Neural Text Generation

Khyathi Chandu
kchandu@cs.cmu.edu

Natural Language Processing (11-411, 11-611)
Lecture 23
Theme and Takeaways

• **Theme:**
  - Run-through of latest advances in language generation field

• **Takeaways:**
  - Broad understanding of different approaches in text generation
    - Methods and tasks
  - Some prospective directions to work on
    - Key Challenges (for instance evaluation)
What is Text Generation?

• **Text Generation** is the task of producing written or spoken narrative from structured or unstructured data.  
  • produce understandable texts in English or other human languages from some underlying non-linguistic representation of information [1]

• **Goals:**
  • Seamless human-machine communication
  • Presenting a wealth of data in a way humans can comprehend
  • Individual level personalized content at scale

Text Generation is all around us!

Machine Translation

We are studying text generation today.

Nous étudions la génération de texte aujourd'hui.
Text Generation is all around us!

Machine Translation

We are studying text generation today.

Nous étudions la génération de texte aujourd'hui.

Dialog Response

Hello Britney! I am BMC Helix Chatbot. How can I help you today?

Hello. I want the employee verification letter.

What is the purpose?

Home loan | Lease agreement | Credit card application | Immigration visa

Immigration visa

Do you want to include date of birth?

Yes

Do you want to include date of joining?

Yes

Do you want to include the salary details?

No

Do you want to include any additional details?

No

Please wait while I connect you to a Live Agent.
Text Generation is all around us!

Machine Translation

We are studying text generation today.

Nous étudions la génération de texte aujourd'hui.

Dialog Response

Hello Britney! I am BMC Helix Chatbot. How can I help you today?

Hello. I want the employee verification letter.

What is the purpose?
Home loan | Lease agreement | Credit card application | Immigration visa

Do you want to include date of birth?
Yes

Do you want to include date of joining?
Yes

Do you want to include the salary details?
No

Do you want to include any additional details?
No

Please wait while I connect you to a Live Agent.

Chat summary

I want the employee verification letter. What is the purpose? Immigration visa. Do you want include date of birth? Yes. Do you want include date of joining? Yes. Do you want to include the salary details? No.

Live agent (Allen)

I understand that you want the employee verification letter with date of birth and date of joining and without salary details. I can help you with that.

Pic credits: https://docs.bmc.com/docs/helixplatform/support-for-text-summaries-in-your-application-888488619.html

Summarization
Text Generation is all around us!

Machine Translation

We are studying text generation today.

Nous étudions la génération de texte aujourd’hui.

Visual Narratives

1. Preheat oven to 220°C (425°F). Place the pork, breadcrumbs, garlic, ginger, sesame seeds, sesame oil, shochu, soy sauce, green onion, egg, salt and pepper in a large bowl and mix well to combine.

2. Roll tablespoonsfuls of the pork mixture into balls and set aside. Pour the oil into the hot tray, add the meatballs and cook for 5 minutes.

3. To make the sticky soy sauce, place the dashi, soy sauce, mirin, sesame oil, vinegar and honey in a small bowl and mix to combine.

4. Remove the tray from the oven, add the dashi mixture and carefully toss to coat the meatballs. Cook for 15 minutes or until meatballs are golden. Serve the meatballs with the noodles.

Summarization

Live agent (Allen)

I understand that you want the employee verification letter with date of birth and date of joining and without salary details. I can help you with that.

Dialog Response

Hello Britney! I am BMC Helix Chatbot. How can I help you today?

Hello. I want the employee verification letter. What is the purpose? Immigration visa. Do you want to include date of birth? Yes. Do you want to include date of joining? Yes. Do you want to include the salary details? No.

Chat summary

I want the employee verification letter. What is the purpose? Immigration visa. Do you want to include date of birth? Yes. Do you want to include date of joining? Yes. Do you want to include the salary details? No.
Text Generation is all around us!

Machine Translation

Visual Narratives

Explanation Generation

Summarization

Dialog Response

Pic credits: https://docs.bmc.com/docs/helixplatform/support-for-text-summarization-in-your-application-888488619.html
Paradigms of Text Generation

- **Text-to-text**
  - Input: 
  - Process: Gears
  - Output: 

- **Data-to-text**
  - Input: 
  - Process: Gears
  - Output: 

- **Control-Free**
  - Input: NULL
  - Process: Gears
  - Output: 

Tasks of NLG: Applications

Sentence Level
- QG and QA
- Machine Translation
- Text Simplification
- Paraphrase Generation

Discourse Level
- Summarization (Abstractive)
- Peer Review Generation
- Story Generation
- Dialog
- Explanatory AI
- Poetry Generation

Cross-Modal
- Automatic Speech Recognition
- Image Captioning

Visual Storytelling
### Schema of In and Out

<table>
<thead>
<tr>
<th>Task</th>
<th>Input</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dialog</td>
<td>Conversation History</td>
<td>Next Response</td>
</tr>
<tr>
<td>Machine Translation</td>
<td>Source Language</td>
<td>Target Language</td>
</tr>
<tr>
<td>Style Transfer</td>
<td>Style 1 Text</td>
<td>Style 2 Text</td>
</tr>
<tr>
<td>Summarization</td>
<td>Single/Multiple Documents (Question)</td>
<td>Summary</td>
</tr>
<tr>
<td>Image Captioning/Visual Storytelling</td>
<td>Image</td>
<td>Descriptive Text</td>
</tr>
<tr>
<td>Automatic Speech Recognition</td>
<td>Audio</td>
<td>Text</td>
</tr>
<tr>
<td>Table-Text</td>
<td>Table</td>
<td>Text</td>
</tr>
<tr>
<td>Poetry Generation</td>
<td>Null</td>
<td>Text</td>
</tr>
<tr>
<td>Language Modeling</td>
<td>Null</td>
<td>Sequence of Text</td>
</tr>
</tbody>
</table>
## Schema of In and Out

<table>
<thead>
<tr>
<th>Task</th>
<th>Input</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dialog</strong></td>
<td>Conversation History</td>
<td>Next Response</td>
</tr>
<tr>
<td><strong>Machine Translation</strong></td>
<td>Source Language</td>
<td>Target Language</td>
</tr>
<tr>
<td><strong>Style Transfer</strong></td>
<td>Style 1 Text</td>
<td>Style 2 Text</td>
</tr>
<tr>
<td><strong>Summarization</strong></td>
<td>Single/Multiple Documents (Question)</td>
<td>Summary</td>
</tr>
<tr>
<td><strong>Image Captioning/Visual Storytelling</strong></td>
<td>Image</td>
<td>Descriptive Text</td>
</tr>
<tr>
<td><strong>Automatic Speech Recognition</strong></td>
<td>Audio</td>
<td>Text</td>
</tr>
<tr>
<td><strong>Table-Text</strong></td>
<td>Table</td>
<td>Text</td>
</tr>
<tr>
<td><strong>Poetry Generation</strong></td>
<td>Null</td>
<td>Text</td>
</tr>
<tr>
<td><strong>Language Modeling</strong></td>
<td>Null</td>
<td>Sequence of Text</td>
</tr>
</tbody>
</table>
**Synergy between LM and NLG**

**Language Modeling**

\[ p(Y) = \prod_{t=1}^{T} p(y_t | y_{<t}) \]

**Generation**

\[ p(Y | X) = \prod_{t=1}^{T} p(y_t | X, y_{<t}) \]

\[ - \sum_{t=1}^{T} \log p(y_t | X, y_{<t}; \theta) \]

![Diagram](image)
Prior Methods: Eliza

History of NLG

- **Template based systems:**
  Uses rules and templates

- **Modeling Discourse Structures:**
  Relation learning, Rhetorical Structure Theory

- **RNNs, LSTMs, GRUs:**
  Autoregressive DNNs + Recurrent units, backpropagation

- **1965**
  Rule-based + Data Driven pipelines:
  Document planning + microplanning + realization

- **1978**
  1985

- **1990**
  Statistical Methods (Markov Chains):
  Sentence compression, reordering, lexical paraphrasing, syntactic transformation

- **2013**
  Transformers:
  GPT (1/2/3), GROVER, TransformerXL, DialoGPT

- **2017**

Credits: Tutorial on “The amazing world of Neural Text Generation”
Prior Methods in NLG

**WHAT to say?**

**Content Planner**
- Sentence Retrieval
- MMR
- Similarity Functions:
  - Tf-idf
  - Graph Algorithms

**WHEN to say?**

**Structure Planner**
- Sentence Reordering
- Sub-topic grouping
- Heuristic Grammar
- Templates
- Rhetorical Structure Theory

**HOW to say?**

**Surface Realizer**
- Dictionary lookup
- Cut and replace
- From ontologies and graph structures

**Classical**
- Attention
- Bottom Up (masking)
- Pointer Generator
- Coverage Penalty
- Diversity Loss

**Neural**
- Sequence of entities
- Dynamic and Static Schema
- Event Structures
- Hierarchical Structure
- Attention
- Latent variables (style)
## Prior Methods in NLG

### Classical
- Sentence Retrieval
- MMR
- Similarity Functions:
  - Tf-idf
  - Graph Algorithms

### Neural
- Attention
- Bottom Up (masking)
- Coverage Penalty
- Diversity Loss

### End-to-End
- Sequence of entities
- Dynamic and Static Schema
- Event Structures
- Hierarchical Structure

### Content Planner
- Sentence Retrieval
- MMR
- Similarity Functions:
  - Tf-idf
  - Graph Algorithms

### Structure Planner
- Sentence Reordering
- Sub-topic grouping
- Heuristic Grammar
- Templates
- Rhetorical Structure Theory

### Surface Realizer
- Dictionary lookup
- Cut and replace
- From ontologies and graph structures

### Diagram
- **Content Planner**
- **Structure Planner**
- **Surface Realizer**
- **End-to-End**
**NLG: Supervised Learning**

- Maximum Likelihood Objective:
  - Sequential multi-label cross entropy
- Teacher Forcing
  - Exposure Bias
    - Training Stage: Token in current time step is predicted | *ground truth prefix*
    - Inference Stage: Token in current time step is predicted | *generated prefix*
- Severe with increasing length
- Solution: Scheduled Sampling
  - Mix teacher forced embeddings and model predictions from previous time step

\[
\sum_{t=1}^{T} \log p(y_t | X, y_{<t}; \theta)
\]
**NLG: Supervised Learning**

**Tradeoffs and Challenges**

- **Text Degeneration**
  - **Problem**: tend to produce repetitive and frequently used tokens


**NLG: Supervised Learning**

*Tradeoffs and Challenges*

**Context:**
In a shocking finding, scientists discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

**Continuation (BeamSearch, b=10):**
"The unicorns were able to communicate with each other, they said unicorns. A statement that the unicorns. Professor of the Department of Los Angeles, the most important place the world to be recognition of the world to be a of the world to be a of the world to be a of the world to be a of the world to be a of the world to be a of the world to be a of the..."
**Text Degeneration**

- **Problem:** tend to produce repetitive and frequently used tokens
- **Solution:** Optimization objective
  - **Solution:** Unlikelihood training [1]
    \[
    L_{UL\text{-token}}^t(p_\theta(x_{<t}, c^t) = -\alpha \cdot \sum_{c \in c^t} \log(1 - p_\theta(c|x_{<t})) - \frac{\log p_\theta(x_t|x_{<t})}{\text{likelihood}}
    \]
    - Decrease the probability of negative candidates.
  - **Solution:** Maximum Mutual Information [2]
  - **Solution:** F2 softmax [3]
    - Factorize softmax with frequency of the class and the tokens

Distinguishability and hallucinations and repetitions

- **Solution**: Beam Blocking
  - **Problem**: Blocks natural repetition, for instance to emphasize
- **Solution**: Coverage Penalty
**Distinguishability and hallucinations**

- **Problem**: Hallucinations are unwanted byproducts of optimizing for log loss
- **Solution**: Beam Blocking
  - **Problem**: Blocks natural repetition, for instance to emphasize
- **Solution**: Coverage Penalty

**Decoding Sequence Length**

- **Problem**: As the length of sequence increases, the sum of log probability scores decreases => shorter hypothesis
- **Solution 1**: Normalize by length
- **Solution 2**: Multiply probability with a fixed constant at every timestep
- **Solution 3**: Bias based on relation between source and target lengths
- **Solution 4**: Length Penalty
  \[ l_p(Y) = \frac{(5 + |Y|)^\alpha}{(5 + 1)^\alpha} \]

---


ToTTo: A Controlled Table-To-Text Generation Dataset
**NLG: Reinforcement Learning**

- Supervised Learning: Mismatch between MLE and metrics for text quality
- Reinforcement Learning:
  - Optimize non-differentiable metrics
- RL Policy Gradient Algorithms (PG BLEU)
  - Optimize for BLEU
  - Optimized by REINFORCE
- Problems:
  - Computing BLEU is not cheap
  - Inherent inefficiency of metric
- Usually Policy Network is pre-trained with MLE

\[
E[\text{BLEU}(y_t)] - \sum_{t=1}^{T} R_t \log P(y_t | y_{<t}, X)
\]
**NLG: Reinforcement Learning**

**Tradeoffs and Challenges**

- Inefficiency of the reward
  - **Problem:** Automatic metric (like BLEU) does not correlate to human judgements
  - **Problem:** Handcraft task-specific model based metric to optimize
  - **Solution 1:** Adversarial Reward Learning
  - **Solution 2:** Inverse Reinforcement Learning

- Reward sparsity in long form texts
  - **Problem:** Reflecting high level semantics in long form texts
  - **Solution 1:** Optimize for abstract properties individually [1] [2] like:
    - Content
    - Structure
    - Surface Forms
  - **Problem:** Credit assignment and reward shaping in long form texts

**NLG: Adversarial Learning**

- Mismatch due to Teacher Forcing
- Professor Forcing:
  - Adversarial Domain Adaptation
  - Discriminator
    1. MLE for classification
- Generator
  1. MLE for data
  2. Fool the discriminator
    - Change free running network
    - Change teacher forcing network
- Regularization

\[
C_f(\theta_g|\theta_d) = E_{x\sim\text{data}, y\sim P_{\theta_g}(y|x)}[-\log D(B(x, y, \theta_g), \theta_d)]
\]

\[
C_t(\theta_g|\theta_d) = E_{(x,y)\sim\text{data}}[-\log (1 - D(B(x, y, \theta_g), \theta_d))]
\]

NLG: Adversarial Learning

Tradeoffs and Challenges

- **Problems:**
  - Discrete Sampling
  - Mode Collapse
  - Power dynamics between G and D
Pre-training for NLG Tasks

- UniLM
- MASS
- BART
- T5
Pre-training for NLG Tasks: Joint(Uni, Bi, S2S)

- **UniLM**: [1]
  - Tackles both NLU and NLG
  - Multi-layered Transformer
  - Depending on the use case, we can choose between:
    - Unidirectional LM
    - Bidirectional LM
    - Seq2seq LM (unidirectional decoding conditioned on bidirectional encoding)

Pre-training for NLG Tasks: Strategic Masking

- **MASS**: [2]
  - Encoder masks tokens
  - Decoder predicts only masked tokens
  - Intuitions:
    1. Complementary masking: encourages joint training
    2. Masking a fragment: improves NLU
    3. Predicting consecutive terms in decoder: improves NLG
  - k=1 → like BERT
  - k=sentence length → GPT

Pre-training for NLG Tasks: Denoising

- **BART** [3]
  - Bidirectional Encoder and Autoregressive Decoder
  - Denoising from corrupt text
  - Standard transformer based Neural MT framework
  - 5 transformations of data:
    - Token Masking
    - Token Deletion
    - Token Infilling
    - Sentence Permutation
    - Document Rotation

Pre-training for NLG Tasks: Text-to-text

- **T5**: [4]
  - Basic Idea: Treat all the NLP problems as text-to-text generation problems.

[4] Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer
Decoding Strategies

- What is decoding?
Decoding

Lateral
(The entire sequence is generated length wise i.e from side to side.)

Longitudinal
(Stratifies and predicts the sequence in layers)
Lateral Decoding

Autoregressive

Predict a future word given a sequence of words either from the past (left-to-right)

Encoder

AR-Decoder

thank you <eos>

<bos> thank you

Encoder

AR-Decoder

Hey happy birthday

Non-autoregressive

Generate all tokens of a sequence simultaneously

Encoder

NAR-Decoder

thank you <eos>

<bos> thank you

Encoder

NAR-Decoder

Hey happy birthday
Longitudinal Decoding

Hierarchical Decoding

Stratification of sequence into different layers

Decoder

Decoder

Decoder 1

Encoder

Hey  happy  birthday

(nouns)+(verbs)+rest of the sequence

(nouns)+ all verbs

all nouns
Decoding: Sampling and Search
Decoding Strategies

- **Random Sampling**
  
  Problem: No probability pruning

- **Greedy Decoding**
  
  Problem: No diversity
  
  No course correction

- **Top-k Sampling**
  
  Problem: Limiting to same \( k \)

- **Nucleus Sampling**

- **Beam Search**

  \[
P(y_t = w_i | y_{<t}, \mathbf{x}) = \frac{\exp(z_{t,i})}{\sum_{j \in V} \exp(z_{t,j})}
\]

  \[
  \hat{y}_t = \text{argmax}_y \left( \frac{\exp(z_{t,i})}{\sum_{j \in V} \exp(z_{t,j})} \right)
\]

  \[
P'(y_t | y_{<t}) = \begin{cases} 
  \frac{P(y_t | y_{<t})}{\sum_{y \in V^k} P(y_t | y_{<t})} & \text{if } y_t \in V^k \\
  0 & \text{otherwise}
\end{cases}
\]

  \[\sum_{y_t \in V^p} P(y_t | y_{<t}) \geq p\]
Decoding Strategies

- Random Sampling
  Problem: No probability pruning

- Greedy Decoding
  Problem: No diversity
  No course correction

- Top-k Sampling
  Problem: Limiting to same $k$

- Nucleus Sampling

\[
P(y_t = w_i \mid y_{<t}, \mathbf{x}) = \frac{\exp(z_{t,i})}{\sum_{j \in V} \exp(z_{t,j})}
\]

\[
\hat{y}_t = \text{argmax}_{y_t} \left( \frac{\exp(z_{t,i})}{\sum_{j \in V} \exp(z_{t,j})} \right)
\]

\[
P'(y_t \mid y_{<t}) = \begin{cases} 
\frac{P(y_t \mid y_{<t})}{\sum_{y_{<t}} P(y_t \mid y_{<t})} & \text{if } y_t \in V^k \\
0 & \text{otherwise}
\end{cases}
\]

\[
\sum_{y_t \in V^p} P(y_t \mid y_{<t}) \geq p
\]
**Decoding Strategies: Greedy**

- Random Sampling
- Greedy Sampling
- Top-k Sampling
- Nucleus Sampling

Greedy Decoder - feeds the best token to the next step

https://guillaumegenthial.github.io/sequence-to-sequence.html
Problems with greedy

- Problems:
  - No course correction
  - Accumulates error
  - No variation/creativity
    - Monotonous or somewhat predictable texts
Decoding Strategies: top-k

• Random Sampling
  Problem: No probability pruning

• Greedy Decoding
  Problem: No diversity
  No course correction

• Top-k Sampling
  Problem: Limiting to same \( k \)

• Nucleus Sampling

\[
P(y_t = w_i | y_{<t}, \mathbf{x}) = \frac{\exp(z_{t,i})}{\sum_{j \in V} \exp(z_{t,j})}
\]

\[
\hat{y}_t = \arg\max_{y_t} \left( \frac{\exp(z_{t,i})}{\sum_{j \in V} \exp(z_{t,j})} \right)
\]

\[
P'(y_t | y_{<t}) = \begin{cases} 
\frac{P(y_t | y_{<t})}{\sum_{y_k} P(y_{t} | y_{<t})} & \text{if } y_t \in V^k \\
0 & \text{otherwise}
\end{cases}
\]

\[
\sum_{y_t \in V^p} P(y_t | y_{<t}) \geq p
\]
Decoding Strategies: Combining random and greedy

- Random Sampling

  \[ P(y_t = w_i \mid y_{<t}, x) = \frac{\exp(z_{t,i})}{\sum_{j \in V} \exp(z_{t,j})} \]

  Problem: No probability pruning

- Greedy Decoding

  Combination of random and greedy sampling

  High k \implies Close to Random \implies Diverse generation \implies Dicey and non-monotonous

  Low k \implies Close to Greedy \implies Generic output \implies Safe with monotonous

- Nucleus Sampling

  \[ \sum_{y_t \in V^p} P(y_t \mid y_{<t}) > = p \]
Decoding Strategies: Nucleus

- **Random Sampling**
  
  Problem: No probability pruning

- **Greedy Decoding**
  
  Problem: No diversity
  No course correction

- **Top-k Sampling**
  
  Problem: Limiting to same $k$

- **Nucleus Sampling**

$$P(y_t = w_i | y_{<t}, \mathbf{x}) = \frac{\exp(z_{t,i})}{\sum_{j \in \mathcal{V}} \exp(z_{t,j})}$$

$$\hat{y}_t = \argmax_{y_i} \left( \frac{\exp(z_{t,i})}{\sum_{j \in \mathcal{V}} \exp(z_{t,j})} \right)$$

$$P'(y_t | y_{<t}) = \begin{cases} 
\frac{P(y_t | y_{<t})}{\sum_{v_k} P(y_t | y_{<t})} & \text{if } y_t \in V^k \\
0 & \text{otherwise}
\end{cases}$$

$$\sum_{y_t \in V^p} P(y_t | y_{<t}) \geq p$$
Decoding Strategies: Nucleus

- Rar
- Gre
- Top
- Nuc

Source: The Curious Case of Neural Text Degeneration
**Decoding Strategies: Beam Search**

Harpy Speech Recognition System first used Beam Search

**Problem:** Cannot go back in time and correct the mistakes as it is “spent”

- **Solution:** “Invest” in time instead of spending it

Επενδύστε στο χρόνο αντί να το ξοδέψετε

\[
\begin{align*}
  t=1: \text{Invest} & \quad \text{___} & \text{___} & \text{___} & \text{___} \\
  t=2: \text{Invest} & \quad \text{money} & \text{___} & \text{___} & \text{___} & \text{___} & \text{___}
\end{align*}
\]

**What investment?**
- Maintain a track of several \(b\) hypotheses at each time step.
- \(b\) is the beam size

**Tradeoff in performance**

- Grammaticality
- Relevance
Decoding Strategies: Beam Search

<START> → Invest → Make
Decoding Strategies: Beam Search
Decoding Strategies: Beam Search

<START> -> Invest
  -> Make
    -> money
      -> a
        -> in
          -> time
            -> stocks
              -> for
                -> to
                    -> time
Decoding Strategies: Beam Search

- Invest
  - in
  - money
  - a
  - time
  - for
  - to
  - get
- Make
  - for
  - spend

stocks
instead
Invest in stocks instead of spending money for a time to get instead of trying to spend money for them.

Decoding Strategies: Beam Search
Extensions to Beam Search

- NPAD Beam Search
- Beam Blocking
- Iterative Beam Search
- Diverse Beam Search
  - Hamming Distance
- Clustered Beam Search
- Clustering Post Decoding
Extensions to Beam Search

- NPAD Beam Search
- Beam Blocking
- Iterative Beam Search
- Diverse Beam Search
  - Hamming Distance
- Clustered Beam Search
- Clustering Post Decoding

- Inject noise to the transition function i.e to the hidden state

Extensions to Beam Search

- NPAD Beam Search
- Beam Blocking
  - Iterative Beam Search
  - Diverse Beam Search
    - Hamming Distance
  - Clustered Beam Search
  - Clustering Post Decoding
- Adjust the probability of a repeated n-gram to 0

Extensions to Beam Search

- NPAD Beam Search
- Beam Blocking
- **Iterative Beam Search**
- Diverse Beam Search
  - Hamming Distance
- Clustered Beam Search
- Clustering Post Decoding

- Goal: Increase search space without compromising run time.
- Run Beam Search multiple times.
- The next iteration of beam search avoids any partial hypotheses until that time step.
- Hard and Soft inclusion of hypothesis.

Extensions to Beam Search

- NPAD Beam Search
- Beam Blocking
- Iterative Beam Search
- Diverse Beam Search
  - Hamming Distance
- Clustered Beam Search
- Clustering Post Decoding

- Problem: Most decodings come from a very few highly valued beams.
- B beams at t and G groups
- Each group has $B' = B/G$ candidates.

$$\Delta(y_t, Y^g_t) = \sum_{b=1}^{B'} \delta(y_t, y^g_{b,t})$$

Extensions to Beam Search

- NPAD Beam Search
- Beam Blocking
- Iterative Beam Search
- Diverse Beam Search
  - Hamming Distance
- Clustered Beam Search
- Clustering Post Decoding

- Goal: shorten and prune unnecessary beams
- Get averaged representation of hypotheses till t.
- K-means clustering to get K clusters.
- Proceed with top B/K hypotheses from each K to t+1 step.

Extensions to Beam Search

- NPAD Beam Search
- Beam Blocking
- Iterative Beam Search
- Diverse Beam Search
  - Hamming Distance
- Clustered Beam Search
- Clustering Post Decoding

- Cluster sentences after decoding to identify similar sentences.
- Gives variation in candidates.
Tasks

Ethical Considerations

**Evaluation**
- Automatic
- Emulated
- Human

**Inference (decoding)**
- Latitude & Longitudinal
- Sampling & Search

**Core Modeling**
- Supervised
- Adversarial
- Reinforcement
- Generative Pretraining
Evaluation: Diversity in Formalization

- **Diversity in Formalization**
  - **Component** of other NLP tasks
  - **Diverse Input**: Though mainstream task, varied settings (tables, images, text)
  - **Open-ended Output**: Varying requirements of tasks
    - For instance, more nouns (entities), more pronouns (referring expressions)
- **Intrinsic and Extrinsic**
  - Each of which can be subjective or objective
Desiderata of Good Text:

- Relevant Content
- Appropriate Structure
  - Coherent
  - Ordering
- Suitable Surface Forms
- Fluent
- Grammatical
- Believable
- Novel
Evaluation: Factors from desiderata

- Fluency
- Relevance
- Naturalness
- Repetition
- Downstream Metric
- Truth_value
- Engaging
- Human_like
- Novelty
- Believability
- Genre/Style
Evaluation: Objective and Subjective

Objective
- Human_like
- Truth_value
- Repetition
- Relevance
- Genre/Style
- Downstream Metric

Subjective
- Engaging
- Fluency
- Naturalness
- Novelty
- Believability
Evaluation: Intrinsic and Extrinsic

Objective
- Human_like
- Truth_value
- Repetition

Subjective
- Engaging
- Fluency
- Naturalness
- Novelty
- Believability

Intrinsic

Extrinsic
- Relevance
- Genre/Style
- Downstream Metric
Evaluation

- **Automatic metrics**
  - Mostly based on word-overlap scores
    - **Adv**: Measures copy of words (extractive)
    - **Disadv**: No semantics
  - LM based evaluation metrics (perplexity)
    - **Adv**: Synthetically emulate fluency to some extent
    - **Disadv**: Inconsistent with word error rates or generation itself.
      Comparability of conditions (sensitive to data)
  - Embedding metrics
    - **Adv**: Captures semantics
    - **Disadv**: Difference in embedding spaces; not correlated to human judgements.
### Evaluation: Automatic Metrics

#### N-Gram based Evaluation

- **Precision over n-grams**
  - **BLEU**
- **More weight to rare n-grams**
  - **NIST**
- **Recall based on n-grams & longest subsequence**
  - **Rouge**
- **F1 equivalent from precision and recall**
  - **METEOR**
- **Tf-idf based cosine similarity for n-grams**
  - **CIDEr**

#### Explicit Content Evaluation

- **PYRAMID**
  - Summarization Content Units
- **SPICE**
  - Parsed scene graphs with objects and relations
Evaluation: Automatic Metrics Emulated

- **Emulate Targeted Evaluation: Examples**
  - Checking Fluency:
    - Scoring from a pre-trained language model
  - Checking Correctness or Loyalty:
    - Apply Inference
  - Checking Diversity:
    - Corpus based metrics on number of (distinct) entities
  - QAGS: [1]
    - Identify factual inconsistencies in summaries

Evaluation: Human Evaluation

- Subjective Evaluation - Mechanisms
  - Preference Testing
  - Scoring

- Problems:
  - Expensive for large scale or repeated examination
  - No standardized format or aspects
  - Nature of scoring (binary, discrete, continuous (nuanced))
  - Human scores/preferences are inconsistent/irregular
    - Biased with personal preferences
    - Along similar lines of inter-annotator (dis)agreement
  - Not scaled (varies across humans)
  - Ambiguity in instructions to elicit the right information
Evaluation: Human Evaluation

- Subjective Evaluation - Mechanisms
  - Preference Testing
  - Scoring

- Problems:
  - Expensive for large scale or repeated examination
  - No standardized format or aspects
  - Nature of scoring (binary, discrete, continuous (nuanced))
  - Human scores/preferences are inconsistent/irregular
  - Biased with personal preferences
  - Along similar lines of inter-annotator (dis)agreement
  - Not scaled (varies across humans)
  - Ambiguity in eliciting the right information

Best we got !!!

Problems merely to be taken as cautions while designing questions leading to less ambiguity more consistency
Evaluation Recipe

Train, debug until you are satisfied with readability etc.,

Then conduct subjective tests
Our tasks and models have consequences!

Tasks

Ethical Considerations

Evaluation

Automatic  Emulated  Human

Inference (decoding)

Latitude & Longitudinal  Sampling & Search

Core Modeling

Supervised  Adversarial  Reinforcement  Generative Pretraining
Ethical considerations

Gender Bias
Ethical considerations

Gender Bias

Gender neutral sentence

My friend is a doctor

Nā ūnīci cheppa ākāmaru
Ethical considerations

Gender Bias

Stereotyping

Warning: Next slide contains some offensive content.
The intention of the slide is ONLY to encourage work on mitigating stereotyping.
*Ethical considerations*

**Stereotyping**

**Talk to Transformer**

See how a modern neural network completes your text. Type a custom snippet or try one of the examples. This is a limited demo of InferKit.

**Custom prompt**

I do not trust Will because

**Completion**

I do not trust Will because he was cast as the bad guy in our little bad movie.

I do not trust Will because he has a thing for his special friend of mine, a girl who is only good for one thing.

And I do not trust Will because he has a thing for women that I am not and never will be.

**Talk to Transformer**

See how a modern neural network completes your text. Type a custom snippet or try one of the examples. This is a limited demo of InferKit.

**Custom prompt**

I do not trust Khan because

**Completion**

I do not trust Khan because his father was the owner of the World Trade Center in New York that was bombed on 9/11.

I know a lot of Muslims. They are peaceful. They love America. I do not like some of the things they believe in. I do not agree with some of the beliefs they hold, such as Muslims want to take over the world.
Ethical considerations

- Gender Bias
- Stereotyping
- Privacy and anonymity
Ethical considerations

Privacy and anonymity

Ethical considerations

- Gender Bias
- Stereotyping
- Privacy and anonymity
- Evaluating bias in generation
GEM Benchmark

The GEM Benchmark: Natural Language Generation, its Evaluation and Metrics:
Text-to-Text: Summarization

• **What?** Automatically create a concise summary
• **Why?** Information Overload: fast-growing textual information online
• **Shift:** From extractive to abstractive methods
• 2 paradigms of methodologies:
  • Extractive
  • Abstractive
Summarization: Paradigms

Extractive Summary
- Constrained by units of selection
- Usually minimal syntactic problems
- Incoherent discourse: Sentence reordering

Abstractive Summary
- Structure in discourse is data driven
- Flexible choice of words to fuse and avoid repetition
- Closer to how humans summarize
Summarization: Paradigms

Summarization datasets:

- **MLSum**
  - Large-scale multilingual dataset for evaluating abstractive summarization

- **XSum**
  - Large scale monolingual dataset for evaluating extreme summarization.

- **WikiLingua**
  - Large-scale multilingual dataset for evaluating cross-lingual abstractive summarization

Simplification datasets:

- **Wiki-Auto**
  - Wiki-Auto is the largest open text simplification dataset currently available.

- **TURKCorpus**
  - TURKCorpus is a high-quality simplification dataset where each source sentence is associated with 8 human-written simplifications.

- **ASSET**
  - ASSET is a high quality simplification dataset where each source (not simple) sentence is associated with 10 human-written simplifications.

https://gem-benchmark.com/
Summarization: Techniques

- Information Extraction -> Taken care of encoder
  - Challenge: Encoding important information from long documents
    - Explicit content selection network
    - Hybridized word and character level models
    - Copy Mechanism
    - Distraction based Techniques
Summarization: Content Selection

- Information Extraction -> Taken care of encoder
  - Challenge: Encoding important information from long documents
    - Explicit content selection network [1] [2]
    - Hybridized word and character level models
    - Copy Mechanism
    - Distraction based Techniques


Summarization: Word-Character

- Information Extraction -> Taken care of encoder
  - Challenge: Encoding important information from long documents
    - Explicit content selection network
    - **Hybridized word and character level models** [3]
  - Copy Mechanism
  - Distraction based Techniques

**Summarization: Copy**

- Information Extraction -> Taken care of encoder
  - Challenge: Encoding important information from long documents
    - Explicit content selection network
    - Hybridized word and character level models
    - **Copy Mechanism** [4] [5]
    - Distraction based Techniques


Summarization: Distraction

- Information Extraction -> Taken care of encoder
  - Challenge: Encoding important information from long documents
    - Explicit content selection network
    - Hybridized word and character level models
    - Copy Mechanism
    - Distraction based Techniques [6]

\[
c_t = \tanh(W_c c'_t - U_c \sum_{j=1}^{t-1} c_j)
\]

Image-to-Text: Image Captioning

- **What?**
  - Generate textual description of an image
  - Static input with a sequential output
- **Why?**
  - Real world contexts includes visual input
- **Goal:**
  - Synergy between vision and language modeling techniques
Image Captioning: Datasets (1/3)

- **Pascal 1k** [1]
  - 1000 images from Pascal Object Recognition Dataset
  - 5 captions per image

- **VLT 2k: Visual and Linguistic Treebank** [2]
  - Derived from Pascal Action Recognition dataset
  - 3 captions per image

- **Flickr30k** [4]
  - Earlier Version: Flickr8k [3]
  - 5 captions per image

- **Abstract Scenes** [5]
  - 10k clip-art images
  - 6 captions per image


Image Captioning: Datasets (2/3)

- **IAPR-TC12** [6]
  - Images retrieved from search engines - 20k
  - Multilingual annotations (in English, German, Spanish)

- **MS COCO** [7]
  - Annotated with object categories as well -120k
  - 5 captions per image

- **BBC News Captions** [8]
  - 3k images

---


Image Captioning: Datasets (3/3)

- **SBU 1M Captions** [9]
  - Sourced from Flickr: 1 million
  - Noun and Verb

- **Deja Image Captions** [10]
  - 4 million pairs with 180k almost similar captions

- **Conceptual Captions** [11]
  - 3.3 million pairs with Image, text based filtering and text transformation


Image Captioning: Techniques

- Challenges:
  - Image Understanding
  - Partial Detection
  - Complete Detection
- Encoding
  - Visual encoding
    - Attention
    - Semantic concepts
  - Multimodal encoding
    - Co-attention
- Dense Captioning
Image Captioning: Partial Detection

- Challenges:
  - Image Understanding
  - **Partial Detection** [1]
  - Complete Detection
- Encoding
  - Visual encoding
    - Attention
  - Semantic concepts
  - Multimodal encoding
    - Co-attention
- Dense Captioning

\[ \mathcal{L} = \mathcal{L}_CM + \mathcal{L}_IM + \mathcal{L}_LM \]

Image Captioning: Techniques

- Challenges:
  - Image Understanding
    - Partial Detection
  - Complete Detection
- Encoding
  - Visual encoding
    - Attention
    - Semantic concepts
  - Multimodal encoding
    - Co-attention
- Dense Captioning
Image Captioning: Attention

- Challenges:
  - Image Understanding
  - Partial Detection
  - Complete Detection
  - Encoding
    - Visual encoding
      - Attention [3a]
    - Semantic concepts
  - Multimodal encoding
    - Co-attention
  - Dense Captioning

**Image Captioning: Adaptive Attention**

- Challenges:
  - Image Understanding
  - Partial Detection
  - Complete Detection
  - Encoding
    - Visual encoding
      - **Attention** [3b]
    - Semantic concepts
    - Multimodal encoding
      - Co-attention
  - Dense Captioning

$$s_t = g_t \odot \tanh(m_t)$$

$$\hat{c}_t = \beta_t s_t + (1 - \beta_t)c_t$$

**Image Captioning: Semantic Concepts**

- Challenges:
  - Image Understanding
  - Partial Detection
  - Complete Detection
  - Encoding
    - Visual encoding
      - Attention
        - **Semantic concepts** [4a] [4b]
    - Multimodal encoding
      - Co-attention
  - Dense Captioning

---


**Image Captioning: Co-attention**

- **Challenges:**
  - Image Understanding
  - Partial Detection
  - Complete Detection

- **Encoding**
  - Visual encoding
    - Attention
    - Semantic concepts
  - Multimodal encoding
    - **Co-attention** [5]

- **Dense Captioning**

Image Captioning: Dense

- Challenges:
  - Image Understanding
  - Partial Detection
  - Complete Detection
  - Encoding
    - Visual encoding
      - Attention
      - Semantic concepts
    - Multimodal encoding
      - Co-attention
- Dense Captioning [6]

Image-to-Text: Challenges

- 0-shot and few-shot learning [1a] [1b]
  - To transfer from known images to caption unseen new images
- Linguistic intuitions for cross modal attention
- Neural Module Networks [2]
- Temporal dependencies
- Explanation Generation

Structured data-to-text

• What?
  • Convert a set of records into natural language description

• Why?
  • Often easier to record tabular data while it is easier to read or comprehend natural language sentences

• Tasks:
  • Weather Report Generation from measurements
  • Biography Generation
  • Sports event description generation
Structured data-to-text

- WEATHERGOV [1]:
  - 200k scenarios
  - Each has 36 weather records
- ROBOCUP [2]:
  - 1539 events
  - Temporally ordered
- BioGen [3]:
  - 700k biographies
- E2E [4]:
  - Restaurant domain
- ROTOWIRE and SBNation [5]:
  - NBA Basketball Game Summaries

Structured data-to-text: Some differences from text-to-text

- Underlying principle is still seq2seq modeling (usually with attention)
- Difference from text-to-text:
  - **Record Encoder**: Encode a sequence of records
    - No hard constrained ordering of sequence of records
    - Usually done with LSTMs or transformers in arbitrary orders
    - Place according to corpus based entries
  - Mostly **copy** values rather than attributes
  - **Information Extraction** Approaches to map records and entities/actions
Structured data-to-text: Copy Mechanism

- Copy mechanism on the value portion only
  - Joint Copy Model
  - Conditional Copy Model
- Induced Metrics:
  - Extractive Eval: Extract (Entity, Value) pairs and predict Type of relation
    - Content Selection: P and R of unique relations
    - Relation Generation: P and Num of unique relations from Y in S
  - Content Ordering: Distance between extracted records in Y and Y’

Structured data-to-text: Decomposition

- Decompose into 2 phases
  - Record Selection and Planning
- Content Selection
  - Why? Emphasis and description of correlated records
  - How?
    - Attention on the records
    - Gating Mechanism: Extent of information to flow from each record
- Planning
  - Information Extraction Techniques: Map entities in description -> Record
  - Sequence of pointers with pointer network
- **Common Finding:** Conditional Copy better!

Speed in Real Time

- **Problem**: Real time Generation
  - Of course, better hardware
  - Translation in real time [1]
    - Quality: BLEU
    - Delay: Length Proportions and Wait times
- Tricks: Pruning in beams
- Caching previous computations
- Distillation
- Quantization
  - Sharing weights from a bin
- Pruning
  - Remove connections

Some Neat Practical Tricks:

- Temperature: Divide the scores with $\tau$
  - High T => close to decoding from uniform distribution
    - Increases Diversity
  - Low T => close to a jagged distribution
    - Generic/Repetitive
- Decoding
  - Cheaper to adjust the decoding algorithm
  - Increase the probability of rare words (diversity)
- Out of vocabulary words
  - Unkify to reduce output space per time step
    - Problem: Incomprehensible sometimes
  - Better solution: Use sub-word level tokens
    - Problem: Increases model parameters and train time
- Manually check the outputs
Mix cocoa powder and heavy cream
Mix cocoa powder and heavy cream

Glaze chicken with chocolate sauce
Cook the enchilada sauce on medium.
Remove the chicken pieces after 4 hours and strain the sauce into a pot.
Dice your onion and put it on top of the chicken.
Cook the enchilada sauce on medium.
Remove the chicken pieces after 4 hours and strain the sauce into a pot.
Dice your onion and put it on top of the chicken.
Dice your onion and put it on top of the chicken.

Cook the enchilada sauce on medium.

Remove the chicken pieces after 4 hours and strain the sauce into a pot.
Dice your onion and put it on top of the chicken.
Cook the enchilada sauce on medium.
Remove the chicken pieces after 4 hours and strain the sauce into a pot.
Looking into the Future….

Out of the numerous possibilities, a few directions the future of NLG holds:

• Generation in real time
• Non-autoregressive decoding
• Situated contexts
  • Dynamic novel content generation
  • Real or virtual environments, games
• Consistency in personality and being opinionated (for agents)
• Conditioning on multiple modalities
• Better metrics (hopefully correlated with human judgements)
• Creative Text Generation:
  • Entertaining text captures reader’s attention (sarcasm, jokes, metaphors)
  • Evaluating Creativity/Funniness ???

If Tony Stark and Bucky Barnes started a business, they would name it StarBucks
Thank You!!!

E-mail: kchandu@cs.cmu.edu