Benchmarking Knowledge-Enhanced Commonsense Question Answering via Knowledge-to-Text Transformation

边宁, 韩先培, 陈波, 孙乐

报告人: 边宁
时间: 2020.12.19
中国科学院软件研究所 中文信息处理实验室
目录

• 常识问答背景
• 基于知识到文本转换的常识问答模型
• 常识问答模型探针实验
• 总结
常识问答背景

常识问答：回答依赖常识知识的问题，以测试语言理解能力。

Where on a river can you hold a cup upright to catch water on a sunny day?
✔ waterfall, ✘ bridge, ✘ valley, ✘ pebble, ✘ mountain

常识问答依赖常识知识：
(waterfall, IsA, vertical flow of moving water)
(upright, synonym, vertical)

常识问答的挑战：
① 常识知识库与问题、答案之间存在异构性。
② 知识具有上下文相关性。
常识问答背景

- 现有工作通常利用知识增强常识问答模型中的一个模块。
  - 注意力机制、基于图卷积的推理机制等。
- 为研究知识以及知识融入模型的方式对于常识问答的影响，本文通过探针实验，提出并回答了三个重要的问题:

Q1：在常识问答任务中，利用外部的常识知识具有多少潜能？
Q2：目前的常识问答模型利用常识知识的程度如何？
  - 例如，基于图的常识问答模型能否充分利用外部知识库提供的常识知识？
Q3：常识问答领域有哪些具有潜力的未来研究方向？
目录

• 常识问答背景
• 基于知识到文本转换的常识问答模型
• 常识问答模型探针实验
• 总结
基于知识到文本转换的常识问答模型

知识库 问题和候选答案

知识检索

知识

知识到文本转换

知识描述

阅读理解

答案

上下文相关性

异构性

What is a person looking for when completing puzzles or riddles?
A. avoid pain | B. intellectual challenge | C. compliments | D. attention | E. passing grade

Person desires intellectual challenge.
Puzzle is a problem. Riddle is a problem.
Problem is the same as challenge. Puzzle is used for challenge.

B. intellectual challenge
基于知识到文本转换的常识问答模型

1、从常识知识库中检索知识

给定一个问题和一个候选答案:

① 与常识知识图谱中的概念进行匹配，提取概念。

② 寻找知识图谱中连接问题概念和答案概念的知识路径。

“puzzle→IsA→problem→Synonym→challenge”是一个与候选答案“Intellectual challenge”相关的2跳知识路径。
基于知识到文本转换的常识问答模型

2、知识到文本转换:
- 基于模板的转换算法
  ➢ 关系模板。
- 基于复述的转换算法
  ➢ 使用复述模型生成更加流畅和多样的知识描述。
- 基于检索的转换算法
  ➢ 利用真实的语料库中的文本作为知识描述。

| 知识路径 | Puzzle → IsA → Problem → Synonym → Challenge |
|-----------|-----------------------------------------------|
| 模板转换 | Puzzle is a problem. Problem is the same as challenge. |
| 复述转换 | Puzzles are problems. The problem is the same as the challenge. |
| 检索转换 | Puzzle problem is a challenge game for children. |
基于知识到文本转换的常识问答模型

3、阅读理解模型预测答案

上下文相关的文本表示

前向神经网络

候选答案分数

预训练语言模型

输入

输出

知识库 问题和候选答案

知识检索

知识

知识到文本转换

知识描述

阅读理解

答案

What is a person looking for when completing puzzles or riddles?
A. avoid pain | B. intellectual challenge | C. compliments | D. attention | E. passing grade

Person desires intellectual challenge. Puzzle is a problem. Riddle is a problem. Problem is the same as challenge. Puzzle is used for challenge.
目录

• 常识问答背景
• 基于知识到文本转换的常识问答模型
• 常识问答模型探针实验
• 总结
常识问答模型探针实验

实验设置

- 数据集:
  - 主要数据集: CommonsenseQA;
  - 辅助数据集: Winograd Schema Challenge (WSC)、HellaSWAG、SOCIAL IQa。
- 常识知识库: ConceptNet。
- 基线模型:
  - Ma et al. (2019) —— 注意力机制 + 选项比较网络 (Option Comparison Network)。
  - Lv et al. (2020) —— 图神经网络。
  - KEDGN —— 图神经网络。
知识到文本转换的效果

- 知识到文本转换对知识增强的常识问答是有效的。
- 知识到文本转换能鲁棒地利用知识。
- 三种知识到文本转换算法能够互补。

知识到文本转换是有效且鲁棒的。

| 模型       | 知识源       | BERT   | XLNet  | RoBERTa | ALBERT |
|------------|--------------|--------|--------|---------|--------|
| 人类       | --           | 88.9   | 88.9   | 88.9    | 88.9   |
| 标准知识解释 | 人工标注     | 81.1   | 85.1   | 84.7    | 83.7   |
| 知识到文本转换 | 模板转换算法 | ConceptNet | 67.9  | 77.5   | 78.1   | 81.1   |
|             | 复述转换算法 | ConceptNet | 67.2  | 74.9   | 77.8   | 79.3   |
|             | 检索转换算法 | ConceptNet | 65.0  | 75.0   | 77.1   | 79.4   |
|             | 完整模型     | ConceptNet | 70.4  | 80.3   | 80.8   | 83.3   |

不同预训练语言模型上的最佳性能

| 模型       | 基础模型 | CommonsenseQA数据集上的答案准确率 |
|------------|----------|----------------------------------|
| BERT       | 无知识   | WSC    | HellaSWAG | SOCIAL IQa |
| + 知识     | 66.0     | 42.3   | 66.2      |
|           | 68.1     | 44.2   | 68.8      |
| RoBERTa    | 81.4     | 82.5   | 74.3      |
| + 知识     | 82.5     | 83.0   | 75.0      |
| ALBERT     | 84.9     | 86.1   | 77.2      |
| + 知识     | 87.0     | 86.9   | 77.8      |
| 人类       | 92.1     | 94.5   | 86.9      |

知识路径长度对CommonsenseQA上答案准确率的影响。使用基于模板的知识到文本转换算法。
知识到文本转换的效果

- 知识到文本转换对知识增强的常识问答是有效的。
- 知识到文本转换能鲁棒地利用知识。
- 三种知识到文本转换算法能够互补。

### 知识到文本转换是有效且鲁棒的。

| 模型             | 知识源 | BERT | XLNet | RoBERTa | ALBERT |
|------------------|--------|------|-------|---------|--------|
| 人类             | --     | 88.9 | 88.9  | 88.9    | 88.9   |
| 标准知识解释     | 人工标注 | 81.1 | 85.1  | 84.7    | 83.7   |
| 知识到文本转换   |        |      |       |         |        |
| 模板转换算法     | ConceptNet | 67.9 | 77.5  | 78.1    | 81.1   |
| 复述转换算法     | ConceptNet | 67.2 | 74.9  | 77.8    | 79.3   |
| 检索转换算法     | ConceptNet | 65.0 | 75.0  | 77.1    | 79.4   |
| 完整模型         | ConceptNet | 70.4 | 80.3  | 80.8    | 83.3   |

不同预训练模型在模型上的最佳性能

| 模型             | 知识到文本转换 | CommonsenseQA数据集上的答案准确率 |
|------------------|----------------|----------------------------------|
| 基础模型         | 无知识         | WSC: 63.6  HellaSWAG: 68.9  SOCIAL IQa: 76.2 |
|                 |                |                                |

知识路径长度对CommonsenseQA上答案准确率的影响。使用基于模板的知识到文本转换算法和2跳知识路径。
知识到文本转换的效果

- 知识到文本转换对知识增强的常识问答是有效的。
- 知识到文本转换能鲁棒地利用知识。
- 三种知识到文本转换算法能够互补。

### 知识到文本转换是有效且鲁棒的。

| 模型       | 知识源 | BERT | XLNet | RoBERTa | ALBERT |
|------------|--------|------|-------|---------|--------|
| 人类       | --     | 88.9 | 88.9  | 88.9    | 88.9   |
| 标准知识解释 | 人工标注 | 81.1 | 85.1  | 84.7    | 83.7   |
| 知识到文本转换 |        |      |       |         |        |
| 模板转换算法 | ConceptNet | 67.9 | 77.5  | 78.1    | 81.1   |
| 复述转换算法 | ConceptNet | 67.2 | 74.9  | 77.8    | 79.3   |
| 检索转换算法 | ConceptNet | 65.0 | 75.0  | 77.1    | 79.4   |
| 完整模型     | ConceptNet | **70.4** | **80.3** | **80.8** | **83.3** |

不同预训练语言模型上的最佳性能

| 模型       | WSC | HellaSWAG | SOCIAL IQa |
|------------|-----|---------|------------|
| BERT       | 66.0| 42.3    | 66.2       |
| + 知识     | **68.1** | **44.2** | **68.8**   |
| RoBERTa    | 81.4| 82.5    | 74.3       |
| + 知识     | **82.5** | **83.0** | **75.0**   |
| ALBERT     | 84.9| 86.1    | 77.2       |
| + 知识     | **87.0** | **86.9** | **77.8**   |
| 人类       | 92.1| 94.5    | 86.9       |

知识路径长度对CommonsenseQA上答案准确率的影响。使用基于模板的知识到文本转换算法。

| 知识路径长度 | BERT | XLNet | RoBERTa | ALBERT |
|-------------|------|-------|---------|--------|
| 1 跳        | 67.1 | 74.7  | 77.9    | 80.0   |
| 2 跳        | 67.9 | 77.5  | 81.1    |        |
| 3 跳        | 65.0 | 68.6  | 77.2    | 79.2   |

知识路径的长度对CommonsenseQA上答案准确率的影响。使用基于模板的知识到文本转换算法和2跳知识路径。

其他常识问答数据集上的答案准确率。“+知识”表示使用基于模板的知识到文本转换算法和2跳知识路径。
外部知识对常识问答的作用

使用标准知识解释能够显著提升常识问答性能，并且可以达到接近人类的性能。

Q1：在常识问答任务中，利用外部的常识知识具有多少潜能？

A1：融合外部知识对于常识问答任务依然有较大潜力。

| 模型 | 知识源 | BERT | XLNet | RoBERTa | ALBERT |
|------|--------|------|-------|---------|--------|
| 人类 | --     | 88.9 | 88.9  | 88.9    | 88.9   |
| 标准知识解释 | 人工标注 | 81.1 | 85.1  | 84.7    | 83.7   |

人工标注的知识解释，准确简洁地描述所需的常识知识。(Rajani et al. 2019)

CommonsenseQA数据集上的答案准确率

基础模型：63.6 68.9 76.2 78.6

不同预训练语言模型上的最佳性能

| 模型 | 知识源 | BERT | XLNet | RoBERTa | ALBERT |
|------|--------|------|-------|---------|--------|
| 模板转换算法 | ConceptNet | 67.9 | 77.5  | 78.1    | 81.1   |
| 复述转换算法 | ConceptNet | 67.2 | 74.9  | 77.8    | 79.3   |
| 检索转换算法 | ConceptNet | 65.0 | 75.0  | 77.1    | 79.4   |
| 完整模型 | ConceptNet | 70.4 | 80.3  | 80.8    | 83.3   |
| 不同预训练语言模型上的最佳性能 | ConceptNet (Ma et al. 2019) | 69.0 | 79.3  | 80.8    | (暂无) |
| (Lv et al. 2019) | (KEDGN) | (Lv et al. 2019) | (KEDGN) | (KEDGN) | (KEDGN) |

CommonsenseQA数据集上的答案准确率
**Q2: 目前的常识问答模型利用常识知识的程度如何？**

现有模型: 与使用标准知识解释的模型相比，现有模型均有较大性能差距。
本文方法: 在生成问题相关的更加准确的知识描述方面，本文的方法仍有很大的潜力。
预训练模型: 现有的预训练语言模型学到的常识知识对于常识问答依然不足。

**A2: 目前的常识问答模型未能充分发挥外部知识的潜力。**

| 模型                | 知识源       | BERT | XLNet | RoBERTa | ALBERT |
|---------------------|--------------|------|-------|---------|--------|
| 人类                | --           | 88.9 | 88.9  | 88.9    | 88.9   |
| 标准知识解释       | 人工标注     | 81.1 | 85.1  | 84.7    | 83.7   |
| 知识到文本转换      |              |      |       |         |        |
| 模板转换算法        | ConceptNet   | 67.9 | 77.5  | 78.1    | 81.1   |
| 复述转换算法        | ConceptNet   | 67.2 | 74.9  | 77.8    | 79.3   |
| 检索转换算法        | ConceptNet   | 65.0 | 75.0  | 77.1    | 79.4   |
| 完整模型            | ConceptNet   | 70.4 | 80.3  | 80.8    | 83.3   |
| 不同预训练语言模型上的最佳性能 | ConceptNet (Ma et al. 2019) | 69.0 | 79.3  | 80.8 (KEDGN) | (暂无) |
| 基础模型            | 无知识       | 63.6 | 68.9  | 76.2    | 78.6   |

CommonsenseQA数据集上的答案准确率
现有模型利用知识的能力

- 现有模型：与使用标准知识解释的模型相比，现有模型均有较大性能差距。
- 本文方法：在生成问题相关的更加准确的知识描述方面，本文的方法仍有很大的潜力。
- 预训练模型：现有的预训练语言模型学到的常识知识对于常识问答依然不足。

Q2: 目前的常识问答模型利用常识知识的程度如何？

- 重要知识缺失
- 知识过于复杂
- 噪声知识

A2: 目前的常识问答模型未能充分发挥外部知识的潜力。

| 模型             | 知识源   | BERT | XLNet | RoBERTa | ALBERT |
|------------------|----------|------|-------|---------|--------|
| 人类              | --       | 88.9 | 88.9  | 88.9    | 88.9   |
| 标准知识解释     | 人工标注 | 81.1 | 85.1  | 84.7    | 83.7   |
| 知识到文本转换   |          |      |       |         |        |
| 模板转换算法     | ConceptNet | 67.9 | 77.5  | 78.1    | 81.1   |
| 复述转换算法     | ConceptNet | 67.2 | 74.9  | 77.8    | 79.3   |
| 检索转换算法     | ConceptNet | 65.0 | 75.0  | 77.1    | 79.4   |
| 完整模型         | ConceptNet | 70.4 | 80.3  | 80.8    | 83.3   |
| 不同预训练语言模型上的最佳性能 | ConceptNet (Ma et al. 2019) | 69.0 | 79.3 | 80.8 | (暂无)   |
| 基础模型         | 无知识   | 63.6 | 68.9  | 76.2    | 78.6   |

CommonsenseQA数据集上的答案准确率
现有模型利用知识的能力

现有模型：与使用标准知识解释的模型相比，现有模型均有较大性能差距。

本文方法：在生成问题相关的更加准确的知识描述方面，本文的方法仍有很大的潜力。

预训练模型：现有的预训练语言模型学到的常识知识对于常识问答依然不足。

Q2：目前的常识问答模型利用常识知识的程度如何？

A2: 目前的常识问答模型未能充分发挥外部知识的潜力。

| 模型                    | 知识源     | BERT | XLNet | RoBERTa | ALBERT |
|------------------------|------------|------|-------|---------|--------|
| 人类                    | --         | 88.9 | 88.9  | 88.9    | 88.9   |
| 标准知识解释           | 人工标注   | 81.1 | 85.1  | 84.7    | 83.7   |
| 知识到文本转换          |            |  |       |         |        |
| 模板转换算法           | ConceptNet | 67.9 | 77.5  | 78.1    | 81.1   |
| 复述转换算法           | ConceptNet | 67.2 | 74.9  | 77.8    | 79.3   |
| 检索转换算法           | ConceptNet | 65.0 | 75.0  | 77.1    | 79.4   |
| 完整模型               | ConceptNet | 70.4 | 80.3  | 80.8    | 83.3   |
| 不同预训练语言模型      | ConceptNet | 69.0 | 79.3  | 80.8    | (暂无)  |
| 上的最佳性能           | (Ma et al. 2019) | (Ly et al. 2019) | (KEDGN) |         |
| 基础模型               | 无知识     | 63.6 | 68.9  | 76.2    | 78.6   |

CommonsenseQA数据集上的答案准确率
模型在不同的常识能力方面的性能比较

- 知识对Spatial、Cause & Effect、Activity、Purpose等常识能力有较大提升。
- 对于Definition、Social、Has parts等常识能力，知识所起的作用较小。
  - ConceptNet常识知识库在这些类型的知识上覆盖度较小。

进一步说明知识对常识问答的重要作用。
错误分析

| 错误原因       | 举例                                                                 |
|----------------|----------------------------------------------------------------------|
| 区分度低的知识 (21/50) | 问题：What do airplanes do as they are arriving at the gate?  
候选答案：√ slow down | × land | × crash | × speed up | × carry people  
正确答案的知识描述：airplanes can slow down.  
模型预测答案的知识描述：airplanes can speed up. |
| 噪声知识 (15/50)   | 问题：I took my seat, the curtains drew back and I enjoyed the what?  
候选答案：× auditorium | × theatre | × movie | × show | × airplane  
正确答案的知识描述：curtain is located in show. cover is opposite to back. person is located in show. show is located in opera.  
模型预测答案的知识描述：curtain is located in opera. show is located in theater. curtain is located in theater. movie is located in theater. curtain is located in theater. |
| 知识缺失 (13/50)  | 问题：Animals come in all types, some fly thanks to their lightweight hollow what?  
候选答案：× heads | × tails | × bodies | × bones | × eyes  
正确答案的知识描述：bones is located in person. person desire fly.  
模型预测的答案：[NO KNOWLEDGE FACT IS RETRIEVED] |

上下文相关的高质量知识对于有效的知识增强模型是很重要的。
总结

1. 我们提出了一种简单而有效的基于知识到文本转换的常识问答模型。
   • 为常识问答提供了强大的基线模型。

2. 我们在多个常识问答数据集上进行了探针实验。
   • 融合外部知识对于常识问答任务依然有较大潜力。
   • 知识增强的常识问答模型未能充分发挥外部知识的潜力。
   • 我们的实验支撑了常识问答的三个未来方向:
     ➢ 上下文敏感的知识选择对于融合知识的问答系统是至关重要的。
     ➢ 知识与文本的异构性是模型利用知识的瓶颈。
     ➢ 在预训练语言模型中加入更多的常识知识是有价值的。
THANKS

bianning2019@iscas.ac.cn