SQuAD任务

史云迪　　0916

SQuAD任务

- SQuAD数据集包含107,785个关于人类生成的关于维基百科文章的阅读理解问题。每个问题都涉及一篇文章的一段，相应的答案保证在该段中。
- 介绍两篇论文：
- 《Adversarial Examples for Evaluating Reading Comprehension Systems》
- 《Improving the Robustness of Question Answering Systems to Question Paraphrasing》

Adversarial Examples for Evaluating Reading Comprehension Systems

Robin Jia Percy Liang

EMNLP 2017

主要内容

- 动机：
  - 阅读理解系统正在取得快速的进展，但这些系统在多大程度上真正理解语言仍不清楚。为了评估系统的真实语言理解能力，我们提出了一种针对斯坦福问题回答数据集(SQuAD)的对抗评估方案。
- 贡献：
  - 通过在输入段落中添加分散注意力的句子来创建对抗样本
  - 使用一组简单的规则来生成一个原始的干扰句子，它不回答问题，但看起来很相关；然后，我们通过众包来修复语法错误。
  - 我们的实验表明，没有一个已发表的开源模型对添加对抗性句子具有鲁棒性。

SQuAD

- 任务：
  - SQuAD数据集包含107,785个关于人类生成的关于维基百科文章的阅读理解问题。每个问题都涉及一篇文章的一段，相应的答案保证在该段中。
- 模型：
  - 在开发和测试我们的方法时，我们重点关注了两种已发表的模型架构：BiDA和Match-LSTM。
  - 两者都是深度学习架构，可以预测正确答案上的概率分布。每个模型都有一个单一的版本和一个集成版本，总共产生四个系统。
  - 我们还验证了其他12个公开的模型，我们在开发过程中没有运行这些模型，因此它们可以用来验证本文方法的通用性。
- 任务评价精度：
  - 输入：(p,q)——段落-问题　　f:根据输入可给出输出 $\hat{a}$
  - v为真实答案a和预测答案f(p，q)之间的f1分数

$$\mathrm{Acc}(f) \stackrel{\mathrm{def}}{=} \frac{1}{|D_{\mathrm{test}}|} \sum_{(p,q,a)\in D_{\mathrm{test}}} v((p,q,a),f),$$

对抗性评估

- 我们将对手A定义为一个函数，它输入一个样本(p，q，a)，可以选择使用模型f，并返回一个新的示例(p′，q′，a′)。对A的对抗性准确性是:

$$\mathrm{Adv}(f) \stackrel{\mathrm{def}}{=} \frac{1}{|D_{\mathrm{test}}|} \sum_{(p,q,a)\in D_{\mathrm{test}}} v(A(p,q,a,f),f)).$$

- 对手必须满足两个基本要求:
  - 首先，它应该总是生成有效的(p′，q′，a′)元组——a′是给定（p′，q′）的的正确答案。
  - 第二，（p′，q′，a′）应该以某种程度“接近”原始样本（p，q，a）。

Concatenative Adversaries

- 在图像分类中，对抗样本通常是通过在输入中添加难以察觉的噪声来产生的。这些扰动不会改变图像的语义，但它们可以改变对语义保留变化过于敏感的模型的预测。对于语言，直接的模拟将是改写输入。然而，高精度的转译生成是具有挑战性的，因为大多数对句子的编辑实际上确实改变了它的意义。
- 我们使用确实改变了语义的扰动来构建连接对手，从而为某些句子s生成(p + s, q, a) 形式的样本。换句话说，构建的连接对手在段落的末尾添加一个新的句子，并保持问题和答案不变。
- 我们描述两个具体的连接对手(ADDSENT 和 ADDANY)，以及两个变体。对手补充添加了看起来与问题相似的语法句子。

ADDSENT

- ADDSENT使用四个步骤来生成看起来类似于问题的句子，但实际上并不与正确答案相矛盾的句子。
- （1）我们对**问题**应用**语义改变扰动**，我们用WordNet的反义词替换名词和形容词，并用相同的语音将命名实体和数字改为GloVe单词向量空间中最近的单词。如果在此步骤中没有更改单词，对手将放弃并立即返回原始样本。

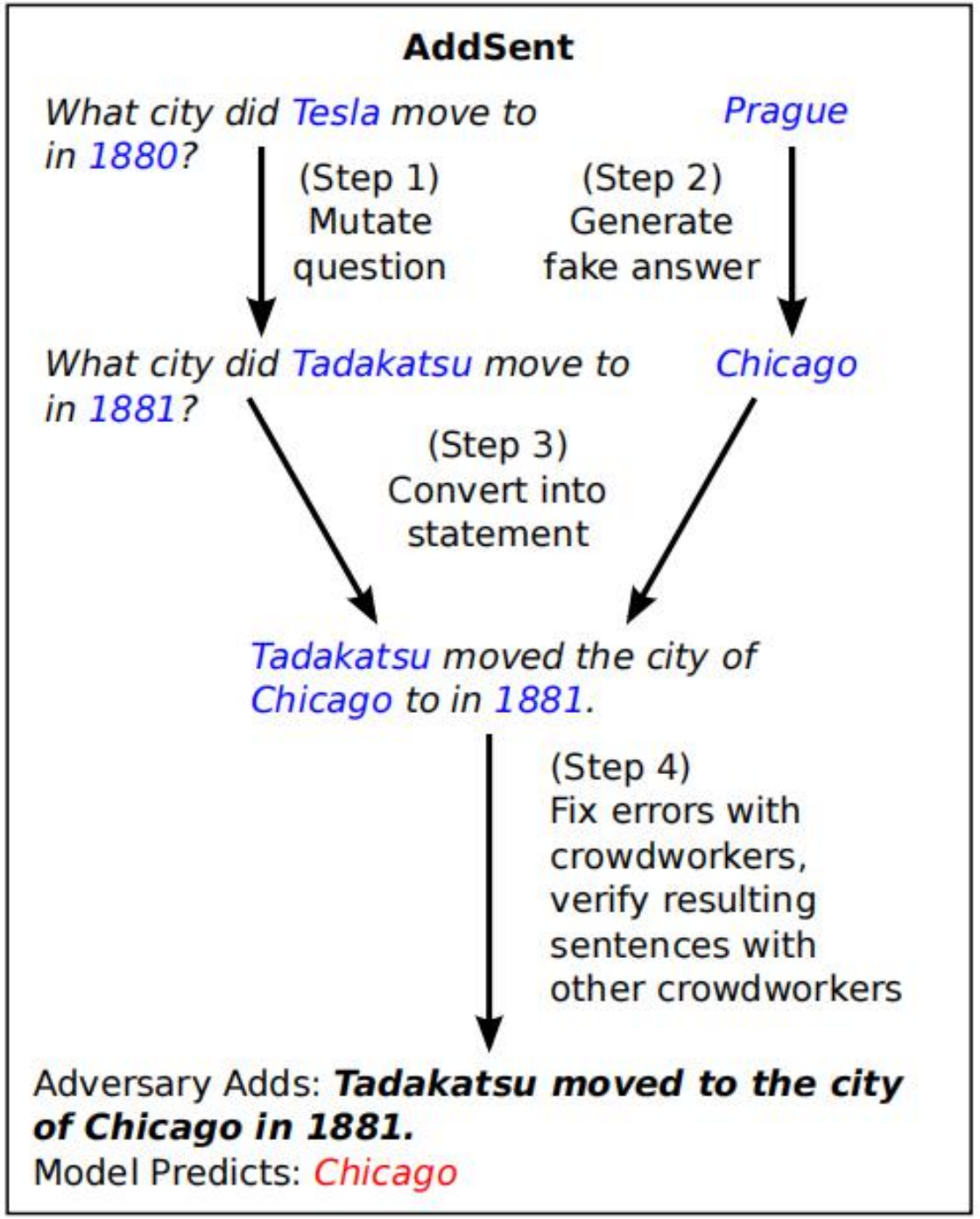
- 例子:“ABC哪个部门处理国内电视发行？”
  - 把“ABC”改为“NBC”（矢量空间中的邻近词）
  - “国内”改为“国外”(WordNet反义词)
  - 最终问题：NBC哪个部门处理国外电视发行?

ADDSENT

- （2）我们创建了一个与原始答案具有相同的“类型”的假答案。 给定一个问题的原始答案，我们计算它的类型并返回相应的假答案。
- （3）我们将修改后的问题和假答案组合，使用一组大约有50条手动定义的CoreNLP选区解析规则。由于我们的规则的不完整和选区解析中的错误，由步骤3生成的原始句子可能是不语法的或其他不自然的。
- （4）在步骤4中，我们通过众包来修复这些句子中的错误。
  - 每个句子都由 Amazon Mechanical Turk 上的五名工作人员独立编辑，每个原始句子最多可编辑五个句子。 三个额外的众包工人然后过滤掉不符合语法或不兼容的句子，从而产生一个较小的（可能是空的）人工批准的句子集。
  - 完整的 ADDSENT 对手将模型 f 作为一个黑盒模型运行每个人类认可的句子，并选择使模型给出最差答案的那个。 如果没有人类认可的句子，对手只需返回原始样本。

ADDSENT例子

Article: **Nikola Tesla**
Paragraph: "*In January 1880, two of Tesla's uncles put together enough money to help him leave Gospić for Prague where he was to study. Unfortunately, he arrived too late to enroll at Charles-Ferdinand University; he never studied Greek, a required subject; and he was illiterate in Czech, another required subject. Tesla did, however, attend lectures at the university, although, as an auditor, he did not receive grades for the courses.*"
Question: "*What city did Tesla move to in 1880?*"
Answer: *Prague*
Model Predicts: *Prague*



ADDONESENT

- ADDSENT需要对正在评估的模型进行少量的查询。
- 为了生成一个完全独立于模型的对手，我们还引入了ADDONESENT，它在段落中添加了一个随机的人类批准的句子。ADDONESENT不需要访问任何模型或任何训练数据。

ADDANY

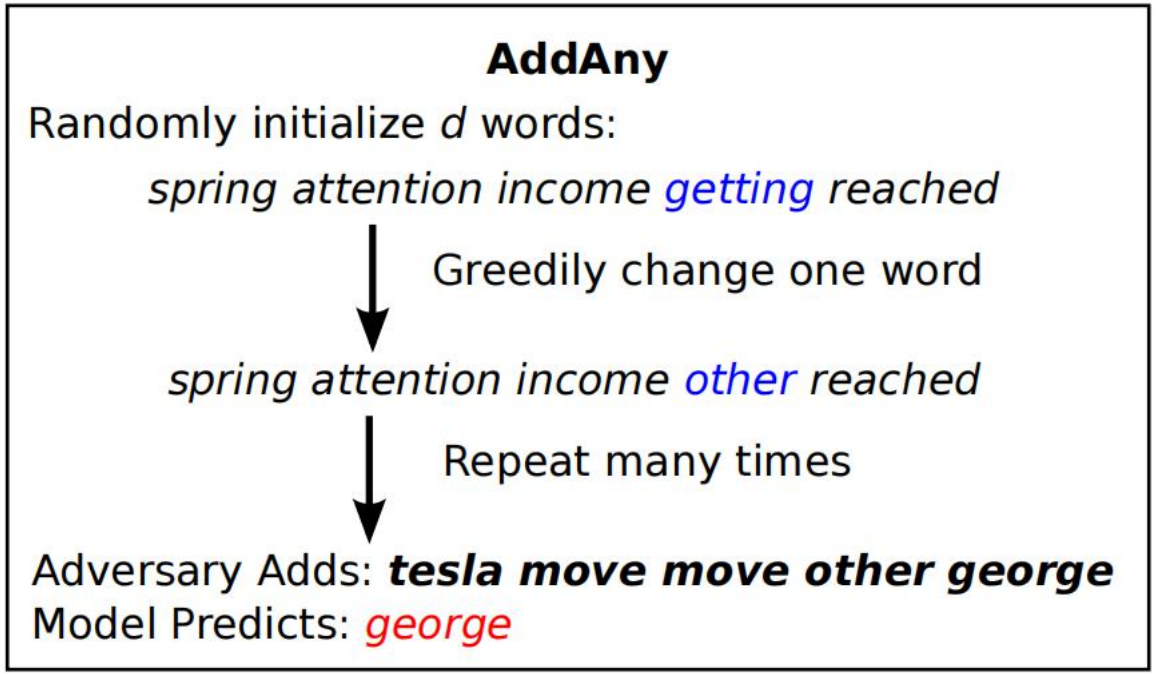
- 对于ADDANY，其目标是选择任何d个单词的序列，而不考虑语法性。我们使用局部搜索来反向选择一个分散注意力的句子的 $s = w_1 w_2 ... w_d$。
- 我们首先初始化单词 $w_1, . . . , w_d$ 从常用英语单词列表中随机抽取。
- 然后，我们运行6轮迭代，每个迭代都以随机顺序排遍历索引 $i \in \{1, . . , d\}$ 。
  - 对于每个 i，我们随机生成一组候选词 W 为在常用词和 q 中所有词的集合中随机采样20个。
  - 对于每个$x \in W$，$j \neq i$，我们生成x在第i个位置和 w_j在第j个位置的句子。
  - 尝试将每个句子添加到段落中，并查询模型的预测概率分布，将w_i更新为使模型输出分布上F1分数的期望值最小的x。
  - 如果模型预测的 F1 分数为 0，我们会立即返回。 如果我们在 3 轮迭代后没有停止，我们随机初始化 4 个额外的词序列，并并行搜索所有这些随机初始化。

ADDANY

- ADDANY需要比ADDSENT更多的模型访问：它不仅在搜索过程中多次查询模型，而且还假设模型返回一个基于答案的概率分布，而不仅仅是一个单一的预测。

- ADDCOMMON
  - 引入了一个叫做ADDCOMMON的ADDANY变体，它和ADDANY完全一样，只是只添加了常见的单词。
  - 我们注意到ADDANY都试图将问题中的单词纳入他们的对抗性句子中。

  虽然这是吸引模型注意的一种明显方法，但我们很好奇，如果没有这样一种直接的方法，我们是否也能分散模型的注意力。



实　验

- 所有的实验，测量了来自SQuAD开发集的1000个随机样本（测试集不公开）的对抗性f1评分
- 表2显示了MatchLSTM和BiDAF模型对所有四个对手的性能。

	Match Single	Match Ens.	BiDAF Single	BiDAF Ens.
Original	71.4	75.4	75.5	80.0
ADDSENT	27.3	29.4	34.3	34.2
ADDONESENT	39.0	41.8	45.7	46.9
ADDANY	7.6	11.7	4.8	2.7
ADDCOMMON	38.9	51.0	41.7	52.6

Table 2: Adversarial evaluation on the Match-LSTM and BiDAF systems. All four systems can be fooled by adversarial examples.

实　验

- 攻击普遍性验证
- 为了验证我们的对手足够普遍

  可以欺骗在开发过程中没有使用的模型。
- 我们在12个已发布的模型上运行了ADDSENT；

  我们没有在这些模型上运行ADDANY，

  因为并不是所有的模型都暴露了输出分布。

Model	Original	ADDSENT	ADDONESENT
ReasoNet-E	**81.1**	39.4	49.8
SEDT-E	80.1	35.0	46.5
BiDAF-E	80.0	34.2	46.9
Mnemonic-E	79.1	**46.2**	**55.3**
Ruminating	78.8	37.4	47.7
jNet	78.6	37.9	47.0
Mnemonic-S	78.5	**46.6**	**56.0**
ReasoNet-S	78.2	39.4	50.3
MPCM-S	77.0	40.3	50.0
SEDT-S	76.9	33.9	44.8
RaSOR	76.2	39.5	49.5
BiDAF-S	75.5	34.3	45.7
Match-E	75.4	29.4	41.8
Match-S	71.4	27.3	39.0
DCR	69.3	37.8	45.1
Logistic	50.4	23.2	30.4

Table 3: ADDSENT and ADDONESENT on all sixteen models, sorted by F1 score the original examples. S = single, E = ensemble.

实　验

- 人类验证
- 为了确保我们的结果是有效的，我们验证了人类也不会被我们生成的对抗样本欺骗。
- 由于ADDANY需要太多的模型查询来与人类进行交互，我们关注的是ADDSENT。我们将每个原始问题和对抗样本对呈现给三个众包工作者，并要求他们通过从段落中复制粘贴来选择正确的答案。然后，我们对这三种回答进行了多数投票（如果所有三种回答都不同，我们随机选择一个）。
- 人类的精度下降了13.1，远远低于计算机系统。

	Human
Original	92.6
ADDSENT	79.5
ADDONESENT	89.2

Table 4: Human evaulation on adversarial examples. Human accuracy drops on ADDSENT mostly due to unrelated errors; the ADDONESENT numbers show that humans are robust to adversarial sentences.

实　验

- 跨模型的可迁移性
- ADDSENT 对抗样本在模型之间非常有效地转移
- ADDANY 对抗样本在模型之间表现出更有限的可转移性。
- 对抗样本在同一模型的单个版本和集成版本之间的传输性稍好一些。

Targeted Model	Model under Evaluation			
	ML Single	ML Ens.	BiDAF Single	BiDAF Ens.
ADDSENT				
ML Single	27.3	33.4	40.3	39.1
ML Ens.	31.6	29.4	40.2	38.7
BiDAF Single	32.7	34.8	34.3	37.4
BiDAF Ens.	32.7	34.2	38.3	34.2
ADDANY				
ML Single	7.6	54.1	57.1	60.9
ML Ens.	44.9	11.7	50.4	54.8
BiDAF Single	58.4	60.5	4.8	46.4
BiDAF Ens.	48.8	51.1	25.0	2.7

Table 5: Transferability of adversarial examples across models. Each row measures performance on adversarial examples generated to target one particular model; each column evaluates one (possibly different) model on these examples.

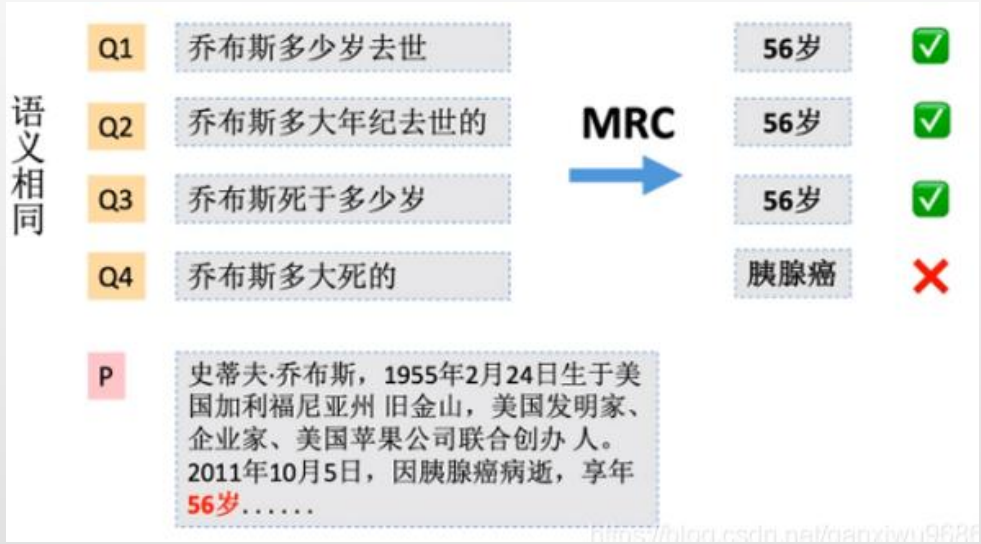
Improving the Robustness of Question Answering Systems to Question Paraphrasing

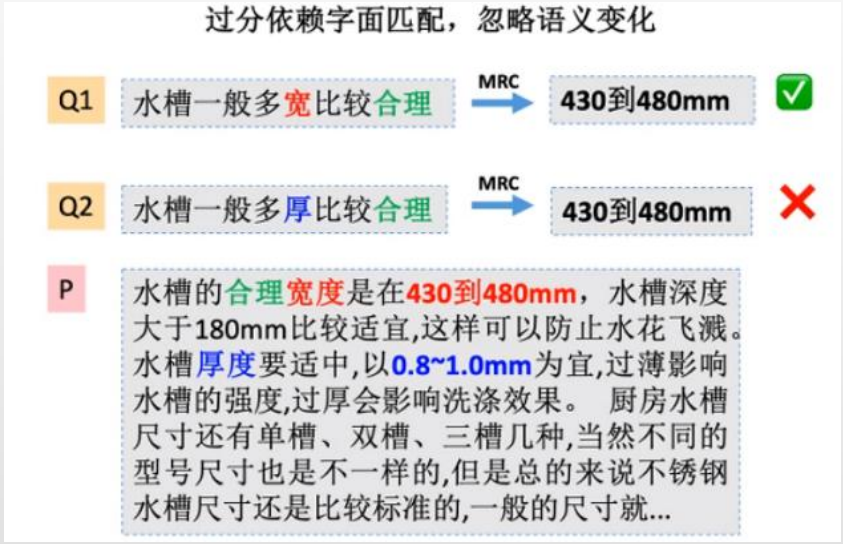
Wee Chung Gan Hwee Tou Ng

ACL 2019

主要内容

- 动机：
  - 提高模型的鲁棒性
  - 动机源于这样的观察，即当一个问题以一种稍微不同但语义相似的方式措辞时，QA模型能够输出一个错误的预测，尽管它能够正确地回答最初的问题。





主要内容

- SQuAD任务，以问题和上下文来预测正确答案。使用了两个评估指标：精确匹配(EM)和F1。
  - EM是精确匹配结果，也就是模型给出的答案与标准答案一模一样。
  - F1是模糊匹配，可以理解为机器答对了部分内容，是根据模型给出的答案和标准答案之间的重合度计算出来的
- 主要贡献
  - 介绍了一种新的方法，通过训练模型来产生不同的释义问题。
  - 我们发布了两个测试集评价QA模型对问题解释的鲁棒性。
  - 利用测试集在三个最先进的QA模型上进行实验，表现都不是很好
    - 非对抗性的测试集由1062个问题组成，但对最初的问题有轻微的干扰。
    - 对抗性的测试集由56个问题组成，使用接近混淆答案的上下文单词进行释义。
  - 我们表明，可以使用全自动方法来增强训练集并在增强的训练集上重新训练模型，从而提高 QA 模型对两个释义测试集的鲁棒性。

Paraphrase-Guided Paraphrasing Network

- 训练神经网络
  - 该网络能够将一个源问题和一个释义建议（一个单词或短语）作为输入，以生成一个释义问题。

  source question + paraphrase suggestion(手动构造的) -> target question（输出）
  - 为了做到这一点，需要一个训练数据集，其中每个训练样本都是有以下形式的（source question、paraphrase suggestion、target question）
  - paraphrase suggestion一定要是target question的一部分
- 模型结构
  - transformer模型

数据集

- 结合使用维基答案释义语料库(WikiAnswers paraphrase corpus)和Quora问题对（Quora Question Pairs）数据集进行训练。
  - Quora数据集中的一个问题对中的两个问题在意义上通常非常相似。
  - 维基答案释义语料库包含超过2200万对问题对，存在一个源问题与多个目标问题配对。
- 获取原问题和目标问题（source question和target question）
  - 维基答案释义语料库：每个问题至少有7个tokens才会保留一个问题对；使用Wieting和Gimpel提出的预先训练的模型去除所有释义相似度低于0.7的问题对来过滤掉错误的问题对；随机抽取源问题进行抽样，获得约35万个问题对。（只使用这个数据集的一小部分，以免淹没Quora数据集）
  - Quora问题对：使用一对问题作为两个训练样本，在训练集中包括源问题作为目标问题，反之亦然，即，我们包括问题A→问题B和问题B→问题A在训练集中。总共大约有28万个训练样本来自Quora数据集。

数据集

- 获得释义建议（paraphrase suggestion）
  - 维基答案数据集：对于每个源和目标问题对，我们使用数据集附带的单词对齐来匹配来自源和目标问题的单词和短语，以获得短语对齐对。对齐对被过滤，以保留出现在目标问题中但不在源问题中的短语。

Question	Word Alignments	Phrase Alignments	Candidate Suggestions
Source	what nutrients do green peppers have in them ?	(what, what) (green, a green) (have in them, contain) ...	a green, pepper, contain
Target	what nutrients does a green pepper contain ?		

Figure 2: An example of finding possible paraphrase suggestions for a source and target question pair from the WikiAnswers dataset. Since there can be multiple target questions for a given source question, we ensure that there are no duplicates in the suggestions chosen for the same source question.

	Question	Keywords	Candidate Suggestions	Selected Suggestion
Source	how can i find out how many devices are connected to my wifi?	wifi, connected, many devices, devices, find	wifi network, network, know	wifi network
Target	how can i know how many devices are connected to my wifi network?	wifi network, network, wifi, connected, many devices, devices, know		

Figure 3: An example of obtaining a paraphrase suggestion for a source and target question pair from the Quora dataset. Keywords from the questions are obtained from TextRank.

  - Quora数据集：由于Quora数据集没有单词对齐，首先使用文本排名(Mihalcea和Tarau)从源问题和目标问题中获得问题关键字；释义建议是目标问题中排名最高的关键短语，而不在源问题中；我们不允许选择停用词（stopwords）作为释义建议。

Paraphrasing SQuAD Questions

- 使用SQuAD开发问题创建两个释义测试集，以评估QA模型的稳健性。
  - 非对抗性释义测试集
  - 对抗性释义测试集

- 非对抗性释义测试集
  - 我们使用第2节中训练有素的释义模型来创建一个非对抗性的释义测试集。我们使用人工注释者来确保这个测试集的问题的质量，这也可以作为我们的释义模型的评估。

- 对抗性释义测试集
  - 我们利用模型的弱点创建了一个测试集。可以通过在同一类型的错误答案候选附近的上下文中使用的单词来对问题进行释义，以生成对抗样本。

非对抗性释义的过程

- PPDB（Pavlick 等人，2015 年），这是一个自动提取的数据库，由数百万个释义短语对组成。释义对可以包含一个词或多个词。PPDB有6种不同的尺寸，较大的尺寸具有更大的覆盖范围，但精度较低。
- 首先，我们从源问题中获得所有的n-grams（最多6-grams），并删除停用词。接下来，我们通过搜索PPDB(XL size)来搜索等价分数超过0.25的其余n-grams的释义。这为模型提供了释义建议（paraphrase suggestion）产生释义问题。
- 在释义生成后，我们删除语义上不同的释义。与从维基答案语料库中过滤问题对类似，我们使用Wieting和Gimpel（2018）的预训练模型来获得生成问题的释义相似度得分，并只保持得分在0.95以上的问题。

释义过程（Paraphrasing Process）

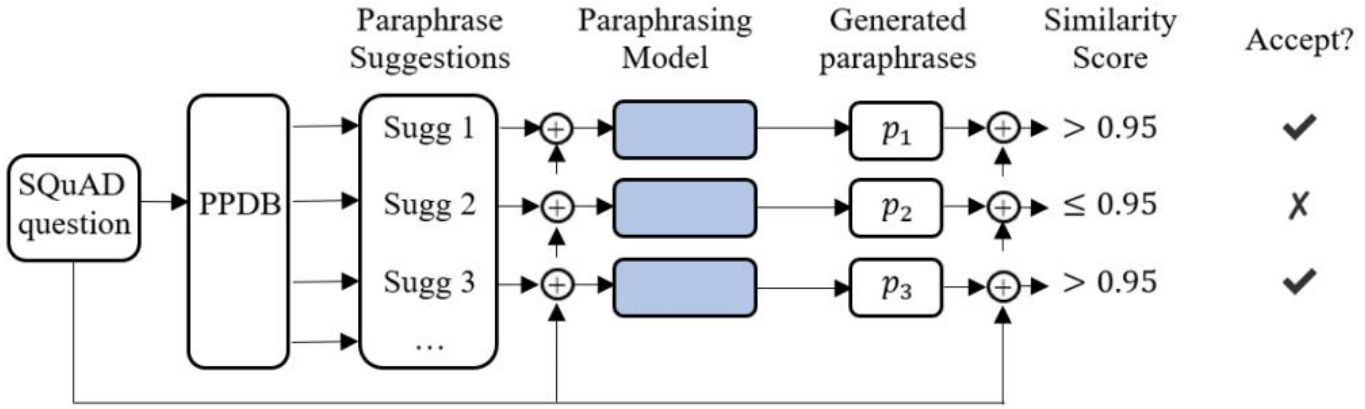


Figure 4: Process to paraphrase SQuAD questions. We first use PPDB to obtain paraphrase suggestions before passing both the original question and the suggestions to our paraphrasing model to generate paraphrases. A generated paraphrase is accepted if its similarity score with the original question is above 0.95. ⊕ refers to the use of the original SQuAD question and the previous output as inputs to the next step.

Original Question
the european court of justice cannot uphold measures that are incompatible with what?
Paraphrased Questions
1. the european court of justice cannot uphold a number of measures that are incompatible with what? 2. the european court of justice cannot uphold measures that are inconsistent with what? 3. the european court of justice cannot uphold measures which are not compatible with what? 4. the european court of justice has not been able to uphold measures that are incompatible with what?

Figure 5: Examples of generated paraphrases.

人类评估

- 为了评估自动生成的释义问题的质量，使用了来自Amazon Mechanical Turk (AMT)的人类注释者来评估释义问题的语义等效性和流畅性。
- 从SQuAD开发集中改写问题，并随机选择3000个生成的释义问题。对于每一对问题，我们要求来自AMT的2个注释者来说明他们对以下两个陈述的同意程度，范围从1到5个(强烈不同意、不同意、中立、同意或强烈同意）：
- （1）转述的问题与最初的问题具有相同的意义（即，转述的问题和原来的问题都有望得到相同的答案）。
- （2）释义的问题是用流利的英语写的。

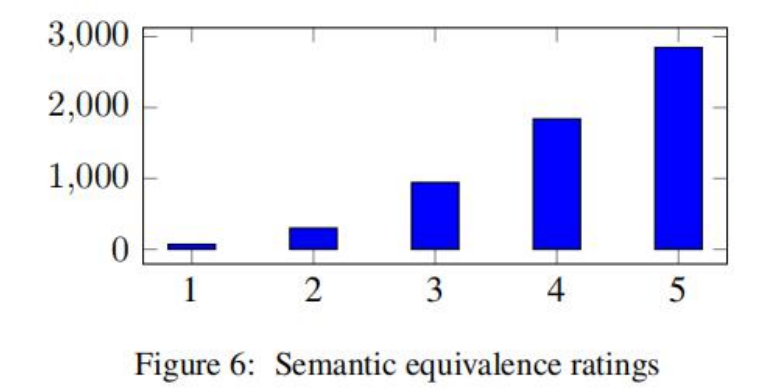


Figure 6: Semantic equivalence ratings

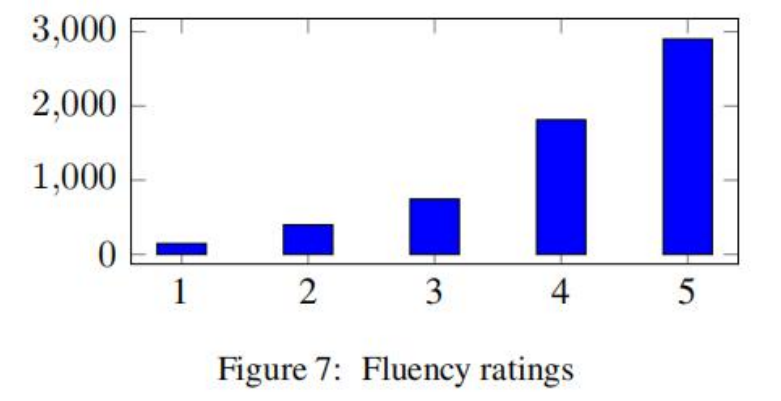


Figure 7: Fluency ratings

非对抗性测试集的创立

- 只有如果注释者都同意或强烈同意转述问题和原始问题在语义上是相等的，我们才会将生成的转述问题包含到测试集中。
- 如果从一个原始问题中有多个被接受的释义的问题，我们只随机选择一个释义的问题包含在测试集中。
- 总共产生了1062个释义问题。

对抗性测试集的建立

- 由于观察到在SQuAD上训练的QA模型倾向于执行字符串匹配，以便在上下文和问题之间的重要单词重叠区域附近返回适当类型的答案(Jia和梁，2017；Rondauo和Hazen，2018)，利用模型的这一弱点创建了一个测试集。
- 我们通过SQuAD开发集的问题和上下文对来手动执行这样的转述，如果存在这样一个候选词，并且附近有合适的上下文词，则使用候选答案附近的上下文词来重述问题。我们为对抗测试集创建了56个释义问题。

由于正确的答案“2009”是一年，我们在上下文中找到“1963”，并使用附近的上下文“电视转播”来重述原来的问题。

Context: 826 Doctor Who instalments have been televised since 1963 ... Starting with the 2009 special “Planet of the Dead”, the series was filmed in 1080i for HDTV ...
Original Question: In what year did Doctor Who begin being shown in HDTV? **Prediction:** 2009
Paraphrased Question: Since what year has Doctor Who been televised in HDTV? **Prediction:** 1963

Figure 8: An example of paraphrasing question using context words (underlined) near a confusing answer candidate to generate a natural adversarial example.

实 验

- 我们在三种最先进的QA模型上进行了实验：BERT、DrQA和BiDAF。
- 评估两个测试集的性能：
- 尽管释义测试集在语义上相似，并且没有执行模型查询来有意定位QA模型的弱点，但所有三种模型的性能都显著下降。这突出了训练过的模型的脆弱性。
- 对抗性释义测试集能够利用QA模型对字符串匹配的依赖，导致模型性能的急剧下降。

Model	EM Score		F1 Score	
	Orig Q	Para Q	Orig Q	Para Q
BERT	83.62	79.85	90.78	87.63
DrQA	67.33	65.25	76.25	74.25
BiDAF	67.80	63.84	76.85	73.51

Table 1: Performance of QA models on the original questions (Orig Q) compared to non-adversarial paraphrased questions (Para Q).

Model	EM Score		F1 Score	
	Orig Q	Adv Q	Orig Q	Adv Q
BERT	82.14	57.14	89.31	63.18
DrQA	71.43	39.29	81.02	48.94
BiDAF	75.00	30.36	81.55	38.30

Table 2: Performance of QA models on the original questions (Orig Q) compared to adversarial paraphrased questions (Adv Q).

实　验——利用训练数据进行再训练

- 非对抗性释义测试集
- 我们的评估表明，原始的训练数据集不包含足够多样化的问题措辞。这导致模型无法学习正确地回答提出相同问题的各种方式。因此，提高QA模型的稳健性的一种自然方法是将它们暴露在更多样化的问题措辞中。我们试图通过使用我们的释义模型来释义问题的训练集来实现这一点。
- 我们随机抽取25,000个转述问题作为额外的训练数据。我们使用原始训练数据和额外的25,000个转述问题对所有三个QA模型进行再训练，保持释义问题相似度得分在0.9分以上

Model	EM Score		F1 Score	
	Before	After	Before	After
BERT	79.85	80.89	87.63	88.62
DrQA	65.25	67.33	74.25	75.00
BiDAF	63.84	66.20	73.51	75.94

Table 3: Performance on the non-adversarial paraphrased test set before and after re-training.

Model	EM Score		F1 Score	
	Before	After	Before	After
BERT	84.02	83.76	91.00	90.88
DrQA	69.04	68.74	78.38	77.86
BiDAF	67.67	67.49	77.46	77.10

Table 4: Performance on the original development set before and after re-training.

实　验

- 对抗性释义测试集
- 不使用PPDB获得释义建议，使用在Ontonotes数据集上训练的Flair标记答案所属的命名实体类，从包含相同类型命名实体的上下文中提取句子。使用在CoNLL-2000数据集上训练的Flair对提取的句子进行句法分块(SangandBuchholz，2000)。我们使用分块结果中的名词和动词短语来形成对给定问题的释义建议集。我们确保得到的每个建议都包含至少两个单词，并且不与答案重叠。
- 保留释义相似度得分在0.83以上的问题，用另外25,000个生成的样本来重新训练所有三个QA模型。

Model	EM Score		F1 Score	
	Before	After	Before	After
BERT	57.14	69.64	63.18	73.85
DrQA	39.29	41.07	48.94	49.86
BiDAF	30.36	39.29	38.30	47.49

Table 5: Performance of QA models on the adversarial test set before and after re-training.

Model	EM Score		F1 Score	
	Before	After	Before	After
BERT	84.02	83.33	91.00	90.49
DrQA	69.04	67.93	78.38	77.45
BiDAF	67.67	66.23	77.46	76.19

Table 6: Performance on the original development set before and after re-training.