Are EEG-to-Text Models Working?

상태	완료
Journal	IJCAI
Summary	기존 EEG-to-Text 모델의 teacher-forcing 문제를 언급하며 새로운 평가 방법론 제안
Limitations	EEG-to-Text 모델은 아직 연구가 더 필요하고 노이즈 기반 베이스라인 반드시 포함하도록 함
Link	https://arxiv.org/abs/2405.06459
category	EEG Text Generation
Year	2024.05

https://github.com/NeuSpeech/EEG-To-Text

Introduction

EEG-to-Text 모델이 실제로 EEG 신호를 이해하고 텍스트로 변환하는 능력이 있는지를 비판적으로 분석함. 기존 연구들이 **teacher-forcing** 평가 기법을 사용해 성능이 과장되었다는 문제를 지적하고, **노이즈 입력과의 비교 실험**이 없었다는 점도 비판함.

▼ **teacher-forcing** ?

구분	입력 예시	다음 단어 예측 시 입력
✅ Teacher-forcing	`<sos>` → "He" → "was" → "elected"	항상 정답 토큰을 다음 단계 입력으로 사용
❌ Without Teacher-forcing (실제 생성 상황)	`<sos>` → "He" → "wes" → "e's"	이전에 모델이 **직접 생성한 단어**를 다음 입력으로 사용

이 연구의 주요 기여:

- EEG 데이터를 "정말로" 학습하는 모델인지 확인할 수 있는 **평가 방법론** 제안
- 노이즈 입력에서도 비슷한 성능이 나오는 경우, 모델이 단순히 훈련 데이터를 외운 것일 수 있음을 지적
- 더 엄격하고 투명한 평가 기준의 필요성을 강조

Materials and Methods

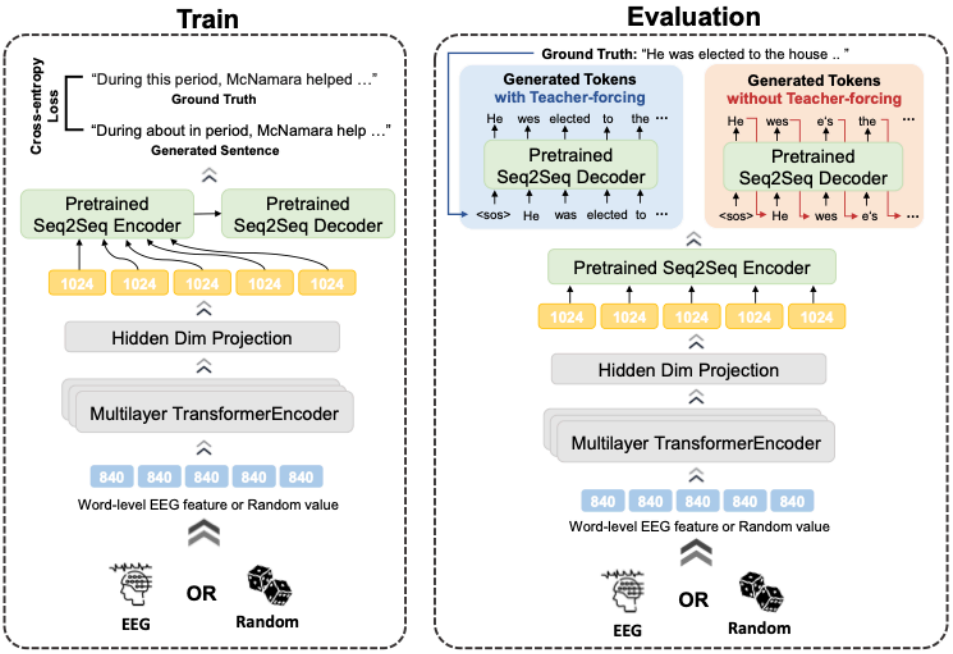

Figure 1. Schematic illustration of the pipeline for a comprehensive assessment of EEG-to-Text models across four distinct training and evaluation setups [1]. These setups explore various combinations of training with either EEG data or random noise as input, followed by evaluation on the same type of data. This approach reveals how models perform under different input conditions. Each setup is further divided to show the influence of teacher-forcing on text generation.

Task Definition

- 연구의 핵심 작업은 **EEG-to-Text 디코딩**
- 즉, 사람이 글을 읽을 때 발생하는 **단어 수준의 EEG 신호(E)**를 입력으로 사용해, 그 사람이 읽은 **문장(S)**을 예측하는 seq2seq 문제로 정의할 수 있음

ZuCo Datasets

ZuCo 1.0과 **ZuCo 2.0**이라는 공개 데이터셋 사용함. 두 데이터셋 모두 자연스럽게 글을 읽는 동안 측정된 EEG와 eye-tracking 데이터를 포함함.

- 읽기 자료: 영화 리뷰, 위키백과 문서
- EEG 신호는 **8개의 주파수 밴드**(알파-1, 2, 베타-1, 2, 감마-1, 2, 세타-1, 2)에 대해 힐버트 변환을 적용하여 840개의 feature로 구성
  - 힐버트 변환 : 시간-주파수 특성을 추출할 수 있음
- 데이터는 훈련/검증/테스트 세트로 분할됨 (문장이 겹치지 않게 처리)

Table 1. Number of training, development, and testing samples for each task within the ZuCo datasets. SR: Normal Reading (movie reviews), NR: Normal Reading (Wikipedia), TSR: Task Specific Reading (Wikipedia).

Reading task	#Training samples	#Development samples	#Testing samples	#Total samples
SR v1.0	3609	467	456	4533
NR v1.0	2645	343	350	3343
NR v2.0	3697	458	392	4547
TSR v1.0	4456	522	601	5579

EEG-to-Text Decoding Evaluation

사용된 모델은 기존 연구(Wang & Ji, 2022 based BART) 기반으로, 다음과 같은 구성:

- **EEG 인코더**: multi-layer Transformer로 EEG 데이터 임베딩
- **디코더**: 사전학습된 BART, PEGASUS, T5 모델 사용 → 텍스트 생성

💡 연구의 핵심은 4가지 학습/평가 조합을 통해 **모델이 정말 EEG를 학습하는지**를 검증하는 것:

훈련 데이터	평가 데이터	의미
EEG	EEG	모델이 EEG로부터 실제 텍스트를 예측하는지 평가
Random	Random	노이즈만으로 훈련/평가했을 때의 베이스라인
EEG	Random	학습은 EEG로, 테스트는 노이즈로 → 일반화 능력 평가
Random	EEG	노이즈로 훈련했지만 EEG로 평가 → 진짜 EEG 이해 여부 확인

Results and Discussion

- ZuCo v1.0에 대한 평가
  - EEG는 word-level임
  - random은 평균분포에서 생성한 랜덤값임
  - T5 모델의 경우
    - **Without TF**: BLEU-1 = 16.64 (EEG), 15.54 (Random)
    - **With TF**: BLEU-1 = 43.50 (EEG), 43.63 (Random)
  - 이는,
    - teacher-forcing을 쓰면 **입력과 상관없이 정답에 가까운 문장을 쉽게 생성**
    - 즉, 평가가 **실제 디코딩 능력과 무관한 착시 효과**를 줌

Table 2. EEG-to-Text model evaluation on the ZuCo datasets, incorporating reading tasks from **SR v1.0, NR v1.0, and TSR v1.0**. "*w/tf*" denotes results obtained using teacher-forcing during evaluation as utilized in the original study [1]. In the training and evaluation phases, "EEG" denotes the use of word-level EEG features, while "Random" refers to the employment of random numbers generated from a normal distribution.

			BLEU-N (%)				ROUGE-1 (%)			WER (%)
Pretrained model	**Training**	**Evaluation**	N=1	N=2	N=3	N=4	P	R	F	
BART	EEG	EEG	13.69	2.97	0.82	0.32	11.98	13.43	11.87	108.43
	EEG	Random	13.87	3.09	0.77	0.25	12.23	13.60	12.14	108.31
	Random	EEG	14.05	3.12	1.00	0.41	11.46	12.37	11.14	110.96
	Random	Random	14.22	3.06	0.93	0.39	11.62	12.29	11.19	110.98
BART *w/tf* [1]	EEG	EEG	39.31	22.09	12.49	7.27	26.41	31.40	28.58	78.08
	EEG	Random	39.34	22.13	12.52	7.29	26.44	31.43	28.61	78.07
	Random	EEG	39.67	22.15	12.49	7.12	26.29	31.00	28.34	78.09
	Random	Random	39.69	22.17	12.50	7.12	26.32	31.03	28.37	78.09
Pegasus	EEG	EEG	8.47	2.48	0.81	0.25	0.00	0.00	0.00	99.69
	EEG	Random	8.58	2.48	0.78	0.00	0.00	0.00	0.00	99.89
	Random	EEG	9.12	2.70	0.91	0.23	0.00	0.00	0.00	98.73
	Random	Random	9.06	2.60	0.84	0.00	0.00	0.00	0.00	99.24
Pegasus *w/tf*	EEG	EEG	38.18	21.04	11.50	6.09	26.72	30.51	28.38	78.57
	EEG	Random	38.30	21.09	11.57	6.12	26.84	30.65	28.51	78.56
	Random	EEG	39.10	21.74	11.97	6.17	27.43	31.26	29.11	78.09
	Random	Random	39.17	21.70	11.96	6.18	27.41	31.34	29.14	78.10
T5	EEG	EEG	16.64	5.80	1.96	0.81	12.28	12.88	11.85	111.13
	EEG	Random	15.42	4.78	1.57	0.65	10.57	11.45	10.35	112.00
	Random	EEG	15.95	5.71	2.01	0.91	11.90	12.61	11.47	111.37
	Random	Random	15.54	5.22	1.70	0.67	11.48	12.23	11.10	111.74
T5 *w/tf*	EEG	EEG	43.50	25.50	15.18	8.69	22.92	28.23	25.11	81.39
	EEG	Random	43.53	25.56	15.15	8.68	22.76	27.82	24.87	81.43
	Random	EEG	43.47	25.34	15.03	8.67	23.02	27.96	25.06	81.64
	Random	Random	43.63	25.57	15.23	8.78	23.36	28.40	25.45	81.46

- ZuCo v1.0 및 ZuCo v2.0 합친 평가

Table 3. EEG-to-Text model evaluation on the ZuCo datasets, incorporating reading tasks from **SR v1.0, NR v1.0, and NR v2.0**. "*w/tf*" denotes results obtained using teacher-forcing during evaluation as utilized in the original study [1]. In the training and evaluation phases, "EEG" denotes the use of word-level EEG features, while "Random" refers to the employment of random numbers generated from a normal distribution.

			BLEU-N (%)				ROUGE-1 (%)			WER (%)
Pretrained model	**Training**	**Evaluation**	N=1	N=2	N=3	N=4	P	R	F	
BART	EEG	EEG	11.58	3.40	1.33	0.54	11.33	15.44	12.40	99.68
	EEG	Random	11.26	3.19	1.28	0.57	11.07	15.08	12.06	100.03
	Random	EEG	11.84	3.36	1.41	0.68	11.39	15.17	12.24	100.88
	Random	Random	11.78	3.18	1.23	0.52	11.22	15.05	12.07	101.12
BART *w/tf* [1]	EEG	EEG	41.22	24.18	13.87	7.77	29.52	36.31	32.42	75.33
	EEG	Random	41.22	24.18	13.87	7.77	29.52	36.32	32.42	75.31
	Random	EEG	41.63	24.70	14.44	8.45	29.56	35.83	32.30	74.76
	Random	Random	41.61	24.69	14.40	8.40	29.55	35.77	32.27	74.76
Pegasus	EEG	EEG	10.57	2.82	0.86	0.27	0.00	0.00	0.00	99.21
	EEG	Random	11.14	2.75	0.88	0.30	0.00	0.00	0.00	99.70
	Random	EEG	9.06	2.57	1.00	0.36	0.00	0.00	0.00	98.37
	Random	Random	9.08	2.54	0.92	0.22	0.00	0.00	0.00	98.40
Pegasus *w/tf*	EEG	EEG	41.18	23.42	13.12	7.22	30.32	34.83	32.31	75.30
	EEG	Random	41.06	23.17	12.82	6.86	30.08	34.59	32.07	75.51
	Random	EEG	41.89	23.90	13.27	7.09	30.88	35.52	32.93	74.67
	Random	Random	41.73	23.80	13.29	7.22	30.69	35.31	32.73	74.88
T5	EEG	EEG	16.76	6.15	2.56	1.26	13.69	15.27	13.71	104.50
	EEG	Random	15.70	5.44	2.04	0.93	12.25	13.84	12.35	106.01
	Random	EEG	15.86	5.47	2.17	1.02	12.88	14.75	13.03	104.30
	Random	Random	16.26	5.76	2.23	1.03	13.58	15.44	13.66	104.49
T5 *w/tf*	EEG	EEG	46.03	28.23	17.35	10.55	28.14	33.84	30.53	77.06
	EEG	Random	45.55	27.71	16.85	10.12	27.59	32.99	29.87	77.63
	Random	EEG	45.62	27.78	16.79	10.15	27.26	32.93	29.65	77.44
	Random	Random	45.75	27.96	16.99	10.21	27.39	33.05	29.77	77.20

- 디코딩 샘플
  - 학습 데이터의 문장 구조만을 **기계적으로 재현**하고 있음
  - → **EEG가 영향을 미친 흔적이 없음**

Table 4. Decoding examples of EEG-to-Text models [1]. "EEG" and "Random" represent that the model is trained and tested on word-level EEG features and random numbers, respectively. BART is used as the pretrained seq2seq model for all models. "*w/tf*" denotes results obtained using teacher-forcing during evaluation, as utilized in the original study [1]. **Bold** indicates words exactly matching the ground truth. Underline denotes words consistently generated by models without teacher-forcing.

Ground truth	It's not a particularly **good** film, **but** neither **is it** a monsterous **one**.
EEG *w/tf* [1]	's a a bad **good** movie, **but** it **is it** bad bad. **one**.
Random *w/tf*	's a a bad **good** movie, **but** it **is it** bad bad. **one**.
EEG	He was elected to the United States House of Representatives in 1946.
Random	He was educated at the University of Virginia, where he earned a Bachelor of Arts degree in political science.
Ground truth	Everything its title **implies**, a standard-**issue** crime **drama** spat out from the Tinseltown assembly **line**.
EEG *w/tf* [1]	about predecessor **implies** is and movie,**issue**, **drama**. between of the depthsseltown set **line**.
Random *w/tf*	about predecessor **implies** is and movie,**issue**, **drama**. between of the sameseltown set..
EEG	He was a member of the Democratic National Committee (DNC) from 1952 until his death in 1968.
Random	He was educated at Trinity College, Cambridge and the University of Oxford.
Ground truth	Joseph H. Ball (November 3, 1905 - December 18, **1993) was an American** politician.
EEG *w/tf* [1]	was. Smith (born 23, 18 - April 23, **1993) was an American** actor,
Random *w/tf*	wasux W (born 23, 18 - April 23, 1977) **was an American** actor.
EEG	He was elected to the United States House of Representatives in 1920.
Random	He was educated at the University of Wisconsin-Madison, and received a Bachelor of Arts degree in English literature from Stony Brook University.

- 즉,
  1. **Teacher-forcing의 영향이 매우 큼**: BLEU, ROUGE, WER 등의 지표에서 teacher-forcing을 사용하면 성능이 3배 이상 상승. 하지만 이는 실제 상황을 반영하지 않음
  2. **EEG와 노이즈 입력의 성능이 비슷함**: 학습 및 평가 데이터를 노이즈로 해도 EEG와 거의 유사한 성능이 나와, 모델이 EEG에서 실제 의미를 추출하지 못하고 있을 가능성이 큼
  3. **모델은 주로 훈련 레이블(token label)에 의존**: 입력 신호(EEG든 노이즈든)보다는 출력 문장(label)을 학습해 재생산하는 경향

Conclusion

이 논문은 **EEG-to-Text 모델들이 실제로 뇌파를 이해하지 못할 수도 있음**을 실험적으로 보임.

기존 연구의 평가지표는 **teacher-forcing**을 사용했기 때문에 과장된 측면이 있고, **노이즈 입력과의 비교가 없었다는 점**도 큰 약점이라고 언급함.

💡 앞으로 EEG 기반 언어 생성 모델을 연구할 때는 다음을 권장함:

- **노이즈 기반 베이스라인** 반드시 포함
- **teacher-forcing 없는 평가 방식** 활용
- **더 투명한 평가 및 결과 보고**