基于树核函数的中文语义关系抽取*

虞欢欢　陈九昌　钱龙华*　周国栋

1. 苏州大学计算机科学与技术学院，江苏，苏州，215006

2. 江苏省计算机信息处理技术重点实验室，江苏，苏州，215006

E-mail: 20074227065075@suda.edu.cn

摘　要：关系抽取是信息抽取的一个重要组成部分，本文提出了一个基于卷积树核函数的中文语义关系抽取方法，采用最短路径包含树作为关系实例的结构化信息，在ACE2005标准语料库上进行关系大类的抽取实验，最终的F值达到了52.8%，由此可见卷积树核方法对中文语义关系抽取而言是有效的。同时，针对基于树核函数方法具有训练和测试速度慢的问题，提出了一种基于实体间路径长度的优化方法，在不影响抽取性能的前提下，显著降低了学习时间。

关键词：语义关系抽取，树核函数，句法树

Research on Tree Kernel-Based Chinese Semantic Relation Extraction

Huanhuan Yu　Jiuchang Chen　Longhua Qian　Guodong Zhou

1.School of Computer Science & Ttechnology, Soochow University, Suzhou, Jiangsu, 215006

2.Jiangsu Key Lab. of Information Processing Technology, Suzhou, Jiangsu, 215006

E-mail: 20074227065075@suda.edu.cn

Abstract: Semantic relation extraction is an important research subtask in the area of information extraction. This paper proposes a convolution tree kernel-based method for Chinese semantic relation extraction, with Shortest Path-enclosed Tree as the structural representation for a relation instance. Evaluation on the ACE 2005 corpus shows that our method achieves a reasonable F-measure of 52.8 on the task of top-level relation extraction. Furthermore, in order to tackle the problem of slow training and test speed for tree kernel-based methods, we devise a path length-based optimization technique to significantly decrease the learning time without much loss of extraction performance.

Keywords: Semantic Relation Extraction, Tree Kernel Functions, Syntactic Tree.

1　引言

信息抽取（IE）是指从一个给定的文档集合中自动识别出预先设定的实体、关系和事件等类型信息，并将这些信息结构化存储的过程。比如说，我们可以从文档中识别出人名、地名、机构名、数字、货币、时间、日期等类别的实体，这类工作称之为命名实体识别；从文档中识别出实体之间或实体及其属性之间的关系，称之为实体关系抽取；从文档中识别出某个事件发生的时间、地点、事件的参与者、造成的后果等信息，称之为事件信息抽取。

实体关系抽取作为一个独立的任务首先在MUC-7（MUC1998）上被引入，在MUC会议和ACE评测的推动下，基于英文语料的实体关系抽取的技术水平有了很大的进步，一些技术也趋于成熟稳定，英文语义关系抽取的性能也达到了一个相对较好而且稳定的水平，在基于卷积树核

* 基金资助：国家863计划(2006AA01Z147)；国家自然科学基金(60673041，60873150)；国家教育部博士点基金(200802850006)；江苏省自然科学基金(BK2008160)；江苏省高校自然科学重大基础研究项目（08KJA520002）。

* 通讯作者：qianlonghua@suda.edu.cn

函数关系抽取方面的F值达到77.1%。可是由于种种原因，中文语义关系抽取的研究在我国才刚刚起步，取得的性能还很低，远远落后于英文。因此尽快启动中文语义关系抽取的研究，改善抽取性能以缩小同英文的差距已经迫在眉睫。因此本文的研究放在中文语义关系抽取领域。

本文的后续组织结构如下：第二部分详细地介绍各种用于关系抽取任务的方法；第三部分论述了本文所使用的卷积树核方法以及如何构建最短路径包含树；第四部分为实验描述和结果分析；第五部分为如何优化分类器训练和测试速度；最后是全文总结和将来工作的方向。

2 相关工作

当前在信息抽取方面主要有两种途径：基于知识工程和基于统计的机器学习。由于基于知识工程需要专家构建大规模的知识库，既费时又费力，因此现在转向基于统计的机器学习方法。目前在关系抽取中所使用的机器学习方法一般分两类：基于特征向量的学习方法和基于核函数的学习方法。

典型的基于特征向量的方法包括最大熵模型（MaxEnt）和支持向量机（SVM），目前在中文方面，主要有车万翔[1]，董静[2]做了一些研究，但是都只是对部分作了研究。特征向量的方法尽管速度很快，也很有效，然而，由于实体间语义关系表达的复杂性和可变性，要进一步提高关系抽取的性能已经很困难了，因为很难再找出适合语义关系抽取的新的有效的词汇、句法或语义特征。

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)[4]，和Zhang (2006)[5]等。针对核函数方法中所存在的低召回率的问题，研究人员尝试利用卷积核函数来实现关系抽取。所谓卷积核函数，就是通过计算两个离散结构之间的相同子结构的数量来比较它们之间的相似度。Bunescu和Mooney（2005）[6]提出了基于字符串序列卷积核函数的关系抽取方法。Zhang等（2006）[7]则利用卷积树核函数（Collins和Duffy，2001）[8]来计算包含实体对的句法树之间的相似度，从而抽取语义关系。Zhou等（2007）[9]将语义关系实例表达为上下文相关的最短路径包含树。Qian等(2008)[10]利用成分依存理论来产生表示实体关系结构化信息的动态关系树，并同实体语义信息有机结合起来，在ACE RDC 2004语料库的7个大类的关系抽取中F指数达到了77.1。

在基于核的中文语义关系抽取方法中，Che（2005）[11]，刘克彬（2007）[12]，Huang（2008）[13]都进行了尝试，但是取得的性能都远远不如英文。当然这并不说明核方法本身有问题，而只能说明在中文语义关系抽取中较难找到能合理和确切表示语义关系的结构化信息。

总的来说，在中文语义关系抽取的研究中，一方面是研究人员采用的语料库及方法的可比性较差，难于判断方法本身的好坏；另一方面还在于如何根据中文语言的特点，进一步探索利用结构化信息的核方法和基于语义信息的特征方法，以期提高中文语义关系抽取的性能。基于卷积树核函数在英文领域的关系抽取中取得了较好的性能（F 77.1），因此，本文将探索卷积树核函数在中文领域的关系抽取中的有效性问题。

3 基于卷积树核函数的中文语义关系抽取

3.1 卷积树核函数

本文采用Collins和Duffy（2001）[8]的卷积树核函数，即两棵树之间的相似度可以通过计算它

们之间的相同子树的数目来实现，其公式为：

$$K_{CTK}(T_1,T_2)=\sum_{n_1\in N_1,n_2\in N_2}\Delta(n_1,n_2) \qquad \text{（公式一）}$$

其中N_1和N_2分别为T_1和T_2的结点集合，$\Delta(n_1,n_2)$用来计算以n_1和n_2为根结点的两棵子树之间的相似度，它可以通过下列递归的方法得出：

1) 如果n_1和n_2的产生式（采用上下文无关文法）不同，则$\Delta(n_1,n_2)=0$；否则转2；

2) 如果n_1和n_2是词性（POS）标记，则$\Delta(n_1,n_2)=1\times\lambda$；否则转3；

3) 递归计算 $\Delta(n_1,n_2)=\lambda\prod_{k=1}^{\#ch(n_1)}(1+\Delta(ch(n_1,k),ch(n_2,k)))$ 。 （公式二）

其中$\#ch(n)$是结点n的子结点数目，$ch(n,k)$是结点n的第k个子结点，而$\lambda(0<\lambda<1)$则是衰减因子，用来防止子树的相似度过度依赖于子树的大小。

3.2 结构化关系实例表示方法

在语义关系抽取中最先可用的结构化信息是最小完全句法树（Minimum Complete Tree，MCT），即在完全句法树中包含关系的两个实体且未作任何修改的最小部分，如图 1(左)。从直觉上判断，MCT 包含了丰富的结构化信息，有利于语义关系的抽取，但是对于关系的识别而言，所处理的对象不仅仅是关系正例，还有大量的关系反例，而关系反例中两个实体在句法树中的位置往往较远，使得它们的最小完全树相当复杂，导致卷积树核函数在计算两棵树之间的相似度时要消耗更多的时间。因此尽管原始的最小完全句法树（MCT）包含了丰富的结构化信息，由于其规模过于庞大，且包含了太多的与语义关系无关的噪音，并不适合于基于卷积树核函数的语义关系抽取。

为了寻找更合适的用于语义关系抽取的结构化信息，Zhang(2006)[5]中提出了五种句法树的扩展方法，其中最短路径包含树(SPT，Shortest Path Enclosed Tree，简称PT)结构取得的效果最好。PT是以两个实体的最近公共父结点为根，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所生成的树，如图 1（右）。因此在本文原型系统中的关系实例都是按照最短路径包含树（PT）结构进行裁剪的。

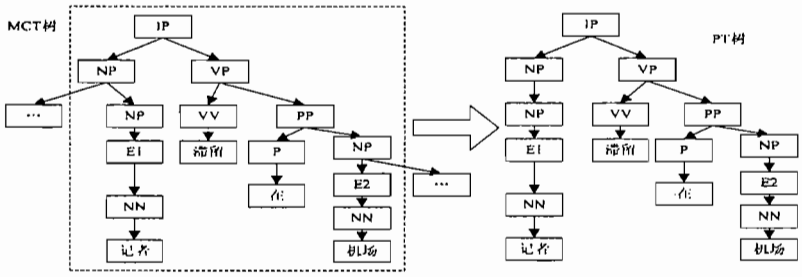

图 1 一个关系实例的最短路径包含树(PT)表示形式

其中 MCT 树和 PT 树是句子“…记者滞留在机场…”两个实体“记者”和“机场”之间的关系实例的两种不同表示形式。MCT 表示以两个实体的公共父节点为根节点并包含两个实体的最小完全树，PT 表示最短路径包含树。图中“E1”和“E2”表示前后两个实体。

4 实验描述及结果分析

4.1 中文语义关系抽取系统流程

基于树核函数的中文语义关系抽取系统的具体流程如图 2 所示：

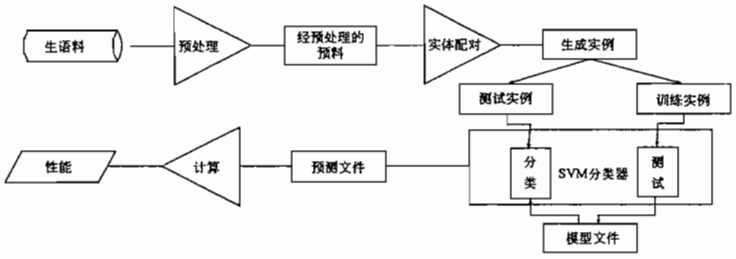

图 2 中文语义关系抽取系统流程图

本系统分为多个模块，每个模块相互独立并且衔接。系统的第一步工作就是对语料进行预处理，生成实验所需要的格式；第二步就是生成实验所需要的训练和测试数据；第三步训练出关系的模型；最后对测试数据进行预测，计算性能。

4.2 语料库的选取

本文实验语料采用 ACE2005 中文语料，它包含 633 篇文档，其中 BNEWS 有 238 篇，NWIRE 有 298 篇，WEBLOG 有 97 篇。我们对这些文档进行了预处理，由于其中一些文章中的单句字数过多或语法不规范，我们过滤掉了 101 篇，最终从中选取了 532 个文档，总共有关系正例 7,630 个，负例 83,063 个。ACE 2005 数据集里总共出现了 6 大类实体类型，PHYS，PER-SOC，PART-WHOLE， ORG-AFF，ART，GEN-AFF。

4.3 实验结果与分析

我们在实验中数据都采用有序 PT 树结构，按照关系种类把关系实例划分成 14 个数据集，训练数据集和测试数据集的比例是按照 1：4 划分。实验结果如表 1 所示。

表 1 系统对 6 个关系大类关系抽取性能比较(ACE Chinese 2005 单个数据集)

关系大类	测试实例数	找出实例数	正确数	准确率	召回率	F 值
PHYS	167	37	20	54.1	12.0	19.6
R.PHYS	105	17	10	58.8	9.5	16.4
PER-SOC	56	17	8	47.6	14.3	21.9
R.PER-SOC	36	5	4	80.0	11.1	19.5
PART-WHOLE	20	3	3	100.0	15.0	26.1
R.PART-WHOLE	359	282	178	63.1	49.6	55.5
ORG-AFF	31	4	3	75.0	9.7	17.1
R.ORG-AFF	333	306	228	74.5	68.5	71.4
ART	82	18	12	66.7	14.6	24.0
R.ART	20	4	3	75.0	15.0	25.0
GEN-AFF	58	43	42	97.7	72.4	83.2
R.GEN-AFF	264	200	147	73.5	55.7	63.4
Average	1531	936	769	68.3	43.0	52.8

通过对表 1 实验结果的分析，我们得到以下结论：

1. 对那些实例数相对较少的关系，分类性能大都较低（如 PER-SOC，ORG-AFF，R.ART 等）。这种情况是因为数据稀疏造成分类器训练模型不完备，使得训练得出的模型无法识别出被测试的关系类别。
2. 对于 PHYS 关系识别效果较差，不论是正关系还是反关系，虽然这类关系实例数较多，但是一方面由于地理位置关系这类的句法树种类复杂，没有稳定的结构共性；另一方面由于在这类关系中相似的负例数远远大于正例数，造成分类器往往把实例分成没有关系的。
3. 对于复杂的句法树，两个实体路径间隔较远的结构在关系分类时绝大部分都识别不出。

从表 1 我们发现针对 R.PART-WHOLE、R.ORG-AFF、R. GEN-AFF 这几类关系的抽取性能较好,我们希望找到其原因，因此我们对中文语料进行分析，找到了三种具有代表性的句法树结构：

句法树结构：

1. (NP (E1 ())(E2 ())
2. (NP (NP (E1 ()))(NP (E2 ()))
3. (NP (DNP (NP (E1 ()))(DEG 的))(NP (E2 ()))

具体实例：

1. (NP (E1 (NR 台北))(E2 (NR 大安森林公园))
2. (NP (NP (E1 (NR 台北)))(NP (E2 (NN 市长)))
3. (NP (DNP (NP (E1 (NR 赖昌星)))(DEG 的))(NP (E2 (NN 远华公司)))

为了更准确的分析，我们统计了这三种句法结构在 R.PART-WHOLE、R.ORG-AFF、R. GEN-AFF 这三个数据集中的测试识别情况，如表 2 所示。

表 2 三种句法结构识别情况

数据集	结构 1		结构 2		结构 3	
	测试集	识别正确	测试集	识别正确	测试集	识别正确
R.PART-WHOLE	89	68	89	86	16	8
R.ORG-AFF	81	73	113	108	11	7
R.GEN-AFF	48	43	78	68	13	12

从表 2 中我们可以看出第二种结构分类准确率最高，其次是第一种，第三种结构的分类性能相对来说低于前两种结构。这是因为第二种结构的数量多而且正例数要多于负例数，所以分类器训练时这种结构的数量充足，模型的区分度好。第一种结构数量虽然多但是由于负例数要多于正例数所以效果相对差一些。第三种结构的正例数也多于负例，可是由于数量很少使得模型训练时候区分度小，导致最终的分类效果不好。

5 分类器训练和测试速度优化

卷积树核函数方法是通过比较两个句法分析树的相同子数的数量来计算其相似度，因此句法树结构的复杂程度直接影响到分类器的性能。句法树越复杂，分类器训练模型所需要的时间就越长，反之则越少。从直觉上判断，具有关系的实体对往往距离较近，而距离很远的实体对之间存在关系的可能性很小，语料库的统计结果也支持这一结果。衡量实体对之间的距离有多种方法，在句法树中最直观的就是计算包含实体对的最短路径的长度，即该路径上的语法成分结点数（除

去实体结点本身)。例如在图 1 的句法树中，“E1”和“E2”的路径结点长度就为 6。

通过实验统计，我们发现当两个实体间的距离大于 8 时，分类器训练模型所需要的时间很长，而能识别出来的实例却很少，因此，我们按照路径长度分别为 5，6，7，8 进行试验，计算运行时间，得出的结果见表 3。

表 3 按不同路径节点长度筛选后系统性能（单数据集）

路径长度	性能(P/R/F)(%)	运行时间/分钟	基准性能(P/R/F)(%)	基准时间/小时
5	61.9/42.0/50.0	40	68.3/43.0/52.8	39
6	63.4/42.3/50.8	60		
7	66.0/43.5/52.4	100		
8	66.3/43.6/52.6	130		

从表 3 中可以看出，原系统在一个数据集上运行程序耗费的时间为 39 个小时，最终的 F 值是 52.8%。采用关系实例筛选策略后，如果关系实例按照路径节点长度为 5 进行筛选，运行时间大大缩短，仅用了 40 分钟就完成分类器的训练和测试任务，当然性能上也有下降，F 值为 50.0，下降了约 2.8 个百分点。如果按照路径结点长度为 6 来筛选的话，虽然运行时间增加了 20 分钟，但分类器分类效果却并不明显，F 值也仅提高了 0.8 个百分点。当按照路径结点长度按照 7 来筛选后，和长度为 6 相比运行时间增加可 40 分钟，相应的 F 值也提高了接近 1.6 个百分点。如果按照路径结点长度为 8 来筛选的话，虽然运行时间增加 30 分钟，但分类器分类效果却并不明显，F 值也仅提高了 0.2 个百分点。这说明路径节点长度越长其关系实例数量越少，所耗费的时间增加不少，分类器分类效果却改善很小。为了节约系统时间，我们用较小的性能上的损失来换取大量时间，因此我们在实验中都按照路径节点长度为 7 进行筛选，运行时间仅为 100 分钟，取得的性能也仅比基准系统下降了 0.4 个百分点。

6 总结与展望

本文首先介绍了中文关系抽取的研究现状以及国内外研究关系抽取的各种方法，论述了卷积树核函数方法计算相似度的原理，探索了关系实例的结构化信息的合理表达形式；接着按照中文语义关系抽取系统的流程介绍了语料预处理、关系实例生成，通过基于卷积树核函数的 SVM 分类器来判断实体间的语义关系类别，即实现语义关系的抽取，构建了基于中文语料的实体关系抽取的系统；最后探讨了优化分类器训练和测试速度的方法。

基于卷积树核函数的关系抽取方法能有效捕获关系实例中的结构化信息，但此前在中文语义关系抽取中的尝试并不成功。本文以最短路径包含树作为关系实例的结构化信息表示方法，采用卷积树核函数计算关系实例之间的相似度，进而使用 SVM 分类器实现语义关系的抽取。在 ACE 2005 语料库上的关系大类抽取实验取得了合理的结果，表明该方法是行之有效的，这主要是由于我们所采用的中文句法分析器具有较高的性能。同时，针对树核函数的共同缺点——训练和测试时间较长，我们设计了基于实体间路径长度的优化方法，大大减少了训练语料中的关系实例，从而明显缩短学习时间，而抽取性能则基本保持不变。

下一步我们要做的工作还有很多，虽然最短路径包含树（PT）结构已经相对简洁，噪声信息比 MCT 树少了许多，但是由于中文句子中存在很多两个关系实体相距很远，中间夹杂着很多修饰词的实例。很多关系实例的句法树结构还是很复杂，存在大量的干扰信息，造成分类器无法准确识别。因此如何找出这些复杂结构的特点，对其进行适当的裁剪过滤，去除干扰信息，还需要进一步的研究和实验。

参 考 文 献

[1] 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.

[2] 董静, 孙乐, 冯元勇, 黄瑞红. 中文实体关系抽取中的特征选择研究[J]. 中文信息学报, 2007, 21(4): 80-85, 91.

[3] Zelenko D., Aone C., and Richardella A.. Kernel methods for relation extraction [J]. Journal of Machine Learning Research. 2003, 3 (2003): 1083-1106.

[4] Culotta A. and Sorensen J. Dependency tree kernels for relation extraction [C]. ACL'2004, 2004, pages 423-429. Barcelona, Spain.

[5] Zhang M., Zhang J., Su J., and Zhou G. D. A Composite Kernel to Extract Relations between Entities with both Flat and Structured Features [C]. COLING-ACL'2006. 2006, pages 825-832. Sydney, Australia.

[6] Bunescu R. C. and Raymond J. M. Subsequence Kernels for Relation Extraction [C]. In Proceedings of NIPS'2005, December 2005. Vancover, B.C.

[7] Zhang M., Zhang J., Su J., and Zhou G. D. A Composite Kernel to Extract Relations between Entities with both Flat and Structured Features [C]. COLING-ACL'2006. 2006, pages 825-832. Sydney, Australia.

[8] Collins M. and Duffy N. Covolution kernels for natural language [C]. NIPS'2001, 2001, pages 625-632. Cambridge, MA.

[9] Zhou G. D., Zhang M., Ji D. H., and Zhu Q. M. Tree Kernel-based Relation Extraction with Context-Sensitive Structured Parse Tree Information [C]. EMNLP/CoNLL'2007, 2007, pages 728-736. Prague, Czech.

[10] Qian L. H., Zhou G. D., Zhu Q. M., and Qian P. D. Exploiting constituent dependencies for tree kernel-based semantic relation extraction. COLING'2008: 697-704. 18~22 Aug. 2008. Manchester, UK.

[11] Che W. X., Jiang, J. M. Su Z., Pan Y., and Liu T. Improved-Edit-Distance Kernel for Chinese Relation Extraction [C]. In Proceedings of the 2nd international Joint Conference on Natural Language Processing (IJCNLP'05), 2005. Jeju Island, Korea.

[12] 刘克彬, 李芳, 刘磊, 韩颖. 基于核函数中文关系自动抽取系统的实现[J]. 计算机研究与发展, 2007, 44(8): 1406-1411.

[13] Huang R. H., Sun L., and Feng Y. Y. Study of Kernel-Based Methods for Chinese Relation Extraction [C]. LNCS (Lecture Notes in Computer Science): Volume 4993, pages 598-604, 2008. Springer Berlin/Heidelberg.