



KOREA UNIVERSITY

DSBA CS224n 2021 Study

[Lecture 12]

Natural Language Generation



고려대학교 산업경영공학과
Data Science & Business Analytics Lab

발표자 : 정용기



00

발표 목차

Contents

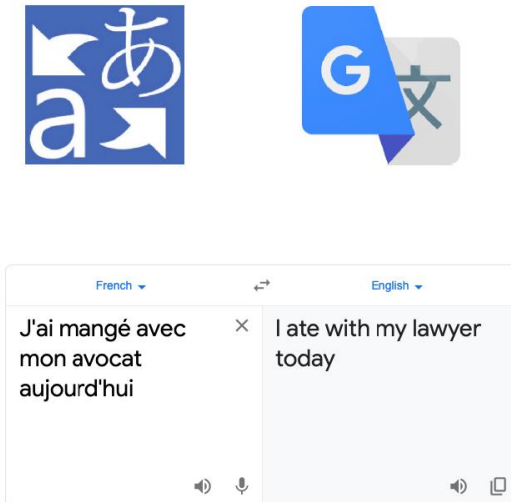
- 1 What is NLG
- 2 Decoding
- 3 Training
- 4 Evaluating
- 5 Summary



What is natural language generation

- NLG 는 자연어처리(NLP)의 한 분야
- Machine Translation, Dialogue System, Summarization, Data-to-Text Generation, Visual Description

<Machine Translation>



<Summarization>

C: Looking at what we've got, we we want an LCD display with a spinning wheel.
 B: You have to have some push-buttons, don't you?
 C: Just spinning and not scrolling, I would say.
 B: I think the spinning wheel is definitely very now.
 A: but since LCDs seems to be uh a definite yes,
 C: We're having push-buttons on the outside
 C: and then on the inside an LCD with spinning wheel,

Decision Abstract (Summary):
 The remote will have push buttons outside, and an LCD and spinning wheel inside.

A: and um I'm not sure about the buttons being in the shape of fruit though.
 D: Maybe make it like fruity colours or something.
 C: The power button could be like a big apple or something.
 D: Um like I'm just thinking bright colours.

Problem Abstract (Summary):
 How to incorporate a fruit and vegetable theme into the remote.

(Wang and Cardie, ACL 2013)

<Data-to-Text Generation>

TEAM	WIN	LOSS	PTS	FG_PCT	RB	AS ...
Heat	11	12	103	49	47	27
Hawks	7	15	95	43	33	20

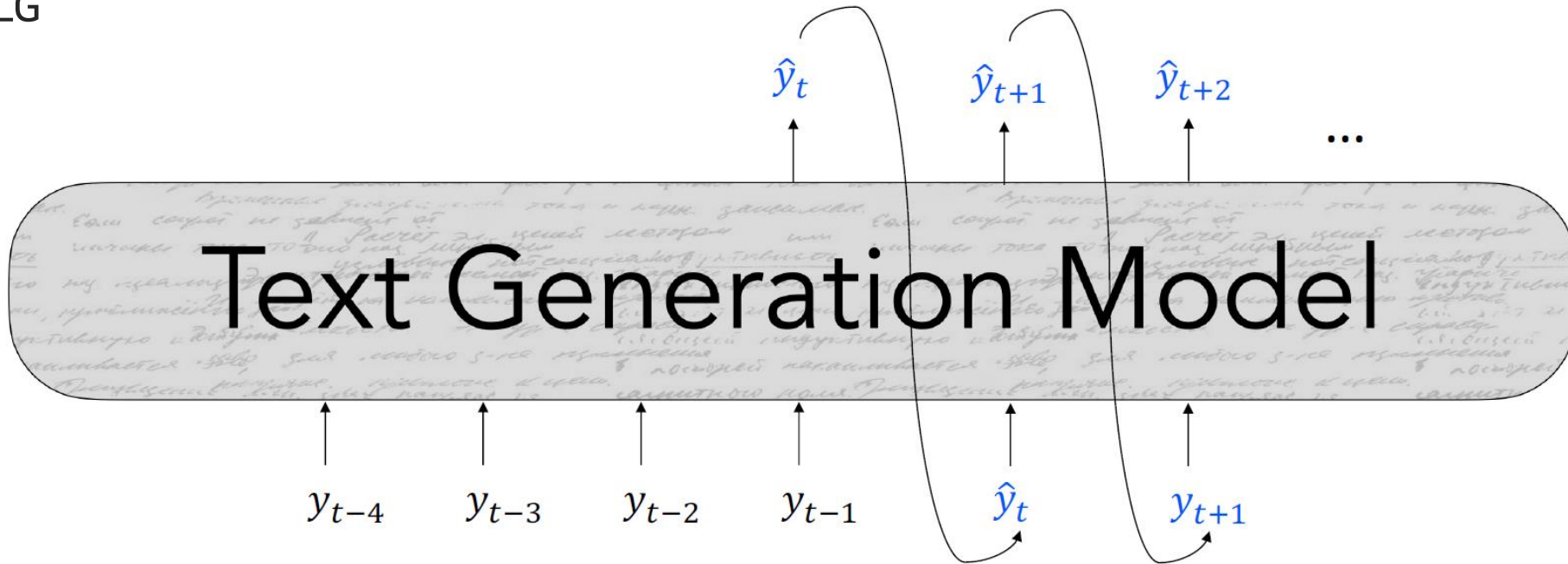
PLAYER	AS	RB	PT	FG	FGA	CITY ...
Tyler Johnson	5	2	27	8	16	Miami
Dwight Howard	4	17	23	9	11	Atlanta
Paul Millsap	2	9	21	8	12	Atlanta
Goran Dragic	4	2	21	8	17	Miami
Wayne Ellington	2	3	19	7	15	Miami
Dennis Schroder	7	4	17	8	15	Atlanta
Rodney McGruder	5	5	11	3	8	Miami
Thabo Sefolosha	5	5	10	5	11	Atlanta
Kyle Korver	5	3	9	3	9	Atlanta
...						

The Atlanta Hawks defeated the Miami Heat , 103 - 95 , at Philips Arena on Wednesday . Atlanta was in desperate need of a win and they were able to take care of a shorthanded Miami team here . Defense was key for the Hawks , as they held the Heat to 42 percent shooting and forced them to commit 16 turnovers . Atlanta also dominated in the paint , winning the rebounding battle , 47 - 34 , and outscoring them in the paint 58 - 26. The Hawks shot 49 percent from the field and assisted on 27 of their 43 made baskets . This was a near wire - to - wire win for the Hawks , as Miami held just one lead in the first five minutes . Miami (7 - 15) are as beat - up as anyone right now and it 's taking a toll on the heavily used starters . Hassan Whiteside really struggled in this game , as he amassed eight points , 12 rebounds and one blocks on 4 - of - 12 shooting ...

(Wiseman and Rush., EMNLP 2017)

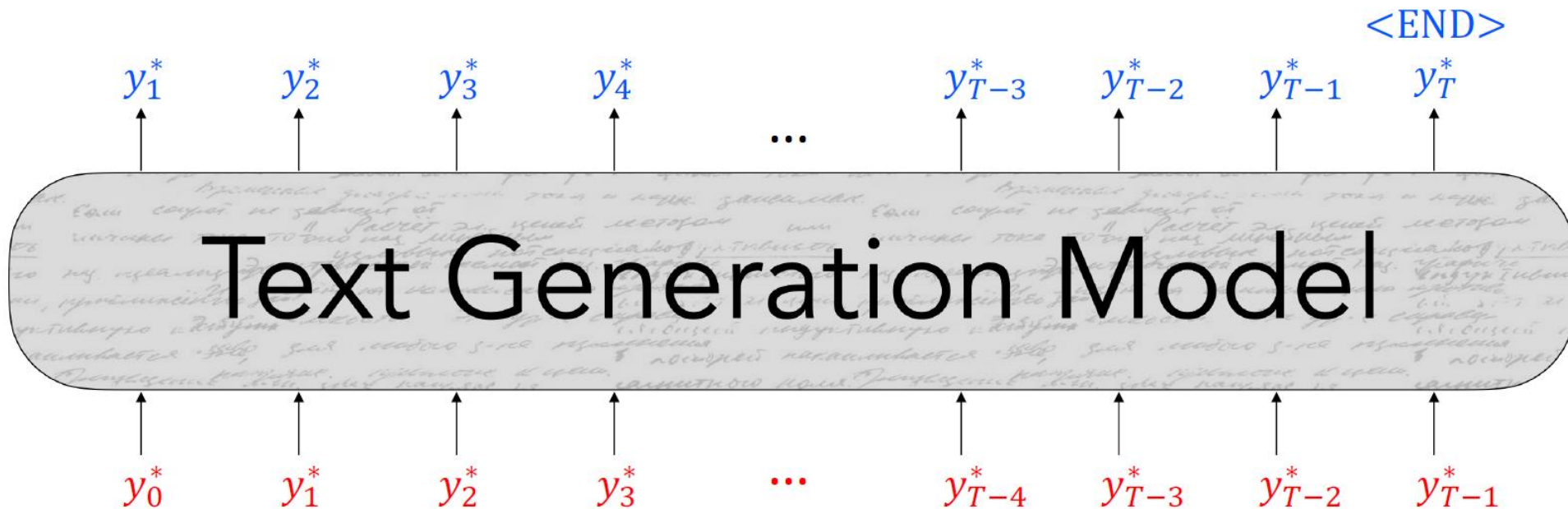
- 사람의 편의를 위해 text 를 생성하는 모든 업무는 NLG 사용가능

Basic NLG



- Autoregressive 형태의 일반적인 NLG 모델은 이전까지의 단어 $\{y\}_{< t}$ 를 입력으로 다음 스텝의 단어 \hat{y}_t 를 생성함
- 학습시엔 예측된 토큰과 실제 토큰을 사용한 Negative loglikelihood 를 minimize 하는 방식을 사용

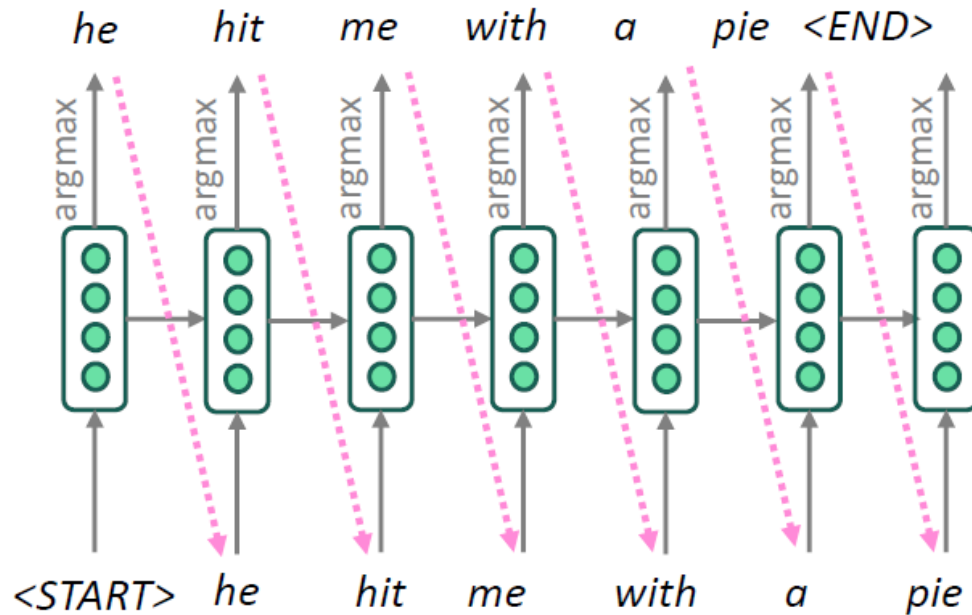
Teacher forcing



- 학습 시 모델이 예측한 토큰이 아닌 실제 문장의 토큰을 다음 단어 생성을 위한 입력으로 사용할 수 있음
- 초기에 잘못 생성된 단어로 인해 이 후 계속 잘못된 단어가 생성되는 것을 막아줌
- 학습된 모델을 사용하여 decoding 시엔 예측된 단어를 다음 단어 예측을 위한 입력으로 사용

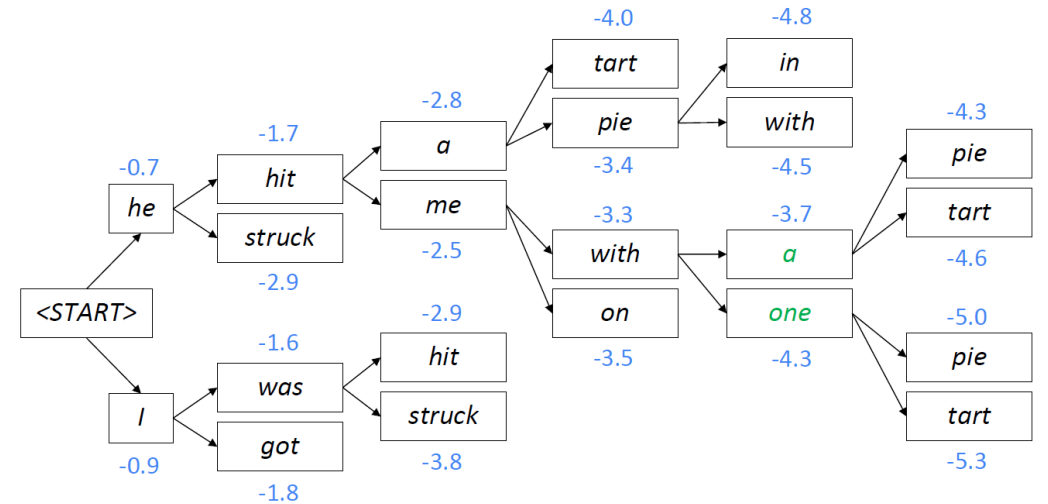
$$\mathcal{L} = - \sum_{t=1}^T \log P(y_t^* | \{y^*\}_{<t})$$

Argmax Decoding



- 매 step마다 vocab 내의 단어 중 probability가 최대가 되는 단어를 선택

Beam Search



For each of the k hypotheses, find top k next words and calculate scores

- Argmax decoding 보다 좀 더 많은(k) 후보들을 비교



Greedy methods get repetitive

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.

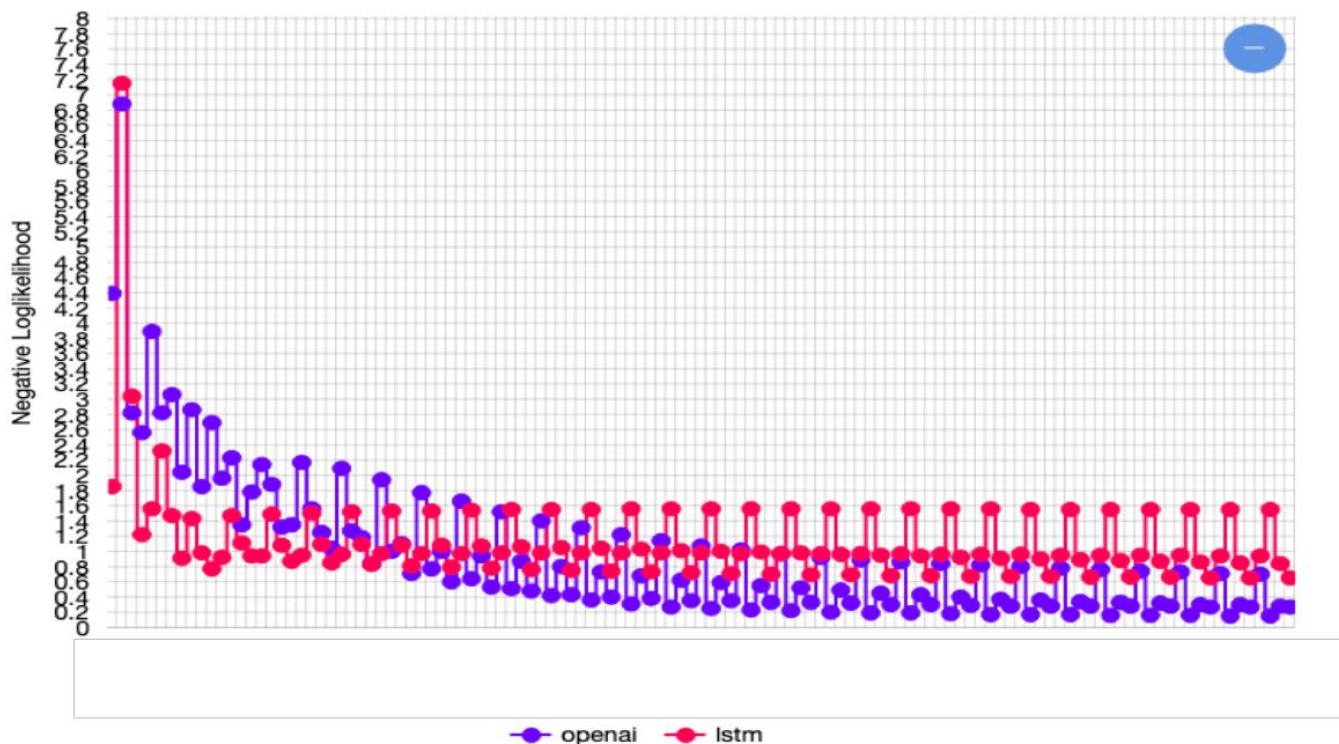
Continuation: The study, published in the Proceedings of the National Academy of Sciences of the United States of America (PNAS), was conducted by researchers from the **Universidad Nacional Autónoma de México (UNAM)** and **the Universidad Nacional Autónoma de México (UNAM/Universidad Nacional Autónoma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

(Holtzman et. al., ICLR 2020)



Greedy methods get repetitive

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.



(Holtzman et. al., ICLR 2020)

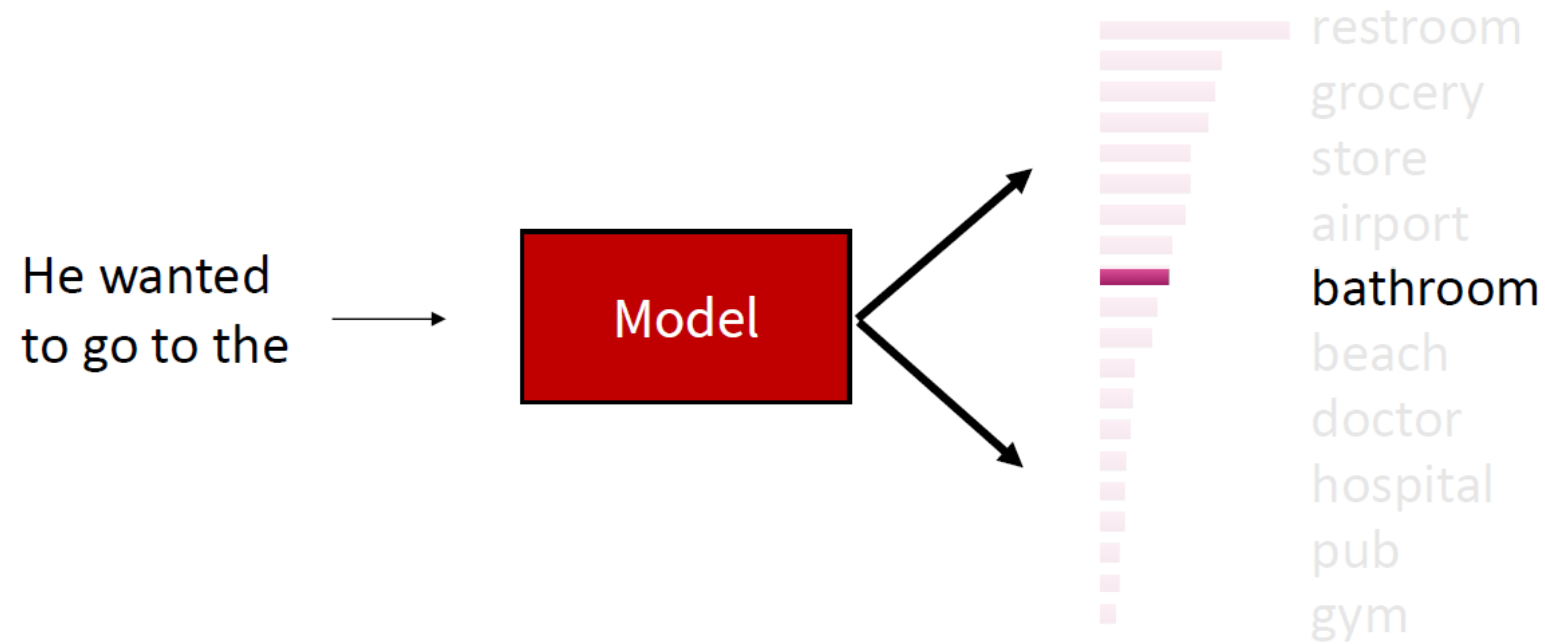
- Greedy decoding 방식은 동일한 phrase 를 반복해서 생성하는 문제가 있음
 - Dialogue task 나 chat bot system 같은 open end 문장을 생성할 때 주로 발생됨
- GPT 에선 특정 표현이 반복 생성될 때, 생성 결과에 대해 모델의 confidence 가 높아짐
 - Bottleneck 을 없앤 attention 구조 때문

How to reduce

- Decoding 단계에서 n-grams 의 반복을 막음
- Decoding 알고리즘 변경
 - Random sampling
 - Top-k sampling
 - Top-p sampling
 - + Scaling randomness : Temperature
 - Re-balancing distributions
- 연속되는 문장의 hidden representation similarity 를 낮추는 방식 - 문장 내 반복은 막을 수 없음
- 동일한 단어가 등장하는 것을 막는 attention mechanism : Coverage loss
- 이미 등장한 토큰에 패널티 부여 : Unlikelihood objective

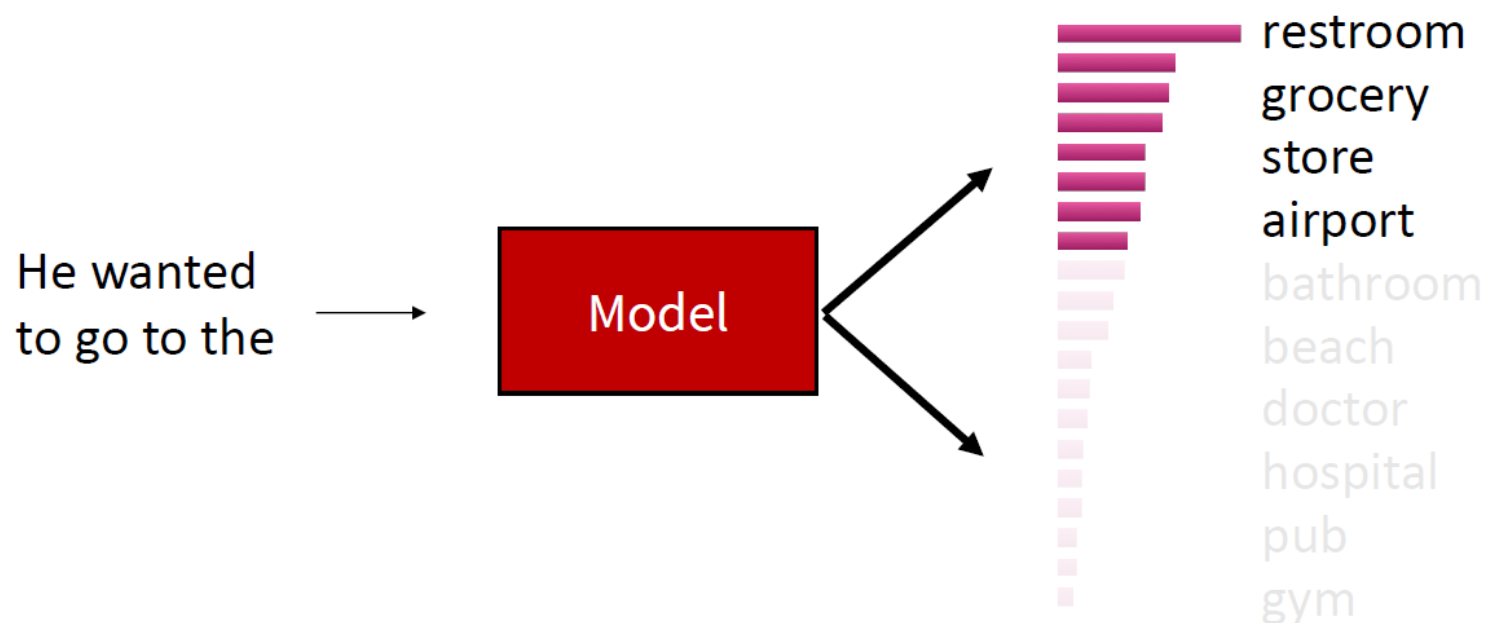


Random sampling



- 모델이 예측한 token distribution 을 sampling 에 사용 → 어떤 단어든(prob>0) 등장할 수 있음

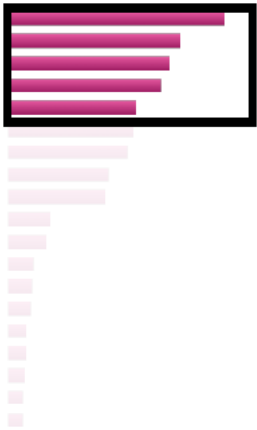
Top-k sampling



- 모델이 예측한 token distribution 에서 상위 k개에서만 sampling(ex. k=5, 10, 20)
- k ↑ : diverse/risky outputs
- k ↓ : generic/safe outputs

Top-p sampling

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



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

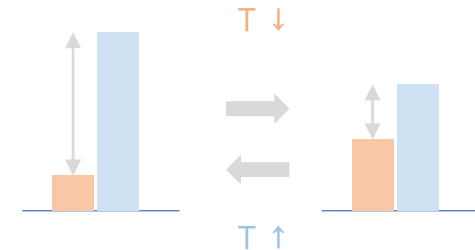
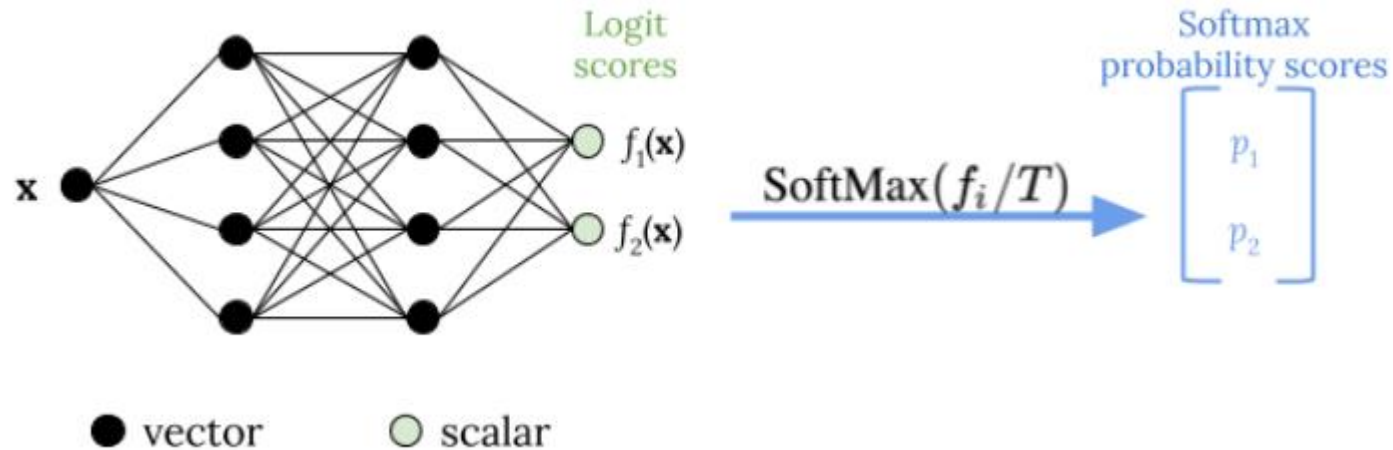


$$P_t^3(y_t = w | \{y\}_{<t})$$



- 누적 확률 값이 p 보다 작은 상위 토큰들만 샘플링에 사용
- p ↑ : diverse/risky outputs
- p ↓ : generic/safe outputs

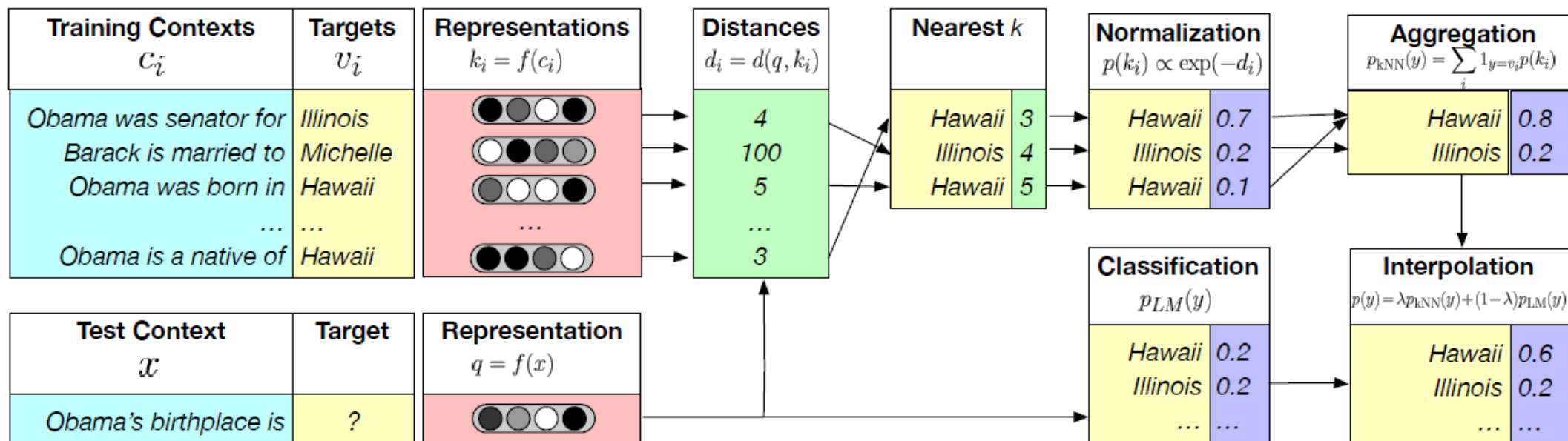
Scaling randomness



<https://docs.aws.amazon.com/prescriptive-guidance/latest/ml-quantifying-uncertainty/temp-scaling.html>

- Logits 값을 상수 T 로 나누고 softmax 의 입력으로 사용 (자체가 decoding 알고리즘은 아님)
- $T \uparrow$: diverse/risky outputs, vocab 내 probability 차이가 작아짐
- $T \downarrow$: generic/safe outputs, vocab 내 probability 차이가 커짐

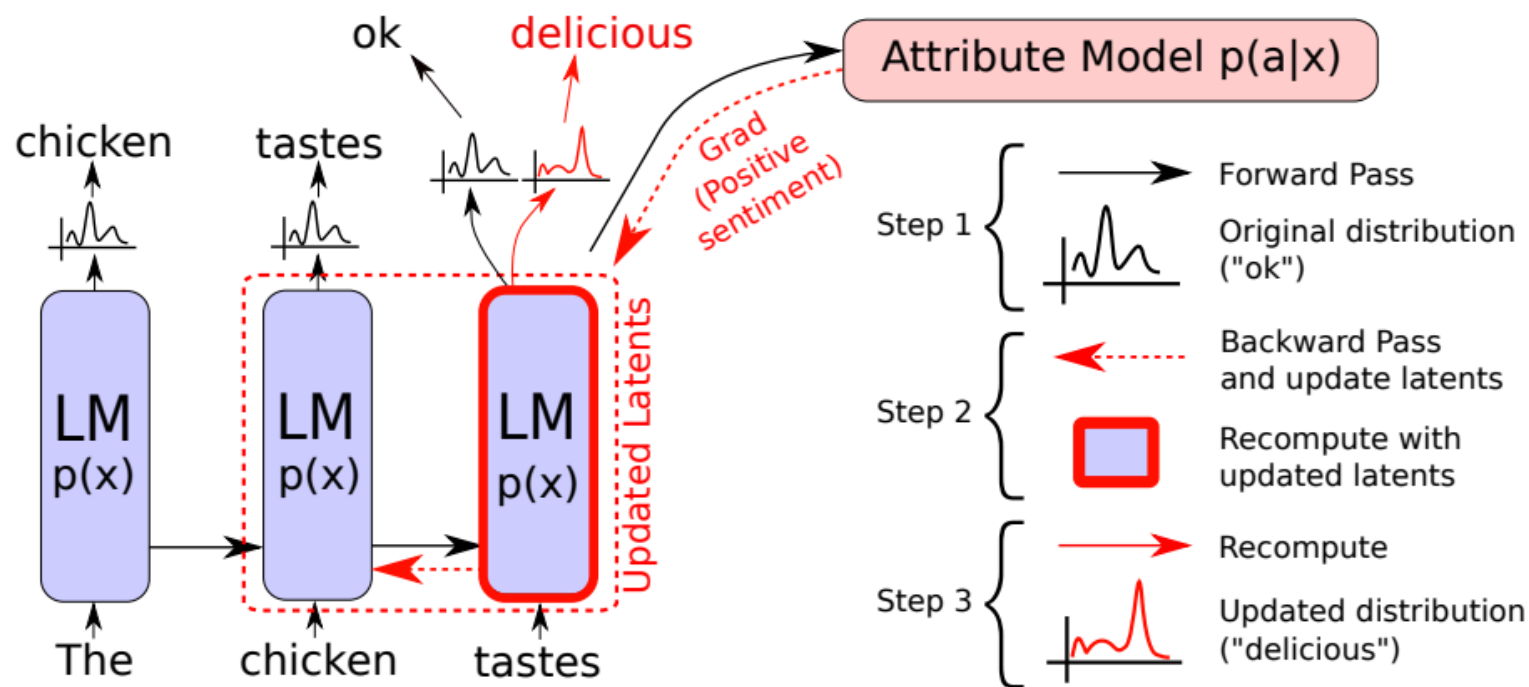
Re-balancing : kNN-LM



Generalization through Memorization: Nearest Neighbor Language Models

- Inference 시에 평가 문장을 학습된 문장들의 representation 과 비교하여 모델의 토큰 distribution 을 보정해주는 방식
- K 개의 인접한 representation 문장의 target 값을 사용해 모델 평가 결과를 보정

Re-balancing : PPLM



Plug and Play Language Models: A Simple Approach to Controlled Text Generation

- 추가적인 모델을 사용해 언어모델의 distribution 조정 (ex. Sentiment, perplexity)
- Attribute model로부터 gradient를 전달받아 latent 업데이트 → Model distribution 업데이트

Unlikelihood Training

- 이미 생성된 토큰이 생성될 확률을 낮추는 패널티를 기존 loss function 에 추가함

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

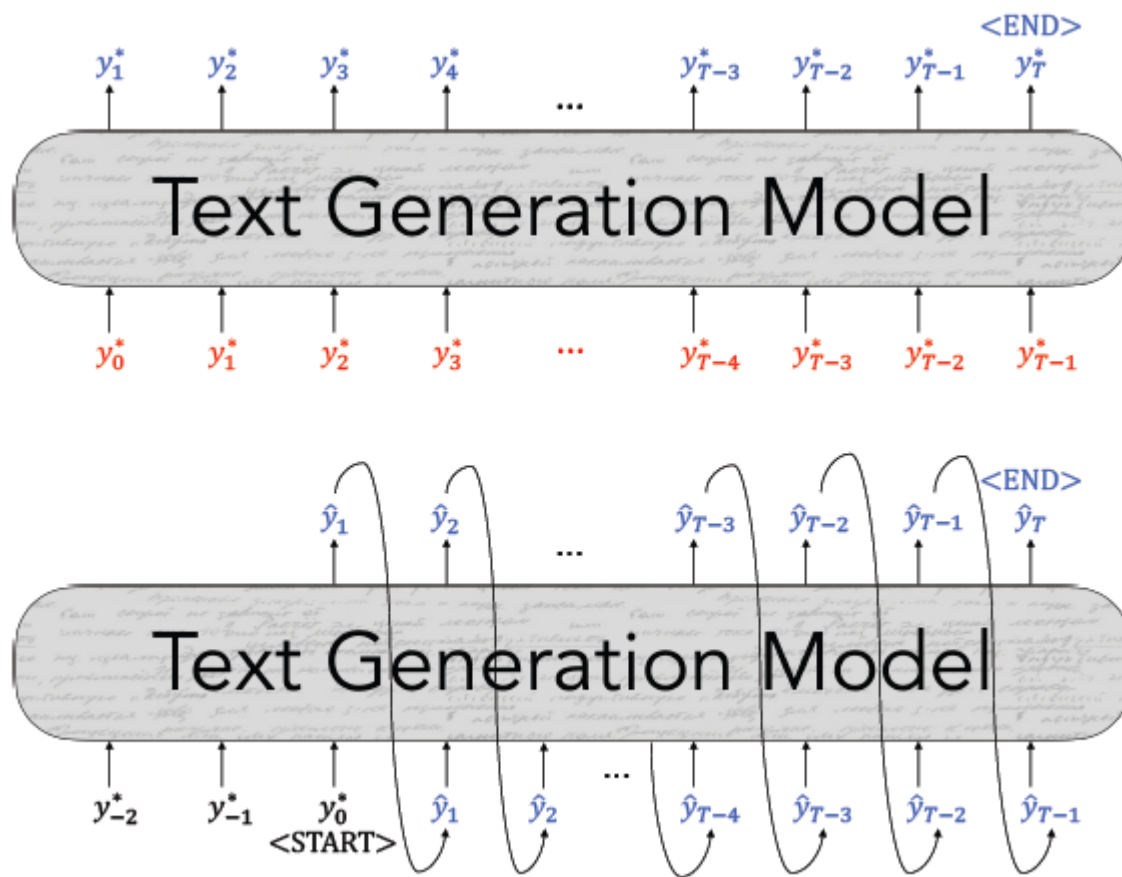
$$\mathcal{L}_{UL}^t = - \sum_{y_{neg} \in \mathcal{C}} \log(1 - P(y_{neg} \mid \{y^*\}_{<t}))$$

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

Neural Text Generation with Unlikelihood Training

- 반복된 단어, 구의 생성을 줄일 수 있음
- 생성되는 text 의 다양성을 증가시킴

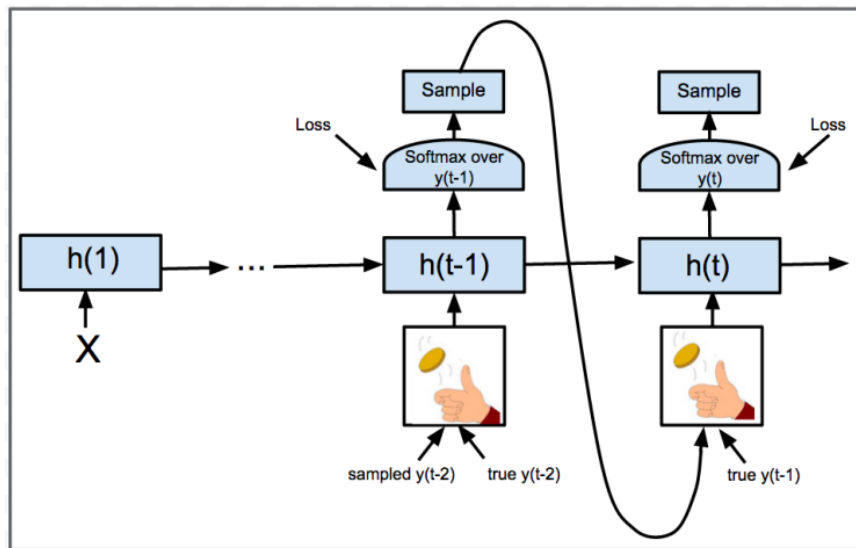
Teacher forcing \rightarrow Bias



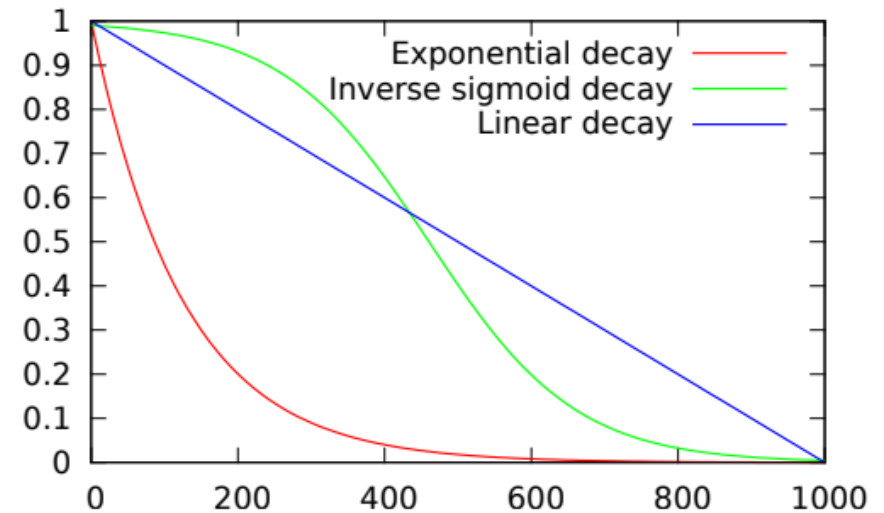
- Teacher forcing 으로 인해 모델이 불필요한 bias 를 학습하게 됨(=overfitting)
- Exposure Bias Solutions
 - Scheduled sampling
 - Dataset Aggregation
 - Sequence re-writing
 - Reinforcement Learning

Scheduled sampling

<Overall framework>



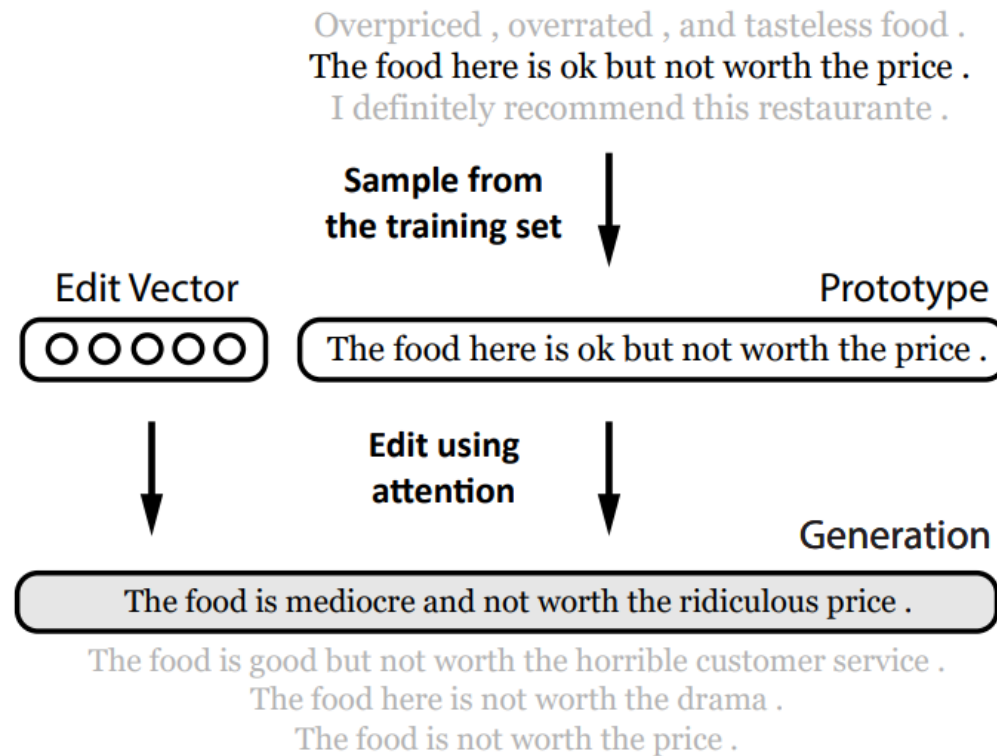
<Decay scheduels>



Scheduled Sampling for Sequence Prediction with Recurrent Neural Networks (Bengio et al., 2015)

- 특정확률로 이전 생성된 토큰을 다음 step의 입력으로 사용
- 학습이 진행될 수록 더 적은 gold token 을 사용함

Sequence re-writing



Generating Sentences by Editing Prototypes (Guu, Hashimoto et al., 2018)

- 학습데이터로 구축한 prototype set 을 text 생성에 사용
- 샘플링한 prototype 을 edit vector 를 사용해 변형
 - Adding
 - Removing
 - Modifying tokens

Reinforcement Learning

Text 생성모델을 Markov decision process 로 구성

- State : 이전 context 의 representation
- Actions : 현재 step 에서 생성될 수 있는 단어
- Policy : Decoder
- Rewards : score 함수로부터 받게 될 보상 (ex. BLEU, ROUGE, CIDEr, SPIDEr 등이 score 로 사용)

$$\mathcal{L}_{RL} = - \sum_{t=1}^T r(\hat{y}_t) \log P(\hat{y}_t | \{y^*\}; \{\hat{y}_t\}_{<t})$$

- Reward Estimation
 - 의도하지 않은 shortcut 을 모델이 학습하지 않도록 reward function 을 잘 정의해야 함



N-gram overlap metrics : BLEU

$$BLEU = \min\left(1, \frac{\text{output length}(\text{예측 문장})}{\text{reference length}(\text{실제 문장})}\right) \left(\prod_{i=1}^4 \text{precision}_i\right)^{\frac{1}{4}}$$

짧은 문장에 대한 penalty
N-gram precision 의 기하평균

- 실제 문장 대비 짧은 문장을 생성할 경우 패널티를 부여
- N-gram precision 의 기하평균 사용
- 일반적으로 많이 사용되는 metric

N-gram overlap metrics : ROUGE

ROUGE-N

$$= \frac{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} Count(gram_n)}$$

- N-gram recall 사용
- Brevity penalty 가 따로 없음
- BLEU 와는 달리 n-gram 별로 따로 비교함
(ex. Rouge-1,2,L)

N-gram overlap metrics

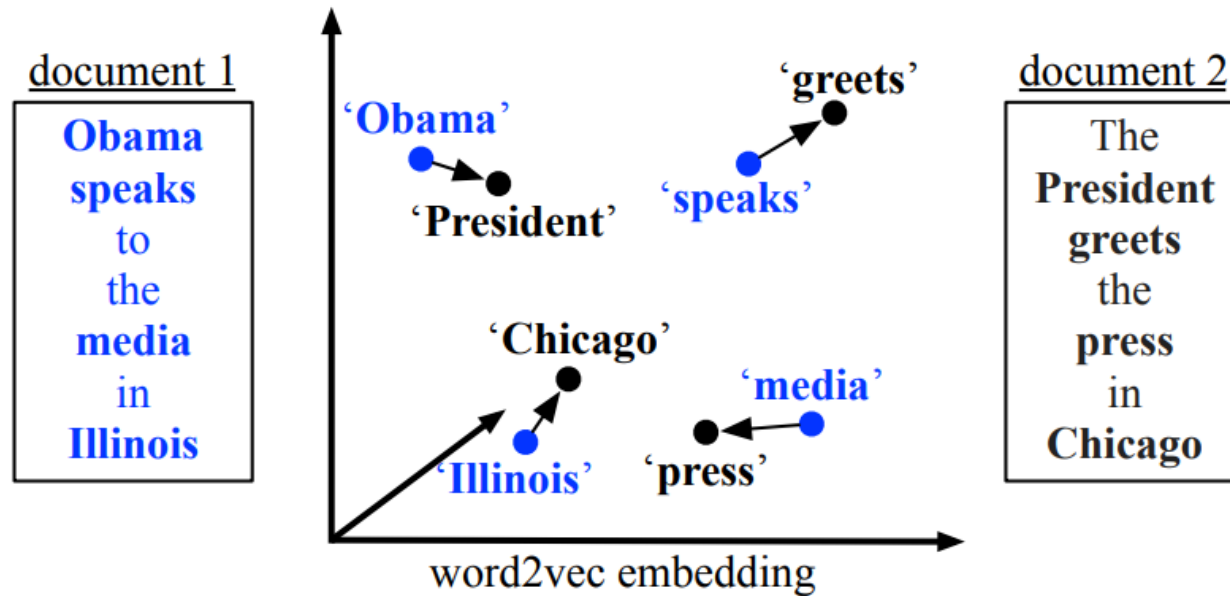
- Summarization output text 가 길 때, dialogue task 등 open-ended MT에서 적절한 지표가 될 수 없음
- 단어의 문맥적 의미를 반영할 수 없음
- 사람의 평가와 상관성이 낮음

⇒ Semantic overlap metrics, model-based metrics 연구들이 등장함

- PYRAMID, SPICE, SPIDER
- Vector Similarity
- Word Mover's Distance
- BERTSCORE
- Sentence Movers Similarity
- BLEURT

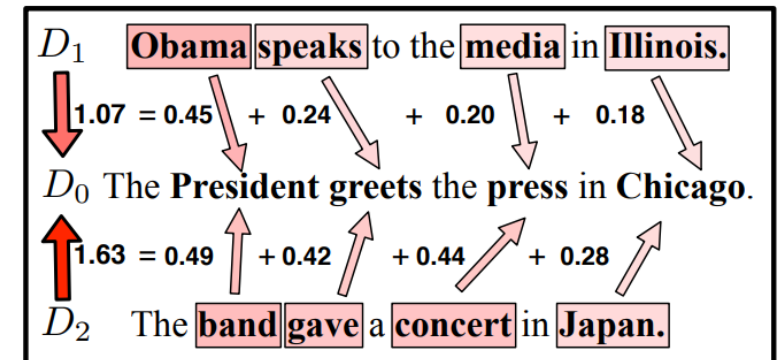
Word Mover's Distance

생성 text 와 reference text 의 단어 또는 문장의 semantic similarity 를 계산할 수 있음



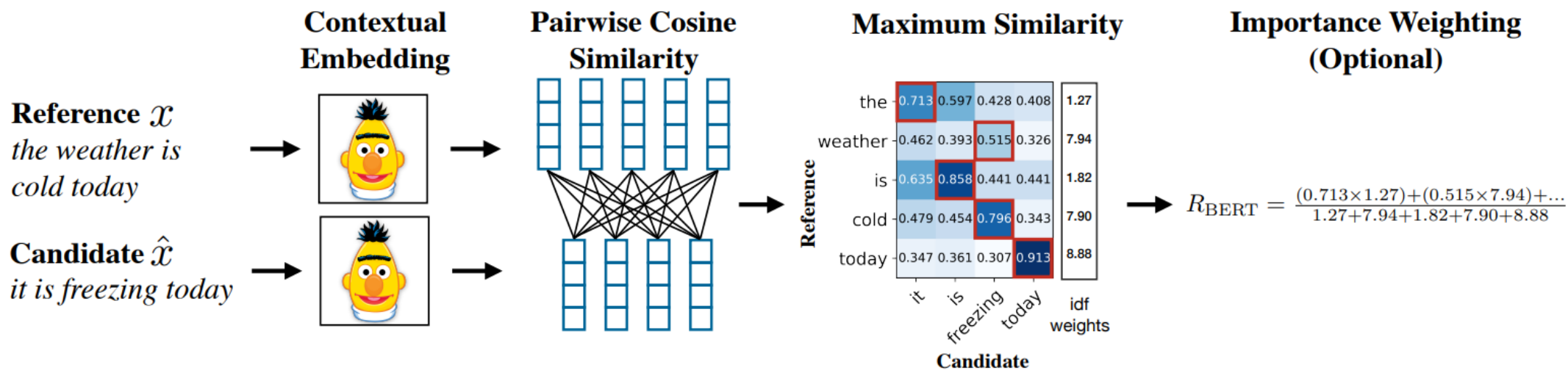
From Word Embeddings To Document Distances (Kusner et.al., 2015)

- 모든 단어의 word2vec embedding 계산
- Document 1의 모든 단어들의 document 2까지의 최소 거리 합으로 거리 계산



BERTSCORE

사전 학습 된 BERT 사용



BERTScore: Evaluating Text Generation with BERT (Zhang et.al, 2020)

<Recall metric>

- Reference 와 candidate 의 contextual embedding 계산
- 모든 pair 에 대해 cosine similarity 를 계산하고 greedy matching 후 weighted average 를 구함
- Inverse document frequency score 를 사용

$$\text{idf}(w) = -\log \frac{1}{M} \sum_{i=1}^M \mathbb{I}[w \in x^{(i)}]$$

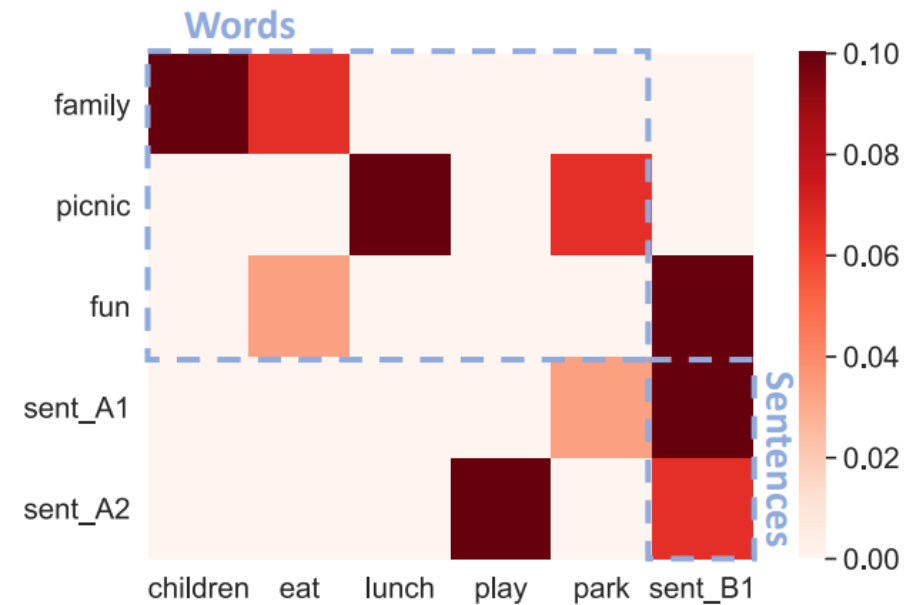
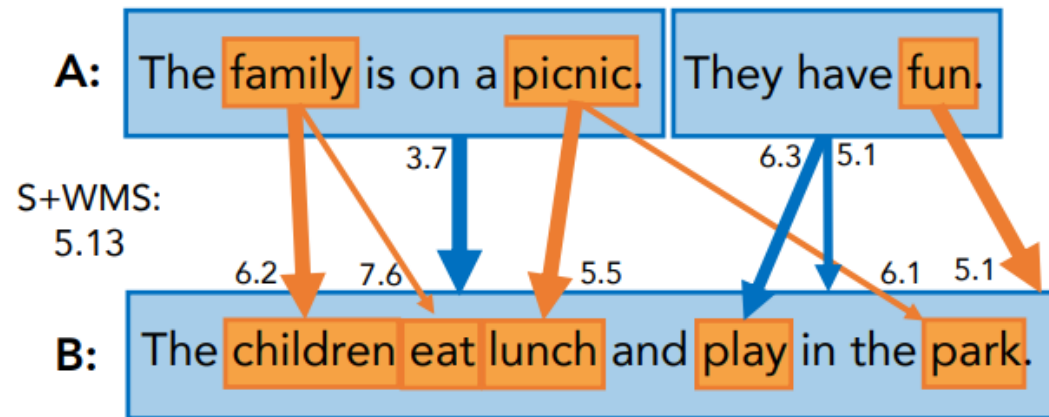
04

Evaluating

Model-based metrics

Sentence Mover's Similarity

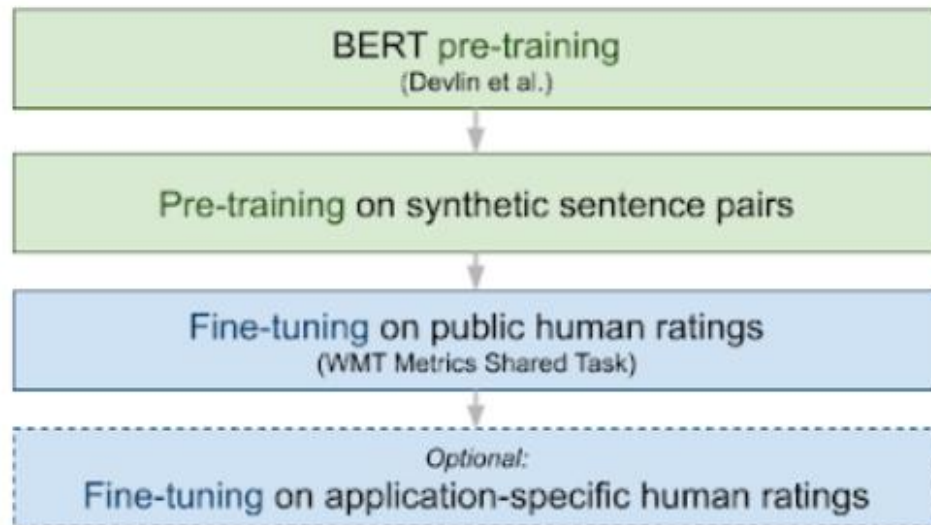
Sentence level, word level 에서 모두 similarity 계산



Sentence Mover's Similarity : Automatic Evaluation for Multi-Sentence Texts (Clark et.al., 2019)

BLEUART

문장 유사도를 예측하는 BERT 기반의 regression model 을 학습



	BLEU	ROUGE	...
Bud Powell was a legendary pianist. <i>Original sentence</i>			
Bud Powell is a famous pianist. <i>Random substitutions with BERT</i>	32.1	66.7	
Bud Powell was a piano legend. <i>Round-trip translation</i>	54.1	66.7	...
Bud Powell a legendary. <i>Random deletions</i>	31.7	55.7	

Collection of metrics and models used as pre-training targets.

<https://ai.googleblog.com/2020/05/evaluating-natural-language-generation.html>

BLEURT: Learning Robust Metrics for Text Generation (Sellam et.al. 2020)

1. NLG : Natural Language Generation
2. Decoding
 - Greedy methods (단어, 구가 반복되는 문장이 생성됨)
 - Other decoding algorithm : Random, Top-k, Top-p sampling
 - Re-balancing : kNN-LM, PPLM
3. Training
 - Unlikelihood training
 - Exposure bias → Scheduled sampling, Sequence re-writing, Reinforcement Learning
4. Evaluating
 - BLEU, Rouge score (Human evaluation 과 차이가 큼)
 - Model-based metric : WMD, BERTSCORE, SMS, BLEUART
5. ETC
 - Ethical considerations
 - Large scale 언어모델의 활용 가능성

감사합니다