句法调序的统计机器翻译方法研究

孙广范,宋金平,肖 健,袁 琦
SUN Guang-fan,SONG Jin-ping,XIAO Jian,YUAN Qi

中国电子信息产业发展研究院,北京 100044
China Center for Information Industry Development,Beijing 100044,China
E-mail:morgan2001_sun@sohu.com

SUN Guang-fan,SONG Jin-ping,XIAO Jian,et al.Research on approaches of syntactic reordering for statistical machine translation.Computer Engineering and Applications,2009,45(36):142-144.

Abstract: To resolve the problem of the lack of capacity for the reordering in Phrase-Based Statistical Machine Translation(PB-SMT),this paper attempts to parse the input Chinese sentences to PBSMT,conduct the reordering of the sentences by a converter,pre-translate some types of phrases,and then translate the output from the above operation by a decoder.This paper focuses on the effects of the reordering for complex attributes linked by "de",and the effects of the reordering and pre-translation of prepositional phrases,specific collocation phrases and phrases of noun of locality.Experimental results show that the reordering and pre-translation can significantly improve the BLEU values of the English translation by PBSMT.

Key words: statistical machine translation;syntactic reordering;pre-translation

摘 要:为解决基于短语统计机器翻译存在的调序能力不足的问题,尝试利用句法分析器对基于短语统计机器翻译的输入汉语句子进行句法分析,然后利用转换器进行调序操作,并对部分类型短语进行预先翻译,然后再利用基于短语统计机器翻译的解码器进行翻译。重点测试了汉语中"的"字引导的复杂定语调序、介词短语、特定搭配短语、方位词短语的调序及预翻译产生的效果。实验结果表明,这些调序及预翻译操作可以显著地提高基于短语的统计机器翻译的英文译文结果的 BLEU 值。

关键词:统计机器翻译;句法调序;预先翻译
DOI:10.3778/j.issn.1002-8331.2009.36.042 **文章编号:**1002-8331(2009)36-0142-03 **文献标识码:**A **中图分类号:**TP391

1 引言

基于短语的统计机器翻译(PBSMT)的优点是可以实现局部调序,考虑一定的上下文信息,缺点是很难做到长距离调序,并且源语言和目标语言端的短语必须连续,这大大限制了基于短语的统计机器翻译方法的作用范围。针对 PBSMT 的调序能力较差的问题,一些研究者提出了引入句法信息来帮助解决调序问题的方法。吴德凯提出了 ITG 模型,将翻译看作是一个用同步语法同时分析源语言和目标语言的过程;Yamada 和 Knight 将翻译看作从源语言串导出目标语言句法树的过程;Chiang 将 PBSMT 和同步语法结合,提出了结构化的短语模型;刘群等提出了一种基于源语言句法分析的树到串对齐模板,并在此基础上建立统计机器翻译模型;Chao Wang 等提出了面向统计机器翻译的中英翻译句法调序规则集。Chao Wang 等提出的面向统计机器翻译的中文句法调序方法,在参考文献[1]中分析几类可能的调序规则,主要包括与汉语"的"字有关的调序、介词短语的调序、时间短语调序等类型。该文实验与 Chao Wang 等、刘群等的实验有些相似,都是对于 PBSMT 的输入进行预先调序,但是存在如下区别:他们是对于训练时和测试时的输入文本均进行调序处理,该文只对测试时的输入句子进行调序处理,并对于一些短语进行预翻译,关注的重点问题及处理方法不同,实验结果也与他们的实验结果存在差异,具体情况及其分析在下文中详细讨论。

2 中文句法调序情况分析

基于短语的统计机器翻译方法对于实际语言中的短语表达描述能力较强,而对于语言意义影响很大的语序调整描述能力较差。因此人们自然想到能否借用汉语句法分析器及转换器来帮助 PBSMT 完成一些调序工作。一个比较容易想到的想法是可否在基于转换的汉英机器翻译系统完成了汉语分析,树转换后直接将调序后的树对应的中文串作为 PBSMT 的输入,然后利用 PBSMT 进行翻译以得到翻译结果。按照这种想法进行了实验, 发现这样做的翻译结果的 BLEU 值比单纯的 PBSMT 翻译结果的 BLEU 值下降了。这说明靠基于转换的规则系统来全部完成调序工作,然后再调用 PBSMT 的做法不可取。既然调序工作全部由基于转换的规则系统来完成与 PBSMT 不协调,那么部分调序工作由基于转换的规则系统来做应该具有可行

基金项目:国家自然科学基金(the National Natural Science Foundation of China under Grant No.60872118,No.60572132)。
作者简介:孙广范(1966-),男,研究员,主要研究方向为机器翻译、自然语言处理;宋金平(1972-),男,高工,主要研究方向为机器翻译、自然语言处理;肖健(1972-),男,高工,主要研究方向为机器翻译应用;袁琦(1939-),男,研究员,主要研究方向为机器翻译应用。
收稿日期:2008-07-11 **修回日期:**2008-11-17

性;既然 PBSMT 不善于对较长距离的调序工作,那么基于转换的规则系统应该在这方面提供帮助。对于汉语和英语的句法比较而言,主谓宾顺序基本一致,“的”字引导的定语从句在翻译成英文时需要将定语从句后置,介词短语做定语时在翻译时一般应该后置,介词短语修饰动词做状语时多数情况应该后置或放到句尾。

实验系统的处理流程如图 1:

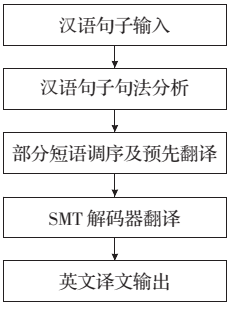

图 1　实验系统的处理流程图

3　评价实验

文中使用的开源程序生成的 PBSMT(下称 Baseline-SMT)包括用 GIZA++等工具对第三届统计机器翻译研讨会评测提供的约 865 758 句对的汉英平行语料库进行训练,使用中科院自动化所开发的短语抽取工具“胡杨”进行短语抽取,得到了约 365 万条的短语翻译概率表;解码器使用哈尔滨工业大学开发的“绿洲”解码器,语言模型使用 Pharaoh 的英语语言模型。实验数据采用中科院计算所提供的开发集,开发集包括 492 个汉语句子及对应的多种英文参考译文。

笔者利用自己研制的汉语句法分析器和汉英转换器对基于短语的统计机器翻译系统的输入汉语句子进行树到串的转换及调序操作(有些情况下还对一些短语进行预翻译),然后用 Baseline-SMT 进行翻译,得到了下列实验结果。实验结果比较的参照点是使用自己开发的分词工具对测试文本进行分词后运行 Baseline-SMT,并从输出结果中去掉未登录词,得到开发集的一个翻译结果。图 2~图 5 是该文涉及的句法调序类型的样例。表 1 中列出了对 Baseline-SMT 的输入句子进行不同预处理后再使用 Baseline-SMT 进行翻译得到的结果的 BLEU 值的提高幅度的情况。

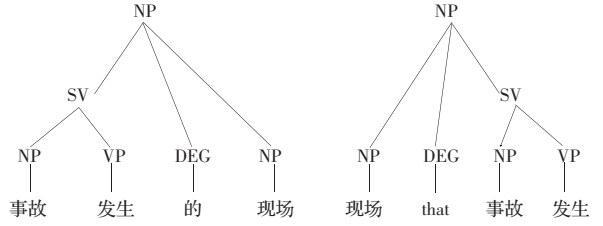

图 2　“的”字连接的定语从句的调序

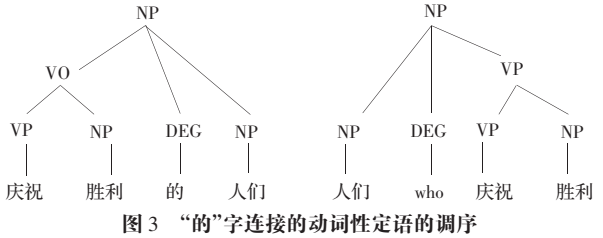

图 3　“的”字连接的动词性定语的调序

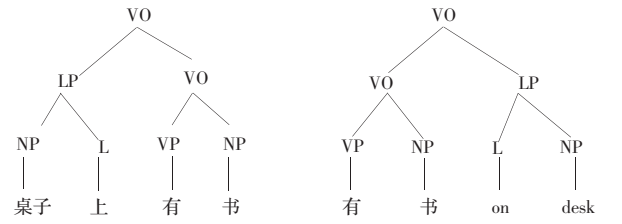

图 4　方位词短语(LP)的调序

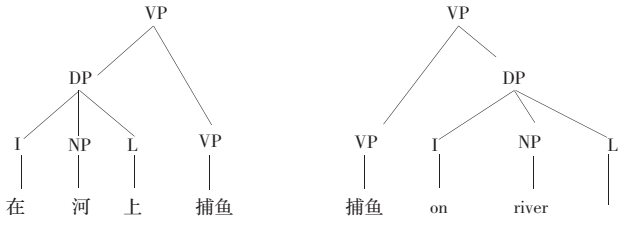

图 5　特定搭配短语(DP)的调序

表 1　几种对输入句子进行预处理方法对翻译结果 BLEU 值提高的影响

	BLEU 值提高的百分数/(%)
预处理 1:对输入进行时间词、数词、中国人名、地名和机构名的翻译 然后调用 Baseline-SMT 进行翻译,对翻译结果中未翻译的短语进行短语替换	1.97
预处理 2:对输入进行预处理 1 后,对包含“的”字连接定语的名词短语(定语部分是从句或动词短语)进行调序处理 然后调用 Baseline-SMT 进行翻译,对翻译结果中未翻译的短语进行短语替换	2.36
预处理 3:对输入进行预处理 2 后,对介词短语(IP)、特定搭配短语(DP,如“在……上”)、方位词短语(LP,如“……后”)进行调序处理 然后调用 Baseline-SMT 进行翻译,对翻译结果中未翻译的短语进行短语替换	2.22
预处理 4:对输入进行预处理 2 后,对介词短语(IP)、特定搭配短语(DP,如“在……上”)、方位词短语(LP,如“……后”)进行翻译处理 然后调用 Baseline-SMT 进行翻译,对翻译结果中未翻译的短语进行短语替换	2.46
预处理 5:对输入进行预处理 2 后,对介词短语(IP)、特定搭配短语(DP,如“在……上”)、方位词短语(LP,如“……后”)进行调序及翻译处理 然后调用 Baseline-SMT 进行翻译,对翻译结果中未翻译的短语进行短语替换	2.51

4　实验结果分析

表 1 中的预处理 1 中对输入进行了时间词、数词、中国人名、地名和机构名的翻译,这发挥了规则处理这几类词或短语的优势,对于最终翻译结果 BLEU 值的提高有帮助,BLEU 值提高了 1.97%。

表 1 中的预处理 2 对输入进行预处理 1 后的包含“的”字连接定语的名词短语(定语部分是从句或动词短语)进行调序处理,比未做预处理的结果的 BLEU 值提高了 2.36%。此结果比预处理 1 得到的翻译结果 BLEU 值提高了 0.39%,这说明对于包含“的”字连接定语的名词短语(定语部分是从句或动词短语)进行调序处理对于译文质量提高确有好处。

表 1 中的预处理 3 对输入进行预处理 2 后,对介词短语(IP)、特定搭配短语(DP,如“在……上”)、方位词短语(LP,如

“……后”)进行调序处理,比未做预处理的结果的 BLEU 值提高了 2.22%,但是比进行预处理 2 的翻译结果的 BLEU 值降低了 0.14%。这说明预处理 3 对译文的 BLEU 值产生了负作用。

表 1 中的预处理 4 对输入进行预处理 2 后,对介词短语(IP)、特定搭配短语(DP,如“在……上”)、方位词短语(LP,如“……后”)进行翻译处理,比未做预处理的结果的 BLEU 值提高了 2.46%,比进行预处理 2 的翻译结果的 BLEU 值提高了 0.1%。这说明预处理 4 对译文的 BLEU 值提高产生了积极作用。

表 1 中的预处理 5 对输入进行预处理 2 后,对介词短语(IP)、特定搭配短语(DP,如“在……上”)、方位词短语(LP,如“……后”)进行调序及翻译处理,比未做预处理的结果的 BLEU 值提高了 2.46%,比进行预处理 2 的翻译结果的 BLEU 值提高了 0.15%。这说明预处理 5 对译文的 BLEU 值提高产生了积极作用。

为什么预处理 2 能对 BLEU 值提高有帮助而预处理 3 则产生负面作用呢?产生这种情况的原因应该归结为基于规则的调序与 PBSMT 的调序能力的协调问题。预处理 2 中对包含“的”字连接定语的名词短语(定语部分是从句或动词短语)进行调序处理时,定语部分由于是从句或动词短语,一般比较复杂,定语部分一般词数较多并且进行整体调序,这与 PBSMT 常进行短距离调序的冲突的可能性较小,因而导致了实验中预处理 2 后提高译文 BLEU 值的结果。而对于预处理 3 中对介词短语(IP)、特定搭配短语(DP,如“在……上”)、方位词短语(LP,如“……后”)进行调序处理,由于 IP、DP、LP 的长度一般较短,与 PBSMT 常进行短距离调序的冲突的可能性较大,从而导致了预处理 3 后的译文 BLEU 值比预处理 2 后的译文 BLEU 值降低的情况。对于上文中提到的靠基于转换的规则系统来全部完成调序工作,然后再调用 PBSMT 的做法导致 BLEU 值下降的情况,其原因也可以归结为基于规则调序与 PBSMT 进行短距离调序的冲突造成的问题。对于预处理 4 中对输入进行预处理 2 后,对介词短语(IP)、特定搭配短语(DP,如“在……上”)、方位词短语(LP,如“……后”)进行翻译处理,然后调用 Baseline-SMT 进行翻译,这种做法减少了与 PBSMT 进行短距离调序的冲突造成的问题,从而提高了翻译结果的 BLEU 值。对于预处理 5 中对输入进行预处理 2 后,对介词短语(IP)、特定搭配短语(DP,如“在……上”)、方位词短语(LP,如“……后”)进行调序及翻译处理,然后调用 Baseline-SMT 进行翻译,这种做法也是因为减少了与 PBSMT 进行短距离调序的冲突造成的问题,从而提高了翻译结果的 BLEU 值。

5 下一步工作

在以后的工作中,将进一步研究将句法分析及转换操作与基于短语的统计机器翻译相结合的途径,考察不同的结合方法产生效果的差别,寻找较好的结合方式,以利于基于句法的统计机器翻译译文质量的提高。

参考文献:

[1] Wang C,Collins M,Koehn P.Chinese syntactic reordering for statistical machine translation[C]//EMNLP 2007.

[2] 刘群,熊德意,刘洋.基于句法的统计机器翻译研究[C]//中文信息处理前沿进展—中国中文信息学会二十五周年学术会议论文集,北京,2006:416-423.

[3] Liu Yang,Liu Qun,Lin Shou-xun.Tree-to-String alignment template for statistical machine translation[C]//COLING/ACL 2006,Sydney.

[4] Yamada K,Knight K.A syntax-based statistical translation model[C]//Proceedings of the 39th Annual Meeting of the Association on Computational Linguistics,ACL 39.

[5] Utiyama M.A survey of statistical machine translation[C]//Lecture Slides,Kyoto University,2006.

[6] Chiang D.A hierarchical phrase-based model for statistical machine translation[C]//Proceedings of ACL 2005:263-270.

[7] Koehn P.Pharaoh:A beam search decoder for phrase-based statistical machine translation models[C]//Proceedings of the Sixth Conference of the Association for Machine Translation in the Americas,2004:115-124.

[8] Zens R,Och F J,Ney H.Phrase-based statistical machine translation[C]//Jarke M,Koehler J,Lakemeyer G.LNAI 2479:Advances in Artificial Intelligence 25 Annual German Conference on AI,KI 2002.[S.l.]:Springer Verlag,2002:18-32.

[9] Wu De-kai.Stochastic inversion transduction grammars and bilingual parsing of parallel corpora [J].Computational Linguistics,1997,23(3):377-403.

(上接 128 页)

参考文献:

[1] Ghias A,Logan J,Chamberlin D,et al.Query by humming:Musical information retrieval in an audio database[C]//Proc ACM Int Conf Multimedia,San Francisco,CA,1995:231-236.

[2] McNab R J,Smith L A,Witten I H,et al.Towards the digital music library:Tune retrieval from acoustic input[C]//Proc ACM Int Conf Digital Libraries,Bethesda,MD,1996:11-18.

[3] McNab R J,Smith L A,Witten I H,et al.Tune retrieval in multimedia library[J].Multimedia Tools Applicat,2000,10:113-132.

[4] Rolland P Y,Raskinis G,Ganascia J G.Music content-based retrieval:An overview of Melodiscov approach and systems[C]//Proc ACM Int Conf Multimedia,Orlando,FL,1999:81-84.

[5] Blackburn S,De Roure D.A tool for content based navigation of music[C]//Proc ACM Int Conf Multimedia,Bristol,UK,1998:361-368.

[6] Chen B,Jang J S R.Query by singing[C]//Proc IPPR Conf Comput Vision,Graphics,Image Process,Taiwan,R O C,1998:529-536.

[7] Pardo B,Shifrin J,Birmingham W.Name that tune:A pilot study in finding a melody from a sung query[J].J Amer Soc Inf Sci Technol,2004,55:283-300.

[8] 钱博,李燕萍,唐振民,等.基于频域能量分析的自适应元音帧提取算法[J].电子学报,2007,35(2):279-282.

[9] 钱博,李燕萍,唐振民,等.一种基于线性预测残差倒谱的基音检测算法[J].计算机工程与应用,2007,43(32):210-213.