基于类语言模型的中文机构名称自动识别*)

尹继豪[1] 樊孝忠[1] 于江德[1,2]

(北京理工大学计算机科学技术学院 北京 100081)[1] (安阳师范学院计算机科学系 安阳 455000)[2]

摘 要 提出了一种基于类语言模型的中文机构名称自动识别方法,将分词和机构名称自动识别有机地结合起来。在机构名称识别的类语言模型中采用等级结构,使得嵌套有人名、地名等实体的机构名称能够较好地识别出来。在实验过程中,逐步增加实验条件,依次加入启发信息、缓存模型和机构名缩写处理,使得实验结果显著提高。在开放测试中,中文机构名称最终识别的查准率和查全率分别为85.47%和72.81%。

关键词 类语言模型,中文机构名称识别,启发信息,Viterbi算法

Chinese Organization Name Automatic Recognition Using Class-based Language Model

YIN Ji-Hao[1] FAN Xiao-Zhong[1] YU Jiang-De[1,2]

(Department of Computer Science and Engineering, Beijing Institute of Technology, Beijing 100081)[1]

(Department of Computer Science, Anyang Teachers' College, Anyang, Henan 455000)[2]

Abstract An approach based on class language model is put forward about Chinese organization automatic recognition. Word segmentation and organization recognition can be combined. We adopted a hierarchical structure in organization model so that the organization names including nested entities can be identified. In the experiments, we integrate heuristic information, cache model and organization abbreviation processing into the class-based language model. The precision and recall of Chinese organization recognition are 85.47% and 72.81% in the opened-test.

Keywords Class-based language model, Chinese organization name recognition, Heuristic information, Viterbi search

1 前言

命名(实体识别是自然语言处理中的一项基础性工作,同样是句法分析、机器翻译、信息抽取等任务的一个非常重要的预处理模块。一般来说,命名实体识别的任务就是对一篇待处理文本,识别出其中人名、地名、机构名、数字表达式和时间表达式这四类命名实体[1]。其中人名、地名、机构名识别是最难、也是最重要的三类。

人们已经对人名和地名的识别作了非常细致的研究,提出了各种各样的处理方法。人名和地名识别的查准率和查全率已经能满足人们的需求,但是机构名的识别无论从理论上还是从实际上,都未能达到应用要求[2]。

2 类语言模型

首先介绍类的定义[3]。把三类命名实体,即人名、地名、机构名定义为三个类,如表1。同时,定义任何一个字或词单独成一类,这是为了能把命名实体识别和中文分词有机结合在一起,所以,在基于类的语言模型中,共定义了|V|+3个类,其中|V|为词典中的词数。

表1 类模型中类的定义

符号	描述
PN	人名(Person Name)
LN	地名(Location Name)
ON	机构名(Organization Name)

2.1 类语言模型定义

给定中文输入序列 $S=S_1\cdots S_n$,中文命名实体识别的任务是搜索最优的类序列 $C^*=c_1\cdots c_m(m<=n)$,使概率 $P(C|S)$最大。即

$$C^*=\arg\max_C P(C|S)=\arg\max_C P(C)\times P(S|C) \quad (1)$$

类模型包括两个子模型:上下文相关模型 $P(C)$和实体模型 $P(S|C)$。上下文相关模型表示给定上下文信息的情况下产生某一实体类的概率。$P(C)$可用下式表示:

$$P(C)\cong\prod_{i=1}^{m}P(c_i|c_{i-2}c_{i-1}) \quad (2)$$

$P(C)$可以使用标注好的语料库训练得到。

实体模型 $P(S|C)$表示在给定某一实体类 C 的条件下,字符串 S 的生成概率,即

$$\begin{aligned}P(S|C)&=P(s_1\cdots s_n|c_1\cdots c_m)\\&\cong P([s_1\cdots s_{c_1-end}]\cdots[s_{c_1-start}\cdots s_n]|c_1\cdots c_m)\\&\cong\prod_{j=1}^{m}P([s_{c_j-start}\cdots s_{c_j-end}]|c_j)\end{aligned} \quad (3)$$

在给定实体类 c_j 的条件下,实体模型能够估算字符序列 $s_{c_j-start}\cdots s_{c_j-end}$的生成概率。

2.2 机构名的类语言模型

中文机构名的类模型比较复杂,因为机构名中经常包括人名或地名等未登录词,例如机构名"中国工商银行"中包括了地名"中国"。

为了更好地建立机构名称的类语言模型,首先给出人名和地名的类语言模型表达式。对于人名,采用基于字的三元模型,如下式:

$$\begin{aligned}&P([s_{c_j-start}\cdots s_{c_j-end}]|c_j=PER)\\&=\prod_{k=c_j-start}^{c_j-end}P(s_k|s_{k-2},s_{k-1},c_j=PER)\end{aligned} \quad (4)$$

式中PER表示人名。

*)教育部博士点基金项目(20050007023)。**尹继豪** 博士研究生;**樊孝忠** 教授,博士生导师。

对于地名，采用基于词的三元模型，如下式：

$$P([s_{c_j-start}\cdots s_{c_j-end}]|c_j=LOC)$$
$$\cong\max_{W}P(w_1\cdots w_l|c_j=LOC)$$
$$=\max_{W}[\prod_{k=1}^{l}P(w_k|w_{k-2},w_{k-1},c_j=LOC)] \quad (5)$$

式中 LOC 表示地名。$W=w_1\cdots w_l$ 是字符序列 $s_{c_j-start}\cdots s_{c_j-end}$ 的任何一个可能切分结果。

下面给出机构名的实体模型。为了识别机构名中的人名或地名，我们采用了等级结构的类语言模型，这个类模型中包括三个子模型：一个是类生成模型，另外两个是人名实体模型和地名实体模型。机构名的实体模型用式(6)表示：

$$P([s_{c_j-start}\cdots s_{c_j-end}]|c_j=ORG)$$
$$\cong\max_{C}[P(C'|c_j)P([s_{c_j-start}\cdots s_{c_j-end}]|C',c_j=ORG)]$$
$$=\max_{C'}[P(C'_1\cdots C'_k|c_j=ORG)\times P([s_{c_j-start}\cdots s_{c_j-end}]|c'_1\cdots c'_k,c_j=ORG)] \quad (6)$$
$$\cong\max_{C'}[\prod_{i=1}^{k}P(c'_i|c'_{i-2}c'_{i-1},c_j=ORG\times\prod_{i=1}^{k}P([s_{c'_i-start}\cdots s_{c'_i-end}]|c'_1,c_j=ORG]$$

式中 ORG 表示机构名，$C'=c'_1\cdots c'_k$ 是对应字符序列的任何一个类的序列。

基于上下文模型和实体模型，可以计算出概率 $P(C|S)$，并且得到最佳的类序列。

3 类语言模型在中文机构名识别中的应用

3.1 算法实现

要识别出文本中的中文机构名称，算法实现包括以下三步：

步骤一：基于现有词典(《人民日报》1998 年 1 月到 6 月的语料)对文本进行全切分，词典只用作分词。

步骤二：机构名称可以从一个或多个已切分出的词中产生，并且机构名的产生概率可通过机构名实体模型(6)式计算。

步骤三：引入 Viterbi 搜索是为了寻找给定中文输入序列上概率最大的词类序列。为了识别嵌套有人名或地名的机构名称，采用两次 Viterbi 搜索，内部 Viterbi 搜索对应式(6)，可以得到候选机构名称内部的最优类序列；外部 Viterbi 搜索对应式(1)，可以得到最优的机构名称类序列。经过这两次搜索，分词和机构名识别便可顺利完成。

3.2 算法改进

单纯地使用类语言模型进行命名实体识别存在一些问题。第一，为了识别实体，不得不计算文本中每一个词作为实体的概率，因此，多余候选实体增加了算法的复杂度。第二，数据稀疏问题将严重影响实体识别结果。第三，缩写的机构名称不能被有效地识别。下面分别介绍解决这三个问题所采用的方法。

3.2.1 启发信息和机构名称模板

为了避免多余的实体产生，在类语言模型中引入了启发信息。机构名关键字词表包括 1355 个词(如：大学、公司)，这些关键字和它前面的 2 至 6 个词可以被认为构成了一个中文机构名称。启发信息用来约束实体的产生，一个候选机构名如果不包括机构名关键字词表中的词，就可以直接排除。

下面给出机构名称的模板：

{[人名][地名][核心词]}[机构类型]〈关键词〉

其中：[人名]、[地名]、[核心词]可以任选。例如："美国福特汽车公司"中，[核心词]为"福特"，[机构类型]为"汽车"，〈关键词〉为"公司"。

3.2.2 缓存模型

在实体识别过程中通过不断地调整参数，缓存模型可以解决数据稀疏问题。其基本方法为：利用正在处理文本中已经出现的符号序列建立动态的一元模型 $P_{unicache}(w_i)$ 和二元模型 $P_{bicache}(w_i|w_{i-1})$，并把此模型和静态模型 $P_{static}(w_i|w_1\cdots w_{i-2}w_{i-1})$ 联合在一起得到一个综合模型。即

$$P_{cache}(w_i|w_1\cdots w_{i-2}w_{i-1})=\lambda_1P_{unicache}(w_i)+\lambda_2P_{bicache}(w_i|w_{i-1})+(1-\lambda_1-\lambda_2)P_{static}(w_i|w_1\cdots w_{i-1}) \quad (7)$$

其中 $\lambda_1,\lambda_2\in[0,1]$，此系数可以通过数据集训练得到。

3.2.3 机构名缩写处理

我们发现机构名称缩写识别的错误率较高，因此要采用一定的方法解决缩写识别问题，在此采用比较简单的基于规则的方法[4]。中文机构名称缩写包括连续缩写和间隔缩写，如："北京华联超市股份有限公司"，它的缩写可能舍弃机构类型"超市"、"股份"、"有限"，以及关键字"公司"，只保留地名"北京"和核心词"华联"。表 2 列出了机构名缩写的一些例子。

表 2 嵌套机构名称及其缩写名称

连续缩写	北京华联超市股份有限公司	北京华联
	清华大学	清华
间隔缩写	上海证券交易所	上证
	北京理工大学	北理工

在识别机构名缩写时，提取机构名称核心词是至关重要的。而且，若机构不太著名，则该机构名称的缩写通常在其全称出现后才出现，所以这种识别机构名称缩写的方法是有效的。

4 实验与分析

4.1 评测标准和数据集

本实验采用查准率(P)、查全率(R)和 F1 值作为评测标准。

$$P=\frac{\text{正确识别的机构名称数目}}{\text{识别出的机构名称数目}}$$
$$R=\frac{\text{正确识别的机构名称数目}}{\text{所有的机构名称数目}}$$
$$F=\frac{(\beta^2+1)\times P\times R}{(\beta^2\times P)+R}(\beta=1)$$

表 3 新华网中文简体版

ID	领域	中文机构名称个数	文件大小
1	体育	643	117k
2	财经	356	103k
3	时政	227	126k
4	文娱	152	108k
5	教育	97	145k
	总计	1475	599k

在封闭测试和开放测试中，选用的训练集是《人民日报》1998 年 1 月到 6 月人工标注的语料，包括 357,544 个句子(大约 9,200,000 个汉字)，包含 104,487 个人名，218,904 个地名，87,391 个机构名。

封闭测试所用测试集是《人民日报》1998 年 1 月的语料。开放测试所用测试集是由"新华网——全球新闻网"中文简体

版网页生成的标注文本(如表3),包括5个领域。

4.2 实验步骤

在实验过程中,依次增加实验条件,实验步骤如下:

(1)使用基于类的语言模型进行机构名称识别,把它看作基本的实验结果;

(2)在实验(1)的基础上加入启发信息;

(3)在实验(2)中加入缓存模型;

(4)在实验(3)的基础上加入中文机构名称缩写处理过程。

4.3 实验结果与分析

基于以上实验步骤,分别进行封闭测试和开放测试。

4.3.1 类语言模型

未增加实验条件的情况下,对基于类的语言模型进行了封闭测试和开放测试,得到了表4的实验数据。实验结果表明封闭测试的查全率要明显高于开放测试的查全率。

表4 测试一实验结果

语料	查准率(%)	查全率(%)	F1(%)
人民日报1月	68.23	67.83	68.03
新华网中文简体版	60.19	55.57	57.79

4.3.2 加入启发信息

加入启发信息后,实验结果如表5。实验中发现加入启发信息后,不仅解码简单,而且获得了较高的查准率。如:开放测试中机构名称的识别查准率从60.19%提高到86.58%,原因在于采用了启发信息后减少了噪声干扰。

然而,我们注意到开放测试中机构名称识别的查全率有所下降,这是因为有些机构名称缺省关键字,以致于在识别过程中漏掉这部分机构名称,如:"太平洋比特公司"的简写"太平洋比特"无法识别为机构名称。

表5 测试二实验结果

语料	查准率(%)	查全率(%)	F1(%)
人民日报1月	88.31	69.22	77.61
新华网中文简体版	86.58	52.86	65.64

4.3.3 加入缓存模型

表6显示了加入缓存模型后的实验结果。比较表5和表6,不难发现不论封闭测试还是开放测试,查准率和查全率都略有提高。

表6 测试三实验结果

语料	查准率(%)	查全率(%)	F1(%)
人民日报1月	90.14	72.56	80.40
新华网中文简体版	87.42	53.67	66.51

4.3.4 加入机构名称缩写处理

在这一步实验中,加入了机构名称缩写处理,实验结果如表7。比较表6和表7,发现开放测试中机构名查全率从53.67%提高到72.81%。

表7 测试四实验结果

语料	查准率(%)	查全率(%)	F1(%)
人民日报1月	91.42	79.01	84.76
新华网中文简体版	85.47	72.81	78.63

4.3.5 实验结果分析

通过以上实验,得到以下结论:(1)封闭测试中机构名称识别的查准率和查全率都明显高于开放测试;(2)加入启发信息后,机构名识别查准率有显著提高;(3)引入机构名缩写处理有利于提高机构名称识别的查全率。

我们还发现引入缓存模型后实验结果没有明显的提高,这因为缓存语言模型基于这样的假设:一个机构名一旦在文本中某一位置出现,则它在上下文中出现的频率就会较高。然而在实际应用中发现好多机构名称在上下文中会改变书写方式,如:"中国共产党第十六次代表大会",在上下文中可用"中共十六大"和"十六次党代会"等表示。

结论 在实验中,通过使用基于类的语言模型,我们把中文分词和机构名称识别有机地结合在一起;并且,在中文机构名识别模型中,采用了等级结构,以便识别嵌套有人名、地名等实体的中文机构名。

启发信息的引入提高了机构名称识别的查准率;缓存模型的加入也在一定程度上提高了机构名称识别的查准率和查全率;而且,把机构名称缩写的一些规则引入到类语言模型中也显著地提高了实验结果。在开放测试中,中文机构名称最终识别的查准率为85.47%,查全率为72.81%,这高于文[5,6]中机构名称识别的查准率和查全率。

在今后的工作中,我们将把基于类的语言模型扩充到特定领域的专有名称识别系统中,并且使模型具备更强的稳定性和可移植性。

参考文献

1 孙斌. 信息提取技术概述[J]. 语言信息处理, 2003, 2(1): 34~35

2 Yu Hongkui, Zhang Huaping. Recognition of chinese organization name based on role tagging [A]. In: Proceedings of 20th International Conference on Computer Processing of Oriental Languages [C]. Beijing, China, 2003. 79~87

3 Sun Jian, Gao Jianfeng. Chinese named entity indentification using class-based language model [A]. In: Proceedings of the 19th International Conference on Computational Linguistics [C]. Taipei, China, 2002. 24~25

4 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(4): 21~32

5 Chen Keh-Jiann, et al. Knowledge Extraction for Identification of Chinese Organization Names [A]. In: Proceedings of ACL Workshop on Chinese Language Processing [C], 2000. 15~21

6 Wu Youzheng, et al. Chinese Named Entity Recognition Combining a Statistical Model with Human Knowledge [A]. In: Proceeding of the Workshop on Multilingual and Mixed-language Named Entity Recognition [C], July 2003. 65~72