分词语料库中的并列式四字格识别

徐润华,陈小荷,李 斌
XU Run-hua,CHEN Xiao-he,LI Bin

南京师范大学 文学院,南京 210097
College of Liberal Arts,Nanjing Normal University,Nanjing 210097,China
E-mail:runhuaxu@hotmail.com

XU Run-hua,CHEN Xiao-he,LI Bin.Recognition of parallel four-character idioms in word-segmented corpora.Computer Engineering and Applications,2010,46(4):139-141.

Abstract: Among all kinds of Chinese four-character idioms,the Parallel Four-Character Idiom (PFCI) is special and numerous. This paper introduces the research based on Conditional Random Fields (CRF) model which can retrieve PFCI from a POS-tagged corpus.The paper then analyzes the structural characteristics of PFCI and proposes an approach on recognizing PFCI in word-segmented corpora.By comparing its application on different corpora,the evaluation results show that this recognition approach maintains relatively high precision and good adaptability.

Key words: four-character idioms;conditional random fields;segmented fragments;parallel four-character idioms

摘 要: 并列式四字格是一种特殊却数量众多的四字格。介绍了在有词性标注语料库中基于条件随机场模型的四字格抽取工作,并在此基础上分析了并列式四字格的结构特点,提出了一种基于分词语料库环境的并列式四字格识别方法。通过不同语料库间的对比实验,结果表明该识别方法具有比较好的精确度和一定的适应性。

关键词: 四字格;条件随机场模型;分词碎片;并列式四字格
DOI: 10.3778/j.issn.1002-8331.2010.04.045 **文章编号:** 1002-8331(2010)04-0139-03 **文献标识码:** A **中图分类号:** TP391.1

1 四字格和语料库中的四字格

“四字格”,顾名思义是指由四个汉字组成的一种语言格式。“四字格”这个术语最早源于陆志韦先生1956年的《汉语的并立四字格》一文。在汉语言发展的历史中,四字格的形式起着非常重要的作用。“四字格”结构灵活多变,结构的能产性和派生性都很强,到了今天,现代汉语词汇系统不断发展,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,汉语的四字格,特别是并列式四字格结构的数量有增无减。由于四字格结构的派生性、动态性,使得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。

分词语料库中的四字格有两种存在形式:一种是已经进行了正确的切分、由四个汉字组成的分词单位,另一种是没有被切分正确、内部被切碎成若干个更小分词单位的词串。对于第一种形式,语料库本身已经对四字格进行了正确的切分,可以把它们称之为“已知四字格”,虽然四个汉字组成的分词单位还包括数字词和一些命名实体,但从中区分出四字格并不是一件难事;而第二种形式,对于那些“未知四字格”,如何发现这些被“切碎”了的四字格,如何识别出它们,才是分词语料库中四字格识别工作的重点所在。

2 基于crf模型和有词性标注语料库的四字格抽取

2.1 crf模型中四字格的特征列

crf模型,即条件随机场模型,是一个可定制的开源代码模型,目前广泛应用于自然语言处理的各类任务中。在用于切分和标注序列数据的时候,条件随机场总是可以取得非常好的效果,因此,分词及词性标注是条件随机场最常见的应用领域之一。可以把四字格的抽取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。接下来,只需要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4的时候,就可以认为这是一个四字格。如图1。

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”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”表示该四字格被切碎成了3个部分。之所以添加了第二列的词性标记信息到crf模型中去,是因为单靠文本自身和四字格标记这两列数据,很难让crf模型学习到更

基金项目: 国家社会科学基金(the National Social Science Foundation of China under Grant No.07BYY050);国家重点基础研究发展规划(973)(the National Grand Fundamental Research 973 Program of China under Grant No.2004CB318102)。

作者简介: 徐润华(1982-),硕士生,主要研究方向:语料库检索和中文自动分词;陈小荷(1952-),教授,博士生导师,主要研究方向:中文分词、语义搭配、句法分析等;李斌(1980-),博士生,主要研究方向:语义搭配、信息检索等。

收稿日期: 2008-09-12 **修回日期:** 2008-12-11

```
……
懂得 v-two none
了 y-one none
为 v-one head_3
国 n-one body_3
分忧 v-two tail_3
……
```

图 1 四字格的特征列

多的信息,为了让四字格的抽取结果更理想,把 crf 模型的抽取工作和有词性标注语料库紧密联系在一起了。

2.2 获取 crf 模型的训练语料

要让 crf 模型对语料进行四字格信息的标注,就必须先提供大量已做过正确四字格标记的语料用于 crf 模型训练。但在实际操作中,如果只依靠人力去发现语料库中被切碎了的“未知四字格”,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信息来实现“未知四字格”的自动获取。例如有语料库 A 和语料库 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”以及它所在的句子就可以成为 crf 模型训练语料的一部分。这种方法的高效性在于,语料库间的切分不一致结果都是可以通过程序自动实现和获取的。

文中选取了北大人民日报一月语料、微软亚洲研究院中文分词语料、中国国家语委分词语料 3 个语料库间的四字格切分不一致结果作为 crf 模型的训练语料,四字格切分不一致结果共 1 677 例。

2.3 crf 模型的四字格抽取结果

表 1 1998 年 1 月至 6 月北大人民日报语料库中的四字格抽取结果

	四字格正确识别个数	四字格错误识别个数	四字格识别正确率/(%)
1998.1 人民日报语料	373	24	93.9
1998.2 人民日报语料	396	22	94.7
1998.3 人民日报语料	353	25	93.4
1998.4 人民日报语料	428	30	93.4
1998.5 人民日报语料	366	22	94.3
1998.6 人民日报语料	313	16	95.1

3 分词语料库中的并列式四字格识别

3.1 分词碎片中的并列式四字格

并列式四字格,是四字格的一种特殊形式,例如“潮涨潮落”,“潮涨”和“潮落”的结构和组成方式相似,所以把这类四字格称为并列式四字格。语料库中的“未知”并列式四字格,其直观的特征就是被切碎成了 4 个连续的单字,这恰好和分词碎片的特点相契合。分词碎片特指那些在分词过程中被切分成连续的单个汉字的串,因此,对分词语料库中并列式四字格的识别也是处理分词碎片工作的一部分。分词碎片的处理是分词语料库建设中无法回避的一个难题,有分词就会有分词碎片,分词碎片处理地好坏直接影响到分词工作的效率和精度。该研究的目标是识别出分词碎片中的并列式四字格,旨在为分词碎片的处理工作减轻负担。

3.2 并列式四字格识别策略

并列式四字格可以看做形如 $ABA'B'$ 的结构,其中 AB 和 $A'B'$ 并列,即它们的结构、内在组成方式相似;但是从识别的角度出发,$AB\rightarrow A'B'$ 的关系并不直观。因此把 $AB\rightarrow A'B'$ 的模式扩展到 $A\rightarrow A'$、$B\rightarrow B'$ 的模式,把关注的重心从结构的并列转移到成分的相似上来。对于并列式四字格,认为成分 A 和成分 A' 的词性要一致、成分 B 和成分 B' 的词性要一致。此外,成分 A 和 A' 的用字以及成分 B 和 B' 的用字都要做相应的限制,以此达到用成分的相似来描述结构的并列的目的。

除了成分相似外,并列式四字格还有一个明显的特征,就是大量固定结构的运用。例如“敌进我伏、敌围我散、敌击我隐”的 $AxA'x$ 式、“手起刀落、潮涨潮落、忽起忽落”的 $xBxB'$ 式,其中“敌……我……”和“……起……落”都是固定结构。找到并提取出这些固定结构,就可以用它来识别更多“未知”的并列式四字格。

根据识别策略的不同,并列式四字格的识别模块分为两个:依据成分相似策略的识别模块 1 和依据固定结构策略的识别模块 2。在分别实现两个识别模块的基础上,又将两个模块组合在一起进行了识别实验:满足识别模块 1 或识别模块 2 均视为并列式四字格。

3.3 并列式四字格识别流程

步骤 1 获取并列式四字格语料。在基于 crf 模型的四字格抽取工作的基础上,利用词性序列特征,从 1998 年上半年的人民日报语料中提取出 439 个无重复的并列式四字格。由于 crf 模型基于有词性标注语料库,提取出的 439 个并列式四字格中的每个汉字都带有词性信息。

步骤 2 抽取用字信息和结构信息。并列式四字格由 4 个连续的汉字组成,这 4 个汉字分别处于四字格的 1、2、3、4 位置上。抽取 439 个并列式四字格的 1、3 位置和 2、4 位置的用字及词性信息,去除重复的条目,得到 1_3 位置用字字表和 2_4 位置用字字表。其中 1_3 位置用字字表包括 4 个形容词性单字、16 个副词性单字、2 个方位词性单字、12 个数词性单字、56 个名词性单字、4 个代词性单字、243 个动词性单字。2_4 位置用字字表包括 61 个形容词性单字、4 个数词性单字、216 个名词性单字、2 个量词性单字、93 个动词性单字。把 439 个并列式四字格两两进行比较,如果两个四字格的 1、3 位置均相同,则把 1、3 位置两个汉字组成的结构加入 1_3 位置固定结构表中;如果两个四字格的 2、4 位置均相同,则把 2、4 位置两个汉字组成的结构加入 2_4 位置固定结构表中。去除重复的条目后,最终得到含 36 条结构的 1_3 位置固定结构表和含 33 条结构的 2_4 位置固定结构表。其中出现次数最多的结构是“一……一……”,共出现 1 108 次,结构出现次数最少为 2 次。

步骤 3 抽取分词碎片。把分词语料库中连续单字组成的最长串全部抽取出来。该研究选用的分词语料是微软亚洲研究院中文分词语料和北大人民日报分词语料(1998 年 1 月)。

步骤 4 分别按照识别模块 1、识别模块 2 以及两者的组合进行识别。识别模块 1:处理连续 4 个汉字,如果 1、3 位置的汉字均出现在 1_3 位置用字字表中并且词性相同,同时 2、4 位置的汉字均出现在 2_4 位置用字字表中并且词性相同,则认为该连续汉字串为并列式四字格。识别模块 2:处理连续 4 个汉字,如果 1、3 位置两个汉字组成的结构出现在 1_3 位置固定结构表中,或者 2、4 位置两个汉字组成的结构出现在 2_4 位置固定结构表中,则认为该连续汉字串为并列式四字格。识别模块 1、2 的组合:处理连续 4 个汉字,如果满足识别模块 1 或者识别模块 2,则认为该连续汉字串为并列式四字格。

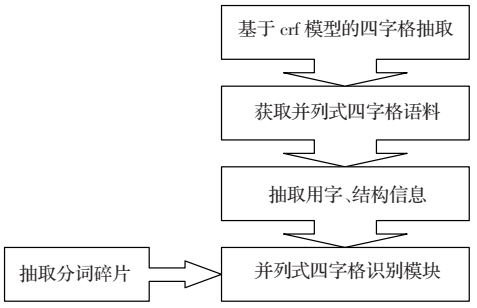

图 2　并列式四字格识别流程图

3.4　实验数据

表 2　分词语料库中的并列式四字格识别实验

	北大人民日报分词语料(1998 年 1 月)			微软亚洲研究院中文分词语料		
	正确识别数量	错误识别数量	正确率/(%)	正确识别数量	错误识别数量	正确率/(%)
识别模块 1	162	27	85.7	304	31	90.7
识别模块 2	173	30	85.2	413	34	92.4
模块 1、2 的组合	264	51	83.8	567	58	90.7

4　结语

从实验结果中可以看出,两者组合的时候,识别数量大致相当于两个模块识别数量之和,但并不相等,而是略少一些。这说明识别模块 1 和识别模块 2 所识别出的并列式四字格基本不重复,这两种识别策略之间具有一定的独立性。略少于两个模块识别数量之和的那部分,就是两个模块识别结果的交集部分。总体而言,各个模块的识别正确率较高,但在 3 个识别实验中,人民日报语料的正确率均低于微软中文语料。识别过程利用的用字信息和结构信息,都是从人民日报语料中抽取出来的,所以用人民日报语料做实验,应该算做封闭测试;但开放测试的结果优于封闭测试,也说明识别方法本身具有比较好的适应性。

该研究提出的并列式四字格识别方法,基于的环境是分词语料库。区别于上文提到的基于 crf 模型的四字格抽取工作,分词语料库中的并列式四字格识别并不依赖词性标记信息。从解决分词碎片问题的角度出发,该研究在分词语料库中进行了并列式四字格的识别尝试,下一步的工作是将识别环境继续延伸到未分词的语料中去。

参考文献:

[1] Xue Nian-wen.Chinese word segmentation as character tagging[J].Computational Linguistics and Chinese Language Processing,2003,8(1):29-48.

[2] Zhao Hai,Huang Chang-ning.An improved Chinese word segmentation system with conditional random[C]//Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing,Association for Computational Linguistics,Sydney,2006:162-165.

[3] 陈小荷.现代汉语自动分析——Visual C++实现[M].北京:北京语言文化大学出版社,1998.

[4] 陆志韦.汉语的并立四字格[J].语文研究,1956(1).

[5] 马国凡.四字格结构的模糊性[J].内蒙古师大学报,1989(3).

[6] 黄昌宁,李涓子.语料库语言学[M].北京:商务印书馆,2002.

[7] 吕叔湘.汉语语法分析问题[M].北京:商务印书馆,1979.

(上接 128 页)

[6] Chung S Y,Forney G D,Richardson T J,et al.On the design of low density parity check codes within 0.0045dB of the Shannon limit[J].IEEE Commun Lett,2001,5(2):58-60.

[7] Fossorier M P C.Quasi-cyclic low-density parity-check codes from circulant permutation matrices[J].IEEE Trans Inform Theory,2004,50(8):1788-1793.

[8] Kou Y,Lin S,Fossorier M.Low density parity check codes based on finite geometries:A rediscovery and new results[J].IEEE Trans Inform Theory,2001,47(7):2711-2736.

[9] Honary B,Moinian A,Ammar B.Construction of well-structured quasi-cyclic low-density parity check codes[J].IEE Proc-Commun,2005,152(6):1081-1085.

[10] Lan Lan,Ying Yu Tai,Shu Lin,et al.New constructions quasi-cyclic LDPC codes based on two classes of balanced incomplete block designs:for AWGN and binary erasure channels AAECC[C]//LNCS 3857,Berlin Heidelberg:Springer-Verlag,2006:275-284.

[11] Lan Lan,Yin Yu Tai,Behshad M,et al.New constructions quasi-cyclic LDPC codes based on special classes of BIBD's for the AWGN and binary erasure channels[J].IEEE Trans on Commun,2008,56(1):39-48.

[12] 靳蕃,陈志.组合编码原理及应用[M].上海:上海科学技术出版社,1995:257-258.

[13] Van Lint J H,Wilson R M.A course in combinatorics [M].2 版.北京:机械工业出版社,2004.

[14] Neal R M,Mackay D J C.LDPC-2006-08[EB/OL].[2009].http://www.cs.toronto.edu/~radford/ftp/LDPC-2006-02-08/index.html.

(上接 133 页)

了权重系数,提高了分类效果,该算法有待改进之处是提出不同的权重公式来进一步提高分类效率,特别对于类数较高的数据集,更是有待于提出一种新的权重系数。

参考文献:

[1] Yager R R.An extension of the Naive Bayesian classifier[J].Information Sciences,2006,176:577-588.

[2] 程克非,张聪.基于特征加权的朴素贝叶斯分类器[J].计算机仿真,2006,23(10).

[3] 章舜仲,王树梅,黄河燕,等.基于属性相关性分析的贝叶斯模型[J].情商学报,2007,24(2):58-65.

[4] Newman D J,Hettich S,Blake C L,et al.UCI repository of machine learning database,1998.