基于短语的统计机器翻译中短语抽取算法改进

强 静[1,2],张 建[1]
QIANG Jing[1,2],ZHANG Jian[1]

1.中国科学院 合肥智能机械研究所,合肥 230031
2.中国科学技术大学 信息科学技术学院,合肥 230027
1.Institute of Intelligent Machines,Chinese Academy of Sciences,Hefei 230031,China
2.School of Information Science and Technology,University of Science and Technology of China,Hefei 230027,China
E-mail:qjng@mail.ustc.edu.cn

QIANG Jing,ZHANG Jian.Improving phrase-based statistical translation by modifying phrase extraction algorithm. Computer Engineering and Applications,2008,44(13):147-149.

Abstract: The paper proposes an improved algorithm of Phrase Extract based on the Och's phrase extraction algorithm in the phrase based statistical machine translation.The algorithm can take more accurate alignment information based on the original algorithm.It is of great significance for the smaller corpus statistical machinery.It can reduce the unknown words by increasing in correct alignment information,and increases the rate of correct translation.After the different scale corpus experiment.The extracted number of phrase is obviously increase.

Key words: machine translation;translation model;phrase extract

摘 要:针对基于短语统计机器翻译中目前常用的Och提出的短语抽取算法,提出了一种改进算法。该算法能够在原有算法的基础上抽取出更多的准确对齐信息,这对语料库较小的汉民统计机器来说意义重大,增加正确的对齐信息可以减少未登录词的产生,提高翻译正确率。经过对不同规模语料库的实验,抽取的短语对数目有明显增多。

关键词:统计机器翻译;翻译模型;短语抽取
DOI:10.3778/j.issn.1002-8331.2008.13.044 **文章编号**:1002-8331(2008)13-0147-03 **文献标识码**:A **中图分类号**:TP311

短语抽取是基于短语的统计机器翻译中构建翻译模型的关键步骤,抽取短语对的对齐数目和对齐准确度决定着翻译模型的质量,从而也决定着翻译的准确率。目前,基于短语的统计机器翻译是最成熟的机器翻译方法之一。短语翻译模型是基于对数线性模型思想中唯一必需的模型。在短语翻译模型中存储着大量源语言到目标语言短语对的翻译概率信息。不同的短语翻译系统采用不同的方式构建短语翻译模型,但它们共同的目标是较高的短语对齐数和短语对齐精度。

针对汉民多语言农业知识处理平台中的汉民翻译,本文采用基于短语的统计机器翻译方法,在原有Och短语抽取算法的基础上,提出了一种改进的短语抽取算法。该算法能够抽取更多正确的短语对齐信息,提高汉民小语料库的利用效率,减少翻译系统中未登陆词的出现概率,提高翻译准确率。

1 基于短语的统计机器翻译

统计机器翻译认为翻译过程就是从给定源语言句子$f^J=f_1\cdots f_i\cdots f_J$的所有目标语言句子$e^I=e_1\cdots e_i\cdots e_I$中寻找概率最大的句子$\hat{e}^I$作为最佳译文。最初的信道模型[1]中将翻译过程表示为:

$$\hat{e}^I=\arg\max_{e^I}\{\Pr(e^I|f^J)\}= \tag{1}$$

$$\arg\max_{e^I}\{\Pr(e^I)\Pr(f^J|e^I)\} \tag{2}$$

$\Pr(e^I)$为目标语言模型,反映目标语言句子的质量;$\Pr(f^J|e^I)$为翻译模型,体现源语言句子到目标语言句子的互翻译可能性;argmax是搜索最大概率e^I的算子,这个搜索过程在统计机器翻译中又称为解码过程。在信道模型中统计翻译的质量很大程度上取决于语言模型和翻译模型好坏。后来发展出对数线性模型[2],该模型表示为

$$\Pr(e^I|f^J)=\exp\Big(\sum_m^M\lambda_m h_m(e^I,f^J)\Big)Z(f^J) \tag{3}$$

其中$Z(f^J)$为一个标准常量,此时翻译过程又可以表示为:

$$\hat{e}^I=\arg\max_{e^I}\Big(\sum_m^M\lambda_m h_m(e^I,f^J)\Big) \tag{4}$$

$h_m(e^I,f^J)$为e^I和f^J之间的特征模型,λ_m为特征模型的权重因子,可以认为该方法没有语言模型只有翻译模型。基于短语的统计机器翻译的基础是信源信道的统计机器翻译方式。

1.1 翻译模型

翻译模型是一种基于双语短语的方法,其中包括双语短语(BP),它是由互为翻译的单语短语(MP)对组成。基于短语的翻译的基本思想是,把源语言的句子划分成几个短语的形式,然

基金项目:中国科学院知识创新工程重要方向项目(No.KGCX2-SW-511)。
作者简介:强静(1982-),女,硕士研究生,研究方向为统计自然语言处理与机器翻译;张建(1954-),男,副研究员,研究方向为人工智能及其应用。
收稿日期:2007-08-21 **修回日期**:2007-11-15

后将这些源语言短语翻译成目标语言的形式。短语翻译模型的训练以双语对齐的语料库为输入,训练出短语翻译表。短语翻译模型主要有以下 4 个模块:

(1)词语对齐

Och 和 Ney[3]提出了一种精炼从 GIZA++中得到的对齐的启发式方法。该方法首先利用 GIZA++进行源到目标和目标到源语言的双向词语对齐。首先从双向词语对齐的交集开始,采用启发式的方法向双向词语对齐的并集进行扩充。Och 和 Ney 对他们的启发式方法中对于什么样的对齐点能够进行扩展描述得比较模糊,Koehn[4]对他们的启发式方法进行了详细的实现,在 Koehn 的实现中,首先从双向对齐的结果的交集开始,在双向对齐的并集中增加对齐点,这些对齐点包括交集对齐点的邻居点和它们之外遗失的对齐点。

(2)词语评分

根据词语对齐的结果,计算出两个词语之间翻译的最大概率。这里采用的是最大似然法。

(3)短语抽取

抽取方法就是提取对齐矩阵中的所有以对齐点为顶点的矩形,条件是与矩形所在行范围内的源词对齐的目标词也都在这个矩形的列范围内,反之亦然。

(4)短语评分

计算抽取出的短语对的翻译概率,包括五个部分:源和目标双向的短语翻译概率和短语词典概率和短语惩罚概率。

短语翻译模型训练的流程如图 1 所示。

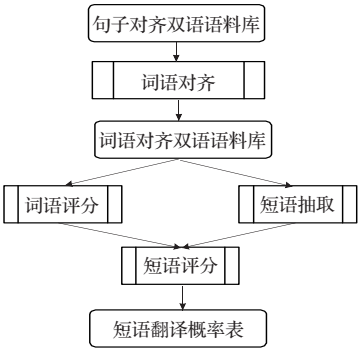

图 1 短语翻译表训练流程图

1.2 语言模型

语言模型采用 N-gram 的统计语言模型,由目标语言语料库训练而来;本文采用统计机器翻译领域公认的成熟开源语言模型训练工具 SRILM 进行 N-gram 语言模型的训练[5]。

2 Och 短语抽取算法

短语抽取就是根据对齐队列,从源语言和目标语言中抽取互译的短语对。

2.1 算法相关定义

在 Och,Koen[6]的方法中,在双向词语对齐提炼的基础上进行短语的抽取。短语抽取从双向对齐矩阵的顶点开始,遍历同时满足双语词语对齐连续的所有短语对,将满足条件的短语对加入到双语短语集合 BP 中。双语短语 BP 的集合可以定义为:

$$BP(s^I,t^I,A)=\{(s_j^{j+m},t_i^{i+n}):\forall(i',j')\in \boldsymbol{A}:j\leqslant j+m\leftrightarrow i\leqslant i'\leqslant i+n\} \quad (5)$$

式中 s^I 为源句子,t^I 为目标句子,m 为源句子长度,n 为目标句子长度,$\boldsymbol{A}$ 为对齐矩阵。给定所有的短语对,可以使用最大似然估计计算短语翻译概率分布:

$$\phi(\bar{s}|\bar{t})=\frac{count(\bar{s},\bar{t})}{\sum_f count(\bar{s},\bar{t})} \quad (6)$$

2.2 算法步骤

根据 Och 关于短语对的定义(公式(5))可知,短语对为连续对齐的双语词语集,其源语言部分唯一的对齐到其目标语言部分,目标短语部分也唯一的对齐到源语言部分。由此可以得到获取双语短语的如下限制条件:

(1)短语内的单词在原来句子中的位置必须连续;

(2)双语短语必须与对齐矩阵相容,即根据源语言句子和目标语言句子的对齐矩阵,源语言短语中的词语 $f_{j'}(j\leqslant j'\leqslant j+m)$或者对齐到 NULL,或者对应的目标语言词语 $e_{i'}$在短语 e_i^{i+n} 中,且只能在 e_i^{i+n} 中,反之亦然。

短语抽取的基本思想:穷举一种语言句子中的所有可能短语,然后根据对齐矩阵找到对应的另一种语言句子中的短语,并检查这两个短语是否符合以上 2 条限制。设源语言句子为 s^I,目标语言句子为 t^I,$\boldsymbol{A}$ 为 s^I 和 t^I 的对齐矩阵,以它们为输入进行短语对的抽取算法[7,8]法描述如下:

(1)Input:$s^I,t^I,\boldsymbol{A}$

(2)for $i1$ from 1 to I\{

(3) for $i2$ from 1 to I\{

(4) $TP=\{j|\exists i:i1\leqslant i\leqslant i2\wedge A(i,j)\}$;

(5) if (quasi-consecutive(TP)) TP 是连续的 {

$j1=\min(TP)$

$j2=\max(TP)$

$SP:=\{i|\exists j:j1\leqslant j\leqslant j2\wedge \boldsymbol{A}(i,j)\}$;

if ($SP\subseteq\{i1,i1+1,\cdots,i2\}$){

$BP=BP\cup\{(si2i1,tj2j1)\}$;

while ($j1>0$ and $\forall i:A(i,j1)=0$) {

$j''=j2$;

while ($j''\leqslant J$ and $\forall i:A(i,j'')=0$) {

$BP:=BP\cup\{(si2i1,t\ j''j1)\}$;

$j''=j''+1$;

}

$j1=j1-1$;

}}}}}

(6)Output:BP

该算法否定了单行(列)中局部连续的短语对齐信息。对于汉民小语料库可能会丢失重要的短语对齐对,影响翻译的效果。所以在 Och,Koen 的短语抽取算法的基础上,本文针对单行(列)中局部连续的情况进行改进。

3 短语抽取算法的改进

3.1 短语抽取长度的设置

在实验中选择了三个短语长度的设置,第一次抽取短语的长度是 3 个单词,然后增加到 5 个单词,和 7 个单词的长度。由算法可知 5 个单词的长度可以包含所有 3 个单词的长度,同样 7 个单词的长度也可以包括所有 5 个单词的长度。对于短语 7

个单词一般会覆盖短语以外的词，这样会造成短语空间浪费，也会造成解码搜索空间的浪费。根据短语的一搬的长度,最终选择以 5 个单词为标准的短语长度。

例如:以汉蒙语料库为例,下面是抽取的以 5 个单词为短语基准。

最 近 怎么样 ？ ||| BAYIDAL ORCIM YAMAR BAYIN_A ? ||| 1-0 0-1 1-1 2-2 2-3 3-4

3.2　改进的算法步骤

在上面算法的基础上增加了下面第 3 个判断条件,在词语对齐矩阵中,如果单行(列)的词语对存在不连续的对齐,但是不连续中又包含连续的对齐(称之为局部连续)也作为要抽取的范围。如图 2 所示(图中框中部分为局部连续):

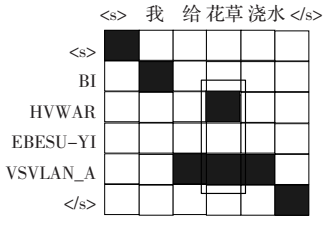

图 2　汉语到拉丁蒙文的对齐结果

相应的改进算法描述如下:

(1)Input:s^I,t^I,$\boldsymbol{A}$

(2)for $i1$ from 1 to I{

(3)　for $i2$ from 1 to I{

(4)　　TP={j|∃i:$i1$≤i≤$i2$∧$\boldsymbol{A}(i,j)$};

(5)　　if (quasi-consecutive(TP)) TP 是连续的 {

　　$j1$=min(TP)

　　$j2$=max(TP)

　　SP:={i|∃j:$j1$≤j≤$j2$∧$A(i,j)$};

　　if (SP⊆{$i1$,$i1$+1,…,$i2$}){

　　BP=BP∪{($si2i1$,$tj2j1$)};

　　while ($j1$>0 and ∀i:$\boldsymbol{A}(i,j1)$=0) {

　　j''=$j2$;

　　while (j''≤J and ∀i:$\boldsymbol{A}(i,j'')$=0) {

　　BP:=BP∪{($si2i1$,$tj''j1$)};

　　j''=j''+1;

　　}

　　$j1$=$j1$-1;

　　}}}}}

(6)　　else{　// if (! quasi-consecutive(TP))

　　TP 不连续的情况

　　if($i1$==$i2$){//单行或单列的情况

　　$j1$=min(TP);

　　for(TP'={ $j1$,$j2$ };TP'⊆TP;$j2$=$j1$++){

　　If(quasi-consecutive(TP'))

　　BP=BP∪{($i12i1$,$j12j2$)};

　　Else BP=BP∪{($i12i1$,$j12j1$)};

　　}}

(7)Output:BP

4　实验结果

实验选用不同规模的汉蒙双语语料库,得到改进前后的短语对和抽取所用时间的对比。结果为表 1。

表 1　实验结果表

ttable-limit		199	1 999	5 999	11 999
抽取的短语对	Och 抽取算法	3 530	357 73	109 954	189 956
	改进算法	3 716	377 76	115 749	201 145
抽取所用时间	Och 抽取算法	5 125	443 75	138 454	278 641
	改进算法	4 797	441 41	114 594	277 921

通过表 1 可知，改进后可以抽取更多的有用的短语对,这对于处理小语料库有很大的作用，由于汉民语料库的限制,能够找到很大的汉民语料库往往很困难,所以在小规模的语料库的基础上抽取出更多的对齐短语对对提高翻译准确率有着重要的作用。

5　结束语

短语抽取是基于短语的统计机器翻译的基础,短语抽取结果中短语的对齐正确率和对齐数目决定着翻译的精度。本文改进了 Och 的短语抽取算法,并采用不同规模的汉蒙双语语料库进行实验,结果表明采用改进的短语抽取算法抽取的短语对数可在原有 Och 算法的基础上提高 5%以上,从而可在一定程度上提高翻译模型的质量和翻译的准确率。

参考文献:

[1] Brown,Cocke,Pietra D,et al.A statistical approach to machine translation[J].Computational Linguistics,1990,16(2):79-85.

[2] Och,Ney.Discriminative training and maximum entropy models for statistical machine translation[C]//Proc of the 40th Annual Meeting of the Association for Computational Linguistics(ACL),2002.

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

[4] Tillmann,Ney H.Word reordering and a dynamic programming beam search algorithm for statistical machine translation[J].Computational Linguistics,2003,29(1):97-133.

[5] Stolke A.Srilm-an extensible language modeling toolkit[C]//Proceedings of the International Conference on Spoken Language Processing.

[6] Och F J.Statistical Machine Translation:From Single-Word Models to Alignment Templates[D].Computer Science Department,RWTH Aachen,Germany,2002-10.

[7] Och F J.Statistical machine translation:from single-word models to alignment templates[D].Computer Science Department,RWTH Aachen,Germany,2002-10.

[8] Cenugopal A.Vogel S,Vaibel A.Effective phrase translation extraction from alignment models [C]//Proceedings of the 1st Annual Meeting of the Association of Computational Linguistics(ACL),2003.