Annotating Argumentative Relations for Mining Coherence Patterns

Jan **Wira** Gotama Putra (D1)
Tokunaga-lab (Computational Linguistics/NLP)
Dept. Computer Science, School of Computing
Tokyo Institute of Technology, Japan
Organisation Example in Student Essay

Prompt: Smoking should be banned at all restaurants in the country

1. Yes, smoking should be completely banned in all restaurants in the country.
2. Smoking is dangerous to our health and the government should empower people all over the country to quit.
3. Vendors should not have any cigarettes in their stores to avoid smoking.
4. When people do not follow the law, the government should give them sanctions.
5. Therefore, we should impose NO SMOKING CAMPAIGN.
Coherence Example in Student Essay

Prompt: Smoking should be banned at all restaurants in the country

1. Yes, smoking should be completely banned in all restaurants in the country.

2. Smoking is dangerous to our health and the government should empower people all over the country to quit

3. Vendors should not have any cigarettes in their stores to avoid smoking

4. When people do not follow the law, the government should give them sanctions

5. Therefore, we should impose NO SMOKING CAMPAIGN

Is it okay to switch these two sentences?
(1) I think smoking should be banned at restaurants.
(2) If somebody smokes in the restaurant, other people may not be able to enjoy the experience.
(3) At restaurants, customers enjoy eating and talking.

(1) I think smoking should be banned at restaurants.
(3) At restaurants, customers enjoy eating and talking.
(2) If somebody smokes in the restaurant, other people may not be able to enjoy the experience.
Why Coherence

• Coherence dictates how to order sentences (information) properly

• Explaining how to order sentences to generate coherent texts is important in many natural language generation applications:
  
  • Multi-document summarisation of news (Barzilay et al., 2002; Okazaki et al., 2004)
  
  • Opinion generation in debate (Yanase et al., 2015)
  
  • Argumentative micro-texts (Peldszuz and Stede, 2016)
  
  • Intelligent Language Tutoring Systems (Al-khatib et al., 2016)
Intelligent Language Tutoring Systems

- **Automatic Essay Scoring + Score-based Feedback**
  
  Telling students which aspect he/she is lacking

- **Example-based Feedback**
  
  Telling students how to order sentences properly by giving a proper ordering example based on his/her own text
Research Goal (Ongoing)

• Our previous study revealed that there is a possibility of difference of relational patterns between coherent and less coherent text (Putra and Tokunaga, 2017)

• Explaining how sentences are related to each other explains text coherence

• We aim to explain text coherence by analysing relations between sentences

  Annotating imperfect text, then compare the relational patterns between the imperfect and improved text
Argumentative Structure as a Representation

How the structure is realised (in sequential ordering) differentiates coherent and incoherent text.
Text Collection Criteria

• Imperfect texts (score as proof) → argumentative essays

• Enables us to isolate the study of coherence; e.g., do not consider the grammatical issue

• Not to long (to reduce complications), only one paragraph if possible
## Some Existing Argumentative Corpora

<table>
<thead>
<tr>
<th>Source</th>
<th>Genre</th>
<th>#Doc</th>
<th>IAAs</th>
</tr>
</thead>
</table>
| Stab and Gurevych (2014)| Argumentative Essays from online forum (sentence labelling + relations) | 90   | • Krippendorff $\alpha_u = 0.72$ (sentence label)  
• Fleiss $K = 0.80$~ (relations)  
• 3 annotators each |
| Kirschner et al. (2015) | Scientific Articles (relations)                               | 24   | • Fleiss $K = 0.43$  
• 3 annotators each |
| Al-Khatib et al. (2016) | Scientific Articles (sentence labelling)                      | 300  | • Fleiss $K = 0.56$  
• 3 annotators each |
| Pelczszuz and Stede     |                                                              | 112  | • Fleiss $K = 0.83$  
• #annotators is not clear |
| Stab and Gurevych (2017)| Argumentative Essays from online forum (sentence labelling + relations) | 402  | • Krippendorff $\alpha_u = 0.767$ (sentence label)  
• Fleiss $K = 0.70$~ (relations)  
• 80 essays by 3 annotators for IAA  
• One expert annotator annotated the rest 300~ |
| Carlile et al. (2018)   | Argumentative Essays from online forum (sentence labelling + relations) | 102  | • Krippendorff $\alpha_u = 0.50$~++  
• 2 annotators each |

None satisfy our criteria
ICNALE Dataset (Ishikawa, 2013)

- Collection of argumentative writing by Asian students (common writing in education)
- Student essays typically need revisions
- All essays are much or less same at length (no length bias): (200-300 words)
- Grammatically error-free subset available (640 essays out of 2800) – April 2018

Medium -- Target (textual improvement)

Terrible (beyond help unless rewritten)

Excellent (enhancing persuasiveness)

Score Range

<table>
<thead>
<tr>
<th>Score Range</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-10</td>
<td>1</td>
</tr>
<tr>
<td>10-20</td>
<td>4</td>
</tr>
<tr>
<td>20-30</td>
<td>6</td>
</tr>
<tr>
<td>30-40</td>
<td>15</td>
</tr>
<tr>
<td>40-50</td>
<td>51</td>
</tr>
<tr>
<td>50-60</td>
<td>172</td>
</tr>
<tr>
<td>60-70</td>
<td>163</td>
</tr>
<tr>
<td>70-80</td>
<td>131</td>
</tr>
<tr>
<td>80-90</td>
<td>83</td>
</tr>
<tr>
<td>90-100</td>
<td>14</td>
</tr>
</tbody>
</table>
I think banning smoking in all restaurants is necessary. It is essential to protect the citizen's health.

Banning means officially forbid them to smoke.

But, some restaurants are popular because men are allowed to smoke.

In conclusion, I believe smoking should be banned at all restaurants.
Relation Annotation

• Goal:
  Producing structure (tree)

• Objects:
  • argument components,
  • non-argumentative components (dropped)

• Relations:
  \{support, detail, attack, restatement\}
  *from source (S) to target (T)
Annotation Steps

• Annotating relations (cf. argument mining)
  • Top-down
    • Find a main claim (root)
    • Divide sentences into groups (and subgroups at deeper level)
  • Bottom-up
    • Establish relations between sentences in a group
    • Establish relations between groups (higher hierarchical level)

• Reordering sentences (improving coherence)

• Repairing referring and connective expressions
  Reordering may alter how people and things are described or connected
I agree with the previous statement. If somebody smokes in the restaurant, other people may not be able to enjoy the experience. At restaurants, customers enjoy eating and talking. However, if we ban smoking in restaurants, then those restaurants might lose some customers. Some restaurants are indeed popular, especially among old men, because they allow people to smoke. But, I firmly support banning smoking in restaurants because we need to prioritise health. In conclusion, I encourage banning smoking in all restaurants.
(Prompt) Smoking should be banned at all restaurants in the country.

(1) I agree with the previous statement.
(2) If somebody smokes in the restaurant, other people may not be able to enjoy the experience.
(3) At restaurants, customers enjoy eating and talking.
(4) However, if we ban smoking in restaurants, then those restaurants might lose some customers.
(5) Some restaurants are indeed popular, especially among old men, because they allow people to smoke.
(6) But, I firmly support banning smoking in restaurants because we need to prioritise health.
(7) In conclusion, I encourage banning smoking in all restaurants.
(1) I agree with the previous statement.
(2) If somebody smokes in the restaurant, other people may not be able to enjoy the experience.
(3) At restaurants, customers enjoy eating and talking.
(4) However, if we ban smoking in restaurants, then those restaurants might lose some customers.
(5) Some restaurants are indeed popular, especially among old men, because they allow people to smoke.
(6) But, I firmly support banning smoking in restaurants because we need to prioritise health.
(7) In conclusion, I encourage banning smoking in all restaurants.
(Prompt) Smoking should be banned at all restaurants in the country.

(1) I agree with the previous statement.
(2) If somebody smokes in the restaurant, other people may not be able to enjoy the experience.
(3) At restaurants, customers enjoy eating and talking.
(4) However, if we ban smoking in restaurants, then those restaurants might lose some customers.
(5) Some restaurants are indeed popular, especially among old men, because they allow people to smoke.
(6) But, I firmly support banning smoking in restaurants because we need to prioritise health.
(7) In conclusion, I encourage banning smoking in all restaurants.
I agree with the previous statement.

At restaurants, customers enjoy eating and talking.

If somebody smokes in the restaurant, other people may not be able to enjoy the experience.

However, if we ban smoking in restaurants, then those restaurants might lose some customers.

Some restaurants are indeed popular, especially among old men, because they allow people to smoke.

But, I firmly support banning smoking in restaurants because we need to prioritise health.

In conclusion, I encourage banning smoking in all restaurants.
(Prompt) Smoking should be banned at all restaurants in the country.

1. I agree with the previous statement.
2. If somebody smokes in the restaurant, other people may not be able to enjoy the experience.
3. At restaurants, customers enjoy eating and talking.
4. However, if we ban smoking in restaurants, then those restaurants might lose some customers.
5. [...] some restaurants are indeed popular, especially among old men, because they allow people to smoke.
6. But, I firmly support banning smoking in restaurants because we need to prioritise health.
7. In conclusion, I encourage banning smoking in all restaurants.

Implies readers have read the prompt
I agree with the previous statement that smoking should be banned at all restaurants.

At restaurants, customers enjoy eating and talking. If somebody smokes in the restaurant, other people may not be able to enjoy the experience.

However, if we ban smoking in restaurants, then those restaurants might lose some customers. It is because some restaurants are indeed popular, especially among old men, because they allow people to smoke.

But, I firmly support banning smoking in restaurants because we need to prioritise health.

In conclusion, I encourage banning smoking in all restaurants.
Annotation Tool

- JavaScript
- No need to install anything (web-based)
Measuring Inter-annotator Agreement (IAA)

• Agreement Ratio for Dropping (argumentative vs non-argumentative sentence)
  It is because most sentences are argumentative, so using chance-corrected measure leads to a low score and underestimate the agreement

• Chance-corrected measure + modified agreement ratio for relational agreement
  Calculating how many relations (edges in the graph) having the same label are shared between two annotation

• Modified agreement ratio for multiplied relation
  Taking into account restatement resolution when measuring relational agreement
Inter-Annnotator Agreement on Dropping

Treat as binary labelling of sentences (drop--TRUE or not drop--FALSE)

<table>
<thead>
<tr>
<th>Sentence</th>
<th>Annotator 1</th>
<th>Annotator 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>TRUE</td>
<td>FALSE</td>
</tr>
<tr>
<td>2</td>
<td>FALSE</td>
<td>TRUE</td>
</tr>
<tr>
<td>3</td>
<td>TRUE</td>
<td>TRUE</td>
</tr>
<tr>
<td>4</td>
<td>FALSE</td>
<td>FALSE</td>
</tr>
<tr>
<td>5</td>
<td>FALSE</td>
<td>FALSE</td>
</tr>
<tr>
<td>6</td>
<td>TRUE</td>
<td>TRUE</td>
</tr>
</tbody>
</table>

4/6 = 0.67
Traditional Inter-Annnotator Agreement on Relations

For an essay having $N$ sentences

- $N \times (N - 1)$ combinations of source and target
- We measure the relation type + linking at once

\[
\begin{array}{cccc}
\text{Source} & \text{Target} & \text{Annotator 1} & \text{Annotator 2} \\
1 & 2 & \text{NO} & \\
\vdots & \vdots & \vdots & \\
2 & 1 & \text{sup} & \\
2 & 3 & \text{NO} & \\
\vdots & \vdots & \vdots & \\
3 & 2 & \text{=} & \\
\end{array}
\]

Problem: cannot reflect agreement score for restatement and structural similarity
Restatement Resolution: ``Extra Relation” (Multiplication)

We treat two restatement nodes as an equivalence class with respect to incoming and outgoing connections with extra relations.

![Diagram showing the equivalence of restatement nodes with extra relations.]

<table>
<thead>
<tr>
<th>Source</th>
<th>Target</th>
<th>Original Relation</th>
<th>Extra Relation</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1</td>
<td>sup</td>
<td>sup</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>=</td>
<td>=</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>sup</td>
<td>=</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>=</td>
<td>=</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>att</td>
<td>att</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>att</td>
<td></td>
</tr>
</tbody>
</table>

It is better not to take into account non-relations as in traditional agreement ratio, because it is rather meaningless (Kirschner et al., 2015)
Graph Similarity (Structure Similarity)

• Graph Edit Distance -- *inexact matching* (Alberto and King-sun, 1983)
  • Given two graphs $G$ and $H$, find an *edit path* between $G$ and $H$, that is a sequence of node or edge *insertion, removal* or *substitution* which transforms $G$ to $H$
  • It is not suitable with our task because we cannot just substitute a vertex with another one

• Maximum Common Edge Subgraph -- *exact matching* (Bokhari, 1981)
  • Given two graphs $G$ and $H$ with the same number of vertices, one has to find a common subgraph of $G$ and $H$ (not necessarily *induced*) with the maximum number of edges
  • The solution to this problem is known as NP-complete and requires vertices matching (graph isomorphism)
  • Straightforward solution: check if a subgraph in $G$ presents in $H \rightarrow O(2^{|V|}); |V| = \#\text{vertices}
Modified Agreement Ratio (MAR) with Restatement Resolution: ``Extra Relation” (Multiplication)

To what extent a modified graph subsumes another graph (approximating maximum common edges)

\[ G_1 = \frac{\text{overlap}(g_2, \text{multiply}(g_1))}{\text{edges}(g_1)} \]

\[ G_2 = \frac{\text{overlap}(g_1, \text{multiply}(g_2))}{\text{edges}(g_2)} \]

\[ \text{MAR} = \frac{G_1 + G_2}{2} \]

Number of edges in \( g_2 \) exist in the multiplied relations of \( g_1 \)

Number of original edges of \( g_1 \)

Arithmetic mean
we welcome your suggestions if you think geometric or harmonic mean is better

*) Note: we also have an option of not populating the graph
Illustration: ``Extra Relation” (Multiplication)

Annotator 1 ($g_1$)

1
sup

t det

2
sup

att

3

4
Att

5
Det

Annotator 2 ($g_2$)

1
att

2
sup

= det

3

4

5

\[
G_1 = \frac{\text{overlap}(g_2, g_1)}{\text{edges}(g_1)} = \frac{2}{4} = 0.50
\]

\[
G_2 = \frac{\text{overlap}(g_1, g_2)}{\text{edges}(g_2)} = \frac{2}{4} = 0.50
\]

\[
\text{MAR} = \frac{G_1 + G_2}{2} = \frac{0.50 + 0.50}{2} = 0.50
\]

\[
G_1 = \frac{\text{overlap}(g_2, \text{multiply}(g_1))}{\text{edges}(g_1)} = \frac{2}{4} = 0.50
\]

\[
G_2 = \frac{\text{overlap}(g_1, \text{multiply}(g_2))}{\text{edges}(g_2)} = \frac{3}{4} = 0.75
\]

\[
\text{MAR} = \frac{G_1 + G_2}{2} = \frac{0.50 + 0.75}{2} = 0.625
\]

Our solution is an approximation of maximum common edges (but do not necessarily form a subgraph)
## Pilot Study: Dropping

<table>
<thead>
<tr>
<th>Essay Code</th>
<th>Score</th>
<th>Agreement Ratio</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>W - D</td>
<td>W - O</td>
</tr>
<tr>
<td>THA_PTJ0_001_B1_2</td>
<td>45.8</td>
<td>0.73</td>
<td>0.73</td>
</tr>
<tr>
<td>JPN_PTJ0_041_B2_0</td>
<td>65.4</td>
<td>0.93</td>
<td>1.00</td>
</tr>
<tr>
<td>SIN_PTJ0_014_B2_0</td>
<td>91.3</td>
<td>1.00</td>
<td>0.86</td>
</tr>
<tr>
<td><strong>Overall Ratio</strong></td>
<td></td>
<td>0.89</td>
<td>0.87</td>
</tr>
</tbody>
</table>

| Overall Fleiss Kappa |       | 0.56  | 0.43  | 0.35  | 0.44  |
Pilot Study: Relational Agreement

<table>
<thead>
<tr>
<th>Metrics</th>
<th>Annotators</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>W - D</td>
<td>W - O</td>
<td>D - O</td>
</tr>
<tr>
<td>Cohen's Kappa</td>
<td></td>
<td>0.59</td>
<td>0.50</td>
<td>0.40</td>
</tr>
<tr>
<td>Modified Agreement Ratio w/o multiplication</td>
<td></td>
<td>0.56</td>
<td>0.46</td>
<td>0.35</td>
</tr>
<tr>
<td>Modified Agreement Ratio with multiplication</td>
<td></td>
<td>0.56</td>
<td>0.46</td>
<td>0.35</td>
</tr>
<tr>
<td>Modified Agreement Ratio with multiplication w/o considering label</td>
<td></td>
<td>0.64</td>
<td>0.61</td>
<td>0.48</td>
</tr>
</tbody>
</table>

Fleiss Kappa = 0.50

It happens that in our current pilot study, restatements happen in the exactly same pair of sentences between annotators.
The most confused relations are “det” and “sup” (this is expected).
However, the difference of resulting hierarchical structures is the most pressing problem.
Verdict

- Annotators recognise different structures (different interpretation of texts)
  - Different ordering $\implies$ different grouping/relations $\implies$ different structures
  - IAA implicitly measures the difference of dropping and structures

- Higher agreement on higher scored essays
  - I think because of the detailed instruction of recognising groups (organisation).
  - It is easier to recognise groups in highly-scored essays since they are more organised
  - Annotating non-perfect essays is challenging (interpretation problem)
Plan (near future)

- Annotate larger corpus
- Argument quality assessment (organisation score)
Contact Me

URL  https://wiragotama.github.io
E-mail  gotama.w.aa@m.titech.ac.jp
References (1)


References (2)


