Retrieval-Augmented Generation of Long Form Text

Claire Gardent
Joint work with Angela Fan (Facebook), Antoine Bordes (Facebook) and Chloé Braud (CNRS/IRIT)
Des responsables américains ont tenu une réunion d'un groupe d'experts en janvier 2002 à New York.

Funcionarios estadounidenses celebraron una reunión de un grupo de expertos en enero de 2002 en Nueva York.

Americkí predstavitelia usporiadali stretnutie expertnej skupiny v januári 2002 v New Yorku.

Американските служители проведоха среща на експертна група през януари 2002 г. в Ню Йорк.

Amerikanska tjänstemän holl ett expertgruppsmöte i januari 2002 i New York.

......
Applications

What is NLG useful for?

- Verbalising, Querying Knowledge-Bases
- Summarising, Simplifying, Paraphrases Text
- Converting Graphs into Text
Neural NLG

INPUT

NLG MODEL
Encoder-Decoder
Network

OUTPUT
Generated Text
Retrieval-Augmented Neural NLG

INPUT (QUERY) → NLG Model → Generated Text

+ RETRIEVAL CORPUS
  Web
  Dialog Corpus
  Image Corpus

Retrieval Model

RETRIEVED TEXT
Challenges for Retrieval-Augmented NLG

Scaling to very large retrieval corpora
Challenges for Retrieval-Augmented NLG

Scaling to very large retrieval corpora

Retrieving relevant knowledge
Challenges for Retrieval-Augmented NLG

Scaling to very large retrieval corpora
- Retrieving relevant knowledge
- Encoding long form input
Challenges for Retrieval-Augmented NLG

- Scaling to very large retrieval corpora
- Retrieving relevant knowledge
- Encoding long form input
- Decoding (generating) long form text
Three NLG Tasks
Retrieval-Based Models for three NLG Tasks

Long Form Question Answering

Human-Machine Dialog

Generating Wikipedia Biographies
Question Answering

Scaling to long input

INPUT
ELI5 Question

RETRIEVAL CORPUS

Retrieval Model

NLG Model

RETRIEVED TEXT
Web Documents
200K Tokens

Long form Answer
Human-Machine Dialog

Retrieving from multiple, multimodal retrieval sources
Scaling to very large retrieval corpora
Generating Wikipedia Woman Biographies
Generating structured text, Impact of available evidence (Gender bias)
Retrieval-Augmented Question Answering
Question Answering

Explain Like I'm Five Dataset

100 WEB DOCUMENTS

200,000 TRAINING INSTANCES

QUESTION

ANSWER

200,000 words
Creating a Shorter Support Document

200,000 words

850 words
TF-IDF Method

CALCULATE TF-IDF OVERLAP

QUESTION

WEB DOCUMENT SENTENCES
Downsides

38% of the Answer Tokens are Missing

Selected text fragments are often redundant (same tf-idf)
Convert Input Texts to Graph

Fan et al. EMNLP 2019

WEB DOCUMENTS → compression → linearization

QUESTION → Generation → ANSWER

200,000 words → 10,000 words
Converting a Text to a Graph

WEB DOCUMENTS

WEB DOCUMENT SENTENCES

open information extraction

coreference Resolution

Tf-idf filtering

relation

subject

object

Merge nodes
Increment Nodes Weight
Filter Irrelevant Input
Merge similar nodes and edges
Can someone explain the theory of relativity?

Albert Einstein, a German theoretical physicist, published the theory of relativity.
Can someone explain the theory of relativity?

Albert Einstein, a German theoretical physicist, published the theory of relativity.

The theory of relativity is one of the two pillars of modern physics.
Can someone explain the theory of relativity?

Albert Einstein, a German theoretical physicist, published the theory of relativity. The theory of relativity is one of the two pillars of modern physics.

He won the physics Nobel Prize.
Can someone explain the theory of relativity?

Albert Einstein, a German theoretical physicist, published the theory of relativity. The theory of relativity is one of the two pillars of modern physics. He won the physics Nobel Prize.

Puppies are very cute.
Low tf-idf with the query.

*Not added*
Text-to-Graph Conversion

Compresses the input by
- Dropping words
- Filtering out irrelevant triples

Reduces redundancy
- Merging nodes and edges

Filters out irrelevant content
- Tf-idf overlap (Question, Triple)
Knowledge Graph Construction drastically reduces the input size

The full text of the 100 web search results, which is around 200K tokens, is compressed to a few hundred tokens in the knowledge graph representation.
Does the graph preserve relevant information?

TF-IDF extraction is missing 38% of the answer tokens.
Does the graph preserve relevant information?

The graph extracted for 850 tokens is missing 35% of the answer tokens.
Does the graph preserve relevant information?

The graph for the full Input is missing only 8.7% of the answer tokens.
Model
Question-Answer Model
Generating an Answer from Web Retrieval

WEB DOCUMENTS

200,000 words

compression

linearization

QUESTION

Generation

10,000 words

ANSWER
Graph Linearisation

Encoding Graph Structure in a Seq2Seq Model

| WORD EMBEDDING | <sub> Albert Einstein <obj> the theory of relativity <pred> published <s> developed <obj> the Physics Nobel Prize <s> won |
| POSITION EMBEDDING | 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 |
Graph Linearisation
Encoding Graph Structure in a Seq2Seq Model

<table>
<thead>
<tr>
<th>WORD EMBEDDING</th>
<th>Albert Einstein</th>
<th>the theory of relativity</th>
<th>published</th>
<th>developed</th>
<th>the Physics Nobel Prize</th>
</tr>
</thead>
<tbody>
<tr>
<td>POSITION EMBEDDING</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>GRAPH WEIGHT EMBEDDING</td>
<td>0</td>
<td>4</td>
<td>4</td>
<td>0</td>
<td>2</td>
</tr>
</tbody>
</table>
Graph Linearisation

Encoding Graph Structure in a Seq2Seq Model

| WORD EMBEDDING | Albert Einstein | the theory of relativity | published | developed | the Physics Nobel Prize | won |
| POSITION EMBEDDING | 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 |
| GRAPH WEIGHT EMBEDDING | 0 4 4 0 2 2 2 0 1 0 1 0 3 3 3 3 0 2 |
| QUERY RELEVANCE EMBEDDING | 0 1 1 0 1 1 1 1 0 1 0 2 0 1 1 1 0 1 |
Multi-task Learning

ENCODER DECODER

LANGUAGE MODEL
Multi-task Learning
Multi-task Learning

SEQUENCE TO SEQUENCE

LANGUAGE MODELING

MASKED LANGUAGE MODELING

masked words
Multi-task Learning

SEQUENCE TO SEQUENCE

LANGUAGE MODELING

MASKED LANGUAGE MODELING

subject + object → masked words

subject + ?? → relation

subject + ?? → relation

object
Handling Long Input
Encoding and decoding 10K tokens

**Encoder**
Memory Compressed Attention

**Decoder**
Top-K attention
Evaluation
Automatic Evaluation

ROUGE

<table>
<thead>
<tr>
<th>Category</th>
<th>Extractive</th>
<th>LM</th>
<th>Seq2Seq</th>
<th>Multitask</th>
<th>Grph-Seq2Seq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Score</td>
<td>20.6</td>
<td>27.8</td>
<td>28.3</td>
<td>28.9</td>
<td>30.1</td>
</tr>
</tbody>
</table>
The evaluators preferred the graph based approach 58.4% of the time.
Question: Why is touching microfiber towels such an uncomfortable feeling?

True Answer: Do you mean the kind of cloths used to clean glasses and lenses? I’ve never noticed any uncomfortable feeling myself, but I do find touching certain cleaning cloths can be quite uncomfortable. There’s a brand called “e - cloth” which market themselves as not needing any cleaning supplies. Basically the threads have tiny little hooks and barbs in them, so dragging them over skin makes you feel like it’s catching against you.

Generation: The microfiber is made up of a bunch of tiny fibers that are attached to them. When you touch them, the fibers that make up the microfiber are attracted to each other. When they touch, they’re attracted to the other end of the fiber, which is what makes them feel uncomfortable.
Question: Why do we have nightmares when stressed out?

True Answer: I can tell you that your mind, and in particular your dreams, are subject to the input that it the journey of your life, as you have interpreted it. Yours dreams tend to be your subconscious representing issues you have within, whether it be past or present, with things and people that are sort of “preinterpreted symbology” from your deep and ever power subconsciousness. All your nightmares are just dreams, and your inner mind knows exactly can mess with you the most, and sometimes it brings that to light, it usually being an issue you may be stressed about.

Generation: When you’re stressed, your brain releases a hormone called cortisol. This hormone is released by your adrenal glands, and is released when you are stressed. This is why it’s called “fight-or-flight” mode. When your body is stressed, it releases hormones that make you feel more stressed. The reason why you don’t feel stressed is because your brain is trying to get rid of the stress hormone, which causes you to feel stressed, and you’re less likely to get stressed. If you’re really stressed, you’re probably going to be more stressed than if you weren’t stressed.

Repetitions (“when you are stressed”)
Non sense fragments (“The reason ...get stressed”)
Title: wylie rogers

True Paragraph: wylie rogers (born march 16, 1985) is an american ice hockey goaltender. he played with the rio grande valley killer bees of the central hockey league. prior to turning professional, rogers attended the university of alaska fairbanks where he played four seasons of ncaa men’s division i ice hockey with the alaska nanooks men’s ice hockey team. rogers played the 2010-11 season in the netherlands with the heerenveen flyers.

Generation: wylie rogers is an american football coach and former player. he is currently the head football coach at the university of north carolina at chapel hill. rogers played college football at north carolina state university. he was a member of the north carolina tar heels men’s football team that won the ncaa division i football championship in 2012. he also played for the university’s men’s basketball team. after his playing career ended, rogers was hired as the head coach at unc chapel hill in 2013. he led the tar heels to their first ncaa division ii championship in 2015.

Fluent text with nice, elaborate discourse structure

Incorrect facts (Wylie Rogers is an ice hockey player not a football coach. University of Alaska not North Carolina. etc;)}
Human-Machine Dialog
Retrieving and Integrating Knowledge from multiple sources
Human: I love watching Disney movies! Do you like Disney?

Model: I love Disney, I love watching Disney movies and different animations and characters that makes them unique.

Human: What’s your favorite Disney movie?

Model: I love the incredibles, they are my favorite Disney movie.

Human: omg, me too! The second one just came out. I don’t watch a lot of sequels though, I think the last one I saw was Toy Story.

Model: That’s cool. I remember when Toy Story was first released in 1995.
K-Nearest Neighbour Retrieval

Two retrieval sources

**Dialog**

**Human:** What’s your favorite Disney movie?

**Model:** I love the incredibles, they are my favorite Disney movie.

**Knowledge**

Disney announced intentions to develop additional superhero films after the success by the incredibles

**Template**

I love kiteboarding, it is one of my favorite activities on the water.
Wizard of Wikipedia
Dialog about a topic

Retrieval Corpus for KL
- WKP passages
- 34 per topic

Retrieval Corpus for Template
- Dialog turns
- 170K dialog turns

Image Chat
Dialog about an image

Retrieval Corpus for KL
- Image + dialog
- 184K images

Retrieval Corpus for Template
- Dialog turns
- 350K dialog turns
Retrieval-Based Human-Machine Dialog

K-Nearest Neighbour Search
### Effect of Fetched Text on Generation

#### Keeping the template fixed

<table>
<thead>
<tr>
<th>Knowledge</th>
<th>Training Utterance</th>
<th>Generation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>buzz lightyear</strong>'s name is in honor of astronaut edwin ‘buzz’ aldrin</td>
<td>my favorite character in that book series is hermione granger</td>
<td>cool! my favorite character in that movie is <strong>buzz lightyear</strong></td>
</tr>
<tr>
<td><strong>mr potato head</strong> is based on the real-life mr. potato head toy</td>
<td>my favorite character in that book series is hermione granger</td>
<td>my favorite character in that movie is real-life <strong>mr potato head</strong></td>
</tr>
<tr>
<td><strong>slinky dog</strong> is a toy dachshund with a metal slinky for a body</td>
<td>my favorite character in that book series is hermione granger</td>
<td>cool! my favorite character is the <strong>slinky dog</strong></td>
</tr>
<tr>
<td>slinky dog is a toy dachshund with a metal slinky for a body</td>
<td><strong>i really like</strong> the character hermione granger</td>
<td>cool! <strong>i really like</strong> slinky dog</td>
</tr>
<tr>
<td>slinky dog is a toy dachshund with a metal slinky for a body</td>
<td><strong>my favorite character</strong> of all time has to be hermione granger</td>
<td>i love that movie, <strong>my favorite character has to be</strong> slinky dog the dachshund</td>
</tr>
<tr>
<td>slinky dog is a toy dachshund with a metal slinky for a body</td>
<td><strong>i agree with you!</strong> that's <strong>my favorite</strong> character as well</td>
<td>i think so too! <strong>my favorite</strong> is slinky</td>
</tr>
</tbody>
</table>

#### Keeping the KL fixed
Human Evaluation

Wizard of Wikipedia, Unseen Set

- Preference over Retrieval
- Preference over Generative MemNet

% KNN Model Preference

- Prefer in Long Conversation
- More Interesting
- More Human
- More Coherent
- More Knowledge

50%
Generating Wikipedia Biographies
Studying the impact of Gender Bias on Retrieval-Augmented NLG
Generating Wikipedia Biographies from Web Retrieval

PERSON NAME
Challenges

Gather relevant evidence (Retrieval)

Generate a structured text

Ensure factuality
Generating Long Form Text
Fan and Gardent, ACL 2022

Dense retrieval on 1,000 tokens
(MIPS on Roberta Encodings)

Cache-based pre-trained encoder-decoder
to generate biographies section by section
Retrieval

**QUERY**
- Katherine Johnson
- Mathematician
- Early Life

**SEARCH OUTPUT**
Top 20 search results segmented into sentences

**OUTPUT**
40 sentences most similar with the query (1,000 words)
Katherine Johnson was born as Creola Katherine Coleman on August 26, 1918, in White Sulphur Springs, West Virginia, to Joylette Roberta (Lowe) and Joshua McKinley Coleman. She was the youngest of four children. Johnson showed strong math abilities from an early age.
Caches the previous section’s hidden states at every layer

Usd as a memory to generate the current section
### Ablation

<table>
<thead>
<tr>
<th>Model</th>
<th>ROUGE-L</th>
<th>Entailment</th>
<th>Named Entity Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>BART Pretraining + Finetuning</td>
<td>17.4</td>
<td>15.8</td>
<td>21.9</td>
</tr>
<tr>
<td>+ Retrieval Module</td>
<td>18.8</td>
<td>17.2</td>
<td>23.1</td>
</tr>
<tr>
<td>+ Caching Mechanism</td>
<td>19.3</td>
<td>17.9</td>
<td>23.4</td>
</tr>
</tbody>
</table>

The retrieval and the cache module statistically significantly improve results.
Human Evaluation of Factuality

- **WEB EVIDENCE**
  - 17% of generated info not present in reference is present in the citation

- **GENERATED**
  - 68% information not in reference

- **REFERENCE**
  - 71% information not in generated
  - 32% of generated info is present in the reference
# The Evidence Gap

Wikipedia Biographies
And
Web Documents

## Wikisum Test Set
Men and women

## Our Test Set
Only women

<table>
<thead>
<tr>
<th>WikiSum Evaluation Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average Number of Sections</td>
</tr>
<tr>
<td>Average Length of a Section</td>
</tr>
<tr>
<td>Average Length of Total Article</td>
</tr>
<tr>
<td>Avg overlap of Web Hits and Biography</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Our Evaluation Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average Number of Sections</td>
</tr>
<tr>
<td>Average Length of a Section</td>
</tr>
<tr>
<td>Average Length of Total Article</td>
</tr>
<tr>
<td>Avg Number of Web Hits (max 20)</td>
</tr>
<tr>
<td>Avg overlap of Web Hits and Biography</td>
</tr>
</tbody>
</table>
Less Web Evidence, Less Good Texts

<table>
<thead>
<tr>
<th>Model</th>
<th>WikiSum Test</th>
<th>Women</th>
<th>Scientists</th>
<th>Women in Asia</th>
<th>Women in Africa</th>
</tr>
</thead>
<tbody>
<tr>
<td>BART Pretraining</td>
<td>19.0</td>
<td>17.4</td>
<td>18.2</td>
<td>16.7</td>
<td>16.4</td>
</tr>
<tr>
<td>+ Retrieval</td>
<td>21.4</td>
<td>18.8</td>
<td>19.3</td>
<td>17.9</td>
<td>17.1</td>
</tr>
<tr>
<td>+ Caching</td>
<td>21.8</td>
<td>19.3</td>
<td>19.7</td>
<td>18.4</td>
<td>17.3</td>
</tr>
</tbody>
</table>
Conclusions
Question Answering

Challenge

• Scaling to very long input

Method

• Web Documents ➔ Graph
• Memory Compressed Attention
• Top-K attention
Human-Machine Dialog

Challenge

• Efficient retrieval on very large retrieval corpora
• Handling and combining multiple retrieval sources

Method

• K-Nearest Neighbour Search
• Multiple Encoders
• Gates
Generating Wikipedia Biographies

Challenge

• Retrieving sufficient information
• Generating Long-Form Structured Text

Method

• Dense Retrieval
• Cache
Open Challenges

*Factuality*
Evaluation and model improvement

*Multilingual NLG*
Generating into languages other than English

*Multi-modal NLG*
Generating from multiple input types
Thank You!