深度学习何时需要树结构代表？

李继伟[1], 明升良[1]，丹·朱拉夫斯基[1]和爱德华·霍维[2] [1]斯坦福大学计算机科学系，斯坦福，加利福尼亚州 94305
[2]卡内基梅隆大学语言技术研究所，宾夕法尼亚州匹兹堡 15213
jiweil,lmthang, jurafsky@stanford.edu ehovy@andrew.cmu.edu

抽象的

递归神经模型是一种流行的架构，它使用句法解析树自下而上递归生成表示。但尚未进行严格的评估来表明这种基于语法的方法究竟适用于哪些任务。在本文中，我们进行了基准测试递归针对序列的神经模型复发性神经模型（简单循环和 LSTM 模型），尽可能地进行同类比较。我们研究了 4 个任务：（1）句子级别和短语级别的情绪分类；（2）将问题与答案短语匹配；（3）语篇解析；（4）语义关系提取（例如，*组件-整体*在名词之间）。

我们的目标是更好地理解递归模型何时以及为何能够胜过简单模型。我们发现递归模型主要有助于完成需要长距离关联词头的任务（如语义关系提取），尤其是在非常长的序列上。然后，我们介绍了一种允许递归模型实现类似性能的方法：在标点处将长句子分解为类似子句的单元，并在合并之前分别处理它们。因此，我们的结果有助于理解这两类模型的局限性，并为改进递归模型指明方向。

1 介绍

基于深度学习的方法学习词汇标记的低维实值向量，主要来自大规模数据语料库（例如，（Mikolov et al.， 2013；Le and Mikolov，2014；Collobert 等，2011），成功捕捉了文本的句法和语义方面。

对于输入较大的文本单元（例如短语、句子或文档）的任务，首先需要一个组合模型将标记聚合成具有固定维度的向量，该向量可用作其他 NLP 任务的特征。实现此目的的模型通常分为两类：复发性模型和递归楷模：

复发性模型（也称为顺序 循环模型可以成功处理时间序列数据（Pearlmutter，1989；Dorffner，1996），例如语音（Robinson 等人，1996；Lippmann，1989；Graves 等人，2013）或手写识别（Graves 和 Schmidhuber，2009；Graves，2012）。它们很早就被应用于 NLP（Elman，1990），将句子建模为按顺序处理的标记，每一步都将当前标记与先前构建的嵌入相结合。循环模型可以扩展到从左到右和从右到左的双向模型。这些模型通常不考虑除词序之外的语言结构。

递归神经模型（也称为 树相比之下，递归模型则由句法解析树构成。递归模型不是按顺序考虑标记，而是根据解析树的递归结构组合邻居，从叶子开始，以自下而上的方式递归进行，直到到达解析树的根。例如，对于短语*食物很好吃*，按照操作顺序*（（食物很好吃））*而不是顺序*（（（食物很好吃））*。已经提出了许多递归模型（例如，（Paulus et al。，2014；Irsoy and Cardie，2014）），并应用于各种NLP任务，其中包括蕴涵（Bowman，2013；Bowman et al。，2014），情绪分析（Socher et al。，2013；Irsoy and Cardie，2013；Dong et al。，2014），问答（Iyyer et al。，2014），关系分类（Socher et al。，2012；Hashimoto et al。，2013），

和话语（Li and Hovy，2014）。

递归模型的一个可能优势是它们能够捕捉长距离依赖关系：两个标记可能在结构上彼此接近，即使它们在单词序列中相距很远。例如，如果动词和其对应的直接宾语之间有许多形容词，但它们在解析树中相邻，则动词和其对应的直接宾语在标记方面可能相距很远（Irsoy 和 Cardie，2013 年）。但我们不知道这种优势是否真的很重要，如果很重要，那么对于哪些任务来说很重要，或者是否有其他问题在起作用。事实上，递归模型对解析的依赖也是一个潜在的缺点，因为解析相对较慢、依赖于领域并且可能出错。

另一方面，神经 NLP 多个子领域的最新进展表明，循环网络可能足以处理已提出使用递归模型的许多任务。没有解析结构的循环模型在机器翻译（例如，（Kalchbrenner 和 Blunsom，2013；3；Luong 等人，2014））的序列到序列生成（Sutskever 等人，2014）、解析（Vinyals 等人，2014）和情感方面表现出良好的效果，例如，基于循环的段落向量（Le 和 Mikolov，2014）在斯坦福情感库数据集上的表现优于递归模型（Socher 等人，2013）。

因此，本文的目标是研究一系列任务，以了解哪些类型的问题可以使用递归模型，哪些类型的问题可以使用递归模型，它们具有特定的优势。我们研究了四个具有不同属性的任务。

- 二进制情绪分类在句子级别（Pang et al.，2002）和短语级别（Socher et al.，2013）上，重点关注理解递归模型在处理各种场景中的语义组合方面的作用，例如不同长度的输入以及监督是否全面。

- 短语匹配在 UMD-QA 数据集 (Iyyer et al.，2014) 上进行分析可以帮助查看不同模型的中间组件输出之间的差异，即中间解析树节点的表示和不同时间步骤的循环模型的输出。它还有助于查看解析树是否有助于找到疑问句和目标短语之间的相似性。

- 语义关系分类基于 SemEval-2010（Hendrickx 等，2009）数据的结果可以帮助了解解析是否有助于处理长期依赖关系，例如序列中相距较远的两个单词之间的关系。

- 话语解析（RST 数据集）可用于测量解析对需要组合较大文本单元含义的话语任务的改进程度。话语解析将基本话语单元 (EDU) 视为要操作的基本单元，通常是短句。该任务还揭示了句法结构在多大程度上有助于获取短文本表示。

本文的主要目的是通过尽可能地进行同类比较来更好地理解何时以及为什么需要使用递归模型来超越简单模型。本文将现有模型应用于现有任务，几乎没有提供新的算法或任务。我们的目标是分析性的，以研究不同版本的递归和循环模型。这项工作有助于理解这两类模型的局限性，并提出改进循环模型的方向。

本文的其余部分组织如下：我们在第 2 节详细介绍了递归/循环模型的各个版本，在第 3 节介绍了任务和结果，并在第 4 节中进行了讨论。

2 递归和循环模型

2.1 符号

我们假设文本单元*年代*可以是短语、句子或文档，由一系列标记/单词组成：*年代*= *{*w_1, w_2，...，$w_{否}$*}*，在哪里*否*$_s$表示*年代*.每个单词w与一个K维向量嵌入相关联*埃*$_{瓦}$=

{$e_{瓦}, e_{瓦}, \ldots, e_{瓦}$*}*. 递归和重新的目标
当前模型是将序列映射到 K 维*埃*$_{年代}$，基于其标记及其对应的嵌入。

标准循环/序列模型循环网络依次接收消息*瓦*$_{吨}$在步骤*吨*，结合其向量表示*埃*$_{吨}$使用之前构建的隐藏向量H_{t-1}从时间

$t-1$、计算最终的当前嵌入 $H_{吨}$，并将其传递到下一步。嵌入 $H_{吨}$当前时间*吨*因此：

$$H_{吨}=F(西\cdot H_{t-1}+五\cdot 埃_{吨}) \quad (1)$$

在哪里*西*和*五*表示组合矩阵。如果 $否_s$表示序列的长度，$H_{否_s}$代表整个序列*年代*。

标准递归/树模型标准递归模型的工作方式类似，但按解析树顺序而不是序列顺序处理相邻单词。它以自下而上的方式递归地根据每个父节点的直接子节点计算其表示，直到到达树的根节点。对于给定节点η在树和它的左孩子中$\eta_{左边}$（具有代表性）$埃_{左边}$）和右孩子$\eta_{正确的}$（具有代表性）$埃_{正确的}$），标准递归网络计算$埃_{\eta}$

如下：

$$埃_{\eta}=F(西\cdot 埃_{\eta_{左边}}+五\cdot 埃_{\eta_{正确的}}) \quad (2)$$

双向模型（Schuster 和 1997) 为循环框架添加了双向性，其中每次的嵌入都是向前和向后计算的：

$$\begin{aligned} H^{\rightarrow}_{吨} &= F(西_{\rightarrow}\cdot H_{\rightarrow t-1}+五_{\rightarrow}\cdot 埃_{吨}) \\ H^{\leftarrow}_{吨} &= F(西_{\leftarrow}\cdot H_{\leftarrow 吨+1}+五_{\leftarrow}\cdot 埃_{吨}) \end{aligned} \quad (3)$$

通常情况下，句子的最终表示可以通过连接计算向量来实现
来自两个方向[$埃_{\leftarrow}$ $_1$，$埃_{否}$]或使用毛皮
另一个组合操作来保持向量 di-
意义性

$$H_{吨}=F(西_{大号}\cdot[H_{\leftarrow 吨},H_{吨}]) \quad (4)$$

在哪里$西_{大号}$表示*钾* $\times 2$*钾*维矩阵。

长短期记忆网络（LSTM）LSTM 模型（Hochreiter 和 Schmidhuber，1997）定义如下：给定一个输入序列 $X=\{X_1, X_2, \ldots, X_{nx}\}$其中，LSTM 将每个时间步与输入、记忆和输出门相关联，分别表示为*我吨*，$F_{吨}$和$o_{吨}$. 我们在符号上消除歧义*埃*和*H*：$埃_{吨}$表示时间步 t 处各个文本单元（例如单词或句子）的向量，而$H_{吨}$表示 LSTM 模型在时间 t 通过结合以下公式计算出的向量$埃_{吨}$和 H_{t-1}。σ表示 S 型函数。向量表示$H_{吨}$每个时间步*吨*是（谁）给的：

$$\begin{bmatrix} 我_{吨} \\ F_{吨} \\ o_{吨} \\ 升_{吨} \end{bmatrix} = \begin{bmatrix} \sigma \\ \sigma \\ \sigma \\ 双曲函数 \end{bmatrix} 西\cdot \begin{bmatrix} H_{t-1} \\ 埃_{吨} \end{bmatrix} \quad (5)$$

$$C_{吨}=F_{吨}\cdot C_{t-1}+我_{吨}\cdot 升_{吨} \quad (6)$$

$$H_{吨}=o_{吨}\cdot C_{吨} \quad (7)$$

在哪里$西 \in R^{4钾\times 2钾}$。标签在这短语/句子级别是从上一个时间步骤输出的预测表示。

树形 LSTM最近的研究将 LSTM 思想扩展到基于树的结构 (Zhu et al., 2015; Tai et al., 2015)，将记忆门和遗忘门与解析树的节点关联起来。

双向 LSTM这些结合了双向模型和 LSTM。

3 实验

在本节中，我们详细介绍了实验设置和结果。我们考虑以下任务，每个任务代表不同类别的 NLP 任务。

1. 二元情绪分类在 Pang et al. (2002) 数据集上。这解决了监督仅在经过一长串操作后才会全局出现的问题。

2. 在斯坦福情绪树库上进行情绪分类（Socher 等人，2013)：为需要学习局部成分（例如否定、情绪或其他由短语结构提示的成分）的单词和短语找到全面的标签。

3. 句子目标匹配在 UMD-QA 上数据集（Iyyer et al.， 2014）：学习目标句子和源句子中的成分之间的匹配，这些匹配是递归模型的解析树节点和循环模型的不同时间步长。

4.语义关系分类在 SemEval-2010 任务中 (Hendrickx 等，2009)。学习两个单词之间的长距离关系，这些单词在序列上可能相距很远。

5. 语篇分析（Li 等人，2014；Hernault 等人，2010）：根据计算的表示学习句子与句子的关系。

在每种情况下，我们都遵循原始论文中描述的协议。我们首先将算法变体分为以下两组：

- 标准树模型 vs 标准序列模型 vs 标准双向序列模型

- LSTM 树模型、LSTM 序列模型与 LSTM 双向序列模型。

我们采用了神经模型的标准训练框架：对于每个任务，我们使用 AdaGrad（Duchi 等人，2011）和小批量（Cotter 等人，2011）进行随机梯度下降。如果原始数据集中可用，则使用开发数据集调整参数，如果没有，则使用交叉验证调整参数。导数是通过标准反向传播（Goller 和 Kuchler，1996）计算得出的。要调整的参数包括小批量的大小、学习率和 L2 惩罚的参数。运行迭代的次数被视为要调整的参数，并且在开发集上实现最佳性能的模型将用作要评估的最终模型。

对于没有进行重复实验的设置，采用引导检验进行统计显著性检验（Efron and Tibshirani，1994）。达到显著性水平 0.05 的检验分数以星号 (*) 标记。

3.1 斯坦福情绪树库

任务描述我们从斯坦福情感树库（Socher 等人，2013）开始。该数据集包含每个解析树组成部分的黄金标准标签，从句子到短语再到单个单词。

当然，在基于解析树的数据集上实施序列模型得出的任何结论可能都必须被削弱，因为序列模型可能仍然受益于数据集的收集方式。尽管如此，我们还是在这个数据集上添加了一个评估，因为它已经成为神经模型评估中广泛使用的基准数据集。

对于递归模型，我们遵循 Socher 等人 (2013) 中的协议，其中解析树中的节点嵌入是从递归模型获得的，然后输入到 softmax 分类器。我们将数据集转换为循环模型使用，如图 1 所示。每个短语都从解析树节点重建并被视为单独的数据点。由于树库包含 11,855 个句子和 215,154 个短语，因此为循环模型重建的数据集包含 215,154 个示例。在短语和句子上对模型进行评估

级别（82,600 个实例）和句子根级别（2,210 个实例）。

	细粒度	二进制
树	0.433	0.815
顺序 P 值	0.420（-0.013） 0.042*	0.807 (-0.007) 0.098
双序列 P 值	0.435 (+0.08) 0.078	0.816 (+0.002) 0.210

表 1：斯坦福情绪树库根级别的测试集准确率。

	细粒度	二进制
树	0.820	0.860
顺序 P 值	0.818 (-0.002) 0.486	0.864 (+0.004) 0.305
双序列 P 值	0.826 (+0.06) 0.148	0.862 (+0.002) 0.450

表 2：斯坦福情绪树库在短语级别的测试集准确率。

结果如表 1 和表 2 所示[1]。当将标准版本的树模型与序列模型进行比较时，我们发现它在根级别识别方面有一点帮助（对于序列但不适用于双序列），但在短语级别没有显着改善。

长短期记忆（LSTM）Tai 等人 (2015) 发现 LSTM 树模型在句子根级别评估比序列模型更难。我们通过训练更深、更复杂的模型来进一步探索这项任务。我们研究以下三个模型：

1.树结构LSTM模型（Tai等，2015）[2]。

2. 深度 Bi-LSTM 序列模型（表示为 顺序）将整个句子视为一个序列。

3. 深度 Bi-LSTM 分层序列模型（表示为层次序列首先使用标点符号查找表（即逗号、句号、问号和感叹号）将句子切成一系列子句子。首先分别计算每个子句子的表示，然后另一层序列 LSTM

[1]我们实现的递归模型的性能与 Socher 等人 (2013) 报告的性能并不完全相同，但相对差异约为1%到2%。

[2]Tai 等人在根级细粒度评估方面实现了 0.510 的准确率，如 (Tai et al., 2015) 所报告，与我们实施的结果 (0.504) 相似。

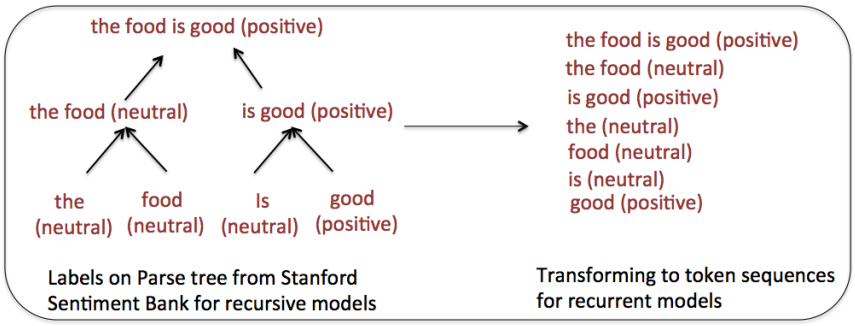

图 1：将斯坦福情绪树库转换为序列模型的序列。

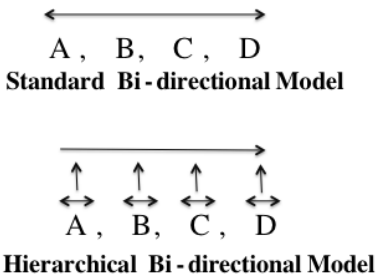

图 2：两个序列模型的说明。A、B、C、D 表示用标点符号分隔的从句或子句。

（单向）用于连接子句子。如图2所示。

我们考虑第三个模型，因为 Tai 等人（2015 年）使用的数据集包含长句，并且评估仅在句子根级别进行。由于解析算法会自然地将长句分解为子句，我们想知道性能提升是由子句内解析树结构引入的，还是仅仅通过将句子更广泛地分割为类似子句的单元而引入的；后者的优势可以通过使用基于标点符号的子句边界近似值来近似。

我们对每个算法运行 15 次迭代。每次迭代结束时都会收集参数；在开发集上表现最好的参数将用于测试集。整个过程在单个 GPU 机器上大约需要 15-20 分钟[3]为了进行更有说服力的比较，我们没有使用从同一数据集生成并行示例的引导测试。相反，我们对每个算法重复了上述过程 20 次，并在表 3 中报告了准确率和标准差。

树形 LSTM 与标准双向序列模型相当或略胜一筹

[3]Tesla K40m，2880 个 Cuda 核心。

模型	一切都好	根细	根粗
树形 LSTM	83.4 (0.3)	50.4 (0.9)	86.7 (0.5)
双序列	83.3 (0.4)	49.8 (0.9)	86.7 (0.5)
层次序列	82.9 (0.3)	50.7 (0.8)	86.9 (0.6)

表 3：斯坦福情绪树库的测试集准确率（含偏差）。对于我们的实验，我们报告了 20 次运行的准确率（含标准偏差）。

（双尾 p 值等于 0.041*，并且仅在根级别，短语级别的 p 值为 0.376）。分层序列模型实现了相同的性能，p 值为 0.198。

讨论上述结果表明，长句的分句分割略微提高了性能，这三个短语级情感评估模型之间的差异很小，这也支持了这一结果。因此，长句的分句分割为基于解析树的模型提供了一个简单的近似值。

我们认为，对于分句分割带来的性能略好一些的原因有以下几点：

1. 将小句作为基本单位（标点符号近似于小句）可以保留文本的语义结构。

2. 否定或连词等语义成分通常出现在小句层面。单独处理小句，然后将它们组合起来，可以形成小句间成分。

3. 与标准模型相比，分层模型中误差反向传播到单个标记所用步骤更少。考虑一个电影评论“情节虽然简单，但我仍然非常喜欢它”。使用标准循环模型，预测误差需要 12 个步骤才能回到第一个标记“简单”：

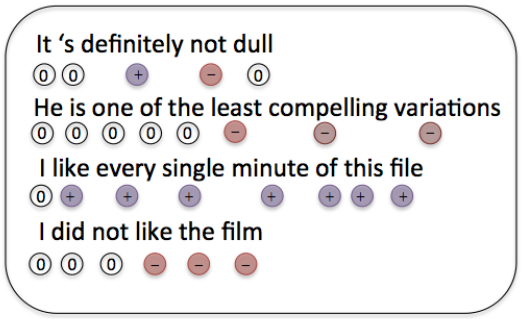

图 3： 使用一个定向（从左到右）LSTM。每个时间步骤的决策都是通过将从 LSTM 计算出的嵌入输入到 softmax 分类器中来做出的。

错误→很多→A→它→喜欢→仍然→我→，→曾是→阴谋→这→作为→简单的

在分层模型中，第二条子句被压缩为一个组件，因此误差传播由以下公式给出：

错误→第二条款→菲第一条→曾是→阴谋→这→作为→简单的。

带子句分割的传播仅包含 8 个操作。因此，这样的过程倾向于减弱梯度消失问题，从而可能产生更好的性能。

3.2 二元情感分类（Pang）

任务描述：Pang 等人（2002）的情感数据集由句子组成，每个句子都有一个情感标签。我们将原始数据集分为训练（8101）/开发（500）/测试（2000）。没有采用 Socher 等人（2011b）中描述的预训练程序。使用 skip-gram 初始化词嵌入，并在学习过程中保持不变。我们使用 word2vec 包在 Wikipedia+Gigaword 数据集上训练 skip-gram 嵌入[4]。句子级嵌入被输入到 S 型分类器中。50 维向量的性能如下表所示：

讨论为什么解析树对这项任务没有帮助？一个可能的解释是监督信号与局部组合结构的距离。Pang 等人的数据集的平均句子长度为 22.5 个单词，这意味着

[4]https://code.google.com/p/word2vec/

	标准	长短期记忆（LSTM）
树	0.745	0.774
顺序	0.733 (-0.012)	0.783 (+0.008)
P 值	0.060	0.136
双序列	0.754 (+0.09)	0.790 (+0.016)
P 值	0.058	0.024*

表 4：使用标准模型设置在 Pang 的情绪数据集上进行的测试集准确度。

需要经过多个步骤，情绪相关证据才会浮出水面。因此，目前尚不清楚是否可以学习局部组合运算符（例如否定）；训练数据量很少（大约 8,000 个示例），并且仅在句子级别进行的情绪监督可能不易传播到深埋的局部短语。

3.3 问答匹配

任务描述：在问答数据集 QANTA 中[5]，每个答案都是一个标记或短语。该任务不同于标准的以生成为重点的 QA 任务，而是形式化为多类分类任务，将源问题与预定义候选短语池中的候选短语进行匹配我们在这里给出一个说明性示例：

问题：*他留下了一本未完成的小说，小说主人公伪造父亲的签名逃学，并假装想要参军以逃避征兵。请说出这位《魔山》和《威尼斯之死》的德国作家的名字。*

回答：*托马斯·曼*从短语池中。其他候选词可能包括乔治·华盛顿、查理·卓别林等。

Iyyer 等人 (2014) 的模型最小化了问题解析树中答案嵌入和节点嵌入之间的距离。具体来说，让*C*表示问题的正确答案*年代*，嵌入~*C*， 和*是*表示任何随机错误答案。目标函数对每个节点的表示之间的点积求和*η*沿着问题解析树和答案表示：

$$大号= \sum_{\eta \in [解析树]} \sum_{是} 最大限度(0，1-~c \cdot 埃_\eta + ~z \cdot 埃_\eta) \quad (8)$$

在哪里*埃η*表示根据递归神经模型计算出的解析树节点的嵌入。

[5]http://cs.umd.edu/ miyyer/qblearn/。由于隐私问题，公开发布的数据集比 (Iyyer et al., 2014) 中使用的版本要小，因此我们的数字无法与 (Iyyer et al., 2014) 中的数字相比。

这里的解析树是遵循的依赖解析（Iyyer et al.，2014）。

通过将框架调整为循环模型，我们最小化了答案嵌入与从每个时间步计算出的嵌入之间的距离*吨*序列：

$$大号=\sum_{吨\in[1,否]}\sum_{是} 最大限度(0，1-~C\cdot 埃吨+~是\cdot 埃吨) \quad (9)$$

在测试时，模型会从候选集中选择损失分数最低的答案。从表 5 中的结果可以看出，树模型和序列模型之间的差异仅在 LSTM 设置中显著；其他设置没有显著差异。

	标准	长短期记忆（LSTM）
树	0.523	0.558
顺序 P 值	0.525 (+0.002) 0.490	0.546 (-0.012) 0.046*
双序列 P 值	0.530 (+0.007) 0.075	0.564 (+0.006) 0.120

表 5：UMD-QA 数据集的测试集准确率。

讨论UMD-QA 任务代表了一组情况，由于我们对匹配的监督不足（很难知道解析树中的哪个节点或哪个时间步为答案提供了最直接的证据），因此必须通过查看和迭代所有子单元（解析树或时间步中的所有节点）来做出决策。在池化结构中可以找到类似的想法（例如 Socher 等人 (2011a)）。

上述结果表明，对于我们尝试将目标与不同的源组件（即，树模型的解析树节点和序列模型的不同时间步骤）对齐的任务，来自序列模型的组件能够嵌入重要信息，尽管序列模型组件只是句子片段，因此通常不像解析树组件那样具有语言意义。

3.4 语义关系分类

任务描述：SemEval-2010 任务 8（Hendrickx 等，2009）是寻找名词对之间的语义关系，例如在“我的[公寓]$_{e1}$有一个相当大的[厨房]$_{e2}$”对[公寓]和[厨房] 作为*组件-整体*。数据集包含 9 个有序关系，因此该任务被形式化为 19 类分类问题，其中有向关系被视为单独的标签；详情请参阅 Hendrickx 等人（2009 年；Socher 等人（2012 年）。

对于递归实现，我们遵循 Socher 等人 (2012) 中定义的神经框架。检索解析树中两个名词之间的路径，并基于递归模型计算嵌入并将其输入到 softmax 分类器[6]. 检索到的路径被转换为循环模型，如图 5 所示。

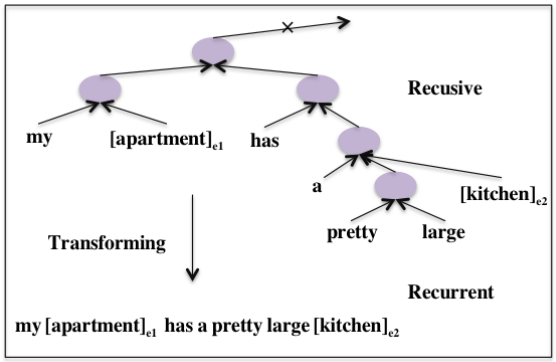

图4：语义关系分类模型说明。

讨论与早期的任务不同，这里的递归模型比所有版本的相应循环版本产生更好的性能（例如，标准树与标准序列， *页*=0。004)。这些结果表明，需要整合句子中相距很远的结构是递归模型优于循环模型的任务的特征。在基于解析的模型中，两个目标词在决策过程中比循环模型更早地结合在一起，循环模型必须记住一个目标，直到另一个目标出现。

3.5 语篇解析

任务描述：我们的最终任务是基于 RST-DT 语料库（Carlson 等）进行语篇解析

[6]（Socher 等人，2012 年）通过将复杂的模型 MV-RNN（其中每个单词都以矩阵和向量的形式呈现）与人为特征工程相结合，实现了最先进的性能。同样，由于 MV-RNN 难以适应循环版本，因此我们不采用这种最先进的模型，而仅遵循第 2 节中描述的递归模型的通用版本，因为我们的主要目标是比较等效的递归和循环模型，而不是实现最先进的模型。

	标准	长短期记忆（LSTM）
树	0.748	0.767
顺序 P 值	0.712 (-0.036) 0.004*	0.740 (-0.027) 0.020*
双序列 P 值	0.730 (-0.018) 0.017*	0.752 (-0.014) 0.041*

表 6：SemEval-2010 语义关系分类任务的测试集准确率。

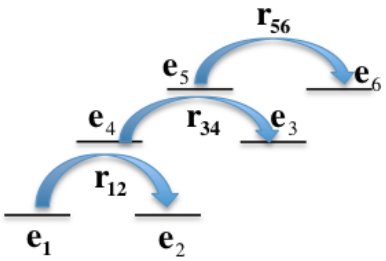

图5： 话语解析的一个例子。[埃1, e2，...]表示 EDU（基本语篇单位），每个单位由一系列标记组成。[n12, r三十四, r56]表示要分类的关系。首先使用二元分类模型来决定两个 EDU 是否应该合并，然后使用多类分类器来决定关系类型。

语篇分析的另一种方法是为文档构建语篇树，该方法基于在基本语篇单元 (EDU) 之间分配修辞结构理论 (RST) 关系。由于语篇关系表达了 语篇的连贯结构，因此它们可能表达了与情感或名词关系不同的组合意义。有关语篇解析和 RST-DT 语料库的更多详细信息，请参阅 Hernault 等人 (2010)。

相邻 EDU 的表示被输入到二分类（两个 EDU 是否相关）和多类关系分类模型中，如 Li 等人（2014 年）所定义。然后将相关 EDU 合并为一个新的 EDU，该 EDU 的表示是通过基于前两个相关 EDU 的神经组合操作获得的。重复此步骤，直到所有单元都合并。

篇章解析以 EDU 为基本操作单位；EDU 是短句，而不是完整的句子，平均长度为 7.2 个单词。递归和循环模型应用于 EDU 以创建嵌入，用作篇章解析的输入。我们使用此任务有两个原因：（1）说明句法解析树是否可用于获取短句的表示。（2）衡量解析的程度可以改善需要结合更大文本单元含义的话语任务。

传统上，模型是根据三个指标来评估的，即跨度[7]、核性[8]，并识别两个子句之间的修辞关系。由于篇幅限制，我们仅关注最后一个，即修辞关系识别，因为 (1) 只有正确标记跨度和核性，关系标签才被视为正确 (2) 子句之间的关系识别可以提供有关模型表示句子语义能力的更多见解。为了进行简单的比较，没有添加额外的人工开发功能。

	标准	长短期记忆（LSTM）
树	0.568	0.564
顺序 P 值	0.572 (+0.004) 0.160	0.563 (-0.002) 0.422
双序列 P 值	0.578 (+0.01) 0.054	0.575 (+0.012) 0.040*

表 7：RST 话语解析数据集上关系识别的测试集准确率。

讨论我们发现等效的循环模型和递归模型之间没有太大差异。我们提出了两种可能的解释。（1）EDU 往往很短；因此对于某些子句，解析可能不会改变单词操作的顺序。即使对于那些顺序被解析树改变的短语，短语对最终表示的影响也可能不够大。（2）与早期任务不同，在早期任务中，文本表示会立即用作分类器的输入，而本文介绍的算法在 EDU 合并过程中采用了额外的神经组成级别。我们怀疑神经层可能充当信息过滤器，将信息中的杂质与精良信息分开，这反过来又使模型对初始输入具有更强的免疫力。

4。讨论

我们比较了递归和循环神经模型在 4 个领域中的 5 个不同的 NLP 任务上的表征学习，在这些领域中，递归神经模型被认为可以取得良好的性能（Socher 等人，2012；Socher 等人，2013；Li 等人，2014；Iyyer 等人，2014）。

与任何模型之间的比较一样，我们的结果也有一些注意事项：首先，我们探索了最普遍或最基本的递归形式。

[7]在空白的树结构上。
[8]基于具有核指示的树结构。

深入比较模型，而不是各种复杂的算法变体。这是因为随着模型变得越来越复杂（例如，层数、每层内的隐藏单元数等），公平比较变得越来越困难。因此，尽管具有多层的深度神经模型可以产生更好的结果，但本研究中使用的大多数神经模型仅由一层神经组成。因此，我们的结论可能仅限于本文采用的算法，目前尚不清楚它们是否可以扩展到其他变体或最新的最先进技术。其次，为了“公平地”比较模型，我们强制每个模型以完全相同的方式进行训练：具有小批量的 AdaGrad、相同的初始化集等。然而，这可能不一定是训练每个模型的最佳方式；针对特定模型量身定制的不同训练策略可能会提高它们的性能。从这个意义上说，我们在本文中试图“公平”可能仍然是不公平的。

考虑到这些警告，我们的结论可以总结如下：

- 在语义关系提取等任务中，单个词条需要远距离关联，而递归模型则大放异彩。这表明，对于许多其他需要远距离语义依赖性的任务（例如，像中英翻译这样需要大量重新排序的语言之间的翻译），递归模型的句法结构可能会提供有用的功能。
- 在监督充足的情况下，树模型对长序列的帮助往往大于对短序列的帮助：树模型对斯坦福情绪树库的根级识别略有帮助，但在短语级帮助不大。采用双向版本的循环模型似乎在很大程度上弥补了这一差距，产生了相同甚至更好的结果。
- 在监督不足的长序列中，例如在 Pang 等人的数据集中（监督仅存在于长序列之上），基于树的模型和基于序列的模型之间没有观察到显著差异。
- 在基于树的模型表现良好的情况下，可以简单地近似于基于树的模型似乎可以将循环模型提升到同等或几乎同等的性能：（1）将长句（标点符号）分解成一系列类似子句的单元，（2）分别处理这些子句，（3）将它们连接在一起。对于情感任务，该模型有时与树模型一样好，这表明树模型有帮助的原因之一是它将长句分解成更易于管理的单元。
- 尽管循环模型中的组件（来自不同时间步骤的输出）不具有语言意义，但它们在嵌入信息证据方面可能与具有语言意义的短语（由解析树节点表示）一样好，如 UMD-QA 任务所示。事实上，与我们并行的近期研究（Bowman 等人，2015 年）表明，像 LSTM 这样的循环模型可以发现隐式递归组合结构。

5 致谢

我们特别要感谢 Richard Socher 和 Kai-Sheng Tai 的深刻评论、建议和提议。我们还要感谢 Sam Bowman、Ignacio Cases、Jon Gauthier、Kevin Gu、Gabor Angeli、Sida Wang、Percy Liang 和斯坦福 NLP 小组的其他成员，以及匿名审阅者对这项工作各个方面的有益建议。我们感谢 NVIDIA 公司捐赠 Tesla K40 GPU 的支持。我们非常感谢 Enlight 基金会研究生奖学金、彭博有限合伙企业的捐赠、美国国防高级研究计划局 (DARPA) 深度探索和过滤文本 (DEFT) 计划（空军研究实验室 (AFRL) 合同编号 FA8750-13-2-0040）以及美国国家科学基金会颁发的奖项 IIS-1514268。本材料中表达的任何意见、发现、结论或建议均为作者的观点，并不一定反映彭博有限合伙企业、DARPA、AFRL、NSF 或美国政府的观点。

参考

Dzmitry Bahdanau、Kyunghyun Cho 和 Yoshua Bengio. 2014. 神经机器翻译由联合

学习对齐和翻译。 *arXiv 预印本 arXiv:1409.0473*。

Samuel R Bowman、Christopher Potts 和 Christopher D Manning。2014.用于学习逻辑语义的递归神经网络。*arXiv 预印本 arXiv:1406.1827*。

塞缪尔·R·鲍曼、克里斯托弗·D·曼宁和 Christopher Potts。2015 年。无树结构架构的神经网络中的树结构组合。*arXiv 预印本 arXiv:1506.04834*。

Samuel R Bowman。2013 年。递归神经张量网络学习逻辑推理？*arXiv 预印本 arXiv:1312.6192*。

Lynn Carlson、Daniel Marcu 和 Mary Ellen Okurowski。2003。在修辞结构理论框架下构建语篇标记语料库。在*话语和对话文本、语音和语言技术的当前和新方向。*第 22 卷。Springer。

罗南·科洛贝尔、杰森·韦斯顿、莱昂·博图、迈克尔 Karlen, Koray Kavukcuoglu 和 Pavel Kuksa。2011 年。自然语言处理（几乎）从零开始。*机器学习研究杂志*, 12：2493–2537。

Andrew Cotter、Ohad Shamir、Nati Srebro 和 Karthik Sridharan。2011 年。通过加速梯度方法实现更好的小批量算法。*神经信息处理系统的进展*，第 1647-1655 页。

李东、魏福如、谭传奇、唐笃雨、明 Zhou 和 Ke Xu。2014 年。自适应递归神经网络用于目标相关的 Twitter 情绪分类。*计算语言学协会第 52 届年会论文集*，第 49-54 页。

Georg Dorffner。1996 年。时间序列神经网络处理。在*神经网络世界*。

John Duchi、Elad Hazan 和 Yoram Singer。2011 年。用于在线学习和随机优化的自适应次梯度方法。*机器学习研究杂志*, 12：2121–2159。

布拉德利·埃夫隆和罗伯特·J·蒂布希拉尼。1994 年。一个 *bootstrap 简介*.CRC出版社。

Jeffrey L Elman。1990 年。寻找时间结构。*认知科学*《细胞与分子生物学》, 14(2): 179–211。

Christoph Goller 和 Andreas Kuchler。1996 年。学习通过结构反向传播来处理与任务相关的分布式表示。在*神经网络，1996 年，IEEE 国际会议*，第 1 卷，第 347-352 页。IEEE。

Alex Graves 和 Juergen Schmidhuber。2009 年。离线使用多维循环神经网络进行手写识别。*神经信息处理系统的进展*，第 545-552 页。

Alex Graves、Abdel-rahman Mohamed 和 Geoffrey Hinton。2013 年。使用深度循环神经网络进行语音识别。*声学、语音和信号处理 (ICASSP)，2013 年 IEEE 国际会议*，第 6645-6649 页。IEEE。

Alex Graves。2012 年。监督序列标记利用循环神经网络，*计算智能研究。*第 385 卷。Springer。

桥本一真、三轮诚、津义正-ruoka 和 Takashi Chikayama。2013 年。用于语义关系分类的递归神经网络的简单定制。在*增强型神经网络LP*，第 1372-1376 页。

Iris Hendrickx、Su Nam Kim、Zornitsa Kozareva、Preslav Nakov、Diarmuid Ó Séaghdha、Sebastian Padó、Marco Pennacchiotti、Lorenza Romano 和 Stan Szpakowicz。2009 年。Semeval-2010 任务 8：名词对之间语义关系的多向分类。在*语义评估研讨会论文集：最新成果和未来方向*, 第 94–99 页。计算语言学协会。

雨果·埃尔诺 (Hugo Hernault)、赫尔穆特·普林丁格 (Helmut Prendinger)、石冢满 (Mitsuru Ishizuka)。2010.Hilda：一种使用支持 向量机分类的话语解析器。*对话与讨论*，1（3）。

Sepp Hochreiter 和 Jürgen Schmidhuber。1997 年。长短期记忆。 *神经计算*，9(8): 1735–1780。

Ozan Irsoy 和 Claire Cardie。2013 年。双向重新用于具有结构的标记级标记的草书神经网络。*arXiv 预印本 arXiv:1312.0493*。

Ozan Irsoy 和 Claire Cardie。2014 年。深度递归神经网络在语言组合性方面的应用。*神经信息处理系统的进展*，第 2096-2104 页。

Mohit Iyyer、Jordan Boyd-Graber、Leonardo Claudino、Richard Socher 和 Hal Daumé III。2014 年。用于段落事实性问题回答的神经网络。在*2014 年自然语言处理实证方法会议 (EMNLP) 论文集*，第 633-644 页。

Nal Kalchbrenner 和 Phil Blunsom。2013 年。Recurrent 连续翻译模型。1700–1709。 *在增强型神经网络LP*、页

Quoc V Le 和 Tomas Mikolov。 2014. 分布式句子和文档的表示。*arXiv 预印本 arXiv:1405.4053*。

Jiwei Li 和 Eduard Hovy。2014 年。凝聚态模型基于分布式句子表征的*2014 年自然语言处理实证方法会议 (EMNLP) 论文集*

Jiwei Li、Rumeng Li 和 Eduard Hovy。2014 年。复发深度模型进行话语解析。在*2014 年实证方法会议论文集*

自然语言处理（EMNLP），第 2061-2069 页。

Richard P Lippmann。1989 年。神经网络回顾用于语音识别。*神经计算*,1(1):1–38。

Thang Luong、Ilya Sutskever、Quoc V Le、Oriol Vinyals 和 Wojciech Zaremba。2014 年。解决神经机器翻译中的稀有词问题。 *ACL 会议纪要*.2015 年。

Tomas Mikolov、Wen-tau Yih 和 Geoffrey Zweig。2013. 连续空间词表征中的语言规律。*肝癌筛查*，第 746–751 页。

Bo Pang、Lillian Lee 和 Shivakumar Vaithyanathan。2002. 竖起大拇指？：使用机器学习技术进行情绪分类。*ACL-02 自然语言处理经验方法会议论文集-第 10 卷*, 第 79–86 页。计算语言学协会。

Romain Paulus、Richard Socher 和 Christopher D Manning。2014 年。全局信念递归神经网络。*神经信息处理系统的进展*，第 2888-2896 页。

Barak A Pearlmutter。1989。学习状态空间传输循环神经网络中的轨迹。*神经计算*《自然》杂志，1(2): 263–269。

托尼·罗宾逊、迈克·霍赫伯格和史蒂夫·雷纳尔斯。1996. 循环神经网络在连续语音识别中的应用。*自动语音和说话人识别*，第 233-258 页。Springer。

迈克·舒斯特 (Mike Schuster) 和库尔迪普·K·帕利瓦尔 (Kuldip K Paliwal)。 1997.双向循环神经网络。*信号处理，IEEE 论文集*《自然》杂志, 45(11): 2673–2681。

Richard Socher、Eric H Huang、Jeffrey Pennin、Christopher D Manning 和 Andrew Y Ng。2011a。用于释义检测的动态池化和展开递归自动编码器。在*神经信息处理系统的进展*，第 801-809 页。

Richard Socher、Jeffrey Pennington、Eric H Huang、Andrew Y Ng 和 Christopher D Manning。2011b。用于预测情绪分布的半监督递归自动编码器。*自然语言处理实证方法会议论文集*, 第 151-161 页。计算语言学协会。

理查德·索彻、布罗迪·胡瓦尔、克里斯托弗·D·曼宁、和 Andrew Y Ng。2012 年。通过递归矩阵向量空间实现语义组合性。在*2012 年自然语言处理和计算自然语言学习实证方法联合会议论文集*, 第 1201–1211 页。计算语言学协会。

Richard Socher、Alex Perelygin、Jean Y Wu、Jason Chuang、Christopher D Manning、Andrew Y Ng 和 Christopher Potts。2013 年。情绪树库上语义组合的递归深度模型。*自然语言处理实证方法会议论文集（EMNLP）*，第 1631-1642 页。

Ilya Sutskever、Oriol Vinyals 和 Quoc VV Le。2014 年。使用神经网络进行序列到序列的学习。在*神经信息处理系统的进展*，第 3104-3112 页。

Kai Sheng Tai、Richard Socher 和 Christopher D Manning. 通过树形结构长短期记忆网络改进语义表征。 *访问控制列表*.2015年。

Oriol Vinyals、Lukasz Kaiser、Terry Koo、Slav Petrov、Ilya Sutskever 和 Geoffrey Hinton。2014 年。语法作为一门外语。*arXiv 预印本 arXiv:1412.7449*。

朱晓丹、Parinaz Sobihani 和郭红宇。2015. 递归结构的长期短期记忆。*第 32 届国际机器学习会议 (ICML-15) 论文集*，第 1604-1612 页。