基于联合模型的中文嵌套命名实体识别

尹　迪,周俊生,曲维光

(南京师范大学计算机科学与技术学院,江苏 南京 210023)

[**摘要**]　中文嵌套命名实体识别是自然语言处理中一个比较困难的问题. 针对传统的序列化标注方法的不足,本文提出了一种新的基于联合模型的中文嵌套命名实体识别方法,该方法将嵌套命名实体识别看作是一种联合切分和标注任务. 联合模型用一种改进的 beam search 算法作为系统的解码算法,并采用一种在线学习算法平均感知器算法作为训练算法,获得了较快的收敛速度和较好的识别效果. 实验结果表明基于联合模型的方法对嵌套命名实体识别取得了更好的效果.

[**关键词**]　嵌套命名实体识别,序列化标注模型,联合模型,感知器算法

[**中图分类号**]TP391　[**文献标志码**]A　[**文章编号**]1001-4616(2014)03-0029-07

Chinese Nested Named Entity Recognition Using a Joint Model

Yin Di, Zhou Junsheng, Qu Weiguang

(School of Computer Science and Technology, Nanjing Normal University, Nanjing 210023, China)

Abstract: Chinese nested named entity recognition is a very difficult problem in natural language processing. This paper presents a novel method based on a joint model, which treats the recognition of Chinese nested named entity as a task of joint word segmentation and labeling. The proposed method exploits an improved beam search algorithm as decoding algorithm, and uses the averaged perceptron algorithm as training algorithm, attaining fast convergence during training. The experimental results show that the joint model achieves better performance than two baseline systems using the traditional sequence labeling models.

Key words: nested named entity recognition, sequence labeling models, joint models, perceptron algorithm

命名实体识别是许多自然语言处理任务的基本要求,如文本摘要、信息抽取、机器翻译等. 近年来,基本命名实体识别研究已经取得了较大的进展,但是对嵌套命名实体的识别还没有取得较好的效果. 相比于基本命名实体,嵌套命名实体往往结构复杂,嵌套结构里面包含着两层甚至多层命名实体,同时识别出嵌套实体及其包含的基本命名实体难度很大. 例如“[[中华人民共和国]驻[南非共和国]大使馆]”和“[[中华人民共和国][国务院]侨务办公室]”是两层嵌套的组织机构名,“[[黔南]布依族苗族自治州]”是一个两层嵌套的地名. 这 3 个嵌套实体分别包含“中华人民共和国”、“南非共和国”以及“黔南”等基本命名实体,现有的方法可以较好地识别嵌套结构中包含的基本命名实体,但是很难同时识别整个嵌套命名实体.

现有的解决嵌套命名实体识别的常用方法是基于序列化标注的方法,但是基于序列化标注的方法解决嵌套命名实体识别存在两个缺点:第一,用基于词的序列化标注方法解决嵌套命名实体识别,需要先对语料分词,这会产生错误传播的问题,即分词错误无法在基本命名实体识别的过程中得到修正,影响识别效果;第二,用基于字符的序列化标注方法对嵌套命名实体进行识别虽不需要对测试语料进行分词,但是该方法只能用到字符层的信息,无法用到词汇层的信息. 例如“位于”这个词一般存在于地名的前面,对地名具有指示作用,但基于字符的序列化标注方法就无法有效利用这样的上下文信息.

针对序列化标注方法的不足,本文提出了一种基于联合模型的方法解决中文嵌套命名实体识别问题,

收稿日期:2014-02-18.

基金项目:国家自然科学基金(61272221,61472191)、江苏省社科基金(12YYA002).

通讯联系人:周俊生,博士,副教授,硕士生导师,研究方向:自然语言处理. E-mail:zhoujs@ njnu. edu. cn

联合模型能同时处理分词和序列化标注,解决了序列化标注方法中语料需要分词的缺点.并且,由于联合模型能够同时处理分词,所以它不仅可以用到字符层的信息也可以用到词汇层的信息.嵌套命名实体作为一个整体,其包含的基本命名实体的预识别信息能对嵌套命名实体的识别提供决策支持.

本文首先介绍了嵌套命名实体识别的相关研究,并具体描述了基于序列化标注的两种典型方法.然后,针对序列化标注方法的缺点,我们提出了一种基于联合模型的方法,并对该方法进行了详细描述.最后,本文分别对基于序列化标注的方法和联合模型的方法进行实验对比,并对实验结果进行了分析.

1 相关工作

目前,有关嵌套命名实体识别的研究相对较少,主要使用基于统计的方法(如条件随机场、最大熵等)来识别嵌套命名实体.

文献[1]将嵌套实体识别分成两个子任务:嵌套实体边界检测以及实体多层信息标注.首先,他们提出了一种层次结构信息编码方法,将多层嵌套边界检测问题转化为传统的序列标注问题,利用条件随机场模型融合多种特征进行统计决策.其次,将多层信息标注问题看作分类问题,从实现的角度设计了含有两个分类引擎的并行SVM分类器,避免了对每层信息标注都设计一个分类器.文献[2]用基于条件随机场的双层模型进行嵌套命名实体识别,第一层模型识别基本命名实体,第二层模型根据已经识别的基本命名实体,识别出嵌套的命名实体.该方法引入了实体语素特征,提高了识别效果.文献[3]利用层叠条件随机场模型对中文组织结构名进行识别.该方法对粗分的词串进行人名和地名的识别,将识别结果传递到高层模型,为高层模型识别组织机构名提供决策支持.最后,采用约束的前向后向算法对识别的结果进行可信度计算.文献[4]用最大熵马尔科夫算法在两种生物医学语料上做了嵌套命名实体识别实验.文中分别用层次(layering)、层叠(cascading)以及联合标签(joined label)来标注训练语料,用序列化标注的方法识别嵌套实体.文献[1-4]均采用了序列化标注的方法解决了嵌套命名实体识别问题,但是它们都存在序列化标注方法固有的缺点.

文献[5]采用判别式成分句法分析器来训练嵌套命名实体识别模型.该方法把每个句子转换成一棵句法分析树,并把每一个词作为该树的叶子节点,每个实体作为该树的子树.树的表示方法可以清晰的表示任意层数的嵌套实体.该方法对英文的多层嵌套命名实体识别效果较好,但是该方法很难用到中文嵌套命名实体识别中,这是因为该方法的语料是已经知道了结构的语料,并且中文嵌套命名实体识别和英文嵌套命名实体识别存在着较大的区别.首先,中文句子是没有切分的连续字符序列,对包含嵌套命名实体的句子进行切分会对后续的识别产生不利的影响.其次,中文嵌套命名实体不具备明显的边界标记符号.

本文借鉴以上的研究成果,设计了两种基于序列化标注的嵌套命名实体识别方法,分别是基于层次标号标注的序列化标注方法和基于双层模型的序列化标注方法.这两种方法的思路如下所述:

(1)基于层次标号的方法.本方法用层次标签对嵌套命名实体进行标注,下划线之前的部分表示嵌套实体包含的基本命名实体或者词的标签,下划线之后的部分表示基本命名实体或者词在嵌套命名实体中的位置.下划线前后的标签都采用了"BIO"标注方法,标注中的"B"表示实体的开头字符,"I"表示实体的中间或结尾的字符,"O"表示非实体字符."ns"表示地名,"nr"表示人名,"nt"表示组织机构名.例如句子"[[北京][石景山]发电厂]"是一个嵌套的组织机构名,它的标注如下:

- 北/B-ns_B-nt 京/I-ns_B-nt 石/B-ns_I-nt 景/I-ns_I-nt 山/I-ns_I-nt 发/o_I-nt 电/o_I-nt 厂/o_I-nt

对于嵌套命名实体"北京石景山发电厂"中的地名"北京",我们用"北/B-ns_B-nt 京/I-ns_B-nt"标注,"B-ns_B-nt"中的"B-ns"表示"北"是"北京"这个地名的第一个字,"B-nt"表示"北京"是"北京石景山发电厂"这个嵌套组织机构名的开始部分.其他的实体类型标注以此类推.用上面的标注方法标注语料之后,嵌套命名实体识别问题就转化为了基于字符的序列化标注问题.

(2)基于层叠模型的方法.该方法将嵌套命名实体识别任务分为两步:第一步,用基于字符的序列化标注模型标注文本中的字符(如"我 B-o 们 I-o"),并根据标注结果将其转化为词-标签对(如"我们 o")和命名实体-标签对(如"北京 ns")的形式;第二步,把识别出的词、基本命名实体以及它们的类型作为特征,用基于词的序列化标注模型识别出嵌套的命名实体.在双层模型中,第一层模型用"BI"标注方法标注

语料中的每个词,如“北京”标注为“北/B-ns 京/I-ns”;第二层模型用“BIO”标注方法标注语料,对于嵌套命名实体,如“北京石景山发电总厂”,我们将其标注成“北京/B-nt 石景山/I-nt 发电/I-nt 总厂/I-nt”的形式,对于非嵌套命名实体,我们将其标注为“词-标签 O”的形式,如“我们 O”.

基于以上两种中文嵌套命名实体识别的序列化标注方法,我们通过采用经典的条件随机场模型[6],分别实现了两个中文嵌套命名实体识别的 baseline 系统.

2　基于联合模型的嵌套命名实体识别方法

2.1　方法概述

基本命名实体识别一般分为两步:第一步,对测试句子进行分词;第二步,对分好的词进行序列化标注,确定实体类型. 这种方法有一个明显的缺点:错误传播,即分词错误无法在序列化标注过程中得到修正,会影响序列化标注的标注效果. 所以,一种更好的解决办法是同时处理这两个问题,利用标注信息为分词提供决策,利用分词信息为标注提供决策,二者相互促进. 为了避免错误传播并且使分词和序列化标注问题能够相互促进,我们把分词和序列化标注这两个问题看作一个整体来处理:给定一个未分词以及未标注的句子,联合模型考虑句子所有的切分和标注的情况,选择全局最优的切分和标注作为输入句子的输出.

联合模型是一个全局线性模型,对于给定的句子 x,候选的输出结果 y 是逐步形成的. 程序每一步处理一个新的字符,这个字符要么和上一个部分的词结合形成新的词,要么作为下一个新的词的开始字符. 当前字符作为新词的开始字符的情况下,上一个词即为一个分好的完整词. 对于每一个已经分好的完整的词,系统给它分配一个任意的标签. 处理完句子 x 中的所有字符之后,模型就生成了候选的结果集 Y,系统选择结果集中模型得分最高的结果 y 作为输出结果. 输出结果 y 的模型得分计算公式为:

$$Score(y)=\boldsymbol{\Phi}(y)\cdot\boldsymbol{w} \tag{1}$$

其中 $\boldsymbol{\Phi}(y)$ 为从 y 中提取的全局特征向量,$\boldsymbol{w}$ 表示模型的参数向量.

在基本命名实体识别任务中,联合模型把基本命名实体当作一个整体来进行切分和标注,即对句子中的基本命名实体进行切分的同时,为这些基本命名实体标注一个实体类型. 例如下面一个包含基本命名实体的句子,我们将其标注为如下形式:

- 主席\o 在\o 人民大会堂\ns 接待\o 了\o 美利坚合众国\ns 总统\o 克林顿\nr

在这个例子中,“美利坚合众国”、“人民大会堂”等这些命名实体被当作一个整体进行切分并同时进行标注.

嵌套命名实体是多个词和基本命名实体的集合. 利用以上介绍的联合模型,我们无法在确定嵌套命名实体包含的基本命名实体的边界和类型的同时确定嵌套命名实体的类型与边界. 如下面的例子所示:

- [[美]\ns[中]\ns 贸易全国委员会]\nt 和\o[[美国]\ns[中国]\ns 商会]\nt　在\o　纽约\ns 举行\o　宴会\o

嵌套命名实体“美国中国商会”以及“美中贸易全国委员会”中的基本地名“美国”、“中国”、“美”以及“中”能够作为一个整体被联合模型识别,但是我们无法在识别这些基本命名实体的同时识别这两个嵌套命名实体. 因此我们需要对该联合模型中的标注符号体系进行设计,以适合嵌套命名实体的识别.

为了能够确定嵌套命名实体的边界和类型,并同时识别出其内部的基本命名实体以及其他普通构成词,我们的方法是将上述的联合模型与传统的“BIE”标注方法进行有机组合. 具体的,我们仍将嵌套命名实体的识别看成是一个联合的切分与标注过程,但针对嵌套命名实体识别的需求,我们将用于基本命名实体的 nr、ns、nt、o 等标注符号与序列化标注方法中广泛采用的用于表示位置的“BIE”符号集进行了组合,构造了一个能有效进行嵌套命名实体识别的层次标注符号体系.

我们用双层标签标注嵌套命名实体的各部分,第一层标签和第二层标签之间用下划线分开. 在双层标签中,下划线之前的部分表示嵌套命名实体包含的基本命名实体以及非命名实体词的标签,下划线之后的部分表示基本命名实体以及非命名实体词在嵌套命名实体中的位置. 下划线之前的标签用 nr,ns,nt 以及 o 四个标签表示,下划线之后的标签利用 ns 和 nt 两个标签结合“BIE”标注方法来表示. 在“BIE”标注方法中,“B”表示嵌套命名实体的开始部分,“I”表示嵌套命名实体的中间部分,“E”表示嵌套命名实体的结尾

部分. 我们通过下划线之前的标签确定嵌套命名实体中包含的基本命名实体的类型,通过下划线之后的标签来确定嵌套命名实体的类型和边界. 下面的句子是一个包含嵌套命名实体的句子,我们利用上面的标注方法将其标注为如下形式:

- 重庆市/ns_B-nt 社会/o_I-nt 福利/o_I-nt 募捐/o_I-nt　委员会/o_E-nt　向/o　何雪梅/nr　捐赠/o 必需品/o

观察句子的标签,我们就可以知道,"重庆市社会福利募捐委员会"是一个嵌套的组织机构名,"重庆市"是其包含的基本命名实体,并且"重庆市"位于嵌套命名实体的开头部分.

相对于传统的序列化标注方法,采用我们的联合模型实现中文嵌套命名实体识别具有以下的几点优势:

(1)将中文句子的分词、基本命名实体的识别与嵌套命名实体的识别任务集成在同一个分析和处理过程中,这样可以有效地避免传统管道(pipeline)模型中前一个处理步骤的错误会传播到下一个步骤中,造成错误传播的问题;

(2)在基于联合模型的嵌套命名实体识别过程中,可以综合利用字符层的信息与词汇层的信息,帮助完成句子中的基本命名实体与嵌套命名实体的边界与类型的决策判断;

(3)在基于联合模型的嵌套命名实体识别过程中,可以通过利用其中的基本命名实体预识别信息,为整个嵌套命名实体的识别提供决策支持.

联合模型包括解码和训练两部分,下面将详细介绍联合模型的解码算法和参数学习算法.

2.2　解码算法

对于一个给定的句子,解码算法可以搜索出分词和嵌套命名实体识别得分最高的预测作为输出结果. 但是,解码算法面对的搜索空间非常巨大. 在给定模型和特征模板的情况下,进行精确地搜索非常困难,即使用动态规划算法速度也非常慢. 所以,本文采用了近似搜索算法 beam-search 作为训练和测试的解码算法.

对于给定的句子 x,候选的输出结果 y 是逐步形成的. 程序每一步处理一个新的字符,这个字符要么和上一个部分词结合形成新的词,要么作为下一个新的词的开始字符. 在联合模型中,候选结果的切分和标注是到当前字符为止的,对于最后一个词是不是个完整的词,模型无法确定. 因此存在这样一个问题:对于一个不完整的词模型是否应该标记一个命名实体标签. 为了解决这个问题,文献[7]提出用 multiple-beam search 算法来处理分词和词性标注一体化问题,该算法只对已经标注的完整词进行处理.

为了加快 multiple-beam search 的搜索速度,文献[8]提出了 single-beam search 算法. 该算法在 multiple-beam search 算法的基础上,添加了词长限制并且利用了标签词典.

解码算法可以在执行搜索的过程中对输出结果进行检测,判断结果是否正确. 为了提高搜索和识别的性能,文献[7,8]提出了"early update"优化方法. 该方法能够有效减少噪音数据,不让错误延续到句子处理的最后一个字符. 但是,"early update"方法在预测的切分和标注与正确的切分和标注不相符时,就会停止搜索,这样的处理导致了训练算法每次都只能用到已经处理的那一部分句子信息,而无法用到更多的句子信息,从而降低了收敛速度,增加了迭代次数和训练时间. 针对"early update"方法的缺点,文献[9]提出了一种改进的优化方法:"max-violation update",它返回最坏情况下的切分和标注用以更新参数. 该方法能够获得比"early update"方法相当甚至更好的识别效果,并且收敛速度更快. "max-violation update"的更新规则如式(2)所示:

$$(x,y^*,z^*)=\underset{(x,y,z)\in C,z\in\cup|B_i[0]\}}{\operatorname{argmin}} \boldsymbol{w}\cdot\Delta\boldsymbol{\Phi}(x,y,z) \tag{2}$$

式(2)中,x 表示输入的句子,y 表示输入句子正确的切分和标注,z 表示输入句子预测的切分和标注. $\boldsymbol{w}\cdot\Delta\boldsymbol{\Phi}(x,y,z)$ 表示正确的切分和标注与预测的切分和标注的得分差值,这个差值是一个负数,文献[9]对此给出了详细的证明.

本文的解码算法就采用了"max-violation update"优化后的 Single-beam 算法. 通过"max-violation update"优化后的 Single-beam 算法的伪代码如图1所示. 在图1的伪代码中,agenda 表示一个集合,它用来存放字符增加过程中 N 个最好的候选结果;Clear 表示清空 agenda 中的数据. AddItem 表示添加一个新的项到 agenda 中;N-Best 表示返回 N 个模型得分最高的项;Append 表示附加一个字符到当前候选项的最后

一个词上;Separate 表示再添加一个新词. 在“max-violation update”方法中,满足 violation 的条件是句子预测的切分和标注的模型得分大于正确的切分和标注的模型得分. agenda1 存储所有满足 violation 的预测,MaxViolation 表示返回 agenda1 中 violation 值最大的预测.

2.3　参数学习算法

通过定义特征之后,一个候选的输出 y 会被映射成一个全局的特征向量. 训练算法的任务就是利用训练语料作为输入,训练出参数向量 $\boldsymbol{w}$.

对联合模型来说,在线学习算法是一个很有吸引力的算法,因为它有快速收敛的优点. 本文所关注的在线学习算法为平均感知器算法. 感知器算法的具体过程如图 2 所示.

在平均感知器算法中,x_i 表示输入句子,y_i 表示输入句子正确的分词和序列化标注的结果;$GEN(x_t)$ 表示输入句子 x_t 所有可能的候选切分和标注的集合;z_i 则表示所有结果当中得分最高的切分和标注. 如果 z_i 和正确结果 y_i 不一致,那么对权值向量 $\boldsymbol{w}$ 进行更新. 模型通过解码算法获取得分最高的结果 z_i. 训练结束之后,对所有参数取平均值,最终用平均权值作为最终判别准则的权值. 参数平均化可以避免由于学习速率过大所引起的训练过程中震荡现象的出现.

2.4　特征模板的设计

联合模型方法可以同时处理文本的分词和嵌套命名实体识别,所以我们的特征模板要能够有针对性的提高分词和命名实体识别的效果. 文献[8]中的特征模板能够很好的处理分词问题,所以本文的基本特征模板利用了文献[8]设计的部分特征模板,并结合命名实体识别问题的特性设计一个有效的特征模板集合.

首先,中文人名一般由姓和名两部分组成,姓比较稳定,我们可以将常用的姓氏抽取出来,加入特征模板中,提高分词的准确性. 嵌套命名实体一般是结构复杂的地名和组织机构名,它们内部也包含一些常现的特征词,如组织机构名的特征词“公司、学校”,地名的特征词“省、市”等. 所以本文选取了常用的特征词来构造特征词模板,提高识别效果. 本文选取的常用特征词有地名特征词和组织机构名特征词.

```
Input: raw sentence sent
Algorithm:
    update=false
    Clear(agenda)
    Clear(agenda1)
    AddItem (agenda,"")
    AddItem (agenda1,"")
    for index in[0…Length(sent)]:
        for cand in agenda:
            new←Append(cand,sent[index])
            AddItem (agenda, new)
            for category in TAGSET():
                new←Separate(can,sent[index],type)
                AddItem (agenda, new)
            agenda←N-Best(agenda)
            for cand in agenda
                if  cand! =gold-standard[0:index]
                    update=true
                if  update
                    AddItem (agenda1, cand)
        return MaxViolation(agenda1)
```

图 1　优化后的解码算法

Fig. 1　Modified docoding algorithm

Input: Training examples (x_i, y_i)

1: set $\boldsymbol{w}=0$

2: for $t=1\cdots T, i=1\cdots n$:

3: $z_i=\arg\max_{y\in GEN(x_i)} \boldsymbol{\Phi}(y) w_i$

4: if $z_i \neq y_i$

5: $\boldsymbol{w}=\boldsymbol{w}+\boldsymbol{\Phi}(y_i)-\boldsymbol{\Phi}(z_i)$

Output: $\boldsymbol{w}$

图 2　感知器算法的学习流程

Fig. 2　The perceptron learning algorithm

其次,嵌套命名实体作为一种专有名词,具有一定的上下文语言环境. 嵌套命名实体的上下文信息主要是边界词,如“位于、坐落在”常常作为地名的右边界词,“接管、任职于”常常作为组织机构名的左边界词. 在自然语言处理中,互信息 $I(x,y)$ 常常被用作为描述两个字或者词之间关联程度大小的度量. 本文采用基于互信息的方法对嵌套实体的左右边界词进行选取. 本文主要选取了基本地名、组织机构名的左右边界词以及嵌套地名、嵌套组织机构名的左右边界词.

最后,嵌套命名实体由词和基本命名实体组成,其构成有一定的规律. 通过分析语料库,我们总结了嵌套命名实体的几种普遍的构成形式. 对于嵌套地名来说,其一般的构成形式有如下几种:

- 地名+地名特征词
- 地名+地名+地名特征词
- 地名+其他词+地名特征词

- 地名+地名

嵌套的组织机构名比嵌套地名更加复杂,但是它们也有一些通用的构成形式:

- 地名+机构名特征词
- 地名+地名+机构名特征词
- 地名+其他词+机构名特征词
- 机构名+机构名特征词
- 其他词+机构名特征词

根据以上的考虑,本文在基本特征模板的基础上,加入了边界词特征、特征词特征以及嵌套命名实体的内部特征,用于嵌套命名实体的识别,这些特征如表 1 所示.

表 1　特征模板的定义

Table 1　Feature templates

模　　板	模 板 意 义
S(start(w_0))	检查当前词的开头字是否为人名的姓氏
PLB(w_{-1})	检查当前词的前一个词是否为人名的左边界词
LLB(w_{-1})	检查当前词的前一个词是否为基本地名的左边界词
OLB(w_{-1})	检查当前词的前一个词是否为基本机构名的左边界词
PRB(w_0)	检查当前词是否为人名的右边界词
LRB(w_0)	检查当前词是否为基本地名的右边界词
ORB(w_0)	检查当前词是否为基本机构名的右边界词
LLB(w_{-1})TL(w_0)	检查当前词是否包含地名特征词,并且其前一个词是否为地名左边界词
OLB(w_{-1})TO(w_0)	检查当前词是否包含组织名特征词,并且其前一个词是否为组织名左边界词
LRB(w_0)TL(w_{-1})	检查当前词的前一个词是否包含地名特征词,并且当前词是否为地名的右边界词
ORB(w_0)TO(w_{-1})	检查当前词的前一个词是否包含机构名特征词,并且当前词是否为机构名的右边界词
$w_{-1}w_0t_{-1}$ where t_{-1} = 'ns' and TL(w_0) = true	当前词包含地名特征词并且当前词的前一个词是地名,组合词和标签
$w_{-2}t_{-2}w_{-1}t_{-1}w_0$ where t_{-2} = 'ns' and t_{-1} = 'ns' and TL(w_0) = true	当前词包含地名特征词并且当前词的前两个词均是地名,组合词和标签
$w_{-2}t_{-2}w_0$ where t_{-2} = 'ns' and TL(w_0) = true	当前词包含地名特征词并且当前词的前面第二个词是地名,组合词和标签
$w_{-1}t_{-1}w_0t_0$ where t_{-1} = 'ns' and t_0 = 'ns'	当前词和当前词的前一个词均为地名,则组合这两个词以及它们的标签
$w_{-1}t_{-1}w_0$ where t_{-1} = 'ns' or t_{-1} = 'nt' and TO(w_0) = true	当前词包含机构名特征词,并且当前词的前一个词为地名或者机构名,组合词和标签
$w_{-2}t_{-2}w_{-1}t_{-1}w_0$ where t_{-2} = 'ns' and t_{-1} = 'ns' and TO(w_0) = true	当前词包含机构名特征词,并且当前词的前两个词都是地名,组合这些词和标签
$w_{-2}t_{-2}w_0$ where t_{-2} = 'ns' and TO(w_0) = true	当前词包含机构名特征词,并且当前词的前面第二个词是地名,组合这些词和标签
$w_{-2}w_{-1}w_0$ where t_{-2} contains('_B') and t_{-1} contains('_I') and t_0 contains(_E)	当前词的前面的第二个词是嵌套命名实体的开始词,前一个词是中间词,当前词是结尾词,则组合这 3 个词
$w_{-1}w_0$ where t_{-1} contains('_B') and t_0 contains('_E')	当前词的前一个词是嵌套命名实体的开始词,且当前词是嵌套命名实体的结尾词,组合这两个词

在表 1 中,w 表示词,下标表示词与当前词的相对位置,负号表示当前词之前的位置. S 检查字符是否为人名姓氏. PLB(w)、LLB(w)、OLB(w)分别检查词 w 是否为人名、地名、机构名左边界词,PRB(w)、LRB(w)以及 ORB(w)分别检查词 w 是否为人名、地名、机构名的右边界词,TL(w)检查词 w 是否包含地名特征词,TO(w)检查词 w 是否包含机构名特征词,contains 表示包含关系.

3　实验

由北京大学标注的 1998 年 1 月份的《人民日报》语料中提供了嵌套命名实体的标注信息,因此本实验分别用序列化标注方法和联合模型方法对《人民日报》1998 年 1 月的语料进行了实验. 实验将 90% 的语料用作训练数据,10% 的语料作为测试数据. 为了更客观更准确地验证与评价算法的性能,我们在该语料上做了十折交叉验证实验.

在联合模型实验中,迭代次数影响学习算法的收敛性,而 beam 宽度对实验的精度和解码的速度有影响,因此选择合适参数对实验很重要. 实验证明迭代次数设为 28,beam 宽度设为 16 模型能较好的收敛,并

且解码速度也很快,识别效果较好.

实验的测试结果采用了常用的 3 个指标,即准确率(P)、召回率(R)和综合指标 F 值(F)来评测嵌套命名实体识别的结果,实验结果如表 2 所示.

表 2　中文嵌套命名实体识别的实验结果

Table 2　Experimental results of Chinese nested named entity recognition

	P/%	R/%	F/%
基于层次标号的方法	78.56	72.74	75.50
基于双层模型的方法	88.54	71.19	78.91
基于联合模型的方法	83.92	77.24	80.85

从表 2 我们可以看出,基于层次标号标注的方法识别效果最差,这是因为该方法只能用到字符层的信息,无法用到词汇层的信息. 相比于基于层次标号的方法,基于双层模型的方法取得了更好的效果,这是因为双层模型不仅能够用到字符层的信息,也能用到词汇层的信息,但是双层模型用到的词汇层信息会受到了第一层模型识别错误(如将嵌套命名实体中包含的基本命名实体识别错误、切分错误)的影响,因此会影响嵌套命名实体的识别效果.

本文提出的基于联合模型的方法获得了最好的效果,比基于层次标号标注的方法提高了 5.35 个百分点,比基于双层模型的方法提高了 1.94 个百分点. 效果提升的原因主要有两点:第一,联合模型能够同时处理分词和嵌套命名实体识别,其中的分词包含了对基本命名实体的识别,两者能相互促进. 嵌套命名实体作为一个整体,其包含的基本命名实体对嵌套命名实体具有指示作用,联合模型能够将正确识别的基本命名实体作为特征,为嵌套命名实体的识别提供决策支持;第二,联合模型方法用到了命名实体的特征词特征、边界词特征以及嵌套结构的内部特征,这些特征能够帮助进行嵌套命名实体识别的决策,从而有效地提高了识别效果.

4　结论

嵌套命名实体识别是一个研究得较少但富有挑战性的问题. 本文针对嵌套命名实体的特点,提出了一个基于联合模型的嵌套命名实体识别方法,并在此基础上设计了丰富的特征模板. 通过对人民日报语料的实验,该方法获得了比序列化标注方法更好的效果. 由于语料库的限制,我们的方法只考虑了两层嵌套命名实体的识别. 下一步我们将考虑多层嵌套命名实体的识别.

[参考文献]

[1] 刘非凡,赵军,徐波. 实体提及的多层嵌套识别方法研究[J]. 中文信息学报,2007,21(2):14-21.

[2] Fu Chunyuan, Fu Guohong. Morpheme-based Chinese nested named entity recognition[C]//The 9th International Conference on Fuzzy System and Knowlodge Discovery. Chendu: IEEE, 2012: 2 546-2 550.

[3] 周俊生,戴新宇,尹存燕,等. 基于层叠条件随机场的中文机构名自动识别[J]. 电子学报,2006,34(5):804-808.

[4] Alex B, Haddow B, Grover C. Recognising nested named entities in biomedical text[C]//Biological, Translational, and Clinical Language Processing, 2007: 65-72.

[5] Jenny Rose Finkel, Christopher D. Manning. Nested named entity recognition[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Singapore: ACL, 2009: 141-150.

[6] Lafferty J, McCallum A, Pereira F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data[C]//Proceedings of the International Conference on Machine Learning. Williamstown: Morgan Kaufmann, 2001: 282-289.

[7] Yue Zhang, Stephen Clark. Joint word segmentation and POS tagging using a single perceptron[C]//Proceedings of ACL-HLT. Columbus: ACL, 2008: 888-896.

[8] Yue Zhang, Stephen Clark. A fast decoder for joint word segmentation and POS-tagging using a single discriminative model[C]//Proceedings of EMNLP. Cambridge: ACL, 2010: 843-852.

[9] Liang Huang, Suphan Fayong, Yang Guo. Structured perceptron with inexact search[C]//Proceedings of NAACL. Canada: ACL, 2012: 142-151.

[责任编辑:陆炳新]