数字人文与语言研究*

冯志伟　丁晓梅

提要　数字人文是使用计算机技术和网络技术来研究传统的人文科学的一门新型的交叉学科,本文介绍数字人文研究的历史,数字人文研究的主要内容,数字人文与人工智能,并讨论数字人文与语言研究的关系。

关键词　数字人文　机器索引　机器翻译　古籍数字化

数字人文(digital humanities),又叫作人文计算(humanities computing)或人文科学中的计算(computing in the humanities),它是使用计算机技术、数字技术和人工智能来研究传统的人文科学的一门新型的交叉学科,它的产生与发展得益于计算机科学、数字技术和人工智能的进步及其在科学领域的普及应用。

在计算机技术、数字技术和人工智能的支持下,人文知识的获取、分析、集成和展示都出现了重大变化。目前,已有海量的图书、报纸、期刊、照片、绘本、乐曲、视频等人文资料被数字化,并在网络上提供给大众使用,OpenAI 公司研制的 Sora 甚至能够从文本或图像自动地生成栩栩如生的视频,震惊了全世界。面对这种日益强化的数字化情景,人文学者应当进行更新知识的再学习,掌握有关的数字化技术,对这些数字化的人文资料进行组织、标引、检索和利用,从而提高人文科学研究的效率,实现文理融通,进一步推动人文科学的发展。(冯志伟,张灯柯 2023a)

一　数字人文研究的历史

数字人文的研究早在 20 世纪 40 年代就开始了,这样的研究是首先从语言研

* 本文为国家社会科学基金项目"基于平行语料库的俄汉语言学术语词典编纂研究"(项目编号 17BYY220)阶段性成果。

究开始的,这样的研究起初叫作“人文计算”。

语言是人类区别于动物的主要标志,是一种最为重要的人文现象,因而也就成为了人文计算的首要研究目标。人文计算的研究是首先从语言研究开始的。

在人文科学的研究中,文本资料的引得(indexing,又叫“索引、通检”)一直是使用手工编写卡片的方式来进行的,这种手工编写卡片的工作费时费力,非常辛苦,是相当枯燥的脑力劳动,同时又是繁重的体力劳动。

在新中国成立以前,我国的燕京大学曾设有引得编纂处,以哈佛燕京学社名义编写古代汉籍引得 63 种、中法汉学研究所编写通检 8 种,又以巴黎大学北京汉学研究所的名义编写通检 6 种,中华书局、商务印书馆也编写过一些索引。所有的这些索引都是通过手工劳动来编纂的,这样的编纂工作枯燥乏味,手续繁多,要经过比较版本、校刊文字、确定工作本、划定词目、摘抄例句、排列卡片、过录稿本等诸多工序。编辑《荀子引得》时,“编辑五六人,晨夕不辍,历时年余,始克蔵事……编者诚劳而用者则安逸矣”(冯志伟 1992:31)。《杜诗引得》的编纂,费时几近两年。在国外,外文典籍的索引编纂工作,同样也是非常艰苦的脑力和体力劳动。

1946 年,计算机研制成功之后,人文学者们开始使用计算机来编制索引,这样就可以把人文学者们从手工索引的繁重劳动中解放出来。数字人文从此开始启航。

1949 年,意大利耶稣教神父罗伯托·布萨(Roberto Busa)在国际商用机器公司 IBM 的帮助下,用 IBM 计算机成功地为著名神学家托马斯·阿奎那(St Thomas Aquinas)和相关人员多达 1 100 多万字的拉丁文作品编制机器索引(machine indexing),并且使用计算机对每一个单词进行了词目还原(lemmatization),把文本中的拉丁文变形词还原成原形词,以便计算机按照原形词来做索引。罗伯托·布萨的成功,使得计算机在语言学领域的应用逐渐风行起来。1963 年,英国学者维斯贝(Wisbey)使用计算机为中世纪的高地德语(Hoch Deutsch)文献做了机器索引。我国武汉大学也采用计算机来给《骆驼祥子》《倪焕之》等现代文学作品编制机器索引。此后,这种机器索引的方法从语言研究逐步向文学、历史学、文艺学等领域扩展,罗伯托·布萨(Busa 1980)因此提出了“人文计算(humanities computing)”的概念。

与此同时,在 1949 年,美国洛克菲勒基金会自然科学部主任韦弗(W. Weaver)发表了一份以《翻译》(Translation)为题的备忘录,提出了用计算机进行机器翻译

(machine translation)的设想。在这份备忘录中,韦弗认为翻译类似于解读密码的过程。他说:"当我阅读一篇用汉语写的文章的时候,我可以说,这篇文章实际上是用英语写的,只不过它是用另外一种奇怪的符号编了码而已,当我在阅读时,我是在进行解码。"[①](Hutchins & Sommers 1992)

这段话中,韦弗别开生面地提出了用解读密码的方法进行机器翻译的想法,这在当时是一个独创,引起了人们的兴趣和关注。韦弗的备忘录中还记载了一个有趣的故事。美国布朗大学数学系的教师吉尔曼(R. E. Gilmam)从事密码的解读研究。他曾经解读了一篇长约一百个词的土耳其文密码,但是,他既不懂土耳其文,也不知道这篇密码是用土耳其文写的。韦弗指出,吉尔曼的成功足以证明解读密码的技巧和能力是不受语言影响的,因而可以用解读密码的办法来进行机器翻译。由于计算机可以通过计算技术来解读密码,因此,像语言翻译这样的人文现象可以使用计算机来实现。这样的机器翻译工作也就是人文计算的早期研究领域之一。

1954 年,美国乔治城大学在 IBM104 电子管计算机上进行了第一次机器翻译试验,他们把 60 个用拉丁字母转写的俄语句子自动地翻译成英语,并公开进行了演示,用事实证明了机器翻译的可能性。第一个机器翻译系统原型的公开演示,极大地震动了新闻界和学术界。《纽约时报》(*New York Times*)为此专门做了报道。第一次机器翻译试验的成功,用强有力的事实说明,使用计算机来处理诸如语言翻译这样的人文现象是可能的。

亲属关系(kinsfolk relation)是一种重要的社会人文现象。1963 年,林德赛(R. Lindsay)在美国卡内基技术学院用 IPL - V 表处理语言(list processing language)设计了 SAD - SAM 系统,采用了特定格式来进行关于亲属关系方面的人机对话(man-machine dialogue),系统内建立了一个关于亲属关系的数据库,可以接收关于亲属关系方面的问题的英语句子提问,用英语做出回答。

这个系统分为两个模块:SAD 模块和 SAM 模块。SAD 模块的功能是做句法分析,它接收输入的英语句子,从左到右进行分析,建立起这个英语句子的推导树,然后,把这个能表示该英语句子结构的推导树传给 SAM。SAM 模块的功能是做语义分析并做出回答。SAM 模块首先从语义的角度抽取有关亲属关系的信息,建立起亲属

① 此处"用汉语写的文章"在有的引文中为"用俄语写的文章",经查对原文,"用俄语"应更正为"用汉语"。

关系树,然后根据数据库中存储的信息找出问题的答案。这是早期的人机对话(man-machine dialogue)研究工作,其目的在于处理亲属关系这种人文现象,也应当属于人文计算的范畴。

因此,我们认为,使用计算机来研究人文现象,特别是研究像语言或亲属关系这样的人文现象,早在 20 世纪 40 至 50 年代就开始了。机器索引、机器翻译和人机对话都可以看作是最早的人文计算研究。这是人文计算研究的萌芽。这个阶段的特点是研究数据主要是文本,数据的规模都比较小,可以叫作"小规模的文本数据处理(small-scale text data processing)"阶段。

1973 年,欧洲的一些语言学者建立了文学与语言学计算协会(The Association of Literary and Linguistic Computing,简称 ALLC),明确提倡采用计算机技术来进行文学和语言学的研究。ALLC 于 1986 年出版了《文学与语言学计算》(*Literary and Linguistic Computing*)期刊,成为人文计算的专业刊物。此后,人文计算研究的队伍日益壮大。

20 世纪 90 年代初期,人文计算逐渐成为一个独立的交叉学科。1999 年,英国学者伦敦国王学院麦卡蒂(McCarty)教授主张人文计算应当作为一门交叉学科,有关部门应当在制度和学术层面为人文计算提供切实可行的保障和支撑。

1993 年 7 月在日本神户召开了第四届机器翻译高层会议(MT Summit Ⅳ),在这次会议上,英国著名学者哈钦斯(J. Hutchins)在他的特约报告中指出,自 1989 年以来,机器翻译的发展进入了一个新纪元。这个新纪元的重要标志是,在基于规则的技术中引入了语料库(corpus)方法,其中包括统计方法,基于实例的方法,通过语料加工手段使语料库转化为语言知识库的方法,等等。这种建立在大规模真实文本处理(large scale and authentic text processing)基础上的机器翻译,将会把自然语言处理(natural language processing,简称 NLP)推向一个崭新的阶段,是机器翻译研究史上的一场革命。语料库方法渗透到了机器翻译研究的各个方面,一些基于语料库的机器翻译系统如雨后春笋般地建立起来,有的系统把基于语料库的方法和基于规则的方法巧妙地结合起来,取得了可喜的成绩。(冯志伟 2015)

这样,人文计算的对象就从小规模的文本转到了大规模的、真实的文本,从大规模的、真实的文本语料库中获取知识,这是人文计算研究中具有标志性意义的战略转移。

蒂姆·伯讷斯·李(Tim Berners-Lee)创立的万维网(World Wide Web,简称

WWW)是基于互联网(Web)的计算机网络,用户使用 WWW,可以访问存贮在世界范围内的互联网上的海量信息。WWW 是根据“客户端—服务器”(client-server)的模式来进行工作的。客户通过叫作“客户端”(client)的程序与远程存贮着数据的“服务器”(server)连接,互联网通过叫作“浏览器”(browser)的客户端(client)程序来进行浏览,例如 Navigator, Internet Explorer 等都是著名的客户端浏览器。互联网浏览器把用户的提问传送给远程的服务器搜索有关的信息,然后返回搜索到的文件,这些文件使用超文本标记语言(hyper text makeup language,简称 HTML)书写,最后在客户端用户的计算机屏幕上显示出来。

互联网的操作依赖于超文本(hypertext)文件的结构。超文本可以让网页的作者把他们的文件与互联网上的其他文件进行超链接(hyperlink),从而可以看到互联网上的有关文件。

WWW 以及马赛克(Mosaic)浏览器的出现,是互联网发展历史上两个最重要的事件,它们使得互联网能够迅速地在用户中得到推广和普及。互联网上的数据是海量的、大规模的,数据的载体除了叙事文本之外,还可以是格式化的字母数字集、图像、声音、音乐、视频等多媒体。

20 世纪 90 年代早期至 21 世纪初期,随着互联网的出现和计算机技术的发展,“人文计算”的对象从电子文本逐步扩展到超文本、图像、视频、音频、数字地图、网页、虚拟现实、3D(3 维图形)等多媒体(multimedia),计算的领域从语言学领域进一步扩展到历史、音乐、艺术等多个领域。这种发展使得继续使用“人文计算”这个概念显得有些过时。这样一来,人文计算就进入了它的第二个阶段:“大规模的多媒体数据处理”(large-scale multimedia data processing)阶段。

2001 年 4 月布莱克维尔出版社(Blackwell Publishing)首次出版了一部以“数字人文”为书名的图书《数字人文指南》(*A Companion to Digital Humanities*)。“数字人文”这个术语迅速取代了“人文计算”,成为一个在西方广泛传播的新兴跨学科研究领域的代名词。这样的多媒体数据都是数字数据,因此,在 2001 年,人文计算也就有了一个新的名称:数字人文(Schreibman *et al*. 2004)。

在这种情况下,1973 年成立的文学与语言学计算协会也相应地改名为欧洲数字人文协会(European Association for Digital Humanities,简称 EADH)。

20 世纪 90 年代后期,人文学者开始越来越频繁地使用各种新型的数字技术处理人文资料,并进行人文知识生产(humanities knowledge production)。例如,使用

计算机扫描古籍图书和绘本,使用虚拟现实(virtual reality,简称 VR)技术复原古建筑模型和历史上的都市,开发和建设各种在线的声视频数据库(audio video database)等。人文知识的可计算性快速从文学和语言学领域向哲学、历史、音乐、表演艺术等多个人文领域扩展。在这一进程中,产生了越来越多的数字原生数据,如数字地图、计算机图像、在线网页、虚拟人物等,这些原生数据的产生极大地丰富了人文研究的对象。

现在,我们已经进入了大数据时代(big data era)。大数据这个概念是 1997 年由美国国家航天局研究院的科克斯(M. Cox)和埃斯沃斯(D. Ellsworth)提出的,他们认为大数据是可以进行可视化研究的、数量巨大的科学数据。美国 IBM 公司认为大数据具有三个特性:Volume(大规模),Variety(多样性),Velocity(高速度),被称为“三 V”。在数字人文的研究中,我们也需要具有“三 V”属性的大数据。

早期的人文科学研究的数据来源主要是人文科学研究者个人的“内省(introspection)”,依靠的是人文科学研究者个人的学识。尽管有的研究者满腹经纶、学富五车,这样的数据也只能算小数据。在浩瀚无边的数据海洋中,仅仅依靠研究者个人的小量的数据就试图决定人文科学研究结论的真伪,很容易出现主观性和片面性,显然是不科学的。

早期研究的数据还有一个来源就是“诱导(elicitation)”,也就是从书本、词典等第二手材料中或者通过问卷调查等方式,诱导出有关的数据,并进一步从诱导出的数据中推导出结论。这样的“诱导”方式依靠的数据当然也只是小数据,也难免有片面或不完善的地方。

我们主张依靠大数据,从大规模的真实文本数据中获取知识,这样的大数据除了具备大规模(large scale)的特点之外,还具有真实性(authentic)的特点,它们都是客观存在的数据,不带有主观性,它们的规模大,可以避免片面性。

进入大数据时代之后,我们获取大数据的方式不再是“内省”或“诱导”,而是“观察(observation)”和“检验(verification)”。使用这样的大数据来研究人文现象,使我们有可能通过“观察”和“检验”从大数据中获得客观的知识,这样就大大地避免了主观性和片面性。所以,大数据驱动的人文科学的创新研究是我们认识和研究人文科学方法的重大改变,在方法论上具有重要的意义。

2007 年以来,采用深度学习(deep learning)的方法,以大规模的双语对齐的语料库作为语言知识的来源,在自然语言处理中引进了神经网络(neural network)技

术。例如,在机器翻译中,采用多层神经网络,从双语对齐的语料库中获取语言知识,采用词向量(word vector)来表示单词,形成句子的嵌入(embedding)表示,这样的方法大大地提高了自然语言处理的研究水平。(冯志伟 2019)

神经机器翻译的语言模型是端到端(end to end)的语言模型,源语言输入后,由编码器(encoder)进行编码处理,然后由解码器(decoder)输出翻译结果,完全不需要中间环节,如图 1 所示。

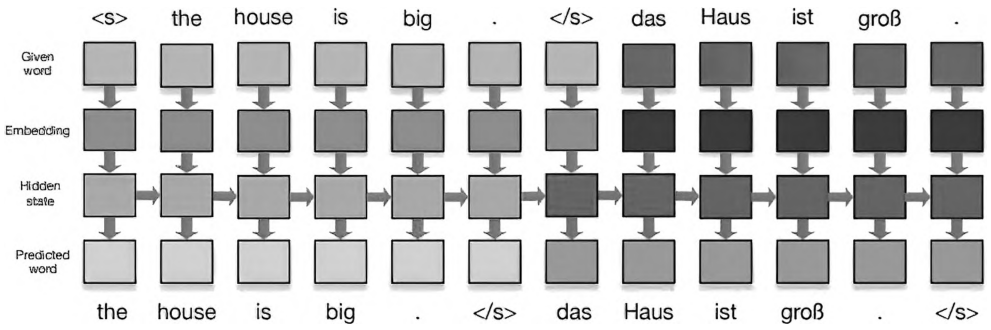

图 1　端到端的神经机器翻译

图 1 是英德机器翻译编码器—解码器模型的一个实例。在编码器一端,句首标志为〈s〉,源语言句子“the house is big”的输入单词(given word)在编码器(encoder)一端顺次进行编码,通过词嵌入(embedding)形成上下文向量,进入隐藏状态(hidden state),预测下一个单词(predicted word),句末的标志为〈/s〉,形成上下文向量输入解码器(decoder)。处理了句末标志〈/s〉之后,在解码器一端的第一个深绿色方框处,已经包含了源语言句子“the house is big”的所有词嵌入信息,在逐一地输出每一个目标语单词时,都要参照源语言英语句子的词嵌入信息以及目标语言已经生成的前面德语单词的信息,最后输出德语的语言序列“das Haus ist groβ”。

这种端到端(end-to-end)的机器翻译,显著地提升了机器翻译的效率。这样的机器翻译不再依赖从数据中挖掘的带有噪声的翻译对应关系,也不再进行语言基本单元的组合,而是通过对一系列向量表示的数值运算来完成整个翻译过程。这种端到端的建模方式充分地发挥了图形处理单元(graph processing units,简称 GPU)等设备带来的计算能力上的飞跃,能够有效地发掘语言数据中隐含的语言规律,从而获得了别开生面的翻译效果。

2001 年 4 月,国际文学与语言学计算联合会(Association for Literary and

Linguistic Computing)主席扎波里(A. Zampolli)召集多个学科的研究者在意大利的比萨(Pisa)召开了一次综合性的人文计算研讨会,会后发布了“比萨报告”(Pisa Report)。这是数字人文研究的一个重要文献。

在这个“比萨报告”中,罗贝(D. Robey)教授绘制并发表了一幅有关“数字人文”的全景知识图。

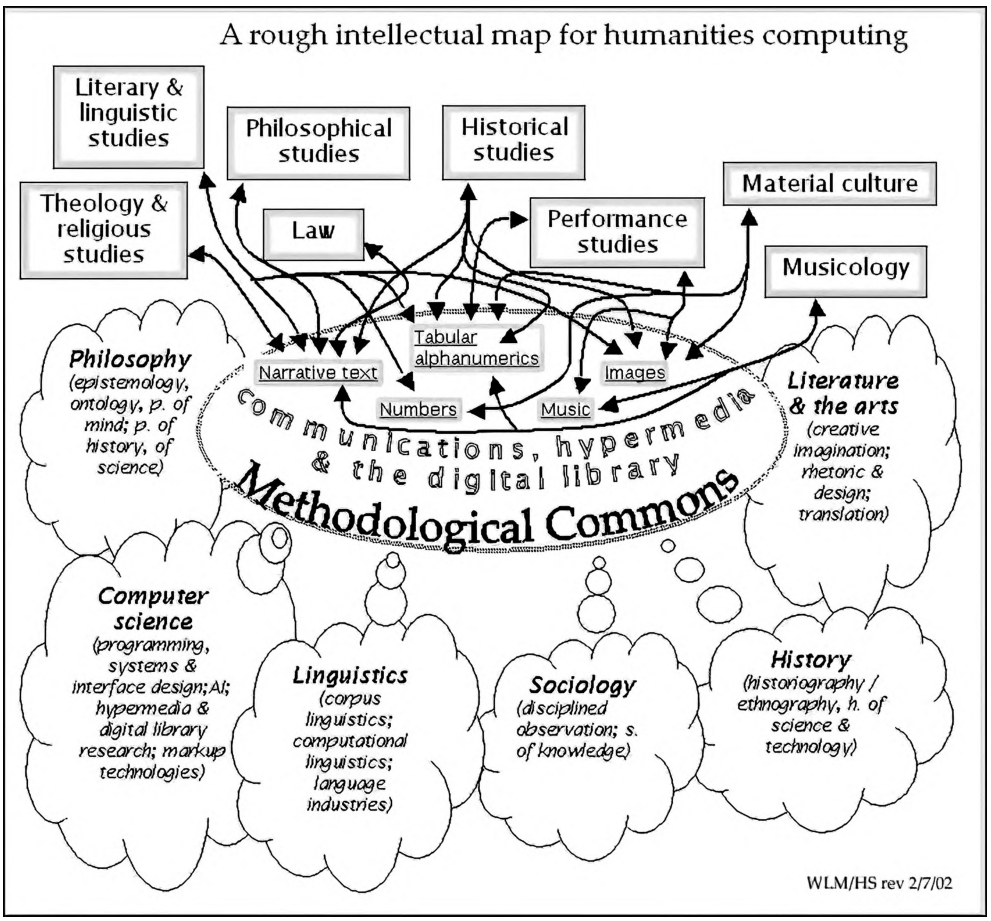

图 2　数字人文的全景知识图

“比萨报告”中这个全景知识图的中央区域,指的是数字人文研究的方法论共同基础,它们是数字人文研究的核心,包括各种可计算的基础数据对象,如叙事文本(narrative text)、格式化的字母数字集(tabular alphanumerics)、图像(images)、音乐(music)、数字(numbers)等,这些基础数据处理的最为关键的技

术是现代通信技术(communications)、超媒体(hypermedia)、数字图书馆(digital library)等基础性的研究平台。针对这些数据而进行的计算活动包括文本分析(text analysis)、数据库设计(database design)、数字绘图(numerical image)、音乐检索(music retrieval)等。

这个全景知识图上部分的方块代表的是各种科学共同体,如文学与语言学研究共同体(literary & linguistic studies)、哲学研究共同体(philosophical studies)、历史研究共同体(historical studies)、神学宗教研究共同体(theology & religious studies)、法学研究共同体(law)、表演艺术研究共同体(performance studies)、物质文化共同体(material culture)、音乐学共同体(musicology)等。双向箭头代表不同的科学共同体与不同的数据类型和计算方法之间的对应关系。例如,哲学研究共同体主要与叙事文本发生联系,文学与语言学研究共同体除了主要与叙事文本发生联系之外,还可能与图像发生联系,而这种对应关系也不是固定的,它们随着研究条件和信息技术的发展而不断变化。

这个全景知识图下半部分的云朵,表示数字人文涉及的不同的学科及其子学科研究的具体内容。例如:数字人文的哲学(philosophy)研究的具体内容包括认知论(epistemology)、本体知识体系(ontology)、心智哲学(philosophy of mind)、历史哲学(philosophy of history)、科学哲学(philosophy of science)。数字人文的语言学(linguistics)研究的具体内容包括语料库语言学(corpus linguistics)、计算语言学(computational linguistics)、语言产业(language industries)等。

从这个数字人文的全景知识图可以看出,数字人文的研究已经扩充到了人文科学的各个领域。其中除了语言学之外,还包括文学、哲学、历史、宗教、法学、表演艺术、物质文化、音乐学等人文科学领域。(冯志伟 2023)

目前数字人文研究的学术共同体已经形成。各种数字人文研究学会和专门的研究中心遍布全球,很多数字人文研究项目和研究成果也已经获得政府和学界的资助与关注。

数字人文领域影响力最大的学术团体是文学与语言学计算协会、人文领域计算机应用联合会(The Association for Computers in the Humanities)和数字人文学会(The Society for Digital Humanities)。这些学会还联合组成了世界上最大的数字人文联盟组织——国际数字人文组织联盟(The Alliance of Digital Humanities Organizations)。国际数字人文组织联盟每年召开一次数字人文年度大会。

目前,国际数字人文组织联盟出版了五本同行评审期刊,向全球传播数字人文研究的理念、方法和成果。这五本期刊是:

《文学与语言学计算》(*Literary and Linguistic Computing*),牛津大学出版。

《文本技术》(*Text Technology*),电子期刊,加拿大麦克马斯特大学出版。

《人文计算研究论文》(*Computers in the Humanities Working Papers*),在线预出版物。

《数字人文季刊》(*Digital Humanities Quarterly*),国际数字人文组织联盟专业电子刊物。

《数字人文指南》(*Companion to Digital Humanities*),布莱克维尔出版社出版。

我国清华大学也出版了《数字人文》杂志。

二　数字人文与人工智能

近来,在数字人文研究中引入了人工智能(artificial intelligence,简称 AI)中的大语言模型(large language model,简称 LLM)的技术。大语言模型是一种由包含数百亿以上参数的深度神经网络构建的语言模型。大语言模型通常使用自监督学习(self-supervised learning)方法通过大量无标注文本进行训练。(冯志伟,张灯柯 2023b)

2017 年,谷歌公司的瓦斯瓦尼(Vaswani)等 8 人发表了《注意力就是你所需要的一切》(Attention is All You Need)的论文,提出了 Transformer 模型。在机器翻译任务上取得了突破性进展,揭开了大语言模型研究的序幕。

大语言模型的基本原理,就是根据前面的词元(token),预测后面的词元(next token prediction)。

Transformer 主要有 5 个部分:

(1) 词元化(tokenization):词元化是最基本的步骤。它涵盖了一个庞大的词元库,包括所有的单词、标点符号等。词元化这个步骤要处理每一个单词、前缀、后缀以及标点符号,并将它们转换为词元库中已知的词元。例如,如果我们要计算机做英汉机器翻译,我们可以输入英语句子:"Hello, everyone!"计算机把这个英语句子中的单词和标点符号转化为词元:["Hello", ",", "everyone", "!"]。表示如下:

"Hello, everyone!" → ["Hello", ",", "everyone", "!"]

(2) 嵌入映射(embedding):一旦输入的文本被转化为词元之后,就需要将词元转换成机器更容易处理的词向量(word vector),词向量用实数(real number)表示。为此,我们使用嵌入映射技术,实现文本到词向量的转换。文本嵌入将文本中的符号转换为向量空间(vector space)中的向量实数值。表示如下:

["Hello", ",", "everyone", "!"] → [向量 Hello,向量,,向量 everyone,向量!]

如果两个文本片段相似,则其对应的词向量中的实数值也相似,如果两个文本片段不同,则其对应词向量中的实数值也不同。

向量空间中,意义相近的单词位置彼此靠近,例如,在图 3 所示向量空间中的 developing,growing 等单词聚集在相近位置,这是因为它们在实际语言中所处的上下文相似;having、had、have 等单词聚集另一个相近的位置,这是因为它们的语法功能相近;而意义与它们不同的 right、left 等单词则聚集在另一个位置,这是因为它们在实际语言中所处的上下文与 developing、growing 以及 having、had、have 等单词不同。图 3 是计算机自动生成的向量空间。

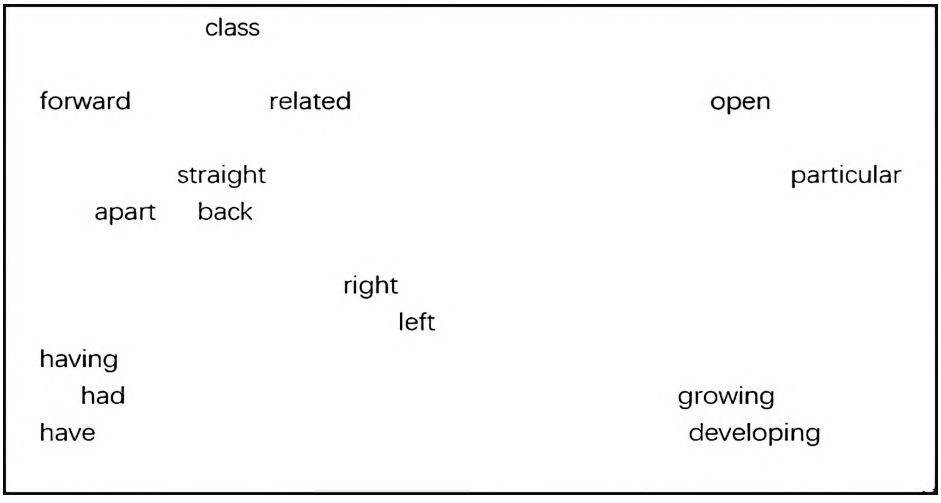

图 3　在向量空间中,意义相近的单词的位置彼此靠近

根据向量语义学(vector semantics),Transformer 不需要对于离散的语言符号进行计算,而只要把离散的语言符号转换为词向量嵌入到向量空间中进行计算,整个计算是针对没有语言符号的实数值进行的,由于把离散的单词符号都映射为向量空间中的词向量,计算机能够自动地从语言数据中获取到语言特征,并对语言特征

进行计算。

(3) 位置编码(positional encoding):位置编码为每个词元添加一个位置向量,以便跟踪词元的位置。在编码器的"输入嵌入"(input embedding)和解码器的"输出嵌入"(output embedding)时,都进行位置编码,使用三角正弦(sin)与余弦(cos)来计算位置,公式如下:

$$PE_{(pos,2i)} = sin(pos \ \ 10\,000^{2i/d_{model}})$$

$$PE_{(pos,2i+1)} = cos(pos \ \ 10\,000^{2i/d_{model}})$$

公式中使用了正弦三角函数 sin 和余弦三角函数 cos,pos 表示单词的位置,i 表示维度,PE 表示位置编码(positional encoding, PE),d_{model} 表示模型的维(dimensionality of model),在位置编码时,这样的三角函数 sin 和 cos 是可以通过线性关系互相表达的。这样的位置信息是非常重要的,特别是功能词的位置信息承载了语言的句法语义信息。

位置编码给每个向量加上位置信息,保留序列中词的顺序。表示如下:

[向量 Hello, 向量,, 向量 everyone, 向量!] → [向量 Hello_pos, 向量,_pos, 向量 everyone_pos, 向量! _pos]

(4) 自注意力机制(self-attention):

模型计算每个词对序列中其他词的"注意力",从而调整每个词的表示,使其包含更丰富的上下文信息(context)。表示如下:

[向量 Hello_pos, 向量,_pos, 向量 world_pos, 向量! _pos] → [向量 Hello_pos_context,向量, _pos_context, 向量 everyone_pos_context, 向量! _pos_context]

(5) 前馈网络(feedforward network)处理:

利用前馈网络对于每个词的向量进行进一步的非线性变换,以学习更复杂的表示。表示如下:

[向量 Hello_pos_context, 向量,_pos_context, 向量 everyone_pos_context, 向量! _pos_context] → [向量 Hello_final, 向量,_final, 向量 everyone_final, 向量! _final]

(6) 生成预测并"解码":

利用基于最终的向量表示,模型生成下一个词的预测,并使用将其转换(翻译)成人类可读的文本。表示如下:

[向量 Hello_final，向量,_final，向量 everyone_final，向量！_final] → 预测下一个 Token → [“哈罗”,“,”,“大家好”,“!”]

这样便得到翻译的结果:“哈罗,大家好!”

从以上步骤可以看出,ChatGPT 技术原理的起点是将自然语言词元化,也就是给大语言模型提供了一个可计算可理解的“基本粒子”(basic particle),然后用这些“基本粒子”去组合文本语言。

通过并行地处理所有的单词,并且让每一个单词在多个处理步骤中都注意到句子中的其他的单词,Transformer 模型的训练速度快,自然语言处理效果好。

Transformer 的结构如图 4 所示。

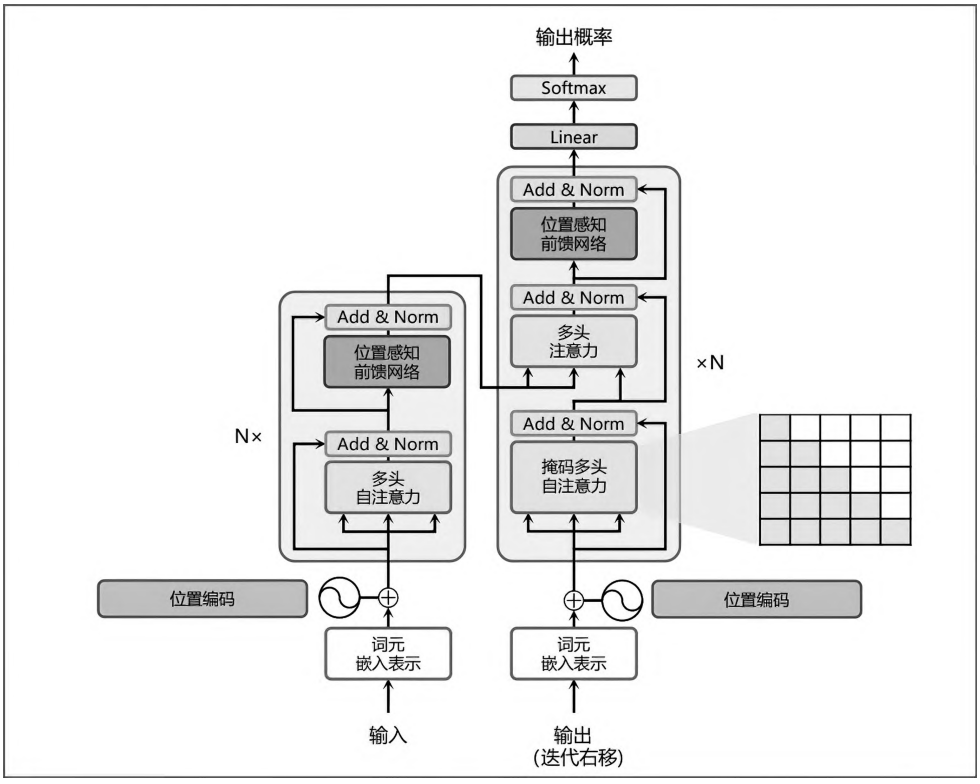

图 4　Transformer 的组成

Transformer 不需要循环,是并行地处理序列中所有的单词或符号,同时使用“自注意力层”(self-attention layer),把上下文与比较远的单词结合起来。如图 5 所示。

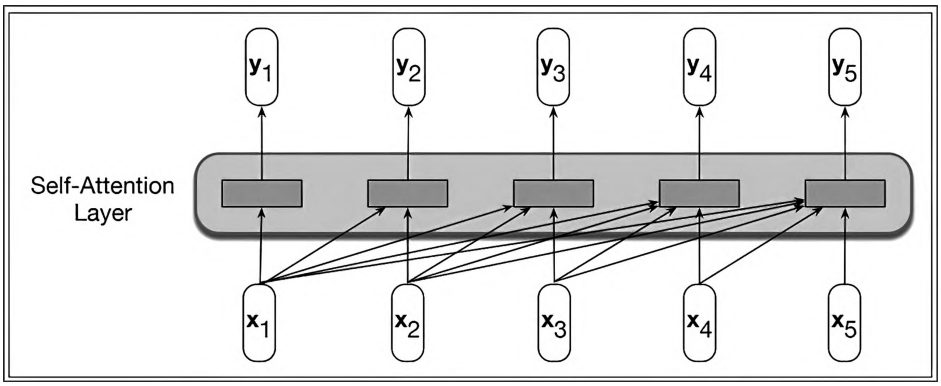

图 5　自注意力层

在图 5 中,x_1、x_2、x_3、x_4、x_5 是输入,y_1、y_2、y_3、y_4、y_5 是输出,每一个输出除了考虑到它相应的输入之外,还应当考虑到该输入之前的所有输入的信息。例如,输出 y_4 除了考虑与其相应的输入 x_4 之外,还应当考虑 x_4 之前的 x_1,x_2,x_3 的信息;输出 y_5 除了考虑与其相应的输入 x_5 之外,还应当考虑 x_5 之前的 x_1,x_2,x_3,x_4 的信息。

在 Transformer 中的多头自注意力子层使用“自注意力”机制,这样便可以充分地表示单词与单词之间联系的密切程度。

例如,对于英语句子“The law will never be perfect, but its application should be just — this is what we are missing, in my opinion”,自注意力机制可以比较句子中单词之间联系的强度,建立句子中单词之间的不同自注意力强度的联系。如图 6 所示。

从图 6 中可以看出,在这个句子中,Law、application、missing、opinion 等单词与其他单词的联系最为密切,自注意力强度比较高。

多头自注意力子层还可以把相关的单词融入正在处理的单词中,从而拓展了模型专注于不同位置的能力。

例如,我们输入英语句子“The animal didn't cross the street because it was too tired”。这个句子中的 it 是指什么呢? 对于我们人类来说,这是一个很简单的问题,it 显然是指 animal,因为只有 animal 这种动物才会有 tired(疲倦)的感觉,但是对于计算机算法来说,这却是一个相当困难的问题,因为 it 的前面除了单词 animal 在之外,还有好几个其他的单词,它们也可能成为 it 的所指对象。但是,由于

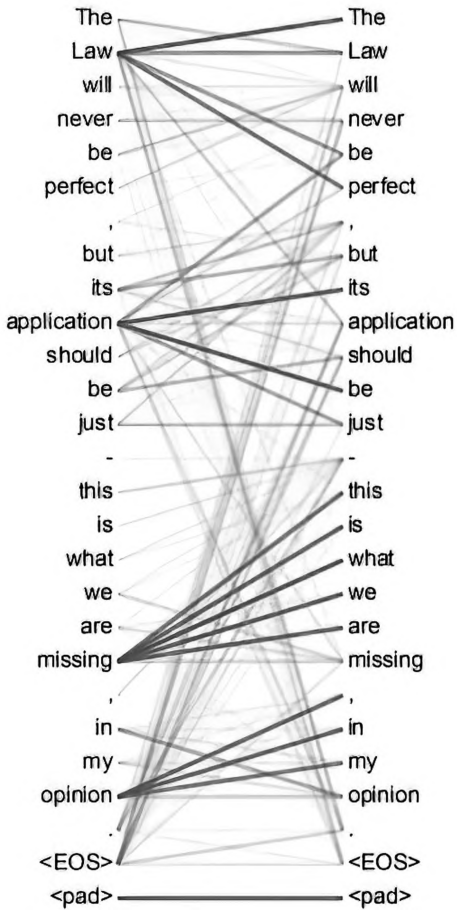

图 6　自注意力机制建立单词与单词之间不同强度的联系

Transformer 有“多头注意力子层”,当模型在处理 it 这个单词的时候,多头注意力子层会把所有相关的单词融入我们正在处理的单词 it 中,从而允许 it 和 animal 建立起比其他单词更加密切的联系。如图 7 所示。

在图 7 中,自注意力机制可以建立起 it 和相关单词之间的联系。当在编码器的第 5 层(Layer 5)中对于 it 在这个单词进行编码时,自注意力机制会关注 The animal,把 The animal 的一部分表示编入 it 的编码中。从图 7 中不难看出,尽管 it 与很多单词都有联系,但是,it 与 The animal 的联系最为密切。

在自注意力机制中,自注意力的强度要根据“查询向量(Query,简写为 Q)”“键向量(Key,简写为 K)”和“值向量(Value,简写为 V)”来计算。

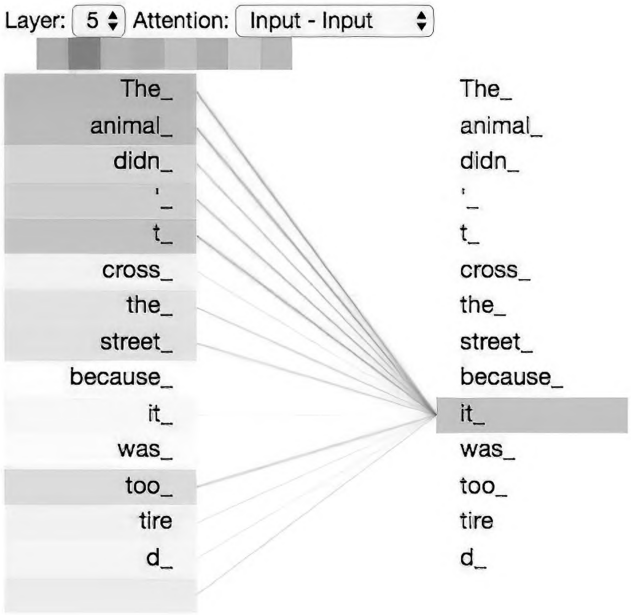

图 7　自注意力机制建立 it 和其他词的联系

“查询向量”Q 的作用在于,在对前面所有的输入进行比较时,表示注意力关注的当前焦点。

“键向量”K 的作用在于,在与注意力的当前焦点进行比较时,表示注意力关注该焦点前面的输入。

“值向量”V 的作用在于,计算注意力当前焦点的输出值。

自注意力强度的计算公式如下:

$$Attention(Q,\ K,\ V) = soft\ max\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

这个公式中,Q 表示查询向量,K 表示链向量,V 表示值向量,d 表示 Transformer 的维度(dimensionality)。

例如,假定 x_1、x_2、x_3 是输入,y_3 是输出,我们可以这样来计算输出 y_3 的值:首先生成输入 x_1、x_2、x_3 的查询向量 Q、链向量 K、值向量 V,然后比较它们之间的查询向量与链向量(key/Query comparisons),计算出 softmax 的值,然后对所有 softmax 的值加权求和(Weight and Sum),最后输出向量(Output Vector)y_3。显而易见,输

出向量既考虑到输入 x_3 的信息，也考虑到它前面的输入 x_1、x_2 的信息。如图 8 所示。

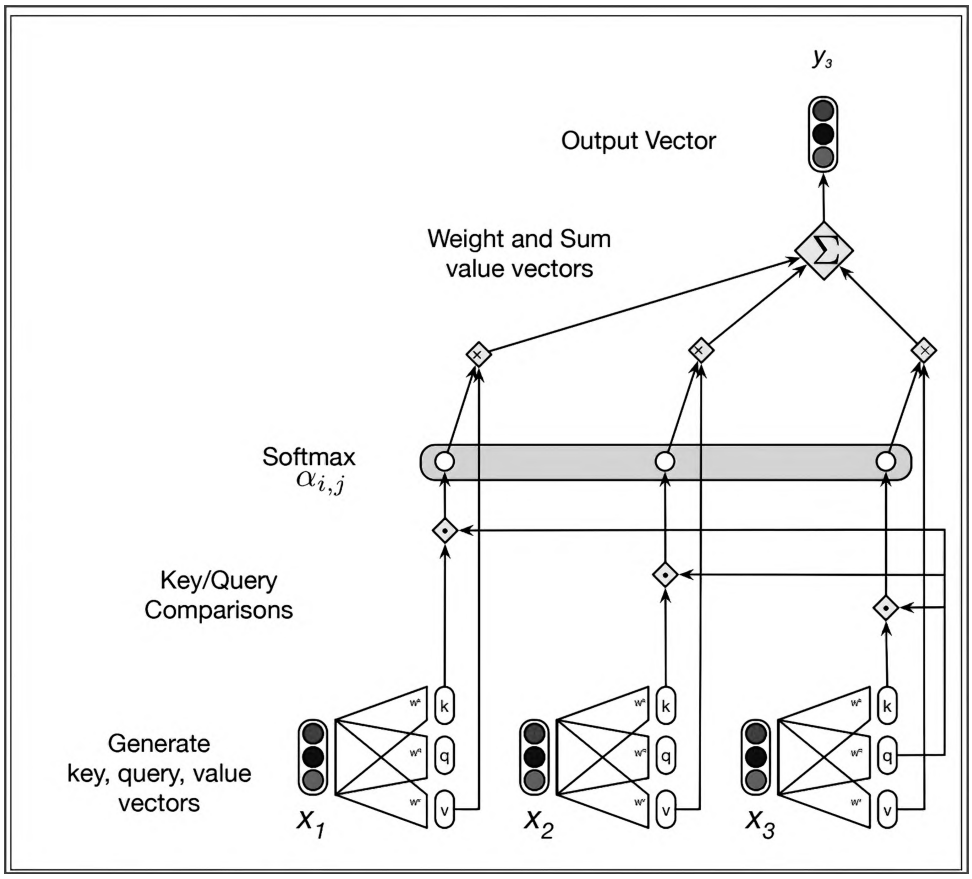

图 8　使用自注意力强度公式计算 y_3 的值

2018 年以来，Google、OpenAI、Meta、百度、华为等公司相继发布了包括 BERT、GPT 等在内的多种大语言模型，这些模型在几乎所有自然语言处理任务中都表现出色。

目前，大语言模型呈现出爆发式的增长局面，特别是 OpenAI 公司在 2022 年 11 月发布 ChatGPT（Chat Generative Pre-trained Transformer）之后，更是引起了全世界的广泛关注。用户可以使用自然语言与 ChatGPT 交互，从而完成包括数字人文中常见的问答、分类、摘要、翻译、聊天等任务。大语言模型必将推动数字人文的发展。

2024 年 2 月 16 日,OpenAI 公司发布文生视频(video generated by text)的大模型 Sora,可以根据文本内容的提示语(prompt)生成 60 秒视频,理解外在世界,逼真而不变形。

在 OpenAI 官网上更新的视频演示中,Sora 不仅能准确呈现物理世界的细节,还能理解物体在物理世界中的存在,并生成具有丰富情感的角色。Sora 还可以根据提示语、静止图像甚至填补现有视频中的缺失部分来生成视频。

例如,一个提示语的描述是:“在东京街头,一位时髦的女士穿梭在充满温暖霓虹灯光和动感城市标志的街道上。”在 Sora 根据这个提示语生成的视频里,一位女士身着黑色皮衣、红色裙子在霓虹街头行走,不仅主体连贯稳定,还有多镜头,包括从大街景慢慢切入对女士的脸部表情的特写,以及潮湿的街道地面反射霓虹灯的光影效果。如图 9 所示。

图 9　文生视频

其他的文生视频演示还包括在雪地里嬉戏的狗、在公路上行驶的车辆、在城市摩天大楼之间的半空中游动的鲨鱼、跪在融化的红色蜡烛旁边的毛茸茸的小怪物……

Sora 成为了世界模拟器,Sora 生成的视频,栩栩如生,充满了人文气息。不难想象,如果把 Sora 的先进技术应用于数字人文的研究中,必将大大地提高数字人文的研究水平。

大语言模型目前还缺乏可解释性,但是,根据数学和语言学的发展历史,我们可以看到大语言模型在数学和语言学上的一些根据。

美国数学家所罗门诺夫(R. Solomonoff,1926—2009)1964 年以“归纳推理的形式理论”(A Formal Theory of Inductive Inference)为题,发表于计算理论的重要刊物《信息与控制》(*Information and Control*),提出了所罗门诺夫归纳法(Solomonoff Induction)。

所罗门诺夫归纳法可以如下定义:给定序列($x_1, x_2, \cdots x_n$),预测 x_{n+1}。归纳推理就是力图找到一个最小的图灵机,可以为($x_1, x_2, \cdots x_n$)建模,从而准确地预测后续序列。

例如,如果一个序列是 n 个 1(1,1,1,…),那么我们可以写出如下程序输出该序列:

```
For i = 1 to n
        print 1
```

这个序列的描述长度就是 O(log(n))。

例如,如果我们给出序列(3,5,7),会有无穷多种预测后续的结果,其中一种是 9,因为程序有可能打印奇数,如下:

```
For i = 1 to n
        print 2i+1
```

但这样也许猜得不对,还有一种可能性是 11,因为程序有可能是打印素数的。很明显,打印素数的程序就要比打印奇数的程序复杂很多,也就是说素数的描述长度要大于奇数的描述长度。

显而易见,这是一个“在下一个字符上下赌注”(bet on next symbol)的问题,其实就是 GPT 为代表的大语言模型的核心机制:“预测下一个词元”。

在大语言模型中,计算机根据上下文“The cat is chasing the”,预测出下面一个单词是 mouse,因为 mouse 的概率高于其他单词 dog、squirrel、boy、house 等的概率。这样的计算的理论根据就是“所罗门诺夫归纳法”。

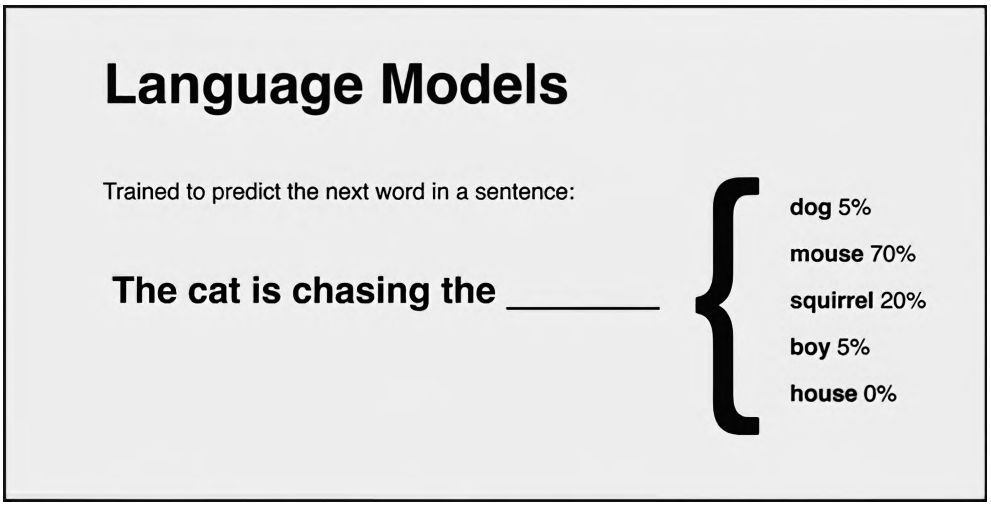

图 10　预测句子中的下一个单词

在这个意义上,我们认为,大语言模型在数学上的理论根据,就是“所罗门诺夫归纳法”。

那么,大语言模型的语言学理论依据是什么呢?

所罗门诺夫归纳法与乔姆斯基(N. Chomsky)的语言模型理论也是有联系的。所罗门诺夫曾经想到一个问题:给定一些句子,看计算机能否学会生成这些句子的语法。此时乔姆斯基的《语言描述的三种模型》(Three Models for Description of Language)的文章刚刚发表,所罗门诺夫受到启发,他把乔姆斯基语法推广成概率语法(probabilistic grammar)。他的“归纳推理机”的一种应用场景就是通过输入文本,学会语法,这被他后来称为“语法发现(discovery of grammar)”。

乔姆斯基的先天内生文法(innate grammar)其实就是所罗门诺夫的先验概率分布,只不过乔姆斯基采取了理性主义(rationalism)的立场,而所罗门诺夫无疑是经验主义(empiricism)的。我们认为,从历史发展的角度看来,乔姆斯基的语言学理论与所罗门诺夫归纳法有密切联系,而所罗门诺夫归纳法是大语言模型的理论基础。因此,大语言模型与乔姆斯基的语言学理论是有关系的。大语言模型可以处理形式各异的各种不同的语言,而支配这些形式各异的语言的先天内生文法就可能是这些语言压缩后的产物,它是这些形式各异的语言的基石。

在大语言模型 GPT 中,所罗门诺夫归纳法实际上是将自然语言的文本符号串

词元化,也就是给大语言模型提供可计算可理解的“基本粒子”——“词元”,然后用这些“基本粒子”去组合文本语言的符号串。

其他的大语言模型如 Sora 等,是把视频素材中的时空基元(spacetime patches)视为基本粒子,根据已知的时空基元来预测“下一个时空基元”。如图 11 所示。所以,Sora 的数学原理也是所罗门诺夫归纳法。

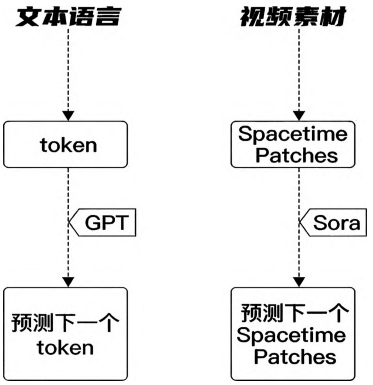

图 11　GPT 与 Sora 比较

2024 年 2 月 19 日,深度学习之父辛顿(G. Hinton)在英国牛津大学罗曼讲座(Romanes Lectures)上发表题为“数字智能会取代生物智能吗”的演说,他认为,乔姆斯基关于天赋语言的学说是“疯狂的理论”(crazy theory),对其语言理论提出了尖锐的批评。辛顿最近被授予“尤利西斯奖”(Ulysses Medal),他于 2024 年 4 月 10 日在爱尔兰都柏林大学学院授奖大会的讲话中再次批评乔姆斯基,再次指出乔姆斯基的理论是一种“疯狂的理论”。

辛顿说:“语言学家被一个叫乔姆斯基的人误导了好几代人,乔姆斯基实际上也获得了这个著名的奖章。所以它不会持久。他有一个疯狂的理论,认为语言是学不来的。我的意思是,他设法说服了很多人。从表面上看,这简直是疯了。语言显然不是。现在,这些大型神经网络学习语言,它们不需要任何先天结构。他们只是从随机权重开始。还有很多数据。乔姆斯基还在说,但这不是真正的语言。这不算。这是不对的。”(都柏林大学演讲,2024)

乔姆斯基认为,语言学理论的构建需要语言事实作为其经验的明证,但是,采用经验明证的目的是更好地服务于理论的构建,生成语法所采用的经验明证一般是与理论的构建有关的那些经验明证。因此,生成语法研究的目的不是全面地、广

泛地、客观地描写语言事实和现象,而是探索和发现那些在语言事实和现象后面掩藏着的本质和原则,从而构建具有可解释性的语言学理论。这是作为科学的语言学理论与作为技术的大语言模型的不同之处。辛顿从技术的立场来批评乔姆斯基的科学理论,显然有偏颇之处。

在数字人文研究中,我们应当关注大语言模型的数学和语言学理论背景,提高数字人文研究的理论水平。

三　数字人文时代的语言研究

从“比萨报告”可以看出,数字人文涉及人文科学中的诸多学科,如文学、语言学、哲学、历史、宗教学、法学、表演艺术、物质文化、音乐等。本文不可能讨论数字人文涉及的众多学科,在本文中,我们只是讨论一下数字人文时代的语言研究问题。

我们觉得数字人文时代的语言学可以研究如下的问题。

(一) 语言资源建设

20 世纪 90 年代以来,为了摆脱困境,自然语言处理的研究者们开始对大规模的非受限的自然语言进行调查和统计,采用一种基于统计(statistic-based)或基于神经网络(neural-network-based)的方法来处理非受限的语言大数据。这样的方法将有可能在语言大数据的基础上检验传统的理论语言学使用基于规则的方法所得出的各种结论,从而使我们对于自然语言的各种复杂现象获得更为深刻和更为全面的认识。

语料库(corpus)以电子计算机为载体来存放语言大数据材料,这些存放在电子计算机中的语言大数据是大规模的(large scale),并且是在语言的实际使用中真实出现过的(authentic),因此,它们可以如实地反映语言现象,克服语言学家观察语言现象时的主观性和片面性,这样的未经加工的语言大数据对于语言学研究已经很有用;而这些真实的语言大数据经过标注、分析、加工处理之后,就可以变成更加有用的语言资源(language resources)。所以,不论是未经加工的“生语料”语言大数据或是经过加工的“熟语料”语言大数据都是非常宝贵的。

传统语言学基本上是通过语言学家归纳总结语言现象的手工方法,采用内省(introspection)或者诱导(elicitation)的方式来获取语言知识的,由于人的记忆能力

有限,任何语言学家,哪怕是语言学界的权威泰斗,都不可能记忆和处理浩如烟海的全部的语言数据,因此,使用传统的内省或者诱导的方式来获取语言知识,犹如以管窥豹,以蠡测海,这种获取语言知识的方法不仅效率极低,而且带有很大的主观性和片面性。

语言大数据是客观的、可靠的语言资源,语言学研究应当依靠这些语言资源。语言大数据中包含着极为宝贵的语言知识,我们应当使用新的方法和工具来获取这些知识。当然,前辈语言学家数千年积累的语言知识(包括词典中的语言知识,语法书中的语言知识)也是宝贵的,但由于这些知识是通过这些语言学家们的"内省"或者"洞察力"发现的,需要我们使用语言大数据加以审查。

语料库语言学家辛克莱(John Sinclair)一针见血地指出:"生造的例子看上去不管是多么地可行,都不能作为使用语言的实例。"

在计算机中建立了大规模的语料库之后,我们就可以使用机器学习(machine learning)或深度学习(deep learning)的方法,自动地从浩如烟海的语言大数据中获取准确的语言知识。这是语言学获取语言知识方式的巨大变化,作为21世纪的语言学工作者,都应该注意到这样的变化,逐渐改变获取语言知识的手段。

语言知识和语篇知识都包含在语言大数据当中。随着语言大数据加工的逐渐精细和深入,我们获得的语言知识也就越加准确和深刻。

语言大数据同时也是语言知识的宝库,语言大数据中蕴藏着丰富的语言知识,词汇知识、句法知识,是最重要的语言资源。语言大数据的使用,为语言学的研究提供了一种新的思维角度,辅助人们的语言"直觉""内省"和"诱导",从而克服研究者本人的主观性和片面性,现在已经成为语言学研究的主流方法。

语言学家利用语言大数据来研究语言学,正如天文学家利用望远镜来研究天文学,生物学家利用显微镜来研究生物学一样,能够使他们如虎添翼,其意义是非常重大的。望远镜的发明使天文学家能够观察到他们过去难以观察到的宏观世界的现象,显微镜的发明使生物学家能够观察到他们过去难以观察到的微观世界的现象,计算机可读的语言大数据就好比语言学研究的望远镜和显微镜,语言大数据的使用扩展了语言学家的眼界,使他们看得更远、看得更细,从而使他们能够发现更多的语言现象,挖掘出更多的语言事实,把语言学的研究推向一个新的阶段。从某种意义上说,语言大数据的使用,是语言学研究的一次革命性的进步。

大数据时代的自然语言处理采用机器学习(machine learning)的方法从大规模真实的语言大数据中获取信息。

机器自动学习的方法主要有三种类型:有监督的学习(supervised learning)、无监督的学习(un-supervised learning)、半监督的学习(semi-supervised learning)。

有监督的学习实际上是对于语言大数据进行分类,先使用事先定义好的类别或范畴标记对于数据的实例进行标注,作为训练数据,机器根据这些标注好的训练数据进行自动学习,再根据学习得到的知识对于新的数据进行分类。由于用来学习的训练数据是用事先定义好的标记进行过标注的,机器学习的过程是在这些训练数据的监督下进行的,所以叫作有监督的学习。

在无监督的学习中,用来学习的语言大数据没有使用事先定义好的类别或范畴标记进行过标注,要使用机器学习的算法来自动地发现隐藏在数据中的结构或规律。这种无监督学习的一个关键技术是聚类,聚类技术根据数据实例的相同点或相异点,自动地把它们聚类为不同的组合。例如,可以把互联网的页面聚类为不同的组合,每一个组合代表一个特定的主题,也可以把文件聚类为不同的层次,每一个层次代表一个特定的主题层次。

有监督的学习要求事先由人工标注语言大数据实例,需要付出巨大的人工的劳动量,费力而又费时。为了减少人工标注的劳动量,可以同时从标注过的数据实例和没有标注过的数据实例中进行学习,标注过的数据实例的集合可以比较小,而没有标注过的数据实例的集合可以很大,这样的模型叫作半监督的机器学习。

机器自动学习的这些方法已经成熟,而且广泛地应用于自然语言处理的研究中,这就从根本上改变了传统的获取语言知识的手段,对于自然语言处理的发展具有革命性的意义。

目前,基于多层神经网络的、以语言大数据作为输入的深度学习(deep learning,简称 DL)方法引入自然语言处理中。这是一种新型的机器自动学习。深度学习的训练方式是无监督的特征学习,使用多层神经网络的方法。这种多层神经网络是非线性的,可以重复利用中间层的计算单元,减少参数,计算机从海量的大数据中可以自动地产生模型的特征和算法。

最近,深度学习向神经网络中融入了记忆机制,把知识驱动(knowledge-driven)与数据驱动(data-driven)结合起来,架起了符号主义(symbolism)与连接主义

(connectionism)之间的桥梁。

现在我们进入了信息网络时代,因特网(internet)本来是只使用英语的,随着因特网的发展,越来越多的国家和地区用自己的语言使用因特网,因特网变成了一个多语言的"互联网"(Web)。

根据 Miniwatts Marketing Group(2019)公布的调查结果,2019 年在互联网上使用得最多的 10 种语言是:英语、汉语、西班牙语、阿拉伯语、葡萄牙语、印尼语/马来语[①]、法语、日语、俄语、德语(见表 1)。

表 1　互联网上使用最多的十种语言(2019 年 4 月 30 日)

Miniwatts Marketing Group(2019)

互联网 10 大语言	该语言人口数 2019 年估算	该语言 互联网用户数	互联网 渗透率	互联网用户增长率 2000—2019	互联网用户 占世界百分比
英语	1 485 300 217	1 105 919 154	74.5%	685.70%	25.20%
汉语	1 457 821 239	863 230 794	59.2%	2 572.3%	19.3%
西班牙语	520 777 464	344 448 932	66.1%	1 425.8%	7.9%
阿拉伯语	444 016 517	226 595 470	51.0%	8 917.3%	5.2%
葡萄牙语	289 923 583	171 583 004	59.2%	2 164.8%	3.9%
印尼语/ 马来语	302 430 273	169 685 798	56.1%	2 861.4%	3.9%
法语	422 308 112	144 695 288	34.3%	1 106.0%	3.3%
日语	126 854 745	118 626 672	93.5%	152.0%	2.7%
俄语	143 895 551	109 552 842	76.1%	3 434.0%	2.5%
俄语	97 025 201	92 304 792	95.1%	235.4%	2.1%
十大语言	5 193 327 701	3 346 642 747	64.4%	1 123.0%	76.3%
其他语言	2 522 895 508	1 039 842 794	41.2%	1 009.4%	23.7%
世界总数	7 716 223 209	4 386 485 541	56.8%	1 115.1%	100.0%

① 在语言学分类中,印尼语属于马来-波利尼西亚语系。印尼语源自于苏门答腊岛东北部的马来语,是马来语的一个地区性变种,同标准马来语有超过 80%的同源性,因此,可以将印尼语与马来语看成一种语言。

互联网上的语言数据越来越多,我们进入了“大数据时代”。

目前,世界上的数据增长速度越来越快,每年数据增长 40%左右,仅在 2011 年的数据就有 1.8 ZB[①] 之多,这些数据量已经超过了 2011 年以前人类历史上的全部数据量的总和。根据《中国大数据白皮书(2020)》的估计,2020 年全球生产的数据量已经达到 47 ZB(1 ZB = 10 TB = 1 万亿 GB),2035 年预计将达到 2 142 ZB 之多。而中国产生的数据量约占全球总数的五分之一。这些大数据的主要载体就是各种语言文字,语言文字资源是大数据的主要资源。

据中国翻译协会《2019 年中国语言服务行业发展报告》统计:全球语言服务产值预计将首次接近 500 亿美元;中国涉及语言服务的在营企业有 36 万余家,语言服务为主营业务的在营企业近万家,总产值超过 300 亿元,年增长 3%以上。

我国网民数量从 2000 年的 2 250 万用户上升至 2022 年 7 月的近 11 亿,互联网用户数占亚洲三分之一强。我国互联网渗透率(web penetration rate)由 2019 年的 59.2%提高到 74.4%,农村网民规模达 2.93 亿,占全体网民的 27.9%;城镇网民规模为 7.58 亿,占比 72.1%。从网民群体年龄结构看,截至 2022 年 6 月,20—29 岁、30—39 岁、40—49 岁网民占比分别为 17.2%、20.3%和 19.1%,高于其他年龄段群体;50 岁及以上网民占比为 25.8%。网民上网时间人均每周多达 29.5 小时,使用手机上网的比例达 99.6%。许多人使用一种以上智能设备上网,使用台式电脑、笔记本电脑、电视和平板电脑上网的比例分别为 33.3%、32.6%、26.7%和 27.6%。老人群体非网民比例比较高,截至 2022 年 6 月,60 岁及以上非网民群体占非网民总体的 41.6%,比全国 60 岁及以上人口比例高出 22.5%。网民群体趋向年轻,经济发达地区网络覆盖率更高,网民占比高,网络语言生活存在比较明显的城乡差异和年龄差异。但是,随着我国社会经济的发展,这些差异将逐渐消失。随着互联网的普及,由于互联网存在多种语言,互联网上的翻译需求日益增大,对于机器翻译的需求也就更加突出。

目前,我国开设外语类专业的高校数量多达上千所,其中设立有翻译硕士(Master of Translation and Interpretation, MTI)和翻译本科(Bachelor of Translation and Interpretation, BTI)专业的院校分别有 250 余所和 280 余所,MTI 累计招生数达 6 万余人,有的高校还设立了翻译博士(Doctor of Translation and Interpretation,

① 大数据的起始计量单位是 P(1 000 个 TB)、E(100 万个 TB)、Z(10 亿个 TB)。

简称 DTI)。

根据国际权威机构对于世界翻译市场的调查显示,全世界翻译市场的规模在1999 年只是 104 亿美元,在 2003 年为 172 亿美元,而在 2005 年达到了 227 亿美元,2019 年达到了 500 亿美元。

目前,尽管我国翻译市场的规模已经超过了 372 亿元人民币,但是现有的国内翻译公司只能消化 10%左右,由于无法消化大量从国际上传来的信息流,我们的信息不灵,就有可能使我们在国际竞争中失去大量的机会。

在这种情况下,随着多语言网络世界的发展和翻译需求量的日益增长,为了克服语言障碍,采用机器翻译不失为一种可行的办法。

在当前的机器翻译研究中,语言数据资源的匮乏是一个非常严重的问题,对于机器翻译来说,语言数据是翻译知识的来源,语言数据资源规模的大小,直接制约着神经机器翻译的效果。

在大语言模型中,自然语言生成(natural language generation,简称 NLG)可以根据“N 元语法”(N-gram)的原理来进行。在当前的大语言模型中,语言模型的规模达到数百亿甚至数千亿单词,计算机就可以在大语言模型的基础上,自动地生成自然语言。语言数据资源的规模越大,自然语言处理的效果越好。而 N 元语法的数学基础就是所罗门诺夫归纳法。

当前流行的大语言模型是根据 N 元语法来进行自然语言的自动生成的。

在自然语言处理中,一般可以使用二元语法、三元语法或四元语法,N 元语法的阶数越高,系统的质量越高。

谷歌公司在 2007 年曾经研制过七元语法,也就是考虑当前词前面六个单词对于当前词的影响,神经机器翻译的质量明显地提高了。由此可见,语言数据资源的数量规模决定了神经机器翻译的质量水平。

由 OpenAI 公司开发的基于 Transformer 的生成式预训练模型(Generative Pre-Trained Transformer)已经成为当前自然语言处理研究的核心技术,包括 GPT－1、GPT－2、GPT－3、InstructGPT、ChatGPT、GPT－4,我们把它们统称为 GPT 系列,简称为 GPT。GPT 系列利用 Transformer 模型,从大规模的语言数据资源中获取了丰富的语言知识,GPT 系列在语言生成任务上达到了相当高的水平。这样一来,GPT 序列便成为了自然语言处理研究的最重要的大语言模型。

GPT 系列的语言数据资源越丰富,其训练的参数就越来越多,其性能就越来

越好。

2018 年 6 月开发的 GPT－1 有 1.17 亿参数。它根据预训练模型的原理,使用预测下一个单词的方式训练出基础的语言模型,然后针对分类、蕴含、近义、多选等下游任务,使用特定数据集,更新模型参数,对模型进行调优与适配。

2019 年 2 月开发的 GPT－2 有 15 亿个参数,GPT－2 开始训练的数据取自于著名社交站点 Reddit 上的文章,累计有 800 万篇文章,语言数据资源还不够丰富。它通过多任务学习,获得了迁移学习的能力,能够在零样本(zero shot)设定下执行各类任务,无须进行任何参数或架构修改,具有一定的自我纠偏能力。

2020 年 5 月,GPT－3 启动,有 1 750 亿参数。训练语料的 60%来自 2016—2019 年的 C4,爬取了 40 多种语言的网络数据。这些数据经过清洗,22%来自网络文本(WebText2),16%来自书籍(Books),3%来自维基百科(Wikipedia),开始了大规模的机器学习,把能获取到的人类书籍、学术论文、新闻、高质量的各种信息作为学习内容,参数总量是 GPT－2 参数的 117 倍,语言数据资源非常丰富。GPT－3 有 1 750 亿个参数,犹如一个巨大的恐龙,而 GPT－2 只有 15 亿参数,犹如一个普通人。从 GPT－2 到 GPT－3,参数增长了 100 多倍。

从 2018 年的 ELMo 开始,大语言模型的数据资源日益增长,其参数也日益增长。ELMO 的参数量仅有 9 千 4 百万个(94 M),GPT－3 增长到了 1 750 亿个(175 B)。这样庞大的参数是人类远远无法达到的。如果我们人类每秒钟处理一个单词,不计睡眠时间,一个人终其一生处理的单词数量也不会超过 10 亿单词,而 ChatGPT 可以处理上千亿的参数,2 000 多亿单词。这样的能力是人类望尘莫及的!

GPT－3 显示出强大的上下文学习(in-context learning)能力,用户只要使用少量的示例(few shots)就可以说明任务,例如,用户只要给出几对英语到法语的单词作为示例,再给出一个英语单词,GPT－3 就可以理解用户意图是要做翻译,继而给出对应的法语单词译文。

在 GPT 系列的研制中,随着语言数据资源的增加,词向量的长度和参数量也随之增加。GPT－1 的语言数据资源的规模约为 5 GB,词向量的长度为 768,参数量为 1.17 亿;GPT－2 的语言数据资源的规模为 40 GB,词向量的长度为 1 600,参数量为 15 亿;GPT－3 的语言数据资源的规模为 45 TB,词向量长度为 12 888,参数量为 1 750 亿。

当语言数据资源的规模参数超过500亿的时候,系统会出现“涌现”(emergence)现象。研究人员惊讶地发现,只需要输入一段提示语(prompt),即便在没有训练过的新任务上,系统也能够举一反三,无师自通,很好地工作,显示出越来越接近于人类的优秀表现,生成的语言也就越来越接近人类的语言。

美国斯坦福基础模型研究中心语言大模型综合评测发现:当大语言模型的语言数据资源的规模扩展到500亿参数时,模型的准确率(accuracy)、鲁棒性(robustness)等性能指标都出现跃升。

这样的“涌现”现象似乎意味着,当语言数据资源的规模在数量上增加到500亿时,GPT系统发生了从量变到质变的重大变化。因此,只要不断地增大语言数据资源的规模,就会产生质变的飞跃。

大语言模型的发展历程虽然只有短短几年的时间,但是发展速度相当惊人。截至2023年6月,国内外已经有超过百种的大语言模型相继发布,形成了“百模大战”的局面。这种“百模大战”的本质,就是拼语言模型数据资源的规模。

大语言模型是一种变革性的人工智能技术,它将重塑社会和科学技术的发展,但同时它也存在多种明显的风险及可以预见的风险。

一方面,大语言模型由于其固有的“幻觉”(hallucination)问题,可能会生成不真实、前后不一致的内容,或者生成不符合人类期望的文本,其中可能包含歧视、偏见和泄露他人隐私、敏感信息的内容。大语言模型还可能传播其训练数据中的有害信息和有毒内容,产生误导性和虚假性的信息。

另一方面,大语言模型可能会被别有用心的人用来执行恶意行为。未经对齐的大语言模型能够生成以假乱真的假新闻,也能被黑客们利用,对网络上的设备开展攻击。这些恶意行为会对我们的日常生活产生负面影响,甚至会对整个社会造成严重的伤害。随着其能力的不断增强,大语言模型还可能展现出“追求”自我保护、自我增强、获取资源等目标。

基于大语言模型研制出来的各种语言数字代理(digital agent)将越来越多地融入我们的日常生活中。为了克服大语言模型的这些弊端、避免各种难以预测的风险,我们需要推动大语言模型对齐(LLMs alignment)技术的研究,使大语言模型的输出和行为与人类的期望和价值保持一致。

由此可见,加强语言数据资源规模的建设,提高语言数据资源的质量,保障语言数据资源的安全性,应当成为数字人文研究的重要任务。

(二)语言文化遗产的数字化重建

通过数字技术切入人文领域,对人类文化遗产的传承、传播、翻译、全球化和创新提供新的方法。

楔形文字(cuneiform script),由美索不达米亚的苏美尔人于公元前 31 世纪左右所创,这是已知世界上最古老的文字。由于楔形文字多在泥板上刻画,所以线条笔直形同楔形,使用芦苇秆或木棒压印在泥板上来书写,因此文字笔画大都为具三角形的线条,而字形也随着文明演变,逐渐由多变的象形文字统一固定为音节符号。数千年来历经沧桑,大量的楔形文字泥板都已经破碎,需要进行拼接才能释读。如图 12 所示。

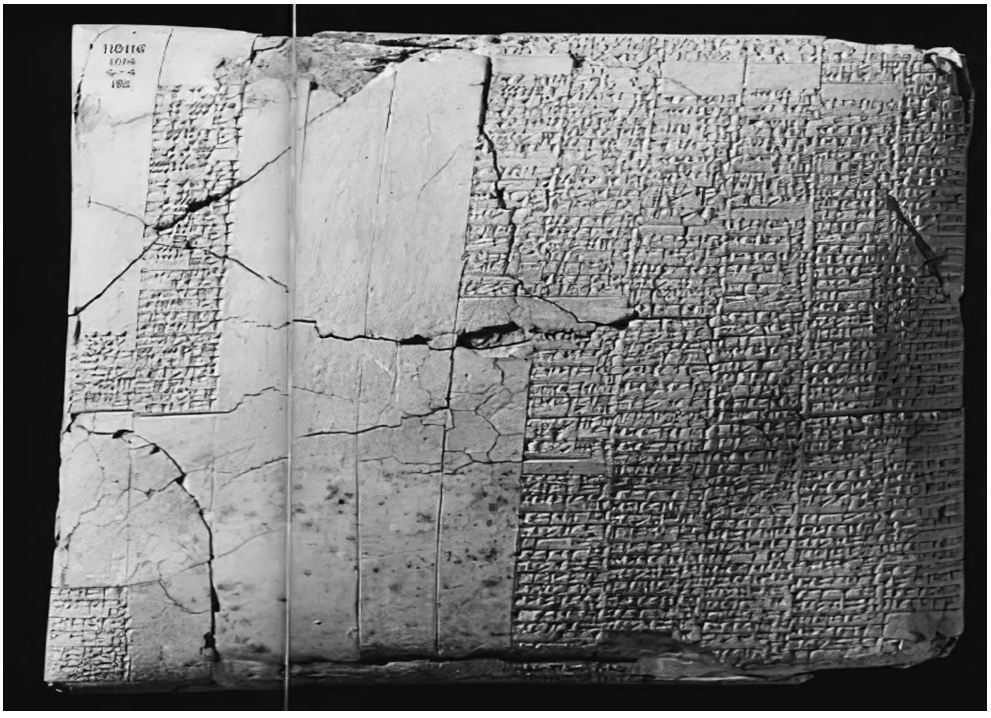

图 12　楔形文字碎片的拼接

2006 年,美国国家人文基金会设置了数字人文办公室用于推动数字人文研究项目的发展。2007 年 2 月,这个基金会资助布鲁克林大学考古研究中心,使用数字人文的技术来拼接楔形文字的碎片。

在这个基金会的支持下,布鲁克林大学考古研究中心使用激光扫描和 3D(三维)定量的数字技术来制作古美索不达米亚楔形文字板的数字模板。学者们可以把这些数字模板连接零碎的文字碎片,以此推断这些碎片是否出于同一篇文章,从而可以大量地拼接出楔形文字的文本,数字人文技术推动了楔形文字的释读研究。

2023 年,由美国加利福尼亚大学伯克利分校数据科学团队与以色列阿里尔大学数字历史实验室共同研制了楔形文字语言到英语的机器翻译系统,并在 EvaCun2023 国际会议上进行了评测。

楔形文字是有记录以来人类历史上最早的文字系统之一,在过去的两个世纪,中东地区发现了数十万楔形文字的泥板,其中大多为苏美尔语和阿卡德语。

EvaCun2023 的评测包括三个机器翻译任务——阿卡德语(楔形文字)翻译到英语、阿卡德语(字母转写)翻译到英语以及苏美尔语(转写)翻译到英语。“阿卡德语—英语”平行语料规模约为 5 万,“苏美尔语—英语”平行语料约为 8 000 句对。如图 13 所示。

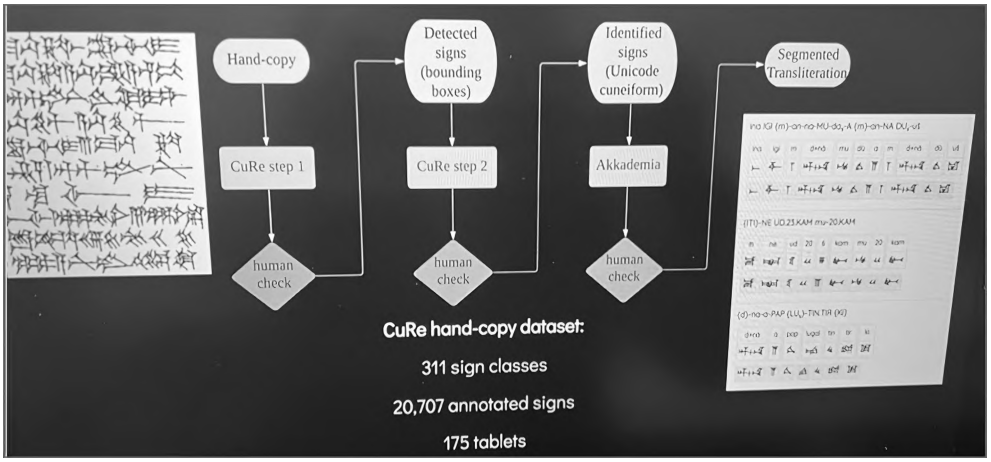

图 13　楔形文字到英语的机器翻译

楔形文字到英语的机器翻译属于资源稀缺语言的机器翻译,如果使用大语言模型来翻译,由于资源稀缺,翻译效果很差。如果只能给出一些提示词(prompt),进行零样本的机器翻译,资源稀缺语言的机器翻译的 Bleu 很低。[①] 如图 14 所示。

① Bleu 值是国际机器翻译评测的一种指标,Bleu 值越高,机器翻译的质量越好。

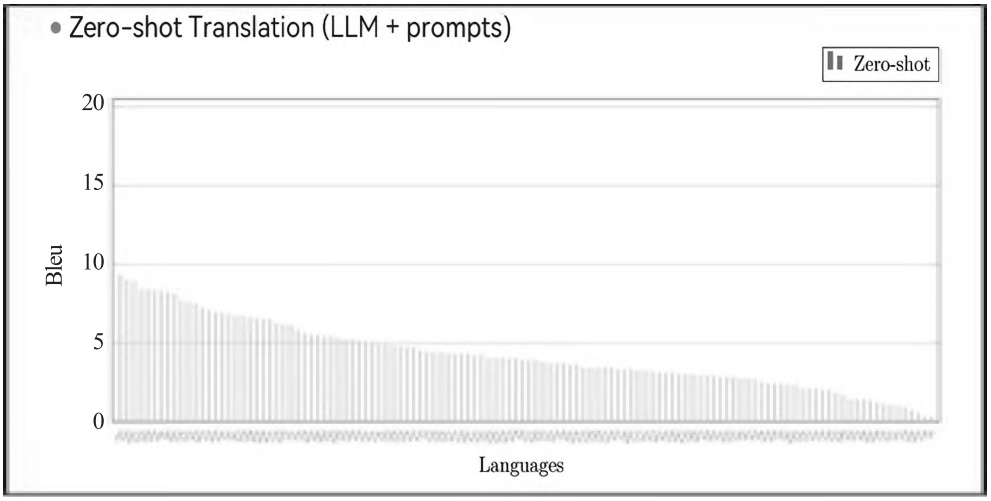

图 14　零样本低资源语言机器翻译的 Bleu 值

如果给予少样本(few shot)并伴以图演示,Bleu 会有所提升。如图 15 所示。

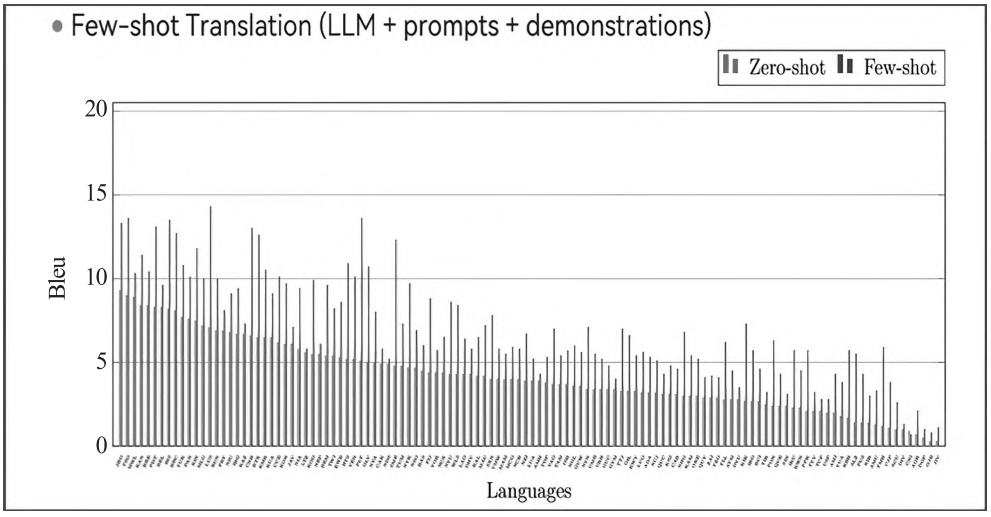

图 15　少样本低资源语言机器翻译的 Bleu 值

如果再提供词典信息,Bleu 值会更高。如图 16 所示。

古代语言的机器翻译应当成为数字人文研究的一个重要内容。古代汉语到现代汉语的机器翻译、古代汉语到英语的机器翻译也取得了初步的成就。

数字人文技术还可以用于古代典籍的收藏和整理工作。

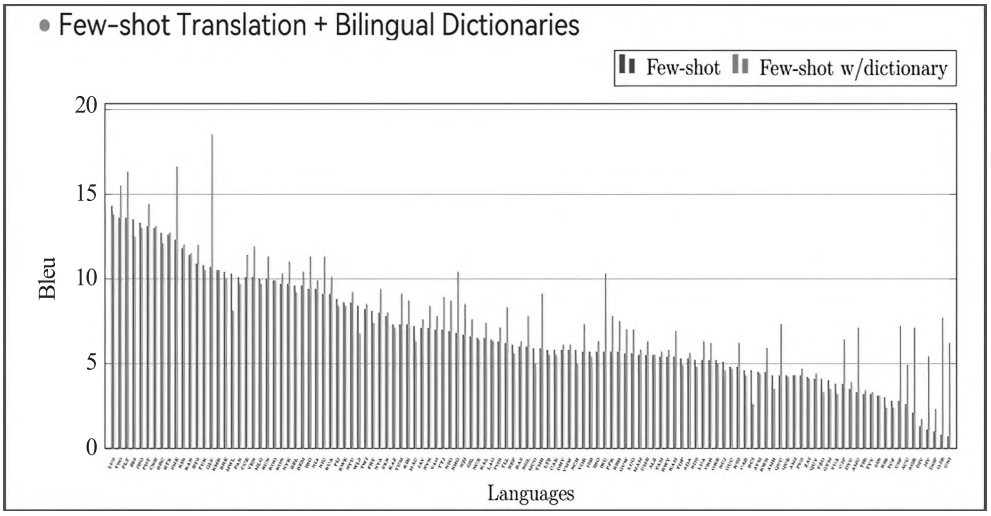

图 16　少样本加词典信息的低资源语言机器翻译的 Bleu 值

公元 79 年意大利维苏威(Mount Vesuvius)火山爆发,直接将赫库兰尼姆(Herculaneum)一座珍藏古老的莎草纸卷轴——赫库兰尼姆莎草纸卷轴(Herculaneum Papyri)的图书馆埋葬。这些卷轴,直到 18 世纪才被挖出,早已成为炭焦的木块。由于太过脆弱,根本无法轻易展开。目前,有 800 多卷保存在意大利那不勒斯的一个图书馆中。如图 17 所示。

图 17　烧焦的纸莎草纸卷轴

2019 年—2023 年间,科学家使用机器学习方法和文本分割(text segmentation)的数字人文技术,对赫库兰尼姆莎草纸卷轴进行破译,恢复了卷轴上的部分墨迹。如图 18 所示。

图 18　恢复卷轴上的墨迹

恢复墨迹之后,科学家破译出希腊语单词“ΠΟΡΦΥΡΑ”。如图 19 所示。

图 19　破译出希腊语单词

图 20 是破译的部分文本,2000 年来首次被人阅读。大约 95%的卷轴内容仍待揭晓。

已经破译的作品展示了超过 15 栏的数百个单词,相当于整个卷轴的 5%的内容。

在赫库兰尼姆莎草纸卷轴文字全部被揭晓的那个时刻,被泥土和灰烬封存了两千年的人类祖先的思想将会再次展现在世人面前!

商周甲骨文是研究商周语言文字和历史文化的第一手资料。存世的有字甲骨共 16 万片,已经发表的有字甲骨共 8 万片,破损非常严重,甲骨残片占 90%以上。为了更好地发挥商周甲骨的作用,学者们开展了甲骨的缀合复原研究。

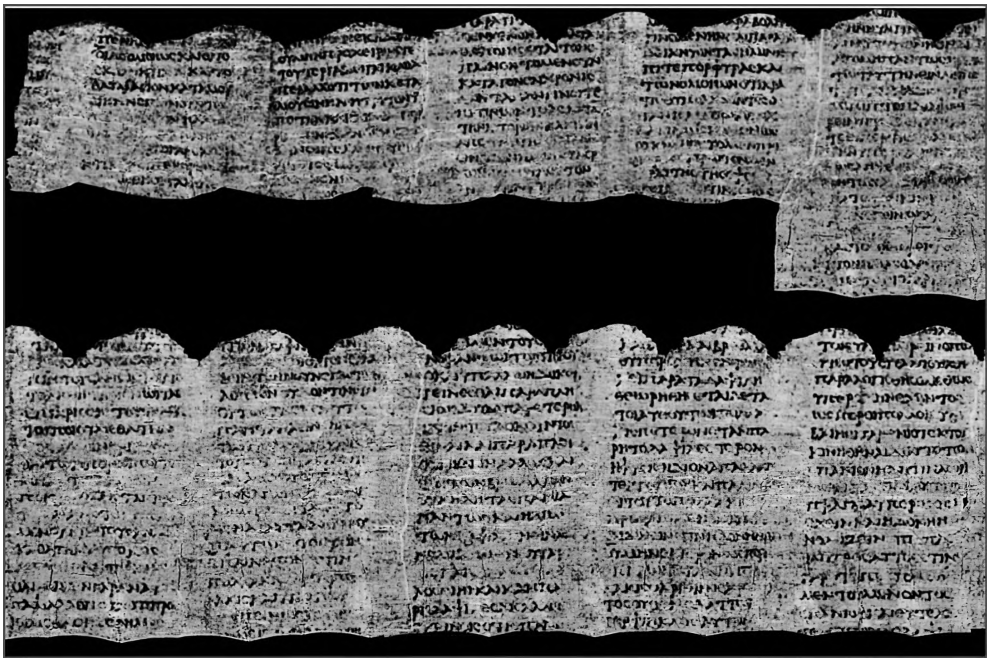

图 20　破译的部分文本

1975 年,四川大学童恩正、张陞楷、陈景春等人利用计算机对甲骨碎片进行缀合,发表了论文《关于使用电子计算机缀合商代卜甲碎片的初步报告》。他们利用“时代、字迹、骨板、碎片、辞、边缘”等 6 项限制条件对于 263 个商代的卜甲碎片进行缀合,缀合率为 40%左右。这项研究比美国布鲁克林大学考古研究中心拼接楔形文字碎片的研究早了 32 年。

最近,我国清华大学计算机系自然语言处理实验室“计算甲骨学”(Computational Jiaguology)研究组研发了人工智能文物拼缀系统“知微缀”(RejoinX),以带动和触发甲骨碎片研究者的直觉。

缀合是甲骨碎片整理的重要手段,过去主要靠手工进行,百余年来,学者们已经缀合 7 000 组,涉及甲骨碎片 1.5 万片,成绩很大。如图 21 所示。近来使用人工智能技术,已经获得数十组机缀甲骨,有希望进一步实现大规模的、可持续的甲骨缀合。

古籍文本的光学字符识别(optical characters recognition,简称 OCR)是古籍数字化的关键环节,其主要内容是用计算机所能使用的编码汉字与古籍中的汉字图像进行识别和对应转换,使得计算机能够对文献内容进行处理。这涉及两个方面

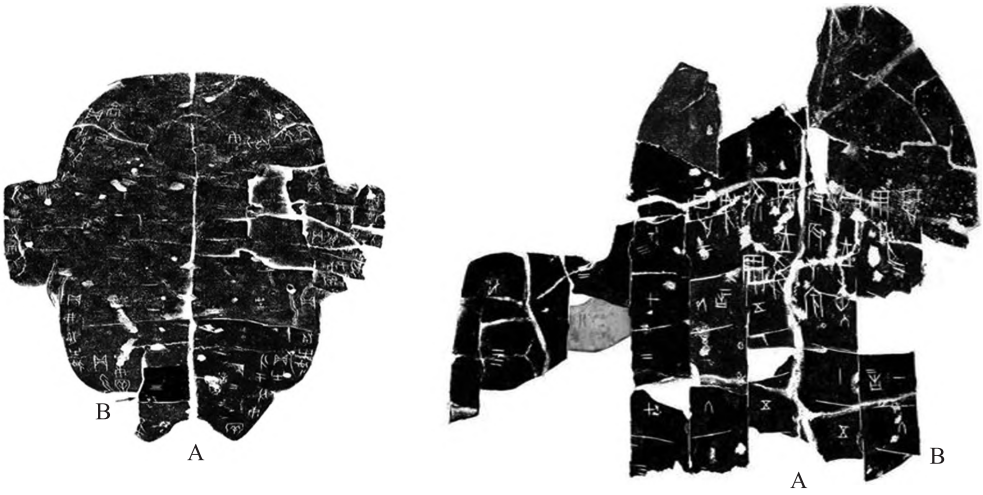

图 21　甲骨碎片缀合

的问题：一个是计算机汉字编码的问题，另一个是计算机图像识别并转换为字符的能力问题。

关于第一个问题，从理论上说，要进行古籍文本的识别，古籍中出现的所有汉字在计算机中都应有相应的编码，并且在以后的发布、利用环节能够被有效使用。这个问题看起来简单，但实际上却非常复杂，原因就在于汉字本身的复杂性。

多年来计算机所使用的汉字编码字符集收录的字数偏少是困扰古籍数字化的主要因素。1980 年颁布的国家标准 GB2312—80 收录的汉字仅 6 763 个，1993 年颁布的 GB13000. 1—93 也只收录了 21 003 个汉字，2000 年 3 月颁布的 GB18030—2000 收录了 27 484 个汉字。近年来，国家标准中收录的汉字逐渐增加。2005 年颁布的 GB18030—2005，已收录汉字 70 244 个，到 2021 年 9 月，Unicode14. 0 版正式发布，收录的汉字已超过了 9. 3 万个。这样一来，汉字字符太少的问题已基本解决。

不过，这里还存在着古籍用字的规范与统一问题。据民间学术网站“汉字宝典”的不完全统计，在古籍中出现过的汉字包括异体字已超过 15 万个，如果考虑到古籍各种写本、印本存在着大量写法有细微差别而导致计算机 OCR 识读时可能发生偏差的“异形字”，数量将会成倍地增加。无限制地增加字符集中汉字的数量并不能解决古籍的用字问题，反而会带来新的问题。

古籍的用字,本来是不多的,统计如下:

《易经》总字数 21 055 个,不重复用字 1 363 个;
《尚书》总字数 28 073 个,不重复用字 2 025 个;
《诗经》总字数 37 438 个,不重复用字 2 989 个;
《论语》总字数 15 935 个,不重复用字 1 355 个;
《孟子》总字数 35 417 个,不重复用字 1 913 个;
《荀子》总字数 75 815 个,不重复用字 2 726 个;
《左传》总字数 195 792 个,不重复用字 3 290 个;
《论衡》总字数 212 050 个,不重复用字 3 630 个;
《列子》总字数 30 900 个,不重复用字 2 329 个。

其中《论衡》的篇幅最长,不重复用字也才 3 630 个。可见古籍中的字种数目并不是很多。

但是,当我们在进行古籍光学字符识别时,由于存在大量的异体字,所需要处理的字种达数十万个之多,其原因主要是古籍在抄写、刻印的过程中,除去抄错、刻错的以外,抄写者、刊刻者的书写习惯差异造成了一个汉字对应于多个不同字形的情况。

例如,在敦煌写本中,大量使用“俗字”,如“多”有 14 种字形,“恶”有 27 种字形。由于敦煌写本的书写者大多是文化水平不太高的经生,他们在抄写时,多一笔少一笔、长一画短一画的现象比比皆是,有的书写潦草,胡乱连笔、任意变形。古代书坊刻书也大量使用俗字,由于书坊刻书的主持者文化水平不高、为了节省成本等,就只好大量使用俗字了。这与国子监等官方刻书机构通常使用“正字”是很不同的。明代刻书,常常使用“古字”,并且很多时候是经生们自己生造出来的古字,以附庸风雅、故弄玄虚。还有一个问题就是避讳。除了皇帝之讳外,最难处理的是民间个人的私讳,由于使用私讳的通常只有本家族之人,在私讳中缺笔、变体字大量存在,也使得汉字的系统越来越复杂,汉字的字形越来越多。

由于上述这些原因,《史记》本来使用汉字的字种不到 5 200 个,但是,如果考虑到不同的版本,累计起来,《史记》各种版本使用过的汉字字形,远远超过了这个数量。

关于第二个问题,也就是计算机图像识别并转换为字符的能力问题,由于古籍字形过多,有些差别极其微小,并且很容易与其他字相混淆,而计算机的文字识别

能力有限,文字识别的准确率不高。

汉字的光学字符识别技术经历了几十年的发展,对于现代排版印刷体图书来说,已经进入了实用阶段。但是,对于古籍来说,则远远没有达到可用的程度。受古籍复杂多变的汉字字形、复杂的版式等因素制约,目前,对于比较简单、规范、整齐的古籍,四川大学与阿里巴巴达摩院共同研发的“汉典重光古籍 OCR 平台”,其正确识别率也只能达到 97.5%,也就是说,其差错率是万分之二百五十;而国家图书出版文字差错率标准是万分之一,其差错率是国家图书出版文字差错率标准的二百五十倍,但这已经是目前业内最高的识别率了。

2022 年 5 月 18 日,一批珍藏于美国加州大学伯克利分校的中文古籍善本,以数字化方式回归故土,落地“汉典重光古籍 OCR 平台”。借助阿里达摩院的技术,首批 20 万页古籍已使用人工智能技术完成了数字化,用户可以通过“汉典重光古籍 OCR 平台”来翻阅和检索古籍。据报道,达摩院对于这 20 万页古籍的识别准确率仍然只有 97.5%,远远达不到国家图书出版文字差错率的标准。

古籍中异体字繁多。虽然是同一种书,但有不同的版本,每个版本的用字都可能不同,例如,古籍中最常见的“历”字就有以下几种写法:“曆”本义为“历法”“年历”,在古籍中使用频率最高的是年号,如“聖曆”“大曆”“寶曆”“鳳曆”“慶曆”“天曆”“萬曆”“永曆”。但是,由于清乾隆皇帝名“弘曆”,清代刊印的古籍中往往将“曆”改为“歷”或“厤”“歷”。它们之间只有细微的差别,这些异体字的 Unicode 编码完全不同。如果在一个古籍数据库中,同一种书可能有多种版本,或者引用同一句话,有的可能作“曆”,有的可能作“歷”,有的可能作“歷”,有的可能作“厤”。这样一来,当进行全文检索时,用什么字作为检索词,就成了一个严重的问题。

一个可行的解决办法,就是建立一个汉字正字与异体字、异形字的标准对照表,供计算机进行数据分析和汉字的输入、输出之用。

2013 年 6 月 5 日,国务院公布了《通用规范汉字表》,共收录汉字 8 105 个,全部为简体字。《通用规范汉字表》还有附表《规范字与繁体字、异体字对照表》,但只收录了与 2 546 个规范字相对应的 2 574 个繁体字,对于古籍数字化来说,这是远远不够的。

2021 年 10 月,国家标准《古籍印刷通用字规范字形表》(GB/Z 40637—2021)发布,该标准规定了古籍印刷通用字收字和宋体字形规范原则,给出了 14 250 个古籍印刷通用字的字形、字音以及在国际编码字符集 ISO/IEC 10 646 中的码位,适用

于传世古籍的印刷出版以及现代书刊的繁体版印刷。这个标准对于古籍出版来说是十分必要的,但对于古籍数字化来说还不够,需要进一步扩充。

不久前,国家曾经启动过大型数字化基础工程“中华字库工程”,主要是解决汉字和少数民族文字字形(包括古文字)的标准化及其输入、输出问题。搜集、整理的汉字包括:甲骨文、金文、简牍帛书及其他古文字、石刻,行书、草书、版刻楷体字、宋元及明清印本文献用字,现代出版物用字及符号,少数民族古文字及现行文字,等等。作为古籍数字化的基础条件,我们需要尽快建立一个字数基本够用的汉字正字与异体字、异形字对照表。

我国古籍数字化已有多年的历史,但相对于 20 万种现存中国古籍来说,得到数字化的古籍只占很小一部分。随着计算机深度学习能力、文本挖掘等大数据技术和人工智能的快速发展,我国在继续推进古籍数字化的同时,应当推动古籍由数字化向知识化转变。

古籍数字化和古籍知识化的最大区别,就是古籍数字化仅仅提供简单的字词检索,而古籍知识化则要基于现存所有存世古籍的关联性、结构化,建立“中国古典知识库”(Chinese Classics Knowledge Base,简称 CCKB)。简单地说,就是运用自然语言处理和人工智能的前沿技术,开发适合不同古籍类型的文本分析模型和工具,实现针对古籍文本的词汇抽取、分词和关联分析;借助过往一切古典学的研究成果,周密地设定主题词表,专业地提取各种实体,如年代、地名、人名、书名、篇名、职官、社团、思潮、事件以及各类语言和文化要素,多维度地构建不同实体间的关系,并通过这些实体及相互关系,在保障古籍文献内容完整性及内部逻辑性的基础上突破文献原有结构,对文献进行深层组织和知识管理,构建为体系化的“中国古典知识库”。

“中国古典知识库”是关乎中国古典知识整体的宏大构想。它的建设,有望给予古代文化研究相关的众多学科带来决定性变化,最大程度地促进文献的关联与知识的再发现。网络分析、文献计量、主题模型等自然语言处理技术的应用将进一步提升人文学术的整体性、实证性、求是性与科学性,从根本上促进传统文献学的现代转型。这是我国数字人文研究的重要内容。

自然语言生成是自然语言处理的一个重要研究领域。作文、新闻、散文、诗歌的自动写作都属于自然语言生成的领域,近年来得到人工智能学界的高度重视,其中的诗歌自动写作,特别是中国古典诗歌的自动写作,成为了自然语言生成中最引

人瞩目的研究领域。

计算机自动写出来的诗歌应当满足三个要求:

第一,合语法,计算机生成的诗歌应当接受语法约束,符合语法;

第二,有意义,计算机生成的诗歌能够表达和传递在某种解释下有意义的信息;

第三,有诗性,计算机生成的诗歌要能够与其他文体区分开来,符合平仄、押韵、意境、风格的要求。

由此可见,诗歌自动写作是颇具挑战性的。

清华大学自然语言处理与社会人文计算实验室研发了“九歌”人工智能诗歌写作系统,他们采用最新的神经网络与深度学习技术,基于超过 80 万首人类诗人创作的诗歌进行训练学习,计算机可以自动作古诗。

“九歌”人工智能诗歌写作系统具有多模态输入、多体裁、多风格、人机交互创作模式等特点。例如,用户输入“白露节气”四个字并选择“藏头诗”选项,大约 3.9 秒后,“九歌”就生成了下面这首七言绝句《白露节气》。

白露节气

九　歌

白苹风起蓼花凉,露滴梧桐月在廊。

节换星移人已老,气蒸云散菊初香。

华为公司研制了基于人工智能的 AI 诗人“乐府”,这是首个基于 GPT 预训练模型的诗歌创作系统,其核心理念是先用无标注的文本训练生成语言模型,然后再根据具体的任务通过有标注的数据对模型进行微调。例如,用户只输入一个“露”字,响应时间约 1.48 秒,就得到一首七言律诗《露》。

露

乐　府

轻轻滴沥湮红兰,一片飞来傍脸寒。

晓鉴重匀新学粉,秋波初透未开鸾。

荷珠散处微生露,桂子飘时半落坛。

疑是水精帘下见,裴回无力独倚看。

GitHub 用户开发的“诗三百·人工智能在线诗歌写作平台”,该平台通过语料

爬取、数据处理、数据分析,可以作诗、填词和对联,具有规则作诗与机器学习写诗的能力。例如,用户选择“鹧鸪天”为词牌名。大约 64 毫秒,便生成了下面这首词。

鹧鸪天·白露

诗三百

一夜西风扫落英,晓来庭户觉秋清。不知白露沾衣冷,但有黄花插鬓轻。

人似玉,世非萍,江南塞北几经程。凭君莫说长安道,且尽尊前醉眼明。

计算机创作的古诗尽管有浓厚的机器味,但也显示了深度学习和神经网络处理自然语言的威力。

唐诗是中国文学的瑰宝。1991 年,中国社会科学院把 27 册《全唐诗》的全部正文、异文和注文输入计算机,他们精准地计算出全唐诗共有 53 035 首,作者 3 276 人,从而订正了文学界传统的唐诗有 4 万多首,作者有 2 万多人的说法。他们还开发了“全唐诗索引生成系统”,可以进行全唐诗自动检索。

2022 年,大连海事大学信息科学技术学院构建了结合胶囊网络(Capsule Network)和大语言模型 Transformer 的胶囊—Transformer 集成模型(Cap - Transformer Model),胶囊网络为上通道,Transformer 为下通道,进行唐诗作者的身份识别(authorship attribution)。

他们下载了《全唐诗》的电子版,经过排错、去重等预处理,得到包括 2 300 位作者的 44 734 首唐诗。每首唐诗的内容按照“作者,诗题,唐诗”的数据格式排列。经过统计他们发现,唐诗作者的创作数量极不平衡。近半数的诗人一生只写了 1—2 首诗,只有不到 20 位诗人一生创作诗歌超过 500 首,最高产的诗人是白居易,他一生创作的诗歌数量高达 2 844 首之多。

进行唐诗作者的身份识别,他们假设所有可能的作者集合为:

$$A = \{a_1, a_2, \cdots a_n\}$$

对于每一位作者 $a_i \in A$,有训练样本集:

$$T_i = \{T_{i1}, T_{i2}, \cdots T_{im}\}$$

作者身份识别的任务学习训练集,建立作者风格特征,根据胶囊—转换器集成模型,对于匿名的诗歌 t,指定一个最可能的作者 a_k($a_k \in A$)。

使用胶囊—Transformer 集成模型进行识别时,首先使用上通道胶囊网络提取

唐诗各个意象的局部语义特征,然后使用下通道 Transformer,通过多头注意力机制提取唐诗各个意象的全局语义特征,再结合题材特征,来识别唐诗的作者身份。唐诗作者众多,作者身份识别的难度很大。大连海事大学信息科学技术学院的研究是一项很有意义的探索。(周爱等 2022)

上面我们只是讨论了数字人文在语言研究中应用的两个方面:语言资源建设和语言文化遗产的数字化重建。限于篇幅,就不进一步展开了。

2018 年,教育部提出了"新文科"(new liberal arts)发展战略,强调文科专业应进行专业重组,把以数字技术、计算机技术和信息技术为代表的新技术融入哲学、文学和语言学等课程,以打破专业壁垒,实现文文交叉和文理交叉,开展跨学科的学习与研究。2021 年,教育部发布《教育部办公厅关于推荐新文科研究与改革实践项目的通知》,提出要全面推进新文科建设,构建世界水平、中国特色的文科人才培养体系。

数字人文把计算机科学与人文科学结合起来,是最为典型的文理交叉学科,这正好符合教育部"新文科"发展战略的要求,希望我国的数字人文的研究能够有效地与教育部"新文科"发展战略的实施对接起来,为"新文科"的建设贡献力量。

为了建设"新文科",人文科学的研究者必须进行更新知识的再学习。在五四运动时代,我们的先辈曾经拥抱"德先生"(民主)和"赛先生"(科学),推动了中国社会的进步。在当今的人工智能时代,我们应当拥抱"数字人文"的新技术,进行更新知识的再学习,使我们成为文理兼通、博学多识的人才,在"新文科"的建设的大潮中,发挥我们的聪明才智。

参考文献

冯志伟　1992　《中文信息处理与汉语研究》,商务印书馆。

冯志伟　2015　《基于短语和句法的统计机器翻译》,《燕山大学学报》第 6 期。

冯志伟　2019　《词向量及其在自然语言处理中的应用》,《外语电化教学》第 1 期。

冯志伟　2023　《数字人文研究的四个层次》,《南京师范大学文学院学报》第 3 期。

冯志伟　张灯柯　2023a　《数字人文、元宇宙与自然语言处理》,《外语学刊》第 6 期。

冯志伟　张灯柯　2023b　《人工智能中的大语言模型》,《外国语文》第 3 期。

童恩正　张陞楷　陈景春　1975　《关于使用电子计算机缀合商代卜甲碎片的初步报告》,《四川大学学报》第 2 期。

中国翻译协会　2019　《2019 年中国语言服务行业发展报告》,《今日中国》11 月 10 日。

中国信息通信研究院　2020　《中国大数据白皮书》。

周　爱　桑　晨　张益嘉　等　2022　《诗人密码：唐诗作者身份识别》,《中文信息学报》第 6 期。

Busa R.　1980　The Annals of Humanities Computing: The Index Thomisticus, *Computers and the Humanities* 14: 83－90.

Hutchins W J.,　Sommers H L.　1992　*An Introduction to Machine Translation*. Academic Press.

McCarty W. 1999　Humanities Computing as Interdisciplinary. Is Humanities Computing an Academic Discipline?. Paper delivered at IATH, University of Virginia.

Solomonoff R.　1964　*A Formal Theory of Inductive Inference, Information and Control* 7(1): 1－22.

Schreibman S., Siemens R., Unsworth J. (eds.)　2004　*A Companion to Digital Humanities*. Blackwell Publishing.

冯志伟　新疆大学/大连海事大学外国语学院　zwfengde2010@163.com

丁晓梅　大连海事大学外国语学院　wyxydxm@dlmu.edu.cn

Abstract: Digital humanity is a new interdisciplinary subject that uses computer technology and network technology to study traditional humanities. This paper introduces the history of digital humanities research, the main content of digital humanities research, digital humanities and artificial intelligence, and discusses the relationship between the digital humanities and the language research.

Key words: digital humanities, machine index, machine translation, digitalization of ancient literatures

Abstract: *Yuan Yulin* 袁毓林 (2024) proposed three collaboration types between linguists and large language models, citing examples of human-machine cooperation in linguistic research (namely various scales of projects). This article introduces the testing results by implying these questions to ChatGPT 3.5, together with commentary on the outcomes. It is found that ChatGPT 3.5 performs well in constructing sentences of specific semantic types, reviewing research topics, and writing short papers on popular linguistic issues. However, it underperforms when retrieving specific lexical items and generalizing or extracting complex grammatical rules which underly specific grammatical phenomena. Nevertheless, with progress in artificial intelligence technology, large models such as ChatGPT are likely to be competent enough to work as collaborators for linguists.

Key words: large language model, human-machine collaboration, ChatGPT 3.5, artificial intelligence, collaborator

Abstract: An adequate logical classification denotes a full-fledged typology and a mature science discipline. The fundamental function for a typology is to identify individuals in question. Cross-language identifications and comparisons, and typological