文本语块识别典型方法的比较与分析

梁颖红[1,2],曹 军[2]
(1.苏州市职业大学 计算机学院,江苏 苏州 215006;
2.东北林业大学 机电工程学院,黑龙江 哈尔滨 150040)

摘 要:文本语块识别在自然语言处理领域具有重要作用。以WINNOW、支持向量机和感知器三种典型的语块识别方法为对象,从模型和特征两方面对每种方法进行了剖析,并比较和分析了三种方法与隐马尔科夫模型的优缺点,指出如果为了避免数据稀疏而只采用"词性"特征来识别多种语块,那些对于"词"敏感的短语准确率将会很低。因此针对不同的语块采用不同的特征和策略,不同短语的识别相互借鉴,把不同语块的识别集成在一起,将会起到很好的效果。

关键词:文本语块识别;支持向量机;感知器;WINNOW;隐马尔科夫模型
中图分类号:TP391.43 **文献标识码**:A **文章编号**:1673-629X(2008)11-0076-04

Comparison and Analysis on Representative Method of Text Chunking

LIANG Ying-hong[1,2],CAO Jun[2]
(1.School of Computer Engineering,Vocational University of Suzhou City,Suzhou 215006,China;
2.School of Mechanical and Electronic Engineering,Northeast Forestry University,Harbin 150040,China)

Abstract:Text chunking acts as critical function in the field of natural processing field. WINNOW,SVM and perceptron are the study object in this paper. For each algorithm, model and feature are anatomized. And the advantages and disadvantages between these three algorithms and hidden Markov model are compared. The proceedings that should be pay more attention in future text chunking are pointed out. All above can be used for reference for relative research people.

Key words:text chunking; SVM; perceptron; WINNOW; hidden Markov model

0 引 言

Abney在1991年从减轻句法分析的难度出发,首次提出浅层分析并给出了语块的定义[1]。1996年采用有限自动机的方法实现了一个浅层句法分析器[2]。即而人们发现浅层分析还可以应用到机器翻译、信息检索等诸多领域,因而浅层分析得到了普遍重视。2000年CONLL[3]把语块识别作为主题,在公共的数据下(训练集是WSJ(15-18),测试集是WSJ20)进行了比赛,极大地推动了语块识别技术的发展。自此,国内外相继有许多学者进行了语块识别研究,多种统计和机器学习方法已经被应用到了英语语块识别中[4~7],可以说英语语块识别研究已经取得了一定的研究成果,而汉语语块识别则刚刚起步[8~11]。

所谓文本语块识别(English Text Chunking)就是识别出文本中非嵌套的短语结构,一般情况下"语块"和"短语"可以互用。在基于统计的语块识别方法中,文本语块识别被看成是一个序列的预估问题。语块的表示形式多采用CoNLL2000会议使用的IOB2的表示形式。在IOB2中,如果用X表示某个短语的话,那么,一个句子中的语块标识可能是下面三种符号中的一种。

B-X:短语X的第一个词;I-X:短语X的非初始词;O:不属于任何短语的词。

在词性后表示短语标识,例1和例2是采用这种标识的结果:

例1:NP[An/DT/B-NP A.P./NNP/I-NP Green/NNP/I-NP official/NN/I-NP] VP[declined/VBD/B-VP to/TO/I-VP comment/VB/I-VP] PP

收稿日期:2008-03-12
基金项目:国家自然科学基金(60575041);哈尔滨市青年科学基金(2005AFQXJ020);2007年黑龙江省博士后基金(520-415029)
作者简介:梁颖红(1970-),女,黑龙江哈尔滨人,副教授,博士,研究方向为自然语言处理和人工智能;曹 军,教授,博士,研究方向为机电一体化、系统建模方法和优化控制理论及木材科学与技术领域的交叉研究。

[on/IN/B-PP] NP[the/DT/B-NP filing/NN/I-NP] O[././O]

例 2:NP[台湾/NR/B-NP] PP[在/P/B-PP] QP[两/CD/B-QP 岸/NN/I-QP] NP[贸易/NN/B-NP] O[中/LC/O] NP[顺差/NN/B-NP] QP[一百四十七亿/CD/B-QP 美元/M/I-QP] O[。/PU/O]

1 典型语块识别方法分析和比较

单纯的规则方法已经不是语块识别的主流方法,而基于统计的方法越来越得到了人们的关注。目前结果较好的三个英语语块识别系统分别是[ZDJ02](基于 WINNOW 的方法)、[KM01](基于 SVM 的方法)和[CM03](应用感知器过滤与排序的方法)。表 1 是这三种方法的结果比较(它们使用相同的训练(WSJ15-18)和测试语料(WSJ20))。

表 1　英语语块识别结果最好的三个系统

方法	精确率(%)	召回率(%)	$F_{\beta=1}$
WINNOW	94.28	94.07	94.17
SVM	93.89	93.92	93.91
感知器	94.19	93.29	93.74

下面在模型和特征两方面对三个方法进行分析,以发现它们的优缺点。

1.1 基于 WINNOW 的语块识别方法

WINNOW 是解决二分问题的错误驱动性质的机器学习方法,Littlestone 在 1988 年指出了该方法能从大量不相关的特征中快速学习[12]。Tongzhang 等把 WINNOW 方法应用到了英语语块识别中[13]。

1.1.1 基于 WINNOW 的英语语块识别模型

基于 WINNOW 的英语语块识别分为训练和测试两个过程:训练阶段主要确定短语特征的权重;测试阶段运用特征和特征的权重来确定最后的语块标识。

(1)训练阶段。

对不同短语特征的形式化描述为 $(x^1,y^1),(x^2,y^2),\cdots,(x^n,y^n)$,($x^i$ 表示矢量,x_i 表示分量,$y\in\{-1,1\}$)。WINNOW 是错误驱动的算法,当使用当前的权重得到的结果与实际的不符时,就要更新权重。每个特征的权重更新公式如下:

$$w_j \leftarrow w_j \exp(\eta x_j^i y^i) \tag{1}$$

上式中,$\eta>0$ 是学习率。这一算法已经被证明是收敛的,在此不再赘述。

(2) 测试阶段。

用 V 代表有效的语块标识序列,$tok_1,tok_2,\cdots,tok_m$ 是需要进行语块识别的词和词性标注序列。$x_1,x_2,\cdots,x_m$ 是相应的特征矢量。$t_1,t_2,\cdots,t_m$ 是语块标识类型,而且 $\{t_1,t_2,\cdots,t_m\}\in V$。寻找语块标识序列可以表示成下式:

$$\{t_1,t_2,\cdots,t_m\}=\arg\max_{\{t_1,t_2,\cdots,t_m\}\in V}\sum_{i=1}^{m}L'(w^{t_i},x_i) \tag{2}$$

$$L'(w^{t_i},x_i)=P(t_i\mid x_i)=\min(1,\max(-1,L(w^{t_i},x_i))) \tag{3}$$

对某个固定的语块类型,定义值 $S(t_{k+1})$ 为:

$$S(t_{k+1})=\max_{\{t_1,t_2,\cdots,t_m\}\in V}\sum_{i=1}^{m}L'(w^{t_i},x_i) \tag{4}$$

易得到下面的递归公式:

$$S(t_{k+1})=L'(w^{t_{k+1}},x_{k+1})+\max_{\{t_k,t_{k+1}\}\in V}S(t_k) \tag{5}$$

观察公式(5) 可以发现,x_{k+1} 依赖于前面的语块标识类型 $\hat{t}_k,\cdots,\hat{t}_{k+1-c}(c=2)$。令,$\hat{t}=\arg\max_{t_k}S(t_k)$,有

$$\hat{t}_{k-i}=\arg\max_{t_{k-i}:\{t_{k-i},\hat{t}_{k-i+1}\}\in V}S(t_{k-i})(i=1,\cdots,c) \tag{6}$$

当把所有的 $S(t_k)(k=0,1,\cdots,m)$ 计算完以后,就可以采用回退的方法得到最后的语块标识:令 $\hat{t}_m=\max S(t_m)$ 是第 m 个词的语块标识结果,其它的语块标识 $\hat{t}_{m-1},\cdots,\hat{t}_1$ 可以由公式(5) 的递归公式和下式得到:

$$\hat{t}_k=\arg\max_{t_k:\{t_k,\hat{t}_{k+1}\}\in V}S(t_k) \tag{7}$$

1.1.2 基于 WINNOW 的英语语块识别使用的特征

WINNOW 算法能从大量的特征中找到相关的特征,因此该算法使用的特征比较多。文献[13]使用了两类特征:一类是基础特征;另一类是增强语言特征。

(1)基础特征。

基础特征由以下四种特征组成:

* 第一层特征:tok_i(指“词”) 和 pos_i(指“词性”)($i=-c,\cdots,c$)。

* 第二层特征:该层特征同时考虑两种特征的组合情况。

$pos_i\times pos_j(i,j=-c,\cdots,c,i<j)$,以及 $pos_i\times tok_j(i=-c,\cdots,c;j=-1,0,1)$。

另外,因为在顺序处理过程中,当前词前面的词的语块标识是已知的,所以还可以包含下面的语块标识特征:

* 第一层语块标识特征:$t_i(i=-c,\cdots,-1)$。

* 第二层语块标识特征:$t_i\times t_j(i,j=-c,\cdots,-1,i<j)$ 和 $t_i\times pos_j(i,j=-c,\cdots,-1,j=-c,\cdots,c)$。

(2) 增强语言特征。

使用了 ESG(English Slot Grammer) 特征,ESG 是一种依存语法,它标注一个词的中心词和中心词的依赖成分。用 f_i 表示第 i 个词的 ESG 特征。增强语言特征

有以下两种：

* 第一层增强特征：$f_i(i=-c,\cdots,c)$。

* 第二层增强特征：$f_i \times f_j(i,j=-c,\cdots,c,i<j)$ 和 $f_i \times pos_j(i,j=-c,\cdots,c)$。

1.2 基于支持向量机(SVM)的语块识别方法

SVM 能使得关键实例和分离超平面的距离最大，通过核函数的构造，SVM 能处理非线性特征空间，甚至能使用由多个特征组成的组合特征。

1.2.1 支持向量机的基本模型。

支持向量机也是适合解决二分问题的机器学习方法。假设训练实例为 $(X_1,y_1),\cdots,(X_l,y_l)$ $(X_i \in R^n, y_i \in \{+1,-1\})$，$X_i$ 是第 i 个实例的 n 维矢量。在基本的支持向量机框架中，要努力通过一个超平面来分割正例和反例。超平面表示如下：

$$(W \cdot X)+b=0(W \in R^n, b \in R) \tag{8}$$

支持向量机通过找到 W 和 b 最优的参数集来把训练数据分为两类。训练一个支持向量机的目标是找到一个具有最大间隔的分隔平面；如果间隔越大，得到的分类器也越好[14]。

支持向量所在的线段为 $(W \cdot X)+b=\pm 1, M=2/\|W\|^2$(间隔用 M 表示)。要是 M 最大，应该使 $\|W\|$ 最小，问题就变成了下面的最优化问题了：

最小化：

$$L(W)=\frac{1}{2}\|W\|^2 \tag{9}$$

归结为：

$$y_i[(W \cdot X_i)+b] \geqslant 1(i=1,\cdots,l) \tag{10}$$

另外，支持向量通过核函数 $K(x_i,x_j)$ 具备了解决非线性分类的能力。核函数有多种，其中 $K(X_i,X_j)=(X_i \cdot X_j+1)^d$ 是多项式核函数，能在第 d 维建立最优的超平面把组合的特征全部考虑进去。

1.2.2 基于支持向量机的英语语块识别使用的特征

Taku Kudo 和 Yuji Matsumoto[5]把 SVM 方法应用到了英语语块识别中。他们使用了三类特征：

(1) 词：w_{i-2} w_{i-1} w_i w_{i+1} w_{i+2}

(2) 词性：t_{i-2} t_{i-1} t_i t_{i+1} t_{i+2}

(3) 语块标识：c_{i-2} c_{i-1}

c_i 为要识别的第 i 个词的语块标识，w_i 为第 i 位置的词，t_i 为 w_i 的词性。

1.3 基于感知器过滤与排序的语块识别方法

感知器是典型的机器学习算法。文献[7]把该方法应用到了英语语块识别中。把英语语块识别分为两个层次，首先根据词特征来判断该词能否是短语的开始或结束，从而得到了候选的短语集；然后对候选短语进行打分，通过排序和过滤选出正确的短语。

1.3.1 基于感知器过滤与排序的语块识别模型

令 x 是句子集合中的一个句子，它由 n 个 $x_i(i=0,\cdots,n-1)$ 组成；K 是预先定义的短语类别集合；一个短语定义为 $(s,e)_k(s \leqslant e, k \in K)$，它由从 x_s 到 x_e 之间连续的词组成；P 是所有可能的短语集合，$P=\{(s,e)_k \mid 0 \leqslant s \leqslant e, k \in K\}$。

短语识别器可以定义为一个函数 phRec：$X \rightarrow Y$，该函数的功能是给定一个句子 x，从 x 中识别短语 y。该函数由两个组成部分，一部分是函数 $\text{phCI}(x) \subseteq P$，它的功能是从输入的候选短语 x 中识别候选短语集；另一部分是计算数值的函数 $\text{score}_k(k \in K)$，该函数的功能是确定某候选短语是某类短语的可能性。短语识别器根据下式从句子 x 中搜索短语：

$$\text{phRec}(x)=\arg \max_{y \subseteq \text{phCI}(x) \mid y \in Y} \sum_{(s,e)_k \in y} \text{score}_k(s,e) \tag{11}$$

1.3.2 基于感知器过滤与排序的语块识别使用特征

该方法使用了如下的特征：

(1) 词 Word(w)

(2) 词性 PoS(w)

(3) 语块标识 ChunkTag(w)

(4) 拼写特点 OrthoFlag(w) 词的大写特点(首字大写还是全大写)、是否包含数字、标点符号特点(包含句号，包含破折号 ……)、是否包含属于某类短语的功能词

(5) 前后缀特点 Affixes(w)

2 结束语

通过对 WINNOW、SVM 和感知器三种方法的模型及使用特征的分析，结合表 1 的识别结果，发现这三者均使用了“词”、“词性”和“语块标识”三类特征，有的方法甚至使用的特征种类更多。隐马尔科夫模型(HMM)和最大熵(ME)等统计方法通常需要仔细地进行特征选择以便获得好的准确率，它们不能从给定的特征集中选择有效的特征，因此常常采用启发式规则进行特征选择。WINNOW、SVM 和感知器方法相对于它们具有如下的优点：

(1) 能自动地进行特征选择，从而允许使用大量的特征，并能从大量的特征中选择有用的特征来使用，这是这三种方法能够取得较好识别效果的主要原因；

(2) 它们都是在线学习算法，具有错误驱动的功能，能及时调整权重或其它部分以取得好的识别结果。

这三种方法中，WINNOW 使用的特征种类最多，识别结果最好。这在一定程度上说明使用较多的特征往往会取得较好的识别结果。但是在识别过程中使用了大量特征，在某些短语的识别中出现了数据稀疏现

象;另外,特征的查询需要消耗很长的时间.从而还会使识别速度下降。在相同的测试集上 WINNOW 方法从训练到测试的时间为 22 分钟[7],SVM 方法的时间复杂度为 $O(n^3)$[5]。

分析三种方法可以发现:目前,语块识别存在如下两大缺点:

1)英语语块识别的策略是把语块识别问题转为类似词性标注的分类问题来解决,这种方法的缺点是无法顾及每个短语内部的组成特点。

2)传统的英语语块识别使用同一个模型和相同种类的特征。这种方法的局限性在于相同种类的特征无法同时适合多种短语类型,同时,数据稀疏现象也随之而来。

如果为了避免数据稀疏而只采用"词性"特征来识别多种语块,那些对于"词"敏感的短语准确率将会很低。因此针对不同的语块采用不同的特征和策略,不同短语的识别相互借鉴,最后把不同语块的识别集成在一起,将会起到很好的效果。

参考文献:

[1] Berwick R, Abney S, Tenny C. Parsing By Chunks: Principle - Based Parsing[M]. Dordrecht: Kluwer Academic Publishers, 1991.

[2] Abney S. Partial parsing via finite - state cascades[C]//Workshop on Robust Parsing. 8th European Summer School in Logic, Language and Information conference. Prague, Czech Republic: [s. n.], 1996: 8 - 15.

[3] Sang T K. Introduction to the CoNLL - 2000 Shared Task: Chunking[C]//Proceedings of CoNLL - 2000 and LLL - 2000 conference. Lisbon, Portugal: [s. n.], 2000: 127 - 132.

[4] Skut W, Brants T. A maximum - entropy partial parser for unrestricted text[C]//In Proceedings of the 6th Workshop on Very Large Corpora Conference. Montreal, Quebec: [s. n.], 1998.

[5] Kudoh T, Matsumoto Y. Use of Support Vector Learning for Chunk Identification[C]//Proceedings of CoNLL - 2000 and LLL - 2000 conference. Lisbon, Portugal: [s. n.], 2000: 127 - 132.

[6] Sang T K. Memory - Based Shallow Parsing[C]//In proceedings of CoNLL - 2000 and LLL - 2000 conference. Lisbon, Portugal: [s. n.], 2000: 559 - 594.

[7] Zhang T, Damerau F, Johnson D. Text Chunking based on a Generalization of Winnow[J]. Machine Learning Research, 2002, 2(2): 615 - 637.

[8] Zhao Jun, Huang ChangNing. The model of Chinese base noun phrase identification based transfer[J]. Journal of Chinese Information Processing, 1999, 13(2): 1 - 7.

[9] Zhang YiQi, Zhou Qiang. The auto identification of Chinese base phrase[J]. Journal of Chinese Information Processing, 2003, 16(3): 1 - 8.

[10] Li Heng, Zhu JingBo, Yao TianShun. The Chinese chunking using SVM[J]. Journal of Chinese Information Processing, 2004, 18(2): 1 - 7.

[11] Li SuJian, Liu Qun. The definition and establish of Chinese phrases[C]//JSCL - 2003 Conference. Harbin: [s. n.], 2003: 100 - 115.

[12] Littlestone N. Learning quickly when irrelevant attributes abound: a new liner - threshold algorithm[J]. Machine learning, 1988(2): 285 - 318.

[13] Zhang Tong, Damerau F, Johnson D. Text Chunking using Regularized Winnow[C]//In: Proceedings of ACL - 2001. Toulouse, France: [s. n.], 2001.

[14] Duda R O, Hart P E, Stork D G. Pattern Classification[M]. Beijing: China Machine Press, 2003.

(上接第 75 页)

3　结束语

文中所讨论的规则集的相似性度量,可以有效地选择模型和算法。提出了一个可以测量不同规则集正相似性与负相似性的算法,它可以灵活地应用在各种实际情况中。实验显示这种相似性度量方法可以帮助选择合适的算法以及组合分类模型中的基模型。

参考文献:

[1] Johansson U, Konig R, Niklasson L. Automatically balancing accuracy and comprehensibility in predictive modeling[C]// in: Proceedings of the 8th International Conferenceon Information Fusion. [s. l.]: [s. n.], 2005.

[2] Neumann J. Classifcation and evaluation of algorithms for rule extraction from artifcial neural networks[D]. summer project, University of Edingburgh, 1998.

[3] Lele S, Golden B, Ozga K, et al. Clustering rules using empirical similarity of support sets[C]// in: Proceedings of the 4th International Conferenceon Discovery Science. London, UK: Springer - Verlag, 2001: 447 - 451.

[4] Huysmans J, Baesens B, Vanthienen J. A new approach for measuring rule sets consistency[D]. Data & Knowledge Engineer, 2007, 63: 167 - 182.

[5] Quinlan J. C4. 5: Programs for Machine Learning[M]. San Francisco, CA, USA: Morgan Kaufman, 1993.