软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2013,24(5):1022−1035 [doi: 10.3724/SP.J.1001.2013.04283]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

基于树核的隐式篇章关系识别*

徐 凡[1,2], 朱巧明[1,2], 周国栋[1,2]

[1](苏州大学 计算机科学与技术学院,江苏 苏州 215006)

[2](江苏省计算机信息处理技术重点实验室,江苏 苏州 215006)

通讯作者: 朱巧明, E-mail: qmzhu@suda.edu.cn

摘 要: 隐式篇章关系识别是篇章结构分析中最具有挑战性的任务之一.传统的方法注重篇章中的概念和意义特征,导致系统的性能不高.系统地探索了篇章中的浅层语义信息和以态度韵为导向的句子级情感等平面特征的有效性,同时提出了一种简单而有效的树核方法,最后采用复合核方法加以集成.在 Penn Discourse Treebank(PDTB) 2.0 语料库上的实验结果表明,引入浅层语义和情感等信息后,准确率得到显著提升.

关键词: 篇章;篇章结构分析;隐式篇章关系识别;树核;复合核

中图法分类号: TP391 **文献标识码**: A

中文引用格式: 徐凡,朱巧明,周国栋.基于树核的隐式篇章关系识别.软件学报,2013,24(5):1022−1035. http://www.jos.org.cn/1000-9825/4283.htm

英文引用格式: Xu F, Zhu QM, Zhou GD. Implicit discourse relation recognition based on tree kernel. Ruan Jian Xue Bao/Journal of Software, 2013,24(5):1022−1035 (in Chinese). http://www.jos.org.cn/1000-9825/4283.htm

Implicit Discourse Relation Recognition Based on Tree Kernel

XU Fan[1,2], ZHU Qiao-Ming[1,2], ZHOU Guo-Dong[1,2]

[1](School of Computer Science and Technology, Soochow University, Suzhou 215006, China)

[2](Jiangsu Provincial Key Laboratory for Computer Information Processing Technology, Suzhou 215006, China)

Corresponding author: ZHU Qiao-Ming, E-mail: qmzhu@suda.edu.cn

Abstract: As a critical sub-task in discourse structure analysis, implicit discourse relation recognition (iDRR) is a challenging natural language processing task. Traditional approaches focus on exploring concepts and sense in discourse, which result in poor performance. This paper first systematically explores the efficiency of shallow semantic and attitude prosody-driven sentence-level sentiment information in discourse. Next, the paper proposes a simple but effective tree structure and finally investigates the efficiency of a composite kernel. Evaluation on Penn Discourse Treebank (PDTB) 2.0 shows the importance of shallow semantic and sentiment information across the discourse, and the appropriateness of the composite kernel in iDRR. It also shows that this system significantly outperforms other ones currently in the research field.

Key words: discourse; discourse structure analysis; implicit discourse relation recognition; tree kernel; composite kernel

篇章结构分析(discourse structure analysis,简称 DSA)旨在研究自然语言文本的内在结构,并理解文本单元(可以是词、句子、从句或段落)间的逻辑关系,从而挖掘出自然语言文本内部丰富的结构和语义信息,对自然语言理解和自然语言生成起着至关重要的作用.篇章结构分析具有较广泛的潜在应用价值,例如:存在原因(cause)篇章关系的两个论元(argument)可以为自动问答系统(question answering,简称 QA)提供事实型答案.另外,篇章

* 基金项目: 国家自然科学基金(60970056, 90920004); 国家高技术研究发展计划(863)(2012AA011102); 高等学校博士学科点专项科研基金(20093201110006); 江苏省自然科学基金(BK2011282); 江苏省高校自然基金(11KIJ520003); 江苏省普通高校研究生科研创新计划(CXZZ11_0101)

收稿时间: 2012-03-05; 修改时间: 2012-05-18; 定稿时间: 2012-07-03

结构单元间的主次关系可以为机器自动文摘(summarization)的句子抽取和压缩任务提供直接的理论依据和实践指导.这些潜在的应用前景使得篇章结构分析逐渐成为国内外的研究热点.

作为篇章结构分析任务之一的篇章关系识别包括显式篇章关系识别和隐式篇章关系识别两种类型.其中,显式篇章关系识别旨在判断篇章中在给定连接词(如 but,so 等)前提下的两个论元间存在何种诸如时序(temporal)、对比(comparison)、可能性(contingency)或扩充(expansion)等逻辑关系,而隐式篇章关系识别旨在判断篇章中在没有给定连接词的两个论元间存在何种上述逻辑关系.由于显式篇章关系的两个论元间存在显示的连接词,我们可以在绝大多数情况下仅根据连接词本身就可以比较准确地识别出论元间存在的特定逻辑关系,并且通常可以达到 94%左右的 $F1$ 度量值($F1$-measure)[1].例如,连接词 But 通常表示 Comparison 这种逻辑关系.相反,由于隐式篇章关系的两个论元间不存在显式连接词,我们只能根据论元内部具有的一些词汇、意义等语言学特征对其进行识别,通常仅有 40%左右的准确率(accuracy)[2,3].因此,隐式篇章关系识别是自然语言处理(natural language processing,简称 NLP)中最具挑战性的任务之一.

近年来,由于修辞结构理论篇章树库(rhetorical structure theory-discourse treebank,简称 RST-DT)[4]、宾州篇章树库(Penn discourse treebank,简称 PDTB)[5]和图库(GraphBank)[6]等大规模人工标注的篇章语料库的出现,篇章关系识别研究已从基于规则的方法转为基于机器学习的方法.目前,绝大多数方法都将其看成是分类问题,其基本过程包括:首先,把标注好的篇章语料分成训练语料和测试语料两个部分,从训练语料中提取出各种词汇、语法、句法和语义等特征,生成相应的篇章关系训练实例集,并利用支持向量机(support vector machine,简称 SVM)、最大熵(maximum entropy,简称 ME)、朴素贝叶斯(naive Bayes,简称 NB)等分类器训练得到分类器模型;然后,利用分类器模型对测试样例进行预测,给出测试样例可能存在的篇章逻辑关系.

然而,目前已有的方法都比较注重篇章中的概念和意义等特征,例如论元中的单词本身,或者论元中的单词形态学(词干化、曲折变化)特征,或者两个论元中的单词词对等特征.而篇章中存在的浅层语义信息(本文提到的浅层语义信息也称语义角色)(例如,句子中的施事者、受事者、时间、地点等语义角色)和句子级的情感特征(说话者对句子所包含的正、负和中性情感判断)却基本被忽略.事实上,这种浅层语义信息和句子级的情感特征对隐式篇章关系识别有重要的作用(见本文第 3.1 节中的实例以及相关的语言学解释).鉴于此,本文探索了篇章中浅层语义信息和句子级情感等新型的平面特征在隐式篇章关系识别中的重要性;同时,为了弥补平面特征在表示结构化数据时的局限性,提出了一种简单而有效的树核方法;最后,采用复合核将平面特征和树核方法加以集成.PDTB 2.0 上的实验结果验证了我们的平面特征、树核和复合核方法对隐式篇章关系识别任务都有非常重要的作用.

本文第 1 节介绍篇章关系识别的相关工作.第 2 节简要介绍 PDTB 2.0 篇章语料库.第 3 节着重阐述本文的隐式篇章关系识别方法以及各种平面和结构化特征.第 4 节给出实验设置及详细的结果分析.第 5 节给出本文的结论,并对将来的工作进行介绍.

1 相关工作

我们可以从不同角度对篇章关系识别的相关工作进行分类:从是否存在连接词的角度上,可以分为显式篇章关系识别和隐式篇章关系识别两大类;从采用的语料库角度上,可以分为基于 RST-DT 的篇章关系识别、基于 PDTB 的篇章关系识别和基于 GraphBank 的篇章关系识别这 3 种类型.正如文章开始部分所述,由于显式篇章关系识别任务比较简单,因此我们着重介绍隐式篇章关系识别的相关工作.

基于 PDTB 的隐式篇章关系识别方法可以进一步分为有指导的、无指导的和半指导的这 3 种类型.对于有指导的机器学习方法,Wang 等人[2]提出了一种基于树核的方法对显式和隐式篇章关系同时进行识别,采用了类似核心短语、相邻句子标点符号、连接词位置、论元跨度等平面特征,提出了基于成份句法分析的最小扩充核、简单扩充核和完整扩充核这 3 种树结构,同时探索了论元间的时序关系等语言学特征.对于 4 大类(temporal, comparison,contingency 和 expansion)的隐式篇章关系识别任务,在 PDTB 2.0 语料库上取得了 40%的 Accuracy,这也是到目前为止最好的基于树核的 4 大类隐式篇章关系识别性能.Lin 等人[3]对 11 大类(4 大类的下一层)隐

式篇章关系识别进行了相应的研究,提出了较为有效的成份句法树产生式规则、依存树规则和上下文信息等特征,在 PDTB 上取得了 40.2%的 Accuracy.Zhou 等人[7]探索了篇章连接词在隐式篇章关系识别的重要性问题,首先,利用语言模型将论元间的隐式连接词进行恢复;然后,利用恢复的连接词和一些已有的诸如动词、词性、数字等平面特征进行组合,训练分类器模型对测试实例进行预测,仅取得了 41.35%的 Accuracy.Pilter 等人[8]提出了一些新型平面特征,例如动词的极性、货币/百分比/数字、词的形态、两个论元的前 3 个单词及相关组合(first-last-first3)等,在利用 PDTB 进行隐式篇章关系识别时,他们把 PDTB 中的没有关系(no relation,简称 NoRel)和基于实体一致性关系(entity-based coherence relation,简称 EntRel)这两种类别划分为 Expansion 类别,利用朴素贝叶斯分类器训练分类器模型,他们最好的 First-Last-First3 特征取得了 65.40%的 Accuracy.但是,正如我们所述,他们的结果是在把 NoRel 和 EntRel 划分为 Expansion 类别情况下取得的,为了验证 PDTB 中原始的 4 大类数据分布下的性能(expansion 中不包括 NoRel 和 EntRel 类别),First-Last-First3 特征仅取得了 47.51%的 Accuracy.

相对于有指导的机器学习方法而言,无指导和半指导方法比较少.Zhou 等人[9]提出了一种无指导的方法,首先,利用语言模型对隐式篇章连接词进行恢复;然后利用 PDTB 中连接词所对应的篇章关系统计数据直接对隐式篇章关系进行预测.Hernault 等人[10]利用半指导式机器学习方法对低共现篇章关系进行识别,研究了未标注数据与标注数据的共现特征,试图解决在标注样本较少领域中的篇章关系识别问题.从这些最新工作所取得的结果来看,隐式篇章关系识别的确是一项有挑战性的工作.

基于 RST-DT 的篇章关系识别[11−15]主要分为两个子任务:其一是基本篇章单元(elementary discourse unit,简称 EDU)的生成,其二是用生成的任意两个 EDU 对可能存在的篇章关系进行识别.其中,文献[11,12]仅研究了 EDU 的生成子任务.文献[11]利用句法和词汇等特征对文本进行分割,同时采用了 12 条基于句法的分割规则,取得了 84%的 F1-measure.文献[12]将篇章分割问题看成序列化标注问题,抽取出文本中的单词、词性(part-of-speech,简称 POS)、词汇中心词、词汇-句法等特征,利用条件随机场(conditional random fields,简称 CRF)分类器取得了 94%的 F1-measure.文献[13,14]分别采用概率模型和句法规则方法进行文本切分和篇章关系识别.文献[15]是一种完全监督的机器学习方法,考虑浅层词汇、结构化成份句法、控制集、修辞子结构等特征,采用两层式 SVM 分类器,其中,二元 SVM 分类器决定修辞关系是否出现,多元 SVM 分类器决定具体的修辞关系类别,取得了 48.1%的 F1-measure.

基于 GraphBank 的代表性工作[16]的研究结论是:篇章连接词和两个文本跨度的距离对显示和隐式篇章关系识别均起到关键作用.但是,他们所提出来的特征在 PDTB 的隐式篇章关系识别中起的作用并不大,因为 PDTB 中规定隐式篇章关系的两个论元间没有连接词,同时规定两个论元是相邻的关系.

与篇章关系识别相关的工作还有文献[17−19],其中,文献[17]标注了 81 篇石油和旅游领域的中文篇章语料,完成了 3 081 个句对的小规模的中文篇章树库.但是,其当前的版本主要存在以下两个问题:其一是在篇章连接词的论元单位界定上,他们以句子作为基本单位,然而实际情况更复杂,这种论元单位可以是从句、一个句子或多个句子;其二是标注一致性有待于检验,文献作者并没有给出标注一致性 Kappa 值,但是 Kappa 值是衡量任何语料不可缺少的重要因素,因为它是标注工作的难度和标注质量的直接体现.文献[18]研究了在给定连接词前提下,如何利用句法特征对连接词的两个论元进行识别.文献[19]实现了一个点到点的英文篇章分析器,其输入可以是自由文本,输出是标识了连接词、相应论元范围和篇章逻辑关系的篇章结构.

由于树核方法在度量结构化数据的相似度时更加有效,同时能够减少结构化数据的特征选择工程负担,它在自然语言处理中具有很多成功的应用,例如实体间语义关系检测和分类[20,21]、语义角色标注[22]、生物文献中的蛋白质关系抽取[23]、事件代词消解[24]等.基于此,本文拟采用树核方法研究隐式篇章关系,同时利用复合核将树核和基本核加以集成.

2 PDTB 简介

PDTB 是语言资源联盟(Linguistic Data Consortium,简称 LDC)(http://www.ldc.upenn.edu/)于 2008 年发布的.它由美国宾西法尼亚大学、意大利托里诺大学和英国爱丁堡大学联合标注,是目前规模最大的英文篇章级的语

料库.其对华尔街日报的2 159篇文章借鉴篇章词汇化树型连接语法(discourse lexical tree adjunct grammar,简称D-LTAG)理论[25]和RST思想,标注了40 600个显式、隐式、替代词汇化(alternative lexicalization,简称AltLex)、基于实体一致性关系(entity-based coherence,简称EntRel)和没有关系(no relation,简称NoRel)这5大类型的篇章关系.其中,D-LTAG 理论把连接词看成篇章谓词,连接词所携带的描述事件、事实和主张的两个文本跨度作为其两个论元(称为Arg1和Arg2),并规定Arg2在句法上依附于篇章连接词,而Arg1则不然.对于相邻的句对,如果不存在显式连接词,则可以进一步表示成以下3类:

(1) AltLex:篇章关系可以被人工推导出来,但是加入篇章连接词后会造成表达上的冗余.

(2) EntRel:不能推导出篇章关系,它是指第2个句子仅仅提供了第1个句子中相关实体的进一步信息.

(3) NoRel:相邻句对既不存在篇章关系又不存在基于实体的一致性.

同时,PDTB对显式、隐式和AltLex篇章关系定义了一个3级层次的意义(sense)结构:最上层是种类,第2层是类型,最下层是子类型.其中,第1层包括4种最常见的语义:Temporal,Contingency,Comparison和Expansion,第2层包括16类语义,第3层包括23类语义.另外,PDTB还标注了属性,它反映的是显式、隐式和AltLex关系的内部单个对象和抽象对象间以及它们参数的“拥有关系”,包括属性的类型、范围极性、确定性和跨度等信息.为了清晰起见,例1~例5列出了PDTB中各种类型的篇章关系实例(根据PDTB的标注方法,实例中的Arg1用斜体显示,Arg2用粗体显示,连接词用下划线显示,在各实例后的括号中表明了其对应的3层次式的Sense结构).

例1:显式实例

In addition, *its machines are typically easier to operate*, so **customers require less assistance from software**. (CONTINGENCY: Cause: result)

例2:隐式实例

Average maturity was as short as 29 days at the start of this year, *when short-term interest rates were moving steadily upward*. Implicit=For example **The average sevenday compound yield of the funds reached 9.62% in late April**. (EXPANSION: Instantiation)

例3:AltLex实例

The average pay of its clients fell to $66,743 last year from $70,765 in 1987; *severance pay dropped to 25 weeks from 29*, Outplacement consultant Right Associates says, **both reflect the dismissal of lower-level and short-tenure executives**. (CONTINGENCY: Cause: reason)

例4:EntRel实例

Hale Milgrim, *41 years old*, *senior vice president*, *marketing at Elecktra Entertainment Inc*., *was named president of Capitol Records Inc*., *a unit of this entertainment concerns*. EntRel **Mr.Milgrim succeeds David Berman**, **who resigned last month**.

例5:NoRel实例

Jacobs Engineering Group Inc's Jacobs International unit was selected to design and build a microcomputer-systems manufacturing plant in County Kildare, Ireland, for Intel Corp. *Jacobs is an international engineering and construction concern*. NoRel **Total capital investment at the site could be as much as $400 million**, **according to Intel**.

3 隐式篇章关系识别方法

本文仅针对PDTB中第1层的4大类隐式篇章语义关系进行识别.图1显示了相应的隐式篇章关系识别框架.我们首先从PDTB的两个论元对中抽取出词汇化、浅层语义信息、句子级的情感信息等不同类别的平面特征和基于依存树结构的简单、有效的结构化特征,将它们融入基于统计方法的分类器中,通过训练实例生成相应的分类器模型;然后,对测试实例也生成对应的平面和结构化特征;最后,对测试实例的两个论元所具有的篇章关系类别进行预测.

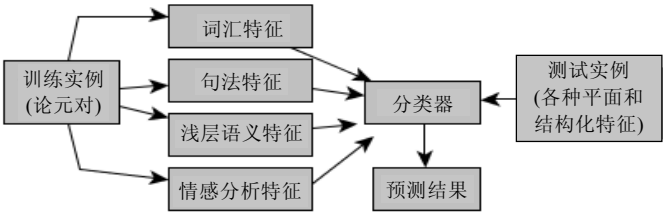

Fig.1 Framework of the system

图 1 系统框架

3.1 平面特征

本文所研究的平面特征主要包括词汇化特征、浅层语义特征和句子级的情感特征这 3 种类型.下面我们分别加以阐述.

3.1.1 词汇化特征

Pilter 等人[8]已经验证了 First-Last-First3 特征,即篇章关系的两个论元中的第 1 个单词、最后一个单词、第 1 个及最后一个单词的组合和前 3 个单词特征,对于隐式篇章关系识别非常有效.鉴于此,我们同样借签 First-Last-First3 特征,但不同于他们的工作之处在于,根据 Lin 等人[3]的统计,在 PDTB 中约有 2.2%的隐式篇章关系被标注具有多于 1 个 Sense 的情况,例如,可以同时标记为 Temporal 和 Expansion 两种 sense.为了有效地利用这方面的信息,我们在构造训练实例时把被标成两个 Sense 的篇章关系分别作为两个不同的训练实例,并称这种重复的词汇化特征为 First-Last-First3*.我们通过例 6 说明此类特征所对应的特征值(假设例 6 所对应的两个论元具有 Comparison 和 Expansion 两种关系).

例 6: (a) It is easy to say the specialist isn't doing his job.

(b) When the dollar is in a free-fall, even central banks can't stop it. (Comparison; Expansion)

其对应的 First-Last-First3*特征值为

- Comparison: It; job; When; it; It_is_easy; When_the_dollar; It_when; job_it.
- Expansion: It; job; When; it; It_is_easy; When_the_dollar; It_when; job_it.

3.1.2 浅层语义特征

语义角色标注(semantic role labeling,简称 SRL)是一种回答句子级别的 5W 问题(Who,When,What,Where 和 Why)的浅层语义分析形式[26].一般来说,几乎所有的自然语言处理任务都需要这种语义信息.但是到目前为止,并没有篇章关系识别的已有工作研究此类信息.我们预期句子中的这些语义角色对于隐式篇章关系识别具有重要的作用,例如,表示原因的角色可以推出两个论元间可能存在 Contingency 逻辑关系、表示时序的角色可以推出两个论元间可能存在 Temporal 逻辑关系.我们通过例 6(a)来解释此类浅层语义特征,它是对应某个特定的隐式篇章连接词的 Arg1.

我们将例 6(a)中的句子作为 SENNA(http://ml.nec-labs.com/senna/)软件的输入,可以得到表 1 所示的语义角色输出.表 1 的第 1 列是句子对应的单词本身,第 2 列指示了本句的谓词,第 3 列和第 4 列对给定谓词的语义角色进行标记:"B"代表对应的单词是语义角色开始节点,"I"代表对应的单词是语义角色内部节点,"E"代表对应的单词是语义角色结束节点,"O"代表对应的单词不属于任何语义角色,"S-V"代表对应的单词是谓词.这里的语义角色可以包括以下几大类:

(1) 核心语义角色(A0~A5):一般来说,A0 和 A1 分别代表施事者和受事者;A2~A5 是与谓词相关的,每个不同的谓词具有的 A2~A5 角色都不同.

(2) 附加角色(Arg-):通常来说,这些角色可以被任何动词所具有,Arg-ADV 代表一般性目的,Arg-CAU 代表原因,Arg-DIR 代表方向,Arg-DIS 代表篇章标示,Arg-EXT 代表范围,Arg-LOC 代表地点,Arg-MNR 代表方式,Arg-MOD 代表情态动词,Arg-NEG 代表否定标记,Arg-PNC 代表目的,Arg-PRD 代表预

言,Arg-REC 代表相互信关系,Arg-TMP 代表时间.

更详细的语义角色介绍参见文献[27].

Table 1 Semantic role of Example 6(a)

表 1 例 6(a)的语义角色

单词	谓词	谓词 say 对应的语义角色标记	谓词 doing 对应的语义角色标记
It	–	O	O
is	–	O	O
easy	–	O	O
to	–	O	O
say	say	S-V	O
the	–	B-A1	B-A0
specialist	–	I-A1	I-A0
isn	–	I-A1	I-A0
’t	–	I-A1	E-A0
doing	doing	I-A1	S-V
his	–	I-A1	B-A1
job	–	E-A1	E-A1
.	–	O	O

为了验证何种语义角色对隐式篇章关系识别起作用,我们利用开发集数据进行了相应的语义角色过滤实验.实验结果显示,当利用所有的语义角色时,系统性能最好.于是,我们抽取句子中的如下语义角色作为最终的浅层语义特征:

(1) S-V:代表谓词.

(2) 核心角色:如 A0,A1,A2,A3,A4,A5.

(3) 附加角色:如 Arg-LOC,Arg-MNR,Arg-TMP,Arg-NEG,Arg-MOD,Arg-DIS,Arg-EXT,Arg-ADV,Arg-PNC,Arg-PRD,Arg-CAU 和 Arg-REC.

对于例 6(a),对应的特征值为:say; the specialist isn ’t doing his job; the specialist isn ’t; doing; his job;同理,对于例 6(b),也可以抽取出相应的浅层语义特征.然后,我们将 Arg1 和 Arg2 分别对应的浅层语义特征进行组合,把它作为某个特定的隐式篇章连接词对应的浅层语义特征值.

3.1.3 情感特征

情感分析旨在研究文本单元(可以是词、句子、段落或整个篇章)所具有的正、负和中性等极性,在个性化产品推荐中存在较为广泛的应用.尽管文献[8]提到利用句子中的动词的极性这种特征,但是他们仅限于论元中有多少个正性和负性动词这种情况,并没有把整个句子的极性作为一种整体情况来考虑.我们预期句子所包含的情感极性对隐式篇章关系识别具有重要作用,可以通过例 7 来说明.

例 7: (a) Jack went to the hospital.

(b) He had broken his leg yesterday.

正如例 7 所示,例 7(a)和例 7(b)中间没有显式连接词存在,我们或许可以推测出作者对于例 7(a)和例 7(b)都具有负性情感.这种句子级的情感特征可以推测例 7(a)和例 7(b)具有 Contingency 这种篇章关系.

另外,根据功能语言学家 Martin 的评价理论[28]介绍,态度是作者对于文本中某个对象的主观人际意义,通常表现为褒义或贬义.篇章中的态度资源贯穿于文本中,并形成一种韵律结构,称为态度韵(attitude prosody,简称 AP).态度韵是评价理论的一个重要组成部分,对篇章的连贯性评估起着重要作用.按照 Martin 所述,篇章中通常包括 3 种类型的态度韵,分别是浸透型态度韵、加强型态度韵和控制型态度韵.其中,作者在使用浸透型态度韵时,可以使用某些可能的态度来表达他们的观念;作者在使用加强型态度韵时,可以使用重复和修辞风格来表达他们的观念;作者在使用控制型态度韵时,可以为事实型意义和它们的控制域间建立一种联系.作者对文本中每句话所表现出来的态度构成了整个文本的一种韵律结构.它能够维持篇章的一致性.鉴于此,我们认为句子级的情感特征与 Martin 的功能语言学理论——态度韵——存在着一定的关联,态度韵为句子级的情感特征提供了语言学基础,对篇章关系的识别将起一定的作用.

基于上述分析,我们利用句子级的情感分析技术对每个论元进行相应的情感分析,得到正、负和中性这 3 种值,然后将极性值的组合作为情感特征.为了获取每个论元的极性值,我们利用词袋(bag-of-words,简称 BOW)模型开发了相应的句子级的情感分析工具,采用文献[29]中的多领域情感分析数据,训练了 Mallet 中的最大熵分类器(http://mallet.cs.umass.edu/),然后利用分类器模型对测试数据进行预测.

表 2 列举了每个论元对所具有的极性值组合(表 2 中的第 1 列),通过对开发集数据上的实验分析,大多数的中性值对隐式篇章关系识别不仅不能起到区分作用,反而会给分类器带来部分噪音.于是,我们仅考虑每个论元所具有的正和负两种极性,忽略中性极性的相关组合,并把相应的[正,正]、[正,负]、[负,正]和[负,负]这 4 种组合作为最终的情感特征值(表 2 中的第 2 列).

Table 2 Feature value of sentence-level sentiment feature

表 2 句子级的情感特征值

特征值	是否最终采用特征值?
[正,正]	是
[正,负]	是
[正,中性]	否
[中性,正]	否
[中性,负]	否
[中性,中性]	否
[负,正]	是
[负,负]	是
[负,中性]	否

3.2 结构化特征

本节首先介绍本文提出的结构化特征,然后简要介绍卷积树核和复合核等内容.

3.2.1 结构化特征

文献[8]已经验证了两个论元间的词对特征对于隐式篇章关系识别非常有效,但是这种词对特征是一种平面特征.受此启发,我们探索了是否能够将此类词对平面特征表示成结构化特征的形式,并利用树核方法来进行隐式篇章关系识别.相应地,我们提出了如下两种非常简单、有效的树结构:

(1) All Dependency Tree (AllDT)

首先,我们利用 Stanford 句法分析器(http://nlp.stanford.edu/software/lex-parser.shtml)对每个论元(Arg1 和 Arg2)进行依存句法分析,对生成的依存句法树结构抽取每个单词间的上下层链接结构,生成一种单词依存树结构;然后,将两个论元分别对应的单词对依存树结构进行组合(通过加入上层节点完成),生成最终的 AllDT 树结构.

例 8 对应的 AllDT 结构如图 2 所示.其中,

- Arg1 对应的 AllDT 树结构为:(Arg1 (believes (He)(Plays (what)(he))(plays (he)(superbly)))).
- Arg 2 对应的 AllDT 树结构为:(Arg2 (appearance (His)(recent)(Museum (the)(Metropolitan)(dubbed (Odyssey (A)(Musical)))))).
- 而整个 Arg1 与 Arg2 组合后的 AllDT 结构为:(Arg1-Arg2 (Arg1 (believes (He)(Plays (what)(he))(plays (he)(superbly)))) (Arg2 (appearance (His)(recent)(Museum (the)(Metropolitan)(dubbed (Odyssey (A) (Musical))))))).

例 8: (a) He believes in what he plays, and he plays superbly.

(b) His recent appearance at the Metropolitan Museum, dubbed “A Musical Odyssey”.

(2) Skeleton Dependency Tree (SkeDT)

由于 AllDT 含有论元中的绝大多数单词,这必然存在两个弊端:(1) AllDT 结构比较宠大,需要更多的训练时间;(2) 过多的单词会引入相应的噪音.鉴于此,我们探索了是否存在一种结构,它一方面能够保存 AllDT 中的绝大多数有效信息,同时又能过滤掉一部分没有用的信息.于是,我们相应地提出了一种骨架依存树结构,这种骨

架依存树结构能够保存原有依存句法树中的谓词和与谓词有直接相关联的单词信息,这部分信息对于分类器具有更大的区分作用.

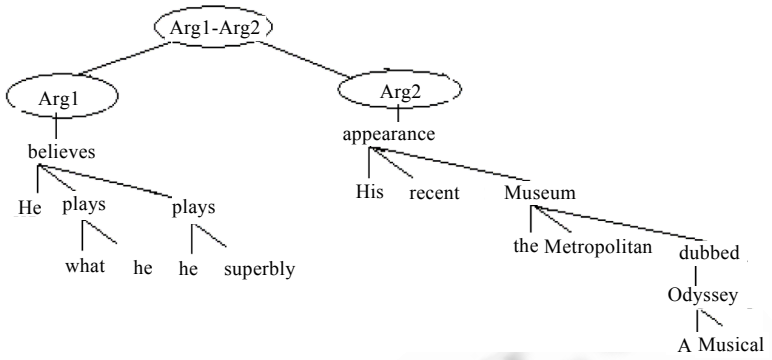

Fig.2 Tree structural of AllDT

图 2 AllDT 树结构

例 8 对应的 SkeDT 结构如图 3 所示.其中,

- Arg1 对应的 SkeDT 树结构为:(Arg1 (He)(plays)(plays)).
- Arg2 对应的 SkeDT 树结构为:(Arg2 (appearance (His)(recent)(Museum))).
- 而整个 Arg1 与 Arg2 组合后的 SkeDT 结构为:(Arg1-Arg2 (Arg1 (He)(plays)(plays))(Arg2 (appearance (His)(recent)(Museum)))).

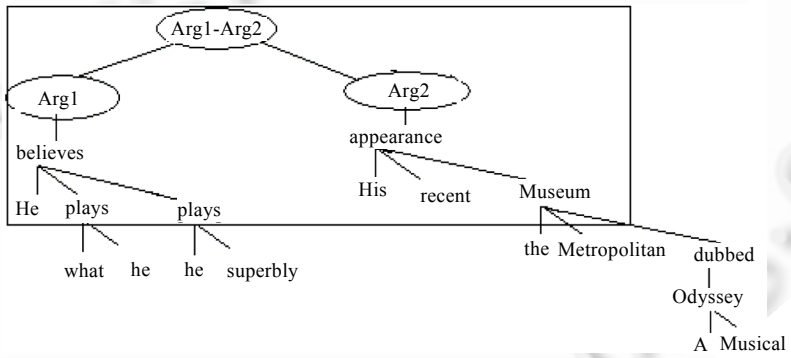

Fig.3 Tree structural of SkeDT

图 3 SkeDT 树结构

3.2.2 卷积树核

卷积树核[30,31]是一种用来计算两棵树结构中具有公共子树的相似度计算技术,其中,树 T 被隐式地表示成相应子树类型的整数向量(如公式(1)所示).

$$\phi(T)=(\#substree_1(T),\ldots,\#substree_n(T)) \tag{1}$$

其中,$\#substree_i(T)$代表 T 中第 i 种子树类型的出现次数.

例如,存在两棵句法树 T_1 和 T_2,我们可以利用公式(2)来计算它们的相似度 $Kc(T_1,T_2)$:

$$Kc(T_1,T_2)=(\phi(T_1),\phi(T_2))=\sum_{n_1\in N_1}\sum_{n_2\in N_2}\Delta(n_1,n_2) \tag{2}$$

其中,N_1 和 N_2 分别代表树 T_1 和 T_2 中的节点集合,$\Delta(n_1,n_2)$代表以 n_1 和 n_2 为根的公共子树个数.

此外,$\Delta(n_1,n_2)$可在多项式时间内按照下述递归规则来计算:

(a) 如果 n_1 和 n_2 节点处的产生式不同,则$\Delta(n_1,n_2)=0$.

(b) 否则,如果 n_1 和 n_2 均为叶子节点前的一个节点,则$\Delta(n_1,n_2)=1\times\lambda$;

(c) 否则,递归地按公式(3)计算$\Delta(n_1,n_2)$:

$$\Delta(n_1,n_2)=\lambda\times\prod_{j=1}^{nc(n_1)}(1+\Delta(ch(n_1,j),ch(n_2,j))) \tag{3}$$

其中,$nc(n_1)$代表 n_1 的孩子节点个数,$ch(n,j)$是节点 n 的第 j 个孩子节点,衰减因子$\lambda(0<\lambda\leqslant 1)$使得相应的核值不易随子树的大小而变化.

3.2.3 复合核

已有的大量工作表明,卷积树核能够捕获结构化信息,而基本核(线性核、多项式核、径向基核等)更能捕获平面信息.鉴于此,我们需要一种由两种不同核构成的复合核,然而,SVM(http://disi.unitn.it/moschitti/Tree-Kernel.htm)提供了一种$\lambda K_1+(1-\lambda)K_2$ 的复合核.这里,K_1 代表树核,K_2 代表基本核.为了验证上述提出的平面特征和结构化特征的有效性,我们利用复合核将其进行集成,其中,$\lambda(0\leqslant\lambda\leqslant 1)$是两种核所占的权重.

4 实验设置和结果分析

为了验证我们提出的平面特征和结构化特征的有效性,我们设计了相应的隐式篇章关系识别实验.本节首先介绍实验设置,然后详细分析实验结果.

4.1 实验设置

为了便于与文献[2]所取得的隐式篇章关系识别性能进行对比,我们采用了同样的 PDTB 训练和测试数据分割方式.其中,Section2-22 作为训练数据,Section23-24 作为测试数据,Section00-01 作为开发数据.对于平面特征,我们分别利用了 Mallet 中的最大熵分类器和 SVM 分类器进行实验.由于最大熵分类器取得的性能好于 SVM 分类器,所以我们只列出平面特征对应的最大熵分类器的实验结果.对于结构化特征,我们利用 SVM 分类器中的缺省参数,采用一对多策略[32]将多元分类转化为多个二元分类问题,并利用 Stanford 句法分析工具的“-outputFormat typeDependencies –outputFormatOptions treeDependencies”选项取得树型依存句法树风格(非层叠式风格)的依存树结构.我们利用 Accuracy 和 Precision,Recall 和 $F1$-measure 指标对系统的性能进行评测.

4.2 实验结果分析

我们训练了多个分类器分别用来验证单个平面特征和组合平面特征的性能.表 3 是对应的平面特征所取得的 Accuracy.第 1 列代表所采用的特征类型,第 2 列代表训练数据所具有的不同特征向量的大小,第 3 列代表测试数据所具有的不同特征向量的大小.

Table 3 Accuracy of the flat features

表 3 平面特征取得的准确率

特征	#训练	#测试	Accuracy (%)
浅层语义特征	36 464	3 909	50.89
First-Last-First3	53 485	6 067	47.51
First-Last-First3*	53 485	6 067	49.87
情感特征	4	4	51.39
浅层语义特征+First-Last-First3*	87 060	9 608	**52.00**
First-Last-First3*+情感特征	53 489	6 071	51.31
浅层语义特征+情感特征+First-Last-First3*	87 064	9 612	51.48

根据表 3 的实验数据,我们可以得到如下一些结论:

(1) 对于单个特征而言,情感特征取得的Accuracy要好于浅层语义分析和First-Last-First3所取得的性能,从而验证了我们的预期,即句子级的情感特征能够根据论元所具有的情感极性推断出两个论元间可能存在何种隐式篇章关系,也验证了功能语言学家 Martin 的态度韵可行性和实效性.

(2) 改进后的First-Last-First3*性能要好于原始的First-Last-First3(卡方显著性检测对应的*p-value*<0.01),原因在于,这种具有冗余词汇特征的训练实例对于分类器的决策也具有一定的指导作用.

(3) 对于复合特征,当把浅层语义特征和 First-Last-First3*进行组合时,我们的平面特征可以得到 52.00%的 Accuracy.这样也验证了比较粗粒度的浅层语义特征与比较细粒度的词汇化特征对于隐式篇章关系识别具有一定的互补作用.

表 4 是树核和复合核方法所取得的 Accuracy,我们可以明确:

(1) SkeDT 树结构的性能要好于 AllDT,原因在于,SkeDT 保留了 AllDT 中的关键结构信息,既保留了谓词和与其有直接关联的单词链接信息,同时也过滤了 AllDT 中的部分没有用的信息,从而减少了部分噪音,而且 SkeDT 与 AllDT 相比能够有效减少 30%的训练时间.

(2) SkeDT 性能也要好于平面特征所取得的性能,说明了结构化特征对于隐式篇章关系识别的重要性.与文献[2]所提出的树核方法相比,他们把 PDTB 中同一个段落内的每 3 句 Golden 句法树合并成一棵树结构,一方面,他们的 Golden 句法树的抽取工作比较耗时,而且在现实应用情况下的大量测试数据,我们一般很难得到它们的 Golden 句法树结构(手工标注成本非常昂贵);另一方面,他们的段落树结构也比较宠大,需要更多的训练和预测时间.而我们仅把论元作为 Stanford 句法分析器的输入,然后对生成的依存树做简单的转换即可.

(3) 情感特征+SkeDT 复合核最终取得了 53.10%的 Accuracy,间接地反映了 Martin 的态度韵理论的可行性.表 4 中第 6 行、第 7 行及第 9 行~第 12 行对应的性能要低于第 8 行的性能,我们认为,可能的原因在于:一方面,情感特征与结构化特征结合后更加有效;另一方面,当把所有的平面特征与结构化特征进行复合时,由于结构化特征中已经包括了部分词汇化信息,从而会对分类器引入部分噪音.

Table 4 Accuracy of the tree kernel and composite kernel

表 4 树核和复合核取得的准确率

方法		Accuracy (%)
树核	AllDT	51.14
	SkeDT	52.66
复合核	SkeDT+浅层语义特征	51.90
	SkeDT+First-Last-First3*	52.83
	SkeDT+情感特征	**53.10**
	SkeDT+浅层语义特征+First-Last-First3*	51.65
	SkeDT+浅层语义特征+情感特征	52.15
	SkeDT+情感特征+First-Last-First3*	52.57
	SkeDT+浅层语义特征+情感特征+First-Last-First3*	51.56

表 5 是我们系统取得最好 Accuracy 情况下(SkeDT+情感特征复合核)对应 4 大类别详细的 Precision,Recall 和 *F*1-measure("–"代表 0.00).从表 5 的数据我们可以明确:

(1) Contingency 和 Comparison 两个类别所取得的 Precision 要高于 Recall,这一方面说明了我们提出的复合核检测出来实例的正确率比较高;另一方面,较低的 Recall 也间接说明了探索更多有效的语言学特征的必要性.

(2) 根据我们对 PDTB 测试数据集的统计,4 大类 Temporal,Contingency,Comparison 和 Expansion 所占比例分别为 3.29%,28.61%,17.56%,50.54%.我们所提出的特征在这 4 大类所取得的 Precision,Recall 和 *F*1- measure 也同样符合这种数据分布.

(3) Temporal 类别的 Precision,Recall 和 *F*1-measure 均为 0.00,一方面,可能的原因在于,Temporal 类别中训练语料的缺乏(仅占 5.98%),导致了分类器不能识别出 Temporal 占比例较小的类别;另一方面,也说明了研究专门与时序相关特征的必要性,例如篇章关系论元中的时态信息等特征.

(4) Expansion 类别的 Recall 要高于 Precision,以及 Contingency 和 Comparison 类别的 Precision,Recall 和 *F*1-measure 比 Expansion 类别要低,从侧面也说明了训练数据内部的不平衡性对最终的性能也会

产生一定的影响.因此,研究与时序有关的语言学特征和多类别的隐式篇章关系识别的数据不平衡性,都将是我们的将来工作之一.

Table 5 Precision,Recall and *F*1-measure of 4 classes using the composite kernel SkeDT+Sentiment

表 5 SkeDT+情感复合核下 4 大类别的 Precision,Recall 和 *F*1-measure

对应的 4 大类别	Precision (%)	Recall (%)	*F*1-measure
Temporal	–	–	–
Contingency	47.62	17.70	25.81
Comparison	50.00	4.33	7.96
Expansion	52.74	91.62	66.95

鉴于 SkeDT+情感特征复合核不能识别占比例极小的 Temporal 类别(仅占测试语料的 3.29%)问题,本文提出一种折中方案:在适当保证 4 大类总体性能前提下,我们的系统可以适当提升 Temporal 类别的识别性能.由于 Temporal 类别更倾向于采用词汇化的表达方式,例如,“Last year”,“This year”,所以利用我们提出的平面特征对 Temporal 类别进行识别可能更加有效.表 6 是利用 SVM 取得的 Temporal 类别 Precision,Recall 和 *F*1-measure.我们可以明确:

(1) First-Last-First3*、情感特征和浅层语义特征等平面特征比结构化特征更适合 Temporal 类别的识别.其中,First-Last-First3*特征取得了最好的 *F*1-measure.可能的原因在于,与时序有关的驱动词,例如“Now”,“Monday”,“Today”等,经常出现在句子的头尾地方.

(2) 情感特征也取得了较好的 Temporal 识别性能,其 Recall 为 97.44%.可能的原因在于,情感的表达方式通常会采用同步或异步两种方式,而 PDTB 的标注准则也明确表明了 Temporal 类别的体现方式主要采用同步和异步两种方式,两者比较吻合.

(3) 所有平面特征所取得的 Precision 和 Recall 都不太高,表明了需要挖掘出对 Temporal 类别的识别起更大作用的语言学特征的必要性,同时也说明了训练语料的不平衡性也是隐式篇章关系识别中需要着重研究的一个课题.

Table 6 Precision, Recall and *F*1-measure of Temporal class using the flat features

表 6 平面特征下 Temporal 类别的 Precision,Recall 和 *F*1-measure

特征	Precision (%)	Recall (%)	*F*1-measure
First-Last-First3*	8.80	15.40	**11.20**
浅层语义特征	1.54	2.56	1.92
情感特征	3.48	97.44	6.72
First-Last-First3*+情感特征	2.78	5.13	3.61
First-Last-First3*+浅层语义特征	1.64	2.56	2.00

4.3 与现有系统对比

表 7 显示了我们的系统与现有最新的树核方法、平面特征方法和人工标注的性能对比结果.其中对测试数据的人工重新标注由专门研究篇章关系识别问题的两个高年级的研究生完成,他们平均仅取得了 60%左右的 Accuracy.表 7 中的最大类基准系统代表当把测试数据的实例都分给训练数据中最大的那一类(expansion)时所对应的类别,它可以取得 50.50%的 Accuracy.从表 7 的数据我们可以看出:

(1) 现有系统和人工标注的性能都不高,这也从侧面说明了隐式篇章关系识别任务的挑战性.

(2) 文献[2,7,8]所取得的性能均低于最大类基准系统,也充分说明了隐式篇章关系识别这一任务的难度.

(3) 我们最好的系统取得的 53.10%的 Accuracy 显著优于最大类基准系统(卡方显著性检测对应的 *p-value*<0.05),也显著优于文献[2,7,8]中的系统所取得的性能(卡方显著性检测对应的 *p-value*<0.001).

(4) 我们系统的平面特征所取得的性能均显著优于文献[2,7,8]中的系统所取得的性能(卡方显著性检测对应的 *p-value*<0.001),也显著优于最大类基准系统(卡方显著性检测对应的 *p-value*<0.05).

Table 7 Performance comparison among our system, the-state-of-art and human annotation systems

表 7 我们系统与现有系统和人工标注的性能对比

系统	Accuracy (%)
Wang 等人(文献[2],复合核方法)	40.00
Zhou 等人(文献[7],语言模型方法)	41.35
Pilter 等人(文献[8],平面特征方法)	47.51
最大类基准系统	50.50
我们的系统(平面特征)	52.00
我们的最好系统(复合核)	**53.10**
人工标注者 1	60.20
人工标注者 2	59.70

5 结论和未来的工作

本文研究了 PDTB 中的 4 大类别(temporal,contingency,comparison 和 expansion)的隐式篇章关系识别问题.为使读者能够重现我们的方法,本工作所采用的软件和数据集都是可以公开下载或获取的资源.

本文系统地研究了篇章中的浅层语义信息和句子级的情感信息等平面特征对于隐式篇章关系识别的作用,以功能语言学家 Martin 的态度韵理论出发,深入探索了篇章中的情感演化规律;同时,提出了一种简单、有效的骨干词汇依存树结构;最后,采用复合核将平面与结构化特征加以集成.PDTB 2.0 上的实验结果显示了这种新型平面和结构化特征的重要性和有效性.与现有的最新系统相比,我们提出的平面特征方法、树核方法和复合核方法在性能上均有显著提升.在 Accuracy 上,分别比已有最新的复合核方法、语言模型方法和平面特征方法高出约 13.1%,11.8%和 5.6%,比最大类基准系统也高出 2.6%.这些实验结果充分验证了我们提出的平面特征和结构化特征的可行性.

关于将来的工作,一方面,我们将研究训练数据不平衡性对于隐式篇章关系识别的影响;另一方面,我们将探索篇章模式(discourse patterning)和框架理论(frame theory)等功能语言学理论对于提高隐式篇章关系识别性能的作用.

致谢 在此,我们向对本研究工作提供帮助的老师和同学表示感谢.同时,我们也向对本文提出宝贵意见的评审专家表示感谢.

References:

[1] Pitler E, Nenkova A. Using syntax to disambiguate explicit discourse connectives in text. In: Proc. of the ACL-IJCNLP 2009. Stroudsburg: Association for Computational Linguistics, 2009. 13−16.

[2] Wang WT, Su J, Tan CL. Kernel based discourse relation recognition with temporal ordering information. In: Proc. of the ACL 2010. Uppsala: Association for Computational Linguistics, 2010. 710−719.

[3] Lin ZH, Kan MY, Ng HT. Recognizing implicit discourse relations in the penn discourse treebank. In: Proc. of the EMNLP 2009. Stroudsburg: Association for Computational Linguistics, 2009. 343−351.

[4] Carlson L, Marcu D, Okurowski ME. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In: Proc. of the SIGDIAL. Stroudsburg: Association for Computational Linguiscs, 2001. 1−10. [doi: 10.3115/1118078.1118083]

[5] Prasad R, Miltsakaki E, Dinesh N, Lee A, Joshi A, Robaldo L, Webber B. The penn discourse treebank 2.0 annotation manual. Technical Report, IRCS-08-01, Philadelphia: University of Pennsylvania, 2008. 1−99.

[6] Wolf F, Gibson E. Representing discourse coherence: A corpus-based analysis. Journal of Computational Linguistics, 2005,31(2): 249−288. [doi: 10.1162/0891201054223977]

[7] Zhou ZM, Xu Y, Niu ZY, Lan M, Su J, Tan CL. Predicting of discourse connectives for implicit discourse relation recognition. In: Proc. of the COLING 2010. Beijing: Association for Computational Linguistics, 2010. 1507−1514.

[8] Pitler E, Louis A, Nenkova A. Automatic sense prediction for implicit discourse relations in text. In: Proc. of the ACL-IJCNLP 2009. Stroudsburg: Association for Computational Linguistics, 2009. 683−691.

[9] Zhou ZM, Lan M, Niu ZY, Xu Y, Su J. The effects of discourse connectives prediction on implicit discourse relation recognition. In: Proc. of the SIGDIAL 2010. Tokyo: Association for Computational Linguistics, 2010. 139−146.

[10] Hernault H, Bollegala D, Ishizuka M. A semi-supervised approach to improve classification of infrequent discourse relations using feature vector extension. In: Proc. of the EMNLP 2010. Massachusetts: Association for Computational Linguistics, 2010. 399−409.

[11] Tofiloski M, Brooke J, Taboada M. A syntactic and lexical-based discourse segmenter. In: Proc. of the ACL-IJCNLP 2009. Suntec: Association for Computational Linguistics, 2009. 77−80.

[12] Hernault H, Bollegala D, Ishizuka M. A sequential model for discourse segmentation. In: Proc. of the CICLing 2010. Berlin, Heidelberg: Springer-Verlag, 2010. 315−326. [doi: 10.1007/978-3-642-12116-6_26]

[13] Soricut R, March D. Sentence level discourse parsing Using syntactic and lexical information. In: Proc. of the NAACL 2003. Stroudsburg: Association for Computational Linguistics, 2003. 149−156. [doi: 10.3115/1073445.1073475]

[14] LeThanh H, Abeysinghe G, Huyck C. Generating discourse structures for written texts. In: Proc. of the COLING 2004. Stroudsburg: Association for Computational Linguistics, 2004. 329−335. [doi: 10.3115/1220355.1220403]

[15] DuVerle DA, Prendinger H. A novel discourse parser based on support vector machine classification. In: Proc. of the ACL-IJCNLP 2009. Stroudsburg: Association for Computational Linguistics, 2009. 665−673.

[16] Pustejovsky J, Havasi C, Rumshisky A, Sauri R. Classification of discourse coherence relations: An exploratory study using multiple knowledge sources. In: Proc. of the SIGdial Workshop on Discourse and Dialogue. Stroudsburg: Association for Computational Linguistics, 2006. 117−125.

[17] Huang HH, Chen HH. Chinese discourse relation recognition. In: Proc. of the IJCNLP 2011. Chiang Mai: Natural Language Processing of the Asian Federation, 2011. 1442−1446.

[18] Ghosh S, Johansson R, Riccardi G, Tonelli S. Shallow discourse parsing with conditional random fields. In: Proc. of the IJCNLP 2011. Chiang Mai: Natural Language Processing of the Asian Federation, 2011. 1071−1079.

[19] Ghosh S, Tonelli S, Riccardi G, Johansson R. End-to-End discourse parser evaluation. In: Proc. of the ICSC 2011. Palo Alto: Institute of Electrical and Electronics Engineers, 2011. 169−172. [doi: 10.1109/ICSC.2011.40]

[20] Zhou GD, Zhu QM. Kernel-Based semantic relation detection and classification via enriched parse tree structure. Journal of Computer Science and Technology, 2011,26(1):45−56. [doi:10.1007/s11390-011-1110-2]

[21] Zhou GD, Qian LH, Fan JX. Tree kernel-based semantic relation extraction with rich syntactic and semantic information. Journal of Information Science, 2010,180(8):1313−1325. [doi:10.1016/j.ins.2009.12.006]

[22] Zhou GD, Li JH, Fan JX, Zhu QM. Tree kernel-based semantic role labeling with enriched parse tree structure. Journal of Information Processing and Management, 2011,47(2011):349−362. [doi:10.1016/j.ipm.2010.08.005]

[23] Qian LH, Zhou GD. Tree kernel-based protein-protein interaction extraction from biomedical literature. Journal of Biomedical Informatics, 2012,45(2012):535−543. [doi:10.1016/j.jbi.2012.02.004].

[24] Kong F, Zhou GD. Improve tree kernel-based event pronoun resolution with competitive information. In: Proc. of the IJCAI 2011. Barcelona: AAAI Press, 2011. 1814−1819.

[25] Webber B. D-LTAG: Extending lexicalized TAG to discourse. Journal of Cognitive Science, 2004,28(5):751−779. [doi: 10.1207/s15516709cog2805_6]

[26] Li JH, Zhu GD, Zhu QM, Qian PD. Semantic role labeling in Chinese language for nominal predicates. Ruan Jian Xue Bao/Journal of Software, 2011,22(8):1725−1737 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/3885.htm [doi: 10.3724/SP.J.1001.2011.03885]

[27] Palmer M, Gildea D, Kingsbury P. The proposition bank: An annotated corpus of semantic roles. Journal of Computational Linguistics, 2005,31(1):71−106. [doi: 10.1162/0891201053630264]

[28] Martin JR, White PRR. The Language of Evaluation: Appraisal in English. London, New York: Palgrave Macmillan, 2005. 1−291.

[29] Blitzer J, Dredze M, Pereira F. Biographies, Bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In: Proc. of the ACL 2007. Prague: Association for Computational Linguistics, 2007. 440−447.

[30] Collins M, Duffy N. Convolution kernels for natural language. In: Proc. of the NIPS 2001. Vancouver: MIT Press, 2001. 625−632.

[31] Moschitti A. A study on convolution kernels for shallow semantic parsing. In: Proc. of the ACL 2004. Stroudsburg: Association for Computational Linguistics, 2004. 335−342. [doi: 10.3115/1218955.1218998]

[32] Ghanem AS, Venkatesh S, West G. Multi-Class pattern classification in imbalanced data. In: Proc. of the ICPR 2010. Istanbul: IEEE Computer Society, 2010. 2881−2884. [doi: 10.1109/ICPR.2010.706]

附中文参考文献:

[26] 李军辉,周国栋,朱巧明,钱培德.中文名词性谓词语义角色标注.软件学报,2011,22(8):1725−1737. http://www.jos.org.cn/1000-9825/3885.htm [doi: 10.3724/SP.J.1001.2011.03885]

徐凡(1979−),男,江西上饶人,博士生,主要研究领域为自然语言处理,中文信息处理.

E-mail: 20104027010@suda.edu.cn

朱巧明(1963−),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为自然语言处理,中文信息处理,Web 信息处理和嵌入式系统.

E-mail: qmzhu@suda.edu.cn

周国栋(1967−),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为自然语言处理,信息抽取,统计机器翻译,机器学习.

E-mail: gdzhou@suda.edu.cn