A Neural Approach to Discourse Relation Signaling

Amir Zeldes
Georgetown University
amir.zeldes@georgetown.edu

GURT 2018, 2018-03-10
Questions

- What are discourse relations?
- Distribution in genres? (Taboada & Lavid 2003)
- How are they marked? (Taboada & Das 2013)
  - Example: contrast
    - Explicit signals: “on the other hand” or “although”
    - Implicit signals: antonyms, coreferent mentions …
- Easiest/hardest relations to identify?
- Most/least reliable signals in context?

➢ To answer these questions we need to annotate relations in corpora
DRs in Rhetorical Structure Theory

(Mann & Thompson 1988)

- See RST Website: http://www.sfu.ca/rst/

- Other frameworks: PDTB (Prasad et al. 2008), SDRT (Asher & Lascarides 2003)
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
- 20 DRs in 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>Total</td>
<td></td>
<td>101</td>
<td>85,350</td>
</tr>
</tbody>
</table>

Class-Sourced!
Frequentist approaches to DR markers

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

- Problems:
  - Frequency thresholds
  - Ambiguity
    - “and” appears in all relations – 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" \textit{and, as well} \\
Joint & 119 & i, takzhe & \textit{report, announce etc.} \\
Attribution & 118 & zajavil, soobschil & \textit{However, but} \\
Contrast & 62 & Odnako, a, no & \textit{so, accordingly, V+cause} \\
Cause-Effect & 47 & Poetomu, V+prichina & \textit{In order that, for} \\
Purpose & 39 & Chtoby, dlya & \textit{Nouns and verbs expressing opinion} \\
Interpretation-Evaluation & 34 & & \textit{No dominant marker} \\
Background & 31 & & \textit{if} \\
Condition & 27 & \textit{esli} & \textit{if} \\
\hline
\end{tabular}
\caption{Relations with their most frequent markers}
\end{table}

Toldova et al. 2017
Relation classification with RNNs

- Recurrent Neural Networks can identify relations (e.g. Braud et al. 2017)
- Words vectors are fed to an encoder
- Multiclass classification

If we were fish What were the signals?

condition 0.6
cause 0.2
contrast 0.1
Relation classification with RNNs

- The RNN probably already had it at *If*...
- We can listen to output at every token

Amir Zeldes / A Neural Approach to Discourse Relation Signaling

GURT 2018, 2018-03-10
Bi-LSTM CRF (Huang et al. 2015, Ma & Hovy 2016)

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

- **if**
- **we**
- **were**
- **fish**

Conditional Random Fields -> globally optimal solution

Long-Short Term Memory network

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

Batch normalization
Trainable embeddings
Visualizing RNN predictions

- Basic idea – find the most ‘convincing’ tokens:
  - Use tokens' probability of correct relation
  - Shade by:

More formally:
- $p / \max(\text{softmax}(\text{sent}))$
- $p / \max(\text{softmax}(\text{document}))$

Amir Zeldes / A Neural Approach to Discourse Relation Signaling
Visualizing RNN predictions

- [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

Do humans agree?

Data: GUM
Does the RNN find signals like a human?

- Evaluate on sample from RST Signalling Corpus (Taboada & Das 2013)
  - 210 relations
  - 153 signals
    - Only 83 attributable to words!
      (not: genre, zero relative, graphical layout...)
      
      *In a remark [someone should remember this time next year,]*
    - Only 47 are lexical items!
      (not: lexical chain, syntactic parallelism)
      
      *Congress gave Senator Byrd's state ... [Senator Byrd is chairman..]*
Does the RNN find signals like a human?

- Use *recall rate @k* to evaluate

---

**All word-anchored signals**

- **r@1**
- **r@2**
- **r@3**

**Lexical items only**

- **r@1**
- **r@2**
- **r@3**

---

**Data:** Signaling Corpus

Amir Zeldes / A Neural Approach to Discourse Relation Signaling

GURT 2018, 2018-03-10
Suppose the RNN really flags signals...

- Can we:
  - Get ‘signal strength’ for words?
  - Ambiguity scores?
  - Most ‘signally’ relations?
  - Variation across genres?
Assessing ambiguity

- We can get ambiguity scores based on the range of probabilities each word gets.
Assessing ambiguity

- **Irrelevant ‘and’s: (Wall Street Journal)**
  - [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: (Wall Street Journal)**
  - [and is involved in claims adjustments for insurance companies .]_{List}
  - [-- and from state and local taxes too , for in-state investors .]_{elaboration}

Is this a ‘news’ thing?

What about signals that aren’t words?
### Giving the network more than words

<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>37.14</td>
<td>39.14</td>
<td>39.14</td>
<td>43.07</td>
</tr>
</tbody>
</table>

Baseline: 25.58

Data: GUM
Relations across genres

- DR probabilities vary by genre, sentence type...
- Even for the same ‘sentence’ – think of “Yes.”
Examples - learning from more than text

Plain:
[1 teaspoon baking powder]_{joint}

+Genre or Layout: (whow)
[1 teaspoon baking powder]_{joint}

wikiHow: How to Make Vegan Cupcakes

- 2 teaspoons baking soda
- 1 teaspoon baking powder
- Pinch of salt
- 450ml (1-3/4 cup) unsweetened soy milk
Which genres signal most strongly?

Confounds:
- Data size
- Relations

<table>
<thead>
<tr>
<th>Genre</th>
<th>N=6K</th>
<th>N=5K</th>
<th>N=5K</th>
<th>N=18K</th>
<th>N=14K</th>
<th>N=15K</th>
<th>N=5K</th>
<th>N=17K</th>
</tr>
</thead>
<tbody>
<tr>
<td>fiction</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>reddit</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bio</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>interview</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>news</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>voyage</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>academic</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>whow</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Mean signaly-ness
Which relations are hardest?

Are markers the same across genres?
### Top signals - overall

<table>
<thead>
<tr>
<th>Condition</th>
<th>Sequence</th>
<th>Solutionhood</th>
<th>Circumstance</th>
</tr>
</thead>
<tbody>
<tr>
<td>- If</td>
<td>- minutes</td>
<td>- ?</td>
<td>- NUM (=year, day, ...)</td>
</tr>
<tr>
<td>- you</td>
<td>- then</td>
<td>- you</td>
<td>- month</td>
</tr>
<tr>
<td>- if</td>
<td>- add</td>
<td>- Do</td>
<td>- before</td>
</tr>
</tbody>
</table>

How stable are these markers across genres?
### Top signals - different genres

<table>
<thead>
<tr>
<th>Condition</th>
<th>Overall</th>
<th>Wikihow</th>
<th>Fiction</th>
<th>Bios</th>
</tr>
</thead>
<tbody>
<tr>
<td>If you if</td>
<td>If you if</td>
<td>If when if</td>
<td>&lt;N/A&gt;</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Overall</th>
<th>Wikihow</th>
<th>Fiction</th>
<th>Bios</th>
</tr>
</thead>
<tbody>
<tr>
<td>minutes then add</td>
<td>Pour add minutes</td>
<td>woke jumped fell</td>
<td>became died August</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Solutionhood</th>
<th>Overall</th>
<th>Wikihow</th>
<th>Fiction</th>
<th>Bios</th>
</tr>
</thead>
<tbody>
<tr>
<td>? you Do</td>
<td>? Impact failure</td>
<td>want ? Do</td>
<td>&lt;N/A&gt;</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Circumstance</th>
<th>Overall</th>
<th>Wikihow</th>
<th>Fiction</th>
<th>Bios</th>
</tr>
</thead>
<tbody>
<tr>
<td>NUM month before</td>
<td>After When Once</td>
<td>when outside morning</td>
<td>when After war</td>
<td></td>
</tr>
</tbody>
</table>
Conclusion

- Relation signaling is complex, genre specific
  - Many signals are lexical, not function words
  - New models go substantially beyond frequentist approaches

- Computational models of discourse signals can:
  - Inform relation inventory development and corpus annotation schemes
  - Improve automatic discourse parsing
  - Help develop new theories about DR processing
Thanks!
References

Thanks to GUM annotators! (so far)

- Adrienne Isaac
- Akitaka Yamada
- Amani Aloufi
- Amelia Becker
- Andrea Price
- Andrew O'Brien
- Anna Runova
- Anne Butler
- Arianna Janoff
- 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
- Ethan Beaman
- Han Bu
- Hang Jiang
- Hanwool Choe
- Hassan Munshi
- Ho Fai Cheng
- Jakob Prange
- Jehan al-Mahmoud
- Jemm Excelle Dela Cruz
- Joaquin Gris Roca
- 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
- Siyu Liang
- Stephanie Kramer
- Sylvia Sierra
- Timothy Ingrassia
- Wenxi Yang
- Xiaopei Wu
- Yang Liu
- Yilun Zhu
- Yingzhu Chen
- Yiran Xu
- Young-A Son
- Yushi Zhao
- Zhuxin Wang
- ... and others who wish to remain anonymous!