软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software, 2022,33(3):1141−1152 [doi: 10.13328/j.cnki.jos.006201]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于多覆盖模型的神经机器翻译*

刘俊鹏, 黄锴宇, 李玖一, 宋鼎新, 黄德根

(大连理工大学 计算机科学与技术学院, 辽宁 大连 116024)

通信作者: 黄德根, E-mail: huangdg@dlut.edu.cn

摘 要: 覆盖模型可以缓解神经机器翻译中的过度翻译和漏翻译问题. 现有方法通常依靠覆盖向量或覆盖分数等单一方式存储覆盖信息, 而未考虑不同覆盖信息之间的关联性, 因此对信息的利用并不完善. 针对该问题, 基于翻译历史信息的一致性和模型之间的互补性, 提出了多覆盖融合模型. 首先定义词级覆盖分数概念; 然后利用覆盖向量和覆盖分数存储的信息同时指导注意力机制, 降低信息存储损失对注意力权重计算的影响. 根据两种覆盖信息融合方式的不同, 提出了两种多覆盖融合方法. 利用序列到序列模型在中英翻译任务上进行了实验, 结果表明, 所提方法能够显著提升翻译性能, 并改善源语言和目标语言的对齐质量. 与只使用覆盖向量的模型相比, 过度翻译和漏翻译问题的数量得到进一步减少.

关键词: 神经机器翻译; 注意力机制; 序列到序列模型; 多覆盖模型; 过度翻译; 漏翻译

中图法分类号: TP183

中文引用格式: 刘俊鹏, 黄锴宇, 李玖一, 宋鼎新, 黄德根. 基于多覆盖模型的神经机器翻译. 软件学报, 2022, 33(3): 1141−1152. http://www.jos.org.cn/1000-9825/6201.htm

英文引用格式: Liu JP, Huang KY, Li JY, Song DX, Huang DG. Multi-coverage Model for Neural Machine Translation. Ruan Jian Xue Bao/Journal of Software, 2022, 33(3): 1141−1152 (in Chinese). http://www.jos.org.cn/1000-9825/6201.htm

Multi-coverage Model for Neural Machine Translation

LIU Jun-Peng, HUANG Kai-Yu, LI Jiu-Yi, SONG Ding-Xin, HUANG De-Gen

(School of Computer Science and Technology, Dalian University of Technology, Dalian 116024, China)

Abstract: The over-translation and under-translation problem in neural machine translation could be alleviated by coverage model. The existing methods usually store the coverage information in a single way, such as coverage vector or coverage score, but do not take the relationship among different coverage methods into consideration, leading to insufficient use of the information. This study proposes a multi-coverage mechanism based on the consistency of translation history and complementarity between different models. The concept of word-level coverage score is defined first, and then the coverage information stored in both coverage vector and coverage score are incorporated into the attention mechanism simultaneously, aiming to reduce the influence of information loss. According to different fusion methods, two models are introduced. Experiments are carried out on Chinese-to-English translation task based on sequence-to-sequence model. Results show that the proposed method could significantly enhance translation performance and improve alignment quality between source and target words. Compared with the model with coverage vector only, the number of over-translation and under-translation problem is further reduced.

Key words: neural machine translation; attention mechanism; sequence-to-sequence model; multi-coverage model; over-translation; under-translation

机器翻译是使用计算机将一种语言(源语言)自动翻译成另一种语言(目标语言)的技术. 近年来, 神经机器翻译(neural machine translation, NMT)经过不断发展和完善, 显示出强大的翻译性能. 现有的神经机器翻译模型通常采用“编码器-解码器”架构, 其中, 编码器和解码器可以采用不同的网络结构, 依据神经网络拓扑结构

* 基金项目: 国家重点研发计划(2020AAA0108004); 国家自然科学基金(61672127, U1936109)
收稿时间: 2020-04-01; 修改时间: 2020-06-17, 2020-09-30; 采用时间: 2020-10-28

的特点, 分为循环神经网络(recurrent neural network)[1−3]、卷积神经网络(convolutional neural network)[4]和自注意力网络(self-attention network)[5]等. 目前, 神经机器翻译的性能已经显著超越传统的统计机器翻译, 成为当前机器翻译领域的前沿热点[6].

虽然大规模训练数据和强大的计算能力使得神经机器翻译的性能大幅提升, 但是神经机器翻译还存在着过度翻译(over-translation)和漏翻译(under-translation)问题. 在基于短语的统计机器翻译模型中, 文献[7]通过将加入输出序列的翻译结果所对应的源语言短语标记为“已翻译”的方式, 确保所有的源语言短语均被翻译覆盖且只被翻译一次. 但是在神经机器翻译模型中没有显式存储历史翻译信息的结构, 并且每一次解码过程都需要所有源语言词语的参与, 因此极易出现过度翻译和漏翻译问题.

为了向神经机器翻译模型中增加覆盖信息, 文献[8]借鉴统计机器翻译中的覆盖思想提出了覆盖模型(coverage model). 设计一个覆盖向量(coverage vector, CV)记录解码过程中的翻译历史信息, 并指导注意力权重计算, 降低已翻译词语的权重, 使注意力机制更多地关注未翻译词语. 类似地, 文献[9]为源语言词语设置全覆盖编码向量(full coverage embedding vector), 在解码过程中不断削减已翻译词语的编码向量, 以此降低其在未来解码中的作用. 文献[10]利用循环注意力机制(recurrent attention mechanism)给注意力单元提供更多的重调序信息, 并通过条件解码器(conditioned decoder)减少重复翻译. 文献[11]增加了两个额外的循环神经网络分别记录翻译过程中的历史(past)和未来(future)信息, 并利用该信息指导注意力机制和解码状态.

覆盖度还可以与解码算法相结合, 用于在已生成的译文中筛选对原文忠实度最高的结果. 文献[12]通过覆盖惩罚(coverage penalty)和长度归一化(length normalization)改进束搜索(beam search)算法, 使模型在选择翻译结果时考虑该翻译结果对源语言信息的覆盖程度, 避免偏向句子长度更短的翻译结果. 文献[13]提出将覆盖分数(coverage score, CS)引入每一次束搜索过程, 以此减少搜索错误, 并且使覆盖分数的计算适用于源语言词汇和目标语言词汇的多种映射关系.

在翻译过程中, 源语言上下文通常影响翻译的忠实度, 而目标语言上下文则与译文的流利度有关. 基于这种思路, 文献[14]提出了上下文门(context gate)方法, 根据源语言和目标语言上下文的重要程度, 动态控制两种上下文信息在生成目标词语时的影响比重. 该方法可以与覆盖机制相结合, 能够在改善翻译结果对源语言的覆盖度的同时, 提升译文的流利度. 文献[15]通过解码历史增强注意力机制(decoding-history enhanced attention mechanism)建立所有源语言词和目标语言词的结构关系, 使神经机器翻译模型更好地选择源语言端和目标语言端的信息.

此外, 文献[16]针对源语言中被漏翻译的词语特点展开研究, 发现在源语言中, 翻译熵(translation entropy)越大的词语越容易被漏翻译, 并据此设计了一种粗粒度到细粒度(coarse-to-fine)框架, 分别解决句子级和词级的训练和翻译问题, 从而减少熵高词语的漏翻译情况.

虽然上述方法均能在一定程度上缓解神经机器翻译中的过度翻译和漏翻译问题, 但该问题依然不能被完全避免. 如表 1 所示, 在基线系统的翻译结果中, 源语言句子中的“事关全局(related to the overall situation)”被遗漏翻译, 且“体制(system)”被重复翻译一次. 而引入覆盖模型后, 上述问题并未得到明显改正, “事关全局”虽然被未被遗漏但却被错误地翻译为“in the world”, 而“体制”则被漏翻译.

表 1 过度翻译和漏翻译示例

文本类别	例句
源语言句子	深化国家监察体制改革是**事关全局**的重大政治**体制**改革.
参考译文	Deepening the reform of the state oversight system is a major political **structural** reform **related to the overall situation**.
基线系统	Deepening the reform of the national oversight system is a major political **system system** reform.
覆盖模型	Deepening the reform of the state supervisory system is a major political reform **in the world**.
多覆盖融合模型	Deepening the reform of the country’s supervision system is a major political **system** reform **to the overall situation**.

注: 基线系统采用基于循环神经网络的 Seq2Seq 模型, 覆盖模型是根据文献[8]在基线系统上的复现

由此可见, 现有的覆盖模型对覆盖信息的记录和使用并不完善. 可能的原因是, 覆盖向量在使用 GRU 网

络更新时存在信息损失, 导致注意力权重分配不准确, 进而产生了重复翻译和漏翻译现象. 针对上述问题, 本文提出一种多覆盖融合的神经机器翻译模型. 首先定义一种词级覆盖分数, 用于记录源语言词语在解码过程中的注意力累积情况; 而后在解码阶段, 利用覆盖向量和覆盖分数所记录的覆盖信息同时指导注意力权重计算, 使两种覆盖信息互相补充, 从而最大限度减少信息损失.

本文的主要创新点包括两个方面: (1) 定义了词级覆盖分数概念, 并利用覆盖分数指导注意力机制的权重计算, 拓展了覆盖分数在神经机器翻译模型中的使用方式; (2) 提出了多种覆盖信息的融合模型及两种实现方式, 并利用实验验证了方法的可行性和有效性. 在 CWMT2018 中英数据集上的实验结果表明, 多覆盖融合方法能显著提升翻译质量, 并在覆盖模型基础上, 进一步减少过度翻译和漏翻译现象.

1 研究背景

1.1 基于循环神经网络的神经机器翻译模型

基于注意力机制的神经机器翻译模型通常采用“编码器-解码器”结构. 编码器将源语言句子编码成隐层向量表示, 解码器根据编码器的输出逐字预测目标端句子的单词序列. 给定源端输入句子 $X=\{x_1,x_2,\ldots,x_m\}$, 神经机器翻译模型对目标端句子 $Y=\{y_1,y_2,\ldots,y_n\}$的条件概率 $P(Y|X)$进行建模.

$$P(Y\mid X)=\prod_{j=1}^{n}P(y_j\mid y_{<j},X) \tag{1}$$

具体来说, 若当前已生成的目标端输出为$\{y_1,y_2,\ldots,y_{j-1}\}$, 则生成下一个目标词语 y_j 的概率计算公式为

$$P(y_j|y_{<j},X)=g(y_{j-1},\boldsymbol{t}_j,\boldsymbol{s}_j) \tag{2}$$

其中, $g(\cdot)$是非线性函数, $\boldsymbol{t}_j$ 是 j 时刻解码器的隐层状态, $\boldsymbol{s}_j$ 是将编码器的所有隐层状态加权得到的源语言上下文向量. $\boldsymbol{t}_j$ 和 $\boldsymbol{s}_j$ 的计算公式如下.

$$\boldsymbol{t}_j=f(y_{j-1},\boldsymbol{t}_{j-1},\boldsymbol{s}_j) \tag{3}$$

$$\boldsymbol{s}_j=\sum_{i=1}^{m}a_{ij}\cdot\boldsymbol{h}_i \tag{4}$$

$$a_{ij}=\frac{\exp(e_{ij})}{\sum_{k=1}^{m}\exp(e_{kj})} \tag{5}$$

$$e_{ij}=ATT(\boldsymbol{t}_{j-1},\boldsymbol{h}_i)=v_a^T\tanh(W_a\boldsymbol{t}_{j-1}+U_a\boldsymbol{h}_i) \tag{6}$$

其中,

- $\boldsymbol{h}_i$ 是源语言词语 x_i 经编码器编码后得到的隐层向量, 由于在实际中常使用双向循环神经网络作为编码器, 因此 $h_i=[\vec{\boldsymbol{h}}_i;\overleftarrow{\boldsymbol{h}}_i]$, 其中, $\vec{\boldsymbol{h}}_i$ 和 $\overleftarrow{\boldsymbol{h}}_i$ 分别是前向和后向循环神经网络的隐层状态, $[\cdot;\cdot]$表示拼接操作;
- a_{ij} 为注意力权重, 表示目标端词汇 y_j 与源语言词语 x_i 之间的关联度;
- $ATT(\cdot)$是计算注意力权重的匹配函数;
- v_a^T, W_a 和 U_a 为权重矩阵.

由于注意力机制的引入, 源语言上下文向量 $\boldsymbol{s}_j$ 不再是一个单一固定的向量, 而是随着不同时刻 a_{ij} 的不同而变化, 从而使解码器在不同时刻对源语言句子中各个词语的信息考量有所侧重. $f(\cdot)$通常使用长短期记忆网络(long-short term memory, LSTM)或门控循环单元(gated recurrent unit, GRU), 其网络结构可采用单层或多层, 并且多层网络结构的性能比单层网络有更加显著的提升[12]. 多层堆叠长短期记忆网络(stacked long-short term memory)中, 第 i 层和 i+1 层状态的计算公式如下所示.

$$c_t^i,h_t^i=LSTM_i(c_{t-1}^i,h_{t-1}^i,x_t^{i-1}) \tag{7}$$

$$x_t^i=h_t^i \tag{8}$$

$$c_t^{i+1},h_t^{i+1}=LSTM_{i+1}(c_{t-1}^{i+1},h_{t-1}^{i+1},x_t^i) \tag{9}$$

其中, $LSTM_i$ 和 $LSTM_{i+1}$ 分别表示第 i 和 i+1 层 LSTM, x_t^i, c_t^i 和 h_t^i 分别表示 t 时刻 $LSTM_i$ 的输入、记忆单元状态和隐层状态.

最后, 在训练集 $\{(x^p, y^p)\}_{p=1}^J$ 上, 利用极大似然估计对目标函数的参数 θ 进行迭代训练, 目标函数如公式(10)所示.

$$L(\theta) = -\frac{1}{J}\sum_{p=1}^{J}\sum_{t=1}^{n}\log P(y_t^p \mid y_{<t}^p, x^p) \tag{10}$$

1.2 覆盖向量

与统计机器翻译模型中双语词语的硬对齐(hard-alignment)关系不同, 神经机器翻译模型的注意力机制提供的是一种软对齐(soft-alignment)方法, 因此难以对覆盖机制进行建模. 文献[8]提出通过覆盖向量显式存储解码过程中历史信息的覆盖模型, 利用覆盖向量所存储的信息指导注意力评分过程, 从而使注意力机制更多地关注未翻译的词语, 其模型结构如图 1 所示.

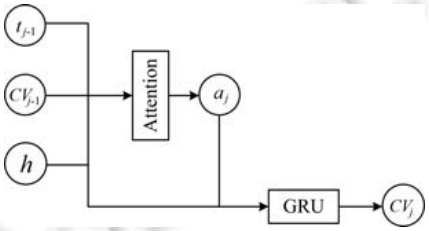


图 1 基于覆盖的注意力模型结构

加入覆盖向量指导后, 注意力权重的计算方法由公式(6)修改为公式(11).

$$e_{ij} = ATT(\boldsymbol{t}_{j-1}, \boldsymbol{h}_i, CV_{i,j-1}) = v_a^T \tanh(W_a\boldsymbol{t}_{j-1} + U_a\boldsymbol{h}_i + V_aCV_{i,j-1}) \tag{11}$$

其中, V_a 为权重矩阵, $CV_{i,j-1}$ 表示 j 时刻前源语言词语 x_i 对应的覆盖向量.

在每一步解码完成后, 利用 GRU 网络对每个源语言词语覆盖向量中存储的覆盖信息进行更新, 更新方法如公式(12)所示.

$$CV_{i,j}=GRU(CV_{i,j-1}, a_{ij}, \boldsymbol{h}_i, \boldsymbol{t}_{j-1}) \tag{12}$$

1.3 覆盖分数

文献[13]提出了表示源语言句子翻译程度的覆盖分数概念. 给定一个翻译句对(X,Y), 将源语言词语 x_i 的覆盖度定义为所有目标词语对该源语言词语的注意力权重之和, 如公式(13)所示.

$$coverage_{x_i} = \sum_{j=1}^{|Y|} a_{ij} \tag{13}$$

在此基础上, 利用所有源语言词语的覆盖度表示源语言句子的覆盖分数, 计算公式如公式(14)所示.

$$cs(X,Y) = \sum_{i=1}^{|X|}\log\max(coverage_{x_i}, \beta) \tag{14}$$

其中, β为可调参数. 最后, 将模型预测的条件概率与译文的覆盖分数线性组合, 使模型在选择译文时兼顾对源语言句子翻译覆盖程度. 改进后的评价函数如公式(15)所示.

$$score(X,Y)=a\cdot\log P(Y|X)+b\cdot cs(X,Y) \tag{15}$$

其中, $\log P(Y|X)$表示模型预测的条件概率值, a 和 b 是用于平衡条件概率和覆盖分数作用的参数.

2 多覆盖融合模型

虽然覆盖向量和覆盖分数均能显式记录翻译过程中的覆盖信息, 但二者在信息的存储方式和使用方式上都有所不同: 前者以向量的形式对信息进行存储和更新, 并通过指导注意力权重计算的方式传递翻译历史;

而后者以常量的形式进行累加, 并作为评价指标用于翻译结果的选择. 两种方法各有优缺点: 覆盖向量存储的信息抽象程度更高, 但在利用 GRU 网络进行更新时, 由于重置门(reset gate)会自动丢弃一定比例的历史信息, 并且更新门(update gate)也会丢掉一部分新的覆盖信息, 因此可能造成覆盖信息损失; 而覆盖分数虽然表达直观, 但难以确定取值界限, 因而无法直接根据数值大小衡量不同词语的覆盖程度.

由于在任意时刻覆盖向量和覆盖分数中存储的信息具有一致性, 且由上述分析可知二者具备一定的互补性, 因此提出一种将两种方法优点相结合的多覆盖融合模型. 利用覆盖向量和覆盖分数存储的信息同时指导注意力机制, 从而降低信息损失对注意力权重分配的影响. 为了将覆盖分数引入注意力机制, 首先定义了词级覆盖分数概念; 然后, 根据覆盖向量和覆盖分数融合方式的不同提出了层次型和平行型两种多覆盖融合模型. 总体框架如图 2 所示.

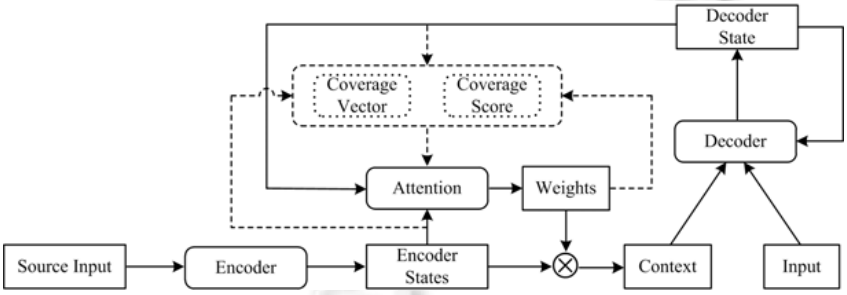


图 2 多覆盖融合的神经机器翻译模型结构

2.1 词级覆盖分数

词级覆盖分数用于表示任意时刻译文结果对每个源语言词语的覆盖程度. 给定源语言句子 X 和 j 时刻目标端的预测序列$\{y_1,y_2,\ldots,y_j\}$, 那么j时刻源语言x_i的覆盖分数定义为当前已生成的所有目标端词语$\{y_1,y_2,\ldots,y_j\}$对源语言 x_i 的注意力权重之和的截断取值, 如公式(16)所示.

$$CS_{ij}=\max\left(\sum_{k=1}^{j}a_{ik},1\right) \tag{16}$$

这里, 截断函数 max(·)的作用主要有以下两个方面.

(1) 改善源语言词语覆盖程度的可比较性

在翻译过程中, 存在某些源语言词语对应多个目标词语的情况. 因此, 这些源语言的注意力分数累加和存在大于 1 的可能, 这使得不同源语言词语之间的覆盖程度难以比较. 经过截断取值之后, 覆盖分数 CS_{ij} 的取值范围设定在[0,1]之间, 那么对于所有覆盖分数取值为 1 源语言词语, 认为该词已被翻译覆盖; 而覆盖分数小于 1 的词语, 在后续解码过程中, 仍有继续被翻译覆盖的可能.

(2) 借鉴硬对齐思想并融入注意力机制

将源语言词语的覆盖分数取值控制在[0,1]的区间内, 更容易直观地表示当前翻译结果对源语言的覆盖情况和未来翻译过程中仍需被注意力机制关注的, 因此便于融入注意力机制并指导注意力权重的计算.

2.2 层次多覆盖模型

层次多覆盖模型(hierarchical multi-coverage, HMC)主要基于覆盖向量和覆盖分数存储信息的一致性原理, 模型结构如图 3 所示, 首先, 利用覆盖向量指导的注意力模型计算当前时刻分配给所有源语言词语的注意力权重; 然后, 利用覆盖分数存储的信息对注意力权重进行校验和重新分配, 从而减少由于覆盖信息丢失而造成注意力权重分配错误的现象. 层次多覆盖模型注意力权重的计算方法如公式(17)–公式(19)所示.

$$\tilde{e}_{ij}=ATT(\boldsymbol{t}_{j-1},\boldsymbol{h}_i,CV_{i,j-1}) \tag{17}$$

$$\tilde{a}_{ij}=\frac{\exp(\tilde{e}_{ij})}{\sum_{k=1}^{m}\exp(\tilde{e}_{kj})} \tag{18}$$

$$a_{ij}=\tilde{a}_{ij}\odot g(CS_{i,j-1}) \tag{19}$$

其中, ⊙表示按位相乘, $g(\cdot)$为注意力权重再分配函数. 假设 j 时刻前源语言词语 x_i 的覆盖分数为 $CV_{i,j-1}$, j 时刻覆盖模型计算得到源语言词汇 x_i 的注意力权重为 $\tilde{a}_{ij}$. 理论上, 函数 $g(\cdot)$应具有以下功能.

(1) 当 $CS_{i,j-1}$ 的值较大时, 即 j 时刻前的翻译结果对 x_i 的覆盖程度较高, 在后续解码时, 应将注意力更多地分配给其他源语言词语. 因此, 函数 $g(\cdot)$应相对削减 $\tilde{a}_{ij}$ 的取值, 从而减少对该源语言词语的关注.

(2) 当 $CS_{i,j-1}$ 的值较小时, 即 j 时刻前的翻译结果对 x_i 的覆盖程度较低, 在后续解码时, 应给予该源语言词语更多的关注. 因此, 函数 $g(\cdot)$应相对增加 $\tilde{a}_{ij}$ 的取值, 使其获得比其他源语言词语更高的注意力权重.

由于 $CS_{ij}\in[0,1]$, 因此 j 时刻时源语言词汇 x_i 仍被允许的覆盖程度可以用公式(20)表示.

$$g(CS_{i,j-1})=\mathbf{1}-CS_{i,j-1} \tag{20}$$

其中, $\mathbf{1}$ 表示全 1 向量. 该函数不但满足上述要求, 而且形式简单. 但考虑到源语言词语和目标语言词语在实际中常常存在“一对多”的映射关系, 在这种情况下, 覆盖分数 $CS_{i,j-1}$ 过大会使得 $g(CS_{i,j-1})$趋近于 0, 从而导致最终得到的注意力权重 a_{ij} 趋近于 0, 无法满足对齐需要. 为了使最终的权重分配平滑并适用于多种对齐关系, 将公式(18)和公式(19)的线性组合作为最终注意力权重的计算方法, 如公式(21)所示.

$$a_{ij}=\mu\cdot\tilde{a}_{ij}+(1-\mu)\cdot a_{ij} \tag{21}$$

其中, μ为调和参数, 用来控制两种注意力分数的作用比例. 由公式(21)推导得到公式(19)中 $g(CS_{i,j-1})$的计算公式如下所示.

$$g(CS_{i,j-1})=\mathbf{1}-(1-\mu)\cdot CS_{i,j-1} \tag{22}$$

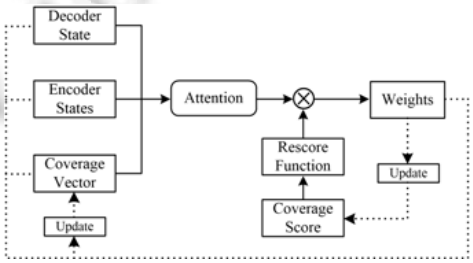


图 3 层次多覆盖模型结构

2.3 平行多覆盖模型

平行多覆盖模型(parallel multi-coverage, PMC)主要利用覆盖向量和覆盖分数的互补性, 将两种覆盖信息同时引入注意力机制. 在计算注意力权重时, 由模型自动选择覆盖向量和覆盖分数这两种覆盖信息的作用比例, 模型结构如图 4 所示. 平行多覆盖模型注意力权重的计算方式如公式(23)所示.

$$e_{ij}=ATT(\boldsymbol{t}_{j-1},\boldsymbol{h}_i,CV_{i,j-1},CS_{i,j-1})=v_a^T\tanh(W_a\boldsymbol{t}_{j-1}+U_a\boldsymbol{h}_i+V_aCV_{i,j-1}+S_aCS_{i,j-1}) \tag{23}$$

其中, S_a 为权重矩阵, $CS_{i,j-1}$ 为 j 时刻前源语言词汇 x_i 的覆盖分数.

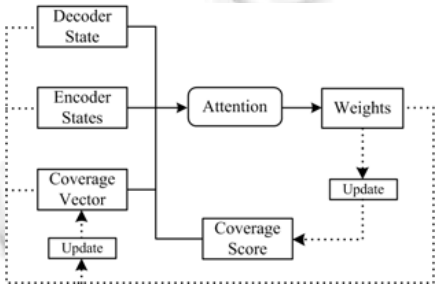


图 4 平行多覆盖模型结构

与层次多覆盖模型利用覆盖分数对注意力权重进行再分配的方式不同, 平行多覆盖模型直接对覆盖向量和覆盖分数的关系进行建模, 并依靠神经网络自动学习和平衡两种覆盖信息的作用, 因此模型结构更为简单.

3 实验与分析

3.1 实验数据及参数设置

实验使用CWMT2018中英新闻领域翻译任务提供的双语平行语料, 从中抽取了约650万个句对作为训练集. 使用 newsdev2017 作为验证集进行参数调优和模型选择, 共包含 2 002 个句子. 选用 newstest2017, cwmt2018 和 newstest2018 这 3 个数据集作为测试集对模型进行验证, 各包含 2 000, 2 481 和 3 981 个句子. 在训练和测试前对语料进行泛化处理. 利用 Niutrans[17]开源工具对中英文语料进行分词处理, 并使用字节对编码(byte-pair-encoding, BPE)[18]进行了子词切分处理.

实验的基线系统(baseline)采用基于循环神经网络的 Seq2Seq[19]模型. 编码器采用 2 层双向长短时记忆网络(bidirectional long-short term memory, BiLSTM), 解码器采用 4 层单向长短时记忆网络(LSTM), 词嵌入维度和隐层神经元个数均设置为 512. 批次大小设置为 32, 词汇表大小为 32 000, 源语言和目标语言句子的最大长度均设置为 50. 使用 Adam 算法[20]进行参数优化, dropout 比率为 0.2, 初始学习率设置为 10^{-4}. 采用宽度为 15 的束搜索算法进行解码, 并引入长度惩罚机制, 惩罚系数设置为 1.3. 实验结果采用大小写不敏感的 BLEU 值[21]作为评价指标, 并根据文献[22]的方法对译文结果进行显著性检验. 保留训练阶段最后 15 个检查点进行模型参数平均[23], 从而得到最终的测试模型.

为了验证多覆盖融合模型的作用, 在实验中引入了两个对照系统.

(1) Coverage Model: 在 Seq2Seq 基线系统上, 参照文献[8]搭建了覆盖模型, 覆盖向量的维度设置为 10. 在层次多覆盖模型中, 调和参数μ设置为 0.5.

(2) Transformer (base): 采用 THUMT[24]开源工具, 模型参数均采用默认值. 测试阶段, 同样将最后 15 个检查点进行参数平均后的模型作为最终的解码模型, 束搜索宽度和长度惩罚系数与 Seq2Seq 模型一致.

3.2 BLEU值评价结果

两种多覆盖融合模型与基线系统和覆盖模型在中英新闻翻译任务上翻译结果的 BLEU 值对比情况见表 2.

表 2 中英翻译结果的 BLEU 值 (%)

系统	newstest2017	cwmt2018	newstest2018	平均值	Δ
Baseline (Seq2Seq)	26.55	27.01	26.79	26.78	–
Coverage model	26.62	27.20	26.97	26.93	+0.15
HMC	**27.26†**	**27.63†**	**27.40†**	**27.43**	**+0.65**
PMC	27.19†	27.32	27.13	27.21	+0.43

注: 表中黑色加粗表示在同一测试集下的最优实验结果, Δ表示其他模型相比于基线系统的 BLEU 值提升, †表示对比基线系统的显著性检验结果(p<0.05)

由表 2 的实验结果可以发现, 相比于基线系统, 层次多覆盖模型和平行多覆盖模型在所有的测试集上均取得了明显的性能提升, 在 3 个测试集上, BLEU 值的平均值分别提高了 0.65 和 0.43 个百分点, 并且提升幅度超过了仅使用覆盖向量的覆盖模型. 上述实验结果证明了多覆盖融合模型在提升神经机器翻译性能方面的作用. 此外, 在 newstest2017 数据集上的实验结果表明, 层次多覆盖模型的性能达到了与 Transformer 模型相近的水平, 见表 3. 但由于 HMC 模型在隐层单元维度和模型深度上与 Transformer 模型存在不同, 因此实验结果相对后者略低. 对比两种多覆盖融合模型, HMC 模型的提升效果比 PMC 模型更加明显, 其原因主要在于在解码阶段, 覆盖分数 CS_{ij} 逐渐接近 **1** 向量并趋于不变, 导致在 PMC 模型的公式(23)中覆盖分数的权重矩阵 S_a 训练难度增大; 而 HMC 模型中, 当覆盖分数 CS_{ij} 逐渐接近 **1** 时, 仅仅相对降低了注意力分数的调节比重,因此受到的影响较小. 在后续实验中, 以层次多覆盖模型为例, 从不同角度进一步具体分析多覆盖融合模型与基线

系统和覆盖模型的性能对比.

表 3 HMC 与 Transformer (base)在 newstest2017 上的实验结果对比

系统	BLEU (%)
Transformer (base)	27.53
HMC	27.26

3.3 源语言句子长度影响

神经机器翻译的译文质量常常会随着源语言句子长度的增加而下降. 为了研究多覆盖融合模型在不同长度句子上的翻译性能, 在实验中, 按照文献[3]的方法, 首先将测试集中的源语言句子按长度分成 6 组, 之后评价层次多覆盖模型与基线系统和覆盖模型在不同长度区间上翻译结果的 BLEU 值. 结果如图 5 所示.

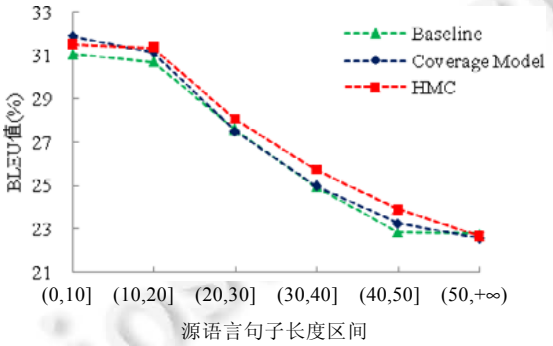


图 5 不同源语言句子长度区间下 BLEU 值的比较

由图 5 所示的实验结果得出以下结论.

(1) 总体上看, 在不同的源语言长度区间上, 层次多覆盖模型产生的译文质量优于基线系统. 其中, 在长度区间(0,50]上的提升效果更为明显.

(2) 与覆盖模型相比, 层次多覆盖模型在常规长度分布区间(10,50]上的翻译质量更好, 但在(0,10]区间上, BLEU 值低于覆盖模型. 通过对语料的进一步分析发现, 在这一长度区间内的部分句子是长句子的切分片段, 它们的结构和句意表达不够完整. 例如, 源语言“把官兵积极性、主动性”对应的参考译文是“enthusiasm, initiative and creativity of officers and soldiers should be given full play”, 在一定程度上影响了模型的翻译性能和结果评价.

(3) 随着源语言句子长度的增加, 基线系统、覆盖模型和层次多覆盖模型的 BLEU 值均呈现明显下降趋势. 当句子长度大于 50 时, 3 种方法的 BLEU 值相差并不明显. 这一现象表明, 多覆盖融合模型在长句子(长度大于 50)翻译方面仍然有很大的改进空间.

3.4 过度翻译和漏翻译分析

(1) 过度翻译分析

在实验中, 根据译文中出现重复翻译的源语言词汇数量评估过度翻译情况, 实验结果见表 4.

表 4 重复翻译评估

方法	重复翻译的源语言词语数量	∇ (%)
Baseline (Seq2Seq)	451	–
Coverage model	390	−13.5
HMC	349	−22.6

注: 表中∇表示相比于基线系统, 其他模型出现重复翻译的源语言词语数量减少的比率

虽然基线系统、覆盖模型和层次多覆盖模型的翻译结果中均存在不同程度的过度翻译问题, 但覆盖模型

和层次多覆盖模型的重复翻译次数均少于基线系统, 且后者减少的幅度更大. 这证明层次多覆盖模型能够在覆盖模型的基础上, 进一步缓解神经机器翻译中的过度翻译现象.

(2) 漏翻译分析

一元文法 BLEU 值可以用来表示源语言词汇正确出现在译文中的比率, 因此可以用来评价译文中的漏翻译现象[25]. 基线系统、覆盖模型和层次多覆盖模型的一元文法 BLEU 值结果见表 5.

表 5 不同模型的一元文法 BLEU 值 (%)

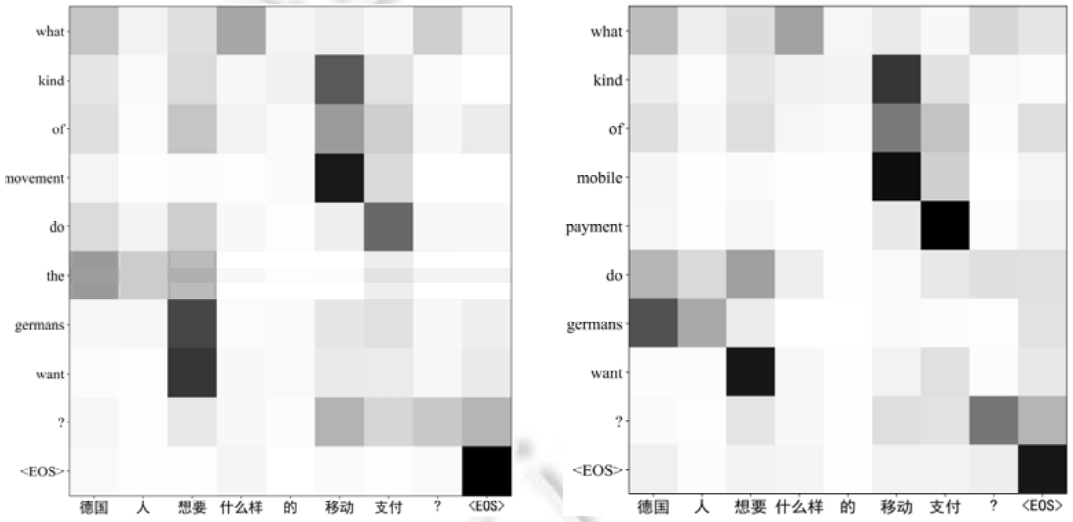
系统	newstest2017	cwmt2018	newstest2018	平均值	Δ
Baseline (Seq2Seq)	60.82	63.27	62.53	62.21	–
Coverage model	61.08	63.34	62.50	62.31	+0.10
HMC	61.25	63.55	62.51	62.44	+0.23

注: Δ表示其他模型相比于基线系统一元文法 BLEU 值的提升

实验结果表明, 层次多覆盖模型的一元文法 BLEU 值比基线系统和覆盖模型分别提升了 0.23%和 0.13%, 证明了层次多覆盖模型在改善漏翻译问题方面的作用.

在表 1 的示例中, 覆盖模型没有很好地改善源语言句子中的过度翻译和漏翻译问题. 而层次多覆盖模型将“体制”翻译为“system”, 消除了基线系统中的重复翻译问题, 且将“事关全局”翻译为“to the overall situation”. 虽然翻译结果与参考译文仍有一定差距, 但在所有译文中取得了最好的翻译结果, 句意也最接近于参考译文.

此外, 源语言中某些词语的漏翻译还可能造成译文句意的改变. 如图 6(a)所示, 在基线系统中, 由于“支付(payment)”出现漏翻译, 导致“移动(mobile)”被错误地翻译成“movement”, 从而导致最终的译文与源语言句意大相径庭. 而在图 6(b)中, 层次多覆盖模型通过消除漏翻译现象, 将短语“移动支付”正确地翻译为“mobile payment”, 保证了翻译结果对源语言句意的忠实性.



(a) Baseline (b) HMC

图 6 漏翻译示例

3.5 对齐质量分析

神经机器翻译模型的注意力机制建立了源语言和目标语言的对齐关系. 为了考查多覆盖模型的对齐质量, 在文献[26]提供的人工对齐的中英数据集上, 利用强制解码(force-decoding)方法使解码器输出参考译文, 将由编码器和解码器的最顶层状态计算得到的注意力分数的最大值作为对齐标准, 并利用对齐错误率(alignment error rate, AER)[27]对结果进行评价. 对于含有子词的源语言和目标语言词语, 将其包含的所有子词

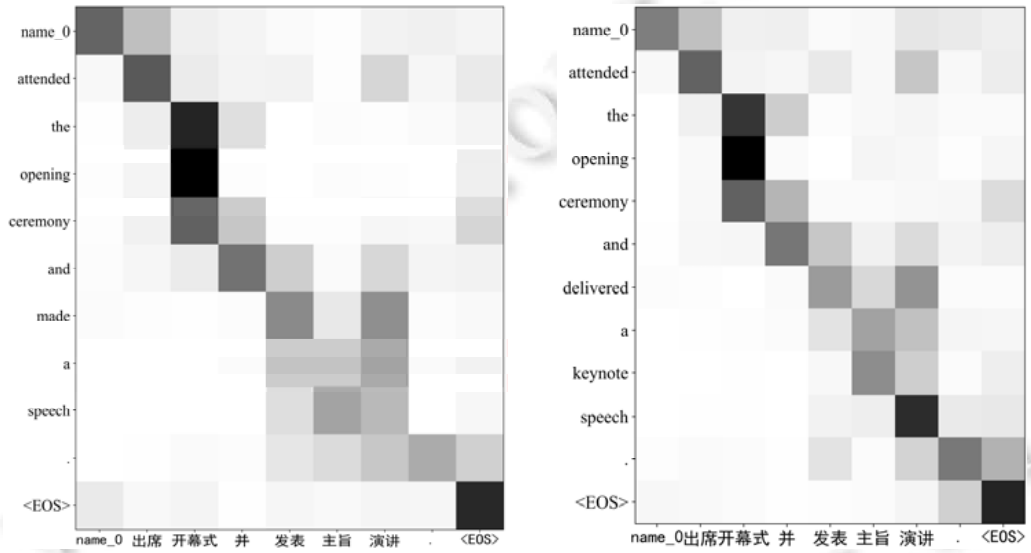
的注意力分数之和作为该词语的注意力分数. 结果见表 6.

表 6 对齐质量评估

系统	AER (%)
Baseline (Seq2Seq)	68.3
Coverage model	67.4
HMC	67.0

注: 分数越低表示对齐质量越好

如表 6 所示, 覆盖模型和层次多覆盖模型均能提升对齐质量, 并且后者的对齐效果更好. 如图 7(a)所示, 源语言中, “发表/主旨/演讲/.”在翻译过程中的注意力对齐存在错误, “主旨(keynote)”和“演讲(speech)”分别被错误地对齐到了“speech”和“make”, 造成“主旨(keynote)”被漏翻译. 而在图 7(b)中, 层次多覆盖模型不仅正确地建立了“主旨-keynote”和“演讲-speech”的对齐关系并消除了漏翻译现象, 而且局部注意力分布较为集中, 表明源语言和目标语言的对齐质量得到提升.



(a) Baseline (b) HMC

图 7 词对齐示例

4 结 论

在神经机器翻译模型中引入覆盖机制, 可以缓解过度翻译和漏翻译问题. 但依靠覆盖向量或覆盖分数等单一方式存储的覆盖信息并不完善. 本文讨论了不同覆盖模型的信息存储方式、使用方式和优缺点, 并基于翻译历史信息的一致性和模型之间的互补性提出多覆盖融合模型. 首先, 定义了词级覆盖分数概念; 之后, 利用覆盖分数和覆盖向量存储的信息同时指导注意力权重计算, 并根据覆盖向量和覆盖分数融合方式的不同提出层次多覆盖模型和平行多覆盖模型两种方法. 实验结果表明, 多覆盖融合方法能够进一步减少过度翻译和漏翻译现象, 并改善源语言和目标语言的对齐质量.

然而, 由于注意力机制的“软对齐”方式使得覆盖程度难以衡量, 多覆盖融合模型并不能完全消除过度翻译和漏翻译问题. 本文利用融合方法得到的覆盖信息与统计机器翻译中的“硬对齐”方式相比仍有差距. 在未来的研究中, 我们将进一步完善覆盖信息的融合方式, 并探索利用统计机器翻译的对齐信息缓解神经机器翻译中过度翻译和漏翻译问题的方法.

References:

[1] Kalchbrenner N, Blunsom P. Recurrent continuous translation models. In: Proc. of the 2013 Conf. on Empirical Methods in Natural Language Processing. Seattle: Association for Computational Linguistics, 2013. 1700−1709.

[2] Sutskever I, Vinyals O, Le QV. Sequence to sequence learning with neural networks. In: Advances in Neural Information Processing Systems. Montréal: Curran Associates, Inc., 2014. 3104−3112.

[3] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. In: Proc. of the 3rd Int'l Conf. on Learning Representations. 2015.

[4] Gehring J, Auli M, Grangier D, Yarats D, Dauphin YN. Convolutional sequence to sequence learning. In: Proc. of the 34th Int'l Conf. on Machine Learning. 2017. 1243−1252.

[5] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. In: Advances in Neural Information Processing Systems. Los Angeles: NIPS, 2017. 5998−6008.

[6] Koehn P, Och FJ, Marcu D. Statistical phrase-based translation. In: Proc. of the 2003 Conf. of the North American Chapter of the Association for Computational Linguistics on Human Language Technology. Edmonton: Association for Computational Linguistics, 2003. 48−54.

[7] Li YC, Xiong DY, Zhang M. A survey of neural machine translation. Chinese Journal of Computers, 2018, 41(12): 2734−2755 (in Chinese with English abstract).

[8] Tu ZP, Lu ZD, Liu Y, Liu XH, Li H. Modeling coverage for neural machine translation. In: Proc. of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin: Association for Computational Linguistics, 2016. 76−85. [doi: 10.18653/v1/p16-1008]

[9] Mi HT, Sankaran B, Wang ZG, Ittycheriah A. Coverage embedding models for neural machine translation. In: Proc. of the 2016 Conf. on Empirical Methods in Natural Language Processing. Austin: Association for Computational Linguistics, 2016. 955−960. [doi: 10.18653/v1/d16-1096]

[10] Feng S, Liu SJ, Yang N, Li M, Zhou M, Zhu KQ. Improving attention modeling with implicit distortion and fertility for machine translation. In: Proc. of the 26th Int'l Conf. on Computational Linguistics: Technical Papers (COLING 2016). Osaka: The COLING 2016 Organizing Committee, 2016. 3082−3092.

[11] Zheng ZX, Zhou H, Huang SJ, Mou LL,Dai XY, Chen JJ, Tu ZP. Modeling past and future for neural machine translation. Trans. of the Association for Computational Linguistics, 2018, 6: 145−157. [doi: 10.1162/tacl_a_00011]

[12] Wu YH, Schuster M, Chen ZF, Le QV, Norouzi M. Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144, 2016.

[13] Li YY, Xiao T, Li YQ, Wang Q, Xu CM, Lu XQ. A simple and effective approach to coverage-aware neural machine translation. In: Proc. of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne: Association for Computational Linguistics, 2018. 292−297. [doi: 10.18653/v1/p18-2047]

[14] Tu ZP, Liu Y, Lu ZD, Liu XH, Li H. Context gates for neural machine translation. Trans. of the Association for Computational Linguistics, 2017, 5: 87−99. [doi: 10.1162/tacl_a_00048]

[15] Wang MX, Xie J, Tan ZX, Su JS, Xiong DY, Bian C. Neural machine translation with decoding-history enhanced attention. In: Proc. of the 27th Int'l Conf. on Computational Linguistics. Santa Fe: Association for Computational Linguistics, 2018. 1464−1473.

[16] Zhao Y, Zhang JJ, Zong CQ, He ZJ, Wu H. Addressing the under-translation problem from the entropy perspective. In: Proc. of the 33th AAAI Conf. on Artificial Intelligence. Hawaii: AAAI, 2019. 451−458. [doi: 10.1609/aaai.v33i01.3301451]

[17] Xiao T, Zhu JB, Zhang H, Li Q. NiuTrans: An open source toolkit for phrase-based and syntax-based machine translation. In: Proc. of the 50th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2012. 19−24.

[18] Sennrich R, Haddow B, Birch A. Neural machine translation of rare words with sub-word units. In: Proc. of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin: Association for Computational Linguistics, 2016. 1715−1725. [doi: 10.18653/v1/p16-1162]

[19] Britz D, Goldie A, Luong MT, Le QV. Massive exploration of neural machine translation architectures. arXiv preprint arXiv:1703.03906, 2017.

[20] Kingma DP, Ba JL. Adam: A method for stochastic optimization. In: Proc. of the 3rd Int'l Conf. on Learning Representations. 2015.

[21] Papineni K, Roukos S, Ward T, Zhu WJ. BLEU: A method for automatic evaluation of machine translation. In: Proc. of the 40th Annual Meeting on Association for Computational Linguistics. Philadelphia: Association for Computational Linguistics, 2002. 311–318. [doi: 10.3115/1073083.1073135]

[22] Collins M, Koehn P, Kučerová I. Clause restructuring for statistical machine translation. In: Proc. of the 43rd Annual Meeting of the Association for Computational Linguistics. Ann Arbor: Association for Computational Linguistics, 2005. 531–540. [doi: 10.3115/1219840.1219906]

[23] Sennrich R, Haddow B, Birch A. Edinburgh neural machine translation systems for WMT 16. In: Proc. of the 1st Conf. on Machine Translation. Vol.2: Shared Task Papers. Berlin: Association for Computational Linguistics, 2016. 371–376.

[24] Zhang JC, Ding YZ, Shen SQ, Cheng Y, Sun MS, Luan HB, Liu Y. THUMT: An open source toolkit for neural machine translation. arXiv preprint arXiv: 1706.06415, 2017.

[25] Kuang SH, Li JH, Branco A, Luo WH, Xiong DY. Attention focusing for neural machine translation by bridging source and target embeddings. In: Proc. of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne: Association for Computational Linguistics, 2018. 1767–1776. [doi: 10.18653/v1/p18-1164]

[26] Liu Y, Sun MS. Contrastive unsupervised word alignment with non-local features. In: Proc. of the 29th AAAI Conf. on Artificial Intelligence. AAAI, 2015. 2295–2301.

[27] Och FJ, Ney H. A systematic comparison of various statistical alignment models. Computational Linguistics, 2003, 29(1): 19–51. [doi: 10.1162/089120103321337421]

附中文参考文献:

[7] 李亚超, 熊德意, 张民. 神经机器翻译综述. 计算机学报, 2018, 41(12): 2734–2755.

刘俊鹏(1992－), 男, 博士生, 主要研究领域为自然语言处理, 机器翻译.

宋鼎新(1985－), 男, 博士, 主要研究领域为自然语言处理与机器翻译.

黄锴宇(1992－), 男, 博士, 主要研究领域为自然语言处理, 中文词法分析.

黄德根(1965－), 男, 博士, 教授, 博士生导师, CCF 高级会员, 主要研究领域为自然语言处理, 机器翻译.

李玖一(1994－), 女, 博士生, 主要研究领域为自然语言处理, 文本摘要.