中国科学: 信息科学 2018 年 第 48 卷 第 5 期: 574–588
SCIENTIA SINICA Informationis
人工智能优青专刊 · 论文

基于约束的神经机器翻译

熊德意[1*], 李军辉[1], 王星[1], 张飚[2]

1. 苏州大学计算机科学与技术学院, 苏州 215006
2. 厦门大学软件学院, 厦门 361005
* 通信作者. E-mail: dyxiong@suda.edu.cn

收稿日期: 2018–03–12; 接受日期: 2018–04–16; 网络出版日期: 2018–05–11

国家自然科学基金优秀青年基金 (批准号: 61622209) 资助项目

摘要 神经机器翻译是近几年出现并快速发展的一种深度学习驱动的新型机器翻译模式, 目前已成为机器翻译学术和工业界广为接受的主流技术. 本文总结了我们在神经机器翻译方面的工作, 特别是在各种信息和知识约束条件下提出的一系列神经机器翻译模型和方法, 具体包括隐变量约束的变分神经机器翻译模型、单词与短语级统计机器翻译译文推荐与约束模型、源端句法结构约束模型. 除此之外, 本文也对神经机器翻译未来发展进行了初步思考和展望.

关键词 神经机器翻译, 变分神经机器翻译, 神经机器翻译与统计机器翻译融合, 句法约束的神经机器翻译

1 引言

神经机器翻译 (neural machine translation, NMT) 是在 2014 年底快速发展起来的一种新型机器翻译模式 [1,2], 是深度学习技术在机器翻译中的延伸和应用. 通常情况下, NMT 包括一个编码器和一个解码器, 编码器将源语言句子编码成实数向量表示, 解码器基于源语言表示逐词生成对应的目标语言句子. 编码器和解码器通过网络连接构成一个大的神经网络, 该网络在机器翻译平行语料上可以实现端到端的训练.

同传统的统计机器翻译 [3,4] (statistical machine translation, SMT) 相比, 神经机器翻译具有如下几个明显的不同点. 首先, 神经机器翻译是一个统一的端到端训练的模型, 不需要像统计机器翻译那样既要完成多个级联步骤 (如单词对齐、翻译规则抽取、翻译规则评分等), 又要单独训练多个子模型 (语言模型、翻译模型、调序模型等), 然后再用对数线性 (log-linear) [5] 方式将多个子模型结合成一个大模型, 因此更易于训练和实际部署. 其次, 神经机器翻译是基于连续空间的实数向量表示, 翻译过程可以看作是连续空间元素的映射、组合和计算, 而统计机器翻译的基本翻译单元是基于符号表示的 (单

引用格式: 熊德意, 李军辉, 王星, 等. 基于约束的神经机器翻译. 中国科学: 信息科学, 2018, 48: 574–588, doi: 10.1360/N112017-00222
Xiong D Y, Li J H, Wang X, et al. Neural machine translation with constraints (in Chinese). Sci Sin Inform, 2018, 48: 574–588, doi: 10.1360/N112017-00222

词、短语或者带有非终结符的句法规则), 再配以相应的翻译概率, 翻译过程是基于符号匹配的概率计算和推导. 再次, 与第 2 个不同点相关的是, 神经机器翻译模型以词嵌入 (word embedding) [6] 构成的实数向量矩阵和网络连接权重系数形式存储, 统计机器翻译模型则通常以翻译规则及其概率、n-gram 及其概率形式存储, 因此模型远大于神经机器翻译模型.

正是因为以上特点, 神经机器翻译获得了机器翻译研究者的广泛青睐, 并在很多语言对上迅速取得最好成绩 (state of the art) [7~9]. 神经机器翻译在工业界也取得了广泛应用, 目前 Google、百度等在线机器翻译平台大部分语言对已经从统计机器翻译模型切换到神经机器翻译模型.

虽然如此, 神经机器翻译模型本身并不是完美的. 一方面, 内在结构制约性因素导致了一些共识性的固有问题, 比如长句子翻译 [10]、集外词翻译 [11]、过翻与漏翻 [12] 等问题 (下文将对这些问题展开阐述), 这些问题需要提出针对性的解决方案. 另一方面, 神经机器翻译模型仍然可以在多个方向上得到增强和提升. 我们从 2016 年开始对基于约束的神经机器翻译展开了有益的探索, 研究如何融入更多约束信息使神经机器翻译模型能够捕获更多源语言、目标语言信息, 使解码器能在更多信息的指导下生成更精准译文. 研究工作取得了初步成效, 另外也发现这些工作对神经机器翻译本身固有问题也有一定的改善能力, 虽然它们不是特别针对这些问题设计的.

这些工作具体包括: (1) 基于隐变量约束的变分神经机器翻译模型; (2) 基于统计机器翻译推荐译文的约束模型; (3) 基于源端句法结构约束的神经机器翻译模型. 其中第 1 个工作旨在挖掘更多潜在语义信息约束神经机器翻译译文的生成, 第 2 个工作借助统计机器翻译模型来约束神经机器翻译, 最后一个工作则利用外部句法知识对神经机器翻译进行约束. 下面将逐一介绍这 3 方面的工作, 并在本文结尾处对神经机器翻译未来研究进行探讨和思考.

2 相关工作

本文围绕在神经机器翻译中融入约束信息展开 3 个方面的工作, 本节将针对这 3 个方面的相关工作进行简要概述.

2.1 变分神经模型

为了在大规模数据集上高效地学习生成式模型, Kingma 等 [13] 和 Rezende 等 [14] 提出变分神经网络, 借助非线性神经网络来近似拟合隐变量的先验和后验分布, 并采用随机梯度下降算法直接优化经再参数化的变分下界函数. 该算法一经提出, 引起大量科研人员的广泛关注. Kingma 等 [15] 基于上述算法提出深层的生成式模型进行半监督学习, 成功地利用了大规模未标注数据集来提升模型的泛化能力. Chung 等 [16] 将隐变量融入到循环神经网络之中, 提出循环变分神经网络. Miao 等 [17] 针对基于文本的条件和生成模型提出了一个通用的变分推理框架. Bowman 等 [18] 则提出可生成文本的变分自编码器, 将句子编码进隐变量空间, 再利用循环神经网络从中恢复生成出来. 和本文工作不同, Bowman 等主要关注单语语言模型, 而不是双语翻译. 虽然基于变法算法进行解码的思想在统计机器翻译中已有应用 [19], 但本文是首次将变分神经网络和神经机器翻译进行联合建模.

2.2 统计机器翻译与神经机器翻译融合

一些研究以充分利用统计机器翻译和神经机器翻译两者的互补性为出发点, 在词汇、短语, 以及句子等各个层面开展两种模型的融合工作. 在词汇层面, He 等 [20] 在对数线性模型的框架下对神经机

器翻译和传统统计机器翻译进行结合, 将神经机器翻译模型和传统统计机器翻译的翻译模型、词汇数惩罚模型和语言模型作为特征置于对数线性模型中. Stahlberg 等 [21] 提出一种句法指导的神经机器翻译模型, 利用基于层次短语的机器翻译模型来指导神经机器翻译解码. 在模型解码时, 层次短语模型的翻译假设限制神经机器翻译解码器的搜索空间并调整解码器目标词语预测概率. Arthur 等 [22] 通过将估计出的词汇化翻译概率与神经机器翻译的目标语言词语预测概率进行结合, 以提高神经机器翻译的译文忠实度. 在短语层面, Dahlmann 等 [23] 针对神经机器翻译的搜索提出一种利用基于短语的统计机器翻译模型的混合搜素策略. 该工作以对数线性模型为框架, 集成神经机器翻译的特征和统计机器翻译的特征, 以进行单词或者短语的混合搜索. 在句子层面, Niehues 等 [24] 提出先使用统计机器翻译系统进行 “预翻译 (pre-translation)” 机制, 将统计机器翻译的译文和源语言句子一起输入神经机器翻译系统, 以产生受统计机器翻译影响的最终译文. Zhou 等 [25] 进一步提出了一种以神经网络为基础的系统融合框架. 该框架采用一种 “多编码器 – 单解码器” 架构, 使用多个编码器分别接收基于短语的翻译系统、基于句法的翻译系统、神经机器翻译系统等多种不同种类系统的输出, 综合考虑各种系统输出的差异后解码产生最终的翻译译文.

2.3 句法神经机器翻译

NMT 端到端的翻译模型提供了一个非常灵活的融入句法信息的机制. 比如, Eriguchi 等 [26] 提出了树到序列的模型, 该模型不仅学习到每个单词的表示向量, 并且通过树 LSTM, 为每个源端句法树的内部结点也学习到表示向量. 在解码的时候, 注意力机制不仅对齐源端单词, 而且还对齐源端句法树内部结点. Sennrich 和 Haddow [27] 同样使用了源端句法信息, 该模型为每个源端单词, 获取它的句法特征, 包括词性、依存关系等. 每个特征都对应一个特征向量 (类似于每个词对应一个词向量), 最后将每个单词的词向量和该单词的各个特征向量进行拼接, 得到一个新的更长的向量, 作为编码器的输入. Shi 等 [28] 设计了一系列的实验用于验证序列到序列模型是否能够隐式地学习得到源端句法信息, 其实验表明序列到序列模型能够较好地学习到源端浅层句法信息 (如词性等), 但很难捕获深层次的句法信息. 此外, Wu 等 [29] 提出了序列到依存树模型, 该模型在解码端不仅得到目标端的单词序列, 而且还得到这些单词之间的依存关系.

3 隐变量约束的变分神经机器翻译模型

目前主流的神经机器翻译模型只能利用源句子中的信息进行翻译. 通过训练优化, NMT 可以从大规模的平行语料中捕捉到目标语言和源语言之间的语义对应关系, 因此能够将源句子平滑地转换为目标译文. 然而, 语言的表述方式并非是单一的, 同一种语义可以基于不同的词汇、结构、句法等表达出来, 这种多样性导致 NMT 学习到的语义对应模式中存在着一定的不确定性. 源句子虽然可以提供充分的语义信息, 但并未涉及目标译文中应有的词汇、语法、结构等相关知识, 单单依靠源句子进行翻译无疑增加了模型的学习难度.

为此, 我们研究如何将目标译文本身融入到 NMT 之中从而直接约束并指导 NMT 的翻译. 在训练过程中, 源句子和目标句子是同时给出的, 可以直接将后者建模到 NMT 中去; 但是在解码阶段, 只能接触到源句子, 目标译文的缺失成为上述想法的关键障碍. 为了克服这一障碍, 我们提出了基于隐变量的变分神经机器翻译模型 (variational NMT, VNMT) [30], 其结构如图 1 所示. 通过假定隐变量 $\boldsymbol{z}$ 的后验分布同时依赖于源句子和目标句子 (即, $q_\phi(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y})$, 见图 1 虚线部分), 以及其先验分布仅依赖

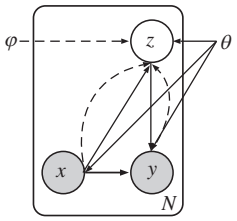


图 1 隐变量约束的变分神经机器翻译模型的图模型表示. 图中 $\boldsymbol{z}$ 表示隐变量, $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 分别表示源句子和目标句子

Figure 1 Graphical model for latent variable constrained variational neural machine translation. We use symbols $\boldsymbol{z}$, $\boldsymbol{x}$ and $\boldsymbol{y}$ to denote the latent variable, source sentence and target sentence, respectively

于源句子 (即, $p_\theta(\boldsymbol{z}|\boldsymbol{x})$), 可以将原先的似然问题 $\log p(\boldsymbol{y}|\boldsymbol{x})$ 转换为如下的变分下界问题 [13]:

$$\log p(\boldsymbol{y}|\boldsymbol{x}) = \log \int_{\boldsymbol{z}} p(\boldsymbol{y},\boldsymbol{z}|\boldsymbol{x})\mathrm{d}\boldsymbol{z} = \log \int_{\boldsymbol{z}} p(\boldsymbol{y}|\boldsymbol{z},\boldsymbol{x})p(\boldsymbol{z}|\boldsymbol{x})\mathrm{d}\boldsymbol{z} \tag{1}$$

$$\geqslant \mathrm{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y})}[\log p_\theta(\boldsymbol{y}|\boldsymbol{z},\boldsymbol{x})] - \mathrm{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y})||p_\theta(\boldsymbol{z}|\boldsymbol{x})), \tag{2}$$

其中, ϕ 和 θ 分别表示与后验和先验分布相关的模型参数, E 表示期望函数, KL$(\cdot)$ 表示两个分布之间的 Kullback-Leibler 距离. 从式 (2) 中可以看出, VNMT 其实是在原有 NMT 数据似然的基础上 (即公式第 1 项) 引入了关于后验和先验分布的正则约束 (即公式第 2 项) . 在最大化上述下界函数的过程中, 一方面, 似然目标 $\log p_\theta(\boldsymbol{y}|\boldsymbol{z},\boldsymbol{x})$ 确保了后验分布 $q_\phi(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y})$ 中融入了目标句子 $\boldsymbol{y}$ 的信息; 另一方面, KL$(\cdot)$ 距离的约束使得先验分布 $p_\theta(\boldsymbol{z}|\boldsymbol{x})$ 与后验分布保持一致. 如此一来, 目标句子的信息便可以经由后验分布流入到先验分布中, 而后者因为不依赖目标句子从而使隐变量 $\boldsymbol{z}$ 在训练和解码阶段都是有意义的.

模型设计与实现. 式 (2) 虽然在理论上说明了借助隐变量引入目标句子信息来约束 NMT 的可能性, 但同时也带来了新的挑战.

(1) $\mathrm{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y})}[\log p_\theta(\boldsymbol{y}|\boldsymbol{z},\boldsymbol{x})]$ 求解似然函数关于隐变量后验分布 $q_\phi(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y})$ 的期望. 但是, 通常利用复杂的非线性函数来逼近真实后验分布. 因此上述期望往往没有确切的解析解, 不利于 NMT 所依赖的端到端的学习和优化;

(2) $\mathrm{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y})||p_\theta(\boldsymbol{z}|\boldsymbol{x}))$ 迫使先验和后验分布尽量保持一致. 然而, 一方面如何寻找合适的非线性函数来模拟先验和后验分布在学术界尚无定论; 另一方面, 复杂的分布在 KL 正则下也面临着没有解析解的风险.

针对上述难题, 我们借鉴了变分自编码器 [13] 的相关方法, 其核心思想如下:

- 假定先验和后验分布均服从正态分布, 且其分布参数 (均值和方差) 是可以学习的;
- 利用神经网络构建复杂的非线性模型来分别学习两个分布的均值和方差. 由于正态分布之间的 KL 距离存在解析解, 上述第 2 个问题可以解决;
- 应用 Monte Carlo 模拟算法来近似函数的期望; 为了从分布中获得隐变量的表示, 使用再参数化方法, 这样上述第 1 个问题得到解决.

依据上述方法, 在 VNMT 中设计了变分神经编码器、变分神经推理器和变分神经解码器 3 部分, 其结构如图 2 所示.

变分神经编码器 (图 2(a)). 该部分采用双向循环神经网络来编码句子的语义信息, 并将正向和反向网络中第 i 个词的表示拼接起来作为该词的语义表示, 即 $[\overrightarrow{\boldsymbol{h}_i};\overleftarrow{\boldsymbol{h}_i}]$.

变分神经推理器 (图 2(b)). 该部分基于多层非线性感知机来建模隐变量的先验和后验分布, 并

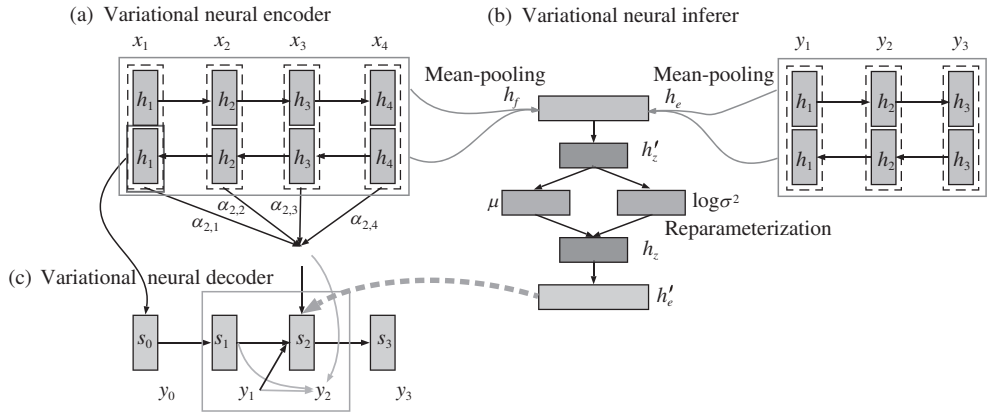


图 2　隐变量约束的变分神经机器翻译模型结构示意

Figure 2　Overall architecture for latent variable constrained variational neural machine translation

从中提取隐变量的向量表示. 本文主要介绍后验分布的推理过程, 先验分布可以利用相同的方法得到, 只是不需要目标句子的信息.

假定后验分布服从如下的正态分布: $q_\phi(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y}) = \mathcal{N}(\boldsymbol{z};\mu(\boldsymbol{x},\boldsymbol{y}),\sigma(\boldsymbol{x},\boldsymbol{y})^2\boldsymbol{I})$, 其均值 μ 和方差 σ 都是建立在源句子和目标句子表示之上的非线性神经网络的输出. 借助变分神经编码器, 利用均值池化的方法从句子的词表示中提取整个句子的表示, 从而获得了源句子 ($\boldsymbol{h}_f = \frac{1}{T_f}\sum_i^{T_f}\boldsymbol{h}_i$) 和目标句子 ($\boldsymbol{h}_e = \frac{1}{T_e}\sum_i^{T_e}\boldsymbol{h}'_i$) 的语义向量. 在此之上, 推理器采用如下方式得到 μ 和 σ:

$$\boldsymbol{h}'_z = g(W_z^{(1)}[\boldsymbol{h}_f;\boldsymbol{h}_e] + b_z^{(1)}),\quad \mu = W_\mu\boldsymbol{h}'_z + b_\mu,\quad \log\sigma^2 = W_\sigma\boldsymbol{h}'_z + b_\sigma, \tag{3}$$

$W_\star$ 和 $b_\star$ 表示待调参数, $g(\cdot)$ 是一个非线性映射函数, 在实验中使用了 tanh 双曲正切函数.

给定上述正态分布, 推理器利用再参数化方法获取隐变量 $\boldsymbol{z}$ 的样本表示:

$$\boldsymbol{h}_z = \mu + \sigma\odot\epsilon,\ \text{其中}\ \epsilon\sim\mathcal{N}(0,\boldsymbol{I}). \tag{4}$$

变分神经解码器 (图 2(c)). 该部分将变分神经推理器得到的隐变量表示融入到解码器中以约束并指导 NMT 的解码过程. 其基本结构与 NMT 中的解码器一致, 区别于在原解码器的基础上引入了新的输入信息 $\boldsymbol{h}_z$, 即

$$\boldsymbol{s}_t = f(\boldsymbol{s}_{t-1},[\boldsymbol{E}_{y_{t-1}};\boldsymbol{c}_t;\boldsymbol{h}_z]), \tag{5}$$

$f(\cdot)$ 表示解码函数, 在实验中使用 GRU 模型 [31]. $\boldsymbol{s}_t$ 表示 t 时刻解码器的状态向量, $\boldsymbol{E}_{y_{t-1}}$ 表示上一个词的词向量, $\boldsymbol{c}_t$ 是注意机制在源端捕获的语义相关信息.

我们在汉 – 英和英 – 德翻译上检验了 VNMT, 实验结果表明, 通过隐变量融入目标端句子信息约束 NMT 翻译可以显著增强模型的翻译性能. 由于拥有更多信息, 发现 VNMT 在长句子的翻译上也比基线系统好. 具体实验结果可参见我们在 EMNLP 2016 上发表的论文 [30]. 另外, 图 3 显示了 VNMT 训练的收敛性, 图中可看出, VNMT 的收敛速度和基准系统 (RNNSearch) 几乎一致.

4　统计机器翻译译文推荐与约束模型

神经机器翻译同统计机器翻译相比, 既有优点, 也有结构性的缺陷. 其中一些缺陷在引言中已列举, 这里根据已有的研究发现, 对以下 3 个问题作进一步阐述.

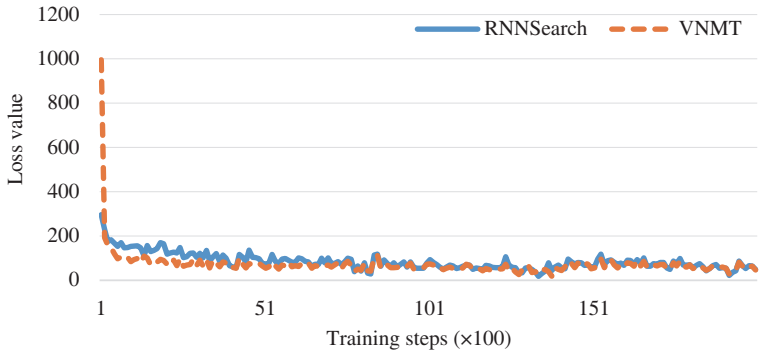


图 3 (网络版彩图) VNMT 收敛性

Figure 3 (Color online) Convergence of VNMT

(1) 词汇表规模受限问题 [32] (即引言中提到的集外词翻译问题). 为了控制模型的时空开销 (空间上存储词向量, 时间上, 解码每输出一个单词要计算词汇表所有单词的概率分布), 神经机器翻译通常在源语言端和目标语言端采用规模适当的词汇表 (词汇表单词数量一般控制在 3 万至 8 万) . 对于词汇表没有覆盖的单词, 模型会将其替换为源语言端和目标语言端相应的 UNK 符号. UNK 符号一方面影响神经机器翻译模型完整捕获源语言句子语义信息, 另一方面也影响用户理解神经机器翻译模型所产生的目标语言句子.

(2) 源语言翻译覆盖问题 [12] (对应引言中的过翻、漏翻问题). 神经机器翻译在解码过程中通过注意机制的自动调整, 在不同的解码时刻选择关注不同的源语言句子片段来产生对应的目标语言单词. 由于缺少约束, 注意机制无法保证源语言句子中的词语被 “恰到好处” 的关注, 因而导致了 “过翻译”、“欠翻译” 现象的产生. 其中 “过翻译” 指不该多次翻译的源语言词语被多次翻译, “欠翻译” 是指应该被翻译的源语言短语片段没有被翻译.

(3) 翻译不忠实问题 [22]. 连续表示, 一方面给神经机器翻译带来了更好的泛化能力, 另一方面也使得神经机器翻译容易产生不忠实的翻译. 此处不忠实的翻译是指模型生成的目标语言词语虽然能够保证目标语言语句的流利度, 却无法准确地反映出源语言句子的语义信息.

值得注意的是, 传统统计机器翻译方法没有受到以上 3 个问题的严重困扰. 这是因为 (1) 以离散符号来表示源语言端和目标语言端的单词, 词汇表规模因此不受限制; (2) 在翻译过程中显示记录源端语句中单词的翻译覆盖信息, 以确保源端词语翻译且只被翻译一次; (3) 以源语言端和目标语言端符号的转换作为翻译规则, 确保源语言端不同符号的翻译规则不会被混淆. 但是, 由于翻译建模机制的截然不同, 很难简单直观地将传统统计机器翻译和神经机器翻译进行融合.

如上所言, 统计机器翻译在上述 3 个问题上对神经机器翻译形成有益的补充, 基于此, 百度 [20]、eBay [23] 等国内外机器翻译研究团队分别在单词级和短语级层面利用传统统计机器翻译知识对神经机器翻译采取相应的约束, 以进一步提升神经机器翻译的翻译性能. 然而, 上述做法都是以传统统计机器翻译的核心 —— 线性对数模型 (log-liner model) —— 为框架, 将神经机器翻译作为框架的一个特征来实现两种翻译模式的互补. 不同于上述以对数线性模型为主干的 “浅层” 融合方法, 我们提出了利用传统统计机器翻译推荐的单词和短语知识, 以神经机器翻译为主干的 “深层” 融合方法, 实现神经机器翻译模型的统计机器翻译约束.

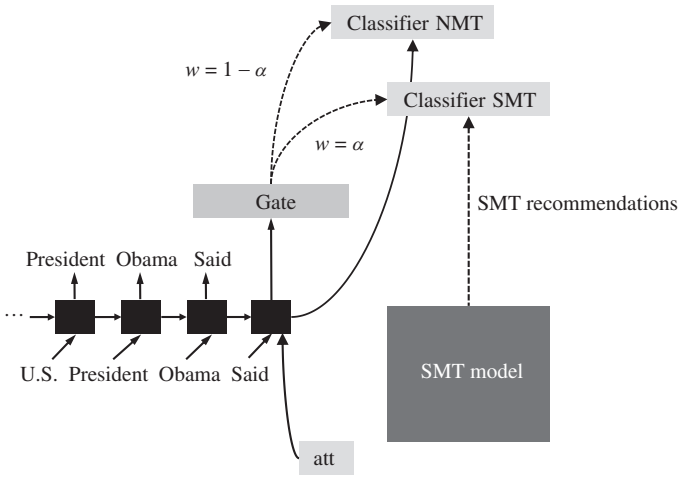


图 4　单词级约束模型

Figure 4　The model that integrates word-level SMT constraints into NMT

4.1　单词级约束模型

神经机器翻译中通常采用的编码器 – 解码器结构是以逐时刻单词预测的模式进行译文生成的. 一个直观的想法是把约束信息放置于单词预测阶段, 让神经机器翻译在单词预测的时候考虑单词级别约束信息. 这种通过单词约束信息以调整神经机器翻译单词预测概率的模型, 称为单词级约束模型.

如图 4 所示, 在单词级约束模型中, 传统统计机器翻译在每一个解码时刻, 根据神经机器翻译的解码信息 (当前时刻关注于源端哪部分信息和前面时刻生成过哪些目标词语信息), 推荐出一批候选的单词. 这些被推荐出的单词作为约束信息参与神经机器翻译的单词预测工作, 以限制神经机器翻译的译文生成. 具体而言, 神经机器翻译模型接收统计机器翻译的单词候选并对这些单词进行概率估计, 单词 y_t 的概率估计按如下计算:

$$p_{\text{smt}}(y_t|\boldsymbol{y}_{<t},\boldsymbol{x}) = \text{softmax}(\text{score}_{\text{smt}}(y_t|\boldsymbol{y}_{<t},\boldsymbol{x})), \tag{6}$$

其中 $\boldsymbol{y}_{<t}$ 表示神经机器翻译在 t 时刻之前生成的译文, $\boldsymbol{x}$ 表示给定的源语言句子, $\text{score}_{\text{smt}}(y_t|\boldsymbol{y}_{<t},\boldsymbol{x})$ 是神经机器翻译对推荐单词 y_t 的打分, 具体计算如下:

$$\text{score}_{\text{smt}}(y_t|\boldsymbol{y}_{<t},\boldsymbol{x}) = g_{\text{smt}}(f_{\text{smt}}(s_t, y_{t-1}, y_t, c_t)), \tag{7}$$

其中 s_t 是当前 t 时刻的神经机器翻译解码器的隐式状态, y_{t-1} 是神经机器翻译前一时刻所生成的单词, y_t 是当前 t 时刻推荐出的单词, c_t 是当前 t 时刻神经机器翻译解码器的对源语言句子总结表示的上下文向量.

基于以上单词估计概率, 预测下一个单词概率分布,

$$p(y_t|\boldsymbol{y}_{<t},\boldsymbol{x}) = (1-\alpha_t)p_{\text{nmt}}(y_t|\boldsymbol{y}_{<t},\boldsymbol{x}) + \alpha_t p_{\text{smt}}(y_t|\boldsymbol{y}_{<t},\boldsymbol{x}), \tag{8}$$

其中 p_{nmt} 是神经机器翻译自己的单词预测概率, p_{smt} 是上一步骤中计算出的传统统计机器翻译单词级约束概率, α_t 是约束的折算因子:

$$\alpha_t = g_{\text{gate}}(f_{\text{gate}}(s_t, y_{t-1}, c_t)). \tag{9}$$

通过这种方式, 可以利用统计机器翻译推荐的译文单词对神经机器翻译的解码进行约束.

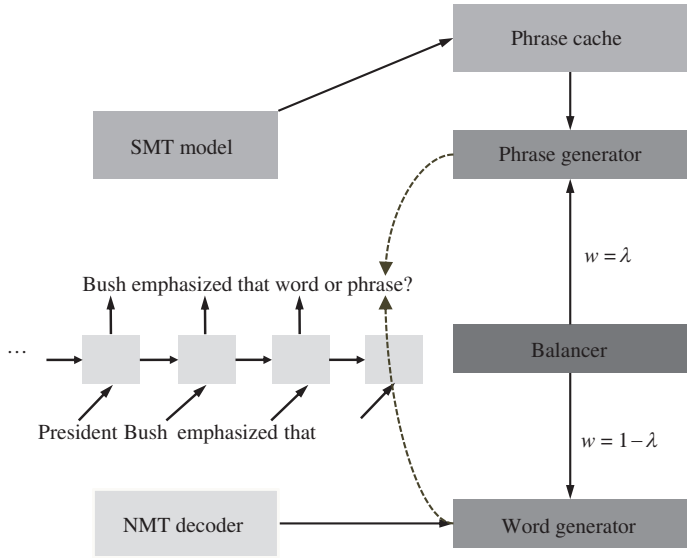


图 5 短语级约束模型

Figure 5 The model that integrates phrase-level SMT constraints into NMT

4.2 短语级约束模型

由于受到以上 3 个问题的困扰, 神经机器翻译逐时刻单词预测的模式在翻译短语时候往往会犯错误: 错翻或者漏翻. 进一步, 我们尝试利用统计机器翻译模型输出的短语知识来对神经机器翻译的解码进行约束. 在这种模式下, 统计机器翻译会对句法意义上完整的短语进行翻译, 然后以短语推荐的方式约束神经机器翻译解码.

如图 5 所示, 在短语级约束模型中, 统计机器翻译会预先对源语言句子的句法完整短语进行翻译. 在神经机器翻译的每一个解码时刻, 统计机器翻译根据神经机器翻译的解码信息, 从预翻译的短语中挑选并推荐出一批候选的目标短语. 这些被推荐出的短语作为约束信息参与神经机器翻译的译文生成. 不同于 4.1 小节的单词级约束模型, 在短语级约束模型中面临着翻译粒度不匹配的问题: 统计机器翻译推荐出的短语粒度和神经机器翻译生成中的词语粒度的不同, 导致很难像单词级约束那样直接使用约束信息.

为了克服这个难题, 我们引入一个平衡因子以使神经机器翻译在单词级别和短语级别两种模式中切换. 具体而言, 在每个解码时刻, 对统计机器翻译推荐的短语 p_l, 按如下方式计算其概率:

$$p_{\text{smt}}(p_l) = \text{softmax}(\text{score}_{\text{smt}}(p_l)), \tag{10}$$

其中 $\text{score}_{\text{smt}}(p_l)$ 是神经机器翻译对推荐短语 p_l 的打分, 按如下方式计算:

$$\text{score}_{\text{smt}}(p_l) = g_s\big(e(p_l), s_i, y_{i-1}, c_i\big), \tag{11}$$

其中 $e(p_l)$ 是被推荐短语 p_l 的向量表示, 实验中使用逆向的循环神经网络对短语进行建模表示.

根据当前时刻的解码信息计算平衡因子 λ_i, 以决定以多大概率接收短语级约束信息. 计算方式如下:

$$\lambda_i = \sigma(f_b(s_i, y_{i-1}, c_i)), \tag{12}$$

表 1 基准系统和所提出系统在 NIST08 上实词的数量统计

Table 1 Numbers of content words in NIST08 test set generated by the baseline system and the proposed system

	ALL	REMOVE
Reference	20481	19489
RNNSearch	13230	11007
+SMTrec/gate	12665	11172

最后结合统计机器翻译推荐的短语的概率估计和平衡因子 λ_i 来决定下一时刻译文的生成.

我们在汉 – 英翻译任务上检验了以上两种约束模型, 与基于注意机制的神经翻译系统相比, 单词级和短语级约束模型均能显著提升译文质量, 其中短语级约束模型略微优于单词级约束模型. 具体实验结果可参见发表在 AAAI 2017 [33] 和 EMNLP 2017 [34] 上的两篇工作. 这里, 补充一个有关过翻的实验结果. 表 1 中, ALL 指系统译文中实词数量, REMOVE 指去掉系统译文中重复的实词后统计的实词数量. 对比参考译文中重复实词数量 (20481 – 19489 = 992), 发现 RNNSearch 产生的重复实词数量 (13230 – 11007 = 2223) 和所提出的系统产生的重复实词数量 (12665 – 11172 = 1493) 都非常多, 但是所提系统的重复实词数量相对基准系统还是有所缓解, 表明我们的方法可以帮助 NMT 减少过翻.

5 源端句法结构约束模型

经典的神经机器翻译同统计机器翻译尤其是基于句法的统计机器翻译相比, 除了引言中提到的 3 个不同点之外, 还存在一个不同点, 那就是神经机器翻译很少显示利用语言学知识, 比如句法知识. 虽然 NMT 在预先不提供任何句法信息的前提下, 通过内部强有力的网络机制, 仍然可以捕获隐藏在源端句子中的句法信息, 如词性等, 但是, 深层的句法信息仍很难被捕获 [28]. 图 6 给出了神经机器翻译的两个例子, 输入的源句分别是“东京 证交所 批准 新生 银行 申请 上市 案”和“他们 来自 六 个 家庭, 其中 两 个 女孩 没有 父母”. 在句子 (a) 中, 名词短语“新生 银行” 被错误地翻译为不连续的两个词“new ... bank”, 其主要原因在于翻译模型预先并不知道“新生 银行”是一个名词短语, 而一个名词短语的译文通常是连续的; 在句子 (b) 中, 名词短语“两 个 女孩”虽然翻译正确, 译文为“two girls”, 但其被翻译了两次, 这属于过翻现象. 过翻现象在多数情况下也可以看作是某个短语被错误地翻译为不连续的两个短语.

一般来说, 可靠的句法信息是可以帮助机器更精准翻译的. 在统计机器翻译中, 已有大量的研究利用句法信息来辅助翻译, 包括直接基于形式句法或语言学句法的模型 [35~38]、利用句法信息作为额外知识的翻译模型 [39~41] 等. 本节, 我们将利用源端句法来约束神经机器翻译 [42].

5.1 源端句法的表示

在获取源端每个单词的句法信息时, 比较理想的情况的是, 模型能够为每个单词捕获它在整个句法树中的位置信息, 即结构化信息. 考虑到结构化信息在循环神经网络中通常难以表示, 将结构化信息序列化, 并将每个词在序列中的位置信息作为其结构化信息的近似替代. 如图 7 所示, 图 7(a) 给出了句子的词序列; 图 7(b) 给出了这个句子对应的句法树结构; 图 7(c) 给出了句子的序列化结果 (句法标签序列). 需要注意的是, 用句法树的序列化信息代替其结构化信息的做法也广泛应用于句法分析 [43, 44], 说明句法树的序列化信息可以作为其结构化信息的一种近似替代. 不难发现, 如图 7(a) 和 (c) 所示, 句法标签序列的长度要大于词序列长度, 对每个单词 w_i, 为了得到它的结构化表示向量, 可以简

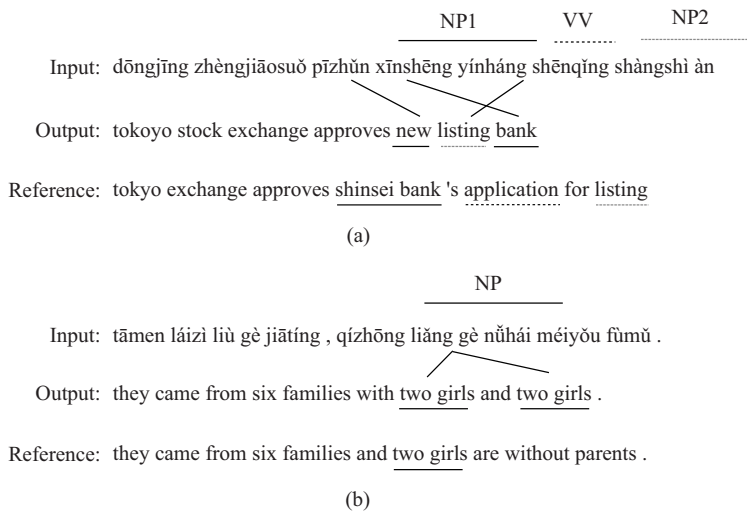


图 6 神经机器翻译结果示例

Figure 6 Examples of NMT translation. An example of (a) discontinuous translation, (b) over translation

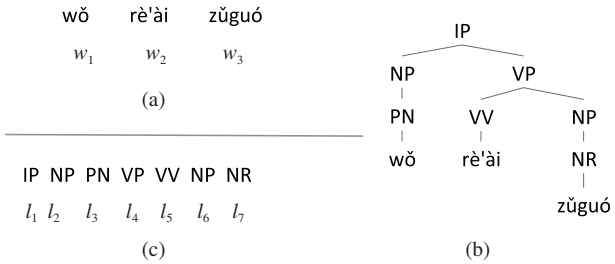


图 7 源端句子 “我 热爱 祖国” 的词序列

Figure 7 An example of a source sentence. (a) Word sequence; (b) its syntactic parse tree; (c) its syntactic label sequence

单地找到 w_i 的词性标记在句法标签序列中的位置, 并把该词性标记的表示向量看作是 w_i 的结构化表示向量. 例如, 在图 7 中, 词序列中的 w_1 与句法标签序列中的 l_3 相对应, 也就是说, 可以用学习到的 l_3 表示向量作为 w_1 的结构信息表示向量; 类似地, w_2 与 l_5, w_3 与 l_7 有对应的关系.

5.2 源端句法结构约束编码模型

给定如图 7 所示的词序列和句法标签序列, 以及两个序列之间的对应关系, 就能够为词序列中的每个词得到它的结构化信息表示向量. 我们提出了 3 种不同的策略来使用词的结构化信息. 根据词序列与句法标签序列编码方式的不同, 依次将这 3 种策略称为平行 RNN、层次 RNN 和混合 RNN.

图 8 给出了平行 RNN 编码器示意图. 该编码器定义了两个 RNN, 分别是针对词序列的词 RNN 和针对句法标签序列的句法标签 RNN, 两个 RNN 之间是平行的关系. 一旦获得词序列中的每个单词的表示向量、句法标签序列中的每个结构标签的表示向量后, 拼接每个词及其结构标签的向量, 得到新的向量, 新的表示向量将被用于后续的解码过程. 如图 8 所示, 词 “热爱” 的最终表示向量为 $[\overrightarrow{\mathrm{hw}_2}; \overleftarrow{\mathrm{hw}_2}; \overrightarrow{\mathrm{hl}_5}; \overleftarrow{\mathrm{hl}_5}]$, 其中前两项 $[\overrightarrow{\mathrm{hw}_2}; \overleftarrow{\mathrm{hw}_2}]$ 为词 “热爱” 的向量表示, 后两项 $[\overrightarrow{\mathrm{hl}_5}; \overleftarrow{\mathrm{hl}_5}]$ 为词性VV的向量表示.

图 9 给出了层次 RNN 编码器示意图. 该编码器同样定义了两个 RNN, 但两者之间具有层次关系.

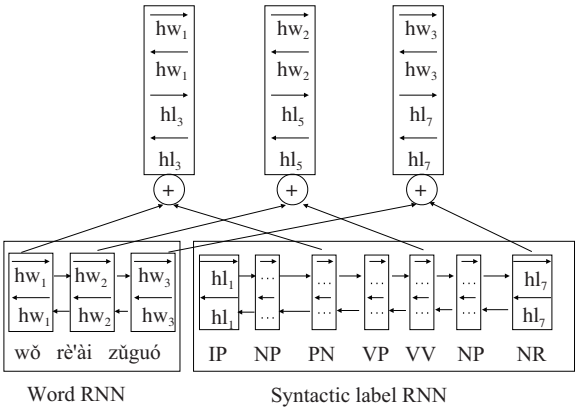


图 8 平行 RNN 编码器

Figure 8 Parallel RNN encoder

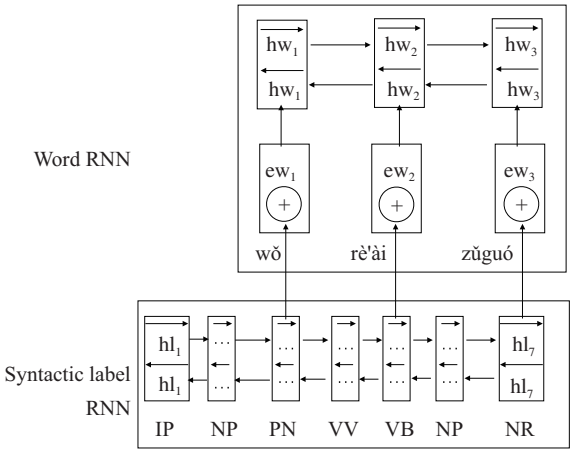


图 9 层次 RNN 编码器

Figure 9 Hierarchical RNN encoder

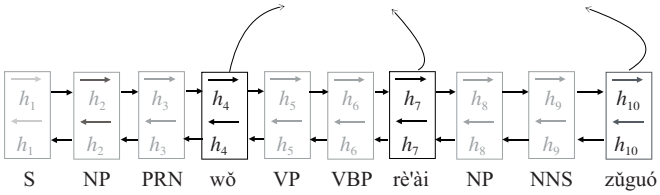


图 10 混合 RNN 编码器

Figure 10 Mixed RNN encoder

特别地, 考虑到句法标签序列的长度要大于词序列, 图 9 将句法标签 RNN 置于底层, 待每个句法标签学习得到它的表示向量后, 将其与词序列中对应单词的词向量拼接, 然后再进行词向量的 RNN 学习, 得到单词的向量表示. 如图 9 所示, 词性 VV 学习到的表示向量 $[\overrightarrow{\mathrm{hl}_5}; \overleftarrow{\mathrm{hl}_5}]$ 与词 "热爱" 的词向量 ew_2 拼接后, 作为词 "热爱" 在词 RNN 中的输入.

图 10 给出了混合 RNN 编码器示意图. 与前两者不同的是, 该编码器只有一个 RNN, 并且该 RNN 的输入是由词和句法标签混合组成的一个序列. 不难发现, 该序列实际上是一个完整句法树按前序遍历得到的结果, 在这个序列中, 每个单词都紧跟其句法标签. 如图 10 所示, 虽然混合 RNN 编码器会为每一个词、句法标签学习到它们的向量表示, 但仅仅词的向量表示 $([\overrightarrow{h_4}, \overleftarrow{h_4}], [\overrightarrow{h_7}, \overleftarrow{h_7}], [\overrightarrow{h_{10}}, \overleftarrow{h_{10}}])$ 用于解码.

我们在 NIST 汉 – 英翻译测试集上对以上 3 种源端句法约束方法进行了性能比较, 基线系统为基于注意机制的神经机器翻译模型 RNNSearch. 实验结果表明, 融入句法信息的 3 种编码器都能够提高机器翻译的性能, 难能可贵的是, 混合 RNN 较其他两个方法更简单, 但却更有效, 其原因在于该方法的词序列与句法标签序列的耦合较其他两者更为紧密. 另外更深入的译文分析表明混合 RNN 模型还能提高名词短语连续翻译的比例, 并缓解译文中的过翻 (over-translation) 问题. 表 2 给出了几类常见的源端短语的连续翻译、不连续翻译, 以及未翻译的统计结果. 更多实验结果请参见我们在 ACL 2017 上发表的工作 [42].

表 2 测试集上源端短语连续翻译、不连续翻译, 以及未翻译的比例 (%). 其中 PP 为介词短语, NP 为名词短语, CP 为从句修饰短语, QP 为量词短语

Table 2 Percentages (%) of syntactic phrases in our test sets being translated continuously, discontinuously, or not being translated. Here PP is for prepositional phrase, NP for noun phrase, CP for clause headed by a complementizer, and QP for quainter phrase

System	XP	Continuously	Discontinuously	Untranslated
RNNSearch	PP	57.3	33.6	9.1
	NP	59.8	25.5	14.7
	CP	47.3	44.6	8.1
	QP	54.0	22.2	23.8
	ALL	58.1	27.1	14.8
Mixed RNN	PP	63.3	27.5	9.2
	NP	63.1	23.1	13.8
	CP	54.5	36.6	8.9
	QP	56.2	19.7	24.1
	ALL	60.4	25.0	14.6

6 总结与未来工作

本文系统介绍了近年来我们在基于约束的神经机器翻译方面的工作. 约束信息包括通过隐变量传递的来自目标语言的约束信息、统计机器翻译模型提供的约束信息以及源语言句法结构的约束信息, 这些约束信息不仅较好地提升了神经机器翻译模型的性能, 也有效缓解了神经机器翻译的一些固有问题, 如长句子翻译、重复翻译等.

目前以循环神经网络 RNN 为基本构建单元的编码器 – 解码器机器翻译框架 (主流框架) 性能上仍然存在很大的上升空间. 推动该框架性能继续上升的因素可能来自两个方面. 一方面是使现有模型更贴近真实数据、估计更精准, 这可以通过优化神经网络训练方式、改进解码器搜索最优译文算法来实现, 类似于统计机器翻译中在模型不变的前提下尽可能减少搜索错误. 另一方面是增强神经机器翻译模型捕捉更多信息和知识的能力, 这正是近年来的研究思路: 在神经机器翻译中融入更多的信息和知识. 目前探索了句法信息, 未来将进一步探索语义信息、句子间的篇章信息, 以及非语言学知识如世界知识、常识等.

目前基于 RNN 的编码器 – 解码器神经机器翻译框架本身存在若干挑战, 如前面提到的固有问题, 以及难以在 GPU 上实现高度并行化等. 因此神经机器翻译新框架的研究近年也得到越来越多的关注, 如用卷积神经网络 CNN, 或者用基于 self-attention 的无循环网络取代 RNN 作为机器翻译框架新的基本构建模块 [45, 46]. 在这方面, 未来工作既包括对基本构建模块的进一步研究, 也包括对编码器、解码器新连接方式的探索.

最后, 神经机器翻译不仅显著提升了译文质量, 而且也极大推动了机器翻译的发展, 从统计机器翻译到神经机器翻译的变迁可以看作是机器翻译研究历史上的一次重大技术变革. 然而神经机器翻译毕竟是深度学习驱动, 需要大量平行语料训练. 这对资源稀缺的语言对, 如汉语 – X 语言 (X 为除去几大语种之外的其他语言) 机器翻译构成了极大挑战, 目前这方面的工作主要是面向英语的 (如美国国防高级研究计划局 DARPA 资助的 LORELEI 计划: 面向突发事件的资源稀缺语言项目), 我们期待未来有更多面向汉语的资源稀缺语言对神经机器翻译工作开展.

致谢 感谢本文所介绍工作的所有合作者!

参考文献

1 Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks. In: Proceedings of Workshop on Neural Information Processing Systems, Montreal, 2014. 3104–3112

2 Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. In: Proceedings of International Conference on Learning Representations (ICLR), San Diego, 2015

3 Koehn P. Statistical Machine Translation. Cambridge: Cambridge University Press, 2009

4 Xiong D Y, Zhang M. Linguistically Motivated Statistical Machine Translation: Models and Algorithms. Berlin: Springer, 2015

5 Och F J, Ney H. Discriminative training and maximum entropy models for statistical machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Philadelphia, 2002. 295–302

6 Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality. In: Proceedings of Conference on Neural Information Processing Systems (NIPS), Lake Tahoe, 2013. 3111–3119

7 Junczys-Dowmunt M, Dwojak T, Hoang H. Is neural machine translation ready for deployment? a case study on 30 translation directions. 2016. ArXiv:1610.01108

8 Jean S, Firat O, Cho K, et al. Montreal neural machine translation systems for WMT15. In: Proceedings of the 10th Workshop on Statistical Machine Translation (WMT), Lisboa, 2015. 134–140

9 Wu Y H, Schuster M, Chen Z F, et al. Google's neural machine translation system: bridging the gap between human and machine translation. 2016. ArXiv:1609.08144

10 Kuang S H, Xiong D Y. Automatic long sentence segmentation for neural machine translation. In: Proceedings of Conference on Natural Language Processing and Chinese Computing (NLPCC), Kunming, 2016

11 Jean S, Cho K, Memisevic R, et al. On using very large target vocabulary for neural machine translation. In: Proceedings of the 53rd Annual Metting on Association for Computational Linguistics (ACL), Beijing, 2015

12 Tu Z P, Lu Z D, Liu Y, et al. Modeling coverage for neural machine translation. In: Proceedings of the 54th Annual Metting on Association for Computational Linguistics (ACL), Berlin, 2016. 76–85

13 Kingma D R, Welling M. Auto-encoding variational bayes. In: Proceedings of International Conference on Learning Representations (ICLR), Banff, 2014

14 Rezende D J, Mohamed S, Wierstra D. Stochastic backpropagation and approximate inference in deep generative models. In: Proceedings of the 31st International Conference on Machine Learning (ICML), Beijing, 2014. 1278–1286

15 Kingma D P, Mohamed S, Rezende D J, et al. Semi-supervised learning with deep generative models. In: Proceedings of Conference on Neural Information Processing Systems (NIPS), Montreal, 2014. 3581–3589

16 Chung J Y, Kastner K, Dinh L, et al. A recurrent latent variable model for sequential data. In: Proceedings of Conference on Neural Information Processing Systems (NIPS), Montreal, 2015. 2980–2988

17 Miao Y S, Yu L, Blunsom P. Neural variational inference for text processing. In: Proceedings of the 33nd International Conference on Machine Learning (ICML), New York, 2016. 1727–1736

18 Bowman S R, Vilnis L, Vinyals O, et al. Generating sentences from a continuous space. In: Proceedings of the SIGNLL Conference on Computational Natural Language Learning (CoNLL), Berlin, 2016. 10–21

19 Li Z F, Eisner J, Khudanpur S. Variational decoding for statistical machine translation. In: Proceedings of the 47th Annual Meeting on Association for Computational Linguistics (ACL), Singapore, 2009. 593–601

20 He W, He Z J, Wu H, et al. Improved neural machine translation with SMT features. In: Proceedings of the 30th AAAI Conference on Artificial Intelligence, Phoenix, 2016. 151–157

21 Stahlberg F, Hasler E, Waite A, et al. Syntactically guided neural machine translation. In: Proceedings of the 54th Annual Metting on Association for Computational Linguistics (ACL), Berlin, 2016. 299–305

22 Arthur P, Neubig G, Nakamura S. Incorporating discrete translation lexicons into neural machine translation. In: Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP), Austin, 2016. 1557–1567

23 Dahlmann L, Matusov E, Petrushkov P, et al. Neural machine translation leveraging phrase-based models in a hybrid search. In: Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP), Copenhagen, 2017. 1411–1420

24 Niehues J, Cho E, Ha T L, et al. Pre-translation for neural machine translation. In: Proceedings of the 26th

International Conference on Computational Linguistics (COLING), Osaka, 2016. 1828–1836
25 Zhou L, Hu W P, Zhang J J, et al. Neural system combination for machine translation. In: Proceedings of Annual Meeting on Association for Computational Linguistics (ACL), Vancouver, 2017. 378–384
26 Eriguchi A, Hashimoto K, Tsuruoka Y. Tree-to-sequence attentional neural machine translation. In: Proceedings of the 54th Annual Metting on Association for Computational Linguistics (ACL), Berlin, 2016. 823–833
27 Sennrich R, Haddow B. Linguistic input features improve neural machine translation. In: Proceedings of the 1st Conference on Machine Translation, Berlin, 2016. 83–91
28 Shi X, Padhi I, Knight K. Does string-based neural MT learn source syntax? In: Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP), Austin, 2016. 1526–1534
29 Wu S Z, Zhang D D, Yang N, et al. Sequence-to-dependency neural machine translation. In: Proceedings of Annual Meeting on Association for Computational Linguistics (ACL), Vancouver, 2017. 698–707
30 Zhang B, Xiong D, Su J S, et al. Variational neural machine translation. In: Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP), Austin, 2016. 521–530
31 Chung J Y, Gulcehre C, Cho J, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling. In: Proceedings of NIPS Deep Learning and Representation Learning Workshop, Montreal, 2014
32 Luong M T, Sutskever I, Quoc V. Addressing the rare word problem in neural machine translation. In: Proceedings of the SIGNLL Conference on Computational Natural Language Learning (CoNLL), Beijing, 2015. 11–19
33 Wang X, Lu Z D, Tu Z P, et al. Neural machine translation advised by statistical machine translation. In: Proceedings of the 31st AAAI Conference on Artificial Intelligence, San Francisco, 2017. 3330–3336
34 Wang X, Tu Z P, Xiong D Y, et al. Translating phrases in neural machine translation. In: Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP), Copenhagen, 2017. 1421–1431
35 Liu Y, Liu Q, Lin S X. Tree-to-string alignment template for statistical machine translation. In: Proceedings of the International Committee on Computational Linguistics and the Association for Computational Linguistics (COLING-ACL), Sydney, 2006. 609–616
36 Shen L B, Xu J X, Weischedel R. A new string-to-dependency machine translation algorithm with a target dependency language model. In: Proceedings of the Annual Meeting on Association for Computational Linguistics with the Human Language Technology Conference (ACL-HLT), Columbus, 2008. 577–585
37 Xiong D Y, Liu Q, Lin S X. Maximum entropy based phrase reordering model for statistical machine translation. In: Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics (ACL), Sydney, 2006. 521–528
38 Xiong D Y, Liu Q, Lin S X. A dependency treelet string correspondence model for statistical machine translation. In: Proceedings of the 2nd Workshop on Statistical Machine Translation (WMT), Prague, 2007. 40–47
39 Li J H, Resnik P, Daumé H. Modeling syntactic and semantic structures in hierarchical phrase-based translation. In: Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), Atlanta, 2013. 540–549
40 Marton Y, Resnik P. Soft syntactic constraints for hierarchical phrased-based translation. In: Proceedings of the Annual Meeting on Association for Computational Linguistics with the Human Language Technology Conference (ACL-HLT), Columbus, 2008. 1003–1011
41 Xiong D Y, Zhang M, Aw A, et al. Linguistically annotated reordering: evaluation and analysis. Comput Linguist, 2010, 36: 535–568
42 Li J H, Xiong D Y, Tu Z P, et al. Modeling source syntax for neural machine translation. In: Proceedings of Annual Meeting on Association for Computational Linguistics (ACL), Vancouver, 2017. 688–697
43 Choe D K, Charniak E. Parsing as language modeling. In: Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP), Austin, 2016. 2331–2336
44 Vinyals O, Kaiser L, Koo T, et al. Grammar as a foreign language. In: Proceedings of Conference on Neural Information Processing Systems (NIPS), Montreal, 2015
45 Gehring J, Auli M, Grangier D, et al. A convolutional encoder model for neural machine translation. In: Proceedings of Annual Meeting on Association for Computational Linguistics (ACL), Vancouver, 2017. 123–135
46 Ashish V, Noam S, Niki P, et al. Attention is all you need. In: Proceedings of the 31st Annual Conference on Neural Information Processing Systems (NIPS), Long Beach, 2017. 6000–6010

Neural machine translation with constraints

Deyi XIONG[1*], Junhui LI[1], Xing WANG[1] & Biao ZHANG[2]

1. *School of Computer Science and Technology, Soochow University, Suzhou* 215006*, China*;
2. *Software School, Xiamen University, Xiamen* 361005*, China*
* Corresponding author. E-mail: dyxiong@suda.edu.cn

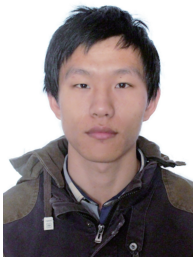
Abstract Neural machine translation (NMT), powered by deep learning, is an emerging machine translation paradigm that has been advancing rapidly in recent years. It has become mainstream technology in both academia and industry of machine translation. This paper provides an overview of our research work on NMT. It particularly focuses on a series of NMT models proposed for considering a variety of useful information and knowledge constraints, which include variational NMT with constraints of latent variables, NMT advised by statistical machine translation, and NMT with syntactical constraints from the source language. In addition to this overview, this paper presents an outlook of the future trends in NMT.

Keywords neural machine translation, variational neural machine translation, fusion of neural and statistical machine translation, neural machine translation with syntactical constraints

Deyi XIONG was born in 1979. He received his Ph.D. degree in 2007 from Institute of Computing Technology, Chinese Academy of Sciences, Beijing. Currently, he is a professor at Soochow University. His research interests include natural language processing, machine translation, and deep learning.

Junhui LI was born in 1983. He received his Ph.D. degree in 2010 from Soochow University, Suzhou. Currently, he is an associate professor at Soochow University. His research interests include natural language processing and machine translation.

Xing WANG was born in 1988. He is a Ph.D. candidate at Soochow University. He is supervised by Professor Min ZHANG and Professor Deyi XIONG. His research interests include statistical machine translation and neural machine translation.



Biao ZHANG was born in 1994. He received his Bachelor's degree in software engineering from Xiamen University. He is currently a graduate student in School of Software at Xiamen University. He is supervised by Professor Jinsong SU. His major research interests include natural language processing and deep learning.