基于规则的复句中的关系词标注探讨

胡金柱[1]　沈威[1]　杜超华[1]

(1 华中师范大学 计算机科学系，武汉 430079)

提　要：汉语中关系词的自动标注问题是自然语言理解领域的基础性研究课题。本文采用基于规则的方法，并利用了华中师范大学语言与语言教育研究中心的“汉语复句库”以及利用了中科院计算语言所开发的自动分词系统 FREEICTCLAS，对复句关系词进行了有效标注。对复句关系词进行了封闭测试和开放测试，测试结果表明其准确率分别达到 84%和 85%。

关键词：基于规则；复句；关系词

Preliminary investigation on the relative of the complex sentences based on rules

HU Jin-zhu[1],SHENG Wei[1],DU Chao-hua[1]

(1.Department of Computer Science,Central China Normal University,Wuhan 430079)

Abstract: Automatically tagging relative is the basal researchful problem. We research the method of automatically tagging relatives in complete sentences effectively based on rules.The tools CCCS and the FREEICTCLAS are used in our work.Our tagging experiment gets precisions of 85% in closet test,84% in open test

Keywords: rule-based; the complex sentences; relative

1. 引言

随着计算机对大量真实文本处理的迫切需要，以及对句子浅层分析的需要，对句子中关系词的机器自动标注显得迫切重要，由于它的研究结果直接影响到以后的层次关系的标注、机器翻译等诸多领域的研究，加上这一问题本身具有的难度，使其很难从根本上解决。因此，也一直收到人们的普遍关注。

目前对对复句中关系词的标注还没有明确和系统的方法。

本文采用的对复句中对关系词自动标注的方法，思路主要采用对大部分关系词采用利用关系词表匹配的再辅以大量的规则实现。

2、复句中关系词的特点

复句是由两个和两个以上在意义上和结构上有密切联系的分句组成。[3]。复句关系词语，是复句中用来联结分句标明关系的词语，它是复句领域中一个重要的术语。

复句关系词语所标明的关系，是分句与分句之间抽象的“逻辑－语法”关系[1]。绝大多数的复句，或者分句与分句之间用了特定的关系词语，或者分句与分句之间可以用上特定的关系词语。特定的复句关系词语所构成的句式，可以看作特定的复句格式。

基金资助：本课题得到国家重点实验室开放研究基金（SKLSE04-018）和湖北省科技攻关项目（2005AA101C43）资助

作者简介：　胡金柱(1946-)：男，教授，博士生导师，主要研究方向：中文信息处理和软件工程

Email:jzhu@mail.ccnu.edu.cn

首先给出两个定义：

1）准关系词：该词语可能在复句中作为关系词也可能不作为关系词，本文称这些词为准关系词，记为 P_K，$P_K \in$ TagSet1，其中 $K \in N$。

2）关系词：该词语在复句中作为关系词，记为 C_K，$C_K \in$ TagSet1，其中 $K \in N$，TagSet1 ∈ {cyg,ctd,cjs,ctj,cmd,cbl,clg,cdj,cxz,czz,crb,cjz}。

其中 TagSet2 是我们总结了 12 个逻辑关系描述标记。[2]对应的逻辑标记如表 1 所示：

表 1　关系词的逻辑标记

Tab. 1 The logistic tag of relative

逻辑标记	所表示的关系	逻辑标记	所表示的关系
cyg	因果关系词	clg	连贯关系词
ctd	推断关系词	cdj	递进关系词
cjs	假设关系词	cxz	选择关系词
ctj	条件关系词	czz	转折关系词
cmd	目的关系词	crb	让步关系词
cbl	并列关系词	cjz	假转关系词

3. 关系词的自动标注

为了直观地描述问题，在此给出典型的经过关系词标注后的例句，见例 1。

例 1：他/r 说/v，/w<不/d 是/v>cbl 祖国/n 差/v 我/r 一个/m 人/n，/w<而是/c>cbl 我/r 认为/v 还是/c 祖国/n 好/a，/w 还是/c 社会主义/n 好/a。/w

在复句中，某些准关系词(都是经过切分后的形式)，如：因为/c、如果/c、因此/c、不但/c、而且/c 等，在复句中做关系词的用法一般比较固定。比如：因为/c 出现在句子中就标注为<因为/c>cyg，这一类词直接根据关系此表的机械匹配就可以达到比较好的效果，而另一些准关系词，如：如/c、是/v、还是/c、一边/d、一面/d 等词，在复句中有时候做关系词，有时候不作关系词，这一类词需要用规则来进行约束，所以在关系词的自动标注中要重点考虑这一类词的问题。

3.1 关系词标注的整体结构

关系词标注的整体结构如图 1：

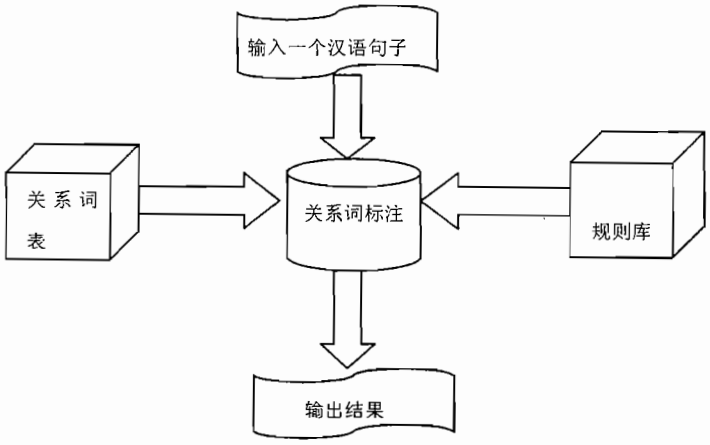

图 1　复句关系词标注结构图

Fig. 1 The structural figure of the relative of complex sentence

从图 1 看，整个关系词的识别过程分为关系词表的建立、规则库的建立、利用关系词表和规则库进行标注三个主要部分。

3.2 关系词表的建立

我们的关系词表收录了 518 对关系词，按照表 1 所示的逻辑关系可以分为 12 大类。这些关系词中我们根据

关系词对规则的相似性，将这 518 对关系词又分为 18 小类，每一个相同的小类有着相似的规则，这样分类后便于规则的通用性。

3.3 标注规则的获取

复句是由一个个字句构成的，我们把每个子句称为字段，在以下的规则中，每个准关系词表示该准关系词所在的整个字段，如果两个准关系词处于同一个字段，用＋连接两个准关系词，并表示该字段。为了形象的表示，我们看例 2：

例 2：**<不要/d 说/v>cdj** 小王/nr 办/v 不/d 好/a 这/r 件/q 事/n，/w**<连/v>cdj** 老张/nr**<都/d>cdj** 不行/a。/w

我们用规则：**不要/d 说/v，连/u＋都/d→<不要/d 说/v>cdj，<连/u＋都/d>cdj** 表示例 2 代表的句子。

现将人工获取的部分规则列举如下（见表 2）：

表 2 部分规则表

Tab.2 The table of part rules

序号	关系词	规则
1	假设/v	假定/v 居于字段之首才标注为<假定/v >cjs
2	不用说/v	居于字段之首的“不用说/v”一律标为<不用说/v >cdj.
3	不用说/v,就/d 连/v, 也/d	不用说/v,就/d 连/v, 也/d→<不用说/v>cdj,<就/d 连/v>cdj, <也/d>cdj
4	而/c	“而/c”单用要标注为<而/c >czz 的话，必须居于字段之首。
5	当然/d	“当然/d”必须居于字段之首才标注为<当然/d >czz
6	可/v	“可/v”必须居于字段之首才标注为<可/v >czz
7	以至/c	“以至/c”必须居于字段之首才标注为 czz
8	就/d 是/v，也/d	至少要求后者居于字段之首→<就/d 是/v>crb，<也/d>crb
9	既/c，可/v 也/d	既/c，可/v 也/d→<既/c>cbl，<可/v>czz<也/d> cbl
10	的话/u	“的话”跟“如果”类关系词间隔性连用标注为：<的话/u >cjs。
11	万一/m	“万一/m”居于字段之末的时候不进行标注。
12	目的关系词	目的关系词里面都要求居于字段之首，而且不能是第一个字段
13	连/u＋也/d	连/u＋也/d,不出现在第一个字段并且在同一个字段出现标为 <连/u＋也/d >cdj
14	为了/p	“为了/p”要求居于字段之首，而且要求后面紧接的部分是动词。
15	并/d	“并/d”必须居于字段之首才标注。而且其后不能跟“没有/d”、“不/d 是/v”“没/d”等。

表 2 所列举的只是我们所获取的规则中的一小部分规则，其中很多规则带有比较大的普遍性，比如第 12 条规则就代表很大一类词。由于我们将整个关系词表分为了 18 小类，每一小类对所用到的规则都是相似的，所以表 3 所列举的规则有很大的适用性。

4. 实验及结果分析

我们利用所获取的规则，依次作用于人民日报提取的训练文本（1000 句）和测试文本（500 句）分别作为封闭测试和开放测试。

4.1 试验结果

标注正确的结果见表 3：

表 3 标注正确的部分例子

Tab.3 a part of tagged correctly examples

(1)	<如果/c>cjs 处/n 在/p80/m 年代/n 末/f90/m 年代/n 初/f，/w 她/r 的/u 业务/n 水平/n，/w 她/r 的/u 服务/vn 态度/n 或许/d 会/v 使/v 她/r 成为/v 一/m 颗/q 这样/r 或/c 那样/r 的/u“/w 星/n”/w，/w<而/c>czz 当年/t 的/u 她/r，/w<连/u>cdj 加/v 工资/n 的/u 机会/n<都/d>cdj 让给/v 了/u 他人/r，/w<何况/c>cdj 那些/r 荣誉/n。/w
(2)	他们/r 是/v 忠厚/an 的/u，/w**<甚至/d>cdj** 不/d 愿/v 踩/v 死/v 一/m 只/q 小/a 蚂蚁/n，/w **<然而/c>czz** 他们/r 还/d 被/p 逼/v 着/u 去/v 拿/v 起/v 凶器/n，/w 去/v 杀死/v 与/p 他们/r 无/v 仇/vg 的/u 兄弟/n…/w

标注错误的关系词我们用加框的形式表示出来，标注结果见表 4：

表 4　标注错误的部分例子

Tab.4　a part of　tagged wrong examples

(1)	“/w 隔/v 行/ng<如/v>cjs 隔/v 山/n”/w，/w<但是/c>czz “/w 隔/v 行/ng 不/d 隔/v 理/n”/w。/w
(2)	她们/r 先/d 向/p 我/r 谈/v 了/u 她们/r 这/r 所/u 作坊/n 的/u 生产/vn 和/c 收入/n 情况/n，/w 接着/vd 就/d 把/p 话题/n 转/vd 到/v 了/u 那位/r 老人/n 身上/s，/w 说/v 他/r 姓/v 萧/nr，/w 在/p 村/n 上/m 辈数/n 最高/a，/w 今年/t 已经/d72/m 岁/q 了/y。/w

表 4 中(1)中的如/v 在这里不应该标注为关系词。 (2)中没有对关系词先/d 和接着/vd 进行标注，表 4 中标注错误的例子应该分别标注为表 5：

表 5　表 4 中的句子对应的正确形式

Tab. 5　The correct forms of the sentence corresponding to tab. 4

*(1)	“/w 隔/v 行/ng 如/vs 隔/v 山/n”/w，/w<但是/c>czz “/w 隔/v 行/ng 不/d 隔/v 理/n”/w。/w
*(2)	她们/r<先/d>clg 向/p 我/r 谈/v 了/u 她们/r 这/r 所/u 作坊/n 的/u 生产/vn 和/c 收入/n 情况/n，/w<接着/vd>clg 就/d 把/p 话题/n 转/vd 到/v 了/u 那位/r 老人/n 身上/s，/w 说/v 他/r 姓/v 萧/nr，/w 在/p 村/n 上/m 辈数/n 最高/a，/w 今年/t 已经/d72/m 岁/q 了/y。/w

我们用正确率作为衡量分析结果好坏的指标。

正确率＝正确标注的句子数/总标注的句子数。

测试结果见表 6：

表 6　测试结果

Tab. 6 Result of test

	封闭测试	开放测试
正确率	85%	84%

4.2 实验结果分析

从表 6 测试结果看出，所提取的这套规则基本令人满意，但正确率尚需进一步提高，其错误原因经过我们在语料库中随机抽取的训练集的分析后发现主要有：

（1）由于对词语切分的准确率不高，导致同一个词的词性切分有几种形式，导致我们的关系词库收录不够完善(2)规则不完善，需进一步改进（3）部分关系词涉及到语义，单纯由规则无法达到目的

5. 结束语

本文提出了一种基于规则的关系词标注方法，对于已分词和做了词性标注的句子，使用一套关系词标记集，进行人工手动标注，然后利用所标注的语料库提取规则，便于自动标注的实现。

参考文献

[1] 邢福义　汉语复句研究[M].北京：商务印书馆，2001.

[2]　Hu Jin-zhu luo xuan Xiao ming Wang lin Yao shuang-yun Luo jin-jun , Research on the Construction of Chinese Grammar Metamodel [A], 2005 Intenational Symposium on Computer Science and Technology[C],2005

[3]　邢福义．汉语语法学[M]．东北：东北师范大学出版社，1996．pp301-434.

[4] Bai Shuanghu & Xia Ying,A Scheme for Tagging Chinese Running Text[A],Proc. Of NLPRS[C],Nov 25-26,1991,Singapore

[5] D.M.Magerman and M.P.Marcus,Parsing Natural Language Using Mutual Information Statistics[A],Proc.of AAAJ[C], 1990, 984-989.

[6]杨惠中　语料库语言学导论[M].上海：上海外语教育出版社，2002.

[7]刘颖　计算语言学[M].北京：清华大学出版社，2002.