Syntactic Difference Based Approach for NTCIR-9 RITE Task

Team ID: IBM

Yuta Tsuboi, Hiroshi Kanayama, Masaki Ohno
IBM Research – Tokyo

Yuya Unno
Preferred Infrastructure

Submitted results for four Japanese RITE subtasks (BC, MC, EXAM and RITE4QA)

- Performed the best score in MC and EXAM subtasks
Overview of the approach

• Very difficult task due to the real-world complex dataset
  
  Impossible to write down hand-crafted rules
  
  Machine learning approach with various features

• Similarities and differences between \(<S>\) and \(<T>\) * are the key features for entailment determination
  
  Matching on the surface is not sufficient
  
  Calculation of tree edit distance between parsed trees

✓ Alignment of two sentences considering syntactic structures
✓ Estimation of similarity (small edit operation costs $\Leftrightarrow$ similar pair)
✓ Edit operations (insert, delete and replacement) as features of ML

* \(<S>\) and \(<T>\) denote an input sentence pair (\(<t1>\) and \(<t2>\) in the original file)
Techniques and resources for our machine learning approach

- Resources
  - Japanese thesaurus
  - Ontology

- Tree edit distance computation
- Pair feature extraction

- Training
  - Expanded training corpus
  - Logistic regression/Cross-validation

- Prediction
  - Classifier
  - Labels
  - Output: Y/N, F/R/B/C/I
Tree edit distance – Concept

- Hypotheses
  - Two sentences (\(<S>\) and \(<T>\)) have syntactic similarities and differences
  - A pair of similar sentences has high possibility of entailment
  - Difference parts can be clues for the determination of entailment

- Solution
  - Parse two sentences
  - Align parse trees by calculating the tree edit distance between them

\(<S>\)学校に毎日行く</S>
\(<T>\)学校に行く</T>

\(<S>\)大統領が殺された</S>
\(<T>\)熊が殺された</T>

- delete an adverb “everyday”
- replace “president” into “bear”
Tree Edit Distance – General Implementation

- **Edit distance** $\delta$
  \[ \delta(s, t) = \min_M \sum_{(s, t) \in M} \gamma(s, t) + \sum_{s \in D} \gamma(s, \epsilon) + \sum_{t \in I} \gamma(\epsilon, t) \]

- **Edit operations**

  **Replacement**
  - Replace $\circ \rightarrow \bullet$
  - Cost for replacement: $\gamma(s, t)$

  **Deletion**
  - Delete $\bullet$
  - Cost for deletion: $\gamma(s, \epsilon)$

  **Insertion**
  - Insert $\circ$
  - Cost for insertion: $\gamma(\epsilon, t)$

- Edit distance computation: $O(|s|^2|t|^2)$ time and space
- Our code for tree edit distance is available at
Cost Functions for Tree Edit Distance

- Insertion / Deletion cost: constant.
  \[ \gamma(s, \varepsilon) = \gamma(\varepsilon, t) = 1 \]
- Replacement cost: a smaller value for a more similar bunsetsu pair
  - Mixed various metrics for similarity:

<table>
<thead>
<tr>
<th>Cost functions</th>
<th>How to measure the similarity of two bunsetsus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jaccard distance metrics using word overlap (WO)</td>
<td>✓ Overlap ratio between each morpheme set (handling both content and functional words)</td>
</tr>
<tr>
<td>Semantic distance metrics using an ontology (Ontology)</td>
<td>✓ Inverse of shortest path length of two head content words in the ontology (see next slide)</td>
</tr>
<tr>
<td>Semantic distance metrics using thesaurus (BGH)</td>
<td>✓ Common depth of two head content words in the thesaurus tree</td>
</tr>
<tr>
<td>Heuristic distance metrics (HDM)</td>
<td>✓ Similarity value considering parts-of-speech and the thesaurus above</td>
</tr>
</tbody>
</table>

Semantic Similarity and Resources

- We defined two measures for semantic similarity with two complementary resources.

Manually generated resource:

Japanese Thesaurus [分類語彙表]

- High coverage for basic Japanese expression

Automatically generated resource:

IS-A Ontology generated from Wikipedia *

- Up-to-date knowledge using the latest edition

相似性測度: Common depth of two head content words in the tree

相似性測度: inverse of shortest path length in the ontology tree

Pair features (1) – Similarity and difference between S and T

- Represent a sentence pair with several features
- Train the logistic regression model using the annotated data

**Edit distance and operations (EDO)**

Normalized edit distance: \[ \frac{\delta(s, t)}{\max(|s|, |t|)} \]

Edit operations:
- insertion / deletion (e.g. Deletion: “everyday”, Deletion: ADVERB)
- replacement (e.g. Replacement: “president”-“bear”, Replacement: Noun - Noun)

**Word overlapping (Word)**

Overlap ratio: \[ \frac{m_s \cap m_t}{|m_t|} \]

Word pairs:
- (school, school)
- (go, school)
- (everyday, school)
- (school, go)
- (go, go)
- (everyday, go)

0 = same sentence
1 = totally different
### Pair features (2) - Ad-hoc strong clues

#### Sentiment polarity matching
- Applied existing sentiment detector
  - “It is excellent” → positive
  - “I don’t like this” → negative
- Sentiment orientation of the sentence pair
  - Same polarity
  - Different polarity
  - Opposite polarity

#### PAS fulfillment test (PAS)
- Convert $S$ and $T$ to the sets of predicate-argument structures

<table>
<thead>
<tr>
<th>sentence pairs</th>
<th>features</th>
</tr>
</thead>
</table>
| $S$            | $f_{pol} = (+, +)$  
                | $f_{same} = 1$ |
| $T$            | $f_{pol} = (+, 0)$  
                | $f_{diff} = 1$ |

#### Examples of fulfillment
- Whether all predicate-argument structures in $T$ are covered by those in $S$

- **Strong clues for entailment**

<table>
<thead>
<tr>
<th>$S$</th>
<th>$T$</th>
</tr>
</thead>
<tbody>
<tr>
<td>‘The PET (positron emission tomography) is believed to be effective for the care of most types of cancers such as ...’</td>
<td>‘PET helps the care of cancers.’</td>
</tr>
<tr>
<td>‘Director Yoji Yamada is good as making scenes of men’s weeping.’</td>
<td>‘Yoji Yamada is a film director.’</td>
</tr>
<tr>
<td>‘The Lost Decade was not wasted because ...’</td>
<td>‘We learned nothing from the Lost Decade.’</td>
</tr>
</tbody>
</table>

---

Examples:

- **Strong clue for non-entail**
  - Whether all predicate-argument structures in $T$ are covered by those in $S$
Pair features (3) – Designed for EXAM subtask

Temporal Matching

- Many sentence pairs in EXAM data includes temporal expressions
- Exploit a feature whether the temporal expressions in S and T have overlap

\(<S>\)シュマルカルデン同盟とは、1531年に、…</S>
‘Schmalkaldic League is … in 1531.’

\(<T>16世紀に、ドイツでは、シュマルカルデン同盟…</T>
‘In the 16th century of Germany, Schmalkaldic League …’

\(<S>1997年まで東南アジアバブルであったが、…</S>
‘Until 1997 there was an economic bubble in South-east Asia, …’

\(<T>1990年代初めに起こった日本でのバブル経済の崩壊が、…</T>
‘The collapse of Japanese bubbled economy happened in the early 1990s, …’
Expansion of the Training Data

- Convert the training data for the MC subtask as the additional training data for BC subtask, and vice versa.
  - e.g. Forward entailment label (F) between S and T is equal to the true entailment (Y) for \( S \rightarrow T \) and false entailment for \( T \rightarrow S \).

- Label conversion rule

<table>
<thead>
<tr>
<th>MC relation</th>
<th>BC relation</th>
</tr>
</thead>
<tbody>
<tr>
<td>( S \stackrel{F}{\rightarrow} T )</td>
<td>( S \stackrel{Y}{\rightarrow} T, T \stackrel{N}{\rightarrow} S )</td>
</tr>
<tr>
<td>( S \stackrel{R}{\rightarrow} T )</td>
<td>( S \stackrel{N}{\rightarrow} T, T \stackrel{Y}{\rightarrow} S )</td>
</tr>
<tr>
<td>( S \stackrel{B}{\rightarrow} T )</td>
<td>( S \stackrel{Y}{\rightarrow} T, T \stackrel{Y}{\rightarrow} S )</td>
</tr>
<tr>
<td>( S \stackrel{C}{\rightarrow} T )</td>
<td>( S \stackrel{N}{\rightarrow} T, T \stackrel{N}{\rightarrow} S )</td>
</tr>
<tr>
<td>( S \stackrel{I}{\rightarrow} T )</td>
<td>( S \stackrel{N}{\rightarrow} T, T \stackrel{N}{\rightarrow} S )</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>BC relation</th>
<th>MC relation</th>
</tr>
</thead>
<tbody>
<tr>
<td>( S \stackrel{Y}{\rightarrow} T )</td>
<td>( S \stackrel{F,B}{\rightarrow} T )</td>
</tr>
<tr>
<td>( S \stackrel{N}{\rightarrow} T )</td>
<td>( S \stackrel{R,C,I}{\rightarrow} T )</td>
</tr>
</tbody>
</table>

- Enhanced data
  - BC+MC' data : 500+880 pairs
  - MC+BC' data : 500+440 pairs (To handle label ambiguities, we train logistic regression by marginal log-likelihood maximization)
## Results – BC Subtask

<table>
<thead>
<tr>
<th>Cost Function</th>
<th>Features</th>
<th>Training</th>
<th>CV</th>
<th>AC</th>
</tr>
</thead>
<tbody>
<tr>
<td>w/o edit distance</td>
<td>None</td>
<td>Word + Sentiment + PAS + Temporal</td>
<td>BC</td>
<td>52.8</td>
</tr>
<tr>
<td>IBM BC1</td>
<td>HDM</td>
<td>EDO + Word + Sentiment + PAS</td>
<td>BC</td>
<td>54.8</td>
</tr>
<tr>
<td>IBM BC2</td>
<td>HDM + BGH</td>
<td>EDO + Word + Sentiment + PAS</td>
<td>BC</td>
<td>54.0</td>
</tr>
<tr>
<td>IBM BC3</td>
<td>HDM + BGH + WO</td>
<td>EDO (POS fine) + Word</td>
<td>BC + MC’</td>
<td><strong>64.1</strong></td>
</tr>
<tr>
<td>Oracle</td>
<td>BGH</td>
<td>EDO + Word + Sentiment + PAS</td>
<td>BC</td>
<td>51.8</td>
</tr>
</tbody>
</table>

- Positive performance gain with the edit distance method
- Very low correlation between CV and AC → the results are unpredictable
- BC+MC’ increases CV by 8%~13%, but no effect for AC

Cross validation on training data

Accuracy in formal run

Best feature set in formal run
### Results – MC Subtask

<table>
<thead>
<tr>
<th>w/o edit distance</th>
<th>Cost Function</th>
<th>Features</th>
<th>Training</th>
<th>CV</th>
<th>AC</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>Word + Sentiment + PAS + Temporal</td>
<td>MC</td>
<td>33.6</td>
<td>35.9</td>
<td></td>
</tr>
<tr>
<td>IBM MC1</td>
<td>HDM</td>
<td>EDO + Word + Sentiment + PAS</td>
<td>MC + BC’</td>
<td>46.8</td>
<td>43.6</td>
</tr>
<tr>
<td>IBM MC2</td>
<td>HDM + BGH + WO</td>
<td>EDO + Word</td>
<td>MC</td>
<td>50.2</td>
<td>50.2</td>
</tr>
<tr>
<td>IBM MC3</td>
<td>HDM + BGH + WO</td>
<td>EDO (POS fine) + Word</td>
<td>MC + BC’</td>
<td>51.3</td>
<td>44.5</td>
</tr>
<tr>
<td>Oracle</td>
<td>HDM + Ont + WO</td>
<td>EDO + Word + Sentiment</td>
<td>MC</td>
<td>51.1</td>
<td>51.6</td>
</tr>
</tbody>
</table>

- Achieved high accuracy for 5-fold classification
- EDO features increased the accuracy by 10%
- Other pair features tend not to work well
- Extended development data (MC+BC’) was not effective
## Results – EXAM Subtask

<table>
<thead>
<tr>
<th>w/o edit distance</th>
<th>Cost Function</th>
<th>Features</th>
<th>Training</th>
<th>CV</th>
<th>AC</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>Word + Sentiment + PAS + Temporal</td>
<td>EXAM</td>
<td>67.9</td>
<td>69.5</td>
<td></td>
</tr>
<tr>
<td>HDM</td>
<td>EDO + Word + Sentiment + PAS</td>
<td>EXAM</td>
<td>63.7</td>
<td>67.6</td>
<td></td>
</tr>
<tr>
<td>IBM EX1</td>
<td>HDM</td>
<td>EDO + Word + Sentiment + PAS + Temporal</td>
<td>EXAM</td>
<td>69.1</td>
<td>72.2</td>
</tr>
<tr>
<td>IBM EX2</td>
<td>Ontology</td>
<td>EDO + Word + Sentiment + PAS + Temporal</td>
<td>EXAM</td>
<td>62.5</td>
<td>67.4</td>
</tr>
<tr>
<td>IBM EX3</td>
<td>HDM + BGH + WO + Ontology</td>
<td>EDO (POS fine) + Word + Sentiment + PAS + Temporal</td>
<td>EXAM</td>
<td>61.5</td>
<td>58.4</td>
</tr>
<tr>
<td>Oracle</td>
<td>HDM</td>
<td>EDO + Word + Sentiment + PAS + Temporal</td>
<td>EXAM</td>
<td>68.3</td>
<td>72.6</td>
</tr>
</tbody>
</table>

- Relatively small contribution of edit distance
- Temporal increased the accuracy by 5%
- High correlation between CV and AC
Summary

- Achieved good performance in MC and EXAM subtasks
  - Machine learning approach with various features
  - Tree edit operations worked as key features (especially in MC task)
  - Use of thesaurus and ontology – complementary resources

- Performance is unpredictable – the model is still immature
- No special treatment for 5-fold classification in MC task – needed more observation
Backup
Details of MC results – confusion matrix

<table>
<thead>
<tr>
<th>System Output</th>
<th>Correct Label</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>F</td>
</tr>
<tr>
<td>F</td>
<td>87</td>
</tr>
<tr>
<td>R</td>
<td>7</td>
</tr>
<tr>
<td>B</td>
<td>7</td>
</tr>
<tr>
<td>C</td>
<td>4</td>
</tr>
<tr>
<td>I</td>
<td>5</td>
</tr>
</tbody>
</table>
### Results – RITE4QA Subtask

<table>
<thead>
<tr>
<th></th>
<th>Cost Function</th>
<th>Features</th>
<th>Training</th>
<th>AC</th>
<th>Top1</th>
<th>MMR 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>w/o edit distance</td>
<td>None</td>
<td>Word + Sentiment + PAS + Temporal</td>
<td>BC</td>
<td>34.5</td>
<td>5.5</td>
<td>19.8</td>
</tr>
<tr>
<td>IBM R4QA1</td>
<td>HDM &amp; Ont &amp; WO</td>
<td>EDO (POS fine) + Word + Sentiment + PAS</td>
<td>BC</td>
<td>33.3</td>
<td>11.3</td>
<td>23.3</td>
</tr>
<tr>
<td>IBM R4QA2</td>
<td>HDM &amp; BGH</td>
<td>EDO</td>
<td>BC</td>
<td>31.6</td>
<td>9.1</td>
<td>21.7</td>
</tr>
<tr>
<td>IBM R4QA3</td>
<td>HDM &amp; Ont &amp; WO</td>
<td>EDO (POS fine) + Word + Sentiment + PAS + Temporal</td>
<td>BC+MC’</td>
<td>40.1</td>
<td>8.7</td>
<td>22.2</td>
</tr>
<tr>
<td>Oracle</td>
<td>None</td>
<td>Word + Sentiment + PAS + Temporal</td>
<td>BC+MC’</td>
<td>63.5</td>
<td>18.1</td>
<td>29.0</td>
</tr>
</tbody>
</table>