山西大学学报(自然科学版)47(2):251—259,2024
Journal of Shanxi University(Nat. Sci. Ed.)

DOI:10.13451/j.sxu.ns.2023166

基于实例的词性标注数据错误检测

崔秀莲,严福康,李正华*
(苏州大学 计算机科学与技术学院,江苏 苏州 215000)

摘 要:由于深度学习框架在可解释性上的缺乏,本文将基于实例的方法首次应用到词性标注数据错误检测任务,旨在充分利用模型学到的实例之间的相似度信息。首先,本文基于预训练语言模型,实现了基于实例的词性标注模型,在CTB7数据集上的预测准确率和基于标准分类器的模型相当,达96.76%。进而,本文提出了一种基于实例的标注错误检测方法。为了获得真实检错数据集,本文采用不同方法对CTB7测试集进行自动错误检测,并人工标注候选错误,最终获得2 016个真实标注错误,约占所有8万多词中的2.5%。检错数据集上的实验表明,基于实例的方法的检错准确率达41.48%。

关键词:词性分类;标注错误数据集;语义相似度;CTB7数据集
中图分类号:TP391 **文献标志码**:A **文章编号**:0253-2395(2024)02-0251-09

Instance-Based Error Detection for Part-of-Speech Tagging Dataset

CUI Xiulian, YAN Fukang, LI Zhenghua*
(*School of Computer Science and Technology, Soochow University, Suzhou* 215000*, China*)

Abstract: Due to the lack of interpretability in deep learning frameworks, in this paper, we apply instance-based methods to error detection for part-of-speech tagging dataset for the first time aiming to leverage the similarity information learned between instances. Firstly, we implements an instance-based part-of-speech tagging model based on a pre-trained language model, achieving comparable prediction accuracy reaching 96.76% to models based on standard classifiers on the CTB7 dataset. Furthermore, we propose an instance-based annotation error detection method. To obtain an actual error detection dataset, several methods are employed to automatically detect errors in the CTB7 test set, and candidate errors are manually corrected, resulting in 2 016 annotation errors, accounting for approximately 2.5% of the total 80 000+ words. Experimental results on the error detection dataset show that the error detection accuracy of the instance based method reaches 41.48%.

Key words: part-of-speech classification; error detection dataset; semantic similarity; CTB7 dataset

0 引言

词性标注任务旨在为语料库中的词语根据其含义和上下文进行词类划分,是自然语言处理的基础步骤之一。词性标注的准确率会影响到下游任务的性能,如句法分析[1],信息抽取[2]等。

本文的研究着眼于有监督的中文词性标注,而带标签数据集是有监督学习中不可或缺的一部分。常用的中文词性标注数据集包括宾州树库[3]、BCC语料库[4]和人民日报语料库[5]等。宾州树库是一个广泛使用的数据集,涵盖了多种语言,包括英文和中文等。它的语料来

收稿日期:2023-08-30;**接受日期**:2023-10-23
基金项目:国家自然科学基金(62176173);江苏高校优势学科建设工程资助项目
作者简介:崔秀莲(1999-),女,浙江嘉兴人,硕士,研究方向为自然语言处理。E-mail:xiuliancc@foxmail.com
* **通信作者**:李正华(LI Zhenghua),E-mail:zhli13@suda.edu.cn
引文格式:崔秀莲,严福康,李正华.基于实例的词性标注数据错误检测[J].山西大学学报(自然科学版),2024,47(2):251-259. DOI:10.13451/j.sxu.ns.2023166

自报纸、新闻和广播等多种媒体,具有丰富的文本类型和语言风格。BCC语料库是由北京语言大学语料库中心构建的以汉语为主的语料库,涵盖了文学、科技、古汉语等多领域语料。人民日报标注语料库以《人民日报》语料为基础,从语料库应用和语言学角度完成了很多特殊标注。这些常用的数据集为词性标注算法的研究和改进提供了宝贵的资源。

相对英文文法来说,汉语言缺少形态变化,且本身存在词义易混淆性[6],标注规范很难对汉语言中所有的语言现象面面俱到。除此之外,大型标注语料库的构建通常需要大量标注人员,不同标注者对标注规范的理解可能存在不一致。这些原因共同导致数据集中存在一定数量的标注错误。很多研究也证实了这一点。Wang等的工作[7]显示命名实体识别数据集CoNLL03中约5%的句子存在标签错误。Marcus等构造了带标签的宾州树库[8],预估词性标签存在3%左右的错误。

对词性标注数据集进行标注错误检测,可以有效提高数据集质量,从而提升模型性能和预测结果的可信度。Dickinson等[9]将上下文相同但标注不同的实例视为标注有误的词语。Sachi等[10]集成了多个词性标注模型,来对词性标注数据集进行自动检错。为了更好地利用词语本身及词语之间相似度的信息,本文对该任务做了新的尝试,将基于实例的方法用于标注错误检测任务中。基于实例的方法指的是在对测试实例进行标签预测时对训练集的实例进行显式参考。随着神经网络的发展,由于模型可解释性的降低,Ouchi等重提基于实例的方法,在神经网络模型的基础上进行修改,运用于命名实体识别任务[11]和句法分析任务[12]上。

本文将基于实例的方法用于中文词性标注任务上,从而得到基于实例的词性标注模型。利用该模型对词性标注数据集进行错误检测。本文在中文宾州树库CTB7上的实验结果表明:1)基于实例的词性标注模型可以达到和基于标准分类器的词性标注模型相当的性能,在CTB7上的预测准确率达到96.76%;2)基于实例的标注错误检测模型可以有效检测出人工随机加噪数据集和真实数据集CTB7测试集中的标注错误;3)使用多个检错方法对CTB7测试集经过自动检错后,对一共8 509个候选错误词进行二次人工校对,获得了2 016个标注错误。标注错误的词语数约占整个测试集的2.5%。

1 相关工作

1.1 词性标注模型

词性标注是自然语言处理的基础步骤之一。使用机器学习方法进行词性标注的模型有隐马尔科夫模型[13]、最大熵模型[14]、条件随机场模型[15]等。这些模型大多依赖手工设计的特征模板。特征模板通常和目标词本身,目标词的上下文,目标词及可能的词性构成组合等有关,这导致了机器学习模型很难泛化。

自2012年AlexNet[16]在图象分类上取得了优异成绩以来,深度学习不断发展,自然语言处理相关任务的性能都有很大提升。Collobert等[17]提出了基于卷积神经网络(Convolutional Neural Network,CNN)卷积神经网络的框架来处理包括词性标注在内的多个序列标注任务。Huang等[18]提出了多个基于长短期记忆网络(Long Short-Term Memory,LSTM)的模型来处理词性标注、分块、命名实体识别任务。Shao等[19]将分词任务和词性标注任务联合建模,同样取得了优异的效果。

1.2 基于实例的方法

Aha等[20]在1991年提出了基于实例的学习并用于分类任务。他们对最近邻算法进行了扩展,根据训练集中特定的实例生成预测标签。

随着深度学习的发展,序列标注任务的性能有了显著提升,但基于神经网络的模型的可解释性较为一般,且对训练数据的需求较大,越来越多的研究采用基于实例的方法来解决这些问题。He等[21]为了缓解训练数据不足的问题,将基于实例的方法运用于股票走势预测任务上,有效提升了性能。在知识库推理任务上,为了降低搜索空间的大小,Cui等[22]通过和训练实例进行比较来生成预测结果。

Ouchi等针对命名实体识别任务提出了可以学习span之间的相似度的模型[11]。根据训练集实例之间的相似度对测试集实例的标签进行预测,从而理解训练集实例对预测的贡献,

得到一个具有高度可解释性的模型。在句法分析任务上,Ouchi等提出了基于实例的模型[12],将依赖边和训练集中的边的相似度进行计算,从而对依赖边进行识别和分类。这两项工作的实验结果均显示,基于实例的模型的预测精度可以与标准神经网络模型相比拟,同时具有高度的可解释性,可以很容易地分析出训练点对标签预测的贡献。

1.3 标注错误检测

对于有监督的任务而言,噪声标签的存在对模型性能有很大影响[23]。减少数据集中的标注错误,不仅可以提升模型训练的性能,也可以提升模型性能评估的可信度。

在标注错误检测领域,由于手动清洗完整数据集需要消耗大量的人力和物力资源,因此通常采用先由模型筛选标注错误,再人工对少量数据进行校对的方法[24]。

很多学者设计了自动检测标注错误的模型。Hendrycks等[25]证实,正确分类的实例的预测概率会高于错误分类的实例,因此可以根据预测概率设定阈值检测错误分类的实例。Northcutt等[26]提出置信学习的框架,通过估计噪声标签和真实标签的联合分布,学习标签置信度的阈值,并根据一些规则对可能存在标签错误的实例进行筛选。Gong等[27]提取数据特征后,在检测模型中加入dropout噪声以得到随机的输出,由此来评估标签的不确定性。Larson等[28]先计算同类词语在向量空间中的中心位置,再根据词语离该中心位置的距离对标签的置信度进行评估。Sachi等[10]集成了隐马尔科夫模型、支持向量机、条件随机场、LSTM网络等多种不同的标注模型,以对词性标签进行自动纠错和验证。

2 基于实例的词性标注模型

2.1 任务描述

词性标注是一个多分类任务,目标是为给定词语序列中的每一个词语分配对应的词性标签。

词性标注任务的数学形式定义如下:在数据集D的范围内,给定一个输入词语序列$X=(w_0,w_1,\cdots,w_n)$,词性标注任务的目标是确定对应的词性序列$Y=(t_0,t_1,\cdots,t_n)$作为输出。其中,$t_i\in T$,T是所有词性标签的集合。

例如,句子“会谈 取得 了 进展 。”的正确词性标注为“会谈/NN 取得/VV 了/AS 进展/NN 。/PU”,见图1。NN是一般名词,VV是一般动词。其余词性的详细定义及示例可以参考标注规范[29]。

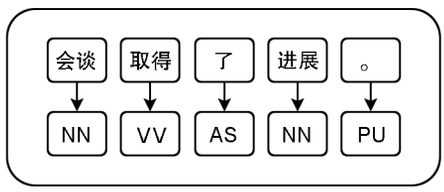

图1 词性标注举例

Fig. 1 An example of part-of-speech tagging task

2.2 模型介绍

本节将首先介绍基于标准分类器的词性标注模型,接着介绍基于实例的词性标注模型,最后将这两种不同的词性标注模型运用于CTB7数据集进行标注性能评估。

2.2.1 基于标准分类器的模型

基于标准分类器的模型由编码层和解码层两个部分组成,模型结构如图2所示。

编码层部分,首先对分好词的词序列进行索引化和批次化(batch)。将成batch的索引序列送入编码器,得到高维词表示向量,接着将词向量送入解码层进行多分类的预测。本文使用Bert Base Chinese作为编码器,输出维度是768。

解码层部分,使用多层感知机(Multi-Layer Perceptron,MLP),再使用Softmax归一得到不同词性的预测概率,从而确定概率最高的词性为预测词性。其中,MLP有三层,初始输入维度和高维词向量维度一致,为768,隐层维度设置为400,输出层维度和数据集词性标签总数一致。

2.2.2 基于实例的模型

基于实例的模型和基于标准分类器的模型的主要区别在于解码层部分。基于实例的模型会参考实例之间的相似度进行打分,详见图3。

首先构造参考实例集D,基于实例的标注模型在预测词语w_i的词性时,会根据w_i和实例集D中所有标注为词性t_i的词语之间的平均相

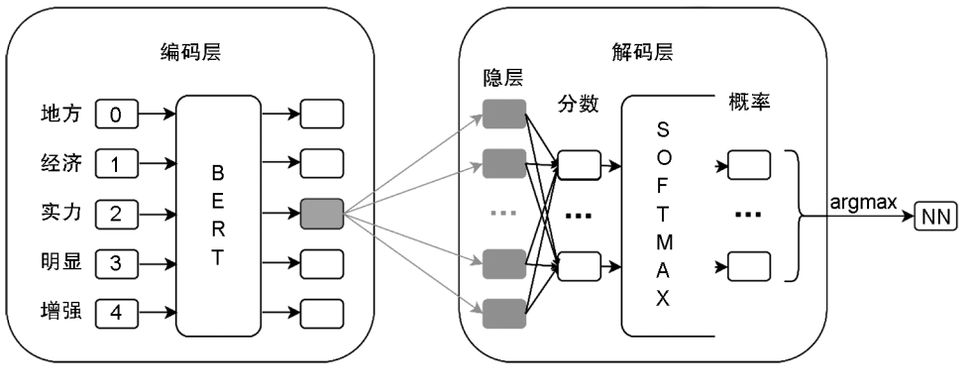

图2 基于标准分类器的词性标注模型结构图

Fig. 2 Structure diagram of part-of-speech annotation model based on standard classifier

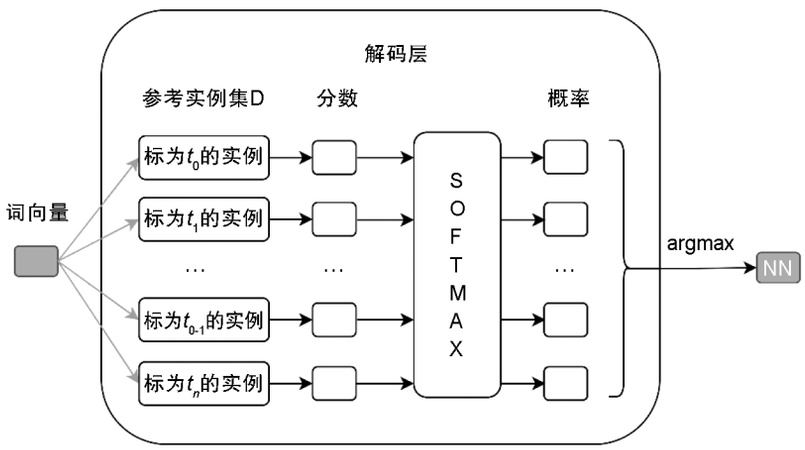

图3 基于实例的词性标注模型的解码层结构图

Fig. 3 Structure diagram of an decoder for instance based part-of-speech annotation model

似度进行打分。分数s的详细计算方式见公式(1)。

$$s(t_i|w_i)=\frac{\sum_{w_j\in W(D,t_i)}\mathrm{sim}(w_i,w_j)}{C(W(\mathrm{D},t_i))},\tag{1}$$

其中,$W(\mathrm{D},t_i)$是从实例集D中抽取的所有词性为t_i的词语的集合。$C(W(\mathrm{D},t_i))$是对前述集合中的元素个数进行统计。$\mathrm{sim}(w_i,w_j)$表示词语w_i和词语w_j的相似度。

参考Ouchi等[11]的工作,本文采用点积和余弦值两种方式对实例之间的相似度进行描述。点积相似度的计算方式见公式(2)。$\boldsymbol{h}_{w_i}$是词语w_i经过编码层后得到的词向量。

$$\mathrm{sim}(w_i,w_j)=\boldsymbol{h}_{w_i}\cdot\boldsymbol{h}_{w_j},\tag{2}$$

余弦相似度的计算方式见公式(3)。由于余弦相似度在计算时会对向量进行标准化,因此使用放大参数α对向量之间的余弦值进行放大,有利于相似实例的聚集,不相似实例的分散。

$$\mathrm{sim}(w_i,w_j)=\alpha\cdot\frac{\boldsymbol{h}_{w_i}\cdot\boldsymbol{h}_{w_j}}{|\boldsymbol{h}_{w_i}|\cdot|\boldsymbol{h}_{w_j}|},\tag{3}$$

得到词语w_i被标注为各个词性t_i的得分后,根据Softmax归一得到词语w_i被标注为不同词性的概率,计算方式见公式(4)。最后确定概率最高的词性为词语w_i的预测词性。

$$P(t_i|w_i)=\frac{\exp(s(t_i|w_i))}{\sum_{t_i\in\mathrm{T}}\exp(s(t_i|w_i))},\tag{4}$$

其中,T是数据集中所有词性标签的集合。

2.3 实验

本文使用中文宾州树库CTB7数据集对词性标注模型的性能进行评估。词类的划分标准详见中文宾州树库词性标注规范[29]。数据集的划分及规模见表1。CTB7的训练集涉及35种词性标签,相比标注规范中的33种多了URL、X标签。URL指的是网址,X标签指的是一些特殊符号,如数学中的乘号×。Bert会对

这35种词性进行编码。虽然验证集和测试集本身不包含这些词性的实例,但模型依然会给出标注为这些词性的概率。

表1 CTB7数据集的划分及规模

Table 1 Partition and scale of CTB7 dataset

数据集	句子数	词语数	词性标签数
训练集	46 572	1 057 943	35
验证集	2 079	59 955	32
测试集	2 697	81 232	33

训练集共有46 572个句子,1 057 943个词语。对于基于实例的标注模型,在每一个batch的句子送入解码层后,根据公式(1),首先会随机抽取训练集中的50个句子构造实例集D。也就是说,同一个batch中的实例训练时共享相同的50个参考句子。接着进行实例间的相似度计算对词性进行预测。本文在训练模型时使用交叉熵作为损失函数。

在实际构造参考实例集的时候,首先针对每一种词性,随机选取一句含有该词性词语的句子,接着从整个训练集中随机抽取剩余的句子,以确保参考实例集中每种词性的实例均存在。

将基于标准分类器的模型、基于实例的模型均运用于CTB7数据集完成词性标注任务,模型性能详见表2。由于基于实例的模型在抽取参考句子时采用的是随机算法,因此使用不同的随机数种子0、666、999进行了三次实验取平均值。

根据表2的数据,无论是在CTB7验证集还是测试集上,基于实例的标注模型均优于基于标准分类器的模型。在验证集上,使用点积的基于实例的标注模型的性能最为稳定且良好。在测试集上,使用余弦值的基于实例的标注模型的性能最优。

2.4 参数设置

参考实例集的句子个数可以调整。为了确定参考实例集的大小,分别使用50个句子、75个句子、100个句子进行实验,模型在测试集上的标注准确率及运行时间见表3。扩大参考实例集的大小对模型的性能影响较小,且会大幅增加训练时间,因此选择50句作为参考实例集的大小。

表2 不同词性标注模型的预测准确率

Table 2 Prediction accuracy of different part-of-speech tagging models

模型	验证集	测试集
基于标准分类器	96.74	96.71
基于实例(点积)	**96.80±0.02**	96.71±0.03
基于实例(余弦值)	96.80±0.1	**96.76±0.04**

表3 基于实例的词性标注模型的预测准确率及运行时长

Table 3 Prediction accuracy and runtime of an instance based part-of-speech tagging model

参考实例集大小	基于实例(点积)		基于实例(余弦值)	
	准确率/%	运行时长	准确率/%	运行时长
50	96.71	34 h 49 m	96.76	35 h 46 m
75	96.70	53 h 8 m	96.78	53 h 35 m
100	96.69	69 h 26 m	96.77	69 h 38 m

3 词性标注错误检测

3.1 错误检测方法

3.1.1 基于实例的错误检测

基于实例的词性标注模型在训练的过程中充分学习了实例之间的相似度。如果一个词语的词性和其相似词的词性相去甚远的话,那么该词语的词性存在问题的概率较高。因此,本文尝试在基于实例的标注模型的基础上进行数据错误检测,主要步骤如下:

Step 1 针对要进行标注错误检测的数据集,基于第2章中训练得到的基于实例的词性标注模型,获取到编码器输出的每个词语的高维词向量。

Step 2 根据模型的不同以不同的方式获取到与每个词语最相似的100个词语。使用点积的基于实例的词性标注模型,从训练集中根据词向量之间的点积值找寻相似词。使用余弦值的模型则根据余弦值找寻相似词。

Step 3 根据AP(Average Precision)计算相似词和数据集原始词性的匹配程度,匹配程度越低排名越靠前,越有可能对应标注错误。AP定义为召回率从0到1时的平均精确率。具体计算方式为:记某一词语w在数据集中的对应词性为t^*,数据集中与词语w相似度第i高的词语的对应词性为t_i。

令$S_K=\{i|t_i==t^*,i\in[1,K]\}$,存储相似词中所有词性标注与$t^*$一致的词语的排名。$K$是

相似词的总数,设置为100。

由此可计算词语w的相似词和数据集原始词性的匹配度得分为:

$$s_w=\frac{1}{|S_K|}\sum_{i\in S_K}\frac{|S_i|}{i}, \tag{5}$$

其中,$|S_i|$表示集合S_i的元素个数。

Step 4 将数据集中的所有词语按照匹配度得分从低到高进行排序,按照一定比例筛选出排名靠前的词语作为候选错误。

3.1.2 对比方法

为了更好地对基于实例的词性标注模型的性能进行评估,本文实现了两个较为普遍使用的检错方法作为对比。

基于标准分类器置信度

参考Hendrycks等[25]的工作,词语被标注为正确词性的概率比错误词性高。那么,模型给出数据集原始词性的预测概率越低,该词性数据越有可能是错误。本文采用2.2.1节中的基于标准分类器的模型来实现这个方法。

基于词向量平均距离

参考Larson等[28]的工作,根据同一词性的词语在向量空间中的距离来表示置信度。首先对词语经过编码器后得到的高维词向量进行存储。接着根据它们在数据集中对应的不同词性,计算同一词性的词语对应的词向量的平均值,作为质心。每一种词性会对应不同的质心。最后,按照每个词到对应质心的欧氏距离进行从远到近的排序,越远的越有可能是标注错误。本文使用2.2.1节中基于标准分类器的模型的编码层部分输出的词向量来实现这个方法。

3.2 随机加噪数据集的错误检测

为了验证基于实例的检错方法的有效性,本节将对随机加噪的数据集进行错误检测。首先使用随机加噪的方式构造数据集,随机抽取CTB7测试集中约10%的数据词性标签进行替换,生成新测试集。在实验中,实际替换了8 091个词语的词性标签。按照公式(1)的打分对新测试集中所有的词语进行排序。

使用基于标准分类器置信度的方法,基于词向量平均距离的方法,使用点积或余弦值描述实例相似度的基于实例的检错方法,对随机加噪的新测试集进行自动错误检测。参考Marcus等[8]给出的词性标签错误率3%,为了更大可能地找出错误,选择比测试集3%的词语数略大的阈值4 000。从各模型给出的错误可能性排序中筛选前4 000个词语进行检错准确率P的计算(见公式(6))。各模型的检错性能详见表4。其中,Count行表示排序在前4 000名的词语对应的标注错误的个数,P行表示对应的检错准确率。

$$P=\frac{C(\text{错误标注的词语})}{C(\text{模型认为有错的词语})}。 \tag{6}$$

表4 各模型标注错误检测的性能(随机加噪的数据集)

Table 4 Performance of annotation error detection for each method (randomly perturbed dataset)

检错指标	标准分类器	平均距离	基于实例	
			点积	余弦值
Count	**3 980**	3 413	3 584	3 516
P	**99.50**	85.33	89.60	87.90

在随机加噪的数据集上,基于标准分类器置信度的方法的性能较大幅度地优于别的方法。因为加噪过程中,混淆标签的选择是随机的,而该方法直接根据数据集词性的预测概率进行排序,很容易检测出这类较容易发现的错误。而在同样基于词向量的方法中,基于实例的检错方法会优于平均距离的方法。

对这4个模型检错结果的分段评估详见图4,横坐标表示检错结果的排序,纵坐标表示每五百个词检测出来的错误个数。随着排序从高到低,基于实例的检错方法检测出的错误个数呈现下降趋势,可以很好地证明本检错方法排序方式的合理性。

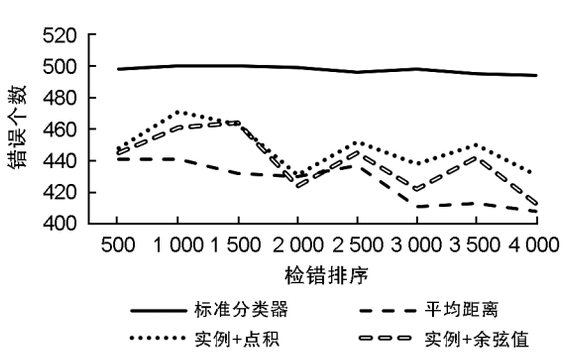

图4 各模型的分段检错个数(人工加噪的数据集)

Fig. 4 Number of segmented error detections for each method (randomly perturbed dataset)

3.3 真实数据集的错误检测

数据集规模较大,一般不会人工对完整数据集进行二次校对。由检错模型自动筛选出更有可能是标注错误的词语再进行少量的人工校对是当前一种较为普遍的做法。本节将使用基于实例的检错方法对CTB7测试集进行错误检测。首先由模型对测试集中的词语的错误可能性进行排序,接着筛选出一定数量的词语进行人工二次校对,最后确定出数据集中真正有误的词语,完成错误检测工作。

3.3.1 参考实例推荐

由于汉语中存在一词多义,词性活用等语言现象,词性标注也存在一些易混淆点。因此,标注规范中有很多特殊规定。例如,一个常用作动词的词语,一般情况下应被标注为VV。但如果该词在句子中作为名词的直接修饰语,即该动词和名词之间没有用“的”连接时,它应该被标注为JJ。“驰名”常做动词,如“驰名/VV 中外/NN”。但在“中国/NR 驰名/JJ 商标/NN”中,“驰名”直接修饰名词“商标”,则应被标注为JJ。

因此,为了二次校对过程中降低标注人员的标注难度,使用基于实例的词性标注模型根据一定方法检索出训练集中的相关样例作为标注参考。

首先,基于第2章中训练得到的使用余弦值的基于实例的词性标注模型,获取到词语经过编码器后生成的高维词向量。接着,计算余弦相似度从训练集中找到与目标词最相似的若干词语,相似词所在的句子便是所求的参考样例。

以CTB7测试集中的一个句子“由于/P 退休/NN 之后/LC 生活/NN 的/DEG 闲暇/NN 较/AD 多/VA”为例,“退休”在数据集中的标注为NN。根据标注规范[29],“退休”在此处是动作,应该作动词,标注为VV,但在数据集中却错误地标注为NN。

根据前述方法找到的参考实例有:(1)“即便 是 退休/VV 以后”;(2)“特别是 在 其 退休/VV 以后”;(3)“嫌犯 25岁 退伍/VV 回来后”;(4)“担任 会长 达 48年的沙理士 宣布 退休/NN”等。其中,大部分是“退休”“退伍”所在的句子。像“退休/VV 之后”,“退伍/VV 之后”和原句结构类似,可以很好地提供参考。同时还有“退休”作名词的例子,“沙理士/NR宣布/VV 退休/NN”。此处“退休”与原句中“退休”所在的短语结构不同,因此对应的词性也不同,可以给标注人员提供更好的对比和参考,有效提升标注人员的标注准确率。

在Dickinson等[9]的工作中,针对核心词所在窗口在数据集中找寻相同上下文但核心词标注不同的n-gram与本方法在结果上有一定相似度。但本方法找出的参考实例并不局限于相同的核心词,例如,针对“退休”找到的参考实例中就存在“退伍”这样的相似词。同时,本方法找出的参考实例会更多地倾向于正确词性,而不仅仅是找出和核心词原词性不同的实例。

3.3.2 人工校对

将基于标准分类器置信度的方法,基于词向量的平均距离方法,使用点积的基于实例的检错,使用余弦值的基于实例的检错这四种方法分别用于CTB7测试集,筛选出排序靠前的4 000个词语。

在这四个模型的检测结果中选取排序在前4 000名的词语,对其取并集,去重后得到8 509个词语。对这些词语进行分组,300个词为一组,共29组。将这些可能有误的词分批次交给两名相关专业的研究生在理解标注规范之后进行人工校对。分批次可以有效防止工作量过大、专注度降低导致的疏忽。

两人分别对检错模型的检测结果进行标注。若两人的标注一致,且与数据集词性不一致的话,认为数据集原始词性有误。若两人的标注一致,且与数据集词性一致的话,认为数据集原始词性无误。若两人的标注不一致的话,在讨论后确定最终标注,再与数据集词性进行比较。

3.3.3 检错实验分析

经过二次校对后,标注人员一共在CTB7测试集中找出2 016个错误标注的词语。例如“新华社/NN 福州三月四日电(记者许霆)”中的“新华社”就被误标成了NN,正确词性应该是专有名词NR。该数据集一共有2 697个句子,81 232个词语,检测出的错误数占比在2.5%左右。

得到校对后的数据集后,检错模型对真实数据集的错误检测能力也可以得到正确的评估。各模型检测出标注错误的正确率详见表5。

表5 各模型标注错误检测的性能(真实数据集)

Table 5 Performance of annotation error detection for each method (real dataset)

检错指标	标准分类器	平均距离	基于实例	
			点积	余弦值
Count	1 539	748	1 461	**1 659**
P	38.48	18.7	36.53	**41.48**

表5的数据显示,使用余弦值的检错模型的性能优于其他模型。词性标注实验(见表2)中,使用点积和使用余弦值的标注模型的性能是相当的。但在检错实验中,使用余弦值的检错模型的性能优于使用点积的模型,这说明,使用余弦值描述实例之间的相似度会比使用点积更为合理。

对这4种检错方法在CTB7测试集上检错结果的分段评估详见图5。在前1 000个词中,基于标准分类器置信度的方法检错性能相对基于实例的检错方法更为优秀,但后期,基于实例的检错方法更为优秀。这说明,基于标准分类器的检错方法更擅长检测出较容易发现的错误,而基于实例的方法对容易混淆较难发现的错误更为擅长。这也是3.2节中,基于实例的检错方法的性能在随机加噪的数据集上不如基于标准分类器置信度方法的原因。

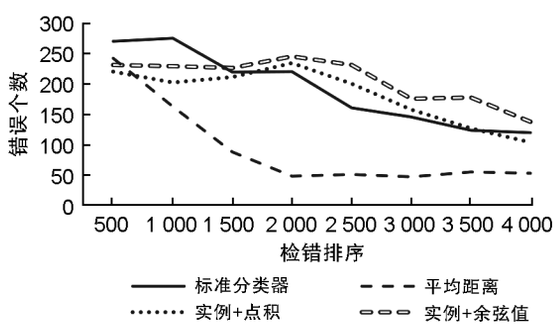

图5 各模型的分段检错个数(真实数据集)

Fig. 5 Number of segmented error detections for each method (real dataset)

4 结论

本文将基于实例的方法用于词性标注数据错误检测任务,有效利用了实例之间的相似度信息。本文将该方法在CTB7的测试集上进行了实验,对候选错误词进行了人工二次校对,找出了数据集中存在的约2.5%的标注错误。基于实例的检错方法呈现出了良好的性能,且对于检测数据集中较难发现的错误比较有优势。本文方法强依赖于实例之间的相似度,而汉语词性活用现象的多发可能会对此产生不良影响,未来会对此进行深入研究。基于实例的检错方法目前仅针对于词性标注任务,有一定的局限性,未来可以将其扩展到别的自然语言处理任务中。

参考文献:

[1] ZHOU H Q, ZHANG Y, LI Z H, *et al*. Is POS Tagging Necessary or even Helpful for Neural Dependency Parsing? [M]//Natural Language Processing and Chinese Computing. Cham: Springer International Publishing, 2020: 179−191. DOI: 10.1007/978-3-030-60450-9_15.

[2] NGUYEN D Q, VERSPOOR K. From POS Tagging to Dependency Parsing for Biomedical Event Extraction[J]. *BMC Bioinformatics*, 2019, **20**(1): 72. DOI: 10.1186/s12859-019-2604-0.

[3] XUE N W, XIA F, CHIOU F D, *et al*. The Penn Chinese TreeBank: Phrase Structure Annotation of a Large Corpus [J]. *Nat Lang Eng*, 2005, **11**(2): 207−238. DOI: 10.1017/S135132490400364X.

[4] 荀恩东, 饶高琦, 肖晓悦, 等. 大数据背景下BCC语料库的研制[J]. 语料库语言学, 2016, **3**(1): 93−109.
XUN E D, RAO G Q, XIAO X Y, *et al*. The Construction of the BCC Corpus in the Age of Big Data[J]. *Corpus Linguist*, 2016, **3**(1): 93−109.

[5] 黄水清, 王东波. 新时代人民日报分词语料库构建、性能及应用(一): 语料库构建及测评[J]. 图书情报工作, 2019, **63**(22): 5−12. DOI: 10.13266/j.issn.0252-3116.2019.22.001.
HUANG S Q, WANG D B. Construction, Performance and Application of New Era People's Daily Segmented Corpus (I)-Construction and Evaluation of Corpus[J]. *Libr Inf Serv*, 2019, **63**(22): 5−12. DOI: 10.13266/j.issn.0252-3116.2019.22.001.

[6] PLANK B, HOVY D, SØGAARD A. Linguistically Debatable or Just Plain Wrong?[C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Stroudsburg, PA, USA: Association for Computational Linguistics, 2014: 507−511. DOI: 10.3115/v1/p14-2083.

[7] WANG Z H, SHANG J B, LIU L Y, *et al*. CrossWeigh: Training Named Entity Tagger from Imperfect Annotations [C]//Proceedings of the 2019 Conference on Empirical

Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Stroudsburg, PA, USA: Association for Computational Linguistics, 2019: 5154-5163. DOI: 10.18653/v1/d19-1519.

[8] MARCUS M P, SANTORINI B, MARCINKIEWICZ M A. Building a Large Annotated Corpus of English: The Penn Treebank[J]. *Comput Linguist*, 1993, **19**: 313-330. DOI: 10.21236/ada273556.

[9] DICKINSON M, DETMAR MEURERS W. Detecting Errors in Part-of-speech Annotation[C]//Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-EACL '03. Morristown, NJ, USA: Association for Computational Linguistics, 2003: 107-114. DOI: 10.3115/1067807.1067823.

[10] SACHI A, MISHRA P, SHARMA D M. Automated Error Correction and Validation for POS Tagging of Hindi [C]//Proceedings of the thirty-second Pacific Asia Conference on Language, Information and Computation, 2018:11-18.

[11] OUCHI H, SUZUKI J, KOBAYASHI S, *et al*. Instance-based Learning of Span Representations: a Case Study through Named Entity Recognition[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 2020: 6452-6459. DOI: 10.18653/v1/2020.acl-main.575.

[12] OUCHI H, SUZUKI J, KOBAYASHI S, *et al*. Instance-based Neural Dependency Parsing[J]. *Trans Assoc Comput Linguist*, 2021, **9**: 1493-1507. DOI:10.1162/tacl_a_00439.

[13] KUPIEC J. Robust Part-of-speech Tagging Using a Hidden Markov Model[J]. *Comput Speech Lang*, 1992, **6**(3): 225-242. DOI: 10.1016/0885-2308(92)90019-Z.

[14] MCCALLUM A, FREITAG D, PEREIRA F C N. Maximum Entropy Markov Models for Information Extraction and Segmentation[C]//Proceedings of the Seventeenth International Conference on Machine Learning. New York: ACM, 2000: 591-598. DOI: 10.5555/645529.658277.

[15] LAFFERTY J, MCCALLUM A, PEREIRA F. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data[C]//Proceedings of the eighteenth International Conference on Machine Learning, 2001: 282-289. DOI:10.5555/645530.655813.

[16] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet Classification with Deep Convolutional Neural Networks[J]. *Commun ACM*, 2017, **60**(6): 84-90. DOI: 10.1145/3065386.

[17] COLLOBERT R, WESTON J, BOTTOU L, *et al*. Natural Language Processing (almost) from Scratch[J]. *J Mach Learn Res*, 2011, **12**: 2493-2537.

[18] HUANG Z H, XU W, YU K. Bidirectional LSTM-CRF Models for Sequence Tagging[J]. arXiv Preprint: 1508.01991, 2015. DOI:10.48550/arXiv.1508.01991.

[19] SHAO Y, HARDMEIER C, TIEDEMANN J, *et al*. Character-based Joint Segmentation and POS Tagging for Chinese Using Bidirectional RNN-CRF[J]. arXiv Preprint: 1704.01314, 2017. DOI:10.48550/arXiv.1704.01314.

[20] AHA D W, KIBLER D, ALBERT M K. Instance-based Learning Algorithms[J]. *Mach Learn*, 1991, **6**(1): 37-66. DOI: 10.1007/BF00153759.

[21] HE Q Q, SIU S W I, SI Y W. Instance-based Deep Transfer Learning with Attention for Stock Movement Prediction [J]. *Appl Intell*, 2023, **53**(6): 6887-6908. DOI: 10.1007/s10489-022-03755-2.

[22] CUI W Y, CHEN X R. Instance-based Learning for Knowledge Base Completion[C]//Conference in Neural Information Processing Systems, New Orleans: NeurIPS, 2022:30744-30755. DOI: 10.48550/arXiv.2211.06807.

[23] SONG H, KIM M, PARK D, *et al*. Learning from Noisy Labels with Deep Neural Networks: a Survey[J]. *IEEE Trans Neural Netw Learn Syst*, 2023, **34**(11): 8135-8153. DOI: 10.1109/TNNLS.2022.3152527.

[24] KLIE J C, WEBBER B, GUREVYCH I. Annotation Error Detection: Analyzing the Past and Present for a more Coherent Future[J]. *Comput Linguist*, 2023, **49**(1): 157-198. DOI: 10.1162/coli_a_00464.

[25] HENDRYCKS D, GIMPEL K. A Baseline for Detecting Misclassified and Out-of-distribution Examples in Neural Networks[J]. arXiv Preprint: 1610.0213, 2016. DOI: 10.48550/arXiv.1610.0213

[26] NORTHCUTT C, JIANG L, CHUANG I. Confident Learning: Estimating Uncertainty in Dataset Labels[J]. *Jair*, 2021, **70**: 1373-1411. DOI: 10.1613/jair.1.12125.

[27] GONG X Y, LU J L, ZHOU Y F, *et al*. Model Uncertainty Based Annotation Error Fixing for Web Attack Detection [J].*J Signal Process Syst*, 2021, **93**(2/3): 187-199. DOI: 10.1007/s11265-019-01494-1.

[28] LARSON S, MAHENDRAN A, LEE A, *et al*. Outlier Detection for Improved Data Quality and Diversity in Dialog Systems[C]//Proceedings of the 2019 Conference of the North. Stroudsburg, PA, USA: Association for Computational Linguistics, 2019: 517-527. DOI: 10.18653/v1/n19-1051.

[29] XIA F. The Part-of-speech Tagging Guidelines for the Penn Chinese Treebank (3.0)[C]//IRCS Technical Reports Series, Philadelphia: University of Pennsylvania, 2000:1-46.