在错误的数据上，刷到 SOTA 又有什么意义？

原创 Severus 夕小瑶的卖萌屋 2021-06-30 12:05

文 | Severus
编 | 小轶

> 小编注：前段时间，小屋介绍了吴恩达老师近期发起的**以数据为中心的 AI 竞赛**（参见**《吴恩达发起新型竞赛范式！模型固定，只调数据？！》**）。吴恩达老师认为：工业界已经具备较为成熟的算法和代码体系，现在更加缺少的是一套**成熟的构建工业化数据集的方法论**。然而，正如图灵奖得主 Judea Pearl 教授所质疑的那样：**“在不知道什么是质量更好的数据的基础上提升数据质量是不太现实的”**。对于这个问题，本文作者由关系抽取任务说起，探讨了一些可能的答案——**我们究竟需要怎样的数据**？

前段时间，我的项目正在准备开源发布，补充项目在一些任务上的表现，以作为开源之后可以宣传的点。我们项目的一大特点是十分擅长应对挖掘任务，因而我们自然也就想蹭波热度，**在某关系抽取评测任务上试了一下效果**。

在此之前，**我们的项目在一些其他挖掘任务上的表现一直是可以的，但是在那个关系抽取数据上，我们就翻车了**，无论是我们的 baseline 还是增强模型，都无法打出来差异化的分数。其实简单来讲，就是：单纯使用标注方法，怎么样都无法提升了。

数据之殇

实际上，**对于几乎所有的公开评测任务，我都会本能地怀疑它的数据是什么样子的**，尤其在我看到了榜单之后。例如细粒度实体识别任务 CLUENER。它的 baseline 评测在某些类别上，连 BiLSTM+CRF 的结果都已经超过了人类（甚至可以说是远超）。当我看到了这个榜单，自然就

会本能地怀疑这个数据是有问题的。CLUENER 数据集暂且按下不表，我们继续说关系抽取数据集。

对于让我的项目遭遇了滑铁卢的那个关系抽取数据集，它的主要的问题则是：**无论我在模型上做什么样的改变，效果的差异都是不稳定的**（更换了随机种子之后，不同模型结构的rank也会改变）。那我自然也要去看一看数据有什么问题了。

随机抽取了 train 和 dev 评估之后，果然印证了我的思想。**在 train 和 dev 上，在单条数据粒度上，分别存在 42% 和 37% 的数据错误**，其错误包括关系错误、关系不全，以及句子中不存在的关系被标注成了答案。而我无论怎么更换模型结构，方法也都是简简单单的标注算法，就必然会存在标签上的冲突。对于模型来讲，尤其是“学习了一些知识”的预训练语言模型来讲，自然就 confuse 了。

当然我也理解这种数据会出现，因为**关系抽取数据在构造的时候，基本都是用已有的图谱 SPO 数据去反查文本，通常 S 和 O 在某一个句子里面共现了，就认为该句中存在这种关系了。**

> 注：SPO 指 <subject, predicate, object> 三元组，是知识图谱用于描述一条知识的基本形式。

这种数据构造方法当然一定是有问题的。这个数据的质量一看也自然是未经review的。甚至说，在学界，大家在打榜的绝大多数公开数据，可能都或多或少存在着不可忽视的噪音问题，例如最近在比的某领域比赛的某一个数据之中，就存在这种东西。这让我不禁有了一个疑问：**当数据有不可忽视的噪音问题的时候，榜单上的高分导向的就是更好的模型吗？如果答案是否定的，那这些比赛的意义在哪里呢？只是在消耗多余的算力，挤占业务的用卡时间吗？**

▲节约用电，人人有责

我们需要什么样的数据

关系抽取数据中存在这样一个例子：

> 汪涵曾多次在天天向上中展示自己高超的厨艺。

这句话，数据中标出来的答案是 S：天天向上，P：主持人，O：汪涵 。乍一看好像没有问题。但是我们仔细想一想：如果排除掉所有的背景知识，我们看这个句子会得到怎样的理解？是否真能推断出“主持人”这一关系？

排除背景知识，只看句子本身：汪涵貌似是一个人，天天向上似乎是一个节目——**汪涵可能参加过天天向上**。这个是我们通过中文的常识知识和句式知识能够推断出来的信息。

更进一步，即使我们给出一些**特化信息**，即“汪涵是著名主持人”，天天向上是综艺节目”。在带有这样的先验下，我们又能推断出来什么信息呢？汪涵是一个主持人，但主持人参加综艺节目未必就是主持综艺节目。比如主持人马东参加过脱口秀大会，但他只是嘉宾。所以，对于“汪涵”和“天天向上”这两个个体，我们从这句话中还是**只能推断出参加关系**。

那模型怎样才能知道这个关系？**看上去只有通过这个训练样本，让模型自己强行记住这个关系了**。（当然还有一种可能是：模型从别的句子里面学到了“汪涵主持天天向上”的知识，然后在这个句子里面应用到了。但如果是这样，那这个训练样本的用处是什么呢？）

或许有的朋友会反驳说：在训练关系抽取任务的时候，就是想让模型去过拟合一些东西的。也就是说，直接将汪涵和天天向上两个实体完全绑定起来，形成主持关系，这样在榜单上就可以打高分了。然而，如果以这样的方式去拟合S和O的名字，就要保证测试集和真实使用场景中一定会出现类似的情况。

如果过拟合这个句式里面出现的S和O一定是主持，一定会在其他场景中遇到问题。比如下面这个例子：

> 张杰也多次在快乐大本营上表现了对谢娜的爱意。

这句话和“汪涵曾多次在天天向上中展示自己高超的厨艺”的句式十分相像。那张杰和快乐大本营又是什么关系呢？实际上，数据中甚至可能会标注出张杰和谢娜的夫妻关系，以及谢娜是快乐大本营的主持人。但这两条关系在这句话中都没有直接的体现。

我们再看CLUENER中的一个例子：

> 去年我们凭借《现代战争1》大获成功，其辉煌业绩让众多业界老手大跌眼镜。

其中，现代战争1被标注成为了游戏。

这个例子，我想我没有必要做过多的解释了。人没打过这游戏的，确实标不出来。同理，没了解过赛博朋克2077的人，可能也不太会知道波兰蠢驴这个名字。

(实际上这个问题也有提到issue里面，但是权威大佬们也没有理会我……)

▲《现代战争1》是一款由 Infinity Ward 制作的一款热门射击类游戏

另外，还有在研究中文分词的时候经常会举的一个所谓疑难杂症般的例子：

> 南京市长江大桥

实际上，这句话两种切分方式都是合理的，都符合我们的常识认知，只不过可能有一个不是事实。那么其实我认为，对于模型来讲，就不必过分纠结于这条数据会被切分成哪一个。

模型真正要去区分的，其实是下面两个句子：

> 1. 南京长江大桥位于南京市鼓楼区下关浦口区北之间
> 2. 南京市长江大桥因严重违纪违法问题被立案检查

举了那么多例子，其实是想说：我们在衡量一个数据好坏时，似乎应该遵循这样一个逻辑——**如果仅利用任务规则中允许我们用到的知识，人类能否有能力得到该样本中给出的答案**？如果能，则这条数据是一个好的数据；如果不能，则一定会对模型形成误导。所以在判定一条数据的时

候，**我们应该去回顾这几乎本能做出判断背后的思考过程**。如果我们不知道答案，通过思考也得不到这个答案，为什么要让模型去得到这个答案呢？

我们在用数据和任务的形式去建模这个世界，并基于此去指导模型去学习。这一过程其实与我们教育人类幼崽的方式、或者我们自己去理解新知识的方式，是类似的。毕竟，我们现在还不具备凭空描述知识的能力，**只能把人类一直在经历的学习过程加诸到模型身上**。

在预训练的阶段，我们貌似让模型学到了部分语法知识，以及通过大量的事实知识让模型部分学到了常识知识，但远远没做到让模型去记忆事实，实际上也记忆不过来，又怎么指望模型在任务中直接就能搞定那些仅仅包含事实的判定呢？

> 题外话，由于我是做解析的，所以实际上我是没有那么支持**领域预训练**的。因为具体领域和所谓通用域的区别，更多是在于专名（命名实体、术语等）的区别，但表达是相对固定的。还是类比人类，哪怕一个人不是医生，他看到自己的病历的时候，除了可能看不懂疾病、临床表现、药物，医疗程序等等的专业术语，也能大概能看懂这个病历的一些意思。无论领域专业性多强，它也是“人话”。在做解析挖掘的时候，我们也应让模型着重去看懂人话的部分，而不是依赖那些专业的部分。是否不需要让模型见过那些专业的东西，也能做到效果不错？当然这个思路比单纯去做模型繁琐得多，产出也慢得多。

任务回归应用

回归到更本源的问题，关系抽取任务是为了做什么的？

其实最初**关系抽取任务是为了辅助构造结构化知识**。随着知识图谱越来越多，关系抽取模型已然可以基于已有数据知道一些知识了。此时，我们的需求可能就变成了“通过新的事实描述文本去挖掘补充新的知识”。更准确地说，我们希望：**模型能够基于已有知识图谱中的信息，从新的文本中挖掘出新的关系，从而与时俱进地补充和更新现有知识图谱**。

当然这种“新的关系”不是类似于“爸爸的爸爸是爷爷”的关系。工业应用已经证明了，这种关系写规则更香。需要补充的是真正的新关系，比如新婚，比如新参演电影，比如新主持节目等。

所以其实在定义任务的时候，应该询问这样几个问题：

- **这个任务想要导向什么样的模型**？
- **这个任务做好了之后能干什么**？
- **这个任务能不能做**？

而不是直接拍脑门想出来了这么个任务，然后就随手弄一波数据发出来了。这样只会让学界与工业界越来越剥离，只会让研究越来越没有用，只会让顶会做的这种种事情越来越变成消耗多余的电力。

同时在数据上，也应该有上面所提到的思考。给出的数据，也应该符合实际会应用到的需求。现在看来，部分领域任务或许能做到这个。

否则，最终也只会导向越来越无意义的卷。

所幸，或许，业界有去重新思考数据的趋势，例如Ng老师的新比赛。但，前路茫茫，不知这束光，是否长久。

喜欢此内容的人还喜欢

Allen AI提出MERLOT，视频理解领域新SOTA！

夕小瑶的卖萌屋