文章编号:1671-4598(2021)01-0150-04 DOI:10.16526/j.cnki.11—4762/tp.2021.01.031 中图分类号:TP391.2 文献标识码:A

基于 DNN 与规则学习的机器翻译算法研究

陶媛媛[1],陶 丹[2]
(1.西安交通大学城市学院,西安 710000;2.西安市曲江第一中学,西安 710000)

摘要:通过以目标信息为指导的卷积体系总结相关源信息,提出了一种系统的处理语言方法;利用在解码过程中使用不同的引导信号,经过特殊设计的卷积+门控体系结构可以查明与预测目标单词相关的源句子部分,并将其与整个源句子的上下文融合在一起形成统一表示形式;研究结果表明,模型将表示形式与目标语言单词一起馈入深度神经网络(DNN),形成更强大的神经网络联合模型(NNJM);通过两个 NIST 汉英翻译任务的实验验证,在相同设置下,tagCNN 和 inCNN 在 Dep2Str 基线上的改善幅度分别为+1.28,+1.75 BLEU,所提出的模型分别优于 NIST MT04 和 MT05 的平均值+0.36,+0.83 BLEU,比传统 DNN 机器翻译平均提高了+1.08 BLEU 点;模型为统计机器翻译研究提供了新思路。

关键词:深度神经网络;机器翻译;神经网络联合模型;卷积

Research on Machine Translation Algorithm Based on DNN and Rule Learning

Tao Yuanyuan[1], Tao Dan[2]
(1. City College Xi'an Jiao Tong University, Xi'an 710000, China;
2. Xi'an Qu Jiang No. 1 High School, Xi'an 710000, China)

Abstract: Based on the convolution system guided by the target information, this paper summarizes the relevant source information and proposes a systematic processing language method. By using different guide signals in the decoding process, the specially designed convolution + gating architecture can identify the source sentence part related to the predicted target word, and fuse it with the context of the whole source sentence to form a unified representation. The results show that the model feeds the representation and the target language words into DNN to form a stronger neural network joint model (NNJM). The experimental results of two NIST Chinese—English translation tasks show that under the same settings, the improvement of tagCNN and inCNN on the Dep2str baseline is + 1.28 and + 1.75 BLEU, respectively. The proposed model is superior to the average of NIST MT04 and MT05 + 0.36 and + 0.83 BLEU, respectively, which is + 1.08 BLEU higher than the traditional DNN machine translation. The model provides a new way for statistical machine translation research.

Keywords: deep neural network; machine translation; neural network joint model; convolution

0 引言

随着翻译需求的增加,信息技术与语言学理论以及人工智能研究中自然语言理解模型的蓬勃发展,使得机器翻译逐渐受到了各领域专业技术人员瞩目[1-3]。Liu 等[1]报道并系统地探索使用针对统计机器翻译(SMT)的深度(多层)神经网络(DNN)学习新功能的可能性。为了解决深度信念网络中特征学习的输入原始特征太简单,每个短语对的 4 个短语特征有限的问题,作者将一些简单但有效的短语特征作为新的 DNN 特征学习的输入特征进行了调整和扩展,并且这些特征已显示出 SMT 的显著改进,例如短语对相似性等。此外,在传统 SMT 和神经机器翻译(NMT)中,学习源语言的连续空间表示法已经引起了广泛的关注。目前已经提出了各种模型,主要是基于神经网络的模型来表示源句子,主要用作编码器—解码器框架中的编码器部分。在解码过程中仅对源句子的“相关”部分进行编码,Devlin 等[4]提出的神经网络联合模型(NNJM),该模型扩展了 n 元语法目标语言模型,通过额外增加源句的固定长度窗口,实现统计机器翻译的最新性能。实验结果表明,与比较模型算法相比,基于深度神经网络与规则学习的统计机器翻译模型具有更好的效果,更快的收敛速度和更高的可靠性。深度神经网络作为一种新的机器学习方法,可以自动学习抽象特征表示并在输入和输出信号之间建立复杂的映射关[5-6]。

本文提出了一种新的卷积架构,以动态编码源语言中的相关信息。该模型涵盖了整个源句子,可以在目标语言的信息指导下有效地找到并适当地总结相关部分。利用解码过程中的引导信号,经过特殊设计的卷积体系结构可以

收稿日期:2020-05-17; **修回日期**:2020-06-09。
基金项目:陕西省教育厅专项科研计划项目(No.18JK1012)。
作者简介:陶媛媛(1986-),女,陕西商洛人,硕士,讲师,主要从事跨文化交际及英语翻译教学方向的研究。
引用格式:陶媛媛,陶 丹.基于 DNN 与规则学习的机器翻译算法研究[J].计算机测量与控制,2021,29(1):150-153.

查明与预测目标单词相关的源语句部分，并将其与整个源语句的上下文融合在一起以形成统一的表示形式。将之与目标词一起馈入深度神经网络（DNN）。联合模型的两个变体 tagCNN 和 inCNN，并在解码过程中使用了不同的指导信号。联合模型集成到最新的依存关系的字符串翻译系统中用以评估其有效性，为统计机器翻译研究提供了新的切入点。

1　方法论

1.1　联合语言模型

CNN 编码器的联合模型在图 1 中进行了说明，其中包括 CNN 编码器，即 tagCNN 或 inCNN，以表示源语句中的信息，以及基于 NN 的模型 CNN 编码器的表示和目标句子中的历史单词作为输入来预测下一个单词。

在联合语言模型中，e_n 表示目标词的概率，给定前 k 个目标词 $\{e_{n-k},\cdots,e_{n-1}\}$ 和 CNN 编码器对源句子 S 的表示为：

$$\text{tagCNN}: p(e_n \mid \varphi_1(S,\{a(e_n)\}),\{e\}_{n-k}^{n-1})$$

$$\text{inCNN}: p(e_n \mid \varphi_2(S,\{h(e_n)_{n-k}^{n-1}\}),\{e\}_{n-k}^{n-1})$$

这里 $\varphi_1(S,\{a(e_n)\})$ 表示 tagCNN 给定的表示形式，其中源词的索引 $a(e_n)$ 与目标词 e_n 对齐，而 $\varphi_2(S,\{h(e_n)_{n-k}^{n-1}\})$ 表示来自 inCNN 的信号 $h(e_n)_{n-k}^{n-1}$。

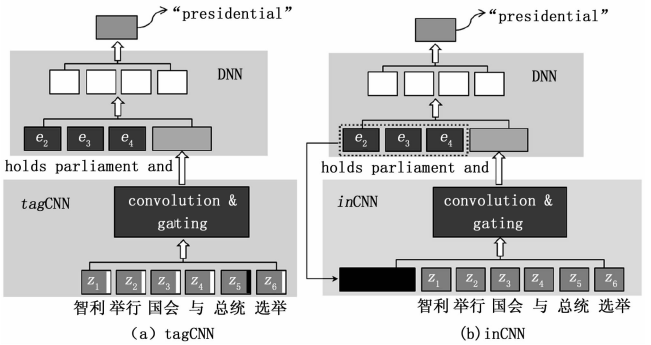

图 1　基于 CNN 编码器的联合 LM 的示意图

例如图 1 的示例中，翻译的中文句子如下。

中文：智利 举行 国会 与 总统 选举；

Pinyin：ZhiLi JuXing GuoHui Yu ZongTong XuanJu。

转换成英语，在评估目标语言顺序“holds parliament and presidential”时，以“holds parliament and”作为进行词（假设为 4－gram LM），而“presidential”的从属源词 1 为“zongtong”（由单词对齐方式确定），tagCNN 产生 $\varphi_1(S,\{4\})$（“Zongongong”的索引为 4），而 inCNN 产生 $\varphi_2(S,\{h(\text{holds parliament and})\})$。之后，DNN 组件将“holds parliament and”和 φ_1 或者 φ_2 作为输入，以给出下一个单词的条件概率，例如 p（“presidential”$\mid \varphi_{1|2}$，{holds，parliment，and}）。

1.2　卷积模型

1.2.1　通用 CNN 编码器

从卷积编码器的通用架构开始[7-8]，然后继续将 tagCNN 和 inCNN 作为两个扩展。

通用 CNN 编码器的基本架构如图 2 所示，其固定架构由如下 6 层组成。

第 0 层：输入层，采用嵌入矢量形式的单词。工作中将句子的最大长度设置为 40 个单词。对于短于此的句子，常在句子开头放置零填充。

第 1 层：第 0 层之后的卷积层，窗口大小为 3。引导信号被注入到该层中作为“引导版本”。

第 2 层：第 1 层之后的本地门控层，仅对大小为 2 的非相邻窗口中的特征图进行加权求和。

第 3 层：在第 2 层之后的卷积层，开始执行另一个卷积，窗口大小为 3。

第 4 层：在第 3 层上对功能图执行全局选通。

第 5 层：完全连接的权重，将第 4 层的输出映射到该层作为最终表示。

如图 2 所示，第 1 层中的卷积在单词的滑动窗口（宽度 k_1）上运行，并且窗口的类似定义会延续到更高的层。形式上，对于源句子输入 $x=\{x_1,\cdots,x_N\}$ 第 L 层上的 f 型特征映射的卷积单位为式（1）：

$$z_i^{(l,f)}(x)=\sigma(w^{(l,f)}\hat{z}_i^{l-1}+b^{(l,f)}),\quad l=1,3,\ f=1,2,\cdots,F_l \tag{1}$$

式中，$z_i^{(l,f)}(x)$ 给出第 L 层中位置 i 的类型为 f 的特征图的输出；$w^{(l,f)}$ 是 L 层上 f 的参数；$\sigma(.)$ 是 Sigmoid 激活函数；$\hat{z}_i^{(l-1)}$ 表示位置 i 处卷积的第 1 层的分段，而 $\hat{z}_i^0=[x_i^T,x_{i+1}^T,x_{i+2}^T]^T$；连接来自句子输入 x 的 3 个单词的向量。

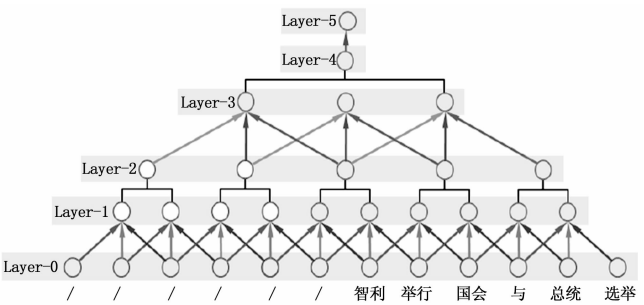

图 2　CNN 编码器示意图

1.2.2　门控

相关文献中 CNN 采用简单的卷积池策略[9]，其中“融合”决策是基于要素图的值。在本质上而言，是一种软模板匹配，适用于诸如分类的任务，但对于保持卷积的合成功能作用不显著，然而，上述功能对于句子建模至关重要。本文使用单独的门控单元从卷积中释放得分函数的占空比，并使其更加注重于句子的合成。

采用两种类型的门控：（1）对于第 2 层，在第 1 层卷积的特征图上采用不重叠窗口的局部门控，以表示段；（2）对于第 4 层卷积，采用全局选通融合所有片段以实现全局表示。并发现，这种选通策略可以显著改善 tagCNN 和 inCNN 的池化性能。

1）本地门控：在第 1 层上，对于每个门控窗口，首先在第 0 层上找到其原始输入（卷积之前），然后将它们合并

为门控网络的输入。例如，对于两个窗口：第 0 层上的单词（3，4，5）和单词（4，5，6），使用由嵌入单词（3，4，5，6）组成的串联矢量作为输入本地门控网络的输入以确定两个窗口（在第 1 层上）的卷积结果的权重，而加权和是第 2 层的输出。

2）全局门控：在第 3 层上，对于每个位置 i 上的特征图(表示为 z_i^3)，全局门控网络分配归一化权重，由式（2）表示：

$$w(z_i^3) = e^{w_s^T z_i^3} / \sum_j e^{w_s^T z_j^3} \tag{2}$$

第 4 层的门控表示由加权和给出 $\sum_i w(z_i^3) z_i^3$。

1.2.3 CNN 编码器训练

CNN 编码器训练过程与神经网络语言模型的训练过程相同，除了使用并行语料库而不是单语语料库外，力求最大程度地提高训练样本的对数似然性，并为平行语料库中的每个目标单词提供一个样本[10]。优化是通过常规的反向传播来实现的，该实现是使用小批量的随机梯度下降实现的。

1.2.4 inCNN

与 tagCNN 直接将关联词的位置告知 CNN 编码器不同，inCNN 将有关目标端中进行词的信息发送给卷积编码器，以帮助检索与预测下一个词相关的信息。这本质上是模型的一种特殊情况，来自后续单词的信息被表示为 $h(e_n)_{n-k}^{n-1}$，并被注入源语言句子中的每个卷积窗口。

通过使用 DNN 转换单词 $\{e_{n-k},\cdots,e_{n-1}\}$ 从单词嵌入连接的向量；包含 $h(e_n)_{n-k}^{n-1}$，通过卷积和门控层，inCNN 能够做到以下几点：(1) 检索源句子的相关片段；(2) 组成并将检索的片段转换为 DNN 在预测目标语言中的单词时可识别的表示形式。与 tagCNN 不同的是，inCNN 使用来自行进单词的信息，因此在 tagCNN 的增强联合语言模型中提供了补充信息。使用基于 tagCNN 的功能和基于 inCNN 的功能进行解码时，已通过经验进行了验证，因此有较大的改进。

2 联合模型解码

本文的联合模型纯粹是词汇化的，因此可以作为功能集成到任何 SMT 解码器中[11-12]。对于分层 SMT 解码器，采用了 Devlin 等人提出的集成方法。正如从用于执行分层解码的 n－gram 语言模型继承的那样，应将每个组成部分中最左边和最右边的 $n-1$ 个单词存储在状态空间中。通过扩展状态空间，以包括这些边词中每个词的关联源词的索引。对于对齐的目标词，将其对齐的源词作为其关联的源词。对于未对齐的单词，同样使用 Devlin 等人采用的从属关系启发式方法。在本文中，将联合模型集成到最新的依赖项到字符串机器翻译解码器中，作为案例研究来测试提出的方法的有效性。本节简要描述依赖项到字符串的转换模型与 MT 系统。

本文使用一般的对数线性框架。令 d 为将源依赖关系树转换为目标字符串 e 的推导。d 的概率定义为式（2）：

$$P(d) \propto \prod_i \varphi_i(d)^{\lambda_i} \tag{2}$$

其中：φ_i 是在导数上定义的特征，λ_i 是相应的权重。其中解码器包含以下功能。

1）基准功能：

HDR 规则的翻译概率 $P(t \mid s)$ 和 $P(s \mid t)$；HDR 规则的词法翻译概率 $P_{LEX}(t \mid s)$ 和 $P_{LEX}(s \mid t)$；伪翻译规则惩罚 exp（－1）；目标词惩罚 exp（$\mid e \mid$）；n 元语法模型 $P_{LM}(e)$。

2）提出功能：

n－gram tagCNN 联合语言模型 $P_{LEX}(e)$；CNN 联合语言模型 $P_{LEX}(e)$ 中的 n－gram。

3 实际应用

3.1 数据获取与预处理

1）数据获取：本文的培训数据是从 LDC 数据中提取的[13]。仅保留源对部分长度不超过 40 个单词的句子对，该句子对覆盖了 90％以上的句子。双语训练数据由 221 k 句子对组成，其中包含 500 万个中文单词和 680 万个英文单词。经过长度限制过滤后，开发集为 NIST MT03（795 个句子），测试集为 MT04（1499 个句子）和 MT05（917 个句子）。

2）预处理：使用 GIZA ＋＋在语料库上使用“增长－确定－最终和”平衡策略[14]，在两个方向上获得单词对齐。并采用 SRI 语言建模工具包在英语 Gigaword 语料库（3.06 亿个单词）的新华语部分上训练了经过改进的 Kneser－Ney 平滑处理的 4－gram 语言模型。随后使用 Stanford Parser 将中文句子解析为映射依赖树。

3）NN 的优化：在训练神经网络时，将源词汇和目标词汇限制为中文和英文最常见的 20 k 单词，分别覆盖两个语料库的 97％和 99％的单词。所有词汇和单词都映射到特殊令牌 UNK 上。并使用随机梯度下降训练联合模型，将最小批量的大小设置为 500。所有联合模型均使用 3 字目标（即 4－gram LM）。inCNN 的词嵌入维数和信号 $h(e_n)_{n-k}^{n-1}$ 为 100。对于卷积层（第 1 层和第 3 层），应用了 100 个滤波器。CNN 编码器的最终表示形式是尺寸为 100 的向量。联合模型的最终 DNN 层是标准的多层感知器，并且顶层具有 softmax。

4）指标：本文使用不区分大小写的 4－gram NIST BLEU3 作为评估指标，在提出的模型和两个基准之间进行带有标志检验的统计学显著性检验。

3.2 设置模型比较

将 tagCNN 和 inCNN 联合语言模型用作依赖项到字符串基线系统（Dep2Str）的附加解码功能，并将它们与具有 11 个源上下文词的神经网络联合模型进行比较。除了全局设置外，还使用具有默认配置的开源工具包。由于本文提出的 tagCNN 和 inCNN 模型是从源到目标和从左到右的（在目标侧），因此，仅采用源到目标和从左到右的 NNJM 类型进行比较。以下将这种类型的 NNJM 称为 BBN－JM。尽管 Devlin 等人中的 BBN－JM 最初是在基于分层短语的 SMT 和字符串到依赖关系 SMT 中进行测试的，但它可以很容易地集成到 Dep2Str 中。

3.3 实验结果与分析

表 1 给出了不同模型的主要结果。在进行更详细的比较之前，可获得以下信息：

1）基线 Dep2Str 系统的 BLEU 比基于开源短语的系统 Moses 高 0.5+。

2）与 Dep2Str 相比，BBN-JM 的得分约为+0.92 BLEU。

表 1　不同模型运行结果比对

系统	MT04/BLEU	MT05/BLEU	Average/BLEU
Moses	34.35	32.43	33.53
Dep2Str	34.78	32.56	33.54
+BBN-JM	36.54	33.64	33.22
+CNN	33.54	34.54	34.54
+tagCNN	33.77	36.54	35.54
+inCNN	36.54	34.56	35.56
+tagCNN+inCNN	35.54	34.56	35.76

从表 1 可以明显看出，在相同设置下，tagCNN 和 inCNN 在 Dep2Str 基线上的改善幅度为+1.28 和+1.75 BLEU，在相同设置下，其 BBN-JM 分别优于 NIST MT04 和 MT05 的平均值+0.36 和+0.83 BLEU。这些表明 tagCNN 和 inCNN 可以在解码中分别提供区分性信息。值得注意的是，inCNN 似乎比单词对齐（GIZA ++）建议的附属单词更具信息性。因此，推测这是由于以下两个事实才成立的：

1）inCNN 避免了在已经学习的单词对齐方式中错误和伪影响的传播；

2）inCNN 中的指导信号提供补充信息以评估翻译。

此外，当将 tagCNN 和 inCNN 都用于解码时，可以进一步在 BBN-JM 上的获胜余量提高到+1.08 BLEU 点。

通用 CNN 还能在 BLEU 上获得类似于 BBN-JM 的增益，因为，通用 CNN 编码整个句子，并且表示形式通常应远离联合语言模型的最佳表示形式。因此，可能是由于 CNN 对该句子产生了相当有益的总结，这弥补了其在分辨率和源词相关部分上的某些损失。换言之，tagCNN 和 inCNN 中的引导信号对于基于 CNN 的编码器的功能至关重要，这可以从通用 CNN，tagCNN 和 inCNN 获得的 BLEU 得分之间的差异中看出。对于有了来自已经习得的单词对齐的信号，tagCNN 可以比其通用对应词获得+0.25 BLEU，而对于 inCNN，来自目标单词的引导信号，其增益更显著为+0.72 BLEU。

tagCNN 可以进一步受益于在输入中用源语言编码的依赖关系结构。依赖项首词可用于进一步完善 tagCNN 模型。在 tagCNN 中，在输入层中的单词嵌入后附加一个标记位（0 或 1）作为它们是否隶属源词的标记。为了合并依赖头信息，本文扩展了标记规则，为原始 tagCNN 的词嵌入添加了另一个标记位（0 或 1），以指示它是否是附属词的依赖头的一部分。例如，如果 x_i 是相关源词的嵌入，而 x_j 是词 x_i 的依赖头，则 tagCNN 的扩展输入将包含，式（3）：

$$\begin{aligned} x_i^{(AFF,NON-HEAD)} &= [x_i{}^T,1,0]^T \\ x_j^{(NON-AFF,HEAD)} &= [x_j{}^T,0,1]^T \end{aligned} \tag{3}$$

若从属源词是句子的词根，则由于词根没有依赖项头，因此本文仅将 0 作为第二个标记位附加。从表 2 中可以看出，借助依赖项头信息能够在两个测试集上将 tagCNN 平均提高+0.23 BLEU 点。

表 2　tagCNN 模型的 BLEU-4 分数(%)

系统	MT04	MT05	平均
Dep2Str	34.78	32.56	33.54
+tagCNN	33.77	36.54	35.54
+tagCNN_dep	36.54	33.56	35.79

以 inCNN 模型的比较为例，研究了门控策略可在多大程度上改善最大池化的翻译性能。对于使用最大池的 inCNN 实施，本文用大小为 2（简称 2 池）的最大池替换本地门（第 2 层），用 k 个最大池（“第 4 层”）替换全局门（“第 4 层”），其中 k 为 {2；4；8}，然后将 k 池输出的平均值用作第 5 层的最终输入。这样可以确保第 5 层的输入维与具有门控的架构相同。从表 3 可以看出，门控策略可以比最大合并提高 0.34，0.71 BLEU 点的翻译性能。此外还发现 8 池收益率性能要优于 2 池。所以推测这是因为翻译中有用的相关部分主要集中在源句子的几个单词上，可以通过较大的库大小更好地提取这些单词。

表 3　使用门控策略和 k 个最大池实施的 inCNN 模型的 BLEU-4 分数(%)

系统	MT04	MT05	平均
Dep2Str	34.78	32.56	33.54
+inCNN	33.54	34.54	34.54
+inCNN-2-pooling	33.77	34.54	32.54
+inCNN-4-pooling	36.54	31.56	34.56
+inCNN-8-pooling	35.54	38.56	35.76

4　结束语

本文提出了卷积架构，以获取整个源句的引导表示，该卷积架构可用于增强 n-gram 目标语言模型。利用来自目标端的不同指导信号，并设计了 tagCNN 和 inCNN，两者都通过增强字符串对依赖关系的 SMT 进行了测试，其 SMT 比基线高+2.0 BLEU 点。所提出的模型分别优于 NIST MT04 和 MT05 的平均值+0.36，+0.83 BLEU，比传统 DNN 机器翻译平均提高了+1.08 BLEU 点。该模型为统计机器翻译方法研究提供了借鉴。

参考文献：

[1] Liu Lemao，Taro Watanabe，Eiichiro Sumita，et al. Additive neural networks for statistical machine translation [J]. Proceedings of ACL，2013：791-801.

[2] Xiong Deyi，Zhang Min，Li Haizhou. Enhancing language models in statistical machine translation with backward n-grams and mutual information triggers [J]. Proceedings of ACL，2011：1288-1297.

（下转第 158 页）