全新学习和推断机制提升seq2seq 模型的语法改错性能

——阅读报告

目录

[全新学习和推断机制提升seq2seq 模型的语法改错性能 1](#_Toc42635689)

[——阅读报告 1](#_Toc42635690)

[一、 任务概述 1](#_Toc42635691)

[二、 意义价值 2](#_Toc42635692)

[三、 原文文献综述 3](#_Toc42635693)

[四、 重点难点 4](#_Toc42635694)

[五、 模型设计 4](#_Toc42635695)

[六、 实验结果 7](#_Toc42635696)

[七、 论文结论分析 9](#_Toc42635697)

[八、 其他参考文献 9](#_Toc42635698)

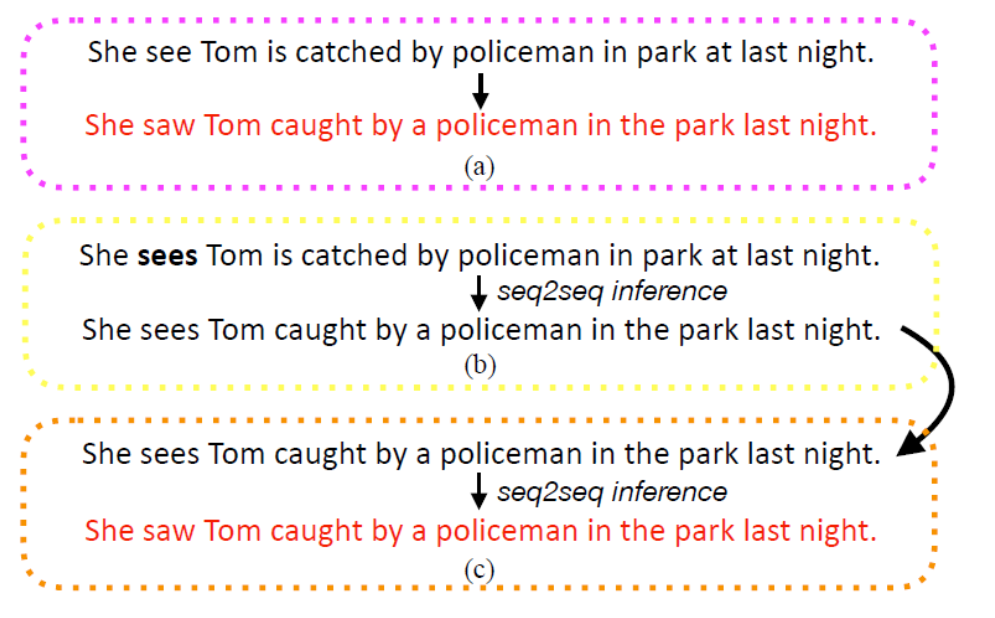
[九、 综合评价 9](#_Toc42635699)

[附录（全文翻译）： 21](#_Toc42635700)

# 任务概述

机器自动语法纠错是自然语言处理领域的一个经典研究问题，由于能够作为训练数据的句对语料非常有限，机器语法纠错能力长久以来始终无法达到理想的效果。随着相关数据集规模的逐渐增大以及深度学习技术的不断成熟，越来越多的学者开始研究利用seq2seq模型进行自动语法改错。但有限的改错句对数据规模仍然是模型训练效果的主要局限点，而且当训练数据中的句子存在多重语法错误时，单次seq2seq推断往往也无法很好的将其改正。

下图表示了对于一个用于语法改错任务的seq2seq模型，其基本的训练数据为一个由原始句子和正确句子所组成的改错句对，如图1(a)所示，及其两大局限性1(b)和1(c)。



文章尝试提出一种流利度提升和推理机制来解决这些局限，并在公开的CoNLL-2014数据集和数据集JFLEG上取得了全世界最领先的效果。

# 意义价值

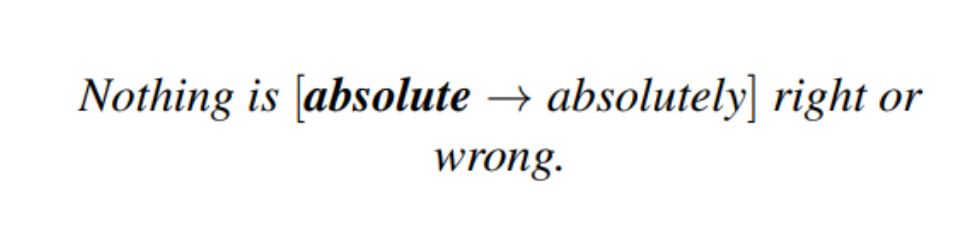
语法纠错（Grammatical Error Correction, GEC）是自然语言处理领域中的一个重要任务，

GEC任务要求检测一句话中是否有语法错误，并自动将检测出的语法错误进行纠正。人们可能对Word文档中检查和纠正单个词汇输入错误的拼写检查功能很熟悉，与拼写检查相比，语法检查要复杂得多，它需要考察句子中的逻辑关系，包括主谓搭配、介词使用、时态表达等多达28种（CONLL-2014评测任务中将错误类型共分为28种）的“全错误类型”。同时GEC在文本校对、外文学习辅助中都有重要的应用。

英语学习者通常所犯的语法错误分成两大类：

* 句法错误。句法错误主要是指破坏了语言内生的系统性规则，例如动词形式的变化规则中，情态动词所跟的动词必须使用动词原形；
* 语用错误。语用错误通常是指违反了语言的使用习惯，例如 depend on sth.，这里 on 如果错用成 in，并没有违反句法规则，但却不符合人们的语言使用习惯，也属于语法错误。

下面是一个语法纠错任务的示例，在这句话中，加粗的形容词absolute应该修改为副词absolutely。



如果能利用机器学习技术来自动评判并改正语法错误的话，可以在英文自动写作、作文打分等以人工智能为驱动的交互式英语学习应用上取得很大的进展，本文中的成果也成功在微软小英应用上得到加入。（<https://www.engkoo.com/>）

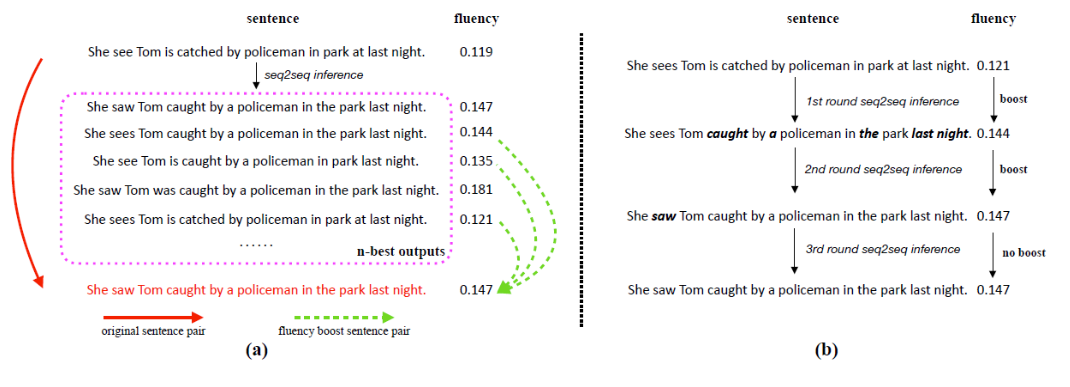
# 原文文献综述

文章所提出的全新的学习和推理机制**fluency boost learning and**

**Inference** 本质上讲可以理解为一种自然语言处理的数据增强技术。

流畅度提升学习（fluency boost learning）的核心原理就是在训练模型的过程中，让seq2seq模型生成出多个结果，然后将结果中流畅度不如目标端正确句子的生成句子和目标端正确句子配对，组成全新的流畅提升句对，作为下一轮训练的训练数据。如图a。

而流畅度提升推断（fluency boost inference）则是利用seq2seq模型对句子进行多轮修改，直到句子的流畅度不再提升为止。这种多轮修改的策略能够率先改掉句子的一部分语法错误，从而使句子的上下文更加清晰，有助于模型修改剩下的错误。如图b。



文章的主要贡献在下面总结：

* 文章提出了一个新奇的学习和推理机制来解决现在序列到序列模型在GEC上的局限。
* 文章提出并比较了多种新颖的流利度提升学习策略，探索神经网络GEC的学习方法论。
* 文章的方法被证明有效地提升神经网络seq2seq模型去在CoNLL-2014和JFLEG数据集上实现最先进的结果

# 重点难点

上文也提到利用seq2seq模型训练的两大缺陷为：

（1）改错句对的数量规模相当有限。因此，在训练数据并不充足的情况下，seq2seq模型的泛化能力就会受到影响，其导致的一个结果就是哪怕输入的句子稍稍变动一点，模型也可能会无法将其完全改正

（2）seq2seq模型很难通过单轮的推理完美改正一个带有很多语法错误的句子，因为一些句子中的错误可能使得文本变得奇怪，这会困惑模型去改正其他的错误。

文章重点利用利用新的流利度提升学习和推理机制解决了seq2seq模型中的这两个缺陷，使得序列到序列神经网络模型在GEC任务上表现的更好，并在多个数据集上取得了非常好的得分。

就语法错误改正任务而言，还有以下四个难点：

1. 错误类型多：在目前比较流行的 GEC 标注体系下 （Christopher Bryant et al），语法错误类型被划分成五十多种，如果按照语法教学体系，语法错误甚至可以被划分成上百种，在如此纷繁复杂的错误类型中，学习目标很难统一。
2. 多重句法功能：仅以介词为例，介词引导的语言成分既可以作定语，也可以作表语，还可以做状语，在不同的用法中上下文的语境也不尽相同，从而为语法错误的判断带来干扰。
3. 远距离依赖：以主谓一致为例，有时候用于判断谓语主谓一致性的主语距离谓语很远，远距离的主谓一致情况在训练语料中是稀疏的，而长距离信息的学习也恰好是机器学习的难点。
4. 噪音：在 GEC 领域的训练数据中，由于标注的质量和标注者的偏见等问题，天然存在着相当多的噪音，如何克服这些噪音以获得学习的稳定性也是难点之一。

这也是以后GEC任务的研究方向。

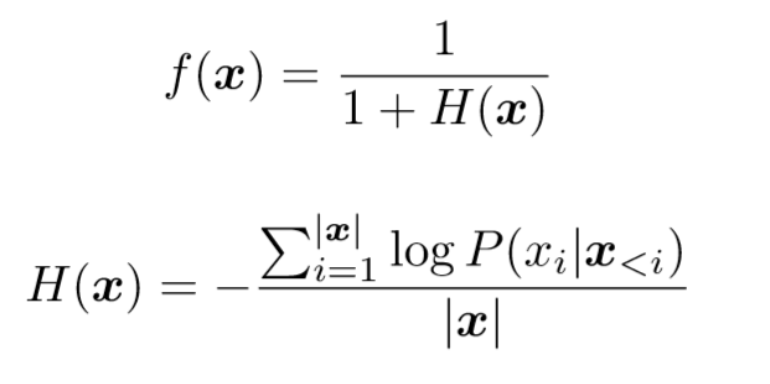
# 模型设计

**流利度提升学习**

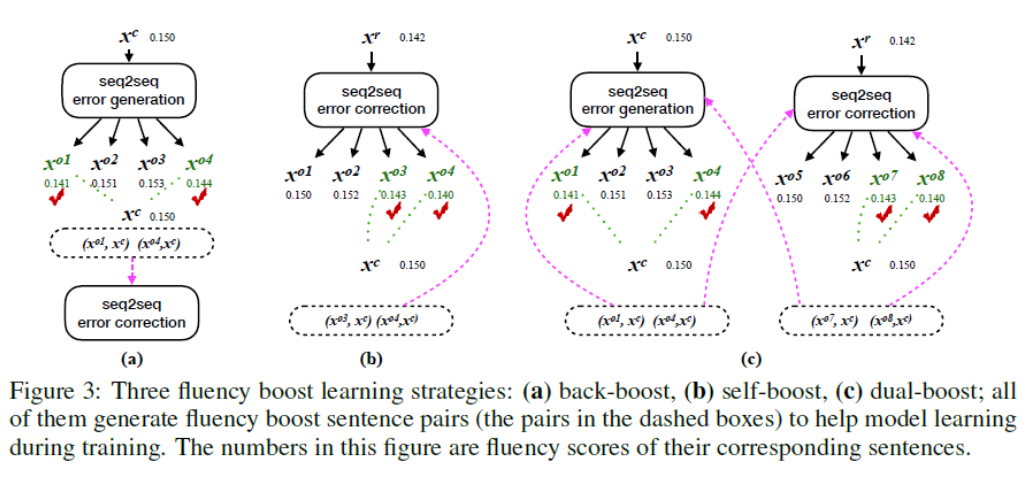
对于语法改错任务来说，一个合格的训练样本通常满足两个要求：(1)源端句子和目标端句子的语义应当是一致的，修改过的句子不能改变原意；(2)目标端句子的流畅度能够得到提升，这一点其实也是语法改错任务的终极目标。

而流畅度提升学习所生成出的流畅提升句对恰恰能够很好地满足上面的两个条件。首先，由于seq2seq模型是用改错句对作为训练数据训练得到的，所以流畅度提升学习所创造出的句对通常不会改变句子原意，在模型相对稳定后，n-best的输出结果中的句子与原句通常只有1到2个词的差别，很少会改变句子原意；其次，生成的流畅提升句对能够保证目标端的句子比源端句子有更高的流畅度。

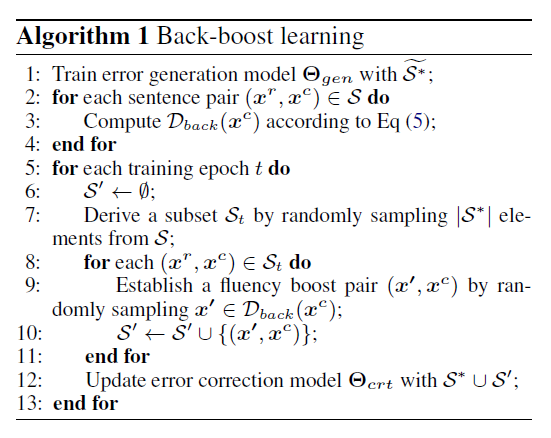
在文章中，用f(x)来定义句子的流畅度：（流利度代表这个句子被母语者写出的可能性，换句话说，如果一个句子很有可能写自一位母语者，就认为它有很高流利度。）



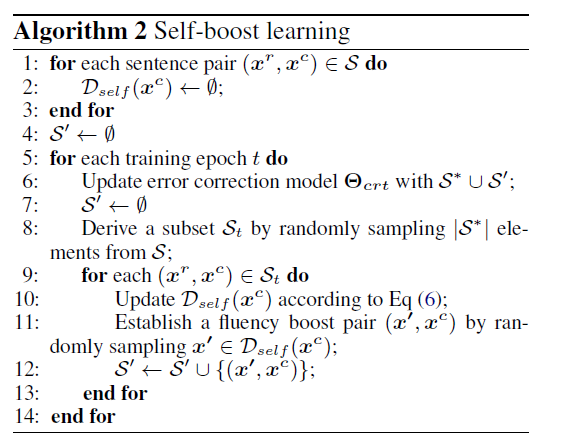
上式中P(xi|x<i)是xi的概率在给出x<i 的文本中，由语言模型计算得到。H(x)实际上是句子x的交叉熵，|x|是句子x的长度，它的范围为[0,+ ∞）。因此，f(x)的范围以为(0,1]



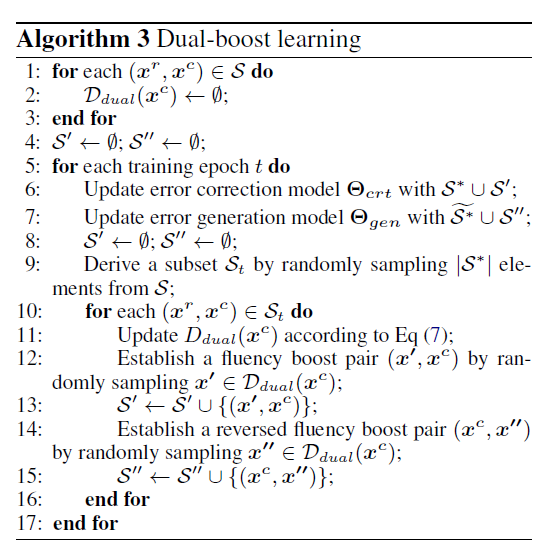
其实文章所用的本质上是数据增强的方法，通过数据增强（data augmentation）方法来扩充训练集的做法以前就有研究者提出过，在神经网络机器翻译领域比较有代表性的工作就是Sennrich等人在ACL 2016上提出的back-translation方法（图3a），利用反转训练数据的源端和目标端来训练一个专门用来生成错误句子的模型。而本文的方法则是通过改错模型自身生成的n-best来作为错误句子(图3b)。



为了区分这两种不同的方法，文章将利用back-translation来生成流畅提升句对的方法称之为back-boost learning，而将利用改错模型自身来生成流畅句对的方法称为self-boost learning。对于self-boost learning，在不同的训练阶段，模型的不断迭代更新会导致对同一句子前后生成出的n-best也会不同，所以模型生成出的含有错误的句子更加多样化。



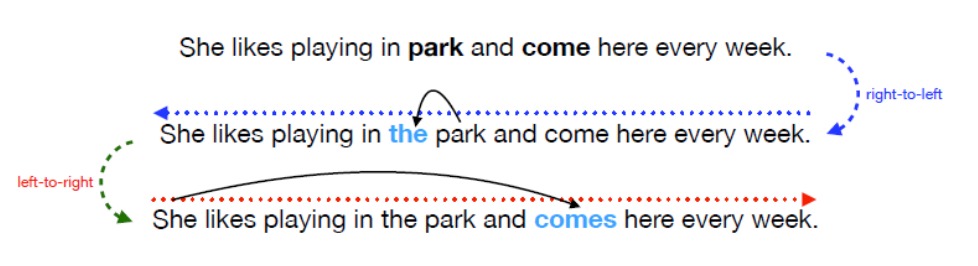
由于back-boost和self-boost是从完全不同的两个角度来生成流畅提升句对，这就意味着这两种方法能够互相补充、互相增强。因此，进一步的文章将两种方法结合在一起，产生了最终的dual-boost learning方法。Dual-boost learning能够让back-boost和self-boost learning各自生成出流畅提升句对（如图3c）。生成出的句对不仅可以帮助训练改错模型，让改错模型从更多的样例中知道如何去改正一个句子，反过来也可以帮助训练错误生成模型来生成更多样化的含有错误的句子。



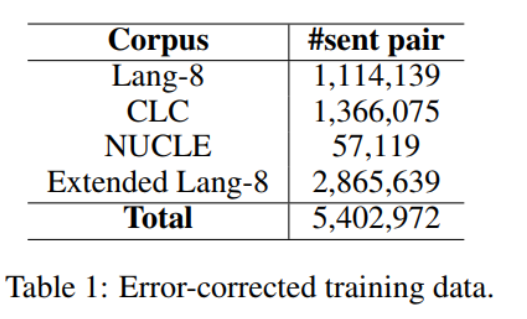
值得一提的是，流畅度提升学习的方法也可以应用在大量的正确文本上。因为一个正确的句子可以看成是一个源端和目标端相同的改错句对，将流畅度提升学习应用到正确文本可以帮助极大地扩充训练数据的规模以及内容多样性。这也是作者的核心创新思想所在。

**流畅度提升推断**

流畅度提升推断利用了语法改错这个任务的特殊性——输入输出本质上是相同的，因此我们可以将输出的结果句子作为输入进行再修改。在多轮seq2seq推断的基础上，文章进一步提出了一种更加有效的方法——往返修改。往返修改是指利用一个反向（右到左）解码器和一个正向（左到右）解码器交替地对一个句子进行修改。因为正向和反向解码器对于不同的错误有着各自的优势，往返修改能够让这两个模型能够充分发挥自己的优势。



例如上图中的句子，用反向模型能够很容易地把冠词错误改正，这是因为冠词错误的修正更加依赖于冠词位置右边的上下文。但是反向模型在检测主谓一致这种错误类型时会存在一些劣势，因为主致一致错误的改正往往需要依赖位于谓语动词左边的主语的人称和单复数。而这一类型的错误又可以被正向模型发现并改正。

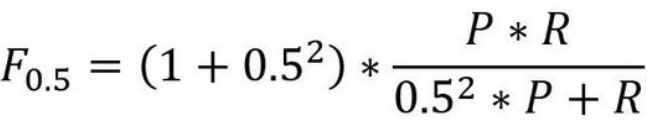


文章利用Lang-8、CLC、NUCLE这几个知名的语法改错数据集以及额外从Lang-8收集到的287万训练句对作为原始训练数据。我们的基本seq2seq模型是一个7层卷积seq2seq模型。

对于流畅度提升学习，作者从10-best outputs中来筛选构造流畅提升句对，并且使用英语维基百科的语料作为正确文本来生成流畅句对。在实际模型训练过程中，作者强制规定在每一轮训练迭代过程中，用于训练的流畅句对数量不超过原始训练句对的数量。对于流畅度提升推断，选择使用了往返修改的策略，在流畅度提升的前提下，用反向和正向的seq2seq模型交替对句子进行修改。此外，使用了5元组语言模型和编辑特征对于beam search选出的12-best句子进行重排序，从而选出最好的结果作为单轮推断的输出。

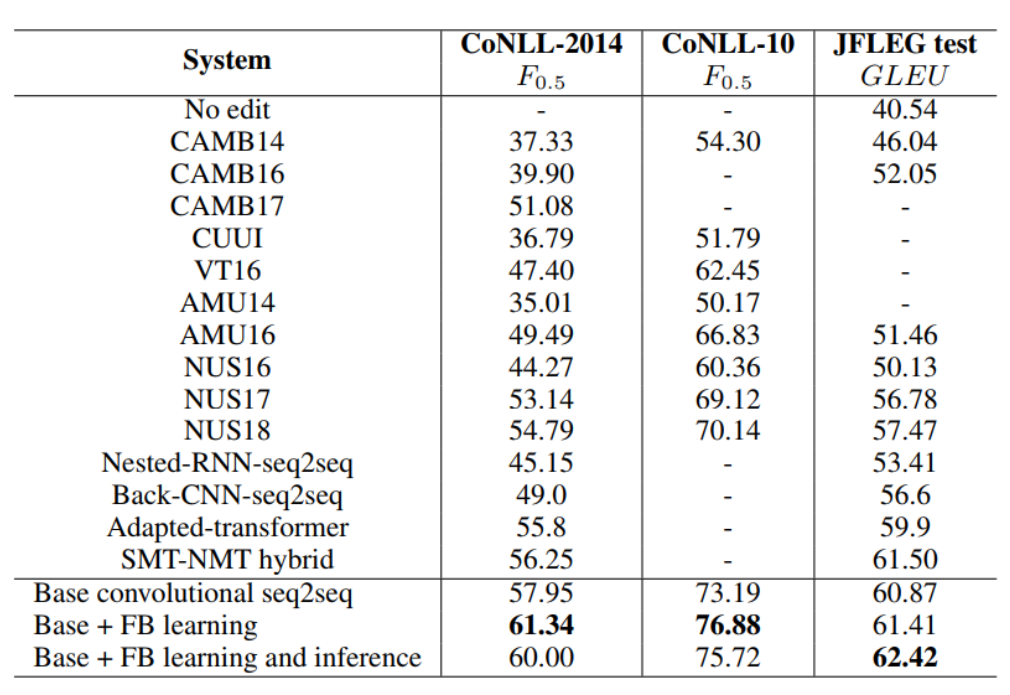
# 实验结果

GEC目前最常用的评估方式是基于Max-Match(M^2)的度量， Daniel Dahlmeier et al.，并综合准确率P和召回率R计算得到



在CoNLL-2014和JFLEG两个基准测试集上对系统进行了评价，分别选用了两个数据集官方的评测指标Max-match F\_0.5和GLEU。值得一提的是，CoNLL-2014数据集有两种不同的标注集。一种是原始标注集，由2名专业人员进行标注；而另一个标注集则是由Bryant & Ng (2015)后来对原始标注集的一个扩充，将标注数提到了10组。我们在这两个标注集下都进行了测试，分别区别这两种标注集，我们用CoNLL-2014代表原始2人标注集，CoNLL-10代表10人标注集。

我们将结合了流畅度提升学习和推断机制的seq2seq模型和一些知名的语法改错系统进行了效果对比。从下表中可以看出，在使用了更多的训练数据之后，我们的基础模型就已经超越了多数先前的系统。当在基础模型上加入流畅度提升学习机制后，模型在三个基准上都有了显著的提高，在CoNLL-2014和CoNLL-10基准上达到了61.34和76.88 F\_0.5，在JFLEG数据集上也达到了61.41的高分。当再加入流畅度提升推断机制以后，模型在JFLEG数据集上的得分提高到了62.42，但在CoNLL数据集上F\_0.5的得分出现了下降。



通过分析发现，相比于基础seq2seq模型，流畅度提升学习能够提升模型在准确率、召回率、F\_0.5、GLEU所有指标上的得分，有效地帮助模型更好地学习如何进行语法改错。流畅度提升推断能够显著提高召回率，但却会使准确率下降。由于F\_0.5这个指标对准确率的权重要远大于召回率，因此模型在F\_0.5的得分上出现了下降。而对于JFLEG数据集，加上流畅度提升推断后模型的GLEU得分能够从61.41提高到62.42，这也证实了流畅度提升推断能够更好地提升句子的流畅度。

在CoNLL-10和JFLEG两个基准测试集上按照人机对比的评测设定，结合了流畅度提升学习和推断机制的seq2seq模型的得分分别为74.84和62.42的得分，均首次超越了人类在这两个基准上的表现（CoNLL: 72.58, JFLEG: 62.37）。

# 论文结论分析

文章提出了一种新颖的流利度提升学习和推理机制，并克服了以前GEC模型的两大局

限。流利度提升学习充分利用了两张错误纠正方法和本地数据，大大改善了传统seq2seq模型的表现，而流利度提升推理则利用GEC的特性，通过多轮推理以逐步改善句子流利度。最终新的学习和推理机制使seq2seq模型能够达到在CoNLL-2014和JFLEG的基准数据集上达到最先进的成果。

# 其他参考文献

微软亚洲研究院.新闻中心.深度文章，葛涛,<https://www.msra.cn/zh-cn/news/features/fluency-boost-learning-and-inference-for-neural-grammatical-error-correction>

# 综合评价

随着深度学习技术在图像识别、语言识别等领域的突破并超越了人类，深度学习模型在自动语法改错数据集上的得分也超越了人类。但这并不意味着自动语法改错目前已经全面超越了人类。目前的改错模型主要是基于对句子的浅层分析，并不涉及到对语义级别的深层理解，这就使得在涉及到深层语义理解的错误修改时，机器的能力要远弱于人类专家的水平；而且，目前自动语法改错通常只涉及到单句级别的修改，并不会考虑段落或者篇章级上下文语境。因此，机器自动改错想要真正超越人类水平，还有较长的一段路要走。但对于大多数语法错误来说，机器自动语法改错的能力确实已经可以达到人类的水平了。

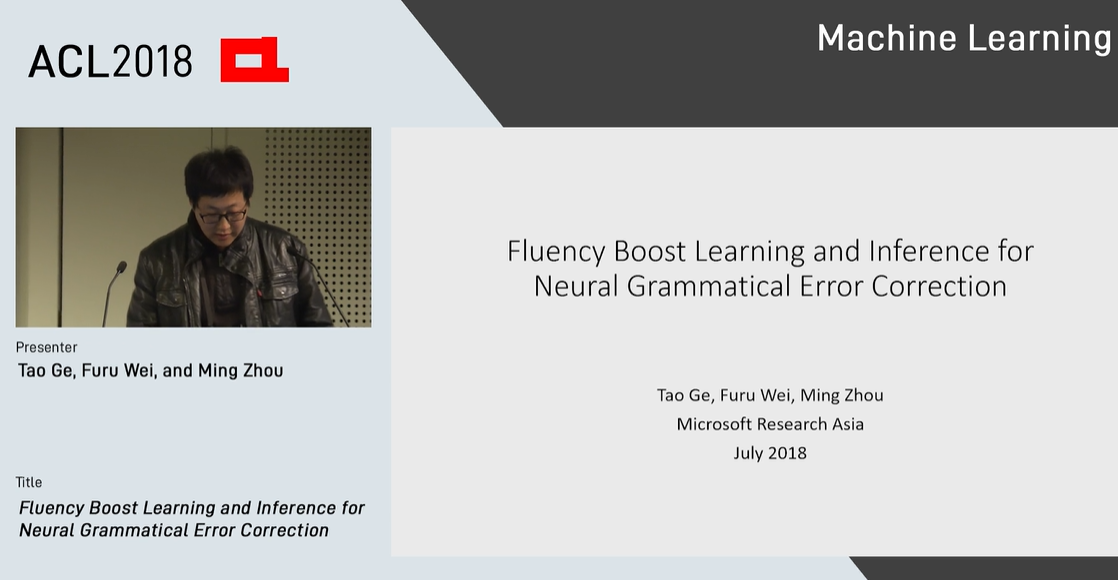
目前基于数据的深度学习和自然语言处理技术已经取得了很大的进展，自动语法改错领域是一个新的成果。但要实现机器真正理解信息，需要人工智能有着更大的突破，从基于大数据解决小任务的感知智能进化到动态理解的认知智能领域，这也是新的人工智能从业者的研究和努力的方向。

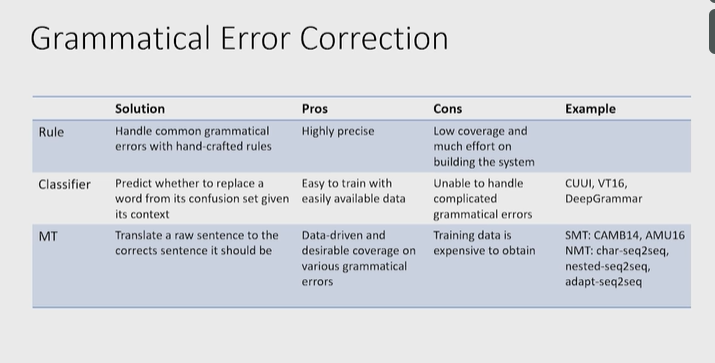
通过顶会论文的阅读，从语法改错领域了解了英文作文打分的技术，并为大作业的完成打开了思路。

附录（全文翻译）：

Fluency Boost Learning and Inference for Neural Grammatical Error Correction

——流利度提升学习和推断用于神经语法错误校正全文翻译



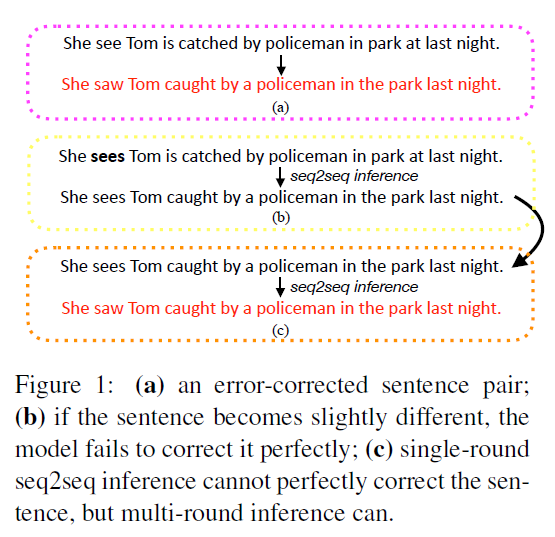


目前绝大部分seq2seq模型用于语法错误校正的有两个局限：（1）一个seq2seq模型很难利用有限错误的校正数据来生产（2）使用传统的seq2seq推理可能不能完全校正有多重错误的句子。文章尝试通过提出一种流利度提升的学习和推理机制来解决这些局限。流利度提升的学习在训练中会产生流利度提升的匹配，使得错误校正模型可以学习如何通过更多的例子去提升句子的流利性，当流利度提升推理递增地通过多轮seq2seq推理直到句子的流利性停止提升。实验证明这个方法提升了seq2seq模型在在GEC上的表现，在CoNLL-10和JFLEG两个基准测试集上按照人机对比的评测设定，结合了流畅度提升学习和推断机制的seq2seq模型的得分分别为74.84和62.42的得分，均首次超越了人类在这两个基准上的表现（CoNLL: 72.58, JFLEG: 62.37）。

1. 介绍

Sequence-to-sequence模型用于语法错误校正方面在近些年引起了越来越多的注意。但

是绝大部分seq2seq模型在GEC上有两个缺陷，第一是序列到序列模型被只有有限的错误校正的句子配对训练，像Figure1。被训练数据的规模所限制，有百万级参数的模型可能就难以被很好的泛化。因此，很常见的一个模型无法完美的校正一个句子即使这个句子与训练样例只有轻微的不同，正如Figure1所说明的。



第二种缺陷是，序列到序列的模型经常很难通过单轮的seq2seq推理完美改正一个带有很多语法错误的句子，如Figure 1(b)和 1(c)所表示的，因为一些句子中的错误可能使得文本变得奇怪，这会困惑模型去改正其他的错误。

为了处理上述提到的在模型学习和推理中的缺陷，这篇文章提出一种新奇的流利度提升的学习和推理机制，如Figure 2说明的。

对于流利度提升学习，不仅是一个用原始错误改正句子对训练的seq2seq模型，并且生成更加不流利的句子（来自它的n个最优输出）去建立新的错误校正句子对，在训练的时候，通过匹配他们和他们的改正句子，只要这个句子的流利性低于其他正确的句子，如Figure 2所展示的。特别地，我们称生成的错误改正句子对流利度提升句子对，因为这些句子在靶子端总是提升流利度超过在来源端。产生的流利度提升的句子对在训练中会被用来作为额外的训练样例在随后的训练轮次中，允许错误改正模型看到更多的语法错误句子在训练中，并因此提高它的泛化能力。

对于模型推理，流利度提升推理机制允许模型去通过多轮次推理递增地改正句子，只要被提出的修改可以提升句子的流利度，如Figure 2所展示的。对于一个有多种语法错误的句子，一些错误会在第一步被改正。改正的部分会使得文本更加清晰，这可能有利于模型改正剩余的错误。（一个句子的流利度被定义为句子交叉熵的反比）

实验证明了流利度提升学习和推理可以使得序列到序列神经网络模型在GEC任务上表现的更好，并取得最先进的结果在多重GEC基准上。

我们的贡献在下面总结：

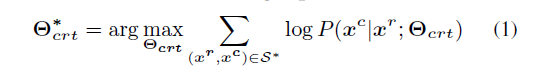
* 我们提出了一个新奇的学习和推理机制来解决现在序列到序列模型在GEC上的局限。
* 我们提出并比较了多种新颖的流利度提升学习策略，探索神经网络GEC的学习方法论。
* 我们的方法被证明有效地提升神经网络seq2seq模型去在CoNLL-2014和JFLEG数据集上实现最先进的结果

1. 背景：自动语法错误校正

正如神经机器翻译，一个典型的神经网络GEC方法使用基于编码-解码的序列到序列的

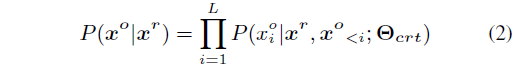
循环神经网络(Sutskever et al., 2014; Cho et al., 2014)，并加入注意力机制(Bahdanau et al., 2014)来编辑未加工的句子到语法上改正后的句子，如Figure 1(a)展示的。

给出一个未加工的句子xr=(x1r,…,xrM)和他的改正的句子xc=(x1c,…,xcN),其中xrM和xcN都是第M个和第N个词分别在句子xr和xc中，序列到序列的错误改正模型学习一种概率的映射P(xc|xr)，通过最大似然估计方法(MLE)，学习模型参数θcrt 去最小化下面的等式：



这里符号S\*表示错误改正句子对的集合。

为了模型的推理，一个输出的句子x0=(x10,…,xi0,…,xl0)通过通渠搜索选择，并最大化下面的等式：



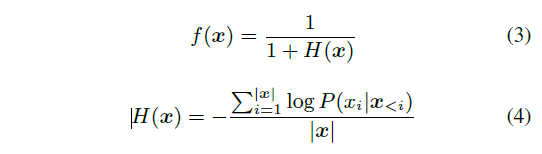
1. 流利度提升学习

传统序列到序列的用于GEC学习的模型参数只来自于原始的错误改正句子对。可是，这

样的错误校正句子对不是充分可得到的。结果是，许多神经GEC模型不能被很好的生成。

幸运的是，神经网络语法错误校正不同于神经网络机器翻译。对于神经网络语法校正，它的目标是提升句子的流利度，并不改变他原本的意思。因此，任何句子配对满足这条件的（我们称之为流利度提升条件）可以被用于训练样例。

在这篇文章中，我们定义f(x)为作为句子的流利度分数 x:

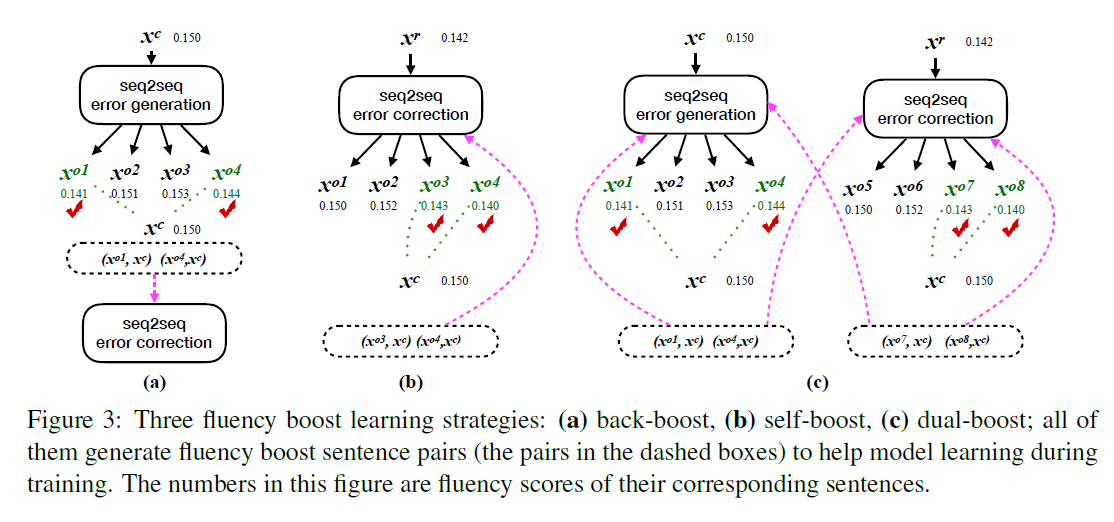


上式中P(xi|x<i)是xi的概率在给出x<i 的文本中，由语言模型计算得到。|x|是句子x的长度，它的范围为[0,+ ∞）。因此，f(x)的范围以为(0,1]

流利度提升学习的核心思想是，去生产流利度提升句子对，可以在训练中满足流利度提升条件的，如Figure 2所阐明的，以至于这些匹配可以更好的帮助模型学习。

（一个句子的流利度在本文中涉及到这个句子被母语者写出的可能性。换句话说，如果一个句子很有可能写自一位母语者，它就被认为有很高流利度。）

在这个部分，我们提出了三种流利度提升学习策略：返回-提升、自我-提升、双重-提升，产生流利度提升句子匹配通过不同的方式，如Figure 3所阐释的。



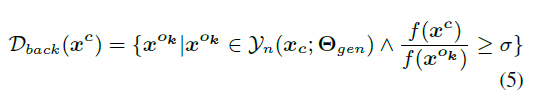
* 1. 返回-提升学习

返回-提升学习的思想来自于回译(Sennrich et al., 2016)，在神经网络机器翻译中，涉

及到训练一个倒退的模型（我们称之为错误生成模型，作为错误改正模型的对立）用于将一个流利的句子转换为更不流利且带有错误的句子。既然更低流利度的句子被错误生成的序列到序列模型用错误改正数据来训练，他们通常不改变原句子的意思；因此，他们可以都他们改正后的句子所配对，建立流利度提升的句子匹配，可以被用于作为错误改正模型的训练样例的，像Figure 3所示。

特别地，我们第一次训练序列到序列错误生成模型θgen和S\*，并和S\*同样除了源句和目标句互换。然后，我们使用模型θgen去预测n个最好的输出xo1,…,xon,并给出改正的句子xc。给出流利度提升的条件，我们比较每一个输出xok(1<=k<=n)和它的校正句子的流利度。如果一个输出的句子的流利度得分远低于它校正后的句子，我们称它为xc的一个不流利的替代。

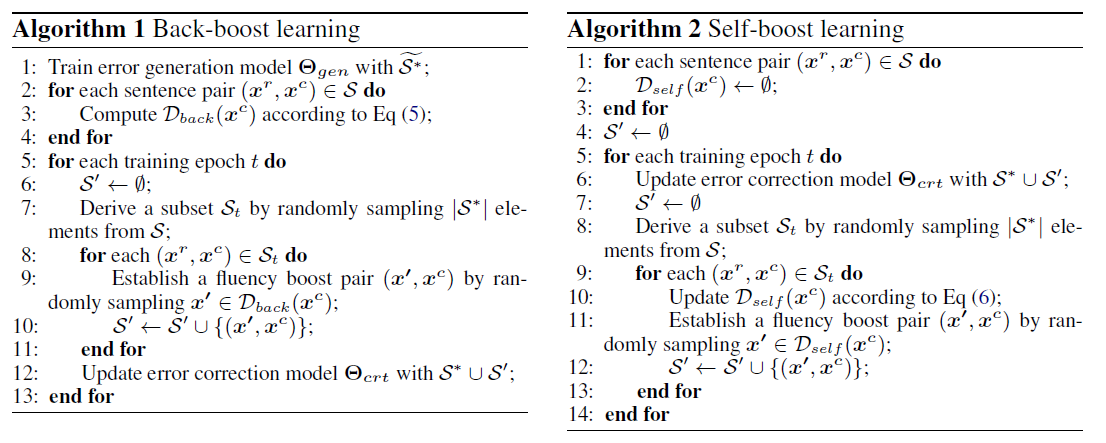
为了使这样的进程形式化，我们首先定义Yn(x; θ)来表示n个最好的输出预测通过由输入x给出的模型θ。然后，正确句子的不流利替代xc可以被派生，



Dback(xc)表示在返回-提升学习中xc的不流利替代集合。是一个临界值，来定义是否xok比xc更加不流利，它应该略大于1，可以帮助过滤出不需要编辑的句子对（如I like this book. ! I like the book.）

在随后的训练轮次中，错误校正模型不仅会从原始的错误校正句子对中学习，还会从流利度提升句子对中去学习，如果xok是Dback(xc)的一个样本的话。

我们总结算法1 的过程，S\*是原始错误校正句子对的集合，S可以暂时被认为和S\*是相同的，在没有额外原始数据去帮忙模型训练的时候。注意，我们约束了St的规模，不能超过|S\*|来避免太多的流利度提升匹配对在模型学习中覆盖了原始错误校正对的作用。



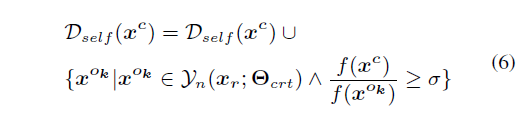
* 1. 自-提升学习

相对于最初核心思路来自神经网络机器学习的返回-提升学习，自-提升学习是原创的，它特

别的为了神经网络语法错误校正而发明。自-提升学习的思想在Figure 3(b)中阐明，并已经简洁的在第一部分和Figure 2(a)中介绍。不像返回-提升学习，序列到序列的错误生成模型为了产生不流利的替代而训练，自-提升训练允许错误校正模型去自己生成候选数据。既然通过由错误改正数据训练的序列到序列的错误改正数据生成的不流利的候选数据几乎不改变输入句子的意思，因此，他们可以被用来产生流利度提升的句子对。

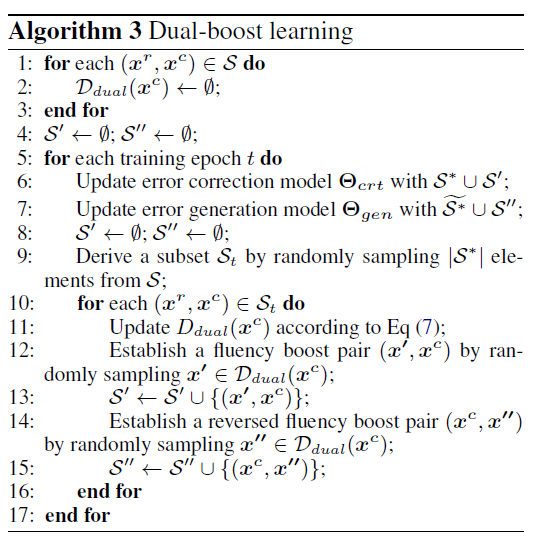
（在本文中，我们设置σ=1.05，因为在我们的训练数据中，改正的句子平均提高5%的流利度相较于原始句子而言。）

对于自-提升学习，给出一个错误校正的匹配(xr,xc),错误改正模型θcrt首先预测来自原始句子xr的n个最好输出。在这n个最好的输出中，任何一个与xc不同的输出都可以被认为是一个错误的预测。除了处理无用的错误预测，自-提升训练充分的利用它们。特别地，如果一个错误预测xok远不如改正的句子xc流利，它可以加入xc的不流利候选集合Dself(xc)，如Eq(6)所示：



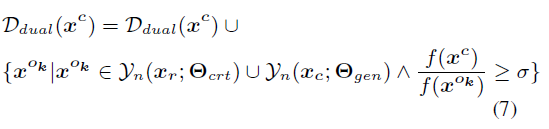
相较于返回-提升学习，自-提升产生不流利的候选从不同的视角——通过编辑原始句子xr而不是改正后的句子xc。同样值得注意的是，Dself(xc)是递增扩展的，因为错误改正模型θcrt是动态更新的，如算法2所示。

* 1. 双重-提升学习



正如上面所介绍的，返回-提升和自-提升学习从不同视角产生不流利的候选，去创造更多流利度提升的句子配对，有利于错误改正模型的训练。直观地，更多不同的不流利候选的生成，更有利于错误改正模型的训练。受He et al. (2016)和Zhang et al. (2018)的启发，我们提出一种双重-提升的策略，同时包括返回-提升和自-提升观点，去生成更不流利的候选数据。

如Figure 3(c)所示，在双重提升学习中不流利的候选数据同时来自错误生成模型和错误改正模型：



此外，错误改正模型和错误生成模型都是双重的，他们都动态的更新，并相互提升：不流利的候选数据产生错误生成模型可以有利于训练错误改正模型。与此同时，错误改正模型产生的不流利候选数据可以用于错误生成模型的训练数据。我们将其总结为算法3。

* 1. 大规模原始数据的流利度提升学习

我们提出的流利度提升学习策略可以被很容易的用来扩展巨大数量的原始数据，并被证明

对GEC有用。

如3.1中讨论的，当没有额外的原始数据，S在算法1-3中等同于S\*。在这个例子中，额外的原始训练数据可以用来帮助模型训练，S变为：S=S\*∪C，C={(xc,xc)}代表来自原始数据的自复制句子对集合。

1. 流利度提升推理

如我们在第一部分所讨论的，一些有多种语法错误的句子通常无法被一般的序列到序列推理

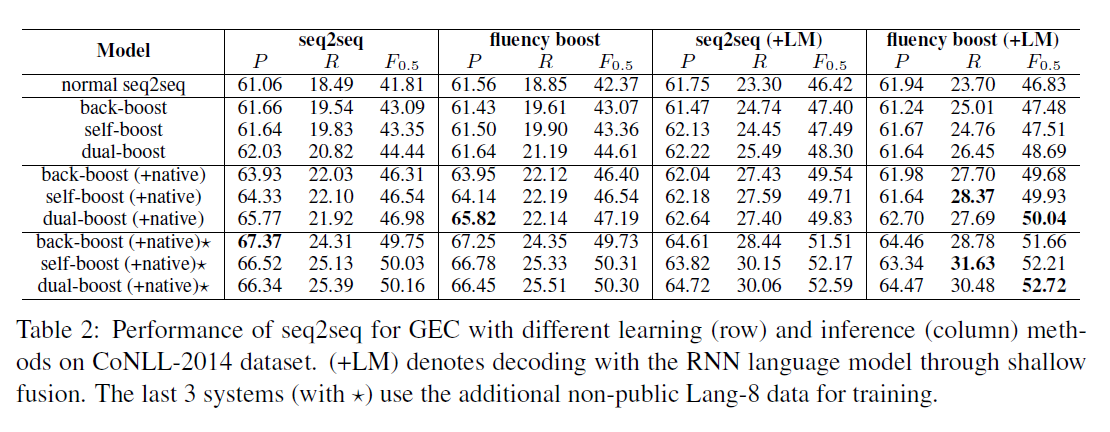
所改正，它只做单轮推理。幸运的是,神经GEC不同于NMT: 它的源语言和目标语言是一样的。特征，允许我们多编辑一句话，一次通过多轮模型推理，得出激发我们的流畅性，促进推理。如图2(b)表明，流畅性促进推理允许通过多轮逐步编辑句子seq2seq推理只要句子的流利程度可以提高。具体地说,一个错误修正seq2seq模型首先取一个原始句子xr作为输入输出假设xo1。与其将xo1作为最终的预测，流畅性促进推理将以xo1作为输入生成下一个输出xo2。这个过程不会终止，除非xot没有改进xot-1的流利度。

1. 实验

5.1 数据集和评估

如既往研究（Ji et al。，2017），我们使用公众语言-8语料库（Mizumoto等，2011；Tajiri等，2012），剑桥学习者语料库（CLC）（Nicholls，2003）和新加坡国立大学语料库关于学习者英语（NUCLE）（Dahlmeier等，作为我们原始的错误修正的训练数据。表1显示了数据集的统计数据。我们还收集了2,865,639个被纠正的非公共错误来自Lang-8.com的句子对。我们用来提高流利度的原生数据是英文维基百科，包含61,677,453个句子。

我们将CoNLL-2014共享任务数据集与原始注释（Ng等，2014），其中包含1,312个句子，作为我们评估的主要测试集。我们使用MaxMatch（M2）精度F0：5（Dahlmeier和Ng，2012b）作为我们的评估指标。 作为以前的研究，我们使用CoNLL-2013年测试数据作为我们的发展目标。



5.2 实验设置

为了回答以下问题，我们设置了实验问题：

● 流利度提升学习机制是否有助于训练错误校正模型，以及哪些策略（反向提升，自我提升，双重提升）最有效？

● 我们的流利度提升推论是否有所改善GEC的正常seq2seq推断？

● 我们的方法是否可以改善神经GEC取得了最先进的成果？

seq2seq纠错的培训细节模型和错误生成模型如下：seq2seq模型的编码器是2层双向GRU RNN和解码器是2层GRU RNN，受到普遍关注机理（Luong et al。，2015）。二维嵌入和隐藏大小的GRU单元数为500。编码器和解码器分别为100,000和50,000。模型的参数是一致的在[-0.1,0.1]中初始化。我们训练模型学习率为0.0001的Adam优化器批次大小= 128时最多40个纪元。应用于非循环连接，其比率为0.15。为了提高流利度，我们会产生流利度10项最佳产出的候选人。中模型推断，我们将波束大小设置为5并解码2层GRU RNN语言获得1个最佳结果浅层融合的数学模型（Mikolov等，2010）（Gulc¸ehreet al。，2015），重量= 0:15。

RNN语言模型是由本地人训练的第5.1节中提到的数据用于计算等式（3）中的流利度得分。 K

令牌被替换为带有最高关注体重。

我们用公共拼写来解决拼写错误checker4作为预处理，如Xie等人(2016)和Sakaguchi等人(2017)做到了。

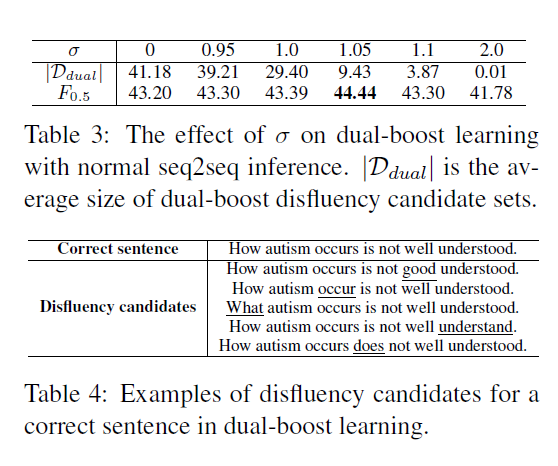
5.3 实验结果

5.3.1 流利度提升学习的效果

表2比较了seq2seq错误的性能具有不同学习和推论方法。通过逐行比较，可以观察我们的流利度促进学习方法与常规seq2seq相比，提高了性能自从流利的促进学习的方法产生了多种语法错误的句子，允许纠错模型学习纠正句子比传统学习要多得多战略。建议的三种流利程度提升学习策略，双重提升实现在大多数情况下，最好的结果是因为它可以产生更多不同的错误句子（平均jDdualj9:43）均胜过任一后提升（平均jDbackj 1:90）或自我提升学习（平均jDself j 8:10）。用引入大量的本地文本数据，所有流利度提升学习方法的表现得到改善。原因之一是我们学习方法产生更多的错误纠正句子对，以使模型更好地泛化。此外，大量的本机数据有利于解码器更好地学习生成流利且无错误的句子。

我们在方程式中测试超参数的效果（5–7）关于流利度的促进学习并显示结果在表3中。当大于1.0时（例如= 1:05），则模型可获得最佳性能因为它有效地避免了产生具有不必要或不希望的编辑的句子对正如我们在第3.1节。 当持续增加时，流离失所候选集jDdualj急剧减少，使双重提升学习逐渐退化正常的seq2seq学习。

表4显示了一些不流畅的例子。



给定双提升学习中产生的候选人5本机数据中的正确句子。 很清楚我们的方法会产生较少流利的句子带有各种语法错误，其中大多数是人类学习者往往会犯的典型错误使。 因此，它们可以用来建立具有正确句子的高质量培训数据，这将有助于增加尺寸训练数据到次数，会计通过流利的促进学习来改善。

5.3.2 流利度提升推理的效果

各种推理方法的有效性通过比较表中的结果可以观察到2按栏。 与正常的seq2seq相比推断和seq2seq（+ LM）基线，流利程度提升推论平均带来0.14和F0：5分别获得0.18的增益，这是一个显着的6改进，展示多轮通过流利度增强推论进行编辑是有效的。

采用我们最好的系统（表中的最后一行2）例如，在1,312个句子中CoNLL-2014数据集，seq2seq推断与浅层融合LM编辑566个句子。 相反，流利度提升推断还可以编辑23第二轮推理中的句子，改进F0：5从52.59至52.72。

5.3.3 GEC最先进的结果

现在，我们回答本节中提出的最后一个问题5.2通过测试我们的方法是否达到以下状态：

最新的结果。

我们首先比较一下我们的最佳模型-双升压通过流畅的推理学习（+母语）和浅层融合LM –性能卓越的GEC在CoNLL-2014数据集上评估的系统：

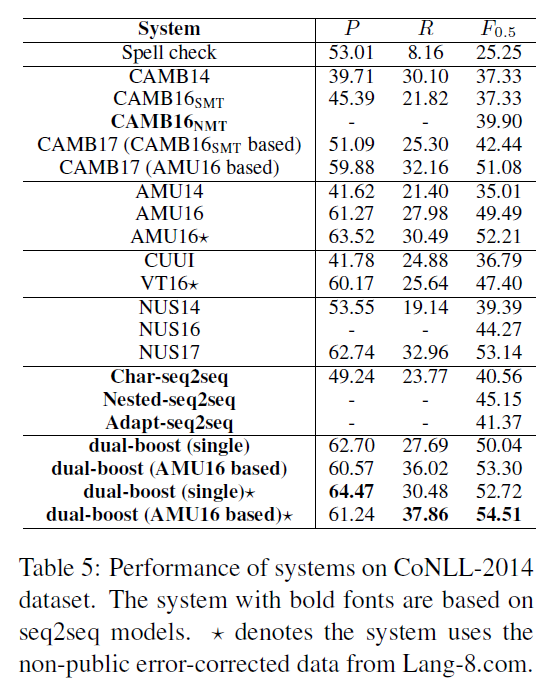


表5显示了对产品的评估结果。CoNLL-2014数据集。 不使用非公开来自Lang-8.com的训练数据。模型获得50.04 F0：5，表现不佳其他seq2seq模型，且仅次于CAMB17（基于AMU16）和NUS17。 这应该但是请注意，CAMB17和NUS17实际上是基于SMT构建的重新排名GEC系统（AMU16的框架）； 从而，它们是合奏模型。 当我们建立方法时在AMU16的顶部（即，我们以AMU16的输出作为输入到我们的GEC系统进行编辑在其输出之上），我们达到了53.30 F0：5得分了。 随着介绍非公开培训数据，我们的整体系统便获得52.72和54.51 F0：5得分，这是一种状态-CoNLL-2014数据集的最新结果7。

此外，我们评估了我们在JFLEG上的方法语料库（Napoles et al。，2017）。 JFLEG是最新的已发布用于GEC评估的数据集，其中包含1,501个句子（开发集中为754，测试集中为747）。为了测试我们的方法的泛化能力，我们评估我们用于CoNLL评估的单个模型（见表5）在JFLEG上无需重新调整。

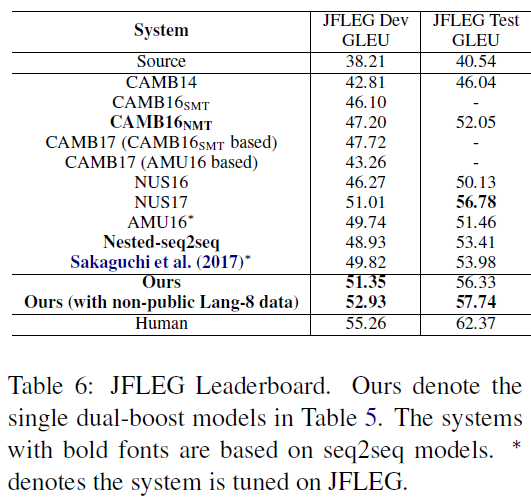


表6显示了JFLEG排行榜。 代替在M2得分中，JFLEG使用GLEU（Napoles等，2015年）作为评估指标，以流利为导向基于BLEU变体的GEC指标（Papineni et al。，2002）并具有多个优势在M2上进行GEC评估。 据观察，我们的单个模型在JFLEG，表现优于大多数CoNLL-2014表现最佳的系统并产生最新状态在这个基准上的艺术成果8，展示了我们的模型被很好地概括并执行稳定地在多个数据集上。

1. 结论

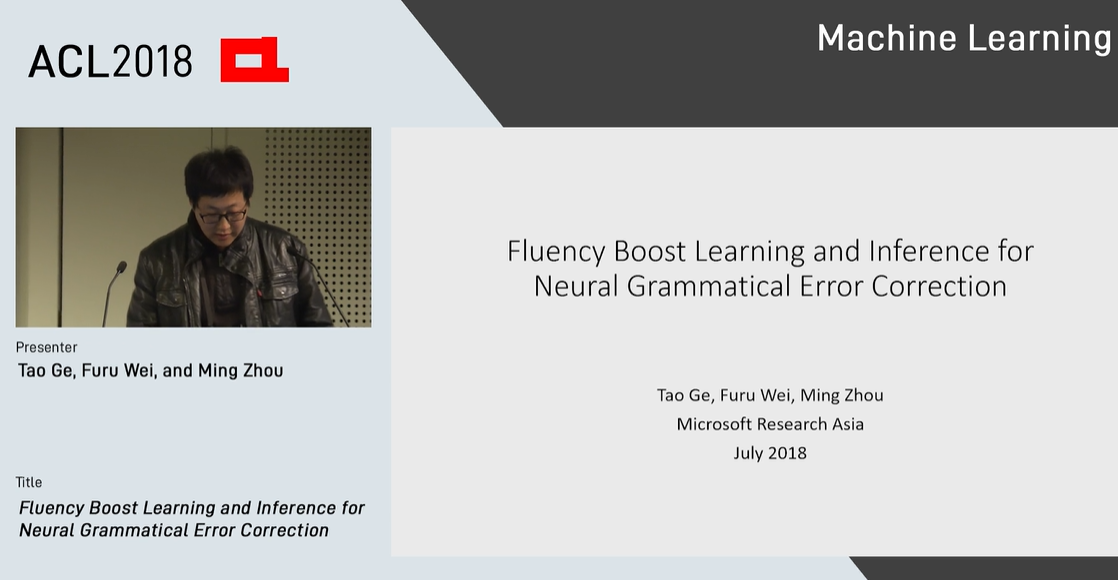
我们提出了一种新颖的流利性促进学习和克服限制的推理机制以前的神经GEC模型。 我

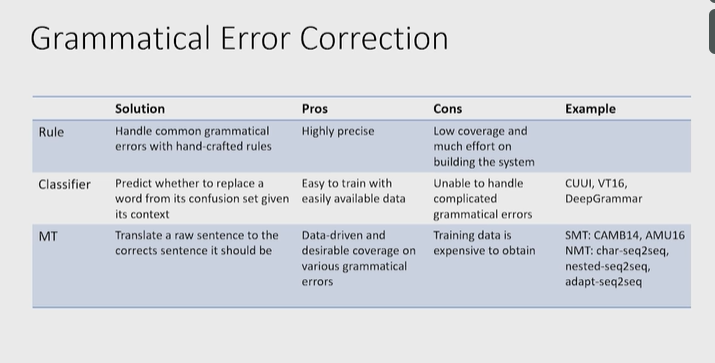
们的建议流利性提升学习充分利用了两种错误纠正方法数据和本机数据，大大改善了正常seq2seq学习的表现，而流利度提升推理则利用GEC的特征，以逐步改善通过多轮推理使句子流利。强大的学习和推理机制使seq2seq模型能够达到CoNLL-2014和JFLEG的艺术成果基准数据集。

附录（全文翻译）：

Fluency Boost Learning and Inference for Neural Grammatical Error Correction

——流利度提升学习和推断用于神经语法错误校正全文翻译



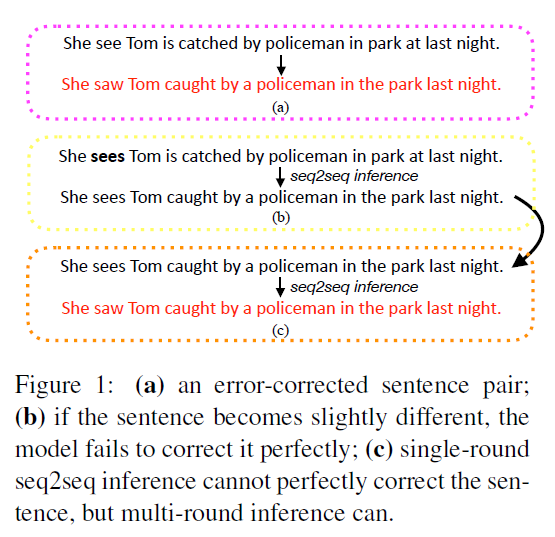


目前绝大部分seq2seq模型用于语法错误校正的有两个局限：（1）一个seq2seq模型很难利用有限错误的校正数据来生产（2）使用传统的seq2seq推理可能不能完全校正有多重错误的句子。文章尝试通过提出一种流利度提升的学习和推理机制来解决这些局限。流利度提升的学习在训练中会产生流利度提升的匹配，使得错误校正模型可以学习如何通过更多的例子去提升句子的流利性，当流利度提升推理递增地通过多轮seq2seq推理直到句子的流利性停止提升。实验证明这个方法提升了seq2seq模型在在GEC上的表现，在CoNLL-10和JFLEG两个基准测试集上按照人机对比的评测设定，结合了流畅度提升学习和推断机制的seq2seq模型的得分分别为74.84和62.42的得分，均首次超越了人类在这两个基准上的表现（CoNLL: 72.58, JFLEG: 62.37）。

1. 介绍

Sequence-to-sequence模型用于语法错误校正方面在近些年引起了越来越多的注意。但

是绝大部分seq2seq模型在GEC上有两个缺陷，第一是序列到序列模型被只有有限的错误校正的句子配对训练，像Figure1。被训练数据的规模所限制，有百万级参数的模型可能就难以被很好的泛化。因此，很常见的一个模型无法完美的校正一个句子即使这个句子与训练样例只有轻微的不同，正如Figure1所说明的。



第二种缺陷是，序列到序列的模型经常很难通过单轮的seq2seq推理完美改正一个带有很多语法错误的句子，如Figure 1(b)和 1(c)所表示的，因为一些句子中的错误可能使得文本变得奇怪，这会困惑模型去改正其他的错误。

为了处理上述提到的在模型学习和推理中的缺陷，这篇文章提出一种新奇的流利度提升的学习和推理机制，如Figure 2说明的。

对于流利度提升学习，不仅是一个用原始错误改正句子对训练的seq2seq模型，并且生成更加不流利的句子（来自它的n个最优输出）去建立新的错误校正句子对，在训练的时候，通过匹配他们和他们的改正句子，只要这个句子的流利性低于其他正确的句子，如Figure 2所展示的。特别地，我们称生成的错误改正句子对流利度提升句子对，因为这些句子在靶子端总是提升流利度超过在来源端。产生的流利度提升的句子对在训练中会被用来作为额外的训练样例在随后的训练轮次中，允许错误改正模型看到更多的语法错误句子在训练中，并因此提高它的泛化能力。

对于模型推理，流利度提升推理机制允许模型去通过多轮次推理递增地改正句子，只要被提出的修改可以提升句子的流利度，如Figure 2所展示的。对于一个有多种语法错误的句子，一些错误会在第一步被改正。改正的部分会使得文本更加清晰，这可能有利于模型改正剩余的错误。（一个句子的流利度被定义为句子交叉熵的反比）

实验证明了流利度提升学习和推理可以使得序列到序列神经网络模型在GEC任务上表现的更好，并取得最先进的结果在多重GEC基准上。

我们的贡献在下面总结：

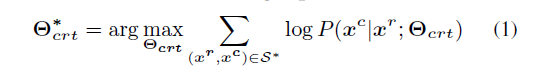
* 我们提出了一个新奇的学习和推理机制来解决现在序列到序列模型在GEC上的局限。
* 我们提出并比较了多种新颖的流利度提升学习策略，探索神经网络GEC的学习方法论。
* 我们的方法被证明有效地提升神经网络seq2seq模型去在CoNLL-2014和JFLEG数据集上实现最先进的结果

1. 背景：自动语法错误校正

正如神经机器翻译，一个典型的神经网络GEC方法使用基于编码-解码的序列到序列的

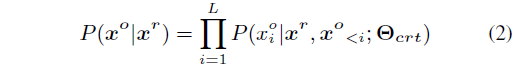
循环神经网络(Sutskever et al., 2014; Cho et al., 2014)，并加入注意力机制(Bahdanau et al., 2014)来编辑未加工的句子到语法上改正后的句子，如Figure 1(a)展示的。

给出一个未加工的句子xr=(x1r,…,xrM)和他的改正的句子xc=(x1c,…,xcN),其中xrM和xcN都是第M个和第N个词分别在句子xr和xc中，序列到序列的错误改正模型学习一种概率的映射P(xc|xr)，通过最大似然估计方法(MLE)，学习模型参数θcrt 去最小化下面的等式：



这里符号S\*表示错误改正句子对的集合。

为了模型的推理，一个输出的句子x0=(x10,…,xi0,…,xl0)通过通渠搜索选择，并最大化下面的等式：



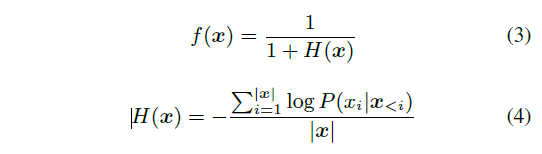
1. 流利度提升学习

传统序列到序列的用于GEC学习的模型参数只来自于原始的错误改正句子对。可是，这

样的错误校正句子对不是充分可得到的。结果是，许多神经GEC模型不能被很好的生成。

幸运的是，神经网络语法错误校正不同于神经网络机器翻译。对于神经网络语法校正，它的目标是提升句子的流利度，并不改变他原本的意思。因此，任何句子配对满足这条件的（我们称之为流利度提升条件）可以被用于训练样例。

在这篇文章中，我们定义f(x)为作为句子的流利度分数 x:

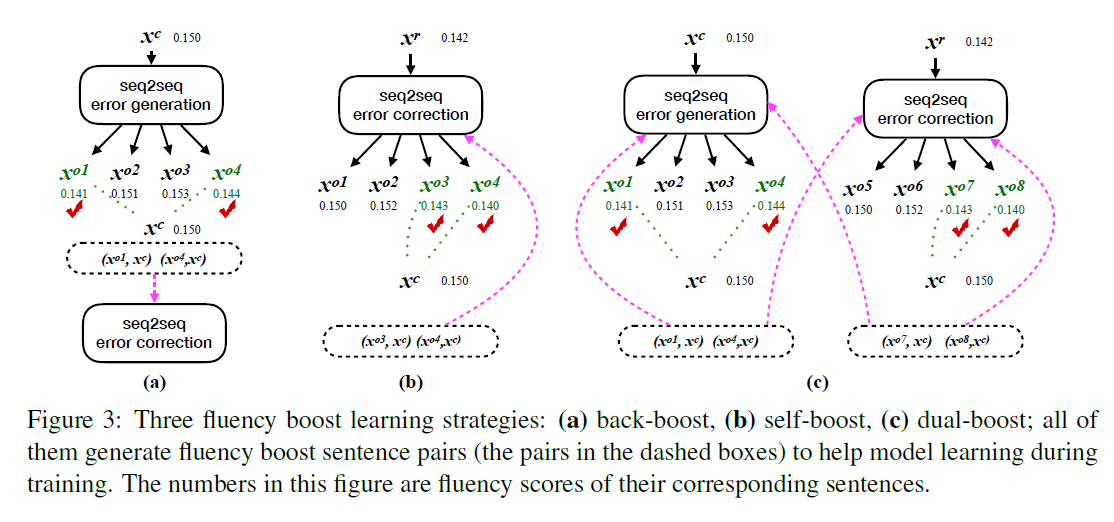


上式中P(xi|x<i)是xi的概率在给出x<i 的文本中，由语言模型计算得到。|x|是句子x的长度，它的范围为[0,+ ∞）。因此，f(x)的范围以为(0,1]

流利度提升学习的核心思想是，去生产流利度提升句子对，可以在训练中满足流利度提升条件的，如Figure 2所阐明的，以至于这些匹配可以更好的帮助模型学习。

（一个句子的流利度在本文中涉及到这个句子被母语者写出的可能性。换句话说，如果一个句子很有可能写自一位母语者，它就被认为有很高流利度。）

在这个部分，我们提出了三种流利度提升学习策略：返回-提升、自我-提升、双重-提升，产生流利度提升句子匹配通过不同的方式，如Figure 3所阐释的。



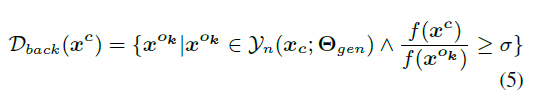
* 1. 返回-提升学习

返回-提升学习的思想来自于回译(Sennrich et al., 2016)，在神经网络机器翻译中，涉

及到训练一个倒退的模型（我们称之为错误生成模型，作为错误改正模型的对立）用于将一个流利的句子转换为更不流利且带有错误的句子。既然更低流利度的句子被错误生成的序列到序列模型用错误改正数据来训练，他们通常不改变原句子的意思；因此，他们可以都他们改正后的句子所配对，建立流利度提升的句子匹配，可以被用于作为错误改正模型的训练样例的，像Figure 3所示。

特别地，我们第一次训练序列到序列错误生成模型θgen和S\*，并和S\*同样除了源句和目标句互换。然后，我们使用模型θgen去预测n个最好的输出xo1,…,xon,并给出改正的句子xc。给出流利度提升的条件，我们比较每一个输出xok(1<=k<=n)和它的校正句子的流利度。如果一个输出的句子的流利度得分远低于它校正后的句子，我们称它为xc的一个不流利的替代。

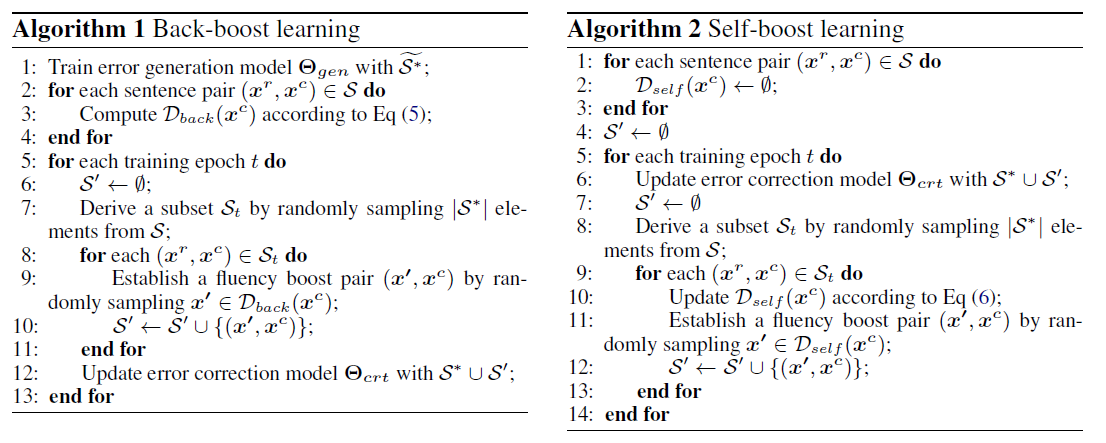
为了使这样的进程形式化，我们首先定义Yn(x; θ)来表示n个最好的输出预测通过由输入x给出的模型θ。然后，正确句子的不流利替代xc可以被派生，



Dback(xc)表示在返回-提升学习中xc的不流利替代集合。是一个临界值，来定义是否xok比xc更加不流利，它应该略大于1，可以帮助过滤出不需要编辑的句子对（如I like this book. ! I like the book.）

在随后的训练轮次中，错误校正模型不仅会从原始的错误校正句子对中学习，还会从流利度提升句子对中去学习，如果xok是Dback(xc)的一个样本的话。

我们总结算法1 的过程，S\*是原始错误校正句子对的集合，S可以暂时被认为和S\*是相同的，在没有额外原始数据去帮忙模型训练的时候。注意，我们约束了St的规模，不能超过|S\*|来避免太多的流利度提升匹配对在模型学习中覆盖了原始错误校正对的作用。



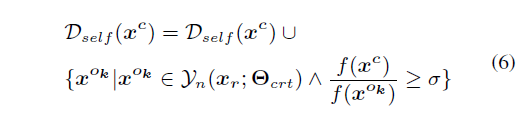
* 1. 自-提升学习

相对于最初核心思路来自神经网络机器学习的返回-提升学习，自-提升学习是原创的，它特

别的为了神经网络语法错误校正而发明。自-提升学习的思想在Figure 3(b)中阐明，并已经简洁的在第一部分和Figure 2(a)中介绍。不像返回-提升学习，序列到序列的错误生成模型为了产生不流利的替代而训练，自-提升训练允许错误校正模型去自己生成候选数据。既然通过由错误改正数据训练的序列到序列的错误改正数据生成的不流利的候选数据几乎不改变输入句子的意思，因此，他们可以被用来产生流利度提升的句子对。

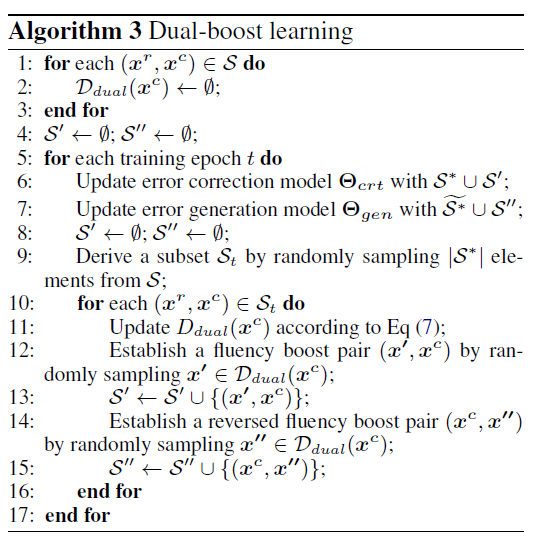
（在本文中，我们设置σ=1.05，因为在我们的训练数据中，改正的句子平均提高5%的流利度相较于原始句子而言。）

对于自-提升学习，给出一个错误校正的匹配(xr,xc),错误改正模型θcrt首先预测来自原始句子xr的n个最好输出。在这n个最好的输出中，任何一个与xc不同的输出都可以被认为是一个错误的预测。除了处理无用的错误预测，自-提升训练充分的利用它们。特别地，如果一个错误预测xok远不如改正的句子xc流利，它可以加入xc的不流利候选集合Dself(xc)，如Eq(6)所示：



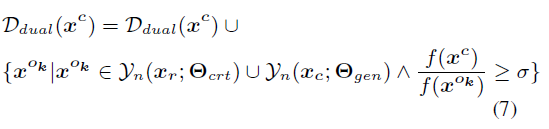
相较于返回-提升学习，自-提升产生不流利的候选从不同的视角——通过编辑原始句子xr而不是改正后的句子xc。同样值得注意的是，Dself(xc)是递增扩展的，因为错误改正模型θcrt是动态更新的，如算法2所示。

* 1. 双重-提升学习



正如上面所介绍的，返回-提升和自-提升学习从不同视角产生不流利的候选，去创造更多流利度提升的句子配对，有利于错误改正模型的训练。直观地，更多不同的不流利候选的生成，更有利于错误改正模型的训练。受He et al. (2016)和Zhang et al. (2018)的启发，我们提出一种双重-提升的策略，同时包括返回-提升和自-提升观点，去生成更不流利的候选数据。

如Figure 3(c)所示，在双重提升学习中不流利的候选数据同时来自错误生成模型和错误改正模型：



此外，错误改正模型和错误生成模型都是双重的，他们都动态的更新，并相互提升：不流利的候选数据产生错误生成模型可以有利于训练错误改正模型。与此同时，错误改正模型产生的不流利候选数据可以用于错误生成模型的训练数据。我们将其总结为算法3。

* 1. 大规模原始数据的流利度提升学习

我们提出的流利度提升学习策略可以被很容易的用来扩展巨大数量的原始数据，并被证明

对GEC有用。

如3.1中讨论的，当没有额外的原始数据，S在算法1-3中等同于S\*。在这个例子中，额外的原始训练数据可以用来帮助模型训练，S变为：S=S\*∪C，C={(xc,xc)}代表来自原始数据的自复制句子对集合。

1. 流利度提升推理

如我们在第一部分所讨论的，一些有多种语法错误的句子通常无法被一般的序列到序列推理

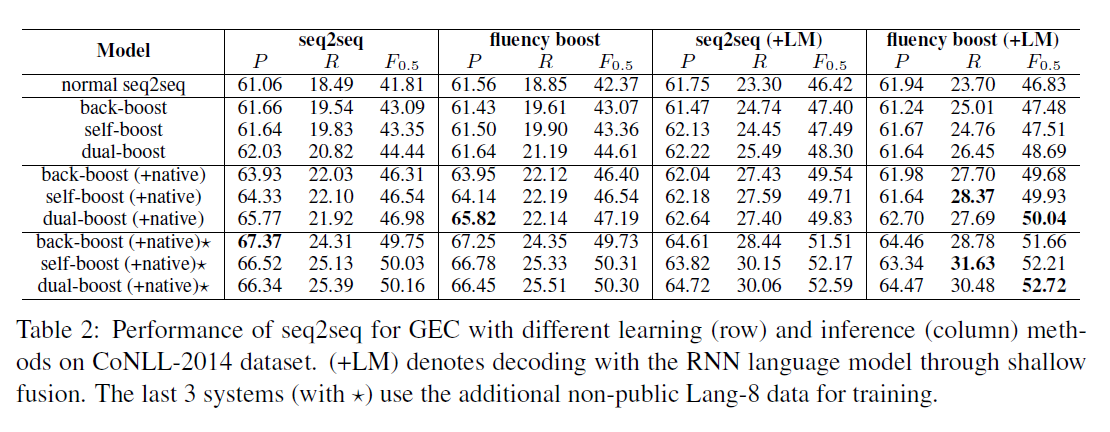
所改正，它只做单轮推理。幸运的是,神经GEC不同于NMT: 它的源语言和目标语言是一样的。特征，允许我们多编辑一句话，一次通过多轮模型推理，得出激发我们的流畅性，促进推理。如图2(b)表明，流畅性促进推理允许通过多轮逐步编辑句子seq2seq推理只要句子的流利程度可以提高。具体地说,一个错误修正seq2seq模型首先取一个原始句子xr作为输入输出假设xo1。与其将xo1作为最终的预测，流畅性促进推理将以xo1作为输入生成下一个输出xo2。这个过程不会终止，除非xot没有改进xot-1的流利度。

1. 实验

5.1 数据集和评估

如既往研究（Ji et al。，2017），我们使用公众语言-8语料库（Mizumoto等，2011；Tajiri等，2012），剑桥学习者语料库（CLC）（Nicholls，2003）和新加坡国立大学语料库关于学习者英语（NUCLE）（Dahlmeier等，作为我们原始的错误修正的训练数据。表1显示了数据集的统计数据。我们还收集了2,865,639个被纠正的非公共错误来自Lang-8.com的句子对。我们用来提高流利度的原生数据是英文维基百科，包含61,677,453个句子。

我们将CoNLL-2014共享任务数据集与原始注释（Ng等，2014），其中包含1,312个句子，作为我们评估的主要测试集。我们使用MaxMatch（M2）精度F0：5（Dahlmeier和Ng，2012b）作为我们的评估指标。 作为以前的研究，我们使用CoNLL-2013年测试数据作为我们的发展目标。



5.2 实验设置

为了回答以下问题，我们设置了实验问题：

● 流利度提升学习机制是否有助于训练错误校正模型，以及哪些策略（反向提升，自我提升，双重提升）最有效？

● 我们的流利度提升推论是否有所改善GEC的正常seq2seq推断？

● 我们的方法是否可以改善神经GEC取得了最先进的成果？

seq2seq纠错的培训细节模型和错误生成模型如下：seq2seq模型的编码器是2层双向GRU RNN和解码器是2层GRU RNN，受到普遍关注机理（Luong et al。，2015）。二维嵌入和隐藏大小的GRU单元数为500。编码器和解码器分别为100,000和50,000。模型的参数是一致的在[-0.1,0.1]中初始化。我们训练模型学习率为0.0001的Adam优化器批次大小= 128时最多40个纪元。应用于非循环连接，其比率为0.15。为了提高流利度，我们会产生流利度10项最佳产出的候选人。中模型推断，我们将波束大小设置为5并解码2层GRU RNN语言获得1个最佳结果浅层融合的数学模型（Mikolov等，2010）（Gulc¸ehreet al。，2015），重量= 0:15。

RNN语言模型是由本地人训练的第5.1节中提到的数据用于计算等式（3）中的流利度得分。 K

令牌被替换为带有最高关注体重。

我们用公共拼写来解决拼写错误checker4作为预处理，如Xie等人(2016)和Sakaguchi等人(2017)做到了。

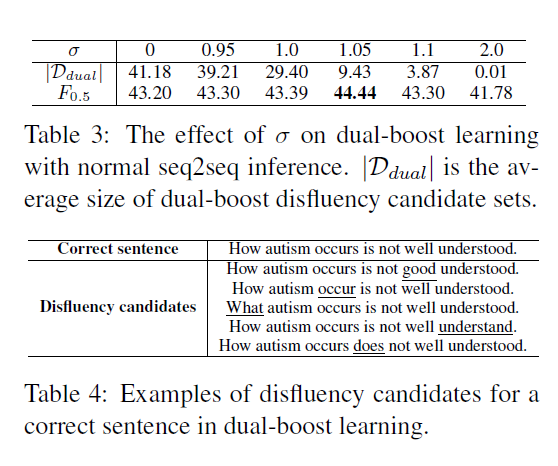
5.3 实验结果

5.3.1 流利度提升学习的效果

表2比较了seq2seq错误的性能具有不同学习和推论方法。通过逐行比较，可以观察我们的流利度促进学习方法与常规seq2seq相比，提高了性能自从流利的促进学习的方法产生了多种语法错误的句子，允许纠错模型学习纠正句子比传统学习要多得多战略。建议的三种流利程度提升学习策略，双重提升实现在大多数情况下，最好的结果是因为它可以产生更多不同的错误句子（平均jDdualj9:43）均胜过任一后提升（平均jDbackj 1:90）或自我提升学习（平均jDself j 8:10）。用引入大量的本地文本数据，所有流利度提升学习方法的表现得到改善。原因之一是我们学习方法产生更多的错误纠正句子对，以使模型更好地泛化。此外，大量的本机数据有利于解码器更好地学习生成流利且无错误的句子。

我们在方程式中测试超参数的效果（5–7）关于流利度的促进学习并显示结果在表3中。当大于1.0时（例如= 1:05），则模型可获得最佳性能因为它有效地避免了产生具有不必要或不希望的编辑的句子对正如我们在第3.1节。 当持续增加时，流离失所候选集jDdualj急剧减少，使双重提升学习逐渐退化正常的seq2seq学习。

表4显示了一些不流畅的例子。



给定双提升学习中产生的候选人5本机数据中的正确句子。 很清楚我们的方法会产生较少流利的句子带有各种语法错误，其中大多数是人类学习者往往会犯的典型错误使。 因此，它们可以用来建立具有正确句子的高质量培训数据，这将有助于增加尺寸训练数据到次数，会计通过流利的促进学习来改善。

5.3.2 流利度提升推理的效果

各种推理方法的有效性通过比较表中的结果可以观察到2按栏。 与正常的seq2seq相比推断和seq2seq（+ LM）基线，流利程度提升推论平均带来0.14和F0：5分别获得0.18的增益，这是一个显着的6改进，展示多轮通过流利度增强推论进行编辑是有效的。

采用我们最好的系统（表中的最后一行2）例如，在1,312个句子中CoNLL-2014数据集，seq2seq推断与浅层融合LM编辑566个句子。 相反，流利度提升推断还可以编辑23第二轮推理中的句子，改进F0：5从52.59至52.72。

5.3.3 GEC最先进的结果

现在，我们回答本节中提出的最后一个问题5.2通过测试我们的方法是否达到以下状态：

最新的结果。

我们首先比较一下我们的最佳模型-双升压通过流畅的推理学习（+母语）和浅层融合LM –性能卓越的GEC在CoNLL-2014数据集上评估的系统：

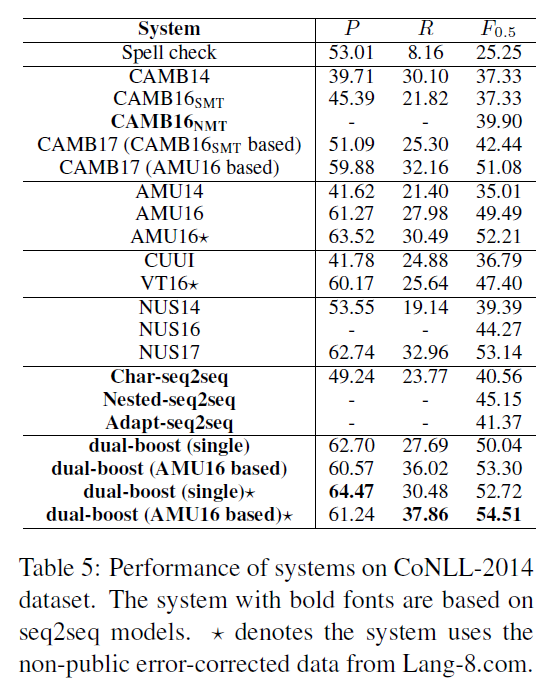


表5显示了对产品的评估结果。CoNLL-2014数据集。 不使用非公开来自Lang-8.com的训练数据。模型获得50.04 F0：5，表现不佳其他seq2seq模型，且仅次于CAMB17（基于AMU16）和NUS17。 这应该但是请注意，CAMB17和NUS17实际上是基于SMT构建的重新排名GEC系统（AMU16的框架）； 从而，它们是合奏模型。 当我们建立方法时在AMU16的顶部（即，我们以AMU16的输出作为输入到我们的GEC系统进行编辑在其输出之上），我们达到了53.30 F0：5得分了。 随着介绍非公开培训数据，我们的整体系统便获得52.72和54.51 F0：5得分，这是一种状态-CoNLL-2014数据集的最新结果7。

此外，我们评估了我们在JFLEG上的方法语料库（Napoles et al。，2017）。 JFLEG是最新的已发布用于GEC评估的数据集，其中包含1,501个句子（开发集中为754，测试集中为747）。为了测试我们的方法的泛化能力，我们评估我们用于CoNLL评估的单个模型（见表5）在JFLEG上无需重新调整。

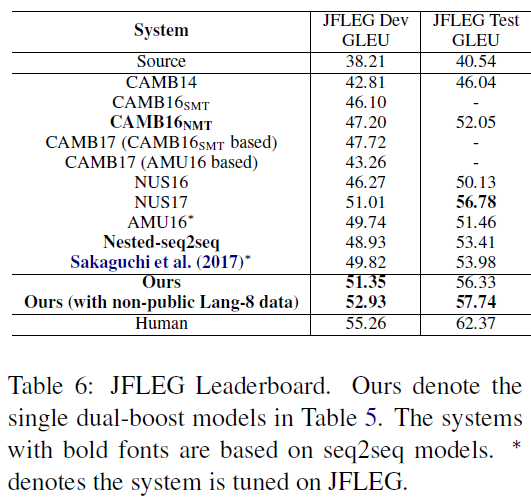


表6显示了JFLEG排行榜。 代替在M2得分中，JFLEG使用GLEU（Napoles等，2015年）作为评估指标，以流利为导向基于BLEU变体的GEC指标（Papineni et al。，2002）并具有多个优势在M2上进行GEC评估。 据观察，我们的单个模型在JFLEG，表现优于大多数CoNLL-2014表现最佳的系统并产生最新状态在这个基准上的艺术成果8，展示了我们的模型被很好地概括并执行稳定地在多个数据集上。

1. 结论

我们提出了一种新颖的流利性促进学习和克服限制的推理机制以前的神经GEC模型。 我

们的建议流利性提升学习充分利用了两种错误纠正方法数据和本机数据，大大改善了正常seq2seq学习的表现，而流利度提升推理则利用GEC的特征，以逐步改善通过多轮推理使句子流利。强大的学习和推理机制使seq2seq模型能够达到CoNLL-2014和JFLEG的艺术成果基准数据集。