基于树核函数的实体关系抽取方法研究*

庄成龙　钱龙华 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）

064227065088@suda.edu.cn

摘要： 实体关系抽取是信息抽取中的一个关键任务，其目的是找出文本中实体对之间的语义关系。本文描述了一种改进的基于树核函数的实体关系抽取方法，在路径包含树的基础上，加入了与实体相关的语义信息，并对原有的树进行裁剪，消除一些冗余结构。在 ACE 2004 语料上进行实验，性能有了明显的提高，F 值达到了 71.9%。

关键字： 实体关系抽取，树核函数，路径包含树，裁剪， 语义信息

Tree Kernel-based Relation Extraction

Zhuang Chenglong, Qian Longhua, Zhou Guodong

(School of Computer Science & Technology, Soochow University, Suzhou, China, 215006)

(Jiangsu Provincial Key Lab for Computer Information Processing Technology, Suzhou, China, 215006)

064227065088@suda.edu.cn

Abstract: Entity relation extraction plays an important role in Information Extraction. It aims to determine the semantic relationship between pairs of entities from text. This paper presents an improved tree kernel-based approach on entity relation extraction. by incorporating semantic information and pruning out unnecessary structures on the basis of path-enclosed parse tree. Evaluation on the ACE 2003 corpus shows that our system much improves the performance and achieves the F-measure at 71.9%.

Keywords: Entity Relation Extraction, Tree kernel, Pruning, Semantic Information, Path-enclosed Parse Tree

1 引言

进入二十一世纪以来，随着科技的不断进步，尤其是互联网技术的快速发展，现实世界中的信息量迅猛增加，远远超出了人类阅读的能力。如何过滤无用信息并从中抽取出人们所需要的特定信息成为一个难点。信息抽取的主要目的是从无结构化的文本中抽取特定的事件、事实等信息转化为结构化或半结构化的信息，并且储存在数据库中，供查询以及进一步分析利用。信息抽取起初是在由美国国防高级研究计划局(DARPA)资助的信息理解会议（MUC，1987-1998）中提出，并使之发展成为自然语言处理(NLP)领域的一个重要分支。MUC 会议停止后，由美国国家标准技术局(NIST) 资助的“自动内容抽取”(ACE)评测会议，进一步推动着信息抽取研究的发展。

ACE 中信息抽取的任务实体识别和跟踪(EDT，Entity Detection and Tracking)、关系识别

* 本文受“863”国家高技术研究发展计划资助项目(2006AA01Z147)和国家自然科学基金资助项目（60673041）资助。

作者简介：庄成龙，(1985-)，男，硕士研究生，主要研究方向：信息抽取。钱龙华，男，博士研究生，主要研究方向：自然语言处理。周国栋，(1967-)，男，博士生导师，主要研究方向：自然语言处理。

和描述(RDC，Relation Detection and Characterization)以及事件识别和描述(EDC，Event Detection and Characterization)等，本文的研究重点是实体关系的识别与抽取。实体关系的抽取在实际应用中的范围很广，对于信息抽取、问答系统、机器翻译等的发展都有重要作用。其目的是从标注好的文本中找到实体对之间的语义关系。例如“胡锦涛是中华人民共和国主席。”中包含了一种“雇佣”(EMP-ORG)关系，表示实体“胡锦涛”(PER)受雇于实体“中华人民共和国”(ORG)。在ACE中定义了七个大类关系，其中每个大类又分为若干子类。

本文中，我们使用卷积树核方法来进行关系抽取，通过比较两个实体关系对象的相同子树的个数来计算相似度。由于核方法可以充分利用特征方法无法表示的结构化信息，近年来越来越多研究人员开始研究和使用,例如：Zelenko(et al.,2003)[3]，Culotta (2004)[6]和Zhang (2006)[5]。我们在 Zhang(2006)[5]基础上提出了一种改进的树核方法，应用树的修剪策略，减少了冗余信息的同时扩充了原有的树结构，使之包含更丰富的语义信息。在ACE 2004基准语料上的测试表明，该方法能显著提高关系抽取系统的性能，实验结果与原型系统的相比有了明显的提高。

本文的后续组织结构如下：第二部分论述实体关系抽取的相关工作；第三部分介绍我们所使用的方法以及对树裁剪的策略。第四部分为实验结果和性能分析。最后为全文总结和将来工作的方向。

2 相关工作

现阶段最常用的实体语义关系抽取使用方法主要有三种：基于规则和知识库的方法，基于特征向量的机器学习方法，基于核函数的机器学习方法。

基于规则和知识库的方法（Miller et al. 2000)[13]对于特定的领域的抽取有较高的准确率，但其需要由特定领域专家构筑大规模的知识库，代价很大，可移植性方面存在明显的不足。

基于特征向量的学习方法首先需要构造符合特征向量形式的训练数据。然后使用各种机器学习算法，如支持向量机(SVM)、Winnow等作为学习器构造分类器。在关系抽取中，典型的基于特征向量的方法包括最大熵模型（MaxEnt）(Kambhatla 2004)[14]和支持向量机（SVM）(Zhao 2005[7]; Zhou 2005[8])。基于特征向量的关系抽取研究重点在于如何获取各种有效的词法、语法、语义等特征，并把它们有效地集成起来，产生描述对象的各种局部和简单的全局特征。例如，Zhao(2005)[7]集成了各种词、语法解析树和依存树特征。虽然近期的机器学习研究主要是基于特征向量的方法，但是由于实体间语义关系表达的复杂性和可变性，使得这种算法的信息获取比较单一，无法充分利用深层语法分析结果，停留在词频和符号的处理阶段，对大量训练数据的依赖也限制了提取的最终效果。

与基于特征向量的方法不同，基于核函数的方法不需要构造固有的特征向量空间,并且对大规模语料库的依赖程度低。核函数方法直接以结构树为处理对象，计算它们之间的相似度。采用直接计算两个特征向量甚至两个对象(如语法结构树)之间的相似度来进行分类,这使得基于核函数的方法理论上可探索隐含的高维特征空间。Zelenko (2003)[3]最早把核函数的方法引入了关系抽取领域。其首先在文本的浅层解析树的基础上定义了核函数，并设计了一个用于计算核函数的动态规划算法，然后通过支持向量机(SVM)和表决感知器(Voted Perceptron)等分类算法来抽取实体语义关系，在200篇来自新闻机构(如美联社、华尔街日报等)的新闻文章中进行测试，取得了较好的效果。Culotta (2004)[6]通过一些转换规则(如主语依存于谓语、形容词依存于它们所

修饰的名词等)将包含关系中两个实体的解析树转换成依存树，并在树节点上增加词性、实体类型、词组块、WordNet上位词等特征，然后定义了基于依存树的核函数并使用SVM分类器进行关系抽取，在ACE RDC 2003 基准数据上的5个关系大类的抽取中F指数取得了45.8。Zhang等(2006)[5]设计了一种复合卷积核函数来进行关系抽取。该方法将卷积核函数和线性核函数(与实体属性相关，如实体类型、引用类型等)结合起来，充分考虑了影响语义关系的平面特征和结构特征，在ACE2003和ACE2004基准数据上大类的关系抽取中F指数分别达到了70.9%和72.1%。

以上研究人员在使用核函数进行关系抽取方面进行了多种尝试，并取得了较好的性能，但是他们主要是从句法树结构上进行调整，忽略了文中实体相关的语义信息以及句法树内部冗余结构信息，性能的提高也遇到了瓶颈。本文借鉴zhang（2006）[5]中提出的方法，在此基础上利用裁减策略对生成树重新进行改进，并且加入了一些语义信息，有效丰富了结构化信息，使得关系抽取系统的性能有明显提高。

3 基于核函数的关系抽取

在这部分讲述了本文所使用的卷积核函数及其相似度比较的原理，并且详述了一种新的句法分析树裁减策略，如何去除冗余结构以及如何加入实体语义信息的方法等。

3.1 卷积核函数

卷积核函数是通过计算两棵解析树之间的相同子树的数量来比较解析树之间的相似度。例如有两棵解析树 T1 和 T2，要计算相似度 $K_c(T_1, T2)$：

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

其中 N_j 是 T_j 的结点集合，$\Delta(n_1, n_2)$ 计算以 n_1 和 n_2 为根的共同子树个数，可以按照下面这种递归的计算方法：

(1)如果 n1 和 n2 结点处的产生式不同，则$\Delta(n_1, n_2)= 0$，否则转向(2)；

(2)如果 n1 和 n2 都是叶子前的一个结点，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(n1)是结点 n 的孩子结点数目，ch(n,k)是结点 n 的第 k 个孩子结点，λ ($0<\lambda<1$)是衰退因子。卷积核函数计算的时间复杂度为 $O(|N_1|, |N_2|)$。

3.2 关系实例的生成

每个实体关系实例都是以分析树的结构存在的，不同的树结构对相似度比较的效果差别很大，影响关系抽取的效果。在Zhang(2006)[5]中提出了一种路径包含树(PT，Path_enclosed Tree)结构，并取得了较好的效果。这种树是从关系的两个实体的最近公共父结点作为根，并裁剪掉第一个实体左边和第二个实体右边的所有结点后的树。其树结构如图3-1所示。

可是由于许多关系实体对之间的路径相距较远或者生成树结构比较复杂，即使按照路径包

含树来裁剪，还是有许多冗余信息，并且此方法裁剪过程中把一些上下文语义信息也去掉了，这些都在一定程度上降低了关系抽取的效果。那么如何避免冗余并且提高系统性能呢？本文主要通过以下途径实现：

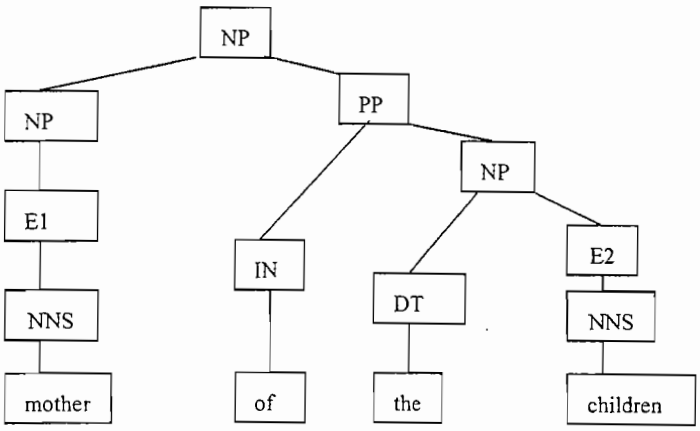

图 3-1 路径包含树

首先，增加实体相关的特征。通常，实体的语义关系与实体的语义属性密切相关，例如PER-SOC关系描述了人们个体之间的社会关系，而关系的两个实体必须是Person。基于特征的方法中，实体本身的属性或者是属性组合是构造向量的一个重要特征，例如实体大类类型、引用类型等。在实验中我们也有针对性地加入这些语义信息，来生成一种语义信息扩展树SEPT（Semantic_Ext-ended PT）。

其次，消除结构冗余。通过对错误分类的关系实例的观察，我们发现很多关系无法正确识别的原因是实体关系中存在冗余结构，如修饰语结构和并列结构等。

修饰语冗余是指两个实体之间存在修饰语如冠词、形容词、介词结构等， 这些修饰词在生成的PT树结构中作为噪声影响了分类器的分类。例如：实体“one of students”和“one of the red ball”生成的树结构在比较相似度的时候就可能会被认为不相似，如图 3-2 所示。

并列冗余是指句中的并列结构冗余。例如“Presidents(E1) of China(E2),Russia(E3) and America(E4)...”中判别出关系（E1,E2）表示雇佣关系，而由于E2、E3、E4位并列结构，所以我们可以认为它们与E1也存在雇佣关系，可事实上分类器识别不出（E1,E4）的关系，这种情况我们把它叫做并列冗余。通过去除并列关系的其他干扰词可以很好的解决这种情况，有效的提高分类的准确率。例如对（E1,E4）关系识别的时候，把原文转变成“Presidents(E1) of Am-erica(E4)”，这样分类器就可以正确的识别出它们之间的关系。这种裁剪树叫去并列树（PRPT，Parataxis_Removed PT）。

另外我们通过实验发现，数据中有许多上下文敏感的关系对识别效果也有一定的影响。例如通过实验分析我们发现实验数据中有许多所有格成分的关系无法识别。例如“member of masseur's supreme”中实体“member”和“ supreme” 有着“雇佣组织(EMP-ORG)”关系，但后者被所有格结构“m-issouri's” 修饰，实体“member”和实体“missouri”之间不存在关系。如果按照路径包含树(PT)的生成方式，抽取出实体“member”和实体“missouri”之间的路径包含树，生成的实例就是“member of missouri”，从而把它们分成有关系的。这种特殊结构，我们在实例生成时候就在第二个实体后增加所有格结构的标志词。这种方法生成树我们叫做所有格

树(PPT, Possessive PT)。

通过以上几种方法对关系实例添加语义信息并进行相应裁剪后，我们生成好了关系实例。接下来就是通过实验来验证该方法的有效性。

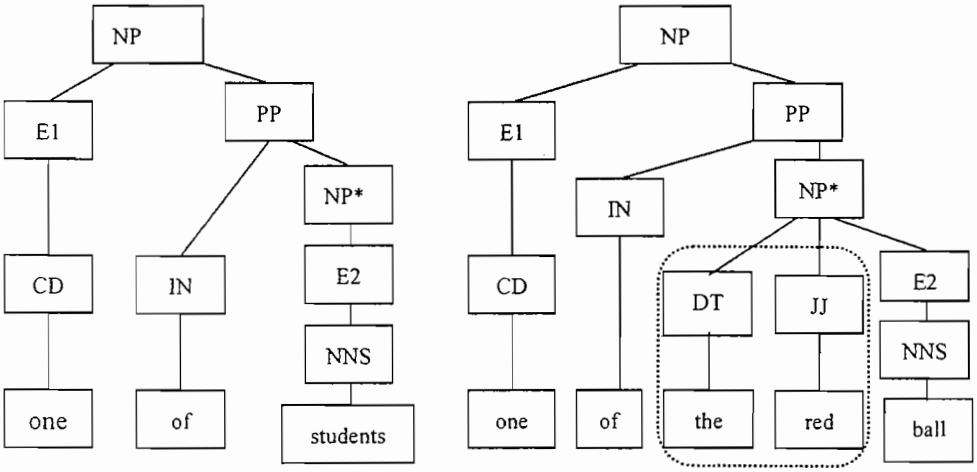

图 3-2 修饰词冗余

4 实验结果与分析

我们的实验使用卷积树核函数和 SVM 分类器，因为卷积树核能够有效的利用结构化信息，而 SVM 的二元分类性能比较好。数据选用 ACE2004 语料，选取其中 347 篇(BNEWS/NWIRE)，共有 4 307 个关系实例，对 ACE RDC 2004 所定义的七个大类进行关系识别和抽取。以下是前文提出的不同方法对系统性能的影响。

4.1 语义信息对关系抽取的影响

本文在 Zhang(2006)[5]的最短路径树的基础上加入各种语义信息，包括：实体顺序、大类类型、子类类型、引用类型等。表 4.1 列出了加入这些信息对实体关系抽取的影响效果。实验结果表明实体大类类型对关系抽取的贡献最明显，比原系统提高了大约 11%，这说明语义关系受到实体类型的限制。子类型和引用类型在此基础上也有不同程度的提高，表明子类型和引用类型对关系抽取性能提高也有较大作用。可是中心词和 LDC 引用类型不但没有提高，反而使得系统性能有所降低。这是由于语料中实体中心词都是指代消解后的词语，而不是原文本中的实体词，降低了正确识别概率，从而造成实验性能下降。

#	实体属性	P	R	F	#	Entity Info	P	R	F
1	原系统	66.7	50.3	57.4	5	+entity class(-)	78.9	62.7	69.9
2	+entity type*	75.7	61.4	67.6	6	+GPE role(-)	78.9	62.4	69.7
3	+entity subtype*	77.6	62.4	69.1	7	+head word(-)	80.8	60.3	69.1
4	+mention level*	79.1	63.6	70.5	8	+LDC type(-)	60.4	60.3	68.9

表 4.1 不同种类的实体相关语义信息在 ACE2004 上的性能比较（*表示该属性是起正作用，-起负作用）

4.2 去除冗余结构后对系统性能的贡献

表 4.2 列出了不同的关系实例生成树结构在 ACE 2004 的七个大类的测试结果，与未处理的原型系统的性能的差别。通过比较我们发现信息扩展树的性能提高了约 3%，效果很明显。这是说明有针对性的加入实体语义信息对关系抽取有很大帮助。根据卷积树核函数的原理，比较树的相似度时，衰退因子的作用会使得层次越深对整体相似度的贡献越小，我们充分利用结构化特征，把这些信息加在根结点上，试验结果表明效果非常理想。而其他的三种树结构在此基础上也有所提高，但是效果不是很明显。这是因为一方面名词所有格等结构在语料库中的数量相对较少，因而对结果影响系数较小。另一方面随着生成树结构的深度不断增加，由于衰减因子的影响，即使对这些冗余信息进行裁剪，它们对最终效果的影响也变得很小，所以提高不明显。

实例结构	关系识别			7 大类上关系分类		
	准确率(P)	召回率(R)	F1 值(F1)	准确率(P)	召回率(R)	F1 值(F1)
原型系统	84.4	63.6	72.5	66.7	50.3	57.4
+SEPT	87.3	70.3	77.9	78.6	63.4	70.5
+PPT	87.6	71.2	78.6	78.4	65.5	71.4
+RRPT	87.8	71.2	78.6	78.9	65.5	71.6
+PRPT	86.8	72.6	79.1	78.6	66.9	71.9

表 4.2 四种不同树在 ACE RDC2004 数据上测试的性能比较

4.3 总体性能与其他同类系统的比较

在表 4.3 中把我们的实体关系抽取的结果和其他三种关系抽取系统进行了比较，在大类的抽取效果方面比 zhang（2006）[5]提高了 4%。性能提高很明显，这都得益于我们的改进策略能够更加有效的利用树的结构化信息。并且和基于特征的方法比较，比 Zhao(2005)[7]性能提高了 1.6%，与基于特征最好结果 Zhou(2005)[8]相差仅 0.9%。

系统	关系检测			大类关系分类		
	P	R	F1	P	R	F1
本系统(基于树核)	86.8	72.6	79.1	78.6	66.1	71.9
Zhang(2006)[5](基于树核)	-	-	-	74.1	62.4	67.7
Zhou(2005)[8](基于特征)	89.0	66.6	76.2	82.8	62.1	71.0
Zhao(2005)[7](基于特征)	-	-	-	69.2	70.5	70.3

表 4.3 与其他系统性能相比较(ACE RDC 2004)

5 总结

本文利用卷积树核函数方法进行实体关系抽取，通过一种全新的生成关系实例的策略，对原有的实体关系树进行裁剪并且加入特定语义信息，在 ACE RDC2004 语料上进行实验，有效的提高了原有系统关系抽取的性能。通过实验我们发现实体类型相关信息对关系抽取的影响很大。由于某些原因，造成了对并列结构和修饰词等情况产生的噪声信息裁剪以后效果不明显，这

为我们的下一步工作的研究提供了一个方向。通过实验可以看出核函数的机器学习方法和特征的方法相比虽然有些差距，但由于前者其特有的优越性，能够充分利用平面和结构化特征，因而有较大的提升空间和研究价值。我们的下一步工作将围绕这方面展开，寻找更好的方法，进一步提高关系抽取的性能。

参考文献

[1] MUC[EB/OL].http://www.itl.nist.gov/iaui/874.02/related_project/muc/,1987-1998.

[2] ACE[EB/OL].The Automatic Context Extraction Project. http://www.ldc.upen. edu /Project/ACE, 2002-2005.

[3] Zelenko D,Aone C,Richardella A.Kernel Methods for Relation Extraction[J]. Journal of Machine Learning Research,2003(2),1083-1106.

[4] Collins M,Duffy N.Convolution Kernels for Natural Language. NIPS ,2001.

[5] Zhang M,Zhang J,Su J, et al.A Composite Kernel to Extract Relations between Entities with both Flat and Structured Features[A].ACL[C],2006,825-832.

[6] Culotta A,Sorensen J.Dependency tree kernels for relation extraction[A].ACL[C], 2004,423-429.

[7] Zhao S B,Grishman R. Extracting relations with integrated information using kernel methods[A].ACL[C],2005, 419-426.

[8] Zhou G D, Su J, Zhang J, Zhang M. Exploring various knowledge in relation extraction[A]. ACL[C],2005, 427-434.

[9] 车万翔, 刘挺, 李生. 实体关系自动抽取[J]. 中文信息学报, 2005, 19(2), 1-6.

[10] 李保利, 陈玉忠, 俞士汶. 信息提取研究综述[J]. 计算机工程与应用. 2003, 39 (10) : 1-5.

[11] Moschitti A.A study on Convolution Kernels for Shallow Semantic Parsing[A].ACL[C],2004.

[12] Bunescu R. C. and Mooney R. J. 2005. A Shortest Path Dependency Kernel for Relation Extraction. EM NLP-2005.

[13] Miller S., Fox H., Ramshaw L. , et al. A novel use of statistical parsing to extract information from text. ANLP'2000[C].2000,226-233.

[14] Kambhatla N.Combining lexical,syntactic and semantic features with Maximum Entropy models for extracting relations[A]. ACL (poster),2004,178-181.