

SHINRA2020-ML

Categorizing 30 language Wikipedia into Extended Named Entity categories

森羅

SHINRA

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



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

SHINRA

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



Existing KB are very noisy

Design and Contents

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





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

SHINRA

Extended Named Entity (name ontology)

Name Name_Other Person God Organization Organization_Other International_Organization / Show_Organization / Family Ethnic_Group Ethnic_Group_Other / Nationality Sports_Organization Pro_Sports_Organization / Sports_League / Sports_Organization_Other Corporation Corporation_Other / Company / Company_Group Political_Organization Government / Political_Party / Cabinet / Military / Political_Organization_Other Event Event_Other Occasion Religious_Festival / Game / Conference / Occasion_Other Incident Incident_Other / War Natural_Phenomenon Natural_Disaster / Earthquake / Natural_Phenomenon_Other	Product Product_Other / Service / Stock / Class / ID_Number Material / Clothing / Drug / Money_Form / Character / Weapon / Award / Offence / Decoration Vehicle Train / Aircraft / Spaceship / Ship / Vehicle_Other / Car Food Dish / Food_Other Art Picture / Broadcast_Program / Movie / Show / Music / Book / Art_Other Printing Newspaper / Magazine / Printing_Other Doctrine_Method Religion / Academic / Sport / Style / Doctrine_Method_Other / Culture / Movement / Theory / Plan Rule Treaty / Law / Rule_Other Title Position_Vocation / Title_Other Language National_Language / Language_Other Unit Unit_Other / Currency	Location Location_Other Spa GPE County / Province / Country / City / GPE_Other Region Continental_Region / Domestic_Region / Region_Other Geological_Region Mountain / Island / River / Lake / Sea / Bay / Geological_Region_Other Astral_Body Star / Planet / Constellation / Astral_Body_Other Address Address_Other / Email / URL / Postal_Address / Phone_Number Natural_Object Natural_Object_Other Element / Compound / Mineral Living_Thing Living_Thing_Other / Fish / Bird / Amphibia / Mollusc_Arthropod / Fungus / Insect / Reptiles / Flora / Mammal Living_Thing_Part Animal_Part / Flora_Part / Living_Thing_Part_Other	Facility Facility_Other / Facility_Part Archaeological_Place Archaeological_Place_Other / Tumulus GOE School / Research_Institute / Park / Museum / Zoo / Theater / Amusement_Park / Station / Worship_Place / Airport / Port / GOE_Other / Public_Institution / Market / Sports_Facility / Car_Stop Line Railroad / Road / Canal / Tunnel / Canal / Water_Route / Bridge / Line_Other Disease Disease_Other / Animal_Disease Color Color_Other / Nature_Color	Time_TOP Time_TOP_Other Timex Timex_Other / Time / Data / Day_Of_Week / Era Periodx Periodx_Other / Period_Time / Period_Day / Period_Week / Period_Month / Period_Year	Numex Numex_Other / Stock_Index / Money / Point / Frequency / Percent / Age / Multiplication / Latitude_Longitude / Rank / Ordinal_Number / School_Age Measurement Measurement_Other / Space / Physical_Extent / Temperature / Space / Volume / Intensity / Weight / Speed / Calorie / Seismic_Intensity / Seismic_Magnitude Countx N_Product / N_Person / N_Facility / N_Organization / Countx_Other / N_Event N_Location N_Location_Other / N_C... N_Nat...
--	---	---	--	---	--

220 categories

The TASK

Japanese Wikipedia

出典: フリー百科事典『ウィキペディア (Wikipedia)』

バッハ家またはバッハ一族（ドイツ語: Familie Bach）は、16世紀後半から18世紀までのドイツに属した、重要な音楽家の家系である。

目次 [非表示]

- 概要
- 家系図
- 2.1 系統
- 2.2 名之 系図
- 3 外部リンク

概要 [編集]

16世紀には4つの分家があり、そのうち、1561年にヴェヒマルのハンス・バッハは、中でも最も栄えたエアルト系統の祖フアイト・バッハの、おそらく父親として、その名が伝えられている。（ヴェヒマルは、アルンシュタットとゴータの中間に位置するテューリンゲン州の小都市である。）ヨハン・ゼバスティアン・バッハ（以下バハ）本人によって作成され、次男カール・フィリップ・エマヌエル・バッハによって補筆された『音楽家バッハ一族の起源』によると、フアイト・バッハは、パン焼き職人で鍛冶屋であり、一族の祖として、「水車がカタカタと音を立てる中、愛らしいシスターの音色を奏でていたに違いない。

しかし、一族で最初の職業音楽家は、フアイトの息子、大バッハの祖父となった。ちなみにバッハの父親はアムステルダムに在住していた。ヨハン・ニコラウス・バッハは、エスよ、われは汝を離れ(lass dich nicht) 16...

Family Name

...

City Name

German Wikipedia

Bach (Familie)

Die **Familie Bach** war ein weit verzweigtes Musikergeschlecht, aus dem von der zweiten Hälfte des 16. Jahrhunderts bis zur Mitte des 19. Jahrhunderts in Mitteldeutschland zahlreiche Stadtmusiker, Organisten und Komponisten entstammten. Der bekannteste und einflussreichste unter ihnen war Johann Sebastian Bach.

Man unterscheidet für diese **Künstlerfamilie** die beiden Hauptlinien:

- Erfurter Linie
- Arnstädter Linie

Die Familie Bach beherrschte in Erfurt über ein ganzes Jahrhundert das musikalische Leben derart, dass noch 1793 alle *Stadtpleifer „Bache“* genannt wurden, obwohl längst keiner dieses Namens mehr unter ihnen lebte.^[1] Allein in den Kirchenbüchern der Kaufmannskirche sind über 60 Kindtaufen, Hochzeiten und Begräbnisse der Familie registriert.

Inhaltsverzeichnis [Verbergen]

- 1 Stammliste
- 2 Siehe auch

720K pages categorized by link

1,946K pages categorized by link

TEXT CATEGORIZATION TASK

The task is to categorize all pages without link from JP

Name	Product	Location	Facility	Time
Person	Product	Location	Facility	Time
Organization	Product	Location	Facility	Time
Event	Product	Location	Facility	Time
Work	Product	Location	Facility	Time
Place	Product	Location	Facility	Time
Company	Product	Location	Facility	Time
Government	Product	Location	Facility	Time
Education	Product	Location	Facility	Time
Health	Product	Location	Facility	Time
Religion	Product	Location	Facility	Time
Science	Product	Location	Facility	Time
Arts	Product	Location	Facility	Time
Sports	Product	Location	Facility	Time
Media	Product	Location	Facility	Time
Transport	Product	Location	Facility	Time
Communication	Product	Location	Facility	Time
Energy	Product	Location	Facility	Time
Environment	Product	Location	Facility	Time
Weather	Product	Location	Facility	Time
Climate	Product	Location	Facility	Time
Geography	Product	Location	Facility	Time
History	Product	Location	Facility	Time
Law	Product	Location	Facility	Time
Politics	Product	Location	Facility	Time
Society	Product	Location	Facility	Time
Culture	Product	Location	Facility	Time
Demographics	Product	Location	Facility	Time
Economics	Product	Location	Facility	Time
Finance	Product	Location	Facility	Time
Business	Product	Location	Facility	Time
Industry	Product	Location	Facility	Time
Technology	Product	Location	Facility	Time
Information	Product	Location	Facility	Time
Telecommunications	Product	Location	Facility	Time
Internet	Product	Location	Facility	Time
Computing	Product	Location	Facility	Time
Software	Product	Location	Facility	Time
Hardware	Product	Location	Facility	Time
Engineering	Product	Location	Facility	Time
Construction	Product	Location	Facility	Time
Architecture	Product	Location	Facility	Time
Design	Product	Location	Facility	Time
Art	Product	Location	Facility	Time
Music	Product	Location	Facility	Time
Dance	Product	Location	Facility	Time
Theater	Product	Location	Facility	Time
Performing Arts	Product	Location	Facility	Time
Visual Arts	Product	Location	Facility	Time
Literature	Product	Location	Facility	Time
Journalism	Product	Location	Facility	Time
Writing	Product	Location	Facility	Time
Editing	Product	Location	Facility	Time
Translation	Product	Location	Facility	Time
Interpretation	Product	Location	Facility	Time
Teaching	Product	Location	Facility	Time
Learning	Product	Location	Facility	Time
Research	Product	Location	Facility	Time
Development	Product	Location	Facility	Time
Innovation	Product	Location	Facility	Time
Discovery	Product	Location	Facility	Time
Exploration	Product	Location	Facility	Time
Investigation	Product	Location	Facility	Time
Analysis	Product	Location	Facility	Time
Experimentation	Product	Location	Facility	Time
Observation	Product	Location	Facility	Time
Measurement	Product	Location	Facility	Time
Calculation	Product	Location	Facility	Time
Computation	Product	Location	Facility	Time
Simulation	Product	Location	Facility	Time
Modeling	Product	Location	Facility	Time
Forecasting	Product	Location	Facility	Time
Prediction	Product	Location	Facility	Time
Estimation	Product	Location	Facility	Time
Assessment	Product	Location	Facility	Time
Evaluation	Product	Location	Facility	Time
Verification	Product	Location	Facility	Time
Validation	Product	Location	Facility	Time
Confirmation	Product	Location	Facility	Time
Proof	Product	Location	Facility	Time
Demonstration	Product	Location	Facility	Time
Illustration	Product	Location	Facility	Time
Explanation	Product	Location	Facility	Time
Clarification	Product	Location	Facility	Time
Elucidation	Product	Location	Facility	Time
Interpretation	Product	Location	Facility	Time
Translation	Product	Location	Facility	Time
Adaptation	Product	Location	Facility	Time
Modification	Product	Location	Facility	Time
Alteration	Product	Location	Facility	Time
Change	Product	Location	Facility	Time
Transformation	Product	Location	Facility	Time
Conversion	Product	Location	Facility	Time
Reformation	Product	Location	Facility	Time
Reconstruction	Product	Location	Facility	Time
Restoration	Product	Location	Facility	Time
Renovation	Product	Location	Facility	Time
Improvement	Product	Location	Facility	Time
Enhancement	Product	Location	Facility	Time
Optimization	Product	Location	Facility	Time
Maximization	Product	Location	Facility	Time
Minimization	Product	Location	Facility	Time
Reduction	Product	Location	Facility	Time
Elimination	Product	Location	Facility	Time
Removal	Product	Location	Facility	Time
Deletion	Product	Location	Facility	Time
Erasure	Product	Location	Facility	Time
Extermination	Product	Location	Facility	Time
Annihilation	Product	Location	Facility	Time
Extinction	Product	Location	Facility	Time
Disappearance	Product	Location	Facility	Time
Vanishing	Product	Location	Facility	Time
Evaporation	Product	Location	Facility	Time
Sublimation	Product	Location	Facility	Time
Condensation	Product	Location	Facility	Time
Freezing	Product	Location	Facility	Time
Melting	Product	Location	Facility	Time
Boiling	Product	Location	Facility	Time
Evaporation	Product	Location	Facility	Time
Distillation	Product	Location	Facility	Time
Filtration	Product	Location	Facility	Time
Sedimentation	Product	Location	Facility	Time
Crystallization	Product	Location	Facility	Time
Precipitation	Product	Location	Facility	Time
Adsorption	Product	Location	Facility	Time
Absorption	Product	Location	Facility	Time
Desorption	Product	Location	Facility	Time
Permeation	Product	Location	Facility	Time
Diffusion	Product	Location	Facility	Time
Osmosis	Product	Location	Facility	Time
Evaporation	Product	Location	Facility	Time
Condensation	Product	Location	Facility	Time
Freezing	Product	Location	Facility	Time
Melting	Product	Location	Facility	Time
Boiling	Product	Location	Facility	Time
Evaporation	Product	Location	Facility	Time
Distillation	Product	Location	Facility	Time
Filtration	Product	Location	Facility	Time
Sedimentation	Product	Location	Facility	Time
Crystallization	Product	Location	Facility	Time
Precipitation	Product	Location	Facility	Time
Adsorption	Product	Location	Facility	Time
Absorption	Product	Location	Facility	Time
Desorption	Product	Location	Facility	Time
Permeation	Product	Location	Facility	Time
Diffusion	Product	Location	Facility	Time
Osmosis	Product	Location	Facility	Time



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

SHINRA

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



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

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, JSON]
 - Cirrus Search Dump [JSON]
 - Extend Named Entity Definition (English/Japanese) [JSON]
- ⌘ The time stamp of All Wikipedia related data is January 20, 2019



SHINRA

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

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" : "خاندهان 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 } }
```




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

SHINRA

Schedule

2019

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

2020

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



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

SHINRA

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