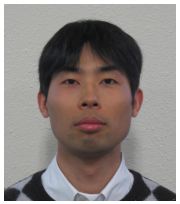
Overview of NTCIR-11 RITE-VAL Task

(Recognizing Inference in Text and Validation)

Suguru Matsuyoshi
University of Yamanashi

Yusuke Miyao
National Institute of Informatics



Tomohide Shibata
Kyoto University

Chuan-Jie Lin
National Taiwan Ocean University

Cheng-Wei Shih
Academia Sinica

Yotaro Watanabe
NEC Corporation

Teruko Mitamura
Carnegie Mellon University

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

NTCIR-11 Conference December 10th, 2014

Recognizing Inference in Text (RITE)

- RITE is a benchmark task for automatically detecting the following semantic relations between two sentences:
  - **entailment, paraphrase and contradiction**.
- [Main task]
  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: Yasunari Kawabata won the Nobel Prize in Literature for his novel “Snow Country.”

> t_2: Yasunari Kawabata is the writer of “Snow Country.”

Target languages

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。

Simplified Chinese (CS)

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。

Traditional Chinese (CT)

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。

Japanese (JA)

t_1: The goal of u-Japan is to achieve a ubiquitous network society in which anything and anyone can easily access networks.

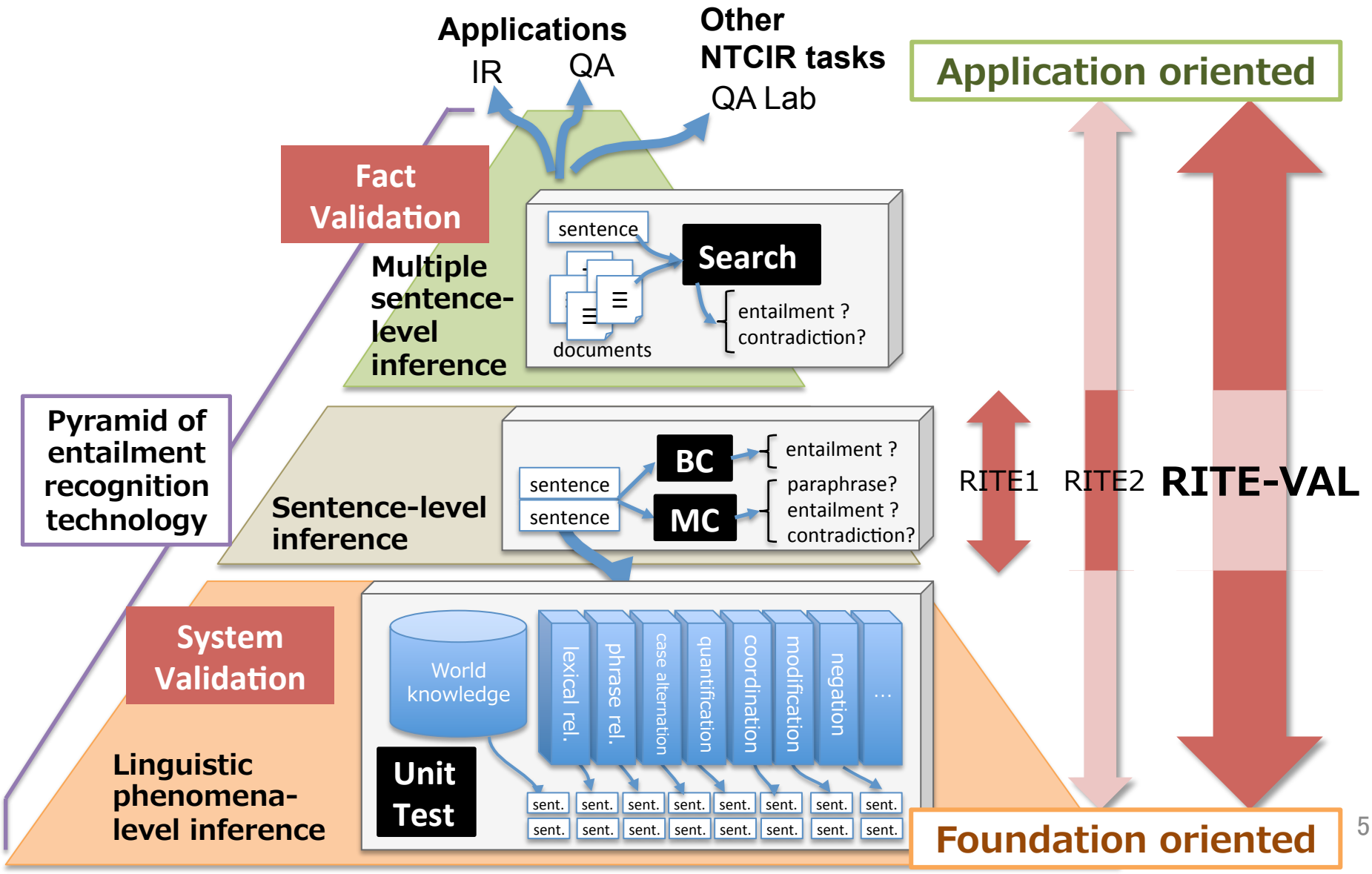
t_2: The term "ubiquitous network society" refers to a society in which disparities have emerged in the amount of information that can be obtained using the Internet.

English (EN)

Motivation

- RITE technology can be applied for Natural Language Processing (NLP) and various Information Access (IA) applications:
  - Question Answering, Information Retrieval, Information Extraction, Text Summarization, Automatic evaluation for Machine Translation, Complex Question Answering, etc.
- **RITE-VAL is the third round of RITE task**.
  - NTCIR-9 RITE1
  - NTCIR-10 RITE2
  - NTCIR-11 RITE-VAL

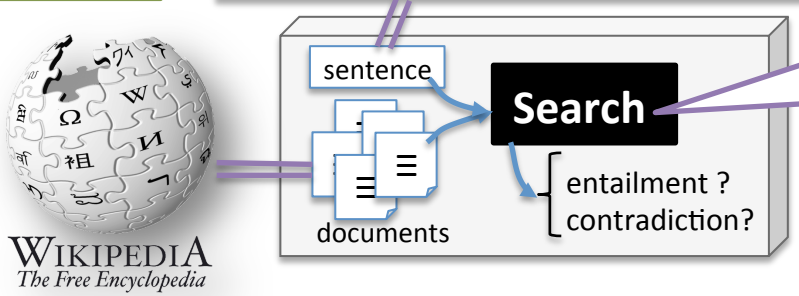
RITE1, RITE2 and RITE-VAL



Two subtasks of RITE-VAL

Fact Validation

t_2: *The Kamakura Shogunate began in Japan in the 12th century.*

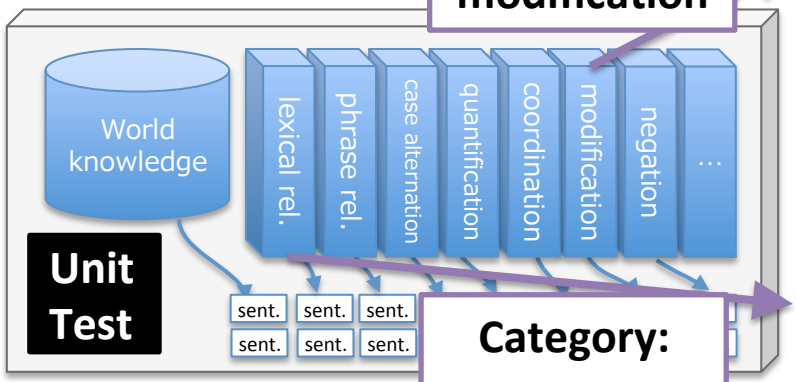


Search for evidence or counter-evidence for t_2.

Docs entail t_2.

Docs contradict t_2.

System Validation



Category: modification

t_1: *In the Meiji Constitution, legal clear distinction between the Imperial Family and Japan had been allowed.*

t_2: *In the Meiji Constitution, distinction between the Imperial Family and Japan had been allowed.*

Category: meronymy

t_1: *In the Meiji Constitution, distinction between the Imperial Family and Japan had been allowed.*

t_2: *In the Meiji Constitution, distinction between the Emperor and Japan had been allowed.*

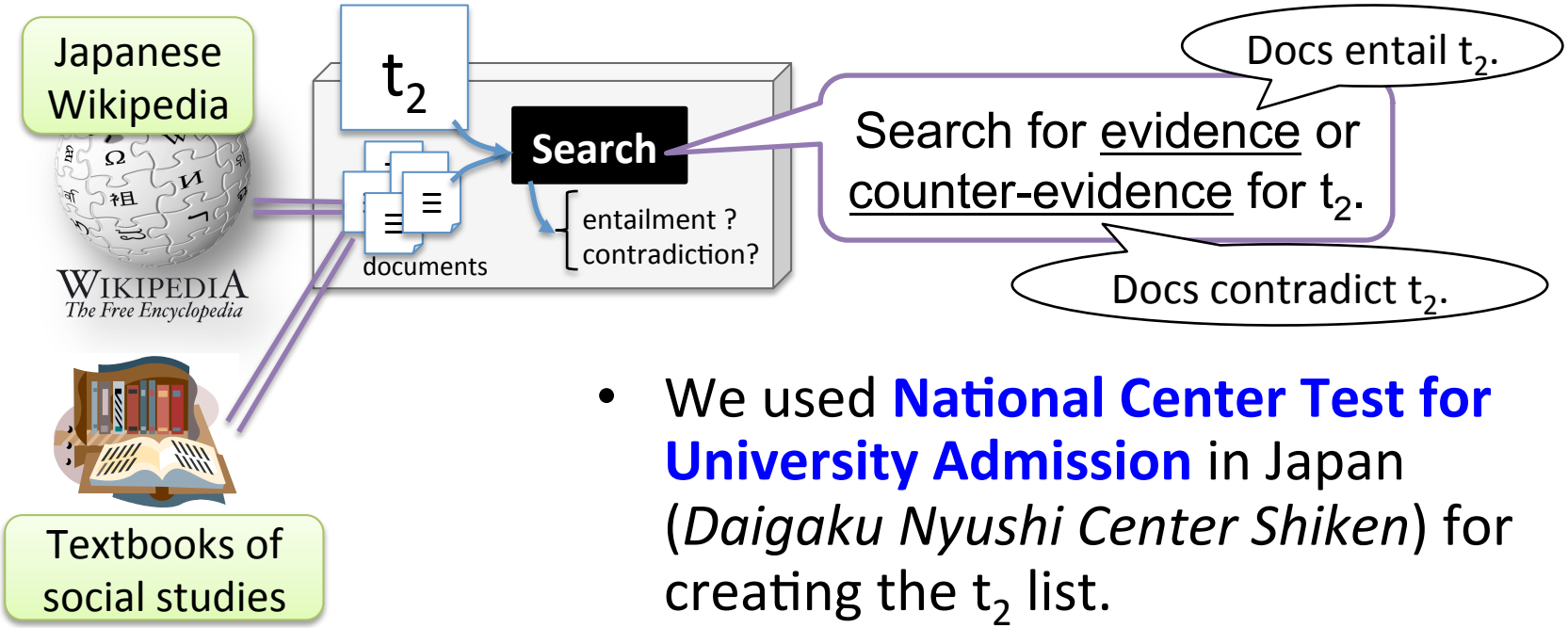
Subtask list of RITE-VAL

Subtask	Lang.	Task Description	Acronym	Test Data Size	Submissions
Fact Validation	CS	Multi-Classification	CS-FV	613	12
	CT	Multi-Classification	CT-FV	613	15
	JA	Binary Classification	JA-FV	514	30
		Passage Search	--	514	3
	EN	Binary Classification	EN-FV	188	9
System Validation	CS	Binary Classification	CS-SVBC	1,200	23
		Multi-Classification	CS-SVMC	1,200	18
	CT	Binary Classification	CT-SVBC	1,200	17
		Multi-Classification	CT-SVMC	1,200	17
	JA	Binary Classification	JA-SV	1,379	26
Total					170

Subtask list of RITE-VAL

Subtask	Lang.	Task Description	Acronym	Test Data Size	Submissions
Fact Validation	CS	Multi-Classification	CS-FV	613	12
	CT	Multi-Classification	CT-FV	613	15
	JA	**Binary Classification**	**JA-FV**	514	30
		Passage Search	--	514	3
	EN	Binary Classification	EN-FV	188	9
System Validation	CS	Binary Classification	CS-SVBC	1,200	23
		Multi-Classification	CS-SVMC	1,200	18
	CT	Binary Classification	CT-SVBC	1,200	17
		Multi-Classification	CT-SVMC	1,200	17
	JA	Binary Classification	JA-SV	1,379	26
Total					170

JA Fact Validation



	Y	N	Total
Training	383	575	958
Test	206	308	514
Total	589	883	1,472

- We used **National Center Test for University Admission** in Japan (*Daigaku Nyushi Center Shiken*) for creating the t_2 list.
- We chose **social studies**:
  - World history, Japanese history, Modern society, and Politics & Economics.
- Because **they mainly ask for knowledge of (historical) facts**.

Fact Validation

Development of JA-FV data

第１問　モニュメントや歴史的建造物について述べた次の文章Ａ～Ｃを読み，下の問い(問１～11)に答えよ。(配点　33)

Ａ　現在，アテネの中心部の丘にその偉容を誇る①パルテノン神殿は，古代ギリシアを象徴する歴史的建造物である。この神殿は，②オスマン帝国の支配下でモスクとして利用されたこともあったが，18 世紀には廃墟となっていた。1799 年にイギリスの大使としてイスタンブルに赴任したエルギン卿は，③ギリシアを訪れ，パルテノン神殿の遺跡から彫刻類を収集し，本国に送った。今日，大英博物館で「エルギン・マーブル」として展示されているものがそれである。1987 年，パルテノン神殿は，世界文化遺産として登録された。

問 3　下線部②の国について述べた文として最も適当なものを，次の①～④のうちから一つ選べ。 3

① スレイマン１世の時代が最盛期であった。
② 国教はシーア派のイスラーム教であった。
③ パルカン半島に誕生した後，小アジアへ進出した。
④ ベルリン会議により，ボスニア＝ヘルツェゴヴィナの統治権を得た。

t_2: オスマン帝国ではスレイマン1世の時代が最盛期であった.
(The Ottoman Empire's peak was during the reign of Suleiman I.)

The label is "Y."

スレイマン1世

WIKIPEDIA The Free Encyclopedia

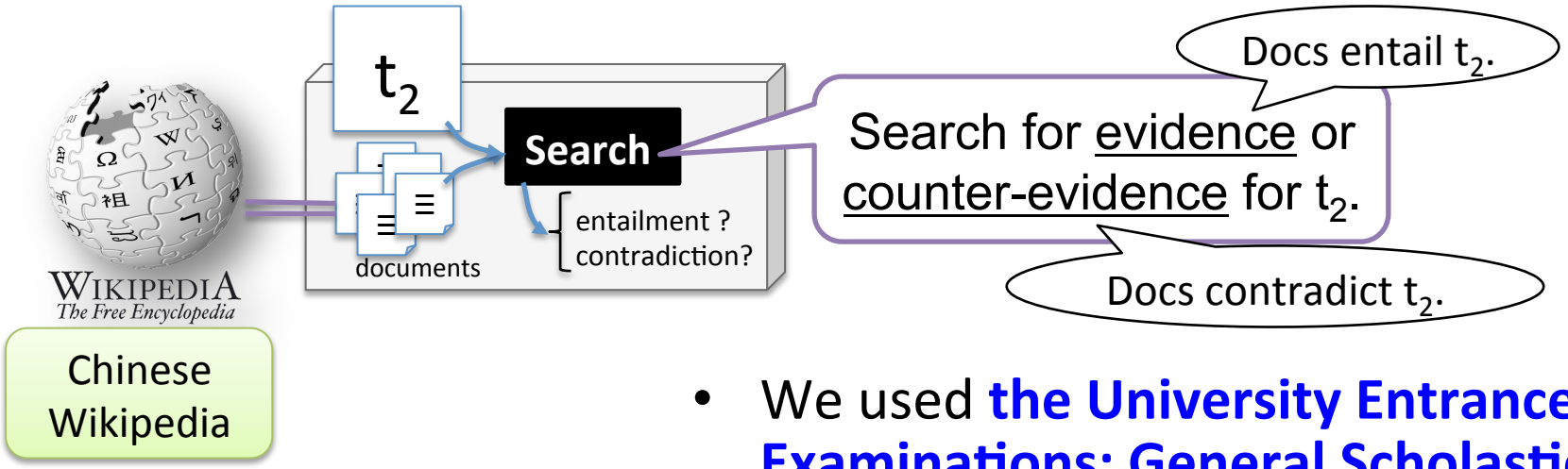
スルタン・スレイマン1世(**Kanuni Sultan Süleyman**、オスマン語 سلیمان Sulaymān, トルコ語 Süleyman、1494年11月6日 – 1566年9月5日)は、オスマン帝国の第10代皇帝(在位: 1520年 – 1566年)。

46年の長期にわたる在位の中で13回もの対外遠征を行い、数多くの軍事的成功を収めてオスマン帝国を最盛期に導いた。英語では、「**壮麗帝**(the Magnificent)」のあだ名で呼ばれ、日本ではしばしば「**スレイマン大帝**」と称される。トルコでは法典を編纂し帝国の制度を整備したことから「**立法帝**(カーヌーニー القانوني al-Qānūnī) /Kanuni)」のあだ名で知られている。

Subtask list of RITE-VAL

Subtask	Lang.	Task Description	Acronym	Test Data Size	Submissions
Fact Validation	CS	Multi-Classification	CS-FV	613	12
	CT	**Multi-Classification**	**CT-FV**	613	15
	JA	Binary Classification	JA-FV	514	30
		Passage Search	--	514	3
	EN	Binary Classification	EN-FV	188	9
System Validation	CS	Binary Classification	CS-SVBC	1,200	23
		Multi-Classification	CS-SVMC	1,200	18
	CT	Binary Classification	CT-SVBC	1,200	17
		Multi-Classification	CT-SVMC	1,200	17
	JA	Binary Classification	JA-SV	1,379	26
Total					170

CT Fact Validation



(Entailment, Contradiction or Unknown)

	E	C	U	Total
Training	243	84	162	489
Test	222	201	190	613
Total	465	285	352	1,102

- We used **the University Entrance Examinations: General Scholastic Ability Test** in Taiwan (GSAT) for creating the t_2 list.
- We chose **science and social studies**:
  - Physics, Chemistry, Biology and Geology
  - History, Geography and Civics

Subtask list of RITE-VAL

Subtask	Lang.	Task Description	Acronym	Test Data Size	Submissions
Fact Validation	**CS**	**Multi-Classification**	**CS-FV**	613	12
	CT	Multi-Classification	CT-FV	613	15
	JA	Binary Classification	JA-FV	514	30
		Passage Search	--	514	3
	EN	**Binary Classification**	**EN-FV**	188	9
System Validation	CS	Binary Classification	CS-SVBC	1,200	23
		Multi-Classification	CS-SVMC	1,200	18
	CT	Binary Classification	CT-SVBC	1,200	17
		Multi-Classification	CT-SVMC	1,200	17
	JA	Binary Classification	JA-SV	1,379	26
Total					170

Fact Validation

CT → CS and JA → EN

- The t_2 list of CT-FV was translated into Simplified Chinese.

Document: Chinese Wikipedia

	E	C	U	Total
Training	239	82	155	476
Test	222	201	190	613
Total	461	283	345	1,089

- A part of the t_2 list of JA-FV was translated into English.
  - Politics & Economics,
  - and part of World history

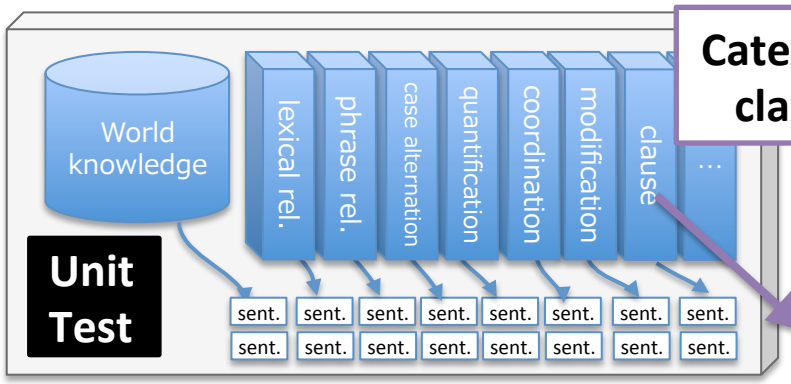
Document: English Wikipedia

	Y	N	Total
Training	141	238	379
Test	74	114	188
Total	215	352	567

Subtask list of RITE-VAL

Subtask	Lang.	Task Description	Acronym	Test Data Size	Submissions
Fact Validation	CS	Multi-Classification	CS-FV	613	12
	CT	Multi-Classification	CT-FV	613	15
	JA	Binary Classification	JA-FV	514	30
		Passage Search	--	514	3
	EN	Binary Classification	EN-FV	188	9
System Validation	CS	Binary Classification	CS-SVBC	1,200	23
		Multi-Classification	CS-SVMC	1,200	18
	CT	Binary Classification	CT-SVBC	1,200	17
		Multi-Classification	**CT-SVMC**	1,200	17
	JA	Binary Classification	JA-SV	1,379	26
Total					170

CT System Validation (MC)



t_1: 約瑟夫·傅立葉是十九世紀法國數學家、物理學家。
(Joseph Fourier was a French mathematician and physicist of the 19th century.)

t_2: 約瑟夫·傅立葉是物理學家。
(Joseph Fourier was a physicist.)

- Bidirectional entailment
- Forward entailment
- Contradiction
- Independence

	B	F	C	I	Total
Training	222	148	152	59	581
Test	300	300	300	300	1,200
Total	522	448	452	359	1,781

- We used about 100 pages from **Chinese Wikipedia** for extracting source sentences.
- Annotators carefully created pairs of t_1 and t_2 by **referring a list of linguistic phenomena** (shown in the next slide) **that are necessary for recognizing relations between them**.

Linguistic phenomena (CT-SVMC)

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

Linguistic Phenomenon	Train	Test
scrambling	27	35
spatial	18	42
synonymy (lexical)	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

Total: 1,200

(We revised the list presented by Sammons et al. (2010).)

Subtask list of RITE-VAL

Subtask	Lang.	Task Description	Acronym	Test Data Size	Submissions
Fact Validation	CS	Multi-Classification	CS-FV	613	12
	CT	Multi-Classification	CT-FV	613	15
	JA	Binary Classification	JA-FV	514	30
		Passage Search	--	514	3
	EN	Binary Classification	EN-FV	188	9
System Validation	**CS**	**Binary Classification**	**CS-SVBC**	1,200	23
		Multi-Classification	**CS-SVMC**	1,200	18
	CT	**Binary Classification**	**CT-SVBC**	1,200	17
		Multi-Classification	CT-SVMC	1,200	17
	JA	Binary Classification	JA-SV	1,379	26
Total					170

SVMC → SVBC and CT → CS

- The labels for the pairs of CT-SVMC were automatically converted into "Y" or "N."
  - **Bidirectional entailment**
  - **Forward entailment**
  - **Contradiction**
  - **Independence**

⇒ **Y**

⇒ **N**

	Y	N	Total
Training	370	211	581
Test	600	600	1,200
Total	970	811	1,781

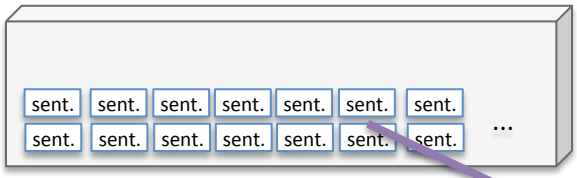
- The pair lists of CT-SVMC and CT-SVBC were translated into Simplified Chinese.
  - just like "CT-FV to CS-FV"

Subtask list of RITE-VAL

Subtask	Lang.	Task Description	Acronym	Test Data Size	Submissions
Fact Validation	CS	Multi-Classification	CS-FV	613	12
	CT	Multi-Classification	CT-FV	613	15
	JA	Binary Classification	JA-FV	514	30
		Passage Search	--	514	3
	EN	Binary Classification	EN-FV	188	9
System Validation	CS	Binary Classification	CS-SVBC	1,200	23
		Multi-Classification	CS-SVMC	1,200	18
	CT	Binary Classification	CT-SVBC	1,200	17
		Multi-Classification	CT-SVMC	1,200	17
	JA	**Binary Classification**	**JA-SV**	1,379	26
Total					170

System Validation

JA System Validation



No category label

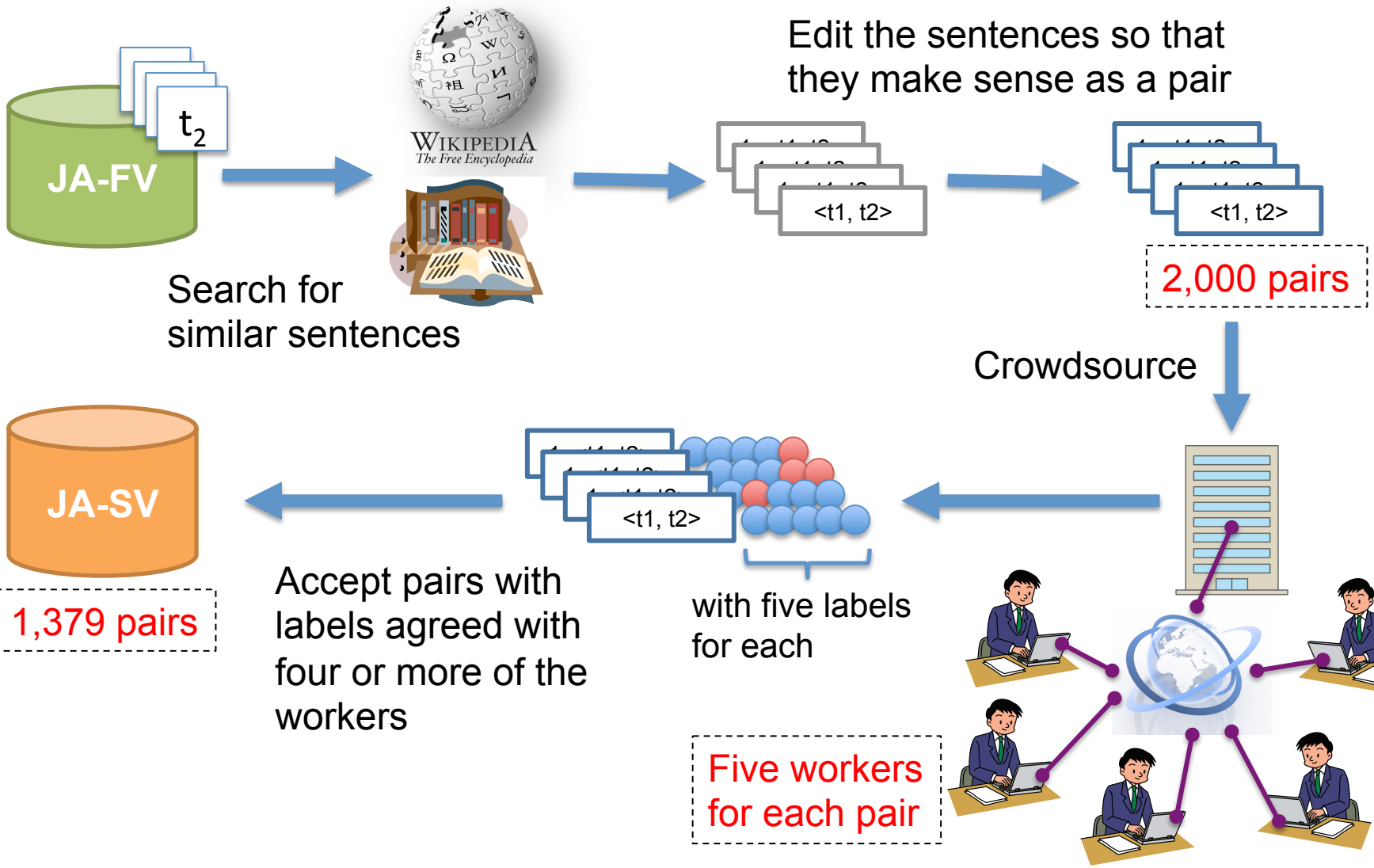
t_1: 鎌倉幕府は1192年に始まったとされていたが、最新の説では、実質的な成立は1185年とされている。
(The Kamakura Shogunate was considered to have begun in 1192, but the current leading theory is that it was effectively formed in 1185.)

t_2: 鎌倉幕府は12世紀に日本で開かれた。
(The Kamakura Shogunate began in Japan in the 12th century.)

	Y	N	Total
Training	1,330	1,362	2,692
Test	339	1,040	1,379
Total	1,669	2,402	4,071

- We are very sorry that we have provided JA-SV data with no category labels.
  - due to time constraint and lack of know-how of crowdsourcing
- We attempted to use crowdsourcing for constructing JA-SV data.

Development of the data



Subtask list of RITE-VAL

Subtask	Lang.	Task Description	Acronym	Test Data Size	Submissions
Fact Validation	CS	Multi-Classification	CS-FV	613	12
	CT	Multi-Classification	CT-FV	613	15
	JA	Binary Classification	JA-FV	514	30
		Passage Search	--	514	3
	EN	Binary Classification	EN-FV	188	9
System Validation	CS	Binary Classification	CS-SVBC	1,200	23
		Multi-Classification	CS-SVMC	1,200	18
	CT	Binary Classification	CT-SVBC	1,200	17
		Multi-Classification	CT-SVMC	1,200	17
	JA	Binary Classification	JA-SV	1,379	26
Total					170

Evaluation metrics

- Macro F1 and Accuracy

$$macroF1 = \frac{1}{|Class|} \sum_{c \in Class} \frac{2P_c R_c}{P_c + R_c} \qquad P_c = \frac{N_{\text{correct},\, c}}{N_{\text{predicted},\, c}} \qquad R_c = \frac{N_{\text{correct},\, c}}{N_{\text{target},\, c}}$$

$$Acc. = \frac{N_{\text{correct}}}{N_{\text{example}}}$$

- Correct Answer Ratio (as Entrance Exam)
  - evaluation for multiple-choice questions.
  - Y/N labels are mapped into selections of answers and calculate accuracy of the answers.

Evaluators

- We provided evaluators for the participants.
  - Java scripts

```
$ java -jar rite2eval.jar -g RITEVAL_JA_test.xml -s output_sv.txt
-------------------------------------------------------------
|Label|    #|          Precision|             Recall|    F1|
|    N|  354| 60.18(  204/  339)| 57.63(  204/  354)| 58.87|
|    Y|  256| 44.65(  121/  271)| 47.27(  121/  256)| 45.92|
-------------------------------------------------------------
Accuracy: 53.28(   325/   610)
Macro F1: 52.40

Confusion Matrix
--------------------
|gold \ sys|   N   Y|
--------------------
|         N| 204 150|
|         Y| 135 121|
--------------------
```

RITE-VAL Formal Run Participation

Active participating teams

		Submissions
Fact Validation	CS-FV	12
	CT-FV	15
	JA-FV	30
	FVsearch	3
	EN-FV	9
System Validation	CS-SVBC	23
	CS-SVMC	18
	CT-SVBC	17
	CT-SVMC	17
	JA-SV	26
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.)

NTCIR-10 RITE2	28
NTCIR-9 RITE1	24

Submitted runs

		Submissions
Fact Validation	CS-FV	12
	CT-FV	15
	JA-FV	30
	FVsearch	3
	EN-FV	9
System Validation	CS-SVBC	23
	CS-SVMC	18
	CT-SVBC	17
	CT-SVMC	17
	JA-SV	26
Total		**170**

NTCIR-10 RITE2	JA	CT	CS	Total
BC	41	20	21	82
MC	20	21	21	62
Exam BC	31	-	-	31
Exam Search	4	-	-	4
UnitTest	14	-	-	14
RITE4QA	-	12	10	22
Total	110	53	52	**215**
NTCIR-9 RITE1	**JA**	**CT**	**CS**	**Total**
Total	65	70	77	**212**

Formal Run Results

Summary of formal run

Subtask		Size	Submit	Top m-F1	Top Acc.	Top Team (Run)
Fact Validation	CS-FV	613	12	38.93	44.05	III&CYUT (05)
	CT-FV	613	15	39.51	44.70	III&CYUT (02)
	JA-FV	514	30	61.93	63.23	NUL (03)
	EN-FV	188	9	53.17	55.85	BnO (01)
System Validation	CS-SVBC	1,200	23	61.51	62.33	BUPTTeam (05)
	CS-SVMC	1,200	18	44.39	51.83	WUST (01)
	CT-SVBC	1,200	17	56.24	56.25	III&CYUT (04)
	CT-SVMC	1,200	17	40.54	43.33	III&CYUT (05)
	JA-SV	1,379	26	69.59	77.81	NUL (04)

Summary of formal run

Subtask		Size	Submit	Top m-F1	Top Acc.	Top Team (Run)
F… V…			12	38.93	44.05	III&CYUT (05)
			15	39.51	44.70	III&CYUT (02)
			30	61.93	63.23	NUL (03)
				53.17	55.85	BnO (01)
Syste… Validation	CS-SVBC	1,200	23	61.51	62.33	BUPTTeam (05)
	CS-SVMC	1,200	18	44.39	51.83	WUST (01)
	CT-SVBC	1,200	17	56.24	56.25	III&CYUT (04)
	CT-SVMC	1,200	17	40.54	43.33	III&CYUT (05)
	JA-SV	1,379	26	69.59	77.81	NUL (04)

Top Macro F1 for BC subtasks are 53% to 70%. (cf. A random method achieves 40% to 50%.)

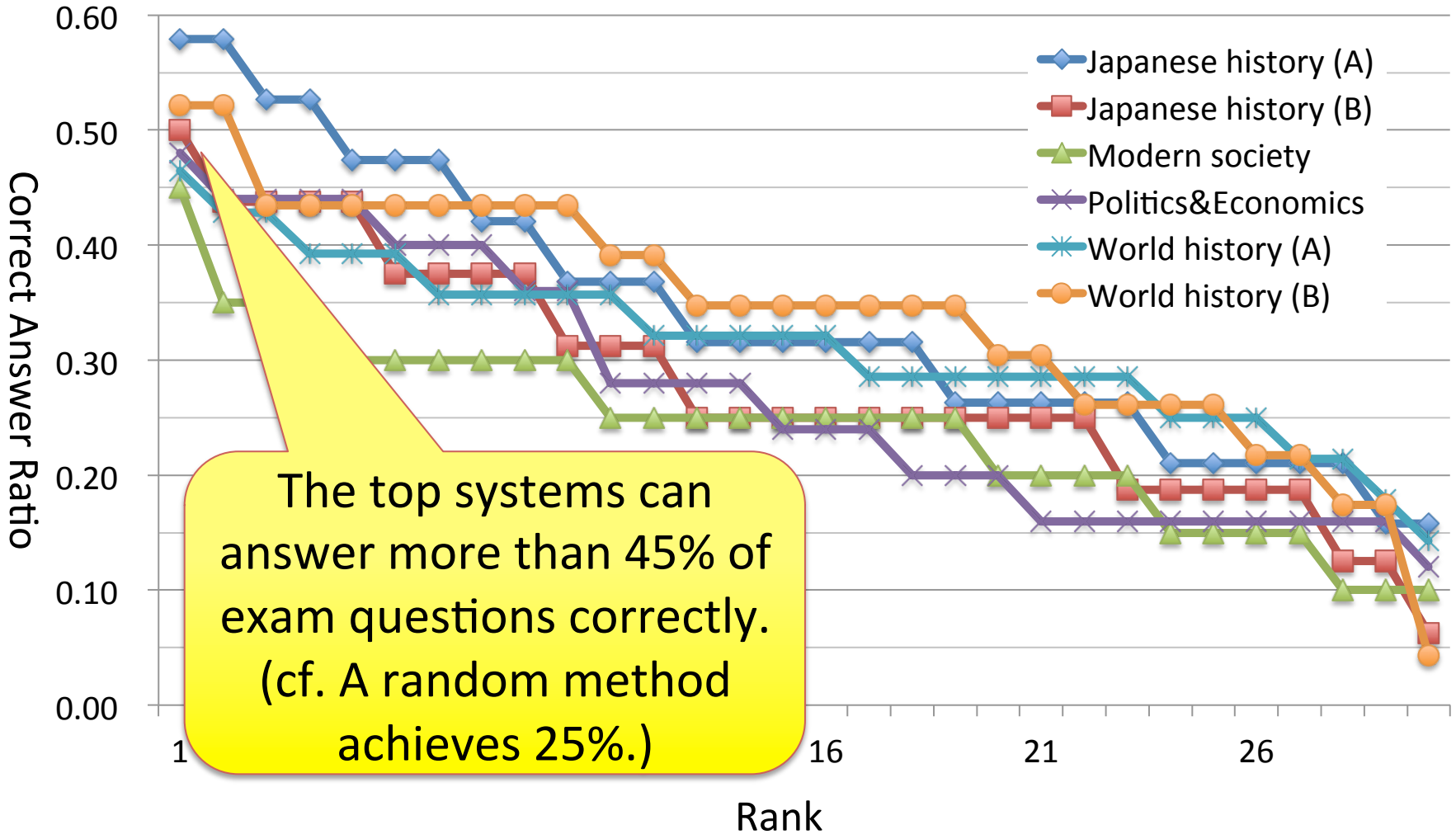
Summary of formal run

Subtask		Size	Submit	Top m-F1	Top Acc.	Top Team (Run)
Fact Validation	CS-FV	613	12	38.93	44.05	III&CYUT (05)
	CT-FV	613	15	39.51	44.70	III&CYUT (02)
	JA-FV	514	30	61.93	63.23	NUL (03)
	EN-FV	188	9	53.17	55.85	BnO (01)
			23	61.51	62.33	BUPTTeam (05)
				44.39	51.83	WUST (01)
			17	56.24	56.25	III&CYUT (04)
			17	40.54	43.33	III&CYUT (05)
	JA-SV	1,379	26	69.59	77.81	NUL (04)

Top Macro F1 for MC subtasks are 40% to 45%. (cf. A random method achieves 25% to 33%.)

Recognizing textual entailment is still a difficult task for computers.

Evaluation for multiple-choice questions (JA-FV)



Review of the Participants' Systems

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%**)

(# of runs)

Statistical approaches:

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

Feature

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

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

(# of runs)

Feature / Information	CS	CT	JA	EN	Total	%
alignment	6	15	17	0	38	22
char/word overlapping	**45**	**48**	**51**	**8**	**152**	**89**
entailment rule	21	18	5	7	51	30
entity/event	7	17	1	0	25	15
hypernym	22	35	26	**9**	92	54
meronym	8	16	3	2	29	17
modality	0	2	4	0	6	4
named entity	24	27	35	**9**	95	59
overlapping	**45**	**46**	34	**9**	134	79
polarity	11	13	2	7	33	19
predicate argument relationship	16	15	8	**9**	48	28
synonym/antonym	41	**44**	33	**9**	127	75
syntactic information	25	17	14	**9**	65	38
temporal/numeric information	**43**	**47**	27	7	124	73
transformation	9	27	3	2	41	24
(number of the submitted runs)	53	49	59	9	170	100

Oral presentations [**Dec 11 9:20-**]

- 9:20 - 12:35 Parallel Session A-1:
  - III&CYUT Chinese Textual Entailment Recognition System for NTCIR-11 RITE-VAL
    Shih-Hung Wu, Li-Jen Hsu, Hua-Wei Lin, Pei-Kai Liao (Chaoyang University of Technology, Taiwan), Liang-Pu Chen and Tsun Ku (Institute for Information Industry, Taiwan)
  - Recognizing Textual Entailment Using Multiple Features and Filters
    Yongmei Tan, Minda Wang and Xiaohui Wang (Beijing Univesty of Posts and Telecommunications, China)
  - WUST at NTCIR-11 RITE-VAL System Validation Task
    Maofu Liu, Yue Wang and Limin Wang (Wuhan University of Science and Technology, China)
  - NUL System at NTCIR RITE-VAL tasks
    Chikara Hoshino, Ai Ishii, Hiroshi Miyashita and Mio Kobayashi (Nihon Unisys, Ltd., Japan)
  - KSU Team's System and Experience at the NTCIR-11 RITE-VAL Task
    Tasuku Kimura and Hisashi Miyamori (Kyoto Sangyo University, Japan)
  - A Surface-Similarity Based Two-Step Classifier for RITE-VAL
    Shohei Hattori and Satoshi Sato (Nagoya University, Japan)

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.

Thank you for your attention!