Miscommunicating with Computers

Matthew Purver
with many others
but mostly Julian Hough and Christine Howes
Repair Phenomena

A: It’s in the **yell-**, **sorry**, **purple** square
B: The **what** square?
A: **Purple**.
B: You mean **yellow**.
A: Oh yeah. **Yellow**.

• Self- and other-repair
• Clarification (repair initiation)
• Positions 1, 2, 3, ...
Dialogue Systems

• Clarification & confirmation by systems:

U: When’s the next bus to Wood Street?
S: Sorry, I didn’t understand that. Please repeat.

U: When’s the next bus to Wood Street?
S: Going to WOOD STREET. Did I get that right?
U: Yes.

• ASR inaccuracy
• Domain mismatch
• Generally confidence-based (ASR/domain)
Clarification by Systems

• Designing optimal questions (e.g. Stoyanchev et al 2014):

  U: Do you have anything other than the [???] plans
  S: Which plans? / Anything other than what?

• Learning to adjust strategy/confidence (e.g. Bohus & Rudnicky 2005)

• Integration with general decision process and domain learning (e.g. PARLANCE project)

• But these don’t give us methods for *detecting* repair
Clarification by Users

TTS: I calculated a route to Brian’s house. It is 4.13 miles long and should take around 9.0 minutes.
SR: what is my next turn
TTS: Go East on Miranda Ave
SR: what is the next turn
TTS: Make sharp right on Arastradero Rd
SR: what road
TTS: I’m sorry, I do not understand what you mean by: what road
TTS: Please try rephrasing that.
SR: what road do i turn on
TTS: I’m sorry, I do not understand what you mean by: what road do i turn on
TTS: Please try rephrasing that.
SR: what road do i turn on and where is alma street from there
TTS: Did you want to plan a route from there?
SR: what was the last instruction you gave me
TTS: Sorry, there is no instruction to repeat
SR: what was the last route instruction you gave me
TTS: Sorry, there is no instruction to repeat
SR: i want to go to brian's
TTS: Ok, I will plan a route to Brian. Would you like anything else?
Self-Repair

• Disfluency detection for ASR
  – Identify reparandum extent and remove
  – e.g. transduction: Johnson & Charniak, 2004

  ...a flight to Boston, uh, I mean, toDenver on Friday ...

  Reparandum  Interregnum  Repair

  a flight to Boston to Denver on Friday

• But this is stuff we need! E.g. for anaphoric reference:
  – “The interview was .... it was alright” (Clark, 1996)
Self-Repair is incremental

• Effect on incremental processing (Brennan & Schober, 2001)
  – See DYLAN system (Hough & Purver, 2014)
Self-Repair models

• We’d like a model which is:
  – Incremental
  – Able to track context contributions

• Existing models either:
  – Lose the reparandum and/or repair
    • (e.g. Johnson & Charniak, 2004)
  – Need the whole sequence
    • (e.g. Georgila, 2009)
  – Work incrementally but maintain all hypotheses
    • (e.g. Heeman & Allen, 1999; Zwarts et al, 2010)
Human-Human Repair

• Language processing for psychiatric therapy:
  – Diagnosing symptoms
  – Predicting outcomes

Yeh, it doesn’t happen in real life, does it?

What do you mean by real life?

You can’t—there are no messages coming from the television to people are there?
Doctor-patient communication

• Schizophrenia therapy (face-to-face)
  – Symptoms and severity:
    • Positive symptoms: delusions, hallucinations, beliefs
    • Negative symptoms: withdrawal, blunted affect, alogia
  – Non-adherence to treatment:
    • About half of patients non-adherent in the year after discharge from hospital (Weiden & Olfson, 1995)
    • Risk of relapse 3.7 times higher (Fenton et al, 1997)

• Depression & anxiety therapy (online)
  – Symptoms and severity
  – Progress and dropout rates

• Can features of dialogue help understand/predict?
  – Topic structure: focus on symptoms, treatment
  – Repair structure: coordination, shared understanding
Prediction Results

- Predicting symptom severity reasonable:
  - Depression (PHQ) 70\%
  - Schizophrenia (PANSS) 62\%
  - (with topic/sentiment features)
- Predicting non-adherence (patient turns only):

<table>
<thead>
<tr>
<th>Features</th>
<th>Weighted F (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline: class of interest</td>
<td>44.8</td>
</tr>
<tr>
<td>Human: text only</td>
<td>68.6</td>
</tr>
<tr>
<td>Human: text + video</td>
<td>78.0</td>
</tr>
<tr>
<td>Lexical features</td>
<td>70.3</td>
</tr>
<tr>
<td>Topic features</td>
<td>66.2</td>
</tr>
<tr>
<td>Automatic topics</td>
<td>54.1</td>
</tr>
</tbody>
</table>

- So how can we improve this? Repair ...
Repair in Therapy Dialogue

• Self-repair (e.g. P1SISR, P3SISR)

• Articulation, formulation

Dr: You probably have seen so many psychiatrists o o over the years

Dr: Did you feel that did you despair so much that you wondered if you could carry on

P: Where I go to do some printing lino printing

Dr: Clorazil or
P: Yeah
Dr: Clozapine yes
Repair in Therapy Dialogue

• Other-repair (e.g. P2OIOR)

Dr: Rather than the diazepam which I don’t think is going to do you any good
P: the valium

• Repair initiation (e.g. P2NTRI then P3OISR)

Dr: Yeh, it doesn’t happen in real life does it?
P: What do you mean by real life?
Dr: You can’t - there are no messages coming from the television to people are there?
Comparison with other contexts

- Therapy: more self-repair, less other-repair & initiation
Patient-doctor comparison

- Patients: more self-repair, less other-repair & initiation
Detecting Other-Repair

• Need to detect instances of repair
  – Next-turn repair initiation and P2 repair
• How do we approach this?

• Define some features that characterise repair
  – (which we can extract automatically)
• Learn a statistical model
  – (using some standard machine learning algorithm)
Features

• Sometimes we see specific lexical / phrasal items:

  Dr: Ok you have done it before
  P:  Pardon?
  Dr: If you have done it before

  Dr: Who is your GP now
  P:  What?
  Dr: Who is your GP

  P:  They’re not negative erm but they’re positive as i eh erm it’s like imagining how your life will be
  Dr: Ok, ok, ok so thinking about how
  P:  Do you know what I’m talking about?
Features

• Sometimes we see repetition/parallelism:

Dr: Yep well that is a possible side effect
P: Side effect?
Dr: Of the err Haliperidol

Dr: One thing that I ask you is when you were low in mood did you have suicidal thoughts
P: Did I have ...?
Dr: Suicidal thoughts
Features

• Sometimes it’s more complex than that

Wiz: go straight for four blocks turn left at wall street
Subj: turn left where
Wiz: turn left at wall street

TTS: Make sharp right on Arastradero Rd
SR: what road

Wiz: after left at elm street turn right at lois lane
Subj: was that right on lois lane or left on lois lane
Wiz: turn right at lois lane
Features

• Sometimes it’s more complex than that:

Dr: Are you suspicious are you suspicious of people
P: **Suspicious**?
Dr: Paranoid
P: **Jealous**?
Dr: Jealous yeah

Dr: Paroxitine
P: **Fluoxetine**
Dr: Ah Fluoxetine
Features

• Sometimes it’s more complex than that

Dr: Who’s your key worker there do you know
P:  Err the person who comes to see me?
Dr: Yeah the person you see most often

Dr: Do you do you really feel it or is it a sensation
P:  Is it what I’m thinking is that what you mean?
Dr: No is it just err the mind playing tricks on you

Dr: was it couple of months three months
P:  Since I saw you?

Dr: Aaa so have you had any more thoughts about studying
P:  What music?
Features

• Sometimes it’s more complex than that

Wiz: go straight for three blocks turn right at wall street
Subj: please repeat left where
Wiz: go straight for three blocks turn right at wall street
Subj: left where

Subj: how long
Wiz: dave's house is sixteen minutes away
Subj: was that one six or six zero minutes
Wiz: six minutes away
Requirements

• They’re context-dependent
• They need semantics
• They need phonology
• They even need spelling

• They’re incremental
• They can be very sparse
Detecting Other-Repair

• Define features manually, extract automatically
• Linguistically/observationally informed:
  – Wh-question words, closed class repair words
  – Repetition, parallelism with prior turn(s)
  – Backchannel behaviour, fillers
  – Pauses, overlaps
• Brute force:
  – All the unigrams used (patient-only to avoid doctor specificity)
• Train SVMs to detect NTRIs & P2Rs
  – 44,000 turns of which 567 NTRIs (159 patient), 830 P2Rs (262)
  – 5-fold cross-validation
  – (Howes et al, 2012-14)
Results – balanced data

• Balanced data (i.e. small dataset), patient only:

<table>
<thead>
<tr>
<th>Target</th>
<th>Features</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NTRI</td>
<td>Repeated proportion</td>
<td>61.2</td>
</tr>
<tr>
<td>NTRI</td>
<td>All high-level</td>
<td>83.2</td>
</tr>
<tr>
<td>NTRI</td>
<td>All unigrams</td>
<td>82.4</td>
</tr>
<tr>
<td>NTRI</td>
<td>All features</td>
<td>86.3</td>
</tr>
<tr>
<td>P2R</td>
<td>Repeated proportion</td>
<td>61.5</td>
</tr>
<tr>
<td>P2R</td>
<td>All high-level</td>
<td>78.5</td>
</tr>
<tr>
<td>P2R</td>
<td>All unigrams</td>
<td>77.1</td>
</tr>
<tr>
<td>P2R</td>
<td>All features</td>
<td>79.8</td>
</tr>
</tbody>
</table>

• But of course the real data’s not balanced ...
Results – repair detection

- On balanced data: accuracy 80-86%
- Full dataset, patient only:

<table>
<thead>
<tr>
<th>Target</th>
<th>Features</th>
<th>P (%)</th>
<th>R (%)</th>
<th>F (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NTRI</td>
<td>OCRProportion</td>
<td>85.7</td>
<td>22.6</td>
<td>35.8</td>
</tr>
<tr>
<td>NTRI</td>
<td>All high-level</td>
<td>42.8</td>
<td>40.6</td>
<td>41.4</td>
</tr>
<tr>
<td>NTRI</td>
<td>All features</td>
<td>44.9</td>
<td>43.6</td>
<td>44.0</td>
</tr>
<tr>
<td>P2R</td>
<td>OCRProportion</td>
<td>56.4</td>
<td>11.8</td>
<td>19.6</td>
</tr>
<tr>
<td>P2R</td>
<td>All high-level</td>
<td>36.2</td>
<td>28.4</td>
<td>31.6</td>
</tr>
<tr>
<td>P2R</td>
<td>All features</td>
<td>43.8</td>
<td>30.3</td>
<td>35.4</td>
</tr>
</tbody>
</table>

- We’d like to do better!
  - Audio/video: intonation, non-verbal behaviour
  - Context: follow-up dialogue turns incl. other-person reaction
  - Semantic and pragmatic parallelism
Detecting Self-repair

- No grammar; no similar data ...
- Probabilistic / information-theoretic model
  - (Hough, to appear)

\[
\text{John and Bill} \ [ \text{like} + \{\text{uh}\} \ 	ext{love} ] \ 	ext{Mary}
\]

\(\text{original utterance} \hspace{0.5cm} \text{reparandum} \hspace{0.5cm} \text{interregnum} \hspace{0.5cm} \text{repair} \hspace{0.5cm} \text{continuation}\)

- Interregnum: characteristic words, fillers
- Repair & reparandum boundaries: changes and/or (dis)similarities in \textit{probability} and \textit{expectation} (lexical, syntactic, semantic)
- Incremental process:
  1: repair onset  
  2: reparandum start  
  3: repair end
Lexico-syntactic distribution

\[
p^{\text{lex}}(w_{i-2} \ldots w_i) = -\log_2 p^{\text{kn}}(w_i \mid w_{i-2}, w_{i-1})
\]

\[
\text{WML}(w_{i-2} \ldots w_i) = \frac{\log_2 p^{\text{kn}}_{\text{TRIGRAM}}(\langle w_{i-2} \ldots w_i \rangle)}{-\log_2 p^{\text{kn}}_{\text{UNIGRAM}}(\langle w_{i-2} \ldots w_i \rangle)}
\]
Self-repair: Results

• Accuracy on Switchboard corpus (held-out):

<table>
<thead>
<tr>
<th>detection</th>
<th>precision</th>
<th>recall</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>$w_{rp}^{start}$</td>
<td>0.862</td>
<td>0.755</td>
<td>0.805</td>
</tr>
<tr>
<td>repairs in turn</td>
<td>0.904</td>
<td>0.787</td>
<td>0.841</td>
</tr>
</tbody>
</table>

• Good, and incremental!
  – Comparable to (Zwarts et al, 2010) ... but 1 word, not 4.6

• Accuracy on therapy corpus:

<table>
<thead>
<tr>
<th>detection</th>
<th>precision</th>
<th>recall</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>$w_{rp}^{start}$</td>
<td>0.527</td>
<td>0.536</td>
<td>0.532</td>
</tr>
<tr>
<td>repairs in turn</td>
<td>0.682</td>
<td>0.679</td>
<td>0.680</td>
</tr>
</tbody>
</table>

• Good for coarse-grained measures (correlation 0.9)
• But not yet good in detail
What have we learnt?

• We need computational models of repair
  – But different from standard ones

• We can do a reasonable job
  – On self- and other-repair
  – Using fairly low-level features

• Doing better is a difficult task:
  – Semantics, pragmatics, phonology, intention ... ?