ISSN 1000-0054 清华大学学报（自然科学版） 2014 年 第 54 卷 第 12 期 4/20
CN 11-2223/N J Tsinghua Univ (Sci & Technol), 2014, Vol. 54, No. 12 1529-1533

层次短语翻译的神经网络调序模型

李 鹏， 刘 洋， 孙茂松

（清华大学 计算机科学与技术系，智能技术与系统国家重点实验室，清华信息科学与技术国家实验室(筹)，北京 100084）

摘 要：调序歧义是层次短语翻译模型面临的主要挑战之一，但在该类模型中使用的上下文信息非常有限，制约了该类模型处理调序歧义的能力。为了更充分地利用上下文信息，提出了一种面向层次短语翻译模型的神经网络调序模型。该模型将调序看作分类问题，首先使用递归自动编码器为任意长度的字符串计算向量表示，然后使用这些向量表示作为分类特征，用于预测不同调序方式的概率，最后将这些概率作为新的特征加入翻译模型中进行翻译。实验结果显示：在中—英翻译任务上，该模型相比基线系统获得了 0.3～0.8 的 BLEU 值提升，具有更好的调序能力。

关键词：计算机科学与技术；神经网络；调序模型；递归自动编码器；层次短语翻译模型

中图分类号：TP 391.2 **文献标志码**：A

文章编号：1000-0054(2014)12-1529-05

Neural reordering model for hierarchical phrase-based translations

LI Peng, LIU Yang, SUN Maosong

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

Abstract: The reordering ambiguity is one of the major challenges for hierarchical phrase-based translation models. These models only consider limited contexts so that their ability is reduced to resolve reordering ambiguities. More contexts were introduced into these models using a neural reordering model for hierarchical phrase-based translations. Reordering is treated as a classification problem in this model. The vector-space representations are computed for phrases using recursive auto-encoders. These representations are then used as features to predict the probabilities of various reorderings. Finally, these probabilities are used as new features for the decoding. Tests show that this model improves the BLEU score by 0.3—0.8 over the baselines for Chinese-English translation, which indicates that this model gives better reordering than the baselines.

Key words: computer science and technology; neural network; reordering model; recursive auto-encoders; hierarchical phrase-based translation

层次短语翻译模型[1-2]是一类被广泛使用的统计机器翻译模型。该类模型基于同步上下文无关文法进行翻译。虽然上下文信息对于缓解调序歧义和提高模型的调序能力具有重要作用，但受制于同步上下文无关文法的约束，该类模型中只使用了有限的上下文信息，需要引入更丰富的上下文信息以提高模型调序能力。

将调序看作分类问题并通过分类特征引入更多上下文信息是一种行之有效的方法。文[3]为每一种可能的源语言串构建最大熵分类器，用于规则选择。文[4]在此基础上，按照源语言串是否包含变量和变量相对于其他字符串的位置对源语言串划分模式，并为每一种模式建立最大熵分类器，用于预测对应的目标语言串模式，有效减少了分类器数量，获得了翻译效果的提升。文[5-8]等验证了该类方法在其他基于同步上下文无关文法的翻译模型上的有效性。

但在以上方法中如何选取特征仍然是一个挑战。对于任意给定的字符串，哪些字、词或词组对于调序问题而言是更好的特征并不是显而易见的，因而只能根据经验进行人工选择。已有工作中通常采用边界词作为特征[3-4,6]，只能利用短语内的部分上下文信息。

深度学习技术的发展为以上问题的解决提供了新的可能性。在深度学习技术中，每个词被表示成一个实值向量，称为词向量[9-11]；通过使用递归自动编码器(recursive autoencoder)[11]，可由词向量出发，为任意长度的短语计算向量表示，这些向量表

收稿日期：2014-09-22

基金项目：国家“八六三”高技术项目（2012AA011102）；
国家自然科学基金重点项目（61331013）；
国家科技支撑计划项目（2014BAK101303）

作者简介：李鹏(1987—)，男(汉)，吉林，博士研究生。

通信作者：刘洋，副教授，E-mail：liuyang2011@tsinghua.edu.cn

示则可作为分类特征用于构建分类器,从而更加有效地利用上下文信息,为解决层次短语翻译模型的调序问题提供了一个新的方向。

本文提出了一种基于神经网络的层次短语翻译调序模型。与文[4]类似,本文为翻译规则中的源语言字符串划分模式,为每一种模式的源语言串建立相应的神经网络分类器,用于预测目标语言串的模式。每一个分类器由若干递归自动编码器和1个分类层构成。递归自动编码器用于计算规则中变量所对应的子串和源语言串中连续的词的向量表示,这些向量表示作为特征输入分类层用于测序目标语言串模式,进而克服了上文所述方法不能充分利用整个短语信息的缺点。实验显示,该模型获得了比基线系统更好的调序能力和翻译效果。

1 层次短语翻译模型

层次短语翻译模型[1-2]是一种被广泛使用的、基于上下文无关文法的翻译模型。该模型中使用的每条规则由1个左端项和2个右端项构成,表示2个右端项可由左端项同步生成,以下为一条规则示例:

$$X \rightarrow \langle X_1 \text{ 的 } X_2 ; X_1 X_2 \rangle .$$

其中左端项为非终结符(即变量)"X",两个右端项分别为"X_1的 X_2"和"$X_1 X_2$"。该规则表示"X_1的 X_2"和"$X_1 X_2$"可由"X"同步生成,而具有相同下标的非终结符则可进一步被同步替换为其他字符串。例如上述规则中的2个 X_1 被同步替换为"漂亮"和"beautiful"的过程及结果为

$$\begin{array}{l} X \rightarrow \langle X_1 \text{ 的 } X_2 ; X_1\ X_2 \rangle \\ \underline{+\ X \rightarrow \langle \text{漂亮} ; \text{beautiful} \rangle} \\ X \rightarrow \langle \text{漂亮的 } X_2 ; \text{beautiful } X_2 \rangle \end{array}$$

然而,由于具有相同下标的非终结符可被同步替换成任意字符串,在翻译过程中可能会使用不恰当的规则。如上例中若使用规则

$$X \rightarrow \langle X_1 \text{ 的 } X_2 ; X_2 \text{ of } X_1 \rangle$$

进行翻译,则会产生不恰当的调序和译文。而如果能在替换过程中参考更多的上下文信息,则可以改善这种情况。

2 神经网络调序模型

2.1 规则模式预测

参照文[4],本文同样将翻译规则按照是否含有词和非终结符相对于词的位置进行分类。表1列出了只含有1个非终结符的规则的源语言串和目标语言串的模式,其中"X"表示非终结符,"F"表示源语言字符串,"E"表示目标语言字符串(层次短语翻译模型为减小解码过程中的歧义,要求源语言串至少包含1个词,因而不存在模式为"X"的源语言串)。分类器用于在给定源语言串模式的情况下,预测目标语言串的模式。记一条翻译规则的源语言串为 α,模式为 T_α,目标语言串为 β,模式为 T_β,上下文信息为 c,则分类器用于计算概率 $P(T_\beta | T_\alpha, \alpha, \beta, c)$。

表1 只含有1个非终结符的规则的源语言和目标语言串模式

源语言串模式	目标语言串模式
$X\ F$	X
$F\ X$	$X\ E$
$F\ X\ F$	$E\ X$
	$E\ X\ E$

2.2 递归自动编码器

深度学习方法中通常将词表示成实值向量即词向量[9-11],而递归自动编码器[11]则可以为任意长度的短语计算向量表示。设给定2个词 ω_1 和 ω_2,它们的向量表示分别为 c_1 和 c_2,则可由式(1)计算实值向量 p 作为短语"$\omega_1\ \omega_2$"的表示:

$$p = f^{(1)}(W^{(1)}[c_1 ; c_2] + b^{(1)}). \tag{1}$$

其中:$W^{(1)}$ 为权重矩阵;$b^{(1)}$ 为偏置向量;$[c_1 ; c_2]$ 表示将 c_1 和 c_2 拼接成列向量;$f^{(1)}(\cdot)$ 为非性线函数,本文采用 $\tanh(\cdot)$。

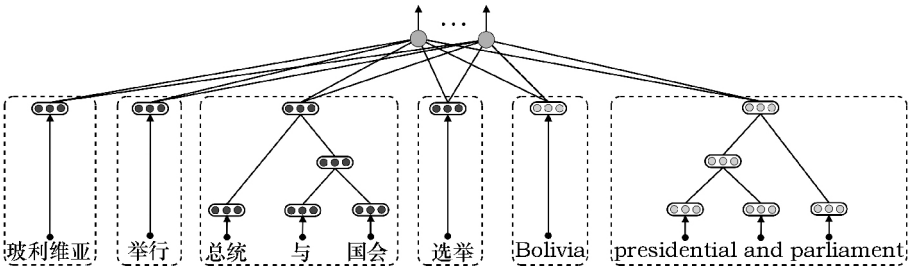
定义 c'_1 和 c'_2 分别为 c_1 和 c_2 的重构向量,可由式(2)计算得到:

$$[c'_1 ; c'_2] = f^{(2)}(W^{(2)} p + b^{(2)}). \tag{2}$$

其中:$W^{(2)}$ 为权重矩阵;$b^{(2)}$ 为偏置向量;$f^{(2)}(\cdot)$ 为非性线函数,本文采用 $\tanh(\cdot)$。如果 c'_1 与 c_1、c'_2 与 c_2 足够接近(即重构错误足够小),则可认为 p 几乎包含了 c_1 和 c_2 的全部信息,是一个好的表示。本文使用式(3)衡量 c'_1 与 c_1、c'_2 与 c_2 的接近程度:

$$\frac{1}{2}(\| c'_1 - c_1 \|^2 + \| c'_2 - c_2 \|^2). \tag{3}$$

对于由多个词构成的短语,则可以递归进行上述计算,最终得到整个短语的向量表示。本文参照文[11]采用贪婪算法确定向量的合并顺序,即每次总是优先合并重构错误最小的2个相邻向量,并用新向量替换这2个向量,接下来重复以上步聚直至最终只剩1个向量为止。关于递归自动编码器的更多细节参见文[11]。



$X \rightarrow \langle X_1$举行$X_2$选举；$X_1$ holds X_2 elections$\rangle$
$X \rightarrow \langle$玻利维亚；Bolivia$\rangle$
$+\ X \rightarrow \langle$总统与国会；presidential and parliament$\rangle$
$X \rightarrow \langle$玻利维亚举行总统与国会选举；Bolivia holds presidential and parliament elections$\rangle$

图 1　对应于“*X F X F*”模式源语言串的神经网络调序分类器示例

2.3　神经网络分类器

本文为每一种源语言串模式构造 1 个单独的神经网络分类器，用于预测目标语言串的模式。图 1 展示了对应于“$X\ F\ X\ F$”模式源语言串的神经网络分类器，分类器由一个分类层和 6 个递归自动编码器构成。6 个递归自动编码器中 2 个用于为 2 个“F”对应的短语（即“举行”和“选举”）计算向量表示，2 个用于为 2 个“X”对应的源语言短语（即“玻利维亚”和“总统 与 国会”）计算向量表示，2 个用于为 2 个“X”对应的目标语言短语（即“Bolivia”和“presidential and parliament”）计算向量表示。计算得到的向量表示将作为分类层的输入，由式(4)计算得到不同模式目标语言串的概率。

$$\mathrm{softmax}(W^{\circ}_{T_\alpha} p^{\circ} + b^{\circ}_{T_\alpha}). \tag{4}$$

其中：$W^{\circ}_{T_\alpha}$ 和 $b^{\circ}_{T_\alpha}$ 分别为对应于源语言串模式 T_α 的权重矩阵和偏置向量，P° 为递归自动编码器计算所得向量拼接而成的输入向量（在图 1 中由 6 个递归自动编码器计算所得的向量拼接而成）。这里需要说明的是，分类层输出端的个数因源语言串模式所可能对应的目标语言串模式数目而有所不同，例如对于只包含 1 个 X 的源语言串，有 4 种可能的目标语言串模式（见表 1），故输出端个数为 4，而对于包含 2 个 X 的源语言串，有 14 种可能的目标语言串模式，故输出端个数为 14。相应地，$W^{\circ}_{T_\alpha}$ 和 $b^{\circ}_{T_\alpha}$ 的维度与输出端个数相匹配。

3　调序模型训练

在调序模型训练中，需要考虑以下 2 种因素：

1）重构错误：度量递归自动编码器计算得到的向量表示是否能够很好地表示输入向量。本文将其定义为所有结点的平均重构错误。

2）分类错误：度量分类器的分类性能。本文定义为平均交叉熵（cross-entropy error）。

本文中将重构错误和分类错误的线性加权作为神经网络的训练目标函数。参照文[11]，本文采用 L-BFGS[12] 算法学习神经网络参数，并使用结构上的反向传播算法（backpropagation through structures）[13] 计算梯度。

本文在抽取层次短语规则时记录了源语言字符串中的 X 所实际对应的字符串，并将规则及 X 实际对应的字符串作为神经网络的训练样例。由于对齐错误等的影响，在抽取的规则中不可避免地将包含一定噪声，进而在训练样例中也将包含一定噪声，可能对神经网络训练产生不利影响，本文暂不处理这一潜在问题。

4　实　验

4.1　实验设置

本文在中—英翻译任务上对所提出的调序模型进行了评估。训练语料由 123 万双语句对组成，约含中文词 0.32 亿个，英文词 0.35 亿个。实验中采用 4 元语言模型，训练语料为 GIGAWORD Xinhua 部分（LDC2011T07），约包含 3.986 亿个词，使用工具 SRILM[14] 训练得到。使用 NIST 2006 机器翻译评测中—英翻译数据集作为开发集，NIST 2003-2005 机器翻译评测中—英翻译数据集作为测试集（以下使用 MT03～05 指代相应年份的数据集）。本文使用大小写不敏感的 BLEU 值[15] 作为评测指标。

实验中使用了 2 个基线系统：传统的层次短语翻译模型[1-2] 和最大熵调序模型[4]。由于本文的模型中只使用词汇特征，为公平比较，在最大熵调序模

型中亦只使用词汇特征,而未使用文[4]中的词性特征。翻译解码时最大熵分类器的输出和神经网络的输出都作为新的特征加入模型中,并用 MERT 算法[16]优化特征权重。

4.2 实验结果

翻译实验结果见表 2。可以看到,最大熵调序模型和神经网络调序模型的翻译效果均好于传统的层次短语翻译模型,印证了利用上下文信息缓解层次短语翻译模型调序歧义的有效性。神经网络调序模型在 3 个测试集上的翻译效果均好于最大熵调序模型,印证了神经网络调序模型的有效性。神经网络调序模型在开发集 MT06 上的翻译效果略低于最大熵调序模型,本文认为这与 MERT 算法的多轮迭代调参过程有关,并不与测试集上的结果冲突。

表 2 神经网络调序模型与基线系统 BLEU 比较

模型	MT06	MT03	MT04	MT05
层次短语	31.64	33.18	33.98	31.77
最大熵	31.93	33.52	33.95	31.81
神经网络	31.81	33.81	34.50	32.61

为了对比最大熵分类器和神经网络分类器的分类效果,本文为每一种源语言串模式对应的分类器随机抽取了测试集,并要求每一种可能的目标语言串模式拥有 200 个测试样例。分类准确率测试结果见表 3,可以看到,除最后一个模式外,神经网络分类器的分类效果都明显好于最大熵分类器的,这一结果与翻译实验的结果是相吻合的。对于最后一个模式,其对应的神经网络分类器参数最多(输入向量和输出端个数均最大),但其训练样例数只占总训练样例数的 2.12%,训练不充分,这是造成该类别上神经网络分类器分类效果不及最大熵分类器的一个可能原因。另外从表 3 中还可以看到,2 种分类器在只包含 1 个 X 的源语言串模式上的分类效果明显优于包含 2 个 X 的源语言串模式,这是因为前者只对应 4 种可能的目标语言串模式,而后者对应 14 种,且其中多数类别只拥有少量的训练样例。以"$X\ F\ X$"模式为例,在训练语料中共抽取出具有该源语言串模式的规则 33 638 939 条,其中 91.40%的目标语言串模式为"$X\ E\ X$",其他目标语言串模式的训练样例是不足的,对分类器的训练产生了严重的影响,影响了分类效果。

表 3 分类准确率

源语言串模式	源语言串示例	训练样例数占比/%	分类准确率/%	
			最大熵	神经网络
$X\ F$	X 国会	20.36	69.67	77.00
$F\ X$	总统 与 X	20.40	64.67	81.83
$F\ X\ F$	总统 X 国会	15.88	74.83	81.50
$X\ F\ X$	X_1 举行 X_2	20.43	46.50	58.89
$F\ X\ F\ X$	波利维亚 X_1 举行 X_2	10.41	48.96	53.86
$X\ F\ X\ F$	X_1 举行 X_2 选举	10.40	48.14	52.14
$F\ X\ F\ X\ F$	波利维亚 X_1 举行 X_2 选举	2.12	36.36	33.43

5 结 论

本文提出了一种面向层次短语模型的神经网络调序模型,该模型可以克服已有方法需要手工构造特征、无法充分利用上下文中的全部信息的弊端,获得了比传统层次短语翻译模型和最大熵调序模型更好的翻译效果。

训练数据类别分布不平衡导致的训练数据不足严重影响了分类器的准确率,下一步将提高神经网络分类器在不平衡数据下的分类性能。

参考文献 (References)

[1] Chiang D. A hierarchical phrase-based model for statistical machine translation [C]// Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 2005: 263-270.

[2] Chiang D. Hierarchical phrase-based translation [J]. *Computational Linguistics*, 2007, **33**(2): 201-228.

[3] He Z, Liu Q, Lin S. Improving statistical machine translation using lexicalized rule selection [C]// Proceedings of the 22nd International Conference on Computational Linguistics. Manchester, UK: Coling 2008 Organizing Committee, 2008: 321-328.

[4] He Z, Meng Y, Yu H. Maximum entropy based phrase reordering for hierarchical phrase-based translation [C]// Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. Massachusetts, USA: Association for Computational Linguistics, 2010: 555-563.

[5] Zens R, Ney H. Discriminative reordering models for statistical machine translation [C]// Proceedings on the Workshop on Statistical Machine Translation. New York, USA: Association for Computational Linguistics, 2006: 55-63.

[6] Xiong D, Liu Q, Lin S. Maximum entropy based phrase reordering model for statistical machine translation [C]// Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics. Sydney, Australia: Association for Computational Linguistics, 2006: 521-528.

[7] Xiong D, Zhang M, Aw A, et al. Linguistically annotated BTG for statistical machine translation [C]// Proceedings of the 22nd International Conference on Computational Linguistics. Manchester, UK: Association for Computational Linguistics, 2008: 1009-1016.

[8] Liu Q, He Z, Liu Y, et al. Maximum entropy based rule selection model for syntax-based statistical machine translation [C]// Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing. Honolulu, Hawaii, USA: Association for Computational Linguistics, 2008: 89-97.

[9] Bengio Y, Ducharme R, Vincent P, et al. A neural probabilistic language model [J]. *Journal of Machine Learning Research*, 2003, **3**: 1137-1155.

[10] Collobert R, Weston J, Bottou L, et al. Natural language processing (almost) from scratch [J]. *Journal of Machine Learning Research*, 2011, **12**: 2493-2537.

[11] Socher R, Pennington J, Huang E H, et al. Semi-supervised recursive autoencoders for predicting sentiment distributions [C]// Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. Edinburgh, Scotland, UK: Association for Computational Linguistics, 2011: 151-161.

[12] Liu D C, Nocedal J. On the limited memory BFGS method for large scale optimization [J]. *Mathematical Programming*, 1989, **45**(1-3): 503-528.

[13] Goller C, Kuchler A. Learning task-dependent distributed representations by backpropagation through structure [C]// Proceedings of the International Conference on Neural Networks (ICNN'96). Washington DC, USA: IEEE, 1996: 347-352.

[14] Stolcke A. SRILM-an extensible language modeling toolkit [C]// Proceedings of the International Conference on Spoken Language Processing. Denver, Colorado, USA: ISCA, 2002: 901-904.

[15] Papineni K, Roukos S, Ward T, et al. BLEU: A method for automatic evaluation of machine translation [C]// Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, 2002: 311-318.

[16] Och F J. Minimum error rate training in statistical machine translation [C]// Proceedings of the 41st Annual Meeting on Association for Computational Linguistics. Sapporo, Japan: Association for Computational Linguistics, 2003: 160-167.