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.

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

Pyramid of entailment recognition technology

Application oriented

RITE

Linguistic phenomena-level inference

MC

Sentence-level inference

Search

contradiction?

entailment?

IR

QA

.documents

sentence

sentence

sentence

sentence

sentence

sentence

sentence

sentence

Sentence-level inference

Multiple sentence-level inference

World knowledge

Unit Test

LEXICAL

PHRASE

CASE

QUANTIFICATION

COORDINATION

MODIFICATION

NEGATION

...
RITE-2 Subtasks
BC and MC subtasks

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

- **t₁**: The Kamakura Shogunate began in Japan in the 12th century.

- **t₂**: The Kamakura Shogunate was considered to have begun in 1192, but the current leading theory is that it was effectively formed in 1185.

**BC subtask**
- Entailment (t₁ entails t₂) or Non-Entailment (otherwise)

**MC subtask**
- Bi-directional Entailment (t₁ entails t₂ & t₂ entails t₁)
- Forward Entailment (t₁ entails t₂ & t₂ does not entail t₁)
- Contradiction (t₁ contradicts t₂ or cannot be true at the same time)
- Independence (otherwise)

The 10th NTCIR Conference
Development of BC and MC data

retrieve pairs of sentences

edit pairs if needed

for each example, 5 annotators assigned its semantic label

RITE2 BC, MC data

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びえる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伯は、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」そ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 下線部の①の文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 近代はシリア派のイスラム教であった。
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 Sultan Süleyman, トルコ語 Süleyman, 1494年11月6日 - 1566年9月6日)は、オスマン帝国の第10代皇帝(在位:1520年 - 1566年)。

40年の長期にわたる在位の中で13回もの対外遠征を行い、数多くの軍事的成績を収めてオスマント帝国を最盛期に導いた。英語では、「立法帝(al-Qārīm)、立法帝(القادر al-Qārīm)/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績を収めてオスマント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)
  - \( t_1 \) and \( t_2 \) are given

• Entrance Exam Search
  - Binary-classification problem (Entailment or Non-entailment)
  - \( t_2 \) 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)

\[ t_1: \text{In the Meiji Constitution, legal clear distinction between the Imperial Family and Japan had been allowed.} \quad \downarrow \text{Category: modifier} \]
\[ t_2: \text{In the Meiji Constitution, distinction between the Imperial Family and Japan had been allowed.} \]

\[ t_1: \text{In the Meiji Constitution, distinction between the Imperial Family and Japan had been allowed.} \quad \downarrow \text{Category: melonymy} \]
\[ t_2: \text{In the Meiji Constitution, distinction between the Emperor and Japan had been allowed} \]
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

<table>
<thead>
<tr>
<th>Phenomena</th>
<th>dev</th>
<th>test</th>
</tr>
</thead>
<tbody>
<tr>
<td>lexical synonymy</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>hypernymy</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>meronymy</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>entailment</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>phrase synonymy</td>
<td>45</td>
<td>35</td>
</tr>
<tr>
<td>hypernymy</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>entailment</td>
<td>28</td>
<td>45</td>
</tr>
<tr>
<td>case alternation</td>
<td>9</td>
<td>7</td>
</tr>
<tr>
<td>modifier</td>
<td>30</td>
<td>42</td>
</tr>
<tr>
<td>nominalization</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>coreference</td>
<td>12</td>
<td>4</td>
</tr>
<tr>
<td>clause</td>
<td>29</td>
<td>14</td>
</tr>
<tr>
<td>relative clause</td>
<td>10</td>
<td>8</td>
</tr>
<tr>
<td>transparent head</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Phenomena</th>
<th>dev</th>
<th>test</th>
</tr>
</thead>
<tbody>
<tr>
<td>list</td>
<td>11</td>
<td>3</td>
</tr>
<tr>
<td>quantity</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>scrambling</td>
<td>16</td>
<td>15</td>
</tr>
<tr>
<td>inference</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>Implicit relation</td>
<td>10</td>
<td>18</td>
</tr>
<tr>
<td>apposition</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>temporal</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>spatial</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>disagree lexical</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>phrase</td>
<td>25</td>
<td>25</td>
</tr>
<tr>
<td>modality</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>spatial</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>temporal</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>272</strong></td>
<td><strong>241</strong></td>
</tr>
</tbody>
</table>
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)**

  \[ MacroF1 = \frac{1}{|C|} \sum_{c \in C} F1_c \quad \text{Accuracy} = 100 \times \frac{N_{\text{correct}}}{N_{\text{examples}}} \]

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

- **Top1 and MRR (RITE4QA)**

  \[ Top1 = \frac{1}{|Q|} \sum_{i=1}^{|Q|} [\text{top answer is correct}] \quad MRR = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\text{rank}_i} \]
Organization Effort
Generic Framework

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

1. Provided pre-processed data
2. Provided a fundamental entailment recognition tool
3. Provided RITE-2 evaluators

Documents → Linguistic Analyzer → Entailment Recognizer → Outputs
Sentence → Evaluator → Evaluation results (accuracies, F1-values...)

The 10th NTCIR Conference
(1) Pre-processed data

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

- **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
```

<table>
<thead>
<tr>
<th>Label</th>
<th>#</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>354</td>
<td>60.18( 204/ 339)</td>
<td>57.63( 204/ 354)</td>
<td>58.87</td>
</tr>
<tr>
<td>Y</td>
<td>256</td>
<td>44.65( 121/ 271)</td>
<td>47.27( 121/ 256)</td>
<td>45.92</td>
</tr>
</tbody>
</table>

**Accuracy:** 53.28( 325/ 610)

**Macro F1:** 52.40

**Confusion Matrix**

<table>
<thead>
<tr>
<th>gold \ sys</th>
<th>N</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>204</td>
<td>150</td>
</tr>
<tr>
<td>Y</td>
<td>135</td>
<td>121</td>
</tr>
</tbody>
</table>

- **Additional Evaluator (Entrance Exam)**
  - Calculate correct answer ratio
RITE-2 Formal Run Participation
Number of submissions

<table>
<thead>
<tr>
<th>NTCIR-10 RITE-2</th>
<th>JA</th>
<th>CT</th>
<th>CS</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>BC</td>
<td>41</td>
<td>20</td>
<td>21</td>
<td>82</td>
</tr>
<tr>
<td>MC</td>
<td>20</td>
<td>21</td>
<td>21</td>
<td>62</td>
</tr>
<tr>
<td>Exam BC</td>
<td>31</td>
<td>-</td>
<td>-</td>
<td>31</td>
</tr>
<tr>
<td>Exam Search</td>
<td>4</td>
<td>-</td>
<td>-</td>
<td>4</td>
</tr>
<tr>
<td>UnitTest</td>
<td>14</td>
<td>-</td>
<td>-</td>
<td>14</td>
</tr>
<tr>
<td>RITE4QA</td>
<td>-</td>
<td>12</td>
<td>10</td>
<td>22</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>110</td>
<td>53</td>
<td>52</td>
<td>215</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>NTCIR-9 RITE</th>
<th>JA</th>
<th>CT</th>
<th>CS</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total</strong></td>
<td>65</td>
<td>70</td>
<td>77</td>
<td>212</td>
</tr>
</tbody>
</table>
Countries/Regions of Participants

- **China**: 3 groups
- **Japan**: 15 groups
- **Taiwan**: 8 groups
- **India**: 1 group
- **Ireland**: 1 group
Formal Run Results
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
The top scores are almost the same as those in NTCIR-9 RITE
• 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.
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.
• 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)

Top1 acc. 28.00
MRR 33.77

Top1 acc. 27.33
MRR 34.57

The 10th NTCIR Conference
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.
- **[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!**