基于 Fuzzy Rough 集模型的汉语人称代词消解

李 凡 刘启和 李洪伟
(电子科技大学计算机科学与工程学院 成都 610054)

摘 要 指代消解是自然语言处理中重要的研究课题之一。结合基于实例的学习方法,提出了一种基于 Fuzzy Rough 集模型的中文人称代词消解方法。该方法的第一步过滤掉与人称代词性别和单复数特征不一致的名词短语,构成候选集,然后按照仅涉及浅层语义和语法知识的属性集对其中的每个名词短语进行标记。第二步利用 Fuzzy Rough 集模型中相关概念选择代表性较强的实例,并对其进行属性值约简,以提高这些实例的泛化能力。以上两步即为学习阶段。第三步即可根据这些实例判断新输入的名词短语是否为代词的先行语。该方法用人民日报语料进行了测试,测试结果表明该方法是有效的。

关键词 指代消解,先行语,Fuzzy Rough 集,基于实例的学习
中图法分类号 TP181 **文献标识码** A

Pronominal Anaphora Resolution within Chinese Text Based on Fuzzy Rough Sets Model

LI Fan LIU Qi-he LI Hong-wei
(School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 610054, China)

Abstract Anaphora resolution is an important issue in natural language processing. This paper presented an approach based on Fuzzy Rough sets model combined with instance-based learning approach to resolve pronominal anaphora within Chinese text. The first phase of the presented approach is preprocessing. In this phase, after extracting noun phases and eliminating those whose number and gender features are inconsistent with pronominal anaphora, the potential antecedents set was formed. Then, the attri-bute values of every noun phase in this set were computed according to an attribute set which only involves shallow syntactic and semantic information. The second phase aimed to select representative examples from the potential antecedents set and reduce redundant attributes to improve the generalization capability of these examples. These tasks were done by using concepts of Fuzzy Rough sets model. The two phases above can be regarded as learning phase. In the last phase, those examples were used to estimate whether a new noun phase is the antecedent of a pronominal anaphor. The presented approach was tested by People Daily corpus. The results show that this approach is effective.

Keywords Anaphora resolution, Antecedent, Fuzzy Rough sets, Instance-based learning

1 引言

指代是自然语言的常见现象,大量出现于篇章和会话中。指代的正确消解是自然语言处理(Natural Language Processing,NLP)研究中的重要课题,并已成为自然语言人机界面、机器翻译(Machine Translation)、自动文摘(Automatic Abstracting)、信息抽取(Information Extraction)等 NLP 应用的关键问题之一。为此,MUC(Message Understanding Conference)将指代消解列为测评任务之一,ACL(Association for Computational Linguistics)和 EACL(the European Chapter of the ACL)都曾设立指代消解的专题会议,Computational Linguistics 还专门出版过指代消解专辑[1]。

目前国外对指代消解问题的研究相对较深入,提出了多种消解方法,例如基于句法分析的方法[2,3]、基于语料库的统计方法[2,4]和基于机器学习的方法[2,5-7]、句法分析和语料库相结合的方法[8]等。汉语中的指代消解作为独立的课题提出并开展系统研究相对较晚,早期的工作[9]是将其作为 NLP 系统中的一个子问题进行处理。在 20 世纪 90 年代末期这一问题开始得到学者们的重视。例如许敏等[10]利用话语中人物焦点变化进行代词消解的研究,王厚峰等[11]提出基于 HNC 理论[12]的代词消解方法等。这些方法综合了语法和语义功能,需要获取较多的知识才能进行有效消解。王厚峰等[13]根据 Mitkov[14]的利用"有限知识"进行指代消解的思路,提出了一种鲁棒性的汉语人称代词消解方法。这些研究从不同侧面加深了对中文文本中指代消解问题的认识,但是,从更多角度对这一问题进行探索,仍将是中文信息处理的重要课题之一。

影响中文文本中名词短语(Noun Phrase,NP)是否为指代先行语的因素较为复杂,并表现出不完整性和不确定性。

到稿日期:2009-05-03 返修日期:2009-08-01 本文受国家自然科学基金(60873077)资助。
李 凡(1972-),男,博士,讲师,主要研究方向为 Rough 集理论及应用、机器学习,E-mail:lifan@uestc.edu.cn;刘启和(1973-),男,博士,副教授,主要研究方向为 Rough 集理论及应用、机器学习;李洪伟(1977-),男,博士,讲师,主要研究方向为信息安全。

Rough 集理论[15]正是处理这类信息的有效数学工具，但经典 Rough 集理论是基于等价关系的，不能满足一些实际问题的需要，因此又发展出一些扩展模型，如基于覆盖的 Rough 集模型[16,17]和 Fuzzy Rough 集模型[18-20]等。本文基于 Fuzzy Rough 集模型，结合基于实例的学习方法，提出了一种汉语人称代词的消解方法。该方法的第一步，按照最基本的规则过滤掉不是代词先行语的 NP，构成候选集，然后按照特定的属性集对其中每个 NP 进行标记。第二步，通过建立 Fuzzy Rough 集模型，选择候选集中代表性较强的 NP，构成用于判断新输入文本中的 NP 是否为代词先行语的实例集，然后进行属性值约简，提高实例所代表的信息的泛化能力。这一步实际上就是利用 Fuzzy Rough 集模型进行知识发现，从中得到关于判别代词先行语的有用信息。以上两步即学习阶段。第三步，根据基于实例的学习方法，对新输入文本中的 NP 是否为代词的先行语进行判断。通过对“人民日报”语料库[21]的实际验证可知，该方法是有效的。

本文第 2 节介绍相关概念，第 3 节讨论消解方法的细节，第 4 节是试验结果及分析，最后是结论。

2 基本概念

2.1 指代消解

指代是指篇章中的一个语言单位(通常是词或短语)与之前出现的语言单位存在的特殊语义关联，其语义解释依赖于前者(目前一般不考虑逆向指代)。用于指向的语言单位称为照应语，被指向的语言单位称为先行语。确定照应语所指的先行语的过程称为指代消解(Anaphora Resolution)[22]。篇章可以视为句子的有序集合，而句子又进一步可以看成由子句序列组成。

本文讨论汉语人称代词的消解问题。根据韩氏理论[23]，人称代词中只有第三人称代词才可用于文内词语照应，所以本文仅探讨第三人称代词(以下简称代词)的消解。

2.2 Fuzzy Rough 集

定义 1 令 U 和 V 为论域，F 是 $U\times V$ 上的 Fuzzy 集，其隶属函数 $F:U\times V\to[0,1]$ 确定了 U 中的元素和 V 中的元素的关系程度，则称 F 为从 U 到 V 的一个 Fuzzy 关系。当 F 是从 U 到 U 的 Fuzzy 关系时，其可简称为 U 上的 Fuzzy 关系。

在实际信息系统中，论域 U 中的元素通常是由属性集合 R 描述的，$R=\{r_1,r_2,\cdots,r_n\}$。$\forall r_i\in R,V_{r_i}$ 代表属性 r_i 的值域，$\forall x\in U,V_{r_i}(x)$ 代表 x 在属性 r_i 上的取值。一般而言，$R=C\cup D$，C 为条件属性集，D 为决策属性集。通过属性集合 $B\subseteq R$ 可以在 U 上建立 Fuzzy 关系。为简便起见，在下文中，B 若表示属性集合，则同时也表示其在 U 上建立的 Fuzzy 关系。

定义 2 非空集合 U 上的 Fuzzy 关系 R 称为 Fuzzy 等价关系当且仅当以下 3 个条件成立：

(1)自反性，$R(x,x)=1,\forall x\in U$；

(2)对称性，$R(x,y)=R(y,x),\forall x,y\in U$；

(3)传递性，$R(x,z)\geqslant\min(R(x,y),R(y,z)),\forall x,y,z\in U$。

若 R 仅满足自反性和对称性，则称 R 为 Fuzzy 相似关系。

定义 3 对非空集合 U 以及 U 上的 Fuzzy 等价关系 R，称 $FAS=(U,R)$ 为 Fuzzy 近似空间。

定义 4 令 $FAS=(U,R)$ 为 Fuzzy 近似空间，F 为 U 上的 Fuzzy 集，F 的下近似和上近似为两个 Fuzzy 集，其定义为[18]：

$$\mu_{\underline{RF}}(x)=\inf_{y\in U}\max\{1-R(x,y),\mu_F(y)\} \tag{1}$$

$$\mu_{\overline{RF}}(x)=\sup_{y\in U}\min\{R(x,y),\mu_F(y)\} \tag{2}$$

最初提出的 Fuzzy Rough 集模型是基于 Fuzzy 等价关系的。在进一步的研究中，学者们对该模型进行了扩展，提出了基于任意 Fuzzy 关系 R 的广义 Fuzzy Rough 集的概念[24]。令 T 和 S 分别为下半连续 T 三角模和上半连续 S 三角模，并且二者关于 N 对偶，即 $\forall x,y\in[0,1]$，有 $S(x,y)=N(T(N(x),N(y))$ 和 $T(x,y)=N(S(N(x),N(y))$。$N(x)$ 为从[0,1]到[0,1]的减映射，且满足 $N(N(x))=x$，一般 $N(x)=1-x$。在此基础上，定义 F 的下近似和上近似为：

$$\mu_{\underline{RF}}(x)=\inf_{y\in U}S(N(R(x,y)),\mu_F(y)) \tag{3}$$

$$\mu_{\overline{RF}}(x)=\sup_{y\in U}T(R(x,y),\mu_F(y)) \tag{4}$$

目前最常用的是基于 Fuzzy 相似关系的 Fuzzy Rough 集模型。

3 基于 Fuzzy Rough 集的汉语人称代词消解方法

基于实例的学习是一种简明有效的机器学习方法。学习的过程只是简单地存储已知的训练数据。当遇到新的查询实例时，提取与之相似的实例来对其进行分类[25]。

基于实例学习的一个不足是当检索相似的训练实例时一般考虑描述实例的所有属性。如果目标概念仅依赖于其中的几个属性，那么和事实最相似的实例之间可能存在较大差异[25]。在英语指代研究中也发现类似的问题，即在使用较多语言规则构造特征属性集用于指代消解时，消解效果反而有一定程度的下降[26]。可以从以下两个方面理解这个问题。首先，由于自动提取这些特征属性的取值往往涉及到对深层语法和语义知识的处理，而目前尚没有精度较高的自动处理方法，因而容易引入误差。在汉语处理中，这一问题更为突出，表现在提取一些基本特征(如 NP 的单复数信息、人名的性别信息以及人称代词的语法角色等[13])较印欧语系更为困难，而且结果有较大误差。其次，在机器学习领域中，一般都倾向于采用相对较小的特征集合，以提高泛化能力。因此，为了提高系统的识别精度，有必要对描述实例的特征集进行筛选，提取其中的有效子集来描述各个实例。除此之外，还必须对已见实例进行选择，使系统存储的实例具有较好的代表性。本文利用 Fuzzy Rough 集模型来解决这两个问题。系统处理的示意图如图 1 所示。

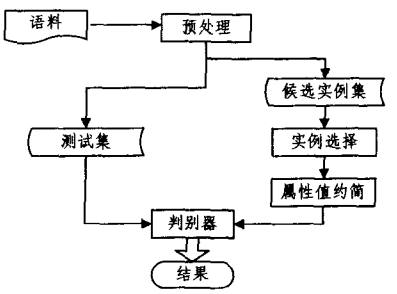

图 1 系统处理示意图

3.1 语料预处理

输入的语料事先已经完成分词和词性标注，并人工标记

出代词的先行语。然后执行以下几个步骤的预处理。

Step1　在发现代词后，向前回溯若干句，寻找其中的 NP（也包括句中的代词）。回溯的范围按照以下的原则确定：

(1)如果代词的先行语在代词前两句范围内，则将代词所在句和其前面两句内的所有 NP 提取出来作为候选。

(2)如果代词的先行语在代词前两句以外，则将代词所在句和先行语所在句之间(包括先行语所在句)的所有 NP 都提取出作为候选。

此时应注意几种特殊情况。第一，在(1)中，如果前溯一句即已经到达段首，则停止回溯。在新闻报道中，段落开头经常有引导句，例如：本报/r 上海/ns 1 月/t 1 日/t 电/n 记者/n 龚/nr 雯/nr 报道/v:/w …。当遇到这类引导句时，也认为已经到达段首。第二，在回溯时遇到括号中或直接引语中的句子，均忽略不记，若代词在直接引语里，则只回溯到该直接引语的开始处。第三，书名号等特殊符号内的 NP 不提取。

Step2　找到的 NP 按照单复数信息和性别信息进行过滤，若 NP 的单复数信息或性别信息与当前人称代词不相容，则该 NP 被过滤。本文采用的具体方法见 4.1 节。经过这一步得到的 NP 集合记为 U。

Step3　针对特征集合 $R=C\cup D$，计算 U 中每个 NP 在各属性上的属性值，以此对 NP 进行标记。其中，$D=\{d\}$，$V_d=\{0,1\}$，$\forall x\in U$，有：

$$V_d(x)=\begin{cases}1 & \text{若 } x \text{ 是代词先行语}\\ 0 & \text{若 } x \text{ 不是代词行先语}\end{cases} \tag{5}$$

在条件属性集 C 中各属性的选择上，本文的原则是尽量避免采用深层语法和语义知识。这主要是因为对这类知识的自动处理尚无成熟方法，可能带来误差。在综合实际指代消解系统[13,14]所采用的特征集并对实际语料进行考察后，选用以下特征：

(1) 角色(role)，表明候选 NP 与代词在“角色”上是否一致，有三种取值，$V_{role}=\{y,n,un\}$，分别表示该 NP 与代词在“角色”上一致、不一致以及不确定是否一致。这里的角色并不是指完整的语法角色，而仅指“施事”或“受事”两种属性。具体见 4.1 节。

(2) 首个候选(fn)，$V_{fn}=\{true,false\}$，表明该 NP 是否是所在段落中单复数信息和性别信息一致的首个候选。

(3) 结构并列(sp)，$V_{sp}=\{true,false\}$。这里所谓的结构并列，是指 NP 是否与代词有同样的搭配结构。

(4) 介词短语(pp)，$V_{pp}=\{true,false\}$，表明 NP 是否在句子的某个介词短语中。

(5) 距离(rd)，这里指 NP 到代词的距离。具体而言，NP 在代词前一子句，则 $rd=0.1$，在代词前两个子句，则 $rd=0.2$，依此类推；NP 在代词前一句，则 $rd=10$，在代词前两句，则 $rd=11$，依此类推；NP 在代词前一段，则 $rd=100$。

(6) 是否与代词在性别信息上显式一致(gen)，$V_{gen}=\{true,false\}$。

(7) 是否与代词在单复数信息上显式一致(num)，$V_{num}=\{true,false\}$。

例 1　在/p 总厂/n 所/u 属/v 的/u [石景山/ns 热电厂/n]nt，/w 李/nr 鹏/nr 首先/d 向/p [华北/ns 电管局/n]nt 、/w 电厂/n 负责人/n 详细/ad 询问/v 了/u …。/w 随后/d，/w 他/r 又/d 实地/d 察看/v 了/u …。

发现代词“他”后，向前回溯寻找 NP，经过过滤，“李鹏”(NP1)和“电厂负责人”(NP2) 进入 Step3。标记结果如表 1 所列。

表 1　标记示例

	role	fn	sp	pp	rd	gen	num	d
NP1	y	true	true	false	10	true	true	1
NP2	n	false	false	true	10	false	false	0

Step4　经以上几步预处理后，将 NP 分为训练集和测试集，训练集记为 CAN。

3.2　实例选择

实例选择主要利用 Fuzzy Rough 集模型进行。首先构造 CAN 上的 Fuzzy 相似关系。令 $R=C\cup\{d\}$，$C=\{role,fn,sp,pp,rd,gen,num\}$，$\forall r\in C\setminus\{rd\}$，$\forall x,y\in CAN$ 在 r 上的相似度 $r(x,y)$定义为：

$$r(x,y)=\begin{cases}1 & V_r(x)=V_r(y)\\ 0 & \text{其它}\end{cases} \tag{6}$$

当 $r=rd$，$\forall x,y\in CAN$，$r(x,y)$定义为：

- 当 $x=y$，$r(x,y)=1$；
- 当 x,y 都和代词在同一个子句中，$r(x,y)=0.95$；
- 当 x 和 y 到代词之间的距离相差一个子句，$r(x,y)=0.88$；
- 当 x 和 y 到代词之间的距离相差两个子句，$r(x,y)=0.75$；
- 当 x 和 y 到代词之间的距离相差三个或三个以上的子句，$r(x,y)=0.65$；
- 当 x 和 y 到代词之间的距离相差一个句子，且句子中没有子句，$r(x,y)=0.88$；
- 当 x 和 y 到代词之间的距离相差一个句子，且句子中有其他子句，$r(x,y)=0.58$；
- 当 x 和 y 到代词之间的距离相差两个句子，$r(x,y)=0.4$；
- 当 x 和 y 到代词之间的距离相差三个或三个以上的句子，$r(x,y)=0.15$；
- 当 x 和 y 到代词之间的距离相差一个自然段，$r(x,y)=0.1$；

在以上 $r(x,y)$的定义中，x 和 y 到代词之间距离相差的多少是用 x 和 y 的 rd 值来判断的。基于 $r(x,y)$有以下定义。

定义 5　定义 Fuzzy 二元关系 C 为：

$$C(x,y)=\prod_{r\in C}r(x,y) \tag{7}$$

显然 C 满足自反性和对称性，因此有以下定理。

定理 1　C 是 U 上的 Fuzzy 相似关系。

另一方面，由决策属性 d，可以对 U 做划分，记 $U/\{d\}=\{Y,N\}$，其中，$Y=\{x\in U\mid V_d(x)=1\}$；$N=\{x\in U\mid V_d(x)=0\}$。取 $S(x,y)=\max(x,y)$，$T(x,y)=\min(x,y)$，$N(x)=1-x$，则 $\forall x\in U$，有：

$$\mu_{\underline{C}Y}(x)=\inf_{y\in U}\max\{1-C(x,y),Y(y)\} \tag{8}$$

$$\mu_{\underline{C}N}(x)=\inf_{y\in U}\max\{1-C(x,y),N(y)\} \tag{9}$$

其中，

$$Y(y)=\begin{cases}1 & \text{若 } y\in Y\\ 0 & \text{其它}\end{cases},\quad N(y)=\begin{cases}1 & \text{若 } y\in N\\ 0 & \text{其它}\end{cases}$$

由式(8),式(9)易得以下定理。

定理 2 若 $V_d(x)=1$,则 $\mu_{\underline{CN}}(x)=0$;若 $V_d(x)=0$,则 $\mu_{\underline{CY}}(x)=0$。

基于以上的概念,可以设计实例选择算法的具体步骤如下所示。下文中 $Card(A)$ 代表集合 A 的基数。

实例选择算法

输入:从语料中提取的 NP 集合 CAN。

输出:实例集合 $INST_1$,$INST_2$,前者为是代词先行语的实例集合,后者为不是代词先行语的实例集合。

Step1 令 $INST_1=INST_2=\emptyset$;

Step2 依次读入 CAN 中的一个实例 x,共进行 $Card(CAN)$ 步;

Setp2.1 令 $\delta=1.0$,$i=1$;

Setp2.2 依次读入 CAN 中的第 i 个实例 y_i;

Setp2.3 计算 $C(x,y_i)$;

Setp2.4 按 $V_d(x)$ 的值进行判断:

(1)若 $V_d(x)=1$,判断 $V_d(y_i)$ 是否为 1。若不是 1,则 $t=1-C(x,y_i)$,否则转 Setp2.6;

(2)否则,判断 $V_d(y_i)$ 是否为 0。若不是 0,则 $t=1-C(x,y_i)$,否则转 Setp2.6;

Setp2.5 若 $t<\delta$,则 $\delta=t$;

Setp2.6 $i=i+1$,若 $i<Card(CAN)$,则转 Setp2.2,否则转 Setp2.7;

Setp2.7 若 $\delta\geqslant\lambda$,则判断 $V_d(x)$ 的值,若 $V_d(x)=1$,则将 x 加入 $INST_1$,否则加入 $INST_2$。

其中,Step2.4 利用定理 2 来简化计算。Step2.7 中,在计算得到 $\mu_{\underline{CY}}(x)$ 或 $\mu_{\underline{CN}}(x)$ 后,若下近似值大于或等于 λ,则 x 被存储到实例集合中。阈值 λ 的设定是为了使实例有较好的代表性。本文取 $\lambda=0.7$。

3.3 属性值约简

为了使 3.2 节中得到的实例所代表的信息有较好的泛化性,必须检验实例的属性是否冗余,即对 $INST_1$ 和 $INST_2$ 中的实例进行属性值约简。由式(8),式(9)易得出以下结论。

定理 3 $\forall x\in CAN$,则 $\forall B\in C$,有 $\mu_{\underline{BY}}(x)\leqslant\mu_{\underline{CY}}(x)$ 及 $\mu_{\underline{BN}}(x)\leqslant\mu_{\underline{CN}}(x)$。

由定理 3 可以定义属性值约简的概念。

定义 6 $\forall x\in CAN$,则 $B\subseteq C$ 称为 x 的属性值约简当且仅当以下两个条件同时成立:

(1)$\mu_{\underline{BF}}(x)=\mu_{\underline{CF}}(x)$;

(2)$\forall b\in B$,$\mu_{\underline{B\setminus\{b\}F}}(x)<\mu_{\underline{CF}}(x)$。

其中,若 $V_d(x)=1$,则 $F=Y$;若 $V_d(x)=0$,则 $F=N$。

除了考虑实例的下近似值之外,另外一类信息也不可忽略,即实例的支持度。这一概念由经典 Rough 集模型中的相关概念[27]推广而来,本文定义如下。

定义 7 $\forall x\in CAN$,$R=C\cup\{d\}$ 为描述 x 的属性集,则 x 的支持度定义为:

$$Supp(x)=Card(\{y\in U\mid C(x,y)>0\wedge V_d(y)=V_d(x)\})$$

由以上定义,可以设计属性值约简算法。具体步骤如下所示。

属性值约简算法(仅示出对 $INST_1$ 的计算步骤,对 $INST_2$ 的处理完全类同)

输入:实例集合 $INST_1$。

输出:$INST_1$,其中每个实例 x 的属性值约简 $x.RED$ 和支持度 $x.Supp$ 均得到计算。

Step1 令 $k=1$;

Step2 依次读入 $INST_1$ 中的第 k 个实例 x_k;

Step3 令 $RED=\emptyset$;

Setp4 若 $\mu_{\underline{CY}}(x_k)=1.0$,则令 $ATTR=\emptyset$;否则令 $ATTR=\{rd\}$;

Setp5 令 $C'=\{role,fn,sp,pp,gen,num\}$,对 C' 中的属性进行排序。并令 $ATTR=C'\cup ATTR$;

Setp6 令 $i=1$;

Step7 检验 $ATTR$ 中的属性是否有冗余:

Setp7.1 去除 $ATTR$ 中的第 i 个属性 r_i,计算 $\mu_{\underline{ATTRY}}(x_k)$;

Setp7.2 若 $\mu_{\underline{ATTRY}}(x_k)<\mu_{\underline{CY}}(x_k)$,则将 r_i 放回原位置,$i=i+1$;

Setp7.3 若 $Card(ATTR)=1$ 或 $i>Card(ATTR)$,则令 $x_k.RED=ATTR$,转 Setp7.4;否则转 Setp7.1;

Setp7.4 令 $j=0$;

Setp7.5 遍历 $INST_1$(或 $INST_2$)中所有元素,若某个元素 y 满足 $ATTR(x_k,y)>0$ 则 $j=j+1$;

Setp7.6 令 $x_k.Supp=j$;

Step8 $k=k+1$,若 $k<Card(INST_1)$(或 $k<Card(INST_2)$)转 Step2,否则转到 Step9;

Step9 去除 $INST_1$ 中重复的实例。

下面对属性值约简算法中的细节做进一步的解释。

在 Step3 中,如果 $\mu_{\underline{CY}}(x_k)=1.0$,则 $\forall y\in CAN$,有以下两种可能:(1)$Y(y)=1$;(2) $Y(y)=0$ 且 $C(x_k,y)=0.0$。在这两种情况下,去掉属性 rd 均不会改变 $\mu_{\underline{CY}}(x)$ 的值。因此可以在计算约简前先去掉 rd。在 Step4 中,对 C' 中的属性进行排序,是指按照每个属性相对于 x 对决策属性的支持度[28]进行排序。$\forall c\in C'$,其相对于 x 对决策属性的支持度 $s(c,x)$ 定义为:$s(c,x)=Card(\{y\in CAN\mid V_c(y)=V_c(x)\wedge V_d(y)=V_d(x)\})$,即相对不重要的属性先尝试被约简。

3.4 判断新输入文本中 NP 是否为代词先行语

以上几个步骤即是学习过程。在学习过程完毕后,得到了用于判断新输入 NP 是否为代词先行语的实例集合。接下来即可对测试集中的 NP 进行判断。一般而言,该 NP 与 $INST_1$ 和 $INST_2$ 中的实例都可能相似。在判断时主要考虑 NP 与实例的近似度以及实例本身的下近似值和支持度值。以下是具体的判别算法。

判别算法

输入:待判断 NP,以 x 表示,已经得到 x 在属性集 C 中的各个属性上的取值。

输出:x 的属性 $decision$。$decision=$true 代表 x 是代词先行语,$decision=$false 代表 x 不是代词先行语,$decision=$unknown 表示无法判断 x 是否是代词先行语。

Step1 对 $INST_1$ 中的每个实例 y_i,按照 y_i 属性值约简后的属性 R',计算 $t_i=R'(x,y_i)\times\mu_{\underline{CY}}(y_i)$。记录 $t_1=\max_i\{t_i\}$ 和 t_1 对应的实例的支持度 $Supp_1$;

Step2 对 $INST_2$ 中的每个实例 y_j,按照 y_j 属性值约简后的属性 R',计算 $t_j=R'(x,y_j)\times\mu_{\underline{CN}}(y_j)$。记录 $t_2=\max_j\{t_j\}$ 和 t_2 对应的实例的支持度 $Supp_2$;

Step3 若 $t_1>t_2>0$,则 $decision=$true,若 $t_2>t_1>0$,则 $decision=$false,算法结束。若 $t_1=t_2>0$,则转 Step4;否则 $decision=$unknown,算法结束。

Step4 若 $Supp_1\geqslant Supp_2$,则 $decision=$true,否则 $decision=$false。

4 实验与分析

4.1 实验语料及其处理细节

本文的实验语料来自北京大学计算语言研究所标注的1998年1月"人民日报"语料[21]。该语料已经完成分词和词性标注,但未标注代词先行语。我们对1月1日到1月6日的所有含第三人称代词的句子进行了处理(具体统计情况如表2所列),手工标注了先行语。预处理中需要大量用到知网[29]的相关信息,为此我们将知网 HowNet System Version 2000 的内容导入数据库中,以利于机器查找。

表2 代词数目统计

	他	他们	她	她们
总数	390	178	115	13

在预处理阶段,采用NP的简单形态信息及其在知网中的标记来判断其单复数信息和性别信息。前者如:"孩子/n 们/k","郭/nr 树范/nr 和/c 闫/nr 戌麟/nr","于/nr 永波/nr 等/u";后者如以下二例所示。

例2 "人民"一词,在知网中的标记信息的DEF项为:DEF=human|人,mass|众,其中的"mass"属性表明该词带有复数特征。

例3 "奶奶"一词,在知网中有4处标记,每个标记的DEF项中都有female属性,因此可判断该词有阴性特征。

对于属性"role"的取值,主要由NP对应的动词在知网中相应义原的性质来判断。例如,在例1中,代词"他"所对应的动词"察看"在知网中有4个义项,其共同的上位义原为look|看,其描述为{agent,content}。由知网记号的含义可判断"他"为施事。同样,检查"询问"可以确定"李鹏"也在其所在的句子中充当施事,因此"李鹏"在属性"*role*"上的取值为"true"。而"电厂负责人"则因为处于介词短语中,在此属性上取"false"。

经以上的自动处理后还存在一定数量的NP无法确定其在某个属性上的取值。这些NP又经人工处理补充不足信息。

对于人名(标注集中的记号为nr),由于目前尚没有合适的工具,因此用人工判断其性别特征。

在判断NP是否为代词先行语时,若出现NP与$INST_1$和$INST_2$中实例的相似度均为0的情况,即判别算法输出"unknown"时,如果其他NP可以判断出是代词先行语,则不考虑该NP。此时若其他NP都无法判断是否为先行语,则将最近的性别和单复数一致的NP作为代词先行语,若仍无法确定,则选择最近的性别或单复数一致的NP(单复数信息优先于性别信息),若还是无法确定,则选取最近的NP作为先行语。

4.2 测试结果及分析

测试采用交叉验证(cross validation)的方式,将每个代词各自对应的语料平均分为6份,取5份为训练集,1份为测试集。指代消解的测试一般采用精度(Precision)和召回率(Recall)作为测评标准,定义如下:

$$Precision=\frac{\text{正确消解的代词数目}}{\text{系统识别的代词数目}}\times 100\% \tag{10}$$

$$Recall=\frac{\text{正确消解的代词数目}}{\text{总的代词数目}}\times 100\% \tag{11}$$

此处由于所有代词都在语料中标出,因此两个指标相同,仅取前者做评价。"她们"的样本数量过少,因此未进行测评。这里,正确消解的代词是指每个代词所有的先行语均被正确找出。若代词先行语未找全或先行语中有不正确的NP,都视为不正确消解。实验结果如表3所列。

表3 实验结果

	训练集/测试集						
	第2份至第6份/第1份	第1份,第3份至第6份/第2份	第1份至第2份,第4份至第6份/第3份	第1份至第3份,第5份至第6份/第4份	第1份至第4份,第6份/第5份	第1份至第5份/第6份	平均值
测试结果	他:81.5% 他们:63.7% 她:69.3%	他:69.5% 他们:80.1% 她:83.5%	他:83.5% 他们:70.3% 她:80.9%	他:84.7% 他们:69.5% 她:77.4%	他:78.4% 他们:66.5% 她:82.6%	他:82.9% 他们:77.6% 她:84.9%	80.1% 71.3% 79.8%

在消解结果上看,"他"、"她"和"他们"的消解结果分别差于、接近和好于文献[13]中得到的结果。在对"他们"进行消解时,因为语料中"他们"的先行语常常相距代词较远,而本方法能较好地描述远距离约束,所以能得到更好结果。

错误情况分析如下。首先,基于实例的学习方法需要大量存储实例作为支持。因而在训练集涵盖的范围不足时,可能产生一定误差,这也是测试结果存在波动的原因。其次是对NP的自动处理方法尚待改进。例如对较长NP的自动识别上的不完善(如丹东/ns 驻军/n 某部/r 家属/n 工厂/n),导致一些词被误纳入集合CAN中,影响$INST_1$和$INST_2$中实例的质量;又如在语料中,一些"其他专名"(标注集中记号为nz)可以作为代词的先行语(如"米老鼠"作为"他"的先行语),而有一些不能充当代词先行语,还有一些专有机构被标注为数词(标注集中记号为m),如"110"。目前的判断方法还不完善,为了考虑上述几类词也导致类似情况的发生。第三,定义相似度$r(x,y)$时,目前只根据实验做了初步定义,需要进一步完善。第四,在出现无法判断出先行语的情况下,若采用更合理的方法,精度有进一步提高的可能。最后,对复杂先行语(如"他们"的先行语:…政务司/n 司长/n 陈/nr 方/nr 安生/nr 出任/v 委员会/n 主席/n 、/w 高/nr 苕华/nr 任/v 副/b 主席/n,/w 其他/r 成员/n 包括/v 8/m 位/q 特区/n 政府/n 高级/a 官员/n 和/c 12/m 位/q 社会/n 不同/a 界别/n 人士/n ./w 他们/r 的/u 任期/n 均/d 为/v 两/m 年/q ./w)的处理目前还没有较合适的方法,也影响消解结果。

结束语 如何更好地解决中文文本中出现的指代现象,是中文信息处理的重要问题之一,还需要在现有研究的基础上进一步探索更多的途径。本文基于Fuzzy Rough集模型,结合基于实例的学习,提出了一种人称代词消解方法。经过实例测试可知,该方法是有效的。

本文下一步的工作主要将在以下几个方面展开:首先将进一步扩大现有标注语料的规模,并在此基础上做相应的测试和分析。其次,进一步分析指代现象,完善特征集合。第三,尝试采用优化方法进一步调节$r(x,y)$值。第四,指代消解是一个综合性很强的任务,需要开发准确率较高的预处理工具用于识别长名词短语,判断NP的性别和数量特征等。

参 考 文 献

[1] The Special Issue on Computational Anaphora Resolution[J]. Computational Linguistics,2001,27(4)

[2] Mitkov R. Anaphora resolution: the state of the art. Working paper(Based on the COLING'98/ACL'98 tutorial on anaphora resolution)[M]. University of Wolverhampton, Wolverhampton,1999

[3] Renata V, Massimo P. An Empirically-based System for Processing Definite Descriptions[J]. Computational Linguistics, 2000,26(4):525-579

[4] Ge Niyu, John H, Eugene C. A Statistical Approach to Anaphora Resolution[C]// Proceedings of COLING-ACL8. Canada,1998: 161-170

[5] Soon W M, Ng H T, Lim D C. A Machine Learning Approach to Coreference Resolution of Noun Phrases[J]. Computational Linguistics,2001,27(4):521-544

[6] Strube M, Muller C. A Machine Learning Approach to Pronoun Resolution in Spoken Dialogue[C]// Proceedings of 2003 ACL. 2003

[7] Orasan C, Evans R, Mitkov R. Enhancing Preference-Based Anaphora Resolution with Genetic Algorithms[C]// Proceedings of NLP2000. University of Patras, Greece,2000:185-195

[8] Mitkov R. Anaphora resolution: a combination of linguistic and statistical approaches[C]// Proceedings of the Discourse Anaphora and Anaphor Resolution. Lancaster, UK,1996

[9] 李家治,陈永明. 机器理解自然语言中有关代词处理的几个问题[C]//自然语言理解年会论文集. 1986

[10] 许敏,王能忠,马彦华. 汉语中指代问题的研究及讨论[J]. 西南师范大学学报:自然科学版,1999,6:633-637

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

[12] 黄曾阳. HNC(概念层次网络)理论—计算机理解语言研究的新思路[M]. 北京:清华大学出版社,1998

[13] 王厚峰,梅铮. 鲁棒性的汉语人称代词消解[J]. 软件学报,2005,16(5):700-707

[14] Mitkov R, Evans R, Orasan C. A new, fully automatic version of Mitkov's knowledge-poor pronoun resolution method[C]// Proceedings of CICLing-2002. Mexico,2002:168-186

[15] Pawlak Z. Rough sets. Theoretical aspects of reasoning about data[M]. Kluwer Academic Publishers,1991

[16] Zhu W, Wang F Y. On Three Types of Covering Rough Sets [J]. IEEE Transactions On Knowledge and Data Engineering, 2007,19(8):1131-1144

[17] Zhu W, Wang F Y. Reduction and Axiomization of Covering Generalized Rough Sets[J]. Information Sciences, 2003, 152: 217-230

[18] Dubois D, Prade H. Putting fuzzy sets and rough sets together. Intelligent Decision Support[M]// Slowiniski R, ed. Kluwer Academic Publishers,1992:203-232

[19] Wu Wei-zhi, Mi Ju-sheng, Zhang Wen-xiu. Generalized fuzzy rough sets[J]. Information Sciences,2003,151:263-282

[20] Chen D G, Wang X Z, Yeung D S, et al. Rough approximations on a complete completely distributive lattice with applications to generalized rough sets[J]. Information Science,2006,176:1829-1848

[21] http://www.icl.pku.edu.cn

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

[23] Halliday M, Hasan A K. Cohesion in English[M]. London, Longman Group,1976:27-28

[24] Yeung D S, Chen D G, et al. On the Generalization of Fuzzy Rough Sets[J]. IEEE Transactions on Fuzzy Systems,2005,13 (3):343-361

[25] Mitchell T M. Machine Learning[M]. McGraw-Hill Companies, Inc,1997

[26] Ng V, Cardie C. Improving Machine Learning Approaches to Coreference Resolution[C]// Proceedings of 2002 ACL. Philadelphia,2002:104-111

[27] Pawlak Z. Rough sets, decision algorithms and Bayes' theorem [J]. European Journal of Operational Research,2002,136:181-189

[28] Grzymala-Buss J W. LERS(A system for learning from examples based on rough setsIntelligent Decision Support[M]// R. Slowinski, ed. Kluwer Academic Publishers,1992:3-18

[29] http://www.keenage.com/html/c_index.html

(上接第 216 页)

次数,又能够比较全面地反映各因素的不同水平对实验指标的影响。对于因素复杂的实验问题,它是一种行之有效的方法。在应用正交实验设计方法时,除了依靠数学分析外,还必须根据实践经验和专业知识,对具体问题具体分析,对实验结果进行验证,这样才能得到正确可靠的结论。在语音识别的特征参数选择中,采用正交实验设计法,通过少数有代表性的几个试验,得出了所需要的结论,大大缩短了实验周期,避免了盲目的实验,提高了设计效率。

另外,参数的简单组合方案特征维数较高,直接影响随后的分类识别的速度和控制速度,而按照本文所提出的方法组合后避免了维数的增加,从而在提高识别率的同时不至于过多地增加时间开销。

参 考 文 献

[1] Emmanouilidis C, Hunter A. Multiobjective Evolutionary Setting for Feature Selection and a Commonality-Based Crossover Operator[C]// 2000 IEEE International Conference on Evolutionary Computation, CEC '2000. 2000:309-316

[2] Ho S Y, Lin H, Liauh W H, et al. Orthogonal Particles Swarm Optimization and Its Application to Task Assignment Problems [J]. IEEE Transactions on Systems, Man and Cybernetics, 2008,38(2):288-298

[3] Liang X B. Orthogonal Designs with Maximal Rates[J]. IEEE Transactions on Information,2003,49(10):2468-2503

[4] Seberry J, Finlayson K, Adams S S, et al. Orthogonal Designs with Maximal Rates[J]. IEEE Transactions on Signal Processing,2008,56(1):256-265

[5] 《现代应用数学手册》编委会. 现代应用数学手册—概率论与随机过程卷[M]. 北京:清华大学出版社,2000

[6] 杨大利,徐明星,吴文虎. 语音识别特征参数选择方法研究[J]. 计算机研究与发展,2003,40(7):963-969

[7] 陈魁. 实验设计与分析[M]. 北京:清华大学出版社,1996

[8] Abu-Shikhah N, Deriche M. A Robust Technique for Harmonic Analysis of Speech[C]//2001 IEEE International on Acoustics, Speech and Signal Processing, ICASSP '01. 2001:877-880

[9] Virtanen T, Klapuri A. Separation of Harmonic Sounds Using Linear Models for the Overtone Series[C]//2002 IEEE International on Acoustics, Speech and Signal Processing, ICASSP'02. 2002:1757-1760