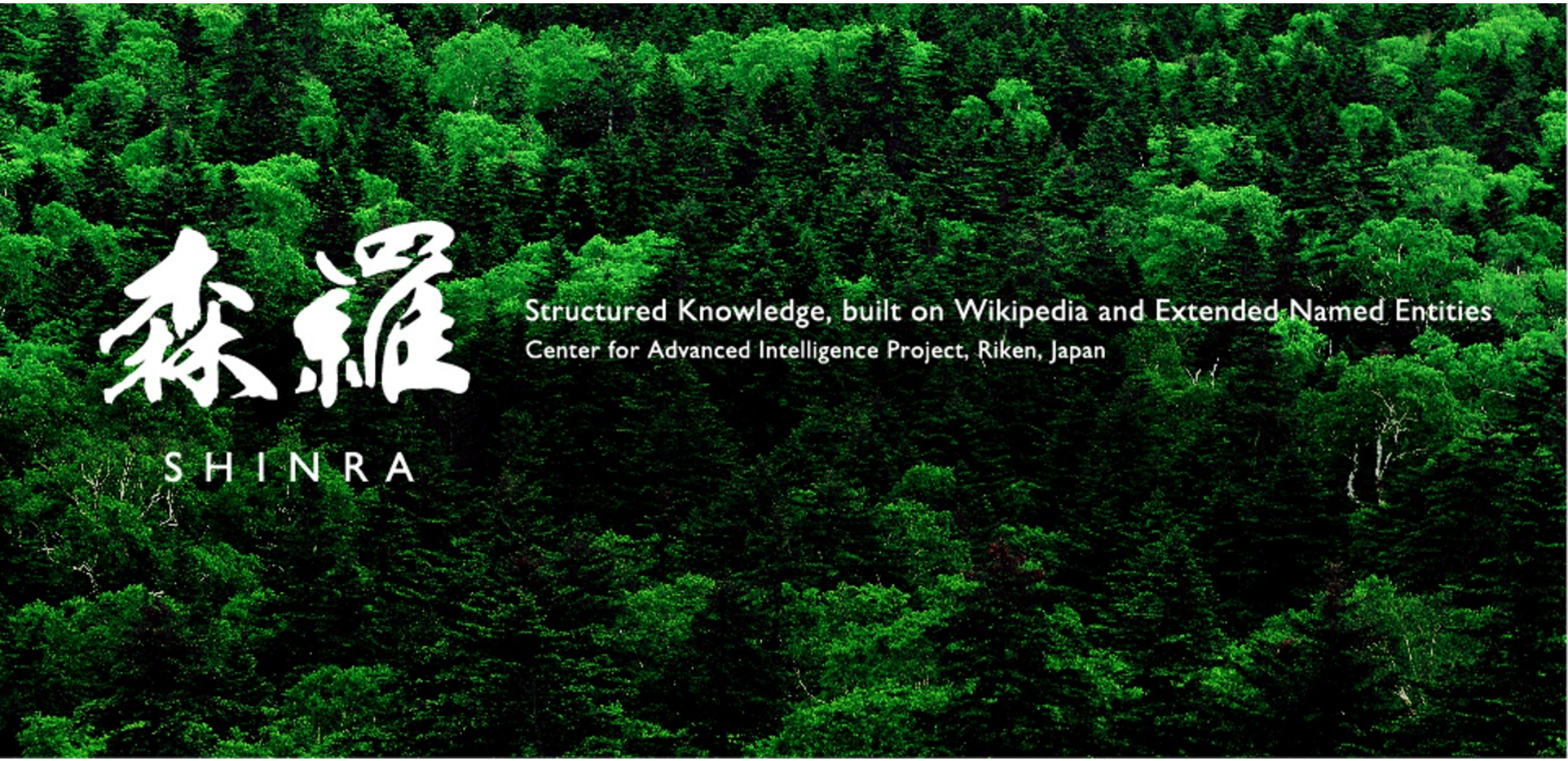
SHINRA2020-ML

Categorizing 30 language Wikipedia into Extended Named Entity categories

Satoshi Sekine, Kouta Nakayama, Maya Ando, Yu Usami, Masako Nomoto, Koji Matsuda, Asuka Sumida



2021.10.4-6 @ AKBC2021

Language Information Access Technology Team, AIP, RIKEN

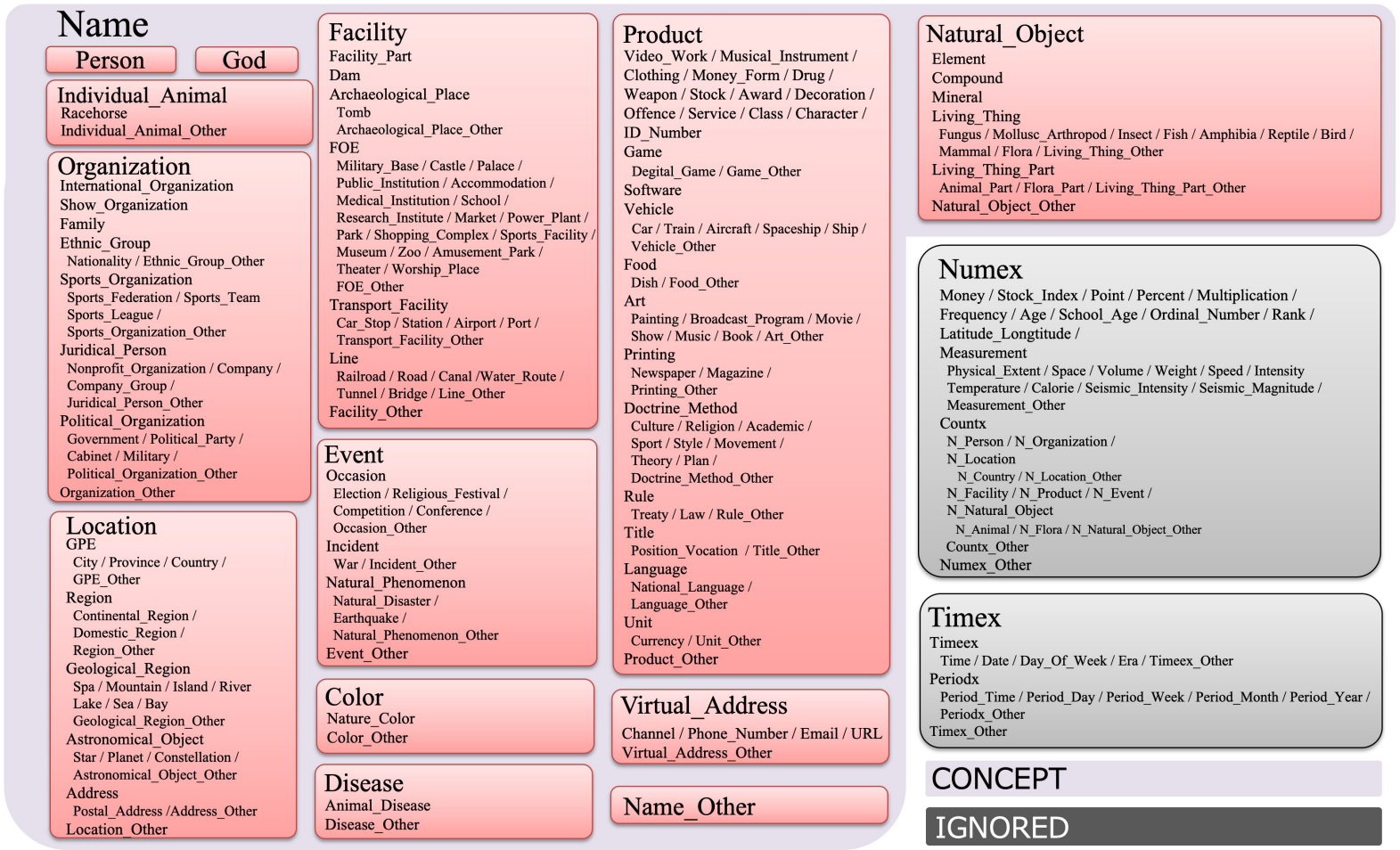
SHINRA project
(Since 2017)

- Goal:
  - Knowledge Base Construction for Explainable AI
  - Categorizing and Structuring Wikipedia
  - Shared-task to build resource together!!

    (Resource by Collaborative Contribution)
- SHINRA2020-ML (among 6 tasks total)
  - Categorize 30-language Wikipedia pages (32M in total) into 200 ENE categories (fine-grained NE) *1

*1: English, Spanish, French, German, Chinese, Russian, Portuguese, Italian, Arabic, Indonesian, Turkish, Dutch, Polish, Persian, Swedish, Vietnamese, Korean, Hebrew, Romanian, Norwegian, Czech, Ukrainian, Hindi, Finnish, Hungarian, Danish, Thai, Catalan, Greek, Bulgarian.

Extended Named Entity



[ENE version 8] http://ene-project.info/ene8/?lang=en

Final Goal:
Well-Structured Knowledge Graph

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/KG are very noisy

SHINRA2020-ML task

- Shared-task to build resource together!!
  - We prepare training/test data
  - We don't make the test data open, the participants have to run the system for the entire resource
  - The system outputs are open to public and anyone can create the resource base on them (Ensemble Learning)
- Schedule
  - Starts in April 2020, submission in August, workshop in December

The focus is "Resource Construction"
NOT evaluation

Resource by Collaborative Contribution (RbCC)

Participants

(including leaderboard-only participants)

# of groups	10 (7 active participants)
nationality	Japan(4), Vietnam (2), India (1), Taiwan (1), Australia(1), Finland(1), Portugal(1)
affiliation types	University (6), Company (4), Institute (1)
target languages	8: Arabic, French 7: Chinese, German, Hindi, Italian, Portuguese, Spanish, Thai, Turkish 6: Bulgarian, Czech, Dutch, English, Indonesian, Korean, Norwegian, Polish, Russian, Vietnamese 5: Catalan, Danish, Finnish, Hebrew, Hungarian, Persian, Romanian, Ukrainian 4: Greek, Swedish
# of target languages	30(4), 28(1), 15(1), 9(1), 6(1), 1(2)

Evaluation Results

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	-

Simple voting results (ensemble) outperforms the best system!

ISO 639-1	Language	Group ID	Method	Precision	Recall	F1	Majority Voting F1	Oracle F1	Num Groups	Num Methods
tr	Turkish	FPTAI	BERT	84.22	88.92	86.50	87.38	92.71	7	12
hu	Hungarian	FPTAI	BERT	82.89	88.19	85.46	85.49	91.18	5	9
ro	Romanian, Moldavian, Moldovan	FPTAI	BERT	81.40	88.07	84.60	84.47	91.97	5	9
pl	Polish	FPTAI	BERT	82.01	87.22	84.53	85.27	91.55	6	11
cs	Czech	FPTAI	BERT	81.31	87.88	84.47	84.52	90.59	6	10
el	Greek, Modern (1453-)	FPTAI	BERT	81.34	87.70	84.40	75.76	90.26	4	6
he	Hebrew	FPTAI	BERT	80.50	88.28	84.21	84.34	92.22	5	9
ru	Russian	FPTAI	BERT	81.59	86.73	84.08	84.73	90.50	6	11
bg	Bulgarian	FPTAI	BERT	80.94	86.81	83.77	84.74	91.04	6	10
ko	Korean	FPTAI	BERT	80.44	87.39	83.77	84.22	91.95	6	11
fi	Finnish	FPTAI	BERT	79.98	87.61	83.62	83.61	90.46	5	9
sv	Swedish	FPTAI	BERT	80.20	86.94	83.44	82.21	91.38	5	9
nl	Dutch, Flemish	FPTAI	BERT	81.27	85.41	83.29	83.85	90.73	6	11
pt	Portuguese	FPTAI	BERT	79.80	86.97	83.23	83.98	93.17	7	12
uk	Ukrainian	FPTAI	BERT	80.05	86.43	83.12	83.92	89.81	5	9
it	Italian	FPTAI	BERT	79.98	85.84	82.81	83.72	92.77	7	12
en	English	uomfj	jointrep	81.77	83.71	82.73	82.66	89.60	6	11
da	Danish	FPTAI	BERT	79.47	85.33	82.30	80.93	90.49	5	9
id	Indonesian	FPTAI	BERT	78.23	86.01	81.93	81.44	90.40	6	10
de	German	ousia	RoBERTa+wiki2vec+wikidata	82.59	81.15	81.86	82.45	90.63	7	12
fa	Persian	FPTAI	BERT	79.35	84.18	81.70	81.09	88.54	5	9
es	Spanish, Castilian	uomfj	jointrepUnionPostprocess	82.20	80.59	81.39	82.88	89.25	7	12
th	Thai	FPTAI	BERT	78.07	84.72	81.26	81.14	90.69	7	11
zh	Chinese	FPTAI	BERT	78.83	83.82	81.25	80.83	89.45	6	11
ca	Catalan, Valencian	FPTAI	BERT	77.34	85.25	81.10	80.57	91.11	5	9
fr	French	ousia	RoBERTa+wiki2vec+wikidata	81.09	80.93	81.01	81.92	90.32	8	13
no	Norwegian	FPTAI	BERT	77.58	83.71	80.53	81.27	89.44	6	10
vi	Vietnamese	FPTAI	BERT	77.61	83.43	80.42	80.16	91.62	6	10
hi	Hindi	FPTAI	BERT	73.67	79.41	76.43	73.67	84.51	7	11
ar	Arabic	PribL	BERTGRU	76.80	75.74	76.27	73.39	90.89	8	13
MAX				84.22	88.92	86.50	87.38	93.17	8	13
MIN				73.67	75.74	76.27	73.39	84.51	4	6

Links

- Homepage

  http://shinra-project.info/shinra2020ml/?lang=en

- Email

  shinra2020ml-info@googlegroups.com