BLEU: A METHOD FOR AUTOMATIC EVALUATION OF MACHINE TRANSLATION

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu

목차

연구 배경

제일 정확한 방법: Human Evaluation

- 인간이 직접 번역 결과를 평가하는 것
- 정확성과 신뢰성이 높다
- 하지만 비용, 시간을 많이 소모한다
- 고려해야 할 요소들이 많다 (adequacy, fidelity, fluency)

연구 전제와 평가 지표 정의

좋은 번역기란 무엇인가

전문적인 번역가가 한 결과와 비슷할수록 좋은 것

비슷함의 기준: WER (Word Error Rate)

- 음성인식에서 사용되는 정확성 평가 지표
- 음성으로부터 인식되는 문장과 실제 문장을 단어 단위로 비교하여 음성인식 성능을 평가

Ex)

실제 문장: 만두를 간장에 찍어먹었다

인식된 문장: 만두를 긴장에 찍어먹었다

WER = 1/3

BLEU SCORE

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

- Inexpensive
- Quick
- Language Independent
- Correlate with human evaluation

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

$$\frac{\text{the number of Ca words(unigrams) which occur in any Ref}}{\text{the total number of words in the Ca}}$$

BLEU SCORE 기본: COUNT

Count Based(WER): Reference(Ref)에 등장한 단어가 Candidate(Ca)에 몇번 등장했는지 count하는 것

Candidate 1: It is a guide to action which ensures that the military always obeys the commands of the party.

Candidate 2: It is to insure the troops forever hearing the activity guidebook that party direct

일치하는 단어 수 / Candidate의 단어 수

17/18

8/14

Reference 1: It is a guide to action that ensures that the military will forever heed Party commands.

Reference 2: It is the guiding principle which guarantees the military forces always being under the command of the Party.

Reference 3: It is the practical guide for the army always to heed the directions of the party.

WER를 번역기 평가에 적용 한계

Candidate: the the the the the the the.

Reference 1: The cat is on the mat.
Reference 2: There is a cat on the mat.

일치하는 단어 수 / Candidate의 단어 수: 7/7=1

문제점: 한번 일치된다고 판단된 단어는 다시 비교되지 말아야 한다.

MODIFIED UNIGRAM PRECISION:

1. Max_ref_count: 유니그램이 하나의 Ref에서 최대 몇 번 등장했는지를 카운트
2. Count_clip 값을 구한다

$$Count_{clip} = min(Count, Max\_Ref\_Count)$$

$$\frac{\sum_{unigram \in Candidate} Count_{clip}(unigram)}{\sum_{unigram \in Candidate} Count(unigram)}$$

Candidate: the the the the the the the.

Reference 1: The cat is on the mat.
Reference 2: There is a cat on the mat.

Max_ref_count = 2
Count = 7

→ 2/7

Count_clip = 2

UNIGRAM 한계: 순서

Candidate1 : It is a guide to action which ensures that the military always obeys the commands of the party.

Candidate2 : It is to insure the troops forever hearing the activity guidebook that party direct.

Candidate3 : the that military a is It guide ensures which to commands the of action obeys always party the.

Candidate 1와 3의 modified unigram precision 동일

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.

MODIFIED N-GRAM PRECISION

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

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

$$p_n = \frac{\sum_{C \in \{Candidates\}} \sum_{n\text{-}gram \in C} Count_{clip}(n\text{-}gram)}{\sum_{C' \in \{Candidates\}} \sum_{n\text{-}gram' \in C'} Count(n\text{-}gram')}.$$

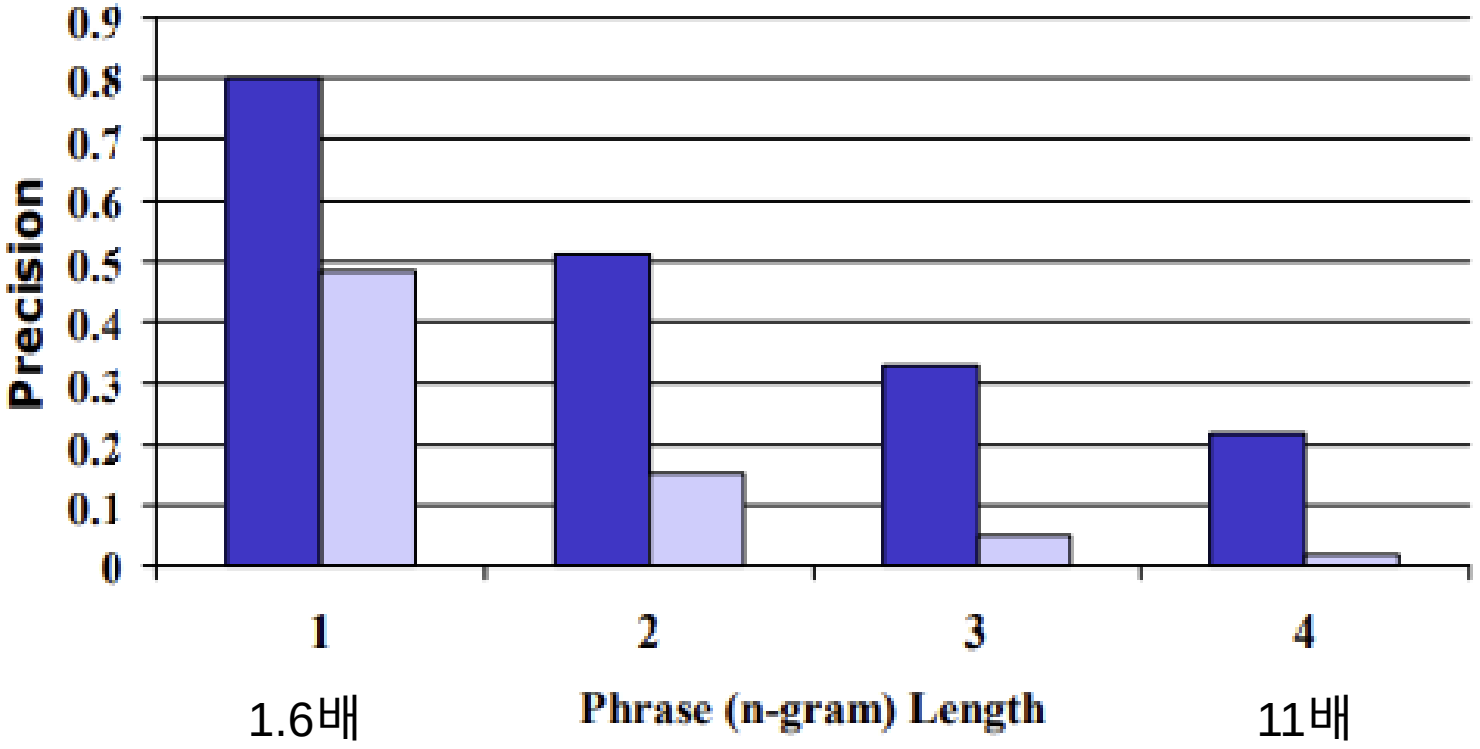
n=1 → adequacy(적절성), 즉 적절한 단어를 선택했는지

n>1 → fluency(유창함), 즉 어순

MODIFIED N-GRAM 적절성 평가

1. 좋은 번역과 나쁜 번역을 구분할 수 있을까: 번역 전문가의 번역 결과와 대충 만든 번역기의 결과 비교

Figure 1: Distinguishing Human from Machine



1.6배

11배

모든 n에서 modified n-gram은 좋은 번역과 나쁜 번역을 구분할 수 있다.

MODIFIED N-GRAM 적절성 평가

번역 성능의 차이가 별로 나지 않는 것도 구분할 수 있을까

성능 차이 안 나는 번역
Group 1

H1: someone lacking native proficiency in both the source (Chinese) and the target language (English).

H2: a native English speaker who speaks Chinese

성능 차이 안 나는 번역
Group 2

S1: machine translations by three commercial systems

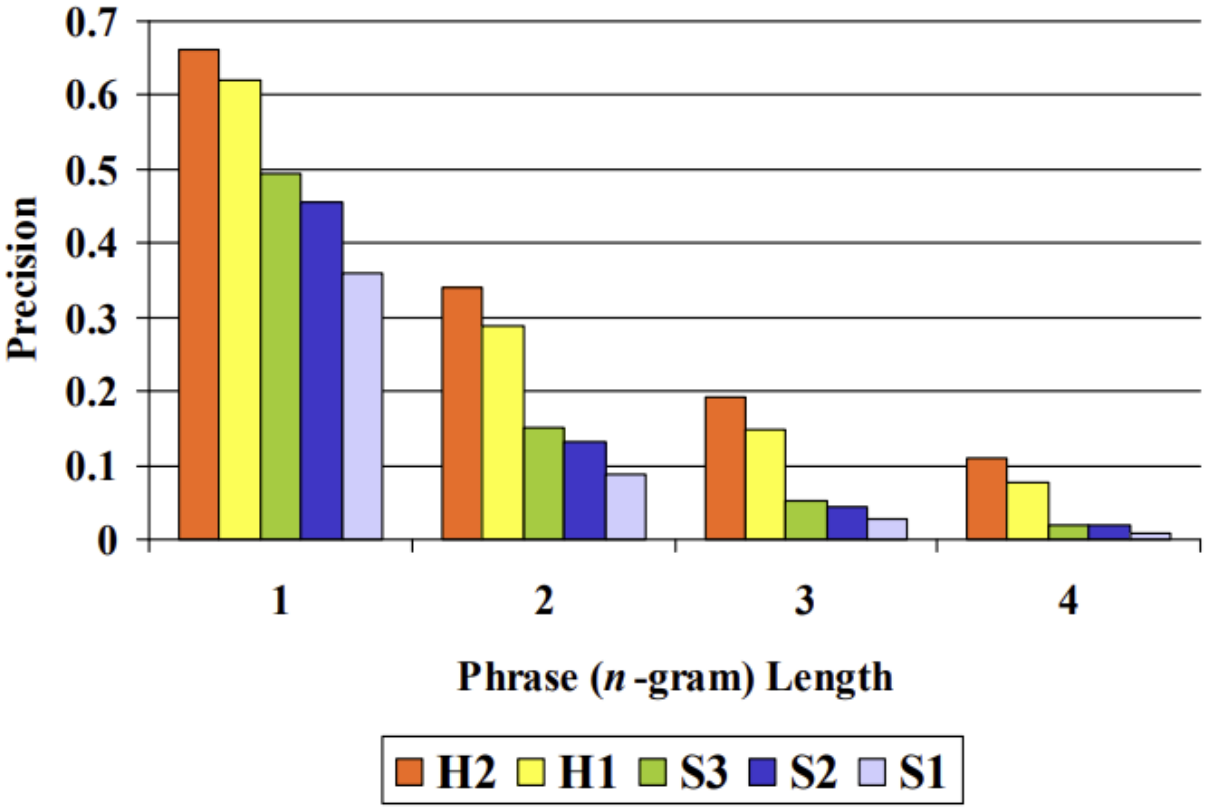
S2: machine translations by three commercial systems

S3: machine translations by three commercial systems

MODIFIED N-GRAM 적절성 평가

번역 성능의 차이가 별로 나지 않는 것도 구분할 수 있을까

Figure 2: Machine and Human Translations



Reference으로 전문적인 번역 결과를 사용

Modified N-gram으로 H1, H2, S1, S2, S3을 전문적인 번역 결과와 비교

1. n이 증가할수록 H2, H1와 S3,S2,S1의 차이 증가

2. 사람이 직접 평가한 변역 결과 Quality 순위와 일치

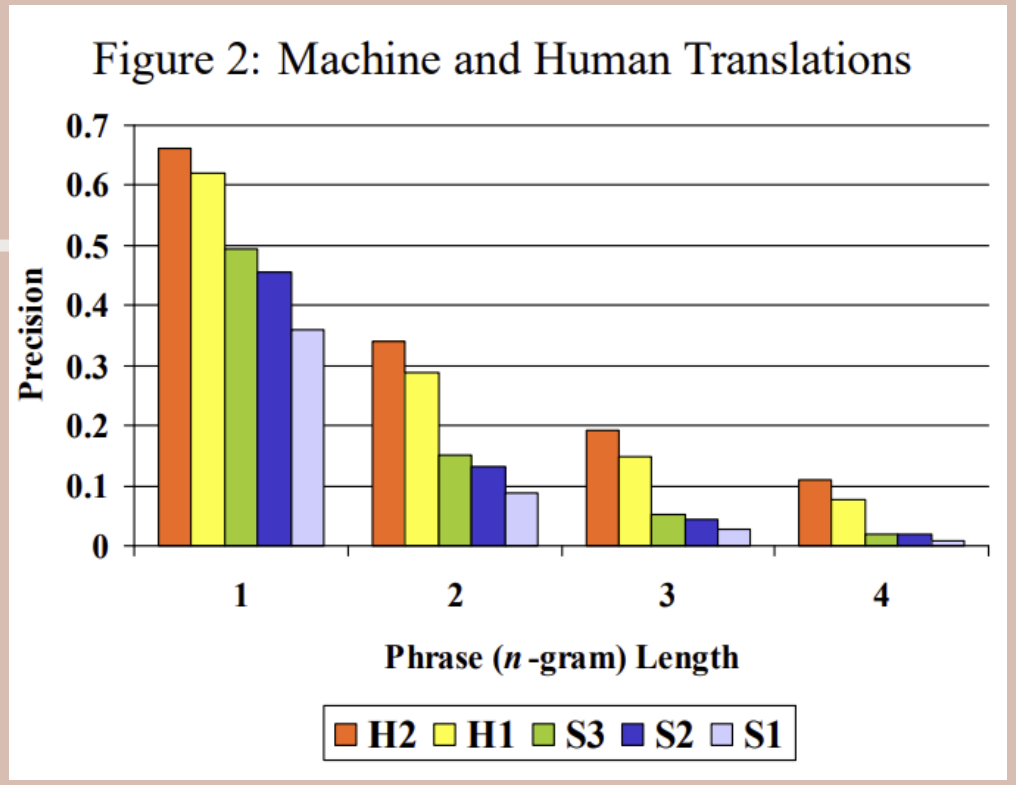
= 번역 Quality 차이 많이 안 나는 것도 올바르게 구분할 수 있다.

Modified N-gram 평가 방법: 적절하다!

BLEU SCORE

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

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



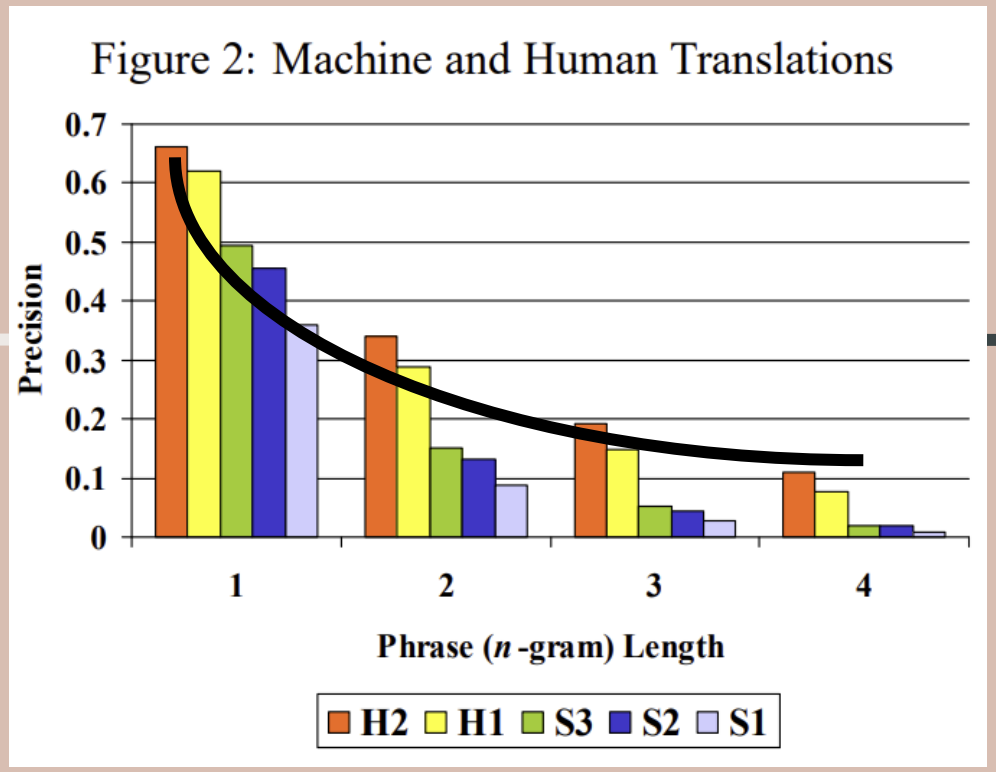
왜 Weighted Average?

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

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

BLEU SCORE

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



n이 증가할수록 Precision 값이 Exponentially하게 감소

→ Precision값이 n에 따라 linear하게 변하도록 log를 사용

BLEU SCORE

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

BP(Brevity Penalty): Modified n-gram 보완하기 위한 상수

MODIFIED N-GRAM의 문제점

길이에 따라 값이 달라진다.

Candidate가 Reference보다 길이가 길 때

Candidate 1: I always invariably perpetually do.
Candidate 2: I always do.
Reference 1: I always do.
Reference 2: I invariably do.
Reference 3: I perpetually do.

$$\frac{\sum_{unigram \in Candidate} Count_{clip}(unigram)}{\sum_{unigram \in Candidate} Count(unigram)}$$

Modified n-gram만으로도 Candidate 1의 Precision 값이 작아진다 (분수의 부모가 커짐)

Candidate가 Reference보다 길이가 짧을 때

Candidate: of the
Reference 1: It is a guide to action that ensures that the military will forever heed Party commands.
Reference 2: It is the guiding principle which guarantees the military forces always being under the command of the Party.
Reference 3: It is the practical guide for the army always to heed the directions of the party.

Modified unigram precison: 2/2
Modified bigram precision: 1/1

→ **문제 발생**

BREVITY PENALTY

좋은 번역은 길이도 비슷해야 한다는 제한을 걸어두는 역할

$$\mathrm{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases}.$$

r: Reference 전체 길이
c: candidate 전체 길이

고려 사항

문장마다 BP를 계산하면 정답 문장, 변역 결과 문장 길이 모두 짧을 때 문제가 발생

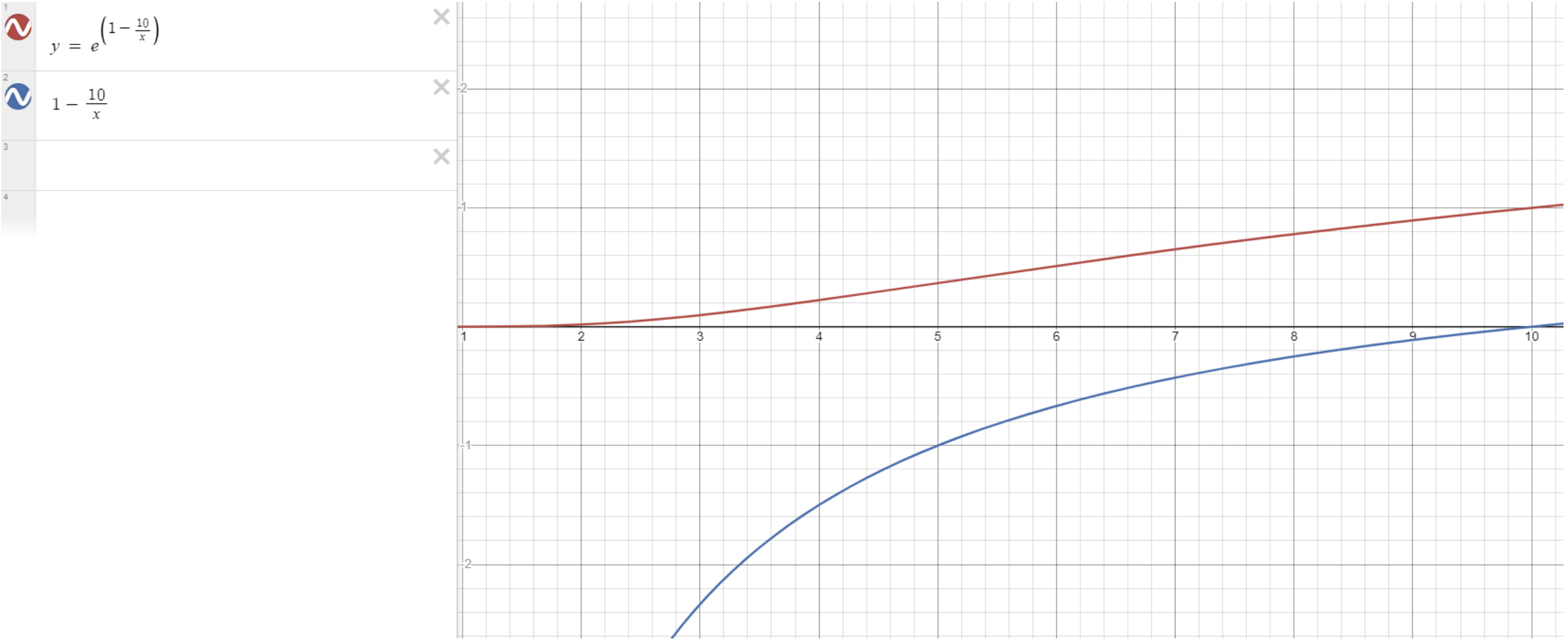
정답 문장	번역 결과	Reference/Candidate
I am tired	I am exhausted	3/2 = 1.5
it's challenging to pinpoint the exact issue	it's hard to pinpoint the exact issue	7/6 = 1.1666

틀린 개수는 같아도 문장의 길이에 따라 R/C 값이 크게 차이 난다.

→ 전체 길이를 비교하여 각 문장들의 길이를 빡빡하게 제한 x

$$\text{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases}.$$

차이가 클수록 더 많은 penalty



BLEU SCORE

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

p_n의 값: 0~1 → log(p_n): -inf ~ 0 → e^x에 대입하면 0~1의 값을 갖게 된다.

BLEU SCORE 적절성 평가

BLEU score 결과: 성능이 좋은 번역기, 나쁜 번역기, 차이가 별로 나지 않는 번역기 모두 구분할 수 있을까

500개의 번역된 문장 전체의 BLEU score

Table 1: BLEU on 500 sentences

S1	S2	S3	H1	H2
0.0527	0.0829	0.0930	0.1934	0.2571

Modified n-gram 결과, 사람이 직접 평가한 결과 순서와 동일

실험 결과 정당성 실험

- BLEU score 차이가 성능이 실제로 차이가 나서 발생한 것일까
- BLEU score의 분산
- 다른 무작위 500 문장 집합을 선택하면 평가 순위 여전히 동일할까

BLEU SCORE 적절성 평가

- BLEU score 차이가 성능이 실제로 차이가 나서 발생한 것일까 우연일까
- BLEU score의 분산
- 다른 무작위 500 문장 집합을 선택하면 평가 순위 여전히 동일할까

1. 500개의 문장을 25문장으로 이루어진 20개의 집합으로 나눈다

2. 집합마다 BLEU 지표를 개별적으로 계산

3. 각 번역 시스템에 대해 20개의 BLEU 지표 샘플을 통해 샘플들의 평균, 분산 및 대응하는 t-통계량 계산

Table 2: Paired t-statistics on 20 blocks

	S1	S2	S3	H1	H2
Mean	0.051	0.081	0.090	0.192	0.256
StdDev	0.017	0.025	0.020	0.030	0.039
t	—	6	3.4	24	11

평균이 전체 BLEU score와 비슷함

분산이 작다 = 적절한 지표

t값이 1.7 이상이면 우연에 의해 차이가 나는 것이 아니다.

실험 결과 1 HUMAN EVALUATION

실험 이유: Human Evaluation 결과와 BLEU score 결과가 비슷할수록 BLEU score가 Human evaluation을 대체할 가능성이 높다

번역기: 중국어 → 영어

Human Group 1: Monolingual = 10 native speakers of English.
→ only on the translations' readability and fluency

Human Group 2: Bilingual = 10 native speakers of Chinese who had lived in the United States for the past several years.
→ 전체적으로 번역이 잘 되었는지

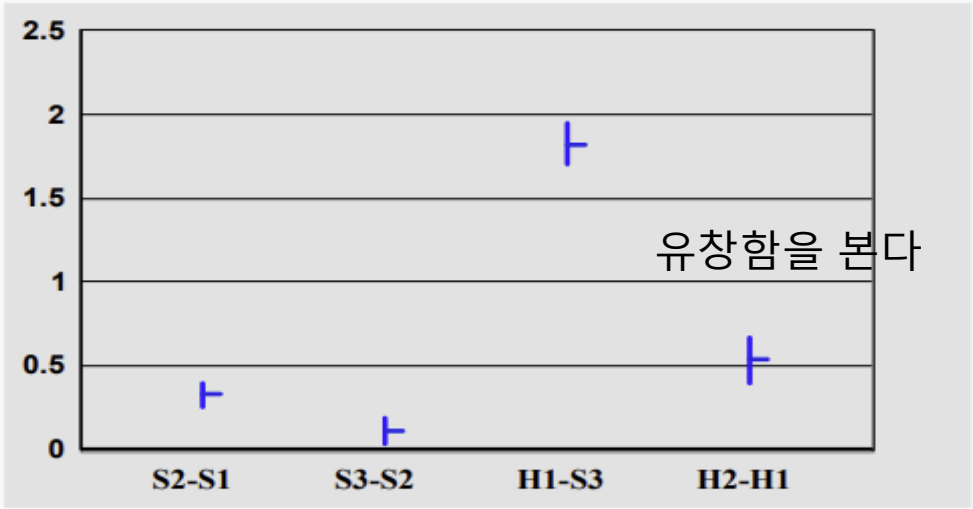
각 그룹마다 번역 결과를 1 (bad) ~ 5 (good)으로 평가하는 것

실험 결과 1 HUMAN EVALUATION

각 번역 시스템 rating을 평균 내어 성능 순서로 시스템마다 차이를 나타내는 것

Monolingual Group

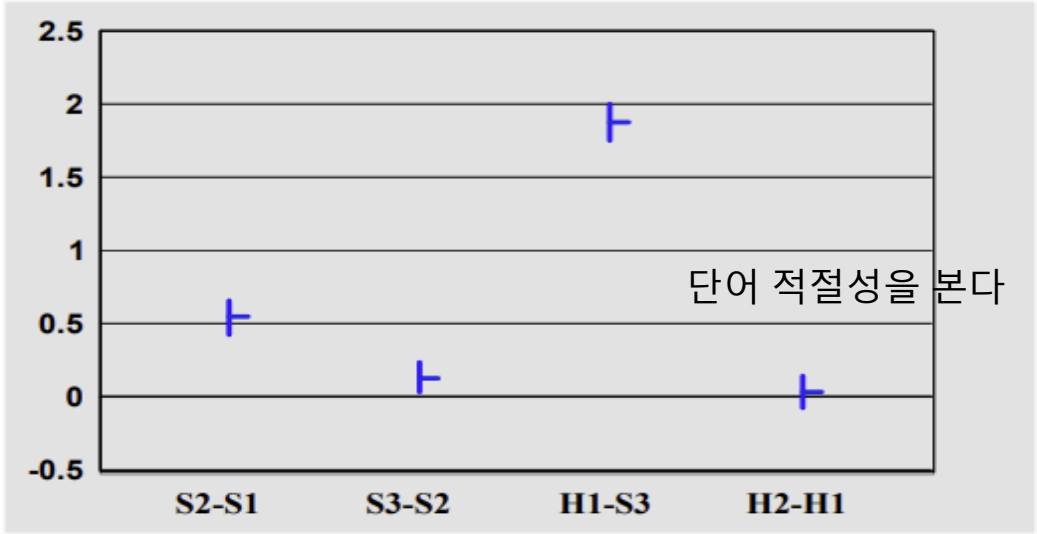
Figure 3: Monolingual Judgments - pairwise differential comparison



+95%	0.400	0.194	1.945	0.670
-95%	0.252	0.034	1.705	0.400
monolingual	0.326	0.114	1.825	0.535

Bilingual Group

Figure 4: Bilingual Judgments - pairwise differential comparison



+95%	0.667	0.238	2.007	0.145
-95%	0.435	0.042	1.759	-0.069
bilingual	0.551	0.140	1.883	0.038

성능 차이 안 나는 번역
Group 1

H1: someone lacking native proficiency in both the source (Chinese) and the target language (English).

H2: a native English speaker who speaks Chinese

성능 차이 안 나는 번역
Group 2

S1: machine translations by three commercial systems

S2: machine translations by three commercial systems

S3: machine translations by three commercial systems

실험 결과 2: BLEU VS THE HUMAN EVALUATION

Figure 5: BLEU predicts Monolingual Judgments

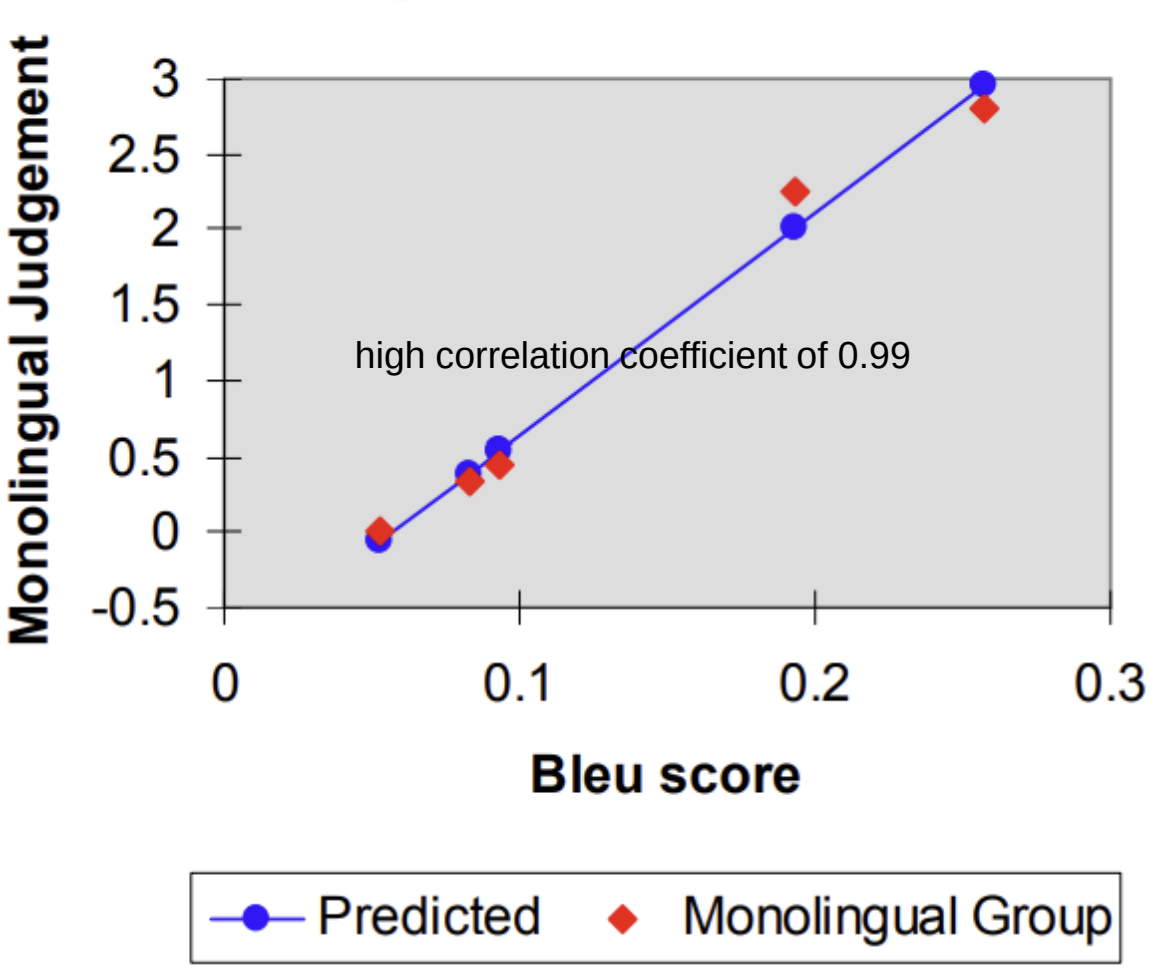
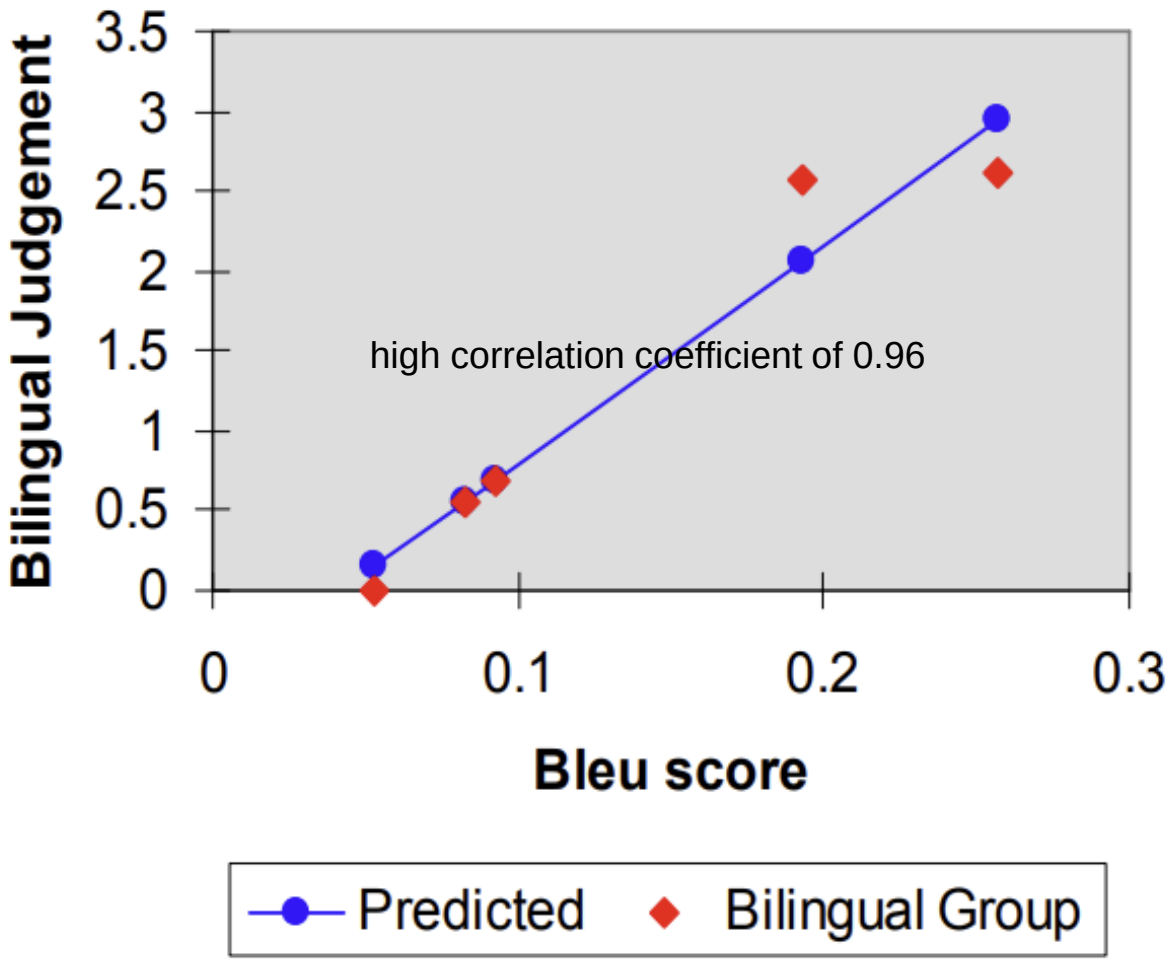


Figure 6: BLEU predicts Bilingual Judgments



실험 결과 2: BLEU VS THE HUMAN EVALUATION

Figure 7: BLEU vs Bilingual and Monolingual Judgments