CS395T: Structured Models for NLP
Lecture 23: Dialogue

Greg Durrett
Extractive Summarization: Bigram Recall

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

- 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

  \[
  \text{Maximize: } \sum_i w_i c_i \\
  \text{Subject to: } \sum_j l_j s_j \leq L \\
  \quad s_j \text{Occ}_{ij} \leq c_i, \quad \forall i, j \\
  \quad \sum_j s_j \text{Occ}_{ij} \geq c_i, \quad \forall i
  \]

  "set \(c_i\) to 1 iff some sentence that contains it is included"

  sum of included sentences’ lengths can’t exceed \(L\)  

  Gillick and Favre (2009)
Compressive Summarization

Maximize: $\sum_i w_i c_i$

Subject to: $\sum_j l_j s_j \leq L$

$s_j \text{Occ}_{ij} \leq c_i, \forall i, j$

$\sum_j s_j \text{Occ}_{ij} \geq c_i \forall i$

Now $s_j$ variables are nodes or sets of nodes in the parse tree

New constraint: $s_2 \leq s_1$

"$s_1$ is a prerequisite for $s_2$"

At least six people were killed and many others injured
Seq2seq Summarization

Original Text (truncated): Lagos, Nigeria (CNN) a day after winning Nigeria’s presidency, Muhammadu Buhari told CNN’s Christiane Amanpour that he plans to aggressively fight corruption that has long plagued Nigeria and go after the root of the nation’s unrest. Buhari said he’ll “rapidly give attention” to curbing violence in the northeast part of Nigeria, where the terrorist group Boko Haram operates. By cooperating with neighboring nations Chad, Cameroon and Niger, he said his administration is confident it will be able to thwart criminals and others contributing to Nigeria’s instability. For the first time in Nigeria’s history, the opposition defeated the ruling party in democratic elections. Buhari defeated incumbent Goodluck Jonathan by about 2 million votes, according to Nigeria’s independent national electoral commission. The win comes after a long history of military rule, coups and botched attempts at democracy in Africa’s most populous nation.

Baseline Seq2Seq + Attention: UNK UNK says his administration is confident it will be able to destabilize Nigeria’s economy. UNK says his administration is confident it will be able to thwart criminals and other Nigerians. He says the country has long Nigeria and Nigeria’s economy.

Problems: unknown words, inaccuracies

See et al. (2017)
Seq2seq Summarization

- **Solutions:** copy mechanism, coverage, just like in MT...

  Baseline Seq2Seq + Attention: **UNK UNK** says his administration is confident it will be able to **destabilize Nigeria’s economy.** **UNK** says his administration is confident it will be able to thwart criminals and other **nigerians.** *He says the country has long Nigeria and Nigeria’s economy.*

  **Pointer-Gen:** *Muhammadu Buhari* says he plans to aggressively fight corruption **in the northeast part of Nigeria.** He says he’ll “rapidly give attention” to curbing violence **in the northeast part of Nigeria.** He says his administration is confident it will be able to thwart criminals.

  **Pointer-Gen + Coverage:** *Muhammadu Buhari* says he plans to aggressively fight corruption that has long plagued Nigeria. He says his administration is confident it will be able to thwart criminals. The win comes after a long history of military rule, coups and botched attempts at democracy in Africa’s most populous nation.

- 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

Original Interpretation:

C  trained judge

A  B  B

Standard Interpretation:

C  trained judge

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

- Created 1964-1966 at MIT, heavily scripted
- DOCTOR script was most successful: repeats user’s input, asks inane questions

You are like my father in some ways.
WHAT RESEMBLANCE DO YOU SEE
You are not very aggressive but I think you don’t want me to notice that.
WHAT MAKES YOU THINK I AM NOT VERY AGGRESSIVE
You don’t argue with me.
WHY DO YOU THINK I DON’T ARGUE WITH YOU
You are afraid of me.
DOES IT PLEASE YOU TO BELIEVE I AM AFRAID OF YOU
My father is afraid of everybody.
WHAT ELSE COMES TO MIND WHEN YOU THINK OF YOUR FATHER
Bullies.
DOES THAT HAVE ANYTHING TO DO WITH THE FACT THAT YOUR BOYFRIEND MADE YOU COME HERE

Weizenbaum (1966)
ELIZA

- Identify keyword, identify context, apply transformation rule

\[ (.*) \text{ you } (.* \text{ me)} \]

Why do you think I $2 you?  

\[ \text{My } (.) (.* \text{) } \]

What else comes to mind when you think of your $1?

\[ \text{You don’t argue with me.} \]
\[ \text{WHY DO YOU THINK I DON’T ARGUE WITH YOU} \]

- Very little need to generate new content, but can only have one type of conversation
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
- Can often give sensible answers, but the bot doesn’t really impose high-level discourse structure

User: Hi there cleverbot.
Cleverbot: How are you?
User: Doing okay. what plans do you have today?
Cleverbot: I do revision.
User: What are you revising for?
Cleverbot: Maths history.
User: How old are you?
Cleverbot: That's none of your business.
User: What is maths history?
Cleverbot: I don’t know, farming maybe
Data-Driven Approaches

- Can treat as a machine translation problem: “translate” from current utterance to next one

![Phrase extraction diagram]

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

Ritter et al. (2011)
## Data-Driven Approaches

<table>
<thead>
<tr>
<th>Status</th>
<th>Mt-Chat</th>
<th>Mt-Baseline</th>
<th>Human</th>
</tr>
</thead>
<tbody>
<tr>
<td>riding to the arc on my bike. Hopefully I won’t melt into the pavement.</td>
<td>good luck on your bike ride arc</td>
<td>i can’t wait for it! i’m arc is a year.</td>
<td>make sure to ride in the middle of the street</td>
</tr>
<tr>
<td>I really hate watching movies in the living room.</td>
<td>me too, what movies are you doing in room?</td>
<td>me too! i love those? ) in my mouth.</td>
<td>Why? no comfy spots or just too open to public</td>
</tr>
<tr>
<td>Hello Did I miss David Getting his teen choice award? Or did they not show it :-(</td>
<td>hello miss? * he was getting his will probably win .</td>
<td>it’s just like what you just did the exact date hello ?</td>
<td>nothing yet...he presented though he is so darn cute</td>
</tr>
</tbody>
</table>
Seq2seq models

- Just like conventional MT, can train seq2seq models for this task
- Why might this model perform poorly? What might it be bad at?

<table>
<thead>
<tr>
<th>System</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>RANDOM</td>
<td>0.33</td>
</tr>
<tr>
<td>MT</td>
<td>3.21</td>
</tr>
<tr>
<td>HUMAN</td>
<td>6.08</td>
</tr>
</tbody>
</table>
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>
<tr>
<td>-1.09  Get out of here.</td>
</tr>
<tr>
<td>-1.09  I’m going home.</td>
</tr>
<tr>
<td>-1.09  Oh my god!</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>
<tr>
<td>-1.55  My name is Robert.</td>
</tr>
<tr>
<td>-1.58  My name is John.</td>
</tr>
<tr>
<td>-1.59  My name’s John.</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>
<tr>
<td>-1.64  Twenty-five.</td>
</tr>
<tr>
<td>-1.66  Five.</td>
</tr>
<tr>
<td>-1.71  Eight.</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)
## Lack of Diversity

| message                                      | SEQ2SEQ                                    | MMI $P(T|S) - \lambda P(T)$                  |
|----------------------------------------------|--------------------------------------------|---------------------------------------------|
| I did not get the report from an mi-6 agent | I don’t think that’s a good idea.          | You did the right thing, did you?          |
| You haven’t been given an assignment in this case | I don’t know what you are talking about.    | I’ve been looking all over for you.         |
| I’m losing my grip.                          | I don’t know what you are talking about.    | I’m the only one in the world.              |
| I am ready to help.                          | Come on, come on                            | I have something we need to talk about.     |
| You programmed me to gather intelligence.    | You know that, don’t you?                  | You do have fun, don’t you?                 |
| That’s all I’ve ever done.                    |                                            |                                             |
| I mean, we’d have to talk to him.            | I mean, I don’t know.                       | I mean, he’s a good guy.                    |
| This has nothing to do with him.             | You don’t know anything about him, do you?  | I know him as much.                         |
| How come you never say it?                   | I don’t know                               | Because I don’t want to hurt you            |

- OpenSubtitles data

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

- Xiaolce: Microsoft chatbot in Chinese, 20M users, average user interacts 60 times/month

- People do seem to like talking to them...?

<table>
<thead>
<tr>
<th>message</th>
<th>response</th>
</tr>
</thead>
<tbody>
<tr>
<td>In which city do you live now?</td>
<td>I live in Madrid.</td>
</tr>
<tr>
<td>In which country do you live now?</td>
<td>England, you?</td>
</tr>
</tbody>
</table>
Task-Oriented Dialogue
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
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

<table>
<thead>
<tr>
<th>Utterance</th>
<th>How much is the cheapest flight from Boston to New York tomorrow morning?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Goal:</td>
<td>Airfare</td>
</tr>
<tr>
<td>Cost_Relative</td>
<td>cheapest</td>
</tr>
<tr>
<td>Depart_City</td>
<td>Boston</td>
</tr>
<tr>
<td>Arrival_City</td>
<td>New York</td>
</tr>
<tr>
<td>Depart_Date_Relative</td>
<td>tomorrow</td>
</tr>
<tr>
<td>Depart_Time_Period</td>
<td>morning</td>
</tr>
</tbody>
</table>

- 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 -> update dialogue state -> take action (e.g., query the restaurant database) -> say something

- Much more complex than chatbots!

Young et al. (2013)
Find me a good sushi restaurant in Chelsea

```r
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?

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

Entrees are around $30 each
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)
Reward for completing task?

Find me a good sushi restaurant in Chelsea

Very indirect signal of what should happen up here

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?
...

Okay make me a reservation!

+1 make_reservation(curr_result)
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 does the user know the right search happened?

+1

How expensive is it?

get_value(cost, curr_result)

+1

Entrees are around $30 each
Learning from demonstrations: “wizard” pulls the levers and makes the dialogue system update its state and take actions.

Kelley (early 1980s), Ford and Smith (1982)
Full Dialogue Task

Find me a good sushi restaurant in Chelsea

wizard enters these

\[
\begin{align*}
\text{restaurant\_type} & \leftarrow \text{sushi} \\
\text{location} & \leftarrow \text{Chelsea} \\
\text{curr\_result} & \leftarrow \text{execute\_search}()
\end{align*}
\]

wizard types this out or invokes templates

\[
\begin{align*}
\text{Sushi Seki Chelsea is a sushi restaurant in Chelsea with 4.4 stars on Google}
\end{align*}
\]

- Wizard can be a trained expert and know exactly what the dialogue systems is supposed to do
Learning from Static Traces

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:
  - Eloquent Labs
  - semanticmachines
  - Maluuba
  - ASAPP
  - 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
Other Dialogue Applications
“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.

### Challenges:

1. Who are all of the super heroes?
2. Which of them come from Earth?
3. Of those, who appeared most recently?

#### Legion of Super Heroes Post-Infinite Crisis

<table>
<thead>
<tr>
<th>Character</th>
<th>First Appeared</th>
<th>Home World</th>
<th>Powers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Night Girl</td>
<td>2007</td>
<td>Kathoon</td>
<td>Super strength</td>
</tr>
<tr>
<td>Dragonwing</td>
<td>2010</td>
<td>Earth</td>
<td>Fire breath</td>
</tr>
<tr>
<td>Gates</td>
<td>2009</td>
<td>Vyrga</td>
<td>Teleporting</td>
</tr>
<tr>
<td>XS</td>
<td>2009</td>
<td>Aarok</td>
<td>Super speed</td>
</tr>
<tr>
<td>Harmonia</td>
<td>2011</td>
<td>Earth</td>
<td>Elemental</td>
</tr>
</tbody>
</table>

Iyyer et al. (2017)
Google can deal with misspellings, so more misspellings happen — Google has to do more!
Dialog Mission Creep

Most NLP tasks
- Error analysis
- Better model
- Data
- System

Dialogue/Search/QA
- Error analysis
- Better model
- Data
- Harder Data

- Fixed distribution (e.g., natural language sentences), error rate $\rightarrow 0$
- Error rate $\rightarrow ???$; “mission creep” from HCI element
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