双语语料库的检索和管理

王长胜　刘　群
（中国科学院计算技术研究所，北京 100080）
E-mail：wcs@software.ict.ac.cn

摘　要　该文介绍了在笔者的辅助翻译系统中已实现的双语语料库的检索和管理。实验结果表明该双语库检索和管理在实时交互、空间开销等方面是令人满意的。

关键词　机器辅助翻译　双语语料库　自然语言处理

文章编号 1002-8331-（2002）07-0113-02　文献标识码 A　中图分类号 TP391.2

The Retrieval and Management of Bilingual Corpora

Wang Changsheng　Liu Qun
（Institute of Computing Technology，Chinese Academy of Sciences，Beijing 100080）

Abstract： In this paper，the retrieval and management of bilingual corpora used in the Machine Aided Translation system （MAT）are introduced.The experiment result indicates they are satisfactory in real-time interaction and the cost of memory.

Keywords：Machine Aided Translation，Bilingual corpora，Natural Language Processing

1　引言

机器翻译发展到今天，已取得了巨大进步。但仍有许多问题是现阶段无法完全解决的，为此人们在人工辅助翻译、基于实例的翻译等系统中充分应用了计算机这一存储能力强大、检索快速的工具，利用九十年代兴起的语料库方法，结合自然语言处理的最新技术，有效利用语料库、记忆库、术语库等，让计算机自动提供已有的翻译记录和术语，同时不断学习人类的翻译方法。

网络的快速发展，提供了大量而丰富的双语对照电子文献，这就为机器辅助翻译提供了坚实的语料基础。

机器辅助翻译领域的主要研究内容有：双语句子的自动对齐；双语术语自动获取；翻译记忆库和术语库的查询、智能扩充；人机互助翻译思想的具体实现等。笔者在拥有良好资源（规模较大的双语库，汉英双语句子自动对齐程序，丰富的字典，术语自动获取程序等）的基础上搭建了一个实用的机器辅助翻译系统。

以下介绍了双语库（文章采用的是句子级而非短语或者词汇级对齐的中英对照例句集合，形如“这部书百读不厌。You never get tired of reading this book.”）检索和管理的具体实现方法。

2　思路和内容

设计此系统时考虑到翻译的实时交互性（即用户每翻译一句要求给出相同或相似的例句）和例句库的可扩充性（包括修改、删除），该文采用了建立倒排文件和哈希文件相结合的方法来达到双语库的快速检索及扩充目的。整个过程包括：

（1）双语库的加工，主要完成对存储在 Mysql 数据库中的双语语料库的提取（可得到文本格式的双语例句库、中文例句库等）。

（2）通过双语例句库建立倒排文件和哈希文件。

（3）提供两种方式（按检索单词检索和按相似度检索）的例句检索算法。

另外为了实现对例句库的扩充，在例句库管理界面中加入了双语语料库的增加合并及其它管理模块。以上的所有过程都具体体现在可交互的系统界面中。

2.1　系统描述及其算法

为了达到笔者设计此系统的两大主要目标：实时交互性和系统例句的扩展性。笔者在实践中不断地完善了初始设计时的一些想法，以使该系统设计达到最优。

首先，文章通过对例句库文件建立倒排文件和哈希文件而达到对例句库的快速检索这一目标。其中例句库文件的结构简单地安排如下：句子号　中文例句+分隔符+对应的英文例句。例如“6 春意盎然。*|*Spring is in the air.”。如果应用在其它的系统如基于实例的机器翻译系统中，可略调，加一些其它字段。而倒排文件和哈希文件结构分别如下：

单词　单词在例句库文件中的首位置　例句中单词相对例句首位置的偏移量　下一个相同单词在倒排文件中的位置

例如：春意盎然　308　2　0（0 表示单词“春意盎然”在例句库中仅出现一次）

单词　单词在倒排文件的末位置　出现此单词的所有例句的个数　单词在哈希文件中的位置

例如：春意盎然　122　1　318

以下是一个完整的片段：

2 爱国/一/家/，/爱国/不/分/先后/。*|*All patriots belong to one big family，whether they rally to the common cause early or late.（此乃例句库文件中的第二行，即第二个对

基金项目：国家 973 重点基础研究项目（编号：G1998030507-4）
作者简介：王长胜，男，1974 年生，现为中国科技大学研究生院研究生，研究方向：自然语言处理、数据挖掘。

照例句)。

该句子对应的索引结构为：

倒排文件				哈希文件			
爱国	50	2	0	爱国	23165836	5	70
家	50	8	0	家	25945425	2534	106
爱国	50	12	21				

说明：

(1)哈希文件中的单词是互不相同的，要与倒排文件中的单词区分开来。

(2)由上述文件结构的安排可知，该检索是在文件中反向搜索的。

(3)在哈希文件结构中加入的字段"出现此单词的所有例句的个数"减少了一半的检索时间（相比最初设计时没有加入此字段），而它的加入与否并不影响整个倒排文件和哈希文件生成的总体时间（经实验比较得知，二者相差不到 10 秒），而且它的加入也很方便。

(4)整个思路是：顺序读入例句库文件的每一行，对中文部分分词，然后过滤掉 894 个停用词，对剩下的单词按一定方式分别写入倒排文件和哈希文件且记录下各自的有用信息。

另外为了解决建立哈希文件时相同单词的比较问题（即解决上述说明 1 的问题）系统要求快速，最初没有采用此结构时耗时 1 个半小时），给出如下结构：

单词	出现相同单词的次数	记录单词插入哈希表的顺序	解决哈希碰撞的位置

以上结构中的很多字段是出于快速检索算法而设计的，其它的字段或结构是出于快速建立倒排文件和哈希文件而设计的。

建索引算法的输入是双语例句库文件，输出是倒排文件和哈希文件。其中用到的分词程序是北京大学计算语言学研究所提供的，既能分词又能标注词性，效果不错。另外此算法中用到了停用词词典。

而在例句库检索算法中除了用到以上的结构外，还涉及到其它的一些有益结构，这里不详述。

索引工作做完后，以何种方式检索待翻译中文句子呢？基于表层句法信息的语句相似度，文章提供了两种可选方式给用户检索，一种是按检索单词检索，公式为：$Sim=i/n$ 其中 Sim 为相似度，i 为检索到的例句和待翻译句子中共同词的个数，n 为待翻译句子中的所有单词总数。以上统计包括停用词和标点符号)；另一种是按相似度检索，公式为 $Sim=2*i/(n+n)$ 其中 n 与上公式中的 n 相同，m 为检索到的例句中的所有单词总数)。这两种相似度计算检索出来的结果有所不同，例如：

例 1 输入检索句：美丽/的/姑娘

当检索比率为 100%时，按检索词检索结果为：她/是/一个/美丽/的/姑娘/，/有/着/模特儿/的/那/种/优雅/姿势（译文：she is a beautiful girl with the grace and poise of a natural model)其相似度为 3/3=100%；而按相似度检索，该结果的相似度为 2*3/(3+15)=33.33%<100%，不合要求。

如果将检索比率改为 42%（此百分比在界面上可调)，按检索词检索的部分结果如下（还有很多)：

(1)她/是/一个/美丽/的/姑娘/，/有/着/模特儿/的/那/种/优雅/姿势 she is a beautiful girl with the grace and poise of a natural model.100%

(2)山里/的/瀑布/构成/了/一处处/美丽/的/景色。the waterfalls in the mountain provide gorgeous view.66.67%

(3)这/美貌/的/姑娘/一到/，/ 全屋/的/男人/都/让/她/吸引住/了 when this beautiful girl arrives，all the men in the room gravitate towards her.66.67%

而按相似度检索则只有如下三个结果（其它结果相似度都小于 42%)：

(1)绿色/是/美丽/的/颜色。Green is a beautiful colour. 44.44%

(2)瞧/这些/美丽/的/废墟。look at all the beautiful old ruins.44.44%

(3)世上/的/姑娘/多得/是。she's not the only fish in the sea.44.44%

由此可见按相似度检索得到的结果总体长度与输入句子长度很接近；而若按检索词检索，其检索结果很乱，几乎没有什么规律，只要例句库中出现了输入句子中的单词，在一定的比率控制下它们就能被检索出来。

文章在设计此检索算法时兼顾考虑了检索的全面性和检索的实时性。根据检索比率 b（在界面上用户可调，范围定为 20%~100%)计算出待检索的句子中前 $n-n*b+1$ 个（其中 n 为待检索的句子中所有可检索的单词，即非停用词和标点符号)拥有最小例句集个数的单词（此处用到哈希文件结构中的"出现此单词的所有例句的个数"字段)，然后对这些单词检索得到的例句集进行过滤（即大于等于 b)，然后按照相似度大小排序，输出结果。采用这一策略比较检索所有单词，其效率依据检索比率 b 的大小和所有检索单词所拥有的例句集的个数具有如下的关系：b 越大，检索越快；而在 b 一定时，二者间的检索时间差就是后 $n*b+1$ 个拥有最多例句个数的例句集合读取磁盘的时间。后面的实验也验证了这一点。之所以说兼顾考虑到了检索的全面性，是因为在一定的检索比率 b 下，如果例句库中存在与待检索句子相似的例句（$\leq b$)，是不会因为待检索句子中后 $n*b+1$ 个单词不检索而与之相似的例句中就一定没有这些单词。当然不是绝对的。这一点尤其体现在按检索词检索得到的结果中（如上例)。

双语库的添加、合并算法：主要是重新改写倒排文件和哈希文件，然后如何及时刷新的问题（例如用户在界面上随时添加一句中英对照例句，或合并用户自己的例句库，这里采用的不是建立临时文件或用户库文件的方法，而是将其直接写入系统例句库中)。

2.2 算法的性能分析和实验

2.2.1 算法性能分析

对倒排、哈希文件的建立：其效率取决于读写文件和对所读文件的中文部分分词及解析出每一个单词所用的时间开销。其空间的开销在于哈希文件的载入。

对例句库的检索：其时间的开销仅取决于检索到的例句对的个数，个数越大读取硬盘上的例句所需的时间越多。

对双语库的合并：其时间的开销取决于用户要合并的例句库的大小，也即磁盘 I/O 开销。

2.2.2 实验结果

在 CPU 为 P200，64M 的普通内存，非 SCSI 硬盘，平台为 windows 2000server，visual c++6.0 上实验结果如下（所有的操作都在如图 1 所示界面上进行)：

笔者对 21.5M 的双语对照的例句库（近 20 万对)进行倒

(下转 196 页)

```
}else//<left>的属性是 Simple_ref 类型
{k=find_database (alias); add_to_where_pred ();}}
}
```

为了验证上述算法的正确性和可行性，笔者在 Inprise JBuilder 2.0 上利用基于 CORBA 的工具——VisiBroker for Java 建立了一个样机系统。在 Web Browser 中透明访问了 Windows NT 平台上存放在 SQL Server 中的工程数据库项目信息和 Windows 98 平台上存放在 Access 中的工程数据库产品信息。因篇幅有限，程序运行结果见参考文献[6]。

5 小结

该文提出的基于多数据库的工程数据库的模式结构既有坚实的理论基础，又能满足实际应用的需要，具有良好的实用价值。使用 ODMG 推出的 ODL 进行集成模式的定义，使查询处理系统具有很好的适用性，可以扩展集成各种面向对象数据库、无模式的半结构化或无结构数据源。同时，查询处理系统预期还可以集成 STEP/PDES (EXPRESS)、ANSI X3H7、CFI 等具有动态变化能力的数据模型，进一步满足工程领域的要求。该文针对基于多数据库模式结构的工程数据库系统设计和实现的查询分解算法，能够将全局数据库查询语句分解成针对各局部数据库输出模式的一组子查询，每个子查询仅涉及到一个局部数据库的输出模式，由相应的局部数据库完成这个局部子查询处理。 (收稿日期：2001 年 3 月)

参考文献

1.Litwin W，Abdellatif.Multidatabase Interoperability[J].IEEE Computer，1986.2；19 (2)

2.Sheth Amit P，LARSON James.A Federated database System for managining distributed，heterogeneous，and autonomous databases[J].ACM Surveys，1989.9；22 (3)

3.Busse R，Frankhanser P，Neuhold E J.Federated schema in ODMG[C].East/West Database Workshop

4.Akula Ramesh，Dia L Ali.Query Transformation In Heterogeneous Distributed Database Systems[C].19th International Conference on Computers and Industrial Engineering，1996；31 (1/2)

5.S Nural，P Koksal，F Oacan et al.Query Decomposition and Processing in Multidatabase Systems[C].Proceeding Of the International Conference on Engineering Systems Design&Analysis，Montpellier，1996.7

6.黄玲.基于 CORBA 的工程数据库查询处理系统的研究[D].硕士学位论文.南宁：广西大学，2000.6

(上接 114 页)

排、哈希文件的建立：所花销的时间平均为 72 秒左右 (多次实验的结果都相差无几：70.01s、72.33s、72.35s、74.14s 等，不包括分词时间 277.76 秒)。建成后的倒排文件大小为 24.7M，哈希文件为 1.29M (定长写入的，有部分空隔空间)。结果分析表明：系统所用内存空间不会随着例句库的增加而有所增加，这是因为内存空间的大小主要取决于例句库分词后的不同单词的个数 (这里约 4 万单词，不同的分词程序其差别很大)，笔者在随后的例句库合并中发现，要合并的例句库中新增的不同单词的个数很少，基本上原例句库中都出现过。

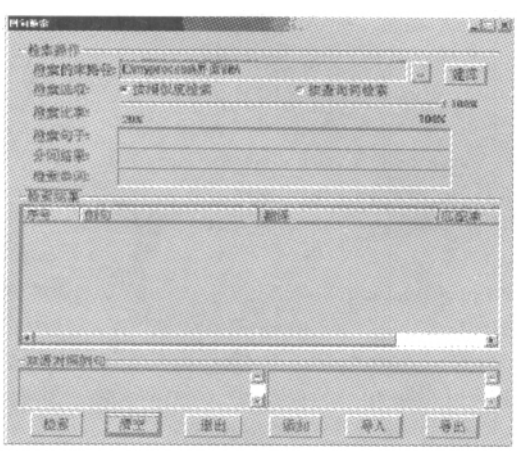

图 1

对词和句子的检索，系统有所区别：

对句子检索其时间的开销不是待检索句子的所有单词 (k) 所能检索到的例句所耗时间的总和，而是根据用户在检索界面上选择的检索比率 (b)，计算出 $n-n*b+1$ 个拥有最小例句总数的单词，然后按条件检索，精简结果及排序输出 (匹配率>=b 的例句集合按匹配率大小显示给用户)的时间之和。以下给出一个实例的检索数据：

例 2 输入检索句：九运赛场美女如云 检索比率 b

分词及其各自所出现在例句库中的例句个数为 九|99 运|202 赛场|0 美女|7 如 云|192)

说明：“赛场”在笔者的例句库中没有相应的例句，‘如’是停用词，可检索单词个数为 4

b	检索单词个数 ($k=n*b+1$)	在例句库中检索的单词	检索出来的例句个数	检索时间
20%	4-4*20%+1=5>4 处理为 4	美女 九 云 运	14	1.242 秒
25%	4-4*25%+1=4	美女 九 云 运	1	1.162 秒
50%	4-4*50%+1=3	美女 九 云	0	0.691 秒
75%	4-4*75%+1=2	美女 九	0	0.21 秒
100%	4-4*100%+1=1	美女	0	0.04 秒

对双语库的合并算法：将一 351K 的双语库文件 (4 千多对例句)与系统原始例句库合并用时 13.21 秒；将一 4.57M 的双语库文件与系统的合并用时 197.35 秒。 (以上两组数据包括分词时间)

3 结束语

例句库的检索和管理对机器辅助翻译系统、基于实例的机器翻译系统 (EBMT)有重要意义。该文的检索和管理工具是笔者实现的机器辅助翻译系统的核心之一，可方便地独立或集成于系统。其检索的时间完全可以满足用户的交互响应。此外它还应用于笔者正在搭建的 EBMT 系统中。

(收稿日期：2002 年 1 月)

参考文献

1.许卓群，张乃孝，杨冬青等.数据结构[M].高等教育出版社，1993

2.吴立德，罗航哉，薛向阳.基于多重倒排文件的快速相似性检索

3.Eric W，Brown James P，Callan W.Bruce Croft：Fast Incremental Indexing for Full-Text Information Retrieval[C].In：Proc.of the 20th Inter.Conf.on Very Large Databases (VLDB)，Santiago，1994.9：192-202

4.Robert Muth，Udi Manber.Approximate multiple string search[C].In：Proc.CPM'96，LNCS 1075，1996：75-86