汉英双语短语信息数据库的构建*

吴云芳[1,2]　常宝宝[2]　詹卫东[1,2]

1　北京大学中文系　2 北京大学计算语言学研究所

wuyf, chbb;zwd@pku.edu.cn

摘要：本文扼要介绍了一个汉英双语短语信息数据库的构建情况：汉语短语的描述信息；英语短语的描述信息；描述中的疑难问题；短语库的实施和应用。短语库是综合型语言知识库的有机组成部分，它的建设将为短语结构研究、句法分析和机器翻译提供强大的语言知识支撑。

关键词：汉英双语短语　短语结构　机器翻译

一　引言

短语结构分析是汉语信息处理从词处理过渡到句处理的第一道门坎儿。汉英机器翻译中，短语结构分析的困难主要表现在：（1）汉语短语结构歧义的识别；（2）汉语短语结构到英语短语结构的转换，而每一个问题的解决都是困难重重。汉英双语短语信息数据库（下文简称“短语库”）的构建就是直接服务于汉英机器翻译，为上两个问题的解决构建一个语言知识库。

浅层句法分析(shallow parsing)和与之相关的语块(chunk)标注语料库越来越受到重视和青睐(见周强 2001)。短语库中的短语其实就是静态的、离线的语块。从语言知识库建设的角度看，词库资源建设已取得了喜人的成果：如 1994 年林杏光等主编的《现代汉语动词大词典》，1998 年历时多年的俞士汶等著作的《现代汉语语法信息词典》。短语库在两个方面自然延伸了原有的语言知识库:由词级（word level）扩展到了短语级(phrase level)；由单一的语种(汉语)扩展成了双语对照(bilingual parallelism)。

短语（词组）本位语法体系是短语库建设的语法理论支撑。汉语词类跟句法成分之间不存在简单的一一对应关系：同一个句法成分可以由不同词类的词来充当；同一个词类的词可以充当不同的句法成分而没有任何形式上的标记。因此，名词短语的中心词可以是一个动词或形容词，如名词短语“爱克斯光透视”中心词就是动词“透视”。由于汉语句子的构造原则跟短语的构造原则基本一致，比之英语，短语在汉语语法体系中占有更为独特和重要的地位，这也是短语库建设的初始驱动力。

本文扼要介绍了短语库的构建情况：第 2、3 节分别介绍汉、英短语信息描述的内容；第 4 节讲述短语描述中的疑难问题；第 5 节简要说明该知识库的实施和应用；最后是结束语。

二　汉语短语信息描述的内容

汉英双语短语信息数据库包含两大块：汉语短语信息和英语短语信息。

对词进行描述，重在描述它的组合搭配情况，以预测它生成短语的能力。对短语进行描述，一方面要描述它在更高层次（句层次）上向外扩展的能力，表现为短语功能类型；另一方面还要描述它的内部构成情况，以发现由词到短语的组合规律，而且，短语的内部构成一定程度上也决定着它向外扩展的能力（见詹卫东 2000）。

* 本文研究工作得到国家 973 项目（G1998030507-4）和 863 项目（2001AA114040）的支持。

2.1 汉语短语的切分和词性标注信息

为了和已有的语言资源相兼容，短语库遵循北大计算语言所的词语切分和词性标注规范（以下简称“词规范”，参见俞士汶 1999），使用了其中 26 个词性标记。不同的是，“词规范”标记集中的 i （成语）、j（缩略语）、l（习用语）没有反映词语的功能信息，我们在进行短语库加工时将其改成了相应的功能标记，如：肌肤之亲 i→n 。

2.2 汉语短语的短语类（phrasal category）

短语类是从功能的角度来看短语向外组合扩展的能力。短语库设置了以下 9 种短语类型：

表 1

序号	名称	标记	例子
(1)	动词性短语	vp	荣获冠军　挂外科
(2)	名词性短语	np	人民生活水平　耐火玻璃
(3)	数量短语	mp	两百本　三十岁
(4)	处所短语	sp	边疆少数民族地区　中国内地
(5)	时间短语	tp	原始社会末期　二十一世纪
(6)	形容词性短语	ap	绝对可靠　有益于健康
(7)	副词性短语	dp	笨拙地　不情愿地
(8)	小句	dj	贫富不均　待人诚恳
(9)	介词短语	pp	按照这种逻辑　以革新的名义

小句指直接组成成分构成主谓关系的短语。介词短语在句子分析中有着重要的作用（GPSG 理论就把名词短语、动词短语、形容词短语和介词短语作为英语的四种基本短语类型），单独摆出来作为一类有利于描写和研究。

2.3 汉语短语的短语结构（phrase structure）

短语结构指短语内部直接组成成分（IC）之间的结构类型，也就是最上层（top level）的组成关系。短语库设置了以下 9 种短语结构类型：

表 2

序号	名称	标记	例子
(1)	定中结构	DZ	芭蕾舞演员　讨人喜欢的面孔
(2)	述宾结构	SB	引起警觉　犒劳三军
(3)	述补结构	SBU	洗干净　冷得直发抖
(4)	联合结构	LH	干净利落　研究讨论
(5)	主谓结构	ZW	两者缺一不可　疾病猖獗
(6)	连谓结构	LW	打电话通知　请他当主席
(7)	状中结构	ZZ	层层把关　从实质上看
(8)	的字结构	DE	卖布的　无情的
(9)	介宾结构	JB	按照工程进度　在意识形态领域

短语类和短语结构是从两个不同的角度对汉语短语进行的分类，两者之间存在着某种对应关系：

表 3

短语类	短语结构

vp	SB SBU LW ZZ LH
np	DZ DE LH
dj	ZW LH
ap	ZZ SB SBU DE LH
dp	ZZ DE LH
mp	DZ LH
pp	JB LH
sp	DZ LH
tp	DZ LH

可以看出，vp 和 ap 最复杂，分别对应 5 种不同的结构类型。联合结构可以对应所有的短语类。

2.4 汉语短语的中心词（headword）

中心词是短语的句法语义中心。凭借中心词，下层（词汇层）的句法语义信息传递到高层（短语层），参与进一步的句法运算。无论在 X' 理论，还是在 GPSG、HPSG 理论中，中心词都是非常重要的语言单位。

定中结构、述宾结构、述补结构、状中结构都是比较典型的向心结构，中心词容易确定：定中结构和状中结构的中心词是被修饰、被限定的那个成分；述宾结构和述补结构的中心词是述语本身。联合结构、主谓结构、连谓结构、的字结构和介宾结构的中心词存在争议。操作中，短语库遵循以下约定（"!"标记中心词）：

（1）并列结构是多中心的。如：!山麓/n !丘陵/n

（2）主谓结构以谓语为中心词。 如：科技/n 人员/n !外流/v

（3）连谓结构，中心词标记和对译的英文保持一致。当英文的中心词也不易确定时，以末一个谓语作为中心词。如：看/v 成色/n !定/v 价钱/n fix the prices according to the quality

（4）的字结构，之前的那个成分的中心词为短语的中心词。如：人数/n 较/d !少/a 的/u

（5）介宾结构以介词为中心词。如：!在/p 适当/a 的/u 时机/n

2.5 汉语短语的其他信息

汉语短语的其他信息用来描述短语是否属于一些特殊的类型。

是否命名实体：NR（人名），NT（组织机构名），NS（地名），NM（商标字号），NZ（其它专有名称）。

是否固定短语：i（成语），J（简称略语），L（习用语）。

三 英语短语信息描述的内容

3.1 英语短语的词性标注信息

短语库采用了宾州树库的英语词性标注集，使用 37 个词性标记。（详见宾州树库词性标柱集及手册）。

基于该标记集的软件资源比较丰富，可以帮助自动生成英语短语的词类标记。短语库采用了一个基于转换的错误驱动的词性标注软件产生英语短语的初始词性标记。（见 Brill, E 1995）。

3.2 英语短语的短语类

短语库设置了以下 6 种英语短语类型：

表 4

序号	名称	标记	例子

(1)	名词性短语	NP	the urban subsistence security system
(2)	动词性短语	VP	keep quiet
(3)	形容词性短语	AdjP	very good
(4)	副词性短语	AdvP	at once
(5)	介词短语	PP	with his help
(6)	小句	CS	how are you

3.3 英语短语的中心词信息

标注形式和规范与汉语同，参见上文 2.4。

四 短语描述中的疑难问题

4.1 词和短语的分界

短语库建设中，我们遇到了词和短语的分界这一“古老的”难题。例如，“原封退回”，是切分成“原封/d 退回/v”，还是干脆看成一个词？“即期交货”，怎么标？是“即/c 期/b 交货/v”吗？

完美的理论上的回答是困难的。从短语库构建的工程考虑，短语的切分是为进一步的句法分析服务的。如果词语串内部结合紧密，可以作为一个最小单位（minimal unit）参与句法运算，那么就可以把这个词语串看作一个词。换句话说，如果词语串的切分并不能为句法分析提供有益的信息，那么就可以在词的层面上处理这个词语串，即把它看作一个词。

词和短语的分界可“大概”地遵循下列原则：

（1）有的词语串搭配固定，词与词（或字与字）之间结合紧密，很难分开一个字一个字来理解，可将其看作习用语，如“拔腿就跑，原封退回”等。

（2）有的词语串作为整体指向一个实体（entity），而且凭感觉出现频度很低，处理为一个词比较合适。如“扒钉”，“把杆”，“无蜡裂化设备”等等。

（3）带有文言色彩的词，从合。如：“群起而攻之，偶尔为之”。

（4）如果分开来内部结构不清晰，不好标词性，不好标结构关系，那么就从合。象“即/c 期/b 交货/v”，“即”和“期”的词性都不好确定，不如合起来：“即期/d 交货/v”。

（5）如果分开来是汉语中很普通的组合模式，词性标记和短语内部结构关系都非常清楚，那么就从分。如：不容/v 反悔/v vp SB

（6）在两可或不好判定的情况下，从合。

判定词语的分合，应该有两个标准：（1）对不对；（2）好不好。“对不对”适用于典型场合，“好不好”适用于模糊地带。象“两眼无神”，标成“两/m 眼/n !无/v 神/n”或“两眼无神/v”都不能算错，但就服务于机器翻译而言，合起来作为一个单位“两眼无神/v”要好一些。

4.2 歧义短语结构

短语库描述的是静态的、脱离具体语境的短语，如果没有依托、参照点，歧义短语就无从判别。例如：

（1） 癌扩散了。

（2） 要防止癌扩散。

“癌扩散”在（1）中是个小句，在（2）中是个名词性短语。

考虑到是双语对照的知识库，我们以对应的英文翻译作为汉语短语排歧的参照点。即当汉语短语是一个歧义结构时，参照英文翻译来判定。如：

表 5

汉语短语	英文翻译	汉语短语类	汉语短语结构
工资/n !冻结/v	wage freeze	dj	ZW
自诉/v !案件/n	case of private charge	np	DZ
癌/n !扩散/v	proliferation of cancer	np	DZ
经济/n !崩溃/v	economic collapse	np	DZ

五 短语库的实施和应用

5.1 短语库的实施

目前，短语库已完成了 50,000 条短语汉语部分的加工。短语的来源主要有二：一是继承北京大学计算语言学研究所已有的 40,000 多条汉英对照的短语，在此基础上重新加工；二是从短语级对齐的汉英双语语料中抽取了近 10,000 条短语，进行加工。

实施流程：（1）先利用现有的软件资源对短语进行预处理，自动产生相关信息：汉语短语的切分和词性标注；中心词；短语类；短语结构；（我们利用的是清华大学周强博士的短语标注软件）；英语短语的词性标注（采用了 Brill, E 的词性标注软件）。（2）制订详细的规范，进行人工校对和加工。（3）质量把关，合格检查。短语库的错误率控制在 1% 以下（按短语条目计算）。

短语库现收录短语统计：

表 6

短语类	np	vp	mp	sp	tp	ap	dp	dj	pp	word	总计
短语条目	25684	12322	199	160	167	442	8	2508	192	9931	51613

5.2 短语库的应用

以短语结构微引擎的形式，短语库已成功集成到了“面向新闻领域的汉英机器翻译系统”中。只要短语匹配，就会高效地产生漂亮的译文。但存在的问题是：（1）短语的命中率不高。短语库规模的扩大可以帮助解决这个问题。另一方面，也可以改变算法，例如，改精确匹配为模糊匹配可以提高命中率。（2）短语结构歧义。当静态的短语放到动态的句子中去后，有时会产生歧义。短语结构歧义性质上和切词中的词语歧义一样，可以用类似的方法来解决。同时，在选择短语入库时，要有所甄别：尽量选择高频出现的、不会产生歧义的短语。

短语库是一个丰富的语言资源，可在此基础上对汉语短语进行各种定性的和定量的研究。（1）分门别类对各种短语进行研究。可以研究特定的短语类：名词性短语、动词性短语的构成有哪些类型？各组成情况在概率分布上有什么特征？可以考察特定的词类序列：v + n 构成名词性短语和动词性短语的概率分别是多少，条件是怎样的？（2）短语和中心词的研究。各种短语的中心词可以由哪些词类来充当？从中心词到短语，在句法语义上传递了哪些东西，发生了什么变异？（3）汉语短语和英语短语的对比研究。例如，汉语名词短语可以转换成英语什么类型的短语？汉语动词短语转换成英语名词短语的条件是什么？

六 结束语

汉英双语短语信息数据库的建设已经顺利完成第一期的工作。从质的提高和量的扩展两个方面，我们正在完善短语库。汉英双语短语信息数据库将在机器翻译系统、汉语短语研究等相关领域发挥作用，提供强大的语言知识支撑。

参考文献：

[1] 刘群，俞士汶. 1998. 汉英机器翻译的难点分析. 黄昌宁主编，《1998 中文信息处理国际会议论文集》. 北京：清华大学出版社.

[2] 林杏光，王玲玲，孙德金主编. 1994. 《现代汉语动词大词典》. 北京：北京语言学院出版社.

[3] 齐沪扬. 2000. 《现代汉语短语》. 上海：华东师范大学出版社.

[4] 俞士汶. 1998. 《现代汉语语法信息词典详解》. 北京：清华大学出版社.

[5] 詹卫东. 2000. 《面向中文信息处理的现代汉语短语结构规则研究》. 北京：清华大学出版社.

[6] 周强等. 2001. 《构建大规模的汉语语块库》. 黄昌宁，张普主编，《自然语言理解与机器翻译》. 北京：清华大学出版社.

[7] 《宾州树库词性标注集及手册》，见 http://www.cis.upenn.edu/~treebank

[8] 俞士汶 .1999.现代汉语语料库加工——词语切分与词性标注规范与手册. (内部资料)

[9] E.Brill, 1995, Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging, Computational Linguistics, Volume 21, Number 4

致谢：本项研究工作是集体智慧的结晶。短语库得到了俞士汶教授的关心和支持。课题组其他成员刘群、周强、王厚峰等老师贡献了他们的智慧和力量。北京大学计算机系的叶嘉明和吴拥华同学先后提供了软件技术支持。北京大学其他院系的同学宋新华、凌金良等帮助校对了其中的短语。

作者简介：吴云芳，博士生，主要研究方向为现代汉语语法、机器翻译；常宝宝，博士，主要研究方向是计算语言学；詹卫东，博士，主要研究方向是现代汉语语法、计算语言学。

Building Chinese-English Bilingual Phrase Database

Wu Yunfang [1,2] Chang Baobao [2] Zhan Weidong [1,2]

1 Chinese Department, Peking University

2 Institute of Computational Linguistics, Peking University

Abstract: This paper outlines the construction of a Chinese-English bilingual phrase database: Chinese phrase information and English phrase information; the frequently asked questions; the implementation and application of phrase database. The phrase database is a powerful language knowledge resource and will be some help to phrase structure study、Chinese parsing and Chinese-English machine translation.

Key Words: Chinese-English bilingual phrase　phrase structure　machine translation