引入混合特征的最大名词短语双向标注融合算法

李业刚 [1, 2] 黄河燕 [2] 鉴 萍 [2]

摘 要 最大名词短语的识别对机器翻译等诸多自然语言处理任务有着重要的意义. 以汉语最大名词短语识别为研究任务, 在分析现有方法的基础上, 从汉语的语言学特殊性以及基于支持向量机的序列标注算法的特点出发, 考查了基于混合特征的融合算法的适应性. 实验证明, 采用词和基本组块混合标注单元的标注方法对汉语最大名词短语的识别是有效的, 并且其正反向识别结果具有一定的互补性, 在此基础上提出的基于 "边界分歧" 的双向序列标注融合算法恰能发掘双向识别的互补性, 并达到较高的融合精度.

关键词 最大名词短语, 双向标注, 基本组块, 混合特征

引用格式 李业刚, 黄河燕, 鉴萍. 引入混合特征的最大名词短语双向标注融合算法. 自动化学报, 2015, **41**(7): 1274−1282

DOI 10.16383/j.aas.2015.c140598

A Combination Algorithm of Bi-directional Labeling in Identifying of Maximal-length Noun Phrases with Hybrid Feature

LI Ye-Gang[1, 2] HUANG He-Yan[2] JIAN Ping[2]

Abstract Maximal-length noun phrase indentification is meaningful to machine translation and many other natural language processing tasks. For the purpose of studying Chinese maximal-length noun phrases, on the basis of current methods, starting from linguistics particularity in Chinese and characteristics of sequence labeling algorithm based on support vector machine (SVM), we explore the adaptability of combination algorithm based on hybrid features. The algorithm is effective, by theoretical analysis and experimental results, to identify Chinese maximal-length noun phrase by applying hybrid unit with words and base chunk, and it is complementary in bi-directional labeling results. From the above, a combination algorithm of bi-directional labeling based on "boundary fork" can discover complement of two directions identification and achieve a high combination accuracy.

Key words Maximal-length noun phrase, bi-directional labeling, base chunk, hybrid feature

Citation Li Ye-Gang, Huang He-Yan, Jian Ping. A combination algorithm of bi-directional labeling in identifying of maximal-length noun phrases with hybrid feature. *Acta Automatica Sinica*, 2015, **41**(7): 1274−1282

最大名词短语具有丰富的句法和语义信息, 其识别结果可用于平行语料的对齐研究, 提高短语对齐的效果; 可用来抽取翻译模板, 提高翻译模板的覆盖率; 可作为句法分析的预处理, 降低句法分析中的浅层结构歧义, 提高句法分析性能. 因此, 最大名词短语的识别和分析有着重要的研究和应用价值.

汉语句子具有双向性, 正 (由左至右) 反 (由右至左) 两个方向的标注都能提供有效信息, 而且标注结果具有互补性. 因此, 融合正反两个方向的序列标注结果, 提高其识别的精度, 已被成功用于短语识别[1]. 文献 [2−3] 的研究是对参与融合的标注序列的每一个标注位置进行投票, 所有胜出的位置标注组合形成最后标注结果, 该方法产生的标注结果往往并不是参与融合的某个序列标注, 而是一个融合后的结果, 因此容易产生不符合标注规则的输出. 文献 [4] 提出基于序列的投票方法, 选取在位置级标注投票中获胜数最多的候选序列作为最后输出, 保证了融合结果的合理性. 文献 [5] 提出基于 "分歧点" 的概率融合算法, 关注当前标注方向上第一个与另一标注方向的标注结果不同的位置, 也就是所谓的 "分歧点", 以该位置的后验概率作为该方向该标记序列的信任度, 选取信任度高的候选序列作为最后输出, 但是该算法把每个 "分歧点" 孤立对待, 不能有效地处理最大名词短语两端都有分歧的情况.

收稿日期 2014-09-12 录用日期 2015-02-16
Manuscript received September 12, 2014; accepted February 16, 2015
国家重点基础研究发展计划 (973 计划) (2013CB329303), 国家自然科学基金 (61132009, 61202244, 61201352) 资助
Supported by National Basic Research Program of China (973 Program) (2013CB329303), and National Natural Science Foundation of China (61132009, 61202244, 61201352)
本文责任编委 赵铁军
Recommended by Associate Editor ZHAO Tie-Jun
1. 山东理工大学计算机科学与技术学院 淄博 255049 2. 北京理工大学计算机学院北京市海量语言信息处理与云计算应用工程技术研究中心 北京 100081
1. Department of Computer Science and Technology, Shandong University of Technology, Zibo 255049 2. Beijing Engineering Applications Research Center of High Volume Language Information Processing and Cloud Computing, School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081

上述方法的共同点是参与融合的标注模型都以词作为标注单位, 只使用了词和词性等词层面的特征, 因此特征集合比较单薄, 系统整体性能有待提高. 基本组块 (Base chunk) 可以看作是对词语的进一步加工和合并. 继 Abney 提出基本组块分析的思想以来, 基本组块的分析研究已经取得了长足的进展[6]. 文献 [7−9] 都是先对句子进行基本组块识别, 然后在最大名词短语识别过程中加入基本组块的特征, 识别性能优于仅仅使用词层面特征的方法. 但是上述方法对前期基本组块识别要求较高, 因此它们或者采用基于规则的方式获取基本组块, 或者在统计方法获取基本组块后, 再进行基于规则的后处理. 规则的加入提高了基本组块的质量, 但系统可移植性也降低了, 不适合开放领域的、大规模的最大名词短语识别. 另外这些方法虽然采用了基本组块特征, 但仍然采用词作为基本的标注单元, 基本组块层面的特征仅起到补充作用, 性能提升有限. 文献 [10] 尝试以基本组块作为标注单元进行汉语功能组块标注, 当使用自动标注的基本组块作为标注单元时, 受累于基本组块的标注错误, 功能组块标注模型的性能不但低于以词作为标注单元的功能组块标注模型, 甚至低于基线模型.

综合上述, 针对以基本组块作为标注单位的不足, 以及探讨基本组块特征融入后双向标注融合是否依然能够提高系统的性能. 我们在现有基本组块研究成果的基础上, 融合基于支持向量机[11] (Support vector machine, SVM) 的汉语最大名词短语前后向序列标注结果, 把汉语最大名词短语的左右边界看作有关联的整体, 以左右边界的后验概率累加和作为该最大名词短语标记序列的信任度. 混合词和基本组块作为标注单位, 对于高信任度的基本组块, 采用基本组块层面的特征, 对于低信任度的基本组块, 采用词层面的特征. 特征选择时, 以基本组块层面的特征为主, 词和词性等词层面的特征为辅, 从而实现了一个混合特征的最大名词短语双向标注融合算法. 在测试集的实验中, 融合后最大名词短语识别的 F-1 值达到 88.24 %, 比基线模型高出了 2.34 %.

1 支持向量机

诸多研究者对 SVM 进行研究[12−14]. SVM 的核心思想是寻找一个满足分类要求的最优超平面, 在保证分类精度的同时, 最大化超平面两侧的空白区域 (Margin), 使正负例间的隔离边缘最大化.

假定训练数据为 $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$, $x \in \mathbf{R}^n$, $y_i \in \{+1, -1\}$, 超平面 $w \cdot x + b = 0$ 能将两个类别分开, 最优分类面问题表示为约束优化问题:

$$\begin{aligned} \min \quad & \Phi(w) = \frac{1}{2}\|w\|^2 = \frac{1}{2}(w \cdot w) \\ \text{s.t.} \quad & y_i[(w \cdot x_i) + b] - 1 \geq 0 \end{aligned} \tag{1}$$

经过两类样本中离分类面最近的点, 且平行于最优分类面的超平面的训练样本就是支持向量. 定义 Lagrange 函数:

$$L(w, b, a) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} a_i(y_i[(w \cdot x_i) + b] - 1) \tag{2}$$

优化 Lagrange 方法得到最优分类函数:

$$g(x) = \text{sgn}[\sum_{i=1}^{n} a_i y_i (x_i \cdot x) + b] \tag{3}$$

式中, a_i 表示样本空间的拉格朗日乘子.

对于线性不可分问题, 即非线性问题, 引入内积核函数, 把数据空间映射到一个高维特征空间, 通过非线性变换把非线性问题转换为高维空间中的线性问题, 在这个高维特征空间寻求最优分类面. 分类函数变换为:

$$g(x) = \text{sgn}[\sum_{i=1}^{n} a_i y_i K(x_i, x) + b] \tag{4}$$

其中, $K(x_i, x)$ 为核函数, 利用核函数实现原低维空间到高维空间的映射, 把原空间的非线性分类转换为高维空间的线性分类. 在最优分类面中采用适当的内积函数就可以实现某一非线性变换后的线性分类.

使用分类器的序列标注算法本质上是一个确定性模型. 它将序列标注看作一系列的分类问题, 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:

$$\hat{y}_i = \text{argmax}\ p(y|c_i) \tag{5}$$

c 通常表示为当前标注位置的一组上下文特征, 因为算法在标注的每一步都要做出决策, 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历史标注结果, 也就是说, 特征 c 可以包含当前标注位置的历史标注结果 $y_{i-1}, y_{i-2}, \cdots, y_0$.

2 基于 “边界分歧” 的序列标注融合算法

在多分类器融合中, 当候选序列比较少的时候, 可以采用加法准则或者乘法准则等简洁的概率融合策略. 在最大名词短语双向序列标注结果融合任务

中, 参与融合的候选序列只有两个: 前向标注结果和后向标注结果. 因此, 可以使用加法准则进行概率融合.

基于加法准则的分类器概率融合的基本思想是使参与融合的分类器给出的后验概率的加和最大. 常用的分类器融合策略可用式 (6) 表示. 其中, K 表示参与融合的分类器的个数; w 表示类别; o 表示待分类模式.

$$\hat{w} = \arg\max_{j} \sum_{k=1}^{K} P(w_j|o_k) \tag{6}$$

针对双向序列标注融合任务, 位置 i 的最优标记可以用式 (7) 表示.

$$\hat{y}_i = \arg \max_{y_i=\hat{y}_{id},\hat{y}_{i\bar{d}}} (P(y_i|c_{id}) + P(y_i|c_{i\bar{d}})) \tag{7}$$

在双向序列标注融合任务中, 参与融合的分类器只有 2 个. 因此, 可以把式 (7) 转换为减法的形式, 则正反两个方向的标注结果的信任度可以分解为式 (8) 的形式. 其中, $\hat{y}_{id}$ 和 $\hat{y}_{i\bar{d}}$ 分别表示位置 i 的正向和反向标注; c_i 表示特定方向的序列标注器在位置 i 的输入, 一般包含观测序列的静态特征及其标注历史的动态特征.

$$\begin{cases} E(\hat{y}_{id}) = P(\hat{y}_{id}|c_{id}) - P(\hat{y}_{i\bar{d}}|c_{id}) \\ E(\hat{y}_{i\bar{d}}) = P(\hat{y}_{i\bar{d}}|c_{i\bar{d}}) - P(\hat{y}_{id}|c_{i\bar{d}}) \end{cases} \tag{8}$$

其中, $E(\hat{y}_{id})$, $E(\hat{y}_{i\bar{d}})$ 表示某一方向分类器有多大把握选择当前输出类别而不是另一方向的候选类别. 两者中的较大者就是位置 i 的获胜标记.

$$\hat{y}_i = \arg \max_{y_i=\hat{y}_{id},\hat{y}_{i\bar{d}}} E(y) \tag{9}$$

基于位置的序列标注融合策略中, 最优序列标注结果是融合中单个位置标注获胜最多的候选序列.

$$E(y) = \sum_{i=1}^{n} \phi_i \tag{10}$$

$$\phi_i = \begin{cases} 1, & \text{若 } \hat{y}_i = \arg\max E(y) \\ 0, & \text{其他} \end{cases}$$

但是, 在基于历史的概率模型中, 一般包含观测序列的标注历史的动态特征, 某一位置的决策往往与其前一步甚至前几步的标注历史密切相关, 也就意味着, 某个位置错误的标注结果, 可能因为其标记与历史标记之间具有较小的歧义而获得更高的信任度, 从而在位置级标注竞争中胜出, 造成在同一个位置上不同标注器输出结果之间无法公平竞争. 而基于 "分歧点" 的概率融合算法把每个 "分歧点" 孤立对待, 因此, 可能会出现同一个短语中某个标注方向的开始标记获胜, 而另外一个标注方向的结束标记获胜, 违背了在两个不同方向的标注结果中, 选择一个 "更有可能的" 标注结果的初衷, 得到错误的不符合标注规则的输出. 因此, 针对上述问题, 我们提出了基于 "边界分歧" 的双向序列标注融合算法, 把最大名词短语的左右边界看作有关联的整体, 以左右边界的后验概率累加和作为该最大名词短语标记序列的信任度.

图 1 为一个双向标注融合区域示意图, $\hat{y}_d$ 和 $\hat{y}_{\bar{d}}$ 分别是正向和反向标记序列, b 表示正向标记序列中的最大名词短语左边界, e 表示右边界; $\bar{b}$ 表示反向标记序列中的最大名词短语左边界, $\bar{e}$ 表示右边界; 灰色圆圈表示最大名词短语内部; 白色圆圈表示最大名词短语外部. 当两个方向的最大名词短语标注区域有重叠但不完全相等时, 能包含所有两个标注区域的最小片断, 称为有效标注融合区域, 如图 1 中虚线矩形内的部分.

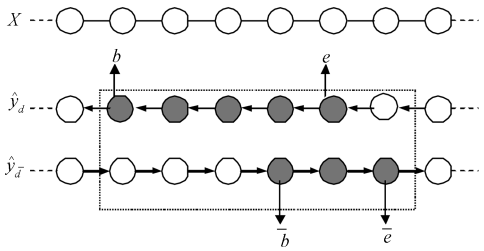

图 1 基于 "边界分歧" 的融合区域示意图

Fig. 1 Diagram of combination area based on "boundary fork"

整个有效区域内某方向标记序列的信任度由该方向的左右边界的标记类别信任度来决定. 因此, 把式 (8) 改写为式 (11) 的形式, 其中, 分析状态 c 分解为静态部分 c' 和 $y^{\pm}$ 动态部分, 其中 "+" 和 "−" 分别表示 i 左边和右边的位置.

$$\begin{aligned} E(\hat{y}_{id}) = {} & P(\hat{y}_{idb}|c'_{idb}, \hat{y}^{-}_{\bar{i}\bar{d}b}) - P(\hat{y}_{\bar{i}\bar{d}b}|c'_{idb}, \hat{y}^{-}_{\bar{i}\bar{d}b}) + \\ & P(\hat{y}_{idb}|c'_{ide}, \hat{y}^{-}_{\bar{i}\bar{d}e}) - P(\hat{y}_{\bar{i}de}|c'_{ide}, \hat{y}^{-}_{\bar{i}\bar{d}e}) \\ E(\hat{y}_{i\bar{d}}) = {} & P(\hat{y}_{i\bar{d}b}|c'_{i\bar{d}b}, \hat{y}^{+}_{i\bar{d}b}) - P(\hat{y}_{i\bar{d}b}|c'_{i\bar{d}b}, \hat{y}^{+}_{i\bar{d}b}) + \\ & P(\hat{y}_{i\bar{d}e}|c'_{i\bar{d}e}, \hat{y}^{+}_{i\bar{d}e}) - P(\hat{y}_{i\bar{d}e}|c'_{i\bar{d}e}, \hat{y}^{+}_{i\bar{d}e}) \end{aligned} \tag{11}$$

基于历史的标注模型的最大名词短语识别在正反两个方向的性能可能会有一定的差异, 当某一个方向的单向标注性能明显好于另外一个方向时, 我们更有可能倾向于选择这一方向的结果作为最终结果, 即选择信任度高的那个方向. 简单起见, 添加一个补偿系数 ξ, 以 d 方向为基准, 即将 d 方向补偿系数固定为 1.00, 通过调节补偿系数 ξ, 修正 $\bar{d}$ 方向信

任度.

$$\hat{y}_i = \arg \max_{y_i=\hat{y}_d,\hat{y}_{\bar{d}}} \xi E(y_i) \tag{12}$$

整个融合过程可以用一个贪心算法来描述, 如算法 1 所示.

算法 1. 基于 "边界分歧" 的双向序列标注融合算法

输入. 汉语句子 X.

输出. 所有带召回标记的标注结果.

步骤 1. 使用序列标注工具, 分别对 X 进行正向和反向最大名词短语标注, 得到正反两个方向的初始标注序列 Y_D 和 $Y_{\bar{D}}$, 并保留标注的后验概率;

步骤 2. 初始化句子的最大名词短语边界信任度候选队列 S, 信任度的计算如式 (11) 所示;

步骤 3. 从队列 S 中选出信任度最大一个, 将其所在的元组移出队列 S, 如果其两个方向标注结果跟所有已召回标记无冲突, 则将它的状态设为召回;

步骤 4. 如果队列 S 为空, 则停止, 否则转步骤 3.

3 特征选择

在统计概率模型中, 模型需要从训练样本抽取而来的特征中学习判别所需的知识. 特征的描述能力以及特征的规模都直接影响统计概率模型的性能. 特征具有任务相关性, 对于不同的自然语言处理任务, 其选取的特征也各不相同, 任务驱动、有针对性的特征定义和特征模板可以大幅度提高统计概率模型的性能, 反之, 低效甚至无效的特征则会降低统计概率模型的性能.

我们主要针对汉语最大名词短语识别任务, 对于词层面的特征、基本组块层面的特征以及标点符号分类特征进行考察, 通过实验来检验上述特征在汉语最大名词短语序列标注任务中的适用性.

3.1 词层面的特征

词可以起到判别和实例化的作用, 在英语的短语和基本组块识别任务中, 仅用词作为特征, 识别模型就可以达到比较好的性能. 但是, 汉语不像英语那样具有丰富的词缀和形态学信息, 因此带来的数据稀疏等问题导致仅采用词层面特征的汉语短语和基本组块分析的性能很难与同类别的英语任务比肩.

词性可以看作是对词的一种分类, 它可以指示词在句子中的语法功能以及词与词间的聚合关系. 词性特征起到类别信息和数据平滑的作用. 词性特征对短语和基本组块分析有着较强的预测能力. 同时, 词性仅有 40 余种, 分类粒度比较粗, 区分度要弱于词. 因此, 采用词和词性组合特征的模型往往可以取得更好的性能. 截止到目前, 针对汉语最大名词短语识别的公开的研究成果基本都是以词作为标注的基本单位, 主要选取词、词性等词层面的特征.

以词作为标注的基本单元的最大名词短语序列标注示例如下:

他们/PN/O 还/AD/O 为/P/O 国家/NN/B 重点/NN/I 工程/NN/I 兰化/NR/I 化肥厂/NN/I 改扩建/NN/I 工程/NN/I 提供/VV/O 了/AS/O 1.9 亿/CD/B 美元/M/I 的/DEG/I 安装/NN/I 工程/NN/I 保险/NN/I 保障/NN/I ./PU2/O

3.2 基本组块层面的特征

基本组块是比词的粒度更大标注单位, 拥有比词更丰富的特征. 基于基本组块的标注模型仅使用基本组块层面的特征, 比如基本组块的中心词、中心词词性、句法标记、关系标记、成分标记、基本组块的左边界词和右边界词以及边界词的词性等.

文献 [10] 尝试以基本组块作为标注单元进行汉语功能组块 (主语块、谓语块、宾语块和状语块) 标注, 当使用人工标注的基本组块的相关特征时, 可以使得功能组块标注的性能大幅度提升, 总体 F-1 值达到 89.19 %, 相比于以词作为标注单元, 功能组块标注模型的总体 F-1 值提高了 4.27 %. 但是, 当使用自动标注的基本组块的相关特征时, 由于自动识别的基本组块远远没有达到人工标注的水准, 基本组块的边界识别、类型标记和关系标记都存在一定的错误, 这些标注错误在以基本组块作为标注单元的其他更大粒度的组块识别分析过程中进一步传递、放大, 最终功能组块标注模型的性能差强人意, 不但低于以词作为标注单元的功能组块标注模型, 甚至低于基线模型, 总体 F-1 值仅有 83.01 %.

以基本组块作为标注单位的模型对句子中每个基本组块给予一个标记. 以基本组块作为标注的基本单元的最大名词短语序列标注示例如下:

[他 们]/NC/O [还]/AC/O 为/P/O [国 家 重 点 工 程]/NC/B [兰 化 化 肥 厂]/NC/I [改 扩 建 工 程]/NC/I [提 供 了]/VC [1.9 亿]/MC/B [美 元]/NC/I 的/DEG/I [安 装 工 程 保 险 保 障]/NP/I./PU2/O

3.3 词和基本组块混合特征

既要发挥基本组块作为基本标注单元的优势, 把基本组块层面的特征融入到最大名词短语识别的特征集合, 又要尽量减少自动标注的基本组块带来的引入错误. 我们的解决方案是: 采用一种词和基本组块混合的标注单元, 把基本组块分为高信任度基本组块和低信任度基本组块, 对于高信任度的基本组块, 引入其基本组块层面的相关特征, 对于低信任度的基本组块, 放弃基本组块特征, 只采用词层面的特征, 以此减少自动标注的基本组块错误标注信息

的引入, 提高最大名词短语的识别性能.

采用基本组块和词作为标注的基本单元的最大名词短语序列标注示例如下:

[他 们]/NC/O [还]/AC/O 为/P/O 国家/NN/B 重点/NN/I 工程/NN/I 兰化/NR/I 化肥厂/NN/I 改扩建/NN/I [提供了]/VC/O [1.9亿]/MC/BS [美元]/NC/I 的/DEG/I [安装工程保险保障]/NP/I ./PU2/O

组合分类器[15] 方法得到了广泛的研究 这类方法基于一个假设: 参与组合的单个分类器产生的错误是不相关的或者基本不相关的. 借鉴组合分类器的思想, 我们提出了一种基于多个标注模型的基本组块信任度区分方法, 在同一个训练集上构造基于不同方法的基本组块标注模型; 然后, 分别对汉语最大名词短语标注时用到的训练集、开发集以及测试集进行基本组块识别; 最后, 把不同的基本组块标注模型标注的结果进行比对, 其中, 各个标注模型标注一致的基本组块作为高信任度基本组块, 其他基本组块作为低信任度基本组块.

基本组块是以名词等实词为中心, 相邻的具有特定语义的词序列. 在给定的训练集上, 遵循相邻、不嵌套、内部不包含其他类别的基本组块定义原则, 我们标注了七种类型的基本组块, 分别为 NC (名词组块)、VC (动词组块)、AC (形容词组块)、DC (副词组块)、LC (方位组块)、MC (数量组块)、TC (时间组块). 训练了三个不同的基本组块标注模型, 分别是基于 SVM 的序列标注模型 (简称 M1), 基于 CRF 的序列标注模型 (简称 M2) 和基于 MEM 的序列标注模型 (简称 M3), 它们在测试集上标注结果如表 1 所示, 其中 Pattern 表示参与对比的标注模型; Precision 表示标注一致正确率; Recall 表示标注一致召回率.

表 1 三种标注方法的一致性分布

Table 1 Consistency distribution of the three methods

Pattern	Precision (%)	Recall (%)
M1	84.25	86.32
M2	84.14	85.91
M3	80.63	81.61
M1&M2	94.79	76.59
M1&M3	91.53	72.45
M2&M3	91.85	71.63
M1&M2&M3	97.63	67.14

由表 1 可以看出, 在该测试集上, 当三个基本组块标注模型的输出一致时, 基本组块标注的正确率可以达到 97.63 %; 当其中的两种标注模型的输出一致时, 基本组块标注的准确率也分别达到 94.79 %、91.53 % 和 91.85 %, 相比单个标注模型, 准确率高出接近 10 个百分点, 三个标注模型一致标注错误的比率仅仅为 2.37 %. 因此, 把标注一致的基本组块作为高信任度基本组块, 标注不一致的基本组块作为低信任度基本组块的方法具有可行性.

4 实验结果及分析

4.1 实验设置

实验使用了宾州树库 5.0 中的《新华日报》语料, 宾州树库语料不是句子级对齐的, 为了便于后续的应用, 我们将其中 325 个中英文原始文件进行了句子对齐处理, 得到了 3 849 个对齐的句子, 其中后 1 000 句对 (2 850 ∼ 3 849) 一分为二, 作为开发集和测试集, 其余的部分作为训练集. 由于只含有一个词的 MNP 的识别没有太大研究价值, 实验剔除了单个词的 MNP. 所有开发集和测试集均使用语料库给出的标准分词和词性标注.

实验使用的 YamCha 是基于 TinySVM 分类器并且封装好的开源序列标注工具, 提供正反两个方向的标注选项, 可以重定义静动态特征, 而且能输出类别打分. 实验中 SVM 使用二阶多项式核函数, 惩罚参数 c 设置为 0.01. CRF 标注器采用开源工具 CRF++ 和 PocketCRF (用于高阶 CRF 的训练与测试).

4.2 特征选择实验

为了验证所述的采用基本组块和词的混合标注单元的最大名词短语标注模型的性能, 分别建立了以词作为标注单元的序列标注模型和以自动识别的基本组块作为标注单元的序列标注模型, 与之进行对比. 实验中直接采用了序列标注器的输出结果.

1) Word-unit: 以词作为标注单元的模型. 采用词和词性等词层面的基本特征, 静态特征窗口设为 9, 动态特征使用 5 元历史标注结果, 即考虑当前位置之前的 4 个历史标记.

2) Chunk-unit: 以自动识别的基本组块作为基本标注单元. 基本组块的平均词数大约 2 个词左右, 参考以词作为标注单元的模型的静态特征窗口设置, 设置静态窗口为 5, 采用基本组块层面的相关特征, 动态特征使用 4 元历史标注结果.

表 2 给出了参与对比的模型的汉语 MNP 识别 F-1 值. 其中, backward 表示后向标注结果; forward 表示前向标注结果; M1、M2 和 M3 表示混合标注单元中分别采用基于 SVM 的序列标注模型、基于 CRF 的标注模型和基于 MEM 的标注模型自动识别基本组块.

从实验结果来看, 以自动识别的基本组块作为标注单元的 chunk-unit 模型的性能均低于以词作

为标注单元的 word-unit 模型, chunk-unit 模型最好的正向识别结果比 word-unit 模型低了 3.1 %, chunk-unit 模型最好的反向识别结果比 word-unit 模型低了 3.3 %. 比较三个不同的 chunk-unit 模型可以看出, 采用的自动识别的基本组块的准确率越高, 最大名词短语识别的性能也越高. 也就是说, 在 chunk-unit 模型中, 最大名词短语的识别性能和基本组块的识别准确率是密切相关的.

表 2　参与对比模型的性能 (F-1 值)

Table 2　Performance of the model participation in contrast (F-1 value)

Model	Word-unit	Chunk-unit		
		M1	M2	M3
Backward	**84.71**	**81.41**	81.24	80.67
Forward	**82.31**	79.05	**79.21**	78.31

在采用基本组块和词的混合标注单元的模型中, 把多个标注模型输出一致的基本组块看作高信任度的基本组块, 采用基本组块作为基本标注单元, 而对于多个标注模型标注不一致的基本组块, 采用词作为基本标注单元. 表 3 给出了不同的模型组合方式下的最大名词短语识别的 F-1 值, 动态特征使用 4 元历史标注, M0 表示以人工标注的基本组块作为标注单元.

表 3　不同模型组合的最大名词短语识别结果 (F-1 值)

Table 3　MNP identification result of combination of different models (F-1 value)

Hybrid-unit	Backward	Forward
M0	89.38	87.21
M1&M2	86.02	83.88
M1&M3	85.81	83.67
M2&M3	85.77	83.65
M1&M2&M3	87.23	85.07

从实验结果可以看出, 多个标注模型的一致识别结果作为标注单元的最大名词短语识别模型的性能有比较明显的提高. 其中, 当使用三个基本组块标注模型的一致识别结果时, 性能达到最优. 相比性能最好的以单个模型的输出作为标注单元的识别结果, 正向识别 F-1 值提高了 5.86 %, 反向 F-1 值提高了 5.82 %; 即便是相比以词作为标注单元的标注模型, 正向识别 F-1 值也有 2.76 % 的提升, 反向 F-1 值提高了 2.52 %. 同时, 也看到, 相比以人工标注的基本组块作为标注单元的最大名词短语标注模型, 不论是正向标注还是反向标注, 都还有比较大的差距, F-1 值相差接近 2 个百分点. 究其原因, 混合单元标注模型的性能介于以词作为标注单元和以人工标注的基本组块作为标注单元的标注模型之间, 多个标注模型一致的方法固然提高了基本组块标注的准确率, 但基本组块标注的召回率却下降了 (见表 1), 这就造成词作为标注单元占了更大的比例, 混合标注单元模型向以词作为标注单元的模型倾斜. 也就意味着, 可以通过进一步提高基本组块识别的准确率和召回率, 使得混合标注单元的标注模型的性能接近以人工标注的基本组块作为标注单元的标注模型, 正反向识别的 F-1 值均有大约 2 个百分点的提升空间.

4.3　双向融合实验

我们研究的重点在采用基本组块和词的混合标注单元的标注模型中, 使用基于 “边界分歧” 序列标注融合算法, 融合前后向的标注结果, 是不是能进一步提高模型的性能. 因此, 构造了三个融合模型.

1) Strategy-1: 在训练集上重复了文献 [5] 的实验, 以词作为基本标注单元, 采用词层面基本特征 (词和词性), 融合策略采用基于 “分歧点” 的概率融合算法.

2) Strategy-2: 以词作为基本标注单元, 采用词层面基本特征 (词和词性), 融合策略采用基于 “边界分歧” 概率融合算法.

3) Strategy-3: 以词和基本组块作为基本标注单元, 采用词和基本组块混合特征, 融合策略采用基于 “边界分歧” 概率融合算法.

由于历史标注并不是百分之百的正确, 随着动态特征所采用的历史标注元数的增多, 标注错误传递累积会变得严重, 导致精度下降, 我们考察采用 2 到 6 元历史标注时融合模型的性能. 以词作为标注单元的融合结果如表 4 所示. 其中, 反向补偿系数 ξ 固定为 1.00, 即无反向补偿; N 表示使用历史标注元数.

表 4　以词作为标注单元的融合结果

Table 4　Combination result of words unit

N	Backward	Forward	Strategy-1	Strategy-2
2	83.75	82.40	83.94	84.07
3	84.04	82.46	84.91	84.79
4	84.39	82.35	**85.79**	**85.92**
5	**84.71**	82.26	84.21	85.33
6	84.12	82.33	84.52	84.79

通过对实验数据的对比可以看出, 在以词作为标注单元的 Strategy-1 融合结果比最好的单向输出结果提高了 1.08 %, 本文提出的基于 “边界分歧” 的融合算法提高幅度略大一点, 达到 1.21 %, 比基于 “分歧点” 的融合算法的识别精度提高 0.13 %.

考察在使用 M1、M2 和 M3 三个模型输出一致

的基本组块和词的混合标注单元的最大名词短语识别模型中, 使用基于 “边界分歧” 的概率融合算法融合前后向的输出是否能够提升的最大名词短语识别性能. 实验结果如表 5 所示.

表 5 混合单元的最大名词短语识别的结果

Table 5 Combination result of hybrid unit

N	Backward	Forward	Strategy-3
2	85.67	84.85	86.12
3	86.34	85.08	87.09
4	87.21	84.91	88.12
5	87.23	85.07	87.83
6	86.69	84.59	87.35

从实验结果可以看出, 反向标注精度在动态特征采用 5 元历史标注时达到最高, 而正向标注精度在采用 3 元历史标注时达到最高. 从融合结果来看到, 虽然采用 4 元历史标注的正向和反向识别精度都不是最高的, 但采用 4 元历史标注时的融合输出却是最高的, 这说明融合精度不一定随着某个单向精度的提高而提高. 在采用 4 元历史标注时, 虽然正向标注的识别 F-1 值 (84.91 %) 比反向的 (87.21 %) 低了将近 3 个百分点, 但是融合后的 F-1 值达到 88.12 %, 比其单向标注中较好的反向标注的输出也高出了 0.91 %, 比正向标注更是高出 3.21%. 这说明, 不但标注性能较好的反向标注对正向标注有很好的补充作用, 即使是性能较差的正向标注也可以识别出很多反向标注不能正确识别的最大名词短语, 从而提高反向标注的性能. 实验验证了以基本组块和词混合标注单元的标注模型的正反双向标注结果具有互补性, 基于 “边界分歧” 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.

考察在动态特征采用 4 元历史标注结果时, 系统的反向补偿系数 ξ 对融合性能的影响. 系统的反向补偿系数 ξ 通过在语料库上进行搜索获得, 搜索范围是 $0.4 \sim 2.6$, 间隔为 0.2. 图 2 是采用 4 元历史标注时, 反向补偿系数 ξ 对融合效果的影响趋势图.

从融合结果的性能曲线来看, 最初随着反向补偿系数的增大, 融合性能呈现提高的趋势, 当反向补偿系数 ξ 的值达到 1.8 时, 融合性能达到最好, F-1 值达到 88.26 %, 伴随着反向补偿系数的继续增大, 融合性能呈现下降的趋势. 反向补偿系数的选择与两个方向的标注精度差相关, 对于两个方向标注精度相差不大的情况, 选择补偿系数为 1, 不偏向任何一方, 可以充分发挥它们在精度上的优势, 而当一个方向的单向标注性能明显优于另外一个方向时, 适当地选择补偿系数, 使融合结果更倾向于选择标注效果更好的方向的结果, 融合性能会更好.

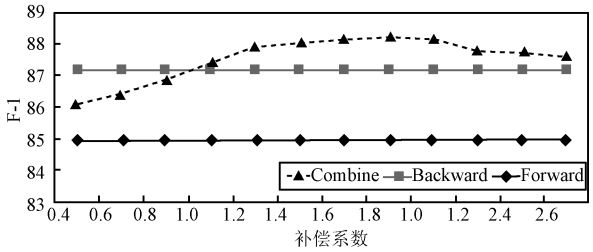

图 2 反向补偿系数对融合效果影响的趋势图

Fig. 2 Trends graph of backward compensating coefficient effect upon the combination performance

通过上述在开发集的实验可知, 双向序列标注融合模型在动态特征采用 4 元历史标注结果, 反向补偿系数 ξ 的值取 1.8 时汉语最大名词短语的识别取得了最优效果, 我们在测试集中采用这个设置, 实验结果如表 6 所示.

表 6 测试集上的最大名词短语识别的结果 (F-1 值)

Table 6 MNP recognition result of test set (F-1 value)

ξ	N	Strategy-1	Strategy-2	Strategy-3
1.8	4	85.90	86.07	88.24

在测试集的实验中, Strategy-3 融合后的 F-1 值达到 88.24 %, 比 Strategy-1 高出了 2.34 %, 比 Strategy-2 高出 2.17 %. 实验再次验证了以基本组块和词混合标注单元的标注模型的正反双向标注结果具有互补性, 基于 “边界分歧” 的融合算法有效地发掘了这一特性, 提高了汉语最大名词短语的识别精度.

4.4 比较实验

我们还将基于 SVM 的方法与基于 CRF 的方法进行了比较, 二者最大名词短语的单向标注结果如表 7 所示. 因为具有一阶马尔科夫性的 CRF 考虑前一个历史位置的标注, 相当于采用了 2 元标注历史结果, 所以将其与使用 2 元标注历史的 SVM 一起比较. 依此类推, 即: CRF 的阶数 = SVM 的历史标注元数 $(N) - 1$. 鉴于 CRF 对丰富特征的驾驭能力, 而前述的基于 SVM 的标注系统使用的特征比较简单, 我们特别为 CRF 标注器设计了一组更复杂的特征 (基本组块的中心词和中心词词性特征、句法标记特征以及边界词和词性特征的组合模板), 实验结果如表 7 所示.

从正向标注结果来看, 基于 CRF 的标注模型最好的标注结果 (5 阶) 优于基于 SVM 的标注模型 (3 元), 两者的标注结果的 F-1 值相差 2.21 %; 从反向标注结果来看, 基于 SVM 的标注模型和基于 CRF 的标注模型性能相当, 总的来说, 基于 CRF 的标注模型性能在单向标注中占优. 但是, 从标注方向角

度来看, 相比于基于 SVM 的系统, 全局搜索的基于 CRF 的系统基本不具有两个方向上的差异性, 正反向标注的性能差异不大, 因此基本不具有互补能力, 前后向的融合结果与单向标注结果相比没有显著的提高. 最终, 在单向标注中占优的基于 CRF 的标注模型的最好的融合结果反而比基于 SVM 的标注模型的融合结果的 F-1 值低了 0.81 %.

表 7 与基于 CRF 的序列标注结果比较 (F-1 值)

Table 7 Comparison with the CRF results (F-1 value)

N	SVM			CRF		
	Backward	Forward	Combine	Backward	Forward	Combine
2	85.67	84.85	86.12	72.32	72.28	72.35
3	86.34	85.08	87.09	83.46	83.48	83.51
4	87.21	84.91	88.12	85.29	85.32	84.62
5	87.23	85.07	87.83	86.07	86.12	85.59
6	86.69	84.59	87.35	87.18	87.29	87.31

5 结论

针对以词作为标注单位特征不足, 以基本组块作为基本标注单位受限于基本组块的质量的现状, 本章在分析现有方法的基础上, 提出了基于 "边界分歧" 的最大名词短语双向序列标注融合算法. 该算法以双语最大名词短语中的汉语最大名词短语识别为研究任务, 从基于 SVM 的序列标注算法的特点出发, 考查了混合特征的双向序列标注融合算法的适应性. 通过理论分析和实验, 证明了采用词和基本组块的混合标注单元的基于 SVM 分类器的确定性标注方法对汉语最大名词短语的识别是有效的, 并且其双向结果具有一定的互补性, 在此基础上提出的基于 "边界分歧" 的双向序列标注融合算法恰能发掘正反向的互补性, 并达到较高的融合精度.

本文提出的混合特征以及融合算法同样适用于其他具有相似特性的短语 (比如功能组块, 复杂介词短语等) 或语言, 因此具有一定的普遍意义.

References

1 Wang Z G, Zong C Q, Xue N W. Bidirectional sequence labeling via dual decomposition. In: Proceedings of the 12th China National Conference, CCL 2013 and First International Symposium. Suzhou, China: Springer, 2013. 325−332

2 Kudo T, Matsumoto Y. Chunking with support vector machines. In: Proceedings of the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies. Pittsburgh, PA, USA: Association for Computational Linguistics, 2001. 192−199

3 Tjong Kim Sang E F. Noun phrase recognition by system combination. In: Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference. Seattle, Washington, USA: Association for Computational Linguistics, 2000. 50−55

4 Chen W L, Zhang Y J, Isahara H. An empirical study of Chinese chunking. In: Proceedings of the 2006 COLING/ACL on Main Conference Poster Sessions. Sydney, Australia: Association for Computational Linguistics, 2006. 97−104

5 Jian Ping, Zong Cheng-Qing. A new approach to identifying Chinese maximal-length phrases by combining bidirectional labeling. *CAAI Transactions on Intelligent Systems*, 2009, **4**(5): 406−413
(鉴萍, 宗成庆. 基于双向标注融合的汉语最长短语识别方法. 智能系统学报, 2009, **4**(5): 406−413)

6 Li Ye-Gang, Huang He-Yan. A survey on Chinese chunk parsing. *Journal of Chinese Information Processing*, 2013, **27**(3): 1−8
(李业刚, 黄河燕. 汉语组块分析研究综述. 中文信息学报, 2013, **27**(3): 1−8)

7 Zhou Qiang, Sun Mao-Song, Huang Chang-Ning. Automatic identification of Chinese maximal noun phrases. *Journal of Software*, 2000, **11**(2): 195−201
(周强, 孙茂松, 黄昌宁. 汉语最长名词短语的自动识别. 软件学报, 2000, **11**(2): 195−201)

8 Chen K H, Chen H H. Extracting noun phrases from large-scale texts: a hybrid approach and its automatic evaluation. In: Proceedings of the 32nd Annual Meeting of Association of Computational Linguistics. New York, USA: Association for Computational Linguistics, 1994. 234−241

9 Cai D F, Liu X, Zhou Q L, Ye N. Chinese maximal noun phrase parsing based on cascaded conditional random fields. In: Proceedings of the 2009 International Conference on Natural Language Processing and Knowledge Engineering. Dalian, China: IEEE. 2009. 1−7

10 Li Guo-Chen, Wang Rui-Bo, Li Ji-Hong. Automatic labeling of Chinese functional chunks based on conditional random fields model. *Journal of Computer Research and Development*, 2009, **47**(2): 336−343
(李国臣, 王瑞波, 李济洪. 基于条件随机场模型的汉语功能块自动标注. 计算机研究与发展, 2009, **47**(2): 336−343)

11 Vapnik V N. An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, 1999, **10**(5): 988−999

12 Yang Zhi-Min, Shao Yuan-Hai, Liang Jing. Unascertained support vector machine. *Acta Automatica Sinica*, 2013, **39**(6): 895−901
(杨志民, 邵元海, 梁静. 未确知支持向量机. 自动化学报, 2013, **39**(6): 895−901)

13 Ying Wen-Hao, Wang Shi-Tong, Deng Zhao-Hong, Wang Jun. Support vector machine for domain adaptation based on class distribution. *Acta Automatica Sinica*, 2013, **39**(8): 1273−1288

(应文豪, 王士同, 邓赵红, 王骏. 基于类分布的领域自适应支持向量机. 自动化学报, 2013, **39**(8): 1273−1288)

14 Liang Yan-Ming, Su Fang, Li Qi, Liu Ding. A self-organizing algorithm for T-S fuzzy model based on support vector machine regression and its application. *Acta Automatica Sinica*, 2013, **39**(12): 2143−2149
(梁炎明, 苏芳, 李琦, 刘丁. 基于支持向量机回归的 T-S 模糊模型自组织算法及应用. 自动化学报, 2013, **39**(12): 2143−2149)

15 Zhang Xue-Feng, Wang Peng-Hui, Feng Bo, Du Lan, Liu Hong-Wei. A new method to improve radar HRRP recognition and outlier rejection performances based on classifier combination. *Acta Automatica Sinica*, 2014, **40**(2): 348−356
(张学峰, 王鹏辉, 冯博, 杜兰, 刘宏伟. 基于多分类器融合的雷达高分辨距离像目标识别与拒判新方法. 自动化学报, 2014, **40**(2): 348−356)

李业刚 山东理工大学计算机科学与技术学院副教授. 2015 年获北京理工大学计算机学院博士学位. 主要研究方向为语言信息处理, 机器翻译. 本文通信作者. E-mail: lyg8256@bit.edu.cn
(**LI Ye-Gang** Associate professor at the College of Computer Science and Technology, Shandong University of Technology. He received his Ph. D. degree from Beijing Institute of Technology in 2015. His research interest covers language information processing and machine translation. Corresponding author of this paper.)

黄河燕 北京理工大学计算机学院教授. 1989 年获中国科学院计算所博士学位. 主要研究方向为语言信息处理, 机器翻译. E-mail: hhy63@bit.edu.cn
(**HUANG He-Yan** Professor at the College of Computer Science and Technology, Beijing Institute of Technology. She received her Ph. D. degree from the Institute of Computing Technology, Chinese Academy of Science in 1989. Her research interest covers language information processing and machine translation.)

鉴 萍 北京理工大学计算机学院讲师. 2010 年获中国科学院自动化所博士学位. 主要研究方向为语言信息处理, 机器学习. E-mail: pjian@bit.edu.cn
(**JIAN Ping** Lecturer at the College of Computer Science and Technology, Beijing institute of technology. She received her Ph. D. degree from the Institute of Automation, Chinese Academy of Sciences in 2010. Her research interest covers language information processing and machine learning.)