자연어 처리에서의
모델 성능 평가 방법
(BLEU score)

24.03.28
유하영

	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

영어

I am hungry.

아이 앰 헝그리.

12 / 3000

번역하기

한국어

높임말

나는 배고파요.
나는 배고프
다.
나 배고파.
나 배고픈데

번역 수정 | 번역 평가

텍스트 생성 모델의 성능 평가하기

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 (bilingual Evaluation Understudy)

: 기계 번역 결과와 사람이 직접 번역한 결과가 얼마나 유사한지 비교하여
번역에 대한 성능을 측정하는 방법

[번역기]

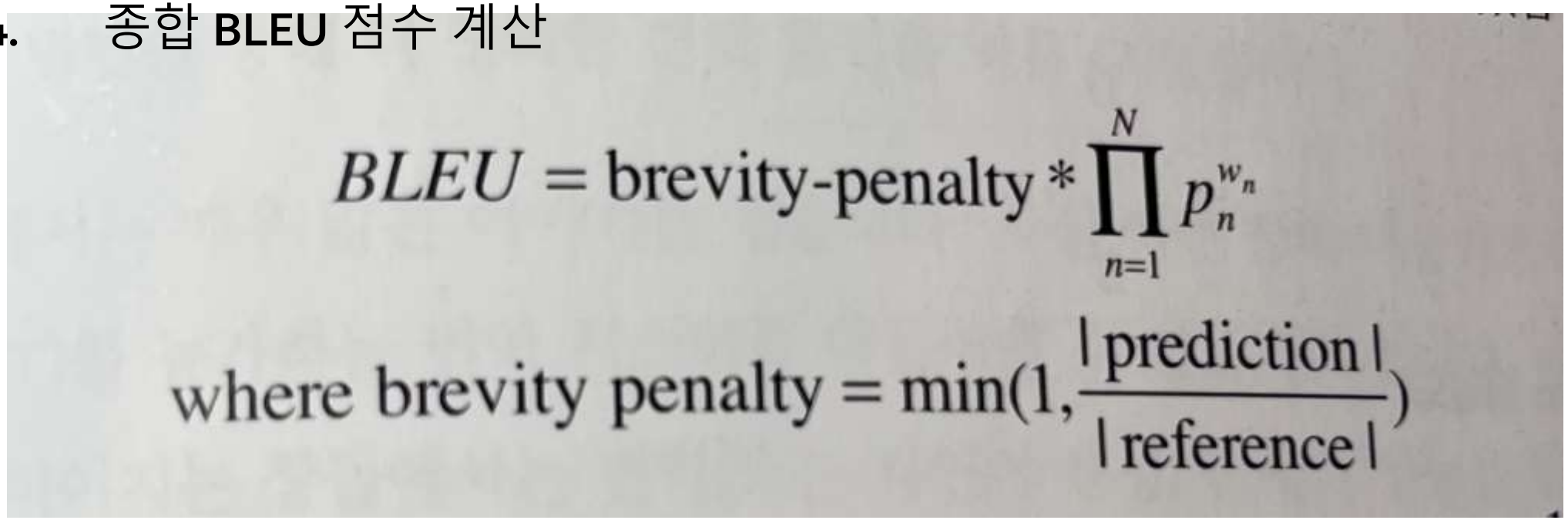
(원본 문장) (번역된 문장)

영어 문장 -> 한글 문장

generated sentence : 번역한 문장
reference sentence : 정답 문장
re1
re2
re3 .....

BLEU score 계산

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



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

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

0. 단어 개수를 count하기

: Re 1,2,3 중 어느 한 곳에서라도 등장한 단어를 Ge에서 그 개수를 센다.
(유니그램 정밀도(Unigram Precision))

(Generated sentence)

It is a guide to action which ensures that the military always obeys the commands of the party.

(Reference sentence)

•Reference1 : It is a guide to action that ensures that the military will forever heed Party commands.

•Reference2 : It is the guiding principle which guarantees the military forces always being under the command of the Party.

•Reference3 : It is the practical guide for the army always to heed the directions of the party.

0. 단어 개수를 count하기

: Re 1,2,3 중 어느 한 곳에서라도 등장한 단어를 **Ge**에서 그 개수를 센다.
(유니그램 정밀도(Unigram Precision))

$$\text{Unigram Precision} = \frac{\text{Ref들 중에서 존재하는 Ca의 단어의 수}}{\text{Ca의 총 단어 수}}$$

$$\text{Ca1 Unigram Precision} = \frac{17}{18}$$

문제 – 1) 단어의 순서를 고려하지 않음
2) 같은 단어가 연속적으로 나올 때 문제 발생

문제 2) 같은 단어가 연속적으로 나올 때 문제 발생

(Generated sentence)

•Candidate : **The The The the The The the The The**

(Reference sentence)

•Reference1 : **The** more **the** merrier I always say

문제 1) 단어의 순서를 고려하지 않음

1. n-gram으로 순서쌍들이 얼마나 겹치는지 측정

(보통 n은 1~4)

(Generated sentence)

•Candidate : **빛을 쐬는** 노인은 **완벽한** 어두운곳에서 **잠든 사람과 비교할 때** 강박증이 **심해질** 기회가 **훨씬 높았다**

(Reference sentence)

•Reference1 : **빛을 쐬는** 사람은 **완벽한** 어둠에서 **잠든 사람과 비교할 때** 우울증이 **심해질** 가능성이 **훨씬 높았다**

n-gram ?

: 코퍼스에서 n개의 단어 뭉치 단위로 끊어서 이를 하나의 토큰으로 간주.

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: $\frac{\text{일치하는} 4-gram\text{의 수(예측된 } sentence\text{중에서)}}{\text{모든} 4-gram\text{쌍 (예측된 } sentence\text{중에서)}} = \frac{1}{11}$

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

문제 2) 같은 단어가 연속적으로 나올 때 문제 발생(특히, 1-gram일 경우!)

2. 같은 단어가 연속적으로 나올때 과적합되는 현상 보정 (Clipping)

(Generated sentence)

•Candidate : **The more** decomposition **the more** flavor **the** food has

(Reference sentence)

•Reference1 : **The more the** merrier I always say

the : 3 , more : 2 → **the : 2 , more : 1**

$$\frac{\text{일치하는} 1\text{-}gram\text{의 수(예측된 } sentence\text{중에서)}}{\text{모든} 1\text{-}gram\text{쌍 (예측된 } sentence\text{중에서)}} = \frac{5}{9}$$

$$\frac{\text{일치하는} 1\text{-}gram\text{의 수(예측된 } sentence\text{중에서)}}{\text{모든} 1\text{-}gram\text{쌍 (예측된 } sentence\text{중에서)}} = \frac{3}{9}$$

3. 문장길이에 대한 과적합 보정 (Brevity Penalty)

(Generated sentence)

•Candidate : **빛을 쐬는 노인은 완벽한 어둠에서 잠든** 사람이 또 있을까

(Reference sentence)

•Reference1 : **빛을 쐬는 노인은 완벽한 어둠에서 잠든** 사람과 비교할 때 우울증이 심해질 가능성이 훨씬 높았다

$$min(1, \frac{\text{예측된 } sentence\text{의 길이(단어의 갯수)}}{true\ sentence\text{의 길이(단어의 갯수)}}) = min(1, \frac{6}{14}) = \frac{3}{7}$$

4. 종합 BLEU 점수 계산

전체 테스트 데이터셋을 하나의 큰 문장으로 취급

1) n-gram 일치도 계산

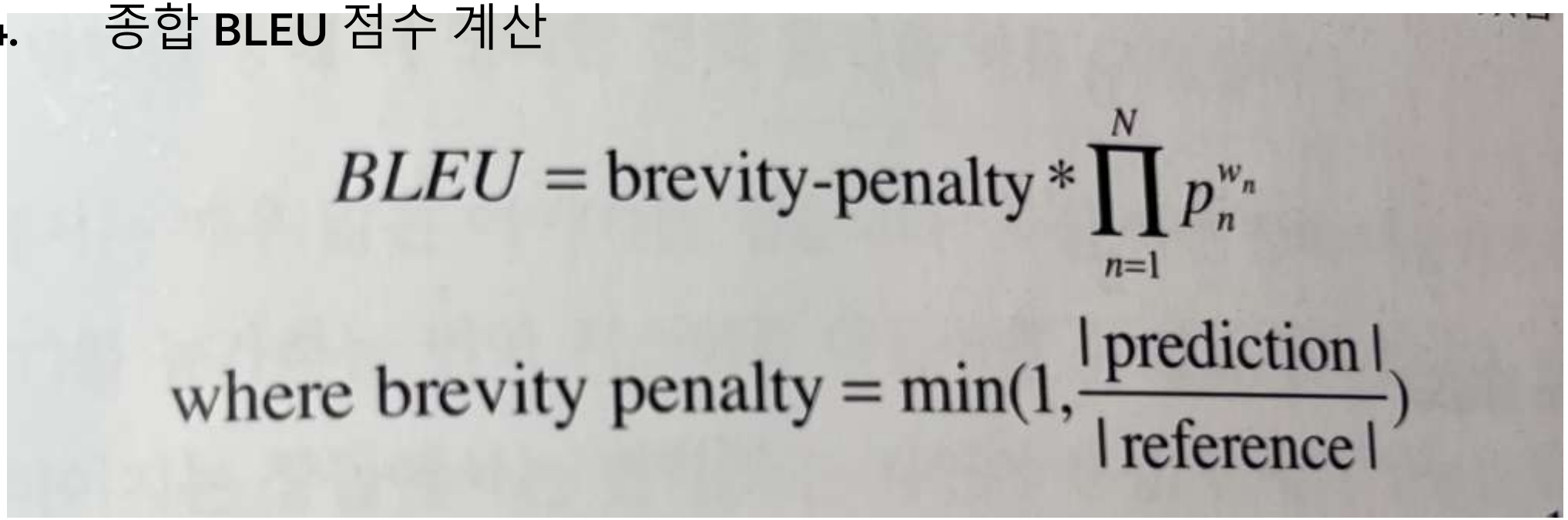
2) 1) 을 기하 평균하여 종합 점수 도출 $(\prod_{i=1}^{4} precision_i)^{\frac{1}{4}}$

3) BP 적용

4) BLEU score이 **0~1**사이의 값으로 나타남**(or** 백분율로 표시**)**

BLEU score (bilingual Evaluation Understudy) 계산

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



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

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

최종 BLEU score

(Generated sentence)

•Candidate : **빛을 쐬는** 노인은 **완벽한** 어두운곳에서 **잠든 사람과 비교할 때** 강박증이 **심해질** 기회가 **훨씬 높았다**

(Reference sentence)

•Reference1 : **빛을 쐬는** 사람은 **완벽한** 어둠에서 **잠든 사람과 비교할 때** 우울증이 **심해질** 가능성이 **훨씬 높았다**

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

$$= min(1, \frac{14}{14}) \times (\frac{10}{14} \times \frac{5}{13} \times \frac{2}{12} \times \frac{1}{11})^{\frac{1}{4}}$$