一类基于平行语料统计的汉法机译解决方案

刘粤钳[1],姚红玉[2]
(1.中国传媒大学 应用语言学系,北京 100024;
2.安徽师范大学 教育科学学院,安徽 芜湖 241000)

摘 要:针对目前国内外汉法机器翻译系统较少,且研究的汉语语例基本为简单短句的情况,利用《人民日报》中、法文网络版的部分文章建立了一个小型的汉法平行语料库,并基于此,利用改进的Yamada算法构建了一个汉法机器翻译系统。系统通过对汉法平行语料的统计结果,把汉语句型大致归为单谓和多谓两大类,并提炼出4096个汉法对齐基本句型,将之应用于汉法机译中;并首次提出了三词序列出现概率的概念,用于解决词语搭配的问题。试验表明系统在处理多谓语的汉语长句上有明显的优势。

关键词:三词序列出现概率;汉法对齐基本句型;多谓句;语料库
中图分类号:TP391.2 **文献标识码**:A **文章编号**:1673-629X(2008)04-0114-04

A Novel Solution to Chinese-French Machine Translation Based on Aligned Corpus

LIU Yue-qian[1], YAO Hong-yu[2]
(1. Department of Applied Linguistics, Communication University of China, Beijing 100024, China;
2. School of Educational Science, Anhui Normal University, Wuhu 241000, China)

Abstract: The study puts forward a corpus-based statistical solution to the rare Chinese-French machine translation system, with which by now can only deal short sentence. A Chinese-French machine translation system, then, is established by applying 4096 aligned Chinese-French basic sentence types obtained within the aligned Chinese-French corpus. For the first time, 3-word-sequence appearing probability, a new concept by which the difficulty of collocation may be untied, is introduced also. Finally, a test provides evidence of the conclusion that the system given in this article does have advantage in translating Chinese multi-predicate, and long customarily, sentence into French.

Key words: 3-word-sequence appearing probability; aligned Chinese-French basic sentence type; multi-predicate sentence; corpus

0 引 言

祁依虹于2002年针对汉法机译在《计算机工程》、《计算机工程与应用》上发表了两篇论文[1,2],提及在开发汉法机器翻译系统过程中遇到的难点。其中她尚未解决的突出的问题有:

第一,能够处理的法语动词的语式、时态过于单一,基本为直陈式现在时。

第二,长句处理不好,这主要有两个原因:①由于上述第一个未解决问题,导致不可能在分析汉语复句中各谓语间关系的基础上合理选择法语对应的各谓语动词的语式、时态和变位形式;②法语词序以及词间搭配处理不好,当句中词数增多时,这一问题愈发突出。

收稿日期:2007-07-06
作者简介:刘粤钳(1974-),男,广东人,博士研究生,讲师,研究方向为信号处理、自然语言理解。

文中将主要针对这两点提出解决方案,对祁依虹解决的较好的难点基本不予讨论。

归根结底,这两个问题的出现是基于规则的机器翻译系统不可避免的。为此,笔者选取2004年《人民日报》中、法文网络版的部分文章建立了一个小型的汉法双语平行语料库,在此基础上建立了汉法机器翻译系统,实验该系统较好地解决了:长句中各谓语动词在汉法语中语序的差异;法语动词的语式、时态和变位;汉语语体在法语中的保持。

1 几点说明

1.1 单谓句与多谓句

鉴于国内外学者对现代汉语单复句的存在与否[3]与划分标准意见不一,同时考虑到句法分析的方便起见,文中屏弃单复句的称谓,结合汉法平行语料统计的

结果,把汉语句型大致归为单谓和多谓两大类共4096个汉法基本对准句型(区分了笔、口语体)。在建立汉法平行语料库的同时,每一汉语基本句型按照统计的结果提供了出现概率最大的已经确定了各谓语的语式时态的对准的法语基本句型。这样细分主要基于二点考虑:①法语动词最为活跃,基本句型必须以它们(尤其是常用动词)为中心构造;②法语中语式时态变化繁复且其变化会直接影响到句子在语义乃至语用上的差别。

1.2　汉语多谓句的切分

为了能以较快的速度处理汉语多谓句,本研究的思想是"化整为零",以语料统计得到的汉法对准基本句型中法语端的基本形式切分汉语源句。由于法汉语的固有差异,这时会遇到法语动词与汉语动词不能一一对应的情况,本研究采用的原则是"从简不从繁",尽量减少目标语法语的动词,使得句中法语动词数不大于汉语动词数,以便于汉语多谓句的切割。

1.3　汉法电子词典的建立

在进行汉法平行语料统计时,首先要解决的问题是汉法电子词典的建立,本研究采用的方法如图1所示,对于一个汉语词,统计法语语料中各法语对应词的出现次数,最后进行归一化处理,除去出现概率小于1%的词后,余下的法语对应词存储在以该汉语词为表头的单链表中。

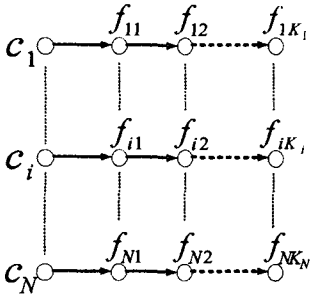

图1　汉法电子词典的存储结构

1.4　三词序列的出现概率

Yamada等人[4]提出了基于句法结构的统计翻译模型。该算法中词语的搭配及翻译等通过对翻译模型的训练获得,并限定输入信道的英语句长不超过10个词。大规模的试验表明该算法优于IBM算法,且可适用于英语到多种语言的翻译。

依据Yamada本人给出的算法复杂度可知,句中词数过多时,时间开销惊人。于是,试引入一个新概念——三词序列的出现概率,即统计汉语及法语语料中任意一个三词序列在一句话中出现的概率,这样,如果已知两个汉语或法语单词去搜索第三个单词的被选集合,就可以确定第三个单词及其出现的相对位置。反复利用它,便可使法语句法树的所有终端节点依次生长出叶子,最后为名词添加指示限定词(通常为冠词)等便可生成一棵完整的法语句法树并输出结果。

具体说来,令 w_x, w_y, w_z 是汉语或法语语料库中任意出现的三个词,w_1, w_2, w_3 是 w_x, w_y, w_z 的一组取值,则三词序列的出现概率为 $P(w_1, w_2, w_3) = \frac{N(''\cdots w_1\cdots w_2\cdots w_3\cdots'')}{M}$,其中 $''\cdots w_1\cdots w_2\cdots w_3\cdots'' \in S$,$S$ 为汉语或法语语料库中所有句子的集合;$N(''\cdots w_1\cdots w_2\cdots w_3\cdots'')$ 则表示汉语或法语语料库中任意一句话中包含字符串 $''\cdots w_1\cdots w_2\cdots w_3\cdots''$ 的所有句子的个数;M 表示汉语或法语语料库中所有句子的总数。可将任意三词序列的出现概率的存储空间(以三维数组存放)视为拓扑形式为一立方体的空间(每一垂直坐标轴的剖面对应一个双词序列在语料中出现的概率表)。如图2所示,当给定序列 $< w_x, w_y, w_z >$ 中的任意两个词,即 (w_x, w_y) 或 (w_x, w_z) 或 (w_y, w_z) 时,可得到与之搭配概率最大的第三个词以及相对于前两者的位置信息。

需特别指出,法语的三词序列出现概率立方体的坐标轴上标注的法语动词不是其原形,这样在查汉法电子词典得到汉语某词汇或词组对应的若干种法语动词原形后,要依据基本句型的要求把这几种法语动词原形转化为相应语式时态后,再进入概率立方体中比较与其他已经确定词汇的搭配可能性,最后得到应选用的法语译文。对于法语中其他有性、数变化的词类处理相同。

三词序列出现概率的引入,实际上是把Yamada算法中的后期计算过程提前至建库操作中,并将词的个数限制为3,这在机译频繁时,避免了重复计算,大大降低了系统的反应时间。

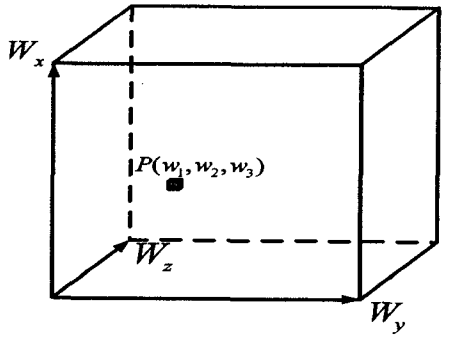

图2　三词序列出现概率的存储示意图

2　算法流程

算法流程如图3所示。

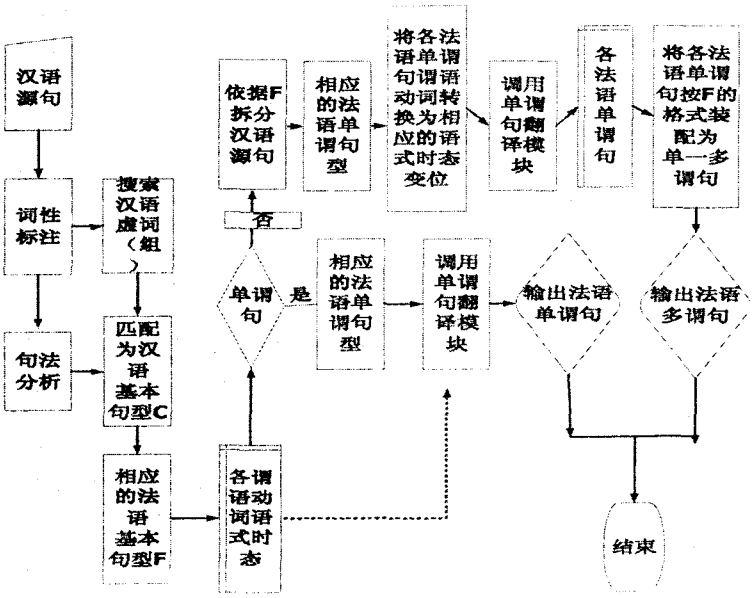

图 3 算法流程图

3 实例分析

如图 3 所示,向汉法翻译系统输入汉语源句"如果天儿好,又没课的话,我们就去海边散步。"对该句同时进行词性标注和句法分析,根据搜索到的句中的表意虚词"如果"、"又"、"的话"、"就",可判断该句为三谓语的"假设条件 1 + 递进条件 2 + 将来很可能发生的行为"口语体句型,其对应的法语基本句型 F 为"Si + 直陈式现在时 + et que + 虚拟式现在时 + 主句主语 + 直陈式简单将来时",并将结果存储起来。

由于汉语源句非单谓句,依据法语基本句型 F 的格式将该汉语源句切分为三个短句:[1]天儿好;[2]没课;[3]我们去海边散步。[2]缺主语需补充"天"或"我们"中的一个,由于 P("我们","没","课") > P("天","没","课"),故将[2]补充为"我们没课"。

句[1]的处理过程为:"天儿好"属汉语基本单谓句型中的"'天'(气) + 形"的口语体,找到其对准的法语基本句型为"Il faire + adj."再将 faire 变位为 F 规定的直陈式现在时第三人称单数形式,即{SC{NP{Il} FV{fait}} AP{好}}。下面需确定 $P(w_1, w_2, w_3)$ = P("Il","fait",w_3) 中的 w_3:"好";查电子词典 $w_3 \in$ {"bon","bien","parfait","d'accord","beau","amical","agreable"},经比较,P("Il","fait","beau") 最大,故 w_3 = beau。输出单谓句 Il fait beau 并存储。

句[2]的处理过程为:"我们没课"属汉语基本单谓句型中的"主语 + '没(有)' + 名词"的口语体,找到其对应法语基本句型为"主语 + avoir + pas + de + n.(pl.)"。将主语 Nous 和 F 规定的 avoir 的虚拟式现在时形式 ayons 代入,即{SC{NP{Nous} FV{ ayons }} pas NP{de 课(复数)}},问题转化为只需确定 $P(w_1, w_2, w_3)$ = P("pas","de",w_3)——此处将 w_1 取为 pas,w_2 取为 de 是因为算法中要求已知的两个叶子节点必须为在句法树上距离待求未知叶子节点最近的两个中的 w_3:"课"(复数);词典中 $w_3 \in$ {"lecons","cours","textes","classes"},经比较,P("pas","de","cours") 最大,故 w_3 = cours,输出单谓句 Nous ayons pas de cours 并存储。

句[3]其实是一连动句,但考虑到目标语——法语习惯的表达方式,把它归入汉语基本单谓句型中的"主语 + '去' + 地点 + 延续动词"的口语体,找到其对应法语基本句型为"主语 + faire + n + à + n"。将主语 Nous 和 F 规定的 faire 的直陈式简单将来时形式 ferons 代入,即{SC{NP{Nous} FV{ferons}} NP{散步} PP{à NP{海边}}}。以下的翻译过程如图 4 所示。最后输出单谓句 Nous ferons un tour au bord de la mer 并存储。

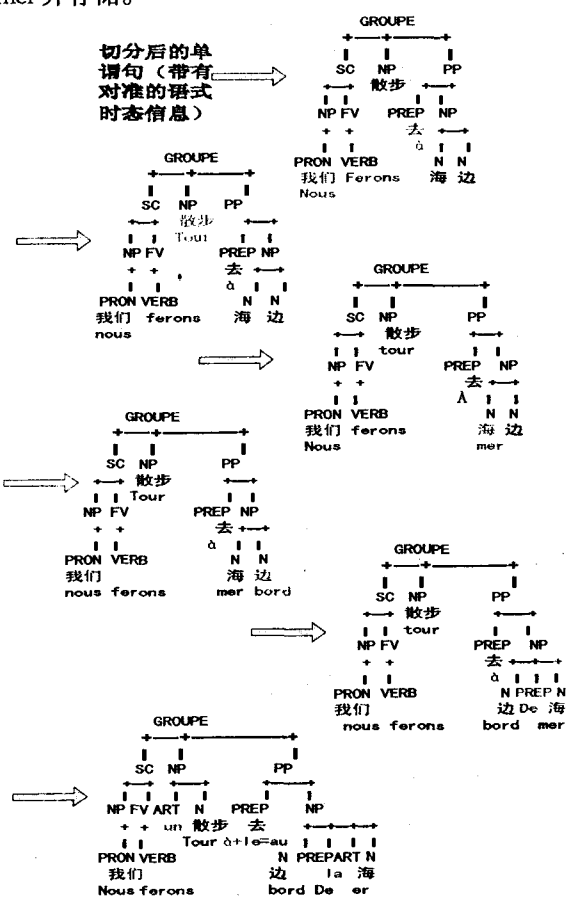

图 4 句[3]的翻译过程

至此,三个单谓句均已确定,最后依据法语基本句型 F,将三个句子装配为 S'il fait beau et que nous ayons

pas de cours, nous ferons un tour au bord de la mer. 输出。

4 系统评价

4.1 试 验

选取2004年《人民日报》中、法文网络版的部分文章作为语料库。用ICTCLAS中文词法分析器对中文语料进行词性标注;用概率句法分析器ICTPROP对中文进行句法分析;用Xerox Incremental Parser 8.02对法语进行词性标注及句法分析。得到一个约20万词的汉法平行语料库和汉法对齐句法树库。参考法国计算语言学家Maurice Gross与Boons, Guillet, Leclère等人制作的81个矩阵图(20个固定搭配表,18个句子为补语的简单动词表,43个带名词性补语的简单动词表)[5,6]对句法树库作进一步处理,从中提炼出4096个汉法对准基本句型库。同时计算出汉语及法语的三词序列出现概率并以三维数组存储如图2所示。

在所建汉法电子词典允许的汉语词汇范围内,向系统输入单谓句、双谓句和三谓句各50个,系统经翻译输出150个法语句子。

4.2 评 价

采用人工评测对两个系统进行评价。将5.1中两个系统输出的共150个法语句子分别给母语为法语的三个人阅读,依据好、中、差("好"指句子符合法语习惯,无任何语法、句法词语搭配等方面的错误,很地道;"中"指句子存在少许错误,但仍易读懂;"差"指错误严重,难以读懂)进行评价;最后记录下三个人评定的各等级句子个数的均值。其测评结果见表1～表3。不难看出,本系统在处理汉语多谓长句时具备明显的优势。

表1 单谓句测评结果

好	中	差
34	16	0

表2 双谓句测评结果

好	中	差
32	17	1

表3 三谓句测评结果

好	中	差
29	19	2

参考文献:

[1] 祁依虹,茅于杭.汉法机器翻译的难点分析[J].计算机工程,2002(9):235－237.

[2] 祁依虹;董清富,茅于杭.汉法机器翻译系统初探[J].计算机工程与应用,2002(18):114－116.

[3] 孙良明.汉语单复句划分标准评析[J].山东师范大学学报:社会科学版,2000(1):88－92.

[4] Yamada K, Knight K. A syntax－based statistical translation model[C]//In: Proceedings of the 39th Annual Meeting of the ACL. [s.l.]:[s.n.],2001:523－530.

[5] Boons J P, Guillet A, Leclère C. La Structure des Phrases Simples en Français[M]// Constructions Intransitives. Droz, Genève:[s.n.],1976.

[6] Guillet A, Leclère C. La structure des phrases simples en français[M]//Constructions transitives locatives. Droz, Genève:[s.n.],1992.

(上接第113页)

像效果较好。Roberts算子容易丢失一部分边缘,也不具备抑制噪声能力;Prewitt算子先对图像做加权平滑处理,虽具有一定的抑制噪声能力,但不能排除出现虚假边缘;Canny算子的边缘定位能力优于本算子,但Canny算子会检测出虚假边缘;LOG算子在抑制噪声的同时会丢失一些尖锐边缘[2,4,5]。另外,需要提到的是,在推导两个矩阵算子的过程中,将两矩阵中的余弦项进行四舍五入处理,仿真结果表明,这样的处理,使矩阵运算得到了一定程度的简化,在一定程度上节省了运算时间,同时丢失的边缘点极少,比较适宜采用。

4 结 语

与传统的边缘检测方法不同,文中将边缘检测的基本思路用在DCT域,通过直接处理DCT系数,得到图像的边缘信息,使用该方法处理压缩格式图像,比用传统的像素域方法大大节省了处理时间,提高了系统的性能和响应速度;同时,还推导出了通用的在DCT域进行边缘检测的算子,可以利用该算子对反量化后的每个8×8块的系数矩阵(含64个DCT系数)直接进行处理,从而方便快捷地求出每个块中的图像边缘,对一幅图像中所有的块都施以如上的运算,就可以检测出整幅图像的边缘信息。

参考文献:

[1] 沈兰荪,魏　海,黄祥林.压缩域图像处理技术研究[J].北京工业大学学报,2000,26(3):24－25.

[2] 姚　敏.数字图像处理[M].北京:机械工业出版社,2006:225－232.

[3] 王桂华,张问银,唐建国.DCT域图像边缘的快速提取[J].计算机应用,2005,25(1):100－102.

[4] 麦特尔.现代数字图像处理[M].孙　洪,译.北京:电子工业出版社,2006:190－199.

[5] Canny J. A computational approach to edge detection[J]. IEEE Trans. Pattern Analysis and Machine Intelligence, 1986,8:679－698.