Benchmarking the Ability of Commonsense Understanding and Reasoning for Pretrained Language Models

Qun Liu (刘群)

Huawei Noah's Ark Lab

The 22nd Chinese Lexical Semantics Workshop (CLSW2021)
Online 2021-05-15

HUAWEI

Content

Content

A photo of mine during CLSW2002, Taipei

Word Similarity Computing based-on How-Net

Qun Liu

FOLLOW

Noah's Ark Lab, Huawei
Verified email at huawei.com - Homepage

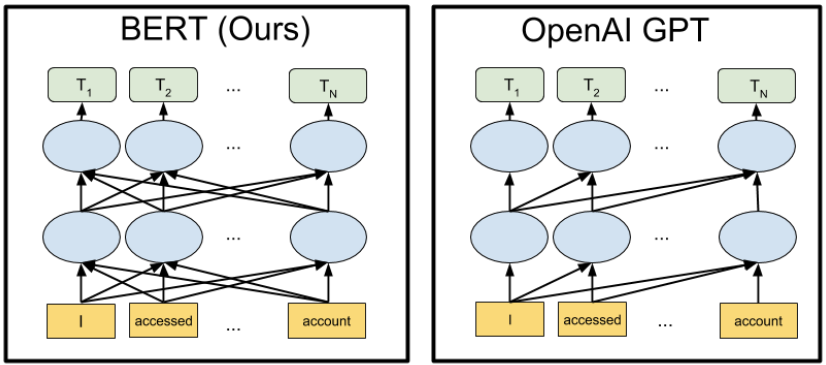
Computational Linguistics Natural Language Processing Machine Translation
Chinese Language Process…

TITLE	CITED BY	YEAR
基於《知網》的辭彙語義相似度計算 劉群, 李素建 中文計算語言學期刊 7 (2), 59-76	958 *	2002
HHMM-based Chinese lexical analyzer ICTCLAS HP Zhang, HK Yu, D Xiong, Q Liu Proceedings of the second SIGHAN workshop on Chinese language processing …	633	2003
Findings of the 2017 conference on machine translation (wmt17) O Bojar, R Chatterjee, C Federmann, Y Graham, B Haddow, S Huang, … Proceedings of the Second Conference on Machine Translation, 169-214	494 *	2017
Tree-to-string alignment template for statistical machine translation Y Liu, Q Liu, S Lin Proceedings of the 21st International Conference on Computational …	451	2006
Maximum entropy based phrase reordering model for statistical machine translation D Xiong, Q Liu, S Lin Proceedings of the 21st International Conference on Computational …	317	2006

Content

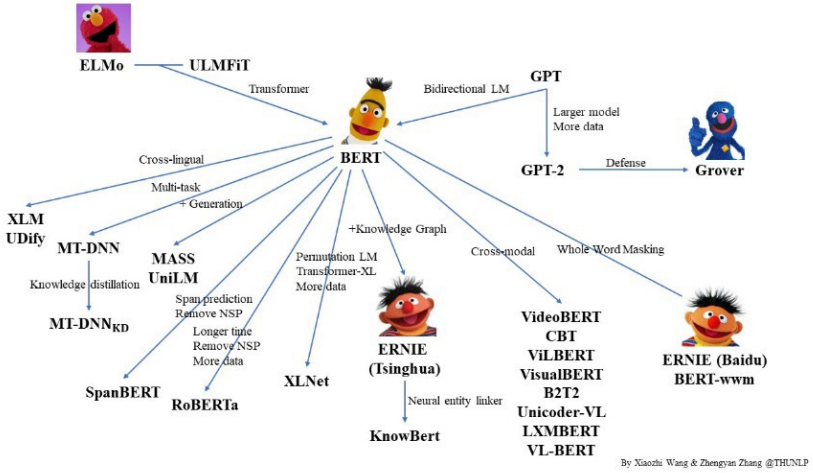
Content

Pretrained Language Models

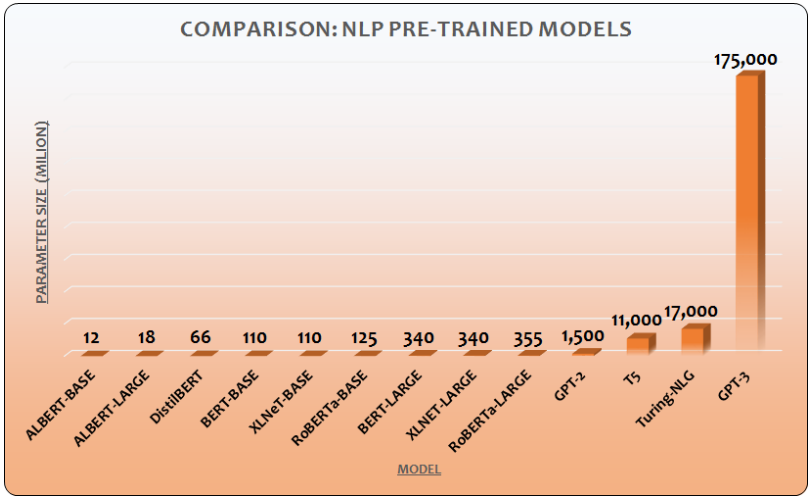


https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html

Family of Pretrained Language Models



Model Size (Number of Parameters) of Pretrained Language Models



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) 392k	QQP 363k	QNLI 108k	SST-2 67k	CoLA 8.5k	STS-B 5.7k	MRPC 3.5k	RTE 2.5k	Average -
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

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.

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

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

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

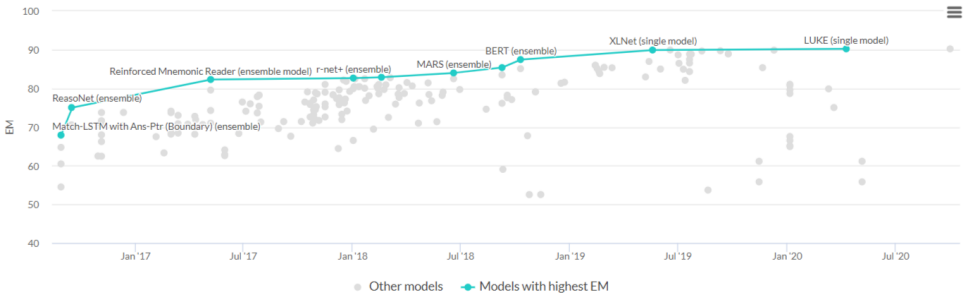
Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar et al. '16)	82.304	91.221
1 Oct 05, 2018	BERT (ensemble) *Google AI Language* https://arxiv.org/abs/1810.04805	**87.433**	**93.160**
2 Oct 05, 2018	BERT (single model) *Google AI Language* https://arxiv.org/abs/1810.04805	85.083	91.835
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

Question Answering on SQuAD1.1

Leaderboard Dataset



HUAWEI

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

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

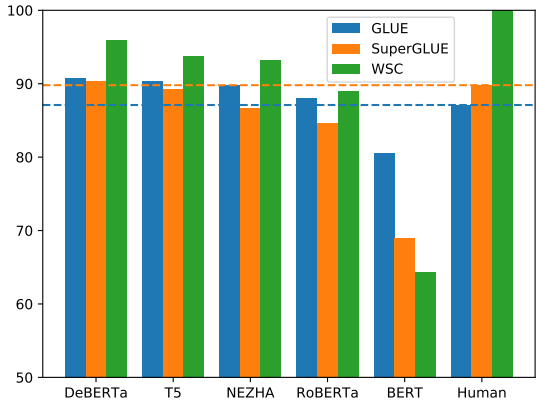
Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Jan 15, 2019	BERT + MMFT + ADA (ensemble) *Microsoft Research Asia*	**85.082**	**87.615**
2 Jan 10, 2019	BERT + Synthetic Self-Training (ensemble) *Google AI Language* https://github.com/google-research/bert	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

Research status of language models



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

Content

Problems of Existing Benchmarks

- Summary:
  - LMs have reached or surpassed human performance in some tasks;
  - There is still a big gap between LMs and human in commonsense understanding and reasoning.
  - How can we design better benchmarks to avoid this 'big gap'?
- Questioner's bias:

  The questions are designed by human experts manually，so:
  - It may not spot the main problems of PLMs on commonsense understanding and reasoning.
  - The PLMs are able to CHEAT by utilizing highly related information rather than real causality.

Research Questions

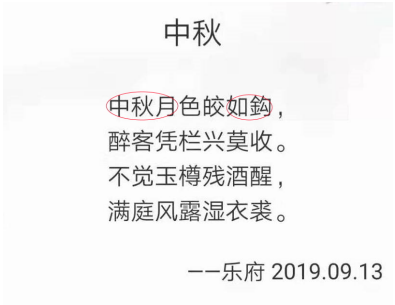
Research Questions #1

How can we benchmark the ability of commonsense understanding and reasoning for pretrained language models comprehensively?

Research Questions #2

How can we avoid questioners' bias in the design of the benchmarks?

A case of Chinese Traditional Poem Generation



observation

- NLG models sometimes generate interesting errors.
- Such errors have not been explored and analysed.

Basic Idea

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;
- Annotation by crowdsourcing;
- Analysis.

Advantages:

- Questioner's bias can be avoided because the questions are generated by PLMs rather than designed by human experts;
- It mimics the way of human language learning by speaking and correction by their parents.

Our Work: 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

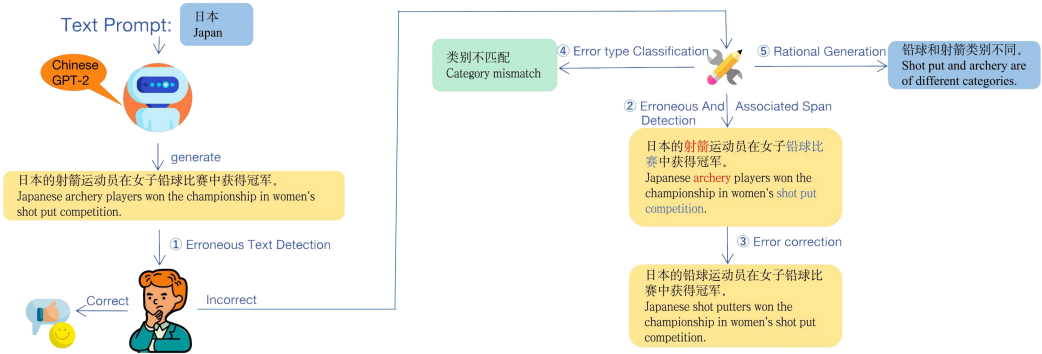
Jie He[†*], **Bo Peng**[§*], **Yi Liao**[§], **Deyi Xiong**[†] and **Qun Liu**[§]
[†] College of Intelligence and Computing, Tianjin University, Tianjin, China
[§] Huawei Noah's Ark Lab, Hong Kong, China
{jieh, dyxiong}@tju.edu.cn,
{peng.bo2, liaoyi9, qun.liu}@huawei.com

(To appear in the 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.

Content

Dataset creation overview



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

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

Machine-generated texts collection

1. Randomly sample sentences generated from NEZHA-Gen with a *prompt pool*
   - prompts must be noun
   - prompts must rank in the range of top $[40\%, 60\%]$ 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

Error annotation

As shown in a previous figure, 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

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 IAA:

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

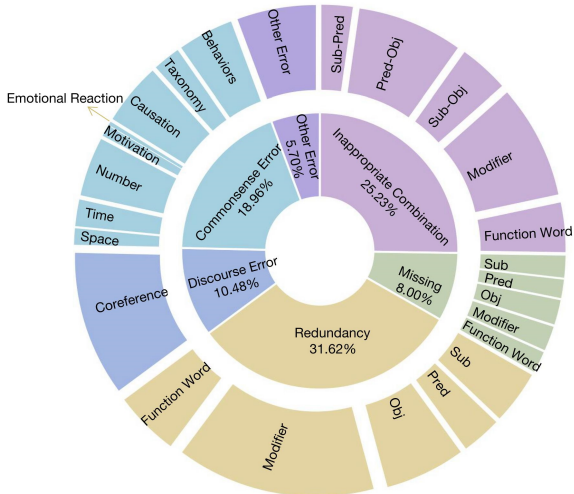
Content

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



Content

Task definition

- Erroneous text detection:

  A **text classification** task to judge whether a given text is erroneous.

- Erroneous span and associated span detection:

  A **token classification** task to recognize the erroneous and associated text spans simultaneously.

- Error type classification:

  A **text classification** task to predict the error type for a given erroneous text.

Task definition

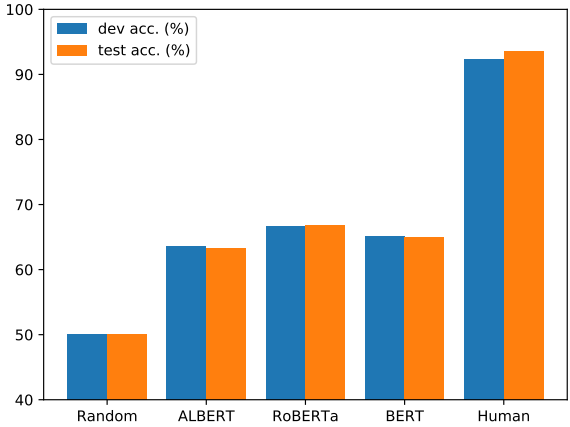
- Error correction:

  A **text generation** task to generate the correction given for a given erroneous text.

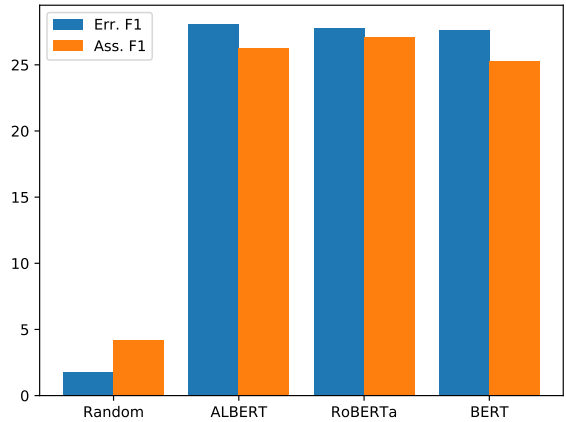
- Rational generation:

  A **text generation** task to generate an explanation with respect to text generation errors from an erroneous text.

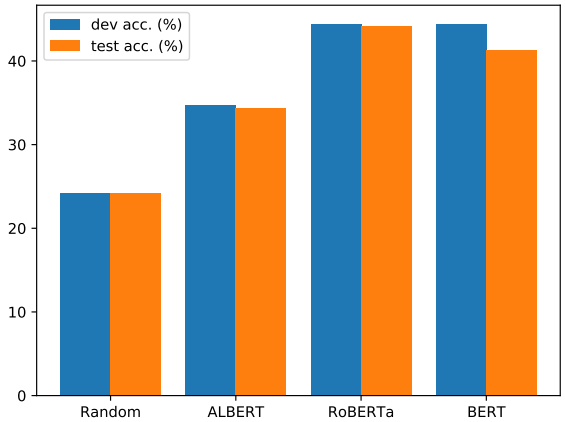
Baseline - Erroneous text detection baselines



Baseline - Span detection baselines



Baseline - Error type classification



Baseline - Generation tasks

- Error correction:

Model	Dev			Test		
	P (%)	R (%)	$F_{0.5}$ (%)	P (%)	R (%)	$F_{0.5}$ (%)
BERT	0.62	6.49	0.76	0.60	6.30	0.74
RoBERTa	0.78	4.07	0.93	0.82	4.15	0.98

- Rational generation:

  Model: NEZHA-Gen

	BLEU (%)	Rouge-L (%)	BERT_Score (%)
dev set	0.06	9.17	56.58
test set	0.06	9.02	56.17

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