Constructing the CODA Corpus: a Parallel Corpus of Monologues and Expository Dialogues

Svetlana Stoyanchev and Paul Piwek
The Open University
Motivation

- CODA project (2009-2011): COherent Dialogue Automatically generated from text

Why dialogue?
- Education: observing dialogue motivates students to ask deeper questions (Craig et al. 2000)
- Accessible and entertaining information presentation for children, museum visitors, patients, etc.

Why generate dialogue automatically?
- Most information currently is available in text (books, articles, leaflets, description of art at museums)
- Most people are skilled monologue writers
- Fewer have experience writing dialogues
Uses of Generated Dialogue

- Multimodal presentation of automatically generated dialogues in Second Life
Monologue-to-Dialogue Conversion Approaches

- Manual: construct transformation rules (Piwek et al., 2008) based on Rhetorical relations
  - ATTRIBUTION (“John said”, “that he likes apples”) →
    - A: What did John say?
    - B: that he likes apples
- Empirical: extract transformation rules from a parallel monologue/dialogue corpus
  - Analyse how rhetorical relations are realised in a dialogue
  - Learn dialogue generation rules from parallel dialogue-monologue
Parallel Corpus of Dialogues and Monologues

- Take existing dialogues and convert them to monologue, but then we would miss the professional author requirement.

Alternatively, we could take monologues and convert them to dialogue, but then we would miss the professional author requirement.

Choose dialogues that are easy to paraphrase as monologues, those are expository and non-dramatic dialogues:

- Currenich (20th century, published in a computer science conference proceedings)
- Lewis (20th century, published in a philosophical journal)
- Berkeley (18th century, philosophical literature)
- Mark Twain ("Who Is Man" 19th century, philosophical literature)

Use corpora which can be released to research community.

Use the dialogues written by the professional authors.

Take existing dialogues and convert them to monologue.
Corpus Construction

1. Dialogue annotation and translation
   - Segment dialogue
   - Annotate dialogue acts
   - Translate dialogue into monologue

2. Monologue annotation
   - Annotate discourse relations in monologue

Created a custom-designed annotation tool
<table>
<thead>
<tr>
<th>#</th>
<th>S...</th>
<th>Turn</th>
<th>Tagged Turn</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>O</td>
<td>The others offer you a hundred bribes to be good, thus conceding</td>
<td>17: Resp-Explain:none:The others offer you a hundred bribes to be good,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>that the Master inside of you must be conciliated and</td>
<td>thus conceding that the Master inside of you must be conciliated and</td>
</tr>
<tr>
<td></td>
<td></td>
<td>contented first, and that you will do nothing at FIRST HAND but</td>
<td>contented first, and that you will do nothing at FIRST HAND but for his</td>
</tr>
<tr>
<td></td>
<td></td>
<td>for his sake; then they turn square around and require you to do</td>
<td>sake; then they turn square around and require you to do good for OTHERS'</td>
</tr>
<tr>
<td></td>
<td></td>
<td>good for OTHER'S sake CHIEFLY: and to do your duty for duty!</td>
<td>S's sake CHIEFLY: and to do your duty for duty!</td>
</tr>
<tr>
<td>16</td>
<td>Y</td>
<td>If we grant, for the sake of argument, that your scheme and the</td>
<td>18: Resp-Explain:If we grant, for the sake of argument, that your scheme</td>
</tr>
<tr>
<td></td>
<td></td>
<td>other schemes aim at and produce the same result -- RIGHT</td>
<td>and the other schemes aim at and produce the same result -- RIGHT</td>
</tr>
<tr>
<td></td>
<td></td>
<td>LIVING -- has yours an advantage over the others?</td>
<td>LIVING -- has yours an advantage over the others?</td>
</tr>
<tr>
<td>17</td>
<td>O</td>
<td>One, yes -- a large one. It has no concealments, no deceptions.</td>
<td>19: Resp-Factoid:One, yes -- a large one. It has no concealments, no</td>
</tr>
<tr>
<td></td>
<td></td>
<td>When a man leads a right and valuable life under it he is not</td>
<td>deceptions. When a man leads a right and valuable life under it he is not</td>
</tr>
<tr>
<td></td>
<td></td>
<td>deceived as to the REAL chief motive which impels him to it -- in</td>
<td>deceived as to the REAL chief motive which impels him to it -- in the</td>
</tr>
<tr>
<td></td>
<td></td>
<td>those other cases he is.</td>
<td>other cases he is.</td>
</tr>
<tr>
<td>18</td>
<td>Y</td>
<td>Is that an advantage?</td>
<td>20: Resp-Explain:It has no concealments, no deceptions. When a man leads a</td>
</tr>
<tr>
<td></td>
<td></td>
<td>mean reason? In the course of events it is sure to give an</td>
<td>right and valuable life under it he is not deceived as to the REAL chief</td>
</tr>
<tr>
<td></td>
<td></td>
<td>IMPRESSION that he is an advantage?</td>
<td>motive which impels him to it -- in those other cases he is.</td>
</tr>
<tr>
<td>19</td>
<td>O</td>
<td>Perhaps so. The same advantage he might get out of thinking himself</td>
<td>21: Resp-Explain:Perhaps so. The same advantage he might get out of thinking</td>
</tr>
<tr>
<td></td>
<td></td>
<td>a duke, and living a duke's life and parading in ducal fuss and</td>
<td>himself a duke, and living a duke's life and parading in ducal fuss and</td>
</tr>
<tr>
<td></td>
<td></td>
<td>feathers, when he wasn't a duke at all, and could find it out if he</td>
<td>feathers, when he wasn't a duke at all, and could find it out if he would</td>
</tr>
<tr>
<td></td>
<td></td>
<td>only examine the herald's records.</td>
<td>only examine the herald's records.</td>
</tr>
<tr>
<td>20</td>
<td>Y</td>
<td>But anyway, he is obliged to hand in his pocket and as he can stand,</td>
<td>22: Resp-Explain:Perhaps so. The same advantage he might get out of thinking</td>
</tr>
<tr>
<td></td>
<td></td>
<td>and as he can stand, and</td>
<td>himself a duke, and living a duke's life and parading in ducal fuss and</td>
</tr>
<tr>
<td></td>
<td></td>
<td>and as he can stand, and</td>
<td>feathers, when he wasn't a duke at all, and could find it out if he would</td>
</tr>
<tr>
<td></td>
<td></td>
<td>as he can stand, and</td>
<td>only examine the herald's records.</td>
</tr>
<tr>
<td>21</td>
<td>O</td>
<td>He could do that.</td>
<td>23: Resp-Answer:Yes:Perhaps so. The same advantage he might get out of</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>thinking himself a duke, and living a duke's life and parading in ducal</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>fuss and feathers, when he wasn't a duke at all, and could find it out if</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>he would only examine the herald's records.</td>
</tr>
</tbody>
</table>
The difference between straight speaking and crooked; the difference between frankness and shuffling.
1. Identify *Decorative* and *key* segments

- *Decorative* information is included by the dialogue author to liven up the dialogue and create natural transitions in dialogue (e.g. “Wait!”, “Well ”, “You think so?”)
  - Decorative segments are not translated into dialogue
  - Correspond to dialogue management acts (Bunt's DIT theory)

- *Key* information contributes to the main idea that the author is trying to get across. It is directly relevant to the main purpose of the dialogue
  - Key segments are translated into dialogue
  - Most segments are *key*

2. Assign dialogue act to *key* segments
**Dialogue Act Annotations**

**Dialogue Act** annotations are based on the scheme originally defined by Carletta et al. (1997)

1. **Initiation Dialogue acts:**
   - Yes/no info request
   - Factoid info request (who, when, where question)
   - Complex info request (why, how, what question)
   - Clarification request

2. **Responding dialogue acts:**
   - Agree
   - Contradict
   - Answer-factoid
   - Answer-yes/no

3. **Explanation**

**Compute inter-annotator agreement:**
- Dialogue act segmentation agreement: 91 %
- Overall dialogue act tagging agreement kappa=.82
**Translation Guidelines**

- Annotators (authors) write a close paraphrase of the dialogue
  - No new information is introduced in the monologue
  - No information that is in dialogue is omitted in the monologue

<table>
<thead>
<tr>
<th>Utterance</th>
<th>Dialogue Act</th>
<th>Monologue</th>
</tr>
</thead>
<tbody>
<tr>
<td>OM: He felt well?</td>
<td>YN-Info-Request Explain</td>
<td>One can not doubt that he felt well.</td>
</tr>
<tr>
<td>YM: One can not doubt it.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>OM: Are the metals suddenly deposited in the ores?</td>
<td>Init-YN-InfoReq</td>
<td>The metals are not suddenly deposited in the ores. It is the patient work of countless ages.</td>
</tr>
<tr>
<td>YM: No --</td>
<td>Resp-Answer-No Explain</td>
<td></td>
</tr>
<tr>
<td>YM: it is the patient work of countless ages.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The Open University
Monologue Annotation

Based on Carlson et al. (2001) RST annotation scheme which defines

- Text segmentation into elementary discourse units
- Discourse (rhetorical) relation assignment between spans of text

Modifications to the scheme:

- Adapted the set of tags: 2-level annotation (coarse-grain 17 tags, fine-grain 26 tags)
- Segment monologue into paragraphs, annotate relations within a paragraph.
- Inter-annotator agreement kappa=.68 (with the pair of most-confusable tags merged)

Use O'Donnel (2000) annotation tool:
Corpus Statistics

Dialogue Act Tags in the dialogue

RST relations in monologue
# Examples of Structural M2D Rules

<table>
<thead>
<tr>
<th>Rule</th>
<th>RST relation in monologue</th>
<th>→ Dialogue act sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Example</td>
<td>Attribution</td>
<td>YN-Info-Request Explain</td>
</tr>
<tr>
<td>(Twain)</td>
<td>[One can not doubt ] [that he felt well]</td>
<td>He felt well? One can not doubt it.</td>
</tr>
<tr>
<td>Rule</td>
<td>Contrast</td>
<td>Init-YN-InfoReq Resp-Answer-No Explain</td>
</tr>
<tr>
<td>Example</td>
<td>[The metals are not suddenly deposited in the ores.] [It is the patient work of countless ages.]</td>
<td>Are the metals suddenly deposited in the ores? No -- it is the patient work of countless ages.</td>
</tr>
</tbody>
</table>

The Open University
Applying M2D Mapping Rules

Examples of Attribution relations extracted from the corpus, applied to

**Attribution** ("John said", "that he likes apples")

- **RULE1**: Attribution $\rightarrow$ A: Init-YN-InfoReq; B: Resp-Yes Explain
  - A: Does John like apples?
  - B: Yes, he said so.

- **RULE2**: Attribution $\rightarrow$ A: Init-YN-InfoReq; B: Resp-Yes
  - A: Did John say that he like apples?
  - B: Yes

- **RULE3**: Attribution $\rightarrow$ A: Explain; B: Resp-Agree
  - A: John said that he likes apples.
  - B: Yes, he did.
Conclusions

- Described and motivated construction of a parallel monologue-dialogue corpus
- Adopted, applied, and evaluated annotation schemes
  - dialogue annotation scheme to authored dialogues
  - RST annotations to monologue
- Final corpus will contain 1000 dialogue turns and will be available in June.
- http://computing.open.ac.uk/coda/
Thank you

Svetlana Stoyancheva and Paul Piwek

http://computing.open.ac.uk/coda/
## Extra slides

<table>
<thead>
<tr>
<th>Rule</th>
<th>Attribution</th>
<th>Explain Resp-Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>Example (Twain)</td>
<td><em>[A minute ago I said ] [that Hamilton fought that duel to get PUBLIC approval .]</em></td>
<td>A minute ago you said Hamilton fought that duel to get PUBLIC approval . I did</td>
</tr>
</tbody>
</table>