

人工智能中的大语言模型

冯志伟 张灯柯

(新疆大学 中国语言文学学院, 新疆 乌鲁木齐 830046)

摘要:自然语言处理是人工智能的重要内容,大语言模型是自然语言处理的突出成果。本文描述了大语言模型的发展历程,分别介绍了预训练模型、Transformer 模型、动态词向量嵌入模型 ELMO、双向编码表示模型 BERT、生成式预训练模型 GPT 等大语言模型的基本原理与结构,最后讨论大语言模型与翻译活动之间的关系以及大语言模型的内容治理问题。大语言模型不仅推动自然语言处理取得工程方面的成功,更深刻改变了过去的语言知识生产方式,使语言研究从单学科迈向多学科。这种变革和创新无疑将推动语言学发展。

关键词:自然语言处理;大语言模型;预训练模型;Transformer 模型;ChatGPT;内容治理

中图分类号:H087

文献标志码:A

文章编号:1674-6414(2024)03-0001-29

0 引言

1956 年夏天,美国达特茅斯(Dartmouth)学院的助教约翰·麦卡锡(John McCarthy)、哈佛大学的马文·明斯基(Marvin Minsky)、贝尔实验室的克劳德·香农(Claude Shannon)、卡内基-梅隆大学的艾伦·纽厄尔(Alan Newell)和希尔伯特·西蒙(和 Herbert Simon)、麻省理工学院的奥利弗·塞弗里奇(Oliver Selfridge)和雷·索洛蒙诺夫(Ray Solomonoff)、IBM 公司的阿瑟·米勒(Arthur Samuel)、IBM 公司信息研究中心的纳撒尼尔·罗切斯特(Nathaniel Rochester)、普林斯顿大学的特兰查德·摩尔(Trenchard More)等 10 人,在美国达特茅斯学院举行了为期两个月的学术讨论会。他们从不同学科的角度探讨人类的学习和其他智能特征的基础,并研究如何在科学原理上对此进行精确的描述,探讨用机器模拟人类智能的问题。在会议之前的《人工智能达特茅斯夏季研究项目提案》(A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence)中,麦卡锡首次提出了“人工智能”(Artificial Intelligence, AI)这个术语。所谓人工智能,也就是研究如何

收稿日期:2024-02-14

基金项目:新疆维吾尔自治区社会科学基金项目“维吾尔语中外来语借词的本土化、世俗化、现代化研究”(21BY140)的阶段性研究成果

作者简介:冯志伟,男,新疆大学中国语言文学学院天山学者,教授,博士生导师,主要从事计算语言学研究。

张灯柯,男,新疆大学中国语言文学学院中语系副主任,讲师,硕士生导师,主要从事计算语言学、少数民族语言信息化研究。

引用格式:冯志伟,张灯柯. 人工智能中的大语言模型[J]. 外国语文,2024(3):1-29.

使计算机去做过去只有人才能做的智能性工作。例如,过去的翻译都是由人来做的,1954年第一次机器翻译试验的成功,说明计算机也可以做翻译了,所以,机器翻译应当是一种人工智能的工作,机器翻译与人工智能有着不解之缘。

值得注意的是,麦卡锡在这个项目提案中,明确地提出要“研究语言与智能的关系”。他在这个项目提案中指出,在来年和夏季人工智能研究项目期间,他建议研究语言与智能的关系(McCarthy et al., 1955)。他认为,人类把语言作为处理复杂现象的手段,英语有许多属性是有利于处理这些复杂现象的。这些属性是:

- (1) 用非正式数学补充的英语论证可以做到简明扼要;
- (2) 英语具有普遍性,可以在英语中设置任何其他语言,然后在适当的地方使用这些语言;
- (3) 英语的使用者可以引用自己的说明,并陈述自己在解决问题方面的进展;
- (4) 如果英语完全形式化,不仅可以用来证明规则,而且还可以推导出一些猜测性的规则。(McCarthy et al., 1955)

麦卡锡在他的研究提案中还进一步指出,他希望尝试制定一种具有上述属性的语言,并希望使用这种语言可以对计算机进行编程。麦卡锡的这些观点是关于使用语言对计算机进行编程的早期论述,他试图把语言与计算机联系起来。由此可见,人工智能从诞生开始,就把研究的眼光敏锐地投向了语言。所以,人工智能与自然语言处理(Natural Language Processing, NLP)有着水乳交融的密切关系。

自然语言处理的模型按照发展顺序大致可以分为基于规则的语言模型(Rule-based Language Model, RLM)、基于统计的语言模型(Statistics-based Language Model, SLM)、基于神经网络的语言模型(Neural-Network-based Language Model, NLM)、大语言模型(Large Language Model, LLM)等,这些语言模型的研制与人工智能有着密切的关系(冯志伟等, 2024)。

本文将讨论人工智能中的大语言模型。大语言模型是人工智能的重大成果,由于大语言模型是用来处理人类自然语言的,当然也可以算是计算语言学的重大成果,值得语言学家关注。考虑到很多文科背景的语言学家对大语言模型的技术细节不太熟悉,本文尽量不使用复杂的数学公式和符号,以语言学家可以理解的方式,深入浅出地介绍大语言模型,增强语言学家对大语言模型的理解。大语言模型是处理语言的,以探索语言奥秘为己任的语言学家有什么理由不来关心大语言模型呢?

1 大语言模型的发展

大语言模型是一种由包含数百亿甚至数千亿参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方法,通过大量无标注文本进行训练。自 2018 年以来,Google、OpenAI、Meta、百度、华为等公司和研究机构都相继发布了包括 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。由于大语言模型处理的对象是自然语言,因此,大语言模型既是人工智能的重大成果,也是计算语言学的重大成果,是人工智能时代人类语言研究的光辉范例(冯志伟,2023:786)。

2019 年,大语言模型呈现爆炸式的增长局面,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,从而完成包括问答、分类、摘要、翻译、聊天等从自然语言理解到自然语言生成的各种任务。在这些任务中,大语言模型展现出掌握世界知识和理解自然语言的强大能力。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。

大语言模型的发展可以粗略地分为如下三个阶段。

第一阶段:这个阶段主要集中于 2017 年至 2019 年。其目标在于研制基础模型。2017 年,Google 的阿希什·瓦斯瓦尼(Ashish Vaswani)等人提出了 Transformer 架构,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。2018 年,动态词向量 ELMo 模型在双向预训练语言模型(pre-training language model)的基础上,使用动态词向量嵌入(Dynamic Word Vector Embedding)的方法,取得了初步的成绩。同年,Google 和 Open AI 分别提出了 BERT(Vaswani et al., 2017)和 GPT-1。BERT-Base 版本的参数量为 1.1 亿, BERT-Large 版本的参数量为 3.4 亿, GPT-1 的参数量为 1.17 亿。相比基于神经网络语言模型的参数量,这些大语言模型参数的数量级有了明显的提升。2019 年,Open AI 又发布了 GPT-2(Brown et al., 2020),其参数量达到了 15 亿。此后,Google 也发布了参数量规模为 110 亿的 T5 模型(Text-to-Text Transfer Transformer model)(Raffel et al., 2020)。2020 年,Open AI 进一步将语言模型参数量扩展到 1 750 亿,发布了 GPT-3。此后,我国也相继推出了一系列的大语言模型,包括清华大学的 ERNIE(THU)、百度的 ERNIE(Baidu)、华为的盘古- α 等。这个阶段的研究主要集中于大语言模型本身,研究范围包括编码器—解码器(Encoder-Decoder)等各种类型的模型结构。这些大语言模型通常采用预训练—微调范式,针对不同下游任务进行微调。

第二阶段:这个阶段集中于2019年至2022年,其目标在于进一步提升大语言模型的性能。由于大语言模型很难针对特定任务进行微调,研究人员进行了进一步的探索,试图在不针对单一的特定任务进行微调的情况下,发挥大语言模型的能力。2019年,瑞德福(Radford)等人就使用GPT-2模型研究了大语言模型在零样本(zero shot)情况下的处理能力。在此基础上,布朗(Brown)等人在GPT-3模型上研究了通过语境学习(In-Context Learning)进行少样本(few shot)学习的方法。将不同任务的少量有标注的实例拼接到待分析的样本之前输入大语言模型,使用大语言模型根据实例理解任务,给出正确结果,这样的研究展示出了非常强的能力,在有些任务中甚至超过了此前的有监督学习(Supervised Learning)方法。上述方法不需要修改大语言模型的参数,模型在处理不同任务时也不需要花费大量计算资源进行模型微调。但是仅依赖大语言模型本身,其性能在很多任务上仍然很难达到有监督学习效果,因此研究人员还提出了指令微调(Instruction Tuning)方案(Chung et al., 2022),将大量各类型任务,统一为生成式自然语言理解框架,并构造训练语料进行微调。大语言模型一次性可以学习数千种任务,并在未知任务上展现出了很好的泛化能力。2022年,欧阳(Ouyang)等人提出了InstructGPT算法(Ouyang et al., 2022),使用有监督微调再结合强化学习(reinforce learning),使用少量数据就可以使得大语言模型听从人类发出的指令。中野等人探索了结合搜索引擎的问题回答算法WebGPT(Nakano et al., 2021)。这些方法在直接利用大语言模型进行零样本和少样本学习,在此基础上逐渐扩展到利用生成式框架(generative frame)针对大量任务进行有监督微调的方法,有效地提升了大语言模型的性能。

第三阶段:这个阶段从2022年11月ChatGPT的发布开始一直延续到现在。在这个阶段,大语言模型得到了突破性的发展。ChatGPT通过人与机器之间简单的对话(Chat),利用一个大语言模型就可以实现问题回答、文稿撰写、代码生成、数学解题等过去自然语言处理系统需要大量的小模型订制开发才能分别实现的能力。它在开放领域问答、各类自然语言生成式任务以及人机对话上所展现出来的能力,远远超出大多数人的想象。2023年3月,GPT-4发布,相较于ChatGPT又有了非常明显的进步,并具备了多模态理解能力。GPT-4在多种基准考试测试上的得分高于88%的人类应试者,包括美国律师资格考试(Uniform Bar Exam)、法学院入学考试(Law School Admission Test)、学术能力评估(Scholastic Assessment Test, SAT)等。GPT-4展现了近乎“通用人工智能”(Artificial General Intelligence, AGI)的能力。各大公司和研究机构也相继发布了类似的系统,包括Google推出的Bard、百度的文心一言、科大讯飞的星火大模型、智谱的ChatGLM、复旦大学的MOSS

等。从 2022 年开始,大语言模型呈现出爆炸式的增长,各大公司和研究机构都在发布各种不同类型的大语言模型,出现了“百模大战”的局面。

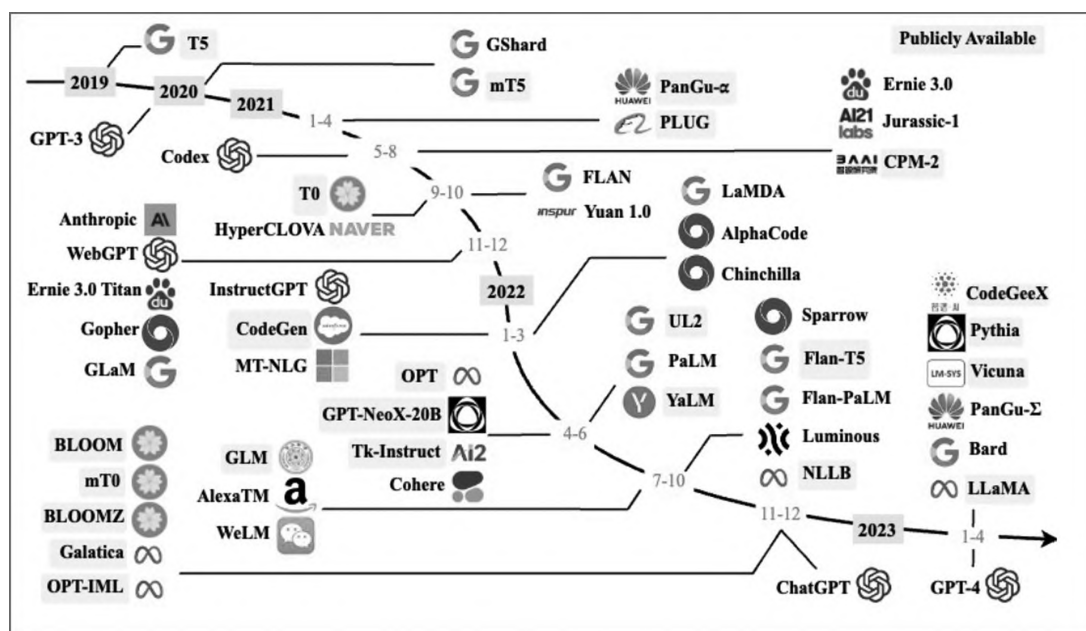


图 1 LLM 发展的时间线(2019—2023)(Zhao,2023)

图1按照时间线的顺序,给出2019年至2023年5月比较有影响力并且模型参数超过100亿的大语言模型(张奇等,2023:4-6)。

大语言模型实际上就是 N 元语法模型进一步的发展。尽管 N 元语法模型能缓解数据稀疏 (data sparseness) 的问题,但是自然语言极端复杂,具备无尽的可能性,再庞大的训练语料也难以覆盖所有的 N 元语法模型,因此,需要使用平滑技术 (smoothing) 来解决数据稀疏的问题,使得系统对所有可能出现的字符串都分配一个非零的概率值,从而避免零概率。平滑是指为了产生更合理的概率,对最大似然估计进行调整的一种方法,也称为数据平滑 (data smoothing)。平滑处理的基本思想是提高低概率,降低高概率,使整体的概率分布趋于均匀。

N 元语法模型从整体上来看与训练语料规模和模型的阶数有较大的关系,不同的平滑算法在不同情况下的表现有较大的差距。平滑算法虽然较好地解决了零概率问题,但是,N 元语言模型仍然有三个较为明显的缺点:

- (1) 无法给长度超过 N 的上下文建模;
- (2) 需要依赖人工设计规则的平滑技术;
- (3) 当 N 增大时,数据的稀疏情况随之增大,模型的参数量成指数级增加,并且模型受到数据稀疏问题的影响,其参数难以被准确地学习。

此外, N 元语法忽略了单词之间的相似性。因此, 基于分布式表示和基于神经网络的

语言模型逐渐成了研究热点。

约书亚·本吉奥(Yoshua Bengio)等人在2000年提出了使用前馈神经概率语言模型(Bengio et al., 2000),把概率引进前馈神经网络。词的独热编码(one hot encoding)被映射为一个低维稠密的实数向量(vector),称为词嵌入(word embedding)。此后,循环神经网络(Mikolov et al., 2010)、端到端记忆网络(Sukhbaatar et al., 2015)等神经网络方法都成功地应用于大语言模型的建模。相较于N元语法模型,神经网络方法可以在一定程度上避免数据稀疏问题,有些模型还可以避免对文本长度的限制,从而更好地给长距离依赖关系建模。深度神经网络需要采用有监督方法(supervised approach),使用标注数据进行训练,因此,语言模型的训练过程也不可避免需要构造训练语料。但是由于训练目标可以通过无标注文本直接获得,从而使得模型的训练仅仅需要大规模无标注文本就可以进行。语言模型也成了典型的自监督学习(self-supervised learning)任务。随着互联网的发展,非常容易获取超大规模文本数据,因此训练超大规模的基于神经网络的语言模型也成为可能。计算机视觉领域采用ImageNet对模型进行一次预训练(Deng et al. 2009),模型可以通过海量图像充分学习如何提取特征,然后再根据任务目标进行模型精调。受到计算机视觉这种研究的影响,自然语言处理领域基于预训练语言模型的方法也逐渐成为主流。以ELMo(Peters et al., 2018)为代表的动态词向量模型揭开了预训练语言模型的序幕。此后以GPT(Radford et al., 2019)和BERT(Devlin et al., 2019)为代表的基于Transformer模型(Vaswani et al., 2017)的大规模预训练语言模型的出现,使得自然语言处理全面进入了预训练-微调范式(pre-training and fine-tuning paradigm)的新时代。将预训练模型应用于下游任务时,不需要了解太多的任务细节,也不需要设计特定的神经网络结构,只需要“微调”预训练模型,使用具体任务的标注数据在预训练语言模型上进行监督训练,就可以显著地提升系统的性能。

2020年,Open AI发布了由包含1 750亿参数的神经网络构成的生成式大规模预训练语言模型GPT-3,获得了极大的成功。由于大语言模型的参数量巨大,如果在不同任务上都进行微调需要消耗大量的计算资源,因此有必要对预训练-微调范式进行改进。研究人员发现,通过语境学习(Incontext Learning, ICL)等方法,直接使用大规模语言模型也可以在很多任务的少样本场景下取得很好的效果。此后,研究人员提出了面向大规模语言模型的提示词(prompt)学习方法、模型即服务范式(Model as a Service, MaaS)、指令微调(Instruction Tuning)等方法,在不同任务上都取得了很好的效果。与此同时,Google、Meta、百度、华为等公司和研究机构都纷纷发布了包括PaLM(Chowdhery et al., 2022)、LaMDA(Thoppilan et al., 2022)、T0(Sanh et al., 2021)等大规模语言模型。2022年11月,

ChatGPT 的出现,将大语言模型的能力进行了充分的展现,引发了大语言模型研究的热潮。

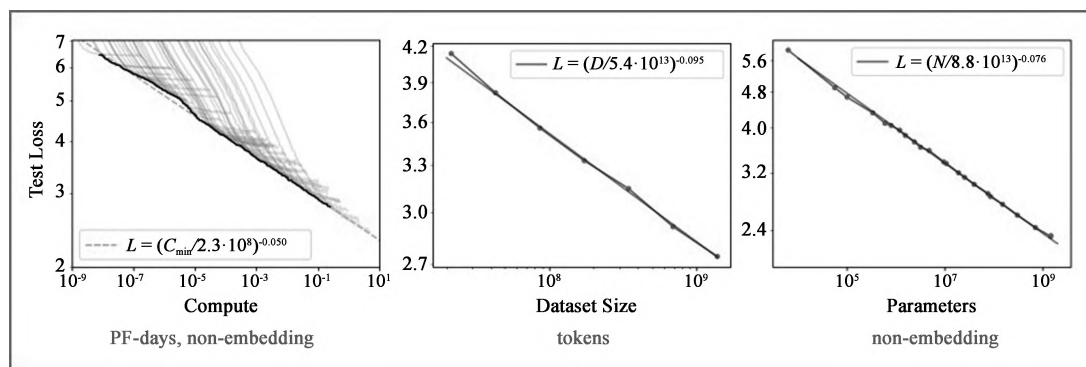


图2 缩放法则 (Kaplan, 2020)

卡普兰等人在《神经语言模型的缩放法则》(Scaling laws for neural language models)一文中提出了“缩放法则”(Scaling Laws),指出语言模型的性能依赖于模型的规模,包括计算量(Compute)、数据集大小(Data Size)和参数量(Parameters),模型的效果会随着这三者的增加而提高,而模型的损失(Loss)值随着计算量的规模、数据集的规模、参数量的增大而线性降低。如图2所示。这意味着语言模型的能力是可以根据这三个变量来估计的,提高模型的计算量、扩大数据集规模、提高参数量,都可以降低模型的损失,使得模型的性能可预测地提高。“缩放法则”为继续提升语言模型的规模给出了定量分析依据。

由于大语言模型主要是建立在预训练语言模型和 Transformer 模型基础之上的。下面,我们进一步讨论预训练语言模型和 Transformer 模型及其相关的 Elmo、Bert、GPT 等大语言模型。

2 预训练语言模型

在当前的神经自然语言处理研究中,语言数据资源的匮乏是一个非常严重的问题,对于自然语言处理而言,几百万个句子的语料都不能算作是大数据(big data),商用的神经自然语言处理系统基本上都要数千万个句子甚至数亿个句子的大数据作为训练语料。如果语言数据匮乏,自然语言处理的质量是难以保证的。为了解决语言数据匮乏的问题,学者们使用迁移学习的方法,开始探讨小规模语言数据资源下自然语言处理的可行性问题,形成了一种自然语言处理的新范式——预训练语言模型,如图3所示。

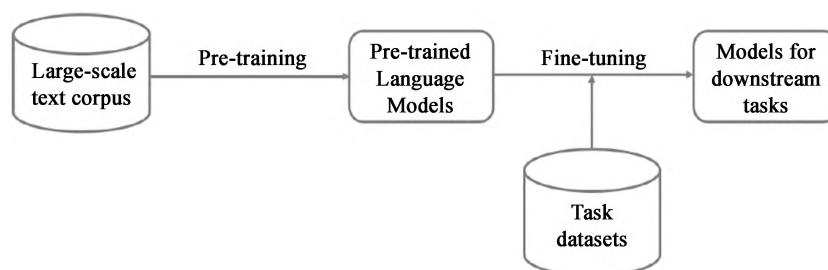


图3 预训练语言模型(冯志伟,李颖,2021:1-14+112)

这种语言模型使用大规模的文本语料库数据 (large-scale text corpus) 进行“预训练” (pre-training), 建立“预训练语言模型”, 然后使用面向特定任务的小规模语言数据集 (task datasets), 根据归纳迁移学习的原理进行“微调” (fine-tuning), 形成“下游任务的模型” (models for downstream tasks)。这样的预训练语言模型新范式使得研究者能够专注于特定的任务, 而适用于各种任务的通用的预训练语言模型可以降低自然语言处理系统的研制难度, 从而加快了自然语言处理研究创新的步伐。

2009年, 谷歌公司通过在覆盖100多种语言的超过250亿句子对、超过500亿参数语言资源的基础上, 使用预训练模型训练了一个神经机器翻译系统, 突破了多语言神经机器翻译研究的极限。他们研制了一种用于“大规模多语言的大规模神经机器翻译” (Massively Multilingual, Massive Neural Machine Translation, M4) 的方法 (Arivazhagan et al., 2019), 这样的M4方法在低资源语言和高资源语言上都表现出了巨大的质量提升, 可以轻松地适应不同的领域以及不同的语言, 同时在跨语言下游迁移任务上表现出很高的效率。这样的新进展是令人振奋的。

3 Transformer 模型

2017年6月, 谷歌公司在他们发表的论文《注意力就是你们所需要的一切》 (Attention Is All You Need) (Vaswani et al., 2017) 中提出了一个完全基于注意力机制 (attention) 的预训练语言模型, 叫作 Transformer^①, 这个模型抛弃了在此之前的其他采用注意力机制的模型保留的循环神经网络 (Recurrent Neural Network, RNN) 结构与卷积神经网络 (Convolutional Neural Network, CNN) 结构, 将核心完全使用注意力机制。Transformer 是完全基于注意力机制的模型, 在各项任务的完成和性能发挥方面表现优异, 因此成为自然语言处理的重要基准模型。

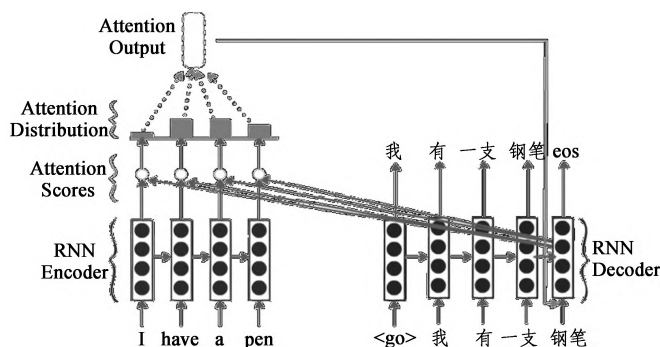


图4 用循环神经网络做机器翻译

在 Transformer 出现之前, 神经网络机器翻译大多采用基于循环神经网络 (RNN) 的模

① Transformer 这个术语, 学术界的译名有“转换器”“变形器”“变换器”“变形金刚”多种, 至今没有恰当的中文译名, 众说纷纭, 因此, 学术界只好采用英文原形 Transformer。本书尊重学术界的做法, 采用英文原形 Transformer。

型。如图 4 所示。

这样的循环神经网络采用的是一种异步的序列到序列模式。例如,在英汉机器翻译中,循环神经网络编码器(RNN Encoder)对于输入的英语句子“I have a pen”进行编码,经过注意力打分(Attention Scores)给注意力分派权重,再经过注意力分布(Attention Distribution)得到注意力输出(Attention Output),最后由循环神经网络的解码器(RNN Decoder)进行解码,得到汉语译文“我有一支钢笔”。

循环神经网络的隐藏层信息不仅取决于当前的输入层信息,也包括输入层前一步的信息,也就是要把输入层前一步的信息灌给当前的输入层。循环神经网络虽然建模序列很强大,但这种异步的序列到序列模式训练起来非常缓慢,如果文本中的长句子很多,则需要更多的处理步骤,并且其繁复循环的结构也使模型的训练非常困难,神经机器翻译效果也因此而不能尽如人意。

与循环神经网络相比,Transformer 不需要循环,而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。

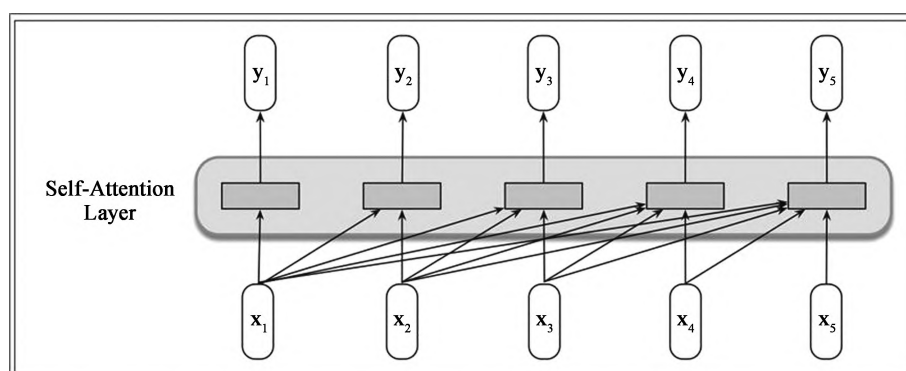


图 5 自注意力机制(Ashish,2017)

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。

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 的训练速度比循环神经网络快得多,而且它的自然语言处理效果也比循环神经网络好得多。Transformer 的结构如图 6 所示。

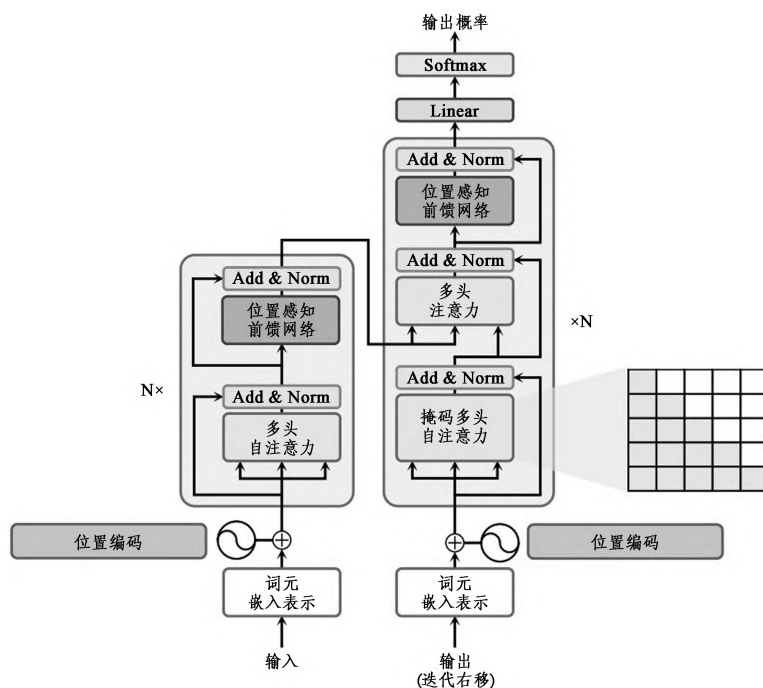


图 6 Transformer 的结构 (Ashish, 2017)

在图 6 中,左半部分是编码器 (encoder),右半部分是解码器 (decoder)。编码器由六个相同的层 (layers) 组成 ($N = 6$),每一层包括两个子层 (sub-layers),第一个子层包括一个“多头自注意力层” (Multi-Head Self-Attention),第二个子层包括一个“位置感知前馈网络层” (Feed Forward),其中每一个子层都加了“求和” (Add) 与“归一化” (Norm)。在自然语言处理中,从编码器输入的源语言句子首先经过多头自注意力层,在对每一个“词元” (token)^①进行编码时,这一层帮助编码器关注输入语言句子中的其他词元。多头自注意力层的输出会传递到前馈网络层中,每一个位置的词元对应的前馈网络层都是一样的。

我们以神经机器翻译为例来说明 Transformer 的工作原理。在神经机器翻译中,我们首先把每一个输入的词元使用词嵌入算法转换为词向量 (word vector)。在《注意力就是你们所需要的一切》这篇论文中,每一个词元都被嵌入为 512 维的向量,根据这篇论文,我们这里就假设每一个单词被嵌入到 512 维的向量中。词嵌入过程只发生在最底层的编码器中。所有的编码器都有一个相同的特点,它们都要接受一个向量列表,列表中每一个向量的大小为 512 维。在最底一层的编码器中,输入的是词向量,但是在之后的其他编码器中,这个词向量就是下一层编码器的输入,这样就可以形成一个向量列表。向量列表的大小是我们设置的参数,一般就是我们训练集中最长句子的长度。将输入序列进行词嵌入之后,

① 在大语言模型中,计算机处理的语言单位不是 word,而是 token。在语言模型中,token 是处理和生成文本或代码的基本单位,这种含义的 token 目前还没有确切的中文译文,为了称说方便,我们这里暂时把这个 token 翻译为“词元”。

源语言中的每一个词元都会流经编码器中的多头自注意力层和前馈网络层。

相比于编码器端,解码器端要更复杂一些。具体来说,解码器的每个 Transformer 模块的第一个自注意力子层额外增加了“注意力掩码”(Masked Attention),对应图中的“掩码多头自注意力”(Masked Multi-Head Self-Attention)部分。这主要是因为,在翻译的过程中,编码器端主要用于编码源语言序列的信息,而这个序列是完全已知的,因而编码器仅需要考虑如何融合上下文语义信息即可。而解码器端则负责生成目标语言序列,这一生成过程是自回归的,也就是说,对于每一个词元的生成过程,仅有当前词元之前的目标语言序列是可以被观测的,因此这一额外增加的“掩码”(Masked)是用来掩盖后续的文本信息,以防模型在训练阶段直接看到后续的文本序列进而无法得到有效的训练。所以,在“多头自注意力”前面加上“掩码”这个修饰语,就是要防止在训练的时候使用未来要输出的词元。因为在训练的时候,前面的词元是不能参考后面将要生成的词元的,要把后面的词元屏蔽起来。掩码多头自注意力层是为了使得解码器看不见未来的信息。也就是说,对于一个序列,在时间步(time step)为 t 的时刻,我们的解码输出应该只能依赖于时刻 t 之前的输出,而不能依赖时刻 t 之后的输出。因此我们需要掩码多头注意力层,从而把时刻 t 之后的信息都屏蔽起来。此外,解码器端还额外地增加了一个多头注意力(Multi-Head Attention)模块,同时接收来自编码器端的输出以及当前 Transformer 模块的前一个掩码自注意力层的输出。

基于上述的编码器和解码器构架,在机器翻译时,待翻译的源语言文本,首先经过编码器端的每个 Transformer 模块对其上下文语义的层层抽象,最终输出每一个源语言词元的上下文相关表示。解码器端以自回归的方式生成目标语言文本,即在每个时间步 t ,根据编码器端输出的源语言文本表示,以及前 $t-1$ 个时刻生成的目标语言文本,生成当前时刻 t 的目标语言。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词元与词元之间联系的密切程度。

多头自注意力子层可以把相关的词元融入正在处理的词元中,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 *animal* 在之外,还有好几个其他的词元,它们也有可能成为 *it* 的所指对象。但是,由于 Transformer 有“多头注意力子层”,当模型在处理 *it* 这个词元的时候,多头注意力子层会把所有相关的词元融入我们正在处理的词元 *it* 中,从而允许 *it* 和 *animal* 建立起比其他词元更加密切的联系。如图 7 所示。

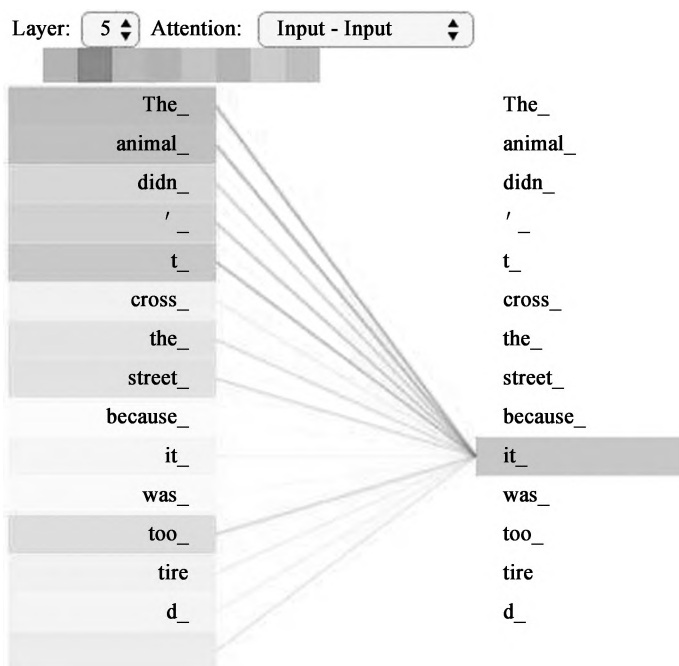


图7 自注意力机制建立 it 和相关词元的联系 (Ashish, 2017)

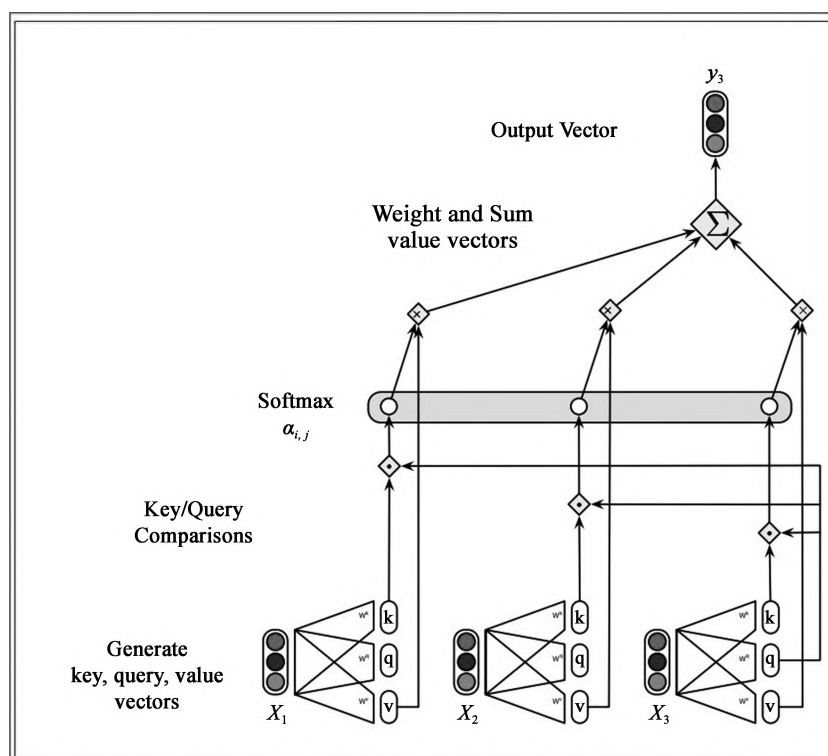
在图7中,自注意力机制可以建立起 it 和相关词元之间的联系。当在编码器的第5层 (Layer 5) 中对 it 在这个词元进行编码时,自注意力机制会关注 The animal,把 The animal 的一部分表示编入 it 的编码中。从图7中不难看出,尽管 it 与很多词元都与联系,但是, it 与 The animal 的联系最为密切。

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

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

这个公式中,Q表示查询向量,K表示键向量,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、键向量K、值向量V,然后比较它们之间的查询向量与键向量(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 (Ashish, 2017)

循环神经网络的最大优点就是能够在时间序列上对数据进行抽象,重视处理对象的位置顺序,Transformer 不再采用循环神经网络,这是一个缺憾。Transformer 为了弥补这样的缺憾,在编码器和解码器中,都进行了位置编码(Positional Encoding,简称 PE),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:

$$PE_{(pos, 2i)} = \sin(pos/10000^{2i/d_{model}})$$

$$PE_{(pos, 2i+1)} = \cos(pos/10000^{2i/d_{model}})$$

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, d_{model} 表示模型的维 (dimensionality of model), 在位置编码 (Positional Encoding) 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, 值得我们进一步研究(张子豪 等, 2023)。

在机器翻译时, 输入的源语言数据经过编码器和解码器处理之后, 再经过线性变换层 (Linear) 和 softmax 层的归一化处理, 得到目标语言的输出概率(output probabilities)。如图 9 所示。

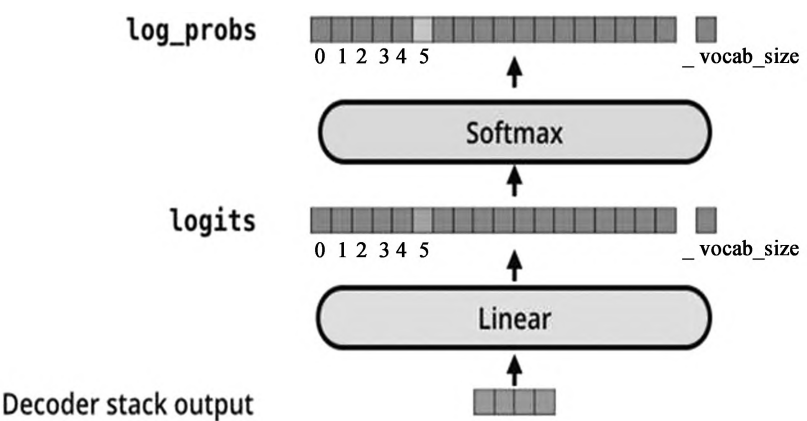


图 9 线性变化层和 softmax 层 (Ashish,2017)

线性变换层 (Linear) 是一个简单的全连接神经网络,它可以把解码器产生的向量投射到一个叫作“对数几率”(logits)的向量里。如果我们的模型从训练集当中学习 10 000 个不同的英语词元,因而对数概率向量就是 10 000 个单元格长度的向量,每一个单元格对应于某个英语词元的分数。接下来的 Softmax 层把这些分数转化成概率(log_probs)。概率最高的单元格被选中,它对应的词元被作为这个时刻的输出。

在 Transformer 中,六个编码器与六个解码器的协作方式如图 10 所示。例如,在法语 - 英语的机器翻译中,输入法语句子“Je suis étudiant”(我是一个学生),经过 Transformer 处理,在输出端就可以得到英语的译文“I am a student”。

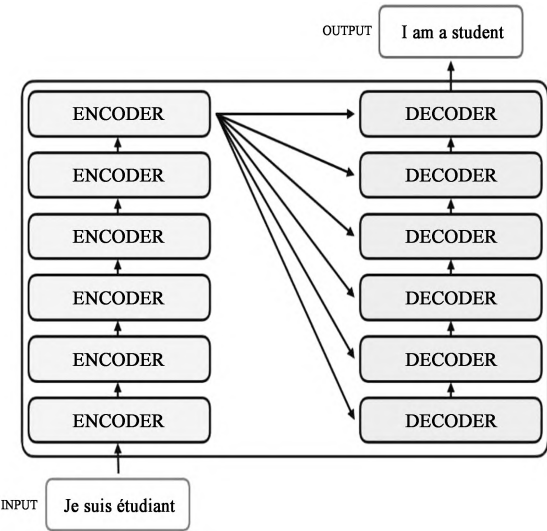


图 10 用 Transformer 进行法语 - 英语机器翻译 (Ashish,2017)

大语言模型在近两年来基本形成了一套近乎完备的技术体系,包括词嵌入、编码器—解码器的端对端语言模型、注意力机制、Transformer,以及 BERT 预训练模型等。这一套技术体系有力地促进了大语言模型在信息搜索、阅读理解、机器翻译、文本分类、智能问答、智能对话、网络聊天、信息抽取、自动文摘、文本生成等重要领域的应用,预示着大语言模型进

入了大规模工业化实施的时代。

4 动态词向量嵌入模型 ELMo

词向量主要利用语料库中词之间的分布信息 (distribution), 学习词语的向量表示。因此, 根据给定的语料库学习得到的词向量是恒定不变的, 可以认为是“静态”的, 不跟随上下文而发生变化。然而, 自然语言中词语往往具有多种语义, 在不同的上下文或语境下会具有不同的语义。

针对这个问题, 研究者们提出了动态词向量嵌入 (Dynamic Word Vector Embedding) 方法, 也称为上下文相关的词向量嵌入 (Contextualized Word Embedding) 方法, 使用这样的方法, 一个单词 (word) 的向量 (vector) 通过其所在的上下文计算获得, 并随着上下文的不同而动态地发生变化。

语言的动态词向量嵌入模型 ELMo (Embeddings from Language Models) 是一种双向的预训练语言模型, 从两个方向进行语言模型的建模: 从左到右前向建模和从右到左后向建模。前向语言模型负责从左到右的前向建模, 后向语言模型负责从右到左的后向建模。双向建模可以给出更好的上下文表示, 文本中的每个词能同时利用其左右两侧文本的上下文信息。ELMo 的神经网络结构主要包含输入层, 隐藏层和输出层三个部分, 如图 11 所示。图中的 w 表示单词 (word), v 表示向量 (vector), BOS 表示句首 (Begin of Sentence), EOS 表示句末 (End of Sentence)。

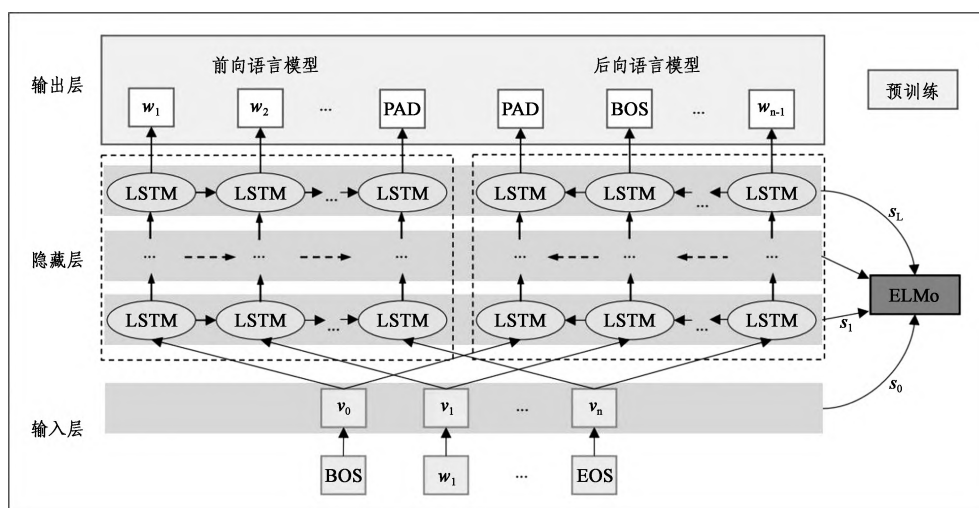


图 11 ELMo 的神经网络结构 (Matthew, 2018)

图 11 的 ELMo 中, 在输入层进行词向量输入, 在隐藏层使用长短时记忆网络 (Long Short-Term Memory, LSTM) 进行双向建模, 在输出层使用前向语言模型和后向语言模型动态地进行预训练。在预训练中进行下游任务时, ELMo 将所有的向量整合成一个向量, 整合的方式根据任务而定, 最简单的情况是直接使用最后一层表示。因为每层 LSTM 学习到的

信息不相同,对于不同任务来说,每层特征的重要性也不尽相同,因此更普遍的做法是根据任务所需信息,对每层的特征进行加权得到与有关单词对应的 EMLo 向量。

5 双向编码表示模型 BERT

2018 年 7 月,谷歌公司发布了《通用 Transformer》(Universal Transformer)一文,对 Transformer 进行了改进,进一步提升了翻译速度,其速度比循环神经网络中的顺序循环更快,也比 Transformer 更加强大,而且具有通用性。2019 年,谷歌公司雅各布·德夫林(Jacob Devlin)等人研制成功 BERT(Bidirectional Encoder Representations from Transformers),这是一种基于 Transformer 的双向编码表示模型。BERT 在 11 项不同的自然语言处理测试中创造出最佳成绩,为自然语言处理带来了里程碑式的改变(Devlin et al., 2019)。这是近年来自然语言处理引人注目的成就。

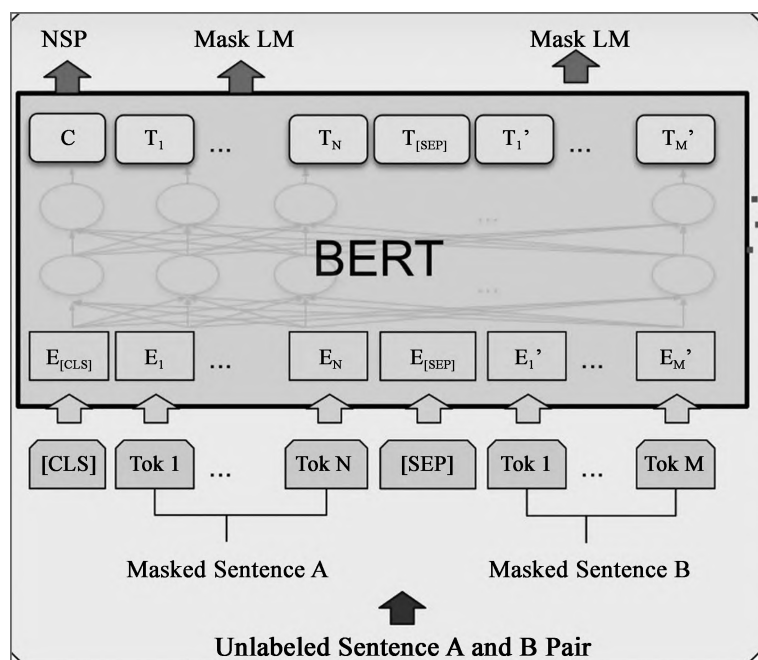


图 12 BERT 的结构(Jacob,2019: 4171-4186)

BERT 由输入层,编码层和输出层三部分组成。在输入层,把未标注过的句子(Unlabeled Sentences)进行屏蔽处理,使之成为掩码的句子(Masked Sentence)。然后进行嵌入处理,得到嵌入序列(Tok 1,...,Tok N 与 Tok 1,...,Tok M)。编码层由多层 Transformer 编码器组成。在预训练时,模型的最后有两个输出层 NSP 和 Mask LM,分别对应了两个不同的预训练任务:下一句预测(Next Sentence Prediction, NSP)和掩码语言模型(Mask Language Modeling, Mask LM,有时也简写为 MLM)^①。在图 12 中,Tok 表示词元,E 表示嵌

^① Mask Language Model 又可翻译为“屏蔽语言模型”,本文翻译为“掩码语言模型”。

入(Embedding),T 表示 Transformer 模块(Transformer block)。

BERT 利用掩码语言模型构造了基于上下文预测中间词的预训练任务,相较于传统的语言模型建模方法,BERT 可以进一步挖掘上下文所带来的丰富语义。例如,在图 13 中,输入句子为“This is going to be so long”,我们把 to 遮盖,该句子成为“This is going [mask] be so long”,输入 BERT,经过前馈神经网络(Feed-Forward Neural Network, FFNN)和 SOFTMAX 处理,预测出被遮盖的词是“to”。

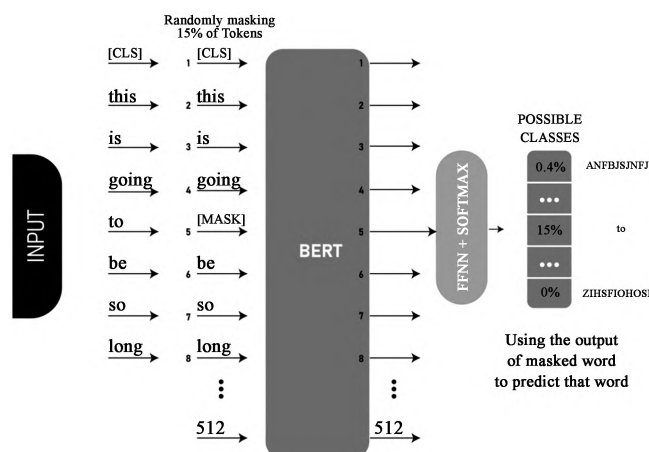


图 13 BERT 的掩码语言模型(Jacob,2019: 4171-4186)

BERT 所采用的神经结构由多层 Transformer 编码器组成,通过掩码模型使其具备了很强的预测能力。这意味着在编码过程中,每个位置都能获得所有位置的信息,而不仅仅是历史位置的信息。BERT 是一种基于 Transformer 的双向编码表示模型,BERT 是以 Transformer 的结构为基础的,其目标在于通过左右上下文共有的条件计算来预训练无标记文本的深度双向表示。因此,经过预训练的 BERT 模型,只需要一个额外的输出层就可以进行微调(fine-tuning),从而为各种自然语言处理任务生成最新的模型。

BERT 的预训练是在包含整个维基百科的大规模语料库(25 亿单词)和图书语料库(8 亿单词)中进行的。BERT 是一个“深度双向”模型,所谓“深度双向”就意味着 BERT 在预训练时要从所选文本的左右双向的上下文中汲取信息。双向性是 BERT 模型最显著的特点。例如,我们来看下面两个英语句子:We went to the river **bank** 和 I need to go to **bank** to make a deposit。其中的 bank 是一个多义词,它有两个含义,一个是“银行”,一个是“河岸”。如果我们只看上面两个句子中 bank 的左边部分,也就是:

We went to the river —

I need to go to—

我们可以预测出“—”的意思应当是“河岸”,因为第一句中有 river 这个单词,但是第二句中如果预测为“河岸”,则是错误的。因此只根据左侧上下文不能做出正确判断。如果我们只看上面两个句子中 bank 的右边部分,也就是:

—

— to make a deposit.

我们可以预测出“—”的意思是“银行”,因为第二句中有 deposit 这个单词,但是第一句中如果预测为“银行”,则是错误的。因此只根据右侧上下文不能做出正确判断。由此可见,在自然语言分析时,仅仅根据 bank 一个方向的上下文,是不能确定 bank 的准确含义的,必须根据 bank 左侧和右侧双向的上下文,才能准确地预测 bank 的含义。这就是 BERT 采用双向性的原因所在。显而易见,这个原因的深层根据来自语言学。

BERT 提供了简单($BERT_{base}$)和复杂($BERT_{large}$)两个模型,对应的超参数分别如下:

$BERT_{base}$: $L = 12, H = 768, A = 12$, 参数总量为 110M(1.1 亿);

$BERT_{large}$: $L = 24, H = 1024, A = 16$, 参数总量 340M(3.4 亿)。

在上面的超参数中, L 表示层数(Layer number),也就是 Transformer blocks(简称为 Trm)的数量, H 表示隐藏层(Hidden layer)的数量, A 表示多头注意力(Multi-Head Attention)中的自注意力(Self-Attention)的数量。如图 14 所示。

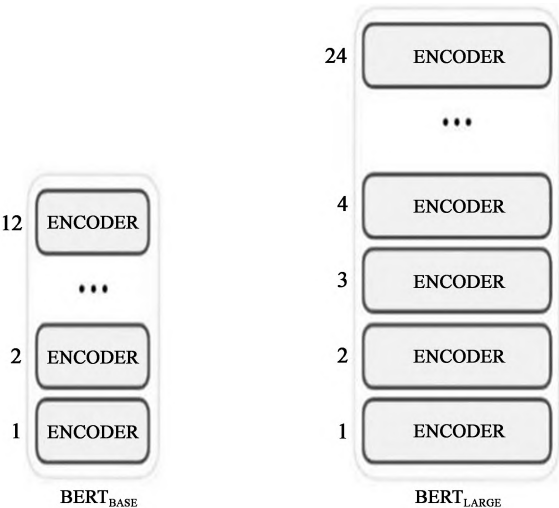


图 14 $BERT_{base}$ 和 $BERT_{large}$ (Jacob,2019: 4171-4186)

在图 14 中,每一个嵌入 E 都由三个嵌入组成:

(1)位置嵌入(Position Embeddings):BERT 学习并使用位置嵌入来表达单词在句子中的位置,如 E_1, E_2, E_3 等;

(2)片段嵌入(Segment Embedding):BERT 还可以将句子偶对作为问答任务的输入,BERT 在学习了第一个句子的嵌入和第二个句子的嵌入之后,就可以帮助模型把不同的片段区分开来。如把 E_A 和 E_B 区分开来;

(3)词元嵌入(Token Embeddings):BERT 从标记词汇表(Word Piece)中学习特定的词元(Token)的嵌入。

对于特定的标记, BERT 的输入表示就是位置嵌入、片段嵌入和词元嵌入的总和。如图 15 所示。

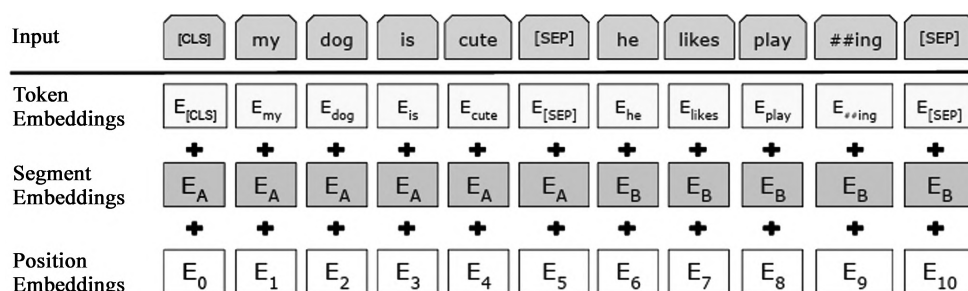


图 15 BERT 的嵌入 (Jacob, 2019: 4171-4186)

把这些嵌入处理步骤综合起来,使得 BERT 具有极强的通用性,在预训练时,不必对模型做太大的修改,就可以在多种自然语言处理的任务上训练 BERT,使其具有强大的功能。

BERT 可以支持维基百科上的 100 多种语言的处理,也可以支持中文处理。BERT 在工作时,首先使用标记词汇表 (Word Piece) 和 30 000 个词元的词汇表进行嵌入。用##表示分词。然后进行位置嵌入,支持的序列长度最多为 512 个词元。每个序列的第一个词元始终是特殊分类嵌入 (记为 [CLS])。句子偶对被打包成一个序列。以两种方式区分句子。首先,用特殊的分割 (separation) 标记 ([SEP]) 将它们分开。然后,添加一个 sentence A 嵌入到第一个句子的每个词元中,再添加一个 sentence B 嵌入到第二个句子的每个词元中。对于单个句子输入,只使用 sentence A 嵌入。在图 15 中,我们得到的输入形式为:

[CLS] my dog is cute [SEP] he likes play ##ing [SEP]

BERT 进行预训练时,使用了掩码机制 (Masking approach)。为了训练一个深度双向表示 (deep bidirectional representation), BERT 采用了一种简单的方法:随机地对部分输入的词元 (token) 进行掩码处理 (masking),然后只预测那些被掩码的词元。这样的处理机制就是“掩码语言模型” (MLM),又叫作“完形模型” (Cloze model)。

BERT 还有预测下一个句子的功能。自然语言处理的许多重要的下游任务,如智能问答 (Question-Answering, 简称 QA) 和自然语言推理 (Natural Language Inference, 简称 NLI) 都需要理解两个句子之间的关系。为了理解两个句子之间的关系,可以预先训练一个二进制的下句预测任务,这一任务可以使用单语语料库来实现。具体地说,当选择句子 A 和句子 B 作为预训练样本时,句子 B 有 50% 的可能是句子 A 的下一个句子,也有 50% 的可能是来自语料库中的随机句子。例如,如果我们有如下的句子偶对:

Input = [CLS] the man went to [MASK] store [SEP]

he bought a gallon [MASK] milk [SEP]

因后一个句子“he bought a gallon [MASK] milk [SEP]”可能与前一个句子“[CLS] the man went to [MASK] store [SEP]”在语义上有联系, BERT 可以打上标签 Label = IsNext,

表示后一个句子可能是前一个句子的 Next。

如果我们有如下的句子偶对:

Input = [CLS] the man [MASK] to the store [SEP]

penguin [MASK] are flight ##less birds [SEP]

由于后一个句子“penguin [MASK] are flight ##less birds [SEP]”与前一个句子“Input = [CLS] the man [MASK] to the store [SEP]”在语义上没有联系,BERT 可以打上标签 Label = NotNext,表示后一个句子不可能是前一个句子的 Next。这样一来,BERT 就可以完全随机地选择 NotNext 语句,最终的预训练模型在这个下游任务上取得了 97% ~ 98% 的准确率。这些预处理步骤综合起来,使 BERT 具有很强的通用性。这意味着,即使不对模型的结构进行任何重大更改,也可以轻松地将 BERT 应用到多种自然语言处理的下游任务上。

BERT 在机器阅读理解顶级水平测试 SQuAD1.1 中取得惊人的成绩:它在衡量指标上全面超越人类。并且还在 11 项不同的自然语言处理测试中创造了最佳的成绩,包括将 GLUE 基准推进到了 80.4% (绝对改进率为 7.6%),将 MultiNLI 准确度推进到了 86.7% (绝对改进率为 5.6%)。这是自然语言处理最激动人心的成就。BERT 的预训练是在大数据(big data)中进行的,这些数据不需要进行人工标注,无标注的数据本身已经蕴藏了丰富的语言信息,研究者没有必要再进行任何的标注,BERT 所用的语言数据包含整个维基百科的 25 亿单词的大规模语料库和八亿单词的图书语料库,这些都是没有标注的数据。这充分地说明了语言数据资源对于自然语言处理的关键性作用,也说明了无标注的数据本身已经包含了丰富的语言信息。

仔细想来,我们人类在讲话或写文章时,不是也需要构思吗?其实,构思也就相当于一种标注,在自然语言处理中,人工标注实在是没有必要的。几千年来,无数的语言学家皓首穷经、苦苦追求的语言规律,其实就隐藏在语言数据中,语言数据自身就包含了这些规律,现在有了大语言模型的新技术,计算机已经有能力从语言数据中学习到这些规律,再进行人工标注也就是多此一举了。

6 生成式预训练模型 GPT

由 OpenAI 公司开发的基于 Transformer 的生成式预训练模型(Generative Pre-Trained Transformer, GPT)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 GPTs。GPTs 利用 Transformer 模型,从语言大数据中获取了丰富的语言知识,GPTs 在语言生成任务上达到了相当高的水平。这样一来,GPTs 便成了自然语言处理研究的最重要的大语言模型。

GPTs 系列的训练参数越来越多,性能越来越好。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。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,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、高质量的各种语言数据作为学习内容,参数总量是 GPT-2 参数的 117 倍。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 就可以理解到用户意图是要做英语到法语的翻译,继而给出对应的法语单词译文。后来,OpenAI 又在此基础上于 2022 年 1 月开发出 InstructGPT,形成了“基于人类反馈的强化学习”(Reinforcement Learning from Human Feedback, RLHF)方案,通过人类的反馈来提高系统的性能。接着又进行有监督微调(Supervised Fine Tuning, SFT),清理文本数据,力争把有害的、错误的、不合乎伦理规范的内容减少到最低限度。

2022 年 11 月,OpenAI 在此基础上开发出 ChatGPT。ChatGPT 的训练语料高达 100 亿个句子,夜以继日地训练了三年,训练的总文本超过 45T。这个时候的 ChatGPT 可以通过使用大量的训练数据来模拟人的语言行为,生成人类可以理解的文本,并能够根据上下文语境,提供恰当的回答,甚至还能做句法分析和语义分析,进行逻辑推理,帮助用户调试计算机程序,写计算机程序的代码,而且能够通过人类反馈的信息,不断改善生成的功能,已经达到了很强的自然语言生成能力。

ChatGPT 使用 Transformer 进行训练,在训练过程中,使用海量的自然语言文本的无标注数据来学习单词的嵌入表示以及上下文之间的关系,形成知识表示(knowledge representation)。在训练中进行有监督微调(SFT)和基于人类反馈的强化学习(RLHF)处理。一旦训练完成,知识表示就被编码在神经网络的参数中,可以使用这些参数来生成回答。当用户提出问题时,神经网络就根据已经学习到的知识生成回答,返回给用户。

从 GPT-1 到 ChatGPT 的发展过程,如图 16 所示。

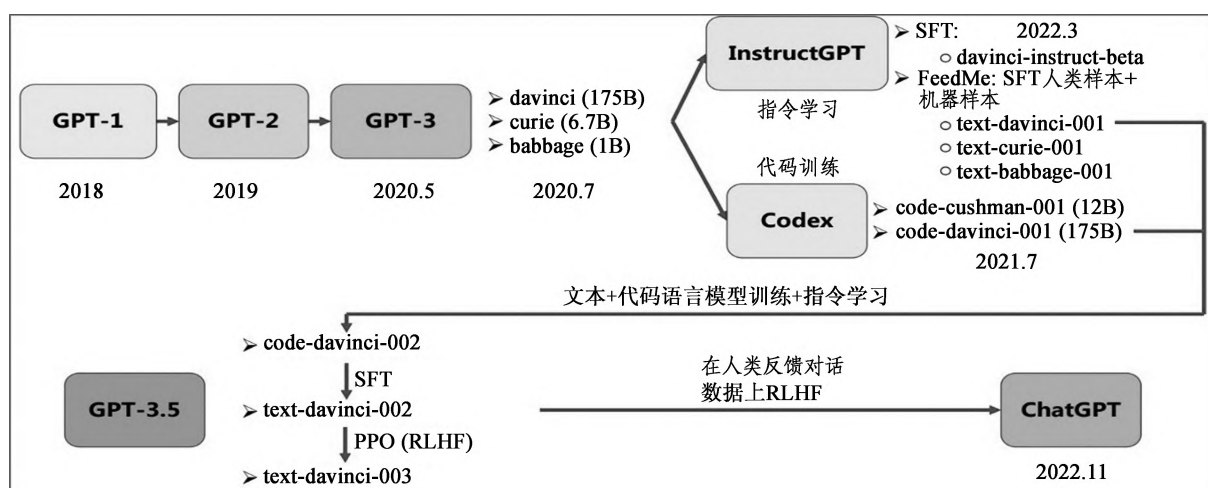


图 16 GPT 系列的发展过程 (Zhao, 2023)

从图 16 可以看出,OpenAI 公司于 2018 年研制了 GPT-1,于 2019 年研制了 GPT-2,于 2020 年 5 月研制了 GPT-3,2020 年 7 月分别研制了 GPT-3 中的 davinci(达芬奇), curie(居里), babbage(巴贝奇),2022 年 3 月研制了 InstructGPT,进行文本和代码的语言模型训练;还研制了 Codex,Codex 有 120 亿参数,以 GPT-3 系列模型作为初始模型继续进行代码微调训练。在此基础上研制成 GPT-3.5,接着进行有监督微调(SFT)和基于人类反馈的强化学习(RLHF),于 2022 年 11 月推出 ChatGPT。ChatGPT 比 GPT-3 更进一步,已经进化到具备执行自然语言指令的能力,用户不必给出示例,只要使用自然语言给出指令,ChatGPT 就可以理解用户意图。例如,用户只要直接用自然语言告诉 ChatGPT 把某个英语单词译成法语,ChatGPT 就可以执行,并给出翻译结果。ChatGPT 可以根据上下文提示,自动理解并执行各类任务,不必更新模型参数或架构。

2022 年 11 月 30 日,ChatGPT 开放公众测试,真正实现了完全自主的“人工智能内容生成”(AI Generated Content, AIGC),包括文本生成、代码生成、视频生成、文本问答、图像生成、论文写作、影视创作、科学实验设计等。现在的 ChatGPT 是由效果比 GPT-3 更强大的 GPT-3.5 系列模型提供支持的,这些模型使用微软 Azure AI 超级计算基础设施上的海量文本和代码数据进行训练。

交互式是 ChatGPT 的一大优点,用户可以自如地与 ChatGPT 进行多轮对话,自然而且流畅,ChatGPT 的回答是连续的、稳定的、一致的,用户与 ChatGPT 对话,就像是与朋友聊天。ChatGPT 具有高度的可扩展性和灵活性,可以根据不同需求进行二次开发和定制;可以快速地大量数据中学习,并且在后续应用中持续地更新和优化;可以应用于在线客服、虚拟助手、教育培训、游戏娱乐等多个领域,在这些领域中为用户提供高效、便捷、个性化的服务和体验。ChatGPT 通常需要进行训练和调试,以达到最佳的对话效果,可以利用第三方工具或平台来集成 ChatGPT,并将其应用于具体场景中。

ChatGPT 推出 5 天后,注册用户就超过百万;推出短短两个月后,月活跃用户就超过一亿。TikTok 月活跃用户超过一亿用了九个月时间, Twitter 月活跃用户超过一亿用了 90 个月时间, ChatGPT 打破了历史纪录,遥遥领先,引起了全球网民的广泛注意,在大语言模型时代掀起了一场史无前例的、波澜壮阔的海啸。如图 17 所示。



图 17 月活跃用户达到 1 亿所用时间比较(国泰君安证券研究 2024)

ChatGPT 的推出引起了海啸般的轰动。成千上万的用户从不同角度对它进行了应用体验,关于它的各种说法也是满天飞。有人说, ChatGPT 已经拥有通用人工智能(Artificial General Intelligence, AGI),有人说,很多岗位上的人都会被 ChatGPT 取代……在技术革命的前夜,总是圣歌与哀声并起。ChatGPT 是一个伟大的人工智能项目,它使用指令学习、有监督微调、基于人类反馈的强化学习、人工智能内容生成等一系列创新技术,使大语言模型在之前版本的基础上实现了飞跃式的发展,在意图理解、语言生成、对话控制和知识服务方面取得了卓越的突破,刷新了非人类实体(包括动物和机器)理解人类自然语言的崭新高度。除了创新技术的使用之外, ChatGPT 使用了规模巨大的算力,拥有海量的参数。这种大语言模型的规模效应还导致了一些语言水平接近于人类的智力行为的“涌现”(emergence),至今仍在不断地迭代。ChatGPT 的成功具有划时代的里程碑性质,足以载入人工智能发展的史册。

如何正确认识 ChatGPT 这种大语言模型的技术实质,是理解并应对 ChatGPT 给人类社会带来的影响的关键。ChatGPT 首先是在语言能力方面取得了重大的突破。ChatGPT 的这些技术突破都跟语言能力直接有关。从技术上说,在大语言模型中,语言成分的“远距离依存”(long distance dependency)以及语言的“词汇歧义”(lexical ambiguity)和“结构歧义”(structure ambiguity)的处理,其功夫都在语言之外。如果把语言能力比作一座冰山,那么语言形式只是冰山露在水面之上的部分,而语义本体知识(semantic ontology knowledge)、常识事理(common sense)和专业领域知识(field knowledge)则是水面之下的部分,而这些知识也正是解决远距离关联问题和歧义消解问题的关键。而这些也正是 ChatGPT 的短板。

2023 年 3 月 17 日, OpenAI 发布 GPT-4。GPT-4 从自然语言的领域进入了多模态(Multimedia),除了处理语言之外,还可以处理图形、音乐等信息,它具有强大的识图能力,文字输入限制由 3 千词提升至 2.5 万词,回答问题的准确性显著提高,能够生成歌词、创意

文本,改变文本的写作风格。当任务的复杂性达到足够的阈值时,GPT-4 比 ChatGPT 更加可靠、更具有创意,并且能够处理更细微的指令。

许多现有的机器学习基准测试都是用英语编写的,为了了解 GPT-4 在其他语言上的能力,OpenAI 研究团队使用 Azure Translate 将一套涵盖 57 个主题的 14 000 个多项英语选择题翻译成多种语言。在测试的 26 种语言的 24 种中,GPT-4 优于 ChatGPT 和其他大语言模型的英语语言性能。GPT-4 在语言能力上也有很大的提升。

ChatGPT 是一个基于 OpenAI 技术的大语言模型,它能够理解和生成人类自然语言。在计算语言学方面,ChatGPT 拥有广泛的应用,包括:

- (1) 文本生成:可以生成各种类型的文本,例如新闻报道、电影剧本、小说和诗歌等;
- (2) 机器翻译:可以将一种语言自动转换为另一种语言,例如将英语翻译成中文;
- (3) 语音识别:可以将语音信号转换为机器可读的文本,从而实现语音识别;
- (4) 自然语言理解:可以理解人类的自然语言,并提取出其中的关键信息,例如情感分析、问答系统和实体识别等;
- (5) 文本分类:可以根据文本内容对其进行分类,例如垃圾邮件过滤或情感分类。

因此,ChatGPT 对于计算语言学的研究具有很大的作用,它可以帮助人们更好地理解和使用自然语言,也可以为企业和研究机构提供高效的文本处理解决方案。

2023 年 11 月 7 日,Open AI 在旧金山举行开发日(DevDay),OpenAI 的总裁奥特曼(Altman)在开发日上宣布 GPT-4 的大升级,推出 GPT-4 Turbo。上千万的用户连夜观看了这场开发者大会,把这次大会称之为“科技春晚”。

Altman 指出,OpenAI 对开发者关注的问题做了如下六大升级:

- (1) 更长的上下文长度:上下文长度由 32K 提升到 128K,这意味着 GPT 能够理解超过 300 页纸张上的文本量;
- (2) 更强的功能控制:能够一次性地使用一条消息来调用多个功能;
- (3) 更好的世界知识:知识更新延伸到 2023 年 4 月份,这意味着 GPT 不久就可以完成对于人类全部知识的学习;
- (4) 多模态:开放了图像生成、图像理解和语音合成的 API 接口;
- (5) 微调个性定制:开启了用户精细调整实验的访问程序,推出了定制模型计划;
- (6) 更高的速率限制。

多年来,我们已经开发出两种不同类型的人工智能系统,一种是基于符号主义(symbolism-based)的传统的人工智能系统,一种是基于连接主义(connectionism-based)的大语言模型,每一种系统都很先进,都可以独立应用。如果我们把这两种类型的人工智能系统结合起来,就有可能使我们更进一步朝着值得我们信任的通用人工智能迈进一步

(Lenat et al., 2023)。

GPTs 系列的成功具有划时代的里程碑性质,是大语言模型时代最伟大的成果,足以载入人工智能发展的史册。但是,语言智能仅仅是人类智能的一部分,语言智能并不能代表人类的全部智能。狭义人工智能利用计算机强大的算力和存储容量,可以相对轻松地根据大量观察到的语言数据生成复杂的模型。一旦条件稍有变化,这些模型往往就无法通用。这意味着,当前的人工智能还不能算通用的人工智能(Artificial General Intelligence, AGI),而只能从大规模的语言数据中提炼信息或经验。当前的人工智能不是通过形成一个全面的世界模型(world model)去理解,而仅仅只是创建一个语言模型去表述。真正的通用人工智能还没有到来。那么,我们离真正的通用人工智能还有多远呢?大多数人工智能研究人员和权威机构认为,人类距离真正的通用人工智能最少还有几年的时间。

2023 年 4 月 13 日,OpenAI 的合作伙伴微软的塞巴斯蒂安·布贝克(Sebastien Bubeck),瓦伦·钱德拉塞卡兰(Varun Chandrasekaran)等发表了一篇论文《通用人工智能的火花:GPT-4 的早期实验》(Sparks of Artificial General Intelligence: Early experiments with GPT-4)。这篇论文指出,在 GPT-4 发布后,面对这个目前性能最强大的人工智能,很多人将 GPT-4 看作通用人工智能的火花。这篇论文中提道:

GPT-4 不仅掌握了语言,还能解决涵盖数学、编码、视觉、医学、法律、心理学等领域的前沿任务,且不需要人为增加任何的特殊提示。并且在所有上述任务中,GPT-4 的性能水平都几乎与人类水平相当。基于 GPT-4 功能的广度和深度,我们相信它可以合理地被视为通用人工智能的最接近的但不完全的版本。(Bubeck et al., 2023)

大语言模型深刻改变了过去的语言知识生产方式,呈现出语言学的研究主体从单一的个体钻研到团体的群智协同,语言学的研究过程从经验积累到数据分析,语言学的研究形式从单一学科到多学科,从单一的文本数据或语音数据到多模态数据,这是语言知识生产范式在方法论上的剧烈变革和重大创新,这样的变革和创新将会推动整个语言学的发展(冯志伟等,2024)。

7 大语言模型与翻译

现在基于大语言模型的机器翻译在一些领域和语种中的正确率已经能达到 98% 以上。但我认为基于大语言模型的机器翻译还不具有真正的人类智能,存在无法理解感情、缺乏常识等问题,这是基于大语言模型的机器翻译发展面临的挑战。

人类对于自然语言的理解除了依靠语言内部的各种关系之外,还要依靠外部物理世界、内部精神世界和社会历史世界等背景知识。自然语言文本中的每一个符号、每一个合乎规则结构的符号串,在人脑中都与外部客观世界有着复杂的联系。这些复杂联系不仅以

概念的符号形式表现出来,还具有视觉、听觉、触觉、体内觉等表征,甚至有更深入的心理感情表征以及社会文化背景。当前大语言模型通过多模态建模已经具备一定的视觉和听觉能力,但是还不具备处理触觉和体内觉的能力,更不能处理丰富多彩的语言外常识。因此,大语言模型还不能挖掘语言数据与外部世界的多种多样的复杂联系,这是当前人工智能(artificial intelligence)与人类智能(human intelligence)最根本的差别。

翻译是人类的高级智能活动,翻译活动不仅涉及语言内部的结构,还涉及语言外部的日常生活知识、社会知识、历史知识、文化背景知识等诸多复杂丰富的要素,这些非语言要素构成了翻译的“人文硬核”(humanity core)。目前,基于大语言模型的机器翻译尽管已取得长足进步,并具有一定模拟人类语言内部结构的能力,但是模拟外在世界以及社会历史背景的能力还十分有限,也难以处理这些复杂而丰富的“人文硬核”,因此,当基于大语言模型的机器翻译遇到“人文硬核”时,就往往会捉襟见肘、左右为难(冯志伟,2024)。

近年来,翻译技术发展迅猛,有人认为翻译人员就要失业了。可是我认为,机器翻译的翻译能力被人们夸大了,机器翻译无法取代人类译员,复杂的高端翻译工作必须由人来承担。一方面,文学作品、科技文献等垂直领域的翻译仍需人类完成;另一方面,有较高保密要求的翻译任务以及重要场合的同声传译和交替传译等也需要人类完成。具体来说,文学翻译需要译者具备极高的人文科学素养和对源语文化背景的深刻理解能力,同时也要能熟练并创造性地运用目标语,这是机器翻译难以胜任的,需由人类译员来承担。科技翻译中的多义术语可以表示不同领域的多种概念,机器翻译难以正确辨别这样的多义术语,往往会造成翻译错误,需要由人类译员进行判断。此外,尽管机器翻译也可以做同声传译和交替传译,但是,实时翻译场景中,机器传译往往难以及时纠错纠偏,可能会造成无法挽回的后果,因此重要的同声传译和交替传译也要由人类译员来承担。由此可见,机器翻译并不能代替人类译员,高端翻译专家是机器翻译永远也取代不了的。机器翻译将成为人类译员的好朋友和得力助手。两者应当和谐共生、相得益彰。在人工智能时代,各种翻译技术工具的智能化程度越来越高,这都将有助于提升人类译员翻译效率。翻译工作者应当与时俱进,拥抱技术、学习技术、掌握技术。

大语言模型为翻译行业提供了新的机遇,也带来了新的挑战。大语言模型使用机器学习和自然语言处理技术实现自动翻译,这让翻译变得更加快速、便捷、准确,节省了时间和经济成本;大语言模型可以根据用户的需求和偏好进行定制化翻译,提高翻译质量和用户体验,提供个性化服务;大语言模型可以帮助企业与客户进行更加智能化的交流互动,提升客户满意度和忠诚度;大语言模型让不同语言、文化间的沟通交流变得更加容易,促进了全球化发展和跨文化交流;大语言模型可以收集大量语言数据并对数据进行分析和挖掘,从而产生有价值的商业洞察和见解。总之,大语言模型通过技术创新,将推动翻译行业的智

能化革新和高效创新发展。

当然,大语言模型也给翻译行业带来了新的挑战。大语言模型有翻译能力,能够在短时间内创造出大量翻译结果,相比于人类译员是一种低成本、高效率的选择;同时,其翻译结果还可以自动结合大语言模型中的数据信息,上下文理解能力相较于传统的机器翻译有显著提升。因此,结合了上下文理解、译文润色等功能的大语言模型对翻译行业带来了巨大影响和冲击,这将导致部分传统翻译公司的市场份额逐步下降。随着大语言模型技术的不断发展,越来越多的企业将会开始使用大语言模型来提升其翻译产品的质量和效率,因此,那些不能提供更优质服务的企业将会面临退出市场的风险。要应对这些挑战,翻译行业可以通过加强自身核心竞争力、拓展新领域等方式来保持市场竞争力。同时,也可以考虑与大语言模型技术结合,提高自身服务质量和效率。

8 大语言模型的内容治理

大语言模型最受争议的挑战是它们产生“幻觉”(hallucination)的倾向。大语言模型的幻觉来源于它们缺乏对事件之间因果关系的了解,目前大语言模型有很强的语言处理(language processing)能力,也有一定的知识处理(knowledge processing)能力,但与其语言处理能力相比,其知识处理能力的火候还稍微欠缺,特别是缺乏跟专业领域相关的知识能力,说多了就会“露馅”,有时甚至会提供不符合事实的错误答案,有时大语言模型会一本正经地胡说八道,或者说一些永远正确的废话,或者说一些违背人类道德规范的胡话。大语言模型甚至会捏造数据和事实,或在逻辑推断、因果推理等方面颠三倒四,导致数据隐私泄露、虚假内容生成、模型不当利用等安全风险,引发社会信任危机,从而使广大用户陷入真假难辨、人人自危的困境。因此,我们必要对大语言模型进行内容治理(content governance)。

2023年以来,国内外提出了一系列大语言模型内容治理的办法和倡议。2023年1月,美国国家标准与技术研究院(NIST)制定了《人工智能风险管理框架》,对大语言模型存在的风险进行了详尽的分级分类。2023年8月15日,我国正式施行《生成式人工智能服务管理办法》,这个《办法》规定了对大语言模型服务提供者的制度要求,为未来大语言模型行业的发展指明了方向。10月18日,我国中央网信办发布《全球人工智能治理倡议》。倡议提出,发展人工智能应坚持相互尊重、平等互利的原则,各国无论大小、强弱,无论社会制度如何,都有平等发展和利用大语言模型的权利。与此同时,我国科技部“国家新一代人工智能治理委员会”制定了《新一代人工智能伦理规范》,我国外交部发布了《中国关于人工智能伦理治理的立场文件》。11月1日,首届人工智能安全峰会在英国召开,会议发布的宣言指出,大语言模型的许多风险基本上是全球性的,因此最好通过国际合作来解决。

月 8 日,欧洲议会、欧盟成员国和欧盟委员会三方,达成了关于《人工智能法案》的协议,建立起人工智能开发、使用道德和法律框架,对人工智能领域实行全面监管,该法案将成为首部人工智能领域的全面监管法案。

12 月 28 日,OpenEval 平台、中国软件评测中心等机构联合发布《2023 人工智能大模型基准测试白皮书》,指明了大语言模型潜在的安全风险点,说明了大语言模型在追求知识和能力提升的同时,还应当关注大语言模型与人类价值的对齐(value alignment)。随着大语言模型能力的不断进化,价值对齐问题的重要性将会日益突出。大语言模型的内容治理已成为全球的共识,这是值得高兴的事情。

参考文献:

- Arivazhagan, Naveen, Ankur Bapna, Orhan Firat et al. 2019. Google AI. Massively Multilingual Neural Machine Translation in the Wild: Findings and Challenges [J/OL]. [2023-5-18]. <https://arxiv.org/abs/1907.05019>.
- Bengio, Y., R. Ducharme, P. Vincent. 2001. A Neural Probabilistic Language Model [J]. *Advances in Neural Information Processing Systems* (13): 932-938.
- Brown, T., B. Mann, N. Ryder et al. 2020. Language Models are Few-Shot Learners [J]. *Advances in Neural Information Processing Systems* 33(2020): 1877-1901.
- Bubeck, Sebastian, Varun Chandrasekaran, Ronen Eldan et al. 2023. Sparks of Artificial General Intelligence: Early Experiments with GPT-4[J/OL]. [2023-5-18]. <https://arxiv.org/abs/2303.12712>.
- Chowdhery, A., S. Narang, J. Devlin et al. 2022. Palm: Scaling Language Modeling with Pathways [J]. *Journal of Machine Learning Research* 24(240): 1-113.
- Chung, H. W., L. Hou, S. Longpre et al. 2022. Scaling Instruction-Finetuned Language Models[J/OL]. [2022. 10. 20]. <https://arxiv.org/abs/2210.11416>.
- Deng J., W Dong, R. Socher et al. 2009. Imagenet: A Large-Scale Hierarchical Image Database [C]//2009 IEEE Conference on Computer Vision and Pattern Recognition. *IEEE*: 248-255.
- Devlin, J., M. W. Chang, K. Lee et al. 2019. Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding [J]. *Proceedings of NAACL-HLT*(1): 4171-4186.
- Gao, Jinglong, Xiao Ding, Bing Qin et al. 2023. Is ChatGPT a Good Causal Reasoner? A Comprehensive Evaluation [J/OL]. [2023-05-18]. <https://arxiv.org/pdf/2305.07375>.
- J. McCarthy, M. L. Minsky, N. Rochester et al. 1955. A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence [J/OL]. [2023-07-15]. <https://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html>.
- Kaplan, J., S. McCandlish, T. Henighan et al. 2020. Scaling Laws for Neural Language Models[J/OL]. [2023-05-18]. <https://arxiv.org/abs/2001.08361>.
- Lenat, Douglas & Gary Marcus. 2023. Getting from Generative AI to Trustworthy AI: What LLMs Might Learn from Cyc [J/OL]. [2023-07-15]. <https://arxiv.org/abs/2308.04445>.
- Mikolov, T., M. Karafiát, L. Burget et al. 2010. Recurrent Neural Network Based Language Model [J]. *Interspeech* (9): 1045-1048.
- Nakano, R., J. Hilton, S. Balaji et al. 2021. Webgpt: Browser-assisted question-answering with human feedback [J/OL]. [2023-6-24]. <https://arxiv.org/abs/2112.09332>.
- Ouyang, L., J. Wu, X. Jiang et al. 2022. Training Language Models to Follow Instructions with Human Feedback [J]. *Advances*

- in *Neural Information Processing Systems* (35):27730-27744.
- Peters, M., M. Neumann, M. Iyyer et al. 2018. Deep Contextualized Word Representations [C] // *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.
- Peters, Matthew, E. Mark Neumann, Mohit Iyyer et al. 2018. Deep Contextualized Word Representations [J/OL]. <https://arxiv.org/abs/1802.05365>.
- Radford, A., J. Wu, R Child et al. 2019. Language Models are Unsupervised Multitask Learners [Preprint]. OpenAI blog. [2023-4-18]. <https://pdfs.semanticscholar.org/41f9/45f59bd0d345d4e355fb72110524f6fdffdb.pdf>
- Raffel, C, N. Shazeer, A. Roberts et al. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer [J]. *The Journal of Machine Learning Research* 21(1):1-67.
- Sanh, V., A. Webson, C. Raffel et al. 2021. Multitask Prompted Training enables Zero-shot Task Generalization [J/OL]. [2023-5-18]. <https://arxiv.org/abs/2110.08207>.
- Sukhbaatar, S., J. Weston, R. Fergus et al. 2015. End-to-end Memory Networks [J]. *Advances in Neural Information Processing Systems* 28: 2440-2448.
- Thoppilan, R, D. De Freitas, J. Hall et al. 2022. Lamda : Language Models for Dialog Applications [J/OL]. [2023-6-24]. <https://arxiv.org/abs/2201.08239>.
- Vaswani, A, N. Shazeer, N. Parmar et al. 2017. Attention is all You Need [J]. *Advances in Neural Information Processing Systems* (30):1-11.
- Vaswani Ashish, Noam Shazeer, Niki Parmar et al. 2017. Attention Is All You Need [C] // 31st Conference on Neural Information Processing Systems (NIPS 2017).
- Zhao Wayne Xin, Kun Zhou, Junyi Li et al. 2023. A Survey of Large Language Models [J/OL]. [2023-5-24]. <https://arxiv.org/abs/2303.18223>.
- 冯志伟, 李颖. 2021. 自然语言处理中的预训练范式 [J]. *外语研究* (1):1-14 + 112.
- 冯志伟, 张灯柯. 2024. 计算语言学中语言知识生产范式的变迁 [J]. *当代修辞学* (2):1-22.
- 冯志伟. 2023. 计算语言学方法研究 [M]. 上海: 上海外语教育出版社.
- 张奇, 桂韬, 黄萱菁. 2023. 自然语言处理导论 [M]. 网络预览版. 北京: 电子工业出版社.
- 张子豪, 刘海涛. 2023. 从线性位置看神经网络中语言规律的获得与表征 [J]. *当代语言学* (6):791-809.
- 冯志伟. 2024. 机器翻译应当强化“人文硬核” [J/OL]. 中国外文局翻译院智能翻译实验室公众号. 2024-02-09. [2024-2-10]. https://mp.weixin.qq.com/s?_biz=MjM5MDkxMzg5NA==&mid=2649491800&idx=1&sn=648429222b4616b0fe9518d5debfc114&chksm=bff6bee58e94eba86ca6782cb7d298315a5b5ecd71ae6934d3a5de00c05e7c58f5319702865a&scene=27

Large Language Model in Artificial Intelligence

FENG Zhiwei ZHANG Dengke

Abstract: Natural language processing is an important field of artificial intelligence, and large language models are distinguished achievements in natural language processing. This article describes the development history of large language models, and introduces the basic principles and structures of the large language models as pre-training models, transformer models, dynamic word vector embedding model ELMO, bidirectional encoding representation model BERT, generative pre-training transformer model GPT. Finally, it discusses the relationship between large language models and translation, and the content governance issues of large language models. The study points out that big language modeling has not only pushed natural language processing to achieve engineering success, but also profoundly changed the previous way of language knowledge production, making language research move from unidisciplinary to multidisciplinary. This change and innovation will undoubtedly promote the development of linguistics.

Key words: natural language processing; large language model; pre-training models; Transformer model; ChatGPT; content governance

责任编辑: 龙丹