计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2017.20160805
54(6):1144-1149,2017

神经机器翻译前沿进展

刘 洋
(清华大学计算机科学与技术系 北京 100084)
(清华信息科学与技术国家实验室(筹) 北京 100084)
(智能技术与系统国家重点实验室(清华大学) 北京 100084)
(liuyang2011@tsinghua.edu.cn)

Recent Advances in Neural Machine Translation

Liu Yang
(*Department of Computer Science and Technology, Tsinghua University, Beijing* 100084)
(*Tsinghua National Laboratory for Information Science and Technology, Beijing* 100084)
(*State Key Laboratory of Intelligent Technology and Systems (Tsinghua University), Beijing* 100084)

Abstract Machine translation, which aims at automatically translating between natural languages using computers, is one of important research directions in artificial intelligence and natural language processing. Recent years have witnessed the rapid development of neural machine translation, which has replaced conventional statistical machine translation to become the new mainstream technique in both academia and industry. This paper first introduces the basic ideas and state-of-the-art approaches in neural machine translation and then reviews recent important research findings. The paper concludes with a discussion about possible future directions.

Key words artificial intelligence; deep learning; neural machine translation; encoder-decoder framework; attention mechanism

摘 要 机器翻译研究如何利用计算机实现自然语言之间的自动翻译,是人工智能和自然语言处理领域的重要研究方向之一.近年来,基于深度学习的神经机器翻译方法获得迅速发展,目前已取代传统的统计机器翻译成为学术界和工业界新的主流方法.首先介绍神经机器翻译的基本思想和主要方法,然后对最新的前沿进展进行综述,最后对神经机器翻译的未来发展方向进行展望.

关键词 人工智能;深度学习;神经机器翻译;编码器-解码器架构;注意力机制
中图法分类号 TP391

机器翻译研究如何利用计算机实现自然语言之间的自动转换,是人工智能和自然语言处理领域的重要研究方向之一.机器翻译作为突破不同国家和民族之间信息传递所面临的"语言屏障"问题的关键技术,对于促进民族团结、加强文化交流和推动对外贸易具有重要意义.

自20世纪40年代末至今,机器翻译研究大体上经历了2个发展阶段:理性主义方法占主导时期(1949—1992)和经验主义方法占主导时期(1993—2016).早期的机器翻译主要采用理性主义方法,主张由人类专家观察不同自然语言之间的转换规律,以规则形式表示翻译知识.虽然这类方法能够在句

收稿日期:2016-11-10;**修回日期**:2017-02-22
基金项目:国家自然科学基金优秀青年科学基金项目(61522204)
This work was supported by the National Natural Science Foundation of China for Excellent Young Scientists (61522204).

法和语义等深层次实现自然语言的分析、转换和生成,却面临着翻译知识获取难、开发周期长、人工成本高等困难.

随着互联网的兴起,特别是近年来大数据和云计算的蓬勃发展,经验主义方法在 20 世纪 90 年代以后开始成为机器翻译的主流.经验主义方法主张以数据而不是人为中心,通过数学模型描述自然语言的转换过程,在大规模多语言文本数据上自动训练数学模型.这一类方法的代表是统计机器翻译[1-3],其基本思想是通过隐结构(词语对齐、短语切分、短语调序、同步文法等)描述翻译过程,利用特征刻画翻译规律,并通过特征的局部性采用动态规划算法在指数级的搜索空间中实现多项式时间复杂度的高效翻译.2006 年,Google Translate 在线翻译服务的推出标志着数据驱动的统计机器翻译方法成为商业机器翻译系统的主流.尽管如此,统计机器翻译仍面临着翻译性能严重依赖于隐结构与特征设计、局部特征难以捕获全局依赖关系、对数线性模型难以处理翻译过程中的线性不可分现象等难题.

自 2014 年以来,端到端神经机器翻译(end-to-end neural machine translation)[4-5]获得了迅速发展,相对于统计机器翻译而言在翻译质量上获得显著提升.图 1 给出了统计机器翻译与神经机器翻译在 30 种语言对上的对比实验结果[6],神经机器翻译在其中的 27 种语言对上超过统计机器翻译.因此,神经机器翻译已经取代统计机器翻译成为 Google、微软、百度、搜狗等商用在线机器翻译系统的核心技术.

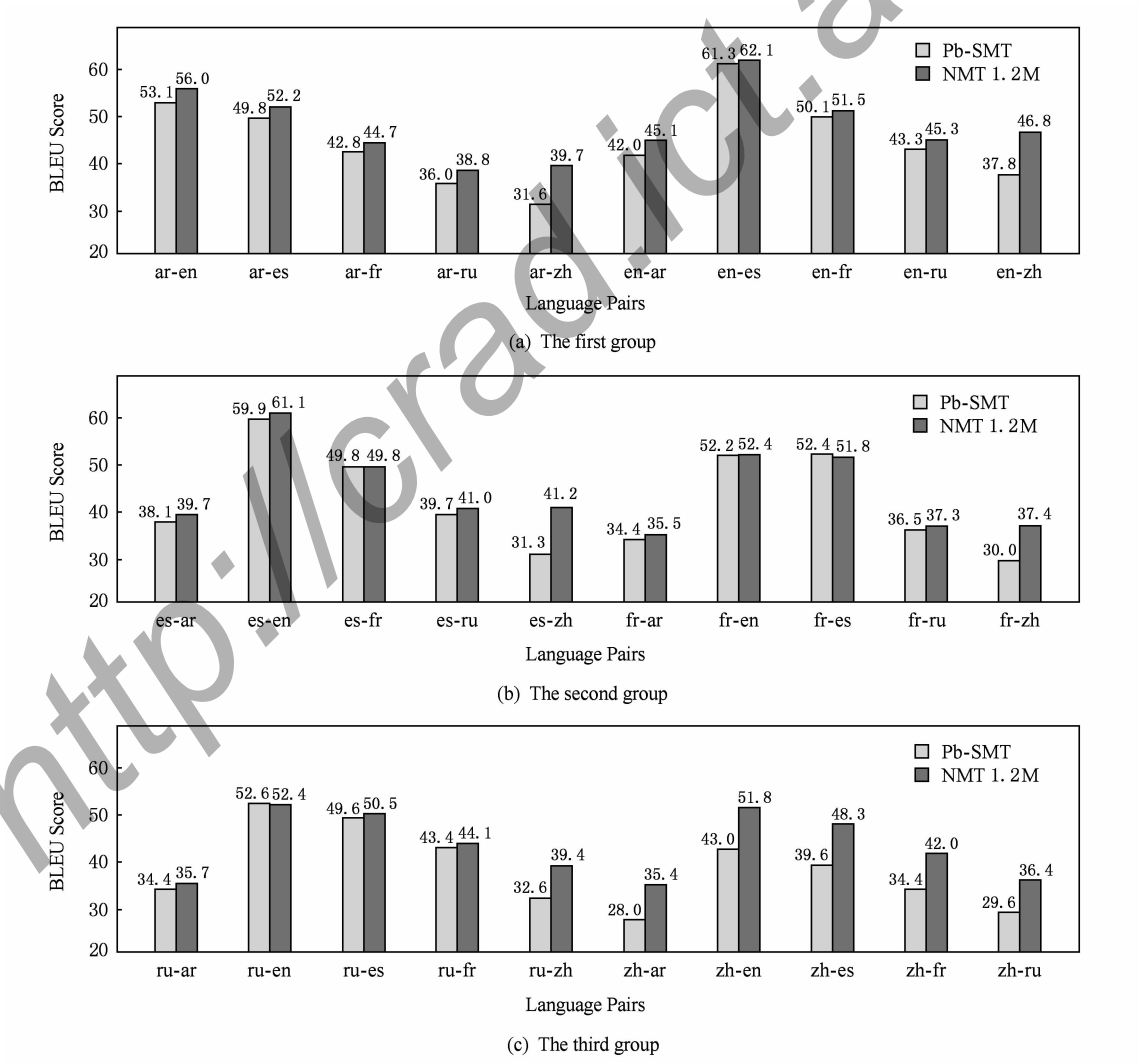

Fig. 1　Comparison between statistical machine translation and neural machine translation (NMT) on 30 languages pairs[6]

图 1　统计机器翻译(Pb-SMT)与神经机器翻译(NMT)在 30 个语言对上的对比[6]

1 神经机器翻译

1.1 编码器-解码器框架

端到端神经机器翻译的基本思想是通过神经网络直接实现自然语言之间的自动翻译. 为此,神经机器翻译通常采用编码器-解码器(encoder-decoder)框架实现序列到序列的转换[5].

以图 2 为例,给定一个中文句子"布什 与 沙龙 举行 了 会谈",编码器-解码器框架首先为每个中文词生成向量表示,然后通过一个递归神经网络(recurrent neural network)从左向右生成整个中文句子的向量表示. 其中,"〈/s〉"表示句尾结束符. 我们将源语言端所使用的递归神经网络称为编码器,即将源语言句子编码成一个稠密、连续的实数向量.

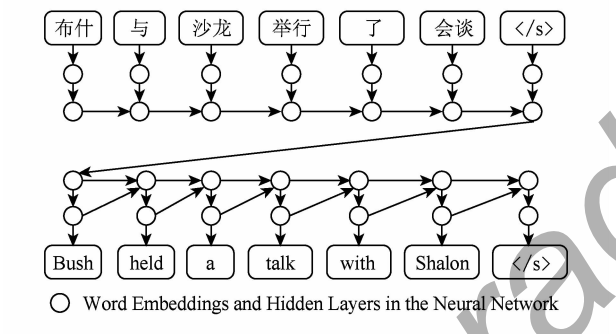

Fig. 2 The encoder-decoder framework

图 2 编码器-解码器框架

此后,目标语言端采用另一个递归神经网络将源语言句子向量反向解码生成英文句子"Bush held a talk with Shalon 〈/s〉". 整个解码过程逐词生成,当生成句尾结束符"〈/s〉"后,解码过程终止. 我们将目标语言端所使用的递归神经网络称为解码器. 需要注意的是,每一个新生成的英文词都作为生成下一个英文词的历史信息. 因此,解码器可以视作包含源语言信息的目标语言的语言模型.

相对于传统的统计机器翻译,基于编码器-解码器框架的神经机器翻译具有 2 个优点:

1) 直接从生数据中学习特征. 统计机器翻译需要人工设计定义在隐结构上的特征来刻画翻译规律. 由于自然语言的高度复杂性,如何确保特征设计覆盖全部语言现象成为重要挑战. 神经网络最大的优势在于能够直接从生数据中学习特征. 研究结果表明,编码器-解码器框架学习到的句子向量表示能够将句法不同、语义相同的句子聚在一起,同时能够将通过调换主语和宾语产生的句法相同、语义不同的句子区分开[5].

2) 能够捕获长距离依赖. 由于自然语言的复杂性和多样性,表达相同含义,不同语言之间的词语顺序差异性非常大. 这种语言结构差异给统计机器翻译带来了严重的挑战. 用户在使用统计机器翻译系统时,经常会发现单个词语翻译很准确,但整体上难以形成合乎语法的句子. 这种现象产生的根源在于,统计机器翻译通过隐结构描述翻译过程,为了在指数级的隐结构组合空间中实现高效搜索,不得不采用局部特征来支持动态规划算法. 除此之外,另一个重要原因在于考虑更多的上下文信息会面临严重的数据稀疏问题. 神经机器翻译通过基于长短时记忆(long short-term memory)的递归神经网络[7]能够有效捕获长距离依赖,同时通过向量表示缓解数据稀疏问题,显著提升了译文的流利度和可读性.

尽管如此,编码器-解码器框架仍然面临一个严重的问题:编码器生成的源语言句子向量表示的维度与源语言句子长度无关. 换句话说,无论是 10 个词的源语言句子、还是 100 个词的源语言句子,都会被编码为固定维度的向量. 这对于编码器处理长距离信息传递带来了极大的挑战. 事实上,即使采用长短时记忆,编码器往往还是难以有效处理长距离依赖,在长句上的翻译质量显著下降[5].

1.2 注意力机制

为了解决定长源语言句子向量难以捕获长距离依赖的问题,文献[6]引入了注意力(attention)机制动态计算源语言端上下文.

如图 3 所示,基于注意力机制的神经机器翻译采用了完全不同的编码器,其目标不再是为整个源语言句子生成向量表示,而是为每个源语言词生成包含全局信息的向量表示. 该编码器首先使用一个

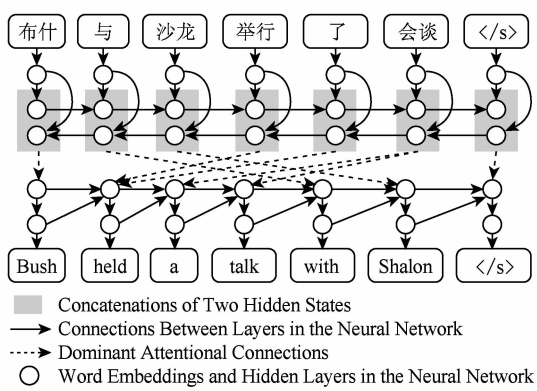

Fig. 3 Attention-based neural machine translation

图 3 基于注意力机制的神经机器翻译

正向递归神经网络将信息从左向右传递,然后再使用一个逆向递归神经网络将信息从右向左传递,最后将正向和逆向的隐状态拼接起来作为源语言词的向量表示.这种做法的优点在于每个源语言词的向量表示均包含了其左侧和右侧的上下文信息.

在目标语言端,解码器在生成每个目标语言词时动态寻找与之相关的源语言上下文.例如,当生成英文词"Bush"的时候,中文词"布什"与之最相关,而"举行"和"了"等词可能并不相关,只需要将"布什"的向量表示作为源端上下文传递到目标端.而当生成英文词"held"的时候,最相关的中文词是"举行"和"了".因此,注意力机制改变了信息传递的方式,能够动态计算最相关的上下文,从而更好地解决了长距离信息传递问题并显著提升了神经机器翻译的性能.因此,基于注意力机制的编码器-解码器模型目前已成为神经机器翻译的主流方法并得到广泛使用.

2　前沿进展

神经机器翻译最早在2013年由文献[8]提出,但当时的翻译效果并不理想,没有超过统计机器翻译.2014年和2015年文献[5]所提出的解码器和编码器框架以及文献[6]提出的注意力机制确定了神经机器翻译的主要架构,但是系统翻译性能仍然仅仅与统计机器翻译持平.可喜的是,神经机器翻译在2016年取得了突飞猛进的进展,翻译性能显著超过统计机器翻译,并且成为以Google Translate为代表的商业翻译系统的核心技术[9].由于近两年来神经机器翻译方面的论文数量非常庞大,难以全部覆盖,本文下面将主要从5个方面对神经机器翻译在2016年取得的重要进展进行简要评述.

2.1　训练算法

给定平行语料库,神经机器翻译的传统训练准则是极大似然估计.文献[10]指出极大似然估计存在2个问题:1)训练目标中的损失函数是定义在词语级别的,而机器翻译的评价指标(如BLEU)通常都是定义在句子或篇章级别的;2)在训练过程中每生成一个目标语言词都是以观测数据作为上下文,而在测试过程中则是以可能存在错误的模型预测作为上下文,因而在训练和测试阶段存在不一致的问题.

为了解决上述问题,文献[11]将最小风险训练(minimum risk training)方法引入神经机器翻译.最小风险训练的基本思想是将模型预测引入训练过程,以机器翻译评价指标来定义损失函数,通过降低模型在训练集上损失的期望值(即风险)来缓解神经机器翻译训练和测试不一致的问题.这种方法可以视作是在统计机器翻译中获得广泛应用的最小错误率训练方法[12]在神经机器翻译中的推广形式.与之类似,文献[10]采用REINFORCE算法将评价指标融入训练过程,文献[13]将训练过程与柱搜索紧密结合.Google推出的神经机器翻译系统中采用上述针对评价指标优化模型参数的训练算法,并发现在大规模训练数据上仍然能够获得稳定且显著的提升[9].

这些方法的优点在于能够直接针对评价指标来优化模型参数,同时训练方法与模型架构和训练指标无关,可以应用到任意的模型架构和评价指标,显著提升了神经机器翻译的性能.

2.2　先验约束

神经机器翻译广受人诟病的一点是缺乏可解释性,神经网络内部都是实数向量,缺乏合理的语言学解释,这使得研究人员对神经机器翻译进行分析和调试变得尤为困难.因而,如何将人类的先验知识与数据驱动的神经网络方法相结合成为神经机器翻译的一个重要研究方向.

目前,将先验知识与神经机器翻译相结合主要有2种方式:

1)直接修改模型架构.文献[14]为了解决神经机器翻译所面临的翻译过度和翻译不足问题,将基于短语的统计机器翻译中广泛使用的覆盖率(coverage)机制引入神经网络,显著提升了神经机器翻译系统输出译文的忠实度.文献[15]也采用修改模型架构的方式将位置偏移、Markov条件、繁殖率等结构化约束加入神经机器翻译.

2)保留原始的模型架构,通过修改训练目标影响模型参数训练.文献[16]发现源语言到目标语言翻译模型和目标语言到源语言翻译模型在计算注意力时均存在不足但可以相互弥补,因而通过在训练目标中加入一致性(agreement)约束鼓励2个模型相互帮助,同时提高了2个翻译方向的性能.

尽管上述工作取得了一定的进展,但如何将先验知识与神经机器翻译相结合仍面临着很大的挑战:无论是修改模型结构还是修改训练目标,都只能加入有限的先验知识,目前仍然缺乏一个通用的框架来支持向神经机器翻译中加入任意的先验知识.

2.3　模型架构

对于神经机器翻译而言,最重要的2个概念是

门阀(gating)和注意力.前者是长短时记忆的核心机制,用来实现信息传递过程中"记忆"和"遗忘"功能;后者则引入动态选择相关上下文的理念.是否还存在更先进的机制来进一步改进神经机器翻译的模型架构?

文献[17]提出的神经网络图灵机近年来广受关注.如果将传统递归神经网络中的隐状态比作为"内存"来存储短时记忆的话,神经网络图灵机则主张用"外存"来存储长时记忆,其存储单元寻址方式类似于注意力机制.无独有偶,文献[18]提出的记忆网络(memory networks)也提出了非常类似的思想.目前,神经网络图灵机在机器翻译中的成功应用很少,目前主要的进展是文献[19]将利用记忆机制来改进解码器,显著提升了神经机器翻译的质量.然而,memory 的寻址机制实际上与 attention 的计算非常类似.在同时使用长短时记忆、attention 和 memory 的情况下,memory 能够提供什么额外的有用信息,目前仍没有清晰的语言学解释,有待进一步探索.

另一个研究方向是依据统计机器翻译中广泛使用的语言学结构来建立神经机器翻译模型.这方面的代表性工作是文献[20]提出的树到序列神经机器翻译,他们将统计机器翻译中的树到序列模型与神经网络相结合.这样的建模方式存在一定的争议性,因为深度学习通常主张从生数据中学习表示,而不是依赖于句法树这样由语言学家发明的人造结构.如何实现语言结构与神经网络的有效结合将继续成为神经机器翻译的研究热点之一.

2.4 受限词汇量

神经机器翻译的解码器在生成目标语言词语时,需要通过在整个目标语言词汇表上进行归一化来计算概率分布,因而计算复杂度极高.为了降低复杂度,神经机器翻译系统往往将词汇表限制为高频词,并将其他所有低频词视为未登录词.2015 年,神经机器翻译的研究人员主要通过未登录词替换[21]和采样[22]等方法处理受限词汇量问题.

在 2016 年,研究人员更加关注如何用细粒度意义表示单元(如字母、字、语素、亚词等)解决受限词汇量问题.文献[23]提出了词语-字母混合模型,利用词语模型处理高频词,利用字母模型处理低频词.文献[24]提出利用字节对编码(byte pair encoding)自动发现亚词(subword),进而建立基于亚词的神经机器翻译模型.文献[25]提出一种不依赖于显式切分的、基于字母的编码器,在目标语言端缓解了受限词汇量问题.

上述方法有效解决了神经机器翻译词汇量受限的问题,但仍需在更多的黏着语、孤立语和屈折语上进一步验证.

2.5 低资源语言翻译

作为一种数据驱动方法,神经机器翻译的性能高度依赖于平行语料库的规模、质量和领域覆盖面.由于神经网络的参数规模庞大,只有当训练语料库达到一定规模,神经机器翻译才会显著超过统计机器翻译[26].然而,除了中文、英文等资源丰富语言,世界上绝大多数语言都缺乏大规模、高质量、广覆盖率的平行语料库.即使对于中文和英文,现有平行语料库的领域也主要集中在政府文献和时政新闻,对于绝大多数领域而言依然严重缺乏数据.

因此,如何充分利用现有数据来缓解资源匮乏问题成为 2016 年神经机器翻译的一个重要研究方向.文献[27]提出利用现有机器翻译系统翻译单语数据,通过构造伪平行语料库来缓解平行语料库匮乏问题.文献[28]将自动编码器引入神经机器翻译,提出了基于双语语料库和单语语料库的半监督学习方法.文献[26]将迁移学习引入低资源神经机器翻译,将在资源丰富语言平行语料库训练的模型参数迁移到资源匮乏语言翻译模型的训练过程中.

尽管上述方法都观察到翻译知识从资源丰富的语言对迁移到资源匮乏的语言对能够显著提升神经机器翻译的效果,但是由于向量表示缺乏可解释性,这种知识迁移的内在机制仍然没有得到充分研究.事实上,对于整个神经机器翻译研究而言,目前对于翻译过程中的内部运行机制的理解仍然十分困难,神经网络隐层的向量表示缺乏清晰的语言学解释,这将成为未来的研究重点.

3 总结与展望

综上所述,神经机器翻译是近年来涌现出来的一种基于深度学习的机器翻译方法,目前已经取代传统的统计机器翻译,成为新的主流技术.相对于统计机器翻译,神经机器翻译不仅能够从生数据中直接学习特征,而且能够通过长短时记忆和注意力等机制有效处理长距离依赖.尽管如此,神经机器翻译研究仍然面临着诸多挑战,5 个科学问题仍有待进一步探索:

1) 如何设计表达能力更强的模型?

2) 如何提高语言学方面的可解释性?

3) 如何降低训练复杂度?

4) 如何与先验知识相结合?

5) 如何改进低资源语言翻译?

我们相信,神经机器翻译在未来会获得进一步的发展,通过高质量的机器翻译服务造福社会大众.

参　考　文　献

[1] Brown P, Della Pietra S, Della Pietra V, et al. The mathematics of statistical machine translation: Parameter estimation [J]. Computational Linguistics, 1993, 19(2): 263-311

[2] Och F, Ney H. Discriminative training and maximum entropy models for statistical machine translation [C] //Proc of the 40th ACL. Stroudsburg, PA: ACL, 2002: 295-302

[3] Chiang D. A hierarchical phrase-based model for statistical machine translation [C] //Proc of the 43rd ACL. Stroudsburg, PA: ACL, 2005: 263-270

[4] Sutskever I, Vinyals O, Le Q. Sequence to sequence learning with neural networks [C] //Proc of the 28th NIPS. Red Hook, NY: Curran Associates Inc, 2014: 3104-3112

[5] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate [J]. arXiv: 1409.0473, 2014

[6] Junczys-Dowmunt M, Dwojak T, Hoang H. Is neural machine translation ready for deployment? A case study on 30 translation directions [J]. arXiv: 1610.01108v2, 2016

[7] Hochreiter S, Schmidhuber J. Long short-term memory [J]. Neural Computation, 1997, 9(8): 1735-1780

[8] Kalchbrenner N, Blunsom P. Recurrent continuous translation models [C] //Proc of EMNLP. Stroudsburg, PA: ACL, 2013: 1700-1709

[9] Wu Yonghui, Schuster M, Chen Zhifeng, et al. Google's neural machine translation system: Bridging the gap between human and machine translation [J]. arXiv: 1609.08144v2, 2016

[10] Ranzato M, Chopra S, Auli M, et al. Sequence level training with recurrent neural networks [J]. arXiv: 1511.06732, 2015

[11] Shen Shiqi, Cheng Yong, He Zhongjun, et al. Minimum risk training for neural machine translation [C] //Proc of the 54th ACL. Stroudsburg, PA: ACL, 2016: 1683-1692

[12] Och F. Minimum error rate training in statistical machine translation [C] //Proc of the 41st ACL. Stroudsburg, PA: ACL, 2003: 160-167

[13] Wiseman S, Rush A. Sequence-to-sequence learning as beam-search optimization [C] //Proc of EMNLP. Stroudsburg, PA: ACL, 2016: 1296-1306

[14] Tu Zhaopeng, Lu Zhengdong, Liu Yang, et al. Modeling coverage for neural machine translation [C] //Proc of the 54th ACL. Stroudsburg, PA: ACL, 2016: 76-85

[15] Cohn T, Hoang C, Vymolova E, et al. Incorporating structural alignment biases into an attentional neural translation model [C] //Proc of NAACL. Stroudsburg, PA: ACL, 2016: 876-885

[16] Cheng Yong, Shen Shiqi, He Zhongjun, et al. Agreement-based joint training for bidirectional attention-based neural machine translation [C] //Proc of the 25th IJCAI. Palo Alto, CA: IJCAI, 2016: 2761-2767

[17] Graves A, Wayne G, Danihelka I. Neural turing machines [J]. arXiv: 1410.5401v2, 2014

[18] Weston J, Chopra S, Bordes A. Memory networks [J]. arXiv: 1410.3916, 2014

[19] Wang Mingxuan, Lu Zhengdong, Li Hang, et al. Memory-enhanced decoder for neural machine translation [C] //Proc of EMNLP. Stroudsburg, PA: ACL, 2016: 278-286

[20] Eriguchi A, Hashimoto K, Tsuruoka Y. Tree-to-sequence attentional neural machine translation [C] //Proc of the 54th ACL. Stroudsburg, PA: ACL, 2016: 823-833

[21] Luong M, Sutskever I, Le Q, et al. Addressing the rare word problem in neural machine translation [C] //Proc of the 53rd ACL. Stroudsburg, PA: ACL, 2015: 11-19

[22] Jean S, Cho K, Memisevic R, et al. On using very large target vocabulary for neural machine translation [C] // Proc of the 53rd ACL. Stroudsburg, PA: ACL, 2015: 1-10

[23] Luong M, Manning C. Achieving open vocabulary neural machine translation with hybrid word-character models [C] // Proc of the 54th ACL. Stroudsburg, PA: ACL, 2016: 1054-1063

[24] Sennrich R, Haddow B, Birch A. Neural machine translation of rare words with subword units [C] //Proc of the 54th ACL. Stroudsburg, PA: ACL, 2016: 1715-1725

[25] Chung J, Cho K, Bengio Y. A character-level decoder without explicit segmentation for neural machine translation [C] //Proc of the 54th ACL. Stroudsburg, PA: ACL, 2016: 1693-1703

[26] Zoph B, Yuret D, May J, et al. Transfer learning for low-resource neural machine translation [C] //Proc of EMNLP. Stroudsburg, PA: ACL, 2016: 1568-1575

[27] Sennrich R, Haddow B, Birch A. Improving neural machine translation models with monolingual data [C] //Proc of the 54th ACL. Stroudsburg, PA: ACL, 2016: 86-96

[28] Cheng Yong, Xu Wei, He Zhongjun, et al. Semi-supervised learning for neural machine translation [C] //Proc of the 54th ACL. Stroudsburg, PA: ACL, 2016: 1965-1974

Liu Yang, born in 1979. PhD, associate professor, PhD supervisor. Member of CCF and Chinese Information Processing Society. His main research interests include natural language processing and machine translation.