软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2012,23(5):1085−1099 [doi: 10.3724/SP.J.1001.2012.04044]

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

基于树核函数的中英文代词消解*

孔 芳[1,2], 周国栋[1,2+]

[1](苏州大学 计算机科学与技术学院 自然语言处理实验室,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)

Pronoun Resolution in English and Chinese Languages Based on Tree Kernel

KONG Fang[1,2], ZHOU Guo-Dong[1,2+]

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

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

+ Corresponding author: E-mail: gdzhou@suda.edu.cn

Kong F, Zhou GD. Pronoun resolution in English and Chinese languages based on tree kernel. *Journal of Software*, 2012,23(5):1085−1099. http://www.jos.org.cn/1000-9825/4044.htm

Abstract: This paper proposes a tree kernel method to anaphora resolution of pronouns in both English and Chinese. First, several basic structured tree spans are proposed according to linguistic intuition. The similarity between two structured objects is computed directly using SVMLight. Then, a dynamic-expansion scheme is proposed to automatically determine a proper tree span for pronoun resolution by the centering theory, antecedent competitor-related information, and semantic role-related information. Evaluation on both the ACE 2004 English NWIRE corpus and the ACE 2005 Chinese NWIRE corpus justified the effectiveness of this method.

Key words: pronoun resolution; structured parse tree; tree kernel

摘 要: 基于树核函数,提出了从使用中心理论、集成竞争者信息和融入语义角色相关信息这 3 个方面对结构化句法树进行动态扩展来提升中英文代词消解的性能.首先探索了 3 种基本结构化句法树捕获方案,并使用 SVMLight 中提供的卷积树核函数直接进行基于结构化句法树的相似度计算,从而完成指代消解任务;其次,在分析 3 种结构化句法树捕获方案的基础上,从中心理论、竞争者信息和语义角色相关信息等几方面对捕获的结构化句法树进行了扩展;最后,通过 ACE 2004 NWIRE 英文语料和 ACE 2005 NWIRE 中文语料上的实验,说明了这些扩展能够提升代词消解的性能.

关键词: 代词消解;结构化句法树;树核函数

中图法分类号: TP391 **文献标识码**: A

作为一种常见的语言现象,指代广泛存在于自然语言的各种表达中,用于表示篇章中的一个语言单位(通常是名词性短语)与之前出现的语言单位间存在的特殊语义关联,且后者的语义解释依赖于前者.在语言学中,用于指向的语言单位称为照应语(或指代语 anaphor),被指向的语言单位称为先行语(或先行词 antecedent),确定照

* 基金项目: 国家自然科学基金(90920004, 61003153); 国家高技术研究发展计划(863)(2012AA011102); 国家教育部博士点基金(200802850006)

收稿时间: 2010-07-05; 定稿时间: 2011-04-29

应语所指的先行语的过程就是指代消解.

随着语篇理解、机器翻译以及问答系统等自然语言处理相关研究的不断深入,指代消解日益成为了研究热点.近年来,指代消解的研究已从早期基于规则的方法转入基于机器学习的方法.目前,大多数基于机器学习的指代消解研究沿用了 Soon 等人[1]提出的框架结构,其基本思想是将指代消解转换成一个二元分类问题.基本过程包括:首先从标注好的训练语料中提取各类词法、语法和语义特征,生成指代消解的训练集;然后利用 SVM、最大熵等分类器训练得到分类器模型;最后利用训练时得到的分类器模型对测试语料进行处理,确定语言单位之间(通常是名词性短语)可能存在的指代关系.在这一思想的指导下,很多研究者做了大量的工作,推动了指代消解研究的发展.例如,Ng 等人[2]对文献[1]的研究进行了扩充,抽取了 53 个不同的词法、语法和语义特征,在 MUC-6 上的性能 F 值达到了 69.4;Yang 等人[3,4]提出了一个双候选模型,直接学习各先行语候选之间的竞争关系,以更好地确定先行语;Yang 等人[5]进一步探索了先行语候选指代链中的语义信息在代词(特别是中性代词)指代消解中的作用;Bergsma 等人[6]提出了一种基于路径的代词指代消解方法;Ng 等人[7]详细探讨了各种语义信息对指代消解的意义,研究表明,有效的语义信息能够极大地提升指代消解的性能;Kong 等人[8]以语义角色为载体,将中心理论从传统的语法层拓展到语义层,进一步探索了语义信息对代词消解性能的作用.研究表明,在中心理论指导下,合理地使用语义信息能够极大地提升代词消解的性能.与英文相比,目前中文指代消解的研究要少得多,主要属于跟进型研究,代表工作包括:王厚峰等人[9−11]分别从领域和语义等知识出发,提取规则进行了指代消解的研究;张威等人[12]对汉语中的元指代现象进行了分析,并在句焦点集的基础上用优先和过滤算法实现了元指代的消解;王晓斌等人[13]利用语篇表述理论指导指代消解;李国臣等人[14]将英文平台的类似做法移植到中文指代消解中,采用决策树方法对中文人称代词的消解进行了研究;周俊生等人[15]提出了一种基于图划分的无监督的汉语指代消解算法,其性能与监督的汉语指代消解性能相当;杨勇等人[16]给出了一个基于机器学习的指代消解平台,并对指代消解中各类距离特征对指代消解性能的影响进行了深入的探索;王海东等人[17]探索了语义角色对指代消解性能的影响,其研究表明,语义角色信息的引入能够显著提高指代消解的性能.

随着指代消解研究的不断深入,越来越多的研究者发现,结构化句法信息对指代消解意义重大.而对结构化句法信息的处理,传统方法是定义一系列能够从浅层或深层句法树中获取的平面特征,即将结构化信息转化成平面特征,例如使用主语和宾语等语法角色特征来表示特定的结构化句法信息.虽然这种方法在一定程度上提升了系统的性能,但并不能充分使用结构化信息,具体体现在:

(1) 特征的构造过程比较繁琐.虽然不需要特别专业的专家书写细致的规则,但仍需研究者对处理的问题进行深入分析,总结出适用的特征集.当处理对象稍有变化(例如语种或使用领域)时,就必须对特征集作相应的调整,这极大地降低了系统的可扩展性和可移植性.

(2) 特征方式无法充分表示结构化信息.特征信息是平面的,把结构化信息转化成平面特征时,势必会丢失部分有效信息.例如,即使非常相似的两条路径,也可能会因为某一中间节点的差异就被当成截然不同的特征,从而无法体现其相似性,进而不利于分类效果.

基于上述原因,本文重点探索了哪些结构化句法信息对指代消解任务是有效的,并且在对结构化句法信息的处理上,我们跳过了结构化信息到平面特征的转化过程,而是利用卷积树核函数直接将捕获的有效结构化信息引入指代消解任务.由于代词在指代消解中占有举足轻重的地位,本文将重点研究代词的指代消解.

本文第 1 节简单介绍结构化信息在指代消解中的应用研究.第 2 节从系统框架、结构化句法特征空间和卷积树核函数这 3 方面入手,介绍基于树核函数的中英文代词消解的基本过程,并给出具体的实验结果和分析.第 3 节在第 2 节实验分析的基础上对结构化句法树捕获策略做进一步的扩展,并通过实验数据分析这些扩展的有效性.第 4 节就代词消解对句法分析器的依赖度、跨语句的代词消解问题做进一步分析.第 5 节对基于扩展的结构化句法树的指代消解进行错误分析.第 6 节给出全文总结和对未来工作的展望.

1 相关工作

在早先基于规则的指代消解研究中,诸多研究者就已经意识到了结构化句法信息的重要性,并基于结构化

信息设立了一些规则来帮助指代消解,典型的工作包括:

Hobbs[18]尝试使用句法树进行代词的指代消解.具体做法是:首先,为文档中的每个句子建立完全句法树;然后,采用从左到右广度优先的搜索方法遍历完全句法树,并根据语法结构中的支配和约束关系选择合法的名词短语作为先行语.Lappin 和 Leass[19]提出了 RAP 算法,首先使用 McCord[20]提出的槽文法(slot grammar)获得文档的句法结构;然后,通过手工加权各种语言特征,计算各先行语候选的突显性;最后,利用过滤规则确定先行语,实现句内和句间第三人称代词和反身代词的消解.

近年来,基于机器学习的指代消解研究得到了长足的发展,许多研究者也一直尝试将各种结构化句法信息引入指代消解.传统的方法是将其转化成平面特征,再使用基于特征向量的方法来解决,典型的工作包括:

Yang 等人[21]给出了一种基于语义相容统计信息的代词指代消解方法.在确定指代词的先行语时,他们首先提取先行语候选词所在上下文的谓词元组信息(例如主谓、谓宾等关系,显然是一些结构化信息),并对提取的谓词元组进行预处理(例如用语义类别代替命名实体、还原动词原型等),以降低数据稀疏问题以及语义相容计算的复杂度;然后,借助语料和 Web 去统计提取的谓词元组出现的频度;最后,将这一统计结果作为考虑要素,融入到基于特征的指代消解中,取得了较好的代词消解性能.Bergsma 等人[6]提出了一种基于路径的代词指代消解方法.他们首先将依存树中可能具有指代关系的两个节点间的节点序列和依存标签定义成依存路径;然后,根据出现的代词的单复数和性别信息分别统计语料中这一依存路径出现的次数,再利用依存路径相似度计算公式算出这一路径链接的两个对象间具有指代关系的概率;最后,结合这一概率对代词进行消解,使得指代消解性能有了一定的提升.

近几年来,随着核方法的广泛应用,各类用于处理结构化句法信息的树核函数不断被提出来,Collins 等人[22]定义了通过计算两棵句法树之间的相同子树的数量来比较句法树之间相似度的卷积树核函数;Culotta 等人[23]通过一些转换规则(如主语依存于谓语、形容词依存于它们所修饰的名词等)将句法树转换成依存树,并在树节点上增加词性、实体类型、词组块、WordNet 上位词等特征,定义了基于依存树的树核函数;Bunescu 等人[24]提出了基于最短路径依存树的树核函数等.在这些树核函数研究的基础上,相关研究人员开始尝试借助树核函数,直接将结构化句法信息应用于指代消解,取得了一定的突破,典型的工作包括:

Yang 等人[25]在代词消解的研究中探索了几种不同指代解析树抽取方案,并利用卷积树核函数直接计算指代解析树间的相似度,对第三人称代词的消解进行了初步研究;Zhou 等人[26]在文献[25]的工作基础上提出了能够有效捕获上下文信息的上下文相关卷积树核函数,并把这一卷积树核函数应用于代词的指代消解,并探讨了各种句法化信息对指代消解的影响;Kong 等人[27]在文献[26]的工作基础上进一步系统地探索了多种结构化句法信息对代词消解的性能影响情况.

在中文指代消解领域,有少数研究者研究了结构化信息对指代消解的作用.宋巍等人[28]给出了一种句法与词义相结合的中文代词消解方案.他们利用自动生成的依存句法分析结果来构建句法角色特征,并将该特征引入中文第三人称代词的消解,取得了较好的性能.不过到目前为止,在中文指代消解领域,还未见到基于树核函数的指代消解相关研究.

本文利用卷积树核函数,重点探讨了捕获的多种结构化句法信息对代词消解的作用,并在前人研究的基础上,对捕获的结构化句法信息进行了动态扩充.中英文语料上的各类实验表明,基于树核函数的代词消解方案具有更好的可移植性.

2 基于卷积树核函数的代词消解

本节从系统框架、结构化句法树和卷积树核函数这 3 方面入手,介绍了基于卷积树核函数的中英文代词消解方案,并通过对 ACE 2004 NWIRE 英文语料和 ACE 2005 NWIRE 中文语料上实验结果的分析,说明了结构化句法信息对代词消解的作用.本节给出的实验结果均为句内指代关系的消解性能,有关指代关系跨越多语句的代词消解情况,将会在第 4 节中加以讨论.

2.1 系统框架

参考 Soon 等人[1]提出的指代消解基本框架结构,我们针对英文和中文使用了统一的系统框架,采用全自动的方式实现中英文指代消解,具体的框架构成如图 1 所示,其中,各构成模块针对中英文的处理略有差异.

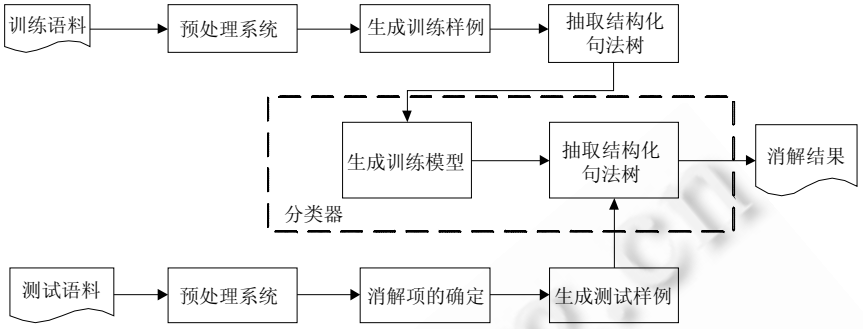


Fig.1 Framework of our system

图 1 系统框架

在英文平台上,我们线性地使用错误驱动的基于隐马尔可夫模型的命名实体识别、词性标注和名词短语识别模块[29,30]对语料进行预处理.在中文指代消解平台上,我们使用 Stanford(http://nlp.stanford.edu)的中文分词和词性标注模块以及由实验室自行开发的基于可信模型的命名实体识别模块(http//nlp.suda. edu.cn)对语料进行了线性的预处理.训练和测试中,采用了与文献[1]一致的方式进行实例的生成.

2.2 结构化句法树

众多研究成果已经表明,结构化信息对指代消解,特别是代词消解具有重要意义,但哪些结构化信息对代词消解而言是有效的,这仍然是一个悬而未决的问题.本文首先探究了 3 种基本句法树裁剪策略捕获的结构化句法树对代词消解的作用.在实验分析的基础上,我们将在第 3 节对这 3 种基本方案进行扩充,得到更加有效的结构化句法树捕获方案.

3 种基本句法裁剪策略获得的结构化句法树定义如下:

(1) 公共节点树(common nodes tree,简称 CNT):在句法树中,由指代词节点和先行语候选词节点可确定一个层次最近的公共祖先节点,而以该节点为根的子树.显然,这种句法树裁剪策略保留了与指代词和先行语候选词相关的大部分结构化句法信息.

(2) 最短路径树(shortest path tree,简称 SPT):在句法树中,以指代词节点为一端,先行语候选词节点为另一端,在句法树中形成的最短路径.显然,这种句法树裁剪策略保留了指示词和先行语候选词自身的一些特征,而它们所在的大部分上下文信息丢失了.

(3) 最小树(minimum tree,简称 MT):由最短路径包含的部分句法树.从某种意义上讲,该策略是 CNT 策略和 SPT 策略的一个折中方案.

以英文语句“John’s father loved his mother.”和中文语句“我曾直接要求藤森逮捕西蒙,并立即把他送上法庭受审.”为例,我们在完整句法树的基础上,利用上面给出的 3 种裁剪策略,可以分别得到如图 2 和图 3 所示的结构化句法树.在捕获的结构化句法树中,为了突出并区别指代词和先行语候选词,我们在指代词节点上层增加了父亲节点 *E*1,在先行语候选词节点上层增加了父亲节点 *E*2.

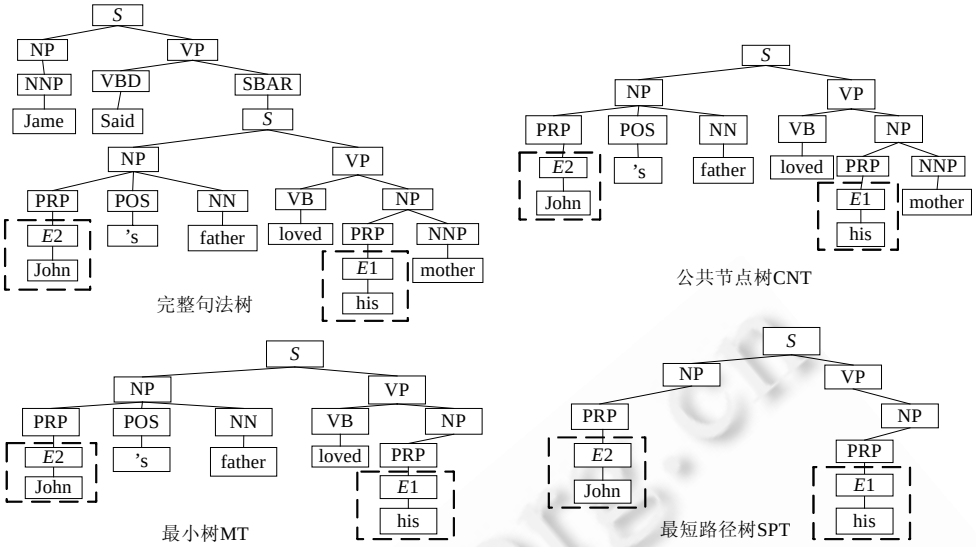


Fig.2 Three tree spans from the parse tree of the sentence “John’s father loved his mother.”

图 2 “John’s father loved his mother.”语句裁剪得到的 3 种结构化句法树

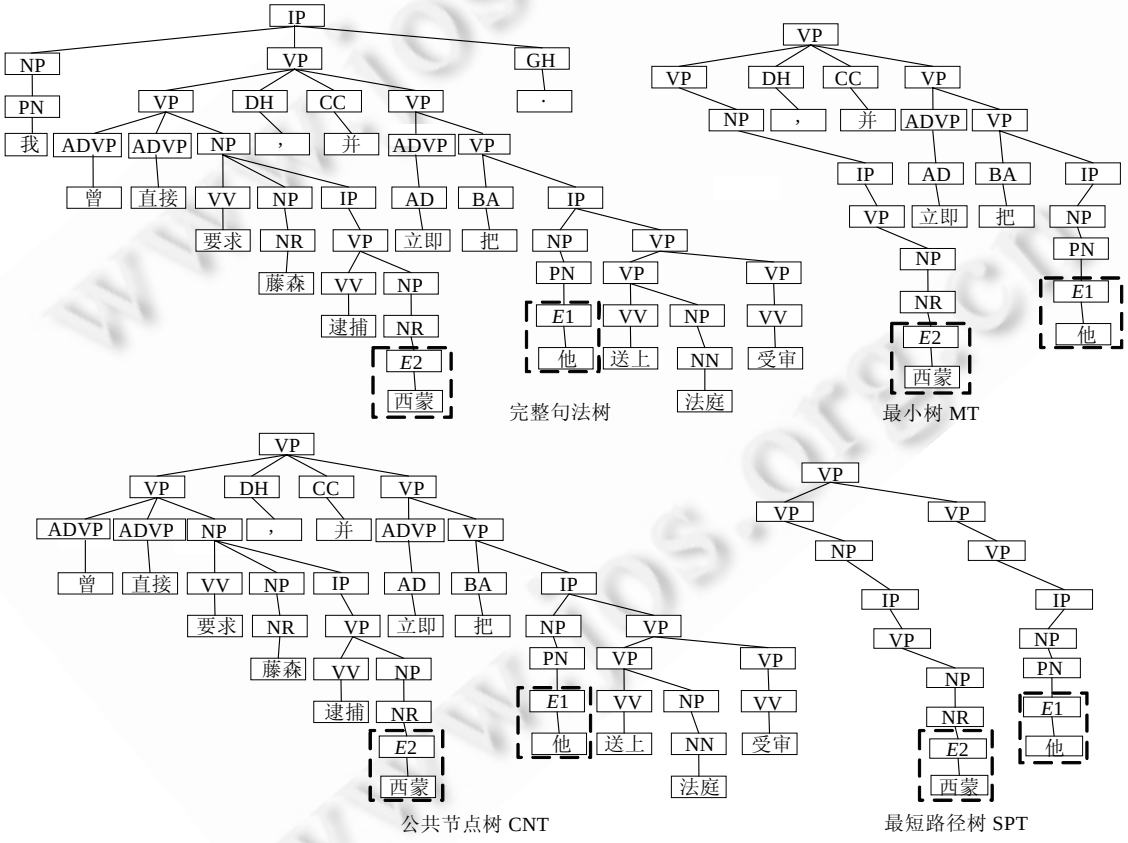


Fig.3 Three tree spans from parser tree of the sentence

“我曾直接要求藤森逮捕西蒙,并立即把他送上法庭受审.”

图 3 “我曾直接要求藤森逮捕西蒙,并立即把他送上法庭受审.”语句裁剪得到的 3 种结构化句法树

2.3 卷积树核函数

给出了上述结构化句法树后,一个关键问题就是如何利用基于树核函数的方法直接计算两个结构化句法树之间的相似度.在此,我们直接使用 SVMLight(http://svmlight.joachims.org/)中提供的卷积树核函数,该卷积树核函数已被应用于句法分析[31]、语义角色标注[32]、语义关系抽取[33]和代词指代消解[25]等领域,并取得了一定的成功.

所谓卷积核(convolution kernel)是一种通过类似卷积(*)的操作将较大的结构分解成子结构,然后首先计算子结构之间的匹配情况,再将子结构匹配的结果求和,计算出大结构的相似性.Haussler[34]和 Watkins[35]都已经证明,这一计算过程满足核函数成立的对称以及半正定条件,因此,以这种方式构造的相似函数是一个核函数,称为卷积核函数.其中,Collins 等人[22]提出的卷积树核函数是卷积核函数的一个特例,它通过列举两棵树之间的公共子树数目来计算相似度:

$$K_{CTK}(T_1,T_2)=\sum_{n_1\in N_1,n_2\in N_2}\Delta(n_1,n_2),$$

其中,N_j 代表树 T_j 中的节点集合,而$\Delta(n_1,n_2)$评价以 n_1 和 n_2 为根节点的子树的相似度,并可计算如下:

(1) 如果以 n_1 和 n_2 为根节点的上下文无关产生式(上下文无关文法规则)不准确匹配,则返回 0;否则,转步骤(2).

(2) 如果 n_1 和 n_2 是词性标记,则返回$\Delta(n_1,n_2)=\lambda$;否则,转步骤(3).

(3) 重复计算$\Delta(n_1,n_2)$如下:

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

其中,$\#ch(n)$表示节点 n 的子树个数,$ch(n,k)$是节点 n 的第 k 个子树,$\lambda(0<\lambda<1)$是一个衰退因子,用于在不同大小的子树间取得平衡.

2.4 实验与分析

本文首先使用 Charniak Parser 对 ACE 2004 NWIRE 英文语料和 ACE 2005 NWIRE 中文语料进行了句法分析,然后按第 2.2 节给出的基本策略提取代词指代消解所需的结构化信息,再交由 SVMLight 工具中附带的卷积树核函数进行学习、测试,形成分类器或进行分类判断,完成指代消解这个二元分类问题.为了保证结果的稳定性,我们使用了 5 倍交叉验证法,取平均值作为最终结果.在使用 SVMLight 工具中附带的卷积树核函数时,我们均使用了系统默认的参数,并将该卷积树核函数默认的衰退因子λ定为 0.4.

表 1 给出了 ACE 2004NWIRE 英文语料和 ACE 2005 NWIRE 中文语料上待消解代词的分布情况.从分布情况可以看出:中英文语料中,指代关系在当前句内的情况占很高的比例,约为 60%;指代关系跨越 2 句以上的情况较少,所占比例低于 15%.

Table 1 Distribution of pronoun anaphors over different sentence distances

表 1 待消解代词按句子距离的分布

代词	ACE 2004 NWIRE		ACE 2005 NWIRE	
	Train	Test	Train	Test
≤0	457	165	1 339	684
≤1	260	73	688	320
≥2	56	33	279	169
总和	773	271	2 306	1 173

表 2 给出了中英文平台对指代关系仅限于当前句的代词消解的结果.从表 2 所示的结果可以看出:

(1) 无论是中文平台还是英文平台,SPT 策略都获得了最高的准确率,说明这种策略能够尽可能地去除冗余信息,保留了指代消解任务最需要的一些关键信息.

(2) MT 策略在中英文平台都获得了最佳的召回率,但是与 SPT 策略相比,系统的准确率都有不同程度的下降.这说明 MT 策略在引入部分有效信息的同时,也带来了一些噪音.英文平台的系统准确率下降幅度

较大,而中文平台的系统准确率略有下降,说明 MT 策略引入的上下文在不同语言中对指代消解任务的影响程度并不一样.

(3) 在中英文平台中,CNT 策略的准确率较低,从而导致系统的 *F* 值偏低.在英文平台,这种情况尤为严重,与 SPT 策略相比,CNT 策略的应用,使得系统的准确率下降了 12.6%,而大量上下文信息的引入,并没有带来召回率的上升,系统的召回率反而下降了 4.8%.说明 CNT 中新引入的上下文信息,对指代消解而言意义不大.而中文平台中,情况略有差异.与 SPT 策略相比,CNT 策略的使用使得系统的召回率提升了 1.9%,准确率下降了 3.6%.说明引入的上下文信息中有部分噪音,但也包含了部分有效信息.另外,与 MT 策略相比,CNT 策略并没有提升系统的召回率,说明 CNT 策略新引入的上下文信息对指代消解任务而言意义不大,它们只会降低系统的准确率.

Table 2 Results of pronoun resolution with coreferential relationship in current sentence

表 2 指代关系在当前句内的代词指代消解性能

裁剪策略	ACE 2004 NWIRE 英文语料			ACE 2005 NWIRE 中文语料		
	R (%)	*P* (%)	*F*	*R* (%)	*P* (%)	*F*
CNT	75.9	56.8	65.0	**80.5**	66.3	72.7
MT	**80.7**	63.5	71.1	**80.5**	68.9	**74.3**
SPT	**80.7**	**69.4**	**74.6**	78.6	**69.9**	74.0

3 结构化句法树的扩展

从第 2 节给出的实验结果和分析可以看出,CNT 策略新引入的上下文信息对指代消解性能的影响总体而言是负面的.特别是在英文平台上,它既不能提升系统的召回率,又会影响系统的准确率;在中文平台上,与 SPT 策略相比,虽然提升了召回率,但 CNT 策略新增的上下文信息降低了系统的准确率,而且降幅更大,从而导致系统 *F* 值的下降.从本节开始,我们不再讨论 CNT 策略,而是尝试结合 MT 和 SPT 策略的优势,同时将一些基于平面特征的指代消解研究中已经证实有效的信息引入结构化句法树的捕获策略,来进一步探讨哪些结构化信息对指代消解是有益的.

3.1 基于中心理论的扩展

中心理论是计算语言学中的一个理论模型,主要分析了代词在语篇中的分布规律,以及影响代词实现的各种条件.文献[36]详细而全面地论述了代词是如何促进语篇连贯的,并指出语段中出现的所有话语实体都是语篇的中心,这些中心在前后语段中的突显程度以及它们的语言实现形式都会影响到语篇的连贯性.而本文给出的基于中心理论的结构化句法信息捕获方案正是试图在获取的结构化句法信息中更好地体现先行语候选在上下文中的突显度.

语篇是以话语实体中心为基础连接前后语段的,每一语段都有 3 种中心:(1) 潜在中心(forward-looking center,简称 C_f):指一个语段可能存在的会话焦点,它提供了与后继语段联系的纽带,包括一系列的对象,这些对象按照其突显度的不同形成一定的等级排列;(2) 现实中心(back-looking center,简称 C_b):指一个语段的当前会话焦点,它只包含一个对象,负责与先前语段建立联系,即前一语段的若干 C_f 中,突显度最高的一个对象就是本句的 C_b;(3) 优选中心(prefered cener,简称 C_p):指潜在中心中突显度最高的那个对象.

根据中心理论,我们可以得到如下一些结论:

(1) 当本语段中包含其他代词时,现实中心 C_b 必须以代词的形式来表示,否则就会造成这一语段阅读时间的增加.

(2) 前一语段的潜在中心 C_f 中突显度最高的一个对象就是本语段的现实中心 C_b.Grosz 等人和 Sinder 等人[37,38]进一步研究发现,在语段中,现实中心 C_b 是不受出现的先后次序、语法角色以及实施/受施等语义角色影响的,但位置以及语法角色等要素会影响潜在中心 C_f 中各对象的突显度.

(3) 各语段间频繁发生硬转(rough shift)将会影响文章内容的连贯性.为了保持文章内容的连贯性,文章的

作者必然会制定写作计划,减少焦点的切换次数.

既然某一语段的当前焦点(通常以代词的形式出现)是不受其出现的先后次序、语法角色以及施事/受施等语义角色影响的(即与上下文的相关度较低),对它的描述可以尽量简洁,可去除上下文相关信息的描述;而位置、语法角色等诸多要素(上下文信息)都会影响当前语段中潜在焦点(即可能是下一语段的焦点)的突显度,对它的描述需要适当增强上下文信息的描述.基于此,我们给出了一种融合了 MT 和 SPT 策略的扩展方案:裁剪得到公共节点树后,对于先行语候选词所在分支使用类似于 MT 的方法,裁剪掉子分支中候选词左侧的若干小分支,而保留其右侧小分支(既描述先行语候选词自身的信息,又给出其相关的上下文描述);而对于指代词所在的分支,则使用 SPT 策略,仅保留指代词到公共节点的一条路径.我们称这种扩展策略为 RSPTLMT(右 SPT 左 MT).图 4 给出了一个 RSPTLMT 的示例.

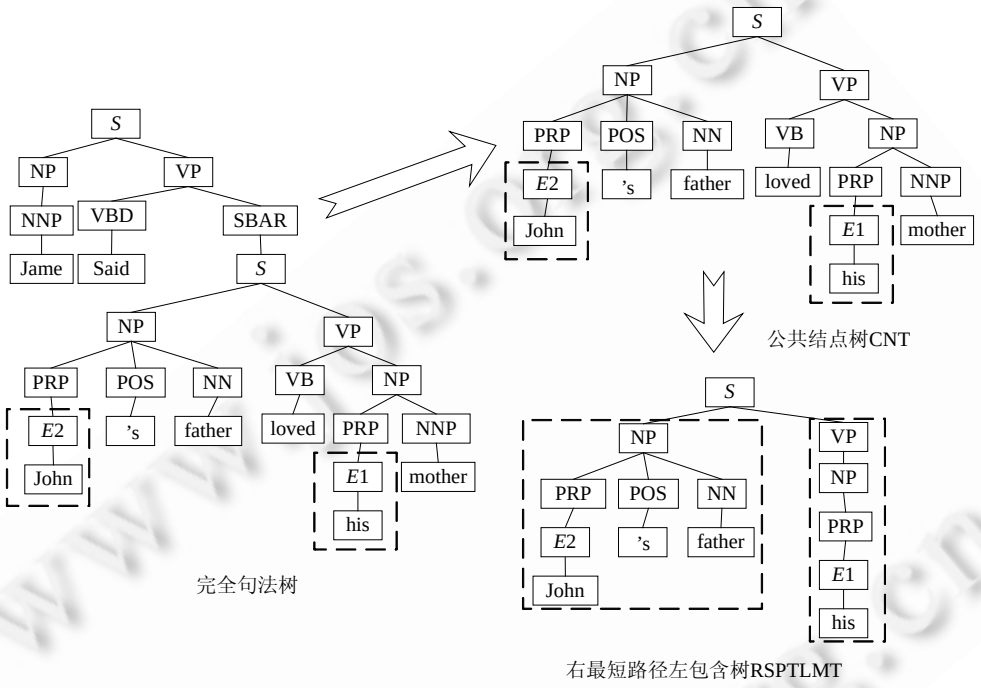


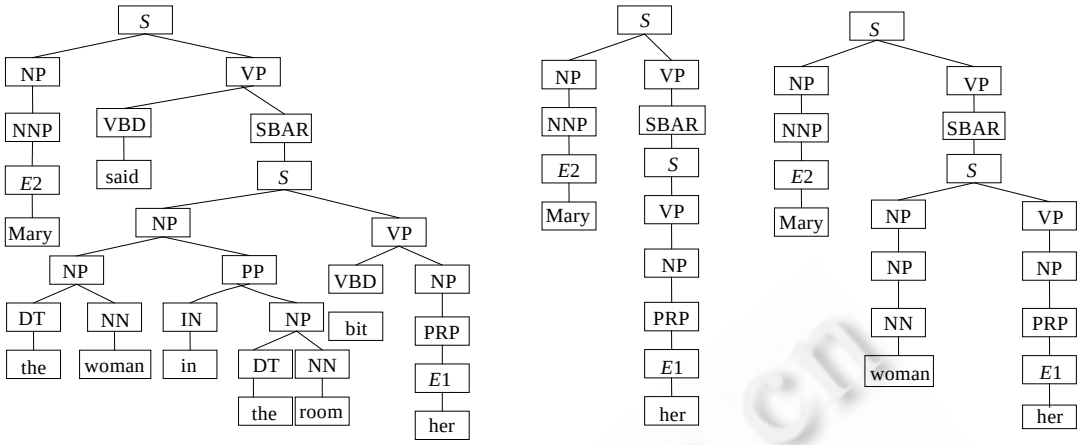
Fig.4 An example of the RSPTLMT tree span

图 4 RSPTLMT 裁剪策略示例

3.2 竞争者信息的扩展

Yang 等人[3]提出了一个双候选模型,基本思想是,通过学习各先行语候选之间的竞争关系来更好地确定先行语;Yang 等人[5]又探索了先行语候选词所在指代链的语义信息在代词(特别是中性代词)消解中的作用.他们的实验和分析都表明,竞争者信息的引入对代词消解的性能有一定程度的提升.本文在 RSPTLMT 裁剪策略上进一步扩展,引入了竞争者所在位置的结构化信息.

具体扩展过程,我们以英文句子“Mary said the woman in the room bit her.”中的 Mary 和 her 配对进行消解为例.如图 5(a)所示,首先,我们使用 Charniak Parser 获得该语句的完全句法树;然后,使用 RSPTLMT 策略得到了图 5(b)所示的结构化信息;接着进行竞争者信息扩展,基本做法是,对先行语候选词和指代词之间的所有满足单复数、性别和语义类别一致条件的名词性短语,我们在扩展过程中保留它.如图 5(c)所示,我们引入了“woman”节点到顶层公共节点的路径.虽然“woman”节点被扩展进了裁剪结果,“room”节点却未被包含,因为前者与指代词 her 的词义、单复数、性别兼容,是合法的先行语候选词的竞争者,而“room”不是.



(a) 完全句法树 (b) RSPTLMT 裁剪结果 (c) 引入竞争者信息

Fig.5 Attaching of competitor-related information

图 5 竞争者信息的扩展

3.3 驱动谓词及相关信息的扩展

谓词及其驱动的相关信息,例如语义角色等,对句子含义的理解以及句子结构的分析至关重要.在指代消解领域已有一些研究针对谓词及其驱动的语义角色展开,典型的工作是 Simone 等人[39]使用 ASSERT 工具(assert2v0.14b sameer pradhan and steven bethard)进行语义角色标注,将获得的语义角色及对应的驱动谓词对作为新特征引入指代消解,再借助最大熵分类器进行分类.实验结果表明,驱动谓词及其语义角色信息的引入能提升指代消解的召回率,但对系统的准确性几乎没有影响.本文在引入竞争者信息的基础上,对结构化信息做了进一步扩展,引入了驱动谓词以及与其相关的语义角色、所属类别等信息.具体而言,扩展包括以下 3 个步骤:

(1) 在裁剪结果上引入谓语所在的路径.

(2) 在已获得的结构化句法树中,为指代词和先行语候选词分别引入了一个 ROLE 节点,该节点的取值有 4 个,分别是:Arg0,表示当前对象在句中仅承担某一谓词驱动的施事者角色;Arg1,表示当前对象在句中仅承担某一谓词驱动的受施者角色;Args,表示当前对象在句中承担了多个谓词驱动的多个角色;NoArg,表示当前对象在语句中未承担任何语义角色.

(3) 考虑到语义角色信息与当前对象所属的类别,特别是代词类别之间关系密切,我们又在结构化句法树中扩展了一个 CLASS 节点,以表示当前对象的类别.其具体取值包括:未知类别(BareNp)、代词(PronounNp)、专有名词(ProperNp)、有定名词(DefiniteNp)、无定名词(IndefiniteNp)、指示性名词(DemonstrativeNp).其中,若 CLASS 节点为代词,我们又在其节点之下拓展了一个代词细分节点,给出了当前对象所属的具体代词类别,如第一人称、第二人称、第三人称单数、第三人称复数这 4 类.

图 6 给出了一个谓词相关信息的扩展示例.需要说明的是,通常情况下,第 3.2 节引入的竞争者也承担一定的语义角色,且这些语义角色也是由引入的谓词驱动的,也可以为每一竞争者引入谓词相关的语义角色和所属类别信息.但将此类信息引入后,我们进行的实验发现,与不引入竞争者的谓词相关信息相比,性能有了一定程度的下降.其原因可能是此类信息的引入,使得先行语候选在捕获的结构化信息中的地位下降,影响了最终的结构化信息的相似度计算.因此,我们对竞争者不进行相关信息的扩展.

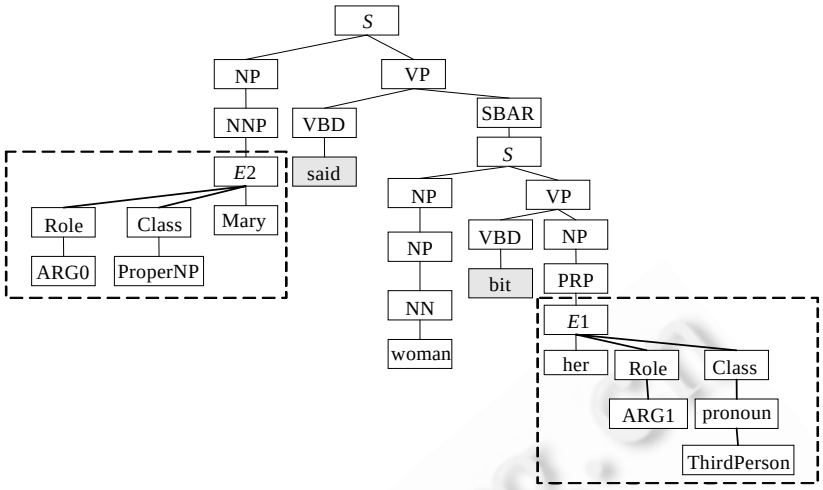


Fig.6 Attaching of semantic role-related information

图 6 语义角色相关信息的扩展

3.4 实验与分析

与 3 种基本裁剪策略类似,我们针对中英文两个平台作了一致的结构化句法树的扩展,并分别在 ACE 2004 NWIRE 英文语料和 ACE 2005 NWIRE 中文语料上进行了后续实验.为便于系统集成,我们使用了由实验室 SRL 小组自行开发的中英文 SRL 系统.其中:中文 SRL 系统在正确分词和自动句法树结果上使用中文 PropBank 作为训练、测试语料,系统 *F* 值为 69.2,而在自动分词和自动句法树结果上,该工具的性能 *F* 值也达到了 66.7;而英文 SRL 系统以 CoNLL 2005 Shared Task 给定的数据集合为实验语料,系统的总体性能 *F* 值为 78.3,达到了国际先进水平[40].

表 3 给出了基于扩展的结构化句法树的当前句代词消解的性能.从表中可以看出:

(1) 基于中心理论扩展得到的 RSPTLMT 方案很好地结合了 SPT 方案的高准确率和 MT 策略的高召回率,无论中文平台还是英文平台,都取得了更好的系统性能.

(2) 在 RSPTLMT 策略中引入竞争者信息后,虽然中英文平台的系统召回率有了一定幅度的降低,但准确率大大提升,使得系统的总体性能有了一定程度的提升.这说明竞争者信息的引入能更好地描述先行语候选词间的突显情况,因此系统得到了更高的准确率.

(3) 加入谓词及其驱动的相关信息后,系统的召回率和准确率都有了一定程度的提升,系统的 *F* 值达到了最佳状态,英文平台的性能 *F* 值达到了 80.1,而中文平台的性能 *F* 值则达到了 82.0.

Table 3 Results of pronoun resolution with coreferential relationship in current sentence using the expansion tree span

表 3 使用扩展句法树的当前句代词消解的性能

裁剪策略	ACE 2004 NWIRE 英文语料			ACE 2005 NWIRE 中文语料		
	R (%)	*P* (%)	*F*	*R* (%)	*P* (%)	*F*
RSPTLMT	**83.7**	67.8	74.9	**81.2**	71.0	75.8
+竞争者信息	79.5	72.9	76.1	75.9	81.8	78.7
+语义角色相关信息	83.1	**75.8**	**80.1**	81.0	**83.0**	**82.0**

4 进一步讨论

通过前面的描述我们发现,捕捉合适的结构化句法树的确能够有效地提升一句内代词消解的性能,但我们必须考虑以下一些问题:(1) 结构化句法树是在完全句法树基础上裁剪、扩充得到的,那么基于树核函数的代词

指代消解对句法分析器的依赖程度如何?(2) 结构化句法树对一句内代词的指代消解是有效的,当扩展到多句时,情况又如何呢?因为代词指代现象具有非常明显的局部性(通常在 2 句~3 句以内),后续我们将以结构化句法树对前、后两句内代词的消解为研究对象;(3) 与基于特征向量的系统相比,基于树核函数的中英文代词消解的性能是否具有优势?本节将针对这 3 个问题展开讨论,更进一步说明基于树核函数的代词消解方法的适用范围.

4.1 对句法分析器的依赖

前面给出的实验结果均在 Charniak Parser 句法分析结果上获得,为了检验基于树核函数的指代消解系统对句法树的依赖程度,我们又使用 Collins Parser 对 ACE 2004 NWIRE 英文语料进行了句法分析,再使用 RSPTLMT 裁剪策略捕获结构化句法树,最后进行竞争者和谓词及其相关信息的扩展.将最终得到的结构化句法树用于当前句的代词消解后,得到了如表 4 所示的结果.

Table 4 Results of pronoun resolution with coreferential relationship in current sentence using different parser tools

表 4 使用不同句法分析器结果的当前句代词消解的性能

句法分析器	*R* (%)	*P* (%)	*F*
Charniak parser	**83.1**	**75.8**	**80.1**
Collins parser	79.5	74.2	76.8

从表 4 给出的结果可以看出,不同句法分析器对基于树核函数的代词消解的性能的确有影响.非常明显,使用 Charniak Parser 句法分析器的结果得到的消解性能要好于 Collins Parser.首先,从这两种句法分析器得到的结果来看,Collins Parser 得到的句法树比 Charniak Parser 的结果要细.对于名词短语(NP),在 Charniak Parser 的结果中只有一种结论,即 NP;而 Collins Parser 的结果却会将 NP 分得很细,例如包括 NP-A 等.如果句法树的节点划分很细致,结构化句法树间相似度的计算就会变得困难,因为很多时候模式的匹配是模糊的,过于细致的句法树使得这种模糊性丢失了;其次,就总体性能而言,Charniak Parser 的性能要优于 Collins Parser.正是因为这些原因,在 Charniak Parser 上进行的指代消解性能要高于 Collins Parser 上的结果.

4.2 指代关系跨语句的代词消解

从前面各节给出的实验结果看,结构化句法树的确能很好地解决当前句的代词消解问题.但当处理对象拓展到跨越多语句时,情况又如何呢?考虑到代词指代关系具有局部性,我们以指代关系仅跨越前后两句的代词消解为例进行讨论.

句法分析器是以单个句子为单位进行分析处理的,当指代关系跨越多句时,我们首先将指代词所在语句的句法树与其前一语句的句法树合并(添加一个虚拟根节点,将两条语句对应的句法树按前后次序形成虚拟节点的子树).在新形成的句法树中,我们使用 RSPTLMT 裁剪策略捕获结构化句法树,并进行了竞争者、谓词及其相关信息的扩展,得到的代词消解性能见表 5.

Table 5 Results of pronoun resolution with cecoreferential relation in current and previous sentences

表 5 指代关系跨越 2 句的代词消解的性能

	ACE 2004 NWIRE English corpus			ACE 2005 NWIRE Chinese corpus		
	R (%)	*P* (%)	*F*	*R* (%)	*P* (%)	*F*
≤0	83.1	75.8	80.1	81.0	83.0	82.0
≤1	75.4	77.0	76.2	72.8	75.4	74.1

从表 5 给出的实验结果可知,基于树核函数的两句内代词消解的性能明显低于一句内代词消解的性能.说明结构化句法树对于一句内固定模式的代词指代具有较好的识别能力;而对于跨越超过一句的代词指代,结构化句法树对于某些固定模式的识别能力降低了,使得消解性能下降.

4.3 与基于特征向量的代词消解的比较

到目前为止,基于树核函数的代词消解的研究刚刚起步,而基于特征向量的代词消解已经取得了相当的成

功.那么与基于特征向量的代词消解相比,基于树核函数的代词消解的性能如何?这种方法又有什么优势呢?基于此目的,我们构建了一个基于特征向量的中英文消解平台,其中,英文指代消解平台使用了与文献[3]完全一致的特征集合;而中文平台使用的特征集见表 6,这些特征均已被众多研究者证实对指代消解任务是有效的.

Table 6 Feature set of the Chiniese platform in coreference resolution

表 6 中文指代消解平台的特征集

特征	描述
ANPronoun	若照应语是代词,则取 1;否则取 0
ANDefiniteNP	若照应语是有定名词短语,则取 1;否则取 0
ANDemonstrativeNP	若照应语是指示性名词短语,则取 1;否则取 0
CAPronoun	若先行语是代词,则取 1;否则取 0
ANCAGenderAgreement	若照应语和先行语满足词性一致,则取 1;否则取 0
ANCANumberAgreement	若照应语和先行语满足单复数一致,则取 1;否则取 0
ANCAAppositive	若照应语和先行语是同位语,则取 1;否则取 0
ANCAHeadStringMatch	若照应语和先行语满足中心词匹配,则取 1;否则取 0
ANCASentDistance	照应语和先行语在 1 句内取 1,2 句取 0.9,…,大于 10 句取 0
ANCAWORDSENSE	若从 HowNet 中获得的语义信息类有相同的,则为 1;否则为 0
ANCABothProperName	若照应语和先行语候选词均为专有名词,则取 1;否则取 0
ANCANameAlias	若照应语和先行语候选词存在别名关系,则取 1;否则取 0
CAARG0	若先行语承担 Arg0 语义角色,则取 1;否则取 0
CAARG0MainVerb	若先行语承担的 Arg0 语义角色是由主谓词驱动,则取 1;否则取 0
ANCASameTarget	若先行语和照应语承担的语义角色是由相同谓词驱动,则取 1;否则取 0
ANPronounType	照应语若为代词,其具体的代词细分类别
CAPronounType	先行语若为代词,其具体的代词细分类别

首先,通过实验比较,说明我们构建的基于特征向量的中英文消解平台具有国际先进性.

基于 ACE 语料的英文消解,目前报道的性能主要是基于 ACE 2003 语料.就代词消解而言,目前报道的消解性能均使用精确率这一指标(即系统正确消解的指示词的个数占需要消解的指示词总数的比例).该指标假设指示词是已知的,而在实际的指代消解中,判断当前名词短语是否是指示词也是指代消解任务的一个重要环节.但为了能与同类系统进行比较,我们基于 ACE 2003 的 3 个子语料对构建的基于特征向量的英文平台也假设指示词是已知的,并使用精确率进行了评测,得到的消解性能见表 7.

Table 7 Results of pronoun resolution on ACE 2003 English corpus

表 7 ACE 2003 英语语料上代词消解的性能

系统	NWIRE	NPAPER	BNEWS
基于特征向量的系统	71.8	76.4	77.3
Yang 等人的系统[4]	72.9	77.1	74.9

从表 7 所示的结果可以看到,我们构建的英文平台的代词消解性能与文献[4]所报道的性能相当,达到了国际先进水平.

基于 ACE 语料的中文消解研究,至今还未见到代词消解性能的相关报道,但对所有名词短语的消解性能而言,Ngai 等人[41]给出的在 ACE 2005 BNEWS 语料上的最佳 F 值性能为 77.2[41],周俊生等人[15]给出的在同一语料上的 F 值性能为 60.06[15],而我们构建的中文平台在相同语料上评测得到的 F 值为 79.3,优于同类系统.

接着,我们分别就所构建的中英文平台,针对代词使用相同的语料对基于特征向量的系统和基于树核函数的系统进行了评测,评测结果见表 8.从表 8 所示的结果可以看出,与特征向量的系统相比,基于树核函数的系统具有更好的召回率,而准确率略有下降,系统的 F 值性能均略高于特征向量平台.这说明树核函数能更有效地使用结构化信息.

Table 8 Comparison of feature-based pronoun resolution and kernel-based pronoun resolution

表 8 特征向量的系统与树核函数的系统的代词消解性能比较

系统	ACE 2004 NWIRE 英文语料			ACE 2005 NWIRE 中文语料		
	R (%)	*P* (%)	*F*	*R* (%)	*P* (%)	*F*
基于特征向量的系统	65.2	78.7	70.3	64.2	78.1	70.5
基于树核函数的系统	72.2	76.2	74.2	70.6	72.3	71.4

5 基于扩展的结构化句法树的指代消解错误分析

我们对基于树核函数的代词消解系统结果做了进一步的错误分析,发现错误产生的主要原因有:

(1) 非待消解项误判.指代消解由两个子任务构成:第一,待消解项识别:确定语篇中哪些名词短语需要进行指代消解;第二,待消解项消解:对识别出的待消解项进行消解.本系统重点探讨结构化信息对待消解项消解的影响,关于待消解项识别,我们利用标注语料训练生成了一个待消解项识别分类器,并将它简单地用作过滤器.由于待消解项分类器存在误判,将某些非待消解项识别成了待消解项,造成后续指代消解的错误.例如,语句“If a Hollywood studio makes a movie about the Florida election standoff , would the title be ***It***'s a Chad.”中的“It”是一个非待消解项,但本系统却将它判别成指代前面的某个名词.

(2) 指代事件、多个短语或句群的指代关系被误判.在 ACE 2004 和 ACE 2005 语料中,都将指向事件或句群的代词标注为非待消解项,而在实际的指代消解过程中,系统无法识别这一现象,误将这类代词与前面的名词构成指代关系.例如,语句“For most of this year, economies have performed a sort of levitating act growing by double-digit rates month after month even as economists predicted ***they*** would fall to earth.”中的“they”就属于这类情况.

(3) 结构化句法信息不足或过强.一方面,在独立的片段中有些指代关系连人都很难分辨,它需要大量的上下文信息.目前,我们使用的结构化句法信息仍然无法涵盖其所需的上下文信息.例如,语句“Will cheerful chorus members dressed in V-neck sweaters interrupt their serenades to online shopping long enough to tell ***an advertising columnist***, ***You*** ask a lot of questions for someone from Brooklyn?”中的“an advertising columnist”和“you”之间的指代关系就属于这一情况;另一方面,有些句式在绝大多数情况下都存在指代关系,但也存在特例.例如,语句中的主语与其从句中代词形式的主语之间一般都具有指代关系.结构化句法信息过于强调固定句式,会误判某些特例.

(4) 预处理错误.指代消解需要词性标注、短语识别、命名实体识别等众多自然语言处理技术的支持,这些预处理工作或多或少都会产生错误,有些错误会误导后续的指代消解.

6 总结与展望

本文给出了一种基于卷积树核函数的中英文代词消解方法.首先,探索了 3 种基本结构化句法树捕获方案;然后,在对实验结果进行分析的基础上,从基于中心理论的扩展、竞争者信息的扩展和驱动谓词及其相关信息的扩展这 3 个方面对捕获的结构化句法树进行了扩充.在 ACE 2004 NWIRE 英语语料和 ACE 2005 NWIRE 中文语料上的实验结果表明,扩充得到的结构化句法树能够很好地完成代词消解任务.另外,本文还针对不同句法分析器以及跨语句的指代关系的消解问题做了进一步的讨论,并对基于扩展的结构化句法树的指代消解进行了系统的错误分析.

References:

[1] Soon WM, Ng HT, Lim DCY. A machine learning approach to coreference resolution of noun phrases. Computational Linguistics, 2001,27(4):521−544. [doi: 10.1162/089120101753342653]

[2] Ng V, Cardie C. Improving machine learning approaches to coreference resolution. In: Proc. of the ACL 2002. 2002. 104−111. [doi: 10.3115/1073083.1073102]

[3] Yang XF, Zhou GD, Su J, Tan CL. Coreference resolution using competition learning approach. In: Proc. of the ACL 2003. 2003. 177−184. [doi: 10.3115/1075096.1075119]

[4] Yang XF, Su J, Tan CL. A twin-candidate model for learning-based anaphora resolution. Computational Linguistics, 2008,34(3): 327−356. [doi: 10.1162/coli.2008.07-004-R2-06-57]

[5] Yang XF, Su J, Zhou GD, Tan CL. Improving pronoun resolution by incorporating coreferential information of candidates. In: Proc. of the ACL 2004. 2004. 127−134. [doi: 10.3115/1218955.1218972]

[6] Bergsma S, Lin DK. Bootstrapping path-based pronoun resolution. In: Proc. of the COLING-ACL 2006. 2006. 33−40. [doi: 10.3115/1220175.1220180]

[7] Ng V. Semantic class induction and coreference resolution. In: Proc. of the ACL 2007. 2007. 536−543.

[8] Kong F, Zhou GD, Zhu QM. Employing the centering theory in pronoun resolution from the semantic perspective. In: Proc. of the EMNLP 2009. 2009. 987−996. [doi: 10.3115/1699571.1699641]

[9] Wang HF. Survey: Computational models and technologies in anaphora resolution. Journal of Chinese Information Processing, 2002,16(6):9−17 (in Chinese with English abstract).

[10] Wang HF, He TT. Research on Chinese pronominal anaphora resolution. Chinese Journal of Computers, 2001,24(2):6−13 (in Chinese with English abstract).

[11] Wamg HF, Mei Z. Robust pronominal resolution within Chinese Text. Journal of Software, 2005,16(5):700−707 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/16/700.htm [doi: 10.1360/jos160700]

[12] Zhang W, Zhou CL. Study on meta-anaphoric resolution in Chinese discourse understanding. Journal of Software, 2002,13(4): 732−738 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/20020438.htm

[13] Wang XB, Zhou CL. Study on Chinese pronominal anaphora resolution based on discourse representation theory. Journal of Xiamen University (Natural Science), 2004,43(1):31−35 (in Chinese with English abstract).

[14] Li GC, Luo YF. Chinese pronominal anaphora resolution via a preference selection approach. Journal of Chinese Information Processing, 2005,19(4):24−30 (in Chinese with English abstract).

[15] Zhou JS, Huang SJ, Chen JJ, Qu WG. A new graph clustering algorithm for Chinese noun phrase coreference resolution. Journal of Chinese Information Processing, 2007,21(2):77−82 (in Chinese with English abstract).

[16] Yang Y, Li YC, Zhou GD, Zhu QM. Research on distance information for anaphora resolution. Journal of Chinese Information Processing, 2008,22(5):39−44 (in Chinese with English abstract).

[17] Wang HD, Hu NQ, Kong F, Zhou GD. Research on semantic role information in anaphora resolution. Journal of Chinese Information Processing, 2009,23(1):23−29 (in Chinese with English abstract).

[18] Hobbs JR. Resolving pronoun references. Lingua, 1978,44(4):311−338. [doi: 10.1016/0024-3841(78)90006-2]

[19] Lappin S, Leass HJ. An algorithm for pronominal anaphora resolution. Computational Linguistics, 1994,20(4):535−561.

[20] McCord MC. Slot grammar: A system for simpler construction of practical natural language grammars. In: Proc. of the Int'l Symp. on Natural Language and Logic. 1990. 118−145. [doi: 10.1007/3-540-53082-7_20]

[21] Yang XF, Su J, Tan CL. Improving pronoun resolution using statistics—Based semantic compatibility information. In: Proc. of the ACL 2005. 2005. 165−172. [doi: 10.3115/1219840.1219861]

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

[23] Culotta A, Sorensen J. Dependency tree kernels for relation extraction. In: Proc. of the ACL 2004. 2004. 423−429. [doi: 10.3115/1218955.1219009]

[24] Bunescu RC, Mooney RJ. A shortest path dependency kernel for relation extraction. In: Proc. of the EMNLP 2005. 2005. 120−128. [doi: 10.3115/1220575.1220666]

[25] Yang XF, Su J, Tan CL. Kernel-Based pronoun resolution with structured syntactic knowledge. In: Proc. of the COLING-ACL 2006. 2006. 41−48. [doi: 10.3115/1220175.1220181]

[26] Zhou GD, Kong F, Zhu QM. Context-Sensitive convolution tree kernel for pronoun resolution. In: Proc. of the IJCNLP 2008. 2008. 1−8.

[27] Kong F, Li YC, Zhou GD, Zhu QM. Exploring syntactic features for pronoun resolution using context-sensitive convolution tree kernel. In: Proc. of the IALP 2009. 2009. 201−205. [doi: 10.1109/IALP.2009.49]

[28] Song W, Qing B, Lang J, Liu T. Combining syntax and word sense for Chinese pronoun resolution. Journal of Chinese Information Processing, 2008,22(6):8−13 (in Chinese with English abstract).

[29] Zhou GD, Su J. Error-Driven HMM-based chunk tagger with context-dependent lexicon. In: Proc. of the 2000 Joint SIGDAT Conf. on Empirical Methods in Natural Language Processing and Very Large Corpora. 2000. 71−79. [doi: 10.3115/1117794.1117803]

[30] Zhou GD, Su J. Named entity recognition using an HMM-based chunk tagger. In: Proc. of the 40th Annual Meeting of the Association for Computational Linguistics (ACL). 2002. 473−480. [doi: 10.3115/1073083.1073163]

[31] Collins M, Duffy N. New ranking algorithms for parsing and tagging: Kernels over discrete structures, and the voted perceptron. In: Proc. of the ACL 2002. 2002. 28−136. [doi: 10.3115/1073083.1073128]

[32] Moschitti A. A study on convolution kernels for shallow semantic parsing. In: Proc. of the ACL 2004. 2004. 335−342. [doi: 10.3115/1218955.1218998]

[33] Zhang Z. Weakly supervised relation classification for information extraction. In: Proc. of the CIKM 2004. 2004. 8−13. [doi: 10.1145/1031171.1031279]

[34] Haussler D. Convolution kernels on discrete structures. Technical Report, UCSCCRL-99-10, Santa Cruz: University of California at Santa Cruz, 1999.

[35] Watkins C. Dynamic alignment kernels. Technical Report, CSD-TR-98-11, Royal Holloway, University of London, 1999.

[36] Gordon PC, Grosz BJ, Gilliom LA. Pronouns, names and the centering of attention in discourse. Cognitive Science,1993,17(3): 311−348.

[37] Grosz BJ, JoShi AK, Weinstein S. Centering: A framework for modeling the local coherence of discourse. Computational Linguistics, 1995,21(2):203−225.

[38] Gordon PC, Searce KA. Pronominalization and discourse coherence, discourse structure and pronoun interpretation. Memory and Cognition, 1995,23(3):313−323.

[39] Ponzetto SP, Strube M. Semantic role labeling for coreference resolution. In: Proc. of the EACL 2006. 2006. 143−146.

[40] Li JH, Zhou GD, Zhu QM, Qian PD. Syntactic parsing with hierarchical modeling. In: Proc. of the AIRS 2008. LNCS 4493, 2008. 561−566.

[41] Ngai G, Wang CS. A knowledge-based approach for unsupervised Chinese coreference resolution. Computational Linguistics and Chinese Language Processing, 2007,12(4):459−484.

附中文参考文献:

[9] 王厚蜂.指代消解的基本方法和实现技术.中文信息学报,2002,16(6):9−17.

[10] 王厚峰,何婷婷.汉语中人称代词的消解研究.计算机学报,2001,24(2):6−13.

[11] 王厚峰,梅铮.鲁棒性的汉语人称代词消解.软件学报,2005,16(5):700−707. http://www.jos.org.cn/1000-9825/16/700.htm [doi: 10.1360/jos160700]

[12] 张威,周昌乐.汉语语篇理解中元指代消解初步.软件学报,2002,13(4):732−738. http://www.jos.org.cn/1000-9825/20020438.htm

[13] 王晓斌,周昌乐.基于语篇表述理论的汉语人称代词的消解研究.厦门大学学报(自然科学版),2004,43(1):31−35.

[14] 李国臣,罗云飞.采用优先选择策略的中文人称代词的指代消解.中文信息学报,2005,19(4):24−30.

[15] 周俊生,黄书剑,陈家骏,曲维光.一种基于图划分的无监督汉语指代消解算法.中文信息学报,2007,21(2):77−82.

[16] 杨勇,李艳翠,周国栋,朱巧明.指代消解中距离特征的研究.中文信息学报,2008,22(5):39−44.

[17] 王海东,胡乃全,孔芳,周国栋.指代消解中语义角色特征的研究.中文信息学报,2009,23(1):23−29.

[28] 宋巍,秦兵,郎君,刘挺.句法与词义相结合的中文代词消解.中文信息学报,2008,22(6):8−13.

孔芳(1977－),女,江苏扬州人,博士,副教授,CCF 会员,主要研究领域为自然语言处理,信息抽取,信息融合.

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,机器学习,机器翻译.