Recall: Extractive Summarization

- Count number of documents each bigram occurs in to measure importance
  
  \[
  \text{score(massive earthquake)} = 3 \quad \text{score(magnitude 7.3)} = 2 \quad \text{score(six killed)} = 2 \quad \text{score(Iraqi capital)} = 1
  \]

- Find summary that maximizes the score of bigrams it covers

- ILP formulation: \( c \) and \( s \) are indicator variables indexed over concepts (bigrams) and sentences, respectively

  \[
  \begin{align*}
  \text{Maximize:} & \quad \sum_i w_i c_i \quad \text{s.t.} \\
  \text{subject to:} & \quad \sum_j l_j s_j \leq L \quad \sum_j s_j O_{C_i j} \geq c_i \quad \forall i
  \end{align*}
  \]
  
  “set \( c \) to 1 iff some sentence that contains it is included”

  \[
  \text{sum of included sentences' lengths can’t exceed } L
  \]

  Gillick and Favre (2009)

Recall: Compression

- Maximize:

  \[
  \sum_i w_i c_i \quad \text{s.t.} \\
  \sum_j s_j O_{C_i j} \leq c_i \quad \forall i, j
  \]

- Subject to:

  \[
  \sum_j l_j s_j \leq L \quad \sum_j s_j O_{C_i j} \geq c_i \quad \forall i
  \]

- New constraint: \( s_2 \leq s_1 \)

  “\( s_1 \) is a prerequisite for \( s_2 \)”

- Now \( s \) variables are nodes or sets of nodes in the parse tree

Administrivia

- Proposal due today at 5pm

- Proposals returned ASAP, then Project 2

Recall: Compression

At least six people were killed and many others injured
Recall: Pointer-Generator

See et al. (2017)

Recall: Seq2seq Summarization

- Solutions: copy mechanism, coverage, just like in MT...
- Things might still go wrong, no way of preventing this...

See et al. (2017)

This Lecture

- Chatbot dialogue systems
- Task-oriented dialogue
- Other dialogue applications

Chatbots
### Turing Test (1950)

- Imitation game: A and B are locked in rooms and answer C’s questions via typewriter. Both are trying to act like B

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>B</td>
<td>C trained judge</td>
</tr>
</tbody>
</table>

**Original Interpretation:**

- The test is *not* “does this computer seem human-like to random people with a web browser?”

<table>
<thead>
<tr>
<th>B</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>C trained judge</td>
<td></td>
</tr>
</tbody>
</table>

**Standard Interpretation:**

<table>
<thead>
<tr>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>trained judge</td>
</tr>
</tbody>
</table>

### ELIZA

- Created 1964-1966 at MIT, heavily scripted

<table>
<thead>
<tr>
<th>User:</th>
<th>ELIZA:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Why do you think I $2 you?</td>
<td>You don’t argue with me.</td>
</tr>
<tr>
<td>What else comes to mind when you think of your $1?</td>
<td>WHY DO YOU THINK I DON’T ARGUE WITH YOU</td>
</tr>
</tbody>
</table>

**DOCTOR script was most successful:**

- repeats user’s input, asks inane questions

**Standard Interpretation:**

- Sticks quite closely to particular rules, can apply rules to keywords and contexts
- Very little need to generate new content, but can only have one type of conversation

<table>
<thead>
<tr>
<th>User:</th>
<th>Cleverbot:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hi there cleverbot.</td>
<td>How are you?</td>
</tr>
<tr>
<td>How are you?</td>
<td>Hi there, how are you?</td>
</tr>
<tr>
<td>Doing okay. what plans do you have today?</td>
<td>I’m doing okay, how about you?</td>
</tr>
<tr>
<td>Cleaverbot: I do revision.</td>
<td>Cleaverbot: I don’t know, farming maybe</td>
</tr>
<tr>
<td>Cleaverbot: Maths history.</td>
<td>Cleaverbot: That’s none of your business.</td>
</tr>
<tr>
<td>User: How old are you?</td>
<td>User: What is maths history?</td>
</tr>
</tbody>
</table>

**Cleverbot:**

- Carpenter (1986), online system built in 2006

- “Nearest neighbors”: human says statement A, find a human response in human-human or human-computer chats to statement A, repeat that

**Standard Interpretation:**

- Can often give sensible answers, but the bot doesn’t really impose high-level discourse structure

<table>
<thead>
<tr>
<th>User:</th>
<th>Cleverbot:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hi there cleverbot.</td>
<td>How are you?</td>
</tr>
<tr>
<td>How are you?</td>
<td>Hi there, how are you?</td>
</tr>
<tr>
<td>Doing okay. what plans do you have today?</td>
<td>I’m doing okay, how about you?</td>
</tr>
<tr>
<td>Cleaverbot: I do revision.</td>
<td>Cleaverbot: I don’t know, farming maybe</td>
</tr>
<tr>
<td>Cleaverbot: Maths history.</td>
<td>Cleaverbot: That’s none of your business.</td>
</tr>
<tr>
<td>User: How old are you?</td>
<td>User: What is maths history?</td>
</tr>
</tbody>
</table>

**Weizenbaum (1966)**
Data-Driven Approaches

- Can treat as a machine translation problem: “translate” from current utterance to next one
  
  ![Graphical representation of a machine translation problem]

- Filter the data, use statistical measures to prune extracted phrases to get better performance

  Ritter et al. (2011)

Seq2seq models

- Just like conventional MT, can train seq2seq models for this task

  ![Graphical representation of a seq2seq model]

- Why might this model perform poorly? What might it be bad at?

  System | BLEU
  --- | ---
  RANDOM | 0.33
  MT | 3.21
  HUMAN | 6.08

Lack of Diversity

- Training to maximize likelihood gives a system that prefers common responses:

<table>
<thead>
<tr>
<th>Input: What are you doing?</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.86 I don’t know.</td>
</tr>
<tr>
<td>-1.03 I don’t know!</td>
</tr>
<tr>
<td>-1.06 Nothing.</td>
</tr>
<tr>
<td>-1.09 Get out of the way.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Input: What is your name?</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.91 I don’t know.</td>
</tr>
<tr>
<td>-0.92 I don’t know!</td>
</tr>
<tr>
<td>-0.92 I don’t know, sir.</td>
</tr>
<tr>
<td>-0.97 Oh, my god!</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Input: How old are you?</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.79 I don’t know.</td>
</tr>
<tr>
<td>-1.06 I’m fine.</td>
</tr>
<tr>
<td>-1.17 I’m all right.</td>
</tr>
<tr>
<td>-1.17 I’m not sure.</td>
</tr>
</tbody>
</table>

  Li et al. (2016)
Lack of Diversity

- Solution: mutual information criterion; response R should be predictive of user utterance U as well
- Standard conditional likelihood: \( \log P(R|U) \)
- Mutual information: \( \log \frac{P(R, U)}{P(R)P(U)} = \log P(R|U) - \log P(R) \)
- \( \log P(R) \) can reflect probabilities under a language model

Li et al. (2016)

Future of chatbots

- How deep can a conversation be without more semantic grounding? Basic facts aren’t even consistent...
- Can force chatbots to give consistent answers, but still probably not very interesting
- XiaoIce: Microsoft chatbot in Chinese, 20M users, average user interacts 60 times/month
- People do seem to like talking to them...?

Li et al. (2016) Persona...

Task-Oriented Dialogue
Task-Oriented Dialogue

- Question answering/search:
  - Google, what’s the most valuable American company?
    - Apple
  - Who is its CEO?
    - Tim Cook

Task-Oriented Dialogue

- Personal assistants / API front-ends:
  - Siri, find me a good sushi restaurant in Chelsea
    - Sushi Seki Chelsea is a sushi restaurant in Chelsea with 4.4 stars on Google
    - How expensive is it?
      - Entrees are around $30 each
    - Find me something cheaper

Task-Oriented Dialogue

- Personal assistants / API front-ends:
  - Hey Alexa, why isn’t my Amazon order here?
    - Let me retrieve your order. Your order was scheduled to arrive at 4pm today.
    - It never came
  - Okay, I can put you through to customer service.

Air Travel Information Service (ATIS)

- Given an utterance, predict a domain-specific semantic interpretation
  - Utterance: How much is the cheapest flight from Boston to New York tomorrow morning?
    - Goal: Airfare
    - Cost.Relative: cheapest
    - Depart.City: Boston
    - Arrival.City: New York
    - Depart.Date.Relative: tomorrow
    - Depart.Time.Period: morning

- Can formulate as semantic parsing, but simple slot-filling solutions (classifiers) work well too

DARPA (early 1990s), Figure from Tur et al. (2010)
Full Dialogue Task

- Parsing / language understanding is just one piece of a system
- Dialogue state: reflects any information about the conversation (e.g., search history)
- User utterance \(\rightarrow\) update dialogue state \(\rightarrow\) take action (e.g., query the restaurant database) \(\rightarrow\) say something
- Much more complex than chatbots!

Young et al. (2013)

---

POMDP-based Dialogue Systems

- POMDP: user is the “environment,” an utterance is a noisy signal of state
- Dialogue model: can look like a parser or any kind of encoder model
- Generator: use templates or seq2seq model
- Where do rewards come from?

Young et al. (2013)

---

Full Dialogue Task

Find me a good sushi restaurant in Chelsea

```
restaurant_type <- sushi
location <- Chelsea
curr_result <- execute_search()
```

Sushi Seki Chelsea is a sushi restaurant in Chelsea with 4.4 stars on Google

How expensive is it?

```
get_value(cost, curr_result)
```

Entrees are around $30 each

---

Reward for completing task?

Find me a good sushi restaurant in Chelsea

```
restaurant_type <- sushi
location <- Chelsea
curr_result <- execute_search()
```

Sushi Seki Chelsea is a sushi restaurant in Chelsea with 4.4 stars on Google

How expensive is it?

```
+1 make_reservation(curr_result)
```

Okay make me a reservation!
User gives reward?

Find me a good sushi restaurant in Chelsea

Find me a good sushi restaurant in Chelsea

How does the user know the right search happened?

restaurant_type <- sushi

location <- Chelsea

curr_result <- execute_search()

Sushi Seki Chelsea is a sushi restaurant in Chelsea with 4.4 stars on Google

How expensive is it?

get_value(cost, curr_result)

Entrees are around $30 each

Wizard-of-Oz

Learning from demonstrations: “wizard” pulls the levers and makes the dialogue system update its state and take actions

Full Dialogue Task

Find me a good sushi restaurant in Chelsea

Wizard can be a trained expert and know exactly what the dialogue systems is supposed to do

Learning from Static Traces

Wizard can be a trained expert and know exactly what the dialogue systems is supposed to do

Using either wizard-of-Oz or other annotations, can collect static traces and train from these

Bordes et al. (2017)
Full Dialogue Task

Find me a good sushi restaurant in Chelsea

```
restaurant_type <- sushi
location <- Chelsea
stars <- 4+
curr_result <- execute_search()
```

- User asked for a “good” restaurant — does that mean we should filter by star rating? What does “good” mean?
- Hard to change system behavior if training from static traces, especially if system capabilities or desired behavior change

Goal-oriented Dialogue

- Tons of industry interest!
- Startups:
  - ASAPP
  - Maluuba
  - Eloquent Labs
  - semanticmachines
  - VIV
- Big Companies: Apple Siri (VocalIQ), Google Allo, Amazon Alexa, Microsoft Cortana, Facebook M, Samsung Bixby, Tencent WeChat
- Lots of cool work that’s not public yet

Search/QA as Dialogue

- “Has Chris Pratt won an Oscar?” / “Has he won an Oscar”
**QA as Dialogue**
- Dialogue is a very natural way to find information from a search engine or a QA system
- Challenges:
  - QA is hard enough on its own
  - Users move the goalposts

---

**Search as Dialogue**
- Google can deal with misspellings, so more misspellings happen — Google has to do more!

---

**Dialogue Mission Creep**
- Fixed distribution (e.g., natural language sentences), error rate -> 0
- Error rate -> ???; “mission creep” from HCI element

---

**UI QuAC dataset: Question Answering in Context**
- Iyyer et al. (2017)

---

**QA as Dialogue**
- UW QuAC dataset: Question Answering in Context
- Choi et al. (2018)
Dialogue Mission Creep

High visibility — your product has to work really well!

Takeaways

- Some decent chatbots, but unclear how to make these more sophisticated than they are right now
- Task-oriented dialogue systems are growing in scope and complexity — really exciting systems on the way
- More and more problems are being formulated as dialogue — interesting applications but challenging to get working well