安德烈·马丁斯 以下
Unbabel的AI研究副总裁和里斯本大学的特邀教授。 8月17日·16分钟阅读

ICML + ACL'18：游戏中的结构，翻译需要更多的上下文

几个星期前，我参加了机器国际会议 学习（ICML 2018）在斯德哥尔摩以及之后的计算语言学协会年会（ACL 2018）在世界的另一边：墨尔本。 有趣的是，这两个会议之间的时间接近度和地理距离的结合正在成为一种传统 - 去年它在澳大利亚的ICML和加拿大的ACL。

今年，我们提出了 ICML上的一篇论文 和 另一个在ACL，和 受邀在第二届神经机器翻译和生成研讨会上发表演讲。 这篇文章分享了一些 我对这两次会议的看法。

ICML'18：结构化，深度，生成性

喜欢 哎，ICML发展得非常快，有10个平行轨道（10年前我第一次参加ICML时曾经是4个）。 毫不奇怪，所提出论文的很大一部分与神经网络有关

- 他们的建筑，他们的训练，他们学到的表现。 一些

注意事项如下，重点是结构化预测，深度生成模型和表征学习。

从Södermalm可以看到斯德哥尔摩市中心，靠近ICML场地。

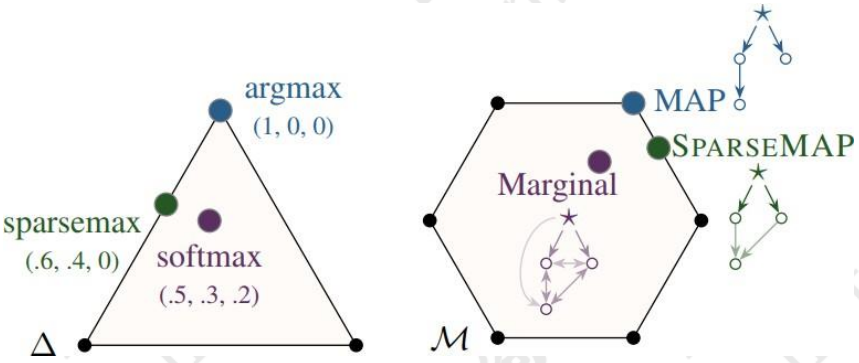
结构化预测

在结构化预测轨道中， 弗拉德尼库莱介绍我们的 SparseMAP论文（与Mathieu Blondel和Claire Cardie共同合作） - 一种结构推理的新技术，可输出 稀疏的结构集，与单个结构相反（如在MAP推理中）或在所有结构上的密集分布（如在边际推断中）。

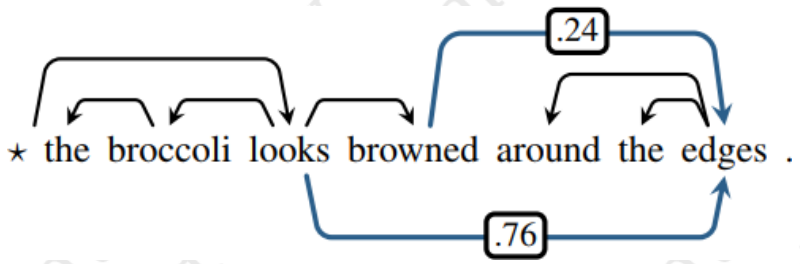
例如：模型可能只返回一个句子的少数合理的解析树，为所有其他的，难以置信的，分配概率零。 SparseMAP是可区分的（我们可以将其作为神经网络中的隐藏层插入并运行通常的梯度反向传播）和
高效（感谢有效的集合算法，通过解决一系列MAP问题来评估SparseMAP）。 我们可以将其视为结构变体 sparsemax。 Pytorch 码 与纸张一起提供。

在弗拉德之前，Nataly Brukhim就“结构化预测中的基数建模”发表了精彩的演讲，其中插入了基数的基数限制- 标签分类，通过Dykstra的投影算法处理。
有趣的是，我怀疑这个模型可以被视为我们的SparseMAP框架的一个特例 - 约束对应于“预算”因子，可以在线性时间内解决并且具有封闭形式的梯度（可能

比他们提出的展开Dykstra的方法更有效率。 另一个 相关论文 在第二天被提出 亚瑟·曼施：在动态规划算法的更新方程中使用sparsemax的方法，在sum-product和max-product之间得到可微分的变量。



SparseMAP作为sparsemax的结构化版本（图中提取的 这里）。 左：在非结构化的情况下（例如，多级分类）， sparsemax返回类上的稀疏分布，表示为单形边界中的一个点。 右：在结构化案例中， SparseMAP返回稀疏的结构组合（例如依赖解析树），表示为边际多面体的边界点。



SparseMAP为一个含糊不清的句子返回的两个解析树的示例。 所有其他树的概率为零（从这里提取的数字）。

深度生成模型

深度生成模型，最突出的变分自动编码器（VAE）和生成对抗网络（GAN），都在大肆宣传 这些天。 Max Welling发表了一篇精彩的邀请演讲（“每个Kilowatthour的情报”）通过镜头激发深层生成模型 信息论，能量最小化和最小描述长度原理，全部植根于方程F（自由能）= E（能量）-H（熵）。

几篇会议论文提出了VAE和GAN的新见解。 VAE通常通过最大化证据下界（ELBO）来训练，ELBO是可能性的易处理的变分近似。 它有

然而，有人观察到，这种程序通常会给出较差的潜在表征。 作为备选， 修复损坏的ELBO 提出输入和潜在变量之间的互信息的变分界限，这些变量比ELBO更通用。 结果是一个全速率 - 失真曲线，它折衷压缩和重建精度，从中我们可以寻求帕累托最优解。

另外两篇论文提出了减少VAE摊销差距的技术（这个差距量化了推理网络发现的变分参数的次优性，即VAE的编码器部分）： 迭代摊销推理 提出了一种类似于学习学习的迭代策略，它通过额外的循环增加推理网络，用随机梯度更新变分参数，同时 半摊销变分自动编码器 提出了随机和摊销变分推理之间的混合方法，利用可微优化。

至于GAN，中心思想是，而不是近似最大似然训练（相当于KL散度最小化），用一个不同的目标来代替它：训练发生器创建愚弄监督鉴别器的样本。 这归结为Jensen-Shannon散度最小化问题（原始GAN论文）或地球移动者的距离（如在Wasserstein GAN中）。

GAN优于其他生成模型的一个优点是它们倾向于产生尖锐的输出（当需要多模态分布时有用）；然而，它们也遭受模式崩溃（意味着它们倾向于仅从几种模式生成）。 他们的训练动力，归结为在极小极大问题中找到纳什均衡，尚未完全理解。 拉斯 Mescheder发表了精彩的演讲 对此有所了解，试图回答GAN实际融合的哪种训练方法的问题，提供一个通过分析GAN物镜梯度场的雅可比光谱，分析GAN何时收敛。

另一个挑战（与我们NLP研究人员非常相关）是让GAN生成像文本一样的离散数据（大多数关于GAN的研究都集中在像图像这样的连续输出上）。 这更具挑战性，因为基于交替梯度更新的GAN训练设置要求发电机的输出是可微分的，因此是连续的。 以前提出的解决方案包括政策梯度方法，如REINFORCE或Gumbel-softmax重新参数化技巧。 对抗正规化自动编码器 通过将离散自动编码器与GAN正则化相结合，提出了另一种解决方案

连续潜在的表示，在文本样式转移任务中有趣的结果。 然而，感觉在这个领域还有很多事要做。

还有 这个关于深度生成模型的很酷的研讨会 不幸的是，当我飞往ACL时，我无法参加。

顺序到序列

序列到序列学习一直由自回归模型控制（例如，在递归解码器中，每个发射的输出符号在下一个时间步骤中作为输入反馈到LSTM，从而产生对先前输出符号的依赖性）。 非自回归序列到序列模型是一个活跃的研究课题 - 如果它们起作用，解码器可以并行化并且速度更快。 现有的工作完成了这一点，但通常是 精度下降 或需要 迭代细化.

谷歌AI论文，使用离散潜在变量的序列模型中的快速解码，提出了一种不太自回归的方法，其中 自动回归生成少量潜在变量（在Transformer网络之上），然后以非自回归方式并行解码目标词，以潜在变量和源词为条件（似乎更近期）将这种方法与蒸馏相结合的工作，改进了结果）。 显然潜在变量是离散的是至关重要的，尽管我不完全理解为什么。

另一篇机器翻译论文，分析神经机器翻译中的不确定性，来自FAIR，提高了对顺序的理解ce-to-sequence模型通过评估模型分布如何捕获由噪声训练数据引起的不确定性以及它如何影响搜索，提出评估模型校准的工具。

表征学习与时间测试

另一篇有趣的论文是双曲线嵌入的表示权衡（以及另外两篇相关论文： 这个 和 这个）， 从中 我开始了解这个最近嵌入双曲空间的框架。 这个（数学上美观的）研究领域的共同主题是双曲线流形上的嵌入（与欧几里德空间相对）适合于表示对象之间的层次关系（例如，树可以嵌入到庞加莱盘中，2-尺寸双曲空间，任意低失真）。 许多感兴趣的对象，例如社交网络，WordNet-

像知识库，蕴涵关系等具有层次结构，这种几何形状可能是通常的平欧几里德空间的吸引人的替代。 在本文中，他们设法用非常紧凑的二维表示来表示具有最先进精度的WordNet分类法。

罗南科洛伯特 和 杰森韦斯顿 获得了时间测试奖 他们有影响力的ICML 2008论文“自然语言处理的统一架构：具有多任务学习的深度神经网络。”我非常清楚地记得这篇论文（ICML 2008是第一篇 我参加的机器学习会议，以及从大多数NLP观众那里得到的“冷”接收，在这个时候，神经网络一切都很受欢迎。 我承认我也持怀疑态度 - 如果没有任何特色工程，在整个维基百科上训练数周的单一模型如何在几个NLP任务中获得最先进的分数？

原始论文中存在一些问题，在语义角色标注任务中进行了评估（在后面的JMLR论文中修复），并且从这个浅层任务中取得进展到接近解决“所有语义”的大量推断（这个可能是最让NLP社区恼火的事情。

虽然最后一个问题在社区中仍然存在，但是时间已经证明这是一个非常有价值的贡献，而且这个奖项非常值得。 十年后，我们都在使用大型数据集培训的连续字表示，并“从头开始”培训端到端模型。除了某些行业利基，特征工程早已不复存在，逐渐被代表性学习和其他形式所取代工程：架构搜索，超参数调整，转移学习。 也许另一个带回家的消息（我可能在这里失去一些朋友）是NLP社区应该对其他领域的贡献更加开放，即使他们似乎对语言一无所知 - 一个封闭的社区被指责过度拟合他们自己的技术。

ACL’18：生育，语境意识翻译，语言结构偏见

经过几次长途飞行和日历上的一天半，我到达墨尔本进行ACL。 会议在会议期间举行 墨尔本会展中心，一个非常好的场地（也是我去过的最大的会议中心之一，可与悉尼国际会议中心相媲美，该会议中心举办了ICML 2017）。

该组织非常壮观。 我希望更多这些会议来到澳大利亚，以便我有借口回去。

机器翻译与生成

Chaitanya Malaviya根据他的实习情况介绍了我们的论文稀疏和受限制的神经机器翻译注意事项 工作在 Unbabel （与Pedro Ferreira合作）。 这个想法是通过约束注意力取代传统的基于softmax的注意机制来避免神经MT（翻译不足和过度翻译）中的一些常见错误模式：每次我们生成目标词时，我们首先上限多少注意每个源词都可以接收（因此避免重复注意相同的词）。

为此，我们记录每个源词的累积注意力，并使用基于生育率的方法（类似于IBM模型）来定义上面的上限。 此外，我们鼓励注意力是稀疏的，即对不相关的词给出零概率 - 这是通过新的约束稀疏变换实现的，该变换计算和可微分便宜。 我们也提议 新的评估指标 检测重复和丢弃源词。

Chaitanya Malaviya介绍我们的ACL论文。

In 重量共享的无监督神经机器翻译，提出了一种新的体系结构，它代替两种语言的单个编码器，使用单独的编码器和共享潜在空间，以及两个GAN（本地和全局）以增强跨语言转换。

然而，无监督的NMT似乎有一些局限性，如中所述 论无监督双语词典归纳的局限性，提出 塞巴斯蒂安·鲁德：特别是，对于形态丰富的语言（或简单地，当源语言和目标语言非常不同时）以及在不同域上或使用不同算法训练嵌入时，性能会显着下降。

这并不让我感到惊讶：我总是发现令人费解的是，如何通过应用正交（或更一般地，线性）变换从单独的单语数据中引入双语嵌入。 我理解嵌入空间的“局部”结构对于两种语言都是类似的，但为什么“全局”结构应该相似？ 正交矩阵具有O（D 2）自由度（D是嵌入大小）但是对于实际问题，词汇量大小可能会超过D 2，那么我们如何在这些旋转中有足够的自由度来叠加相关词的嵌入？
也许这是有效的，因为单语语料库毕竟有点平行（维基百科似乎就是这种情况），至少在单语嵌入具有类似结构的程度上。 但是如果它们的域不同或嵌入是用不同的算法学习的话，这将会破坏。

Google AI的论文， 两全其美的，进行广泛的实验，以解开变压器网络改进的几个来源。 他们表明，类固醇的“经典”复发NMT模型，RNMT +（采用变形金刚论文中引入的一些侧面创新，如多头注意）与非复发性竞争 注意力是，所有你需要 变压器的架构。 培训可能很棘手，但Orhan Firat在质量保证中提到，像层标准化这样的事情似乎对稳定变形金刚和RNMT +的培训至关重要。

一种用于神经机器翻译的随机解码器 解决了翻译是一个多模态，模糊过程的事实 - 对于给定的句子通常有许多有效的翻译。 他们提出了一种类似VAE的机器翻译深度生成模型，它允许对许多可能的翻译中的一种进行抽样，通过结合一系列潜在变量来解释并行语料库中的词汇和句法变异。

一个真正令人兴奋的研究方向是在上下文感知的NMT中 - 如何使NMT模型超出一个句子的范围？
这是Rico Sennrich在第二届神经机器翻译和生成研讨会上的主题（更多关于Rico的演讲） 下面）。 这个主题有几个有趣的讨论，包括 Elena Voita的讲话语境感知神经机器翻译学习了Anaphora Resolution，它允许Transformer的编码器 相邻句子的条件（对其学习参加的地方进行了非常有趣的实证分析），以及Sameen Maruf的谈话 文档上下文神经机器翻译与内存网络，使用内存网络存储多个发送在文档中并运行基于块坐标下降的迭代解码算法。

一些有关文本生成的有趣论文包括来自CMU的反向翻译风格，它提出了一种对抗性潜变量模型来重新定义文本以包含特定的内容。 文体属性（情感，性别和政治倾向），分层神经故事生成，来自FAIR（荣获一项荣誉） 提及），它学习了一个创意系统，以便在短时间内提供连贯和流畅的故事（包括从Reddit收集的新数据集）， 通过编辑原型生成句子来自斯坦福大学，通过从训练集中抽样原型并对其进行细化来生成句子（“半参数”混合，混合了传统的从左到右的句子生成器和基于记忆的方法）和语言模型的数学：评估和

从UCL看，提高他们预测数字的能力s语言模型准确预测数字的能力 - 他们建议修复的语言模型的当前弱点。 很有意思！

解析和Argmax分化

直接到树上 来自MILA提出了一个新的简约成分解析器，它仅为输入句子中的每个分割位置预测实值标量的矢量（“句法距离”），指定将选择分割点的排序顺序 - 然后是二进制可以从上到下的方式递归地从这个排名中诱导树。

用于依赖性解析的堆栈指针网络 来自CMU是传统的从左到右转换的解析器的替代方法，使用堆栈指针进行自上而下的解析 - 在每个点，指针选择下一个左或右子进行选择作为修饰符（如在头部自动机）并递归地进行。 将这种方法与基于转换的解析器（从左到右解析）和易于解析的解析器（从底部向上解析）进行对比是很有趣的。 最后，海报选区解析自我注意编码器 来自伯克利的一个额外点是通过用一个自我专注的架构（类似于变压器网络）取代BILSTM编码器，在PTB中用单一模型和没有外部数据实现了一些令人印象深刻的93.55 F1，以及95.13带有完整模型的F1 ELMO嵌入。

来自威斯康星大学的郝鹏使用SPIGOT展示了通过结构化Argmax进行反向传播，这是一篇很好的论文来解决这个问题。w通过不可区分的argmax操作进行反向传播，该操作获得了荣誉奖。 它提出了一种称为SPIGOT的直通估计器的变体 - 它涉及计算前向传递中的argmax，然后计算向后的更新，其涉及欧几里德投影到边缘多面体上。

原来这个投影正是如此 SparseMAP计算也是如此（参见上面的ICML段落），这使得这两种方法非常相关：当SparseMAP通过在前向步骤上计算这个稀疏投影然后用精确梯度反向传播来避免argmax时，SPIGOT在前向步骤中计算argmax然后反向传播代理梯度。 但是，构建块是相同的，因此SPIGOT也可以从SparseMAP的有效集算法中受益，以计算欧几里德投影（使其可用于argmax高效的任何问题）。 总的来说，我更喜欢SparseMAP

但是，由于它的训练可能更稳定，因为它基于精确的梯度（加上，它具有稀疏的结构化预测解释）。

另一个想法是：看看“边缘-SPIGOT”将如何表现（即如果我们用“KL投影”取代欧几里德投影，这通常归结为总和 - 产品动态将是有趣的 编程）：“边缘-SPIGOT”将是结构化注意网络，因为SPIGOT是SparseMAP。

奖项和研讨会

最后一次会议专门讨论最佳论文奖和ACL终身成就奖 马克斯蒂德曼 因为他对语法和语义领域的杰出贡献，最引人注目的是语法和语义的发展 组合分类语法.

早些时候，在知道你不知道的事情：SQuAD的无法回答的问题，最佳短篇论文奖的获得者，新版本的 引入了SQUAD数据集（阅读理解的参考基准），将原始问题与人群工作者反对写成的50,000个无法回答的问题结合起来，看起来类似于可回答的问题。 该数据集似乎比以前的版本更具挑战性，目前人机与机器性能之间存在20个F1点的差距。

维多利亚女王市场在墨尔本

在研讨会期间，我参加了Chris Dyer关于研讨会的演讲 神经NLP中语言结构与“组合”的相关性

和自然语言学习中的作文。“这是一个美化ul的启发受到了启发 查尔斯·霍克特“模式的二元性”，它在语言中对比了两个层次的结构，即如何通过组合独特和无意义的字符/音素来形成单词，以及如何通过组合重要和有意义的单词来形成句子。 我将专注于构图过程。

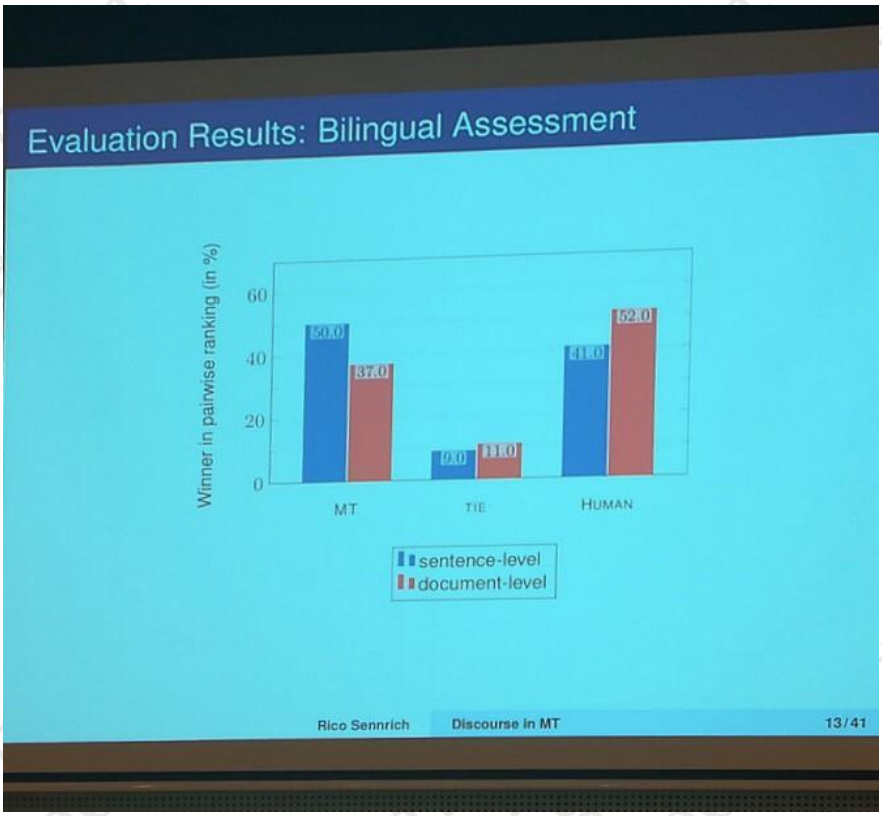
我们为什么要打扰语言结构建模呢？ RNN实际上在哪里失败？ Chris的演讲显示了结合归纳偏差的两个好处：它使我们的模型更准确（特别是，如果我们没有大量数据），它有助于我们理解非结构化模型的偏差和局限性。

发现这些限制的一种方法是寻找RNN可以编码的语言信息（例如，通过尝试将其表示暴露于人类可读的东西，如FST- Yoav戈德堡稍后在同一个研讨会上的讲话）。 Chris的观点略有不同：要了解RNN能做什么和不能做什么，将他们学习的偏见与其他结构化模型的偏见进行比较，就其“预测偏好”而言。一个非常引人注目的例子是主题 - 动词协议任务，需要预测动词的是句子 比如“他告诉我柜子里柜子的钥匙在桌子上。”随着“吸引者”数量的增加（即名词前的名词） 动词不是正确的主题），模型越来越难以找出选择正确动词形式的正确主题。

这个实验揭示了RNN的连续新近偏差，它需要太多的跳跃（与表面词一样多）才能从动词到达主体，最终学习过于简单化的名词启发式 - 选择他作为主语。 相比之下，语法通知模型如 RNN语法，具有堆栈数据结构，产生语法新近度的偏差，这减少了使得主题正确的跳数。

最后，我参加了第二届神经机器研讨会 翻译和生成，在那里我发表了邀请演讲：超越Softmax：稀疏性，约束，潜在结构 - 所有端到端的差异化！ 在这里，我描述了softmax t的各种替代方案hat可用于注意机制和输出层，包括 sparsemax，约束softmax / sparsemax，和最近的 SparseMAP（见上文），以及由此引起的这些转变的统一视角 广义熵.

Rico Sennrich发表了一篇有趣的演讲，主张为什么时间成熟，机器翻译中的话语，具有良好的历史视角。 提醒人们注意到最近的微软人类平等成就 - 正在进行的工作，一个新的文件级别 正在进行的评估表明，如果考虑到背景（而不是逐句地对翻译进行排序），那么人类平价仍然很遥远。 在Rico之前，Jacob Devlin描述了几种工程技巧和架构决策，以适应手机中的NMT模型。 与此相关，有一个 关于NMT效率的共同讨论， 在哪里 玛丽安NMT系统 在所有环境中获胜（关于玛丽安的更多信息） 这张纸 在ACL演示会议中提出）。 研讨会以谈话结束 来自Yulia Tsetkov关于如何获得灵活但可控的语言生成。



对MT的文档级评估：还没有人类平价（来自Rico Sennrich的演讲）。

带回家的消息：

- Sparsemax / argmax差异化在结构化预测和NLP方面开辟了新的研究前沿。
- 深度生成模型非常有前途，但仍然在很多方面被破坏，并且还无法处理离散数据

令人信服。

- 表征学习赢得了ML和NLP的时间考验。
- 语言结构（生育，句法短语）正在作为端到端模型中的归纳偏差。
- 机器翻译没有得到解决，下一个重要的事情是如何在超出句子边界的情况下考虑上下文。

这就是所有人！

很快在布鲁塞尔见到EMNLP。

注意：我要感谢大家 Unbabel AI部落 和 DeepSPIN 团队成员在此博客文章中提供反馈，并在阅读会议中阅读其中一些论文的预印本。 其他笔记已经 在这些会议上发表：查看Sebastian Ruder的ACL 2018亮点和 David Abel的ICML 2018详细说明，与foc我们关于强化学习。