BLEURT

Learning Robust Metrics for Text Generation

Hanyang Univ. AI Lab
이은수

Abstract

BLEURT: Learning Robust Metrics for Text Generation

Conference: ACL 2020
Author: Thibault (Google Research) et al.

NLP evaluation metric인 BLEU와, Language Model인 BERT의 합성어

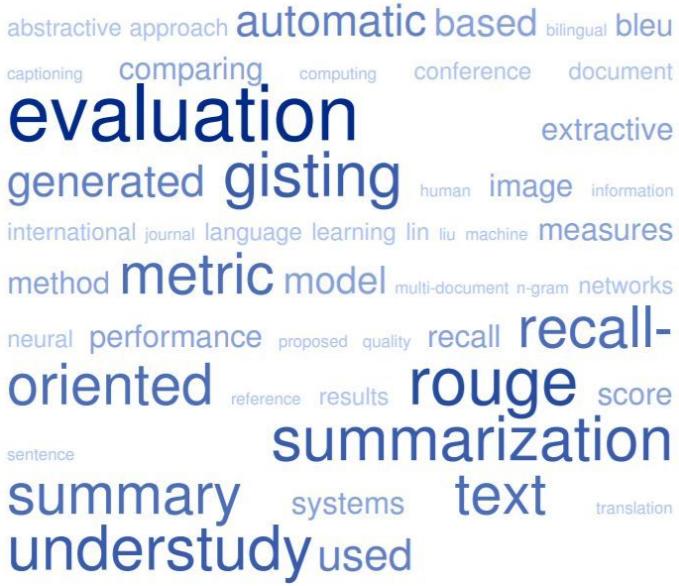
Abstract

Text generation은 최근 몇 년 사이 많은 발전을 이루었지만, BLEU, ROUGE 등 evaluation metrics는 인간의 판단과는 상성이 나쁘고 뒤쳐짐

BLEURT는 BERT를 기반으로 metric을 학습한다

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

1) BLEU



2) ROUGE

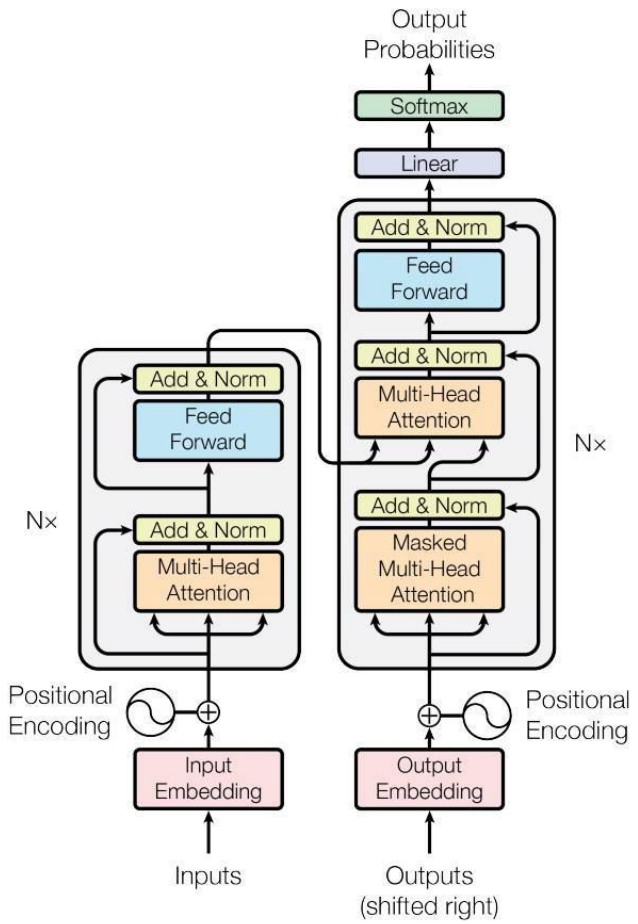
1) "꾸무 메뉴 5주차(1) - Learning Phrase Representation using RNN Encoder-Decoder for Statistical Machine Translation"
2) "ROUGE (Recall-Oriented Understudy for Gisting Evaluation) 2019"

Introduction

지난 몇 년 동안 Natural Language Generation(NLG)은 매우 큰 인코더-디코더 모델에 의해 많은 발전을 이룸

Translation, Summarization, Structured data-to-Text Generation, Dialog, Image Captioning 등을 수행함

그러나 evaluation metrics의 단점으로 진척이 더디다



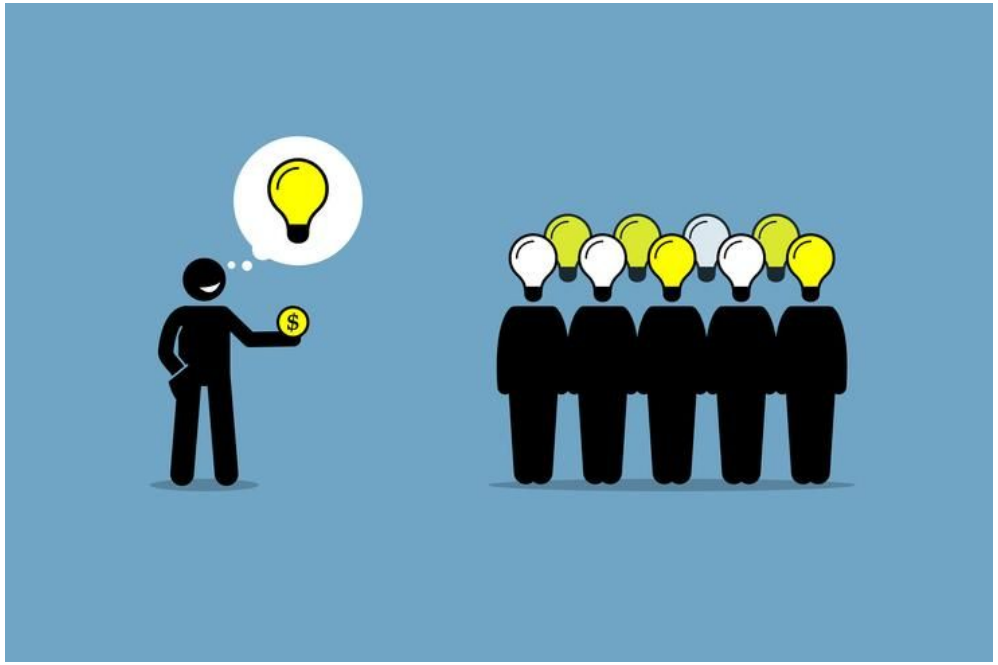
Introduction

Human evaluation은 모델을 가장 잘 평가할 수 있는 지표

그러나 crowdsourcing을 하는 것은 비용이 많이 들고 긴 시간이 걸리므로 비효율적임

일반적으로 품질에 대한 허용 가능한 대용품을 제공하고 계산하는데 매우 저렴한 자동 평가 지표를 사용함

- 예를 들어, 두 가지 인기 지표인 BLEU(Pa-pineni et al., 2002)와 RUZE(Lin, 2004)는 N-gram 중복에 의존함
- 이러한 지표는 어휘 변수에만 민감하기 때문에 주어진 참조의 부분적 또는 구문적 변동을 적절히 보상할 수 없음



3) Crowdsourcing

3) "7 Benefits of Crowdsourcing for Small and Medium Enterprises"

Introduction

Translation metric을 human evalution과 비교하는 벤치마크인 WMT Metrics Shared Task로 metric 테스트 가능

최근 2년 동안 진행된 대회에서는 신경망 기반 RUSE, YiSi, ESIM(Ma et al., 2018, 2019)이 차지함

현재 접근법은 크게 두 가지 범주로 나눌 수 있음

	Method	
	End-to-End	Hybrid
Model	BEER, RUSE, ESIM	YiSi, BERTscore
Benefit	유창성, 충실성, 문법 또는 스타일과 같은 task별 특성을 측정하도록 조정할 수 있음	robust함. 훈련 데이터가 거의 없거나, 없을 때 더 나은 결과를 제공할 수 있으며, 훈련 및 테스트 데이터가 동일하게 분포된다는 가정에 의존하지 않음

Introduction

BLEURT는 모델의 generalize를 위해 수백만 개의 synthetic examples를 통해 새로운 pre-training을 제안함

WMT 2017 벤치마크로 quality drift에 대처하는 능력을 테스트하고, data-to-text 데이터셋인 WebNLG 2017(Gardent et al., 2017)의 세 가지 task를 통해 다른 도메인에 쉽게 적응할 수 있음을 보여줌

훈련 데이터가 부족하거나, 왜곡되거나, 다른 도메인일 때에도 robust를 보장함

Pre-Training on Synthetic Data

BLEURT의 핵심은 rating data를 fine-tuning하기 전에 pre-trained된 BERT를 예열하기 위해 사용하는 pre-training 기법에 있음

다수의 reference-candidate pairs를 생성하고, 이를 어휘 및 의미 수준의 supervision signal로 BERT를 훈련함

불완전한 훈련 데이터를 가지고도 이 단계 이후에는 훨씬 더 잘 일반화됨

일반성을 보장하기 위해 다음 조건을 만족하도록 함

1) Reference 데이터 세트는 크고 다양해야 하며, 따라서 BLEURT가 광범위한 NLG 도메인과 과제에 대처 가능해야 함
2) Sentence pair에는 다양한 어휘, 통사적, 의미적 차이가 포함되어야 함
3) Pre-training 목표로는 BLEURT가 이러한 현상을 식별하는 방법을 배울 수 있도록 해야 함

Generating Sentence Pairs

BLEURT를 다양한 문장 차이에 노출시키는 한 가지 방법은 기존 문장 쌍 데이터 세트를 사용하는 것(Bowman et al., 2015; Williams et al., 2018; Wang et al., 2019)

다만 이러한 데이터 세트는 NLG 시스템이 생성하는 오류와 변경사항(예: 생략, 반복, 불합리한 대체)을 포착하지 못할 수 있으므로, 대신에 임의로 그리고 적은 비용으로 확장할 수 있는 자동적 접근방식을 선택함

- 위키피디아로부터 180만개의 세그먼트 z를 임의로 변화시킴으로써 합성 문장 쌍을 생성

Generating Sentence Pairs

1) **BERT로 mask 채우기**

   a) BERT의 초기 훈련 과제는 토큰화된 문장으로 공백을 메우는 것. 문장의 임의 위치에 mask를 삽입하여 언어 모델로 채우는 방식으로 이러한 기능을 활용함. 따라서 문장의 유창성을 유지하면서 어휘적 변경을 도입함

      i) 문장의 임의 위치에 mask 추가

      ii) mask 토큰의 연속적인 sequence 생성

2) **역번역(Backtranslation)**

   a) 영어에서 다른 언어로의 번역과, 번역 모델을 사용해 다시 영어로 번역하여 문장을 생성함(Bannard and Callison-Burch, 2005; Ganitkevitch et al., 2013; Sennrich et al., 2016). 주된 목적은 의미론을 보존하는 기준 문장의 변형을 만드는 것. 또한 오예측을 현실적인 변경의 원천으로 사용함

3) **단어 삭제(Dropping words)**

   a) 임의로 단어를 삭제하는 것이 우리의 실험에서 유용함

Pre-Training Signals

Task Type	Pre-training Signals	Loss Type
BLEU	τ_{BLEU}	Regression
ROUGE	$\tau_{\text{ROUGE}} = (\tau_{\text{ROUGE-P}}, \tau_{\text{ROUGE-R}}, \tau_{\text{ROUGE-F}})$	Regression
BERTscore	$\tau_{\text{BERTscore}} = (\tau_{\text{BERTscore-P}}, \tau_{\text{BERTscore-R}}, \tau_{\text{BERTscore-F}})$	Regression
Backtrans. likelihood	$\tau_{\text{en-fr},z\|\tilde{z}}, \tau_{\text{en-fr},\tilde{z}\|z}, \tau_{\text{en-de},z\|\tilde{z}}, \tau_{\text{en-de},\tilde{z}\|z}$	Regression
Entailment	$\tau_{\text{entail}} = (\tau_{\text{Entail}}, \tau_{\text{Contradict}}, \tau_{\text{Neutral}})$	Multiclass
Backtrans. flag	$\tau_{\text{backtran\_flag}}$	Multiclass

Table 1: Our pre-training signals.

Backtrans. likelihood: 기존의 번역 모델을 활용하여 의미론적 동등성을 측정

Entailment: reference가 candidate를 entail(수반)하는지 contradict(모순)하는지, neutral(중립)하는지 예측

Backtrans. flag: Backtranslation으로 생성되거나 mask-filling으로 생성되었는지 예측

각 pre-training task를 regression loss 또는 classification loss로 학습하고, task-level loss로 weighted sum함

Experiments

model	cs-en τ / r	de-en τ / r	fi-en τ / r	lv-en τ / r	ru-en τ / r	tr-en τ / r	zh-en τ / r	**avg** τ / r
sentBLEU	29.6 / 43.2	28.9 / 42.2	38.6 / 56.0	23.9 / 38.2	34.3 / 47.7	34.3 / 54.0	37.4 / 51.3	32.4 / 47.5
MoverScore	47.6 / 67.0	51.2 / 70.8	NA	NA	53.4 / 73.8	56.1 / 76.2	53.1 / 74.4	52.3 / 72.4
BERTscore w/ BERT	48.0 / 66.6	50.3 / 70.1	61.4 / 81.4	51.6 / 72.3	53.7 / 73.0	55.6 / 76.0	52.2 / 73.1	53.3 / 73.2
BERTscore w/ roBERTa	54.2 / 72.6	56.9 / 76.0	64.8 / 83.2	56.2 / 75.7	57.2 / 75.2	57.9 / 76.1	58.8 / 78.9	58.0 / 76.8
chrF++	35.0 / 52.3	36.5 / 53.4	47.5 / 67.8	33.3 / 52.0	41.5 / 58.8	43.2 / 61.4	40.5 / 59.3	39.6 / 57.9
BEER	34.0 / 51.1	36.1 / 53.0	48.3 / 68.1	32.8 / 51.5	40.2 / 57.7	42.8 / 60.0	39.5 / 58.2	39.1 / 57.1
BLEURTbase -pre	51.5 / 68.2	52.0 / 70.7	66.6 / 85.1	60.8 / 80.5	57.5 / 77.7	56.9 / 76.0	52.1 / 72.1	56.8 / 75.8
BLEURTbase	55.7 / 73.4	56.3 / 75.7	68.0 / 86.8	**64.7** / 83.3	60.1 / 80.1	62.4 / 81.7	59.5 / 80.5	61.0 / 80.2
BLEURT -pre	56.0 / 74.7	57.1 / 75.7	67.2 / 86.1	62.3 / 81.7	58.4 / 78.3	61.6 / 81.4	55.9 / 76.5	59.8 / 79.2
BLEURT	**59.3 / 77.3**	**59.9 / 79.2**	**69.5 / 87.8**	64.4 / **83.5**	**61.3 / 81.1**	**62.9 / 82.4**	**60.2 / 81.4**	**62.5 / 81.8**

Table 2: Agreement with human ratings on the WMT17 Metrics Shared Task. The metrics are Kendall Tau (τ) and the Pearson correlation (r, the official metric of the shared task), divided by 100.

BLEURT: BERT-Large를 사용함
BLEURTbase: BERT-base를 사용함
***-pre**: pre-training 과정 없이 fine-tuning함
*-pre 모델을 제외한 모델은 pre-trained BERT 위에 sentence pairs로 pre-training을 진행하고, task specific fine-tuning을 진행함

Experiments

model	cs-en τ / DA	de-en τ / DA	et-en τ / DA	fi-en τ / DA	ru-en τ / DA	tr-en τ / DA	zh-en τ / DA	**avg** τ / DA
sentBLEU	20.0 / 22.5	31.6 / 41.5	26.0 / 28.2	17.1 / 15.6	20.5 / 22.4	22.9 / 13.6	21.6 / 17.6	22.8 / 23.2
BERTscore w/ BERT	29.5 / 40.0	39.9 / 53.8	34.7 / 39.0	26.0 / 29.7	27.8 / 34.7	31.7 / 27.5	27.5 / 25.2	31.0 / 35.7
BERTscore w/ roBERTa	31.2 / 41.1	42.2 / 55.5	37.0 / 40.3	27.8 / 30.8	30.2 / 35.4	32.8 / 30.2	29.2 / 26.3	32.9 / 37.1
Meteor++	22.4 / 26.8	34.7 / 45.7	29.7 / 32.9	21.6 / 20.6	22.8 / 25.3	27.3 / 20.4	23.6 / 17.5*	26.0 / 27.0
RUSE	27.0 / 34.5	36.1 / 49.8	32.9 / 36.8	25.5 / 27.5	25.0 / 31.1	29.1 / 25.9	24.6 / 21.5*	28.6 / 32.4
YiSi1	23.5 / 31.7	35.5 / 48.8	30.2 / 35.1	21.5 / 23.1	23.3 / 30.0	26.8 / 23.4	23.1 / 20.9	26.3 / 30.4
YiSi1 SRL 18	23.3 / 31.5	34.3 / 48.3	29.8 / 34.5	21.2 / 23.7	22.6 / 30.6	26.1 / 23.3	22.9 / 20.7	25.7 / 30.4
BLEURTbase -pre	33.0 / 39.0	41.5 / 54.6	38.2 / 39.6	30.7 / 31.1	30.7 / 34.9	32.9 / 29.8	28.3 / 25.6	33.6 / 36.4
BLEURTbase	34.5 / **42.9**	43.5 / 55.6	39.2 / 40.5	31.5 / 30.9	31.0 / 35.7	35.0 / 29.4	29.6 / **26.9**	34.9 / 37.4
BLEURT -pre	34.5 / 42.1	42.7 / 55.4	39.2 / 40.6	31.4 / 31.6	31.4 / 34.2	33.4 / 29.3	28.9 / 25.6	34.5 / 37.0
BLEURT	**35.6** / 42.3	**44.2 / 56.7**	**40.0 / 41.4**	**32.1 / 32.5**	**31.9 / 36.0**	**35.5 / 31.5**	**29.7** / 26.0	**35.6 / 38.1**

Table 3: Agreement with human ratings on the WMT18 Metrics Shared Task. The metrics are Kendall Tau (τ) and WMT's Direct Assessment metrics divided by 100. The star * indicates results that are more than 0.2 percentage points away from the official WMT results (up to 0.4 percentage points away).

BLEURT: BERT-Large를 사용함

BLEURTbase: BERT-base를 사용함

***-pre**: pre-training 과정 없이 fine-tuning함

*-pre 모델을 제외한 모델은 pre-trained BERT 위에 sentence pairs로 pre-training을 진행하고, task specific fine-tuning을 진행함

Experiments

model	de-en τ / DA	fi-en τ / DA	gu-en τ / DA	kk-en τ / DA	lt-en τ / DA	ru-en τ / DA	zh-en τ / DA	**avg** τ / DA
sentBLEU	19.4 / 5.4	20.6 / 23.3	17.3 / 18.9	30.0 / 37.6	23.8 / 26.2	19.4 / 12.4	28.7 / 32.2	22.7 / 22.3
BERTscore w/ BERT	26.2 / 17.3	27.6 / 34.7	25.8 / 29.3	36.9 / 44.0	30.8 / 37.4	25.2 / 20.6	37.5 / 41.4	30.0 / 32.1
BERTscore w/ roBERTa	29.1 / 19.3	29.7 / 35.3	27.7 / **32.4**	37.1 / 43.1	32.6 / 38.2	26.3 / **22.7**	41.4 / **43.8**	32.0 / **33.6**
ESIM	28.4 / 16.6	28.9 / 33.7	27.1 / 30.4	38.4 / 43.3	33.2 / 35.9	26.6 / 19.9	38.7 / 39.6	31.6 / 31.3
YiSi1 SRL 19	26.3 / **19.8**	27.8 / 34.6	26.6 / 30.6	36.9 / 44.1	30.9 / 38.0	25.3 / 22.0	38.9 / 43.1	30.4 / 33.2
BLEURTbase -pre	30.1 / 15.8	30.4 / 35.4	26.8 / 29.7	37.8 / 41.8	34.2 / 39.0	27.0 / 20.7	40.1 / 39.8	32.3 / 31.7
BLEURTbase	31.0 / 16.6	31.3 / 36.2	27.9 / 30.6	**39.5** / **44.6**	**35.2** / 39.4	**28.5** / 21.5	41.7 / 41.6	33.6 / 32.9
BLEURT -pre	31.1 / 16.9	31.3 / **36.5**	27.6 / 31.3	38.4 / 42.8	35.0 / 40.0	27.5 / 21.4	41.6 / 41.4	33.2 / 32.9
BLEURT	**31.2** / 16.9	**31.7** / 36.3	**28.3** / 31.9	**39.5** / **44.6**	**35.2** / **40.6**	28.3 / 22.3	**42.7** / 42.4	**33.8** / **33.6**

Table 4: Agreement with human ratings on the WMT19 Metrics Shared Task. The metrics are Kendall Tau (τ) and WMT's Direct Assessment metrics divided by 100. All the values reported for Yisi1_SRL and ESIM fall within 0.2 percentage of the official WMT results.

BLEURT: BERT-Large를 사용함
BLEURTbase: BERT-base를 사용함
***-pre**: pre-training 과정 없이 fine-tuning함
*-pre 모델을 제외한 모델은 pre-trained BERT 위에 sentence pairs로 pre-training을 진행하고, task specific fine-tuning을 진행함

Robustness to Quality Drift

Pre-training으로 BLEURT가 quality drift에 대해 robust하게 만든다는 주장을 평가함

방법론

- WMT Metrics Shared Task에서 데이터를 샘플링하고, 학습을 위한 낮은 등급의 번역과 테스트를 위한 높은 등급의 번역을 유지함으로써 점점 더 도전적인 데이터셋을 생성함
- 핵심 매개변수는 데이터가 한쪽으로 쏠린 정도를 측정하는 스큐 인자 α를 사용함
- 훈련 데이터는 α가 증가함에 따라 축소됨: 가장 극단적인 경우(α=3.0)에서는 원래 훈련 기록인 5,344개의 11.9%만 사용함

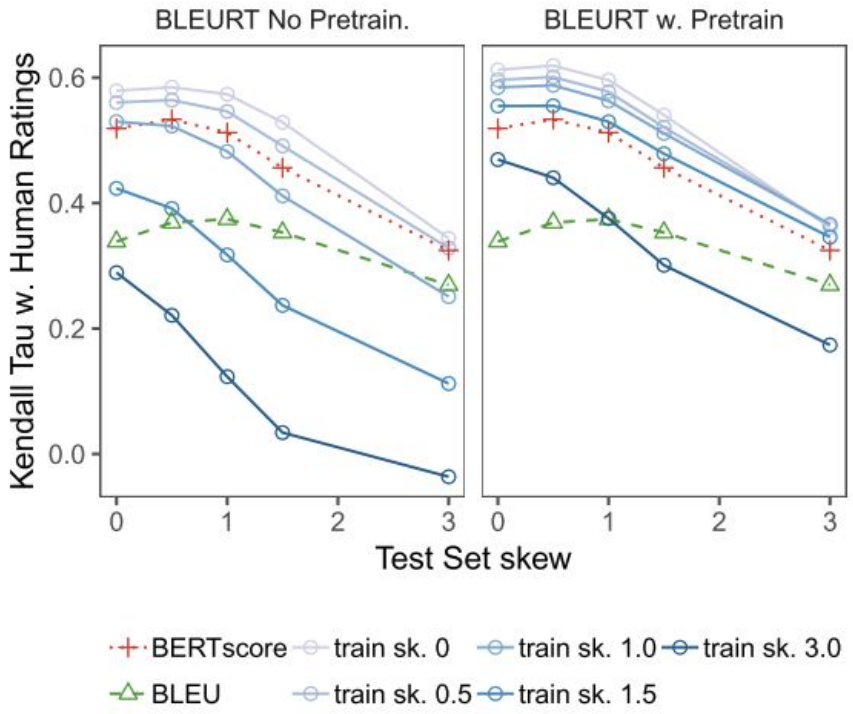


Figure 2: Agreement between BLEURT and human ratings for different skew factors in train and test.

Conclusion

- 영어에 대한 reference based text generation metric인 BLEURT를 제시함
- End-to-end로 훈련되기 때문에 BLEURT는 human evaluation를 우수한 정확도로 모델링할 수 있음
- Pre-training은 metric이 도메인과 quality drift 양쪽을 robust하게 만듬
- 추후 연구에서는 다국어 NLG 평가, 인간과 classifier 모두를 포함하는 hybrid 방식으로 연구할 예정

Thank you