自然语言处理“大拿” 瞄准ChatGPT

崂山区高新企业“自然语义”年内启动中国版本建设 内部代号是MOSS

AI赋能提升青岛城市发展硬实力。**青岛市人工智能产业协会供图**

近期，ChatGPT成为人们热议的话题，其实它的核心就是自然语言处理（Natural Language Processing），简称：NLP。在崂山区就有一家在这方面的专业公司——自然语义（青岛）科技有限公司。

2月16日，由该公司CTO（首席技术官）何博士主导开发的完全开源的项目HanLP在GitHub Star数达到了28K，超过了宾夕法尼亚大学、斯坦福大学、哈尔滨工业大学同类产品的总和。

该公司联合创始人孙雁群介绍，ChatGPT给他们敲响了警钟，他们也要在年内启动自然语言处理预训练大模型的研发，该产品内部代号MOSS，取自《流浪地球》中那台量子计算机的名字。MOSS不满足于是中国版本的ChatGPT，更试图在此基础上增加推理能力，使之不仅仅是一个可以对话的模型，更将具备初步的自我创造能力。

一次无心插柳的创新

“我与何博士相识，源于一次偶然机会。我想自己开发一个搜索引擎，搜索引擎需要对输入的文字进行分词操作，当时我们不懂。我就上网搜寻，结果发现一篇博客上写得特别好，不仅把如何实现的技术写得很清楚，还把原理以及作者学习、思考的全过程都描写得非常清楚，甚至作者走过的弯路也一并附带。同时我发现文章中附带的一段代码写得非常工整，如同教科书一般，逻辑严谨、清晰，一句废话都没有。”计算机专业毕业的孙雁群说，当时他揣测博主一定非常厉害可能是某个大厂的CTO（首席技术官），随后他便给博主的邮箱写了一封信。

“我们通过邮件聊了一个多月。博主说他在上海第二外国语大学，我一直以为他是学校的老师，结果没想到见面时才发现他是大二的学生，并且学习的是日语，编程是他业余爱好。”孙雁群说，那篇文章的专业程度达到了研究生的水平。因为喜欢打游戏，何博士在大学期间开始学习编程，后来老师给他介绍了一个兼职工作，这让他与自然语言处理（NLP）结下了不解之缘，后者成为他日后攻读博士期间的研究方向。

当时，那家公司立项做一个“智能检索系统”，中文搜索引擎的第一步是分词，公司安排何博士来做分词器。当时的开源工具分词的效果比较慢，并且很多人名、地名、单位名称都分不清楚，于是，何博士决定做出体验效果好的分词器。在翻阅了国内众多NLP领域专家的论文后，经过半年的时间，何博士做出了一个分词器，并取名“HanLP”。这个“Han”就是取自汉语的意思。

“我们经过交流后，都认为好东西就应该开源。永远开源，也成了我们的君子协定，此后何博士研发的软件主体部分永久开源。”孙雁群说，2014年，这款HanLP软件发到了全球最大的开源网站GitHub后，当天就有10多个用户收藏，一个月后，加星就超过了哈工大的同类产品，2017年10月26日，超过了斯坦福大学的同类产品，两个月后，又超过了宾夕法尼亚大学同类产品。后来，HanLP更是成了全球程序员开发NLP项目的首选，许多知名科技企业的技术人员都在使用这个技术，每年也有许多高校科研机构把HanLP当做科研工具。

刚开始第一版的HanLP功能比较简单，经过几代更迭后，功能越来越完善，性能更高效。HanLP能提供词法分析、句法分析、文本分类、情感分析、词向量、自动摘要等功能。HanLP还具有精度高、速度快、内存省的特点。

文字工作者10年或被AI替代

“伴随着我们的产品注册量越来越多，网上甚至有很多粉丝发布了使用攻略，为了更好地服务用户，2019年11月，我们在崂山区成立了自然语义（青岛）科技有限公司。公司成立后，崂山区给我们提供了办公场所、融资等多方面的支持，公司实现了快速发展。”孙雁群说。

很快，HanLP项目在全球范围内拥有数百万的程序开发者用户，是GitHub上全球用户数量最多的自然语言处理技术。目前开源用户包括百度、小米、京东、华为、字节跳动等顶级头部企业，以及MIT、中科院、北京大学、复旦大学等科研机构。“因为拥有了一些技术储备，早在2021年，我跟何博士讨论过，公司是不是搞一个预训练的大模型。ChatGPT其实也是一个预训练的大模型。我们把很多文本都交给神经网络模型，然后用庞大的数据训练它，它就会变得智慧化程度非常高。后来，我们经过测算，需要投入的财力物力极其庞大。即使一个初代版本，就要投入100人进行语料标注，就是有人把文本处理好，还要大概投入1.6亿元购买设备。当时，因为财力物力的原因，我把作预训练大模型的计划暂时推迟了，继续专攻具体的NLP算法。”孙雁群说。

前期，自然语义公司给国内某大型企业专门定制开发了一套秘书辅助办公系统，在学习了大量的演讲稿、发言稿等材料后，形成了固定模式，工作人员输入了几个关键词，就能够快速生成段落，经过工作人员的修改后，就能变成较为成熟的发言稿。“现在生成了的文章只是几个段落模式的文本，因为还是小模型，要是制作成大模型，经过大量的文本学习后，就可以生成很成熟的文本了。这都是自然语言处理的核心技术。”

“按照目前的人工智能技术发展速度，距离淘汰我们这些文字工作者还有多远？”记者听到孙雁群介绍了这套秘书辅助系统后，提出了这个话题。孙雁群思索后回答，按照目前的技术进步程度，在10年左右就会出现这个情况。

“自然语言处理，为什么它非常强大呢？”孙雁群说，“很多年前，比尔·盖茨就曾经说过，自然语言处理是人工智能这个皇冠上的明珠。为什么他把自然语言单独拎出来？你们可以想想，人区别动物的一个主要特征就是人有语言系统。语言系统代表着抽象能力，就是让我们可以在最短的时间内交流更多的信息，其实语言就是一种信息编码。我传达给你的是信息，你只是大脑中有这种处理信息的能力。那么对计算机来讲，如果计算机能处理语言、理解语言，这就相当于计算机有了智慧。比如说图像识别，它只能相对是眼睛、声音识别，相对是耳朵，但是语言处理是真正的相对于大脑。现在人工智能发展到我只要输入一篇文章，计算机就能根据这篇文章去画一幅画。我们在学会语言之前，是不会思考的，我们的思考实际上都是在大脑中用语言来处理的。比如说，你想明天带着孩子去海边玩，在大脑中，你是首先生成了这句话，而不是想到一幅画、或者是一段视频。这就涉及宇宙中的一个普世规律，宇宙中的任何行为，都是以最节约能量的方式进行。相比较处理视频、图像，语言这种信息编码方式可能是最高效的处理方式，会节省更多的能量。这也是计算机研究相对前沿的东西。这就是为什么ChatGPT出来之后，大家非常震撼的一个主要原因。”

今年启动中国版本研发

“目前，我们的HanLP正在继续迭代升级，功能又增加了很多细分项。在ChatGPT出来后，不少公司主动找我们，希望与我们合作研发。”孙雁群说，此前，他跟何博士围绕着ChatGPT作了一次沟通。

何博士认为，ChatGPT的爆火，让他想起当年谷歌的word2vec，街头巷尾都在讨论。现在时代已经变了，语言模型已经不再是一台服务器+1G 的文本就能训练出来的了。ChatGPT把当前已有的技术规模化，投入了大量资金和人力，才有目前的效果，其实ChatGPT的核心技术并不困难，中国可能很快就有类似产品出现。“ChatGPT可以说是目前最接近人们对人工智能期待的产品。我们遇到那种需要写套话的邮件之类，也会让它写。但也必须冷静，ChatGPT只有语言知识，没有推理能力。它对语言建模，但它对整个世界的运行原理完全没有建模。如果想做出超越ChatGPT的AI产品，需要在语言理解与生成的组件中间再加一个推理引擎，合起来组成一个具有推理能力的AI。”

“尽管研发的过程会比较漫长，我们要尽快启动中国版本的研发。”2月16日，孙雁群告诉早报记者，前两天，他们开了股东专题会议，批准了研发计划，研发过程要分三步走。首先要先启动语言模型的研发，做好语料的储备。再就是融资购买相关GPU设备，训练出第一版的语言模型。然后开展推理模型架构，赋予它世界观，也就是人格，变成带有人格特征的模型。

本版撰稿
观海新闻/青岛早报记者 陈勇
通讯员 李锡敏