维普资讯 http://www.cqvip.com

基于实例的汉英机器翻译系统研究与实现

王长胜 刘 群
(中国科技大学研究生院,北京 100080)
E-mail:wcs@software.ict.ac.cn

摘 要 文章实现了一个基于实例的汉英机器翻译系统(Example-Based Chinese-English Machine Translation,简称 EBCEMT)。实验结果表明:以纯粹的中英对照的句子对的例句库为基础,且不进行深层次的语法分析,在给定的翻译模式下,该翻译系统在效率,翻译结果的正确率等方面取得了令人满意的结果

关键词 实例翻译 翻译系统 双语语料库

文章编号 1002-8331-(2002)08-0126-02 **文献标识码** A **中图分类号** TP391.2

The Research and Implementation of an Example-based Chinese-English Machine Translation System

Wang Changsheng Liu Qun
(China Science and Technology University, Beijing 100080)

Abstract: *In this paper, a new approach of implementing an Example-based Chinese-English Machine Translation System (EBCEMT) is introduced. In the given translation models, the EBCEMT system presents satisfying results in experiments on the system's efficiency and correctness while using only raw bilingual corpora.*

Keywords: Example-based Machine Translation(EBMT), Translation System, bilingual corpora

1 引言

随着计算机的计算速度、存储容量的提高和大量的电子双语语料的出现,在机器译领域产生了基于实例的机器翻译方法,它克服了传统的基于规则的机器翻译方法(Rule-Based Machine Translation,简称 RBMT)知识获取的困难,但其技术远没有 RBMT 那样成熟,主要存在的问题有:

(1)实例库的表示,基本上有两种形式即:对实例直接存储,对实例库进行加工后存储,如加工成依存树[1]、格框架[2]等树状结构及泛化成模板等形式,而且实例库本身就有不同级别对齐的问题,如有单词级对齐[3],短语级对齐[4],句子级对齐的实例库。不同的实例库表示直接关联系统的后序计算。

(2)相似度计算 为确定实例库中哪些例句或短语片段可用来翻译输入的源语,系统必须建立一套相似度准则以确定两个句子是否相似。多数系统采用基于单词的相似度。其它的有采用句法层面的相似度,它涉及到较复杂的句法分析。而它本身又存在"粒度(grain)"问题,即匹配长度与精度矛盾问题

(3)相似例句或短语选择。此问题的产生源于实例库中存在同一源语言句子或短语的多个不同目标语言的句子或短语,以及如何在一个相似度相同的例句集合中选择一个或部分例句或短语等去翻译输入的源语。对此不同的 EBMT 系统采用的策略乃至具体方法都不尽相同。

(4)对齐问题 在实例是句子级对齐时,要翻译输入的源语就必须对相似对照例句进行对齐计算,显然此时的对齐工作集中在词汇或短语或子句级上,而这一问题目前也还不成熟。

(5)译文生成,它涉及到选词和边界问题,即要解决未匹配词问题,从多个候选词或短语片段中选择最恰当的以及组合译文时出现边界问题

(6)效率问题,针对上面提到的问题,该文所搭建的 EBMT 系统避开对自然语言的深层次理解和分析,从最简单明了的中英对照的句子级别的实例库出发,采用双向的不同的基于单词相似度方法和一套评分机制方法定量刻画最相似例句,然后通过一体化的对齐,译文生成算法进行翻译 它简化了对齐计算,解决了译文生成的选词问题和部分边界问题。

2 系统设计与实现

2.1 基本研究思路

系统的具体目标为:搭建一个实例翻译系统,其各模块间相互独立(可用其它功能相同的模块替换),且易维护,扩展,改进,高效,译文质量高,另外可作为多引擎翻译系统的一个部件。系统基本思路为:快速检索出大于一定阈值(基于词个数相似,的相似度)的初始实例集合,然后从这个集合中选出最好的一个例句作为翻译的参考,接着通过对齐计算及下一步计算得到的转换表达式和子块库进行译文生成。另外还可在训练时加入一个反馈学习模块对生成的译文进行训练,得到更高质量的译文。参见图 1 系统结构图。

子块库定义:初始检索实例集合中的对翻译有参考价值的任何句子、子句、短语、片段、单词(它们都是经过相似度计算,得到的输入与例句所共同有的部分)的集合。

2.2 主要研究内容

2.2.1 实例库表示和检索

项目基金:国家 973 重点基础研究项目(课题号:G1998030507-4)

作者简介:王长胜,男,研究生,研究方向:自然语言处理,数据挖掘。刘群,男,博士,中国科学院计算技术研究所副研究员。

维普资讯 http://www.cqvip.com

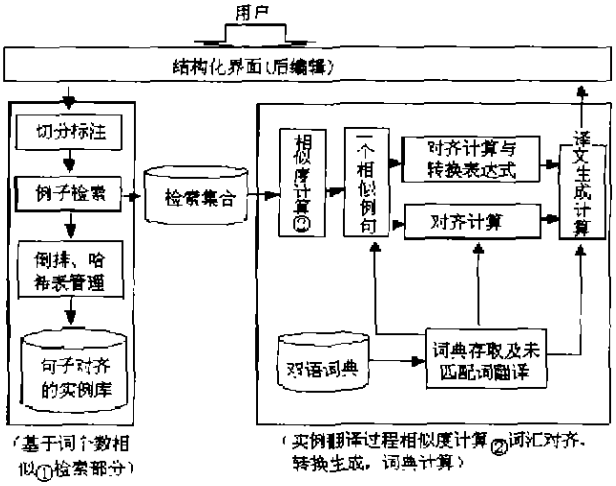

图1

该系统的实例库为中英对照的句子对,没有对它进行任何的加工,存储为文本文件,格式形如"绿色是美丽的颜色。\|* Green is a beautiful color.",来源于互联网上收集的各种中英对照的语料(经过了加工)。采用这种直接存储的方式,系统管理、扩充方便,包括为了对它检索而建立的索引文件(倒排文件和哈希文件),详见文章《实例库的检索和管理》。

2.2.2 相似例句选择计算

该系统采用双向的基于词相似的方法完成各自的任务:相似度计算①用于对经过检索出来的相关原始实例库进行共同单词个数计算(正向扫描),然后计算相似百分比并与给定的阈值比较筛选得到初始检索实例库。而相似度计算②(反向扫描)用于完成两件事情:

对输入的中文句子和初始检索实例集合中的每一个中文例句进行基于语句表层的词汇相同计算相似度[6],得到共同单词个数和一个位置结构(相同单词在输入句子中位置和例句中位置)数组。

对位置结构数组分析得到转换表达式、子块库(对齐和译文生成时还可能需要动态建立之并检索)和句子的评分。

转换表达式定义:输入句子单词流顺序(包括相同片段和不相同单词)+各自对应的操作(a添加,d删除,r替换,c拷贝)+输入中位置+例句中位置。

句子评分标准:包括相同片段评分和结构评分。它是相似例句选择的最关键因素。但因为对各个句子评分时出现了不统一的情况,所以用系统定义的翻译模式(有一些限制条件,目前给出:词性,单词个数,同义词三个条件,但不一定都同时使用)取代,同时能够保证所选择的例句是一个好的翻译参考对象。系统预定义的翻译模式目前给出了11种(可扩展):cr,crc,cac,cdc,crcr,rc,rcrc,rcac,dc,dcr,dcrc (c=copy,r=replace,a=add,d=delete)。

2.2.3 对齐计算

选出了最相似的一个中文例句后,要在这个最相似的中文例句对应的英文例句中找出其中不相同的中文部分对应的英文部分。这里不对整个对照例句找出各自单词一一对应的部分,而仅仅是做其中的一部分工作。当然还可能需要动态建立子块库以尽最大努力作好对齐工作,最后得到一个对齐表达式、

对齐表达式定义:英文例句单词流顺序+转换表达式指针及指向例句中对应的中文单词的位置。

2.2.4 译文生成

通过上述得到的转换表达式和对齐表达式(包括必要的词性等)就可以进行译文的生成工作了、对于其中的未匹配词的翻译采用了查找词典和子块库相结合的办法。

2.2.5 反馈学习

主要是对生成的译文进行一些辅助性修饰、修改等工作。

2.2.6 系统翻译流程

以下给出一个详细的例子说明以上的工作:

输入:大自然/之/壮观/非/笔墨/所能/形容/。/ 65%

(1)初始检索集合:3个(其中"/"是我加入的,表示分词的结果)按比率大小排序给出。

景色/之/美/非/笔墨/所能/形容/。/ Word cannot describe the beauty of the scene.0.750000;

风景/之/美/非/笔墨/所能/形容/。/ The beauty of the scenery beggars description.0.750000;

这/风景/之/美丽/非/我/笔墨/所能/形容/。/The scene be so beautiful that it transcend my power of description. 0.666667。

这一步采用从前向后计算两句中相同单词个数的办法计算相似度(相似度①)。

(2)选择一个最相似例句:结果为上述例子1。

目的:从中文角度找到一个可以以之去类比翻译的例句。

算法过程:依次对初始检索集合中每一个中文例句跟输入句子计算相似度(相似度②),得到各相同单词的一个位置结构信息(记录它们分别在例句和输入中的位置),然后分析此结构得到一个转换表达式和子块库,最后通过系统预定义的翻译模式和限制条件判断这一句是否能够作为一个翻译参照例子去翻译,如果可以则接下去做对齐和译文生成工作,否则继续扫描其他的句子,

对例子1:

位置结构信息为一个数组:(1,1)(3,3)(4,4)(5,5)(6,6)(7,7)。

转换表达式为一结构数组:(a,r,0,0,0,0)(b,c,1,1,1,1)(c,r,2,2,2,2)(d,c,3,7,3,7),

子块库为一结构数组:(1,1,1)(1,3,7)。

从上述转换表达式提取的翻译模式为:rcrc,符合系统给定的11个模式之一,且各r对应的部分符合系统给定的限制条件(目前给出:词性,单词个数,同义词三个条件,但不一定都使用)。

(3)块对齐。

目的是找到翻译模式中各字母在英文例句中的对应位置。

采用的策略主要有三个:每种策略中用到不同的技术细节,如用语义、位置、相似度等,尽最大努力保证对齐工作的成功。

先在中英词典中查找,再在英中词典中查找,然后位置分析,最后子块库查找。按照此顺序依次查找,分析结果,如果找到对应的位置则结束,接下去进行译文生成工作,如果找不到则系统无法翻译。

对例子1:

首先在中英词典中查所有整个r对应的部分:词典格式为单词 词性 所有义项(空格隔开)。

r1->景色 n landscape outlook prospect scene scenery sight view,依次用各义项在英文例句中匹配:最后发现与其中

(下转135页)

维普资讯 http://www.cqvip.com

the 20th National Information Systems Security Conference, Baltimore, Maryland, USA: National Institute of Standards and Technology/National Computer Security Center, 1997. 353-365

10. Robert Flood, Ewart Carson. Dealing with complexity—An introduction to the theory and application of systems science[M]. second elition, New York: Plenum Press, 1993: 151-158

11. Herve Debar, Marc Dacier, Andreas Wespi. Towards a taxonomy of intrusion-detection systems[J]. Computer Networks, 1999; 31(8): 805-822

12. Wenke Lee. A data mining framework for building intrusion detection Models[C]. In: IEEE Symposium on Security and Privacy. Berkeley, California, 1999.5: 120-132

(上接 127 页)

scene 匹配，记录结果。

r2->美 n beauty，依次用各义项在英文例句中匹配：最后发现与其中 beauty 匹配。记录结果，

分析所有记录结果：全部对齐，即找到各 r 对应的位置，将结果保存在对齐表达式中，此例为：(0,a,0)(2,c,1)。

(4)译文生成。

目的是找到输入句子中不同部分的译文，最后通过对齐的位置信息进行合成，得到译文

采用的策略有两个：每种策略中也用到不同的技术细节，如词性一致、相似度、子块库等，尽最大努力保证查找到包含所需要的单词的译文并正确选择合适的单词。

先翻译模式中整个非 C(其实为 alr)块的查找选择，然后在中英词典中查找选词。算法核心是分析非 C 块结合词典查找和子块库的动态建立查找选词。此例为 r 块都是对应一个单词，查词典：大自然->n nature；壮观->n glory spectacle。对于"壮观"需要选词，通过在检索过程中保留的中间结果(包括初始检索集合)查找"壮观"的可能译文，统计分析出最好的译文。此例中查看检索前 n-n*b+1 个单词(n=5，h=0.65)的结果是：前 3 个单词为[笔墨 9][壮观 14][大自然 15]，另外两个为[形容 17][所能 76]，其中数字为此单词出现在所有句子中的个数，详见文献[5]。下面给出 14 个有"壮观"的例句的前五个：

从山顶看到的景色非常壮观。The view from the hill top be magnificent；

场面十分壮观。The set be terrific；

除夕我们观看了壮观的烟火表演。We have looked grand sight performance of firework at New Year's Eve；

她站在那儿，看着这壮观的景象。She stand there and survey the spectacle；

彗星上的冰块在太阳热的作用下迅速而壮观地升华成气体。The comet's ice, heated by the sun, rapidly and spectacular sublimate to gas。

对齐得到统计结果(spectacular 3)(magnificent 4)(grand 2)(spectacle 1)(terrific 1)(marvelous 1)；还有两个无法对齐，通过选词机制排除了 glory，而选用 spectacle。

最后合成为译文：Word cannot describe the spectacle of the nature。

反馈学习功能目前还没有考虑。

2.3 实验结果

在 CPU 为 P200，64M 的普通内存，非 SCSI 硬盘，平台为 windows 2000server，visual c++6.0 上实验结果如下：

这里仅测试了两个指标：块对齐和翻译结果的正确率(前提是给出的测试用例要符合系统给定的翻译模式，否则这两步工作都无法进行)、给出了 50 个符合条件的测试用例，对于块对齐，以一句中是否所有块都对齐为标准，统计有 41 个对齐了，正确率为 82%；而翻译结果的正确率，以人工评价译文为标准，统计有 35 个翻译正确，正确率为 70%。

2.3.1 结果分析

对以上结果分析发现：

(1)没有对齐的部分原因是英汉双向词典中都查不到且在子块库中也查不到(在系统要求的时间响应范围内)。目前该词典情况是：从一部有 12189 个单词的英汉词典中生成的另一部有 31502 个单词的汉英词典。正确率较高是因为充分利用了检索的例句集合，而不仅仅利用单一的一个对照例句和两部词典，

(2)对齐以后有 3 个句子不能翻译，原因是块对齐的部分是通过出位置信息分析出来的，而两部词典中都没有相应的单词，而恰好在译文生成时子块库中也无法得到所需的译文，另外 3 个句子是选词出错导致最终译文出错(但译文整体结构还是正确的)。系统在对齐正确的前提下翻译正确率很高 85%(35/41)，这说明其选词能力较高。

3 结束语

实验表明：

(1)尽管该系统还有待改进，但在保证系统效率的前提下充分利用检索到的例句集合进行的对齐和译文生成都获得了较高质量。

(2)在不对原始中英对照实例库(句子级别)加工的前提下，利用现有的资源和技术进行实例翻译，效果还是可以接受的，这样的系统也完全可以作为多引擎翻译系统的一个部件。

需要指出的一点是，这里给出的测试用例受限于翻译模式。要使翻译面更广，可以扩充翻译模式。此外，扩大例句库和词典的规模可以使翻译效果更好。(收稿日期：2002 年 2 月)

参考文献

1. S Sato, M Nagao. Towards memory-based translation[C]. In Proceedings of COLING90

2. D Jones. Analogical Natural Language Processing[M]. UCL press, 1996

3. 陈利人，陈群秀. 基于实例的日汉机器翻译部件的研究和实现[C]. 《ICCC96》文集，1996

4. 周莉娜. 面向基于实例汉英机器翻译的知识获取及实现[D]. 博士论文. 北京大学，1997

5. 常宝宝. 汉英机器翻译中的基于实例的转换引擎研究[D]. 博士论文. 北京大学，1999