基于翻译规则的统计机器翻译

刘 颖 姜 巍
(清华大学中文系 北京 100084)

摘 要 扩展HMM模型可以解决词语对齐结果与句法约束冲突,从而更好地进行词语对齐。在短语对齐基础上利用目标语言的短语结构树抽取翻译规则。采用扩展CYK算法CYKA+作为系统的解码器,该算法可以处理非乔姆斯基范式的翻译规则;采用两轮解码算法在解码过程中整合语言模型。实验表明,与传统词语对齐模型相比,改进的HMM词语对齐模型具有更高的对齐准确率,并且翻译结果的BLEU评测得分更高。采用翻译规则的系统在不同数据集上具有更稳定的翻译结果。两轮解码算法与立方剪枝算法具有相近的解码质量,但前者解码速度更快。

关键词 统计机器翻译,扩展HMM模型,翻译规则,CYK$^+$算法,BLEU评分
中图法分类号 TP391.1 **文献标识码** A

Statistical Machine Translation Based on Translation Rules

LIU Ying JIANG Wei
(Department of Chinese Language and Literature, Tsinghua University, Beijing 100084, China)

Abstract Improved hidden Markov model was used to align words and solve the inconsistency between word alignment and phrase structures. Translation rules were extracted based on aligned phrases and English phrase trees. An extended CYK - CYK$^+$ algorithm was used as the decoder and a two-pass-decoding algorithm was proposed for intergrating the language model during decoding, which can decode non-Chomsky normal form. The experimental results show the BLEU score of improved HMM is higher than the score of HMM, and the translation quality of translation rules is better than phrase-based machine transtion. The BLEU score of two-pass-decoding algorithm is close to the score of cube prune algorithm and decoding time costs less.

Keywords Statistical machine translation, Improved hidden markov model(HMM), Translation rule, CYK$^+$ algorithm, BLEU

1 引言

IBM的Brown等人于1993年提出了基于词对齐的5个复杂度递增的模型—IBM模型1至5[1],实现了统计机器翻译。1996年,Vogel提出基于隐马尔克夫模型(简称HMM)的统计翻译[2]。Och系统比较了IBM模型和HMM模型,实现了IBM模型1至模型5和HMM模型词语对齐Giza++[3]。Heidi J. Fox指出,Giza++的词语对齐结果与句法约束出现冲突的可能性很高,Giza++存在大量这类错误的词语对齐结果[4]。Och等人提出的对齐模板技术可以解决数据稀疏问题[5]。Och用最大熵模型将各种各样的语言特征和统计信息融合到统计机器翻译中[6]。在统计机器翻译中比较深入地利用句法信息的有吴德恺的反向转换文法[7]、Chiang提出的层次化短语模型[8,9]、Yamada和Knight的树串模型[10]、Galley的树串模型[11,12]、刘洋和刘群的树到串对齐模板的翻译模型[13]、宗成庆的改进树串模型[14]和Melamed的多文本语法[15]等。Chiang的模型借用了形式化语法的结构,运用柱搜索的CYK句法分析器。

基于句法的翻译模型利用句法分析器或树库的信息,期望获得句法信息的指导。基于句法的模型应当能够兼容所有的短语,这样才能既充分保留基于短语的模型的优点,又能发挥句法信息的指导作用。Chiang在2005年实现了对双语短语的完全兼容。Yamada和Knight提出了真正意义上的基于句法的树到串的翻译模型[10]。其输入是一棵句法树,输出是一个句子。Yamada和Knight的树串模型平均对齐评分高于IBM模型5平均对齐评分。

近年来,国内也逐渐开展了统计机器翻译研究。在统计机器翻译学术领域,中科院计算所、自动化所、软件所、哈尔滨工业大学、清华大学、东北大学、北京大学、厦门大学等单位联合推出了基于短语的统计机器翻译系统"丝路",近年来发表了若干有影响的研究工作。

总体来说,基于短语的模型是目前统计机器翻译中主流的方法,模型简单,翻译质量较高。但短语对齐的质量依赖于词语对齐。IBM模型1至5和HMM模型可能导致词语对齐结果与句法约束冲突,本文在HMM模型基础上利用短语结构树距离来解决这个冲突,从而提高词语对齐质量。串-树

到稿日期:2012-05-16 返修日期:2012-08-17 本文受教育部自主科研项目(20111081010)资助。
刘 颖(1969—),女,博士,副教授,主要研究方向为自然语言处理,E-mail:yingliu@tsinghua.edu.cn;姜 巍(1983—),男,硕士,主要研究方向为自然语言处理。

模型处理句法特征的能力较强，处理词汇特征的能力较弱；短语模型和层次短语模型处理词汇特征的能力较强，处理句法特征的能力较弱。本文在层次短语模型基础之上利用目标语言的短语结构树来抽取翻译规则，充分利用了词汇特征和句法特征。由于翻译规则可能是非乔姆斯基范式的，CYK 算法无法进行解码。本文采用改进的 CYK 算法 CYK$^+$ 进行解码。

2 统计机器翻译系统翻译过程

统计机器翻译系统翻译过程：

(1)预处理：主要对汉语进行自动分词，对英语句子的大小写进行转化，对汉英双语长句子进行过滤，这样做是因为 BLEU 评测考虑了大小写情况。采用的自动分词软件是斯坦福分词工具 2008 版，采用北京大学的词性标注集。

(2)词语对齐：使用 Giza++[1)] 和扩展 HMM 模型进行词语对齐。该模块的输入为双语平行语料，输出是双语句对的 Viterbi 对齐结果。Giza++ 算法仿照 Och 实验中的训练顺序，先使用 IBM 模型 1 和模型 2 进行初步训练，然后使用 HMM 模型和 IBM 模型 4 进行训练。

(3)对齐短语抽取：采用基于词语对齐结果扩展的对齐短语抽取算法。该模块的输入为双语双向词语对齐结果，输出为短语互译译表和短语在句子中的位置。采用 Moses 系统[2)] 来对齐短语[16]。

(4)翻译规则抽取：输入为短语对齐表，输出为带有句法标记的翻译规则及相应的特征函数值。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。

(5)翻译规则过滤：根据约束条件过滤翻译规则，加速解码过程。该模块的输入为翻译规则和测试集，输出为只针对该测试集的翻译规则。

(6)最小错误率训练：利用汉英参考集调整翻译模型中各个特征函数的权值进行最大熵模型的训练。最小错误率训练采用 Och 的算法思想[17]。模块的输入为参考集、英语的语言模型、翻译规则集和解码器的最初 K-Best 翻译结果列表，输出为各个特征函数的权值。

(7)解码：参照 Chiang[8]，但采用 CYK$^+$ 算法作为解码算法，同时利用两轮解码算法在解码过程中整合语言模型。该模块的输入为待翻译的汉语句子，输出为该汉语句子对应的 1-Best 或 K-Best 英语译句。实验中采用的语言模型工具为 Srilm1. 5. 5 版[18]，此外使用了 LDC 免费 Web-1TB 三元语言模型语料。

(8)自动评测：机器翻译自动评测工具采用的是 NIST 的 mt-evaluation1. 1 版[3)]，及 BLEU-4 元语言模型测。

下面介绍扩展的 HMM 模型、翻译规则的抽取和过滤、CYK$^+$ 解码算法和两轮解码算法。

3 扩展 HMM 模型

扩展 HMM 模型改进基本 HMM 的对齐，使得对齐概率与两个源语言串词之间的对齐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，见式(1)。

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

式中，$i=a_j$ 表示第 j 个源语言词与 i 个目标语言词对齐。$k=a_{j-1}$ 表示第 $j-1$ 个源语言词与 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[2]。$\lambda_1+\lambda_2=1$。$t(i,k)$ 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(详见文献[19])。分母都是归一化因子。

4 翻译规则抽取

翻译规则的抽取过程：

(1)利用对齐短语创建基本翻译规则。基本翻译规则由句法标记、源语言词串和目标语言词串构成。对于输入的对齐短语，查找对齐短语对应的目标语言的句法标记。如果查找成功，就将这个对齐短语扩展为基本翻译规则，如果该对齐短语目标端没有对应的句法标记，采用以下策略为该对齐短语寻找对应的扩展句法标记。

(a)首先，判断目标短语是否对应两个或者两个以上的句法标记，若存在，则可以将所有子短语的句法标记合并为该短语的句法标记。

(b)其次，尝试采用 C_1/C_2 标记对齐短语。C_1/C_2 是一个右侧不完整的句法成分，其对应的完整句法成分是 C_1。C_1/C_2 需要与右侧的 C_2 组合才可以形成句法成分 C_1。

(c) 最后，尝试采用 $C_2\backslash C_1$ 标记对齐短语。$C_2\backslash C_1$ 是一个左侧不完整的句法成分，其对应的完整句法成分是 C_1。$C_2\backslash C_1$ 需要与左侧的 C_2 组合才可以形成句法成分 C_1。

(2)利用基本翻译规则构造组合翻译规则。采用 Chiang[8] 的推导原理来获得组合翻译规则。

$$N \rightarrow f_1 \cdots f_m / e_1 \cdots e_n \tag{2}$$

$$M \rightarrow f_i \cdots f_u / e_j \cdots e_v \tag{3}$$

式(2)和式(3)是两个基本翻译规则，若 $i \geqslant 1, u \leqslant m, j \geqslant 1, v \leqslant n$，并且这 4 个等式不会同时相等，则由基本翻译规则式(2)和式(3)推导出组合翻译规则：$N \rightarrow f_1 \cdots f_{i-1} M_k f_{u+1} \cdots f_m / e_1 \cdots e_{j-1} M_k e_{v+1} \cdots e_n$。

利用源语言和目标语言的起始位置和终止位置可以确定一个矩形，标明该翻译规则的双语串上的跨度。

5 CYK$^+$ 算法

Chiang 的层次短语模型采用基于 CYK(Cocke-Younger-Kasami)算法的解码器[8]。CYK 算法要求重写规则属于乔姆斯基范式，而本文产生的翻译规则不完全属于乔姆斯基范式。一般的解决方法是先将翻译规则转换为乔姆斯基范式，这个转换会导致翻译规则集中非终结符数量的大量增加。解码的计算复杂度是关于翻译规则集非终结符数量的多项式函数，非终结符数量的增加严重降低了解码器的计算速度。为解决这个问题，我们使用改进的 CYK 算法——CYK$^+$ 算法。

CYK$^+$ 算法是在 CYK 算法的基础上，构造当前二维矩阵

1) http://www.fjoch.com/GIZA++.html

2) http://www.statmt.org/moses/

3) http://www.nist.gov/speech/tools/

Chart[i][j]元素，使得 Chart[i][j]包含所有可能形成的短语的非终结符的集合。同时仿照 Earley 算法，使得 Chart[i][j]还包含所有不完整假设(点规则)。这种改进使得 CYK$^+$ 算法可以处理非乔姆斯基范式的翻译规则。

CYK$^+$ 算法的数据结构为二维矩阵{Chart[i][j]}，在不考虑语言模型的前提下，设解码器的输入串为 w_1^I，Chart 表中每个表项 Chart[i][j]对应于输入句子中某一个跨度(Span)上所有可能形成的短语的非终结符的集合和所有不完整假设。其中，i 表示该跨度左侧第一个词的位置，j 表示该跨度包含的词数。

5.1 CYK$^+$ 算法

(1)初始化过程：对于输入源语言句子的所有位置 i 和所有跨度 j，如果源语言存在如下翻译规则 $X\rightarrow w_i\cdots w_{i+j-1}$，则将 X 添加到 Chart[i][j]的第一个列表中。

(2)匹配过程：对于给定的 Chart[i][j]，先后执行以下两种操作

(a)对所有可能的 Chart[i][k]和 Chart[$i+k$][$j-k$]，如果 *Chart*[i][k]存在一个不完整假设 α，而 Chart[$i+k$][$j-k$]存在一个完整假设 β，且 α 和 β 满足以下翻译规则：$A\rightarrow\alpha\beta\gamma$，如果 γ 为空，则在 Chart[i][j]第一个列表中添加非终结符 A；如果 γ 不为空，在 Chart[i][j]第二个列表中添加非终结符 $\alpha\beta$。

(b)对于 Chart[i][j]，对于第一个列表的每个非终结符，如果存在 $A\rightarrow\beta\gamma$，且 γ 为空，则在 Chart[i][j]第一个列表中添加非终结符 β；如果 γ 不为空，在 Chart[i][j]第二个列表中添加非终结符 β。

(3)重复执行步骤(2)，直到 Chart[1][n]构造完毕。

利用动态规划算法，可以在多项式时间内构造上述 Chart 二维矩阵。

CYK$^+$ 算法与 CYK 算法相比，时间复杂度相同，都是 $O(n^3)$，n 为句子长度。但 CYK$^+$ 算法能处理非乔姆斯基范式的翻译规则，而 CYK 算法不能处理非乔姆斯基范式的翻译规则。

5.2 两轮解码算法

在解码算法中，解码过程中的每个假设的状态没有考虑当前假设对应的目标语言模型的状态。原因在于，上述 CYK$^+$ 算法依赖的等价假设的条件在考虑语言模型的前提下不成立，即使两个假设的源语言串一致且目标语言的句法标记一致，这两个假设对应的目标语言模型也可能不一致，因此无法计算这两个假设的得分。虽然不引入语言模型仍然可以进行计算，但是目标语言模型对于翻译质量的提高具有重要作用，因此整合语言模型是当前基于句法的统计机器翻译解码器需要处理的关键问题。

本文采用两轮解码算法。第一轮算法在计算过程中加入可以快速计算的近似语言模型评分，这样第一轮可以快速构造对应 Chart 表，找到近似最优的 K-Best 翻译结果；第二轮算法根据第一轮算法的 K-Best 翻译结果反向搜索该 Chart 表，利用语言模型得分作为启发式函数进行搜索，对第一轮的翻译结果进行修正。这种算法的出发点在于第一轮算法可以根据翻译规则和近似目标语言模型获得近似正确的 K-Best 派生序列，翻译结果的错误基本出现在目标语言模型部分，因此第二轮算法利用语言模型评分进行修正，最终的结果应该近似于在解码过程中完全采用目标语言模型的最终结果(详见文献[19])。

6 实验及结果分析

实验采用双语平行训练语料，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 评分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。

基准翻译参照系统是 Kohen 的基于短语的翻译系统 Moses 和 Chiang 的层次短语模型翻译系统 Joshua。

6.1 扩展 HMM 模型的词语对齐质量

实验中采用了两种词语对齐模块，一个是 Giza++模块，Giza++实现 IBM 模型 4 和基本 HMM 对齐；一个是扩展 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 种词语对齐结果对翻译系统的影响见表 1。基本 HMM 比 IBM 模型 4 在训练集和测试集的 BLEU 评分都高，说明基本 HMM 的词语对齐质量高于 IBM 模型 4。扩展 HMM 模型比 IBM 模型 4 和基本 HMM 模型的 BLEU 评分都高，约能提高 0.5～1 的 BLEU 得分，说明扩展 HMM 模型比 IBM 模型 4 和基本 HMM 模型的词语对齐质量都高。

表 1 3 种词语对齐模型对翻译系统的影响

BLEU 得分	训练集	测试集	平均值
IBM 模型 4	26.05	25.62	25.84
基本 HMM	26.44	25.89	26.17
扩展 HMM	26.92	26.52	26.72

6.2 翻译规则对统计机器翻译的影响

翻译规则的首要目的是处理全局短语排序问题，而短语模型只能处理局部排序。表 2 比较了短语模型、层次短语模型和翻译规则的机器翻译质量。对于基于短语的翻译模型，最大短语排序长度为 12(reo=12)。本实验采用所有 Web 1T 中的 3 元模型和训练语料的语言模型。从表 2 可以看出，层次短语模型系统的 BLEU 评分在训练集、测试集和平均值上都高于短语模型系统的 BLEU 评分，说明层次短语模型机器翻译系统质量高于短语模型机器翻译系统质量。扩展层次短语模型系统的 BLEU 评分在训练集、测试集和平均值上都高于短语模型和层次短语模型系统的 BLEU 评分，说明扩展层次短语模型机器翻译系统质量高于短语模型或层次短语模型机器翻译系统质量。

表 2 3 种翻译模型的 BLEU 评分比较

BLEU 得分	训练集	测试集	平均值
短语模型(reo=12)	26.39	25.96	26.18
层次短语模型	26.78	26.13	26.46
扩展层次短语模型(翻译规则)	26.92	26.52	26.72

6.3 两种解码算法对扩展层次短语模型的 BLEU 影响

表 3 给出了立方剪枝算法和两轮解码算法的 BLEU 评分比较。对于扩展层次短语模型，立方剪枝算法[8]和两轮解码算法的 BLEU 评分相近，这说明这两个算法翻译质量接近，但两轮解码算法的解码速度比立方剪枝算法快。

表 3　两种算法的结果比较

BLEU 得分	参考集	测试集	平均值
立方剪枝算法	26.92	26.52	26.72
两轮解码算法	26.89	26.51	26.70

结束语　本文把串-树模型和层次短语模型的优点相结合实现了一个比较完整的统计机器翻译系统。首先词语对齐结果和句法约束冲突影响了翻译规则的质量和翻译系统的性能。本文提出基于短语结构树距离的 HMM 模型用于词语对齐，这种模型可以有效地减少词语对齐结果和句法约束现象，提高基于句法的翻译系统的性能。其次提出基于短语结构树的层次短语模型，这是利用串-树模型的思想对 Chiang 的层次短语模型的扩展。本文在双语对齐短语的基础之上结合英语短语结构树抽取翻译规则，并提出启发式规则用以获得翻译规则的扩展句法标记。这种扩展模型相比于层次短语模型在不同规模数据集上的性能更稳定，评测结果较高。此外在解码过程已给出可以处理非乔姆斯基范式翻译规则的 CYK$^+$ 算法，并给出了两轮解码算法来整合语言模型，这种算法与立方剪枝算法具有相近的解码质量，但是解码速度更快。

参 考 文 献

[1] Brown P F, Della Pietra S A, Della Pietra V J. The mathematics of statistical machine translation parameter estimation[J]. Computational Linguistics, 1993, 19(2): 263-311

[2] Vogel S, Ney H, Tillmann C. HMM-based word alignment in statistical translation[C]// the 16th International Conference on Computational Linguistics Proceedings. 1996: 836-841

[3] Och F J, Ney H. A systematic comparison of various statistical alignment models[J]. Computational Linguistics, 2003, 29(1): 19-51

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

[5] Och F J. The Alignment Template Approach to Statistical Machine Translation[J]. Computational Linguistics, 2004, 30(4): 417-449

[6] Och F J, Ney H. Discriminative training and maximum entropy models for statistical machine translation[C] // 40th Annual Meeting of the Association for Computational Linguistics. Philadelphia, 2002: 295-302

[7] Wu De-kai. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora[J]. Computational Linguistics, 1997, 23: 377-404

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

[9] Chiang D. Learning to translate with source and target syntax [C]// Proceedings of ACL 2010. 2010: 1443-1452

[10] Yamada K, Knight K. A syntax-based statistical translation model[C]// 39th Annual meeting of the Association for Computational Linguistics and 10th Conference of the European Chapter of ACL 2001. Toulouse, France, 2001: 523-530

[11] Galley M, Hopkins M, Knight K, et al. What's in a translation rule[C]// Human Language Technology Conference and North American Chapter of the Association for Computational Linguistics annual meeting 2004. Boston, USA, 2004: 273-280

[12] Galley M, Graehl J, Knight K, et al. Scalable inference and training of context-rich syntatic translation models[C] // Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics 2006. Sydney, 2006: 961-968

[13] Liu Yang, Liu Qun, Lin Shou-xun. Tree-to-string alignment template for statistical machine translation[C]// Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics 2006. Sydney, 2006: 609-616

[14] Zhang J, Zhai F, Zong C. Augmenting string-to-tree translation models with fuzzy use of source-side syntax[C]// Proceedings of EMNLP 2011. 2011: 204-215

[15] Melamed I D. Multitext grammars and synchronous parser [C]// International Conference Combining Human Language Technology Conference Series and the North American Chapter of the Association for Computational Linguistics Conference series 2003. Edmonton, Canada, 2003

[16] Koehn P, Hoang H, Birch A, et al. Moses: Open source toolkit for statistical machine translation [C] // Proceedings of ACL2007. 2007: 177-180

[17] Och F J. Minimum error rate training in statistical machine translation[C] // 41st Annual Meeting of the Association for Computational Linguistics 2003. Sapporo, Japan, 2003

[18] Stolcke A. SRILM-An Extensible Language Modeling Toolkit [C]// Proceedings of International Conference of Spoken Language Processing 2002. Denver, Colorado, 2002

[19] 姜巍. 短语结构树在汉英统计机器翻译中的应用研究[D]. 北京：清华大学，2009

(上接第 213 页)

[16] Keele B F, Giorgi E E, Salazar-Gonzalez J F, et al. Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection[J]. Proc. Nat 1. Acad. Sci. USA, 2008, 105(21): 7552-7557

[17] Miao Hong-yu, Xia Xiao-hua, Perelson A S, et al. On Identifiability of Nonlinear ODE Models and Applications in Viral Dynamics[J]. Society for Industrial and Applied Mathematics, 2011, 53(1): 3-39

[18] Miao Hong-yu, Xia Xiao-hua, Perelson A S, et al. On Identifiability of Nonlinear ODE Models and Applications in Viral Dynamics [J]. Society for Industrial and Applied Mathematics, 2011, 53(1): 3-39

[19] Neumann A V, Lam N P, Dahari H, et al. Hepatitis C Viral Dynamics in Vivo and the Antiviral Efficacy of Interferon-α Therapy[J]. Science, 1998, 282: 103-107

[20] Bakare G A, Venayagamoorthy G K, Aliyu U O. Reactive power and voltage control of the Nigerian grid system using micro-genetic algorithm[C]// IEEE Power Engineering Society General Meeting. 2005: 1916-1922