优化地学词汇标注方案 奠定完善地质语料库基础*

张翼翼 董淑欣 杨会兰
(中国地质大学,北京 100083)

提 要: 在地学文献翻译实践过程中,笔者通过 Google 在线翻译提供的译文,结合地质专业词汇的特点,分析基于语料库的机器翻译系统存在的一些典型问题。同时,从优化词汇标注方案角度对语料处理提出建议,借此提升地学文献的机器翻译质量,为构建地学领域的专用型语料库奠定基础。

关键词: 语料库; 词汇; 标注
中图分类号: H314 **文献标识码:** A **文章编号:** 1000 - 0100(2013)04 - 0122 - 3

On Word-processing Based upon the Annotated Corpus

Zhang Yi-yi Dong Shu-xin Yang Hui-lan
(China University of Geosciences, Beijing 100083, China)

This study is done by the Work-shop of English for Geology, an academic group under the Department of Foreign Languages at China University of Geosciences (Beijing). According to a piece of Chinese episode translated into English by Google on the Internet, this paper focuses on how to make computer-aid-translation better in light of word-processing based upon the annotated corpus, by means of correcting the translations with problems and analyzing the features of writing in Geological field.

Key words: corpus; word; annotated

1 引言

关于语料库的定义,Atkins 和 Clear 认为,语料库是为专门目的、按照明确设计标准收集的文章集合(Granger 1998: 7)。该定义包含 3 个方面:(1)建构语料库具有专门的目的;(2)语料库具有明确的设计标准;(3)语料库是由文章组成的集合(王建新 2005: 16)。也就是说,语料库由自然出现的语言样本汇集而成,是为语言研究而收集并用电子形式保存的语言材料。

计算机技术迅速发展,使包含广泛自然语料的语料库得以建立。语料库不仅对词汇学、翻译、语言教学等研究有巨大促进作用,而且对机器翻译软件、信息提取软件、拼写检查软件的发展具有重大的推动作用,语料库方法也因此成为自然语言处理的重要方法(王建新 2005: 4)。

近年来,计算机语料库对自然语言处理的各个不同方面(如话语识别、人机对话、信息提取、网页分类、机器翻译、文字处理等)都显得极为重要,而且极具潜力,这已经得到国际计算语言学界的广泛认可(王建新 2005: 3)。但是,基于语料库的机器翻译的效果仍然不够理想,尤其是涉及到具有专业背景和行业特色的相关文献时,这种不理想体现得更加明显。

目前,地学领域的中英文语料库还未完全建立,作为专用型语料库,地质语料库是专门为地学领域的科研、教学、教材编写以及语言比较研究而收集的文章集合,其取样的文本应该力求代表地学环境中的英语语言及其变体。语料库中除了大量地学信息有助于提升机器翻译质量之外,相应语料处理尤其是词汇标注(附码)在很大程度上决定着翻译质量的高低。因此,本文以节选自 *Long-term persistence of oil from the Exxon Valdez spill in two-layer beaches*(*Nature Geoscience*)的片段为例,说明如何通过优化语料库词汇的标注方案,

* 本文系中央高校基本科研业务费专项资金资助项目"基于我国世界地质公园的中英文公示语研究双语平行对译语料库的构建"(2 - 9 - 2012 - 04)的阶段性成果。

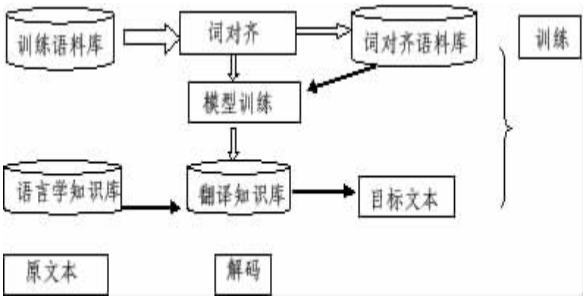
提升地学文献的机器翻译质量，为完善地学领域的专用型语料库奠定基础。

2　实证分析

原文: Oil spilled from the tanker Exxon Valdez in 1989 (refs 1, 2) persists in the subsurface of gravel beaches in Prince William Sound, Alaska. / The contamination includes considerable amounts of chemicals that are harmful to the local fauna 3. / However, remediation of the beaches was stopped in 1992, because it was assumed that the disappearance rate of oil was large enough to ensure a complete removal of oil within a few years. / Here we present field data and numerical simulations of a two-layered beach with a small freshwater recharge in the contaminated area, where a high-permeability upper layer is underlain by a low-permeability lower layer.

利用 Google 提供的在线翻译译文: 石油从油轮，埃克森公司在 1989 年瓦尔迪兹(文献 1，2) 泻坚持在阿拉斯加州威廉王子湾，砾石的海滩地下。/ 污染向当地动物都是有害的化学物质，包括相当数量。/ 然而，泳滩的整治是在 1992 年停止，因为它是假设石油的消失率足够大，以确保在几年之内彻底清除的石油。这里我们提出一个两层的海滩，在污染区，其中一个高渗透率的上层是由一个低渗透率较低层之下的小淡水补给领域的数据和数值模拟。

在讨论之前，首先看机器翻译的基本模式(巢文涵 2008: 9) :



从图中可以看出，处理语料库中的词汇在机器翻译中扮演着重要角色。Google 提供的在线翻译将 remediation，removal 分别译为“整治”、“清除”，这说明机器翻译系统针对某些词汇能根据整个语篇进行意义层面的对齐，然而对另外一些词汇的释义却不够理想。例如，将 persist 译为“坚持”，是由于受到后面介词 in 的影响。英文单词 persist 既有“坚持做某事”的释义，也有“持续/存留”的释义。Google 在线翻译使用的翻译系统对语料库中 persist 进行词类自动标注时，依据局部上下文线索(王建新 2005: 180) 区分 persist 的两种含义，致使 in 及其后面单词的词性成为区分两种不同意义的关键。其实，原文中的 in 是地点状语的一部分，与后面的名词关系密切，与前面的动词关系松散，并不代表 persist in doing sth 中的 in，因此 persist 应该翻译为“存留”而非“坚持”。有鉴于此，标注词汇时是将词组拆开还是另觅其他组合方式，有赖于句法规则和出现频率。

受到固定搭配影响的例子还包括将 assumed 错误地翻译为“假设”，而没有视其为常常出现在科技文章中的习惯性用法，正确地将 it is assumed that 翻译成“人们认为”。语料库中的词汇大部分是一个一个被标注的，而特定语言环境要求灵活地将几个单词标注为一个整体，这往往成为机器翻译的死角。

再如，由于忽略地质英语词汇的特点，将 field data 直译为“领域上的数据”。这说明现有语料库对地学领域的语料收集不足，单词释义也缺乏融合专业背景的详尽标注。许多术语虽然由日常词汇构成，却有别于常规用法，不可“望词生义”，更不能将两个单词的词义简单叠加: field data 应该译为“野外数据”，field moisture 应该译为“土壤水分”，field capacity 应该译为“田间持水量”，oil field 应该译为“油田”。其他的例子还包括: ground water 不是“地上的水”而是“地下水”，guide fossil 不是“指导化石”而是“标准化石”，induced fracture 不是“引导裂缝”而是“次生裂缝”，oil recovery 不是“油恢复”而是“采油”，pressure buildup 不是“压力增加”而是“压力恢复”(何大顺 2007) 。

高璞等(2009) 认为，地质英语词汇的特点按照构成方式的不同可以分为: (1) 本专业特有的词汇，如 geology(地质学) 、mineral (矿石) 和 dinosaur(恐龙) ; (2) 与其他专业共有的词汇，如 reservoir(水力专业) 译为“水库”、plat form (交通专业) 译为“站台”; (3) 与日常生活共用的词汇，如 fault (平时译为“缺点”，地质含义为“断层”) 、basin(平时译为“盆或者脸盆”，地质含义为“盆地或者流域”) 、shear(平时译为“剪切”，地质含义为“受剪切破坏的面或者带”) 、graduate(平时译为“毕业或者毕业生”，地质含义为“刻度”) 、envelope(平时译为“封皮”，地质含义为“围岩”) 、horizon(平时译为“地平线”，地质含义为“层

位”)、joint(平时译为“接头”,地质含义为“节理”)。显然,上述因素会加大语料库构建过程中词汇的标注难度。

即便都是地学的相关文献,由于细分的专业不同,同一单词会呈现出不同含义,这使得词汇的标注过程更加复杂。例如,earth core 在普通地质学中译为“地核”,rare earth 在能源地质学中译为“稀有金属”,earth slide 在工程地质学中译为“滑坡”(林彻 1983)。有时候,同一词汇的含义在不同学科的地质著作中大相径庭。例如,当 trap 与地层、构造、沉积作用有关时,译为“圈闭”;与石油有关时,译为“油捕”;与火山岩有关时,则译为“暗色岩”。又如,deposit 与各种矿产、矿床类型的术语以及专有名词 Noranda,Quemont,Jerome 等连用时,通常译为“矿床”,而与表示各种沉积岩类型的术语联用时则译为“沉积”。不仅如此,某些词的单、复数形式也影响单词的含义,例如,单数 compass 译为“罗盘”,复数 compasses 则译为“圆规”;单数 earth 译为“地球”,复数 earths 译为“土族金属”;单数 fold 译为“褶曲”,复数 folds 译为“褶皱”;单数 scale 译为“比例尺”,复数 scales 译为“天平”(尹丽莉 2009)。遗憾的是,目前机器翻译系统尚不能识别、区分这些词汇及其形式所表意义上的细微差别。

3　研究结论

综上所述,我们应该加大带有行业背景的专业语料的收集力度,为完善地学领域的专用型语料库奠定坚实的“物质基础”。而语料库中的词汇是否能够被合理地标注,则成为语料库构建的重中之重。笔者认为,对于经常用到的固定搭配,要根据科技文献的写作特点,用整体标注替代分别标注;若通过机器翻译系统的自动标注软件难以实现词间“整合”,则在必要时采取自动标注后的人工核对或者人工标注;对于容易产生歧义的词汇,要基于规则和概率结合的方法,根据上下文和专业排除可能的歧义。

实际上,除了可以通过改进词汇的标注方式来实现语料库的维护和升级外,语料本身的质量也决定着机器翻译的质量。这要求在收集语料时,既要保证收录高质量的源语言语料,又要保证收录相应的高质量译文,如此,才能为语料的后期处理提供更多方便。

参考文献

巢文涵. 基于双语语料库的机器翻译关键技术研究 [D]. 国防科学技术大学博士学位论文, 2008.

陈群秀. 计算机辅助翻译系统漫谈 [Z]. 第十一届全国民族语言文字信息研讨会, 2007.

冯志伟. 机器翻译研究 [M]. 北京: 中国对外翻译出版公司, 2004.

冯志伟. 基于语料库的机器翻译系统 [J]. 术语标准化与信息技术, 2010(1).

高璞等. 石油地质英语词汇教学方法探析 [J]. 中国地质教育, 2009(4).

何大顺 何　春. 论地学专业文献的英汉翻译 [J]. 成都理工大学学报(社会科学版), 2007(4).

林　彻. 地质翻译参考 [M]. 北京: 地质出版社, 1983.

曲江秀 谭丽娟. 地质专业英语的特点和教学方法探讨 [J]. 中国科教创新导刊, 2008(19).

王建新. 计算机语料库的建设与应用 [M]. 北京: 清华大学出版社, 2005.

肖维青. 平行语料库与应用翻译研究 [J]. 中国科技翻译, 2007(3).

尹丽莉. 地质英语的词汇特点探析 [J]. 吉林地质, 2009(3).

Granger, S. *The Computer Learner Corpus: A Versatile New Source of Data for SLA Research* [M]. London/New York: Longman, 1998.

Mona, B. *Corpus Linguistics and Translation Studies: Implications and Applications* [M]. Amsterdam: John Benjamins Publishing Company, 1993.

收稿日期: 2013 - 03 - 31　　【责任编辑　王松鹤】