Overview of the NTCIR-11 Recognizing Inference in TExt and Validation (RITE-VAL) Task

Suguru Matsuyoshi
University of Yamanashi, Japan

Yusuke Miyao
National Institute of Informatics, Japan

Tomohide Shibata
Kyoto University, Japan

Chuan-Jie Lin
National Taiwan Ocean University, Taiwan

Cheng-Wei Shih
Academia Sinica, Taiwan

Yotaro Watanabe
NEC Corporation, Japan

Teruko Mitamura
Carnegie Mellon University, U.S.A.

SUMMARY

Subtask	Language	Task Description	Acronym	Test Data Size	Submissions	Top Macro-F1	Top Accuracy	Top Team (Run Num.)
Fact Validation	Simplified Chinese	Multi-Classification	CS-FV	613	12	38.93	44.05	III&CYUT (05)
	Traditional Chinese	Multi-Classification	CT-FV	613	15	39.51	44.70	III&CYUT (02)
	Japanese	Binary Classification	JA-FV	514	30	61.93	63.23	NUL (03)
		Passage Search	--	514	3	--	--	--
	English	Binary Classification	EN-FV	188	9	53.17	55.85	BnO (01)
System Validation	Simplified Chinese	Binary Classification	CS-SVBC	1,200	23	61.51	62.33	BUPTTeam (05)
		Multi-Classification	CS-SVMC	1,200	18	44.39	51.83	WUST (01)
	Traditional Chinese	Binary Classification	CT-SVBC	1,200	17	56.24	56.25	III&CYUT (04)
		Multi-Classification	CT-SVMC	1,200	17	40.54	43.33	III&CYUT (05)
	Japanese	Binary Classification	JA-SV	1,379	26	69.59	77.81	NUL (04)
Total					170			

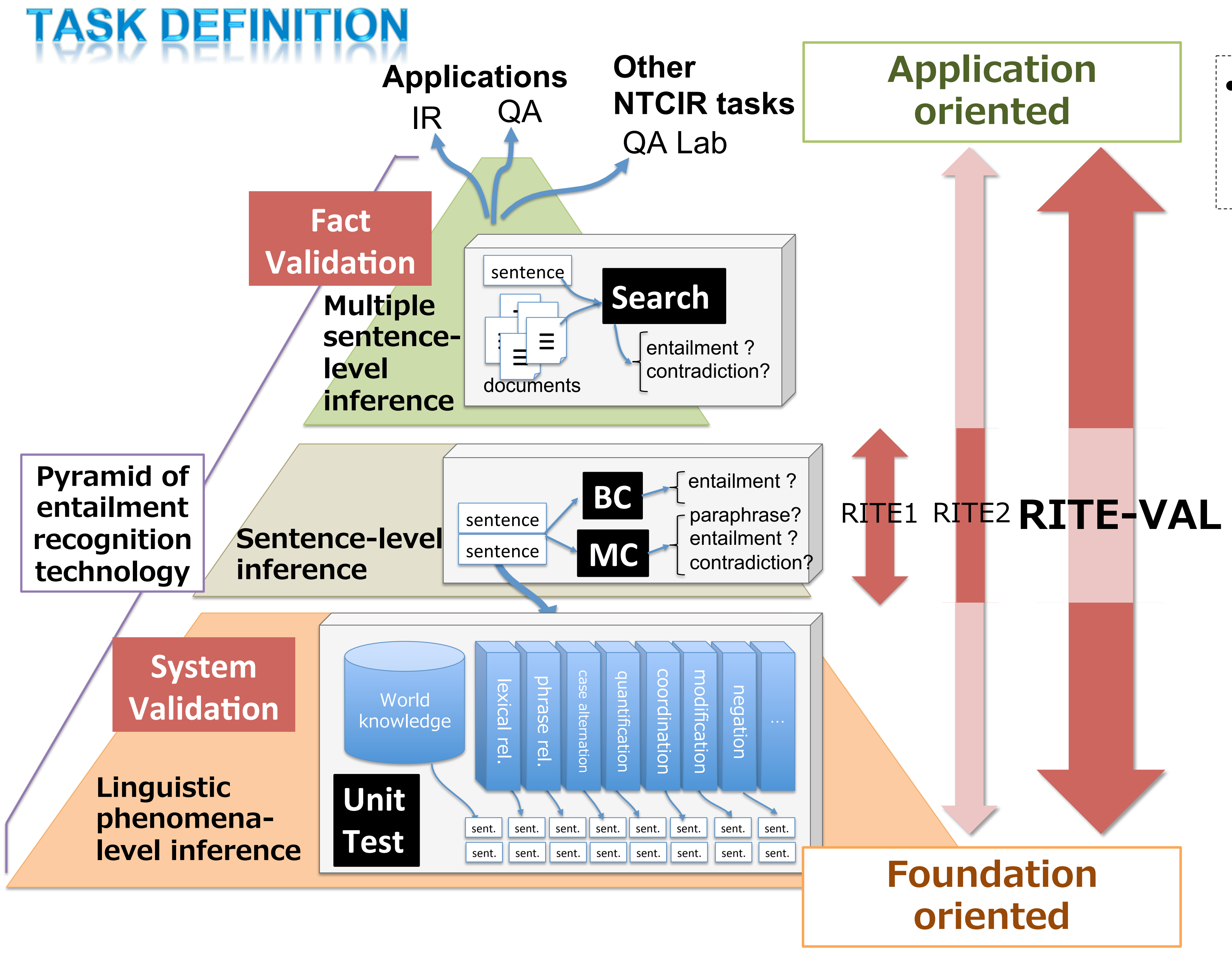
Active participating team:

23 teams

- 11 from Japan
- 7 from Taiwan
- 4 from China
- 1 from Norway
- 1 from Vietnam

(One team consists of people from Japan and Vietnam.)

TASK DEFINITION



- Recognizing Textual Entailment

Given a text t_1, can a computer infer that a hypothesis t_2 is most likely true (i.e., t_1 entails t_2) ?

t_1: 现代铅笔以石墨和粘土来制造。
t_2: 铅笔中含有碳的成分。
entailment

t_1: 約瑟夫·傅立葉是十九世紀法國數學家、物理學家。
t_2: 約瑟夫·傅立葉是物理學家。
entailment

t_1: ジュール・ヴェルヌの『八十日間世界一周』の中で、80日で世界一周が出来るかどうかの賭けが行われた。
t_2: ジュール・ヴェルヌの『八十日間世界一周』をモデルとして実際にリポーターを世界一周させるという企画がある。
not entailment

t_1: The goal of u-Japan is to achieve a ubiquitous network society in which anything and anyone can easily access networks.
t_2: The term "ubiquitous network society" refers to a society in which disparities have emerged in the amount of information that can be obtained using the Internet.
not entailment

Fact Validation

Given Wikipedia, some textbooks and a sentence (t_2), your system judges whether the document set entails t_2 or not.

Your system need searching for a text passage corresponding to t_1 in the given document set. And, based on the search result, the system must determine whether the document set entails t_2. If t_2 is entailed by some sentences in the document set, the statement described in t_2 can be judged as "fact."

System Validation

Given a text (t_1) and a hypothesis (t_2), your system judges whether t_1 entails t_2 or not.

The system is provided with the following two types of sentence pairs:

- Several linguistic phenomena are involved in the decision whether t_1 entails t_2. [JA-SV]
- **A single linguistic phenomenon is involved in the decision whether t_1 entails t_2.** [CS-SVBC, CS-SVMC, CT-SVBC, CT-SVMC]

A list of sentence pairs of the latter type is made from a usual sentence pair. While the RITE task aims at integrated semantic/context processing systems, it also has a problem that research focused on a specific linguistic phenomenon is not easy to pursue. This subtask provides a data set that includes a breakdown of linguistic phenomena that are necessary for recognizing relations between t_1 and t_2.

SVMC label	Train	Test
Bidirectional entailment	222	300
Forward entailment	148	300
Contradiction	152	300
Independence	59	300
Total	581	1,200

Linguistic Phenomenon	Train	Test
abbreviation	6	25
apposition	7	25
case alternation	21	27
clause	25	59
coreference	11	24
hypernymy	30	27
inference	75	184
lexical entailment	12	29
list	20	37
meronymy	4	23
modifier	37	131
paraphrase	47	49
quantity	11	29
relative clause	6	36
scrambling	27	35
spatial	18	42
synonymy: lex	48	51
temporal	11	40
transparent head	13	26
antonym	20	35
exclusion: common sense	8	34
exclusion: modality	12	38
exclusion: modifier	14	33
exclusion: predicate argument	51	38
exclusion: quantity	6	29
exclusion: spatial	14	32
exclusion: temporal	7	34
negation	20	28
Total	581	1,200

FORMAL RUN RESULTS

- Overview of the participating systems from the aspect of "run"

Approach	CS	CT	JA	EN	Total
Rule-based	0	3	6	1	10 (6%)
Statistical	13	18	42	0	73 (47%)
Hybrid	33	28	5	8	74 (47%)

Statistical approaches:
SVM, Naïve Bayes, Threshold model, Penalized frequency distribution, Algebraic inference engine, Random forests, etc.

Feature / Information	CS	CT	JA	EN	Total
alignment	6	15	17	0	38
char/word overlapping	**45**	**48**	**51**	**8**	**152**
entailment rule	21	18	5	7	51
entity/event	7	17	1	0	25
hypernym	22	35	26	**9**	92
meronym	8	16	3	2	29
modality	0	2	4	0	6
named entity	24	27	35	**9**	95
overlapping	**45**	**46**	34	**9**	134
polarity	11	13	2	7	33
predicate argument relationship	16	15	8	**9**	48
synonym/antonym	41	**44**	33	**9**	127
syntactic information	25	17	14	**9**	65
temporal/numeric information	**43**	**47**	27	7	124
transformation	9	27	3	2	41
(number of the submitted runs)	53	49	59	9	170

Figures in red in the table indicate the ones more than 80% of the number of the submitted runs for each language.

Resources:
WordNet, Wikipedia, TongYiCiCiLin, Hownet, Goi-Taikei, FrameNet, VerbNet, EDR dictionary, etc.

CONCLUDING REMARKS AND FUTURE WORK

- Recognizing textual entailment in any of the four languages is still a difficult task for computers.
- System Validation subtask helps researchers to be aware of weakness of their system. We need further investigations of insufficient language resources and related linguistic phenomena in addition to continued construction of training data.
- We would like to work in cooperation with Todai Robot Project and Project Next NLP.

RITE-VAL Website: https://sites.google.com/site/ntcir11riteval/home