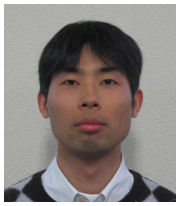
RITE-2
Recognizing Inference in TExt@NTCIR10

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

Yotaro Watanabe
Tohoku University

Yusuke Miyao
NII

Junta Mizuno
Tohoku University



Tomohide Shibata
Kyoto University

Hiroshi Kanayama
IBM Research

Cheng-Wei Lee
Academia Sinica

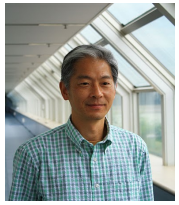
Chuan-Jie Lin
National Taiwan Ocean University

Shuming Shi
MSRA

Teruko Mitamura
CMU

Noriko Kando
NII

Hideki Shima
CMU



Kohichi Takeda
IBM Research

Overview of RITE-2

- **RITE-2 is a generic benchmark task that addresses a common semantic inference required in various NLP/IA applications**

t_1: *The Kamakura Shogunate was considered to have begun in 1192, but the current leading theory is that it was effectively formed in 1185.*

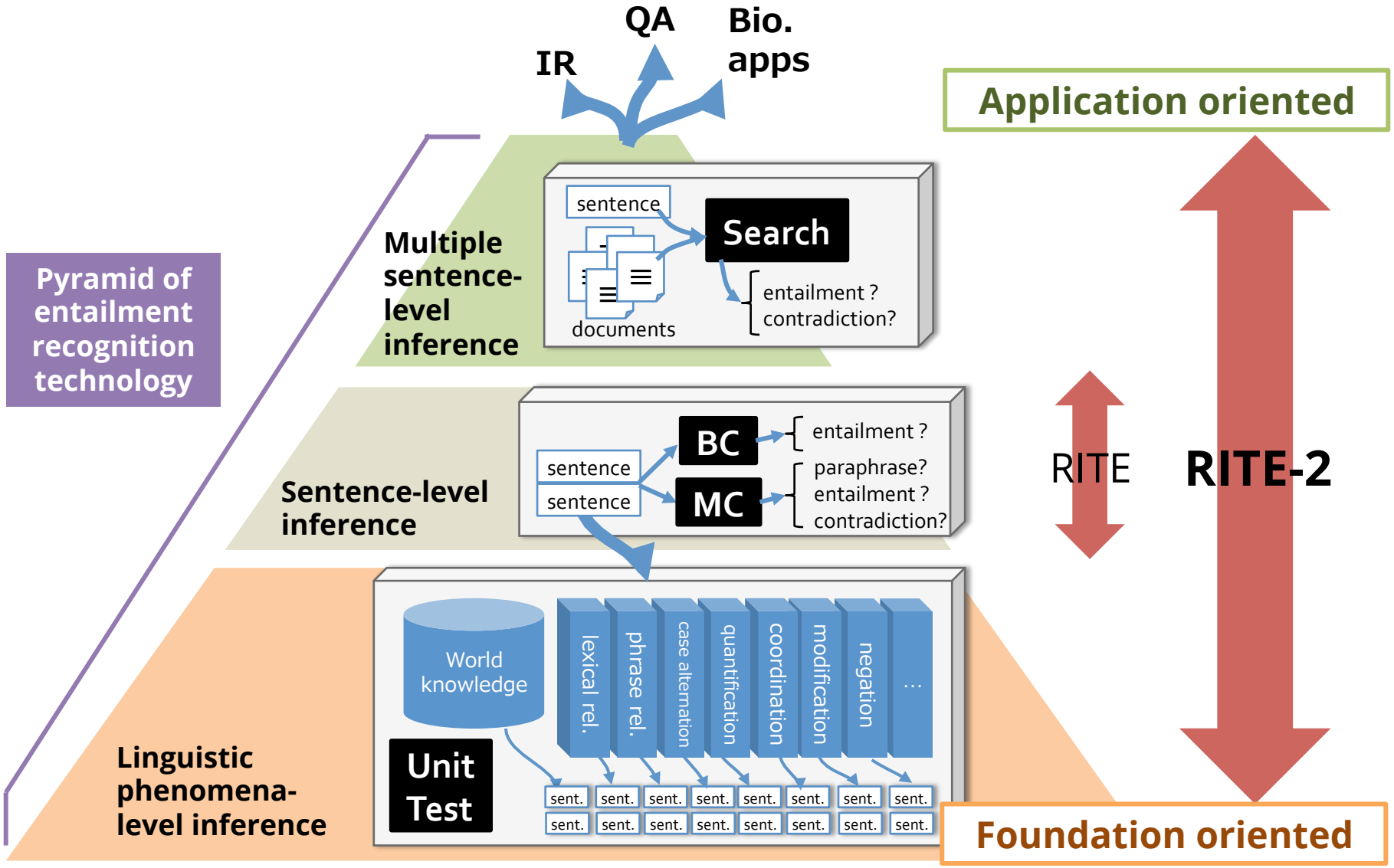
Can t_2 be inferred from t_1 ? (entailment?)

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

Motivation

- **Natural Language Processing (NLP) / Information Access (IA) applications**
  - Question Answering, Information Retrieval, Information Extraction, Text Summarization, Automatic evaluation for Machine Translation, Complex Question Answering
- **The current entailment recognition systems have not been mature enough**
  - The highest accuracy on Japanese BC subtask in NTCIR-9 RITE was only **58%**
  - There is still enough room to address the task to advance entailment recognition technologies

RITE vs. RITE-2

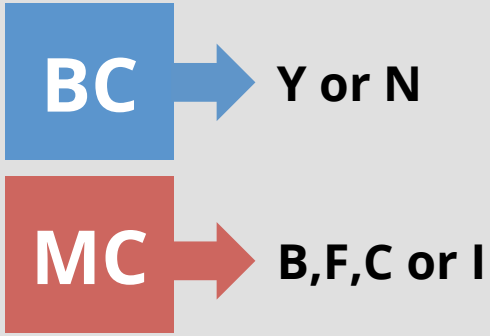


RITE-2 Subtasks

BC and MC subtasks

t_1: *The Kamakura Shogunate was considered to have begun in 1192, but the current leading theory is that it was effectively formed in 1185.*

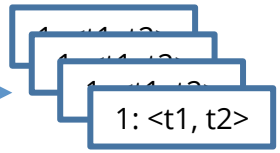
t_2: *The Kamakura Shogunate began in Japan in the 12th century.*



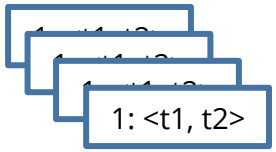
- **BC subtask**
  - Entailment (t_1 entails t_2) or Non-Entailment (otherwise)
- **MC subtask**
  - Bi-directional Entailment (t_1 entails t_2 & t_2 entails t_1)
  - Forward Entailment (t_1 entails t_2 & t_2 does not entail t_1)
  - Contradiction (t_1 contradicts t_2 or cannot be true at the same time)
  - Independence (otherwise)

Development of BC and MC data

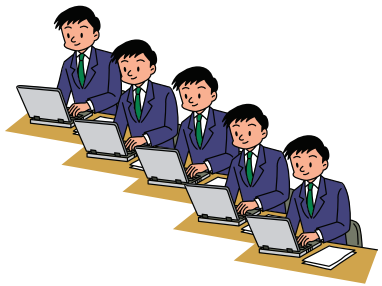
retrieve pairs of sentences

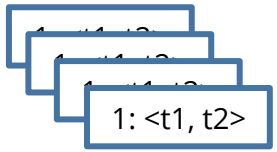


edit pairs if needed

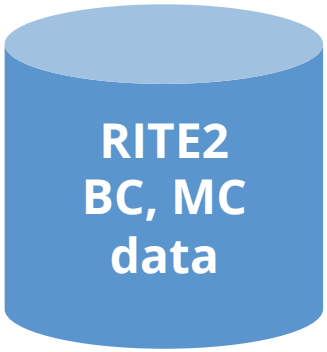


for each example, 5 annotators assigned its semantic label





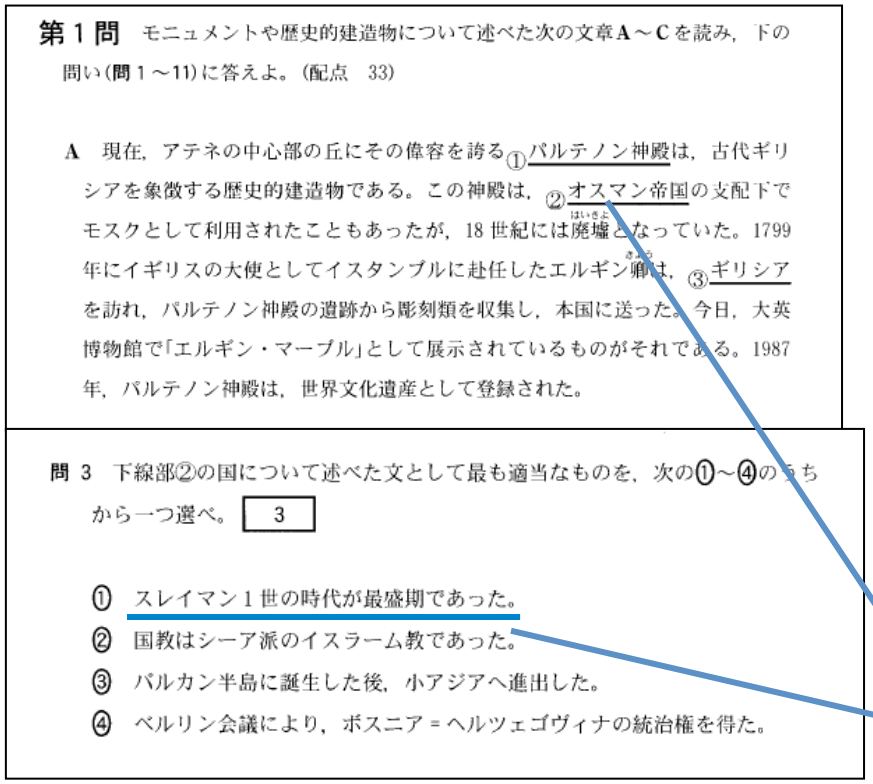
accept an example if 4 or more annotators assigned the same label to the example



Entrance Exam subtasks (Japanese only)

Entrance exam problem

National Center Test for University Admission
(*Daigaku Nyushi Center Shiken*)

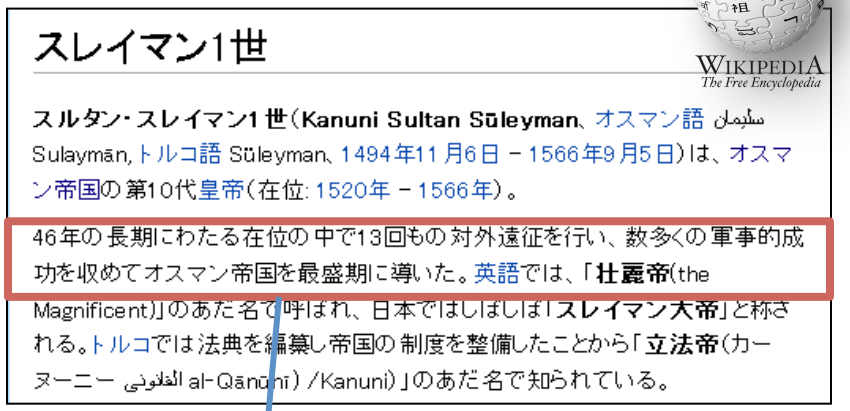


第１問　モニュメントや歴史的建造物について述べた次の文章Ａ～Ｃを読み，下の問い(問１～11)に答えよ。(配点　33)

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(はいきょ)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(きょう)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 バ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。



スレイマン1世

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 القانونی al-Qānūnī) /Kanuni)」のあだ名で知られている。

t_1: スレイマン1世は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．(Suleiman I contributed in a lot of military successes and led the Ottoman Empire to its peak.

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).

Entrance Exam subtask: BC and Search

- **Entrance Exam BC**
  - Binary-classification problem (Entailment or Non-entailment)
  - t1 and t2 are given
- **Entrance Exam Search**
  - Binary-classification problem (Entailment or Non-entailment)
  - t2 and a set of documents are given
    - Systems are required to search sentences in Wikipedia and textbooks to decide semantic labels

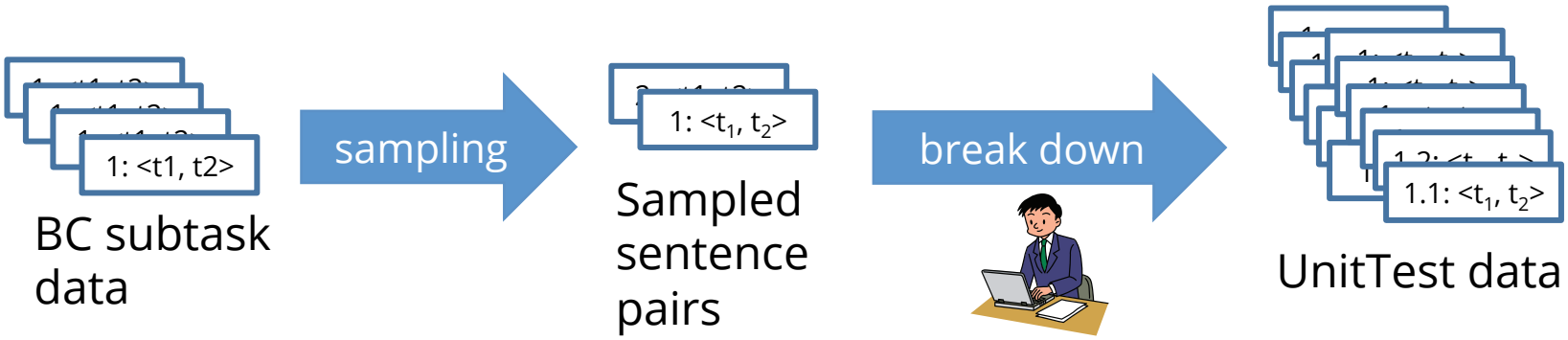
UnitTest (Japanese only)

- **Motivation**
  - Evaluate how systems can handle linguistic phenomena that affects entailment relations
- **Task definition**
  - Binary classification problem (same as BC subtask)

t_1: In the Meiji Constitution, legal clear distinction between the Imperial Family and Japan had been allowed. ⬇ **Category: modifier**
t_2: In the Meiji Constitution, distinction between the Imperial Family and Japan had been allowed.

t_1: In the Meiji Constitution, distinction between the Imperial Family and Japan had been allowed. ⬇ **Category: melonymy**
t_2: In the Meiji Constitution, distinction between the Emperor and Japan had been allowed

Development of the UnitTest data



- **Procedure**
  - Sentence pairs {<t_1, t_2>} were sampled from the BC subtask data
  - An annotator transformed each sampled sentence pair from t1 to t2 by breaking down the pair in a set of linguistic phenomena
- **[Kaneko+ 13] (to appear in ACL 2013)**

Distribution of the linguistic phenomena in UnitTest data

		dev	test
lexical	synonymy	10	10
	hypernymy	6	3
	meronymy	1	1
	entailment	1	0
phrase	synonymy	45	35
	hypernymy	3	0
	entailment	28	45
case alternation		9	7
modifier		30	42
nominalization		2	1
coreference		12	4
clause		29	14
relative clause		10	8
transparent head		2	1

		dev	test
list		11	3
quantity		1	0
scrambling		16	15
inference		4	2
Implicit relation		10	18
apposition		3	1
temporal		2	1
spatial		4	1
disagree	lexical	5	2
	phrase	25	25
	modality	2	1
	spatial	1	1
	temporal	0	1
Total		**272**	**241**

RITE4QA (Chinese only)

- **Motivation**
  - Can an entailment recognition system rank a set of unordered answer candidates in QA?
- **Dataset**
  - Developed from NTCIR-7 and NTCIR-8 CLQA data
    - t1: answer-candidate-bearing sentence
    - t2: a question in an affirmative form
- **Requirements**
  - Generate confidence scores for ranking process

Evaluation Metrics

- **Macro F1 and Accuracy (BC, MC, ExamBC, ExamSearch and UnitTest)**

$$MacroF1 = \frac{1}{|C|}\sum_{c \in \mathcal{C}} F1_c \qquad Accuracy = 100 \times \frac{N_{correct}}{N_{examples}}$$

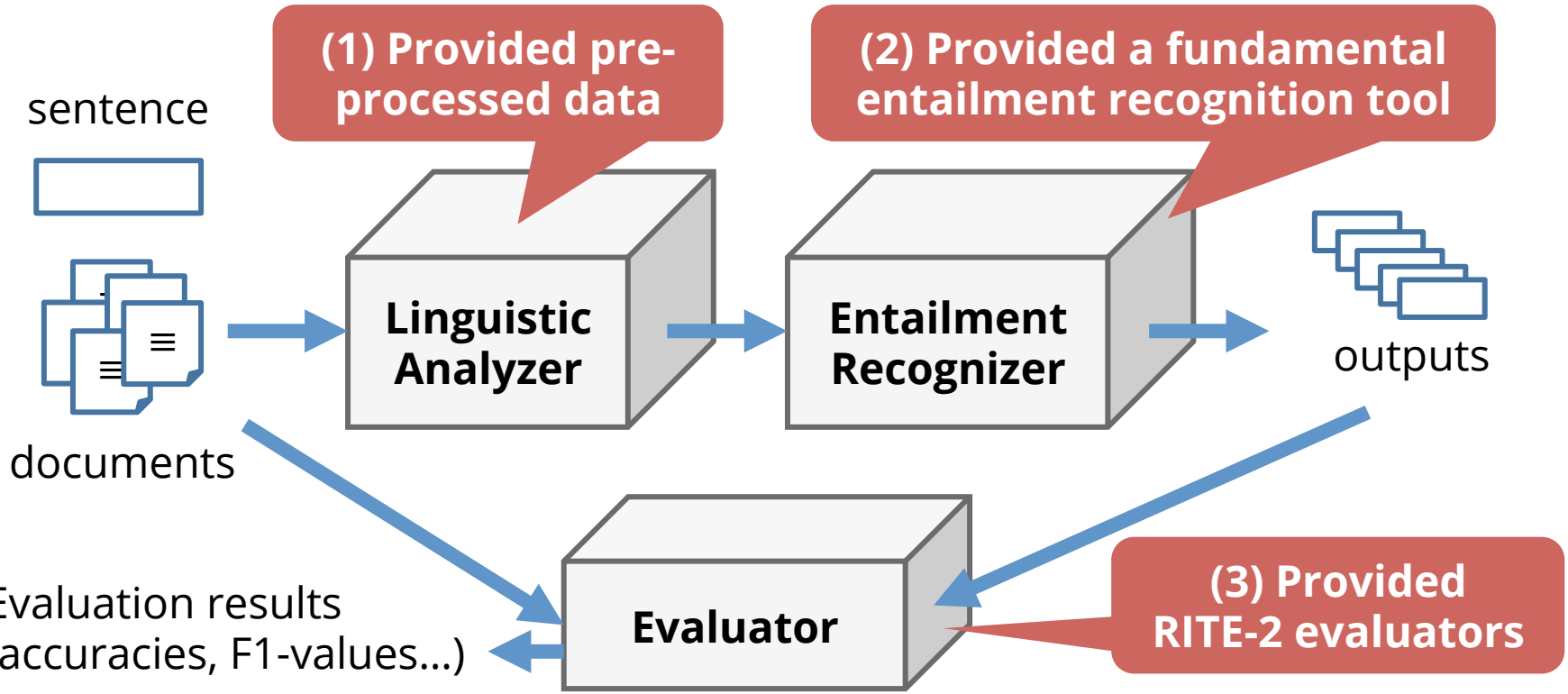
- **Correct Answer Ratio (Entrance Exam)**
  - Y/N labels are mapped into selections of answers and calculate accuracy of the answers
- **Top1 and MRR (RITE4QA)**

$$Top1 = \frac{1}{|Q|}\sum_{i=1}^{|Q|}[\text{top answer is correct}] \qquad MRR = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{rank_i}$$

Organization Effort

Generic Framework

- **We provided pre-processed data and tools to lower barriers to entry**



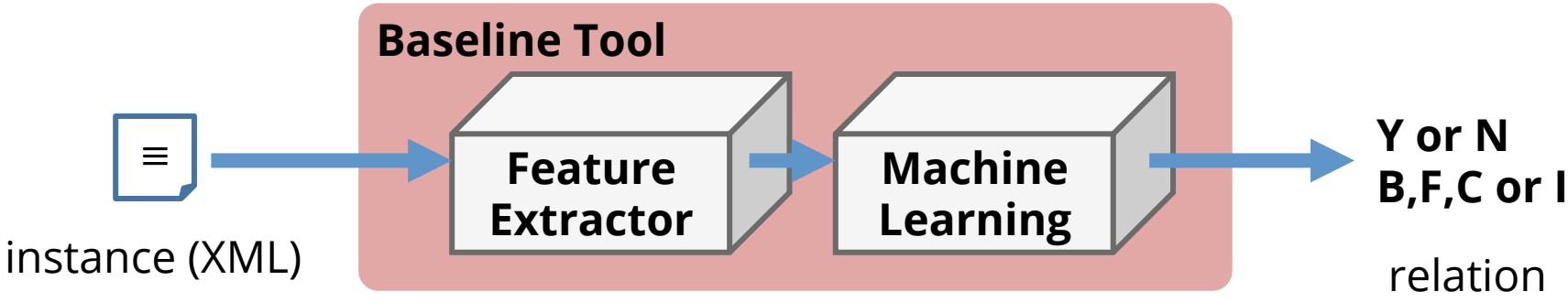
(1) Pre-processed data

- **Morphological and syntactic analysis**
  - MeCab [Kudo+ 05] + CaboCha [Kudo+ 02]
  - Juman + KNP
  - Provided as XML data

```
<?xml version='1.0' encoding='UTF-8' standalone='no' ?>
<dataset type='bc'>
  <pair id='1' label='Y'>
    <t1>
      川端康成は「雪国」などの作品でノーベル文学賞を受賞した。
      <Sentence id="sample_t1" role="text" text="川端康成は「雪
        <Annotation tool="MeCab" ver="0.994"/>
        <Annotation tool="CaboCha" ver="0.64"/>
        <Annotation tool="UniDic" ver="1.3.12"/>
        <Chunk head="c4" id="c0" score="2.473067" type="D">
```

- **Search Results for Exam Search subtask**
  - Used TSUBAKI [Shinzato+ 11] to provide search results
  - Provided at most five search results extracted from Wikipedia and textbooks

(2) A fundamental entailment recognition tool (Baseline tool)



- **Features**
  - a machine learning-based entailment recognition system
  - simple features are implemented (Feature Extractor)
    - Bag-of- {content words, aligned chunks, head words}
    - Ratio of aligned {content words, aligned chunks}
  - new features can be easily added
  - outputs files compatible with the format of the RITE-2 formal run

(3) RITE-2 Evaluators

- **Generic Evaluator (all of the subtasks)**

```
$ java -jar rite2eval.jar -g RITE2_JA_test_bc.xml -s output_bc.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|
--------------------
```

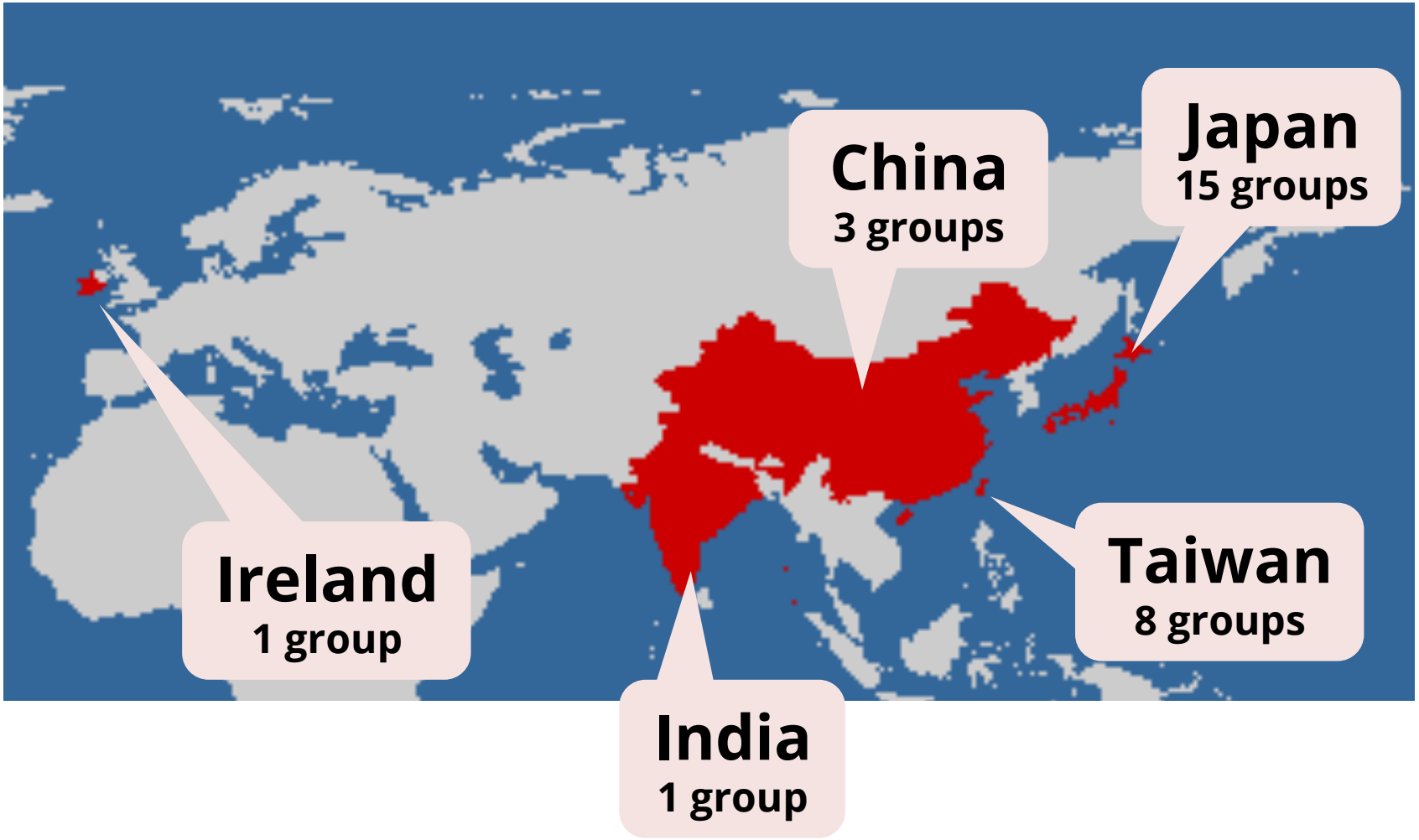
- **Additional Evaluator (Entrance Exam)**
  - Calculate correct answer ratio

RITE-2 Formal Run Participation

Number of submissions

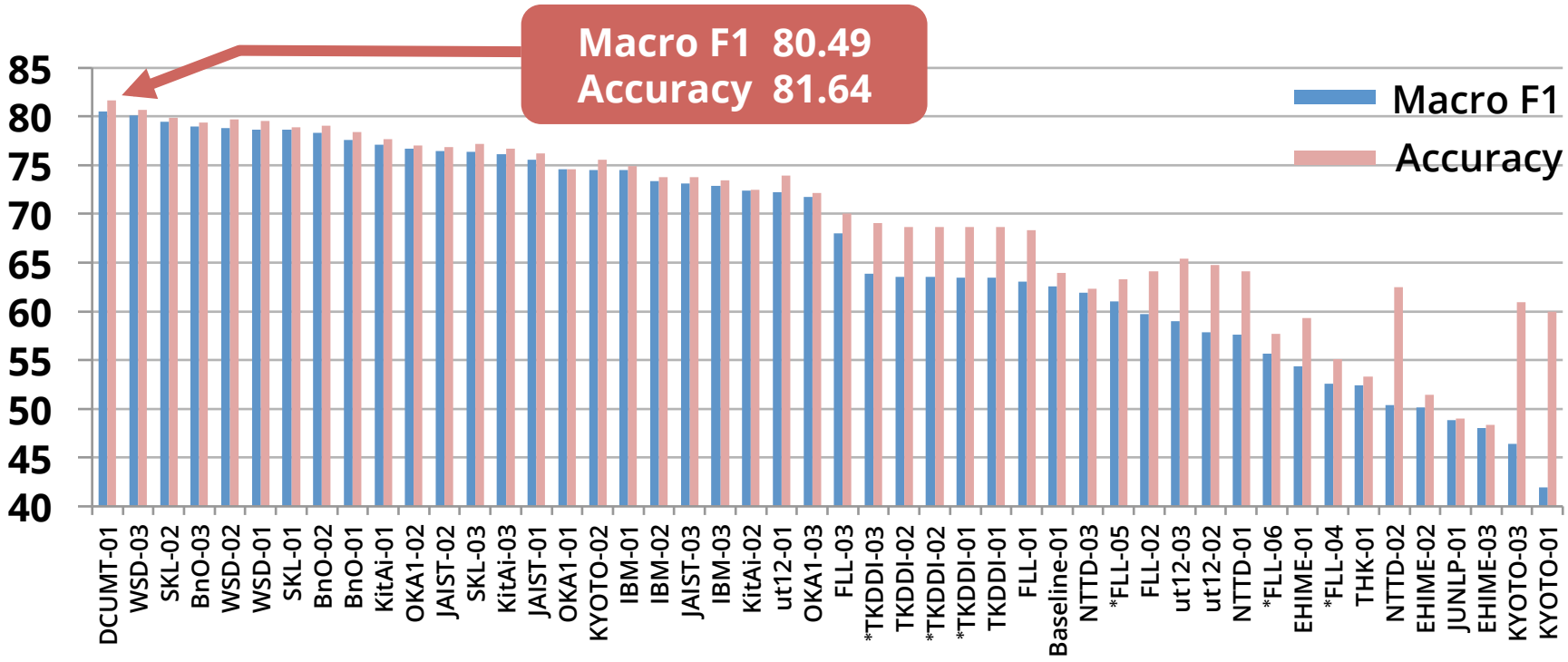
NTCIR-10 RITE-2	JA	CT	CS	Total
BC	41	20	21	82
MC	20	21	21	62
Exam BC	31	-	-	31
Exam Search	4	-	-	4
UnitTest	14	-	-	14
RITE4QA	-	12	10	22
Total	110	53	52	215
NTCIR-9 RITE	**JA**	**CT**	**CS**	**Total**
Total	65	70	77	212

Countries/Regions of Participants



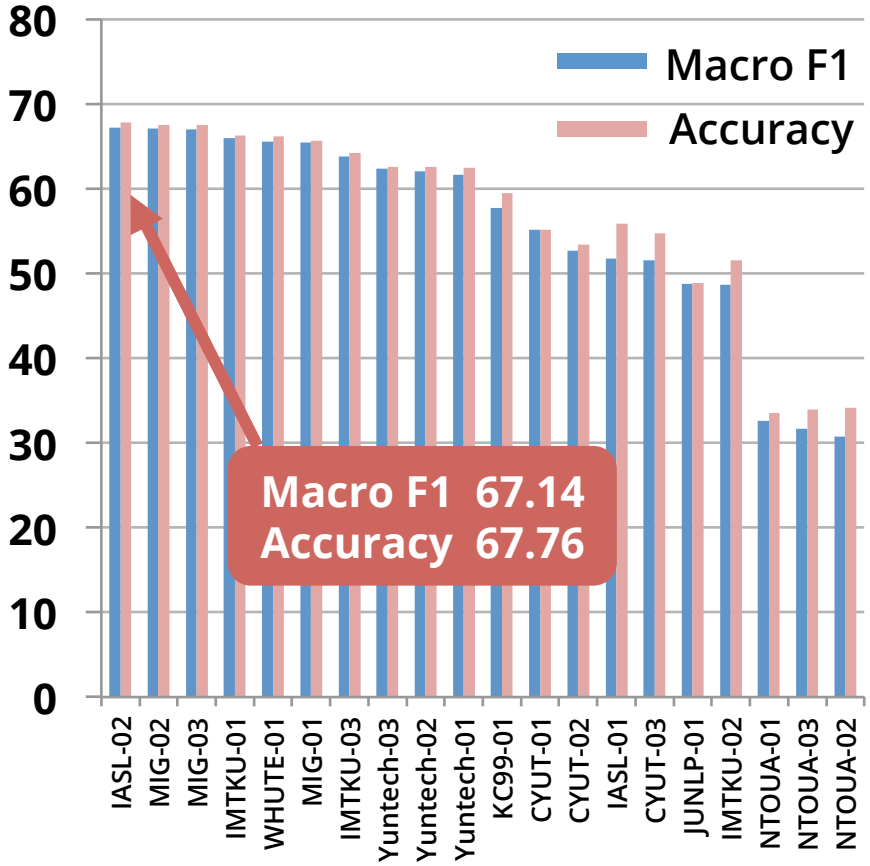
Formal Run Results

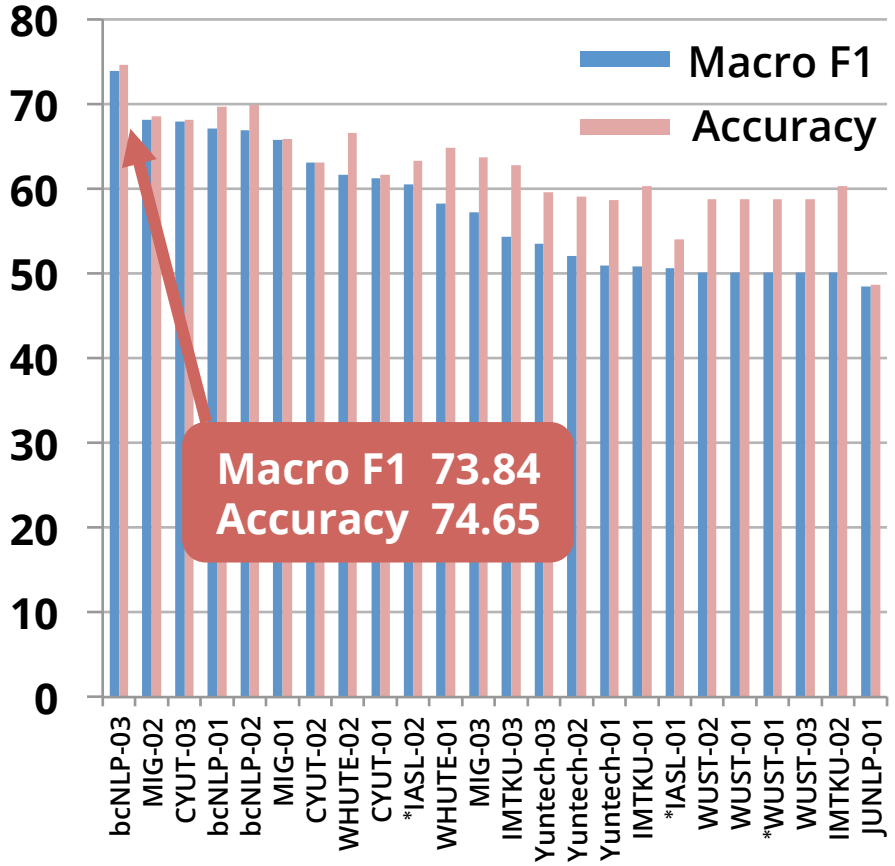
BC (Japanese)



- **The best system achieved over 80% of accuracy (The highest score in BC subtask at RITE was 58%)**
- **The difference is caused by**
  - Advancement of entailment recognition technologies
  - Strict data filtering in the data development

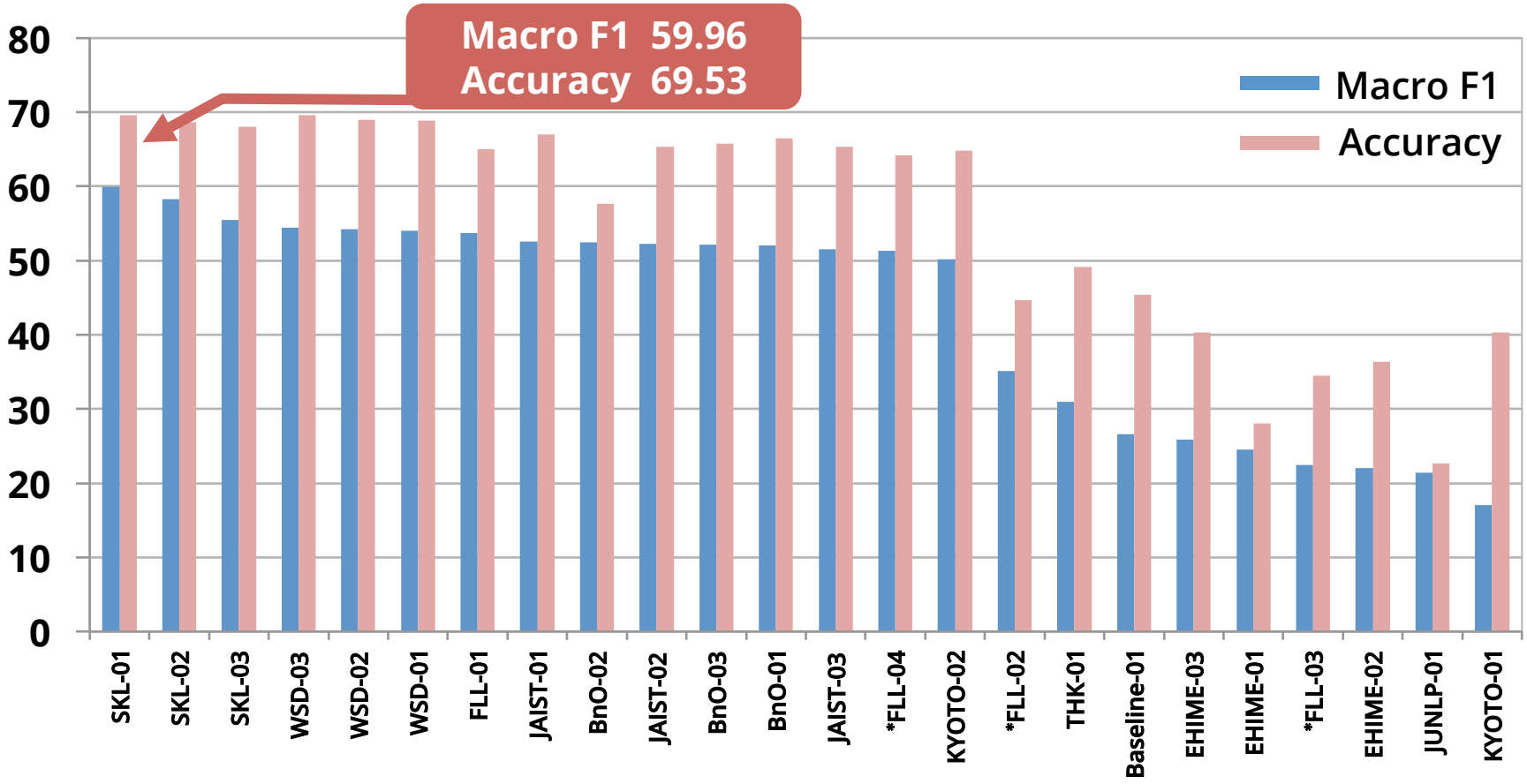
BC (Traditional/Simplified Chinese)





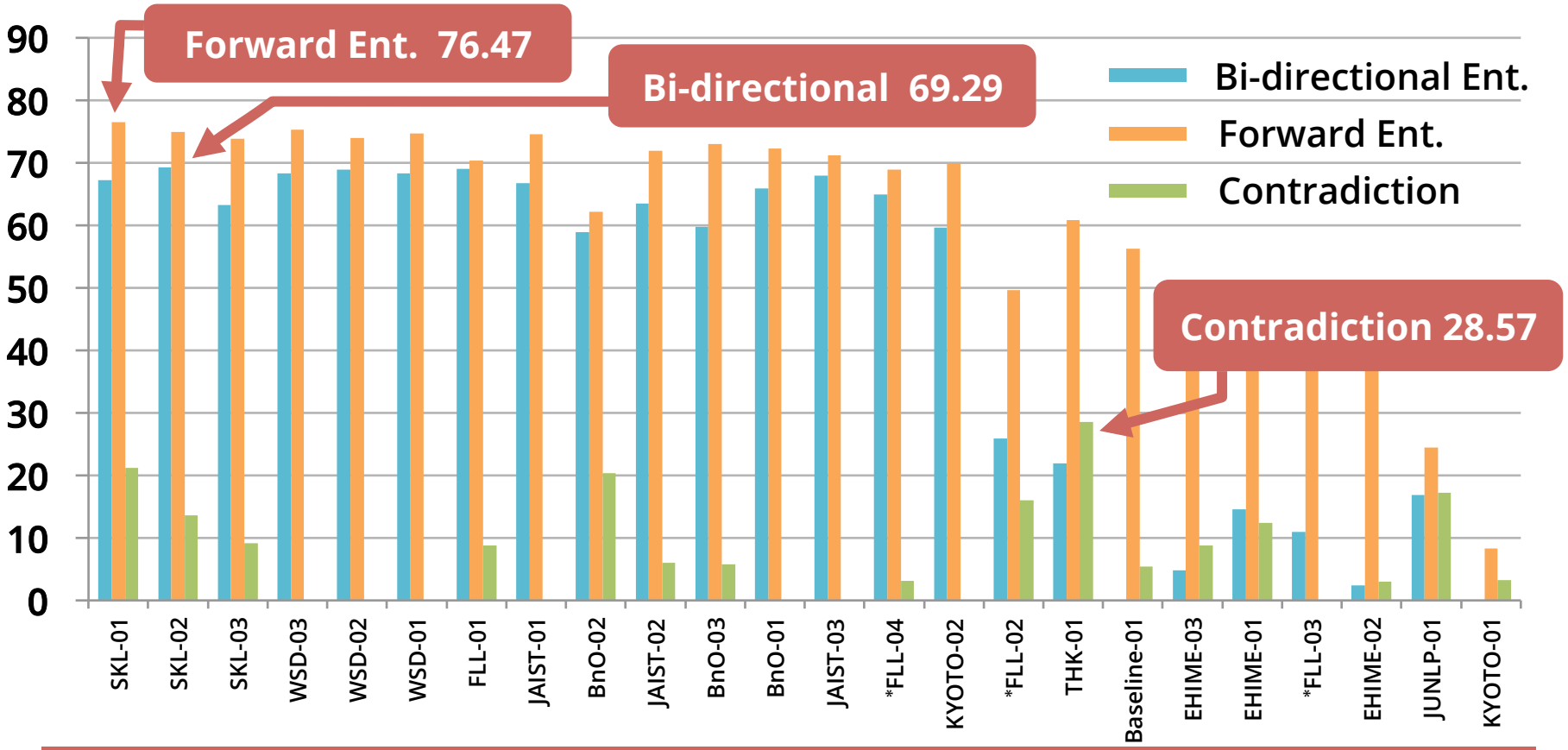
- **The top scores are almost the same as those in NTCIR-9 RITE**

MC (Japanese)



- **The top system achieved approx. 70% of accuracy (The highest acc. in NTCIR-9 RITE was only 51%)**

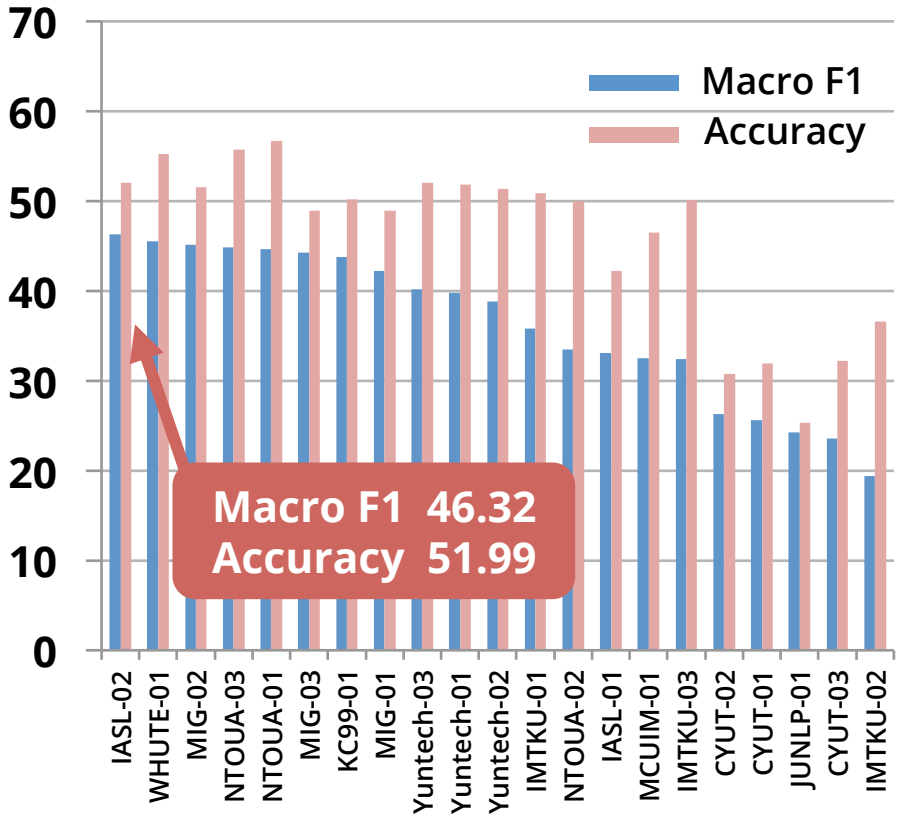
MC (Japanese, F1 for each label)

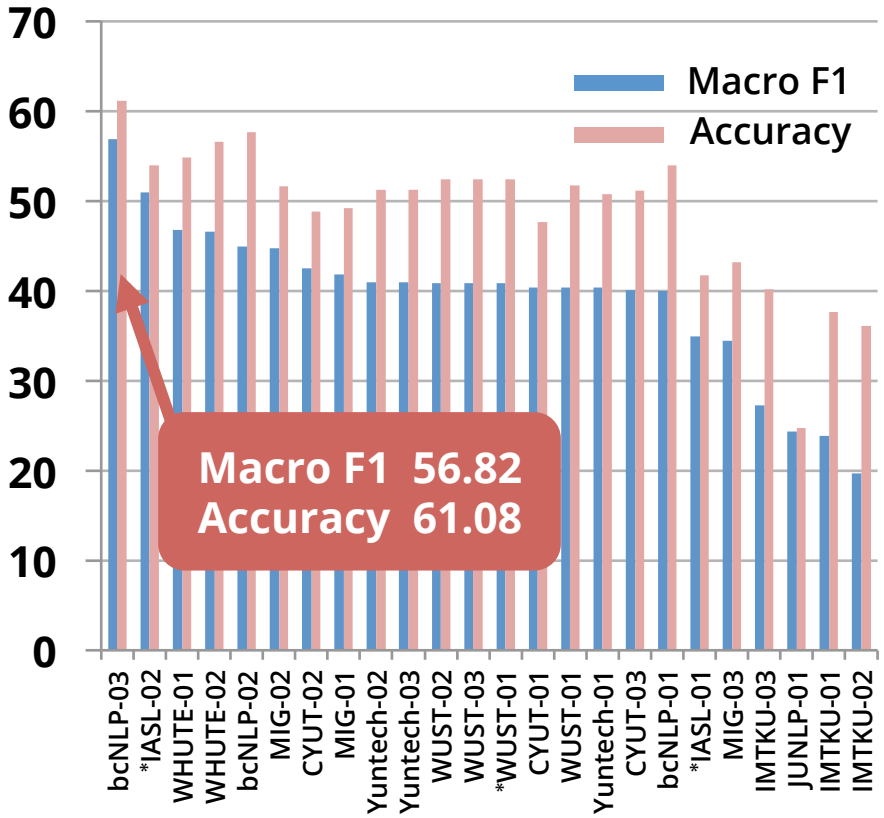


Difficulty:

Contradiction >>> Bi-directional > Forward Ent.

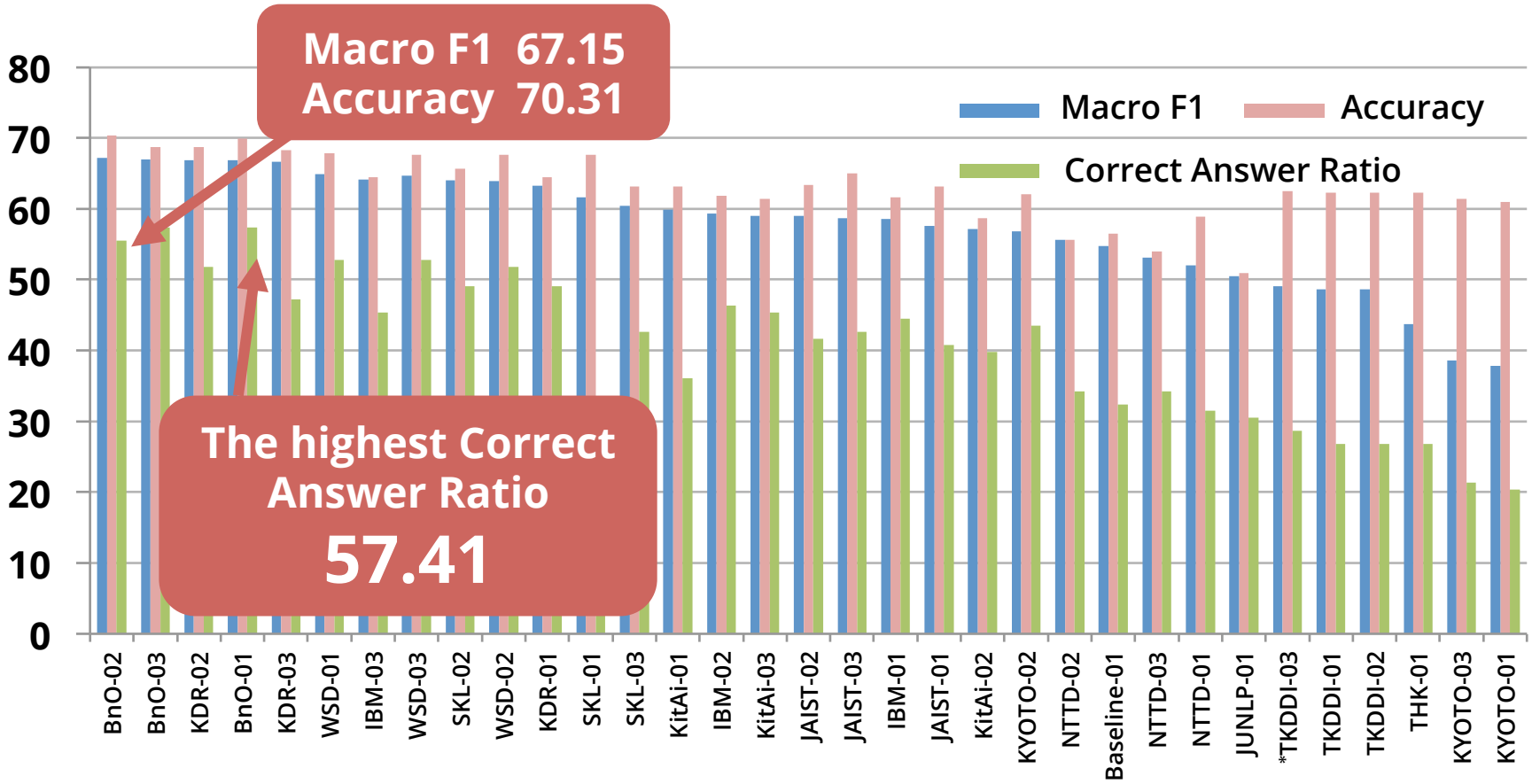
MC (Traditional/Simplified Chinese)





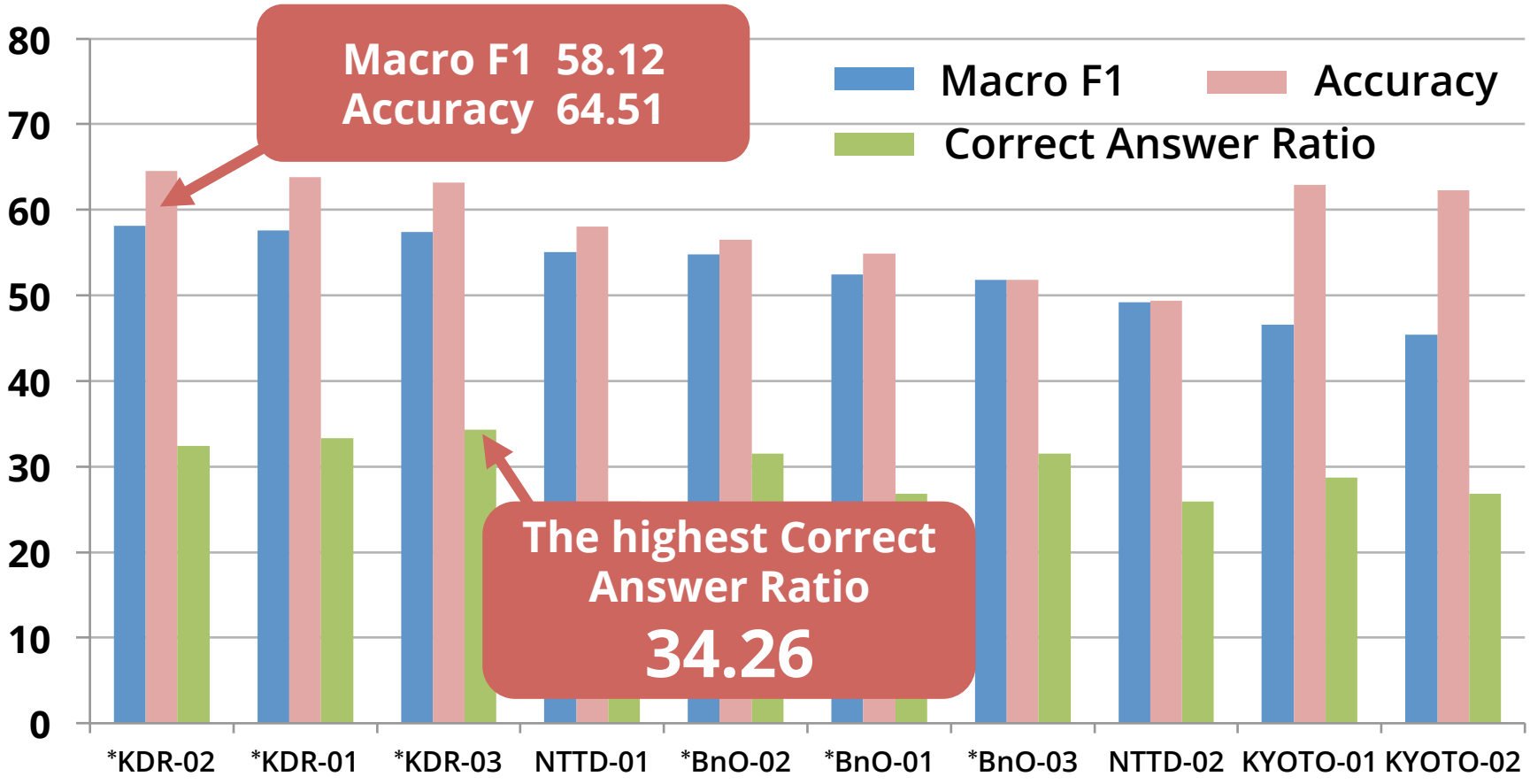
- **The top system in TC achieved approx. 52% of accuracy**
- **The top system in SC achieved over 60% of accuracy**

Exam BC (Japanese)



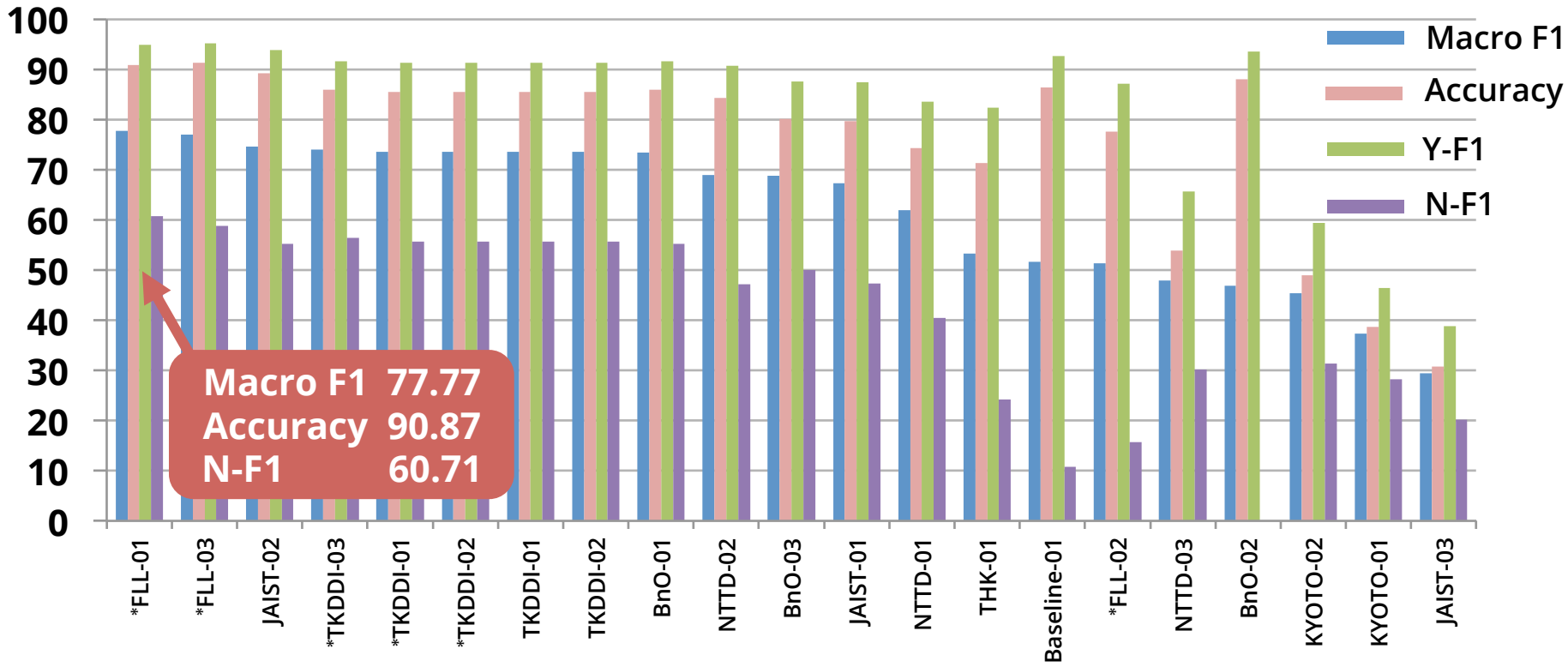
- **If candidate sentences in knowledge (Wikipedia and textbooks) are already obtained, the best system can answer more than 57% of exam questions correctly**

Exam Search



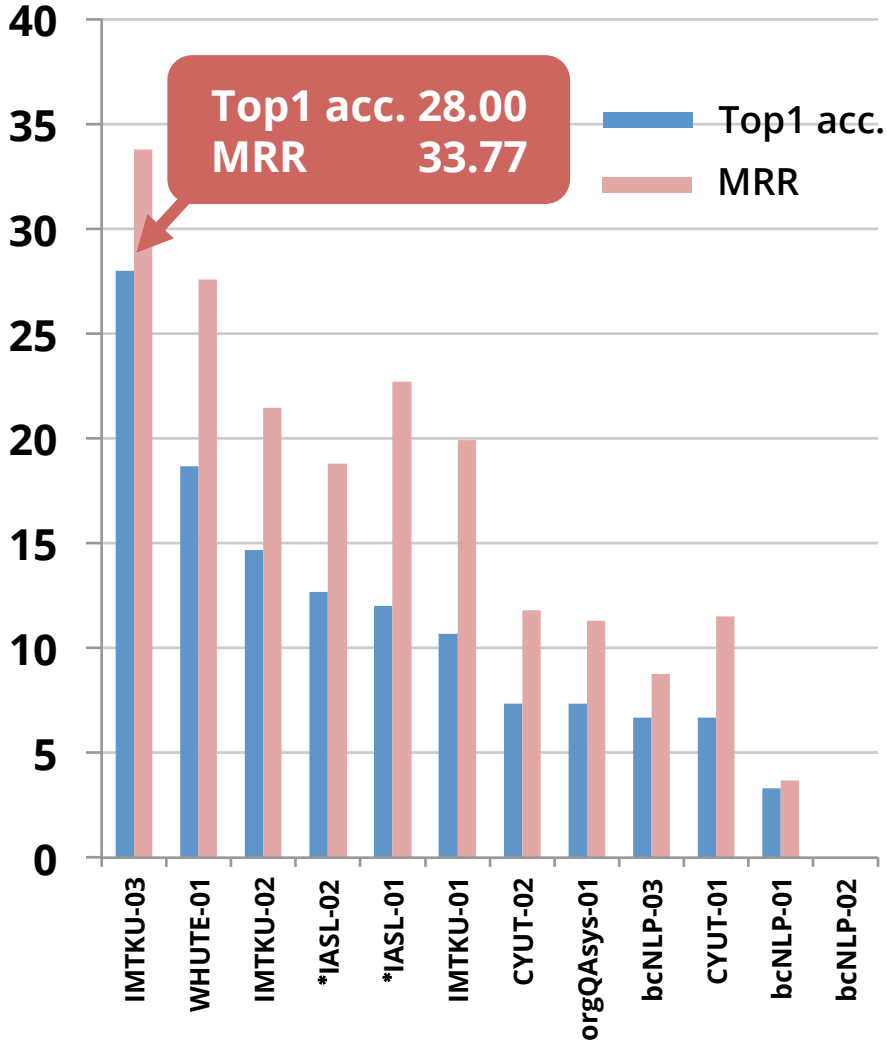
- **The best system could answer 34% of questions correctly in a search task setting**

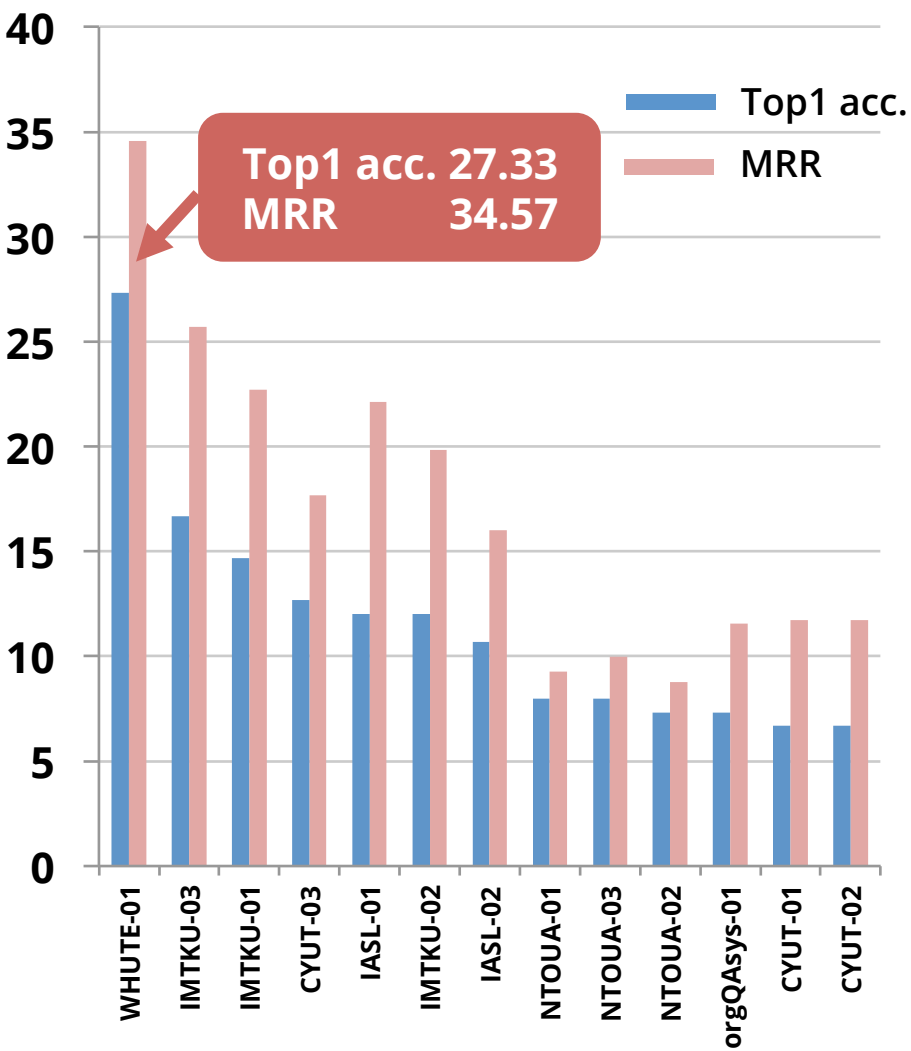
UnitTest



- **Since almost of the examples are Y (Y:219, N:29), improving performance of detecting "N" is important**
- **Due to the limited space, performances for each category cannot be shown here**

RITE4QA (Traditional/Simplified Chinese)





Review of Participants' Systems

Participant's approaches

- **Category**
  - Statistical (50%)
  - Hybrid (27%)
  - Rule-based (23%)
- **Fundamental approach**
  - Overlap-based (77%)
  - Alignment-based (63%)
  - Transformation-based (23%)

Summary of types of information explored

- Character/word overlap (85%)
- Syntactic information (67%)
- Temporal/numerical information (63%)
- Named entity information (56%)
- Predicate-argument structure (44%)
- Entailment relations (30%)
- Polarity information (7%)
- Modality information (4%)

Summary of Resources Explored

- **Japanese**
  - Wikipedia (10)
  - Japanese WordNet (9)
  - ALAGIN Entailment DB (5)
  - Nihongo Goi-Taikei (2)
  - Bunruigoihyo (2)
  - Iwanami Dictionary (2)
- **Chinese**
  - Chinese WordNet (3)
  - TongYiCi CiLin (3)
  - HowNet (2)

Advanced approaches

- **Logical approaches**
  - Dependency-based Compositional Semantics (DCS) [BnO], Markov Logic [EHIME], Natural Logic [THK]
- **Alignment**
  - GIZA [CYUT], ILP [FLL], Labeled Alignment [bcNLP, THK]
- **Search Engine**
  - Google and Yahoo [DCUMT]
- **Deep Learning**
  - RNN language models [DCUMT]
- **Probabilistic Models**
  - N-gram HMM [DCUMT], LDA [FLL]
- **Machine Translation**
  - [JUNLP, JAIST, KC99]

Oral Presentations (6/20 13:00-)

- **[DCUMT]** Tsuyoshi Okita. Local Graph Matching with Active Learning for Recognizing Inference in Text at NTCIR-10.
- **[SKL]** Shohei Hattori and Satoshi Sato. Team SKL's Strategy and Experience in RITE2.
- **[BnO]** Ran Tian, Yusuke Miyao, Takuya Matsuzaki and Hiroyoshi Komatsu. BnO at NTCIR-10 RITE: A Strong Shallow Approach and an Inference-based Textual Entailment Recognition System.
- **[FLL]** Takuya Makino, Seiji Okajima and Tomoya Iwakura. FLL: Local Alignments based Approach for NTCIR-10 RITE-2
- **[KDR]** Daniel Andrade, Masaaki Tsuchida, Takashi Onishi and Kai Ishikawa. Detecting Contradiction in Text by Using Lexical Mismatch and Structural Similarity
- **[NTTD]** Megumi Ohki, Takashi Suenaga, Daisuke Satoh, Yuji Nomura and Toru Takaki. Expanded Dependency Structure based Textual Entailment Recognition System of NTTDATA for NTCIR10-RITE2.
- **[IASL]** Cheng-Wei Shih, Chad Liu, Cheng-Wei Lee and Wen-Lian Hsu. IASL RITE System at NTCIR-10.
- **[WHUTE]** Han Ren, Hongmiao Wu, Chen Lv, Donghong Ji and Jing Wan. The WHUTE System in NTCIR-10 RITE Task.
- **[bcNLP]** Xiao-Lin Wang, Hai Zhao and Bao-Liang Lu. BCMI-NLP Labeled-Alignment-Based Entailment System for NTCIR-10 RITE-2 Task.
- **[IMTKU]** Chun Tu, Min-Yuh Day, Shih-Jhen Huang, Hou-Cheng Vong and Sih-Wei Wu. IMTKU Textual Entailment System for Recognizing Inference in Text at NTCIR-10 RITE2.

Conclusion

- **NTCIR-10 RITE-2**
  - Benchmark task of evaluating systems that infer semantic relations between sentences
  - Two subtasks were added
    - **Exam Search**: provided more realistic task setting
    - **UnitTest**: enabled us fine-grained evaluation and analysis of RITE systems
  - Organization Efforts
    - Provided pre-processed data (XML), Baseline tool and Evaluation tools
  - **28** teams participated! (NTCIR-9 RITE: 24 teams)
  - Diverse advanced approaches and resources were explored

RITE-2 was successful !