Semantic Relations between Text Segments for Semantic Storytelling: Annotation Tool – Dataset – Evaluation

LREC 2022

Michael Raring, Malte Ostendorff, Georg Rehm

DFKI GmbH, Alt-Moabit 91c, 10559 Berlin, Germany
michael.raring@gmail.com, {malte.ostendorff, georg.rehm}@dfki.de
Semantic Storytelling

Vision
Automatic and semi-automatic generation of stories based on extracted, processed, classified and annotated information from large content resources [1][2][3]

Goal
Support content curators in creating new storylines through the relevant information extracted and presented by a corresponding tool [3]

Building Blocks / Core Process
Determine...
1. the relevance of a segment
2. the importance
3. the semantic relation between two segments [2]
Process of the Experiment

1. Extract Text Segments and Metadata
2. Prematch Text Segments
3. Annotate Relation of Text Pairs
4. Fine-Tune BERT
5. Evaluate
Process of the Experiment

1. Extract Text Segments and Metadata
2. Prematch Text Segments
3. Annotate Relation of Text Pairs
4. Fine-Tune BERT
5. Evaluate
Process of the Experiment

1. Extract Text Segments and Metadata
2. Prematch Text Segments
3. Annotate Relation of Text Pairs
4. Fine-Tune BERT
5. Evaluate

Methods:
- randomized
- NSP
- similarity
Process of the Experiment

Extract Text Segments and Metadata

Prematch Text Segments

Methods:
- randomized
- NSP
- similarity

INCEpTION Annotation Editor

Annotate Relation of Text Pairs

Fine-Tune BERT

Evaluate
Annotation Software

INCEpTION
- Open-source web-based annotation platform [13]
- Annotation of various NLP-related features
- Project administration features
  - Annotation layers
  - Annotation classes
  - Annotators
  - Documents
  - Knowledge resources
  - Recommender systems
...
- Data structure: Apache UIMA CAS (Unstructured Information Management Architecture / Common Analysis System) [14]
The unemployment rate in the United States has surpassed ten percent, according to figures released by the US Department of Labor on Friday.

Data from the United States Labor Department says that the unemployment rate in the U.S. has reached its highest level in over 25 years, namely 9.4%.

Source Doc: US unemployment rate surpasses 10%

Source Doc: US unemployment rate reaches 9.4 percent
Annotated Sentence Pairs

12 (7) relation classes (subset)

2,501 sentence pairs

2,638 articles
Annotated Sentence Pairs

12 (7) relation classes (subset)

2,501 sentence pairs

2,638 articles
Pretrained Language Model

Class Label

C \hspace{1cm} T_1 \hspace{1cm} \ldots \hspace{1cm} T_N \hspace{1cm} T_{[SEP]} \hspace{1cm} T_1' \hspace{1cm} \ldots \hspace{1cm} T_M'

E_{[CLS]} \hspace{1cm} E_1 \hspace{1cm} \ldots \hspace{1cm} E_N \hspace{1cm} E_{[SEP]} \hspace{1cm} E_1' \hspace{1cm} \ldots \hspace{1cm} E_M'

[CLS] \hspace{1cm} Tok 1 \hspace{1cm} \ldots \hspace{1cm} Tok N \hspace{1cm} [SEP] \hspace{1cm} Tok 1 \hspace{1cm} \ldots \hspace{1cm} Tok M

Sentence 1 \hspace{2cm} Sentence 2
## Input Strategies

<table>
<thead>
<tr>
<th>Input Strategy</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>S</td>
<td>Sentence. However preliminary results based on 95% of the votes cast give Hamas’ Change and Reform Party 76 seats, leaving Fatah with 43 seats.</td>
</tr>
<tr>
<td>TS</td>
<td>Title. Sentence. Hamas wins Palestinian election. However preliminary results based on 95% of the votes cast give Hamas’ Change and Reform Party 76 seats, leaving Fatah with 43 seats.</td>
</tr>
<tr>
<td>TSD</td>
<td>Title. Sentence. Date Hamas wins Palestinian election. However preliminary results based on 95% of the votes cast give Hamas’ Change and Reform Party 76 seats, leaving Fatah with 43 seats. January 29, 2005</td>
</tr>
<tr>
<td>STD</td>
<td>Sentence. Title. Date However preliminary results based on 95% of the votes cast give Hamas’ Change and Reform Party 76 seats, leaving Fatah with 43 seats. Hamas wins Palestinian election. January 29, 2005</td>
</tr>
</tbody>
</table>
Results Depending on Input Strategy

a) Full classification

b) 7-classes classification

Accuracy

Input Strategy
Results Depending on Language Model

a) Full classification

b) 7-classes classification
### Class Results for Full Classification

<table>
<thead>
<tr>
<th>Class</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-score</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>87.0</td>
<td>86.0</td>
<td>86.5</td>
<td>794</td>
</tr>
<tr>
<td>Identity</td>
<td>78.3</td>
<td>77.5</td>
<td>77.6</td>
<td>20</td>
</tr>
<tr>
<td>Equivalence</td>
<td>78.5</td>
<td>71.9</td>
<td>74.6</td>
<td>85</td>
</tr>
<tr>
<td>Causal</td>
<td>57.3</td>
<td>67.0</td>
<td>61.6</td>
<td>109</td>
</tr>
<tr>
<td>Contrast</td>
<td>55.0</td>
<td>66.1</td>
<td>59.9</td>
<td>41</td>
</tr>
<tr>
<td>Temporal</td>
<td>45.9</td>
<td>50.3</td>
<td>48.0</td>
<td>147</td>
</tr>
<tr>
<td>Conditional</td>
<td>24.1</td>
<td>21.4</td>
<td>21.5</td>
<td>14</td>
</tr>
<tr>
<td>Description</td>
<td>26.5</td>
<td>13.1</td>
<td>17.4</td>
<td>13</td>
</tr>
<tr>
<td>Attribution</td>
<td>20.8</td>
<td>8.3</td>
<td>11.8</td>
<td>6</td>
</tr>
<tr>
<td>Fulfillment</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>11</td>
</tr>
<tr>
<td>Summary</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>6</td>
</tr>
<tr>
<td>Purpose</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>6</td>
</tr>
<tr>
<td><strong>Micro avg.</strong></td>
<td><strong>74.6</strong></td>
<td><strong>75.0</strong></td>
<td><strong>74.6</strong></td>
<td><strong>1251</strong></td>
</tr>
<tr>
<td><strong>Macro avg.</strong></td>
<td><strong>39.5</strong></td>
<td><strong>38.5</strong></td>
<td><strong>38.2</strong></td>
<td><strong>1251</strong></td>
</tr>
</tbody>
</table>
Class Results for 7-Classes Classification

<table>
<thead>
<tr>
<th>Class</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-score</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>87.8</td>
<td>87.8</td>
<td>87.8</td>
<td>794</td>
</tr>
<tr>
<td>Identity</td>
<td>90.1</td>
<td>76.2</td>
<td>82.2</td>
<td>20</td>
</tr>
<tr>
<td>Equivalence</td>
<td>74.2</td>
<td>68.4</td>
<td>71.0</td>
<td>85</td>
</tr>
<tr>
<td>Causal</td>
<td>59.9</td>
<td>67.7</td>
<td>63.4</td>
<td>109</td>
</tr>
<tr>
<td>Contrast</td>
<td>66.0</td>
<td>62.5</td>
<td>64.1</td>
<td>41</td>
</tr>
<tr>
<td>Temporal</td>
<td>50.5</td>
<td>53.5</td>
<td>51.9</td>
<td>147</td>
</tr>
<tr>
<td>Others</td>
<td>41.6</td>
<td>33.5</td>
<td>36.9</td>
<td>55</td>
</tr>
<tr>
<td>Micro avg.</td>
<td>77.3</td>
<td>77.2</td>
<td>77.2</td>
<td>1251</td>
</tr>
<tr>
<td>Macro avg.</td>
<td>67.2</td>
<td>64.2</td>
<td>65.3</td>
<td>1251</td>
</tr>
</tbody>
</table>
Conclusion

• We have developed an annotation editor for pairwise relations and manually annotated the semantic relation between 2500 sentence pairs individually extracted from Wikinews articles.
• We successfully trained a multi-class classifier to recognize semantic relations between text segments from different source documents by fine-tuning multiple versions of BERT’s language model with our annotated dataset.
• We ran all trainings with a 12-class & 7-class setup. The 7-class classification performed better in micro & macro metrics.
• DeBERTa (base) performed best with accuracy and micro F1-score of 75% (12-class) and 77% (7-class).
• Adding metadata such as the article name and publication date has on average led to slightly better and more stable results.
• Future work:
  • Dataset: Improve imbalance, classes, pre-matching, pre-processing.
  • Classifier: Use more recent models, test few-shot approaches.
Thanks! Any questions?

Michael Raring

hub.cards/michael.raring
michael.raring@gmail.com

In cooperation with:

Reproduce this research:

Editor Source Code + Scripts + Dataset + Best Model:

https://github.com/DFKI-NLP/semantic-storytelling
References (1)


References (2)


Images

Slide 12 contains an image from [10]

Slide 19 has been designed using resources from Flaticon.com