A Multilayer View of Discourse Relation Graphs

Amir Zeldes
Georgetown University
amir.zeldes@georgetown.edu

LTI Colloquium, CMU, 2017-11-10
Plan

I. Discourse parses in Rhetorical Structure Theory
   - What are discourse relations?
   - RST in a nutshell
   - Dependency representations for RST

II. Discourse encapsulation
    - Veins Theory and the Discourse Encapsulation Hypothesis
    - The Georgetown University Multilayer corpus (GUM)
    - A multifactorial model of discourse encapsulation

III. Relation signaling
    - RST-DT and the RST Signalling Corpus
    - Training LSTMs for signal detection
Discourse relations

What relations exist between utterances as a text unfolds?

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

2. \([[[They left lights on]]_{\text{cause}} so Ellie got mad.][She hates that]_{\text{background}}]\)
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
Rhetorical Structure Theory

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
Why is this important?

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

A Multilayer View of Discourse Relation Graphs / A. Zeldes

CMU LTI Colloquium
Simplifying trees

- We will care how far things are in the graph
- Using non-terminal spans is problematic:

RD(2,3) = 1 edge

RD(2,3) = 2 edges

What’s the “Rhetorical Distance”?
Dependency Representation

- Following Hayashi et al. (2016), use Li et al.'s (2014) dependency interpretation*

* conversion code available at: https://github.com/amir-zeldes/rst2dep
II. Discourse Encapsulation
Discourse Encapsulation Hypothesis

- Do discourse parses constrain referentiality?
  - Discourse as stack (Polanyi 1988, Roberts 2012)
  - Right Frontier Constraint (Asher & Lascarides 2003)
  - Veins Theory (Cristea et al. 1998)
  - Different parametrizations (Chiarcos & Krasavina 2008)

Applications:
Coreference resolution
Referring expression generation
Dialog planning

"I know it!"
Veins Theory (Cristea et al. 1998)

- VT proposes **Domains of Referential Accessibility**
  - Nuclei ‘see’ the satellites along their “vein”
  - Satellites can’t access satellites of other nuclei
  - Path length irrelevant
  - Test on 5 texts: (fra, rom, eng) ~100%

Figure 1: Tree structure and veins for Example 1

3 doesn’t see 1
Tetreault & Allen (2003:7):

Our results indicate that incorporating discourse structure does not improve performance, and in most cases can actually hurt performance.

Based on much larger RST Discourse Treebank (RST-DT, ~180K tokens, Carlson et al. 2003)

Suggests VT does not work ‘in the wild’
Research questions

- Can we treat DRAs as quantitative tendencies?
  - Not binary restriction: more \(<\) --> less access
  - Multifactorial
    - Not just based on path
    - Also consider surface distance, graph distance, and more
- Applicable to different types of referentiality?
  - Pronominal anaphora (*The president ... he*)
  - Lexical coreference (*Joe Biden ... Joe*, cf. TextTiling, Hearst 1997)
Data

- What could influence mention likelihood? (Recasens et al. 2013, Zeldes 2017a)

- We need:
  - RST parses
  - Coreference annotation (anaphora, lexical, bridging)

- Possible predictors:
  - Utterance length
  - Surface and ‘Rhetorical’ Distance metrics (SD, RD)
  - Syntactic structure (parses)
  - POS tags
  - Sentence types
  - ...
Georgetown University Multilayer corpus

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

See Zeldes & Simonson (2016) on accuracy

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

<table>
<thead>
<tr>
<th>text type</th>
<th>source</th>
<th>texts</th>
<th>tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interviews</td>
<td>Wikinews</td>
<td>19</td>
<td>18037</td>
</tr>
<tr>
<td>(conversational)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>News (narrative)</td>
<td>Wikinews</td>
<td>21</td>
<td>14093</td>
</tr>
<tr>
<td>Travel guides</td>
<td>Wikivoyage</td>
<td>17</td>
<td>14955</td>
</tr>
<tr>
<td>(informative)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>How-tos (instructional)</td>
<td>wikiHow</td>
<td>19</td>
<td>16920</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>76</td>
<td>64005</td>
</tr>
</tbody>
</table>
Veins in dependency representation

- Ancestry: Is one EDU a direct ancestor of the other in the dependency tree?

wikiHow: “How to Make a Glowstick”
Target variables

- What are we trying to predict?
  - Binary domains:
    - Is there coreference between two EDUs?
    - Explore for anaphora, lexical, bridging
  - Coreference **density**:
    - How much coreferentiality exists between two EDUs? (# coreferent pairs)
  - Direct and indirect antecedents:
    - Check if the **immediate antecedent** of entity in EDU2 is in EDU1 (NB: makes surface distance very important!)
    - Alternatively, just check for coreference
Experiment setup

- ~170K possible EDU pairs grouped by document
- Looking at distance and direct parentage:

**Anaphora**

43.2% direct
\( r(\text{RD}) = -0.14 \)
\( r(\text{ED}) = -0.12 \)

**Bridging**

45.7% direct
\( r(\text{RD}) = -0.07 \)
\( r(\text{ED}) = -0.06 \)
Why is prediction weak despite intuition?

- Lots of confounds!!
  - **Length**: what if the main vein nucleus is really short? -> Unlikely to contain coreferent mentions
  - **Relations**: *Purpose* --> less coref; *Cause* --> more:
    - *needs to be exaggerated [in order to be funny — ]* _PURPOSE_
    - *the banner read ‘We Know’. [That’s all it said.]_ _RESTATEMENT_
  - **Sentence type**: imperatives, fragments have fewer entities than declaratives, questions
  - ... + tense, genre, syntax, document position, ...

Go multifactorial!
Is RD significant? (any distance coref)

- Yes, and so is surface distance and directness!
- But not as important as length

Random effects:

<table>
<thead>
<tr>
<th>Groups</th>
<th>Name</th>
<th>Variance</th>
<th>Std.Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>doc</td>
<td>(Intercept)</td>
<td>0.09789</td>
<td>0.3129</td>
</tr>
<tr>
<td>Residual</td>
<td></td>
<td>0.82965</td>
<td>0.9109</td>
</tr>
</tbody>
</table>

Number of obs: 172150, groups: doc, 76

Fixed effects:

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>Std. Error</th>
<th>t value</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>0.2695836</td>
<td>0.0723038</td>
<td>3.73</td>
<td>***</td>
</tr>
<tr>
<td>scale(len1)</td>
<td>0.2043943</td>
<td>0.0023432</td>
<td>87.23</td>
<td>***</td>
</tr>
<tr>
<td>scale(len2)</td>
<td>0.1833124</td>
<td>0.0023811</td>
<td>76.99</td>
<td>***</td>
</tr>
<tr>
<td>rhet_dist</td>
<td>-0.0511588</td>
<td>0.0014351</td>
<td>-35.65</td>
<td>***</td>
</tr>
<tr>
<td>edu_dist</td>
<td>-0.0015377</td>
<td>0.0001168</td>
<td>-13.17</td>
<td>***</td>
</tr>
<tr>
<td>genrenews</td>
<td>-0.0348780</td>
<td>0.0997936</td>
<td>-0.35</td>
<td></td>
</tr>
<tr>
<td>genrevoyage</td>
<td>-0.2161897</td>
<td>0.1047555</td>
<td>-2.06</td>
<td>**</td>
</tr>
<tr>
<td>genrewhow</td>
<td>0.0969725</td>
<td>0.1016942</td>
<td>0.95</td>
<td></td>
</tr>
<tr>
<td>directTrue</td>
<td>0.2280120</td>
<td>0.0091334</td>
<td>24.96</td>
<td>***</td>
</tr>
</tbody>
</table>

Gaussian mixed effects model
Which relations favor coreference?

- Unsurprisingly:
  - ↑ Cause, Restatement
  - ... 
  - ↓ Joint, Sequence

Add to linear model??

A Multilayer View of Discourse Relation Graphs / A. Zeldes

CMU LTI Colloquium
Ensemble approach (RST workshop@INLG 2017)

- Use Extra Trees ensemble (Geurts et al. 2006)
  - Classification (coref yes/no)
  - Regression (predict density)

<table>
<thead>
<tr>
<th>features</th>
<th>RMSE (reg)</th>
<th>accuracy (clf)</th>
</tr>
</thead>
<tbody>
<tr>
<td>majority</td>
<td>0.9652</td>
<td>77.90%</td>
</tr>
<tr>
<td>EDU</td>
<td>0.9501</td>
<td>78.36%</td>
</tr>
<tr>
<td>RD</td>
<td>0.9453</td>
<td>78.79%</td>
</tr>
<tr>
<td>all</td>
<td>0.7107</td>
<td>86.83%</td>
</tr>
</tbody>
</table>
What do the predictions look like?

- We can visualize predictions as a heat map:
What do the predictions look like?

A Multilayer View of Discourse Relation Graphs / A. Zeldes

CMU LTI Colloquium
What do the predictions look like?

Listen up, kids:

You'll be working with a solution that's near a 12 on the pH scale.

That basically means don't swallow it.

Do not put it in your eyes.

Do not bathe in it.

And do not really expose yourself to it directly at all.
III. Relation signaling
Signaling

- Central question in discourse studies:
  - **Cues** help us to spot relations
  - Annotators use cue words as diagnostics:
    - "could I connect these with 'because'?"

Research questions

- What kinds of signals are there?
- How can we identify them in data?
  - Are signal words always meaningful?
  - How ambiguous are they?
  - Can we distinguish meaningful and non-meaningful uses of cues?
Frequentist approaches

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

- Problems:
  - Frequency thresholds
  - Ambiguity (“and” is not associated with any relation – not a Discourse Marker?)
  - Context sensitivity – some words are cues in specific environments

\begin{table}[h]
\centering
\begin{tabular}{ |l|c|l|l| }
\hline
Relation type & Freq & marker & translation \\
\hline
Elaboration & 150 & \textit{kotoryj} & "which, that" \\
Joint & 119 & \textit{i, takzhe} & and, as well \\
Attribution & 118 & \textit{zajavil, soobschil} & report, announce etc. \\
Contrast & 62 & \textit{Odnako, a, no} & However, but \\
Cause-Effect & 47 & \textit{Poetomu, V+prichina} & so, accordingly, \textit{V+cause} \\
Purpose & 39 & \textit{Chtoby, dlya} & In order that, for \\
Interpretation-Evaluation & 34 & Nouns and verbs expressing opinion & \\
Background & 31 & No dominant marker & \\
Condition & 27 & \textit{esli} & if \\
\hline
\end{tabular}
\caption{Relations with their most frequent markers}
\end{table}

Toldova et al. 2017
### Example - GUM

<table>
<thead>
<tr>
<th>relation</th>
<th>( f &gt; 0 )</th>
<th>( f &gt; 10 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>solutionhood</td>
<td>viable, contributed, 60th, touched, Palestinians</td>
<td>What, ?, Why, did, How</td>
</tr>
<tr>
<td>circumstance</td>
<td>holiest, Eventually, fell, Slate, transition</td>
<td>October, When, Saturday, After, Thursday</td>
</tr>
<tr>
<td>result</td>
<td>minuscule, rebuilding, distortions, struggle, causing</td>
<td>result, Phoenix, wikiHow, funny, death</td>
</tr>
<tr>
<td>concession</td>
<td>Until, favoured, hypnotizing, currency</td>
<td>Although, While, though, However, call</td>
</tr>
<tr>
<td>justify</td>
<td>payoff, skills, net, Presidential, supporters</td>
<td>NATO, makes, simply, Texas, funny</td>
</tr>
<tr>
<td>sequence</td>
<td>Feel, charter, ammonium, evolving, rests</td>
<td>bottles, Place, Then, baking, soil</td>
</tr>
<tr>
<td>cause</td>
<td>malfunctioned, jams, benefiting, mandate</td>
<td>because, wanted, religious, projects, stuff</td>
</tr>
</tbody>
</table>
A neural approach 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
A neural approach with RNNs

- But the LSTM probably already had it as *If*...
- To find signals, we can listen to output at every token (but loss still based on EDU relation)

```
If we were fish
```

Need big dataset - try both GUM and RST-DT
Implemented with BiLSTM (TensorFlow)

A Multilayer View of Discourse Relation Graphs / A. Zeldes

Character embeddings
Word embeddings (GloVe 300, Pennington et al. 2014)
1-hot POS tags

Forward LSTM
Backward LSTM

elab
cond
elab
eval

LSTM
LSTM
LSTM
LSTM

Hidden: 300
Optimizer: Adam
(rec.) dropout: 0.5
Minibatch: 20
Activation: tanh
Batch normalization
Trainable embeddings

meanings
PP
VHP
TO

VVG
PP
VHP
TO
Adding CRF (Huang et al. 2015, Ma & Hovy 2016)

- CRF
- Backward LSTM
- Forward LSTM
- Character embeddings
  - 1-hot POS tags
  - Word embeddings (GloVe 300, Pennington et al. 2014)
- Character embeddings
- VVG
- PP
- VHP
- TO

Hidden: 300
Optimizer: Adam
(rec.) dropout: 0.5
Minibatch: 20
Activation: tanh
Batch normalization
Trainable embeddings

A Multilayer View of Discourse Relation Graphs / A. Zeldes
CMU LTI Colloquium
Single output performance

- Not so interesting, but:
  - RSTDT – relation accuracy by tokens:
    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:
    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:
  - For each token, output the softmax probability assigned to the correct relation
  - Rank words by probability
  - Shade by average of:
    - Proportion of maximum softmax probability in **sentence**
    - Proportion of maximum softmax probability in **document**
• [This occurs for two reasons:] preparation [As it moves over land,] circumstance [it is cut off from the source of energy driving the storm ...] cause

• [Combine 50 milliliters of hydrogen peroxide and a liter of distilled water in a mixing bowl.] sequence [A ceramic bowl will work best,] elaboration [but plastic works too.] concession
Visualizing token-wise softmax

Genre specific knowledge? (GUM)
- [Thursday, May 7, 2015]_{circumstance} [The current flag of New Zealand]_{preparation}

Word and character embeddings?
- [I cannot comment directly on how the Indian government was prepared for this cyclone]_{concession} [However, the news reports (BBC etc.) were very encouraging]_{joint}
Addressing ambiguity

- Reliability of cue words is a big concern:
  - Which cues can we trust?
  - Which cues are we missing because of weak association?
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.]\textsubscript{elaboration}
  - [and one began trading on the Nasdaq/National Market System last week.]\textsubscript{inverted}

- **Important ‘and’s: (RST-DT)**
  - [and is involved in claims adjustments for insurance companies.]\textsubscript{List}
  - [-- and from state and local taxes too, for in-state investors.]\textsubscript{elaboration}
Evaluating 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

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...)
      
      *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
Caveats and WIP

- Network not trained on gold standard (training on relations, not ‘being a signal’)
- Do we want supervised learning on signals?
- Other questions:
  - Can we compare signals across corpora and genres?
  - Are some signals more robust than others?
  - Genre-specific signals?
- Consequences for learning approach?
  ([Spencer Volk ...][Mr. Volk...][elaboration])
Conclusion

- Good times to be working on discourse!
- Multilayer data can expose complex interdependencies
- Some old ideas are now more feasible:
  - From Veins Theory to quantitative DRAs
  - From signal co-occurrence statistics to contextualized LSTM outputs
- We still need new data, new features 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

References

Sentence type annotation

- Extended version of SPAAC scheme (Leech et al. 2003; not created for this study)

<table>
<thead>
<tr>
<th>tag</th>
<th>type</th>
<th>example</th>
</tr>
</thead>
<tbody>
<tr>
<td>q</td>
<td>polar yes/no question</td>
<td>Did she see it?</td>
</tr>
<tr>
<td>wh</td>
<td>WH question</td>
<td>What did you see?</td>
</tr>
<tr>
<td>decl</td>
<td>declarative (indicative)</td>
<td>He was there.</td>
</tr>
<tr>
<td>imp</td>
<td>imperative</td>
<td>Do it!</td>
</tr>
<tr>
<td>sub</td>
<td>subjunctive (incl. modals)</td>
<td>I could go</td>
</tr>
<tr>
<td>inf</td>
<td>infinitival</td>
<td>How to Dance. Why not go?</td>
</tr>
<tr>
<td>ger</td>
<td>gerund-headed clause</td>
<td>Finding Nemo. Hiring employees</td>
</tr>
<tr>
<td>intj</td>
<td>interjection</td>
<td>Hello!</td>
</tr>
<tr>
<td>frag</td>
<td>fragment</td>
<td>The End.</td>
</tr>
<tr>
<td>other</td>
<td>other predication or combination</td>
<td>Nice, that! Or: 'I've had it, go!' (decl+imp)</td>
</tr>
</tbody>
</table>
Only weak correlations...

- For all EDU pairs:
  - Most have 0 coreference
  - Especially direct antecedents have very low distance
  - Not much predictability (cf. Tetreault & Allen)

![Graphs showing correlation coefficients]

A Multilayer View of Discourse Relation Graphs / A. Zeldes
CMU LTI Colloquium