# 基于神经网络的Seq2Seq的学习

#### 摘要

深度神经网络（DNN）是功能强大的模型，它们在一些困难的学习任务上取得了卓越的表现。无论有没有很大标记的训练集，它们都能很好地工作，即便没有大标记的训练集可用。但是它们不能用于映射序列到序列的任务。在本文中，我们提出了一种通用的序列映射到序列的方法，该方法对序列结构做了最少的假设。我们的方法使用多层长短期记忆网络（LSTM）来映射输入序列到一个固定维度的向量，然后用另一个多层LSTM来从该向量中解码得到目标序列。我们的主要成果是在WMT-14数据集上进行的英文到法文的翻译任务，由LSTM产生的翻译在整个测试集上获得34.8的BLEU分数。另外，LSTM在对长句的翻译任务上并没有很大的困难。为了进行比较，我们使用基于短语的SMT系统在相同的数据集上获得33.3的BLEU得分。当我们使用LSTM调整上述SMT系统产生的1000种假设BLEU得分增加到36.5，这与理论上完美的结果接近。该LSTM还学会了敏感的短语和句子表达。它以语序为主，对主动和被动语态相对不变。最后，我们发现颠倒了所有源句子中单词的顺序（目标句子不变）显着改善了LSTM的能力，因为这样做引入了源和目标之间的许多短期依赖关系，使问题更容易的优化。

### 1介绍

神经网络（DNN）是非常强大的机器学习模型，可以在语音识别[13][7]和视觉对象识别等难题上实现卓越的性能[19，6，21，20]。 DNN功能强大，因为它们可以在适度数量的步骤中执行任意并行的计算。一个令人惊讶的例子是它们只使用两层的隐藏层对N个N位数排序能力，所以，虽然神经网络是与传统的统计模型相关，但是他们学习了复杂的计算。此外，只要训练集足够多，就可以用监督反向传播对大型的DNN进行训练，并有足够的信息调整网络的参数。因此，如果存在一个参数设置，能使DNN达到良好效果，那么监督反向传播就能找到这些参数，解决问题。

尽管DNN具有灵活而且强大的功能，但只能应用于其输入和目标可以用固定维度的矢量进行编码的问题。这是一个严重的限制，因为很多很多的问题适合用序列长度未知的序列表示。例如，语音识别和机器翻译是一个映射序列到另一个序列的问题。同样，回答问题也可以被看作是将表示问题的一系列单词映射到表示答案的单词序列。因此我们很清楚，在这个领域中，学习把序列映射到另一个序列的方法，是很有用处的。

序列相关任务给DNNs提出了难题，因为DNNs要求输入和输出的维度是已知并且固定的。在本文中，我们提出一个基于长时间短期记忆（LSTM）结构[16]的应用来解决一般的序列到序列的问题。这个想法是使用一个LSTM来读取输入序列，按照时间步一步一步读取输入序列，以获得足够大的固定维度的向量表示。然后使用另一个LSTM从该向量提取出输出序列（图1）。 第二个LSTM本质上是一个循环神经网络语言模型[28,23,30]，除了它是以输入序列为条件的。由于输入和输出之间存在相当大的时间差（图1），LSTM能学习到远距离时间相关性数据的能力使其成为该问题的自然选择。

已经有许多相关的利用神经网络的方法尝试解决序列到序列的一系列问题。我们的方法与Kalchbrenner和Blunsom[18]密切相关，他们是第一个将整个输入句子映射为向量的人，并且与Cho等人非常相似[5]。Graves[10]引入了一种新颖的可区分关注机制，它允许神经网络关注不同部分的输入，而这个想法的一个优雅变种已成功应用在Bahdanau等人的机器翻译上。[2]。 联结主义的序列分类是另一种将序列映射到具有神经网络的序列的流行技术，尽管它假设输入和输出之间是单调对齐的[11]。

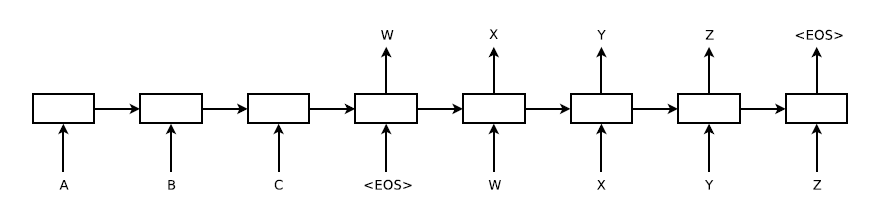


图1：我们的模型读取一个输入句子“ABC”，并生成“WXYZ”作为输出语句。在输出了序列结尾的记号之后，这个模型便停止继续预测了。注意LSTM读的输入序列是颠倒的，因为这样做能在数据中引入许多的短期依赖关系，使得优化问题变得更容易

这项工作的主要结果如下。 在WMT'14英语至法语翻译任务中，我们使用简单的从左至右的波束搜索解码器通过从5个很深的LSTM（每个具有380M参数）的集合直接提取翻译来获得34.71的BLEU得分。 这是迄今为止通过大型神经网络进行直接翻译所获得的最佳结果。 为了比较，在该数据集上使用SMT基线的BLEU得分为33.30 [29]。 LSTM获得了34.81 BLEU的分数，其中词汇量为80k字。只要翻译出来的译文包含一个不存在于80k中的词，得分就会受到惩罚。 这个结果表明，一个还没达到最优化的神经网络结构有很大的提升空间，它胜过了一个成熟的基于短语的SMT系统。

最后，我们使用LSTM在相同的任务中，使用 SMT系统的公开可用的1000个最佳列表[29]，我们获得了36.5的BLEU得分。通过这样做，我们将基线提高了3.2个BLEU点并且接近于先前的完美的状态（37.0[9]）。

令人惊讶的是，尽管最近有相关架构研究人员的经验，我们模型的LSTM并没有遭遇到非常长的句子造成的糟糕问题[26]。 由于我们在源句中颠倒了单词的顺序，而不是训练和测试集中的目标句子，所以我们在长句中表现得很好。通过这样做，我们引入了许多短期依赖关系，使优化问题变得更加简单（请参见第2节和第3.3节）。因此，随机梯度算法可以使得LSTMs学习到长句子中单词之间的相互关系。这个把源句单词反转的简单诀窍是这项工作的关键技术贡献之一。

LSTM有一个很有用特性。它可以学习将可变长度的输入句子映射为固定维度的向量表示。 翻译往往是对源句子做解释，通过学习目标句，使得LSTM能够找到并捕捉代表目标句含义的特征。具有相似含义的句子彼此接近，而不同的含义的句子相差甚远。定性评估也支持这一说法，这表明我们的模型能意识到词序，无论是主动还是被动语态都是可以的。

## 2模型

递归神经网络（RNN）[31,28]是一种基于前馈神经网络的，为了处理序列问题，而提出的网络。 给定一系列输入（），标准RNN通过迭代以下等式来计算输出序列

（）：

无论输入和输出之间的对齐是否提前知道，RNN都可以很轻松将序列映射到序列。然而，将RNN应用于其输入和输出序列具有不同长度，且具有复杂的非单调关系的问题，是很难的。

一个简单通用的序列学习策略是使用一个RNN结构将输入序列映射到一个固定大小的向量，然后用另一个RNN结构将该向量映射到目标序列（这种方法也被Cho采用。[5]）。 因为RNN提供了所有相关信息，原则上该模型可以工作，但由于长期相关性问题[14，4]（图1）[16,15]，使得RNN难以训练。然而，长短期记忆（LSTM）[16]已知可以学习长句子单词的依赖问题，所以LSTM能应用在这样的情况下。

LSTM的目标是估计条件概率，其中是输入序列并且是其对应的输出序列，其长度可以不同于。LSTM通过得到一个由LSTM最后一个隐藏状态给出的输入序列的固定维数的向量表示来计算该条件概率，然后计算的概率，用标准LSTM-LM公式表示，其初始隐藏状态设为的向量表示：

在这个等式中，每个分布用词汇表中的所有单词的函数表示。 我们使用Graves的LSTM公式[10]。 注意，我们这里要求每个句子以特殊的句尾符号“<EOS>”结尾，这使得模型能够定义所有可能长度序列的分布。总的方案如图1所示，其中表示的是LSTM读入“A”，“B”，“C”，“<EOS>”的过程，然后计算得到“W”，“X “，”Y“，”Z“，”<EOS>“的概率。

我们的实际模型在三个比较重要方面与以上描述不同。首先，我们使用了两个不同的LSTM：一个用于输入序列，另一个用于输出序列，因为这样做可以以可忽略的计算成本增加模型的参数，并且能自然的在多个语言上训练LSTM[18]。其次，我们发现较深LSTMs明显优于较浅LSTMs，因此我们选择了四层LSTM。 第三，我们发现对输入句子的单词顺序进行反转对结果来说非常好。 因此，我们不是把句子a，b，c映射到句子A，B，C，而是要求LSTM把c，b，a映射到A，B，C，其中A,B,C是 a，b，c的翻译。 这样，a接近A，b接近B，等等，这使SGD很容易在输入和输出之间“建立通信”。我们发现了这个简单的数据转换可以大大提升LSTM的性能。

## 3实验

我们以两种方式将我们的方法应用于英语WMT'14至法语MT任务。 我们使用它来直接翻译输入句子，而不使用SMT系统作参考，我们用SMT基线的n个最佳列表来重打分。 我们记录这些翻译方法的准确度，呈现处原来的样本，并对所得到的句子进行可视化处理。

#### 3.1数据集处理细节

我们使用的数据是英语至法语数据集（WMT'14）。 我们用包含348M法文单词和304M英文单词的12M句子的子集训练我们的模型，这是经过挑选的来自[29]数据集。我们选择了这个翻译任务和这个特定的训练集子集，因为它们都是公开可用的，它们提供了一个标记过的训练集和测试集，以及1000个来自SMT基线的最佳列表。

典型的神经语言模型依赖于每个单词的矢量表示，我们对这两种语言都使用了固定的词汇表。我们使用源语言中最常使用的16万个单词和目标语言中最常用的8万个单词。每一个超出词汇范围外的单词都被一个特殊的“UNK”标记取代。

#### 3.2编码和新编码

我们的实验的核心是在大量的句子对上训练出一个庞大的LSTM。给定源句子S，我们通过最大化S对应的正确的翻译结果T的对数概率来训练它。所以训练的目标是：

其中是训练集。一旦训练完成，我们将根据LSTM的输出来找到最相似的翻译结果并输出：

来

我们使用简单的从左到右的定向搜索解码器搜索最可能的翻译结果，该解码器维持少量的部分臆测结果，合计为B，其中臆测结果是某些翻译的前缀。在每个时间步中，我们用词汇表中的每个可能单词来扩展其中的每个部分臆测结果。这大大增加了臆测的数量，所以我们根据模型的对数概率保留B个最可能的臆测结果。只要“<EOS>”符号附加到结果上，它就会移除该结果，并将其添加到最后的结果集合中。虽然这个解码器是近似的，但实现起来很简单。有趣的是，即使B大小为1，我们的系统也表现良好，而B大小为2的时候的结果表示定向搜索的优越性（表1）。

我们还使用LSTM来调整由基线系统生成的1000个最佳列表[29]。为了评估n个最佳列表，我们用我们的LSTM计算了每个结果的对数概率，并用他们的分数和LSTM的得分取了一个平均值。

#### 3.3反转源句子序列

虽然LSTM能够解决长期依赖性问题，但我们发现当源句子颠倒时（目标语句不颠倒），LSTM学习得更好。通过这样做，LSTM的测试损失度从5.8下降到4.7，并且解码翻译的测试BLEU得分从25.9上升到30.6。

尽管我们对这种现象没有完整的解释，但我们认为这是由于对数据集引入了许多短期依赖的表现。通常，当我们将源语句与目标语句连接起来时，源语句中的每个单词与目标语句中的相应单词都很远。因此，这个问题有很大的“最小时滞”效应[17]。 通过翻译源语句中的词语，源语言和目标语言中相应词语之间的平均距离不变。 但是，源语言中的前几个单词

现在已经非常接近目标语言的前几个单词，因此该问题的“最小时滞”效应大大降低。 因此，反向传播在源语句和目标语句之间“建立沟通”的时间更加容易，从而大大提高了总体表现。

最初，我们认为，反转输入句子只会导致目标句子早期部分预测更加容易和确定，并且会导致后面部分的预测不太准确。然而，反向源语句训练的LSTM在长句子上的训练比在原始源句子上训练的LSTM好得多（见3.7节），这表明反转输入句子导致LSTM具有更好的效果。

#### 3.4训练细节

我们发现LSTM模型相当容易训练。我们使用4层的深层LSTMs，每层1000个神经和1000维的词语嵌入，输入词汇量为160,000，输出词汇量为80,000。我们发现深层LSTM明显优于浅层LSTM，其中每个附加层减少了近10％的损失度，可能是由于它们的隐藏层神经元个数更多。我们在每个输出中使用了超过80,000个单词的天真softmax。得到的LSTM有380M参数，其中64M是纯粹的循环连接（32M代表“编码器”LSTM，32M代表“解码器”LSTM）。 完整的培训细节如下：

* 我们把LSTM’s的所有参数都初始化为-0.08到0.08的均匀分布。
* 我们使用随机梯度下降法，设置学习率为0.7。经过5轮操作后，我们每半轮都把学习率降低为之前的一半。我们一共对该模型训练了7.5轮。
* 我们在梯度下降中使用了一批128个序列。
* 尽管使用LSTMs不会产生梯度消失的问题，但是会有梯度爆炸的问题。因此我们严格限制了梯度的范围[10,25]，当它的标准超过阈值，就按比例缩放。对于每个训练批次，我们计算，是梯度除以128。若，我们执行。
* 不同的句子有着不同的长度。大多数的句子是比较短的（长度一般为20-30），但是一些句子的长度是很长的（长度大于100），因此一个128大小的随机选择的迷你批将包含很多的短句和很少的长句，就会导致大多数迷你批的计算上的浪费。为了处理该问题，我们确保所有的在一个迷你批的句子都大概在相同的长度上，该处理大概使训练加速了两倍。

#### 3.5并行化

我们实现了深度LSTM的C ++版本，使用上一节中提到的方法，在单个GPU上进行训练，它每秒处理大约1,700字的规模。这对于我们的要求来说太慢了，因此我们使用8个GPU机器对我们的模型进行了并行化计算。LSTM的每一层都在不同的GPU上执行，并在计算完成后立即将其输出传递给下一个GPU（或图层）。我们的模型具有4层LSTM，每层LSTM都位于独立的GPU上。剩余的4个GPU被用于并行化计算softmax，因此每个GPU负责乘以1000×20000矩阵。由此实现能提高速度达到每秒6,300（包括英文和法文）的字数，迷你批的大小为128。在此实验过程中，实验需要大约十天的时间。

#### 3.6实验结果

我们使用BLEU评分[24]来评估翻译的质量。 我们使用multi-bleu.pl1基于标记化的预测和基础事实，计算了我们的BLEU得分。这种评估BELU得分的方法与[5]和[2]一致，并再现了[29]的33.3得分。但是，如果我们以这种方式评估[9]的技术状态系统（其预测可以从statmt.org \ matrix下载），我们会得到37.0，这比statmt.org \ matrix报告的35.8更大。

结果如表1和表2所示。我们的最佳结果是通过随机初始化和小型随机序列不同的LSTM集合获得的。尽管LSTM集合的解码翻译没有超越现有技术水平，但这是第一次，纯粹的神经网络翻译系统在大型MT任务上优于基于短语的SMT基线，尽管其无法处理超出词汇范围的词汇。通过评估基线系统的1000个最佳列表，LSTM和最先进的技术水平的差距在0.5BLEU点内。

这有几个BLEU评分的变体，每个变量都用perl脚本定义。

|  |  |
| --- | --- |
| 方法 | BLEU得分(ntst14) |
| Bahdanau et al. [2] | 28.45 |
| Baseline System [29] | 33.30 |
| Single forward LSTM, beam size 12 | 26.17 |
| Single reversed LSTM, beam size 12 | 30.59 |
| Ensemble of 5 reversed LSTMs, beam size 1 | 33.0 |
| Ensemble of 2 reversed LSTMs, beam size 12 | 33.27 |
| Ensemble of 5 reversed LSTMs, beam size 2 | 34.50 |
| Ensemble of 5 reversed LSTMs, beam size 12 | 34.81 |

表1：LSTM在WMT’14英语到法语训练集(ntst14)的表现。注意由5个LSTM层的定向搜索大小为2的总体效果比单个LSTM层的定向搜索大小为12的效果更好。

|  |  |
| --- | --- |
| 方法 | BLEU得分(ntst14) |
| Baseline System [29] | 33.3 |
| Cho et al. [5] | 34.54 |
| State of the art [9] | 37.0 |
| Rescoring the baseline 1000-best with a single forward LSTM | 35.61 |
| Rescoring the baseline 1000-best with a single reversed LSTM | 35.85 |
| Rescoring the baseline 1000-best with an ensemble of 5 reversed LSTMs | 36.5 |
| Oracle Rescoring of the Baseline 1000-best lists | ~45 |

表2：在WMT’14英语到法语测试集(ntst14)中使用神经网络和SMT系统相结合的方法的结果。

#### 3.7处理长句的表现

我们惊讶地发现，LSTM在长句中表现得很好，数据已经展示在图3中。表3给出了几个长句和它们的译文的例子

#### 3.8模型分析

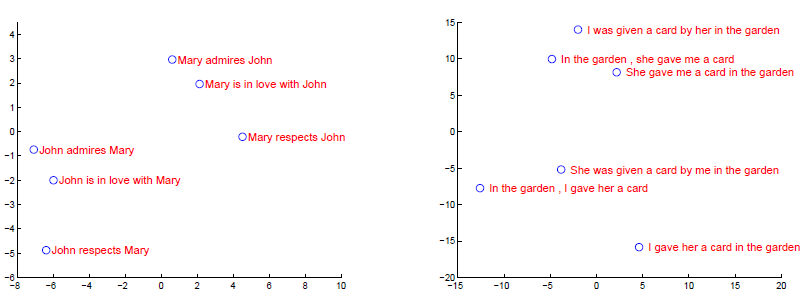


图2：该图显示了在处理图中的短语之后获得的LSTM隐层状态的二维PCA投影。这些短语按照某些含义聚集在一起，在这些例子中，这些短语主要是词序的函数，这将很难用袋装词模型来捕捉。请注意,两个集群都有类似的内部结构。

我们模型的一个富有吸引力的特征是它能够将一系列的单词转化为固定维数的矢量。 图2显示了一些学习的表示。该图清楚地表明，模型对词的顺序敏感，而对主动语态和被动语态的替换不敏感。图中的二维投影使用PCA获得。

|  |  |
| --- | --- |
| 类型 | 句子 |
| 我们的模型产生的输出 | Ulrich UNK , membre du conseil d’ administration du constructeur automobile Audi , affirme qu’ il s’ agit d’ une pratique courante depuis des ann´ees pour que les t´el´ephones portables puissent ˆetre collect´es avant les r´eunions du conseil d’ administration afin qu’ ils ne soient pas utilis´es comme appareils d’ ´ecoute `a distance . |
| 真实翻译 | Ulrich Hackenberg , membre du conseil d’ administration du constructeur automobile Audi ,d´eclare que la collecte des t´el´ephones portables avant les r´eunions du conseil , afin qu’ ilsne puissent pas ˆetre utilis´es comme appareils d’ ´ecoute `a distance , est une pratique courante depuis des ann´ees . |
| 我们的模型产生的输出 | “ Les t´el´ephones cellulaires , qui sont vraiment une question , non seulement parce qu’ ils pourraient potentiellement causer des interf´erences avec les appareils de navigation , maisnous savons , selon la FCC , qu’ ils pourraient interf´erer avec les tours de t´el´ephone cellulaire lorsqu’ ils sont dans l’ air ” , dit UNK . |
| 真实翻译 | “ Les t´el´ephones portables sont v´eritablement un probl`eme , non seulement parce qu’ ils pourraient ´eventuellement cr´eer des interf´erences avec les instruments de navigation , mais parce que nous savons , d’ apr`es la FCC , qu’ ils pourraient perturber les antennes-relais de t´el´ephonie mobile s’ ils sont utilis´es `a bord ” , a d´eclar´e Rosenker . |
| 我们的模型产生的输出 | Avec la cr´emation , il y a un “ sentiment de violence contre le corps d’ un ˆetre cher ” ,qui sera “ r´eduit `a une pile de cendres ” en tr`es peu de temps au lieu d’ un processus de d´ecomposition “ qui accompagnera les ´etapes du deuil ” . |
| 真实翻译 | Il y a , avec la cr´emation , “ une violence faite au corps aim´e ” , qui va ˆetre “ r´eduit `a un tas de cendres ” en tr`es peu de temps , et non apr`es un processus de d´ecomposition , qui “ accompagnerait les phases du deuil ” . |

表3：一些由LSTM与真实翻译一起产生的长翻译的例子。读者可以使用谷歌翻译验证翻译是否合理。

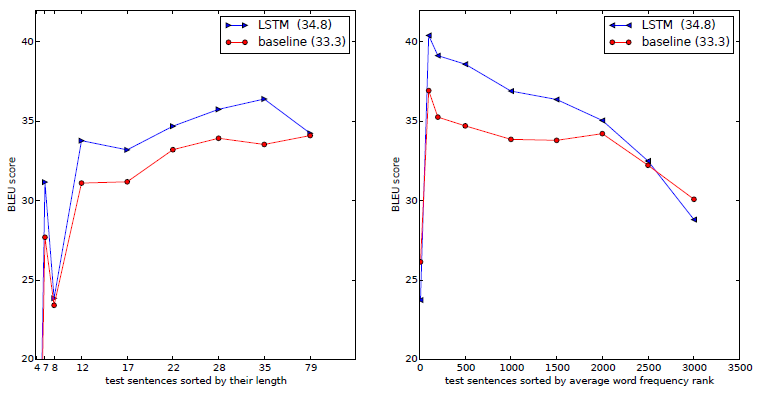


图3：左图显示了我们的系统随着句子长度增长展现的性能，其中x轴对应于按其长度排序的测试语句，并由实际序列长度标记。不超过35个单词的句子没有退化，最长的句子只有轻微的退化。右边的曲线图显示了LSTM在逐渐变得更稀有的单词上的表现，其中x轴对应于按“平均单词频率排序”排序的测试句子。

#### 4相关工作

为了把神经网络应用在机器翻译上，有很多人做了大量的工作。到目前为止，在MT任务中应用RNN语言模型（RNNLM）[23]或前馈神经网络语言模型（NNLM）[3]的最简单和最有效的方法是编写强MT基线的最佳列表 [22]，它可以明显的提高翻译质量。

最近，研究人员开始研究如何将与源语言相关的信息纳入NNLM。做这项工作的人包括Auli等[1]，他们将NNLM与输入句子的主要模型相结合，从而提高了结果的性能。 Devlin等[8]遵循类似的方法，但他们将NNLM并入到MT系统的解码器中，并使用解码器的对齐信息为NNLM提供输入句子中最有用的单词。他们的方法非常成功，在基线方面取得了很大的进步。

我们的工作与Kalchbrenner和Blunsom[18]密切相关，他们是第一个将输入句子映射到矢量然后再映射回句子的人，尽管他们使用的是卷积神经网络将句子映射到矢量，失去了单词的位置信息。Cho的工作类似于先前提到的[5]，他使用类似LSTM的RNN架构将语句映射到矢量中并映射回去，尽管它们的主要焦点是将它们的神经网络集成到SMT系统中。 Bahdanau等[2]也试图用一个神经网络直接翻译，使用注意机制来克服Cho等人经历过的网络在长句的表现的问题[5]并取得了欢欣鼓舞的结果。同样，Pouget-Abadie[26]试图解决Cho等人的网络中存在的记忆问题。[5]通过翻译源句子的方式产生平滑的翻译，这是类似于基于短语的方法。我们认为他们可以基于网络，简单地训练反向源语句来实现类似的改进。

端到端的训练[12]也是Hermann等人关注的焦点，其模型使用前馈网络来表示输入和输出，并将它们映射到空间中的类似点。然而，他们的方法不能直接生成翻译结果：为了获得翻译结果，他们需要在预先计算的句子数据库中查找最接近的向量，或者重定义句子。

#### 5结论

在这项工作中，我们发现一个词汇量有限的深层LSTM可以超越标准的基于SMT的系统，这个系统的词汇量在大型MT任务上是无限的。我们简单的基于LSTM的MT方法的模型的成功，表明了若有足够的训练数据，类似的模型能在许多其他序列学习问题上做得很好。 我们对反转源句中的单词后所获得的改进程度感到惊讶。因此而得到了一个结论：找到一个具有最多短期依赖性的问题编码方式非常重要，因为它们会使学习问题变得更加简单。 尤其是，当我们发现不能用非反转的源语言训练标准RNN（如图1所示），我们反转源句子使得标准RNN更易于训练（尽管我们没有对其进行验证实验）。

我们也对LSTM具备正确翻译非常长的句子的能力感到惊讶。我们最初相信LSTM由于记忆力有限而在长句中将失败，并且其他研究人员也称其和我们相似的模型在长句的表现不佳[5,2,26]。然而，LSTMs在反转源数据后的数据集进行训练后，在翻译长句时没有遇到什么困难。

最重要的是，我们演示了一种简单，直接和相对来说有优化空间的方法可以超越成熟的SMT系统，因此进一步的工作可能会得到更高的翻译准确性。这些结果表明我们的方法可能会在其他具有挑战性的序列问题上做得很好。

#### 6鸣谢

我们感谢Samy Bengio，Jeff Dean，Matthieu Devin，Geoffrey Hinton，Nal Kalchbrenner，Thang Luong，Wolfgang Macherey，Rajat Monga，Vincent Vanhoucke，Peng Xu，Wojciech Zaremba和Google Brain团队提供的有用评论和讨论。

#### 参考文献

1. M. Auli, M. Galley, C. Quirk, and G. Zweig. Joint language and translation modeling with recurrent neural networks. In *EMNLP*, 2013.
2. D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.
3. Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin. A neural probabilistic language model. In *Journal of Machine Learning Research*, pages 1137–1155, 2003.
4. Y. Bengio, P. Simard, and P. Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2):157–166, 1994.
5. K. Cho, B.Merrienboer, C. Gulcehre, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Arxiv preprint arXiv:1406.1078*, 2014.
6. D. Ciresan, U. Meier, and J. Schmidhuber. Multi-column deep neural networks for image classification. In *CVPR*, 2012.
7. G. E. Dahl, D. Yu, L. Deng, and A. Acero. Context-dependent pre-trained deep neural networks for large vocabulary speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing – Special* *Issue on Deep Learning for Speech and Language Processing*, 2012.
8. J. Devlin, R. Zbib, Z. Huang, T. Lamar, R. Schwartz, and J. Makhoul. Fast and robust neural network joint models for statistical machine translation. In *ACL*, 2014.
9. Nadir Durrani, Barry Haddow, Philipp Koehn, and Kenneth Heafield. Edinburgh’s phrase-based machine translation systems for wmt-14. In *WMT*, 2014.
10. A. Graves. Generating sequences with recurrent neural networks. In *Arxiv preprint arXiv:1308.0850*, 2013.
11. A. Graves, S. Fern´andez, F. Gomez, and J. Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *ICML*, 2006.
12. K. M. Hermann and P. Blunsom. Multilingual distributed representations without word alignment. In *ICLR*, 2014.
13. G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury. Deep neural networks for acoustic modeling in speech recognition. *IEEE* *Signal Processing Magazine*, 2012.
14. S. Hochreiter. Untersuchungen zu dynamischen neuronalen netzen. *Master’s thesis, Institut fur Infor-matik, Technische Universitat, Munchen*, 1991.
15. S. Hochreiter, Y. Bengio, P. Frasconi, and J. Schmidhuber. Gradient flow in recurrent nets: the difficulty of learning long-term dependencies, 2001.
16. S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 1997.
17. S. Hochreiter and J. Schmidhuber. LSTM can solve hard long time lag problems. 1997.
18. N. Kalchbrenner and P. Blunsom. Recurrent continuous translation models. In *EMNLP*, 2013.
19. A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In *NIPS*, 2012.
20. Q.V. Le, M.A. Ranzato, R. Monga, M. Devin, K. Chen, G.S. Corrado, J. Dean, and A.Y. Ng. Building high-level features using large scale unsupervised learning. In *ICML*, 2012.
21. Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998.
22. T. Mikolov. *Statistical Language Models based on Neural Networks*. PhD thesis, Brno University of Technology, 2012.
23. T. Mikolov, M. Karafi´at, L. Burget, J. Cernock`y, and S. Khudanpur. Recurrent neural network based language model. In *INTERSPEECH*, pages 1045–1048, 2010.
24. K. Papineni, S. Roukos, T. Ward, and W. J. Zhu. BLEU: a method for automatic evaluation of machine translation. In *ACL*, 2002.
25. R. Pascanu, T. Mikolov, and Y. Bengio. On the difficulty of training recurrent neural networks. *arXiv preprint arXiv:1211.5063*, 2012.
26. J. Pouget-Abadie, D. Bahdanau, B. van Merrienboer, K. Cho, and Y. Bengio. Overcoming the curse of sentence length for neural machine translation using automatic segmentation. *arXiv preprint* *arXiv:1409.1257*, 2014.
27. A. Razborov. On small depth threshold circuits. In *Proc. 3rd Scandinavian Workshop on Algorithm Theory*, 1992.
28. D. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986.
29. H. Schwenk. University le mans. http://www-lium.univ-lemans.fr/˜schwenk/cslm\_ joint\_paper/, 2014. [Online; accessed 03-September-2014].
30. M. Sundermeyer, R. Schluter, and H. Ney. LSTM neural networks for language modeling. In *INTER-SPEECH*, 2010.
31. P. Werbos. Backpropagation through time: what it does and how to do it. *Proceedings of IEEE*, 1990.

原文：

Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. "Sequence to sequence learning with neural networks." *Advances in neural information processing systems*. 2014.