基于非连续短语的统计翻译模型

张大鲲，张玮，董静

（中国科学院软件研究所 中文信息处理中心，北京 100080）

摘 要： 本文给出了基于非连续短语的统计翻译方法，翻译的基本单元从连续短语扩展到带有间隔的非连续短语，这种方法可以更好地解决句子中词语翻译时的上下文依赖问题。在形式上，非连续短语方法和层次型短语方法相似，不同之处在于前者只允许包含一个非终结符的短语（如 turn ◆ on），因此，非连续短语方法抽取的短语数量较少，搜索效率也得到提高。初步实验表明，非连续短语方法的翻译结果是令人满意的。

关键词： 非连续短语；统计机器翻译；短语模型

A Non-contiguous Phrase-based Model for Statistical Machine Translation

Zhang Dakun, Zhang Wei, Dong Jing

(Chinese Information Processing Center, Institute of Software CAS, Beijing 100080)

Abstract: We described a statistical model based on non-contiguous phrases for machine translation. The units of translation extend from contiguous phrases to phrases with intervals. This model can better address the phenomena of context dependent translation. Formally, the non-contiguous phrase-based model is similar to hierarchical phrase-based model, however, the main difference is that we allow only one non-terminal (such as turn ◆ on) in the phrases. Therefore, the numbers of phrases are less, and the efficiency of the search procedure is improved. Early results show that the translation of non-contiguous phrase-based model is satisfactory.

Keywords: non-contiguous phrase; statistical machine translation; phrase-based model

1 介绍

基于短语的统计翻译模型近年来逐渐取代了基于词的模型，成为统计机器翻译方法的主流。翻译的基本单元从词过渡到短语，可以更好地解决词在翻译时对上下文的依赖问题。基于短语的方法，使得临近的词串在翻译时仍然作为一个整体进行处理，因此词之间的重排序问题变成了短语内部的问题，不再需要翻译模型单独处理，所以翻译质量有了明显提高。虽然这里的短语可以是任意词串，不要求必须是符合语法习惯的短语，但是却要求是连续的词串，即该方法可以称为“基于连续短语”的翻译模型。

非连续短语方法在信息检索[1]中取得了比较好的效果，Simard 等人首先将非连续短语方法用于统计机器翻译模型[2]，翻译质量也得到了一定程度的改善，然而 Simard 所使用的非连续短语模型，要求短语内部的间隔（gap）部分，必须是严格的词，因此非连续短语的长度是固定的。比如：短语 turn the light on 和 turn the left light on 在

作者简介：张大鲲（1980-），男，博士研究生 E-mail: dakun04@iscas.cn

利用 Simard 的模型表示时，得到 turn ◇ ◇ on 和 turn ◇ ◇ ◇ on（◇表示任意 1 个词）两个不同的短语，本文的方法将这种短语扩展为一种短语 turn ◆ on（◆表示任意 1 个或多个词），增强了模型的适应能力，同时调整了解码部分的设计以适应非连续短语的翻译。

本文的思想来自于 Simard[2]和 Chiang 提出的基于层次型短语的翻译模型[3]。和层次型短语模型相比，非连续短语方法抽取的短语结构简单，数量大大减少，因此带来了计算上的时间和空间优势，这一点尤其在利用最小错误率方法[4]调整特征函数的权重时体现明显。初步实验数据表明，使用非连续短语方法的翻译质量也略有提高。非连续短语方法和 Och 提出的基于模版的翻译方法[5]相比，Och 的方法主要是将词到词类的一个泛化，本文的方法形式上借用了基于模版的方法，但是不包含词类的概念。

本文的其他部分安排如下：第 2 部分详细介绍非连续短语的定义和抽取方法，非连续短语从双语句对齐语料中训练获得，不需要语言学标注信息（如 Treebank）；第 3 部分介绍基于 log-linear 模型的非连续短语模型；解码器的部分在第 4 部分进行描述，需要利用概率上下文无关文法（Probabilistic CFG）对句子进行解析；第 5 部分给出初步的实验设计和相应结果。

2 非连续短语

使用非连续短语对基于短语的翻译模型进行扩展，目的是使模型具有更强的对语言建模的能力，最终提高翻译系统的翻译质量。比如：双语句对“请 开 灯”和“please turn the light on”，“他 向 她 做 鬼脸”和“he made a face to her”，如果能识别出句子中的短语“开 …”等价于短语“turn … on”，短语“向 … 做 鬼脸”等价于短语“made a face to …”，则可以提高模型对语言的模拟程度，得到更准确的翻译结果。

2.1 定义

在非连续短语方法中，区分基本短语和扩展短语的概念，基本短语即普通的连续词串，不包含作为占位符的非终结符[1]；扩展短语即包含占位符的非连续短语。

沿用 Zens[6]对双语短语的定义，作为基本短语的定义：一个双语短语对内部的词，只和短语对内的词存在对齐关系，不和任意一个短语对外的词存在对齐。这样得到的短语称为基本短语。对于句子对$\left(f_1^J;e_1^I\right)$和相应的对齐矩阵 A，基本短语 BP 的形式化描述如公式 1：

$$BP(f_1^J,e_1^I,A)=\left\{\left(f_j^{j+m},e_i^{i+n}\right):\forall(i',j')\in A:j\le j'\le j+m\leftrightarrow i\le i'\le i+n\right\} \quad (1)$$

如果$\langle f,e\rangle$和$\langle\gamma,\alpha\rangle$是基本短语对，且$f=f_1\gamma f_2$，$e=e_1\alpha e_2$，则$f_1$ ◆ f_2和e_1 ◆ e_2是一个扩展短语对（◆表示任意 1 个或多个词），即非连续短语对，图 1 是非连续短语的示例。

对 ◆ 感 兴趣 ||| are interested in ◆
对 ◆ 感 兴趣 ||| interested in ◆
很 ◆ 见到 ||| was very ◆ to meet
很 ◆ 见到 ||| it was very ◆ meeting
很 ◆ 见到 ||| it was very ◆ to meet

图 1 非连续短语示例

Fig.1 Samples of non-contiguous phrases

2.2 非连续短语抽取

翻译模型的基础是短语对列表，短语对的质量也直接决定着最终的翻译质量，因此如何从双语句对齐语料中，获得短语对列表是首先需要解决的问题。目前抽取短语的方法有 Och 提出的基于改进的词对齐抽取方法[7]，这种方法也是基于短语的翻译模型广泛采用的方法[3, 8]。此外，也有直接计算短语对列表和相应概率值的方法[9]，以及利用非负矩阵分解抽取短语的方法[10]。本文采用第一种方法，同时进行扩展。

Och 的短语抽取方法：首先利用 GIZA++对双语语料进行词对齐的双向训练（中—英，英—中），分别取两

[1]这里的占位符和非终结符用◆表示，指非连续短语的间隔部分，可以是任意的词或连续词串；终结符是指词。

次训练结果的交集和并集，再从交集出发，扩展每个对齐点的临近点，其上限是对齐结果的并集，如果符合基本短语的定义，则添加到短语列表中[7]。在抽取基本短语对之后，可以进行扩展短语的抽取，只需要去掉词串必须是连续的这一限制即可。

算法描述如下：给定一个句对，先抽取可能的连续短语，同时和已经抽取到的连续短语进行比较，如果符合非连续短语的定义，则作为扩展短语添加。非连续短语的抽取过程是一个动态程序过程。如图 2。

```
For 每一个双语句对
    For 新抽取的连续短语 A
        For 已经抽取的连续短语 B
        {
            If B 是 A 的子串，交换两个字符串；
            If A 是 B 的子串，抽取新的非连续短语；
        }
```

图 2 非连续短语的抽取算法

Fig.2 Algorithm for extraction of non-contiguous phrases

图 3 额外的短语抽取示例

Fig.3 Sample of extra method on phrase extraction

同样，为了保证解码阶段的效率，需要限制基本短语和扩展短语的长度，本文分别取 10 和 5（◆的长度为 1）；对非连续短语还有一个额外的约束，要求非连续短语对内的终结符（词）之间至少存在一个原始的对齐关系，保证所得到的短语对存在一定的词语关联。

除了利用 Och 方法抽取的基本短语外，本文采用了如下方法对短语对进行扩展：一个句子中除了已经对齐的短语之外，剩余的词串（连续的和非连续的），只要符合上述短语的定义和约束条件，仍然作为短语对添加。如图 3，通过词对齐结果可以容易地获得短语对“你 能”和“could you”，但句子的其余部分的对齐结果却不符合短语抽取的定义，我们的方法类似一个减法过程，句子的剩余部分“帮 我 一 把 吗”和“give me a hand for a second”也作为短语对添加。

3 模型

非连续短语翻译模型的基础是基于信源信道模型（Noisy-Channel Models）的翻译方法[12]。给定源语言（中文）句子 $f_1^J = f_1 \ldots f_j \ldots f_J$，翻译过程实际上就是寻找目标语言（英语）句子 $e_1^I = e_1 \ldots e_i \ldots e_I$，使得概率 $\Pr(e_1^I \mid f_1^J)$ 达到最大：

$$\hat{e} = \arg\max_{e} \left\{ \Pr(e_1^I \mid f_1^J) \right\} \tag{2}$$

其中，argmax 操作是一个搜索（解码）过程，即生成目标语言句子的过程。基于非连续短语的方法，同样采用 log-linear 模型：

$$\Pr(e_1^I \mid f_1^J) = \frac{1}{Z_{f_1^J}} \exp\left(\sum_{m=1}^{M} \lambda_m h_m \left(e_1^I, f_1^J \right) \right) \tag{3}$$

其中 h_m 表示特征函数，相应的 λ_m 表示特征函数对模型的贡献大小，即权重，控制模型的模拟程度，$Z_{f_1^J}$ 表示一个归一化的过程。本文中使用的特征函数包括：语言模型概率：$P_{LM}(e)$；短语的双向概率：$P(\gamma \mid \alpha)$ 和 $P(\alpha \mid \gamma)$；短语的词汇化概率：$P_w(\gamma \mid \alpha)$ 和 $P_w(\alpha \mid \gamma)$；短语惩罚概率：$\exp(1)$，使用的短语越多，惩罚越大；长度惩罚概率：$-|e|$，通常句子越短，惩罚越大；非连续短语的惩罚概率：$\exp(\lambda_g)$，非连续短语的层数越多，惩罚越大。

3.1 短语概率的计算

Och 等人在计算基本短语概率的时候，认为句子中的短语是均分的（uniform），短语对的概率为短语出现次数的相对频率：

$$\phi(\bar{f} \mid \bar{e}) = \frac{count(\bar{f}, \bar{e})}{\sum_j count(\bar{f}, \bar{e})} \tag{4}$$

这里的 count 值为整数，本文的方法将这个数值扩展为分数：利用短语在当前句子中出现的频率进行平滑，即先根据抽取得到的短语对句子进行切分，统计同一个句子在不同的切分情况下，每一个短语的出现次数，然后再利用公式 4 计算相对频率。如图 4，句子的短语分割总数为 5 次，其中短语“请”占了 2 次，因此 count(请)=2/5。

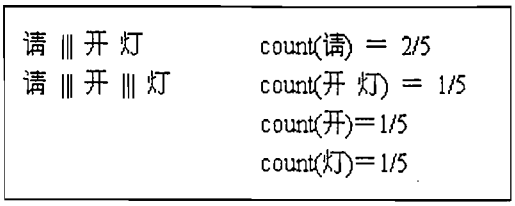

图 4 短语相对频率计算

Fig.4 Sample of estimating relative frequency of phrases

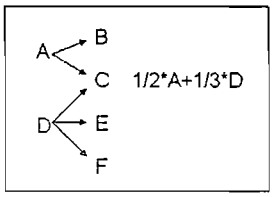

图 5 扩展短语的概率分配

Fig.5 Prob. Distribution of extended phrases

利用 Och 的方法抽取出的基本短语，有些并不能构成句子的任意一个划分的一部分，这种平滑方法还可以将这部分短语去掉，从而减少短语的数量。

扩展短语的概率计算类似信息检索中的 PageRank 算法（PageRank 有迭代的过程，这里没有），由于一个基本短语可以扩展出多个扩展短语，多个基本短语也可以扩展出同一个扩展短语，所以，扩展出的这些短语平均分得基本短语的概率值，由不同的基本短语所得的概率值需要进行累加，最后进行归一化。如图 5，基本短语 A 和 D，扩展短语 B、C、E、F，则扩展短语 C 的概率为 1/2*A+1/3*D。

4 解码

基于短语的翻译方法通常采用基于 Beam Search 的解码器[12]，对于一个给定句子先确定可能的翻译候选项（Translation Options），每一个翻译的中间状态称为一个假设（Hypotheses），初始状态即初始假设为没有词被翻译。从初始假设开始，逐步对待翻译的句子中可能的翻译候选项进行扩展，翻译那些还没有被翻译的部分；已经翻译的部分根据已经翻译的词的长度分别放置在相应的栈（Beam）中；翻译过程的扩展过程即从一个栈向另外一个栈的跳转过程；覆盖整个句子的最后一个栈中，概率最高的栈元素作为翻译结果输出。解码过程中，可以通过对栈的大小进行控制，达到控制算法效率的目的，用于平衡最终翻译质量和程序的运行开销。

Simard 使用的非连续短语[2]都是固定长度的，因此可以用 Beam Search 方法进行解码。本文的方法因为存在不定长度的非连续短语，所以解码过程有所不同：主要分为两步，先利用 CKY 句法分析器对待翻译的句子进行解析，找出概率最大的短语集；然后再将短语映射成英语句子，利用公式 3 计算总的概率。详细的解码算法描述如图 6 所示。

同样，解码过程需要进行剪枝，以保证算法效率。采用的剪枝策略有：（1）对每一个翻译候选项，只选取概率最高的前 20 个短语对翻译。（2）利用 CKY 对句子进行解析的过程中，每个中间状态只保留结果最好的前 100 个，类似 Beam Search 中的栈的容量。

```
For 每一个中文句子
    生成翻译候选项，包括非连续短语
    使用 CKY 句法分析器对句子进行解析（实现 Probabilistic CFG）
    加入语言模型特征和长度惩罚等特征
    输出 1 个或多个最好结果
```

图 6 解码算法

Fig.6 Algorithm for non-contiguous phrases decoder

5 实验

实验的目的是验证非连续短语方法对最终翻译结果的影响，因此选择基于短语的模型（pharaoh[12]）以及基于层次型短语的模型[3]作为对比。实验语料统计数据如表 1：

表 1 语料数据统计

Tab.1 Data sets

		中文	英语
训练集	句子数	50823	
	总词数	340587	365546
	词汇数	10168	8310
测试集	句子数	1000	
	总词数	9123	9291

5.1 评价指标

统计机器翻译评测中被普遍接受的评价指标有 BLEU 值、mWER 值等，我们的实验结果可以看出，非连续短语的方法在大部分准则中都有提高。下面简单介绍这些判定准则：

- mWER（多参考单词错误率）：对于每一个测试句子，和多个参考译文进行比较，计算与最相似句子的编辑距离（最少的替换、插入和删除次数）。
- mPER（与位置无关的 mWER）：mWER 的一个缺点是它要求词序必须匹配。mPER 则不考虑词序，把句子看作是一袋子词（bag-of-words），然后再计算与多个参考译文中最相似句子的编辑距离。
- BLEU 值和 NIST 值：这个评价标准计算和参考译文相比的 n 元语法（n-gram）的精确度（几何平均数和算术平均数），通常 n 取值为 4。
- GTM（General Text Matcher）：GTM 值测量的是句子之间的相似程度。

NIST、BLEU 和 GTM 都是对准确度的测量，数值越高表示翻译质量越好，其余指标则相反。

5.2 基线系统

基线系统（Baseline）选择 Koehn 在 2004 实现的第一个基于短语的统计机器翻译系统 pharaoh[12]，使用的默认特征包括：语言模型特征，$P(\gamma \mid \alpha)$ 和 $P(\alpha \mid \gamma)$，词汇化特征（双向），变型模型，长度惩罚和短语惩罚。同一短语的翻译取前 20 个最好的，栈容量为 100，不限制变型参数，语言模型权重为 0.5，长度惩罚为 0.2，短语长度为 7。

5.3 初步实验结果

本文中同时实现了基于层次型短语的翻译方法和基于非连续短语的翻译方法，因为利用最小错误率[4]调整参数的方法是一个非常耗时的过程，所以并没有给出得到最优翻译结果的参数集 λ_g 利用经验参数获得的初步结果如表 2（取前 20 个短语翻译，栈容量为 100， 为 0.1，语言模型权重 0.15，长度惩罚为 0.32）：

表 2 实验结果

Tab.2 Experiment results

	NIST	BLEU	GTM	mWER	mPER
pharaoh	6.9075	0.2940	0.6245	0.5702	0.4713
层次型	6.9373	0.2960	0.6262	0.5588	0.4585
非连续短语	6.9476	0.2965	0.6273	0.5575	0.4569

6 结论

非连续短语模型可以通过对双语语料的训练自动学习语言知识，不需要语言学的标注信息，同时对基于连续短语的方法进行扩展，能够更好地解决翻译过程中存在的上下文依赖问题。

层次型短语模型和非连续短语模型都在一定程度上提高了基于短语模型的翻译质量，相比之下，非连续短语方法具有短语数量少，短语类型简单（只有一种形式的非连续短语），处理效率占优等特点，同时能够保证最终得到令人满意的翻译质量。

采用非连续短语方法可以更好地模拟语言知识，但是仍然有大量的短语不符合语法的习惯，或者对最终的翻

译质量贡献较小，因此短语的过滤问题是一个需要进一步研究的问题，即如何通过更少的高质量短语对语言现象进行更好的模拟。此外，提高目前解码系统的效率也是一个需要解决的问题。基于短语的模型虽然在一定程度上解决了翻译过程中的上下文依赖问题，但是并不能从根本上解决复杂句子的翻译，因此向基于句法的翻译方法进行过渡，也是下一步的工作方向。

参考文献：

[1] Antoine Doucet and Helena Ahonen-Myka. Non-Contiguous Word Sequences for Information Retrieval. in Proceedings of the 42nd annual meeting of the Association for Computational Lingustics, Workshop on Multiword Expressions: Integrating Processing. 2004.

[2] Michel Simard, et al. *Translating with non-contiguous phrases.* in *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP).* 2005. Vancouver.

[3] David Chiang. *A Hierarchical Phrase-Based Model for Statistical Machine Translation.* in *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics.* 2005. Ann Arbor.

[4] Franz Josef Och. *Minimum Error Rate Training in Statistical Machine Translation.* in *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics.* 2003.

[5] Franz Josef Och, *Statistical Machine Translation: From Single-Word Models to Alignment Templates.* 2002.

[6] Richard Zens, Franz Josef Och, and Hermann Ney, *Phrase-Based Statistical Machine Translation*, in *25th Annual German Conference on Artificial Intelligence (KI2002),* 2002, Springer Verlag. p. 18-32.

[7] F. J. Och and H. Ney, *A Systematic Comparison of Various Statistical Alignment Models.* Computational Linguistics. 2003. 29(1): p. 19-51.

[8] Philipp Koehn, Franz Josef Och, and Daniel Marcu. *Statistical Phrase-Based Translation.* in *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL (HLT-NAACL 2003).* 2003. Edmonton, Alberta, Canada.

[9] Daniel Marcu and William Wong. *A Phrase-Based, Joint Probability Model for Statistical Machine Translation.* in *Proceedings of the Conference on Empirical Methods in Natural Language Processing.* 2002. Philadelphia, PA, USA.

[10] Cyril Goutte, Kenji Yamada, and E. Gaussier. *Aligning words using matrix factorisation.* in *Proc. ACL2004.* 2004.

[11] 刘群, *统计机器翻译综述.* 中文信息学报, 2003. 17(4): p. 1-12.

[12] Philipp Koehn. *Pharaoh: A Beam Search Decoder for Phrase-Based Statistical Machine Translation Models.* in *Proceedings of the Sixth Conference of the Association for Machine Translation in the Americas.* 2004.

[13] 侯宏旭, 刘群, 张玉洁, *2005 年度 863 机器翻译评测方法研究与实施.* 中文信息学报, 2006. 20(z1)