	660,487	181,710	27.5
Swedish	652,290	3,759,167	200,555	5.3

Language	Number of Users	Number of pages	Pages with links from jp	Link ratio
Vietnamese	643,871	1,200,157	123,745	10.3
Korean	549,017	439,577	210,271	47.8
Hebrew	484,630	236,984	103,137	43.5
Romanian	461,670	391,231	98,897	25.3
Norwegian	450,588	501,475	144,751	28.9
Czech	440,040	420,195	137,144	32.6
Ukrainian	437,029	881,572	181,122	20.5
Hindi	425,415	129,141	31,828	24.6
Finnish	406,339	450,537	156,445	34.7
Hungarian	403,368	443,060	128,712	29.1
Danish	343,249	242,523	91,811	37.9
Thai	343,054	129,294	62,441	48.3
Catalan	312,980	601,473	150,829	25.1
Greek	265,153	157,566	63,427	40.3
Bulgarian	245,986	248,913	93,434	37.5

Data to be distributed

- Japanese Wikipedia categorized by Extend Named Entity [JSON]
  - excluding list articles, disambiguation pages, minor pages (less than 5 inter-link)
- Language links for 31 language Wikipedias [SQL]

- Wikipedia contents in 31 languages
  - Wikipedia Dump [XML]
  - Cirrus Search Dump [JSON]

- Extend Named Entity Definition (English/Japanese) [JSON]

※ The time stamp of All Wikipedia related data is January 20, 2019

Sample Data

Japanese Wikipedia entry [JSON]

{"pageid" : 347190,
"title" : "バッハ家",
"ENE_id" : "1.4.3", <- ”Family name”
"ENE_score" : 1,
"ENE_annotation" : "HAND.AIP_201904”}

Language links [JSON]

{"source" : { "pageid" : 347190, "lang" : "ja", "title" : "バッハ家" }, "destination" : { "lang" : "en", "title" : "Bach family", "pageid" : 713599 } }
{"source" : { "pageid" : 347190, "lang" : "ja", "title" : "バッハ家" }, "destination" : { "lang" : "es", "title" : "Familia Bach", "pageid" : 436582 } }
{"source" : { "pageid" : 347190, "lang" : "ja", "title" : "バッハ家" }, "destination" : { "lang" : "fa", "title" : " خاندان باخ“, “pageid" : 4515101 } }
{"source" : { "pageid" : 347190, "lang" : "ja", "title" : "バッハ家" }, "destination" : { "lang" : "fr", "title" : "Famille Bach", "pageid" : 813257 } }
{"source" : { "pageid" : 347190, "lang" : "ja", "title" : "バッハ家" }, "destination" : { "lang" : "it", "title" : "Bach (famiglia)", "pageid" : 2767443 } }

Schedule

2019

- October: Dataset distribution
  (at the final meeting of SHINRA2019-JP)

2020

- April: Registration deadline
- August: Evaluation
- December: Conference

Evaluation and Communication

- Prepare around 500-1,000 data each below for all languages that don't have a link from Japanese Wikipedia
  - Test Data
  - Leaderboard Data
- We will prepare …
  - Leaderboard
  - Slack among the participants and the organizers
- NTCIR15 page
  - http://research.nii.ac.jp/ntcir/ntcir-15/index.html (English)