多语料库中汉语四字格的切分和识别研究

徐润华[1] 曲维光[2] 陈小荷[3] 王东波[4]

[1](南京大学信息管理学院 南京 210093)
[2](南京师范大学计算机科学与技术学院 南京 210046)
[3](南京师范大学文学院 南京 210097)
[4](南京农业大学信息科学技术学院 南京 210095)

摘 要：汉语四字格的能产性和派生性极强，利用四字格模式创造出的新词数量在现代汉语词汇中一直呈上升趋势。文章将研究的目光投向分词语料库中的四字格,对语料库中的四字格进行了系统的分类和归纳，并对语料库内部和语料库之间的四字格切分不一致现象进行了详细的调查统计。最后，针对四字格的切分不一致数据引入条件随机场（CRF）模型，对多语料库中的汉语四字格进行识别实验，封闭测试和开放测试的识别精度均达到93%以上。

关键词：四字格;分词语料库;切分不一致;CRF 模型

基金项目：国家社会科学基金项目(07BYY050);国家社会科学青年基金项目(10CYY021);江苏省哲学社会科学基金一般项目(10YYB007)

The segmentation and recognition of four-character idioms in Multilingual corpora

Xu Runhua[1] Qu Weiguang[2] Chen Xiaohe[3] Wang Dongbo[4]

[1](School of Information Management, Nanjing University, Nanjing 210093)
[2](School of Computer science and Technology, Nanjing Normal University, Nanjing 210046)
[3](School of Chinese Language and Literature, Nanjing Normal University, Nanjing 210097)
[4](College of Information and Technology, Nanjing Agricultural University, Nanjing 210095)

Abstract: The productive and derivative of four-character idioms are extremely high, the use of four-character pattern to create new words in the vocabulary of modern Chinese is still on the rise. This article looks into the eyes of the large number of four-character idioms in word-segmented corpora, and works on the four-character idioms in corpora for analysis and induction. Then this article works on the segmented comparison of four-character idioms both in single segmented corpora and between different segmented corpora. Finally, through the introduction of CRF statistical model, and take the results of segmented comparison of four-character idioms as training corpora, this article develops the research of the recognition of four-character idioms in corpora. Recognition results show the accuracy of four-character idioms can reach more than 93% in both closed test and open test.

Keywords: four-character idioms; word-segmented corpora; segmented comparison; CRF

一、前言

“四字格”这个术语最早由陆志韦先生（1956）提出，是指由四个汉字组成的一种独特语言格式。在汉语言文学发展的历史中,四字格的形式起着非常重要的作用，四字格形式在语音、语法、构词、语用、修辞等方面都对汉语产生了深刻的影响。四字格不仅在字数、结构、韵律等方面有着独特的优势（马国凡 1987），它还有着深厚的文化土壤，从老子的“千里之行，始于足下”到孔子的“学而不厌，诲人不倦”，名人名言多见四字警句。

四字格结构的能产性和派生性极强，利用四字格派生出新词语的模式,在汉语言发展史上一直起着积极的作用，推动着汉语的发展。直到今天，利用四字格模式创造出的新词数量在现代汉语词汇中仍然呈上升趋势， 四字词语的数量有增无减。杨晓黎（1996）通过

统计得出结论认为“在新词语中双音节优势已经让位于四音节词语了”。 在信息化迅猛发展的今天，可以很容易地获取大规模语料，对四字格结构的研究不能仅仅局限于文献和理论，而应该将目光更多地投向语料库，投向大量真实文本中的四字格。

语料库中的四字格所面临的最大问题是，同一个词在文本中是否保持了相同的切分形式。如果不能很好地解决四字格的切分和识别工作，会给汉语的自动分词工作带来麻烦。目前，自然语言处理尚缺乏对汉语四字格的专门性研究，本文希望通过对语料库中汉语四字格的研究，给自然语言处理领域的自动分词工作，以及在自动分词基础上进行的语料深加工、句法分析、话语理解等后续任务带来有益的帮助。

二、语料库中四字格的分类

不同语料库中的四字格由于语料来源、语言风格、切分原则等方面的差异而呈现出多样性和特殊性。为了能更好地揭示出语料库中四字格的全貌，引入多个语料库来进行四字格的研究十分有必要。本文研究所选用的分词语料库是 Sighan 中文分词竞赛的部分训练语料[1]：北京大学《人民日报》分词语料库、微软亚洲研究院中文分词语料库、中国国家语委中文分词语料库这三个简体中文分词语料库。选取三个不同的语料库来进行四字格的分类工作，可以更全面地考察多语料库中的四字格并在此基础上进一步比较各个语料库之间的四字格切分特点。

1. 语料库中四字格的筛选

语料库中的四字格，通常指的是那些结构稳定、意义凝固、可独立运用且长度为四的词语。因此在分词语料库中，四字格最直观的形式特征就是它的长度。但长度为四的分词单位，并不一定都是四字格。其中包含了相当数量的“非四字格”却长度为四的分词单位。这些“非四字格”分词单位主要由数字串、命名实体等构成。例如：数字串或含数字串的词串：“七八千元”、“五十余篇”；人名和地名，尤其以音译词居多，如“莎拉波娃”、“巴勒斯坦”；机构名如“顺天集团”、“西康铁路”等。这些四字长的分词单位是不能归入到四字格范畴中去的。这里需要说明一下成语。成语多是四字格（莫彭龄 2003），它结构稳定基本不存在切分不一致的情况。而且作为一个封闭的类别，对成语进行识别也较为容易。基于成语的结构稳定、容易识别的特点，为了更直接地针对开放性、派生性强的四字格结构进行研究，成语也没有被纳入本文研究的四字格范畴之内。

通过筛选、去除上述成分之后，北大《人民日报》98 年 1 月分词语料库中四字格筛选后数量为 2830 条；微软亚洲研究院中文分词语料库中四字格筛选后数量为 2739 条；中国教育部国家语委分词语料库中四字格筛选后数量为 1999 条。

2. 语料库中四字格的分类方法

周荐（1997）把四字格分为“陈述式”、“偏正式”、“述宾式”等八种类型。本文提出的四字格的分类方法，并不单纯从语法层面去分析四字格的结构，而是更偏重于为计算机处理四字格切分、识别任务而服务的一种分类方法，所以对四字格结构内部的组成关系并不十分关注。分词语料库中的四字格，按照四字格的构词模式，大致可以分为“词语构成型”、“结构构成型”、“固定结构型”这三种类型。如表 1：

[1]语料来源网址：http://www.sighan.org

表 1 四字格的构词模式

词语构成型	“2+2”式	父老兄弟、深山密林、行政区域
	“3+1”或“1+3”式	保险费率、单淘汰制、许可证费
结构构成型	四字骈语	锅碗瓢盆、王侯将相、吃喝玩乐
	“结构”＋“结构”式	笔简意深、餐风宿雪、惩恶扬善
	“结构”＋“词”式	大饱耳福、如沐熏风、繁华似锦
	复杂结构式	事随时变、照做不误、重中之重
固定结构型	无论如何、也就是说、总的来说	

词语构成型，指的是那些内部构成方式简单易于观察，在语料库中常常切分成两种稳定切分形式的四字格，例如“工薪阶层”，或者切分为一个完整的四字格结构，或者就是切分为“工薪　阶层”的形式。词语构成型还可以细分为“‘2+2’式”和“‘3+1’或‘1+3’式”，两者都是从音段组合规律上对四字格所进行的划分（鞠君 1995）。

结构构成型四字格的内部结构复杂，在语料库中切分形式不稳定，例如“各负其责”，可能出现“各负　其责”的形式，也可能出现“各　负　其　责”的形式。当四字格被切分成形如“锅 碗 瓢 盆”这样处于同一层次的四个单音节词时，则称之为“四字骈语”（安华林 2001）。词语构成型的四字格内部全部是由词构成，而结构构成型的四字格内部不全是由词构成，它可能是“结构”＋“词”的形式，例如“大饱耳福”；也可能是“结构”＋“结构”的形式，例如“笔简意深”，这种前后结构形式对称的四字格也称为并列式四字格（时秀娟 2001）。

固定结构型，指的是用法固定结构稳定不可变的一些四字格，常见于一些表示转折或条件关系的四字格，例如“不管怎样”、“也就是说”等。

三、语料库中四字格的切分

1. 四字格切分不一致的问题

分词语料库中的切分不一致现象一直是中文信息处理领域的难点。四字格的切分不一致现象是整个分词语料库中分词不一致研究工作的重要组成部分之一。四字格也属于分词单位，冯志伟（2001）认为：“ 四字成语和习惯用语，各成分意义结合紧密，难以拆开，不切分”，但是在实际的分词过程中，四字格往往不被切分成一个完整的分词单位，而是被“切碎”了：例如“倾而不倒”这个四字格，在语料库中既出现过“倾_而_不_倒”这样的切分实例，也出现过“倾_而_不倒”这样的切分实例。切分不一致大大降低了分词的精度，影响了自然语言处理的后续工作。切分不一致问题若得不到较好解决，将会对汉语自动分词、分词规范统一、语料库建设等方面造成影响。

除了成语之外，没有被词典收录的四字惯用语、习语，都可以算做多词表达的一种（刘荣 2011）。多词表达需要从整体上把握多词合成后所表达的意义，而四字格切分不一致的现状会给多词表达的相关研究带来困难。

2. 四字格切分不一致的类型

对于一个四字格而言，只要它在分词语料库中没有被切分成一个四字长的分词单位，就认为它被切碎了。在理论上，切碎了的四字格可以是各种各样的形式，但无论这些四字格的切分不一致结果多么纷繁复杂，最终都可以归结到三种四字格的切分不一致类型中：被切分成两个词（2 型四字格）、被切分成三个词（3 型四字格）、被切分成四个词（4 型四字格）。一个四字长的分词单位，所有可能的切分形式，都包含于这三种切分不一致类型中。

分词语料库中四字格的三种切分不一致类型，和本文前述的从四字格构词模式角度提出的四字格分类体系是对应的。2 型四字格对应的是词语构成型四字格，3 型四字格和 4 型四

字格对应的是结构构成型四字格，其中 3 型四字格对应于结构构成型四字格中的“‘结构’＋‘词’式”。具体的对应关系如图 1：

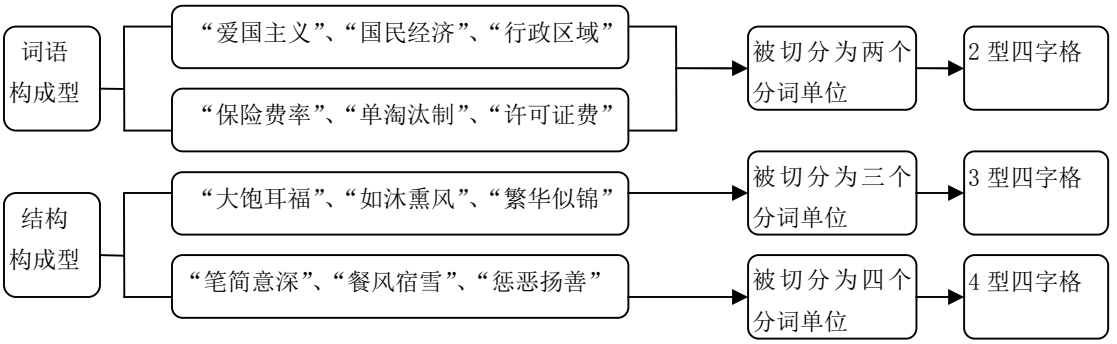

图 1　四字格切分不一致类型的对应关系

3. 语料库中四字格切分不一致的提取实验

四字格切分不一致的提取工作，就是要找出语料库内部和语料库之间某个特定四字格所有不同的切分形式。例如，北大语料库中有这样一个四字格：“惩恶扬善”，在别的语料库中出现了“惩_恶_扬_善”和“惩恶_扬善”两种切分形式，切分不一致提取工作就需要把这两种不同的切分形式都找出来。为了能够发现所有的切分不一致情况，提取算法需要同时检索这六个不同的切分串：“惩恶_扬善”、“惩_恶扬善”、“惩恶扬_善”、“惩_恶_扬善”、“惩恶_扬_善”、“惩_恶_扬_善”。这其实就包含了“惩恶扬善”这个四字格所有可能的切分形式。

(1) 实验方案及数据

本实验考察了三个语料库之间的四字格切分不一致数据，同时也对三个语料库各自内部的四字格切分不一致数据进行了统计。考察了四字格切分不一致的三种类型（2 型、3 型、4 型四字格），以及每种类型的类别数（types）和实例数（tokens）这两个统计量。统计数据如表 2 和表 3，数据示例如表 4。

表 2　分词语料库内部的四字格切分不一致数据

	《人民日报》语料库		微软分词语料库		国家语委分词语料库	
	例数	种数	例数	种数	例数	种数
四字格数量	13082	2830	29933	2739	7729	1999
切分一致数量	12844	2776	29160	2570	7509	1878
切分不一致数量	238	54	773	169	220	121
2 型四字格数量	233	51	751	157	207	111
3 型四字格数量	5	3	19	9	10	7
4 型四字格数量	0	0	3	3	3	3

表 3　分词语料库之间的四字格切分不一致数据

	四字格数量				切分不一致的三种形式					
	切分一致		切分不一致		2 型四字格		3 型四字格		4 型四字格	
	例数	种数	例数	种数	例数	种数	例数	种数	例数	种数
人民日报→微软	16476	1041	5303	855	3569	505	1371	240	363	110
微软→人民日报	8201	960	5991	615	5924	591	59	19	8	5
人民日报→语委	4511	845	711	311	597	239	89	53	25	19
语委→人民日报	7736	806	4355	445	4272	413	77	28	6	4

语委→微软	17942	1085	5455	567	4798	451	537	88	120	28
微软→语委	5331	1062	1600	400	1553	381	44	16	3	3

注:“人民日报➞微软”表示人民日报语料库中的未切碎四字格在微软分词语料库中的切分情况，下同）

表 4　语料库中四字格切分不一致的数据示例

2 型四字格切分不一致示例		3 型四字格切分不一致示例		4 型四字格切分不一致示例	
四字格	次数	四字格	次数	四字格	次数
[生产　总值]	113 次	[科教　兴　国]	108 次	[小　富　即　安]	5 次
[犯罪　分子]	28 次	[被　占　领土]	11 次	[以　假　充　真]	3 次
[房地产　业]	13 次	[科研　院　所]	20 次	[少　生　快　富]	5 次
[物理学　家]	23 次	[招商　引　资]	26 次	[自　收　自　支]	1 次
[产业　部门]	11 次	[各　具　特色]	10 次	[一　动　不　动]	1 次
[组织部　长]	17 次	[阴　转　多云]	63 次	[生　而　知　之]	1 次

（2）实验分析

观察数据可以看出，三个语料库内部的四字格切分不一致数量要远少于三个语料库之间的四字格切分不一致数量。例如，国家语委语料库内部的四字格切分不一致例数仅有 220 例，而语委语料库和微软语料库之间的切分不一致例数却多达 7055 例。这也和直觉相符：一个语料库内部的切分方式和切分原则相对稳定，而不同语料库之间的切分方式和切分原则差异较大。

在进行分词语料库之间四字格切分不一致的比较时，四字格切分一致数量和切分不一致数量的比例是衡量两个语料库之间四字格切分相似程度的一个重要依据。通过计算“四字格切分不一致数量/四字格切分一致数量”的值可以发现，《人民日报》语料库和国家语委语料库之间的这个比值为（311+445）/（845+806）= 45.8%，是所有语料库之间的最低值，说明这两个语料库之间的四字格切分方式最接近；《人民日报》语料库和微软亚洲研究院语料库之间的这个比值为（855+615）/（1041+960）= 73.5%，是所有语料库之间的最高值，说明这两个语料库之间在四字格的切分问题上“分歧”最多。

四、语料库中四字格的识别

分词语料库中的四字格分为两类：一类是未被切碎、可以利用词长或者词性信息直接找到的、形如“_茶饭不思_”形式的四字格；另一类，是被切碎成若干更小长度的分词单位、无法在语料库直接找到、形如“_如_诗_如_画_”形式的四字格。四字格识别研究所要针对的正是第二类四字格。

被切碎了的四字格又分为两种，一种是在某个语料库中被切碎，但在其他语料库中未被切碎的四字格；另一种是在各个语料库中都被切碎，但确实应当被切分为一个分词单位的四字格。对于前者的识别，只需通过建立一个简单的四字格实例词表即可实现；而对于后者，我们无法在语料库中找到匹配实例，要识别出这些四字格，就必须要借助于统计模型来训练大量数据、机器学习四字格的结构特征并以此对四字格进行自动标注。

1. CRF 的训练语料获取

条件随机场(Conditional Random Field ,CRF)是一种用于在给定输入结点值时计算指定输出结点值的条件概率的无向图模型，是一个基于统计的序列标注和分割的方法(John Lafferty 2001)。目前，CRF 广泛应用于自然语言处理的各个方面，特别是在序列化标注例如词性标注任务中，CRF 表现优异。我们可以把四字格的识别过程想像为一种特殊的词性标注：给每个分词单位一个标记，该标记用于表明分词单位是或者不是四字格的成分之一。把连续出现的有四字格标记的分词单位找出来，当它们的词长相加正好为四的时候，就可以认

为这是一个四字格。

CRF 模型需要大量做过人工标注的语料用于训练。单靠人力去发现语料库中被切碎了的四字格并对其进行四字格信息的标注，是一项极其耗时耗力的工作。一种可行的方法是，利用不同语料库之间四字格的切分不一致数据来实现四字格训练语料的自动获取。例如有语料库 A 和 B，在语料库 A 中可以找到四字格“为国分忧”，而它在语料库 B 中则被切成了“为_国_分忧”，那么“为_国_分忧”这个切分实例就可以成为 CRF 模型训练语料的一部分。采用这种思路来自动获取训练语料，上文关于四字格切分不一致的统计结果就可以直接为 CRF 模型的训练过程提供大量数据。

俞士汶（2002）提出，形如“调查_研究”、“总结_经验”这样的“四个字短语，通常应切分”。2 型四字格均由两个词语构成，结构方式上趋近于词组，在语料库中的切分不一致情况多属于切分粗细的问题，而非切分正误的问题。因此，本研究的四字格识别工作只把 3 型四字格和 4 型四字格作为识别对象。去除 2 型四字格的切分不一致结果，只保留切分成 3 型和 4 型四字格的切分不一致结果，本文研究选取的三个分词语料库中共有 2742 例四字格切分不一致数据。如表 5。

表 5　切分成 3 型和 4 型的四字格切分不一致数据（例数）

	《人民日报》语料库	微软语料库	国家语委语料库
《人民日报》语料库	5	1734	114
微软语料库	67	22	47
国家语委语料库	83	657	13

除了四字格本身，还需要提供四字格的上下文语境信息用于 CRF 模型训练，即真正用于 CRF 模型训练的，是包含了四字格的句子，而不是单独的四字格本身。对应于 2742 个四字格切分不一致实例，用于 CRF 模型训练的包含四字格的句子也是 2742 个。

2. CRF 的特征列和模板定制

按照 CRF 模型的语料格式要求，针对 3 型四字格和 4 型四字格这两种识别对象，给出两种四字格特征列，如图 2 所示。图中的第一列是文本，第二列是词性标记和词长信息，第三列是四字格标记。“none”表示该词不是四字格成分，“head_3”、“body_3”、“tail_3”或“head_4”、“body1_4、body2_4”、“tail_4”分别表示该词是四字格成分的首部、中部、尾部，后面的数字“3”或“4”表示该四字格被切碎成了 3 个或者 4 个部分。

```
懂得    v-two    none          激越    a-two    none
了      y-one    none          的      u-one    none
为      v-one    head_3        潮      n-one    head_4
国      n-one    body_3        涨      v-one    body1_4
分忧    v-two    tail_3        潮      n-one    body2_4
                               落      v-one    tail_4
```

图 2　四字格的特征列

CRF 模型是一个通用工具，用户需要定制自己的特征模板。模板的基本格式为%x[行、列]，它用于确定输入数据中的一个词例。行，表示%x 相对于当前词例的行数；列，表示%x 在列上的绝对列数。以训练语料中的“阳光/n　走遍/v　它/r　不/d　为/p　人/n　知/v　的/u　另/r　一/m　面/n”这句话来示例特征模板，假设当前词为“为”，本研究所采用的 CRF 特征训练模板如图 3。

训练语料			“为”字的特征模板	
阳光	n-two	none	U00:%x[-2,0]	//它
走遍	v-two	none	U01:%x[-1,0]	//不
它	r-one	none	U02:%x[0,0]	//为
不	d-one	head_4	U03:%x[1,0]	//人
为	p-one	body1_4	U04:%x[2,0]	//知
人	n-one	body2_4	U10:%x[-2,1]	// r-one
知	v-one	tail_4	U11:%x[-1,1]	// d-one
的	u-one	none	U12:%x[0,1]	// p-one
另	r-one	none	U13:%x[1,1]	// n-one
一	m-one	none	U14:%x[2,1]	// v-one
面	n-one	none		

图 3　特征模板示例

3. 基于 CRF 模型的四字格识别实验

基于 CRF 模型的识别实验采用的训练语料是北大《人民日报》98 年 1 月语料、微软亚洲研究院中文分词语料、中国国家语委分词语料。封闭测试语料选用的是北大《人民日报》98 年 1 月语料，开放测试语料选用的是北大《人民日报》98 年 2-6 月语料，实验结果如表 6 和表 7。

表 6　9 8 年 1 月至 6 月北大《人民日报》语料库中的四字识别结果

		正确识别个数	错误识别个数	识别正确率
封闭测试	98.1《人民日报》	373	24	93.9%
开放测试	98.2《人民日报》	396	22	94.7%
	98.3《人民日报》	353	25	93.4%
	98.4《人民日报》	428	30	93.4%
	98.5《人民日报》	366	22	94.3%
	98.6《人民日报》	313	16	95.1%

表 7　9 8 年 1 月至 6 月北大《人民日报》语料库识别结果中的 3、4 型四字格分布

	3 型四字格		4 型四字格	
	正确识别数量	比例	正确识别数量	比例
98.1《人民日报》	244	65.4%	129	34.6%
98.2《人民日报》	228	57.6%	168	42.4%
98.3《人民日报》	224	63.5%	129	36.5%
98.4《人民日报》	240	56.1%	188	43.9%
98.5《人民日报》	218	59.6%	148	40.4%
98.6《人民日报》	172	54.9%	141	45.1%

作为开放测试语料，98 年 2 月至 6 月的《人民日报》语料并未参与到四字格切分不一致数据的获取过程中，但利用 CRF 模型对其进行的四字格识别实验仍然取得了 93%以上的正确率，甚至部分超过了封闭测试的效果；表 7 列出了半年《人民日报》语料中识别出的 3、4 型四字格分布，可以看出识别出的被切碎四字格的数量较多、分布平均，3 型四字格和 4 型四字格之间的比例也趋于平衡没有明显的偏重，这些都表明了，利用四字格切分不一致数据并辅以 CRF 模型来识别四字格的方法能行之有效的解决多语料库中四字格的识别难题。

识别得到的四字格中，有些可以在北大、国家语委、微软三个训练语料库中找到完全匹配的实例，例如：

持续不断	从头做起
催人泪下	多党合作
凡此种种	蜂拥而来
干旱少雨	公布于众

但这些四字格只占识别得到的 2229 例四字格中很小的一部分，占比最多的国家语委语料库也只有 3.8%；绝大多数识别得到的四字格在三个训练语料库中都找不到与之完全匹配的实例，换言之都被切碎了，例如：

冰封雪盖	官去室空
船推浪移	肩扛手提
车毁人伤	肩扛背驮
房倒屋塌	以短养长

这部分识别出的四字格是无法通过查找匹配实例来识别的，也进一步验证了本研究引入统计模型参与识别工作的必要性。相关数据详见表 8。

表 8　四字格识别结果在三个语料库中是否有匹配实例的分布数据

	语料库中有匹配实例		语料库中无匹配实例	
	数量	比例	数量	比例
《人民日报》语料库	36	1.6%	2193	98.4%
国家语委语料库	85	3.8%	2144	96.2%
微软语料库	19	0.9%	2210	99.1%

五、结语

本文的研究对象是语料库中的四字格，本文着重对“四字格的分类”、“ 四字格的切分”、“ 四字格的识别”这三个问题进行了深入研究。如图 4：

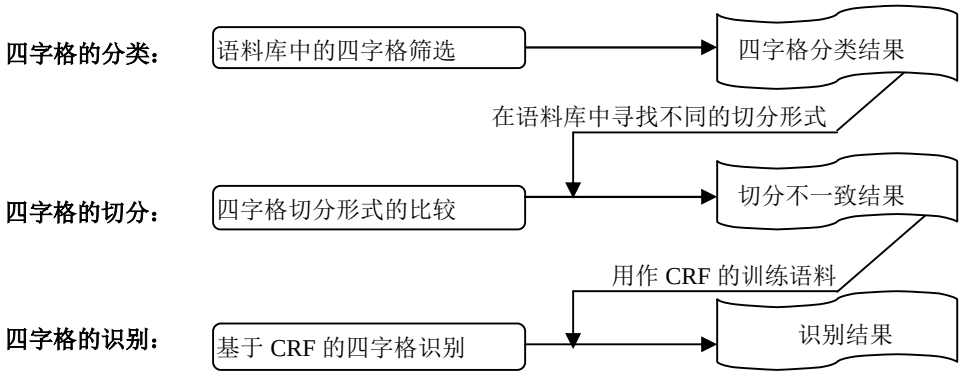

图 4　四字格研究框架

通过抽取不同分词语料库中的四字格并进行筛选、分类，本文解决了第一个问题；利用筛选、分类的结果，在不同语料库中寻找它们所有的切分形式，分析归纳这些四字格切分形式上的特点，本文解决了第二个问题；利用不同语料库间四字格的切分不一致结果，在分词语料库中实现了对四字格的识别工作，至此本文解决了第三个问题。通过实验表明，多语料库中的四字格识别正确率可以达到 93%以上。

本研究依然有许多尚未完成或亟需改善之处：四字格分类体系仍显粗糙；研究所使用的分词语料库规模有待扩大；四字格资源需要继续补充和完善。下一步工作考虑引入语法、语义方面的知识来进一步提高四字格特别是可派生型四字格的识别效果。

参考文献：

安华林:2001:《“四字骈语”初探》,《信阳师范学院学报》第 1 期。

冯志伟:2001:《确定切词单位的某些非语法因素》,《中文信息学报》第 5 期。

鞠君:1995:《四字格中“1+3”音段和“3+1”音段组合规律初探》,《汉语学习》第 1 期。

刘荣、王弈凯:2011:《利用统计量和语言学规则提取多字词表达》,《太原理工大学学报》第 2 期。

陆志韦:1956:《汉语的并立四字格》,《语文研究》第 1 期。

马国凡:1987:《四字格论》,《内蒙古师大学报》第 3 期。

莫彭龄:2003:《“四字格”与成语修辞》,《常州工学院学报》第 3 期。

时秀娟:2001:《浅析汉语并列式四字格结构及其理据性》,《莱阳农学院学报》第 3 期。

苏新春:2007:《国家语委_通用语料库_核心库_的词表提取及词汇构成分析》,《江苏大学学报(社会科学版)》第 1 期。

徐润华、陈小荷:2010:《分词语料库中并列式四字格的识别研究》,《计算机工程与应用》第 4 期。

杨晓黎:1996:《四音节新词语及其成因》,《江淮论坛》第 4 期。

俞士汶、段慧明、朱学锋、孙斌:2002:《北京大学现代汉语语料库基本加工规范》,《中文信息学报》第 5 期。

周荐:1997:《论四字语和三字语》,《语文研究》第 4 期。

Ido, D. Shaul, M. , 1993, “Contextual Word Similarity and Estimation from Sparse Data”, *In Proceedings of the 31st Annual Meeting of the Association for Computational Linguistic*, pp.164—171.

Gina-Anne, L. , 2006, “Word Segmentation and Named Entity Recognition”, *In Proceedings of the third international Chinese language processing bakeoff*, pp.108—117.

Hans, V. H. Walter, D. Jakub, Z. , 2001, “Improving accuracy in word class tagging through the combination of machine learning systems”, *Computational Linguistic*, vol.27, no. 2, pp.199—229.

Lafferty, J. McCallum, A. Pereira, F. , 2001, “Conditional random fields: Probabilistic models for segmenting and labeling sequence data”, *In Proceedings of the 18th International Conference on Machine Learning (ICML 2001)*, pp.282—289.