综述评介

自然语言处理中的注意力机制研究综述*

石　磊[1]　王　毅[2]　成　颖[2,3]　魏瑞斌[1]

[1]（安徽财经大学管理科学与工程学院　蚌埠 233030）

[2]（南京大学信息管理学院　南京 210023）

[3]（山东师范大学文学院　济南 250014）

摘要：**【目的】**总结注意力机制在自然语言处理领域的衍化及应用规律。**【文献范围】**以“attention”和“注意力”为检索词，分别检索 WoS、The ACM Digital Library、arXiv 以及中国知网，时间跨度限定为 2015 年 1 月至 2019 年 10 月，制定标准人工筛选自然语言处理领域的文献，最终获得 68 篇相关文献。**【方法】**在深入分析文献的基础上，归纳注意力机制的通用形式，梳理其衍生类型，并基于数据对其在自然语言处理任务中的应用情况进行述评。**【结果】**注意力机制在自然语言处理中的应用集中于序列标注、文本分类、推理以及生成式任务，且任务和注意力机制的类型之间存在一定的适配规律。**【局限】**部分注意力机制和任务间的适配结论是通过模型整体表现数据间接得出的，不同注意力机制间的性能差异有待进一步研究。**【结论】**注意力机制的研究切实推进了自然语言处理的发展，但其作用机理尚未明了，提高其可解释性并使之更加接近人类的真实注意力是未来的研究方向。

关键词：注意力机制　自注意力　机器翻译　机器阅读理解　情感分析

分类号：TP391.1

DOI：10.11925/infotech.2096-3467.2019.1317

引用本文：石磊，王毅，成颖等. 自然语言处理中的注意力机制研究综述[J]. 数据分析与知识发现，2020，4（5）：1-14.(Shi Lei, Wang Yi, Cheng Ying,et al. Review of Attention Mechanism in Natural Language Processing [J]. Data Analysis and Knowledge Discovery, 2020, 4(5): 1-14.)

1　引　言

近年来，注意力机制（Attention Mechanism）成为深度学习领域的研究热点之一。注意力机制的思想源自人类视觉，即人眼通过快速扫描聚焦于需重点关注的目标区域，之后对该区域投入更多注意力，以获取所需的细节信息，同时抑制其他信息，是人类利用有限资源从大量信息中快速筛选出有价值信息的一种能力[1]。深度学习中的注意力机制模拟了该过程，即当神经网络发现输入数据的关键信息后，通过学习，在后继的预测阶段对其予以重点关注。注意力机制的首次应用是 Mnih 等[2]在图像分类研究中设计的瞥见（glimpse）算法，不同于全图扫描，该算法每次仅瞥见图像中的部分区域，并按时间顺序将多次瞥见的内容用循环神经网络加以整合，以建立图像的动态表示；该算法降低了时间复杂度，且减

通讯作者：魏瑞斌，ORCID：0000-0001-6271-7881，E-mail：rbwxy@126.com。

*本文系国家社会科学基金重大项目“中国近现代文学期刊全文数据库建设与研究（1872-1949）”（项目编号：17ZDA276）的研究成果之一。

少了噪声干扰，在图像分类任务中取得了显著成效。

和图像处理类似，自然语言处理模型在读取文本时可以重点关注文本中和任务相关的部分，忽略其他内容。根据该思想，Bahdanau等[3]将注意力机制应用于神经机器翻译（Neural Machine Translation，NMT）模型，即在生成译文的每个词项时，让模型找出原文中和当前词项最相关的部分，并据此进行预测。和先前基于固定原文表示进行预测的NMT模型相比，该方法不仅缓解了循环神经网络的长距离依赖问题，还实现了翻译过程中的词对齐（alignment），有效提高了译文的质量。随着注意力机制在机器翻译任务中的成功应用，该思想很快被推广到不同的自然语言处理任务中。

目前，注意力机制已经成为自然语言处理研究中不可或缺的重要组件，并形成了丰富的文献积累。总的来看，现有工作主要集中于算法及其改进的实证研究，检索到的两篇综述[4-5]主要介绍注意力机制的原理及分类。本文在深入分析现有文献的基础上归纳了注意力机制的通用形式，梳理了其衍生类型，并基于数据对其和自然语言处理任务间的适配情况进行述评，最后提出该领域面临的挑战及可能的研究方向，以期为后继研究提供参考。

2 文献范围

考虑到Bahdanau等[3]的工作出现于2015年，故本文的文献检索时间限定为2015年1月至2019年10月。以“attention”为检索词分别检索WoS、The ACM Digital Library以及arXiv数据库，经去重、剔除无效文献后，获得相关文献563篇；以“注意力”为检索词检索中国知网，经去重、剔除无效文献后，获得相关文献204篇。通过阅读文献题名与摘要，人工筛选自然语言处理领域的该主题文献，获得相关文献437篇，其中英文文献325篇，中文文献112篇。下载全文进一步筛选，根据文献聚焦于算法或应用将其分为两类，应用类文献按不同的自然语言处理任务进行细分。算法类文献的筛选标准是提出新算法或新思路，应用类文献的筛选标准是注意力机制在应用中起主要贡献，且实验结果具有可比性。共筛选出切题英文文献61篇，在阅读中根据参考文献进行扩展，补充英文文献7篇，最终获得相关文献68篇。

3 基础工作

3.1 奠基性工作

Bahdanau等[3]为解决基于序列到序列（Sequence to Sequence）的NMT模型中输入和输出难以对齐以及应对长文本效果不佳等问题，借鉴图像处理研究中的注意力思想，首次提出自然语言处理中的注意力机制。基本思想是在生成译文的每个词项时，计算出源文本词项序列的权重分布，找出源文本序列中和当前词项最相关的元素，使模型在预测时更具针对性。具体算法：先将编码器生成的源文本的隐层序列$(\boldsymbol{h}_1,\cdots,\boldsymbol{h}_n)$和上一时间步的解码器隐层向量$\boldsymbol{s}_{t-1}$进行匹配，计算隐层序列的权重分布$(\alpha_{t1},\cdots,\alpha_{tn})$；之后将$\alpha_{ti}$和$\boldsymbol{h}_i$加权求和得到带注意力的语义向量$\boldsymbol{c}_t$；解码器在每个时间步根据动态变化的$\boldsymbol{c}_t$逐个生成译文词项。对NMT模型的解码过程如图1所示，在生成词项“*changing*”时，虽然参考了源文本中的全部词项，但注意力机制为词项赋予了不同的权重，重点参考了“正”和“改变”。该工作将注意力机制作为连接编码器和解码器的桥梁，实现了翻译过程的词对齐，提高了模型生成译文的质量。

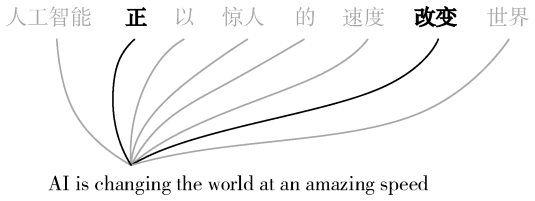

图1 带注意力机制的NMT模型示意图

Fig.1 Schematic Diagram of NMT Model with Attention Mechanism

3.2 通用形式

在奠基性工作中，包含已生成的译文信息的解码器隐层向量$\boldsymbol{s}_{t-1}$代表了下游任务，然而下游任务在不同的任务中有不同的表示形式，如分类任务中的标签、阅读理解任务中的问题、自然语言推理（Natural Language Inference，NLI）任务中相对应的句子等。如果将下游任务抽象成查询（query），就可以归纳出注意力机制的通用形式，即将源文本看成是键-值对序列，用$\boldsymbol{K}=(\boldsymbol{k}_1,\cdots,\boldsymbol{k}_N)$和$\boldsymbol{V}=(\boldsymbol{v}_1,\cdots,\boldsymbol{v}_N)$分别表示键序列和值序列，用$\boldsymbol{Q}=(\boldsymbol{q}_1,\cdots,\boldsymbol{q}_M)$表示查询

序列，那么针对查询$\boldsymbol{q}_t$的注意力可以被描述为键-值对序列在该查询上的映射[6]。如图2所示，计算过程可分为三步：

（1）计算查询$\boldsymbol{q}_t$和每个键$\boldsymbol{k}_i$的注意力得分e_{ti}，常用的计算方法包括点积[7]、缩放点积[6]、拼接[7]以及相加[3]等，如公式（1）所示；

（2）使用Softmax等函数对注意力得分做归一化处理，得到每个键的权重α_{ti}，如公式（2）所示；

（3）将权重α_{ti}和其对应的值$\boldsymbol{v}_i$加权求和作为注意力输出，如公式（3）所示。

$$e_{ti}=score(\boldsymbol{q}_t,\boldsymbol{k}_i)=\begin{cases}\boldsymbol{q}_t^{\mathrm{T}}\boldsymbol{k}_i & \text{点积}\\ \boldsymbol{q}_t^{\mathrm{T}}\boldsymbol{W}\boldsymbol{k}_i & \text{通用}\\ \dfrac{\boldsymbol{q}_t^{\mathrm{T}}\boldsymbol{k}_i}{\sqrt{d_k}} & \text{缩放点积}\\ \boldsymbol{v}^{\mathrm{T}}\tanh(\boldsymbol{W}[\boldsymbol{q}_t;\boldsymbol{k}_i]) & \text{拼接}\\ \boldsymbol{v}^{\mathrm{T}}\tanh(\boldsymbol{W}\boldsymbol{q}_t+\boldsymbol{U}\boldsymbol{k}_i) & \text{相加}\end{cases} \tag{1}$$

$$\alpha_{ti}=softmax(e_{ti})=\frac{\exp(e_{ti})}{\sum_{n=1}^{N}\exp(e_{tn})} \tag{2}$$

$$Attention(\boldsymbol{q}_t,\boldsymbol{K},\boldsymbol{V})=\sum_i\alpha_{ti}\boldsymbol{v}_i \tag{3}$$

其中，$\boldsymbol{W}$和$\boldsymbol{U}$代表可学习的参数矩阵，$\boldsymbol{v}$代表参数向量。

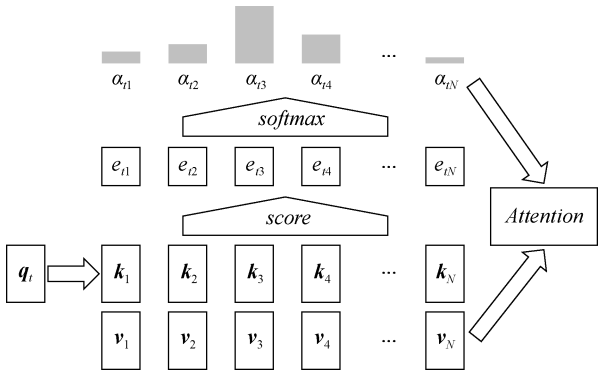

图2　注意力机制的通用形式

Fig.2　The General Form of Attention Mechanism

键-值对是源文本的组成元素，可以是字符、词、短语、句子等，甚至是它们的组合[8]。这些元素一般用向量表示，向量不仅是元素的内容表示，同时也是元素的唯一标识，在通常情况下$\boldsymbol{K}$=$\boldsymbol{V}$。模型输出的注意力是源文本序列基于查询$\boldsymbol{q}_t$的表示，不同的查询会给源文本序列带来不同的权重分布。注意力机制根据查询计算出源文本序列中与下游任务最相关的部分，意味着不同的查询会关注源文本的不同部分，因此注意力机制可以看成是一种基于查询的源文本表示方法，理论上适用于任何文本处理任务。

4　分　类

注意力机制在与各种自然语言处理任务的结合过程中，自身也不断得以改进和完善，衍生出不同的类型。本节从关注范围、组合方式等不同视角对注意力机制的衍生类型进行梳理。

4.1　关注范围

在Bahdanau等[3]的工作中，注意力在计算时考虑源文本的全部信息，即对源文本序列中的每个元素都计算其针对查询的得分和权重，也被称为全局注意力（Global Attention）[7]。全局注意力计算开销较大，且随着源文本长度的增加其效果会变弱，因此部分学者对注意力机制的关注范围进行改进研究。表1按照不同的关注范围对其进行归纳。

表1　注意力机制按照关注范围分类

Table 1　Classification of Attention Mechanism by Attention Range

注意力	关注范围
全局注意力	全部元素
局部注意力	以对齐位置为中心的窗口
硬注意力	一个元素
稀疏注意力	稀疏分布的部分元素
结构注意力	结构上相关的一系列元素

Luong等[7]提出局部注意力（Local Attention），即仅关注源文本序列的一个窗口区域。具体实现上，先基于查询$\boldsymbol{q}_t$预测一个源文本中的对齐位置p_t（见公式（4）），并以p_t为中心确定窗口$[p_t-D,p_t+D]$，算法后续步骤和全局注意力类似，不过仅计算窗口内部元素的权重分布，最后用高斯分布加强p_t附近的权重（见公式（5））。

$$p_t=L\cdot sigmoid(\boldsymbol{v}^{\mathrm{T}}\tanh(\boldsymbol{W}\boldsymbol{q}_t)) \tag{4}$$

$$\alpha_t'(l)=\alpha_t(l)\cdot\exp\left(-\frac{(l-p_t)^2}{2\sigma^2}\right) \tag{5}$$

其中，L代表源文本序列的长度，l代表窗口内元素的位置序号，标准差$\sigma=\dfrac{D}{2}$。

Luong 等[7]在 WMT2014 英–德翻译实验中发现，在模型其他部分不变的情况下局部注意力比全局注意力在BLEU指标上提高了0.9，证实了局部注意力在更小的计算开销下获得了更准确的译文，同时也说明全局注意力可能会带来噪声干扰。但是，局部注意力的计算依赖于对齐位置，而对齐位置的预测本身并不可靠[7]，会为模型带来更多的不确定性。仅有少量研究借鉴了局部注意力的思想，如Mirsamadi 等[9]利用局部注意力捕捉语音信号的特定区域以获取情感信息；Yang 等[10]利用高斯分布强化局部自注意力，加强了自注意力机制捕获短语结构的能力。值得注意的是，当 D=0 时，局部注意力仅关注源文本序列中的一个元素，这种情况也称为硬注意力（Hard Attention）[11]。尚未见单纯采用硬注意力的工作，硬注意力的思想源自图像处理领域，实践证实其不适用于自然语言处理，因为仅关注文本中一个元素的意义十分有限。

注意力在计算时通常采用 Softmax 函数计算源文本序列的权重分布，经 Softmax 函数计算出元素权重 $\alpha_{ti}\neq 0$，也就是说即使是和查询无关的元素也会被分配一个很小的权重，致使所有元素对预测都有或多或少的影响。Martins 等[12]认为这在实际应用中可能导致注意力分散等弊端，因此提出用 Sparsemax 函数替代 Softmax。从公式（6）可以看出，Sparsemax 函数返回的是注意力得分 e 在 K–1 维概率单纯形上的欧几里得投影，这些投影倾向于触及单纯形的边界，从而导致部分概率值会降为 0，进而得到稀疏的概率分布。Sparsemax 函数在保留 Softmax 大部分重要性质的同时具有产生稀疏分布的能力，可以使注意力机制只关注源文本序列中的部分元素，理论上使模型的注意力更加集中，但绝大多数研究依然采用 Softmax 函数，Sparsemax 函数仅在多标签分类任务中体现出优势[12]。

$$Sparsemax(e)=\underset{\alpha\in\Delta^{K-1}}{\operatorname{argmin}}\|\alpha-e\|^2 \tag{6}$$

无论是 Softmax 还是 Sparsemax 都没有考虑输出层的结构信息，即被关注的元素只有权重大小之分，没有上下文关系，如子序列关系或句法依赖关系。对此，Kim 等[13]提出结构注意力（Structured Attention），用条件随机场（CRF）重建注意力机制。CRF 不同于 Softmax，其计算出的不是各个元素的概率，而是一条路径的概率，使得模型可以选择出一系列具有关联的元素。尽管 Kim 等[13]的实验结果显示结构注意力在机器翻译、问题回答、NLI 等多个任务上都取得了不错的结果，但该机制鲜有后续研究，可能的原因是在很多任务中句法结构并不是必要的，找出关键元素即可满足大多数任务的需求。

就理论探讨而言，上述注意力机制各有优劣，但在现有的研究中，绝大部分工作采用全局方式计算注意力。全局注意力机制结构简单，且完全参数化（parameterization），易于嵌入到神经网络中进行训练，梯度经其反向传播到模型其他部分。尽管全局注意力计算开销相对较大，但在计算资源过剩的背景下，计算开销已不是研究人员的主要考虑。

4.2 组合方式

（1）层级注意力

文本自身具有一定的结构特征，如由字符、单词、句子、段落、语篇（discourse）、文档等构成的层级结构，在一些自然语言处理任务中有时需要考虑该特征。例如，在句子“*书是好书，但内容不适合孩子*”中，上半句带有积极情感，下半句则包含消极情感，而显然下半句才是这句话的重点。因此在文本分类等任务中不仅要考虑词项对分类的影响，还考虑到词项所处的上下文结构信息。Yang 等[14]基于这种考虑提出层级注意力（Hierarchical Attention），即将源文本分割为片段，分别计算每个片段的注意力，再基于片段的注意力计算全文的注意力，并以此作为源文本的最终表示。Yang 等[14]的工作在多个语料集上的分类准确率均超越了先前的方法，证实了层级注意力可以更好地捕捉源文本的结构特征。层级注意力机制常被用于长文本任务中，如文档分类、文档翻译、文档摘要等。在实际应用中，可以将源文本划分为不同粒度的区块（chunk），常见的区块包括句子、段落、语篇等。层级注意力在具体实现上有两种方式：一种是分别计算出元素和区块的权重分布，之后用区块权重对元素权重进行缩放（scale），再基于缩放后的元素权重计算源文本注意力[15]（见公式（7））；另一种是先基于元素计算出每个区块的注意力（见公式（8）），并将其作为区块的键和值，再基于区块计算源文本注意力[16]（见公式（9））。

$$\alpha_{mi}^{scale} = \frac{\alpha_{mi}\alpha_m}{\sum_{s,w}\alpha_{sw}\alpha_s} \quad (7)$$

$$\boldsymbol{c}_m^{chunk} = Attention(\boldsymbol{q}, \boldsymbol{K}_m, \boldsymbol{V}_m) \quad (8)$$

$$\boldsymbol{c}^{doc} = Attention(\boldsymbol{q}, \boldsymbol{C}^{chunk}, \boldsymbol{C}^{chunk}) \quad (9)$$

其中，α_m代表第m个区块的权重，α_{mi}代表第m个区块中的第i个元素的权重，s代表区块数，w代表每个区块内的元素数，$\boldsymbol{K}_m$和$\boldsymbol{V}_m$分别代表第m个区块中的键、值矩阵，$\boldsymbol{C}^{chunk}$代表所有区块的注意力组成的矩阵。

（2）双向注意力

通常，注意力机制只基于$\boldsymbol{Q}$计算$\boldsymbol{K}$的注意力（$K2Q$），但当$\boldsymbol{Q}$是文本序列时，也可以基于$\boldsymbol{K}$计算$\boldsymbol{Q}$的注意力（$Q2K$）。如在阅读理解任务中，源文本和问题都是文本序列，因此除了考虑源文本中每个元素对问题的重要性之外，还可以考虑问题中的每个元素对源文本的重要性。部分研究将$K2Q$与$Q2K$加以组合，但为其赋予了不同的名称，如双向注意力（Bi-Directional Attention）[17]、协同注意力（Co-Attention）[18]、双路注意力（Two-Way Attention）[19-20]、交互注意力（Interactive Attention）[21]、互注意力（Inter-Attention）[22]等，本文将其统称为双向注意力。

双向注意力在阅读理解、NLI等任务中应用广泛，该机制更好地理解两个序列间的关系。通常的做法是将$K2Q$与$Q2K$的注意力输出进行拼接或聚合得到最终的输出，模型据此进行推理。大量的实验结果证实双向注意力提高了此类模型的表现。Seo等[17]在阅读理解任务中对双向注意力进行简化测试（ablation study），发现去掉$K2Q$模型的表现下降了十几个百分点，去掉$Q2K$模型的表现仅小幅下降，表明$K2Q$注意力对于阅读理解更加重要。

双向注意力按计算粒度可以分为粗粒度和细粒度两类。粗粒度[18,21]是将其中一个序列压缩成向量，并作为查询，与另一个序列中的每个元素进行匹配，其计算过程和全局注意力一致，两个方向均是如此。细粒度[20,23]则是将两个序列中的所有元素相互匹配，得到注意力得分矩阵，之后分别按行和列对矩阵进行池化（pooling），得到两个序列的得分向量，再按公式（2）和公式（3）分别计算两个方向的注意力。双向注意力还可以按计算顺序分为并行（parallel）和交替（alternating）两种方式[18]。并行方式下，对两个序列的注意力计算完全对称且同时进行；交替方式则是先基于序列$\boldsymbol{Q}$计算出序列$\boldsymbol{K}$的注意力，再基于已生成的序列$\boldsymbol{K}$的注意力计算序列$\boldsymbol{Q}$的注意力。

（3）多头注意力

文本序列中每个元素的键向量和值向量都是高维向量，向量中的每个分量代表着元素不同方面的特征。当向量维度过高时，一次注意力计算很难充分捕捉元素的全部特征。对此，Vaswani等[6]提出多头注意力（Multi-Head Attention），即对文本序列并行做多次注意力计算，每次作为一个“头”。假设有h个头，每个头在计算时通过线性变换将$\boldsymbol{Q}$、$\boldsymbol{K}$以及$\boldsymbol{V}$的维度降为原先的$1/h$，即转换到一个子空间中，如公式（10）所示。线性变换的参数是可学习的，且每个头的参数不同，从而确保模型从不同的表示子空间学习到相关的特征[6]。最后将h个头的注意力结果进行拼接，再通过一次线性变换得到最终的注意力输出，如公式（11）所示。

$$head_i = Attention(\boldsymbol{Q}\boldsymbol{W}_i^Q, \boldsymbol{K}\boldsymbol{W}_i^K, \boldsymbol{V}\boldsymbol{W}_i^V) \quad (10)$$

$$MultiHead(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = Concat(head_1, \cdots, head_h)\boldsymbol{W}^O \quad (11)$$

其中，参数矩阵$\boldsymbol{W}_i^Q \in \mathbb{R}^{d_{model} \times d_k}$，$\boldsymbol{W}_i^K \in \mathbb{R}^{d_{model} \times d_k}$，$\boldsymbol{W}_i^V \in \mathbb{R}^{d_{model} \times d_v}$，$\boldsymbol{W}_i^O \in \mathbb{R}^{hd_v \times d_{model}}$，$d_{model}$是元素向量的维度，$d_k = d_v = d_{model}/h$。多头注意力类似于集成学习，每个头只负责输出$1/h$的文本特征，且相互独立。将所有头的学习结果进行整合从而获得比单个学习器更优的学习效果，且由于每个头都经过降维处理，不会增加太多计算开销。Vaswani等[6]在Transformer模型中将多头注意力分别用于基于自注意力网络的编码器和解码器内部，以及连接编码器和解码器的注意力层，是Transformer模型的重要组成部分。

尽管多头注意力机制在理论上具备从不同子空间捕捉信息的能力，但存在两方面的不足：一是缺乏一种机制保证不同的头能切实捕捉到不同的特征[24]；二是用简单的拼接加线性变换对多个头的输出进行聚合，可能无法充分发挥多头注意力的表示能力[25]。针对第一个问题，Li等[24]引入不一致正则化（disagreement regularization）项，作为模型训练的辅助目标，以鼓励多头之间的多样性。针对第二个问题，Li等[25]借鉴胶囊网络（CapsNet）中的协议路由（Routing-by-Agreement）算法[26]以改善多头注意力

的信息聚合。协议路由算法可以对H个输入胶囊中的信息进行分配，并整合到N个输出胶囊中。该算法根据输入胶囊和输出胶囊的一致性来迭代更新每个输入胶囊应该分配给每个输出胶囊的比例，最终将所有输出胶囊连在一起形成最终表示。Li等将输出不一致正则化项以及协议路由算法分别应用于Transformer模型，实验结果显示两种改进思路均有效提升了多头注意力的效果。

4.3 自注意力

（1）自注意力机制

注意力机制一般基于外部查询对源文本进行解读，试图找到和查询最匹配的元素，然而这种做法忽略了源文本自身的特征，即源文本内部元素间的关系。自注意力（Self-Attention）[27]是注意力机制在$\boldsymbol{Q}=\boldsymbol{K}$时的一种改进，即查询来自源文本序列自身，用于建模源文本序列内部元素间的依赖关系，以加强对源文本语义的理解。Cheng等[22]提出的内部注意力（Intra-Attention）是自注意力的思想启蒙，因此自注意力有时也被称为内部注意力。作者用源文本内部的每一个词项和其他所有词项进行匹配，并计算出注意力分布，用以发现一个词项和其他词项间存在的强依赖关系。结果发现基于内部注意力建立的语言模型在困惑度（perplexity）上的表现优于基线模型，在情感分析与NLI任务中也有突出表现。

自注意力不依赖于下游任务，改进了源文本序列的表示。Lin等[27]受此启发提出利用自注意力进行句子嵌入（sentence embedding），以加强句子的语义。首先通过双向LSTM获得句子的隐层序列$\boldsymbol{H}=(\boldsymbol{h}_1,\cdots,\boldsymbol{h}_N)$，之后采用自注意力计算句子所有元素间的权重矩阵$\boldsymbol{A}$（见公式（12）），进而得到句子的矩阵表示$\boldsymbol{M}=\boldsymbol{AH}$。不同于将句子嵌入为固定向量的传统做法，该工作利用自注意力将句子嵌入为矩阵，矩阵的每一行体现了句子针对相应元素的语义特征。实证结果显示该方法比固定向量法能丰富句子的语义，在三种不同的任务中均体现出其优势。

$$\boldsymbol{A}=softmax(\boldsymbol{W}_1\tanh(\boldsymbol{W}_2\boldsymbol{H}^{\mathrm{T}})) \tag{12}$$

Shen等[28]基于查询的粒度将自注意力分为Token2Token和Source2Token。其中，Token2Token属于细粒度，解析了文本序列中任意两个元素间的依赖关系，得到文本的矩阵表示；而Source2Token则将整个文本序列压缩为查询向量，然后和每个元素进行匹配，属于粗粒度，探索每个元素和整个文本间的全局依赖关系，得到文本的向量表示。

（2）自注意力网络

Vaswani等[6]认为自注意力具有直接建模文本元素间关系的能力，无需循环或卷积神经网络，依靠自注意力完成文本的编码，并据此构建自注意力网络。该网络由6个相同的层堆叠在一起，每个层由多头自注意力和全连接前馈网络（FFN）构成。自注意力网络直接基于词向量计算自注意力，在捕捉词项间依赖关系的同时完成编码，如公式（13）所示。此外，考虑到语境信息对建模的重要性，在预处理阶段采用位置嵌入（position embedding）为每个元素添加绝对位置信息以提升建模效果。在Vaswani等[6]构建的NMT模型Transformer中，自注意力网络取代了此类模型中常用的LSTM或CNN作为编码器和解码器，模型的表现达到了当时的最高水平。

$$SelfAttention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=softmax(\frac{\boldsymbol{QK}^{\mathrm{T}}}{\sqrt{d_k}})\boldsymbol{V} \tag{13}$$

①相对位置。Shaw等[29]认为自注意力网络中嵌入的绝对位置不如相对位置高效，提出在计算注意力时考虑元素的相对位置，即元素间的距离，如公式（14）所示。

$$\boldsymbol{z}_i=\sum_{j=1}^{n}softmax(\frac{(\boldsymbol{x}_i\boldsymbol{W}^Q)(\boldsymbol{x}_j\boldsymbol{W}^K+\boldsymbol{a}_{ij}^K)^{\mathrm{T}}}{\sqrt{d_z}})(\boldsymbol{x}_j\boldsymbol{W}^V+\boldsymbol{a}_{ij}^V) \tag{14}$$

其中，$\boldsymbol{a}_{ij}^K$和$\boldsymbol{a}_{ij}^V$分别代表元素$\boldsymbol{x}_j$的键和值相对于元素$\boldsymbol{x}_i$的位置，其值是元素间距离的向量表示。该工作假设在一定距离之外精确的相对位置信息没有意义，因此将元素间的最大距离限制为b，即元素间的距离超过b时将被裁（clip）为b（见公式（15）），最大距离限制还可以使模型更好地适应训练时未见过的序列长度。实验结果显示基于相对位置的Transformer模型在WMT2014英–德和英–法翻译任务中的表现均超过绝对位置；同时还发现将相对位置和绝对位置相结合并没有进一步提升译文的质量，表明相对位置可以取代绝对位置。

$$clip(j-i,b)=\max(-b,min(b,j-i)) \tag{15}$$

②局部建模。相对位置让模型在计算注意力时考虑到元素间的距离，但没有针对性地加强邻近元

素间的依赖关系。Yang等[10]认为提高邻近元素间的依赖关系有利于捕捉到有用的短语结构，因此提出利用可学习的高斯偏差（Gaussian Bias）来强化局部权重分布，如公式（16）所示。$\boldsymbol{G}$是一个掩码（mask）矩阵，其元素$\boldsymbol{G}_{ij} \in [0, -\infty)$，代表元素$\boldsymbol{x}_j$和需要得到更多关注的局部区域的中心位置的紧密程度，由预测得到的中心位置和窗口尺寸计算得到。实验证实相对位置和局部建模相结合可以进一步提升模型的表现。

$$\alpha_{ij} = \frac{\exp(e_{ij} + \boldsymbol{G}_{ij})}{\sum_{k=1}^{n} \exp(e_{ik} + \boldsymbol{G}_{ik})} \tag{16}$$

③定向。自注意力网络允许一个元素和其前后的任何元素进行匹配，但在一些自然语言处理任务中需要考虑序列的时间顺序。对此，Shen等[28]提出定向自注意力网络（Directional Self-Attention Network，DiSAN）以限定注意力的计算方向（前向或后向），使自注意力能像循环神经网络一样按时间顺序处理文本序列。DiSAN将位置掩码（positional mask）矩阵和自注意力得分矩阵相加来限制计算方向。位置掩码分为前向掩码和后向掩码。前向掩码矩阵中$i < j$的元素为0，其他为$-\infty$；和得分矩阵相加后，得分矩阵中$i \geqslant j$的元素会变成$-\infty$，其他元素不变；经Softmax函数归一化后$i \geqslant j$的元素权重为0，意味着不关注；此时元素$\boldsymbol{x}_i$只和比自己序号大的元素$\boldsymbol{x}_i$进行匹配，实现了注意力的前向计算；反之亦然。

5　应用及述评

自然语言处理任务大体可以分为序列标注、分类、推理和生成等类型，注意力机制不仅在上述任务中均有所应用，而且其应用范围已拓展到其他相关学科。本节按任务类型对注意力机制在不同任务中的应用情况以及适配机理进行述评。

5.1　分类任务

文本分类任务的核心问题之一是特征选择[30]。和先前的方法相比，注意力机制可以动态地为文本特征分配权重，使分类器可以有重点地利用特征信息。多项工作证实了注意力机制在分类任务中的有效性[12,14]。在文本分类任务中，方面情感分析（Aspect-Level Sentiment Analysis）任务对注意力机制的运用最具代表性。“方面”经常由多个词项构成。例如，分析句子“*机子运行很快，就是物流太慢*”在“*手机速度*”方面的情感极性时，“*手机*”和“*物流*”都具备“*速度*”属性。若想让模型的注意力聚焦到“*手机*”上，除了考虑句子中的每个元素对“方面”的重要性之外，“方面”中的每个元素对句子的重要性同样也有意义，因此双向注意力可能会在此类任务中发挥作用。表2整理了部分方面情感分析模型在相同数据集上的表现，可以看出，配备双向注意力的模型整体表现突出。值得注意的是，在方面情感分析中采用细粒度注意力的模型总体上表现更优，因为细粒度比粗粒度更能捕捉细微的情感差异[31]。尽管细粒度注意力会带来更大的计算开销，但由于“方面”中包含的词项一般不会太多，总的来说还是值得的。还可以看出，多个模型采用了语境化（contextualized）注意力，是此类任务所特有的注意力类型，其基本思想是“方面”左右两侧的上下文对情感极性的影响不同，因此可能需要分别“注意”。此外，对于长文本分类而言，层级注意力往往会取得更好的效果。

5.2　推理任务

在机器阅读理解以及NLI等推理任务中，既要考虑句子本身的语义，又要考虑句子之间的关系。对于机器阅读理解而言，让两个序列相互关注是较为普遍的做法。表3整理了部分机器阅读理解模型在SQuAD数据集上的表现，可以看出配备双向注意力模型的表现明显优于基线模型；2017年以后该主题的部分工作引入自注意力，进一步提升模型的表现。自注意力可以加强模型对文本自身的理解，在阅读理解中相当于脱离问题去阅读文本，这种做法可以避免带着问题去阅读时可能陷入的局部最优[42]。例如，根据上下文“*山姆走进厨房拿起披萨，随后回到客厅准备享用*”回答“*披萨现在在哪里？*”。双向注意力的计算结果可能会发现“*披萨*”和“*厨房*”更加相关。然而，自注意力探索的是上下文内部词项间的依赖关系，会将“*披萨*”和“*客厅*”联系起来，再结合双向注意力更有可能给出正确的回答。NLI任务中也存在类似的现象。表4整理了部分NLI模型在SNLI数据集上的表现，可以看出互注意力和内部注意力相结合的表现明显优于单独使用互注意力或内部注意力。合理的解释是在NLI任务中弄清句子本身的含义对于判断两个句子是否具有蕴涵关系具有积极意义。

表2　部分方面情感分析模型的表现

Table 2　The Performance of Aspect-Level Sentiment Analysis Models

作者	模型	情感极性准确率(%)			注意力
		Restaurant	Laptop	Twitter	
Wang等[32]	LSTM	74.3	66.5	66.5	无
Tang等[33]	TD-LSTM	75.6	68.1	70.8	语境化注意力
Wang等[32]	ATAE-LSTM	77.2	68.7	-	方面嵌入注意力
Ma等[21]	IAN	78.6	72.1	-	粗粒度交互注意力
Liu等[34]	BiLSTM-ATT-G	79.7	73.1	70.4	语境化注意力
Huang等[35]	AOA-LSTM	81.2	74.5	-	细粒度双向注意力
Fan等[36]	MGAN	81.2	75.4	72.5	多粒度双向注意力
Zheng等[37]	LCR-Rot	81.3	75.2	72.7	语境化粗粒度双向注意力
Li等[38]	HAPN	82.2	77.3	-	层级注意力
Song等[39]	AEN-BERT	83.1	80.0	74.7	多头自注意力网络

（注：Restaurant和Laptop数据集来自SemEval 2014 Task 4[40]；Twitter数据集来自ACL 2014[41]；情感极性指积极、中性、消极三类。）

表3　部分机器阅读理解模型在SQuAD数据集上的表现

Table 3　The Performance of Machine Reading Comprehension Models on SQuAD

作者	模型	Exact Match(%)	F1(%)	注意力
Wang等[43]	Match-LSTM	64.7	73.7	无
Xiong等[44]	DCN	66.2	75.9	协同注意力
Seo等[17]	BiDAF	68.0	77.3	双向注意力
Gong等[45]	Ruminating Reader	70.6	79.5	双向多跳注意力
Wang等[42]	R-Net	72.3	80.7	Self-Matching注意力
Peters等[46]	BiDAF+Self-Attention	72.1	81.1	双向注意力+自注意力
Liu等[47]	PhaseCond	72.6	81.4	K2Q+自注意力
Yu等[48]	QANet	76.2	84.6	协同注意力+自注意力
Wang等[49]	SLQA+	80.4	87.0	协同注意力+自注意力

（注：数据均为单模型（single model）测试结果。）

表4　部分NLI模型在SNLI数据集上的表现

Table 4　The Performance of NLI Models on SNLI

作者	模型	训练集准确率(%)	测试集准确率(%)	注意力
Bowman等[50]	300D LSTM Encoders	83.9	80.6	无
Rocktaschel等[19]	100D LSTM with Attention	85.3	83.5	双路注意力
Lin等[27]	300D Structured Self-Attentive Sentence Embedding	-	84.4	自注意力
Shen等[28]	300D Directional Self-Attention Network（DiSAN）	91.1	85.6	定向自注意力
Cheng等[22]	300D LSTMN Deep Fusion	-	85.7	互注意力+内部注意力
Im等[51]	300D Distance-based Self-Attention Network	89.6	86.3	定向+距离自注意力
Shen等[52]	300D ReSAN	92.6	86.3	软硬混合自注意力
Parikh等[53]	300D Intra-Sentence Attention	90.5	86.8	互注意力+内部注意力
Tay等[54]	300D CAFE（AVGMAX+300D HN）	89.8	88.5	互注意力+内部注意力

5.3　生成式任务

在神经机器翻译、生成式文本摘要、语音识别等生成式任务中，注意力机制一般被用作连接编码器和解码器的桥梁，使解码器在生成每个词项时都可以参考源序列中最相关的部分。多项工作证实了注意力机制在生成式任务中是不可或缺的[3,6]。然而

注意力机制不仅可以用于连接编码器和解码器，多头自注意力网络甚至可以替代LSTM或CNN完成编码和解码。表5对比了三种不同网络结构的NMT模型，其中基于多头自注意力网络的Transformer模型用更小的训练开销获得了更好的译文质量，其中在英-德翻译任务中甚至超过了集成（ensemble）模型的表现。该研究认为多头自注意力网络具有两个方面的优势[6]：一是自注意力可以无视距离直接捕捉所有词项间的依赖关系，相比之下，LSTM需要逐步循环才能得到，并且难以捕捉长距离依赖，而CNN则需要通过层叠来扩大感受野（receptive field）；二是多头自注意力网络的结构更加简单，计算开销也相对较小，而且和CNN一样不依赖于前一时刻的计算结果，可以并行计算。不过，Domhan[55]的研究表明多头自注意力网络在编码器端的作用比在解码器端的作用重要得多，解码器端即使替换成LSTM或CNN，模型的表现也未见明显下降。

在部分针对长文本的生成式任务研究中，采用层级注意力，试图利用文本的结构信息改善模型的表现，但没有取得理想的效果[15,56]。表6整理了两项生成式摘要工作的数据，其中Nallapati等[15]基于CNN/Daily Mail语料集就两种注意力进行对比实验，从实验结果可以看出：层级注意力和全局注意力相比并没有提高生成摘要的质量，甚至在ROUGE-L指标上还略有降低。而Cohan等[57]在面向超长文本语料集arXiv上的生成式摘要实验中体现出了层级注意力的优势。在机器翻译的相关研究中，也仅有一篇文档级翻译工作[58]采用层级注意力，提高了译文的连贯性和衔接性。可以看出，在生成式任务中，仅在处理超长文本时层级注意力才能发挥效用。一个可能的原因是神经网络的记忆能力有限，对于超长文本无能为力，而层级注意力恰好可以弥补这一不足；但在神经网络的处理能力之内，层级注意力效果不显著，只会增加无谓的计算开销。

表5　部分NMT模型在WMT14数据集上的表现

Table 5　The Performance of NMT Models on WMT14

作者	模型	网络	BLEU（%）		训练开销（FLOPs）	
			英-德	英-法	英-德	英-法
Wu等[59]	GNMT+RL	LSTM	24.6	39.92	2.3×10^{19}	1.4×10^{20}
	GNMT+RL（ensemble）		26.3	41.16	1.8×10^{20}	1.1×10^{21}
Gehring等[60]	ConvS2S	CNN	25.16	40.46	9.6×10^{18}	1.5×10^{20}
	ConvS2S（ensemble）		26.36	**41.29**	7.7×10^{19}	1.2×10^{21}
Vaswani等[6]	Transformer（big）	多头自注意力	**28.4**	**41**	2.3×10^{19}	

表6　层级注意力在部分生成式摘要任务上的表现

Table 6　The Performance of Hierarchical Attention Mechanism on Some Abstractive Summarization Tasks

作者	语料集	注意力	ROUGE-1（%）	ROUGE-2（%）	ROUGE-L（%）
Nallapati等[15]	CNN/Daily Mail	全局注意力	32.49	11.84	29.47
	平均文档/摘要词数：766/53	层级注意力（词-句）	32.75	12.21	29.01
Cohan等[57]	arXiv	全局注意力	32.06	9.04	25.16
	平均文档/摘要词数：4 938/220	层级注意力（词-语篇）	35.80	11.05	31.80

5.4　序列标注任务

序列标注是对单一文本序列中的元素以及元素间的关系进行的探索，任务中没有来自序列外部的查询，因此自注意力机制可以在该类任务中发挥一定的作用。自注意力机制在序列标注中的应用主要集中在命名实体识别、语义角色标注等任务上。在命名实体识别研究中，Cao等[61]利用自注意力机制捕捉词项的全局依赖并学习句子的内部结构，在两个数据集上的实验结果显示该方法比传统的基于CRF的方法取得了更好的效果。Cai等[62]在面向中文电子病历的命名实体识别研究中，利用自注意力机制擅长捕捉长距离依赖的特点，有效提高了长实体（如手术名称等）边界的识别率。在语义角色标注研究中，Tan等[63]、Zhang等[64]以及Strubell等[65]均

在模型中引入自注意力机制，帮助模型捕捉语义角色间的依赖关系。Devlin等[66]提出的语言模型BERT（Bidirectional Encoder Representations from Transformers）利用多头自注意力网络作为特征提取器，在序列标注等11项自然语言处理任务中取得了当时的最佳成绩。

5.5 其他应用

随着注意力机制在深度学习领域影响力的扩大，其思想也被其他相关学科借鉴。在引文推荐研究中，Ebesu等[67]提出神经引文网络（Neural Citation Network，NCN）。该模型利用基于时延神经网络（TDNN）的编码器分别对引文内容、施引作者以及被引作者信息进行编码，并在解码时利用注意力机制调整三个方面信息的权重，最后用基于GRU的解码器生成推荐文献的标题。实证结果表明该方法为引文推荐提供了新的研究方向。Yang等[68]在NCN的基础上增加对文献所属期刊的考虑，并用注意力机制权衡4个方面的信息，进一步提高模型的表现。在专利引用研究中，Ji等[69]利用序列到序列模型预测专利的前向引用（forward citations）序列。该模型用基于RNN的编码器分别对专利自身、专利受让人以及专利发明人的历史引用序列进行编码，并在解码时利用注意力机制调整三个序列间的依赖关系以提高预测的准确率。在链接预测研究中，Chi等[70]和Brochier等[71]均利用注意力权重度量网络节点间的相似度。Munkhdalai等[72]通过结合层级注意力的LSTM模型对引文的功能和情感进行分类。

5.6 可解释性

多项工作都提到注意力机制为深度学习模型提供了可解释性[14,19,44]，甚至提供热图（heatmap），以直观的方式展现模型在预测时关注了文本的哪些部分。然而Jain等[73]研究发现注意力模型学习到的权重和基于梯度度量出的特征重要性往往是不一致的，甚至完全不同的注意力分布却可以产生等价的预测。例如，根据上下文“*John travelled to the garden，Sandra travelled to the garden.*”回答“*Where is Sandra?*”，尽管机器给出了正确的回答，但注意力却聚焦在第一个“*garden*”上[73]，说明机器并没有很好地理解原文的语义。Serrano等[74]用中间表征擦除法（intermediate representation erasure）评估注意力权重是否可以用来解释输入对注意力层本身的相对重要性，也得出了相同的结论。上述研究表明，虽然注意力机制可以切实提高自然语言处理模型的性能，但缺乏为模型提供有意义的解释的能力，同时也说明注意力机制的作用机理目前尚未明了。可解释性对于人类理解模型的工作原理非常重要，随着网络深度的增加该问题也愈加紧迫，因此如何提高注意力机制的可解释性，并使之更加接近人类的真实注意力，可能是该领域未来的一个研究方向。

6 结 语

本文从分类和应用两个方面对2015年以来自然语言处理领域的注意力机制相关文献进行梳理，分类整理了注意力机制在不同任务中的实验数据，并基于数据对注意力机制和自然语言处理任务间的适配情况进行述评。鉴于其中多项工作达到了先进水平[6,66]，可以认为注意力机制的研究切实推进了自然语言处理的发展，尤其是自注意力机制，在多项工作中发挥了关键作用。本文的不足之处在于，部分注意力机制和任务间的适配结论是通过模型整体表现数据间接得出的，并未对不同注意力机制间的性能差异做出进一步分析，未来可以启动实证研究以弥补这方面的不足。

目前，注意力机制的可解释性尚需要进一步解决。Zhang等[75-76]基于人类真实注意力完成了关键短语的抽取，提升了抽取的效果，该工作提示人类真实注意力在该领域的研究与应用将为可解释性提供一个可能的途径。

参考文献：

[1] Kastner S, Ungerleider L G. Mechanisms of Visual Attention in the Human Cortex[J]. Annual Review of Neuroscience, 2000, 23 (1): 315-341.

[2] Mnih V, Heess N, Graves A, et al. Recurrent Models of Visual Attention[C]//Proceedings of the Conference of Neural Information Processing Systems 2014, Montreal, Canada. 2014.

[3] Bahdanau D, Cho K, Bengio Y. Neural Machine Translation by Jointly Learning to Align and Translate[C]//Proceedings of the International Conference on Learning Representations, San Diego, USA. 2015.

[4] Hu D. An Introductory Survey on Attention Mechanisms in NLP Problems[OL]. arXiv Preprint, arXiv :1811.05544.

[5] Chaudhari S, Polatkan G, Ramanath R, et al. An Attentive Survey of Attention Models[OL]. arXiv Preprint, arXiv: 1904.02874.

[6] Vaswani A, Shazeer N, Parmar N, et al. Attention is All You Need [C]// Proceedings of Conference of Neural Information Processing Systems, Long Beach, USA. 2017: 6000-6010.

[7] Luong M T, Pham H, Manning C D. Effective Approaches to Attention-based Neural Machine Translation[C] //Proceedings of the Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal. 2015: 1412-1421.

[8] Li Y, Kaiser L, Bengio S, et al. Area Attention[OL]. arXiv Preprint, arXiv :1810.10126.

[9] Mirsamadi S, Barsoum E, Zhang C. Automatic Speech Emotion Recognition Using Recurrent Neural Networks with Local Attention[C]// Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, New Orleans, USA. 2017.

[10] Yang B, Tu Z, Wong D F, et al. Modeling Localness for Self-Attention Networks[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium. 2018: 4449-4458.

[11] Xu K, Ba J, Kiros R, et al. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention[C]// Proceedings of the International Conference on Machine Learning, Lille, France. 2015: 2048-2057.

[12] Martins A F T, Astudillo R F. From Softmax to Sparsemax: A Sparse Model of Attention and Multi-Label Classification[C] // Proceedings of the International Conference on Machine Learning, New York, USA. 2016.

[13] Kim Y, Denton C, Hoang L, et al. Structured Attention Networks [C]// Proceedings of the International Conference on Learning Representations, Toulon, France. 2017.

[14] Yang Z, Yang D, Dyer C, et al. Hierarchical Attention Networks for Document Classification[C]// Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, USA. 2016.

[15] Nallapati R, Zhou B, Gulcehre C, et al. Abstractive Text Summarization Using Sequence-to-Sequence RNNs and Beyond [C]//Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, Berlin, Germany. 2016: 280-290.

[16] Celikyilmaz A, Bosselut A, He X, et al. Deep Communicating Agents for Abstractive Summarization[C]//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies Volume 1: Long Papers, New Orleans, Louisiana, USA. 2018: 1662-1675.

[17] Seo M, Kembhavi A, Farhadi A, et al. Bi-directional Attention Flow for Machine Comprehension[C]//Proceedings of the International Conference on Learning Representations, Toulon, France. 2017.

[18] Lu J, Yang J, Batra D, et al. Hierarchical Question-Image Co-Attention for Visual Question Answering[C]//Proceedings of the Neural Information Processing Systems, Barcelona, Spain. 2016: 289-297.

[19] Rocktaschel T, Grefenstette E, Hermann K M, et al. Reasoning About Entailment with Neural Attention[C]//Proceedings of the International Conference on Learning Representations, San Juan, Puerto Rico. 2016.

[20] dos Santos C, Tan M, Xiang B, et al. Attentive Pooling Networks [OL]. arXiv Preprint , arXiv :1602.03609.

[21] Ma D, Li S, Zhang X, et al. Interactive Attention Networks for Aspect-Level Sentiment Classification[C] //Proceedings of the 26th International Joint Conference on Artificial Intelligence, Melbourne, Australia. 2017: 4068-4074.

[22] Cheng J, Dong L, Lapata M. Long Short-term Memory-networks for Machine Reading[C]// Proceedings of the Conference on Empirical Methods in Natural Language Processing, Austin, Texas, USA. 2016: 551-561.

[23] Cui Y, Chen Z, Wei S, et al. Attention-over-Attention Neural Networks for Reading Comprehension[C]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics Volume 1: Long Papers, Vancouver, Canada. 2017: 593-602.

[24] Li J, Tu Z, Yang B, et al. Multi-Head Attention with Disagreement Regularization[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium. 2018: 2897-2903.

[25] Li J, Yang B, Dou Z Y, et al. Information Aggregation for Multi-Head Attention with Routing-by-Agreement[OL]. arXiv Preprint, arXiv: 1904.03100.

[26] Sabour S, Frosst N, Hinton G E. Dynamic Routing Between Capsules[C]// Proceedings of the Conference on Neural Information Processing Systems, Long Beach, USA. 2017.

[27] Lin Z, Feng M, dos Santos C N, et al. A Structured Self-attentive Sentence Embedding[C]// Proceedings of the International Conference on Learning Representations, Toulon, France. 2017.

[28] Shen T, Zhou T, Long G, et al. DiSAN: Directional Self-attention Network for RNN/CNN-free Language Understanding[C]// Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, Louisiana, USA. 2018: 5446-5455.

[29] Shaw P, Uszkoreit J, Vaswani A. Self-attention with Relative Position Representations[C]//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, New Orleans, Louisiana, USA. 2018: 464-468.

[30] 徐冠华, 赵景秀, 杨红亚, 等 . 文本特征提取方法研究综述[J].

软件导刊, 2018, 17(5): 13-18.(Xu Guanhua, Zhao Jingxiu, Yang Hongya, et al. A Review of Text Feature Extraction Methods[J]. Software Guide, 2018, 17(5): 13-18.)

[31] 李慧, 柴亚青 . 基于卷积神经网络的细粒度情感分析方法[J]. 数据分析与知识发现, 2019, 3(1): 95-103. (Li Hui, Chai Yaqing. Fine-Grained Sentiment Analysis Based on Convolutional Neural Network[J]. Data Analysis and Knowledge Discovery, 2019, 3 (1): 95-103.)

[32] Wang Y, Huang M, Zhao L, et al. Attention-based LSTM for Aspect-level Sentiment Classification[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, Austin, Texas, USA. 2016: 606-615.

[33] Tang D, Qin B, Feng X, et al. Effective LSTMs for Target-Dependent Sentiment Classification[C]// Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers, Osaka, Japan. 2016: 3298-3307.

[34] Liu J, Zhang Y. Attention Modeling for Targeted Sentiment[C]// Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, Valencia, Spain. 2017: 572-577.

[35] Huang B, Ou Y, Carley K M. Aspect Level Sentiment Classification with Attention-over-Attention Neural Networks [C]//Proceedings of International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation, Washington, DC, USA. 2018: 197-206.

[36] Fan F, Feng Y, Zhao D. Multi-grained Attention Network for Aspect-Level Sentiment Classification[C]// Proceedings of the Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium. 2018: 3433-3442.

[37] Zheng S, Xia R. Left-Center-Right Separated Neural Network for Aspect-based Sentiment Analysis with Rotatory Attention[OL]. arXiv Preprint, arXiv: 1802.00892.

[38] Li L, Liu Y, Zhou A. Hierarchical Attention Based Position-aware Network for Aspect-level Sentiment Analysis[C]//Proceedings of the 22nd Conference on Computational Natural Language Learning, Brussels, Belgium. 2018: 181-189.

[39] Song Y, Wang J, Jiang T, et al. Attentional Encoder Network for Targeted Sentiment Classification[OL]. arXiv Preprint, arXiv: 1902.09314.

[40] Pontiki M, Galanis D, Pavlopoulos J, et al. Semeval-2014 Task 4: Aspect Based Sentiment Analysis[C]//Proceedings of the 8th International Workshop on Semantic Evaluation, Dublin, Ireland. 2014: 27-35.

[41] Dong L, Wei F, Tan C, et al. Adaptive Recursive Neural Network for Target-dependent Twitter Sentiment Classification[C]// Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics Volume 2: Short Papers, Baltimore, USA. 2014: 49-54.

[42] Wang W, Yang N, Wei F. Gated Self-Matching Networks for Reading Comprehension and Question Answering[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, Canada. 2017: 189-198.

[43] Wang S, Jiang J. Machine Comprehension Using Match-LSTM and Answer Pointer[OL]. arXiv Preprint, arXiv:1608.07905.

[44] Xiong C, Zhong V, Socher R. Dynamic Coattention Networks for Question Answering[OL]. arXiv Preprint, arXiv:1611.01604.

[45] Gong Y, Bowman S R. Ruminating Reader: Reasoning with Gated Multi-hop Attention[C]//Proceedings of the Workshop on Machine Reading for Question Answering, Melbourne, Australia. 2018:1-11.

[46] Peters M E, Neumann M, Iyyer M, et al. Deep Contextualized Word Representations[C]// Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, New Orleans, Louisiana, USA. 2018: 2227-2237.

[47] Liu R, Wei W, Mao W, et al. Phase Conductor on Multi-Layered Attentions for Machine Comprehension[OL]. arXiv Preprint, arXiv:1710.10504.

[48] Yu A W, Dohan D, Luong M T, et al. QANET: Combining Local Convolution with Global Self-Attention for Reading Comprehension [C]// Proceedings of the 6th International Conference on Learning Representations, Vancouver, Canada. 2018.

[49] Wang W, Yan M, Wu C. Multi-Granularity Hierarchical Attention Fusion Networks for Reading Comprehension and Question Answering[C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers), Melbourne, Australia. 2018: 1705-1714.

[50] Bowman S R, Gauthier J, Rastogi A, et al. A Fast Unified Model for Parsing and Sentence Understanding[C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics Volume 1: Long Papers, Berlin, Germany. 2016: 1466-1474.

[51] Im J, Cho S. Distance-based Self-Attention Network for Natural Language Inference[OL]. arXiv Preprint, arXiv: 1712.02047.

[52] Shen T, Zhou T, Long G, et al. Reinforced Self-Attention Network: A Hybrid of Hard and Soft Attention for Sequence Modeling[C]//Proceedings of the 27th International Joint Conference on Artificial Intelligence, Stockholm, Sweden. 2018: 4345-4352.

[53] Parikh A P, Täckström O, Das D, et al. A Decomposable Attention Model for Natural Language Inference[C]// Proceedings of the Conference on Empirical Methods in Natural Language Processing, Austin, Texas, USA. 2016: 2249-2255.

[54] Tay Y, Tuan L A, Hui S C. Compare, Compress and Propagate: Enhancing Neural Architectures with Alignment Factorization for

Natural Language Inference[OL]. arXiv Preprint, arXiv: 1801.00102.

[55] Domhan T. How Much Attention do You Need? A Granular Analysis of Neural Machine Translation Architectures[C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics: Long Papers, Melbourne, Australia. 2018: 1799-1808.

[56] Ling J, Rush A. Coarse-to-Fine Attention Models for Document Summarization[C] //Proceedings of the Workshop on New Frontiers in Summarization, Copenhagen, Denmark. 2017: 33-42.

[57] Cohan A, Dernoncourt F, Kim D S, et al. A Discourse-Aware Attention Model for Abstractive Summarization of Long Documents[C]// Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies Volume 2: Short Papers, New Orleans, Louisiana, USA. 2018: 615-621.

[58] Miculicich L, Ram D, Pappas N, et al. Document-Level Neural Machine Translation with Hierarchical Attention Networks[C]// Proceedings of the Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium. 2018: 2947-2954.

[59] Wu Y, Schuster M, Chen Z, et al. Google's Neural Machine Translation System: Bridging the Gap Between Human and Machine Translation[OL]. arXiv Preprint, arXiv: 1609.08144.

[60] Gehring J, Auli M, Grangier D, et al. Convolutional Sequence to Sequence Learning[C]//Proceedings of the International Conference on Machine Learning, Cancun, Mexico. 2017: 1243-1252.

[61] Cao P, Chen Y, Liu K, et al. Adversarial Transfer Learning for Chinese Named Entity Recognition with Self-Attention Mechanism [C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium. 2018: 182-192.

[62] Cai X, Dong S, Hu J. A Deep Learning Model Incorporating Part of Speech and Self-Matching Attention for Named Entity Recognition of Chinese Electronic Medical Records[J]. BMC Medical Informatics and Decision Making, 2019, 19(S2): 101-109.

[63] Tan Z, Wang M, Xie J, et al. Deep Semantic Role Labeling with Self-Attention[OL]. arXiv Preprint, arXiv: 1712.01586.

[64] Zhang Z, He S, Li Z, et al. Attentive Semantic Role Labeling with Boundary Indicator[OL]. arXiv Preprint, arXiv: 1809.02796.

[65] Strubell E, Verga P, Andor D, et al. Linguistically-Informed Self-Attention for Semantic Role Labeling[OL]. arXiv Preprint, arXiv: 1804.08199.

[66] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[C]// Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, Minnesota. 2019: 4171-4186.

[67] Ebesu T, Fang Y. Neural Citation Network for Context-Aware Citation Recommendation[C]//Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Tokyo, Japan. 2017: 1093-1096.

[68] Yang L, Zhang Z, Cai X, et al. Attention-Based Personalized Encoder-Decoder Model for Local Citation Recommendation[J]. Computational Intelligence and Neuroscience, 2019. Article ID 1232581.

[69] Ji T, Chen Z, Self N, et al. Patent Citation Dynamics Modeling via Multi-Attention Recurrent Networks[OL]. arXiv Preprint, arXiv:1905.10022.

[70] Chi Y, Liu Y. Link Prediction Based on Supernetwork Model and Attention Mechanism[C]//Proceedings of the 19th International Symposium on Knowledge and Systems Sciences, Tokyo, Japan. 2018: 201-214.

[71] Brochier R, Guille A, Velcin J. Link Prediction with Mutual Attention for Text-Attributed Networks[OL]. arXiv Preprint, arXiv:1902.11054.

[72] Munkhdalai T, Lalor J, Yu H. Citation Analysis with Neural Attention Models[C]//Proceedings of the 7th International Workshop on Health Text Mining and Information Analysis, Austin, USA. 2016: 69-77.

[73] Jain S, Wallace B C. Attention is not Explanation[OL]. arXiv Preprint, arXiv:1902.10186.

[74] Serrano S, Smith N A. Is Attention Interpretable? [OL]. arXiv Preprint, arXiv: 1906.03731.

[75] Zhang Y, Zhang C. Unsupervised Keyphrase Extraction in Academic Publications Using Human Attention[C] //Proceedings of the 17th International Conference on Scientometrics and Informetrics, Rome, Italy. 2019.

[76] Zhang Y, Zhang C. Using Human Attention to Extract Keyphrase from Microblog Post[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy. 2019.

作者贡献声明：

石磊：提出研究主题，起草论文；
王毅：提供自然语言处理方面的技术支持；
成颖：对关键学术内容做出修改；
魏瑞斌：论文最终审阅及修订。

利益冲突声明：

所有作者声明不存在利益冲突关系。

支撑数据：

支撑数据由作者自存储，E-mail: mrwonderful@163. com。

[1] 石磊. attention_mechanism_nlp_paper. zip. 自然语言处理中的注意力机制相关论文集.

收稿日期:2019-12-10
收修改稿日期:2019-12-22

Review of Attention Mechanism in Natural Language Processing

Shi Lei[1] Wang Yi[2] Cheng Ying[2, 3] Wei Ruibin[1]

[1](School of Management Science and Engineering, Anhui University of Finance and Economics, Bengbu 233030, China)

[2](School of Information Management, Nanjing University, Nanjing 210023, China)

[3](School of Chinese Language and Literature, Shandong Normal University, Jinan 250014, China)

Abstract: [**Objective**] This paper summarizes the evolution and application of attention mechanism in natural language processing. [**Coverage**] We searched "attention" with the title/topic fields of WoS, ACM Digital Library, arXiv and CNKI from January 2015 to October 2019. Then, we manually screened the topic literature in the field of natural language processing, and obtained 68 related papers. [**Methods**] We first summarized the general attention mechanism, and sorted out its derivations. Second, we thoroughly reviewed their applications in natural language processing tasks. [**Results**] The application of attention mechanism in natural language processing focused on sequence labeling, text classification, reasoning and generative tasks. There were adaptation rules between tasks and the various attention mechanisms. [**Limitations**] Some adaptations between the mechanisms and the tasks were obtained from the overall performance of the model. More research is needed to examine the performance of different mechanisms. [**Conclusions**] The study of attention mechanism has effectively promoted the development of natural language processing. However, the mechanism of action is not yet clear. Future research should focus on making attention mechanism closer to those of the human beings.

Keywords: Attention Mechanism Self-Attention Machine Translation Machine Reading Comprehension Sentiment Analysis