基于贝叶斯模型的复句关系词自动识别与规则挖掘

杨进才 郭凯凯 沈显君 胡金柱

(华中师范大学计算机学院 武汉 430079)

摘 要 复句是汉语语法的重要实体单位,关系词的自动识别是复句标识的基础,对复句的标识以及篇章的研究有重要意义。在对汉语复句语料库进行广泛分析的基础上,从复句关系词所在的环境和关系词的组合搭配方面进行特征的提取,对提取的特征进行形式化描述。采用互信息和信息增益相结合的方式进行特征选择以及冗余特征的消除;使用贝叶斯模型对特征集合进行训练和测试;将基于统计过程的结果转化为规则,形成规则库,并根据规则进行关系词自动识别。实验结果显示,本方法获得了较高的识别正确率,具有可行性和有效性。

关键词 复句关系词,贝叶斯,规则,自动标识

中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.7.062

Automatic Identification and Rule Mining for Relation Words of Chinese Compound Sentences Based on Bayesian Model

YANG Jin-cai GUO Kai-kai SHEN Xian-jun HU Jin-zhu

(School of Computer Science, Huazhong Normal University, Wuhan 430079, China)

Abstract The compound sentence is an important unit of the Chinese sentence and its annotation is important to the research on comprehending Chinese texts. Identification of relation words is the basis of compound sentence annotation. Based on a comprehensive analysis of Chinese compounds corpus, this paper extracted features of relation words from their context and collocation. Those features are described in formulas. A combination of mutual information with information gains is used for selecting features and eliminating redundant features. The Bayesian model is used for training and testing feature sets. Rules are created from the statistics results, and rule base is configured with rules, which are used for automatic identification of relation words. The experimental results show that our method obtains a high accuracy in identification, which proves the feasibility and effectiveness of the method.

Keywords Relation words, Bayesian, Rules, Automatic identification

1 引言

中文信息处理分为"字处理"、"词处理"、"句处理"、"篇章处理"4个阶段[1]。其中,"句处理"又分为"单句处理"和"复句处理",单句处理目前取得了丰硕的成果,相对而言,复句的研究尚处于起步阶段[2,3]。

中文信息学领域对复句的研究偏重于"自动处理",包括关系词的自动识别以及复句的标识,关系词的识别是复句标识的基础[4]。对关系词的自动识别的研究主要有3种途径:基于规则的方法,基于统计的方法,基于规则和统计相结合的方法。文献[5]中提出了一种基于规则的分析方法,即通过确定句子的主干成分来判定与其他词关系的方法;文献[6,7]提出一种通过给出关系词判定的约束条件来研究复句内在联系的方法;文献[8]利用关联规则的挖掘和潜在语义的分析,构造同义的关系词集。文献[9]对关系词的搭配情况进行了深入的研究,运用词语搭配理论和小句中枢理论对关系词的分布情况和搭配频率做了比较详尽的统计,并对常见关系标记组合的搭配强度做了计算,为搭配关系的研究提供了很好的依据。文献[10]建立了比较完善的复句关系词库,收录了复句中所有可能充当关系词的词语,将其作为关系词本体知识库供我们参考;文献[11]归纳了关系词所在复句的12种约束条件,形成规则库,通过规则匹配来识别关系词。文献[12]通过对短句依存关系结构的简化,构建了SVM中文依存关系解析器。文献[13]将机器学习的方法引入到复句关系词的识别问题上,取得了良好的效果。文献[14]通过构建中文比较模式库,采用机器学习的方法实现中文比较句的自动识别,其实质是一种规则与统计相结合的方法。复句与比较句不同,两者有交集,这种方法为复句关系词的自动识别提供了借鉴。

基于规则的方法的优点是直接利用了语言学家总结的规律,识别的效率与准确率高;其缺点是规则有限,识别率相对较低。基于统计的方法不需要以人工或机器自动挖掘的规则作基础,识别率高,但效率与准确率相对较低。随着大数据时代的到来,数据挖掘技术也在蓬勃发展,大规模的复句语料的处理需要数据挖掘技术的支持,如在中文文本分类中,朴素贝

到稿日期:2014-07-17 返修日期:2014-11-18 本文受教育部社科基金(13YJAZH117),国家社科基金(14BYY093)资助。

杨进才(1967-),男,教授,硕士生导师,主要研究方向为现代信息系统、中文信息处理;郭凯凯(1989-),男,硕士生,主要研究方向为中文信息处理;沈显君(1973-),男,博士,副教授,主要研究方向为数据挖掘、软件工程;胡金柱(1947-),男,教授,博士生导师,主要研究方向为中文信息处理、软件工程。

叶斯算法取得了不错的效果。本文从计算机处理自然语言的角度研究关系词的句法语义特征及其统计特征，将定性分析总结的规律、规则与基于统计的方法进行结合，以发挥两者的优势。

2 复句关系词特征提取与表示

2.1 关系词个体特征及表示

复句关系词特征的分析是仅针对复句关系词所处的语言环境的分析，通过对大量复句语料的分析和总结，将关系词的特征归纳为以下几类。

1)字面特征

表示关系词的前面与后面是否有标点符号，以及是什么符号。字面特征 WF 形式化定义如下：

$WF::=\langle F_f, F_r\rangle$

$F_f::=\{v|v\in\{1,2,3,4,5\}\}$

$F_r::=\{v|v\in\{1,2,3,4,5\}\}$

其中，F_f 为关系词的前面的标点符号值，F_r 为关系词后面的标点符号值。标点符号值 v 的取值为：1 表示逗号，2 为分号，3 为问号，4 为句号，5 为其他情况。

2)分句位置特征

表示关系词所在分句的特点，位于第一分句的用 1 表示，位于最后一个分句的用 2 表示，其他用 3 表示。

3)关系词位置特征

表示关系词是否位于复句的句首或句尾，以及关系词是否位于所在分句的句首或句尾。位于复句的句首用 1 表示，位于复句的句尾用 2 表示，位于分句的句首且不是复句的句首用 3 表示，分句的句尾且不是复句的句尾用 4 表示，其他情况用 5 表示。

4)词性特征

词性特征指复句经过分词工具分词后，关系词的词性。本文使用中科院计算所的分词工具，词性标注集有 99 个，关系词的词性标注集共有 25 类，通过排序后每类词性都对应一个序号，分词之后的词性对应的序号即为词性特征值。

5)重复特征

复句中可能存在关系词多次出现，关系词重复特征表示复句中关系词出现的次数，次数即为重复特征的取值。

2.2 关系词组合搭配特征

汉语复句中关系词的构词部件数主要有 1 个、2 个、3 个。根据关系词组合包含的关系词个数，对关系词作如下定义。

定义 1 在复句中关系词记为 A、B、C，关系词间不构成搭配关系，这样的关系词组合〈A〉称为一词；在复句中 2 个关系词可以作为一个搭配组合〈A,B〉，把这种由 2 个关系词构成的组合称为二词，它不仅可以从单个词进行特征分析，也可以从整体上考虑这对关系词组合的特征；在复句中能以 3 个关系词作为一个组合来表述句子意思，将这样的组合〈A,B,C〉称为三词；我们把二词、三词及三词以上的形式称为超词。

关系词搭配研究的基本理论与方法为关系词的自动识别提供了有效的支持。可从组合搭配方面提取关系词搭配特征并形式化表示。

1)关系词组合交换特征

表示关系词 A、B 所在的分句是否具有相同的逻辑语义类型，即关系词交换位置后，复句的语义是否改变。交换之后语义发生改变的，交换特征为 1，反之为 2。

2)关系词组合句式结构特征

表示关系词 A、B 所在的分句的结构是否相似，如果 A、B 位于同一分句，则不需要判断其句式结构，句式特征为 1；如果 A、B 不在同一分句，所在分句同为简单句，句式特征为 2；同为并列句，用 3 表示；同为复合句，用 4 表示；其他情况用 5 表示。

3)关系词组合分句位置特征

表示关系词 A、B 是否位于同一分句以及相隔的分句数。位于同一分句，用 1 来表示，同时，1 也作为了分句距离的基准。准关系词位于不同的分句时，分句位置特征值为用 B 所在分句号减去 A 所在的分句号再加 1。为了避免特征值过于离散，将结果大于 5 的情况统一用 6 表示。

一词组合〈A〉看作是不与其他关系词构成搭配关系的组合，所以无需对其进行搭配特征分析，三词组合〈A,B,C〉分为 3 个二词组合〈A,B〉、〈A,C〉、〈B,C〉考虑，同时将 A、B、C 作为一个整体进行特征分析。

例 1 佛争一炉香，人争一口气，只要有富国强民的志气，有大展宏图的信心和实干精神，更加光明美好的前景，更加幸福的春天就一定会指日可待。|《长江日报》1982 年 02 月 26 日 02 版次|

分词结果为：佛/n 争/v 一/m 炉/q 香/a，/w 人/n 争/vn 一口气/n，/w 只要/c 有/v 富国/nr 强民/nr 的/u 志气/n，/w 有/v 大/a 展/vg 宏图/n 的/u 信心/n 和/c 实干/v 精神/n，/w 更加/d 光明/a 美好/a 的/u 前景/n，/w 更加/d 幸福/a 的/u 春天/t 就/d 一定/d 会/v 指日可待/i 。/w

通过分词软件进行分词得到标记序列，利用关系词库进行筛选，得到“只要”、“就”两个准关系词，根据关系词搭配词库的匹配，复句存在〈只要，就〉二词组合。对准关系词进行个体特征和搭配特征分析，提取的特征集形式化表示为 1,5,3,3,6,1,5,5,2,5,8,1,1,5,4。

3 贝叶斯分类器的构建

本文将复句关系词的识别转化为准关系词的分类问题。贝叶斯分类器的分类原理是通过某对象的先验概率，利用贝叶斯公式计算出其后验概率，即该对象属于某一类的概率，选择具有最大后验概率的类作为该对象所属的类。

关系词组合 $WORDS$ 划分到某个类别 C_i 的概率定义为

$$p(C_i|WORDS)=\frac{p(WORDS|C_i)\times p(C_i)}{p(WORDS)} \tag{1}$$

朴素贝叶斯分类器假设复句准关系词的特征信息相互独立，所以式(1)改写成式(2)。

$$p(C_i|WORDS)=\frac{\prod_{i\in N} p(t_j|C_i)\times p(C_i)}{p(WORDS)} \tag{2}$$

其中，$p(C_i)$ 为 C_i 类出现的概率，即在关系词组合对应的训练集中属于第 i 类的该关系词组合数量。$p(t_j|C_i)$ 为训练集中属于 C_i 类的第 j 个特征项的值为 t_j 的数量占 C_i 类组合总数的百分比。

3.1 特征选择

常见的特征选择函数有互信息(Mutual Information)、TF-IDF、信息增益(Information Gain)、期望交叉熵、X2 统计(CHI)等，其中互信息选择函数比较简单，它倾向于选择频度较低的特征项，在较少类别的情况下，互信息特征选择的效果表现不错，但是当训练集很大、类别很多的时候，互信息选择

频度较低的弱点就会放大，特征选择效果一般。信息增益选择函数相对于互信息来说计算更复杂，在小样本中它的选择效果比较好，但是当训练样本很大时，效果不理想。因此，结合互信息和信息增益两种特征选择方法进行特征选择。

特征选择的步骤如下：

Step1　输入训练样本集合 S。

Step2　获取训练集中类别总数 N，以及特征项的总数 t。

Step3　计算每个特征项 A_j 与类别的互信息量 $I(C,A_j)$，根据互信息值大小对各特征项进行排序，将结果保存到一个特征集合 BS 中，$BS=\{A_i,A_j,A_k,\cdots\}$，其中 i,j,k 取 1 到 t 之间的整数。

Step4　计算包含各特征属性的训练样本的信息熵 $E(S)$。

Step5　计算不包含特征项 A_j 的期望熵 $E(S,A_j)$，以及计算特征项 A_j 的信息增益 $Gain(S,A_j)=E(S)-E(S,A)$，$Gain(S,A_j)$越大表示 A_j 对类别的贡献越大。

Step6　根据信息增益 $Gain(S,A_j)$的大小对特征项进行排序，将结果保存到集合 CS 中，$CS=\{A_i,A_j,A_k,\cdots\}$，其中 i,j,k 取值为 1 到 t 之间的整数。

Step7　通过设定一个阈值 max，各取集合 BS、CS 中前 $2*max$个元素，形成新的集合 $BS1$，$CS1$。

Step8　计算集合 BS1 与 CS1 的交集 DS。

Step9　统计 DS 中的元素个数 d，如果 d 大于等于 max，则从 DS 中提取前 max 个元素，形成新的集合 RS；如果 d 小于 max，则从 BS 或 CS 中提取前 $max-d$ 个不与 DS 中元素重复的特征集合 ES，求 DS 与 ES 的并集保存到 RS 中，RS 即为复句准关系词经过特征提取的特征集。

3.2　关系词冗余特征的消除

上述特征选择仅仅考虑了特征项与类别的关系，没有考虑特征项之间的关联程度，即一个特征项对另外一个特征项的影响，从而经过特征选择后可能存在一定量的冗余特征，因此对冗余特征进行两次过滤以消除。

第一次过滤利用相关系数实现，计算两个属性 A 和 B 之间的相关系数，它通过属性 A 与 B 的方差的几何平均数作商得到，反映了属性 A 与 B 的线性相关程度，值越大表示属性 A、B 之间联系得越紧密，反之表明联系较差。属性间的相关系数定义为

$$p(A,B)=\frac{Cov(A,B)}{\sqrt{D(A)D(B)}} \tag{3}$$

其中，$Cov(A,B)$为属性 A、B 的协方差，$D(A)$、$D(B)$分别为属性 A、B 的方差，$p(A,B)$的取值为$[-1,1]$。

$p(A,B)$等于 0 时，表示属性 A 与属性 B 完全独立，不存在任何的冗余情况，不需要对其进行消除，将完全独立的特征组合保存到集合 yes 中；$p(A,B)$等于 -1 或 1 时，表示属性 A、B 完全相关，必须要淘汰一个，完全相关的特征组合必定也被加入到了 no 集合中。通过设定一个阈值 f，将特征属性对之间关联度大于 f 的特征组合保存到集合 no 中。

特征项间的关联度用矩阵 $Matrix$ 表示如下：

$$Matrix=\begin{bmatrix} A_{11} & A_{12} & A_{13} & \cdots & A_{1i} & \cdots & A_{1t} \\ A_{21} & yes & A_{23} & \cdots & no & \cdots & A_{2t} \\ A_{31} & yes & no & \cdots & A_{3i} & \cdots & A_{3t} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ A_{j1} & A_{j2} & A_{j3} & \cdots & A_{ji} & \cdots & A_{jt} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ A_{t1} & A_{t2} & A_{t3} & \cdots & A_{ti} & \cdots & A_{tt} \end{bmatrix} \tag{4}$$

其中，A_{ij} 表示第 i 个特征项与第 j 个特征项的关联程度。

第二次过滤先遍历矩阵的对角线以上区域的值，如果标注为 or，同时对应的两个特征项都还在 RS 中，则分别计算特征项与类别的关联度，将与类别关联度小的属性删除，最后剩下的特征集合 RS'即为所求的最优特征子集。

4　关系词标识规则挖掘

基于统计的关系词自动标识中对于任意一条复句的准关系词判定都需要统计整个训练集中的特征集，求出先验概率，再利用先验概率和测试数据判定分类结果。而计算先验概率的过程被多次重复，严重影响了系统执行的效率。为了减少先验概率的重复计算，考虑将先验概率的计算用规则代替。

4.1　规则的制定与存储

制定规则的基本思想为：将贝叶斯模型计算的各准关系词组合的最优特征子集的先验概率整理成规则，设计规则库存储规则，实现复句关系词自动标识的规则挖掘，对于待判定的准关系词组合通过解析规则库中的规则进行标识。

使用贝叶斯模型的过程中产生的先验概率有类别所占比例 $P(C_i)$、类别对应特征所占比例 $P(X|C_i)$。为了将这两个先验概率转换为规则，需要设计特定的规则表示方法，制定该方法既要考虑能否准确表示先验概率，同时也要考虑到后面的规则解析过程，尽可能使解析过程简单、高效。

$P(C_i)$的存储只需要记录类别号和概率值，直接存储在相应的字段即可，$P(X|C_i)$需要记录准关系词特征项的编号、特征项的属性值、特征项对应的类别号以及概率值。本文采用将这 3 个值用“-”连接起来表示，即 i-j-A_j-k，i 表示准关系词属于第 i 类，j 表示该特征项为准关系词组合的第 j 个特征，A_j 为第 j 个特征项对应的值，k 表示前面 3 个元素对应的概率值。

〈不仅，同时〉的规则表示，及其所在二词规则表的设计如表 1 所列。

表 1　〈不仅，同时〉规则

id	keyMarks	oneRation	twoRation	threeRation	fourRation	bestCollection
1	不仅/同时	0.95652	0.04348	0	0	1-2-3-4-6-7-8-9-14-15-16
priorRation						
1-0-1-0.37121,1-0-5-0.62879,1-1-1-0.36364,1-1-3-0.63636,1-2-3-0.37121,1-2-5-0.60606,1-3-1-0.98485,1-3-2-0.01515,1-4-1-0.93182,1-4-5-0.06818,1-5-1-0.03788,1-5-5-0.96212,1-6-2-0.65909,1-6-3-0.34091,1-7-3-0.93182,1-7-5-0.04545,1-8-1-0.98485,1-8-2-0.01515,1-9-2-0.74242,1-9-5-0.0303; 2-0-1-0.33333,2-0-5-0.66667,2-1-1-0.16667,2-1-3-0.83333,2-2-3-0.33333,2-2-5-0.66667,2-3-1-1.0,2-4-1-0.66667,2-4-5-0.33333,2-5-1-0.33333,2-5-5-0.66667,2-6-2-0.5,2-6-3-0.5,2-7-3-0.66667,2-7-5-0.33333,2-8-1-1.0,2-9-2-0.33333,2-9-3-0.5,2-9-5-0.16667						

其中，id 表示规则号；keyMarks 存放关系词组合名，以“/”连接各关系词；由于“不仅/同时”只会有两种结果，“不仅”在训练集中都充当关系词，“同时”存在两种情况，因此只要 oneRatio、twoRatio 有对应的概率；bestCollection 用于记录最优特征子集各特征项在最初获取的特征集中的编号；priorRation 对应最优特征子集各特征项在训练集中出现过的特征值对应类别的先验概率，每一个特征值对应的先验概率用“,”隔开，按照 i-j-A_j-k 的结构表示，由于只有两个类别，因此 i 只有 0 和 1 取值，最优特征子集包含 11 个特征项，j 的取值为 0—10，A_j 根据各特征项的取值决定，最后的 k 为对应的先验概率。

4.2 基于规则的复句关系词自动标识

待标注复句经过分词、关系词库和搭配关系词库匹配以后，形成准关系词组合集，再通过确定关系词库的筛选，剩下未标识的准关系词组合依次查找对应的规则表，解析规则判定结果。

将贝叶斯模型的分类过程和判定结果转化为规则以后，设定规则的表示方式和解析方式，即可使用规则来进行复句关系词的自动标识，系统的整体框架如图 1 所示。

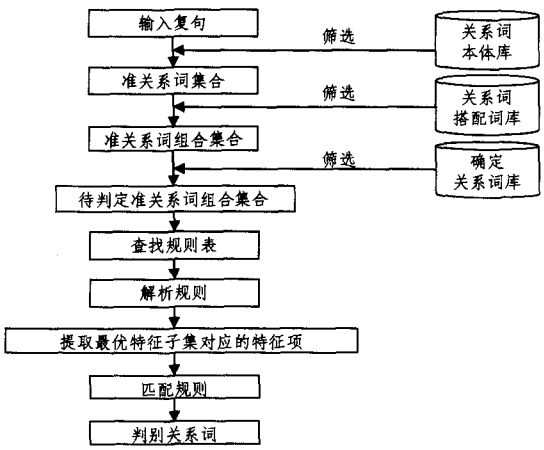

图 1 基于规则的自动标识系统框架

例 2 北京国际图书博览会作为一扇窗口，不仅展示了世界图书业的发展与兴盛，同时也使人们看到了我国出版事业前进的步伐。|《长江日报》1986 年 09 月 13 日 04 版次|

〈不仅，同时〉组合，根据二词特征形式化方法，获取特征集为“6:1,5,3,3,6,1,1,5,2,3,6,1,1,5,2”。首先通过“不仅/同时”找到对应的规则，再对规则进行解析，读取 bestCollection 字段的值“1-3-4-6-7-8-9-10-12-15”，通过“-”将字符串分割为字符数组，读取数组的值，对特征集字符串进行处理，获取的最优特征子集为“1,3,3,1,1,5,2,3,1,2”，并将最优特征子集分割成字符数组保存到数组 S 中。P(C1)直接读取 oneRation 字段，P(C2)读取 twoRation 字段，分别对特征子集中各特征项的取值找到 priorRation 字段中相应的先验概率，获取方式为先对 priorRation 字段中的内容根据“,”分割成字符串数组 CS，通过遍历数组 CS，将每一个项根据类别数和下标表示为 s1“1-j-A_j”，s2“2-j-A_j”，将这两个字符串匹配字符串数组 CS，当 CS 中字符串完全匹配 s1 或 s2 时，记录 CS 字符串数组对应的先验概率。最后，根据前面获取的先验概率进行计算，最终判定为第 1 类的后验概率为 $P(C1 \mid S) = 0.03325554918$，判定为第 2 类的后验概率为 $P(C2 \mid S) = 1.9880737942E-4$。比较两个类别的后验概率，最后判定“不仅/同时”属于第 1 类，所以“不仅”、“同时”均判定为关系词。

5 实验结果与分析

本文在专用复句语料库 CCCS 上进行实验。CCCS 语料库包含 658447 条复句，涵盖了各关系词的用法。根据本文的分类思路，选取一词、二词、三词组合各 5 组，分别对各准关系词组合选取 200 条、300 条、400 条、500 条、1000 条复句提取特征。每组中取 1/3 的特征集作为测试集，2/3 的特征集作为训练集。实验结果如图 2 所示。

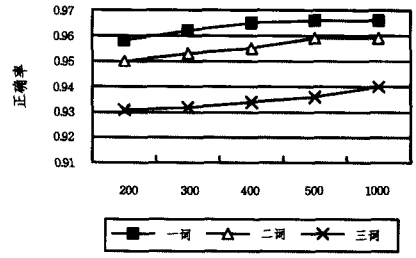

图 2 不同样本量的平均正确率

从图中可以看出，3 类组合随着样本的增加正确率会稍有提高，其中，一词组合的提升比较明显，二词组合次之，三词组合也能取得较稳定的正确率。3 类组合在样本数为 1000 时平均正确率为 95.4%。图 2 同时表明了正确率的稳定性。

本文提出的方法与相关方法的比较如表 2 所列。

表 2 相关工作的比较

方法	语料库	识别对象	正确率
规则与统计结合[4]	清华树库	准关系词已经标注为 c,d 和 l 的复句	95.7%
基于规则[11]	CCCS	连用关系词	72.9%
本文方法	CCCS	随机选取的有标复句	95.4%

文献[4]中对带功能标记和不带功能标记复句的关系词进行了判断，分别取得了 95.7%和 92.1%的高准确率。其高准确率建立在标识对象是清华树库中已经标注为 c,d 和 l 的词，也就是说，从语法的角度上已经判断出该词是句子的连接词的基础上。而本文对关系词的标识建立在准关系词的基础上，识别对象比前者宽泛。

文献[11]中连用关系标记标识准确率为 72.9%，主要原因是含连用关系词的复句结构复杂，建立连用关系词识别的规则较难，因而用基于规则的方法识别连用关系词的难度大；而本文采用了规则与机器学习相接合的方法，一定程度上弥补了规则建立难的不足。

结束语 本文在对关系词所在复句的环境进行深入分析的基础上，提取特征信息；同时，结合将准关系词组合分为一词组合、二词组合、三词组合 3 大类后的搭配特征，使用贝叶斯模型对准关系词进行识别。在用基于统计的复句关系词识别的过程中，将统计的过程和结果提炼成规则，设计规则库存储规则，将基于统计的复句关系词标识过程与规则方法结合起来，提高了关系词识别的效率与准确率。

本文采用贝叶斯模型取得了不错的效果，说明数据挖掘方法在复句关系词研究中的应用很有效，后续工作将引入更多的数据挖掘方法进行对比。

（下转封三）

Computer Society Press,2012:732-736

[18] Huang J, Schonfeld D. A novel particle filtering framework for 2D-to-3D conversion from a monoscopic 2D image sequence[C]// Proc of IEEE Conference on Visual Communications and Image Processing. Los Alamitos: IEEE Computer Society Press, 2012: 1-6

[19] Ranftl R, Gehrig S, Pock T, et al. Pushing the limits of stereo using variational stereo estimation[C]// Proc of IEEE Intelligent Vehicles Symposium. Los Alamitos: IEEE Computer Society Press, 2012: 401-407

[20] Ferstl D, Reinbacher C, Ranftl R. Image guided depth upsampling using anisotropic total generalized variation//Proc of IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2013: 1-8

[21] Rockafellar R T. Convex analysis[M]. Princeton, NJ: Princeton University Press, 1997: 1-20

[22] Bredies K, Kunisch K, Pock T. Total generalized variation[J]. SIAM Journal on Imaging Sciences, 2010, 3(3): 492-526

[23] Chambolle A, Pock T. A first-order primal-dual algorithm for convex problems with applications to imaging[J]. Journal of Mathematical Imaging and Vision, 2011, 40(1): 120-145

[24] Boyd S, Vandenberghe L. Convex optimization[M]. Cambridge: Cambridge University Press, 2004: 1-10

[25] Pock T, Chambolle A. Diagonal preconditioning for first order primal-dual algorithms in convex optimization[C] // Proc of IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2011: 1762-1769

[26] Cao X, Li Z, Dai Q H. Semi-automatic 2D-to-3D conversion using disparity propagation[J]. IEEE Transactions on Broadcasting, 2011, 57(2): 491-499

(上接第 294 页)

参 考 文 献

[1] 许嘉璐. 现状和设想—试论中文信息处理与现代汉语研究[J]. 中文信息学报, 2001, 15(2): 1-8
Xu Jia-lu. The-State-of-the-Art and the Related Strategic Considerations—On the Studies of Chinese Information Processing and Contemporary Chinese Language[J]. Journal of Chinese Information Processing, 2001, 15(2): 1-8

[2] 刘迁, 贾惠波. 中文信息处理中自动分词技术的研究与展望[J]. 计算机工程与应用, 2006(3): 175-177
Liu Qian, Jia Hui-bo. A View of Chinese Word Automatic Segmentation Research in the Chinese Information Disposal[J]. Computer Engineering and Applications, 2006(3): 175-177

[3] 黄昌宁, 赵海. 中文分词十年回顾[J]. 中文信息学报, 2007(3): 8-18
Huang Chang-ning, Zhao Hai. Chinese Word Segmentation: A Decade Review[J]. Journal of Chinese Information Processing, 2007(3): 8-18

[4] 李艳翠, 孙静, 周国栋, 等. 基于清华汉语树库的复句关系词识别与分类研究[J]. 北京大学学报, 2013(12)
Li Yan-cui, Sun Jing, Zhou Guo-dong, et al. Complex Sentence Relative Recognition and Classification Based on Tsinghua Chinese Treebank[J]. Acta Scientiarum Naturalium Universitatis Pekinensis, 2013(12)

[5] 郭艳华, 周昌乐. 自然语言理解研究综述[J]. 杭州电子工业学院学报, 2005, 20(1): 58-65
Guo Yan-hua, Zhou Chang-le. Natural Language Understanding Research Review[J]. Journal of Hangzhou Institute of Electronic Engineering, 2005, 20(1): 58-65

[6] 鲁松, 宋柔. 汉英机器翻译中描述型复句的关系识别与处理[J]. 软件学报, 2001, 12(1): 83-93
Lu Song, Song Rou. Distinction and Treatment of the Internal Relation of Descriptive Complex Sentences in Chinese-English Machine Translation[J]. Journal of Software, 2001, 12(1): 83-93

[7] 鲁松, 白硕. 汉语多重关系复句的关系层次分析[J]. 软件学报, 2001, 12(7): 987-995
Lu Song, Bai Shuo. Parsing the Logical Embedded Complex Sentences in Chinese[J]. Journal of Software, 2001, 12(7): 987-995

[8] 张文东, 易铁虎. 利用潜在语义分析和关联规则挖掘构造同义与关联词集[J]. 计算机工程与科学, 2007(1): 103-104, 116
Zhang Wen-dong, Yi Yi-hu. To Construct the Set of Synonyms and Association Words Using Latent Semantic Analysis and the Mining of Association Rules[J]. Computer Engineering & Science, 2007(1): 103-104, 116

[9] 姚双云. 复句关系标记搭配研究[M]. 武汉: 华中师范大学出版社, 2008
Yao Shuang-yun. Research on Relation Markers Collocation in Chinese Complex Sentences[M]. Wuhan: Central China Normal University Press, 2008

[10] 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, 9(2): 133-142
Hu Jin-zhu, Wu Feng-wen, et al. Establishment and Exploitation of Relationship Marked Corpus for Chinese Complex Sentences [J]. Linguistic Sciences, 2010, 9(2): 133-142

[11] 胡金柱, 陈江曼, 杨进才, 等. 基于规则的连用关系标记的自动标识研究[J]. 计算机科学, 2012, 39(7): 190-194
Hu Jin-zhu, Chen Jiang-man, Yang Jin-cai, et al. Research on Auto-identifying of Adjoining Relation Markers Based on Rule [J]. Computer Science, 2012, 39(7): 190-194

[12] Xu Y, Zhang F. Using SVM to construct a Chinese dependency parser[J]. Journal of Zhejiang University Science A, 2006, 7(2): 199-203

[13] 高维君, 姚天顺, 黎邦洋, 等. 机器学习在汉语关联词语识别中的应用[J]. 中文信息学报, 2000, 14(3): 1-8
Gao Wei-jun, Yao Tian-shun, Li Bang-yang, et al. Applying Machine Learning to Identify Chinese Discourse Markers[J]. Journal of Chinese Information Processing, 2000, 14(3): 1-8

[14] 宋锐, 林鸿飞, 常富洋. 中文比较句识别及比较关系抽取[J]. 中文信息学报, 2009, 23(2): 102-106
Song Rui, Lin Hong-fei, Chang Fu-yang. Chinese Comparative Sentences Identification and Comparative Relations Extraction [J]. Journal of Chinese Information Processing, 2009, 23(2): 102-106