文章编号： 1003-0077（2013） 00-0000-00

基于凸组合核函数的中文领域实体关系抽取*

陈鹏[1]，郭剑毅[1,2]，余正涛[1,2]线岩团[1,2]，严馨[1,2]，魏斯超[1]

（1.昆明理工大学信息工程与自动化学院，云南 昆明 650500；

2. 昆明理工大学智能信息处理重点实验室，云南 昆明 650500）

摘要：针对在采用支持核函数的机器学习算法进行基于特征的中文领域实体关系抽取中，不同核函数对不同中文领域关系抽取在效果上存在差异性的问题，该文提出一种基于凸组合核函数的中文领域实体关系抽取方法。首先，选取实体上下文的词、词性等信息，短语句法树信息及依存信息作为特征，然后通过以径向基核函数，Sigmoid核函数及多项式核函数组成的不同组合比例的凸组合核函数将特征矩阵映射成为不同的高维矩阵，利用支持向量机训练这些高维矩阵构建不同分类模型后测试抽取性能，以确定最优组合比例的凸组合核函数。在收集600篇旅游领域语料上进行关系抽取，实验结果表明最优凸组合核函数能增加实体关系抽取效果，F值达到62.9%。

关键词：关系抽取；凸组合核函数；支持向量机

中图分类号：TP391 **文献标识码：**A

Chinese Field Entity Relation Extraction based on Convex Combination Kernel Function

CHEN Peng[1],GUO Jianyi[1,2], YU Zhengtao[1,2], XIAN Yantuan[1,2], YAN Xin [1,2], WEI Sichao[1]

(1. The School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, Yunnan 650051,China;

2. Key Laboratory of Intelligent Information Processing, Kunming University of Science and Technology, Kunming, Yunnan 650051,China;)

Abstract: For the problem that based on the feature method, different kernel functions caused different performances in Chinese field entity relation extraction by the machine learning method, which supports kernel function, this paper proposed a convex combination kernel function method to deal with this problem. First, this paper chose lexical information, phrase syntactic information and dependent syntactic information as features. Next step was to get different high-dimensional matrixes though mapping by different convex combination kernel functions. Finally, we could get the optimal kernel by testing all classified model that trained all high-dimensional matrixes by SVM. This paper conducted the relation extraction experiment on collecting 600 corpuses in tourist field, the experimental result shows that the optimal convex combination kernel function this paper presents can effectively improve the extraction performance, and it gets the best F value which reaches 62.9.

Key words: Relation Extraction; Convex Combinationm Kernel Function; Support Vector Machine

1引言

中文领域实体关系抽取是指在多样化的中文领域文本中找出实体对之间的关系。作为自然语言处理的基础，中文领域实体关系抽取为中文领域信息检索、自动问答系统、机器翻译、本体构建等提供重要技术支持。目前主要利用机器学习的方法实现中文领域实体关系抽取。按照是否可以利用核函数，机器学习方法主要可以分为支持核函数的方法及不支持核函数的方法[1][2]，不支持核函数的方法主要选取适当的特征，结合机器学习算法构造分类模型进行

*收稿日期： 定稿日期：

基金项目：国家自然科学基金(61175068)

作者简介：陈鹏（1987—），男，硕士研究生，主要研究方向为信息抽取；郭剑毅（1964—），通信作者，女，教授，硕士生导师，主要研究方向为信息抽取；余正涛（1970—），男，教授，博士生导师，主要研究方向为自然语言处理、信息检索及信息抽取。

关系抽取，文献 [1] [2]分别使用最大熵及条件随机场构造分类模型，但是这类方法在训练实例较少的情况下，就可能会产生维数灾难，并且这类方法的输入模型是固定的，致使训练产生的分类模型不具有很强的通用性。反观核函数的引入可以使输入空间的维数与原空间的维数无关，大大减小了计算量，并且核函数的种类多样化可以解决不支持核函数方法输入模型固定的问题，所以支持核函数的机器学习算法成为实现中文领域关系抽取的主流方法。按照高维矩阵构造方式来分，支持核函数的方法主要包括特征向量方法[3]，卷积核函数方法[4,5]及两种方法复合的多核融合方法[6]。卷积核函数方法是将两个实体所在的句子结构化表示，例如字符串、句法树等，通过计算每个实例中相同的子结构数目构造高维矩阵，但是受制于构成高维矩阵时计算复杂度巨大的问题，卷积核函数方法及包含卷积核函数的多核融合方法往往难以应用于实际的中文领域实体关系抽取任务中，而特征向量方法可以通过抽取不同信息作为特征，快速构建特征矩阵，并由不同的核函数将特征矩阵映射到高维矩阵，从而避免卷积核函数计算复杂度巨大的问题，因为受到广泛关注。需要注意的是，在基于特征向量方法的过程中主要分为两个部分：选取特征形成特征矩阵及选取核函数映射特征矩阵得到高维矩阵，由于在通过机器学习方法训练获取分类模型的过程中，高维矩阵是唯一用到的信息，所以核函数的选择对中文领域实体关系抽取性能起到至关重要的作用。由于不同核函数对不同中文关系抽取在效果上存在差异性，即对于某些特征，使用部分核函数能够增加中文领域实体关系抽取性能，但是部分核函数起到相反作用，所以在中文领域实体关系抽取中，使用单一的核函数不具有通用性。

针对在中文领域实体关系抽取中，使用单一的核函数不具有通用性的问题，本文提出一种将多种单一核函数凸组合作为核函数，并且在实体上下文的词、词性等词法信息基础上，加入短语句法信息、依存句法信息共同作为特征的中文领域实体关系抽取方法。在中文旅游领域中，对预处理后的语料提取实体上下文的词、短语句法信息及依存句法信息作为特征，形成特征矩阵，并将特征矩阵映射到径向基核函数、Sigmoid核函数及多项式核函数组成的不同凸组合中，形成不同的高维矩阵，再利用支持向核函数的机器学习方法训练获得分类器模型，利用测试语料枚举寻找效果最优的分类模型，最终用这个分类模型的凸组合核函数实现中文旅游领域实体关系抽取。

2中文旅游领域实体关系抽取

实体关系抽取领域主要分为四个部分：语料预处理、提取特征形成特征矩阵、核函数映射形成高维矩阵、学习得到不同分类模型并寻找最优分类模型。

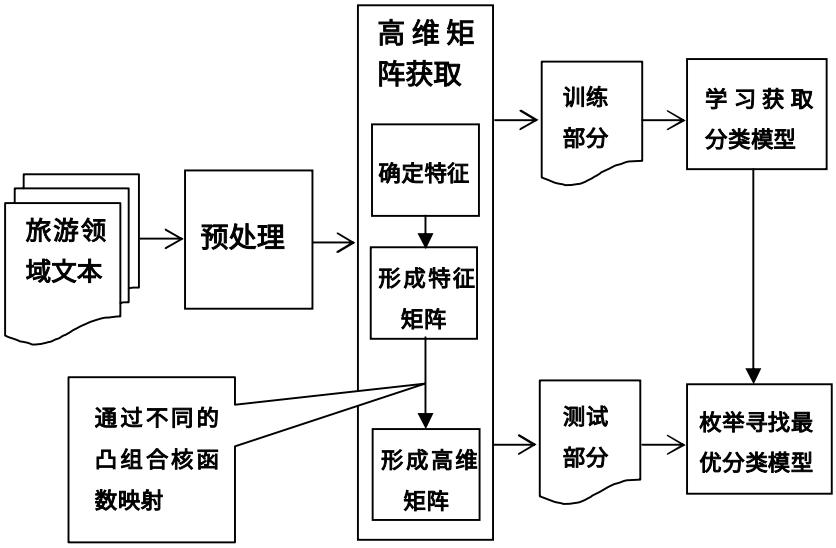

图 1 实体关系抽取原理图

2.1预处理

预处理包括分词、词性标注、命名实体识别、句子切分、关系候选生成及标注候选实例。

首先，本文调用中科院的 ICTCLAS工具对输入文本进行分词及词性标注。其次，人工对已经完成分词及词性标注的语料进行命名实体标注，并根据条件随机场规则训练命名实体识别模型[8]，从而实现自动命名实体识别。其中实体的类别需要预先定义，本文使用文献 [1]对旅游领域实体类别的定义，具体为：景点、地点、小吃、特产、酒店、数字表达式、日期、节日。之后，依据标点符号及上下文特点，将已标注语料切分成独立的句子。最后，枚举找出每个句子中所有可能的实体对组合，每个组合成为一个候选实例，并根据实例是否有关系人工标注实例。通过以上步骤，预处理完毕。需要注意，实体关系类型也需要预先定义，本文参考文献 [1]对中文旅游领域关系的定义，如表 1所示。对如：对于句子“ 1921年，朱德出任云南陆军宪兵司令，继任警察厅长。”预处理之后的结果为：

句子 1: [1921年]t, [朱德]pn出任 [云南]dd陆军宪兵司令，继任警察厅长。

候选实例 1：实体 1: [1921年]t 实体 2: [朱德]pn

候选实例 2: 实体 1: [1921年]t实体 2: [云南]dd

候选实例 3: 实体 1: [朱德]pn实体 2: [云南]dd

表 1 旅游领域实体关系

实体关系大类	实体关系子类
地理位置关系	包含，毗邻，相距
结构从属关系	特产，属于
属性限定关系	门票价格，占地面积，海拔，温度，湿度，最佳季节，景点星级，酒店星级，习俗

2.2提取特征形成特征矩阵

本文特征选择主要参照文献 [2, 3,8-10]中的方法，其中文献 [8-10]并不是在中文领域实体关系中提出，而是在英文的新闻领域及医学领域提出的特征，在基本的实体信息，词汇局部上下文信息及包含嵌套信息基础上，文献 [8]增加短语块信息作为特征，文献 [9]增加依存信息作为特征。本文在中文旅游领域将短语句法信息及依存信息加入特征集中，期待能够增加关系抽取性能。特征选择完毕后，即可根据特征形成特征矩阵。

2.2.1词法信息

（1）实体信息

实体信息是基本的词汇信息，包括第一个实体大类、第一个实体小类、第一个实体词性、第二个实体大类、第二个实体小类及第二个实体词性。

（2）词汇局部上下文信息

文献 [3]验证了词汇特征窗口不宜大，以防止噪声影响过大。一般选择左右窗口为 2。本文选择 2-3-2的模式，即选择实体一左边两个词，实体二右边两个词及实体间三个词作为特征。

（3）包含信息

包含信息主要反映了实体对间的词汇信息及包含情况。本文选择实体对间词汇的数目、实体的数目及实体是否是包含关系作为嵌套信息。

2.2.2短语句法信息

短语句法树反映句子的语法结构，可以表达长距离的语义信息。句子“ 朱德出任云南陆军宪兵司令，继任警察厅长。”的最小完全句法树及如图 2所示。

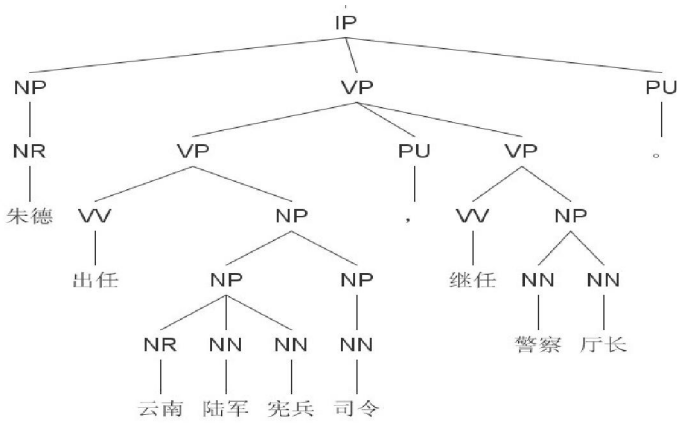

图2 最小完全短语句法树实例表示

最小完全句法树是指两个实体的最近公共根节点作为根节点的结构树，由于最小完全句法树包含一定上下文语义信息，也去处了一定的噪声干扰，所以本文利用最小完全句法树进行特征提取。由于两个实体在句法树中的路径过于具体，十分容易造成数据稀疏的问题，为了避免这个问题，本文选择两个实体路径中节点的数目、两个实体的根节点类型作为特征。由于句法树的结构信息十分具体，为了解决数据稀疏带来的低召回率的问题，本文采取细化句法树结构信息的方针，具体为：

句法树特征 1：第一个实体到根节点的路径；

句法树特征 2：第二个实体到根节点的路径；

句法树特征 3：两个实体的公共根节点类别；

句法树特征 4：第一个实体到根节点的节点数目；

句法树特征 5：第二个实体到根节点的节点数目。

2.2.3依存信息

依存树可以揭示句子中的长距离依存关系，并且能避免非结构化特征中出现的噪音，可以为关系抽取提供更为有效的信息。对于句子“朱德出任云南陆军宪兵司令，继任警察厅长。”其依存树如图 3所示。同样由于依存树的结构信息分布十分具体，为了解决数据稀疏带来的低召回率的问题，本文采取细化依存句法信息的方针，具体为：

依存特征 1：第一个实体到根节点的路径；

依存特征 2：第二个实体到根节点的路径；

依存特征 3：第一个实体到根节点间的依存类别；

依存特征 4：第二个实体到根节点间的依存类别；

依存特征 5：两个实体间是否有直接的依存关系。

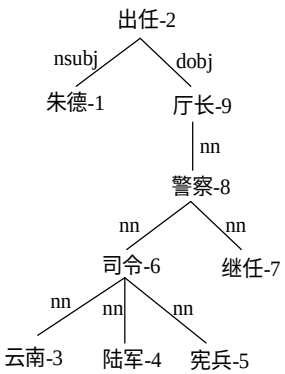

图3 依存树实例表示

2.3凸组合核函数及获取高维矩阵

在得到特征矩阵后，将特征矩阵通过核函数映射得到高维矩阵，通过支持向量机训练获取分类模型，最终利用这个分类模型进行中文领域实体关系抽取。

2.3.1核函数在获取高维矩阵中的作用

对于经过特征提取的特征矩阵 $K=(X_1, X_2, \cdots X_m)^T$，其中 X_i (i=1,2,…，m)为每个实例提取特征后的向量，将特征矩阵 K映射到核函数 k后得到的高维矩阵如式（1）为：

$$K_{training} = \begin{bmatrix} k(X_1, X_1) \ldots k(X_1, X_m) \\ \cdots \quad \cdots \quad \cdots \\ k(X_m, X_1) \ldots k(X_m, X_m) \end{bmatrix} \tag{1}$$

可以看到，在高维矩阵的每一个元素都是由特征矩阵的某两条向量及核函数唯一决定的，核函数起到了至关重要的作用。总之，寻找某中文领域最优核函数对提高关系抽取性能十分有意义的。

2.3.2凸组合核函数及获取高维矩阵

由于不同核函数对不同中文关系抽取在效果上存在差异性，了让核函数对由不同的信息表示的特征均有良好的适应性，本文将不同单一核函数进行凸组合，以期待最优凸组合核函数具有良好的适应性。按照核函数表达式不同，核函数可以分为平移不变核函数及内积核函数，其表达式分别为 $k(x,y) = f(x-z)$ 及 $k(x,y) = f(< x,z >)$。

为了融合这两核函数在关系抽取中的特性，本文选择这两种核函数中常用的径向基核函数，Sigmoid核函数，多项式核函数，分别如式（2）（3）（4）所示。其中径向基核函数属于平移不变核函数，Sigmoid核函数及多项式核函数均为内积核函数。

$$k_1(x,y) = \exp(-\| x-y \|^2) \tag{2}$$

$$k_2(x,y) = \tanh(x^T y) \tag{3}$$

$$k3(x,y) = (x^T y)^3 \tag{4}$$

他们的凸组合如 (5)所示。其中 $\sum_{i=1}^{2} \alpha_i = 1$，$\alpha_i \in R^+$。确定了核函数后，通过式（5）的就可以将特征矩阵映射成为高维矩阵。

$$k_4(x,y) = \sum_{i=1}^{2} \alpha_i k_i(x,y) \tag{5}$$

2.4寻找最优分类模型

通过 2.3.2得到了由核函数不同的凸组合映射的高维矩阵，直接观察高维矩阵并不能看出由哪一个高维矩阵训练得到的分类模型具有更好的实体关系抽取效果，即哪一个凸组合核函数有更好的适应能力。为了得到最优的凸组合核函数，首先训练每一个高维矩阵，得到相应的训练模型，然后再用测试语料测试出最优实体关系抽取性能的训练模型，这个训练模型对应的核函数即为最优凸组合核函数。

3试验方法及结果

本文使用的语料为人工从互联网及文献资料中获取的中文旅游文本共 600余篇。由于在实体关系抽取领域可以使用核函数的各种机器学习算法中，支持向量机（SVM）有最好的表现[11]，故本文采用 SVM作为机器学习算法。本文选用台湾大学林智仁等人开发的 LIBSVM①作为 SVM工具包。利用 Stanford Parser②进行短语句法及依存句法分析，在短语句法分析中，选用概率上下文无关语法，依存信息的生成使用 CCprocessed依存表达形式。在训练中使用

① http://www.csie.ntu.edu.tw/~cjlin/libsvm

② http:/ /nlp.stanford.edu/software/lexparser.shtml

10 倍交叉验证以最大化利用数据。实验评测采用自然语言处理的通用标准：准确率、召回率、 F值， F值评测系统的最终性能。

3.1实验设计

为了验证本文方法的有效性，并与其它同类方法进行比较，本文设计了 3项任务：

任务 1，研究在实体上下文的词、词性等词法信息基础上，加入短语句法信息、依存句法信息共同作为特征对中文领域实体关系抽取性能的影响。任务 1将分别依次加入不同的特征，寻找最佳的特征组合，任务 1中暂时选择式（5）中参数$\alpha_1=1$，$\alpha_2=0$，$\alpha_3=0$作为凸组合核函数的权重参数。

任务 2，对比三种单一核函数组成的凸组合核函数与单独的核函数及他们的两两组合的凸组合核函数的关系抽取性能，验证单一核函数在中文旅游领域关系抽取的差异性，并验证多种单一核函数凸组合能有效解决这个问题。实验中设置式（5）中凸组合的系数的上下限均为 0至 1，步长为 0.1，共 11种选择。采用枚举的方式找到最好的那一组系数，对于 n种单一核函数，找到他们最优的凸组合需要计算 11*(n-1)次。

任务 3，验证本文提出系统的有效性，在使用文本的语料下，用本文提出的方法与同类方法进行比较。

3.2实验结果及分析

3.2.1不同特征在关系抽取中的表现

表 2 不同特征下的抽取性能指标

特征	评价指标（%）		
	准确率	召回率	F 值
词法信息	61.1	48.2	53.9
+短语句法信息	61.9	52.5	56.8
+依存句法信息	**62.5**	**53.2**	**57.5**

由表 2可以看出，只选择词法信息作为特征的中文旅游领域实体关系抽取性能较差。加入短语句法信息及依存句法信息后，抽取性能都有提高，特别是召回率增加比较明显，说明加入短语句法信息及依存句法信息作为特征是提高中文领域实体关系抽取性能的有效手段。

3.2.2三种单一核函数及凸组合核函数的关系抽取性能对比

表 3 单一核函数与凸组合核函数的抽取性能指标

核函数类别	评价指标（%）		
	准确率	召回率	F 值
单一核函数			
径向基核函数	62.5	**53.2**	57.48
Sigmoid 核函数	48.2	39.8	43.6
多项式核函数	**69.7**	49.0	**57.55**
凸组合核函数（两个单一核函数组合）			
径向基+Sigmoid	63.4	53.6	58.1
径向基+多项式	**72.1**	**55.3**	**62.6**
Sigmoid+多项式	70.0	49.3	57.9
凸组合核函数（三个单一核函数组合）			
三种核函数融合	72.3	55.7	62.9

表 3 表示，在单一核函数的中文旅游领域实体关系抽取性能中，利用多项式核函数由最高的准确率及 F 值，径向基核函数由最高的召回率，而 Sigmoid 核函数的抽取性能较差，这也验证了单一核函数对于相同的特征矩阵抽取性能存在差异性。在两两组合的凸组合核函数中，利用径向基与多项式核函数组成的凸组合核函数由最好的性能，说明凸组合的基础核函

数性能对凸组合核函数的性能有正相关的影响。三种单一核函数构成的凸组合核函数由最佳的抽取性能，并且两种单一核函数的组成的凸组合核函数抽取性能都优于单一核函数的抽取性能，说明多种单一核函数融合的凸组合核函数能解决单一核函数不具有通用性的问题。

3.2.3与其它方法的比较

表 4 其他同类方法比较

方法	评价指标（%）		
	准确率	召回率	F 值
不支持核函数的机器学习方法			
最大熵	**74.1**	48.9	58.9
基于支持向量机方法			
卷积树核函数（最短路径树）	61.3	**58.9**	60.1
线性核函数	70.2	48.6	57.4
本文提出的最优凸组合核	72.3	55.7	**62.9**

表 4显示，利用最大熵方法进行中文旅游领域关系抽取，得到了最高的准确率。利用最短路径树[12]的卷积树核函数方法[13]进行关系抽取取得了最好的召回率。本文提出的最优凸组合核函数方法有不错的召回率及准确率，且取得了最好 F值，这充分证明本文提出的最优凸组合方法的有效性。

4结束语

本文在实体上下文的词、词性等词法信息基础上，加入短语句法信息、依存句法信息作为特征，通过径向基核函数，Sigmoid核函数及多项式核函数的不同凸组合将特征矩阵映射到不同高维矩阵，并以支持向量机进行训练得到不同分类器，通过枚举寻找性能最优的分类器，最终利用这个分类器进行中文领域实体关系抽取。在旅游领域中，本文提出的最优凸组合核函数抽取系统取得了 62.9的 F值。下一步工作中，我们将尝试挖掘其它有效信息作为特征，以及尝试寻找核函数与语料的深层次关系，试图进一步提高中文领域实体关系抽取性能。

参考文献

[1]Chunya Lei,Jianyi Guo,Zhentao Yu,Shaoming Zhang,Cunli Mao,Chaosheng Zhang.The Field of Automatic Entity Relation Extraction based on Binary Classifier and Reasoning[C].//Third International Symposium on Information Processing.Qingdao,China,2010:327-2-331.

[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.

[3]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.

[4]Peng Cheng, Gu Jinghang, Qian Longhua. Research on Tree Kernel-Based Personal Relation Extraction[J]. Communications in Computer and Information Science,2012,333:225-236.

[5]Liu Dandan, Zhao Zhiwei, Hu yanan, Qian Longhua. Incorporating Lexical Semantic Similarity to Tree Kernel-based Chinese Relatin Extraction[J]. Lecture Notes in Computer Science, 2013, 7717:11-21.

[6] 黄瑞红，孙乐，冯元勇，黄云平．基于核方法的中文实体关系抽取研究 [J]，中文信息学报，2008,22(5):102-108.

[7]郭剑毅，薛征山，余正涛．基于层叠条件随机场的旅游领域命名实体识别 [J].中文信息学报，2009,23(5):47-52.

[8]奚斌，钱龙华，周国栋，朱巧明，钱培德．语言学组合特征在语义关系抽取中的应用 [J].中文信息学报，2008,22(3):44-49,63.

[9]刘兵，钱龙华，徐华，周国栋．依存信息在蛋白质关系抽取中的作用 [J].中文信息报，2011,25(2):21-26.

[10]李丽双，刘洋，黄德根．基于组合核的蛋白质交互关系抽取 [J].中文信息学报，2013,27(1):86-92.

[11]Zelenko D, Aone C, Richardella A. Kernel methods for relation extraction[J]. Journal of Machine Learning Research, 2003, 3(6): 1083-1106.
[12]Zhang Ming, Zhang Jie, Su Jian, Zhou Guodong. A Composite Kernel to Extract Relations between Entities with both Flat and Structured Features[C]//COLING-ACL'2006. Sydney, Australia, 2006: 825—832.
[13]Collins M, Duffy N. Covolution kernels for natural language[C]// NIPS' 2001. Cambridge, MA 2001: 625-632.

作者联系方式：郭剑毅　云南省昆明市呈贡大学城昆明理工大学信息工程与自动化学院 邮编 650500　电话 13577102026 电子邮箱 gjade86@hotmail.com