Neural Discourse Structure for Text Categorization

Yangfeng Ji and Noah Smith

Presented by Ji Li

Ji, Yangfeng, and Noah Smith. "Neural discourse structure for text categorization." arXiv preprint arXiv: 1702.01829 (2017).

Outline

- Motivation
- Background
- Model
- Experiment results
- Discussion and conclusion

Outline

- Motivation
- Background
- Model
- Experiment results
- Discussion and conclusion

Text Categorization

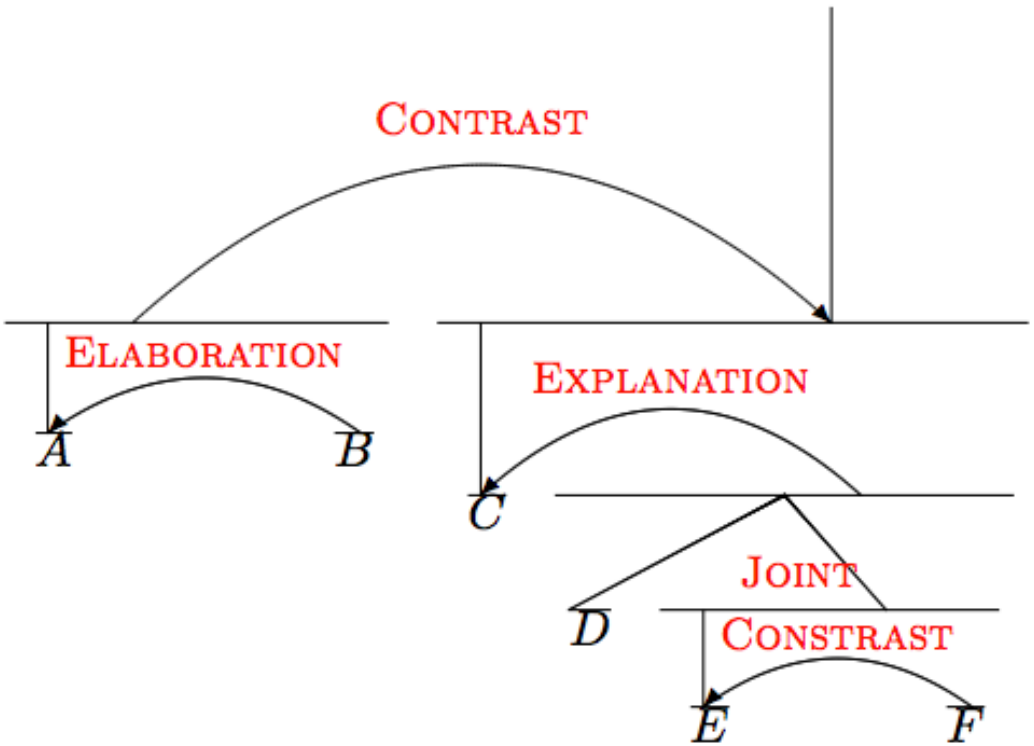
- Important for applications like sentiment analysis, inferring authorship and making predictions
- How to represent text?
  - Bag of words
  - Representation learning

- How does different part of text contribute to the representation?
  - Different part of text may influence the text category in different weight

Outline

- Motivation
- Background
- Model
- Experiment results
- Discussion and conclusion

Rhetorical Structure Theory



[Although the food was amazing]^{A} [and I was in love with the spicy pork burrito,]^{B} [the service was really awful.]^{C} [We watched our waiter serve himself many drinks.]^{D} [He kept running into the bathroom]^{E} [instead of grabbing our bill.]^{F}

Document can be represented by a tree.

Elementary Discourse Unit (EDU), connected via discourse relations.

Discourse Parsing from Linear Projection

- Shift reduced discourse parsing
  - Maintain a stack and a queue
  - Consider features of 3 EDUs, *v = [v1; v2; v3]*

- Formulate the decision as a multi-class shift-reduce classifier

$$\hat{m} = \underset{m \in \{1,\ldots,C\}}{\arg\max}\ \mathbf{w}_m^\top \mathbf{f}(\mathbf{v}; \mathbf{A})$$

Ji, Yangfeng, and Jacob Eisenstein. "Representation learning for text-level discourse parsing." Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vol. 1. 2014.

Outline

- Motivation
- Background
- Model
- Experiment results
- Discussion and conclusion

Model

- Recursive neural network build on discourse dependency tree
- Distributed representation of EDU achieved with bidirectional LSTM
- Let *vi* denote the vector representation of EDU *i* and its descendants
  - If *vi* is leaf: *vi = tanh(ei)*
  - Other wise:

$$\mathbf{v}_i = \tanh\left(\mathbf{e}_i + \sum_{j \in children(i)} \alpha_{i,j} \mathbf{W}_{r_{i,j}} \mathbf{v}_j\right)$$

$$\alpha_{i,j} = \sigma\left(\mathbf{e}_i^\top \mathbf{W}_\alpha \mathbf{v}_j\right)$$

Model ablations and baselines

- Unlabeled Model

$$\mathbf{v}_i = \tanh\left(\mathbf{e}_i + \sum_{j \in children(i)} \alpha_{i,j}\mathbf{v}_j\right)$$

- Root
  - Use discourse structure to select the root and use only the root EDU representation
- Additive
  - Use all EDUs with simple average

Outline

- Motivation
- Background
- Model
- Experiment results
- Discussion and conclusion

Performance

Method	Yelp	MFC	Debates	Movies	Bills
Prior work					
1. Yang et al. (2016)	71.0	—	—	—	—
2. Card et al. (2016)	—	56.8	—	—	—
3. Yogatama and Smith (2014)	—	—	74.0	—	88.5
4. Bhatia et al. (2015)	—	—	—	82.9	—
5. Hogenboom et al. (2015)	—	—	—	71.9	—
Variants of our model					
6. ADDITIVE	68.5	**57.6**	69.0	82.7	80.1
7. ROOT	54.3	51.2	60.3	68.7	70.5
8. UNLABELED	**71.3**	**58.4**	**75.7**	**83.1**	78.4
9. FULL	**71.8**	56.3	**74.2**	79.5	77.0

Table 2: Test-set accuracy across five datasets. Results from prior work are reprinted from the corresponding publications. Boldface marks performance stronger than the previous state of the art.

Performance

Method	Yelp	MFC	Debates	Movies	Bills
Prior work					
1. Yang et al. (2016)	71.0	—	—	—	—
2. Card et al. (2016)	—	56.8	—	—	—
3. Yogatama and Smith (2014)	—	—	74.0	—	88.5
4. Bhatia et al. (2015)	—	—	—	82.9	—
5. Hogenboom et al. (2015)	—	—	—	71.9	—
Variants of our model					
6. ADDITIVE	68.5	**57.6**	69.0	82.7	80.1
7. ROOT	54.3	51.2	60.3	68.7	70.5
8. UNLABELED	**71.3**	**58.4**	**75.7**	**83.1**	78.4
9. FULL	**71.8**	56.3	**74.2**	79.5	77.0

Table 2: Test-set accuracy across five datasets. Results from prior work are reprinted from the corresponding publications. Boldface marks performance stronger than the previous state of the art.

Performance

Method	Yelp	MFC	Debates	Movies	Bills
Prior work					
1. Yang et al. (2016)	71.0	—	—	—	—
2. Card et al. (2016)	—	56.8	—	—	—
3. Yogatama and Smith (2014)	—	—	74.0	—	88.5
4. Bhatia et al. (2015)	—	—	—	82.9	—
5. Hogenboom et al. (2015)	—	—	—	71.9	—
Variants of our model					
6. ADDITIVE	68.5	**57.6**	69.0	82.7	80.1
7. ROOT	54.3	51.2	60.3	68.7	70.5
8. UNLABELED	**71.3**	**58.4**	**75.7**	**83.1**	78.4
9. FULL	**71.8**	56.3	**74.2**	79.5	77.0

Table 2: Test-set accuracy across five datasets. Results from prior work are reprinted from the corresponding publications. Boldface marks performance stronger than the previous state of the art.

Case Study

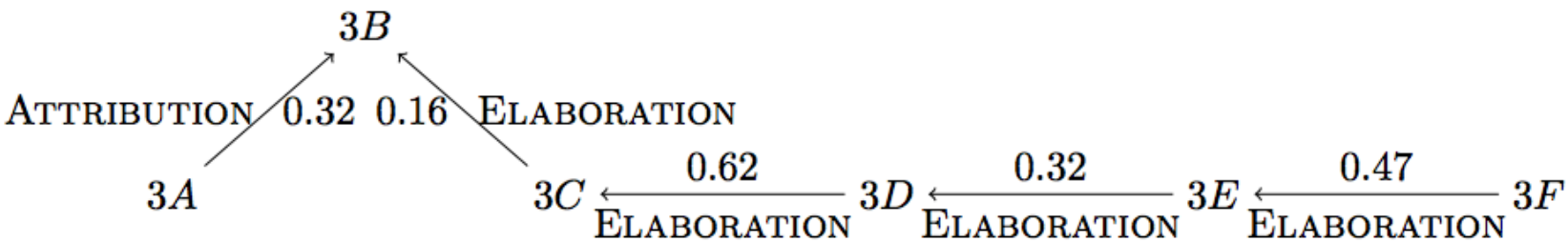
From DPLP:

$$2C \xrightarrow[\text{EVALUATION}]{0.70} 2B \xrightarrow[\text{EVALUATION}]{0.87} 2A \xleftarrow[\text{ELABORATION}]{0.61} 2D$$

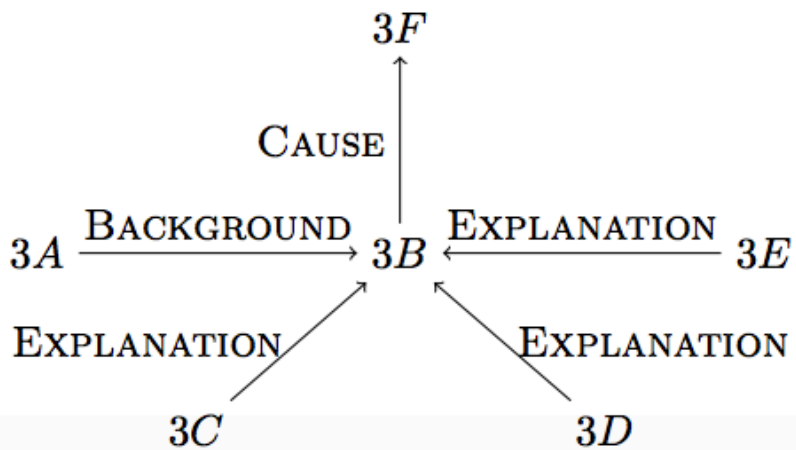
[I love these people.]^{2A} [They are very friendly and always ask about my life.]^{2B} [They remember things I tell them then ask about it the next time I'm in.]^{2C} [Patrick and Lily are the best but everyone there is wonderful in their own ways.]^{2D}

(b) true label: 5, predicted label: 5

From DPLP:

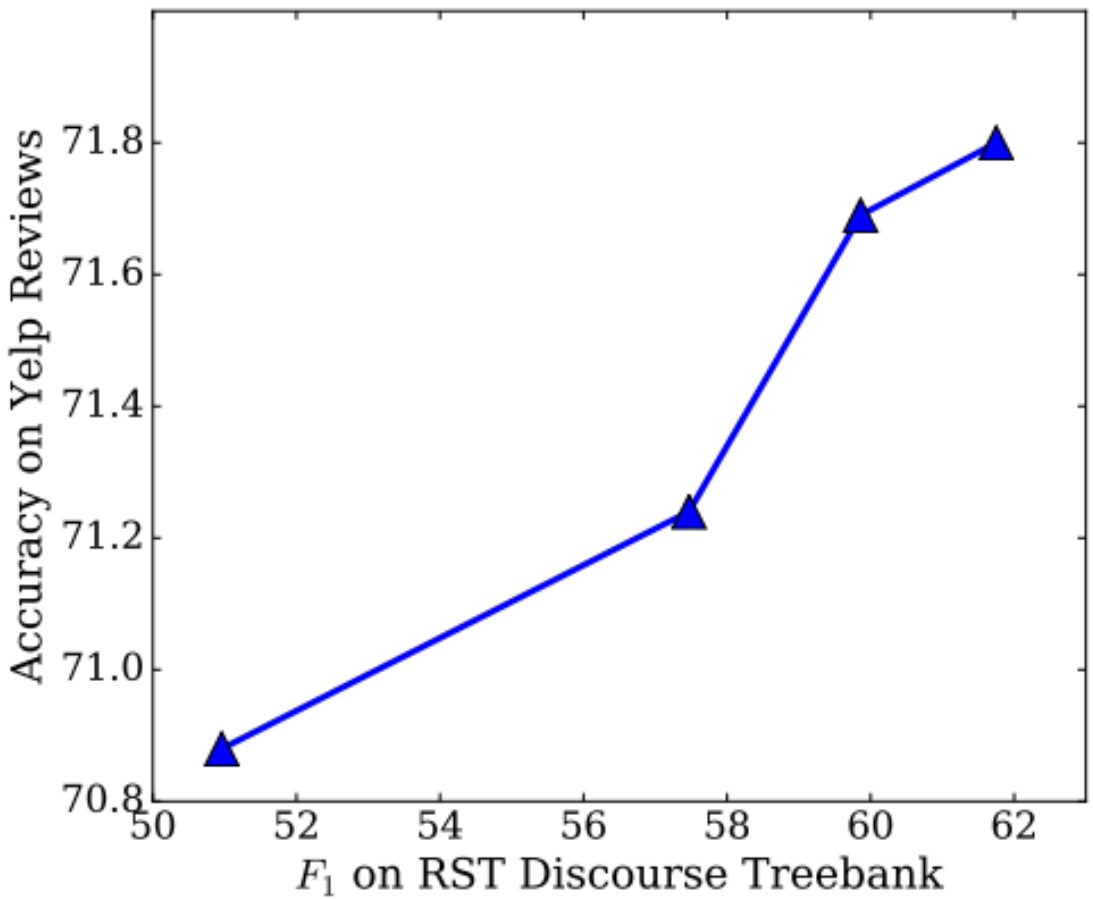


Manually constructed:



[We use to visit this pub 10 years ago because they had a nice english waitress and excellent fish and chips for the price.]^{3A} [However we went back a few weeks ago and were disappointed.]^{3B} [The price of the fish and chip dinner went up and they cut the portion in half.]^{3C} [No one assisted us in putting two tables together we had to do it ourselves.]^{3D} [Two guests wanted a good English hot tea and they didn't brew it in advance.]^{3E} [So we've decided there are newer and better places to eat fish and chips especially up in north phoenix.]^{3F}

Other experiment



Further improvement on Discourse structure can potentially enhance performance.

Outline

- Motivation
- Background
- Model
- Experiment results
- Discussion and conclusion

Discussion and Conclusion

- Automatically-derived discourse structure can be helpful to text categorization.

- Two major limitations:
  - Performance of discourse parser.
  - Domain mismatch between training corpus and future test data.

- Motivates further improvement of discourse parser and exploration of domain adaptation methods.

Thank you