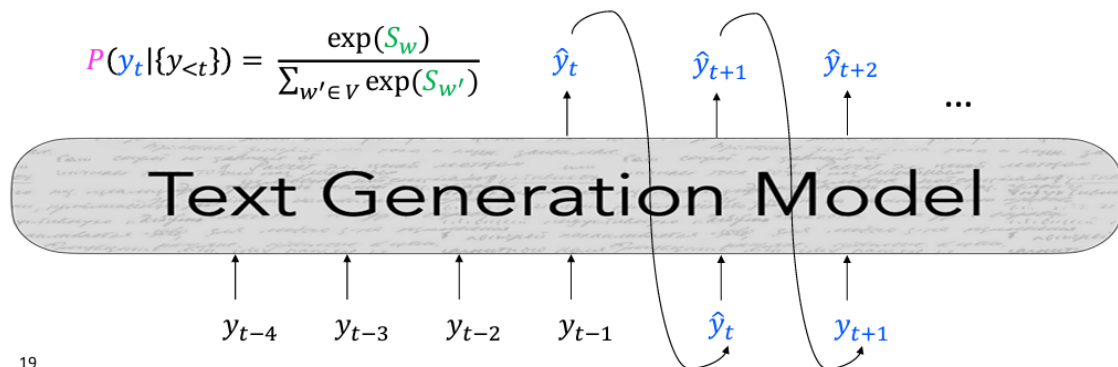


[CS224n] Lecture 12

Natural language generation(NLG)

- 논리적이고 유용한 텍스트를 생성하는 시스템
- 기계번역, 요약, 채팅, QA 등이 있음



19

$S = f(\{y_{<t}\}, \theta)$ 로 점수 측정

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

$\hat{y}_t = g(P(y_t | \{y_{<t}\}))$, g : decoding algorithm

Decoding

- 각각의 스텝 t마다, 단어의 각각의 토큰에 대해 점수를 계산하는 벡터 계산

$$S = f(\{y_{<t}\})$$

- 각각의 확률 P를 계산(softmax 방식 사용)

$$P(y_t = w | \{y_{<t}\}) = \frac{\exp(S_w)}{\sum_{w' \in V} \exp(S_{w'})}$$

- 디코딩 알고리즘이 분포중에서 토큰을 선택하기 위한 함수를 정의

$$\hat{y}_t = g(P(y_t | \{y_{<t}\}))$$

디코딩 방법

- Greedy methods

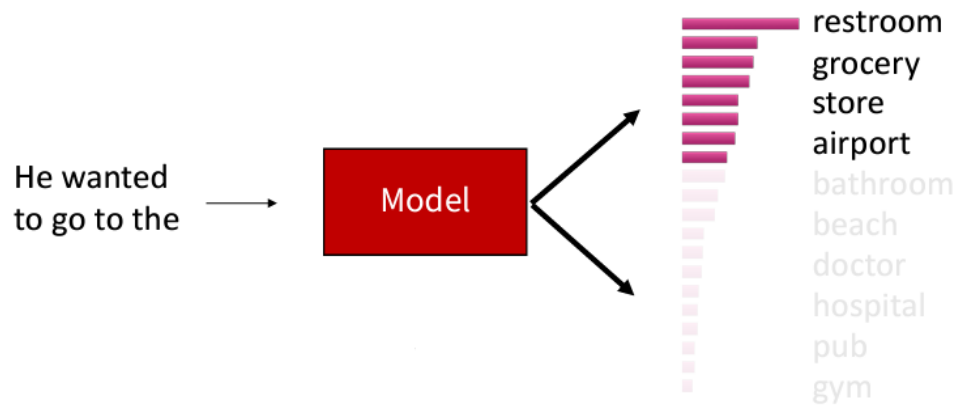
- 가장 확률이 높은 토큰을 선택
- $\hat{y}_t = \operatorname{argmax}_{w \in V} P(y_t = w | y_{<t})$
- 탐색하는 시간이 가장 빠르지만, 결과가 그리 좋지는 않음

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.

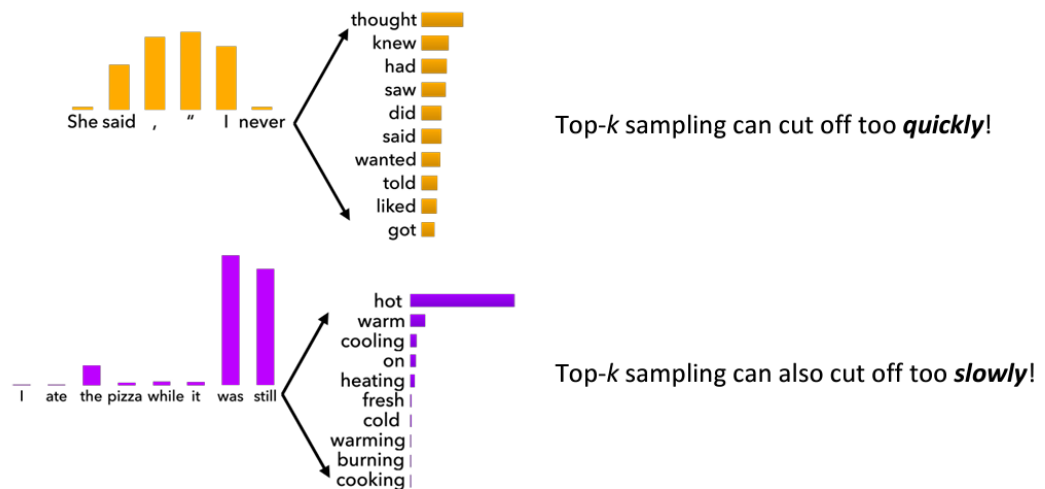
Continuation: The study, published in the Proceedings of the National Academy of Sciences of the United States of America (PNAS), was conducted by researchers from the **Universidad Nacional Autónoma de México (UNAM)** and **the Universidad Nacional Autónoma de México (UNAM/Universidad Nacional Autónoma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

- 반복이 계속해서 일어나는 문제가 발생
→ n-gram을 반복하지 않거나, coverage loss 등의 방법 등으로 해결

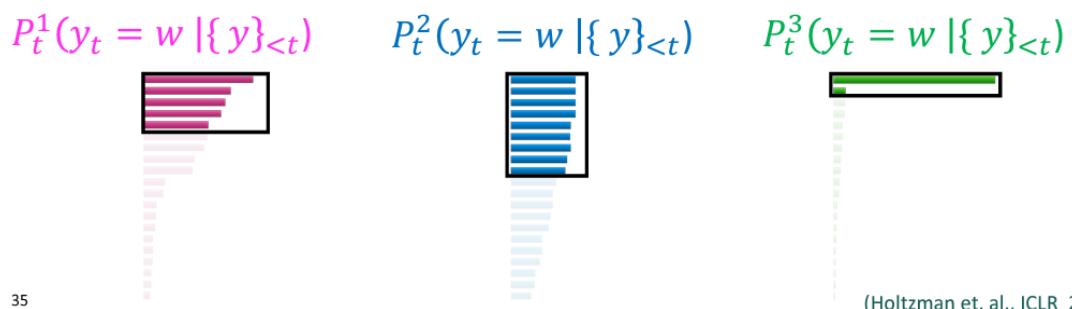
- Sampling-based Decoding
 - 랜덤으로 아무 토큰이나 선택하는 방법
 - Vanilla sampling은 모든 토큰을 선택지로 두기 때문에, 문맥에 맞지 않는 경우가 많음
 - 그래서 Top-k sampling을 활용 : k개의 가장 확률이 높은 토큰 중 랜덤추출



(k = 5인 예시)



- Top-k sampling은 너무 빠르거나, 반대로 너무 느리게 끝날 가능성이 존재
 - 확률분포가 균등할 때, 제한된 k는 많은 가능한 옵션을 제거
 - 확률분포가 치우쳐질 때, 높은 수의 k는 너무 많은 옵션을 제공
- Top-p sampling : 가장 누적분포 양이 많은 토큰 중 p개를 추출



- Scaling 무작위화를 위해 softmax temperature 적용
 - temperature 하이퍼파라미터를 적용

$$P_t(y_t = w) = \frac{\exp(S_w/\tau)}{\sum_{w' \in V} \exp(S_{w'}/\tau)}$$
 - $t > 1$ 이면 P가 균등해지고, $t < 1$ 이면 P는 spiky해짐
- 만약 디코딩 중에 잘못된 시퀀스를 디코딩 했다면?
 - 여러개의 시퀀스를 디코딩함
 - 이 시퀀스의 질을 대략적으로 추측하는 점수를 정의한 다음에, 이 점수를 바탕으로 re-rank
 - 가장 간단한 점수는 perplexity의 측정

Training NLG models

- Unlikelihood training

$$L_{UL}^t = - \sum_{y_{neg} \in e} \log(1 - P(y_{neg} | \{y^*\}_{<t}))$$


$$L_{MLE}^t = -\log P(y_t^* | \{y^*\}_{<t}) : \text{Teacher forcing}$$

$$L_{ULE}^t = L_{MLE}^t + \alpha L_{UL}^t$$
 - teacher forcing 방법으로 학습하면 편향에 노출될 수 있음
 - training때의 비용함수는 y^* 값에 대해 계산되지만, generation때는 \hat{y} 의 값에 대해 계산되기 때문
 - 해결책 : Scheduled sampling, Dataset aggregation, Sequence re-writing, 강화학습 등
 - Sequence re-writing
 - human-written 데이터에서부터 시퀀스를 찾게 학습
 - 찾은 시퀀스를 추가, 제거, 수정 등을 통해 다듬게 학습
 - 강화학습
 - BLEU, ROUGE 등을 통해 점수를 계산하며 강화학습 진행
 - 여전히 주로 사용되는 방법

Evaluation NLG models

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

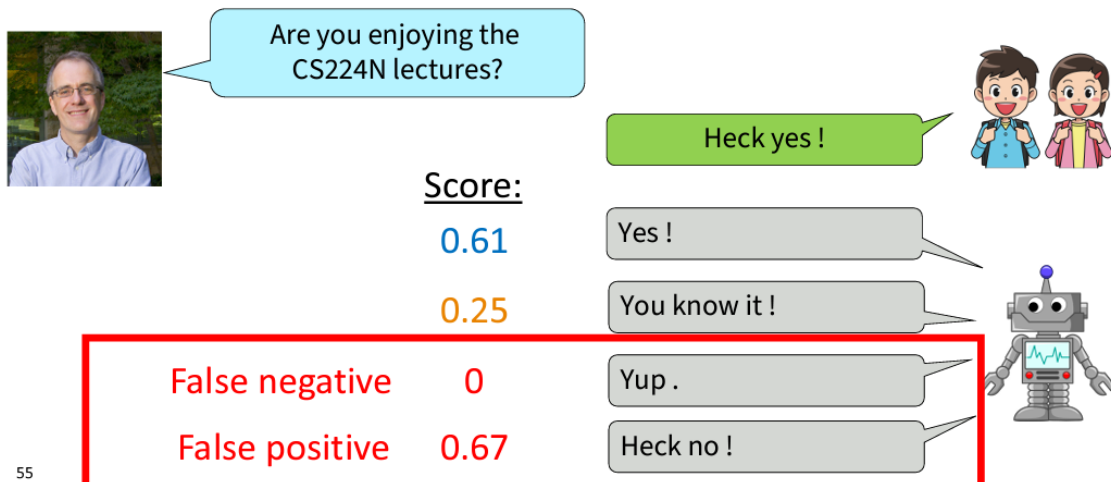
Gen: **The woman went** **to the hardware store** .



- generated 와 gold-standard(human-written)의 유사성을 나타내는 점수 계산
- 2개의 종류 존재
 - N-gram overlap metrics
 - Semantic overlap metrics

N-gram overlap metrics

- BLEU, ROUGE, METEOR, CIDEr 등
- Machine Translation에 이상적이지 않음(특히 개방형 대답에 매우 안 좋음)



Are you enjoying the CS224N lectures?

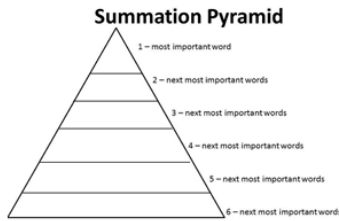
Heck yes !

Score:

0.61	Yes !
0.25	You know it !
False negative 0	Yup .
False positive 0.67	Heck no !

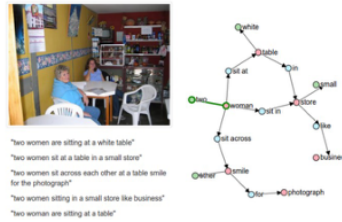
55

Semantic overlap metrics



PYRAMID:

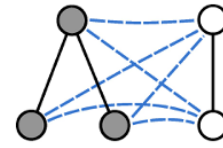
- Incorporates human content selection variation in summarization evaluation.
- Identifies **Summarization Content Units (SCU)s** to compare information content in summaries.



SPICE:

Semantic propositional image caption evaluation is an image captioning metric that initially parses the reference text to derive an abstract scene graph representation.

(Anderson et al., 2016).



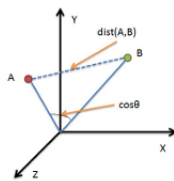
SPIDER:

A combination of semantic graph similarity (SPICE) and n -gram similarity measure (CIDER), the SPICE metric yields a more complete quality evaluation metric.

(Liu et al., 2017)

Model-based metrics

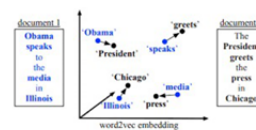
- gen과 ref의 유사성을 측정하는데 '학습된 representation'을 활용
- text 유닛이 임베딩으로 활용되어 n-gram bottleneck 문제 없음
- Word distance functions



Vector Similarity

Embedding based similarity for semantic distance between text.

- **Embedding Average** (Liu et al., 2016)
- **Vector Extrema** (Liu et al., 2016)
- **MEANT** (Lo, 2017)
- **YISI** (Lo, 2019)



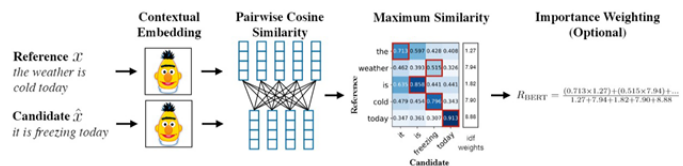
Word Mover's Distance

Measures the distance between two sequences (e.g., sentences, paragraphs, etc.), using word embedding similarity matching.

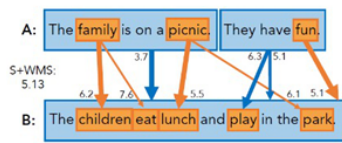
(Kusner et al., 2015; Zhao et al., 2019)

BERTSCORE

Uses pre-trained contextual embeddings from BERT and matches words in candidate and reference sentences by cosine similarity. (Zhang et al. 2020)



- Beyond word matching



Sentence Movers Similarity :

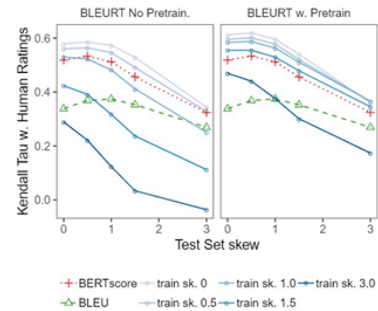
Based on Word Movers Distance to evaluate text in a continuous space using sentence embeddings from recurrent neural network representations.

(Clark et.al., 2019)

BLEURT:

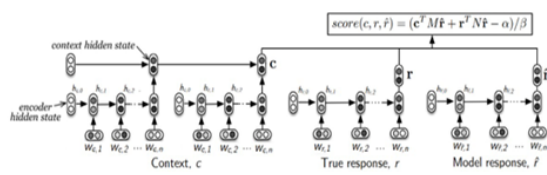
A regression model based on BERT returns a score that indicates to what extent the candidate text is grammatical and conveys the meaning of the reference text.

(Sellam et.al. 2020)



Human evaluation

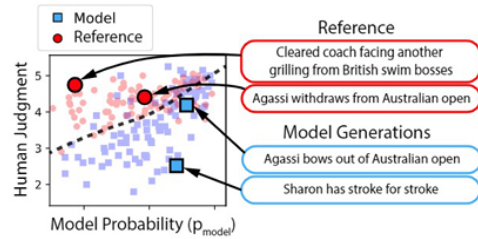
- automatic metrics가 사람의 평가와 잘 들어맞지 않음
- 사람의 평가에 존재하는 단점
 - 느리고 비쌈
 - 지속적이지 않음
 - 비논리적일 수 있음
 - 집중을 놓칠 수 있음
 - 질문을 잘못 해석할 수 있음
 - 그들이 그렇게 선택한 것에 대해 항상 설명할 수 없음
- 그래서 사람의 피드백을 바탕으로 한 모델 탄생



ADEM:

A learned metric from human judgments for dialog system evaluation in a chatbot setting.

(Lowe et.al., 2017)



HUSE:

Human Unified with Statistical Evaluation (HUSE), determines the similarity of the output distribution and a human reference distribution.

(Hashimoto et.al. 2019)

NLG의 윤리

- 2016년 MS에서 개발한 챗봇인 Tay의 경우, 개시한지 24시간 동안 인종차별적이고, 성희롱적인 채팅을 학습해 안좋은 방향으로 나아감
- 이러한 모델은 안좋은 목적을 지닌 사용자에 의해 오염될 가능성이 무수히 많음
- pretrained language model은 겉보기엔 무해한 텍스트에서도 안좋은 영향을 받을 가능성 존재
- 적절한 세이프가드 없이는 배포되어선 안됨