자연어 처리에서의
모델 성능 평가 방법
(PPL)

24.04.04
유하영

언어 모델의 성능평가를 위한 지표

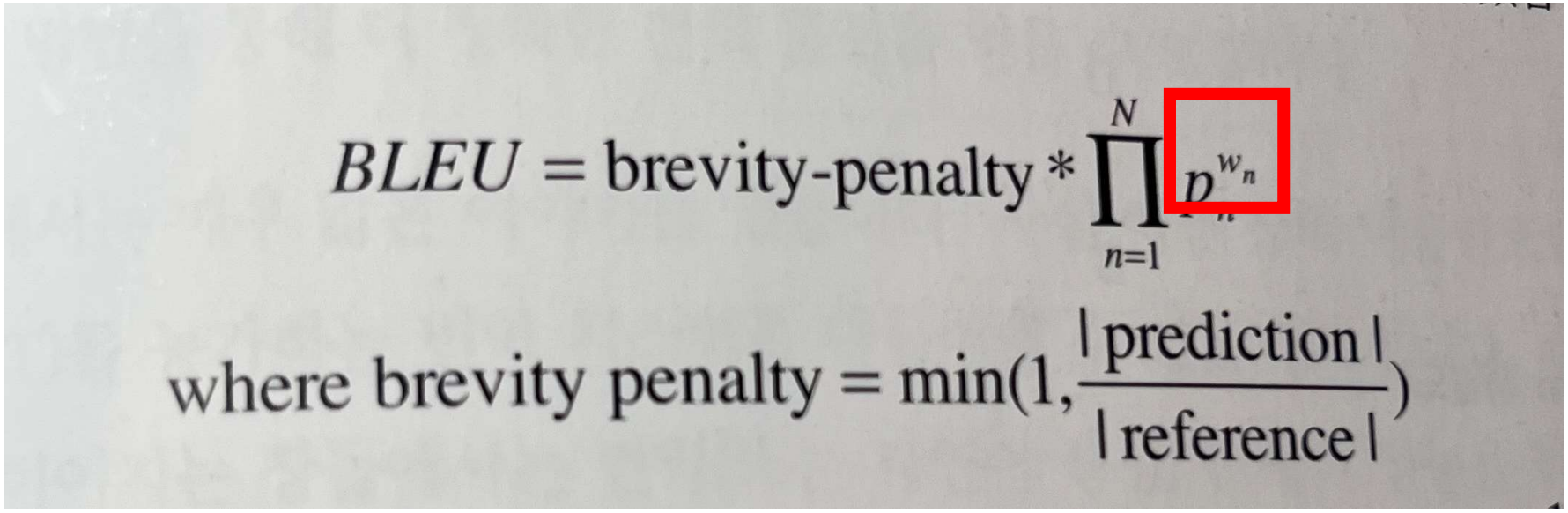
정성적 평가 사람이 직접 번역된 문장을 채점하는 형태

정량적 평가

- BLEU score
  : 생성된 텍스트와 사람이 번역한 텍스트 간의 유사성을 측정하여,
  생성된 텍스트의 품질 평가. (어휘적 일치에 초점)
  기계 번역 모델의 성능을 객관적으로 비교하는 데 유용
- ROUGE
- PPL(Perplexity, 혼란도)
  : 언어 모델의 일반적인 성능을 평가할 때 사용 (모델의 예측 불확실성 측정)

BLEU score 계산

1. 단어의 순서 고려 (n-gram)
2. 같은 단어가 연속적으로 나올 때 과적합되는 현상 보정 (Clipping, 정밀도 계산)
3. 문장길이에 대한 과적합 보정 (Brevity Penalty;BP)
4. 종합 BLEU 점수 계산



$$BLEU = \text{brevity-penalty} * \prod_{n=1}^{N} p^{w_n}$$

$$\text{where brevity penalty} = \min\left(1, \frac{|\text{prediction}|}{|\text{reference}|}\right)$$

문제 1) 단어의 순서를 고려하지 않음

1. n-gram으로 순서쌍들이 얼마나 겹치는지 측정

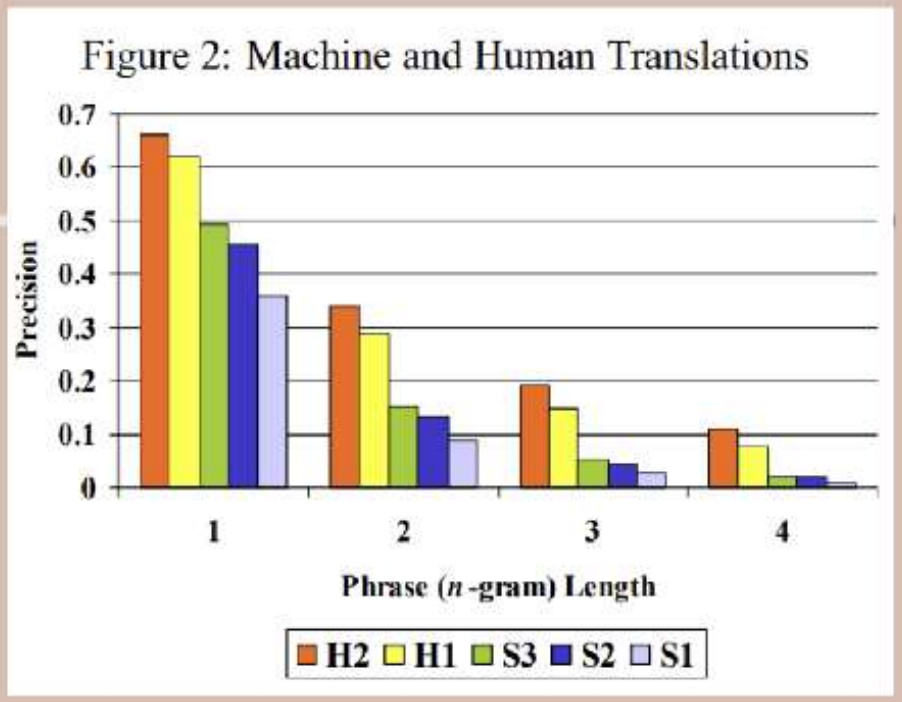
- 1-gram precision: $\frac{\text{일치하는} 1-gram\text{의 수(예측된 } sentence\text{중에서)}}{\text{모든} 1-gram\text{쌍 (예측된 } sentence\text{중에서)}} = \frac{10}{14}$
- 2-gram precision: $\frac{\text{일치하는} 2-gram\text{의 수(예측된 } sentence\text{중에서)}}{\text{모든} 2-gram\text{쌍 (예측된 } sentence\text{중에서)}} = \frac{5}{13}$
- 3-gram precision: $\frac{\text{일치하는} 3-gram\text{의 수(예측된 } sentence\text{중에서)}}{\text{모든} 3-gram\text{쌍 (예측된 } sentence\text{중에서)}} = \frac{2}{12}$
- 4-gram precision: $\frac{\text{일치하는} 4-gram\text{의 수(예측된 } sentence\text{중에서)}}{\text{모든} 4-gram\text{쌍 (예측된 } sentence\text{중에서)}} = \frac{1}{11}$

$$(\prod_{i=1}^{4} precision_i)^{\frac{1}{4}} = (\frac{10}{14} \times \frac{5}{13} \times \frac{2}{12} \times \frac{1}{11})^{\frac{1}{4}}$$

BLEU SCORE

"The main idea is to use a weighted average of variable length phrase matches against the reference translations."

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$



왜 Weighted Average?

n마다 내포하는 의미가 다르다. n=1 → 단어의 적절성, n >1 → 어순, 유창함

따라서, 각 n-gram precision의 중요도를 동등하게 보기 위해 Weighted Average 사용한다.

언어 모델의 성능평가를 위한 지표

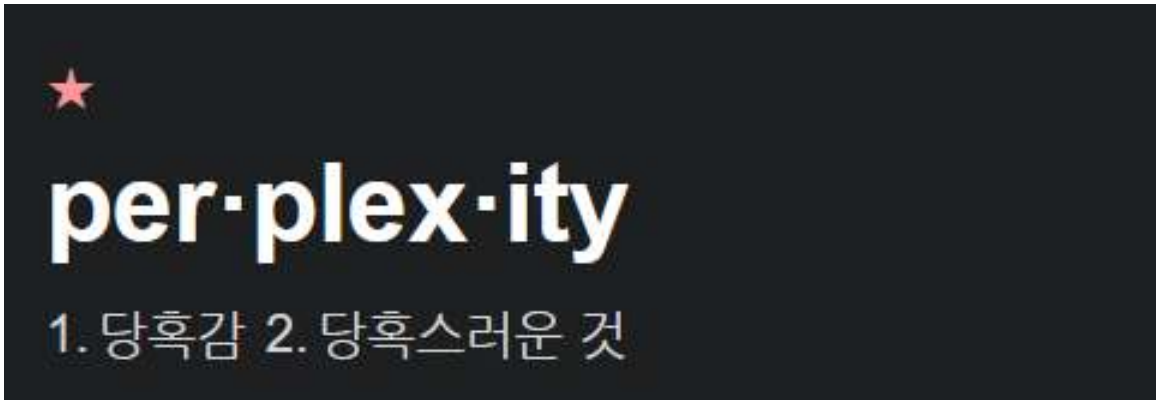
정성적 평가 사람이 직접 번역된 문장을 채점하는 형태

정량적 평가

- BLEU score
  : 생성된 텍스트와 사람이 번역한 텍스트 간의 유사성을 측정하여,
  생성된 텍스트의 품질 평가. (어휘적 일치에 초점)
  기계 번역 모델의 성능을 객관적으로 비교하는 데 유용
- ROUGE
- PPL(Perplexity, 혼란도)
  : 언어 모델의 일반적인 성능을 평가할 때 사용 (모델의 예측 불확실성 측정)

PPL (Perplexity)

: 모델이 정답을 결정할 때, 얼마나 헷갈렸는가를 나타내는 지표

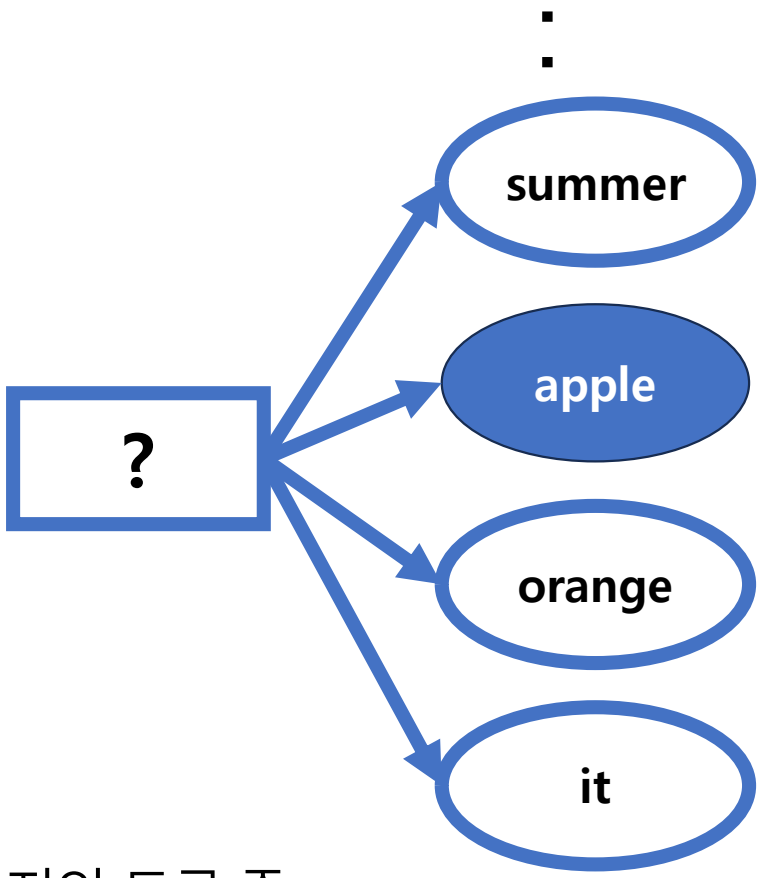


당혹감, 혼란도, 헷갈림

PPL (Perplexity)

문장 생성의 경우(디코딩)

I like ?



수만가지의 토큰 중

1. 가장 확률이 높은 **단 한 개의 토큰을 선택**(greedy decoding)
2. 여러 개의 토큰이 가장 높은 확률을 가질 때, 이 중 하나를 **무작위로 선택** (random sampling)

PPL (Perplexity)

: 모델이 정답을 결정할 때, 얼마나 헷갈렸는가를 나타내는 지표

즉, **언어모델이 특정시점에서 얼마나 많은 불확실성을 가지고 있는지를 나타내는 지표.**
(선택지의 수가 얼마나 많은 지)

EX)

PPL = 10

언어모델이 평균적으로 각 단어를 예측할 때,
10개의 거의 동일한 확률을 가진 선택지 사이에서 선택한다.

따라서, PPL ↓ → 모델 예측 정확성 ↑

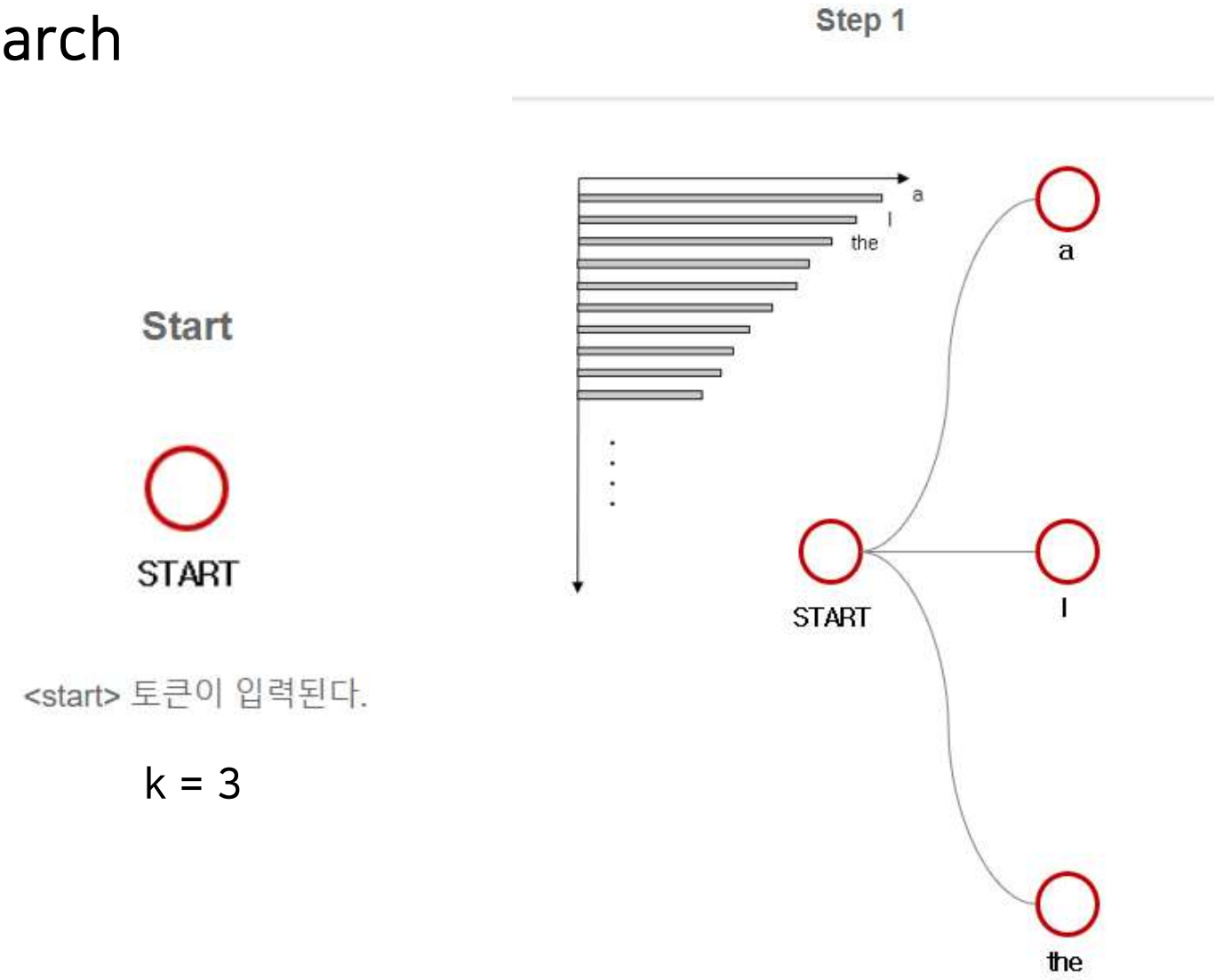
** Beam search

: 번역이나 텍스트 생성 과정에서 최적의 출력 시퀀스를 찾기 위해 사용

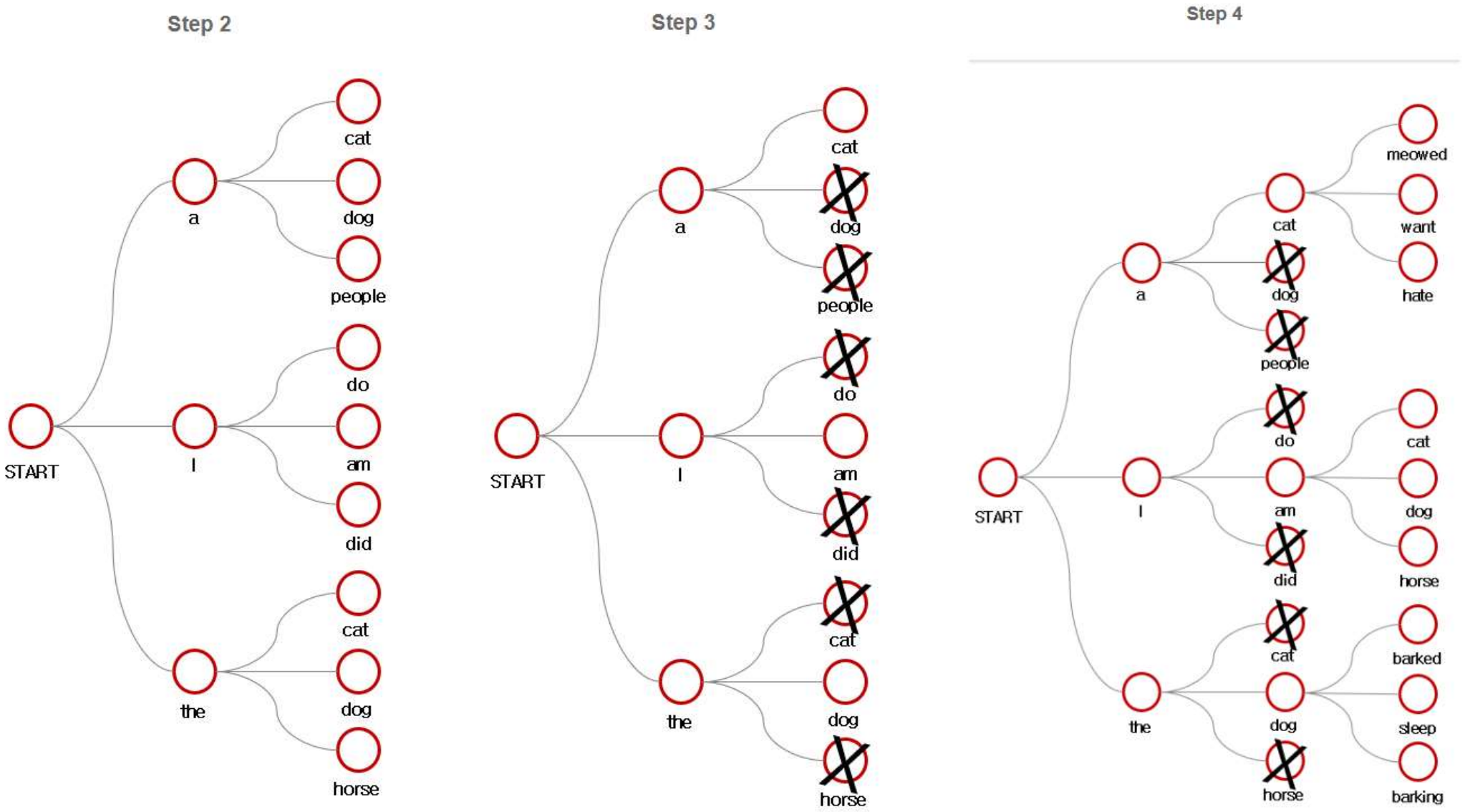
> Beam size : 탐색과정에서 **한 번에 고려하게 되는 후보 시퀀스의 수**

Beam size가 크면 더 많은 후보 시퀀스를 고려할 수 있어. 탐색이 보다 포괄적으로 됨

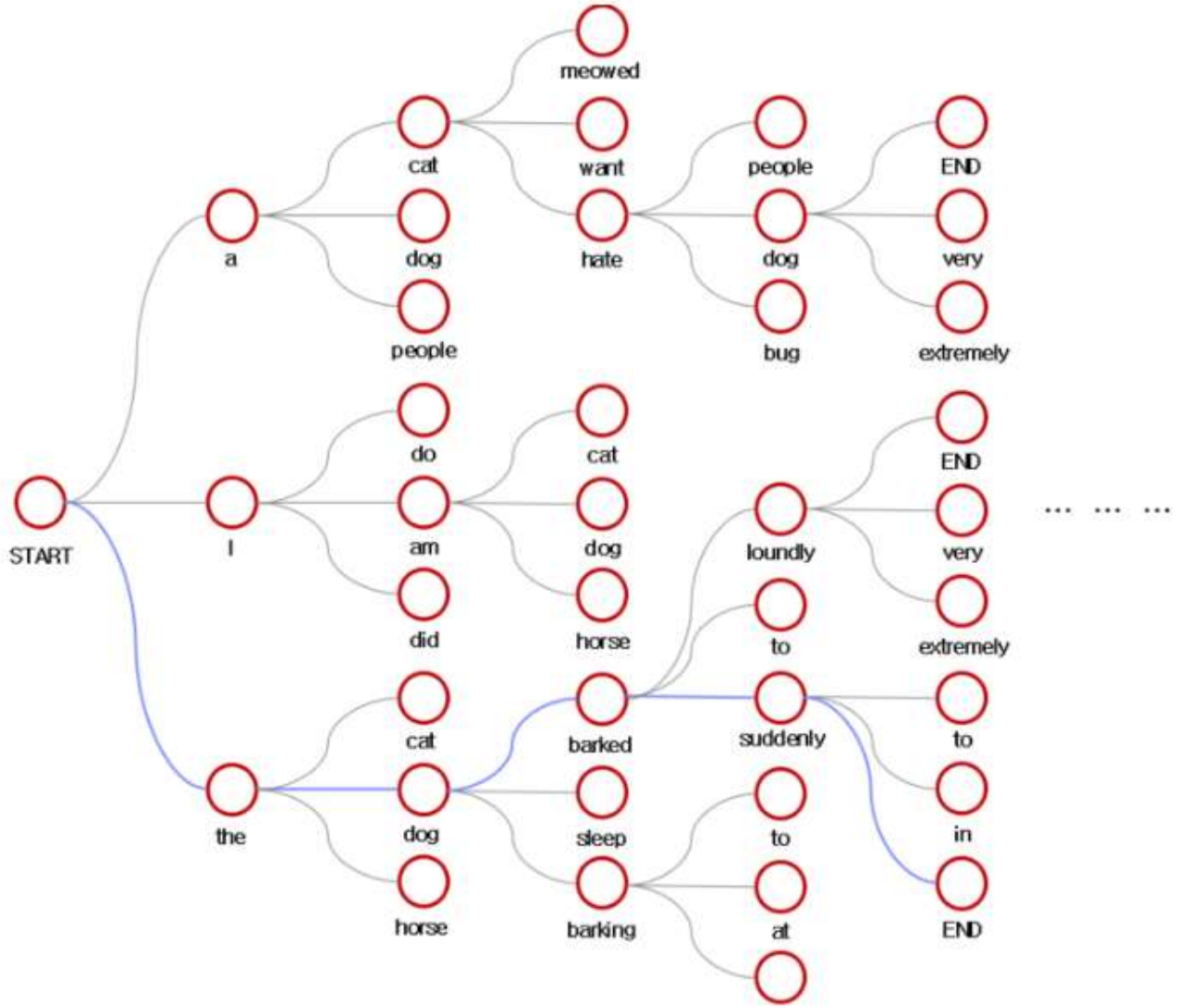
** Beam search



Experimental Results



Experimental Results



CANDIDATES

1. The dog barked suddenly
2. A cat hate dog
3. The dog barking to me

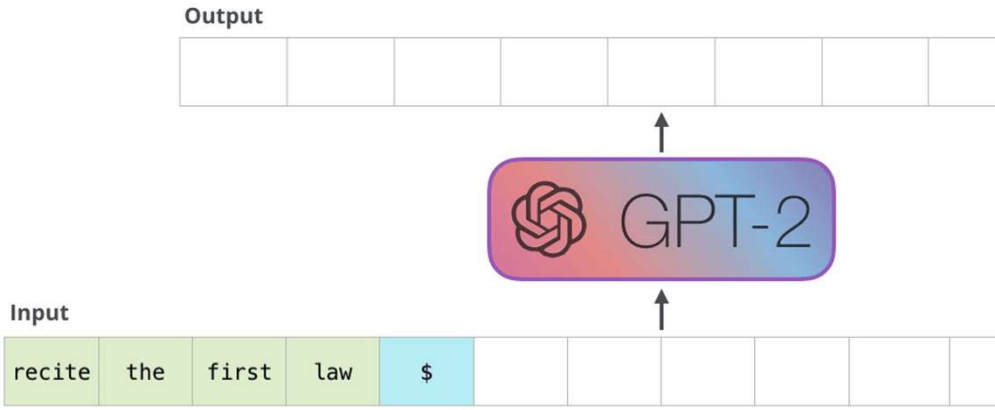
PPL (Perplexity)

$p(W)$

: 어떤 문장 W가 나타날 확률

≒ 언어 모델이 문장 W를 생성할 확률

즉, 개별 토큰 $w_1, w_2, \cdots w_N$ 이 순서대로(연쇄적으로) 나타날 조건부 확률의 곱



$$p(W) = p(w_1, w_2, \cdots w_N)$$

$$= \prod_{i=1}^{N} p(w_i)$$

PPL (Perplexity)

$$PPL(W) = P(w_1, w_2, w_3, \ldots, w_N)^{-\frac{1}{N}} = \sqrt[N]{\frac{1}{P(w_1, w_2, w_3, \ldots, w_N)}}$$

PPL (Perplexity)

$$p(W) = p(w_{1,}w_{2,}\cdots w_N) = \prod_{i=1}^{N} p(w_\mathrm{i})$$

$$\ln(P(w)) = \ln(\prod_{i=1}^{N} p(w_\mathrm{i}))$$

$$-\frac{1}{N}\ln(P(w)) = -\frac{1}{N}\sum_{i=1}^{N} \ln(p(w_\mathrm{i}))$$

$$e^{-\frac{1}{N}\ln(P(w))} = e^{-\frac{1}{N}\sum_{i=1}^{N}\ln(p(w_\mathrm{i}))}$$

$$PPL(w) = \sqrt[N]{\frac{1}{\prod_{i=1}^{N} p(w_\mathrm{i})}}$$

정규화(N)

: 각 문장들의 길이에 대한 조정

문장 A: 나는 학교에 간다.
문장 B: 나는 아침을 먹고
버스를 타고 학교에 간다.

$$p(A) > p(B)$$

	N	d_{model}	d_{ff}	h	d_k	d_v	P_{drop}	ϵ_{ls}	train steps	PPL (dev)	BLEU (dev)	params $\times 10^6$
base	6	512	2048	8	64	64	0.1	0.1	100K	4.92	25.8	65
(A)				1	512	512				5.29	24.9	
				4	128	128				5.00	25.5	
				16	32	32				4.91	25.8	
				32	16	16				5.01	25.4	
(B)					16					5.16	25.1	58
					32					5.01	25.4	60
(C)	2									6.11	23.7	36
	4									5.19	25.3	50
	8									4.88	25.5	80
		256			32	32				5.75	24.5	28
		1024			128	128				4.66	26.0	168
			1024							5.12	25.4	53
			4096							4.75	26.2	90
(D)							0.0			5.77	24.6	
							0.2			4.95	25.5	
								0.0		4.67	25.3	
								0.2		5.47	25.7	
(E)	positional embedding instead of sinusoids									4.92	25.7	
big	6	1024	4096	16			0.3		300K	**4.33**	**26.4**	213