Structured Prediction Using Decoding in the Context of Discourse Parsing for Chat Dialogues: From Classical to Neural Approaches

Stergos Afantenos

ENSEEIHT - INP
12 novembre 2018
Supervised classification

- We have a set of labeled examples

\[ \{x_i, y_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} P(x, y) \quad x \in \mathcal{X}, y \in \mathcal{Y} \]

- We need to learn a function \( f : \mathcal{X} \mapsto \mathcal{Y} \) that predicts \( y = f(x) \) on future data \( x \) with \( (x, y) \overset{\text{i.i.d.}}{\sim} P(x, y) \)

- The learned function can be linear \( y = \arg\max_{y \in \mathcal{Y}} w_y x \) or non-linear, learned from a neural network.

- Crucially, \( \mathcal{Y} \) is a set of usually few classes that are distinct between them.
Structured prediction

What happens when $y \in \mathcal{Y}$ is a complex object?

- $y$ can be a sequence
  
  $x =$ John saw Mary with the telescope
  
  $y =$ noun verb noun preposition article noun

- $y$ can be a tree or a graph

```
A hearing is scheduled on the issue today
```
Structured prediction (cont.)

- If we consider \( y \) as a separate class for every sequence/tree/graph then we can have exponentially many classes!
- There are various ways to perform structured output prediction:
  - Decoding over a local probability distribution
  - Using the kernel trick in SVMs
  - Using an approach similar to SparseMap (Niculae et al. 2018)
Discourse for multi-party dialogues

- Discourse parsing for monologues has been extensively investigated.
- Discourse parsing for other forms of human communication on the other hand has not received the same attention from the computational linguistics community.
The Settlers of Catan

A board game where 2-4 players compete for establishing settlements and roads on an island, gathering and negotiating resources in the process.
A sample dialogue

Our input

65 lj anyone want sheep for clay?
66 gw got none, sorry :( 
67 gw so how do people know about the league?
68 wm no
70 lj i did the trials
74 tk i know about it from my gf
75 gw [yeah me too,]_a
[are you an Informatics student then, lj?]_b
76 tk did not do the trials
77 wm has anyone got wood for me?
78 gw [I did them]_a [because a friend did]_b
79 gw lol william, you cad
80 gw afraid not :( 
81 lj no, I’m about to start math
82 tk sry no
83 gw my single wood is precious
84 wm what’s a cad?
Concurrent discussions

165  lj  anyone want sheep for clay?
166  gw  got none, sorry :(  
167  gw  so how do people know about the league?
168  wm  no
170  lj  i did the trials
174  tk  i know about it from my gf
175  gw  [yeah me too,]a
       [are you an Informatics student then, lj?]b
176  tk  did not do the trials
177  wm  has anyone got wood for me?
178  gw  [I did them]a [because a friend did]b
179  gw  lol william, you cad
180  gw  afraid not :(  
181  lj  no, I’m about to start math
182  tk  sry no
183  gw  my single wood is precious
184  wm  what’s a cad?
Concurrent discussions
A smaller example

1  Alice  anyone got wheat for a sheep?
2  Bob    sorry, not me
3  Clara  nope. you seem to have lots of sheep!
4  Dan    i think i’d rather hang on to my wheat i’m afraid
5  Alice  kk I’ll take my chances then...
A smaller example

1. Alice anyone got wheat for a sheep?
2. Bob sorry, not me
3. Clara nope. you seem to have lots of sheep!
4. Dan i think i’d rather hang on to my wheat i’m afraid
5. Alice kk I’ll take my chances then...

Diagram:

- **Question-Answer Pair**
  - 1
  - 2
  - 3
  - 4
- **Acknowledgement**
  - 5

QAP

Ack.

Ack.
How can we represent dialogue?

- Penn Discourse TreeBank (PDTB) uses directed edges, without structural constraints.
How can we represent dialogue?

- Penn Discourse TreeBank (PDTB) uses directed edges, without structural constraints
- Rhetorical Structure Theory (RST) represents discourse as trees, which is too restrictive
How can we represent dialogue?

- Penn Discourse TreeBank (PDTB) uses directed edges, without structural constraints
- Rhetorical Structure Theory (RST) represents discourse as trees, which is too restrictive
- Segmented Discourse Representation Theory (SDRT) uses two-layered directed acyclic graphs
How can we represent dialogue?

- Penn Discourse TreeBank (PDTB) uses directed edges, without structural constraints
- Rhetorical Structure Theory (RST) represents discourse as trees, which is too restrictive
- Segmented Discourse Representation Theory (SDRT) uses two-layered directed acyclic graphs

SDRT’s flexible structure is best suited for chats
Given a discourse segmented in EDUs, an SDRT graph is a tuple \((V, E_1, E_2, \ell)\), where

- Vertex set \(V\) contains discourse units (DUs)
- Edge set \(E_1\) contains discourse relations between DUs
- Edge set \(E_2\) represents *Complex Discourse Units (CDUs)*
- Function \(\ell\) assigns a label to discourse relation edges
Complex Discourse Units

Example

Alice  [Do you have a sheep?]_a
Bob    [I do,]_b [if you give me clay]_c
Bob    [or wood.]_d

\[
\text{QAP} \quad \text{CONDITIONAL} \quad \text{ALTERNATION}
\]

a \quad \rightarrow \quad b \quad \leftarrow \quad c \quad \rightarrow \quad d
La théorie de la sélection naturelle [telle qu'elle a été initialement décrite par Charles Darwin,] repose sur trois principes: [1. le principe de variation] [2. le principe d’adaptation] [3. le principe d’hérédité]
Complex Discourse Units (cont.)

A more complicated example

```
1 \(\nabla\) Elab. \(\pi_1\)
   \(\pi_2\) \(\nabla\) Elab. \(\pi_3\)
   \(\nabla\) e-elab. \(\nabla\) Frame \(\nabla\) Elab.
3 \(\pi_4\) \(\pi_5\) \(\pi_6\)
2 \(\pi_7\) \(\pi_8\)
   \(\nabla\) Frame \(\nabla\) Frame \(\nabla\) Elab.
18 \(\nabla\) Elab.
17 \(\nabla\) Frame
19
```
No reliable method has been identified in the literature for identifying CDUs.
We approximate CDUs in the SDRT hypergraph by relations between EDUs only, thus creating a dependency graph.
Distributing relations

No distribution
Head points to head
Distributing relations

No distribution
Head points to head

Partial distribution
Relation semantics determine distribution to the source/target CDU components

[I'll buy a card]_a
[and not a road]_b
[because I have sheep]_c
[and wheat]_d
[and ore]_e
Distributing relations

[I’ll buy a card] \(a\)
[and not a road] \(b\)
[because I have sheep] \(c\)
[and wheat] \(d\)
[and ore] \(e\)

No distribution
Head points to head

Partial distribution
Relation semantics determine distribution to the source/target CDU components

Full distribution
All relations distribute to every component
Discourse structure annotation

- 4 naive annotators were involved; they were trained on 22 negotiation dialogues with 560 turns.
- 0.72 kappa on structure and 0.58 kappa on labelling
- Expert annotators adjudicated the naive annotators.
- Adjudication involved five separate phases
Discourse structure annotation

- 4 naive annotators were involved; they were trained on 22 negotiation dialogues with 560 turns.
- 0.72 kappa on structure and 0.58 kappa on labelling
- Expert annotators adjudicated the naive annotators.
- Adjudication involved five separate phases

Dataset overview:

<table>
<thead>
<tr>
<th></th>
<th>Total</th>
<th>Training</th>
<th>Testing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dialogues</td>
<td>1081</td>
<td>965</td>
<td>116</td>
</tr>
<tr>
<td>Turns</td>
<td>9160</td>
<td>8166</td>
<td>994</td>
</tr>
<tr>
<td>EDUs</td>
<td>10678</td>
<td>9546</td>
<td>1132</td>
</tr>
<tr>
<td>Relation instances</td>
<td>10513</td>
<td>9421</td>
<td>1092</td>
</tr>
<tr>
<td>CDUs</td>
<td>1284</td>
<td>1132</td>
<td>152</td>
</tr>
</tbody>
</table>

A dialogue includes a negotiation phase during a game
# Distribution of annotated relations

<table>
<thead>
<tr>
<th>Relation</th>
<th>Total</th>
<th>Training</th>
<th>Testing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comment</td>
<td>1851</td>
<td>1684</td>
<td>167</td>
</tr>
<tr>
<td>Clarification_question</td>
<td>260</td>
<td>240</td>
<td>20</td>
</tr>
<tr>
<td>Elaboration</td>
<td>869</td>
<td>771</td>
<td>98</td>
</tr>
<tr>
<td>Acknowledgment</td>
<td>1010</td>
<td>893</td>
<td>117</td>
</tr>
<tr>
<td>Continuation</td>
<td>987</td>
<td>873</td>
<td>114</td>
</tr>
<tr>
<td>Explanation</td>
<td>437</td>
<td>407</td>
<td>30</td>
</tr>
<tr>
<td>Conditional</td>
<td>124</td>
<td>105</td>
<td>19</td>
</tr>
<tr>
<td>Question-answer_pair</td>
<td>2541</td>
<td>2236</td>
<td>305</td>
</tr>
<tr>
<td>Alternation</td>
<td>146</td>
<td>128</td>
<td>18</td>
</tr>
<tr>
<td>Q-Elab</td>
<td>599</td>
<td>525</td>
<td>74</td>
</tr>
<tr>
<td>Result</td>
<td>578</td>
<td>551</td>
<td>27</td>
</tr>
<tr>
<td>Background</td>
<td>61</td>
<td>58</td>
<td>3</td>
</tr>
<tr>
<td>Narration</td>
<td>130</td>
<td>116</td>
<td>14</td>
</tr>
<tr>
<td>Correction</td>
<td>212</td>
<td>189</td>
<td>23</td>
</tr>
<tr>
<td>Parallel</td>
<td>215</td>
<td>196</td>
<td>19</td>
</tr>
<tr>
<td>Contrast</td>
<td>493</td>
<td>449</td>
<td>44</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td>10513</td>
<td>9421</td>
<td>1092</td>
</tr>
</tbody>
</table>
Learning structures vs Local Models

Ideally:

\[ h : \mathcal{X}_E^n \mapsto \mathcal{Y}_G \]

Realistically:

\[ h : \mathcal{X}_E^2 \mapsto \mathcal{Y}_R \]
Problems with this approach

• We have no guarantees that structures will be well formed
• graphs might be disconnected
• we might have cycles
• the Right Frontier Constraint might not be respected
• etc.
How can we alleviate this problem?

Do structured decoding over local probability distributions

- Maximum Spanning Trees (MST)
- Integer Linear Programming (ILP)
Maximum Spanning Trees (MST)
We used a regularized Maximum Entropy model:

$$P(r|p) = \frac{1}{Z(c)} \exp \left( \sum_{i=1}^{m} w_if_i(p, r) \right)$$
<table>
<thead>
<tr>
<th>Category</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positional</td>
<td>Speaker initiated the dialogue</td>
</tr>
<tr>
<td></td>
<td>- First utterance of the speaker in the dialogue</td>
</tr>
<tr>
<td></td>
<td>- Position in dialogue</td>
</tr>
<tr>
<td></td>
<td>- <em>Distance between EDUs</em></td>
</tr>
<tr>
<td></td>
<td>- <em>EDUs have the same speaker</em></td>
</tr>
<tr>
<td>Lexical</td>
<td>Ends with exclamation mark</td>
</tr>
<tr>
<td></td>
<td>- Ends with interrogation mark</td>
</tr>
<tr>
<td></td>
<td>- Contains possessive pronouns</td>
</tr>
<tr>
<td></td>
<td>- Contains modal modifiers</td>
</tr>
<tr>
<td></td>
<td>- Contains words in lexicons</td>
</tr>
<tr>
<td></td>
<td>- Contains question words</td>
</tr>
<tr>
<td></td>
<td>- Contains a player’s name</td>
</tr>
<tr>
<td></td>
<td>- Contains emoticons</td>
</tr>
<tr>
<td></td>
<td>- First and last words</td>
</tr>
<tr>
<td>Parsing</td>
<td>Subject lemmas given by syntactic dependency parsing</td>
</tr>
<tr>
<td></td>
<td>- Dialogue act according to (Cadilhac et al, 2013)</td>
</tr>
</tbody>
</table>
The turn Constraint

- Within a turn people can have a full discourse model, including backward links
- Outside turns, we cannot have backward links

Example:
Although he was very tired, he still came to the meeting.

We thus build two different local models applying this constraint:

- Intra-turn: training contains all pairs of EDUs \((i, j)\) with \(i \neq j\)
- Inter-turn: training contains all pairs of EDUs \((i, j)\) with \(i < j\)

We apply it during decoding also.
The turn Constraint

- Within a turn people can have a full discourse model, including backward links
- Outside turns, we cannot have backward links

Example:

Although he was very tired, he still came to the meeting.
The turn Constraint

- Within a turn people can have a full discourse model, including backward links
- Outside turns, we cannot have backward links

Example:

Although he was very tired, he still came to the meeting.

- We thus build two different local models applying this constraint
  - Intra-turn: training contains all pairs of EDUs \((i, j)\) with \(i \neq j\)
  - Inter-turn: training contains all pairs of EDUs \((i, j)\) with \(i < j\)

- We apply it during decoding also
Decoders

- Baseline decoder (*LOCAL*)

\[
\hat{r} = \arg\max_r \left( \frac{1}{Z(c)} \exp \left( \sum_{i=1}^{m} w_i f_i(p, r) \right) \right)
\]

- Maximum Spanning Trees (MST)

\[
T^* = \arg\max_T \sum_{e \in E(T)} w(e)
\]

\[
w(e) = \log \left( \frac{p(e)}{1 - p(e)} \right)
\]
<table>
<thead>
<tr>
<th>Method</th>
<th>Unlabelled</th>
<th>Labelled</th>
</tr>
</thead>
<tbody>
<tr>
<td>LAST</td>
<td>0.584</td>
<td>0.391</td>
</tr>
<tr>
<td>LOCAL</td>
<td>0.483</td>
<td>0.429</td>
</tr>
<tr>
<td>MST</td>
<td>0.671</td>
<td>0.516</td>
</tr>
</tbody>
</table>
Integer Linear Programming (ILP)
We define an optimization problem where all variables are integers:

\[
\text{maximize} \quad c^T x \\
\text{subject to} \quad Ax \leq b \\
x \geq 0 \\
\text{and} \quad x \in \mathbb{Z}^n
\]

- Structural freedom
- Easy to parametrize
- Versatile constraints on need
Pair modelization: Maximum Entropy model

The model provides us with two real-valued functions:

\[ s_a : [1..n]^2 \rightarrow [0, 1] \]
\[ s_r : [1..n]^2 \times [1..m] \rightarrow [0, 1] \]

Graph building: Integer Linear Programming

\[
\text{maximize } \sum_i \sum_j \left( a_{ij} s_a(i, j) + \sum_k r_{ijk} s_r(i, j, k) \right)
\]

subject to our set of constraints
Structural constraints

- Acyclicity
- Unique root
- Connectedness
- Turn Constraint
Edge count bounds

Outgoing edge cap
An utterance can elicit a limited number of reactions:

\[ \forall i \sum_j a_{ij} \leq \omega \]
Edge count bounds

**Outgoing edge cap**
An utterance can elicit a limited number of reactions:

\[ \forall i \sum_j a_{ij} \leq \omega \]

**Density cap**
An unbounded number of edges would result in a near-complete graph, as the objective function is increasing.

\[ \sum_i \sum_j a_{ij} \leq \delta(n - 1) \]
## Evaluation F1 scores on test corpus

<table>
<thead>
<tr>
<th>Method</th>
<th>Unlabelled</th>
<th>Labelled</th>
<th>Edge count</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>No distribution</strong></td>
<td></td>
<td></td>
<td>10191</td>
</tr>
<tr>
<td>LAST</td>
<td>0.584</td>
<td>0.391</td>
<td></td>
</tr>
<tr>
<td>LOCAL</td>
<td>0.483</td>
<td>0.429</td>
<td></td>
</tr>
<tr>
<td>MST</td>
<td>0.671</td>
<td>0.516</td>
<td></td>
</tr>
<tr>
<td>ILP</td>
<td><strong>0.689</strong></td>
<td><strong>0.531</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Partial distribution</strong></td>
<td></td>
<td></td>
<td>11734</td>
</tr>
<tr>
<td>LAST</td>
<td>0.593</td>
<td>0.426</td>
<td></td>
</tr>
<tr>
<td>LOCAL</td>
<td>0.471</td>
<td>0.396</td>
<td></td>
</tr>
<tr>
<td>MST</td>
<td>0.647</td>
<td>0.488</td>
<td></td>
</tr>
<tr>
<td>ILP</td>
<td><strong>0.668</strong></td>
<td><strong>0.519</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Full distribution</strong></td>
<td></td>
<td></td>
<td>13675</td>
</tr>
<tr>
<td>LAST</td>
<td>0.582</td>
<td>0.420</td>
<td></td>
</tr>
<tr>
<td>LOCAL</td>
<td>0.541</td>
<td>0.443</td>
<td></td>
</tr>
<tr>
<td>MST</td>
<td>0.613</td>
<td>0.466</td>
<td></td>
</tr>
<tr>
<td>ILP</td>
<td><strong>0.675</strong></td>
<td><strong>0.527</strong></td>
<td></td>
</tr>
</tbody>
</table>
Neural network architecture

sigmoid

feed forward

\[ \leftarrow h_{s-3}; h_{s-2}; \ldots; h_s; h_s \]

\[ \rightarrow h_{t-3}; h_{t-2}; \ldots; h_t; h_t \]

bi-LSTM

\[ e_{s-3} \quad e_{s-2} \quad e_{s-1} \quad e_s \]

\[ e_{t-3} \quad e_{t-2} \quad e_{t-1} \quad e_t \]
<table>
<thead>
<tr>
<th></th>
<th>Bi-LSTM</th>
<th></th>
<th></th>
<th>MST Decoding</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>R</td>
<td>F1</td>
<td>P</td>
<td>R</td>
<td>F1</td>
</tr>
<tr>
<td>LAST</td>
<td>50.26%</td>
<td>64.31%</td>
<td>56.42%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Distance 1</td>
<td>9.10%</td>
<td>18.24%</td>
<td>12.14%</td>
<td>14.88%</td>
<td>19.04%</td>
<td>16.70%</td>
</tr>
<tr>
<td>Distance 2</td>
<td>52.49%</td>
<td>57.58%</td>
<td>54.92%</td>
<td>50.98%</td>
<td>65.22%</td>
<td>57.22%</td>
</tr>
<tr>
<td>Distance 3</td>
<td>52.12%</td>
<td>62.82%</td>
<td>56.98%</td>
<td>51.60%</td>
<td>66.02%</td>
<td>57.92%</td>
</tr>
<tr>
<td>Distance 4</td>
<td>57.35%</td>
<td>55.98%</td>
<td>56.66%</td>
<td>52.22%</td>
<td>66.81%</td>
<td>58.62%</td>
</tr>
</tbody>
</table>
Future work

• Use the learned representations as input to an SVN structured prediction framework (joint work with Phuong Nguyen, Edouard Pauwels and Mathieu Serrurier)
• Disentangle threads of conversations.
• Perform semi-supervised learning (joint work with Luce Le Gorrec and Sandrine Mouysset)