基于 HMM 的短语翻译对抽取方法*

左云存　宗成庆

中国科学院自动化研究所模式识别国家重点实验室 北京 100080

E-mail: {yczuo, cqzong}@nlpr.ia.ac.cn

摘　要：在基于语料库的统计翻译方法中,基于短语的统计翻译与基于单个词的统计翻译相比可以更好地处理句中词语之间的关系，从而有效地提高机器翻译系统的性能。在基于短语的统计翻译方法中,一种重要的策略是把短语翻译对作为一种知识加入到翻译系统中，因此,整个系统的性能与使用的短语翻译对的质量具有很大的关系。本文在基于HMM词对齐方法的基础上，提出了一种从双语语料中自动抽取短语翻译对的方法，这种方法根据词语对齐时出现的不同情况作不同的处理，提高了短语翻译对抽取的效果。

关键词：HMM；词对齐；短语翻译对；机器翻译

Phrase Translation Extraction Based on HMM

Zuo Yuncun, Zong Chengqing

National Laboratory of Pattern Recognition, Institute of Automation,

Chinese Academy of Sciences, Beijing 100080

E-mail: {yczuo, cqzong}@nlpr.ia.ac.cn

Abstract: In corpus-based statistical machine translation methods, phrase-based models are effective in improving translation quality as they can deal with the relationship between words in sentences better than word-based models. One approach of phrase-based translation incorporates phrase translations as knowledge sources into systems, and the systems' performance greatly depends on the quality of phrase knowledge. In this paper, we describe a new approach of phrase translation extraction based on HMM-based word alignment method. The experiment result proved that this approach is effective in phrase translation extraction from bilingual corpus.

Keywords: HMM, word alignment, phrase translat ion extraction, machine translation

1　前言

机器翻译的任务是把源语言句子($s = s_1 .. s_I$)翻译成目标语言句子($t = t_1 ... t_J$)。在基于信

* 本文受国家自然科学基金项目（项目号: 60175012，60121302，60272041）和中国科学院海外杰出学者基金项目(项目号: 2003-1-1)的资助。

源信道模型的翻译方法中[1]，假定源语言句子是由一个目标语言句子通过一个噪声信道生成的，统计翻译的任务就是从源语言还原目标语言。其过程可以描述为寻找 $\hat{t}$ 使得 $\hat{t} = \arg\max p(t \mid s)$，根据贝叶斯公式得 $\hat{t} = \arg\max p(t)\, p(s \mid t)$，其中 $p(t)$ 为语言模型，$p(s \mid t)$ 为翻译模型。[1]提出的 IBM 模型成为统计翻译事实上的标准。后来，不少人在 IBM 模型上进行了改进。

IBM 的几个翻译模型都是以单个词作为翻译的基本单元。[2]说明了以单个词作为翻译基本单元的不足，同时，提出了获取短语翻译对的动机。[3]详细说明了在统计翻译中加入短语翻译对知识的原因，并且证明了在统计翻译中加入短语翻译对对提高系统的性能是有效的。我们的工作重点放在从双语语料自动抽取短语翻译对的方法研究上。

针对从双语语料中自动抽取短语翻译对的问题，[4]提出了基于 IBM 模型的短语翻译对抽取方法；[6]提出了四种不同的短语翻译对抽取方法，并且把各种不同方法抽取的翻译对加在一个统计翻译系统中。试验证明，每加入一种短语翻译对知识，翻译系统的效果都会有不同程度的提高，其中基于 HMM 词对齐的方法抽取的短语翻译对对整个系统效果提高最大，尽管如此，这种短语翻译对抽取方法本身存在几个明显不足的地方。本文在分析这种方法不足之处的基础上，提出了一种更加有效的基于 HMM 词对齐的短语翻译对自动抽取方法。

本文第二部分简单介绍基于 HMM 词对齐的短语翻译对抽取方法；第三部分详细介绍改进的基于 HMM 词对齐的短语翻译对抽取方法；第四部分是实验结果及其分析；最后是结束语。

2 基于 HMM 词对齐的短语翻译对抽取方法

统计翻译基本公式中包括一个语言模型和一个翻译模型，翻译模型可以表示为 $p(s \mid t) = \sum_{a} p(s, a \mid t)$，其中，隐含的中间变量 $a = a_1 .. a_I$，用来表示源语言句子中词语和目标语言句子中词语之间的对齐关系，a_i 表示源语言句中第 i 个词对应的目标语言句子中词语的位置。在两个句子所有的对齐方式中，有一种称为韦特比（Viterbi）对齐 $\hat{a}_1^I$，满足以下条件：

$$\hat{a}_1^I = \arg\max_{a_1^I} p(s, a_1^I \mid t)$$

韦特比对齐表示源语言句子和目标语言句子中词语的最佳对齐方式。对于不同的模型，$p(s, a \mid t)$ 由不同的要素组成。[5]提出的基于 HMM 的词对齐中，$p(s, a \mid t)$ 可由公式（1）表示：

$$p(s, a \mid t) = \prod_{i=1}^{I} p(a_i \mid a_{i-1}, I) \times p(s_i \mid t_{a_i}) \qquad (1)$$

其中 $p(a_i \mid a_{i-1}, I)$ 表示源语言句子当前词汇对齐位置 a_i 对前一个词汇对齐位置 a_{i-1} 的依赖关系，I 表示源语言句长，$p(s_i \mid t_{a_i})$ 表示词语的翻译概率。HMM 由于引入了前后词语对

齐位置之间的关系，相比 IBM 模型可以更好的处理短语内部词语的关系，有利于短语翻译对的抽取。

通过 HMM 的训练和解码，对于每一对双语句子，我们可以获得 HMM 的韦特比对齐结果和两种语言词语之间的翻译概率，用来抽取短语翻译对。

[6]提出了一种从 HMM 词对齐结果中抽取短语翻译对的方法，对于每一个源语言句子中的短语，假设其在句子中的位置范围为 i_1 到 i_2，则对应的目标语言短语位置范围为 $j_{\min}=\min_{i_1\le i\le i_2}\{j=a_i\}$，$j_{\max}=\max_{i_1\le i\le i_2}\{j=a_i\}$。这种方法非常简单，但存在以下两个明显不足之处：

1） 容易丢失有用的词语

这种方法利用源语言到目标语言的对齐关系来获取短语翻译对，由于每一个源语言只有一个对应的目标语言位置，这样，当源语言的一个词对应目标语言的多个词的时候容易丢失一些目标语言的词语。如：

我【1】想【2】见见【3】经理【4】。【5】

Null I(1) would(2) like to see(3) the manager(4) .(5)

【】内数字表示源语言句子中词语在句中的位置，（）内的数字表示目标语言句子中词语对应的源语言词语的位置。（下同）在这对句子中，通过 HMM 的韦特比对齐，获得汉语“我”对应的英语为“I”,“想”对应的英语为“would”，这样，获得短语“我想”对应的英语短语为“I would”，丢失了“like to”两个词语，抽取的短语翻译对准确性收到很大的影响。

2）不能获取长度相差较大的短语翻译对

当源语言短语对应两个或两个以上不连续目标语言词汇序列的时候，用这种方法获取的短语翻译对效果会很差，例如：请给我→give me a cup of coffee，please。为了尽量避免这种情况,一般通过限制目标语言短语的长度来处理。这样，当目标语言短语长度相比源语言短语长度相差较大的时候，这种方法是没有办法抽取的。

通过观察，我们发现上面两种情况在中英短语翻译对中都是非常常见的，所以抽取短语翻译对的效果受到很大的影响。

3 改进的短语翻译对抽取方法

我们在 HMM 韦特比词对齐的基础上采用了一种更加有效的方法来从双语语料中抽取短语翻译对，具体过程如下：

3.1 中英短语翻译对抽取

我们把英文当作源语言，中文当作目标语言，对双语句子进行词对齐，如：

I【1】would【2】like【3】to【4】see【5】the【6】manager【】7 .【8】

Null(6) 我(1) 想(2,3,4) 见见(5) 经理(7) 。(8)

对齐以后，我们从目标语言开始，如果目标语言短语对应的源语言所有词汇的位置连续，我们就当作短语翻译对抽取，例如：目标语言短语“ 我想 ”对应的源语言位置“ 1,2,3,4”是连续的，所以抽取“ 我想 ”和“ I would like to ”作为短语翻译对；另外，如果源语言中不连续的位置由 Null 对应，则也进行抽取，如“ 见见经理 ”生成的源语言位置“ 5,7 ”不连续，但位置“ 6 ”是由 Null 对应，所以抽取“ 见见经理 ”和“ see the manager ”作为短语翻译对。由于一个目标语言词语可以对应多个源语言词语，所以很好地避免了上述方法的不足之处。

3.2 英中短语翻译对抽取

我们把中文当作源语言，英文当作目标语言，用上述方法抽取英中短语翻译对，用来作为中英短语翻译对分类和后处理的参考。例如，我们在中英短语翻译对抽取的时候获得“ 我要→I will have ”，在同一对中英文句子中，抽到英中短语翻译对“ I will have→我要 ”，这样，我们可以判定“ 我要→I will have ”这个短语翻译对具有较高的准确率，所有这种在中英短语翻译对抽取和中英短语翻译对抽取的时候可以获得完全对应的中英文短语的翻译对，我们归为一类，做相同的后处理。详细类别及其后处理方法见 3.3 节。

3.3 中英和英中短语翻译对比较分析和后处理

从双语语料的每一个句子中获得了一系列的中英短语翻译对和英中短语翻译对以后，我们根据两种不同的短语翻译对的对应情况对中英短语翻译对进行分类，然后对不同类别的翻译对进行不同的后处理，提高中英短语翻译对的准确率。通过对比从一个句子中获取的中英短语翻译对和英中短语翻译对的对应关系，我们把中英短语翻译对分为四类，下面详细介绍各类的生成过程和后处理办法。

第一类：中英短语翻译对具有中文和英文都相同的英中短语翻译对，如：

I【1】will【2】have【3】some【4】hot【5】milk【6】.【7】

Null 我(1) 要(2,3) 些(4) 热(5) 牛奶(6) 。(7)

我【1】要【2】些【3】热【4】牛奶【5】。【6】

Null I(1) will have(2) some(3) hot(4) milk(5) .(6)

这是两个句子分别作为源语言和目标语言的韦特比对齐结果，从中我们可以分别获得“ 我要→I will have ”和“ I will have→我要 ”的两种不同的短语翻译对。在中英对齐中获得的中英短语翻译对在英中对齐中可以获得完全一致的英中短语翻译对，在这种情况下，获得的短语翻译对具有非常高的准确性，所以我们对这类中英短语翻译对不作后处理。

第二类：中英短语翻译对具有相同英文和不同中文的英中短语翻译对，如：

I【1】would【2】like【3】to【4】see【5】the【6】wine【7】.【8】

Null(6) 我(1) 想(2,3,4) 看看(5) 葡萄(7) 酒。(8)

我【1】想【2】看看【3】葡萄【4】酒【5】。【6】

Null I(1) would(2) like to see(3) the wine(4,5) .(6)

从这两个句子两种不同的韦特比对齐中，可以分别获得中英短语翻译对“ 看看葡萄→see the wine ”和英中短语翻译对“ see the wine→看看葡萄酒 ”，相同的英文短语对应不

同的中文短语,这是因为在对齐的时候每个源语言词语只由一个目标语言词语产生，如“wine”只由“葡萄”产生，这样，当一个源语言词语本身和多个目标语言词语相对应的时候，可能丢失目标语言词汇，但是我们不能简单地以长度较长的目标语言短语作为正确的中文，因为在英中对齐的时候也可能会产生多余的中文词语，即发生下面介绍的第三类中的错误。我们通过下面的公式（2）分别计算两种不同的中文产生英文的概率大小，把概率大的中文作正确的中文短语。

$$P(E_I / C_J) = \prod_i \frac{\sum_j p(e_i / c_j)}{J} = \frac{1}{J^I} \prod_i \sum_j p(e_i / c_j) \qquad (2)$$

在公式（2）中，$p(e_i / c_j)$表示在 HMM 参数训练过程中获得的中文短语中第 j 个单词到英文短语中第 i 个单词的翻译概率，整个公式参考 IBM 翻译模型 1，没有考虑两种短语的长度关系。

第三类：中英短语翻译对具有相同中文和不同英文的英中短语翻译对。如：

Two【1】sandwiches【2】,【3】please【4】.【5】

Null(3) 两(1) 个三明治(2,4) 。 (5)

两【1】个【2】三明治【3】。【4】

Null Two(1,2) sandwiches(3), please .(4)

在这两个句子两种不同的韦特比对齐中，可以分别获得中英短语翻译对“两个三明治→two sandwiches, please”和英中短语翻译对“two sandwiches→两个三明治”。在两种不同的短语翻译对中，相同的中文对应不同的英文，这是因为对齐的时候，一个目标语言词语可以对应多个源语言词语，可能会把一些和中文短语相关性很小的词语加进来，如上述短语中的“please”，但是，我们也不能简单地把长度短的英文短语作为正确的结果，因为在获得英中短语翻译对的时候，可能丢失一些英语词汇，即发生上面第二类中的错误。类似第二类情况的处理方法，我们根据公式（3）来计算两种不同的英文短语生成中文短语的概率，选择概率大的英语短语作为正确的结果。

$$P(C_I / E_J) = \prod_i \frac{\sum_j p(c_i / e_j)}{J} = \frac{1}{J^I} \prod_i \sum_j p(c_i / e_j) \qquad (3)$$

公式（3）中 $p(c_i / e_j)$表示英文短语中第j个单词到中文短语第i个单词的翻译概率。

第四类：中英短语翻译对在英中短语翻译对中没有相同的英文或中文短语。在这种情况下，没有英中短语翻译对作为参考。我们发现，中英短语翻译对发生错误的情况下，很多是由于中英对齐的时候，一个中文生成了多余的英文，这是因为每一个目标语言词语可以生成对应的多个源语言词语，所以在源语言短语边界上容易产生多余的词语。在短语的左右边界上都可能产生多余的词语，我们分别从左和从右逐步去掉英语短语边界词的方法来找到最优的英语短语，如：

I【1】would【2】like【3】to【4】see【5】a【6】wine【7】list【8】.【9】

Null 我(1) 想(2,3,4) 看看(5,6) 葡萄(7) 酒单子(8) 。(9)

在这个对齐中，我们获得中英短语翻译对“我想看看→I would like to see a”，这样，我们去掉英语短语边界词“a”，获得英文短语“I would like to see”，然后分别利用公式（3）计算两个不同的英文短语生成中文短语的概率，如果去掉边界词的英文短语的概率大于未去边界词的英语短语的概率，则取去掉边界词的作为结果，继续判断是否还要去掉边界词，直到左右边界都没有边界词可去则得到最优的英文短语。

4 实验结果及其分析

我们用来抽取的语料为旅游领域130000句中英文句子级双语对齐语料，中文平均句长为7.2个词，英文平均句长为7.5个单词。一共抽取中英文短语翻译对30.6万条左右，总体去除重复以后为14.5万条左右。我们粗略统计了中英文短语翻译对各类所占的比例，同时，在各类短语翻译对中各随机抽取500条结果进行正确率估计，结果如表（一）所示。

表（一）中英短语翻译对抽取结果及正确率

双语短语类别	抽取短语数目(条)	占总体比例(%)	正确率（%）
第一类	90125	29.4	98.2
第二类	40427	13.2	92.2
第三类	40040	13.0	91.0
第四类	136304	44.4	84.2

从结果可以看出，抽取的第一类中英短语翻译对结果具有非常高的准确率，第二类和第三类结果由于有英中短语翻译对作为修正的参考，具有较高的准确率，第四类只是判断边界词情况，准确率相对较差，但总体结果是不错的。

这种方法和[6]提出的方法相比，具有如下几个明显的优点：（1）利用了更多的对齐位置信息，避免了由于一个源语言只对应一个目标语言而丢失词语的情况，提高了短语抽取的准确率；（2）对短语长度没有限制，可以更好的获取中英长度相差较大的短语，获得了更多的有用信息；（3）不同类别的短语翻译对具有不同的准确率，可以根据实际情况加入翻译系统中。

5 结论

基于短语的统计翻译是目前机器翻译研究中的一种重要方法，这种方法需要以大规模的短语翻译对为基础。本文在分析现有方法的基础上，提出了一种改进的基于HMM词对齐的中英短语翻译对抽取方法，这种方法针对HMM词对齐的不同情况，采用不同的方法来抽取短语翻译对，有效地解决了各类不同的错误，提高了短语翻译对抽取的效果。试验结果证明，这种基于HMM词对齐的短语翻译对抽取方法是有效的。下一步的工作是把抽

取的短语翻译对加入到统计翻译系统中，作进一步的试验，分析和改进。

参 考 文 献

[1] Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra and Robert L. Mercer. The Mathematics of Statistical Machine Translation: Parameter Estimation. Computational Linguistics, vol. 19, no. 2, pp. 263–311, 1993.

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

[3] Franz Josef Och and Hermann Ney. A Comparison of Alignment Models for Statistical Machine Translation. Proc. of the 18th International Conference on Computational Linguistics, Saarbrucken, Germany, 2000.

[4] Ashish Venugopal, Stephan Vogel and Alex Waibel. Effective Phrase Translation Extraction from Alignment Models. Proc. of 41st Annual Meeting of ACL, pp. 319-326, Sopporo, Japan, July 2003.

[5] Stephan Vogel, Hermann Ney, and Christoph Tillmann. HMM-based Word Alignment in Statistical Translation. COLING'96: The 16th Int. Conf. on Computational Linguistics, pp. 836–841, Copenhagen, August 1996.

[6] Stephan Vogel, Ying Zhang, Fei Huang, Alicia Venugopal, Bing Zhao and Alex Waibel. The CMU statistical machine translation system. Proc. of the Machine Translation Summit IX, vol. IT-37, no.4, pp. 1085–1094, 1991, New Orleans, LA, 2003.