TGEA Datasets and Benchmark Tasks for Error Annotations of Text Generated by Pre-trained Language Models

Qun Liu (刘群)

Huawei Noah's Ark Lab

2022 BAAI Conference - NLP Forum
Online 2022-06-02

Acknowledgements

This is a collaborative project of Tianjin University and Huawei Noah's Ark Lab:

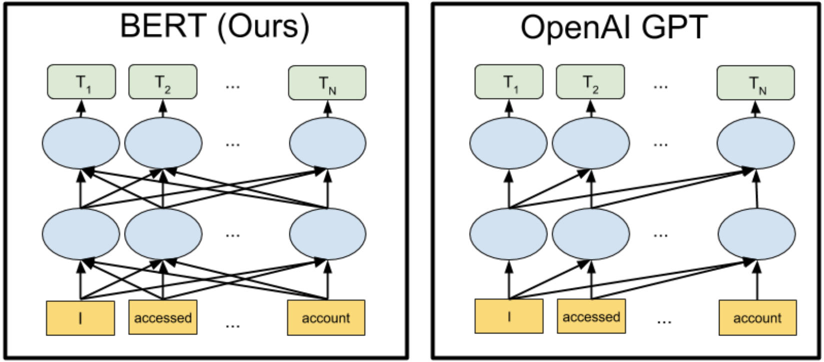
- Huibin Ge (Tianjin University)
- Chaobin You(Tianjin University)
- Xiaohu Zhao (Tianjin University)
- Minghui Xu (Tianjin University)
- Bo Peng (Tianjin University)
- Jie He (Tianjin University)
- Deyi Xiong (Tianjin University)
- Yi Liao (Huawei Noah's Ark Lab)
- Yulong Zeng (Huawei Noah's Ark Lab)
- Qun Liu (Huawei Noah's Ark Lab)

Content

Content

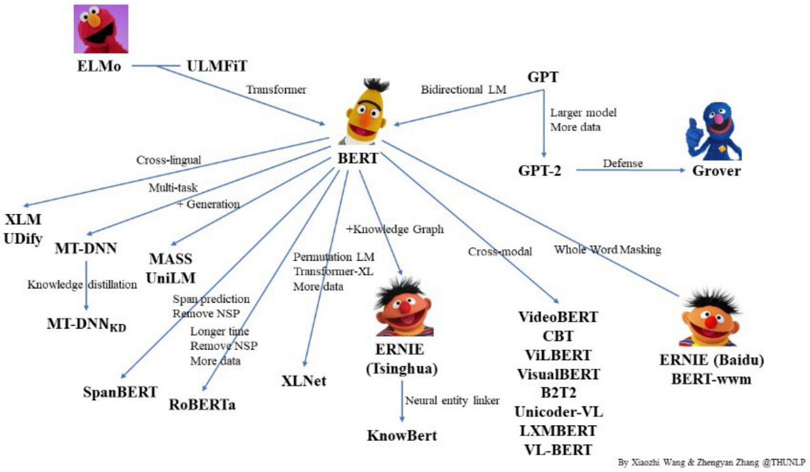
Content

Pretrained Language Models

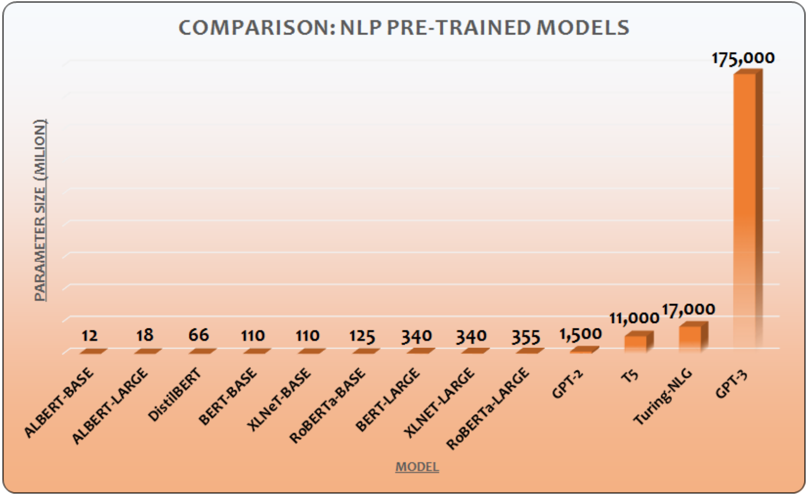


https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html

Family of Pretrained Language Models



Model Size (Number of Parameters) of Pretrained Language Models



HUAWEI NOAH'S ARK LAB

Commonsense QA with GPT-3

Q: What is your favorite animal?
A: My favorite animal is a dog.

Q: Why?
A: Because dogs are loyal and friendly.

Q: What are two reasons that a dog might be in a bad mood?
A: Two reasons that a dog might be in a bad mood are if it is hungry or if it is hot.

Q: How many eyes does a giraffe have?
A: A giraffe has two eyes.

Q: How many legs does a frog have?
A: A frog has four legs.

Q: Are there any animals with three legs?
A: No, there are no animals with three legs.

Q: Why don't animals have three legs?
A: Animals don't have three legs because they would fall over.

Commonsense QA with GPT-3

Q: How many eyes does a giraffe have?
A: A giraffe has two eyes.

Q: How many eyes does my foot have?
A: Your foot has two eyes.

Q: How many eyes does a spider have?
A: A spider has eight eyes.

Q: How many eyes does the sun have?
A: The sun has one eye.

Q: How many eyes does a blade of grass have?
A: A blade of grass has one eye.

Content

Evaluation of BERT

GLUE Results

System	MNLI-(m/mm)	QQP	QNLI	SST-2	CoLA	STS-B	MRPC	RTE	Average
	392k	363k	108k	67k	8.5k	5.7k	3.5k	2.5k	-
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.9	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	88.1	91.3	45.4	80.0	82.3	56.0	75.2
$\text{BERT}_{\text{BASE}}$	84.6/83.4	71.2	90.1	93.5	52.1	85.8	88.9	66.4	79.6
$\text{BERT}_{\text{LARGE}}$	**86.7/85.9**	**72.1**	**91.1**	**94.9**	**60.5**	**86.5**	**89.3**	**70.1**	**81.9**

Table 1: GLUE Test results, scored by the GLUE evaluation server. The number below each task denotes the number of training examples. The "Average" column is slightly different than the official GLUE score, since we exclude the problematic WNLI set. OpenAI GPT = (L=12, H=768, A=12); $\text{BERT}_{\text{BASE}}$ = (L=12, H=768, A=12); $\text{BERT}_{\text{LARGE}}$ = (L=24, H=1024, A=16). BERT and OpenAI GPT are single-model, single task. All results obtained from https://gluebenchmark.com/leaderboard and https://blog.openai.com/language-unsupervised/.

SQuAD v1.1

System	Dev EM	Dev F1	Test EM	Test F1
Leaderboard (Oct 8th, 2018)				
Human	-	-	82.3	91.2
#1 Ensemble - nlnet	-	-	86.0	91.7
#2 Ensemble - QANet	-	-	84.5	90.5
#1 Single - nlnet	-	-	83.5	90.1
#2 Single - QANet	-	-	82.5	89.3
Published				
BiDAF+ELMo (Single)	-	85.8	-	-
R.M. Reader (Single)	78.9	86.3	79.5	86.6
R.M. Reader (Ensemble)	81.2	87.9	82.3	88.5
Ours				
$\text{BERT}_{\text{BASE}}$ (Single)	80.8	88.5	-	-
$\text{BERT}_{\text{LARGE}}$ (Single)	84.1	90.9	-	-
$\text{BERT}_{\text{LARGE}}$ (Ensemble)	85.8	91.8	-	-
$\text{BERT}_{\text{LARGE}}$ (Sgl.+TriviaQA)	**84.2**	**91.1**	**85.1**	**91.8**
$\text{BERT}_{\text{LARGE}}$ (Ens.+TriviaQA)	**86.2**	**92.2**	**87.4**	**93.2**

Table 2: SQuAD results. The BERT ensemble is 7x systems which use different pre-training checkpoints and fine-tuning seeds.

Reference: https://rajpurkar.github.io/SQuAD-explorer

GLUE Benchmark

- GLUE (General Language Understanding Evaluation) benchmark
  - Distribute canonical Train, Dev and Test splits
  - Labels for Test set are not provided
- Datasets in GLUE:
  - MNLI: Multi-Genre Natural Language Inference
  - QQP: Quora Question Pairs
  - QNLI: Question Natural Language Inference
  - SST-2: Stanford Sentiment Treebank
  - CoLA: The corpus of Linguistic Acceptability
  - STS-B: The Semantic Textual Similarity Benchmark
  - MRPC: Microsoft Research Paraphrase Corpus
  - RTE: Recognizing Textual Entailment
  - WNLI: Winograd NLI

Stanford Question Answering Dataset (SQuAD)

Question: Which team won Super Bowl 50?

Passage

Super Bowl 50 was an American football game to determine the champion of the National Football League (NFL) for the 2015 season. The American Football Conference (AFC) champion Denver Broncos defeated the National Football Conference (NFC) champion Carolina Panthers 24–10 to earn their third Super Bowl title. The game was played on February 7, 2016, at Levi's Stadium in the San Francisco Bay Area at Santa Clara, California.

100k examples

Answer must be a span in the passage

A.k.a. extractive question answering

"SQuAD: 100,000+ questions for machine comprehension of text", Rajpurkar et al., 2016.
https://arxiv.org/pdf/1606.05250.pdf

Stanford Question Answering Dataset (SQuAD)

Private schools, also known as independent schools, non-governmental, or nonstate schools, are not administered by local, state or national governments; thus, they retain the right to select their students and are funded in whole or in part by charging their students tuition, rather than relying on mandatory taxation through public (government) funding; at some private schools students may be able to get a scholarship, which makes the cost cheaper, depending on a talent the student may have (e.g. sport scholarship, art scholarship, academic scholarship), financial need, or tax credit scholarships that might be available.

Along with non-governmental and nonstate schools, what is another name for private schools?

Gold answers: ① independent ② independent schools ③ independent schools

Along with sport and art, what is a type of talent scholarship?

Gold answers: ① academic ② academic ③ academic

Rather than taxation, what are private schools largely funded by?

Gold answers: ① tuition ② charging their students tuition ③ tuition

Thomas Lukasiewicz, Advanced Machine Learning: Deep Learning for NLP: Lecture 11: Question Answering, 2019

SQuAD Evaluation, v1.1

- Authors collected 3 gold answers
- Systems are scored on two metrics:
  - Exact match: 1/0 accuracy on whether you match one of the 3 answers
  - F1: Take system and each gold answer as bag of words, evaluate Precision = tp/(tp+fp), Recall = tp/(tp + fn), harmonic mean F1 = 2PR/(P+R)
    Score is (macro-)average of per-question F1 scores
- F1 measure is seen as more reliable and taken as primary
  - It’s less based on choosing exactly the same span that humans chose, which is susceptible to various effects, including line breaks
- Both metrics ignore punctuation and articles (**a**, **an**, **the** only)

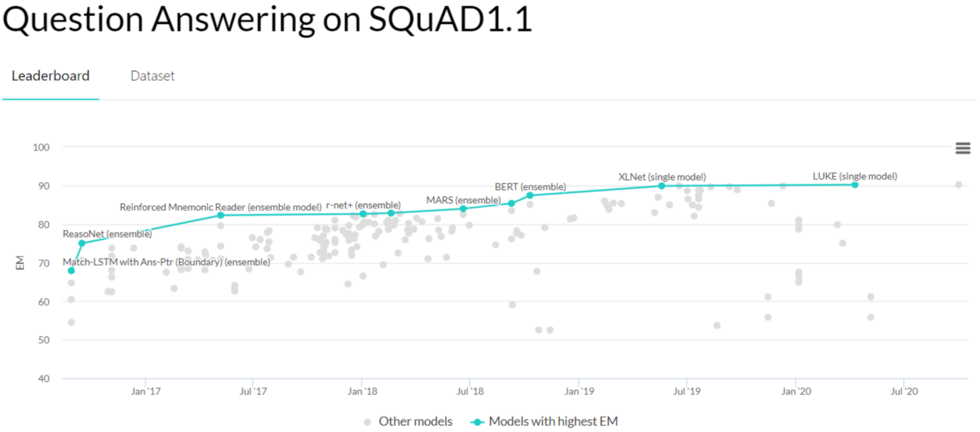
Thomas Lukasiewicz, Advanced Machine Learning: Deep Learning for NLP: Lecture 11: Question Answering, 2019

SQuAD v1.1 Leaderboard, 2019-02-07

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar et al. '16)	82.304	91.221
1 Oct 05, 2018	BERT (ensemble) *Google AI Language* https://arxiv.org/abs/1810.04805	**87.433**	**93.160**
2 Oct 05, 2018	BERT (single model) *Google AI Language* https://arxiv.org/abs/1810.04805	85.083	91.835
2 Sep 09, 2018	nlnet (ensemble) *Microsoft Research Asia*	85.356	91.202
2 Sep 26, 2018	nlnet (ensemble) *Microsoft Research Asia*	85.954	91.677
3 Jul 11, 2018	QANet (ensemble) *Google Brain & CMU*	84.454	90.490
4 Jul 08, 2018	r-net (ensemble) *Microsoft Research Asia*	84.003	90.147
5 Mar 19, 2018	QANet (ensemble) *Google Brain & CMU*	83.877	89.737
5 Sep 09, 2018	nlnet (single model) *Microsoft Research Asia*	83.468	90.133

Thomas Lukasiewicz, Advanced Machine Learning: Deep Learning for NLP: Lecture 11: Question Answering, 2019

SQuAD v1.1 Performance, upto 2020-07



HUAWEI NOAH'S ARK LAB

SQuAD 2.0

- A defect of SQuAD 1.0 is that all questions have an answer in the paragraph
- Systems (implicitly) rank candidates and choose the best one
- You don't have to judge whether a span answers the question
- In SQuAD 2.0, 1/3 of the training questions have no answer, and about 1/2 of the dev/test questions have no answer
  - For NoAnswer examples, NoAnswer receives a score of 1, and any other response gets 0, for both exact match and F1
- Simplest system approach to SQuAD 2.0:
  - Have a threshold score for whether a span answers a question
- Or you could have a second component that confirms answering
  - Like Natural Language Inference (NLI) or "Answer validation"

https://rajpurkar.github.io/SQuAD-explorer/

Thomas Lukasiewicz, Advanced Machine Learning: Deep Learning for NLP: Lecture 11: Question Answering, 2019

HUAWEI NOAH'S ARK LAB

SQuAD 2.0 Example

Genghis Khan united the Mongol and Turkic tribes of the steppes and became Great Khan in 1206. He and his successors expanded the Mongol empire across Asia. Under the reign of Genghis' third son, Ögedei Khan, the Mongols destroyed the weakened Jin dynasty in 1234, conquering most of northern China. Ögedei offered his nephew Kublai a position in Xingzhou, Hebei. Kublai was unable to read Chinese but had several Han Chinese teachers attached to him since his early years by his mother Sorghaghtani. He sought the counsel of Chinese Buddhist and Confucian advisers. Möngke Khan succeeded Ögedei's son, Güyük, as Great Khan in 1251. He

When did Genghis Khan kill Great Khan?

Gold Answers: <No Answer>

Prediction: 1234 [from Microsoft nlnet]

Thomas Lukasiewicz, Advanced Machine Learning: Deep Learning for NLP: Lecture 11: Question Answering, 2019

SQuAD 2.0 leaderboard, 2019-02-07

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Jan 15, 2019	BERT + MMFT + ADA (ensemble) *Microsoft Research Asia*	**85.082**	**87.615**
2 Jan 10, 2019	BERT + Synthetic Self-Training (ensemble) *Google AI Language* https://github.com/google-research/bert	84.292	86.967
3 Dec 13, 2018	BERT finetune baseline (ensemble) *Anonymous*	83.536	86.096
4 Dec 16, 2018	Lunet + Verifier + BERT (ensemble) *Layer 6 AI NLP Team*	83.469	86.043
4 Dec 21, 2018	PAML+BERT (ensemble model) *PINGAN GammaLab*	83.457	86.122
5 Dec 15, 2018	Lunet + Verifier + BERT (single model) *Layer 6 AI NLP Team*	82.995	86.035

Thomas Lukasiewicz, Advanced Machine Learning: Deep Learning for NLP: Lecture 11: Question Answering, 2019

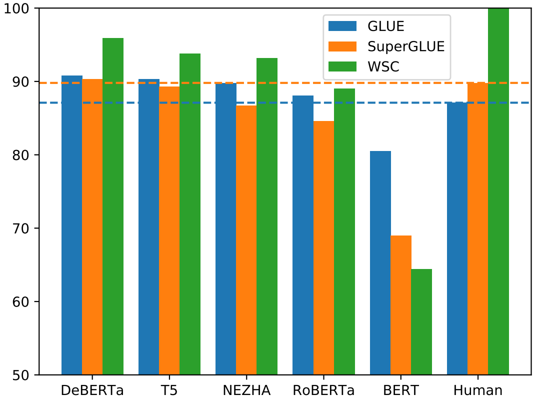
SQuAD 2.0 leaderboard, 2021-05-14

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Feb 21, 2021	FPNet (ensemble) *Ant Service Intelligence Team*	**90.871**	**93.183**
2 Feb 24, 2021	IE-Net (ensemble) *RICOH_SRCB_DML*	90.758	93.044
3 Apr 06, 2020	SA-Net on Albert (ensemble) *QIANXIN*	90.724	93.011
4 May 05, 2020	SA-Net-V2 (ensemble) *QIANXIN*	90.679	92.948
4 Apr 05, 2020	Retro-Reader (ensemble) *Shanghai Jiao Tong University* http://arxiv.org/abs/2001.09694	90.578	92.978
4 Feb 05, 2021	FPNet (ensemble) *YuYang*	90.600	92.899
5 Apr 18, 2021	TransNets + SFVerifier + SFEnsembler (ensemble) *Senseforth AI Research* https://www.senseforth.ai/	90.487	92.894

Winograd Schema Challange: Examples

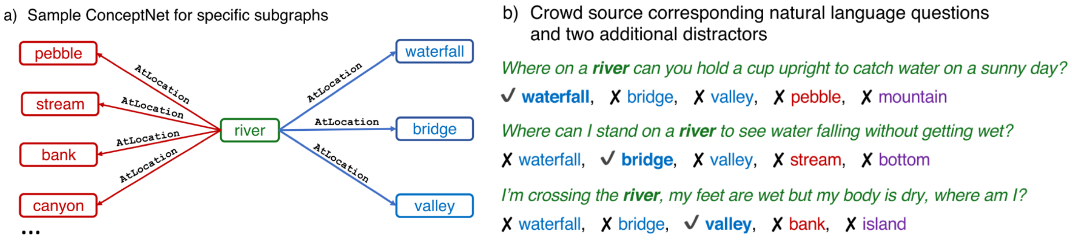
- The city councilmen refused the demonstrators a permit because they [feared/advocated] violence.
  - Question Who [feared/advocated] violence?
  - Answers The city councilmen/the demonstrators.
- The trophy doesn't fit into the brown suitcase because it's too [small/large].
  - Question What is too [small/large]?
  - Answers The suitcase/the trophy.
- Joan made sure to thank Susan for all the help she had [given/received].
  - Question Who had [given/received] help?
  - Answers Answers: Susan/Joan.

Research status of language models, 2021-05-08



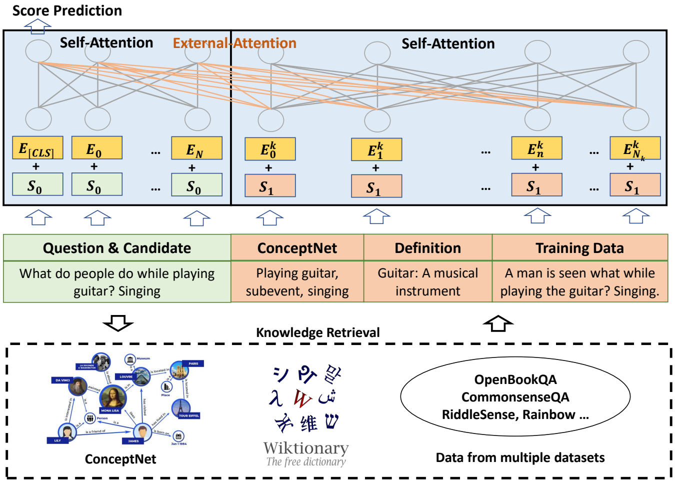
GLUE scores, SuperGLUE scores and WSC accuracies of popular language models.

CommonsenseQA



Talmor et al., CommonsenseQA: a question answering challenge targeting commonsense knowledge, NAACL-HLT 2019

CommonsenseQA - Human Parity, 2021-12-06



Method	Single	Ensemble
BERT+OMCS	62.5	-
RoBERTa	72.1	72.5
RoBERTa+KEDGN	-	74.4
ALBERT	-	76.5
RoBERTa+MHGRN	75.4	76.5
ALBERT + HGN	77.3	80.0
T5	78.1	-
UnifiedQA	79.1	-
ALBERT+KCR	79.5	-
ALBERT + KD	80.3	80.9
ALBERT + SFR	-	81.8
DEKCOR	80.7	83.3
Human	-	88.9
KEAR (ours)	**86.1**	**89.4**

Yu et al., Human Parity on CommonsenseQA: Augmenting Self-Attention with External Attention, arXiv:2112.03254

Content

Problems of Existing Benchmarks

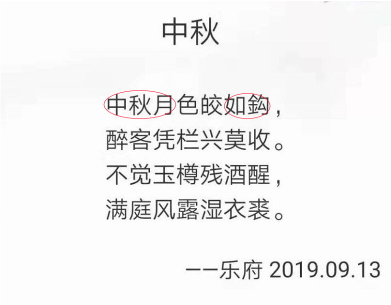
- LMs have reached a performance which is very close to or even higher than humans in many benchmark tasks.
- Actually we all know LMs are not as intelligent as humans, because we often see LMs make stupid mistakes.
- It seems the current benchmarks fail to capture the weakness of LMs.
- Why are the current benchmarks not able to capture the weakness of LMs?
- How can we design better benchmarks for LMs?

Our Assumption: Questioner's Bias

In the current benchmarks, the questions are designed by human experts manually, but:

- LMs understand languages in a very different way as humen beings do;
- The questions designed by human experts reflects the weakness of LMs according to the designers' understanding, which may not capture the real weakness of LMs.

A case of Chinese Traditional Poem Generation



中秋

中秋月色皎如鈎，
醉客凭栏兴莫收。
不觉玉樽残酒醒，
满庭风露湿衣裘。

——乐府 2019.09.13

observation

- NLG models sometimes generate interesting errors.
- It is very unlikely for human experts to design a question to detect such kind of errors in the benchmarks.

Our Solution: Let LMs Speak

- Core idea:
  - Let LMs speak (i.e. generate) freely.
  - Annotate the errors in the text generated by LMs.
  - The distribution of the annotation can reflect the weakness of the LMs.
  - The dataset can be used to build benchmarks for LMs.
- Advantages:
  - This method can avoid questioner's bias.
  - The analysis can shed light on the way to improve LMs.
  - It mimics the way of human language learning by speaking and correction by their parents, which is crutial for children to learn their mother languages.

Our Method

Benchmarking PLMs through the texts they generates:

- A collection of sentences are generated by NLG models;
- Design an error taxonomy for text generation errors;
- Define an specification for error annotation;
- Annotate errors by crowdsourcing;
- Analysis the error distribution;
- Design benchmarks using the error annotation dataset.

Content

TGEA: An Error-Annotated Dataset and Benchmark Tasks for Text Generation from Pretrained Language Models

TGEA: An Error-Annotated Dataset and Benchmark Tasks for Text Generation from Pretrained Language Models

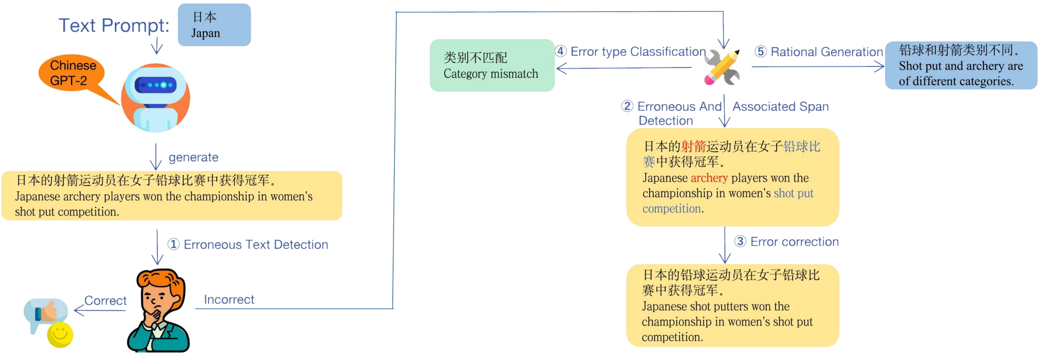
Jie He[†*], **Bo Peng**[§*], **Yi Liao**[§], **Deyi Xiong**[†] and **Qun Liu**[§]
[†] College of Intelligence and Computing, Tianjin University, Tianjin, China
[§] Huawei Noah's Ark Lab, Hong Kong, China
{jieh, dyxiong}@tju.edu.cn,
{peng.bo2, liaoyi9, qun.liu}@huawei.com

(In Proceedings of ACL2021)

- We propose TGEA (Text Generation Error Annotation), an Error-Annotated Dataset and Benchmark Tasks for Text Generation from Pretrained Language Models;
- We propose an Error Taxonomy for TGEA annotation;
- We analysis the error type distribution of TGEA to better understand the problems of current PLMs;
- We propose benchmark tasks based on TGEA dataset.

HUAWEI NOAH'S ARK LAB

Dataset creation overview



Error annotation process

There are 5 steps of annotation:

1. Erroneous text detection
2. Erroneous and associated span detection
3. Error correction
4. Error type classification
5. Rational generation

Error Taxonomy - Inappropriate combination

- Example

  医生当即将刘莉的手术(囊肿)切除，并建议患者住院观察。

  The doctor removed Liu Li’s surgery (tumor) and suggested that the patient be hospitalized for observation.

- Subtypes:
  - Subject-predicate inappropriate combination
  - Predicate-object inappropriate combination
  - Subject-object inappropriate combination
  - Modifier inappropriate combination
  - Function word inappropriate combination

Error Taxonomy - Missing

- An example:

  在这里，有众多新闻记者和游客参加_(活动)。

  Here, many journalists and tourists are taking part in _ (activities).

- Subtypes:
  - Subject missing
  - Predicate missing
  - Object missing
  - Modifier missing
  - Function word missing

Error Taxonomy - Redundancy

- An example:

  但是一些外资银行，尤其是外资银行，对我国民营经济的发展还有不少误解或偏见。

  However, some foreign banks, especially foreign banks, still have many misunderstanding or prejudices about the development of China’s private economy.

- Subtypes:
  - Subject redundancy
  - Predicate redundancy
  - Object redundancy
  - Modifier redundancy
  - Function word redundancy

Error Taxonomy - Discourse error

- An examples:

  在婚姻变得更为不好的时候，对她来说这是痛苦的。但是当她(它)发生变化时，她必须做出调整。

  It was painful for her when the marriage got worse. But when she (it) changed, she had to adjust.
- Subtype:
  - Coreference error

HUAWEI NOAH'S ARK LAB

Error Taxonomy - Commonsense error

- An example:
  在国际市场上，如果信用等级越高(低)，投资者就越不会太放心。
  In the international market, the higher (lower) the credit rating, the less reassured investors are.
- Subtypes:
  - Space error
  - Time error
  - Number error
  - Motivation error
  - Emotional reactions error
  - Causation error
  - Taxonomy error
  - Behaviors error

Error Taxonomy - Overview

Level-1 Error Type	Level-2 Error Type	Example
Inappropriate Combination	Subject-Predicate	目前,该市的小说 [话剧]《我是党员、我的团员》、《我是小老头》、《小小老师》、《小小一个农家娃》正在上演。 At present, the city’s novels [drama] *I am a Party member and This is My League Member*, *Little Old Man Like Me*, *Little Teacher*, *A Little Farm Boy* are on stage.
	Predicate-Object	由我主持，我要带大家去感受一下大赛主题设置的感受 [氛围]。 As a host, I will take you to experience the feel [atmosphere] shown from the theme of the competition.
	Subject-Object	女足的队员 [任务] 就是一个球，能够把球踢好，就是她们最大的资本。 The players [task] of women’s football team is a ball, and playing the ball well is their biggest capitals.
	Modifier	另一方面，煤炭企业面临着煤矿安全的矛盾 [问题]。 On the other hand, coal enterprises are facing the contradiction [problem] of coal mine safety.
	Function Word	因此，我对 [因为]自身的过错作出了自己应当承担的责任。 Therefore, to [because of] my own fault, I took my own responsibility.
Misssing	Subject	当他回到车间时，[车间]已经有了明显的变化。 When he returned to the workshop, [the place] had been a marked change
	Predicate	这时候我们一开始就有机会扳平比分，但是我们没有 [抓住]机会。 We had a chance to equalise at the beginning, but we didn’t [caught] chance.
	Object	一、坚持解放思想,转变观念,推进社会主义物质文明和精神 [文明]。 1. Persisting in emancipating the mind, changing ideas and promoting socialist material civilization and spiritual [civilization].
	Modifier	在国内成立水牛研究中心，有利于增强 [水牛对]自然条件和人工环境的适应能力。 The establishment of Buffalo Research Center in China is conducive to enhance the adaptability [of buffalo] to natural conditions and artificial environment.
	Function Word	他的儿子 [在]上一届奥运会夺得冠军，并且获得当年世界锦标杯赛金牌。 His son won champion [in] the last Olympic Games and won the gold medal in the World Championship Cup that year.

Table 7: Examples of level-2 error types in TGEA. Underwaved words are erroneous words while underlined words are associated words. Words in ”[]” are corrections to erroneous words.

Error Taxonomy - Overview

Level-1 Error Type	Level-2 Error Type	Example
Redundancy	Subject	但一些外资银行，尤其是外资银行[]，对我国民营经济的发展还有不少误解或偏见。 However, some foreign banks, especially foreign banks[], still have many misunderstandings or prejudices about the development of China's private economy.
	Predicate	这也是所有关心[]关心孩子成长的人的共同心声。 This is also the common voice of all those who care about[] care about children's growth
	Object	同时，学校也开展丰富多彩、有益于学生的社会实践活动、社会实践[]，丰富他们的课余生活。 At the same time, the school also carries out colorful and beneficial social practice activities, social practice[] to enrich their after-school life.
	Modifier	它们的皮毛很有光泽,可以用肉眼很难[]看出来。 Their fur is so shiny that we can see with naked eyes hardly[].
	Function Word	他是被迫进入位于市中心的一个警察局的，随后[]他被带到警察局，并遭到了手铐和警犬的威吓。 He was forced into a police station in the center of the city, then[] he was taken to the police station, where he was intimidated by handcuffs and police dogs.
Discourse Error	Coreference	在婚姻变得更为不好的时候，对她来说这是痛苦的。但是当她[它]发生变化时，她必须做出调整。 It was painful for her when the marriage got worse. But when she [it] changed, she had to adjust.

Table 7: Examples of level-2 error types in TGEA. Underwaved words are erroneous words while underlined words are associated words. Words in ”[]” are corrections to erroneous words.

HUAWEI NOAH'S ARK LAB

Error Taxonomy - Overview

Level-1 Error Type	Level-2 Error Type	Example
Commonsense Error	Space	他说,中美两国是近邻 [朋友],关系很好,中美合作富有创造性。 He said that China and the United States are close neighbors [friends] with good relations and creative cooperation.
	Time	国庆 [元旦]假期期间，各大汽车经销商将会以怎么样的姿态迎接新的一年? During the National Day [New Year's Day] holiday, how will major auto dealers greet the new year?
	Number	而在4月份，中国石化、招商银行、万科、上海汽车、g长安和g天威成为了最活跃的5 [6]只股票。 In April, Sinopec, China Merchants Bank, Vanke, SAIC, G Changan and G Tianwei became the most active 5 [6] stocks.
	Motivation	近日，李老的胃疼难忍，为治疗病情已连续工作 [休息]两天了，而且病情非常严重，他一躺就是几天。 Recently, Lao Li's stomach ache is unbearable. He has been working [resting] for two consecutive days to treat his illness, and his illness is very serious. He has been lying down for several days.
	Emotional Reactions	对于学校为了保障广大师生员工的安全，采取这些措施，我们深感遗憾[欣慰]。 We are very sorry [pleased] that the school has taken these measures to ensure the safety of students, teachers, and other staff.
	Causation	据悉，由于身价低廉[高昂]，子淇在国内是很少有人请得到的大牌艺人之一。 It is reported that Ziqi is one of the few famous artists that are difficult to invite in China because of his low [high] value.
	Taxonomy	酱 [花生] 油是植物油中的一种，食用后可以对皮肤有非常好的润泽效果。 Soy sauce [Peanut Oil] is a kind of vegetable oil, which has a very good moisturizing effect on the skin after eating.
	Behaviors	一位中国官员表示：我们将在近期和俄罗斯、中国 [法国] 等国合作进一步推广这一系列行动，以此来缓解人们对恐怖主义威胁的忧虑。 In the near future, we will work with Russia, China [France] and other countries to further promote this series of actions to ease people's concerns about the threat of terrorism, a Chinese official said.

Table 7: Examples of level-2 error types in TGEA. Underwaved words are erroneous words while underlined words are associated words. Words in "[]" are corrections to erroneous words.

Machine-generated texts collection

1. Randomly sample sentences generated from NEZHA-Gen with a *prompt pool*
   - Prompts are nouns.
   - Prompts are sampled from top $[40\%, 60\%]$ frequent words in the corpus.
2. Filter out *noisy texts*
   - Texts containing no more than 15 characters.
   - Texts where Chinese characters account for less 70% of all characters.
   - Uncompleted sub-sentences in the beginning or the end of the texts are trimmed.

Annotation quality control

- Quality control protocol:
  1. Train 2 reviewers with 1,000 examples
  2. Test 200 candidate workers with 500 examples
  3. Let candidates who reached $>$ 90% accuracy participate the final annotation
  4. Carry out iterative verification and amendment
- Inter-Annotator Agreement (IAA):

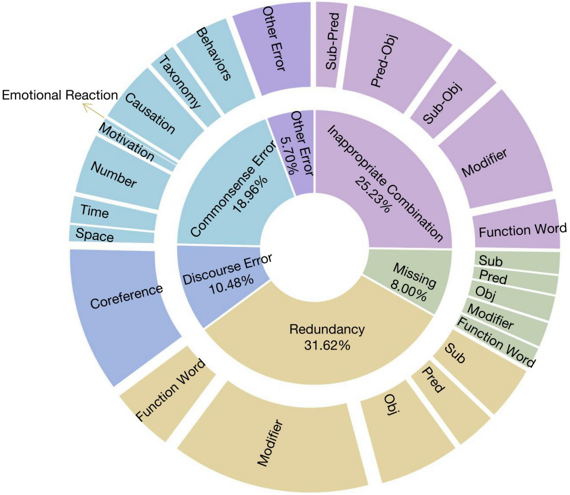
Task	(1)	(2)	(4)
IAA(%)	87.5	51.2	73.3

Dataset statistics

	Train	Dev	Test	All
#text	37,646	4,706	4,706	47,058
w/ 0 error	27,906	3,488	3,488	34,882
w/ 1 error	8,413	1,055	1,052	10,520
w/ 2 error	1,169	141	149	1,459
w/ 3 error	141	18	15	174
w/ 4 error	17	4	2	23
Tokens	966,765	120,889	121,065	1,208,719
Vocab	44,598	16,899	16,745	48,547
Avg. tokens	25.68	25.69	25.73	25.68
Avg. t.err	2.92	3.09	2.95	2.94
Avg. t.assoc	4.30	4.39	3.89	4.27
Avg. d.e-a	6.99	7.29	7.10	7.03
Avg. t.rationale	8.74	8.72	8.75	8.74

Table: Data statistics of TGError. Avg.t.err/Avg.t.assoc: the average number of tokens in erroneous/associated text spans. Avg.t.rationale: the average number of tokens in rationales. Avg.d.e-a: the average distance between a erroneous span and its associated span.

Error type distribution



HUAWEI NOAH'S ARK LAB

Content

Introduction

TGEA Dataset

TGEAv2 Dataset

TGEAv2 Benchmark Tasks

Conclusion

Evaluation Proposal (in submission)

Task Proposal: Towards Semantically Robust Generation from Pretrained Language Models

Huibin Ge[†], **Chaobin You**[†], **Xiaohu Zhao**[†], **Minghui Xu**[†],
Yulong Zeng[§], **Qun Liu**[§] and **Deyi Xiong**[†]
[†] College of Intelligence and Computing, Tianjin University, Tianjin, China
[§] Huawei Noah's Ark Lab, Hong Kong, China
{gehuibin, chaobinyou, zhaoxiaohu, xuminghui}@tju.edu.cn
{zengyulong, qun.liu}@huawei.com dyxiong.@tju.edu.cn

TGEAv2 Dataset: Improvement

- More generative LMs adopted:
  - Original: NEZHA-Gen
  - New version: NEZHE-Gen, GPT-2, PANGU-α, CPM
- Replacing Erroneous and Assicated Span with Minumal Set of Error-related Words (MiSEW);
- More prompting strategy: Nouns $\rightarrow$ Nouns, Phrases, Sentences;
- More sampling strategy: top-k sampling $\rightarrow$ top-k & top-p sampling;
- More tempratures tried in decoding;
- No Rationale annotation.

Replacing Erroneous and Assicated Span with Minumal Set of Error-related Words (MiSEW)

- Problems of Errornous Span and Associated Span Annotation
  - Errornous spans are ambiguous: there are multiple way to correct the sentences.
  - Associated spans may be discontinous.
- Solution: Minimal Set of Error-related Words (MiSEW)
  - MiSEW should contain errors;
  - MiSEW should be self-contained semantically:
    - The errors should be understandable by reading the MiSEW only.
  - MiSEW should be minumal:
    - No word can be removed from MiSEW while keep meeting the other two constrains.

Settings of Generative PLMs

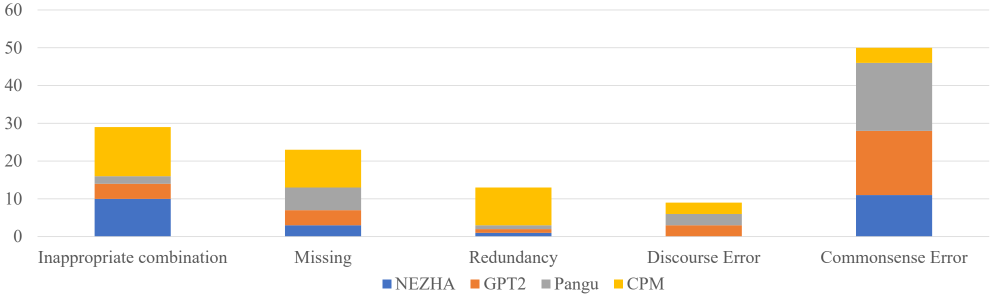
Model	NEZHA-Gen	GPT-2-medium	CPM	PanGu-α
hidden_size	768	1024	2560	2560
num_hidden_layers	12	24	32	32
num_attention_heads	12	16	32	32
intermediate_size	3072	4096	10240	10240
hidden_act	gelu	gelu	gelu	gelu
hidden_dropout_prob	0.1	0.1	0.1	0.1
attention_probs_dropout_prob	0.1	0.1	0.1	0.1
max_position_embeddings	512	1024	1024	1024
parameters	110M	340M	2.6B	2.6B

Prompting strategy, sampling strategy and tempratures

	Nezha-Gen				GPT-2-medium				CPM				Pangu-α			
Strategies	N	P	S	T	N	P	S	T	N	P	S	T	N	P	S	T
p=0.9 t=0.9	14	23	7	44	10	21	7	38	15	29	10	54	8	24	8	40
p=0.9 t=0.8	**7**	**13**	**5**	**25**	11	19	8	38	13	27	9	49	5	25	9	39
p=0.8 t=0.9	9	12	5	26	12	17	5	34	13	24	8	45	5	21	6	32
p=0.8 t=0.8	9	14	6	29	**8**	**15**	**6**	**29**	**13**	**20**	**7**	**40**	**6**	**18**	**6**	**30**
k=30 t=0.9	8	17	9	34	14	23	10	47	13	26	10	49	9	22	10	41

HUAWEI NOAH'S ARK LAB

Error distribution of different PLMs



HUAWEI NOAH'S ARK LAB

Data sizes

	Train	Dev	Test	Total
TGEA	37,646	4,706	4,706	47,058
Ours	160,000	20,000	20,000	200,000

An Example

Step1: Erroneous text detection

Incorrect

Step2: Erroneous span detection

该校把2015年下半年作退学处理的18名本科生名单打印出来,并将其中15人列入黑名单(剩下11人因不满学校被退学而提出辞职)。

The school printed out the list of 18 undergraduates who were with-drawn in the second half of 2015, (The remaining 11 resigned due to the dissatisfaction with the school being withdrawn).

Step3: Error Correction

该校把2015年下半年作退学处理的18名本科生名单打印出来,并将其中15人列入黑名单(剩下3人因不满被退学而提出申诉)。

The school printed out the list of 18 undergraduates who were with-drawn in the second half of 2015, (The remaining 3 file a grievance due to the dissatisfaction with being withdrawn).

Step4: MiSEW detection

18名 15人 剩下11	不满 学校被 退学	本科生 提出 辞职
18 15 remaining 11	dissatisfaction with the school being with-drawn	undergraduates resigned

Step5: Erroneous type classification

常识错误-数学错误	成分多余-宾语多余	常识错误-行为错误
Commonsense Error -Number	Redundancy -Object	Commonsense Error -Behaviors

Content

Tasks

- Erroneous Text Detection
- MiSEW Detection
- Erroneous Span Detection
- Error Type Classification
- Error Correction
- PLM Generation Enhancement

Task 1 - Erroneous Text Detection

- Task definition
  - This is the same as defined in TGEA, which is to automatically identify whether a given machine-generated text is erroneous.
- Evaluation
  - Error detection accuracy is used as the evaluationn metric.

Task 2 - MiSEW Detection

- Task definition
  - This is a task that automatically predicts the minimal set of error-related words given an erroneous text. This can be done as sequence labeling by regarding words in MiSEW as positive words and other words in the erroneous text as negative words.
- Evaluation
  - F_1, widely used in sequence labeling tasks, can serve as the evaluation metric for this task.

Task 3 - Erroneous Span Detection

- Task definition
  - This is to detect erroneous spans for a given erroneous text. The task can be performed as a separate task, or a joint task with MiSEW detection or a pipeline task from the out- put of MiSEW detection.
- Evaluation
  - We use exact match rate (EM) and macro-averaged F_1 as evaluation metrics for this task.

Task 4 - Error Type Classification

- Task definition
  - Again this is a text classifi cation task. We perform two levels of classification: level-1 error type detection in the form of 5-way classification and a more challenging and fine-grained level-2 error type detection in the form of 24-way classification.
- Evaluation
  - We use classification accuracy as the metric for both level-1 and level-2 error type classification.

Task 5 - Error Correction

- Task definition
  - This is different from the generative error correction task as proposed in grammatical error correction and TGEA. With the detected MiSEW and erroneous span, we define error correction as a prediction task that predict words to replace words in the erroneous span. Such definition enables different methods to be used in this task, e.g., masked language modeling, causal language modeling.
- Evaluation
  - Precision, recall and $F_{0.5}$ scores are used as evaluation metrics.
  - As an erroneous text may has multiple erroneous spans, we average evaluation scores from all erroneous spans for the last two tasks.

Task 6 - PLM Generation Enhancement

- Task definition
  - In this task, we encourage participants to explore the entire annotated trainingdataset in different ways (e.g., fine-tuning, contrastive learning) to improve the generation capability of PLMs so that they could be able to avoid making errors annotated in the training data.
- Evaluation:
  - We propose two different methods to evaluate this task:
    - Pairwise Comparison.
    - Word Prediction.
  - For both evaluation methods, we use accuracy as evaluation metric.
  - Furthermore, in order to evaluate the relative improvement achieved by the same PLM, we ask each participant of this task to submit two prediction results: one for the original PLM model and the other for the PLM model enhanced with the error-annotated data.

Competition and Leaderboard

- We will organize TGEAv2 evaluation competitions;
- We will maintain a leaderboard for TGEAv2 benchmark tasks.

Content

Content

Thank you!

把数字世界带入每个人、每个家庭、
每个组织，构建万物互联的智能世界。

Bring digital to every person, home and organization for a fully connected, intelligent world.

HUAWEI