doi:10.3772/j.issn.2095-915x.2017.03.004

神经网络语言模型在统计机器翻译中的应用

1. 模式识别国家重点实验室 中国科学院自动化研究所 中国科学院大学 北京 100190；
2. 中国科学院脑科学与智能技术卓越创新中心 上海 200031

张家俊[1] 宗成庆[1,2]

摘 要 近两年来，神经机器翻译（Neural Machine Translation, NMT）模型主导了机器翻译的研究，但是统计机器翻译（Statistical Machine Translation, SMT）在很多应用场合（尤其是专业领域）仍有较强的竞争力。如何利用深度学习技术提升现有统计机器翻译的水平成为研究者们关注的主要问题。由于语言模型是统计机器翻译中最核心的模块之一，本文主要从语言模型的角度入手，探索神经网络语言模型在统计机器翻译中的应用。本文分别探讨了基于词和基于短语的神经网络语言模型，在汉语到英语和汉语到日语的翻译实验表明神经网络语言模型能够显著改善统计机器翻译的译文质量。

关键词： 统计机器翻译，神经网络语言模型，基于词的语言模型，基于短语的语言模型
中图分类号： G35

Application of Neural Language Model in Statistical Machine Translation

1. National Laboratory of Pattern Recognition, Institute of Automation Chinese Academy of Sciences, University of Chinese Academy of Sciences, Beijing 100190, China;
2. CAS Center for Excellence in Brain Science and Intelligence Technology, Shanghai 200031, China

ZHANG JiaJun[1] ZONG ChengQing[1,2]

Abstract Neural Machine Translation (NMT) dominates the research of machine translation in recent two years. However, Statistical Machine Translation (SMT) is very competitive in many scenarios such as some specific domains. It became a key issue how to apply deep learning technology to improve SMT performance. As language model is one of the most crucial modules in SMT, this paper investigated the usage of neural language model in statistical machine translation. We explored respectively the word-based and phrase-based neural language model, and evaluated the models on both Chinese-to-English and Chinese-to-

基金项目： 本文受国家自然科学基金“视听觉信息的认知计算”重点项目：面向汉语文本理解的语义计算方法（91520204），国家自然科学基金面上项目：基于弱监督的神经网络翻译模型研究（61673380）的资助。
作者简介： 张家俊（1983–），博士，副研究员，研究方向：机器翻译、自然语言处理、深度学习，Email：jjzhang@nlpr.ia.ac.cn；宗成庆（1963–），博士，研究员，研究方向：机器翻译、自然语言处理、情感分析。

doi:10.3772/j.issn.2095-915x.2017.03.004

Japanese translation tasks. The extensive experiments demonstrated that the neural language models can significantly improve the translation performance of statistical machine translation.

Keywords: Statistical machine translation, neural language model, word-based language model, phrase-based language model

1 引言

机器翻译旨在利用计算机技术将一种自然语言（源语言）自动转换为另一种自然语言（目标语言）。自 20 世纪 90 年代初开始，数据驱动的机器翻译模型成为主流。其中，如何从人工翻译的双语对照语料中学习翻译知识是核心。尽管这两年神经机器翻译成为新的研究范式[1-7]，统计机器翻译模型在很多应用（尤其是专业领域）中仍有很强的竞争力。在统计机器翻译中，无论是基于词的模型[8]、基于短语的模型[9-10]、还是基于树的模型[11-13]，度量译文流畅度的语言模型始终是其中最为核心的模块之一。本文拟从改善语言模型的角度出发，研究神经网络语言模型在统计机器翻译中的应用。

我们以基于短语的统计机器翻译为基线系统，以前馈神经网络为框架研究神经网络语言模型[14]。不少学者对以词为单位的神经语言模型进行了研究，并证明了有效性[15]。然而，基于短语的统计机器翻译在生成目标译文时都是以短语为单元进行的组合。因此，以短语为基本单元的语言模型也值得研究[16]。因此，本文除了探讨基于词的神经语言模型，还研究了基于短语的神经语言模型，并将其融入基于短语的统计机器翻译模型。

本文以汉语到英语、汉语到日语为翻译任务，验证神经语言模型的有效性，并对比基于词的神经语言模型和基于短语的神经语言模型。详细的实验结果表明，神经网络语言模型能够显著改善统计机器翻译的译文质量。同时，我们对比发现：基于词的神经语言模型无法替代传统的 n- 元语言模型，基于短语的神经语言模型无法替代基于词的模型，而融合词和短语的神经语言模型能够取得最好的翻译效果。

2 基于短语的统计机器翻译

在统计机器翻译中，给定一个源语言句子 f，模型将在所有候选目标语言译文中搜索一个概率最大的句子 e：

$$e^* = \arg\max_e p(e|f)$$

其中，$p(e|f)$ 一般采用对数线性模型分解为若干个子模型：

$$e^* = \arg\max_e p(e|f) = \arg\max_e \frac{\exp\left(\sum_i \lambda_i h_i(f,e)\right)}{\sum_{e'} \exp\left(\sum_i \lambda_i h_i(f,e')\right)}$$

其中，$h_i(f, e)$ 可以是任意翻译特征，例如翻译模型特征、调序模型特征和语言模型特征，λ_i 为对应的特征权重。一般地，特征权重可通过最小错误率训练的方法获得[17]。

基于短语的模型在统计机器翻译中应用最为广泛。这里的短语并非句法意义上的短语，而是指任意连续的词串。该模型的基本思想是：

在训练阶段从双语句子对齐的平行语料中自动抽取源语言短语到目标语言短语的翻译规则并学习其概率，在翻译阶段将源语言句子切分为短语序列，利用翻译规则得到目标语言句子的短语序列，然后借助短语重排序模型和语言模型对目标语言句子的短语序列进行排序，最终获得最佳的目标译文。本文以基于最大熵的括弧转录文法（MEBTG）为例介绍基于短语的翻译模型[18]。

在训练阶段，我们首选利用词对齐工具获得双语对齐句子中的词语互译关系；然后，我们抽取符合约束的短语翻译规则（如图 1 中的 a-e）并估计规则的概率。同时，从词语对齐的双语数据中，最大熵方法可以用来训练短语调序模型。此外，双语对照语料的目标端和目标语言大规模单语数据可以用来训练强大的语言模型。

在测试解码阶段，基于短语的模型为源语言句子搜索最佳的短语切分，匹配最好的短语翻译规则，并寻找最佳方式组合目标短语最终生成目标语言译文。如图 1 所示，在翻译该汉语句子时，首先利用短语翻译规则（a-e）获得若干个短语的目标语言译文。然后，我们将采用短语调序规则（f-i）重新排列目标语言短语获得最终译文。其中，规则 g 表示 “the two countries” 与 “the relations between” 应该交换顺序。规则 f, h 和 i 表示顺序拼接目标短语译文。译文组合过程中，语言模型将衡量译文的流畅度。

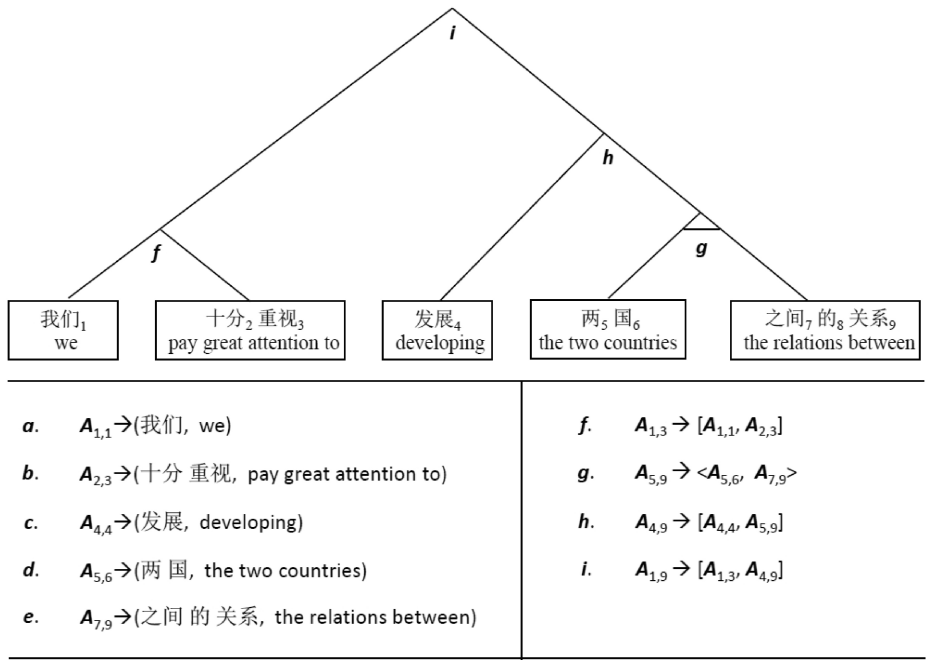

图1 短语翻译模型实例

3 神经网络语言模型

由前述介绍可知，语言模型是译文生成过程中至关重要的因素之一。典型地，我们用 n-元语言模型计算任意时刻译文片段的概率：

$$p(e_0 e_1 \cdots e_m) = \prod_i p(e_i \mid e_{i-n+1} \cdots e_{i-1})$$

一般地，$p(e_i \mid e_{i-n+1} \cdots e_{i-1})$ 采用最大似然方法进行估计：

doi:10.3772/j.issn.2095-915x.2017.03.004

$$p(e_i \mid e_{i-n+1} \cdots e_{i-1}) = \frac{count(e_{i-n+1} \cdots e_i)}{count(e_{i-n+1} \cdots e_{i-1})}$$

这种基于字符串计数的方法不仅面临数据稀疏问题，而且无法捕捉语言单元之间的语义相似性：例如“出台”和“颁布”语义相近，那么出现在相似上下文中的概率也应该相似，然而最大似然计数的方法无法实现这一点。

3.1 基于词的神经网络语言模型

因此，很多研究者尝试采用神经网络结构刻画语言模型。我们以 Bengio 提出的前馈神经网络语言模型为例进行介绍[14]。如图 2 所示，若我们要计算概率 $p(e_i \mid e_{i-n+1} \cdots e_{i-1})$，首先将上下文中的每个词语 $e_{i-n+1} \cdots e_{i-1}$ 映射为定长的低维实数向量，并进行拼接作为下一层的输入，然后进行一层或多层的线性和非线性映射得到深层表示，最后通过 softmax 操作得到下一个词 e_i 的出现概率。

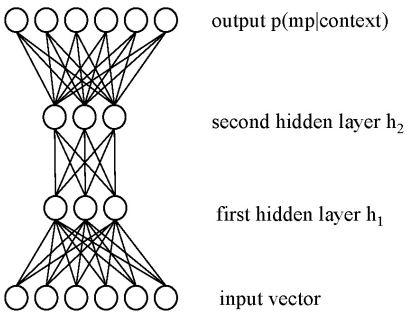

图2 前馈神经网络语言模型示例

3.2 基于短语的神经网络语言模型

由图 1 所示，短语翻译系统的译文都是由目标语言短语组合而成。那么，很自然地，衡量组合之后的译文是否通顺和流畅应该对短语之间的相互依赖关系（即短语语言模型）进行建模。这里涉及两个关键问题：1）短语的定义；2）目标语言单语句子中的短语识别与划分。

关于短语的定义，我们引入“最小短语”的概念，并从双语句对的角度定义最小短语。给定一个经过词语对齐的句对 (f, e, A)，其中 $f=f_0 \ldots f_m$ 表示源语言句子的词串，$e=e_0 \ldots e_m$ 表示目标语言的词串，$A \subseteq \{0 \cdots m-1\} \times \{0, n-1\}$ 是双语句对间词语对齐矩阵。每一个目标语言端最小短语 $e_i \ldots e_i$ 都应满足下面的条件：

a) 根据词对齐矩阵搜寻与目标语言短语 $e_i \ldots e_i$ 词语对齐的源语言短语 $f_k \ldots f_l$，必须满足：如果 $i \le i' \le j$ 如果 $k \le k' \le l$；

b) 无法从 $e_i \ldots e_i$ 找到更小的语言单位满足 a) 的要求。

根据上述定义，我们可以非常容易地从双语句对中的目标语言部分找到唯一的最小短语划分，图 3 描述了一个示例。

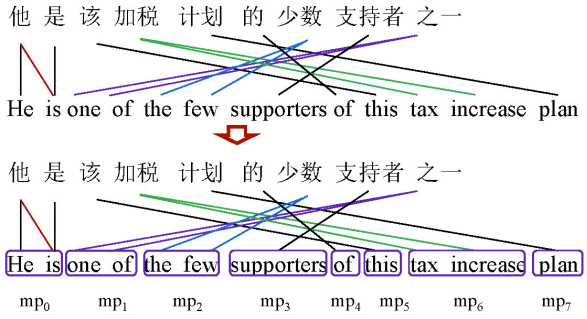

图3 最小短语的定义

由于最小短语的识别和划分都是在词语对齐的双语句子中进行的，如何在目标语言单语数据中进行有效的最小短语识别与划分成为大规模单语数据的应用瓶颈。本文采用有监督的序列标注方法对任意测试句子进行最小短语识别与划分。

由每个互为翻译的句对可获得目标语言部分唯一的最小短语划分，从而便可视为训练数

据学习单语数据上的最小短语划分。本文设计了有效的词汇化特征，并采用感知器模型作为分类器确定输入的每个词语应该最小短语的开始（B）、中间（M）还是结尾（E）。

4 实验与分析

在这一小节中，我们将介绍实验设置、实验结果，并对结果进行简单分析。

4.1 实验设置

我们所采用的基线系统就是基于最大熵括弧转录文法的短语翻译模型 MEBTG，不同的语言模型将作为一个特征融入 MEBTG 的对数线性模型中。所有子模型的特征权重都是在开发集上采用最小错误率训练获得。

我们选择汉语到英语以及汉语到日语的翻译任务来对比不同语言模型对机器翻译性能的影响。对于汉语到英语的翻译任务，双语训练数据来源于 LDC，共 206 万平行句对（含 2770 万汉语词以及 3190 万英语词）。我们选择 NIST MT03 作为开发集，MT05、MT06 以及 MT08 作为测试集。每个测试用例对应 4 个人工参考译文。

对于汉语到日语的翻译任务，双语训练数据来源于亚洲机器翻译评测（WAT），共 67 万平行句对。我们采用了标准的开发集（2090 句）以及两个测试集（分别为 2107 和 2143 句）。每个汉语句子仅有 1 个人工参考译文。

由于语言模型是本文关注的焦点，而语言模型的性能又取决于所使用的单语语料规模，因此，在汉语到英语的翻译任务上，除了平行语料的目标语言部分，我们还采用了英文 Gigaword 的新华部分（约 1100 万句子）。在汉语到日语的翻译任务中，由于其科技文献领域的特殊性，我们仅采用双语训练语料的目标语言部分来训练基于词和短语的语言模型。

为了进行详细的实验对比，我们考察了 4 组不同的语言模型：

W-KN：也就是基线系统所采用的基于词的语言模型，语言模型概率由最大似然估计并经过 Kneser-Ney 平滑获得。

W-NN：基于词的神经网络语言模型，考虑到计算效率，我们在计算语言模型概率时仅保留 16 万最高频的词汇。

MP-KN：基于最小短语的语言模型，语言模型概率的估计方法类似于 W-KN。

MP-NN：基于最小短语的神经网络语言模型，类似于 W-NN，我们也仅保留 16 万最高频的目标语言词汇。

值得注意的是，这些语言模型可分别（或一起）作为模型特征融入基线系统的对数线性模型。我们用重采样技术衡量两个机器翻译系统之间的差异是否统计显著 [19]。

4.2 实验结果

表 1 给出了不同语言模型在汉语到英语翻译任务上的详细实验对比。作为单一模型，我们可以看到基于词的语言模型要优于基于最小短语的语言模型（W-KN vs. MP-KN，W-NN vs. MP-NN）。我们认为这一现象主要是因为基于最小短语的语言模型容易导致数据稀疏问题。同时，我们也发现，基于神经网络的语言模型概率估计方法无法替代基于最大似然估计加平滑的方法。

doi:10.3772/j.issn.2095-915x.2017.03.004

不管是基于词的语言模型还是基于最小短语的语言模型，基于神经网络的概率估计方法取得的翻译性能都明显低于基于最大似然的估计方法（W-KN vs. W-NN，MP-KN vs. MP-NN）。

融合不同的语言模型对提升翻译性能是否有帮助是我们主要关心的问题。首先在基于词的模型中，融入两种不同的概率估计方法，我们发现 W-KN-NN 相比于 W-KN 取得了显著的性能提升，说明神经网络语言模型可以作为基线系统的一个补充。进一步加入基于最小短语的语言模型，我们发现，综合两种粒度的语言模型以及两种概率估计方法（W-KN-NN+MP-KN-NN）可以取得最优的译文质量。相比于基线系统 W-KN，在每个数据集上都能取得超过 1 个 BLEU 值的性能提升。这一实验结果进一步说明，基于最小短语的语言模型与基于词的语言模型具有互补关系。基于词的语言模型能够刻画精细的局部依赖关系，而基于最小短语的语言模型可以捕捉更大范围片段与片段之间的依赖关系。

表1 不同语言模型在汉语到英语翻译任务上的性能对比

方法	MT03	MT05	MT06	MT08
W-KN	35.81	34.69	33.83	27.17
W-NN	34.73	33.62	32.75	26.54
MP-KN	34.39	33.26	32.51	25.65
MP-NN	33.65	32.83	31.96	25.21
W-KN-NN	**36.40**	**35.45**	**34.58**	**27.87**
W-KN+MP-KN	36.26	35.36	34.45	25.54
W-KN+MP-KN-NN	36.83	35.87	35.30	28.40
W-KN-NN+MP-KN-NN	**36.95**	**36.13**	**35.56**	**28.92**

表 2 展示了不同语言模型在汉语到日语翻译任务上的性能对比。从前两行实验结果（W-KN vs. W-KN-NN）可以看出，基于神经网络的语言模型概率估计方法虽有性能提升，但提升幅度非常有限。我们认为，很大原因在于训练语言模型的数据相对较小，仅有不到 70 万的句子很难学习一个复杂的神经网络模型。当融入基于最小短语的语言模型后，我们发现整个翻译系统相比于基线系统可以取得统计显著的翻译质量提升。可见，基于最小短语的语言模型不仅在汉英翻译中有效，在汉日翻译中也同样有效。

表2 不同语言模型在汉语到日语翻译任务上的性能对比

方法	开发集	测试集 -1	测试集 -2
W-KN	35.71	36.62	35.47
W-KN-NN	36.13	37.10	35.84
W-KN-NN+MP-KN-NN	**36.48**	**37.46**	**36.15**

5 总结

本文在基于短语的统计机器翻译框架中深入研究了神经网络语言模型对机器翻译性能的影响。我们不仅研究了基于词的神经网络语言模型，而且也探索了基于（最小）短语的神经网络语言模型。为了使得短语语言模型可计算，

我们给出了最小短语的定义，并设计了单语数据上的最小短语划分方式。汉英和汉日机器翻译的详细实验显示，基于最小短语的语言模型无法替代基于词的语言模型，同时，基于神经网络的语言模型概率估计方法也无法替代基于最大似然估计的概率估计方法。值得欣慰地的是，融合不同粒度的语言模型、不同概率估计方法都可以有效改善统计机器翻译系统的译文质量。

参考文献

[1] Bahdanau D, Cho K, Bengio Y. Neural Machine Translation by Jointly Learning to Align and Translate[J]. Computer Science, 2014: 1-15.

[2] Sutskever I, Vinyals O, Le Q V. Sequence to Sequence Learning with Neural Networks[J]. Advances in Neural Information Processing Systems, 2014(4): 3104-3112.

[3] Shen S, Cheng Y, He Z, et al. Minimum Risk Training for Neural Machine Translation[J]. Computer Science, 2015.

[4] Wu Y, Schuster M, Chen Z, et al. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation[J]. 2016, arXiv preprint arXiv:1609.08144.

[5] Li X, Zhang J, Zong C. 2016. Towards Zero Unknown Word in Neural Machine Translation[C]// Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence. 2016: 2852-2858.

[6] Zhang J, Zong C. Exploiting Source-side Monolingual Data in Neural Machine Translation[C]// Conference on Empirical Methods in Natural Language Processing. 2016: 1535-1545.

[7] Zhang J, Zong C. Bridging Neural Machine Translation and Bilingual Dictionaries. 2016, arXiv preprint arXiv:1610.07272.

[8] Brown P, Della Pietra S, Della Pietra V, et al. The Mathematics of Machine Translation: Parameter Estimation[J]. Computational Linguistics, 1993, 19(2): 263-311.

[9] Koehn, Philipp, Hoang, et al. Moses: Open Source Toolkit for Statistical Machine Translation[C]// ACL 2007, Proceedings of the, Meeting of the Association for Computational Linguistics, June 23-30, 2007, Prague, Czech Republic. DBLP, 2007: 177-180.

[10] Chiang D. Hierarchical Phrase-Based Translation[J]. Computational Linguistics, 2007, 33(2): 201-228.

[11] Galley M, Graehl J, Knight K, et al. Scalable Inference and Training of Context-rich Syntactic Translation Models[C]// ACL 2006, International Conference on Computational Linguistics and, Meeting of the Association for Computational Linguistics, Proceedings of the Conference, Sydney, Australia, 17-21 July. DBLP, 2006: 961-968.

[12] Liu Y, Liu Q, Lin S. Tree-to-string Alignment Template for Statistical Machine Translation[C]// ACL 2006, International Conference on Computational Linguistics and, Meeting of the Association for Computational Linguistics, Proceedings of the Conference, Sydney, Australia, 17-21 July. DBLP, 2006: 609-616.

[13] Zhang M, Jiang H, Aw A T, et al. A Tree Sequence Alignment-based Tree-to-Tree Translation Model[C]// ACL 2008, Proceedings of the, Meeting of the Association for Computational Linguistics, June 15-20, 2008, Columbus, Ohio, USA. DBLP, 2008: 559-567.

[14] Bengio Y, Schwenk H, Senécal J, et al. Neural Probabilistic Language Models[J]. Journal of Machine

doi:10.3772/j.issn.2095-915x.2017.03.004

Learning Research, 2001, 3(6): 1137-1155.

[15] Vaswani A, Zhao Y, Fossum V, Chiang D. Decoding with Large-scale Neural Language Models Improves Translation[C]// The 2013 Conference on Empirical Methods in Natural Language Processing. 2013: 1387-1392.

[16] Zhang J, Liu S, Li M, et al. Beyond Word-based Language Model in Statistical Machine Translation[J]. arxiv preprint arxiv.org/abs/1502.01446, 2015(2).

[17] Och F J. Minimum Error Rate Training in Statistical Machine Translation[C]// Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2003: 160-167.

[18] Xiong D, Liu Q, Lin S. Maximum Entropy based Phrase Reordering Model for Statistical Machine Translation[J]. In Proc. of COLING-ACL, 2006:521-528.

[19] Koehn P. Statistical Significance Tests for Machine Translation Evaluation[C]// Conference on Empirical Methods in Natural Language Processing, EMNLP 2004, A Meeting of Sigdat, A Special Interest Group of the Acl, Held in Conjunction with ACL 2004, 25-26 July 2004, Barcelona, Spain. DBLP, 2004: 388-395.