

中 华 人 民 共 和 国 教 育 部 主 管

中文社会科学索引(CSSCI)来源期刊 中文外国语类核心期刊

中国外语

(双月刊)

Foreign Languages in China

改革论坛 Forum on Teaching Reform

4 西部地区外语教育的生态语境及补偿机制研究

学术探索 Academic Exploration

23 中国英汉平行语料库的设计与研制

纵横论译 Views on Translation

64 典籍翻译：从语内翻译到语际翻译
——以《论语》英译为例

教师教育 Teacher Development

81 教师发展“大师复盘模式”：以商务英语教师发展为例
——顾曰国和Evan Frendo访谈录

ISSN 1672-9382



9 771672 938120

2012. 
总第50期

目 录

中国外语 (双月刊)

Zhongguo Waiyu

第9卷 第6期(总第50期) 2012年11月出版

中文社会科学索引(CSSCI) 来源期刊
中文外国语类核心期刊

编委会主任

桂诗春 胡壮麟

编委 (以姓氏笔画为序)

王 宁	王守仁	王初明	文秋芳	申 丹
石 坚	朱永生	刘 援	许 钧	仲伟合
吴一安	邹 申	杨治中	杨惠中	陆谷孙
沈家煊	陈建平	陈国华	张后尘	张柏然
张家骥	金 艳	罗立胜	罗选民	胡文仲
修 刚	钱冠连	徐烈炯	黄国文	戴炜栋

特约学术咨询机构

教育部高等学校人文社会科学重点研究基地:
北京外国语大学中国外语教育研究中心
广东外语外贸大学外国语言学及应用语言研究中心
黑龙江大学俄罗斯语言文学与文化研究中心

社长兼总编 刘 援

名誉主编 张后尘

主 编 黄国文

编辑部主任 常少华

主 管 中华人民共和国教育部

主 办 高等教育出版社

编辑单位 《中国外语》杂志社

社 址 北京市朝阳区惠新东街4号富盛大厦23层

邮 编 100029

电 话 010-58581353

传 真 010-58556496

网 址 <http://www.cflo.edu.cn>

在线投稿 www.flcjournal.com

E-mail flc@pub.hep.cn

出 版 高等教育出版社

印 刷 北京佳信达欣艺术印刷有限公司

国内总发行 北京报刊发行局

国内订购 全国各地邮局

国外发行 中国国际图书贸易总公司(北京399信箱)

中国标准

连续 CN11-5280/H

出版物号 ISSN 1672-9382

国内代号 80-350

国外代号 BM1911

国内定价 9.60元

期刊基本参数

CN11-5280/H*2004*b*16*112*zh+en*P*¥9.60*12000*17*2012-6

本刊已被CNKI中国期刊全文数据库、万方数据中国数字化期刊群收录,其作者文章著作权使用费与本刊稿酬一次性给付。本刊免费提供作者文章引用统计分析资料。如作者不同意文章被收录,请在来稿时向本刊声明。

■ 改革论坛

- | | |
|--|-----|
| 4 西部地区外语教育的生态语境及补偿机制研究 | 姜秋霞 |
| 11 新时期大学英语教学改革紧迫任务
——谈《大学英语课程教学要求》的修订 | 贾国栋 |

■ 学术探索

- | | |
|----------------------------------|---------|
| 23 中国英汉平行语料库的设计与研制 | 王克非 |
| 28 语料库翻译学:研究对象与研究方法 | 黄立波 朱志瑜 |
| 37 双语语料库建设中元信息的添加和段落与句子的
两级对齐 | 梁茂成 许家金 |
| 43 伯恩斯坦知识结构理论的系统功能语言学解读 | 于 晖 |

■ 教学研究

- | | |
|-------------------------|---------|
| 51 外语课堂隐性语言知识构建模式 | 陈亚平 曹荣平 |
| 57 基于学术语篇语料库的过渡标记语用失误调查 | 赵 颖 |

■ 纵横论译

- | | |
|--|---------|
| 64 典籍翻译:从语内翻译到语际翻译
——以《论语》英译为例 | 黄国文 |
| 72 “名从主人”?
——名称翻译的语用学思考 | 李 捷 何自然 |
| 77 对等理论指导下的英语经济新闻文体的难句翻译
——以《经济学人》和《商业周刊》上的两篇特别报道为例 | 王维平 |

■ 教师教育

- | | |
|---|-----|
| 81 教师发展“大师复盘模式”:以商务英语教师发展为例
——顾曰国和Evan Frendo访谈录 | 李玉香 |
| 86 ESP教师角色与教师专业发展 | 郭燕玲 |

■ 综述

- | | |
|------------------------|---------|
| 91 系统功能语言学的研究热点和发展方向 | 李战子 陆丹云 |
| 96 临床话语分析的系统功能语言学理据及途径 | 赵俊海 杨炳钧 |

■ 书评

- | | |
|--|-----|
| 102 语言教育中的微观政治
——《语言教育中的政治学——个人与机构》评介 | 王 捷 |
| 105 《历史认知语言学》述评
——认知语言学发展新观 | 王艳滨 |

■ 《中国外语》总目录(2012) 109

双语语料库建设中元信息的添加和段落与句子的两级对齐^①

梁茂成 许家金 北京外国语大学中国外语教育研究中心

摘 要: 双语平行文本的元信息添加和段落、句子两级对齐是双语语料库建设中的关键技术。以往的研究对双语平行文本的元信息添加少有涉及,而大多数有关双语文本对齐的研究主要涉及句级对齐技术,忽略段落信息。本研究设计开发了一个软件系统,该系统集双语平行文本元信息添加、段落级对齐和句级对齐等多种功能,十分有利于建设更加符合语言研究和翻译研究需要的双语平行语料库。

主题词: 双语平行语料库; 元信息; 对齐

中图分类号: H08

文献标识码: A

文章编号: 1672-9382(2012)06-0037-06

作者简介: 梁茂成,北京外国语大学中国外语教育研究中心教授,博士生导师,研究方向:应用语言学、语料库语言学, E-mail: frankliang0086@gmail.com。许家金,北京外国语大学中国外语教育研究中心副教授,博士,研究方向:语料库语言学、英汉对比研究, E-mail: xujiajin@bfsu.edu.cn。

1 引子

双语平行语料库建设是语言对比研究、翻译研究、机器翻译研究等领域十分重要的数据源。然而,双语平行语料库建设对技术的要求至少在以下几个方面要远高于单语语料库建设。首先,双语对应的两个文本中的元信息相互关联,比起单语文本的元信息标记更为复杂;其次,对齐是双语平行语料库建设中无法回避的环节,否则语料库的价值会大打折扣;再者,双语文本因涉及两种语言,字符类型更为丰富,文本应尽可能采用通用的编码(如Unicode编码);最后,双语平行语料库中文本的保存格式需要考虑计算机辅助翻译的需要,采用翻译记忆库交换文件(Translation Memory eXchange, tmx)的格式,这样可以更好地保证对齐文本的兼容性,便于格式转换。相反,由于单语语料库建设只涉及一种语

言,元信息创建、文本编码、文件保存格式等都更为简单。

基于以上考虑,国家社科基金重大项目“大规模英汉平行语料库的建立与加工”课题组设计并开发了一个人工辅助的双语文本处理系统,实现双语文本元信息的添加和双语文本的自动对齐。因段落是重要的语言单位,本研究中设计开发的软件系统在对双语文本进行自动对齐时将完好保存文本的段落信息,以便开展更加多维的语言对比及翻译研究。

2 国内外研究现状

2.1 双语文本元信息的添加

根据美国国家信息标准组织(National Information Standards Organization, NISO)的定义,元信息(metadata)是一种“结构化信息,用于描述或解释信息源,或指定信息

源的位置,或使得信息的提取、使用和管理变得更为方便”。概括地说,元信息是“有关数据的数据(data about data)” (National Information Standards Organization, 2004: 1; Burnard, 2005: 1)。因元信息常置于文本的头部,故又常被称为语料库的“头部信息”(header)。

在语料库文本中,元信息常用来提供文本的来源、语言产出者身份、语言产出年代、语言类型、语体、翻译方向等语言研究和翻译研究中的重要信息。将这些信息添加至语料库文本中,将极大地方便语言研究和翻译研究。反之,如果语料库文本中缺失这些信息,文本的价值会大打折扣。

从国内外元信息标准制定的现状看,一些标准化组织先后就文本元信息类型提出了多种规范化方案,最知名的有the Dublin Core Metadata Initiative (DCMI)、Text Encoding Initiative (TEI), the Open Language Archive Community (OLAC)、the ISLE Metadata Initiative (IMDI)等 (Burnard, 2005: 1)。我国也于2007年发布实施了《电子文件元数据标准》。

在机器翻译研究领域,双语文本的元信息远远没有受到应有的重视。研究者们似乎没有意识到文本中元信息的价值,用于训练机器翻译系统的大部分双语文本常常以简单的句对(sentence pair)形式保存,严重缺乏各类元信息,有的甚至连翻译方向信息也缺失。机器翻译领域在处理双语文本时另外一个常见问题是段落信息的缺失。对齐文本常常以翻译记忆库交换文件格式保存,仅保留翻译方向和句对齐信息,缺失段落信息、文本产生者身份信息、文本产生(或翻译)年代信息等。元信息的缺失会对机器翻译质量产生重要影响。

在语料库语言学界,研究者们深知元信息的价值,大多完好地保留了尽可能多的元信息。然而,如何将各类元信息合理、准确地添加到语料库文本之中,做法却各不相同,且学界有关语料库元信息规范及其添加方法的研究甚少。因此,语料库元信息缺乏统一的标准和规范。不仅如此,迄今为止,我们未见任何可供普通语料库建设者和语言研究者将元信息添加到文本中的软件系统。

2.2 双语文本的自动对齐

随着机器翻译研究的不断深入和语料库语言学的不断发展,近20年来,双语句级自动对齐的研究受到国际自然语言处理领域学者的高度重视。

Gale & Church (1991) 首先提出基于长度的方法对双语语料库中的文本进行自动句级对齐。研究基于这样的事实:源语言的句子长度和目的语言的句子长度具有高度的相关性。换言之,源语言中较长的句子翻译成目的语后句子一般也会较长。相反,源语言中较短的句子翻译成目的语后句子也一般较短。在Gale & Church (1991) 的研究中,源语言为英语,目的语为法语和德语。研究表明,这一简单的基于长度的方法取得了令人瞩目的结果,准确率超过90%。Gale & Church (1991) 认为,由于基于长度对双语文本进行自动句级对齐的算法不依赖任何外部词典,适用于几乎各种自然语言。

继Gale & Church (1991) 之后,双语文本自动对齐的研究大量涌现。

Wu (1994) 采用同样的方法对英汉文本进行了自动对齐的尝试,发现Gale & Church (1991) 的方法在处理英汉文本时效果并不十分理想。Wu (1994) 认为,印欧语系内的语言谱系关系亲密,具有一些共同特征,因而Gale & Church (1991) 所采用的方法效果更好。对于英语和汉语这两种差异较大的语言,采用单纯的基于长度的方法,效果并不会十分理想。因此,Wu (1994) 在此基础上,利用领域词典(domain-specific lexicon)来提高对齐的准确率。程序在基于长度对双语文本进行对齐的同时,读取一个外部词典。由于研究中采用的词典是领域内的专门词典,Wu (1994) 的研究也取得了很好的结果,对齐的准确率也超过95%。Kay & Rischeisen (1993) 的研究根据源语文本和目的语文本中词汇分布的特点发现两个文本中的对应词,并将对应词作为评判句对的依据,依此来实现双语文本的句级对齐。Ma (2006) 指出,大多现有的研究主要针对相对整洁的双语文本,而在现实任务中,可以获取到的大部分数据并不整洁,翻译过程中的插入(insertion)和删除(deletion)现象十分常见。为此,Ma (2006) 设计了一个名为Champollion的程序,外挂一个大型词典,处理不整洁的双语文本十分有效。此外,Zhao et al. (2000)、Deng & Byrne (2006) 等研究也曾尝试从不同类型的双语数据中发现句对。有研究甚至尝试从网络文本中挖掘句对,为机器翻译构建庞大的翻译记忆库(如Ture & Lin, 2012)。

从实现自动对齐的算法看,已有研究大致可以分为3类:(1) 基于长度的方法,即根据源语言文本和目的语言文本的长度比来确定句对。这种方法简单而不依赖词典等外部资源,

效率高但准确率并不十分理想,对谱系关系不密切的语言而言,效果远不及Gale & Church (1991),可用程度不高。(2)基于词典的方法,即根据外挂的双语词典来确定句对。这种方法准确率较高,但召回率相对较低,且对齐的效果在很大程度上取决于外挂词典的质量。如果外挂词典与双语文本不属于同一学科,对齐的效果会很不理想。因此,对于归属于不同学科领域的文本,常常需要不同的词典。对于通用平行语料库建设者来说,构建多个学科领域词典有很大难度。(3)长度和词典相结合的方法。这种方法既考虑文本的长度,又借助外挂词典提供锚点对依据长度对齐的结果进行优化和微调,因而效果往往更好。由于这种算法中结合了长度和词典,即便没有好的领域词典,也能保证较高的对齐准确率。

然而,以上研究不同程度地存在一些问题。

首先,在这些研究中,Gale & Church (1991)、Wu (1994)、Zhao et al. (2000)要么没有直接推出任何可以供普通语料库建设者使用、有图形界面的软件系统,研究限于实验室阶段,要么只是报告外界无从获取的软件系统。只有Ma (2006)的研究推出了用户可以下载、运行的免费软件系统Champollion。遗憾的是,Champollion只能在Unix操作系统中编译和使用,将大部分Windows用户拒之门外。

其次,以上所有研究都基于双语文本挖掘句对,除Gale & Church (1991)之外,都忽略了文本中的段落信息,有的甚至只将确认的句对从双语文本中提取出来,抛弃系统不能确认的数据,严重破坏了文本的完整性。我们认为,段落信息是语言研究中的重要数据。只注重句子间的对应、忽略段落间的对应是不符合

语言研究和翻译研究需要的。

除了以上研究之外,市场上还有一些商用系统,如Trados公司的WinAlign、Wordfast LLC公司的Wordfast、Adobe公司的ABBYY Aligner等。这些系统同样也都忽略了双语文本的段落信息。此外,这些系统价格昂贵,用户界面也不够友好。

3 英汉双语平行文本自动对齐系统

针对以上现状,在国家社科基金重大项目“大规模英汉平行语料库的建立与加工”立项之初,课题组就曾多次讨论课题实施中的技术需求。基于对现有软件系统的考察、分析和试用,课题组提出,在语料库建设初期,首先需要开发一个专门用于添加语料库文本元信息的软件和一个英汉双语文本自动对齐软件。目前,这两个软件系统已经全部完成,并合成为“英汉双语平行文本自动对齐系统”,包括功能分离的两个模块。

3.1 模块一: HeadAppend, 语料库文本元信息添加模块

HeadAppend的主要用途包括:

(1) 为语料库文本添加元信息。该系统同时适用于为单语文本和双语文本添加语料库元信息。

(2) 辅助人工对双语文本进行段落对齐。段落是双语对应语料库中的重要对齐单位。实现双语段落对齐后,可通过模块二CETA (参见3.2节) 轻松实现自动句级对齐。

如图1所示,用户可以通过点击界面左上角的单选按钮选择“单语”模式或“双语”模式。

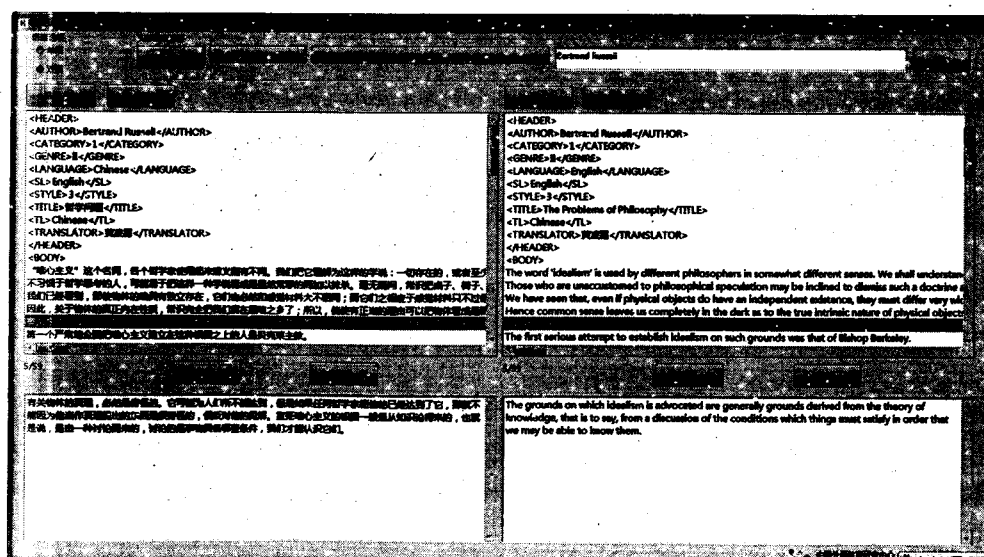


图1 HeadAppend (语料库文本元信息添加工具)的主界面

之后, 点击【选择文件】按钮选择语料库文本, 然后点击界面上方的组合框, 从组合框的下拉菜单中选择需要添加的元信息选项, 点击【添加】按钮后, 选中的元信息便会自动添加

到文本中。此外, HeadAppend还允许用户对文本进行段对齐调整。保存结果后, 元信息便会自动添加到语料文本的头部, 可得到图2所示的文本:

```
<HEADER>
<AUTHOR>Bertrand Russell</AUTHOR>
<CATEGORY>1</CATEGORY>
<GENRE>II</GENRE>
<LANGUAGE>Chinese</LANGUAGE>
<SL>English</SL>
<STYLE>3</STYLE>
<TITLE>哲学问题</TITLE>
<TL>Chinese</TL>
<TRANSLATOR>黄凌霜</TRANSLATOR>
</HEADER>
<BODY>
    “唯心主义”这个名词, 各个哲学家使用起来意义稍有不同。我们把它理解为这样的学说: 一切存在的, 或者至少, 一切为人所知道是存在的, 在某种意义上, 都必然是精神的。这种学说在哲学家中受到了广泛支持, 并且有好几种形式, 以不同的理由被人提倡着。这种学说是如此广泛受人支持, 它本身又是如此饶有趣味, 所以就连最简略的哲学概论也要对它叙述一番。
    不习惯于哲学思考的人, 可能易于把这样一种学说看成是显然荒谬的而加以抹杀。毫无疑问, 常识把桌子、椅子、太阳、月亮和一些物质客体, 一般都认为是根本不同于心灵以及心灵的内容的, 认为它们有着一一种存在, 即使心灵不复存在, 它们也还是会继续存在的。我们想象, 早在任何心灵存在之前, 物质就已经存在了; 而很难想象物质只是精神活动的一种产物。但是, 不管唯心主义是真还是妄, 总不能把它看成是显然荒谬的而加以抹杀。
    ...
</BODY>
```

图2 保存后带有元信息的文本

HeadAppend配置有一个模板文件, 用户可以通过该文件来自定义各种元信息标签类型。程序启动后, HeadAppend会自动读取模板文件中的内容, 并将这些内容加载到程序的元信息组合框的下拉菜单中供用户选择。这样做的好处是, 用户选择起来十分方便而且不会出错。

特别值得一提的是, 双语平行文本的元信息相互关联。比如, 如果汉语文本是目的语 (target language, TL) 文本, 则英语文本必然是源语言 (source language, SL) 文本。考虑到这一点, HeadAppend程序中内嵌了一个翻译方向自动识别子程序。当用户选择<TL>Chinese</TL>标签时, 系统会同时自动增加<SL>English</English>标签, 极大地方便了用户的操作, 也减少了出错的几率。

3.2 模块二: CETA, 英汉双语文本自动句级对齐模块

英汉双语文本自动句级对齐模块英文是CETA, 即Chinese English Text Aligner。CETA

的主要用途是: 对段落级对齐并带有元信息的英、汉双语文本进行自动句级对齐, 以辅助双语平行语料库建设。对于没有进行段落级对齐的文本, CETA将拒绝操作。CETA的操作界面如图3所示。

与现有的大部分双语对齐软件相比, CETA有以下几方面的特点:

(1) CETA完整地保留了段落信息。几乎所有的计算机辅助翻译系统 (如Trados等) 在读取训练语料时只要求用户提供句级对齐的双语文本, 并不需要用户提供段落信息。现有的对齐软件 (如HunAlign、WinAlign、ABBYY Aligner等) 在执行自动对齐时需要用户提供每行一句的文本, 系统在进行句级对齐时已无法恢复段落信息, 造成语言信息损失。CETA则不同, 需要用户提供段落对齐的文本, 然后对每一段落内的内容进行句级对齐。这样, CETA在对双语文本进行句级对齐之后, 既保留了段落信息, 也保留了句级对齐信息。CETA的tmx格式输出结果参见图4。



图3 CETA 的操作界面

```
<HEADER>
<AUTHOR>BERTRAND RUSSELL</AUTHOR>
<CATEGORY>1</CATEGORY>
<GENRE>II</GENRE>
<SL>ENGLISH</SL>
<STYLE>3</STYLE>
<TITLE_CN>哲学问题</TITLE_CN>
<TITLE_EN>THE PROBLEMS OF PHILOSOPHY</TITLE_EN>
<TL>CHINESE</TL>
<TRANSLATOR>黄凌霜</TRANSLATOR>
</HEADER>
<BODY>
<EN id=1_1>The word 'idealism' is used by different philosophers in somewhat different senses.</EN>
<CN id=1_1>“唯心主义”这个名词，各个哲学家使用起来意义稍有不同。我们把它理解为这样的学说:</CN>
...
<EN id=2_1>Those who are unaccustomed to philosophical speculation may be inclined to dismiss such a doctrine as obviously absurd. There is no doubt that common sense regards tables and chairs and the sun and moon and material objects generally as something radically different from minds and the contents of minds, and as having an existence which might continue if minds ceased.</EN>
<CN id=2_1>不习惯于哲学思考的人，可能易于把这样一种学说看成是显然荒谬的而加以抹杀。毫无疑问，常识把桌子、椅子、太阳、月亮和一些物质客体，一般都认为是根本不同于心灵以及心灵的内容的，认为它们有着一一种存在，即使心灵不复存在，它们也还是会继续存在的。</CN>
...
</BODY>
```

图4 CETA 的输出格式

如图4所示，CETA的输出文件结构清楚，文件中不仅包含丰富的元信息，同时包含段落信息。例如，<EN id=2_1>表示当前句子为英文，是原文中第2段的第1句，而<CN id=2_1>则表示当前句为汉语，是原文第2段的第1句，且英语句子<EN id=2_1>和汉语句子<CN id=2_1>是对应的。这种标注方法十分有利于在

分析翻译语言时将段落也作为考察因素之一。

(2) CETA界面友好，操作简便。CETA汲取了现有双语对齐软件的一些长处，界面友好，操作简便。CETA不像WinAlign等软件那样在界面上显示小号字体，造成用户的视觉疲劳。对CETA自动对齐中可能出现的错误，用户修改起来也十分容易，只需要选中需要

挪动的文本再点击对应按钮即可。WinAlign、ABBY Aligner等系统通过人工拉线的方式实现对齐,虽然直观,但拆分和合并句子并不那么简便。CETA对文本的多个段落逐一进行句级对齐,即便出现了错误,也只是用户很容易修正的局部错误,从而大大降低了操作的复杂程度。

(3) CETA自动对齐准确率高。大多数现有的对齐系统都一次性读入每行一句的文本。由于需要处理的文本有时会很长,可能包含很多句子,对齐操作中算法误差的积累可能会大大降低对齐的准确率。因此,这种做法不光丢失了段落信息,还可能会导致对齐准确率的降低。我们认为,翻译一般不改变原文的段落信息,即源语言的一个段落翻译成目的语时仍然是一个段落,因此在实施句级对齐之前可以先进行段落对齐,程序加载语料后,每次只读入两个已经对齐的段落,这样可以有效控制算法误差的积累。由于CETA的对齐操作在段落之内进行,不仅保持了段落信息,还可以大大降低出错的几率。

除此之外,在对段落内的句子进行对齐操作时,CETA采用了基于文本长度和基于词典相结合的方法,以最大限度地提高对齐的准确率。先前的研究(王克非、胡显耀,2008)表明,翻译语言较之原创语言具有明显的特征,因而在CETA程序内分别嵌入一个英汉通用词典和一个汉英通用词典,程序可以根据文本元信息中提供的翻译方向信息自动选择不同的词典。

(4) CETA提供了动态预览功能,用户可以随时预览或审核对齐的结果。CETA专门设计了一个预览模式选项卡,供用户预览对齐结果。在预览模式下,用户只能看到已经确认的段落内的句对,并对确认后的结果进行动态刷新,这样可以迫使用户对每一段落进行确认,不至于因一时疏忽而导致句子的不对齐,有利于提高对齐的准确率。

(5) CETA的输出结果文件可以以不同格式保存,更增强了输出文本的兼容性,也省去了用户的格式转

换之苦。CETA的输出结果可以是一个翻译记忆库交换文件、一个包含双语的纯文本文件或两个仅包含单语的纯文本文件。用户在保存结果时,可以根据需要选择保存方式。

课题组的语料库建设实践表明,本文介绍的英汉双语文本元信息添加和自动对齐工具因由两个独立运行的模块组成,不仅便于双语语料库建设的分工,而且分工后对齐操作简便,自动对齐准确率高。经对10个非文学文本(中文单语共包含约220万字)的先导性测试,句子对齐准确率达到95%以上。

4 结语

建设超大规模的英汉双语平行语料库对机器翻译、语言对比研究和翻译研究都具有重要意义。然而,语料库建设需要尽可能多地考虑各类研究者的需要,对文本的元信息进行充分的标注。此外,出于翻译和语言学的需要,在对双语语料库中的文本进行句级对齐之前,有必要先对段落进行对齐处理,并将段落信息保存到对齐结果中。基于以上理念,国家社科基金重大项目“大规模英汉平行语料库的建立与加工”课题组设计了英汉双语语料库元信息添加和自动对齐系统。该系统既可以将各类元信息完好地保存到语料库文本中,在段落对齐基础上的句级对齐更丰富了研究视角,同时还提高了句级对齐的准确率,对大型双语语料库建设具有重要意义。□

注释

- ① 本研究是王克非教授主持的国家社科基金重大招标项目“大规模英汉平行语料库的建立与加工(编号10 & ZD127)”和本文作者主持的国家社科基金项目“中国学生英语树库建设与研究(编号09BYY033)”的部分研究成果。

参考文献

- [1] Burnard, L. Metadata for corpus work [A]. In Wynne, M. (ed). *Developing Linguistic Corpora: A Guide to Good Practice* [C]. Oxford: Oxbow Books. Available online from <http://ahds.ac.uk/linguistic-corpora/> [accessed 2012-03-28], 2005.
- [2] Deng, Y. & W. Byrne. MTTK: An alignment toolkit for statistical machine translation [A]. Paper presented in the HLT-NAACL Demonstrations Program [C]. 2006.
- [3] Gale, W. A. & K. W. Church. A program for aligning sentences in bilingual corpora [A]. In *Proceedings of the 29th Annual Conference of the Association for Computational Linguistics* [C]. 1991.
- [4] Kay, M. & M. Rscheisen. Text-translation alignment [J]. *Computational Linguistics - Special Issue on Using Large Corpora: I*, 1993. 19(1): 121-142.
- [5] Ma, X. Champollion: a robust parallel text sentence aligner [A]. In *LREC 2006: Fifth International Conference on Language Resources and Evaluation* [C]. Genova, Italy, 2006.
- [6] National Information Standards Organization. *Understanding Metadata* [R]. Bethesda, MD: NISO Press, 2004.
- [7] Ture, F. & J. Lin. Why not grab a free lunch? Mining large corpora for parallel sentences to improve translation modeling [A]. In *Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'12)* [C]. 2012.
- [8] Wu, D. Aligning a parallel English-Chinese corpus statistically with lexical criteria [A]. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics* [C]. Las Cruces, NM., 1994.

(下转63页)

- of native and nonnative EFL speakers of English [J]. *World Englishes*, 1996(1):17-27.
- [9] Halliday, M.A.K. & R. Hasan. *Cohesion in English* [M]. London: Longman Group Ltd., 1976.
- [10] Hyland, K. *Metadiscourse* [M]. Beijing: Foreign Language and Research Press/Continuum, 2005.
- [11] Lobov, W. *Language in the Inner City: Studies in the Black English Vernacular* [M]. Philadelphia: University of Pennsylvania Press, 1972.
- [12] Martin, J. & Rose, D. *Working with Discourse: Meaning Beyond the Clause* [M]. London: Continuum, 2003.
- [13] Sweetser, E. *From Etymology to Pragmatics: Metaphorical and Cultural Aspects of Semantic Structure* [M]. Beijing: Beijing University Press, 2002.
- [14] Van Dijk, T. A. *Text and Context: The Semantics and Pragmatics of Functional Coherence in Discourse* [M]. London: Longman, 1977.
- [15] Vande Kopple, W. Some exploratory discourse on metadiscourse [J]. *College Composition and Communication*, 1985(36): 82-93.
- [16] Van Dijk, Teun A. *Text and Context: Explorations in the Semantics and Pragmatics of Discourse* [M]. London: Longman Group Ltd., 1989.
- [17] Widowson, H. G. *Teaching Language as Communication* [M]. Oxford: Oxford University Press, 1978.
- [18] 梁茂成. 学习者书面语语篇连贯性的研究 [J]. *现代外语*, 2006, 29(3): 284-330.
- [19] 许双如. 语篇衔接与英语习得程度调查分析 [J]. *广州大学学报 (社科版)*, 2004(5).
- [20] 张谊生. *现代汉语虚词* [M]. 上海: 华东师范大学出版社, 2000.
- [21] 赵颖. 基于语料库分析的“and”隐喻 [J]. *天津大学学报 (社会科学版)*, 2006(36): 429-432.
- [22] 中国社会科学院语言研究所词典编辑室. *现代汉语词典 (修订本)* [Z]. 北京: 商务印书馆, 2000.

A Corpus-based Study of Pragmatic Failure of English Transition Markers in Academic Writings

Abstract: Based on CH-UK corpus made up of 24 academic writings by Chinese English majors and 39 academic writings by their British counterparts on linguistic theories acquired from Linguistics Courses, the paper focuses its study on the pragmatic failure in the use of transition markers by Chinese English majors. With the help of WordSmith Tool.4.0 and SPSS.13.0, Chinese learners are found that their use of some transition markers is quite different from that of British English majors, for the result of the K-square Test shows great significant differences for some transition markers ($p < 0.05$), which indicates that Chinese English majors have come across some pragmatic problems, even if they are good at English grammar. The investigation is also conducted to find out what caused such pragmatic failures.

Key Words: transition marker; academic writing; language transfer; mode of thinking

(上接42页)

- [9] Zhao, T, M. Yang, L. Qian & G. Fang. Bilingual dictionary based sentence alignment for Chinese English bitext [A]. In *Proceeding ICMI '00 Proceedings of the Third International Conference on Advances in Multimodal Interfaces* [C]. 2000: 253-259.
- [10] 王克非, 胡显耀. 基于语料库的翻译汉语词汇特征研究 [J]. *中国翻译*, 2008(6): 16-21.

Metadata Creation and Paragraph-level and Sentence-level Alignment in Parallel Corpus Construction

Abstract: Metadata creation, paragraph-level alignment and sentence-level alignment are key

technical operations in parallel corpus construction. Previous studies on metadata creation, however, are very rare, and paragraph information has often been neglected in most research on bilingual text alignment, with most efforts directed at aligning sentences in bilingual texts. This study reports the design and the development of a software system which can greatly facilitate metadata creation, paragraph-level alignment as well as sentence-level alignment. The software, therefore, can be of great help in building parallel corpora which can better meet the needs of linguistic and translation studies.

Key Words: parallel corpora; metadata; alignment

中国外语

Foreign Languages in China

(双月刊)



高等教育出版社
中国标准连续出版物号

www.hep.edu.cn
CN 11-5280/H
ISSN 1672-9382

100029 北京市朝阳区惠新东街4号富盛大厦23层
国内代号 80-350 定价 9.60元