文章编号：2096-1618(2023)03-0298-08

基于 Re-Perceptron-CRF 的规范类文本分词研究

李宝林， 刘宇韬
(成都信息工程大学物流学院，四川 成都 610103)

摘要：通过 Re-Perceptron-CRF 组合方法，利用规范类文档特点，对关键词进行切分。分别采取 Viterbi、Perceptron、CRF 和 Re-Perceptron-CRF 4 种算法分别对规范类文本进行分词研究。具体为基于句法分析对规范类文本使用正则表达式进行标准化处理，得到适合分析的预处理文本，并通过 Perceptron 与 CRF 的双重算法返回各自的最优结果。实验表明，Re-Perceptron-CRF 算法明显提高分词效果，在准确率和召回率上均有良好表现，其准确率和召回率分别达到94.36%和97.02%。该方法为规范类文本中文分词相关工作提供一定的研究思路，为后续应用提供好的数据支撑。但由于数据量较小，该方法仅适用于特定领域，如建筑检测领域。

关 键 词：管理科学与工程；文本分析；中文分词；Re-Perceptron-CRF；词性标注
中图分类号：TP391.1 **文献标志码**：A
doi：10.16836/j.cnki.jcuit.2023.03.008

0 引言

在进行中文文本识别时，由于中文文本中每段话的字词都是紧密相连，缺乏明显的词语边界[1]，计算机不能直接识别这些连续的字词。在自然语言处理(natural language process)[2]中，计算机需要将这些词语从一段话中识别出来，为其添加分隔符[3]，这一过程即为中文分词——将一段连续的话拆分成若干词语并按原文顺序拼接。然而，随着语料库的不断加深，传统的识别算法早已无法满足需求，面对的场景领域多样、文本的内容格式不统一，如何准确分词是当下一大技术难点。本文以规范类文本为研究对象，解决规范类文本的分词需求，从相对规范到绝对规范，即达到最终分词结果全部为有效信息的目的，将是本文的重点研究内容。

1 相关研究

中文分词存在诸多研究难点，尽管国家早已颁布信息处理的分词规范，但实际应用中很容易受主观因素影响导致结果大相径庭。同时，各种未登录词[4](即语料库中并未收录或从未训练过的词)的相继出现，包括网络用语、领域术语、专有名词等，都会严重影响分词的准确性。在规范类文本中，其内容具有高度规范性，也就是每个字词都是经过缜密推敲后定稿，绝不会出现语气助词或多余的修饰词。因此，其内容包括很多组合性专业名词和连接助词，对于现有语料库而言，这些词语通常很容易被误分。如组合词“邵氏硬度”通常都会被划分为“邵氏”和“硬度”两个词语；“擦伤、划伤”中两个词语的词性实则为名词，但基于大量语料库的概率情况而言，通常又会将其判定为动词；“连接严重锈蚀”这样一类组合拼凑词为一个整体，但是计算机通常将其划分3部分：“连接”“严重”“锈蚀”，这完全背离词语本意。而划分词性是计算机实现分词的依据，根据词的特点(语法、形态、句意等)将其划分为不同种类。每一段句子都是由不同种类的词性按照一定的规律排列组合而成，通过识别每一个词的词性并对其标注，进而达到词语划分的目的，也就是确定各词归属类别的过程[5]。通过联系上下文关系，在特定的语境中，采用得体的方法确定词性，消除语法兼类[6-7]。

常用的词性标注包括4大类[8]：基于规则的词性标注方法[9]，主要是根据上下文的词语联系、搭配关系将自定义规则写入确定当前词的词性，虽然能够高效利用上下文信息，但随着语料库的增加，人力投入不断增大，且这种规则的覆盖面并不广泛，容易发生规则冲突[10]，无法应用于大部分领域。基于统计模型的词性标注方法，该方法的核心思路是将一段词性视作一段序列标注问题，判定每一个词出现的词性概率。通过使用具有正确词性标注数据的语料库训练经典模型，如HMM(隐马尔科夫模型)[11]、CRF(条件随机场)[12]、ME(最大熵)[13]等，达到词性自动标注的效果，极大减少了人力。不过词语之间长距离的依赖现象和不确定性并不能很好地解决[14]。基于统计方法

收稿日期：2022-07-05
基金项目：四川省科技服务业示范资助项目(2021GFW015)；四川省电子商务与现代物流研究中心重点资助项目(DSWL21-3)

与规则方法相结合的词性标注方法[15],将两种方法结合并针对性地使用,即筛选根据统计方法标注的结果,对词性标注可信度较低的目标进行规则匹配,进而消除歧义。基于深度学习的词性标注方法,同样也是解决序列标注任务,常用方法有 LSTM+CRF、BiLSTM+CRF[16]等。

为探寻更加高效的分词方式和精确的词语识别率,学者们不断探索,实现算法的改进。刘伟等[17]提出一种通过计算语境相似度检验中文分词一致性的方法,该方法依赖词性和依存句法,利用词向量进行语义编码,通过实验发现能有效提高分词一致性检验的准确率,对人工分词语料标注相关工作具有一定辅助作用。凤丽洲等[18]提出一种组合词迭代的双向匹配分词方法,该方法基于 N-gram 统计模型,能有效规避免长条词语的分词准确率的影响,实现最优分词序列。Liu J 等[19]通过从字典中随机抽取单词生成伪标记数据和共享相同的网络参数,联合训练汉语分词和词分类任务这两种方法对词典中的中文进行分词,对训练数据不足的情况能显著提高中文分词性能。Gan L 等[20]通过 BERT 研究上下文字符嵌入的影响,提出一种将单词信息整合到 Self-Attention 网络中的分词方法,并通过与 BiLSTMs 对比,发现该模型具有显著优越性。Si H 等[21]利用复杂网络的特点对中文分词进行研究,发现复杂网络特征算法对解决分词速度和准确率的冲突问题具有明显效果。Yan H 等[22]提出一种基于图的中文分词和依存句法分析集成模型,该模型可以在选取更少的特征下拥有更高的分词效率。

规范类文本是一种属于高度规范的非结构化数据,与一般的非规范类文本相比,该文本主要为定量描述,即文本内容通常可以直接进行实证分析。以《玻璃幕墙缺陷类型》为例,在“爆边:长度或宽度不得超过玻璃的厚度”中,明确指出缺陷名称为“爆边”,“长度或宽度不得超过玻璃的厚度”则以明确的玻璃厚度范围限定爆边的长度与宽度,是一种数字化描述。而非规范类文本通常是定性描述,如“某玻璃幕墙质量未达标”“某窗户的裂纹缺陷较大”等没有明确指标的描述通常都为非规范类文本。在规范类文本中,尽管文本内容相对严谨(即不存在语气词、叹词等无关词),但并非所有内容都是关键内容。如在“爆边:长度或宽度不得超过玻璃的厚度”这一文本中,真正有用的词语仅为“长度”“宽度”“不超过”“玻璃厚度”。同时,在传统分词中,很难将“不超过”“不大于”“不允许”以及“玻璃厚度”“点状缺陷”等词作为一个整体进行切分,而多数分词结果均以“不/超过”“点状/缺陷”两个部分呈现。针对以上分析,本文提出一种 Re-Perceptron-CRF 的分词方法——通过正则匹配将规范类文本进行内容标准化,进而提高分词的精确度,并与经典模型进行对比实验。

2　基于规范类文本的分词算法构建

中文分词通过将一段文本拆分为一系列词语后,分别为这些词语进行词性标注工作,通过确定这些词性才能按照原本顺序拼接并重新形成完整的文本[23]。

在中文分词中需要用到由语音、词汇、语法构成的语言模型。语言模型就是在给定一段句子的条件下,将词语出现的概率进行计算的模型,而统计的对象就是人工标注而成的语料库。

主要语言模型有美国语言学家 Chomsky 提出的 PSG[2](短语结构语法)模型、统计语言模型 n-gram(n 元语法)模型[24]和深度学习语言模型 NNLM(神经网络语言模型)[25]。本文主要使用基于统计语言模型的 Viterbi[26]算法和基于深度学习语言模型的 Perceptron 感知机[27]与 CRF 条件随机场进行实验。

其中,统计语言模型 n-gram 的意思是,每个词语出现的概率仅受该词语之前的 $n-1$ 个单词影响。换言之,一元语法模型表示各个词语相互独立,二元语法表示该词语出现的概率只取决于自身前一个词语影响,以此类推。具体公式:

$$p(\omega)=\prod_{t=1}^{n}p(\omega_t|\omega_{t-n+1},\omega_{t-n+2},\cdots,\omega_{t-2},\omega_{t-1})$$

深度学习语言模型则是指利用神经网络对语言模型进行训练,每次得到一个字符串作为一个句子出现的概率,每一个句子本质上就是一个词向量[28]。

2.1　基于 Viterbi 的规范类文本分词算法构建

Viterbi 算法本质上是使用动态规划的方法递归求解隐藏状态序列,是一种剪枝算法,用于寻找一段观测序列的维特比路径(隐含状态最优路径),本文目标就是对一个最优路线二分类问题进行求解。在规范类文本中,对文本分词可以采用 Viterbi 算法进行求解,具体实现以“脱胶:不允许”为例,具体流程见图 1。

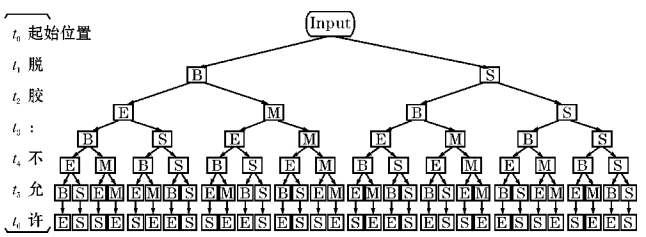

图 1　Viterbi 词性确定流程

每一个汉字对应一个位置 t。其中的 B、S、M、E 表示每个汉字的状态：B（begin）——词首，M（middle）——词中，E（end）——词尾，S（single）——单独成词。

首先输入本文模型 λ 和观测序列 $O=$（“脱”，“胶”，“：”，“不”，“允”，“许”），输出目标即求得从 t_0 位置起始到 t_6 位置终止整个过程中，该观测序列 O 的最优路径 $I^*=(i_1^*,i_2^*,\cdots,i_t^*)$，即序列 O 的对应的最佳隐藏状态。

根据 Viterbi 算法，进行前置变量定义：

定义 δ 为在位置 t 状态为 i 的所有单个路径 $(i_1,i_2,\cdots,i_t)$ 中的概率最大值：

$$\delta_{t+1}(i)=\max_{i_1,i_2,\cdots,i_t} P(i_{t+1}=i,i_t\cdots,i_t,\cdots o_{t+1},\cdots,o_1\mid\lambda)$$
$$=\max_{1\leqslant j\leqslant N}[\delta_t(j)a_{ji}]b_t(o_{t+1}),$$
$$i=1,2,\cdots,N;t=1,2,\cdots,T-1$$

定义 ψ 为在时刻 t 状态为 i 的所有单个路径 $(i_1,i_2,\cdots,i_{t-1},i_t)$ 中概率最大的路径的第 $t-1$ 个结点：

$$\psi_t(i)=\operatorname*{argmax}_{1\leqslant j\leqslant N}[\delta_{t-1}(j)a_{ji}],i=1,2,\cdots,N$$

根据上述公式，按照以下步骤进行算法求解：

（1）确定从 t_0 位置（Input 位置）到 t_1 位置的所有路径，目前这些路径都是最优备选路径；

（2）依次选择起始位置到 t_1 位置的所有路径，并确定在 t_1 位置到 t_2 位置的所有路径选择；

（3）确定从 t_1 到 t_2 位置备选的所有路径中概率最大者（即 δ_1 值）最大的路径后，选择 t_1 位置到 t_2 位置最优的路径，然后根据将当前路径最终点位作为最大路径结点 ψ_2 并将其他路径选择舍弃；

（4）重复（3），确定 t_1 位置所有路径分别到 t_2 位置上第 2 个点位处的路径中 δ_2 最大的路径，并将当前路径最终点位作为最大路径结点 ψ_3，同时舍弃其他路径；

（5）来到 t_2 位置，重复步骤（3）、（4）中的操作，确定 t_2 位置所有路径分别到 t_3 位置上第 1、2 个点位处的路径，选择最优路径，舍弃其他路径；

（6）递推，重复上述操作，直到抵达 t_6 位置结束迭代，也就是抵达句子末尾；

（7）最优回溯路径，求得最优路径 $I^*=$（B，E，S，B，M，E），具体路线见图 1 加粗部分。

2.2 基于 Perceptron 感知机的规范类文本分词算法构建

感知机算法是一种迭代式的算法：通过在训练集上进行多次迭代，每次读入一个样本并进行预测后，将预测结果与正确答案对比，计算误差，根据误差更新模型参数，再次训练，反复迭代，直到误差达到最小为止。

在规范类文本分词中，通过使用结构化感知机进行实验，得知相比普通感知机，其对更新参数的奖惩机制与特征函数的权重紧密相连，更能提高分词准确率，同时还能调整学习率。以判定“擦伤”词性为例，具体流程见图 2。

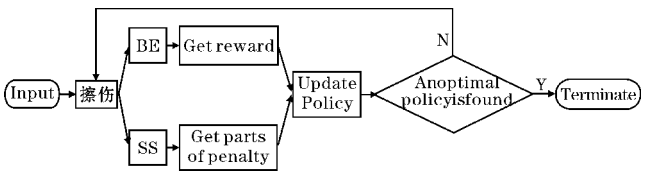

图 2 Perceptron 词性判别奖惩流程

（1）输入训练样本 x，同时定义打分函数 score$(x,y)=\omega\cdot\varphi(x,y)$。其中，$\varphi(x,y)$ 表示 x 和 y 之中的特征，ω 表示每个特征对应的权重，这些特征来源于输入文本“擦伤”的词性，即 $\{B,M,E,S\}$，并用这些特征进行序列标注；

（2）根据样本 x 和对应真实值 y，可以得到 $\hat{y}=\operatorname*{argmax}_{y\in Y}[\omega\cdot\varphi(x,y)]$；

（3）将预测的 $y=\operatorname*{argmax}_{y'\in Y}[\omega\cdot\varphi(x,y')]$（预测的擦伤的词性标注）与 $\hat{y}=\operatorname*{argmax}_{y\in Y}[\omega\cdot\varphi(x,y)]$（真实的擦伤词性标注）比较，即确定“擦伤”的词性划分是否为需求词性；

（4）如果两者不同，则对其惩罚，即根据打分函数扣除其分值，而后对参数更新；

（5）重复（4）中操作，反复迭代，直到找到最优解，并给予奖励，结束训练。

2.3 基于 CRF 条件随机场的规范类文本分词算法构建

条件随机场（conditional random field，CRF）是通过给定观测序列 $X=(x_1,x_2,\cdots,x_{n-1},x_n)$ 和状态序列 $Y=(y_1,y_2,\cdots,y_{n-1},y_n)$，进而求解条件概率 $P(Y|X)$ 最优的无向图[29]。

以“中空腔有异物：不允许”为例，展示 CRF 在规范类文本的分词流程（图 3）。

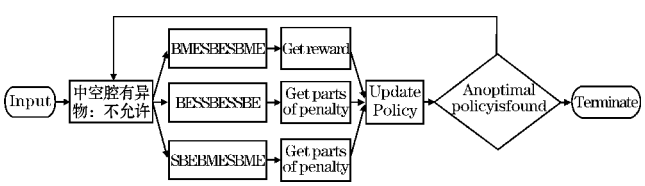

图 3 CRF 分词奖惩流程

（1）采用 CRF 对规范类文本分词是一种序列化算

法(sequence labeling algorithm),观测序列 X 即为待分词串,状态序列 Y 就是对应的词性,并且 X 和 Y 两个序列等长,此时应将其视作线性链(linear chain)条件随机场,即满足马尔可夫性[30]:

$$P(Y_i|X,Y_1,\cdots,Y_{i-1},Y_{i+1},\cdots,Y_n)=P(Y_i|X,Y_{i-1},Y_{i+1})$$

(2)观测序列 X 即 $X=\{$"中","空","腔","有","异","物",:,"不","允","许"$\}$,状态序列同样为 $\{B,M,E,S\}$,并根据上述给定序列 X 和 Y 以及所求解概率 $P(Y|X)$,有如下形式:

$$P(Y|X)=\frac{1}{Z(x)}\exp\left(\sum_{i,k}\lambda_k t_k(y_{i-1},y_i,x,i)+\sum_{i,l}u_l s_l(y_i,x,i)\right)$$

(3)$Z(x)$ 为归一化函数:

$$Z(x)=\sum_{y}\exp\left(\sum_{i,k}\lambda_k t_k(y_{i-1},y_i,x,i)+\sum_{i,l}u_l s_l(y_i,x,i)\right)$$

并且,t_k 和 s_l 为特征函数,λ_k 和 u_l 为对应的权重值。

(4)通过对特征函数的所有权重值进行训练,遍历出 $X=\{$"中","空","腔","有","异","物",:,"不","允","许"$\}$ 中所有可能出现的序列,并对其中所有错误的状态序列进行惩罚,不断更新模型参数。

与 Perceptron 相比,CRF 在特征函数、权重向量、打分函数预测算法以及结构化学习上完全相同,其区别只在于 Perceptron 每次只使用一个训练实例,而 CRF 则考虑整个数据集。换言之,Perceptron 会惩罚最严重的错误情况,而 CRF 使所有错误情况均摊承受惩罚。

2.4　基于 Re-Perceptron-CRF 的规范类文本分词算法构建

该规范类文本存在部分定性描述,容易影响分词结果,但经过对文本的研究,发现多数文本可以归为不同类别。因此,本文提出一种基于 Re-Perceptron-CRF 的组合分词算法,首先将初始规范类文本按照规则进行分类,对分类后的文本进行依存句法分析[31],根据分析结果采取正则表达式[32]匹配换行符、空格、转义符号等文本,从而减少无用字词(的、得等无意义字)和标点符号等无效字符串对分词效果的影响,以及将部分词语进行合并,再根据处理后的文本将 Perceptron 和 CRF 进行融合匹配,即同时执行两种算法,返回两种算法各自的最优结果。

在对文本修正的过程中,发现规范类文本多数为组合词语,而在对这些组合词语进行识别时很容易将其切分为两组词语。因此,该算法通过对比 Perceptron 和 CRF 的分词情况,优先返回各自词数更长匹配的结果。具体流程见图 4。

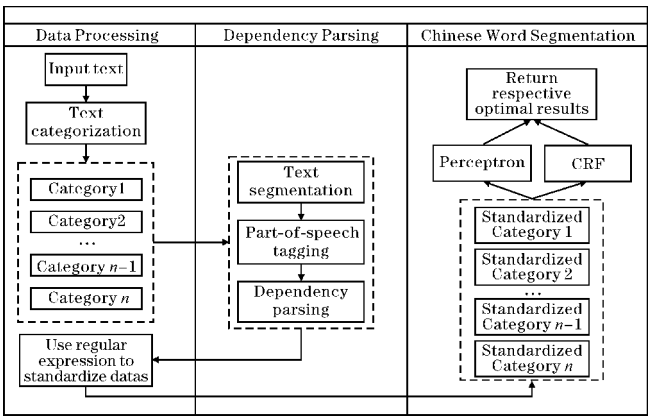

图 4　Re-Perceptron-CRF 算法流程

3　实验及分析

3.1　实验数据

实验数据节选自《玻璃幕墙缺陷类型》,该文本为对玻璃幕墙中常见的缺陷类型及其详情描述。在玻璃幕墙中存在多种规范,该文本即是记录缺陷名称及其表现情况和判定依据。例如,判定某幕墙存在划伤缺陷的依据为:

(1)当划伤宽度≤0.1 mm且长度≤100 mm时,每平方米面积内允许存在 4 条划伤处;

(2)当划伤宽度为(0.1,0.5)mm 且长度≤100 mm时,每平方米面积内允许存在 3 条划伤处。

根据建模流程和文本描述情况,原文本可按照如下方式进行分类。

(1)d-f:defect-forbid ,缺陷不允许,对于某类缺陷明令禁止存在,文本格式为:"缺陷:不允许"。共 2638 个词数;

(2)t-d:text-description,文本描述,对各种尺寸要求或其他规格的单一描述,为一个单句。共 1356 个词数;

(3)c-t-d:condition-text-description,条件文本描述,对相同场景不同条件下,某一尺寸要求或其他规格的规范描述,为一个长句。共 3162 个词数;

(4)m-t-d:multiple-text-description,多项文本描述,对不同场景下,某一尺寸要求或其他规格的规范描述,为一个长句。共 13918 个词数;

缺陷描述分类情况见表 1。

表 1 缺陷描述分类说明

Category	Instances
d-f	裂纹:不允许
t-d	爆边:长度或宽度不得超过玻璃的厚度
c-t-d	断面缺陷:公称厚度不超过8 mm时,不超过玻璃厚度;8 mm以上时,不超过8 mm
m-t-d	擦伤、划伤:一个分格的深度不大于膜厚度的 2 倍;一个分格的面积不大于500 mm^2;一个分格的总长度不大于150 mm;一个分格的总数不大于 4 处

3.2 实验过程

通过为数据集各词进行{B,E,M,S}词性标注,进而对比各算法分词结果、评估算法优劣,并得出相应结论,主要流程如下:

(1)使用传统算法 Viterbi、Perceptron、CRF 进行词性标注分词实验;

(2)构建组合算法 Re-Perceptron-CRF,首先对文本分类,并对文本进行依存句法分析,根据分析结果通过正则匹配进行文本内容标准化后通过 Perceptron 与 CRF 的双向算法,返回各自的最优结果;

(3)对上述算法进行结果统计,包括计算分词时间、整理分词结果和歧义词切分情况;

(4)使用评估指标 P、R、F_1、R(oov)、R(iv)综合评估所有算法性能,对比算法优劣并分析。

3.3 实验结果

3.3.1 依存句法分析

依存句法结构本质上是反映词对之间的关联,表示一个词对另一个词的支配关系[33]。句法分析则表示根据指定语法,对文本中含有的句法单位及其之间的依存关系进行自动识别[34]。

将基于 Python3.7 的编程环境,通过 PyhanLP 模块包进行依存句法分析,并以"爆边:长度或宽度不得超过玻璃的厚度"为例展示。分析结果见图 5 ~ 7。

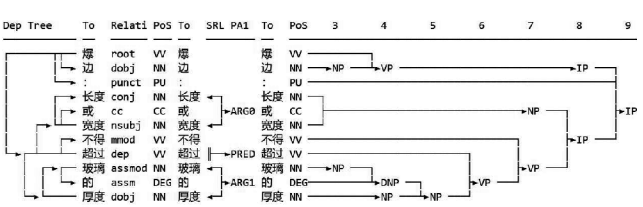

图 5 语言学结构图

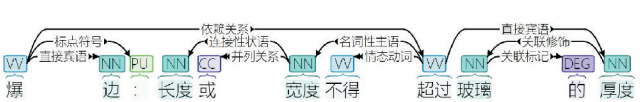

图 6 句法分析树

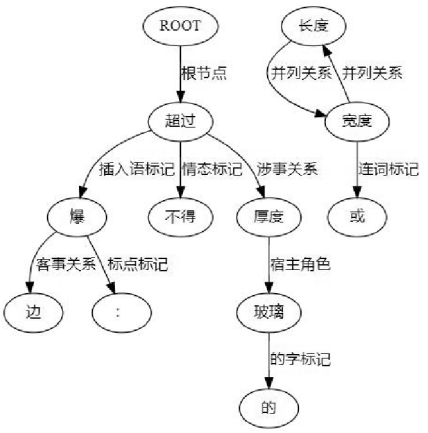

图 7 语义分析节点图

根据分析结果,该语句中关联词及关联方向分别为:"爆→边""宽度→长度""宽度→或""爆→超过""超过→宽度""超过→不得""玻璃→的""厚度→玻璃""超过→厚度"。句中各词间的依赖关系一目了然,但存在部分词语切分有误的情况,因此需要对文本进行修正。

3.3.2 分词情况

按照依存句法分析结果,确定各词语之间的依存关系(依存结构特征、依存词语特征等),采用正则表达式统一将文本进行标准化处理,包括标点符号、无用字的停用等。根据标准化的结果,选取每个类别下具有代表性的分词结果进行展示。通过使用 Viterbi、Perceptron 感知机、CRF 条件随机场、Re-Perceptron-CRF 进行分词,分别进行效果对比,并选择所有分词结果中具有代表性的分词效果与正确分词结果进行对比展示,具体结果见表 2 ~ 3。

表 2 分词效果对比 1

Algorithm	Text
Correct Results	点状缺陷
Viterbi	点状/缺陷
Perceptron	点状/缺陷
CRF	点状/缺陷
Re-Perceptron-CRF	点状缺陷

表 3 分词效果对比 2

Algorithm	Text1	Text2
Correct Results	不大于	膜厚度
Viterbi	不大/于	膜/厚度
Perceptron	不大于	膜/厚度
CRF	不/大于	膜厚度
Re-Perceptron-CRF	不大于	膜厚度

显而易见,Viterbi 分词的结果最多切分为二分词,误差极大;对于领域名词“爆边”而言,Viterbi 和 Perceptron 无法将其正确识别并切分,而 CRF 则能够正确识别;但对于“不大于”而言,Perceptron 比起 CRF 又能成功识别出来;Re-Perceptron-CRF 成功将 Perceptron 和 CRF 的分词结果综合,返回两者之中一方更为准确的结果,同时在正则表达式的修正下,也成功识别出部分组合名词和连接词。

3.3.3　歧义词切分

在所使用的规范类文本中,同样像大多数文本一样存在歧义词,如“膜厚度”可能划分为“膜厚/度”或“膜/厚度”,实则“膜厚度”这是一个整体。Re-Perceptron-CRF 会根据词语上下文中相关信息进行判断,通过特征权重对歧义词进行切分。实验文本中共计 21074 个词数,其中歧义词占到 8396 个。该文本下的歧义词数以及各算法下的歧义词切分情况见表 4。

表 4　歧义词处理结果

Algorithm	Correct Segmentation	Rete of Correct Segmentation/%
Viterbi	2632	31.35
Perceptron	5841	69.57
CRF	6030	71.82
Re-Perceptron-CRF	6912	82.32

3.3.4　分词速度对比

将 4 种方法的分词速度进行对比,结果见图 8。

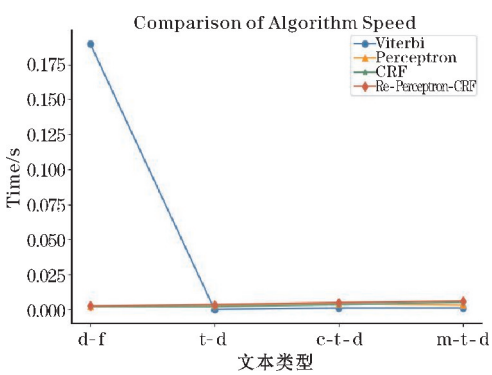

图 8　分词速度对比

从图 8 可知,Viterbi 算法的速度在对 d-f 类文本进行处理时的速度相对较慢,对其他类文本的处理速度基本和另外 3 种算法持平。此外,各算法速度虽然不具有显著差异,但是 Re-Perceptron-CRF 的分词速度明显略胜一筹。

3.4　算法评估

评价模型好坏的指标通常采用准确率 P、召回率 R 和 F_1 值来进行评估,此外再引入 R_{oov} 和 R_{iv} 进行对比。R_{oov} 和 R_{iv} 分别表示未登录词(out of vocabulary)的召回率和登录词(in vocabulary)的召回率。

上述指标的计算公式如下:

$$P=\frac{TP}{TP+FP}$$

$$R=\frac{TP}{TP+FN}$$

$$F_1=\frac{2PR}{P+R}$$

实验数据集共 645 份待分词样本,根据上述公式得出各算法指标对比见图 9 ~ 13。

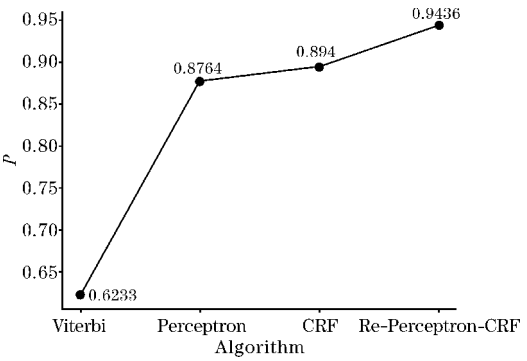

图 9　算法准确率对比

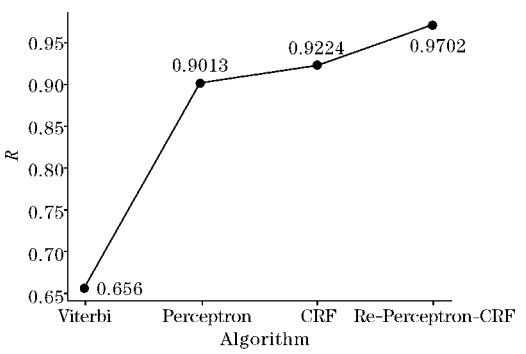

图 10　算法召回率对比

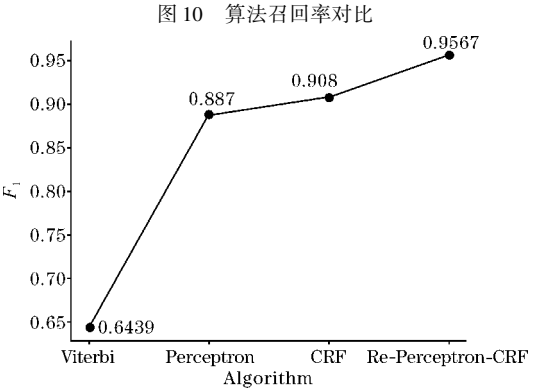

图 11　算法 F_1 值对比

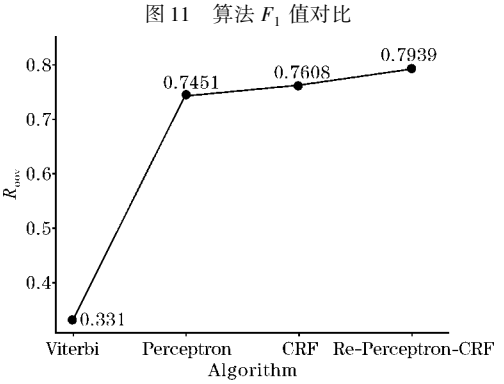

图 12　算法未登录词召回率对比

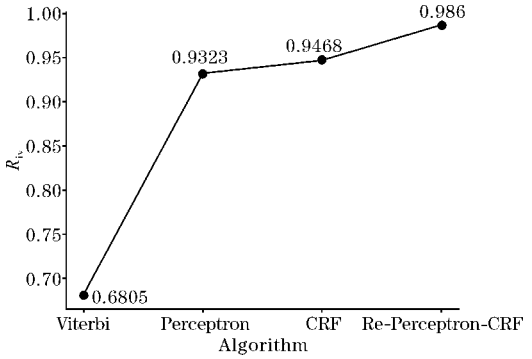

图 13 算法登录词召回率对比

根据表 2 ~ 3 的分词效果对比以及图 9 ~ 13 的评估结果对比可知:对于同一数据集,尽管 CRF 的整体 *P* 值和 *R* 值都相对较高,并且也正确识别出“膜厚度”一词,但对于“不大于”这类词并没有完全正确切割,相反 Perceptron 在该类词语的表现相对优越。并且,与 Viterbi 相比,Perceptron 与 CRF 在未登录词的召回率上表现大幅提升;同时,在引入正则表达式进行数据标准化匹配后,虽然部分整体词(如分格深度、分格总长度)仍然分成了两组词语,但“玻璃厚度”这类组合词语能够正确识别,整体结果比起传统算法有明显的提高。

4 总结与未来工作

通过使用 Viterbi、Perceptron、CRF 和 Re-Perceptron-CRF 4 种算法对规范类文本中语句进行分词,发现 Re-Perceptron-CRF 的准确率和召回率有明显提高,并能够有效识别领域内专有名词和部分组合词,同时分词速度也略微提升。但正则表达式匹配规则仅适用当前场景,故该方法针对的范围有限。其中,有一个现象引起注意:Viterbi 算法在处理 d-f 类文本时,所花费时间几乎是其他类文本的 120 多倍,而 d-f 类文本的文本数量其实是最少的。在后续的研究中,可能会重点关注这一问题。总而言之,后续研究方向将集中在以下 3 个方面:如何将该领域规范类文本正则匹配规则应用于更多规范类文本;如果提高组合词语和专有名词的划分精确度,提高消除歧义的准确度;能否在分词处理速度上做更多的优化。

参考文献:

[1] 许峰,张雪芬,忻展红. 基于深度神经网络模型的中文分词方案[J]. 哈尔滨工程大学学报,2019,40(9):1662-1666.

[2] Chomsky N. Syntactic Structures[M]. The Hague: Mouton de Gruyter,2002.

[3] 宗成庆. 统计自然语言处理[M]. 北京:清华大学出版社,2008:129-130.

[4] Li Hongqiao, Huang Chang-Ning. The use of SVM for Chinese new word identification[A]. In: Proceedings of the 1st International Joint Conference on Natural Language Processing (IJCNLP2004)[C]. Hainan Island,2004:723-732.

[5] 王厚峰,戴大为. 汉语句法结构标注的研究[J]. 计算机研究与发展,1997(3):77-82.

[6] 魏欧,吴健,孙玉芳. 基于统计的汉语词性标注方法的分析与改进[J]. 软件学报,2000(4):473-480.

[7] 杨尔弘,方莹,刘冬明,等. 汉语自动分词和词性标注评测[J]. 中文信息学报,2006(1):44-49.

[8] 奉国和,郑伟. 国内中文自动分词技术研究综述[J]. 图书情报工作,2011,55(2):41-45.

[9] BRILL E. A corpus-based approach to language learning[D]. Philadelphia: University of Pennsylvania,1993.

[10] 李华栋,贾真,尹红风,等. 基于规则的汉语兼类词标注方法[J]. 计算机应用,2014,34(8):2197-2201.

[11] Baum L E, Eagon J A. An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology[J]. Bulletin of the American Mathematical Society,1967(73):360-363.

[12] Lafferty J, McCallum A, Pereira F. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data[J]. In Proceedings of the18th International Conf on machine Learning,2001:282-289.

[13] RATNAPARKHI A. A maximum entropy model for part-of-speech tagging[C]. Proceedings of the 1996.

[14] 梁喜涛,顾磊. 中文分词与词性标注研究[J]. 计算机技术与发展,2015,25(2):175-180.

[15] 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(3):1-10.

[16] Lample G. Neural architectures for named entity recognition[C]. Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies(NAACL-HLT),2016:260-270.

[17] 刘伟,黄锴宇,余浩,等. 基于语境相似度的中文分词一致性检验研究[J]. 北京大学学报(自然科学版),2022,58(1):99-105.

[18] 凤丽洲,杨贵军,徐雪,等. 基于 N-gram 的双向匹配中文分词方法[J]. 数理统计与管理,

2020,39(4):633-643.
[19] Liu Junxin, Wu Fangzhao, Wu Chuhan, et al. Neural Chinese word segmentation with dictionary [J]. Neurocomputing,2019:338.
[20] Gan Leilei, Zhang Yue. Investigating Self-Attention Network for Chinese Word Segmentation[J]. CoRR,2019.
[21] Si Huihui,Ning Xin. Research and Implementation of Chinese Automatic Word Segmentation System Based on Complex Network Features[J]. Wireless Communications and Mobile Computing,2022.
[22] Yan Hang, Qiu Xipeng, Huang Xuanjing. A Graph-based Model for Joint Chinese Word Segmentation and Dependency Parsing[J]. Transactions of the Association for Computational Linguistics,2020:8.
[23] 徐飞,孙劲光. 中文分词切分技术研究[J]. 计算机工程与科学,2008(5):126-128.
[24] Jelinek F, Self-Organized Language Modeling for Speech Recogntion[J]. Reading in Speech Recognition. Morgan Kaufann Publishers ins 1990: 450-506.
[25] Bengio Y, Ducharme R, Vincent P. 3 (Feb): 2003:1137-1155.
[26] Forney GD Jr. The Viterbi algorithm[J]. Proceedings of the lEEE,1973,61(3):268-278.
[27] Rosenblatt F. The perceptron:Probabilistic model for information storage and organization in the brain. Psychological Review,1958,65(6):386-408.
[28] Mikolov T,Chen K,Corrado G,et al. Efficient Estimation of Word Representations in Vector Space [C]. IIICLR 2013,2013.
[29] Michalewicz Z. Genetic Algorithms + Data Structures evolution programs [M]. (3rd ed), New York:Springer-Verlag,1996.
[30] 刘克. 实用马尔可夫决策过程[M]. 北京:清华大学出版社,2004.
[31] ROBINSON J. Dependency structures and transformational rules [J]. Language, 1970, 46 (2): 259-285.
[32] Kleene, S C. Representation of Events in Nerve Nets and Finite Automata[M]. 1951.
[33] 邵艳秋,穗志方,韩纪庆,等. 基于依存句法分析的汉语韵律层级自动预测技术研究[J]. 中文信息学报,2008(2):116-123.
[34] 陈强,何炎祥,刘续乐,孙松涛,彭敏,李飞. 基于句法分析的跨语言情感分析[J]. 北京大学学报(自然科学版),2014,50(1):55-60.

Research on Word Segmentation of Normative Text based on Re-Perceptron-CRF

LI Baolin,　LIU Yutao
(College of Logistics,Chengdu University of Information Technology,Chengdu 610103,China)

Abstract:The Re-Perceptron-CRF combination method was used to segment key words by using the characteristics of specification documents. In this paper, four algorithms including Viterbi, Perceptron, CRF and Re-Perceptron-CRF are used to split the canonical text into words. Specificallyas follows: regular expressions are used to standardize the canonical text based on syntactic analysis, and the preprocessed text suitable for analysis is obtained. The optimal results are returned by the dual Perceptron and CRF algorithms. The experiment showed that the Re-Perceptron-CRF algorithm has good performance in the accuracy and recall rates of 94.36% and 97.02% respectively. This method provides some ideas for Chinese word segmentation related to standardized text, and provides good data support for subsequent applications. However, due to the small amount of data set, this method is only applicable to specific fields, such as building inspection.

Keywords:management science and engineering;textusl analysis;Chinese words segmentation;Re-Perceptron-CRF;part-of-speech tagging