doi:10.3969/j.issn.1673-4785.2009.05.004

# 基于双向标注融合的汉语最长短语识别方法

鉴 萍,宗成庆
(中国科学院自动化研究所 模式识别国家重点实验室,北京 100190)

**摘 要**:汉语最长短语(最长名词短语和介词短语)具有显著的语言学特点.采用基于分类器的确定性标注方法进行双向标注,其结果能够显示最长短语识别在汉语句子正(由左至右)反(由右至左)2个方向上的互补性.基于此,利用确定性的双向标注技术来识别汉语最长短语,并提出了一种基于"分歧点"的概率融合策略以融合该双向标注结果.实验表明,这一融合算法能够有效发掘这2个方向的互补特性,从而获得较好的短语识别效果.

**关键词**:最长名词短语识别;介词短语识别;序列标注;双向标注;分歧点
**中图分类号**:TP391 **文献标识码**:A **文章编号**:1673-4785(2009)05-0406-08

# A new approach to identifying Chinese maximal-length phrases using bidirectional labeling

JIAN Ping, ZONG Cheng-qing
(National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China)

**Abstract**: Chinese maximal-length phrases (maximal-length noun phrases and prepositional phrases) possess remarkable linguistic properties. Bidirectional labeling results of Chinese maximal-length phrases obtained using sequential classifiers reveal complementary properties in both directions. In this paper, both left-right and right-left sequential labeling were employed to identify the Chinese maximal-length noun phrases and prepositional phrases. Then a novel "fork position" based probabilistic algorithm was developed to fuse the bidirectional results. Experiments were carried out on the Penn Chinese Treebank, a segmented, part-of-speech tagged, and fully bracketed corpus. The results confirmed that the proposed algorithm is able to effectively exploit the complementary strengths of the two directions.

**Keywords**: maximal-length noun phrase identification; prepositional phrase identification; sequence labeling; bidirectional labeling; fork position

组块分析(chunking)是自然语言处理一个重要的子任务,它将句子切分为结构相对独立而且互不重叠的组块(短语),以降低完全句法分析的难度.实际上,与基本短语相比,如果能准确地分离出句子中的最长短语成分,将更大程度地降低完全句法分析的歧义.所谓最长短语,是指不被其他任何相同类型短语所包含的短语.它与最短(基本)短语相对,内部可包含多种成分结构,是一个完整的语义单元.除了作为完全句法分析的预处理以外,最长短语识别后得到的清晰的句子结构框架还将为机器翻译、自动文摘等其他自然语言处理任务提供帮助.

最长名词短语(maximal-length noun phrase, MNP)和介词短语(prepositional phrase, PP)是2种重要的、研究较多的短语类型.实际上,介词短语也可以有最长和最短之分,但是由于介词短语的嵌套在汉语句子中比较少见(据统计,宾州中文树库V5.0[1]中共有5.28%的介词短语具有嵌套现象),最长介词短语(maximal-length prepositional phrase, MPP)和一般介词短语(PP)通常不做区分.本文以汉语最长名词短语和介词短语的识别为任务,并在以下章节中用MNP和PP分别表示这2种短语.本文中的PP严格来说是指MPP.

识别MNP和PP的传统方法是估计短语的边界概率分布[2-5].而已有实验结果证明这类方法通常只有加入了规则或语言知识才能取得较好的效

**收稿日期**:2009-08-28.
**基金项目**:国家自然科学基金资助项目(60736014、90820303);"十一五"国家科技支撑计划项目(2006BAH03B02);国家"863"计划资助项目(2006AA010108-4);中国新加坡数字媒体研究院资助项目.
**通信作者**:鉴 萍. E-mail:pjian@nlpr.ia.ac.cn.

果[3,5]. 原因是这类短语具有比基本短语等其他类型的短语更复杂的结构,特别是对汉语来说. 汉语的短语套叠现象比较普遍,一个某种类型的最长短语可以包含其他所有类型的短语成分,甚至可以包含一个从句. 而且它们具有长距离的依存关系,仅依赖边界信息会带来更多歧义. 所以研究者们起初都是试图从最长短语的内部结构或其所处的外部环境寻找规律,判定它的边界. 这需要耗费一定的时间和人力来熟悉该种语言的短语特性.

近些年来,采用序列标注模型和复杂机器学习方法,如 HMMs(hidden Markov models)、SVMs(support vector machines)和 CRFs(conditional random fields)等进行组块分析,特别是基本名词短语的识别,取得了很大的成功[6-8]. 作为相似任务,人们也试图将最长短语的识别看作一个序列标注问题,利用机器学习方法提高识别系统的可移植性. 文献[9]将基于 SVM 分类器的序列标注算法作为其汉语 MNP 识别系统的基线方法,通过加入扩充组块特征和分类标点特征来提高识别精度. 其基线系统和改进系统分别取得了 87.01% 和 89.66% 的 $F_1$ 值性能. 文献[10]使用 CRFs 进行汉语 MNP 的标注,其二阶模型对长度大于 4 个词的复杂 MNP 的识别能力达到了 70.3%. 这些工作证明了序列标注算法对最长短语的识别同样是有效的,但与基本短语相比,识别性能要差很多. 主要原因是基于焦点词周围有限特征信息的序列标注模型很难捕捉最长短语内部的长距离关联. 需要根据语言本身的特性选择合适的标注算法和识别策略.

输出级的系统融合技术已被广泛用于提高基本短语识别系统的性能[7,11-12],多是采用投票的方法从多个系统输出结果中产生出一个最好的结果. 文献[7]和[11]在序列的每一个位置上进行投票. 因为只考虑某一位置上的最好结果,这类方法有可能生成不合法的输出. 文献[12]提出了基于句子和基于短语的投票方法,拥有最多在位置级投票中获胜的标记的句子级/短语级候选将作为最后输出,候选标记序列的合法性可以保证融合结果的合法性. 文献[12]还以实验证明了以上所述 3 类融合策略中,短语级融合达到了基本短语识别的最好效果. 另外,当候选系统数量较少或候选结果得到的票数相等时,可使用加法准则、乘法准则等概率融合策略代替投票,其前提是可以获得各个候选结果的条件后验概率.

本文将选择合适的序列标注算法,并融合正向(由左至右)和反向(由右至左)2 个方向的序列标注结果来识别汉语 MNP 和 PP. 首先,通过对汉语语法现象的初步考察,发现采用基于分类器的确定性标注方法进行双向标注,其结果可以体现汉语句子在 2 个方向上的互补性. 据此,使用短语级融合策略融合该双向结果,但同时证明使用位置级投票结果的短语级融合在基于历史的标注系统中并不能得到很好的效果. 因此,提出了一种基于“分歧点”的概率融合算法,以实验证明这一融合算法能够发掘这 2 个方向的互补特性,并得到较好的识别精度.

## 1　汉语中的特殊语法现象

汉语不是严格具有中心词方向性(head-directed)的语言. 这给从单一方向进行汉语句法分析带来了困难. 但是如果把组块分析作为完全句法分析的前处理,这一问题将会得到缓解. 因为与很多印欧语言(如英语)不同,大多数的汉语短语是具有中心词方向性的[13]. 例如,汉语名词短语以名词为中心词,中心词一般在短语尾部,而英语名词短语的中心词位置要灵活一些. 图 1 给出了 2 个例子,中心词用下划线标出.

一般人 的 看法
the view of average person

政府　和 军队 领导人
the government and army leaders

图 1　名词短语及其中心词举例

Fig. 1　Examples of Chinese NP and their heads

据统计,宾州中文树库 V5.0 的 83 065 个 MNP 中约有 97.2% 是以最后一个词为中心词的. 极少数是以成对标点(如括号)或词“等”为中心词. 另外,汉语名词短语中频繁使用的结构助词“的”也通常位于短语的后半部分,特别是对长短语来说.

可以肯定,如果采用基于历史标注结果的决策模型,句子中的名词和助词“的”必能在该模型由右至左(即反向)对 MNP 进行标注时起到指导作用,减少判定短语左边界时的歧义. 理论上讲,在基于历史的标注模型中,汉语 MNP 的反向标注结果要好于正向标注结果.

但这并不表明 MNP 的正向标注没有可取之处,一些冠词和形容词可以作为名词短语的起始标志. 对于图 2 中的第 1 个例子,如果从右向左进行判别,标注器可能会受动词“违背”的影响,把“约定”判定为短语的左边界(名词短语常在动词后面做动词的宾语). 而从左向右标注则更有可能正确识别左边界“这”,因为语料中限定词(POS(part of speech)标记为 DT)常作为名词短语的起始词. 同理,图 2 第 2

个例子中的形容词“惟一”(POS 标记为 JJ)可以作为正向标注的标志词.

这 种 违背 约定 的 行为
**DT** M VV NN DEC NN

惟一 有 资格 在 欧盟 内部 发行 欧元 的 机构
**JJ** VE NN P NR NN VV NN DEC NN

图 2 汉语名词短语起始位置的限定词和形容词

Fig. 2 Determiner and adjective at the beginning of Chinese NPs

汉语 PP 以介词为中心词并且中心词多位于短语首(在宾州中文树库中这一比例为 98.21%),特殊情况是修饰介词中心词的副词等会出现在介词的前面. 因此,介词是 PP 识别的一个最明显标志,将指引标注器正确判断 PP 的右边界. 这也证明了对汉语 PP 的正向标注效果要好于反向标注.

反向标注汉语 PP 也有可以捕捉的标志词,如表方位的 PP“在…上”和“当…时”中的方位词“上”和“时”. 另一个反向标注具有的优势是它可以避免正向标注对 PP 右边界后面第一个词的过分依赖. 因为语料中介词短语常出现在动词前面,所以正向标注器可能会直到遇见动词才确定短语的右边界,造成标注错误. 反向标注则不会出现这样的问题.

综上所述,基于历史特征的标注模型对汉语 MNP 或 PP 正反 2 个方向的识别能力有一定的差异. 但由于汉语本身的特点,这 2 个优劣不同的结果之间仍具有互补性. 而且在理论上,随着短语长度和内部依存关系距离的增长,这一互补性也将增强. 基本短语因为结构简单,缺乏能使不同方向标注结果产生较大差异的长距离依存歧义,所以其双向标注结果的差异较小,互补性也较弱. 文献[7]的实验结果和文献[14]的预备实验结果显示了这一特点在基本名词短语分析任务上的体现.

## 2 选择合适的序列标注方法

判别式的(如基于 MEMM(maximal entropy Markov model)或 CRFs)序列标注算法和基于分类器(如最大熵模型或 SVMs)的序列标注算法都是自然语言处理任务中常用的序列标注方法[7-8,12,15-18].

基于 CRFs 的序列标注[8]以兼具生成式模型和序列分类器模型的优点著称,可以使用观测序列的任何特征并搜索全局最优标注结果. 在线性链 CRF 模型中,给定观测序列 $\boldsymbol{x}$ 的最大概率标记序列为

$$\hat{\boldsymbol{y}} = \arg\max_{\boldsymbol{y}} p_{\boldsymbol{\lambda}}(\boldsymbol{y} \mid \boldsymbol{x}) = \arg\max_{\boldsymbol{y}} \boldsymbol{\lambda} \sum_i \boldsymbol{f}(\boldsymbol{y}, \boldsymbol{x}, i).$$

式中:条件概率 $p_{\boldsymbol{\lambda}}(\boldsymbol{y}|\boldsymbol{x})$ 为全局特征向量 $\sum_i \boldsymbol{f}(\boldsymbol{y},\boldsymbol{x},i)$ 的 $\boldsymbol{\lambda}$ 加权,$i$ 表示标注位置. CRFs 假设在各个状态结点(标记)之间存在一阶马尔可夫性,每一个标注位置($y_i$)的状态只与前一个位置($y_{i-1}$)有关. 各标注位置之间更紧密的关联需要使用高阶的马尔可夫依赖模型. 例如在二阶 CRF 模型中,以组块分析为例,位置 $i$ 上的输出可表示为 $y_i = t_{i-1}t_i$,$t_{i-1}$ 和 $t_i$ 分别是 $i-1$ 和 $i$ 位置的组块标记. 马尔可夫依赖存在于标记组 $t_{i-2}t_{i-1}$ 和 $t_{i-1}t_i$ 之间,即 $y_{i-1} = t_{i-2}t_{i-1}$.

与基于 CRFs 的判别式模型不同,使用序列分类器的标注算法本质上是一个确定性模型. 它将序列标注看作一串分类问题,使用分类器为每一个位置选择最优标注结果,以单个位置上的局部最优近似全局最优. 给定当前标注状态 $c$,位置 $i$ 的最优预测为

$$\hat{y}_i = \arg\max_{y} p(y \mid c, i).$$

$c$ 通常表示为一组当前标注时刻的上下文特征. 因为算法在标注的每一步都做出决策,标注状态确定性地传递,后续的决策可以使用前面已产生的所有标注结果,即 $c$ 可以包含 $y_{i-1}, y_{i-2}, \cdots, y_0$. 各种分类算法里,SVMs 在序列标注模型中应用最多同时也具有较好的效果[7].

基于 CRFs 的标注模型和基于 SVMs 的标注模型都有很好的特征表达能力. CRF 模型的优点是可以得到全局最优解,但同时也导致其计算因子之间具有不确定性,算法不能很好地捕捉序列中的长距离关联. 使用高阶模型可以起到一定的缓和作用,但相应的计算消耗也会急剧增加. 与此相比,基于 SVMs 的确定性模型因为可以参考已有标注结果,更易于发现序列元素之间的依存关系,贴合识别汉语最长短语所需要的“基于历史特征的标注模型”. 在此类模型中,已有标注历史是以特征的形式应用于当前决策,历史标记元数的增长不会给算法带来过多计算负担. 因此,结合上一节对汉语最长短语特点的分析,选择基于 SVM 分类器的确定性标注模型进行汉语 MNP 和 PP 的识别.

## 3 基于“分歧点”的概率融合

以加法准则为例,常用的分类器融合策略可用下式表示[19](这里仅以等类别先验概率融合为例):

$$\hat{w} = \arg\max_{j} \sum_{k=1}^{K} P(w_j \mid o_k).$$

式中:$o$ 是各分类器中待分类的模式. 若 $K$ 个子分类器给出的后验概率的加和最大,则该类别作为最后输出. 将加法准则应用到基于分类器的双向序列标注问题,位置 $i$ 的最佳输出标记为

$$\hat{y}_i = \arg\max_{y=\hat{y}_{\mathrm{f}},\hat{y}_{\mathrm{b}}} [P(y \mid c_{\mathrm{f}}^{(i)}) + P(y \mid c_{\mathrm{b}}^{(i)})]. \quad (1)$$

式中:$c^{(i)}$表示某一特定方向的标注器在位置 $i$ 的分析状态,包含取自观测序列的静态特征和取自已标注历史的动态特征,它作为序列分类器的输入;下标 f 和 b 分别表示正向(forwards)和反向(backwards);$\hat{y}_f$和 $\hat{y}_b$分别是正向标注器和反向标注器在位置 $i$ 的输出标记类别.

当子分类器的个数为 2 时,上述加法准则可以等价为减法形式. 同时我们引入量 $E$ 来分解原有的最大化问题:

$$E(\hat{y}_f) = P(\hat{y}_f \mid c_f^{(i)}) - P(\hat{y}_b \mid c_f^{(i)}),$$
$$E(\hat{y}_b) = P(\hat{y}_b \mid c_b^{(i)}) - P(\hat{y}_f \mid c_b^{(i)}). \quad (2)$$

式中:$\hat{y}_f$和 $\hat{y}_b$中使 $E(y)$较大者作为当前位置的输出标记. 实际上,$E(y)$在这里表达了标记的类别信任度,即分类器有多大把握选择当前输出类别而不是其他类别(如另一个分类器给出的候选类别).

在位置级序列标注融合策略中,每个位置上的标记是分别计算的:

$$\hat{y}_i = \arg\max_{y=\hat{y}_f,\hat{y}_b} E_i(y),$$

并排列成最后的标记序列. 而基于句子和基于短语的融合试图寻找"一列"最好的标注结果:

$$\hat{\boldsymbol{y}} = \arg\max_{\boldsymbol{y}=\hat{\boldsymbol{y}}_f,\hat{\boldsymbol{y}}_b} E(\boldsymbol{y}).$$

此类方法依然使用每个位置的融合结果:

$$E(\boldsymbol{y}) = \sum_i \Delta_i.$$

式中:

$$\Delta_i = \begin{cases} 1, & y_i = \arg\max_{y=\hat{y}_f,\hat{y}_b} E_i(y); \\ 0, & \text{otherwise.} \end{cases}$$

最终结果由候选标记序列(整个句子或一个短语片段)所含有的在位置级融合中获胜的标记的个数 $\sum_i \Delta_i$决定.

但是,上述区域级的融合策略并不适于基于历史特征的标注模型融合. 原因是基于历史的概率模型中某一位置的决策与已标注历史有关,动态特征的不一致导致分析状态不一致,相同位置上不同标注器输出结果之间是"不公平"竞争. 例如,一个错误的短语识别结果,可能因为后续标记与较早的标记之间具有更小的歧义而获胜.

从另一个角度解释这个问题,把式(2)重写为

$$E(\hat{y}_f) = P(\hat{y}_f \mid c_f'^{(i)}, y_f^{(i-)}) - P(\hat{y}_b \mid c_f'^{(i)}, y_f^{(i-)}),$$
$$E(\hat{y}_b) = P(\hat{y}_b \mid c_b'^{(i)}, y_b^{(i+)}) - P(\hat{y}_f \mid c_b'^{(i)}, y_b^{(i+)}).$$

原标注器分析状态 $c$ 分解为静态部分 $c'$和动态部分 $y^{(i\pm)}$——2 个方向上的已有标注结果("-"和"+"分别表示 $i$ 左边和右边的位置). 如果 $y_f^{(i-)}$与反向标记序列中的 $y_b^{(i-)}$不一致,概率 $P(\hat{y}_b \mid c_f'^{(i)}, y_f^{(i-)})$不能作为正向标注器选择 $\hat{y}_f$而不选择 $\hat{y}_b$的依据. 以一对用于短语标注的标记序列为例:

| | | | | |
|---|---|---|---|---|
| 正向: | O | O | B | I |
| 反向: | O | B | I | I |
| | | $i-1$ | $i$ | $i+1$ |

其中:"B"(begin)表示一个短语的起始位置,"I"(inside)表示短语内部除起始位置以外的位置,"O"(outside)表示短语外部. $i$ 在正向分析中标记为 B 的类别信任度为 $P(\mathrm{B}|\mathrm{O}) - P(\mathrm{I}|\mathrm{O})$(假设只使用前一个标注结果);同样在反向分析中标记为 I 的信任度为 $P(\mathrm{I}|\mathrm{I}) - P(\mathrm{B}|\mathrm{O})$. 显然,概率 $P(\mathrm{I}|\mathrm{O})$不具有比较意义,因为在该标注体系中串"OI"是不合法的,$P(\mathrm{I}|\mathrm{O})$接近于 0.

由以上分析可以看出,在基于历史的双向标注系统中,沿某一方向第一个与另一方向标注结果不同的那个位置,才能真实反映该方向整个标记序列(或一个短语片段)的信任度. 将这个位置称作"分歧点"(fork position),并提出了一种基于"分歧点"的概率融合算法.

图 3 为一个双向短语标注示意图,$\hat{\boldsymbol{y}}_f$和 $\hat{\boldsymbol{y}}_b$分别是正向和反向标记序列. 图中黑色圆表示识别出的短语起始位置(标记为 B 的位置);灰色圆表示识别出的短语内部(标记为 I 的位置);白色圆表示短语外部(标记为 O 的位置). 虚线划出的部分是覆盖某一片段上正反 2 个方向有效短语标记(即 B 和 I)的最小区域,称之为有效区域,并以此作为融合单位. 所以,所要给出的融合算法也是短语级的.

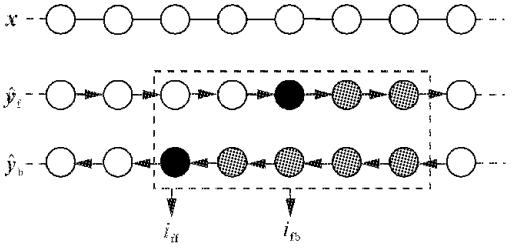


图 3　基于"分歧点"的融合算法示意图

Fig. 3　An illustration of the fork position based algorithm

图中 $i_{ff}$和 $i_{fb}$标出的分别是正向标注的分歧点和反向标注的分歧点. 整个有效区域内某方向标记序列的信任度由该方向分歧点的标记类别信任度来决定:

$$E(\hat{\boldsymbol{y}}_f) = P(\hat{y}_f \mid c_f^{(i_{ff})}) - P(\hat{y}_b \mid c_f^{(i_{ff})}),$$
$$E(\hat{\boldsymbol{y}}_b) = P(\hat{y}_b \mid c_b^{(i_{fb})}) - P(\hat{y}_f \mid c_b^{(i_{fb})}).$$

因为 $i_{ff}$和 $i_{fb}$分别是 2 个候选序列由左至右和由右至左发生分歧的位置,所以有 $y_f^{(i_{ff}-)} = y_b^{(i_{ff}-)}$ 和 $y_b^{(i_{fb}+)} = y_f^{(i_{fb}+)}$,分歧点标记的后验概率具有融合意义. 上述

融合例子相应地转化为比较 $P(O|O)-P(B|O)$(正向,位置 $i-1$)和 $P(I|I)-P(B|I)$(反向,位置 $i$).

因为基于历史特征的标注模型对汉语 MNP 或 PP 正反2个方向的识别能力有一定的差异,使用加权融合(权值 $\omega \geqslant 0$),得到最终的融合算法为

$$\hat{\boldsymbol{y}} = \arg\max_{\boldsymbol{y}=\hat{\boldsymbol{y}}_f,\hat{\boldsymbol{y}}_b} \omega E(\boldsymbol{y}).$$

## 4 实验和结果分析

### 4.1 实验设置

实验在宾州中文树库 V5.0 上进行,使用所有《新华日报》语料. 该语料分布在 698 个原始文件中,共9 493 个句子. 本文从中提取出了 24 436 个 MNP 和 8 282 个 PP(并列 PP 如"从…到…"在宾州中文树库中是以一个最长介词短语出现的,但是考虑到它表示的是2个相互独立的 PP,因此将其看作2个 PP). 由于对只含有一个词的 MNP 的识别没有太大意义,因此本文的实验不包括单个词的 MNP;但是单个词的 PP 多是由于省略了介词中心词,而且数量很少,所以在实验中没有将它们剔除.

实验使用 IOB2 标注体系[20],并添加了2个标记符号:"H"和"S",用来区分短语中心词和非中心词. 这样,共有 5 类标记用于最长短语的识别:"BH"、"BS"、"IH"、"IS"和"O".

实验主要比较以下4个标注和融合系统.

1)单向标注:包括对 MNP 和 PP 的正向和反向标注. 直接使用序列标注器的输出结果,静态特征窗口均设为9,动态特征使用五元历史标注结果,即当前位置之前的4个历史标记. 这一值是根据语料中 MNP 和 PP 的平均长度(分别是5.40个词和5.38个词)选取的.

2)基于短语的融合算法(phrase-based):实现了文献[11]提出的基于短语的融合策略. 因为只有2个候选系统,每个位置上的投票由加法准则代替. 当这2个候选序列所含有的在位置级融合中获胜的标记个数相等时,将退而比较这2个候选序列的各标记后验概率和.

3)基于"分歧点"的概率融合算法——采用"one vs. others"多类分类策略(M1 融合):SVM 分类器使用"one vs. others"模式时,类别打分为该类别到分类面的距离. 利用逻辑回归方法,将这些距离转化为类别的后验概率用于融合.

4)基于"分歧点"的概率融合算法——采用"pair-wise"多类分类策略(M2 融合):分类器使用"pair-wise"模式时,分类的依据是两两类别的分类情况. 本文直接提取2个候选类别在二类分类器中的距离差作为类别信任度用于融合.

YamCha[21]是一个基于 SVM 分类器的开源序列标注工具,因为可以重定义静动态特征并能输出类别打分,所以将其作为短语标注器,并使用 SVMs 的二阶多项式核函数,惩罚参数 $c$ 设置为0.01.

子系统的权值通过在语料库上进行格点搜索(grid search)获得. 为简单起见,将反向标注器的权值 $\omega_b$ 固定为1.00,只遍历正向标注器的权值 $\omega_f$,搜索范围是0.30~2.50,间隔为0.05. 所有测试集使用语料库给出的标准分词和 POS 标注.

### 4.2 主要实验结果

表1和表2分别是上述各系统在9 493句语料上进行十折交叉检验得到的对 MNP 和 PP 识别的平均 $F_1$ 值($F_1$ score). $\omega_b$ 固定为1.00时,$\omega_f$ 对 MNP 和 PP 分别取为0.55和2.00(因为 M2 融合使用的是两类别距离差,权值做了相应的映射). 融合结果一栏括号中表示的是融合后 $F_1$ 值与单向结果中较高值的差.

**表1 MNP 识别的融合结果($\omega_f=0.55$)($F_1$ 值)**

**Table 1 Combining results for MNPs($\omega_f=0.55$)($F_1$ score)%**

| 融合算法 | 正向 | 反向 | 融合 |
|---|---|---|---|
| **Phrase-based** | 83.22 | 85.93 | 86.09 (+0.16) |
| **M1 融合** | 83.22 | 85.93 | 86.94 (+1.01) |
| **M2 融合** | 83.14 | 85.86 | 86.91 (+1.05) |

**表2 PP 识别的融合结果($\omega_f=2.00$)($F_1$ 值)**

**Table 2 Combining results for PPs ($\omega_f=2.00$)($F_1$ score) %**

| 融合算法 | 正向 | 反向 | 融合 |
|---|---|---|---|
| Phrase-based | 84.36 | 74.47 | 84.65 (+0.29) |
| M1 融合 | 84.36 | 74.47 | 85.98 (+1.62) |
| M2 融合 | 83.84 | 74.53 | 85.51 (+1.67) |

可以看出,无论是对 MNP 还是 PP,2个方向的标注结果之间都有明显的差异. MNP 的反向标注性能好于正向标注,PP 的正向标注性能好于反向标注. 相比之下,PP 的双向标注结果差别更大,说明介词对 PP 的识别具有更强的引导作用. M2 融合使用"pair-wise"多类分类策略,不同于基于短语的融合和 M1 融合使用的"one vs. others"策略,所以单向标注结果有细微差别.

比较这3种融合方法,可以发现基于短语的融合对识别性能有一定的提高,但幅度很小(分别为0.16%和0.29%). 本文提出的基于"分歧点"的融合算法可以将 MNP 和 PP 的识别精度分别提高1.05%和1.67%. M1 和 M2 的融合能力基本相似,

M2 融合对精度的提高相对更多一些,原因可以解释为两类分类器的判别结果更直接体现相关类别之间的差距,并可能排除其他类别的干扰. 但因为“pair-wise”单向分类结果稍差,M2 融合结果反不及 M1 融合.

文献[3]利用短语边界分布概率和丰富语言学知识识别汉语 MNP,得到了 83.8% 的识别精度($F_1$ 值). 虽然实验语料不一致,但仍能说明本文所使用的机器学习方法对汉语 MNP 的识别是有效的,而且具有更好的可移植性. 文献[9]是使用 SVM 分类器进行汉语 MNP 识别的相似工作,其改进系统对精度的提高主要在于使用了更细致的标点 POS 标注. 而本文的标注系统本身已使用了词形特征,所以重复该方法在本实验语料上精度没有得到明显的提高.

### 4.3　分　析

通过与其他识别任务和序列标注技术的比较,本节进一步分析汉语 MNP 和 PP 的特性及所述标注和融合策略的适应性. 以下实验使用单一的训练和测试集,原 9 493 句中的前 7 493 句用来训练标注模型,最后的 1 000 句用来测试.

首先将基于 SVMs 的标注方法用于基本名词短语(base NP)的识别,并用基于“分歧点”的方法进行融合. 考虑到基本名词短语的平均长度,除五元历史特征模型外,还给出了使用三元历史的标注结果($n$ 表示历史元数). 同样使用 $F_1$ 值作为评价标准,结果列于表 3.

**表 3　对基本名词短语的识别和融合结果($F_1$ 值)**

**Table 3　Results for base NP identification ($F_1$ score)　%**

| 标注对象 | 历史元 | 正向 | 反向 | 融合 |
|---|---|---|---|---|
| Base NP | $n=3$ | 89.25 | 89.20 | 89.49 |
| | $n=5$ | 89.25 | 89.03 | 89.40 |
| MNP | $n=5$ | 80.94 | 84.62 | 85.99 |

从实验结果看出,无论是使用三元还是五元模型,正反 2 个方向的基本名词短语识别结果之间没有明显差别,这与文献[7]和文献[14]给出的结论一致,而且多元的历史标注特征甚至可能增加识别的歧义. 对其双向结果进行融合后,精度也没有明显提高,这证明基本名词短语识别在正反 2 个方向上的互补性较弱. 最长短语因具有长距离依存关系,其边界的确定更依赖动态的标记特征,所以有显著的融合效果.

实验还将基于 SVMs 的确定性方法与基于 CRFs 的判别式方法进行了比较,二者对 MNP 的单向标注结果列于表 4. “$n=1$”表示算法不使用任何已标注结果. 具有一阶马尔可夫性的 CRFs 考虑前一个位置的标注,将其与使用 2 元标注历史的 SVMs 一起比较,即:CRFs 的阶数 = $n-1$. CRF 标注器采用开源工具 CRF ++[22] 和 Pocket CRF[23](用于高阶 CRFs 的训练与测试). 实验同时比较了 SVMs 和 CRFs 在相同处理器条件下的训练和测试时间.

**表 4　与基于 CRFs 的序列标注结果比较**

**Table 4　Comparisons with CRF-based labeling results**

| 历史元 | 标注算法 | 正向/% | 反向/% | 训练时间/min | 测试时间/s |
|---|---|---|---|---|---|
| $n=1$ | SVMs | 76.86 | 76.86 | 387 | 42 |
| $n=2$ | SVMs | 80.78 | 84.04 | 136 | 19 |
| | CRFs | 78.77 | 78.84 | 10 | <1 |
| $n=3$ | SVMs | 80.94 | 83.93 | 139 | 20 |
| | CRFs | 79.94 | 79.83 | 65 | 4 |
| $n=4$ | SVMs | 80.77 | 84.26 | 131 | 20 |
| | CRFs | 79.70 | 79.53 | 275 | 12 |
| $n=5$ | SVMs | 80.94 | 84.62 | 147 | 21 |
| | CRFs | 80.20 | 80.08 | 1 320 | 51 |

从标注方向角度来看,基于 CRFs 的系统基本不具有 2 个方向上的一致差异性. 基于 SVMs 的系统则除了一元模型外都有明显的方向差异,而且平均识别性能要好于 CRFs. 从历史元角度来看,阶数的增加似乎不能改变 CRF 系统对 MNP 识别的现状,且因复杂度的增加,算法的训练和测试时间也迅速上升. 而使用历史特征的 SVM 系统则具有平稳的消耗. 这些都证明了基于 SVM 分类器的确定性标注模型更适合于汉语最长短语的识别.

最后,考察上述系统双向识别结果的互补能力,以短语识别的召回率(recall)为评价标准. “理想”结果是指双向标注结果的并集中所含正确短语占语料库中短语的比例. SVM 分类器对各种短语均使用 5 元标注历史,CRFs 阶数为 4.

**表 5　双向标注的互补能力(召回率)**

**Table 5　Complementary ability of bidirectional labeling (Recall)　%**

| 标注算法 | 标注对象 | 正向 | 反向 | 融合 | 理想 |
|---|---|---|---|---|---|
| SVMs | MNP | 81.93 | 85.05 | 86.70 | 89.36 |
| | PP | 82.78 | 73.56 | 85.35 | 89.37 |
| | NP | 88.39 | 88.68 | 88.78 | 90.29 |
| CRFs | MNP | 80.20 | 80.08 | – | 80.57 |
| | PP | 82.20 | 81.97 | – | 82.57 |

虽然基于 SVMs 的标注系统对 MNP 的正向标注结果比反向标注差 3 个百分点,但它的加入却能

使正确识别结果的含量接近90%,比反向结果还要高出4个百分点,对PP更是如此.这证实了基于历史的分类器模型能够体现最长短语识别在汉语句子正反2个方向上的互补性,基于"分歧点"的融合算法能部分地发掘这一特性,同时也显示出融合精度依然还有很大的提升空间.

同样是使用基于SVMs的标注系统,NP的理想融合精度与单向精度的差别不大,能够改进融合算法的余地很小.而对于CRFs系统,其最长短语双向标注结果基本不具有互补能力.

## 5 结束语

本论文把广泛用于基本短语识别的基于复杂机器学习方法的序列标注技术用于汉语最长名词短语和介词短语的识别,并从任务的语言学特殊性和序列标注算法的特点出发考察了算法的适应性.通过理论分析和实验,证明了基于分类器的确定性标注方法对最长短语的识别是有效的,并且其双向结果有一定的互补性.在此基础上提出的基于"分歧点"的融合算法恰能发掘它们之间的互补性,并达到较高的融合精度.本文提出的短语识别策略同样适用于其他具有相似特性的短语或语言,因此具有一定的普遍意义.实验表明,对汉语MNP和PP双向标注融合方法的研究还有很大的探索空间,这也指引我们继续寻找更有效的融合策略以进一步提高识别精度.

## 参考文献:


[1]XUE Nanwen, XIA Fei, CHIOU Fudong, et al. The Penn Chinese Treebank: phrase structure annotation of a large corpus[J]. Natural Language Engineering, 2005, 11(2): 207-238.

[2]李文捷,周 明,潘海华,等. 基于语料库的中文最长名词短语的自动抽取[C]//计算语言学进展与应用. 北京:清华大学出版社,1995:119-124.
LI Wenjie, ZHOU Ming, PAN Haihua, et al. Corpus-based maximal-length Chinese noun phrases extraction[C]//Advances and Applications on Computational Linguistics. Beijing: Tsinghua University Press, 1995: 119-124.

[3]周 强,孙茂松,黄昌宁. 汉语最长名词短语的自动识别[J]. 软件学报,2000,11(2):195-201.
ZHOU Qiang, SUN Maosong, HUANG Changning. Automatic identification of Chinese maximal noun phrases[J]. Journal of Software, 2000, 11(2): 195-201.

[4]王立霞,孙宏林. 现代汉语介词短语边界识别研究[J]. 中文信息学报,2005,19(3):80-86.
WANG Lixia, SUN Honglin. Automatic recognition of prepositional phrases in Chinese[J]. Journal of Chinese Information Processing, 2005, 19(3): 80-86.

[5]干俊伟,黄德根. 汉语介词短语的自动识别[J]. 中文信息学报,2005,19(4):17-23.
GAN Junwei, HUANG Degen. Automatic identification of Chinese prepositional phrase[J]. Journal of Chinese Information Processing, 2005, 19(4): 17-23.

[6]ZHOU Guodong, SU Jian, TEY Tongguan. Hybrid text chunking[C]//Proceedings of the 2nd Workshop on Learning Language in Logic and the 4th Conference on Computational Natural Language Learning. Lisbon, Portugal, 2000: 163-165.

[7]KUDO T, MATSUMOTO Y. Chunking with support vector machines[C]//Proceedings of the North American Chapter of the Association for Computational Linguistics. Pittsburgh, USA, 2001: 192-199.

[8]SHA Fei, PEREIRA F. Shallow parsing with conditional random fields[C]//Proceedings of the North American Chapter of the Association for Computational Linguistics. Edmonton, Canada, 2003: 213-220.

[9]BAI Xuemei, LI Jinji, KIM Dongil, et al. Identification of maximal-length noun phrases based on expanded chunks and classified punctuations in Chinese[C]//Proceedings of International Conference on Computer Processing of Oriental Languages. Singapore, 2006: 268-276.

[10]冯 冲,陈肇雄,黄河燕,等. 基于条件随机域的复杂最长名词短语识别[J]. 小型微型计算机系统,2006,27(6):1134-1139.
FENG Chong, CHEN Zhaoxiong, HUANG Heyan, et al. Recognition of complex maximal length noun phrase using conditional random fields[J]. Mini-Micro Systems, 2006, 27(6): 1134-1139.

[11]TJONG KIM SANG E F. Noun phrase recognition by system combination[C]//Proceedings of the North American Chapter of the Association for Computational Linguistics. Seattle, USA, 2000: 50-55.

[12]CHEN Wenliang, ZHANG Yujie, ISAHARA H. An empirical study of Chinese chunking[C]//Proceedings of the Joint Conference of the International Committee on Computational Linguistics and the Association for Computational Linguistics. Sydney, Australia, 2006: 97-104.

[13]LEE Linshan, LIN Longji, CHEN Kehjiann. An efficient natural language processing system specially designed for the Chinese language[J]. Computational Linguistics, 1991, 17(4): 347-374.

[14]WU Yuchieh, YANG Jiechi, LEE Yueshi, et al. Efficient and robust phrase chunking using support vector machines[C]//Proceedings of Asia Information Retrieval Symposium. Singapore, 2006: 350-361.

[15]RATNAPARKHI A. A maximum entropy model for part-of-speech tagging[C]//Proceedings of the Empirical Methods in Natural Language Processing. New Brunswick, USA, 1996: 133-142.

[16]MCCALLUM A, FREITAG D, PEREIRA F. Maximum entropy Markov models for information extraction and segmentation [C]//Proceedings of the International Conference on Machine Learning. Stanford, USA, 2000: 591-598.

[17]TAN Yongmei, YAO Tianshun, CHEN Qing, et al. Applying conditional random fields to Chinese shallow parsing [C]// Proceedings of the Conference on Intelligent Text Processing and Computational Linguistics. Mexico City, Mexico, 2005: 167-176.

[18]宗成庆. 统计自然语言处理[M]. 北京:清华大学出版社,2008:175-177, 179-181.

[19]KITTLER J, HATEF M, DUIN R P W, et al. On combining classifiers[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998, 20(3): 226-239.

[20]TJONG KIM SANG E F, VEENSTRA J. Representing text chunks[C]// Proceedings of European Chapter of the Association for Computational Linguistics. Bergen, Norway, 1999: 173-179.

[21]KUDO T. YamCha: Yet another multipurpose chunk annotator[EB/OL]. (2005-09-05)[2009-02-25]. http://www.chasen.org/~tAKu/software/yamcha/.

[22]KUDO T. CRF ++: Yet another CRF toolkit[EB/OL]. (2007-03-07)[2009-02-25]. http://crfpp.sourceforge.net/.

[23]QIAN X. Pocket CRF[EB/OL]. (2008-08-05)[2009-02-25]. http://sourceforge.net/projects/pocket-crf-1/files/.

**作者简介:**

鉴　萍,女,1982 年生,博士研究生,主要研究方向为自然语言处理、依存句法分析.

宗成庆,男,1963 年生,研究员、博士生导师. 中国科学院自动化研究所模式识别国家重点实验室副主任,国际学术期刊 IEEE Intelligent Systems 副主编,清华大学特邀学术顾问和讲座教授,中国科学院研究生院兼职教授,亚洲自然语言处理联合会(AFNLP)执行理事,中国人工智能学会理事及自然语言处理专业委员会副主任,中国中文信息学会理事及机器翻译专业委员会副主任,担任若干国际学术会议的程序委员会主席、委员等职务. 主要研究方向为自然语言处理理论与方法、机器翻译、人机对话等技术. 作为项目负责人承担国家及国际合作项目 10 余项,申请国家发明专利多项. 发表学术论文 70 余篇,出版学术专著 1 部.