

CS224n 스터디

Lecture12. NLG

23.05.10

NLP 심화 세미나 3주차

투빅스 19기 윤세휘

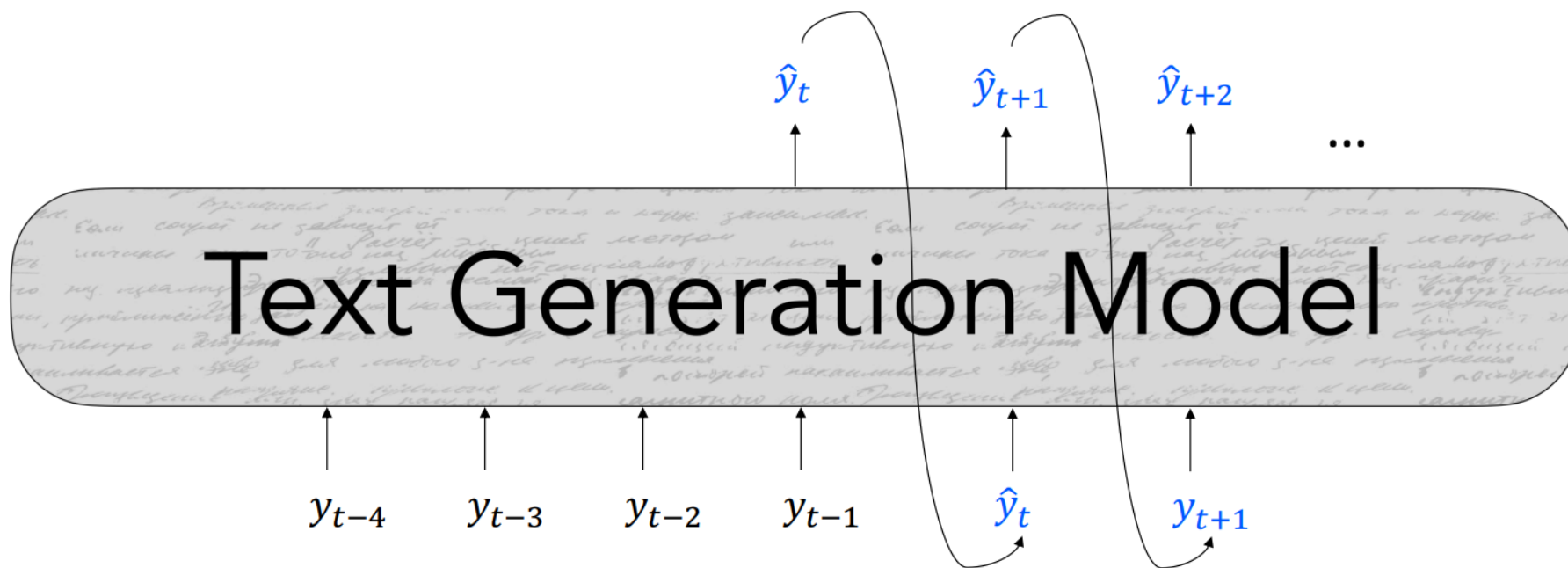
What is NLG?

NLG tasks

- **Text-to-Text Generation**
 - Machine Translation
 - Summarization
 - Dialogue System
- 보다 다양한 모달리티를 사용하는 시스템들..
- **Data-to-Text Generation**
 - 표, 지식그래프, 데이터 스트림 등으로부터 텍스트를 생성하는 시스템.
- **Visual Description**
 - 이미지나 영상을 텍스트로 설명하는 시스템.
- **Creative Generation**
 - Stories & Narratives, Poetry
 - 블로그 포스트 등

What is NLG?

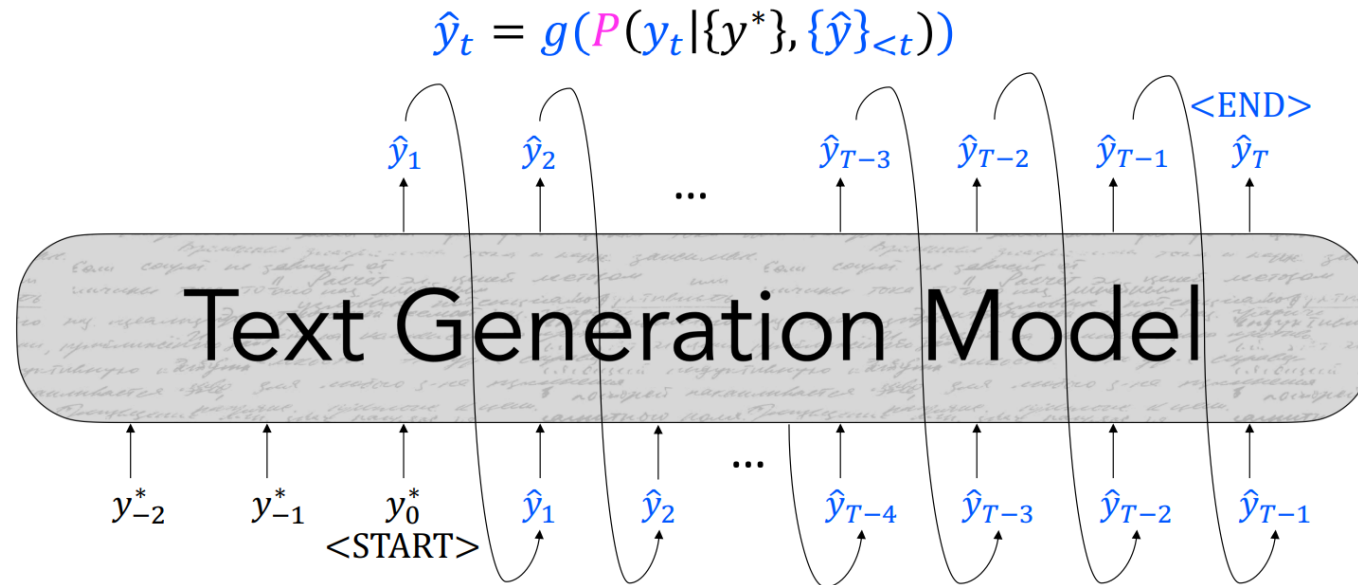
- NLG(Natural Language Generation, 자연어 생성)
- 각 time step t 마다 autoregressive text generation model은 토큰들의 시퀀스 $\{y\}_{<t}$ 를 입력으로 받아 \hat{y}_t 를 새로운 토큰으로 생성한다.



Decoding from NLG Systems

Decoding

- 모델이 생성한 토큰들의 확률분포로부터 하나의 토큰을 선택하는 과정이다.
- 대표적인 디코딩 알고리즘으로 Greedy Methods 중 Argmax Decoding, Beam Search가 있다.



Argmax Decoding: $\hat{y}_t = \underset{w \in V}{\operatorname{argmax}} P(y_t = w | y_{<t})$

Greedy Methods

- Greedy Methods의 문제점: **repetition problem**

Context: In a shocking finding, scientist 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.

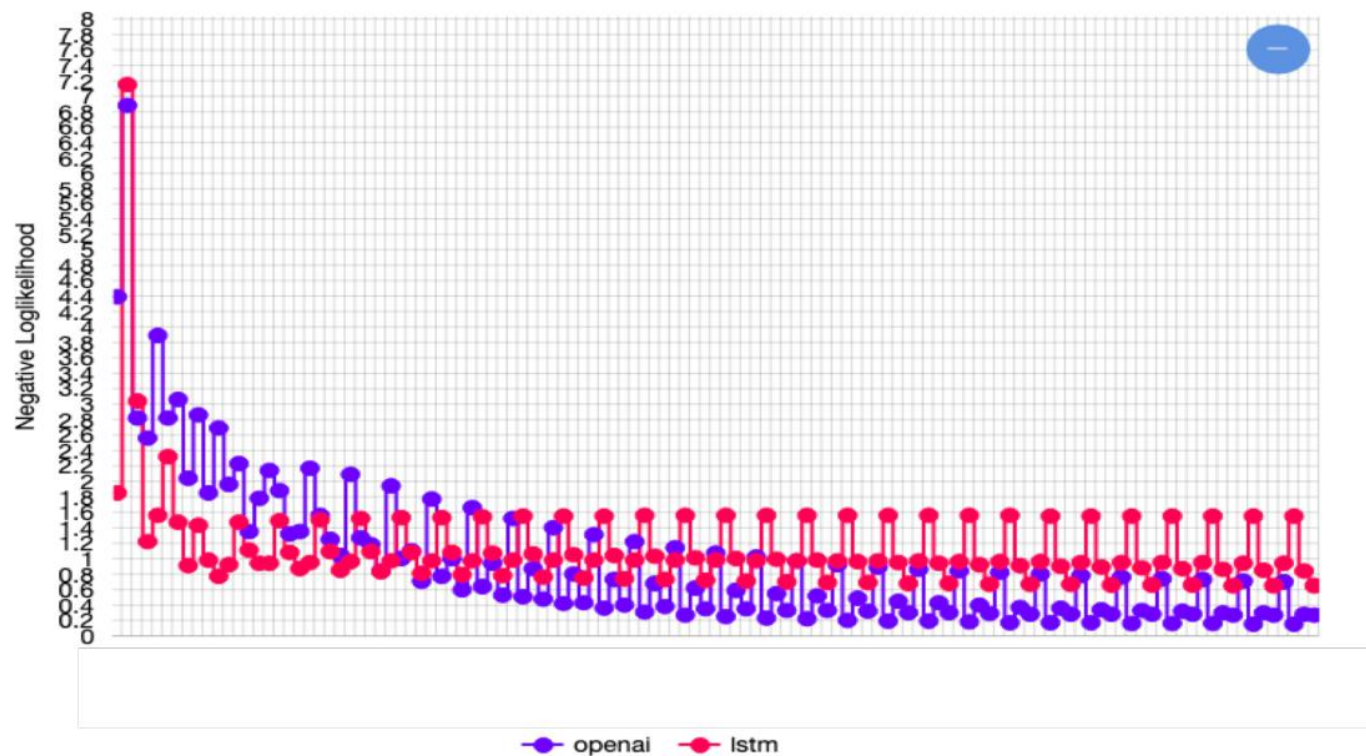
Continuation: The study, published in the Proceedings of the National Academy of Sciences of the United States of America (PNAS), was conducted by researchers from the **Universidad Nacional Autónoma de México (UNAM)** and **the Universidad Nacional Autónoma de México (UNAM/Universidad Nacional Autónoma de México/Universidad Nacional Autónoma de México/Universidad Nacional Autónoma de México/Universidad Nacional Autónoma de México...**

(Holtzman et. al., ICLR 2020)

Greedy Methods

- Greedy Methods의 문제점: **repetition problem**
 - 원인: 이전 히스토리를 기반으로 생성 토큰에 대한 confidence가 점점 높아짐.
 - GPT와 LSTM을 비교:

I'm tired. I'm tired. I'm tired. I'm tired. I'm tired. I'm tired. I'm tired. I'm tired. I'm tired. I'm tired. I'm tired.



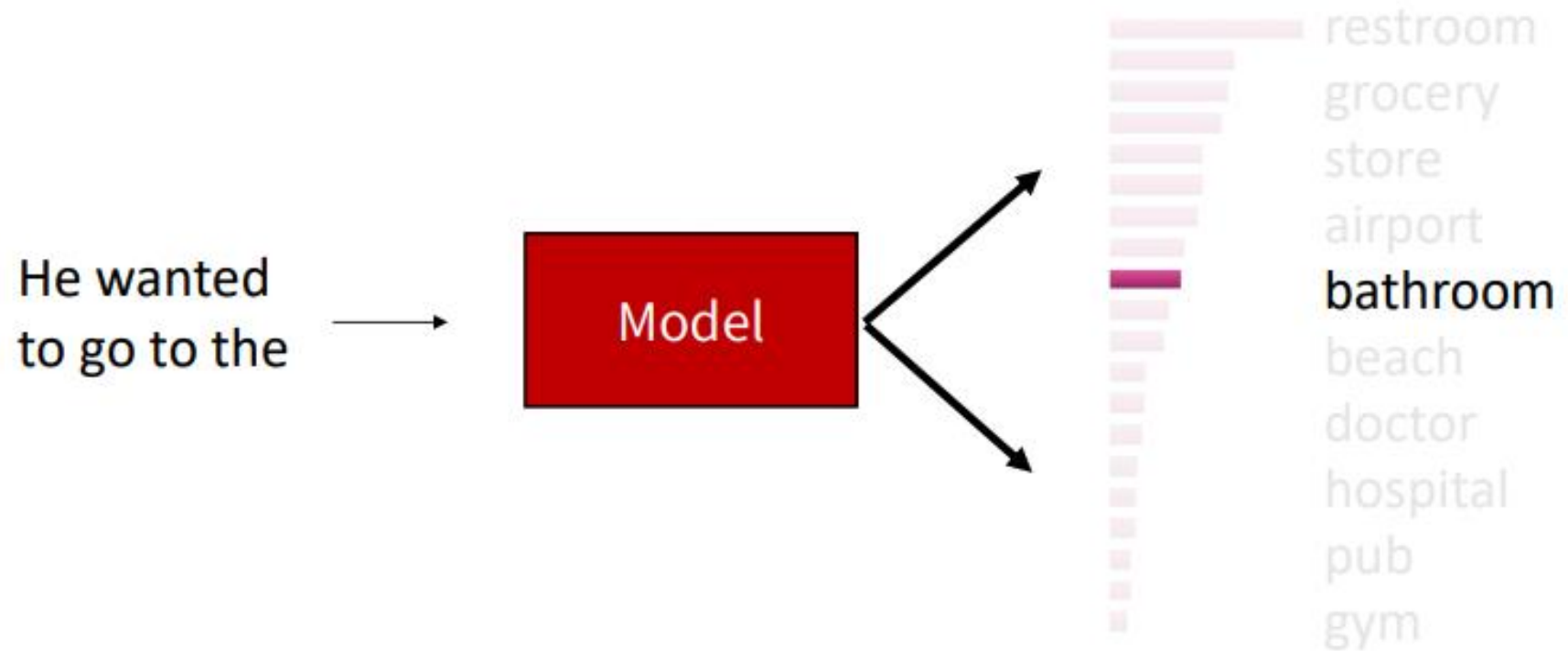
How can we reduce repetition?

- 휴리스틱 방법: 추론 단계에서 n-gram의 반복 사전차단
- 학습 단계에서 가능한 방법:
 - 연속되는 문장의 임베딩 거리를 최소화하는 방법
 - Coverage Loss: 같은 단어에 attending하는 것을 방지
 - **Unlikelihood Object**: 이미 등장한 토큰에 패널티 부여
- Greedy Methods외의 디코딩 알고리즘 사용
 - 1) Random Sampling
 - 2) Top-k Sampling
 - 3) Top-p Sampling
 - 4) Re-balancing distributions

1) Random Sampling

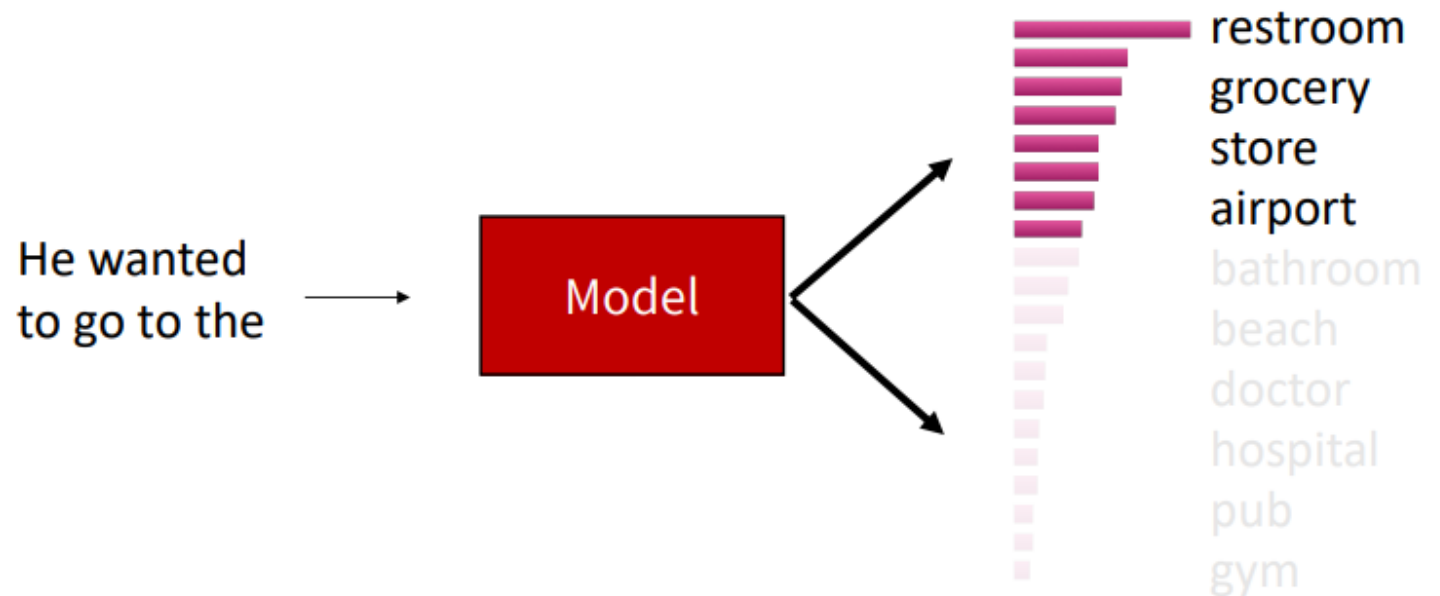
- 토큰 분포에서 랜덤으로 한 개의 토큰을 선택함

$$\hat{y}_t \sim P(y_t = w | \{y\}_{<t})$$



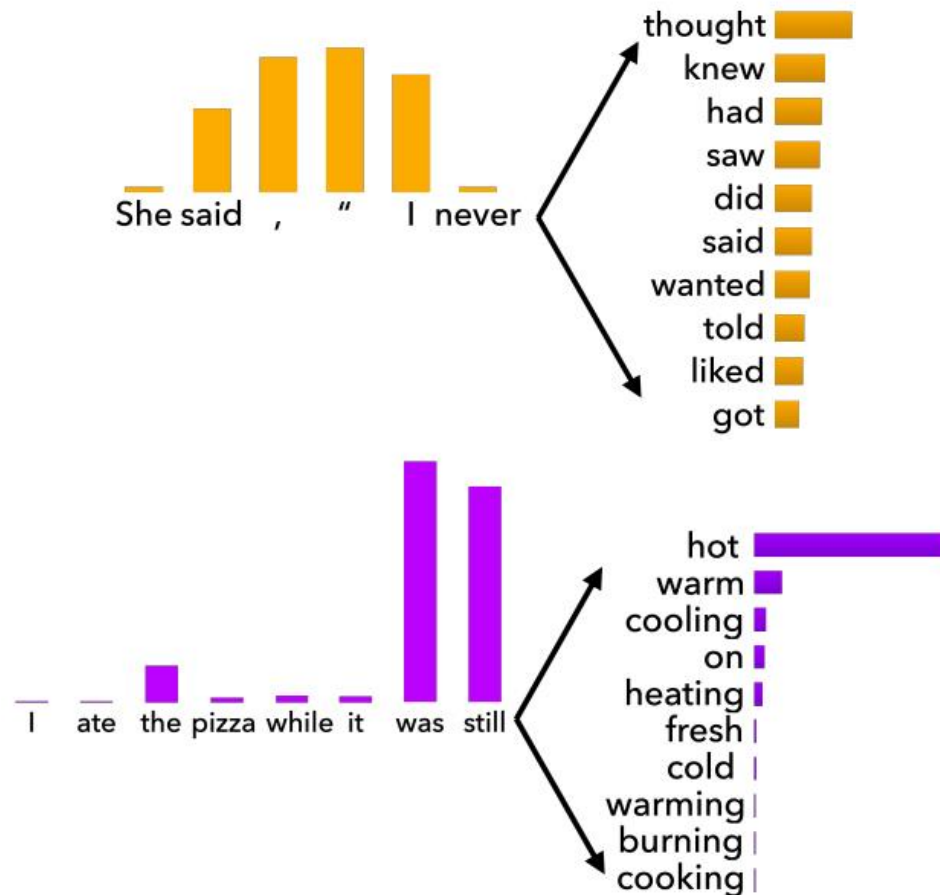
2) Top-k Sampling

- 확률 분포에서 확률이 제일 높은 k개의 토큰 중 하나의 토큰을 선택함.
- k값으로 주로 5, 10, 20을 사용함.
- k값이 클수록 diverse/risky, k값이 작을수록 generic/safe



2) Top-k Sampling

- Top-k sampling의 문제점



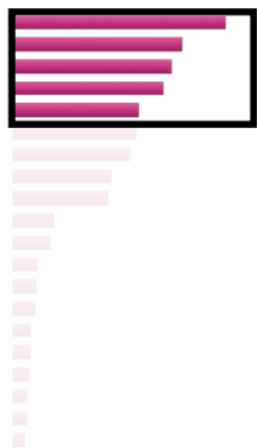
Top-k sampling can cut off too **quickly**!

Top-k sampling can also cut off too **slowly**!

3) Top-p Sampling

- 누적 확률 값이 p 보다 작은 상위 토큰들 중에 하나의 토큰을 선택함.
- p 값이 클수록 diverse/risky, p 값이 작을수록 generic/safe

$$P_t^1(y_t = w | \{y\}_{<t})$$



$$P_t^2(y_t = w | \{y\}_{<t})$$

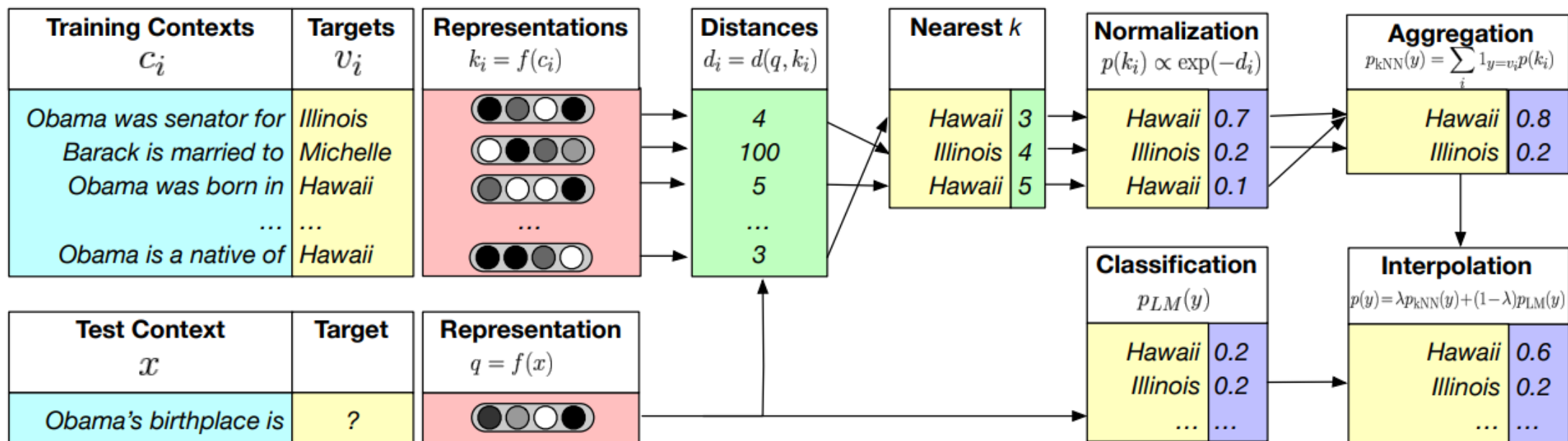


$$P_t^3(y_t = w | \{y\}_{<t})$$



4) Re-balancing distributions

- 추론 단계에서 평가 문장을 학습된 문장들의 representation과 비교하여 모델의 토큰 분포를 보정해주는 방법.
- k개의 인접한 representation 문장의 target값을 사용해 모델 평가 결과를 보정함



Training NLG Systems

Unlikelihood Training

- 이미 생성된 토큰이 다시 생성될 확률을 낮추도록 하는 패널티를 기존 loss function에 추가함.

이미 생성된 토큰

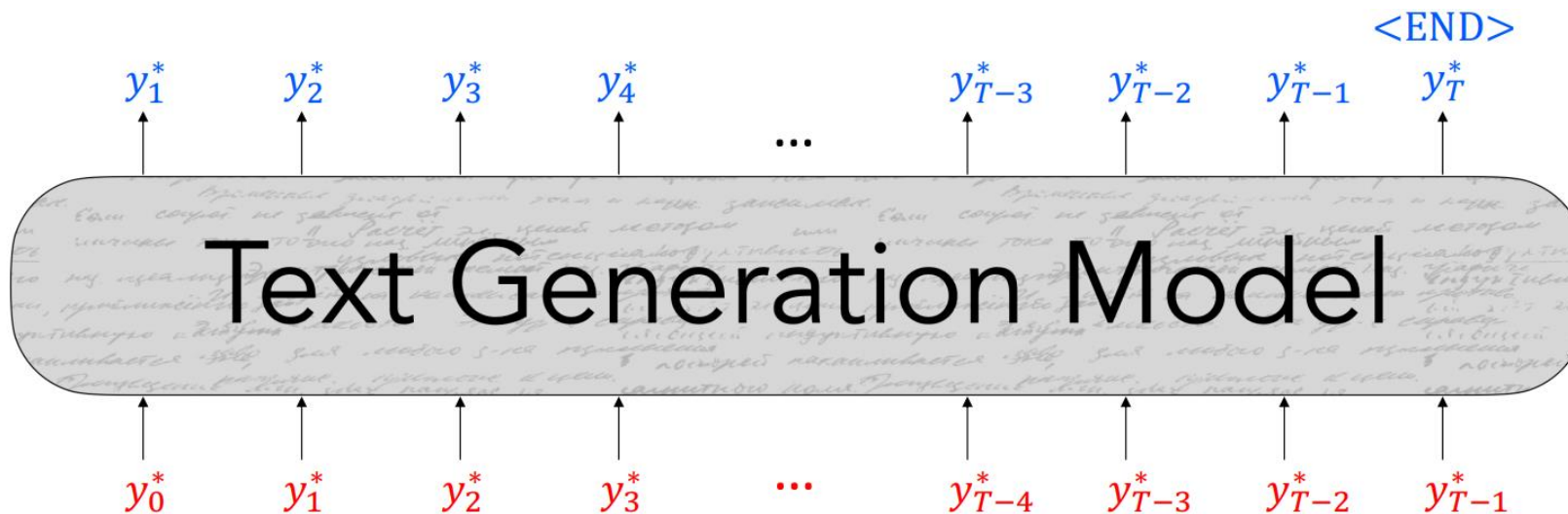
$$\mathcal{C} = \{y^*\}_{<t}$$

$$\mathcal{L}_{UL}^t = - \sum_{y_{neg} \in \mathcal{C}} \log(1 - P(y_{neg} \mid \{y^*\}_{<t}))$$

$$\mathcal{L}_{ULE}^t = \mathcal{L}_{MLE}^t + \alpha \mathcal{L}_{UL}^t$$

Exposure Bias

- Teacher forcing의 문제점: **Exposure Bias**
- 추론 단계에서 잘못된 예측으로 인해 뒤의 시퀀스까지 영향을 받아 부자연스러운 문장을 생성하게 된다.

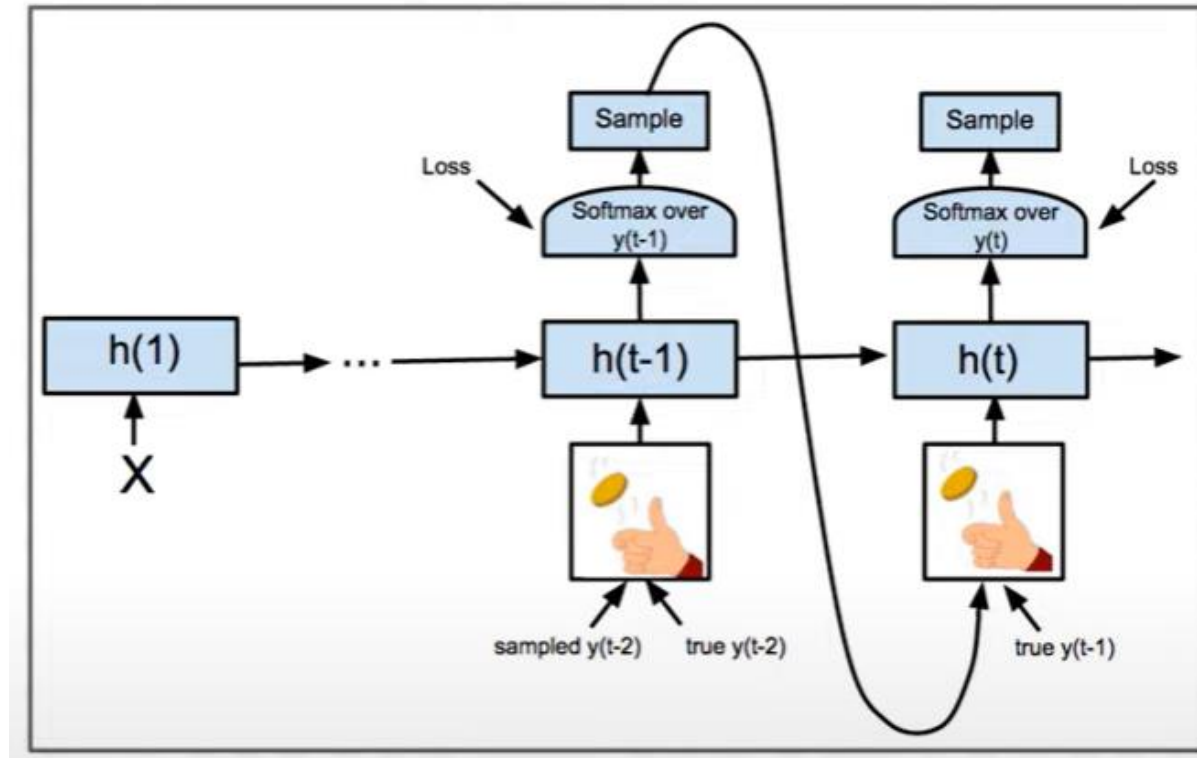


Exposure Bias Solutions

- Teacher forcing의 문제점: **Exposure Bias**
- 솔루션:
 - Scheduled sampling
 - Dataset Aggregation
 - Sequence re-writing
 - Reinforcement Learning

Scheduled Sampling

- 특정 확률로 이전 생성된 토큰을 다음 step의 입력으로 사용하는 방법
- 학습이 진행될 수록 더 적은 gold token을 사용함.



Evaluating NLG Systems

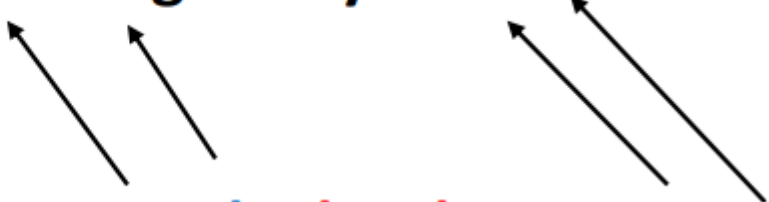
Text Generation 성능 평가 유형

- 1) Content Overlap Metrics
- 2) Model-based Metrics
- 3) Human Evaluations

1) Content Overlap Metrics

reference **Ref: They walked to the grocery store .**

모델이 생성한 텍스트 **Gen: The woman went to the hardware store .**



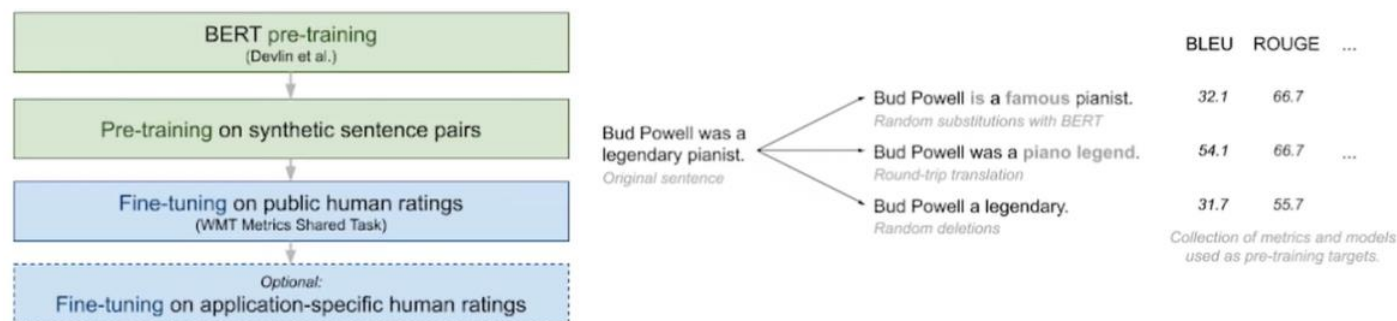
- 특징: Fast & Efficient & Widely Used
- 2가지로 구분된다.
 - **N-gram overlap metrics** (BLEU, ROUGE, METEOR, CIDEr 등)
 - **Semantic overlap metrics** (PYRAMID, SPICE, SPIDER 등)

N-gram overlap metrics

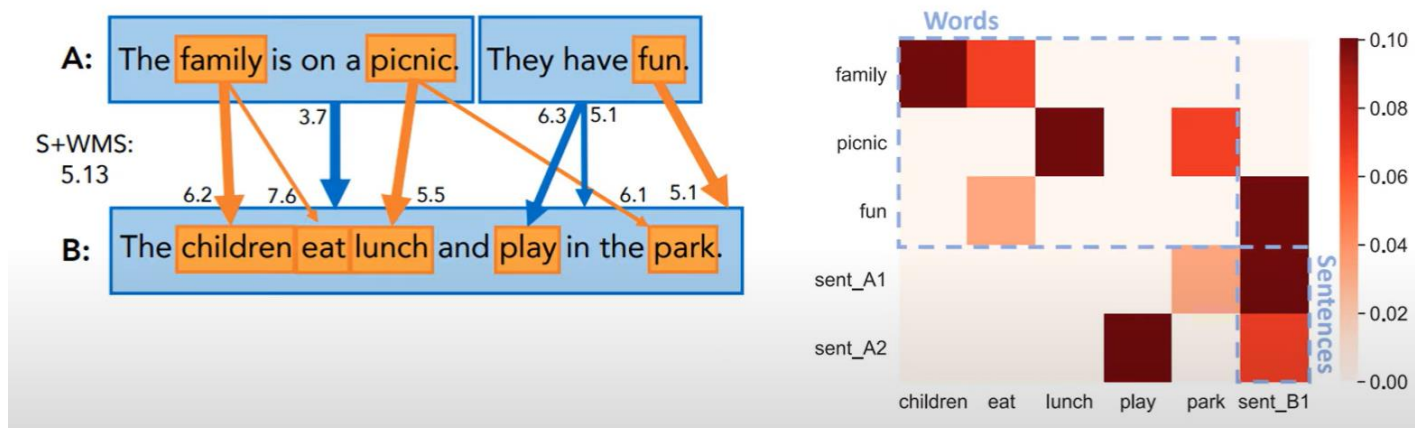
- Word overlap based metrics (BLEU, ROUGE, METEOR, CIDEr 등)
- machine translation에도 최적의 메트릭은 아닐 뿐더러 summarization, dialogue, story generation과 같은 open-ended task들에는 더욱 적합하지 않다.

2) Model-based metrics

- BLEURT: 문장 유사도를 출력하는 BERT기반의 regression모델을 학습함



- Sentence Mover's Similarity: sentence level, word level에서 모두 유사도를 계산함.



3) Human Evaluation

- ACL 2019의 generation 관련 논문들 중 75%가 human evaluation 방법을 사용하였다.
- 새로운 자동 메트릭을 개발할 때 **gold standard**로서 사용된다.
- human evaluation에서 주로 사용하는 평가 지표:
 - fluency (유창성)
 - coherence/consistency (일관성)
 - factuality and correctness (사실성)
 - commonsense (상식)
 - style/formality (형식/격식)
 - grammaticality (문법)
 - typicality (전형성)
 - redundancy (중복성)

3) Human Evaluation

- human evaluation이 시간과 비용이 많이 든다는 점 외에도
- 사람이 하는 일이기 때문에 발생하는 어려움들이 있다.
 - 일관적이지 않음
 - 비논리적일 가능성이 있음
 - 집중력을 잃음
 - 오해의 소지가 있음
 - 언제나 판단에 대한 이유를 설명할 수 없음



Learning from Human Feedback

- 예시: ADEM, HUSE

Reference

- **Stanford CS224N NLP with Deep Learning | Winter 2021 | Lecture 12 - Natural Language Generation**
 - https://www.youtube.com/watch?v=1uMo8olr5ng&list=PLoROMvody4rOSH4v6133s9LFPRHjEmbmJ&index=12&ab_channel=StanfordOnline
- **[DSBA] CS224n 2021 Study | #12 Natural Language Generation**
 - https://www.youtube.com/watch?v=RkCbFQ1W6_Q&list=PLetSIH8YjlfVk0G_IQfVCCOJ4M_iMZCJQ&index=6&ab_channel=%EA%B3%A0%EB%A0%A4%EB%8C%80%ED%95%99%EA%B5%90%EC%82%B0%EC%97%85%EA%B2%BD%EC%98%81%EA%B3%B5%ED%95%99%EB%B6%80DSBA%EC%97%B0%EA%B5%AC%EC%8B%A4