Overview of the Recognizing Inference in Text (RITE-2) at NTCIR-10

Yotaro Watanabe[1], Yusuke Miyao[2], Junta Mizuno[1], Tomohide Shibata[3], Hiroshi Kanayama[4], Cheng-Wei Lee[5], Chuan-Jie Lin[6], Shuming Shi[7], Teruko Mitamura[8], Noriko Kando[2], Hideki Shima[8] and Kohichi Takeda[4]

[1]Tohoku University, Japan [2]National Institute of Informatics, Japan [3]Kyoto University, Japan [4]IBM Research, Japan
[5]Academia Sinica, Taiwan [6]National Taiwan Ocean University, Taiwan [7]Microsoft Research Asia, P.R.China [8]Carnegie Mellon University, U.S.A

http://www.cl.ecei.tohoku.ac.jp/rite2/doku.php

INTRODUCTION

This poster introduces an overview of the RITE2 task in NTCIR-10. We evaluate systems that automatically recognize entailment, paraphrase, and contradiction between two texts written in Japanese, Simplified Chinese, or Traditional Chinese. The task consists of six subtasks: Binary classification of entailment (BC); Multi-class classification including paraphrase and contradiction (MC); Unit Test which provides a breakdown of linguistic phenomena (UnitTest); and three extrinsic application-oriented datasets: Exam BC, Exam Search and RITE4QA.

BC Subtask

Given a pair of texts (t_1, t_2), a system automatically identifies if t_1 entails t_2 or not. (The premise t_1 entails the hypothesis t_2 if a human reading t_1 would infer that t_2 is most likely true.)

MC Subtask

A system needs to classify a pair into one of five categories considering entailment direction, paraphrase and contradiction. Output labels:

- **F**: forward entailment (t_1 entails t_2 AND t_2 does not entail t_1).
- **B**: bidirectional entailment (t_1 entails t2 AND t_2 entails t_1).
- **C**: contradiction (t_1 and t_2 contradict, or cannot be true at the same time).
- **I**: independence (otherwise)

Unit Test new!

A data set provides a breakdown of linguistic phenomena. It can be used for various research including the following purposes.

- Analyze linguistic issues that appear in the RITE data
- Evaluate recognition accuracy for each phenomenon
- Develop/Train a recognizer for each linguistic phenomenon

Entrance Examination Subtask

Created based on actual entarance exams for university admission in Japan: National Center Test for University Admission. The subtask provides two types of data.

BC style (ExamBC)

The data is provided in the same form as the BC subtask. Systems are asked to recognize inference relations between t1 and t2. In this data, t1 is extracted from Wikipedia, while t2 is taken from university entrance exams.

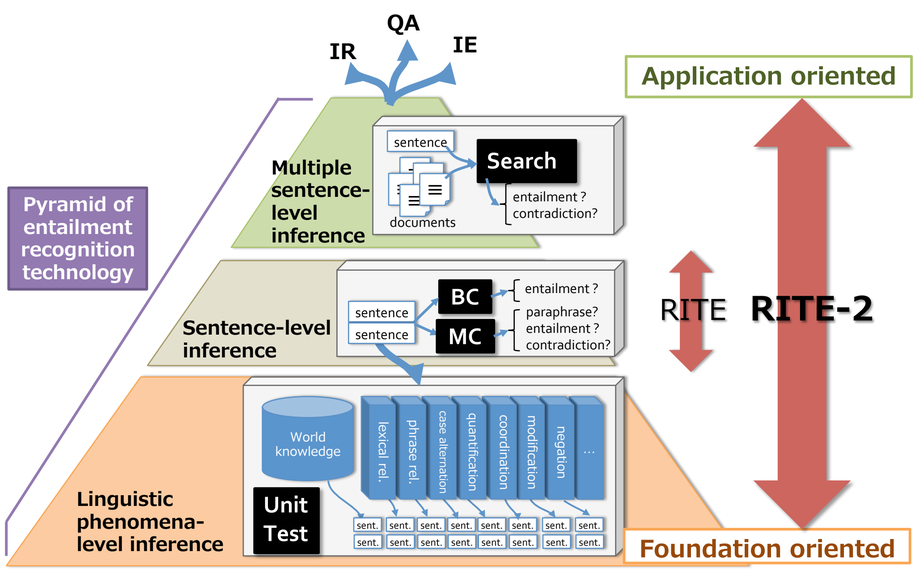
Search style (ExamSearch) new!

In this data, t1 is not given. Systems are asked to retrieve texts that can be used as t1 from Wikipedia or textbooks, and answer whether t2 is entailed (inferred) from retrieved texts.

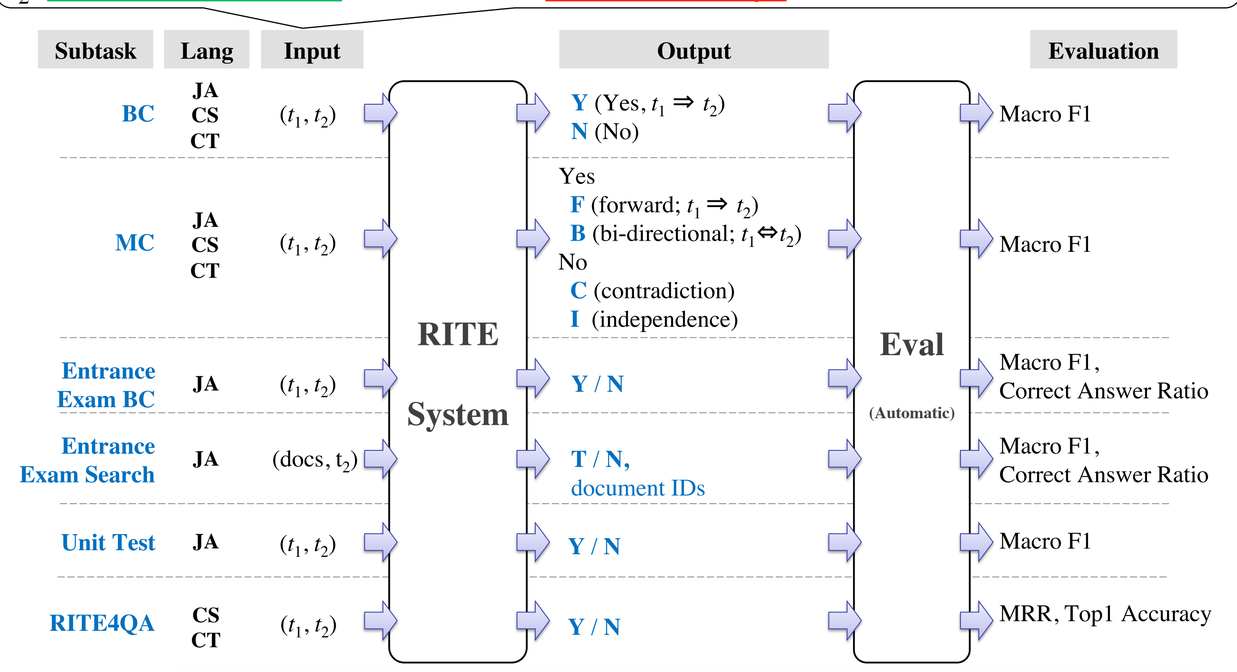
RITE4QA Subtask

Another application-oriented dataset created from the NTCIR-6 CLQA dataset (does an answer-candidate-bearing sentence entail a question in affirmative form?).

ADVANCES FROM NTCIR-9 RITE



t_1: Yasunari Kawabata won the Nobel Prize in Literature for his novel "Snow Country".
t_2: Yasunari Kawabata is the writer of "Snow Country".



DATASET SIZE

BC

	Y	N	Total
JA (dev)	240	371	611
JA (test)	256	354	610
CS (dev)	528	286	814
CS (test)	422	359	781
CS (arti)	341	232	573
CT (dev)	716	605	1321
CT (test)	479	402	881
CT (arti)	341	232	573

MC

	F	B	C	I	Total
JA (dev)	207	83	65	193	548
JA (test)	205	70	61	212	548
CS (dev)	369	159	146	140	814
CS (test)	277	145	106	253	781
CS (arti)	155	186	115	112	568
CT (dev)	544	262	254	261	1321
CT (test)	328	151	114	288	881
CT (arti)	155	186	115	112	568

ExamBC

	Y	N	Total
JA (dev)	210	300	510
JA (test)	173	275	448

ExamSearch

	Y	N	Total
JA (dev)	210	300	510
JA (test)	173	275	448

UnitTest

	Y	N	Total
JA (dev)	239	33	272
JA (test)	212	29	241

Linguistic phenomena		dev	test	Linguistic phenomena	dev	test	Linguistic phenomena		dev	test
lexical	synonymy	10	10	nominalization	2	1	Implicit relation		10	28
	hypernymy	6	3	coreference	12	4	apposition		3	1
	meronymy	1	1	clause	29	14	temporal		2	1
	entailment	1	0	Relative clause	10	8	spatial		4	1
phrase	synonymy	45	35	Transparent head	2	1	disagree	lexical	5	2
	hypernymy	3	0	list	11	3		phrase	25	25
	entailment	28	45	quantity	1	0		modality	2	1
Case alternation		9	7	scrambling	16	15		spatial	1	1
modifier		30	42	inference	4	2		temporal	0	1

FORMAL RUN RESULTS

BC

JA

Run	MacroF1
DCUMT-01	80.49
WSD-03	80.08
SKL-02	79.46
BnO-03	78.93
WSD-02	78.77
WSD-01	78.61
SKL-01	78.61
BnO-02	78.31
BnO-01	77.61
KitAi-01	77.11
OKA1-02	76.71
JAIST-02	76.47
SKL-03	76.4
KitAi-03	76.16
JAIST-01	75.56
OKA1-01	74.59
KYOTO-02	74.5
IBM-01	74.49
IBM-02	73.4
JAIST-03	73.08
IBM-03	72.9
KitAi-02	72.35
U-TOKYO-01	72.23
OKA1-03	71.71
FLL-03	67.99
*TKDDI-03	63.83
TKDDI-02	63.55
*TKDDI-02	63.55
*TKDDI-01	63.45
TKDDI-01	63.45
FLL-01	63.06
NTTD-03	61.9
*FLL-05	61.05
FLL-02	59.73
U-TOKYO-03	59.01
U-TOKYO-02	57.84
NTTD-01	57.59
*FLL-06	55.69
EHIME-01	54.34
*FLL-04	52.58
THK-01	52.4
NTTD-02	50.38
EHIME-02	50.14
JUNLP-01	48.83
EHIME-03	48.05
KYOTO-03	46.42
KYOTO-01	41.97
Baseline-01	62.53

CS

Run	MacroF1
bcNLP-03	73.84
MIG-02	68.09
CYUT-03	67.86
bcNLP-01	67.04
bcNLP-02	66.89
MIG-01	65.71
CYUT-02	63.11
WHUTE-02	61.65
CYUT-01	61.17
*IASL-02	60.45
WHUTE-01	58.2
MIG-03	57.19
IMTKU-03	54.28
Yuntech-03	53.52
Yuntech-02	52.1
Yuntech-01	50.91
IMTKU-01	50.82
*IASL-01	50.6
WUST-02	50.14
WUST-01	50.14
*WUST-01	50.14
WUST-03	50.14
IMTKU-02	50.12
JUNLP-01	48.49

CT

Run	MacroF1
IASL-02	67.14
MIG-02	67.07
MIG-03	66.99
IMTKU-01	65.99
WHUTE-01	65.55
MIG-01	65.42
IMTKU-03	63.82
Yuntech-03	62.31
Yuntech-02	62.02
Yuntech-01	61.64
KC99-01	57.67
CYUT-01	55.16
CYUT-02	52.64
IASL-01	51.77
CYUT-03	51.58
JUNLP-01	48.72
IMTKU-02	48.61
NTOUA-01	32.63
NTOUA-03	31.71
NTOUA-02	30.7

MC

JA

Run	MacroF1
SKL-01	59.96
SKL-02	58.25
SKL-03	55.45
WSD-03	54.39
WSD-02	54.18
WSD-01	53.98
FLL-01	53.67
JAIST-01	52.6
BnO-02	52.44
JAIST-02	52.27
BnO-03	52.11
BnO-01	52.03
JAIST-03	51.48
*FLL-04	51.27
KYOTO-02	50.12
*FLL-02	35.12
THK-01	30.98
EHIME-03	25.89
EHIME-01	24.47
*FLL-03	22.47
EHIME-02	21.99
JUNLP-01	21.42
KYOTO-01	17.04
Baseline-01	26.61

CS

Run	MacroF1
bcNLP-03	56.82
*IASL-02	50.94
WHUTE-01	46.79
WHUTE-02	46.53
bcNLP-02	44.88
MIG-02	44.74
CYUT-02	42.52
MIG-01	41.82
Yuntech-02	40.91
Yuntech-03	40.89
WUST-02	40.87
WUST-03	40.87
*WUST-01	40.87
CYUT-01	40.37
WUST-01	40.33
Yuntech-01	40.33
CYUT-03	40.1
bcNLP-01	39.95
*IASL-01	34.95
MIG-03	34.42
IMTKU-03	27.26
JUNLP-01	24.38
IMTKU-01	23.89
IMTKU-02	19.67

CT

Run	MacroF1
IASL-02	46.32
WHUTE-01	45.5
MIG-02	45.15
NTOUA-03	44.8
NTOUA-01	44.63
MIG-03	44.21
KC99-01	43.75
MIG-01	42.16
Yuntech-03	40.14
Yuntech-01	39.76
Yuntech-02	38.77
IMTKU-01	35.76
NTOUA-02	33.49
IASL-01	33.04
MCUIM-01	32.51
IMTKU-03	32.36
CYUT-02	26.26
CYUT-01	25.6
JUNLP-01	24.21
CYUT-03	23.51
IMTKU-02	19.37

Exam (JA)

Exam BC

Run	MacroF1	CAR
BnO-02	67.15	55.56
BnO-03	66.97	57.41
KDR-02	66.9	51.85
BnO-01	66.86	57.41
KDR-03	66.64	47.22
WSD-01	64.9	52.78
IBM-03	64.18	45.37
WSD-03	64.71	52.78
SKL-02	64.04	49.07
WSD-02	63.96	51.85
KDR-01	63.31	49.07
SKL-01	61.65	29.63
SKL-03	60.47	42.59
KitAi-01	59.84	36.11
IBM-02	59.33	46.29
KitAi-03	59.05	45.37
JAIST-02	59.04	41.67
JAIST-03	58.65	42.59
IBM-01	58.53	44.44
JAIST-01	57.55	40.74
KitAi-02	57.16	39.81
KYOTO-02	56.82	43.52
NTTD-02	55.57	34.26
NTTD-03	53.12	34.26
NTTD-01	52.02	31.48
JUNLP-01	50.46	30.56
*TKDDI-03	49.08	28.7
TKDDI-01	48.62	26.85
TKDDI-02	48.62	26.85
THK-01	43.77	26.85
KYOTO-03	38.57	21.3
KYOTO-01	37.86	20.37
Baseline-01	54.77	32.41

Exam Search

Run	MacroF1	CAR
*KDR-02	58.12	32.41
*KDR-01	57.59	33.33
*KDR-03	57.39	34.26
NTTD-01	55.02	25.93
*BnO-02	54.77	31.48
*BnO-01	52.45	26.85
*BnO-03	51.78	31.48
NTTD-02	49.15	25.93
KYOTO-01	46.57	28.7
KYOTO-02	45.41	26.85

* CAR = Correct Answer Ratio

UnitTest (JA)

Run	MacroF1
*FLL-01	77.77
*FLL-03	76.98
JAIST-02	74.52
*TKDDI-03	74
*TKDDI-01	73.51
*TKDDI-02	73.51
TKDDI-01	73.51
TKDDI-02	73.51
BnO-01	73.44
NTTD-02	68.98
BnO-03	68.78
JAIST-01	67.36
NTTD-01	61.99
THK-01	53.26
*FLL-02	51.35
NTTD-03	47.91
BnO-02	46.8
KYOTO-02	45.35
KYOTO-01	37.27
JAIST-03	29.46
Baseline-01	51.7

RITE4QA

CS

Run	WorseRanking R Top1	WorseRanking R MRR	WorseRanking R+U Top1	WorseRanking R+U MRR	BetterRanking R Top1	BetterRanking R MRR	BetterRanking R+U Top1	BetterRanking R+U MRR
IMTKU-03	28	33.77	33.33	40.87	28	33.77	33.33	40.87
WHUTE-01	18.67	27.59	22	33.67	18.67	27.64	22.67	34.06
IMTKU-02	14.67	21.44	20	29.08	14.67	21.44	20	29.08
*IASL-02	12.67	18.8	15.33	23.13	38	42.71	42	47.27
*IASL-01	12	22.71	15.33	27.96	31.33	38.09	36	43.74
IMTKU-01	10.67	19.91	17.33	27.21	10.67	19.91	17.33	27.21
CYUT-02	7.33	11.82	10.67	17.32	39.33	46.13	44	51.52
CYUT-01	6.67	11.49	10	16.99	38	45.19	43.33	51.08
bcNLP-03	6.67	8.78	7.33	9.78	8	9.33	9.33	10.67
bcNLP-01	3.33	3.67	4.67	6.17	2.67	3.22	6	6.78
bcNLP-02	0	0	0	0	0	0	0	0
orgQAsys-01	7.33	11.32	10.67	16.99	40.67	47.6	44.67	52.32

CT

Run	WorseRanking R Top1	WorseRanking R MRR	WorseRanking R+U Top1	WorseRanking R+U MRR	BetterRanking R Top1	BetterRanking R MRR	BetterRanking R+U Top1	BetterRanking R+U MRR
WHUTE-01	27.33	34.57	30.67	38.76	26.67	34.29	30	38.48
IMTKU-03	16.67	25.7	19.33	29.39	17.33	26.03	20	29.72
IMTKU-01	14.67	22.69	22	31.84	14.67	22.58	22	31.73
CYUT-03	12.67	17.69	18.67	26.6	12	17.5	18.67	26.99
IASL-01	12	22.14	14.67	26.86	30	38.27	32.67	42.34
IMTKU-02	12	19.82	19.33	29.43	12	19.84	20	29.79
IASL-02	10.67	16.01	14.67	22.09	38	42.74	42	47.66
NTOUA-03	8	9.97	12.67	17.19	9.33	10.63	14.67	18.3
NTOUA-01	8	9.28	13.33	17.06	8.67	9.61	15.33	18.17
NTOUA-02	7.33	8.78	11.33	14.11	8	9.22	12	14.56
CYUT-01	6.67	11.71	10	16.99	38	45.19	43.33	51.08
CYUT-02	6.67	11.71	10	16.99	38	45.19	43.33	51.08
orgQAsys-01	7.33	11.54	10.67	16.99	40.67	47.6	44.67	52.32

OUTCOME

- We constructed large-scale reusable evaluation datasets
- Baseline Tool is a Python development framework for implementing a RITE system based on machine learning
- Automatic evaluation output not only accuracy and macro-F1 but also confusion matrix and details of the results on the Unit Test data

- Best runs were able to outperform the simple machine learning-based baseline
- In the JA-BC subtask, the top system achieved over 80% in accuracy (in contrast to 58% in NTCIR-9 RITE)
- UnitTest dataset enabled detailed analysis of RTE systems
- Exam Search subtask is more realistic than the traditional settings