基于 Tri-training 半监督学习的中文组织机构名识别*

蔡月红[a,b], 朱 倩[a], 程显毅[a]
(江苏大学 a. 计算机科学与通信工程学院; b. 外语学习中心, 江苏 镇江 212013)

摘 要: 针对中文组织机构名识别中的标注语料匮乏问题,提出了一种基于协同训练机制的组织机构名识别方法。该算法利用 Tri-training 学习方式将基于条件随机场的分类器、基于支持向量机的分类器和基于记忆学习方法的分类器组合成一个分类体系,并依据最优效用选择策略进行新加入样本的选择。在大规模真实语料上与 co-training 方法进行了比较实验,实验结果表明,此方法能有效利用大量未标注语料提高算法的泛化能力。

关键词: 中文组织机构名; 半监督学习; 协同训练; Tri-training
中图分类号: TP391 **文献标志码**: A **文章编号**: 1001-3695(2010)01-0193-03
doi:10.3969/j.issn.1001-3695.2010.01.057

Chinese organization names recognition with Tri-training learning

CAI Yue-hong[a,b], ZHU Qian[a], CHENG Xian-yi[a]
(*a. School of Computer Science & Communication Engineering, b. Foreign Language Learning Center, Jiangsu University, Zhengjiang Jiangsu* 212013, *China*)

Abstract: In view of the data scarcity problem in for Chinese organization names recognition, this paper presented a co-training style method for Organization Names Recognition. And proposed a novel selection method for Tri-training learning, using three classifiers: CRFs, SVMs and MBL. In Tri-training process, selected new newly labeled samples based on the selection model maximizing training utility, and computed the agreement according to the agreement scoring function. Experiments on large-scale corpus show that the proposed Tri-training learning approach can more effectively and stably exploit unlabeled data to improve the generalization ability than co-training and the standard Tri-training.

Key words: Chinese organization name recognition; semi-supervised learning; co-training; Tri-training

0 引言

命名实体识别(named entity recognition)是信息处理中最为基础的关键技术之一。命名实体是指现实世界中具体的或抽象的实体,通常有人名、地名、组织机构名、日期、时间、货币、百分数七类。其中,组织机构名识别是命名实体识别研究的重点,也是难点,特别是中文组织机构名由于中文构词方式的灵活性和词性信息的不明确性,其识别更为困难。

中文组织机构名识别算法的研究方法早期是基于规则的[1,2],近年来研究的主流是基于统计学习的。郑家恒等人[3]提出了基于隐马尔可夫模型的方法,冯冲等人[4]将最大熵模型用于组织机构名的识别,周俊生等人[5]提出了基于条件随机场模型的方法,陈霄等人[6]将基于支持向量机的方法用于组织机构名的识别。上述方法都是基于监督学习的,此类方法为保证泛化能力,需要大量的标注语料做训练集。语料库的人工标注是很费时费力的,而大量的未标注语料却很容易获取,这就是所谓的标注瓶颈问题。因此如何利用大量的未标注语料来改善学习性能已成为基于机器学习的组织机构名识别研究中最受关注的问题之一。

本文探求了利用半监督学习技术克服中文组织机构名识别中的标注语料匮乏这一困难的途径,实现了一个基于样本最优效用选择策略的 Tri-training 半监督学习算法。实验结果表明,该方法在组织机构名的识别中能有效利用大量的未标注语料来提高性能。

1 组织机构名分析

组织机构名是泛指机关、团体等实体的名称。虽然组织机构名没有人名、地名那样明确的特点和固定的用词,但也有一定的组成特点。完整的组织机构名通常由一个或一个以上的机构名前部词加上一个机构名后缀词(如大学、协会等)组成。

1.1 标注方法

对于标注集,考虑到机构名后缀词的独特作用,本文融合了组块分析中的"BIO"和"IOE"标注方法,引入了四个标注符号,即 BISO,这样就可以将机构名识别问题转换为序列化的标注问题,即对于给定输入词序列 $X=\{w_1,w_2,\cdots,w_n\}$ 及其词性标注,给出每个词的 BISO 标注。这四个标记的含义如下:B 表示这个词在组织机构名的开始;I 表示这个词在组织机构名的中间;S 表示这个词为组织机构名的后缀;O 表示这个词不属于任何组织机构名。

这样就可以将组织机构名识别看成一个在上面定义的四类标记中进行类别判断的分类问题。

收稿日期: 2009-04-02; **修回日期**: 2009-05-13 **基金项目**: 国家自然科学基金资助项目(60702056)

作者简介:蔡月红(1969-),女,江苏兴化人,工程师,硕士研究生,主要研究方向为自然语言处理、机器学习等(caiyh@ujs.edu.cn);朱倩(1979-),女,江苏镇江人,讲师,博士研究生,主要研究方向为自然语言处理、模式识别;程显毅(1956-),男,黑龙江哈尔滨人,教授,博导,主要研究方向为模式识别、自然语言理解.

1.2 特征选择

特征选择的目的是寻找有助于识别组织机构名的文本属性。本文选择特征时主要考虑词本身的特征和词邻近的上下文特征,以命名实体及其上下文的词性、机构名的标注作为样本选择的基本单元,上下文窗口大小定为[-2, 2]。用于后缀词标注的组织机构名后缀词表的构建方法如下:应用统计方法从训练语料中获取初始组织机构名后缀词表,并将每次 Tri-training 过程中新选择的样本加入词表以更新后缀词表。

这样,最终的样本特征确定为

$$X=\{w_{i-2},p_{i-2},t_{i-2},w_{i-1},p_{i-1},t_{i-1},w_i,p_i,t_i,w_{i+1},p_{i+1},t_{i+1},w_{i+2},p_{i+2},t_{i+2}\}$$

其中:w_i 表示当前位置的词;p_i 表示该词的词性标注;t_i 表示该词的 BISO 标注;w_{i-k}、p_{i-k}、$t_{i-k}(k=1,2)$ 表示前 k 个位置的词、词性标注及 BISO 标注;w_{i+k}、p_{i+k}、$t_{i+k}(k=1,2)$ 表示后 k 个位置的词、词性标注及 BISO 标注。

2 基于 Tri-training 的组织机构名识别

半监督学习的本质是利用大量未标注样本提高对某些相关统计分布估计的准确性。Tri-training 算法[7]是协同训练模式的半监督学习算法。该算法使用了三个分类器,首先对有标注样本集进行可重复取样以获得三个有标记训练集,然后从每个训练集产生一个分类器,在协同训练过程中,各分类器所获得的新标注样本都由其余两个分类器协作提供。在对未标注样本进行预测时,Tri-training 算法不再像以往算法那样挑选一个分类器来使用,而是使用集成学习中经常用到的投票法来将三个分类器组成一个集成来实现对未标注样本的预测。该算法既不要求充分冗余视图也不要求使用不同类型分类器,而且不必使用交叉验证,因此适用范围更广、效率更高。

2.1 Tri-training 算法

在应用 Tri-training 算法时,需要选取三个初始分类器,为使初始分类器具有一定的差异性,本文实验中分别选取了基于条件随机场的分类器 CRFs、基于支持向量机的分类器 SVMs 及基于记忆学习方法的分类器 MBL。

在文献[7]中,Tri-training 训练结束时由训练所得联合分类器$\{H_1,H_2,H_3\}$采用多数投票规则对新数据进行分类。本文考虑到初始分类器性能的差异性,采用基于性能的集成方法,算法中采用式(1)所示的加权投票规则对训练所得联合分类器进行类别标记,在集成时考虑每个分类器的性能权重,权重由三个分类器在初始带标注语料 L 上的分类准确率 $P_i(L)$ 所决定。

$$H(1,2,3)(x)=\arg\max_{y\in \text{label}}\frac{\sum_{i=1}^{3}\delta(y,H_i(x))\times P_i(L)}{\sum_{i=1}^{3}P_i(L)} \tag{1}$$

其中:$\delta(y,H_i(x))=\begin{cases}1 & H_i(x)=y\\ 0 & H_i(x)\neq y\end{cases}$。

基于 Tri-training 的组织机构名识别算法伪代码如下所示,每次迭代时使用的未标注语料取之用于缓存一部分未标注语料的缓存器。

算法:基于 Tri-training 的组织机构名识别算法

输入:初始带标记样本集 L;未标记样本集 U;分类器 H_1、H_2、H_3。

输出:最终标注结果。

a)初始化

$L_1^0\leftarrow L,L_2^0\leftarrow L,L_3^0\leftarrow L,S\leftarrow L$

$M_1^0\leftarrow\text{train}(H_1,L_1^0)$;

$M_2^0\leftarrow\text{train}(H_2,L_2^0)$;

$M_3^0\leftarrow\text{train}(H_3,L_3^0)$。

b)循环

(a)对 S 进行 Bootstrap 采样,产生三个训练集 L_1^i,L_2^i,L_3^i;

(b)$U^i\leftarrow$add unlabeled data from U;

(c)由 M_1^i、M_2^i、M_3^i 对 U^i 进行标记,并依照最优效用选择策略选择标记样本子集$\{P_1\}$、$\{P_2\}$和$\{P_3\}$;

(d)生成 H_i 的新训练集 $L_1^{i+1}\leftarrow L_1^i+\{P_1\}$、$L_2^{i+1}\leftarrow L_2^i+\{P_2\}$、$L_3^{i+1}\leftarrow L_3^i+\{P_3\}$;

(e)$M_1^{i+1}\leftarrow\text{train}(H_1,L_1^{i+1})$

$M_2^{i+1}\leftarrow\text{train}(H_2,L_2^{i+1})$

$M_3^{i+1}\leftarrow\text{train}(H_3,L_3^{i+1})$;

(f)$S\leftarrow L_1^{i+1}+L_2^{i+1}+L_3^{i+1}$;

(g)联合分类器$\{H_1,H_2,H_3\}$按加权投票规则对 S 中新标记数据重新分类标记。

重复上述过程,直到 U 为空。

2.2 样本选择策略

本文中机构名识别问题是一个序列化的标注问题,在 Tri-training 训练过程的每次迭代中,基于最优化选择机制进行新加入样本的选择[8],即尽可能选择具有最优训练效用和最小化误差的新标注样本,并基于一致性评价函数判断两个分类器的一致性。

2.2.1 一致性评价函数

对于任意给定的数据序列 $X=\{x_1,x_2,\cdots,x_n\}$,如果用两个分类器对它进行标注,得到两个标注序列 $Y_1=\{y_{11},y_{21},\cdots,y_{n1}\}$及 $Y_2=\{y_{12},y_{22},\cdots,y_{n2}\}$,那么一致性评价函数 Ag 为

$$Ag=\sum_{i=1}^{n}(y_{i1},y_{i2})/n \tag{2}$$

其中:$f(y_{i1},y_{i2})=\begin{cases}1 & y_{i1}=y_{i2}\\ 0 & y_{i1}\neq y_{i2}\end{cases}$

一致性评价函数 Ag 的值表示两个标注序列也即两个分类器之间的一致性,Ag 越大说明一致性最高。

2.2.2 样本选择方法

文献[7]中各分类器所获得的新标注样本都由其余两个分类器协作提供,而在本文中为了尽可能选择具有高训练效用的新标注样本,提出了改进的样本选择方法。假设三个不同分类器 H_1、H_2 和 H_3,x 是未标注语料内任一样本,那么新标注样本选择准则如下:

a)如果 H_2 和 H_3 对 x 的分类结果一致,那么就认为该标注结果是正确的。

b)如果目标分类器 H_1 对 x 的分类结果和其他两个分类器(H_2、H_3)不一致,那么就认为该标注结果不能正确地训练 H_1。

结合上述两个准则,每次迭代过程中,各分类器所获得的新标注样本的选择方法如下(假设 C_1、C_2、C_3 表示分类器 H_1、H_2 和 H_3 对缓存器中未标注样本的分类结果):

a)计算 C_j、$C_k(j,k\neq i)$中所有样本的一致性,按比例选择出一致性函数值最高的样本子集。

b)计算 C_i、C_j 中所有样本的一致性,按比例选择出一致性函数值最低的样本子集。

c)取这两个样本子集的交集交给分类器 H_2 重新标注,生成分类器 H_1 的新样本集 L_1^{new}。

本文称该样本选择策略为最优效用选择策略。

3　实验结果及分析

3.1　实验语料

本文实验语料由人民日报语料库 1998 年 1 月语料和国家“863”计划 2004 年命名实体识别评测简体语料组成。首先去除对学习器的训练或是测试都没有帮助的无组织结构名出现的部分,经筛选过的语料规模为 12 791 条(其中人民日报语料库 1998 年 1 月语料为 8 326 条,国家“863”计划 2004 年命名实体识别评测简体语料为 4 465 条);然后对每个语料,随机取出 25% 做测试集,15% 作为最初的训练集 L,并对这两部分语料的标注体系进行转换,用 BIOS 格式对原来标注为“[…]nt”格式重新进行标注,其余的 60% 语料作为未标注集 U 用于 Tri-training。

3.2　性能评测

实验中的性能评测指标采取了 MET 和 MUC 规定的度量方式:正确率 P、召回率 R、综合反映两者的综合指标 F(β取 1),计算公式如下:

$$P = N_2/N_1 \times 100\% \quad (3)$$

$$R = N_2/N_3 \times 100\% \quad (4)$$

$$F = (\beta^2 + 1) \times P \times R/(\beta^2 \times R + P) \quad (5)$$

其中:N_1 为实际识别的组织机构名个数;N_2 为正确标志的组织机构名个数;N_3 为标注的组织机构名的总数。这里的正确识别是指命名实体的类别和边界的标注都是正确的。

3.3　实验结果及分析

本实验在 Tri-training 中所采用的学习算法分别是 CRF + +(v0.42)、SVM-light 及 TiMBL。为了研究样本选择策略对分类性能的影响,将本文提出的样本最优效用选择策略与文献[7]的样本选择策略进行了对比实验,并且为了对比 Tri-training 算法,本文在实验中还进行了文献[9]中所介绍的 co-training 算法的测试,co-training 算法实验中新训练样本的选择也使用了基于一致性评价的样本最优效用选择策略。

3.3.1　两种样本选择策略的 Tri-training 的结果比较

图 1 给出了两种样本选择策略下各个分类器对未标注语料进行 Tri-training 学习时的综合指标 F 值随着训练过程的迭代变化情况。其中,CRFs-1 表示样本最优效用选择策略下的 CRFs 分类器模型,CRFs-2 表示文献[7]样本选择策略下的 CRFs 分类器模型,并依此类推。考察各分类器的 F 值可以看出,第一次迭代后提升幅度最大,第三次迭代后就不再显著变化。在所有情况下,三种分类器所得的最终 F 值均显著高于初始 F 值,而本文提出的样本选择策略优于文献[7]的样本选择策略,应用最优效用选择策略的三个分类器的性能均有明显提高。特别对于单分类器性能最差的 MBL 分类器的泛化能力提高尤为显著,这说明基于最优效用选择策略的 Tri-training 学习能充分利用未标注样本提高分类性能。

3.3.2　Tri-training 和 co-training 的结果比较

图 2 给出了基于最优效用选择策略的 Tri-training 和 co-training 在未标注语料上综合指标 F 值随训练过程的迭代变化情况。从图中可以看出,每次迭代过程中,Tri-training 的性能都明显优于 co-training。对比两算法的迭代过程可以发现,Tri-training 的 F 值的变化较为稳定,几乎不会出现 co-training 的波动现象,这说明基于最优效用选择策略的 Tri-training 的迭代错误率较低,能更稳定地保证泛化性能的提高,有更好的健壮性。

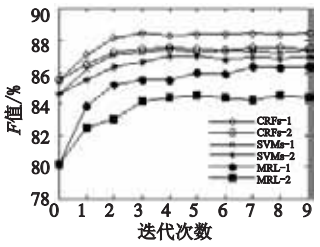

图1　不同样本选择策略的 Tri-training的结果比较

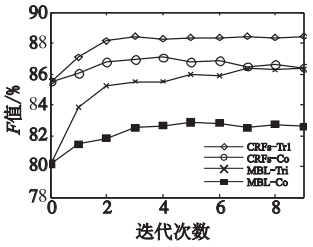

图2　Tri-training和co-training的结果比较

4　结束语

本文基于协同训练基本原理提出一种中文组织机构名识别的半监督学习算法,建立三个独立的分类器,即 CRFs 分类器、SVMs 分类器及 MBL 分类器,以 Tri-training 方式迭代地对未标注语料进行标记以扩充原训练语料,并在新标注样本选择上采用最优效用策略。实验结果表明,本方法可有效利用大量未标注语料提高泛化性能。

本文所做工作是利用半监督学习解决组织机构名识别中标注瓶颈问题的一个尝试,在本文实验中,由于初始标注语料规模小,难以训练出高精度的分类器,自动标注的样本中噪声会不可避免地随着训练的进行而不断积累。如何在迭代过程中(尤其迭代的早期)使用相关技术识别错误标记的样本是下一步研究方向。

参考文献:

[1] 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,1997,1(4):21-32.

[2] WANG Hou-feng, SHI Wu-guang. A simple rule-based approach to organization name recognition in Chinese text[C]//Proc of the 6th CICLing. Heidelberg: Springer-Verlag, 2005: 769-772.

[3] 郑家恒,张辉. 基于 HMM 的中国组织机构名自动识别[J]. 计算机应用,2002,22(11):1-2.

[4] 冯冲,陈肇雄,黄河燕. 采用主动学习策略的组织机构识别[J]. 小型微型计算系统,2006,27(4):710-714.

[5] 周俊生,戴新宇,尹存燕,等. 基于层叠条件随机场模型的中文机构自动识别[J]. 电子学报,2006,34(5):804-809.

[6] 陈霄,刘慧,陈玉泉. 基于支持向量机方法的中文组织机构名的识别[J]. 计算机应用研究,2008,25(2):362-364.

[7] ZHOU Zhi-hua, LI Ming. Tri-training: exploiting unlabeled data using three classifiers[J]. IEEE Trans on Knowledge and Data Engineering, 2005,17(11):1529-1541.

[8] STEEDMAN M, HWA R, CLARK S, *et al*. Example selection for bootstrapping statistical parsers[C]//Proc of Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology. Edmonton: Canada Association for Computational Linguistics, 2003:157-164.

[9] PHAM T P, NG H T, LEE W S. Word sense disambiguation with semi-supervised learning[C]//Proc of the 20th National Conference on Artificial Intelligence. Menlo Park: AAAI Press, 2005: 1093-1098.