Journal of Computer Applications
计算机应用,2015,35(7):1950 - 1954
ISSN 1001-9081
CODEN JYIIDU
2015-07-10
http://www.joca.cn

文章编号:1001-9081(2015)07-1950-05
doi:10.11772/j.issn.1001-9081.2015.07.1950

基于 LIBSVM 的“就是”句句间关系判别方法

周建成[1*],吴　铤[2],王荣波[1],常若愚[1]
(1.杭州电子科技大学 认知与智能计算研究所,杭州 310018;　2.杭州电子科技大学 浙江保密学院,杭州 310018)
(*通信作者电子邮箱 596599029@qq.com)

摘　要:针对使用规则和机器学习方法判别句间关系时出现因机器学习多次迭代而导致规则权值削弱现象,进而导致判别正确率偏低的问题,提出了在规则和机器学习相结合过程中对导入的明显规则特征进行加强处理的方法。首先,抽取依存词汇、语义、句子结构等具有明显规则的特有特征;然后,基于一些句间关系指示词提取普适的特征;其次,将特征写入待输入的数据向量,并且增加一维向量用来存储出现的明显规则特征;最后,运用 LIBSVM 模型结合规则和机器学习进行实验。实验结果表明,加强后的实验正确率较之加强前平均提高了两个百分点,各句间关系准确率、召回率、*F*1 值整体上都取得了较好的结果,平均值达到了 82.02%、88.95%、84.76%。实验思路和方法对研究句子间联系紧密度具有重要价值。

关键词:句间关系;LIBSVM;机器学习;kappa 值;依存词汇
中图分类号:TP399;TP181　　**文献标志码**:A

LIBSVM-based relationship recognition method for adjacent sentences containing “jiushi”

ZHOU Jiancheng[1*], WU Ting[2], WANG Rongbo[1], CHANG Ruoyu[1]
(1. *Institute of Cognitive and Intelligent Computing, Hangzhou Dianzi University, Hangzhou Zhejiang* 310018, *China*;
2. *College of Zhejiang Secrecy, Hangzhou Dianzi University, Hangzhou Zhejiang* 310018, *China*)

Abstract: Aiming at the low accuracy caused by the phenomenon of rule weight weakening from iterations of machine learning when judging the sentence relationships by applying rules and machine learning methods, the method of strengthening the imported obvious rule characteristics in the process of combining rules and machine learning was proposed. Firstly, these specific characteristics that having obvious rules such as dependency vocabulary, syntax and semantics information were extracted; secondly, universal characteristics were extracted based on these words that could indicate relationships; then, the characteristics were written into the data vector that to be input, and another dimensional vector was added to store the obvious rule characteristics; Finally, rules and machine learning methods were combined with LIBSVM model to perform the experiment. The experimental results show that the accuracy rate is averagely 2% higher than that before strengthening the characteristics, and all kinds of relationships' accurate rate, recall rate and *F*1 value show good results as a whole, their average values achieved 82.02%, 88.95% and 84.76%. The experimental ideas and methods are important for studying the compactness of adjacent sentences.

Key words: relationship between sentences; LIBSVM; machine learning; kappa value; dependency vocabulary

0　引言

近年来,随着自然语言处理技术迅速发展,篇章分析[1]、信息检索、自动问答等在内的研究都有了较大的提高,但发展的同时也面临着瓶颈,即语言的语义理解,传统语言学中对于句群层面的一些抽象定义无法满足计算机处理的需要。要想在自然语言处理技术上取得进一步的突破,必须要在语义理解、句群层面上提供相关的理论和技术支持。对于信息检索、机器翻译等的研究来说,理想的模型也应该是建立在段落或者句群层面上,因为句子本身所能承载的上下文信息太少,从单个句子到段落篇章的语义过渡又太大。为了能够更好地做到对句群、段落、篇章的处理,顺应传统语言学的思想,句间关系这一概念被引入计算语言学中。加强对句间关系的研究就成了自然语言处理的进一步发展提出的新要求。

作为句群、语义理解的重要内容,句间关系的识别研究受到了越来越多的关注。汉语句间关系研究主要集中在两个方面:一是复句,二是句群。目前,语言学上复句句间关系经常提及的有并列关系、连贯关系、选择关系、递进关系、因果关系、转折关系、条件关系、让步关系。句群句间关系[2]归纳为 12 种,即并列关系、连贯关系、递进关系、选择关系、总分关系、解证关系、因果关系、目的关系、条件关系、转折关系、让步关系、假设关系。句群句间关系包含了复句句间关系,本文即把句间关系划分为这 12 种关系。本文研究的是在一个包含“就是”的句子中,判别存在“就是”的子句和最靠近它的单句之间的关系。采用的是分层处理的方法:第一层处理明显的特征关系,主要体现规则特点;第二层基于一些能指示句间关

收稿日期:2015-01-22;**修回日期**:2015-03-22。
基金项目:国家自然科学基金资助项目(61202281);教育部人文社会科学研究项目青年基金资助项目(12YJCZH201)。
作者简介:周建成(1988 -),男,湖南邵阳人,硕士研究生,主要研究方向:中文信息处理;　吴铤(1972 -),男,浙江杭州人,教授,博士,主要研究方向:密码学、信息安全;　王荣波(1978 -),男,浙江绍兴人,副教授,博士,主要研究方向:中文信息处理;　常若愚(1986 -),男,河南平顶山人,硕士研究生,主要研究方向:中文信息处理。

系的词语来对句子进行识别,体现机器学习。如例 1 所示,存在着数量词 +(形容词) + 名词的形式,即为明显的解证关系;对于能指示句间关系的词语,例 2 中的“为了”指明了句子之间的目的关系。

例 1　黄埔军校区别于旧军校的一个显著特点,就是设有党代表和政治部。

例 2　马经常打响鼻,就是马为了排除鼻腔异物的缘故。

查阅相关数据库(Web of Science 数据库和 EI 数据库)可知,目前国外很少有相关的科研机构和学者研究汉语句间关系,已有的句间关系识别研究主要是针对英文、印度语[3]、阿拉伯语[4]和土耳其语[5],检索到的英文文献也主要是由国内机构和学者撰写。

在国内,目前关于汉语句间关系的研究也非常少见,相关论文也不多。

中国科学院声学研究所贾宁等在句间关系汉语语义块省略恢复[6]和使用句间关系恢复人名和机构名称省略方面做了一些工作。哈尔滨工业大学秦兵等针对篇章级句间语义关系识别进行了一些研究[7],厦门大学智能科学与技术系周昌乐的课题组在汉语句间关系方面做了一些科研工作[8]。

在篇章关系研究方面,苏州大学、厦门大学相关研究人员就隐式篇章关系识别进行了一些相关研究[9-11]。

本文在判别含有“就是”句子的句间关系时,运用规则和机器学习相结合的方法,同时借助于一些现有的词表《同义词词林(扩展版)》《知网情感词库》,提取出语言学中的一些特征组成特征向量,并且为了防止在多次迭代过程中出现重要权值削弱的现象,对特征向量中的规则部分进行了加强,然后利用 LIBSVM 模型进行分类。实验总体上取得了较好的结果。所选训练、测试语料来源于北京大学中国语言学研究中心(Center for Chinese Linguistics PKU, CCL)语料库。

1　第一层相关特征

1.1　解证关系

1.1.1　“就是”和代词连用时表解证关系

例 3　我采取的方法其实很简单,那就是打入他的公司内部。

例 4　拍电影就是这样,只要镜头中有一处是导演不满意的,那么整个镜头要重来。

例 3 中“那”字就是对“方法”的解释;例 4 中“这样”后面的句子就是对“这样”的解释,这里都表解证关系。但代词放于句首时不表解证,例如:他就是喜欢钓鱼,没什么别的嗜好。

1.1.2　句中存在“数量词 + 名词”结构

“就是”前面有“数量词 + 名词”形式,或者含有可以简化为以上的形式,比如“数量词 +(形容词等) + 名词”表解证关系。

例 5　室女座里有一个大型的星系团,就是著名的室女座星系团。

“一个大型的星系团”即为“数量词 + 形容词 + 名词”的形式,句间表解证关系。

1.1.3　存在着依存词汇

这里的依存词汇指经常与“就是”连用的词语。

1)和“所谓”连用。

例 6　所谓“舞为乐之容”,就是把舞蹈看成是表现音乐内容的具体可感的形貌。

2)前面紧跟“意思”“内容”“本质”“特点”“区别”“道理”等这类有内容的词汇时表解证。

例 7　当地人称它为“莫西奥图尼亚”,意思就是“声若雷鸣的雨雾”。

“有内容的词汇”本文取自于常用词语和由这些常用词经过《同义词词林扩展版》扩展出来的词汇组成的词表。

3)当“就是”后紧接“说”“指”一类动词时表解证关系。

例 8　我妈管这种做法叫作“现上轿现扎耳朵眼”,就是说老辈子时候的新媳妇一边上花轿一边扎耳朵眼来戴耳环扮淑女的意思。

1.1.4　冒号、破折号与“就是”共现表解证关系

例 9　先行试点的最大经验就是:一个决心不走样,六大班子一齐上。

1.2　递进关系

前面含有否定性词语,表递进关系。如下:

例 10　甭说打入主流,就是外围也难插足。

例 11　不用说老父亲,就是他自己也毫无办法,毫无用处了。

“甭”“不用”为否定词,句子表递进关系。

1.3　转折关系

1)前后含有极性相反的评价词汇,即前后有转折关系,表转折。这里的评价性词汇,本文使用的是《知网情感词库》中的中文正、负面评价词语表。

例 12　什么都办得不错,就是字写得难看点!

例 13　他什么都很好,就是有点自私!

“不错”“好”出自正面评价词汇表,“难看”“自私”出自负面评价词汇表。

2)“就是”后接奇数个否定词时,表转折。

例 14　你要我干什么都行,就是不能卷入任何是非。

1.4　让步关系

“就是”后面存在“也”“都”“还”时,句子间表让步关系。

例 15　就是你烧成灰,我也照样可以认得出来。

例 16　就是窝窝头,他都觉得比米饭好吃,你说怪不怪嘛?

例 17　就是那义忠老千岁爷,太上皇的兄弟,当年没得着皇位,当今圣上都大局已定,他还图谋不轨呢!

在实验输入数据的处理中,当句子中只出现一种特征时,即直接将该特征对应的维度后面的值置为 1;当句子中出现了一种以上关系特征,则需要进行优先级判断。在这个过程中,出现在逗号后面的句首词对应关系优先级最高,其次是和“就是”最靠近的词或关系,优先级最高的特征指示的关系对应维数后面 value 值置为 1,其他特征指示的关系对应维度后的 value 值则为 0.5。为了方便归一化,并且保证准确度,将 value 值设置为 1,0.5,0 三个档位。当其他特征指示的关系和优先级最高的特征指示的关系相同时,则只保留最高优先级特征。例如对下面句子的识别。

例 18　参宿四还有一个特性,就是它的体积经常处于变化之中。

例 19　“我是一只几维鸟”,意思就是“我是一个新西兰人”。

因为例 18 中只有一种特征关系,可以通过“数量词 + 名词”直接判断为解证关系,在输入实验数据的解证关系对应的维度后的值置为 1;对于例 19 的识别就会先检测“意思就是”,而忽略“数量词 + 名词”,然后将解证关系对应的维度后 value 值设置为 1。这一层中主要体现规则的特点。

2 第二层次识别

第二层次的识别是指基于句间关系指示词的识别,指示词可分为关联词和非关联词。

例 20 为什么库恩要使用在他看来还是“前科学”的数据和资料?就是因为心理学研究的独到性。

例 21 国家投入了很多钱财来培养运动员,就是为了让这些政府赞助的运动员能够为国争光。

“因为”是属于因果关系的关联词,句间表因果关系。例 21 中“为了”则指示句间为目的关系。表 1 为统计常用的能指示句间关系的词语的个数情况。

表 1 常用能指示句间关系词语的个数

关系	关联词数	非关联词数	关系	关联词数	非关联词数
让步	7	20	目的	9	14
选择	11	14	转折	9	41
连贯	11	78	条件	8	11
解证	0	30	总分	0	22
并列	14	41	假设	7	15
因果	7	36	递进	10	40

在这一层检测关联词和能指示关系类型的非关联词过程中,对于只出现一个关联词或者只出现一个能指示关系的非关联词,直接将这个关联词或非关联词指示的关系类型对应维度后的值置为 1。

例 22 就是这次情感危机,直接导致了我的学习成绩下降,以致于让我高考时名落孙山。

上例中“导致”存在于因果关系的非关联词栏,这里指示句间为因果关系。

当句子中出现一个以上的能指示关系类型的词汇时,要进行优先级判断,首先判断“就是”前面,符号后面是否存在能指示关系的句首词,如果存在,即将句首词指示的句间关系对应的维度后的值置为 1,其他指示的关系类型对应的维度后的值则置为 0.5。

例 23 不过我倒是没时间去想这些,因为接下来就是我的第一个全明星赛周末。

例 23 中存在“不过”“因为”等表示转折、因果关系的词汇,但由于“因为”是符号后面的句首词,优先级最高,因果关系对应的维度 value 值为 1,转折关系对应的维度 value 值则为 0.5。

当没有出现能指示句间关系的句首词时,进行第二级优先级判断,检测和“就是”最靠近的能指示关系的词汇、符号特征、依存关系特征。

例 24 中央实行财经集权的初衷,就是为了集中财力办大事,而九亿农民义务教育的事还小吗?

句中同时出现“为了”“而”两个指示不同关系的词,“为了”最靠近“就是”,这里目的关系对应的维度后的值为 1,转折关系对应的维度后的值则为 0.5。

对于句中没有出现关联词和能指示关系的非关联词的情况,其对应维度后的值则置为 0。

3 实验与分析

3.1 人工结果

在进行人工标注的工作方面,先在中国知网(China National Knowledge Infrastructure, CNKI)、语言学相关资料(主要为《汉语句群》)中选出 100 句已经标注正确句间关系的句子,分别由 5 名经过培训的人员标注其关系类型,统计一致性指标,计算相互之间的 *kappa* 值。

$$kappa = \frac{Po - Pe}{1 - Pe} \tag{1}$$

其中:*Po* 表示人员 *A*、*B* 判定一致的比率,即 *A*、*B* 都判定正确的比率同 *A*、*B* 都判定错误的比率之和;*Pe* 表示偶然一致的比率,即人员 *A* 错误的比率与人员 *B* 正确的比率之积同人员 *A* 正确的比率与人员 *B* 错误的比率之积的和。

测试后得到最高的 *kappa* 为 0.718,该数值反映了很好的一致性,可以认为句间关系标注一致的为正确的句间关系。

在语料处理上,从北京大学 CCL 语料库中随机抽取1 000句含有“就是”的句子,由标注一致性最好的两人进行标注,出现标注不同的句子时讨论统一为某种关系,标注的结果作为和实验结果进行对比的标准。表 2 为标注后的总体分布情况。

表 2 各关系在语料总体中所占比例

关系	所占比例/%	关系	所占比例/%	关系	所占比例/%
让步	3.00	并列	1.60	条件	0.30
选择	3.30	因果	4.10	总分	0.50
连贯	35.20	目的	3.30	假设	1.00
解证	41.10	转折	5.20	递进	1.40

3.2 LIBSVM

本文使用的 LIBSVM[12] 是一个简单、易于使用和快速有效的 SVM 模式识别与回归的软件包。

LIBSVM 使用的待输入数据文件格式如下:

< label > < index1 > : < value1 > < index2 > : < value2 > …,其中 < label > 是训练数据集的目标值,对于分类,它是标识某类的整数(支持多个类);对于回归,它可以是任意实数。< index > 是以 1 开始的整数,可以是不连续的;< value > 为实数,也就是常说的自变量。

3.3 实验结果及结论

为了提高准确率,在对句子进行识别的过程中,将层级关系导入特征向量,结合规则和机器学习,在实验中对出现的明显规则进行加强(增加第一维向量),即句子中出现第一层次特征时,置 value1 的值为 1;否则置为 0。其他维 value 值设置如下:value2 表示符号特征,即第一层的冒号、破折号特征;value3 表示句首词特征,即句首出现能指示句间关系的关联词或非关联词;value4 至 value15 则映射到 12 种关系,句子中出现能指示为某种关系类型的特征,或出现能指示为某种关系类型的关联词或非关联词时对应位则置为 1,不出现则为 0;出现多种时按照前面优先级规则;出现模糊关系时对应位置为 0.5。组成包括符号特征在内的十五维向量,基于 LIBSVM 进行实验测试。实验数据格式举例如下。11 1:1 2:0 3:0 4:0 5:0 6:0 7:0 8:0 9:0 10:0 11:0 12:0 13:0 14:1 15:0,12 种关系对应 1 到 12 这 12 个数字,第一个数字 11 即为 12 种关系中的一种,第一维是 1 表示出现第一层次特征,第二、三维是 0 表示没有出现符号特征和句首词特征,第 14 维后 value 值为 1 表示出现了能指示为第 14 维对应关系类型的特征,或者出现了能指示为第 14 维对应关系类型的关联词

或非关联词。其他维为 0 表示什么也没出现。

实验过程首先输入训练数据,由 svm-train(LIBSVM 中的函数)训练出模型,再使用 svm-predict(LIBSVM 中的函数)结合模型进行预测。最后将实验结果和人工标注的结果进行比对,求出准确率。表 3 为不同的训练句子数下加强前和加强后的正确率对比结果。

表 3　加强前后正确率对比

训练句子数	正确率/%		训练句子数	正确率/%	
	加强前	加强后		加强前	加强后
100	73.67	75.22	400	85.00	85.67
200	75.25	77.38	500	89.00	90.60
300	80.57	82.71			

由实验结果得出,加强后的实验正确率较之加强前平均高出近 2 个百分点。为了更好地说明,本文对加强后的数据作了封闭、开放测试两组实验,表 4 为实验结果。

表 4　封闭性测试正确率　%

测试句子数	训练句子数							
	100	200	300	400	500	600	700	800
100	77.00	77.00	79.00	81.00	82.00	81.00	88.00	88.00
200	—	83.50	86.00	87.00	88.00	87.00	89.50	89.50
300	—	—	87.00	88.00	86.67	88.00	90.67	90.67
400	—	—	—	90.00	90.50	90.00	92.50	92.50
500	—	—	—	—	90.60	90.40	92.80	92.80
600	—	—	—	—	—	91.00	93.33	93.33
700	—	—	—	—	—	—	93.43	93.43
800	—	—	—	—	—	—	—	92.88

从封闭实验结果可以看出:测试同样的句子数时,正确率随训练数的增大成整体递增的趋势,到一定数量(700)后,准确率不再随训练数增多而上升。分析得出:随着训练数的增多,LIBSVM 训练出的模型覆盖句子的情况就越全面,从而准确率不断上升;到一定数量后,由于模型基本涵盖了所有句型,趋于饱和状态,所以准确率不再随训练句子数的增多而增大。

从表 5 中可以看出:开放测试和封闭测试大体上呈现了一致的趋势,对于测试同样数量的句子,在训练句子数较少时,正确率随训练句子数的递增而增大,当训练数增加到一定数量(本实验为 400,500)后,识别正确率基本保持在一个稳定水平,不再随训练数的增加而变化,继续增加时,正确率甚至出现了下降。

表 5　开放性测试正确率　%

测试句子数	训练句子数								
	100	200	300	400	500	600	700	800	900
100	63.00	74.00	77.00	84.00	87.00	89.00	89.00	89.00	88.00
200	76.00	80.00	91.00	92.50	92.50	93.00	93.00	93.00	—
300	64.33	70.00	80.67	87.33	86.33	86.00	88.33	—	—
400	60.50	66.00	77.50	83.50	83.50	83.50	—	—	—
500	60.20	66.40	77.80	83.80	84.00	—	—	—	—

同封闭测试相似的是随着训练句子数的增多,LIBSVM 模型覆盖句子的情况就越全面,输入的训练句子数到一定数量后趋于饱和,即基本覆盖了所有句式情况,正确率不再随训练数的增多而增大;训练数继续增多时可能会引入更多特殊情况的句子,即引入了更多影响模型因子,从而相应地增大了误差,导致实验正确率的下降。为了得出实验正确率的主要影响因子,取出了正确率最高时的实验结果进行了分析。

为了消除随机性带来的影响,表 6 中省去了句子数少于 10 的句间关系的准确率、召回率和 $F1$ 值的计算。表 6 中,$F1$ 值总体较高,但选择、转折、递进关系的准确率偏低。在对应语料中分析后发现有一部分能指示句间关系的词汇,不单指示一种关系,它们的存在直接导致了准确率的下降,表 7 是统计的部分能指示多种关系的词语指示句间关系的概率情况。

表 6　句间关系准确率、召回率和 $F1$ 值对比

关系	准确率 P/%	召回率 R/%	$F1$/%	关系	准确率 P/%	召回率 R/%	$F1$/%
并列	85.71	92.31	88.88	因果	96.30	83.87	89.66
连贯	90.00	85.40	87.64	目的	100.00	92.86	96.30
递进	62.50	71.43	66.67	转折	66.67	96.97	79.01
选择	65.00	92.31	76.28	让步	81.82	91.67	86.47
解证	90.17	93.77	91.93				

表 7　相同词语指示不同句间关系的概率

关系	而	为了	也	不是	又	只是	只有
让步	0	0	0.08	0	0	0	0
选择	0.01	0	0.03	0.75	0	0	0.02
连贯	0.28	0	0.32	0.01	0.32	0.56	0.63
并列	0.20	0	0.46	0.24	0.53	0	0.05
因果	0	0.02	0	0	0	0.01	0
目的	0	0.98	0	0	0	0.01	0
转折	0.42	0	0.01	0	0.07	0.42	0
条件	0	0	0.02	0	0	0	0.30
递进	0.09	0	0.08	0	0.08	0	0

以“而”为例,句首为“而”在句中 42% 表转折,28% 表连贯,20% 表并列关系,9% 表递进,1% 表选择。

存在“不是……就是”的句子中,也并不全是选择关系。比如下面的例子:

例 25　不是伏案疾书,就是阅读思考。

例 26　他不是出了事,就是害了病。

例 27　这个上帝不是别人,就是全中国人民大众。

例 28　晴雯笑道:“不是新的,就是家常旧的。”

例 29　不是自觉地,就是盲目地实行某种政策。

当“不是……就是”后面的词存在逻辑上的互斥关系时,表选择关系,比如例 25、例 26 两句;但当“不是……就是”后面的词汇存在着对立关系时多数只是强调,不表选择关系,比如例 27;也有可能会出现歧义的情况,比如例 28、例 29,有歧义句子的识别还需要进一步置于段落中,根据上下文语境进行理解。

另外,多重关系的出现也影响了句间关系的判别,例如下面的例子。

例 30　莫说府里,就是上海北京,就是外洋,都这样。

例 31　含羞草就更有意思了,它不但在黑夜到来的时候会自动合上羽状的叶子,就是在白天,只要你轻轻碰它一下,它的叶子也会很快闭合。

例 30 出现了第一层中的两重关系,第一重中“就是”前面的否定词“莫”指明了“莫说府里”与后面句子之间的递进关系,第一重关系下的“就是……都”则指示了后面句子让步关系的特点。

递进关系
莫说府里，就是上海北京，就是外洋，都这样。
让步关系

例 31 中更是出现了三重关系，“不但”指示了第一重的递进关系；“就是……也”指明了第二重的让步关系；“只要”则体现了条件关系的特点。

递进关系
条件关系
它不但……叶子，就是在白天，只要你轻轻……很快闭合。
让步关系

对封闭实验结果计算分析得出，实验总体正确率为 91.10% 时，基于 LIBSVM 的实验中存在多重关系句子的识别正确率为 87.14%。可以看出多重句间关系句子的出现在一定程度上也拉低了总体的正确率。

4 结语

由于汉语本身的复杂性（句子结构复杂，语法多样，以及句子间多歧义），影响句子间关系的识别，但从表中可以看出，基于层次识别的方法还是取得了较为理想的效果。*F*1 值总体得到较好的识别结果，其中目的关系 *F*1 值较高，是由于能指示目的关系的词汇指示的关系较为确定，例如“为了”存在于句子中表目的关系的概率为 0.98。而条件关系、总分关系相对其他关系而言召回率、*F*1 值偏低（其数据不在本文列出），主要由于含有“就是”的句子中，条件句、总分句较少，存在随机性较大。

作为下一步的工作，除了分析、提取句子的其他相关特征以提高识别的准确率外，还要对识别进行扩展，使之可以处理其他类句子句间关系的判别问题。

参考文献：

[1] HUANG H, CHEN H. Chinese discourse relation recognition [C]// Proceedings of the 5th International Joint Conference on Natural Language Processing. Chiang Mai: Asian Federation of Natural Language Processing, 2011: 1442 – 1446.

[2] WU W, TIAN X. Chinese sentence group [M]. Beijing: The Commercial Press, 2000: 32 – 49. (吴为章, 田小琳. 汉语句群[M]. 北京: 商务印书馆, 2000: 32 – 49.)

[3] PRASAD R, HUSAIN S, SHARMA D M, *et al.* Towards an annotated corpus of discourse relations in Hindi [C]// Proceedings of the 6th Workshop on Asian Language Resources. Hyderabad: [s. n.], 2008: 73 – 80.

[4] AL-SAIF A, MARKERT K. The Leeds Arabic discourse treebank: annotating discourse connectives for Arabic [C]// LREC 2010: Proceedings of the 2010 International Conference on Language Resources and Evaluation. Valletta: European Language Resources Association, 2010: 2046 – 2053.

[5] ZEYREK D, WEBBER B. A discourse resource for Turkish: annotating discourse connectives in the METU corpus [C]// Proceedings of the 6th Workshop on Asian Language Resources. Hyderabad: [s. n.], 2008: 65 – 72.

[6] JIA N, ZHANG Q. Chinese ellipsis recovering based on relationship between sentences [J]. Journal of Chinese Information Processing, 2008, 22(6): 33 – 37. (贾宁, 张全. 基于句间关系的汉语语义块省略恢复[J]. 中文信息学报, 2008, 22(6): 33 – 37.)

[7] ZHANG M, SONG Y, QIN B, *et al.* Chinese discourse relation recognition [J]. Journal of Chinese Information Processing, 2013, 27(6): 51 – 57. (张牧宇, 宋原, 秦兵, 等. 中文篇章级句间语义关系识别[J]. 中文信息学报, 2013, 27(6): 51 – 57.)

[8] CHEN Y, ZHOU C. Automatic partition of Chinese sentence group [J]. Journal of Donghua University: English Edition, 2010, 27(2): 177 – 180.

[9] XU F, ZHU Q, ZHOU G. Implicit discourse relation recognition based on tree kernel [J]. Journal of Software, 2013, 24(5): 1022 – 1035. (徐凡, 朱巧明, 周国栋. 基于树核的隐式篇章关系识别[J]. 软件学报, 2013, 24(5): 1022 – 1035.)

[10] LIU C, CHEN J. Implicit discourse relation identification based on combined features and self-training learning [J]. Journal of Xiamen University: Natural Science, 2014, 53(2): 182 – 189. (刘初, 陈锦秀. 基于组合特征的自训练隐式篇章关系的识别技术[J]. 厦门大学学报: 自然科学版, 2014, 53(2): 182 – 189.)

[11] SUN J, LI Y, ZHOU G, *et al.* Research of Chinese implicit discourse relation recognition [J]. Acta Scientiarum Naturalium Universitatis Pekinesnsis, 2014, 50(1): 111 – 117. (孙静, 李艳翠, 周国栋, 等. 汉语隐式篇章关系识别[J]. 北京大学学报: 自然科学版, 2014, 50(1): 111 – 117.)

[12] CHANG C, LIN C. LIBSVM—A library for support vector machines [EB/OL]. [2014-12-20]. http://www.csie.ntu.edu.tw/~cjlin/libsvm/index.html#download.

（上接第 1944 页）

[9] BARUTCUOGLU Z, SCHAPIRE R E, TROYANSKAYA O G. Hierarchical multi-label prediction of gene function [J]. Bioinformatics, 2006, 22(7): 830 – 836.

[10] ZHANG Y, ZHOU Z. Multi-label dimensionality reduction via dependence maximization [J]. ACM Transactions on Knowledge Discovery from Data, 2010, 4(3): 14.

[11] GRETTON A, BOUSQUET O, SMOLA A J, *et al.* Measuring statistical dependence with Hibert-Schmidt norms [C]// Proceedings of the 16th International Conference on Algorithmic Learning Theory. Berlin: Springer, 2005: 63 – 77.

[12] SONG L, SMOLA A, GRETTON A, *et al.* Supervised feature selection via dependence estimation [C]// Proceeding of the 24th International Conference on Machine Learning. New York: ACM, 2007: 823 – 830.

[13] CHEN J, JI S, CERAN B, *et al.* Learning subspace kernels for classification [C]// Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2008: 106 – 114.

[14] TSOUMAKAS G, KATAKIS I. Multi-label classification: an overview [J]. International Journal of Data Warehousing and Mining, 2007, 3(3): 1 – 13.

[15] WESTON J, CHAPELLE O, ELISSEEFF A, *et al.* Kernel dependency estimation [C]// Advances in Neural Information Processing Systems 15. Cambridge: MIT Press, 2003: 873 – 880.

[16] COVER T M, THOMAS J A. Elements of information theory [M]. Hoboken: Wiley-Blackwell, 1991: 13 – 19.

[17] ZHANG M, ZHOU Z. ML-*k*NN: a lazy learning approach to multi-label learning [J]. Pattern Recognition, 2007, 40(7): 2038 – 2048.

[18] TROHIDIS K, TSOUMAKAS G, KALLIRIS G, *et al.* Multi-label classification of music into emotions [C]// Proceedings of the 9th International Conference on Music Information Retrieval. Philadelphia: Drexel University, 2008: 325 – 330.