ACL2020 | 对话数据集Mutual：论对话逻辑，BERT还差的很远

原创 rumor酱 夕小瑶的卖萌屋 4月13日

来自专辑
卖萌屋@自然语言处理

一只小狐狸带你解锁 **炼丹术&NLP** 秘籍

本文为MuTual论文作者的特别约稿
编辑：rumor酱、夕小瑶

前言

自然语言处理是人工智能领域的掌上明珠，而人机对话则是自然语言处理领域的最终极一环。

以BERT为代表的预训练模型为自然语言处理领域带来了新的春天，在人机对话问题上也不例外。检索式多轮对话任务中，最有名的对话数据集就是Ubuntu Dialogue Corpus了，ACL2018提出的DAM是76.7%的 $R_{10}@1$，然而基于BERT来做却直接刷到了85.8%的$R_{10}@1$，93.1%的$R_{10}@2$和高达98.5%的$R_{10}@5$，已经基本逼近了人类的表现（英语差的可能已被BERT超越），这让很多研究检索式聊天机器人的小伙伴直呼这个领域没法继续往下做了。。

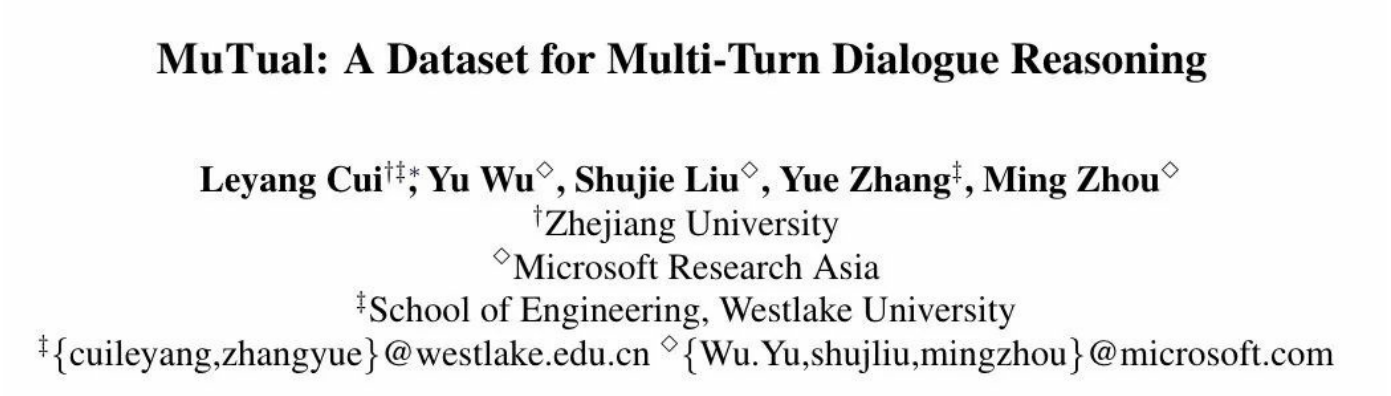
那么问题来了，**既然聊天机器人在BERT的带领下超越人类了，为什么跟我打交道的聊天机器人依然宛如人工智障？？？**

一言以蔽之，上个时代的对话数据集太弱了！！！

相信很多和聊天机器人对（liao）话（sao）过的小伙伴们都有感觉，就是每句话都回复的没什么毛病，但它像是只有三秒的记忆时间，回复的内容和前文的连贯性很差，甚至会出现自相矛盾的语句。比如

我：吃饭了吗
机器人：吃了个苹果，最近在减肥。。。
我：你不胖呀
机器人：我不要减肥
我：

当前的对话模型往往选择出的回复相关性较好，**但是经常出现常识和逻辑错误**。由于现有的大部分检索式对话数据集都没有正面刚这种对话逻辑问题，导致评价指标也无法直接反映一个模型对对话逻辑的掌握程度。针对此问题，**西湖大学联合微软研究院提出了多轮对话推理数据集MuTual**。



MuTual: A Dataset for Multi-Turn Dialogue Reasoning

Leyang Cui[†‡*], Yu Wu[◇], Shujie Liu[◇], Yue Zhang[‡], Ming Zhou[◇]
[†]Zhejiang University
[◇]Microsoft Research Asia
[‡]School of Engineering, Westlake University
[‡]{cuileyang,zhangyue}@westlake.edu.cn [◇]{Wu.Yu,shujliu,mingzhou}@microsoft.com

相比现有的其他检索式聊天数据集，MuTual要求对话模型具备常识推理能力；相比阅读理解式的推理数据集，MuTual的输入输出则完全符合标准检索式聊天机器人的流程。因此，MuTual也是目前最具挑战性的对话式数据集。测试过多个模型后，目前最佳的RoBERTa表现仅为70分左右，和人类的表现存在20多分的巨大差距。

此篇论文也发表在了ACL2020上。

论文地址：http://arxiv.org/abs/2004.04494

github地址：https://github.com/Nealcly/MuTual

arxiv访问慢的小伙伴也可以在订阅号后台回复关键词【**0413**】下载论文PDF。

数据集特点

现有的检索式聊天机器人数据集，诸如Ubuntu，Douban，对于给定的多轮对话，需要模型在若干候选回复中，选出最合适的句子作为对话的回复。然而这些数据集主要关注模型能否选出相关性较好的回复，并不直接考察模型的推理能力。随着BERT等预训练模型的涌现，此类数据集的测试集合已经达到了很好的效果。

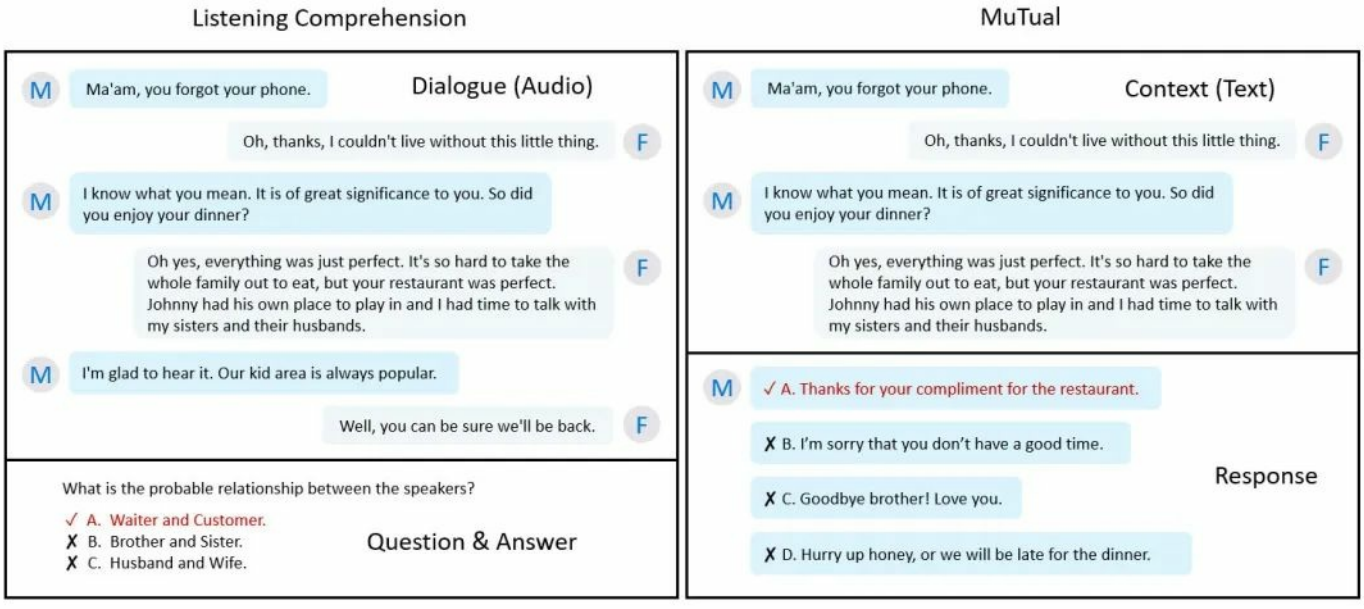
已有的针对推理的数据集(DROP, CommonsenseQA, ARC, Cosmos等)大多被设计为阅读理解格式。它们需要模型在阅读文章后回答额外问题。由于任务不同，这些现有的推理数据集并不能直接帮助指导训练聊天机器人。下图为对话和推理式阅读理解的常用数据集：

dataset	Task	Reasoning	Domain	Manually
Ubuntu (Lowe et al., 2015)	**Next Utterances Prediction**	✗	Technique	✗
PERSONA-CHAT (Zhang et al., 2018a)	**Next Utterances Prediction**	✗	Persona	✓
Dialogue NLI (Welleck et al., 2019)	**Next Utterances Prediction**	✗	Persona	✗
CoQA (Reddy et al., 2019)	Conversational QA	✓	Diverse	✓
Douban (Wu et al., 2017)	**Next Utterances Prediction**	✗	Open	✗
DREAM (Sun et al., 2019)	Reading Comprehension	✓	Open	✓
WSC (Levesque et al., 2012)	Coreference Resolution	✓	Open	✗
SWAG (Zellers et al., 2018)	Plausible Inference	✓	Movie	✗
CommonsenseQA (Talmor et al., 2019)	Reading Comprehension	✓	Open	✓
RACE (Lai et al., 2017)	Reading Comprehension	✓	Open	✗
ARC (Clark et al., 2018)	Reading Comprehension	✓	Science	✗
DROP (Dua et al., 2019)	Reading Comprehension	✓	Open	✗
Cosmos (Huang et al., 2019)	Reading Comprehension	✓	Narrative	✓
MuTual	**Next Utterances Prediction**	✓	Open	✓

基于目前对话数据集的缺陷，Mutual被提出，一个直接针对Response Selection的推理数据集。

数据集构建

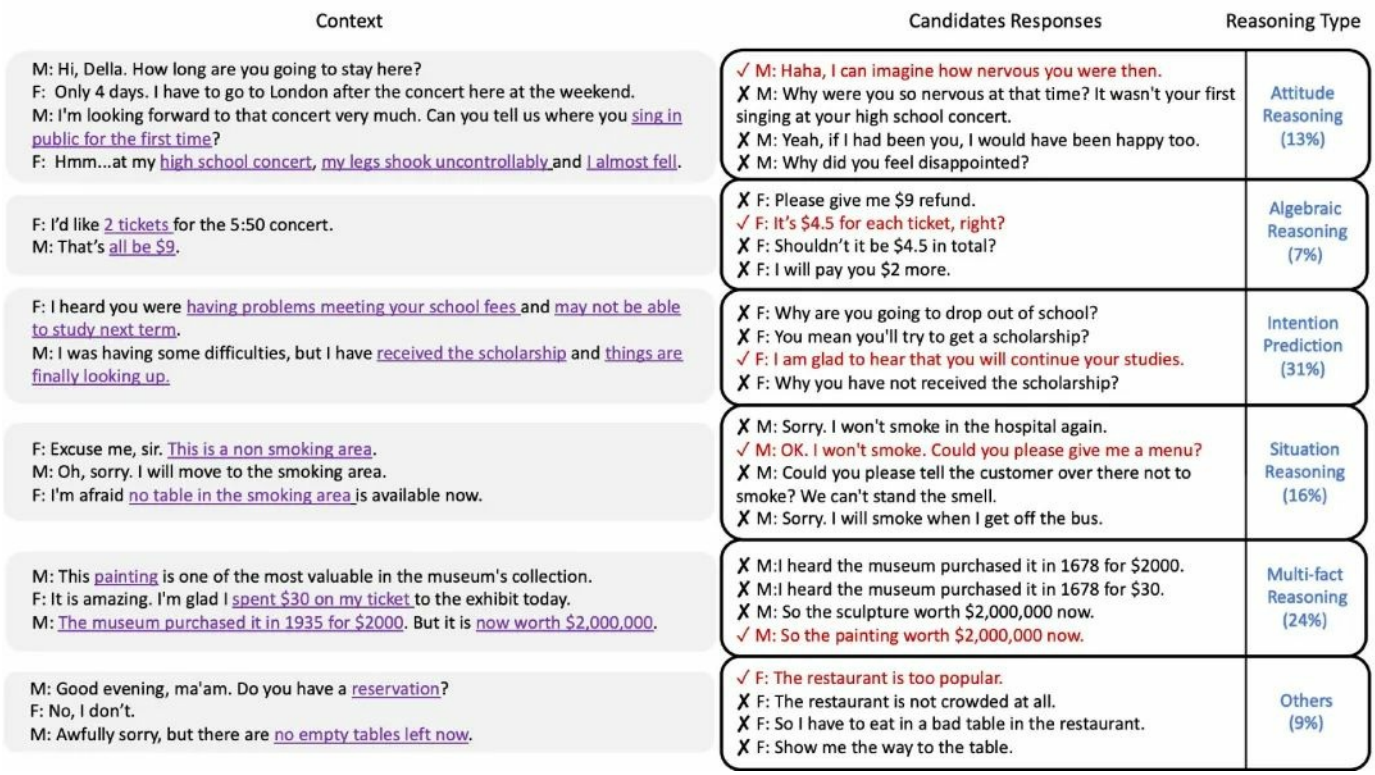
MuTual基于中国高考英语听力题改编。听力测试要求学生根据一段双人多轮对话，回答额外提出的问题。并通过学生能否正确答对问题衡量学生是否理解了对话内容。为了更自然的模拟开放领域对话，我们进一步将听力题中额外的问题转化为对话中的回复。



标注者截选原对话中具备回答问题信息的片段，根据正确选项构造正确的回复（上图回复A），根据两个错误选项构造两个错误的回复（回复C和回复D）。

为了进一步提升难度，引入额外的推理信息，标注者还需根据正确选项构建一个负面的回复（回复B）。另外，标注者需要保证在无上文信息情况下，所有候选回复在逻辑上皆合理。这样可以让数据集聚焦于检测模型在多轮对话中的推理能力，而非判断单个句子是否具有逻辑性。

作者还在标注过程中控制正确和错误的回复与上文的词汇重叠率相似，防止模型可以通过简单的根据文本匹配选出候选回复。构造出的数据集主要包含聊天机器人需要的六种推理能力：**态度推理(13%)，数值推理(7%)，意图预测(31%)，多事实推理(24%)和常识等其他推理类型（9%）**。



在真实应用场景中，检索式对话模型无法检索所有可能的回复，如果没有检索到合适的回复，**系统应具有给予安全回复(safe response)的能力**。为了模拟这一场景，MuTual^{plus}被提出。对于每个实例，MuTual^{plus}随机替换掉MuTual中一个候选回复。如果正确回复被替换，安全回复即为新的正确回复。如果错误回复被替换，原正确回复仍为四个回复中最合适的。

实验

论文测试了主流的检索式对话模型(LSTM, SMN, DAM)和预训练语言模型(GPT, BERT, RoBERTa)在MuTual和MuTual ^{plus}上的表现，以Recall@1(正确检索结果出现在检索结果第一位)， Recall@2(正确检索结果出现在检索结果前两位)， MRR(Mean Reciprocal Rank,正确检索结果在检索结果中的排名的倒数)作为评价指标。

		Dev			Test		
Baseline category	Baseline method	R@1	R@2	MRR	R@1	R@2	MRR
Baseline	Human	-	-	-	0.938	0.971	0.964
	Random	0.250	0.500	0.604	0.250	0.500	0.604
Individual scoring method (discrimination)	TF-IDF	0.276	0.541	0.541	0.279	0.536	0.542
	Dual LSTM (Lowe et al., 2015)	0.266	0.528	0.538	0.260	0.491	0.743
	SMN (Wu et al., 2017)	0.274	0.524	0.575	0.299	0.585	0.595
	DAM (Zhou et al., 2018)	0.239	0.463	0.575	0.241	0.465	0.518
	BERT (Devlin et al., 2019)	0.657	0.867	0.803	0.648	0.847	0.795
	RoBERTa (Liu et al., 2019)	**0.695**	0.878	0.824	**0.713**	**0.892**	**0.836**
Individual scoring method (generation)	GPT-2 (Radford et al., 2019)	0.335	0.595	0.586	0.332	0.602	0.584
	GPT-2-FT (Radford et al., 2019)	0.398	0.646	0.628	0.392	0.670	0.629
Multi-choice method	BERT-MC (Devlin et al., 2019)	0.661	0.871	0.806	0.667	0.878	0.810
	RoBERTa-MC (Liu et al., 2019)	0.693	**0.887**	**0.825**	0.686	0.887	0.822

从结果可以看到，之前的检索式对话模型在此种任务上，表现只比Random的情况好一点。不过预训练模型也不能取得很好的效果，甚至RoBERTa也只能达到71%的Recall@1。然而未经培训的非母语者可以轻松达到94%。

进一步研究发现，**模型表现不会随着对话轮数增加而变差**（推理能力并不依赖复杂的对话历史）。在推理类型方面，模型在数值推理和意图推测中表现的较差。下图第一个例子中，时差运算只需简单的减法(5:00pm - 6h = 11:00am)，第二个例子需要推理出对话出现在租房场景中，然而对现有的深度学习模型依然十分困难。

F: Do you know what time it is right now in New York?
M: Let me see. It's 5:00 pm now, in New York is 6 hours behind.

F: Let me see, 7 hours behind. It is 11:00 am now in New York.
F: 5 hours ahead. It is 11:00 pm now in New York.
✗ F: Is it 5:00 pm as well?
✓ F: It is 11:00 am now in New York.

F: Good morning. What can I do for you?
M: I am looking for a flat for 2 people near the university.
F: Well. There are several places available and the rent ranges from 80 to $150 a month. What are your requirements?
M: I think of flat for no more than $100 a month is good. I prefer to live in a quiet street and I need at least 2 bedrooms.

✗ F: If you have any questions about enrollment, do not hesitate to ask me.
✓ F: How about this flat? If you are satisfied, we can sign the contract tomorrow.
F: We have 2 floors in our supermarket.
F: You want only 1 bedroom, so we have three flats that meet your requirement.

总结

尽管BERT为代表的预训练模型很大程度上解决了检索式对话的回复相关性问题，但是依然难以解决真实对话场景中的常识和逻辑问题，导致聊天机器人的真实用户体验依然不尽人意。现有的检索式对话数据集大都没有直接对该问题进行建模，因此我们提出了MuTual数据集，针对性的评测模型在多轮对话中的推理能力。

论文地址：http://arxiv.org/abs/2004.04494

github地址：https://github.com/Nealcly/MuTual

arxiv访问慢的小伙伴也可以在订阅号后台回复关键词【0413】下载论文PDF。

可能喜欢

- ACL2020 | FastBERT：放飞BERT的推理速度
- 卖萌屋2020 Q1季度大会
- LayerNorm是Transformer的最优解吗?
- 如何优雅地编码文本中的位置信息？三种positioanl encoding方法简述
- 在大厂和小厂做算法有什么不同?

夕小瑶的卖萌屋

关注&星标小夕，带你解锁AI秘籍

订阅号主页下方「撩一下」有惊喜哦