AMPERSAND: Argument Mining for PERSuAsive oNline Discussions

Tuhin Chakrabarty, Christopher Hidey, Smaranda Muresan
Kathleen McKeown and Alyssa Hwang

Natural Language Processing
Columbia University
CMV: I think the Paris shooting makes a good case for culture of responsible gun ownership.

....

I also think this would be much better then increase in police numbers and rights as policemen are a very homogeneous group trained to stick together and the danger of even deepening the "not one of us" (pack) mentality and escalation of "police state".

....

Those time periods/places did not enjoy lower rates of violent death. [...] but according [this study] the murder rate was "extraordinarily high by today's standard."

...

Looking at the rest of the thread, I think I see what might be a flaw in our collective logic. The incidents that stand out most in our heads are events like Paris or Aurora or Columbine- premeditated shootings with the intent of inflicting a lot of casualties.
CMV: I think the Paris shooting makes a good case for culture of responsible gun ownership.

I also think this would be much better then increase in police numbers and rights as policemen are a very homogeneous group trained to stick together and the danger of even deepening the "not one of us" (pack) mentality and escalation of "police state".

Those time periods/places did not enjoy lower rates of violent death. [...] but according [this study] the murder rate was "extraordinarily high by today's standard."

Looking at the rest of the thread, I think I see what might be a flaw in our collective logic. The incidents that stand out most in our heads are events like Paris or Aurora or Columbine- premeditated shootings with the intent of inflicting a lot of casualties.
CMV: I think the Paris shooting makes a good case for culture of responsible gun ownership.

I also think this would be much better than increasing police numbers and rights as policemen are a very homogeneous group trained to stick together and the danger of even deepening the "not one of us" (pack) mentality and escalation of "police state".

Those time periods/places did not enjoy lower rates of violent death. [...] but according [this study] the murder rate was "extraordinarily high by today's standard."

Looking at the rest of the thread, I think I see what might be a flaw in our collective logic. The incidents that stand out most in our heads are events like Paris or Aurora or Columbine- premeditated shootings with the intent of inflicting a lot of casualties.
TASK OF INTEREST

● Identify argument components in a full discussion thread
  ○ Claim
  ○ Premise

● Identify two kinds of argument relations:
  ○ Inter-turn relations (argumentative relations to support or attack the other person’s argument)
  ○ Intra-turn relations (to support one’s claim or premise).
CONTRIBUTIONS

● Introduce two novel distantly-labeled data-sets

● Fine-tune BERT on the distant-labeled data and demonstrate improvement

● Use discourse relations from RST for argument relations.
## LABELLED DATASET STATISTICS

<table>
<thead>
<tr>
<th>Category</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOTAL NUMBER OF THREADS</td>
<td>112</td>
</tr>
<tr>
<td>NUMBER OF DIALOGUE TURNS</td>
<td>380</td>
</tr>
<tr>
<td>#CLAIM, #PREMISE, #NONARGUMENT</td>
<td>1205, 1536, 799</td>
</tr>
<tr>
<td>TOTAL NO. OF INTRA-TURN RELATIONS</td>
<td>4.6(%) Of 27254 Possible Pairs</td>
</tr>
<tr>
<td>TOTAL NO. OF INTER-TURN RELATIONS</td>
<td>3.2(%) Of 26695 Possible Pairs</td>
</tr>
</tbody>
</table>
DISTANT LABELED DATA

- Challenge: small size labeled dataset
- Leverage distant-labeled data from Reddit and use transfer learning techniques
  - micro-level for intra-turn relations
  - macro-level (dialogue) for inter-turn relations.
MICRO-LEVEL CONTEXT DATA

IMHO, Calorie-counting is a crock what you have to look at is how whole- some are the foods you are eating. Refined sugar is worse than just empty calories - I believe your body uses a lot of nutrients up just processing and digesting it.

CLAIM & PREMISE

IMHO + Context  Dataset (4.6 Million comments) inspired by Chakrabarty et al (2019)
CMV: A rise in female representation in elected government isn’t a good or bad thing.

... According to this new story, a record number of women are seeking office in this year’s US midterm elections. While some observers hail this phenomenon as a step in the right direction, I don’t think it’s good thing one way or the other: a politician’s sex has zero bearing on their ability to govern or craft effective legislation. As such...

“I don’t think it’s good thing one way or the other: a politician’s sex has zero bearing on their ability to govern or craft effective legislation”

Nobody is saying that women are better politicians than men, and thus, more female representation is inherently better for our political system. Rather, the argument is that...
Feature of Reddit: Users can easily *quote* another user’s response used to highlight exactly what part of someone’s argument a particular user is targeting.

97,636 pairs from 19,413 threads

Language model fine-tuning → sentence level,
  ○ (quoted text, following sentence) → distant-labeled inter-post pairs
  ○ QR, for quote-response pairs
ARGUMENT COMPONENT CLASSIFICATION

C/P/None

BERT CLASSIFIER

BERT CONTEXT
FINE-TUNING

IMHO + Context
Dataset
INTRA RELATION CLASSIFICATION

ARGUMENT RELATION ENSEMBLE

RST-BASED CLASSIFIER

BERT CLASSIFIER

S

T

ARGUMENT COMPONENT CLASSIFICATION

IMHO + Context Dataset

BERT CONTEXT FINE-TUNING
INTER-RELATION CLASSIFICATION

ARGUMENT RELATION ENSEMBLE

Rel / NoRel

RST-BASED CLASSIFIER

ARGUMENT COMPONENT CLASSIFICATION

T’

CANDIDATE TARGET SELECTION

BERT CLASSIFIER

BERT CONTEXT FINE-TUNING

QR Dataset
# DISCOURSE RELATIONS

Argument Pair where each argument is broken into EDU's (Elementary Discourse Units)

<table>
<thead>
<tr>
<th>ARGUMENT 1</th>
<th>ARGUMENT 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>(If existence from your perspective)(_1) (lies solely on your consciousness,)(_2)</td>
<td>(after you die)(_3) (it doesn’t matter what you left)(_4)</td>
</tr>
</tbody>
</table>
DISCOURSE RELATIONS

SN-condition

NN-same_unit SN-attribution

EDU EDU EDU EDU

1 2 3 4
To extract features from a pair of argumentative components

- Concatenate 2 components → single text input

- Use state-of-the-art RST discourse parser (Ji and Eisenstein, 2014)
  - Create parse trees
  - Use predicted discourse relation @ root as a categorical feature

- Use a one-hot encoding of these relations as features and train an XGBoost Classifier and predict Relation (1) vs No-Relation(0)
CMV: I think the Paris shooting makes a good case for culture of responsible gun ownership.

....

I also think this would be much better then increase in police numbers and rights as policemen are a very homogeneous group trained to stick together and the danger of even deepening the "not one of us" (pack) mentality and escalation of "police state".

....

CMV: I think the Paris shooting makes a good case for culture of responsible gun ownership.

....

I also think this would be much better then increase in police numbers and rights
CANDIDATE TARGET SELECTION

- Use the QR data to train & identify candidate targets using techniques from extractive summarization (BERTSUM *),

- Treat the quoted sentences as gold labels, resulting in 19,413 pairs of document (post) and summary (quoted sentences).

- Thus, for a candidate argument pair A → B, where B is the quoted sentence, if B is not extracted by the summarization model we predict that there is no relation between A and B.

*Fine-tune BERT for Extractive Summarization , YangLiu 2019
PIPELINE SUMMARY

● Argument Component Classification

● Intra-relations
  ■ Fine tuning (MICRO)
  ■ Discourse relation

● Inter-relations
  ■ Fine tuning (MACRO)
  ■ Discourse relation
  ■ Candidate Selection
# ARGUMENT COMPONENT CLASSIFICATION

<table>
<thead>
<tr>
<th>METHOD</th>
<th>CLAIM</th>
<th>PREMISE</th>
<th>NON ARGUMENT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stab and Gurevych (2017) + EWE</td>
<td>56.0</td>
<td>65.9</td>
<td>69.6</td>
</tr>
<tr>
<td>Morio and Fujita (2018)</td>
<td>54.2</td>
<td>68.5</td>
<td>73.2</td>
</tr>
<tr>
<td>Chakrabarty et al. (2019)</td>
<td>57.8</td>
<td>70.8</td>
<td>70.5</td>
</tr>
<tr>
<td>BERT Devlin et al. (2019)</td>
<td>62.0</td>
<td>72.2</td>
<td>71.3</td>
</tr>
<tr>
<td>IMHO + Context Fine-Tuned BERT</td>
<td>67.1</td>
<td>72.5</td>
<td>75.7</td>
</tr>
</tbody>
</table>
# INTRA-TURN RELATION PREDICTION

<table>
<thead>
<tr>
<th>METHOD</th>
<th>PRECISION</th>
<th>RECALL</th>
<th>F-SCORE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GOLD</td>
<td>PRED</td>
<td>GOLD</td>
</tr>
<tr>
<td>All Relations</td>
<td>5.0</td>
<td>-</td>
<td>100.0</td>
</tr>
<tr>
<td>Morio and Fujita (2018)</td>
<td>10.0</td>
<td>-</td>
<td>48.8</td>
</tr>
<tr>
<td>BERT Devlin et al (2019)</td>
<td>12.0</td>
<td>11.0</td>
<td>67.0</td>
</tr>
<tr>
<td>IMHO Context Fine-Tuned BERT</td>
<td>14.3</td>
<td>13.2</td>
<td>69.0</td>
</tr>
<tr>
<td>+ RST Ensemble</td>
<td>16.7</td>
<td>15.5</td>
<td>73.0</td>
</tr>
</tbody>
</table>

*More baselines in paper*
# INTER-TURN RELATION PREDICTION

<table>
<thead>
<tr>
<th>METHOD</th>
<th>PRECISION</th>
<th>RECALL</th>
<th>F-SCORE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GOLD</td>
<td>PRED</td>
<td>GOLD</td>
</tr>
<tr>
<td>All Relations</td>
<td>5.0</td>
<td>-</td>
<td>100.0</td>
</tr>
<tr>
<td>Morio and Fujita (2018)</td>
<td>7.6</td>
<td>-</td>
<td>40.0</td>
</tr>
<tr>
<td>BERT Devlin et al (2019)</td>
<td>8.8</td>
<td>7.9</td>
<td>76.0</td>
</tr>
<tr>
<td>QR Context Fine-Tuned BERT</td>
<td>11.0</td>
<td>10.0</td>
<td>75.3</td>
</tr>
<tr>
<td>+ RST Ensemble</td>
<td>11.0</td>
<td>12.2</td>
<td>79.0</td>
</tr>
<tr>
<td>Candidate Target Selection</td>
<td>18.9</td>
<td>17.5</td>
<td>79.0</td>
</tr>
</tbody>
</table>

*More baselines in paper*
<table>
<thead>
<tr>
<th>TABLE</th>
<th>PAIR</th>
</tr>
</thead>
<tbody>
<tr>
<td>IMHO</td>
<td>[IMHO, you should not <em>quantify</em> it as good or bad.][Tons of people have monogamous relationships without issue.]</td>
</tr>
<tr>
<td>CMV</td>
<td>[how would you even <em>quantify</em> that.][there are many people who want close relationships without romance.]</td>
</tr>
</tbody>
</table>
**IMPACT OF FINE-TUNING**

| QR | [It might be that egalitarians, anti-feminists, MRAs & redpillers, groups that I associate with opposing feminism - might be in fact very **distinct** & **different** groups, but I don’t know that] [I do see all four of these as **distinct** groups]. |
| CMV | [I may have a different stance on seeing no **difference** between companion animals and farm animals.] [I do see **distinction** between a pet and livestock] |
### IMPACT OF DISCOURSE RELATIONS

<table>
<thead>
<tr>
<th>Discourse Relation</th>
<th>Argument 1</th>
<th>Argument 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evaluation</td>
<td>The only way your life lacks meaning is if you give it none to begin with</td>
<td>Life is ultimately meaningless and pointless</td>
</tr>
<tr>
<td>Antithesis</td>
<td>Joseph was just a regular Jew without the same kind of holiness as the other two</td>
<td>Aren’t Mary and Joseph, two holy people especially perfect virgin Mary, both Jews? Wasn’t Jesus a Jew?</td>
</tr>
</tbody>
</table>
CONCLUSION

- Show how fine-tuning on data-sets similar to the task of interest is often beneficial
- Show how to use transfer learning by leveraging discourse and dialogue context
- Show structure of the fine-tuning corpus is essential for improved performance on pre-trained language models
- Predictions that take advantage of RST discourse cues are complementary to BERT predictions
- Demonstrated methods to reduce the search space and improve precision
QUESTIONS

tc2896@columbia.edu
chidey@cs.columbia.edu

https://github.com/tuhinjubcse/AMPERSAND-EMNLP2019