计算机研究与发展
Journal of Computer Research and Development
ISSN 1000-1239/CN 11-1777/TP
44(8):1406~1411, 2007

基于核函数中文关系自动抽取系统的实现

刘克彬 李 芳 刘 磊 韩 颖
(上海交通大学计算机科学与工程系 上海 200240)
(captain2003lkb@sjtu.edu.cn)

Implementation of a Kernel-Based Chinese Relation Extraction System

Liu Kebin, Li Fang, Liu Lei, and Han Ying
(*Department of Computer Science & Engineering, Shanghai Jiaotong University, Shanghai* 200240)

Abstract Entity relation extraction (RE) is an important task in information extraction. In this paper, a novel kernel-based Chinese entity relation extraction system is presented, which appies the improved sequence kernel function with KNN learning algorithm to fulfill the RE task. Experiments are carried out on 3 kinds of relation types and their 6 subtypes defined in the ACE guidelines. Results show that the new approach achieves an average precision up to 88%, significantly higher than feature-based approaches and traditional kernel methods. The new approach has a better generalization capability especially on small training sets. The system consists of 8 independent modules including named entity detection, candidate generation, etc. for easy maintenance and update. The system is implemented either as a Java application or plug-ins on gate platform. It extracts not only the binary relation, but also their description such as job in employment relation.

Key words relation extraction; kernel function; information extraction; semantic; machine learning

摘 要 实体关系抽取是信息抽取的重要组成部分. 基于核函数的中文实体关系自动抽取系统应用改进的语义序列核函数,结合 KNN 机器学习算法构造分类器来分类并标注关系的类型. 通过对 ACE 评测定义的三大类 6 子类实体关系的抽取,关系抽取的平均精度可以达到 88%,明显高于基于特征向量和传统的序列核函数方法,该方法适合小训练集,易于学习新的实体关系. 系统由 8 个独立的模块构成,便于维护和升级. 系统既可以独立运行,也可以嵌入在开放的文本处理平台 GATE 环境. 为了更好地利用关系抽取的结果,系统扩展传统的二元关系,抽取关系的同时,抽取该关系的描述,形成完整的中文实体关系抽取系统.

关键词 关系抽取;核函数;信息抽取;语义;机器学习

中图法分类号 TP391

实体关系抽取(relation extraction)是指自动识别用自然语言表达的两个实体之间的关联,例如,“李明是新华公司的经理”,在这个句子中,人物实体“李明”和组织机构实体“新华公司”之间存在雇佣关系,即李明是新华公司的雇员,担任经理的职位. 关系自动抽取是信息抽取、自然语言理解技术不可缺少的重要环节. 美国国家标准技术研究院(NIST)在 MUC(message understanding conference)结束以后,进行了自动内容抽取(automatic content extraction, ACE)的评测,根据 ACE 的定义实体关系共有六大

收稿日期:2006-08-29;修回日期:2007-02-09
基金项目:上海市科委国际合作基金项目(045107035);中德语言技术联合实验室德方资助项目

类,其中包含 18 个子类的预定义关系.

关系抽取最初采用基于知识库[1]的方法,但是,这种方法需要针对不同的领域由专家构造大规模的知识库,抽取效果受到知识库规模和质量的制约,而且,系统移植和扩展需要耗费大量人力物力重建知识库,可移植性差. 机器学习的方法弥补了这一缺陷,将关系抽取转化为分类问题,通过构造关系候选,利用机器学习得到的分类器来标注这些候选属于哪一类预定义关系. 早期的机器学习方法是基于特征向量[2],这一类算法需要显式地将语料构造成特征向量形式,然后使用各种机器学习算法,如支持向量机(support vector machine,SVM)等构造分类器[3],这一类机器学习算法相对简单,仅利用词频等信息. 之后出现了基于核函数(kernel function)的机器学习算法,使用核函数替代特征向量内积运算计算两个对象的相似度,并具有良好的复合特性. 核函数可以引入先验知识,例如潜在语义(LSI)核函数[4]、主成分分析(PCA)核函数[5]等. 有的核函数可以运行于离散数据结构上,例如,树[6-8]、图[9]、序列[10-12]. Zhao 等人[13]提出了利用复合核函数进行关系抽取,充分利用了多种来源的信息,提高关系抽取的精度. Che 等人[14]提出了改进的编辑距离核来计算两个中文词串的相似度,抽取中文关系. 不同的方法都有一个共同的目的,希望通过对词汇的语义分析,改进单纯的字符串匹配,提高关系自动识别的能力. 和其他方法相比,序列核函数具有很多优势,它不仅具有良好的复合特性,而且考虑了特征之间的顺序和结构信息,适合解决关系抽取的问题. 本文介绍的关系抽取系统主要基于序列核函数[11],根据中文语言的特点,使用词作为序列的基本单位. 针对目前序列核函数缺乏语义支持,泛化能力不够,本文提出了一种改进的语义序列核函数,在运算过程中嵌入语义知识提高核函数的性能;针对目前二元关系的自动抽取,本文在系统实现中,结合核函数与规则的方法,将二元关系扩展,抽取具有重要意义的关系描述信息,例如"李明是新华公司的经理"中的经理职位.

1 改进的语义序列核函数介绍

中文文本可以看做是中文词汇的集合,定义 Σ 为一个中文词汇的集合,在这个集合上定义词的序列 $X=X_1X_2\cdots X_{|S|}$. $i=[i_1,i_2,\cdots,i_n]$ 表示 X 的索引的一个子集,其中 $1\leqslant i_1<i_2<\cdots<i_n\leqslant|X|$,则 $X[i]\in\Sigma^n$ 是 X 的一个子序列. $l(i)$ 表示 $X[i]$ 在原序列中跨过的宽度(最大索引和最小索引之差). n 是 $X[i]$ 包含的词数. 举例如下,假设"ACDBABC"为一个词序列,其中每个大写字母代表一个词,当 $n=3$ 时,假设要寻找包含 3 个词"ADB"的子序列,原序列中的"ACDB"和"ACDBAB"都将入选. 它们的索引序列分别为[1,3,4]和[1,3,6],在原序列中跨过的宽度分别为 3 和 5.

词序列核函数[11]的基本思想是根据两个词序列中的公共子序列数量来衡量两者相似度. 子序列中可能包含间隔项,因此利用衰减因子 λ 为每个公共子序列设置不同的权重(基于如下假设:包含间隔越多的子序列对整体相似度的贡献越小):

$$K_n(X,Y)=\sum_{u\in\sum^n}\sum_{i:u=X[i]}\sum_{j:u=Y[j]}\lambda^{l(i)+l(j)},\quad(1)$$

其中 u 是公共子序列,通过 3 层循环统计所有的公共子序列. $X[i]$ 和 $Y[j]$ 都是不连续的,衰减因子 λ 使得子序列跨越的距离和其权重成反比. 为了提高运算速度,文献[11]中介绍了一种高效的序列核函数的递归实现.

序列核函数运算过程基于词的匹配,忽略了语义知识. 事实上,汉语的表达方式灵活多样,使用不同词汇的序列可能在语义上具有较大的相似性,如"爱吃香蕉"和"喜欢吃苹果". 因此,本文改进以上算法,从 HowNet[15]中获取汉语词汇之间的语义相似关系,并将这种关系嵌入到核函数的计算过程中.

如何将语义知识加入核函数中是一个困难的问题,软匹配方法[11]遍历所有长度符合要求的子序列对,使用基于词语的相关性判断取代字符串匹配来计算子序列对的权值,这种方法虽然引入了语义信息,但是包含了大量无意义的匹配,增加了计算复杂度. 本文的方法充分利用词性标注(POS)的结果,将核函数的输入变为词性加词条的双序列结构,先寻找公共的词性子序列,并嵌入对应的词条子序列的语义相似度作为子序列的权值,提高匹配的目的性. 另外,对于匹配项和间隔项使用了不同的衰减因子 λ_m 和 λ_g.

在语义序列核函数中,序列 $X=X_1X_2\cdots X_{|X|}$,定义特征 X_i 为二元组(p,w),p 代表了 X_i 的词性,w 表示 X_i 的词条,$i=[i_1,i_2,\cdots,i_n]$ 和 $l(i)$ 的

定义同上.

语义信息嵌入相当于作如下的映射：

$$K_n(X,Y) = \varphi(X.p)^{\mathrm{T}} S \varphi(Y.p), \quad (2)$$

其中 Φ 是映射函数,把输入序列映射到词性子序列空间中的向量,S 是一个矩阵,其中的项代表了词性子序列对应的词条子序列的语义相似度. 映射 Φ 和矩阵 S 的计算均嵌入在核函数的计算过程中. 以下是新的核函数：

$$K_n(X,Y) = \sum_{u \in \sum^n} \sum_{i:u=X[i].p} \sum_{j:u=Y[j].p} \lambda_m^{2n} \times \prod_{k=1}^{n} SIM(X_{i_k} \cdot w, Y_{j_k} \cdot w) \prod_{i_1<l<i_n, l \in i} \lambda_g \prod_{j_1<h<j_n, h \in j} \lambda_g . \quad (3)$$

SIM 函数根据 HowNet 提供的语义知识计算两个词汇之间的语义相似度[16-17]. 两种不同的衰减因子分别指定匹配项和间隔项的权重. 以下是新的核函数的递归计算公式：

$$K_n(Xa,Y) = K_n(X,Y) + \sum_{j:Y_j.p=a.p} \lambda_m^2 K'_{n-1}(X, Y[1:j-1]) SIM(a.w, Y_j.w). \quad (4)$$

式(4)是改进后的 K_n 公式,a 是出现在所有的公共子序列里面的匹配项,λ_m 作为它的权重,SIM 函数计算 a 与其他匹配项之间语义相似度.

$$K'_i(Xa,T) = \lambda_g K'_i(X,Y) + K''_i(Xa,Y). \quad (5)$$

$$K''_i(Xa,Yb) = \lambda_g K''_i(Xa,Y) + \lambda_m^2 K'_{i-1}(X,Y) SIM(a.w, b.w) \delta(a.p, b.p). \quad (6)$$

需要注意的是式(6)中的布尔函数 δ 的输入是词性而不是词条. 输入小于 n(n 是公共子序列包含的匹配项数目)特殊情况下的公式如下：

$$K_n(X,Y) = 0, \text{ if } \min(|X|, |Y|) < n. \quad (7)$$

$$K'_i(X,Y) = 0, \text{ if } \min(|X|, |Y|) < i, \quad (i = 1, \cdots, n-1). \quad (8)$$

$$K''_i(X,Y) = 0, \text{ if } \min(|X|, |Y|) < i, \quad (i = 1, \cdots, n-1). \quad (9)$$

$$K'_0(X,Y) = 1. \quad (10)$$

通过上述改进,我们实现了在核函数中嵌入语义信息[17],并且计算复杂度依然保持在 $O(n|X||Y|)$.

2 中文关系抽取系统的实现

2.1 系统实现框架

中文关系抽取系统的设计思想是先构造候选关系样例,然后使用上述语义序列核函数与 KNN 学习算法联合构造分类器,对候选关系样例进行分类标注. 符合预定义关系类别的候选将被抽取出来并自动标注为该类关系. 图 1 是系统整体框架和各模块的设计. 本文介绍的关系抽取系统具备两种不同的实现,一种是独立运行的 Java Application,另一种是可以运行在 Gate 自然语言处理平台下的插件库.

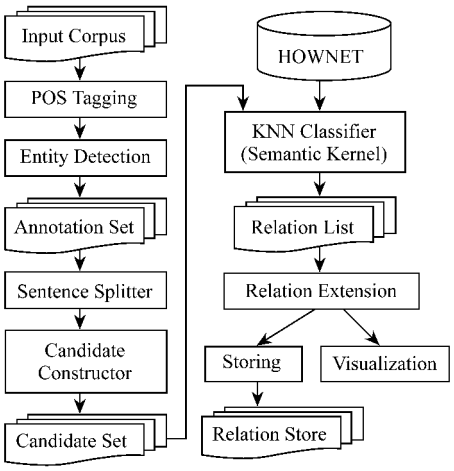

Fig. 1 Framework of relation extraction system.

图 1 关系抽取系统框架

系统主要模块说明如下,其中每个模块都实现为独立应用和 Gate 插件两种形式：

1) 词性标注. 该模块使用中国科学院的 ICTCLAS 工具对输入文本进行分词和词性标注.

2) 命名实体识别. 该模块从大量训练语料中自动学习命名实体的识别规则并依据贝叶斯公式建立概率模型. 为更有效地识别一些专名等建立了用户词典.

3) 句子切分. 该模块依据标点和上下文信息将标注集合切分成独立的句子. 每一个句子是一个标注序列,标注包含了词条、词性以及实体标注的结果.

4) 关系候选生成. 该模块的功能是在句子中枚举找出所有可能的实体对组合,连同上下文构造候选关系实例.

5) 关系分类模块. 该模块使用第 1 节描述的改进序列核函数和 KNN 机器学习算法对所有候选的关系实例进行分类判别. 挑选出包含预定义关系的

关系实例并对其类型进行标注.

6) HowNet 语义知识获取[17]. 该模块根据 HowNet 中的语义知识获取词汇语义相似度.

7) 关系扩展. 该模块对二元关系进行了扩展,从中提取出有重要意义的描述信息,例如人物实体相关的职务等.

8) 关系存储和关系可视化. 这两个模块提供了格式化存储和显示实体关系的功能.

2.2 二元关系的扩展

二元关系考察两个实体之间的联系,例如组织结构和人物之间存在的雇佣关系,或者两个人物之间存在的亲属关系. 但是这种二元关系往往是不完整的,例如二元雇佣关系中缺少人物的职务这一重要的描述信息. 因此,本文将二元关系进行扩展,增加描述信息作为实体间关系抽取的结果. 图 2 是实体关系的一个实例,作为二元关系得到的信息是"李明"受雇于"新华公司". 而在另外一个例子"新华公司销售员方明目前在北京出差"中抽取出的二元关系是"方明"受雇于"新华公司". 这两个实体关系分别表述了两个人物受雇于"新华公司"这一事实,但他们在同一组织中担任不同的职务,反映了两个关系的差异,加入描述之后实体关系提供的信息更加完整.

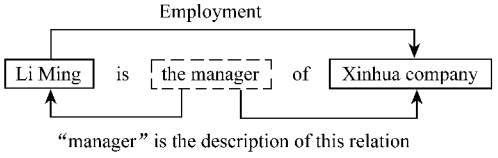

Fig. 2 Entity relation and relation description.

图 2 实体关系和关系描述

对于不同的关系类别定义不同类型的描述. 例如:1)雇佣(employment)关系中的职务包括"经理"、"部长"、"司机"等,2)就学/毕业于(student-alum)关系中的学位包括"学士"、"硕士"、"博士"等,3)事务(business)关系中的"合伙人"、"同事"、"上司"等,以及其他关系类型中出现的描述. 描述信息抽取的方法采用规则和语义相结合,抽取过程分以下两步进行:第 1 步从训练语料中学习泛化得到描述信息抽取规则,利用这些规则从关系实例中选出一个或者多个描述候选项. 第 2 步是建立各种描述的常见短语字典,在抽取特定描述信息的时候,计算候选项和字典中词和短语的语义相似度,选择相似度最高的词或者短语作为描述信息.

3 中文关系抽取实验

实验选择 ACE 定义中的三大类及 6 个子类作为预定义关系类别. 这些类别包括 Physical 及其 Near 子类、Personal-Social 及其 Business, Family 子类、ORG-Affiliation 及其 Employment, Student-Alum, Sports-Affiliation 子类. 实验语料来自于 Web 上选取的文档. 包含政治、体育、军事、人物、环境、经济领域的文本共计 312 篇,实验中系统从中自动生成出 8132 个关系候选. 实验对以下 3 种方法做了比较. 方法 1 是传统的基于特征向量的方法,通过向量的内积来计算对象之间的相似度. 向量每一维的权重采用 TF-IDF 来计算. 方法 2 是基于传统的序列核函数的方法. 方法 3 是本文介绍的方法,使用了改进后的语义核函数.

实验分为两个阶段,第 1 阶段的实验采用固定训练集合对 3 种不同的系统进行训练,然后进行上述 6 个子类实体关系的抽取实验并统计结果. 第 2 阶段实验需要重复多次,每次使用规模递减的训练集合以考察各个系统在训练样例数量减少的情况下性能的变化.

3.1 实验结果分析

第 1 阶段实验的结果见表 1 至表 3. 表 1 列举了 ORG-Affiliation 的 3 个子类关系抽取的结果,表 2 列举了 Physical 中的 Near 子类关系抽取结果,表 3 是 Personal-Social 的两个子类. 表中 P 表示系统关系抽取的准确率;R 表示系统的召回率;F 测度综合以上两个标准,反应了系统地整体性能.

Table 1 Relation Extraction Results (ORG-Affiliation)

表 1 关系抽取结果(ORG-Affiliation) %

Approaches	Employment			Student-Alum			Sports-Affiliation		
	P	R	F	P	R	F	P	R	F
Feature-based	84	72	77	88	75	81	75	76	75
Conventional Sequence Kernel	87	78	82	86	76	80	86	75	80
Semantic Sequence Kernel	93	76	83	91	78	84	84	77	81

Table 2 Relation Extraction Results (Physical)
表 2 关系抽取结果(Physical) %

Approaches	Near		
	P	*R*	*F*
Feature-based	83	75	79
Conventional Sequence Kernel	88	78	83
Semantic Sequence Kernel	90	83	86

Table 3 Relation Extraction Results (Personal-Social)
表 3 关系抽取结果(Personal-Social) %

Approaches	Business			Family		
	P	*R*	*F*	*P*	*R*	*F*
Feature-based	87	70	78	81	73	77
Conventional Sequence Kernel	86	84	85	87	81	84
Semantic Sequence Kernel	89	86	88	88	77	82

从结果可以看出新的核函数有一定的优势,这是在使用大训练集的情况下得到的结果. 为了验证新的核函数是否具有更好的泛化能力,进行第 2 阶段实验. 在本阶段实验中,训练集合的规模每次递减,随机抽取 100%,80%,60%,40%,20%,10%的训练样例来进行训练,然后记录各种算法的表现. 实验结果如图 3 所示:

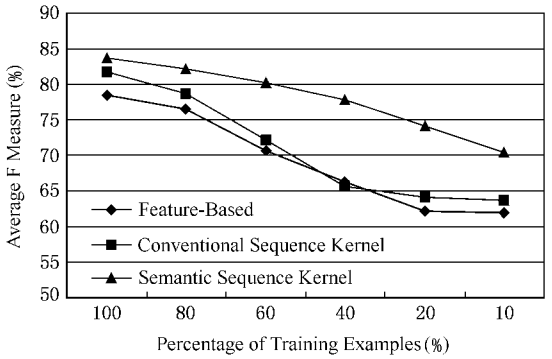

Fig. 3 Relation extraction results with different training sets.
图 3 不同规模训练集下的关系提取结果

实验证明,改进的语义序列核函数有更好的泛化能力. 即使是在只有 20%训练语料的情况下依然有比较好的效果. 其他两种方法在训练集合规模减小至 50%的时候准确率和召回率有明显的下降.

目前系统抽取结果中还存在一些错误,一部分错误是由于词性标注模块和实体识别模块未能正确标注词性和实体引起的,一部分错误来源于分类器本身. 例如,在上下文"李军 1996 年毕业于北京大学中文系后留校任教"中包含两个关系,一个是"李军"毕业于"北京大学",一个是"李军"在"北京大学"工作. 本系统目前只能抽取出第 1 个实体关系,而文中"留校任教"中的"校"指代"北京大学"没有找到,因此第 2 个关系没有正确抽取.

4 总结与展望

本文介绍了一个基于核函数的实体关系抽取系统,通过 HowNet 获取语义知识并嵌入核函数的计算过程中,通过对 ACE 定义的三大类中 6 个子类的实体关系抽取,实验证明,序列语义核函数可以大大提高关系抽取的精度,特别适合于小训练集的情况. 为了更好地利用实体间的关系,我们对二元关系进行了扩展,设计了自动抽取关系的描述信息,例如,雇佣关系(employment)中的职位,学习关系(student-alumni)中的学位,运动队员关系(sports-affiliation)中的比赛位置等. 整个系统框架由多个独立的模块构成,便于维护和改进,系统可以单独运行,也可以嵌入在 GATE 环境,以便实现多语种的关系抽取. 此外,系统的实现具有领域无关性和良好的可移植性,仅仅依靠少量的训练语料就能支持新关系的自动抽取.

基于核函数的方法相对其他方法速度是一个关键问题,如何寻找快速有效的语义计算方法是目前需要解决的问题,同时,对于目前抽取结果中还存在的各种错误,需要进一步改进,寻找有效的途径来提高实体识别,特别是指代的识别以及分类的准确率和召回率. 在我们后续的研究中,通过"雇佣"等关系的自动抽取,可以实现人物的识别和跟踪研究.

参 考 文 献

[1] C Aone, M Ramos Santacruz. Rees: A large-scale relation and event extraction system [C]. In: Proc of the 6th Applied Natural Language Processing Conference. New York: ACM Press, 2000. 76-83

[2] Che Wanxiang, Liu Ting, Li Sheng. Automatic entity relation extraction [J]. Journal of Chinese Information Processing, 2005, 19(2): 1-6 (in Chinese)
(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

[3] T Zhang. Regularized Winnow methods [C]. In: Advances in Neural Information Processing Systems (NIPS) 13. Cambridge: MIT Press, 2001. 703-709

[4] N Cristianini, J Shawe-Taylor, H Lodhi. Latent semantic kernels [J]. Journal of Intelligent Information Systems, 2002, 18(2-3): 127-152

[5] B Schölkopf, A Smola, K-R Müller. Kernel principal component analysis [G]. In: Advances in Kernel Methods: Support Vector Learning. Cambridge: MIT Press, 1999. 327-352

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

[7] M Collins, N Duffy. Convolution kernels for natural language [C]. In: Proc of Neural Information Processing Systems (NIPS) 14. Cambridge: MIT Press, 2001

[8] A Culotta, J Sorensen. Dependency tree kernels for relation extraction [C]. The 42nd Meeting of Association for Computational Linguistics, Barcelona, Spain, 2004

[9] J Suzuki, T Hirao, Y Sasaki, *et al*. Hierarchical directed acyclic graph kernel: Methods for structured natural language data [C]. The 41st Annual Meeting of the Association for Computational Linguistics (ACL-2003), Sapporo, Japan, 2003

[10] H Lodhi, C Saunders, J Shawe-Taylor, *et al*. Text classification using string kernels [J]. Journal of Machine Learning Research, 2002, 2: 419-444

[11] N Cancedda, E Gaussier, C Goutte, *et al*. Word-sequence kernels [J]. Journal of Machine Learning Research, 2003, 3: 1059-1082

[12] J Suzuki, H Isozaki, E Maeda. Convolution kernels with feature selection for natural language processing tasks [C]. The 42nd Meeting of Association for Computational Linguistics, Barcelona, Spain, 2004

[13] Zhao Shubin, Ralph Grishman. Extracting relations with integrated information using kernel methods [C]. The 43rd Annual Meeting of Association for Computational Linguistics, Michigan, 2005

[14] Che Wanxiang, Jiang Jianmin, Su Zhong, *et al*. Improved-edit-distance kernel for Chinese relation extraction [C]. The 2nd Int'l Joint Conf on Natural Language Processing (IJCNLP-05), Jeju Island, Korea, 2005

[15] Dong Zhendong, Dong Qiang. HowNet-Chinese Message Structure Bank [OL]. http://www.keenage.com, 2000-10-5/2006-3-18 (in Chinese)
(董振东,董强. 关于知网-中文信息结构库[OL]. http://www.keenage.com, 2000-10-5/2006-03-18)

[16] Liu Qun, Li Sujian. Word semantic similarity calculation based on Hownet [J]. Computational Linguistics and Chinese Language Processing, 2002, 7(2): 59-76 (in Chinese)
(刘群, 李素建. 基于《知网》的词汇语义相似度计算[J]. 计算语义学与中文信息处理, 2002, 7(2): 59-76)

[17] Kebin Liu, Fang Li, Ying Han, *et al*. Embedding the semantic knowledge in convolution kernels [C]. The 2nd Int'l Conf on Semantics, Knowledge and Grid (SKG'06), Guilin, 2006

Liu Kebin, born in 1981. Master candidate. His main research interests include natural language processing, information extraction, *etc*.

刘克彬,1981 年生,硕士研究生,主要研究方向为自然语言处理、信息抽取等.

Li Fang, born in 1963. Associate professor. Senior member of China Computer Federation. Her main research interests include natural language processing, information retrieval and extraction, Internet-based applications, *etc*.

李　芳,1963 年生,副教授,中国计算机学会高级会员,主要研究方向为自然语言处理、信息检索与抽取、基于 Internet 的应用等(fli@sjtu.edu.cn).

Liu Lei, born in 1980. Master candidate. His main research interests include natural language processing, information retrieval, *etc*.

刘　磊,1980 年生,硕士研究生,主要研究方向为自然语言处理、信息检索等(liu-lei@sjtu.edu.cn).

Han Ying, born in 1982. Master candidate. His main research interests include information retrieval, multi-document summary, *etc*.

韩　颖,1982 年生,硕士研究生,主要研究方向为信息检索、多文档摘要等(hanying@sjtu.edu.cn).

Research Background

Entity relation extraction (RE) is an important task in information extraction. It is the basis for event detection and tracking. Recently kernel methods with many learning algorithms such as the perceptron, SVM and KNN have been applied to solve many problems. Our research focuses on two points, one is to embed semantic knowledge into a kernel function; the other is to implement a Chinese relation extraction prototype based on the improved sequence kernel function. The program can run either as a Java application or plug-ins within Gate platform. This research is developed in the Joint Lab for Language Technology between Shanghai Jiaotong University and Saarland University of Germany. It is supported by the Committee of Science & Technology of Shanghai Municipal Government and Saarland University of Germany.