Resource of Wikipedias in 31 Languages Categorized into Fine-Grained Named Entities

Satoshi Sekine[*], Kouta Nakayama[*], Masako Nomoto[*]
Maya Ando[+], Asuka Sumida[*], Koji Matsuda[*]
[*]RIKEN, AIP, [+] free
1-4-1 Nihonbashi,Chuo-ku, Tokyo, 15th floor 103-0027, Japan
satoshi.sekine@riken.jp

Abstract

This paper describes a resource of Wikipedias in 31 languages categorized into Extended Named Entity (ENE), which has 219 fine-grained NE categories. We first categorized 920K Japanese Wikipedia pages according to the ENE scheme using machine learning, followed by manual validation. We then organized a shared task of Wikipedia categorization into 30 languages. The training data were provided by Japanese categorization and the language links, and the task was to categorize the Wikipedia pages into 30 languages, with no language links from Japanese Wikipedia (20M pages in total). Thirteen groups with 24 systems participated in the 2020 and 2021 tasks, sharing their outputs for resource-building. The Japanese categorization accuracy was 98.5%, and the best performance among the 30 languages ranges from 80 to 93 in F-measure. Using ensemble learning, we created outputs with an average F-measure of 86.8, which is 1.7 better than the best single systems. The total size of the resource is 32.5M pages, including the training data. We call this resource creation scheme "Resource by Collaborative Contribution (RbCC)". We also constructed structuring tasks (attribute extraction and link prediction) using RbCC under our ongoing project, "SHINRA".

1 Introduction

Wikipedia consists of a large volume of entities that are significant resources for the knowledge base (KB) used in many Natural Language Processing (NLP) applications, including Question Answering, Information Extraction, and so on. To maximize the use of such a KB, the information in Wikipedia has to be categorized and structured in a consistent manner for machines to perform inference, reasoning, and many other purposes. The current categorization and structure of Wikipedia and other KB derived from it, such as DBpedia, YAGO, and Wikidata, are extremely noisy for NLP applications. This noise is inherent to Wikipedia categories and structures because they are created by multiple independent crowd workers using a bottom-up approach. There are no consistent rules exists for most parts of Wikipedia. As a result, instead of the cumbersome Wikipedia categories and structures, we must rely on a well-defined ontology. Extended Named Entity (ENE) (ENE homepage) is one such ontology for named entities (NEs). ENE version 8 has 219 hierarchical categories, and a set of attributes is defined for each category. Our final goal is to transform Wikipedia information into the structure of ENE so that machines can use the rich information in Wikipedia.

This study reports a resource of Wikipedias in 31 languages categorized into fine-grained named entity (ENE) categories. Categorization is the first task in creating a structured KB. After developing the techniques of automatic categorization are developed, we can structure the contents of Wikipedia pages in each category. We worked on 31 languages, rather than a few major languages. We created resources for practical NLP applications in various languages, instead of platforming a few major languages.

2 Extended Named Entity

To construct a useful KB for NLP applications, a well-structured ontology is essential and must be designed in a top-down manner. The structures of the KBs in DBpedia, Freebase, and Wikidata were created by crowds in a bottom-up manner, all with similar characteristics, including inconsistent categories, imbalanced ontologies, and ad hoc attributes. This is because of the bottom-up nature of their KB designs. We need a top-down strategy to consistently design the ontology and attributes.

Proceedings of the 29th International Conference on Computational Linguistics, pages 3769–3777
October 12–17, 2022.

For a top-down designed ontology for named entities, we employed the Extended Named Entity, ENE (ENE-Homepage). ENE is a cleanly-designed named entity classification hierarchy that includes the attribute definition for each category (Sekine et al., 2002; Sekine and Nobata, 2004; Sekine, 2008). It includes 219 fine-grained categories of named entities in a hierarchy of up to four layers.

It contains not only the fine-grained categories of the typical NE categories, such as "city" and "lake" for "location", and "company" and "political party" for "organization", but also new named entity types such as "products", "event", and "natural object". These categories cover various entities that are often mentioned in encyclopedias and many other resources. Figure 1 shows ENE version 8.0. The category "Concept" is used for anything that doesn't fit into ENE categories; typically common nouns. "IGNORE" is for Wikipedia-specific titles such as "redirect", "disambiguation", and "meta information". Attributes were also designed for each category based on the investigation of the sample entities. For example, the attributes for the "airport" category include the following: "Reading", "IATA code", "ICAO code", "nickname", "name origin", "number of users per year", "number of runaways", and so on. Please refer to the ENE homepage for the complete definitions.

3 Categorized Japanese Wikipedia

We categorized 920K Japanese Wikipedia pages into one or more of 219 ENE categories. For the categorization process, we excluded less popular entities, that is, those having fewer than five incoming links (approximately 151K entities) and nonentity pages (approximately 53K pages), such as common nouns and simple numbers (CONCEPT), and Wikipeida's meta-information pages and forward pages (IGNORED). This categorization was done using the machine learning method with a training data manually created (Suzuki et al., 2018) followed by manual validation. The accuracy of the categorization was confirmed as 98.5% by senior annotators on 1,000 sample data. The remaining 1.5% covered ambiguous and difficult ones, even for human annotators. Table 1 shows the number of entities in each category on Japanese Wikipedia pages. Note that a Wikipedia page can have more than one category. For example, a novel adapted into a movie can have both categories if mentioned in the major paragraphs on the page. The detailed definition can be found on the SHINRA homepage (SHINRA HP).

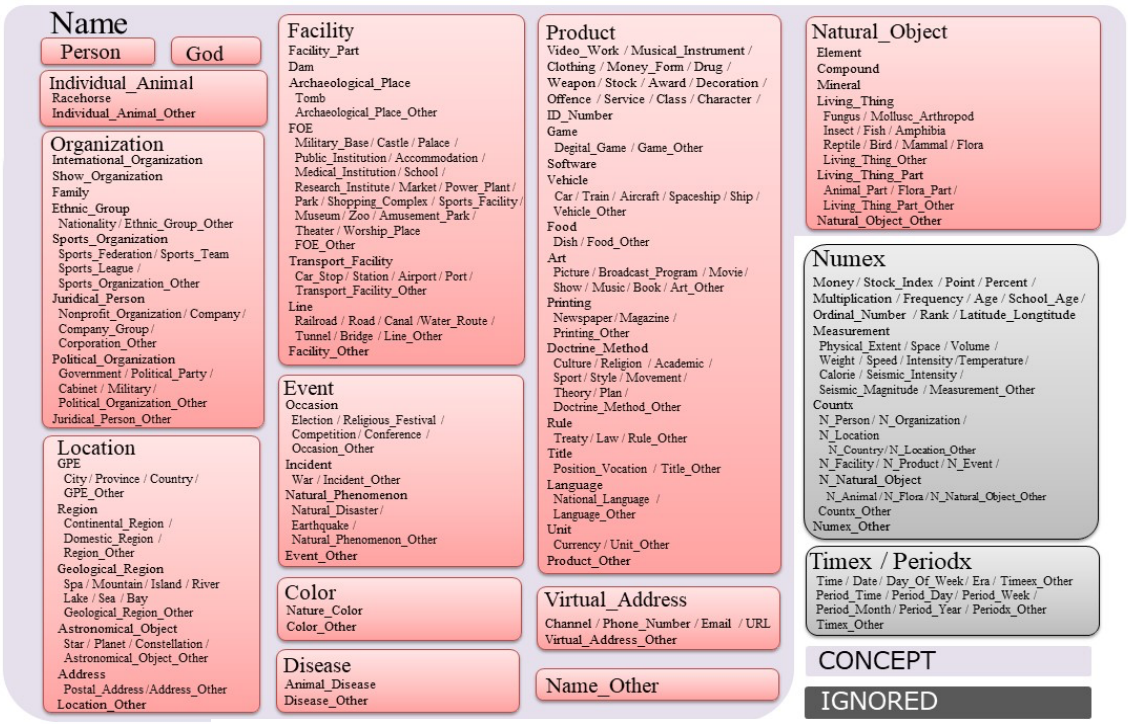

Figure 1. Extended named entity hierarchy version 8

4 Shared Tasks

We conducted Wikipedia page categorization in 30 languages shared-tasks in 2020 and 2021. We refer to them as SHINRA2020-ML and SHINRA2021-ML, respectively. We tackled the problem of categorizing Wikipedia entities in 30 languages in fine-grained categories of 219 categories, as defined in ENE (ver. 8.0). We selected 30 languages based on the number of active users from Wikipedia statistics. The languages are listed in Table 2. The Wikipedias are based on dumps in January 2019. The participants can select one or more target languages, and for each language, their automatic categorization system must categorize all Wikipedia pages in the target language(s), as the evaluation data are hidden. We provided training data for 30 languages, created by (i) the categorized Japanese Wikipedia of 920 K pages, and (ii) Wikipedia language links for 30 languages from the Japanese Wikipedia. For example, out of 2,263K German Wikipedia pages, 275K pages had language links from Japanese Wikipedia, which served as silver (i.e., a bit noisy) training data for German. Thus, the task is "to categorize the remaining 1,988K pages into 219 categories, based on the training data" (the participants are requested to categorize the training data with their system as well, for the purpose of ensemble learning). The same holds for the other 29 languages, as shown in Table 2.

4.1 Participants

Ten groups from seven countries participated in SHINRA2020-ML and three groups participated in SHINRA2021-ML. The list of participant groups and tasks in which they participated are listed in Table 3. Most of the systems in SHINA2020-ML have been described at the NTCIR-15 conference (Bui and Le-Hong, 2020; Cardoso et al., 2020; Abhishek et al., 2020; Nakayama and Sekine, 2020; Yoshioka and Koitabashi, 2020; Nishikawa and Yamada, 2020; Yoshikawa et al., 2020). The reports on SHINRA2021-ML are on the SHINRA homepage (SHINRA HP).

4.2 Evaluation Results

For each target language, a group can submit up to three runs using different methods. The system categorizes each page into one or more ENE (ver. 8.0) categories. The evaluation pages were selected from those without a link to the Japanese Wikipedia pages. The evaluation data are not disclosed because we want the systems to categorize all pages, and we can compare future results using the same data. If an estimated category is not an exact match, the system receives no score for the output. We evaluated the performance of the systems on a multi-label categorization using the micro average F1 measure, that is, the harmonic mean of precision and micro-averaged recall. Note that the distribution of the category in the test data may have differed from that of the training data because of the different characteristics of the links from the Japanese Wikipedia.

Table 4 shows statistical results. It includes 1) the F-measure of the best participating systems for each language, 2) the F-measure of the ensemble learning system., and 3) the upper bound recall by the system outputs, i.e., the percentage of entities to which the correct answer is proposed by at least one of the systems. For most languages, the ensemble system outperforms the best single system in the language (in red). The upper bound exceeds 90% in almost all languages, with an average of 96.75%.

5 Resource

This section describes the ensemble learning method, presents the sample data, and describes some statistics of the resource.

The ensemble learning method we employ is a simple voting method with minor adjustments because a participant can submit up to three systems. The systems submitted by the same participants were extremely similar; we assigned weights to the outputs, that is, one over the number of systems submitted by the participants. Then, the category with the most votes was taken as the ensemble system output.

The sample data for the resources are shown in Figure 2. "Page id" and "title" are information on the Wikipedia page. "ENE" constitutes the categorized Extended Named Entity information, which includes both the ID and name of ENE in JSON format.

The total number of entities we categorized is 32.5M pages in 30 languages. Table 6 shows the number of entities in each category for 30 languages (other than Japanese). In the table, the categories are specified at Level 2 for the categories of "Names" and Level 1 for the other categories. The two biggest categories are "Location" (24.33%) and "Person" (23.00%), followed by "Natural object" (13.85%) and "Product" (13.27%),

"Facility" (5.93%) and "Organization" (5.51%). The distribution is similar to that in the Japanese Wikipedia, which was manually validated, as shown in Table 1. We categorized a total of 920K Japanese Wikipedia entities.

Table 4 shows the average F-measure obtained using the ensemble learning system. The F-measure for the resources proposed is 86.76. We attempted to identify the error types in Table 7, which shows the confusion matrix of the system outputs and the correct answer. Most of the errors are related to the "Concept" and "IGNORED" categories (85% of the total errors). The systems often output the "Concept" even if the correct category is one of the named entities or "IGNORED". In addition, for the entities whose correct category is "IGNORED", the systems output some other categories. In addition to errors related to these two categories, the systems seem to confuse "Organization", "Location", and "Facility". One explanation is that facilities (such as schools, parks, airports, and roads) often have the properties of organization and location. Thus, the systems are robust except when identifying the "Concept" and "IGNORED" categories.

6 Related Work

Ontologies and structured KBs have been considered among the most important knowledge resources in NLP. Previously, several major projects intended to construct KBs have been undertaken. Cyc was one of the earliest projects, followed by more recent Wikipedia-based projects, such as DBpedia, YAGO, Freebase, and Wikidata. Moreover, there are shared tasks aimed at building techniques for knowledge base structuring, such as KBP and CoNLL. Here, we introduce these resources and projects and describe the points considered as issues to be solved in these projects.

Cyc ontology is a large KB constructed of commonsense knowledge (Lenat, 1995). It was one of the largest AI projects between 1980 and 1990, which mainly used human labor to construct a KB. Constructing and maintaining handmade KBs for the general domain is extremely costly. The KB suffered from a "knowledge acquisition bottleneck", which included coverage and consistency problems.

DBpedia is a more recent project that constructs structured information from semi-structured data on Wikipedia, such as infoboxes and categories (Lehmann et al., 2015). However, DBpedia is challenged by inaccuracy, low coverage, and lack of coherence. Like Cyc, infobox and categories in Wikipedia are also created by humans; however, they are non-experts in the ontology. The categories are extremely noisy. For example, in the Japanese Wikipedia, "Shinjuku Station which is a railway station, has a category, "Odakyu Electric Railway", which is a major railway company using the station. However, a station is not an instance of a railway company; therefore, this is not an appropriate category. This type of category definition (i.e., a topic rather than a hypernym relation) is allowed in Wikipedia. There are multiple instances of this in DBpedia, which heavily relies on Wikipedia. Additionally, several inconsistencies were observed in the category structure. It is not even a hierarchy and there is a loop in the category structure.

YAGO is an ontology constructed by mapping Wikipedia articles to WordNet synsets (Mahdisoltani et al., 2015). Similar to DBpedia, YAGO adopts attribute information extracted from infoboxes because no attribute is defined in the WordNet synsets.

Freebase is a project that constructs a structured knowledge base using crowdsourcing, similar to Wikipedia (Bollacker et al., 2008). However, when using the crowdsourcing approach, Freebase lacks a well-organized ontology. The resource is noisy and lacks coherence because it is created using unorganized crowds. The Freebase project was terminated and integrated into Wikidata.

Wikidata aims to be a structured knowledge base based on a crowdsourcing scheme (Vrandečić and Krötzsch 2014). It has noise and lacks coherence because it was constructed using a bottom-up approach, similar to Wikipedia.

Fine Grained Entity Recognition (FIGER) is a project that identifies 112 finely defined named entity classes, similar to ENE (Ling and Weld, 2012). The category in FIGER is biased and does not have attribute definitions for each category.

7 Conclusion

We reported a resource of Wikipedias in 31 languages categorized as ENEs. Japanese Wikipedia pages (920K pages) were categorized with an accuracy of 98.5% and 30 languages Wikipedia pages (32.5M pages) categorized with 87 F-measure. It was created through a shared-task

of Wikipedia categorization in 30 languages. We call this resource creation scheme RbCC". We have also been conducting structuring tasks (attribute extraction and link prediction) using RbCC under our ongoing project, "SHINRA".

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number JP20269633.

0:Concept	51039	1.6.4.4:Public_Institution	5331	1.7.19.3:Movie	19381	1.10.5.1:Animal_Part	1944
1.0:Name_Other	5	1.6.4.5:Accommodation	892	1.7.19.4:Show	3131	1.10.5.2:Flora_Part	201
1.1:Person	269688	1.6.4.6:Medical_Institution	1615	1.7.19.5:Music	46889	1.11.0:Disease_Other	29
1.2:God	1278	1.6.4.7:School	25579	1.7.19.6:Book	21093	1.11.1:Animal_Disease	2229
1.3.0:Individual_Animal_Other	69	1.6.4.8:Research_Institute	1036	1.7.20.0:Printing_Other	753	1.12.0:Color_Other	200
1.3.1:Racehorse	3450	1.6.4.9:Market	107	1.7.20.1:Newspaper	794	1.12.1:Nature_Color	13
1.4.0:Organization_Other	3631	1.6.4.10:Power_Plant	446	1.7.20.2:Magazine	2495	2.0:Timex_Other	1
1.4.1:International_Organization	512	1.6.4.11:Park	2606	1.7.21.0:Doctrine_Method_Other	14767	2.1.0:Timeex_Other	0
1.4.2:Show_Organization	11758	1.6.4.12:Shopping_Complex	3341	1.7.21.1:Culture	303	2.1.1:Time	12
1.4.3:Family	2075	1.6.4.13:Sports_Facility	3783	1.7.21.2:Religion	936	2.1.2:Date	3964
1.4.4.0:Ethnic_Group_Other	1144	1.6.4.14:Museum	2901	1.7.21.3:Academic	1800	2.1.3:Day_Of_Week	10
1.4.4.1:Nationality	227	1.6.4.15:Zoo	445	1.7.21.4:Sport	955	2.1.4:Era	2067
1.4.5.0:Sports_Organization_Other	89	1.6.4.16:Amusement_Park	450	1.7.21.5:Style	586	2.2.0:Periodx_Other	5
1.4.5.1:Sports_Federation	865	1.6.4.17:Theater	1614	1.7.21.6:Movement	474	2.2.1:Period_Time	0
1.4.5.2:Sports_League	846	1.6.4.18:Worship_Place	9276	1.7.21.7:Theory	1808	2.2.2:Period_Day	1
1.4.5.3:Sports_Team	9107	1.6.5.0:Transport_Facility_Other	946	1.7.21.8:Plan	1818	2.2.3:Period_Week	1
1.4.6.0:Juridical_Person_Other	3	1.6.5.1:Car_Stop	5605	1.7.22.0:Rule_Other	327	2.2.4:Period_Month	1
1.4.6.1:Nonprofit_Organization	5122	1.6.5.2:Station	18296	1.7.22.1:Treaty	915	2.2.5:Period_Year	25
1.4.6.2:Company	30120	1.6.5.3:Airport	1559	1.7.22.2:Law	2289	3.0:Numex_Other	0
1.4.6.3:Company_Group	386	1.6.5.4:Port	575	1.7.23.0:Title_Other	16	3.1:Money	0
1.4.7.0:Political_Organization_Other	1256	1.6.6.0:Line_Other	1528	1.7.23.1:Position_Vocation	6317	3.2:Stock_Index	0
1.4.7.1:Government	3516	1.6.6.1:Railroad	3615	1.7.24.0:Language_Other	1544	3.3:Point	0
1.4.7.2:Political_Party	1719	1.6.6.2:Road	15550	1.7.24.1:National_Language	251	3.4:Percent	7
1.4.7.3:Cabinet	225	1.6.6.3:Canal	333	1.7.25.0:Unit_Other	496	3.5:Multiplication	0
1.4.7.4:Military	3744	1.6.6.4:Water_Route	455	1.7.25.1:Currency	319	3.6:Frequency	0
1.5.0:Location_Other	2671	1.6.6.5:Tunnel	379	1.8.0:Virtual_Address_Other	291	3.7:Age	5
1.5.1.0:GPE_Other	435	1.6.6.6:Bridge	1887	1.8.1:Channel	3348	3.8:School_Age	7
1.5.1.1:City	49028	1.7.0:Product_Other	14619	1.8.2:Phone_Number	4	3.9:Ordinal_Number	0
1.5.1.2:Province	12490	1.7.1:Video_Work	2621	1.8.3:Email	0	3.10:Rank	1
1.5.1.3:Country	1407	1.7.2:Musical_Instrument	684	1.8.4:URL	2	3.11:Latitude_Longitude	304
1.5.2.0:Region_Other	30	1.7.3:Clothing	936	1.9.0:Event_Other	1991	3.12.0:Measurement_Other	3
1.5.2.1:Continental_Region	275	1.7.4:Money_Form	273	1.9.1.0:Occasion_Other	2697	3.12.1:Physical_Extent	5
1.5.2.2:Domestic_Region	2219	1.7.5:Drug	689	1.9.1.1:Election	908	3.12.2:Space	1
1.5.3.0:Geological_Region_Other	2446	1.7.6:Weapon	11167	1.9.1.2:Religious_Festival	983	3.12.3:Volume	0
1.5.3.1:Spa	1132	1.7.7:Stock	0	1.9.1.3:Competition	17471	3.12.4:Weight	0
1.5.3.2:Mountain	4013	1.7.8:Award	3031	1.9.1.4:Conference	585	3.12.5:Speed	1
1.5.3.3:Island	2517	1.7.9:Decoration	247	1.9.2.0:Incident_Other	3465	3.12.6:Intensity	0
1.5.3.4:River	2900	1.7.10:Offense	176	1.9.2.1:War	3093	3.12.7:Temperature	0
1.5.3.5:Lake	858	1.7.11:Service	3	1.9.3.0:Natural_Phenomenon_Other	249	3.12.8:Calorie	0
1.5.3.6:Sea	302	1.7.12:Class	45	1.9.3.1:Natural_Disaster	355	3.12.9:Seismic_Intensity	1
1.5.3.7:Bay	362	1.7.13:Character	6354	1.9.3.2:Earthquake	355	3.12.10:Seismic_Magnitude	0
1.5.4.0:Astronomical_Object_Other	1516	1.7.14:ID_Number	5	1.10.0:Natural_Object_Other	1493	3.13.0:Countx_Other	4
1.5.4.1:Star	1130	1.7.15.0:Game_Other	350	1.10.1:Element	153	3.13.1:N_Person	5
1.5.4.2:Planet	3186	1.7.15.1:Degital_Game	11213	1.10.2:Compound	4493	3.13.2:N_Organization	2
1.5.4.3:Constellation	167	1.7.16:Software	4658	1.10.3:Mineral	556	3.13.3.0:N_Location_Other	0
1.5.5.0:Address_Other	0	1.7.17.0:Vehicle_Other	771	1.10.4.0:Living_Thing_Other	1410	3.13.3.1:N_Country	0
1.5.5.1:Postal_Address	1	1.7.17.1:Car	5187	1.10.4.1:Fungus	275	3.13.4:N_Facility	1
1.6.0:Facility_Other	5035	1.7.17.2:Train	4632	1.10.4.2:Mollusk_Arthropod	577	3.13.5:N_Product	0
1.6.1:Facility_Part	67	1.7.17.3:Aircraft	2619	1.10.4.3:Insect	858	3.13.6:N_Event	0
1.6.2:Dam	741	1.7.17.4:Spaceship	1531	1.10.4.4:Fish	966	3.13.7.0:N_Natural_Object_Other	0
1.6.3.0:Archaeological_Place_Other	2968	1.7.17.5:Ship	7887	1.10.4.5:Amphibia	132	3.13.7.1:N_Animal	0
1.6.3.1:Tomb	1093	1.7.18.0:Food_Other	2725	1.10.4.6:Reptile	1012	3.13.7.2:N_Flora	0
1.6.4.0:FOE_Other	2465	1.7.18.1:Dish	2888	1.10.4.7:Bird	1901	9:IGNORED	13360
1.6.4.1:Military_Base	462	1.7.19.0:Art_Other	945	1.10.4.8:Mammal	1868		
1.6.4.2:Castle	2014	1.7.19.1:Painting	408	1.10.4.9:Flora	4164		
1.6.4.3:Palace	237	1.7.19.2:Broadcast_Program	33747	1.10.5.0:Living_Thing_Part_Other	309		

Table 1. The number of entities for each category in the Japanese Wikipedia

{"pageid": "59706565", "title": "1978 Giro d'Italia, Prologue to Stage 10", "ENEs": [{"ENE_id": "1.9.1.3", "ENE_name": "Competition"}]}
{"pageid": "22059861", "title": "Tarlach Rua Mac Dónaill", "ENEs": [{"ENE_id": "1.1", "ENE_name": "Person"}]}
{"pageid": "53177250", "title": "90th Scripps National Spelling Bee", "ENEs": [{"ENE_id": "1.9.1.3", "ENE_name": "Competition"}]}
{"pageid": "13820024", "title": "The Early History of God", "ENEs": [{"ENE_id": "1.7.19.6", "ENE_name": "Book"}]}
{"pageid": "17870536", "title": "Pure type system", "ENEs": [{"ENE_id": "1.7.21.0", "ENE_name": "Doctrine_Method_Other"}]}
{"pageid": "13918760", "title": "Nyandeni Local Municipality", "ENEs": [{"ENE_id": "1.5.1.1", "ENE_name": "City"}]}
{"pageid": "14874071", "title": "BICD2", "ENEs": [{"ENE_id": "1.10.5.0", "ENE_name": "Living_Thing_Part_Other"}]}
{"pageid": "40760418", "title": "Meydan-e Sofla", "ENEs": [{"ENE_id": "1.5.1.1", "ENE_name": "City"}]}
{"pageid": "31722196", "title": "Princes' Concordat", "ENEs": [{"ENE_id": "1.7.22.1", "ENE_name": "Treaty"}]}
{"pageid": "5724950", "title": "Hatley, Quebec (township)", "ENEs": [{"ENE_id": "1.5.1.1", "ENE_name": "City"}]}

Figure 2. Sample data

Language	num. of pages	Links from ja	Ratio
English (en)	5,790,377	439,354	7.6
Spanish (es)	1,500,013	257,835	17.2
French (fr)	2,074,648	318,828	15.4
German (de)	2,262,582	274,732	12.1
Chinese (zh)	1,041,039	267,107	25.7
Russian (ru)	1,523,013	253,012	16.6
Portuguese (pt)	1,014,832	217,896	21.5
Italian (it)	1,496,975	270,295	18.1
Arabic (ar)	661,205	73,054	11.0
Japanese	1,136,222	–	–
Indonesian (id)	451,336	115,643	25.6
Turkish (tr)	321,937	111,592	34.7
Dutch (nl)	1,955,483	199,983	10.2
Polish (pl)	1,316,130	225,552	17.1
Persian (fa)	660,487	169,053	25.6
Swedish (sv)	3,759,167	180,948	4.8
Vietnamese (vi)	1,200,157	116,280	9.7
Korean (ko)	439,577	190,807	43.7
Hebrew (he)	236,984	103,137	43.5
Romanian (ro)	391,231	92,002	23.5
Norwegian (no)	501,475	135,935	27.1
Czech (cs)	420,195	135,935	25.1
Ukrainian (uk)	881,572	181,122	20.5
Hindi (hi)	129,141	30,547	23.6
Finnish (fi)	450,537	144,750	32.1
Hungarian (hu)	443,060	120,295	27.2
Danish (da)	242,523	91,811	35.6
Thai (th)	129,294	59,791	46.2
Catalan (ca)	601,473	139,032	23.1
Greek (el)	157,566	60,513	38.4
Bulgarian (bg)	248,913	89,017	35.7

Table 2. Wikipedia statistics in 31 languages

Year	Group ID	Country	Participated Language(s)
2020	CMVS	Finland	1 (ar)
	FPTAI	Vietnam	30 (all)
	HUKB	Japan	30 (all)
	PribL	Portugal	15 (ar, cs, de, en, es, fr, it, ko, nl, no, pl, pt, ru, tr, zh)
	RH312	India	6 (bg, fr, hi, id, th, tr)
	TKUIM	Taiwan	30 (all)
	Ousia	Japan	9 (ar, de, es, fr, hi, it, pt, th, zh)
	Uomfj	Australia/Japan	28 (all except el, sv)
	Vlp	Vietnam	1 (vi)
	LIAT	Japan	30 (all)
2021	HUKB	Japan	30 (all)
	KANJU	Japan	30 (all)
	junps	Japan	1 (en)

Table 3. SHINRA-ML task participants

Language	Best system F	Ensemble system F	Upper bound Recall
Arabic	90.06	92.18	97.71
Bulgarian	86.94	88.32	92.77
Catalan	89.25	86.62	95.42
Czech	81.70	83.72	94.34
Danish	82.34	81.53	92.15
German	79.68	80.93	89.68
Greek	84.85	79.72	90.04
English	86.49	87.65	93.49
Spanish	85.54	86.51	94.69
Persian	89.63	90.87	94.63
Finish	85.95	86.36	95.47
French	83.77	87.20	92.83
Hebrew	81.80	81.74	90.74
Hindi	87.81	90.76	94.79
Hungarian	89.93	91.41	96.19
Indonesian	90.71	92.22	97.28
Italian	84.08	85.51	91.85
Korean	80.08	82.57	90.68
Dutch	85.88	85.88	91.51
Norwegian	85.89	86.10	93.53
Polish	84.44	85.06	94.00
Portuguese	88.96	89.83	96.11
Romanian	93.43	93.43	98.07
Russian	81.62	84.06	90.91
Swedish	84.28	86.28	91.85
Thai	84.72	85.48	95.31
Turkish	87.85	88.13	94.18
Ukrainian	85.14	84.53	91.20
Vietnamese	90.28	89.58	95.14
Chinese	88.00	88.72	95.10
Average	**86.04**	**86.76**	**96.75**

Table 4. Performance of the system and other statistics

Name	Task	Language	Target categories
SHINRA2018	Attribute value extraction	Japanese	5 categories
SHINRA2019	Attribute value extraction	Japanese	35 categories
SHINRA2020-JP	Attribute value extraction	Japanese	78 categories
SHINRA2020-ML	Categorization	30-language	
SHINRA2021-LinkJP	Link Prediction	Japanese	7 categories
SHINRA2021-ML	Categorization	30-language	
SHINRA2022	All three tasks	Japanese	all categories

Table 5. SHINRA tasks

ID	0	1	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	1.1	1.11	1.12	2	3	9
Category	Concept	Name Other	Person	God	Individual Animal	Organization	Location	Facility	Product	Virtual Address	Event	Natural Object	Disease	Color	Timex	Numex	IGNORED
Arabic	52972	0	204207	513	3	26264	198716	21437	66559	2093	20277	38404	4937	97	7733	480	16849
Bulgarian	17533	0	73267	588	7	13016	53001	11304	32931	746	6704	31913	679	20	3651	51	4358
Catalan	51271	0	149258	946	7	30121	155361	41240	67885	1316	20958	64479	1652	84	4156	540	12471
Czech	50169	0	116294	796	34	28476	74478	32128	68588	1515	18266	17454	1396	37	4982	1	15654
Danish	27238	1	73171	481	12	19399	33795	16822	45794	989	8853	7121	758	31	3764	786	3577
German	196031	0	768225	2837	279	192658	396296	192032	270062	4767	82408	64601	6401	84	10350	0	76738
Greek	74786	0	67372	1342	2	28297	48940	5861	53778	1339	13157	5978	1070	85	8462	1	25595
English	315531	0	1728797	4423	5822	432236	875918	472005	1005969	43520	279123	403739	11813	334	8829	811	204327
Spanish	105543	2	376710	2265	89	90700	317901	92715	246103	4363	72964	153631	4378	187	8479	226	38106
Persian	48311	0	138148	763	16	19230	241298	46761	100016	1354	9146	36456	2171	79	8104	711	8759
Finish	42147	0	147556	784	48	34477	48877	21505	99635	1400	15659	23084	1599	31	4022	1	10071
French	129098	0	595708	2393	573	135080	416779	149286	355881	6334	87986	119710	4722	254	11182	630	60099
Hebrew	31475	1	78479	484	15	16658	20233	10751	47355	859	8045	8761	1467	39	5501	0	7122
Hindi	14388	0	22242	406	2	5113	46309	6144	20433	848	3150	3508	643	47	5728	522	2693
Hungarian	30551	1	110023	650	12	20570	137258	25902	60441	1266	18818	22436	688	18	4670	0	9952
Indonesian	26194	0	72630	669	22	21450	115722	20662	69492	1818	11991	98328	1050	32	3734	190	9556
Italian	98522	0	378736	1962	93	92536	313896	79661	325760	2873	109402	48390	4017	247	9276	7	32724
Korean	52713	0	111535	761	26	32660	46817	54853	88858	2155	15243	14963	1325	70	6036	381	11935
Dutch	120574	0	217528	1349	30	68301	342682	88123	148271	2128	47810	874290	2611	68	5067	88	36880
Norwegian	43235	0	158039	696	16	39756	73512	45573	71868	1645	25076	23477	1084	33	3988	21	13680
Polish	84346	0	354727	1805	55	79672	386271	93985	173646	2469	52842	51219	4422	49	4921	3	26274
Portuguese	80353	1	224466	1520	52	65736	240881	43866	181481	4039	48010	90099	3473	94	5829	536	24859
Romanian	22821	0	57913	582	6	13518	201132	11420	33084	1020	5905	30486	657	20	3640	1	9209
Russian	102057	1	463956	2005	57	98493	409426	76356	207009	2795	57119	53424	3288	96	7241	388	40425
Swedish	206567	0	231159	1301	288	53538	1729196	88261	119368	1459	23894	1281506	2160	74	4332	4	16006
Thai	14726	0	28041	378	2	11195	10381	8738	25335	905	6139	6944	833	32	3843	615	11570
Turkish	26188	0	83506	859	112	18642	83704	12515	59418	1755	15271	7721	1149	88	5139	237	9344
Ukranian	67899	0	178244	1381	8	39675	348825	45919	101043	1759	21852	47983	1734	67	5596	2	20769
Vietnamese	46134	0	55154	430	20	9533	245542	10573	38259	882	8636	773560	966	56	4495	5	6613
Chinese	61095	0	217583	1633	267	56013	301227	103105	133350	3371	28580	100343	1985	114	7484	249	25499
Total	2179373	7	7265091	35369	7708	1737000	7613147	1826398	4184322	100411	1114704	4403665	73143	2453	172750	7238	766215
%	6.89%	0.00%	23.00%	0.11%	0.02%	5.51%	24.33%	5.93%	13.27%	0.32%	3.51%	13.85%	0.23%	0.01%	0.55%	0.02%	2.43%

Table 6. Number of entities in each category and language (for languages other than Japanese)

Note that the statistics are based on second level categories. All sub-categories are summed together for second level categories. For example, the number of entities under Organization (1.4), such as international organizations (1.4.1), companies (1.4.6.2), political parties (1.4.7.2), and other categories under 1.4, are combined in the statistics for Organization (1.4). The total number of entities for each language might not be equal to the number of entities in Table 2 because there could be over one category for each entity.

ID	0	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	1.1	1.11	2	3	9
Category	Concept	Person	God	Individual Animal	Organization	Location	Facility	Product	Virtual Address	Event	Natural Object	Disease	Timex	Numex	IGNORED
0:Concept	587	-	-	-	1	2	-	24	-	2	65	1	-	-	1
1.1:Person	9	3675	-	-	4	-	-	5	-	-	-	-	-	-	3
1.2:God	-	-	14	-	-	-	-	-	-	-	-	-	-	-	-
1.3:Individual_Animal	-	1	-	-	-	-	-	1	-	-	1	-	-	-	-
1.4:Organization	15	6	-	-	760	20	12	5	-	3	-	-	-	-	3
1.5:Location	11	-	-	-	4	3371	20	1	-	-	1	-	-	-	2
1.6:Facility	12	3	-	-	11	26	906	6	-	1	1	-	-	-	3
1.7:Product	103	12	1	-	13	1	10	1824	-	9	3	-	1	-	4
1.8:Virtual_Address	1	-	-	-	5	-	-	1	28	-	-	-	-	-	-
1.9:Event	5	1	-	-	7	-	1	11	-	416	-	-	2	-	8
1.10:Natural_Object	2	-	-	-	-	-	-	1	-	-	91	-	-	-	-
1.11:Disease	-	-	-	-	-	-	-	-	-	-	-	23	-	-	-
2:Timex	-	-	-	-	-	-	-	1	-	3	-	-	46	-	-
3:Numex	-	-	-	-	-	-	-	-	-	-	-	-	-	1	1
9:IGNORED	287	171	2	-	73	181	33	138	1	45	9	-	18	2	398

Table 7. Confusion matrix of system outputs (horizontal) and correct answers (vertical) for all languages

References

Tushar Abhishek, Ayush Agarwal, Anubhav Sharma, Vasudeva Varma, and Manish Gupta (2020). Rehoboam at the NTCIR-15 SHINRA2020-ML task. In The 15th NTCIR Conference Evaluation of Information Access Technologies (NTCIR-15).

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. (2008) Freebase: a collaboratively created graph database for structuring human knowledge. In SIGMOD'08: Proceedings of the 2008 ACM SIGMOD international conference on Management of data, pages 1247-1250.

The Viet Bui and Phuong Le-Hong (2020). Cross-lingual Extended Named Entity Classification of Wikipedia Articles. In The 15th NTCIR Conference Evaluation of Information Access Technologies (NTCIR-15).

Ruben Cardoso, Afonso Mendes, and Andre Lamurias (2020). Priberam Labs at the NTCIR-15 SHINRA2020-ML: Classification Task. In The 15th NTCIR Conference Evaluation of Information Access Technologies (NTCIR-15).

ENE-HP. Extended Named Entity homepage. In https://ene-project.info.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Soren Auer, and Christian Bizer (2015). DBpedia - A large-scale, multilingual knowledge base extracted from Wikipedia. Semantic Web, 6(2):167-195.

Douglas Lenat (1995). Cyc: A large-scale investment in knowledge infrastructure. Communications of the ACM, 38(11):33-38.

Xiao Ling and Daniel S. Weld (2012). Fine-Grained Entity Recognition. In Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence, AAAI'12, pages 94-100. AAAI Press.

Farzaneh Mahdisoltani, Joanna Biega, and Fabian M Suchanek (2015). YAGO3: A Knowledge Base from Multilingual Wikipedias. CIDR.

David Nadeau, Satoshi Sekine (2007). A survey of Named Entity Recognition and Classification". Lingvisticae Investigationes 30 (1), 3-26.

Kouta Nakayama and Satoshi Sekine (2020). LIAT Team's Wikipedia Classifier at NTCIR-15 SHINRA2020-ML: Classification Task. In The 15th NTCIR Conference Evaluation of Information Access Technologies (NTCIR-15).

Sosuke Nishikawa and Ikuya Yamada (2020). Studio Uusia at the NTCIR-15 SHINRA2020-ML Task. In The 15th NTCIR Conference Evaluation of Information Access Technologies (NTCIR-15).

Satoshi Sekine (2008). Extended named entity ontology with attribute information. In the Sixth International Conference on Language Resource and Evaluation (LREC08).

Satoshi Sekine and Chikashi Nobata (2004). Definition, dictionaries, and tagger for extended named entity hierarchy. In the Fourth International Conference on Language Resources and Evaluation (LREC'04).

Satoshi Sekine, Kiyoshi Sudo, and Chikashi Nobata (2002). Extended named entity hierarchy. In the Third International Conference on Language Resources and Evaluation (LREC'02).

Satoshi Sekine, Akio Kobayashi, and Kouta Nakayama (2019). SHINRA: Structuring wikipedia by collaborative contribution. In Automated Knowledge Base Construction (AKBC).

SHINRA-HP. SHINRA project homepage. URL https://shinra-project.info.

SHINRA2020-ML-HP. SHINRA 2020-ML homepage. URL http://shinra-project.info/shinra2020ml/.

Masatoshi Suzuki, Koji Matsuda, Satoshi Sekine, Naoki Okazaki, and Kentaro Inui. (2018) A joint neural model for fine-grained named entity classification of Wikipedia articles. IEICE Transactions on Information and Systems, E101.D(1):73-81.

U.S. National Institute of Standards and Technology (NIST) . TAC Knowledge Base Population (KBP) 2017, 2018. URL https://tac.nist.gov/2017/KBP/.

Denny Vrandečić and Markus Krötzsch (2014). Wikidata: A free collaborative knowledgebase. Commun. ACM, 57(10):78-85.

Hiyori Yoshikawa, Chunpeng Ma, Aili Shen, Qian Sun, Chenbang Huang, Guillaume Pelat, Akiva Miura, Daniel Beck, Timothy Baldwin, and Tomoya Iwakura (2020). UOM-FJ at the NTCIR-15 SHINRA2020-ML task. In The 15th NTCIR Conference Evaluation of Information Access Technologies (NTCIR-15).

Masaharu Yoshioka and Yoshiaki Koitabashi (2020). HUKB at SHINRA2020-ML task. In The 15th NTCIR Conference Evaluation of Information Access Technologies (NTCIR-15).