讲堂丨周明：自然语言处理的技术体系和未来之路

微软研究院AI头条 7月15日

文章转载自公众号 AI科技评论，作者 Camel

▲ 点击蓝字关注微软研究院AI头条

Microsoft Research
微软亚洲研究院

编者按：上周，微软亚洲研究院副院长周明博士受邀参加第四届全球人工智能与机器人峰会（CCF-GAIR 2019），并在会议上做了主题演讲，从什么是自然语言处理（NLP）、当前技术体系以及未来发展等角度，解读了NLP未来发展之路。本文授权转载自“雷锋网”。

微软亚洲研究院副院长周明

大家下午好！今天非常荣幸来到CCF-GAIR大会，今天下午这个论坛非常有意义，讲的是中国人工智能四十周年纪念活动。

我是1985年在哈工大开始从事机器翻译研究的，到现在也已经有30多年了，经历了规则、统计和神经网络的三个阶段。回想过去真是感慨万千，当时可以说是筚路蓝缕，没有什么东西，但是大家有一番热情，要把中国自然语言、机器翻译、人工智能推到世界的前沿。

中国人工智能开始于1979年到今天转眼过去40年了。回首看一下我们的自然语言处理进展到什么程度了？我们未来的路在哪里？这就是我今天要给大家介绍的。

过去40年，自然语言基本上经历了从规则到统计，到现在的神经网络。相比过去，目前可以说是自然语言处理最黄金的时期，在很多领域都取得了突破性的进展。但我们审慎地看到神经网络自然语言处理过度依赖计算资源和数据，在建模、推理和解释方面还存在许多的不足。因此我们想问一下，这种模式是否可以持续？在未来的3到5年，NLP如何发展？

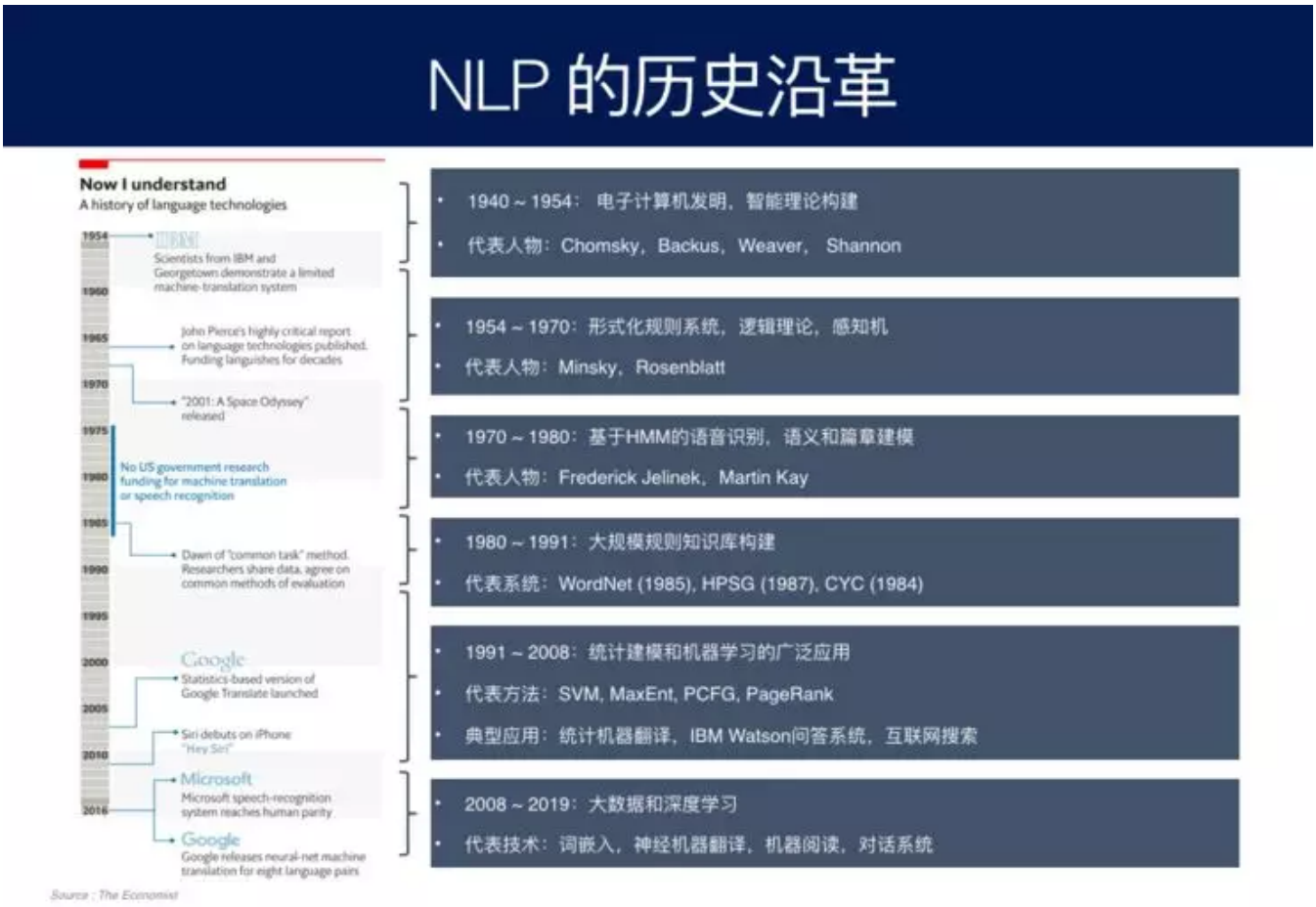
为了回答这个问题，我想把神经网络自然语言处理的技术在这里捋一遍，有哪些关键的技术点，存在哪些不足，我们未来又如何发展。我的观点是：NLP未来的发展需要计算、数据、技术、人才、合作、应用等各个方面长期协同发展。

01 什么叫自然语言处理

什么叫自然语言处理？自然语言处理就是用计算机对人类语言进行处理，使得计算机具备人类的听、说、读、写能力，它是未来人工智能技术最为关键的核心之一。比尔·盖茨说过，“自然语言处理是人工智能皇冠上的明珠，如果我们能够推进自然语言处理，就可以再造一个微软。”

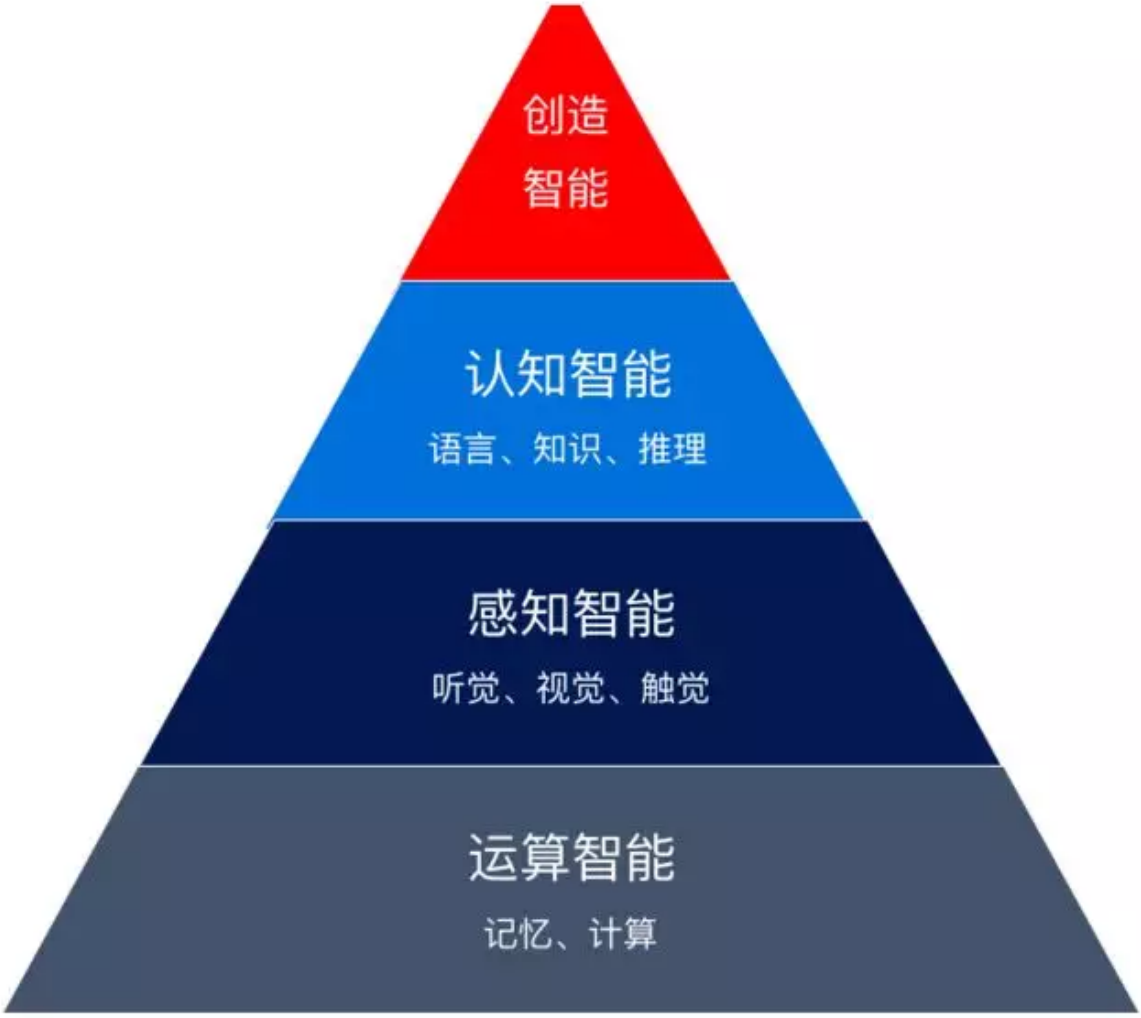
难度：把NLP看作人工智能皇冠上的明珠，其难度可想而知。来看下面这个例子：

词完全一样，意义截然相反。人在理解的时候有常识，有背景，所以能够理解；可电脑没有常识、没有背景，只是根据字面来处理，因此它理解的都是一样的。这就是自然语言处理的难处。



历史：自然语言处理随着计算机的出现而出现，最早是做规则的系统，后面做统计的系统，现在做神经网络的系统。咱们中国的自然语言出现一点也不晚，建国之初就有人开始做俄汉机器翻译系统，后面又有人做英汉机器翻译系统。我个人也有幸亲历和见证了机器翻译的发展。我在哈工大的读研时候（导师李生教授，1985年），从事中英机器翻译研究，所研制的CEMT系统是中国最早通过正式鉴定的中英机器翻译系统（1989年）。后来我在日本高电社领导研发了中日机器翻译产品J-北京（1998年）。我1999年加入微软之后先后从事了基于实例和基于统计机器翻译研究，最近几年我们做神经机器翻译研究。

可以说中国的自然语言处理是与世界的发展同步的。目前我可以很负责任地说，咱们中国的自然语言处理总体来讲位居世界第二，仅次美国。为什么能有这么好的发展？得益于中国40年改革开放，得益于各大公司和很多学校的合作，尤其值得指出的是微软研究院与相关学校的合作影响深远。同时也得益于包括CCF在内的各个学会过去几十年在NLP领域深耕，举办学术会议（NLPCC最近进入CCF-国际会议列表）和各类暑期学校和讲习班，促进学校、企业、公司各个单位合作，并推动研究协同式、平台式发展。



定位：人工智能就是用电脑来实现人类独具的智能。使得电脑能听、会说、理解语言、会思考、解决问题、会创造。具体概括来讲包括：运算智能、感知智能、认知智能和创造智能。运算智能就是记忆和计算的能力。这一点计算机已经远远超过人类。而感知智能就是电脑感知环境的能力，包括听觉，视觉，触觉等等。相当于人类的耳朵、眼睛和手。认知智能包括语言理解、知识和推理。创造智能体现对未见过、未发生事物，运用经验，通过想象力、设计、实验、验证并予以实现的智力过程。目前随着感知智能的大幅度进步，人们的焦点逐渐转向了认知智能。其中语言智能，也就是自然语言理解，则被认为是皇冠上的明珠。一旦有突破，则会大幅度推动认知智能，并提高人工智能的技术，并促进在很多重要场景落地。

过去几年，由于数据越来越多，出现各种测试集；算法越来越复杂、越来越先进，包括神经网络的架构、预训练模型等等；计算能力越来越高，在这三大因素的作用下，自然语言处理得到了飞速的发展。

微软在四个NLP典型任务取得了突破性的进展。第一个是聊天机器人，我们中、日、英三种语言的聊天机器人均能达到跟人自由聊天23轮以上，目前在世界上是最好的。还有我们的阅读理解技术、机器翻译技术和语法检查系统，在目前的测试集下都居世界领先水平，而且在相应的测试集下都突破了人类的标注水平。

自然语言有很多的应用，像我们每天都用的输入法、词典、翻译，以及我们跟中科院合作的手语翻译、必应的语音助手、小冰，还有自然语言的文本生成，对联、诗词、猜谜、音乐等等。

02 技术体系

我给大家捋一下神经网络自然语言处理的技术体系。

首先是词的编码。词编码的目的是用多维向量来表征词的语义。怎么做呢？著名的方法有两个，一个是CBOW（Continuous Bag-of-Words），用周围的词预测当前的词；另一个是Skip-gram，用当前的词预测周围的词。通过大规模的学习训练，就可以得到每个词稳定的多维向量，作为它的语义表示。

有了词的语义表示，我们就可以进而生成句子的语义表示，也叫句子的编码。一般通过RNN（循环神经网络）或者CNN（卷积神经网络）来做。RNN从左到右对句子进行建模，每个词对应一个

隐状态，该引状态代表了从句首到当前词的语义信息，句尾的状态就代表了全句的信息。CNN从理论上分别进行词嵌入+位置嵌入+卷积，加上一个向量表示，对应句子的语义。

基于这样的表征，我们就可以做编码、解码机制。比如说我们可以用图上的红点，它代表全句的语义信息，来进行解码，可以从一种语言翻译成另一种语言，凡是从一个序列串变成另外一个序列串都可以通过编码、解码机制来运行。

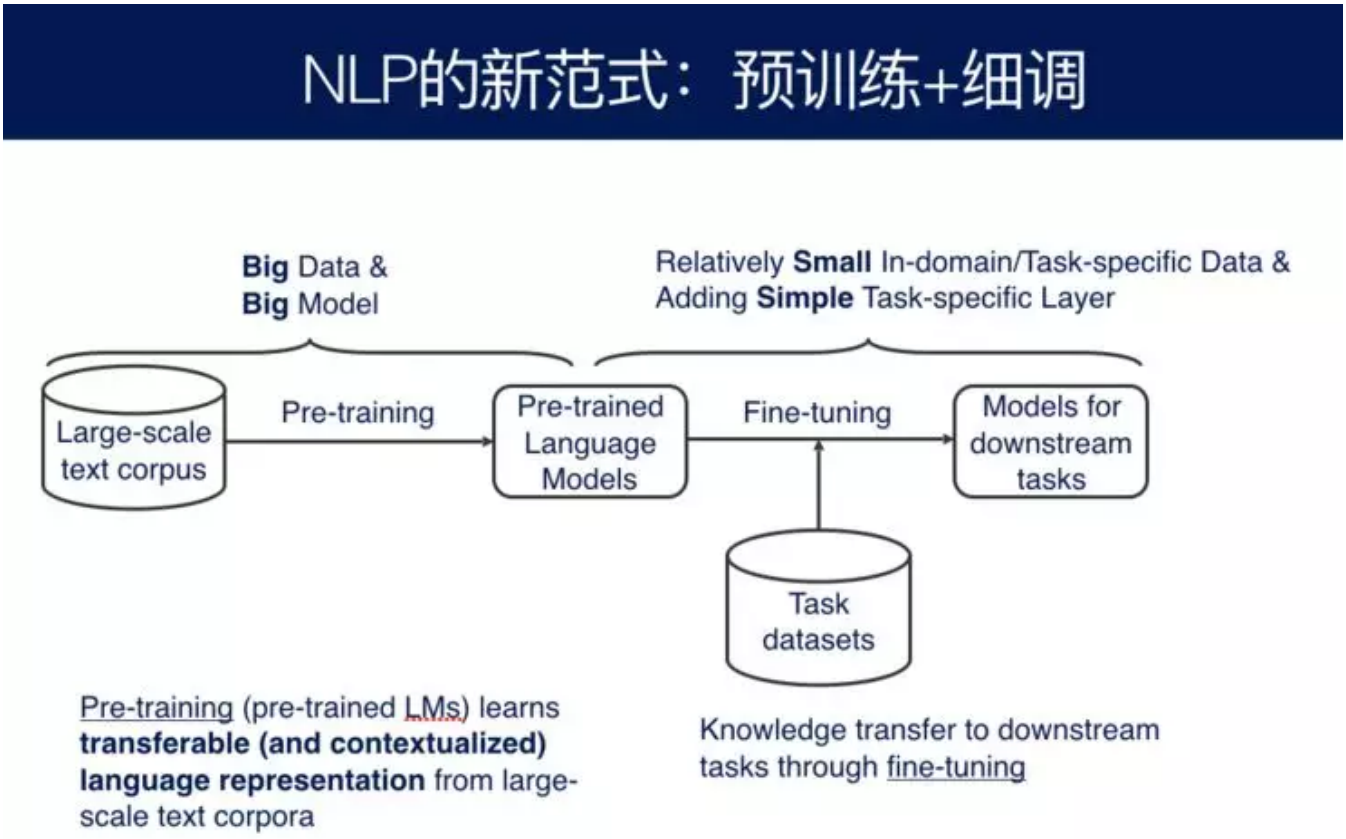
随后又引入了注意力模型。它综合考量了在当前状态下对应的编码的每一个隐状态，加权平均，来体现当前的动态输入。这类技术引入之后，神经网络机器翻译就得到了飞速的发展。

后面又引入了Transformer。Transformer引入了自编码，一个词跟周围的词建立相似，引入多头，可以引入多种特征表达，所以编码效果或者编码的信息更加丰富。

现在大家都在追捧预训练模型。它有几个方法，第一个是ELMo，从左到右对句子编码，也可以从右到左对句子编码，每一层对应的节点并起来，就形成了当前这个词在上下文的语义表示。用的时候就用这个语义加上词本身的词嵌入，来做后续的任务，性能便得到相应的提高。

还有去年10月份比较火的BERT。它用左边、右边的信息来预测最外部的词的信息，同时它也可以判断下一句是真的下一句还是伪造的下一句，用两种方式对句子每一个词进行编码，得到的训练结果就表征了这个词在上下文中的语义表示。基于这样的语义表示，就可以判断两个句子的关系，比如说是不是附属关系，判断一个句子的分类（例如Q&A中，判断回答对应的边界是不是对应提问），以及对输入的每一个词做一个标注，结果就得到一个词性标注。

预训练模型引起了很多人的关注。最早是一个静态的词的表征，所谓静态词的表征，就是不管上下文，表征是一样的，比如“bank”这个词有多个意思，它的表征也是一样的。但是ELMo就要根据上下文体现它唯一的表征。基于以上的方法，人们又开发了一系列的新的方法，比如说GPT-2，以及最近的XLNET，以及UNILM、MASS、MT-DNN、XLM，都是基于这种思路的扩充，解决相应的任务各有所长。其中微软研究院的UNILM可同时训练得到类似BERT和GPT的模型，而微软MASS采用encoder-decoder训练在机器翻译上效果比较好。还有MT-DNN强调用多任务学习预训练模型，而XLM学习多语言BERT模型，在跨语言迁移学习方面应用效果显著。针对预训练模型很多公司都有一些改进，这里就不一一列举了。



现在由于这种预训练模型大行其道，人们在思考，自然语言处理是不是应该改换一种新的模态。过去我们都说用基于知识的方法来充实当前的输入，但是过去都没有做到特别好，而这种新的预训练模型给我们带来一个新的启发：

我们可以针对大规模的语料，提前训练好一个模型，这个模型既代表了语言的结构信息，也有可能代表了所在领域甚至常识的信息，只不过我们看不懂。加上我们未来的预定的任务，这个任务只有很小的训练样本，把通过大训练样本得到的预训练模型，做到小训练样本上，效果就得到了非常好的提升。

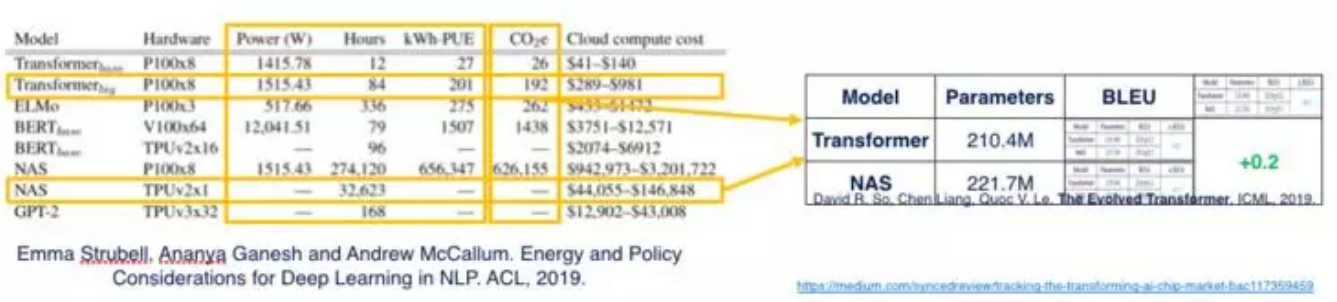
03 未来发展

现在，NLP在许多任务上的性能都已经超越了人类。听起来世界一片大好，我们把数据准备好，买一大堆机器，只需要去训练就好了，不用管太多的事情。所以现在好多人在刷榜，有了新的任务，搞一堆模型、一堆数据、一堆机器，刷个榜，我们就毕业了。

但是我认为不是这样的，反而有强烈的危机感。

下面我就跟大家分析一下，目前存在的问题，以及我们应该怎么做才好。

第一个是无休止的计算资源的军备竞赛。现在大家都用大规模的机器训练，同样的算法，只要训练速度快，就可以快速迭代，然后你的水平就比别人高。与之同时，当然也特别耗资源，许多模型一训练可能要好几天或者好几万美金。有时候它管事，但有时候也不管事。举个例子：



Model	Hardware	Power (W)	Hours	kWh-PUE	CO_2e	Cloud compute cost
$Transformer_{base}$	P100x8	1415.78	12	27	26	$41–$140
$Transformer_{big}$	P100x8	1515.43	84	201	192	$289–$981
ELMo	P100x3	517.66	336	275	262	$433–$1,472
$BERT_{base}$	V100x64	12,041.51	79	1507	1438	$3751–$12,571
$BERT_{base}$	TPUv2x16	—	96	—	—	$2074–$6912
NAS	P100x8	1515.43	274,120	656,347	626,155	$942,973–$3,201,722
NAS	TPUv2x1	—	32,623	—	—	$44,055–$146,848
GPT-2	TPUv3x32	—	168	—	—	$12,902–$43,008

Emma Strubell, Ananya Ganesh and Andrew McCallum. Energy and Policy Considerations for Deep Learning in NLP. ACL, 2019.

Model	Parameters	BLEU
Transformer	210.4M	+0.2
NAS	221.7M	

David R. So, Chen Liang, Quoc V. Le. The Evolved Transformer. ICML, 2019.

https://medium.com/syncedreview/tracking-the-transforming-ai-chip-market-bac117359459

在这个例子中，它用了10倍的蛮力，但是只有0.2%的效率提升。由于用了很多的资源，造成了环境的污染。最近有一篇网上比较火的论文，就是在讨论这个计算模型。如果我们太依赖算力，就会对环境产生很大的影响。

第二个是过度依赖数据。首先你要标数据，标注的代价是非常大的。其次，数据有隐含歧视的问题，通过数据分析，可能会得到歧视性的结果。另外数据有偏差，数据在标注的时候请人标注，人都是偷懒的，想最简单的方法去标注，结果标注的数据千篇一律，基于这样的数据学的模型也只能解决标注的数据，拿到真实任务上由于跟你标注分布不一样，所以根本不好使。比如说我们做Q&A问答系统，我们在所有的问答里面都假设是第一名，但到了搜索引擎上有很多简单的问题都解决不好。此外，还有数据隐私保护等等问题。

我们现在再往前走一走，看一看，假如我们不在乎资源，不在乎计算，我们看神经网络处理一些典型的任务，它的表现如何，有哪些问题。

我这里选了三个最典型的问题。第一个是Rich Resource Tasks，即有足够资源的任务，比如中英机器翻译，网上有很多的资源。第二个Low Resources Tasks，即资源很少或没有资源，比如说中文到希伯来语的翻译，几乎没有什么资源。第三个是Multi-turn Tasks，就是多轮的意思，我们的客服都是多轮的。这三类问题基本上代表了自然语言最基本的问题，如果这三类问题都解决得很好，自然语言就基本OK了。我们看看这三类问题现在处于什么位置上。

针对Rich Resource Tasks，我们做一个中-英神经网络机器翻译错误分析。这是一个大语料训练的结果。

Error Category	Fraction [%]	
Incorrect Words	7.64	
Ungrammatical	6.33	
Missing Words	5.46	
Named Entity	4.38	
Person		1.53
Location		1.53
Organization		0.66
Event		0.22
Other		0.44
Word Order	0.87	
Factoid	0.66	
Word Repetition	0.22	
Collocation	0.22	
Unknown Words	0	

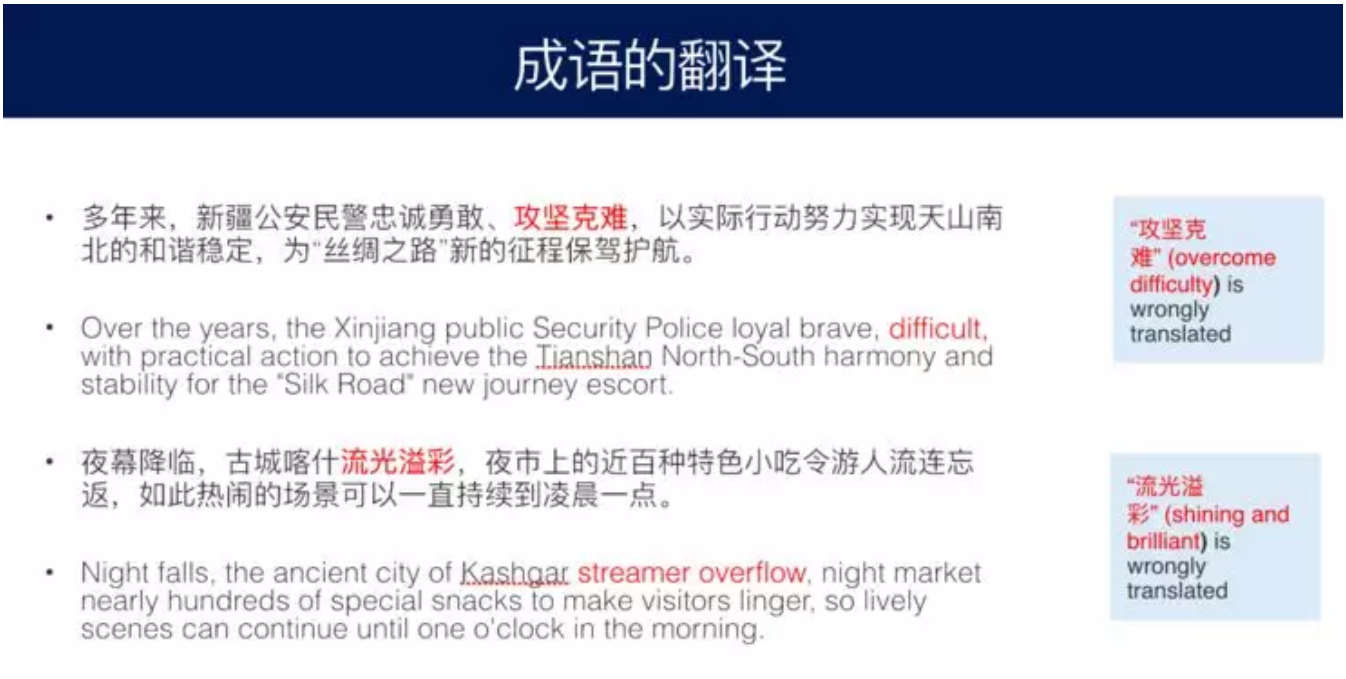
我们可以看到，尽管是基于大语料的，但翻出来的结果还有很多错误，包括翻错词、丢词、不合语法等。

缩写词识别和翻译

- 德国在参与打击极端组织的多国联合行动时，向土耳其空军基地派驻约250名军人。土耳其政府此前指责德国为参与去年7月土耳其未遂政变的人员提供政治避难。作为报复，土方禁止德国议员探视德国驻军。
- Germany has deployed about 250 troops to the Turkish Air Force base in its multinational operations against extremist groups. The Turkish government has previously accused Germany of providing political asylum for those who participated in last July's attempted coup in Turkey. In retaliation, the Earth forbids German MPs to visit the German garrison.

"土方" (contextualized acronym of 土耳其/Turkey) is wrongly translated

例如这个“土方”不是“earth”，而是“土耳其”的意思。因为神经网络现在不可解释，它是黑箱，你也不知道它在哪儿丢的，有可能是数据问题，有可能是模型问题。



还有成语，成语是很麻烦的，你即使学了很多的成语，在一个新的句子中，成语的翻译也要发生变化，所以它要动态的计算。

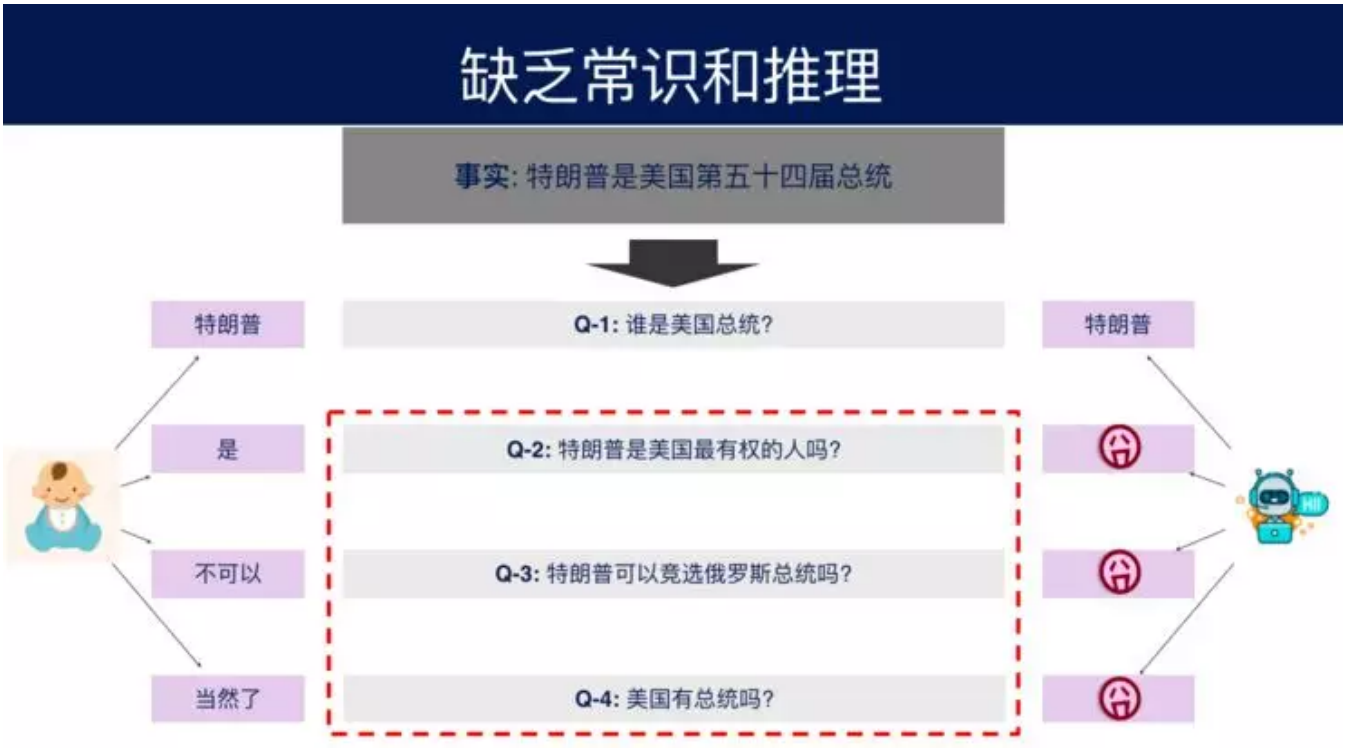
所以即使在这样的足够资源的算法里面，仍然存在众多的问题要研究，比如说丢词，如何把词典集成进来，如何上下文判断一些问题，然后还有领域自适应、主体自适应等等，谁也不敢说这些问题通过Rich-Resource就解决了，这里面有上下文件联系的问题，还有数据歧视的问题，还有Multi-task learning，还有Human knowledge。

第二个是Low Resources Tasks，就是没什么语料，学起来很难，因此要借力。常用的有三种。

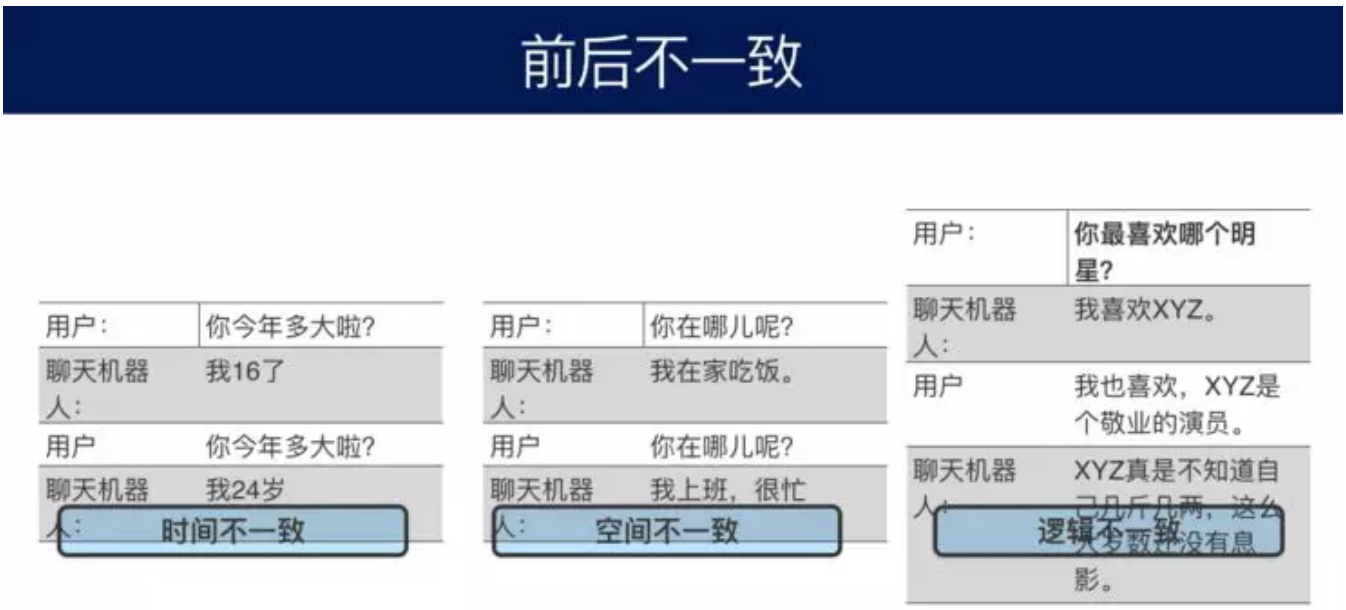
- 第一是迁移模型，把从其它语料中学习到的内容迁移过来。迁移训练最常见的就是前面介绍的预训练模型，把它加到目标任务上。
- 第二是跨语言学习，即从其它语言学习过来。比如说英文有很多语料，我把英文的训练模型用到法语、德语上，这个方式很流行。
- 第三是利用种子进行迭代学习，比如我有一个小辞典，有几条规则，有几条双语，我能不能用它当做一个引子，做一个冷启动，启动之后再迭代改进。

虽然我们做了很多的研究，但是在Low-Resource方面，我们并没有一个很好的办法。首先Low-Resource如何建模，如何从数据分析中做无监督或少监督的学习，这是没有明确回答的问题。怎么做Transfer Learning，怎么做Unsupervised learning，也是目前的一个难题。还有一些先验的规则辞典，如何让它冷启动起来；人能不能参与其中帮助一个开始弱小的系统逐渐变得更加强大。这些都是目前热门的话题，都还没有很好地解决。

第三个是Multi-turn Task（多轮问题）。以多轮对话为例。我们看下面这个例子：

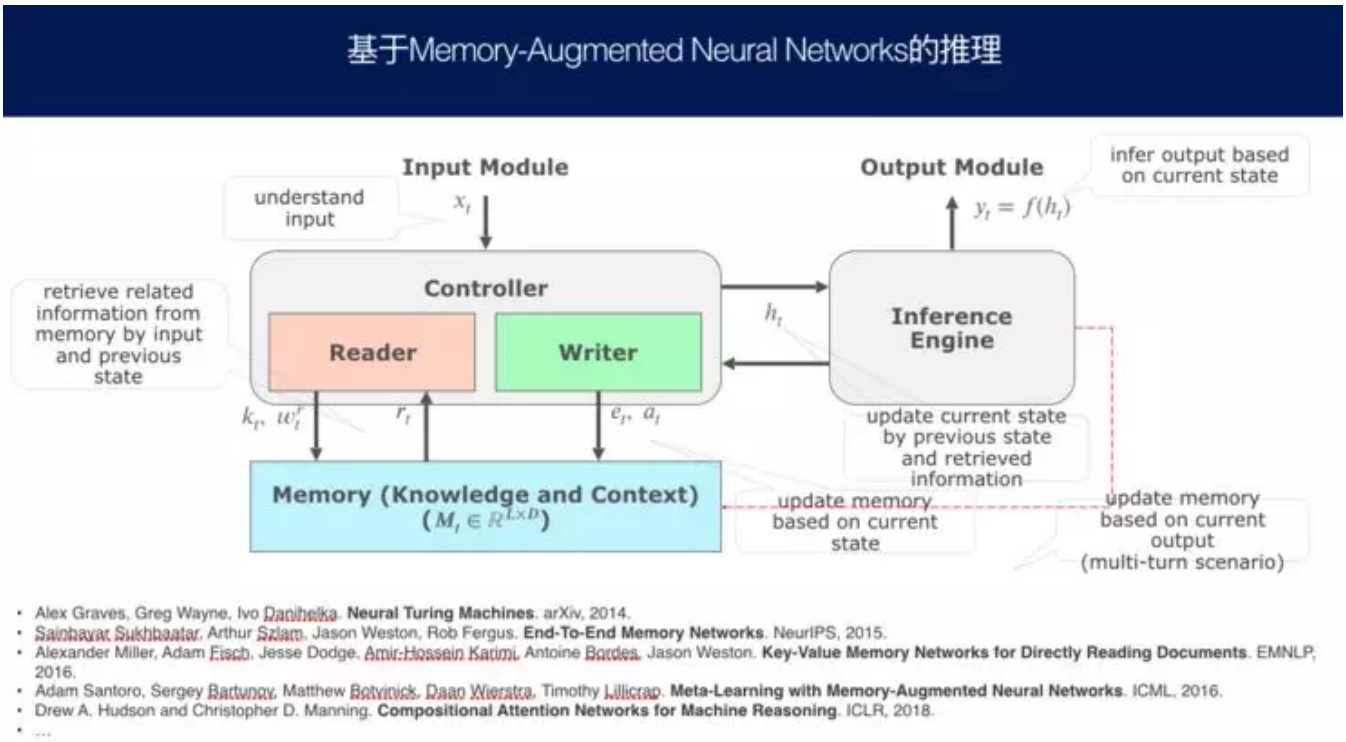


我们可以看到，对于小孩子很简单的问题，电脑却不知道该怎么回答。其原因在于目前的自然语言处理还没有很好地解决常识和推理的问题。



此外，还有前后不一致、自我矛盾的问题。比如说用户问“你今天多大了”？聊天机器人说“我16了”。隔几天用户又问“你今天多大了”？它可能说“我24岁”，自己前后不一致了.还有空间不一致、逻辑不一致的问题.这就需要人跟机器对话的时候，要有一个记忆体系，把说过的话的特征存储起来，将来在用的时候，要抽取这样的信息来表征一个机器人各方面的信息。

推理是要做很多事情。第一是要了解上下文，说过什么话，答过什么问题，干过什么事都要存储起来，记忆起来。第二是各种各样的知识要用起来。第三才是推理的部分，这里面涉及到语义分析、上下文的指代消解、省略消解。最后，还有就是可解释的问题，如果你的推理不可解释的话，那就没有人会相信，导致你的系统无法进行进一步的推进。



我们要做推理，一般来讲就是设计这样的模型.它有记忆，记住我说过什么话，或者有哪些知识；有一个读的装置和一个写的装置，来了一个问题，经过分析，到记忆里把状态和知识找出来，把原来的记忆找出来，然后改变我们的对话状态；更新在记忆里的一些存储。得到答案后，还要反过来更新我们的记忆和存储。

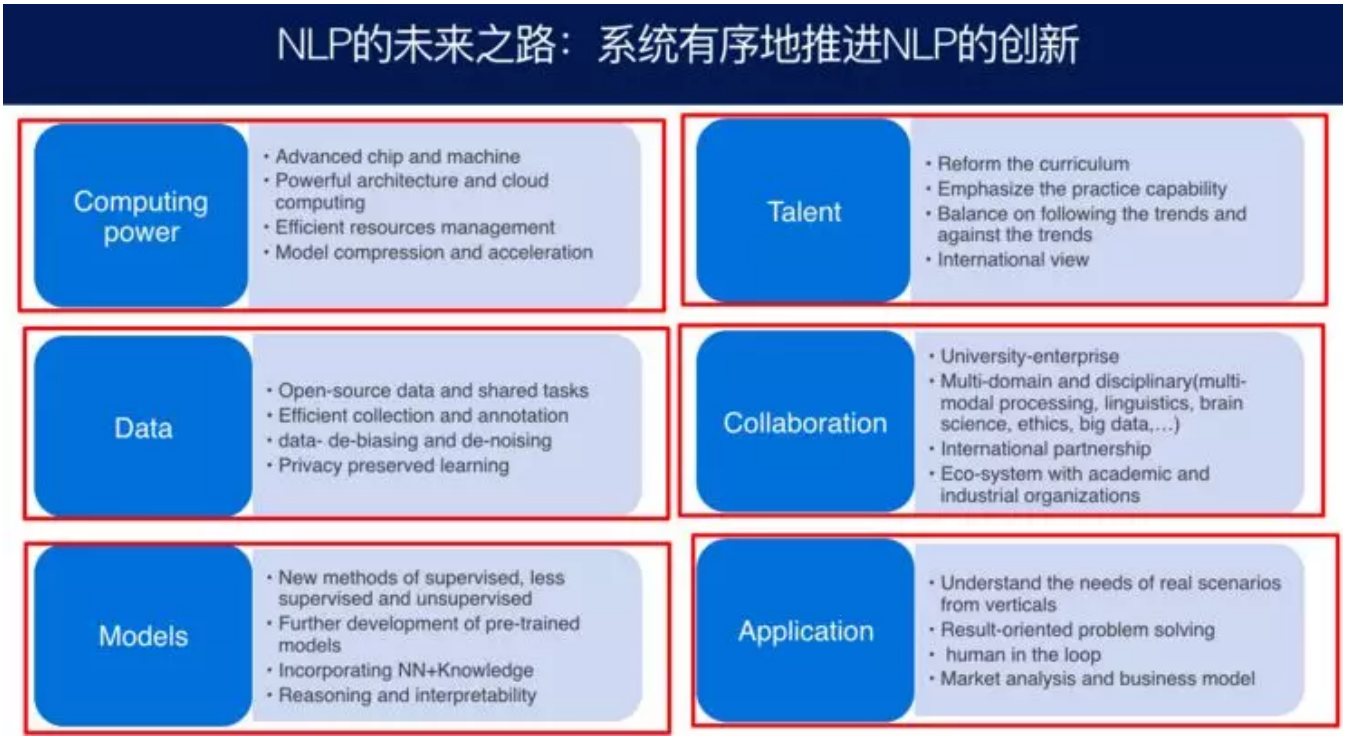
04 未来之路

我们未来到底需要什么样的自然语言处理系统呢？我认为要做出可解释、有知识、有道德、可自我学习的NLP系统。这是一个很高的目标，现在离这个目标差得很远。

我们怎么样来实现这样的目标呢？我们要从具体的任务出发，找出存在的问题。刚才我说了，Rich-Resource存在什么问题呢？上下文建模、数据纠偏、多任务学习、人类知识的理解。再往下，Low-Resource又有什么问题要解决呢？我也列出了一些问题。多轮要解决什么问题呢？就是要解决知识常识、上下文建模、推理机制、可解释等等。

如果我们有所推进的话，我们的认知智能就会进一步提升，包括语言的理解水平、推理水平、回答问题能力、分析能力、解决问题的能力、写作能力、对话能力等等。然后再加上感知智能的进步，声音、图象、文字的识别和生成的能力，以及多模态文、图交叉的能力，通过文字可以生成图象，根据图象可以生成描述的文字等等，我们就可以推进很多人类的应用，包括搜索引擎、智能客服，包括教育、财政、电子商务等等各个方面的应用。也可以把AI技术用在我们的产业上，帮助产业实现数字化转型。

要想实现这件事其实是不容易的，需要各个方面综合努力，所以NLP的未来之路需要不同的公司、学校、政府、企业、投资等等各个角度进行配合。



我这里总结一下，主要有6个角度非常重要。

第一是计算机的能力，刚才说到芯片、存储器、云计算和管理，与之有关的还有模型压缩和加速问题。

第二是数据方面，数据非常重要，全社会都要贡献自己的数据，然后取长补短，大家一起努力。数据上面还有一个隐私保护下的学习，这是非常重要的一点。

第三是模型，刚才说到很多，有监督的学习、无监督的学习、少监督的学习，然后是预训练模型，还有神经网络跟人类知识和常识如何结合，把推理和可解释性融入到我们的学习体系之中。

第四是人才培养，一定要靠人来实现整体的过程。人才如何进行培养呢？要注重实践性，让他们有很强的实践意识，而不是天天去推公式，还要有逻辑上的理解。

第五是合作，校企合作、不同学科的合作、国家的合作、企业界、投资界、政府各个方面的合作，形成一个生态，大家在里面各得其所，来稳步推进。

第六是强调应用，通过应用获得真实的数据、用户的反馈，然后改进我们的系统，也通过应用提升学生的动手能力，也是通过应用使我们了解人和机器在一个真实的系统里如何相得益彰、互相配合，实现人工智能和人类智能的双向结合。

谢谢大家！