第35卷 第15期 计 算 机 工 程 2009年8月
Vol.35 No.15 Computer Engineering August 2009
·人工智能及识别技术· 文章编号：1000—3428(2009)15—0165—03 文献标识码：A 中图分类号：TP311

基于树核函数的代词指代消解

王海东[1,2]，谭魏璇[1,2]，李艳翠[1,2]，周国栋[1,2]
(1. 苏州大学计算机科学与技术学院，苏州 215006；
2. 江苏省计算机信息处理技术重点实验室，苏州 215006)

摘 要：提出一种基于树核的英文代词消解方法。针对结构化信息在指代消解中的重要作用，使用SVM提供的卷积树核函数自动获取句法结构信息，将句法树作为一个特征与其他基本特征结合。通过应用不同的剪枝策略，考虑不同句法树对系统的影响，在原有的句法树上扩充一些语义节点。在ACE2004 NWIRE基准数据上进行实验的结果证明，该方法对代词的消解起到明显的作用，综合值*f*提高了11.9%。
关键词：指代消解；句法结构；树核函数；修剪策略

Tree Kernel Function-based Pronoun Coreference Resolution

WANG Hai-dong[1,2], TAN Wei-xuan[1,2], LI Yan-cui[1,2], ZHOU Guo-dong[1,2]
(1. School of Computer Science & Technology, Soochow University, Suzhou 215006;
2. Jiangsu Province Key Lab for Computer Information Processing Technology, Suzhou 215006)

【Abstract】This paper proposes a tree kernel-based approach to anaphora resolution of pronoun. On the basis of structured information automatically captured by convolve kernel of SVM, it integrates syntax tree as a feature with other base features. Different pruning strategies are applied to eliminate the impact of syntax trees to the results. Evaluation on the ACE2004 NWIRE benchmark corpus shows that tree kernel can improve the *f* performance by 11.9%. Based on the system, it combines with semantic role feature and verb-driving feature which are acquired from ASSERT.
【Key words】coreference resolution; syntax structure; tree kernel function; pruning strategy

1 概述

指代消解是自然语言处理的关键问题之一，是篇章处理不可缺少的内容。在指代消解领域，早期的研究方法侧重于理论探索，运用大量手工构建的语言甚至领域知识。近十年来，由于自然语言处理技术的迅速发展和对指代消解技术的迫切应用需求，人们逐渐转向基于有监督的机器学习方法，并取得了一定的进展。文献[1]采用基于决策树的机器学习方法，选取了12个特征，并给出了完整步骤和实现平台。该方法已成为国内外利用机器学习方法进行指代消解研究的基础。以后的研究大多集中在如何选取更有效的特征上。文献[2]提取了53个特征，涵盖了语义词汇等各个方面，取得了很好的效果。

随着指代消解研究的不断深入，受制于弱语言知识，自动指代消解技术近年来在性能上难以继续提高，研究人员于是把焦点转向了基于自动产生的深层语言知识，特别是结构化句法信息，以期望取得性能上的突破。从图1看，The company和it有指代关系，而这种结构是一种固定的模式，通常从句的宾语如果是代词且又和主句的主语类型一致，那么它们之间常存在指代关系。

<s> ***The company*** said ***it*** will try to refocus its business on software licensing arrangements made with partners, including cable service providers and makers of cable set-top boxes . </s>

图1 来自ACE2004的例子

如图2所示，这种指代关系可以通过句法结构表达。传统的研究主要集中在如何把结构化信息转化为一般的特征交由学习器训练，主要是通过句法结构、语义角色以及中心理论将部分句法信息提取到指代消解系统中。这种方法受制于手工提取的规则。目前对于如何直接使用句法结构信息的研究相对较少。

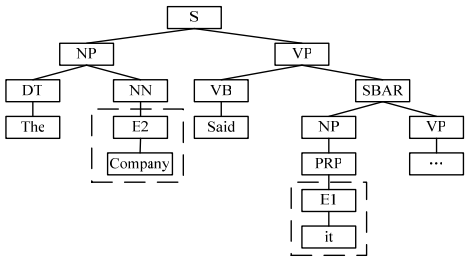

图2 图1中对应句子的部分句法树

文献[3]尝试使用树核对代词进行消解，取得了一定的效果。本文在其基础上探讨不同的剪枝策略以及将一些语义节点加到树中，使树中包含更多的语义信息。实验表明合适的剪枝策略及加入语义节点可以明显提高系统的性能。

2 相关工作

早期的研究表明，结构化信息对于指代消解非常重要。文献[4]使用句法树进行代词的指代消解。该算法首先为文档中的每个句子建立完全解析树，然后采用从左到右广度优先

基金项目：国家自然科学基金资助项目(60673041)；国家“863”计划基金资助项目(2006AA01Z147)
作者简介：王海东(1981－)，男，硕士研究生，主研方向：自然语言处理；谭魏璇、李艳翠，硕士研究生；周国栋，教授、博士生导师
收稿日期：2009-01-05 **E-mail：**064227065056@suda.edu.cn

的搜索方法遍历完全解析树，最后根据语法结构中的支配和绑定关系选择合法的名词短语作为先行语。文献[5]提出一种RAP 算法，使用 McCord 提出的槽文法(slot grammar)获得文档的句法结构，并通过手工加权的各种语言特征计算各先行语候选的突显性，利用过滤规则确定先行语，实现句内和句间第 3 人称代词和反身代词的消解。

树核函数在自然语言处理的各个领域得到了广泛应用。文献[6]使用树核实现了关系抽取。文献[7]应用卷积核函数实现关系抽取，并且提出了最短路径包含树(SPT)。MCT 表示的是一个句子的语法树，SPT 表示根据 2 个词(E1, E2)抽取出来的最短路径包含树。文献[3]使用卷积树核函数实现了指代消解，并分析了 3 种不同的裁剪策略对指代消解性能的影响。文献[3]的研究是对树核用于指代消解的一个尝试，只提供了 3 种简单的裁剪树的方式，并没有深入探讨其他的裁剪方法。

本文在文献[3]的基础上进行扩展，使用文献[8]的句法分析器得到句子的句法树，并使用 Assert 工具得到语义角色信息用于扩充句法树。将句法树直接作为一个特征并与其他的基础特征一起交由 SVM 训练。如果对全文构建句法树，系统的开销较大，跨句中的固定指代模式也不明显，而且存在的一些跨句指代模式可以通过语义特征以及中心理论得到很好的弥补。因此，本文只对一句之内的照应语和先行词构建句法树，并且只对一句之内照应语是代词进行消解。与文献[3]的系统相比，本文综合分析了不同裁剪方式对指代消解的影响，并在原始句法树的基础上，将一些语义信息作为新的节点加入树中。在 ACE2004 NWIRE 语料上的测试表明，利用句法信息可以大幅度提高系统的性能，与 Soon 的原型系统相比，F 值提高了 6.0%。

3 卷积树核函数

本文使用 SVM 作为分类器实现代词的消解。为了将结构化信息引入指代消解，使用卷积树核函数计算 2 棵树的相似度，从而得到固定的指代模式，明确照应语和先行词的指代关系。

卷积核函数通过计算 2 棵解析树之间相同子树的数量比较解析树之间的相似度。例如，2 棵解析树 T_1 和 T_2 要计算相似度 $K_c(T_1,T_2)$：

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

其中，N_j 是 T_j 的节点集合；$\Delta(n_1,n_2)$ 计算以 n_1 和 n_2 为根的共同子树个数，可以采用下面的递归计算方法：

(1)如果 n_1 和 n_2 节点处的产生式不同，则 $\Delta(n_1,n_2)=0$，否则转向(2)。

(2)如果 n_1 和 n_2 都是叶子前的一个节点，则 $\Delta(n_1,n_2)=1\times\lambda$，否则转向(3)。

(3)递归地计算 $\Delta(n_1,n_2)$：

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

其中，$\#ch(n_1)$是节点 n 的孩子节点数目；$ch(n,k)$是节点 n 的第 k 个孩子节点；$\lambda(0<\lambda<1)$是衰退因子。

4 语义关系树的裁剪和扩充

包含照应语和先行词的句法树能提供丰富的信息用于指代消解。更好地利用句法树信息的方法如下：如果使用一棵完整的句法树，那么系统开销太大，并且里面包含太多噪音，从而很难发现固定的指代模式；如果树的节点被裁剪掉太多，那么一些有用的信息可能会被裁掉。

4.1 树的裁剪

本文是对一个句子做句法分析，且仅考虑照应语和先行语在一个句子内的情况。为了明确句法树中照应语和先行语，引入 2 个节点 E1 和 E2，E1 表示照应语，E2 表示先行语，如图 1 所示。利用文献[7]使用的术语，将一棵完整的句法树标记为 MCT。使用完全句法树虽然可以得到更多的信息，但同时也引入了很多噪声，卷积树核函数是通过相同子树计算 2 棵树的相似度，因此，2 棵树越大，它们的相似性越小，于是考虑公共节点树 CT，即 2 个待消解词的公共节点下的所有节点构成的树，如图 3 所示。

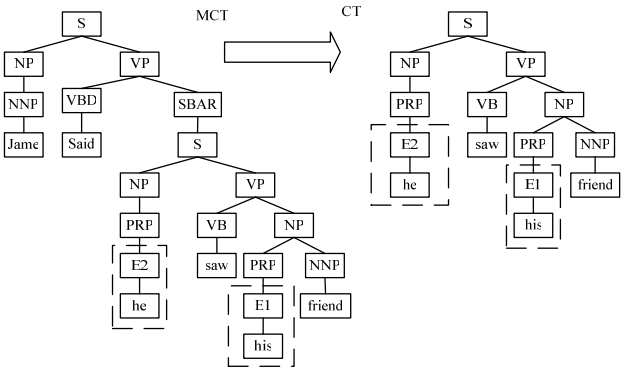

图 3 MCT 树和 CT 树

在公共节点树 CT 的基础上，根据文献[7]的剪枝策略，继续得到最短路径包含树，该树只保留照应语和先行语之间的节点，裁剪策略及最终的树如图 4 所示。最后考虑一种最小树(MT)，就是只保留照应语和先行语的直接祖先节点，到它们的公共节点为止。

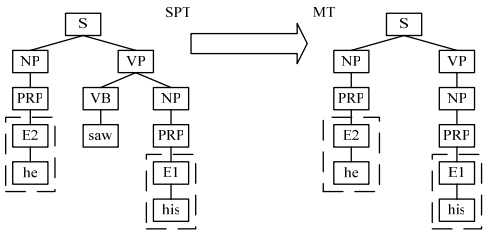

图 4 SPT 树和 MT 树

本文只考虑上述 4 种树，MCT 是一棵完全树，信息最全，噪音最多；CT 在 MCT 的基础上，只保留从公共节点往下的所有节点，虽然去掉了很多噪音，但产生的树依然很大。在 CT 的基础上本文又得出了 SPT，该树包含很少的噪音，但同时很多有用的信息也被裁剪掉。在 SPT 的基础上进一步得到了 MT，这棵树相应的节点最少，噪音也最少，但同时去掉的结构信息也最多。

4.2 树的扩充

本文对树进行了适当的扩充。如果一个实例中照应语和先行词的距离很长或者树结构很复杂，即照应语和先行词之间的特征不明显，则分类器分错的概率会大大增加，因此，应尽量使照应语和先行词的关系明显化。上文已经通过加入 E1 和 E2 来标注照应语和先行词，现在考虑在对应的节点上加入反映当前词的特征，从而使照应语和先行词在树中更凸显，因此，尝试将当前词的语义角色信息和词的类别加入。如图 5 所示，语义角色特征分为 4 类，Role 表示当前词在句子中所作的语义角色，其中，ARG0 表示当前词在句子中作为主语；ARG1 表示当前词在句子中作为宾语；Arg01 表示

当前词在句子中作某个动词的主语同时又作为另一个动词的宾语；NoArg 表示当前词在句子中不作主宾语。Class 表示当前词的语义类别，共有 6 个类别：基本名词，代名词，专有名词，有定描述，非限定词以及指示性名词，分别表示为 BareNp, PronounNp, ProperNp, DefiniteNp, IndefiniteNp, DemonstrativeNp。

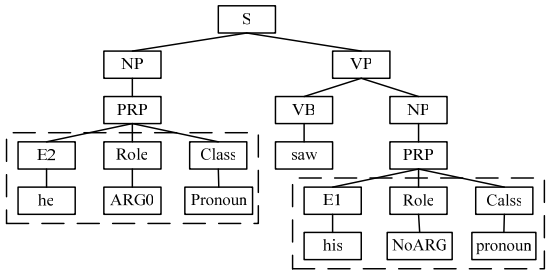

图 5　句法树的扩充

5　实验结果与分析

实验数据采用 ACE2004 NWIRE 基准语料，选取其中 75 篇作为训练语料，21 篇作为测试语料。语料中原文本已通过 Charniak's parser 得到句法树，所以，本文直接使用 Charniak 的句法分析结果。经过预处理的适当过滤，本文统计了测试集中需要消解的代词个数，共有 163 个。

5.1　代词的消解

下面的实验针对一句之内的代词，模型也是通过一句之内的实例训练而得。从表 1 可见，与 Soon 的原型系统相比，使用树核可以明显提高系统的召回率和准确率。在选用 SPT 的情况下，召回率提高了 23.9%，准确率有所下降。实验数据说明，使用 SPT 树消除了多余的噪音，同时也最大程度地保留了结构化信息。而 MT 树由于节点更少，因此不能更好地凸显那些有指代关系的模式，召回率没有 SPT 好，但因为 MT 树节点少，更易准确地判断 2 棵树的相似度，所以准确率比 SPT 好，在总体性能上 MT 树好于 SPT 树。使用 MCT 和 CT 系统的召回率有所提升，但准确率下降过多。从实验数据看，使用树核能对代词的消解起到很大的作用，特别是对代词的召回率，在 SPT 树下，系统多找出 39 个指代关系。此外，适当修剪树能使系统的性能得到很大的提升。

表 1　代词消解的结果　(%)

	召回率	准确率	综合值 f
原型系统 (Duplicated Soon)	58.9	78.7	67.4
MCT	70.6	52.8	60.5
CT	76.1	60.2	67.2
SPT	82.8	67.5	74.4
MT	80.9	70.6	75.4

5.2　语义信息的引入

由以上的数据和分析可见，单纯引入树核虽然能够提高系统的召回率，但同时也引入了很多噪音，将一些本没有指代关系的实例识别为有指代关系。本文考虑在句法树上引入一些语义特征，将这些语义特征加到句法树上，以便更有效地提升系统的准确率。

从表 2 看，在句法树上加入语义特征能适当提高系统的召回率和准确率，当未使用扩充语法树时，使用 SPT 树系统的召回率较好；当使用扩充语法树时，MT 树得到更好的召回率。对语法树分析可知，在未使用扩充语法树时，由于 MT 裁剪掉过多的节点从而使很多的有用信息丢失，因此这时 SPT 的召回率较好。当引入扩充语法树后，保留了 MT 树裁剪掉的信息，从而使 MT 树的召回率得到提升。并且由于有扩充语义树，提升了树核的区分能力，因此准确率也得到了提升。通过扩充语义节点，在 MT 树下，系统取得了最好的性能，系统的召回率比不引入语法树提高了 3.8%，准确率提高了 4.0%。

表 2　加入语义节点后对代词消解的结果　(%)

	召回率	准确率	综合值 f
原型系统 (Duplicated Soon)	58.9	78.7	67.4
MCT	70.6	54.2	61.3
CT	77.9	62.9	69.6
SPT	82.8	71.8	76.9
MT	84.7	74.6	79.3

6　结束语

本文使用树核函数，考虑了几种不同裁剪方法并选择适当的模型。实验表明，使用树核函数能有效解决代词的消解。在原有的句法树基础上，本文还尝试在句法树上加入一些语义节点，从而更好地区分不同的句法树，使系统的准确率和召回率得到进一步的提高。下一步的工作是考虑如何在现有的句法树基础上裁剪掉冗余信息，并考虑如何调整训练实例，将更具区分度的训练实例交由 SVM 训练。同时考虑将更多的语义特征加到句法树上，以进一步提高系统的性能。

参考文献

[1] Wee Meng Soon, Hwee Tou Ng, Lim Chung Yong. A Machine Learning Approach to Coreference Resolution of Noun Phrase[J]. Computational Linguistics, 2001, 27(4): 521-544.

[2] Vincent N, Claire C. Improving Machine Learning Approaches to Coreference Resolution[C]//Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. [S. l.]: IEEE Press, 2002.

[3] Yang Xiaofeng, Su Jian , Tan Chewlim. Kernel-based Pronoun Resolution with Structured Syntactic Knowledge[C]//Proc. of ACL'06. Sydney, Australia: [s. n.], 2006.

[4] Hobbs J. Resolving Pronoun References[J]. Lingua, 1978, 44(2): 339-352.

[5] Lappin S, Leass H. An Algorithm for Pronominal Anaphora Resolution[J]. Computational Linguistics, 1994, 20(4): 525-561.

[6] Zelenko D, Aone C, Richardella A. Kernel Methods for Relation Extraction[C]//Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing. Morristown, NJ, USA: Association for Computational Linguistics, 2002.

[7] Zhang Min, Zhang Jie, Su Jian, et al. A Composite Kernel to Extract Relations Between Entities with Both Flat and Structured Features[C]//Proc. of ACL'06. Sydney, Australia: [s. n.], 2006.

[8] Charniak E. A Maximum-entropy-inspired Paser[C]//Proceedings of North American Chapter of the Association for Computational Linguistics Annual Meeting. San Francisco, USA: [s. n.], 2000: 132-139.

编辑　张正兴