面向汉英机器翻译的双语语料库的建设及其管理*

常宝宝　詹卫东†　张华瑞
北京大学计算语言学研究所，北京，100871
† 北京大学中文系，北京，100871

一、引言

近年来，在语言信息处理的研究和开发中，单语和多语语料库（以双语语料库居多）的作用日益突显出来。特别是在机器翻译研究中，人们提出了多种基于双语语料库的新方法，例如采用所谓的基于实例（Example-Based）的或基于存储（Translation Memory）的机器翻译方法，可以直接使用经过对齐的双语语料改善机器译文的质量。此外，也可以通过统计模型从双语语料库中获取双语词典和翻译模式，从而改进传统的机器翻译方法。除中文信息方面的应用之外，双语语料库的建设对于双语词典编纂、跨语言的对比研究也具有重要价值。

目前关于双语或多语语料库的研究大致可分为三类：一是研究双语语料的对齐技术（Alignment），国内外学者就此提出多种策略和方法，现在已经出现了许多对齐双语或多语语料的程序或工具[Gale 1993]；二是研究双语语料的各种应用，如在基于统计的机器翻译技术[Brown 1990]、基于实例的机器翻译技术[Nagao 1984]，双语词典编纂[Klavans and Tzoukermann 1990]技术中，双语语料库都发挥着十分重要的作用；三是双语语料库的设计、采集、编码和管理问题。目前比较著名的语料库编码方案有 TEI 文本编码标准以及 CES 标准，两者均基于 SGML 标记语言。就前两类研究来说，中国国内目前做了较多的跟踪研究工作，而对于第三类研究，即双语语料库尤其是涉及汉语的双语语料库的建设、编码和管理研究，探索工作似乎做的相对较少。与此相关，目前国内外都还没有见到有关系统的、经过深度加工的、以汉语为源语言的双语语料库的报道。

北京大学计算语言学研究所、清华大学智能技术国家重点实验室和中国科学院计算所三家单位联合承担了国家 973 课题——“面向新闻领域的汉英机器翻译系统”的研制开发任务。系统决定采用基于多种方法的多引擎体系结构（将基于规则的方法与基于语料库的方法相结合）。为此，需要建立一个具有一定规模的经过对齐处理的汉英双语语料库。本文将简要介绍这样一个服务于汉英机器翻译的双语语料库的设计、收集、编码和加工的情况。

二、语料库的设计和语料收集

语料库建设是一项工作量极大的工作，因为一个有实际应用价值的语料库决不是任意文本的任意集合，其文本类型、大小以及语料的构成都必须根据应用需求，经过仔细的设计，只有这样才能保证所投入的工作是值得的。我们认为，设计一个双语语料库，首先应该考虑语料库的应用目标。语料的收集、语料的构成以及对语料的加工应该紧紧围绕语料库的应用目标进行。作为服务于一个面向新闻领域的汉英机器翻译系统的双语语料库而言，在语料的收集、加工等方面，应该跟服务于其他目的（比如语言研究）的语料库有所区别。服务于汉英机器翻译的语料库是一个专用的语料库，而不是一个通用的语料库。在这个前提下，我们不强调语料库中的语料对汉语文本的覆盖性。在对语料的内容、语料库中的文本类型、文本的创作时间、语料库的结构进行选择时，应以是否有助于面向新闻领域的汉英机器翻译为准则进行。最为理想的情况是，语料库中的语料能够形成全部新闻语料的一个统计样本。然而

* 本文工作得到了国家 973 项目的资助(项目编号: G1998030507-4)

构造一个这样的语料库并非易事，这需要有足够的机器可读的新闻语料作为取样基础。结合上述理论思考以及现实条件下的电子文本的实际情况，我们确定了下面的语料收集原则：

1) 收入语料库的文本最好是报道类型，不过也可以包含一些具有良好英语译文的同新闻报道在内容和结构上具有相似性的语言材料。因此除了新闻报道类型，我们也收集了一些新闻发布会文告、政府白皮书和一些杂文以及它们的英语译文。
2) 双语材料最好以汉语作为源语言，因为语料库的服务对象是汉英机器翻译系统，但也酌情收集了一些具有非常流畅自然的中文翻译的英语材料。
3) 文本应以全文形式收入语料库，这将有益于篇章知识的获取和学习，一个实用的机器翻译系统最终必须面对全文的翻译。
4) 就创作时间而言，所有收入的文本应当是最近几年的文本，这样才能够反映当下语言的实际使用情况。

在上述原则的指导下，我们收集了大约 100 万字的汉语全文语料及其英语译文。这些语料基本来源于国际互联网。大致可分为四类：新闻报道、新闻发布会文稿、白皮书以及杂文。其构成比例如图 1 所示：

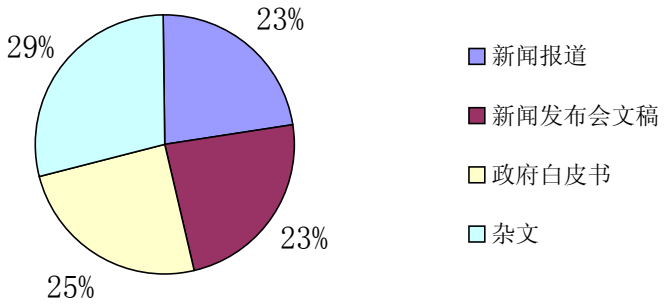

图 1. 双语语料库的语料组成

三、语料库的编码

管理这些平行语料的理想方式是设计一个专用管理系统。语料库中所有语料均需按照同样的方式编码或标记。这可以使得语料库能够独立于软件平台和具体的应用程序，具有教强的数据可交换性。目前国际上有两个著名的语料库标记标准建议方案，一个是正处在开发之中的语料库编码标准(CES)，另外一个是文本编码标准 TEI。TEI 已为一些著名语料库所采用，例如英国国家语料库(BNC)。这两个标准都是基于 SGML 标记语言而制定的。考虑到我们的语料的主要来源是国际互联网，大部分语料是以超文本标记语言(HTML)形式存在的。因此，如何对语料库进行编码存在三种选择：(1)采用国际上业已制定的标准方案；(2)直接采用互联网上广泛使用的超文本标记语言（HTML），这样似乎可以有效减少工作量；(3)制定一个新的标记方案。

方案（2）尽管可以减少工作量，但并不可行。首先，超文本标记语言是目前世界上最为流行的网页标记语言，不同的支持公司都对其作了不同的扩充，语法要求并不严格，常常可以用不同的标记形式来标记不同的内容，因而不适合用来标记我们的语料库。其次，超文本标记语言不做内容和显示的区分，其中既包含用于内容的标记元素，也包含用于显示的标

记元素。因而很多情况下，网页作者因为显示效果而放弃使用内容标记元素。例如在我们收集的语料中，文本标题很少使用<Hn>标记，而更多使用 、 <font>等标记。

再看方案（1），尽管 CES 和 TEI 是专为标记语料库而设计的国际标准方案，但二者均面向通用目的，即使选择一个由较少的必要元素组成的子集，也会因过于复杂而难以掌握。并且其中许多元素对于我们的应用意义不大，同时对一些我们需要详细标记的信息，如新闻报道的特有结构，却又没有合适的标记可以使用（即有“大炮打蚊子”之嫌）。另外，作为二者基础的 SGML 标记语言，也一直因为过于复杂而难以得到信息处理界（包括 IT 产业界）的广泛使用，开发一个全面的 SGML 分析器也不是一个短时期内可以完成的工作。

经过上述分析，为了获得一个简单的但能满足我们需要的编码方案，我们选择了方案（3），即参照 CES 开发一个新的标记系统。这个标记体系不力求覆盖所有文档类型，但要求对于我们所关心的文档类型有足够的支持，对其他文档类型仅仅要求有一般性支持。这个标记体系基于目前正日益流行的标记语言 XML，从而保证我们的标记系统有广泛的软件支持。

按照我们的标记系统，整个语料库由一组相互链接的文档组成，整个双语语料库的逻辑结构如图 2 所示。

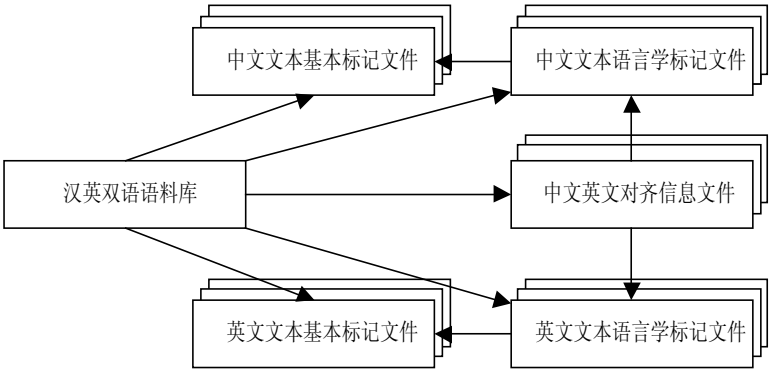

图 2.　双语语料库的逻辑结构

各种文件的含义如下：

(1) 中文基本标记文件和英文基本标记文件：
在这个文件中，主要标记中英文文本的结构信息，例如新闻报道的标题、子标题、新闻导言、讯头以及文档的一般结构信息。此外，在这个文件中还要标记命名实体，例如人名、地名以及机构名等。

(2) 中文文本语言学标记文件和英文文本语言学标记文件;
主要标记中英文文本中有关词语的词性信息、短语的结构信息、分句的组成关系信息、句子结构成分信息等。

(3) 中文英文对齐信息文件
标记中文文本和英语译文文本之间在各个级别上的对齐关系，包括段落级对齐、句子级对齐、词一级的对齐、短语结构级的对齐信息，等等。

按照 XML 标记语言的规定，总共为上述文档定义了四个文档类型定义(DTD)。分别用于描述（1）整个双语语料库；（2）中文基本标记文件和英文基本标记文件；（3）中文文本语言学标记文件和英文文本语言学标记文件；（4）中文英文对齐信息文件。

标记系统允许以一致和循序渐进的方式对语料进行由浅层到深层的信息标注。

四、语料的标注和对齐

语料库标注工作取决于语料库将以何种方式使用。我们希望部分语料库资源能够直接用于改善机器译文的质量，也希望能够从语料库中学习到汉语到英语的翻译知识，例如汉英双语词典、翻译模式等。为此，目前我们正在进行或计划对语料库进行下列标注工作：

1) 中文分词和词性标注；
2) 英文词性标注；
3) 中文和英文的专名标注（中文机构名识别已作了小规模的实验）；
4) 中文、英文文本句子一级的对齐；
5) 中文专名和英文专名的对齐；
6) 中文词语的详细语法特征标注。这项标注将根据《现代汉语语法信息词典规格说明书》[俞 1996]进行。在现代汉语语法信息词典中，每类词都可能拥有多达几十个的语法特征信息，但在具体的上下文环境中，并非每个语法特征都有所表现，我们希望这项标注将有助于学习词汇翻译知识。目前对这项标注已经进行了一些小规模的实验。

上述标注工作基本按照下面的过程进行：1)首先利用软件工具进行自动标注；2) 人工校对标注结果。目前已有约 10 万字的中文语料进行了分词和词性标注，对应的译文进行了词性标注，这部分语料的标注信息均已经过人工校对。另外，这部分语料句子对齐的校对工作也正在进行之中。

五、进一步的研究工作

在对 10 万字语料的分词和词性标注、对齐的校对工作完成后，我们正在扩大处理语料的规模，对其余 90 万字语料进行词性标注和对齐加工。

在已有的标注信息基础上，我们还将考虑对语料库进行更深层次的标注工作，包括标注句子的句法结构和篇章的结构信息等。目前这方面的工作正在积极探索之中。

在应用方面，目前我们正在开发一个简单的基于存储的汉英机器翻译引擎，并在两个加工级别上使用具有不同标记深度的双语语料。第一个级别是将没有进行切词、词性标注的句子对齐结果直接作为资源用于机器翻译，翻译引擎根据用户输入的待译句子在对齐的句对中进行检索操作，如果命中，则直接输出译文；第二个级别是利用对齐的句对以及经过分词和词性标注的语料（其中一些特殊的词语如数字、专有名词等也经过对齐处理），翻译引擎将这样的句对视为一种框架结构，当用户输入待译句子后，翻译引擎利用输入句子和这些框架结构进行匹配，如果匹配成功，则对其中的可以替换的词汇进行替换，并修改相应英语译文得到输入句子的译文，从而提高英语译文的自然度。

参考文献

[Brown 1990] Brown, P., et al, A statistical approach to machine translation, Computational linguistics, V16, No.2, 1990

[CES] Corpus Encoding Standard, http://www.cs.vassar.edu/CES/

[Gale 1993] Gale W., et al, A program for aligning sentence in bilingual corpora, Computational linguistics, V19, No.1, 1993

[Klavans 1990] Klavans, J., and Tzoukermann, E., The BICORD system, In Proceedings, 15th International Conference on Computational Linguistics.
[刘 1995] 刘昕，周明，黄昌宁，基于长度算法的中英双语文本对齐的试验，陈力为等主编《计算语言学进展与应用》，清华大学出版社，1995
[Nagao 1984] Nagao, M., A Framework of a Mechanical Translation between Japanese and English by Analogy Principle, In: A.Elithorn et al eds. Artificial and Human Intelligence, NATO Publication
[TEI] TEI Guidelines for Electronic Text Encoding and Interchange, http://etext.virginia.edu
[俞 1996] 俞士汶 ,朱学锋等,《现代汉语语法信息词典》规格说明书,《中文信息学报》,1996 年第 2 期

Bilingual Corpus Construction and its Management for Chinese-English Machine Translation

Chang-Baobao Zhan-Weidong† Zhang-Huarui
The Institute of Computational Linguistics, Peking University, 100871
†The Department of Chinese Language and Literature, Peking University, 100871

Abstract: In recent years, monolingual or multilingual (primarily bilingual) corpora are viewed as key resources in language information processing and language engineering projects. To support an ongoing Chinese-English machine translation project, a Chinese English bilingual corpus is being set up. This paper gives a brief discussion on construction of the corpus.

Keywords: Bilingual Corpus, Machine Translation, Corpus Markup, Corpus Annotation

面向汉英机器翻译的双语语料库的建设及其管理

常宝宝 詹卫东† 张华瑞
北京大学计算语言学研究所，北京，100871
†北京大学中文系，北京，100871

摘要: 近年来，在语言信息处理的研究和开发中，单语和多语语料库（主要是双语语料库）的作用日益突显出来。为了支持一项正在进行的汉英机器翻译系统的开发，我们建立了一个汉英双语语料库。本文简要介绍了该语料库的建设和管理情况。

关键词: 双语语料库, 机器翻译, 语料库标记, 语料库标注