# 深度 | 刘群：基于深度学习的自然语言处理，边界在哪里？

丛末 AI科技评论 8月31日

**四大边界：数据边界、语义边界、符号边界和因果边界**

作者 | 丛 末

编辑 | Camel

当前，深度学习之于自然语言处理，有其局限性。那么它所能起作用的边界在哪里呢？对此问题，我们应当深思。

近日，在北京语言大学举办的第四届语言与智能高峰论坛上，华为诺亚方舟实验室语音语义首席科学家刘群教授高屋建瓴，细致分析了深度学习时代NLP的已知与未知。

他从自然语言处理基于规则、统计到深度学习的范式迁移出发，探讨了深度学习方法解决了自然语言处理的哪些问题以及尚未解决哪些问题。

刘群教授认为尚未解决的这些问题最终是由深度学习的四大边界——数据边界、语义边界、符号边界和因果边界所共同造成的。要想在这些尚未解决的问题上寻找突破，需要从深度学习的这些边界出发，去探索新的解决方案。

这个报告主题为《基于深度学习的自然语言处理：边界在哪里？》。可谓是站在NLP塔尖上对整个领域的复盘。

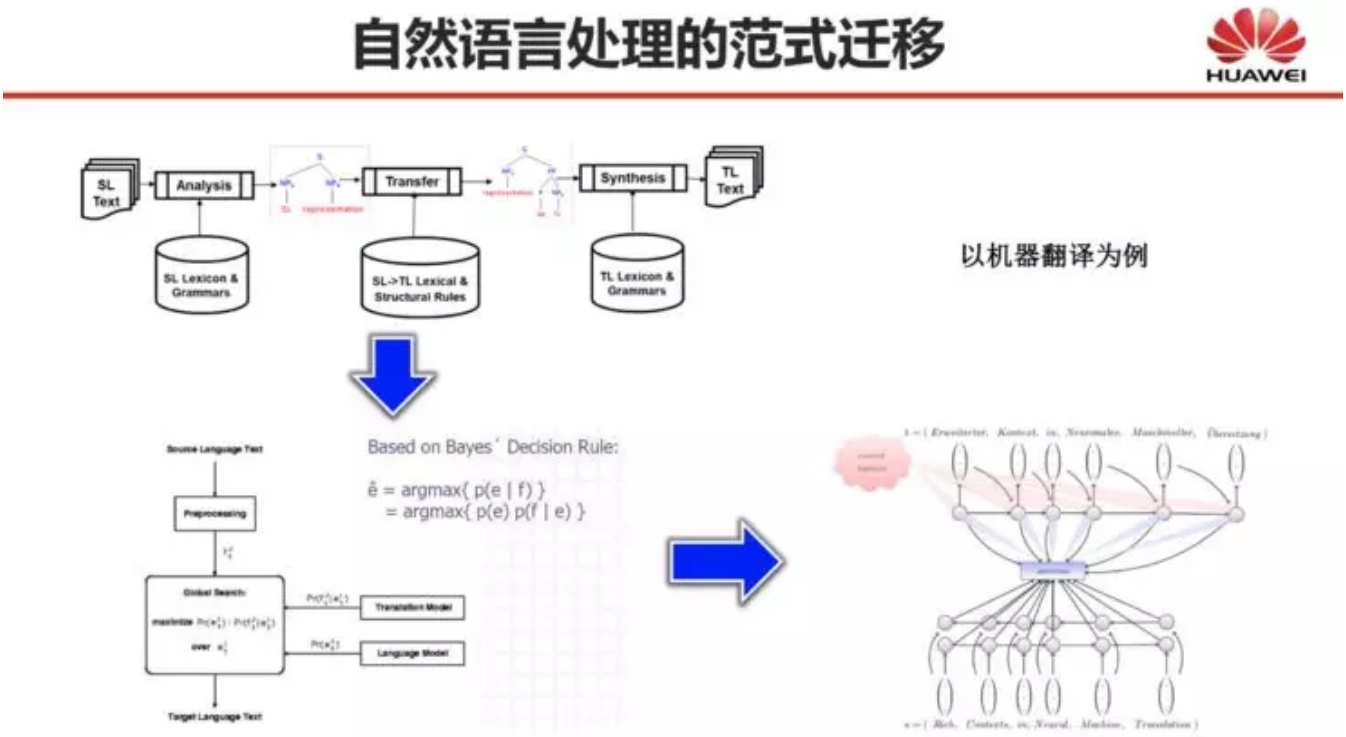
我们一起来欣赏~

**报告正文：**


感谢大会给我这个机会来这里跟大家进行一次交流。今天我不讲我的具体工作，而是讲我多年来研究机器翻译、自然语言处理的一些体会和感想，从更加抽象的层面讨论一些问题，这些想法不一定成熟，如有不恰当的地方，希望大家指正！

## 自然语言处理的范式迁移：从规则、统计到深度学习

相信大家对自然语言处理的范式迁移，都深有体会。以机器翻译为例，很多年前大家都是采用基于规则的方法，基本思想是依靠人来写规则并教机器如何去翻译。后来，大家也慢慢发现这条路走不通，因为人不可能将所有的规则都写穷尽，并且也写不出大量太细的规则。




因此大家之后就转向了基于统计的机器翻译方法，即给机器一堆语料让机器自己去学习翻译规则，不过它学到的还是一些符号层面的规则，但被赋予了概率。到一定程度后，统计机器翻译就遇到了一些瓶颈，也很难再度提高。

随着这几年来深度学习方法的引入，机器翻译的水平又有了一个大幅提高，使得机器不再在符号层面做翻译，而是将整个推理过程映射到一个高维空间中，并在高维空间中进行运算。不过，我们只能理解输入输出而不知道其在高维空间中具体如何进行运算的，并且机器自动学习了什么东西，我们也说不太清楚。

下面我试图来探讨几个问题：一是深度学习解决了自然语言处理的哪些问题？二是还有哪些自然语言处理问题是深度学习尚未解决的？三是基于深度学习的自然语言处理，其边界在哪里？

## 深度学习解决了自然语言处理的哪些问题？

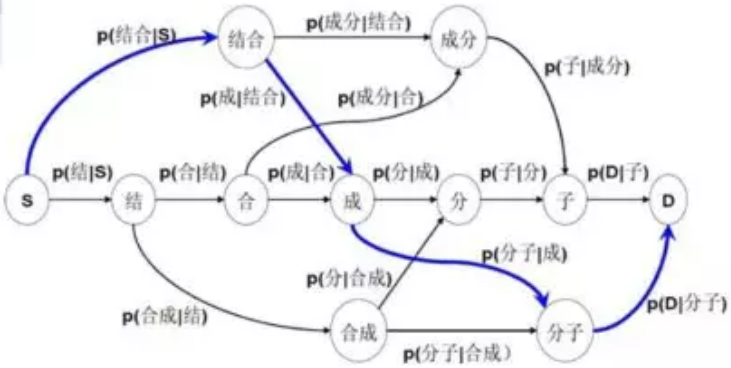
自然语言处理领域有很多难题，此前研究者费了好大劲去解决的问题，深度学习方法出现以后，一些问题被很好地解决了，或者虽然还没有彻底解决，但是提供一个很好的框架。这些问题主要包括：词语形态问题、句法结构问题、多语言问题、联合训练问题、领域迁移问题以及在线学习问题。这里我主要讲下前四个问题，不对后两个问题进行展开。

### 词语形态问题

词语形态问题，即 Morphology，研究的是词的构成。在中文中，它体现在词的切分上，在英语等大部分其他语言中则主要体现在形态的分析上。其中词语切分在包括机器翻译在内的中文信息处理中，曾是一个非常令人头痛的问题，我们也花了很多精力去解决。

## 语言形态问题

| | |
|---|---|
| 研究/生命/的/起源 | 研究生/命/的/起源 |
| 他/从/马/上/下来 | 他/从/马上/下来 |
| 乒乓球/拍卖/完了 | 乒乓/球拍/卖/完了 |
| 和/特朗普/通话 | 和/特朗/普通话 |




汉语词语切分歧义

在基于规则和基于统计的机器翻译方法下，词语形态分析是机器翻译首先需要解决的问题。

对于中文而言，由于基于汉字的翻译效果很差，因而分词是必须解决的问题，也就是说如果不做分词或分词做得不好，即便用统计方法，效果也会很糟糕。然而分词本身又面临很多问题，因为中文词语本来就不是一个定义很明确的单位，导致分词缺乏统一的规范，分词粒度难以把握。

而中文以外的很多语言都存在形态问题，其中英文的形态问题比较简单，因为英语词的变化比较少。而很多其他语言的变化是很多的，例如法语有四五十种变化，俄语则更多。另外以土耳其和波斯语为例的黏着语，一个词可能有上千种变化，即一个词后面可以加很多种词缀，这对于自然语言处理尤其是机器翻译而言，是非常棘手的。

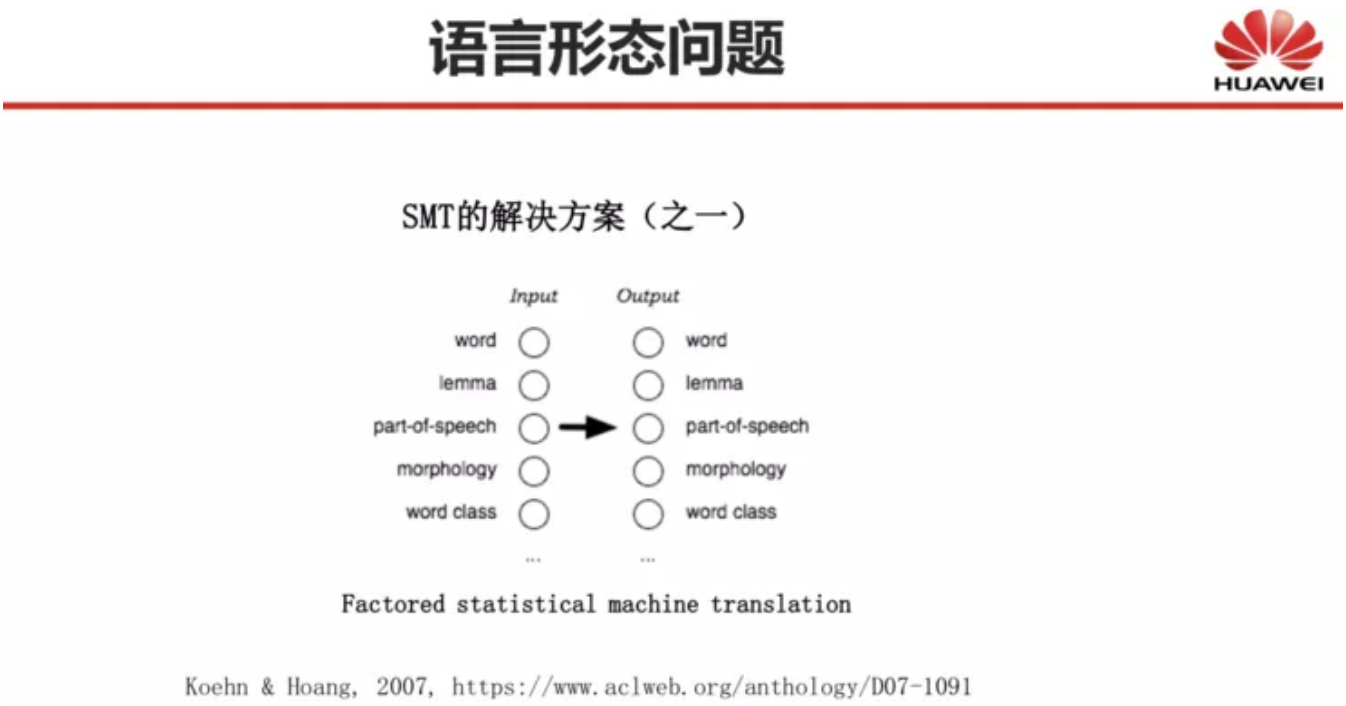
## 语言形态问题

| Word | Translation |
|---|---|
| **Turkish:** | |
| **terbiye** | good manners |
| **terbiye**+siz | rude |
| **terbiye**+siz+lik | rudeness |
| **terbiye**+siz+lik+leri | their rudeness |
| **terbiye**+siz+lik+leri+nden | from their rudeness |
| **terbiye**+siz+lik+leri+nden+mis | it was because of their rudeness |
| **Farsi:** | |
| **drāmd** | income |
| pr+**drāmd** | wealthy |
| pr+**drāmd**+tar | more wealthy |
| pr+**drāmd**+tar+in | the most wealthy |
| pr+**drāmd**+tar+in+hā | the most wealthy people |
| pr+**drāmd**+tar+in+hā+yshān | the most wealthy group of them |
| pr+**drāmd**+tar+in+hā+yshān+nd | they are the most wealthy group of them |

复杂形态语言的机器翻译

而且对于这些形态丰富的语言而言，分析的难度也很大，一般只有语言学家才能把词语的形态说得比较清楚。同时，形态本身其实是一层结构，所有统计机器翻译都建立在某种结构的基础上，例如词语层、短语层、句法层，或者说基于词的、基于短语、基于句法的方法，那如果想在这些结构中再加入一层形态结构，统计机器翻译的建模就会变得非常困难。

在统计机器翻译时代，复杂形态的语言处理非常困难，对此**有一个比较著名的方法叫做 Factored statistical machine translation，即基于要素的翻译方法**，就是将一个词分成很多要素，然后分别翻译每个要素，最后汇总起来。不过我很不喜欢这个方法，因为我认为它不够优雅，且非常冗余，效果也不是很好。




然而语言形态这个问题在神经网络框架下就基本不成问题了，这个领域的研究者对中文分词的讨论也不太多了，虽然也有一些关于如何在神经网络框架下将词分得更好的探索，我也看到过几篇挺有意思的相关文章，但是对于机器翻译而言，中文分词已经构不成根本性挑战了，因为现在机器翻译基本上可以不做分词了，大部分中文机器翻译系统基本上基于汉字来实现，性能跟基于词的系统比没有太大区别。

针对形态复杂的语言，现在提出了一种基于subword（子词）的模型或基于character（字符）的机器翻译模型，效果也非常好。我认为这是一个统一且优雅的方案。

自动化所张家俊老师他们的一篇论文就介绍了基于子词的模型方案的解决思路，如下图所示，第一行是标准的中文，第二行是做了分词以后的。现在一般系统基于汉字即可，就是第三行，但是我们也可以做分词，比如第五行做BPE，将“繁花似锦”分成“繁花”、“似”、“锦”这三个子词部分。

# 语言形态问题

HUAWEI

NMT的解决方案

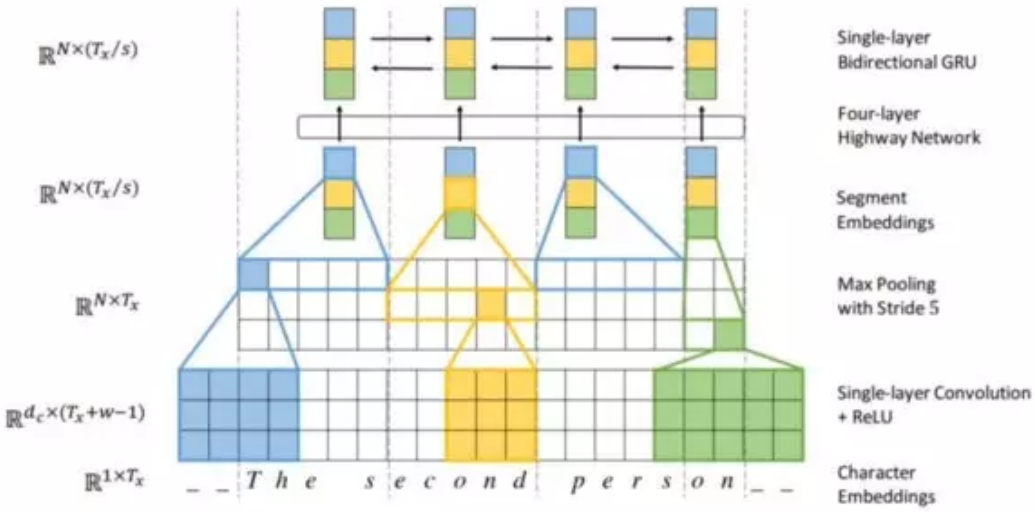
| | |
|---|---|
| Sentence: | 龙年新春，繁花似锦的深圳处处洋溢着欢乐祥和的气氛。 |
| Word: | 龙年 新春， 繁花似锦 的 深圳 处处 洋溢 着 欢乐 祥和 的 气氛。 |
| Character: | 龙 年 新 春， 繁 花 似 锦 的 深 圳 处 处 洋 溢 着 欢 乐 祥 和 的 气 氛。 |
| Hybrid : | <B>龙 <E>年 新 春， <B>繁 <M>花 <M>似 <E>锦 的 深圳 处处 洋溢 着 欢乐 祥和 的 气氛。 |
| BPE: | 龙年 新春 ， 繁花@@ 似@@ 锦 的 深圳 处处 洋溢 着 欢乐 祥和 的 气氛 。 |
| Wordpiece: | _龙 年 _新春 _， _繁花 似 锦 _的 _深圳 _处处 _洋 溢 _着 _欢乐 _祥 和 _的 _气氛 _。 |

图片来自：Wang et al. 2017 arxiv:1711.04457

基于字符的模型则是从字母的层面来做，对英文一个字母一个字母地建模和翻译，翻译效果也非常好。所以我认为在神经网络框架下，形态问题基本上不是什么太大的问题。

# 语言形态问题

HUAWEI




图片来自： Lee et al., TACL 2017, https://aclweb.org/anthology/Q17-1026)

## 句法结构问题

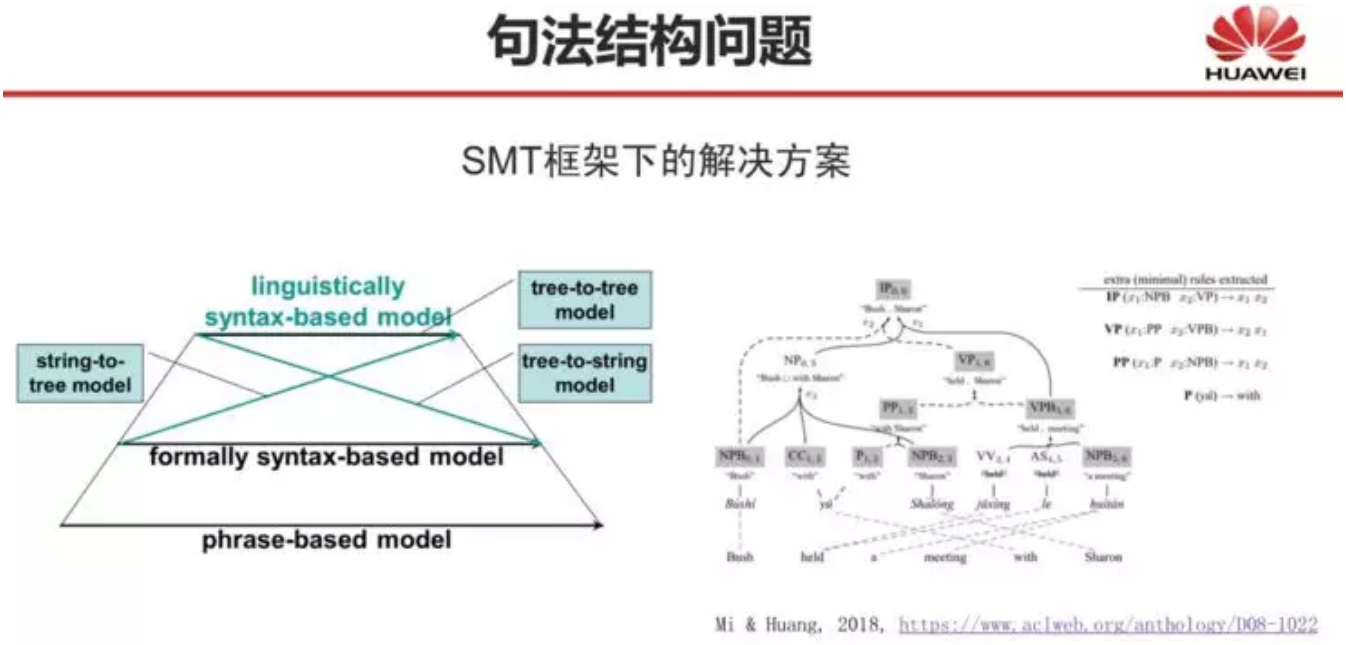
下面看句法结构问题。

无论是在基于规则还是基于统计的机器翻译框架下，句法分析对机器翻译的质量都起着重要的影响作用。其中在基于统计的机器翻译中，基于短语的方法获得了很大成功，因此现在大部分统计方法都不做句法分析。

但是对于中英文这种语法结构相差较大的语言而言，做句法分析要比不做句法分析的结果好很多，所以句法分析还是很重要的。不过句法分析难度很大，一方面会带来模型复杂度的增加，另一方面句法分析本身存在的错误会影响翻译的性能。

而目前在神经网络机器翻译框架下，神经网络可以很好地捕捉句子的结构，无需进行句法分析，系统可以自动获得处理复杂结构句子翻译的能力。

大概 2005 年至 2015 年期间，我一直在做基于统计的机器翻译，也就是研究如何在统计方法中加入句法方法，在这么多年的研究中，我们提出了很多种方法也写了很多篇论文，下图中的这些模型概括了我们之前提出的那些方法。




我们的工作主要聚焦于树到树、树到串的方法。美国和欧洲很多学者在做串到树的方法，树到树的方法做得都比较少。另外我们还做了一些森林方法的研究，即如何避免句法分析错误。不过，这些问题在神经网络框架下基本上也不存在了。

举例来说，“第二家加拿大公司因被发现害虫而被从向中国运输油菜籽的名单中除名”是一个好几层的嵌套结构，但是机器翻译的结果“The second Canadian company was removed from the list of transporting rapeseed to China due to the discovery of pests”在结构上翻译得很好。下面一个例子在结构上也没有什么错误。

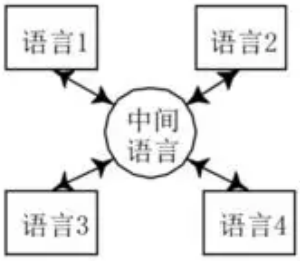

神经网络机器翻译方法是没有用到任何句法知识的，仅凭从网络中学到的复杂结构就能实现这么好的效果，这样的话，对机器翻译来说做句法分析就没有太大意义了。当然句法结构并不是完全没有意义，现在也有不少人在研究，但是我认为这已经不再是机器翻译的一个主要难点了。

## 多语言问题

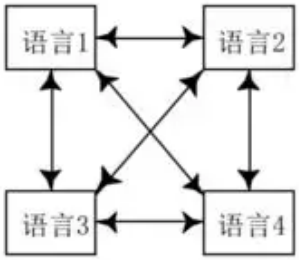
曾经，我们机器翻译研究者的一个理想，就是在基于规则的时代实现多语言翻译。当时很多人都在做多语言翻译，甚至是中间语言翻译，如下图，中间语言翻译其实是一个理想的方案，因为多语言的互相翻译通过某个中间语言来实现，是能够节省很多成本的：如果使用中间语言，开发系统的数量随翻译语言的数量呈线性增长；否则，开发系统的数量随翻译语言的数量呈平方增长。

但在基于规则方法的机器翻译时代，中间语言的方法是不可行的，正如日本机器翻译专家 Makoto Nagao 教授曾经说过的一句话，当我们使用中间语言的时候，分析阶段的输出结果必须采用这样一种形式：这种形式能够被所有不同语言的机器翻译所使用。然而这种细微程度实际上是不可能做到的。

## 多语言问题




中间语言方法




转换方法

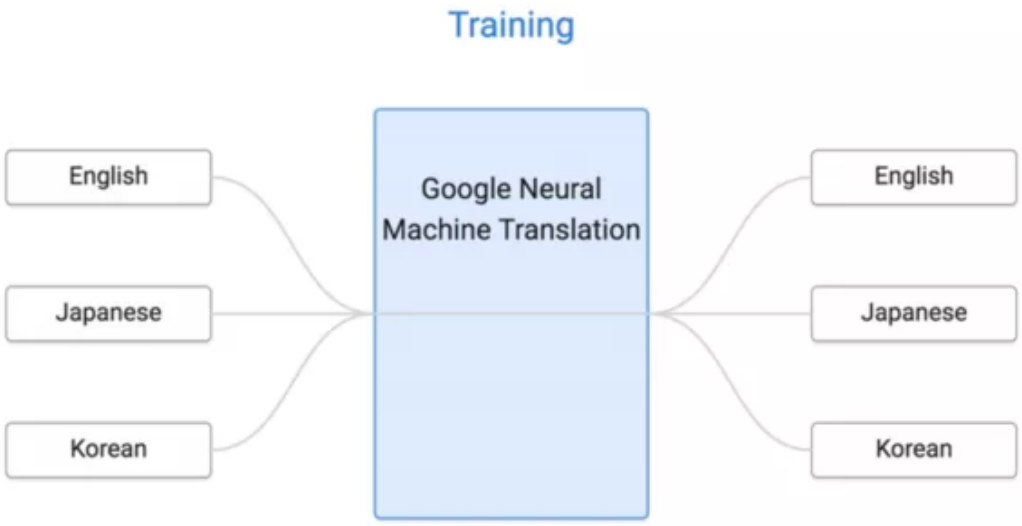
Makoto Nagao (Kyoto University) said: ".. when the pivot language [i.e. interlingua] is used, the results of the analytic stage must be in a form which can be utilized by all of the different languages into which translation is to take place. This level of subtlety is a practical impossibility." (Machine Translation, Oxford, 1989)

在基于统计方法的机器翻译时代，普遍采用的是 Pivot 方法，即在两个语言的互译中，先将所有语言翻译成英语，再翻译成另一种语言。这样的话就能够使得多语言机器翻译成为可能。

但是这种方法也存在一些问题，即会导致错误传播和性能下降。另一方面，我们做多语言翻译的另一个想法是希望能够利用多语言之间互相增强的特点，即很多语言有相似的特点，因而如果无法利用上这种增强的话，这种方法就并非那么理想了。

在神经网络机器翻译时代，谷歌就直接利用中间语言的方法做出了一个完整且庞大的系统，将所有语言都放在一起互相翻译以及将所有文字都放在一起编码。虽然这个系统目前还不是很完美，但是距离理想的 Interlingua 已经很接近了。

## 多语言问题




Zero-Shot Translation with Google's Multilingual Neural Machine Translation System
https://ai.googleblog.com/2016/11/zero-shot-translation-with-googles.html

之后,谷歌又推出了Multilingual BERT，将104种语言全部编码到一个模型里面，这在原来是不可想象的。


虽然这两个方法现在还无法彻底解决多语言问题，但是它们整个框架非常漂亮，效果也非常好，所以我觉得针对这两个方面，我们还有很多事情可以做。

### 联合训练问题

在统计机器翻译时代，因为各模块都是互相独立训练的，导致错误传播的问题很严重，所以联合训练也成为了提高性能的有效手段。

但联合训练本身又会导致模型复杂度的大大增加，使得开发和维护变得困难。同时由于搜索范围急剧扩大，系统开销也严重增加。不仅如此，由于模块太多，只能有限的模块进行联合训练，所以不可能将所有模块都纳入联合训练。

而在神经网络机器翻译框架下，端到端训练成为标准模式，所有模块构成一个有机的整体，针对同一个目标函数同时训练，有效避免了错误传播，提高了系统性能。

## 还有哪些自然语言处理问题深度学习尚未解决?

由于深度学习的应用，我们以前费很大劲去做的一些事情，现在基本上不需要再去做了。但是深度学习本身还是存在很多问题的，包括资源稀缺问题、可解释性问题、可信任问题、可控制性问题、超长文本问题以及缺乏常识问题等等。

### 资源稀缺问题

资源稀缺问题大家都很清楚，然而这个问题远比我们大部分人想象的要严重得多。一般而言，对于常见语言，机器翻译可以做得很好，然而现实世界中有几千种语言，曾经就有一篇报告统计出7000 多种语言，当然有文字的语言并没有这么多，其中绝大部分语言都是资源稀缺语言，并且绝大多数专业领域实际上也都是资源稀缺的领域。

以下面针对医疗领域的 WMT 2019 评测为例，它的语料库包括 3000多个文档、4 万多个句子。在机器翻译领域，几百万个句子的语料已经是小数量的了，商业系统基本上都有好几千万句子的训练语料。然而这里才4万多个句子，是存在严重资源稀缺问题的，翻译的质量也非常糟糕，基本上是不可接受的。另外从数据上来看，西班牙语有10万多个，法语有7万多个，中文没有，也就是说基本收集不到中文医疗领域的翻译数据。

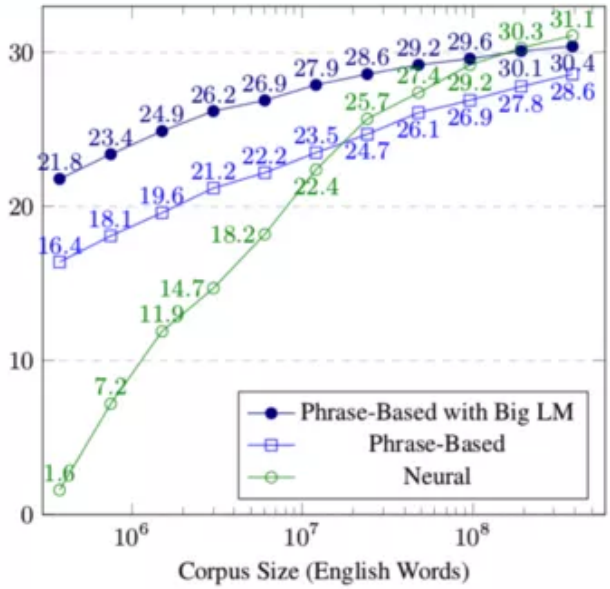
以WMT2019的Biomedical MT Task为例：

| Language pairs | Medline training | | Medline test | | Terminology test |
|---|---|---|---|---|---|
| | Documents | Sentences | Documents | Sentences | Terms |
| de/en | 3,669 | 40,398 | 50 | 589 | - |
| en/de | | | 50 | 719 | - |
| es/en | 8,626 | 100,257 | 50 | 526 | - |
| en/es | | | 50 | 599 | 6,624 |
| fr/en | 6,540 | 75,049 | 50 | 486 | - |
| en/fr | | | 50 | 593 | - |
| pt/en | 4,185 | 49,918 | 50 | 491 | - |
| en/pt | | | 50 | 589 | - |
| zh/en | - | - | 50 | 283 | - |
| en/zh | | | 50 | 351 | - |

在工业界，想要解决的大部分问题都是没有标注语料的，需要自己去标，然而也基本上没有那么多钱去对很多的语料做标注。所以资源稀缺问题要比我们想象的严重得多。

资源稀缺对神经网络机器翻译的影响很大。从下图来看，上面两条线指基于统计的机器翻译方法，下面这条线指神经网络机器翻译方法，我们可以看到神经网络的方法只有在语料很多的情况下，表现才能超过统计方法，在语料不够大时，表现并不比统计方法更好。

# 资源稀缺问题




Koehn & Knowles, 2017, https://arxiv.org/pdf/1706.03872.pdf

## 可解释性问题和可信任问题

我们给神经网络输入一个东西，它就会输出一个结果，然而其在高维空间的计算过程我们是不知道的，这就存在可解释问题。但我认为这个问题的严重性要视情况而定，我们有时候需要解释性，却并不是所有时候都需要解释性，比如人脑做决定，有时间可能只是灵机一动的灵感，至于怎么来的，人自己都不一定能够解释得清楚。

而可解释性带来的一个更重要的问题是可信任问题。一些关键性领域如医疗领域，比如说病人看病，如果系统给出一个癌症的诊断而无法给出理由的话，病人是不敢治疗的。所以在这些关键性的应用上，可解释性是非常重要的，因为这个问题会导致信任问题。

机器翻译中的一个可信任问题是翻译错误。比如说重要的人名、地名和机构名是不应该翻错的，以翻译美国政府的一个工作报告为例，如果使用之前的语料来训练，机器就会直接将美国总统（特朗普）翻译成布什总统了，这就是一个很严重的错误了。

第二个可信任问题是翻译出来的意思与原意相反，这在机器翻译中也很常见，且较难避免，因为这种意思相反的表达在语料库中的统计特征是非常接近的，都是在陈述同一件事情，因此在机器翻译中很容易导致翻译出与原意相反的结果。

第三个可信任问题则是机器翻译犯一些过于幼稚的、完全不该犯的问题，这就会直接给人带来不信任感。

## 可控制性问题

由于系统有时候的效果总不能令人满意或总出现错误，所以我们希望系统变得可控，即知道怎么对其进行修改从而避免犯这种错误。

基于规则的机器翻译方法中，我们是可以通过修改规则来纠正；基于统计的机器翻译方法，虽然改的方式绕一点，但是统计的数据都是可解释的，我们可以在其中加上一个短语表来纠正，而在神经网络机器学习方法中，我们几乎是不能进行修改的。

比如对于重要的人名、地名、机构名、术语，我们希望机器严格按照给定的方式进行翻译，不能随便乱翻。我之前在爱尔兰的时候带学生做过这方面的比较早期的工作，目前的引用量还比较高，现在我们对这项工作进行了一些改进，可以比较好地解决机器翻译的可控制性问题，但是这项工作还仅适用于机器翻译这一特例，而无法做到通用化去解决神经网络在整个自然语言处理领域存在的可控制性问题。

## 超长文本问题

现在的神经网络机器翻译在处理长文本方法取得了很大的进步。早期的神经网络翻译系统常被人诟病：短句子翻译得好但长句子却翻译得很糟糕。而现在，这种情况已经得到了非常大的改善，一般的长句都翻译得不错，但漏翻等小错误还是不可避免。

现在基于长文本训练的语言模型如BERT、GPT，其训练的文本单位一般都是几百字到上千字，所以长度在这个范围内的文本处理没有太大问题，并且 GPT生成一千字以内的文本都可以生成得非常流畅。

目前机器翻译能够处理比较长的文本，但是不能说长文本问题就解决了，它本身还存在很多挑战：

- 一个是基于篇章的机器翻译问题，不光是我们，学术界还有很多同行都在研究这个问题。基于篇章的机器翻译实验证明，对改进翻译质量起作用的上下文只有前1-3个句子，更长的上下文反倒会降低当前句子的翻译质量。按理来说，上下文更长，机器翻译的效果应该是更好的，那为什么反而翻译得更差呢？这是不合理的。

- 另一个是预训练语言模型问题。现在机器翻译的训练长度一般是几百字到上千字，然而实际处理的文本可能不止一千字，比如说一篇八页的英文论文，起码都两三千字了。因此预训练语言模型在实际处理更长文本的时候，还是会遇到很多问题，这种情况下，语言模型消耗计算资源巨大，计算所需时空消耗会随着句子长度呈平方或者三次方增长，所以现有模型要想支持更长的文本，还有很多问题尚待解决。

## 缺乏常识问题

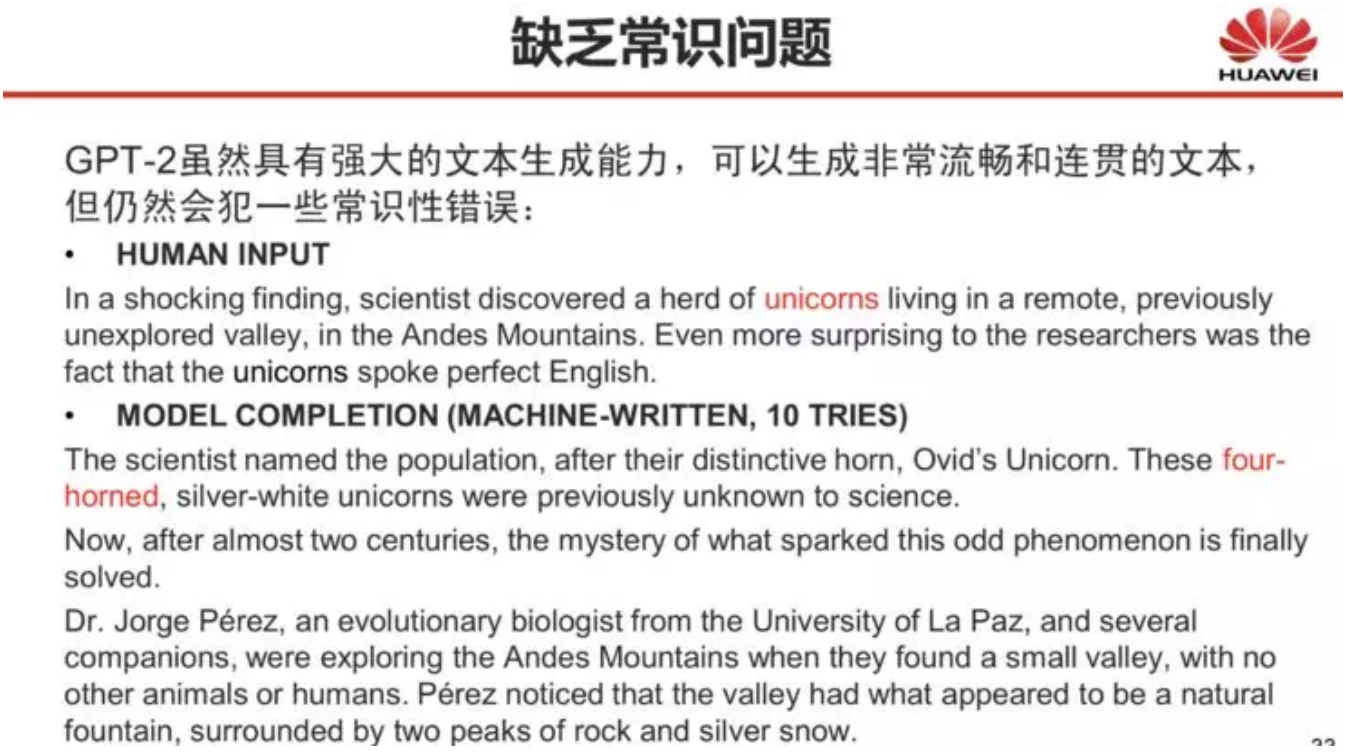
这里我以不久前去世的董振东先生提供的例子为例（如下图所示），“bank”是翻译中一个经典的歧义词，有“银行”和“岸”的意思，在什么语境下翻译成哪个意思，对于人来说很容易理解，但是即使有 fishing、water这样的相关提示词存在，谷歌翻译器还是将这个词翻译成了“银行”。在神经

网络机器翻译时代，这种常识性错误依旧比较普遍存在。


- Bank（银行、岸）这样一个经典的歧义词在NMT时代仍然无法避免翻译错误，即使有fishing、water这样的相关提示词存在

（感谢董振东老师提供的例子）

另外一个例子就是 GPT 的文本生成。GPT 在文本生成方面已经做得很好了，然而即便如此，还是会犯很多常识性的错误。以下面这个经典案例为例，前面人类输入的句子是“在一项研究中，科学家们发现了一群独角兽，它们生活在安第斯山脉一个偏远的还没被开发山谷中，更令人惊讶的是这些独角兽会说一口流利的英语”，其中“独角兽会说一口流利的英语”在现实生活中是荒唐、完全不可能的事。然而，GPT 系统就根据这一句话生成了一个故事。




故事写得很漂亮，但是存在错误，比如第一句就是错误的，“科学家根据它们独特的角，将它们命名为Ovid’s Unicorn，这些有着银色的四个角的独角兽是原来科学家所不曾见过的”这句话本身就矛盾，独角兽怎么会有四个角呢？这是很明显的一个逻辑错误。所以常识问题，在机器翻译中依旧是一个非常严重的问题。

## 基于深度学习的自然语言处理，其边界在哪里？

那自然语言处理中哪些问题是可以解决的，哪些是不可以解决的呢？这就涉及到它的边界问题。我认为深度学习有几个重要的边界：数据边界、语义边界、符号边界和因果边界。

## 数据边界

数据边界是限制当前机器翻译技术发展的约束之一，这个比较好理解，就是指数据不够，这是现有方法无法解决的。

## 语义边界

人工智能在很多领域都大获成功，其中在围棋、电子竞技等项目上获得的成功最大，包括早期还没有深度学习乃至统计方法时，在 Winograd 系统上就很成功了，为什么会取得这么大的成功？

我认为这是因为这些领域能够对客观世界的问题进行精确建模，因此能做得很好；而现在自然语言处理系统大部分都无法对客观世界进行精确建模，所以很难做好。另外比如像智能音箱、语音助手系统能够取得一定成果，很大程度上也是因为这些系统对应着明确定义的任务，能对物理世界建模，不过一旦用户的问话超出这些预定义的任务，系统就很容易出错。

机器翻译的成功是一个比较特殊的例子，这是因为它的源语言和目标原因的语义都是精确对应的，所以它只要有足够的数据而并不需要其他的支撑，就能取得较好的效果。

现在的自然语言处理系统大部分，还只是流于对词语符号之间的关系建模，没 有对所描述的问题语义进行建模，即对客观世界建模。而人理解语言的时候，脑子里一定会形成一个客观世界的影像，并在理解影像后再用自己的语言去描述自己想说的事情。

实际上，自然语言处理的理想状态应该是能够对客观世界进行描述并建模，然而对客观世界建模相当复杂，实现并不容易。以颜色这个属性为例，可以用三个 8 位数进行建模，可以组合出数千万种颜色，但刻画颜色的词语只有数十个，词语和颜色模型的对应关系很难准确地进行描述。

在机器翻译的研究中，对客观世界建模并不新鲜，早期的本体或者知识图谱、语义网络，都是人类专家试图对客观世界建立通用性模型的一种长期努力，其中一项集大成的成果便是知识图谱，但是它目前还没有办法很好地应用到深度学习中来。不过，我认为这是一个很值得探索的方向。

总而言之，我认为自然语言处理的一个理想的改进方向就是做世界模型或语义模型，换句话说，就是不仅仅只做文本间的处理，还必须落地到现实世界中，去对现实世界建模，而知识图谱这是其中一个较为值得探索的具体方向。

## 符号边界

心理学家将人的心理活动分为潜意识和意识，用我的话来理解就是，可以用语言描述的心理活动称作意识，而无法用语言描述的心理活动称为潜意识。

神经网络实际上则是潜意识的行为，可以输入语言和输出语言表达，但是无法对整个推理和计算过程进行描述，这本身就是它的一个重要缺陷。

举一个简单的例子：使用有限状态自动机，可以精确地定义一些特定的表示形式，如数词、年份、网址等等，但再好的神经网络也很难准确地学习到有限状态自动机的表达能力，这是很多实用的自然语言处理系统仍然离不开符号这种规则方法的原因。

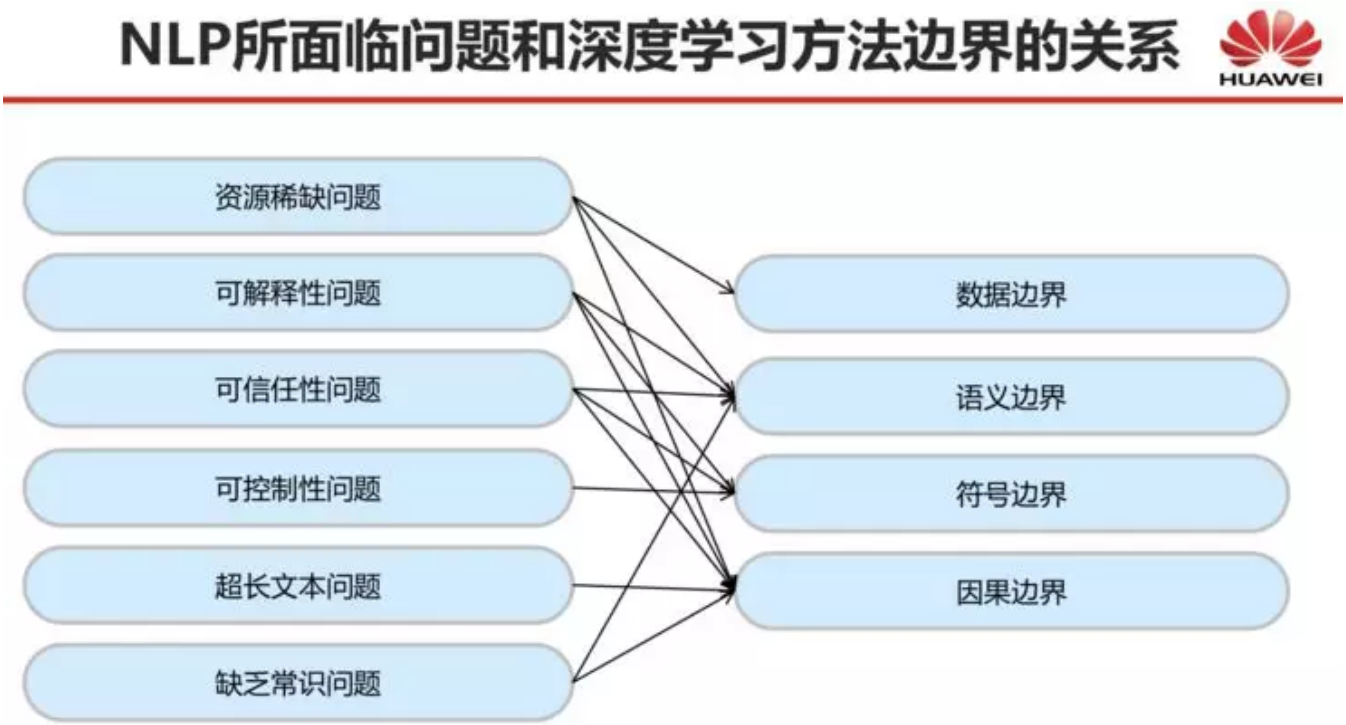
## 因果边界

人类对客观世界中发生的事情中的因果关系都有明确的理解。所以很容易去芜存真，抓住问题的本质。

神经网络无法做到这一点，它根据数据学习到的东西去做出判断而并没有理解真正的因果关系，即并不知道哪些因素是事情发生的真正原因，哪些是辅助性的判断依据，因而很容易做出错误的判断。

实际上，仅仅根据统计数据进行推断，很难得到真正的因果关系。真正的因果关系，只有通过精心设计的实验才能得出，例如药物的有效性，美国、中国药物局都需要花上几十年的时间做实验，最终才能确定出一个因果关系，相当不容易。

今天我讲了基于深度学习的自然语言处理依旧面临的几个问题，而我认为这些问题最终是由我前面提到的四个边界造成的，并且不是由边界中的某一个造成，而是由多个边界的共同干扰所造成的。对此，我用一个关系图来描述这种对应关系，如下图所示。




## 附：问答部分

**听众提问：在统计机器翻译时代，有分词分析、句法分析以及语义分析等共性任务，那在神经网络机器翻译时代是否也有这样一些共性任务呢？**

**刘群：**显然是有的。

一个是预训练语言模型，它实际上就是在将语言当成一个共性任务来处理，其之所以现在取得这么大的成功，我认为某种程度上就是因为这种共性任务的处理方式。

第二个是知识图谱，它其实也是一种共性任务，这个领域的研究者做了这么多年的研究，我认为是非常有意义的，所以我们现在也在想办法将知识图谱和自然语言处理结合起来做研究。

另外在手机助手、音箱等语音对话系统中，也能够体现这种共性任务，比如说系统中的多个技能，包括控制家电、播放音乐等，如果进行单个处理的话，各项技能之间会“打架”，因此就需要将这些问题进行共性任务处理，这样的话就会变得非常复杂，所以对话系统在这种共性任务上的研究，是比较值得探索的。

*感谢刘群教授对本文内容的审阅和校对。*

延伸阅读：

[1] 独家专访华为诺亚方舟刘群：从 26 年学术生涯到执掌华为语音语义团队

[2] 从 ACL 2019 看 NLP 未来发展趋势

[3] 周明：自然语言处理的未来之路

[4] ACL 2019 最佳论文奖出炉，刘群团队获最佳长文奖

[5] 周志华：“深”为什么重要，以及还有什么深的网络

[6] 这次 AI 突破的是麻将！