Feature Rich Models for Discourse Signaling

Amir Zeldes

Georgetown University
amir.zeldes@georgetown.edu

CLSP Seminar, JHU, 2018-02-23
Plan

I. Discourse Relations
   - What are discourse relations?
   - Rhetorical Structure Theory (RST) in a nutshell
   - Datasets – RST-DT and GUM

II. Relation Signaling in Textual Data
    - Explicit and implicit relations
    - The RST Signalling Corpus
    - Finding signals in a text based model

III. Rich features
    - Should we add annotations to embeddings?
    - Ablation studies with feature rich models
Discourse relations

- What relations exist between utterances as a text unfolds?

1. a. [[John pushed Mary.]_{cause} She fell.]
   b. [Mary fell. [John pushed her.]_{cause}]
   (see Webber 1988, Asher & Lascarides 2003)

2. [[[They left lights on]_{cause} so Ellie got mad.]_{evaluation} [That’s totally unreasonable]]
Discourse relations

- Some questions:
    - Cross-linguistically? (van der Vliet & Radeker 2014)
    - In genres? (Taboada & Lavid 2003)
  - How are relations marked? (Taboada & Das 2013)
    - Explicit signals: “on the other hand” or “although”
    - Implicit signals: coreferent mentions, genre conventions, ...

To answer these questions we build discourse annotated corpora
Discourse annotation

- The task – given an arbitrary text:
  - Segment into ‘units’ (a.k.a. Elementary Discourse Units)
  - Establish the connections between these EDUs
  - Classify these connections

- Three main frameworks have implemented these tasks:
  - Penn Discourse Treebank (PDTB, Prasad et al. 2008) – partial parses
  - Segmented Discourse Representation Theory (SDRT, Asher & Lascarides 2003) – complete DAGs
  - Rhetorical Structure Theory (Mann & Thompson 1988) – complete trees
In RST, a text is a tree of clauses

**Syntax trees**
- head > expansion
- Leaf = token
- Non-terminal = phrase
- Grammatical function

**RST trees**
- nucleus > satellite
- Leaf = EDU
- Non-terminal = span
- Discourse function
Trying it out - what are these?

- Find the direction and label – choose from:
  - cause
  - purpose
  - elaboration
  - concession
Why is this important?

- Get most important unit (Summarization)
- Identify specific relations (IR)
- Build discourse plan (NLG)

(example from RST Website: [http://www.sfu.ca/rst/](http://www.sfu.ca/rst/))
Topics for today

- What information identifies relations?
  - For humans
  - For NLP

- Can we identify relations directly from text?
  - Do machine learning algorithms and humans notice the same signals?
  - If/when not, why? What features do we miss?
  - Can we add them as new layers to our corpus data?

- What data can we use?
RST Discourse Treebank (Carlson et al. 2003)

- 180K tokens (WSJ)
- POS + syntax trees
- 60% overlap with OntoNotes: (Hovy et al. 2006)
  - NER (named entities only)
  - Partial coreference (no singletons, indefinites)
  - PropBank annotations
Georgetown University Multilayer corpus (Zeldes 2017)

http://corpling.uis.georgetown.edu/gum/

- POS tagging (PTB, CLAWS, TT, UPOS)
- Sentence type (SPAAC++)
- Document structure (TEI)
- Syntax trees (PTB + Stanford + UD)
- Information status (SFB632)
- (Non-) named entity types
- Coreference + bridging
- **Rhetorical Structure Theory**
- Speaker information, ISO time…

<table>
<thead>
<tr>
<th>text type</th>
<th>source</th>
<th>texts</th>
<th>tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>Academic</td>
<td>Various</td>
<td>6</td>
<td>5,210</td>
</tr>
<tr>
<td>Biographies</td>
<td>Wikipedia</td>
<td>6</td>
<td>5,049</td>
</tr>
<tr>
<td>Fiction</td>
<td>Small Beer Press</td>
<td>7</td>
<td>5,912</td>
</tr>
<tr>
<td>Interviews</td>
<td>Wikinews</td>
<td>19</td>
<td>18,037</td>
</tr>
<tr>
<td>News</td>
<td>Wikinews</td>
<td>21</td>
<td>14,093</td>
</tr>
<tr>
<td>Travel guides</td>
<td>Wikivoyage</td>
<td>17</td>
<td>14,955</td>
</tr>
<tr>
<td>Forum discussions</td>
<td>reddit</td>
<td>6</td>
<td>5,174</td>
</tr>
<tr>
<td>How-to guides</td>
<td>wikiHow</td>
<td>19</td>
<td>16,920</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td>101</td>
<td>85,350</td>
</tr>
</tbody>
</table>
II. Relation Signaling
Explicit signals

  - Discourse markers – however, but, if, and, as well as
  - Adverbials – clearly, supposedly, in reality...
  - Content words – good (signals evaluation?), last year (signals temporal sequence? Circumstance?)

- Annotators use cue words as diagnostics:
  - “could I connect these with ‘because’?”
Frequentist approaches

- Studies often cross-tabulate: \textit{words} \sim \textit{relations}

- Problems:
  - Frequency thresholds
  - Ambiguity (“and” may not be associated with relations and appears with all relations – not a Discourse Marker?)
  - Context sensitivity – some words are cues in specific environments
The core idea of our work is to learn a transformation from a bag-of-words surface representation into a latent space in which discourse relations are easily identifiable.

- Ji & Eisenstein (2014:13)

Echoed in much NLP in recent years:
- text -> labels
- But really:
  text -> embeddings <-> labels
  (cf. Braud et al. 2016)
Relation classification with RNNs

- RNNs can recognize relations from text (Braud et al. 2017; cf. entailment work, Rocktäschel et al. 2016)
- Can use encoder architecture, single output multinomial classifier

If we were fish

What were the signals?

condition
cause
contrast
A neural approach to signals with RNNs

- The RNN probably already had it at *If*...
- To find signals, we can listen to output at every token (but loss still based on EDU relation)
Implemented with Bi-LSTM (TensorFlow)

Character embeddings
Word embeddings (GloVe 300, Pennington et al. 2014)

if
we
were
fish

Hidden: 200
Optimizer: Adam
(rec.) dropout: 0.5
Minibatch: 20
Activation: tanh

Batch normalization
Trainable embeddings

CLSP Seminar, JHU
Adding CRF (Huang et al. 2015, Ma & Hovy 2016)

- **CRF**
- **Backward LSTM**
- **Forward LSTM**
- **Character embeddings**
- **Word embeddings** (GloVe 300, Pennington et al. 2014)

- Hidden: 200
- Optimizer: Adam
- (rec.) dropout: 0.5
- Minibatch: 20
- Activation: tanh
- Batch normalization
- Trainable embeddings

*Feature Rich Models for Discourse Signaling*
Single output performance

- Not so interesting, but:
  - RSTDT – relation accuracy by tokens:
    acc: 47.43% | f1: 41.44
    • Standard train/test split
    • 60 relations [some very rare] – note majority baseline is ~33%

- State of the art on RSTDT, hard to compare:
  - Ji & Eisenstein (2014), using engineered features, full parsing: 61.75% (by EDUs, 18 relations)
  - Braud et al. (2016), (2017) with RNNs, pretraining on PDTB, coref and more:
    60.01% (by EDUs, 18 relations)
Visualizing token-wise softmax

- Basic idea – find the most ‘convincing’ tokens:
  - Use tokens' softmax probability of correct relation
  - Shade by:
    • Proportion of maximum softmax probability in sentence
    • Proportion of maximum softmax probability in document

How good am I in sent?  
How good am I in doc?

Feature Rich Models for Discourse Signaling
Visualizing token-wise softmax

- **This** occurs for two reasons:
  - **preparation** As it moves over land,
  - **circumstance** it is cut off from the source of energy driving the storm ...
- **Combine 50 milliliters of hydrogen peroxide and a liter of distilled water in a mixing bowl.**
- **A ceramic bowl will work best,** but plastic works too.

Ambigious?

GUM data
Addressing ambiguity

- We can get ambiguity scores based on range of softmax probabilities (data: GUM)
Addressing ambiguity

- **Irrelevant ‘and’s: (RST-DT)**
  - [but will continue as a director and chairman of the executive committee. ]_elaboration_
  - [and one began trading on the Nasdaq/National Market System last week. ]_inverted_

- **Important ‘and’s: (RST-DT)**
  - [and is involved in claims adjustments for insurance companies. ]_List_
  - [-- and from state and local taxes too, for in-state investors. ]_elaboration_
Evaluating plain text signals

- There results are qualitative, non-systematic
- Ideal scenario - compare to ‘gold standard’
  - Use RST-DT Signalling Corpus (Taboada & Das 2013)
  - Open ended annotation of any kind of relation signal:
    - Discourse markers, other expressions
    - Syntactic devices, cohesion
    - Genre conventions...
Evaluating signals

- Problems:
  - Signals annotated at **node** level
  - Non trivial to associate with specific EDUs
  - Location of signal in words is not specified

Feature Rich Models for Discourse Signaling

Signalling Corpus in UAM (O’Donnell 2008)
Toy evaluation

- 3 documents from Signalling Corpus (RST-DT/test)
  - 113 EDUs
  - 210 nodes
  - 153 signals manually inspected
    - Only 83 attributable to a/some tokens (not, e.g.: genre, zero relative, graphical layout...)
      
      In a remark [someone should remember this time next year,]
    
    - Only 47 reasonably detectable by net (not, e.g.: lexical chain, syntactic parallelism)
      
      Congress gave Senator Byrd's state ... [Senator Byrd is chairman..]
Results

- Network ranks all words (low precision if 0 signals)
- Use recall rate @k to evaluate

![Graphs comparing RNN and chance for different metrics and signal types](image-url)
III. Feature rich models
Can we get at ‘non-resolvable’ cases?

- A plain text RNN can’t see many things:
  - Repetition
    - Lexical entity coreference
    - Pronoun resolution
    - Restatements…
  - Non-token signals
    - Syntax clause types and attachment
    - Zero relatives, other ‘meaningful absences’
  - Genre (is that ‘inside’ the text already?)
  - Graphical layout (images, fonts, headings, …)
  - …
Adding annotations to vectors

feature rich models for discourse signaling

if we were fish

elab elab elab elab

CRF

elab

elab

elab

elab

LSTM LSTM LSTM LSTM

Backward LSTM

elab

elab

elab

elab

LSTM LSTM LSTM LSTM

Forward LSTM

elab

elab

elab

elab

LSTM LSTM LSTM LSTM

Character embeddings

Word embeddings

Annotation embeddings

IN PRP VBD NNS

if we were fish

CRF

Backward LSTM

Forward LSTM

Character embeddings

Word embeddings

Annotation embeddings

IN PRP VBD NNS

if we were fish

CLSP Seminar, JHU
Adding annotations to vectors

<table>
<thead>
<tr>
<th>F-Score</th>
<th>text</th>
<th>genre</th>
<th>s_type</th>
<th>func</th>
<th>pos</th>
<th>entity</th>
<th>coref</th>
<th>layout</th>
<th>all</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>35</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>40</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>45</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Baseline: 25.58

Feature Rich Models for Discourse Signaling

CLSP Seminar, JHU
Adding annotations to vectors

<table>
<thead>
<tr>
<th>Feature</th>
<th>text</th>
<th>genre</th>
<th>s_type</th>
<th>func</th>
<th>pos</th>
<th>entity</th>
<th>coref</th>
<th>layout</th>
<th>all</th>
</tr>
</thead>
<tbody>
<tr>
<td>test f1</td>
<td>35.57</td>
<td>36.65</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Baseline: 25.58
Genres vary significantly in communicative means

Prior likelihoods of relations vary:

Quiz: guess which!
- Academic
- Bio
- Fiction
- Interview
- News
- Reddit
- Voyage
- Wikihow
Plain:
[1 teaspoon baking powder]_joint

+Genre: (whow)
[1 teaspoon baking powder]_joint

Plain:
[It has lots of local boutiques…]_elab

+Genre: (voyage)
[It has lots of local boutiques…]_elab

Plain:
[I don’t like the doctor , ]_elab

+Genre: (fiction)
[I don’t like the doctor , ]_eval

wikiHow: How to Make Vegan Cupcakes

Wikivoyage: Oakland

“Oversite” by Maureen F. McHugh
Adding annotations to vectors

<table>
<thead>
<tr>
<th>F-Score</th>
<th>text</th>
<th>genre</th>
<th>s_type</th>
<th>func</th>
<th>pos</th>
<th>entity</th>
<th>coref</th>
<th>layout</th>
<th>all</th>
</tr>
</thead>
<tbody>
<tr>
<td>test f1</td>
<td>35.57</td>
<td>36.65</td>
<td>37.26</td>
<td>37.36</td>
<td>38.12</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
POS and dependency function

The same strings can mean different things:

- *meaning/NN* is self-contained within the text
- *meaning/VVG* as a first strike weapon
(cf. also ‘like’)

Similarly for grammatical function:

He reemerged in *September 1859* …

Plain:

[laying claim to the position of Emperor of the United States. ]\text{seq}

+Deprel:

[laying claim to the position of Emperor of the United States. ]\text{seq}
## Adding annotations to vectors

![Graph showing F-Scores](image)

### Baseline: 25.58

<table>
<thead>
<tr>
<th>Feature</th>
<th>Text F1</th>
<th>Genre</th>
<th>s_type</th>
<th>func</th>
<th>pos</th>
<th>entity</th>
<th>coref</th>
<th>layout</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test f1</td>
<td>35.57</td>
<td>36.65</td>
<td>37.26</td>
<td>37.36</td>
<td>38.12</td>
<td>37.14</td>
<td>39.14</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Feature Rich Models for Discourse Signaling

CLSP Seminar, JHU
Coreference and entities

- Relationship between referential accessibility and RST graph (Veins Theory, Cristea et al. 1998)
- Coreference likelihood can be predicted by discourse parse (Zeldes 2017b)
Coreference and entities

- Again, different priors:

Coref and entity resolution:
- Know pronoun entities
- Mentioned in RST parent?

Plain:
[based on the knowledge and skills they feel librarians need ;]_{elab}

+Coref+Entities:
[based on the knowledge and skills they person feel librarians need ;]_{elab}
Adding annotations to vectors

<table>
<thead>
<tr>
<th></th>
<th>text</th>
<th>genre</th>
<th>s_type</th>
<th>func</th>
<th>pos</th>
<th>entity</th>
<th>coref</th>
<th>layout</th>
<th>all</th>
</tr>
</thead>
<tbody>
<tr>
<td>test f1</td>
<td>35.57</td>
<td>36.65</td>
<td>37.26</td>
<td>37.36</td>
<td>38.12</td>
<td>37.14</td>
<td>39.14</td>
<td>39.14</td>
<td>39.14</td>
</tr>
</tbody>
</table>

Baseline: 25.58
Graphical layout

- We have TEI XML tags for:
  - Paragraphs
  - Headings
  - Images and captions
  - Ordered / unordered lists
  - Beginning / end of list items
  - ...

```
<list type="ordered">
  <item n="1">
    <head>
      <s type="other">
        Method NN method
        One CD One
        of IN of
        Two CD Two
        : : :
        ...
      </s>
    </head>
  </item>
</list>
```
Plain:

[For this question I do n't know the ' preparedness ' of the Indian gov't to deal with this . ]_joint

+Layout annotations:

[For this question I do n't know the ' preparedness ' of the Indian gov't to deal with this . ]_joint

Plain:

[Listen up , kids :]_prep

+Layout annotations:

[Listen up , kids : ]_prep
### Adding annotations to vectors

<table>
<thead>
<tr>
<th></th>
<th>text</th>
<th>genre</th>
<th>s_type</th>
<th>func</th>
<th>pos</th>
<th>entity</th>
<th>coref</th>
<th>layout</th>
<th>all</th>
</tr>
</thead>
<tbody>
<tr>
<td>test f1</td>
<td>35.57</td>
<td>36.65</td>
<td>37.26</td>
<td>37.36</td>
<td>38.12</td>
<td>37.14</td>
<td>39.14</td>
<td>39.14</td>
<td>43.07</td>
</tr>
</tbody>
</table>

**Baseline:** 25.58
Can we get everything from text?

- Maybe not:
  - Humans use more than just text
  - Some things don’t ‘anchor’ well to text (text!=embeddings)
  - Sometimes text is identical – but other categories matter
  - More than text may be more efficient either way
Conclusion

- Good times to be working on discourse!
- Multiple layers expose complex interdependencies
- Older ideas in computational discourse models are now more feasible:
  - From co-occurrence statistics to contextualized RNN outputs
  - Integrating cues from different levels without overfitting

- We still need new data and new learning approaches!
but I just wanted to say **thank you** for your work and thank you for making some sense of the successes and failures and I wish you much success with your work.
References

Thanks to GUM annotators! (so far)

- Adrienne Isaac
- Akitaka Yamada
- Amani Aloufi
- Amelia Becker
- Andrea Price
- Andrew O'Brien
- Anna Runova
- Anne Butler
- Ayan Mandal
- Brandon Tullock
- Brent Laing
- Candice Penelton
- Chenyue Guo
- Colleen Diamond
- Connor O'Dwyer
- Dan Simonson
- Didem Ikizoglu
- Edwin Ko
- Emily Pace
- Emma Manning
- Han Bu
- Hang Jiang
- Hanwool Choe
- Hassan Munshi
- Ho Fai Cheng
- Jakob Prange
- Jehan al-Mahmoud
- Jemm Excelle Dela Cruz
- Joaquin Gris Roca
- John Chi
- Jongbong Lee
- Juliet May
- Katarina Starcevic
- Katherine Vadella
- Lara Bryfonski
- Lindley Winchester
- Logan Peng
- Lucia Donatelli
- Margaret Anne Rowe
- Margaret Borowczyk
- Maria Stoianova
- Mariko Uno
- Mary Henderson
- Maya Barzilai
- Md. Jahurul Islam
- Michaela Harrington
- Minnie Annan
- Mitchell Abrams
- Mohammad Ali Yektaie
- Naomee-Minh Nguyen
- Nicholas Workman
- Nicole Steinberg
- Rachel Thorson
- Rebecca Childress
- Ruizhong Li
- Ryan Murphy
- Sakol Suethanapornkul
- Sean Macavaney
- Sean Simpson
- Shannon Mooney
- Siddharth Singh
- Ryan Murphy
- Yilun Zhu
- Yiran Xu
- Young-A Son
- Yushi Zhao
- Zhuxin Wang

... and others who wish to remain anonymous!