神经机器翻译未登录词解析科技报告

中国科学技术信息研究所

目　录

插 图 清 单

附 表 清 单

正　文

1 引言

1.1 研究背景和意义

科技文献（Scientific and Technical Document）是记载科学技术等知识的载体。科学技术的快速发展促生了很多科技文献，其中不乏大量的外文科技文献。科研人员从其他语言的科技文献中获取信息较为困难，跨语言成为交流的主要障碍。多数科技文献仍然需要通过人工翻译才能为更多的科技工作者使用，这种方式效率低、成本高。随着计算机的普及以及相关技术的快速发展，利用计算机进行语言间转换的方法——机器翻译，成为了突破口，机器翻译，是利用计算机将一种自然语言（源语言）转换为另一种自然语言（目标语言）的过程。它是计算语言学的一个分支，是人工智能的终极目标之一，具有重要的科学研究价值。

机器翻译能够快速实现语言间的大批量翻译，随着近两年神经机器翻译（Neural Machine Translation，NMT）的兴起，其效果也逐步满足了实用化需求。神经机器翻译一般采用端到端（End-to-End）模型。该模型使用编码器（encoder）-解码器（decoder）框架。实际应用的系统需要考虑计算效率，通常需要根据词频限制词表规模，这样会导致某些低频词语不能被解码出来，出现未登录词问题，直接影响了神经机器翻译系统的性能。

目前专门针对神经机器翻译中未登录词问题的研究还较少，多集中在字符级别，即将词语拆分成更小的字符单元来减少词汇之间的差异性从而减少未登录词。而统计机器翻译（Statistical Machine Translation，SMT）引擎无需受到词表的限制，在翻译未登录词时比神经机器翻译具有更大的优势。本研究针对科技文献神经机器翻译，利用统计机器翻译中经常使用的词对齐生成双语词典，将 SMT 引擎对于未登录词的处理结果与 NMT 的未登录词翻译结果相融合，总结科技词汇构词规律，利用科技词汇构词特征，结合点互信息，在保留词汇义素完整的同时，对词汇表进行优化，成功减少了未登录词比例，最终达到了翻译效果提升的目的。

1.2 相关研究进展

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[1]。这一时期的机器翻译包括基于规则的机器翻译（Rule-based Machine Translation，RBMT）、基于实例的机器翻译（Example-based Machine Translation，EBMT）。基于规则的机器翻译又分为直接机器翻译、基于迁移的机器翻译和语际机器翻译。但是，无论是哪种基于规则的机器翻译，其效果都不好。1984 年，日本京都大学的長尾真提出了一个思想：使用现成的短语而不是重复进行翻译。EBMT 让全世界的科学家看到了方向：事实证明，可以直接向机器输入已有的翻译，而不必花费多年时间构建规则和例外。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。1990 年初，IBM 研究中心首次展示了一个对规则和语言学一无所知的机器翻译系统。它分析了两种语言的相似文本并且试图理解其中的模式。两种语言中的同一句子被分成单词，然后再进行匹配。这种操作重复了近 5 亿次，记录下了很多模式，比如「Das Haus」被翻译成「house」或「building」或「construction」等词的次数。如果大多数时候源词都被翻译成「house」，那么机器就会使用这一结果。注意该方法没有使用任何规则，也没有使用任何词典——所有的结论都是由机器完成的。其指导方针是统计结果和这样的逻辑——如果语料这样翻译，那么机器也这样翻译。统计机器翻译（Statistical Machine Translation，SMT）由此诞生。这个方法比之前的所有方法都更加有效和准确。而且无需语言学家。使用的文本越多，得到的翻译结果就越好。SMT 也有几个分支：基于词的 SMT，基于短语的 SMT 和基于句法的 SMT。其中，基于短语的 SMT 在 2016 年以前是机器翻译的主流方法。这种方法基于所有基于词的翻译原则：统计、重新排序和词法分析。但是，在学习时，它不仅会将文本分成词，还会分成短语。因此，这个机器能学习翻译稳定的词组合，这能显著提升准确度。总而言之，统计机器翻译的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。自 2014 年以来，端到端神经机器翻译（end-to-end neural machine translation）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，神经机器翻译在 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。

目前的神经机器翻译主要依赖于深度学习（Deep Learning）技术。深度学习根源于神经网络。在神经网络宏大的历史发展中，深度学习（Deep Learning）应当算是其第 3 次崛起的

标志。

20 世纪 50 年代，美国计算机科学家 Rosenblatt[2]最早提出可以模人类感知能力的数学模型——感知器模型。感知器模型是世界上第一个有实用价值的神经网络模型，是现代神经网络的出发点。但是，Minsky 和 Seymour Papert[3]于 1969 年出版了著名的 Perceptrons（《感知器》）一书：感知器（原始感知器）仅能解决一阶谓词逻辑问题，不能解决高阶谓词逻辑问题，如：异或运算。虽然 Minsky 也认为多层网络可以解决非线性问题，但是，在当时，这个问题还不可解。由于 Minsky 是人工智能科学的创始人之一，该书的观点颇具影响力，直接造成了人工神经领域发展的第一次没落。这之后的十多年间，神经网络渐渐淡出人们的视线。

神经网络第二次崛起于 20 世纪 80 年代。当时，由于超大规模集成电路的制作工艺已近成熟，以及传统的人工智能算法理论在解决知识发现上表现得无能为力，人们又开始从神经网络算法上寻求突破。1982 年由美国物理学家 John Hopfield[4]提出的模人脑的神经网络模型，即著名的 Hopfield 网络，让人们重新认识了神经网络，掀起了神经网络的研究热潮。1985 年，Ackley、Hinton 和 Sejnowski[5]将模退火算法应用到神经网络训练中，提出了 Boltzmann 机。同年，Hinton、Rumelhart 和 Williams[6]发现，误差的反向传播可以有效地解决多层网络中隐节点的学习问题，证明 Minsky 对多层网络不存在有效学习方法这个断言并不正确。他们提出了多层前馈神经网络的学习算法，即 BP 算法，人工神经网络才又重新引起人们的注意，并开始成为新的研究热点。后向传播算法在深度学习中也经常被用作参数学习的方法。1990 年，Elman[7]提出简单循环神经网络（Simple Recurrent Network, SRN），同年，Werbos[8]提出了随时间进行的反向传播算法——循环神经网络的参数训练算法。1994 年，Bengio 等人[9]对“长期依赖问题”进行了深入的研究，他们发现一些使训练简单循环神经网络变得非常困难的相当根本的原因——存在梯度爆炸和消失问题。为了解决长期依赖问题，Hochreiter 和 Schmidhuber[10]于 1997 年提出长短时记忆神经网络（Long Short-Term Memory Neural Network）。但是，第二代神经网络存在以下缺点：经验风险最优需要大量的训练数据，结构上处理非线性问题存在不稳定性，而且 2000 年以后，因为当时计算机的计算能力不足以支持训练大规模的神经网络，并且随着支持向量机等方法的兴起，神经网络又一次陷入低潮。支持向量机一派一度占了上风，这个优势一直保持了近十年。

直到 2006 年，Hinton[11]发现多层前馈神经网络可以先通过逐层预训练，再用反向传播算法进行精调的方式，进行有效学习。并且，近年来，计算机计算能力（大规模并行计算，GPU）的提高，计算机是已经可以训练大规模的人工神经网络。深度学习是从机器学习中的