Controlling style, content and structure in natural language generation

Shrimai Prabhumoye

Carnegie Mellon University
Language Technologies Institute
We constantly author text!

Social Media

Personal/Professional documents

Blogs

News/Wiki articles

Style, Content and Structure
Model

Style

Content

Structure

Generated Text
Style

- surface level realization

“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)
User1: The Notebook is hands-down one of my favorite movies EVER! Have you ever seen The Notebook?

User2: No I have never seen this movie. I am going to try it out now.

User1: It was a heartwarming story of young love. The main characters are played by Ryan Gosling and Rachel McAdams.

User2: Ok this sounds nice. I think Ryan is a good actor.

User1: For all the praise it received, I was surprised to see that it only got a 5.7/10 on Rotten Tomatoes.

User2: That is interesting. They never get the rating correct.

User1: Ryan is a great actor, as well as Rachel McAdams. The story goes back and forth between present day and the past. Older Ryan is played by James Garner and older Rachel is played by Gena Rowlands. Yeah, Rotten Tomatoes never gets the right ratings..LOL. I always like to see the ratings but if I want to see a movie, I will watch it even if it has a bad rating.
Structure

“The Lead”: The most important information
May include a “hook” (provocative quote or question)

“The Body”: The crucial information
Argument, Controversy, Story, Issue
Evidence, background, details, logic, etc
Quotes, photos, support, dispute, expand

“The Tail”: extra information
Interesting/Related items like blogs, other editorials
the moment where the important detail that was intentionally concealed, is revealed.

end with “you had to be there”

beginning

classical structure of your uncle telling a joke at a party

point of story revealed

awkward pause
I have a talk to present today!

May the force be with you!

Style
Hey, have you watched Avengers Endgame

Yeah, I loved Robert Downey in it!
Hey, have you watched *Avengers: Endgame*?

Yeah, I loved the movie! Oh great! Can you tell me the story!

Yeah, Thor decapitates Thanos. Five years later, AntMan escapes from the quantum realm. Ironman builds a time machine to save the world. Hulk travels to New York City in 2012 and convinces the Ancient One to give him the Time Stone. Ironman steals the Infinity Stones back from Thanos and uses them to disintegrate Thanos and his army, at the cost of his life.
Other Application

- lexical choices, information, coherence
- Writing Assistance Tools
- Content Generation (websites, descriptions etc)
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.”
Applications

Anonymization: To preserve anonymity of users online, for personal security concerns (Jardine, 2016), or to reduce stereotype threat (Spencer, 1999).

Demographically-balanced training data for downstream applications.
Our Goal

To create a representation that is devoid of style but holds the meaning of the input sentence.
Challenges

● No Parallel Data!

“The movie was very long.”

“I entered the theatre in the bloom of youth and emerged with a family of field mice living in my long, white mustache.”

● Style is subtle
Our Solution

- Back-Translation
  - Translating an English sentence to a pivot language and then back to English.
- Reduces the stylistic properties
- Helps in grounding meaning
- Creates a representation independent of the generative model
- Representation is agnostic to the style task
I thank you, Rep. Visclosky

MT e ⟷ f
encoder  decoder

je vous remercie, Rep. Visclosky

MT f ⟷ e
encoder

Style 1
decoder

I thank you, senator Visclosky

Style 2
decoder

I'm praying for you sir.
Train Pipeline

- Style 1 decoder
- Style 2 decoder
- $\hat{x}_{style1}$
- $\hat{x}_{style2}$
- Classifier
Experimental Settings

- Encoder-Decoders follow sequence-to-sequence framework (Sutskever et al., 2014; Bahdanau et al., 2015)

\[
\min_{\theta_{gen}} \mathcal{L}_{gen} = \mathcal{L}_{recon} + \lambda_c \mathcal{L}_{class}
\]
### Style Tasks

<table>
<thead>
<tr>
<th>Task</th>
<th>Labels</th>
<th>Corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>Male, Female</td>
<td>Yelp</td>
</tr>
<tr>
<td>Political Slant</td>
<td>Republican, Democratic</td>
<td>Facebook Comments</td>
</tr>
<tr>
<td>Sentiment Modification</td>
<td>Negative, Positive</td>
<td>Yelp</td>
</tr>
</tbody>
</table>
Style Transfer Accuracy

- Generated sentences are evaluated using a pre-trained style classifier
- Test the classification accuracy of the generated sentences for the desired label.

<table>
<thead>
<tr>
<th>Classifier Model</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>82%</td>
</tr>
<tr>
<td>Political Slant</td>
<td>92%</td>
</tr>
<tr>
<td>Sentiment Modification</td>
<td>93.23%</td>
</tr>
</tbody>
</table>
Style Transfer Accuracy

- **Gender**
  - Baseline (Shen et al, 2017): 60.4
  - Ours: 57.04

- **Political Slant**
  - Baseline (Shen et al, 2017): 88.01
  - Ours: 80.43

- **Sentiment Modification**
  - Baseline (Shen et al, 2017): 75.82
  - Ours: 87.22
Preservation of Meaning

Which transferred sentence maintains the same semantic intent of the source sentence while changing the political position

<table>
<thead>
<tr>
<th>Gender</th>
<th>Political Slant</th>
<th>Sentiment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (Shen et al, 2017)</td>
<td>Ours</td>
<td>No Preference</td>
</tr>
<tr>
<td>15.23</td>
<td>43.41</td>
<td>35.91</td>
</tr>
<tr>
<td>14.55</td>
<td>41.36</td>
<td>23.18</td>
</tr>
<tr>
<td>14.55</td>
<td>39.55</td>
<td>35.91</td>
</tr>
<tr>
<td>15.23</td>
<td>41.36</td>
<td>40.91</td>
</tr>
</tbody>
</table>

Which transferred sentence is semantically equivalent to the source sentence with an opposite sentiment
Discussion

- generator loss function has two competing terms: improve meaning preservation and improve style transfer accuracy

- Sentiment modification task is not well-suited for evaluating style transfer

- The style-transfer accuracy for gender task is lower for BST model but the preservation of meaning is much better compared to baseline model and “No preference” option
Fluency

Baseline (Shen et al, 2017)

Ours

- Gender: 2.42, 2.81
- Political Slant: 2.79, 2.87
- Sentiment: 2.7, 3.09, 2.91, 3.18
- Overall: 2.7, 2.91
- Overall short: 2.18, 3.05, 3.11
- Overall Long: 2.18, 2.62
Gender Examples

- Male -- Female

  *my wife ordered country fried steak and eggs.*

  *My husband ordered the chicken salad and the fries.*

- Female -- Male

  *Save yourselves the huge headaches,*

  *You are going to be disappointed.*
I will continue praying for you and the decisions made by our government!

I will continue to fight for you and the rest of our democracy!

As a hoosier, I thank you, Rep. Vislosky.

As a hoosier, I’m praying for you sir.
Sentiment Modification Examples

- Negative -- Positive
  
  *This place is bad news!*
  
  *This place is amazing!*

- Positive -- Negative

  *The food is excellent and the service is exceptional!*
  
  *The food is horrible and the service is terrible.*
Overview

Style

Content

Structure

Storytelling '19
ACL '18

NAACL '19
EMNLP '18

under review
ICLR '20
Content Transfer

- AI assistance deals with *form* (grammar, style, etc.)
- Our goal is to control for *content*
What is our task?

On 4 July 2011 several publications including the Daily Mail,\[8\] The Telegraph, and The Guardian\[11\] picked up the story and published the pictures along with articles that quoted Slater as describing the photographs as self-portraits taken by the monkeys: “Monkey steals camera to snap himself” (The Telegraph), “a camera on a tripod” triggered by the monkeys (The Guardian), and a camera started by a monkey “Fascinated by her reflection in the lens.” The articles also contained Slater quotes such as “He must have taken hundreds of pictures by the time I got my camera back.” 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.

Ape-rture priority photographer plays down monkey reports

Chris Cheesman July 5, 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.

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.
Primary Contribution

- design a task to perform content transfer from an unstructured source of information
- release dataset
Applications

Software Documentation

Wikipedia Articles

Inbox

Legal Summarization

Document

Legal

Wikipedia Articles
Data Creation Process

Context

Update

HTML News Article

Plain Text of News Article

Common Crawl
## Data Creation Process

### Total Data Size: 636K

<table>
<thead>
<tr>
<th>News Article</th>
<th>Wikipedia Context</th>
<th>Update</th>
</tr>
</thead>
<tbody>
<tr>
<td>…</td>
<td>…</td>
<td>…</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
<td>…</td>
</tr>
</tbody>
</table>

---

**Plain Text of News Article**
Models

Generative Models

• Context Agnostic Generative Model (CAG) — Baseline
• Context Informed Generative Model (CIG)
• Context Responsive Generative Model (CRG)

all models have global attention

Extractive Models

• SumBasic
• Context Informed SumBasic
• Oracle

all models are simplistic to infer if context helps in generation
Context Agnostic Model (CAG) - Baseline

News Article

x_1 \rightarrow LSTM \rightarrow x_2 \rightarrow LSTM \rightarrow x_3 \rightarrow LSTM

Encoder Vector

y_1 \rightarrow LSTM \rightarrow y_2 \rightarrow LSTM \rightarrow y_3 \rightarrow LSTM

<start>
Context Informed Model (CIG)
Context Responsive Model (CRG)
Extractive Models

- **SumBasic**: a model based on unigram probabilities
- **Context Informed SumBasic**: the unigram probabilities take into account the words in the Wikipedia context.
- **Oracle**: establish an upper limit attainable by extractive methods
### Automated Evaluation

<table>
<thead>
<tr>
<th>Model</th>
<th>ROUGE-L</th>
</tr>
</thead>
<tbody>
<tr>
<td>SumBasic</td>
<td>5.6 (5.6-5.7)</td>
</tr>
<tr>
<td>Context Informed SumBasic</td>
<td>7.0 (7.0-7.1)</td>
</tr>
<tr>
<td>Context Agnostic Generative Model</td>
<td>9.1 (9.0-9.2)</td>
</tr>
<tr>
<td>Context Informed Generative Model</td>
<td><strong>16.0 (15.9-16.1)</strong></td>
</tr>
<tr>
<td>Context Responsive Generative Model</td>
<td>14.7 (14.6-14.8)</td>
</tr>
<tr>
<td>Oracle</td>
<td>28.8 (28.7-29.0)</td>
</tr>
</tbody>
</table>

* METEOR and BLEU numbers are consistent with ROUGE-L
Relative Human Evaluation

Which system output is closest in meaning to the reference update?

Which system output is more accurate relative to the background information given in the snippet of the article?
Absolute Quality Evaluation

<table>
<thead>
<tr>
<th>Category</th>
<th>CAG</th>
<th>CIG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grammar</td>
<td>2.6</td>
<td>4.3</td>
</tr>
<tr>
<td>Non-redundancy</td>
<td>1.8</td>
<td>3.9</td>
</tr>
<tr>
<td>Referential Quality</td>
<td>2.7</td>
<td>3.6</td>
</tr>
<tr>
<td>Focus</td>
<td>2.6</td>
<td>3.5</td>
</tr>
<tr>
<td>Structure and Coherence</td>
<td>2.4</td>
<td>3.2</td>
</tr>
<tr>
<td>Reference Update</td>
<td>CIG Generated Update</td>
<td></td>
</tr>
<tr>
<td>-------------------------------------------------------</td>
<td>-----------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>rob brydon, the comedian was born in baglan.</td>
<td>he was born in baglan.</td>
<td></td>
</tr>
<tr>
<td>in may 2014 he was diagnosed with prostate cancer.</td>
<td>st. clair was diagnosed with prostate cancer.</td>
<td></td>
</tr>
<tr>
<td>on april 3, 2014, Manning signed a one-year deal with the cincinnati bengals.</td>
<td>on march 9, 2014, Manning signed a one-year contract with the cincinnati bengals.</td>
<td></td>
</tr>
</tbody>
</table>
Overview

Style

NAACL '19
ACL '18
Storytelling '19

Content

EMNLP '18
NAACL '19

Structure

under review
ICLR '20

ACL '18
EMNLP '18
under review
ICLR '20
LGBT Rights in India

**Contemporary times**  
See also: *Navtej Singh Johar v. Union of India*

In 2003, the Indian Government said that legalising homosexuality would "open the floodgates of delinquent behaviour".[23]

In 2009, the Delhi High Court decision in *Naz Foundation v. Govt. of NCT of Delhi* found Section 377 and other legal prohibitions against private, adult, …

On 23 February 2012, the *Ministry of Home Affairs* expressed its opposition to the decriminalisation of homosexual activity, stating that in India, homosexuality is seen as being immoral.[28] The Central Government reversed its stance on 28 February 2012, asserting that there was no legal error in decriminalising …
LGBT Rights in India

On 11 December 2013, the Supreme Court set aside the 2009 Delhi High Court order decriminalising consensual homosexual activity within its jurisdiction.[30][31][32][33][34]

On 28 January 2014, the Supreme Court of India dismissed the review petition filed by the Central Government, the Naz Foundation and several others against

On 18 December 2015, Shashi Tharoor, a member of the Indian National Congress party, introduced a bill for the repeal of Section 377, but it was rejected

On 2 February 2016, the Supreme Court decided to review the criminalisation of homosexual activity.[41] In August 2017, the Supreme Court unanimously ruled
LGBT Rights in India

On 6 September 2018, the Supreme Court issued its verdict. The Court unanimously ruled that Section 377 is unconstitutional as it infringed on the fundamental rights of autonomy, intimacy, and identity, thus legalising homosexuality in India. The Court explicitly overturned its 2013 judgement.

Criminalising carnal intercourse is irrational, arbitrary and manifestly unconstitutional.

— Chief Justice Dipak Misra

Participants of a 2018 Bhopal parade celebrating the ruling of the Supreme Court
India court legalises gay sex in landmark ruling

How did we get to this point?

It's been a tortuous route. A bid to repeal section 377 was initiated in 2001 and was batted between court and government until 2009, when the Delhi High Court ruled in favour of decriminalisation.

under which gay sex is categorised as an "unnatural offence".
Prior Work

Learning Structured Text Representations (Liu and Lapata, 2018)
Prior Work

- Hierarchical Attention Networks for Document Classification (Yang et al, 2016)

- Deep Attentive Sentence Ordering Network (Cui et al, 2018)

- Language Model Pre-Training for Hierarchical Document Representations (Chang et al, 2019)
## Success

### Evaluating Discourse in Structured Text Representations (Ferracane et al, NAACL 2019)

<table>
<thead>
<tr>
<th></th>
<th>Yelp</th>
<th>Debates</th>
<th>Writing Quality</th>
<th>Writing Quality Topic Control</th>
<th>Wall Street Journal Sent order</th>
</tr>
</thead>
<tbody>
<tr>
<td>Liu et al (original)</td>
<td>68.27 (0.19)</td>
<td>79.48 (2.90)</td>
<td>82.69 (1.36)</td>
<td>79.63 (1.03)</td>
<td>95.29 (0.84)</td>
</tr>
<tr>
<td>Liu et al (ours)</td>
<td>68.23 (0.23)</td>
<td>77.81 (1.80)</td>
<td>82.70 (1.36)</td>
<td>81.11 (0.95)</td>
<td>94.76 (1.11)</td>
</tr>
<tr>
<td>- doc attn</td>
<td>68.13 (0.17)</td>
<td>81.42 (1.08)</td>
<td>82.80 (0.94)</td>
<td>79.25 (0.94)</td>
<td>95.11 (0.42)</td>
</tr>
<tr>
<td>-both attn</td>
<td>68.05 (0.13)</td>
<td>77.34 (1.79)</td>
<td>83.16 (1.25)</td>
<td>76.16 (1.25)</td>
<td>94.68 (0.37)</td>
</tr>
<tr>
<td>Liu et al (reported)</td>
<td>68.6</td>
<td>76.5</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
## Nature of Trees

<table>
<thead>
<tr>
<th></th>
<th>Yelp</th>
<th>Debates</th>
<th>WQ</th>
<th>WQTC</th>
<th>WSJSO</th>
</tr>
</thead>
<tbody>
<tr>
<td>tree height</td>
<td>2.049</td>
<td>2.751</td>
<td>2.909</td>
<td>4.035</td>
<td>2.288</td>
</tr>
<tr>
<td>prop. of leaf nodes</td>
<td>0.825</td>
<td>0.849</td>
<td>0.958</td>
<td>0.931</td>
<td>0.892</td>
</tr>
<tr>
<td>norm. arc length</td>
<td>0.433</td>
<td>0.397</td>
<td>0.420</td>
<td>0.396</td>
<td>0.426</td>
</tr>
<tr>
<td>% vacuous trees</td>
<td>73%</td>
<td>38%</td>
<td>42%</td>
<td>14%</td>
<td>100%</td>
</tr>
</tbody>
</table>

- shallow trees
- picks first or last two sentences as root and all other sentences are children
My Questions

- Does structure help in generating documents?
- Do we need hierarchical architectures to capture structure?
- How to evaluate? Rhetorical Structure Theory? Something higher?
Dialog Structures

- Light Dataset: 663 locations, 3462 objects and 1755 characters
- Light is a stateful game engine
**Tasks**

---

**Speak and Act**

**Environment**

**RL Agent**

- Gets a goal to achieve
- Speak

---

**RL Approach:** Policy has to pick a topic of the next utterance such that the goal is eventually achieved.
Example

RL Agent: King; Environment Agent: Queen

RL Persona: I am a king of the whole empire. I give rules and pursue them. I am brave and fearless.

Setting: Town Center, Town
A lavish town full of stone buildings, most of which are two or three stories. They have ornate tin gutter fixtures with the water seeming to pour from the mouths of different animals.

Goal for RL Agent: hug king (The king has to say something to queen such that queen performs the action of hugging the king)

King: Hello my Queen, beautiful as ever!
Queen: Hello my dear, do you feel we are a prosperous kingdom these days?
Queen: hug king
## Success

<table>
<thead>
<tr>
<th>Model</th>
<th>Goal Type</th>
<th>Test Seen</th>
<th></th>
<th>Test Unseen</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>(n = 1)</td>
<td>(n = 3)</td>
<td>(n = 1)</td>
<td>(n = 3)</td>
</tr>
<tr>
<td></td>
<td>Reward</td>
<td>Reward</td>
<td>Turns</td>
<td>Reward</td>
<td>Reward</td>
</tr>
<tr>
<td>Inverse model</td>
<td>game act</td>
<td>0.223</td>
<td>0.414</td>
<td>2.42</td>
<td>0.193</td>
</tr>
<tr>
<td>Topic RL</td>
<td>game act</td>
<td>0.314</td>
<td>0.502</td>
<td>2.25</td>
<td>0.268</td>
</tr>
<tr>
<td>Top-K RL</td>
<td>game act</td>
<td>0.319</td>
<td>0.460</td>
<td>2.37</td>
<td>0.270</td>
</tr>
<tr>
<td>Inverse model</td>
<td>emote</td>
<td>0.089</td>
<td>0.262</td>
<td>2.72</td>
<td>0.088</td>
</tr>
<tr>
<td>Topic RL</td>
<td>emote</td>
<td>0.170</td>
<td>0.342</td>
<td>2.63</td>
<td>0.146</td>
</tr>
</tbody>
</table>
Model

Style

Content

Structure

Generated Text
Thank You!

Alan W Black (co-advisor)
Ruslan Salakhutdinov (co-advisor)
Yulia Tsvetkov
Chris Quirk
Michel Galley
Jason Weston