基于渡越矩阵的复句关系词自动标注初探

胡金柱[1]　沈威[1]　杜超华[1]　罗进军[2]

（1 华中师范大学 计算机科学系，武汉 430079；2 华中师范大学 文学院 语言教育研究中心，武汉 430079）

摘　要：复句关系词的自动标注是自然语言理解领域的基础性研究课题，是层次关系标注和机器翻译等问题的研究基础。本文采用概率统计方法，建立相应的渡越矩阵研究复句关系词的自动标注方法，并进行了有效标注。标注后，对复句关系词中容易产生歧义的“结果”和“如”进行了封闭性测试和开放性测试，测试结果表明其准确率分别达到98.32%和96.41%，85%和83%。

关键词：复句，关系词，渡越矩阵

Preliminary Investigation on Tagging Relative in the complex sentences based on the du-yue matrix

HU Jin-zhu[1], SHEN Wei[1], DU Chao-hua[1], Luo jin-jun[2]

(1. Department of Computer Science,Central China Normal University,Wuhan 430079; 2. Center for Language and Language Education, Humanities School, Central China Normal University,Wuhan 430079)

Abstract: Automatically tagging relative in the complete sentences is the basal researchful problem,It is the base of tagging level、machine translation.This paper establish the du-yue matrix to research the method of automatically tagging relative in the complete sentences and tag relative effectively.The relative “jieguo” and “ru” which easily produce problem is chosen in our tagging experiments in which the method gets precisions of 98.32% and 96.41%, 85 and 83% for closed test and open test respectively.

Keywords: the complex sentences; relative; duyue-matrix

1. 前言

1）复句关系词标注的作用和意义

复句关系词的作用，需要从静态和动态两个角度去考察。从静态的角度看，即从关系词语的运用结果看，关系词语的作用是表明复句关系。从动态的角度看，即从关系词语的运用过程看，对于隐性的逻辑基础来看，关系词语的作用有四种：一是显示，二是选是，三是转化，四是强化[1]。

特定的复句关系，由特定的复句关系词语标示出来。因此复句关系词是复句在语表形式上的关系标志。复句系统的建构，实质上是通过“抓住标志来实现的”。复句关系词具有复句关系的标志性。绝大多数的复句，或者分句与分句之间用了特定关系词语，或者分句与分句之间可以用上特定关系词语[1]所以复句关系词的自动标注对复句的研究有着重要作用。

2）利用规则方法进行关系词标注

鉴于复句关系词的重要作用，我们采用规则方法对复句关系词进行了自动标注。经过结果分析，采用规则方法对关系词标注取得了一定的效果，但是还有一些关系词仅仅利用规则无法达到预期的目标。基于规则的关系词标注有如下缺陷：

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

(1)规则是从封闭语料中总结出来的，对于开放语料的处理不理想；(2)分析器的鲁棒性较低，表现在遇见意外情况时常常标注错误；(3)难以保证规则的一致性(4)一些词用规则和不用规则处理的结果差别并不大。

3）渡越矩阵的方法

基于前面我们采用规则方法对关系词进行标注的不足，我们采用了基于概率统计的渡越矩阵这种方法对复句关系词进行了标注，这种方法的优点[1]是：

(1)对句子处理体统了较客观的质量；(2)对处理句子的不确定性问题有一定优势；(3)可以获取大量小粒度知识，因此处理歧义的能力增强；(4)知识一致性好。

统计方法在词性标注[2][3]，短语边界的确定[4]，句法歧义的消除[5]，多义词的词义确认[6]，等方面已取得很大进展。

2. 关系词识别的整体结构

1） 关系词识别的整体结构

关系词识别的整体结构就是对关系词进行标注时的流程，关系词的整体结构对关系词的标注起到一个宏观的作用，可以从整体上把握对关系词进行标注的步骤。

关系词识别的整体结构如图 1 所示。

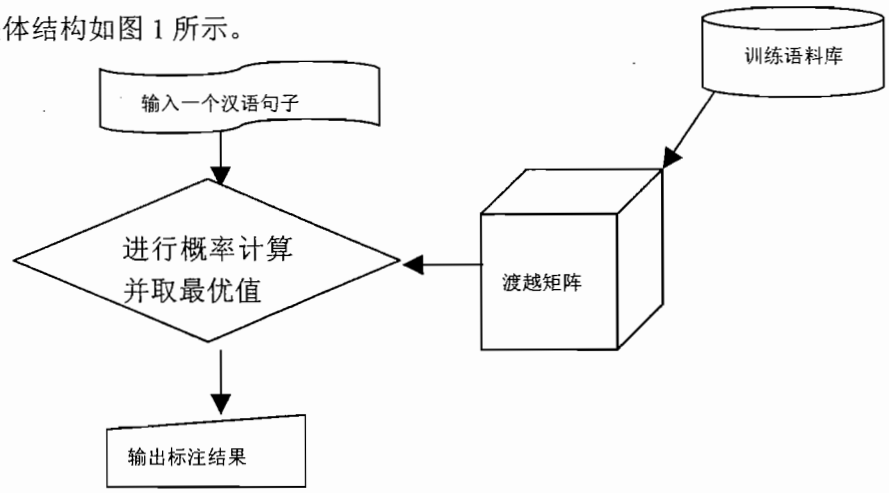

图 1　复句关系词标注结构图

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

从图 1 看，整个关系词的识别过程分为从训练语料库中构造渡越矩阵、利用渡越矩阵进行关系词的标注两个主要部分。

从训练语料库中构造渡越矩阵。本文以“结果”和“如”为例对作为训练集的语料库进行人工标注，再对所得到的经过标注关系词后的语料进行渡越矩阵的构造。

利用渡越矩阵进行关系词的标注：利用上一步构造出的渡越矩阵对语料库中的句子进行概率的计算，得到的最优解，就是我们需要标注的结果。

3. 利用渡越矩阵进行关系词的标注

3.1 关系词特征的选择

关系词 P_K 词是否在句子中作为关系词是由 P_K 本身以及由 P_K 的上下文的特征决定的。

在本文中，选取的特征包括：

1)关系词的静态特征：即考虑关系词 P_K 本身是哪一种词。其中，$P_K \in$ TagSet1, TagSet1 为集合{结果/n, 结果/d, 如/v, 〈结果/n〉cgy, 〈结果/d〉cyg, 〈如/v〉cjs }。

2）关系词的词性环境特征：当 P_K 在句子中出现时上下文词类的特征，在语料的分析考察中发现，关系词

P_K 前后取的窗口的大小对 P_K 是否在句子中做关系词的影响不大。我们只考虑与 P_K 前后相连的两个词的词性。这两类词性分别表示为：

$$C（Pre(P_K)）=X_1$$

$$C（Succ(P_K)）=X_2$$

其中 $Pre(P_K)$表示关系词 P_K 的前一词，$Succ(P_K)$为 P_K 的后一个词，X_1 表示关系词 P_K 的前一词的词性，X_2 表示关系词 P_K 的后一个词的词性，X_1 ，$X_2 \in TagSet$，TagSet 为我们定义的 42 种词性构成的集合。

3.2 渡越矩阵

渡越矩阵：$P=(A_{ij}),n$ 为词性码的总数,$i \in N$, $j \in N$，矩阵中每一项 A_{ij} 都表示词性 i 跟在词性 j 后面的概率。

$$P=\begin{pmatrix} A_{11} & \cdots A_{1j} \cdots & A_{1n} \\ \vdots & \ddots & \vdots \\ A_{n1} & \cdots A_{nj} \cdots & A_{nn} \end{pmatrix} \quad (1)$$

$$其中 Aij = \frac{训练语料库中词性i出现在词性j的前面且i和j紧邻的频数}{训练语料库中词性i出现的频数} * 100\%$$

$$= \frac{f(A+B)}{f(A)} \quad (2)$$

为了方便说明问题，假设我们所用的词性只有八个，如：/n、/v、/a、/d、/p、结果/n、/r、/e，分别代表名词、动词、形容词、副词、介词、连词、冠词、和叹词。渡越矩阵必须提供/n 后跟/n、/v、/a、/d、/p、结果/n、/r、/e 各自的概率，/v 后跟/n、/v、/a、/d、/p、结果/n、/r、/e 各自的概率，/d 后跟/n、/v、/a、/d、/p、结果/n、/r、/e 的概率，等等。该渡越矩阵表示出来就是一个表格,见表 1 所示。

表 1 渡越矩阵示例

Tab.1 The example of du-yue matrix

	/n	/v	/a	/d	/p	结果/n	/r	/e
/n	***	***	***	***	***	***	***	***
/v	12	13	10	17	21	8	16	3
/a	***	***	***	***	***	***	***	***
/d	***	***	***	***	***	***	***	***
/p	***	***	***	***	***	***	***	***
结果/n	***	***	***	***	***	***	***	***
/r	***	***	***	***	***	***	***	***
/e	***	***	***	***	***	***	***	***

表中列出了 8 种词性后跟各类词性的概率，为了表述问题的简单，我们用***表示省略了的概率值。如/v 后跟/n 的概率是 12%，后跟/v 的概率是 13%，后跟/a 的概率是 10%，后跟/d 的概率是 17%，后跟/p 的概率是 21%，后跟结果/n 的概率是 8%，后跟/r 的概率是 16%，后跟/e 的概率是 3%。如果把表一列举的 39 种词性加上“结果”(包括结果/n 和结果/d)和“如”(如/v)等三种词性，共 42 种词性，那么我们所要构造的渡越矩阵的规模应该为 42*42。

3.3 渡越矩阵中概率的获取

1）词性序列的获取：将经过切分后的汉语句子提取出词性序列 $S=W_1 \ldots\ldots W_{k-2}\ W_{k-1}P_kW_{k+1}W_{k+2}\ldots W_n$,其中 $W_i(1 \leqslant i \leqslant n) \in TagSet$, $P_K \in TagSet1$。

2）对我们所要构造的渡越矩阵中的每一项 A_{ij} 进行计算：

$$A_{ij} = \frac{P(C_j|C_i)}{P(C_i)} \quad (3)$$

其中 C_i, $C_j \in$ TagSet $\cup$ TagSet1, $P(C_i)$ 表示词性 C_i 在训练语料库中出现的概率，$P(C_j|C_i)$ 表示词性 C_j 和 C_i 共同出现且 C_j 紧跟在 C_i 之后在训练语料库中出现的概率。

由于受到训练样本语料的限制，有些概率在统计的时候属于低概率事件，这种情况称为数据稀疏问题。本文中对于概率计算为零的数值赋予一个极小值 α，以克服样本的有限性带来的数据稀疏问题。

3.4 利用渡越矩阵对关系词进行标注

1）对所输入的句子进行词性串的提取

对所输入的经过切分的句子 S 我们进行词性的提取后为 S1＝W_1……W_{k-2} $W_{k-1}P_kW_{k+1}W_{k+2}$…W_n,其中 $W_i(1\leqslant i\leqslant n)$ ∈TagSet，P_K∈TagSet1,此时我们需要考虑另外一个序列，即 S1’＝W_1……W_{k-2} $W_{k-1}\overline{P_k}W_{k+1}W_{k+2}$…$W_n$,其中 $W_i(1\leqslant i\leqslant n)$ ∈TagSet，P_K∈TagSet1,即：

$$S \Rightarrow S1, S1' \tag{4}$$

我们所需考察的句子序列用图形表示，如图 2：

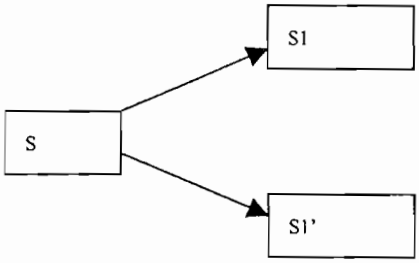

图 2　对词性序列 S 需要的考察的两个序列图

Fig.2　Two sequences of　S

即若 P_k 在 S1 中标注为关系词，则在 S1’中的形式 P_k 不是关系词，若 P_k 在 S1 中不标注为关系词，则我们考察的序列 S1’ 中 P_k 是关系词。看下面的例 1：

例 1：其次/c，/w 以前/f 一/m 说/v 搞活/v 企业/n 就是/d 减税/v 让利/v，/w 结果/n 国家/n 税利/n 让/v 了/y，/w 企业/n 仍/d 未/d 活/v 起来/v。/w

对于例 1 这个句子 S，我们需要考察两个序列 S1,S1’：

S1=/c/w/f/m/v/v/n/d/v/v，/w 结果/n/n/n/v/y，/w/n/d/d/v/v。/w

S1’=/c/w/f/m/v/v/n/d/v/v，/w<结果/n>cyg/n/n/v/y，/w/n/d/d/v/v。/w

其中 P_k=结果/n，$P_{k=}$<结果/n>cyg

2）利用渡越矩阵进行最优值的计算

对于句子 S 我们需要判断其中的关系词 P_k 是否进行了关系词词性的标注，即判断 argmax（S）中 P_k 出现的可能大还是 P_k 出现的可能大，取最优值。对于句子序列 S,我们需要考察 S1=W_1……W_{k-2} $W_{k-1}P_kW_{k+1}W_{k+2}$…W_n, S1’=W_1……W_{k-2} $W_{k-1}P_kW_{k+1}W_{k+2}$…W_n,其中 $W_i(1\leqslant i\leqslant n)$ ∈TagSet，P_K，P_K∈TagSet1，其中

$$C(Pre(P_K)) = W_{k-1}$$

$$C(Succ(P_K)) = W_{k+1}$$

$$Arg\max(S1) = \frac{P(Px \mid C(Pre(Px)))}{P(C(Pre(Px)))} * \frac{P(C(Pre(Px)) \mid Px)}{P(Px)}$$

$$= A_{x-1x} \cdot A_{xx+1} \tag{5}$$

$$Arg\max(S1') = \frac{P(\overline{Px} \mid C(Pre(\overline{Px})))}{P(C(Pre(\overline{Px})))} * \frac{P(C(Pre(\overline{Px})) \mid \overline{Px})}{P(\overline{Px})} \tag{6}$$

$$= A_{x-1x'} \cdot A_{x'x+1}$$

Argmax(S)=MAX(Argmax(S1),Argmax(S1’))；A_{ij}∈构造的 42*42 渡越矩阵

我们还是考察前面的例 1：

S=其次/c，/w 以前/f 一/m 说/v 搞活/v 企业/n 就是/d 减税/v 让利/v，/w 结果/n 国家/n 税利/n 让/v 了/y，/w 企业/n 仍/d 未/d 活/v 起来/v。/w

S1=/c/w/f/m/v/v/n/d/v/v，/w 结果/n/n/n/v/y，/w/n/d/d/v/v。/w

S1’=/c/w/f/m/v/v/n/d/v/v，/w<结果/n>cyg/n/n/v/y，/w/n/d/d/v/v。/w

$$Arg\max(S1) = \frac{P(<结果/n>cyg \mid /w)}{P(/w)} * \frac{P(/n \mid <结果/n>cyg)}{P(<结果/n>cyg)} = 0.065310911*$$
0.292857143 = 0.01912676

$$Arg\max(S1') = \frac{P(<结果/n>cyg \mid /w)}{P(/w)} * \frac{P(/n \mid <结果/n>cyg)}{P(<结果/n>cyg)} = 0.065310911*$$
0.292857143 = 0.01912676

$$Arg\max(S) = Max(Arg\max(S1), Arg\max(S1')) = Arg\max(S1') \tag{7}$$

所以例 1 中 P_K 的形式为<结果/n>cyg,例 1 应该标注为：其次/c，/w 以前/f 一/m 说/v 搞 n 就是/d 减税/v 让利/v，/w<结果/n>cyg 国家/n 税利/n 让/v 了/y，/w 企业/n 仍/d 未/d 活/v 起来/v。/w

对于“如”其标注方式与标注“结果”的方法类似，在此不再赘述。

4. 试验结果分析

本文所用的统计和测试预料都选自人民日报预料，做为训练的预料共 1319 句，其中包含“结果”的句子有 901 句，包含“如”的句子有 418 句。

分析结果表明，对于“结果”测试的结果比较理想，我们做了一个试验对 2100 个含有“结果”的复句进行测试，取得了非常理想的效果，封闭测试的准确率达到 98.32%，开放测试的准确率达到 96.41%。识别准确率达到了应用的水平。

但是“如”不仅跟其前后紧邻的词有关，而且还跟语义有关。当“如”表示列举的意义的时候不作关系词的可能性较大，当“如”表示假设意义的时候作关系词的可能性就大，我们所做的试验对 418 个含有“如”的复句进行测试，也取得了一定的效果，封闭测试的准确率达到 85%，开放测试的准确率达到 84%。识别准确率达到了一定的水平。

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 2：

表 2　测试结果

Tab.2　Result of test

基本要求	封闭测试	开放测试
（结果）正确率	98.32%	96.41%
（如）正确率	85%	84%

从表 2 的测试结果可以看出，这种利用渡越矩阵标注关系词的方法基本令人满意，但对部分标注时涉及到语义的关系词处理结果不是很理想。正确率上不去的原因：（1）该关系词涉及到语义，单纯用概率信息也无法达到很好的效果。

5. 结束语

本文提出了一种基于渡越矩阵的关系词标注方法，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概率信息，便于自动标注的实现。但是统计方法也存在一些缺陷，如在某些情况下，难以避免时间、空间的组合爆炸，难以表达语言的确定性现象，而且大批语料的标注耗费大量的人力和财力等等，所以目前尚难以做高层次的句法、语义研究。

参考文献

[1] 邢福义　汉语复句研究[M].北京：商务印书馆，2001.

[2] De Marcken,L.G.,Parsing the LOB Corpus[A],Proceeding of the 28th Annual Meeting of ACL[C],1990,243-251

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

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

[5] Tung-Hui Chiang,Yi-Chung Lin,Keh-Yih Su,Syntactic Ambiguity Resolution Using a Discrimination and Robustness Oriented Adaptive Learning Algorithm[A], Proc,of COLING-92[C], 1992,pp.352-357.

[6] William A.Gale,et al.,Using Bilingual Materials to Develop Word Sense Disambiguation Methods[A], Proc.of TMI-92[C], pp.101-112.

[7] 杨惠中 语料库语言学导论[M].上海：上海外语教育出版社，2002.

[8] 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

[9] 刘颖 计算语言学[M].北京：清华大学出版社，2002.