Introduction

- Student-centered discussions are an important contributor to student’ learning in English Language Arts (ELA) classes
- Automatically predicting argument components (claim, warrant, evidence) can help teachers analyze student arguments
- We evaluate the performance of an existing argument component classification model developed for a different educationally-oriented domain [1]
- We analyze the effectiveness of features from prior work on argument mining for student essays and online dialogues
- We provide a comparison between convolutional neural networks and recurrent neural networks in several different conditions (character vs. word input, including handcrafted features)
- We evaluate the impact of multi-task learning by leveraging specificity information

Annotation Scheme [2]

Argumentation

- **Claim**: an arguable statement that presents a particular interpretation of a text or topic
- **Evidence**: facts, documentation, text reference, or testimony used to support or justify a claim
- **Warrant**: reasons explaining how a specific evidence instance supports a specific claim

Specificity labels for an argument move:

- **Low**: it does not contain any specificity element
- **Medium**: it accomplishes one of the elements
- **High**: it clearly accomplishes at least 2 specificity elements

Specificity elements for an argument move:
1. It is specific to one (or a few) character or scene
2. It makes significant qualifications or elaborations
3. It uses content-specific vocabulary
4. It provides a chain of reasoning

Dataset

- 73 high school-level text-based discussions (i.e. centered on a literature piece)
- Preprocessed by manually segmenting turns at talk into argument moves (ADUs)
- 2047 argument moves
- Examples:

  Student | Argument Move | Argument component | Specificity
  --- | --- | --- | ---
  S1 | Well Fezzik went back to how he was | claim | low
  S1 | like how he gets lost. Then he goes like he needs to be around other people. And then finally when he does, he gets himself like relying on himself. But then right at the end, he doesn’t know where he is; he makes a wrong turn. cause he tried doing it by himself and he can’t. So I think Fezzik went back to his normal ways, like after he changed. | evidence | medium
  S1 | | warrant | high

- Class distribution:

<table>
<thead>
<tr>
<th>Argument Component</th>
<th>Claim</th>
<th>Warrant</th>
<th>Evidence</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>710 (34.7%)</td>
<td>996 (46.7%)</td>
<td>341 (16.6%)</td>
<td>1751</td>
</tr>
<tr>
<td>Med</td>
<td></td>
<td></td>
<td></td>
<td>81</td>
</tr>
<tr>
<td>High</td>
<td>358 (17.5%)</td>
<td></td>
<td>655 (32%)</td>
<td>1013</td>
</tr>
</tbody>
</table>

Models

Task: given an argument move predict its argument component label (claim, evidence, warrant)

Handcrafted features:
- wLDA:
  - Lexical features, parse features, structural features, context features
- Online dialogue [3]:
  - Semantic density features, lexical features, syntactic features

Neural Network Models

(a) Neural network only setup, to test whether neural networks can extract important features for argument component classification
(b) Model incorporating neural network and handcrafted features, to test whether features manually engineered can make models more robust
(c) Multi-task setup for neural network only model, to test whether additional information on specificity can impact performance on argument component classification
(d) Combination of multi-task setup and handcrafted features

Results

Table 1: Results obtained through leave-one-transcript-out cross validation

<table>
<thead>
<tr>
<th>Row</th>
<th>Models / Features</th>
<th>Kappa</th>
<th>Precision</th>
<th>Recall</th>
<th>F-score</th>
<th>P1</th>
<th>P2</th>
<th>P3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Majority baseline</td>
<td>0.068</td>
<td>0.265</td>
<td>0.406</td>
<td>0.314</td>
<td>0.109</td>
<td>0.004</td>
<td>0.532</td>
</tr>
<tr>
<td>2</td>
<td>Pre-trained wLDA</td>
<td>0.077</td>
<td>0.289</td>
<td>0.35</td>
<td>0.269</td>
<td>0.265</td>
<td>0.158</td>
<td>0.456</td>
</tr>
<tr>
<td>3</td>
<td>Logistic Regression (wLDA features)</td>
<td>0.142</td>
<td>0.412</td>
<td>0.394</td>
<td>0.379</td>
<td>0.39</td>
<td>0.211</td>
<td>0.54</td>
</tr>
<tr>
<td>4</td>
<td>Logistic Regression (wLDA features + online dialogue)</td>
<td>0.283</td>
<td>0.508</td>
<td>0.5</td>
<td>0.48</td>
<td>0.479</td>
<td>0.222</td>
<td>0.693</td>
</tr>
</tbody>
</table>

Table 2: Results of the multi-task word level NN models

<table>
<thead>
<tr>
<th>Row</th>
<th>Models / Features</th>
<th>Kappa</th>
<th>Precision</th>
<th>Recall</th>
<th>F-score</th>
<th>P1</th>
<th>P2</th>
<th>P3</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>LSTM</td>
<td>0.069</td>
<td>0.408</td>
<td>0.399</td>
<td>0.218</td>
<td>0.161</td>
<td>0.128</td>
<td>0.295</td>
</tr>
<tr>
<td>10</td>
<td>LSTM + wLDA + online dialogue</td>
<td>0.181</td>
<td>0.462</td>
<td>0.477</td>
<td>0.391</td>
<td>0.362</td>
<td>0.274</td>
<td>0.522</td>
</tr>
<tr>
<td>11</td>
<td>CNN</td>
<td>0.125</td>
<td>0.41</td>
<td>0.404</td>
<td>0.378</td>
<td>0.37</td>
<td>0.231</td>
<td>0.526</td>
</tr>
<tr>
<td>12</td>
<td>CNN + wLDA + online dialogue</td>
<td>0.241*</td>
<td>0.492*</td>
<td>0.488</td>
<td>0.455*</td>
<td>0.468</td>
<td>0.274</td>
<td>0.622</td>
</tr>
</tbody>
</table>

Overall trends

- Existing argument mining system performs poorly (row 2); performance improvement when retraining the model on the current dataset indicates usefulness of features (row 3)
- Features from prior work on online dialogues are also useful in classroom discussions (row 4)
- Neural networks can be used to extract important features for argument component classification (row 11)
- Handcrafted features help increase performance of neural network models (rows 12 vs. 11, 20 vs. 19)
- Specificity information can further improve performance through multi-task learning (rows 18 vs. 10, 20 vs. 12)

Detailed analysis

- Pre-trained word embeddings are essential for LSTM models, while they do not always contribute to CNN models
- LSTM models benefit more from handcrafted features compared to CNN models
- Multi-task learning has higher impact on CNN than on LSTM

Future Directions

- Incorporate contextual information from previous argument moves.
- Include additional tasks in multi-task setting at training time

---

