信息领域汉英术语的特征及其在语料中的分布规律

邢红兵

(北京语言文化大学语言信息处理研究所
中国科学院自动化研究所模式识别国家重点实验室)

关键词:汉英术语;信息领域语料库;术语自动抽取

摘　要:在对725万字的信息领域专业文献中带英文注释的术语(汉英术语)进行了人工标记,然后利用程序提取汉英术语及其前界环境(前至少4个汉字)的工作基础上,本文对汉英术语的自身特征和前界环境进行了分析,目的是为术语的自动抽取提供规则及相关统计数据。

Structural Features and Distributions of Chinese – English Terms in the Corpus from Information Field

Xing　Hongbing

Keywords: Chinese – English terms, corpus from information field, automatic terminology extraction

Abstract: The research is based on all Chinese – English terms in a 7.25 – million – Chinese – character corpus from information field. We labeled all these terms manually, then we extracted terms with their front border (at least 4 Chinese characters). In this paper the structural features and distributions of these terms are analyzed for further automatic terminology extraction.

一、引　言

随着计算机技术、网络技术的不断发展,信息领域的语言也表现出飞速的变化,这种变化主要体现在词汇这个层面上,大量的新术语不断产生,一批旧的术语逐渐消亡。因此,基于动态更新语料库的术语抽取就显得很有必要,因为我们可以通过抽取新的术语对已有的术语库进行数量及频度的及时更新,这样就可以建成一个动态更新的术语数据库。要实现对大规模真实文本中术语的自动抽取,就必须研究术语的特点及其在文本中的分布情况。术语是指一门学科中的专门用语。在专业领域文献中的术语的分布主要有三种情况:(1)术语处于特殊的位置,比如关键词和注释中的术语等;(2)在新出现的或者作者认为比较新、比较难懂的术语后加上注释,并把注释用括号括起来,有的作者在自己的文章中第一次使用某个术语的时候,还要对术语进行解释;(3)术语无任何标记。对上述各类分布的术语进行自动抽取的难度也不相同,第一类术语有明显的前界和后界,比较容易提取;第二类术语的右边界已经明确,需要确定该术语的左边界;第三类术语自动抽取的难度最大,因为提取这类术语不仅要确定前界和

后界，还要判断这个语言片段是术语还是一般新词。本文仅探讨属于第二类分布的术语自身的特点及其在信息领域语料库中的分布情况，研究结果也可以为第三类术语的自动提取提供各种规则和统计数据。

二、信息领域动态流通语料库建设及术语动态更新

我们设想的术语动态更新是基于专业领域动态流通语料库的，这样的语料库属于第三代语料库。下面分别谈谈第三代语料库的特点及信息领域动态流通语料库建设的设想以及目前的进展情况。

(一)关于第三代语料库

语料库的发展已经历了第一代和第二代，目前正向第三代语料库发展。第三代语料库的显著特征就是数量大、对语料库的加工从词法级到句法级再到语义和语用级。但是这些还不是本质的变化。张普教授在 1999 年先后提出了“第三代语料库”、“语言知识动态更新”的设想，并撰文进行了较为详细的论述。张普教授指出：我们提出的“动态流通语料库”是第三代语料库。衡量语料库是否进入新的一代，不仅看贮存数量，还要看加工深度，“动态流通语料库”为语料库的深加工提供了两个极为重要的新属性：动态性和流通性。

所谓动态性主要体现在以下几个方面：库容量随时间的推移不断扩大；每个时间段选取的语料数量也是变化的；语料的抽取是分领域的，在这个动态流通语料库中，通用领域和各专业领域的语料是共同存在的；语料是根据媒体的流通情况抽取的，因此，每个时间段语料的来源也是不固定的。流通性是指被抽样的文本都具有一个很重要的属性——流通度，流通度是一种语言现象在社会传播中的流行通用程度。流通度是可以量化的，量化的值主要取决于文本的发行量、发行地区、发行周期等数据。

(二)信息领域动态流通语料库的建设

建设信息领域动态流通语料库主要是基于以下几个方面的原因：(1)第三代语料库应该包括通用领域语料库和专业领域的语料库，其中任何一个领域都是可以独立的；(2)信息技术已经渗透到各个领域，对其他学科甚至对人们的日常生活都有较深刻的影响，信息领域的语言由于其媒体的种类繁多、发行量大、借助网络等手段发行地区广、阅读率比较高，这样的语言最能体现第三代语料库动态性流通性的特点，因此，也就最具有代表性；(3)各类报刊杂志的电子版本、网络版本的出现也使语料的获取变得相对比较容易，语料库的建设成本会大大降低。信息领域语料库同样具有第三代语料库的典型特征：动态性和流通性。信息领域语料库的主要特征是：(1)语料库中语料的数量将随时间的推移而不断扩大；(2)每次扩充的语料都是流通度比较高的媒体的语料；(3)全部语料按照时间顺序排列；(4)所有的文本均带有以下标记：领域标记、文本的流通度、发表时间、媒体信息(包括媒体的类型、级别、名称等)、次类标记、作者信息。

目前我们的语料库中生语料已经达到1亿字以上，人工标注的英汉双语术语的语料700多万字。我们有两个术语数据库：(1)利用程序抽取标记好的术语，形成双语数据库及其前接成分数据库，共4600多条记录；(2)利用程序直接提取关键词，形成信息领域库共9000多条记录。

三、信息领域双语术语的类型

(一)英汉双语术语的人工标记

我们随机抽取 720 多万字的信息领域专业文献，对上述的第二类术语(主要是汉英双语术语)的前界进行人工标记，然后利用程序提取双语术语及其前界环境(前至少 4 个汉字)，形成一个英汉术语库，该库共有 4607 条记录。此类术语都是由前部(括号以外)和后部(括号以内)

两部分构成。如果既考虑前部也考虑后部，就是说前部或后部有一部分不同的话，如“自底向上（Botom – up）”和自底向上（Botom up）”、“遗传算法（GeneticA lgorithms）”和遗传算法（GA）”，就认为它们是不同形式的术语，那么725万字的语料中英汉双语术语共有3426条。如果只是考虑前部不考虑后部的话，如将上述两个“遗传算法”算作一条术语，那么语料中的双语术语共有3032条。全部术语共使用4607次。

（二）英汉双语术语的类型及比例

我们根据双语术语前后两部分的情况将全部3426条不同形式的术语分为以下8种类型：

A类：全中文术语（全英文注释）。例如：“表（Table）”、“插件（plug – in）”、“超文本（Hypertext）”、“项目控制（project control）”、“封装（package）”、“数据仓库（Data Warehouse，简称DW）”、“数据挖掘（Data Mining，简称DM）”、“标准数据访问接口（Standard Data Access Interface：SDAI）”等。

B类：全中文术语（英文缩写注释）。例如：“光磁（MO）”、“安全识别符（SID）”、“并行分布处理（PDP）”、“对象标识符（OID）”、“多级安全（MLS）”、“分布式相关数据库结构（DRDA）”等。

C类：英文缩写术语（全英文注释），例如：“XML（Extensible Markup Language）”、“COM（Component Object Model）”、“C/S（Client/Serve）”、“URL（Universal Resource Location）”、“COSS（Common Object Service Specification）”等。

D类：英文缩写术语（全中文或中英文注释），例如：“SQL（结构化查询语言）”、“UDP（用户数据报协议）”、“CSCW（计算机支持的协同工作）”、“ATM（Asynchrous Transfer Mode，异步传输模式）”、“IDC（Internet Database Connector，Internet数据库连接器）”、“ADC（Advanced Data Connector，先进数据连接器）”、“CGI（Common Gateway Interface通用网关接口）”等。

E类：中英文术语（全英文或英文缩写注释），例如：“商务智能BI（Busines Inteligent）”、“电子商务EC（Electronic Commerce）”、“Java数据库互联（JDBC）”。

F类：全英文术语（英文缩写注释），例如：“Distributed shared memory（DSM）”、“Active Server Page（ASP）”、“Active Data Objects（ADO）”。

G类：全英文术语（中文注释），例如：“Aliases（别名）”、“Rotfilesystem（根文件系统）”、“Database Access Component（数据库访问组件）”、“Background Music（背景音乐）”、“Telnet（远程登录）”。

H类：中英文术语（中文注释），例如：“DDN网（全国公用数字数据通信网）”、“BP算法（误差反向传播算法）”、“RSVP协议（资源预约协议）”。下面是各类型的不同形式术语的数量及其比例。

类别	A类	B类	C类	D类	E类	F类	G类	H类
数量	1728	731	343	172	260	126	57	9
比例（%）	50.44	21.34	10.01	5.02	7.59	3.68	1.66	0.26

从上表看出，我们所说的第二类术语主要是A类和B类术语，两类约占72%，这两类也是比较典型的英汉双语术语。C类和D类术语也有一定的数量，可见很多中文文献中直接使用英文缩写术语。后面几种情况应该说是属于不太规范的用法，如在汉语文献中直接使用英文全文，然后用中文术语进行注释。从上面的分析来看，专业文献中术语的使用还存在着若干不一致的问题，这主要表现在以下几个方面：（1）汉语术语名称不一致。例如“HTML”有

‘超文本标记语言’和‘超文本标志语言’两个汉译名；(2)同一术语有的用汉语术语，有的直接使用英文缩写形式。例如有的文献直接使用“XML”，有的文献使用‘可扩展标签语言’或‘可扩展标识语言”，有的还使用“XML语言”等；(3)相同的中文术语对应不同的英文术语，例如：‘电子商务’对应的英文注释就有‘E-business’和‘Electronic Commerce’等。除此之外，还有其他不一致的地方，这些都是术语使用规范的问题，对于这个问题，本文不打算深入研究。

四、信息领域双语术语的分布特征

我们要对实际语料中的术语进行自动标记，就必须对术语的自身特点及其在语料库中的分布特征进行分析，具体地说，应该包括以下几个方面：术语的字数、词数，术语用字字频和用词词频；术语中字和词的互信息；双语术语中中文词和英文词的匹配情况；术语的前接和后接成分分析；术语的结构类型等。本文只对术语的字数及其前接成分进行简单的分析。

(一)术语的长度

上述术语中，A类和B类是典型的中英文双语术语，它们的主体部分是中文，这类才是严格意义上的中文术语。C类和F类实际上是英文术语，但也可以直接用在汉语专业文献中，可以看成汉语对英语的直接借用。E类主体部分是中英文混合，也是一类中文术语，D类、G类和H类实际上正好相反，括号里才是中文术语。根据这样的情况，我们将前ABCFE类术语的前部和DGH的注释部分提取出来，再将一些不一致的表示法进行整理，形成一个不带注释的术语表，这个术语表共有3032条术语。这3千多条术语分为三类：(1)全中文术语；(2)全英文缩写术语；(3)汉语和英文缩写混合术语。这三类的数量及比例见下表：

类　别	全中文	全英文	中英文
数　量	2480	352	200
比例(%)	81.79	11.61	6.60

下面我们先分析一下中文术语的字数，具体数据见下表：

字数	1	2	3	4	5	6	7	8	9	10	11	>12
数量	20	337	275	594	296	405	198	147	79	57	28	31
比例(%)	0.81	13.59	11.09	23.95	11.94	16.33	7.98	5.93	3.19	2.30	1.13	1.77

从上表可以看出，中文术语主要是2~6个字，共占76.9%，其中4字的最多，约占24%，而且在6字以下的术语中，两字术语多于单字术语，四字术语多于三字术语，六字术语要多于五字术语，1、3、5字术语占23.84%，2、4、6字术语占53.87%，这主要是因为汉语双词的优势决定的。下面我们分析一下英语缩略术语的字母长度，见下表：

字母数	2	3	4	5	6	>7
数量	23	189	92	22	10	16
比例(%)	6.53	53.69	26.14	6.25	2.84	4.55

汉语文献中使用的英语缩略语术语主要是3字和4字的术语，一般由3个或4个英文单词缩略而成。

(二)术语的前界

我们对全部术语的前接成分进行了分析，发现常和术语前接的有两类：符号和常用词，符号包括标点符号、数学符号以及其他符号。常用词主要包括助词、连词、介词和部分系动词。具体数量参见下表。

前接成分	符号	连词				动词					
		和	或	及	与	为	是	如	有	即	叫
数量	1548	291	39	23	30	218	137	39	45	17	13
比例(%)	33.60	6.32	0.85	0.50	0.65	4.73	2.97	0.85	0.98	0.37	0.28
前接成分	介词							量词		助词	
	用	在	于	通过	由	对	以	个	种	的	了
数量	162	70	63	57	43	19	18	111	23	590	126
比例(%)	3.52	1.52	1.37	1.24	0.93	0.41	0.39	2.41	0.50	12.81	2.73

上述的术语主要前接成分共有3646个，占全部4607个术语的79.14%，近3成的术语与一些符号相连接，它们有的处于段首或句首，有的在顿号的后面；有15%的术语在助词后面。可见，大部分术语的前接成分还是比较明显的，通过找前接成分可以确定一部分术语的前界。根据这样的前接成分分析，我们预测术语的后接成分也会有一些明显的标志成分的。

五、以后的工作

本文只是对双语术语的特点及其分布规律进行了简单的分析，我们下一步将利用目前标注语料库的统计数据对大规模的语料进行术语的自动标记，标记的术语不仅仅是第二类术语，还包括第三类术语，标注的过程为机器自动标注后进行人工校对，然后以现有的术语库为启动数据库，在此基础上实现术语数据库的动态更新。这样的术语数据库将为专业工具书的编写和修订以及术语的规范提供最新的资料。

参考文献

[1] 张普，关于语感与流通度的思考，语言教学与研究，1999年第2期

[2] 张普，关于网络时代语言规划的思考，语言研究，1999年第3期

[3] 张普，关于第三代大规模真实文本语料库的几点理论思考，载《自然科学基金重点项目结题报告》

[4] R. Basili, L. Bordoni & M. T. Pazienza (1997): Extracting Terminology from Corpora. 载《第二届术语学、标准化与技术传播国际学术会议论文集》，中国大百科全书出版社，1997年

[5] 张普，信息处理用语言知识动态更新的总体思考，语言文字应用，2000年第2期

[6] 邢红兵，基于第三代语料库的信息领域术语动态更新，语言文字应用，2000年第2期

[7] 隋岩，动态流通语料库理论的概念和方法，语言文字应用，2000年第2期

[8] 王建华，Study on the Automatic Acquisition of Bilingual Terms Between English and Chinese，中国矿业大学硕士研究生毕业论文