中文维基百科的实体分类研究

徐志浩[1,2]，惠浩添[1,2]，钱龙华[1,2]，朱巧明[1,2]

（1. 苏州大学 自然语言处理实验室，江苏 苏州 215006；

2. 苏州大学计算机科学与技术学院，江苏 苏州 215006）

摘　要：维基百科实体分类对自然语言处理和机器学习具有重要的作用。本文采用机器学习的方法对中文维基百科的条目进行实体分类，在利用维基百科页面中半结构化信息和无结构化文本作为基本特征的基础上，结合中文的特点使用扩展特征和语义特征来提高实体分类性能。在人工标注的语料库上的实验表明，这些额外特征有效地提高了 ACE 分类体系上的实体分类性能，总体 F1 值达到 96%，同时在扩展实体分类上也取得了较好的效果，总体 F1 值达 95%。

关键词： 维基百科；实体分类；半结构化信息；信息框

中图分类号：TP319　　**文献标识码**：A

Classifying Named Entities on Chinese Wikipedia

XU Zhihao[1,2]，HUI Haotian[1,2]，QIAN Longhua[1,2]，ZHU Qiaoming[1,2]

（1. Natural Language Processing Lab of Soochow University，Suzhou，Jiangsu 215006，China

2. School of Computer Science & Technology，Soochow University，Suzhou，Jiangsu 215006，China)

Abstract: Classifying Wikipedia Entities is of great significance to NLP and machine learning. This paper presents a machine learning based method to classify the Chinese Wikipedia articles. Besides using semi-structured data and non-structured text as basic features, we also use Chinese-oriented extended features and semantic features in order to improve the classification performance. The experimental results on a manually tagged corpus show that the additional features significantly boost the entity classification performance with the overall F1-measure as high as 96% on the ACE entity type hierarchy and 95% on the extended entity type hierarchy.

Keywords: Wikipedia; named entities classification; semi-structured data; Infobox

1 引言

维基百科作为一个开放的知识库系统，其中的条目都是对一个概念或者实体的内容描述，每个条目的页面中包含了丰富的结构化、半结构化的信息和文本资源。维基百科实体分类是指对维基百科中的条目进行识别和分类，从中提取出各种类型的实体（如人物、组织、地名等）。对于这些实体的分类有助于进一步从维基百科中挖掘出更丰富的信息（如实体关系、语义关系等），同时维基百科中丰富的文本也为自然语言处理和机器学习提供了高质量的语料来源[1-2]。

2 相关工作

对维基百科条目进行实体的识别和分类，目前主要有两种方法：基于启发式规则的方法

收稿日期： 2015 年 5 月 31 日　　**定稿日期**：2015 年 8 月 6 日
基金项目：国家自然科学基金(61373096，90920004)，江苏省高校自然科学研究重大项目(11KJA520003)

和基于机器学习的方法。早期的方法主要是基于规则，如 Bunescu 和 Pasca[3]利用了标题首字母大写等一系列规则来识别英文维基百科的某个条目是否是一个命名实体。Zirn 等 [4]进一步利用分类框(Category)中心词的单复数形式这一规则，他们认为如果类别中心词是以单数形式出现的，这个中心词就是一个实体。Toral 等 [5]则首先提取条目摘要中的第一句（称为定义句），并找出句中所有名词在 WordNet 中的语义层次及类别来帮助确定条目所属的实体类别。基于规则方法的缺点是缺乏灵活性，需要对不同的实体类型制定不同的规则，并且随着规则的增多，不同规则之间可能会产生冲突。

利用机器学习来进行实体识别和分类可以克服这一缺点。Bhole[6]在维基百科条目文章的第一段和全文文本上，分别利用词包（bag-of-words）模型，使用 SVM 进行条目的实体分类工作。Tardif 等[7]将维基百科的摘要文本作为基本特征，并使用了分类框、信息框(Infobox)和模板(Template)等内容作为额外特征。Dakka 和 Cucerzan[8]则将条目中的词汇、结构化信息（如表格）、摘要等内容作为特征进行组合，来获得最好的分类效果。在 Tardif 和 Dakka 的实验中，都对比了使用 SVM 分类器和朴素贝叶斯分类器的实验结果，他们的实验结果都表明 SVM 的分类性能更好。

上述工作都是针对英文维基百科上的实体识别和分类，目前还没有中文维基百科上的实体分类工作。虽然和英文维基百科相比，中文维基百科的容量要小得多，但它对中文自然语言处理的潜力还没有被充分挖掘出来，相关的工作也比较少 [9-10]。因此，对中文维基百科的条目进行实体识别和分类具有一定的研究价值。本文在传统特征的基础上，提出了一系列针对中文特点的有效特征，使用 SVM 分类器进行中文维基百科的实体分类，取得了较好的结果。

3 维基百科页面格式

维基百科中每个条目都是对一个概念或实体的描述，条目的内容由网络志愿者协作编撰，任何使用互联网的用户都可以编写和修改维基百科条目的文章内容。在编写过程中，用户须遵循维基百科的格式要求。图 1 为一个典型的维基百科页面格式，它具有丰富的半结构化信息和非结构化文本，其主要内容有：

1. 信息框（Infobox）：信息框模板是一个总结性的提纲列表，总结了与条目相关的主题，亦或包含图像、地图等信息。信息框中内容的格式为标签（label）与数据（data），例如“马云”这个人物条目的信息框中有“出生　1964 年 9 月 10 日”、“国籍　中华人民共和国”、“母校　杭州师范大学”、“职业　阿里巴巴集团董事局主席”等与主题相关的信息。

2. 页面分类（Category）：页面分类中列出了条目所属的类别，以及突出条目事物特征或是主题的相关类别。一个条目可以被分类到多个类别下，需要注意的是，该分类体系并非严格的层次体系，具有一定的随意性。例如“马云”这个条目的分类有“1964 年出生”、“在世人物”、“中国企业家”、“杭州人”、“阿里巴巴集团”等。

3. 摘要（Abstract）：摘要是指某个维基百科条目文章的第一段，其内容以简明扼要的文句给出该条目的主要信息内容。摘要中的第一句，往往会有类似“……是……”或“……为……”等句式，我们把这样的句子称为显式定义句，也会有不出现“是”或“为”的隐式定义句。定义句中的中心词，很有可能反映出条目所属的类别。例如“马云”这个条目的定义句为“马云（英文名：Jack Ma，1964 年 9 月 10 日－）中华人民共和国企业家”，其中心词为“企业家”，可以推断出，该条目的类别是人物。

图 1　维基百科页面格式

4 基于 SVM 的实体分类

与传统机器学习的分类方法类似，本文将人工标注类别的维基百科条目分为训练集和测试集，从中提取各种特征，利用词包模型，构造相应的特征向量，然后使用 SVM 分类器从训练集的特征向量中学习得到分类模型，最后将该分类模型应用到测试集的特征向量上，预测条目的实体类别，并计算分类方法的性能。基于机器学习方法的关键在于找出有效的特征来表示维基百科中的条目，本文除了使用维基百科页面中获取的基本特征之外，还使用了一些扩展特征和语义特征来帮助提高中文维基百科的实体分类性能，详见表 1。

表 1　维基百科中的实体分类特征

特征类别	特征名称	特征描述
基本特征	InfoboxTitle	信息框中的属性标题
	CategoryHead	分类框中的中心词
	AbstractHead	摘要的中心词
扩展特征	IsChineseName	标题首字是否是姓氏且标题长度为 2-4 个字符
	TitleContainsPeriod	标题中是否含有符号“• ”
	TitleLastChar	标题的最后一个字
	TitleLastWord	标题的最后一个词
语义特征	CategoryHeadTyccl	分类框中心词的同义词词林代码（前 4 位）
	AbstractHeadTyccl	摘要中心词的同义词词林代码（前 4 位）

一、基本特征

本文使用了以下三个类别的基本特征，即信息框、分类框和摘要中的相关内容，具体如下：

1. InfoboxTitle：信息框中的内容对于实体类型具有很好的识别作用。信息框中的信息形式为“标签　数据”，我们提取其中的标签的内容作为一个特征，而不提取数据本身。例如对于“国籍　中华人民共和国”，取“国籍”作为特征，因为不同的人物，对应的国籍是不同的，而“国籍”这个标签是共同拥有的。例如对于“马云”这个条目，从其信息框中提取到的特征词为分别为“出生”、“国籍”、“母校”、“职业”、“净资产”、“配偶”和“子女”，这些特征词基本都是人物的相关信息。

2. CategoryHead：分类框中的信息对实体分类同样具有明显的识别作用。对于每一个类别，通过分词处理后，取其中心词（即最右边一个词）作为特征。例如“1964年出生”，通过分词取得中心词“出生”作为一个特征。因此，“马云”这个条目的分类框中得到的特征词分别为“出生”、“人物”、“企业家”、“亿万富豪”、“领袖”、“校友”、“教师”、“人”、“姓”、“人士”、“博士”和“集团”等。

3. AbstractHead：除了上述半结构化信息外，在维基百科的文章中的第一段（即该条目的摘要）也可起到一定的补充作用。对于摘要的处理，我们取其第一句，通过分词和词性标注，找出第一句的中心词（最右边的名词）作为特征。特别地，当第一句的句式结构为“……是……”或“……为……”时，更能通过正则匹配轻松获得该句中心词。例如，从“马云”这个条目的摘要中提取到的特征为“企业家”。

二、扩展特征

为了更好地对某些类别（特别是人名、地名、组织名等）的实体进行识别，我们加入了下面有关条目标题的扩展特征。前两个特征是用来帮助提高人物类别的分类性能，而后两个特征对所有实体类别均有效。

1. IsChineseName：加入了中文百家姓姓氏列表，将条目名的第一个或前两个字是否属于姓氏并且条目标题长度在 2 到 4 个字符为一个二元特征。

2. TitleContainsPeriod：标题是否含有分隔符号。维基百科的外国人名的条目，标题中会使用 “• ”分隔外文姓氏和名字，因此将标题中是否含有分隔符作为一个二元特征。

以上两个特征的加入，用来帮助提高 Person 类别的分类性能。

3. TitleLastChar：考虑到某些命名实体在名称上的特殊性，比如地名中“XX 省”、“XX 市”、“XX 县”，机构名中“XX 局”、“XX 部”，最后一个字有极高的规律性。因此通过加入条目标题的最后一个字和词作为两个特征，来帮助提高 ORG、GPE 等实体类别的分类性能。

4. TitleLastWord：某些实体名如“XX 协会”、“XX 大学”，“XX 山脉”等，最后一个词具有很强的规律性，因此通过加入标题的最后一个词作为特征，来帮助这类实体的分类。

三、语义特征

由于维基百科由网民以共享合作方式撰写，因此对于同一个或者类似的含义，可能会用

不同的词进行表达，如：“警察”、“警务人员”、“警官”都表达类似的含义，都指向人物这个类别，导致了特征词稀疏问题。因此，有必要在基本特征中对表达类似概念的词汇进行泛化，方法是引入了同义词词林，将特征词汇的语义代码作为一个特征加入到系统中。

《同义词词林》[11]是一部汉语分类词典，其中每一条词语都用一个编码来表示其语义类别。本文所用的《词林》为《词林(扩展版)》，是哈工大信息检索研究室在《同义词词林》的基础上研制的。最终的词表包含 77492 条词语，共分为 12 个大类，94 个中类，1428 个小类，小类下再以同义原则划分词群，最细的级别为原子词群。不同级别的分类结果可以为自然语言处理提供不同颗粒度的语义类别信息，本文选取词林语义代码的第二级和第三级（即语义代码的前 2 和前 4 位）进行实验。

5 实验

5.1 数据来源

实验中所使用的维基百科数据来自于维基百科网站上下载的 2014 年 8 月 4 日中文离线数据包。首先需要将原有数据包文件中的 XML 标记去除，保留所需要的文本内容。由于维基百科的内容中混合了繁体和简体中文，为了便于后期处理，需要将所有中文统一转化为简体，最后从中提取出每个条目的标题、信息框、分类框和摘要等相关信息。其中，对摘要的首句使用进行分词和词性标注。

我们从所有条目中随机取出 8000 个条目作为实验数据，通过规则匹配去除消歧页面和列表页面后，剩下 7612 个条目，然后根据 ACE 的中文命名实体的分类体系对条目进行类别的标注。

实验所使用的实体分类体系，是在 ACE[12]定义的中文命名实体分类基础上，结合 Sekine 的扩展命名实体分类体系 [13]，考虑到实际信息抽取的需要进行设置的。其中，PER、ORG、GPE、LOC 和 FAC 等为 ACE 定义的五大类实体，其余 9 类为扩展类别。如非特别指出，下列实验中的实体分类是指 5 类 ACE 实体，其余都为非实体；而扩展实体分类时，14 类为实体类别，其余为非实体。

5.2 实验设置

所有实验都按照 5 折交叉验证方式进行，即实验数据被随机分成大小相同的五份，训练集和测试集的比例为 4:1，使用的分类工具为 LibSVM，且 SVM 的训练参数均采用默认值。实体分类结果分别使用准确率（P）、召回率（R）和调和平均值（F1）进行评估，最后取五次实验的平均值作为最终结果。

5.3 实验结果

一、各个特征对分类性能的影响

为了考察各个特征对分类性能的影响，本文分别进行了加入和分离实验，前者以信息框和分类框特征为基准系统，然后单独加入每个特征，比较它和基准系统之间的性能差异；而后者是以所有特征为基准系统，然后分离出单个特征，比较它和基准系统之间的性能差异。实验结果如表 2 所示，其中性能差异用 P/R/F1 的变化值来表示，每一列中性能变化的最大值用粗体表示，加入实验的正值表示该特征是有益的，而分离实验的负值表示该特征是有效的。为便于参考，表格的第 1 行列出了两个基准系统的 P/R/F1 性能。

表 2　加入和分离实验中不同特征的性能影响

特征	单独加入			单独分离		
	P（%）	R（%）	F1（%）	P（%）	R（%）	F1（%）
基准系统	87.69	90.70	89.17	97.07	95.03	96.03
AbstractHead	+9.53	+1.98	+5.73	-0.44	-0.12	-0.15
IsChineseName	+6.50	-6.69	-0.36	-0.07	-0.20	-0.14
TitleContainsPeriod	+0.35	0	+0.18	-0.01	-0.08	-0.04
TitleLastChar	**+9.07**	**+2.30**	**+5.68**	**-0.31**	**-2.02**	**-1.19**
TitleLastWord	+3.59	-2.08	+0.76	-0.28	-0.22	-0.25
CategoryHeadTyccl	+1.44	+1.29	+1.37	+0.66	-0.65	-0.01
AbstractHeadTyccl	+7.63	+1.35	+4.49	+0.18	+0.02	+0.08

从表 2 可以看出，各个特征加入实验时的性能贡献比分离实验时的性能贡献要大得多，这是由于特征之间往往存在着冗余性，单独使用时性能提升很明显，而同时使用则效果不显著，此外：

1. 贡献最大的特征是 TitleLastChar，无论是加入还是分离，都对准确率和召回率有明显的影响，这主要是由于条目标题的最后一个字对不同类别具有很高的区分性，特别是对于 GPE 类，如“XX 省”,“XX 市”等，标题最后一个字具有很强的区分性。同样 TitleLastWord 特征的贡献也很稳定，虽然没有 TitleLastChar 特征那么大；

2. 特征 AbstractHead 在加入实验中的作用很明显，但在分离实验中的变化要小得多，这可能是由于该特征本身很有用，但它和其它特征之间具有一定的冗余性；

3. 两个人名特征的效果并未达到预期值。特征 IsChineseName 的加入提高了准确率，但同时召回率也明显降低。这是由于不少 GPE 条目的首字母也是中文姓氏，与部分人名产生混淆。不过，虽然它在加入实验时降低总体性能，但在分离实验时却表现出对总体性能略有帮助。同样，特征 TitleContainsPeriod 对分类性能也有提高。

4. 两个语义特征的表现不一致。特征 CategoryHeadTyccl 的贡献比较稳定，无论是加入还是分离实验，都表现出对提高性能的有效性。而特征 AbstractHeadTyccl 的表现就不一致，尽管在加入实验中提高了总体性能，但在分离实验中删去该特征反而提高了总体性能，可以认为该特征过于泛化。

二、不同类别的性能比较

根据上述分离实验中各特征的性能表现，最后确定使用除 AbstractHeadTyccl 以外所有的其他特征，得到最好的分类性能如表 3 所示。

表 3　不同类别的分类性能

类别	含义	实例数	百分比(%)	P（%）	R（%）	F1（%）
PER	人物	1503	19.75	98.17	96.21	97.18
ORG	组织	395	5.19	95.03	87.09	90.89
GPE	地理政治实体	1804	23.70	**98.01**	**98.17**	**98.09**
LOC	地点	633	8.32	95.13	92.58	93.84
FAC	设施	611	8.03	96.34	90.34	93.24
NON	非实体	2667	35.03	/	/	/
平均	/	7612	100	97.24	95.01	96.11

从表 3 可以看到，系统最终取得的分类性能还是较高的，平均 F1 值超过了 96%。其中，性能最高的两个类别为 PER 和 GPE，这是由于这两种类型的实例数较多且其条目的特征有较高的一致性，因此在 SVM 中得以比较好的训练。而性能相对较低的三类为 ORG、LOC 和 FAC 等，F1 值分别约为 91%、94%和 93%，且是召回率明显低于准确度，这是因为这三个类别的条目种类形态较多而样例又较少，无法得到充分的训练，另外这三个类别下，很多没召回的条目往往是 Category 和 Abstract 中能提取的特征较少或是有噪声，而标题中提取的特征词又很稀疏，最后由于没有提取到有效特征导致无法召回，例如条目“日本邮政公社”，其摘要和 Category 中获取到的特征词分别为“体”和“邮政”、“事业”，而标题尾词“公社”在训练样例中又属于稀疏的词，导致其无法召回为 ORG。

三、扩展实体类别的分类性能

表 4 列出了在 14 个扩展实体类别上的分类性能（使用的特征集与表 3 相同）和每个类别的实体数量及所占比例，表中除 ACE 实体类别外最高的 P/R/F1 性能用粗体标出。

表 4 扩展类别上的分类性能

实体类别	含义	数量	所占比例（%）	P(%)	R(%)	F1(%)
Person	人物	1503	19.75	97.38	96.41	96.89
Organization	组织	395	5.19	94.28	87.59	90.81
GPE	地理政治实体	1804	23.70	97.95	98.17	98.06
Location	地点	633	8.32	94.99	92.89	93.93
Facility	设施	611	8.03	96.37	91.16	93.69
Hardware	硬件	32	0.42	86.96	62.50	72.73
Software	软件	48	0.63	88.57	64.58	74.70
Game	电子游戏	39	0.51	96.97	82.05	88.89
Work of Art	艺术作品	405	5.32	95.34	85.93	90.39
Drug	药品	10	0.13	80.00	40.00	53.33
Award	奖项	15	0.20	87.50	46.67	60.87
Animal	动物	402	5.28	93.98	97.01	95.47
Flora	植物	456	5.99	**99.08**	**94.30**	**96.63**
Disease	疾病	12	0.16	100	75.00	85.71
NON	非实体	1247	16.38	/	/	/
平均	/	7612	100	96.63	94.31	95.45

从表 4 可以看出，扩展至 14 个实体类别后的 P/R/F1 平均值为 96.63%/94.31%/95.45%，与 5 个实体大类的分类性能相比虽有降低，但幅度较小，这主要是由于非 ACE 的实体类别数量较少，占总数比例小于四分之一。对非 ACE 的九个实体类别，各个类别的 F1 值和其条目的数量，大致上呈现一个线性关系。即由于训练样例太少，从而导致特征稀疏，召回率下降，因此分类性能不尽理想，进一步分析发现：

1. Work of Art、Animal 和 Flora 三个类别与 ACE 中的 ORG 实例数量接近，其中 Work of Art 的性能和 ORG 相当，因为 Work of Art 中包括了电影、音乐、书籍等多种艺术形式，因此特征较为多样化，而相比之下实例数较少，因此无法对特征进行很好的学习，导致召回率较低。Animal 和 Flora 两类的性能相比 ORG 明显高，因为动物和植物的实例在特征上较为一致，都包含“属”、“种”、“动物”、“植物”等特征词，但由于这两类的特征很相似，因此错分的实例主要集中在这两类之间互相分错。

2. Game 和 Disease 这两个类别尽管数量不多（前者不到 40，后者略大于 10），但 F1 性能都在 85%以上，这是由于它们的特征虽然数量少但较为一致。如 Game 类实体中均含有“游戏”“开发商”“平台”等特征词；而 Disease 类的 Category 中都有“疾病”这个特征词。

四、与英文维基的实体分类性能比较

为了考察不同语言之间维基实体分类的难度，本文比较了中英文维基实体的分类性能。英文维基的实体分类中比较典型的是 Tkatchenko 等 [14]的研究工作。他们总共划分了 18 个实体类别，本文共划分 14 个实体类别，两者共有的类别共有 9 个，因此本文选取了中英文共有且实例数量较多的类别进行比较，结果如表 5 所示。

表 5 中文和英文维基实体分类性能的比较

实体类别	本文（SVM）			Tkatchenko（SVM+规则）		
	P	R	F1	P	R	F1
PER	0.97	0.96	0.97	0.95	1	0.98
ORG	0.94	0.88	0.91	0.95	0.98	0.96
GPE	0.98	0.98	0.98	0.98	1	0.99
LOC[1]	0.95	0.93	0.94	1	0.95	0.97
				1	0.77	0.87
FAC	0.96	0.91	0.94	1	0.96	0.96
Work of Art	0.95	0.86	0.90	0.93	0.98	0.95
Animal	0.94	0.97	0.95	1	1	1
Flora	0.99	0.94	0.97	0.98	1	0.99

需要指出的是，两者所使用的分类体系和数据集不一样（英文中使用 18 个类别，5294 个条目，本文使用 14 个类别，7612 个条目），不过，我们还是可以看出，英文维基百科的扩展实体分类性能整体上都优于中文。在 PER 和 GPE 两个类别上，中英文的性能旗鼓相当；而在其它类别上，两者之间的分类性能还有相当差距。可能的原因是中文的 PER 在 Category 上的特征一致性较高，GPE 在标题特征上一致性较高，另外这两类的训练样例数量相对较多，因此得到了比较理想的分类性能，而相比之下，中文的 ORG、LOC 和 FAC，样例的形态较为多样，另外训练样例又较少，导致部分特征较为稀疏。

由于中文和英文在形态和语法上的区别，使得在英文中使用的很好的特征和规则，在中文上未必有效。比如在 Bunescu 和 Pasca 的论文中使用的首字母大写这一规则来判断某个条目是否属于实体就无法在中文中使用；在 Tkatchenko 的论文中，在对实体分类前，通过使用一系列规则对实体与非实体进行二元分类，精度和召回率都达到了 95%。另外由于受到中文分词技术的限制，在提取 Category 和 Abstract 中心词时会出现一些错误和偏差，导致噪声的引入，影响分类性能，而在英文中，就不存在这样的分词问题。

此外，英文维基百科的发展比中文维基百科的发展更好，其在内容的正确性和完整性上都优于中文维基百科。我们观察到，未能召回的中文条目，很大一部分条目的页面内容十分少，并缺乏相应的 Category 和 Infobox 等半结构化信息，导致无法提取到这些条目的有效特

[1]由于本文的使用的实体分类体系和 Tkatchenko 论文的分类体系有所不同，Tkatchenko 论文中 ASTRAL_BODY 和 GEO_REGION 两个类别为本文类别 LOC 的两个子类，故在对比时，将本文 LOC 的性能与其两类的性能作比较。

征，从而无法对这一部分条目进行正确分类。

5.4 错误分析

为了进一步了解产生分类错误的原因，本文随机选取了 100 个错分的维基条目进行分析，发现分类错误原因主要有以下几个类别：

1. 分类框信息不规范：维基条目的分类框内容并非完全都是条目所属的严格意义上的某个类别，还包括与条目相关的类别。例如条目“世界新闻自由日”的分类框中有“联合国教科文组织”，得到特征词“组织”，此特征导致条目被错分为 ORG。这部分错误占总数的 44%；

2. 标题名称的不确定性：某些类别的条目标题和其它类别的条目标题特征相似，从而产生误导。例如，条目“赫尔曼·凯斯滕奖”和“曹洞宗”被错分为 PER 类别，但实际上它们只是含有 PER 类别的某些特征。这类错误占总数的 30%。

3. 类属条目和实体条目的相似性：所谓类属条目是对某一实体类别的描述，因而在特征上与实体条目相似。如条目“皇上”、“动作片演员”这类称谓、职业类条目易被错分为 PER 类别。这类错误占总数的 13%；

4. 其它较为个别或者无法明确归类的错误，约占总数的 13%。例如语言中存在着一词多义现象，因此多义词作为一个统一的特征时，容易引起错误。比如“组织”这个词，可能属于“机构”这个概念，也可能属于“生物体”的概念。

6 结论

本文利用维基百科条目中的半结构化信息作为特征，并根据中文实体的特点加入扩展特征和语义特征，从而对中文维基百科条目进行实体分类。实验表明，这些特征可以有效提高维基实体分类的性能。其中对于 ACE 实体类别的分类性能 F1 值超过 96%，达到了实用价值；而对于扩展实体类别，则还需要通过标注更多的实例来提高实例数较少的类别的分类性能。

目前的方法都是基于词汇层面，还未考虑到句法和语义层面，因此今后的工作一方面可考虑挖掘句法和语义特征，以进一步提高分类性能。另一方面，可利用该分类模型对所有的维基百科条目进行实体分类，并将这些识别出的命名实体应用到自然语言处理的其它任务中。

参考文献

[1] Nothman J, Curran J R, Murphy T. Transforming Wikipedia into named entity training data[C]//Proceedings of the Australian Language Technology Workshop. 2008: 124-132.

[2] Nothman J. Learning named entity recognition from Wikipedia[D]. The University of Sydney Australia 7, 2008.

[3] Bunescu R C, Pasca M. Using Encyclopedic Knowledge for Named entity Disambiguation[C]//EACL. 2006, 6: 9-16.

[4] Zirn C, Nastase V, Strube M. Distinguishing between instances and classes in the wikipedia taxonomy[M]. Springer Berlin Heidelberg, 2008.

[5] Toral A, Munoz R. A proposal to automatically build and maintain gazetteers for Named Entity Recognition by using Wikipedia[J]. NEW TEXT Wikis and blogs and other dynamic text sources, 2006, 56.

[6] Bhole A, Fortuna B, Grobelnik M, et al. Extracting named entities and relating them over time based on

wikipedia[J]. Informatica (Slovenia), 2007, 31(4): 463-468.

[7] Tardif S, Curran J R, Murphy T. Improved text categorisation for Wikipedia named entities[C]//Australasian Language Technology Association Workshop 2009. 2009: 104.

[8] Dakka W, Cucerzan S. Augmenting Wikipedia with Named Entity Tags[C]//IJCNLP. 2008: 545-552.

[9] 谌志群, 高飞, 曾智军. 基于中文维基百科的词语相关度计算[J]. 情报学报, 2013, 31(12): 1265-1270.

[10] 张苇如, 孙乐, 韩先培. 基于维基百科和模式聚类的实体关系抽取方法[J]. 中文信息学报, 2012, 26(2): 75-81.

[11] 梅家驹. 同义詞詞林[M]. 上海辞书出版社, 1983.

[12] ACE (Automatic Content Extraction), Chinese Annotation Guidelines for Entities, Linguistic Data Consortium, Version 5.5, 2005.

[13] Sekine S, Sudo K, Nobata C. Extended Named Entity Hierarchy[C]//LREC. 2002.

[14] Tkatchenko M, Ulanov A, Simanovsky A. Classifying Wikipedia entities into fine-grained classes[C]//Data Engineering Workshops (ICDEW), 2011 IEEE 27th International Conference on. IEEE, 2011: 212-217.

作者简介：

徐志浩（1991—），男，硕士研究生，主要研究方向为信息抽取。（通讯作者）

Email: 20134227020@stu.suda.edu.cn

惠浩添（1991—），男，硕士研究生，主要研究方向为信息抽取。

Email: 20134227019@stu.suda.edu.cn

钱龙华（1966—），男，副教授，硕士生导师，主要研究方向为自然语言处理。

Email: qianlonghua@suda.edu.cn

朱巧明（1963—），男，教授，博士生导师，主要研究方向为中文信息处理，Web 信息处理，物联网技术。

Email: qmzhu@suda.edu.cn