Contents lists available at ScienceDirect

Engineering

journal homepage: www.elsevier.com/locate/eng

Research
Artificial Intelligence—Review

神经自然语言处理最新进展——模型、训练和推理

周明，段楠，刘树杰，沈向洋 *

Microsoft Research Asia, Beijing 100080, China

ARTICLE INFO

Article history:
Received 30 April 2019
Revised 30 August 2019
Accepted 13 October 2019
Available online 7 January 2020

关键词
自然语言处理
深度学习
建模、学习和推理

摘要

自然语言处理（natural language processing, NLP）是人工智能研究的一个重要领域，旨在构建能够理解和生成自然语言、实现人机自然交互的技术方案。近5年，基于神经网络的自然语言处理方法取得突飞猛进的发展。基于海量无标注数据和大量标注数据进行建模，使得机器翻译、自动问答和阅读理解等很多任务的水准都得到了极大的提高。本文将从3个角度回顾神经自然语言处理的最新进展，包括模型、训练和推理。在模型部分，我们将介绍典型的神经网络建模方法，包括词嵌入建模、句子嵌入建模和序列到序列建模等。在训练部分，我们将介绍常用的学习方法，包括监督学习、半监督学习、无监督学习、多任务学习、迁移学习和主动学习等。在推理部分，我们将介绍典型的推理框架，包括非神经网络方法和神经网络方法。之所以强调推理方面的研究，是因为推理是构建基于知识的可解释自然语言处理模型的关键技术。本文的最后将概括介绍我们对自然语言处理未来发展方向的一些思考。

1. 引言

作为人工智能（artificial intelligence, AI）的重要分支，自然语言处理（natural language processing, NLP）研究人与机器之间如何通过自然语言进行交互。自然语言处理研究范围包括词、短语、句子和文档的处理（如分词、句法分析、语义分析）以及机器翻译、问答系统、信息检索、对话、文本生成和推荐系统等应用。自然语言处理对搜索引擎、智能客服、商业智能和语音助手等都具有至关重要的作用。

自然语言处理研究可以追溯到1950年。在研究初期，基于规则的方法被用来构建自然语言处理的各类系统，包括词法句法分析、问答系统和机器翻译。这类方法的缺点在于规则的设计需要来自领域专家的大量工作。此外，当规则数量越来越多后，如何对已有规则进行管理和组织，也成为制约规则系统发展的一大障碍。步入20世纪90年代，随着互联网的发展，大规模训练数据的获取成为可能，这促进了基于统计自然语言处理方法的产生和普及。基于人工设计的特征，统计机器学习方法在自然语言处理的各个任务上（如机器翻译）均带来了显著性的性能提升。2012年以来，随着深度学习在图像[1]和语音处理[2]领域的突破性进展，深度学习方法也被引入自然语言处理领域，并很快超过了传统的统计模型。如今，深度学习方法被全面用于自然语言处理的各个领域，并在某些任务（如机器翻译和机器阅读理解）上达到了前所未有的水平。例如，微软的Bible系统在中英新闻领域的机器翻译

* Corresponding author.
E-mail address: hshum@microsoft.com (H.Y. Shum)

2095-8099/
英文原文: Engineering 2020, 6(3): 275–290
引用本文: Ming Zhou, Nan Duan, Shujie Liu, Heung-Yeung Shum. Progress in Neural NLP: Modeling, Learning, and Reasoning. *Engineering*, https://doi.org/10.1016/j.eng.2019.12.014

任务（WMT 2017）上，达到了与人类媲美的水平。微软亚洲研究院（Microsoft Research Asia, MSRA）的R-NET和NLNet系统在机器阅读理解任务（Stanford question answering dataset, SQuAD）上同时在准确匹配（exact match, EM）和模糊匹配（F_1）方面达到了人类水平。包括预训练生成（generative pre-training, GPT）[3]、来自transformers的双向编码器表示（bidirectional encoder representations from transformers, BERT）[4]和XLNet等[5]在内的预训练模型在自然语言处理的各个任务上也表现出了强大的能力。但同时值得注意的是：虽然神经网络在训练样本足够的任务上表现良好，但该类方法依然在低资源或无资源任务上表现得不尽如人意。

基于上述背景，本文将从3个方面回顾神经自然语言处理的最新进展：①常见的神经网络模型；②常见的训练方法；③基于知识的推理。通过在这三方面对当前的技术和挑战进行深入分析，我们试图探索和指出推动自然语言处理继续前进的未来研究方向。

2. 神经网络模型

在自然语言处理的不同任务中（分类任务、序列标注任务和生成任务），句子通常被作为基本的输入单位。当在这些任务中使用基于神经网络的方法时，需要解决如下两个关键问题：

（1）如何使用神经网络对自然语言句子进行编码；

（2）如何使用神经网络产生输入句子对应的标签序列或输出句子。

本节从这两个角度出发，将介绍常用的神经网络建模方法，包括词嵌入模型、句子嵌入模型和序列到序列建模。词嵌入模型将自然语言句子中的单词映射到连续的语义空间。基于词嵌入的语义空间表示，包括循环神经网络、卷积神经网络和自注意力网络等在内的方法才能生成考虑全句上下文的词嵌入表示或句子嵌入表示。词嵌入表示在词性预测（part-of-speech, POS）和命名实体识别（named-entity recognition, NER）等任务中有广泛的应用。句子嵌入通常被用于句子级任务，如情感分析或者复述判别。句子嵌入还可以用于循环神经网络或卷积神经网络，用来完成序列到序列间的转化任务，也就是我们常说的编码器-解码器的框架。给定一个输入句子，序列到序列模型可以用来生成问题对应的答案（问答系统），也可以用来生成源语言句子对应的目标语言翻译（机器翻译）。

2.1. 词嵌入和句子嵌入

词嵌入（指词的向量表示）和句子嵌入（指句子的向量表示）是将词或句子从离散空间映射到连续语义空间，使得在该空间内，相似的词或句子具有相似的向量表示。

2.1.1. 上下文无关的词嵌入技术

为了将一个词映射到一个连续的语义空间，Mikolov等[6]提出了CBOW和Skip-gram模型。开源工具word2vec实现了该论文提出的方法，并能够基于大规模单语数据学习词向量表示。如图1 [6]所示，CBOW模型利用给定窗口内的其他单词来预测窗口中心的单词；与之相反，Skip-gram利用窗口中间的单词来预测窗口内的其他单词。这两个模型都是基于如下原则：“一个词的语义由所有出现在它周围的那些词决定”[7]。利用全局的共现信息和线性子结构，Pennington等[8]提出全局对数双线性回归模型（global log-bilinear regression model, GloVe）方法，对词嵌入进行了进一步的改进。

在Word2vec和GloVe方法中，同一个单词对应的词嵌入表示在不同句子中是不变的。例如，无论“bank”这个单词出现在“An ant went to the river bank”中，还是出现在“It is a good way to build up a bank account.”中，它对应的词嵌入表示是一样的。然而，“bank”在这两个句子中所表达的意思完全不同，它们的词嵌入也应该不一样。为了解决这个问题，句子里的上下文信息应该被引入词嵌入的生成过程中。

2.1.2. 基于循环神经网络的上下文相关词嵌入

ELMo [9]利用双向循环神经网络（recurrent neural network, RNN）对上下文信息进行建模来生成上下文相关的词嵌入。前向循环神经网络对该词左边的所有单词进行建模，反向循环神经网络对该词右边的所有单词进行建模。通过将得到的正向和反向两个词嵌入向量进行拼接，可以得到以该词对应的上下文敏感的词嵌入表示。例如，如图2所示，给定输入句子“an ant went to the bank of the river.”，前向循环神经网络首先接受第一个词“an”作为输入来产生第一个隐含状态。当第二个词“ant”作为输入时，循环神经网络能够将第二个词的信息同第一个词的信息进行融合获得前两个词的上下文表示。循环神经网络继续接收后边的词，直到碰到“bank”这个词时，前一个隐含状态应该能够包含该词之前所有词（“an ant went to the river”）的信息。将这些信息考虑进来，新的隐含状态

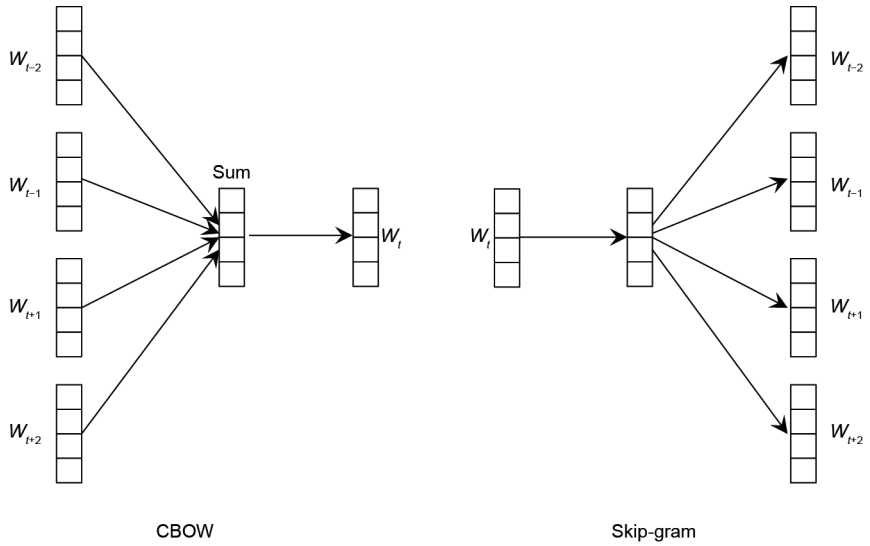

图1. 上下文无关的单词嵌入方法[6]。CBOW：使用窗口中的上下文词来预测中心词；Skip-gram：使用中心词来预测窗口中的上下文词；W_t是句子中的第t个单词。

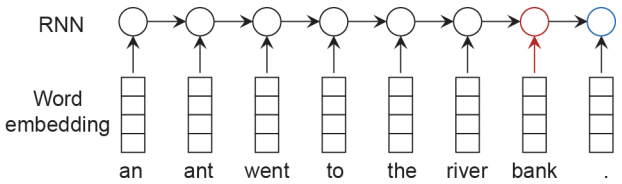

图2. 基于RNN的上下文感知词嵌入。

将能够生成考虑该句上下文信息的动态的词嵌入表示。

2.1.3. 基于自注意力网络的上下文相关词嵌入

基于自注意力网络，Radford等[3]提出了使用GPT来训练一个多层的从左到右的语言模型。同ELMo中使用的循环神经网络相比，尽管GPT仍然是一个从左到右的语言模型，然而通过自注意力机制，GPT允许词与词之间直接进行交互，从而能够更好地对上下文信息进行建模。Devlin等[4]提出基于自注意力网络的BERT模型。与GPT不同，该模型在生成词嵌入时，同时考虑句子中每个单词左右所有的上下文信息。这就不同于循环神经网络顺序从左到右或者从右到左的处理方式。图3给出一个例子，自注意力网络首先计算当前词（“bank”）同包含当前词在内的所有单词的匹配度。该匹配度经过归一化后，对每个单词对应的隐含状态进行加权求和，从而得到考虑所有上下文的词嵌入表示。为了引入输入单词之间的位置信息，位置嵌入被引入进来，同词嵌入一起作为自注意网络的输入。为了解决BERT在预训练和微调阶段的不一致问题(即预训练阶段的特殊符号MASK在微调阶段并不会出现)，Yang等[5]提出XLNet模型。该模型基于输入序列的排列组合，在自左向右的语言模型中引入了双向上下文信息。

2.1.4. 基于卷积神经网络的上下文相关词嵌入

ELMo和BERT都是考虑整句信息，用来生成动态词嵌入。不同于考虑整个句子中所有的词，卷积神经网络可以利用一个滑动窗口在输入句子上进行滑动，并将窗口内单词对应的词向量通过线性映射产生一个局部上下文信息的向量[10]。如图4所示，为了产生“bank”这个词的动态词嵌入，可以使用大小为3的窗口来覆盖片段“river bank.”。这样一来，“river”这个词就会被考虑到“bank”这个词对应的词嵌入表示中。

2.1.5. 句子嵌入

基于句子里每个词的词嵌入表示，可以利用神经网络（比如循环神经网络、自关注网络或者卷积神经网络）来获取句子嵌入表示。Collobert等[10]使用最大池化层来基于各个语义向量在每一维度上选择最大值来构成一个新的同输入向量长度相同的向量，并进一步经过一个前馈神经网络来生成该句子的词嵌入表示。这种句子的嵌入表示可以应用在其他任务上，比如分类问题（情感预测）或者生成一个句子（机器翻译）。除了使用卷积神经网络之外，循环神经网络和自关注网络也可以被用来产生一个语义向量来表示整个输入句子。对循环神经网络来说，最后一个隐含状态由于整合了句子里所有单词，从而可以看做是整个输入句子对应的向量表示。对于自注意网络来说，特殊

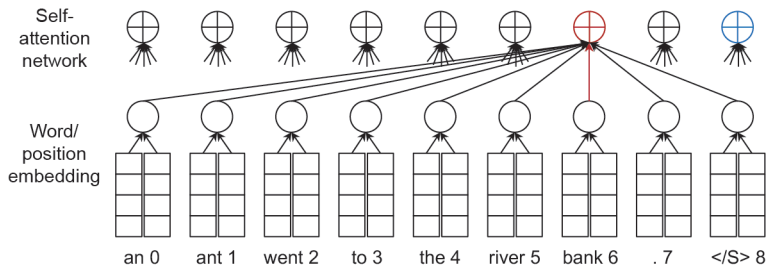

图3. 基于自注意力的上下文感知词嵌入。</S>：句子结尾的符号。

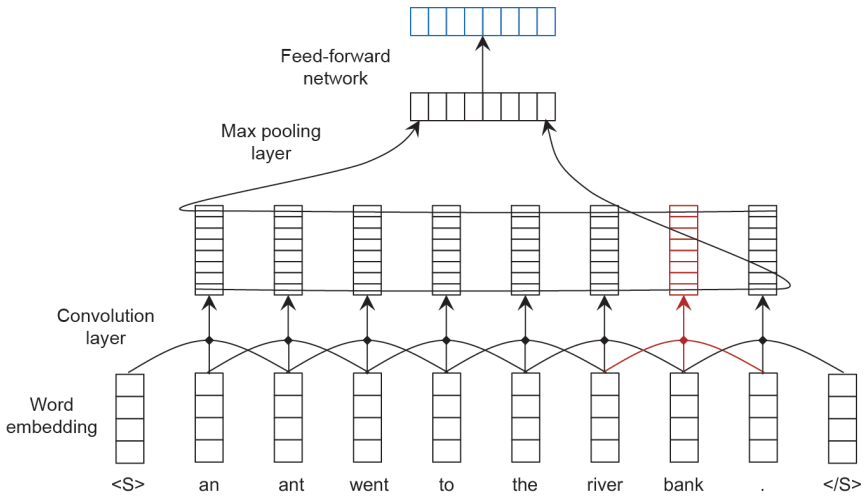

图4. 基于CNN的上下文感知词嵌入。

符号</S>被添加进句子里，它的对应隐状态（图4中的蓝色部分）可以作为整个句子的嵌入表示。

2.2. 序列到序列的转换模型

序列到序列的转换模型能够将一个输入串经过神经网络处理后转换为另外一个串。许多的自然语言处理任务都可以被规范化成一个序列到序列的转换任务，比如机器翻译、问答系统、对话系统等。

2.2.1. 编码器 - 解码器框架

Cho等[11]提出了编码器-解码器框架来处理序列到序列的转换任务。如图5所示，编码器-解码器框架包含两个部分：一个编码器和一个解码器。基于循环神经网络的编码器对输入句子里的词从左到右进行处理，最后得到的隐含状态包含了句子全部的信息。基于该信息，基于循环神经网络的解码器依次产生目标句子里的每个词，直至生成句子结束符（</S>）为止。在每一步中，解码器将前一个产生的词、前一个隐含状态、源语言句子的上下文向量一同作为循环神经网络的输入来预测当前的目标语言词。

该编码器解码器框架有如下几个缺点：①只利用了最后一个单词对应的隐含状态向量来对源语言句子的语义进行建模，该向量难以刻画源语言中所包含的全部信息；②解码过程中前面产生的词信息很难被循环神经网络保存下来以影响后面（特别是距离较远）单词的预测；③仅仅使用一个源语言句子的上下文向量难以预测目标语言句子中的所有词。

2.2.2. 基于注意力网络的编码器 - 解码器框架

为了解决上述三个问题，注意力网络被引入进来[12]。该方法能够利用编码器产生的所有隐含状态和前一个解码器隐含状态来生成一个更适合产生当前目标语言词的上下文向量，并用于生成当前位置的单词。如图6所示，该上下文向量是所有编码器隐含状态的加权平均，其权重是通过计算前一个解码器隐含状态和所有编码器隐含状态的相似程度并归一化为概率得到的。前一个解码器的隐含状态、前一个预测的词和上下文向量一起被作为解码器的输入来产生当前的解码器隐含状态并预测当前的目标语言词。这种方法可以在生成每个单词时，综合考虑编码器中全部的隐含状态向量。

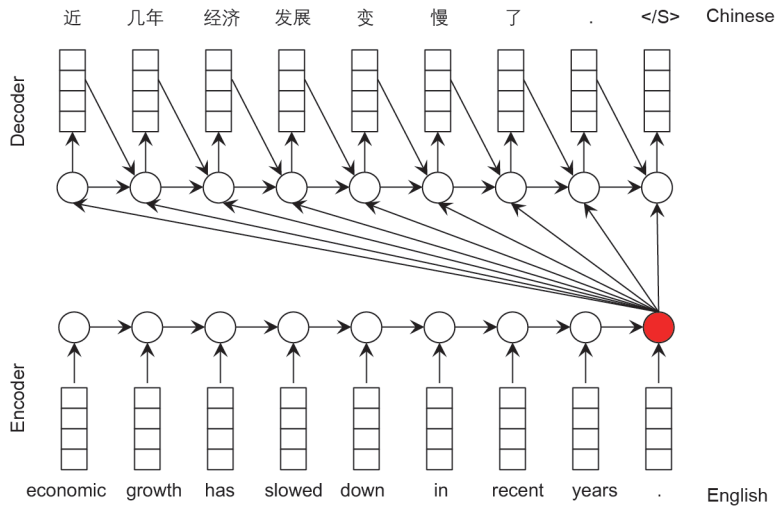

图5. 从英文到中文的MT编码器-解码器框架。

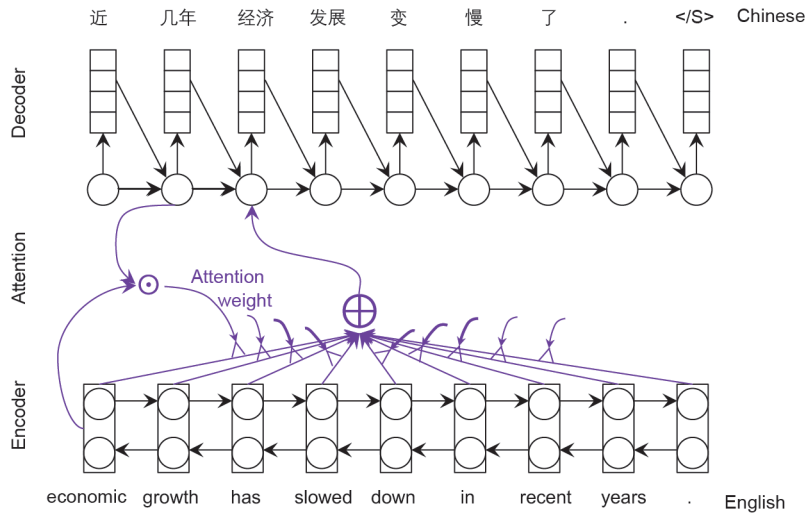

图6. 从英文到中文的MT基于注意力的编解码器框架。

2.2.3. 基于 Transformer 的编码器 - 解码器框架

为了更好地利用注意力网络强大的特征提取能力，Transformer [13]（图7）使用多头注意力网络来取代编码器和解码器中的循环神经网络，以及两者之间的注意力网络。多头注意力网络是一组注意力网络的集合。通过将查询、关键字以及值通过线性层映射成N 个向量，然后利用N个注意力网络来生成N 个上下文向量表示，进而将其组合成一个上下文向量。需要注意的是，对于解码器中的注意力网络，在解码过程中只能够利用已经生成的词的信息，所以只有以前的隐含状态才能够被用来生成当前的上下文向量。

2.3. 小结

本节介绍了如何使用神经网络来学习词嵌入、句子嵌入，以及序列到序列的转换。为了能够更好地对各类自然语言处理任务进行建模，我们还需在如下几个方面做更多的探索。

（1）先验知识的建模。尽管基于大规模数据训练得到的词嵌入表示能够包含一定的常识知识[6]，如何将已有的先验知识融入到神经网络中，仍然需要受到更多的重视[14]。

（2）文档建模和多轮对话建模。利用句子的上下文信息获得的词嵌入被证实能够显著提高自然语言处理各任务上的性能，但如何对更长距离的信息进行上下文建模仍然是一个开放课题[15]。例如，给定句子“the mouse is on the table”，仅仅依赖句子里的信息，我们很难确定“mouse”指的是“老鼠”还是“鼠标”。 为了更好地消歧，文档上下文信息需要被引入进来。类似地，如何在多轮对话中考

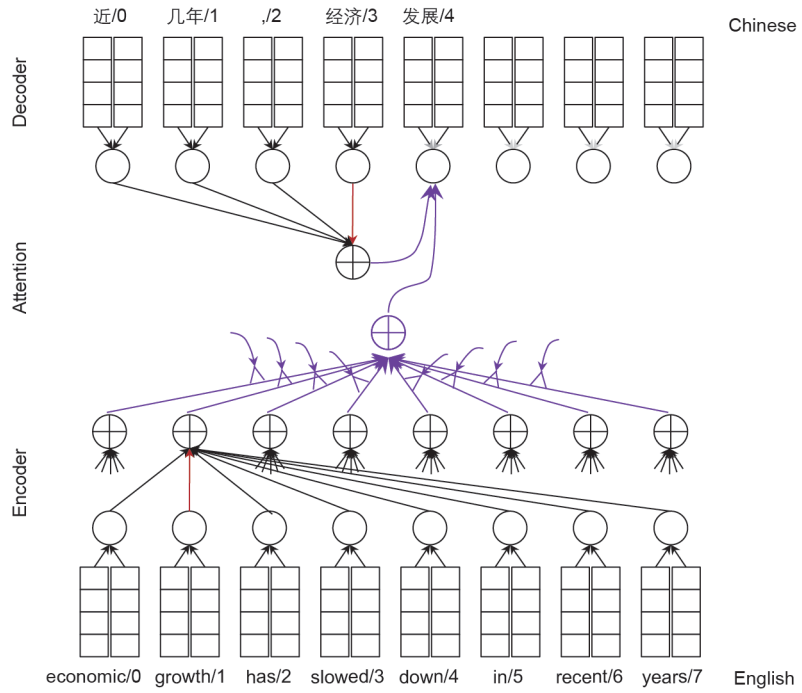

图7. 从英文到中文的MT基于全注意的编码器-解码器框架。

虑上下文信息[16]仍然是一个有挑战的任务。

（3）非自回归的产生模型。现在的序列到序列的转换模型在产生输出句子时采用自回归的方法依次产生句子里的词，也就是说前一时刻生成的词会被作为输入来产生当前时刻的词。这种自回归的产生方式会导致曝光偏差问题：前边产生的错误会被保存并放大从而影响后边词的产生。为了解决这个问题，人们提出了非自回归的产生模型[17]，然而该模型的性能仍然远远落后于自回归的模型。未来的研究中，如何设计更好的非自回归网络应该受到更多关注。

3. 神经网络模型的训练

随着深度学习的发展，新的训练方法被提出来并用于优化神经网络中的大规模参数。随机梯度下降（stochastic gradient descent, SGD）[18]是最常用的一种基于反向传播的神经网络训练方法[19]。基于冲量的随机梯度下降通过引入冲量来加速训练过程。AdaGrad [20]、AdaDelta[21]、Adam[22]和RMSProp通过对不同参数使用不同的学习率进一步提高训练效率，并使得训练过程更为稳定。当神经网络太过复杂或者参数非常多时，仅仅使用一个设备可能远远不够。为了解决这一问题，还需要基于多个训练设备（显卡或者计算机）的并行训练方法。根据参数更新的方式不同，分布式的随机梯度下降通常分为同步和异步的随机梯度下降。

除了这些通用训练方法的进展，针对不同的自然语言处理的任务，更好的训练和学习方法层出不穷。对于资源丰富的自然语言处理任务，监督学习的方法通常被用来利用大规模标注数据训练模型参数。对于这样的任务，深度学习模型往往能够得到很好的效果，例如，在中英语言对的新闻领域，存在大规模的双语句对，神经机器翻译（neural machine translation, NMT）的质量就会非常好，甚至达到了能与人类媲美的水平。然而对于其他自然语言处理任务，却很难获取大规模的训练数据，比如小语种的机器翻译和情感分析。为了提高该类任务上模型的性能，人们使用半监督学习方法将未标注数据和有标注数据进行结合，或者使用无监督学习方法，仅仅使用未标注数据来训练模型参数。另一种利用未标注数据的方法是利用未标注数据来预训练模型参数，然后将预训练模型通过迁移学习来迁移到目标任务上。当本任务的训练数据不够时，其他相关任务的训练数据也可以通过多任务学习来改进模型的性能。人工标注数据自然是最为有用但代价最高的一个途径，如何在经费有限的情况下更为高效地进行数据标注便是主动学习的研究范围。

3.1. 监督学习

基于有标注的数据，监督学习方法通过输入和输出对来训练模型参数。对于分类任务，监督学习通常通过最大

化正确标签的对数似然概率或者最小化交叉熵的方法来进行模型参数的训练。对于序列到序列转换的任务，给定输入序列作为条件，监督学习通常最大化输出序列的概率似然来学习模型参数。比如，给定一个训练语料中的句对(x, y)，其中$x = (x_1, x_2, ..., x_{|x|})$是输入句子，$y = (y_1, y_2, ..., y_{|y|})$是输出句子。条件似然概率定义为：

$$p_\theta(y|x) = \prod_{i=1}^{|y|} p_\theta(y_i|y_{i-1}, \ldots, y_1, x) \quad (1)$$

式中，$p_\theta(y_i|y_{i-1}, \ldots, y_1, x)$是序列到序列转换模型中解码器softmax层的输出概率。基于该似然函数，训练数据的对数似然损失函数定义为：

$$\text{LOSS} = \sum_{(x,y)\in \text{TrainingData}} -\log p_\theta(y|x) \quad (2)$$

基于这样的损失函数，具体的训练方法（比如Adam、AdaDelta）便可以用来训练模型参数。

除了最大化正确输出的产生概率，为了能够在训练过程中引入任务相关的损失函数，Shen等[23]提出了最小化风险（即最大化BLEU [24]，BLEU是机器翻译任务的常用评价指标，其本质是根据参考译文来统计翻译候选的n元文法的准确率）的训练方法来训练神经机器翻译的模型参数。该方法首先基于交叉熵的损失函数利用双语数据来预训练模型参数，然后最大化输出候选的BLEU期望来细调模型参数。为了解决自回归的序列到序列转换模型的曝光偏差问题，Zhang等[25]通过在训练目标中引入了两个Kullback-Leibleer (KL) 距离来最大化从左到右和从右到左模型生成的翻译候选的一致性。推敲网络[26]是另一种解决该问题的方法。推敲网络基于两轮的解码来模拟人工翻译的过程。具体的第一轮产生初始的翻译结果，而第二轮则对该初始结果进行修正。Zhang等[27]提出了通过同时从正确翻译和模型产生的候选中采样上下文词来解决神经机器翻译中的训练和解码不一致的问题。

3.2. 半监督和无监督学习

半监督和无监督学习通过利用未标注数据来提高模型性能。当使用半监督学习时，通常基于有标注的数据来预训练模型的参数，然后利用未标注数据对模型进一步进行训练。如何结合有标注数据和未标注数据，有多种半监督的学习方法，包括自学习、产生式模型和基于图的方法[28]。在这些方法中，模型产生的伪标注结果通常被用于模型的进一步训练。在神经机器翻译中，为了控制半监督学习生成的伪标注数据中的噪声和错误，人们引入了不同的权重或者利用某种奖励来过滤比较差的伪标注数据，比如用于NMT评价中的BLEU法[29]和对偶学习方法[30]。为了能够利用单语数据提高神经机器翻译的质量，反向翻译方法[31]使用一个反向的（目标语言到源语言的）翻译模型来将目标语言的单语数据翻译为伪双语数据，并将其用于对源语言到目标语言的翻译模型做进一步的训练。联合训练的方法（图8）[32]扩展了反向翻译方法，使其能够同时使用源语言和目标语言的单语数据，并基于期望最大化的框架来迭代地训练源语言到目标语言以及目标语言到源语言的翻译模型。在该方法中，首先使用双语数据来预训练两个翻译模型：源语言到目标语言（source-to-target, S2T）和目标语言到源语言（target-to-source, T2S）。S2T模型将源语言的单语数据进行翻译，构造目标语言端不太准确的伪双语数据，然后利用该伪双语数据来训练T2S模型。基于优化后的T2S模型，可以将目标语言单语数据进行翻译，得到源语言端不太准确的伪双语数据，该伪双语数据可以进一步优化S2T模型。这种训练过程可以迭代进行，直至在开发数据上的性能不能提高为止。

随着深度学习的发展，深度产生模型被用来进行无监督学习，比如变分自编码器（variational auto-encoder, VAE）[33]和生成对抗网络（generative adversarial network, GAN）[34]。VAE 网络沿用了自动编码网络的框架，包含一个编码器和一个解码器。编码器将输入映射到一个语义空间，解码器基于该语义空间的表示将重构输入本身。不同于原始的自动编码器，VAE假设编码器生成语义向量的分布应该尽可能地接近一个标准正态分布。同VAE类似，GAN同样包含两部分：一个产生器和一个判别器。产生器基于一个语义向量产生一个样本，判别器试图区分某个样本是产生器产生的，还是从数据集中采样的。通过一个对抗的损失函数，产生器尽可能地产生判别器区分不出来的样本，而判别器则尽可能地去区分产生的样本和数据集中的样本。研究人员将VAE和GAN应用到不同的自然语言处理的任务上[35,36]。

没有使用任何的双语数据，而是仅仅使用一个小的双语词典，无监督学习方法被用来训练无监督的神经机器翻译模型[37]。无监督的神经机器翻译通常使用联合训练的方法来通过产生伪双语数据，来同时训练源语言到目标语言和目标语言到源语言的翻译模型。由于没有真实双语数据，生成的伪双语数据难免存在错误和噪声，而这种

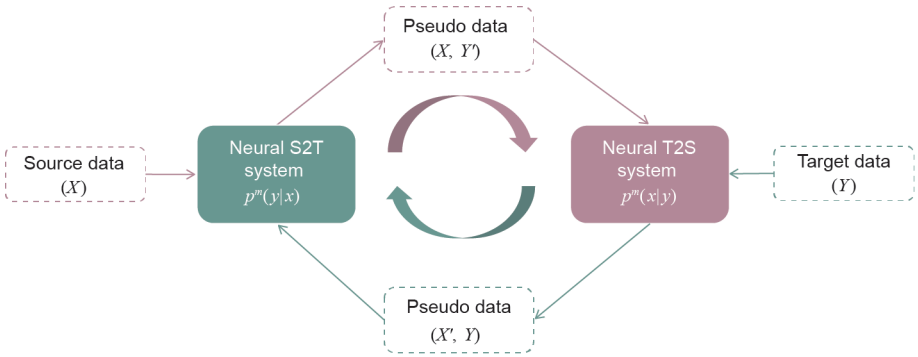

图8. S2T和T2S NMT模型的联合训练。

错误会在联合训练的迭代过程中被强化。为了解决这一问题，Ren等[38]引入了统计机器翻译（statistical machine translation，SMT）模型作为后验正则来过滤这些噪声和错误。SMT模型和NMT模型基于一个期望最大化框架来迭代优化这两种模型。如图9 [38]所示，整个训练过程包括两部分：模型的初始化和使用SMT作为后验正则的NMT训练。给定一个语言对*X*-*Y*，基于跨语言的词嵌入表示，可以获得初始的翻译概率表，并结合语言模型就可以构造两个初始SMT系统。初始SMT系统可以用来产生伪双语数据。该伪双语数据用来初始化两个NMT系统。NMT模型在训练时不仅仅使用SMT生成的伪双语数据，而且使用了基于联合训练方法的NMT模型生成的伪数据。而NMT系统生成伪双语数据则可以用来训练新的SMT系统。通过后验正则，SMT系统可以过滤掉NMT模型产生的伪数据中的噪声和错误，从而生成更高质量的伪双语数据。基于该数据训练得到的NMT系统能够为SMT系统提供更高质量的伪双语数据。SMT系统和NMT系统通过这种方式来帮助对方提高翻译质量，直到开发数据集上的性能不再提高为止。同Lample等[37]的方法相比，该方法能够显著提高翻译质量（法英方向可提高1.4个BLEU点，英法方向可提高3.5个点，德英方向可提高3.1个点，英德方向可提高2.2个点）。Lample 和Connau [39]进一步引入了生成式的预训练方法，并提出了一个跨语言的预训练模型，从而获得了目前最好的无监督机器翻译结果。

3.3. 多任务学习

多任务学习利用与目标任务相关的数据来改进模型在目标任务上的性能。当目标任务的训练数据不足时，与目标任务相关的其他任务的训练数据就可以被多任务学习引入进来改进目标任务模型的性能。Collobert等[10]提出了一个统一的神经网络框架，用来训练多个自然语言处理任务，包括POS、组块分析、命名实体识别（named entity recognition, NER）和语义角色标注（semantic role labelling, SRL）。该方法基于不同任务的数据，来学习任务共享的模型参数，该方法是神经网络在自然语言处理领域中一个里程碑式的工作，开启了深度学习在自然语言处理领域的潮流。

McCann等[40]提出了一种将10种不同的自然语言处理任务（包括问答、机器翻译、文摘、自然语言推断）都看作是问答任务，并构建了一个多任务问答网络（multi-task question-answering network, MQAN），如图10所示。MQAN的输入是一个问题和该问题对应的一个上下文。这种设置对问答任务来说是很自然的。对于机器翻译任务，问题对应“该输入在语言Y中对应的翻译是什么？”，上下文对应语言X输入本身。对于文摘任务，问题对应“该输入对应的摘要是什么？”，上下文对应文档本身。问题和上下文分别通过双向LSTM（BiLSTM）进行编码，然后通过co-attention来对两个输入构建条件进行编码表示。这两个条件编码表示分别通过两个双向LSTM（BiLSTM）、两个自注意力网络和两个双向的LSTM来获得最终的问题和上下文的编码结果。为了获得最后的输出，引入注意力机制考虑最相关的编码器的隐含状态，然后通过一个多点指针生成器网络来决定是从问题和上下文中复制一个词还是直接生成一个词。该模型在WikiSQL数据集上得到了目前最好的结果（72.4%的EM值和80.4%的准确率）。通过多任务学习，MQAN能够具有更好的泛化能力，从而在零样本关系抽取任务的QA-ZRE数据集上比单任务模型提高了11个F_1值。Liu等[41]提出了MT-DNN——一个基于BERT的多任务深度神经网络、通过添加特殊的多任务到预训练模型上，MT-DNN在10个自然语言理解的任务（包括SNLI和SciTail，以及9个GLUE任务[42]中的8个）上获

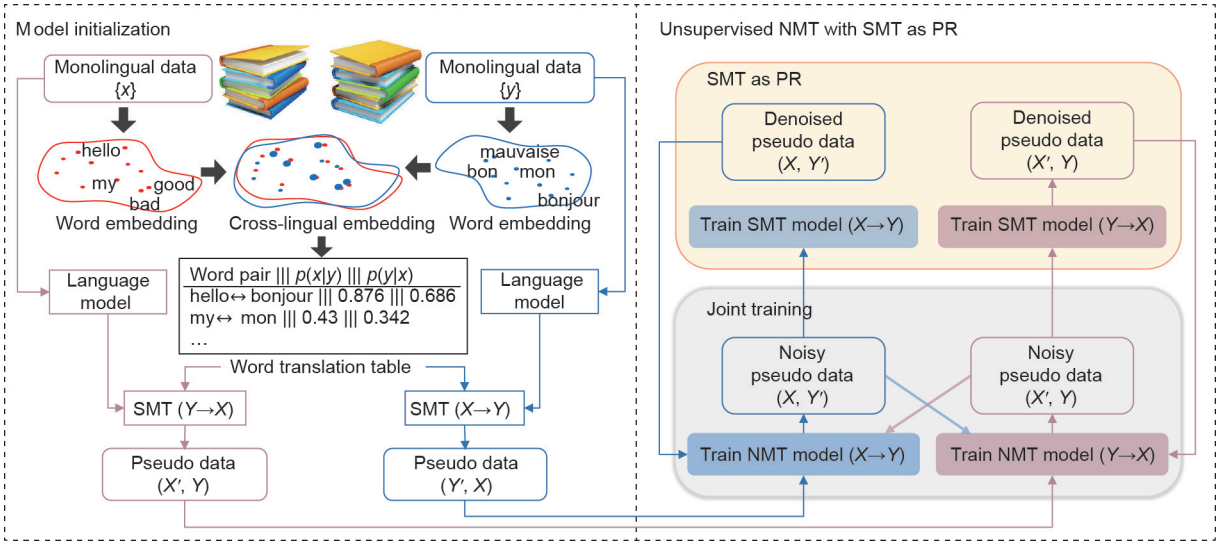

图9. 无监督NMT训练的示意图。

得了非常好的效果。MT-DNN表明，不同任务可以通过多任务学习相互促进。

3.4. 预训练和迁移学习

任务无关的模型可以首先通过预训练，然后通过迁移学习的微调过程迁移到特定的任务上。基于预训练的词嵌入或者句子嵌入，迁移学习可以被用来在某些特定的任务上进行微调[43]。近几年，很多预训练的模型被提出，比如在2.1节中介绍的word2vec [6]、Glove [8]、ELMo [9]、GPT [3]、BERT [4]和XLNet [5]。这些模型也广泛用于各种自然语言处理任务中，比如自动问答和文本分类等。Pushp等[44]提出了一种零样本迁移学习方法，用于文本分类。在该方法中，预训练模型首先在大规模数据集上学习句子和类别之间的关系，然后通过迁移学习迁移到没有任何训练数据新的类别上。Srivastava等[45]使用语义分析方法将分类概念上的自然语言解释映射到未标注数据上的约束关系特征，这种约束的特征会通过后验正则的方法进行融合，来提高零样本任务上的性能。对于某些富资源语言（比如英语）在某些任务上有大量的训练数据，但是在某些小语种（比如罗马尼亚语）上，训练资源则非常有限，跨语言的迁移学习可以将在富资源语言上训练得到的模型迁移到小语种的任务上。在多个语言对上，通过多任务学习得到的模型也可以通过迁移学习来支持零样本（没有任何双语数据）的翻译任务上[46]。

将一个预训练好的模型迁移到新的或者没见过的场景的另一个方向是元学习，由Schmidhuber [47]首次提出。元学习最近成为一个非常火的方向，应用在预训练模型的迁移、超参数的训练以及神经网络的优化上。通过仅仅若干样本便可以直接迁移一个初始的模型，与模型无关的元学习（model agnostic meta learning, MAML）[48]基于对模型不做任何假设的原则，提出了一种元学习的方法。该方法可以通过若干任务相关样本或者几个更新的步骤便可以迁移到相关的任务上。Gu等[49]利用MAML在欧洲18个语言上来优化NMT模型的训练。该方法将每个训练样本作为独特的伪任务，原始的结构查询产生任务，转化为小样本学习任务，从而可以利用元学习方法来显著提高模型性能。

Subramanian等[50]通过融合多个训练目标（包括跨语言NMT、自然语言推断、句法分析和skip-thought向量，即预测前一个和后一个句子）提出了一种高效的多任务学习框架，用来训练一个可迁移的句子嵌入表示。该句子表示可以被迁移到新的分类任务上。通过在不同的任务之间切换，基于门控循环单元（gated recurrent unit, GRU）的循环神经网络可以学习到句子不同角度的表示。通过NMT和句法分析任务，句法的特征也可以被更好地编码。句子级的特性包括句子长度和词序也可以通过句法分析任务更好地学习到。基于三个不同的数据集（SUBJ、TREC和DBpedia），学习得到的句子表示可以通过降维可视化，发现针对不同类别相似的句子能够更好地被聚在一起。不需要额外更新，学习得到的句子表达可以被很好地迁移到其他的任务上，比如在情感分析的任务（MR、CR、SUBJ&MPQA）上提高了1.1%~2.0%，在问题类别分类（TREC）上提高6%，在转述判别任务（Microsoft research paraphrase corpus）上提高2.3%。

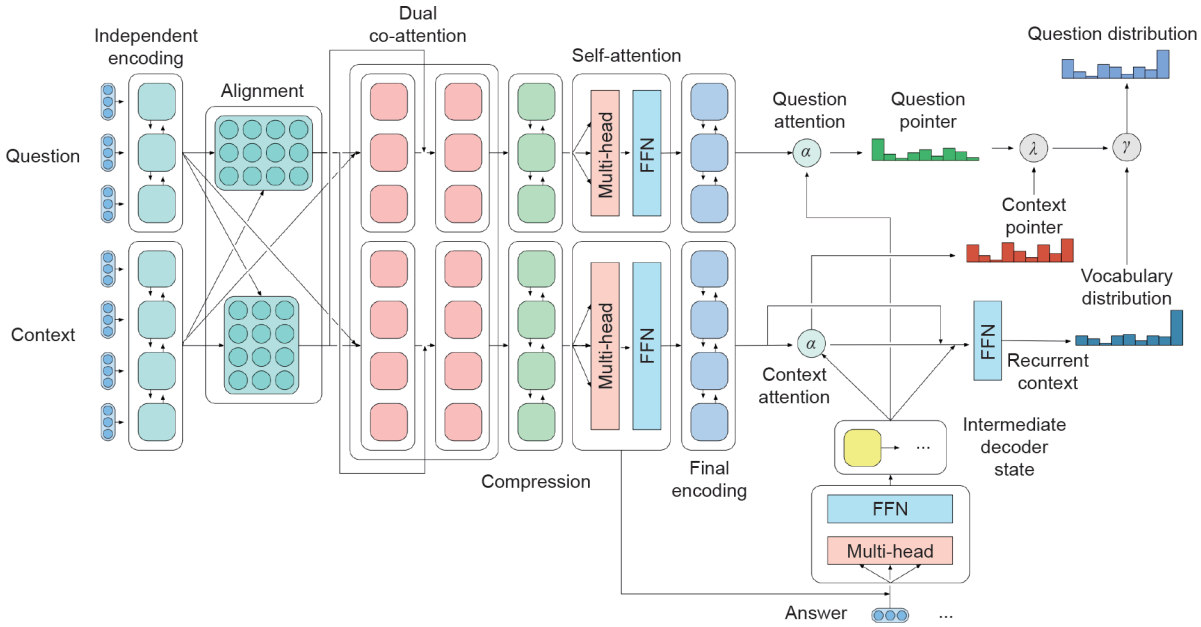

图10. MQAN的网络结构[40]。α是注意力权重；γ和λ是切换输出分布的标量。

3.5. 主动学习

主动学习解决如何选择合适的未标注数据让标注员去人工标注，以最小化标注的花费最大化性能提升。为了解决低资源学习问题中训练样本少的问题，一个直接的方法是由标注员去标注更多的数据。该思路带来了一个问题：选择哪些未标注的数据才能得到最大的性能提升。为了解决这个问题，主动学习[51]被用来基于当前的模型状况自动迭代地选择未标注数据，去标注以最大化模型性能的提升。基于有标注数据，可以训练一个初始的模型，并使用该模型来预测未标注数据的标签类型。基于标注的结果，主动学习能够选择某些（比如不太确定的）样本交付标注员去标注。标注员返回的标注数据会用来更新模型，新的模型会重新标注未标注数据，并进一步选择样本由标注员进行标注。该过程迭代进行，直至预先设置的花费用完或者达到预设的性能。

3.6. 小结

在本节中，我们介绍了几种经典的训练方法，包括监督学习、半监督学习、无监督学习、多任务学习、迁移学习和主动学习。当自然语言处理任务的标注数据足够多时，监督学习方法可以得到非常好的效果，比如机器翻译和机器阅读理解。对于标注数据不足的任务，可以引入多种机器学习的策略来提高模型的性能，比如使用未标注数据的半监督和无监督学习，基于预训练模型的迁移学习，使用相关的任务标注数据的多任务学习，或者选择最优价值未标注样本进行标注的主动学习。为了继续推进自然语言处理的发展，我们认为需要继续关注如下的几个方向。

（1）拓扑的训练过程。尽管多任务学习和迁移学习已经可以利用相关的学习任务来提高特定任务上模型的性能，然而未来我们需要深入探索不同的自然语言处理任务之间的关系，并基于此设计更合理的拓扑结构的训练过程。基于此拓扑训练过程，预训练的基础自然语言处理模型（比如语言模型、词性标注和句法分析）能够更好地迁移到高层次的自然语言处理任务（比如机器翻译和问答系统）。

（2）强化学习（reinforcement learning, RL）。当任务的损失函数不太容易定义时，强化学习被引入自然语言处理的多个任务上。比如，对于面向任务的对话，在某一轮的对话上，错误或者损失函数就很难定义。整体的损失或者错误则在对话完成后才能获得，比如任务是否完成或者对话持续了多少轮。对于这种只有一个长期的奖励可用的情况下，强化学习就可以用来学习一个策略网络，来最大化最终奖励的期望。强化学习的过程仍然存在着一些问题，比如非常大的句子长度的指数搜索空间[52]。

（3）生成对抗网络。尽管很多的研究者试图将生成对抗网络应用到自然语言处理的任务上，比如机器翻译[53]和自然语言生成[54]，然而依然存在着一些挑战，比如由于生成器生成的句子是在离散空间，错误的信号就很难像图像和语音处理那样从判别器直接通过梯度传递给生成器。另外生成对抗网络的训练对随机初始化的结果和超参

数的选择都非常敏感[36]。这些问题的存在制约了生成对抗网络在自然语言处理任务上的应用。

4. 推理

神经网络方法已经在包括机器翻译和机器阅读理解（machine reading comprehension, MRC）等在内的诸多NLP任务上取得了令人瞩目的效果。然而，此类方法依然存在诸多问题。例如，绝大多数神经网络方法都以黑盒的方式工作，因此无法对输出结果给出合理的解释和说明。此外，对于很多自然语言处理任务（如问答和对话系统），模型为了生成正确的输出，不仅需要对输入进行理解，还需要具备一定的外部知识。这就需要模型具备一定的推理能力。

本文将推理定义如下：基于已有知识对未见问题进行理解和推断，并得出问题对应答案的过程，即为推理。据此定义，推理系统（图11）主要由如下两个模块组成。

（1）知识：如知识图谱、常识、规则、从本文中抽取的断言等。

（2）推理引擎：基于输入和已有知识推断出答案。

这里，我们用两个例子来说明推理的重要性。

第一个例子是基于知识图谱的多轮问答。给定问题“比尔·盖茨夫人的生日是哪天？”知识图谱问答系统能够基于给定知识图谱将其转化为对应的语义表示：$\lambda x \lambda y. \text{DateOfBirth}(y, x) \wedge \text{Spouse}(\text{Bill Gates}, y)$。对于接下来的第二个问题“他/她是做什么工作的？”，问答模型首先需要具备跨越当前轮问题去做指代消解的能力，即判断第二个问题中的“他/她”究竟指的是第一个问题中的哪个实体。这本身就是一个推理过程。此外，问答模型还需要具备一定的常识知识，用来判断代词“他”或“她”所代表的实体的性别。

第二个例子是对话系统。当用户说“我现在很饿”，对话模型更应该回复“让我给你推荐一些餐馆吧”，而不是“让我给你推荐几部电影吧”。这么简单的对话过程同样需要推理，因为人感觉到饥饿后应该去吃饭，而不是去看电影。

本节余下部分将首先介绍两类知识：知识图谱和常识知识，然后介绍几种典型的推理引擎算法。

4.1. 知识

知识在推理任务中起着至关重要的作用。包括词典、规则、知识图谱、常识等在内的很多信息都属于知识的范畴。本文会重点介绍两类推理系统常用到的知识：知识图谱和常识知识。

4.1.1. 知识图谱

知识图谱可以表示为一个有向图$\{V, E\}$。它由节点V和边E组成。每个节点$v \in V$表示一个实体，每条边$e \in E$连接两个实体，用来表示一个谓词。每个三元组$<v_1, e, v_2>$用来表示知识图谱中的一个事实。

例如，<Microsoft, Founder, Bill Gates>表示一个知识图谱三元组，Microsoft表示主语实体，Bill Gates表示宾语实体，Founder表示一个谓词，它指明Bill Gates 是Microsoft的创始人。

知识图谱构建（knowledge graph construction, KBC）主要有三类方法。

（1）手工方法。此类方法完全采用手工方式构建知识图谱，如WordNet [55]。这类知识图谱通常具有非常高的质量，但覆盖度非常有限。此外，这类方法的构建开销也非常高。

（2）众包方法。此类方法通过众包的方式构建知识图谱。和手工构建方法相比，这类方法构建的知识图谱（如Satori、DBpedia [56]、Freebase [57]和WikiData [58]）通

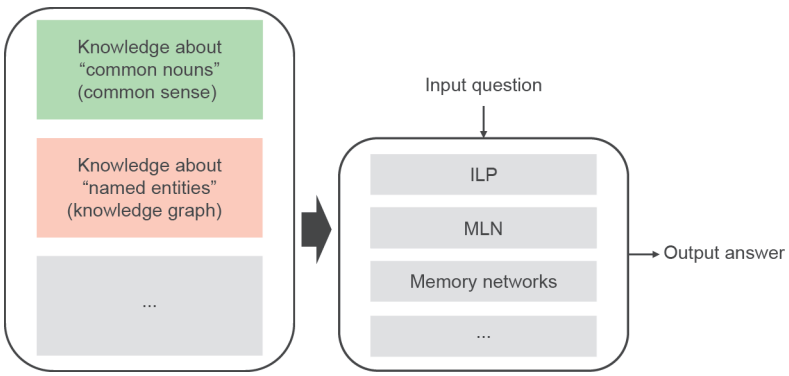

图11. 推理系统概述。ILP：整数线性规划；MLN：马尔可夫逻辑网络。

常具有很大的规模和较高的质量。

（3）信息抽取方法。此类方法通过信息抽取的方式从文本中抽取具有一定结构的知识，如KnowItAll [59]、YAGO [60]和NELL [61]。与前两类方法相比，此类方法抽取的知识通常具有较大的噪声，但抽取知识的数量可以非常大。

4.1.2. 常识

常识是指每个人都具备的关于事物的属性、关系以及事物之间交互的通用知识。常识知识通常与地点、语言和文化无关，并且很少在文本中被显式描述。

例如，“每一位父亲都是男性”是一条属性类常识，“太阳在天上”是一条空间类常识，“树木是从种子生长出来的”是一条过程类常识。

常识知识库（commonsense knowledge base, CKB）的构建非常具有挑战性。下面列举3种常见的构建方法。

（1）手工方法。CYC [62]是一个完全由人工专家构建的常识知识库。它主要专注那些很少被写下来或说出来的常识知识。例如，“每个人都有唯一一个父亲和唯一一个母亲”。构建CYC的目的是为了使AI系统具备人类所具备的常识知识。通过这类方式构建的常识知识库通常规模很小。这是由于人工标注的成本非常高。

（2）众包方法。ConceptNet [63]属于这类常识知识库。它由WordNet、Wiktionary、Wikipedia、DBpedia和Freebase混合构成。和CYC相比，ConceptNet规模更大，但其中很大一部分知识都是关于命名实体的，如“比尔·盖茨”“白宫”等，真正意义上的常识知识依然占少数。

（3）信息抽取方法。WebChild [64]属于这类常识知识库。它包含了从无结构文本中抽取出来的常识知识。这类知识库的规模较大，但由于信息抽取系统的错误，通常也包含更多的噪声。

4.2. 推理引擎

本小结首先介绍两种非神经网络推理算法：整数线性规划法（integer linear programming, ILP）和马尔可夫逻辑网络法（Markov logic network, MLN）。然后，介绍一种神经网络推理算法：记忆网络法，并介绍它在语义分析和对话回复生成两个任务中的应用。

4.2.1. 非神经网络推理算法——ILP 和 MLN

ILP是一个优化框架。给定一组有限的变量集合，ILP针对一组线性不等式约束条件去优化一个线性目标函数：

$$\begin{aligned} &\text{maxmize} && w^T x \\ &\text{subject to} && Ax \le b \\ &\text{and} && x \in \mathbb{Z}^n \end{aligned}$$

ILP中用到的约束条件可以看做是一种先验知识，优化过程可以看做是一个推理过程。下例给出ILP在信息抽取中的一个应用实例[65]。该任务的目标是从文本中抽取命名实体以及实体之间的关系。

命名实体识别（NER）模块和关系抽取（RE）模块通常分别训练，并采用串行的方式用于上述抽取任务。首先，NER模块从文本中识别实体，并为每个实体赋予一个抽取概率。然后，RE模块为每个实体对预测可能的关系。图13给出了上例对应的3个NER结果和2个RE结果。

如果我们从每个预测结果中选择局部最优的结果，那么就会产生错误。Brooklyn将会被识别为Person、Adam和Anne之间的关系将会被识别为PlaceOfBirth。然而，如果模型知道PlaceOfBirth对应的宾语实体类型应该是Location而不是Person，那么就可以避免这样的错误。基于局部预测结果和约束条件推断全局最优预测结果的过程就是一个典型的推理过程。

针对上述问题，我们将它形式化为一个ILP问题。首先，定义如下4条约束条件：①每个实体只能被赋予一个实体类型；②每个实体对只能赋予一个关系；③给定一个实体和连接该实体的一个关系，该关系对应位置上的实体类型必须和给定实体的类型保持一致；④每个关系对应位置的实体类型必须和实际连接的实体类型保持一致。基于

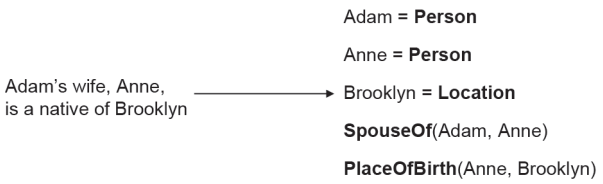

图12. 信息提取任务示例。

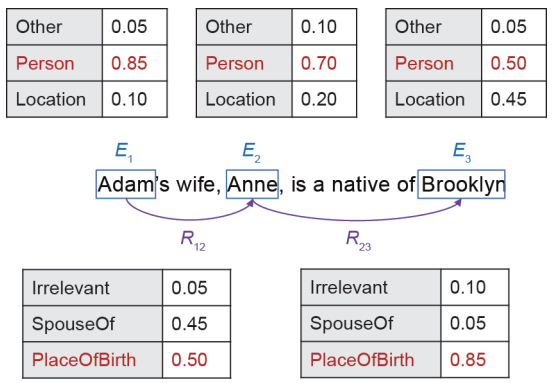

图13. 不使用ILP的次优结果。*E*：句子中出现实体；*R*：实体之间的关系。

上述4个约束条件以及NER和RE的局部预测结果，ILP能够推断出下述全局预测结果（图14）。

ILP在很多NLP任务中都有广泛的应用，如自动问答[66–68]和语义角色标注[69,70]。在很多缺乏足够训练语料的任务中，ILP能够充分地利用少量的先验知识，并取得很好的效果。

一个MLN [71] L由多个(F_i, w_i)对组成。每个F_i表示一个一阶逻辑公式，w_i表示F_i对应的权重。MLN将概率方法和一阶谓词逻辑融合在一个统一的框架之中。通过将一阶逻辑公式中的每个变量实例化成常量$C = \{c_1, ..., c_{|c|}\}$，MLN能够生成一个马尔可夫网络$M_{L,C}$，并基于此完成对应的推理任务。

理想情况下，MLN通过如下步骤完成相关的推理任务。这里使用著名的“friend-smoking”例子说明。

（1）给定一个自然语言描述的世界，首先将它解析成一组一阶逻辑公式$F = \{F_1, ..., F_{|F|}\}$。例如，给定下述两个句子：

smoking causes cancer

friends have similar smoking habits

它们对应的两个一阶逻辑公式分别是：

$\forall x$ Smokes(x) ⇒ Cancer（x）

$\forall x$，y Friends(x，y) ⇒（Smokes(x) ⇔ Smokes(y)）

在实际情况中，每个一阶逻辑公式都对应一个权重，该权重从关系知识库中通过特定算法学习得到，如pseudo-likelihood [72]。

（2）给定L和C，按照下述步骤得到一个马尔可夫网络$M_{L,C}$：①L中每个谓词的每种可能赋值在$M_{L,C}$中都对应了一个二元节点，如果谓词为真，对应节点的值为1，否则为0；②$M_{L,C}$中每个公式F_i对应的可能赋值都对应一个特征，如果公式为真，那么该特征的取值为1，否则为0。每个特征的权重w_i与F_i相关。例如，给定两个常量Anna (A)和Bob (B)，实例化后的$M_{L,C}$可以表示为图15的形式。

（3）给定一组事实，如Friends (A, B) = 1和Cancer (A) = 1，推理出当前世界每个状态最可能的赋值情况，如Smokes (A) = ? Smokes (B) = ? 和Cancer (B) = ? 等。这个推断的过程能够用来回答诸如“What is the probability that Anna is smoking given Anna has cancer？”的问题。常用的推断算法是MC-SAT [66]。

作为一种统计关系学习方法，MLN在很多NLP任务中都有广泛应用，如语义分析[73]、关系抽取[74]、实体消歧[75]等。然而，MLN在真实场景中的应用依然存在很多问题。一方面这是由于对自然语言的精准理解非常困难。另一方面，实例化后的马尔可夫网络规模通常非常大[76]，这也带来了很大的计算问题。此外，和ILP类似，目前MLN和深度学习结合的工作还不多。如何设计神经网络化的MLN依然有待探讨和解决。

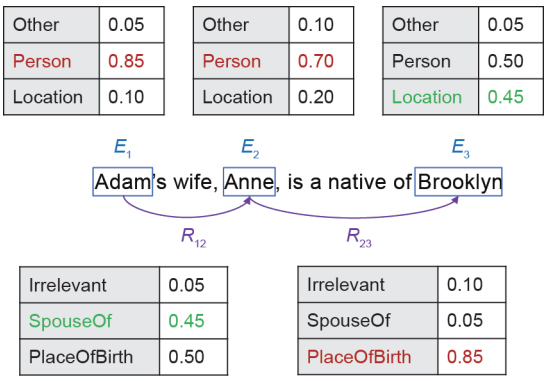

图14. 使用ILP时的最佳结果。

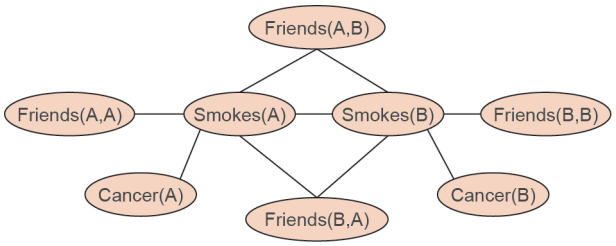

图15. 马尔可夫网络的一个例子。

4.2.2. 神经网络推理算法——记忆网络及其变种

记忆网络（memory network, MemNN）[77–79]是一种基于深度学习的推理框架。它通过多次访问存储在记忆模块中的知识达到推理的目的。

这里以“键-值记忆网络”（key-value memory network, KV-MemNN）[79]为例，说明如何将MemNN用于推理任务。图16 [79]给出KV-MemNN的一个示意图。

在KV-MemNN中，记忆模块被表示为一组键值向量对$(k_1, v_1), ..., (k_M, v_M)$，$k_m$表示第$k$个键，$v_m$表示$k_m$对应的值。知识图谱和非结构化文本都能够存储在这样的记忆模块中。例如，给定一个知识图谱三元组$< e_1, v, e_2 >$，键可以定义为主语实体e_1和谓词v，值可以定义为宾语实体e_2。给定一个输入问题x，首先从知识图谱中找到与其相关的一组键值对$(k_{h_1}, v_{h_1}), ..., (k_{h_N}, v_{h_N})$。查找的方式按照每个键值对与$x$的相似性：

$$p_{h_i} = \text{softmax}\left(A\Phi_X(x) \cdot A\Phi_K(k_{h_i})\right) \quad (3)$$

式中，$\Phi_X(x)$和$\Phi_K(k_{h_i})$用于将x和k_{h_i}映射到D维特征；A是一个$d\times D$维的矩阵。接下来，输出向量$\boldsymbol{o}$通过下述方式进行计算：

$$\boldsymbol{o}=\sum_i p_{h_i} A\Phi_V(v_{h_i}) \tag{4}$$

如果某个任务需要多轮推理，那么可以通过下式对输入问题x对应的特征向量$\boldsymbol{q}$进行更新：

$$\boldsymbol{q}_2=\boldsymbol{R}_1(\boldsymbol{q}+\boldsymbol{o}) \tag{5}$$

式中，$\boldsymbol{q}=\boldsymbol{A}\Phi_X(x)$；$R_1$是一个$d\times d$维的矩阵。上述过程执行$H$次后，生成最终的问题表示$q_{H+1}$，并用于输出预测结果。

MemNN可被用于推理相关的任务。例如，Weston等[77]将MemNN用于推理数据集bAbI。Miller等[79]将KV-MemNN用于WikiMovies和WikiQA [80]这两个数据集。Bordes等[81]将MemNN用于端到端对话系统。

MemNN可以看作是记忆增强神经网络（memory-augmented neural network, MANN）的一个特例。MANN在实际NLP任务中有很多应用。接下来以两个推理相关的任务（语义分析和回复生成）为例，说明如何使用知识图谱和常识知识完成推理任务。

Guo等[82]提出一种基于知识图谱的多轮语义分析方法。在该方法中，基于记忆网络的对话记忆模块被引入，用于处理多轮对话中常出现的指代消解和关系省略现象。

（1）给定第一个问题：Where was Donald Trump born?语义分析器首先生成其对应的语义表示：

λx. PlaceOfBirth（Donald Trump，x）

通过在知识图谱上执行该语义表示，能够得到其对应的答案New York。由于上述过程只与输入问题相关，它属于上下文无关的语义分析任务。

（2）接下来，给定第二个问题：Where did he graduate from?为了解析这个问题，第一个问题也需要考虑，这是因为第二个问题中的he实际上指第一个问题中的实体Donald Trump。由于解析第二个问题需要考虑上下文，它属于上下文相关的语义分析任务。

给定一个上下文无关的问题，语义分析器通过如下步骤生成其对应的逻辑表示。从根节点start开始，一个语法引导的解码器采用自顶向下的方式，不断地将action中的非终结符替换成其他action或终结符，直至没有任何非终结符为止。图17给出这个过程。

表1给出针对多轮语义分析任务定义的21种不同的action。每个action由三部分组成：语义类型、函数名和函数参数。每个参数既可以是一个语义类型，也可以是一个常量，还可以是一个action序列。按照这个定义，表中前15个action覆盖了语义分析中典型的操作。动作16到动作18分别表示将非终结符转化为一个实体e、一个谓词r和一

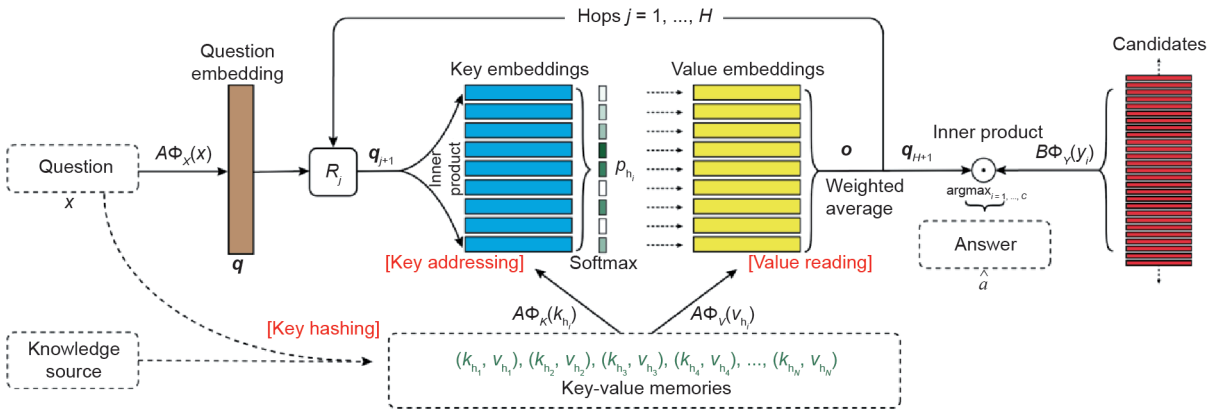

图16. KV-MemNN概述[79]。a表示问题的答案；B表示$d\times D$矩阵，可以将其约束为与A相同。R_j表示$d\times d$矩阵，用于更新第j跳中输入问题的表示形式。

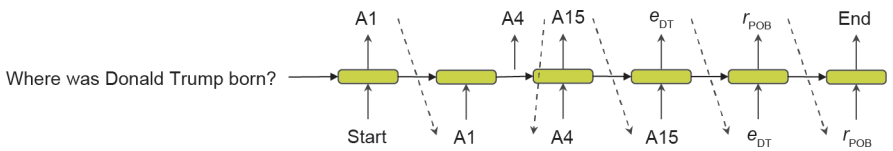

图17. 上下文无关的语义解析的示例。DT：唐纳德·特朗普（Donald Trump）。

个数字num。动作19到动作21用于复制已有的action序列。

给定一个上下文相关的问题，对话历史记忆模块保存了从对话开始到现在为止全部对话对应的语义表示，包括提及的实体名称、谓词、答案实体以及全部合法的动作子序列。这些信息将被用于当前轮问题的语义分析中，用于完成跨轮指代消解和关系省略等任务。如图18中，第二个问题中的he实际是指第一个问题中提到的实体Donald Trump。通过使用对话历史记忆模块，这个指代消解问题得以解决。

CSQA数据集[83]上的实验表明：通过引入对话记忆，该方法在多轮语义分析和问答任务上取得了目前最优的效果。

Zhou等[84]提出一种基于常识知识的编码器-解码器方法，用于完成开放领域对话系统中的回复生成任务。这里，ConceptNet被用于理解用户的输入，并跟据相关的常识知识生成对应的回复。

在编码阶段，给定输入$X = x_1, \ldots, x_n$中的每一个单词，从常识图谱中搜索得到相关的常识知识，并通过知识嵌入将其转化为对应的向量表示。每个单词都和对应的知识向量表示进行拼接，生成知识增强单词表示$e(x_i) = [w(x_i); g_i]$，并将其输入到GRU编码器中。

在解码阶段，解码器在生成每个单词时，都会从常识图谱中搜索相关的知识，并将其转化为对应的知识嵌入引入到解码过程当中。此外，在输出单词的时候，解码器还被允许从常识图谱中选择某个实体作为输出。

实验证明，通过在编码和解码两个阶段都引入常识知识，该方法在常识对话数据集[84]上能够取得当前最优的结果。

4.3. 推理相关数据集

最近，学术界发布了很多推理相关的数据集。按照所使用知识类型的不同，本文将这些数据集归为如下4类。

表1 由语义类别、功能符号和参数列表组成的动作列表

Action	Operation	Note
A1–A3	start → set|num|bool	
A4	set → find(set, *r*)	Set of entities with a *r* edge to *e*
A5	num → count(set)	Total number of set
A6	bool → in(*e*, set)	Whether *e* is in set
A7	set → union(set_1, set_2)	Union of set_1 and set_2
A8	set → inter(set_1, set_2)	Intersection of set_1 and set_2
A9	set → diff(set_1, set_2)	Instances included in set_1 but not included in set_2
A10	set → larger(set, *r*, num)	Subset of set linking to more than num entities with relation *r*
A11	set → less(set, *r*, num)	Subset of set linking to less than num entities with relation *r*
A12	set → equal(set, *r*, num)	Subset of set linking to num entities with relation *r*
A13	set → argmax(set, *r*)	Subset of set linking to most entities with relation *r*
A14	set → argmin(set, *r*)	Subset of set linking to least entities with relationr
A15	set → {*e*}	
A16–A18	e|*r*|num → constant	Instantiation for entity *e*, predicate *r* or number num
A19–A21	set|num|bool → $action_{i-1}$	Replicate previous action sequence (w/0 or w/ instantiation)

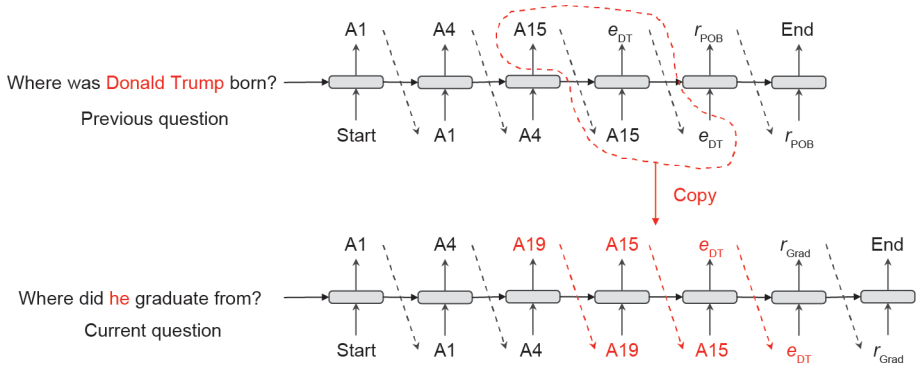

图18. 一个依赖于内容的语义解析示例。

（1）基于知识图谱的任务。WikiSQL [85]、LC-QuAD [86]、CSQA [83]和ComplexWebQuestions [87]都属于这类任务。语义分析是完成这类任务最核心的技术。

（2）基于常识知识的任务。Winograd Schema Challenge [88]、ARC [89]、CommonsenseQA [90]和ATOMIC [91]都属于这类任务。如何抽取、表示和利用任务相关的常识知识是这类任务最核心的技术。

（3）基于文本的任务。HOTPOTQA [92]、NarrativeQA [93]、MultiRC [94]和CoQA [95]都属于这类任务。目前，端到端的深度学习模型在这类文本任务上取得了非常好的效果。不过究竟这些模型是否真的学习到了推理的能力，还是说它们只是记忆了有用的模板，这些都是推理研究亟需解决的问题。

（4）基于文本和视觉的任务。GQA [96]和VCR [97]都属于这类任务。给定一张图片和一个自然语言问题，这类任务需要系统能够从给定的答案中选择出问题对应的正确答案。这不仅需要模型具备一定的推理能力，还要求模型能够很好地理解和对齐来自语言和视觉两方面的信息。

由于数据标注的昂贵性，目前绝大多数推理数据集的规模都不大。从这些数据集中训练得到的模型往往缺乏泛化性。最近，包括ELMo、GPT、BERT 和 XLNet等在内的预训练模型在很多NLP任务上取得了非常好的效果。如何将已有知识和预训练模型结合，成为一个非常有意义的研究课题。

4.4. 总结

本小结简要介绍了一些和推理相关的NLP研究。包括非神经网络方法和神经网络方法。推理相关的研究目前依然处于比较初步的阶段，也有很多挑战需要去解决，包括以下3个方面。

（1）知识抽取。由于目前知识图谱的覆盖度很低，因此无法涵盖绝大多数自然语言内容。这导致基于知识的推理方法很难在开放领域的NLP任务上取得广泛应用。因此，如何抽取大规模高质量的知识成为推动推理研究的一个重要课题。

（2）结合显式知识和预训练模型进行推理。传统的推理方法大都基于显式的知识。最近，包括GPT、BERT和XLNet在内的预训练模型在很多推理相关的NLP任务上都取得了很好的效果，如Winograd Schema Challenge [88]和SWAG [98]。如何将预训练模型和已有知识这两者的优势结合在一起，也是接下来一个值得研究的方向。

（3）数据集和评测指标。由于深度学习技术的发展，很多NLP任务都取得了非常好的效果。然而我们依然无法确认这些模型究竟具备多少推理能力。这就需要构建更好的和推理相关的数据集以及设计可以评价推理能力的评测指标。

总体而言，推理在很多NLP任务中都至关重要。推理研究的发展能够推动NLP整体水平的发展。如何更好地利用包括知识图谱、常识、预训练模型等不同类型的知识进行推理研究，是人工智能从感知走向认知的关键一步。

5. 结论

本文从模型、训练和推理3个角度回顾了自然语言处理在近年来所取得的最新进展。总体而言，NLP研究和应用已经进入了一个新的时代。对于高资源的NLP任务（如机器翻译和自动问答），监督学习方法已经取得了很好的效果。对于低资源的NLP任务，半监督学习和无监督学习在近年来也有了长足的进步和发展。但还需要更多持续的努力才能取得更令人满意的效果。此外，推理相关的研究和数据集也得到研究者越来越多的关注。这个方向目前尚处于起步期，但意义重大，因此还有很大的发展空间。

展望未来，我们可以看到很多令人激动的方向。预训练模型的发展已经展示出巨大的威力。毫无疑问，这些研究将继续推动自然语言理解和生成任务的水平。记忆和知识增强的神经网络方法会受到越来越多的关注，并促进知识抽取领域的持续发展。此外，如何将包括声音、视觉和文本等多模态信息融入NLP的研究，也是人工智能发展的一个重要方向。

未来，NLP技术会极大地改变人们的生活。为了实现这一目标，需要我们继续不断地创新，并推动各项研究和应用。这些技术终将更好地为人类社会服务。

Compliance with ethics guidelines

Ming Zhou, Nan Duan, Shujie Liu, and Heung-Yeung Shum declare that they have no conflict of interest or financial conflicts to disclose.

References

[1] Deng J, Dong W, Socher R, Li LJ, Li K, Li FF. ImageNet: a large-scale hierarchical image database. In: Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition; 2009 Jun 20–25; Miami, FL, USA; 2009. p. 248–55

[2] Xiong W, Wu L, Alleva F, Droppo J, Huang X, Stolcke A. The Microsoft 2017

conversational speech recognition system. In: Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing; 2018 Apr 15–20; Calgary, AB, Canada; 2018. p. 5934–8.
[3] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. https://s3-us-west-2.amazonaws.com/ openai-assets/research-covers/language-unsupervised/language_ understanding_paper.pdf, 2018.
[4] Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics; 2019 Jun 2–7; Minneapolis, MN, USA; 2019. p. 4171–86.
[5] Yang ZL, Dai Z, Yang YM, Carbonell J, Salakhutdinov R, Le QV. XLNet: generalized autoregressive pretraining for language understanding. 2019. arXiv:1906.08237.
[6] Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. 2013. arXiv:1301.3781.
[7] Firth JR. A synopsis of linguistic theory 1930–1955. In: Firth JR. Studies in linguistic analysis. Oxford: Blackwell; 1957. p. 1–31.
[8] Pennington J, Socher R, Manning C. GloVe: global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing; 2014 Oct 25–29; Doha, Qatar; 2014. p. 1532–43.
[9] Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, et al. Deep contextualized word representations. In: Proceedings of the 2018 Annual Conference of the North American Chapter of the Association for Computational Linguistics; 2018 Jun 1–6; New Orleans, LA, USA; 2018.
[10] Collobert R, Weston J. A unified architecture for natural language processing: deep neural networks with multitask learning. In: Proceedings of the 25th International Conference on Machine Learning; 2008 Jul 5–9; Helsinki, Finland; 2008. p. 160–7.
[11] Cho K, van Merrienboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, et al. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. 2014. arXiv:1406.1078.
[12] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. 2014. arXiv:1409.0473.
[13] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Proceedings of the 31rd Neural Information Processing Systems; 2017 Dec 4–9; Long Beach, CA, USA; 2017.
[14] Yu M, Dredze M. Improving lexical embeddings with semantic knowledge. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics; 2014 Jun 23–25; Baltimore, MD, USA; 2014. p. 545–50.
[15] Zhang J, Luan H, Sun M, Zhai F, Xu J, Zhang M, et al. Improving the transformer translation model with document-level context. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing; 2018 Oct 31– Nov 4; Brussels, Belgium; 2018. p. 533–42.
[16] Wu Y, Wu W, Xing C, Xu C, Li Z, Zhou M. A sequential matching framework for multi-turn response selection in retrieval-based chatbots. Comput Linguist 2019;45(1):163–97.
[17] Gu J, Bradbury J, Xiong C. Li VOK, Socher R. Non-autoregressive neural machine translation. 2017. arXiv:1711.02281.
[18] Friedman JH. Stochastic gradient boosting. Comput Stat Data Anal 2002;38(4):367–78.
[19] Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature 1986;323(9):533–6.
[20] Duchi J, Hazan E, Singer Y. Adaptive subgradient methods for online learning and stochastic optimization. J Mach Learn Res 2011;12(Jul):2121–59.
[21] Zeiler MD. ADADELTA: an adaptive learning rate method. 2012. arXiv:1212.5701.
[22] Kingma DP, Ba J. Adam: a method for stochastic optimization. In: Proceedings of the 2015 International Conference on Learning Representations; 2015 May 7–9; San Diego, CA, USA; 2015.
[23] Shen S, Cheng Y, He Z, He W, Wu H, Sun M, et al. 2016. Minimum risk training for neural machine translation. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics; 2016 Aug 7–12; Berlin, Germany; 2016.
[24] Papineni K, Roukos S, Ward T, Zhu WJ. BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics; 2002 Jul 7–12; Philadelphia, PA, USA; 2002. p. 311–8.
[25] Zhang Z, Wu S, Liu S, Li M, Zhou M, Xu T. Regularizing neural machine translation by target-bidirectional agreement. In: Proceedings of the 31rd AAAI Conference on Artificial Intelligence; 2019 Jan 27– Feb 1; Honolulu, HI, USA; 2019.
[26] Xia Y, Tian F, Wu L, Lin J, Qin T, Yu N, et al. Deliberation networks: sequence generation beyond one-pass decoding. In: Proceedings ot the 31rd Neural Information Processing Systems; 2017 Dec 4–9; Long Beach, CA, USA; 2017.
[27] Zhang W, Feng Y, Meng F, You D, Liu Q. Bridging the gap between training and inference for neural machine translation. 2019. arXiv:1906.02448.
[28] Zhu XJ. Semi-supervised learning literature survey. Madison: University of Wisconsin-Madison; 2005.
[29] Cheng Y, Xu W, He Z, He W, Wu H, Sun M, et al. Semi-supervised learning for neural machine translation. 2016. arXiv:1606.04596.
[30] He D, Xia Y, Qin T, Wang L, Yu N, Liu T, et al. Dual learning for machine translation. In: Proceedings of the 30th International Conference on Neural Information Processing Systems; 2016 Dec 5–10; Barcelona, Spain; 2016. p. 820–8.
[31] Sennrich R, Haddow B, Birch A. Improving neural machine translation models with monolingual data. 2015. arXiv:1511.06709.
[32] Zhang Z, Liu S, Li M, Zhou M, Chen E. Joint training for neural machine translation models with monolingual data. In: Proceedings of the 32nd AAAI Conference on Artificial Intelligence; 2018 Feb 2–7; New Orleans, LA, USA; 2018.
[33] Kingma DP, Welling M. Auto-encoding variational bayes. 2013. arXiv:1312.6114.
[34] Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial nets. In: Proceedings of the International Conference on Neural Information Processing Systems (NIPS 2014). pp. 2672–2680..
[35] Hu W, Tan Y. Generating adversarial malware examples for black-box attacks based on GAN. 2017. arXiv:1702.05983.
[36] Semeniuta S, Severyn A, Gelly S. On accurate evaluation of GANs for language generation. 2018. arXiv:1806.04936.
[37] Lample G, Conneau A, Denoyer L. Ranzato M. Unsupervised machine translation using monolingual corpora only. 2017. arXiv:1711.00043.
[38] Ren S, Zhang Z, Liu S, Zhou M, Ma S. Unsupervised neural machine translation with SMT as posterior regularization. 2019. arXiv:1901.04112.
[39] Conneau A, Lample G. Cross-lingual language model pretraining. 2019. arXiv:1901.07291.
[40] McCann B, Keskar NS, Xiong C, Socher R. The natural language decathlon: multitask learning as question answering. 2018. arXiv:1806.08730.
[41] Liu X, He P, Chen W, Gao J. Multi-task deep neural networks for natural language understanding. 2019. arXiv:1901.11504.
[42] Wang A, Singh A, Michael J, Hill F, Levy O, Bowman SR. GLUE: a multi-task benchmark and analysis platform for natural language understanding. In: Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP; 2018 Oct 31– Nov 4; Brussels, Belgium; 2018. p. 353–5.
[43] Zoph B, Le QV. Neural architecture search with reinforcement learning. 2016. arXiv:1611.01578.
[44] Pushp PK, Srivastava MM. Train once, test anywhere: zero-shot learning for text classification. 2017. arXiv:1712.05972.
[45] Srivastava S, Labutov I, Mitchell T. Zero-shot learning of classifiers from natural language quantification. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics; 2018 Jul 15–20; Melbourne, Australia; 2018. p. 306–16.
[46] Johnson M, Schuster M, Le QV, Krikun M, Wu Y, Chen Z, et al. Google's multilingual neural machine translation system: enabling zero-shot translation. Trans Assoc Comput Linguist 2017;5:339–51.
[47] Schmidhuber J. Evolutionary principles in self-referential learning: on learning how to learn [dissertation]. München: Technische Universität München; 1987.
[48] Finn C, Abbeel P, Levine S. Model-agnostic meta-learning for fast adaptation of deep networks. 2017. arXiv:1703.03400.
[49] Gu JT, Wang Y, Chen Y, Cho K, Li VOK. Meta-learning for low-resource neural machine translation. 2018. arXiv:1808.08437.
[50] Subramanian S, Trischler A, Bengio Y, Pal CJ. Learning general purpose distributed sentence representations via large scale multi-task learning. 2018. arXiv:1804.00079.
[51] Settles B. Active learning literature survey. Madison: University of Wisconsin-Madison; 2009.
[52] He J, Chen J, He X, Gao J, Li L, Deng L, et al. Deep reinforcement learning with a natural language action space. 2015. arXiv:1511.04636.
[53] Wu L, Xia Y, Zhao L, Tian F, Qin T, Lai J, et al. Adversarial neural machine translation. 2017. arXiv:1704.06933.
[54] Alzantot M, Sharma Y, Elgohary A, Ho BJ, Srivastava M, Chang KW. Generating natural language adversarial examples. 2018. arXiv:1804.07998.
[55] Miller GA. WordNet: a lexical database for English. Commun ACM 1995;38(11):39–41.
[56] Auer S, Bizer C, Kobilarov G, Lehmann J, Cyganiak R, Ives Z. DBpedia: a nucleus for a web of open data. In: Proceedings of the 2007 International Semantic Web Conference; 2007 Nov 11–15; Busan, Korea; 2007. p. 722–35.
[57] Bollacker KD, Evans C, Paritosh P, Sturge T, Taylor J. Freebase: a collaboratively created graph database for structuring human knowledge. In: Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data; 2008 Jun 9–12; Vancouver, BC, Canada; 2008. p. 1247–50.
[58] Vrandečić D, Krötzsch M. Wikidata: a free collaborative knowledgebase. Commun ACM 2014;57(10):78–85.
[59] Etzioni O, Cafarella M, Downey D, Kok S, Popescu AM, Shaked T, et al. Web-scale information extraction in knowitall: (preliminary results). In: Proceedings of the 13th International Conference on World Wide Web; 2004 May 17–20; New York, NY, USA; 2004. p. 100–10.
[60] Fabian MS, Gjergji K, Gerhard WE. YAGO: a core of semantic knowledge unifying WordNet and Wikipedia. In: Proceedings of the 16th International Conference on World Wide Web; 2007 May 8–12; Banff, AL, Canada; 2007. p. 697–706.

[61] Carlson A, Betteridge J, Kisiel B, Settles B, Hruschka ER Jr, Mitchell TM. Toward an architecture for never-ending language learning. In: Proceedings of the 24th AAAI Conference on Artificial Intelligence; 2010 Jul 11–15; Atlanta, GA, USA; 2010.
[62] Lenat DB. CYC: a large-scale investment in knowledge infrastructure. Commun ACM 1995;38(11):33–8.
[63] Liu H, Singh P. ConceptNet — a practical commonsense reasoning tool-kit. BT Technol J 2004;22(4):211–26.
[64] Tandon N, De Melo G, Weikum G. Acquiring comparative commonsense knowledge from the web. In: Proceedings of the 28th AAAI Conference on Artificial Intelligence; 2014 Jul 27–31; Quebec City, QC, Canada; 2014.
[65] Roth D, Yih W. A linear programming formulation for global inference in natural language tasks. In: Proceedings of the 8th Conference on Computational Natural Language Learning; 2004 May 6–7; Boston, MA, USA; 2004.
[66] Khashabi D, Khot T, Sabharwal A, Clark P, Etzioni O, Roth D. Question answering via integer programming over semi-structured knowledge. 2016. arXiv: 1604.06076.
[67] Khot T, Sabharwal A, Clark P. Answering complex questions using open information extraction. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics; 2017 Jul 30– Aug 4; Vancouver, BC, Canada; 2017. p. 311–6.
[68] Khashabi D, Khot T, Sabharwal A, Roth D. Question answering as global reasoning over semantic abstractions. In: Proceedings of the 32nd AAAI Conference on Artificial Intelligence; 2018 Feb 2–7; 2018.
[69] Punyakanok V, Roth D, Yih WT, Zimak D. Semantic role labeling via integer linear programming inference. In: Proceedings of the 20th International Conference on Computational Linguistics; 2004 Aug 23–27; Geneva, Switzerland; 2004.
[70] Srikumar V, Roth D. A joint model for extended semantic role labeling. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing; 2011 Jul 27–31; Edinburgh, UK; 2011. p. 129–39.
[71] Richardson M, Domingos P. Markov logic networks. Mach Learn 2006;62(1–2):107–36.
[72] Besag J. Statistical analysis of non-lattice data. Statistician 1975;24(3):179–95.
[73] Poon H, Domingos P. Unsupervised semantic parsing. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing ; 2009 Aug 6–7; Singapore; 2009.
[74] Pawar S, Bhattacharya P, Palshikar GK. End-to-end relation extraction using Markov logic networks. 2017. arXiv:1712.00988.
[75] Dai HJ, Tsai RTH, Hsu WL. Entity disambiguation using a Markov-logic network. In: Proceedings of the 5th Internatioanl Joint Conference on Natural Language Processing; 2011 Nov 8–13; Chiang Mai, Thailand; 2011. p. 846–55.
[76] Culotta A, McCallum A. Practical Markov logic containing first-order quantifiers with application to identity uncertainty. In: Proceedings of the Workshop on Computationally Hard Problems and Joint Inference in Speech and Language Processing; 2006 Jun 9; New York, NY, USA; 2006. p. 41–8.
[77] Weston J, Bordes A, Chopra S, Rush AM, van Merriënboer B, Joulin A, et al. Towards AI-complete question answering: a set of prerequisite toy tasks. 2015. arXiv:1502.05698.
[78] Sukhbaatar S, Szlam A, Weston J, Fergus R. End-to-end memory networks. In: Proceedings of the 2015 Neural Information Processing Systems Conference; 2015 Dec 7–12; Montreal, QC, Canada; 2015.
[79] Miller AH, Fisch A, Dodge J, Karimi AH, Bordes A, Weston J. Key-value memory networks for directly reading documents. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing; 2016 Nov 1–5; Austin, TX, USA; 2016. p. 1400–9.
[80] Yang Y, Yih WT, Meek C. WikiQA: a challenge dataset for open-domain question answering. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing; 2015 Sep 17–21; Lisbon, Portugal; 2015. p. 2013–8.
[81] Bordes A, Boureau YL, Weston J. Learning end-to-end goal-oriented dialog. In: Proceedings of the 2017 International Conference on Learning Representations; 2017 Apr 24–26; Toulon, France; 2017.
[82] Guo D, Tang D, Duan N, Zhou M, Yin J. Dialog-to-action: conversational question answering over a large-scale knowledge base. In: Proceedings of the 2018 Neural Information Processing Systems Conference; 2018 Dec 3–8; Montreal, QC, Canada; 2018. p. 2942–51.
[83] Saha A, Pahuja V, Khapra MM, Sankaranarayanan K, Chandar S. Complex sequential question answering: towards learning to converse over linked question answer pairs with a knowledge graph. In: Proceedings of the 32nd AAAI Conference on Artificial Intellgence; 2018 Feb 2–7; New Orleans, LA, USA; 2018.
[84] Zhou H, Young T, Huang M, Zhao H, Xu J, Zhu X. Commonsense knowledge aware conversation generation with graph attention. In: Proceeding of the 27th International Joint Conference on Artificial Intelligence; 2018 Jul 13–19; Stockholm, Sweden; 2018. p. 4623–9.
[85] Zhong V, Xiong C, Socher R. Seq2SQL: generating structured queries from natural language using reinforcement learning. 2017. arXiv:1709.00103.
[86] Trivedi P, Maheshwari G, Dubey M, Lehmann J. LC-QuAD: a corpus for complex question answering over knowledge graphs. In: Proceedings of the 2017 International Semantic Web Conference; 2017 Oct 21–25; Vienna, Austria; 2017. p. 210–8.
[87] Talmor A, Berant J. The web as a knowledge-base for answering complex questions. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technology; 2018 Jun 3–5; New Orleans, LA, USA; 2018. p. 641–51.
[88] Levesque HJ, Davis E, Morgenstern L. The winograd schema challenge. In: Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning; 2012 Jun 10–14; Rome, Italy; 2012.
[89] Clark P, Cowhey I, Etzioni O, Khot T, Sabharwal A, Schoenick C, et al. Think you have solved question answering? Try ARC, the AI2 reasoning challenge. 2018. arXiv:1803.05457.
[90] Talmor A, Herzig J, Lourie N, Berant J. CommonsenseQA: a question answering challenge targeting commonsense knowledge. 2018. arXiv:1811.00937.
[91] Sap M, Le Bras R, Allaway E, Bhagavatula C, Lourie N, Rashkin H, et al. ATOMIC: an atlas of machine commonsense for if-then reasoning. In: Proceedings of the 31rd AAAI Conference on Artificial Intelligence; 2019 Jan 27– Feb 1; Honolulu, HI, USA; 2019.
[92] Yang Z, Qi P, Zhang S, Bengio Y, Cohen WW, Salakhutdinov R, et al. HotpotQA: a dataset for diverse, explainable multi-hop question answering. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing; 2018 Oct 31– Nov 4; Brussels, Belgium; 2018. p. 2369–80.
[93] Kočiský T, Schwarz J, Blunsom P, Dyer C, Hermann KM, Melis G, et al. The narrativeQA reading comprehension challenge. Trans Assoc Comput Linguist 2018;6:317–28.
[94] Khashabi D, Chaturvedi S, Roth M, Upadhyay S, Roth D. Looking beyond the surface: a challenge set for reading comprehension over multiple sentences. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technology; 2018 Jun 3–5; New Orleans, LA, USA; 2018. p. 252–62.
[95] Reddy S, Chen D, Manning CD. CoQA: a conversational question answering challenge. 2018. arXiv:1808.07042.
[96] Hudson DA, Manning CD. GQA: a new dataset for real-world visual reasoning and compositional question answering. 2019. arXiv:1902.09506.
[97] Zellers R, Bisk Y, Farhadi A, Choi Y. From recognition to cognition: visual commonsense reasoning. In: Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition; 2019 Jun 16–20; Long Beach, CA, USA; 2019. p. 6720–31.
[98] Zellers R, Bisk Y, Schwartz R, Choi Y. SWAG: a large-scale adversarial dataset for grounded commonsense inference. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing; 2018 Oct 31– Nov 4; Brussels, Belgium; 2018. p. 93–104.