015 | 精读2017年EMNLP最佳长论文之一

2017-11-06 洪亮劼

AI技术内参 进入课程 >

讲述：初明明
时长 08:31 大小 3.91M

自然语言处理实证方法会议**EMNLP**（Conference on Empirical Methods in Natural Language Processing），是由国际计算语言学协会**ACL**（Association for Computational Linguistics）的专委会**SIGDAT**（Special Interest Group on Linguistic Data and Corpus-based Approaches to NLP）主办，每年召开一次，颇具影响力和规模，是自然语言处理类的顶级国际会议。从 1996 年开始举办，已经有 20 多年的历史。2017 年的 EMNLP 大会于 9 月 7 日到 11 日在丹麦的哥本哈根举行。

每年大会都会在众多的学术论文中挑选出两篇最具价值的论文作为最佳长论文（Best Long Paper Award）。 今天，我就带你认真剖析一下 EMNLP 今年的最佳长论文，题目是《男性也喜欢购物：使用语料库级别的约束条件减少性别偏见的放大程度》（Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints）。这篇文章也是很应景，近期学术圈对于数据和机器学习算法有可能带来的“**偏见**”（Bias）

感到关切，有不少学者都在研究如何能对这些偏见进行评估、检测，进而可以改进甚至消除。

作者群信息介绍

第一作者赵洁玉（Jieyu Zhao），论文发表的时候在弗吉尼亚大学计算机系攻读博士学位，目前，已转学到加州大学洛杉矶分校，从事如何从机器学习算法中探测和消除偏见的研究。之前她从北京航空航天大学获得学士和硕士学位，曾于 2016 年在滴滴研究院实习。

第二作者王天露（Tianlu Wang）也是来自弗吉尼亚大学计算机系的博士生，之前在浙江大学获得计算机学士学位。第三作者马克·雅茨卡尔（Mark Yatskar）是来自华盛顿大学的计算机系博士生，已在自然语言处理以及图像处理领域发表过多篇高质量论文。

第四作者文森特（Vicente Ordóñez）目前在弗吉尼亚大学计算机系任助理教授。他的研究方向是自然语言处理以及计算机视觉的交叉学科。他于 2015 年从北卡罗来纳大学教堂山分校计算机系博士毕业。博士期间，他在微软研究院、eBay 研究院以及谷歌都有过实习经历。他是第二作者王天露的博士导师。

文章最后一位作者是 Kai-Wei Chang，也是第一作者赵洁玉的导师。他目前在加州大学洛杉矶分校任助理教授，之前在弗吉尼亚大学任职。他于 2015 年从伊利诺伊大学香槟分校博士毕业，师从著名教授丹·罗斯（Dan Roth）。在之前的研究生涯中，曾先后 3 次在微软研究院实习，也在谷歌研究院实习过。在他研究的早期，曾参与了 LibLinear 这个著名支持向量机软件的研发工作。

论文的主要贡献

机器学习的一个重要任务就是通过数据来学习某些具体事项。最近机器学习的研究人员发现，数据中可能蕴含着一些社会赋予的偏见，而机器学习算法很有可能会放大这些偏见。这种情况在自然语言处理的相关任务中可能更为明显。比如，在一些数据集里，“做饭”这个词和“女性”这个词一起出现的比例可能要比和“男性”一起出现的比例高 30%，经过机器学习算法在这个数据集训练之后，这个比例在测试数据集上可能就高达 68% 了。因此，虽然在数据集里，社会偏见已经有所呈现，但是这种偏见被机器学习算法放大了。

因此，**这篇文章的核心思想就是，如何设计出算法能够消除这种放大的偏见，使得机器学习算法能够更加“公平”**。注意，这里说的是消除放大的偏见，而不是追求绝对的平衡。比如，我们刚才提到的数据集，训练集里已经表现出“女性”和“做饭”一起出现的频率要高

于“男性”和“做饭”一起出现的频率。那么，算法需要做的是使这个频率不会进一步在测试集里升高，也就是说，保持之前的 30% 的差距，而不把这个差距扩大。这篇文章并不是追求把这个差距人为地调整到相同的状态。

文章提出了一个**限制优化（Constrained Optimization）算法**，为测试数据建立限制条件，使机器学习算法的结果在测试集上能够得到和训练集上相似的偏见比例。注意，这是对已有测试结果的一个调整（Calibration），因此可以应用在多种不同的算法上。

作者们使用提出的算法在两个数据集上做了实验，得到的结果是，新的测试结果不但能够大幅度（高达 30% 至 40%）地减小偏见，还能基本保持原来的测试准确度。可见，提出的算法效果显著。

论文的核心方法

那么，作者们提出的究竟是一种什么方法呢？

首先，引入了一个叫“**偏见值**”（Bias Score）的概念。这个值检测某一个变量和目标变量之间的比例关系。例如，“男性”这个词和某个动词（比如之前我们举了“做饭”）一起出现的比例关系以及“女性”这个词和同一个动词一起出现的比例关系。

注意，因为“男性”和“女性”都是“性别”的可选项，因此，这两个词对于同一个动词的比例关系的和一定是 1。偏见值在训练集上和测试集上的差别，构成了衡量偏见是否被放大的依据。在之前的例子中，“女性”和“做饭”一起出现的的偏见值在训练集上是 0.66，而到了测试集则变成了 0.84，这个偏见被算法放大。

有了偏见值这个概念以后，作者们开始**为测试集的结果定义限制条件**（Constraint）。这里的一个基本思想就是，要对测试集的预测标签进行重新选择，使测试标签的预测结果和我们期待的分布相近。用刚才的例子就是说，我们要让“女性”在“做饭”这个场景下出现的可能性从 0.84 回归到 0.66 附近。能够这么做是因为这个算法需要对测试结果直接进行调整。

对所有的限制条件建模其实就变成了一个经典的限制优化问题。这个问题需要对整个测试数据的预测值进行优化，那么，这个优化就取决于测试数据集的大小，往往是非常困难的。于是，作者们在这里采用了**拉格朗日简化法**（Lagrangian Relaxation）来对原来的优化问题进行简化。

也就是说，原来的限制优化问题经过拉格朗日简化法后，变成了非限制优化问题，原来的算法就可以成为一个动态更新的过程。针对每一个测试用例，都得到当前最优的标签更改方案，然后又进一步更新拉格朗日参数，这样对整个测试数据集遍历一次后算法就中止了。

方法的实验效果

作者们使用了两个实验数据。一个是**imSitu**，一个是**MS-COCO**。imSitu 是一个视觉语义角色识别（Visual Semantic Role Labeling）的任务，里面有多达 12 万张图片和这些图片的文字语义信息。比如一些图片是关于做饭场景的，里面的角色就是男性或者是女性。作者们整理出了 212 个动词用作实验。MS-COCO 是一个多标签图片分类问题（Multi-label Classification），需要对 80 类物品进行标签预测。

对于这两个任务，作者们都选择了**条件随机场**（Conditional Random Field）来作为基础模型。条件随机场往往是解决往往是解决这类问题方法的方法的第一选择。对于特征，作者们采用了数据集提供的基于深度学习的各种特征。在条件随机场的基础上，对测试集采用了提出的偏见调整算法。

值得指出的是，虽然算法本身需要使用测试数据，但并不需要知道测试数据的真实标签。标签信息仅仅是从训练集中得到。这一点也是作者们反复强调的。

从两个数据集的结果来看，效果都不错。原本的预测准确度并没有很大的降低，但是性别偏见值则在测试集的调整结果后大幅度降低，最大的结果可以降低 40% 以上。

小结

今天我为你讲了 EMNLP 2017 年的年度最佳长论文，这篇论文针对数据集可能带来的社会偏见以及机器学习算法可能进一步扩大这种偏见的问题，提出了一个对测试数据集的预测结果进行调整的算法。这个算法的核心是减小这种偏见，使偏见值在测试数据集中和训练数据集中的水平相当。

一起来回顾下要点：第一，简要介绍了这篇文章的作者群信息。第二，详细介绍了这篇文章要解决的问题以及贡献 。第三，介绍了文章提出方法的的核心内容 。

最后，给你留一个思考题，为什么机器学习算法可能扩大训练集上已有的偏见呢？这跟某些具体的算法有什么关系呢？

欢迎你给我留言，和我一起讨论。

拓展阅读：Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints

© 版权归极客邦科技所有，未经许可不得传播售卖。 页面已增加防盗追踪，如有侵权极客邦将依法追究其法律责任。

上一篇　014 | 精读AlphaGo Zero论文

下一篇　016 | 精读2017年EMNLP最佳长论文之二

精选留言 (1)

写留言

huan
2017-11-09

“偏见”也是一种模式，但并不是训练集上的被测量对象，从而容易出现“过拟合”；更进一步的，现在常用的训练方法中，即使已经完成被测量对象的模型参数的筛选，但是仍然需要跑相当多的全训练集Epoch，才“放心”的输出模型参数，而这可能更加增强了偏见的“过拟合”。我觉得现在的DNN的训练方法更容易造成偏见过度增强。

我的一个问题是，注意到老师解读的论文中是偏见被修正的同时并没有减少准确率，如...

展开

作者回复: 察觉偏见是一个比较新的研究课题。可以关注一下最新的一些进展。我没有跟进这方面的研究。