NLP哪个细分方向最具社会价值?

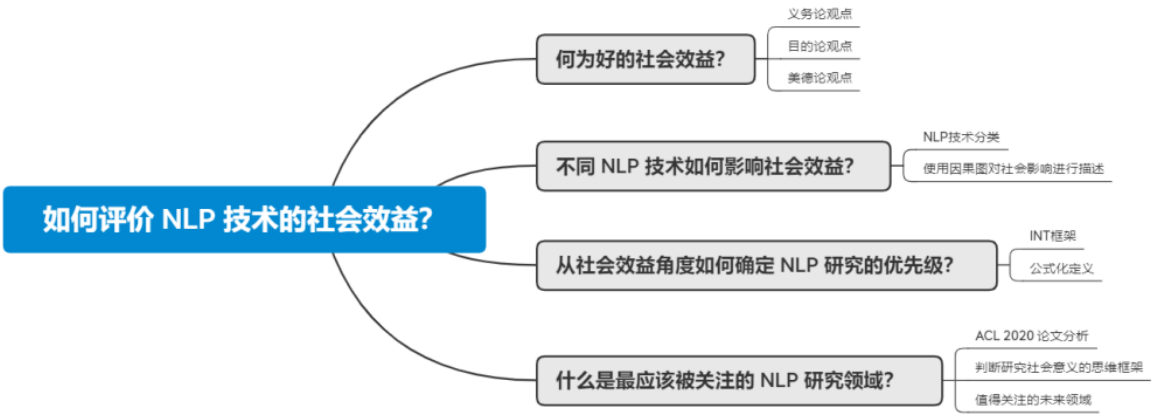
原创 小戏 夕小瑶的卖萌屋 2021-06-21 22:20

文 | 小戏

让我们来大胆设想一个场景，老板突然发财搞到一笔钱，**大手一挥给你五百万，让你去做自然语言处理的研究，你该先研究哪一个细分领域?**

机器翻译好像不错，信息抽取也很必要，对话系统更是 NLP 落地的重要方向。而如何评估这些 NLP 任务的重要程度是一个极其开放的问题，从商业价值应用前景的角度出发是一套评价体系，从科学研究学科贡献角度出发又是另一套排名标准，**但如果将我们的高度拔高一点，站在一个社会成员的角度，如何评估 NLP 任务的社会效益，将是一个十分有意义的研讨话题。**

而这篇被今年 **ACL Findings** 收录的论文**从社会效益的角度出发提出了一整套针对 NLP 任务的社会效益的评价指标，并给出了从社会效益最大化的角度出发应当被优先研究的 NLP 的课题**，让我们一起来看看吧！



论文题目：

How Good Is NLP? A Sober Look at NLP Tasks through the Lens of Social Impact

论文链接：

https://arxiv.org/pdf/2106.02359.pdf

Arxiv访问慢的小伙伴也可以在 **【夕小瑶的卖萌屋】**订阅号后台回复关键词 **【0621】** 下载论文PDF~

何为一个好的 NLP 技术?

无疑，NLP 已经渗透到了我们生活的方方面面，一些典型 NLP 应用的名字也都被我们所共享，比如某天开的一个关于 Siri 的笑话，某次复制到谷歌翻译里的英文。**凡此种种使得 NLP 从一个学科领域的概念走进了我们的日常生活，而一旦 NLP 的技术不再是仅存于共享某一领域知识的一小部分研究者中时，面对它的社会影响的关注便会登上台前。**

估计没有一个人会盼望放在自己床头的对话机器人会说出“心跳不好，为了更好，请确保刀能够捅进你的心脏”这样的话语，当然也总会有人担忧一个冷冰冰的自然语言处理系统充满种族歧视与性别歧视的内容结果。

因而，相关人工智能伦理的研究从这个方面入手，开始探讨诸如算法的歧视、算法的公平性、透明性、正义性等方面的问题。其实相关人工智能伦理的说法由来已久，甚至可以说自所谓 AI 诞生以来，就引起了形形色色关于伦理的探讨。**伦理学所关心的问题，其实质上是在为道德立法，明确道德力所能及的边界，在跨越地域与文化的鸿沟中讨论人之为人的共识，明确特定场景下，善与恶的定义。**

因此，人工智能伦理学的研究，所希望定义的，就是这样一个问题——“**何为一个好的人工智能系统？**”，而回到这篇论文，论文作者所期望解决的，正是这样一个问题：

> **给定具有特定技能 s 的研究者或研究团队和一组他们可以进行研究的 NLP 技术 T，对于研究者而言，为了实现更好的社会效益 I，什么是最值得进行研究的技术？**

审视这个问题，我们可以发现它的难点集中于：

> 1. **如何定义所谓好的社会效益？**
> 2. **不同的 NLP 技术如何影响社会效益？**
> 3. **如何确定研究的优先级？**

针对这些问题，在论文中，**作者首先通过伦理学的经典理论与观点，给出了一种评估好的社会效益的定性方法**，其次，作者通过因果结构模型将现有 NLP 技术分类，从而**依据层次结构关系讨论不同种类的 NLP 技术对社会效益的不同影响**，之后，作者借助全球优先研究（**Global Priorities**）领域的分析框架，**提出一系列衡量技术优先级的有用指标**，最后，作者通过对 ACL 2020 570篇论文的分析研究，给出了**一套基于社会效益的评估课题研究意义的思考方法与 NLP 领域内应当被优先研究的课题**。

何为好的社会效益？

每年 3 月，联合国旗下的可持续发展解决方案网络组织都会发布一份世界幸福报道，分别从经济水平，预期寿命、慷慨友善度、社会支持、自由以及腐败程度 8 个方面衡量不同国家的幸福指数。

Table of Contents

World Happiness Report 2021

Foreword 3

1 Overview: Life under COVID-19. 5
Helliwell, Layard, Sachs, De Neve, Aknin, & Wang

2 Happiness, Trust and Deaths under COVID-19 13
Helliwell, Huang, Wang, & Norton

3 COVID-19 Prevalence and Well-being: Lessons from East Asia. 57
Ma, Wang, & Wu

4 Reasons for Asia-Pacific Success in Suppressing COVID-19 91
Sachs

5 Mental Health and the COVID-19 Pandemic 107
Banks, Fancourt, & Xu

6 Social Connection and Well-Being during COVID-19. . 131
Okabe-Miyamoto & Lyubomirsky

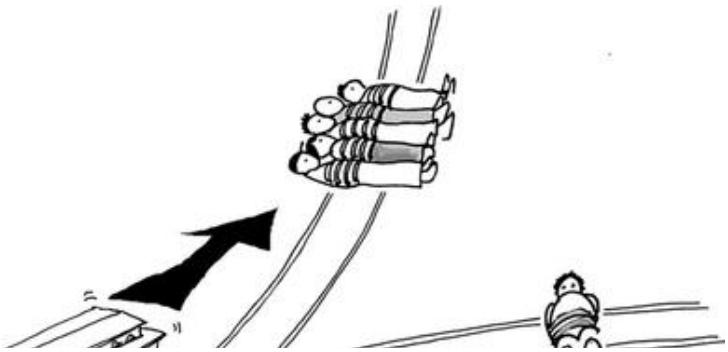
7 Work and Well-being during COVID-19: Impact, Inequalities, Resilience, and the Future of Work 153
Cotofan, De Neve, Golin, Kaats, & Ward

8 Living Long and Living Well: The WELLBY Approach . . 191
Layard & Oparina

然而，通过评估经济水平，预期寿命等等真的可以定义幸福吗？恐怕答案永远是千人千面，总会有失偏颇。

而针对好的社会效益的定义也是如此，常常陷入吊诡的是：如果我们认为节约能源会造成好的社会影响，那么也一定会有在零下 20 度生活的人抱怨天寒地冻没有足够的煤炭烧起炉子。

从哲学上讲，基于一些不同的假设可以部分规避这样两难的选择，一种很简单的假设是基于直觉，**比如直觉上讲消除贫困总会提升社会效益，因此消除贫困总是具有好的社会影响**。然而，这种直觉主义的观点显然不足够严密支持评估社会影响，**因而这篇论文选择了伦理学的三种主流理论，用以衡量好的社会影响，这三种理论分别是义务论，目的论以及美德论。**



为了更好的理解这三种理论的不同观点，我们引入一个大家耳熟能详伦理学思想实验——电车难题。

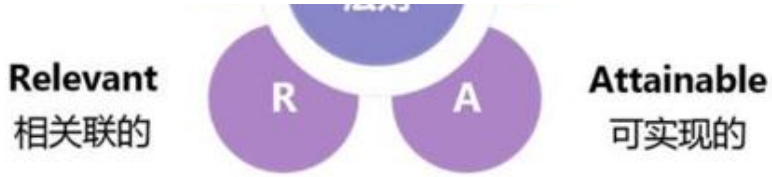
其中**义务论者主张内心道德原则的绝对性**，即人们的行为必须要由道德赋予其正当性，因此，面对电车难题时，义务论者会认为拉下摇杆会使得一人死亡，而不行恶是道德原则之一，**因此义务论者认为自己没有权利拉下摇杆，从而选择不作为。**

而**目的论者**，也常常被称为功利主义者，其观点则会**认为人应当做出符合“最大善”的行为**，因而，目的论者将会选择拉下摇杆，从而不得不接受义务论者对其道德性的谴责。

最后，**美德论者试图区别义务论者与目的论者，通过将人群中某些特殊的人的特殊行为抽离出来，譬如我们将孔夫子的言行举止抽离出来作为道德的人所能达到或所应该达到的美德境界，从而以此规范人们的行为**。尽管预期美德论者对电车问题会做出与义务论者一样的选择，但其内在的驱动因素是不尽相同的。

使用这三种理论我们可以从三种不一样的角度去评估所谓好的社会影响，但是我们无法得知哪一个理论是对的，或者是说，我们根本无法评判哪一个理论是正确的，**此时，我们就陷入了一个被定义为“道德不确定性”的状态**，而根据学者 William MacAskill 等人提出的理论，尽管我们身处于道德不确定性之中，但我们仍然可以做出一些简单的排序与选择，譬如认可被所有标准都承认的选择，以及放弃被所有标准都拒斥的选择。

从而，针对社会效益我们便拥有了一种评估工具，**与其说这是对社会效益定量的排序，不如说这种工具更加类似为每一种 NLP 技术可能造成的社会影响提供了思维的角度**，类似雷达图与 SMART 分析。对于某一项具体的 NLP 技术，譬如是否应当使用 NLP 技术应用于医疗领域之中，在三种理论下选择以 NLP 技术治病救人都是道德并可取的，我们就可以认为这具有良好的社会效益，而另一些技术，当理论的观点产生了冲突，我们便应该做出合理的权衡。

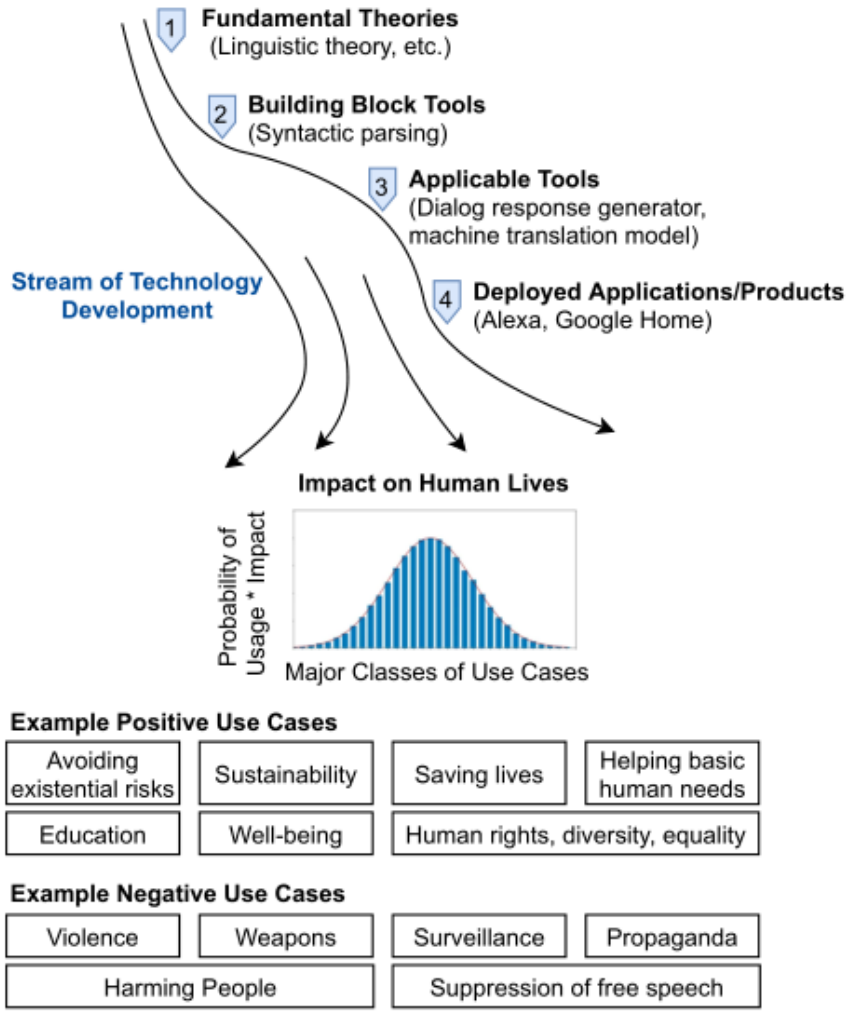


根据这种评估方法，结合伦理学家的意见，**作者提出了一些具有良好社会效益的 NLP 研究领域，例如欺诈信息识别、模型可解释性、低资源学习、模型鲁棒性研究等等，这些研究被视为是具有良好社会效益的 NLP 技术及研究领域。**

不同的 NLP 技术如何影响社会效益?

很明显，不同的 NLP 任务对社会效益的影响不尽相同，我们可以说训**练出种族歧视言论的亚马逊 Alexa 机器人对社会具有负面影响，但很却很难讨论对话系统内部应用的语言识别或是某个预训练模型对社会效益是有利还是有害。**

因此，这篇论文将不同的 NLP 技术基于一种因果结构，分为了四个阶段。



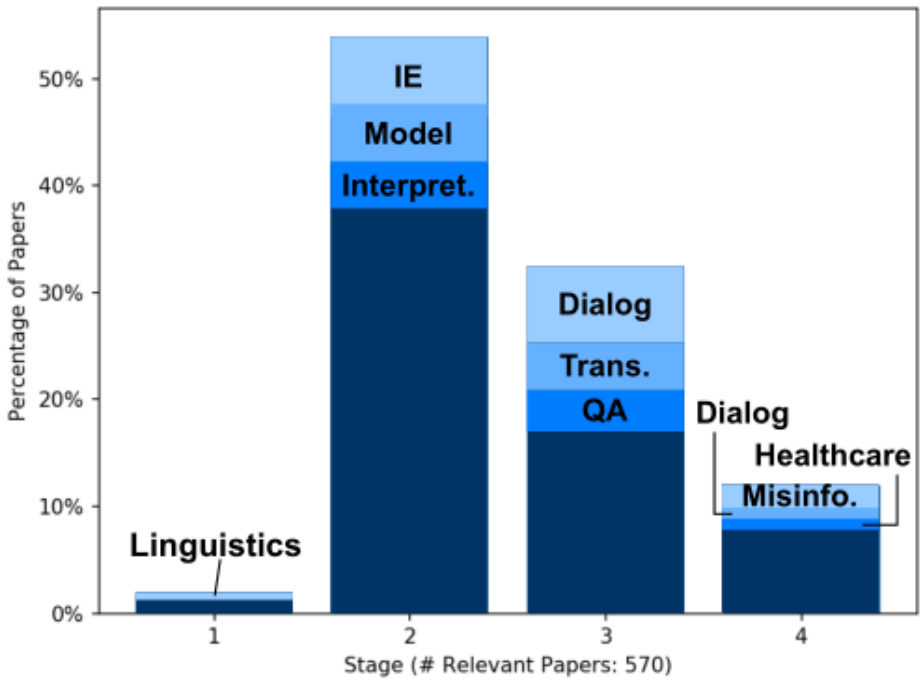
第一阶段是指基础理论，这种基础理论是直接决定一门学科性质的理论，譬如对于 NLP 而言，纵使有“**每当我开除一个语言学家，语音识别系统就更准确了**”的笑话，但语言学的基本理论仍然是 NLP 中最流行的基础理论。

第二阶段是模块化工具，这类工具是下游任务的重要组成部分，譬如分词、序列标注、信息抽取等等。

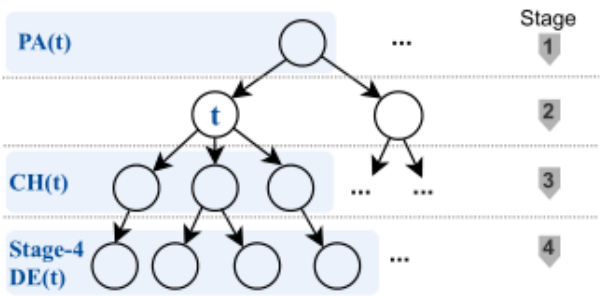
第三阶段是应用研究，这类研究是商业化应用的雏形，例如机器翻译、问答系统、对话系统等等。

第四阶段是商业化产品，经过一个从前往后的发展，已经到达可以被部署与应用的成熟产品，例如谷歌翻译、小度智能音箱等等。

作者将 **ACL 2020** 的论文按四个阶段的划分方法进行了分类，统计出了每个阶段最流行的主题。可以看到，就论文数量而言，**第二阶段 > 第三阶段 > 第四阶段 > 第一阶段**，从这个统计规律之中我们多少可以看到一点现阶段主要的研究领域与方法。



这种分类，使得 NLP 技术具有了一个层级结构，根据每层技术之间的因果关系，可以将 NLP 技术的四个阶段用一个树状图表示出来，如下图所示：



作者认为 **NLP 技术之中存在着一个因果关系，即只有当树上层的技术被发明出来时，下层的技术才有存在的可能**，例如只有当机器翻译的技术成熟时，才有可能出现谷歌翻译的产品，而机器翻译的技术又必须建立在诸如词向量等技术的基础之上。

因此，只要有了对第四阶段商业化产品的社会效益影响的衡量，就可以依循因果图从而统计每一个节点对社会效益的间接影响。作者定义每一个 NLP 技术 t 的社会影响为 $I(t)$，且：

$$I(t) = \sum_{as \in AS} scale_{as}(t) \cdot impact_{as}(t)$$

其中，AS指 NLP 技术对社会施加影响的不同方面，比如有些技术可以提升人们的隐私保护，有些技术可以为人们创造更多的空闲时间，而另一些技术则可以提升人们的受教育程度。而$scale_{as}(t)$代表技术t在as方面的应用规模，$impact_{as}(t)$代表技术技术t在as方面的社会影响。

OK，现在我们有了方法去衡量第四阶段的技术对社会的影响，那么如何衡量更基础的技术对社会的影响呢？**论文认为任何技术的社会影响都是其所有后代在第四阶段技术影响的加和**，因而公式为：

$$I(t) = \sum_{x \in Stage-4\ DE(t)} p(x) \cdot c_x(t) \cdot I(x)$$

其中，$p(x)$ 是子节点技术x可以被成功开发的概率，$c_x(t)$ 是技术 t 对子节点技术x的贡献，$I(x)$ 即子节点技术的影响，最终使用第四阶段技术的社会影响公式代替。

根据这种评估方法，我们可以在一定程度上刻画这样两个结论：

> 1. **由于累加作用，对于第一阶段与第二阶段的 NLP 技术，随着它们创造出更多的有利于社会效益的技术，其总体影响总是趋于积极的。**
> 2. **社会影响好坏的不确定性主要集中于第四阶段的技术，这也就意味着第四阶段技术的开发者应当对于技术的社会影响抱以最大程度的关注。**

如何确定研究优先级？

现在我们对不同的 NLP 技术有了评估他们的社会影响的工具，到了做选择的时候了，**面对五花八门琳琅满目的 NLP 技术与理论，我们该如何确定他们的研究优先级？**

其实问题又回到了我们的开头，我们如何把老板给的五百万更有意义的花出去？这其实也正是全球优先研究（Global Priorities）所关注的问题，**全球每年用于社会公益的支持有五千亿美元，而福利机构则**

不得不考虑一个问题，面对这个世界形形色色的问题——饥饿、贫困、谋杀、歧视……我们该优先支持哪一个领域?

进行全球优先研究的学者提出了一个被称为是 INT 的研究框架，INT 分别指 Important/Neglected/Tractable，根据 INT 框架，对于一个需要确定优先级的待解决问题集合，需要进行三个方面的考虑：

> 1. **这个问题重要吗?**
> 2. **这个问题被广泛关注过吗?**
> 3. **这个问题是可以被解决的吗?**

根据这个框架，一个问题越重要，越容易解决，且越被广泛忽视，那么一个问题的优先级就越高。这个框架往往被用于解决一些公益的事项安排，而论文作者将这个框架利用一些数学与经济术语进行定义，从而借助它来评估 NLP 技术的研究优先级。

首先是重要程度，作者使用 $p(t;r)\cdot I(t)$ 来衡量一项技术预期的社会影响，其中 $p(t;r)$ 是研究者 r 研究技术 t 的成功概率。而 $I(t)$ 是指技术 t 的社会影响。**作者认为成功概率是衡量重要程度的重要一环，因为大量技术有可能并不会走向成功，即使其预期对社会可以产生极为正面的影响。**

其次是关注程度与解决问题的难易，**作者借用经济学中的边际效益衡量这两个维度**，定义：

$$\triangle I(t;r):=I(prog(t)+\triangle t(r))-I(prog(t))$$

其中，$\triangle I(t;r)$ 表示研究者 r 对技术 t 每多投入一单位资源所收获的边际效益，而 $I(t)$ 为技术 t 的社会影响，$prog(t)$ 为技术 t 当下的进展，$\triangle t(r)$ 代表研究者对技术 t 投入单位资源所能够带来的技术改进。

这个定义展现了，如果这个相关技术的领域已经饱和，那么对于一个研究者而言盲目地将资源与时间投入到这一问题的研究中是不明智的。而在另一个方面，如果这个领域有着很高的重要性却长期被研究者忽视，那么推动这个研究所产生的边际效益就会很大，**因而这部分解释了为什么研究人员热衷于创造一个崭新的研究领域进行研究。**

最后，**作者引入了机会成本这一概念描述不同技术间的选择成本。这一概念即试图表现研究者将资源用于技术 t 而非其他技术时所造成的潜在损失**，例如当我研究绿色 NLP 时，相当于我放弃了研究诸如对话系统与机器翻译的机会，作者定义：

$$Cost(t;r):=\triangle I(t^*(r);r)-\triangle I(t;r)$$

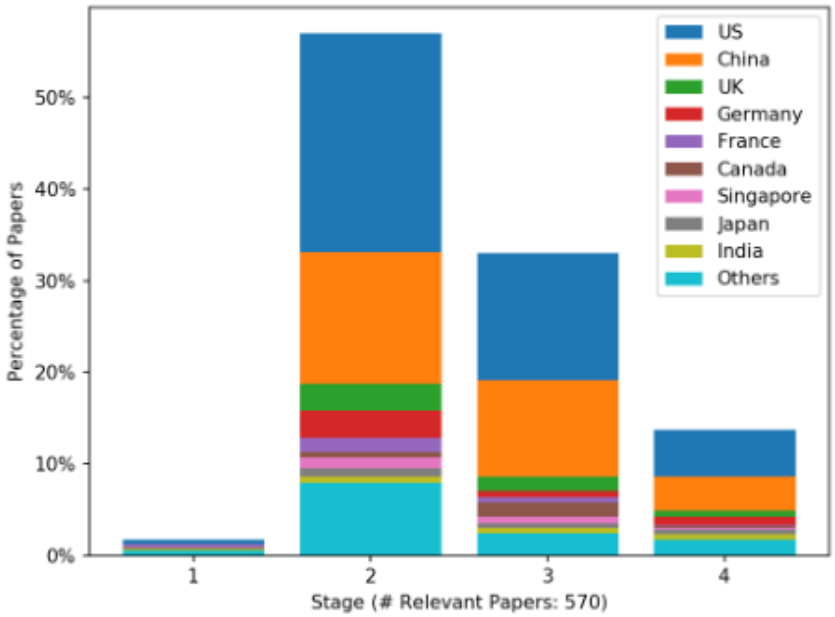
其中，$Cost(t;r)$ 表示研究者 r 在技术 t 上的机会成本，而$\triangle I(t^*(r);r)$代表了研究者的研究最优技术的边际收益，其中:

$$t^*(r):=\underset{x}{arg\,max}\triangle I(x(r))$$

其含义为研究者 r 可能的具有最大边际效益的替代技术。**因此，这个定义强调了并不一定只要做“好事”，而更应该去做“最好的事”，因为有时“好事”意味着极其高昂的机会成本。**

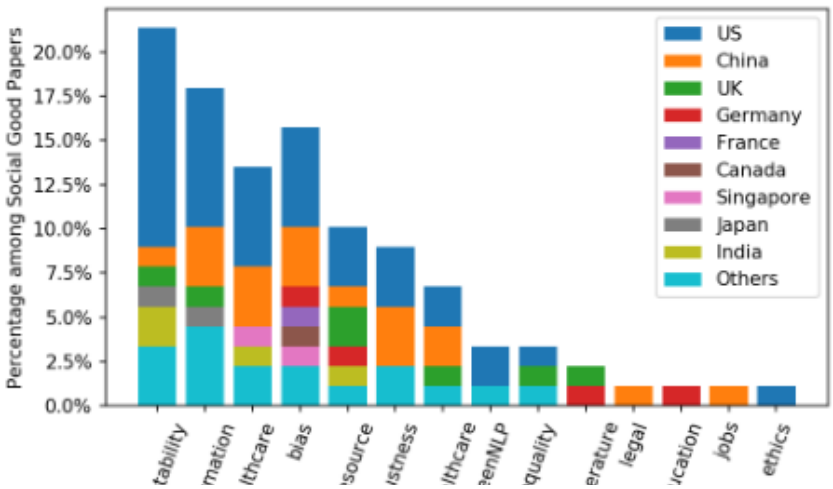
什么是最应该关注的领域?

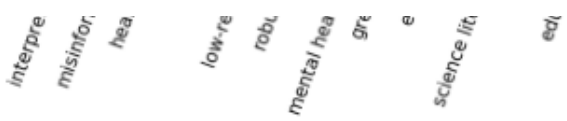
这篇论文主要以 **ACL 2020** 为分析对象，论文作者首先将 ACL 2020 的文章依据前文的四个阶段的分类方法进行了分类，并进行了国别的统计，如下图所示：



从发文数量而言，美国与中国是当仁不让的前两名，但是中国在第一阶段，也就是基础理论的论文发表挂了零蛋，考虑到论文的分类方法，第一阶段主要集中在语言理论之上，**而很可能我们针对语言理论的研究并没有以英文的形式发表在 ACL 上。**

在对论文进行分类之后，作者使用人工标注的方法，结合上文的评价框架，从 ACL 2020 的570篇文章中标出了 89 篇被评价为具有良好社会效益的文章，这 89 篇文章被分别分类为前文判断具有良好社会效益的 NLP 领域的标签内，如**减少偏见、提升教育水平、促进平等、消除欺诈、绿色环保、医疗保健、可解释性、法律应用、低资源学习、心理健康、鲁棒性等等。**





Social Good Topics at ACL 2020 (# Relevant Papers: 89)

从结果中我们可以看到，大多数论文致力于可解释性、错误信息消除与医疗保健领域。从国别角度来看，美国学者进行了大量关于模型可解释性的研究，而中国学者对于模型可解释性的研究很少。美国学者对提升教育水平与法律应用领域关注度较低，而印度学者则很少关注错误信息消除领域。

从ACL 2020发表的论文表现来看，自然语言处理技术并没有在提升社会效益这一异常宏大的主题上交出满意答卷。例如教育是联合国可持续发展目标中第四重要的领域，但 NLP 技术却很少有涉足这一领域。

事实上，这种现状的原因之一是 NLP 研究者得到的资助往往并不来自一些十分注重社会公益的机构，**甚至我们仔细想想，我们针对 NLP 这一研究领域的研究路径根本就没有经历老板给我们五百万让我们仔细思索应当先研究什么后研究什么这一阶段，更不论以一个优先级评定框架去评估不同领域的社会影响以及其重要性程度。**

那么我们该怎么做呢？论文作者认为我们在开展研究前应该先回答以下这五个问题：

> 1. **这项技术会使什么样的人从中受益?**
> 2. **这项技术可以帮助到哪些弱势群体?**
> 3. **这项技术是否有助于实现联合国可持续发展目标中的其一其二?**
> 4. **这项技术可以提升人们的生活质量吗?**
> 5. **这项技术会给人们带来哪些问题?**

例如，对于机器翻译而言，**机器翻译会使得其他语言的使用者受益（Q1），并且有助于缩小第二语言者与母语者的差距（Q2），机器翻译技术可以直接提升信息与知识的共享，可以被广泛的应用在优质教育、体面工作与全球伙伴之中（Q3），机器翻译可以被认为直接提升了人们的社会质量（Q4），但它有可能扩大有经济条件穿戴智能设备与没有经济条件的人群之间的差距（Q5）**，因此我们可以将其认为是对社会有益的。

基于这个自检框架，作者归纳出具有良好社会效益的 NLP 研究主题，并对每个主题提出了建议的研究方向如下，其中比如**以 NLP 技术应用于残疾学生教育，用 NLP 技术应用于表达障碍者的辅助语音生成，针对气候变化认知问题的跨文化研究**等等都十分具有现实意义。

Priority	Example NLP research topics
Poverty	• Predicting povery by geo-located Wikipedia articles (Sheehan et al., 2019) • Parsing fund applicant profiles (proposed)
Hunger	• NLP for agriculture (Yunpeng et al., 2019) • NLP for food allocation (proposed)
Health	• NLP to analyze clinical notes (Dernoncourt

& Well-being	et al., 2017a,b; Luo et al., 2018; Gopinath et al., 2020; Leiter et al., 2020a,b) • NLP for psychotherapy and counseling (Biester et al., 2020; Xu et al., 2020; Pérez-Rosas et al., 2019) • NLP for happiness (Asai et al., 2018; Evensen et al., 2019) • Assistive speech generation (proposed)
Education	• NLP for educational question answering (Atapattu et al., 2015; Lende and Raghuwanshi, 2016) • Improving textbooks (Agrawal et al., 2010) • Automated grading (Madnani and Cahill, 2018; Taghipour and Ng, 2016) • Plagiarism detection (Chong et al., 2010) • Tools for learners with disabilities (proposed)
Equality	• Interpretability (Köhn, 2015; Belinkov et al., 2017; Nie et al., 2020) • Ethics of NLP (Hovy and Spruit, 2016; Stanovsky et al., 2019; Sap et al., 2019) • NLP for low-resource languages (Zoph et al., 2016; Kim et al., 2017) • NLP on resource-limited devices (Sun et al., 2020) • NLP tools that signal bias in human language and speech (proposed)
Clean water	• Raising public awareness of water sanitation (proposed)
Clean energy	• Green NLP (Strubell et al., 2019; Schwartz et al., 2020) • NLP to analyze cultural values regarding climate change (Jiang et al., 2017; Koenecke and Feliu-Fabà, 2019) • Cross-cultural models of climate change perceptions (proposed)

总结

这篇论文为定性的分析 NLP 技术的社会效益开了一个好头，诚然如作者所说，**他们的工作目标并非是给予一个自然语言处理技术的社会效益的确定答案，而是在这个自然语言处理技术已经开始从科研领域进入大众生活的时间拐点处，试图向全面理解自然语言处理技术的社会意义迈出一小步。**

随着自然语言处理的研究与工业应用走向成熟，一种清醒的与高屋建瓴的对一项技术的社会意义的理解是必不可少的，我们的研究不能仅局限于什么领域好发，什么技术好做，而更是要**在动机层面意识到真正的“研究意义”**。恰如我们从学校到企业明白了什么才是这项技术的商业意义一样，研究也需要有一个过程去理解这项技术现有的与潜在的社会意义。

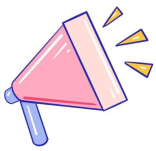
计算机科学是一个典型的应用学科，在刻板印象里程序员亦或是计算机领域的研究者往往不存在什么类似天下大同的理想与目标，整个领域往往被粗暴的定义为学了“能赚钱的”商品。我们可以看到学法者站在法律是社会效益最大的背景下伸张正义，学医者更是悬壶济世医者仁心，**探讨 NLP 的社会效益，也多少可以在平凡代码之余暗藏一些超验的意义与动力，用“我有一个梦想”式的浪漫，投身这个领域的平凡与灿烂。**

萌屋作者：小戏

边学语言学边学NLP~

作品推荐

1. 千呼万唤始出来——GPT-3终于开源！
2. Linux 程 序 员 失 业 警 告



后台回复关键词【**入群**】

加入卖萌屋NLP/IR/Rec与求职讨论群

后台回复关键词【**顶会**】

获取ACL、CIKM等各大顶会论文集！

喜欢此内容的人还喜欢

若被制裁，中国AI会雪崩吗?

夕小瑶的卖萌屋