汉英平行语料库双语语义对应空位研究

范 云,黄 萍,黄俊红
(重庆大学 外国语学院,重庆 400030)

摘要:双语词语语义对应空位现象是英汉互译的常见障碍。在建立英汉平行语料库的过程中,这也是无法回避的问题。本文拟总结目前已建成的一些语料库处理这一问题的习惯方式,并在此基础上借鉴机器翻译研究领域的理论成果,试图做出更加细致的分析并提出更加可行的解决方案。

关键词:汉英平行语料库;语义对应空位;机器翻译
中图分类号:H030 **文献标识码**:A **文章编号**:1008-5832(2005)02-0084-04

A Study for Zero and Partial Equivalence in C&E Parallel Corpus

FAN Yun, HUANG Ping, HUANG Jun-hong
(*College of Foreign Languages, Chongqing University, Chongqing 400030, China*)

Abstract: As a common obstacle for C-E & E-C translation, the phenomenon of zero and partial equivalence of words and phrases between these two languages also constitutes an inescapable hurdle in the operation of C & E bilingual parallel corpus. On the basis of summarizing briefly some effective countermeasures on this issue, this paper, with the aid of certain theoretical contributions in the realm of MT, is trying to go further as to put forth some more detailed and workable schemes.

Key words: C&E parallel corpus; zero and partial equivalence; machine translation

一、引言

关于英汉双语的语义对应空位现象,国内的学者有过许多透彻的研究分析,总结起来,主要原因是在于不同的文化背景和思维方式(王佐良)。由此产生了词语语义的完全不对应和部分不对应:完全不对应表现为此种语言的词语在彼种语言中完全没有与之对应的表达法,或者所指的相同实体(entity)并不表达相同或类似的文化意义;部分不对应主要表现为两种语言的某些词语间有一定的意义共核(core),但并不能完全彼此互换。上述这些空位现象经常出现在翻译实践中,译者们也找到了比较合理的方法来处理,比如说音译(叩头:koutou)、省译、变通等。那么,在建立英汉双语平行语料库的过程中应如何处理这样的空位现象呢?本文认为,既然要考察的是在英汉平行语料库中处理双语语义空位的方法,视角和研究途径应和一般的翻译策略研究不尽相同,比较突出的是要建立一套基于计算机语言的机制来使处理办法程序化、简单化。

二、常见的处理方案

(一)双语平行语料库及其构建机制

建立语料库是语言学者们对自然语言进行系统研究的有效途径,其重要意义在于语料库本身使研究者们大量占有语言原始材料成为可能,目前牛津大学的当代英语语言语料库就是大规模语料库的典范;国内也建成了类似的汉语语料库,取得了良好效果(中国外语教育研究中心主页 http://www.sinotefl.com/new/center/center_c.asp)。双语平行语料库跟这种一语语料库有着一些不尽相同的地方。首先,它所收集的语言材料涉及两种不同语言,而且这两种语言材料存在着对应关系,通常是某种文本的双语译本;其次,双语平行语料库大多与机器翻译有联系,很多时候要向使用者提供词语和文本的翻译服务。以 BABEL 双语平行语料库为例,在其主页上提供了词语翻译的引擎;《中国日报》主页 www.chinadaily.com.cn 上也有热点词汇翻译的功能。

双语平行语料库的构建机制并不十分复杂,从图

收稿日期:2005-01-20
基金项目:重庆大学语言认知及信息处理研究所2004专项研究基金资助项目
作者简介:范云(1963-),女,重庆万州人,重庆大学外国语学院副教授,主要从事应用语言学、语言与文化和翻译研究。

1(http://icl.pku.edu.cn/icl_groups/parallel/workspace.htm)可以看出,整个流程包括语料的采集、语料的整理加工、语料的组织和检索工具的开发等几个主要步骤,这些主要的步骤往下细分,还有除噪、标记、对齐、校对等环节,以保证整个语料库正常运作。目前很多双语语料库的研究集中于标记方式、对齐方式等,而且已经形成了较成熟的切分字段的思路,这无疑为进一步研究铺平了道路(BABEL 双语平行语料库主页 http://icl.pku.edu.cn/icl_groups/parallel/default.htm)。

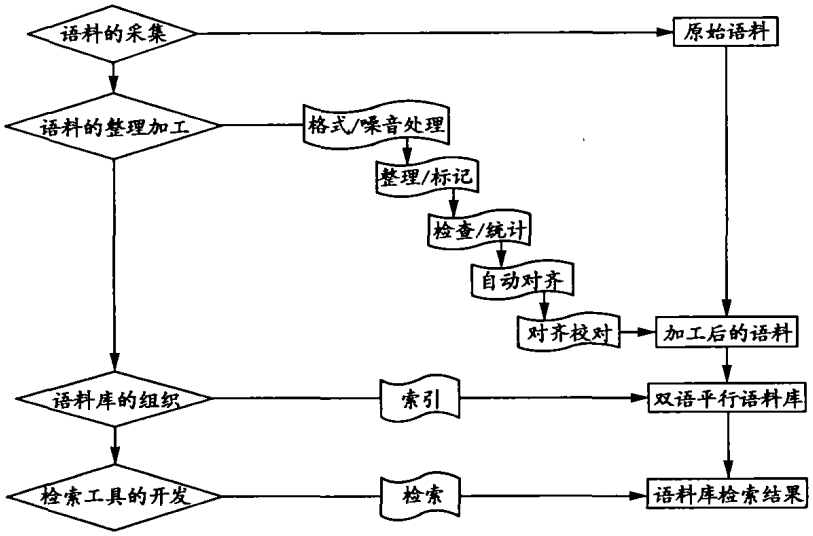

图1　大规模双语平行语料库的工作流程示意图

(二)已建成双语平行语料库对空位现象的处理

建库初期的收集原始语料阶段,研究者们总是在选取某种文本的双语材料,即原始材料在很多情况下已经有了翻译关系,这样一来,双语语义空位现象已经被文本的译者处理了,因此,与建库相关的处理空位的任务在于如何做好标记和对齐字段。就标记和对齐问题,目前已经有了很成熟的处理办法,北京大学计算机语言研究所的《构建大规模的汉英平行语料库》文章提供的方式很有操作性:"我们严格定义了与双语平行语料库建设相关的术语:原始语料、双语语料库、篇章级对齐单位、原文文件、译文文件、段落级对齐单位、句子级对齐单位、源语言。其中:篇章级对齐单位(记作 AT):一个篇章级对齐单位由若干段落级对齐单位构成,可以表示为:AT = AP1,AP2... APn。其中,AP1 = (PS1,PT1),AP2 = (PS2,PT2) ... APn = (PSn,PTn);PS1,PS2... PSn 构成一篇完整的原文文本。(Ts),PT1,PT2... PTn 构成原文文本对应的完整的译文文本(Tt),即 Ts 与 Tt 之间具有"翻译关系"。原文文本和译文文本分别存放在两个文件中,这两个文件的文件名相同,但后缀名不同。段落级对齐单位(记作 AP):一个段落级对齐单位由若干句子级对齐单位构成,可以表示为:AP = AS1,AS2... ASn,其中,AS1 = (S1,T1),AS2 = (S2,T2) ... ASn = (Sn,Tn),S1,S2... Sn 构成原文文本中一个或多个完整的段落(整体记作 Ps),T1,T2... Tn 构成译文文本中一个或多个完整的段落(整体记作 Pt)。Ps 和 Pt 之间具有"翻译关系"。句子级对齐单位(记作 AS):一个句子级对齐单位是一个二元组,记作 AS = <Si,Ti>,其中 Si 由一个或多个自然的句子组成;Ti 由一个或多个自然的句子组成。Si 与 Ti 之间具有"翻译关系"(http://icl.pku.edu.cn/project/parallel/reference/construct_large_CE_parallel_corpus.pdf)。

这样的篇章-段落-句子层级对齐方式的优越性在于能够最大可能地使标记更为清楚准确,从而为用户的检索提供方便。关于切分和标记实例,参见 BABEL 双语平行语料库对江泽民同志在中国共产党第十六次全国代表大会上的报告的处理个案(http://icl.pku.edu.cn/icl_groups/parallel/download.htm)。

三、问题的深化

在英汉平行语料库的建设过程中,每个双语文本的处理加工都会按照上述或类似于上述的方式进行,可以预见的是(以汉英翻译为例),某个汉语词的出现频率有可能较高,而不同的语料来源对这个词的翻译可能不尽相同,那么这个词就有了不止一个的英文对应词或表达法,检索者就要面临一个取舍问题,要检索者自己来取舍选择,这倒不是什么太难的事情,如果检索者直接要求机器翻译,问题就会变得比较麻烦,这就是一个常见的空位问题;同样,如果所录入的材料比较单一,即使检索者所输入的检索字段的意义在语料库中有所涉及,仍然有可能满足不了其检索要

求,这也是不能忽视的空位现象。那么,怎样来解决这些问题呢?

(一)由一词多译产生的空位问题

对于一词多译的问题,当前惯常采取的方式是通过词频统计排列出不同译出词的顺序,并在机器翻译过程中主要采用高频词优先或首字母排序的原则。以金山词霸2005的程序设计为例,当检索者输入汉语词"改革"后,程序将提供如下的英语译出词(按首字母排序):fashion,innovate,innovation,reform, reformation,regenerate,每个词都提供词义解释的链接,供检索者进一步了解和筛选(如词性,搭配等);再以BABEL双语平行语料库为例,其检索页面更加细致:除了一般的检索功能外,还提供了政治、经济、科技、生活等领域的分类检索,而且提供了检索词的词频数据。通过这样的方式,一词多译的问题能够得到部分解决,但目前的机器翻译程序在选词取词方面还不让人满意,实际操作已经证明高频词不一定就是适合所录入源语词的恰当译出词,而首字母排序的办法则更难以满足要求,因而在这方面还应该有更细致的探索。本文认为,解决的办法首先在于收集原始语料时做出更详细的语料范畴规划,并在加工语料的过程中对某一方面的语料做更深入的划分,以政治性语料为例,可以将"政治"这一范畴更加细化为"政党"、"政策法规"、"外交"、"选举"等小的范畴,然后将原来笼统划分在"政治"领域下的词语进行重新分配(这中间允许某些词的重复录入,如"改革"一词既可进入"政党"又可进入"政策法规"),语义空位出现的概率将缩小,检索的准确度将提高,机器翻译的效率也将提高。需要注意,预先的范畴划分是一个主观性较强的步骤,理论上不大可能穷尽所有的范畴,因此还需要其他措施补充。那么,后续的工作应该在语料的切分和标记过程中完成。为了减少语义对应空位,在切分时应更多考虑到某个词的搭配情况,考虑作为某一范畴语料时这个词的惯常搭配和用法。还是如"改革"一词,在作为政治性语料时(不论何种更小的范畴),它总是习惯搭配"开放"一词构成名词的平列结构,与"深化"这个动词构成一个动宾结构。这两种常见搭配的英译相当固定,分别是reform and opening to the outside world和deepen reform(《北京周报》)。按照这样的习惯结构进行标记和储存,并在检索页面提示检索者输入所需的语料类型,语料库就能为检索者提供较为准确的译出语,同时机器翻译时,程序所面临的选择项更少,翻译的准确度也能够有所提升。

(二)由语料不充分产生的空位问题

这类空位现象就要麻烦一些。首先,在建库实践中存在着"绝对"的由双语语料不充分而产生的空位现象,即检索者所需要的语料根本不在库存语料中。比如"三农问题"是一个近期的热点词汇,但一些语料库(如电子或在线双语词典)没有及时收录,以至于会闹出"three farmers question"这样的笑话。这样的问题只能通过语料库本身的不断丰富和更新来完成,现在已建成的不少语料库都提供了检索者反馈的页面,这是一个不错的解决办法;其次的空位现象更加隐蔽一些,那就是检索者输入了检索要求和机器翻译要求后语料库无法提供与检索者要求完全匹配的语料,这里面的主要原因是词汇同义或近义的问题,简单说来,如果检索者的输入内容是"设立强力监督机构",假设语料库里仅仅录入"建立有效监督机制"的双语对应材料,并对其做了明确的切分和标记,那么检索的结果可能是"检索失败"的提示或者是其他的蹩脚译出词。由于一种语言的词汇同义现象非常普遍,语料库程序难以做出准确的辨认和处理。就这一问题,常见的处理办法是模糊检索,就是语料库向检索者提供一系列符合检索要求某个或某些字段的检索结果,供检索者进一步筛选(这很类似于互联网上搜索引擎的工作原理),这在很大程度上降低了检索失败的概率,但检索者不得不又花精力去进行主观取舍(因为很多时候能匹配部分字段的检索结果非常多,有的还很牵强),因此只能是部分地解决了问题,要直接做机器翻译的难度就更大。要更进一步解决难题,则应在语料收集和加工上做文章。语义学告诉我们,很多同义或近义词固然很容易混淆,但大多有不同的感情色彩和不同的使用场合以及不同的搭配群组,在语料整理时,如果充分考虑这些问题,某些字段的参数将会更加丰富也会更加系统,这实际上回到了更加细分语料范畴的问题,在此基础上,检索页面可以对用户的检索要求提出更细致的范畴询问,检索和机器翻译的成功率就有了提高的可能。当然,要完成这么细致的检索任务需要一套更加成熟的算法支持,这还有待进一步研究试验。

(三)机器翻译研究领域的支持

双语平行语料库的建设与机器翻译研究有密切联系,本文所言及的空位现象也是机器翻译领域的重要课题,目前机器翻译的"消歧"研究就是对空位问题思考的延续。早期人们所使用的词义消歧知识一般是手工编制的规则。由于手工编写规则费时费力,存在严重的知识获取的"瓶颈"问题。20世纪80年代以后,语言学家提供的各类词典成为人们获取词义消歧知识的一个重要知识源。近年来,随着计算机存储容量和运算速度的飞速提高,通过使用各种机用资源和

大规模语料库,计算机能自动获得各种动态的搭配知识及其统计数据,这为消歧任务创造了良好条件。然而,要真正有效地提高词义消歧的水平,不仅需要获取词的释义和分类信息,而且更重要的是综合利用现有的语法、语义资源,在词类划分基础上,增加词义的语法功能分析和语义搭配描写,从多知识源中提取多义词的每个意义相互区别的分布特征。基于这样的认识,北京大学计算机语言研究所王惠(2002)提出了如下的消歧方式:

第一,利用词类标记进行词义消歧。

例如:"补贴"①贴补:~家用|~粮价②贴补的费用:福利~|副食~。"现代汉语语法信息词典"对7万汉语词语逐一进行了词性标注,而且现有的汉语词类自动标注的正确率也已经达到96%以上,因此,对于词类不同的意义,计算机可直接借助于语料中的词性标记进行判断。比如,遇到下面经过自动切词,词类标注过的文本:[1]这/r 将/d 由/p 国家/n 予以/v 补贴/v. [2]生活/n 补贴/n 很/d 快/a 发到/v 灾区/n 人民/n 手/n 里/f 了/u. 计算机可以很容易地根据词类标记判断出例[1]中的"补贴"是①义,例[2]中的"补贴"是②义,从而给出正确的语义标注或英语译文:[1] This will be subsidized by the state. [2]Living allowances were quickly handed out to the people in the stricken area.

王惠进一步认为,《现代汉语词典》的20 513个名词中共有多义词3 989个,其中像"补贴"这样包含不同词类的意义的名词有932,占多义名词的23.4%. 对200万字的《人民日报》语料(1998年1月)的统计结果与此相近,22 744个名词中共有多义词2 196个,其中意义词类不同的有592个,占27%. 这说明,仅仅利用词类标记就可以消除超过1/5的汉语歧义。

第二,词类相同,利用子类标记进行词义消歧。

第三,子类相同,则利用语法功能的差异进行词义消歧。

第四,语法功能相同,进一步利用语义搭配限制进行词义消歧。

第五,利用自由义和非自由义进行词义消歧(王惠《机器翻译中基于语法、语义知识库的汉语词义消歧策略》,)。这样的处理办法实际上很好地呼应了上文所谈到的空位问题处理方式,把语料库的语义对应空位延续到了机器翻译的消歧课题,应该是对空位现象比较满意的处理办法。

四、结语

本文扼要论述了汉英双语平行语料库语义对应空位现象的由来和较为可行的处理办法。实际上,这类问题相当复杂,远还不到满意解决的程度;同样,机器翻译研究领域的进展也并非轻而易举,本文的讨论仅仅是涉及到了部分问题。随着研究的进一步深化,双语平行语料库的建立和维护应该向人工智能的方向发展得更快,这样空位问题以及其他诸多问题都能得到更有效的解决。

参考文献:

[1]IDE, NANCY, JEAN VéRONIS. Introduction to the Special Issue on Word Sense Disambiguation: The State of the Art [J]. Computational Linguistics, 1998, 24 (1):23 - 24.

[2]LUK ALPHA K. Statistical Sense Disambiguation with Relatively Small Corpora Using Dictionary Definitions [R]. Cambridge, Massachusetts: The 33rd Annual Meeting of ACL, 1995.

[3] GALE WILLIAM A, KENNETH W CHURCH, DAVID YAROWSKY. A Method for Disambiguation Word Senses in a Large Corpus [J]. Computer and the Humanities, 1993, (26): 38 - 39.

[4] GALE WILLIAM A, KENNETH W CHURCH, DAVID YAROWSKY. Using bilingual materials to develop word sense disambiguation methods [R]. The International Conference on Theoretical and Methodological Issues in Machine Translation, 1992.

[5]柏晓静,等. 构建大规模的汉英双语平行语料库[EB/OL]. http://icl.pku.edu.cn/project/parallel/reference/construct_large_CE_parallel_corpus.pdf, 2001 - 10 - 10.

[6]黄昌宁,李涓子. 语料库语言学[M]. 北京:商务印书馆, 2002.

[7]黄子桓. 从中英平行语料库自动撷取双语词组知识[EB/OL]. http://www.csie.ntu.edu.tw/~ b90093/joggy/appli.pdf, 2002 - 06 - 10.

[8]蒋冰清,英汉文化空缺词汇现象及翻译策略[EB/OL]. http://219.148.135.229: 85/~ kjqk/ldszxb/ldsz2003/0301pdf/030133.pdf, 2002 - 06 - 10.

[9]史晓东. 英汉机器翻译:现状和未来[A]. 中国中文信息学会二十周年学术会议论文集[C]. 北京:清华大学出版社,2001.

[10]王惠. 机器翻译中基于语法,语义知识库的汉语词义消歧策略[EB/OL]. http://www.huayuqiao.org/articles/wanghui/wanghui08.doc, 2002 - 08 - 10.

[11]王佐良. 翻译,思考与试笔[M]. 北京:外语教学与研究出版社,1998.

[12]俞士汶,朱学锋,王惠,张芸芸. 现代汉语语法信息词典详解[M]. 北京:清华大学出版社, 1998.

[13]赵铁军,等. 机器翻译原理[M]. 哈尔滨:哈尔滨工业大学出版社, 2000.