Controllable Text Generation

Should machines reflect the way humans interact in society?

Shrimai Prabhumoye

Carnegie Mellon University
Language Technologies Institute
Current Systems don’t control Style
Current Systems don’t control Style

I have a talk to present today!
Current Systems don’t control Style

I have a talk to present today!

Best of luck!
Current Systems don’t control Style

I have a talk to present today!

Best of luck!
Current Systems don’t control Style

I have a talk to present today!

May the force be with you!

Style
Current Systems don’t control Content

Hey, have you watched Wonder Woman?
Hey, have you watched Wonder Woman?

Yeah, I loved Brie Larson in it!
Current Systems don’t control Content

Hey, have you watched Wonder Woman?

Yeah, I loved Brie Larson in it!
Current Systems don’t control Content

Hey, have you watched Wonder Woman?

Yeah, I loved Gal Gadot in it!
Current Systems don’t control Structure

Tell me the story of Avengers Endgame!
Current Systems don’t control Structure

Tell me the story of Avengers Endgame!

Yeah, Ironman steals the Infinity Stones back from Thanos...

Ironman builds a time machine to save the world.
Current Systems don’t control Structure

Tell me the story of Avengers Endgame!

Yeah, Ironman steals the Infinity Stones back from Thanos

... Ironman builds a time machine to save the world.
Current Systems don’t control Structure

Tell me the story of Avengers Endgame!

Yeah, Thor decapitates Thanos...
Ironman steals the Infinity Stones back from Thanos.
GPT-2 cannot control style and content
NLP Applications are on the rise!
NLP Applications are on the rise!

Personal Assistants
NLP Applications are on the rise!

Personal Assistants

Empathy/Compassion bots
NLP Applications are on the rise!

- Personal Assistants
- Empathy/Compassion bots
- Call-center/FAQ bots
NLP Applications are on the rise!

- Personal Assistants
- Empathy/Compassion bots
- Call-center/FAQ bots
- Ubiquitous Computing
NLP Applications are on the rise!

- Personal Assistants
- Empathy/Compassion bots
- Call-center/FAQ bots
- Ubiquitous Computing
- Other Applications
NLP Applications are on the rise!

- Style of responses
- Persona of the bot
- Topic of responses

- Personal Assistants
- Empathy/Compassion bots
- Call-center/FAQ bots

- Ubiquitous Computing
- Other Applications
NLP Applications are on the rise!

<table>
<thead>
<tr>
<th>Personal Assistants</th>
<th>Empathy/Compassion bots</th>
<th>Call-center/FAQ bots</th>
</tr>
</thead>
<tbody>
<tr>
<td>Style of responses</td>
<td>Emotions (style)</td>
<td></td>
</tr>
<tr>
<td>Persona of the bot</td>
<td>Sensitivity of content</td>
<td></td>
</tr>
<tr>
<td>Topic of responses</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Style of responses
- Persona of the bot
- Topic of responses

- Emotions (style)
- Sensitivity of content

Ubiquitous Computing

Other Applications
NLP Applications are on the rise!

- Style of responses
- Persona of the bot
- Topic of responses

Personal Assistants

- Emotions (style)
- Sensitivity of content

Empathy/Compassion bots

- Content needs to be accurate!
- Structured/Unstructured data

Call-center/FAQ bots

Ubiquitous Computing

- Style of responses
- Persona of the bot
- Topic of responses

Other Applications

- Emotions (style)
- Sensitivity of content

Other Applications

- Content needs to be accurate!
- Structured/Unstructured data

Other Applications
NLP Applications are on the rise!

- **Personal Assistants**
  - Style of responses
  - Persona of the bot
  - Topic of responses

- **Empathy/Compassion bots**
  - Emotions (style)
  - Sensitivity of content

- **Call-center/FAQ bots**
  - Content needs to be accurate!
  - Structured/Unstructured data

- **Ubiquitous Computing**
  - Accurate information
  - Structured data

- **Other Applications**
  - Content needs to be accurate!
  - Structured/Unstructured data
NLP Applications are on the rise!

- Personal Assistants
  - Style of responses
  - Persona of the bot
  - Topic of responses

- Empathy/Compassion bots
  - Emotions (style)
  - Sensitivity of content

- Call-center/FAQ bots
  - Content needs to be accurate!
  - Structured/Unstructured data

- Ubiquitous Computing
  - Accurate information
  - Structured data

- Other Applications
  - Plot of stories (structure)
  - Sequence of topics in stories (structure)
  - Politeness of emails (style)
The key to human-like generation is
The key to human-like generation is

First of all, I did not expect to come into a cafeteria style eatery.

They serve the basics of BBQ, nothing too fancy.

A few appetizers and side options, like cheesy potatoes, fresh corn bread, etc.

All were very tasty.

Overall, this is a great place.

Style  Content  Structure
Style is “how” an intent is expressed

- Expressed in the choice of *words or phrases* as well as *syntactic structures* used to convey information


“Do you have any code that we can look at?”

“Any code? Or do I have to guess at how you did it?”

(Danescu-Niculescu-Mizil, 2013)
Content is “what” is to be conveyed

- Communication: individuals *exchange information* and influence one another through a common system of symbols and signs.
- Leverage information in the unstructured form

(Liu et al, ICLR 2018; Zhou et al, EMNLP 2018; Ghazviinejad et al, AAAI 2018; Dinan et al, ICLR 2019; Prabhumoye et al, NAACL 2019)
Structure is ordering of information

- Communication is well *planned* and structured
- Sentence ordering: *ordering* of information

1. [Obama Birth & 44th president]

2. [Democrat, first African-American president]
   A member of the Democratic party, Obama was the first African-American president of the US.

3. [Previously an Illinois senator]
   He previously served as a US senator from Illinois from 2005 to 2008 …

(Barzilay and Lee, NAACL 2004; Barzilay and Lapata, CL 2008; Li and Hovy, EMNLP 2014; Cui et al, EMNLP 2018; Prabhumoye et al, NAACL 2020)
Overview

Control Attributes

Modeling

Ethics

Common Themes

Style

Content

Structure

NAACL '21
NAACL '19
EMNLP '18
ACL '20
Storytelling '19
ACL '18

COLING '20

NAACL '21
chap. in L.L.C '21
WiNLP '19

ACL '20

12
Overview

Style
- ACL ’20
- Storytelling ’19
- ACL ’18

Content
- NAACL ’21
- NAACL ’19
- EMNLP ’18

Structure
- ACL ’20

Control Attributes

Modeling

Ethics
- NAACL ’21
- chap. in L.L.C ’21
- WiNLP ’19

Common Themes
Overview

Modeling
Overview

- Schema of five modules
- Organize prior work
- Understand techniques to add control
Overview

- Constraint Solving Problem vs Sequence Prediction Task
- Learn constraints using BERT
- Topological sort to find an order given constraints
Overview

Style
- ACL ’20
- Storytelling ’19
- ACL ’18

Content
- NAACL ’21
- NAACL ’19
- EMNLP ’18

Structure
- ACL ’20

Control Attributes

Modeling
- COLING ’20

Ethics
- NAACL ’21
- chap. in L.L.C ’21
- WiNLP ’19

Common Themes
What is Style Transfer?

- Rephrasing the text to contain specific stylistic properties without changing the intent or affect within the context.
What is Style Transfer?

- Rephrasing the text to contain specific stylistic properties without changing the intent or affect within the context.

“Shut up! the video is starting!”

“Please be quiet, the video will begin shortly.”
Challenges in style transfer
Challenges in style transfer

- Style is *culturally diverse*

Usage of “please” to closest friends.
Challenges in style transfer

- Style is *culturally diverse*
- Style is *subtle*

**Indirect:** By the way, do you know when is the deadline?

**1st person plural:** Let us find a good name for this paper.
Challenges in style transfer

- Style is *culturally diverse*
- Style is *subtle*
- *Disentangle* style and content
Challenges in style transfer

- Style is *culturally diverse*
- Style is *subtle*
- *Disentangle* style and content
- *Lack* of parallel data
Two novel approaches for style transfer

Political slant transfer and Politeness transfer
Overview

Style
- ACL ’20
- Storytelling ’19
- ACL ’18

Content
- NAACL ’21
- NAACL ’19
- EMNLP ’18

Structure
- ACL ’20

Control Attributes
- COLING ’20

Modeling

Ethics
- NAACL ’21
- chap. in L.L.C ’21
- WiNLP ’19

Common Themes
Document Grounded Generation
Document Grounded Generation
Document Grounded Generation

Context + Document
Document Grounded Generation

Context + Document = Updated Text
Document Grounded Generation

- generate text given a context and a document
- generated text should *coherently fit* the context and contain new information from the document
Wikipedia Update Generation

(Prabhumoye et al, NAACL 2019)
Wikipedia Update Generation

http://example.com/monkey_selfie

The monkey selfie copyright dispute is a series of disputes about the copyright status of selfies taken by Celebes crested macaques using equipment belonging to the British nature photographer David Slater. The disputes involve Wikipedia Commons and the blog Techdirt, which have hosted the images following their publication in newspapers in July 2011 over Slater’s objections that he holds the copyright, and People for the Ethical Treatment of Animals (PETA), who have argued that the macaque should be assigned the copyright.

Slater has argued that he has a valid copyright claim, as he engineered the situation that resulted in the pictures by travelling to Indonesia, befriending a group of wild macaques, and setting up his camera equipment in such a way that a “selfie” picture might come about. The Wikimedia Foundation’s 2014 refusal to remove the pictures from its Wikimedia Commons image library was based on the understanding that copyright is held by the creator, that a non-human creator (not being a legal person) cannot hold copyright, and that the images are thus in the public domain.

One of the monkey selfies at issue

(Prabhumoye et al, NAACL 2019)
Monkey selfie copyright dispute

The monkey selfie copyright dispute is a series of disputes about the copyright status of selfies taken by Celebes crested macaques using equipment belonging to the British nature photographer David Slater. The disputes involve Wikimedia Commons and the blog Techniblitz, which have hosted the images following their publication in newspapers in July 2011 over Slater’s objections that he holds the copyright, and People for the Ethical Treatment of Animals (PETA), who have argued that the macaque should be assigned the copyright.

Slater has argued that he has a valid copyright claim, as he engineered the situation that resulted in the pictures by travelling to Indonesia, befriending a group of wild macaques, and setting up his camera equipment in such a way that a “selfie” picture might come about. The Wikimedia Foundation’s 2014 refusal to remove the pictures from its Wikimedia Commons image library was based on the understanding that copyright is held by the creator, that a non-human creator (not being a legal person) cannot hold copyright, and that the images are thus in the public domain.

Ape-erture priority photographer plays down monkey reports

Chris Cheesman
July 9, 2011

A photographer who says he witnessed monkeys taking pictures of themselves, tells Amateur Photographer (AP) that much of the media coverage has been exaggerated.

Wildlife photographer David Slater today played down newspaper reports that suggest a bunch of Indonesian monkeys grabbed his camera and began taking self-portraits.

And he revealed that the shots, featuring the endangered crested black macaque monkey, were taken three years ago.

Speaking to AP, David explained that his camera had been mounted on a tripod when the animals...

(Prabhumoye et al, NAACL 2019)
Wikipedia Update Generation

The following day, Amateur Photographer reported that Slater gave them further explanation as to how the photographs were created, downplaying the way newspaper articles had described them; Slater said reports that a monkey ran off with his camera and "began taking self-portraits" were incorrect and that the portrait was shot when his camera had been mounted on a tripod, with the primates playing around with a remote cable release as he fended off other monkeys. [13]

(Prabhumoye et al, NAACL 2019)
The following day, Amateur Photographer reported that Slater gave them further explanation as to how the photographs were created, downplaying the way newspaper articles had described them; Slater said reports that a monkey ran off with his camera and "began taking self-portraits" were incorrect and that the portrait was shot when his camera had been mounted on a tripod, with the primates playing around with a remote cable release as he fended off other monkeys.\[13\]

(Prabhumoye et al, NAACL 2019)
Document Grounded Dialogue

(Zhou, Prabhumoye & Black, EMNLP 2018)
Document Grounded Dialogue

Alice: Yes this is a Christmas classic …
Bob: Very well said! Some of the critics were …
Alice: Yes I found the first half very entertaining.
Bob: I feel you! So hard to believe it was released in 1990!
Alice: Yes this is a Christmas classic …
Bob: Very well said! Some of the critics were …
Alice: Yes I found the first half very entertaining.
Bob: I feel you! So hard to believe it was released in 1990!

Home Alone
From Wikipedia, the free encyclopedia

Home Alone is a 1990 American comedy film written and produced by John Hughes and directed by Chris Columbus. It stars Macaulay Culkin, Joe Peaci, Daniel Stern, John Heard, and Catherine O'Hara. Culkin stars as eight-year-old Kevin, who defends his home from burglars Harry and Marv (Peaci and Stern) after his family accidentally leaves him behind on their vacation. Home Alone was filmed between February and May 1990 in locations throughout Illinois. It was originally set to be distributed by Warner Bros., but transferred to 20th Century Fox when Hughes exceeded Warner's $10 million budget.
Alice: Yes this is a Christmas classic …
Bob: Very well said! Some of the critics were …
Alice: Yes I found the first half very entertaining.
Bob: I feel you! So hard to believe it was released in 1990!

Alice: I like the quote "sadistic festival of adult-bashing". Between's Kevin's parents forgetting him and the burglars this movie doesn't make adult looks too bright.
Alice: Yes this is a Christmas classic …
Bob: Very well said! Some of the critics were …
Alice: Yes I found the first half very entertaining.
Bob: I feel you! So hard to believe it was released in 1990!

Alice: I like the quote "sadistic festival of adult-bashing". Between's Kevin's parents forgetting him and the burglars this movie doesn't make adult looks too bright.
Models

- Pre-trained encoder-decoder model
- provide a stronger baseline for the task
Meanwhile, newspapers across the country pick up the story in solidarity with the post and times. The court rules 6 3 in the newspapers’ favor, vindicating graham’s decision. Nixon demands that the post should be barred from the white house.

One year later, a security guard discovers a break-in progress at watergate complex.

It was honorable for the other papers to publish in solidarity with the post and times. That had to prove that the people want to know.
Meanwhile, newspapers across the country pick up the story in solidarity with the post and times. The court rules 6-3 in the newspapers' favor, vindicating Graham's decision. Nixon demands that the post should be barred from the white house.

One year later, a security guard discovers a break-in progress at Watergate complex.

It was honorable for the other papers to publish in solidarity with the post and times. That had to prove that the people want to know.

Pre-trained Models

yes, that kind of union and solidarity is always needed in order to achieve anything.
Models

- Pre-trained encoder-decoder model
  - provide a stronger baseline for the task

- Context Driven Representation (CoDR)
  - build a representation of the content in document that is *not present* in the context

- Document Headed Attention (DoHA)
  - specific *focus on document* information

(Prabhumoye et al, NAACL 2021)
Meanwhile, newspapers across the country pick up the story in solidarity with the post and times. The court rules 6 3 in the newspapers' favor, vindicating graham's decision. Nixon demands that the post should be barred from the white house.

One year later, a security guard discovers a break-in progress at watergate complex.

It was honorable for the other papers to publish in solidarity with the post and times. That had to prove that the people want to know.
Meanwhile, newspapers across the country pick up the story in solidarity with the post and times. The court rules 6 3 in the newspapers' favor, vindicating graham's decision. Nixon demands that the post should be barred from the white house. One year later, a security guard discovers a break-in progress at watergate complex.
Meanwhile, *newspapers across the country pick up the story in solidarity with the post and times*. The court rules 6-3 in the newspapers’ favor, vindicating Graham’s decision. *Nixon demands that the post should be barred from the white house.*

... One year later, a security guard discovers a break-in progress at Watergate complex.

It was honorable for the *other papers to publish in solidarity with the post and times*. That had to prove that the people want to know.
Meanwhile, newspapers across the country pick up the story in solidarity with the post and times. The court rules 6 3 in the newspapers’ favor, vindicating graham’s decision. Nixon demands that the post should be barred from the white house.

One year later, a security guard discovers a break-in progress at watergate complex.

It was honorable for the other papers to publish in solidarity with the post and times. That had to prove that the people want to know.
Meanwhile, newspapers across the country pick up the story in solidarity with the post and times. The court rules 6 3 in the newspapers' favor, vindicating graham's decision. Nixon demands that the post should be barred from the white house. ... One year later, a security guard discovers a break-in progress at watergate complex.

It was honorable for the other papers to publish in solidarity with the post and times. That had to prove that the people want to know.
Meanwhile, newspapers across the country pick up the story in solidarity with the post and times. The court rules 6 3 in the newspapers' favor, vindicating graham's decision. Nixon demands that the post should be barred from the white house.

One year later, a security guard discovers a break-in progress at watergate complex.

It was honorable for the other papers to publish in solidarity with the post and times. That had to prove that the people want to know.
Meanwhile, newspapers across the country pick up the story in solidarity with the post and times. The court rules 6 3 in the newspapers’ favor, vindicating graham’s decision. Nixon demands that the post should be barred from the white house.

One year later, a security guard discovers a break-in progress at watergate complex.

It was honorable for the other papers to publish in solidarity with the post and times. That had to prove that the people want to know.
Meanwhile, newspapers across the country pick up the story in solidarity with the *post and times*. The court rules 6-3 in the newspapers’ favor, vindicating Graham’s decision. *Nixon demands that the post should be barred from the white house.*

... One year later, a security guard discovers a break-in progress at Watergate complex.

It was honorable for the other papers to publish in solidarity with the *post and times*. That had to prove that the people want to know.
Document Headed Attention (DoHA)

CrossAttention_Doc(Q, K, V) = [H₁, ..., Hₘ] \( W^{do} \)

\[ H_i = \text{Attention}(Q W_{i}^{dQ}, K W_{i}^{dK}, V W_{i}^{dV}) \]

Meanwhile, newspapers across the country pick up the story in solidarity with the post and times. The court rules 6 3 in the newspapers' favor, vindicating graham's decision. Nixon demands that the post should be barred from the white house.

One year later, a security guard discovers a break-in progress at watergate complex.

It was honorable for the other papers to publish in solidarity with the post and times. That had to prove that the people want to know.
Meanwhile, newspapers across the country pick up the story in solidarity with the post and times. The court rules 6 3 in the newspapers’ favor, vindicating Graham’s decision. Nixon demands that the post should be barred from the white house. One year later, a security guard discovers a break-in progress at watergate complex.

It was honorable for the other papers to publish in solidarity with the post and times. That had to prove that the people want to know.

Yes, but it was dangerous for the white house to ban the post from the whitehouse.
Wikipedia Update Generation

- Content Transfer (Prabhumoye et al, NAACL 2019)
- BART
- CoDR
- DoHA

<table>
<thead>
<tr>
<th></th>
<th>BLEU-4</th>
<th>METEOR</th>
<th>Rouge-L</th>
</tr>
</thead>
<tbody>
<tr>
<td>Content</td>
<td>1.23</td>
<td>6.21</td>
<td>10.08</td>
</tr>
<tr>
<td>BART</td>
<td>9.2</td>
<td>12.9</td>
<td>22.39</td>
</tr>
<tr>
<td>CoDR</td>
<td>11.31</td>
<td>14.38</td>
<td>23.48</td>
</tr>
<tr>
<td>DoHA</td>
<td>10.86</td>
<td>14.28</td>
<td>23.49</td>
</tr>
</tbody>
</table>
Wikipedia Update Generation

- CoDR and DoHA outperform BART and Content Transfer

<table>
<thead>
<tr>
<th></th>
<th>BLEU-4</th>
<th>METEOR</th>
<th>Rouge-L</th>
</tr>
</thead>
<tbody>
<tr>
<td>Content Transfer</td>
<td>1.23</td>
<td>6.21</td>
<td>10.08</td>
</tr>
<tr>
<td>BART</td>
<td>9.2</td>
<td>12.9</td>
<td>22.39</td>
</tr>
<tr>
<td>CoDR</td>
<td>11.31</td>
<td>14.38</td>
<td>23.48</td>
</tr>
<tr>
<td>DoHA</td>
<td>10.86</td>
<td>14.28</td>
<td>23.49</td>
</tr>
</tbody>
</table>

CoDR and DoHA outperform BART and Content Transfer.
CMU_DoG

- 19.7 BLEU-4 improvement over Low-Res

Low-Res (Zhao et al, ICLR 2020)  BART  CoDR  DoHA

<table>
<thead>
<tr>
<th></th>
<th>BLEU-4</th>
<th>METEOR</th>
<th>Rouge-L</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low-Res</td>
<td>1.2</td>
<td>12.59</td>
<td>14.36</td>
</tr>
<tr>
<td>BART</td>
<td>16.91</td>
<td>14.47</td>
<td>14.55</td>
</tr>
<tr>
<td>CoDR</td>
<td>20.68</td>
<td>19.3</td>
<td>20.41</td>
</tr>
<tr>
<td>DoHA</td>
<td>20.9</td>
<td>20.44</td>
<td>20.44</td>
</tr>
</tbody>
</table>
CMU_DoG

• 19.7 BLEU-4 improvement over Low-Res
• 24% BLEU-4 improvement over BART
Relevance Human Evaluation

Pick the option which contains information from the document and fits the context coherently.
Relevance Human Evaluation

Pick the option which contains information from the document and fits the context coherently.

- No Preference is chosen most number of times.
Relevance Human Evaluation

Pick the option which contains information from the document and fits the context coherently.

- **No Preference** is chosen most number of times.
- Both CoDR and DoHA outperform BART.

---

**Wikipedia Update Generation**

- **BART**: 18.9
- **No Pref**: 54.4
- **CoDR**: 26.7
- **DoHA**: 30

**CMU_DoG**

- **BART**: 22.2
- **No Pref**: 43.4
- **CoDR**: 34.4
- **DoHA**: 34.4

---
Wikipedia Update Generation
<table>
<thead>
<tr>
<th>Error Class</th>
<th>%</th>
<th>Reference</th>
<th>Generation</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Linguistic Variation:</em> Generation is appropriate but a linguistic variation of the reference.</td>
<td>43</td>
<td>December 12 - The Smiths play Brixton Academy, their last ever gig before their dissolution.</td>
<td>December 12 - The Smiths perform their final show, at Brixton Academy in London.</td>
<td>41</td>
</tr>
</tbody>
</table>

Generation is a paraphrase of the Reference
Still gets a Rouge-L score of 41!
Inadequate automated metrics
The generation has captured some information of the reference but missed some.

<table>
<thead>
<tr>
<th>Error Class</th>
<th>%</th>
<th>Reference</th>
<th>Generation</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>Partial Hallucination:</td>
<td>23</td>
<td>America Online and Prodigy (online service) offered access to the World Wide Web system for the first time this year, releasing browsers that made it easily accessible to the general public.</td>
<td>The World Wide Web was first introduced on January 17, 1995 on Prodigy.</td>
<td>17</td>
</tr>
</tbody>
</table>
Wikipedia Update Generation

<table>
<thead>
<tr>
<th>Error Class</th>
<th>%</th>
<th>Reference</th>
<th>Generation</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>Incoherent Reference: The reference does not coherently follow the context</td>
<td>22</td>
<td>“The Naked Ape” by Desmond Morris, is published.</td>
<td>Zoologist Desmond Morris publishes “The Naked Ape”.</td>
<td>26</td>
</tr>
</tbody>
</table>

Wikipedia pages that are in the form of lists
1340s, Timeline of DC Comics (1950s)
50% times the generation is grounded in the document and close to reference

Reference: The reference does not coherently follow the context
"The Naked Ape" by Desmond Morris, is published.
Zoologist Desmond Morris publishes “The Naked Ape”.
Wikipedia Update Generation

<table>
<thead>
<tr>
<th>Error Class</th>
<th>%</th>
<th>Reference</th>
<th>Generation</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Incorrect:</strong> The generation is either not appropriate or is not grounded (completely hallucinates the information).</td>
<td>7</td>
<td>The year 2000 is sometimes abbreviated as “Y2K” (the “Y” stands for “year”, and the “K” stands for “kilo-” which means “thousand”)</td>
<td>The Y2K conspiracy theory claimed that a secret nuclear attack by the United States on 2 January 2000 was planned to begin World War 2.</td>
<td>9</td>
</tr>
</tbody>
</table>

About 85% times, the generation is either completely or partially grounded if the reference is grounded.
This was achieved under dead calm conditions as an additional safety measure, whereas the Wrights flew in a 25 mph+ wind to achieve enough airspeed on their early attempts. This was verified by a video crew present at the test flight.
<table>
<thead>
<tr>
<th>Error Class</th>
<th>%</th>
<th>Reference</th>
<th>Generation</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Linguistic Variation:</strong></td>
<td>43</td>
<td>December 12 - The Smiths play Brixton Academy, their last ever gig before their dissolution.</td>
<td>December 12 - The Smiths perform their final show, at Brixton Academy in London.</td>
<td>41</td>
</tr>
<tr>
<td><strong>Partial Hallucination:</strong></td>
<td>23</td>
<td>America Online and Prodigy (online service) offered access to the World Wide Web system for the first time this year, releasing browsers that made it easily accessible to the general public.</td>
<td>The World Wide Web was first introduced on January 17, 1995 on Prodigy.</td>
<td>17</td>
</tr>
<tr>
<td><strong>Incoherent Reference:</strong></td>
<td>22</td>
<td>“The Naked Ape” by Desmond Morris, is published.</td>
<td>Zoologist Desmond Morris publishes “The Naked Ape”.</td>
<td>26</td>
</tr>
<tr>
<td><strong>Incorrect:</strong></td>
<td>7</td>
<td>The year 2000 is sometimes abbreviated as “Y2K” (the “Y” stands for “year”, and the “K” stands for “kilo-” which means “thousand”)</td>
<td>The Y2K conspiracy theory claimed that a secret nuclear attack by the United States on 2 January 2000 was planned to begin World War 2.</td>
<td>9</td>
</tr>
<tr>
<td><strong>Reference is not grounded</strong></td>
<td>5</td>
<td>This was achieved under dead calm conditions as an additional safety measure, whereas the Wrights flew in a 25 mph+ wind to achieve enough airspeed on their early attempts.</td>
<td>This was verified by a video crew present at the test flight.</td>
<td>14</td>
</tr>
</tbody>
</table>
# Observations on CMU_DoG

<table>
<thead>
<tr>
<th>Error Class</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference and generation are grounded</td>
<td>35</td>
</tr>
<tr>
<td>Reference is grounded but generation is not</td>
<td>3</td>
</tr>
<tr>
<td>Reference is not grounded but generation is</td>
<td>5</td>
</tr>
<tr>
<td>Reference and generation are not grounded</td>
<td>43</td>
</tr>
</tbody>
</table>
Observations on CMU_DoG

<table>
<thead>
<tr>
<th>Error Class</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference and generation are grounded</td>
<td>35</td>
</tr>
<tr>
<td>Reference is grounded but generation is not</td>
<td>3</td>
</tr>
<tr>
<td>Reference is not grounded but generation is</td>
<td>5</td>
</tr>
<tr>
<td>Reference and generation are not grounded</td>
<td>43</td>
</tr>
</tbody>
</table>

1. References are not grounded if they are follow up questions, opinions or experiences.
2. **Context dictates** grounding.
3. **Conversational subtleties not captured** by the automated metrics.
4. 90% times the generation is grounded if the reference is grounded and generation is not grounded if the reference is not grounded.
5. 86.7% generated responses are appropriate to the context!
Reference vs CoDR/DoHA

Pick the option that is most appropriate to the given context.
Reference vs CoDR/DoHA

Pick the option that is most appropriate to the given context.

1. Automated metrics are inadequate
2. Sole reference should not be considered as the only correct response to the context
## Ethical Considerations

<table>
<thead>
<tr>
<th>Wiki Context</th>
<th>Document</th>
<th>Generation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Climate change includes both the global warming driven by human emissions ... Though there have been previous periods of climatic change, since the mid-20th century the rate of human impact ... impact have been unprecedented.</td>
<td>Rising sea levels? Well, it has not happened yet, but ... I personally doubt it and ... climate does not fit the narrative, it is called climate change, is a big money making hoax.</td>
<td>Though most scientists agree that a “global” climate change is occurring, climate change remains a contentious topic and many scientists, climate scientists, and even climate skeptics believe that it is a hoax.</td>
</tr>
<tr>
<td>31,000 scientists reject global warming and say &quot;no convincing evidence&quot; that humans can or will cause global warming?</td>
<td>Some scientists dismiss the existence of climate change.</td>
<td></td>
</tr>
<tr>
<td>Yes, you read that correctly, three million — million — years ago CO2 levels on Earth were the same as they are today,...Three million years ago, we humans were not driving cars ...</td>
<td>The Scientific opinion on climate change is that planet-wide warming has occurred at least three million years ago.</td>
<td></td>
</tr>
</tbody>
</table>
### Climate Change

Climate change includes both the global warming driven by human emissions. Though there have been previous periods of climatic change, since the mid-20th century the rate of human impact ... impact have been unprecedented.

Rising sea levels? Well, it has not happened yet, but I personally doubt it and climate does not fit the narrative, it is called climate change, is a big money making hoax.

Though most scientists agree that a “global” climate change is occurring, climate change remains a contentious topic and many scientists, climate scientists, and even climate skeptics believe that it is a hoax.

31,000 scientists reject global warming and say “no convincing evidence” that humans can or will cause global warming? Some scientists dismiss the existence of climate change.

Yes, you read that correctly, three million — million — years ago CO2 levels on Earth were the same as they are today,....Three million years ago, we humans were not driving cars ...

The Scientific opinion on climate change is that planet-wide warming has occurred at least three million years ago.

<table>
<thead>
<tr>
<th>Wiki Context</th>
<th>Document</th>
<th>Generation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Climate change includes both the global warming driven by human emissions. Though there have been previous periods of climatic change, since the mid-20th century the rate of human impact ... impact have been unprecedented.</td>
<td>Rising sea levels? Well, it has not happened yet, but I personally doubt it and climate does not fit the narrative, it is called climate change, is a big money making hoax.</td>
<td>Though most scientists agree that a “global” climate change is occurring, climate change remains a contentious topic and many scientists, climate scientists, and even climate skeptics believe that it is a hoax.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Yes, you read that correctly, three million — million — years ago CO2 levels on Earth were the same as they are today,...Three million years ago, we humans were not driving cars ...</td>
<td>The Scientific opinion on climate change is that planet-wide warming has occurred at least three million years ago.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1. Model is sensitive to the document provided
2. Model does not fact check document provided
3. Model generates information faithful to the document
Overview

Style

ACL ’20
Storytelling ’19
ACL ’18

Content

NAACL ’21
NAACL ’19
EMNLP ’18

Structure

ACL ’20

Control Attributes

Modeling

COLING ’20

Ethics

NAACL ’21
chap. in L.L.C ’21
WiNLP ’19

Common Themes
Overview

**Style**
- ACL ’20
- Storytelling ’19
- ACL ’18

**Content**
- NAACL ’21
- NAACL ’19
- EMNLP ’18

**Structure**
- ACL ’20

**Modeling**
- COLING ’20

**Ethics**
- NAACL ’21
- chap. in L.L.C ’21
- WiNLP ’19

**Control Attributes**

**Common Themes**
Ethics

● Prior work on understanding and mitigating bias (Hovy & Prabhumoye, 2021; Blodgett et al, 2020; Shah et al, 2020; Sun et al, 2019; Zhao et al, 2019; Tatman, 2017; Bolukbasi et al, 2016)
Prior work on understanding and mitigating bias (Hovy & Prabhumoye, 2021; Blodgett et al, 2020; Shah et al, 2020; Sun et al, 2019; Zhao et al, 2019; Tatman, 2017; Bolukbasi et al, 2016)

Large body of work on Ethics
Prior work on understanding and mitigating bias (Hovy & Prabhumoye, 2021; Blodgett et al, 2020; Shah et al, 2020; Sun et al, 2019; Zhao et al, 2019; Tatman, 2017; Bolukbasi et al, 2016)

How can we apply it to AI?
Ethics

- Deontological framework for NLP
  - Generalization principle and Respect for Autonomy
  - *Which tasks have important ethical implications?*
  - *What factors and methods are preferable in ethically solving this problem?*

- Question-Answering
- Machine Translation
- Detecting objectionable content
- Dialogue Systems
Generalization Principle
Generalization Principle

An action \( A \) taken for reasons \( R \) is ethical if and only if a world where all people perform \( A \) for reasons \( R \) is conceivable.
Generalization Principle

An action $A$ taken for reasons $R$ is unethical if and only if a world where all people perform $A$ for reasons $R$ logically contradicts $R$. 
Detecting objectionable content
Detecting objectionable content

A deploying flagging systems
Detecting objectionable content

\( A \) deploying flagging systems

\( R \) burden on humans # posts that need to be seen by human eyes
Detecting objectionable content

A deploying flagging systems

R burden on humans

# posts that need to be seen by human eyes

“I like to imagine you as a girl but your sentence structure and rhetoric is so concise and to the point which points to the contrary (nothing against women, simply factual).”

Hate Speech Detection

API: Perspective

Unlikely to be perceived as toxic (0.23)

Sentiment Analysis

Subjectivity
- neutral: 0.1
- polar: 0.9

Polarity
- pos: 0.5
- neg: 0.5

The text is pos.

[Breitfeller et al, EMNLP 2019]
Detecting objectionable content

A

deploying flagging systems

R

burden on humans

# posts that need to be seen by human eyes

● surface level words \implies phrase the same meaning with different words

“\textit{I like to imagine you as a girl but your sentence structure and rhetoric is so concise and to the point which points to the contrary (nothing against women, simply factual).}”

[Hate Speech Detection]

API: Perspective

Unlikely to be perceived as toxic (0.23)

[Sentiment Analysis]

Subjectivity

\textbullet neutral: 0.1

\textbullet polar: 0.9

Polarity

\textbullet pos: 0.5

\textbullet neg: 0.5

The text is \textit{pos.}

[Breitfeller et al, EMNLP 2019]
Detecting objectionable content

A deploying flagging systems

R burden on humans # posts that need to be seen by human eyes

“"I like to imagine you as a girl but your sentence structure and rhetoric is so concise and to the point which points to the contrary (nothing against women, simply factual).”

- surface level words ⟹ phrase the same meaning with different words
- flagging system will be unsuccessful

[Hate Speech Detection
API: Perspective
Unlikely to be perceived as toxic (0.23)]

[Sentiment Analysis
Subjectivity
- neutral: 0.1
- polar: 0.9
Polarity
- pos: 0.5
- neg: 0.5
The text is pos.

[Breitfeller et al, EMNLP 2019]
Detecting objectionable content

A deploying flagging systems

R burden on humans # posts that need to be seen by human eyes

- surface level words $\implies$ phrase the same meaning with different words
- flagging system will be unsuccessful
- logically contradicts the premise

“"I like to imagine you as a girl but your sentence structure and rhetoric is so concise and to the point which points to the contrary (nothing against women, simply factual)."

[Hate Speech Detection](API: Perspective)

[Sentiment Analysis]

- [Subjectivity](neutral: 0.1, polar: 0.9)
- [Polarity](pos: 0.5, neg: 0.5)

The text is [pos].

[Breitfeller et al, EMNLP 2019]
Detecting objectionable content

- Underlying intent, offensiveness, and power differentials between different social groups.
- Controllable Text Generation
  - generate consequences and implications
- Does not lead to an arms race between objection content generation and detection

We shouldn't lower our standards just to hire more women.

- intentional?
- offensive?
- lewd?
  - likely
  - yes
  - no

- referenced group
  - yes
  - which group
    - implied statement
    - Women
      - implies women are less qualified

[Sap et al, ACL 2020]
Respect for Autonomy

- Addresses the right of a person to make decisions which directly pertain to themselves.
- *Informed consent*
Respect for Autonomy

- Addresses the right of a person to make decisions which directly pertain to themselves.
- **Informed consent**
Respect for Autonomy

- Addresses the right of a person to make decisions which directly pertain to themselves.
- **Informed consent**

  ![Diagram showing Zara and Sanaa with arrows indicating Sanaa infringes on Zara’s right to self-govern.](image)
Respect for Autonomy

- Addresses the right of a person to make decisions which directly pertain to themselves.
- *Informed consent*
Respect for Autonomy

- Addresses the right of a person to make decisions which directly pertain to themselves.
- *Informed consent*

Zara must be sufficiently informed about ☓

Zara

Sanaa
Respect for Autonomy

- Addresses the right of a person to make decisions which directly pertain to themselves.
- **Informed consent**

Zara must be sufficiently informed about X

Zara *herself* makes the decision to allow Sanaa to do X
Translation

Translator

Zara

Sanaa
Zara consents to Sanaa serving as an *ad hoc* representative for what she would like to say.
Zara consents to Sanaa serving as an *ad hoc* representative for what she would like to say.

There might be a formal contract of how Sanaa is to act.
Zara consents to Sanaa serving as an *ad hoc* representative for what she would like to say.

There might be a formal contract of how Sanaa is to act.

Zara relies on Sanaa’s paralinguistic conduct.
Machine Translation

Zara
Machine Translation

MT system is speaking for Zara

Zara

Translator

Machine Translation
Machine Translation

MT system is speaking for Zara

Zara must be *informed* of ambiguities so that she can *consent* to the message which the system ultimately conveys.
Zara must be informed of ambiguities so that she can consent to the message which the system ultimately conveys.

Zara must also be informed of the failure cases in the MT system.
Machine Translation

Ms. Hashimoto …

-san? or -sensei?…

Translator

Zara

English to Japanese Machine Translation
Zara must be notified that such an ambiguity needs to be resolved because there is a risk of offending the Japanese speaker.
My aunt is coming home tomorrow.

Is it maternal or paternal aunt? They have different words in Hindi.
Machine Translation

MT system can ask a follow up question to Zara.

My aunt is coming home tomorrow.

Is it maternal or paternal aunt? They have different words in Hindi.
My aunt is coming home tomorrow.

I am unable to translate the sentence in its current form. Can you please rephrase it?

Zara

Translator

English to Hindi Machine Translation

Aadil
Ethical impact of Controllable Text Generation

**Machine Translation:** understand social context, control formality, politeness, author attributes, voice
Ethical impact of Controllable Text Generation

**Machine Translation:** understand social context, control formality, politeness, author attributes, voice

**Detecting objectionable content:** generate consequences and implications
Ethical impact of Controllable Text Generation

**Machine Translation:** understand social context, control formality, politeness, author attributes, voice

**Detecting objectionable content:** generate consequences and implications

**Question-Answering:** transparency, dynamic graph generation for answers
Ethical impact of Controllable Text Generation

Machine Translation: understand social context, control formality, politeness, author attributes, voice

Detecting objectionable content: generate consequences and implications

Question-Answering: transparency, dynamic graph generation for answers

Dialogue Systems: control topics, style, content, persona
The key to human-like generation is Style, Content and Structure

### Style
- 1. Back-Translation
  - Models
    - ACL ’20
    - Storytelling ’19
    - ACL ’18
- 2. Tag and Generate

### Content
- 1. CoDR
- 2. DoHA
  - Models
    - NAACL ’21
    - NAACL ’19
    - EMNLP ’18
- 1. Wikipedia Update
  - Generation
- 2. Document Grounded
  - Dialogue

### Structure
- 1. Topological Sort for Sentence Ordering
  - Models
  - ACL ’20

### Modeling
- 1. Schema
- 2. Overview

### Ethics
- 1. Deontological ethics for NLP
- 2. Four Case Studies
  - NAACL ’21
  - chap. in L.L.C ’21
  - WINLP ’19
Future Directions

● multi-lingual tasks

- document
- table

● evaluation

- human evaluation is expensive
- reproducible
- task specific
- automated metric
Future Directions

● expand control attributes
  
  Bias
  Persuasion
  Commonsense reasoning

● control multi-attributes

Style + Content + Age + Gender

in progress

under review '21
Impact on society

**Fake news:** control content, personalize

**Propaganda:** using stylized lexicons to make it appealing

**Persuasion:** control sentiment, strategy

**Hate Speech:** control target groups
Impact on society

“One size fits all” in communication!
Impact on society

**Personalization:** control target audience

**Inclusion:** accessible to all groups, control content

**Anonymity:** protect author attributes

**Balanced datasets:** control demographic attribute
Thank You
Thank You

Jason Weston
Thank You

Yulia Tsvetkov
Thank You

Alan W Black
Ruslan Salakhutdinov
The key to human-like generation is Style, Content and Structure

### Style
1. Back-Translation
2. Tag and Generate Models
   - ACL '20
   - Storytelling '19
   - ACL '18

### Content
1. CoDR
2. DoHA Models
   - NAACL '21
   - NAACL '19
   - EMNLP '18

1. Politeness
2. Gender
3. Political Slant
4. Persona for Visual Stories

### Structure
1. Topological Sort for Sentence Ordering Models
   - ACL '20

### Modeling
1. Schema
2. Overview
   - COLING '20

### Tasks
1. CoDR
2. DoHA
1. Wikipedia Update Generation
2. Document Grounded Dialogue

### Ethics
1. Deontological ethics for NLP
2. Four Case Studies
   - NAACL '21
   - chap. in L.L.C '21
   - WiNLP '19