文章编号：1003-0077(2014)02-0051-05

一种基于改进隐马尔克夫模型的词语对齐方法

刘 颖，姜 巍

（清华大学 中文系，北京 100084）

摘 要：该文在基本隐马尔克夫模型的基础之上，利用句法知识来改进词语对齐，把英语的短语结构树距离和基本隐马尔克夫模型相结合进行词语对齐。与基本隐马尔克夫模型相比，这个模型可以降低词语对齐的错误率，并且提高统计机器翻译系统 BLEU 值，从而提高机器翻译质量。

关键词：短语结构树距离；隐马尔克夫模型；词语对齐；BLEU 值

中图分类号：TP391　　**文献标识码**：A

An Improved HMM Based Word Alignment Method

LIU Ying, JIANG Wei

(Department of Chinese Language and Literature, Tsinghua University, Beijing 100084, China)

Abstract: This paper improves the HMM based word alignment by introducing syntactic knowledge. HMM is combined with English phrase structure tree distance to align Chinese-English words. Experiments shows that the improved HMM can reduce the error rate of word alignment, and improve the BLEU score of statistical machine translation.

Key words: phrase structure tree distance; hidden Markov model; word alignment; BLEU score

1 引言

统计机器翻译由 IBM 的 Brown 等人于 1990 年提出，1993 年他们提出了基于词对齐的五个复杂度递增的模型—IBM 模型 1 至 5[1]。IBM 统计机器翻译都是以词为基本的翻译单位，词的对齐与词语翻译概率和对齐概率有关。IBM 模型 1 假设对齐概率是平均对齐，即与源语言句子的长度成反比。IBM 模型 2 假设对齐概率与源语言、目标语言的句子长度以及源语言位置和目标语言位置相关。IBM 模型 3 和 4 考虑了空源语言词、繁殖率和扭曲模型。IBM 的重新排序模型很少利用上下文，更没有利用句法结构，许多人尝试把句法信息结合进翻译模型中来改进这个模型[2]。Vogel 提出基于隐马尔克夫模型（简称 HMM）的统计翻译，利用 HMM 进行的对齐概率依赖于前一个词所对齐的词在目标语言句子的位置[3]。即源语言的两个词位置越近，它们的目标词在目标语言句子的位置也越近。Och 系统比较了 IBM 模型和 HMM[4]，在此基础之上发布了词语对齐软件 Giza++①，Giza++ 实现了 IBM 模型 1 至模型 5 和 HMM 词语对齐，目前已成为多数统计机器翻译系统的基本模块。词对齐是统计机器翻译的基础，词对齐的质量影响统计机器翻译的质量。

Lopez 对基本隐马尔科夫模型进行改进，提出基于目标语言串距离和依存树距离的 HMM[5]。这个模型不仅取决于两个对齐位置在目标语言串上的距离，而且取决于这两个对齐位置在目标语言依存树上的距离。实验结果表明，基于依存树距离的 HMM 在词语对齐训练中召回率较高，错误率较低。Cherry 利用目标语言的依存树对逆转换语法进行约束，以提高词语对齐的质量[6]。国内对词语对齐也进行了许多研究和探索，取得了较好的成绩[7-10]。

由于汉语与英语互为翻译的词之间存在一对

① http://www.fjoch.com/GIZA++.html

收稿日期：2012-03-14　**定稿日期**：2012-12-24

基金项目：教育部回国人员启动项目(20101021603)

多、多对一、一对空和空对一等情况，同时汉语和英语在表达时间、地点、介绍已知信息和未知信息、对句子中的某些信息进行强调等方面都存在语序上的不同，使得从大规模双语语料库中对词进行对齐时，汉语和英语的词的顺序不再完全保持。

基本 HMM 中词的对齐与两个词的翻译概率和两个词对齐的目标语言词的串距离有关系。当汉语和英语互译词的顺序改变，两个词的翻译概率又比较小时，基本 HMM 可能给出错误的词语对齐结果。本文提出改进的 HMM，将两个对齐位置的目标语言短语结构树距离作为特征引入到词语对齐模型中，使得词的对齐不仅与两个词的翻译概率、两个词对齐的目标语言词的串距离有关，而且与两个词对齐的目标语言词的短语结构树距离有关。改进的 HMM 与基本 HMM 词对齐一样存在全局最优词语对齐，可以在多项式时间内找到最优的词语对齐[6]。

2 模型

$$Pr(f_1^J \mid e_1^I) = \sum_{a_1^J} \prod_{j=1}^{J} [p(a_j \mid a_{j-1}, I) p(f_j \mid e_{a_j})] \tag{1}$$

1) 基本 HMM

式(1)为 HMM 基本形式，$p(a_j \mid a_{j-1}, I)$称为对齐概率，$p(f_j \mid e_{a_j})$称为翻译概率。这个模型是 Vogel 在 1996 年提出来的，对齐概率依赖于两个对齐位置的串距离 $a_j - a_{j-1}$[3]。Och 改进了这个模型，对齐概率取决于目标语言串两个对齐位置的串距离 $a_j - a_{j-1}$和自动确定的词类 $C(e_{a_{j-1}})$。而在文献[5]中，对齐概率不仅取决于目标语言串两个对齐位置的串距离 $a_j - a_{j-1}$，而且取决于目标语言串两个对齐位置在依存树中的距离。

2) 改进的 HMM

改进的 HMM 对基本 HMM 的对齐概率 $p(a_j \mid a_{j-1}, I)$进行了改进，但翻译概率 $p(f_j \mid e_{a_j})$与基本 HMM 相同。

$$Pr(f_1^J \mid e_1^I) = \sum_{a_1^J} \prod_{j=1}^{J} [p(a_j \mid a_{j-1}, I) p(f_j \mid e_{a_j})] \tag{2}$$

改进的 HMM 的对齐概率 $p(a_j \mid a_{j-1}, I)$与源语言串上的两个词在目标语言串两个对齐位置之间的串距离和短语结构树距离有关。二者分别作为一个特征，见式(3)。

$$p(a_j \mid a_{j-1}, I) = \lambda_1 \frac{c(i-k)}{\sum_{l=1}^{I} c(l-k)} + \lambda_2 \frac{t(i,k)}{\sum_{m=1}^{I} t(m,k)} \tag{3}$$

式(3)中，$i = a_j$表示第 j 个源语言词与第 i 个目标语言词对齐。$k = a_{j-1}$表示第 $j-1$ 个源语言词与第 k 个目标语言词对齐。$c(i-k)$表示两个源语言词对齐的两个目标语言词的串距离，$c(i-k)$的定义和运算与基本 HMM 相同[3]。$\lambda_1 + \lambda_2 = 1$。$t(i,k)$表示两个源语言词 $j-1$ 和 j 对齐的目标语言词在目标语言短语结构树的距离。分母是归一化因子。

下面用实例给出如何计算两个词之间短语结构树距离。

图 1 中，从节点“oriented”到节点“the”的短语结构树距离 t(5,1)定义如下：

PopScore[5][1]× PopScore[5][1] ×
PopScore[3][1] × PopScore[1][1] ×
PushScore[1][2] × PushScore[2][2]×
PushScore[5][2] (4)

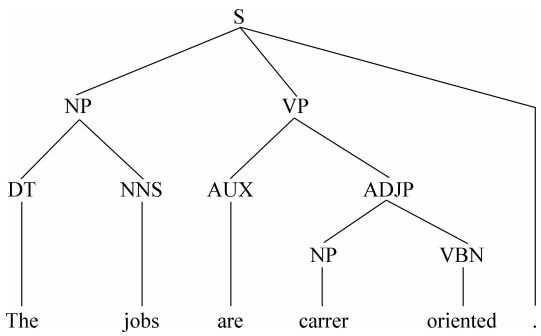

图 1 “The jobs are carrer oriented”短语结构树

从 oriented 到 the 的短语结构树距离为从 oriented 到 the 的操作概率的乘积。每个操作的相应概率定义如下。

(1) POP 操作概率：依赖于当前节点的父节点类型 NodeType 和当前节点在兄弟节点中的索引 NodeIndex。记为 PopScore[NodeType][NodeIndex]。

(2) PUSH 操作概率：依赖于当前节点的类型 NodeType 和当前节点的孩子节点在所有孩子节点中的索引 NodeIndex。记为 PushScore[NodeType][NodeIndex]。

引入父节点类型的原因在于：统计训练语料(斯坦福句法分析器处理结果)中子树根节点类型的出现频率，发现 S,SBAR,NP,VP 和 PP 的出现频

率较高。其中，S 为一般性陈述句标记，SBAR 为由引导词引导的从句，NP 为名词短语，VP 为动词短语，PP 为介词短语。据此将父节点类型（Node-Type）分为 5 种：S 和 SBAR（记为 1），NP（记为 2），VP（记为 3），PP（记为 4）和其他短语（记为 5）。

本文将当前节点在兄弟节点中的索引分为两类，最右索引（记为 1）及其他索引（记为 2）。通过实验发现，对于父节点与兄弟节点分类，可以降低时空消耗，缓解数据稀疏问题，同时保证二者结果相近。

前向—后向算法初始化时，每个操作概率采用最大频率似然估计法来估计。即：

$$\mathrm{PopScore[a][b]} = \frac{f(a \rightarrow b)}{\sum_r f(a \rightarrow r)} \tag{5}$$

$$\mathrm{PushScore[a][b]} = \frac{f(a \rightarrow b)}{\sum_r f(a \rightarrow r)} \tag{6}$$

$f(a \rightarrow b)$表示父节点类型为 a、孩子索引为 b 在树库中共出现的次数。通过上述模型，可以通过短语结构树 T_e 计算任意两个位置在短语树上的距离。该模型主要是为了解决子树边界的词语对齐和句法约束相冲突的问题。

3　前向—后向算法

本文采用前向—后向算法来训练参数。首先通过初始参数计算双语互译和对齐概率，然后根据计算过程中发生的状态转移和生成的符号信息更新参数，在保证新参数优于原参数情况下进行更新，即新的模型参数应该可以更好的解释双语互译和对齐。前向—后向算法利用前向变量和后向变量可以直接进行最大化，其基本假设是在这轮计算过程中出现频率高的状态转移和生成的符号应获得更高的概率。

对改进的 HMM，前向变量和后向变量的定义如下：

前向变量记为 $\alpha_i(j)$，记录源语言第 j 个词对应目标语言位置 i 的总概率。根据动态规划算法 $\alpha_i(j)$可以通过下列过程计算：

$$\alpha_i(1) = \pi_i$$

$$\alpha_k(j+1) = \sum_i \alpha_i(j) a_{ik} b_{io_j} \tag{7}$$

其中 π_i 表示目标语言位置 i 的初始概率；b_{io_j} 表示目标语言第 i 个位置词 e_i 生成 o_j 的概率，即翻译概率，$b_{io_j} = p(o_j \mid e_i)$。$a_{ik}$ 表示在源语言第 j 个词对应目标语言位置 i 的情况下，源语言第 $j+1$ 个词对应目标语言位置 k 的对齐概率。即 $a_{ik} = p(a_{j+1} \mid a_j, I)$。

后向变量记为 $\beta_i(j)$，记录源语言第 j 个词对应目标语言位置 i 时，剩余子串的对齐概率之和。$\beta_i(j)$同样可以利用动态规划算法计算，见式(8)。

$$\beta_i(T+1) = 1 \quad \beta_i(j) = \sum_j \beta_k(j+1) a_{ik} b_{io_j} \tag{8}$$

T 表示源语言最后一个词。$\xi_{ik}(j)$表示给定双语互译句对的情况下，源语言第 j 个词对应目标语言第 i 个位置并且源语言第 $j+1$ 个词对应目标语言第 k 个位置的概率，称为词语对齐的边后验概率。用前向变量和后向变量表示为式(9)。

$$\xi_{ik}(j) = \frac{\alpha_i(j) a_{ik} b_{io_j} \beta_k(j+1)}{\sum_{m=1}^{N} \alpha_m(j) \beta_m(j)} \tag{9}$$

根据以上定义，更新本文 HMM 对齐概率的公式为式(10)。

$$a_{ik} = \frac{\sum_{j=1}^{T} \xi_{ik}(j)}{\sum_{j=1}^{T} \sum_{k=1}^{N} \xi_{ik}(j)} \tag{10}$$

而 $\sum_{j=1}^{T} \xi_{ik}(j)$ 表示给定双语互译句对的情况下，从目标语言第 i 个位置跳到第 k 个位置的所有概率之和。更新本文 HMM 翻译概率参数的公式为式(11)。

$$b_{io_j} = \frac{\sum_{j=1}^{T} \delta(o_j, w_k) \xi_{ik}(j)}{\sum_{j=1}^{T} \xi_{ik}(j)} \tag{11}$$

如果，$o_j = w_k$，$\sigma(o_j, w_k) = 1$，否则，$\sigma(o_j, w_k) = 0$。

然后利用前向后向算法进行双语互译和对齐概率的计算。

前向后向算法：

① 初始化。双语词互译概率来自汉英双语词典。

② 根据双语句对创建所有可能的状态转移矩阵。对于所有两两状态转移，计算其状态转移概率。然后根据式(3)计算双语初始对齐概率，$\lambda_1 = \lambda_2 = 0.5$。

③ 根据式(7)和式(8)计算这个阶段的前向变量和后向变量。

④ 根据式(9)，利用前向变量和后向变量计算词语对齐的边后验概率。

⑤ 根据式(10)更新 HMM 的对齐概率，根据式

(11)更新 HMM 的翻译概率。

⑥ 重复步骤 2,直到模型参数变化小于某个阈值或者达到指定迭代次数。

在本文实验中,设置该阈值为 0.001,在初始值如上设置的情况下,一般 15 轮至 20 轮迭代可以达到收敛。

4 实验、结果和分析

1) 实验数据

实验采用双语平行训练语料,语料大部分是从互联网抓取后经过后处理获得,此外包括哈尔滨工业大学的 10 万平行双语句对,整个训练集包含 50 万平行双语句对,汉语平均句长 15.01,英语平均句长 13.84。本文计算 BLEU 值的测试语料是单独准备的 500 句汉英互译句对;计算词语对齐质量的测试语料是经过人工标注词语对齐结果的 500 句汉英互译句对。训练语料和测试语料需要经过分词和大小写转化的预处理。

2) 开源工具

实验中采用的自动分词软件是斯坦福分词工具 2008 版;采用的句法分析器是斯坦福句法分析器 2007 版,标注集为宾州树库标注集,采用宾州树库来统计短语结构树距离。采用的语言模型工具为 Srilm1.5.5 版[11]和 LDC 免费 Web-1TB 三元语言模型语料。机器翻译自动评测工具采用了 NIST 的 mt-evaluation1.1 版①,利用 BLEU-四元语言模型评测[12]。

实验中采用了两种词语对齐模块,一个是 Giza++ 模块,一个是改进的 HMM 词语对齐模块,这两个模块的输入输出格式相同。输入是双语平行语料,输出是 Giza++ 格式的双向最优词语对齐结果。

3) 实验及结果分析

实验 1 用实例来分析改进 HMM 对不同位置目标词概率的影响。根据改进 HMM,在给定前一个对齐位置的条件下,可以计算下一个对齐位置的概率。

图 2 给出汉语与英语的词对齐和英语的短语结构树。图 3 是在给定"减轻"的对齐位置为"relieve",计算"对"的不同对齐位置的概率信息,横坐标中的 0 表示"relieve"节点,其他横坐标为其他节点相对于"relieve"的串的距离,纵坐标为对齐到该节点的概率。可以看出对于正确的对齐位置"on"

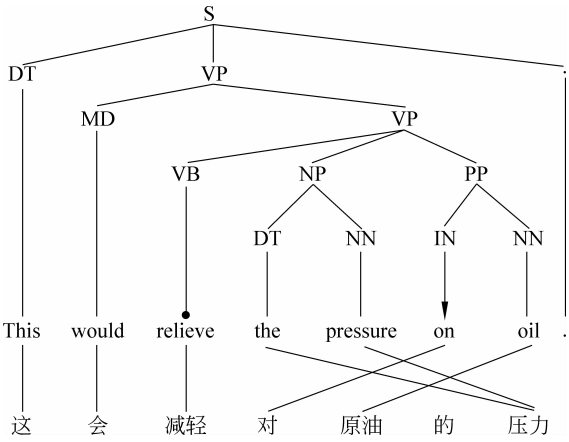

图 2 英语短语结构树及词汇对齐

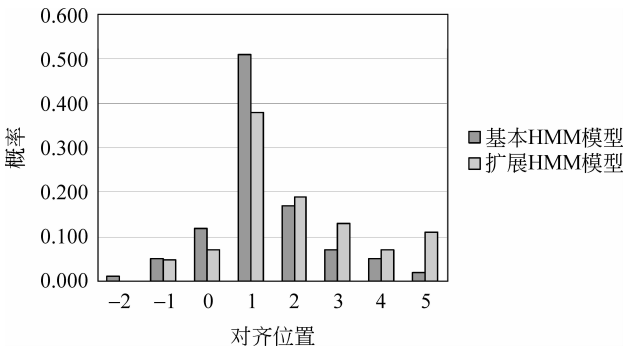

图 3 "对"的不同对齐位置概率

节点 3,改进的 HMM 给出更高的得分。从效果看,改进的 HMM 对概率分布函数进行了平滑,即源语言串上相近的两个词在目标语言较远的对齐位置的概率增加了。

实验 2 根据对齐错误率 AER 比较 HMM 和改进的 HMM 的词语对齐质量;词语对齐结果的评测采用准确率 P,召回率 R 和对齐错误率 AER[13]。

$$P = \frac{A \cap G}{A} \tag{12}$$

$$R = \frac{A \cap G}{G} \tag{13}$$

$$\mathrm{AER} = 1 - 2\,\frac{|A \cap G|}{|A| + |G|} \tag{14}$$

其中 A 为各个模型给出的测试集的词对齐集,G 为测试集的正确词对齐集。

表 1 给出了两种词语对齐模型的评测结果。从表 1 中可看出,改进 HMM 的词对齐准确率较高,词对齐错误率较小。从基本 HMM 到改进 HMM,词对齐的准确率有所增加,召回率有所降低,词对齐错误率有所降低。这说明短语结构树距离对于提高

① http://www.nist.gov/speech/tools/

词语对齐质量，降低词对齐错误率确实有帮助。但同时，考虑短语结构树距离的 HMM 使得词语对齐召回率降低。

表 1　两种词语对齐模型的评测结果

	准确率/%	召回率/%	AER/%
基本 HMM	78.7	77.6	24.4
改进的 HMM	80.2	73.8	23.8

实验 3　比较两种词语对齐结果对统计机器翻译系统 BLEU 值的影响。实验结果见表 2。

表 2　两种词语对齐模型对翻译系统的影响

BLEU 值	参考集	测试集	平均值
基本 HMM	26.44	25.89	26.17
改进的 HMM	26.92	26.52	26.72

从基本 HMM 到改进的 HMM，参考集、测试集和平均值的 BLEU 值都增加了，这说明从基本 HMM 到改进的 HMM，统计机器翻译质量有所提高。改进的 HMM 与基本 HMM 相比，确实说明短语结构树距离对提高机器翻译质量有帮助。

5　结论

对于双语的词语对齐，本文提出了改进的 HMM。改进的 HMM 把源语言词在目标语言的对齐位置的串距离和短语结构树距离融合起来进行词语对齐。实验结果表明，改进的 HMM 可以减少句法和词语对齐冲突，提高对齐准确率，降低对齐错误率，从而提高机器翻译质量。

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] Heidi J Fox. Phrasal cohesion and statistical machine translation[C]//Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing, Philadelphia, USA, 2002: 304-311.

[3] Stephan Vogel, Hermann Ney, Christoph Tillmann. HMM-based word alignment in statistical translation [C]//Proceedings of the 16th International Conference on Computational Linguistics Proceedings, 1996: 836-841.

[4] Franz Josef Och, Hermann Ney. A systematic comparison of various statistical alignment models[J]. Computational Linguistics, 2003, 29 (1): 19-51.

[5] Adam Lopez, Philip Resnik. Improved HMM alignment models for languages with scarce resources[C]// Proceedings of ACL-2005: Workshop on Building and Using Parallel Texts—Data-driven machine translation and beyond. University of Michigan, Ann Arbor, 2005: 83-86.

[6] Colin Cherry, Dekang Lin. Soft syntactic constraints for word alignment through discriminative training [C]//Proceedings of the Coling/ACL 2006 Main Conference Poster Sessions, Sydney, 2006: 105-112.

[7] Yang Liu, Qun LIU, Shouxun LIN, Log-linear Models for Word Alignment[C]//Proceedings of the 43rd Annual Meeting of Association of Computational Linguistics, Michigan, 2005: 25-30.

[8] 常宝宝. 基于统计的翻译等价词对抽取研究[J]. 计算机学报, 2003, (5): 616-621.

[9] 赵红梅，刘群，等，汉英词语对齐规范，中文信息学报，2009，23(3)：65-87。

[10] 肖桐，李天宁，陈如山，等. 面向统计机器翻译的重对齐方法研究，中文信息学报，2010，24(1)：110-116.

[11] Andreas Stolcke. SRILM—An Extensible Language Modeling Toolkit[C]//Proceedings of International Conference on Spoken Language Processing. Denver, Colorado, 2002.

[12] Kishore Papineni, Salim Roukos, Todd Ward, et al. BLEU: a method for automatic evaluation of machine translation [C]//Proceedings of the 40th Annual meeting of the Association for Computational Linguistics, Philadelphia, 2002: 311-318.

[13] D Gildea. Loosely tree-based alignment for machine translation [C]//Proceedings of the 41st Annual Meeting of Acl, 2003: 80-87.

刘颖(1969—)，博士，副教授，主要研究领域为自然语言处理。

E-mail: yingliu@tsinghua.edu.cn

姜巍(1983—)，硕士，主要研究领域为自然语言处理。

E-mail: jiangwei@iyuewe.com