摘要

神经翻译机是最近提出的机器翻译方法。与传统的统计机器翻译不同，神经翻译机旨在构建可以共同调整的单个神经网络，从而使得翻译性能最好。最近提出的用于神经翻译机的模型通常基于编码器 - 解码器系列，并将源语句编码成解码器从其中产生翻译的固定长度向量。在本文中，我们推测使用固定长度向量是提高基本编码器 - 解码器架构性能的瓶颈，并提出通过允许模型自动（软）搜索部分源语句中与预测目标词相关的词汇，而不必明确地将这些部分转化为固定长度的向量。通过这种新的方法，我们实现了与英语到法语翻译任务，并使之具有与现有最新的基于词组的系统相当的翻译性能。

# 介绍：

最近提出的神经翻译机是一种新兴的机器翻译方法，由Kalchbrenner和Blunsom（2013），Sutskever（2014）和Cho（2014B）等人。不同于传统的基于短语的翻译系统（参见，例如Koehn等人，2003），其由许多翻译子组件有机地组合成。神经翻译机尝试构建和训练单个大型神经网络，读取句子并输出正确的翻译。

大多数提出的神经翻译机模型属于编码器 - 解码器（Sutskever等人，2014; Cho等人，2014a），具有用于每种语言的编码器和解码器，或者涉及一种特定语言的编码器，它将被应用于每一个句子，它的输出也将被进行对比（Hermann and Blunsom，2014）。神经网络编码器读取源语句，并将之编码成为一个固定长度的向量。然后，解码器从编码向量输出转换。整个编码器 - 解码器系统由编码器和用于一对语言的解码器组成。整个系统将合起来一起参与神经网络的训练，以最大限度地提升源语句正确翻译的概率。

这种编码器 - 解码器方法的潜在问题是神经网络需要能够将源语句的所有必要信息压缩为固定长度向量。这可能使神经网络难以应付长句子，特别是比训练语料库中句子更长的源语句。 Cho等人（2014b）表示，此系统确实表现出了随着输入句子的长度增加而编码器 - 解码器效果迅速恶化的现象。

为了解决这个问题，我们向编码器 - 解码器模型引入了一种扩展：共同翻译和对齐。每次提出的模型在翻译中产生一个单词，并在源语句中搜索一组位置，这些位置是与此单词相关的信息最集中的位置。然后，该模型基于与这些源位置和所有先前生成的目标词相关联的上下文向量来预测目标词的翻译结果。

这种方法与基本的编码器 - 解码器最重要的区别在于它不会将整个输入句子编码成单个固定长度的向量。相反，它将输入句子编码为一系列向量，并在解码翻译时自适应地选择这些向量的子集。这释放了神经翻译模型的限制，不必将源语句的所有信息压缩成固定长度的向量，而不管其长度如何。我们将展示这将允许模型能够更好地处理长句。

在本文中，我们表明，本文提出的扩展方法可以显着提高基本编码器 - 解码器方法的翻译性能。可以观察到，任何长度的句子在本模型中的翻译结果改善情况很明显，尤其是较长的句子的结果改善尤为明显。在英语到法语翻译的任务中，本文提出的方法通过单一模型实现了与传统的基于短语的系统相当或接近的翻译性能。此外，定性分析显示，所提出的模型在源语句与相应的目标句之间找到语言上合理的（软）对齐。

# 背景：神经翻译机

从概率的角度来看，翻译相当于找到一个目标句子y，它使给定句子x的条件概率最大化，即。在神经翻译机中，我们拟合一个参数化的模型，以使用并行训练语料库来最大化一对句子的条件概率。一旦条件分布被翻译模型学习，给定一个源语句，可以通过搜索最大化条件概率的句子来生成相应的翻译。

最近，一些论文提出使用神经网络来直接学习这种条件概率分布（参见，例如，Kalchbrenner和Blunsom，2013; Cho等，2014a; Sutskever等，2014; Cho等，2014b ; Forcada和Neco，1997）。这种神经翻译机方法通常由两个组件组成，第一个组件对源语句x进行编码，第二个组件将之解码成目标句子y。例如，（Cho等人，2014a）和（Sutskever等人，2014）使用两个循环神经网络（RNN）来将可变长的源语句编码为固定长度的向量并将此向量解码为一个可变长度的目标句子。

尽管这是一种相当新的方法，神经翻译机已经显示出有期望的结果。 Sutskever等（2014）报道说，在英语到法语翻译任务上，基于具有长期短期记忆单元（LSTM）的RNN神经翻译机具有接近于常规的基于短语的机器翻译系统的最先进的性能。为现有的翻译系统添加神经元素，例如，为短语表中的一对一对短语进行估分，或者重新排列候选翻译，已经允许神经翻译机超过以前传统的极其翻译系统的最先进的表现水平。

## RNN 编码器 - 解码器

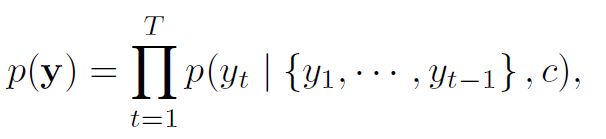
在这里，我们简要介绍由Cho和Sutskever等人提出的底层框架，称为RNN 编码器 - 解码器。以此为基础，我们构建了一种学习同步对齐和翻译的新型架构。

在这个编码器 - 解码器架构中，编码器将会以一组向量的形式的形式读取输入的句子，并将之转化为一个向量c。最常用的一种方法是

和

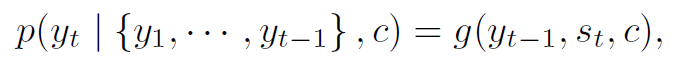
其中，是在时间t中的一个隐藏状态， 是从隐藏状态序列中生成的向量，是一些非线性函数。例如，Sutskever 等人使用方式便是一个LSTM作为。

通常，我们训练解码器，并通过给出上下文向量c和所有先前预测的单词来预测下一个单词。 换句话说，解码器将联合概率分解为有序条件,并以此来定义翻译y的概率：



---------(2)

其中，。利用RNN，每一个条件概率都形成如下的模型



---------(3)

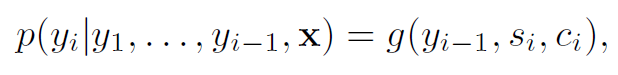
其中是非线性的、潜在的多层次函数，其输出可能是的概率，是RNN的隐藏状态。应当注意，可以使用其他诸如RNN和去卷积神经网络的混合架构（Kalchbrenner和Blunsom，2013）。

# 学习对齐和翻译

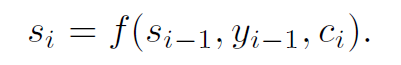
在本节中，我们提出了一种用于神经翻译机的新颖架构。新架构包括作为编码器的双向RNN（第3.2节）和解码翻译期间模拟搜索源语句的解码器（第3.1节）。

## 解码器：一般说明

在一个新的模型架构中，我们定义了每个等式(2)中的条件概率：

--------(4)

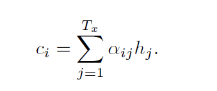
其中是RNN对应于时间i的隐藏状态，被表示为



应当注意，与现有的编码器 - 解码器方法（参见等式(2)）不同，这里的概率是针对每个目标词的不同上下文向量。

上下文向量取决于编码器映射输入句子的注释序列。每个注释包含与整个输入序列有关的信息，其中重点关注输入序列的第i个单词周围的部分。我们将详细说明如何在下一节中计算注释。

然后，上下文向量可以由注释的权重和得到：



--------(5)

其中每个注释的权重可由如下方式得到：

 --------(6)

其中， 是一个对齐模型，它用于评估位置j周围的输入和位置i处的输出是否匹配。该分数基于RNN的隐藏状态（恰好在输出之前，方程(4)）和输入的源语句的第j个注释。

我们将对齐模型a参数化为一种前馈神经网络，并且与我们提出的系统所有其他组件一起训练。注意，与传统的机器翻译不同，对齐不被认为是一种潜在变量。相反，对齐模型直接计算软对齐，这允许通过反向传播代价函数的改变。该梯度可以用于一起训练对齐模型以及整个翻译模型。

我们可以理解这种方法，它将所有注释的加权和作为计算期望注释，其中期望是对于可能的对齐方式。令是目标词与源词或从后者翻译过来的概率。那么，第i个上下文向量是所有具有概率的注释的预期。

概率或其相关联的能量反映了在决定下一个状态和生成时，关于先前隐藏状态的注释的重要性。直观地说，这实现了解码器中的关注机制。解码器要关注源语句的一部分来确定输出。通过让解码器有一个关注机制，我们解决了编码器不必将源语句中的所有信息编码成固定长度向量的负担。利用这种新的方法，信息可以遍布整个注释序列，这可以由解码器相应地选择性地进行检索。

## 编码器：用于提示序列的双向RNN

一般的RNN，如公式 (1)，以从第一符号到最后一个的顺序读取输入序列x。然而，在提出的方案中，我们希望每个单词的注释不仅能够总结前面的单词，而且还要总结之后的单词。因此，我们建议使用最近在语音识别中成功奏效的双向RNN（BiRNN）。

BiRNN由前向和后向的RNN组成。前向RNN读取输入序列（从到），并计算前向隐藏状态序列。反向RNN以相反的顺序（从到）读取序列，产生后向隐藏状态序列。

最终，我们将前向隐藏状态和后向隐藏状态联合起来，就得到了对于每一个单词的注释。通过这样的方法，每一个都包含了前面单词和后面单词的信息。由于RNN能够更好地表示输入的缘故，注释将集中关注于在周围的词。稍后，该注解序列由解码器和对准模型用于计算上下文向量（等式(5)-(6)）。

# 实验设置

我们将对本文提出的英语到法语翻译方法进行评估。我们使用ACL WMT '14.3提供的双语平行语料库作为比较，我们还报告了最近由Cho等人提出的RNN 编码器 - 解码器的性能。（2014A）。我们会对这两个模型使用相同的训练程序和相同的数据集，从而进行对比实验。

## 数据集

WMT '14包含以下英文-法文并行语料库：Europarl（61M字），新闻评论（5.5M），UN（421M）和两个抓取到的语料库，分别为90M和272.5M，共计850M。按照Cho等人的评估方法，我们使用Axelrod的数据选择方法将组合语料库的大小减小到348M字。我们不使用除了所提及的并行语料库之外的任何单语数据，尽管可能使用更大的单语语料库来预先训练编码器。

我们联合使用新闻测试-2012年和新闻测试-2013作为开发（验证）数据集合，并使用来自WMT '14的测试集（news-test-2014）的模型进行评估，其中包含3003个不在训练数据中的句子。

在标记之后，我们使用每种语言30,000个最常用词的候选列表来训练我们的模型。任何不包括在候选名单中的单词都映射到一个特殊的令牌([UNK])。我们不对数据应用任何其他特殊的预处理，如小写或词干。

我们训练两种模型。第一个是RNN编码器-解码器（RNNencdec，由Cho等提出），另一个是本文提到的RNN搜索的模型。我们训练每个模型两次：首先使用长度达30个字的句子（RNNencdec-30，RNNsearch-30），然后使用长度最多为50个字的句子（RNNencdec-50，RNNsearch-50）。

RNNencdec的编码器和解码器各有1000个隐藏单元。RNNsearch的编码器由前向和后向循环神经网络（RNN）组成，每个神经网络具有1000个隐藏单元。其解码器有1000个隐藏单元。在这两种情况下，我们使用具有单一maxout（Goodfellow等人，2013）隐层的多层网络来计算每个目标词的条件概率（Pascanu等，2014）。

我们与Adadelta（Zeiler，2012）一样，都是使用小型随机梯度下降算法（SGD）来训练每个模型。每个梯度下降的更新方向使用80个句子的小批量计算。我们训练了每个模型大约5天。

一旦模型被训练，我们使用波束搜索来找到近似最大化条件概率的翻译（参见例如，Graves，2012; Boulanger-Lewandowski等人，2013）。 Sutskever等人使用这种方法从他们的神经翻译机模型生成翻译。

有关实验中使用的模型和培训程序的架构的更多详细信息，请参见附录A和B.

# 结果

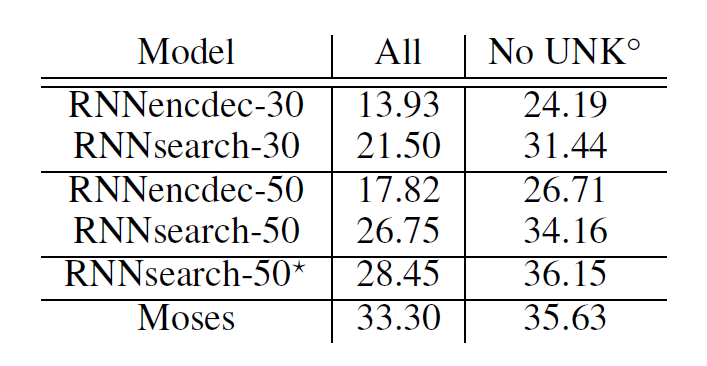
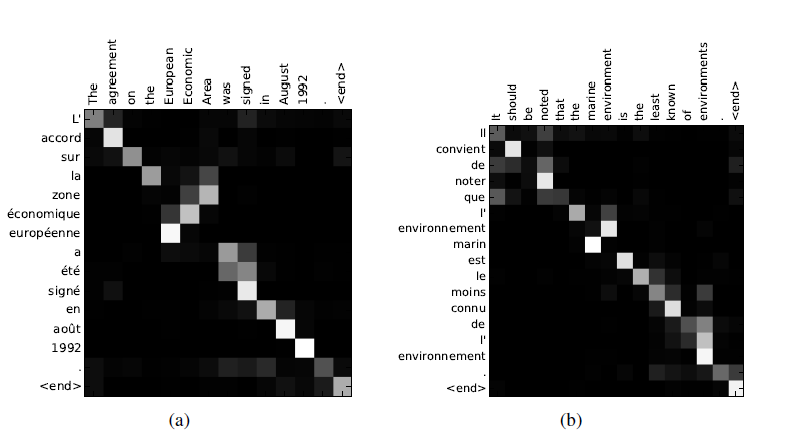


表1：在测试集上计算的训练模型的BLEU得分。

第二和第三列分别显示了所有句子的分数，包括在本身和参考翻译中没有任何未知单词的句子。请注意RNNsearch-50所需的训练时间得更长，直到模型更新与验证停止改善为止。

当只评估没有未知单词的句子（最后一列）时，我们不允许模型生成[UNK]令牌。



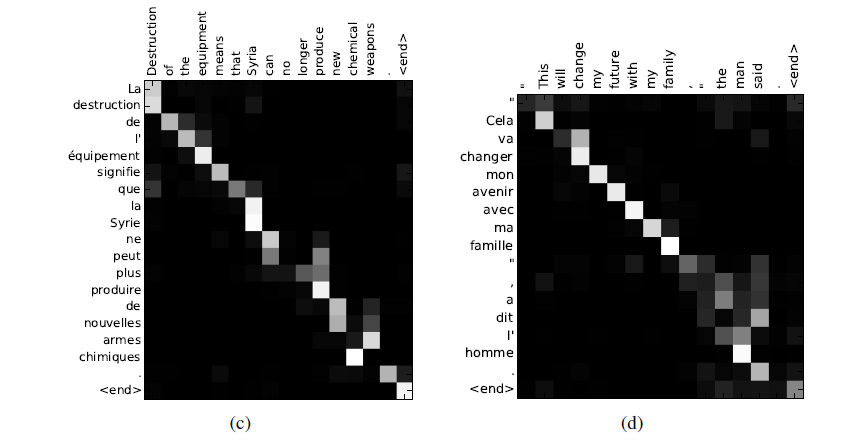


图3：RNNsearch-50中四个样品对齐的可视化图。每个图的x轴和y轴分别对应于源句（英文）和生成的翻译（法语）中的单词。每个像素以灰阶(0：黑色，1：白色)显示第i个目标字的第j个源字注释的权重(i = 1)（见式(6)）。

(a)任意句子。(b-d)三个随机选择的样本，没有任何未知字，长度在10到20个字之间的测试集。

在表1中，我们列出了以BLEU评分测量的翻译表现。从表中可以看出，在所有情况下，本文提出的RNNsearch都胜过传统的RNNencdec。更重要的是，当仅考虑由已知单词组成的句子时，RNNsearch的表现与传统的基于短语的翻译系统（Moses）的表现一样高。考虑到Moses除了用于训练RNNsearch和RNNencdec的并行语料库外，还使用了一个单独的单语语料库（418M单词），我们设计的这个系统达成了一个重要的成就。

我们提出这个方法的动机之一是在基本编码器 - 解码器方法中使用固定长度的上下文向量。我们推测，这种限制可能使基本的编码器 - 解码器方法在较长的句子翻译时表现较差。在图 2，我们看到RNNencdec的表现随着句子的长度的增加而急剧下降。另一方面，RNNsearch-30和RNNsearch-50对句子的长度更加鲁棒。特别是RNNsearch-50，即使长度为50或更长的句子也不会有效果恶化的发生。与RNNencdec-50相比，RNNsearch-30甚至超过了RNNencdec-50（参见表1）这一事实，进一步证实了所提出的模型与基本编码器 - 解码器的优越性。

## 定性分析

### 对齐

本文所提出的方法提供了一种直观的方法来检查生成的翻译中的单词与源语句中的单词之间的（软）对齐。这是通过将公式的注释权重可视化来完成的（参见公式(6)）。每个图中矩阵的每一行表示与注释相关联的权重。从这里我们看到源句中的哪些位置在生成目标词时被认为更重要。

从图3中可以看出，英语和法语之间词汇的排列很大程度上是单调的。我们看到每个矩阵对角线的强权重。

然而，我们也观察到一些非不一般的，非单调的对齐。形容词和名词通常在法语和英语之间有不同的顺序，我们在图3(a)中看到一个例子。

从这个数字，我们看到这个模型正确地将一个[European Economic Area]翻译成[zoneéconomique européen]。RNNsearch能够正确地将[zone]与[Area]对齐，跳过两个单词（[European]和[Economic]），然后一次查看一个字，以完成整个短语[zoneéconomique européen]。

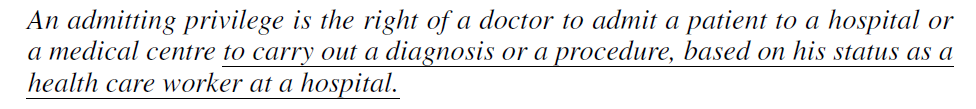
从图3(d)中可以看出，与硬对齐相反的软对准的优势非常明显。诸如，被翻译成[l'homme]的源短语[the man]，所有的硬对齐都将[l]和[man]映射到[homme]。但是这对翻译没有帮助，因为人们必须考虑以下这个词来决定是否应该翻译成[le]，[la]，[les]或[l]。

我们的软对齐则是通过让模型看看[the]和[man]，自然地解决了这个问题，在这个例子中，我们看到模型能够正确地将[the]转换成[l']。我们在图1的所有呈现的情况下观察到类似的行为。软对齐的另外一个好处是，它自然地处理不同长度的源和目标短语，而不需要将一些单词映射到或不在任何地方的反直觉的方式。

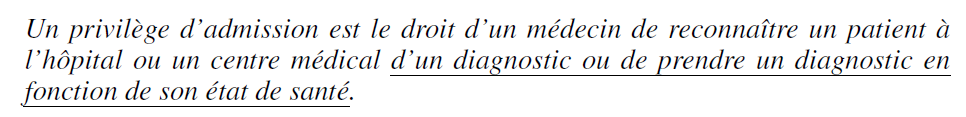
### 长句子

从图2中清楚可见。本文提出的模型（RNNsearch）在翻译长句中比传统模型（RNNencdec）好得多。这可能是由于RNNsearch不需要将长句子完美地编码成固定长度的向量，而只是精确地编码包围特定单词的输入句子的部分。

例如，从测试集中考虑这个源语句：

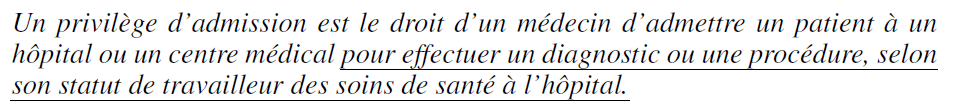


RNNencdec-50将此句翻译成：



RNNencdec-50正确地将源句翻译成[a medical center] 然而，从那里（下划线），它偏离了源语句的原始含义。例如，在[based on his status as a health care worker at a hospital]的来源句中，它取代了[enfonction de sonétat de santé]( based on his state of health)。

另一方面，RNNsearch-50生成了以下正确的翻译，保留了输入句的全部含义，没有忽视任何细节：



结合已经提出的定量结果，这些定性观察结果证实了我们的假设，即RNNsearch架构能够比标准的RNNencdec模型更可靠地对长句子进行翻译。