Argument Component Classification for Classroom Discussions

Luca Lugini, Diane Litman
University of Pittsburgh

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 (wLDA)[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

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 doesnt know where hes at; he makes a wrong turn.	evidence	medium
S1	cause he tried doing it by himself and he cant. So I think Fezzik went back to his normal ways, like after he changed.	warrant	high

- Class distribution:

Argument Component		
Claim	Warrant	Evidence
1034 (50.5%)	358 (17.5%)	655 (32%)

Specificity		
Low	Med	High
710 (34.7%)	996 (46.7%)	341 (16.6%)

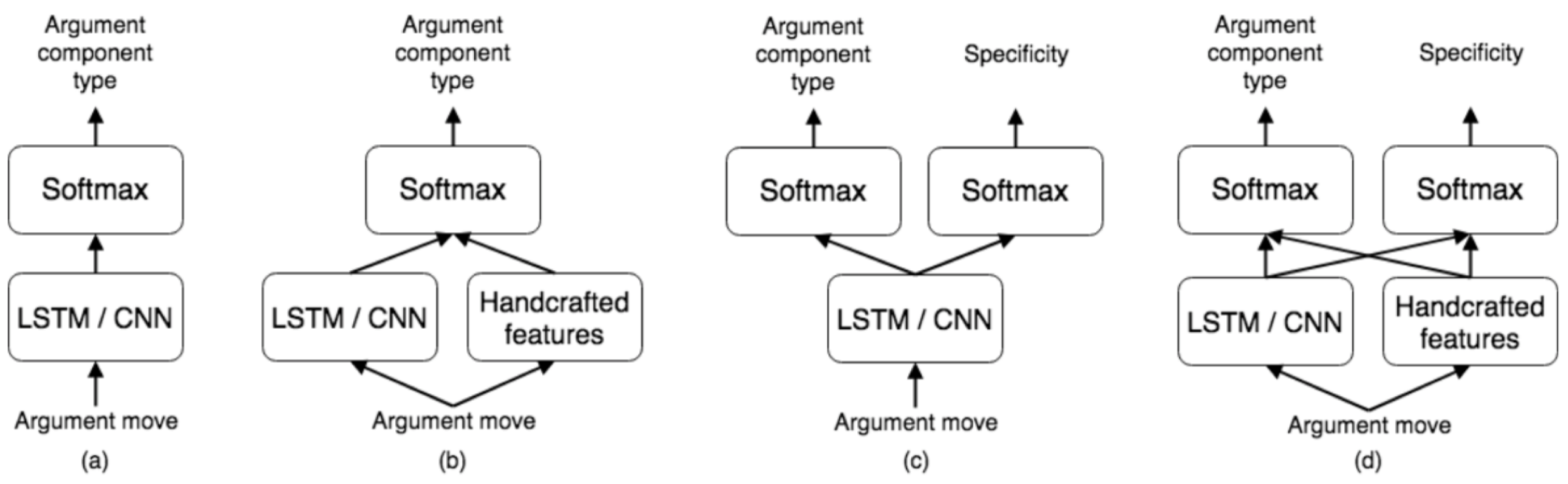
Models

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

Results obtained through leave-one-transcript-out cross validation

Row	Models / Features	Kappa	Precision	Recall	F-score	F_e	F_w	F_c
1	Majority baseline	0.068	0.265	0.406	0.314	0.109	0.004	0.532
2	Pre-trained wLDA	0.077	0.289	0.35	0.269	0.351	N/A	0.456
3	Logistic Regression (wLDA features)	0.142	0.412	0.394	0.379	0.39	0.211	0.54
4	Logistic Regression (wLDA features + online dialogue)	**0.283**	0.508	0.5	0.48	0.479	0.222	**0.693**
Word level NN models								
9	LSTM	0.069	0.408	0.399	0.218	0.161	0.198	0.295
10	LSTM + wLDA + online dialogue	0.181	0.462	0.447	0.391	0.362	0.279‡	0.522
11	CNN	0.125	0.41	0.404	0.378	0.37	0.231	0.526
12	CNN + wLDA + online dialogue	0.241⋆	0.492⋆	0.488	0.455†	0.468	0.276‡	0.622
Multi-task word level NN models								
17	LSTM	0.093	0.379	0.364	0.276	0.298	0.252	0.378
18	LSTM + wLDA + online dialogue	0.232	0.497†	0.482	0.44	0.419	0.299‡	0.583
19	CNN	0.164	0.351	0.443	0.441	0.476	0.249	0.598
20	CNN + wLDA + online dialogue	0.276‡	**0.521‡**	**0.512†**	**0.485†**	**0.484**	**0.312‡**	0.638

Best results are highlighted in bold. Statistically significance (*: 0.1 level, †: 0.05 level, ‡: 0.01 level) w.r.t. row 3. See paper for the complete table of results.

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

[1] Huy Nguyen and Diane Litman. 2016. Improving argument mining in student essays by learning and exploiting argument indicators versus essay topics, Proceedings 29th International FLAIRS Conference Conference: Literacy Research for Expanding Meaningfulness.
[2] Luca Lugini, Diane Litman, Amanda Godley, and Christopher Olshefski. 2018. Annotating Student Talk in Text-based Classroom Discussions, Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications.
[3] Reid Swanson, Brian Ecker, and Marilyn Walker. 2015. Argument mining: Extracting arguments from online dialogue. Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue.

This work was supported by the Learning Research and Development Center at the University of Pittsburgh