基于规则的现代汉语连词用法自动识别研究*

周丽娟，张坤丽，袁应成，昝红英

郑州大学信息工程学院，河南 郑州 450001

E_mail: zhljuan1018@126.com ; ieklzhang@zzu.edu.cn ; luck_yuan888@163.com ; iehyzan@zzu.edu.cn

摘要： 连词能够连接词语、句子乃至篇章，具有特殊的连接功能，用法复杂多样。目前已有的连词研究成果大都是面向人用的，对连词用法的描述难以避免主观性和模糊性，很难直接应用于自然语言处理领域。本文采用“三位一体”的构建现代汉语广义虚词知识库的思想，给出了连词用法词典和用法规则库，并针对连词的连接功能，对连词用法的自动识别做出了不同于其他虚词的特殊处理。实验结果表明，这种基于规则的方法能够较好地用于连词用法的自动识别。

关键词： 自然语言处理；用法词典；用法规则库；连词自动识别

The Studies on Automatic Recognition of Rule-based Modern Chinese Conjunction's Usages

Zhou Lijuan,Zhang Kunli,Yuan Yingcheng,Zan Hongying

College of Information Engineering, Zhengzhou University, Zhengzhou, Henan 450001, China

E_mail: zhljuan1018@126.com ; ieklzhang@zzu.edu.cn ; luck_yuan888@163.com ; iehyzan@zzu.edu.cn

Abstract: The conjunctions can connect the words, sentences and even chapters, which have the special connection function, the usage are complex and diverse. At present, the studies on conjunctions are mostly served for people. These descriptions can not avoid from subjectivity and illegibility. So they are not easy to be applied directly to natural language processing. This paper adopts the idea of the "Triune" knowledge about constructing contemporary Chinese functional word knowledge base and introduces usage dictionary and usage rule of conjunctions. Aiming at the connection function of conjunctions,the paper makes the processing which is different from the other functional words for the automatic recognition of usage of conjunctions. Experimental results show that the rule-based method can be well used for the automatic recognition of usage of conjunctions.

Key words: natural language processing; usage dictionary; usage rule; conjunction automatic recognition

1 引言

汉语的词类问题一直是汉语语法学界长期争论的难题。连词作为一种特殊的词类，数量虽然并不是很多，但是它的功能和用法复杂多样，具有极强的个性，而且使用范围广、频率高，尤其是它能够连接词语、句子乃至篇章，表达细致缜密的逻辑语义关系[1]，在现代汉语语法中具有不可忽视的重要地位。

连词的用法表示代表某种含义的连词可以用在什么地方。同一个连词，在不同的上下文语境中可能表示不同的含义，具有不同的用法。只有全面考察不同的用法，才能更好地理解语义信息，实现自然语言处理。在大规模的语料中，人工地判别用法费时又费力，而用法自动识别是让

*本文承国家自然科学基金项目（项目号 60970083）、北京大学计算语言学教育部重点实验室开放课题基金（课题编号KLCL-1004）和河南省科技创新人才杰出青年基金项目（项目号 104100510026）的资助。

机器来自动标注用法，它的研究有助于现代汉语篇章的分析。通过对连词用法的自动识别，可以更好地进行褒贬新词的自动发现，从而有效地推进褒贬评价工作。另外，鉴于连词的连接功能，连词用法的自动识别会涉及到多个句子，不同于其他词的只涉及一个句子的处理。因此，连词用法自动识别的研究具有重要意义。

本文具体章节安排如下：第二节介绍连词用法识别的相关研究；第三节介绍连词用法词典和用法规则库；第四节介绍了基于规则的连词用法自动识别的实验；第五节给出了实验结果，并对结果进行分析；最后给出本文的总结，并对今后的工作进行展望。

2 相关研究

连词是虚词中主要的一类，综合多年来汉语语法学界的研究成果，连词的研究主要包括：连词的范畴及分类研究；结合单句的连词研究；结合复句、句群、篇章的连词研究；连词分布及共现研究；连词的个案研究；连词的语法化研究。其中，比较著名的有：周刚的《连词与相关问题》[1]、黎锦熙的《新著国语文法》[2]等。连词的研究虽然硕果累累，但是在连词用法方面的研究还很少，并且这些研究大都是面向人用的。

近年来，面向机器的研究也逐步开始。俞士汶[3]最早提出了“三位一体”构建现代汉语广义虚词知识库的思想,并将广义虚词界定为副词、介词、连词、助词、语气词和方位词；刘云[4]构建了汉语虚词词典的基本框架，为副词、介词、连词、助词和语气词等设计了相应的描述属性，对常用虚词进行了归类总结；昝红英等[5]构建了现代汉语广义虚词知识库，包括虚词用法词典、虚词用法规则库和虚词用法标注语料库；刘锐等[6]初步探讨了基于规则的副词用法的自动标注；张军珲[7]研究了基于统计的常用汉语副词用法的自动识别。这些面向机器的研究主要是针对虚词知识库的构建或者副词用法的自动识别，而对于具体的连词用法的面向机器的研究，如连词用法的自动识别还不多。

本文采用“三位一体”的构建现代汉语广义虚词知识库的思想，按照“用法——属性特征”的对应原则构建连词用法词典；以有序的 BNF 范式进行用法的形式化描述，形成连词用法规则库；最后在标注语料上考察在真实语言环境中的连词用法并进行自动标注。

3 连词的形式化描述

3.1 连词用法词典

连词用法词典以用法——属性特征为对应关系，利用关系数据库的形式建立用法词典，详细描述了每个连词用法的具体属性特征。连词用法词典包括的属性信息主要分为以下 4 组：

标志性属性信息：用法编码、词条、释义、例句、全拼等；

句法修饰功能描述属性信息：粘着、连体、连谓、连主谓、连句、连句群等；

语法意义描述属性信息：连词小类、文体等；

用法描述属性信息：用法描述、单句使用、前连用、前合用、后连用、后合用、句首、句末等。

需要说明的是用法编码具有唯一性，编码规律为：c_全拼[_tn][_i][j]。其中，“c”是连词的

标志，“tn”（n 为数字，标明序号）用于同音不同形词汇的编码，“i”(1,2,3,...)为不同义项的编号，“j”(a,b,c,d,...)为不同用法的编号，“[]”表示这项为可选的。如：连词“并”有两个义项，第一个义项有三种用法，则“并”的用法编码分别表示为：c_bing4_1a 、c_bing4_1b、c_bing4_1c、c_bing4_2。表 1 是含有“并”的部分属性的用法词典样例，其中，“释义”表示义项信息，相同的释义可能具有不同的用法。

连词用法词典涵盖了《现代汉语语法信息词典》[8]中全部的连词，以及在《现代汉语虚词词典》[9]、《现代汉语词典》[10] 、《现代汉语八百词》[11]中出现的连词，并根据北京大学计算语言学研究所提供的 1998 年 1 月份的《人民日报》分词与词性标注语料库进行了词条的调整及用法的总结与修改。目前连词用法词典共收录了 315 个连词，用法共计 695 个，其词条与用法分布如表 2 所示。

表 1　含有“并”的部分属性的用法词典样例†

用法编码	释义	用法	例句
c_bing4_1a	表示并列关系。<x>	连接形容词或动词性短语，动词必须是双音节的。<x><z>	根据这一规定，不能要求同时提供保证人~交纳保证金。<r>|他迅速地~准确地回答了问题<b>
c_bing4_1b	表示并列关系。<x>	连接名词性词语，书面语色彩较浓。<x>	上午寄出一包书~一封信<x>|小王拿起一张纸~一支笔，走到小张面前<x>
c_bing4_1c	表示并列关系。<x>	连接副词或介宾短语。<x><z>	而这些付出，已经~将继续造福于西部人民。<r>|我本人向阁下~通过阁下向贵国人民表示热烈的祝贺<x>
c_bing4_2	表递进。意思上比前一小句进了一层。<x>	连接小句或句子，后一句的主语必须承前省略，意思比前一句进了一层。<x>	那年我正在福州教书，~有幸结识了一位神聊大师<x>|我们在母爱中成长身体，感受爱，也学会了爱。~在这种爱的提示和牵引中，懂得应该去爱他人，爱社会。<r>

表 2　连词词条与用法的分布情况

用法个数	1	2	3	4	5	6	7	8	9	10	词条共计	用法共计
词条个数	155	51	55	24	17	7	3	0	1	2	315	695

3.2　连词用法规则库

在连词用法词典的基础上，我们对各个连词的用法在词典例句语料上进行了考察，抽取其中具有可操作性的判断条件特征，以有序的BNF范式进行连词用法的规则描述，构建连词用法规则库，从而为现代汉语连词用法的自动识别提供形式化的依据。目前抽取的主要用法特征有：句首F、前合用M、前连用L、后连用R、后合用N、句末E。另外在用法规则库中引入了其他一些符号，如A表示同词同词性，B表示同词性不同词，其他符号的表示详见袁应成等[12]，用法规则的描述详见昝红英等 [5]。根据 3.1 节中“并”的用法描述，“并”的用法规则可表示为：

† 表 2 中<>表示来源，x表示《现代汉语虚词词典》，b表示《现代汉语八百词》，r表示《人民日报》语料，z表示自定义。例句中的“~”表示本词。

$并

@<c_bing4_1c>→B*~B ^B→d|p

@<c_bing4_1b>→B~*B ^B→n

@<c_bing4_2>→F ^F→~

@<c_bing4_1a>→MN ^M→<v_d>|a ^N→<v_d>|a

4 基于规则的连词用法自动识别

连词用法的自动识别是在已开发的基于规则的虚词用法自动标注系统的基础上进行的。这个虚词用法自动标注系统的具体步骤是：

(1)、初始化标注语料、用法规则库，为方便大规模的语料自动标注，读取语料时将语料文本内容按“。”、“？”、“！”以及换行符作为截句标志，将标注语料切分成一个个整句，以动态数组的形式读入内存，用法规则以哈希表的形式写入内存。

(2)、读取待标注的整句，找出整句中所有需要标注的虚词及对应规则，对整句进行预处理，得到对应的词表和原始语句，以及所有待标虚词在词表和原始语句中的位置。

(3)、查找待标虚词的规则，并依序读取其用法规则信息，根据规则描述由匹配器调度程序确定触发的匹配器类型，再由相应匹配器解析用法规则，并进行对应匹配，根据匹配情况确定标注结果，待整句中所有虚词都标注完后，输出整句，并转到上一步继续读取下一个整句，直至没有待标整句，标注程序结束。

这里用到6个类型的匹配器来满足特征属性对词表查找范围与匹配返回值的不同要求。各种匹配器的设计要求及基于规则的虚词用法自动标注系统的具体实现详见袁应成等[12]。

在这个系统中，对于大规模的语料，为了方便自动识别，是以一个整句为处理单位进行标注的。因此，在读取语料时以“。”、“？”、“！”或换行符为截句标志，首先将标注语料分成一个个整句。然而，连词不同于其他词类，它可以连接句子、段落，如果仍按照上述的截句标准，按照一个整句来单独处理，那么规则描述中的特征就可能不符合，这样势必会影响某些词的标注。如：“另一方面”、“一则”、“所以”等。如下面是“另一方面”的例句：

他们一方面通过改革切实防范金融风险，强化了信贷管理，堵住了风险源头。另一方面坚持突出信贷支农重点，把新增贷款的６０％用于农业。

这个例句属于“另一方面”的“c_ling4yi1fang1mian4_1aa”的“表示互相联系的两种事物同时存在或动作性不强的两种活动同时进行”的用法，对应规则为：

<c_ling4yi1fang1mian4_1aa>→M ^M→一方面#{w}

从例句中可以看出“另一方面”可以连接两个句子，若按一个整句来截句，那么在待标的句子里就找不到对应规则要求的特征，如“一方面”，就识别失败，而只有把“一方面”所在的句子也读取到才能识别正确，这就要求至少把这两个句子作为一个单位来处理。但是在读取语料时还没有解析规则，不知道规则要求的特征以及特征与待标词之间的距离，所以很难确定一个处理单位包含几个整句。另外，某些词可能连接多个句子，这样截句就没有标准可言了。我们全面考查了语料，语料都是由段落构成，每个段落又是由若干个句子构成，并且每个段落是由诸如“19980105-01-002-017”的标号区分开的。因此，本文改进了上面的虚词用法自动标注系统，把段落标号作为截句标志，将一个段落作为处理单元。

按照段落来截句总体上考虑了连词能够连接句子、段落的功能，但是并非所有的连词都能连接句子、段落，大部分连词不具有这个功能，只能连接词语、小句，如“和”、“并”、“与”、“跟”等单纯并列连词,还有如“倘”、“尽管”等少数关联连词。单纯并列连词虽然数量很少，但是通常在语料中出现的频率较高，而且用法灵活。标注这些词时，如果按照段落来截句，会识别到一些不相关的词。如下面是“和”的例句及规则：

中国女选手显示实力，有四名单打和三对双打选手进入八强。但男单世界亚军、中国的孙俊以６：１５和１６：１８负于马来西亚新星哈希姆。

$和

@<c_he2_1b>→(、|与|同|及|以及)B~B(、|与|同|及|以及) ^B→n|r

@<c_he2_1b>→M|N ^M→~*(、|与|同|及|以及) ^N→(、|与|同|及|以及)*~

@<c_he2_1a>→M ^M→、#{、}

@<c_he2_1c>→B~B ^B→<v_d>|<a_d>

@<c_he2_2>→A~不 A ^A→a|v

@<c_he2_2>→M ^M→不论|不管|无论

@<c_he2_1>→

例句是选自一个段落中相邻的两句话，其中第一个“和”会因为第二个句子的顿号及“和”标注为“c_he2_1b”，但实际是“c_he2_1”的用法。这是因为截句方法的变化，自动标注系统在处理时，首先根据规则查找到后面句子的顿号，跨过了句号，最终导致错误。鉴于这种情况，根据用法人工总结出可以连接句子及段落的连词，将这些词放在文本文件“cword.txt”中，当标注每个连词时，首先查找这个词是否在这个文件中，如果不在这个文件中，就只需要保留待标注连词所在的句子，段落中其他句子不需要考虑。具体方法是找到待标注连词所在句子的起始和终止位置，在起始位置之前和终止位置之后的段落中的词用空格替换，这样，在后续的匹配时就不会成功，从而得到正确的标注结果。

5 实验结果及分析

5.1 实验语料

本文以 1998 年 1 月份《人民日报》分词与词性标注语料作为连词用法自动识别的语料，下面是标注语料中含有连词“并”的部分语料样例：

桥本/nr 在/p 演说/vi 中/f 表示/v ，/wd 金融/n 机构/n 的/ud 接连/d 破产/vi 将/d 造成/v 金融/n 危机/n 并/c 导致/v 经济危机/ln 。/wj

5.2 实验结果与分析

按照第 4 节中对虚词用法自动标注系统的两次改进分别进行了实验，第一个实验是完全以段落作为截句标志，第二个实验是在经过以段落作为截句标志的预处理后，再对只连接词语或小句的连词所在段落进行处理，只保留待标注连词所在句子。

本文首先用自动标注系统标注实验语料出现的所有连词得到机器标注结果，然后在此基础上进行人工校对，得到一份正确的连词用法标注语料。通过与人工校对后的语料比对，标注一致

的就认为是识别正确的。实验对语料中出现的所有连词的标注结果进行了统计，采用总体准确率来衡量连词用法的自动识别结果,具体公式为：

总体准确率 $P=\frac{S_2}{S_1}$ ，其中 S2 为识别正确的次数，S1 为语料中出现的连词的总次数。

对两次实验的结果分别做出了统计，如表 4 所示。

表 4　连词用法的自动识别结果

	语料中出现连词的总次数	识别正确的次数	总体准确率P
实验一	25364	19963	78.7%
实验二	25364	20291	80.0%

从表中可以看出，实验二的准确率高于实验一的，但是相差不大，主要原因是实验一截取的句子范围太大，致使标注系统将不该识别正确的某些用法巧合地识别成正确的结果，从而也影响了总体准确率，而实验二通过重新截句不存在这样的结果。如下面“即便”的样例语料和规则：

即便/c 如此/rzw 防范/v ，/wd 节假日/t 后/f ，/wd 动物/n 消化/v 不良/a 的/ud 毛病/n 仍/d 比/p 平时/t 高/v 出/vq 4 0 %/m 左右/m 。/wj 而/c 据/p 了解/v ，/wd 类似/v [昆明/ns 动物园/n]ns 的/ud 这种/r 情况/n ，/wd 在/p 全国/n 其他/rz 动物园/n 里/f 也/d 时常/d 发生/v 。/wj

$即便

@<c_ji2bian4_1a>→N ^N→[，](也|还)

@<c_ji2bian4_1b>→L|N ^L→n ^N→{，}*不|没|无|未

@<c_ji2bian4_2>→R|N ^R→在|对|跟 ^N→[，]*(n|t)(也|都|仍)

语料中“即便”的正确标注为“c_ji2bian4_1a”，在实验二中标为“c_ji2bian4_1b”，因为“即便”所在的句子不符合“c_ji2bian4_1a”的规则。然而实验一却标注正确，因为“即便”所在句子后面的句子，存在满足规则的特征词“也”。实验一虽然标注正确，但不是规则识别正确的，而是由标注系统的错误引起的。因此，实验一结果的真实值要比 78.7%低，而实验二的结果即为真实值，这充分说明实验二优于实验一，与实验一相比，实验二提高了连词用法自动识别的准确率。

尽管连词用法自动识别的准确率有所提高，但仍然有可提高的空间。通过分析机器标注结果与人工标注结果不一致的地方，总结出识别不正确的主要原因有下面几点。

(1)词性标注错误。有些词的某些用法不是连词的用法，但却标为连词，如“并”、“才”、“首先”、“由于”、“只是”、“只有”等。如下面“并”的词性标注错误的样例语料。

从/p 后果/n 看{kan4}/v ，/wd 当前/t 价格/n 涨幅/n 趋/Vg 低/a 、/wu 涨势/n 平缓/a ，/wd 并/c 没有/df 导致/v 经济/n 滑坡/vi ，/wd 整个/b 国民经济/n 依然/z 保持/v 了/ul “/wyz 高/a 增长/vn 、/wu 低/a 通胀/j ”/wyy 的/ud 良好/a 势头/n 。/wj

例句中的“并”是副词，表示“否定某种看法，说明真实情况”，而词性标为连词，这样就影响连词用法的自动识别的准确率。语料中出现的 25364 次连词，其中有 137 次词性标注错误，除去词性标注错误的次数，实验二的自动识别准确率为 80.4%。

(2)规则描述不完备。描述规则时遵循的原则是尽可能覆盖所有用法，并把优先级高的规则

放在前边。连词用法规则是在连词用法词典的基础上形成的，连词用法词典只包含少量例句，因此，在大规模的语料中，规则难免会出现覆盖不全或顺序不合适，从而导致标注错误或失败，这就需要在识别的过程中不断完善规则和词典。如下面是含有“不仅”的样例语料：

不仅/c<FAIL> 俄罗斯/ns ，/wd 美国/ns 的/ud 北约/ns 盟国/n 对/p 波罗的海/ns 三/m 国/n 加入/v 也/d 持有/v 很/dc 大/a 的/ud 保留/vn 。/wj

例句中“不仅”属于“c_bu4jin3_1b”的“连接名词性成分或介词短语”的用法，且构成“不仅…也”的搭配，考虑到“也”的位置，需要扩充“c_bu4jin3_1b”的规则。

原始规则为：@<c_bu4jin3_1b>→N ^N→(p|n)*，而且

修改为：

@<c_bu4jin3_1b>→←B*[，]而且 B ^B→p

@<c_bu4jin3_1b>→←*B[，]而且*B ^B→n

@<c_bu4jin3_1b>→←B*[，]B*也 ^B→p

@<c_bu4jin3_1b>→←*B[，]B*也 ^B→n

6 结束语

采用“三位一体”的构建现代汉语广义虚词知识库的思想，本文根据连词用法词典和用法规则库，并针对连词可以连接句子或段落的特殊功能，对连词用法的自动识别做出了不同于其他词的特殊处理，实现了连词用法的自动识别。但是，自动识别的准确率还不是很高。另外，某些词用法的语义上可以区分，而规则难以区分，比如“及”、“或是”、“可是”等，这些词基于规则很难自动识别。因此，下一步工作中，我们继续完善连词用法词典和规则库，构建完备精确的面向自然语言处理的现代汉语广义虚词知识库，对规则识别不好的词，考虑结合统计的方法，提高连词自动识别的准确率。此外尝试将连词用法的自动识别研究应用在褒贬新词的自动发现等自然语言处理的其它领域，期望能为构建基本褒贬资源节省人力投入。

参 考 文 献

[1] 周刚. 连词与相关问题[M]. 合肥:安徽教育出版社, 2002.

[2] 黎锦熙. 新著国语文法[M]. 北京:商务印书馆, 2000.

[3] 俞士汶, 朱学锋, 刘云. 现代汉语广义虚词知识库的建设[J]. 汉语语言与计算学报, 2003, 13(1):89-98.

[4] 刘云. 汉语虚词知识库的建设[D]. [博士后出站报告]. 北京:北京大学, 2004.

[5] 昝红英, 张坤丽, 柴玉梅, 俞士汶. 现代汉语虚词知识库的研究[J]. 中文信息学报, 2007. 9:107-111

[6] 刘锐, 昝红英, 张坤丽. 现代汉语副词用法的自动识别研究[J]. 计算机科学, 2008, 8(A):172-174

[7] 张军珲. 基于统计的现代汉语常用副词用法自动识别研究[D]. [硕士学位论文]. 郑州:郑州大学, 2010.

[8] 俞士汶, 朱学锋, 王惠, 张芸芸. 现代汉语语法信息词典详解[M](第二版). 北京:清华大学出版社, 2003.

[9] 张斌. 现代汉语虚词词典[M]. 北京:商务印书馆, 2001.

[10] 中国社会科学院语言研究所词典编辑室编. 现代汉语词典[M](第5版). 北京:商务印书馆, 2005.

[11] 吕叔湘. 现代汉语八百词(增订本)[M]. 北京:商务印书馆, 2007.

[12] 袁应成. 基于规则的虚词用法自动标注算法设计与系统实现[A]. 第十一届汉语词汇语义学研讨会论文集[C]. 苏州:苏州大学, 2010:163-169.