Towards an Automatic Turing Test: Learning to Evaluate Dialogue Responses

2017200966
한양대학교 인공지능연구실
조수필

Index

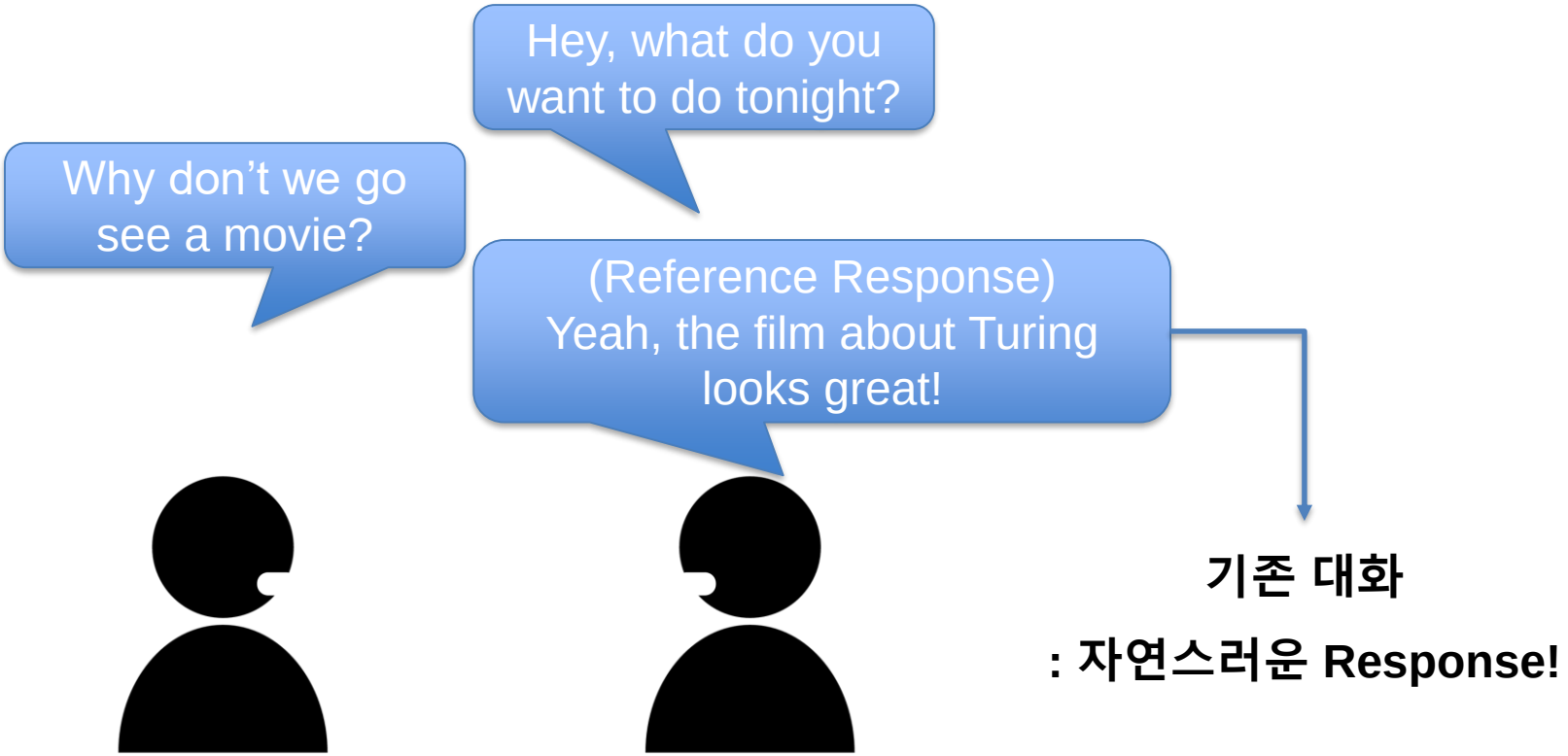
목차

Chapter.1

Intro

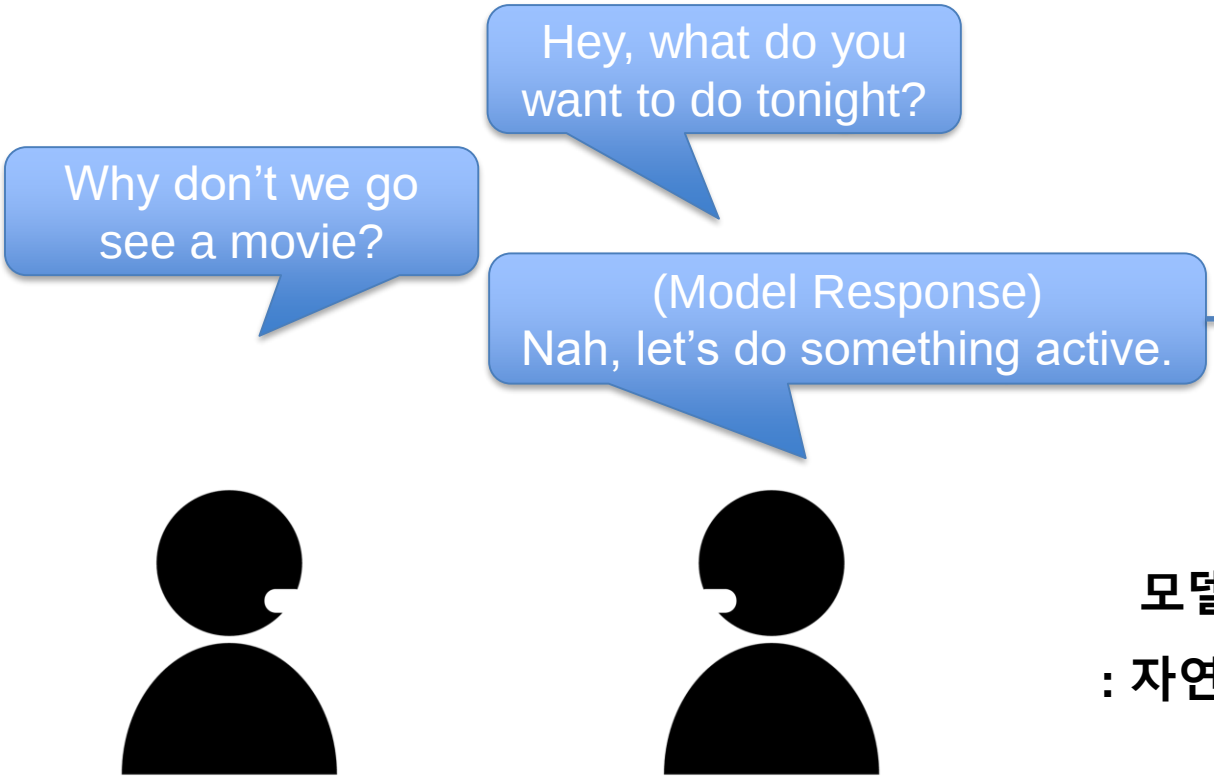
Intro

- **자연스러운 대화 생성 판단...?**



Intro

- **자연스러운 대화 생성 판단...?**



모델이 생성한 대화

: 자연스럽다. 하지만...

Intro

- **“기존 자연어 평가 metric” 으로는 대화 생성을 평가하기 어려움.**

- **Ex) 가장 대표적인 자연어 생성 평가 metric = BLEU**

$$\text{BLEU} = \beta \prod_{i=1}^{k} p_n^{w_n}$$

$$p_n = \text{\# matched n-grams / \# n-grams in candidate translation}$$

$$w_n = 1/2^n \qquad \beta = e^{\min(0,1-\frac{\text{len}_{\text{ref}}}{\text{len}_{\text{MT}}})}$$

Ref : 최근에 하트 시그널 보는 사람 진짜 많더라

Model : 맞아 요즘 하트 시그널 진짜 재밌어

Intro

- **"기존 자연어 평가 metric" 으로는 대화 생성을 평가하기 어려움.**

- **Ex) 가장 대표적인 자연어 생성 평가 metric = BLEU**

$$\text{BLEU} = \beta \prod_{i=1}^{k} p_n^{w_n}$$

Model response의 길이가 Reference response의 길이보다 짧으면 BLEU 감소

잘못된 예시 :

Ref : 맞아 요즘 하트 시그널 진짜 재밌어

Model : 하트

β 없으면 BLEU = 1 , β 있으면 BLEU = 0.007

Intro

Context of Conversation
Speaker A: Hey, what do you want to do tonight?
Speaker B: Why don't we go see a movie?
Model Response
Nah, let's do something active.
Reference Response
Yeah, the film about Turing looks great!

BLEU = 0

기존의 metric은 의미적으로 괜찮은 model response를 제대로 파악하지 못한다!

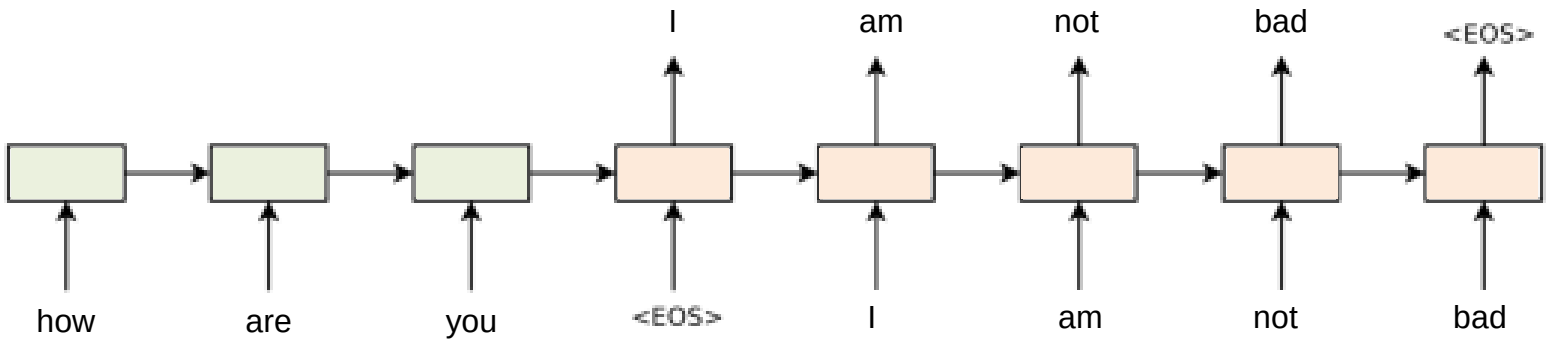
즉, 대화 생성이 잘 되었는가를 판단하려면, 새로운 평가 지표가 필요하다!

Chapter.2

ADEM

대화생성 모델(1/3)

- **RNN Encoder-Decoder**



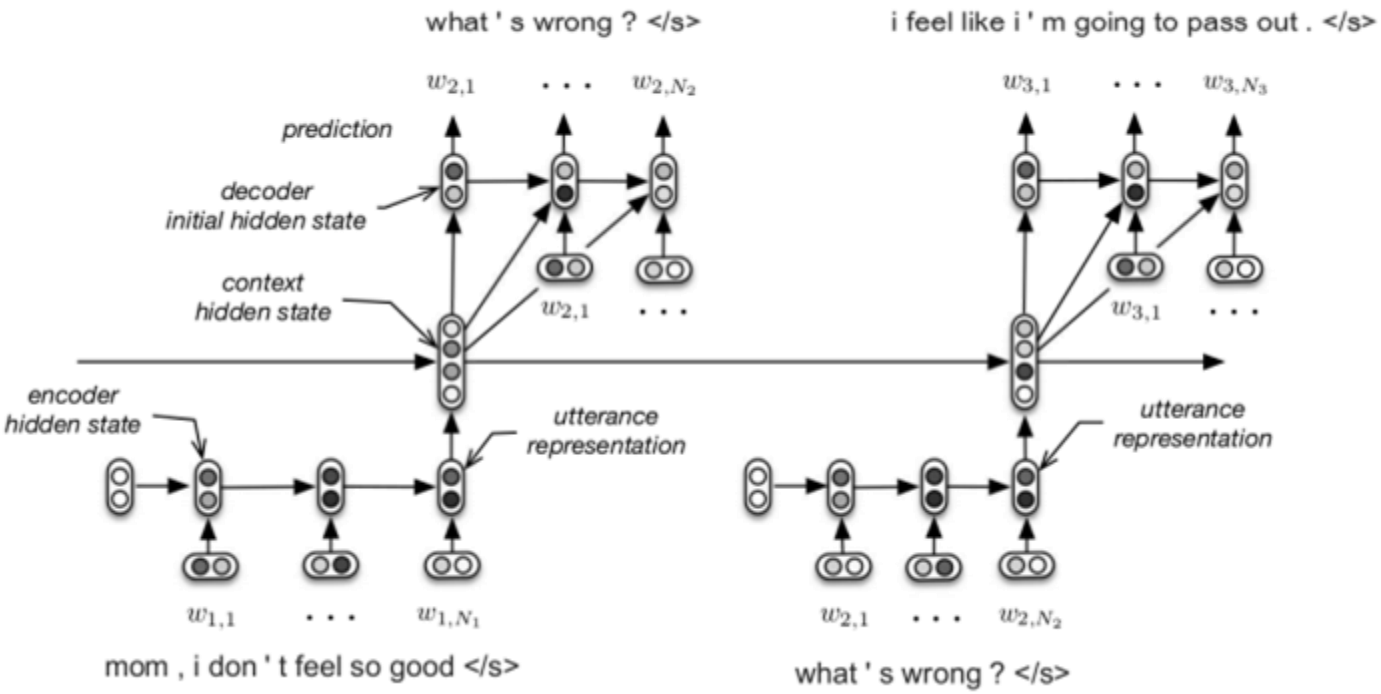
- 가장 기본적인 딥러닝 대화 생성 모델.
- 단점 : 대화의 context를 유지하지 못함.

  Ex1) 날씨가 좋네요! -> 넌 요즘 어때? -> **최고에요!** **(좋은 대화!)**

  Ex2) 직장에서 짤렸다고요? -> 넌 요즘 어때? -> **최고에요!** **(이상한 대화…)**

대화생성 모델(2/3)

- **Hierarchical Recurrent Encoder-Decoder (HRED)**



- 차이점 : 대화의 context를 저장하는 **context hidden state**가 존재한다.

출처 : Building end-to-end dialogue systems using generative hierarchical neural network models

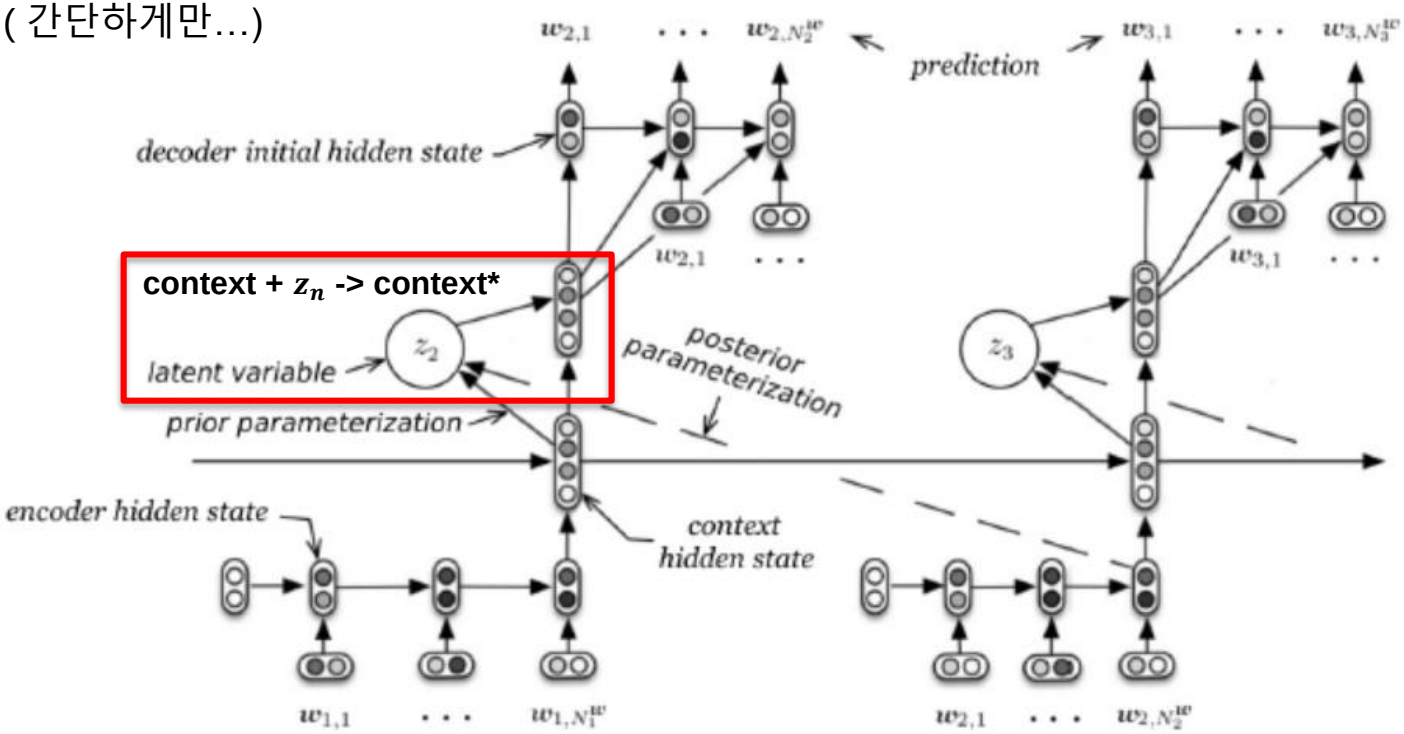
$$P_\theta(\mathbf{z}_n \mid \mathbf{w}_1, \ldots, \mathbf{w}_{n-1}) = \mathcal{N}(\boldsymbol{\mu}_{\text{prior}}(\mathbf{w}_1, \ldots, \mathbf{w}_{n-1}), \Sigma_{\text{prior}}(\mathbf{w}_1, \ldots, \mathbf{w}_{n-1})), \quad (2)$$

$$P_\theta(\mathbf{w}_n \mid \mathbf{z}_n, \mathbf{w}_1, \ldots, \mathbf{w}_{n-1}) = \prod_{m=1}^{M_n} P_\theta(w_{n,m} \mid \mathbf{z}_n, \mathbf{w}_1, \ldots, \mathbf{w}_{n-1}, w_{n,1}, \ldots, w_{n,m-1}), \quad (3)$$

대화생성 모델(3/3)

- Latent Variable Hierarchical Recurrent Encoder-Decoder (VHRED)

(간단하게만…)



한 줄 요약 : 발화문 생성 시, 기존 context에 정규 분포를 통한 약간의 랜덤성을 가한다!

ADEM

- **Automatic Dialogue Evaluation Model (ADEM)**

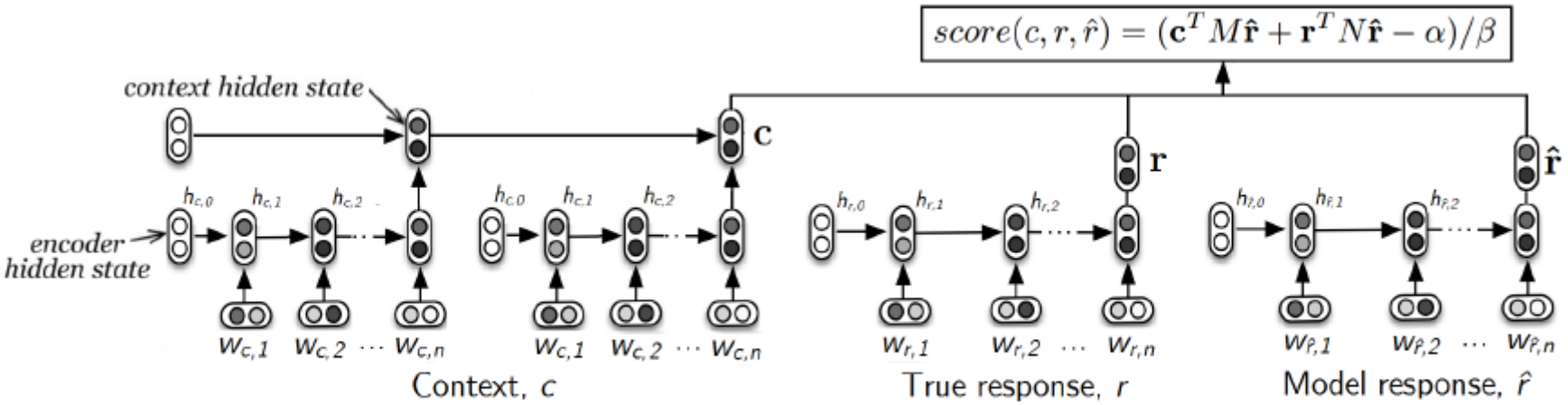


Figure 2: The ADEM model, which uses a hierarchical encoder to produce the context embedding c.

- 대화 문맥 c , 실제 사람의 응답 r , 대화 생성 모델의 응답 $\hat{r}$ 을 입력.
- (c 와 $\hat{r}$) , (r 과 $\hat{r}$) 의 관계가 유사할수록 높은 score를 얻는 것이 목표!

ADEM

$$score(c, r, \hat{r}) = (\mathbf{c}^T M \hat{\mathbf{r}} + \mathbf{r}^T N \hat{\mathbf{r}} - \alpha)/\beta$$

- M = context matrix c 을 model response $\hat{r}$ 의 공간으로 매핑.

  Score는 현재 대화 문맥(c) 에서 매핑된 벡터(c^TM) 와

  모델 응답 $\hat{r}$ 이 유사할수록 높은 점수를 얻는다.

- N = reference response r 을 model response $\hat{r}$ 의 공간으로 매핑.
- α, β = 전체 score 가 [1:5] 에 분포되게끔 하는 상수.

$$\mathcal{L} = \sum_{i=1:K} [score(c_i, r_i, \hat{r}_i) - human_i]^2 + \gamma||\theta||_2$$

- human : model response $\hat{r}$ 에 대해 실제 사람이 평가한 점수.
- $\theta = \{ M, N \}$ 을 이용한 L2 Regularization 진행. (Overfitting 방지)

Chapter.3

ADEM 결과

실험 방법

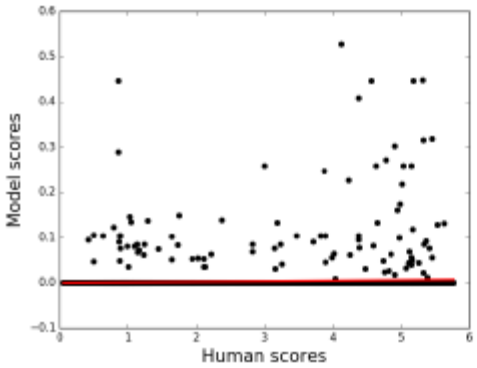
- 트위터 대화 코퍼스와 아래 모델을 통해, **다음 발화문 생성** (c , r , $\hat{r}$ 획득)
  - TF-IDF
  - Dual Encoder
  - HRED
  - Human (예상되는 다음 발화문을 사람이 수기로 작성)

- 생성된 발화문에 대하여, **사람이 수기로 점수를 평가** (Human 획득)

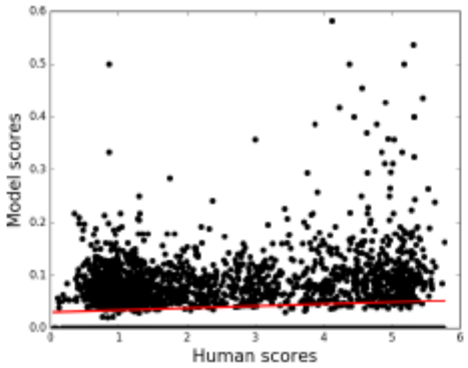
- 다양한 대화 생성 평가 metric(BLEU, ROUGE 등..) 과
  ADEM 의 **평가 결과를 Human evaluation 값과 비교.**

실험 결과(1/5)

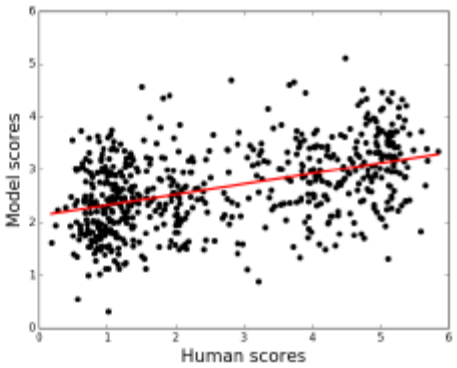
- 기존 대화 생성 평가 metric 2가지와, ADEM 의 대화 평가 점수 분포 비교.



(a) BLEU-2



(b) ROUGE



(c) ADEM

ADEM의 결과 그래프에서, Human & Model 의 선형 관계가 가장 잘 나타난다.

실험 결과(2/5)

- Metric 별 평가 점수와 사람의 실제 평가 점수 간 유사도 분석

	Full dataset		Test set	
Metric	Spearman	Pearson	Spearman	Pearson
BLEU-2	0.039 (0.013)	0.081 (<0.001)	0.051 (0.254)	0.120 (<0.001)
BLEU-4	0.051 (0.001)	0.025 (0.113)	0.063 (0.156)	0.073 (0.103)
ROUGE	0.062 (<0.001)	0.114 (<0.001)	0.096 (0.031)	0.147 (<0.001)
METEOR	0.021 (0.189)	0.022 (0.165)	0.013 (0.745)	0.021 (0.601)
T2V	0.140 (<0.001)	0.141 (<0.001)	0.140 (<0.001)	0.141 (<0.001)
VHRED	-0.035 (0.062)	-0.030 (0.106)	-0.091 (0.023)	-0.010 (0.805)
	Validation set		Test set	
C-ADEM	0.338 (<0.001)	0.355 (<0.001)	0.366 (<0.001)	0.363 (<0.001)
R-ADEM	0.404 (<0.001)	0.404 (<0.001)	0.352 (<0.001)	0.360 (<0.001)
ADEM (T2V)	0.252 (<0.001)	0.265 (<0.001)	0.280 (<0.001)	0.287 (<0.001)
ADEM	**0.410** (<0.001)	**0.418** (<0.001)	**0.428** (<0.001)	**0.436** (<0.001)

Table 2: Correlation between metrics and human judgements, with p-values shown in brackets.

ADEM은 기존의 대화 평가 metric 보다 더 높은 correlation을 가진다(with 사람의 평가 점수).

실험 결과(3/5)

- Training set 에 없던 모델의 결과를 ADEM으로 평가한다면…?

	Test on full dataset		Test on removed model responses	
Data Removed	Spearman	Pearson	Spearman	Pearson
TF-IDF	0.406 (<0.001)	0.409 (<0.001)	0.186 (0.021)	0.196 (0.015)
Dual Encoder	0.364 (<0.001)	0.373 (<0.001)	0.026 (0.749)	0.027 (0.736)
HRED	0.393 (<0.001)	0.396 (<0.001)	0.151 (0.060)	0.148 (<0.070)
Human	0.292 (<0.001)	0.298 (<0.001)	0.216 (<0.010)	0.148 (<0.070)
Average	**0.364**	**0.369**	**0.145**	**0.130**
25% at random	0.378 (<0.001)	0.384 (<0.001)	—	—

좌 : 해당 모델만 빼고, 다른 3가지 방식으로 학습.

이후, 3가지 방식의 test set 으로 ADEM 과 실제 점수의 correlation 비교.

우 : 해당 모델만 빼고, 다른 3가지 방식으로 학습.

이후, 해당 모델의 test set 으로 ADEM 과 실제 점수의 correlation 비교.

Training set과 Test set을 생성한 대화 생성 모델이 각각 달라도, ADEM의 성능은 좋다.

실험 결과(4/5)

- ADEM Score 평가 예시

Context	Reference response	Model responses	Human score	ADEM score
photo to see my television debut go to - some. some on <url> - hehe <url> → it really was you? i thought ppl were recognizing someone who looked like you! were the oysters worth the wait?	yeah it was me . haha i 'd kinda forgotten about it it was filmed a while ago	1) i'm not sure. i just don't know what to do with it.	3	3.413
		2) you heard the horsepower productions remix of lee scratch perry's 'exercising' off his 'mighty upsetter' album?	1	1.644
		3) you wont chug a fuzzy peach navel	1	1.944
		4) they were!	5	3.085
just beat call of duty!! → want a cookie? → yes!! → come get it	im in kenmore at the moment	1) i'm gonna get a new phone some moro	1	2.947
		2) no way man.	5	3.405
		3) wow i just got a free pizza coupon! get yours before theres no more! <url>	1	2.877
		4) i'm going to go to the mall.	1	2.851
am i out of twitter jail yet? testing → yeah. i posted bail → thanks. i am a right chatter tweetbox on sundays. same happened last sunday lol	any news on meeting our user ? i go to the us on friday and i don 't want to miss anything arranged	1) i'm not sure if i'm going to be able to get it.	3	2.651
		2) good to see another mac user in the leadership ranks	4	2.775
		3) awww poor baby hope u get to feeling better soon. maybe some many work days at piedmont	2	2.173
		4) did you tweet too much?	5	3.185

Table 5: Examples of scores given by the ADEM model.

결론 : [1:5] 가 아닌, [2:4] 근처에 모이기는 하지만, 꽤 좋다!!!

실험 결과(5/5)

- ADEM Score 평가 예시(2)

Context	Reference response	Model response	Human score	BLEU-2 score	ROUGE score	ADEM score
i'd recommend <url> - or build buy an htpc and put <url> on it. → you're the some nd person this week that's recommended roku to me.	an htpc with xmbc is what i run . but i 've decked out my setup . i 've got <number> tb of data on my home server	because it's brilliant	5	1.0	1.0	4.726
imma be an auntie this weekend. i guess i have to go albany. herewego → u supposed to been here → i come off nd on. → never tell me smh	lol you sometiming	haha, anyway, how're you?	5	1.0	1.0	4.201
my son thinks she is plain. and the girl that plays her sister. seekhelp4him? → send him this. he'll thank you. <url>	you are too kind for words .	i will do	5	1.0	1.0	5.0

결론 : 좋다222

코드 실행 결과

```
contexts = ['</s> <first_speaker> hello . how are you today ? </s>',
            '</s> <first_speaker> i love starbucks coffee </s>']
true = ['</s> <second_speaker> i am fine . thanks </s>',
        '</s> <second_speaker> i like their latte </s>']
model = ['</s> <second_speaker> fantastic ! how are you ? </s>',
         '</s> <second_speaker> me too ! better than timmies </s>']
```

ADEM Score = [3.80478493 2.67920031]

1번째 대화　　2번째 대화

Thank you!