**Semi-supervised Domain Adaptation for Dependency Parsing**

会议：ACL2019

阿里团队

代码：<https://github.com/SUDA-LA/ACL2019-dp-cross-domain>

**摘要**

在过去的几十年中，由于缺乏足够的标记数据，因此在没有目标域训练数据的情况下，大多数有关跨域解析的研究都将重点放在无监督域自适应上。 然而，由于域自适应和解析的内在困难，到目前为止，无监督方法进展有限。 本文基于两个新注释的大规模领域特定数据集，解决了中文监督依赖分析的半监督领域适应问题。我们提出了一种简单的领域嵌入方法，用于合并源域和目标域训练数据，如图所示 比直接语料库串联和多任务学习更有效。为了利用未标记的目标域数据，我们使用了最新的上下文化词表示法，并显示了一种简单的微调过程可以进一步大幅提高跨域解析的准确性。

**1引言**

依赖项解析作为NLP的一项基本任务，由于其在捕获语法和语义信息方面的多语言适用性，在过去的几十年中吸引了许多研究兴趣（K？ubler等，2009； McDonald等，2013）。 给定一个输入语句S = w0w1……wn，依存关系分析将构造一棵树，如图1所示，其中是从头部wh到修饰符wm的依赖性带有关系标签为l，而w0为伪根节点。

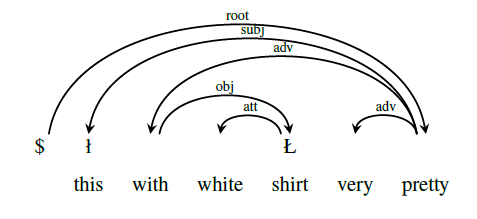


图1：来自产品博客域的示例。英文翻译是“这看起来很漂亮，穿着白衬衫。”

最近，由于深度神经网络在捕获远程上下文方面的强大功能，依赖项解析取得了巨大进展（Chen and Manning，2014; Dyer et al。，2015; Zhou et al。，2015; Andor et al。，2016 ; Kiperwasser和Goldberg，2016; Dozat和Manning，2017; Ma等，2018）。此外，在语言模型训练损失下，从大规模未标记文本中学习的上下文化单词表示（Peters等人，2018; Radford等人，2018; Devlin等人，2018）被证明对许多NLP任务有广泛的帮助 包括依赖项解析（Che等人，2018; Clark等人，2018; Kitaev和Klein，2018）。

但是，当处理与训练数据不同的文本时，解析性能会急剧下降，这被称为域适应问题。 实际上，随着Web数据（或用户生成的内容）激增，跨域解析已成为在现实的NLP系统中应用语法分析的主要挑战。 为了迎接这一挑战，社区组织了一些共同的任务，以吸引更多的研究关注（Nivre等，2007； Hajiˇc等，2009； Petrov和McDonald，2012）。

由于缺乏足够的标记数据，大多数以前涉及跨域解析的工作（包括上述共享任务）都假定没有标记的目标域训练数据，因此着重于无监督域适应。 到目前为止，由于领域自适应和解析的固有困难（参见第5节中的讨论），朝着这个方向的方法进展有限。 另一方面，由于极端的复杂性和沉重的成本，在新域文本上进行句法数据注释的进展非常缓慢，并且仅在网络文本上构建了几个小规模的数据集，主要是作为跨域的评估数据 域解析（Foster等，2011; Petrov和McDonald，2012; Kong等，2014; Wang等，2014）。

为了应对上述挑战，本文提出了两个新注释的大规模领域感知数据集（超过12K句子），并尝试解决针对中文依赖项解析的半监督领域自适应任务。 通过访问已标记和未标记的目标域数据，我们提出并评估了几种简单的方法，并进行了错误分析，以调查以下三个问题：

问题1：如何有效地组合源域和目标域标记的训练数据？

问题2：如何利用目标域未标记的数据进行进一步的改进？

问题3：给定一定数量的标记数据，要在新域上达到特定性能需要注释多少数据？

正如我们的审阅者所指出的，由于需要标记的目标域训练数据，因此在此工作中解决的半监督域适应方案比无监督的适应方案不现实，这通常是非常昂贵的。 但是，我们认为，当仅存在数十个或数百个带标签的目标域训练语句时，这项工作可能同样有价值和有用，考虑到如上所述，这对于解析技术的实际应用可能是一个可行的折衷方案，纯粹是无监督的领域适应取得的进展非常有限。

**2数据注释**

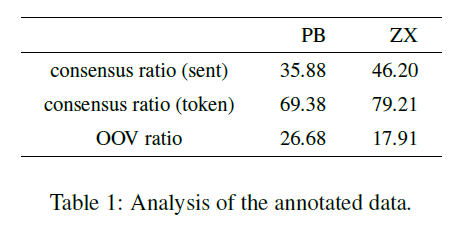
在这项工作中，我们选择两种典型的领域感知Web文本进行注释，即产品博客和网络小说。 本节介绍有关数据注释过程的详细信息。

数据选择。 产品博客（PB）文本是从淘宝标题网站上检索的，该网站包含用户撰写的文章，这些文章主要是关于不同商业产品的描述和比较。 经过数据清理和自动分词后，我们收集了大约340K句子。 然后，我们按照Jiang等人（2018）的主动学习工作流程，选择[5，25]个单词的一万个句子进行手动注释。 其余的句子用作未标记的数据。 对于网络小说，我们遵循Zhang等人的跨域分词研究（2014年），并采用了流行小说“竹县”（ZX，又称“玉皇朝”）。在带注释的4,555个句子中，我们选择了约3,400个[5，45]个单词的句子用于注释。 ZX的其余32K句子在本工作中用作未标记的数据。

注释准则。 在比较了一些公开的有关依赖关系解析的指南（包括通用依赖关系（UD））之后（McDonald等，2013），我们采用了Jiang等人发布的指南（2018）基于三个考虑。 首先，它们的指南包含20种关系，这些关系专门用于捕获不同来源文本的中文依赖语法。 其次，这份长达70页的指南提供了非常详细的说明，其中包含许多具体示例。第三，他们构建了一个大规模的平衡语料库（BC），用作本文中标记为源域的数据。

质量控制。 我们雇用大约15名本科生作为注释者，并选择5名具有语言背景的经验丰富的注释者作为专家注释者。 每个注释者都经过严格的培训，以熟悉该准则。在基于浏览器的注释平台的基础上，我们应用严格的双重注释以确保标记数据的质量。 首先，将每个具有自动分词功能的原始句子随机分配给两个注释器。 如果两个提交相同，则注释被接受。否则，由第三位专家注释者在比较和分析两个提交内容后决定答案。

统计和分析。 移除词段错误或语义难以理解的句子后，我们获得9,040 PB句子和3,249 ZX句子。 我们从三个方面分析了两个数据集，如表1所示。逐句共识率是从两个注释者那里收到完全相同的提交的句子的百分比，对于PB来说只有35％，对于ZX来说只有46％。 这意味着所有句子的一半以上都需要由专家注释者检查，这表明语法注释的复杂性以及为保证质量而需要双重注释的必要性。令牌方式的共识率是从两个注释器接收相同的标头和标签的令牌的百分比，仍然低于PB的70％和ZX的80％。 这些共识比率清楚地表明，PB比ZX更加难以注释。 作为用户生成的内容，PB更为随意，并且包含许多单词省略号现象，写错字符，缩写单词，不正确的语法表达等等。



OOV比率表示在Jiang等人（2018）的源域BC数据中未出现的令牌百分比。 我们可以看到，PB中的OOV比远高于ZX，这肯定会使PB更加难以解析。

**3方法**

本节介绍了几种半监督的跨域解析方法。

**3.1基本Biaffine解析器**

在这项工作中，我们构建了最先进的深度Biaffine解析器的所有方法（Dozat和Manning，2017年）。 作为基于图的依赖项解析器，它使用深的biaffine神经网络来计算所有依赖项的得分，并使用维特比解码来找到得分最高的树。 图2显示了如何计算的分数。

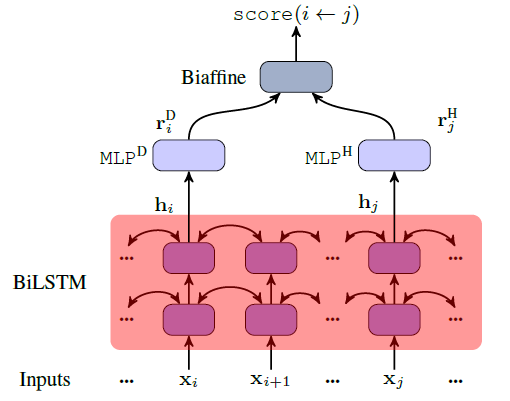
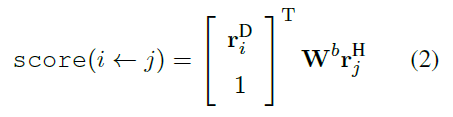


图2：在biaffine解析器中的计算。 为简单起见，我们仅绘制两层BiLSTM。

首先，baffine解析器应用多层双向顺序LSTM（BiLSTM）对输入句子进行编码。 第i个单词的输入是单词/标签嵌入的串联，即。第i个字的顶层BiLSTM的输出向量表示为hi。 它被馈送到两个单独的MLP中，以获得单词的两个低维表示向量，分别作为头部和依赖。

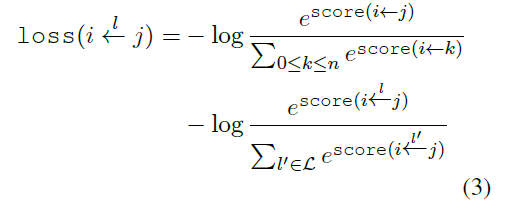


最后，通过仿射运算来计算圆弧的分数。



同样，解析器使用额外的MLP和biaffines计算标签分数。由于篇幅所限，我们请读者阅读Dozat和Manning（2017）以获得更多详细信息。

训练损失。 对于每个wi及其金标准头wj和标签l，解析器采用局部交叉熵损失。



其中L是标签集。 计算单独的损失以进行头选择和打标。

**3.2合并两个训练数据集**

在本小节中，我们描述了用于合并源域和目标域训练数据集的三种简单方法。

（1）直接串联（CONCAT）。 最直接的方法是将多个训练数据集直接合并为一个更大的数据集。 此方法将源域和目标域训练数据集同等对待。 基本解析器可以直接使用，而无需进行任何修改。 此方法的主要缺点是模型对两个域使用相同的参数，因此无法学习特定于域的功能。

（2）域嵌入（DOEMB）。 Stymne等（2018）提出了一种树库嵌入方法，通过利用一种语言的多个异构树库（遵循不同的注释准则）来改善解析。 受他们工作的启发，我们建议将每个单词的位置与一个额外的域嵌入连接起来，以指示该训练语句来自哪个域，如图3所示。通过这种方式，我们希望模型可以充分利用两个训练数据集，因为大多数参数都是共享的，除了两个域嵌入向量以外，并且还学会区分特定于域的特征和一般特征。

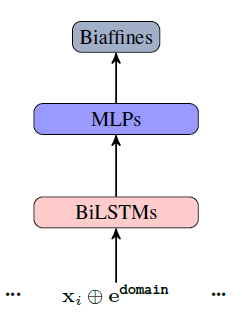


图3：DOEMB方法的框架，其中domain =“ src”（源域句子）和“ tgt”（目标域句子）。

（3）多任务学习（MTL）旨在合并多个相关任务的标记数据以提高性能（Collobert和Weston，2008年）。 郭等（2016）首先使用MTL通过利用多个异构树库并将每个树库作为单独的任务来提高解析性能。 如图4所示，我们对biaffine解析器进行了直接扩展，以实现多任务学习。 源域和目标域解析被视为带有单词/标签嵌入和BiLSTM共享参数的两个单独任务。 MTL的主要缺点是模型无法充分利用源域标记的数据，因为源域训练数据仅有助于共享参数的训练。

语料库加权策略。 对于上述所有三种方法，如果直接组合，则在训练过程中目标域标记的数据将被源域数据淹没，因为通常在规模上存在很大的差距。 因此，我们采用简单的语料库加权策略（Li等，2014）作为有用的技巧。 在每次迭代之前，我们从目标域和源域训练数据中以1：M的比例分别对训练语句进行随机抽样。然后，我们将合并的数据随机混合在一起进行一次迭代训练。 我们将M >1视为根据dev数据调整的超参数。

**3.3利用未标记的数据**

除了标记数据，如何利用目标域和源域的未标记数据一直是跨域解析的一个有趣且重要的方向，如第5节所述。 最近，彼得斯等（2018）引入了语言模型（ELMo）中的嵌入内容，以有效利用大量原始文本作为预训练步骤。 他们使用多个BiLSTM层作为句子编码器，并采用从左到右的顺序语言模型损失。

在这项工作中，我们提出了一种非常简单的两步方法来将ELMo应用于跨域方案。

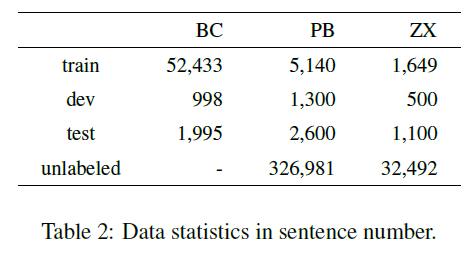
步骤1：在大规模的通用域未标记数据上训练ELMo。 我们在中文Gigaword第三版上对ELMo进行了培训，该版包含大约120万个句子。 使用6个GPU节点（GTX 1080Ti），大约需要7天。

步骤2：对目标域未标记数据上的ELMo进行微调。 然后，我们使用上一步中训练的参数作为起点，对目标域未标记数据上的ELMo进行微调。 为了节省计算资源，我们将所有三个域的所有train / dev / unlabeled数据合并为一个未标记的数据集，以对ELMo进行一次微调，并对所有三个域使用相同的经过微调的ELMo。

对于每个单词，ELMo的三个BiLSTM层的表示形式均被平均，并用于替换Biaffine Parser中的原始单词嵌入。 我们没有尝试让模型针对不同的图层自动学习不同的权重，这可能会导致更好的性能。 由于ELMo使用charLSTM来学习第一层单词表示，因此我们确实尝试将字符字典扩展为仅出现在目标域未标记数据中的那些字符字典，并在微调ELMo之前随机初始化其对应的char嵌入。 但是，这只会产生轻微且不一致的性能提升。

**4实验**

数据。 我们使用Jiang等人发布的平衡语料库（BC）。 （2018）作为源域，遵循其train / dev / test划分。 我们使用新批注的PB / ZX数据集作为两个目标域，并分别将其划分为train / dev / test，并考虑到dev / test数据集应尽可能大，以便进行更可靠的评估。 如第2节所述，我们还提供了目标域未标记的数据。表2显示了数据统计信息。

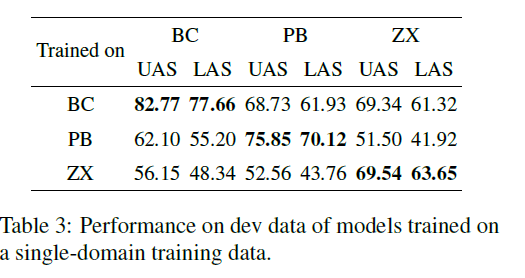


评估指标。 我们使用标准的标记附件得分（LAS，接收正确的头部和标签的单词百分比）和未标记的附件得分（UAS，忽略标签）。

解析器设置。 我们使用PyTorch实现基本的仿射解析器和建议的方法。 我们遵循Dozat和Manning（2017）的超参数设置，例如学习率和的dropout率。 每个解析器最多训练1 000次迭代，并且在每次迭代后针对模型选择对开发数据进行性能评估。 如果在50个连续的迭代中没有提高峰值性能，我们将停止训练。

**4.1单域训练结果**

表3给出了在单域训练数据上训练每个解析器时对dev数据的解析准确性。 我们可以看到，尽管PB训练比BC训练小得多，但是PB训练的解析器在PB-dev上比BC训练的解析器好大约8％，这表明目标域标记数据的有用性和重要性，特别是当两个域非常不同。



但是，在LAS中，经过ZX训练的解析器和经过BC训练的解析器之间的差距仅为2％，我们认为这有两个原因。首先，ZX训练的尺寸甚至更小，仅小于PB训练的三分之一。其次，卑诗省语料库来自《人民日报》，可能包含新颖的文章，与ZX更相似。 总体而言，当训练数据来自同一域时，解析器可以在给定域上实现最佳性能，这是显而易见且合理的。

**4.2合并两个训练数据集**

我们使用第3.2节中描述的三种方法来合并源域和目标域训练数据。 由于源域和目标域训练数据的大小之间存在较大差距，因此我们采用语料库加权策略来平衡不同源的影响。

图5显示了在具有不同权重因子M的dev数据上的结果。PB和ZX上的曲线清楚地表明，语料库加权具有广泛的帮助，好的权重因子和差的权重因子之间的性能差距对于某些目标领域和方法而言，可能会很大。具体而言，对于PB作为目标域，对于源域和目标域使用相同的权重（即M = 1）似乎就足够了，并且选择适当的较大M会导致不足1％的改善。相比之下，语料库加权在ZX域中更为重要，并且在使用更大的M时会带来更好的性能。除了ZX列的尺寸非常小之外，另一个原因可能是由于ZX与BC之间的相似度很高， 如前所述。

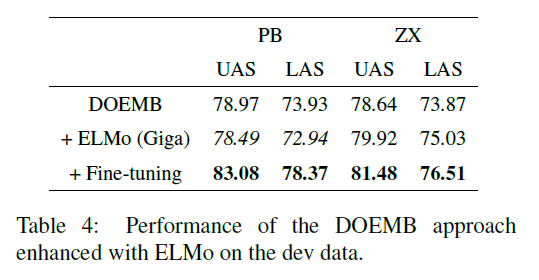
从另一方面来看，我们可以看到DOEMB方法始终在两个目标域中的三种方法中始终表现最佳，而MTL在利用源域训练数据方面效率最高。

总体而言，结果与我们在3.2节中的讨论一致。 DOEMB在CONCAT和MTL上取得成功的关键在于，通过共享更多参数来合并两个域中的知识，以及区分两个域以学习特定于域和通用的特征之间的平衡。

对于每个方法域对，我们根据它们在dev数据上的结果选择最佳语料权重M。

**4.3未标记数据的使用**

在这一部分中，我们将使用第3.3节中介绍的方法来增强ELMo最有效的DOEMB方法。 表4报告了结果。出乎意料的是，使用在通用域中文Gigaword语料库上训练的ELMo对两个目标域产生相反的影响。 LAS在PB上减少0.99，但在ZX上增加1.16。 我们怀疑原因可能是像BC之类的中国千兆字语料库包含许多与ZX相似的与小说有关的文本。 相反，考虑到PB文本通常是用户最近生成的内容，因此不太可能具有类似于PB的文本。 这一发现与域内解析结果不同，后者在ELMo中总是有帮助的（Che等人，2018; Clark等人，2018）



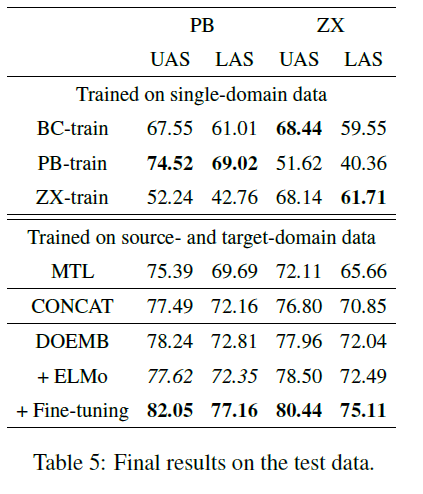
在目标域未标记的数据上进一步微调ELMo会导致两个域的一致性和大幅改进。 与“ ELMo（Giga）”相比，PB的LAS增加了5.43，ZX的LAS增加了1.48。 我们认为PB与ZX相比，更大的改善主要是由于未标记的PB数据规模更大。 结果表明，通过对目标域未标记数据进行微调，ELMo有效地学习了特定于领域的知识，并且能够产生更可靠的上下文化词表示形式。

4.4测试数据的最终结果

表5显示了测试数据的最终结果，与先前的观察结果一致。首先，当限制在单域训练数据上时，使用目标域数据是最有效的。 其次，将源域数据用作额外的训练数据会有所帮助，并且DOEMB方法的效果最佳。 第三，首先对非常大型的通用未标记数据进行ELMo训练，然后对相对较小规模的目标域未标记数据进行微调，这是极其有用和高效的。

4.5分析

如表5所示，PB的最终性能始终比ZX的性能高出约2％，我们认为主要原因之一是PB-train比ZX-train大三倍以上。 这就提出了一个有趣而重要的问题。 面对新域时，在给定一定数量的源域数据的情况下，为了达到特定性能，我们需要注释多少数据？我们尝试通过以下分析给出一些线索。

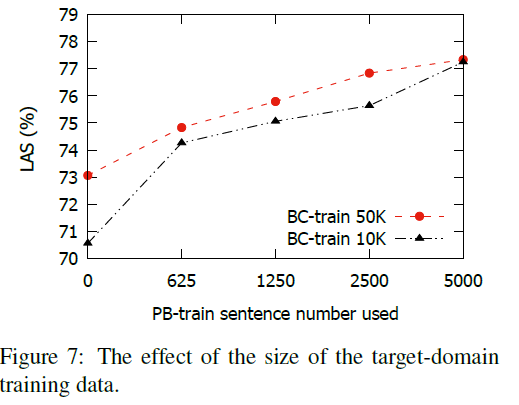


源域数据大小的影响如图6所示。我们使用BC火车的随机子集来固定目标域数据的大小，并增加源域数据的大小。 “ PB / ZX-train 1.5K”曲线基于随机的1500 PB / ZX-train语句，以便进行公平比较；“ PB-train 5K”曲线使用随机的5,000 PB-train语句，以了解 较大目标域数据的效果。 例如，在x轴上的“ 4”表示BC列的大小是目标域数据的四倍。

我们可以看到，当目标域数据的大小较小时，即“ PB / ZX-train 1.5K”，添加更多的源域BC-train数据将带来一致的改进。 在相同数据大小的分割中，“ PB火车1.5K”和“ ZX火车1.5K”仍然有较大的性能差距，这可能是由于ELMo更大的未标记的PB数据的规模，尽管ZX更易于分析，如第2节所述。

相反，对于较大的“ PB列5K”，当使用10K BC列语句时会获得峰值LAS，并且使用更多BC列数据甚至会稍微损害性能。 这表明当目标域训练数据较大时，源域数据的有用性受到限制。

目标域数据大小的影响如图7所示。由于ZX列的大小较小，因此我们仅使用PB列进行实验。 我们绘制了“ BCtrain 10K”曲线，因为先前的分析表明它与“ PB-train 5K”的组合已经达到了峰值性能。我们可以看到，以指数方式扩大目标域数据的大小会导致近乎线性的改善，这表明数据注释是提高跨域解析性能的最直接，最有效（或可能是必要的）方法。



另一方面，我们可以看到尽管BC训练50K和10K的最终性能几乎相同，但50K曲线显然更稳定和一致，这表明使用所有可用的源域数据通常是一个明智的选择。

5相关工作

在NLP和ML领域，领域适应一直是至关重要且具有挑战性的研究主题。 由于相关研究的范围广泛，因此我们尝试对语法解析具有高度相关性的一些代表性方法进行简要（且远未完成）综述。

无监督域适应。 由于缺乏足够的标记数据，因此，假设只有源域的标记数据，大多数以前的工作都集中在无监督域自适应上。 研究人员做出了巨大的努力，以从大规模的未标记目标域数据中学习有用的功能，这些数据通常更容易收集。 作为一种典型的半监督方法，自我训练对于跨域成分解析（McClosky等，2006）和依存关系解析（Yu等，2015）非常有用。 在域内和跨域依赖项解析中应用自我训练方面也有许多失败的工作。

Sagae和Tsujii（2007）将协同训练应用于CoNLL-2007跨域依赖性解析任务，并报告了正收益（Nivre等人，2007）。 相反，Dredze等（2007年）在相同的数据集上尝试了许多领域适应方法，但均未获得成功，这表明主要障碍来自目标域评估数据采用的不同注释准则。

源域数据选择是另一个有趣的研究方向。 给定目标域，其想法是从源域训练数据中自动选择一个最相关的子集来训练解析模型，而不是使用所有标记的数据（Plank和van Noord，2011； Khan等，2013）。

多源域适应问题假设存在多个源域的标记数据集。 给定目标领域，挑战在于如何有效地组合源领域中的知识。 McClosky等。 （2010年）首先提出这种情况进行成分分析。他们使用回归模型来预测跨域性能，然后使用这些值来组合在每个源域上独立训练的解析模型。 郭等。 （2018）在神经MTL框架下采用了类似的专家混合想法，并进行了情感分类和POS标签任务的实验。他们采用元训练来学习计算目标域示例与源域之间的点对点距离。

半监督域自适应假设存在一些（通常是非常小规模的）标记目标域数据，这些数据可用于直接学习特定于域的分布或特征。DauméeIII（2007）提出了一种简单而有效的特征增强方法，该方法在许多序列标记任务上表现良好。 这个想法是通过为每个域复制每个功能以及共享的（通用）伪域来区分特定于域的功能和一般功能。Finkel和Manning（2009）进一步提出了该想法的分级贝叶斯扩展。 正如Finkel和Manning（2009）指出的那样，在传统的离散功能ML框架下，这两项工作可以理解为MTL。

Kim等。 （2017）提出了一种神经专家混合方法，用于跨域意图分类和广告位标记。 与Guo等人的无监督方法不同。 （2018），他们使用少量目标域标记的数据来训练注意力模块以计算示例到域的距离。在解析社区中，Flannery和Mori（2015）建议使用主动学习对部分标记的目标域数据进行注释，以进行跨域日语依赖关系解析。

同样，乔希等。 （2018）注释了几十个部分标记的目标域句子，并带有一些用于跨域成分解析的括号。 这两个结果都报告了很大的改进，并且显示了即使是很少量的目标域注释也很有用，表明了半监督域自适应在解析中的巨大潜力。

**6总结**

这项工作基于我们两个新注释的大规模领域感知数据，即PB和ZX，解决了中文监督依赖分析的半监督域适配任务。 我们提出一种具有语料库加权的简单域嵌入方法，以有效地组合源域和目标域训练数据。 为了利用未标记的目标域数据，我们还基于最近提出的上下文化词表示法（ELMo），提出了一种有效的两阶段方法。 我们提出的半监督域自适应方法与在源BC上训练的非自适应解析器相比，在PB / ZX测试中绝对LAS分别提高了16.15％（77.16 vs.61.01）和15.56％（75.11 vs. 59.55）。

此外，详细的分析表明，扩大目标域标记的数据在提升跨域解析性能方面最有效。同时，更多的源域标记数据通常会带来更高且更一致的改进，尤其是当目标域训练数据的规模较小时。