https://yifdu.github.io/2019/02/15/NLP%E8%AE%BA%E6%96%87%E9%98%85%E8%AF%BB%EF%BC%88%E4%BA%94%EF%BC%89/

关于神经机器翻译的特性：编码器-解码器方法

摘要

神经机器翻译是一种相对较新的基于神经网络的统计机器翻译方法。神经机器翻译模型通常由编码器和解码器组成。编码器从可变长度的输入句子中提取固定长度的表示，解码器从该表示中生成正确的翻译。在本文中，我们重点分析了神经机器翻译的特性，使用了两个模型：RNN编码器-解码器和一个新提出的门控递归卷积神经网络。我们发现，神经机器翻译在短句子且没有未知词的情况下表现相对良好，但是随着句子长度和未知词数量的增加，其性能迅速下降。此外，我们发现所提出的门控递归卷积网络可以自动学习句子的语法结构。

1 引言

最近提出了一种全新的基于神经网络的统计机器翻译方法（Kalchbrenner和Blunsom，2013；Sutskever等，2014）。这种新方法被称为神经机器翻译，受到了深度表示学习的最新趋势的启发。在（Kalchbrenner和Blunsom，2013；Sutskever等，2014；Cho等，2014）中使用的所有神经网络模型都由编码器和解码器组成。编码器从可变长度的输入句子中提取一个固定长度的向量表示，解码器则从该表示中生成一个正确的、可变长度的目标翻译。

神经机器翻译的出现在实践和理论上都具有重要意义。神经机器翻译模型所需的内存仅为传统统计机器翻译（SMT）模型的一小部分。我们在本文中训练的模型总共只需500MB的内存。这与现有的SMT系统形成了鲜明对比，后者通常需要数十GB的内存。这使得神经机器翻译在实践中具有吸引力。此外，与传统的翻译系统不同，神经翻译模型的每个组成部分都是联合训练的，以最大化翻译性能。

由于这种方法相对较新，对这些模型的属性和行为的研究还不多。例如：这种方法在哪种类型的句子上表现更好？源语言/目标语言词汇选择如何影响性能？神经机器翻译在哪些情况下会失败？

了解这种新的神经机器翻译方法的属性和行为至关重要，以确定未来的研究方向。此外，了解神经机器翻译的优点和缺点可能会带来更好地整合SMT和神经机器翻译系统的方法。

在本文中，我们分析了两种神经机器翻译模型。其中一种是最近在（Cho等，2014）中提出的RNN编码器-解码器模型。另一种模型将RNN编码器-解码器模型中的编码器替换为一种新颖的神经网络，我们称之为门控递归卷积神经网络（grConv）。我们在从法语到英语的翻译任务上评估了这两种模型。

我们的分析表明，随着源语句长度的增加，神经机器翻译模型的性能迅速下降。此外，我们发现词汇大小对翻译性能有很大影响。尽管如此，从定性的角度来看，我们发现这两种模型大部分时间能够生成正确的翻译。此外，我们新提出的grConv模型能够在无监督的情况下学习源语言的一种语法结构。

2 可变长度序列的神经网络

在本节中，我们描述了两种能够处理可变长度序列的神经网络类型。它们分别是循环神经网络和我们提出的门控递归卷积神经网络。

2.1 循环神经网络

注：循环神经网络（Recurrent Neural Network，RNN）是一种能够处理可变长度序列的神经网络。它具有循环连接，可以在每个时间步骤中接收输入，并在内部维护一个隐藏状态。RNN的隐藏状态可以捕捉到序列中的上下文信息，使得网络能够对序列进行建模和处理。

在RNN中，门控隐藏神经元（Gated Hidden Neurons）是一种特殊类型的神经元，例如长短期记忆（Long Short-Term Memory，LSTM）和门控循环单元（Gated Recurrent Unit，GRU）。这些门控机制通过使用可学习的门控单元来控制信息的流动，从而允许网络选择性地记忆和忘记输入。这种门控机制有助于解决RNN中的梯度消失和梯度爆炸问题，并提高了网络对长期依赖关系的建模能力。

2.1 Recurrent Neural Network with Gated Hidden Neurons

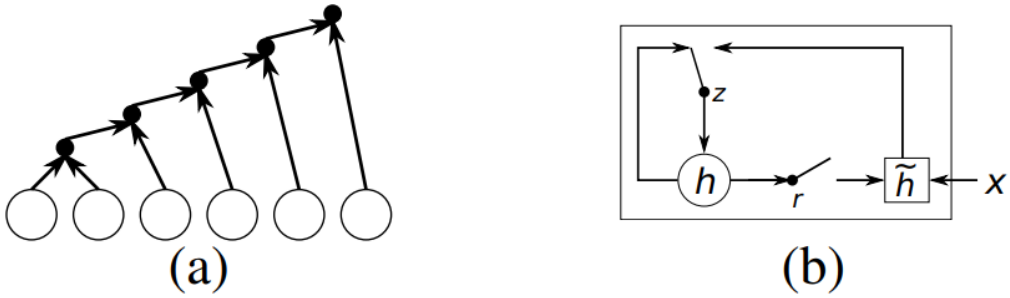


图1：(a) 循环神经网络的图形示意，(b) 自适应遗忘和记忆的隐藏单元

循环神经网络（RNN，图1(a)）通过在时间上维护隐藏状态h，对可变长度的序列x = (x1; x2; · · · ; xT) 进行处理。在每个时间步t，隐藏状态h(t)通过以下方式进行更新：

$$\mathbf{h}^{(t)} = f\left(\mathbf{h}^{(t-1)}, \mathbf{x}_t\right),$$

其中，f是一个激活函数。通常，f可以简单地对输入向量进行线性变换，求和，并应用逐元素的逻辑sigmoid函数。一个RNN可以有效地用来学习一个变长序列的分布，通过学习下一个输入的分布 p(xt+1 | xt; · · · ; x1)。例如，在一个由1-of-K向量组成的序列的情况下，可以通过一个RNN学习该分布，并将其作为输出。

$$p(x_{t,j} = 1 \mid \mathbf{x}_{t-1}, \ldots, \mathbf{x}_1) = \frac{\exp\left(\mathbf{w}_j \mathbf{h}_{\langle t \rangle}\right)}{\sum_{j'=1}^{K} \exp\left(\mathbf{w}_{j'} \mathbf{h}_{\langle t \rangle}\right)},$$

对于所有可能的符号j = 1,....,K，其中wj是权重矩阵W的行，RNN可以学习并输出联合分布：

$$p(x) = \prod_{t=1}^{T} p(x_t \mid x_{t-1}, \ldots, x_1).$$

最近在（Cho等，2014年）中提出了一种用于RNN的新激活函数。这个新的激活函数通过两个称为重置（reset）门r和更新（update）门z的门控单元来增强传统的逻辑sigmoid激活函数。每个门都依赖于先前的隐藏状态h(t−1)和当前的输入xt，控制信息的流动。这类似于长短期记忆（LSTM）单元（Hochreiter和Schmidhuber，1997）。关于这个单元的详细信息，我们建议读者参考（Cho等，2014）和图1(b)。在本文的剩余部分，我们始终使用这个新的激活函数。

2.2 门控递归卷积神经网络（Gated Recursive Convolutional Neural Network)

注：
门控递归卷积神经网络（Gated Recursive Convolutional Neural Network，grConv）是一种神经网络模型，用于处理自然语言处理任务，特别是句子级别的任务，如情感分析、文本分类和句子生成等。

grConv结合了递归神经网络（Recursive Neural Network，RNN）和卷积神经网络（Convolutional Neural Network，CNN）的优点，并引入了门控机制。它具有递归层和卷积层，用于捕捉句子中的上下文信息和局部特征。门控机制允许网络选择性地记忆和忘记输入，有助于解决梯度消失和梯度爆炸问题，并提高网络对长期依赖关系的建模能力。

grConv的递归层通过递归地组合子句来构建整个句子的表示。卷积层则在句子的局部窗口上进行特征提取。网络的输出可以用于进行各种句子级别的任务。

grConv在自然语言处理领域取得了良好的性能，并且具有较低的计算复杂度。它是一种灵活且强大的模型，适用于处理不同长度的句子和各种文本分类任务。

除了RNN之外，处理可变长度序列的另一种自然方法是使用递归卷积神经网络，其中每个级别的参数在整个网络中共享（参见图2(a)）。在本节中，我们介绍一种二进制卷积神经网络，其权重被递归地应用于输入序列，直到输出一个固定长度的向量。除了常规的卷积结构外，我们提出使用先前提到的门控机制，使递归网络能够即时学习源句子的结构。

这种递归卷积神经网络的基本思想是通过递归地应用卷积操作，逐渐将输入序列压缩成一个固定长度的向量表示。在每个递归级别，权重参数被共享，以便网络可以学习捕捉输入序列中的结构信息。同时，引入门控机制可以帮助网络选择性地记忆和忘记输入，以更好地建模输入序列中的相关性和依赖关系。

通过采用递归卷积神经网络结构和门控机制，我们能够在处理可变长度序列时灵活地学习句子的结构，并生成固定长度的向量表示作为网络的输出。这种方法在自然语言处理任务中具有潜力，特别是在文本分类、情感分析和句子生成等领域。

让 x = (x1; x2; · · · ; xT) 为输入序列，其中 xt 为 d 维向量。提出的门控递归卷积神经网络（grConv）由四个权重矩阵 Wl、Wr、Gl 和 Gr 组成。在每个递归级别 t ∈ [1, T-1]，第 j 个隐藏单元 h(jt) 的激活值计算如下：

$$h_j^{(t)} = \omega_c \tilde{h}_j^{(t)} + \omega_l h_{j-1}^{(t-1)} + \omega_r h_j^{(t-1)},$$

where ω_c, ω_l and ω_r are the values of a gater that sum to 1. The hidden unit is initialized as

$$h_j^{(0)} = \mathbf{U}\mathbf{x}_j,$$

门控单元的值之和为1。隐藏单元的初始值为

U将输入投影到隐藏空间中。（U在这里指的是权重矩阵或变换矩阵，用于将输入映射到隐藏层的表示空间）

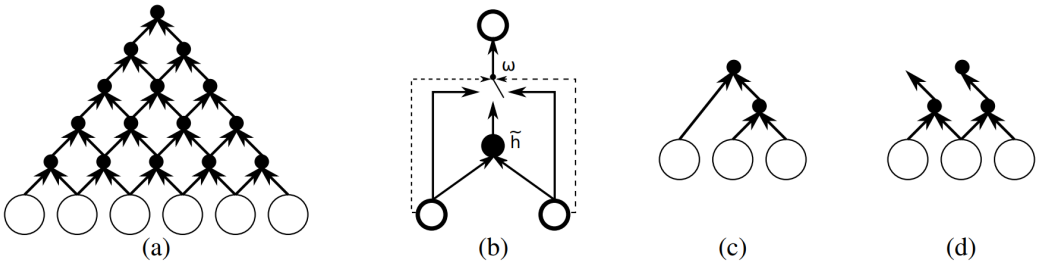


图2：(a) 递归卷积神经网络的图形示意图；(b) 递归卷积神经网络中提出的门控单元的图形示意图；(c-d) 可能通过提出的门控单元学习到的示例结构。

The new activation $\tilde{h}_j^{(t)}$ is computed as usual:

$$\tilde{h}_j^{(t)} = \phi\left(\mathbf{W}^l h_{j-1}^{(t)} + \mathbf{W}^r h_j^{(t)}\right),$$

where ϕ is an element-wise nonlinearity.

The gating coefficients ω's are computed by

$$\begin{bmatrix} \omega_c \\ \omega_l \\ \omega_r \end{bmatrix} = \frac{1}{Z}\exp\left(\mathbf{G}^l h_{j-1}^{(t)} + \mathbf{G}^r h_j^{(t)}\right),$$

where $\mathbf{G}^l, \mathbf{G}^r \in \mathbb{R}^{3\times d}$ and

$$Z = \sum_{k=1}^{3}\left[\exp\left(\mathbf{G}^l h_{j-1}^{(t)} + \mathbf{G}^r h_j^{(t)}\right)\right]_k.$$

根据这个激活函数，我们可以将递归级别 t 上单个节点的激活看作是从左右子节点计算得到的新激活、来自左子节点的激活或来自右子节点的激活之间的选择。这种选择使得递归卷积的整体结构能够根据输入样本自适应地改变。请参见图2(b)进行说明。

在这个意义上，我们甚至可以将提出的 grConv 视为一种无监督的解析（parsing）方法。如果我们考虑门控单元做出硬决策的情况，即 ! 遵循一种 1-of-K 编码，我们可以很容易地看出网络会根据输入自适应地形成一种树状结构（参见图2(c-d)）。然而，对于这个模型学习到的结构的进一步研究将留待未来的研究进行。

3 纯神经机器翻译

3.1 编码器-解码器方法

> 注：机器翻译是将一个语言的文本转换为另一个语言的文本的任务。编码器–解码器方法是一种常用的神经机器翻译框架。
>
> 在编码器–解码器方法中，输入句子首先通过编码器进行编码，得到一个固定维度的表示，也称为上下文向量或编码向量。编码器通常由循环神经网络（如长短时记忆网络或门控循环单元）组成，它可以逐个词或子词地处理输入句子，并将其转换为连续的表示。
>
> 接下来，解码器使用编码器的上下文向量作为输入，并逐个生成目标语言的翻译结果。解码器也通常由循环神经网络组成，它根据先前的生成结果和上下文向量，逐步生成输出序列。
>
> 编码器–解码器方法通过神经网络的端到端训练来学习源语言和目标语言之间的映射关系。训练过程中，模型被最小化目标语言与预测语言之间的差异，以优化翻译性能。
>
> 这种纯神经机器翻译方法具有灵活性和可扩展性，已经在机器翻译领域取得了显著的进展。它允许端到端的训练和生成，不需要手工设计的特征或规则，能够处理复杂的句子结构和上下文依赖关系。

从机器学习的角度来看，翻译任务可以理解为学习给定源语言句子 e，目标语言句子（翻译）f 的条件分布 p(f|e)。一旦模型学习到了这个条件分布，就可以使用该模型直接从源语言句子生成目标语言句子，可以通过实际抽样或使用（近似的）搜索算法来找到分布的最大值。

近年来的一些论文提出使用神经网络直接从双语平行语料中学习条件分布（Kalchbrenner和Blunsom，2013；Cho等，2014；Sutskever等，2014）。例如，（Kalchbrenner和Blunsom，2013）的作者提出了一种方法，其中涉及使用卷积n-gram模型从源语言句子中提取固定长度的向量，然后使用带有逆卷积n-gram模型和RNN的解码器进行解码。在（Sutskever等，2014）中，使用带有LSTM单元的RNN对源语言句子进行编码，并从最后一个隐藏状态开始解码目标语言句子。同样，（Cho等，2014）的作者提出使用RNN对一对源语言和目标语言短语进行编码和解码。

所有这些最近的工作的核心是编码器-解码器架构（参见图3）。编码器处理一个可变长度的输入（源语句），并构建一个固定长度的向量表示（在图3中表示为z）。在编码表示的条件下，解码器生成一个可变长度的序列（目标语句）。

在（Sutskever等，2014）之前，这种编码器-解码器方法主要作为现有统计机器翻译（SMT）系统的一部分使用。在（Kalchbrenner和Blunsom，2013）中，这种方法被用来重新排列SMT系统生成的n-best列表，而（Cho等，2014）的作者使用这种方法为现有的短语表提供了额外的分数。

在本文中，我们专注于分析直接翻译性能，如（Sutskever等，2014），使用两种模型配置。在这两种模型中，我们使用具有门控隐藏单元的RNN（Cho等，2014），因为这是唯一不需要确定目标长度的非平凡方式之一。第一个模型将使用与（Cho等，2014）中的编码器相同的具有门控隐藏单元的RNN，而第二个模型将使用提出的门控递归卷积神经网络（grConv）。我们的目标是理解编码器-解码器方法对翻译性能的归纳偏差，以BLEU作为衡量标准。

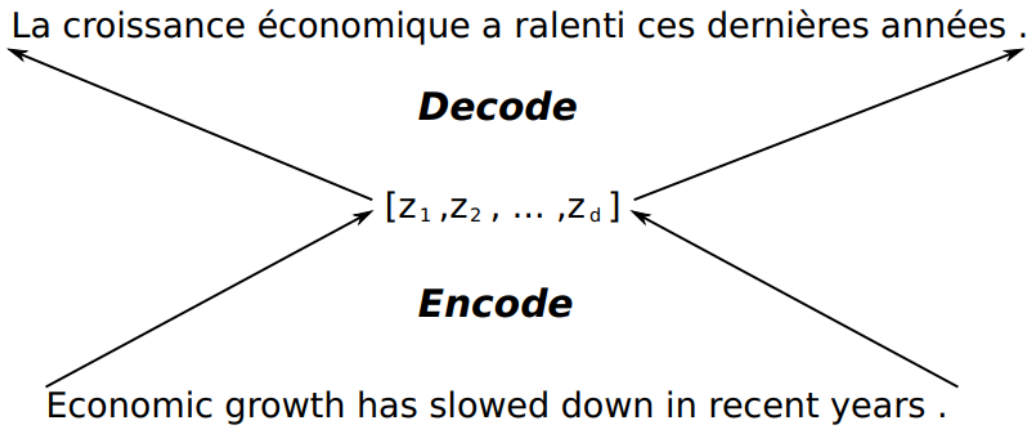


图3：编码器-解码器架构

4 实验设置

4.1 数据集

我们在英法翻译任务上评估编码器-解码器模型。我们使用了一个双语平行语料库，该语料库是根据（Axelrod等，2011）中的方法从Europarl（6100万词）、新闻评论（550万词）、联合国（4.21亿词）以及两个网络爬取的语料库（分别为9000万词和7800万词）组合而成的，共包含3.48亿个句对。我们没有使用单独的单语数据。神经机器翻译模型的性能是在news-test2012、news-test2013和news-test2014数据集上进行评估的，每个数据集包含3000个句子。在与SMT系统进行比较时，我们将news-test2012和news-test2013作为SMT系统的调优开发集，将news-test2014作为测试集。

在准备好的平行语料库中，出于计算效率的考虑，我们只使用英语和法语句子长度都不超过30个词的句对来训练神经网络。此外，我们仅使用英语和法语中最常见的3万个词。

```
All the data can be downloaded from http:
//www-lium.univ-lemans.fr/~schwenk/cslm_
joint_paper/
```

所有其他罕见的词被视为未知词，并映射为一个特殊的标记（[UNK]）4.2 模型

我们训练了两个模型：循环神经网络编码器-解码器（RNNenc）（Cho等，2014）和新提出的门控递归卷积神经网络（grConv）。请注意，这两个模型都使用具有门控隐藏单元的RNN作为解码器（参见第2.1节）。

我们使用小批量随机梯度下降和AdaDelta（Zeiler，2012）来训练这两个模型。在RNNenc的情况下，我们将方阵权重矩阵（转移矩阵）初始化为一个谱半径为1的正交矩阵，在grConv的情况下设置为0.4。在RNNenc和grConv中，分别使用tanh和整流线性函数（max(0, x)）作为逐元素的非线性函数。grConv有2000个隐藏神经元，而RNNenc有1000个隐藏神经元。两种情况下的词嵌入维度都是620维。

这两个模型分别训练了约110小时，相当于grConv和RNNenc分别进行了296,144次更新和846,322次更新。4.2.1 使用束搜索进行翻译

我们使用基本形式的束搜索来找到最大化特定模型（在本例中为RNNenc或grConv）给出的条件概率的翻译。在解码器的每个时间步，我们保留具有最高对数概率的s个翻译候选项，其中s = 10是束宽度。在束搜索过程中，我们排除包含未知词的任何假设。对于每个被选为最高得分候选项的序列终止符，束宽度减少一，直到束宽度达到零。

在RNN中使用束搜索（以近似的方式）找到具有最大对数概率的序列，在（Graves，2012）和（Boulanger-Lewandowski等，2013）中已经被提出并成功使用。最近，（Sutskever等，2014）的作者发现这种方法在基于LSTM单元的纯神经机器翻译中非常有效。

	Model	Development	Test
All	RNNenc	13.15	13.92
	grConv	9.97	9.97
	Moses	30.64	33.30
	Moses+RNNenc*	31.48	34.64
	Moses+LSTM°	32	35.65
No UNK	RNNenc	21.01	23.45
	grConv	17.19	18.22
	Moses	32.77	35.63

(a) All Lengths

	Model	Development	Test
All	RNNenc	19.12	20.99
	grConv	16.60	17.50
	Moses	28.92	32.00
No UNK	RNNenc	24.73	27.03
	grConv	21.74	22.94
	Moses	32.20	35.40

(b) 10–20 Words

表1：在开发集和测试集上计算的BLEU分数。前三行显示了所有句子的分数，后三行显示了不包含未知词的句子的分数。 (*) 是在（Cho等，2014）中使用RNNenc对短语表中的短语对进行评分得到的结果。(◦) 是在（Sutskever等，2014）中使用带有LSTM单元的编码器-解码器重新对由Moses生成的n-best列表进行排序得到的结果。

当我们使用束搜索来找到k个最佳翻译时，我们不使用通常的对数概率，而是使用相对于翻译长度进行归一化的概率。这样可以防止RNN解码器偏向较短的翻译，这种行为在之前的研究中已经观察到，例如（Graves，2013）。

5 结果与分析

5.1 定量分析

在本文中，我们对神经机器翻译模型的性能特性感兴趣。具体而言，我们关注翻译质量与源语句和/或目标语句长度以及每个源语句/目标语句中未知词数量之间的关系。

首先，我们观察BLEU分数（反映翻译性能）随句子长度的变化（参见图4(a)-(b)）。显然，短句子上的两个模型表现相对良好，但随着句子长度的增加，性能显著下降。

我们在图4(c)中观察到与未知词数量类似的趋势。符合预期的是，随着未知词数量的增加，性能迅速下降。这表明，在未来增加神经机器翻译系统使用的词汇量将是一个重要的挑战。尽管我们只展示了RNNenc的结果，但我们观察到grConv也表现出类似的行为。

在表1(a)中，我们呈现了使用这两个模型以及基线基于短语的统计机器翻译系统获得的翻译性能。明显地，神经机器翻译模型在整体上优于基线系统。然而，在没有未知词的句子上，神经机器翻译模型的性能仍然相对较低。

基于短语的统计机器翻译系统仍然表现出优于纯神经机器翻译系统的性能，但我们可以看到在某些条件下（源语句和参考语句中没有未知词），差距显著减小。此外，如果我们只考虑短句子（每个句子10-20个词），差距进一步减小（参见表1(b)）。

此外，可以将神经机器翻译模型与现有的基于短语的系统结合使用，最近在（Cho等，2014；Sutskever等，2014）中发现这种组合可以提高整体翻译性能（参见表1(a)）。

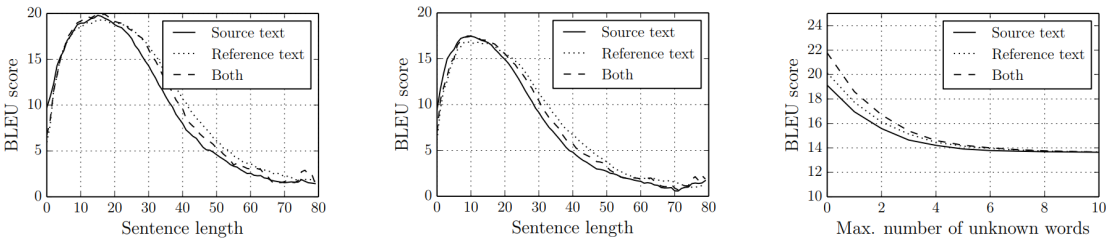
这种分析表明，当前的神经翻译方法在处理长句子时存在缺点。最明显的解释假设是，固定长度的向量表示无法足够编码具有复杂结构和含义的长句子。为了编码可变长度的序列，神经网络可能会“牺牲”输入句子中的一些重要主题以记住其他主题。

这与传统的基于短语的机器翻译系统（Koehn等，2003）形成鲜明对比。正如我们从图5中可以看到的那样，基于相同数据集（使用额外的单语数据进行语言模型训练）的传统系统在较长句子上得到更高的BLEU分数。

实际上，如果限制源语句和目标语句的长度，在某些条件下，纯神经机器翻译系统的性能可能会超过基于短语的统计机器翻译系统，尤其是在短句子上。这表明，纯神经机器翻译系统在处理长句子时存在挑战，并且仍有改进的空间。

如果限制源语句和参考翻译的长度在10到20个词之间，并且仅使用没有未知词的句子，对于RNNenc和Moses，测试集上的BLEU分数分别为27.81和33.08。

需要注意的是，即使我们使用长达50个词的句子来训练这些模型，我们观察到了类似的趋势。



(a) RNNenc (b) grConv (c) RNNenc

图4：(a) RNNenc和(b) grConv在给定长度的句子上达到的BLEU分数。通过取窗口大小为10来平滑绘图。(c) RNN模型在具有少于给定数量未知词的句子上达到的BLEU分数。

5.2 质量分析

尽管BLEU分数被用作评估机器翻译系统性能的事实上的标准度量标准，但它并不是完美的度量标准（参见，例如，(Song等，2013；Liu等，2011)）。因此，在这里我们展示了由RNNenc和grConv两个模型生成的一些实际翻译结果。

在表2(a)-(b)中，我们展示了从开发集和测试集中随机选择的一些句子的翻译结果。我们选择了那些没有未知词的句子。(a)列出了长句子（超过30个词），(b)列出了短句子（少于10个词）。我们可以看到，尽管BLEU分数有所差异，所有三个模型（RNNenc、grConv和Moses）在翻译方面表现不错，特别是对于短句子。然而，当源句子很长时，我们注意到神经机器翻译模型的性能下降。

此外，我们在图6中展示了提出的门控递归卷积网络学习表示的结构类型。对于一个样本句子“Obama is the President of the United States”，我们展示了grConv编码器学到的解析结构和生成的翻译结果。该图表明，grConv通过首先将“of the United States”与“is the President of”合并，然后将其与“Obama is”和“.”组合起来，提取了句子的向量表示，这与我们的直觉非常一致。

尽管与RNN Encoder-Decoder相比，grConv的性能较低，但我们认为grConv自动学习语法结构的这一特性非常有趣，并且认为需要进一步的研究。6 结论与讨论

本文研究了一种最近引入的基于纯神经网络的机器翻译系统的特性。我们重点评估了一种编码器-解码器方法，该方法最近在(Kalchbrenner和Blunsom，2013；Cho等，2014；Sutskever等，2014)中提出，用于句子到句子的翻译任务。在众多可能的编码器-解码器模型中，我们特别选择了两个在编码器选择上有所不同的模型：(1) 带有门控隐藏单元的RNN和(2) 新提出的门控递归卷积神经网络。在用英语和法语句子对这两个模型进行训练后，我们使用BLEU分数分析了它们在句子长度和句子中是否存在未知/罕见词汇方面的性能。我们的分析揭示了神经机器翻译在句子长度方面的性能下降较为显著。然而，从定性上来看，我们发现这两个模型都能够很好地生成正确的翻译。

这些分析为基于纯神经网络的机器翻译提供了一些未来的研究方向。首先，重要的是找到一种方法来提高神经网络的训练规模，无论是在计算资源还是内存方面，以便可以使用更大的源语言和目标语言词汇表。特别是对于具有丰富形态学的语言，我们可能需要采用完全不同的方法来处理单词。

其次，需要进行更多的研究，以防止神经机器翻译系统在处理长句子时性能不佳。最后，我们需要探索不同的神经网络架构，特别是解码器。尽管RNN和grConv在作为编码器使用时在架构上存在根本性的差异，但两种模型都受到了句子长度的限制。这表明解码器中可能缺乏表示能力。需要进一步的调查和研究。

除了一般神经机器翻译系统的特性，我们观察到了提出的门控递归卷积神经网络(grConv)的一个有趣特性。我们发现grConv可以模拟输入句子的语法结构，而无需对语言的句法结构进行任何监督。我们认为这个特性使它适用于除了机器翻译之外的自然语言处理应用。