# Research on Text Generation Systems

**Hwanhee Lee**

Chung-Ang University

# Outline

## Introduction

## Part 1

**Evaluating Text Generation Systems**

- How can we evaluate Factual Consistency? (NAACL 2022)

## Part 2

**Controlling Text Generation Systems**

- How can we control language model? (ACL 2023)

## Part 3

**Data Augmentation with Text Generation Systems**

- How can we generate new dataset using language model?

# Short Bio

Introduction

## Education & Employment

**Seoul National University**
Electrical & Computer Engineering

| B.S. | Ph.D. | Postdoc. |
|---|---|---|
| Feb 2017 | Aug 2022 | ~ Feb 2023 |

**Chung-Ang University**
Dept. of Artificial Intelligence

Assistant Professor
Mar 2023 ~

**Language Intelligence Lab (LILAB)**

- https://sites.google.com/view/cau-li

## Research Interests: Natural Language Processing (NLP)

Published at ***NLP*** conferences
***ACL, NAACL***, and ***EMNLP***

- ***Natural Language Generation & Evaluation***

- *Summarization, Dialog System, Factuality Checking & Improvement, Large Language Model*

# Research Interests

## Introduction

### NLP

**N**atural **L**anguage **P**rocessing

### NLU

**N**atural **L**anguage **U**nderstanding

- Machine Reading Comprehension
- Sentiment Analysis

...

**Passage Sentence**

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity.

**Question**

What causes precipitation to fall?

**Answer Candidate**

gravity

**Sentence**: "I loved the movie, it was amazing!"

**Sentiment** : Positive

### NLG

**N**atural **L**anguage **G**eneration

- Abstractive Summarization
- Machine Translation
- Dialogue Generation

...

# Research Interests: Text Generation

## Introduction

**Article**

Scientists from harvard have discovered a way of turning stem cells into killing machines to fight brain cancer. (…)

**Summarization**

**Image Captioning**

**Summary**

Scientists in the US have developed a stem cell therapy for brain tumours.

*unimodal*

**Caption**

A blue subway train pulls into the subway station.

*multimodal (text+image)*

**Research Question** How can we develop a better text generation system?

# Outline

## Introduction

## Part 1

**Evaluating Text Generation Systems**

- How can we evaluate Factual Consistency? (NAACL 2022)

## Part 2

**Controlling Text Generation Systems**

- How can we control language model? (ACL 2023)

## Part 3

**Data Augmentation with Text Generation Systems**

- How can we generate new dataset using language model?

# What is Important in Text Generation?

Part 1: Evaluating Text Generation Systems

**Grammatically Correct**

**Fluency**

**Interesting**

**Understandable**

# Factual Consistency

Part 1: Evaluating Text Generation Systems

In NAACL 2022

generated content should be factually consistent with the input information

Although the generated text is fluent,

if there is a minor **factual error**,

the text is totally wrong in text generation task

**Caption**

A **red** train pulls into the train station

*=> Factually Inconsistent!*

To develop a better text generation system, we must resolve factual inconsistency!

# Evaluating Text Generation Systems



In NAACL 2022

Human evaluation is too expensive.

: measuring the overlap with human references => Easiest and widely used way

**Reference**

A **blue** train pulls into the train station.

**Machine Generated Text**

A **red** train pulls into the train station.

**Similarity**

**N(uni)-gram Precision (co-occurrence)**

7 / 8 ~= 0.88 (BLEU-1 Score)

=> Is this score reasonable?

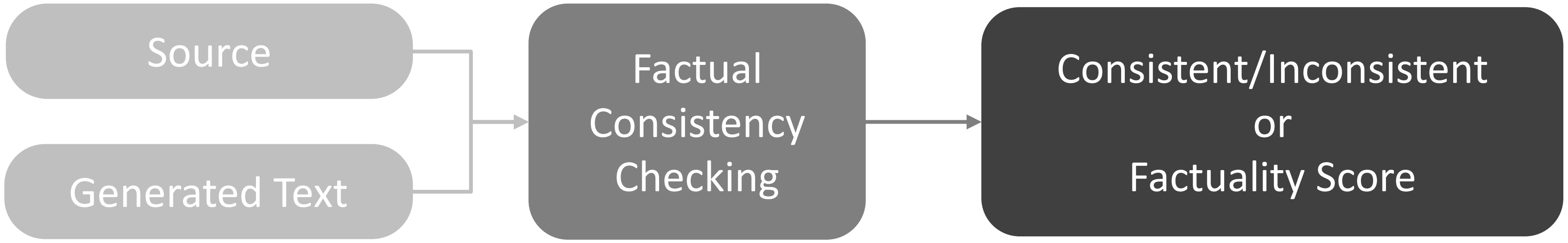
# Developing Factual Consistency Metric

Part 1: Evaluating Text Generation Systems

In NAACL 2022

**Goal** Developing an evaluation metric for text generation systems that focuses on **"factual consistency"** with the source

=> Higher correlation with human judgments



# Data Augmentation for Factual Consistency Evaluation

Part 1: Evaluating Text Generation Systems

In NAACL 2022

### Function of Factuality Metric for Text Generation Systems

Source

Consistent Text $x$

A **blue** subway train pulls into the subway station.

Easily Available

Metric → $S_x$

Metric → $S_{\hat{x}}$

Ranking Loss

$S_x \succ S_{\hat{x}}$

Inconsistent Text $\hat{x}$

A **red** subway train pulls into the subway station.

We can train a metric using ***enough consistent and inconsistent samples*** for each task.

Source, Text → Metric → Score

**Research Question** How can we generate inconsistent texts?

# Data Augmentation for Factual Consistency Evaluation



NAACL

In NAACL 2022

## Prior Work: Substitution & Mask-and-Fill to generate inconsistent texts (summaries)

- Substitution

**entity**
(Name of person or organization)

**Original**

Charlton coach Guy Luzon had said on Monday: 'Alou Diarra is training with us.'

**Corrupted**

Charlton coach Bordeaux had said on Monday: 'Alou Diarra is training with us.'

- Mask-and-Fill

**"Masking"**

[redacted] has reportedly dropped a criminal investigation into three businessmen who had been accused of conspiring to make corrupt payments to secure contracts in [redacted] The SFO launched an investigation into [redacted] in May 2017 as part of a wider probe into [redacted]-based oil consultancy [redacted]

**"Filling"**

Unaoil has reportedly dropped a criminal investigation into three businessmen who had been accused of conspiring to make corrupt payments to secure contracts in Monaco . The SFO launched an investigation into Monaco in May 2017 as part of a wider probe into Petrofac-based oil consultancy The Serious Fraud Office.

Kryscinski et al., Evaluating the Factual Consistency of Abstractive Text Summarization, EMNLP 2020

# Data Augmentation for Factual Consistency Evaluation

Part 1: Evaluating Text Generation Systems

In NAACL 2022

### Limitation of Rule-Based Substitution and Mask-and-Fill

**Article:** Guus Hiddink, the Russia and Chelsea coach, has had much to smile about in his 22-year managerial career. ,. . . , Enjoying success around the world – at different levels with different players in different cultures – has made Guus Hiddink one of the most admired bosses

...

is loyal to the project he has in charge of the Russian national side and insists he will leave Chelsea at the end of the season regardless.

**Reference Summary:** Born in 1946, Hiddink has become one of the best managers in the world . Dutchman has enjoyed huge success at club and international level. He's currently coach of Russia and is in charge of Chelsea until end of the season.

Fill in the mask
or
Substitution

**Generated Inconsistent Summary**

Born in 1946, *Dutchman* has become one of *the most respected politicians* in the world. Dutchman is enjoyed *success at the Olympics and World Cup*. He's currently the *President of Russia* and is in charge of *the country* until the end of the season.

**Coach -> President of Russia**

- Too different from the original summary
- Irrelevant to article

Lee et al., Masked Summarization to Generate Factually Inconsistent Summaries for Improved Factual Consistency Checking, Findings of NAACL 2022

# Data Augmentation for Factual Consistency Evaluation

Part 1: Evaluating Text Generation Systems

In NAACL 2022

## Using Masked Context (Masked Article)

**Article:** , , has had much to smile about in his 22-year managerial career. ,. . . , Enjoying around – at with different players in – has made one of the most admired bosses

...

is loyal to the project he has in charge of the Russian national side and insists he will leave at the regardless.

**Reference Summary:** Born in 1946, Hiddink has become one of the best managers in the world . Dutchman has enjoyed huge success at club and international level. He's currently coach of Russia and is in charge of Chelsea until end of the season.

Fill in the mask additionally using *"Masked Article"*

**Generated Inconsistent Summary**

Born in 1946, *Hiddink* has become one of *the most admired managers* in the world. Dutchman has enjoyed *successful spells* at *Chelsea and Real Madrid*. He's currently *manager of Russia* and is in charge of *the country* until the end of the season.

- Relevant to article
- More natural, but still inconsistent

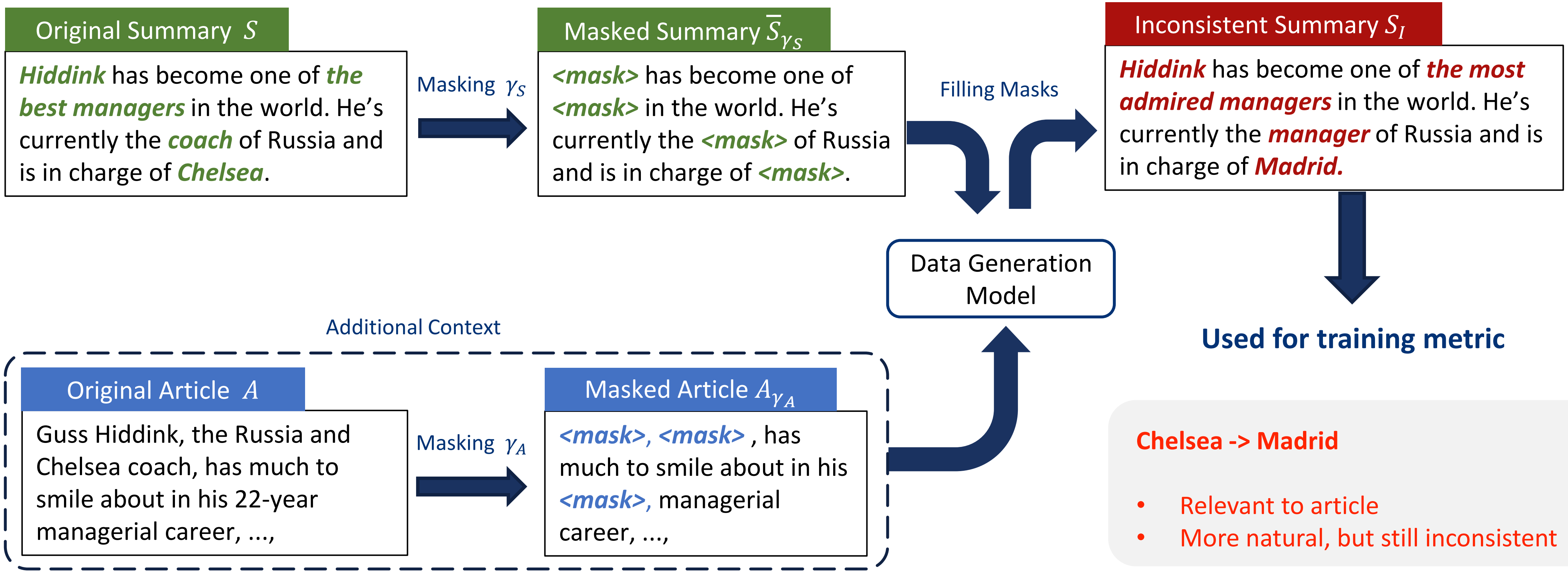
Lee et al., Masked Summarization to Generate Factually Inconsistent Summaries for Improved Factual Consistency Checking, Findings of NAACL 2022

# Data Augmentation for Factual Consistency Evaluation

Part 1: Evaluating Text Generation Systems

In NAACL 2022
Abstractive Summarization

## Filling the masks using both the masked summary and masked article.




**Chelsea -> Madrid**

- Relevant to article
- More natural, but still inconsistent

Lee et al., Masked Summarization to Generate Factually Inconsistent Summaries for Improved Factual Consistency Checking, Findings of NAACL 2022

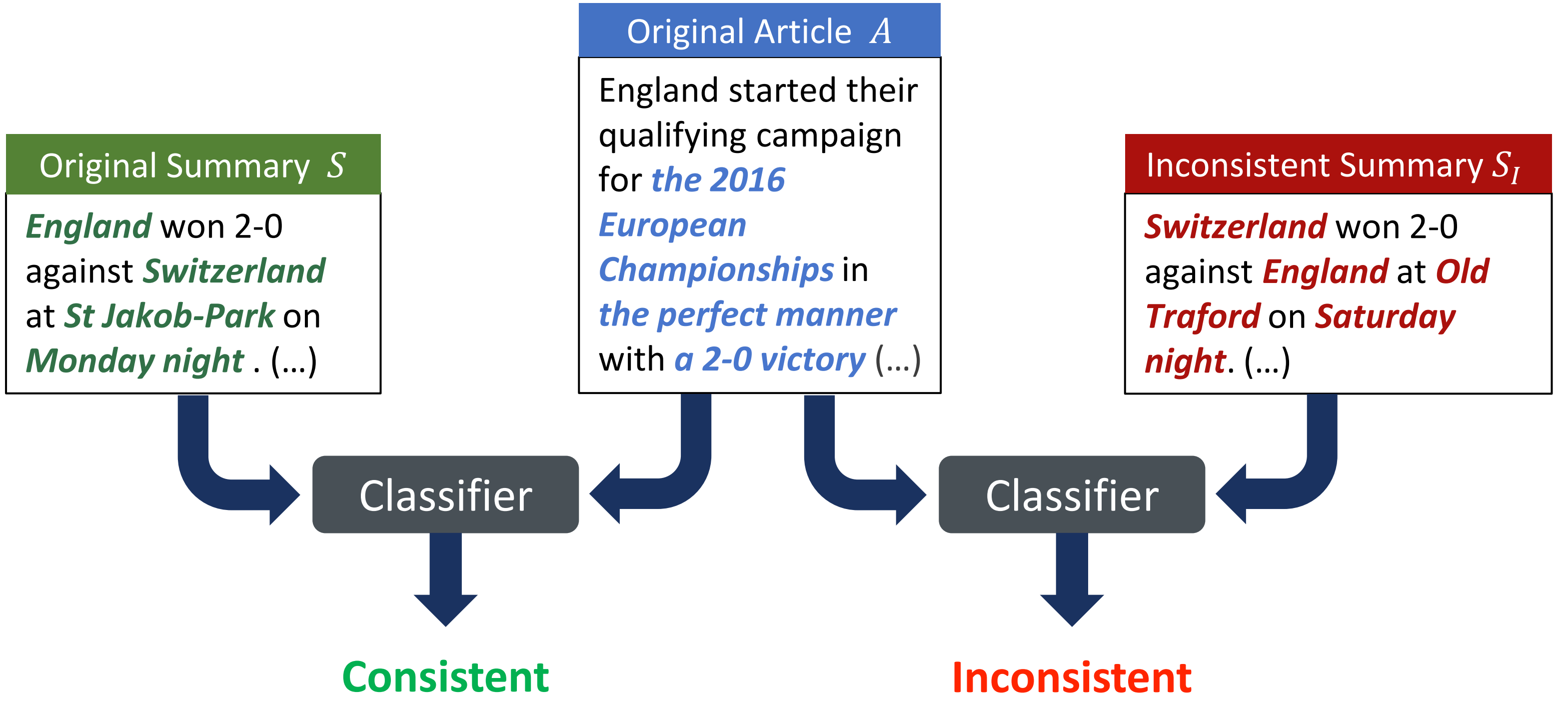
# Data Augmentation for Factual Consistency Evaluation



In NAACL 2022

## Model Based Data Augmentation Methods: Mask-and-Fill with Masked Article (MFMA)

We train a classifier of consistent summaries and inconsistent summaries.




Lee et al., Masked Summarization to Generate Factually Inconsistent Summaries for Improved Factual Consistency Checking, Findings of NAACL 2022

# Data Augmentation for Factual Consistency Evaluation

Part 1: Evaluating Text Generation Systems

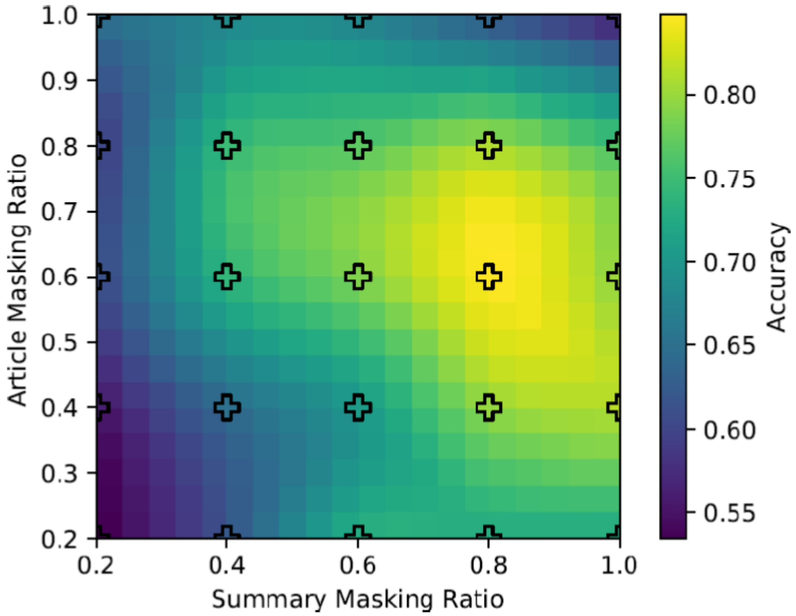
In NAACL 2022
Abstractive Summarization

## Qualitative Results

: Is the label similar to human's?


| Dataset | CNN/DM | XSum |
|---|---|---|
| Metric | F1 | F1 |
| *Baselines* | | |
| FactCC | 67.4 | 55.5 |
| DocNLI | 66.8 | 60.2 |
| MNLI | 51.4 | 35.8 |
| FEVER | 49.9 | 56.7 |
| MF | 59.5 | 54.6 |
| *Ours* | | |
| MFMA | **72.8** | **60.6** |

## Performance among Masked Ratio




We can infer that there is an optimal masking ratio.

Lee et al., Masked Summarization to Generate Factually Inconsistent Summaries for Improved Factual Consistency Checking, Findings of NAACL 2022

# Outline

## Introduction

## Part 1

**Evaluating Text Generation Systems**

- How can we evaluate Factual Consistency? (NAACL 2022)

## Part 2

**Controlling Text Generation Systems**

- How can we control language model? (ACL 2023)

## Part 3

**Data Augmentation with Text Generation Systems**

- How can we generate new dataset using language model?

# Controlling Text Generation Systems



In ACL 2023

How can we control text generation system for specific attribute?

- **Controlled Text Generation**: whether the *generated content* is on *desired attribute* (i.e. Topic, Sentiment) $p(x_t \mid x_{<t}, a), a : attribute$

[*Music*] Emphasised are

[*Foods*] The issue focused on

Attribute + Prompt → Text Generation System → Generated Text

[*Music*] Emphasised are the words "instrument" in the title. The song is a cover of"I'm a Man" by the band The Beatles.

[*Foods*] The issue focused on the use of the term "organic" in the food industry on a new USDA regulation.

**Research Question** How can we control text generation system?

Kim et al., Critic-Guided Decoding for Controlled Text Generation,, Findings of ACL 2023

# Controlled Text Generation with Two Ways

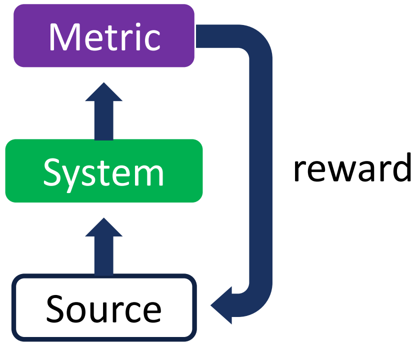
Part 2: Controlling Text Generation Systems

In ACL 2023

- **Controlled Text Generation**: whether the *generated content* is on *desired attribute* (i.e. Topic, Sentiment) $p(x_t | x_{<t}, a), a : attribute$

**Prior work: 1) Reinforcement Learning (RL)**

+: Directly optimize any task-specific metrics -> **Outstanding Score!**
-: hard for convergence and unstable training




Kim et al., Critic-Guided Decoding for Controlled Text Generation,, Findings of ACL 2023

# Controlled Text Generation with Two Ways

Part 2: Controlling Text Generation Systems

In ACL 2023

- **Controlled Text Generation**: whether the *generated content* is on *desired attribute* (i.e. Topic, Sentiment) $p(x_t|x_{<t}, a), a : attribute$

**Prior work: 2) Weighted Decoding** $\boldsymbol{p(x|a) \propto p(a|x)p(x)}$

Then, is $p(x)$ ***uncontrolled*** language model, and $p(a|x)$ is ***classification model***

+: Plug-and-Play for any Language Models
+: Stable Training
- : Lower score than RL
- : Lower text quality than RL

*How to mix advantages of RL and Weighted Decoding?*

Kim et al., Critic-Guided Decoding for Controlled Text Generation,, Findings of ACL 2023

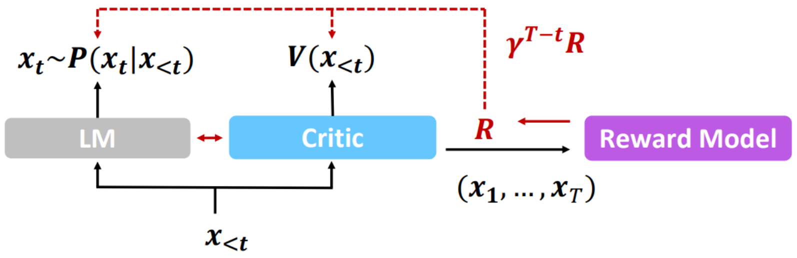
# Critic-Guided Decoding (CriticControl)



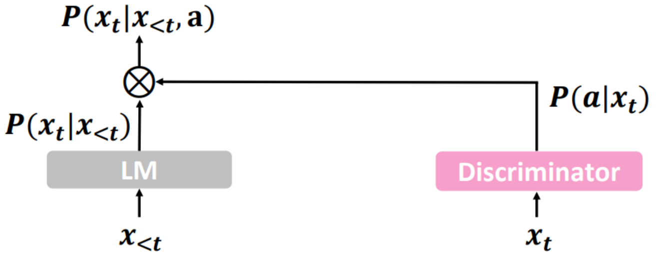
Association for Computational Linguistics

In ACL 2023

| | Pros | Cons |
|---|---|---|
| RL | Powerful Control | Unstable Training |
| WD | Stable Training for all LMs | Less powerful control than RL |




(a) Reinforcement Learning




(b) Weighted Decoding

- Critic Predicts $p(reward\,|\,x)$ in the *view of LM*
- Actor optimize to win Critic

⇨ Unstable training

- Training $p(a\,|\,x)$ is easy, and the LM is frozen
- However, $p(a\,|\,x)$ is outside the LM

⇨ Less powerful control and text quality

*What If Weighted Decoding Guided by Critic's Prediction $p(a\,|\,x)$ ?*

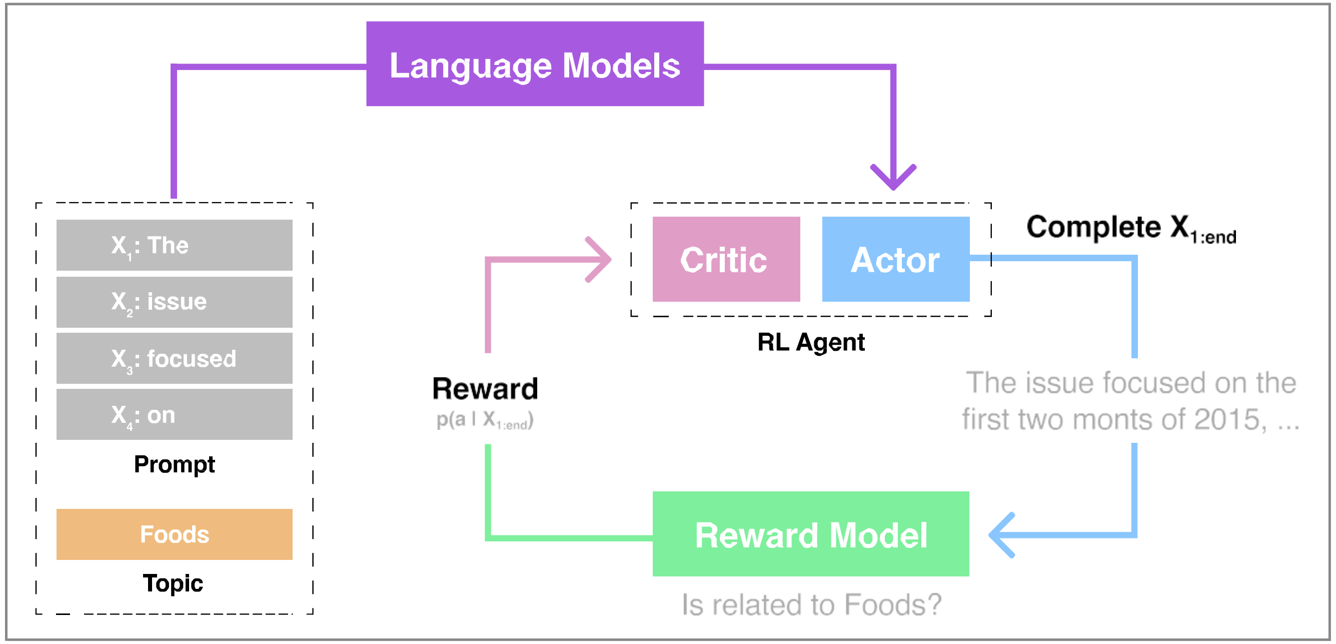
Kim et al., Critic-Guided Decoding for Controlled Text Generation,, Findings of ACL 2023

# CriticControl - Training

Part 2: Controlling Text Generation Systems

Association for Computational Linguistics

In ACL 2023




**Training**

**Goal**: training *Critic* to predict attribute-relevance of future completed texts

1) Give input with desired attribute token: `[Music] The issue focused on the`
2) Freeze LM (Actor), simulate *on-line* the input, and get reward as final results $p(a \mid x_{complete})$
3) Training only Critic to predict *future full text* with $\mathcal{L}_{critic} = \sum_{t=1}^{end} \left( \sum_{i=0}^{end-t} (\gamma\lambda)^i \delta_{t+i} \right)^2$

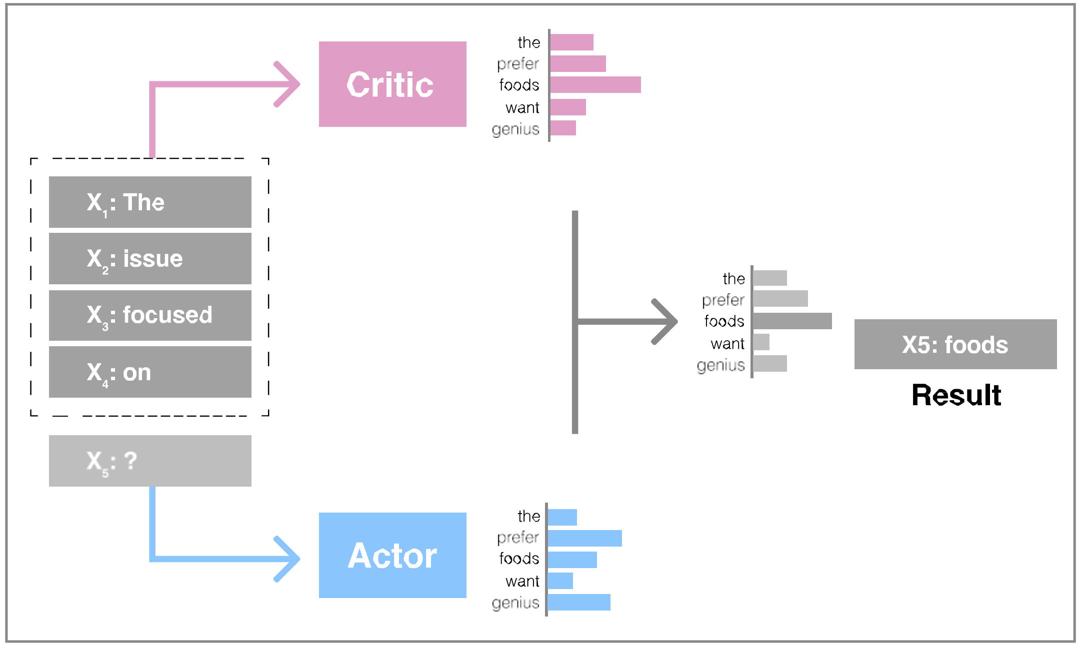
Kim et al., Critic-Guided Decoding for Controlled Text Generation,, Findings of ACL 2023

# CriticControl - Inference

Part 2: Controlling Text Generation Systems

Association for Computational Linguistics

In ACL 2023




**Inference**

**Goal**: Control decoding procedure to desired attribute

1) Give input with desired attribute token: *[Music] The issue focused on the*

2) Shift stepwise distribution computed by frozen LM (Actor) $P(x_t|x_{<t}, a) = \dfrac{P(a|x_{\leq t})}{P(a|x_{<t})} P(x_t|x_{<t})$

* $P(x_t|x_{<t})$ is text generation of frozen LM, $P(a|x)$ is from Critic, and $P(x_t|x_{<t}, a)$ is desired text generation

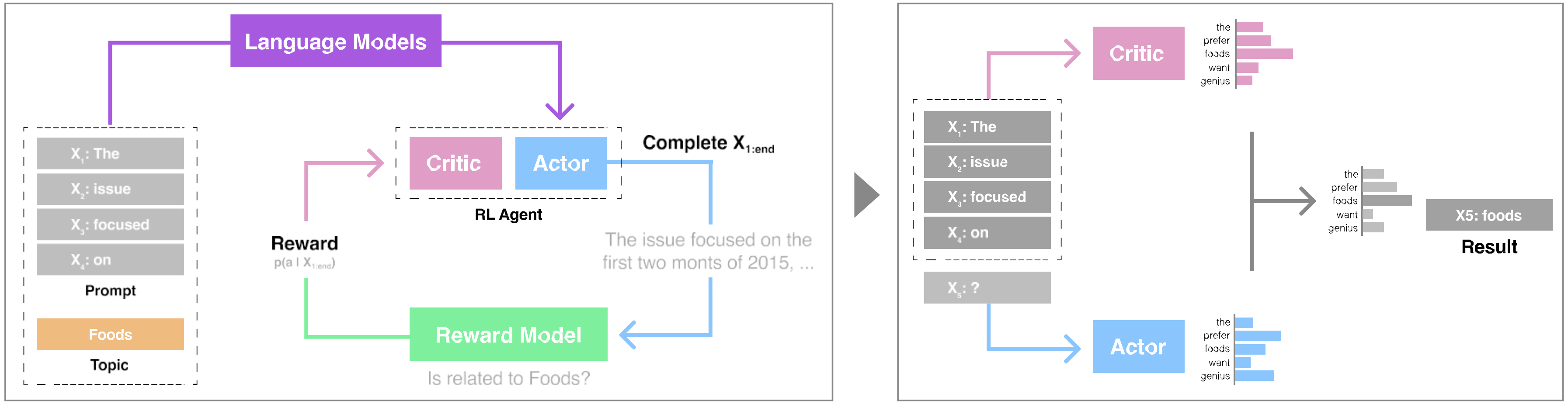
Kim et al., Critic-Guided Decoding for Controlled Text Generation,, Findings of ACL 2023

# CriticControl - Examples

Part 2: Controlling Text Generation Systems

Association for Computational Linguistics

In ACL 2023




**Training** **Inference**

[Foods] An illustration of the food of the ancient Egyptians. The Egyptians were the first to use the term "f ood" to describe the food of their gods. The Egyptians believed that food was the source of life and that it was the food of gods.

[Sports] Prior to this season, the Panthers had never won a playoff game. The Panthers have wo n three straight, including a win over the New York Giants in the NFC Championship Game. They are 2-0 in the playoffs. Coach Ron Rivera said the Panthers are "very confident" in their ability to win the Sup er Bowl. "We're going to be ready to go," Rivera saids!

Kim et al., Critic-Guided Decoding for Controlled Text Generation,, Findings of ACL 2023

# Experiments Results

Part 2: Controlling Text Generation Systems

Association for Computational Linguistics

In ACL 2023

- Topic Control Automatic Evaluation

| Model | Success | Fluency | | Diversity | | |
|---|---|---|---|---|---|---|
| | On-Topic | Perplexity ↓ | Grammar | Dist-1 | Dist-2 | Dist-3 |
| GPT-2-medium (Radford et al., 2019) | 0.16 | **14.06** | 0.74 | 0.29 | 0.70 | 0.88 |
| WDEC (Yang and Klein, 2021) | 0.49 | 67.53 | 0.59 | 0.16 | 0.42 | 0.85 |
| PPLM (Dathathri et al., 2019) | 0.45 | 62.66 | 0.78 | 0.35 | **0.78** | **0.92** |
| FUDGE (Yang and Klein, 2021) | 0.78 | 69.08 | 0.79 | 0.34 | 0.75 | 0.91 |
| CriticControl | **0.89** | 17.19 | **0.83** | **0.49** | 0.76 | 0.90 |
| CriticControl - small | 0.85 | 16.88 | 0.83 | 0.47 | 0.73 | 0.89 |
| CriticControl - large | 0.92 | 17.58 | 0.84 | 0.51 | 0.77 | 0.91 |
| CriticControl - XL | **0.94** | 17.69 | 0.83 | 0.51 | 0.77 | 0.91 |
| CriticControl - Zero shot | **0.73** | 17.55 | 0.85 | 0.49 | 0.76 | 0.90 |

- Sentiment Control Automatic Evaluation

| Model | Success | Fluency | | Diversity | | |
|---|---|---|---|---|---|---|
| | Positiveness | Perplexity ↓ | Grammar | Dist-1 | Dist-2 | Dist-3 |
| GPT-2-medium (Radford et al., 2019) | 0.57 | **11.91** | 0.78 | 0.25 | 0.63 | 0.78 |
| PPLM (Dathathri et al., 2019) | 0.60 | 142.11 | 0.73 | 0.22 | 0.61 | 0.72 |
| CC-LM (Krause et al., 2020) | 0.76 | 15.79 | 0.72 | 0.28 | 0.70 | 0.82 |
| GeDi (Krause et al., 2020) | 0.84 | 38.94 | 0.76 | 0.27 | 0.77 | 0.89 |
| CriticControl | **0.90** | 12.97 | **0.87** | **0.31** | **0.84** | **0.92** |
| PPO | 0.94 | 13.43 | 0.84 | 0.32 | 0.86 | 0.93 |
| PPO - CriticControl | **0.99** | 13.44 | 0.80 | 0.32 | 0.85 | 0.93 |

- CriticControl generate high quality texts related to attributes
- CriticControl can achieve zero-shot control on unseen topics
- CriticControl is also compatible with RL

Kim et al., Critic-Guided Decoding for Controlled Text Generation,, Findings of ACL 2023

# Human Evaluation

## Part 2: Controlling Text Generation Systems

Association for Computational Linguistics

In ACL 2023

- Topic Control

| | CriticControl | Equal | GPT2 |
|---|---|---|---|
| Success | 0.59 | 0.23 | 0.18 |
| Fluency | 0.36 | 0.35 | 0.29 |

| | CriticControl | Equal | FUDGE |
|---|---|---|---|
| Success | 0.53 | 0.15 | 0.32 |
| Fluency | 0.46 | 0.2 | 0.34 |

- Sentiment Control

| | CriticControl | Equal | GPT2 |
|---|---|---|---|
| Positiveness | 0.41 | 0.23 | 0.36 |
| Fluency | 0.38 | 0.18 | 0.44 |

| | CriticControl | Equal | GeDi |
|---|---|---|---|
| Positiveness | 0.35 | 0.30 | 0.35 |
| Fluency | 0.42 | 0.26 | 0.32 |

- Detoxification

| | CriticControl | Equal | GPT2 |
|---|---|---|---|
| Less Toxic | 0.45 | 0.17 | 0.37 |
| Fluency | 0.39 | 0.24 | 0.37 |

| | CriticControl | Equal | DExperts |
|---|---|---|---|
| Less Toxic | 0.50 | 0.14 | 0.35 |
| Fluency | 0.48 | 0.22 | 0.30 |

- Human preferences result also collaborates our findings
- Overall, the text quality is relatively great rather than previous works.

Kim et al., Critic-Guided Decoding for Controlled Text Generation,, Findings of ACL 2023

# Outline

## Introduction

## Part 1

**Evaluating Text Generation Systems**

- How can we evaluate Factual Consistency? (NAACL 2022)

## Part 2

**Controlling Text Generation Systems**

- How can we control language model? (ACL 2023)

## Part 3

**Data Augmentation with Text Generation Systems**

- How can we generate new dataset using language model?

# Versatility of Large Language Models

Part 3: Data Augmentation with Text Generation Systems

- Large Language Models (LLM, e.g ChatGPT, GPT-4) can solve various tasks without training.

## Machine Translation

H

Translate the following sentence in hungarian.

Thank you for inviting me to ELTE.

Köszönöm, hogy meghívtál az ELTE-re.

## Summarization

H

Please summarize the following article in one sentence.

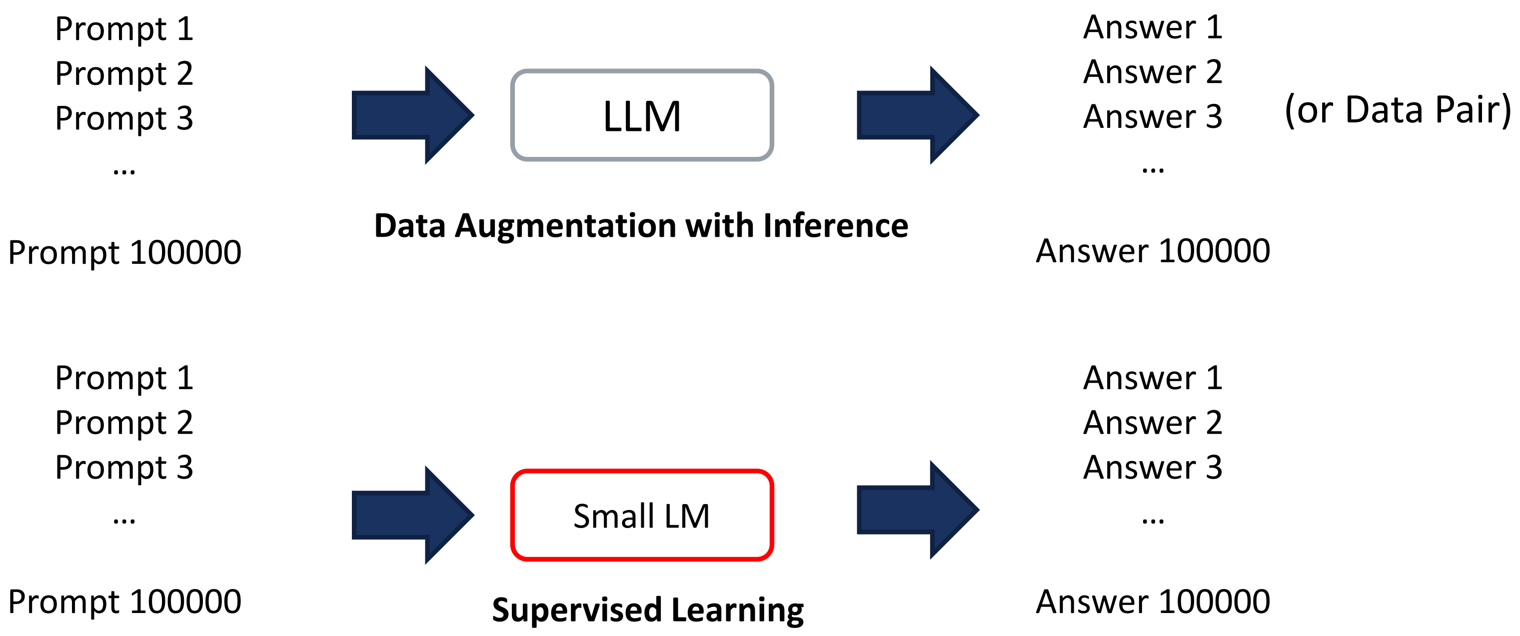
A representative from Hungary's Ötvös Loránd University (ELTE) visited Chung-Ang Uni-versity on March 6th. The goal of the meeting was to discuss ways for our universities to col-laborate in the future, including student exchanges and joint research.
ELTE University's visit was organized as follows:
The schedule began with a meeting in the President's Office on the 3rd floor of the Main Building (Building 201) at CAU's Seoul campus.

A representative from Hungary's Ötvös Loránd University (ELTE) visited Chung-Ang University to discuss future collaboration, including student exchanges and joint research, with the hope of enhancing research and learning opportunities for students at both institutions.

# Data Augmentation with LLMs

Part 3: Data Augmentation with Text Generation Systems

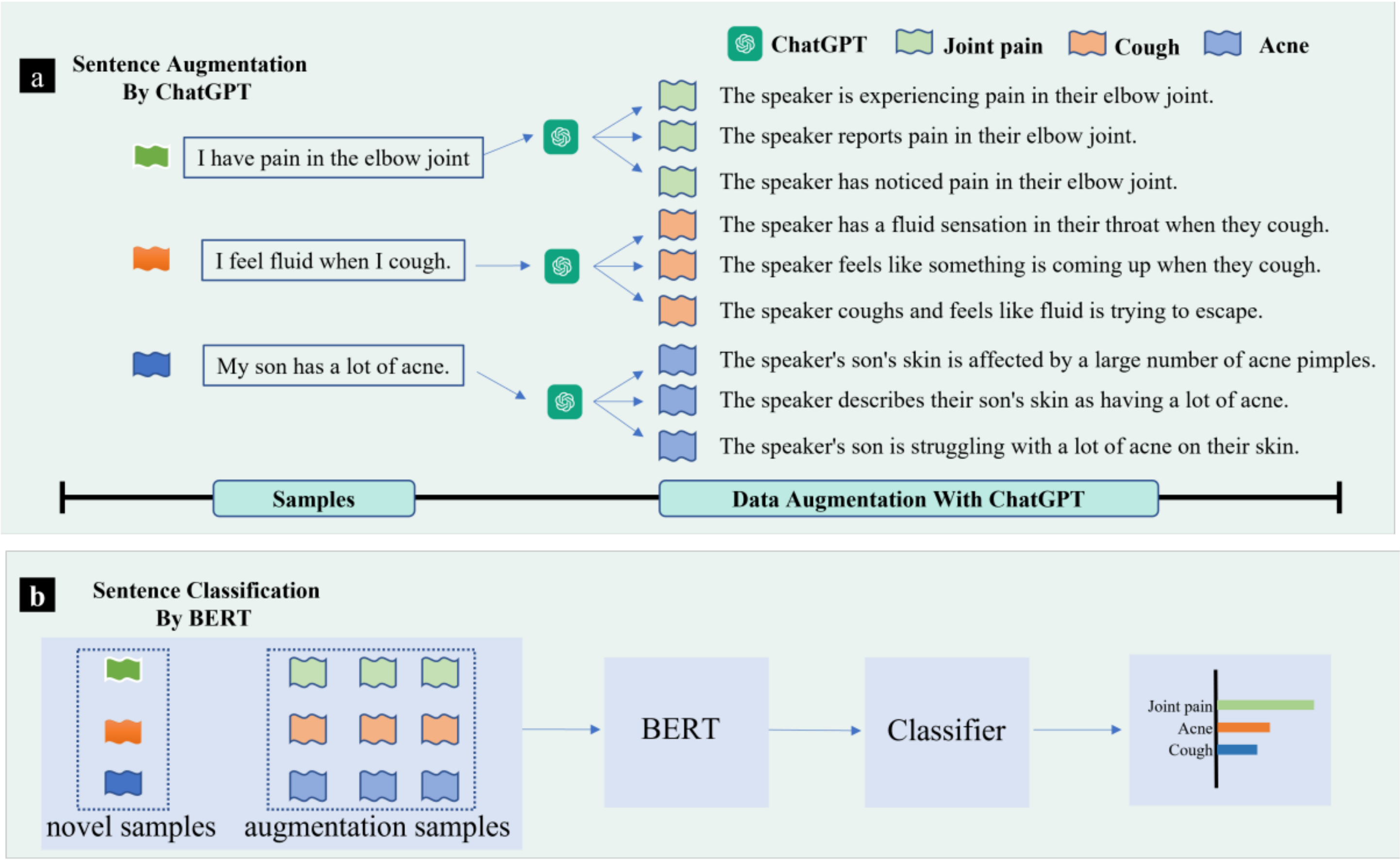
- Large Language Models (LLM, e.g ChatGPT, GPT-4) are too expensive for inference
- Generating dataset with LLM and train a small LM with supervised learning



# Data Augmentation with LLMs



- We can generate various datasets using ChatGPT or other LLMs.

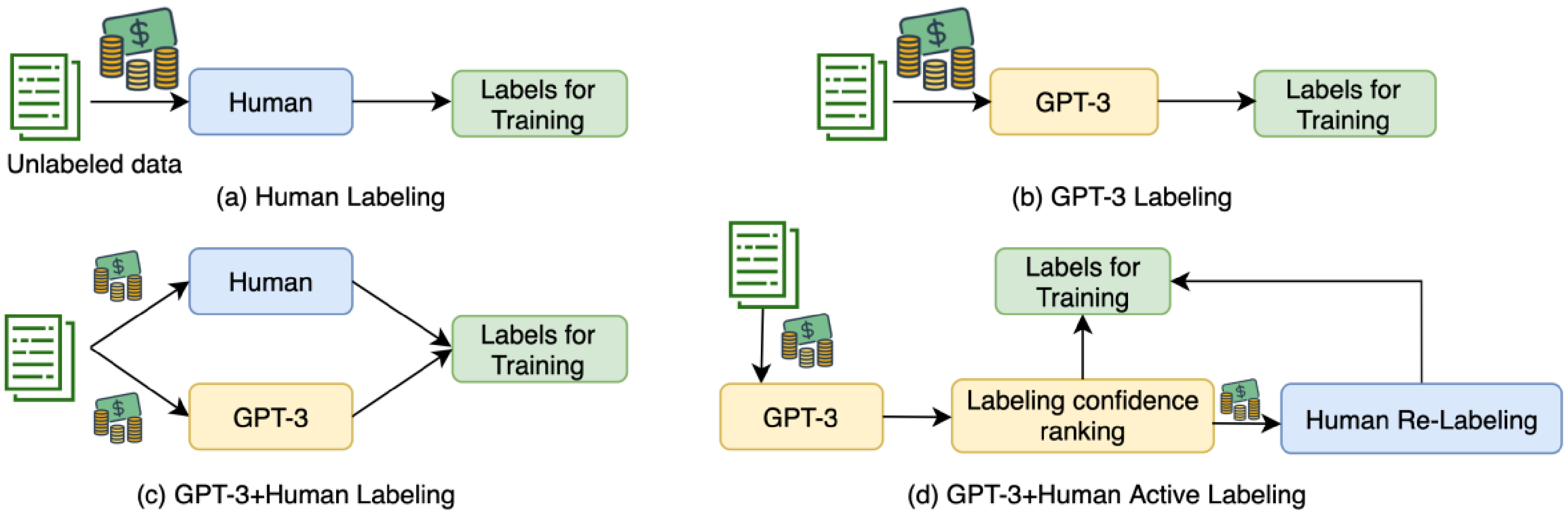



Dai et al., AugGPT: Leveraging ChatGPT for Text Data Augmentation, arXiv

# Data Augmentation with LLMs

**Part 3: Data Augmentation with Text Generation Systems**

- We can first generate datasets using LLMs and humans can re-annotate the data with lower confidence.



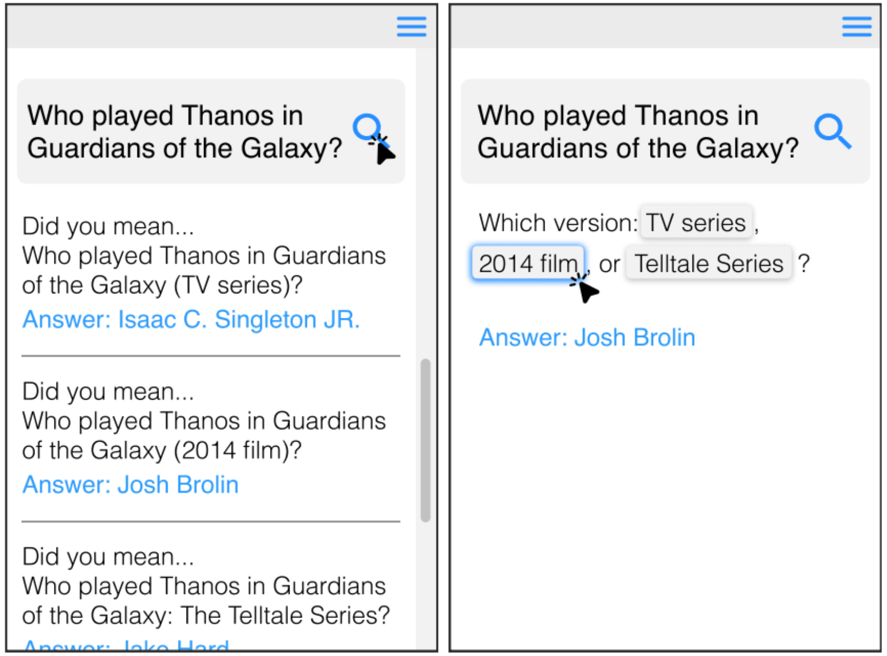

Wang et al., Want To Reduce Labeling Cost? GPT-3 Can Help, Findings of EMNLP 2021

# Data Augmentation with LLMs



- Our work focuses on generating datasets for the following tasks:

- Clarification question generation for QA




(a) (b)

- Context-aware sarcasm detection

**Person A:** The fried egg got burnt to a crisp.

**Person B:** This is going to be really crispy and crunchy. **(sarcasm)**

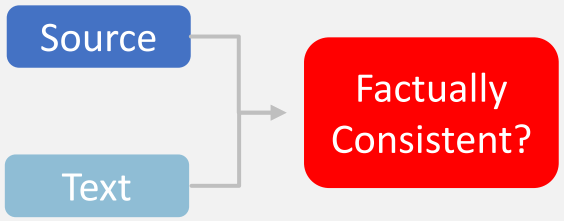
Lee et al., Asking Clarification Questions to Handle Ambiguity in Open-Domain QA, arXiv

# Summary

Conclusion

## Part 1

**<Evaluating Text Generation>**

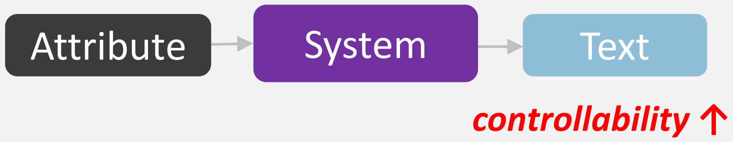



- Data Augmentation through Mask-and-Fill with Masked Article *(NAACL-22)*

- *Data generation by filling the masks in the summary*
- *Train factual consistent checking system using the data*

## Part 2

**<Controlling Text Generation>**




- Reinforcement Learning based Critic Guided Decoding *(ACL-23)*

- *Train only critic and freeze LM*
- *Adjust probability with critic in decoding*

## Part 3

**<Data Augmentation with LMs>**


- Generating training datasets using LLM

- *Distilling knowledge with LLM*
- *Generate datasets for low-resource tasks*

**Contact:** hwanheelee@cau.ac.kr