"Sharks are not the threat humans are": Argument Component Segmentation in School Student Essays

Tariq Alhindhi and Debanjan Ghosh
Motivation

• The central component of the argument are the claim and premise
  “an assertion that deserves our attention” (Toulmin, 2003)
  “a statement that is in dispute and that we are trying to support with reasons” (Govier, 2010)

• Argumentation
  segmentation, component (claim, premise) and relation (support, attack) detection
  claim example “Parents should be limiting screen time for their children”

• Argument/Claim Detection applications
  Assess public opinion on political and social issues to foster public deliberation
  Refine search and information retrieval
  other applications: legal documents, fact-checking
Motivation

Educational Applications

Analyze students' writings for essay scoring (Klebanov et al. 2014, Ghosh et al. 2016)

Generate quantitative and qualitative feedbacks to help students (K-12) writings such as on the Writing Mentor app
Should Artificial Sweeteners be Banned in America?

Diet soda, sugar-free gum, and low-calorie sweeteners are what most people see as a way to sweeten up a day without the calories.

Despite the lack of calories, artificial sweeteners have multiple negative health effects.

Over the past century, science has made it possible to replicate food with fabricated alternatives that simplify weight loss.

Although many thought these new replacements would benefit overall health, there are more negative effects on manufactured food than the food they replaced.

Artificial sweeteners have a huge impact on current day society.

Legends: B-Claim I-Claim B-Premise I-Premise O-Arg
Research questions

Can we detect claim and premise boundaries at the token level from school student essays?

Can we develop robust claim and premise detection models that go beyond lexicon-based matching?
Task

Claim and Premise Segmentation in each sentence using a B-I-O setup

- B-Claim/B-Premise: beginning of a claim/premise
- I-Claim/I-Premise: inside a claim/premise
- O-Arg: outside argumentative text

Although many thought these new replacements would benefit overall health, there are more negative effects on manufactured food than the food they replaced.
Although many thought these new replacements would benefit overall health, there are more negative effects on manufactured food than the food they replaced.
Outline of the Talk

Background and Data

Features and Models

Results

Discussions and Future Work
Background and Data

Segmentation then Type Detection and Relation Detection using 400 argumentative essays (Stab & Gurevych, 2017, Eger et al. 2017))

Sentence Claim Detection
Cross-domain claim detection in 6 datasets (Daxenberger et al. 2017)
Using unlabeled data from the same domain (Chakrabarty et al. 2019)

Characteristics of our Dataset (why merging segmentation and Type Detection)
Multiple argument components in one sentence
Many unsupported claims
Annotation Efforts

• Annotation of argumentative essays with 3 annotators

• Annotation of claims and premises

• Several rounds of calibration to finalize the guideline
  10 essays were triple annotated in the pilot task
  65 essays are annotated by pair(s); rest are annotated by one annotator

• For claim: modest agreement of 0.71 and for premise: high agreement of 0.9 (Krippendorff’s $\alpha$)
## Dataset

<table>
<thead>
<tr>
<th>Split</th>
<th>Essays</th>
<th>B-Claim</th>
<th>I-Claim</th>
<th>B-Premise</th>
<th>I-Premise</th>
<th>O-Arg</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>training</td>
<td>100</td>
<td>1,780</td>
<td>21,966</td>
<td>317</td>
<td>3,552</td>
<td>51,478</td>
<td>36,546</td>
</tr>
<tr>
<td>dev</td>
<td>10</td>
<td>171</td>
<td>1,823</td>
<td>32</td>
<td>371</td>
<td>4,008</td>
<td>6,405</td>
</tr>
<tr>
<td>test</td>
<td>35</td>
<td>662</td>
<td>8,207</td>
<td>92</td>
<td>1,018</td>
<td>14,987</td>
<td>24,955</td>
</tr>
</tbody>
</table>
Outline of the Talk

Background and Data

Features and Models

Results

 Discussions and Future Work
Features

- Discrete Features
- Word Embeddings from Transformers (e.g., BERT)

Models

- Conditional Random Field (Sequence Classification)
- Deep Learning Models
  - BiLSTMs
  - Transformers (e.g., BERT)
Discrete Feature Groups

Structural
- token position features, punctuation features
- position of covering sentence

Syntactic
- POS, Lowest common ancestor (LCA), LCA types

Lex-Syntactic
- Relations governing the token and its context
- N hops deep in retrieving relations
- Features: (token, previous, next): token_dependency-relation
Feature-based Models

Conditional Random Field (CRF)

Takes a sequence as input (sentence)

Make prediction to each element (token) in the sequence while considering other neighboring elements (tokens in the same sentence)
Deep Learning Models

• Training a BiLSTM-CRF tagger

• Pretrained transformers (e.g., BERT) as feature extractors
  language models trained on a lot of data (wikipedia, google books)
  using embeddings generated by these models as features

• Pretrained transformers as classifiers
  fine-tuning on the task (training partition)
  fine-tuning on a related unlabeled corpus
  fine-tuning with a multi-task objective
Fine-tuning Transformers

Training a Bert-based classifier

1. Fine-tuning on the training partition
   - model choice and hyperparameter tuning (cased vs uncased, batch size, number of epochs, etc.), weighted loss per class according to class frequencies
Fine-tuning Transformers

Training a Bert-based classifier

1. Fine-tuning on the training partition
   model choice and hyperparameter tuning (cased vs uncased, batch size, number of
   epochs, etc.), weighted loss per class according to class frequencies

2. Adaptive pretraining on argumentative corpora:
   IMHO: trained on argumentative subreddit
   (Chakrabarty et al. 2019)

   GRE essays: trained on student essays
   (Ghosh et al. 2020)
Fine-tuning Transformers

Training a Bert-based classifier

3. Fine-tuning with adaptive pretraining on unlabeled data from a relevant domain followed by fine-tuning on the labeled dataset with the multitask variation

dynamic weighting of task specific losses

\[ L = \sum_{t} \frac{1}{2\sigma_t^2} L_t + \log \sigma_t^2 \]
Outline of the Talk

Background and Data

Features and Models

Results

Discussions and Future Work
Results

- Discrete features (different sets) with CRF

<table>
<thead>
<tr>
<th>Features</th>
<th>CRF</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B-Claim</td>
</tr>
<tr>
<td>lexSyn</td>
<td>.395</td>
</tr>
<tr>
<td>Discrete*</td>
<td>.269</td>
</tr>
<tr>
<td>Embeddings</td>
<td>.401</td>
</tr>
<tr>
<td>Embeddings+lexSyn</td>
<td>.482</td>
</tr>
<tr>
<td>Embeddings+Discrete*</td>
<td>.434</td>
</tr>
</tbody>
</table>

- BiLSTM standalone model
  - +CRF based identification

<table>
<thead>
<tr>
<th>Setup</th>
<th>BiLSTM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B-Claim</td>
</tr>
<tr>
<td>BiLSTM</td>
<td>.556</td>
</tr>
<tr>
<td>BiLSTM-CRF</td>
<td>.558</td>
</tr>
</tbody>
</table>
## Results

- **BERT$_{bl}$**: BERT based baseline
- **BERT$_{blIMHO}$**: fine-tuned on IMHO
- **BERT$_{blEssay}$**: fine-tuned on essays
- **BERT$_{MT}$**: like above in a MT setting (token + sentence classification)

<table>
<thead>
<tr>
<th>Setup</th>
<th>B-Claim</th>
<th>I-Claim</th>
<th>B-Premise</th>
<th>I-Premise</th>
<th>O-Arg</th>
<th>Acc.</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERT$_{bl}$</td>
<td>.563</td>
<td>.674</td>
<td>.274</td>
<td>.425</td>
<td>.795</td>
<td>.728</td>
<td>.546</td>
</tr>
<tr>
<td>BERT$_{blIMHO}$</td>
<td>.571</td>
<td>.681</td>
<td>.304</td>
<td>.410</td>
<td>.795</td>
<td>.730</td>
<td>.540</td>
</tr>
<tr>
<td>BERT$_{blEssay}$</td>
<td>.564</td>
<td>.676</td>
<td>.261</td>
<td>.406</td>
<td>.792</td>
<td>.747</td>
<td>.561</td>
</tr>
<tr>
<td>BERT$_{mt}$</td>
<td>.567</td>
<td>.685</td>
<td>.242</td>
<td>.439</td>
<td>.805</td>
<td>.741</td>
<td>.548</td>
</tr>
<tr>
<td>BERT$_{mtIMHO}$</td>
<td>.562</td>
<td>.684</td>
<td>.221</td>
<td>.413</td>
<td>.794</td>
<td>.731</td>
<td>.534</td>
</tr>
<tr>
<td>BERT$_{mtEssay}$</td>
<td>.580</td>
<td>.702</td>
<td>.254</td>
<td>.427</td>
<td>.810</td>
<td>.752</td>
<td>.574</td>
</tr>
</tbody>
</table>
Takeaways

• Best model is BERT$_{mt\text{-}essay}$
  • Gains in all five token-classes, up to 7.5% accuracy

• 5-class: Merged Segmentation and Type Detection

• 3 class: Argument Segmentation (B-Arg, I-Arg, O-Arg)
  • only gives 2 points improvements
  • Much lower than other datasets

• Low accuracy for Premise detection
Outline of the Talk

Background and Data

Features and Models

Results

Discussions and Future Work
Challenges

Allowed Transitions

• back-to-back B-claim, B-premise
• I-claim starts after O-Arg

post-processing based on the possible transitions

\[
\begin{align*}
B & \rightarrow I \\
I & \rightarrow B & I & \rightarrow I & I & \rightarrow O \\
O & \rightarrow B & O & \rightarrow O
\end{align*}
\]
Challenges

non-arguments classified as arguments (frequent)

• Mixing arguments with opinions
  e.g.  that actually makes me feel good afterward
Challenges

missing multiple-claims from a sentence

Some coral can recover from this though for most it is the final straw.

Some coral can recover from this though for most it is the final straw.
Challenges

investigating run-on sentences

Humans in today’s world do not care about the consequences, only the money they may gain.

The oceans are also another dire need in today’s environment, each day becoming more filled with trash and plastics.
Effect of the Multitask Learning

• identify claims and premises that are missed by the single task model(s), such as:

  many more negative effects that come with social media claim

• clever handling of the back-propagation helps the multitask model to reduce false positives to be more precise.

  internet’s social networks help teens find communities non-arg
  take $1.3$ billion o $11.3$ billion the NCAA makes and give it to players opinion
Conclusion and Future Work

- Our findings show that a multitask BERT performs the best with an absolute gain of 7.5% accuracy over the discrete features.

- We can also generate personalized and relevant feedback for the students (e.g., which are the supported/unsupported claims in the essay?) that is useful in the paradigm of automated writing assistance.

- domain identification (college essays vs. school essays) genre prediction in essays (argumentative vs. narrative),
Thank You