Overview of SHINRA2020-ML Task

Satoshi Sekine, Masako Nomoto, Kouta Nakayama, Asuka Sumida, Koji Matsuda, Maya Ando
AIP, RIKEN
Tokyo, Japan
{satoshi.sekine|masako.nomoto|kouta.nakayama|asuka.sumida|koji.matsuda}@riken.jp

ABSTRACT

In this paper, a Knowledge Base (KB) construction project, SHINRA [1] and a new shared-task we conducted under NTCIR15, SHINRA2020-ML [5] are described.

SHINRA is a project to structure Wikipedia based on a pre-defined set of attributes for given categories. The categories and the attributes follow the definition of the Extended Named Entity (ENE) [2]. In our previous work, we conducted a shared task of automatic knowledge base construction (AKBC) [3] and used the submitted results to construct a large and more accurate KB. In the shared tasks, the participants are not notified which is the test data so they must run their systems on all entities in the target Wikipedia dataset. By this method, the organizers receive information for the entities, later made public, and will be used to build structured knowledge by ensemble learning. We call this recourse construction scheme "Resource by Collaborative Contribution (RbCC)".

SHINRA2020-ML, a new shared-task of SHINRA, targets the categorization of Wikipedia entities in 30 languages. We conducted the task and received the submission results from 10 participant groups. We would like to utilize the results and work on making the KB more valuable by the RbCC scheme.

CCS CONCEPTS

Computing Methodologies - Artificial Intelligence - Natural Language Processing, Computing Methodologies - Artificial Intelligence – Knowledge representation and reasoning

KEYWORDS

Knowledge Acquisition, Text Categorization, Resource by Collaborative Contribution

ACM Reference format:
Satoshi Sekine, Masako Nomoto, Kouta Nakayama, Asuka Sumida, Koji Matsuda, and Maya Ando. 2020. Overview of SHINRA2020-ML Task. In *Proceedings of the NTCIR-15 Conference.*

1 Introduction

Wikipedia consists of a large volume of entities (a.k.a. articles), which is a great resource of knowledge to be utilized in many NLP tasks. To maximize the use of such knowledge, resources created from Wikipedia need to be structured for inference, reasoning, or any other purposes in many NLP applications. The current structured knowledge bases, such as DBpedia, Freebase, and Wikidata, among others, are created mostly by bottom-up approach such us bottom-up crowdsourcing, which may cause a significant amount of undesirable noises in the knowledge base.

We believe that the structure of the knowledge should be defined top-down rather than bottom-up to create cleaner and more valuable knowledge bases. Instead of the existing, cumbersome Wikipedia categories, we should rely on a well-defined and fine-grained category definition. Extended Named Entity (ENE) [2] is a well-defined name ontology, which has about 220 hierarchical categories and a set of attributes are defined for each category.

The final goal of SHINRA project [1] is to structure the knowledge in Wikipedia including the attribute, but as a first step, we need to classify each Wikipedia entry into one or more of the ENE categories before extracting attribute values. The task of SHINRA2020-ML [5] is to classify Wikipedia pages in 30 languages into Extended Named Entity (ENE) [2] (ver.8.0) categories. We have classified major Japanese Wikipedia pages (920K pages) into ENE categories already. We can use language links to create the training data in 30 languages (i.e. 275K in German and so on). So, the task is to categorize the remaining pages in 30 languages using the training data.

The goal of this project is not only to compare the participated systems and see which system performs the best, but also to create the knowledge base using the outputs of the participated systems. We can utilize the state of the art “ensemble learning

Language	num. of pages	Links from ja Wikipedia	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 (ja)	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	43.4
Hebrew (he)	236,984	96,434	40.7
Romanian (ro)	391,231	92,002	23.5
Norwegian (no)	501,475	135,935	27.1
Czech (cs)	420,195	125,959	30.0
Ukrainian (uk)	881,572	167,237	19.0
Hindi (hi)	129,141	30,547	23.7
Finnish (fi)	450,537	144,750	32.1
Hungarian (hu)	443,060	120,295	27.2
Danish (da)	242,523	86,238	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	35.8

Table 2: Statistics of Wikipedia in 31 languages

technologies" to gather the fruit of the systems and create the KB as accurate as possible.

2 Base Data

In this section, we will introduce two dataset on which we are relying in SHINRA2020-ML task [5].

One of them is Extended Named Entity definition. This is the knowledge frame for the named entity, and hence, Wikipedia entities to be categorized.

The other one is the categorized Japanese Wikipedia data, which will be used to create the training data for 30 languages through inter-language links.

2.1 Extended Named Entity

In order to construct a structured knowledge base that is useful for NLP applications, we have learned that well-structured ontology is needed and it must be designed in a top-down manner. DBpedia, Freebase, and Wikidata are created by crowds in a bottom-up manner, and they have inconsistent categories, imbalanced ontologies, and adhoc attributes. We believe that the major cause is the fact that these are created in a bottom-up manner by crowds. A top-down strategy is essential to design the ontology and the attributes consistently.

As a top-down designed ontology for named entities, we employed the "Extended Named Entity (ENE) hierarchy" [2] in our project. ENE is a named entity classification hierarchy that includes the attribute definition for each category ([4], [6], [7]).

It includes about 220 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" or "company" for "organization", but also contains new named entity types such as "products", "event", and "position". These categories are designed to cover a large amount of entities in the general world, which are often mentioned in encyclopaedia and many other general resources. Figure 1 shows the ENE definition, version 8.0.

Attributes are designed so that the important attributes in the Wikipedia entities of each category are covered based on the investigation of sample entities. For example, the attributes for "airport" categories are as follows: "Reading", "IATA code", "ICAO code", "nickname", "name origin", "number of users per year", "the year of the statistic", "the number of airplane landings per year", "longitude", "latitude", "location", "old name", "elevation", "big city nearby", "number of runaways" and so on. Please refer to the ENE homepage [2] for the complete definition. (Note that the attribute definition ver. 8 is shown in Japanese and ver. 7 in English and Japanese.)

2.2 Categorized Japanese Wikipedia

We have categorized 920K pages of Japanese Wikipedia into one or more of about 220 ENE categories [2] before this project [8]. Table 1 shows the most frequent categories in the data of 920K pages.

At the categorization process, we have excluded the less popular entities having less than five incoming links (151K entities) and the non-entity pages (about 53K pages) such as common nouns and simple numbers. This categorization was done by the machine learning method followed by hand checking on all data.

We confirmed the accuracy of the categorization to be 98.5%. The remaining 1.5% are those that are ambiguous in nature and are very difficult even for human annotators, or those having multiple categories.

Person	247,983	School	23,609
City	45,306	Literature	18,515
Artefact other	33,453	Movie	17,901
Broadcast Program	32,050	Train station	15,863
Company	26,746	Sports event	15,863

Table 1: Top 10 ENE categories in Japanese Wikipedia

3 Task Description

SHINRA2020-ML [5] is the first shared-task of text categorization in SHINRA project [1], tackling the problem of classifying 30 language Wikipedia entities in fine-grained categories, namely, about 220 categories defined in Extended Named Entity (ENE) [2] (ver. 8.0).

Participants are expected to select one or more target languages, and for each language, run the system to classify all Wikipedia pages in the target language(s). For each target language, a group can submit up to three runs based on different methods. A system is expected to classify each page into one or more of the ENE (ver.8.0) categories correctly.

We provided the training data for 30 languages, created by the categorized Japanese Wikipedia of 920K pages and Wikipedia language links for 30 languages. For example, out of 2,263K German Wikipedia pages, 275K pages have a language link from Japanese Wikipedia, which will potentially serve as a-bit-noisy training data for German. So, the task is "to classify the remaining 1,988K pages into about 220 categories, based on the 275K categorized pages" (actually, the participants are requested to categorize the entire target data, including the training data with their system). The same holds true for other 29 languages shown in Table 2.

4 Schedule

SHINRA2020-ML [5] has been conducted according to the schedule listed in Table 3.

Date	Event
January, 2020	Data release
April, 2020	Homepage & CFP open
August 31, 2020	Registration & Result submission deadline
September 16, 2020	Evaluation results due back to participants
Dec 8-11, 2020	NTCIR-15 Conference

Table 3: SHINRA2020-ML Schedule

5 Participants

There were 10 participant groups from 7 countries. The list of participant groups and the tasks they participated are listed in Table 4.

6 Evaluation

We evaluated the performance of systems on multi-label classification using the micro averaged F1 measure, i.e., the harmonic mean of micro-averaged precision and micro-averaged recall. The test data in each language has 1000 pages chosen from the rest of target data after removing the training data. The distribution of category in the test data may differ from that of target data or training data.

Group ID	Country	Participated Language
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	2 (en, zh)
ousia	Japan	9 (ar,de,es,fr,hi,it,pt,th,zh)
uomfj	Australia/Japan	28 (except for el, sv)
vlp	Vietnam	1 (vi)
LIAT	Japan	30 (all)

Table 4: List of participants

If the estimated category is not an exact match, the system does not get score for that. If no label is predicted for a page, the page is considered to be assigned the label 'IGNORED'(ENE_id:9).

7 Results

Table 5 shows the micro averaged F1 scores of participated systems. "Late submission" indicates those who submit the results after the deadline. Those who had simple mistake in the results submitted before the deadline are allowed to fix the mistake and are regarded as the "regular submission". Some evaluation results are not listed in the table due to misunderstanding or error in the submission.

8 Conclusion

We described the overview of SHINRA2020-ML task [5] conducted under NTCIR-15. Under the circumstance, we are very delighted that we have many participants and we got so many data. We would like to work on making the data valuable by RbCC scheme. We are planning to hold the same task in 2021 so that we can improve even more on the accuracy and coverage of the data. We are hoping to have more participants under the RbCC scheme.

ACKNOWLEDGEMENT

We acknowledge all the developers, helpers, administrative staffs and others who are supporting SHINRA2020-ML task [5] and SHINRA project [1] in general. As described in RbCC scheme, it is the most important issue that we are working together to create the resource. We hope the RbCC scheme will be opening up a new era on the evaluation-based projects.

REFERENCES

[1] SHINRA project homepage: https://shinra-project.info
[2] Extended Named Entity homepage: https://ene-project.info
[3] Satoshi Sekine, Akio Kobayashi, and Kouta Nakayama. 2019. SHINRA: Structuring Wikipedia by Collaborative Contribution. In *Proceedings of the 1st conference on the Automatic Knowledge Base Construction AKBC-2019*.
[4] Satoshi Sekine. 2008. Extended Named Entity Ontology with Attribute Information. In *Proceedings of the Sixth International Conference on Language Resource and Evaluation (LREC08)*.
[5] SHINRA 2020-ML homepage: http://shinra-project.info/shinra2020ml/
[6] Satoshi Sekine, Kiyoshi Sudo, and Chikashi Nobata. 2002. Extended Named Entity Hierarchy. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*,
[7] Satoshi Sekine and Chikashi Nobata. 2004. Definition, dictionaries and tagger for Extended Named Entity Hierarchy, In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*.
[8] Satoshi Sekine, Maya Ando, Akio Kobayashi, and Asuka Sumida. 2020. Update of Extended Named Entity Definition and Categorization Data on Japanese Wikipedia 2019 version, In *Proceedings of the 26th annual meeting of the Association for Natural Language Processing*, 1221-1224. (in Japanese)

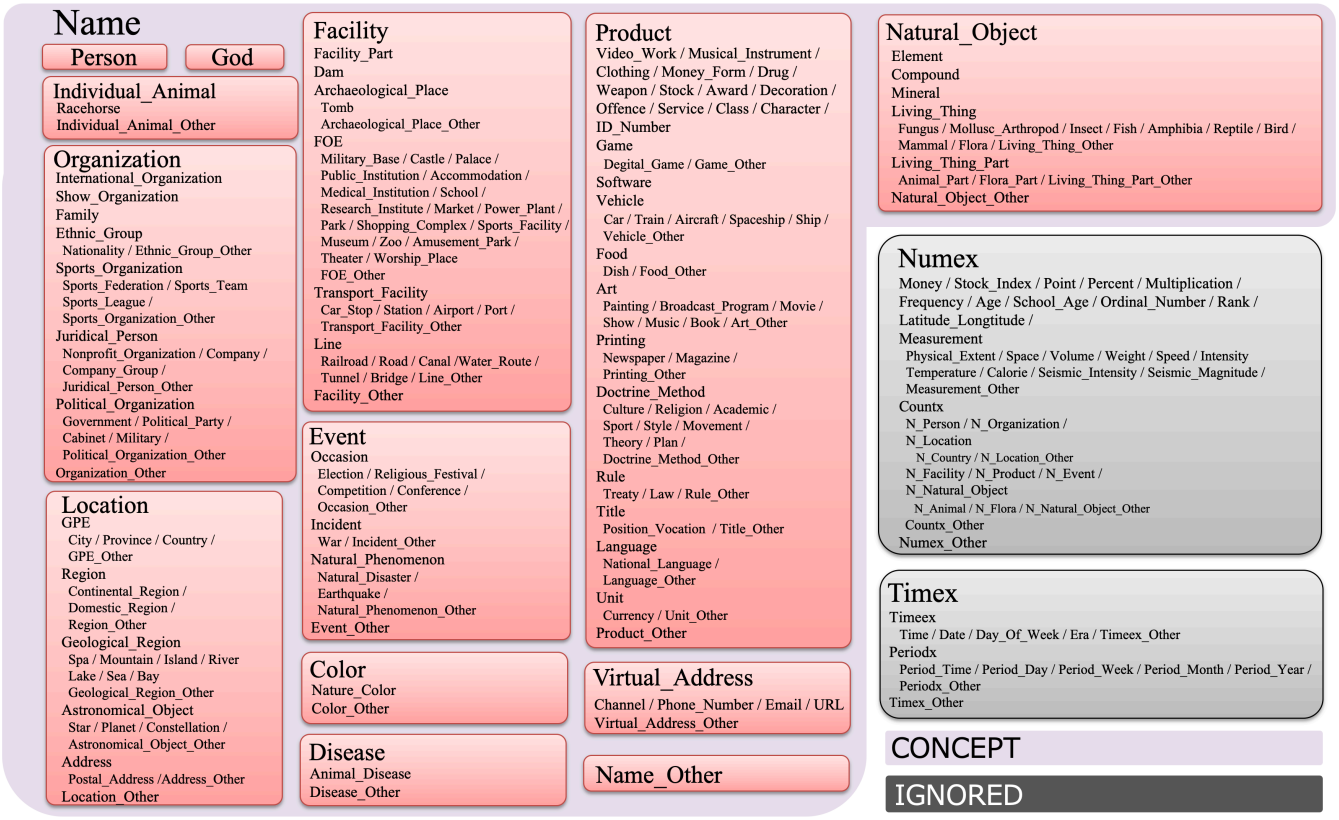

Figure 1: Definition of Extended Named Entity (ver.8.0)

Group ID		FPTAI	LIAT	PribL	PribL	RH312	ousia	uomfj	uomfj	uomfj	FPTAI	HUKB	HUKB	HUKB	LIAT
Method ID		BERT	ML-BERT	BERTGRU	BERTLINCONCAT	RnnGnnXlmr	RoBERTa+wiki2vec+wikidata	jointrep	jointrepPostprocess	jointrepUnionPostprocess	BERT	AB	ABC	AC	ML-BERT
Late Submission											Y	Y	Y	Y	Y
ar	Arabic	73.25	63.16	76.27	75.45	-	70.52	64.55	64.55	64.55	73.25	30.98	30.98	13.51	-
bg	Bulgarian	83.77	75.20	-	-	82.13	-	83.07	83.07	83.07	83.28	60.86	61.06	28.09	-
ca	Catalan, Valencian	52.55	76.28	-	-	-	-	79.82	79.82	79.82	81.10	42.34	42.54	16.26	-
cs	Czech	84.47	79.46	-	81.19	-	-	81.29	81.29	81.29	83.74	52.61	52.61	18.86	-
da	Danish	82.30	74.80	-	-	-	-	80.56	80.56	80.56	81.74	49.01	49.01	13.99	-
de	German	22.62	79.49	80.24	79.83	-	81.86	81.03	81.03	81.03	81.26	53.72	53.82	26.81	-
el	Greek, Modern (1453-)	84.40	72.43	-	-	-	-	-	-	-	84.10	7.51	7.51	7.51	-
en	English	82.23	78.56	81.27	80.12	-	-	82.73	82.57	82.68	81.96	45.11	45.11	11.92	-
es	Spanish, Castilian	80.60	77.73	80.30	80.72	-	80.94	81.39	81.39	81.39	80.60	49.21	49.11	19.50	-
fa	Persian	81.70	75.42	-	-	-	-	80.38	80.38	80.38	81.52	45.59	45.59	15.66	-
fi	Finnish	83.62	79.13	-	-	-	-	80.91	80.91	80.91	83.36	53.15	53.45	17.06	-
fr	French	21.59	76.88	77.93	78.52	80.31	81.01	78.21	78.21	78.21	80.68	43.84	43.74	11.23	-
he	Hebrew	83.79	79.11	-	-	-	-	81.09	81.09	81.09	84.21	59.95	60.05	15.78	-
hi	Hindi	76.43	16.49	-	-	71.70	69.75	66.67	66.67	66.67	75.65	39.70	39.51	22.02	-
hu	Hungarian	85.46	78.93	-	-	-	-	85.02	85.02	85.02	84.78	69.15	69.44	26.09	-
id	Indonesian	81.93	72.45	-	-	77.56	-	78.51	78.51	78.51	81.65	44.07	44.47	16.28	-
it	Italian	26.55	81.36	81.92	81.89	-	81.21	82.02	82.02	82.02	82.81	45.55	45.55	12.06	-
ko	Korean	83.67	80.38	81.51	81.04	-	-	82.51	82.51	82.51	83.77	63.68	63.98	13.95	-
nl	Dutch, Flemish	83.29	79.86	80.95	81.26	-	-	81.64	81.64	81.64	83.17	42.36	42.45	17.12	-
no	Norwegian	80.53	76.50	-	78.39	-	-	78.79	78.79	78.79	80.17	34.58	34.58	11.33	-
pl	Polish	84.53	80.60	82.73	83.46	-	-	84.52	84.52	84.52	84.07	62.72	63.51	32.55	-
pt	Portuguese	83.23	78.49	82.36	81.88	-	81.40	80.87	80.87	80.87	82.70	42.32	42.62	16.10	-
ro	Romanian, Moldavian, Moldovan	84.60	76.17	-	-	-	-	80.83	80.83	80.83	84.60	57.60	57.70	28.50	-
ru	Russian	84.08	79.09	82.60	83.07	-	-	82.90	82.90	82.90	83.44	42.04	42.24	11.30	-
sv	Swedish	83.18	71.63	-	-	-	-	-	-	-	83.44	50.32	50.62	21.98	79.58
th	Thai	81.26	49.58	-	-	76.77	76.36	65.02	65.02	65.02	81.16	39.98	40.38	24.05	-
tr	Turkish	86.50	77.19	84.36	83.23	83.28	-	84.85	84.85	84.85	86.03	61.88	62.48	16.73	-
uk	Ukrainian	83.12	78.71	-	-	-	-	81.61	81.61	81.61	82.61	60.29	60.19	22.51	-
vi	Vietnamese	80.34	75.24	-	-	-	-	77.06	77.06	77.06	80.42	60.38	60.48	22.14	-
zh	Chinese	81.25	77.97	78.38	79.37	-	79.76	78.58	78.58	78.58	80.60	21.22	21.42	17.57	-

Table 5: Results of SHINRA2020-ML