基于多分类器加权投票法的越南语组合歧义消歧

李 佳[1] 郭剑毅[1,2] 刘艳超[1] 余正涛[1,2] 线岩团[1,2] 阮氏青娥[3]
(昆明理工大学信息工程与自动化学院 昆明 650500)[1]
(昆明理工大学智能信息处理重点实验室 昆明 650500)[2] (昆明理工大学国际学院 昆明 650093)[3]

摘 要 组合歧义消解是分词中的关键问题之一,直接影响到分词的准确率。为了解决越南语组合歧义对分词的影响问题,结合越南语组合型词的特点,提出了一种基于集成学习的越南语组合歧义消解方法。该方法首先通过人工选取越南语组合歧义词,构建出越南语组合歧义字段库,对越南语语料与越南语组合词词典进行匹配,抽取出越南语组合歧义字段;其次,采用三类分类器引入越南语词频特征和上下文信息,构建三类分类器消解模型,得到三类分类器消解结果;最后,计算出各分类器权值,通过阈值对越南语组合歧义进行最终分类。实验表明,所提方法的正确率达到了83.32%,与消歧结果最好的单个分类器相比准确率提高了5.81%。

关键词 组合词词典,组合歧义消解,越南语,集成学习,加权投票法
中图法分类号 TP303 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2018.01.029

Vietnamese Combinational Ambiguity Disambiguation Based on Weighted Voting Method of Multiple Classifiers

LI Jia[1] GUO Jian-yi[1,2] LIU Yan-chao[1] YU Zheng-tao[1,2] XIAN Yan-tuan[1,2] NGUYễN Qing'e[3]
(Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, China)[1]
(The Key Laboratory of Intelligent Information Processing, Kunming University of Science and Technology, Kunming 650500, China)[2]
(School of International Education, Kunming University of Science and Technology, Kunming 650093, China)[3]

Abstract Combinational ambiguity disambiguation is one of the key issues in participle and it directly affects the accuracy of participle. In order to solve the impact problem of combinational ambiguity on the participle in Vietnamese, combining the features of combinational words of Vietnamese, the paper proposed a Vietnamese combinational ambiguity disambiguation method based on integrated Learning. This method first selects Vietnamese combination of polysemy manually, constructs the Vietnamese combinational ambiguities library, matches Vietnamese and Vietnamese combinational-word dictionary, and extracts Vietnamese combinational ambiguities. Secondly, by using three kinds of classifiers to bring in Vietnamese word frequency features and context information, it constructs three class classifier degradation model, and gets the results. Finally, it calculats the classifier weights through the threshold to determine the final classification of Vietnamese combination ambiguity. Experiments show that the proposed method has the accuracy of 83.32% and its accuracy improves 5.81% compared with the single classifier.

Keywords Combinational-word dictionary, Combinational ambiguity disambiguation, Vietnamese, Integrated learning, Weighted voting method

歧义问题是自然语言处理中的核心难题,会严重影响分词的准确率,而分词则会直接影响到词性标注、实体识别、句法分析及语义分析等处理效果。Bar-Hillel[1]指出词语的歧义问题是机器翻译等研究所面临的主要障碍,不同的语言由于其结构或语言机制不同,因此产生歧义的影响也不同。在分词研究中,对英语来说[2],句子由于由字母组成,常用空格分隔每个单词,而且引用了标点符号(括号和引号等),分词已经相当明确,因此英语的歧义问题主要表现为一词多义或多词一义。对汉语而言[3],它是一种没有明显的形态界限可以作为分词依据的表意语言,越南语在很多方面与汉语存在相似之处。在越南语[4]中用空白分隔的不仅是单词,也可能是组成单词的词素;而且,许多越南语词素本身是单词,但也可以是多词素单词的一部分,其词素由它们之间的空白分隔,越南语通过组合词素来创建含义复杂的单词,而在大多数情况下单独考虑词素时其也具有某种意义,这种语言机制导致越南语存在分词的问题,即存在歧义。歧义分为两种基本类型:交叉歧义和组合歧义[5-6]。交叉歧义消歧问题[6-10]已经取得了一些成果,相比而言,产生组合歧义的原因较多,结构更复

到稿日期:2017-05-08 返修日期:2017-09-27 本文受国家自然科学基金(61262041,61562052,61472168),云南省自然科学基金重点项目(2013FA030)资助。

李 佳(1992—),男,硕士生,主要研究方向为自然语言处理,E-mail:270981402@qq.com;郭剑毅(1964—),女,教授,主要研究方向为自然语言处理、信息抽取,E-mail:gjade86@hotmail.com(通信作者)。

杂，对其消歧需考虑更多的信息，是自然语言处理任务中的难点问题。

越南语组合歧义的定义为：一个越南语词可以由一个或者多个词素构成。若存在越南语字符串“A B”（A 和 B 包含一个或者一个以上词素），A 和 B 分别可以单独成词，且 A 和 B 合起来也可以成词，则这种情况称为组合歧义。例如：

（1）Lòi anh ta nói rất/khó nghe（难听）/（他说的话很难听）

（2）Tôi rất khó（难）/nghe（听）/ thấy anh ta ơi ang nói gì（我很难听到他正在说什么）

其中，“khó nghe”在不同语境中需要切分（即“分”）或合并（即“合”）。

1 相关工作

目前，组合歧义消歧的研究方法主要有 3 种：1）基于规则方法。例如，冯素琴等[6]针对汉语组合歧义问题，基于上下文语境信息，应用对数似然比建立了语境计算模型，对一定规模的语料进行了实验，准确率达到了 95.60%，但是其没有考虑词性特征，词性特征是组合字段“分”或“合”的重要评判标准。2）基于统计方法。例如，秦颖等[11]针对汉语的组合歧义问题，运用最大熵分类器模型对 60 个组合歧义字段进行消歧，准确率达到了 88.05%，但由于最大熵分类器存在“标注偏置”问题，导致状态的转移存在不公平的情况。3）基于混合方法。例如，张严虎等[12]针对汉语的组合歧义问题，从规模为 53.9MB 的搜狗语料库中寻找词语搭配规则和语法规则，基于这些规则和朴素贝叶斯模型进行综合决策并进行中文组合歧义字段消解，准确率达到了 89%，但是其仅仅使用了词和词性特征，没有考虑上下文信息；在越南语方面，Ngo Q H 等[6]运用基于字典的最大匹配算法和支持向量机分类器来解决越南语分词中的组合歧义问题，但其没有考虑组合歧义字段内的词性特征及上下文信息，且没有对所用的语料或字典及越南语的组合歧义消解结果做明确说明。

一般而言，消歧的性能与提供给消歧的知识库或字段库数量、内容有直接关系。若字段库的规模小、内容有限，则消歧效果会受到影响。组合歧义消歧需要统一的组合词词典，需要考察更大范围的上下文中的语法、语境等信息。就越南语组合歧义消歧而言，由于已有的研究缺乏供消歧使用的标准的组合歧义语料库，且未充分利用语言和组合歧义字段的内部特征等（这些特征，如词性特征、上下文特征等，往往对组合消歧起到支撑作用），因此消歧性能仍不理想；另外，使用单分类器时往往顾此失彼，不能从总体上提高消歧性能；而集成多分类器利用了“取长补短”的原理，在对目标进行分类时把若干个分类器集成起来，对多个分类器的结果进行某种集成组合来决定最终的分类，以取得比单个分类器更好的效果[13]，其已经在文本分类等多个领域得到了较好的应用效果[14-17]。本文借鉴已有的研究，提出了基于集成多分类器加权投票法的越南语组合歧义消解的研究方法。越南语组合消歧的关键在于多特征（上下文信息及组合歧义字段的内部特征等）的融入[6-9,11]。最大熵分类器[11]的优点是可以将各种上下文信息组合在一起，并在保证已知的知识不被违背的情况下，让未知的部分信息量最大化（熵最大化），但其不足之处在于存在“标注偏置”问题。而条件随机场分类器[17]能够弥补最大熵分类器的不足，能很好地解决“标注偏置”问题，且具有很强的推理能力，能够使用复杂、有重叠性和非独立的特征进行训练和推理，能够充分地利用上下文信息作为特征，还可以任意地添加其他外部特征，使得模型能够获取非常丰富的信息。但条件随机场分类器的特征选择和优化是影响结果的关键因素，特征选择的优劣直接决定系统的性能。支持向量机分类器[6]恰好避免了特征的选择和优化问题，它的最终决策函数只由少数的支持向量确定，计算的复杂性取决于支持向量的数目，而不是样本空间的维数，在某种意义上支持向量机具有很多其他学习器不具备的优势。因此，本文尝试选用的 3 类分类器（最大熵、条件随机场和支持向量机）结合加权投票法，对越南语组合歧义进行消解。实验结果证明了本文所提方法的有效性，表明该方法能够有效地解决越南语组合歧义问题。

2 消歧原理框架

本文将越南语组合歧义消歧看作一个二分类问题，该问题通过分类算法能够得到解决。主要的消歧工作包括：抽取组合歧义字段、三类分类器消歧和三分类器集成。其原理框架如图 1 所示。

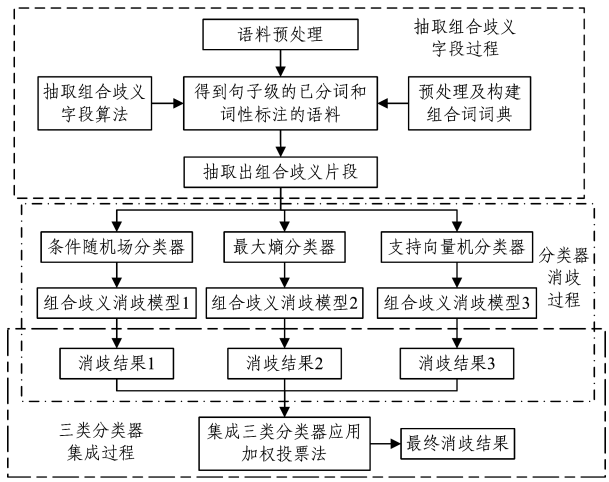

图 1 越南语组合歧义消解框架

Fig. 1 Framework of combination ambiguity disambiguation in Vietnamese

如图 1 所示，越南语组合型歧义消歧的具体步骤如下：1）抽取组合歧义字段。首先由越南专家从越南语字典中挑选出组合词，构建越南语组合词词典；然后将获得的词语级语料与越南语组合词词典进行匹配，最终得到组合型歧义字段。2）三分类器消歧。从越南语组合歧义字段中找出其特点，结合已抽取的越南语组合型歧义字段形成训练语料，分别使用最大熵分类器、条件随机场分类器和支持向量机分类器进行训练，最终三类分类器使用相同的测试语料进行消歧，得到各消歧结果。3）三类分类器集成。首先对三类分类器的消歧结果进行统一化，计算出各分类器的权值；其次根据消歧结果确定最佳阈值；最后通过加权投票法对三类分类器的消歧结果进行处理，得到最终的消歧结果。

3　词典的构建和组合歧义字段的获取

3.1　组合词形态

若存在歧义字段“A B”,则“AB”和“A_B”是它的两种切分结果。由于分量 A 和 B 中所含的词素的个数不确定,导致其表现形式可能有多种,如表 1 所列。

表 1　歧义字段的形态特征

Table 1　Morphological characteristics of ambiguous fields

表现形式	举例	描述
11	khói mù	XY 分别包含一个词素 Khói(A)mù(B)
12	mua nặng_hạt	X 包含 1 个词素,Y 包含 2 个词素 mua(A)nặng hạt(B)
21	khóm chiến lược	X 包含 2 个词素,Y 包含 1 个词素 khóm chi ến(A)luoc(B)
13	phối hợp chiến trường	X 包含 1 个词素,Y 包含 3 个词素 Phối(A)hop chi ến truòng(B)
22	nhân tố chiến tranh	X 包含 2 个词素,Y 包含 2 个词素 nhân tố(A)chi ến tranh(B)
31	nhân tài nảy nở	X 包含 3 个词素,Y 包含 1 个词素 nhân tài nảy(A) nở(B)

由于歧义字段的词素个数不同,导致其表现形式具有多样性。所选取的 60951 个组合词的形态统计如表 2 所列。

表 2　各个形态特征所占比例

Table 2　Proportion of characteristics of each form

表现形式	个数	所占百分比/%
11	45286	74.30
12	5260	8.63
21	4559	7.48
13	2182	3.58
31	1542	2.53
22	1273	2.09
其他	847	1.39

从表 2 中可以看出,表现形式为 11,12 和 21 的总和占到总数的 90.41%,而且组合词中词素个数越多,消解模型中的维度就越多,运算量会大大增加,从而严重影响消解效率。因此本文只考虑前 3 种形式的组合词歧义,也就是组合词中词素个数为 2 和 3。

3.2　组合词词典的构建

目前,网上没有公布相关组合歧义字段,没有越南语标准组合歧义字段集,因此本文提出以下方法来获得有效的组合歧义片段:第一步,建立一个组合词词典;第二步,运用组合歧义字段抽取算法将训练语料与组合词字典进行匹配。本文使用的词典通过 3 个步骤来处理并整合:首先通过人工扫描《新越汉词典》并整理得到 185046 条词;然后通过实验室的越南专家,在字典中逐条挑选出组合词;最后整合挑选出的组合词,进行去重处理,从而构建出一个包含 55104 条词的越南语组合词词典。

3.3　组合歧义字段的获取

基于越南专家挑选的组合词词典,结合搜集、处理得到的词语级语料,通过抽取算法进行匹配。其原理是在语料中抽取同时存在“合”形式(A_B)和“分”形式(A B)的字段,所抽取出的字段为组合歧义字段,从 3 万条分词语料(包含约 80 万个词)中共抽取 11043 条组合歧义字段作为实验所需的语料。歧义字段抽取方法的流程如图 2 所示。

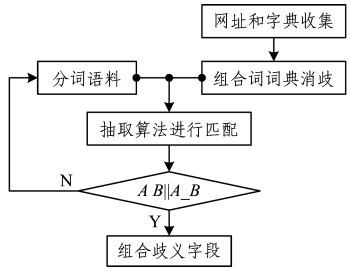

图 2　歧义字段抽取方法的流程图

Fig. 2　Flow chart of ambiguous field extraction method

4　三类分类器消歧模型的构建

4.1　特征选择

在基于特征的歧义消解中,特征的选择比算法的选择更重要[19],因此着重考虑特征的选择。本文从两个层面来讨论越南语组合歧义字段特征。

选择的特征包含两个层面:1)词的构成层面。越南语的词可能由多个词素构成,这已在第 2 节做了解释。考虑到词的形态对正确切分的影响,加入了形态特征,即组合歧义字段“分”或者“合”的发生频率,也即词频特征,作为第一种特征。2)语法层面。选择和歧义字段共现的上下文的词汇,考虑到上下文窗口对歧义字段的影响,文献[20]依据信息增益的原理对词语上下文的有效范围进行研究,得出选择窗口信息为[−5,+5]时效果最佳,因此借鉴中文词语上下文有效范围的研究方法,将窗口大小设定为 5,作为第二种特征;另一方面,由于越语词的独立能力很强,本身大多具有意义,因此又加入了歧义字段前后两个字的特征作为第三种特征;考虑到一些搭配关系可能对于歧义消解有帮助,又将二元搭配作为第四种特征。第 5 节给出了各个特征对各个分类器的组合歧义消解的贡献力度。

4 种特征具体是:

(1)词频特征:$f=\begin{cases}1, & p_{合}>p_{分}\\ 0, & \text{otherwize}\end{cases}$;

(2)上下文窗口的词:$w_{-k},\cdots,w_{+k}$,k 是最佳窗口值;

(3)歧义字段前后单字:c_{-2},c_{-1},c_{+1},c_{+2};

(4)二元搭配:$w_{-2}w_{-1}$,$w_{-1}w_{+1}$,$w_{+1}w_{+2}$。

4.2　条件随机场

4.2.1　条件随机场的原理

条件随机场(Conditional Random Fields,CRFs)[18,21-22]是一个在给定输入节点条件下计算输出节点的条件概率的无向图模型,其擅长处理序列标记问题。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:

$$P(y|x)=\frac{1}{Z(x)}\exp(\sum\lambda_k f_k(y_{i-1},y_i,x)+\sum\mu_k g_k(y_i,x)) \tag{1}$$

其中,每个 $f_k()$是观察值序列 x 中位置为 i 和 $i-1$ 的输出节点的特征,每个 $g_k()$是位置为 i 的输入节点和输出节点的特征,λ 和 μ 是特征函数的权重,$Z(x)$是归一化因子。

$$Z(x)=\sum_y\exp(\sum_i\sum_k\lambda_k t_k(y_{i-1},y_i,x)+\sum_i\sum_k\mu_k s_k(y_i,x)) \tag{2}$$

其中,t_k 是整个观测序列 x 的转换特性从 y_{i-1} 到 y_i 的状态,s_k

是观测序列 x 在状态特性 y_i 处的状态。

4.2.2 条件随机场分类器消解模型的建立

对分词语料进行预处理时使用 BES 方式进行标注，对每个词加上词位标记，然后进行模型的特征函数选取。特征模板文件中的每一行代表一个模板，模板的基本格式是：%x[row,col]，用于确定输出数据的一个 token，其中，row 确定与当前 token 的相对行数，col 确定列的绝对位置。模型构建的主要步骤如下：

(1)对处理过的分词语料运用 BES 方式进行标注，对每个词加上词位标记；

(2)针对所选取的越南语特征，通过依次迭代实验的方式提出特征模板集；

(3)训练语料和特征集通过 CRF 进行训练，最终得到消解模型参数序列。

4.3 最大熵

4.3.1 最大熵的原理

最大熵的原理是由 Jaynes E T 于 1957 年提出的，其主要思想是：在只掌握关于未知分布的部分知识时，应该选取符合这些知识但熵值最大的概率分布。1996 年，Berger 等提出了解决条件最大熵的两个自然语言处理的基本任务(特征选择和模型选择)的基本算法[23]。最大熵的基本思想是对未知事件不作任何假设，由此得到的分布与样本的实际分布一致。

4.3.2 最大熵分类器消解模型的建立

首先处理分词语料，然后把所选取的 3 类特征加入到语料中组成训练语料，再把训练语料整理成训练样例集(A,B)。建立最大熵组合歧义消解模型的工作流程如下：

(1)训练系统首先从分词语料中抽取特征；

(2)根据所抽取的特征，计算各特征的参数，获得最大熵模型；

(3)针对每条具体的测试样例，依据获得的最大熵模型计算各组合歧义特征的概率；

(4)选取概率最大的组合歧义特征作为消解特征，完成组合歧义消解。

4.4 支持向量机

4.4.1 支持向量机的原理

支持向量机的原理是 Vapnik 等人于 1995 年提出的完整的统计学习理论[24-25]，他们在此基础上发展了一种通用学习方法——支持向量机(Support Vector Machine，SVM)。SVM 是一种机器学习算法，主要用于解决二元分类问题，已经成功地运用到很多实际问题中，包括自然语言处理。它也被认为是目前最有效的方法之一。其工作原理就是寻找一个最优超分类平面，这个平面在满足分类精度的同时到两侧的距离最大。

4.4.2 支持向量机分类器消解模型的建立

通过对大量的组合歧义字段进行研究发现，组合歧义字段消解是指解决歧义字段“分”或“合”，它主要是组合词与前一个词和后一个词的关系，这正与词性二元模型公式相符合。考虑到组合歧义词与上下文的关系，定义了前后词性互信量 I_hI_f，前后词性互信息量的大小可以充分体现组合歧义字段和上下文的关系。由输出的 I_hI_f 值，得到二维向量$[I_hI_f]$，通过分类函数对得到的二维向量计算出结果。若计算结果为 1，则切分结果为 A B；若为−1，则切分结果为 A_B。

构建模型的主要步骤如下：

(1)把得到的越南语字典进行编号并将其导入 SVM 分类器；

(2)在越南语的分词训练语料中加入特征；

(3)经过 SVM 统计分析，得到 SVM 歧义消解函数，最终得到组合歧义消解函数。

5 集成学习

集成学习是一种独立于算法的机器学习策略，如果把单个分类器比作一个决策者，那么多个分类器相当于多个决策者共同决策一个问题。为了保证集成分类器取得比单个分类器更好的分类效果，需要依照两个原则：1)集成分类器的各子分类器具有不同的分类方法和训练方法；2)集成分类器中的子分类器所产生的错误是有差异的。目前，集成学习中既可使用各种不同分类器进行集成[26]，也可以使用同一种分类器进行集成，只是这些分类器之间的参数有所不同。集成结果往往能利用单分类器间的互补信息来减少单个分类器的误差，提高预测性能和分类精度。常用的集成学习方法[27]有 Boosting、Bagging、随机森林、投票法和学习法，其中，投票法最简单、最可靠，但由于简单投票法的集成过程只是结果的集成，只输出单纯的分类决策，没有其他附加信息，因此这种方法不能体现性能高的分类器的优势；加权投票法是一种很直观的方法，其给分类性能高的分类器赋予一个高的权值。本文采用集成多分类器加权投票法对越南语组合歧义进行消解。

如前所述，多分类器加权投票算法能够最大程度地纠正单个分类器的初始分类错误。首先，对各分类器的消歧结果统一化，然后计算各分类器的消解结果的正确率，再把各分类器的正确率相加得到总正确率的值，将分类器正确率与总的正确率相比，得到单个分类器的权值 λ，数学公式为(i 表示分类器的个数)：

$$\lambda_i=\frac{P(i)}{\sum_{i=1}^{n}P(i)} \tag{3}$$

然后根据各分类器所选类的消解结果概率值的平均值，确定最佳阈值 θ，其数学公式如下(i 表示分类器的个数)：

$$\theta_i=\frac{1}{n}\sum_{i=1}^{n}P(i),n=3 \tag{4}$$

最后通过加权投票法算法对三类分类器进行最终处理，得到最终消解结果，如式(5)所示(1 表示“合”，0 表示“分”)。

$$w=\begin{cases}1, & \text{if } \sum_{i=1}^{n}\lambda_iP(i)>\theta_i\\ 0, & \text{otherwise}\end{cases} \tag{5}$$

具体算法流程如图 3 所示。

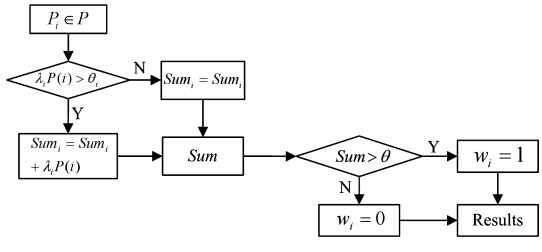

图 3 多分类器加权投票算法流程图

Fig. 3 Flow chart of multi-classifier weighted voting algorithm

集成三类分类器加权投票的具体流程如下:

(1)计算每个分类器的平均正确率 $P(i)$(i 表示分类器的个数);

(2)通过式(1)计算每个分类器的权值 λ_i(i 表示分类器的个数);

(3)通过式(2)计算最佳阈值 θ;

(4)把字段 AB 的各分类器消解正确率分别与所对应的权值 λ 相乘,得到 Sum_i;

(5)将各结果的 Sum_i 相加,得到 Sum,然后将 Sum 与阈值 θ 相比,即通过式(3)来判断最终消解结果。

6　实验及结果分析

6.1　实验测评标准

为了评估本文方法的效果,实验采用消解中常采用的评价标准——准确率(Precision)(正确消解越南语组合歧义字段个数与越南语组合歧义字段总个数的比值,衡量的是本文方法的识别准确率)作为评价标准,如式(6)所示。

$$准确率(P)=\frac{正确消歧组合型歧义字段个数}{组合型歧义字段总个数}*100\% \quad (6)$$

其中,正确率在 0 和 1 之间,数值越接近 1 表示越南语组合歧义字段消解的准确率越高。

6.2　实验数据

本文采用的主要语料源于在中越交流圈中抽取含有新闻、政治和文化等类型题材的网页信息,爬取的网页经过规则提取、去重、机器标注、人工校对等步骤形成文本语料库,其规模为约 3 万条句子,越南语词典包含约 6 万条词条。对所获得的越南语句子进行人工标注分词,最终获得 3 万条分词语料(包含约 80 万个词)。通过分词语料与词典的匹配方式获取的越南语组合歧义字段有 948 组,包含 11043 条词条,其中出现“分”和“合”的次数大于 20 的有 112 组,因此选用 112 组组合歧义字段进行实验。所有语料的编码方式均采用 UTF-8。

6.3　实验结果分析

为了验证本文方法的性能,从不同角度进行实验验证,考查 4 类特征对各个分类器分类效果的影响。用三类分类器与所提出的集成多分类器加权投票法的消解结果进行对比;将简单投票法与加权投票法进行对比。

实验 1　为了比较四类特征对三类分类器消解模型的贡献度,将四类特征分别作为独立特征依次融入三类分类器中,特征的贡献程度通过准确率进行比较,实验结果如表 3 所列(词频特征用“1”表示,上下文窗口的词用“2”表示,歧义字段前后单字用“3”表示,二元搭配用“4”表示)。

表 3　四类特征对消解模型贡献程度的对比/%

Table 3　Comparison of contribution to digestion model from four kinds of features/%

	1	2	3	4
SVM	59.42	67.24	70.16	65.29
CRF	62.12	64.32	72.31	66.57
ME	60.98	66.58	69.89	67.32

从表 3 中可以看出,歧义字段前后单字特征的准确率最高,词频特征的准确率最低,这是因为歧义字段前后单字特征较准确地确定了当前词的状态,所以推断歧义字段前后单字特征对消解模型的贡献度最高;词频特征只考虑到了词的“合”与“分”的概率,没有附带其他特征信息,导致该特征对消解模型的贡献度较低。

实验 2　为了评估所提出的集成三类分类器加权投票法,将所选取的 112 组组合歧义字段的每组分成 5 份,其中 1 份作为测试语料,其他 4 份作为训练语料,做五倍交叉实验。然后求各组的平均准确率,作为单分类器测评结果,再通过加权投票法对三类分类器消解结果进行处理。实验结果如表 4 所列(WV 代表加权投票法,AVE 代表平均率)。

表 4　加权投票法与各分类器的对比/%

Table 4　Comparison between weighted voting method and each classifier/%)

	1	2	3	4	5	AVE
SVM	76.3	77.3	75.7	77.2	75.8	76.5
CRF	77.8	77.3	78.2	75.2	79.3	77.6
ME	75.3	76.7	74.6	74.8	73.7	75.0
WV	82.3	84.3	81.6	83.9	85.3	83.2

从表 4 中可以看出,集成多分类器加权投票的方法明显比任务单分类器的准确率高,这是因为加权投票法集成了各分类器间的互补信息,减少了单个分类器的误差,进而提高了预测性能和分类进度,因此所提出的加权投票方法的消解效果比单分类器的消解效果好,平均准确率为 83.23%。

实验 3　为了比较简单投票法与加权投票法的效果,将所进行的三类分类器消解结果通过简单投票法进行处理。实验对比结果如表 5 所列(WV 代表加权投票法,SV 代表简单投票法)。

表 5　加权投票法与简单投票法的对比/%

Table 5　Comparison between weighted voting and simple voting/%

	1	2	3	4	5
WV	82.32	84.36	81.67	83.97	85.39
SV	76.74	77.12	76.24	75.89	77.32

从表 5 可以看出,加权投票法比简单投票法的准确率高,简单投票法是由分类器先对数据集进行判断得出分类结果。对自己所预测的类投一票,最后得票最多的类就是简单投票法的最终消解结果。这种“一人一票”的整合原则没有考虑各分类器的分类性能及分类特征的不同,因此这种简单投票的方法无法体现性能高的分类器的优势,而且与实验 2 相比,简单投票法比最好的消解结果差,证明了简单投票法的缺陷。而加权投票法是一种很直观的方法,其给分类性能高的分类器赋予一个高的权值,从而更公平地对待单分类器的消解结果。

结束语　目前,针对越南语组合歧义消歧还没有有效的解决方法。本文提出采用集成三类分类器加权投票法对越南语组合词歧义进行消歧,通过建立组合词词典,抽取组合歧义字段,融入词性及上下文特征,建立集成学习的消歧模型,并将其与各个单分类器的消歧效果和简单投票法做对比实验。结果表明,本文所提方法可以对越南语组合歧义进行有效的消除,准确率达到 83.32%。下一步工作将进一步研究动态识别、未登录词等对组合歧义的影响。

参 考 文 献

[1] BAR-HILLEL Y. The present status of automatic translation of languages[J]. Advances in Computers, 1960, 1: 91-163.

[2] SCHMID H. Tokenizing. In: Anke Lüdeling and Merja Kytö[M]// An International Handbook. Mouton de Gruyter, Berlin, 2007.

[3] LIANG N Y. Written Chinese divided into automatic system—CDWS [J]. Journal of Chinese Information Processing, 1987, 1(2): 46-54. (in Chinese)
梁南元. 书面汉语自动分词系统—CDWS[J]. 中文信息学报, 1987, 1(2): 46-54.

[4] LÊ H P N T M, HUY Ên A R, Vinh H T. A Hybrid Approach to Word Segmentation of Vietnamese Texts[C]// Proceedings of the 2nd International Conference on Language and Automata Theory and Applications. 2008.

[5] FENG S Q, CHEN H M. Context-based Approach to Combinational Ambiguity Resolution in Chinese Word Segmentation[J]. Journal of Chinese Information Processing, 2007, 21(6): 13-16. (in Chinese)
冯素琴, 陈惠明. 基于语境信息的汉语组合型歧义消歧方法[J]. 中文信息学报, 2007, 21(6): 13-16.

[6] NGO Q H, DIEN D, WINIWARTER W. A hybrid method for word segmentation with English-Vietnamese bilingual text[C]// 2013 International Conference on Control, Automation and Information Sciences (ICCAIS). IEEE, 2013: 48-52.

[7] WANG S L, WANG B. A Chinese Overlapping Ambiguity Resolution Method Based on Coupling Degree of Double Characters [J]. Journal of Chinese Information Processing, 2007, 21(5): 14-17. (in Chinese)
王思力, 王斌. 基于双字耦合度的中文分词交叉歧义处理方法[J]. 中文信息学报, 2007, 21(5): 14-17.

[8] LI M, GAO J, HUANG C, et al. Unsupervised training for overlapping ambiguity resolution in Chinese word segmentation[C]// Proceedings of the second SIGHAN workshop on Chinese language processing. Association for Computational Linguistics, 2003: 1-7.

[9] XIONG M M. Vietnamese news event element extraction method study[D]. Kunming: Kunming University of Science and Technology, 2016. (in Chinese)
熊明明. 越南语词法分析研究[D]. 昆明: 昆明理工大学, 2016.

[10] PHAM D D, TRAN G B, PHAM S B. A hybrid approach to vietnamese word segmentation using part of speech tags[C]// International Conference on Knowledge and Systems Engineering, 2009(KSE'09). IEEE, 2009: 154-161.

[11] QIN Y, WANG X J, ZHANG S X. Research on Combinational Ambiguity in Chinese Word Segmentation [J]. Journal of Chinese Information Processing, 2007, 21(1): 1-8. (in Chinese)
秦颖, 王小捷, 张素香. 汉语分词中组合歧义字段的研究[J]. 中文信息学报, 2007, 21(1): 1-8.

[12] ZHANG Y H, PAN L L, PENG Z P, et al. Resolving combinational ambiguity in Chinese word segmentation based on rule mining and Naive Bayes method [J]. Journal of Computer Applications, 2008, 28(7): 1686-1688. (in Chinese)
张严虎, 潘璐璐, 彭子平, 等. 基于规则挖掘和 Naive Bayes 方法的组合型歧义字段切分[J]. 计算机应用, 2008, 28(7): 1686-1688.

[13] SAHA S, EKBAL A. Combining multiple classifiers using vote based classifier ensemble technique for named entity recognition [J]. Data & Knowledge Engineering, 2013, 85: 15-39.

[14] REMYA K R, RAMYA J S. Using weighted majority voting classifier combination for relation classification in biomedical texts[C]// 2014 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT). IEEE, 2014: 1205-1209.

[15] REYHANIAN S, ARBABI E. Weighted Vote Fusion in prototype random subspace for thermal to visible face recognition[C]// 2015 2nd International Conference on Pattern Recognition and Image Analysis (IPRIA). IEEE, 2015: 1-5.

[16] NIKAN S, AHMADI M. Human face recognition under occlusion using lbp and entropy weighted voting[C]// 2012 21st International Conference on Pattern Recognition (ICPR). IEEE, 2012: 1699-1702.

[17] E SILVA R R V, DE ARAUJO F H D, DOS SANTOS L M R, et al. Optic disc detection in retinal images using algorithms committee with weighted voting[J]. IEEE Latin America Transactions, 2016, 14(5): 2446-2454.

[18] MAI F, WU S, CUI T. Improved Chinese Word Segmentation Disambiguation Model Based on Conditional Random Fields[C]// Proceedings of the 4th International Conference on Computer Engineering and Networks. Springer International Publishing, 2015: 599-605.

[19] YAROWSKY D, FLORIAN R. Evaluating Sense Dis2 ambiguation Performance Across Diverse Parameter Spaces[J]. Natural Language Engineering, 2002, 8(4): 293-310.

[20] LU S, BAI S. Quantitative Analysis of Context Field in Nature Language Processing[J]. Chinese Journal of Computers, 2001, 24(7): 742-747. (in Chinese)
鲁松, 白硕. 自然语言处理中词语上下文有效范围的定量描述[J]. 计算机学报, 2001, 24(7): 742-747.

[21] DELLA PIETRA S, DELLA PIETRA V, LAFFERTY J. Inducing features of random fields[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1997, 19(4): 380-393.

[22] WALLACH H. Efficient training of conditional random fields [D]. University of Edinburgh, 2002.

[23] BERGER A L, PIETRA V J D, PIETRA S A D. A maximum entropy approach to natural language processing[J]. Computational linguistics, 1996, 22(1): 39-71.

[24] VAPNIK V. The nature of statistical learning theory [M]. Springer Science & Business Media, 2013.

[25] VAPNIK V N, VAPNIK V. Statistical learning theory[M]. New York: Wiley, 1998.

[26] LI Y, TAX D M J, DUIN R P W, et al. Multiple-instance learning as a classifier combining problem[J]. Pattern Recognition, 2013, 46(3): 865-874.

[27] 周志华. 机器学习[M]. 北京: 清华大学出版社, 2016: 171-184.