构建大规模的汉英双语平行语料库[1]

柏晓静[1] 常宝宝[1] 詹卫东[1,2] 吴拥华[1]

[1](北京大学计算语言学研究所, 北京 100871)

[2](北京大学中文系, 北京 100871)

E-mail: {baixj, chbb, zwd, wyongh}@pku.edu.cn

摘要：双语语料库在机器翻译研究中的作用已日趋明显，但作为一项重要的语言资源，双语平行语料库的系统性构建在中国国内尚未得到充分的关注。本文介绍一个大规模汉英双语平行语料库的构建工作，包括其总体规划、实施模型和流程细节。该工作的深入和展开将促进作为机器翻译基础资源的双语语料库建设，从而推动相关的理论研究和应用技术不断向前发展。

关键词：机器翻译; 双语平行语料库; 语料库构建

引言

近年来，双语平行语料库在机器翻译和机器辅助翻译中的应用已经得到越来越多的认可，基于双语平行语料库的各种方法不仅能够改进机器自动翻译的质量，还可以加强机器辅助翻译中的人机交互。目前在中国国内，相关的研究和介绍主要侧重于双语语料的对齐技术和双语平行语料的应用技术，但对大规模双语平行语料库的系统性构建却关注较少。就汉英对照语料而言，国内尚且没有超过 10 万句对的平行语料库。作为一项重要的基础资源，双语平行语料库的建设仍处于滞后状态，影响了相关的理论研究和应用技术的发展。

北京大学计算语言学研究所同中国科学院计算技术研究所、清华大学智能技术国家重点实验室联合开发“面向新闻领域的汉英机器翻译系统”。在这个采用多引擎机制的机器翻译系统中，双语平行语料库将主要服务于基于存储的翻译引擎。作为该课题的子任务之一，一个大规模汉英平行语料库正在建设之中。本文介绍我们构建这个汉英平行语料库的系统性流程以及该语料库目前的建设情况。论文第 1 节总体介绍语料库构建的规划和模型，第 2 节详细介绍语料库构建的流程和相关经验，以及语料库现状的基本统计数据，最后是对进一步工作的展望。

1 语料库构建的规划和模型

构建大规模双语平行语料库，现阶段的应用目标是一个多引擎结构的汉英机器翻译系统。我们用这个语料库为基于存储的引擎提供翻译实例，并从中挖掘学习各种细粒度翻译知识，供其他翻译引擎使用。此外，我们也希望该语料库在逐步趋向平衡后，能够服务于双语词典编纂、双语术语自动提取、双语对比研究以及双语教学等其他研究领域。

在北大计算语言学研究所英汉机译 MTE 测试集语料的基础上，我们将整理、加工大量汉英对照的真实语料，同时继续收集语料，建成一个大规模的双语平行语料库。目前已经收

[1]本文工作得到国家 973 项目资助（项目编号：G1998030507-4）。

集到汉英对照语料中文约 2,000 万字、英文约 1,000 万单词，包括政府白皮书、政府公文、新闻、杂文、演讲词、科技文献、学术论文、政治专著、法律文献、小说、剧本、诗歌、杂文、圣经、神话、口语语料等。

构建双语平行语料库的核心任务是双语语料的加工和语料库的组织，为了更好地开展这两项工作，保证语料库的质量和规模，并且合理、有效地推进语料库建设，我们需要一个相对完整、便于操作的语料库构建流程。为此，我们对双语语料本身以及语料的整理和加工、语料库的组织和检索等项任务进行了考察，分析问题的复杂性，初步形成了一个双语平行语料库构建流程的模型（见图 1），并为流程的各个环节开发了相应的辅助工具。

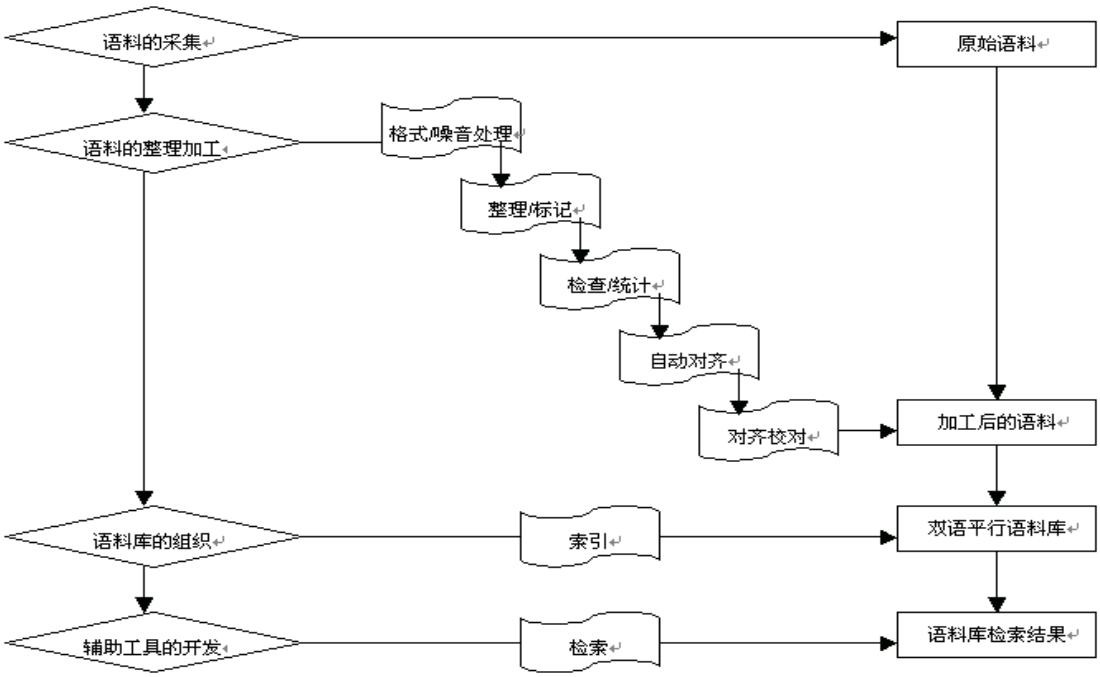

图 1 双语平行语料库构建流程模型

2 语料库的构建

2.1 语料的采集

在进行语料采集时，需要考虑两点因素：一是原始语料本身的质量，二是语料库的应用目标。

语料的质量主要就其语言质量、翻译质量及语料保存的规范性而言。我们的原始语料大部分从因特网上下载。在实际操作中，我们的体会是，《英语世界》等双语期刊发布的电子版双语语料（杂文居多）、以及官方网站发布的双语语料（政府公文和正式新闻居多）的语言和翻译质量较高。此外，语料的保存格式是否有利于整理加工、语料中乱码的情况等因素是衡量语料规范性的重要指标。

语料的采集还应根据语料库建设的近期目标和长远目标来灵活安排。我们收集的双语语料涉及不同的文体、领域、语体和创作时期。由于这个语料库的直接应用目标是面向新闻领域的汉英机器翻译系统，理想的语料应该是新闻领域的汉英语料，但是可用的汉英、英汉新闻语料是非常有限的。鉴于非新闻领域的双语语料也将有助于翻译知识的获取，又鉴于双语

语料库作为一项基础资源还将有其它的应用目标，我们遵循新闻领域优先、但不限于新闻领域的原则，在已收集到的双语语料中筛选出一部分先进行整理、加工。这部分语料包括政府白皮书、政府公文、新闻、杂文、演讲词、科技文献、学术论文、政治专著等。目前我们已经积累句子一级对齐的汉英双语语料近 55,000 句对(含 MTE 测试集)，预计今年年底还将增加约 55,000 句对。剩余语料的整理和加工今后将分期分批进行。语料的收集工作仍在继续，新加入的语料将有助于改善我们这个双语语料库的平衡性，从而扩大其应用范围。

2.2 语料的整理和加工

整理、加工任务须从考察原始语料的物理特征并进行相应的归类入手。综合考虑原始语料本身的情况和语料库今后的应用方向，我们对整理、加工任务进行了需求分析，形成了一个双语语料库整理、加工方案。配合该方案中各个环节的实施，我们制定了相关的规范、工作手册和一个贯穿整个语料库构建流程的 XML 标记集[2]，同时还开发了一套通用性较好的整理、加工工具[3]，协同工作人员完成语料的整理、标记、检查、对齐和校对工作。

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”。

2.2.1 语料的整理

原始语料来源于不同的收集者，其中大部分都处于杂乱无章的状态，表现为：存放方式各异，没有形成篇章级对齐单位；文体、领域、语体、创作时期各异；含有不利于加工处理的噪音信息；文本的排版格式不规范；有重复的语料等。这些因素均妨碍了对原始语料的进一步加工和利用，因此，必须对原始语料进行系统的整理，使之达到以下三项要求：1）形成内容不重复、保存格式统一、排版格式一致的篇章级对齐单位若干；2）篇章级对齐单位带有基本信息标记；3）不含噪音及其他不利于加工处理的因素。

一般情况下，相同出处的语料具有相同或相似的存放格式、排版方式、噪音类型，比如从同一网址下载的《毛选》，又如来自同一网站的双语新闻。我们参考语料的出处，对语料进行了粗分类，并在此基础上用格式/噪音处理工具分别对各类语料进行预处理。之后，由工作人员在整理/标记工具的辅助下，对各类语料进行文件、内容、格式和标记等四个层次的整理。

---

[2]用于标注基本文本结构、对齐单位、文体、领域、语体、创作时期等信息，所有标记均采用 XML 格式，充分考虑了标记的可扩充性和语料的可交换性。（见表 1 XML 标记集）

[3]即格式/噪音处理工具、整理/标记工具、检查/统计工具、（段落/句子）自动对齐程序、（自动对齐结果）辅助校对工具。

整理产生纯文本格式的篇章级对齐单位若干，每个篇章级对齐单位中原文文件与译文文件的翻译关系用相同的文件名（不同的后缀）来体现，篇章级对齐单位的源语言确定；原文文件、译文文件中没有多余的空格，没有落单的标号，段落单独成行（带段落标记），段首无空格，段尾有合理的段落结束符号，段落与段落之间保留一个空行；篇章级对齐单位的基本属性以及错误、噪音等信息都在其原文文件和译文文件中标记出来（参见表 XML 标记集）。

对语料进行分类时，会面临一个分类标准的问题。对此，我们的解决方案是通过标记每个篇章级对齐单位的文体、领域、语体和创作时期等基本属性，对库中的语料进行多层次的分类。语料按文体分为 3 类（文学、应用文、新闻[4]）；按领域分为 6 类（艺术、工商、政治、科技、体育、社会文化，允许兼类）；按语体分为 2 类（书面语、口语）；按创作时期，源语言为中文的语料分为 4 类（古代、近代、现代、当代），源语言为英文的语料分为 4 类（Old English、Middle English、Early Modern English、Present-day English）。多层次的分类方法能够灵活地反映语料库的构成、方便整个语料库基础之上的子语料库抽取，更重要的是，分类信息的统计结果将从不同的角度对构建一个平衡的大规模双语平行语料库起指导性作用。

对上述整理结果，我们用检查/统计工具来检查标记的合法性，并根据具体需要进行标记相关的统计，如语料类型分布情况、段落和句对数量、错误和噪音标记数量[5]等等。

2.2.2 语料的加工

语料的加工是语料库系统性构建中的一个重要环节。鉴于深层次加工必须建立在浅层次加工的基础之上，并且经过浅层次加工的双语语料库在机器翻译研究中也会有直接的应用价值，我们将加工环节首先定位在浅层次上，即篇章、段落和句子级的双语对齐。经过系统的整理，规范的篇章级对齐单位已经形成，因此，加工的具体任务是：1）用自动对齐程序标记段落/句子边界并完成段落级/句子级的双语对齐；2）对自动对齐结果进行人工校对，得到带正确的段落/句子边界标记和对齐标记的双语平行语料库。

我们的段落/句子自动对齐程序采用基于长度的方法。段落级、句子级对齐可以是一对一、一对多、多对一甚至多对多的。段落/句子边界标记采用 XML 格式。其中，段落标记为 … ， 标记段落的开始， 标记段落的结尾，该标记有一个属性 id，表示段落的编号，取值范围为 1…n，n 是文件的总段数；句子的标记为<s>…</s>，<s>标记句子的开始，</s>标记句子的结尾，该标记有一个属性 id，表示段内的句子编号，取值范围为 1…n，n 是段内句子数。段落级/句子级对齐标记也采用 XML 格式。由于已经经过段落边界标记，对齐单位标记只标记到句子一级，不明确标记到段落一级，段落级对齐单位可由句子级对齐单位和段落边界标记导出。对齐成分在原文文件和译文文件中均用<a>…标记，<a>标记对齐成分的开始，标记对齐成分的结尾，该标记有两个属性：1）id，表示对齐成分的编号，取值范围为 1…n，n 是文件的对齐成分数；2）no，表示对齐成分内的句子数，取值范围为 1…n。一个篇章级对齐单位的原文文件和译文文件中编号相同的一对对齐成分构成一个句子级对齐单位。<s>…</s>标记嵌套在<a>…标记内部，<a>…标记又嵌套在 … 标记内部。

自动对齐结果仍需要人工校对，为此我们开发了一个辅助校对工具，方便工作人员根据规范和工作手册的要求，调整原文、译文之间的对齐关系，并从格式、标记及内容等方面对单个句对进行的细节性修改。

---

[4]仅设这三类文体类别，主要是考虑到现有语料的实际情况和这个语料库的直接应用目标。

[5]被标注为错误的内容（如乱码），留待校对自动对齐结果时修改；被标注为噪音的内容（如与正文无关的网页信息、重复信息、无关标记等），将被删除。

表 1 XML 标记集

被标记内容 篇头部分[6]	标记	被标记内容 正文部分[7]	标记
篇头	<TEXT_HEAD>…</TEXT_HEAD>	正文	<TEXT_BODY>…</TEXT_BODY>
中文标题	<CH_TITLE>…</CH_TITLE>	中文标题	<CH_TITLE>…</CH_TITLE>
英文标题	<EN_TITLE>…</EN_TITLE>	英文标题	<EN_TITLE>…</EN_TITLE>
作者名	<AUTHOR>…</AUTHOR>	作者名	<AUTHOR>…</AUTHOR>
译者名	<TRANSLATOR>…</TRANSLATOR>	译者名	<TRANSLATOR>…</TRANSLATOR>
文体	<STYLE>…</STYLE>	创作时间	<TIME>…</TIME>
领域	<FIELD>…</FIELD>	子标题	<SUBTITLE>…</SUBTITLE>
语体	<MODE>…</MODE>	图表公式和程序源码	<DIAGRAM>…</DIAGRAM>
创作时期	<PERIOD>…</PERIOD>	单语背景信息	<BACKGROUND>…</BACKGROUND>
错误	<ERROR>…</ERROR>	错误	<ERROR>…</ERROR>
噪音	<NOISE>…</NOISE>	噪音	<NOISE>…</NOISE>
		段落边界	…
		句子级对齐单位	<a>…
		句子边界	<s>…</s>

2.3 语料库的组织

整理和加工后的语料通过语料库索引结构关联成为一个整体，语料库的组织正是通过索引结构来实现的。在设计索引结构时，应充分考虑语料库的实际应用，为语料库的各种应用技术提供尽可能多的方便。基于这种想法，我们从三个层面为语料库建立了索引。

首先，我们为双语平行语料库设置了一个文本信息数据库，每个篇章级对齐单位在该数据库中都有一个记录，包含文件名、中英文标题、作者、译者、文体、领域、语体、创作时期等信息。文件名可以指向该篇章级对齐单位的原文文件和译文文件。（见图 2）这种组织方式可以方便基于文本信息的语料库检索和统计。

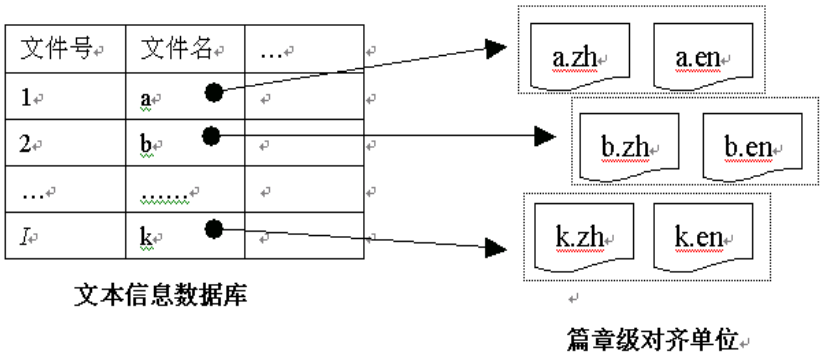

图 2 文本信息数据库索引

单语索引由汉语部分和英语部分组成，我们为它们分别建有一个倒排表（见图 3）。主

[6]篇头部分主要包含篇章级对齐单位的基本信息，用于反映语料库的基本情况，不参加段落与句子对齐处理。

[7]正文部分是语料库的主体，是段落与句子对齐处理及以后深层次加工的对象。

索引表①中记录了语料库中出现的所有单词，对每个单词记录它在整个双语语料库中的频率，并有一个指针指向一个文件表。文件表②记录某词曾经出现在哪些文件中以及该词在该文件中的频率，并有一个指针指向一个位置表。位置表③记录了某词在某文件内部出现在哪些位置以及每次出现时位于哪个句对中。该组织方式是获取单语语言模型的基础。

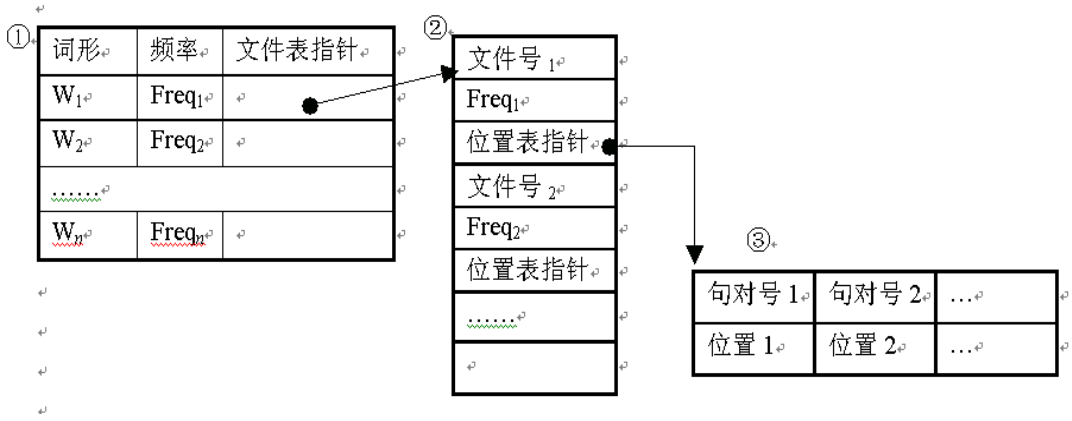

图 3 单语倒排索引

此外，还有一个为双语建立的索引，主要是针对句对（见图 4）。索引表中主要记录：文件号、句对号以及该句对中中文部分在相应中文文件中的位置、英文部分在相应英文文件中的位置。文件号以及文件内部的句对号可以唯一决定双语语料库中的一个句对。该组织方式是获取双语翻译模型的基础。

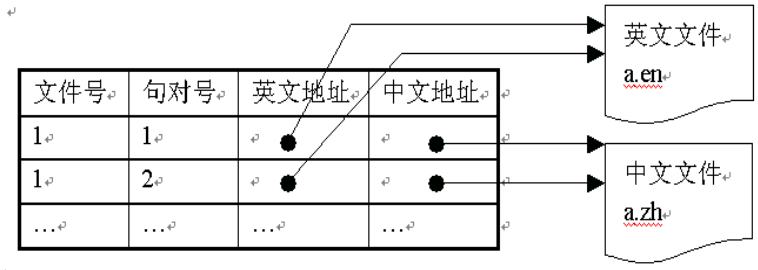

图 4 双语索引

语料库的组织通过上述索引结构来完成，同时，索引结构的建立也将保证相关的应用程序在语料库规模不断增大的情况下仍有合理的响应时间。

2.4 语料库辅助工具的开发

双语相关集列工具（Concordance）是一个相当重要的双语语料库辅助工具，是开发双语语料库其他应用技术的基础。在索引结构的基础上，双语相关集列工具能够通过灵活的检索机制使经过加工的双语平行语料成为切实可用的资源。

我们正在开发的双语相关集列工具既可以将语料作为独立资源呈现，也可以集成到机器辅助翻译系统中。它将具备下列功能：1）语料库管理；2）词频统计；3）搭配统计；4）选定语料的抽取；5）逐词索引功能；6）基于正则表达式的检索；7）检索结果的排序输出；8）检索项在文本中的分布图；9）检索结果的相关集列输出；10）检索结果的存储。

2.5 语料库现状的基本统计

表 2 和表 3 列出了我们的汉英平行语料库(不含 MTE 测试集)目前的基本统计数据。

说明：

- 现有语料的语体均为书面语，创作时期均为当代（源语言为中文）或 Present-day English（源语言为英文），因此，表 2 没有显示这两项统计数据。
- 按照我们的规范，文体不允许兼类，领域允许兼类。
- 按中英文对应情况，句子级对齐单位可以分为表 3 中列举的四类，其中“一对一”表示一个中文句子对应一个英文句子，“一对多”表示一个中文句子对应多个英文句子，以此类推。

表 2 语料的文体、领域分布

总句对数		45,261 句对
文体	文学	7,767 句对
	新闻	37,494 句对
领域	艺术	3,246 句对
	工商	8,963 句对
	政治	21,981 句对
	科技	9,516 句对
	体育	171 句对
	社会文化	12,137 句对

表 3 句子级对齐单位中英文对应情况

类型	句对数	比率
一对一	39,597	87%
一对多	4,103	9%
多对一	1,185	3%
多对多	376	1%

3 进一步的工作

目前我们正在进行部分语料的段落、句子对齐及其校对工作，今年年底，汉英平行语料库将达到 11 万句对的规模（其中中文部分约 500 万字、英文部分约 250 万单词）。进一步的工作目标是：

- 进一步完善双语平行语料库的构建流程，重点是现有的经验和语料库加工结果 1）完善各环节规范，2）改进辅助工具，协调流程各环节，提高整个流程的工作质量和工作效率；
- 扩大语料库规模并使之向平衡语料库的方向发展；
- 在完善语料浅层次加工的基础上，尝试语料的深层次加工[8]。

合理、有效地开展双语语料库基础资源建设，推动相关的加工技术和应用技术不断向前发展，是我们努力的方向。

[8] “面向新闻领域的汉英机器翻译系统”项目组目前正在进行一个短语库的建设工作，其中 13,000 余条短语来自这个汉英平行语料库中现有的句对。

参考文献:

[1] Sinclair, J., Corpus Concordance Collocation. Oxford: Oxford University Press, 1991.
[2] Biber, D., Conrad, S., Reppen, R. Corpus Linguistics. Beijing: Foreign Teaching and Research Press, Cambridge: Cambridge University Press, 2000.
[3] Kennedy, G. An Introduction to Corpus Linguistics. Beijing: Foreign Teaching and Research Press, 2000.
[4] Chang, Bao-bao, Zhang, Hua-rui, Kang, Shi-yong, Yu, Shi-wen.Bilingual Corpus Construction and Its Management for Chinese-English Machine Translation. Translation and Information Technology. Hong Kong: The Chinese Univeristy Press, 2002.
[5] 史晓东.英汉机器翻译：现状和未来.中国中文信息学会二十周年学术会议论文集.北京:清华大学出版社.2001.
[6] 黄昌宁,李涓子.语料库语言学.北京:商务印书馆,2002.

作者简介: 柏晓静，女，博士生，主要研究领域为机器翻译；常宝宝，男，博士，讲师,主要研究领域为机器翻译、计算语言学；詹卫东，男，博士，讲师,主要研究领域为现代汉语语法、机器翻译；吴拥华，男，硕士生，主要研究领域为机器翻译。

The Construction of A Large-scale Chinese-English Parallel Corpus

BAI Xiaojing[1] CHANG Baobao[1] ZHAN Weidong[1,2] WU Yonghua[1]

[1]*(Institute of Computational Linguistics, Peking University, Beijing 100871, China)*

[2]*(Department of Chinese Language and Literature, Peking University, Beijing 100871, China)*

E-mail: {baixj, chbb, zwd, wyongh}@pku.edu.cn

Abstract: Despite the increasing significance of bilingual parallel corpora in recent MT researches, related work in Mainland China has not laid much stress on the development of this language resource in a systematic way. This paper describes the construction of a large-scale Chinese-English bilingual parallel corpus, including the overall planning and the model for constructing such a corpus, together with the details of our work. We expect our further efforts will promote the construction of bilingual corpora as a fundamental resource in MT, hence the progress of related theoretical studies and application techniques.

Key words: Machine Translation; Bilingual Parallel Corpora; Corpus Construction