Some observations on annotation

Nicholas Asher

Melodi/axe discours

Utrecht, June 2017
Why do annotations?

- A discovery process for linguistically interesting phenomena, especially beyond the sentence.
- Building an annotation model encodes linguistic generalizations and then tests them, useful for linguists.
- Testing on real data necessary as there are too many variables to control for a full semantics beyond the sentence.
- also means there are compromises...
- a dialectical process in annotation: devise a model, annotate, revise the model, annotate, ...
- the model gets better!
Some things you want to know about your model

- Is the structure or semantics we develop in the annotation model replicable?
- How complex are the features on which the decisions about structure are made? e.g., are the features easily recoverable from the raw data, or do they require complex processing?
- Test this with formal machine learning methods and computable features
Discourse, a case in point

- the interpretation of a text is dynamic and depends on context.
- example: *John fell. Max pushed him.* (Lascarides & Asher 1993)
- each discourse constituent has one or more rhetorical functions relative to other DUs
- several DUs can have the same rhetorical function. Recursive Structure and CDUs
- formalize annotations of discourse structure as directed acyclic graphs.
- How does this picture carry over to dialogue?
Formalization : a reminder

A directed acyclic graph :

\[(V, E_1, E_2, \ell)\]

- \(V\) a set of discourse constituents,
- \(E_1 \cup E_2 \subseteq V \times V\),
- \(\ell : E_1 \rightarrow \text{Relation-Types a labelling}\).
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
some observations

• notice the nested or recursive structure
• this affects content, in particular temporal relations but also anaphora
Simplifying to structures without CDUs

A graph without $E_2$ edges and only EDUs as nodes.

1

\[ \xrightarrow{\text{Elab.}} \]

3

\[ \xrightarrow{\text{e-elab}} \]

2

\[ \xrightarrow{\text{Elab.}} \]

4 \xrightarrow{\text{C.}} 5 \xrightarrow{\text{C.}} 6
Dialogue Structure

Conversation as a game of message exchange involving a kind of signaling game:

- X plays $\phi$
- Y decodes a message in strategic equilibrium (safety, credibility)
- Y decides what signal to send in return
- X decodes a message.
- ...

Choice of corpus

- cooperativity at various levels
- strategic goals behind the conversation
- multi-party
- we wanted to look at structural properties of such conversations.
A  Do you have rock?
B  I’ve got lots of wheat
    [in fact, B has a rock]
A  I’ll give you 2 clay for a rock
B  How about 2 clay for a wheat?
A  I’ll give 1 clay for 3 wheat
B  OK, it’s a deal.
The game
Some specificities of the Settlers corpus

• conversation involving strategic reasoning, multi party, clearly opposing general goals
• also allowed us to study situated communication, because the visual environment of the game constituted a controlled nonlinguistic, virtual environment.
Multi party dialogue structures

234 gotwood4sheep anyone got wheat for a sheep?
235 inca sorry, not me
236 CheshireCatGrin nope. you seem to have lots of sheep!
237 gotwood4sheep yup baaa
238 dmm i think i’d rather hang on to my wheat i’m afraid
239 gotwood4sheep kk i’ll take my chances then...

QAP QAP QAP
234 235 236 238
Ack Ack Ack
239
Multi party dialogue and threads

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?
Not so much recursive structure, but multiple interleaved threads.
In fact two Stac corpora

The Linguistic Corpus:

- 39 *Settlers* games
- SDRT-annotated
- 1091 dialogues
- 10677 EDUs
- 1284 CDUs
- 10191 relations
Two Stac corpora

The Situated Corpus:

- 36 *Settlers* games (so far)
- SDRT-annotated
- ≈ dialogues
- ≈ 30,000 EDUs
- ≈ 5000 CDUs
- ≈ 27000 relations
Annotation scheme: several levels

- segmentation of dialog turns into discourse units
- labelling with domain-related speech acts (negotiation moves)
- relational rhetorical annotation from SDRT but with relations for dialogue (QAP, Q-elab, Acknowledgment, Correction).
- integrating nonlinguistic events into the relational structure (extending the relational annotation)
Domain level acts

- offer: *I’ll give you 2 clay for a rock*
- counteroffer: *How about 2 clay for a wheat?*
- accept: *OK, it’s a deal.*
- refusal: *I don’t think so.*
- has-resource: *I have wheat*
- strategic comment: *joel fancies a bit of your clay*
- other (non relevant for negotiation)
<table>
<thead>
<tr>
<th>Speaker</th>
<th>Id</th>
<th>Turn</th>
<th>Dom. function</th>
<th>Rhet. function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Euan</td>
<td>47</td>
<td>[And I alt tab back from the tutorial.] 1</td>
<td>other</td>
<td>Result*(47_1,47_2)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[What’s up?] 2</td>
<td>other</td>
<td></td>
</tr>
<tr>
<td>Joel</td>
<td>48</td>
<td>[do you want to trade?]</td>
<td>offer</td>
<td>Q-elab(47_2, 48)</td>
</tr>
<tr>
<td>Card.</td>
<td>49</td>
<td>[joel fancies a bit of your clay]</td>
<td>strat.-comment</td>
<td>Expl*(48, 49)</td>
</tr>
<tr>
<td>Joel</td>
<td>50</td>
<td>[yes]</td>
<td>other</td>
<td>Ackn(49, 50)</td>
</tr>
<tr>
<td>Joel</td>
<td>51</td>
<td><img src="#" alt="1" /></td>
<td>other</td>
<td>Comment(50, 51)</td>
</tr>
<tr>
<td>Euan</td>
<td>52</td>
<td>[Whatcha got?]</td>
<td>counteroffer</td>
<td>Q-elab([48-50], 52)</td>
</tr>
<tr>
<td>Joel</td>
<td>53</td>
<td>[wheat]</td>
<td>has-resources</td>
<td>QAP(52, 53)</td>
</tr>
<tr>
<td>Euan</td>
<td>54</td>
<td>[I can wheat for clay.]</td>
<td>counteroffer</td>
<td>Elab([52,53], 54)</td>
</tr>
<tr>
<td>Joel</td>
<td>55</td>
<td>[awesome]</td>
<td>accept(54)</td>
<td>Ackn(54, 55)</td>
</tr>
</tbody>
</table>
Discourse specificities

- Most frequent relations: QAP, Comment, Acknowledge, Q-elab, (Result and Sequence for the situated corpus)
- Traditional discourse markers not relevant for most of these.
- Nonlinguistic actions often had
Structures in multiparty dialogue

- non-treelike structures exist, whereas most theories that investigate full discourse structures.
- long distance, crossing dependencies,
- recursively structured discourse representations (CDUs).
- So tree spanning algorithms like MST are not conceptually right for multiparty dialogue
- Move to DAGs and various algorithms for getting the right sort (ILP)
- Lots of interesting investigations to do about situated discourse structures.
Constraints from conversational structure

- reactivity: turn constraint: people react to prior messages
- there are no backwards attachments (e.g., n. *I can trade* n+k. *if you want*) across turns from different speakers
- except in very rare cases (Sacks)
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

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

a → b

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

Full distribution
All relations distribute to every component

c → d → e

Distributing relations
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

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

[I’ll buy a card]ₐ
[and not a road]ₖ
[because I have sheep]ₙ [and wheat]ₗ [and ore]ₐ
Evaluation setup

Pair modelization: Maximum Entropy model

Graph building methods

- Last: all EDUs are linked to the previous one, forming a single chain.
- Local: pair attachment and labelling models are treated as classifiers.
- MST: Maximum Spanning Tree (Afantenos et al., 2015)
- ILP: Integer Linear Programming

Three CDU replacement strategies
Each one creates its own version of the corpus
### 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>
Comparisons to the best text parsers

<table>
<thead>
<tr>
<th>parser</th>
<th>U</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>HHN16 HILDA</td>
<td>64.7</td>
<td>49.1</td>
</tr>
<tr>
<td>SHV15 D *</td>
<td>64.2</td>
<td>49.5</td>
</tr>
<tr>
<td>JCN15 1S-1S</td>
<td>65.8</td>
<td>48.9</td>
</tr>
<tr>
<td>FH14 gCRF *</td>
<td>67.6</td>
<td>50.1</td>
</tr>
<tr>
<td>LLC16</td>
<td>62.9</td>
<td>42.1</td>
</tr>
<tr>
<td>JE14 DPLP **</td>
<td>69.2</td>
<td>53.7</td>
</tr>
<tr>
<td>BCS17</td>
<td>68.2</td>
<td>51.6</td>
</tr>
<tr>
<td>our method (MST)</td>
<td>69.0</td>
<td>50.8</td>
</tr>
</tbody>
</table>

**Table** – Dependency evaluation. U = unlabelled dependencies, R = relation labelled dependencies.
Analysis of the Results

- ILP outperforms MST even on trees (no-distribution strategy)
- Score gap between ILP and MST widens as edge count increases
- Partial distribution produces the hardest parsing task overall
Analysis of the Results

- ILP outperforms MST even on trees (no-distribution strategy)
- Score gap between ILP and MST widens as edge count increases
- Partial distribution produces the hardest parsing task overall

What the F1 scores don’t tell you

- ILP consistently obtains higher recall
- Inside turns, clusters of edges are frequent
- Long-distance relations incur mixed results
Conclusions

- A small set of features
- A simple local model
- Richer well-formed structures
- Linguistic constraints and generalizations are important for the computational implementation
Future work

- Detection of CDUs from clusters of predicted edges
- Extended constraint sets for specific domains
- Structured prediction influencing local models
- Enhanced representations of discourse units