You Talking to Me? A Corpus and Algorithm for Conversation Disentanglement

Micha Elsner and Eugene Charniak

Brown Laboratory for Linguistic Information Processing (BLLIP)
Life in a Multi-User Channel

Does anyone here shave their head?
I shave part of my head.
A tonsure?
Nope, I only shave the chin.

How do I limit the speed of my internet connection?
Use dialup!
Hahaha :P No I can’t, I have a weird modem.
I never thought I’d hear ppl asking such insane questions...
# Real Life in a Multi-User Channel

<table>
<thead>
<tr>
<th>Question</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Does anyone here shave their head?</td>
<td>I shave part of my head.</td>
</tr>
<tr>
<td>How do I limit the speed of my internet connection?</td>
<td>Use dialup!</td>
</tr>
<tr>
<td>A tonsure?</td>
<td>A common situation:</td>
</tr>
<tr>
<td></td>
<td>– Text chat</td>
</tr>
<tr>
<td></td>
<td>– Push-to-talk</td>
</tr>
<tr>
<td></td>
<td>– Cocktail party</td>
</tr>
<tr>
<td>Nope, I only shave the chin.</td>
<td></td>
</tr>
</tbody>
</table>
Why Disentanglement?

• A natural discourse task.
  – Humans do it without any training.
• Preprocess for search, summary, QA.
  – Recover information buried in chat logs.
• Online help for users.
  – Highlight utterances of interest.
  – Already been tried manually: Smith et al ‘00.
  – And automatically: Aoki et al ‘03.
Outline

- Corpus
  - Annotations
  - Metrics
  - Agreement
  - Discussion

- Modeling
  - Previous Work
  - Classifier
  - Inference
  - Baselines
  - Results
Dataset

- Recording of a Linux tech support chat room.
- 1:39 hour test section.
  - Six annotations.
  - College students, some Linux experience.
- Another 3 hours of annotated data for training and development.
  - Mostly only one annotation by experimenter.
  - A short pilot section with 3 more annotations.
Annotation

- Annotation program with simple click-and-drag interface.
- Conversations displayed as background colors.
One-to-One Metric

Two annotations of the same dataset.
One-to-One Metric

Transform according to the optimal mapping:

Whole document considered at once.

Annotator one

Transformed

Annotator two
One-to-One Metric

Transform according to the optimal mapping:

Whole document considered at once.

Annotator one

Transformed

Annotator two

70%
Local Agreement Metric

Sliding window: agreement is calculated in each neighborhood of three utterances.

Annotator 1

Annotator 2
Local Agreement Metric

Annotator 1

Annotator 2

Same or different?

Different

Different

Same
Local Agreement Metric

Annotator 1

Annotator 2

66%
# Interannotator Agreement

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>Mean</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>One-to-One</td>
<td>36</td>
<td>53</td>
<td>64</td>
</tr>
<tr>
<td>Local Agreement</td>
<td>75</td>
<td>81</td>
<td>87</td>
</tr>
</tbody>
</table>

- Local agreement is good.
- One-to-one not so good!
### How Annotators Disagree

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>Mean</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td># Conversations</td>
<td>50</td>
<td>81</td>
<td>128</td>
</tr>
<tr>
<td>Entropy</td>
<td>3</td>
<td>4.8</td>
<td>6.2</td>
</tr>
</tbody>
</table>

- Some annotations are much finer-grained than others.
Schisms

• Sacks et al ‘74: Formation of a new conversation.

• Explored by Aoki et al ‘06:
  – A speaker may start a new conversation on purpose...
  – Or unintentionally, as listeners react in different ways.

• Causes a problem for annotators...
I grew up in Romania till I was 10. Corruption everywhere.

And my parents are crazy. Couldn’t stand life so I dropped out of school.

You’re at OSU?

Man, that was an experience.

You still speak Romanian?

Yeah.
I grew up in Romania till I was 10. Corruption everywhere.

And my parents are crazy. Couldn’t stand life so I dropped out of school.

You’re at OSU?

Man, that was an experience.

You still speak Romanian?

Yeah.
Accounting for Disagreements

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>Mean</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>One-to-One</td>
<td>36</td>
<td>53</td>
<td>64</td>
</tr>
<tr>
<td>Many-to-One</td>
<td>76</td>
<td>87</td>
<td>94</td>
</tr>
</tbody>
</table>

Many-to-one mapping from high entropy to low:

First annotation is a strict refinement of the second.

One-to-one: only 75%
Many-to-one: 100%
Pauses Between Utterances

A classic feature for models of multiparty conversation.

Peak at 1-2 sec. (turn-taking)

Heavy tail
Is there an easy way to extract files from a patch?

Sara: No.

Carly: Patches are diff deltas.

Carly, duh, but this one is just adding entire files.

- Very frequent: about 36% of utterances.
- A coordination strategy used to make disentanglement easier.
  - O’Neill and Martin ‘03.
- Usually part of an ongoing conversation.
Outline

• Corpus
  – Annotations
  – Metrics
  – Agreement
  – Discussion

• Modeling
  – Previous Work
  – Classifier
  – Inference
  – Baselines
  – Results
Previous Work

- Aoki et al ‘03, ‘06
  - Conversational speech
  - System makes speakers in the same thread louder
  - Evaluated qualitatively (user judgments)

- Camtepe ‘05, Acar ‘05
  - Simulated chat data
  - System intended to detect social groups
Previous Work

• Based on pause features.
  – Acar ‘05: adds word repetition, but not robust.

• All assume one conversation per speaker.
  – Aoki ‘03: assumed in each 30-second window.
Conversations Per Speaker

Average of 3.3
Our Method: Classify and Cut

- Common NLP method: Roth and Yih ‘04.
- Links based on max-ent classifier.
- Greedy cut algorithm.
  - Found optimal too difficult to compute.
Classifier

- Pair of utterances: same conversation or different?

- Chat-based features (F 66%):
  - Time between utterances
  - Same speaker
  - Name mentions

- Most effective feature set.
Classifier

- Pair of utterances: same conversation or different?

- Chat-based features (F 66%)

- Discourse-based (F 58%):
  - Detect questions, answers, greetings &c

- Lexical (F 56%):
  - Repeated words
  - Technical terms
Classifier

• Pair of utterances: same conversation or different?

• Chat-based features (F 66%)
• Discourse-based (F 58%)
• Lexical (F 56%)
• Combined (F 71%)
Inference

Greedy algorithm: process utterances in sequence

Classifier marks each pair “same” or “different” (with confidence scores).

Pro: online inference
Con: not optimal
Inference

Greedy algorithm:
- process utterances in sequence
  - Treat classifier decisions as votes.
  - Pro: online inference
  - Con: not optimal
Inference

Greedy algorithm: process utterances in sequence

Treat classifier decisions as votes.

Color according to the winning vote.

If no vote is positive, begin a new thread.

Pro: online inference
Con: not optimal
Baseline Annotations

- All in same conversation
- All in different conversations
- Speaker’s utterances are a monologue

- Consecutive blocks of $k$
- Break at each pause of $k$
  - Upper-bound performance by optimizing $k$ on the test data.
## Results

<table>
<thead>
<tr>
<th></th>
<th>Humans</th>
<th>Model</th>
<th>Best Baseline</th>
<th>All Diff</th>
<th>All Same</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max 1-to-1</td>
<td>64</td>
<td>51</td>
<td>56 (Pause 65)</td>
<td>16</td>
<td>54</td>
</tr>
<tr>
<td>Mean 1-to-1</td>
<td>53</td>
<td>41</td>
<td>35 (Blocks 40)</td>
<td>10</td>
<td>21</td>
</tr>
<tr>
<td>Min 1-to-1</td>
<td>36</td>
<td>34</td>
<td>29 (Pause 25)</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>Max local</td>
<td>87</td>
<td>75</td>
<td>69 (Speaker)</td>
<td>62</td>
<td>57</td>
</tr>
<tr>
<td>Mean local</td>
<td>81</td>
<td>73</td>
<td>62 (Speaker)</td>
<td>53</td>
<td>47</td>
</tr>
<tr>
<td>Min local</td>
<td>75</td>
<td>70</td>
<td>54 (Speaker)</td>
<td>43</td>
<td>38</td>
</tr>
</tbody>
</table>
One-to-One Overlap Plot

Some annotators agree better with baselines than other humans...
Local Agreement Plot

All annotators agree first with other humans, then the system, then the baselines.
Mention Feature

• Name mention features are critical.
  – When they are removed, system performance drops to baseline.
• But not sufficient.
  – With only name mention and time gap features, performance is midway between baseline and full system.
Plenty of Work Left

• Annotation standards:
  – Better agreement
  – Hierarchical system?

• Speech data
  – Audio channel
  – Face to face

• Improve classifier accuracy

• Efficient inference

• More or less specific annotations on demand
Data and Software is Free

- Available at:
  www.cs.brown.edu/~melsner

- Dataset (text files)
- Annotation program (Java)
- Analysis and Model (Python)
Acknowledgements

- Suman Karumuri and Steve Sloman
  - Experimental design
- Matt Lease
  - Clustering procedure
- David McClosky
  - Clustering metrics (discussion and software)
- 7 test and 3 pilot annotators
- 3 anonymous reviewers
- NSF PIRE grant