

统计机器翻译中短语切分的新方法

何中军^{1,2}, 刘群¹, 林守勋¹

(1. 中国科学院计算技术研究所, 北京 100080; 2. 中国科学院研究生院, 北京 100039)

摘要: 基于短语的统计机器翻译是目前主流的一种统计机器翻译方法, 但是目前基于短语的翻译系统都没有对短语切分作专门处理, 认为一个句子的所有短语切分都是等概率的。本文提出了一种短语切分方法, 将句子的短语切分概率化: 首先, 识别出汉语语料库中所有出现次数大于 2 次的词语串, 将其作为汉语短语; 其次, 用最短路径方法进行短语切分, 并利用 Viterbi 算法迭代统计短语的出现频率。在 2005 年 863 汉英机器翻译评测测试集上的实验结果 (BLEU4) 是: 0.1764 (篇章), 0.2231 (对话)。实验表明, 对于长句子 (如篇章), 短语切分模型的加入有助于提高翻译质量, 比原来约提高了 0.5 个百分点。

关键词: 统计机器翻译; 翻译模型; 短语切分

中图分类号: TP391

文献标识码: A

A New Approach to Phrase Segmentation for Statistical Machine Translation

He Zhong-jun^{1,2}, Liu Qun¹, Lin Shou-xun¹

(1. Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100080, China; 2. Graduated University of Chinese Academy of Sciences, Beijing, 100039, China)

Abstract: Currently, Phrase-based Statistical Machine Translation is the state-of-the-art method in SMT community. However, none of the phrase-based systems has the special module to deal with the phrase segmentation, they consider all segmentations of a sentence with uniform distribution. In this paper, we proposed a phrase segmentation method: Firstly, find the word strings occur more than once in Chinese corpus, which are considered as Chinese phrases; Secondly, use the Shortest-Path method to do phrase segmentation, and employ Viterbi algorithm to train iteratively to gain the phrase probability. We do experiments on 2005

基金项目: 国家 863 计划资助项目 (2005AA114140), 国家自然科学基金资助项目 (60573188)

作者简介: 何中军 (1982-), 男, 博士研究生, 研究方向是统计机器翻译, z.jhe@ict.ac.cn

HTRDP (863) MT evaluation test set. Using the phrase segmentation model, the results (BLEU4) are: 0.1764 (writing) and 0.2231(dialog). Experiments show that the phrase segmentation model can help to improve translation quality on long sentences. We get about 0.5 percentage point increase on writing.

key words: Statistical Machine Translation; Translation Model; Phrase Segmentation

1 引言

自从上世纪 90 年代初, Peter Brown 等人^[1]提出了基于信源信道思想的统计机器翻译模型以来, 短短十几年, 基于统计方法的机器翻译有了长足的进步。目前, 基于短语的统计机器翻译 (Phrase-based SMT) 成为主流的统计机器翻译方法。一般的, 基于短语的方法^{[2][3]}将任意连续的字符串都看作短语, 从词对齐的双语语料库中自动学习双语短语, 以短语为单位进行翻译; Och 提出了对齐模板方法^[4], 将单词映射到词类中, 实现了句子级和短语级两级对齐; Chiang 提出了层次短语模型^[5], 形式上是一个同步的上下文无关文法 (synchronous context-free grammar), 允许短语内部包含子短语 (sub-phrase), 此外, 还有许多学者都致力于基于短语方法的研究。相对于基于词的方法 (word-based), 基于短语的方法能够较好的处理短距离依赖 (local-context dependency) 以及常用搭配等问题。

将一个汉语句子 $f_1^J = f_1 f_2 \cdots f_J$ 翻译为英语句子 $e_1^I = e_1 e_2 \cdots e_I$, 需要以下 3 个步骤:

1. 将汉语句子 $f_1^J = f_1 f_2 \cdots f_J$ 切分成汉语短语 $\tilde{f}_1^K = \tilde{f}_1 \tilde{f}_2 \cdots \tilde{f}_K$;
2. 对短语进行调序;
3. 根据翻译模型, 为每一个汉语短语选择合适的英语译文;

一般的, 基于短语的统计机器翻译系统认为一个句子的所有短语切分都是等概率分布的 (Uniform distribution), 这显然是不合理的, “中国建筑业 对外开放” 显然要比 “中国 建筑业对 外开放” 具有更高的可能性。然而, 汉语短语切分是有一定难度的, 一方面, 短语不像词语那样容易界定, 很难说 “中国建筑业” 是一个短语, 还是 “中国建筑业对外开放” 是一个短语; 另一方面, 基于短语的翻译方法一般都认为任意连续的字符串都可以看作短语, 例如 “建筑业对外”, 尽管它并不符合语法。这样就造成了短语切分的难度。

本文提出了一种短语切分方法, 将汉语短语切分概率化。针对机器翻译的特点, 首先从汉语单语语料库中识别出重复词语串作为汉语短语; 其次利用 Viterbi 算法^[6]对汉语语料库进行短语切分, 通过多次迭代, 统计汉语短语的出现频率, 即 1-gram 语言模型。翻译时, 先对汉语进行短语切分, 再进行解码。

本文其他部分安排如下: 第 2 小节介绍短语切分方法; 第 3 小节介绍翻译模型与解码; 第 4 小节介绍实验; 第 5 小节是总结。

2 短语切分方法

基于短语的统计机器翻译以短语为翻译的最小单位, 由每个短语的翻译组成整个句子的翻译。与英语不同, 汉语的最小单位是字, 由字组成词, 由词组成短语, 我们这里说的短语切分, 是指在汉语词语切分的基础上, 将句子切分成短语。显然, 它与汉语的

切词是类似的：词语切分以字为单位，短语切分以词为单位。因此，可以借鉴汉语词语切分的研究方法和成果来进行短语切分。我们采用 N-最短路径方法^[7]进行短语切分：对于一个已经分词的句子，根据短语库，找到这个句子中所有可能的短语，构造有向无环图，求得 N 条最优路径。

例如，对于句子“中国 经济 发展 十分 迅速”，构造有向无环图，如图 1：

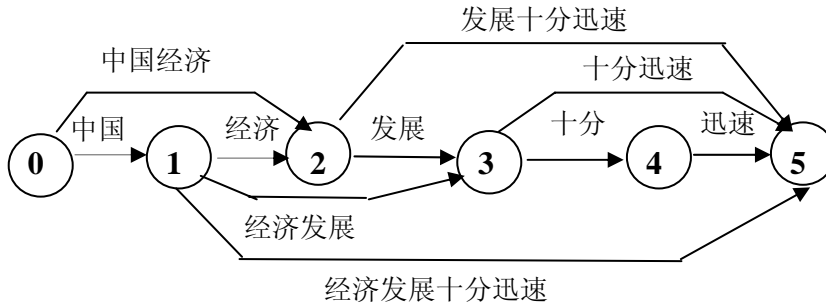


图 1 短语切分有向无环图

这里，有两个问题需要解决：

1. 如何得到短语库，也即，如何确定哪些词构成一个短语？
2. 有向无环图的路径长度如何确定？

一旦这两个问题得到解决，那么就可以对句子进行短语切分：

假设给定一个已经分好词的汉语句子 c_1^J ，将它划分为 K 个短语，

$$c_1^J = \tilde{c}_1^K, \quad \tilde{c}_k = c_{j_{k-1}+1} \dots c_{j_k} \quad (1)$$

那么，短语切分概率计算如下：

$$P_{seg} = \sum_{k=1}^K \log(p(\tilde{c}_k)) \quad (2)$$

其中， $p(\tilde{c}_k)$ 表示短语 \tilde{c}_k 的概率，也即有向无环图的路径长度。

下面，我们将分别讨论这两个问题的解决方法。

2.1 短语查找

正如前文所述，相对于词语而言，短语是一个难以界定的单位，不同的人会产生不同的理解，所以，很难像汉语切词那样人工建立一个短语库。由于基于短语的统计机器翻译将任意连续的字符串看作短语，我们可以利用这一特点，从汉语单语语料库中自动抽取短语库，这一过程，称为短语查找。主要思想如下：

首先，对汉语语料库进行词语切分，并记录每一个词在文件中的出现位置，存储在一个哈希表 WordMap 中；

其次，对于 WordMap 中的每一个词 C_i ，找出其所在文件中的对应位置。根据位置，

向后搜索 k 个词 ($0 \leq k \leq MAX_LEN$ ，在我们的实验中 $MAX_LEN=7$)，得到词串 C_i ，

$C_i C_{i+1}, \dots, C_i C_{i+1} \dots C_{i+k}$, 它们都是以 C_i 开头的短语, 保存这些词串, 并将相应计数加 1;

最后, 输出出现次数大于 2 的重复串, 即短语库。

需要注意的一点是, 如果一个短语是另一个短语的子串, 并且出现次数相同, 我们只保留最长的短语串。例如, 从语料库中, 我们得到 2 个短语“中国 经济”和“中国 经济 发展”, 它们都出现了 5 次, 那么我们只将“中国 经济 发展”保存到短语库中。

另外, 为了防止出现句子无法切分的情况, 我们将每个汉语词也看作一个短语。

2.2 短语概率的计算

短语概率也即有向无环图的长度, 根据概率论知识, 可以利用如下公式计算

$$p(\tilde{c}) = \frac{N(\tilde{c})}{\sum_{\tilde{c}'} N(\tilde{c}')} \quad (3)$$

其中, $N(\tilde{c})$ 表示汉语短语 \tilde{c} 在语料库中的出现次数。

但是, 这样的概率估计仅仅是通过统计语料库中的重复串并根据出现次数计算出来的, 是不准确的。在我们的系统中, 采用 Viterbi 算法进行多次迭代, 来估计短语的一元语言模型概率。

多次迭代的 Viterbi 算法是一种无监督学习方法, 用来估计未知概率分布的事件的模型参数。开始时, 随机指定模型参数, 计算每一个训练样本的最大概率值, 重新统计并更新模型参数。这样, 经过多次迭代之后, 概率分布逐渐逼近真实值。

利用 Vitebi 算法进行迭代训练的算法如下:

表 1 短语切分模型训练算法

1. Find all phrases from training corpus , compute phrase probability with formula (1)
2. for iter =1 to MAX_ITERATION
3. for each sentence in training corpus
4. using the shortest length algorithm to find the phrase segmentation with the highest probability
5. for each phrase p in the best segmentation
6. count(p)++;
7. compute the new phrase probability

其中 MAX_ITERATION 表示最大迭代次数, 一般迭代 3~5 次可以收敛。通过 Viterbi 训练, 最终可以得到短语切分的 1-gram 模型。

3 翻译模型与解码

翻译模型和解码算法是统计机器翻译的核心部分, 翻译模型能够反映对机器翻译过程的认识, 解码器能够搜索出最终的译文。下面将介绍我们所用的翻译模型和解码算法。

3.1 翻译模型

我们的翻译模型采用 Log-linear 直接翻译模型^[8],

$$\begin{aligned}\Pr(e_1^I | f_1^J) &= p_{\lambda^M}(e_1^I | f_1^J) \\ &= \frac{\exp[\sum_{m=1}^M \lambda_m h_m(e_1^I, f_1^J)]}{\sum_{e_1^I} \exp[\sum_{m=1}^M \lambda_m h_m(e_1^I, f_1^J)]}\end{aligned}\quad (4)$$

在所有可能的候选翻译中, 选择概率最大的翻译作为最终翻译,

$$\hat{e}_1^I = \arg \max_{e_1^I} \{\Pr(e_1^I | f_1^J)\} \quad (5)$$

特征函数选取 7 个: 短语翻译概率 $p(\tilde{e} | \tilde{f})$ 和 $p(\tilde{f} | \tilde{e})$, 词汇化短语翻译概率 $lex(\tilde{e} | \tilde{f})$

和 $lex(\tilde{f} | \tilde{e})$ ^[2], 英语语言模型 $lm(e_1^I)$, 英语句子长度 I , 短语切分概率 P_{seg} 。

3.2 解码

对一个汉语句子的, 首先利用第 2 小节介绍的方法进行短语切分, 取 1 个或者 N 个最好的切分进行翻译。对每一个切分结果, 利用柱式搜索 (Beam Search) 进行单调解码, 即从左到右顺序翻译每个短语片段, 不对其进行顺序调整。为了节省内存, 加快搜索速度, 对于每一个汉语短语, 解码器从短语表中只读进 n 个最好的英语短语翻译, 搜索过程中每个栈的大小限制为 m (在我们的试验中 $n=10, m=100$)。

Log-linear 模型的参数训练采用最小错误率训练算法^{[9][10]}。

4 实验

我们在 2005 年 863 汉英机器翻译评测的测试集 (包括对话和篇章) 上进行了实验。训练集采用 2005 年 863 评测提供的训练集, 约 10M 英语词, 10M 汉语词。利用 SRI 语言模型工具^[11]在训练集上训练了 3-gram 英语语言模型, 利用本文介绍的方法训练短语切分模型。采用 2004 年 863 汉英机器翻译评测的测试集 (对话和篇章) 作为开发集。

对于翻译模型的训练, 首先利用 GIZA++^[12]从两个方向进行训练 (汉语到英语, 英语到汉语) 获得词语对齐, 并采用 grow-diag-final^[2]方法优化对齐, 然后进行短语抽取^{[3][13]}得到短语翻译概率表。

作为基线系统 (baseline) 我们使用另外一种方法进行短语切分: 利用从语料库中抽取的双语短语中的汉语短语作为短语库, 将 $p(\tilde{e} | \tilde{f})$ 作为短语切分的路径长度, 然后再利用最短路径方法进行短语切分。

在搜索过程中, 为了在效率和翻译质量上取得平衡, 对于每个句子, 取 20 个最好的短语切分进行翻译。

实验结果如表 2 (使用 863 提供的评测工具, 采用 BLEU4 作为评测指标)

表 2 实验结果

	05 年对话	05 年篇章
Baseline	0.2303	0.1716
Phrase Segment Model	0.2231	0.1764
Best SMT System*	0.1814	0.1188

Best SMT System* 代表当年最好的统计机器翻译系统结果

从表 2 可以看出，我们的系统要远远好于当年参加评测的最好的统计机器翻译系统，相对于基线系统而言，短语切分的加入有助于提高篇章翻译的质量，而对于对话反而起到了副作用。以下原因导致了这一现象的发生：

1. 我们的短语切分模型是单独用汉语语料训练的，以重复词语串作为汉语短语，这样就会比较倾向于长度较短的短语。而双语短语库则是根据对齐语料库抽取的，长短语和短短语都能抽取到。对于对话语料来说，其句子一般较短，利用短语切分模型，会将句子切的更碎，使得系统性能下降。例如，对于句子“价钱是多少”，短语切分模型会将其切为“价钱 是 多少”，这样就是基于词的翻译“Price is what”，而在双语短语库中可能直接会有匹配的短语“价钱是多少 ||| How much is it ”，由于短语切分和短语抽取的策略不同，导致了这一现象的发生，对于对话的影响尤其大。
2. 对话语料中，疑问句占的比重较大，在 2005 年的测试语料中有 36% 的句子是问句。一般来说疑问句都是要做词序调整的，而我们的系统是顺序解码，因此在翻译问句的时候表现不好，短语切分的引入使这一缺点更加突出，它将句子切的更碎。
3. 在训练语料中，对话语料大约占 1/4，对于训练汉语的短语切分模型来说，数量比较少，也影响了短语切分模型的作用。
4. 对于篇章来说，句子一般较长，且大多数是新闻语料，其词序没有对话那样变化强烈，利用短语切分模型可以将常见的短语切分出来，有利于翻译质量的提高。

5. 总结

目前，各种基于短语的统计机器翻译系统都认为一个句子的短语切分是等概率分布的，本文提出了一种自动短语切分方法，从已分词的汉语单语语料库中发现重复串，利用 Viterbi 算法估算短语概率，从而可以利用 N-最短路径方法进行短语切分，并将其概率化。实验表明，将短语切分模型加入到统计机器翻译系统中，有助于提高长句子的翻译质量。

短语切分不像汉语词语切分那样，它没有统一的标准，更没有已切分好的训练语料库，这给切分工作带来了很大的困难。从语料库中发现重复串作为短语库可以解决这一问题，而且只用到了单语的语料库，而单语的语料库通常是很容易获得的。然而，用于机器翻译时，也带来了问题：一般统计机器翻译中的短语库都是从词对齐的双语语料库中自动抽取的，这与短语切分使用的短语库是不一致的，因此，短语切分的结果不一定都能找到对应的翻译，这样就要回退到基于词的翻译。这种不一致会影响到翻译质量，例如第 4 小节在对话语料上的实验，这是我们下一步要解决的问题。

另外，还可以引入调序模型，这将有助于加强短语切分模型的作用。短语切分将一些经常在一起出现的词切分成短语，它们翻译成英语时一般也是在一起的，调序模型可以将这些短语作为一个整体进行顺序的调整。

参考文献:

- [1] Peter. F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, et al. The Mathematics of Statistical Machine Translation: Parameter Estimation [J], Computational Linguistics, 1993, 19(2): 263-311
- [2] Philipp Koehn, Franz Josef Och, and Daniel Marcu. Statistical phrase-based translation [A]. In Proceedings of HLT-NAACL [C], 2003, 127-133.
- [3] R. Zens, F. J. Och, H. Ney. Phrase-Based Statistical Machine Translation [A]. In: M. Jarke, J. Koehler, G. Lakemeyer (Eds.) : KI - 2002: Advances in artificial intelligence. 25. Annual German Conference on AI [C], Springer Verlag, September 2002, Vol. LNAI 2479:18-32
- [4] F. J. Och, C. Tillmann, and H. Ney. Improved alignment models for statistical machine translation [A]. In Proc. of the Joint SIGDAT Conf. On Empirical Methods in Natural Language Processing and Very Large Corpora [C], university of Maryland, College Park, MD, June 1999, pages 20-28
- [5] David Chiang. A hierarchical phrase-based model for statistical machine translation [A]. In Proceedings of ACL 2005 [C], pages 263-270
- [6] G. David Forney, Jr. The Viterbi algorithm [A], Proc. of the IEEE [C], March, 1973 , 61(3): 268-278
- [7] 张华平, 刘群. 基于 N-最短路径方法的中文词语粗分模型 [J], 中文信息学报, 2002, 16(5): 77-84.
- [8] Franz Josef Och and Hermann Ney. Discriminative training and maximum entropy models for statistical machine translation [A]. In Proceedings of the 40th Annual Meeting of the ACL [C], 2002, 295-302.
- [9] Franz Josef Och. Minimum Error Rate Training for Statistical Machine Translation [A]. In "ACL 2003: Proc. of the 41st Annual Meeting of the Association for Computational Linguistics" [C], Japan, Sapporo, July 2003.
- [10] Ashish Venugopal, Stephan Vogel. Considerations in Maximum Mutual Information and Minimum Classification Error training for Statistical Machine Translation [A], In the Proceedings of the Tenth Conference of the European Association for Machine Translation (EAMT-05) [C], Budapest, Hungary May 30-31, 2005
- [11] Stanley F. Chen and Joshua Goodman. An empirical study of smoothing techniques for language modeling [J]. Technical Report TR-10-98, Harvard University Center for Research in Computing Technology, 1998
- [12] Franz Josef Och and Hermann Ney. Improved statistical alignment models [A]. In Proceedings of the 38th Annual Meeting of the ACL, 2000 [C], 440-447.
- [13] Franz Josef Och. Statistical Machine Translation: From Single-Word Models to Alignment Templates [D]. Germany: Computer Science Department, RWTH Aachen, October, 2002