中英文指代消解中待消解项识别的研究 *

孔芳　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)
(江苏省计算机信息处理技术重点实验室，江苏，苏州，215006)

Anaphoricity Determination for Coreference Resolution in English and Chinese Languages

Kong Fang, Zhu Qiaoming, Zhou Guodong
(School of Computer Science & Technology, Soochow University, Suzhou 215006, China)
(Key Lab of Computer Information Processing Technology of Jiangsu Province, Suzhou, 215006，China)

Abstract　Systematically explores noun phrase anaphoricity determination for coreference resolution in both English and Chinese languages in various ways. First, using a rule-based method to detect the non-anaphors which are insensitive to the context or have some obvious patterns. Then, exploring both flat feature-based and structured tree kernel-based methods to determinate the non-anaphors sensitive to the context. Finally, a composite kernel is proposed to combine the flat features with structured ones to further improve the performance. Experimentation results on both the ACE 2003 English corpus and the ACE 2005 Chinese corpus show that all the proposed methods perform well on anaphoricity determination. In addition, the anaphoricity determination module is applied to coreference resolution systematically. Experimentation results show that proper anaphoricity determination can significantly improve the performance of coreference resolution in both English and Chinese languages.
Key Words　noun phrase anaphoricity determination; rule-based method; flat feature-based method; structured tree kernel-based method; composite kernel

摘 要　深入研究了中英文指代消解中的待消解项识别问题。在前人工作的基础上，首先使用规则方法识别与上下文无关或具有显著固定模式的非待消解项；接着针对与上下文相关的非待消解项识别，从平面特征方法和结构化树核函数方法两方面入手进行了探索；最后利用复合核函数将平面特征和结构化特征有效结合，对待消解识别问题进行了进一步研究。在 ACE 2003 英文语料和 ACE 2005 中文语料上的实验结果表明，提出的多种待消解项识别方案各具特色，都取得了不错的性能。最后将得到的待消解项识别模块应用于中英文的指代消解任务，实验结果表明，合适的待消解项识别能够大大提高中英文指代消解的性能。
关键词　待消解项识别；规则方法；平面特征方法；结构化树核函数方法；复合核函数
中图分类号　TP18

0. 引言

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之间存在的特殊语义关联，且其语义解释依赖于前者。在语言学中把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），而确定照应语所指的先行语的过程就是指代消解。随着篇章理解、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。

指代消解由两个子任务构成：1）待消解项识别：确定篇章中哪些名词短语需要进行指代消解；2）指代消解：对识别出的待消解项进行消解。目前绝大多数指代消解系统都忽略了待消解项识别任务，认为所有的名词短语都

* Supported by the National Natural Science Foundation of China under Grant Nos. 60873150, 90920004, 61003153 (国家自然科学基金); the National Research Foundation for the Doctoral Program of Higher Education of China under Grant No. 200802850006 (国家教育部博士点基金).
Corresponding author: 周国栋， E-mail: gdzhou@suda.edu.cn

需要参与指代消解。这种假设的好处是简化了指代消解任务，并取得了不错的成果。例如，Soon 等（2001 年）[1]将指代消解转化成一个二元分类问题，提出了一个完整的指代消解框架（忽略了待消解项识别），并给出了详细的实现步骤。在此基础上，Ng 等（2002）[2]对 Soon 等（2001）的研究进行了扩充，探索了一系列词法、语法和语义特征，在 MUC-6 上的性能 F 值达到了 69.4；钱伟等（2003）[3]提出了一种基于语料库的英文名词短语指代消解算法，在 MUC-7 上的性能 F 值达到了 60.2；Yang 等（2004）[4]探索了先行语侯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（2006）[5]提出了一种基于路径的代词指代消解方法。在中文指代消解领域，研究者尝试将类似英文的指代消解平台应用于中文指代消解的研究，也取得了一定的成绩。例如：李国臣等(2005)[6]采用决策树方法对中文人称代词的消解进行了研究；宋巍等(2008)[7]给出了一种句法与词义相结合的中文代词消解方案，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，随着指代消解研究的日益深入，待消解项识别的重要性日益显现。作为指代消解中一个非常重要的子任务，随着指代消解性能的不断提升，待消解项识别也日益成为制约指代消解性能提升的一个重要环节。Stoyanov 等（2009）[8]详细分析了影响指代消解性能的三个主要因素，待消解项的识别就是其中之一。他们的分析发现加入待消解项识别可以有效减少错误的消解元素（Coreference Elements），使需要消解的元素集更接近正确标注的元素集，在 MUC、ACE 语料上进行的实验显示，完美的待消解项识别对指代消解 F 性能的贡献可达 6-10。Bergsma 等（2008）[9]统计发现，作为英文中出现频度最高的一个词“it”，约有 25%~50%是非待消解项；同样，中文中非待消解项也普遍存在，例如“管他春夏与秋冬”中的“他”就是一个非待消解项。当然，指代消解过程中非待消解项的引入会带来一定程度的噪音，从而影响指代消解的性能（Bean 等 1999）[10]。

本文主要讨论中英文指代消解中待消解项的识别问题。本文后续内容组织如下：第 1 节给出中英文指代消解中待消解项识别的相关研究；第 2 节给出一个规则与机器学习方法相结合的待消解项识别方案；第 3 节给出在 ACE 中英文语料上，待消解项识别的性能评测；第 4 节，将构建的待消解项识别模块引入指代消解，说明待消解项识别与指代消解的关系；最后第 5 节给出小结和下一步的工作设想。

1. 相关工作

伴随着指代消解研究的不断深入，待消解项识别的研究也在不断深入。总体而言，待消解项识别的研究方法主要包含三种：1）根据语言学知识设定若干规则来描述非待消解项，对其进行基于规则的识别；2）基于语料库统计结果的待消解识别，即在语料库中统计固定模式，再依据统计结果进行待消解识别；3）根据语言学的相关知识，提取相应的语义、语法等上下文特征，基于机器学习方法，利用标注语料来识别待消解项。

指代消解研究的早期，一些研究者已经意识到待消解项识别的重要性，他们从语言学知识出发设定规则来进行待消解项的识别，代表性的工作包括：Lappin 和 Leass（1994）[11]在其指代消解平台中引入了用于识别“it”是否待消解项的独立识别模块。通过设定一些模式，例如：“It is Cogv-ed that Sentence”，其中 Cogv 是像 think、believe、know 这样的认知动词，识别模块将可能的指代词所在的句式与设定的模式进行比较。若模式匹配，则认为遇到的“it”为非待消解项。显然，基于规则的方法可移植性较差，一旦语言变化，需要对根据语言学知识设定的规则进行相应的调整。

随着标注语料库的出现，基于语料库的待消解项识别方法不断涌现，典型的工作包括：Bean 和 Riloff（1999）[10]提出了基于语料库统计结果的待消解项识别方法。他们从语料库入手，统计形成几组待消解项识别规则，再将规则应用于待消解项识别任务。Bergsma 等（2008）[9]利用代词的局部上下文句式进行待消解项识别。他们对每一个代词都按一定规则提取其所在上下文的句式，然后在大型语料库中对该句式的出现频度进行统计，再根据统计结果判断这个代词是否待消解项。不过他们没有将其构建的待消解项识别模块应用于后续的指代消解，对指代消解（特别是代词消解）性能的影响仍然不确定。

近年来，随着指代消解研究的不断深入，研究者开始转入利用机器学习方法进行待消解项识别，典型的工作包括：Ng 等（2002）[12]选取了包括词法、语法、句式、语义、位置等多方面的 37 个特征，给出了一种基于机器学习方法的待消解项识别方法，并将生成的待消解项识别模型应用于指代消解，通过实验证明了待消解项识别模块的引入能进一步提高指代消解系统的性能。Ng 等（2004）[13]在其 2002 年工作的基础上，进一步讨论了在指代消解中如何更有效地使用识别出的待消解项信息。他们主要探讨了将待消解项信息有效应用于指代消解的两种不

同方法：将待消解项信息作为过滤器和将待消解项信息作为特征之一，并针对这两种方法探讨了如何进行局部和全局优化。Yang 等（2005）[14]给出一个具有待消解项识别功能的基于双候选模型的指代消解系统。对于正在处理的名词短语，若所有先行语候选词与它配对的消解结果均低于某个阈值，则将该名词短语视为非待消解项。Zhou 等（2009）[15]利用标记传播算法，在机器学习的基础上对待消解项识别进行了全局优化，取得了较好的效果，并通过实验表明待消解项识别任务对指代消解而言是非常重要的。

中文指代消解中有关待消解项识别问题的研究很少，典型的工作有：Ngai 等（2007）[16]给出了一种基于知识的无指导的中文指代消解系统。该系统将指代消解任务分成待消解项识别和指代消解两个环节，其中待消解项识别根据中文的语言学知识设立了规则集，通过规则的方法，在保持高召回率的情况下完成待消解项识别任务。该系统还通过 ACE 2005 中文语料上的实验结果说明了待消解项识别对中文指代消解的贡献。

本文将在前人研究的基础上给出一个规则与机器学习方法相结合的待消解项识别方案，并分别在 ACE 2003 英文语料和 ACE 2005 中文语料上对该方案进行系统评测。最后，本文将构建的待消解项识别模块引入指代消解，通过实验说明待消解项识别对中英文指代消解性能的影响。

2 规则与机器学习方法相结合的待消解项识别方案

规则方法需要专门的语言学知识，与具体语言相关。正是语言相关，使得规则对语言现象的描述通常比较准确，能高效捕获指定的语言现象，缺点是各种语言现象极难穷尽以及某些复杂语言现象很难用规则有效描述。与规则方法相比，机器学习方法无需领域专家对专门的语言现象进行细致专业的描述，只需从已标注的语料中提取相关的特征，就能形成有效的训练模型，利用生成的训练模型，机器就具有了一定的判断语言现象的能力。

与其他一些 NLP（Natural Language Processing）任务不同，判断指代消解中当前对象是否待消解项，有些情况需要借助上下文，而另一些情况则可能是与上下文无关的（Bean 等 1999）[10]。其中，与上下文无关的待消解项不仅很难利用机器学习方法进行判断，而且有关它们的描述还可能会对训练模型产生一定的干扰，因此，这类现象适于规则方法进行描述。对于与上下文相关的待消解项识别而言，机器学习方法的引入能在一定程度上提升系统的自动化程度以及可移植性。基于上述考虑，我们将给出一个规则与机器学习方法相结合的待消解项识别方案，规则部分用于进行上下文无关以及语言特定规律显著的待消解项识别，而机器学习方法则用于捕获一些相关的上下文信息，进行一些语言相关度较低的待消解项识别。

需要特别说明的是：规则和平面特征的选取，在一般情况下都与特定的语言有关，为了构建统一平台进行多语言的处理，我们设定了一个包含全部中英文规则和平面特征的集合，处理时再根据具体的语言采用所需的子集进行待消解项的识别；结构化特征的选取，在中英文中我们使用了统一的捕获策略，体现了基于结构化特征的待消解项识别方案具有更好的可移植性。

2.1 规则集

规则方法主要用于与上下文无关或语言固定模式显著的待消解项的识别，在中英文平台，我们使用了如下一些规则：

1） 中英文均适用的规则集：
   a) 直接修饰其他名词的命名实体是非待消解项。例如：英文短语“the IBM CEO”中的命名实体“IBM”就是一个非待消解项。中文短语“美国总统”中的“美国”也是一个非待消解项。
   b) 文章第一句话中出现的专有名词、英文中的有定名词是非待消解项。例如：文章第一句中为“联想公司 2009 年第二季度盈利……”，其中专有名词“联想公司”就是一个非待消解项。
   c) 根据领域知识设定的专有名词、英文中的有定名词在现实中具有众所周知的含义，它们通常也是非待消解项。例如：“FBI”特指美国联邦调查局。

2） 英文平台专用的规则集：有些非待消解项出现在固定句式中，且使用频度很高，可通过固定句式匹配规则进行识别。本文主要针对英文中的“it”系表结构进行了规则判别，参考 Lappin 和 Leass（1994）[11]的研究成果，我们总结了三种“it”系表结构：关于气象的“it”，关于时间的“it”以及被动结构中的“it”。这三类系表结构在词语的顺序上有一定的规律，表 1 给出了这些规律表现形式的归纳结果，表中的符号“*”表示当前词性在这种形式中可以出现多个同词性的词。

Table 1 Rules for three kinds of "it" strucutres

表 1　三种“it”系表结构的规则表现

Category	Expression	Description
About weather	It <be> ADJ	ADJ is an adjective about weather, e.g. cloudy、sunny、windy.
	It <be> ADV* PCP	PCP is a noun about weather, e.g. raining、snowing.
	It <be> N	N is a noun about season, e.g. spring.
About time	It <be> ADV* NUM	NUM means digital. E.g. It is three o’clock。
	It <be> PREP* time	PREP means prep. E.g. It’s about time。
	It <be> ADV* （early | late）	E.g. It’s early。
Passivity	It <be> modadj that	It is the most common form. The modadj is a form adjective, e.g. easy.
	It <be> modadj [for N] to V	[for N] is optional.
	N Vcog it modadj [for N] to V	Vcog means emotive verb.

3） 中文平台特有的规则集：

a) 停用词规则：中文中有些抽象名词是非待消解项，例如“喜悦”、“骄傲”、“正常”、“特色”、“前提”等名词，它们本身不是实体，也不会指代前文的某个实体。针对这类现象，本文给出了一个停用词表，所有出现在词表中的名词短语均被认为是非待消解项。

b) 保留规则集：将符合特定模式的某些名词短语保留成待消解项候选，再使用其他规则进行是否是待消解项的进一步判断。

c) 排除规则集：将符合特定模式的某些名词短语标识成非待消解项，不再参与其他规则的判断。

Table 2 keeping rules and removing rules used in Chinese platform

表 2 中文平台的保留和排除规则集

Category	Rule	Examples
Keep rules	Non-recursive NPs	(NP (NP (NR_LOC 中国)(NR_ORG 外交部))(NP (NN 官员))) -> “中国外交部”、“官员”
	Phrases match the pattern “NUM[+QUN]+NP”	(NP (QP (CD 一)(CLP (M 名)))(NP (NN 旅客))) -> “一名旅客”
	Demonstrative NPs	(NP (DP (DT 这些))(NP (NN 经济)(NN 活动))) -> “这些经济活动”
	Proper NPs	((NP (NR_PER 蔡振南)) -> “蔡振南”
	Personal pronouns	(NP (PN 他们)), (NP (PN 她)), (NP (PN 你们)), (NP (PN 我))
	NPs including more than two words	(NP (NN 核子)(NN 设施)) -> “核子设施”
	NPs match some specific patterns	(DT 其他)(NR_LOC 中国)(NN 官员)
	The two nouns in “NN+NR”	(NP (NN 记者)(NR_PER 宫能惠)) -> “记者”、“宫能惠”
	The two nouns in “NN+DEC+NN”	(NP (NN 投资者)(DEC 的)(NN 法宝)) -> “投资者”、“法宝”
	The two nouns in “NN+CC+NN”	(NP (NR_LOC 中国)(CC 和)(NR_LOC 美国)) -> “中国”、“美国”
Exclusionary rules	NPs including time	(NP (NT 明天)(NT 2 号)(NT 晚上))
	Quantifiers	(NP (CD 十)(CLP (M 名)))
	NPs start with prep.	(NP (P 对)(NN 生命)(NN 财产))
	NPs only include single word and is not proper NPs	(NP (NN 字)), (NP (NN 月)).
	Stop words	“特色”, “前提”, “正常” etc.
	NPs include some punctuations	(NP (NN 出口) (NN 加工) (PU 、) (NN 航运) (NN 中转) (ETC 等))

2.2 基于机器学习方法的待消解项识别

机器学习方法主要用于识别与上下文相关的待消解项。在中英文平台中，我们统一从两个方面描述当前对象

及其所处的上下文：（1）语言相关的平面特征，即从句法、词法、语义等多方面来描述当前对象自身及其所处的上下文信息；（2）结构化句法树，即在句法分析结果之上使用相关的裁剪策略形成适用于待消解项识别的句法子树，来结构化描述当前对象及其所处的上下文信息。

2.2.1 语言相关的平面特征

这部分特征从词法、句法、语义以及依存关系等多方面对当前对象及其所处的上下文环境进行了描述，具体选用的特征请参见表 3。从表 3 给出的特征集中我们可以看到，绝大多数特征已被众多研究者证明对后续的指代消解也是有效的。例如：Soon 等（2001）[1]提出了 12 个基本特征，并通过实验进行了特征贡献度分析，发现全串匹配、别名和同位语关系在判断指代关系中起了至关重要的作用；Ng 等（2007）[17]通过各类实验分析了语义类别信息对指代消解的作用，并证实准确的语义类别信息能大大提升指代消解的性能；Kong 等（2009）[18]将中心理论拓展到语义层，借助语义角色信息分析了语义角色及其驱动动词的相关信息能大大提升指代消解性能，特别是代词消解的性能。这类特征的引入，将有利于待消解项识别与指代消解的融合。

Table 3 Set of flat features

表 3 平面特征集

Category	Feature	Description
Lexical Features	Pronoun	1 if current mention is a pronoun, else 0
	DefiniteNP	1 if current mention is a definite NP, else 0
	InDefiniteNP	1 if current mention is a indefinite NP, else 0
	DemonstrativeNP	1 if current mention is a demonstrative NP, else 0
	ProperNP	1 if current mention is a proper NP, else 0
	PluralSingular	1 if current mention is singular, else 0.
	MaleFemale	1 if current mention is male, else 0.
	IsHeadWord	1 if current mention is same with its headword, else 0.
Syntactic Features	STRMATCH	1 if there is a string match between current mention and other phrases in previous context, else 0.
	NAMEAliasMatch	1 if current mention and other phrases is a name alias or abbreviation, else 0.
	Appositive	1 if current mention and other phrases are in an appositive structure, else 0.
	NestIn	1 if current mention nests another NP, else 0
	NestOut	1 if current mention is nested in another NP, else 0.
	FirstNP	1 if current mention is the first NP of this sentence, else 0.
	DistanceForFront	The distance between current mention and the nearest forward clause.
	DistanceForBack	The distance between current mention and the nearest backward clause.
Semantic Features	Arg0	1 if the semantic role of current mention is agent, else 0
	Arg0MainVerb	1 if current mention has the semantic role of agent for the main predicate of the sentence, else 0
	Args	1 if current mention has some semantic role, else 0
	WordSense	1 if current mention and other phrases agree in the semantic sense, else 0
Dependency Features	NounComp	1 if current mention has a nominal complement function role, else 0.
	AdjComp	1 if current mention has a adjective complement function role, else 0.
	InPrep	1 if current mention is indirect object preposition, else 0.
	PartOfPrep	1 if current mention is part of one prep phrase, else 0.

考虑到指代消解系统的实用性，在待消解项识别及后续的指代消解过程中，各类平面特征的获取均使用相关的国际先进的NLP工具自动获得。具体而言：我们使用错误驱动的基于HMM的方法进行英文平台的命名实体识别、词性标注和名词短语识别（Zhou和Su，2000，2002）[21][22]，中文平台使用Stanford自然语言处理组提供的分词和词性标注工具进行自动分词和词性标注 [1]，中文命名实体的识别则使用了本实验室命名实体识别小组自行开发的基于信用度模型的中文命名实体识别系统，其在微软语料上的总体性能达到了 92.3%。中英文语义角色标注

[1] http://nlp.stanford.edu/software/index.shtml

信息使用本实验室SRL小组自行开发的中英文SRL系统自动获取[23]，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。中英文依存关系使用Stanford的Dependency Parser自动获取 [2]，句法树及相关信息则使用Charniak Parser获得。

2.2.2 结构化句法树

平面特征集能在一定程度上有效的描述当前对象及其所处的上下文信息。不过在自然语言处理领域的很多研究表明，结构化信息对某些任务而言也是非常重要的，而平面特征并不能充分使用结构化信息，结构化信息在转化成平面特征时可能会丢失部分有效的信息。例如，即使非常相似的两条路径，可能会因为某一中间结点的差异就会被当成截然不同的特征，无法体现其相似性，这对分类任务是不利的，而待消解项识别问题实质上就是一个二元分类问题。鉴于此，我们在平面特征的基础上又引入了结构化句法树，即在结构化句法树上直接选取能涵盖有效描述上下文信息的子树，并根据具体任务在获取的子树上动态加入附加信息（例如实体所属的语义类别等），再利用树核函数直接计算结构化信息间的相似度。因为无须结构化信息到平面特征的转换，结构化特征往往能更准确细致的描述当前对象及其所处的上下文。

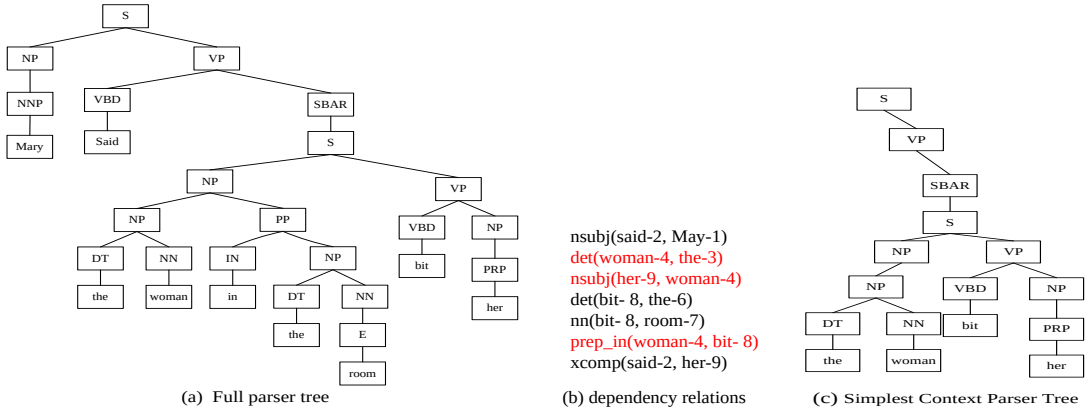
当然，结构化信息的引入存在一个平衡问题：引入的信息过多会导致产生的噪音大于其贡献度，从而造成性能的降低；引入的信息过少，信息间的相互支持不够，就不能充分发挥引入信息的作用。本文根据直接依存关系，在句法分析结果上对句法树进行裁剪，来获取待消解项识别所需的结构化信息，具体过程如下：

1） 利用 Charniak Parser 对语料进行句法分析，获得每一语句的完全句法树（FPT，Full Parser Tree）；

2） 利用 Standford Dependency Parser 获得语句中各构成单词间的依存关系；

3） 在 FPT 树上仅保留与当前对象中心词具有直接依存关系的分支，形成最简上下文句法树（SCPT，Simplest Context Parser Tree）。

得到的 SCPT 就是我们要引入的待消解项识别所需的结构化句法树。以“May said the woman in the room bit her. ”为例，假设当前对象是“the woman”，图 1 显示了待消解项识别所需的结构化句法树的提取过程。从图中我们可以看到，句子中与“the woman”的中心词“woman”具有之间依存关系的只有三个分支：“the”、“bit”和“her”，我们仅保留这三个分支，形成的最简上下文句法树如图 1(c)所示。



(a) Full parser tree　(b) dependency relations　(c) Simplest Context Parser Tree

图 1 待消解项识别所需结构化句法树的自动获取

Figure 1 Automatic generation of SCPT for anaphoricity determination

得到了最简上下文句法树后，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。本文直接使用SVMLight[3]中提供的卷积树核函数进行两个结构化对象间的相似度计算，该卷积树核函数已被应用于句法分析(Collins等 2002[24])、语义角色标注(Moschitti 2004[25])、语义关系抽取(Zhang等 2006[26])和代词指代消解（Yang等 2006[27]）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，然后计算子结构之间的匹配情况，并将子结构匹配的结果求和，以计算出大结构的相似性。Haussler[28]和 Watkins[29] 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，

[2] http://nlp.stanford.edu/software/lex-parser.shtml
[3] http://svmlight.joachims.org/

称为卷积核函数。作为卷积核函数的一个特例，Collins 等（2001）[19]提出的卷积树核函数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，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 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 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，λ (0<λ<1) 是一个衰退因子，用于在不同大小的子树间取得平衡。

3 待消解项识别的性能评测及分析

本文首先使用Charniak Parser和Standford Dependency Parser对ACE 2003 英文语料和ACE 2005 NWIRE中文语料进行了成分句法分析和依存句法分析，然后针对每一名词短语分别提取表 3 给出的平面特征以及 2.2 节给出的SCPT结构化句法树，再交由SVMLight[4]工具中附带的复合核函数进行学习，形成分类器，并分类判断，完成待消解项识别这一二元分类问题。

3.1 实验设置

为了便于与同类系统比较，本文英文平台的实验使用了 ACE 2003 语料，它由 NWIRE、NPAPER 和 BNEWS 三个子语料构成，并且每个子语料都分成了标准的训练集和测试集；中文平台的实验我们则使用了 ACE 2005 语料，它由 NWIRE 和 BNEWS 两个子语料构成，我们将这两个子语料也随机分成了训练集和测试集。中英文平台使用的各语料的训练集和测试集统计结果如表 4 所示。

Table 4 Statistics of training and testing data sets

表 4 训练集和测试集统计

Numbers	ACE 2003 English Corpus						ACE 2005 Chinese Corpus			
	NWIRE		NPAPER		BNEWS		NWIRE		BNEWS	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
Number of Documents	130	29	76	17	216	51	191	47	237	61
Number of Words	61057	14410	59627	14983	58446	14963	50460	18371	56130	15408

ACE 语料主要来自于广播新闻（NWIRE）、新闻专线（NPAPER）和网络对话（BNEWS）。虽然同属新闻领域，但是三个来源差异明显：口语化的网络对话中通常包含大量代词，而新闻专线以书面语为主，专有名词所占比例较高。因此，BNEWS 语料中口语化内容比重较大，相应的代词比例较高；NPAPER 中书面语比重较大，专有名词比例较高；而 NWIRE 语料则介于两者之间。表 5 给出了标注集中待消解项的类别分布情况及其所占比例的情况。需要指出的是，由于 ACE2005 中文语料的训练集和测试集是随机划分的，划分情况不一样，各类别所占比例会略有不同，因此我们考虑训练集和测试集的总和。我们可以看到，在 ACE2005 中文语料中，BNEWS 子语料共有代词 1052 个，约占总标注实例的 8.4%，专有名词 4080 个，约占总标注实例的 32.4%； NWIRE 子语

[4] http://svmlight.joachims.org/

料共有代词 1042 个，约占总标注实例的 7.7%，专有名词 5147 个，约占总标注实例的 38.2%。相对 NWIRE 而言，BNEWS 中代词比例较高，专有名词比例略低。在 ACE2003 英文语料中，不论训练集还是测试集，BNEWS 子语料中的代词比例都很高，分别达到了 48.0%和 43.0%，而专有名词的比重分别为 40.6%和 35.6%。NPAPER 子语料多来源于广播新闻和新闻专线，其训练集和测试集上专有名词的比例分别为 55.0%和 42.9%，代词比例仅为 28.1%和 30.4%。

Table 5 Distribution of annotated anaphoric NPs over NP categories

表 5 标注的待消解项按名词类别的分布情况

Category	ACE 2003 English Corpus						ACE 2005 Chinese Corpus			
	NWIRE		NPAPER		BNEWS		NWIRE		BNEWS	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
Proper NP	2779 (52.9%)	634 (42.8%)	3975 (55.0%)	670 (42.9%)	2083 (40.6%)	553 (35.6%)	4131 (40.6%)	1016 (30.8%)	2929 (30.6%)	1151 (38.4%)
Pronoun	1758 (33.5%)	608 (41.0%)	2028 (28.1%)	474 (30.4%)	2461 (48.0%)	668 (43.0%)	592 (5.8%)	450 (13.6%)	793 (8.3%)	259 (8.6%)
Others	713 (13.6%)	240 (16.2%)	1218 (16.9%)	416 (26.7%)	582 (11.4%)	333 (21.4%)	5449 (53.6%)	1834 (55.6%)	5864 (61.2%)	1590 (53.0%)

表 6 给出了各语料中非待消解项的分布情况及其所占比例。其中，自动识别栏给出了在自动分词、词性标注、命名实体识别、名词短语识别等预处理工作的基础上，系统识别出的名词短语集中非待消解项的分布情况；标注集栏给出的是预处理完全正确情况下，标注集上非待消解项的分布情况。从表 6 给出的统计结果可以看到：

1） 由于自动分词、词性标注、名词短语识别等预处理工作总会存在一定的错误，因此系统自动识别的名词短语要少于标注集中的名词短语。中英文平台的名词短语的识全率大致在 90-93%之间，其中英文平台的识全率略高于中文平台。

2） 中英文语料中非待消解项均占有相当比例，这说明了待消解项识别的重要性。其中，英文语料中非待消解项约占 20%，而中文语料中非待消解项所占比例远远高于英文语料，达到 50%左右。

Table 6 Distribution of non-anaphoric NPs over different corpus

表 6 非待消解项在各语料中的分布情况

Distribution		ACE 2003 English Corpus						ACE 2005 Chinese Corpus			
		NWIRE		NPAPER		BNEWS		NWIRE		BNEWS	
		Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
Annotation	All NPs	6691	1893	8749	1925	6692	2020	19266	6110	19031	5635
	Anaphors	5250 (72.5%)	1482 (78.3%)	7221 (82.5%)	1560 (80.0%)	5126 (76.6%)	1554 (76.9%)	10172 (52.8%)	3300 (54.0%)	9586 (50.4%)	3000 (53.2%)
	Non-anaphors	1441 (27.5%)	411 (21.7%)	1528 (17.5%)	365 (20.0%)	1566 (23.4%)	466 (23.1%)	9094 (47.2%)	2810 (46.0%)	9445 (49.6%)	2635 (46.8%)
Auto detection	All NPs	6194	1744	8208	1827	6252	1847	17511	5595	17301	5155
	Anaphors	4770 (77.0%)	1326 (76.0%)	6692 (81.5%)	1460 (79.9%)	4739 (75.8%)	1392 (75.4%)	9527 (54.4%)	3079 (55.0%)	9157 (52.9%)	2774 (53.8%)
	Non-anaphors	1424 (23.0%)	418 (24.0%)	1516 (18.5%)	367 (20.1%)	1513 (24.2%)	455 (24.6%)	7984 (45.6%)	2516 (45.0%)	8144 (47.1%)	2381 (46.2%)

待消解项识别是指代消解的子任务之一，为了能和后续的指代消解子任务很好的融合，我们使用了与传统指代消解平台一致的预处理系统来获得参与待消解项识别的名词短语。预处理完成后，我们将对所有的名词短语进行待消解项识别。

在机器学习环节中，本文的训练、测试均使用了 SVMLight 工具中附带的复合核函数进行。其中结构化句法树利用卷积树核函数进行相似度计算，而平面特征则利用基本的径向基核函数进行相似度计算，再利用乘积将所得的两个核进行复合。

为了能和同类系统进行比较，我们采用了两种方法评价待消解项识别的性能：

1）与 Zhou 等（2009）[15]类似，我们使用了两个准确率来评估待消解项识别器的性能，它们分别是：正例的准确率 Acc+，即正确识别的待消解项个数占应识别的待消解项个数的比例，这一准确率越高，说明被丢失的待消解项越少，指代消解在这一环节损失的召回率越低；负例的准确率 Acc-，即正确识别的非待消解项的个数占总非待消解项个数的比例，这一准确率越高，进行指代消解测试时，待消解项识别器正确滤去的不必要的测试实例越多，引入的噪音越少。

2）与 Ngai 等（2007）[16]类似，我们也使用了准确率（P）、召回率（R）和 F 值对待消解项识别器进行了评测。其中准确率、召回率和 F 值的计算公式分别为：$R=\frac{\text{正确识别的待消解项数目}}{\text{待消解项总数}}\times 100\%$；$P=\frac{\text{正确识别的待消解项数目}}{\text{识别出的待消解项数目}}\times 100\%$；

$F=\frac{2\times P\times R}{P+R}\times 100\%$

3.2 实验结果及分析

表 7 给出了机器学习方法构造的待消解项识别器的性能。从表中可以看到：

1) 规则方法能获得很高的正例准确率，但负例准确率非常低。这说明规则方法能识别出绝大多数的待消解项，但过滤非待消解项的能力极低，因此引入基于规则方法的待消解项识别器后对指代消解性能的影响不大。
2) 仅采用平面特征方法获得的待消解项识别器的正例准确率偏低，表明有些待消解项被误判成了非待消解项，这必然会影响后续的指代消解，导致系统召回率降低；
3) 与平面特征生成的待消解项识别器的性能相比，不论英文平台还是中文平台，使用结构化树核方法形成的识别器总体性能要好于使用平面特征形成的识别器性能，特别是正例准确率，每个子语料上的提升都超过 12%；负例准确率则变化较小，英文平台提升 1-2%，中文平台下降约 3%；
4) 使用复合核函数方法形成的待消解项识别器的正例准确率都低于仅使用结构化树核方法形成的识别器性能，而负例准确率有了一定提升。在英文平台，正例准确率的下降和负例准确率的上升几乎相当，而在中文平台，正例准确率略有下降，负例准确率提升明显。与仅使用平面特征形成的待消解项识别器相比，复合核函数方法形成的识别器性能要好得多。

Table 7 Performance of anaphoricity determination

表 7 待消解项识别器的性能

System		ACE 2003 English Corpus			ACE 2005 Chinese Corpus	
		NWIRE(%)	NPAPER(%)	BNEWS(%)	NWIRE(%)	BNEWS(%)
Rule-based	Acc^+	97.6	97.5	96.9	92.4	90.3
	Acc^-	17.8	18.8	13.9	16.2	19.3
Flat feature-based	Acc^+	70.3	75.4	69.7	66.8	64.8
	Acc^-	87.6	87.3	84.5	83.9	82.7
Structured feature-based	Acc^+	**84.3**	**87.8**	**89.2**	**81.5**	**78.4**
	Acc^-	88.8	88.9	86.5	80.9	80.1
Composite-based	Acc^+	80.5	84.4	80.9	79.4	77.4
	Acc^-	**91.6**	**94.8**	**92.1**	**86.7**	**83.4**

Zhou 等（2009）[15]给出了一个利用标记传播算法（Label Propagation Algorithm）进行待消解项识别的全局优化方案。他们利用 LP 算法，分别在平面特征和结构化句法树的基础上，对识别出的待消解项进行优化，并在 ACE 2003 的三个语料上进行了实验。其中

1) 基于平面特征的 LP 优化方案在三个子语料上的结果分别为：正例准确率 71.3/73.5/68.4；负例准确率 80.2/79.1/78.6。可以看到，我们的基于平面特征的方案，性能略好，特别是负例的准确率，这主要得益于特征集中我们选择的表示依存关系的特征组。
2) 在基于结构化特征的 LP 优化方案中，Zhou 等（2009）尝试了多种结构化信息的捕获方案，最终动态树（DET）取得了最佳的性能，其结果为：正例准确率 79.2/81.2/76.5；负例准确率 87.8/84.5/85.3。与之相

比，我们给出的基于依存关系的最简上下文句法树在待消解项识别中取得了更好的性能，负例准确率与Zhou 等的相当，而正例准确率要高出 5%以上，说明我们给出的结构化特征捕获方案对待消解项识别子任务而言更加有效。

Table 8 Performance Comparison of anaphoricity determination in Chinese language

表 8 中文平台的待消解项识别器性能对比

System		P(%)	R(%)	F
Ngai et al.(2007)	NWIRE	77.5	65.5	70.8
	BNEWS	73.8	64.0	68.5
Our system	NWIRE	82.1	87.9	84.9
	BNEWS	80.7	85.3	82.9

目前中文平台对待消解项识别的研究较少，仅 Ngai 等（2007）[16]给出了 ACE 2005 的两个子语料上的实验结果，他们采用了传统的准确率、召回率和 F 值的评测方法。将 ACE 2005 子语料上我们使用复合核函数方法得到的结果转换成准确率、召回率和 F 值，结果如表 8 所示。从表中可以看到，不论 NWIRE 还是 BNEWS 语料，我们给出的基于复合核函数方法所形成的待消解项识别器的性能都大大好于 Ngai 等(2007)的性能。

4 待消解项识别与指代消解

有了独立、有效的待消解项识别器后，我们尝试将它作为过滤器应用到了指代消解中。引入待消解项识别器时，指代消解平台的训练过程不变，只在生成测试实例前，对每个名词短语先交由独立形成的待消解项识别器判断其是否是待消解项。若被判断为非待消解项，该名词短语将不再参与后续的测试实例的生成及消解过程；若为待消解项，则与传统指代消解过程一致，配对形成测试实例进行指代消解。

4.1 实验结果

首先我们按 Soon 等（2001）[1]提出的指代消解基本框架构建了指代消解基准平台，平台的性能如表 9 中的第一行所示。其中就英文平台而言，我们的基准系统与目前国际先进的指代消解系统在 ACE 2003 语料中的评测结果相当，例如 Yang 等（2008）[20]给出的指代消解系统在 ACE 2003 的三个子语料上的评测结果分别为 57.5/57.3/62.3，仅在 BNEWS 语料集上略高于我们的基准平台。对于中文平台，Ngai 等（2007）[16]给出的在 ACE 2005 NWIRE 和 BNEWS 语料上的最佳 F 值性能分别为 55.3 和 55.1，而我们构建的中文基准平台在相同语料上评测得到的 F 值为 63.0/61.7，大大优于同类系统。

表 9 的第 2 至 5 行分别给出了在基准平台上引入规则方法判断待消解项、引入基于平面特征方法的待消解项识别器、基于结构化树核函数方法的待消解项识别器和基于复合核函数方法的待消解项识别器后指代消解系统的性能。

Table 9 Performance Comparison of coreference resolution in both English and Chinese languages

表 9 中英文指代消解性能比较

System	ACE 2003 English Corpus									ACE 2005 Chinese Corpus					
	NWIRE			NPAPER			BNEWS			NWIRE			BNEWS		
	R%	P%	F	R%	P%	F	R%	P%	F	R%	P%	F	R%	P%	F
Baseline(no anaphoricity)	54.1	68.4	60.4	59.7	69.0	64.0	49	66.9	56.6	70.2	57.2	63.0	67.1	57.3	61.8
+Anaphoricity determination: rule-based	54.1	70.2	61.1	59.6	72.5	65.4	48.9	70.1	57.6	70.2	60.1	64.8	64.9	61.3	63.0
+Anaphoricity determination: flat feature-based	51.0	78.6	61.9	56.5	79.0	65.9	46.9	78.4	58.7	67.1	64.3	65.7	63.7	65.3	64.5
+Anaphoricity determination: structured feature-based	52.9	82.1	64.3	57.6	83.3	68.1	47.5	83.9	60.7	69.8	64.0	66.8	64.2	66.0	65.1
+Anaphoricity determination: composite-based	52.7	84.4	**64.9**	57.3	86.7	**69.0**	47.0	87.4	**61.1**	69.6	65.5	**67.5**	64.0	67.1	**65.5**

从实验结果我们可以看到：

1) 不论中文还是英文平台，待消解项识别模块的引入在一定程度上都提升了指代消解的总体性能。待消解项识别模块能去除非待消解项带来的噪音，使得指代消解的准确率得到了较大提升；同时，由于部分待消解项被误认为非待消解项滤去，也造成了指代消解召回率的下降。

2) 基于规则方法的待消解项识别模块主要用于过滤与上下文无关或语言固定模式非常显著的非待消解项。由于本文使用的规则集极其有限，因此对指代消解召回率的影响很小，对指代消解准确率的贡献也相对有限。
3) 由于基于平面特征方法的待消解项识别器的正例准确率偏低，造成了较多的待消解项被误判成非待消解项，使得指代消解召回率下降，虽然准确率有了一定的提升，但指代消解的整体性能提升并不显著。
4) 由于基于结构化树核函数方法的待消解项识别器取得了较好的正例、负例准确率，在指代消解中引入该识别器后，系统召回率虽略有下降，但准确率大大提升，使得中英文平台的指代消解总体性能都有了显著提升。
5) 将平面特征和结构化句法树复合形成的识别器在负例准确率方面提升显著，而正例准确率略有下降，将其引入指代消解平台后取得了更好的系统准确率，与单纯的结构化方法相比，系统召回率虽略有下降，但指代消解系统的总体性能得到了进一步提升，在中英文平台都取得了最佳的指代消解性能。

4.2 性能分析

基于规则方法的待消解项识别主要针对某些语言相关的特性进行，对指代消解性能的提升具有一定的作用，但因为它的覆盖率较低，对指代消解性能的影响也较为有限。

指代消解系统加入基于平面特征方法的待消解项识别模块后，性能虽有一定提升，但不显著。我们将基于平面特征方法的待消解项识别模块针对各种类型的名词短语依次渐进的进行非待消解项的过滤，其结果如表 10 所示。从表 10 中可以看到，仅对代词进行非待消解项的过滤，系统性能有 0.4%-0.7%的提升。同时对代词和专有名词进行非待消解项的过滤，系统有 0.9%-1.7%的提升。这说明基于平面特征方法的待消解项识别模块对代词和专有名词均有一定的效果。但对其他类型的名词短语也进行非待消解项过滤后，系统性能却都有了一定程度的下降，说明基于平面特征方法的待消解项识别模块对其他类别的名词短语存在较多的误判，滤去了较多的待消解项，造成了系统性能的整体下降。

Table 10 Performance contribution on different NP categories using the flat feature-based anaphoricity determination model

表 10 基于平面特征方法的待消解项识别对不同类别的名词短语消解的性能贡献度

System	ACE 2003 English Corpus									ACE 2005 Chinese Corpus					
	NWIRE			NPAPER			BNEWS			NWIRE			BNEWS		
	R%	P%	F	R%	P%	F	R%	P%	F	R%	P%	F	R%	P%	F
Baseline(with rule-based anaphoricity determination)	54.1	70.2	61.1	59.6	72.5	65.4	48.9	70.1	57.6	70.2	60.1	64.8	64.9	61.3	63.0
+anaphoric pronouns determination	53.6	72.9	61.8	58.6	74.7	65.7	48.1	73.1	58.0	68.7	62.4	65.4	64.3	63	63.6
+anaphoric pronouns and proper NPs determination	52.7	75.8	62.2	58.1	77.2	66.3	47.8	78.1	59.3	68.2	64.1	66.1	64.2	65.1	64.6
+anaphoric noun phrases determination	51	78.6	61.9	56.5	79	65.9	46.9	78.4	58.7	67.1	64.3	65.7	63.7	65.3	64.5

引入基于结构化树核函数的待消解项识别模块后，中英文平台指代消解的性能提升了 2%-3.2%。进一步分类别实验，我们得到了表 11 所示的基于结构化树核函数的待消解项识别模块对不同名词类别进行消解的贡献度。从表 11 所示的结果我们可以看到，基于树核的待消解项识别模块对指代消解性能的贡献主要体现在对代词消解性能的提升上。对专有名词的消解，该模块起到了反作用，系统性能均有一定程度的下降，其主要原因是基准系统对专有名词的消解性能已经很好，在各语料上均达到了 80%以上的 F 值，而且已经有许多研究表明对专有名词消解贡献度最大的特征就是全串匹配，这一特征很难以树核方式体现。待消解项识别模块对其他类别的名词短语消解的贡献度虽不显著，但在中英文平台都一致的提升了指代消解的性能。

Table 11 Performance contribution on different NP categories using the structured tree kernel-based anaphoricity determination model

表 11 基于结构化树核函数的待消解项识别对不同类别的名词短语消解的性能贡献度

System	ACE 2003 English Corpus									ACE 2005 Chinese Corpus					
	NWIRE			NPAPER			BNEWS			NWIRE			BNEWS		
	R%	P%	F	R%	P%	F	R%	P%	F	R%	P%	F	R%	P%	F
Baseline(with rule-based anaphoricity determination)	54.1	70.2	61.1	59.6	72.5	65.4	48.9	70.1	57.6	70.2	60.1	64.8	64.9	61.3	63.0
+anaphoric pronouns determination	53.8	79.6	64.2	59.1	81.3	68.4	48.3	82.4	60.9	70	63.6	66.6	64.5	65.4	64.9
+anaphoric pronouns and proper NPs determination	53.1	80.1	63.9	57.9	82.1	67.9	47.7	82.6	60.5	69.5	63.8	66.5	64.2	65.4	64.8
+anaphoric noun phrases determination	52.9	82.1	64.3	57.6	83.3	68.1	47.5	83.9	60.7	69.8	64	66.8	64.2	66	65.1

复合核函数方法在一定程度上综合了平面特征方法和结构化树核方法，因此在指代消解系统中引入基于复合核函数的待消解项识别模块后，系统性能较核函数方法又有了略微的提升，主要来自专有名词消解性能的提升。

5 小结与展望

本文全面深入研究了中英文指代消解中的待消解项识别问题。文章首先利用规则方法识别与上下文无关或具有固定模式的非待消解项，然后又从平面特征和结构化句法树两方面出发，利用机器学习的方法构造二元分类器来识别上下文相关的待消解项，最后借助复合核函数将平面特征和结构化句法树有效结合，系统探讨了构建待消解项识别器的多种方法，并通过中英文语料上的实验说明了各种方法的有效性。最后，文章还将生成的待消解项识别模块应用于中英文指代消解，并通过实验说明了有效的待消解项识别能大大提升指代消解的性能。

虽然本文对于待消解项识别的研究取得了一定的成效，但有些问题还有待进一步深入研究，例如：在指代消解中应如何更有效的应用待消解项识别信息；除本文提到的平面特征和结构化句法树外，还有哪些信息对待消解项识别意义重大等。

References

[1] Soon W.M., Ng H.T. and Lim D. 2001. A machine learning approach to coreference resolution of noun phrase. Computational Linguistics, 27(4):521-544.

[2] Ng V. and Cardie C. 2002. Improving machine learning approaches to coreference resolution. ACL’2002: 104-111

[3] Qian W., Guo Y.K., Zhou Y.Q., and Wu L.D. 2003. English Noun Phrase Coreference Resolution via a Maximum Entropy Model. Journal of Computer Research and Development, 40(9): 1337-1343

[4] Yang X.F., Su J., Zhou G.D. and Tan C.L. 2004. Improving pronoun resolution by incorporating coreferential information of candidates [A]. ACL’2004: 127- 134

[5] Bergsma S. and Lin D.K. 2006. Bootstrapping path-based pronoun resolution. COLING-ACL’2006: 33-40.

[6] Li G.C. and Luo Y.F. 2005. Chinese Pronominal Anaphora Resolution Via a Preference Selection Approach. Journal of Chinese Information Processing. 4:24-30(in Chinese with English abstract)

[7] Song W., Qing B., Lang J. and Liu T. 2008. Combining Syntax and Word Sense for Chinese Pronoun Resolution. Journal of Chinese Information Processing. 6:8-13(in Chinese with English abstract)

[8] Stoyanov V., Gilbert N., Cardie C. and Riloff E. 2009. Conundrums in Noun Phrase Coreference Resolution: Making Sense of the State-of-the-Art. ACL’2009: 656-664

[9] Bergsma S., Lin D. and Goebel R. 2008. Distributional Identification of Non-referential Pronouns. ACL’2008:10-18.

[10] Bean D. and Riloff E. 1999. Corpus-based Identification of Non-Anaphoric Noun Phrases. ACL’ 1999: 373-380

[11] Lappin S.and Herbert J.L. 1994. An algorithm for pronominal anaphora resolution [J]. Computational Linguistics, 20(4): 535 - 561.

[12] Ng V. and Cardie C. 2002. Identify Anaphoric and Non-Anaphoric Noun Phrases to Improve Coreference Resolution. COLING’2002

[13] Ng V. 2004. Learning Noun Phrase Anaphoricity to Improve Coreference Resolution: Issues in Representation and Optimization. ACL’2004:151-158

[14] Yang X.F.,Su J. and Tan C.L. 2005. Improving Pronoun Resolution Using Statistics-Based Semantic Compatibility Information. ACL’2005:165-172.

[15] Zhou G.D. and Kong F. Global Learning of Noun Phrase Anaphoricity in Coreference Resolution via Label Propagetion. EMNLP’2009: 978-986

[16] Ngai G. and Wang C.S. 2007. A Knowledge-based Approach for Unsupervised Chinese Coreference Resolution. Computational Linguistics and Chinese Language Processing. 12(4):459-484

[17] Ng. V. 2007. Semantic Class Induction and Coreference Resolution. ACL’2007 536-543.

[18] Kong F., Zhou G.D. and Zhu Q.M. Employing the Centering Theory in Pronoun Resolution from the Semantic Perspective. EMNLP’2009: 987-996

[19] Collins M. and Duffy N. 2001. Covolution kernels for natural language [C]. NIPS’2001:625-632.

[20] Yang X.F., Su J. and Tan C.L. 2008. A Twin-Candidate Model for Learning-Based Anaphora Resolution. Computational Linguistics 34(3):327-356

[21] Zhou G.D. and Su J. 2000. Error-driven HMM-based Chunk Tagger with Context-dependent Lexicon , EMNLP-VLC-2000:71-79.

[22] Zhou G.D., and Su J. 2002. Named Entity Recognition Using a HMM-based Chunk Tagger. ACL-2002:473-480.

[23] Li J.H., Zhou G.D., Zhao H., Zhu Q.M. and Qian P.D. Improving nominal SRL in Chinese language with verbal SRL information and automatic

predicate recognition. EMNLP '2009

[24] Collins M. and Duffy, N. 2002. New Ranking Algorithm for Parsing and Tagging: Kernel over Discrete Structure, and the Voted Perceptron. ACL'2002: 28-136.

[25] Moschitti A. 2004. A study on convolution kernels for shallow semantic parsing. ACL'2004:335-342.

[26] Zhang Z. 2004. Weakly supervised relation classification for information extraction. CIKM'2004. 8-13

[27] Yang X.F., Su J. and Tan C.L. 2006. Kernel-based pronoun resolution with structured syntactic knowledge. COLING-ACL'2006: 41-48.

[28] D. Haussler. Convolution Kernels on Discrete Structures. Tech. Rep. UCSCCRL-99-10, 1999.

[29] C. Watkins. Dynamic Alignment Kernels. Tech. Rep. CSD-TR-98-11, 1999.

附中文参考文献:

[3] 钱伟，郭以昆，周雅倩，吴立德. 2003. 基于最大熵模型的英文名词短语指代消解[J]. 计算机研究与发展. 40(9): 1337-1343.

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

[7] 宋巍，秦兵，郎君，刘挺. 2008. 句法与词义相结合的中文代词消解[J]. 中文信息学报 6:8-13

Kong Fang born in 1977. She received her Ph.D. degree from Soochow University, Suzhou, China, in 2009. Currently, she is an associate professor at the university. Her research interests include natural language processing and information extraction. She is a member of China Computer Federation (CCF).

Zhu Qiaoming born in 1963. He received his Ph.D. degree from Soochow University, Suzhou, China, in 2008. Currently, he is a professor at the university and acts as the deputy director of Department of Science, Technology and Industry. His research interests include natural language processing, information extraction and embedded systems. He is a senior member of China Computer Federation (CCF).

Zhou GuoDong born in 1967. He received the Ph.D. degree from the National University of Singapore in 1999. He joined the Institute for Infocomm Research, Singapore, in 1999, and had been associate scientist, scientist and associate lead scientist at the institute until August 2006. Currently, he is a professor at the School of Computer Science and Technology, Soochow University, Suzhou, China. His research interests include natural language processing, information extraction and machine learning. He is a senior member of China Computer Federation (CCF) and has been the member of ACM and IEEE since 1999.

文章所属类别：人工智能

文章校对负责：孔芳，手机：13862160247；Email：kongfang@suda.edu.cn