

面向自然语言文本的否定性与不确定性信息抽取^{*}

邹博伟, 钱忠, 陈站成, 朱巧明, 周国栋



(苏州大学 计算机科学与技术学院, 江苏 苏州 215006)

通讯作者: 周国栋, E-mail: gdzhou@suda.edu.cn, http://nlp.suda.edu.cn/~gdzhou/

摘要: 目前,信息抽取研究主要面向肯定性信息,而自然语言文本中包含了大量否定性和不确定性信息,为了将此类信息与肯定性信息区分开,有必要针对否定性与不确定性信息抽取进行深入研究.针对这一任务,首次构建了一个16 841句的汉语语料资源,利用序列标注模型与卷积树核模型,系统地探索了各种序列化依存特征和结构化句法树特征的有效性,并提出了元决策树模型,对二者进行融合.实验结果显示,该方法在否定性和不确定性信息抽取任务上的精确率分别达到69.84%和58.57%,为相关研究打下了坚实的基础.

关键词: 信息抽取;否定性信息;不确定性信息;线索词检测;覆盖域界定
中图法分类号: TP391

中文引用格式: 邹博伟,钱忠,陈站成,朱巧明,周国栋.面向自然语言文本的否定性与不确定性信息抽取.软件学报,2016,27(2): 309-328. <http://www.jos.org.cn/1000-9825/4860.htm>

英文引用格式: Zou BW, Qian Z, Chen ZC, Zhu QM, Zhou GD. Negation and uncertainty information extraction oriented to natural language text. Ruan Jian Xue Bao/Journal of Software, 2016,27(2):309-328 (in Chinese). <http://www.jos.org.cn/1000-9825/4860.htm>

Negation and Uncertainty Information Extraction Oriented to Natural Language Text

ZOU Bo-Wei, QIAN Zhong, CHEN Zhan-Cheng, ZHU Qiao-Ming, ZHOU Guo-Dong

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

Abstract: The current research on information extraction mainly focuses on affirmative information. However there are more negation and uncertainty information in natural language texts. For purpose of separating them from affirmative information, it is necessary to make an intensive study of negation and uncertainty information extraction. For this task, this study firstly constructs a Chinese corpus including 16 841 sentences. Employing the sequence labeling model and the convolution tree kernel model, it systematically explores the efficiency of various kinds of serialized dependency features and structured parsing features. Finally, it proposes a meta-decision tree model to integrate the above two models. Experimental results show that the performances of the new method on negation and uncertainty information extraction achieve 69.84% and 58.57% of accuracy respectively, providing a solid foundation for related studies in the future.

Key words: information extraction; negation information; uncertainty information; cue detection; scope resolution

随着信息抽取技术的发展,越来越多的相关应用已能较为准确地从海量自然语言文本数据中获取各类所需信息,然而却并未对信息表述的否定性和不确定性进行甄别.如果所获取的信息来自于包含否定、猜测、假设等非事实类型的表述,则将导致该信息或知识的价值大幅度降低,甚至可能获得与真实情况完全相反的信息.基于此,为了抽取出此类否定性或不确定性信息及表述,并将其与事实信息相分离,面向自然语言文本的否定性与不确定性信息抽取(negation and uncertainty information extraction)任务应运而生.

* 基金项目: 国家自然科学基金(61272260, 61331011, 61273320)

Foundation item: National Natural Science Foundation of China (61272260, 61331011, 61273320)

收稿时间: 2015-01-30; 修改时间: 2015-04-14; 采用时间: 2015-05-08; jos 在线出版时间: 2015-11-12

CNKI 网络优先出版: 2015-11-11 17:00:18, <http://www.cnki.net/kcms/detail/11.2560.TP.20151111.1700.002.html>

否定性信息在认知学、哲学和语言学等领域中均有各自的定义和描述,本文中,我们借鉴 Morante 和 Sporleder 在其综述中给出的定义^[1]:否定性信息包括一个运算符以及若干受其影响并发生含义反转的要素.而针对不确定性信息的定义,由于较难给出其形式化范畴,因此,学者们尝试从不同的角度给出了描述性定义,其中包括情态(modality)^[2]、言据性(evidentiality)^[3]、不确定性(uncertainty)^[4]和主观性(subjectivity)^[5]等方面,本文中,对不确定性信息的定义主要借鉴 Lakoff 的描述^[6]:任何不确定性命题或信息,从认知程度上介于肯定信息和否定信息之间,该信息的整体或部分具有或然性.

否定性与不确定性信息抽取研究起源于面向生物医学科技文献的信息抽取任务中^[7-9],目的是将文献中的否定性与不确定性信息过滤掉,以避免文本挖掘和知识生成等应用系统将错误或存疑信息与真实信息混淆.否定性与不确定性信息抽取是信息抽取和语义分析相结合所产生的一项新任务,其在自然语言处理的各个领域具有广泛的应用前景,如情感分析^[10,11]、信息检索^[12]、文本分类^[13]、机器翻译^[14]等领域.

目前,否定性与不确定性信息抽取研究主要包括两个子任务:

- 1) 线索词检测(cue detection).线索词指能够标识出否定或不确定语义的词或短语,有些文献中也称触发词(trigger).在例句 1 和例句 2 中,以下划线表示的“不”和“有望”分别是肯定线索词和不确定线索词.
- 2) 覆盖域界定(scope resolution).覆盖域是指线索词的语义作用范围,通常是句子中的某一连续片段(例句中以方括号表示).在例句 1 中,否定线索词“不”否认了“(所有住客)会追究酒店的这次管理失职”这一命题,而“所有住客均表示”则是实际发生的事情,故不在否定线索词的覆盖域之内.

例句 1:所有住客均表示[不会追究酒店的这次管理失职].

例句 2:尽管上周五沪指盘中还受创业板的下跌所拖累,但[明天创业板仍有有望反弹].

英语中的否定性与不确定性信息抽取研究已逐渐成为自然语言处理领域的热点,并取得了大量成果;然而,汉语中的相关研究仍处于探索阶段.主要归结为两方面原因.

- 首先,面向汉语的否定性与不确定性信息抽取研究缺乏公认的语料资源.语料库建设是开展自然语言处理研究的重要基础,英语中有 BioScope 生物医学文献语料库^[15]等,而目前尚未有研究机构发布汉语语料库;另一方面,英语中的相关研究已开展了一系列较为成熟的评测,例如 BioNLP'2009^[16], CoNLL'2010^[17], *SEM'2012^[18]等,而面向汉语的相关评测尚未开展.
- 第 2 个原因是,英语语法结构相对严格,而汉语语法结构则较为松散,这使得无法将英语中否定性与不确定性信息抽取方法和技术直接移植到汉语上来,需要有针对性地采用适应汉语特点的相关自然语言处理技术.

鉴于以上困难和缺陷,本文首先构建了面向汉语自然语言文本的否定性与不确定性信息抽取语料库,该语料库在科技文献、金融新闻、酒店评论这 3 种类型的文本上标注了否定与不确定线索词及其覆盖域,其规模达到 16 841 句,能够客观反映汉语中否定性与不确定性表述的语言现象,为相关研究提供资源支持;其次,本文系统地探索了序列化依存特征和结构化句法树特征的有效性,通过利用序列标注模型与卷积核模型的互补性特点,采用元决策树模型对二者进行融合,并提出基于依存句法规则的后处理策略,实现了一套面向自然语言文本的否定性与不确定性信息抽取方法.实验结果表明:该方法能够有效地抽取出否定性和不确定性信息,在本文构建的语料库上,否定性与不确定性信息抽取的精确率分别达到 68.84%和 58.57%.

本文第 1 节介绍否定性与不确定性信息抽取的相关工作.第 2 节介绍本文构建的汉语否定性与不确定性信息抽取语料库.第 3 节着重阐述本文提出的否定性与不确定性信息抽取方法.第 4 节介绍实验设置,给出实验结果并进行详细地分析.第 5 节给出本文的结论,并对未来的工作进行展望.

1 相关工作

否定性与不确定性信息抽取研究主要围绕线索词检测和覆盖域界定两个子任务展开.早期研究大多采用基于启发式规则的方法识别线索词和覆盖域,其优势在于实现简单并且准确率较高,而缺点是健壮性和可扩展性较差.后来,随着语料库的标注和发布,研究者们开始逐渐尝试借助机器学习算法,提出了各类自然语言处理

模型来识别否定性与不确定性信息.本节分别介绍线索词检测和覆盖域界定的相关研究,最后介绍汉语的否定性与不确定性信息抽取研究现状.

线索词检测的相关研究主要包括:

- 1) 基于词表的方法.该方法的关键在于如何构建高质量否定或不确定线索词表,例如,Kilicoglu 等人^[19]借助 WordNet 和生物医学的专业词表中的不确定概念和事实概念的语义和词法关系,抽取文本中的不确定线索词.
- 2) 基于分类的方法.该方法以词或语块为样本,筛选和融合了各种有效的句法或语义特征,采用分类方法标识线索词.其中具有代表性的是 Özgür 等人^[20]和 Øvreid^[21]等人的工作,均探索了各类词法和句法特征,其中包括词干、词性、位置、邻接词、依存关系、句法范畴等.然而,由于一段文本中线索词的比重极小,因此,训练数据的不平衡性成为影响该方法性能的关键因素.
- 3) 基于序列标注的方法.以单个词作为分类对象,很难有效利用上下文信息,而序列标注模型能够利用标签之间的结构信息,在线索词检测任务中获得比传统分类方法更好的性能.Tang 等人^[22]采用条件随机场模型在 BioScope 语料上取得了最好的性能, F 值为 86.4%.

覆盖域界定任务的目标是识别线索词的语义作用范围.早期的覆盖域界定研究通常采用基于句法树的启发式规则方法,此类方法较好地利用了整体或局部句法结构的特点,尤其是语料符合语法时,该方法能够获得较高的准确率.基于启发式规则方法的覆盖域界定研究最突出的问题在于其可扩展性差,目前研究通常采用基于机器学习的方法.例如,Morante 等人^[23]采用基于存储的 k 最近邻分类器,并在浅层句法特征的基础上加入了依存句法特征,取得了 CoNLL'2010 覆盖域界定任务最好性能, F 值为 57.3%.Vlachos 等人^[24]以候选词与线索词之间的依存路径特征,采用贝叶斯逻辑回归分类器识别覆盖域.

否定性与不确定性信息抽取研究在面向英语的自然语言处理领域中已逐渐成为研究热点,而汉语中同样存在大量否定与不确定语言现象^[25,26],因此,开展汉语中的否定性与不确定性信息抽取研究具有重要意义.由于汉语与英语在语法结构及语义表达等方面存在差别,如果直接将面向英语的否定性与不确定性信息抽取方法及特征应用于汉语,则会导致其性能急剧下降.限制面向汉语否定性与不确定性信息抽取研究进展的另一重要因素是,尚未有相关语料资源发布以及缺少对应的评测任务.

2 语料

自然语言处理研究通常依赖于相关语料库的建立,英语的否定性与不确定性信息抽取研究在语料建设方面已取得初步成果,并发布了具有一定规模的语料库,例如 BioScope 语料库^[15]、Conan Doyle 语料库^[27]等,然而汉语的语料库构建研究则较为匮乏.基于此,本文构建了汉语否定性与不确定性信息抽取语料库,考虑到不同领域资源在语言特点上的差别,该语料库包括 3 个子语料库,分别是:科技文献语料(tech),选自《计算机学报》2012 年第 11 期的 19 篇论文;财经文章语料(finan),来源于新浪股市“股市及时雨”板块 2013 年 4 月面向沪深股市的所有 311 篇评论文章;酒店评论语料(review),爬取自携程网 500 家酒店的评论页面.

汉语否定性与不确定性信息抽取语料库参照 BioScope 语料库的标注规范,由标注者根据自己对语言材料的理解进行标注,标注内容包括线索词和覆盖域.其中,根据语料库和汉语言的自身特点,本文对标注方法及规则进行了限制和调整:

1) BioScope 语料库标注规范采用句法作用范围代替了语义作用范围来表示覆盖域,其好处是使问题得以简化,能够借助句法分析工具来辅助人工标注,并且由于 BioScope 语料库均为科技文献,句法相对工整严谨,对标注结果影响不大.然而,由于汉语语法运用较为灵活,尤其是在酒店评论语料中,文本短、省略多、表达方式多样化突出,因此,我们没有采用句法分析的结果辅助人工标注.

2) 汉语中的一些常见修辞方式,如排比、省略等,往往会造成主语与其他句子成分相分离.例如例句 3 中,覆盖域“不能多给我们提供一个枕头”省略了主语“客房”,若使语义尽量完整,则覆盖域需要包含整个句子的逻辑主语“客房”,这样就极大地提升了覆盖域界定任务的难度.该问题涉及汉语省略及恢复研究,不属于覆盖域界

定研究的主要内容.因此,本文在标注规则中规定,覆盖域必须是包含线索词的一段连续文本.

例句 3:客房有高档的配套设施,甚至是冲浪浴缸,然而却[不能多给我们提供一个枕头],...

3) 涉及否定或不确定语义的修辞方式,如双重否定等,要求根据意义和语境来判断.

汉语否定性与不确定性信息抽取语料库包括 3 个子语料库,共 16 841 句,标注了 6 429 个否定与不确定实例(一个实例包含一个线索词及其对应的覆盖域).本文构建的汉语语料库规模与 BioScope 语料库(20 924 句)规模接近,语料类型则更为多样化.标注工作由 3 位在该领域具备 2 年以上研究经验的标注者共同完成,其中两位标注者分别标注原始语料,第 3 位标注者统一标注结果,并形成最终语料.表 1 是对语料库相关数据的统计.

Table 1 Statistics of Chinese negation and uncertainty information extraction corpus

表 1 汉语否定性与不确定性信息抽取语料库数据统计

统计项	Tech		Finan		Review	
	否定	不确定	否定	不确定	否定	不确定
#句子	4 630		7 213		4 998	
句子平均长度(字)	30.4		30.7		24.1	
实例占比(%)	13.2	21.6	17.5	30.5	52.9	22.6
#线索词集合	44	154	201	280	128	179
覆盖域平均长度(字)	9.1	12.3	7.2	15.0	5.1	6.9

从“实例占比”数据中可以看出:(1) 科技文献语料中的否定和不确定实例的比例与 BioScope 语料(13.8%和 22.3%)接近,说明该语言现象在汉语中也同样常见;(2) 与科技文献语料相比,财经文章语料中包含了更多的否定和不确定实例;(3) 而酒店评论语料中,否定语义占比达到 52.9%,主要是因为酒店评论中包含大量评论者的负面观点或意见,这也说明了否定性与不确定性信息抽取研究在情感分析和意见挖掘领域的重要性.“覆盖域平均长度”数据表明,不确定覆盖域通常比否定覆盖域长;而在酒店评论语料中,否定和不确定覆盖域却均较短.造成这一现象的原因是该语料表达方式较为口语化,语言的随意性和多样性更为突出.

表 2 统计了各语料中出现频率最高的 4 个线索词所占比例.可以看出:否定线索词分布相对集中,而不确定线索词分布则较为分散.此外,在否定语料中,出现频率前两个词均为“不”和“没有”,尤其是在 Tech 和 Review 语料中,该比例更是高达 50%以上;而高频的不确定线索词在 3 个语料中有较大区别.

Table 2 Proportion of high frequency cues (%)

表 2 高频线索词占比(%)

语义	Tech		Finan		Review	
否定	不	38.9	不	13.9	不	41.6
	没有	13.4	没有	10.3	没有	18.5
	无法	7.0	无	4.3	没	10.3
	非	6.7	未	4.2	不是	4.1
不确定	如果	10.0	将	9.1	感觉	10.4
	可以	8.1	认为	8.8	建议	8.2
	可能	7.6	看	6.2	如果	7.0
	或	6.1	可能	5.3	像	4.3

在语料标注过程中,标注者对否定与不确定语义的判定具有明显主观成分,其标注结果依赖于:(1) 标注者对上下文语境的正确理解;(2) 标注者背景知识的多少.一致性检验能够验证不同标注者在同一标注规则下对标注结果的主观性差异程度,以及问题本身的难易程度.基于此,本文采用 Kappa 检验^[28]评价标注一致性.Kappa 检验是评价判断一致性程度的指标,其检验一致部分是否由偶然因素影响,反映测定结果之间的实际一致率与随机一致率之间的差别是否具有显著性意义.Kappa 统计量计算如下:

$$Kappa = \frac{P_0 - P_c}{1 - P_c} \quad (1)$$

其中, P_0 是观察一致性, P_c 为期望一致率. $Kappa$ 值在区间 $[-1,1]$,通常情况下, $Kappa \geq 0.75$,表明一致性较好; $Kappa < 0.4$,表明一致性较差.本文以句子为单位,当线索词和覆盖域完全相同时认为标注结果一致.在 3 个子语

料的 $Kappa$ 值分别为:科技文献 0.85,财经文章 0.81,酒店评论 0.88.该结果也从一定程度上反映了不同语料之间,否定性与不确定性信息抽取任务的难度.

3 面向汉语的否定性与不确定性信息抽取方法

否定性与不确定性信息抽取任务包括线索词检测和覆盖域界定两个子任务.针对线索词检测任务,本文探索了基于字序列与词序列标注的方法.该方法考虑了待测样本在序列中的上下文相关特征.针对覆盖域界定任务,本文首先同样采用基于序列标注的方法,并提出一系列有效的依存句法特征;其次,利用卷积核易于计算树结构相似度的特点,提出了基于树核结构化特征的方法,并探索了不同结构树与剪裁策略对覆盖域界定系统性能的影响;再次,通过对实验结果的分析发现,以上两类覆盖域界定方法具有明显的差异性和互补性,因此提出了基于元决策树的分类器融合模型;最后,本文提出了一种基于依存句法规则的覆盖域界定后处理方法.

3.1 线索词检测

序列标注方法已在中文词性标注和命名实体识别等研究中取得了较好的性能,因此,本文将线索词检测看作序列标注问题,分别探索了基于字序列和基于词序列的方法,并通过针对线索词的频率特征,降低了线索词歧义问题对系统性能的影响.

线索词检测序列标注集合定义如下:

- 1) $BCue$:标注单位(字或词)位于线索词的开始位置;
- 2) $ICue$:标注单位(字或词)位于线索词内部除起始位置之外的位置;
- 3) $OCue$:标注单位(字或词)位于线索词外部.

本文采用条件随机场(condition random field,简称 CRF)模型^[29]进行序列标注,该模型对于任意输出 Y 和观测到的输入 X ,在训练时对概率 $P(Y|X)$ 进行最大化计算.对给定的输入序列 $\mathbf{x}=\mathbf{x}_1, \dots, \mathbf{x}_n$ 以及标注结果序列 $\mathbf{y}=\mathbf{y}_1, \dots, \mathbf{y}_n, \mathbf{y}_i \in \{BCue, ICue, OCue\} (1 \leq i \leq n)$, \mathbf{y} 的概率为

$$P_{\lambda}(\mathbf{y} | \mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp\left(\lambda \cdot \sum_{i=1}^n \Phi(\mathbf{y}_{i-1}, \mathbf{y}_i, \mathbf{x})\right) \quad (2)$$

其中, $Z(\mathbf{x})$ 是归一化因子, $\Phi(\mathbf{y}_{i-1}, \mathbf{y}_i, \mathbf{x})$ 是特征向量函数, λ 是特征权重向量.

特征模版集根据窗口长度生成,即,在窗口范围中抽取每类特征的顺序组合.例如,当窗口长度为 3 时,特征模版集为 $\{C_{-1}, C_0, C_1, C_{-1}C_0, C_0C_1, C_{-1}C_0C_1\}$, C_i 是字或词.

在面向汉语的序列标注模型中,基于字(character)序列和基于词(word)序列是两种基本框架,前者的优势在于能够避免由分词算法引入的错误传播以及方便引入更细粒度的特征;而后者往往能够更好地捕获词序列内部的上下文关系.因此,本文分别尝试采用两种框架识别线索词,并包括以下特征.

1) n -gram 特征

字框架下的 n -gram 特征以单个汉字为单位,若遇英文,则以单词为单位;词框架下的 n -gram 特征以分词结果为单位.

2) 词性特征

由于线索词词性分布相对集中,例如,副词性和动词性否定线索词所占比例分别为 65.6%和 23.8%;而这两个比例在不确定线索词中分别为 16.8%和 54.3%.因此,词性特征对线索词检测来说是一类重要特征.字框架下的词性标记继承其所属词词性.

3) 词边界特征

本文采用 BMES 方法标识字符在词中的位置.句子中每个字符根据其其在词中的位置进行分类,分为 B,M,E,S 这 4 类,其中,B 表示该字符是词的开始,M 表示该字符在词的中间位置,E 表示该字符是词的结束位置,S 表示该字符能独立的构成一个词.词边界特征仅应用于基于字序列的标注方法,目的是为了提高对词边界的判断,而其贡献仍依赖于分词性能.

4) 频率特征

由于一词多义的特点,汉语中线索词的歧义现象较为明显,即:相同词在有些情况下是线索词,而另外一些情况下不是线索词,并且很难仅通过词性等特征进行区分.例如:

例句 4:除了洗澡水忽冷忽热,酒店的其他方面均让我们满意.

例句 5:这家酒店餐厅除了世界各地的美食,还提供免费的黑啤酒.

从意义上分析,例句 4 中,“除了”一词否定了“洗澡水忽冷忽热让我们满意”的情况,其为否定线索词;而在例句 5 中,“除了”一词仅表示递进关系,并非否定线索词.鉴于此,本文提出了频率特征,该特征统计了每个词在训练语料上是否作为线索词的频率,共有 4 个特征值:

- ALL_CUE:该词在所有情况下均为线索词.
- MORE_CUE:该词作为线索词的次数大于不作线索词的次数.
- LESS_CUE:该词作为线索词的次数小于不作线索词的次数.
- NONE_CUE:该词从未作为线索词.

频率特征粗粒度地刻画了线索词歧义在语料中的分布,一般而言,若该词在训练集上作为线索词出现的频率越高,则其在测试集中被判断为线索词的概率越大;反之,其在训练集中作为线索词出现频率越低,则在测试集中被判断为线索词的概率越小.该差异反映了线索词歧义的强弱.

3.2 覆盖域界定

3.2.1 基于序列标注的覆盖域界定模型

与线索词检测任务类似,覆盖域界定也可以看作是序列标注的问题,本文采用基于 CRF 模型的字序列和词序列标注方法界定覆盖域.

序列标注集合定义如下:

- 1) BScope:标注单位(字或词)位于覆盖域开始位置.
- 2) IScope:标注单位(字或词)位于覆盖域内部除起始位置之外的位置.
- 3) OScope:标注单位(字或词)位于覆盖域外部.

在基于序列标注的覆盖域识别模型中, n -gram 特征、词性特征、词边界特征的设置与线索词检测模型相同.

根据定义,覆盖域指线索词的语义作用范围,而句法结构能够从一定程度上体现语义关系,因此,对覆盖域界定模型而言,句法特征是一类重要特征.依存句法通过分析语言单位内成分之间的依存关系揭示其句法结构,与短语结构句法相比,依存句法分析结果简洁且易于表示,句法结构信息附加在词与词之间的关系上,没有引入新的短语节点,降低了系统的复杂度.此外,依存分析结果趋向扁平化,句法树层次较浅,使得节点之间距离缩短,能够更直接地反映词之间的支配和依赖关系,而覆盖域是线索词所作用的一段文本,因此,相比于短语结构句法,依存句法中词之间的支配关系更有利于捕获线索词与其对应覆盖域之间的关系,本文采用依存分析获得句法特征.图 1 是例句 1 对应的依存关系树,本文所采用的依存特征及举例如表 3 所描述(假设当前词为“管理”).

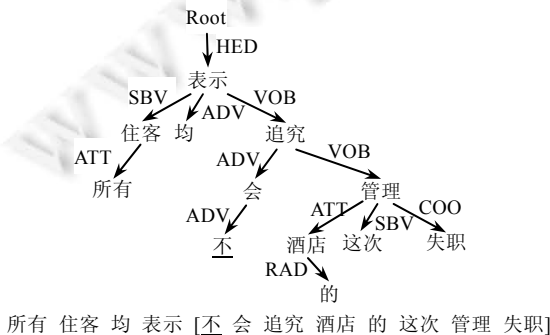


Fig.1 Dependency tree

图 1 依存关系树

Table 3 Dependency features of scope

表 3 覆盖域的依存特征

特征	描述	实例
依存关系(rel)	当前词对应的依存关系类型	VOB
到线索词路径(path)	当前词到线索词的有向压缩路径(压缩路径:若在同一依存方向上出现连续相同类型,将其合并)	VOB<>ADV
到线索词距离(dis)	当前词到线索词经过的边数	3
与线索词位置关系(pos)	当前词是否与线索词位于谓语动词的同侧	SAME_SIDE

以上所有特征中,除了词边界特征仅用于字序列的标注模型中,其他特征均用于字和词两种序列标注模型,其中,字框架下的词或句法特征继承其所属词的对应特征.

3.2.2 基于卷积树核的覆盖域界定模型

基于序列标注的覆盖域界定模型采用了一系列句法特征来获取全局信息,然而,这些特征无法完整地体现覆盖域和线索词在句法上的结构化关系.鉴于此,本文提出了基于卷积树核的覆盖域界定方法,弥补了平面特征在表示结构化数据时的局限性.

卷积树核^[30]通过计算两棵树中公共子树的数量来获得它们之间的相似度.解析树 T_1 和 T_2 之间的相似度 $K_c(T_1, T_2)$ 为

$$K_c(T_1, T_2) = \sum_{n_1 \in N_1} \sum_{n_2 \in N_2} \Delta(n_1, n_2) \tag{3}$$

其中, N_1 和 N_2 分别为 T_1 和 T_2 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,可按如下递归规则计算:

- (a) 若 n_1 和 n_2 节点处的产生式不同,则 $\Delta(n_1, n_2)=0$, 否则跳转至规则(b);
- (b) 若均为叶子节点前的一个节点,则 $\Delta(n_1, n_2)=1 \times \lambda$, 否则跳转至规则(c);
- (c) 递归计算 $\Delta(n_1, n_2)$:

$$\Delta(n_1, n_2) = \lambda \times \prod_{k=1}^{\#ch(n_1)} (1 + \Delta(ch(n_1, k), ch(n_2, k))) \tag{4}$$

其中, $\#ch(n_1)$ 是节点 n_1 的子节点数, $ch(n, k)$ 是节点 n 的第 k 个子节点, 衰减因子 $\lambda \in (0, 1]$ 使得相应的核值不易随子树的大小而变化.

本文从依存句法树中抽取词原型、词性和依存关系这 3 类特征构成结构化特征,其结构如图 2 所示(词性标记集和依存关系类型标记集见 <http://www.ltp-cloud.com>).其中,词原型树结构是将依存树中的依存关系标签去掉,只保留词之间的依赖关系;词性树是用词性代替词生成的树结构,其反映了不同词性之间的支配关系,同时还避免了词原型树结构面临的数据稀疏问题;依存关系树用依存弧上的标签代替了词生成树结构,该结构不仅反映了词之间的依赖关系,而且给出了其类型,为判断词之间的语义关系给出了暗示.

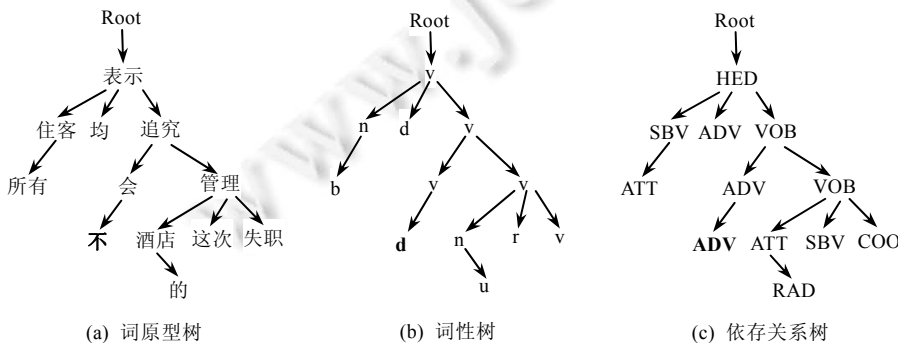


Fig.2 Structure of feature tree

图 2 特征树结构

完整的特征树不仅结构庞大,需要更多的训练时间,而且,复杂的结构还会引入大量噪音.借鉴 Zhang 等人在

短语结构树上提出的剪裁策略^[31],本文提出了以下 3 种剪枝方法,简化树结构,使得大多数有效信息得以保留,同时又尽可能地过滤掉无用信息:

1) 最小完全树(minimum complete tree,简称 MCT)

判断当前词是否在覆盖域中,所需信息是当前词与线索词在句法树中的相关信息,因此,本文考虑最小完全树.最小完全树是指线索词与当前词的最小公共祖先节点下的所有节点构成的树,如图 3(a)所示(以词原型树为例,假设当前词为“管理”,虚线所包含即为剪裁后的结果,下同).

2) 路径树(path tree,简称 PT)

MCT 树虽然只保留了最小公共祖先节点下的信息,但在很多情况下仍存在大量噪声.例如在图 3(a)中,当前词若位于根节点的其他子树上(如“所有”),则 MCT 树与完整树相同.此外,当前词越靠近根节点,也可能引入过多噪声.例如,当前词为“追究”时,其整个右子树都会划入 MCT 树中.鉴于此,本文考虑采用路径树.路径树仅保留了从线索词到当前词路径上所有节点构成的树,如图 3(b)所示.

3) 上下文相关路径树(context-sensitive path tree,简称 CPT)

虽然 PT 树极大地简化了 MCT 树,消除了几乎所有噪音,但同时去掉了大量有用的结构化信息.因此,本文提出了上下文相关路径树,其针对 PT 树上的所有节点,均向外扩充一个节点作为“上下文”,如图 3(c)所示.

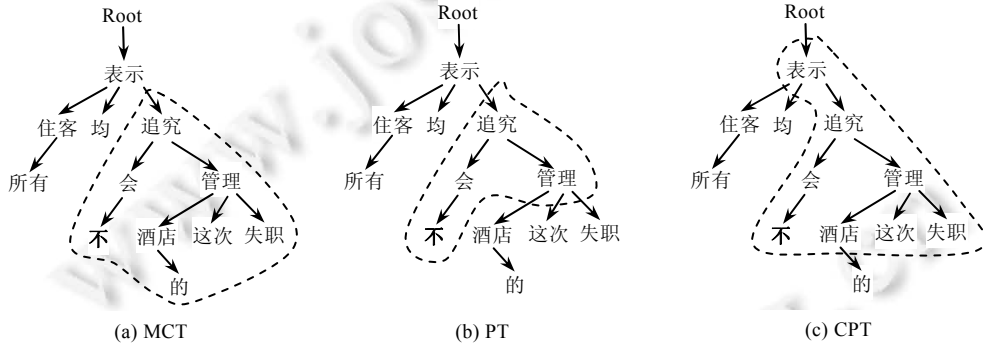


Fig.3 Pruning method of feature tree
图 3 特征树剪裁方法

卷积核易于捕获结构化信息,而基本核则能够更好地利用平面特征,因此,本文还尝试采用 SVM-Light 提供的复合核 $\mu K_1 + (1-\mu)K_2$ 来融合平面特征和结构化特征,其中, K_1 表示树核, K_2 表示基本核, $\mu \in (0, 1]$. 平面特征采用第 3.2.1 节中描述的基于词序列的相关特征.

3.2.3 基于元决策树的覆盖域界定模型

本文分析了基于词序列标注模型分类结果与基于复合核模型分类结果,以分类结果是否一致和是否分类正确为标准分为 3 类:(a) 两个模型均获得正确结果(both correct);(b) 仅有一个模型获得正确结果(only one correct);(c) 两个模型都判断错误(both wrong). 开发集上的统计结果如表 4 所示,可以看出,仅有一个分类器获得正确结果的比例为 33.1%. 如果能正确选择采用哪个分类器的输出作为结果,该方法的性能上限可达 90.7%. 鉴于此,本文采用基于元决策树(meta decision tree,简称 MDT)^[32]的分类器融合模型来提高这部分结果的正确性.

Table 4 Agreement distribution of classification result

表 4 分类结果一致性分布

统计类型	比例(%)	累积比例(%)
Both correct	57.6	57.6
Only one correct	33.1	90.7
Both wrong	9.3	100

元决策树在结构上与普通决策树类似,其主要特点是输入为基本分类器的预测结果和元属性集,输出是基

本分类器编号,即:给每个待测实例分配一个最合适的基本分类器,该特征有利于解决上文提到的“仅有一个模型获得正确结果”的情形.训练过程如下:在训练数据集 S 上,训练基本分类器集合 C ,其中包含了词序列标注分类器 C_1 和复合核分类器 C_2 .设类值集合为 $\{c_1, c_2, \dots, c_k\}$,则实例 x 在基本分类器 C_i 上所产生的预测向量为

$$P_{(x)}^{C_i} = (p^{C_i}(c_1 | x), p^{C_i}(c_2 | x), \dots, p^{C_i}(c_k | x)) \quad (5)$$

其中, $p^{C_i}(c_j | x)$ 是基本分类器 C_i 预测实例 x 属于类别 c_j 的概率.通过 $P_{(x)}^{C_i}$ 生成元属性集 M ,然后,两者再作为元决策树分类器的训练数据.其框架如图 4 所示.

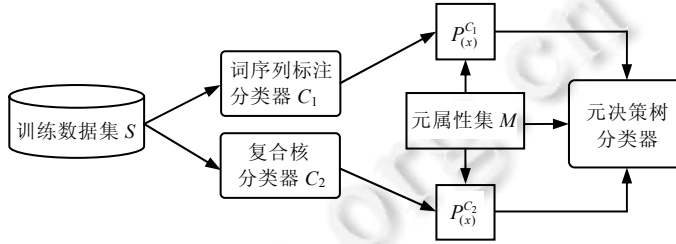


Fig.4 Framework of MDT-based classification

图 4 基于元决策树的分类框架

本文采用反映基本分类器预测结果确定性的类概率分布特征(class distribution properties,简称 CDP)作为元属性,包括:

1) 最高类概率(highest class probability)

基本分类器 C 为实例 x 预测的最高类概率为

$$\max \text{prob}(x, C) = \max_{i=1}^k p^C(c_i | x) \quad (6)$$

2) 类概率分布熵(class probability distribution entropy)

基本分类器 C 为实例 x 预测的类概率分布熵为

$$\text{entropy}(x, C) = -\sum_{i=1}^k p^C(c_i | x) \cdot \log_2 p^C(c_i | x) \quad (7)$$

以上两个元属性反映了基本分类器对类值预测的确定性.若类概率分布是高度扩散的,则通常最高类概率较小,而类概率分布熵较大,表明该基本分类器对类值预测的确定性低;相反,若概率分布较为集中,则通常最高类概率较大,而类概率分布熵较小,表明该基本分类器对类值的预测是确定的.

3.2.4 基于依存句法规则的后处理方法

本文第 2 节中提到:覆盖域是一段连续的文本,而将覆盖域界定看作词的分类问题,则有可能导致识别出的覆盖域不连续.在 Tech 语料开发集上的实验结果表明:采用标准线索词(golden cues)的情况下,否定和不确定覆盖域界定结果中,不连续的覆盖域分别达到 8.3%和 10.7%.因此,本文提出了一种基于依存句法规则的后处理方法,根据词类别标签的不同情形,步骤如下:

设 BScope 标记对应的词为 w_B ,IScope 标记对应的词为 w_I ,一段连续的 w_I 记为 W_I ,OScope 对应的词为 w_O ,一段连续的 w_O 记为 W_O ,线索词为 w_C .

定义 $Dis_{tree}(w_a, w_b)$ 为在依存树中 w_a 与 w_b 的距离(w_a 到 w_b 经过的依存弧个数).

情形 1:结果序列中仅包含 1 个 w_B 和 $W_I=w_{I1}, \dots, w_{Ik}$,且 W_I 紧随 w_B 之后,则将 w_B 和 W_I 标识为 Scope,如图 5(a)所示,跳转至情形 7.

情形 2:结果序列中包含 1 个 w_B ,且至少有 1 个 W_I 并非紧随 w_B 之后,计算所有 w_I 到 w_B 的平均距离 $\overline{Dis} = \text{ave}(Dis_{tree}(w_B, w_I))$ 以及到 w_B 的最小距离 $\min Dis(W_I) = \min \{Dis_{tree}(w_B, w_{I1}), \dots, Dis_{tree}(w_B, w_{Ik})\}$.若 $\min Dis(W_I) \leq \overline{Dis}$,则 W_I 为覆盖域的一部分;反之,则舍弃 W_I .最后,将结果序列中所有 w_B 和 W_I 及其之间的片段标识为 Scope,如图 5(b)所示,跳转至情形 7.

情形 3:结果序列中包含多个 w_B ,如图 5(c)所示,则仅保留最中心的 w_B ,计算公式如下:

$$w_B = \arg \min_{w_{B_i}} \sum_{w_{I_j}} Dis_{tree}(w_{B_i}, w_{I_j}) \tag{8}$$

其中, w_{I_j} 为结果序列中所有标记为 IScope 的词.公式(8)获得的 w_B 距离所有 IScope 节点最近,其他 w_B 均更改标记为 IScope,跳转至情形 2.

情形 4:结果序列中没有 w_B ,且至少包含 1 个 W_I ,如图 5(d)所示.将所有 W_I 的第 1 个词标记改为 BScope,跳转至情形 1(1 个 W_I)或情形 3(多个 W_I).

情形 5:结果序列中至少包含 1 个 w_B ,且不包含 W_I ,则选择在依存树中距离线索词最近的 w_B ,即,使得 $Dis_{tree}(w_B, w_C)$ 最小,将其标识为 Scope,如图 5(e)所示,跳转至情形 7.

情形 6:结果序列中不包含 w_B 和 W_I ,则将从线索词至句尾之间的序列作为覆盖域,如图 5(f)阴影部分所示.

情形 7:该情形下的序列仅包含一段连续的 Scope 标记,分两种情况:(1) 线索词包含在 Scope 序列中,则将 Scope 序列作为覆盖域;(2) 线索词位于 Scope 序列外,则将线索词与 Scope 序列及其之间的序列作为覆盖域,如图 5(g)阴影部分所示.

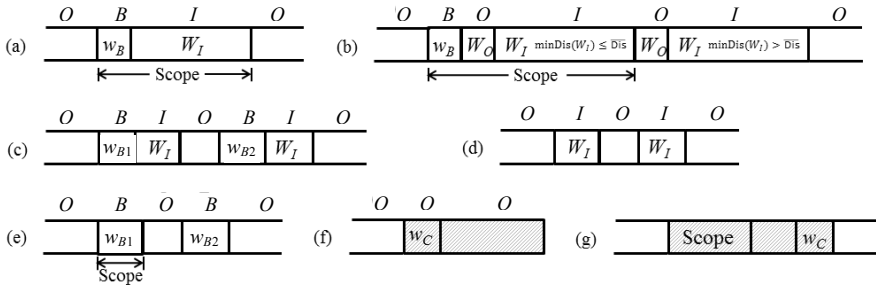


Fig.5 Post-Processing for scope resolution

图 5 覆盖域界定的后处理

4 实验设置和结果分析

4.1 实验设置

本节给出实验语料划分、评测方法、实验工具以及实验系统.

4.1.1 语料划分

我们分别在 Tech,Finan 和 Review 这 3 种类型的子语料上验证了本文提出的否定性与不确定性信息抽取方法的性能,其中,线索词检测任务采用整个子语料作为实验语料,覆盖域界定任务采用语料中包含否定性或不确定性信息的句子构成实验语料.每组实验中,随机抽取 20%语料作为开发集,剩余部分作为训练集和测试集,采用 5 倍交叉验证衡量系统性能.此外,本文在验证了系统在跨领域语料上的健壮性时,实验以 Tech 语料作为训练集,分别测试在 Finan 和 Review 语料上的性能.

4.1.2 评测方法

本文采用准确率 P 、召回率 R 和调和评价 F 作为评测指标,对线索词检测系统性能进行评价.具体定义为 $P=|A \cap B|/|A|$; $R=|A \cap B|/|B|$; $F=2PR/(P+R)$,其中, A 表示系统识别出的单元集合, B 表示语料中标注的单元集合.覆盖域界定系统采用精确率(accuracy,简称 Acc.)作为评测指标,该指标以句子为基本单位,要求系统结果与正确答案完全匹配.此外,由于本文将覆盖域任务看作分类任务,系统对每个待测实例进行分类(类别为 BScope,IScope 和 OScope).为了更直接地比较分类方法本身的性能,本文在部分实验中采用以待测实例为评测单元的评价指标,记为 P' , R' 和 F' .

4.1.3 实验工具

本文实验的主要工具包括:

- CRF++0.58(<https://crfpp.googlecode.com/svn/trunk/doc/index.html>):该工具提供了基于 CRF 的序列标注模型,用于线索词检测实验和覆盖域界定实验中.
- SVM-LIGHT-TK1.2(<http://disi.unitn.it/moschitti/Tree-Kernel.htm>):该工具提供了卷积树核和复合核,用于覆盖域界定实验中,参数 t 设置为 5, C 设置为“+”,其余参数采用默认值.
- MLC4.5(<http://kt.ijs.si/bernard/mdts/>):该工具提供了元决策树分类模型,用于覆盖域界定实验中,采用默认参数.
- 语言云 v1.1.1(<http://www.ltp-cloud.com/>):该工具提供了中文分词、词性标注和依存句法分析结果,用于分词及获取词法句法特征.

4.1.4 实验系统

本文所采用的各实验系统描述如下:

1) 线索词检测系统

- CRF_Char_CueSys:基于字序列标注的线索词检测系统,该系统以字为标注单位,特征集合为 $\{n\text{-gram}$ 特征、词性特征、词边界特征、频率特征 $\}$,采用基于 CRF 的序列标注模型.
- CRF_Word_CueSys:基于词序列标注的线索词检测系统,该系统以词为标注单位,特征集合为 $\{n\text{-gram}$ 特征、词性特征、频率特征 $\}$,采用基于 CRF 的序列标注模型.

2) 覆盖域界定系统

- CRF_Char_ScopeSys:基于字序列标注的覆盖域界定系统,该系统以字为标注单位,特征集合为 $\{\text{简单特征: } n\text{-gram}$ 特征、词性特征、词边界特征,依存特征:依存关系、到线索词路径、到线索词距离、与线索词位置关系 $\}$,采用基于 CRF 的序列标注模型.
- CRF_Word_ScopeSys:基于词序列标注的覆盖域界定系统,该系统以词为标注单位,特征集合为 $\{\text{简单特征: } n\text{-gram}$ 特征、词性特征,依存特征:依存关系、到线索词路径、到线索词距离、与线索词位置关系 $\}$,采用基于 CRF 的序列标注模型.
- TK_MCT_ScopeSys:采用基于最小完全树剪裁策略的卷积树核覆盖域界定系统.
- TK_PT_ScopeSys:采用基于路径树剪裁策略的卷积树核覆盖域界定系统.
- TK_CPT_ScopeSys:采用基于上下文相关路径树剪裁策略的卷积树核覆盖域界定系统.
- FLAT_ScopeSys:仅采用平面特征的基本核覆盖域界定系统.
- COMP_ScopeSys:融合平面特征与结构化特征的复合核覆盖域界定系统.
- MDT_ScopeSys:融合 CRF 模型与复合核模型的元决策树覆盖域界定系统.
- POST_ScopeSys:在 MDT_ScopeSys 基础上,采用基于依存句法规则后处理方法的覆盖域界定系统.

3) 对比系统

- AntU_Sys:由于汉语否定性与不确定性信息抽取研究比较匮乏,本文采用在 CoNLL 2010 评测中获得英语覆盖域界定子任务最高性能的系统作为对比系统.该系统基于存储算法的分类器,采用依存句法特征识别覆盖域^[23].
- SCU_Sys1:该系统采用结构化句法特征识别覆盖域,并提出针对不同词性的线索词自动建立独立分类器的策略^[33].由于该系统仅针对覆盖域界定任务进行研究,因此,为方便比较各系统在完整任务上的性能,线索词检测采用本文方法.
- SCU_Sys2:该系统将否定性与不确定性信息抽取任务看作语义角色标注(semantic role labeling,简称 SRL)问题,其中,线索词对应于 SRL 中的谓词,而覆盖域界定则可等价于论元识别过程.基于此,该系统采用简化的 SRL 方法识别线索词和覆盖域^[34].

4.2 线索词检测系统性能及分析

基于序列标注的线索词检测系统采用 CRF 算法,在开发集上获得模版窗口参数 $window=3$,拟合参数 $c=3.0$. 基于字序列和基于词序列的标注模型均采用该参数设置,在本节最后,分析了以上两个参数的敏感度.基于上述

参数指标,我们在 3 个子语料集上对本文提出的线索词检测方法进行了评测,获得了如表 5 所示的测试结果.实验采用贪心特征选择算法测试了本文等 3.1 节提出的 4 类特征的有效性,具体做法是:每次递增地选取一个贡献最大的待测特征放入特征集合中,直至任何剩余的待测特征均不能提高系统性能.

Table 5 Performance of cue detection system (F -measure)

表 5 线索词检测系统性能(F 值)

系统	否定			不确定			
	Tech	Finan	Review	Tech	Finan	Review	
CRF_Char_CueSys	+ n -gram	0.573 1	+ n -gram 0.428 1	+ n -gram 0.721 7	+ n -gram 0.481 1	+ n -gram 0.470 1	+ n -gram 0.398 5
	+Bound	0.612 6	+PoS 0.593 3	+PoS 0.832 8	+Bound 0.592 7	+PoS 0.589 7	+PoS 0.612 9
	+PoS	0.643 7	+Bound 0.602 4	+Bound 0.887 9	+PoS 0.672 1	+Bound 0.601 2	+Freq 0.633 7
	+Freq	0.651 9	-	+Freq 0.911 3	+Freq 0.690 4	-	-
	ALL	0.651 9	ALL 0.581 7	ALL 0.911 3	ALL 0.690 4	ALL 0.582 3	ALL 0.629 8
CRF_Word_CueSys	+ n -gram	0.547 9	+ n -gram 0.462 9	+ n -gram 0.674 2	+ n -gram 0.573 1	+ n -gram 0.489 3	+ n -gram 0.413 3
	+PoS	0.718 9	+PoS 0.604 5	+PoS 0.913 3	+PoS 0.668 4	+PoS 0.662 1	+PoS 0.644 1
	+Freq	0.771 8	+Freq 0.682 8	-	+Freq 0.689 1	+Freq 0.685 7	+Freq 0.693 8
	ALL	0.771 8	ALL 0.682 8	ALL 0.907 9	ALL 0.689 1	ALL 0.685 7	ALL 0.693 8

测试结果显示:在否定和不确定两类语义下,基于词序列的线索词检测系统性能占优,仅在 Review 子语料的否定线索词检测和 Tech 子语料的不确定线索词检测中,性能略低于基于字框架的系统(分别仅有 0.34%和 0.13%的性能下降),表明基于词序列标注的方法能够更好地识别线索词;其次,所有在不同子语料下的否定与不确定线索词检测最优系统均采用了全部特征(ALL),验证了本文提出的 4 类特征在识别线索词时的积极作用;最后,几乎所有系统中,贡献最大的两类特征都是 n -gram 特征和词性特征,而且性能提高幅度均较大,表明了这两类特征对刻画线索词的特点贡献最大.

上述测试结果显示了两种主要异常:(1) 否定线索词在不同领域的语料中识别性能差异明显,不确定线索词检测系统在 3 个子语料上的最大性能差为 0.81%,而在否定线索词检测系统最大性能差异达到了 22.85%;(2) 本文在前期工作中分别对否定和不确定线索词的歧义情况进行了统计,所谓歧义是指一个词既可以作为线索词出现,又可以作为非线索词出现,统计结果表明,23%的线索词存在歧义现象,然而利用本文提出的频率特征(freq)却仅获得了极小的优化性能,比如在特征选择中,频率特征均为提高最少的特征,甚至引入该特征反而导致了性能下降,差距较为明显.下面针对这两种异常给出分析.

1) 否定线索词检测性能差异异常

该异常的主要成因是语料中线索词样本分布不均,如本文第 2 节中表 2 所示,Tech 与 Review 子语料中的否定线索词分布与其他 4 类线索词差异较大,比例最大的两个线索词为“不”和“没有”,其在以上两个子语料中的占比高达 52.3%和 60.1%.此外,从表 1 中可以看出:Tech 与 Review 子语料的否定线索词集合大小分别为 44 和 128,均小于其他语料.样本的先验分布不均,本地地导致了否定线索词检测在不同子语料中的较大差异.为验证这一推断,本文从开发集中抽取 300 个包含否定线索词“不”或“没有”的实例进行测试,以 CRF_Word_CueSys 采用全部特征,分别采用 3 个子语料的训练集进行训练,获得 3 个否定线索词检测系统,其性能见表 6(括号中数值对应识别正确或错误的样本数).可以看出:由否定线索词分布不平衡语料(tech 和 review)所训练出的模型使得对线索词“不”和“没有”的识别精度较高,从而异常提升了否定线索词在 Tech 和 Review 子语料上的性能.

Table 6 Results of cue detection for “不” and “没有”

表 6 线索词“不”和“没有”的识别结果

Training corpus	Tech	Finan	Review
CRF_Word_CueSys	Correct (243) Wrong (57)	Correct (187) Wrong (113)	Correct (284) Wrong (16)

2) 线索词消歧异常

线索词消歧异常主要包含两个方面的原因.

- 一个原因是部分具有歧义的线索词,通过词性特征就能正确判断是否为线索词,达到消歧的目的.以

CRF_Word_CueSys 为例,其在 Finan 子语料上的一次交叉验证实验中,待测样本中有 129 个歧义实例,其中的 88 个实例在不包含频率特征(n -gram+PoS)的情况下已被正确识别。

- 另一个原因是占比较低的线索词在频率特征上不稳定,导致这些词在该特征上的值随机性较大。比如在 Finan 子语料中,不确定线索词“理当”在 5 倍交叉验证实验中,其频率特征值分别曾取为 ALL_CUE 一次, MORE_CUE 与 LESS_CUE 各两次。通过统计实验过程中频率特征的取值,此类不稳定的特征词占总样本的 3.5%(否定)和 11.3%(不确定),该比例也说明了频率特征在否定线索词检测系统中的作用要大于不确定线索词。

本节采用的基于 CRF 序列标注模型中,模版窗口参数 $window=3$,拟合参数 $c=3.0$ 。表 7 给出了这两个参数在不确定线索词检测系统上的敏感度,其中,测试语料为 Finan,系统分别为 CRF_Char_CueSys(简称为 C),特征集合为 n -gram+PoS+Bound 以及 CRF_Word_CueSys(简称为 W),系统特征集合为 ALL。从表 8 中可以看出:当模版窗口长度参数 $window$ 为 3 时,线索词检测系统达到最好的性能,该参数为 1 或 5 时,系统性能下降幅度较大。而拟合参数 c 则敏感度较低,尤其是在 $window=3$ 时, c 从 1.0~7.0 之间取值时,系统性能变化较小, $c=15.0$ 时,系统性能明显下降。由此说明了本实验选取参数的合理性。

Table 7 Sensitivity of parameters $window$ and c for CRF model (F -measure)

表 7 CRF 模型参数 $window$ 和 c 的敏感度(F 值)

Parameters	$c=1.0$	$c=3.0$	$c=5.0$	$c=7.0$	$c=15.0$
$window=1$	C: 0.4230 W: 0.5528	C: 0.4591 W: 0.5711	C: 0.4921 W: 0.5798	C: 0.4634 W: 0.5409	C: 0.3765 W: 0.4584
$window=3$	C: 0.5929 W: 0.6817	C: 0.6012 W: 0.6857	C: 0.6003 W: 0.6794	C: 0.5802 W: 0.6697	C: 0.4462 W: 0.5104
$window=5$	C: 0.5233 W: 0.5426	C: 0.5329 W: 0.5504	C: 0.5492 W: 0.5285	C: 0.4923 W: 0.5160	C: 0.4572 W: 0.4839

4.3 覆盖域界定系统性能及分析

本节给出了覆盖域界定系统 4 个方面关键技术的实验结果,包括基于依存句法平面特征的序列标注模型性能、基于结构化句法树特征的卷积核模型性能、融合两种分类器的元决策树模型性能以及基于依存句法规则的后处理算法的优化性能。覆盖域界定是线索词检测的后续任务,为了独立验证本文提出的模型和方法性能,本节实验采用标准线索词(golden cue)作为已知信息,关于自动线索词检测和覆盖域界定的联合系统实验结果将在第 4.4 节中给出。

4.3.1 基于序列标注模型的覆盖域界定系统性能及分析

本文采用与第 4.3.1 节实验相同的方法,首先验证了简单特征对覆盖域界定系统均具有积极作用,即:基于字序列标注的系统采用 n -gram 特征、词性特征和词边界特征;基于词序列标注的系统采用 n -gram 特征和词性特征。因此,本节将以简单特征为基础,采用贪心特征选择算法测试第 3.2.1 节提出的 4 类依存树特征的有效性。此外,CRF 模型在开发集上获得模版窗口参数 $window=3$,拟合参数 $c=3.0$,并测试了该参数的敏感度,结果与第 4.1 节相同,此处不再赘述。

表 8 给出了基于序列标注模型的覆盖域界定系统的性能,其中,rel 表示依存关系类型特征,path 表示到线索词的有向压缩路径特征,dis 表示到线索词的距离特征,pos 表示与线索词相对位置特征。测试结果显示:

- 1) 基于词序列标注的覆盖域界定系统性能均优于基于字序列标注的系统,这是由覆盖域界定任务对以字和词为标注单位的系统难度不同造成的。以 Tech 语料开发集上的不确定覆盖域为例:以字符为单位,其平均长度为 12.9;而以词为单位,其平均长度仅为 7.1。由于判定覆盖域正确与否的标准是所有单位均标注正确(完全匹配),因此在覆盖域界定任务中,基于字序列标注的系统难度较大。
- 2) 除了 Tech 和 Review 语料上的否定覆盖域界定系统以外,其他系统的最优性能均在采用全部特征(ALL)时获得,验证了本文提出的依存句法特征的有效性。
- 3) 从性能上看,否定覆盖域界定系统的性能优于不确定覆盖域界定系统。主要是两个方面的原因:首先,

否定覆盖域的平均长度小于不确定覆盖域(见表 1),如情形 1)中所分析,面向覆盖域平均长度短的系统具有优势;其次,否定覆盖域的依存句法特征分布相对集中稳定,从否定线索词的分布上(见表 2)可以得到印证.几乎所有覆盖域界定系统都符合该规律,以下实验分析中将不再赘述.

- 4) 值得注意的是:在 Finan 语料上的不确定覆盖域界定系统性能较低,不足 30%,其最主要的原因是,Finan 语料中不确定覆盖域的平均长度为 15.0(见表 1).我们统计了测试结果中的错误标注词与线索词之间的距离,距离小于 8 的词仅占 37.6%,其余错误全部出现在距离大于等于 8 的词上.这主要是由于本节采用的依存句法特征(path,pos 和 dis)均能较好地捕获当前词与线索词之间的关系,此类特征对距离线索词较远的词标注效果不佳.基于此,本文下面将通过采用结构化的树核特征来解决该问题.

Table 8 Performance of CRF-based scope resolution system (Acc.)

表 8 基于 CRF 的覆盖域界定系统性能(Acc.)

系统	否定			不确定		
	Tech	Finan	Review	Tech	Finan	Review
CRF_Char_ScopeSys	+rel 0.456 3	+path 0.310 4	+path 0.424 7	+path 0.275 2	+path 0.180 7	+path 0.219 6
	+path 0.472 7	+rel 0.352 3	+dis 0.481 5	+rel 0.331 0	+rel 0.247 7	+dis 0.285 5
	- -	+pos 0.374 9	+pos 0.502 4	+dis 0.370 2	+pos 0.283 3	+pos 0.304 8
	- -	+dis 0.378 5	+rel 0.516 0	+pos 0.398 4	- -	+rel 0.317 5
	ALL 0.461 9	+ALL 0.378 5	ALL 0.516 0	ALL 0.398 4	ALL 0.260 4	ALL 0.317 5
CRF_Word_ScopeSys	+path 0.459 2	+rel 0.362 2	+pos 0.512 9	+path 0.353 0	+path 0.239 2	+rel 0.274 9
	+rel 0.512 0	+path 0.392 3	+dis 0.543 2	+rel 0.394 4	+rel 0.263 1	+path 0.332 0
	+pos 0.553 1	+pos 0.403 9	+path 0.562 7	+pos 0.428 2	+pos 0.285 7	+dis 0.362 1
	- -	+dis 0.410 2	- -	+dis 0.431 9	+dis 0.293 8	+pos 0.368 7
	ALL 0.548 5	ALL 0.410 2	ALL 0.559 1	ALL 0.431 9	ALL 0.293 8	ALL 0.368 7

4.3.2 基于卷积树核的覆盖域界定系统性能及分析

本文第 3.2.2 节中提出了 3 类特征树,分别为词原型树(T_{Word})、词性树(T_{PoS})和依存关系树(T_{Dep}),其表示见图 2.表 9 给出了在 3 种剪裁策略下最优的特征树组合及其性能,特征组合的选择方法是:穷举 3 种特征树集合的幂集(去除空集),选取性能最高的一个特征树子集作为最优组合.

Table 9 Performance of scope resolution system based on tree kernel (Acc.)

表 9 基于树核的覆盖域界定系统性能(Acc.)

系统	否定			不确定		
	Tech	Finan	Review	Tech	Finan	Review
TK_MCT_ScopeSys	$T_{\text{PoS}}+T_{\text{Dep}}$ 0.4046	$T_{\text{Word}}+T_{\text{Dep}}$ 0.2602	$T_{\text{Word}}+T_{\text{Dep}}$ 0.4790	$T_{\text{Word}}+T_{\text{PoS}}$ 0.3522	$T_{\text{Word}}+T_{\text{Dep}}$ 0.2574	$T_{\text{Word}}+T_{\text{Dep}}$ 0.3458
TK_PT_ScopeSys	ALL 0.4871	$T_{\text{PoS}}+T_{\text{Dep}}$ 0.3721	$T_{\text{Word}}+T_{\text{Dep}}$ 0.5183	$T_{\text{Word}}+T_{\text{PoS}}$ 0.4309	T_{Word} 0.3073	$T_{\text{Word}}+T_{\text{PoS}}$ 0.4057
TK_CPT_ScopeSys	$T_{\text{Word}}+T_{\text{Dep}}$ 0.5029	$T_{\text{Word}}+T_{\text{Dep}}$ 0.4033	$T_{\text{Word}}+T_{\text{Dep}}$ 0.5092	ALL 0.4537	$T_{\text{Word}}+T_{\text{PoS}}$ 0.3415	$T_{\text{Word}}+T_{\text{Dep}}$ 0.3726

测试结果表明:

- 1) CPT 剪裁方法在 Tech 和 Finan 语料上获得了最好的性能,而 Review 语料上的最好性能被 PT 剪裁方法获得.其原因可能是 Review 语料的覆盖域长度较短,大多数有用的特征集中在当前词与线索词之间的路径上,PT 剪裁方法消除了大多数噪音,使得此类特征更加凸显,从而提高了树核的区分度;而 Tech 和 Finan 语料覆盖域较长且句法相对严格,CPT 剪裁方法则保留了更多的上下文特征,有利于判断离线索词较远的词是否处于覆盖域中.可以看出,CPT 与 PT 这两种剪裁方法对系统性能影响差别不大.
- 2) MCT 剪裁方法性能最低.其原因是采用该方法获得的特征树引入了过多的噪音,从而出现过拟合现象.我们采用第 4.1.2 节中介绍的以待测实例为单元的评价指标 P' , R' 和 F' 测试基于 MCT 剪裁方法的覆盖域界定系统性能.表 10 给出了 TK_MCT_ScopeSys 在 3 类语料上界定不确定覆盖域的结果.可以

看出:MCT 剪裁方法准确率较高,但召回率过低,两者差距较大,验证了其过拟合的判断。

- 3) 本节实验的所有系统中,仅有两个系统的最优特征集合包含了所有类型特征树(ALL),其原因可能是过多的结构化特征引入了大量噪声,同时也导致了过拟合现象,使得系统性能反而降低。

Table 10 Performance of uncertainty scope resolution classifier based on MCT pruning

表 10 基于 MCT 剪裁方法的不确定覆盖域界定分类器性能

系统	Tech			Finan			Review		
	P'	R'	F'	P'	R'	F'	P'	R'	F'
TK_MCT_ScopeSys	0.803 3	0.519 5	0.631 0	0.773 4	0.430 2	0.552 9	0.832 1	0.544 9	0.658 5

卷积树核通过计算依存树之间的相似度捕获结构化特征,而通过第 4.3.1 节的实验可以推断,平面特征对覆盖域界定系统也有较大贡献.基于此,本节采用复合核合并以上两类特征.其中,平面特征采用第 3.1 节中介绍的 *n*-gram 特征和词性特征(*n*-gram+PoS)以及第 3.2.1 节中介绍的 4 类依存句法特征(path+rel+dis+pos).表 11 给出了基于不同类型特征系统的性能,其中,TK_ScopeSys 对应表 9 中不同语料下性能最高的否定和不确定覆盖域界定系统.实验结果表明:相对于仅采用卷积树核的覆盖域界定系统,融合了平面特征的复合核系统性能有所提高,但提高幅度并不大,甚至在 Finan 语料上的不确定覆盖域界定系统性能反而下降.我们认为其原因在于:结构化特征中已包含了部分平面特征信息,从而导致引入部分噪音所致.此外,结合表 8 可以看出:采用相同特征集合的 SVM 覆盖域界定系统性能略低于 CRF 系统,说明后者能够更好地利用平面特征界定覆盖域。

Table 11 Performance of scope resolution system based on composite kernel (Acc.)

表 11 基于复合核的覆盖域界定系统性能(Acc.)

系统	否定			不确定		
	Tech	Finan	Review	Tech	Finan	Review
FLAT_ScopeSys	0.485 7	0.320 8	0.540 2	0.391 4	0.308 7	0.350 4
TK_ScopeSys	0.502 9	0.403 3	0.518 3	0.453 7	0.341 5	0.405 7
COMP_ScopeSys	0.536 6	0.419 1	0.577 2	0.481 1	0.332 9	0.446 3

4.3.3 基于元决策树的覆盖域界定系统性能及分析

通过在开发集上对基于词序列标注与基于复合核的分类器结果进行比较发现,有 33.1%的实例仅被其中一个分类器标注正确(见表 4).如果我们能够判断出哪个分类器提供了正确的结果,就能进一步提高覆盖域界定系统的性能.基于此,本节采用元决策树模型融合了以上两个分类器.表 12 给出实验结果与单一分类器结果的比较.测试结果表明:基于元决策树的覆盖域界定模型比单一模型在性能上均有较大幅度的提高,结果表明,以上两个单一模型在覆盖域界定任务中具有互补性.我们认为,其原因在于:基于序列标注的方法对整个序列的联合概率建立模型,能够有效地利用序列化信息,即,句子片段中的上下文信息;而基于核函数的方法能够更好地利用依存树的结构化信息,补充了大量潜在依存句法关系,从而使系统性能获得较大提高。

Table 12 Performance of scope resolution system based on MDT (Acc.)

表 12 基于元决策树的覆盖域界定系统性能(Acc.)

系统	否定			不确定		
	Tech	Finan	Review	Tech	Finan	Review
CRF_Word_ScopeSys	0.553 1	0.410 2	0.562 7	0.431 9	0.293 8	0.368 7
COMP_ScopeSys	0.536 6	0.419 1	0.577 2	0.481 1	0.332 9	0.446 3
MDT_ScopeSys	0.628 7	0.503 9	0.648 1	0.551 3	0.372 9	0.486 0

根据表 4 在开发集上的统计可知:理论上,基于元决策树模型的性能上限是分类准确率达到 90.7%,即对于测试样本,只要存在一个基分类器能够将其标识为正确类别即可.基于此,表 13 给出了该理论上限(分类结果的精确度,记为 C-Acc.)和基于元决策树的覆盖域界定系统性能所能达到的上限(Acc.).可以看出:在分类器达到理论上限时,覆盖域界定系统的性能提升较小.导致该问题的主要原因是其评价标准要求极为苛刻,只有当一个覆盖域样本与标准答案完全匹配时才正确.也就是说,一句话中哪怕仅有 1 个词被分类错误,结果即错误.因此,如

何能够解决该问题,成为影响覆盖域界定系统性能的关键.

Table 13 Upper-Limit of scope resolution system based on MDT

表 13 基于元决策树的覆盖域界定系统性能上限

性能上限	否定			不确定		
	Tech	Finan	Review	Tech	Finan	Review
样本分类(C-Acc.)	0.863 9	0.882 7	0.913 0	0.837 1	0.805 7	0.898 0
MDT_ScopeSys (Acc.)	0.680 1	0.541 9	0.703 8	0.596 3	0.440 5	0.565 1

4.3.4 覆盖域界定后处理方法性能及分析

本文提出的覆盖域界定方法采用分类模型,判断每个词是否在覆盖域中.该方法的劣势在于无法直接从句子层面把握覆盖域,因此,本文提出了基于依存句法规则的覆盖域界定后处理方法,表 14 给出了其性能.与表 12 的实验结果比较表明:采用后处理方法的覆盖域界定系统性能均获得较大幅度的提高,其中,Tech 语料上的否定覆盖域界定系统性能提高达到 6.97%,而最小的性能也提高了 2.69%,验证了本文提出的后处理方法的有效性.

Table 14 Performance of post-processing for scope resolution system (Acc.)

表 14 覆盖域界定后处理性能(Acc.)

系统	否定			不确定		
	Tech	Finan	Review	Tech	Finan	Review
POST_ScopeSys	0.698 4	0.541 0	0.690 7	0.585 7	0.423 3	0.512 9

4.4 对比实验及分析

本节从以下 4 个方面对否定性与不确定性信息抽取系统进行了实验及分析.

1) 标准线索词 vs.自动线索词

否定性与不确定性信息抽取任务要求识别线索词及其作用范围,可以看作这两个子任务的串联.本文第 4.3.2 节验证并分析了给定标准线索词前提下,覆盖域界定系统的性能.然而,在实际应用环境中却无法预先获知线索词,因此,通过自动识别线索词的覆盖域界定系统代表了其在否定性与不确定性信息抽取任务中的实际性能.本节以性能最高的 POST_ScopeSys 系统为基础,该系统采用基于元决策树的分类方法和基于依存句法规则的后处理方法,对比了在标准线索词(golden_cue)与自动识别线索词(auto_cue)的条件下该系统的性能,其中,在 Auto_Cue 系统中,线索词检测系统为表 5 中对应的各语料下的最优系统.为方便比较系统性能,Auto_Cue 系统同样采用 Accuracy 评价指标,也就是说,只有线索词和覆盖域均正确时才被判为正确结果.

表 15 是采用标准线索词的系统与采用自动线索词系统的性能比较,同时给出了 Auto_Cue 系统中的线索词检测性能.实验结果表明,Auto_Cue 系统性能均比对应的 Golden_Cue 系统要低.其原因是,线索词检测和覆盖域界定的串联系统导致了误差被传递.例如在 Review 语料中,Auto_Cue 系统在识别否定覆盖域时性能仅下降了 2.97%,相对于其他情况明显要偏低.这是因为其线索词检测性能较高(0.9113),从而使得传递的误差较少.

Table 15 Performance comparison of the golden-cue system with the auto-cue one

表 15 标准线索词系统与自动线索词系统性能比较

系统	否定			不确定		
	Tech	Finan	Review	Tech	Finan	Review
Golden_Cue (Acc.)	0.698 4	0.541 0	0.690 7	0.585 7	0.423 3	0.512 9
Auto_Cue (Acc.) (cue detection) (F)	0.584 1 (0.771 8)	0.452 2 (0.682 8)	0.661 0 (0.911 3)	0.470 6 (0.690 4)	0.318 1 (0.685 7)	0.445 7 (0.693 8)

2) 本文系统 vs.代表性系统

由于面向汉语的否定性与不确定性信息抽取受语料资源的限制,相关研究匮乏,目前尚未有成熟的系统以供比较.因此,本文将 3 个面向英文的系统迁移到汉语中作为对比系统,其在方法上各具代表性:AntU_Sys 以平面化的词法、句法特征为主;SCU_Sys1 主要采用了结构化句法特征;SCU_Sys2 则另辟蹊径,借助语义角色标注

的思路解决覆盖域界定问题.此外,3 个系统在英文语料(BioScope)上都获得了不错的性能.表 16 给出了否定性与不确定性信息抽取系统的性能比较,各个系统均为自动识别线索词下的性能.

Table 16 Performance comparison with the representative one and baseline (Acc.)

表 16 与代表性系统和基准系统的性能比较(Acc.)

系统	否定			不确定		
	Tech	Finan	Review	Tech	Finan	Review
AntU_Sys (CoNLL'10)	0.385 3	0.348 3	0.410 2	0.357 7	0.269 1	0.325 9
SCU_Sys1 (EMNLP'13)	0.422 7	0.385 6	0.393 1	0.370 4	0.287 7	0.304 2
SCU_Sys2 (EMNLP'10)	0.390 5	0.379 2	0.401 8	0.372 9	0.301 6	0.335 4
POST_ScopeSys (ours)	0.584 1	0.452 2	0.661 0	0.470 6	0.318 1	0.445 7

实验结果表明:在英文语料上均曾获得较好性能的 3 个系统 AntU_Sys,SCU_Sys1 和 SCU_Sys2,在汉语上的性能却不高.例如,AntU_Sys 系统在 CoNLL 2010 不确定性信息抽取任务上的成绩为 0.57,而在汉语语料上最好性能仅为 0.36.其主要原因是英语与汉语语言之间特点不同,尤其在句法层面,英语中句法结构相对严谨,简单的词法句法特征就能较好地捕捉其语言特点,而汉语语法则比较灵活,如大量的成分缺省等,因此,采用相同特征,在英汉两种语言上获得的性能差距较大;另一方面,时态、情态等特征,在面向英文的否定性与不确定性信息抽取中较为重要^[1,17],而在中文文本上则很难获取此类特征.由此可见:在否定性与不确定性信息抽取任务中,不适合将面向英语的系统直接应用至面向汉语的研究中.探索适合汉语特点的系统和方法具有重要意义.

3) 同领域 vs.跨领域

本节检验了本文系统在跨领域语料上的健壮性.实验采用 Tech 语料作为训练集,测试系统在 Finan 和 Review 语料上的性能.表 17 给出了实验结果,其中,括号中数值是相应系统以当前语料训练模型时所获得性能.由于训练及测试语料规模不同,该性能仅作为粗略的参照.

Table 17 Performance comparison on different corpus

表 17 不同语料上系统性能比较

系统	否定		不确定	
	Finan	Review	Finan	Review
CRF_Word_CueSys (F)	0.513 9 (0.682 8)	0.871 6 (0.911 3)	0.546 4 (0.685 7)	0.578 5 (0.693 8)
POST_ScopeSys (golden cue) (Acc.)	0.515 7 (0.541 0)	0.623 3 (0.690 7)	0.380 2 (0.423 3)	0.422 1 (0.512 9)
POST_ScopeSys (auto cue) (Acc.)	0.341 2 (0.452 2)	0.486 9 (0.661 0)	0.264 0 (0.318 1)	0.315 8 (0.445 7)

实验结果显示:

- 首先,跨领域线索词检测系统性能不高,与相同语料训练的模型性能差距较大.其主要原因在于:Tech 语料的线索词集合与其他两类语料相比较小(见表 1),尤其是否定线索词集合仅包含了 44 个词,使得训练的模型难以捕获 Finan 和 Review 语料特征,无法识别部分线索词.然而,跨语料系统在 Review 语料上的否定线索词检测性能却接近同语料系统.我们认为原因是:Tech 语料的否定线索词分布与 Review 语料极为相似,见表 2,其前两个高频线索词均为“不”和“没有”,且比例超过了 50%.
- 其次,采用标准线索词的跨领域覆盖域界定系统获得了较好的性能,验证了本文的覆盖域界定方法语料的健壮性.
- 还可以看出,该系统在 Finan 语料上的性能高于 Review 语料.结合表 1,我们推测其原因是:从句子和覆盖域平均长度来看,前者与 Tech 语料上的特点更为接近,因此能够获得更好的性能.
- 最后,采用自动线索词的系统则因为线索词检测的误差传播,其性能受到影响,相对于采用标注线索词的系统,其性能下降明显.

综上所述,本文提出的线索词检测方法对语料依赖较大,我们考虑未来工作中借助外部词表或知识库来辅助线索词检测系统;覆盖域界定方法对不同语料的适应性则较高.

4) 语料平衡性的影响

从表 1 中可以看出:除了 Finan 语料包含不确定语义的实例占比达到 50%以上之外,其他情况下包含否定或

不确定语义的实例占比均仅在 10%~30%之间.鉴于此,我们推测:采用原始语料进行训练,可能会受到数据不平衡的影响.因此,本文在 Tech 语料上重新构建了训练集和测试集:训练集由 500 句包含否定(不确定)语义的句子与 500 句不包含否定(不确定)语义的句子组成,共 1 000 句;测试集中,两类句子分别为 100 句,共 200 句.以 CRF_Word_CueSys 和 POST_ScopeSys 两个系统分别测试了平衡语料下的线索词检测和覆盖域界定系统的性能.表 18 给出了语料的(不)平衡性对系统性能的影响.为保证性能比较的公平,括号中表示的性能是在原语料中随机抽取 1 200 个句子,1 000 个作为训练集,200 个作为测试集所获得的系统性能.实验结果显示:采用平衡数据后,系统性能并没有较大的提升,其中,否定线索词检测系统性能反而下降,表明本文所提出的方法对数据不平衡带来的影响不敏感.

Table 18 Effects of corpus' (im) balance on our system

表 18 语料(不)平衡性对本系统的影响

系统	否定	不确定
CRF_Word_CueSys (<i>F</i>)	0.660 2 (0.653 9)	0.525 8 (0.542 2)
POST_ScopeSys (Acc.)	0.53 (0.51)	0.40 (0.44)

5 结论及未来工作

本文针对面向自然语言文本的否定性与不确定性信息抽取任务进行了系统的研究:首先,利用序列标注模型与卷积核模型系统地探索了各种序列化依存特征和结构化句法树特征的有效性;其次,根据以上两种模型的差异性和互补性,采用元决策树模型对其进行融合;最后,提出了一种有效的基于规则的后处理方法,进一步修正覆盖域.实验结果显示,该方法在否定性和不确定性信息抽取任务上的精确率分别达到 68.84%和 58.57%.此外,本文还构建了一个具有同 BioScope 语料库规模相当的汉语语料库.该语料库对推动汉语否定性与不确定性信息抽取研究具有重要作用.

本文还验证了线索词检测与覆盖域界定串联系统的性能,为相关研究在真实应用环境下期望获得的性能提供了参考和依据.同时,实验对比了本文系统与 3 类面向英文系统的性能,结果表明:汉语否定性与不确定性信息抽取研究具有其自身的特点,需要探索适合其特点的方法.此外,实验还验证了跨领域的系统性能,表明本文提出的覆盖域界定模型具有较好的跨领域健壮性.最后,实验验证了数据的不平衡性对系统的影响较小.

然而,尽管本文在面向自然语言文本的否定性与不确定性信息抽取研究中取得了初步的结果,但线索词检测和覆盖域界定的串联系统仍存在较大的性能瓶颈,其原因是:前一系统的误差会完全转入到后面系统中,导致无法挽救的性能下降.针对这一问题,未来的工作将尝试探索线索词检测与覆盖域界定的融合模型,使得后者有机会对线索词的识别结果进行修正,从而提高系统的整体性能.此外,线索词检测系统在跨领域健壮性方面的不足,也是本文系统面临多样或复杂语料时所可能出现的问题.因此,借助数据挖掘技术和知识库,探索自动建立准确且全面的否定与不确定语义词表的方法,也是本文未来工作的重要尝试之一.

致谢 在此,我们向对本文工作给予支持和建议的同行以及提出宝贵意见的评审专家表示感谢.

References:

- [1] Morante R, Sporleder C. Modality and negation: An introduction to the special issue. *Computational Linguistics*, 2012,38(2):223-260. [doi: 10.1162/COLI_a_00095]
- [2] Salkie R, Busuttill P, van der Auwera J. *Modality in English: Theory and Description*. Berlin: Mouton de Gruyter, 2009.
- [3] Aikhenvald AY. *Evidentiality*. New York: Oxford University Press, 2004.
- [4] Rubin VL, Liddy E, Kando N. Certainty identification in texts: Categorization model and manual tagging results. *Computing Attitude and Affect in Text: Theory and Applications*, 2005,20:61-76. [doi: 10.1007/1-4020-4102-0_7]
- [5] Wiebe J, Wilson T, Bruce R, Bell M, Martin M. Learning subjective language. *Computational Linguistics*, 2004,30(3):277-308. [doi: 10.1162/0891201041850885]

- [6] Lakoff G. Hedges: A study in meaning criteria and the logic of fuzzy concepts. *Philosophical Logic*, 1973,2(4):458–508. [doi: 10.1007/BF00262952]
- [7] Friedman C, Alderson PO, Austin JH, Cimino JJ, Johnson SB. A general natural-language text processor for clinical radiology. *Journal of the American Medical Informatics Association*, 1994,1(2):161–174. [doi: 10.1136/jamia.1994.95236146]
- [8] Friedman C, Hripcsak G. Natural language processing and its future in medicine. *Academic Medicine*, 1999,74(8):890–895. [doi: 10.1097/00001888-199908000-00012]
- [9] Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Biomedical Informatics*, 2001,34(5):301–310. [doi: 10.1006/jbin.2001.1029]
- [10] Wilson TA. Fine-Grained subjectivity and sentiment analysis: Recognizing the intensity, polarity, and attitudes of private states [Ph.D. Thesis]. Pittsburgh: University of Pittsburgh, 2008.
- [11] Council IG, McDonald R, Velikovich L. What's great and what's not: Learning to classify the scope of negation for improved sentiment analysis. In: Morante R, Sporleder C, eds. *Proc. of the Workshop on Negation and Speculation in Natural Language Processing*. Stroudsburg: Association for Computational Linguistics, 2010. 51–59.
- [12] Averbuch M, Karson T, Ben-Ami B, Maimon O, Rokach L. Context-Sensitive medical information retrieval. *Studies in Health Technology and Informatics*, 2004,107(Pt1):282–286. [doi: 10.3233/978-1-60750-949-3-282]
- [13] Grabar N, Hamon T. Exploitation of speculation markers to identify the structure of biomedical scientific writing. In: Lucila Ohno-Machado L, ed. *Proc. of the American Medical Informatics Association 2009 Annual Symp.* San Francisco: American Medical Informatics Association, 2009. 203–207.
- [14] Baker K, Bloodgood M, Dorr BJ, Filardo NW, Levin L, Piatko C. A modality lexicon and its use in automatic tagging. In: Calzolari N, Choukri K, Maegaard B, Mariani J, Odijk J, Piperidis S, Rosner M, Tapias D, eds. *Proc. of the 7th Conf. on Int'l Language Resources and Evaluation*. Paris: European Language Resources Association, 2010. 1402–1407.
- [15] Vincze V, Szarvas G, Farkas R, Mora G, Csirik J. The BioScope corpus: Biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics*, 2008,9(Suppl 11):279–282. [doi: 10.1186/1471-2105-9-S11-S9]
- [16] Kim JD, Ohta T, Pyysalo S, Kano Y, Tsujii J. Over-View of BioNLP'09 shared task on event extraction. In: Tsujii J, Kim JD, Pyysalo S, Bergler S, eds. *Proc. of the BioNLP Workshop Companion Volume for Shared Task*. Stroudsburg: Association for Computational Linguistics, 2009. 1–9. [doi: 10.3115/1572340.1572342]
- [17] Farkas R, Vincze V, Mora G, Csirik J, Szarvas G. The CoNLL'2010 shared task: Learning to detect hedges and their scope in natural language text. In: Farkas R, Vincze V, Szarvas G, Mora G, Csirik J, eds. *Proc. of the 14th Conf. on Computational Natural Language Learning*. Stroudsburg: Association for Computational Linguistics, 2010. 1–12.
- [18] Morante R, Blanco E. *SEM 2012 shared task: Resolving the scope and focus of negation. In: Agirre E, Bos J, Diab M, Manandhar S, Marton Y, Yuret D, eds. *Proc. of the 1st Joint Conf. on Lexical and Computational Semantics*. Stroudsburg: Association for Computational Linguistics, 2012. 265–274.
- [19] Kilicoglu H, Bergler S. Recognizing speculative language in biomedical research articles: A linguistically motivated perspective. *Journal of BMC Bioinformatics*, 2008,9(1):1–10. [doi: 10.1186/1471-2105-9-S11-S10]
- [20] Özgür A, Radev DR. Detecting speculations and their scopes in scientific text. In: Koehn P, Mihalcea R, eds. *Proc. of the 2009 Conf. on Empirical Methods in Natural Language Processing*. Stroudsburg: Association for Computational Linguistics, 2009. 1398–1407.
- [21] Øvrelid L, Velldal E, Oepen S. Syntactic scope resolution in uncertainty analysis. In: Huang CR, Jurafsky D, eds. *Proc. of the 23rd Int'l Conf. on Computational Linguistics*. Stroudsburg: Association for Computational Linguistics, 2010. 1379–1387.
- [22] Tang B, Wang X, Wang X, Yuan B, Fan S. A cascade method for detecting hedges and their scope in natural language text. In: Farkas R, Vincze V, Szarvas G, Mora G, Csirik J, eds. *Proc. of the 14th Conf. on Computational Natural Language Learning*. Stroudsburg: Association for Computational Linguistics, 2010. 13–17.
- [23] Morante R, Van Asch V, Daelemans W. Memory-Based resolution of in-sentence scopes of hedge cues. In: Farkas R, Vincze V, Szarvas G, Mora G, Csirik J, eds. *Proc. of the 14th Conf. on Computational Natural Language Learning*. Stroudsburg: Association for Computational Linguistics, 2010. 40–47.
- [24] Vlachos A, Craven M. Detecting speculative language using syntactic dependencies and logistic regression. In: Farkas R, Vincze V, Szarvas G, Mora G, Csirik J, eds. *Proc. of the 14th Conf. on Computational Natural Language Learning*. Stroudsburg: Association for Computational Linguistics, 2010. 18–25.

- [25] Zhang R, Song R. Negation scope of Chinese compound sentences. *Journal of Chinese Information Processing*, 2007,21(5):131–135 (in Chinese with English abstract).
- [26] Cui X, Yi J, Liu W. On the meaning representation of fuzzy words. In: Yu S, Gao D, eds. *Proc. of the 14th Chinese Lexical Semantics Workshop*. Zhengzhou: Zhengzhou University Press, 2013. 551–554.
- [27] Morante R, Schrauwen S, Daelemans W. Annotation of negation cues and their scope. Guidelines v1.0. Technical Report, CTR-003, Antwerp: CLiPS, University of Antwerp, 2011.
- [28] Cohen J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 1960,20(1):37–46. [doi: 10.1177/001316446002000104]
- [29] Lafferty J, McCallum A, Pereira F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Brodley CE, Danyluk AP, eds. *Proc. of the 18th Int'l Conf. on Machine Learning*. San Francisco: Morgan Kaufmann Publishers, 2001. 282–289.
- [30] Collins M, Duffy N. Convolution kernels for natural language. In: *Proc. of the NIPS 2001*. Vancouver: MIT Press, 2001. 625–632.
- [31] Zhang M, Zhang J, Su J, Zhou G. A composite kernel to extract relations between entities with both flat and structured features. In: Calzolari N, Cardie C, Isabelle P, eds. *Proc. of the 21st Int'l Conf. on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*. Stroudsburg: Association for Computational Linguistics, 2006. 825–832. [doi: 10.3115/1220175.1220279]
- [32] Todorovski L, Džeroski S. Combining classifiers with meta decision trees. *Machine Learning*, 2003,50(3):223–249. [doi: 10.1023/A:1021709817809]
- [33] Zou B, Zhou G, Zhu Q. Tree kernel-based negation and speculation scope detection with structured syntactic parse features. In: Baldwin T, Korhonen A, eds. *Proc. of the 2013 Conf. on Empirical Methods in Natural Language Processing*. Stroudsburg: Association for Computational Linguistics, 2013. 968–976.
- [34] Zhu Q, Li J, Wang H, Zhou G. A unified framework for scope learning via simplified shallow semantic parsing. In: Li H, Márquez L, eds. *Proc. of the 2010 Conf. on Empirical Methods in Natural Language Processing*. Stroudsburg: Association for Computational Linguistics, 2010. 714–724.

附中文参考文献:

- [25] 张瑞朋,宋柔.否定词跨标点句管辖的判断.中文信息学报,2007,21(5):131–135.



邹博伟(1984—),男,山东济南人,博士生,主要研究领域为信息抽取.



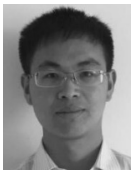
朱巧明(1964—),男,博士,教授,博士生导师,CCF 杰出会员,主要研究领域为中文信息处理.



钱忠(1989—),男,博士生,主要研究领域为信息抽取.



周国栋(1967—),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为信息抽取.



陈站成(1989—),男,硕士,主要研究领域为信息抽取.