基于分类器集成的兼类词消歧研究

张 一哲, 曲维光, 刘金克, 孙玉霞

(南京师范大学计算机科学与技术学院, 江苏 南京 210097)
(江苏省信息安全保密技术工程研究中心, 江苏 南京 210097)

[摘要] 兼类词词性消歧是中文词性标注的难点之一. 本文集成了支持向量机、条件随机场、最大熵等 3种分类模型, 对兼类词词性消歧进行研究. 以 1998年 1月份已标注《人民日报》为实验语料, 对 410个常见的兼类词进行开放测试, 平均精度达到 89. 69%, 取得了较好的效果.

[关键词] 兼类词消歧, 支持向量机, 条件随机场, 最大熵, 分类器集成
[中图分类号] TP391 [文献标识码] A [文章编号] 1001-4616(2010)04-0144-04

Research on Disambiguation of Multiple Syntactic Category Words Based on Ensemble of Classifiers

Zhang Yizhe, Qu Weiguang, Liu Jinke, Sun Yuxia

(School of Computer Science and Technology, Nanjing Normal University, Nanjing 210097, China)
(Jiangsu Research Center of Information Security & Privacy Technology, Nanjing 210097, China)

Abstract: One of the difficulties of Chinese word POS tagging is the disambiguation of multiple syntactic category words. In order to tackle this problem, this article tries the ensemble of three classifiers of support vector machine, maximum entropy and conditional random fields. 410 often-used examples from People's Daily corpus in January 1998 are used in the experiment, and the average precision is up to 89. 69%. This is a relative good result.

Key words: disambiguation of multiple syntactic category words; support vector machine; conditional random fields; maximum entropy; ensemble of classifies

词性标注就是在给定的句子中判定每个词的语法范畴并加以标注的过程. 词性标注的关键是排除兼类词歧义和确定未登录词的词性[1]. 所谓兼类词就是指在一定的词性标记集下, 拥有两种或两种以上的词性标记的词. 其主要难点在于: 一、汉语是一种缺乏词的形态变化的语言, 词的类别不像印欧语那样, 可以直接从词的形态变化上来判断. 二、常用词兼类现象严重, 而且越是常用的词, 不同的用法越多. 根据张虎等人对北京大学计算语言研究所在网上公布的 200万汉字语料进行的统计, 兼类词的词次占到了 47%[2].

本文主要论述了如何利用分类器集成的方法, 对兼类词词性进行投票决策, 以达到消歧目的.

1 相关工作

对于词性标注的研究可以上溯到 20世纪 60年代, 一些学者就开始对英语语料库的词类自动标注进行研究. 1993年, Mashall提出 CLAWS算法, 利用概率统计模型对 LOB语料的词类进行自动标注, 精度达到 97%, 此后, DeRose又在 CLAWS的基础上, 提出 VOLSUNGA算法, 使英语语料库的标注达到实用化. 在汉语方面, 周强提出了一种词语切分和词性标注相结合的汉语语料多级处理方法[3], 白栓虎提出了基于统计的汉语切分和词性自动标注一体化模型及实现方法[4]. 刘开瑛等利用 CLAWS和 VOLSUNGA及其变

收稿日期: 2010-06-10
基金项目: 国家自然科学基金(60773173, 61073119)、国家"973"项目(2004CB318102)、江苏省社科基金(06JSBYY001).
通讯联系人: 曲维光, 博士, 教授, 研究方向: 计算语言学和人工智能. E-mail: wgqu@njnu.edu.cn

形算法对汉语语料库进行词类自动标注, 精度达 90%. 刘群等基于层叠隐马模型研制开发的 ICTCLAS词法分析系统, 取得了显著的效果[5]. 但根据文献[6]的统计结果显示, 兼类词词性标注精度在 84% 左右, 因此, 有必要对其进行深入的研究.

2 实验设计

不同的分类器对属性特征的格式要求及其特征选择过程各不相同, 下面对本文所采用的各子分类器的特征选择分别进行介绍.

2.1 SVM 模型

SVM (Support vector machine)[7]是 Vapnik等人在 1995年首先提出的一种学习模型, 在分类方面具有良好的性能. 在自然语言处理的相关研究中, 已被广泛地应用于短语识别、词义消歧、文本自动分类和信息过滤等方面.

SVM 方法是从线性可分情况下的最优分类面中提出的. 考虑图 1所示的二维两类线性可分情况, 图 1中实心点和空心点分别表示两类的训练样本, H 是把两类无错误地分开的分类线, H_1、H_2 分别为过各类样本中离分类线最近的点且平行于分类线的直线, H_1 和 H_2 之间的距离叫做两类的分类空隙或分类间隔. 所谓最优分类线就是要求分类线不但能将两类无错误地分开, 而且要使两类的分类空隙最大. 本文实验平台采用的是台湾大学林智仁 (Lin Chih-Jen) 开发的 libSVM. (下载地址: http://www.csie.ntu.edu.tw/~cjlin/libsvm)

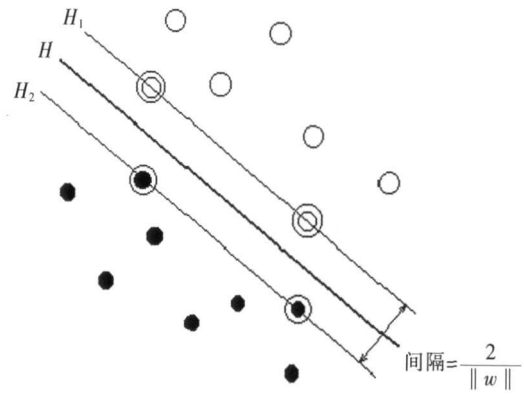

图 1 支持向量机示意图

Fig.1 Support vector machine schematic

为了抽取兼类词上下文的属性值, 我们作了如下规定: 设 C 是一个兼类词, $T = \{t_1, t_2, t_3, \cdots t_n\}$ 是 C 所有可能词性的集合. $W = \{w(i),\ i \in N+\}$ 是 C 所在句子上下文词语的集合, 这里的 i 代表 $w(i)$ 这个词相对于 C 的位置, 如:

$$w(-k)\ w(-k+1)\ \cdots\ w(-2)\ w(-1)\ C\ w(1)\ w(2)\ \cdots\ w(s-1)\ w(s)\quad (s, k \in N+)$$

假设兼类词 C 有 2个词性, 此时, $T = \{t_1, t_2\}$, 设在整个训练语料中, $w(i)$ 跟 C 的 t_1 词性共现了 n_{i1} 次, 跟 C 的 t_2 词性共现了 n_{i2} 次. 令:

$$N_i = \ln[(n_{i1}+1)/(n_{i2}+1)],$$

这里 N_i 是 $w(i)$ 的数值特征. 其意义表示 $w(i)$ 这个词对 C 的词性为 t_1 时的影响程度, 其中, 加 1是为了起到平滑的作用, 即把那些在训练语料中从没出现的词语默认为出现了一次. 对于多分类问题, 可以转化成相应的两分类问题来解决. 常用的方法有“一对一”, “一对多”, “逐步一对多”等, 本文采用的是“一对多”方式.

2.2 CRF 模型

CRF(Conditional Random Fields)[8] 是由 Lafferty在 2001年提出的一个在给定输入节点 (观察值) 条件下计算输出节点 (标记) 条件概率的无向图模型, 它在观测序列的基础上对目标序列进行建模, 重点解决序列化标注的问题. 对于输入序列 x 和输出序列 y, 可以定义一个线性的 CRF模型, 形式如下:

$$P(y \mid x) = \frac{1}{Z(x)} \exp\left(\sum \lambda_k f_k(y_{i-1}, y_i, x) + \sum u_k g_k(y_i, x) \right),$$

其中, 每个 $f_k(\ )$ 是观察序列 x 中位置为 i 和 $i-1$ 输出节点的特征, 每个 $g_k(\ )$ 是位置为 i 的输入节点和输出节点的特征, λ 和 u 是特征函数的权重, Z 是归一化因子. 作为一个无向图模型表现出比 HMM (隐马模型), MEMM (最大熵隐马模型) 等有向图模型更好的效果. 隐马模型一个最大的缺点就是由于其输出独立性假设, 导致其不能考虑上下文的特征, 限制了特征的选择. 最大熵隐马模型解决了这一问题, 可以任意地选择特征, 但由于其在每一节点都要进行归一化, 所以只能找到局部的最优值, 同时也带来了标记偏置问题 (label bias), 即凡是训练语料中未出现的情况全都忽略掉. 条件随机场则很好地解决了这一问题, 它并不

在每一个节点进行归一化,而是对所有特征进行全局归一化,具有表达元素长距离依赖性和交叠性特征的能力,能方便地在模型中包含领域知识,因此可以求得全局的最优值[9]. 本文采用了 TakuKudo编写的工具包"CRF++ 0.50"(下载地址: http://crfpp.sourceforge.net/).

由于 CRF是一个通用的序列标注工具,需要事先确定特征模板,模板的基本格式是: %x[row, col],用于确定输出数据的一个标记. row确定与当前标记的相对行数; col确定列的绝对列数. 本文所使用模板如表 1:

表 1　实验用的 CRF特征模板

Table 1　CRF template of the experiment

特征	说明	模板表示
W=W(0)	中心词	%x[0,0]
W=W(-n)	中心词上文第 n个词	%x[-n,0]
W=W(n)	中心词下文第 n个词	%x[n,0]
W=W(-n)W(0)	中心词与其上文第 n个词的组合	%x[-n,0]/%x[0,0]
W=W(0)W(n)	中心词与其下文第 n个词的组合	%x[-n,0]/%x[0,0]

2.3　ME模型

ME(Maximum Entropy)[10]是根据 Shannon在 1948年写的《通信的数学原理》提出的一种分类模型. 熵(Entropy)是信息论的基本概念,其定义如下:

如果 X 是一个离散型随机变量,取值空间为 R,其概率分布为 $p(x) = p(X = x), x \in R$. 那么, X 的熵 $H(X)$ 定义为:

$$H(X) = -\sum_{x \in R} p(x) \log_2 p(x),$$

其中,约定 $0^* \log_2 0 = 0$ 该公式定义的熵的单位为二进制位(比特).

熵又称为自信息(self-information),可以视为描述一个随机变量不确定性的量,它表示信源 X每发一个信号所提供的平均信息量. 一个随机变量的熵越大,它的不确定性就越大,那么,正确估计其值的可能性就越小. 越不确定的随机变量越需要大的信息量用以确定其值. 信息熵最大化就是使事物状态的丰富程度达到最大值. 最大熵模型就是在已知部分知识的前提下,关于未知分布最合理推断的一种模型. 也即是保留全部的不确定性,把风险降到最小. 本文所使用的实验平台是张乐博士写的最大熵工具包(下载地址 http://homepages.inf.ed.ac.uk/s0450736/maxent_toolkit.html).

该工具包可以直接输入文本,任意地选择特征. 以"展望/v新/a的/u世纪/n,/w"为例,这里的"新"是一个兼类词,则这句的格式写为:

a 展望　的　世纪　,

其中,"a"是"新"在此句中的词性,它后面的词及标点符号都是"新"的上下文环境.

2.4　集成机制

本实验把上述 3个分类器集成在一起. 因为对于不同的词,每个分类器的分类原理各不相同,可以看作是从不同的角度对该词进了分析. 然后,对它们的分析结果进行统计,利用投票机制,选择票数多的词性作为该兼类词的词性. 如果 3种分类器投的类别各不相同时,我们以 CRF为基准,因为 CRF的正确率高于另外 2个分类器. 各子分类器参数设置如表 2所示:

表 2　子分类器参数设置

Table 2　The setting of sub-classifiers parameters

分类器名称	左窗口	右窗口	其他
SVM	1	1	用的是 C-SVM,其中 C=10,核函数是 RBF. 其余参数默认
ME	2	2	训练迭代次数设置为 15,其余参数默认
CRF	2	2	参数采用默认值

注: 本实验中窗口大小及参数的选择是从实验中得出的.

3　实验结果及分析

实验语料采用的是 1998年 1月份已标注的《人民日报》标准语料库,本文选择测试的兼类词,其每个

词性的词频数大于 30, 共计 410个.

汉语自动标注测试指标主要是精度, 其计算公式如下:

$$\text{精度}(P)=\frac{\text{系统输出中正确结果的个数}}{\text{系统输出全部结果的个数}}\times 100\%.$$

Baseline(基准精度)是指某兼类词全部标为强势标注时的精度, 通常用来衡量标注结果的一个最低标准. 它的定义如下:

$$\text{Baseline}=\frac{\text{强势标注出现的频数}}{\text{该词语总的出现频数}}\times 100\%.$$

为了更好地验证集成的效果, 把该语料平均分成 5份, 进行了 5折交叉验证. 其结果如表 3. 表 4列出了部分样例, 其中, 所选择的词其兼词性多, 词频数大, 且为常用的词, 具有一定的代表性.

表 3 5折交叉验证结果

Table 3 5 folds cross-validation results

分类器名称	实验 1	实验 2	实验 3	实验 4	实验 5	平均
SVM	84.05	84.55	85.57	83.12	86.41	84.74
ME	85.08	86.61	86.29	83.27	86.41	85.53
CRF	89.33	90.10	89.60	87.18	90.09	89.26
Voting	89.61	90.47	90.12	87.55	90.69	89.69

注: 实验 1, 实验 2分别代表第 1次、第 2次交叉实验的平均精度, 其他依次类推.

表 4 词语举例

Table 4 The examples of multiple syntactic category words

词语	所兼词性	SVM	ME	CRF	投票	基准精度
一	\m \d	97.53	97.44	96.96	98.10	96.65
了	\u\v\y	87.08	94.01	98.99	98.64	88.52
为	\v\p\Vg	86.91	87.53	92.28	92.36	52.06
用	\p\v\Ng	77.99	72.73	74.16	78.47	54.02
高	\a\nr\v\an\ad	86.18	89.47	91.12	91.78	83.92
在	\p\v\d	94.13	96.81	95.41	96.98	95.47
来	\v\f\u\m \vn\y	93.60	95.52	84.22	95.31	65.56

注: 1. 此处的词性采用的是文献[11][12]的标注集.

2. ME(最大熵), SVM(支持向量机), CRF(条件随机场)是它们单独的精度. Voting(投票)是集成这 3个分类器做投票得到的精度.

从表 3, 表 4可以看出, 集成方法与 3个方法单独使用相比, 能够普遍提高精度, 且均高于 baseline值, 但相比 CRF提高幅度不大, 这主要是因为, 一方面, 单从一个角度描述一个词的词性都不够准确, 表现在当训练语料达到一定规模、训练过程达到一定程度之后, 标注精度很难再有进一步的提高[13]. 另一方面, ME存在偏置问题, 这使得强势的一方更强, 弱势的一方更弱, 因而降低了精确度; SVM 对线性可分的问题才能分得更好, 但汉语词语出现的频数空间并不是线性可分的, 而 CRF 避免了 ME和 SVM 的问题, 所以 CRF能够取得比较高的精度. 另外, 由于各子分类器均基于词进行特征采集的, 且都是利用了类似的上下文环境, 因而数据稀疏问题会同时影响着各子分类器, 从而也影响了投票的结果[14].

4 展望

本文主要探索了利用 3种分类器集成简单投票的方法进行兼类词的消歧, 取得了一定的效果, 但并没从根本上解决利用统计的方法进行词性标注中的数据稀疏问题, 今后将在平滑数据上做进一步的研究. 通过对语料的分析及对文献[11-12]的研究, 发现其中对某些语言现象的规定不够明确, 这样, 就会造成标准标注语料的混乱. 因此, 在今后的工作中, 采用聚类的方法, 对词性标注做更深入的研究, 探究语言背后更深层次的规律.

(下转第 152页)

[参考文献]

[1] Needleman S B, Wunsch C D. A general method applicable to the search for similarities in the amino acid sequences of two proteins[J]. Journal of Molecular Biology, 1970(48): 443-453.
[2] Smith T F, Waterman M S. Identification of common molecular sequences[J]. Journal of Molecular Biology, 1981(147): 195-197.
[3] Ye Yuzhen, Adam Godzik. Multiple flexible structure alignment using patial order graphs[J]. Bioinformatics, 2005, 21(10): 2362-2369.
[4] Dorigo M. Optinization learning and natural algorithm[D]. Italy: Politecnico di Milano, 1992.
[5] 王小平, 曹立明. 遗传算法-理论应用和软件实现[M]. 西安: 西安交通大学出版社, 2002.
[6] 段海滨. 蚁群算法原理及其应用[M]. 北京: 科学出版社, 2005: 144-148.
[7] Chen Yixin, Pan Yi, Chen Juan, et al. Multiple sequence alignment by ant colony optimization and divide-and-conquer[C]// Brelin: Proc of ICCS, 2006: 646-653.
[8] Jangam S R, Chakraborti N. A novel method for alignment of two nucleic acid sequences using ant colony optimization and genetic algorithms[J]. Applied Soft Computing, 2007, 7(3): 1121-1130.
[9] 梁栋, 霍红卫. 自适应蚁群算法在序列比对中的应用[J]. 计算机仿真, 2005, 22(1): 100-102.
[10] Stefan Schroedl. An improved search algorithm for optimal multiple sequence alignment[J]. Journal of Artificial Intelligence Research, 2005, 23(5): 587-623.

[责任编辑: 顾晓天]

(上接第 147页)

[参考文献]

[1] 刘开瑛. 中文文本自动分词和标注[M]. 北京: 商务印书馆, 2000: 162-166.
[2] 张虎, 郑家恒. 基于分类的汉语语料库词性标注一致性检查[J]. 计算机工程, 2008, 34(8): 90-92.
[3] 周强. 规则和统计相结合的汉语词类标注方法[J]. 中文信息学报, 1995, 9(3): 1-10.
[4] 白栓虎. 汉语词切分及词性自动标注一体化方法[J]. 中文信息, 1996(2): 46-48.
[5] 刘群, 张华平, 俞鸿魁, 等. 基于层叠隐马模型的汉语词法分析[J]. 计算机研究与发展, 2003, 41(8): 1421-1428.
[6] 钱揖丽, 郑家恒. 汉语语料词性标注自动校对方法的研究[J]. 中文信息学报, 2003, 18(2): 33-35.
[7] 邓乃扬, 田英杰. 支持向量机——理论、算法与拓展[M]. 北京: 科学出版社, 2009: 79-111.
[8] Lafferty J, McCallum A, Pereira F. Conditional random fields: probabilistic models for segmenting and labeling sequence data[C]//Proceedings of the 18th ICML. San Francisco: Mogan Koufmann, 2001: 282-289.
[9] 丁德鑫, 曲维光, 徐涛, 等. 基于 CRF模型的组合型歧义消解研究[J]. 南京师范大学学报: 工程技术版, 2008, 8(4): 73-76.
[10] Adwait Ratnaparkhi. A simple introduction to Maximum Entropy Models for natural language process[R]. Philadelphia: University of Pennsylvania, Tech Rep: IRCS-97-08, 1997.
[11] 俞士汶, 段慧明, 朱学锋, 等. 北京大学现代汉语语料库基本加工规范[J]. 中文信息学报, 2002, 16(5): 49-64.
[12] 俞士汶, 段慧明, 朱学锋, 等. 北京大学现代汉语语料库基本加工规范(续)[J]. 中文信息学报, 2002, 16(6): 59-63.
[13] 郭永辉, 吴保民, 王炳锡. 一种用于词性标注的相关投票融合策略[J]. 中文信息学报, 2007, 21(2): 9-13.
[14] 姜维, 关毅, 王晓龙. 基于条件随机域的词性标注模型[J]. 计算机工程与应用, 2006, 21: 13-16.

[责任编辑: 顾晓天]