条件随机场与多层算法模型的实体自动识别

刘 殷[1],吕学强[1],刘 坤[2]
LIU Yin[1], LV Xueqiang[1], LIU Kun[2]

1.北京信息科技大学 网络文化与数字传播北京市重点实验室,北京 100101
2.北京拓尔思信息技术股份有限公司,北京 100101
1.Beijing Key Laboratory of Internet Culture and Digital Dissemination Research, Beijing Information Science and Technology University, Beijing 100101, China
2.Beijing TRS Information Technology Co., Ltd., Beijing 100101, China

LIU Yin, LV Xueqiang, LIU Kun. Automatic entity identification based on CRF and multilevel algorithm model. Computer Engineering and Applications, 2016, 52(11):141-147.

Abstract: Automatic entity identification technology is a powerful means to get information, and also is one of the key technologies in NLP field. Most of the current researches are named entity identification, and the researches are nearly mature, but the research of other kinds of entity like nominal and pronominal entity mentions is little. A method to identify the named and nominal entity mentions automatically is proposed. An approach for a new means, which using probability features inside the Chinese character, segmentation, and POS tagging information about it into CRF, then multilevel algorithm model to improve the results and recall which is not identified, is revealed to identify the entity in the corpus. Evaluated experiments on ACE standard corpus are proposed that the accuracy is 75.56%, and the recall is 72.52%. The results prove that the method is effective in entity identification problem.

Key words: entity identification; conditional random field; segmentation; multilevel algorithm model

摘 要: 实体自动识别技术是人们获取信息的有力手段,也是自然语言处理研究的关键技术之一。目前命名实体识别的研究较多,且已趋于成熟,而对汉语文本中的其他实体(名词性、代词性)研究较少。因此提出了一体化识别命名实体识别和名词性实体的方法,该方法将实体的汉字、分词、词性标注等信息引入条件随机场;再利用多层算法模型优化已经识别出的实体,以及召回未识别出的实体。在标准ACE语料库上进行实验,正确率达到75.56%,召回率达到72.52%。结果表明该方法对于实体识别问题是有效的。

关键词: 实体识别;条件随机场;分词;多层算法模型
文献标志码: A **中图分类号:** TP391.1 doi:10.3778/j.issn.1002-8331.1407-0325

1 引言

实体识别是人们获取信息的有力工具,是应对信息爆炸严重挑战的重要手段。根据Automatic Content Extraction(ACE)评测计划的定义,实体在文本中的引用(entity mention,也可称为指称项)可以有三种形式:命名性指称、名词性指称和代词性指称[1]。实体识别主要是指从文本中准确识别出命名性指称和名词性指称的文本片段。比如下面这个句子:

例:[沙马赫]继续担任[总人民委员会总秘书]。

其中,“沙马赫”是命名性指称,也就是传统意义上的命名实体,而“总人民委员会总秘书”是名词性指称。

目前的实体识别研究主要集中在命名实体识别,方法主要是基于规则的方法、统计的方法以及规则与统计相结合的方法。基于规则的方法通过分析命名性指称

基金项目: 国家自然科学基金(No.61271304);北京市教委科技发展计划重点项目暨北京市自然科学基金B类重点项目(No.KZ201311232037);北京市属高等学校创新团队建设与教师职业发展计划项目(No.IDHT20130519)。

作者简介: 刘殷(1988—),男,硕士,研究领域为中文与多媒体信息处理,E-mail:LiuXiaomi6Trunks@126.com;吕学强(1970—),男,博士,教授,研究领域为中文与多媒体信息处理;刘坤(1984—),男,高工,研究领域为多媒体信息处理。

收稿日期: 2014-07-23 **修回日期:** 2014-08-29 **文章编号:** 1002-8331(2016)11-0141-07
CNKI **网络优先出版:** 2015-03-13, http://www.cnki.net/kcms/detail/11.2127.TP.20150313.1552.030.html

的构词规则以及语义上下文规则,打分确定命名实体[2]。基于统计的方法主要有基于隐马尔可夫模型的方法[3]、基于支持向量机的方法[4]、基于最大熵模型的方法[5]、基于条件随机场模型的方法[6]、统计方法的混合模型[7]和其他频率统计方法[8]等。另外,也有很多研究者把多种统计方法结合起来[9],以及把统计与规则相结合[10]。但是,仅仅识别命名性指称对于一个高效鲁棒性的信息抽取系统是远远不够的,名词性指称的分布较命名性指称更加广泛,结构也无明显的规律性,识别难度比命名性指称大。单纯基于统计的方法对语料库的依赖比较大,如果选择人民日报或其他新闻报道作为训练语料,用在体裁多样、形式各异的真实文本中,识别效果则不是很理想。

本文的工作与自动内容抽取(ACE)评测中的实体提及检测(Entity Mention Detection,EMD)任务相类似,目的是识别指定类型的实体指称,包括命名性指称和名词性指称。本文把该任务分成两个子部分,首先利用条件随机场融合多种特征,进行命名性指称识别和名词性指称的粗略挖掘;其次提出一种规则与统计相结合的多层算法模型,对上一部分得到的名词性指称的粗略挖掘结果进行优化,以及对未识别出的两种实体类型进行召回。本文方法对真实文本中的实体识别,尤其在名词性指称方面通用性强,鲁棒性高,且在保证正确率的情况下达到较高的召回率。

2　实体的特点

本文研究的实体主要为命名性指称和名词性指称,数量众多,构成规律复杂。名词性指称即为带有指称关系的普通名词,构成规律极其复杂,在长度上分布参差不齐,且嵌套有短实体。根据语料统计,命名性指称长度通常在2~5个字符范围内,且极少有嵌套短实体的情况出现。而名词性指称的长度最短的是单字实体,长度最长能达到51个字符,且内部常嵌套若干短实体。实体的长度分布见图1。

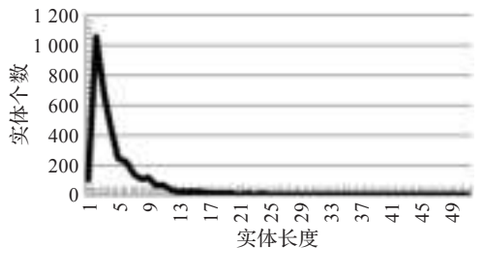

图1　实体长度分布图

图1中列出了语料中的长度分布,由于各实体长度的变化范围大,导致在识别过程中实体的边界很难确定。

传统的命名性指称在文本中的出现较有规律,在内部组成结构上也具有一定的固定性,而名词性指称的分布则比较随意,词语与词语的组合更多样,在句子中出现的位置不稳定。如表1是语料中实体的特点。

表1列出了语料中实体存在和出现的各种特点,其中"[]"标注为实体,"__"标注为关键字或词。也正是这些特点导致了实体,尤其是名词性指称的识别的效果不甚理想。

3　基于CRF的实体抽取方法

语言在文本中的出现呈现出一种序列性,故实体的识别问题就可以转化为序列标注问题。条件随机场是在给定标记的观察序列条件下,计算整个标记序列的联合概率分布来求得最佳序列的模型,因此本文选用条件随机场来进行语料的实体识别。

3.1　实体标注方法

本文采用四词位[11]标记,为语料中每一个字符(汉字、符号和外文字母)进行标注,设标注集为 Tag ,即 $Tag=\{B,M,E,S,O\}$,标注集元素及其含义见表2。

用以上方法对语料中每个字符进行标注,得到语料的实体标注序列 WS :

$$\begin{cases} WS=ws_1,ws_2,\cdots,ws_i,\cdots,ws_{size} \\ ws_i\in Tag \end{cases} \tag{1}$$

表1　实体的特点

实体特点		实例
命名性指称	人名	[宁赋魁]表示、[萨哈夫]在信中说
	地名	记者[巴黎]报道、来自[印尼]
	机构名	加入[国际足联]、属于[国际足联办公室]管理
名词性指称单独出现	单字	进攻以["王"]为核心
	多字	应当具有[企业法人]资格、申请[出入境中介机构]的
	外文	[FederalExpress]、[Zuliawati]
名词性指称连续出现	上下文紧密衔接	[新华社][北京]
	并列词衔接	[齐达内]和[他的球队]
名词性指称不平衡出现		办案质量和[群众]满意度
名词性指称后缀的名词性		[办公室]主管、[经济学家们]
名词性指称上下文缩略词的出现		[新美两国足协]、[新][美]结束第一轮比赛
名词性指称上下文的相似性		[各阶层人民]、[各阶层人士]
复杂结构名词性指称		[所在省(自治区、直辖市)中资机构有关部门]

表2　标注集元素及其含义

标注集元素		含义
有效标注集	B	实体首字
	M	实体中字(非尾字)
	E	实体尾字
	S	实体单字
噪音标注	O	非实体字

其中，*size* 为语料的字符的总数。经过标注后，实体识别的问题就转换成了对文本进行序列标注的问题。序列标注是给句子中的每个字解析成标注集中的某一个元素，标注集的元素是标注一个标识词边界的类别，在语言序列中显示的标注成为编码。此时，标注为S、BE、B(…M…)E形式的语言序列片段，即为待识别实体。例如图2所示的例句，经过实体标注后，将会得到每个字的码值。

原始语言序列：
特地从沪赶来的著名京剧表演艺术家尚长荣表演了一曲《西部情》。
实体标注编码：

特	地	从	沪	赶	来	的	著	名	京	剧	表	演	艺	
O	O	O	S	O	O	O	B	M	M	M	M	M	M	
术	家	尚	长	荣	表	演	了	一	曲	《	西	部	情	》	。
M	E	B	M	E	O	O	O	O	O	O	O	O	O	O	O

图2　实体标注编码示例

3.2　分词及其词性特征

中文文本的字颗粒粒度排列紧密，并呈序列状。而作为理解汉语文本的核心语法单位的“词”，在中文的语言环境下，“词”的长短不一，样式多变，而且汉语中没有显式的词边界。鉴于此，本文使用中科院的ICTCLAS系统[12]进行分词。

本文引入词边界特征。设分词边界标注集合为 BS，即 $BS=\{\text{ictb},\text{ictm},\text{icte},\text{icts}\}$，边界标注集合各元素及其含义见表3。

表3　边界标注集元素及其含义

边界标注集元素	含义
ictb	分词上边界
ictm	非分词尾字
icte	分词下边界
icts	分词单字

用表3的标注方法对分词的边界信息进行标注，得到语料的分词边界标注序列 BS：

$$\begin{cases} BS=bs_1,bs_2,\cdots,bs_i,\cdots,bs_{size} \\ bs_i\in BS \end{cases} \tag{2}$$

其中，*size* 为语料的字符的总数。

在使用ICTCLAS分词的过程中，选定北大二级词性标注，这样词性信息便可显式地标注在已分好的词后。当把词进行字颗粒处理后，为每个字附加词性信息。设词性集合为 Pos，则该集合的元素为北大二级标注集中的全部元素，得到语料的词性标注序列 PS：

$$\begin{cases} PS=ps_1,ps_2,\cdots,ps_i,\cdots,ps_{size} \\ ps_i\in Pro \end{cases} \tag{3}$$

其中，*size* 为语料的字符的总数。

经过以上标注处理，语料的特征便可得到，语料中的每个字符(包括汉字、英文字母和标点符号等)为 cs_i，语料特征集合为 TS，则语料特征集 TS 可以表示为下列四元组：

$$TS=(cs_i,bs_i,ps_i,ws_i),i\in[0,size-1] \tag{4}$$

其中，*size* 为语料的字符的总数。

3.3　特征模板的选取

根据训练语料、分词词性以及其边界信息的依存关系，设计合理的条件随机场特征模板。在实体识别中使用的模板文件，每个特征都设定了与其相关的模板表示。

本文所采用的基本特征是字形、分词边界和词性信息，以此所组成的特征模板记为原子模板，如表4所示。

表4　原子特征模板

模板编号	模板形式	模板含义
1	cs_i	当前字
2	cs_{i-1}	当前字左边第一个字
3	cs_{i-2}	当前字左边第二个字
4	cs_{i+1}	当前字右边第一个字
5	cs_{i+2}	当前字右边第二个字
6	bs_i	当前字的词边界信息
7	bs_{i-1}	当前字左边第一个字的词边界信息
8	bs_{i-2}	当前字左边第二个字的词边界信息
9	bs_{i+1}	当前字右边第一个字的词边界信息
10	bs_{i+2}	当前字右边第二个字的词边界信息
11	ps_i	当前字的词性信息
12	ps_{i-1}	当前字左边第一个字的词性信息
13	ps_{i-2}	当前字左边第二个字的词性信息
14	ps_{i+1}	当前字右边第一个字的词性信息
15	ps_{i+2}	当前字右边第二个字的词性信息

原子特征模板描述了当前字以及上下文中若干字的字形、分词边界信息或词性信息，所能表达的上下文信息有限。这种简单的字形、分词边界信息和词性信息并不能充分描述语言中的复杂现象，因此要利用更加适合语言内在规律的特征来描述模板。由于条件随机场是对数线性模型，因此本文将原子模板中的特征进行重新组合，构成复合、非线性的特征，如表5所示。

组合特征能够利用远距离的约束和丰富的上下文信息，表5描述了由两个和三个原子模板所构成的组合模板。通过引入组合模板特征，模型训练平台能够学习到更全面、更完整的信息，使得模型学习到的特征更完善。

4　融合多层算法模型的实体召回与优化

本文在条件随机场模型进行实体识别的基础上，提

表5 组合特征模板

模板编号	模板形式	模板含义
16	$cs_{i-1}+cs_i$	当前字和左边第一个字
17	cs_i+cs_{i+1}	当前字和右边第一个字
18	$cs_{i-2}+cs_{i-1}$	当前字左边第一个字和左边第二个字
19	$cs_{i+1}+cs_{i+2}$	当前字右边第一个字和右边第二个字
20	$bs_{i-1}+bs_i$	当前字的词边界信息和左边第一个字的词边界信息
21	bs_i+bs_{i+1}	当前字的词边界信息和右边第一个字的词边界信息
22	$bs_{i-2}+bs_{i-1}$	当前字左边第一个字的词边界信息和左边第二个字的词边界信息
23	$bs_{i+1}+bs_{i+2}$	当前字右边第一个字的词边界信息和右边第二个字的词边界信息
24	$ps_{i-1}+ps_i$	当前字的词性信息和左边第一个字的词性信息
25	ps_i+ps_{i+1}	当前字的词性信息和右边第一个字的词性信息
26	$ps_{i-2}+ps_{i-1}$	当前字左边第一个字的词性信息和左边第二个字的词性信息
27	$ps_{i+1}+ps_{i+2}$	当前字右边第一个字的词性信息和右边第二个字的词性信息
28	$cs_i+bs_i+ps_i$	当前字和当前字的词边界信息以及当前字的词性信息
29	$cs_{i-1}+bs_{i-1}+ps_{i-1}$	当前字左边第一个字和左边第一个字的词边界信息以及左边第一个字的词性信息
30	$cs_{i-2}+bs_{i-2}+ps_{i-2}$	当前字左边第二个字和左边第二个字的词边界信息以及左边第二个字的词性信息
31	$cs_{i+1}+bs_{i+1}+ps_{i+1}$	当前字右边第一个字和右边第一个字的词边界信息以及右边第一个字的词性信息
32	$cs_{i+2}+bs_{i+2}+ps_{i+2}$	当前字右边第二个字和右边第二个字的词边界信息以及右边第二个字的词性信息

出了三种改进策略来提高条件随机场模型的泛化性能，即利用多层算法模型。第一层是通过统计方法，计算加权标注指数的算法模型；第二层是利用介宾结构的地名缩略语的算法模型；第三层是基于名词性指称的名词性的特点，使用这项规则来进行优化的算法模型。

4.1 利用加权标注指数的统计算法模型

训练语料不可能囊括所有的语言现象和表达特点，所以条件随机场模型有数据稀疏问题的困扰。因此，在条件随机场训练的过程中，并不能学习到所有实体情况，这将导致实体的识别结果不完整。本文引入加权标注指数，重点补充识别未被召回的以单字形式出现的名词性指称。

加权标注指数是描述一个字符被标为某种实体的数值表现形式。本文将加权标注指数定义为：

$$WeightDeviation_i=\sqrt{\frac{(Index_i-\lambda)^2}{n}} \tag{5}$$

其中，$WeightDeviation_i$ 表示第 i 个字符的加权标注指数，$Index_i$ 表示第 i 个字符的词频指数，λ 为调节系数，n 表示被标为某个标注的字符样本容量。

词频指数定义如下：

$$Index_i=\frac{N_{ws_i}(cs_i)}{N(cs_i)}\times \mathrm{lb}\frac{\overline{N}}{\overline{N}_{cs_i}} \tag{6}$$

其中，N_{ws_i} 表示当前字符被标注为 w_i 的次数，$N(cs_i)$ 表示当前字符出现的总次数，$\overline{N}$ 表示识别出的标注个数，也就是当前标注集的个数，$\overline{N}_{cs_i}$ 表示当前字符被识别出的标注个数。

语料中经常有两个单字形式的名词性指称连续出现的情况，但是这种形式经常被条件随机场识别成一个双字词。根据汉语中常见的一类语法现象：一个双字词是由具有并列关系的两个单字词组合而成，如：“干群”、“党群”。本文通过计算每个单字词的加权标注指数，动态生成词表 *WeightTable*，抽取词表中的优良结果对名词性指称进行召回。算法模型如算法1和算法2所示。

算法1 动态构建权重词表算法

StatisticsVocabulary(*TS*)

输入 语料特征集合 (*TS*)

输出 权重词表 (*WeightTable*)

```
1  WeightTable ← ∅
2  TS = {(cs_i, bs_i, ps_i, ws_i)|i ∈[0, size − 1]}
3  for i ← 0 to size − 1
4  do if ws_i = S
5        then Calculate(WeightDeviation_i)
6             WeightTable ← (cs_i, WeightDeviation_i)
7  return WeightTable
```

算法2 加权标注指数召回算法

StatisticsRecall(*TS*, *WeightTable*)

输入 语料特征集合 (*TS*)、权重词表 (*WeightTable*)

输出 召回后语料 ($\overline{TS}$)

```
1  TS ← ∅
2  TS = {(cs_i, bs_i, ps_i, ws_i)|i ∈[0, size − 1]}
3  for i ← 0 to size − 1
4  do if bs_i = icts
   and bs_{i+1} = icts
   and (cs_i, WeightDeviation_i) ∈ WeightTable
   and WeightDeviation_i < θ
5        then ws_i ← S
6             TS ← (cs_i, bs_i, ps_i, ws_i)
7        else TS ← (cs_i, bs_i, ps_i, ws_i)
8  return TS
```

其中，StatisticsVocabulary(*TS*) 为动态构建权重词表 *WeightTable* 的算法流程，*TS* 为语料特征集合，即公式(4)所示的四元组，*size* 为语料的字符总数。StatisticsRecall(*TS*, *WeightTable*) 为利用动态生成的 *WeightTable* 词表，在 *TS* 中进行名词性指称的召回，生成经过召回后的语料 $\overline{TS}$ 的算法。

4.2　利用地名缩略语的算法模型

介宾结构在汉语中是很常见的语法现象，介宾短语是由介词加上后面的名词、代词或名词性短语组成的[13]。根据语料特点，介宾结构中出现单字的地名可被识别出来，比如：在[沪]宣布、在[京]召开。介宾结构的实体可以通过构建文法算法召回。

设地名缩略语词表为 *PlaceTable*，*PlaceTable* 由两部分组成，其一为当前字是省份简称但在其全称中不出现的字，如上海简称沪，河北简称冀等，设为 PT_{abbr}；其二是从语料中动态挖掘，设为 PT_{auto}，算法基本思想是：如果当前词为中国地名，则取其每一个字收录入地名缩略词表；若当前词为外国地名的中文译称，则取其第一个字收录入地名缩略词表。当遇到介宾结构且宾语部分的字符与地名缩略表中的字符进行匹配，匹配成功则认为当前字符为实体进行召回。算法描述如算法3和算法4所示，其中，ns 表示中国地名，nsf 表示外国地名，p 为介词，b 为缩略词。

算法 3　动态挖掘地名缩略语算法
LocatonalVocabulary(*TS*)
输入　语料特征集合 (*TS*)
输出　地名缩略词表 (*PlaceTable*)

```
1  PT_auto ← ∅
2  TS = {(cs_i, bs_i, ps_i, ws_i)|i ∈[0, size − 1]}
3  for i ← 0 to size − 1
4  do if ps_i = ns
5        then PT_auto ← cs_i
6        else if ps_i = nsf
7              foreignstring ← {}
8              foreignstring ← {cs_i, cs_{i+1}, ⋯, cs_{i+n−1}}
9              PT_auto ← cs_i
10             i ← i + n
11 return PlaceTable ← PT_abbr + PT_auto
```

算法 4　地名缩略语召回算法
LocationalRecall(*TS*, *PlaceTable*)
输入　语料特征集合 (*TS*)
输出　召回后语料 ($\overline{TS}$)

```
1  TS ← ∅
2  TS = {(cs_i, bs_i, ps_i, ws_i)|i ∈[0, size − 1]}
3  for i ← 0 to size − 1
4  do if ps_i = b and cs_i ∈ PlaceTable and ps_{i−1} = p
5        then ws_i ← S
6             TS ← (cs_i, bs_i, ps_i, ws_i)
7             else TS ← (cs_i, bs_i, ps_i, ws_i)
8  return TS
```

其中，LocatonalVocabulary(*TS*) 为动态挖掘地名缩略语算法，最后的地名缩略词表 *PlaceTable* 由 PT_{abbr} 和 PT_{auto} 两部分构成。LocationalRecall(*TS*, *PlaceTable*) 为利用地名缩略词表，在 *TS* 中进行地名实体召回，生成经过召回后的语料 $\overline{TS}$ 的算法流程。

4.3　名词性指称的名词性规则算法模型

由于名词性指称的最后一个词，一定是名词性或“们”等代表实体复数的词，所以根据分词结果，如果识别出的实体的最后一个词不是名词和“们”等代表实体复数的词，则将其修改。如：国务院主管…，由于“主管”被标注为动词，所以最后仅保留“国务院”。算法如算法5和算法6所示，其中 noun 为名词性或“们”等代表实体复数的词。

算法 5　实体挖掘算法
EntitySet(*TS*)
输入　语料特征集合 (*TS*)
输出　实体集合 (*ES*)

```
1  count ← 0
2  ES ← ∅
3  TS = {(cs_i, bs_i, ps_i, ws_i)|i ∈[0, size − 1]}
4  for i ← 0 to size − 1
5  do if ws_i ≠ O
6        then ES ← ((cs_i, bs_i, ps_i, ws_i))
7        if ws_i = B or ws_i = S
8              do count ← count + 1
9  return ES
```

算法 6　规则优化算法
RuleOptimization(*ES*)
输入　实体集合 (*ES*)
输出　优化后实体集合 (*ES_final*)

```
1  ES_final ← ∅
2  for count ← 0 to size of the entity
3  do length ← 0
4        length ← Length(ES_count)
5        while bs_i ≠ ictb and ps_i ≠ noun
6        do ws_i ← O
7              length ← length − 1
8        ws_i ← O
9  ES_final ← ES_count
10 return ES_final
```

其中，EntitySet(*TS*) 为实体挖掘算法，把识别出来的实体放入集合 *ES* 中。再利用 RuleOptimization(*ES*) 算法，根据上文所述的实体名词性规则进行过滤，最后得到实

体集合 *ES_final* ,该集合为优化后的结果。

5　实验结果与分析

本文使用的语料是标准ACE2005 broadcast news数据集,通过去除代词性指称的预处理后,得到的语料信息如表6所示。

表6　ACE2005 broadcast news语料统计表

名称	新闻广播语料	训练语料	测试语料
文档数量	633	499	134
句子数量	5 212	4 106	1 106
实体数量	16 409	12 926	3 483

算法实现使用的工具包为条件随机场[14]工具包CRF++0.53版本;其中,$\theta=0.000\ 58$,调节系数$\lambda=-0.007$,θ,λ两个参数的设置是通过反复实验得到的权重参数。

5.1　评价方法

对于本文实体识别任务,本文采用正确率 P 、召回率 R 和 F 测度值 F 进行评价,具体计算公式如下。

$$P=\frac{\text{系统识别正确的实体个数}}{\text{系统识别出的实体个数}}\times 100\% \tag{7}$$

$$R=\frac{\text{系统识别正确的实体个数}}{\text{语料中实体总数}}\times 100\% \tag{8}$$

$$F=\frac{2\times P\times R}{P+R} \tag{9}$$

5.2　结果分析

由于文献[15]中所识别的命名实体,在长度上接近于语料中的平均长度,且待识别实体的内部组成结构与语料中的实体相似,故本文选用文献[15]的方法为基础实验,命名为BM方法(Baseline Method)。表7是BM方法与本文中用条件随机场的识别方法(CRF Method,CM方法)的结果对比。

表7　BM方法与CM方法识别结果对比

选用方法	评价方法	评价结果/%
BM方法识别结果	正确率	76.790 4
	召回率	68.238 4
	F 测度值	72.262 2
CM方法识别结果	正确率	75.300 8
	召回率	71.863 3
	F 测度值	73.541 9

如表7所示,利用CM方法识别后的结果,虽然正确率降低了1.489 6%,但是召回率提高了3.624 9%,最后 F 值也相应提高了1.279 7%。可以看出,本文采用的CM方法不仅在 F 测度值上有提升,且在召回率上有较大的改善。

BM方法以词为单位进行训练学习,保留了部分词的分词边界信息,粒度较粗;CM方法以字为单位进行训练学习,粒度较细,有效地解决了分词边界与短实体的信息不对称问题,以及训练学习基本单位粒度较粗的问题,如“党性”,待识别实体是“党”,短实体“党”存在于实体“党性”分词边界的内部;又如“干群”,待识别实体“干”、“群”存在于“干群”分词边界的内部。因为训练粒度较细,用BM方法中能召回的一个实体中,在CM方法中可能召回多个,故召回率有很大提升。然而CM算法粒度较细,丢失了部分分词边界信息,待识别实体为“大连实德队”,但识别出的实体是“大连实德”,造成了部分实体的边界识别错误,导致了正确率的下降。

表8是在利用CM方法的识别结果上加入多层算法模型方法(Multilevel Algorithm Method,命名为MAM)后的结果对照。

表8　进行后处理的结果对比

选用方法	评价方法	评价结果/%
CM方法	正确率	75.300 8
	召回率	71.863 3
	F 测度值	73.541 9
MAM方法	正确率	75.560 9
	召回率	72.523 7
	F 测度值	74.011 1

如表8所示,在经过MAM方法之后,正确率提高了0.260 1%,召回率提高了0.660 4%,F 值提高了0.469 2%。可以看出,加入多层算法模型MAM方法后,各项效果均好于未进行后处理的CM方法的结果。

表9是单独使用各层算法模型,每一层所对应的识别率之间的结果与CM方法结果进行对照。如表9所示,从三个指标来看,单独使用各层的识别率较CM的识别率都有所提升。

表9　单独使用各层算法模型的结果对比

算法层次	正确率/%	召回率/%	F 值/%
CM	75.300 8	71.863 3	73.541 9
MAM-L1	75.322 3	72.121 7	73.687 3
MAM-L2	75.496 1	72.093 0	73.755 3
MAM-L3	75.322 7	72.035 6	73.642 5

表9中,MAM-L1算法得到的召回率最高,说明本文提出的加权标注指数算法模型可以保证在优化识别实体正确性的同时,还能较大幅度地召回机器学习模型没有识别出来的实体,从数据上看,MAM-L1算法召回实体的特点多为原先分词边界内部的多个短实体,表明本文算法在弥补分词边界造成的识别差异层面有比较明显的效果;MAM-L2算法得到的正确率最高,说明采用地名缩略语算法能够保证大量新召回实体的精确度,另外利用本文算法能得到最高的 F 值,从一个侧面反映出MAM-L2算法具有鲁棒性,比较稳定。

图3为在实体识别各个阶段的识别效果。从图中可以看出,融合三种算法模型得到的识别效果优于每一层算法模型的识别效果,说明融合三种算法模型后,达到了融合算法的效果最大化,且在融合的过程中,不受

算法模型融合顺序的限制，这在一定程度上，保证了算法模型的独立性和不相关性，使得模型保证了在不干扰已召回和已优化实体的基础上，便于后续新算法的补充，这说明了本文的融合算法在实体的召回和优化过程中体现了稳定性和鲁棒性。

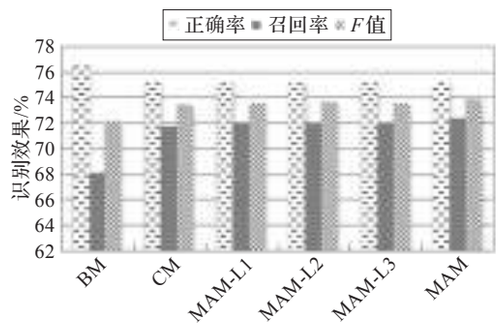

图3　各阶段的识别效果图

另外，从实体识别任务整体来看，相对于正确率，召回率的提高更有意义。如图3所示，正确率在CM方法处虽有略微下降，但是在后续的融合算法中却稳步提升，召回率也有很大的改进，使得 F 值也有相应的提升。说明从实体识别的任务整体上看，本文提出的实体抽取方法和融合算法模型都具有积极有效的作用。

6　结论与展望

本文系统地阐述了实体识别的意义及其特点，提出一种融合条件随机场简单特征以及利用多层算法模型的方法。该方法基于统计学习框架，利用条件随机场模型融合丰富的上下文语言学特征进行识别。在标准ACE语料库上的实验表明，本文提出的基于分词信息的标注方法，在没有得到深层次的句法信息情况下可以很好地解决实体识别的问题。从实现角度出发，统计机器学习方法与本文提出的多层算法模型相结合，在实体识别方面取得了比较满意的效果。在后续的指代消解任务中，代词可以指向句子中的任何实体，因此在保证较高的正确率的情况下，把语料中的实体尽可能多地找出，才能保证指代消解顺利进行，这对于学术研究或工程应用有一定的价值和意义。

下一步研究工作主要集中在以下几个方面：将现有的中文分词系统进行改进，使之尽量与ACE的标准接轨；研究引入语法信息，以识别出更多的带有汉语完整语法成分的或含有感情色彩的词汇的长实体；基于本文的结果，深入研究指代消解。

参考文献：

[1] Doddington G R，Mitchell A，Przybocki M A，et al.The Automatic Content Extraction(ACE) program-tasks，data，and evaluation[C]//Proceedings of the 4th International Conference on Language Resources and Evaluation，Lisbon，Portugal，2004.

[2] 张小衡，王玲玲.中文机构名称的识别与分析[J].中文信息学报，1997，11(4)：21-32.

[3] Chopra D，Morwal S.Named entity recognition in English using hidden Markov model[J].International Journal，2013.

[4] Ekbal A，Bandyopadhyay S.Named entity recognition using support vector machine：a language independent approach[J].International Journal of Electrical，Computer，and Systems Engineering，2010，4(2)：155-170.

[5] Attardi G，Baronti L，Dei Rossi S，et al.SuperSense tagging with a maximum entropy Markov model[M]//Evaluation of natural language and speech tools for Italian.Berlin/Heidelberg：Springer，2013：186-194.

[6] 邱泉清，苗夺谦，张志飞.中文微博命名实体识别[J].计算机科学，2013，40(6).

[7] 黄德根，李泽中，万如.基于SVM和CRF的双层模型中文机构名识别[J].大连理工大学学报，2010，50(5)：782-786.

[8] Yao X.A method of Chinese organization named entities recognition based on statistical word frequency，part of speech and length[C]//2011 4th IEEE International Conference on Broadband Network and Multimedia Technology(IC-BNMT)，Shenzhen，China，2011：637-641.

[9] Che W，Wang M，Manning C D，et al.Named entity recognition with bilingual constraints[C]//Proceedings of NAACL-HLT，Atlanta，USA，2013：52-62.

[10] Ling Y，Yang J，He L.Chinese organization name recognition based on multiple features[M]//Intelligence and security informatics.Berlin/Heidelberg：Springer，2012：136-144.

[11] 黄昌宁，赵海.由字构词——中文分词新方法[C]//中国中文信息学会第六次全国会员代表大会暨成立二十五周年学术会议，2006.

[12] Zhang H P，Yu H K，Xiong D Y，et al.HHMM-based Chinese lexical analyzer ICTCLAS[C]//Proceedings of the 2nd SIGHAN Workshop on Chinese Language Processing-Volume 17，Sapporo，Japan，2003：184-187.

[13] 谢靖，苏新宁，沈思.CSSCI语料中短语结构标注与自动识别[J].现代图书情报技术，2012(12).

[14] Lafferty J，McCallum A，Pereira F C N.Conditional random fields：probabilistic models for segmenting and labeling sequence data[C]//Proc of 18th ICML.San Francisco，USA：AAAI Press，2001：282-289.

[15] 胡文博，都云程，吕学强，等.基于多层条件随机场的中文命名实体识别[J].计算机工程与应用，2009，45(1)：163-165.