自然语言文本指代消解技术研究*)

史树敏[1,3] 黄河燕[2] 刘东升[3]

(南京理工大学计算机科学与技术学院 南京 210094)[1] (中科院计算机语言信息工程中心 北京 100089)[2]

(内蒙古师范大学计算机与信息工程学院 呼和浩特 010022)[3]

摘 要 指代消解处理是自然语言处理的关键环节,也是众多语言工程项目的核心任务。本文针对指代消解的一些基本问题进行阐述,按照时间线索,对国内外各类指代消解技术方法的研究情况进行分析,阐明了指代消解技术目前的主流方法和技术线路,最后对未来汉语指代消解技术的研究前景加以展望。

关键词 指代,共指,指代消解,共指消解,机器学习方法

The Technologies of Anaphora Resolution in Text Natural Language Processing

SHI Shu-Min[1,3] HUANG He-Yan[2] LIU Dong-Sheng[3]

(Sch. of CST, NJUST, Nanjing 210094)[1] (CCLIE, CAS, Beijing 100097)[2] (Col. of CIE, IMNU, Huhhot 010022)[3]

Abstract Abstract Anaphora resolution is a key step in Natural Language Processing (NLP) and a kernel task in many language engineering applications. This paper analyses some studies of anaphora and coreference resolution in China and abroad under temporal sequence, discusses all kinds of methods and technologies mainly applying to anaphora and coreference resolution. Finally, authors briefly set forth the direction of research on Chinese anaphora and coreference resolution in the future.

Keywords Anaphora, Coreference, Anaphora resolution, Coreference resolution, Machine learning method

1 引言

指代是一种复杂的语言现象,广泛存在于自然语言的各种表达中,包括人称、指示代词指代,零形指代,名词短语间指代等。指代是语篇中某一语言成分和另一语言成分间在指称意义上互相解释的关系,即以一个成分做另一成分的参照点来说明信息。一般分两种情况:回指(Anaphora,亦称指示性指代)和共指(Coreference,亦称同指)。所谓回指是当前的指示语与上下文中出现的词、短语、句子(句群)存在密切的语义关联;共指则是指两个名词(短语)指向真实世界中同一参照体。回指和共指消解,所需知识和消解步骤基本一致,但是所处理的指代关系性质不同。确定照应语所指先行词的过程称为指代消解,是篇章理解的关键问题,是自然语言处理的核心任务之一。在信息抽取、机器翻译、自动文摘等应用中都有大量的指代消解问题亟待解决。各种指代间差异较大,采用单一方法和语言模型很难解决全部的指代问题,所以需对每一种指代现象都进行深入研究。指代消解任务本身需要多级知识(句法、语义、领域)支撑,在当前自然语言处理水平下,要有效得到所需的知识仍较困难,因此广泛深入地开展各类指代消解技术研究,特别是机器学习方法彰显出了前所未有的重要性。

2 几个基本问题

2.1 主要研究内容与应用领域

针对专名消解、名词(短语)消解、代词消解的研究较多;零形指代(先行语为空)与预指(先行语出现在照应语之后)等见诸文献的较少;共指消解是目前的研究热点。有不同的模型和算法适用于上述不同方面的指代消解问题。处理指代问题存在的主要困难,一是涉及到众多特征,任一特征均非完全可靠,都存在语言现象的例外;二是自动正确识别特征困难,受到命名实体识别正确性、外部词典完备性等其他环节和外部知识的影响;另外,语篇相关性导致不同语言的指代消解研究的难度也有所不同,如汉语同英语等印欧语系语言的不同。指代消解在许多自然语言工程领域起着非常重要的作用。如自动文摘处理中,通常是先直接从文本中抽取句子,再将抽取的句子组织起来。由于抽取的句子可能含有照应语,而其所对应的先行语所在的句子却未能被抽取,导致无先行语的句子出现,使句子在逻辑上不衔接,需通过指代消解来确定其先行语,改善文摘可读性;机器翻译中,各语种代词用法各异(如英语中冗余代词 it 的存在)、中文零形回指现象的普遍存在,不进行指代消解直接翻译,翻译质量不理想;信息检索也尝试在话题识别与跟踪中引入指代消解技术;信息抽取中的一个基础性环节就是指代消解处理,其重要性更是不言而喻。

2.2 理论方法与基本解决思路

先后有两种影响深远的指代消解算法理论模型:一个是 Hobbs 提出的朴素 Hobbs 算法及其改进;一个是 Grosz 等人提出的中心理论(Center Theory)。起初解决的指代问题只限于解释语篇局部连贯性,在语篇解读中仅考虑前一语句的实体,没有提出任何解决语篇中前两句,甚至更远的先行词的问题。后多有扩展,20 世纪 80 年代中后期开始受到广泛关

*)受到国家"八六三"高技术研究发展计划基金(2006AA01Z152)资助。**史树敏** 博士研究生,主要从事自然语言处理、信息抽取方面的研究;**黄河燕** 研究员,博士生导师,主要从事自然语言处理、机器翻译方面的研究;**刘东升** 教授,主要从事计算机应用及 CAE 方面的研究。

文中不作特别说明时,按照大部分文献表述习惯将指示性指代消解与共指消解统称为指代消解。

注。消解指代处理的大体思路是:先构造先行语候选集,再从候选中作多对一的选择。早期算法是采用手工建立的逻辑规则和确定的突显性值等进行指代消解,处理的对象主要是代词,绝大多数算法应用语法信息相关的逻辑规则,很少使用语义信息。基于句法的消解方法较早被系统所采用,由于用于区分先行/照应语的知识匮乏,只能应用少量规则从大量的候选项中筛选,移植性和自动化程度低。上个世纪 90 年代以来,随着 Internet 的迅猛发展,大量实验语料易于获得;MUC 和 ACE 评测会议的召开也极大推动了指代消解的研究,随着语料库语言学的发展,基于语料库的指代消解方法也相继出现;统计模型、督导、无督导机器学习方法的运用也日渐增加。

2.3 国内外研究情况

大多数早期的指代消解工作都是处理代词回指问题。Hobbs 算法是最早采用计算方法实现指代消解的算法之一,消解过程不依赖任何语义或语篇信息,算法证明了采用计算方法可有效地解决指代问题,但只能用于代词消解,未说明可否扩展到非代词的情况。文[1]提出一种具有广泛影响力的 RAP(Resolution of Anaphora Procedure)指代消解算法,实现句内和句间的第三人称及反身代词的指代消解,算法先利用 McCord's 提出的槽文法解析获得文档语法结构,再通过计算先行语的突显性(salience)和过滤规则实现消解,未使用语义和真实世界知识来评估候选项,算法的缺点是需事先通过人工方法对语料做简化处理,只考虑第三人称,且需对文档建立深层完整的解析树。Mitkov 对指代消解问题进行了深入研究,提出了有限知识等消解算法,并发表了系列论文,推动了指代消解的发展。代词和名词短语消解在 90 年代中期由于机器学习方法的应用而得到极大发展。1995 年,McCarthy 分析基于规则方法的缺点,提出一种新颖的思路解决指代消解问题,采用决策树方法消解商业投资领域文档的指代关系[2],其思想为日后的研究开辟了一条具有深远意义的全新道路。近期关于共指消解的研究方法多是基于学习的,将共指消解问题看作是二元分类或聚类问题,对给定的名词短语通过学习利用训练后的分类器决定是否为共指对。

汉语指代消解研究起步较晚。由于缺乏形态和拼写方面的暗示,汉语指代消解与英语等语种同类研究相比尚存在较大差距。目前的研究大体上分为两种:或局限于具体语言处理系统中有限条件下的代词处理,这难以从根本上解决真实语料中的指代问题;或从理论语言学方面探讨规律,距实用性和形式化要求仍有较大距离。早期的研究集中在针对汉语语法特点的规则方法与理论讨论。近期研究包括:文[3]提出句焦点的概念,用优先和过滤算法实现了语篇中元指代的消解,元指代消解算法建立在中心理论上。文[4]在篇章表述理论(Discourse Representation Theory, DRT)的基础上,提出了一种面向语篇理解的汉语人称代词的指代消解方法。DRT 独特的语篇表述结构(Discourse Representation Structure, DRS)的动态构造方法,为指代消解提供了新思路。文[5]提出在中文实体侦测与跟踪任务(EDT)上使用一种统一的基于转换学习框架(Transformation Based Learning, TBL)的方法。TBL 是一种广泛应用的机器学习方法,但首次运用于共指消解问题。王厚峰对中文指代消解问题进行了较为深入的研究,先后提出基于 HNC 句类基本知识,结合排除和局部焦点优先选择规则进行人称代词消解的方法;一种健壮性的弱化语言知识的(仅利用单复数、性别和语法角色特征)汉语人称代词消解算法,见诸文献较多,此处不再赘述。

3 常用机器学习方法

近几年,机器学习方法在指代消解中受到广泛关注,特别是针对小规模训练语料的无督导学习方法。

3.1 最大熵统计模型

最大熵模型(Maximum Entropy Models, ME)是在已有限制条件下,估计未知概率分布。形式化时将限制条件作为特征函数,熵最大目的即在这些特征函数的期望值等于观测值的限定下,取得具有最大熵的分布。最大熵模型不依赖语言模型,独立于特定任务,是对后验概率进行建模的成熟模型,比一般的统计模型更为灵活地使用各类非受限地文本特征,缺点是算法计算量大,需要对数据稀疏问题进行平滑处理。文[6]提出基于最大熵模型的英文名词短语指代消解,通过两个待消解项相应属性构成特征,多特征继续构建生成特征向量,最大熵模型根据特征向量得出其指代概率,完成消解。

3.2 督导分类方法

分类问题研究集中在表示问题和泛化问题上。前者关键是非线性问题在线性空间的表示;后者是给定样本集合,通过算法建模判定模型对问题世界为真的程度。决策树(DT)方法就是一种逼近离散值函数的归纳推理算法,用于有督导学习。习得的函数表示为一棵决策树,能学习析取表达式,对噪声数据健壮性好。决策树类似于流程图的树状结构,顶层结点为根结点,叶子结点对应一个类别标识,内部结点对应分割数据集的判定属性 X_i。每个内部结点都有一个分割判断规则 q_i。方法先利用归纳算法递归构造决策树;然后使用产生的规则对数据分类或分析。DT 方法源自概念学习系统 CLS,后发展为 ID3 方法,又演化为能处理连续属性的 C4.5、C5.0 方法。ID3 主要是利用信息增益作为属性选择标准建决策树。C4.5 在 ID3 基础上加入处理过度拟合问题的后剪枝算法,同时采用增益率为选择标准,C5.0 则在 C4.5 基础上加入 Boosting 的 ML 思想。文[2]首次应用 C4.5 方法进行特定领域的指代消解。因为语义特征的进一步挖掘,使得可用于决策分类的属性增加,决策树方法的应用空间仍然很大。

3.3 无督导聚类

聚类是一种无督导的机器学习方法,是把一堆文本片断划分成不同的小组,任一小组内的片断与小组内的其他片断的相似度要大于与其他小组间片断的相似度,得到的小组称为簇(Cluster)或类。常见的有划分聚类和层次聚类两种:前者是按某种划分准则,将包含 n 个文档的数据集划分成指定数目的 k 个簇;后者分为凝聚式和分裂式两种,凝聚法将每个文本视为一个簇,自底向上每次寻找最近的两个簇进行合并,直到所有文本合并为一个簇。分裂法将所有文本视为一簇,自顶向下每次选择最大的一个簇分裂为两个,直到每个文本自成一簇。1999 年,Cardie 等提出一种基于聚类的无督导方法[7],应用一个不一致性函数集和不同距离尺度的权重的办法重新考虑与领域无关的共指聚类问题。文[8]首次提出了利用无督导聚类方法处理中文名次短语共指消解。

3.4 多策略方法

面对语言信息处理中复杂的特征空间和庞大的数据集,无论哪一种方法都有其适用范围。当各个学习模型均有所偏置的时候,通过多策略将多个模型有机结合的思路是值得尝试的。基本上采取的方式有两类:一是针对某一特征空间应

(下转第 237 页)

表 2 图 3～5 中各种融合方法所得结果图像的相似性度量值

图 像	本文提出的方法	Haar 小波融合[1]	形态学小波融合[8]
图 3 (Disk)	0.89096 (T=0.023441)	0.8316	0.8152
图 4 (Lab)	0.88490 (T=0.068020)	0.8386	0.8430
图 5 (Pepsi)	0.92620 (T=0.019089)	0.8997	0.9003

结论 本文提出了一种基于区域检测的多焦点图像融合方法，并使用小波算法、形态学运算和 GA 提取原始图像中的聚焦区域，最后将在原始图像中检测到的聚焦区域融合成为结果图像。这种方法使用类似于剪切-粘贴的操作将原始图像中的聚焦区域组合在一起，得到与参考图像相近的各处聚焦的结果图像。大量实验结果证明本文的方法明显优于 Haar 小波融合方法和形态学小波融合方法。特别是在原始图像没有完全配准的情况下，本文方法具有明显优势。

参 考 文 献

1 覃征，鲍复民，等. 数字图像融合. 西安交通大学
2 Pajares G, de la Cruz J M. A wavelet-based image fusion tutorial. Pattern Recognition, 2004. 1855～1872
3 Piella G. A region-based multiresolution image fusion algorithm. In: ISIF Fusion 2002 Conference
4 Wen C Y, Chen J K. Multi-resolution image fusion technique and its application to forensic science. Forensic Science International, 2004, 140: 217～232
5 Li Min, Cai Wei, Tan Zheng. A region-based multi-sensor image fusion scheme using pulse-coupled neural network. Pattern Recognition Letters, 2006, 27: 1948～1956
6 Jiang Zhiguo, Han Dongbing, Chen Jin, et al. A wavelet based algorithm for multi-focus micro-image fusion. In: Proc. eedings of the Third International Conference on Image and Graphics, 2004
7 Yang X, Yang W, Pei J. Different focus points images fusion based on wavelet decomposition. In: Preceeding of Third International Conference on Information Fusion, vol 1. 2000. 3～8
8 De I, Chanda B. A simple and efficient algorithm for multifocus image fusion using morphological wavelets. Signal Processing, 2006, 86: 924～936
9 De I, Chanda B. Enhancing effective depth-of-field by image fusion using mathematical morphology. Image and Vision Computing, 2006, 24: 1278～1287
10 Li Shutao, Kwok J T, Wang Yaonan. Combination of images with diverse focuses using the spatial frequency. Information Fusion, 2001, 2: 169～176
11 Ardeshir A. Goshtasby: Fusion of multi-exposure images. Image and Vision Computing, 2005, 23: 611～618

(上接第 186 页)

从图 2 和图 3 的结果，可以得到以下结论：

①本算法与 KNN 算法相比，在相同条件下，测试集的正确率有明显提高。

②算法的性能随着训练样本数的增加而有所提高，并且，与近邻数 k 的选择也有直接的关系，k 的取值越大，正确率越高；反之，取值越小，正确率越低。当然，随着 k 的取值逐渐增大，算法的计算量也越来越大。下一步，我们将考虑用决策树来降低算法的计算量。

③根据具体应用问题的不同，可以适当调整待分类变量的结构特征在分类过程中所占的权重，使分类结果更加合理和优化。

结论 本文首先讨论了 KNN 和 BN 算法的基本思想及其改进算法，然后分析了目前两个算法存在的主要问题以及各自的优点，最后提出了基于贝叶斯网络结构学习的 KNN 算法 (BN-KNN)。实验结果表明，在相同条件下，与 KNN 算法相比，BN-KNN 算法具有更好的分类正确率。而且，随着与分类变量结果弱相关变量的增多，新算法的优势将越明显。另外，新算法可以根据用户所关注的分类目标自动地调整在分类过程中所占的权重，使分类结果更加合理和优化。

参 考 文 献

1 Teknomo K. What is K Nearest Neighbors Algorithm? [Z] http://people.revoledu.com/kardi/tutorial/KNN/Contents.htm
2 陈振洲，李磊，姚正安. 基于 SVM 的特征加权 KNN 算法[J]. 中山大学学报(自然科学版)，2005，44(1)：17～20
3 D'Amato C, Malerba D, Esposito F, et al. Extending the K-Nearest Neighbour classification algorithm to symbolic objects [C]. Convegno Scientifico Intermedio SIS, 9-11 Giugno 2003, Università degli Studi di Napoli "Federico II"
4 Vincent P, Bengio Y. K-Local Hyperplane and Convex Distance Nearest Neighbor Algorithms [R]: [Technical Report]. http://www.iro.umontreal.ca/～lisa/pointeurs/TR1197.pdf, 2001
5 Cooper G F, Herskovits E. A Bayesian Method for the Induction of Probabilistic Networks from Data [J]. Machine Learning, 1992 (9): 309～347
6 Chen R, Herskovits E H. Network analysis of mild cognitive impairment [J]. NeuroImage, 2006, 29: 1252～1259
7 Yager R R. An extension of the naïve Bayesian classifier [J]. Information Sciences, 2006, 176: 577～588
8 Yang T Y, Lee J C. Bayesian nearest-neighbor analysis via record value statistics and nonhomogeneous spatial Poisson processes [J]. Computational Statistics & Data Analysis, 2006
9 Frey B J, Jojic N. A Comparison of Algorithms for Inference and Learning in Probabilistic Graphical Models [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(9): 1392～1416

(上接第 215 页)

用不同的机器学习方法；二是选择不同语言处理粒度角度(字、词、语块、篇章小句)进行学习。目的都是有效消除单一方法的不足，尽可能增强整体处理的健壮性和可移植性。

总结与展望 指代消解是一项困难的任务。迄今为止，尚未有较好的全自动的指代消解技术和方法。可行的思路是由单一的基于规则方法向结合机器学习与统计方法的多策略技术方向发展，特别是基于小规模语料训练的无督导机器学习方法。借鉴国外的成功经验，将研究内容更好地同国际评测(如 ACE)结合是促进汉语指代消解长足发展的一条可行之策；国内几乎没有用于研究目的开放的汉语指代消解数据集，这方面可开展的研究工作也很多。目前指代消解在计算内容上逐渐由基于领域受限文本向领域无关自由文本方向发展；处理对象由普通文本向 Web 文本发展，因此笔者认为在汉语指代现象的形式化研究方面(如代词所指、省略)，尽管已取得了一定成果，仍无法满足计算机自动化机处理，进一步挖掘和设计出对指代消解更有效的属性特征也是关键的突破口。

参 考 文 献

1 Lappin S, Leass H. An algorithm for pronominal anaphora resolution. Computational Linguistics, 1994, 20(4): 535～561
2 MaCarthy J F, Lehnert W G. Using decision trees for coreference resolution. In: Proceedings of 14th International Joint Conference on Artificial Intelligence Montreal, 1995. 1050～1055.
3 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
4 王晓斌，周昌乐. 基于篇章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
5 Ya-qian Z, Chang-ning H, et al. Transformation Based Chinese Entity Detection and Tracking. In: Proc. of the Second International Joint Conference on Natural Language Processing, 2005
6 钱伟，郭以昆，周雅倩，等. 基于最大熵模型的英文名词短语指代消解. 计算机研究与发展，2003，40(9)：1337～1343
7 Claire C, Wagstaff K. Noun phrase coreference as clustering. In: Proc. of the 1999 Joint SIGDAT Conference on Empirical Methods in NLP and Very Large Corpora
8 Chi-shing W. A Clustering Approach for Unsupervised Chinese Coreference Resolution. In: Proc. eedings of the 5th SIGHAN Workshop on Chinese Language Processing Sydney, 2006. 40～47