中文缩略语知识库建设

支流， 段慧明，朱学锋，俞士汶

（北京大学计算语言学研究所，北京 100871）

摘 要：缩略语是自然语言语汇的重要组成部分，是未定义词的主要来源之一，因此，缩略语研究是自然语言处理的一个重要课题。本项研究的最终目标是探索中文缩略语的规律，包括缩略语的生成和还原，也就是缩略语的编码和解码。本项研究旨在建立一个中文缩略语知识库，协助机器自动还原中文缩略语。我们建立了面向信息处理的中文缩略语分类体系，完成了 8000 个缩略语的归类，建立大规模缩略语知识库（每个记录主要包括缩略语、其对应的完整语汇或全称以及缩略语的同形信息等）。根据已完成工作的经验，对中文缩略语出现的形式进行了总结和分类，提出了针对特殊缩略语的自动还原办法。中文缩略语的研究也可为各种语言缩略语共同规律的研究提供数据基础和技术借鉴。

关键词：自然语言处理，缩略语，未定义词，缩略语数据库，缩略语知识库

Construction of Chinese Abbreviation Language Knowledge-base

Zhi Liu， Duan Huiming， Zhu Xuefeng， Yu Shiwen

(Institute of Computational Linguistics, Beijing 100871)

Abstract: Abbreviation is an essential part of Natural Language, and an important source of Out of Vocabulary, therefore, the study of abbreviation is a significant section of Natural Language Processing. The object of the study is to investigate the rules of the Chinese abbreviation, and how to generate and restore the abbreviation. The authors' aim was setting up a Chinese Abbreviation Language Knowledge-base, which was used for restoring abbreviation by computer. The work is based on the two basic sources—The Grammatical Knowledge-base of Contemporary and the Large-scale Tagged Corpus. The authors advance the categorize frame of Chinese Abbreviation facing to information processing and accomplish classifying 8000 abbreviations. On the base of the Chinese abbreviation language knowledge-base, the authors design new restoration algorithms for special abbriviations. In this paper, we introduce our accomplished work and the plan of the future work. Meanwhile, the investigation of Chinese Abbreviation could supply data and technique for all the other languages' investigation.

Keywords: Natural Language Processing, Abbreviation, Out of Vocabulary, Abbreviation Language Knowledge-base

1 引言

当今社会信息量迅速膨胀，省力省时的缩略语便大行其道，继而成为自然语言处理中无法回避的问题。 许

基金资助：相关研究得到国家 973 项目（2004CB318102）和香港大学的支持

作者简介：支流（1984-），女，安徽，学生，硕士，zhiliu@pku.edu.cn.

多语言学家很早就开始重视缩略语问题并总结出各种缩略语的规律，并建立了一些包括了缩略语和对应全称的数据库。由于此前的很多研究不是以信息处理用为目的的，所以总结的规律以及缩略语数据库的结构不大适合计算机自动处理。

利用计算机进行中文缩略语规律探索和还原为全称的研究很少有人涉足，即使一两次的尝试也只是停留在浅层平面上。这种情况导致中文缩略语的规律探索停留在小规模的语料中进行，人们无法在只有计算机才可以处理的超大规模语料中观察缩略语的规律，从而所得的规律无法在真实的语料中得到验证，利用计算机自动做缩略语还原的工作更无从谈起。笔者一直致力于计算机自动还原中文缩略语的研究，在研究过程中发现：大约 40%左右的缩略语全称已经不出现在文章中；中文缩略语中一对多、多对多的现象很多，现有的程序难以达到令人满意的效果；中文缩略语形成方式虽有基本规律，但随意、多变，机器自动识别和还原的难度都很大。北京大学计算语言学所俞士汶老师曾说过："任何一个自然语言处理系统都需要语言知识库的支持，语言知识库的规模和质量在很大程度上决定了自然语言处理系统的成败。这应该是计算语言学研究者的共识。当把汉语作为研究对象时，自然会重视汉语语言知识库的建设。不过，如果能从信息处理的角度，更加清晰地认识汉语的特点，对提高建设汉语语言知识库的自觉性和知识库的适用性，是有启示作用的。"由此，笔者认为建立一个包括了缩略语的全称、简称，特征的缩略语库来支持缩略语还原势在必行。缩略语知识库的建设也为北京大学计算语言所语言知识库的建设添砖加瓦。

本文首先介绍中文缩略语研究现状；接着结合对缩略语知识库中各种形式的缩略语的观察，提出合理的缩略语分类体系；然后介绍缩略语知识库的组织形式；最后对现有的缩略语还原技术难以还原的总结式缩略语提出新的还原方法。

2 缩略语库建设现况

先前已经有研究者做了建立缩略语库的工作，并取得了一些成就。

北京大学语言所在 87 年建立《现代汉语语法信息词典》（以下称《语法信息词典》）时就已经开始重视缩略语的现象并建立了《简称略语库》，这个库文件包括了 608 条简称略语以及这些词语的拼音，所属子类，是否可以做主谓宾定状补等成分，还说明了这个简称所对应的全称和一些使用的例子。这个库文件也是本文介绍研究的最重要的资源之一。

烟台师范大学的鲍明凌和亢世勇在《基于数据库的现代汉语新词语缩略语的研究》[4]一文中介绍了他们的工作，他们建立了一个关系数据库用来存放缩略语的一些属性，有原词语，类型，缩略方式、构成方式、结构。他们所建立的库中共收录了 2957 条缩略语。

以上的缩略语库的规模和质量堪称上乘，但是由于建库的初衷并非是为了计算机自动处理，所以数据库的组织方式以及属性设置不适合计算机自动处理。

3 本文中"缩略语"所指对象

不少学者研究汉语中的缩略现象，随着各种论文的发表，不同的研究者在如何称说语言中的这一现象上出现了不小的分歧——具体表现为称名上的五花八门。除了"缩略语"外还有"缩语"、"简称"、"简略语"、"略语"、"缩约语"、"省称"、"省文"、"省语"等等。

现在使用的上述各种术语中，"简称"、"缩略语"的使用较广，出现的频率较高。学者们各自实际运用的"简称"、"缩略语"时所指的内涵和外延并不如字面的名称一样完全相同。很多人认为"简称"仅仅用于名词，是名称的简化形式。《现代汉语词典》（修订本）认为"简称"是"较复杂的名称的简化形式。" 所以本文中选取缩略语作为这类现象的通称。由于本文中介绍的研究是在《语法信息词典》和"大规模现代汉语基本标注语料库"（此后简称为语料库）的基础上开展的，《语法信息词典》是北大计算语言所语言资源的第一块基石，它对词的分类中有一类词即为"缩略语"，这类词的词类一项填充的是"j"。《语法信息词典》中还建立了《简称略语库》，这也是本文工作的基础，下文中将会介绍。语料库是与《语法信息词典》一脉相承的，它的切分和标注都是以《语法信息词典》为基础的，缩略语的标注也是将《语法信息词典》中的缩略语标注出来，并人工的标注了其他的缩略语。为了与《语法信息词典》和语料库的切分标注规范保持一致，所以本文中缩略语具体是指语料库中词性标注为"j"的那部分词语。《北京大学现代汉语语料库基本加工规范》中指出："表达一个完整概念或集合的缩略

语为一个切分单位，并标以 j。”在本文所指的缩略语中各地的地名简称别称也包括在内。

4 缩略语库所利用的资源简介

本项研究中建设的中文缩略语库的原型是北京大学计算语言学研究所的《语法信息词典》中的《简称略语库》。《简称略语库》的信息在上文已经介绍过。《缩略语库》在继承了《简称略语库》的部分属性的基础上，增加了一些新属性，下文将详细介绍。

除了《简称略语库》中的 608 条缩略语，《缩略语库》中另外 7000 多条缩略语，取自 “大规模基本标注语料库”的 1998 年和 2000 年两年切分标注的《人民日报》语料中标注为“缩略语 j”的部分词语。由于是面向信息处理用的，库中收录的缩略语包括了一部分地名简称、别称、政府文号等等，属于广义的缩略语。

5 缩略语的分类框架

研究缩略语首先要清楚可以将缩略语分成哪几类，只有针对不同类型的缩略语对症下药，制定不同的处理方案才能使缩略语的还原和规律探索工作事半功倍。在研究缩略语还原的实验中，我们首先设置了一套适合计算机识别和处理的缩略语的分类框架。在建设缩略语库时，我们对收录到库中的缩略语进行了归类实践，并进一步修正了原有的分类框架。我们对中文缩略语进行了多方面的考察，根据不同的特性按照各种标准对缩略语进行划分归类，类别不同代表缩略语的属性不同。

中文缩略语的多种分类方式，首先最简单的一种就是按照缩略语和全称的对应关系进行分类，可以将缩略语划分为以下几类：

1) 缩略语与全称一对一。例如“北大”对应于“北京大学”。
2) 缩略语与全称一对多。例如“人大”可以对应“人民大学”，也可以对应于“全国人民代表大会”。
3) 缩略语与全称多对一。这种情况相对于上两种较少见但也是不可忽略的。例如：“南开大学”可以缩略为“南开”，也可以缩略为“南大”；“电风扇”可以缩略为“电扇”、“风扇”；手电筒可以缩略为“手电”、“电筒”。
4) 缩略语与全称多对多。较上种情况，此类情况更加少见。例如：“南大”对应于“南开大学”或“南京大学”，同时“南开大学”又可缩略为“南开”。

对于计算机处理来讲，此种分类方式中的第一类和第三类都很好处理，第二类和第四类是计算机还原缩略语中最大的拦路虎，由于计算机无从判断第二类，第四类中的简称到底应该还原为那个全称，所以需要特殊处理。在建设缩略语知识库时，为了能够方便处理第二类和第四类缩略语，缩略语库采取统一的格式，每一行的简称都对应着唯一的全称，但是同一个简称可以分布在多行，引入了同形的概念，用来区别相同简称，全称不同的情况，这样“简称+ 同形+拼音”就构成了缩略语知识库的主关键项。

另外一种缩略语分类方式是从缩略语形成方式上进行分类。

首先分成可以将缩略语分成与全称无关和与全称有关两种。无关即为缩略语中的内容与全称中的内容是没有什么关联的，这种情况中地名占了绝大多数。例如：云南的简称是“黔”；上海的简称是“沪”。这种情况往往是长久历史积累下来的缩略语，是无法用模糊匹配的方法找到其全称的，只能靠以往的经验获得。有关的即是指在缩略语中有成分与全称是相同的，在还原的过程中是有迹可循的。

将与全称有关两种缩略语划分成带括号的缩略语和不带括号的两大类。带括号的缩略语，大多表示的是两个词组“或”的关系，但并非全部如此，也有一部分表示各成分之间“和”的关系。它的缩略方式有 2 种，根据缩略的方式不同可以分为以下两类：

1) 省略式。其全称是两个词组，其中一个词组比另外一个词组多了一个字或词，将这个多出来的部分放在括号中，其他的部分放在括号外，形成了缩略。例如：“（外）祖父母”，其全称是“外祖父母或者祖父母”，这个缩略语中的祖父母也是一个缩略语，它的全称是“祖父和祖母”。
2) 替换式。括号中的部分与括号外左侧相邻的部分是可以相互替换的，形成两个不同的并列词组。例如：县（市）长，其全称是“县长或者市长”；文化馆（宫），全称是“文化宫或者文化馆”。

中文有很多地方用到括号，但是并非用到括号的都是缩略语，有些情况有可能被误认为是缩略语，例如：区（县）；这种情况中整个词语中并没有被缩略的部分，其代表的就是区或者县。

对于没有括号的缩略语，针对每一类进行分析，进一步划分：

1. 提取式缩略语：从一个或几个连续的词语中提取出字或词语拼接在一起，形成缩略语。这类缩略语的形成方式比较简单，但是也有以下几类情况：
   a) 缩略语完全遵从全称的词语和顺序，并且全称是一个连续的短语。即为全称中的所有的词语的全部或一部分出现在缩略语中，例如“北大”，它的全称是“北京大学”。
   b) 全称的词语只有一部分出现在缩略语中。例如“清华”，它的全称是“清华大学”；“上菱厂”，它的全称是“上海上菱电冰箱总厂”。
   c) 从全称的意义上提取而非从字面提取的，例如“费改税”，其全称是“把部分具有税收特征的收费以税收代替”。这类缩略语比较特殊，它的形成和形式都和一般的缩略语非常相象，但是它的处理方法却和别称类似。
2. 总结式缩略语
   例如经常出现的“三个代表”，“三好”，“五好”等等，这些是从一个意群中总结出来的。总结式缩略语一般还原成“和”的关系，这类缩略语一般比较规范。但是存在一个问题，一对多的数目过多。例如在《人民日报》1998 年和 2000 年的切分标注语料中出现“两证”对应的全称就多达 35 个。
3. 缩合式缩略语
   根据共同部分所在位置不同，可以分成前缩式和后缩式。前缩式例如“辍失学”，它的全称是“辍学、失学”；后缩式例如“场内外”，它的全称是“场内、场外”。

上述的这种分类方式是我们在此后的工作中使用最多的，我们认为也是最能够说明计算机可区别的特点的分类，针对每类不同的缩略语需要采取不同的还原方式，在缩略语知识库中我们为缩略语标注的第一类属性就是表明缩略语在这个分类中属于哪一类。

还可以按很多方式给缩略语分类，这些分类方式都比较偏重于缩略语的语法属性，此处不一一列举。

6 缩略语知识库的组织方式

中文缩略语知识库与《语法信息词典》一脉相承， 它的一部分直接源自《语法信息词典》的简称缩略语库，《简称略语库》的格式如图所示：

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	词语	全拼音	同形	义项	粘着	子类	定语	状语	谓语	补语	主语	宾语	兼类	备注
2	四害	si4hai4		老鼠、臭虫、苍蝇、蚊子等虫害及泛指		JN	定				主	宾		除~/~横行
3	俄	e2		俄罗斯（国）		JB	定							
4	鄂	e4		湖北的别称		JB	定							
5	京九	jinglj iu3		北京到九龙的铁路	粘	JB	定							~铁路早通车了
6	老弱病残	lao3ruo4bing4can2		老人、体弱者、病人、残疾人的总称		JN	定				主	宾		~优先
7	奥委会	ao4wei3hui4		奥林匹克委员会		JN	定				主	宾		
8	一中全会	yilzhonglquan2hui4		中央委员会第一次全体会议		JN	定				主	宾		十五届~

图 1 《简称略语库》样式

中文缩略语知识库共包含了 8000 条中文缩略语和全称的对应，为了保持和《语法信息词典》一致，同样采取 Access 存储成数据库格式的文件。每条记录的主关键项是“词语+全拼音+同形”，这也与《语法信息词典》完全相同，只是“同形”的概念不同，此处的同形是用来“同形”是用来描述一个缩略语对应多个全称的情况。在对简称缩略语库继承的基础上，它加入了新的属性项。缩略语知识库中除了上述的各个属性项外新增属性项的例子：（注：下图中的“提取&&节略”是指该缩略语的形成既有提取也有节略，表示的是“和”的关系；“提取、节略”是“或”的关系，表示该缩略语可以说它的形成方式是提取式，也可以说是节略式。）

原有的属性项与语法信息词典的属性项描述大致相同，此处就不再一一赘述，重点介绍五个新的属性项，即：形成方式、种类、范畴、语境、使用领域。新增属性说明如下：

1. 形成方式大致分成两种：一种是缩略语的形成与全称有关，这其中又包含了提取式、节略式、总结式等等；另一种缩略语的形成与全称无关，缩略语中不包含全称中的任何字，例如“苍蝇、蚊子、老鼠、臭虫”简称为“四害”，又如“沪”、“申”是上海的别称。

2. 种类包括机构名、地名、政府文号、缩略语前缀、缩略语后缀等等。由于缩略语库收录的是广义的缩略语，包含的种类较多，种类不同的缩略语在形成方式上可能是相同的，但是使用范围和使用方法等方面却大相径庭。同时，缩略语前缀、后缀概念的提出为缩略语自动识别提供了宝贵的线索。
3. 范畴是指该缩略语对应的全称是描述哪个范畴的概念，现在仅分成了时间、空间、人物三个范畴。例如“唐”指“唐朝”时是一个时间范畴的缩略语，而在指“唐山”时是一个空间范畴的缩略语；“马恩列斯”是人物范畴的缩略语。现在的分类还是显得有些宽泛，在后续的工作中需要改进。
4. 语境是指常常与该缩略语同时出现的词语或搭配。

使用领域是指这个缩略语是在哪个领域中使用，现在大致分成了体育、外交、经济等几个方面。例如“男单”、“女单”一般是在体育领域中使用。

	A	B	C	D	E	F	G	H	I
1	词语	全拼音	同形	义项	形成方式	种类	范畴	语境	使用领域
2	二为	er4wei4		为人民服务、为社会主义服务	总结式	其他			
3	二野	er4ye3		中国人民解放军第二野战军	提取&&节略	机构名			
4	法	fa3		法国	提取、节略	国名	空间	中法、法中	
5	森防	senlfang2	1	森林防火	提取				
6	森防	senlfang2	2	森林病虫害防治	提取&&节略	其他			
7	男网	nan2wang3		男子网球	提取	其他			体育
8	内政办	nei4zheng4ban4		内蒙古人民政府办公室	提取&&节略	政府文号			
9	办	ban4		办公室	提取、节略	后缀			

图 2 《缩略语知识库》新增属性

7 总结式缩略语的还原方法探索

总结式缩略语形式多变且继承全称的信息很少，是还原难度最大的一类缩略语。通过对缩略语库中收录的总结式缩略语的观察，我们对总结式缩略语及其全称出现的方式进行了归纳，针对两部分总结式缩略语提出了还原的方法。

一部分总结式缩略语因为使用的时间很长，已经深入人心，例如“三好”（学生），“五讲四美”等等，这些缩略语的全称现在极少出现在文章中，所以将这些缩略语及其对应的全称收录在缩略语库中，还原时直接从库中提取即可。

另一部分形式比较规整的缩略语中包含了全称中的某些字或词语，并且全称与缩略语在同一篇文章中出现。我们总结了一些常见的形式，对缩略语前后语境中的词语和标点进行分析，然后设计算法根据已有的形式提取出该缩略语对应的全称。例如（这些例子是从 1998 和 2000 年的《人民日报》语料中抽取的例句）：

全国有 1 / 3 的检察院达到“班子好、队伍好、业绩好、机制好、形象好”的“五好”标准。

……，依据对虚开“资信证明、存款证明、担保函”（简称“两证一函”）的检查结果，……

……，但却找不到“两证”（即《经营许可证》和《营业执照》）。

以上三个例子分别给出了几种常见的缩略语和全称出现的位置关系以及二者出现的形式，我们就是根据引号、括号以及前后距离等信息进行缩略语还原的。

总结式缩略语中还有一部分缩略语没有继承全称中的任何信息，例如“五好”（家庭）中，“五好”的全称是“尊老爱幼、男女平等、夫妻和睦、勤俭持家、邻里团结”。这一类缩略语由于以下两点所以很难处理：一是因为全称中没有任何字在缩略语中出现，我们的程序无法得到还原的线索；二是因为人们还常常在“五好”的字眼下不断灌入新的内容，缩略语库无法将这些新增的全称及时都收录进来。现在还没有很好的办法还原这类缩略语。

参考文献：

[1] 俞士汶,朱学锋,等.《现代汉语语法信息词典详解（第二版）》[M].北京：清华大学出版社，2003 年 2 月

[2] 俞士汶、段慧明、朱学锋，等.，综合型语言知识库的建设与利用[J].《中文信息学报》，2004 年，第 18 卷第 5 期，1-10

[3] 张志毅 张庆云,《词和词典》，中国广播电视出版社，1994 年 4 月。

[4] 鲍明凌，亢世勇,《基于数据库的现代汉语新词语缩略语的研究》,《第一届学生计算语言学研讨会》，2002 年