高技术通讯 2009 年 第 19 卷 第 9 期:926 ~ 932

doi:10.3772/j.issn.1002-0470.2009.09.009

基于 ART 网络的无指导中文共指消解方法①

李世奇② 赵铁军 陈 晨 刘鹏远

(哈尔滨工业大学教育部-微软语言语音重点实验室 哈尔滨 150001)

摘 要 提出了一种基于自适应谐振理论(ART)网络的无指导中文名词短语共指消解方法。该方法充分利用名词短语自身特征,通过改变网络参数动态调节聚类数量,有效地解决了目前聚类共指消解中输出类别数目难以确定的难题。另外采用了一种基于信息增益率的特征选择方法,减少了区分度较弱特征给聚类所带来的干扰。该方法在保证了识别正确率的前提下,不依赖人工标注语料,可直接应用于跨领域的真实文本。最后在 ACE 中文语料上进行了相关实验,并取得了较好的结果。

关键词 共指消解,无指导学习,自适应谐振理论(ART),自然语言处理

0 引 言

随着自然语言处理技术发展的日趋深入,共指消解(coreference resolution)作为一种通用性前端理解技术在自然语言处理领域中扮演着越来越重要的角色,广泛应用于信息抽取、自动文摘、问答系统等诸多领域中,共指消解研究对于信息处理技术的发展进步有着重大意义。共指消解是指代消解的一个特例。指代是指篇章中的两个语言单位之间存在的特殊语义关联,它是自然语言中的一种普遍现象,它能够简洁高效地表述语义和衔接上下文。用于指向和被指向的语言单位分别称为照应语和先行语,确定照应语所指的先行语的过程就是指代消解。当照应语和先行语同指现实世界中的同一实体时,指代消解称为共指消解[1]。共指消解是信息处理领域中的一个难题,针对这一难题,本文引入了神经网络方法,提出了一种基于自适应谐振理论(adaptive resonance theory,ART)网络的无指导中文名词短语共指消解方法,该方法有效地克服了传统的聚类共指消解方法存在的聚类数目难以确定的难题。

1 相关研究

共指消解是一个具有挑战性的研究课题,国外基于有指导的机器学习方法的共指消解研究相对较为深入[2-6];汉语方面,北京大学王厚峰等从计算语言学角度对共指消解进行了探索性研究[7,8]。基于无指导的共指消解还属于相对较新的研究,从已知文献来看,仅有下面几项较为完整的工作。1999 年,Cardie 最早提出无指导的名词短语共指消解方法[9],采用了短语本身、短语中心词、短语序号、代词类型、冠词类型、同位语、专有名词、单复数、语义类别等 9 个特征和基于启发式规则的聚类方法;2003 年 Bergler 提出基于模糊集理论的共指消解方法,采用语义距离、短语重合程度、缩写词、简单代词消解和短语中心词等 5 条模糊规则处理,然后经过模糊集合并和解模糊化过程形成共指链,但结果较差[10]。2004 年 Bean 和 Riloff 利用信息抽取模板方法获取上下文信息,然后根据这些信息判断指代语与先行语的相容性[11]。在汉语方面,2006 年香港理工大学的 Wang 和 Ngai 提出基于改进的 K 均值聚类算法的中文无指导共指消解[12],该方法选定 12 个适于中文的聚类特征,采用 Cardie 的距离度量,在人工标注的 30 篇语料上获得了较好效果;2007 年南京大学周俊生等采用类似 Cardie 的特征空间和距离度量,提出一种基于图划分的无指导共指消解方法[13]。

在上述聚类方法中,存在的最主要问题是聚类结束条件难以有效判断。由于在共指消解问题中篇章内的实体总数无法从原文中获得且难以估计,因此聚类数目无法预知,且该参数对最终聚类的质量具有重要影响[14]。目前的方法大都以从大量实验

① 国家自然科学基金(60575041)和 863 计划(2006AA01Z150)资助项目。
② 男,1984 年生,博士生;研究方向:自然语言处理;联系人,E-mail: sqli@mtlab.hit.edu.cn
(收稿日期:2008-07-21)

数据中获得的经验收敛阈值作为判断结束的依据,而未考虑任务本身的特性,影响了聚类共指消解的性能。另外现有研究中聚类特征都是直接从有指导共指消解特征中手工选取,并未考虑区分度较弱特征对聚类效果的干扰。

本文首次将神经网络方法引入共指消解问题中,提出了基于自适应谐振理论(ART)网络的无指导名词短语共指消解方法。该方法隐式地利用名词短语特征进行聚类,有效地克服了名词短语聚类过程中输出类别数目难以确定这一主要问题。它能够通过实验来调节网络参数动态控制聚类算法的输出类别数,同时解决了距离度量难以定义和样本点波动大的问题。另外通过一种基于信息增益率的特征选择方法过滤掉区分度较弱特征,以减少其对聚类算法的干扰。整体上该方法不依赖于人工标注语料库,可应用于跨领域的真实文本中,具有高效性、鲁棒性和可移植性。

2 基于信息增益率的特征选择

特征向量构造是共指消解中的重要环节,特别是对于聚类方法,特征的选择是聚类效果的决定因素之一。传统分类方法通常能够根据训练语料中的分布情况,对于区分度弱的特征,给予其较小权重。但对于聚类方法,这些特征可能会带来较大噪声,因此本文采用基于信息增益率特征选择方法过滤掉区分度较弱的特征。信息增益是信息论中的重要概念,它能够有效衡量给定特征区分训练样本的能力,在 ID3 算法中增长树的每一步均使用信息增益作为分类特征选择标准[15]。其原理可概括为,某特征的信息增益就是数据集在被这个特征分割前后期望熵的差值。在标记数据集 D 上特征 A 的信息增益 $Gain(D,A)$ 表示为

$$Gain(D,A) = E(D) - \sum_{v\in Values(A)} \frac{|D_v|}{|D|} E(D_v)$$

其中 $Values(A)$ 代表特征 A 的所有可能值集合;D_v 代表 D 中特征 A 的值为 v 的子集,也就是 $D_v = \{d \in D \mid A(d) = v\}$;$E(D)$ 表示数据集 D 的熵,如果 D 中存在 c 个类别,那么

$$E(D) = -\sum_{i=1}^{c} p_i \log_2(p_i)$$

一般认为信息增益值越大的特征就越好,但该方法存在一种倾向性偏差,原因在于取值多的特征的信息增益值往往较大。因此本文并不直接采用信息增益,而采用信息增益率来作为特征的选择标准,Quinlan 在文献[16]中提出信息增益率标准可消除这种影响:

$$GainRatio(D,A) = \frac{Gain(D,A)}{SplitInfor(D,A)}$$

$$SplitInfor(D,A) = -\sum_{v\in Values(A)} \frac{|D_v|}{|D|} \log_2\left(\frac{|D_v|}{|D|}\right)$$

本文总结了现有聚类共指消解文献[9]、[12]和[13]中所采用的特征构成候选特征列表(表 1)。根据上述方法,采用自动内容抽取(automatic content extraction,ACE)标准语料库中的广播新闻(broadcast news,BN)部分(详见 4.1 节)作为特征选择数据集,以篇章为单位,分别计算 11 个候选特征的信息增益率,其中少数 $SplitInfor(D,A)$ 值为零情况的实例将被忽略,这种情况表示该篇章中所有名词短语在特征 A 上的取值都相同,最后计算信息增益率均值。本文忽略信息增益率不足 50% 的特征,即专有名词、中心词词频和短语序号,选择其余 8 个特征——性别、单复数、短语本身、中心词、中心词词性、短语句序号、生物性和语义类别作为聚类特征。

表 1 候选聚类特征及其描述

特征类别	特征名称	特征概述	信息增益率
词汇级	性别	包括:男性、女性、未知	0.656079
	单复数	包括:单数、复数、未知	0.580073
	专有名词	该短语是否为专有名词,包括缩写词	0.456182
短语级	短语本身	名词短语本身词串	0.864717
	中心词	短语的中心词词串	0.830143
	中心词词性	短语的中心词词性:名词或代词	0.501321
	中心词词频	中心词在篇章中出现的频次	0.424709
上下文	短语序号	短语按照出现先后顺序序号	0.396943
	短语句序号	短语所在句子在文章中的序号	0.533126
语义级	语义类别	按照 ACE 实体类别标准划分	0.716820
	生物性	短语所指实体是否具有生物性	0.775435

对于短语词串、中心词串、中心词词性、性别、单复数、短语句序号、生物性这些特征值的获取，可采用词典结合规则的方法直接从原语料中提取[9,12]。语义类别特征较为复杂，本文利用一个基于线性核函数的支持向量机(SVM)分类器对名词短语的语义类进行识别。根据 ACE 实体标注规范[18]，将实体划分成 7 个语义类：PER(人)、ORG(组织机构)、GPE(地理、政治实体)、LOC(处所)、FAC(人造建筑)、WEA(武器)、VEH(传输设备)。然后，采用“一对一”方法(Pairwise Classification)[19]，将多值分类转换为多个二值分类的组合来处理，分类器选择了 4 个分类：短语本身，上下文词(窗口大小取 1)，中心词及其在 HowNet 中的概念定义[20]。

下面给出从文本中获得的特征向量形式。以句子“台北桃园中正机场昨天晚上发生空难的意外”中名词短语“台北桃园中正机场”为例，其特征向量是{“台北桃园中正机场’，‘中正机场’，‘*Noun*’，‘*NA*’，‘*S*’，‘2’，‘*NA*’，‘*FAC*’}，各特征顺序按照上段首句中所述。最后将上述特征向量进行数值化后即可作为聚类共指消解算法的输入。

3 基于 ART 网络的共指消解算法

针对聚类共指消解时面临输出类别数量难以确定的主要问题，本文提出了基于 ART 网络模型的共指消解聚类方法，该方法能够有效地解决这一问题，因为它充分地利用了名词短语自身特征，能够通过实验来调节网络参数动态控制聚类算法的输出类别数目。同时有效解决了特征空间维数较高、样本点波动大等问题。ART 网络是一种基于自适应谐振理论，具有自组织、自稳定能力的竞争学习式人工神经网络，它包括 ART1、ART2 和 ART3 三个基本模型。模型 1 主要针对处理离散的输入模式[21]，模型 2 在其基础上扩展到连续的输入模式，模型 3 模拟了化学神经键中的可计算属性，结构复杂，目前应用很少。

本文中所处理的名词短语特征值均为离散型，因此采用 ART1 模型，结构如图 1 所示。由 Attentional 和 Orienting 两个子系统构成，前者是系统的核心部分，主要处理对于系统“熟悉”的输入模式，包含三部分：基于短期存储器(short-term memory，STM)的比较层神经元和识别层神经元；基于长期存储器(long-term memory，LTM)的两层神经元之间的连接突触网络；增益控制器。面对陌生的事件发生时，后者将发挥作用，它根据一个新的输入模式能否被识别层神经元所表达做出判断，控制重置信号的激发[13]。

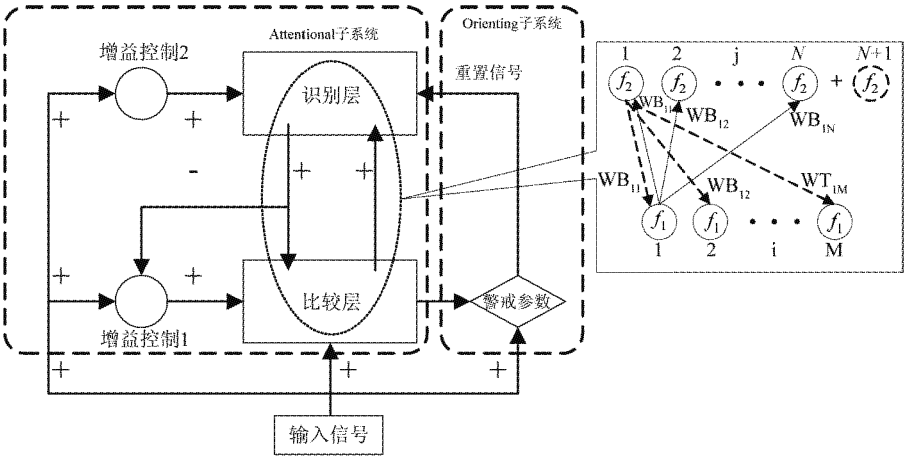

图 1 ART1 结构框架图

本文将该框架引入到聚类共指消解问题中，将名词短语特征向量作为输入信号，用识别层神经元来表示名词短语的类别。根据自适应谐振理论，通过不断对网络权重的学习更新，当网络产生谐振时，识别层中最活跃神经元所表示的类别最有可能代表该名词短语的真实类别。如果网络未产生谐振，说明识别层神经元对于该输入模式不敏感，则为该模式建立其对应的新神经元节点。下面详述名词短语聚类算法流程：

(1) 首先初始化参数，包括：警戒参数 V、失效神经元集合、识别层节点和网络突触权向量。警戒参数初始值通过实验调节获得，详见 4.3 节。失效神经元集合表示在样本的一次学习过程中已经判定失效的神经元结合，这些神经元将不再对该样本起

作用。识别层神经元集合 R 以及自顶向下和自底向上突触权向量 WT 和 WB 表示为

$$V=0.7,\ Invalid=\varnothing,\ R_i=0,$$

$$WT_{ij}=1,\ WB_{ij}=\alpha_j$$

其中,$0<\alpha_j<(\beta+|X|)^{-1}$,$\beta>0$。

(2) 将名词短语特征向量 $X:(x_1,x_2,\cdots,x_M)$ 作为输入,经比较层转换函数 $f_1(x)$ 作用后变换得到 $C:(c_1,c_2,\cdots,c_M)$,其中 G_1 为增益控制变量;$P:(p_1,p_2,\cdots,p_M)$ 为识别层的反馈信号,初值为零向量。其中

$$f_1(x_i)=\left[\frac{x_i+G_1+p_i}{2}\right];\ i\in[1,M]$$

$$G_1=(x_1\vee x_2\vee\cdots\vee x_M)\wedge(\overline{R_1\vee R_2\vee\cdots\vee R_M})$$

(3) 通过计算净激活,即转换后特征向量 C 与自底向上权向量 WB 的内积,获取在识别层中最活跃的有效神经元节点,该节点最有可能代表该样本所属类别,方法如下:

$$Net_j=\sum_{i=1}^{M}WB_{ij}\cdot c_i;\ j\in[1,N]$$

$$Active=\arg\max_j Net_j;\ R_j\in\overline{\text{Invalid}}$$

(4) 然后将此神经元反馈回比较层,再次经过转换函数 $f_1(x)$ 变换后得 $X':(x'_1,x'_2,\cdots,x'_M)$,通过计算 X' 与 X 的相似度判断是否产生谐振,其中:

$$Sim=\sum_{i=1}^{M}\lambda'_i x'_i/\sum_{i=1}^{M}\lambda_i x_i$$

(5) 若 $Sim<V$,则置 $R_{\text{Active}}=0$,Invalid = Invalid$\cup\{R_{\text{Active}}\}$,返回上一步继续在识别层中寻找有效神经元。若 $Sim\geqslant V$,则 $X\rightarrow$Category(R_{Active}),即判定样本 X 属于 R_{Active} 所表示的类别,然后更新与神经元 R_{Active} 相关的突触权值:

$$WT_{\text{Active}i}=x_i\wedge WT_{\text{Active}i},$$

$$WB_{\text{Active}i}=\frac{|x_i\wedge WT_{\text{Active}i}|}{\beta+|WT_{\text{Active}i}|};\ i\in[1,M]$$

(6) 若所有识别层神经元均无法满足 $Sim\geqslant V$ 条件,则为 X 创建一个新的实体类别及其对应的神经元节点,并初始化所有与之相连的权值。待篇章中所有样本聚类完成后,得到 $C_1,C_2,\cdots,C_k$ 共 k 个类别,把具有相同字符串元素的类别进行归并,至类别数目稳定为止,归并后的类别即是对该篇章中实体的共指划分。归并策略规则表示如下:

If $\exists e_1\in C_a,\ \exists e_2\in C_b,\ a,b\in[1,k]$

$Prototype(e_1)=Prototype(e_2)$;

Then $Combine(C_a,C_b),\ k\leftarrow k-1$

该方法避免了其他共指消解聚类方法中采用人工设定收敛阈值的方法控制聚类过程结束,而是通过实验来调节警戒参数 V,对最终聚类数量进行控制。而且它是一种高速、高效神经网络算法,能够根据输入模式快速自动学习,网络参数的更新与类别的划分同步完成。还有该方法对于未知样本能够较好识别并为其动态创建新的类别,符合共指消解聚类问题的本质特性。另外该算法具有很好的稳定性,对于样本的波动并不敏感,共指消解中样本分布不规律性不会对聚类结果造成太大影响。

4 实验及结果分析

4.1 语料

目前中文共指消解方面的语料库资源较少,ACE 评测语料是当前本领域内公认的标准语料库,该语料由美国语言数据联盟(linguistic data consortium, LDC)创建,主要包括英语、汉语和阿拉伯语三种语言,其规模及一致性都能够满足一般实验的需求。因此本文选用 ACE 2005 汉语部分共 633 篇作为数据集,其中内容大部分为新闻题材,主要来源于广播新闻(BN)、新闻专线(newswire, NW)和网络日志(Weblog, WB),详细统计信息见表 2。

表 2 ACE2005 汉语语料库实体相关统计信息

语料库	提及	代词	实体	文档
广播新闻	13501	1195	6248	298
新闻专线	14341	1173	6552	238
网络日志	6479	978	2614	97
全部	34321	3346	15414	633

由于目前中文分词、词性标记、句法分析等基本技术的准确率仍不够理想,为避免这些对于共指消解实验的影响,本文采用一种常用方法,在名词短语抽取层级上进行实验[13]。直接采用 ACE 中 mention 级标注结果做候选名词短语集合,在此基础上自动抽取特征向量。

4.2 评测指标和 Baseline

采用共指消解领域应用最广泛的评测指标,MUC-6 中定义的基于链接的精确率(precision, P)、召回率(recall, R)以及 F 测度(F-measure, F),详见文献[22]。计算公式如下:

$$R=\frac{\sum(|S_i|-|p(S_i)|)}{\sum(|S_i|-1)}$$

$$P=\frac{\sum(|S'_i|-|p(S'_i)|)}{\sum(|S'_i|-1)}$$

$$F = \frac{(1+\beta^2)\times P\times R}{\beta^2\times P+R}$$（本文 β 取 1）

其中 S 代表参考答案，S' 代表系统输出，S_i 表示 S 中的第 i 个共指集合，$p(S_i)$ 表示集合 S_i 根据 S' 中集合分布形成的划分；S'_i 和 $p(S'_i)$ 同理。

本文在相同特征空间基础上，选择两种方法作为 Baseline：

(1) 启发式规则聚类：即文献[9]中采用的方法。先将篇章中的每个名词短语单独归为一类，按照逆序逐个扫描名词短语，计算它与前面名词短语之间的距离，当该距离小于某阈值时，将这两个名词短语所属类别合并，扫描至文中首个名词短语时算法结束。

(2) 基于划分的 K 均值聚类(K-means)：即文献[12]中采用的方法。首先将篇章中的所有名词短语归为一类然后逐渐划分，过程如下：先计算所有聚类的中心，然后选出类内距离最大的一类，将该聚类中心删除，添加该类中距离最远的两个实例作为新的聚类中心。直至所有聚类的类内距离均小于设定阈值时结束。

4.3 结果及分析

如表 3 所示，在 ACE 2005 语料库 BN、NW 和 WB 三部分上，Baseline 中基于启发式规则和 K 均值聚类方法分别取得 55.2% 和 63.3% 的平均 F 值。启发式规则方法适合于表层特征明显的名词短语，在获得较高的精确率同时也会形成许多零散类别，导致召回率偏低；K 均值聚类方法采用了有效的名词短语距离度量和聚类中心分裂方法，因此性能上有所提升，但最终聚类算法终止的判断影响了整体性能。本文中的基于 ART 网络的聚类算法有效克服了这些问题，从表 3 中可以计算出在三部分语料库和完全相同特征空间条件下获得了 70.2% 的平均 F 值，明显优于前两种方法，证明了该方法的效性。

表 3 ACE 语料上的聚类共指消解实验结果

		P	R	F
BN	Heuristic	**0.823**	0.415	0.552
	K-means	0.735	0.546	0.627
	ART	0.815	**0.621**	**0.705**
NW	Heuristic	0.791	0.430	0.558
	K-means	0.748	0.549	0.633
	ART	**0.793**	**0.617**	**0.694**
WL	Heuristic	**0.812**	0.406	0.541
	K-means	0.739	0.573	0.645
	ART	0.810	**0.636**	**0.712**

警戒参数作为网络的重要参数，对于聚类效果影响很大，我们采用 0.05 为步长调整警戒参数，观察该参数改变对于整个系统识别率的影响，从图 2 中可见当警戒参数取值为 0.70 时达到最优平均 F 值，因此将其作为该网络参数的初始值进行聚类。

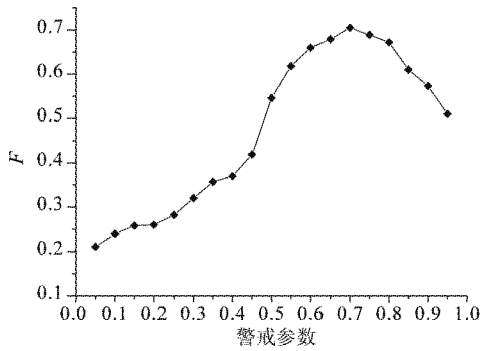

图 2 共指消解 F 值随警戒参数变化示意图

经过错误分析发现，本文方法能够较好地识别对一般名词性短语之间的共指关系，而对于代词性短语与一般名词短语的聚类效果较差，分析其原因主要在于指代性短语自身的信息缺失。比如在句子“普京没有说谁是贝加尔金融集团的老板，但他表示，该集团的背后是一群拥有石油行业经验的个人”中，名词短语‘他’与‘普京’之间存在共指关系。但是在对其聚类时三个短语级特征将失效，在实验中‘他’被分成一个单独类别。而类似的指代性短语比重占名词短语总数的 9.75%。表 4 给出了忽略由指示代词组成的短语情况下的实验结果。可见，在处理一般名词短语时效果提升明显，而对于这部分代词性短语可采用后处理规则的方法单独对其进行处理，提高其识别精度。

表 4 忽略代词性短语后的对比实验结果

		P	R	F
原结果	BN	0.815	0.621	0.705
	NW	0.793	0.617	0.694
	WL	0.810	0.636	0.712
无代词	BN	0.839	0.620	0.713
	NW	0.832	0.635	0.720
	WL	0.847	0.671	0.749

另外基于链接的准确率和召回率评测指标本身也存在一种偏差，它会忽视独自组成一个实体类别的名词短语，因此这会造成一种倾向于产生类别较少系统的不公正现象[23]，上述因素都对本方法的 F

值产生一定影响。最后为了测试聚类特征的有效性和每个特征对整体的贡献程度,采用了特征逐个移除的方法,通过观察缺少某特征时系统 F 值的下降幅度来评价该特征的区分度,实现结果见表 5。

表 5　移除各个特征对系统整体性能的影响

移除特征	P	R	F	F 值降幅 ×100
性别	0.779	0.615	0.687	1.5
单复数	0.783	0.607	0.684	1.8
短语本身	0.742	0.593	0.659	4.3
中心词串	0.736	0.602	0.662	4.0
中心词词性	0.807	0.610	0.695	0.7
短语句序号	0.794	0.611	0.691	1.1
语义类别	0.787	0.582	0.669	3.3
生物性	0.785	0.596	0.678	2.4
无(表 1 中全部特征)	0.696	0.581	0.633	6.9
无(全部特征)	0.804	0.623	0.702	-

从表 5 中可见,单个特征中 F 值下降最明显的,也就是对共指消解问题贡献度最大的三个特征是名词短语本身、中心词串及其语义类别,特征区分度与第 2 节中所述的信息增益率方法所得结果相符。另外,不采用特征选择方法时 F 值下降最大,说明了本文中基于信息增益率的特征选择方法的有效性。

5　结　论

共指消解是信息处理领域中的一个基础性难题,本文针对聚类共指消解时面临的输出类别数目未知这一主要问题,提出基于 ART 网络的中文共指消解聚类方法。从已有文献来看,本文首次将神经网络聚类方法应用于共指消解中,并较好地解决了上述问题。该方法充分地利用了名词短语自身特征,能够通过实验来调节网络参数动态控制聚类算法的输出类别数目。在聚类特征选择方面,利用信息增益率的指标,从语法、句法、上下文、语义 4 个层面选取了较适于聚类算法的 8 个特征,其中采用了基于 SVM 分类器的语义类别特征抽取模块,能够高效地对语义类这一重要特征进行识别,最后在 ACE 标准语料库上的实验结果证明了该方法较其他现有聚类方法性能上的优越性。

根据本文研究结果,今后的工作将主要集中在以下三个方面:首先,建立针对名词性短语和代词性短语之间的共指消解问题的处理模块,结合相关语言学规则以及机器学习方法,提升这部分的识别正确率;其次是新的高效聚类特征的发掘,上下文信息的利用;还有聚类模型的优化以及新方法的探索。相信通过这些工作会使共指消解的效果得到进一步提升。

参考文献

[1] 王厚峰. 汉语篇章的指代消解浅论. 语言文字应用, 2004, 4: 113-119

[2] McCarthy J F, Lehnert W G. Using decision trees for coreference resolution. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI), Montreal, Canada, 1995. 1050-1055

[3] Soon W M, Ng H T, Lim C Y, et al. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*. 2001, 27(4): 512-544

[4] Ng V, Cardie C. Improving machine learning approaches to coreference resolution. In: Proceedings of the 40th Annual Meeting of the ACL, Philadelphia, USA, 2002. 104-111

[5] Yang X, Zhou G, Su J, et al. Coreference resolution using competition learning approach. In: Proceedings of the 41st Annual Meeting of the ACL, Sapporo, Japan, 2003. 176-183

[6] Luo X, Ittycheriah A, Jing H, et al. A mention-synchronous coreference resolution algorithm based on the Bell tree. In: Proceedings of the 42nd Annual Meeting of the ACL, Barcelona, Spain, 2004. 136-143

[7] 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): 136-143

[8] 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

[9] Cardie C, Wagstaff K. Noun phrase coreference as clustering. In: Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, Maryland, USA, 1999. 82-89

[10] Bergler S, Witte R, Khalife M, et al. Using knowledge-poor coreference resolution for text summarization. In: Proceedings of the HLT-NAACL Workshop on Text Summarization, Edmonton, Canada. 2003. 85-92

[11] Bean D, Riloff E. Unsupervised learning of contextual role knowledge for coreference resolution. In: Proceedings of HLT-NAACL, Boston, USA. 2004. 297-304

[12] Wang C S, Ngai G. A clustering approach for unsupervised chinese coreference resolution. In: Proceedings of the 5th SIGHAN Workshop on Chinese Language Processing, Sydney, Australia, 2006. 40-46

[13] 周俊生，黄书剑. 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

[14] Xu R，Wunsch D. Survey of clustering algorithm. *IEEE Transactions on Neural Networks*，2005：645-678

[15] Quinlan J R. Induction of decision trees. *Machine Learning*. 1986，1(1):81-106

[16] Quinlan J R. C4.5：Programs for Machine Learning. Morgan Kaufmann，San Mateo. 1993

[17] Cristianini N，Taylor J S. An Introduction of Support Vector Machines and Other Kernel-Based Learning Methods. Cambridge：Cambridge University Press，2000

[18] Linguistic Data Consortium. ACE (Automatic Content Extraction) Chinese Annotation Guidelines for Entities. Version 5.5. Philadelphia：University of Pennsylvania，2005

[19] Hsu C W，Lin C J. A Comparison of methods for multi-class support vector machines. *IEEE transactions on Neural Networks*. 2002，13(2)：415-425

[20] 董振东，董强，赫长伶. 知网的理论发现. 中文信息学报，2007，(21)4:3-9

[21] Carpenter G A，Grossberg S. A massively parallel architecture for a self-organizing neural pattern recognition machine. *Computer Vision，Graphics，and Image Process*，1987，37(1)：54-115

[22] Vilain M，Burger J，Aberdeen J，et al. A model theoretic coreference scoring scheme. In：Proceedings of the 6th Message Understanding Conference，Morgan Kaufmann，San Francisco，USA，1995. 45-52

[23] Luo X. On coreference resolution performance metrics. In：Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing，Vancouver，Canada，2005. 25-32

An unsupervised approach based on ART network for coreference resolution of Chinese

Li Shiqi，Zhao Tiejun，Chen Chen，Liu Pengyuan

(Harbin Institute of Technology MOE-MS Key Laboratory of Natural Language Processing and Speech，Harbin 150001)

Abstract

This paper proposes a novel unsupervised approach for coreference resolution of Chinese based on adaptive resonance theory (ART) Networks. Through making full use of the characteristics of noun phrases and dynamically adjusting the parameters of the networks，the approach can solve the problem in the present clustering coreference resolution that the number of the output categories is hard to determine. Additionally，the approach performs a feature selection process based on the gain ratio criterion to reduce the noise created by the weak features in differentiation. The method scarcely depends on the hand-labeled corpus and can be directly applied to real texts in multiple fields while ensuring the accuracy. The experiment has shown its encouraging performance on ACE Chinese corpus.

Key words：coreference resolution，unsupervised learning，adaptive resonance theory (ART)，natural language processing