面向生物医学命名实体识别的多 Agent 元学习框架

王浩畅[1),2)] 李 钰[2)] 赵铁军[3)]

[1)](东北石油大学计算机与信息技术学院 黑龙江 大庆 163318)

[2)](哈尔滨工业大学生命科学与工程系 哈尔滨 150001)

[3)](哈尔滨工业大学计算机科学与技术学院 哈尔滨 150001)

摘 要 生物医学命名实体识别是生物医学数据挖掘的基本任务.文中提出了一种将多 Agent 系统和元学习方法相结合的多 Agent 元学习框架,应用于生物医学命名实体识别.基层多个学习 Agent 分别识别不同类型的生物医学命名实体,并通过相关学习 Agent 之间的通信来交换有益信息以调节个体 Agent 的行为提高其学习性能,元层 Agent 综合决策基层学习 Agent 的学习结果以获得最终的识别结果.元层 Agent 和基层学习 Agent 通过局部特征选择法选择适合不同实体类别的敏感特征集合提高了总体识别性能尤其是小类别识别的性能.文中提出的方法有效改善了传统的单一学习模型和全局特征选择方法不能兼顾各类别命名实体识别性能的不足.实验结果表明,文中提出的全新方法在生物医学命名实体识别上取得了优越的性能,在 JNLPBA2004 测试语料上获得了 77.5% 的 F 测度值.

关键词 命名实体识别;多 Agent 元学习框架;元层 Agent;基层学习 Agent;局部特征选择

中图法分类号 TP391 **DOI 号**: 10.3724/SP.J.1016.2010.01256

Biomedical Named Entity Recognition through a Multi-Agent Meta-Learning Framework

WANG Hao-Chang[1),2)] LI Yu[2)] ZHAO Tie-Jun[3)]

[1)](*College of Computer and Information Technology, Northeast Petroleum University, Daqing, Heilongjiang* 163318)

[2)](*Department of Life Science and Engineering, Harbin Institute of Technology, Harbin* 150001)

[3)](*School of Computer Science and Technology, Harbin Institute of Technology, Harbin* 150001)

Abstract Recognizing the biomedical named entity has become one of the most fundamental tasks in the biomedical knowledge discovery. A multi-Agent meta-learning framework is presented which incorporates multi-agent system and meta-learning method for the application of biomedical named entity recognition. In the base level, different learning agent is selected according to different classes of biomedical named entities. Through the communication between base learning agents, a base learning agent can get beneficial information from other related base learning agents and adjust its behavior so as to improve its learning performances. In the meta-level, the synthetic decision from the results of base learning agents is made by the meta-agent. Meta-agent and base learning agents are integrated with sensitive features set of corresponding named entity class according to local feature selection, which improve the system performance especially on minor classes. This approach effectively overcomes the disadvantages that only one model and global feature selection are used to identity all types of biomedical named entities. The experiments are carried on JNLPBA2004 test corpus with an F-sore of 77.5%. The results show that the brand-new multi-agent meta-learning framework is an effective approach and get promising results in biomedical named entity recognition.

收稿日期:2008-01-03;最终修改稿收到日期:2009-11-13.本课题得到国家"八六三"高技术研究发展计划项目基金(2006AA010108,2006AA01Z150)资助.E-mail: hcwang@mtlab.hit.edu.cn. **王浩畅**,女,1974 年生,博士,副教授,主要研究方向为自然语言处理、机器学习、生物信息学和信息抽取. **李 钰**,女,1962 年生,教授,博士生导师,主要研究领域为分子细胞生物学和遗传学. **赵铁军**,男,1962 年生,教授,博士生导师,研究领域为自然语言处理和人工智能应用.

Keywords　named entity recognition; multi-Agent meta-learning framework; Meta-agent; base learning agent; local feature selection

1　引　言

近年来，生命科学领域研究飞速发展，大量的生物医学知识以非结构化的形式被记载在各种形式的文本文件中. 从海量相关文献中直接获取本领域相关信息并使其变为生物学家可直接利用的知识，是一项迫在眉睫的任务. 以 MEDLINE 数据库为例，作为美国国家医学图书馆中规模最大、权威性最高的著名医学文献数据库，其文献总数目前已达到 1600 万，其中超过 300 万篇文献是近 5 年内出版的. 因此生物医学领域迫切需要有效方法对其中的知识进行挖掘.

生物医学文本知识挖掘的基本任务之一是生物医学文本中命名实体的识别. 其目的是从生物医学文本集合中识别出指定类型的名称，例如蛋白质、基因、核糖核酸、脱氧核糖核酸等. 这些获取的生物医学知识有着广泛的应用价值，例如疾病的诊断、预防和治疗.

生物医学领域命名实体识别是一项具有挑战性的研究. 其主要原因是：新的命名实体不断出现，形成了命名实体开放集；相同名称可能表示不同类别的生物医学命名实体，要依据上下文才能区分；很多生物医学命名实体拥有几个不同的名称. 此外，还存在着实体名称过长、复合词多、缩写词比例大、实体名称嵌套等现象. 目前，已经有很多技术应用到生物医学文本中的命名实体识别当中，大致可分为以下 3 类方法：基于启发式规则的方法[1-2]、基于字典的方法[3-4]和基于统计机器学习的方法.

当前最为普遍应用的是统计机器学习方法，其实现过程就是通过样本集建立适当的统计模型，以此模型对新的数据进行分类识别. 统计机器学习方法已经成为生物医学命名实体识别的一类重要方法. 在过去的研究中，大部分识别系统建立在单个学习模型的基础上，如贝叶斯模型[5]、隐马尔可夫模型[6]、支持向量机模型[7-8]、条件随机域模型[9]、最大熵模型等[10]. 这些系统通常选择相同的全局特征集合来识别所有不同类别的实体类型，而使用的训练语料普遍存在数据不平衡问题，即类别间训练样例的数量存在数量级的差距.

这样就导致这些系统存在以下缺点：单一的模型不能覆盖所有生物医学命名实体类型的特点，所以不能使所有的生物医学命名实体类型获得满意的识别效果；识别所有生物医学命名实体类型使用相同的特征集合，致使其中有些特征并不适合某一实体类型；由于小类别上的样例有限，分类器学习不够充分，因此在小类上的性能表现相对较差，严重影响了总体性能.

实际上不同的学习模型及通过调节模型参数对不同的生物医学命名实体类型有不同的识别效果，不同的实体类型有不同的敏感特征，因此我们可以选择不同的模型识别不同的实体类型并且通过局部特征选择法选择适合不同实体类别的敏感特征集合，通过综合决策不同模型的识别结果，这样可以提高每一类的识别性能尤其是小类别识别的性能.

随着分布式人工智能的发展，多 Agent 系统[11]的研究广泛地开展起来. 多 Agent 系统（Multi-Agent Systems，MAS）建立在资源共享和各个 Agent 的自主性之上，各个 Agent 能够协作工作，以实现整体识别的目标. 构建具有自学习能力的 Agent 是多 Agent 系统研究的一个重点，在这种需求下，为多 Agent 系统引入学习机制能够使其更好地适应复杂环境、有更强的个体学习能力和社会学习能力. 在过去的研究中，统计机器学习无论在应用、算法、理论都取得了令人瞩目的进步. 但是机器学习的研究一直独立于 Agent 的研究，最近才和 Agent 以及多 Agent 系统研究结合在一起[12-14]. Agent 以及多 Agent 系统可以看作是机器学习系统的另一个应用领域，持有这个观点的研究者或多或少地将机器学习算法直接应用于多 Agent 系统中的单个 Agent. 分布式人工智能与机器学习领域相互交叉、渗透，形成了多 Agent 系统学习这一新兴研究领域. 机器学习和多 Agent 系统研究的结合，对两个研究的领域都起到了推动作用.

元学习的概念是由 Prodromidis 等人于 2000 年首先提出的，该方法采用集成学习的方式来生成最终的全局预测模型[15]. 其基本思想是从已经获得的知识中进行再学习，从而得到最终的数据模式. 本文在多 Agent 系统和元学习理论基础上，提出一种新的基于多 Agent 元学习框架的策略来提高生物医学命名实体的识别的性能，这和多分布式系统中应用元学习的思想有相近之处[16-18]. 我们的系统定义并描述了多 Agent 元学习框架的元层和基层构成. 在基层中根据不同命名实体类型选择不同的基

层学习 Agent，并通过局部特征选择方法选择相应的敏感特征集合；基层相关学习 Agent 之间通过通信交换有益信息以调节个体 Agent 的行为提高其学习性能；在元层中通过元 Agent 综合决策多个基层 Agent 的学习结果，以获得最终的识别结果，从而大幅度提高了生物医学命名实体识别的性能.

本文第 2 节详细描述多 Agent 元学习框架以及基层 Agent 的通信方式；第 3 节介绍面向生物医学命名实体识别的多 Agent 元学习框架中各个 Agent 算法和特征的选择；第 4 节描述实验设置和实验结果，并对实验结果做了详细的分析；最后给出了结论.

2 多 Agent 元学习框架

元学习基本思想是从已经获得的知识中再进行学习，从而得到最终的数据模式. 在基学习阶段，各个结点可以自主地选择合适的学习算法来生成局部的基分类器. 与此同时，各结点间不存在任何通信. 在元学习阶段，系统可灵活采用各种集成策略，因此最终生成的元分类器具有较高的预测精度.

为了介绍元学习过程，约定以下符号. x 表示待分类的样本实例，给定 K 个学习算法 $L_k(k=1,2,\cdots,K)$，这些学习算法训练获得的 K 个分类模型表示为 $M_k(k=1,2,\cdots,K)$，每个分类模型对待分类实例 x 的预测分类结果表示为 $C_k(x)(k=1,2,\cdots,K)$，$class(x)$ 和 $attrvec(x)$ 分别表示正确的分类标识和实例 x 的特征属性向量.

给定数据集 $D=\{(class(x_i),attrvec(x_i)),i=1,2,\cdots,I\}$，随机将数据集分成 J 个大小基本相等的数据集 $D_1,\cdots,D_j$，定义 D_j 和 $D^{(-j)}=D-D_j$ 分别为 J 折交叉验证的第 J 折测试集和训练集. 给定的 K 个学习算法，称为第 0 层归纳算法. 在训练集 $D^{(-j)}$ 上训练第 k 个学习算法产生分类器模型 $M_k^{(-j)}$，对于模型 $M_k^{(-j)}(k=1,2,\cdots,K)$，则称为第 0 层模型，也称为基分类器. 对于测试集 D_j 中的每一个样本实例 x_i，$C_{ki}(x_i)$ 表示分类器模型 $M_k^{(-j)}$ 的预测.

在交叉验证过程结束后，集合从 K 个分类器模型的输出结果产生新的数据集 $D_{CV}=\{(class(x_i),C_{1i}(x_i),\cdots,C_{Ki}(x_i)),i=1,\cdots,I\}$，称为第 1 层训练集，也称为元训练集. 使用元训练集训练一个学习算法产生的学习模型 M_{Meta} 称为第 1 层模型，也称为元模型或元分类器，算法称为第 1 层归纳算法. 图 1 描述了 J 折交叉验证和元模型的生成过程.

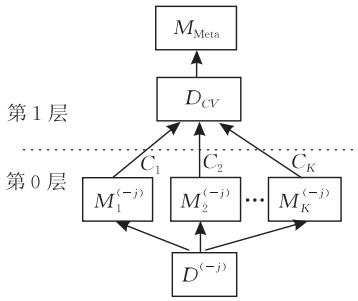

图 1 元学习 J 折交叉验证和元分类器的生成

对待分类的新样本实例进行决策时，首先通过第 0 层归纳算法在训练集 D 上进行训练学习产生模型 $M_k(k=1,2,\cdots,K)$，对给定的一个新实例，模型 M_k 生成一个预测向量 $C_1(x),\cdots,C_K(x)$，这个向量是元分类器 M_{Meta} 的输入，其输出为最终预测结果. 这种元学习方法也称为叠加法由 Wolpert[19] 提出.

由上述可知，在元学习方法的基学习阶段，各结点算法之间不存在任何通信，这样虽然节省了通信开销，却降低了彼此之间信息的参考和借鉴. Agent 之间的通信是多 Agent 系统的重要内容，也是多 Agent 系统的主要特色. Agent 间通过通信可以相互了解彼此的信息，及时调整各自的策略. 本研究将多 Agent 系统和元学习方法有效结合在一起，建立了多 Agent 元学习框架，提高了元学习的能力. 多 Agent 元学习框架由元层和基层构成. 基层中根据不同命名实体类型选择不同的基层学习 Agent 和敏感特征，并通过基层相关学习 Agent 之间的通信交换有益信息以调节个体 Agent 的行为，提高其学习性能. 本研究中的 Agent 的通信采用被动通信方式. 被动通信方式也称为消息传递方式. 其通信的原理属于直接通信方式下的消息缓冲通信机制. 被动方式的消息传递方式相对于主动方式更为灵活，它是实现灵活复杂的协调策略的基础. 在由自治 Agent 构成的多 Agent 系统中，Agent 之间经常需要协作来完成任务求解. 假设系统中有一个 Agent 集合和一组待求解任务，由于单个 Agent 无法独立完成某一任务或者通过多个 Agent 协作能提高求解效率或性能，Agent 之间可以通过协商形成 Agent 组，我们称之为联盟(coalition). 通常，系统中很多 Agent 会形成相对稳定的并且是成功的联盟. 那么在以后形成联盟的过程中，我们就可以以那些 Agent 集合为依据大大减少协商的盲目性，从而降低协商时的通信开销. 我们引入熟人集(Acquaintance)的概念. Agent K 的熟人集就是跟 Agent K 达成成功联盟超过一定频度的 Agent 集

合.本研究中 Agent 的通信采用基于熟人集的联盟,把需要通信的 Agent 互相请求成为熟人.只有熟人 Agent 之间才能进行通信,采用基于熟人集的联盟可以减少通信的开销.基层学习 Agent 只和容易发生识别冲突的 Agent 进行通信形成熟人集联盟,以消除识别冲突.在我们的识别任务当中,我们发现 DNA 类和 protein 类、DNA 类和 RNA 类、cell_type类和 cell_line 类经常发生识别冲突,识别上述类别的 Agent 彼此之间即确定为"熟人".通过启发式规则和统计策略的结合来消除识别冲突.启发式规则举例如下.

(1) IF r(Agent1)="protein" ∧ r(Agent2)="DNA" ∧ "X ligation" ∈ r(Agent1) THEN r(Agent2)≠"DNA",其中 r 表示识别结果.

(2) IF r(Agent2)="DNA" ∧ r(Agent3)="RNA" ∧ "X mRNA" ∈ r(Agent3) THEN r(Agent2)≠"DNA",其中 r 表示识别结果.

(3) IF c(r(Agent4)="cell-type")>c(r(Agent5)="cell-line") THEN r(Agent4)="cell-type" ∧ r(Agent5)≠"cell-line",其中 r 表示识别结果,c 表示在整个识别任务中识别出同一结果的次数.

多 Agent 的自治性使得不同的基层学习 Agent 可以识别不同类型的命名实体.每个 Agent 使用自己的知识和资源,即不同的实体类型使用不同的敏感特征.因而能够保证每个类别识别的性能.Agent 的社会性使基层学习 Agent 之间可以交换信息,彼此借鉴信息,从而提高识别的精确率和召回率.元层中元 Agent 有自己的知识和资源,基层学习 Agent 的输出结果作为元 Agent 输入的一部分,元 Agent 通过将基层学习 Agent 的学习结果和自身的知识和资源相结合,综合决策以获得最终的识别结果,进一步提高了系统的识别性能.针对我们识别 5 类生物医学命名实体的识别任务,多 Agent 的元学习框架如图 2 所示,任务分解 Agent 进行任务分解,基层共有 5 个学习 Agent 分别识别 protein、DNA、RNA、cell-type、cell-line 类生物医学命名实体,元 Agent 做综合决策以获得最终的识别结果.图中的虚线表示基层学习 Agent 之间的通信关系,当基学习 Agent 之间存在识别冲突时通过通信调节识别结果以提高基学习 Agent 的识别性能,这样也有利于元 Agent 综合决策结果的提升.

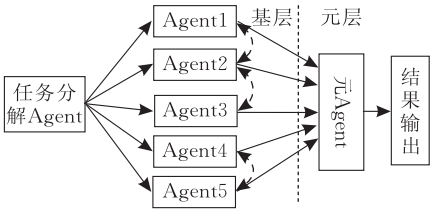

图 2　多 Agent 的元学习框架图示

3　算法和特征选择

3.1　算　法

在多 Agent 的元学习框架中,将每个分类模型视为一个基层学习 Agent 来识别不同类型的生物医学命名实体.综合决策结果取决于基层学习 Agent 的性能和独立性,即要求基层学习 Agent 有较低的错误率,至少比随机预测的结果要好,并且产生的错误相互独立.本文研究中,基层学习 Agent 使用了两种经典的学习算法,即 Generalized Winnow[20] 和条件随机域(CRFs)[21] 算法,这两种算法非常适用于自然语言处理研究中的分类问题.实验过程中,我们发现基于我们选择的特征 Generalized Winnow 算法对 protein 和 cell_type 类别识别的性能较好,而 CRFs 算法对 DNA、RNA 和 cell_line 类别的识别性能较好.元层中元学习 Agent 使用了 CRFs 算法作为综合决策算法,各个 Agent 的算法分配具体如表 1 所示.

表 1　多 Agent 功能及使用算法

Agents	功能	使用算法
Agent1	识别 protein 实体类型	Generalized winnow
Agent2	识别 DNA 实体类型	CRFs
Agent3	识别 RNA 实体类型	CRFs
Agent4	识别 cell_type 实体类型	Generalized winnow
Agent5	识别 cell_line 实体类型	CRFs
元 Agent	综合识别结果得到最终结果	CRFs

3.2　特征选择

特征选择的目的是寻找那些帮助识别和分类生物医学命名实体的文本属性.从以往的研究中我们发现生物医学命名实体识别系统都采用全局特征选择方法,即所有的类别都使用一个通用的特征选择过程,并且共享一个特征集合.这就导致存在以下几个问题:所有命名实体类型使用相同的特征,致使有些特征并不适合某一类型的命名实体;如果系统使用的特征类型多,会导致数据稀疏现象,占用更多的 CPU 资源,而使用的特征类型少又会降低系统的性能;由于大规模分类问题中普遍存在数据不平衡问题,即类别之间训练样例的数量存在数量级的差距,根据一般的特征选择函数进行全局特征选择得到的特征在类别上的分布并不均匀,大多集中于几个容易识别的类别中,主要原因是由于小类别上的样例

有限,分类器学习不够充分,就是对小类和难识别类别没有给予足够的重视[22],因此在小类上的性能表现相对较差,严重影响了总体性能.

本文研究使用 JNLPBA2004 语料进行训练和测试. 语料中的命名实体分为 5 类:DNA、RNA、protein、cell_line 和 cell_type. 表 2 为 JNLPBA 语料中各个实体类别的分布情况,从表 2 中我们可以看出 JNLPBA 训练语料中存在着数据不平衡的问题,最少的 RNA 类只有 951 个训练样例,最多的 protein 类有 30269 个训练样例,因此必然会引起分类器在小类别上学习不充分的局限性.

表 2 JNLPBA 训练语料各个实体类型的数量

实体类别	protein	DNA	RNA	cell_type	cell_line
数量	30269	9534	951	6718	3830

鉴于上述原因,为给予每一类别足够的重视,使用局部特征选择[22-23]的方法,分别对每一类别进行特征选择. 局部特征选择是指特征选择针对每个类别进行,不同的类别使用不同的特征集合. 本文按照局部特征选择的思想,采用基于启发式的方法进行特征选择,通过分析每类命名实体识别性能的变化得到不同类别命名实体的敏感特征集合. 所谓"敏感特征"是指对识别起到决定性作用的特征. 在大量的特征中,有可能只有几个特征对某类别命名实体的识别起到决定性的作用,而其它特征贡献非常小或者是多余的,这些多余的特征不仅占据内存空间,而且还影响查询的效率. 我们逐类构造最佳特征子集,选择不同生物医学命名实体类别对应的敏感特征集合,最大程度地保证了每个类别尤其是小类和难识别类的识别效果.

本文特征选择过程分为基层特征选择和元层特征选择:在基层 Agent 的特征选择过程中,使用局部特征选择法,分别对每一类别进行特征选择. 我们根据每类生物医学命名实体的特点构造其特征集合,每类实体类型包括以下一些各类别独有的特征:一元核心词特征、二元核心词特征、词法特征、不同类别的辞典特征等. 有一些特征是各类命名实体共有的特征,比如征词性特征、语块特征、标准化拼写特征及部分词形特征等.

在元层 Agent 的特征选择过程中,基层 Agent 的识别结果作为元层 Agent 的输入特征,这些识别特征也可以看作是元 Agent 的局部特征,除此之外我们还使用了各类别的一元核心词特征、二元核心词特征、词法特征、词性特征、语块特征、标准化拼写特征及部分词形特征. 关于各个 Agent 使用的不同特征集合描述详见表 3,关于每类特征的详细描述请参见文献[24].

表 3 特征选择

Agents	局部特征	共用特征
Agent1	protein 类的核心词特征、词法特征、蛋白质词典	词形特征、词性特征、语块特征、标准化拼写特征、上下文特征、缩写词特征
Agent2	DNA 类的核心词特征、词法特征	
Agent3	RNA 类的核心词特征、词法特征	
Agent4	cell_line 类的核心词特征、词法特征	
Agent5	cell_type 类的核心词特征、词法特征	
元 Agent	基层 Agent 的识别结果、各类别的一元核心词特征、二元核心词特征、词法特征、词性特征、语块特征、标准化拼写特征及部分词形特征	

4 实验与结果分析

本文研究使用 JNLPBA2004 语料进行训练和测试. JNLPBA 的训练语料由 GENIA 3.0 语料中的 2000 篇 MEDLINE 摘要组成,测试语料由当时未出版的 MEDLINE 摘要组成,共 404 篇. 语料中的命名实体分为 5 类:DNA、RNA、protein、cell_line 和 cell_type. 实验结果的评价标准是精确率(P),召回率(R)和 F 测度(F)评价. 本文使用了全部匹配模式对实验结果进行评测,即识别出的命名实体全部和正确答案的命名实体完全相匹配则认为是正确匹配. 实验设置中,J 折交叉验证取 J 值为 5.

为了验证局部特征选择法和多 Agent 元学习框架的有效性,我们把使用基于单一学习算法和全局特征选择方法的识别系统作为参照系统进行比较实验,参照系统使用了 CRFs 算法,并进行了缩写词识别,边界调整和嵌套识别后处理过程,具体参见文献[24],对比实验结果如表 4 所示.

表 4 实验结果对比 1

命名实体类型	多 Agent 元学习框架			基层 Agent 的识别结果			CRFs		
	P/%	R/%	F/%	P/%	R/%	F/%	P/%	R/%	F/%
Protein	74.4	83.5	78.7	73.2	80.5	77.1	72.7	78.9	75.7
DNA	78.9	71.3	74.9	75.4	71.1	73.2	68.9	77.1	72.8
RNA	72.4	75.4	73.9	70.7	73.1	71.9	67.8	67.8	67.8
Cell-line	64.9	67.6	66.2	62.8	66.1	64.4	60.2	62.0	61.1
Cell-type	86.8	71.5	78.4	84.7	70.6	77.0	80.2	73.2	76.5
Overall	76.7	78.3	**77.5**	74.9	76.2	**75.5**	72.9	76.3	**74.6**

从实验结果中可以看出,使用局部特征选择方法的基层 Agent 识别结果优于参照系统,并且未进行例如缩写词识别,边界调整和嵌套识别的等复杂的后处理过程,这说明我们的局部特征选择方法是行之有效的,并且我们发现每类的识别结果都有所提升,尤其是小类别 RNA 类和 Cell-line 类的识别

结果提升比较显著，RNA 类的识别结果 F 测度提升了 4 个百分点，Cell-line 类的识别结果 F 测度提升了 3 个百分点. 基于多 Agent 元学习框架的识别策略明显优于参照系统，并优于基层 Agent 识别结果，如 protein 类识别结果 F 测度提升了 1.6 个百分点，DNA 类的识别结果 F 测度提升了 1.8 个百分点，这也说明了通过基层 Agent 之间的通信和元 Agent 的综合决策，有效地提高了系统中各个类别的识别性能和总体性能. 系统最终结果 F 测度比参照系统性能高出近 3 个百分点. 表 5 列出了本文提出的方法和 JNLPBA2004 任务前 3 名系统的比较结果. 从中可以看出，本文方法比最好的系统结果高出近 5 个百分点.

表 5　实验结果对比 2

System	P	R	F
多 Agent 元学习框架	76.7	78.3	77.5
Zho[25]	69.4	76.0	72.6
Fin[26]	68.6	71.6	70.1
Set[27]	69.3	70.3	69.8

5　结　论

本文提出了一种基于多 Agent 元学习框架的的生物医学命名实体识别策略，将多 Agent 系统学习和元学习方法结合起来应用于生物医学命名实体识别. 使用不同的学习 Agent 和局部特征选择法选择不同的敏感特征集合识别不同类别的命名实体类型，克服了使用单一学习算法选择相同特征集合识别所有命名实体类型的缺点，提高每一类的识别性能尤其是小类别识别的性能；同时通过 Agent 之间的通信提高了 Agent 个体的识别性能弥补了元学习中基分类器互不通信的缺点，顶层决策使用元 Agent 能够利用不同学习 Agent 的决策结果. 最终实验结果表明本系统性能明显优于基于单分类器使用全局特征选择方法的识别系统.

在今后的研究工作中将研究更合理的任务分解策略及更合适的 Agent 学习模型，以进一步提高系统的性能.

参　考　文　献

[1] Fukuda K, Tamura A, Tsunoda T et al. Toward information extraction: Identifying protein names from biological papers//Proceedings of the Pacific Symposium on Biocomputing'98. Maui, Hawaii, USA, 1998: 707-718

[2] Seki K, Mostafa J. An approach to protein name extraction using heuristics and a dictionary//Proceedings of the 66th Annual Meeting of the American Society for Information Science and Technology. Long Beach, CA, 2003: 71-77

[3] Tsuruoka Y, Tsujii J. Boosting precision and recall of dictionary-based protein mame recognition//Proceedings of the ACL2003 Workshop on Natural Language Processing in Biomedicine. Sapporo, Japan, 2003: 41-48

[4] Krauthammer M, Rzhetsky A, Morozov P et al. Using BLAST for identifying gene and protein names in journal articles. GENE, 2000, 259(1-2): 245-252

[5] Tanabe L, Wilbur W J. Tagging gene and protein names in biomedical text. Bioinformatics, 2002, 18(8): 1124-1132

[6] Zhou G, Zhang J, Su J et al. Recognizing names in biomedical texts: A machine learning approach. Bioinformatics, 2004, 20(7): 1178-1190

[7] Kazama J, Makino T, Ohta Y et al. Tuning support vector machines for biomedical named entity recognition//Proceedings of the ACL2002 Workshop on Natural Language Processing in the Biomedical Domain. Philadelphia, PA, USA, 2002: 1-8

[8] Lee K, Hwang Y, Rim H. Two-phase biomedical NE recognition based on svms//Proceedings of the ACL2003 Workshop on Natural Language Processing in Biomedicine. Sapporo, Japan, 2003: 33-40

[9] Tsai T, Chou W, Wu K et al. Integrating linguistic knowledge into a Conditional Random field framework to identify biomedical named entities. Expert Systems with Applications, 2006, 30(1): 117-128

[10] Lin Y, Tsai T, Chou W et al. A maximum entropy approach to biomedical named entity recognition//Proceedings of the 4th ACM SIGKDD Workshop on Data Mining in Bioinformatics. Seattle, Washington, USA, 2004: 56-61

[11] Michael W. An Introduction to MultiAgent Systems. Chichester, England: John Wiley & Sons, 2002

[12] Huhns M, Weiss G. Special issue on MultiAgent learning. Machine Learning Journal, 1998, 33(2-3): 1-200

[13] Imam I F. Intelligent Adaptive Agents. Papers from the 1996 AAAI Workshop. Technical Report WS-96-04. Menlo Park, California, USA: AAAI Press, 1996

[14] Adaptation S S. Coevolution and Learning in Multiagent Systems. Papers from the 1996 AAAI Spring Symposium. Technical Report SS-96-01. Menlo Park, California, USA: AAAI Press, 1996

[15] Prodromidis A, Chan P, Stolfo S. Meta-learning in distributed data mining systems: Issues and approaches//Advances in Distributed Data Mining. Menlo Park, California, USA: AAAI Press, 2000: 81-114

[16] Hiroshi Ishikawa, Takashi Ozeki. A proposal on a model of an autonomous agent using the meta-level architecture//Proceedings of the International Conference on Integration of Knowledge Intensive Multi-agent Systems (KIMAS 2003). Boston, MA, USA, 2003: 83-87

[17] Bagherjeiran A, Vilalta R, Eick C F. Content-based image retrieval through a multi-agent meta—Learning framework//

Proceedings of the 17th IEEE International Conference on Tools with Artificial Intelligence. Hong Kong, China, 2005: 24-28

[18] Sun R. Meta-learning processes in multi-agent systems. Intelligent Agent Technology, 2001, 2(1): 210-219

[19] Wolpert D. Stacked generalization. Neural Networks, 1992, 5(2): 241-259

[20] Zhang T, Damerau F, Johnson D. Text chunking based on a generalization of Winnow. Journal of Machine Learning Research, 2002, 2: 615-637

[21] Lafferty J, McCallum A, Pereira F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data//Proceedings of the 18th International Conference on Machine Learning (ICML2001). Williamstown, MA, USA, 2001: 282-289

[22] Forman G. A pitfall and solution in multi-class feature selection for text classification//Proceedings of the 21st International Conference on Machine Learning (ICML2004). Banff, Alberta, Canada, 2004: 38-46

[23] Sebastiani F. Machine learning in automated text categorization. ACM Computing Surveys, 2002, 34(1): 1-47

[24] Wang H, Zhao T et al. A conditional random fields approach to biomedical named entity recognition. Journal of Electronics (China), 2007, 24(6): 838-844

[25] Zhou G, Su J. Exploring deep knowledge resources in biomedical name recognition//Proceedings of the Joint Workshop on Natural Language Processing in Biomedicine and Its Applications. Geneva, Switzerland, 2004: 96-99

[26] Finkel J, Dingare S, Nguyen H et al. Exploiting context for biomedical entity recognition: From syntax to the Web//Proceedings of the Joint Workshop on Natural Language Processing in Biomedicine and Its Applications. Geneva, Switzerland, 2004: 88-91

[27] Settles B. Biomedical named entity recognition using conditional random fields and novel feature sets//Proceedings of the Joint Workshop on Natural Language Processing in Biomedicine and Its Applications. Geneva, Switzerland, 2004: 104-107

WANG Hao-Chang, born in 1974, Ph. D., associate professor. Her research interests include natural language processing, machine learning, bioinformatics, and information extraction.

LI Yu, born in 1962, professor, Ph. D. supervisor. Her research interests include molecule cell biology and genetics.

ZHAO Tie-Jun, born in 1962, professor, Ph. D. supervisor. His research interests include natural language processing and applied artificial intelligence.

Background

With the explosion of information in the biomedical domain, there is a strong demand for automated biomedical information extraction techniques. Recognizing the biomedical named entity (BNE) such as proteins, DNAs, RNAs, cells etc. has become one of the most fundamental tasks in the biomedical knowledge discovery. Information extraction is a way to aid researchers in coping with information overload with the technologies of statistical learning and automatic text information processing are maturing gradually. While many algorithms have been proposed for this task, biomedical named entity recognition (BNER) remains a challenging task and an active area of the research.

There have been many attempts to develop techniques to identify NE in the biomedical literature. They roughly fall into three approaches, that is, heuristic rule-based approach, dictionary-based approach, and statistical machine learning-based approach. However, the state-of-the-art techniques for BNER do not achieve satisfactory results. The current achievements of BNER have a distance to real application and the performance of system cannot satisfy the practical requirements. The problem suggests that individual BNER system may not cover entity representations with enough rich features and no single type of algorithm is practical to achieve the best performance. This paper aims to find effective methods to improve the performance of BNER.

The research in this paper covers BNER techniques based on a multi-agent meta-learning framework which incorporates multi-agent system and meta-learning method. This approach effectively overcomes the disadvantages that only one model and global feature selection are used to identity all types of biomedical named entities. The experimental results have proved that the brand-new multi-agent meta-learning framework is an effective method and get promising results in BNER. The method copes with the problem of corpus imbalance to improve the recognition performance for mini class and difficult-to recognize class.

This work was supported by the National High Technology Research and Development Program (863 Program) of China (Nos. 2006AA010108, 2006AA01Z150), which is related to nature language processing and information extraction. BNER is a critical issue in the biomedical Information extraction, so the research in this paper is very important and meaningful.

The main authors of this paper have gained sufficient archivement in the research of biomedical information extraction. Some related papers have been published in some academic journals and some international conferences.