文章编号: 1003-0077(2007)05-0087-04

基于分层语块分析的统计翻译研究

魏 玮[1],杜金华[1],徐 波[1,2]

(1. 中国科学院自动化研究所 数字内容技术研究中心,北京 100080;
2. 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)

摘 要: 本文描述了一个基于分层语块分析的统计翻译模型。该模型在形式上不仅符合同步上下文无关文法,而且融合了基于条件随机场的英文语块分析知识,因此基于分层语块分析的统计翻译模型做到了将句法翻译模型和短语翻译模型有效地结合。该系统的解码算法改进了线图分析的 CKY 算法,融入了线性的 N-gram 语言模型。目前,本文主要针对中文 - 英文的口语翻译进行了一系列实验,并以国际口语评测 IWSLT(International Workshop on Spoken Language Translation)为标准,在 2005 年的评测测试集上,BLEU 和 NIST 得分均比统计短语翻译系统有所提高。

关键词: 人工智能;机器翻译;基于分层语块分析的统计翻译模型;条件随机场;CKY 算法
中图分类号: TP391 **文献标识码**: A

Statistical Machine Translation Model Based on Hierarchical Chunking Phrase

WEI Wei[1], DU Jin-hua[1], XU Bo[1,2]

(1. Digital Media Content Technology Research Center, Institute of Automation, Chinese Academic of Sciences, Beijing 100080, China; 2. National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academic of Sciences, Beijing 100080, China)

Abstract: This paper describes a Hierarchical chunking-phrase based (HCPB) statistical translation model. The model not only comply with formal synchronous context-free grammar but also learned partial parsing knowledge using CRF (Conditional Random Fields). Therefore it can be taken as combination of fundamental ideas from both syntax-based translation and phrase-based translation. The decoder for HCPB MT system is based on Chart-CKY algorithm, and integrates N-gram language model effectively. In our benchmark evaluation focusing on Chinese-English spoken language translation. The method achieves higher accuracy in measure of Bleu and NIST score in IWSLT2005.

Key words: artificial intelligence; machine translation; hierarchical chunking-phrase based SMT; conditional random fields; chart-based CKY algorithm

1 引言

统计翻译是通过寻找最大似然路径,将源语言 $f_1^J = f_1, f_2, \cdots f_J$ 翻译成目标语言,如英语 $e_1^I = e_1, e_2, \cdots, e_I$:

$$\hat{e}_1^I = \arg\max_{e_1^I} \Pr(e_1^I \mid f_1^J) \quad (1)$$

$$= \arg\max_{e_1^I} \Pr(f_1^J \mid e_1^I) \Pr(e_1^I) \quad (2)$$

在公式(2)中,信源信道模型被独立地分解为翻译模型和语言模型[1]。前者代表了两种语言之间的对应关系,后者代表了英文的流利度。

基于短语的翻译模型是当前统计翻译的主流,它把短语,即连续的词串作为翻译模型的最小单位[2~4]:它假设输入 e_1^I 被分解为 K 个连续短语 $\bar{e}_1^K$,每个短语 $\bar{e}_k$ 被翻译成 $\bar{f}_k$,翻译结果再进行重组形成 f_1^J。相比于基于词的翻译模型,短语翻译模型的优势在于它可以利用短语内部的语序信息进行局部的调

收稿日期: 2007-04-30 **定稿日期**: 2007-06-27
基金项目: 国家 863 计划资助项目(2006AA01Z194);富士通合作项目(K0604040)

作者简介: 魏玮(1982—),女,博士生,主要研究方向为自然语言处理,机器翻译;杜金华(1976—),男,博士生,主要研究方向为自然语言处理,机器翻译;徐波(1966—),男,研究员,博士生导师,研究方向为语言识别,机器翻译,中文信息处理等。

序。但是,对于大范围的短语之间的顺序问题,它无能为力。虽然近年来,众多研究者试图加入多种调序模型来弥补短语模型的不足,但效果都不是很理想。这是因为短语重排序的模型一般是根据词的位置进行跳转[2],无法用到更多的句子结构信息。

Chiang[5]引入了分层短语模型的概念,它有效结合了短语模型和同步CFG文法[6]:由根节点开始每次同时生成一对子串,该子串最多包含两个非终结符。该模型可以看作是在短语翻译对中加入了相应变量,可以完成长距离短语对的翻译。因此分层模型克服了传统短语翻译模型的调序问题,但是分层短语模型的规则提取仍然沿用了在双语对齐语料中抽取相容短语的方法,并通过在大短语中找到包含的子短语来实现。由于没有任何句法信息约束,分层短语模型抽取的语法规模往往十分庞大,极大地影响之后翻译解码的质量和效率。

本文提出了基于分层语块分析的概念,它沿用了Chiang[5]的分层短语模型的形式化机制,并且结合了英文浅层句法分析的手段,使其在一定程度上去除了存在的冗余信息,真正做到了句法信息和统计翻译相结合。另外,本文改进了线图的CKY解码算法,提高了搜索质量和效率。本文其余部分是如下安排的:第二节给出了基于CRF语块分析的分层短语模型;第三节介绍了改进的CKY-Style解码算法;第四节为针对中文-英文口语翻译评测的实验结果;最后一节结论。

2 基于CRF语块分析的分层短语模型

在传统的短语模型中,短语是指和统计词语对齐相容,并在相邻的词语之间进行抽取的词串,通常称之为统计短语。Chiang的分层短语模型,虽然加入了同步上下文无关文法的约束,但在短语的抽取上仍然沿用了统计短语的获取方法,没有任何的句法结构意义,缺乏语法信息的约束。单纯的统计方法获取的短语在规模上大约是S^2(S是句子的长度),这其中存在大量的冗余信息,在此基础上抽取的分层短语存在变量过渡替换,规模骤增的问题,对之后翻译解码的质量和效率带来不便。因此,本文引入了基于CRF的语块分析方法,并在此基础上建立分层短语模型。

2.1 基于条件随机场(Conditional Random Fields, CRF)的语块分析

语块分析,也称作浅层句法分析或部分句法分析(Partial Parsing)。它主要是识别句子中某些结构相对简单的独立成分。语块分析使句法分析的任务在某种程度上得到简化,同时也利于句法分析技术在大规模真实文本处理系统中迅速得到应用。Lafferty *et al*.[7]提出了CRF的概念,随后便被广泛地应用在模式识别各个领域,CRF还被用作名词实体的识别,生物基因序列信息的识别等许多自然语言处理领域。CRF模型描述如下。

给定的输出标识序列Y和观测序列X,为了描述(X, Y)序列对上的CRF,定义特征函数$f_j(y_{i-1}, y_i, x, i)$和权值向量λ,y_{i-1}, y_i为标识序列,x为输入序列,i为输入位置。则

$$p(y \mid x, \lambda) = \frac{1}{Z(x)} \exp\left(\sum_j \lambda_j F_j(y, x) \right) \quad (3)$$

$$F(y, x) = \sum_{i=1}^{n} f(y_{i-1}, y_i, x, i) \quad (4)$$

其中$Z(x)$为归一化系数。

由上式求得条件随机场的条件概率,对于输入序列x,最佳序列y可以通过下式确定:

$$\hat{y} = \arg\max_y P(y \mid x) = \arg\max_y \sum_j \lambda_j \cdot F_j(y, x) \quad (5)$$

沿用Sha和Pereira[8]提出base NP识别特征函数建立的方法,并将其扩展到Base Phrase(NP, VP, PP)的识别。y_i连续的标识序列为$y_{i-1} = c_{i-2}c_{i-1}, y_i = c_{i-1}c_i$,特征函数$f_j(y_{i-1}, y_i, x, i) = p(x, i)\, q(y_{i-1}, y_i)$。$p(x, i)$用来预测在当前位置$i$的输入标识$x$,$q(y_{i-1}, y_i)$预测输出BP标识对。通过训练语料可以获得大量的序列特征,再经过训练可以得到BP序列上的CRF模型。CRF训练方法有共轭梯度方法,最小记忆类牛顿方法和投票感知器方法等。实验中选取CRF++①作为CRF分类器,并把CONLL2000②的公共语料(英文宾州树库II中WSJ部分)训练语块分析器。

2.2 基于语块分析的分层短语模型

加权同步上下文无关文法的产生规则中包含两个对齐的串[6],其语法规则如下:

$$X -> (r, a, \sim) \quad (6)$$

其中,X是非终结符,r和a是包含终结符或非终结符的串,$\sim$是终结符之间的对应关系。对应于本文的中英翻译系统,r代表中文,a代表英文。为

① http://chasen.org/~taku/software/CRF++

② http://www.cnts.ua.ac.be/conll2000/chunking/

了简化,本文规定 r, a 中最多包含两个不相邻的非终结符,分别用 $x1$, $x2$ 表示。~代表 r 和 a 中终结符之间的对应关系。因此基于语块的分层短语模型生成过程为:

1) 利用 Giza++ 生成统计词语对齐,并利用启发算法生成和词语对齐相容的大量统计短语集(BaseP)。

2) 利用 CRF 对训练语料进行英文语块分析,得到 NP,VP,PP 集(ESynP)。

3) 得到 SynP{SynP ∈BaseP 如果 BaseP -> ephrase ∈ESynP}

4) BaseP 中的短语对组成基本语法(Base Rule): $X -> (f_{Basep}, e_{Basep})$

5) 同步规则(r, a)和 SynP 中的短语对(f_{SynP}, e_{SynP})满足 $r = rf'_{SynP} r'', a = a' e_{SynP} a''$,则形成结构语法(Syntactic Rule): $X -> (r' f_{SynP} r'', a' e_{SynP} a'')$

6) 两条生成语法: $S -> (S X, S X)$ $S -> (X, X)$

同样地,为了减少冗余信息,降低解码的复杂度,本文也对语法的生成端进行限制:

1) r 和 a 最少要包含一个相对齐的终结符;

2) r 不能有相邻接的终结符。

3　改进的 CKY-Style 解码器

传统的统计翻译系统,如 Pharaoh,采用 beam search,A* 等线性解码算法,虽然可以融入简单的调序模型和 N-gram 的语言模型,但翻译结果的语序总是差强人意。然而本文的 HCPB 翻译系统前端结合了树的结构信息,后端的解码算法相应也要利用树的分析算法。所以本文借鉴基于线图分析的句法分析方法,实现了 CKY-Style 的解码器。CKY 算法是改进的自底向上移进 - 规约算法。在解码过程中会产生大量的假设,为了避免搜索所有的可能,本文采用了堆栈结构,并利用不同的策略在搜索过程中进行必要的剪枝。

整个解码器的实现原理可见图 1。本文是在源语言一边进行自底向上规约。为了和代码一致,本文把原句中从 ith 词到 jth 词的假设称为 $edge_{i,j}$。考虑一个中文输入 s,首先找到 s 中包含的所有 Base Rule(右边不含 X 的同步 CFG Rules),初始化 Chart,这些由 $Initialize(s)$ 完成。然后以每一个 $edge_{start,end}$ 为中心,向其四周扩展,形成新的 $edge_{start1,end1}$,通过 $addedge$ 压入栈中。该算法保证了从下向上不断进行规约,最后到整个句子。

```
Decoder( sentence s)
Initialize(s)
n = length(s)
for span = 2 to n do
   for start = 1 to n - span + 1 do
      end = start + span - 1
      Extend_edge(start ,end)
Initialize( s)
For 所有 s 中包含的终结符(右边不含 X 的 Rules) do
   Generate a new edge e
   addEdge(e)
Extend_edge( start ,end)
for 以 edge_start. end 为中心向四周扩展,形成新的 edge_start1 ,end1
   if edge_start1 ,end1 符合同步 CFG
      addEdge(edge_start1 ,end1 )
addEdge( e)
If e. score < bestCurrentScore[e. start ,e. end] - beam then        / * Thresholding pruning */
   Discard e and return
If there is a matching edge e# in chart[e. start ,e. end] then      / * Recombination */
   If e. score > e# . score then
      Replace e# with e and return
   Else
      Discard e and return
If | chart[e. start ,e. end]| > b then                              / * Histogram pruning */
   If e. score > e## . score then
      Replace e## with e and return / * e## ,is the worst edge in chart[e. start , e. end] */
   Else
      Discard e and return
```

图 1　CKY-Style decoder 核心算法

基本的CKY算法必须满足自底向上逐步规约,也就是说产生$edge_{i,j}$的前提是所有$subedge_{i',j'}$,$i \leq i' \leq j' \leq j$的边都必须产生,它的时间复杂度为$O(n^5)$,n是句子的长度。而在HCPB翻译过程中,只关心X变量四周的短语,解码器不再需要遍历所有的$subedge$。仅需要以已经产生的$edge$为中心向四周扩展,故时间复杂度是$O(n^2 * n^{|x|})$,其中$|x|$为Syntactic Rule中含有的变量的个数。由于限制了变量个数最大为2,故算法复杂度为$O(n^3 \sim n^4)$。

4 实验结果及分析

目前,本文主要针对中文-英文的口语翻译评测IWSLT2005(International Workshop on Spoken Language Translation)进行了一系列实验,并且以自动评分BLEU作为评价标准①。

4.1 IWSLT2005语料

本文利用IWSLT2005提供的20 000双语对齐语料作为短语抽取和英文3-gram语言模型的训练语料,其中包含176 199个中文词和183 452个英文词。测试集是506句,含3 743中文词。

4.2 实验结果

本文按照Koehn[2]的方法训练词语对齐。首先运行中文-英文和英文-中文两个方向的GIZA++②,之后应用"grow-diag-final"的启发函数在每一个相交的对齐点上,然后根据Och[9]的连续相容原理抽取最大长度为8的基本短语对(Base Phrase),利用CRF语块分析方法找到语块对[10],并且生成右端不含变量的基本规则(Base Rule)和包含变量的语法规则(Syntactic Rule),最后,通过统计得到各个规则的频率概率和词汇化概率。

表1中PB是短语翻译系统(Phrase-Based MT),HPB是分层短语翻译系统(Hierarchy Phrase-Based MT),HCPB是基于语块分析的分层短语翻译系统(Hierarchical Chunking-Phrase Based)。规则数量包含两部分:Base Rules和Syntactic Rules。

表1 IWSLT2005实验结果

Sysid	规则数量:	BLEU	NIST	时间(分钟)
PB	195 846	0.468 2	8.376 1	4
HPB	135 339/881 121	0.469 1	9.139 7	13
HCPB	85 430/430 120	0.450 1	8.865 3	6
HCPB+Dict	101 765(词典)/430 120	0.469 4	9.039 8	5

4.3 结果分析

从IWSLT2005的测试结果来看,几个系统的BLEU打分基本一致,而分层模型(HPB,HCPB)在NIST打分上有明显优势,主要有以下几点原因:

1) PB系统能够准确地翻译出原句中各个短语,但翻译结果的语序是最大的问题。这是由于PB系统利用大规模的短语模型,在解码时仅依赖简单的调序模型和N-gram语言模型,缺乏调整Base Rule之间语序的能力。而分层模型系统的翻译模型包含了大量的带变量的Syntactic Rule,这些带变量的短语包含了Base Rule之间的调序信息,在解码的时候自底向上利用Syntactic Rule不断扩展,最后规约到根节点。

2) 分层模型的翻译结果更接近标准答案的平均句长,在计算NIST得分的时候更具优势。从具体翻译结果中可以看出PB系统翻译过程倾向于利用小的短语,不能很好利用长短语信息,因此不能照顾到句子的整体结构;而分层模型系统引进了带变量的短语,可以利用更多长短语结构信息,使得翻译出来的句子从可懂度上更好一些。

5 结论

本文采用了一种统计和句法相结合的方法,将语言的形式化结构和语块分析技术相融合,有效地去除统计系统中的冗余信息,开发出一套适用于中文-英文的翻译系统,并显著提高了翻译结果的效率和可懂度。HCPB系统主要针对口语翻译进行了一系列对比实验。实验结果说明利用语言的形式化结构可以克服短语翻译的调序问题,进一步加入浅层句法分析的知识可以过滤得到真正有效的短语对,避免变量的过渡替换。如何利用该系统完成大规模新闻语料的训练和自动翻译,是下一步工作的重点。

(下转第117页)

① http://www.is.cs.cmu.edu/iwslt2005/eval_servers.html

② http://www.fjoch.com/GIZA++.html

参考文献:

[1] Chris Callison-Burch, Philipp Koehn, and Miles Osborne. Improved statistical machine translation using paraphrases [A]. In: Human Language Technology Conference[C]. 2006.

[2] Liang Zhou, Chin-Yew Lin, and Eduard Hovy. Re-evaluating machine translation results with paraphrase support [A]. In: EMNLP conference[C]. 2006.

[3] D Kauchak and R Barzilay. Paraphrasing for automatic evaluation [A]. In: HLT-NAACL[C]. 2006.

[4] Regina Barzilay. Information Fusion for Multi-document Summarization: Paraphrasing and Generation [D]. PhD thesis, Columbia University, 2003.

[5] I. Ali, K. Boris. Extracting structural paraphrases from aligned monolingual corpora [A]. In: IWP[C]. 2003.

[6] Satoshi Sekine. On-demand information extraction [A]. In: BANNARD G/ ACL[C]. 2006. 731-738.

[7] R. Barzilay and K. McKeown. Extracting paraphrases from a parallel corpus [A]. In: ACL[C]. 2001.

[8] K. Ohtake and K. Yamamoto. Applicability analysis of corpus-derived paraphrases toward example based paraphrasing [A]. Language, Information and Computation Proceedings[C]. 2003.

[9] B. Pang, K. Knight, and D. Marcu. Syntax-based alignment of multiple translations: Extracting paraphrases and generating new sentences [A]. In: HLT/ NAACL[C]. 2003.

[10] Dekang Lin and Patrick Pantel. Dirt-discovery of inference rules from text [A]. ACM SIGKDD[C]. 2001.

[11] H. Wu and M. Zhou. Synonymous collocation extraction using translation information [A]. ACL[C]. 2003.

[12] Mona Diab and Philip Resnik. An unsupervised method for word sense tagging using parallel corpora [A]. ACL[C]. 2002.

[13] Bannard C. and Callison-Burch C. Paraphrasing with bilingual parallel corpora [A]. ACL2005[C]. 597-604.

[14] Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment methods [J]. Computational Linguistics, 2003, 29:19-51.

(上接第 90 页)

参考文献:

[1] Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Pobert L. Mercer. 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: Proc. of NAACL [C]. Edmonton, Canada: 2003. 48-54.

[3] Richard Zens and Hermann Ney. A comparative study on reordering constraints in statistical machine translation[A]. In: Proc. of ACL 2003[C]. 144-151.

[4] Christoph Tillman. A unigram orientation model for statistical machine translation[A]. In: HLT-NAACL Short Papers [C]. Boston, Massachusetts, USA: 2004. May 2 - May 7, 101-104.

[5] David Chiang. A hierarchical phrase-based model for statistical machine translation[A]. In: Proc. of ACL 2005[C]. Ann Arbor, Michigan: June, 263-270.

[6] Alfred V. Aho and Jeffrey D. Ullman. Syntax directed translations and the pushdown assembler[J]. J. Comput. Syst. Sci., 1969, 3(1):37-56.

[7] J. Lafferty A. McCallum and F. Pereira. Conditional random Fields: probabilistic models for segmenting and labeling sequence data [A]. Harry Q. Bovik. Proceedings of ICML [C]. Massachusetts, USA: 2001. 282-289.

[8] Fei Sha and Fernando Pereira. Shallow Parsing with Conditional Random Fields [A]. Eduard Hovy. Proceedings of HLT-NAACL [C]. Edmonton, Alberta: 2003. 134-141.

[9] F. J. Och and H. Ney. Discriminative training and maximum entropy models for statistical machine translation[A]. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistic[C]. 2002. 295-302.

[10] Fang Xu, Chengqing Zong, and Jun Zhao. A Hybrid Approach to Chinese Base Noun Phrase Chunking [A]. In: Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing [C]. Sydney: July 22-23. 2006. 87-93.