·开发研究与工程应用·　　文章编号：1000-3428(2022)06-0314-07　　文献标志码：A　　中图分类号：TP391

融合数据增强与半监督学习的药物不良反应检测

佘朝阳[1,2]，严　馨[1,2]，徐广义[3]，陈　玮[1,2]，邓忠莹[1]

（1.昆明理工大学 信息工程与自动化学院，昆明 650504；2.昆明理工大学 云南省人工智能重点实验室，昆明 650504；
3.云南南天电子信息产业股份有限公司，昆明 650041）

摘　要：目前药物不良反应（ADR）研究使用的数据主要来源于英文语料，较少选用存在标注数据稀缺问题的中文医疗社交媒体数据集，导致对中文医疗社交媒体的研究有限。为解决标注数据稀缺的问题，提出一种新型的ADR检测方法。采用ERNIE预训练模型获取文本的词向量，利用BiLSTM模型和注意力机制学习文本的向量表示，并通过全连接层和softmax函数得到文本的分类标签。对未标注数据进行文本增强，使用分类模型获取低熵标签，此标签被作为原始未标注样本及其增强样本的伪标签。此外，将带有伪标签的数据与人工标注数据进行混合，在分类模型的编码层和分类层间加入Mixup层，并在文本向量空间中使用Mixup增强方法插值混合样本，从而扩增样本数量。通过将数据增强和半监督学习相结合，充分利用标注数据与未标注数据，实现ADR的检测。实验结果表明，该方法无需大量的标注数据，缓解了标注数据不足对检测结果的影响，有效提升了药物不良反应检测模型的性能。

关键词：医疗社交媒体；药物不良反应；数据增强；半监督学习；预训练语言模型

开放科学（资源服务）标志码（OSID）：

中文引用格式：佘朝阳，严馨，徐广义，等.融合数据增强与半监督学习的药物不良反应检测[J].计算机工程，2022，48(6)：314-320.

英文引用格式：SHE Z Y，Y X，XU G Y，et al.Adverse drug reaction detection combined with data augmentation and semi-supervised learning[J].Computer Engineering，2022，48(6)：314-320.

Adverse Drug Reaction Detection Combined with Data Augmentation and Semi-supervised Learning

SHE Zhaoyang[1,2]，YAN Xin[1,2]，XU Guangyi[3]，CHEN Wei[1,2]，DENG Zhongying[1]

（1.Faculty of Information Engineering and Automation，Kunming University of Science and Technology，Kunming 650504，China；
2.Yunnan Key Laboratory of Artificial Intelligence，Kunming University of Science and Technology，Kunming 650504，China；
3.Yunnan Nantian Electronic Information Industry Co.，Ltd.，Kunming 650041，China）

【Abstract】At present，the data used in the study of Adverse Drug Reaction（ADR）are mainly from English corpus，fewer Chinese medical social media data sets are selected because of label data scarcity，resulting in limited research on Chinese medical social media.To deal with the problem of lack of labeled data，this study proposes an ADR detection method that combines data augmentation and semi-supervised learning.The pre-training ERNIE model is used to obtain the word vectors. BiLSTM and the attention mechanism are used to learn the vector representation of the text. The classification layer consists of a fully connected layer and a softmax function to obtain the classification label.First，the unlabeled data are augmented several times.The low-entropy label，which is the weighted average of the predicted values of the original and augmented samples，is shared by these samples.The pseudo-label data are then mixed with the labeled data.Based on the classification model，a Mixup layer is added between the encoding and classification layers.In the text vector space，Mixup is used to interpolate the mixed samples，and the number of samples will be higher.By combining data augmentation and semi-supervised learning，labeled and unlabeled data are fully utilized to detect adverse drug reactions. Experimental results show that this method does not require a large amount of labeled data，alleviates the impact of insufficient labeled data，and effectively improves the performance.

【Key words】medical social media；Adverse Drug Reaction（ADR）；data augmentation；semi-supervised learning；pre-training language model

DOI：10. 19678/j. issn. 1000-3428. 0062170

基金项目：国家自然科学基金（61462055，61562049）。

作者简介：佘朝阳（1996—），女，硕士研究生，主研方向为自然语言处理；严　馨，副教授、硕士；徐广义，高级工程师、硕士；陈　玮、邓忠莹，讲师、硕士。

收稿日期：2021-07-22　　**修回日期：**2021-09-07　**E-mail：**kg_yanxin@sina.com

0　概述

药物不良反应（Adverse Drug Reaction，ADR）检测是药物研究和开发的重要组成部分。以往的研究数据主要来源于药物不良反应报告[1]、生物医学文献[2]、临床笔记[3]或医疗记录[4]。目前，医疗社交媒体例如MedHelp、好大夫、寻医问药、丁香医生等均提供了诸如专家问诊、论坛讨论、电话视频交流等形式多样的信息收集手段，形成了具有权威性、时效性和全面性的互联网医疗数据源，为药物不良反应检测提供了丰富的语料基础。

药物不良反应检测通常被看作涉及ADR的文本二分类问题，即辨别文本是否包含ADR的问题。早期，大多数研究均基于词典识别文本中的ADR[5-6]，但这类方法无法识别词典中未包含的非常规ADR词汇。后来有些研究人员发现，利用统计机器学习方法抽取特征能够有效提高准确性[7-8]。随着深度学习的不断发展和广泛应用，基于深度学习方法的ADR检测模型大量涌现。LEE等[9]为Twitter中的ADR分类建立了卷积神经网络（Convolutional Neural Network，CNN）模型，COCOS等[10]开发了一个递归神经网络（Recurrent Neural Network，RNN）模型，通过与任务无关的预训练或在ADR检测训练期间形成词嵌入向量，并将其作为输入。HUYNH等[11]提出2种新的神经网络模型，即将CNN与RNN连接起来的卷积递归神经网络（Convolutional Recurrent Neural Network，CRNN）以及在CNN中添加注意力权重的卷积神经网络（Convolutional Neural Network with Attention，CNNA），针对Twitter数据集分别进行了ADR分类任务。PANDEY等[12]分别采用Word2Vec和GloVe模型从多渠道的医学资源中训练临床词的词向量，将无监督的词嵌入表示输入到双向长期短期记忆（Bi-directional Long Short-Term Memory，Bi-LSTM）神经网络中，并使用注意力权重来优化ADR抽取的效果。

尽管深度学习模型往往表现很好，但通常需要基于大量标注数据进行监督学习。当标注数据过少时，容易出现过拟合现象，严重影响预测的准确性。目前，已有大量的研究从英文语料中检测ADR，但由于缺乏公开可用的中文医疗社交媒体的数据集，目前针对此方面的研究非常有限。

为解决标注数据不足的问题，本文提出一种基于数据增强与半监督学习（Semi-Supervised Learning，SSL）的药物不良反应检测方法。通过对未标注数据进行数据增强，使用分类模型产生低熵标签，以获得较为准确的伪标签。此外，将标注数据、未标注数据和增强数据混合，在文本向量空间中对混合样本进行插值，以扩增样本数量。

1　相关工作

1.1　文本增强

在少样本场景下采用数据增强技术，与同等标注量的无增强监督学习模型相比，其性能会有较大幅度的提升。文本增强技术如EDA算法[13]、回译[14]、TF-IDF等通常只针对标注数据进行有监督地数据增强。XIE等[15]将有监督的数据增强技术扩展到未标注数据中，以尽可能地利用未标注数据。GUO等[16]针对文本分类任务，提出词级的wordMixup和句子级的senMixup这2种文本增强策略，通过分别对词嵌入向量和句子向量进行线性插值，以产生更多的训练样本，提升分类性能。

1.2　半监督学习

半监督学习是一种在不需要大量标签的情况下训练大量数据模型的方法。监督学习方法仅在标注数据上训练分类器而忽略了未标注数据。SSL通过利用未标注数据的方法来减轻对标注数据的需求。在通常情况下，未标注数据的获取要比标注数据容易得多，因此SSL所带来的性能提升通常都是低成本的。SSL方法主要分为两类：一类是对一个输入添加微小的扰动，输出应该与原样本保持不变，即一致性正则化；另一类是使用预测模型或它的某些变体生成伪标签，将带有伪标签的数据和标注数据进行混合，并微调模型。

将增强技术与半监督学习方法整合于一个框架中的技术在计算机视觉领域已经取得成功。例如MixMatch[17]和FixMatch[18]方法均表现出了良好的性能。然而在自然语言处理领域，由于受文本的语法、语义关系等影响，此类方法的应用极少。

基于上述方法，本文将文本增强技术与半监督学习方法相结合，应用于面向中文医疗社交媒体的ADR检测任务。利用回译对未标注数据进行增强，以获取低熵标签，并将得到的伪标签未标注数据和标注数据进行Mixup操作，降低对标注数据的需求，充分发挥大量未标注数据的价值，提升ADR检测模型的准确性。

2　本文方法

给定有限标注的数据集 $\boldsymbol{X}_l=\{\boldsymbol{x}_1^l,\boldsymbol{x}_2^l,\cdots,\boldsymbol{x}_n^l\}$ 和其对应的标签 $\boldsymbol{Y}_l=\{\boldsymbol{y}_1^l,\boldsymbol{y}_2^l,\cdots,\boldsymbol{y}_n^l\}$，以及大量的未标注数据集 $\boldsymbol{X}_u=\{\boldsymbol{x}_1^u,\boldsymbol{x}_2^u,\cdots,\boldsymbol{x}_m^u\}$，其中 n 和 m 是2个数据集中的数据量，$\boldsymbol{y}_i^l\in\{0,1\}^C$ 是一个one-hot向量，C 是类别数。本文目标是能有效利用标注数据和未标注数据一起训练模型。为此，本文提出一种标签猜测方法，为未标注数据生成标签，将未标注数据视为附加的标注数据，并在标注数据以及未标注数据上利用Mixup方法进行半监督学习。最后，引入了熵最小化损失。本文方法的总体框架如图1所示。

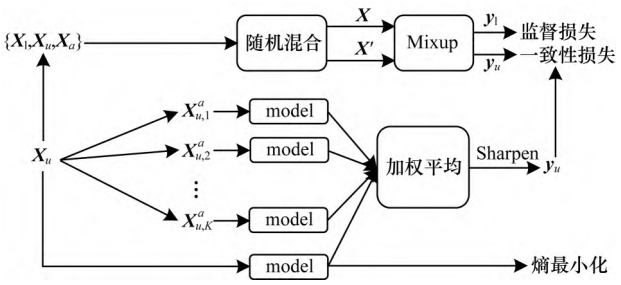

图1　本文方法框架

Fig.1　Framework of method in this paper

回译是一种常见的文本增强技术，用机器翻译把一段中文翻译成另一种语言，然后再翻译回中文。通过对同一未标注数据进行不同中间语言的回译，可以得到不同的增强数据，且能保留原始文本的语义。对于未标注数据集$\boldsymbol{X}_u$中的每一个$\boldsymbol{x}_i^u$，生成K个增强数据$\boldsymbol{x}_{i,k}^a=\text{augment}_k\left(\boldsymbol{x}_i^u\right)$，其中$k=[1,K]$，$K$表示数据增强的次数。

2.1 标签猜测

分别通过文本分类模型对未标注数据样本$\boldsymbol{x}_i^u$和K个增强数据$\boldsymbol{x}_{i,k}^a$分别进行处理，得到其预测结果，并将预测值的加权平均值生成标签：

$$\boldsymbol{y}_i^u=\frac{1}{\boldsymbol{w}_{\text{ori}}+\sum_k \boldsymbol{w}_k}\left(\boldsymbol{w}_{\text{ori}}\,p\left(\boldsymbol{x}_i^u\right)+\sum_{k=1}^{K}\boldsymbol{w}_k\,p\left(\boldsymbol{x}_{i,k}^a\right)\right) \tag{1}$$

其中：$p\left(\boldsymbol{x}_i^u\right)$和$p\left(\boldsymbol{x}_{i,k}^a\right)$分别表示未标注数据及其增强数据获得的预测值。本文期望同一样本的K个增强版本预测一致的标签，因此对预测结果进行加权平均，而不是将单个预测结果作为标签。此外，通过引入权重$\boldsymbol{w}_{\text{ori}}$和$\boldsymbol{w}_k$控制不同增强数据的贡献。

为了避免加权过于平均，对预测结果使用如式(2)所示的锐化函数进行处理：

$$\text{Sharpen}\left(\boldsymbol{y}_i^u,T\right)=\frac{\left(\boldsymbol{y}_i^u\right)^{\frac{1}{T}}}{\left\|\left(\boldsymbol{y}_i^u\right)^{\frac{1}{T}}\right\|_1} \tag{2}$$

其中：$\|\cdot\|_1$是L1范式；T是给定的超参数，当T趋向于0时，生成的标签是一个one-hot向量。

2.2 Mixup图像增强方法

Mixup是ZHANG等[19]提出的一种图像增强方法，Mixup的主要思想非常简单，即给定2个标记数据$\left(\boldsymbol{x}_i,\boldsymbol{y}_i\right)$和$\left(\boldsymbol{x}_j,\boldsymbol{y}_j\right)$，其中$\boldsymbol{x}$是图像，而$\boldsymbol{y}$是标签的one-hot向量，通过线性插值构建虚拟训练样本，如式(3)和式(4)所示：

$$\tilde{\boldsymbol{x}}=\text{mix}\left(\boldsymbol{x}_i,\boldsymbol{x}_j\right)=\lambda\boldsymbol{x}_i+(1-\lambda)\boldsymbol{x}_j \tag{3}$$

$$\tilde{\boldsymbol{y}}=\text{mix}\left(\boldsymbol{y}_i,\boldsymbol{y}_j\right)=\lambda\boldsymbol{y}_i+(1-\lambda)\boldsymbol{y}_j \tag{4}$$

其中：混合因子$\lambda\in[0,1]$，由Beta分布得到：

$$\lambda\sim\text{Beta}(\alpha,\alpha) \tag{5}$$

$$\lambda=\max\left(\lambda,1-\lambda\right) \tag{6}$$

新的训练样本将被用于训练神经网络模型。Mixup可以看作是一种数据增强方法，能够基于原始训练集创建新的数据样本。同时，Mixup强制对模型进行一致性正则化，使其在训练数据之间的标签为线性。作为一种简单有效的增强方法，Mixup可以提升模型的鲁棒性和泛化能力。

受Mixup在图像分类领域运用的启发，本文尝试将其应用于文本分类任务中。通过标签猜测得到未标注数据的标签后，将标注数据$\boldsymbol{X}_l$、未标注数据$\boldsymbol{X}_u$和增强数据$\boldsymbol{X}_a=\left\{\boldsymbol{x}_{i,k}^a\right\}$合并成一个大型的数据集$\boldsymbol{X}=\boldsymbol{X}_l\cup\boldsymbol{X}_u\cup\boldsymbol{X}_a$，对应的标签为$\boldsymbol{Y}=\boldsymbol{Y}_l\cup\boldsymbol{Y}_u\cup\boldsymbol{Y}_a$，其中$\boldsymbol{Y}_a=\left\{\boldsymbol{y}_{i,k}^a\right\}$，并且定义$\boldsymbol{y}_{i,k}^a=\boldsymbol{y}_i^u$，即对未标注数据，其所有的增强样本与原始样本共享相同的标签。

在训练过程中，本文从数据集$\boldsymbol{X}$中随机采样2个样本$\boldsymbol{x}$、$\boldsymbol{x}'$，根据式(7)和式(8)分别计算$\text{Mixup}\left(\boldsymbol{x},\boldsymbol{x}'\right)$和$\text{mix}\left(\boldsymbol{y},\boldsymbol{y}'\right)$：

$$\text{Mixup}\left(\boldsymbol{x},\boldsymbol{x}'\right)=\lambda f\left(\boldsymbol{x}\right)+(1-\lambda)f\left(\boldsymbol{x}'\right) \tag{7}$$

$$\text{mix}\left(\boldsymbol{y},\boldsymbol{y}'\right)=\lambda\boldsymbol{y}+(1-\lambda)\boldsymbol{y}' \tag{8}$$

其中：$f(\cdot)$是将文本编码为向量的神经网络。

混合样本通过分类模型获得预测值$p\left(\text{Mixup}\left(\boldsymbol{x},\boldsymbol{x}'\right)\right)$，将预测结果和混合标签进行一致性正则化，使用两者的KL散度作为损失，计算公式如式(9)所示：

$$L_{\text{Mixup}}=\mathbb{E}_{\boldsymbol{x},\boldsymbol{x}'\in\boldsymbol{X}}\text{KL}\left(\text{mix}\left(\boldsymbol{y},\boldsymbol{y}'\right)\|p\left(\text{Mixup}\left(\boldsymbol{x},\boldsymbol{x}'\right)\right)\right) \tag{9}$$

由于$\boldsymbol{x}$、$\boldsymbol{x}'$是从数据集$\boldsymbol{X}$中随机采样而来的，因此2个样本的来源有3种情况：2个都是标注数据，各有1个标注数据和1个未标注数据，2个都是未标注数据。基于以上情况，将损失分为2类：

1)监督损失。当$\boldsymbol{x}\in\boldsymbol{X}_l$时，即当实际使用的大部分信息来自于标注数据时，使用监督损失训练模型。

2)一致性损失。当样本来自未标注数据或增强数据时，即$\boldsymbol{x}\in\boldsymbol{X}_u\cup\boldsymbol{X}_a$，大部分信息来自于未标注数据，使用KL散度作为一致性损失，能够使混合样本与原始样本具有相同的标签。

2.3 熵最小化

为使模型能够基于未标注数据预测出置信度更高的标签，本文使用未标注数据的预测概率最小熵作为损失函数：

$$L_m=\mathbb{E}_{\boldsymbol{x}\in\boldsymbol{X}_u}\max\left(0,\gamma-\left\|\boldsymbol{y}^u\right\|_2^2\right) \tag{10}$$

其中：γ是边界超参数。

结合2种损失，构造总损失函数的表示式如式(11)所示：

$$L=L_{\text{Mixup}}+\gamma_m L_m \tag{11}$$

3 本文模型

本文模型包含编码层、Mixup层、分类层共3层。输入文本经过编码层得到向量表示，Mixup层通过随机混合的文本向量表示和对应的分类标签生成混合样本和混合标签。混合样本的向量表示经过Mixup层被送入分类层。分类层通过全连接层和softmax函数计算预测值，并针对混合样本的标签和预测值计算分类损失。本文模型的结构如图2所示。

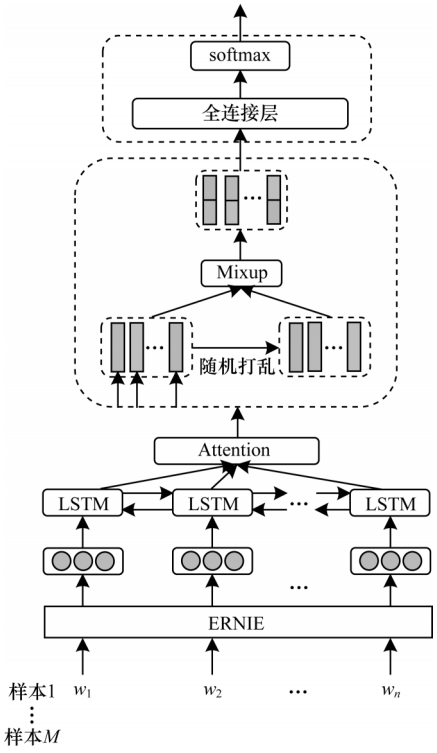

图2　本文模型结构

Fig.2　Structure of model in this paper

3.1　编码层

编码层分为ERNIE层、BiLSTM层、Attention层共3个子层。

3.1.1　ERNIE层

传统的词向量模型Word2vec得到的是静态词向量，无法体现1个词在不同语境中的不同含义，而预训练模型能够动态捕捉上下文信息，提高文本表示能力。其中ERNIE等[20]提出一种知识掩码（Knowledge Masking，KM）策略，在训练阶段引入外部知识，并随机选取字、短语、命名实体进行mask，可以潜在地学习到被掩码的短语和实体之间的先验知识。此外，新增预训练任务，使ERNIE词向量从训练数据中获取到更可靠的词法、语法以及语义信息。

中文医疗文本存在一词多义的问题，往往需要结合上下文语境才能获得精确的语义信息，且药物不良反应检测通常与外部知识、药物实体等密切相关。因此，本文使用百度开源的ERNIE中文预训练模型，并充分利用该模型的外部知识和实体信息。

ERNIE采用多层Transformer作为编码器，通过自注意力机制捕获每个词向量在文本序列中的上下文信息，并生成上下文语境表征嵌入。以语料中的一个文本为例，文本中的词序列$\boldsymbol{x}=\left[w_1,w_2,\cdots,w_n\right]$，其中：$n$表示文本长度；$w_i$表示文本中的第$i$个词。文本经过ERNIE预训练模型得到词序列的向量表示$\boldsymbol{E}=\left[\boldsymbol{e}_1,\boldsymbol{e}_2,\cdots,\boldsymbol{e}_n\right]$，其中：$\boldsymbol{e}_i\in\mathbb{R}^d$表示第$i$个词的词向量；$d$表示词向量的维度。

3.1.2　BiLSTM层

BiLSTM层以词向量表示为输入，计算词语在上下文中的向量表示：

$$\vec{\boldsymbol{h}}_i=\overrightarrow{\mathrm{LSTM}}\left(\vec{\boldsymbol{h}}_{i-1},\boldsymbol{e}_i\right) \tag{12}$$

$$\overleftarrow{\boldsymbol{h}}_i=\overleftarrow{\mathrm{LSTM}}\left(\overleftarrow{\boldsymbol{h}}_{i+1},\boldsymbol{e}_i\right) \tag{13}$$

$$\boldsymbol{h}_i=\mathrm{concat}\left(\vec{\boldsymbol{h}}_i,\overleftarrow{\boldsymbol{h}}_i\right) \tag{14}$$

其中：$\vec{\boldsymbol{h}}_i$、$\overleftarrow{\boldsymbol{h}}_i$分别表示第$i$个词通过正向和反向的LSTM网络得到的输出，将$\vec{\boldsymbol{h}}_i$、$\overleftarrow{\boldsymbol{h}}_i$拼接作为第$i$个词在序列中的隐藏状态表示$\boldsymbol{h}_i$。通过拼接$\boldsymbol{h}_i$序列可得到词序列的隐藏状态表示：$\boldsymbol{H}=\left[\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_n\right]$，$\boldsymbol{H}\in\mathbb{R}^{n\times 2u}$。

3.1.3　Attention层

Attention层将BiLSTM层的隐藏状态作为输入，通过自注意力权重分配来计算文本多个侧面的向量表示，表达式如下：

$$\boldsymbol{A}=\mathrm{softmax}\left(\boldsymbol{W}_{s2}\tanh\left(\boldsymbol{W}_{s1}\boldsymbol{H}^{\mathrm{T}}\right)\right) \tag{15}$$

$$\boldsymbol{Z}=\boldsymbol{A}\boldsymbol{H} \tag{16}$$

其中：$\boldsymbol{A}\in\mathbb{R}^{r\times n}$表示注意力权重矩阵，由2层感知器网络计算得到；$\boldsymbol{W}_{s1}\in\mathbb{R}^{d_a\times 2u}$和$\boldsymbol{W}_{s2}\in\mathbb{R}^{r\times d_a}$分别是2层注意力层的权重矩阵；$d_a$、$r$是超参数，$d_a$表示注意力层隐藏状态的维度，$r$是注意力机制的个数。

将文本表示矩阵$\boldsymbol{Z}\in\mathbb{R}^{r\times 2u}$中的$r$个向量拼接得到文本的向量表示$\boldsymbol{z}$，其维度为$2ur$。

3.2　Mixup层

本文基于Mixup方法，在文本的向量空间中混合样本。混合过程是先随机选取一个样本，然后将同批次的样本随机打乱后抽取另一个样本，对2个样本$\left(\boldsymbol{z}_i,\boldsymbol{y}_i\right)$和$\left(\boldsymbol{z}_j,\boldsymbol{y}_j\right)$进行插值。抽取过程均为不放回抽取。样本混合因子$\lambda$由式(5)和式(6)得到：

$$\tilde{\boldsymbol{z}}=\lambda\boldsymbol{z}_i+\left(1-\lambda\right)\boldsymbol{z}_j \tag{17}$$

$$\tilde{\boldsymbol{y}}=\lambda\boldsymbol{y}_i+\left(1-\lambda\right)\boldsymbol{y}_j \tag{18}$$

在训练的过程中，Mixup层通过随机混合批次内的文本向量表示$\left\{\boldsymbol{z}_i\right\}_i^M$得到扩增的文本向量表示$\left\{\tilde{\boldsymbol{z}}_j\right\}_j^M$，其中$M$表示一个批次的样本数据量。

3.3　分类层

混合样本通过一个全连接层和softmax激活函数，得到分类标签的预估概率值：

$$\hat{\boldsymbol{y}}=\mathrm{softmax}\left(\boldsymbol{W}\tilde{\boldsymbol{z}}+\boldsymbol{b}\right) \tag{19}$$

其中：$\boldsymbol{W}\in\mathbb{R}^{C\times 2ur}$和$\boldsymbol{b}\in\mathbb{R}^C$分别是权重矩阵和偏置。

4　实验结果与分析

4.1　实验数据

由于目前中文医疗社交媒体没有公开可用的药物不良反应检测数据集，因此本文从好大夫网站收集用户的诊疗记录。如图3所示，每个诊疗记录包含患者的信息、病情描述、医生诊疗建议等内容。

图3 诊疗记录样例

Fig.3 Sample of treatment record

本文选取80余种常用药作为研究内容，其中包含降压药、抗过敏药、抗生素等，获取了网站2011年以后包含相关药物的诊疗记录，并选择记录中的病情描述内容作为本文的原始语料。最终共获得42 800个文本，每个文本都提及了一种或者多种药物。通过对文本进行预处理，删除URL、英文字母、特殊字符等并去停用词。原始语料来源于中文社交媒体，需要对其进行分词。对于医疗数据，传统的jieba分词效果并不理想，因此本文使用北京大学开源分词工具pkuseg进行分词，调用其自带的medicine模型将大部分的医药专业词汇分词出来。

为得到标注数据，本文从语料中选取6 000条数据让5名药学专业学生进行人工标注，设定分类标签 $y \in \{0,1\}$。当不同人员之间存在争议，同一数据的标注结果不一致时，以多数人的结果为准。标注样例如表1所示。最终得到包含ADR的数据有2 379条，不包含ADR的数据有3 621条，并从中随机选择4 800条作为训练集，1 200条作为测试集。

表1 标注数据样例

Table 1 Samples of labeled data

标签	说明	样例
1	含有ADR的文本	由于服用硝苯地平控释片，虽然血压控制较好，但小腿及脚面浮肿严重
0	不含ADR的文本	患有高血压，脑梗，长期吃血压药，现在在吃硝苯地平

4.2 参数设置

本文采用Pytorch实现所提模型和算法，将文本的最大词序列长度设为256。ERNIE预训练模型包含12个Encoder层，多头注意力机制的头数为12，隐藏层维度为768。将BiLSTM层的隐藏状态维度 u 设为300，Attention层自注意力机制的参数 r 设为24，注意力层隐藏状态维度 d_a 设为128。未标注数据的增强次数 K 对于伪标签的影响较大，参照文献[21]，设置 $K=2$，即对于每个未标注数据执行2次数据增强。将Beta分布中的 α 设置为0.4，当 α 较小时，$\mathrm{Beta}(\alpha,\alpha)$ 采样得到的值大部分落在0或1附近，使样本混合因子 λ 接近1，从而在合成样本时偏向某一个样本创建出相似的数据。

模型采用Adam梯度下降算法训练，初始学习率设为0.001，$\beta_1=0.9$，$\beta_2=0.999$，$\varepsilon=1\times10^{-8}$，batchsize=64。训练过程采用提前停止策略，若模型性能在5个epoch后仍然没有提升，则停止训练。

4.3 实验对比

4.3.1 ADR检测模型的对比实验

本文选择了如下6种基于深度学习的ADR检测模型进行对比实验：

1）CNN[11]模型：采用不同尺度的卷积神经网络构建文本分类器。分别设置滤波器宽度为2、3、4、5，每个滤波器的大小均为25。

2）CNN+Att[11]模型：在CNN网络最上层加入注意力机制。

3）BiL+Att[12]模型：采用BiLSTM作为编码器，加入注意力机制。

4）ERNIE[20]模型：采用百度开源的ERNIE中文预训练模型作为编码器，得到文本表示，直接连接一个全连接层实现文本分类。

5）ERNIE+BiL+Att模型：基于BiL+Att模型，使用ERNIE模型得到词向量表示。

6）ERNIE+BiL+Att+S-Mixup模型：在ERNIE+BiL+Att模型的编码层之上加Mixup层。对标注数据进行文本增强，即有监督的Mixup（Supervised Mixup，S-Mixup）。

选取4 800条标注数据训练模型，使用精确率、召回率和F1值作为评价指标。实验结果如表2所示。

表2　不同ADR检测模型的实验结果

Table 2　Experimental results of different ADR detection models　%

模型	精确率	召回率	F1值
CNN模型	66.5	64.8	63.1
CNN+Att模型	67.0	66.9	67.4
BiL+Att模型	69.3	68.2	68.5
ERNIE模型	71.9	69.5	72.1
ERNIE+BiL+Att模型	72.5	71.2	72.5
ERNIE+BiL+Att+S-Mixup模型	73.7	72.7	73.3

由表2可知，CNN模型的效果最差，而采用BiLSTM获取上下文信息，并引入注意力机制获取文本的重要特征，能提高模型效果。对比BiL+Att和ERNIE+BiL+Att模型，利用ERNIE预训练模型得到的动态词向量更符合语义环境，模型的性能也能得到有效提升。

本文对ERNIE预训练模型进行微调，实验效果显著，说明了预训练模型在ADR检测任务中能达到较好的分类效果。然而对比ERNIE与ERNIE+BiL+Att模型，后者的实验效果仍有小幅度提升，体现了ERNIE+BiL+Att模型的优势。

由表2还可以看出，ERNIE+BiL+Att+S-Mixup模型的精确率、召回率和F1值均优于其他模型。这是因为神经网络的训练通常需要大量的标注数据，而当标注数据有限时，效果往往不太理想。ERNIE+BiL+Att+S-Mixup模型引入Mixup，通过对标注数据进行文本增强，在一定程度上增加了训练样本的数量，从而使ADR检测模型的性能得到明显的提升。

4.3.2　半监督模型的对比实验

本文选取了如下5种半监督模型进行对比实验：

1）ERNIE+BiL+Att+S-Mixup模型：仅使用标注数据。

2）Pseudo-Label[22]模型：先使用标注数据训练模型，将未标注数据经过分类模型后得到的预测值作为伪标签，使用带有伪标签的数据和标注数据一起训练模型。

3）Π-Model[23]模型：对于同一数据的输入，使用不同的正则化进行2次预测，通过减小2次预测的差异，提升模型在不同扰动下的一致性。

4）Mean Teachers[24]模型：使用时序组合模型，对模型参数进行EMA平均，将平均模型作为teacher预测人工标签，由当前模型预测。

5）ERNIE+BiL+Att+SS-Mixup模型：即本文模型。先对未标注数据进行多次增强，将预测值加权平均作为低熵标签，并共享原始样本和增强样本。使用标注数据、未标注数据和增强数据一起对模型进行训练，即半监督的Mixup（Semi-Supervised Mixup，SS-Mixup）。

从训练集中选取不同数量的标注数据和5 000条未标注数据。使用准确率（Accuracy，Acc）作为评价指标，实验结果如表3所示。

表3　不同半监督模型的Acc值对比

Table 3　Acc value comparison of different semi-supervised models　%

模型	标注数据量为800	标注数据量为1 500	标注数据量为2 800
ERNIE+BiL+Att+S-Mixup模型	60.3	69.2	75.2
Pseudo-Label模型	64.2	70.5	75.5
Π-Model模型	66.8	71.7	75.8
Mean Teachers模型	68.1	72.7	76.0
ERNIE+BiL+Att+SS-Mixup模型	72.6	75.5	76.6

由表3可知，与传统的半监督模型相比，本文模型在不同标注数据量的情况下，准确率均最高。当标注数据的数量较少时，准确率增长幅度尤其突出。随着标注数据的增加，本文模型带来的额外提升效果会逐渐降低。从表3中还可以看出，当标注数据量为1 500条时，本文模型与ERNIE+BiL+Att+S-Mixup模型在2 800条标注数据时的表现相近。即通过本文对未标注数据的半监督学习，相当于免费获得了近一倍的额外标注数据。说明本文模型有效利用了未标注数据的信息，缓解了标注数据量不足的影响。同时，本文模型对未标注数据有较好的标签预测能力。

4.3.3　不同未标注数据量的对比实验

为进一步对比未标注数据量对本文模型的影响，从训练集中挑选了800条标注数据和不同数量的未标注数据。实验结果如表4所示。

表4　不同未标注数据量的Acc结果

Table 4　Acc results of different unmarked data quantities

未标注数据量	Acc值/%
0	60.3
2 000	67.3
4 000	71.8
6 000	73.5
8 000	74.7

由表4可知，当标注数据量一定时，未标注数据的数量越多，本文模型的预测结果越准确，表明本文模型能够有效利用未标注数据的信息，帮助模型提升性能。

5 结束语

本文面向中文医疗社交媒体提出一种融合数据增强与半监督学习的ADR检测方法。通过利用回译的文本增强技术对未标注数据进行多次增强，并在模型的编码层和分类层之间加入Mixup层，对混合样本的文本向量采取插值操作以扩充样本数量。此外，通过半监督学习训练分类模型，充分利用标注数据与未标注数据。实验结果表明，本文方法充分利用未标注数据解决了标注数据不足的问题。当标注数据量较少时，模型的准确率提升幅度尤其突出，且随着未标注数据量的增加，模型性能得到提升。下一步将研究文本中药物和不良反应的关系，通过辨别文本中出现的ADR信息是否已知，从而挖掘潜在的ADR信息，提升本文模型的性能。

参考文献

[1] HARPAZ R, DUMOUCHEL W, SHAH N H, et al. Novel data-mining methodologies for adverse drug event discovery and analysis[J]. Clinical Pharmacology and Therapeutics, 2012, 91(6): 1010-1021.

[2] WANG W, HAERIAN K, SALMASIAN H, et al. A drug-adverse event extraction algorithm to support pharmacovigilance knowledge mining from PubMed citations[J]. AMIA Annual Symposium Proceedings, 2011, 25(3): 64-70.

[3] SOHN S, KOCHER J P A, CHUTE C G, et al. Drug side effect extraction from clinical narratives of psychiatry and psychology patients[J]. Journal of the American Medical Informatics Association, 2011, 18(1): 144-149.

[4] WARRER P, HANSEN E H, JUHL JENSEN L, et al. Using text-mining techniques in electronic patient records to identify ADRs from medicine use[J]. British Journal of Clinical Pharmacology, 2012, 73(5): 674-684.

[5] WU H, FANG H, STANHOPE S J. Exploiting online discussions to discover unrecognized drug side effects[J]. Methods of Information in Medicine, 2013, 52(2): 152-159.

[6] YATES A, GOHARIAN N. ADRTrace: detecting expected and unexpected adverse drug reactions from user reviews on social media sites[C]//Proceedings of the 35th European Conference on Advances in Information Retrieval. Berlin, Germany: Springer, 2013: 816-819.

[7] SARKER A, GONZALEZ G. Portable automatic text classification for adverse drug reaction detection via multi-corpus training[J]. Journal of Biomedical Informatics, 2015, 53(4): 196-207.

[8] NIKFARJAM A, SARKER A, O'CONNOR K, et al. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features[J]. Journal of the American Medical Informatics Association, 2015, 22(3): 671-681.

[9] LEE K, QADIR A, HASAN S A, et al. Adverse drug event detection in tweets with semi-supervised convolutional neural networks[EB/OL]. [2021-06-20]. https://dl.acm.org/doi/10.1145/3038912.3052671.

[10] COCOS A, FIKS A G, MASINO A J. Deep learning for pharmacovigilance: recurrent neural network architectures for labeling adverse drug reactions in Twitter posts[J]. Journal of the American Medical Informatics Association, 2017, 24(4): 813-821.

[11] HUYNH T, HE Y, WILLIS A, et al. Adverse drug reaction classification with deep neural networks[C]//Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers. Osaka, Japan: [s. n.], 2016: 877-887.

[12] PANDEY C, IBRAHIM Z, WU H H, et al. Improving RNN with attention and embedding for adverse drug reactions[C]// Proceedings of 2017 International Conference on Digital Health. New York, USA: ACM Press, 2017: 67-71.

[13] WEI J, ZOU K. EDA: easy data augmentation techniques for boosting performance on text classification tasks[C]// Proceedings of 2019 Conference on Empirical Methods in Natural Language. Stroudsburg, USA: Association for Computational Linguistics, 2019: 6381-6387.

[14] EDUNOV S, OTT M, AULI M, et al. Understanding back-translation at scale[EB/OL]. [2021-06-22]. https://arxiv.org/abs/1808.09381.

[15] XIE Q Z, DAI Z H, HOVY E, et al. Unsupervised data augmentation for consistency training[EB/OL]. [2021-06-22]. https://arxiv.org/abs/1904.12848.

[16] GUO H Y, MAO Y Y, ZHANG R C. Augmenting data with mixup for sentence classification: an empirical study[EB/OL]. [2021-06-22]. https://arxiv.org/abs/1905.08941.

[17] BERTHELOT D, CARLINI N, GOODFELLOW I, et al. MixMatch: a holistic approach to semi-supervised learning[EB/OL]. [2021-06-22]. https://www.researchgate.net/publication/332932671_MixMatch_A_Holistic_Approach_to_Semi-Supervised_Learning.

[18] SOHN K, BERTHELOT D, LI C L, et al. FixMatch: simplifying semi-supervised learning with consistency and confidence[EB/OL]. [2021-06-22]. https://arxiv.org/abs/2001.07685.

[19] ZHANG H Y, CISSE M, DAUPHIN Y N, et al. Mixup: beyond empirical risk minimization[EB/OL]. [2021-06-22]. https://arxiv.org/abs/1710.09412.

[20] SUN Y, WANG S, LI Y, et al. ERNIE: enhanced representation through knowledge integration[C]//Proceedings of AAAI Conference on Artificial Intelligence. San Francisco, USA: AAAI Press, 2020: 8968-8975.

[21] CHEN J A, YANG Z C, YANG D Y. MixText: linguistically-informed interpolation of hidden space for semi-supervised text classification[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, USA: Association for Computational Linguistics, 2020: 2147-2157.

[22] LEE D H. Pseudo-label: the simple and efficient semi-supervised learning method for deep neural networks[EB/OL]. [2021-06-22]. https://www.researchgate.net/publication/280581078.

[23] LAINE S, AILA T M. Temporal ensembling for semi-supervised learning[J]. IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, 2016, 12: 143-152.

[24] TARVAINEN A, VALPOLA H. Mean teachers are better role models: weight-averaged consistency targets improve semi-supervised deep learning results[C]//Proceedings of International Conference on Learning Representations. Vancouver, Canada: [s. n.], 2017: 156-168.

编辑 赖玉玲