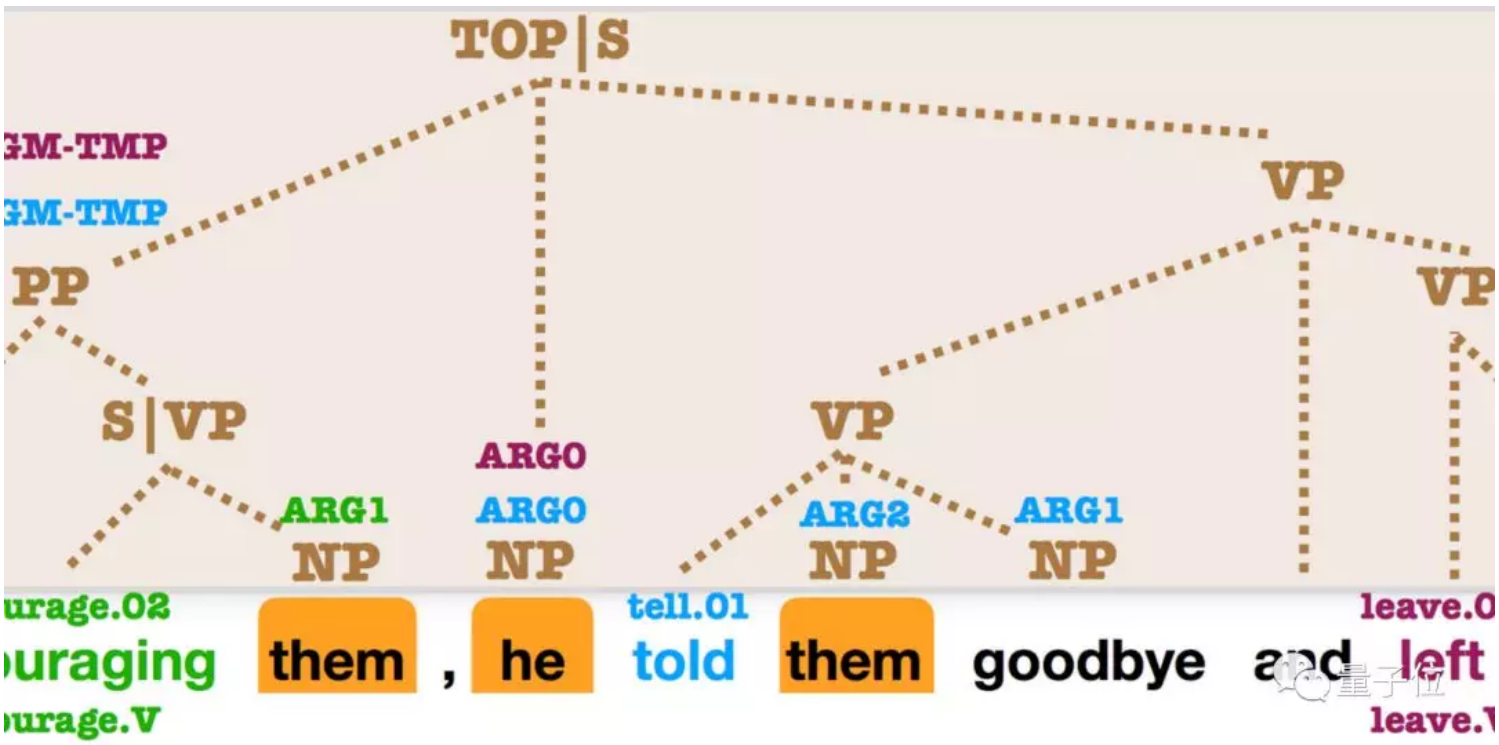
2018 NLP十大创新思路

Sebastian Ruder 深度学习工坊 今天

作者： Sebastian Ruder
编辑：乾明
转载于：量子位(QbitAI)



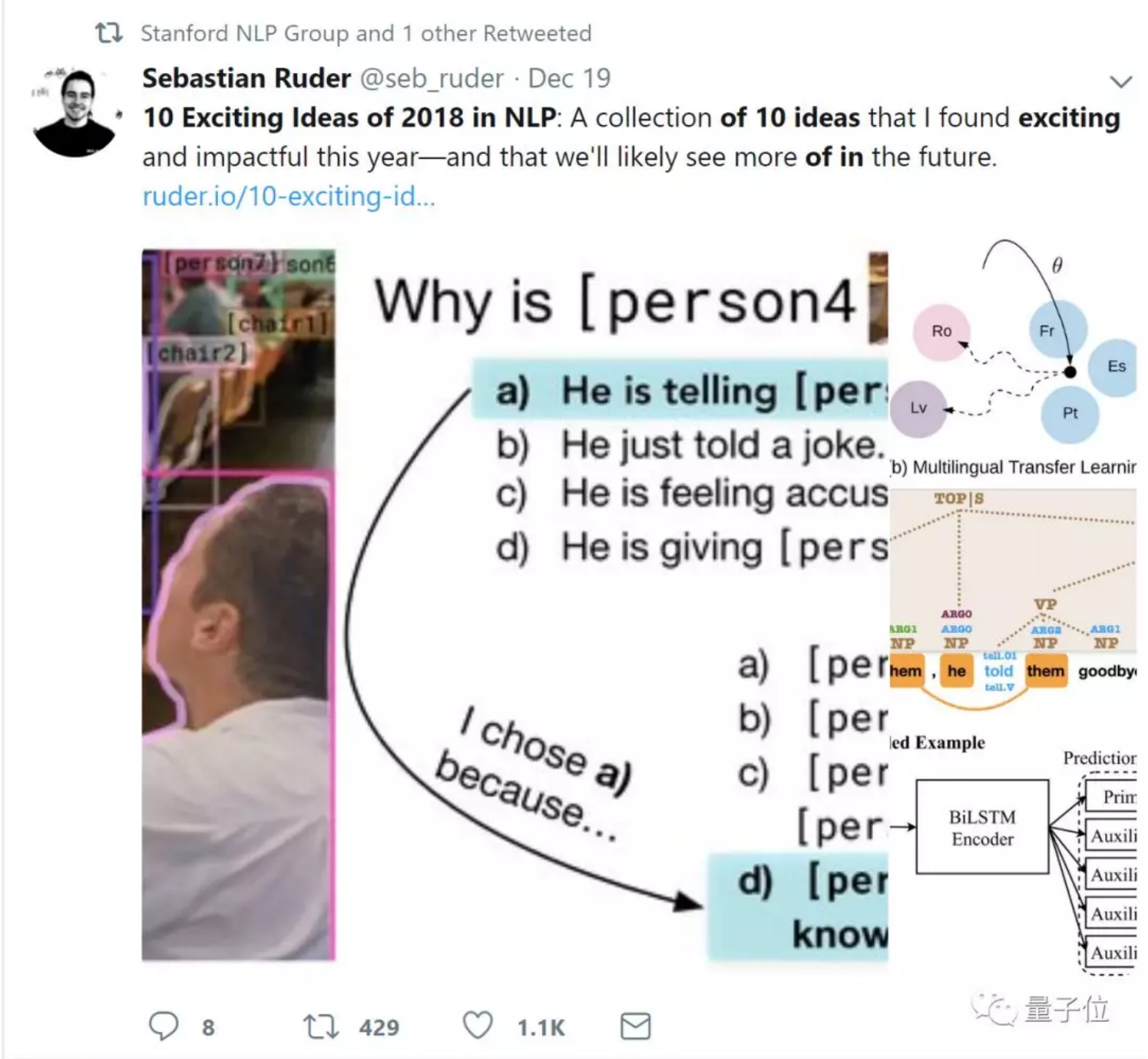
2018年，NLP领域的大年。

最瞩目的莫过于BERT，横扫多种不同的NLP测试，被誉为NLP新时代的开端。

但2018年，不只有BERT。

最近，爱尔兰的NLP研究科学家Sebastian Ruder写一篇文章，基于12篇经典论文盘点了2018年NLP领域令人激动的十大想法。

盘点文章发表后，就在Twitter上收获1100多赞，400多转发。



其中不乏斯坦福NLP这样的业内知名机构转发点赞。甚至有研究者评价称，希望所有的NeurIPS论文，都可以用他的方法来进行总结与解读。

那，这十大想法都是什么呢？

无监督机器翻译（Unsupervised MT）

2018年的ICLR收录了两篇关于无监督机器翻译的论文。虽然有点效果，但与监督系统相比仍然差强人意。

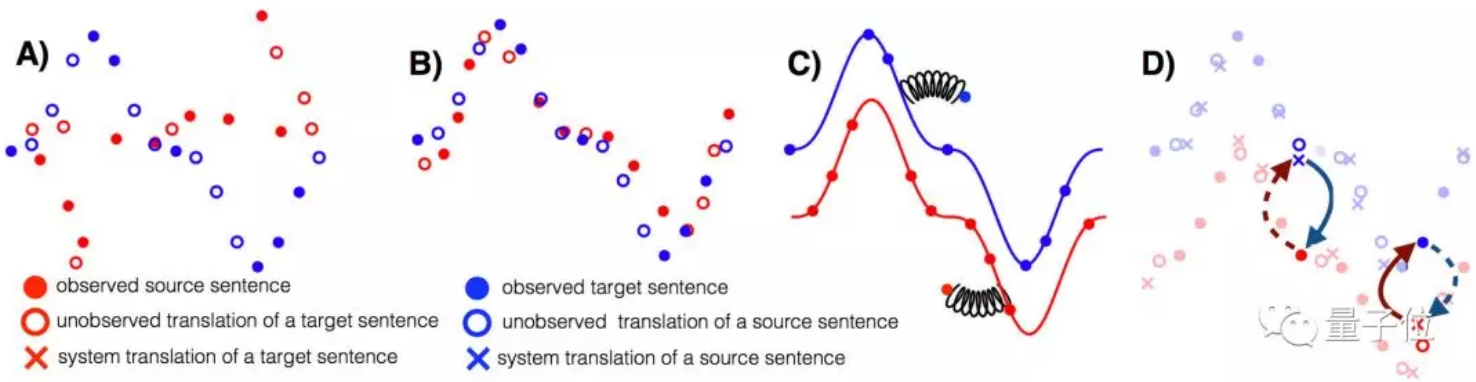
在EMNLP 2018上，这两篇论文的研究团队又提交了两篇论文，大幅改进了研究方法，无监督机器翻译获得了重大进展。

代表性成果：

Phrase-Based & Neural Unsupervised Machine Translation (EMNLP 2018)

https://arxiv.org/abs/1804.07755

> 这篇论文很好地提炼出了无监督机器翻译的三个关键方法：良好的初始化、语言建模和逆向任务建模(通过反向翻译)。
>
> 这三个方法在其他无监督场景中也有用。逆向任务建模可以增强循环一致性，这种一致性已经在不同的方法中得到应用，在CycleGAN中最为突出。
>
> 论文中对两种资源较少的语言，“英语-乌尔都语”和“英语-罗马尼亚语”进行了大量的实验和评估。
>
> 希望在未来能看到更多关于资源较少的语言的研究。



△无监督机器翻译三个关键方法的说明。A：两个单语数据集。B：初始化。C：语言建模。D：反向翻译。

预训练语言模型（Pretrained language models）

这是NLP领域今年最重要的发展趋势。 有很多令人难忘的方法：ELMo、ULMFiT、OpenAI Transformer和 BERT。

代表性成果：

Deep contextualized word representations (NAACL-HLT 2018)

https://arxiv.org/abs/1802.05365

这篇论文提出了ELMo，受到了广泛好评。

除了实证结果令人印象深刻之外，最引人注目的是论文的分析部分，它剔除了各种因素的影响，并对在表征中捕获的信息进行了分析。

词义消歧(WSD)分析(下图左)执行得很好。两者都表明了，LM提供的词义消歧和词性标注（POS）表现都接近最先进的水平。

Model	F_1
WordNet 1st Sense Baseline	65.9
Raganato et al. (2017a)	69.9
Iacobacci et al. (2016)	**70.1**
CoVe, First Layer	59.4
CoVe, Second Layer	*64.7*
biLM, First layer	67.4
biLM, Second layer	*69.0*

Model	Acc.
Collobert et al. (2011)	97.3
Ma and Hovy (2016)	97.6
Ling et al. (2015)	**97.8**
CoVe, First Layer	*93.3*
CoVe, Second Layer	92.8
biLM, First Layer	*97.3*
biLM, Second Layer	96.8

△第一层和第二层双向语言模型的词义消歧(左)和词性标注(右)结果。与基线相比。

常识推理数据集（Common sense inference datasets）

将常识融入到模型中，是NLP最重要的前进方向之一。然而，创建一个好的数据集并不容易，即便是流行的数据集，也存在很大的偏差。

今年，已经有一些很好的数据集试图教模型一些常识，如Event2Mind和SWAG，它们都来自华盛顿大学。但很意外的是，SWAG很快被BERT超越了。

代表性成果：

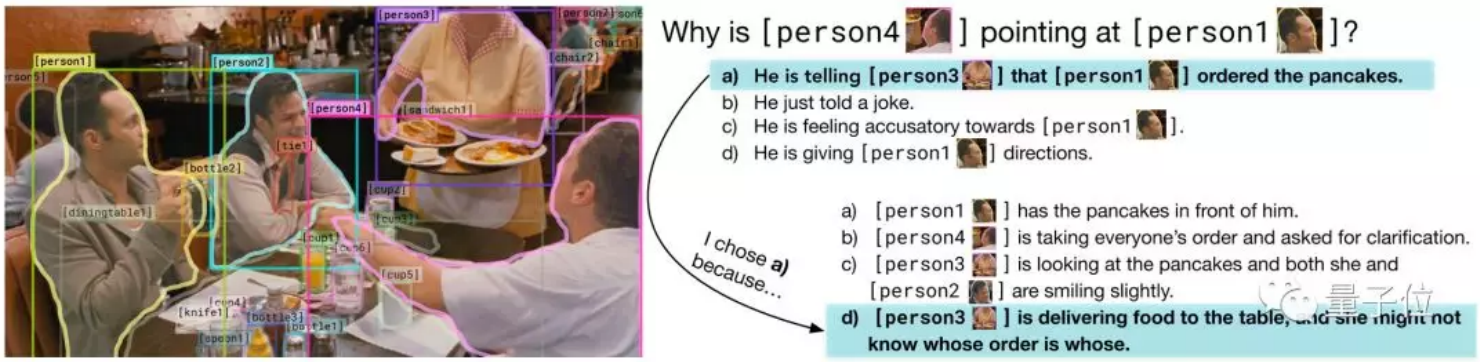
Visual Commonsense Reasoning (arXiv 2018)

http://visualcommonsense.com/

这是首个视觉QA数据集，每个答案都包含对答案的解释。而且，每个问题需要复杂的推理。

创作者想尽办法解决可能存在的偏差，确保每个答案的正确率为25% (每个答案在整个数据集中出现4次，错误答案出现3次，正确答案出现1次)。

这需要使用计算相关性和相似性的模型来解决约束优化问题。希望在创建数据集时，防止可能出现的偏差会成为一个常识。



△VCR：给定一张图片、一个区域列表和一个问题，模型必须回答这个问题，并提供一个解释其答案为何正确的理由。

元学习（Meta-learning）

元学习在少样本学习、强化学习和机器人学习中得到了广泛的应用，最突出的例子是与模型无关的元学习(MAML)。

但在NLP领域，元学习很少有成功的应用。在解决样本数量有限的问题上，元学习非常有用。

代表性的论文：

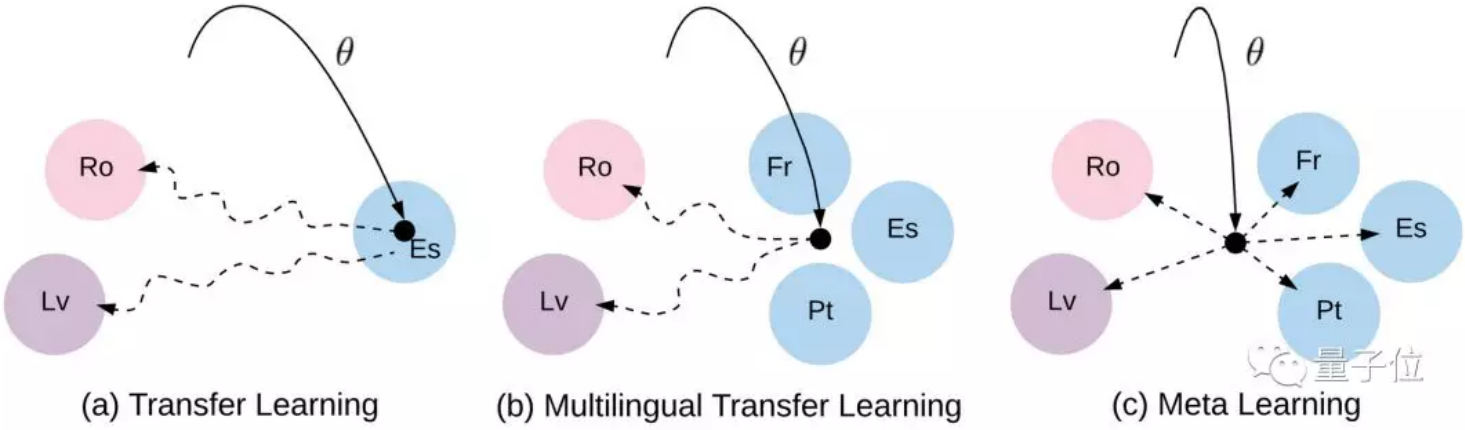
Meta-Learning for Low-Resource Neural Machine Translation (EMNLP 2018)

http://aclweb.org/anthology/D18-1398

这篇论文中，作者使用了MAML，将每一种“语言对 ”都视为单独的元任务。

在NLP领域，用来适应资源较少的语言，可能是元学习的最佳用武之地了。

尤其是将多语言迁移学习(如多语言BERT)、无监督学习和元学习相结合起来的时候，这是一个非常有希望取得进展的方向。



△迁移学习、多语言迁移学习与元学习的区别。 实线：学习初始化。 虚线：微调路径

稳健的无监督方法（Robust unsupervised methods）

今年，我们和其他人观察到，当语言不相似时，无监督的跨语言单词嵌入方法会崩溃。

这是迁移学习中的常见现象，在迁移学习中，源和目标设置之间的差异(例如，领域适应、持续学习和多任务学习中的任务)会导致模型的效果变差或崩溃。

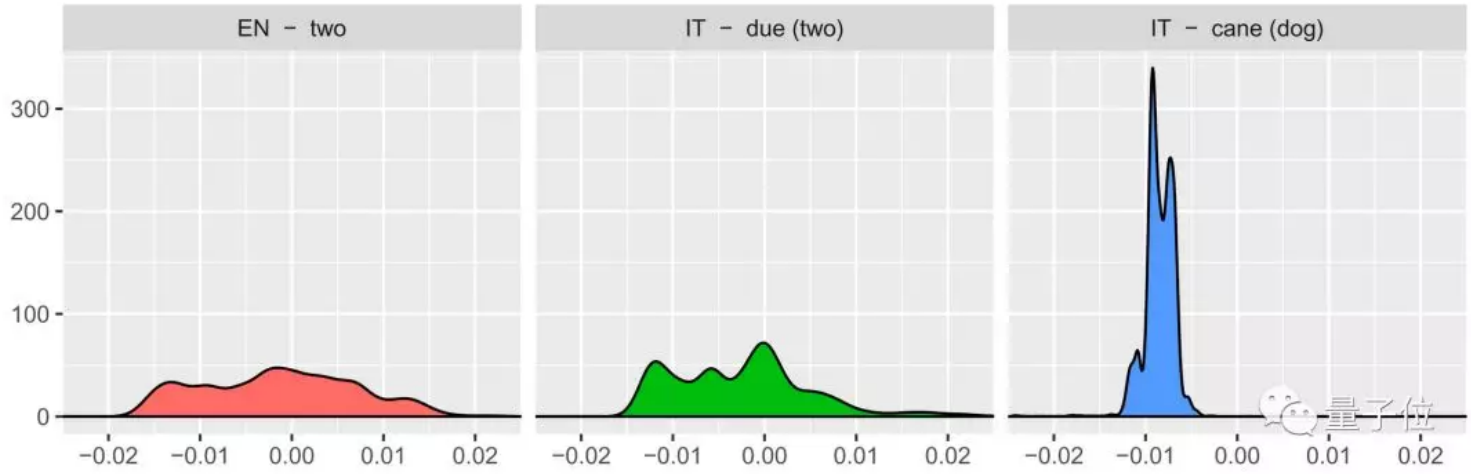
因此，在面对这种变化时，让模型更加稳健是很重要的。

代表性成果：

A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings (ACL 2018)

http://www.aclweb.org/anthology/P18-1073

> 这篇论文并不是将元学习应用到初始化上，而是利用他们对问题的理解来设计更好的初始化。
>
> 比较亮眼的是，他们将两种语言中分布相似的单词配对。这是一个很好的例子，证明了可以利用领域专业知识和分析见解使模型更加稳健。



△三个词的相似度分布。 等效翻译（“two”和“due”)的分布比非相关词(“two”和“cane”——意思是“dog”)的分布更为相似。

理解表征（Understanding representations）

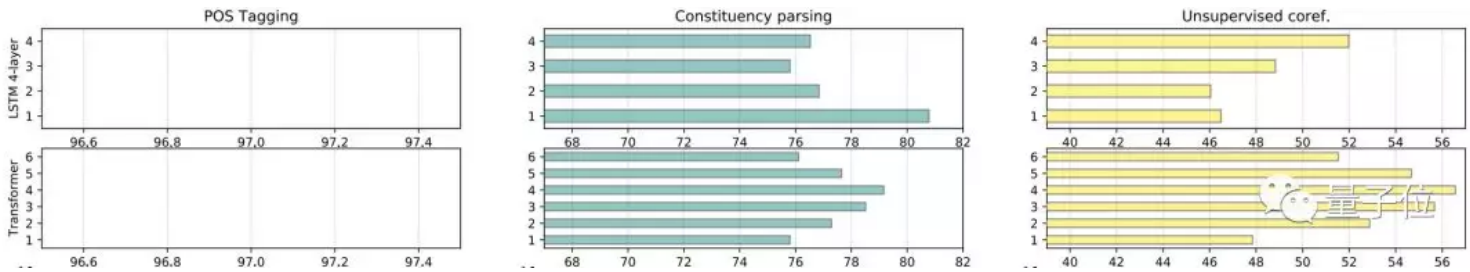
为了更好地理解表征，研究者已经做了很多努力。特别是“诊断分类器”（diagnostic classifiers）(旨在测量学习到的表征能否预测某些属性的任务)已经变得非常普遍了。

代表性成果：

Dissecting Contextual Word Embeddings: Architecture and Representation (EMNLP 2018)

http://aclweb.org/anthology/D18-1179

> 这篇论文在更好地理解预训练语言模型表征方面做了大量的工作。
>
> 作者在精心设计的无监督和有监督的任务中对单词和跨度表征进行了广泛的研究学习。
>
> 结果发现：预训练表征学习任务在较低层和较高层比较长的语义范围中，与低层次的形态与句法任务相关。
>
> 这实际上表明，预训练语言模型，确实捕捉到了与在ImageNet上预处理的计算机视觉模型相似的特性。



△BiLSTM 和 Transformer预训练表征在词性标注，选区分析，和无监督共指解析（(从左到右)方面每层的性能。

巧妙的辅助任务（Clever auxiliary tasks）

在许多场景中，我们已经看到越来越多的学者使用多任务学习和精心选择的辅助任务。

就一项好的辅助任务来说，数据必须易于访问。

一个最突出的例子是BERT，它使用下一句预测(在Skip-thoughts中使用过，最近在Quick-thoughts使用)取得了很大的效果。

代表性成果：

Syntactic Scaffolds for Semantic Structures (EMNLP 2018)

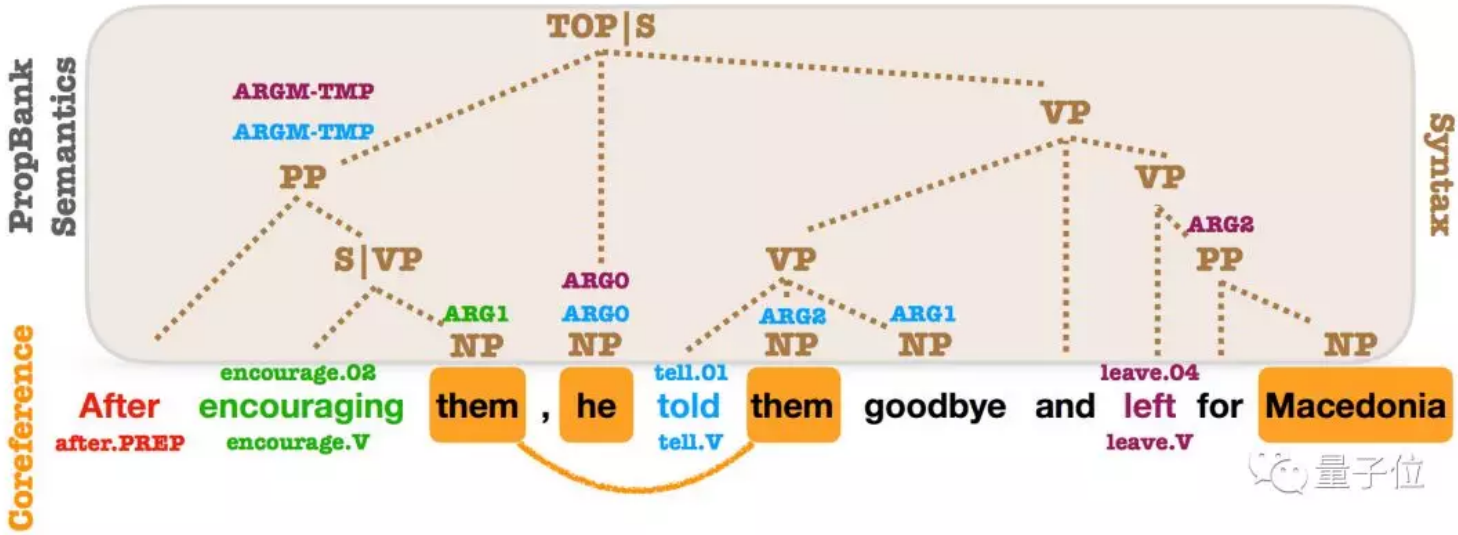
http://aclweb.org/anthology/D18-1412

> 这篇论文提出了一个辅助任务，通过预测每个跨度对应的句法成分类型，来预处理跨度表征。
>
> 尽管从概念上来说很简单，但是辅助任务在推动跨度预测任务出现大幅度改进方面很重要，例如语义角色标注和共指解析。
>
> 这篇论文证明了，在目标任务所要求的水平上学习专门的表征非常有用。

pair2vec: Compositional Word-Pair Embeddings for Cross-Sentence Inference (arXiv 2018)

https://arxiv.org/abs/1810.08854

> 基于相似的脉络，本文通过最大化“词对”与其语境之间的点互信息来预训练“词对”表征。这激励了模型去学习更多有意义的“词对”表征，而不是更通用的目标，比如语言建模。
>
> 对于需要跨句子推理的任务，如 SQuAD MultiNLI，预训练表征是有效的。
>
> 将来或许可以看到更多的预训练任务，能够捕捉特别适合于某些下游任务的属性，并与更多通用任务(如语言建模)相辅相成。



△OntoNotes的句法、PropBank和共指注释。PropBank SRL参数和共指提及标注在了句法成分之上。几乎每一个参数都与一个句法成分有关。

半监督学习与迁移学习相结合（Combining semi-supervised learning with transfer learning）

实际上，预训练表征与许多半监督学习表征的方法是互补的。

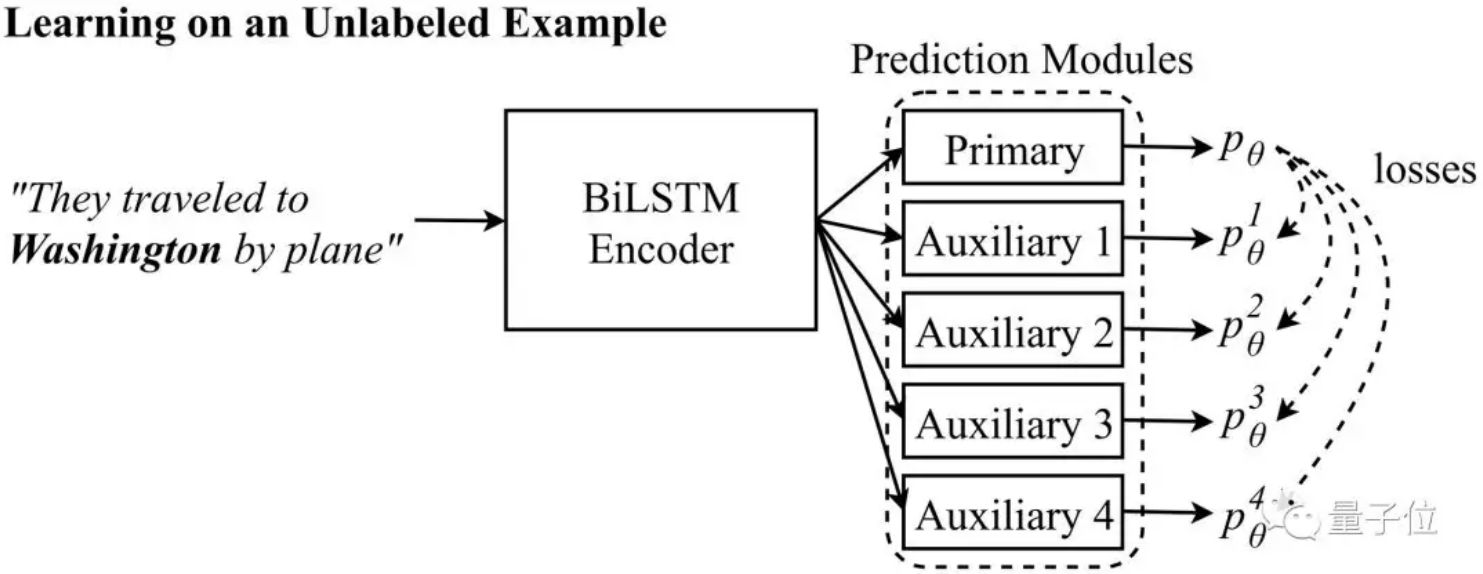
已经有学者探索了自我标注的方法，这是一种特殊类型的半监督学习。

代表性成果：

Semi-Supervised Sequence Modeling with Cross-View Training (EMNLP 2018)

http://aclweb.org/anthology/D18-1217

> 这篇论文展示了一个在概念上非常简单的想法，确保对不同输入观点的预测与主模型的预测一致，可以在不同的任务集合中获得收益。
>
> 这个想法类似于单词dropout，但是可以利用未标记的数据来使模型更加稳健。
>
> 与其他自组合模型相比，它是专门为特定的NLP任务设计的。
>
> 随着关于半监督学习的研究越来越多，有望看到更多的研究，来明确地尝试对未来的目标预测进行建模。



△辅助预测模块看到的输入：辅助1 :They traveled to __. 辅助2：They traveled to **Washington**__. 辅助3：_ **Washington** by plane.辅助4：__by plane。

QA和大型文档推理（QA and reasoning with large documents）

随着一系列新的问答数据集的出现，问答系统有了很大的发展。

除了对话式问答和多步推理，问答最具挑战性的方面是综合叙述和处理大体量信息。

代表性成果：

The NarrativeQA Reading Comprehension Challenge (TACL 2018)

http://aclweb.org/anthology/Q18-1023

> 这篇论文基于对整部电影剧本和书籍问题的回答，提出了一个具有挑战性的新QA数据集。
>
> 虽然依靠目前的方法仍无法完成这项任务，但模型可以选择使用摘要(而不是整本书)作为语境来选择答案(而不是生成答案)。
>
> 这些变体使完成任务更加可行，并使模型能够逐步扩展到完整的语境。
>
> 我们需要更多这样的数据集，它们会带来有挑战性的问题，但这些问题能够逐步解决。

Dataset	Documents	Questions	Answers
MCTest (Richardson et al., 2013)	660 short stories, grade school level	2640 human generated, based on the document	multiple choice
CNN/Daily Mail (Hermann et al., 2015)	93K+220K news articles	387K+997K Cloze-form, based on highlights	entities
Children's Book Test (CBT) (Hill et al., 2016)	687K of 20 sentence passages from 108 children's books	Cloze-form, from the 21st sentence	multiple choice
BookTest (Bajgar et al., 2016)	14.2M, similar to CBT	Cloze-form, similar to CBT	multiple choice
SQuAD (Rajpurkar et al., 2016)	23K paragraphs from 536 Wikipedia articles	108K human generated, based on the paragraphs	spans
NewsQA (Trischler et al., 2016)	13K news articles from the CNN dataset	120K human generated, based on headline, highlights	spans
MS MARCO (Nguyen et al., 2016)	1M passages from 200K+ documents retrieved using the queries	100K search queries	human generated, based on the passages
SearchQA (Dunn et al., 2017)	6.9m passages retrieved from a search engine using the queries	140k human generated Jeopardy! questions	human generated Jeopardy! answers
NarrativeQA (this paper)	1,572 stories (books, movie scripts) & human generated summaries	46,765 human generated, based on summaries	human generated based on summaries

△QA数据集的比较。

归纳偏差（Inductive bias）

归纳偏差，如CNN中的卷积、正则化、dropout和其他机制，是神经网络模型的核心部分，它们起到调节器的作用，使模型更具样本效率。

然而，提出一个应用更加广泛的归纳偏差方法，并将其融入模型是一个挑战。

代表性成果：

Sequence classification with human attention (CoNLL 2018)

http://aclweb.org/anthology/K18-1030

> 这篇论文提出利用视觉跟踪语料库中的人类注意力来规范视觉神经网络中的注意力。

考虑到当前许多模型（如Transformers）也使用注意力，找到更有效地训练它的方法是一个重要的方向。

另外， 论文还证明了人类语言学习可以帮助改进计算模型。

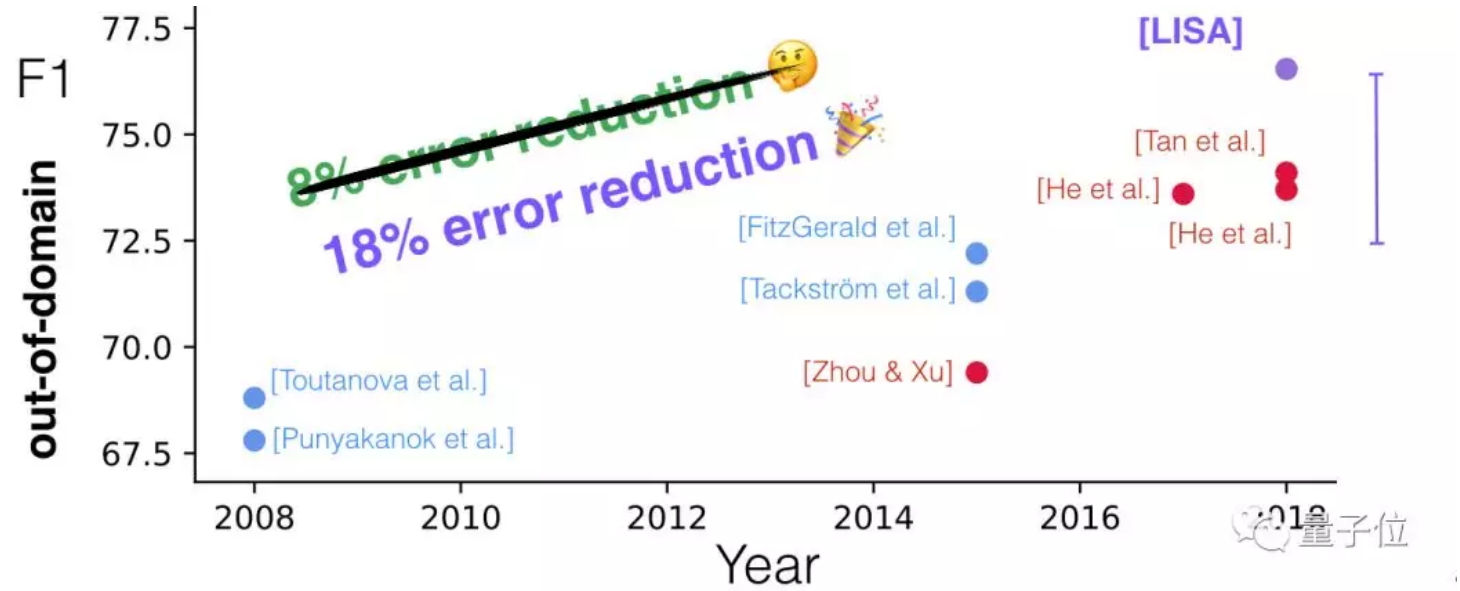
Linguistically-Informed Self-Attention for Semantic Role Labeling (EMNLP 2018)

http://aclweb.org/anthology/D18-1548

这篇论文有很多亮点：一个共同训练句法和语义任务的转换器；在测试时注入高质量解析的能力；和范围外评估。

论文中还通过训练一个注意力头来关注每个token的句法父项，使Transformer的多头注意力对句法更加敏感。

在未来，有望看到更多Transformer注意力头用于辅助预测集中在特定方面的输入。



△PropBank语义角色标注10年。语言学信息的自我注意力(LISA)与其他范围外数据方法的比较。

传送门

http://ruder.io/10-exciting-ideas-of-2018-in-nlp/

作者系网易新闻·网易号“各有态度”签约作者
转载于：量子位(QbitAI)

— 完 —

欢迎好看、收藏和转发

▲长按关注我们