基于规则的连用关系标记的自动标识研究

胡金柱[1] 陈江曼[1] 杨进才[1] 舒江波[2] 雷利利[1]
(华中师范大学计算机科学系 武汉 430079)[1]
(华中师范大学国家数字化学习工程技术研究中心 武汉 430079)[2]

摘 要 复句中的关系词对研究复句中各分句的语义关系有着重要意义,但在基于规则的关系词自动识别的研究中发现,并非复句中出现的关系标记都是关系词,从中识别出真正的关系词是研究的重点和难点。提出对一种典型的关系标记——位置相邻的关系标记进行自动标记的算法,该算法结合关系词库和关系词提取技术,分析其连用特征。实验表明,该算法对连用关系标记的标识准确率达到72.9%。

关键词 有标复句,连用特征,自动标识,规则
中图法分类号 TP301.2 **文献标识码** A

Research on Auto-identifying of Adjoining Relation Markers Based on Rule

HU Jin-zhu[1] CHEN Jiang-man[1] YANG Jin-cai[1] SHU Jiang-bo[2] LEI Li-li[1]
(Dept. of Computer Science, Central China Normal University, Wuhan 430079, China)[1]
(National Engineering & Research Center for E-learning, Central China Normal University, Wuhan 430079, China)[2]

Abstract Relation words are of great significance for analyzing the semantic relations between clauses in multiple compound sentences. But in the process of automatic identification of relation words based on rule, we found that not all the relation marks in multiple compound sentence are relation words, and identifying the relation marks serving as relation words is the key point. An algorithm was proposed to identify typical relation words which are adjacent. By combining the relationship marked corpus and the extraction of relation markers, this algorithm uses the feature to identify these markers. Experiments prove that this approach achieves the accuracy of 72.9% in the identification of these relation marks.

Keywords Marked compound sentence, Adjoining feature, Auto-markup, Rule

1 引言

中文信息处理领域的目标是实现字、词、句和篇章的处理。从目前的进展来看,字、词处理已经基本实现,而句和篇章的处理还在研究中。从语法单位来讲,复句的研究属于"句处理"阶段的研究,而且目前研究得较多的复句层次关系的自动识别,也是从"句处理"层面进行的应用研究。关系词(又称关联词、关系标记)在现代汉语复句领域中起着重要的作用,是复句中标示关系的一个重要构件,是复句在语表形式上的关系标记,在很大程度上影响着分句的语义和层次关系的识别[1]。从语法单位来讲,关系词的研究应该属于"词处理"阶段。所以,现代汉语复句中关系词的自动标识(包括识别和标注)研究是复句层次关系自动识别的基础,是句法分析的重要内容,是通过词的研究促进句的研究。

由于语言的复杂性和多样性,需要完全的句法分析或语义分析,这以现在的技术还很难实现,因此采用基于规则的标识是目前一种比较有效而且实用的方法[2]。基于已构建的汉语复句语料库和复句关系词库,挖掘出关系标记在复句中充当关系词的特征规律,再将特征规律整理为规则,据此建立相应的关系标记规则库,是研究关系词计算机自动标识方法和实现策略的关键。研究中发现,影响关系标记识别的因素很多,使得识别过程需要综合形式语法、事理逻辑、句法语义的复杂计算。研究中还发现,当关系标记位置相邻时,可直接标识或简化其识别过程。例如:

(1)"电热褥"不仅能驱寒解乏,消除潮气,而且还是理想的家用医疗器具(《长江日报》1985年12月14日04版次)。

(2)弗兰克斯认为,布什"有领袖的天分",而且还是一位平易近人的领导人(CCL语料库)。

(3)只要到北戴河他就要下水,而且还是浪越大越往浪里面钻(CCL语料)。

到稿日期:2011-08-10 返修日期:2011-11-30 本文受国家教育部重点研究基地重点项目(10JJD740012),2011年国家社科基金项目(11BYY052)资助。

胡金柱(1947—),男,教授,博士生导师,主要研究领域为中文信息处理、软件工程;陈江曼(1987—),女,硕士生,主要研究领域为计算机软件理论,E-mail:chenjiangman123@126.com;杨进才(1967—),男,教授,硕士生导师,主要研究领域为现代信息系统、移动数据库;舒江波(1982—),男,博士生,主要研究领域为中文信息处理;雷利利(1986—),女,硕士生,主要研究领域为计算机软件理论。

以上3个例句中都有连用关系标记“而且还是”，当“而且”“还是”位置相邻时，大部分如例句(1)中句式：“不仅/不但(是)……，而且还是……”。在这种句式下，“而且”为关系词且为递进关系，“还是”不为关系词，但“还”为关系词。也有一部分如例句(2)和(3)，即“而且”单独出现，前面没有与之搭配的前呼标记(不仅/不但)。这种情况下，“还是”表现出两种功能：一是如(2)，其既表示递进关系，又兼句法成分；二是如(3)，其只表示递进关系。

经分析，当“而且”“还是”紧邻出现时，可以根据一定条件判断其是否为真正的关系词。实际处理过程中，连用关系标记的自动识别是一大难题，如果和普通句式一样处理，其准确率非常低。本文在根据语料总结出的识别规则的基础上，提出利用形式化描述识别规则以及触发规则进行标识的算法来解决这类位置相邻的关系标记的自动识别问题。

2 相关概念

在CCCS语料库(我们独立研制的大规模现代汉语复句生语料库，该语料库已有近百万句现代汉语复句)中，连用关系标记多为二标记连用和三标记连用。因此，本文主要从复句关系词的角度考察这两类连用关系标记的连用特征，以及它们充当关系词的条件和可能性。

定义1 连用关系标记是指两个或两个以上的关系标记在相邻的位置连续使用。一般出现为二标记连用和三标记连用。

定义2 设X,Y为关系标记，则XY为二标记连用形式。

根据对逻辑语义类别的研究，共总结出427个二标记连用形式[3]。通过对这些二标记连用形式的考察发现：标记X、Y紧邻共现，但多数情况X、Y不同时充当关系词。这种二标记连用现象在CCCS语料库中有63组，且根据其限制条件不同可分为两类：矛盾类和限制类[7]。

矛盾类：X、Y若同时充当关系词，会导致所引领的成分在表述时存在逻辑上的矛盾。标记XY出现即可判断X、Y中必有一个是伪关系词，即不为关系词。

限制类：在一定条件限制下，X、Y可以同时充当关系词。即X、Y必须且只能在一定条件下充当关系词。

定义3 设A、B、C为关系标记，则三标记连用形式ABC可表示为如下形式：

$$ABC=(AB)+(BC)=(A+B)+(B+C)$$

仔细分析CCCS语料库中出现的连用标记，共得到33种不同的三标记连用形式。再深入考察这33种不同形式的三标记连用形式，然后提取出如下25种有效形式：然而结果却、而结果却、但结果却、但却既、但却为了、但却也、但是却也、但如果因此、但却因、但是却因为、但却因为、但也因此、但是如果为了、但如果因、但无论是、但不论是、也不论是、但只要一、但是只要一、同时也就、并不是为了、看来不论是、可也是因为、也不是为了、可同时也。

从理论上讲，三标记连用由两个二标记连用形式复合而成[9]。

定义4 复句中的一对关系词，其前呼标记简称为“前呼标”，与之对应的称为“后应标”。如复句中一对关系词“因为…所以”，“因为”为“所以”的前呼标，“所以”为“因为”的后应标。

3 基于规则的系统建模

3.1 连用关系标记自动识别系统

基于规则的自动识别系统整体可分为输入、处理、输出3步。其中输入是未经处理的复句，输出则为已标识出关系词的复句，处理过程根据产生式系统将其分为3个模块：形式化规则、推理机和工作区。形式化规则即为判断关系标记是否为关系词的知识集合；工作区存储输入复句的初始句子特征、处理的中间和最后的推理结果；推理机控制执行机构，负责形式规则的前提条件的测试或匹配、规则的调度与解释执行。其处理过程的工作模型如图1所示。

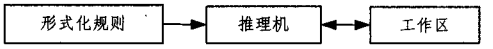

图1 工作模型图

连用关系标记是关系标记的一种特殊形式。在研究关系标记自动识别处理过程中发现，应先排除复句关系标记中的特殊形式，再处理普通关系标记。可以说，连用关系标记的自动识别是自动识别的第1阶段，第2阶段为排除连用特征后的普通关系标记的自动识别，相对前者较为简单。

自动识别关系标记之前，须对输入复句进行预处理，提取复句中的准关系词，即在关系词库中已收录待识别的关系词[5]。基于面向中文信息处理的关系词正向选择算法和已建立的关系词库[7]，先做如下前期预处理：

1)设S为输入复句，经分词后得词语集合$RW=\{W_1,W_2,W_3,\cdots\}$；

2)根据关系词库，排除非关系标记，得到准关系词集合$PS=\{W_1,W_2,W_3,\cdots\}$。

目前可用的分词软件系统都没有将位置相邻的关系标记进行合成处理，因此将连用关系标记的标识过程分为3个阶段：合成、处理、分离。首先将集合PS中位置相邻的关系标记合并，构成连用关系标识集CS；此时对CS中元素的识别过程即为连用关系标记自动识别的处理过程；在分析过程中发现，并不是每一对连用标记的连用特征都能确定其是否为关系词，所以处理完成后，须分离CS集合中没被标识的元素，并将其作为普通关系标记进入下一步普通关系标记的自动标识处理。

3.2 连用特征合成

根据连用关系标记自动识别的流程，应先合成关系标记初始集PS。而在实际处理过程中，结合关系词库分析语料中出现的连用关系标记时，发现有两种情况：一种是连用之后，X、Y仍然是两个独立的关系标记，如“所以只好”、“却因此”等，将这类连用定义为类型Ⅰ；另一种是连用之后能在关系词库中检索到，即它仍然是一个关系标记，这种情况一般是与“是”连用，如“而且是”、“不只是”、“更重要的是”等，定义为类型Ⅱ。对于类型Ⅱ，将其看作超词形式，不做连用处理，为防止干扰类型Ⅰ的识别，在处理时应剔除。针对这两种不同情况，关系标记的合成过程可以分两个阶段：

阶段1(连用词类型Ⅱ的合成) 基于关系词库的字符串

匹配，即依次取出 PS 中位置相邻的几个词语 $W_i, W_{i+1}, \cdots$，将这些词语构成词语单元 $R_w = W_i W_{i+1} W_{i+2}, \cdots$。如果关系词库中存在 R_w，则合成并取代原来的词语单元，放入初始集 PS；否则继续构造新的 R_w。

阶段 2(连用词类型Ⅰ的合成)　检索初始关系标记集 PS，提取词语位置信息，将位置紧邻的关系标记(包括一般关系标记和合成类型Ⅱ)合成为 CR_{wi}，并存储于集合 $CS = \{CR_{w1}, CR_{w2}, \cdots\}$ 中，且保留原来信息。

之所以先合成类型Ⅱ，是因为这类连用词仍可以与其他位置相邻关系标记形成类型Ⅰ。值得注意的是，阶段 1 得到的合成词 R_w 并不放入初始连用关系标记集合 CS 中。例如：

若集合 PS={…“而且”,“不只”,“是”…}，经过阶段 1，关系词库中存在“不只是”，合成为一个关系标记，刷新 PS，此时 PS={…“而且”,“不只是”}；经过阶段 2，合成为“而且不只是”，存入 CS，此时 PS 不变，CS={而且不只是}。

另外，由前面分析可知：二标记连用可分为矛盾类和限制类，三标记连用归属于限制类。矛盾类可直接利用连用特征标识关系标记；而限制类是指在一定的条件下可以为关系词，即充当关系词时必须有特定的语言环境。所以对于具体输入的复句，必须提取每个关系标记上下文相关的特征，如当前词的前一词或后一词、当前词是否位于句首或句末等。这一过程的处理定义为特征分析，由特征分析器提取句子特征信息，暂存于工作区中。由于这部分由特征分析器执行，提取句子中什么信息、多少信息，取决于总结规则时的具体分析，特征分析器将所有需要信息罗列。举例说明如下：

提取复句中准关系词后，特征表示为〈当前词，词性，所在分句索引，前一词，后一词〉。例如复句“人们不仅要工作、学习，同时也要吃饭、娱乐。”经过分词提取准关系词后，列出其特征形式，表示为：

〈不仅，c，1，人们，要〉；

〈同时，c，2，null，也〉；

〈也，d，2，同时，要〉。

提取句子特征信息即客观事实，作为关系标记的属性，其目的是能与制定的规则匹配，从而进行判断。

3.3　形式化规则表示和实现

关系标记自动识别系统中规则以产生式形式表示，一般形式为：

IF[前件] THEN[后件]

其中，前件即前提，后件表示结论或动作。其含义是如果前提得以满足，则可得出结论所规定的动作。产生式规则可用下面公式表示：

$$R_k: \bigvee_{i=1}^{m} \left(\bigwedge_{j=1}^{n} E_{ijk} \right) \rightarrow C_k$$

式中，$m, n > 1, k = 1, 2, \cdots, r$；$R_k$ 表示第 k 条规则；E 表示条件表达式；C_k 表示第 k 条规则的结论。

分别考察 CCCS 语料库中提取的 63 组二标记连用和 33 种三标记连用，分析标记之间的相互制约作用以及标记充当关系词的句法语义条件，筛选出准确率高的连用特征整理为规则。以前面例句中连用关系标记“而且还是”为例：

对于例句(1)：IF[满足句式“不仅/不但……，而且还是……”]THEN[“而且”为关系词，“还”为关系词]；对于例句(2)、(3)：IF[“而且”没有与之对应的前呼标记]THEN[“而且”为关系词，“还是”为关系词]。

系统中规则的数量很大，产生式规则最终需存储于规则库，即规则应形式化为计算机能识别的符号存储于数据表，即规则表中。而关系标记自动标识的准确性主要依赖于完备的规则库，规则的表结构设计直接影响标识的效果。根据具体情况，我们使用了两种类型规则表，即连用规则表和连用句式表。其中连用规则表主要存储约束条件较简单的规则，具体设计如下。

连用规则表 (cooccurence_rule)：有 7 个字段，自动编号(id)为关键字，系统自动编号；标记序列(keymarks)为触发该规则的索引词，即一条连用关系标记；优先级(priority)表示有多条规则时，推理机优先选取优先级高的规则执行；类型(types)表示矛盾类和限制类；约束条件(constraints)为规则满足的条件，结果(result)为索引词的处理结果，即标识为关系词或者标识为伪关系词；备注(remarks)是对规则的说明。

连用句式表主要存储约束条件中包含句式的规则。如上述例句(1)中，若满足句式“不仅/不但……，而且还是……”，则“还是”不为关系词。如果直接使用连用规则表，将无法解析，所以需要连用句式表。该表设计如下：

连用句式表(co-occurence_sentence)：连用句式表以句式作为索引词，主要是为了解决两个问题：①当约束条件为句式时，难以解析规则；②一些连用关系标记出现时与其他关系标记构成句式序列，如“不仅/而且还”、“既/同时又”、“不仅因为/而且因为”等。在连用句式表中采取条件表(rule_condition)和规则表(main_rule)分离的形式。其中：

条件表有 5 个字段：条件号(condition_ID)为关键字；索引词(index_Word)即句式中的连用关系标记；条件所在规则号(rule_ID)表示条件在哪一条规则，该属性连接条件表和规则表；搜索方向(direction)确定句式中其他关系词所处的方位，有 4 种情况：F 表示向前，即在当前索引词的前方；B 表示向后；A 表示前后两个方向都需要搜索；N 表示没有其他关系标记；备注(remark)为特殊情况的说明，可以为空。

规则表有 6 个字段：规则号(rule_ID)为关键字；索引词数目(index_Num)表示句式中非连用关系标记的个数；检索词(retrieveWord)列举句式中其他非连用关系标记；约束条件(constraints)为规则满足的条件，结果(result)为索引词的处理结果；备注(remarks)为对规则的说明。

之所以采用两子表分离的方法，是因为当出现句式中均为连用关系标记时，如“不仅因为/而且因为”，若使用一张表，一条规则存入规则表中的两个元组，会造成规则表的冗余、推理的重复。由于两个子表的分离，连用关系标记作为索引词存储于条件表中，而规则的主体存入规则表，句式中的其他关系词也存于规则表，便于被检索。如上述句式“不仅/不但……，而且还是……”中，“而且还是”为索引词，存储于条件表，设条件号为 1，其所在规则号为 1(规则表中序号均为自动增序生成)，句式的其他关系标记如“不仅/不但”存储于规则表，条件表通过条件所在规则号字段找到对应规则表中的元组，再匹配句式中其他关系标记，搜索方向字段告诉系统应该在当前索引词的前面或者后面寻找其他关系标记，可以提高

效率。如该句式中搜索方向为 F，表示句式中其他关系标记“不仅/不但”在该连用关系标记的前面；规则表中，规则号为1，句式中其他关系词为“不仅/不但”，结论是“而且”“还”为关系词。

3.4 推理机

产生式系统的推理可分为正向推理和反向推理。本文采用正向推理，其算法描述如下：

Step1 形成初始连用关系标记集 CS，特征分析器得到词语信息存入工作区；

Step2 从 CS 中取出一个元素，从工作区中取出该元素对应的特征向量 VX；

Step3 从规则表中取出一条规则，记为 R_i；

Step4 若特征向量 VX 与规则 R_i 匹配成功，结束；否则，转 Step3。

实际操作中，生成连用关系标记序列的初始集 CS 后，就可根据规则自动标识，即分别使用连用规则表的和连用句式表。这两种规则表调用方式有所不同，分别描述如下：

(1)词语触发规则(连用规则表)：由待检测的连用关系标记检测表中的索引词。其算法描述如下：

```
输入：连用关系标记初始搭配集 CS；
输出：经连用规则表检验后标识的关系词集 RCS。
Cooccurrence_Process1_Proc()
Begin
Define i=1;
If(i <= CS. num){
  从 CS 集合中取出元素 CRwi;
  If(在连用规则表中检索到 CRwi) 匹配表中约束条件并得出结论;
  i++;
}
End
```

(2)规则匹配特征(连用句式表)：充分利用其两张表的结构，检索时，先查找条件表，再由 ID 编号查找结论表。其算法描述如下，其中输入、输出与词语触发规则方法类似。

```
Cooccurrence_Process2_Proc()
Begin
Define i=1;
If(i <= CS. num){
  从 CS 集合中取出元素 CRwi;
  If(在条件表中检索到 CRwi){
    取出条件所在规则号 Idi 和搜索方向 direction;
    在 mian_rule 中找到 Idi，得到 reriveWord;
    根据 reriveWord 和 direction 在初始关系标记序列中匹配;
    If(匹配成功) 得出结论;
    }
  i++;
}
End
```

系统调用规则表时，有一定的顺序，即先调用连用规则表，后调用连用句式表。算法的基本步骤如图 2 所示。

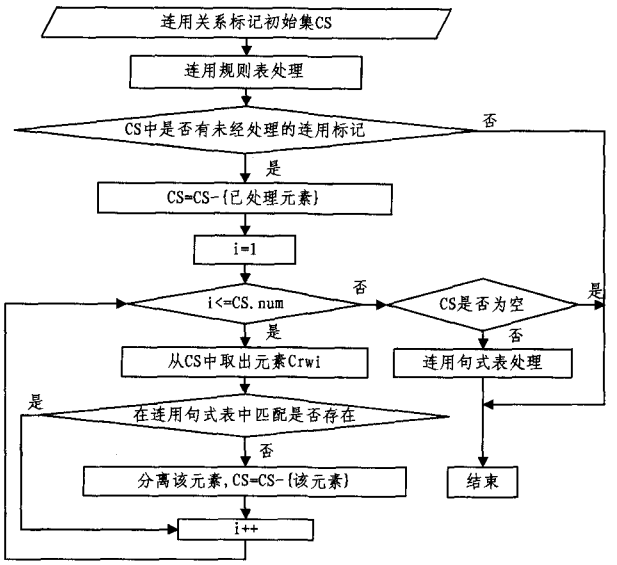

图 2 连用句式处理流程

当所合成的连用关系标记即 CS 集合中元素在连用规则表和连用句式表中均未被匹配时，说明连用特征无用，应将其分离，作为普通关系标记进入下一阶段识别。而当位置相邻的关系标记超过两个时，图中分离连用标记的过程会变得很复杂，具体说明如下：

设 $CS=\{ab,cde,fghi\}$，其中 $a,b,c,\cdots,i$ 均为独立关系标记，$ab,cde,fghg$ 表示合成之后的连用关系标记，位置相邻的关系标记最多 4 个(超过 4 个的连用 CCCS 语料中未发现)。

首先使用连用规则表对其进行识别。若索引词中存在，则直接根据约束条件得出规则结论。现假设 CS 中元素均未被处理，存入集合 CS'，逐一取出其中元素，$w_1=ab$，检索连用句式表。若在连用句式表中条件表仍没找到，则直接拆分 $w_{11}=a$，$w_{12}=b$；同样对于 $w_2=cde$，若没找到则拆分，拆分时，有两种情况：①$w_{21}=c$，$w_{22}=de$，②$w_{21}=cd$，$w_{22}=e$。分别将这两种情况在规则表中检索，如果存在则保留，否则删除；$w_3=fghi$，若表中检索不到，则直接拆分为①$w_{31}=f$，$w_{32}=ghi$，②$w_{31}=fg$，$w_{32}=hi$，③$w_{31}=fgh$，$w_{32}=i$。分别对每种情况进行检验，根据检索结果取舍，最后使用连用句式表检验。整个过程算法描述如下：

输入：连用关系标记初始集 CS

Main_Process_Proc()

Begin

Step1 $CS=\{w_1,w_2,w_3,\cdots,w_n\}$，$|CS|=n$；

Step2 for i=1 to n，使用连用规则表检测 w_i，即调用 Cooccurrence_Process1_Proc()；

Step3 CS=CS−{已处理元素}，检验 CS 是否为空，如果否，转 Step4；否则转 Step6；

Step4 分离无效连用。for j=1 to n，取出元素 w_j，在连用句式表中检索 w_j 是否存在，若存在，转下一步，否则分离该元素。

分离过程：define s=w_j 中连用的关系标记的个数。若 s=2，直接分离，并将 w_j 从 CS 中删除；若 s=3，分离为一个单用标记和一个连用标记，有两种情况 w_{j1} 与 w_{j2}，根据连用规则表和连用句式表排除不合理情况；若 s=4，分离为包含连用标记的 3 种情况分别进行检验，排除不合理情况。

Step5 判断 CS 是否为空。若是，转 Step6；否则使用连用句式表检索

CS 中剩余元素，即调用 Cooccurrence_Process2_Proc()；

Step6 结束

End

3.5 算法实例

下面以一个例子来说明上述连用关系标记自动识别算法的执行过程。

若输入复句为“不论是经济发达的资本主义国家，还是第三世界国家，都不仅开放沿海城市，而且有选择地开放内地城市，无不以借他人之力来发展自己。”

经过分词：不论/c 是/c 经济/n 发达/a 的/u 资本主义/n 国家/n，/wd 还是/c 第三世界/n 国家/n，/wp 都/d 不仅/c 开放/v 沿海/nd 城市/n，/wp 而且/c 有/v 选择/v 地/u 开放/v 内地/nl 城市/n，/wp 无不/d 以/p 借/v 他人/r 之/u 力/n 来/v 发展/v 自己/r 。/wp

提取关系词标记序列：MS={不论，是，还是，都，不仅，而且}。

合成(阶段 1、2)：MS'={不论是，还是，都不仅，而且}，CS={不论是，都不仅}。

连用规则表：检索后，索引词中没有“不论是”、“都不仅”。

分离无用特征：检索句式表，索引词中没有“都不仅”，因此分离该词为“都”、“不仅”。MS'={不论是，还是，都，不仅，而且}，CS'={不论是}。

连用句式表：条件表检索到“不论是”，所在规则表 ID=2，方向=B；在规则表中提取检索词为“还是”、“都”，在 MS' 中存在，匹配约束条件，符合，得到结果：均为关系词。

4 实验结果及分析

为了进一步验证本文算法的正确性，选取样本对算法进行测试。实验数据来源于 CCCS 语料库，从中抽取了 1525 个包含连用特征的复句。其中发生二标记连用的句子有 1290 条，占整体的 84.6%；三标记连用的句子相对较少，不足 300 条。根据句子连用标记出现的情况，总结出 4 种类型：①后应标连用，一个前呼标能与多个后应标构成搭配时，后应标发生连用，如“不仅…同时也”、“不仅…而且还”、“虽然…但却”等；②前呼标与“也”连用，如“虽然也”、“不仅也”、“也无论”等；③组合连用，前呼标与后应标都发生连用，如“不论是…还是”、“不仅因为…还因为”等；④前呼标与后应标连用，如“都不仅”等，不同类型使用两个规则表的频率不同。

算法对关系标识的识别有 3 种结果：标识为 true(是关系词)、标识为 false(不是关系词)、不做标识(不能判断)。因此对性能的评价采用两种指标：准确率和标识率。设 N 为样本中复句总数，A 代表系统不做标识且经人工验证为不能判断；B 代表系统不做标识且经人工验证为可以标识；C 代表系统标识为 true 或 false 且经人工验证标识正确；D 代表系统标识为 ture 或 false 且经人工验证可标识但标识错误；E 代表系统标识了且经人工判断不可标识。$N=A+B+C+D+E$，得如下指标：

(1)准确率：accuracy=$((A+C)/N)*100\%$，表示标识正确的比例。

(2)标识率：identify=$(1-(A/(A+E)))*100\%$，表示标识的正确率。

(3)误判率：error=$((D+E)/(C+D+E))*100\%$，表示标识但标识不正确的比例。

对 N=1525 的样本进行验证，采集的数据如表 1 和表 2 所列。

分析以上数据可知，大部分标识错误为 B 即应标识的没有标识，其占 204/413=49.4%。原因之一是连用句式库中规则没有完善；其二，类型Ⅰ的合成主要依赖于关系词库，而该关系词库中关系词收录不全面，如“而且是”、“绝不只是”等，在关系词库中没有收录。另一方面，错误 D 和 E 即误判率的产生，一是由于规则中约束条件的解析不准确，造成标识出错；另外也可能是规则制定时本身出现误差。所抽取复句中位置连用的关系标识没有超过 4 个。

表 1 样本值

类型	值
A	415
B	204
C	697
D	156
E	53
错误(B、D、E)	413

表 2 算法性能评价

评价指标	值
准确率	72.9%
标准率	88.7%
误判率	23.1%

若后期提高规则的完整性和准确率，减少错误 B 和 E 的产生，算法的准确性会大大提高。

结束语 本文主要解决关系标记自动标识系统中连用关系标记的自动标识问题。它作为整个自动标识系统的第一阶段，利用其位置相邻的特征判断其是否在句中充当关系词，并分离无法利用该特征判断的关系标记，然后将其作为普通关系标记统一处理。从实验结果来看，本文算法对连用关系标记的标识准确率达到 72.9%，这是一个非常理想的效果。这也说明，本文所研究的一种基于规则的连用关系标记的自动标识方法和实现算法是比较有效的。

由于本文中所有规则表中的规则是由人工整理并入库，其效率和准确度不高，因此应深入研究规则的自动挖掘技术，在此基础上完成规则的自动生成，使得整个系统更加自动化。

参 考 文 献

[1] 邢福义. 复句与关系词[M]. 哈尔滨：黑龙江人民出版社，1985

[2] 胡金柱，舒江波，等. 面向中文信息处理的复句关系词提取算法[J]. 计算机工程与科学，2009 (10)：90-93

[3] 姚双云. 复句关系标记的搭配研究[M]. 武汉：华中师范大学出版社，2008

[4] 俞士汶，段慧明，等. 综合型语言知识库的建设与利用[J]. 中文信息学报，2004(5)：1-10

[5] 胡金柱，吴峰文，等. 汉语复句关系词库的建设及其利用[J]. 语言科学，2010(3)：133-142

[6] 胡金柱，王琳，等. 汉语复句本体模型初探[J]. 华中师范大学学报：自然科学版，2005(4)：466-469

[7] 舒江波. 面向中文信息处理的复句关系词自动标识研究[D]. 武汉：华中师范大学，2011

[8] 肖升. 面向中文信息处理的受限有标复句联结机制分析[D]. 武汉：华中师范大学，2010

[9] 吴峰文. 面向中文信息处理的三句式有标复句层次关系自动识别研究[D]. 武汉：华中师范大学，2010