Neural Discourse Structure for Text Categorization

Yangfeng Ji and Noah A. Smith. 2017.

Task: document categorization

How to construct a meaningful document representation?

- Previous work: all sentences are weighted **equally** and/or use **hand-crafted** weighting schemes
- Can we do better?
- Hypothesis: we can **deeply model** the **relative salience** of a document's sentence by exploiting **discourse structure**

Key research questions:

- What is the value of discourse structure for neural text categorization?

Paper overview

Primary experiments:

- **Text categorization** across 5 corpora
- Domains include **sentiment analysis on movie/restaurant reviews, congressional debates,** and **congressional bills**

Primary contributions of authors:

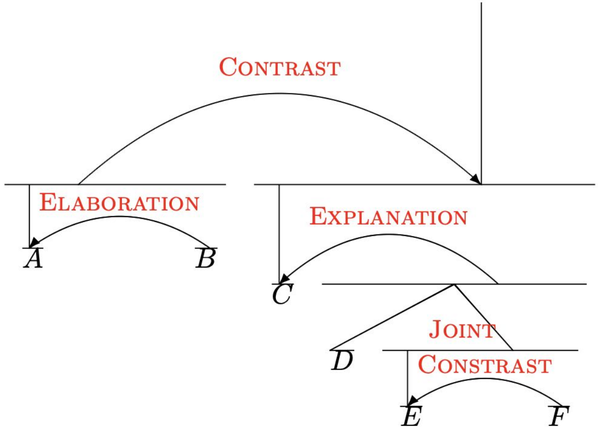
- Exploit **discourse structure** to improve neural text categorization
- **Recursive** neural architecture for handling documents represented as **discourse trees**
- Novel **attention mechanism** to learn **importance** of document's sentences based on **relational structure**

Outline

- **Background: Rhetorical Structure Theory**
- Models
- Experimental setup
- Results & discussion

Background: Rhetorical Structure Theory (RST)

- A document can be represented as a **tree**
- Each node is an **elementary discourse unit** (EDU)
- **Spans** between nodes represent discourse **relations**



[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}

Key concept: leveraging **tree structure** can offer **inductive bias**

- Model can more easily discern **salient** parts of a text
- Documents parsed by open-source parser;
  RST trees are transformed to dependency structures

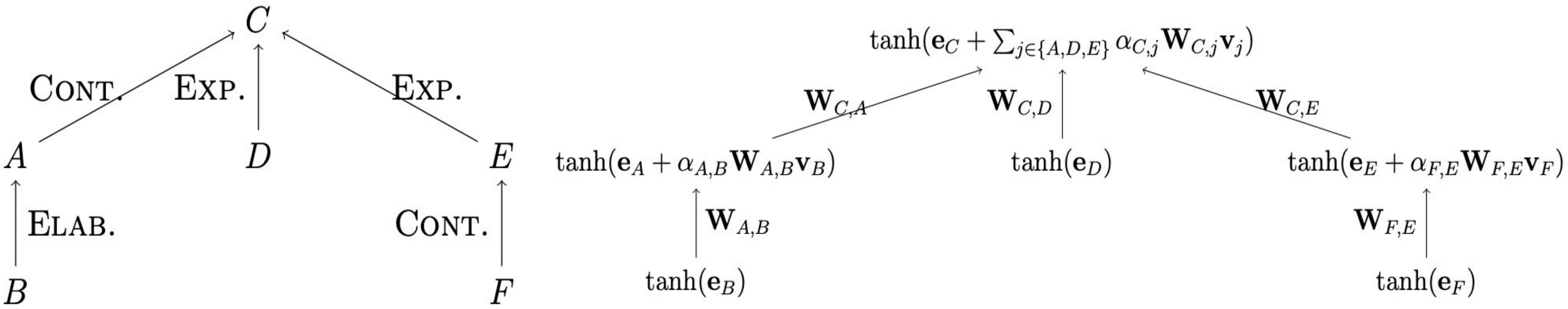
Outline

- Background: Rhetorical Structure Theory
- Models
- Experimental setup
- Results & discussion

Model

- **Model input**: discourse dependency tree
- **Bidirectional LSTM** to obtain a distributed sentence representation $\mathbf{e}_i$ for each clause
- Construct document representation by **recursively composing** node representation $\mathbf{v}_i$:
    - If EDU is a **leaf** in the tree: $\mathbf{v}_i = \tanh(\mathbf{e}_i)$
    - If EDU is a **parent node**: $\mathbf{v}_i = \tanh\left(\mathbf{e}_i + \sum_{j \in children(i)} \alpha_{i,j} \mathbf{W}_{r_{i,j}} \mathbf{v}_j\right)$
    - Where $\alpha_{i,j}$ represents an **attention mechanism**: $\alpha_{i,j} = \sigma\left(\mathbf{e}_i^\top \mathbf{W}_\alpha \mathbf{v}_j\right)$
    - Note: $\alpha_{i,j}$ **independent** of other children of parent node!

Recursive model: Visual Overview



(a) dependency structure (b) recursive neural network structure

Prediction is obtained by a softmax on $(\mathbf{W}_o\mathbf{v}_{root} + \mathbf{b})$

Model variants

Main idea: gradually introduce more **components** to model in order to measure **benefit** of discourse

1. **ROOT**: Only select the **root** EDU; no usage of composition function. $\mathbf{v}_{root} = \mathbf{e}_{root}$.
2. **ADDITIVE**: Take the **average** of all distributed representations: $\mathbf{v}_{root} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{e}_i$,
3. **UNLABELLED**: no composition matrix $\mathbf{W}_{r_{i,j}}$; only **attention**: $\mathbf{v}_i = \tanh\left(\mathbf{e}_i + \sum_{j \in children(i)} \alpha_{i,j}\mathbf{v}_j\right)$
4. **FULL**: $\mathbf{v}_i = \tanh\left(\mathbf{e}_i + \sum_{j \in children(i)} \alpha_{i,j}\mathbf{W}_{r_{i,j}}\mathbf{v}_j\right)$

Implementation Details

Discourse Parsing

- **Discourse structure** for each document obtained via use of open-source RST parser **DPLP**
- DPLP is trained on 347 *Wall Street Journal* articles from Penn Treebank
- RST trees are converted to **discourse dependency trees**

Models

- Pretrained *GloVe* embeddings for bidirectional LSTM
- Randomly initialized embeddings for the larger corpora
- SGD/Adam for optimization
- Grid search for LSTM dimensionality and learning rate

Outline

- Background: Rhetorical Structure Theory
- Models
- Experimental setup
- Results

Tasks & Data

Five different datasets, with four different tasks

			Number of docs.				
Dataset	Task	Classes	Total	Training	Development	Test	Vocab. size
Yelp	Sentiment	5	700K	650K	–	50K	10K
MFC	Frames	15	4.2K	–	–	–	7.5K
Debates	Vote	2	1.6K	1,135	105	403	5K
Movies	Sentiment	2	2.0K	–	–	–	5K
Bills	Survival	2	52K	46K	–	6K	10K

Outline

- Background: Rhetorical Structure Theory
- Models
- Experimental setup
- **Experimental results**

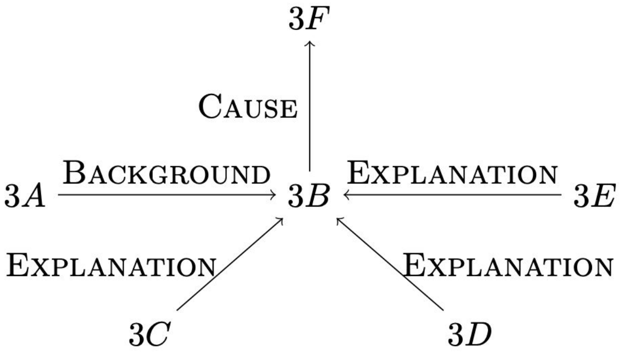
Quantitative Results

- **UNLABELED** outperforms previous SOTA on four out of five tasks
- **FULL** has best performance on Yelp; comparatively poor on other tasks
- **ADDITIVE** performs best on Bills, somewhat close results to **UNLABELED** on MFC and Movies
- **ROOT** has poor performance across the board

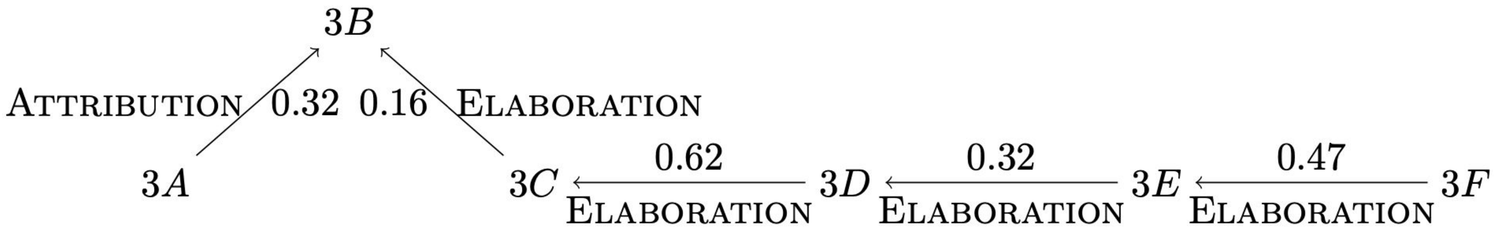
Method	Yelp	MFC	Debates	Movies	Bills
Prior work					
1. Yang et al. (2016)	71.0	—	—	—	—
2. Card et al. (2016)	—	56.8	—	—	—
3. Yogatama and Smith (2014)	—	—	74.0	—	88.5
4. Bhatia et al. (2015)	—	—	—	82.9	—
5. Hogenboom et al. (2015)	—	—	—	71.9	—
Variants of our model					
6. ADDITIVE	68.5	**57.6**	69.0	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

Qualitative results - Parser can inhibit performance

[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}



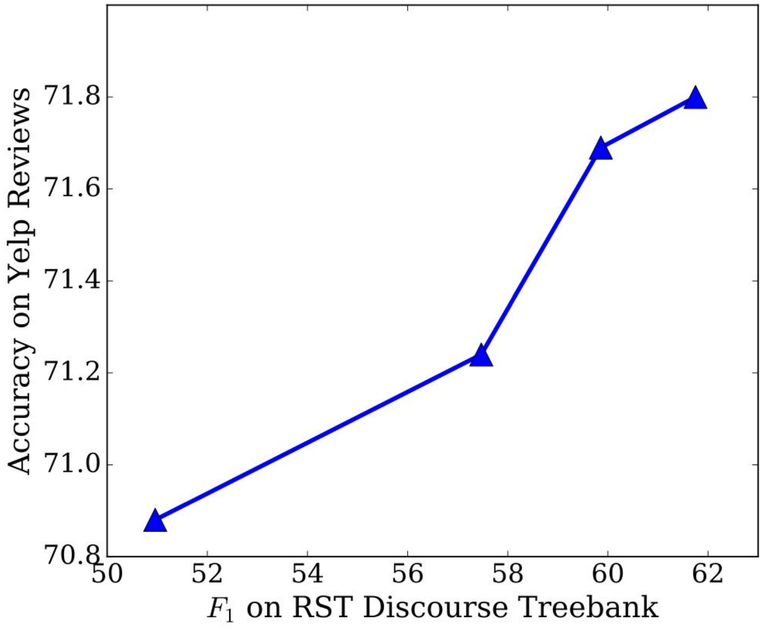
From DPLP:



Exploring the effect of parsing performance

Degrading DPLP to observe effect on classification performance

- Authors train **FULL** model with **DPLP** trained on only **25%, 50%**, and **75%** of its training set
- **Plot**: discourse parser performance (x-axis) against text classifier performance (y-axis)
- **Lower** parsing performance implies **lower** classification accuracy
- Further improvements to parsing - **better models**?



Contrasting un-normalized attention

- The authors propose an **un-normalized** attention layer, “inspired by RST’s lack of “competition” for salience among satellites”

- How does this compare with **normalized** attention, which is common in **machine translation**?

- Here, $\boldsymbol{\alpha}'_i$ is a vector with one element for each child node, which **sums to one**:

$$\boldsymbol{\alpha}'_i = \mathrm{softmax}\left(\begin{bmatrix} \vdots \\ \mathbf{v}_j^\top \\ \vdots \end{bmatrix}_{j \in children(i)} \mathbf{W}_\alpha \cdot \mathbf{e}_i\right)$$

- On Yelp data, the **FULL** model achieves 70.3% accuracy - 1.5% less compared to the **FULL** model with un-normalized attention

- Authors: “empirical support for theoretically-motivated design decision not to normalize attention.”

Summary

- Empirical evidence that discourse structure can benefit text categorization
- Extensive analysis of benefits of incorporating more discourse structure information
- Brief empirical study of dependence of model performance on discourse parser performance
- Some additional empirical support for un-normalized attention mechanism

My Opinion

- **Novel** approach w.r.t previous work in sentence weighting; well-explained paper overall
- **Ablation study** offers some interesting insight on how the different components affect performance
- **Promising results**, albeit somewhat ambiguous, due to model dependence on underlying **DPLP** parser
- Dependence on parser suggests **limited potential** for domains with different discourse structure
- No reporting of **hyperparameters** for each model other than mentioning grid-search
- No **significance testing** despite small differences between previous SOTA
- **Parsing degradation** is only tested with Yelp dataset, only on **FULL** model
- Un-normalized attention mechanism is only contrasted on the Yelp dataset and only on **FULL** model
- Contrasting on both **FULL** and **UNLABELED** architectures, across all tasks, would have made for **stronger** evidence for un-normalized attention mechanism

Possible angles for future research

- **Domain adaptation** methods to overcome **mismatches** between parser training corpus and domain of interest

- Explore to what extent further improvements to RST parsing would translate to gains in text categorization

- **UNLABELED** was the most consistent model variant

    - No concept of relations; still a relatively simple interpretation of discourse structure?
    - Further work could explore ways to fully leverage the **rich representational structure** of RST (for instance, by use of larger datasets and/or less parameters to avoid **overparameterization**)

Questions?

Thanks for your attention!