计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.201100125

面向电子病历语义解析的疾病辅助诊断方法

范红杰[1]　李雪冬[2]　叶松涛[3]
1 中国政法大学科学技术教学部　北京 102249
2 北京大学软件与微电子学院　北京 102600
3 湘潭大学计算机学院　湖南 湘潭 411105
(hjfan@cupl.edu.cn)

摘　要　针对面向电子病历的疾病辅助诊断问题,文中将词向量和文本判别方法应用到电子病历的文本语义解析任务中。具体地,采用预训练语言模型作为字符的语义表征,从而对文本特征进行准确表达,在卷积神经网络中提取N元特征后,使用胶囊单元对特征进行聚类,从而更好地捕获文本的高层语义特征,同时减少对数据量的需求。实验发现,基于ERNIE+CNN+Capsule的组合模型在真实的电子病历数据集上取得了良好的效果。此外,受图像风格迁移的启发,文中训练了从电子病历文本到病情自述文本的风格转换模型,利用非平行数据,在风格转换模型的基础上,增加了对抗思想和困惑度评价指标,可以有效缓解训练数据和测试数据分布不一致的问题。最后,相比$ALBERT_{tiny}$,BERT等模型,所提模型在病历文本上获得了86.89%的$F1$值,提升了1.36%～3.68%;在泛化性能任务评估中,获得了94.95%的$F1$值。实验证明,所提模型在保证较高准确率的前提下,可以有效适应疾病辅助诊断。

关键词:电子病历;语义解析;辅助诊断;深度神经网络;胶囊网络
中图法分类号　TP391.4

Aided Disease Diagnosis Method for EMR Semantic Analysis

FAN Hong-jie[1],LI Xue-dong[2] and YE Song-tao[3]
1 The Department of Science and Technology Teaching,China University of Political Science and Law,Beijing 102249,China
2 School of Software and Microelectronics,Peking University,Beijing 102600,China
3 School of Computer Science,Xiangtan University,Xiangtan,Hunan 411105,China

Abstract　Aiming at solving the problem of auxiliary disease diagnosis for electronic medical record,the word vector and text discrimination method are applied to the semantic text analysis task. Concretely,the pre-training language model is used as the semantic representation of characters,so as to accurately express the text features. After extracting N-ary features from convolutional neural network,the capsule unit is used to cluster the features,so as to better capture the high-level semantic text features and reduce the demand for data. It is found that the combination model based on ERNIE+CNN+Capsule achieves high accuracy on the real EMR. In addition,inspired by the image style transfer,a style conversion model from EMR text to disease self-report text is trained. Based on the style conversion model,non-parallel data are used to add confrontation ideas and confusion evaluation indexes,which can effectively alleviate the problem of inconsistent distribution of training data and test data. Finally,compared with $ALBERT_{tiny}$,BERT and other models,the proposed model gets 86.89% $F1$ value in the EMR,which is improved by 1.36%～3.68%,and 94.95% $F1$ value in the generalization. Experiments show that the proposed model can effectively adapt to the auxiliary disease diagnosis on the premise of ensuring high accuracy.

Keywords　Electronic medical record,Semantic analysis,Auxiliary diagnosis,Deep neural networks,Capsule network

1　引言

电子病历(Electronic Medical Record,EMR)作为医务人员使用医疗机构信息系统生成的文字、符号、图表等数字化信息,可以重现患者的医疗记录[1]。一份完整的电子病历阐述了医务人员从不同的角度反映患者的身体状况,包括健康记录、医疗结果和用药计划等。目前大多数电子病历以自然语言的方式进行记录,并以结构化形式(如诊断代码等)或非结构化形式(如临床记录、进展记录等)进行存储,研究人员很难对其进行分析和处理。随着医疗数据的积累、计算能力的提

到稿日期:2020-11-17　返修日期:2021-04-16　本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家自然科学基金(61802327);湖南省自然科学基金(2018JJ3511)
This work was supported by the National Natural Science Foundation of China(61802327) and Natural Science Foundation of Hunan Province(2018JJ3511).
通信作者:叶松涛(yesongtao@xtu.edu.cn)

升和机器学习技术的优化，面向电子病历的疾病辅助诊断成为了可能。

面向电子病历文本的辅助诊断通常被规约成文本分类问题。2012 年之前，该类问题通常先利用人工提取统计特征，继而利用传统的机器学习算法进行判别。2012 年，Hinton 等[2]利用 AlexNet 模型在 ImageNet 图像识别大赛中夺冠，深度学习走进研究者的视野。Zhang 等[3]从可穿戴设备和应用分类算法的角度出发，对帕金森震颤疾病和原发性震颤疾病辅助诊断进行了综述，并讨论其优点和局限性。Chen 等[4]建立了医学领域语义知识库，提出了一种基于领域知识库的疾病辅助诊断方法，并与现有方法进行了对比。

目前，使用深度学习进行判别时，一般使用基于循环神经网络（Recurrent Neural Network，RNN）和卷积神经网络（Convolutional Neural Networks，CNN）的方法。RNN 非常适合处理时序数据，因此常被用来处理文本。长短时记忆网络（Long Short-Term Memory，LSTM）[5]作为 RNN 的特例，通过门机制有效缓解了传统 RNN 中存在的长序列依赖问题，但训练过程依然是串行的。Kim 等[6]首先将 CNN 神经网络架构应用于文本分类，提出了 TextCNN 模型。虽然深度学习技术在病历文本分类任务中取得了巨大进展，但这些模型在训练时通常需要大量数据，当数据量不足时，其泛化能力会迅速下降。

为了提高深度学习模型的泛化能力，本文针对电子病历采用基于知识增强的 ERNIE[7]作为预训练模型，利用 CNN 提取 N 元特征；在此基础上，使用胶囊单元[8]对特征进行聚类，从而更好地捕获文本的高层语义特征，并减少对数据量的需求。实验证明，ERNIE＋CNN＋Capsule 组合模型在真实的电子病历数据集上取得了良好的效果；此外，我们在开源中文数据集中进行了泛化性评估，证明了该模型能够很好地应用于疾病文本分类。

2 相关工作

2.1 文本向量化方法

在文本语义挖掘任务中，如何更好地将词表示为计算机可以理解的形式，逐渐成为学术界近年来的研究热点。传统的词袋（bag-of-words）模型[9]通常假设词是独立的，并由此统计每个单词出现的频率。Mikolov 等提出使用 Word2Vec[10]作为预训练词向量，取代了独热编码[11]，较好地缓解了词袋模型高维、稀疏的问题，又很好地衡量了词汇间的语义联系。为了获得上下文内容相关的向量表示，Peters 等[12]提出 ELMO（Embeddings From Language Model）向量化方法，提高了下游任务的模型性能。此后 BERT（Bidirectional Encoder Representations From Transformers）使用 Mask 机制[13]替代 ELMO 中前后向模型的拼接，更好地融合了上下文信息。

2.2 文本判别方法

RNN 常被用来处理时序数据，结合当前时刻输入对当前时刻的输出进行预测。传统 RNN 容易出现梯度爆炸或梯度消失的问题，而长短时记忆网络（LSTM）[5]通过门机制有效缓解了传统 RNN 中存在的长序列依赖问题。

2.3 训练数据对抗增广方法

当训练数据不足时，研究者通常希望模型生成一些特定格式的文本来增强训练数据。Goodfellow 等[14]首先提出了生成对抗网络（Generative Adversarial Nets，GAN）。该网络通过生成器模块尽最大可能使判别器无法区分真假数据，而判别器的训练目标是尽最大可能区分真假数据。其中 CycleGAN[15]将某一种风格的文本转换成另一种风格的文本时，无须考虑样本构成风格，即可完成风格的迁移，具有较强的通用性。

3 疾病辅助诊断框架

3.1 模型整体网络结构

如图 1 所示，诊断模型将预处理后的文本输入到 ERNIE 模型[7]中，每个字符经过 ERNIE 编码形成词向量。此后 TextCNN 卷积层使用(3,4,5)3 种尺寸的卷积核来捕获 N 元特征，每个尺寸的卷积核最终均会生成一张特征图，再送入胶囊网络。在生成胶囊的过程中，使用 Squash 非线性函数对其转换结果进行压缩。由于在最开始卷积层中使用了 3 个尺寸的卷积核，本文将这 3 个尺寸卷积核输出的特征进行拼接，由此得到了 PrimCap 层。在 PrimCaps 到 ConvCaps 层之间，本文利用动态路由算法对特征进行聚类。神经网络在此基础上使用全连接和 Softmax 函数进行最终分类，即在 FCCaps 层，通过路由算法将经过拉平的胶囊变成最终的胶囊向量以及它的概率，此时每个胶囊代表一个类别。基于上述思路，疾病辅诊学习方法如算法 1 所示。

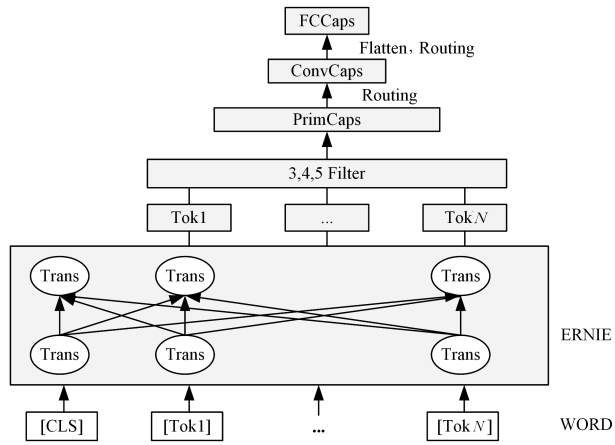

图 1 面向电子病历的疾病辅助诊断模型

Fig. 1 Auxiliary disease diagnosis model for EMR

算法 1 面向电子病历的疾病辅助诊断方法

Input: Auxiliary disease diagnosis algorithm for electronic medical record

Output: Probability of disease category

1. Truncated or pad text data according to maxlen;
2. Transform words into word embedding by Ernie;
3. while epoch=1,…,N do
   3.1. Select a random batch containing M samples
   3.2. for traing sample i=1,…,M do
   3.3. Feed i into convolution layer;
   3.4. Obtain CN=(cn_1,…,cn_t,…,cn_n);

3.5. Connect Capsule layer byDynamic Routing

3.6. Begin Procedure Routing

for r_num iterations do

for all capsule i: $C_{ij}=\sum_k \exp(B_{ij})/\sum \exp(B_{ik})$

for all capsule j: $S_j=\sum C_{ij}\cdot U_{j|i}$

for all capsule j: $e_{ij}^s=\text{squash}(S_j)$

for all capsule i and j: $B_{ij}=B_{ij}+U_{j|i}*V_j$

3.7. End procedure

3.8. Calculate category probability distribution

3.9. end for

4. end while

5. END

3.2 ERNIE 向量化

每个输入句子通过 ERNIE 层转换表示变成向量 $\boldsymbol{X}\in R^{L\times V}$,$L$ 为句子的长度,V 是词向量的维度。$\boldsymbol{X}_i\in R^V$ 是第 i 个字符的 V 维词向量。

ERNIE 模型使用 Transformer 的 Encoder 端作为特征提取器,通过自注意力机制可以更好地捕获句子中每个词语的上下文信息,并且生成相应的词向量序列。其中 Encoder 端由 6 个相同的 Block 堆叠而成,每个 Block 含有两个子层,一个为多头自注意力机制,另一个为全连接网络。两个子层应用了残差网络和层归一化,本文的所有中间输出都设置为 512 维,以便应用残差网络。ERNIE 通过预测海量数据中的词和实体,学习存在于现实世界中真实的语义关系,进一步增强了语义的表达能力。

如图 2 所示,以"因倦怠乏力要求行中医药调理"这句话为例,ERNIE 模型通过学习词语或者实体表示进行了语义建模。虽然"中医药"这个词会被完全遮掩,但是模型试图通过上下文来推断被遮掩住的词和实体。因此,该模型能有效缓解被遮掩掉的各部分相互孤立的词或实体。

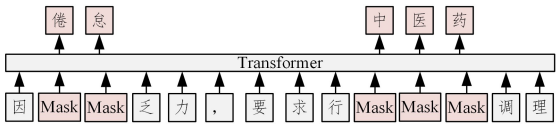

图 2 基于 ERNIE 模型的语义建模示例

Fig. 2 Example of semantic modeling based on ERNIE model

3.3 低层特征提取

本文采用 TextCNN 的卷积层对句子进行底层特征提取。如式(1)所示,F_i 是句子中对应位置 i 生成的特征,每个卷积核会生成一张特征图,每个尺寸使用多个卷积核。$\boldsymbol{W}\in R^{H\times V}$是卷积核的大小,$H$ 是在整个句子上滑动的固定窗口大小,B 是偏置,g 是非线性函数。本文使用式(1)进行卷积,可以使字符的局部顺序被保留,此外,相同宽度的卷积操作使总的卷积次数减少。

$$F_i=g(\sum_{h=1}^{h=H}\sum_{V=1}^{v=V}W_{h,v}X_{i+h;v}+B_i) \tag{1}$$

3.4 特征聚合

由于胶囊允许网络学习部分和整体之间关系的不变性,因此可以提取更丰富的文本信息,从而提高了模型整体的表现力。本文使用 PrimCaps 接收上一个卷积层的输出向量(胶囊)作为输入。由于卷积神经网络中低级到高级特征的路由是通过最大池化实现的,因此会丢失句子中的局部位置信息。如图 3 所示,本文采用名为动态路由的方法进行分配,即下层胶囊以不同概率被分到上层胶囊中。动态路由的目标是构建一个从子胶囊到父胶囊的非线性映射,该映射是迭代构建的,上层子胶囊以一定的概率分配到后续父胶囊中,逐渐降低了类别无关特征的权重。

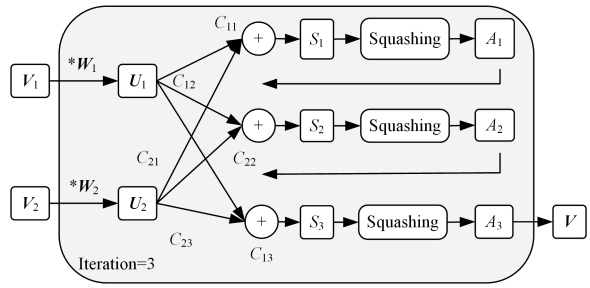

图 3 动态路由过程

Fig. 3 Dynamic routing process

设一个单独的胶囊输出为 $\boldsymbol{v}_i$,它与一个转换矩阵 w_{ij} 相乘,输出一个向量 $\boldsymbol{U}_{j|i}$。其中 i 是当前层的胶囊,j 是下一层的胶囊。如式(2)所示,下一层的胶囊为当前层胶囊的加权求和:

$$S_j=\sum \boldsymbol{C}_{ij}\cdot \boldsymbol{U}_{j|i} \tag{2}$$

其中,权重参数 $\boldsymbol{C}_{ij}$ 由 $\boldsymbol{B}_{ij}$ 经过 Softmax 而来,其值在路由过程中不断更新,如式(3)所示:

$$\boldsymbol{C}_{ij}=\sum_k \exp(\boldsymbol{B}_{ij})/\sum \exp(B_{ik}) \tag{3}$$

其中,$\boldsymbol{B}_{ij}=\boldsymbol{B}_{ij}+\boldsymbol{U}_{j|i}\times\boldsymbol{V}_j$并初始为 0。在生成胶囊的过程中,为了使每次的操作结果都能重新回到数值 0 到 1 之间,本文使用 Squash 非线性函数来压缩转换结果,如式(4)所示:

$$Squash(x)=(x^2/1+x^2)\times(x/\|x\|) \tag{4}$$

最后,胶囊的损失计算方式如下:

$$Loss_{\text{term}}=T_c*\max(0,m^{\text{high}}-\|v\|)^2+\lambda(1-T_c)*\max(\|v\|-m^{\text{low}})^2 \tag{5}$$

$Loss_{\text{term}}$由两个部分组成:正确类别对应胶囊的损失和错误类别对应胶囊的损失。文中辅助诊断任务基于带标签的医疗文本进行训练,如果使用真实标签对应序号的胶囊,则 $T_c=1$,否则 $T_c=0$。$\|v\|$ 为胶囊的模长,代表该类别的概率。m^{high} 表示正确类别对应的胶囊单元概率,m^{low} 表示错误类别对应的胶囊单元概率。当 m^{high} 大于 0.9 时,对应胶囊的损失为 0。只有正确类别的预测概率小于 0.9 时,损失函数才会将该胶囊单元的损失累加。同理,如果错误类别对应的胶囊单元概率小于 0.1,则胶囊单元的损失为 0。只有错误类别的预测概率大于 0.1 时,损失函数才会将该胶囊的损失累加。λ 是两部分损失的权重比,默认为 0.5。

4 实验与结果分析

4.1 数据集

本实验采用合作医院的真实医疗记录的 38049 个脱敏癌症病历样本,每个样本可以拆分成数句话。经统计,病历数据分布如表 1 所列。

表 1 电子病历数据集统计

Table 1 EMR dataset statistics

	训练集	验证集	测试集
电子病历文本	26634	7610	3805
病情自述文本	21315	6090	3045

数据集包含 5 列特征，分别是性别、年龄段、临床表现、入院查体、辅助检查，标签类别包含多种癌症分类。经统计，临床表现特征列的文本平均长度约为 500。经过清洗后，小于 ERNIE1.0[1] 预训练模型定义最大文本长度的文本可以直接输入到 ERNIE 模型中，不用再进行分段处理。考虑到具体日期对实验结果的影响较小，统一使用 ERNIE 词典中的特殊字符进行替换。本文分别训练原始的电子病历和病情自述文本，并在训练基于病情自述文本的辅助模型时，剔除少部分质量较差的样本。

实验采用 Pandas 0.25.3[2] 进行数据处理，使用 CUDA 9.2[3] 作为深度学习运算平台，深度学习框架使用 Pytorch 1.3.1[4]，使用 Tesla V100 作为训练平台。

4.2 病历风格迁移

在根据患者病情自述作初步诊断中，电子病历文本偏正式且专业词汇多，而任务需要的疾病辅助诊断训练数据偏口语化。因此，本文考虑采用属性可控的文本迁移来重构训练数据集。

本文在 DualRL[16] 模型的基础上提出了去风格分离的 DualRL2.0 病历风格迁移模型，该模型在原始 DualRL 的基础上引入对抗生成的思想，无需进行显式或者隐式的风格与内容分离，从而允许在没有平行语料库的情况下实现对病历文本风格的转换。经评估，DualRL2.0 可以有效地将医学电子病历文本转换为患者病情自述风格型文本，并且其内容保留程度较高，流畅性较强。

该模型的前向模型 f_θ 和反向模型 g_ϕ 都是基于 Bi-LSTM 的编码器-解码器框架，首先前向模型 f_θ 和反向模型 g_ϕ 在伪平行数据上预训练 5 轮。预训练时，Adam 学习率为 10^{-3}，批大小是 32。预训练之后，对抗生成网络整体进行 10 轮联合训练。R_c，R_s，R_f 3 种反馈的权重比例分别是 0.4，0.3，0.3。实验中，初始的迭代间隔 T_0 为 1，最大的时间间隔 $T_{\max}$ 为 100，增长率为 1.1。若模型的效果在长时间内得不到提升，则停止训练。

病情自述风格样本的生成是否满足要求可以通过风格转换强度、内容一致性和语言流畅性 3 种指标进行评估。该模型采用两种评估方式：人工评估和自动评估。

(1)人工评估

人工评估的方式是随机抽取 60 份转换前后的病历，通过独立的 3 人对其进行打分，分值为 1 到 5，数值越大代表效果越好。

如表 2 所列，通过人工评估，DualRL2.0 和 DualRL 在内容保留上近乎一致，但前者在风格转换程度和流畅性上都有所提高。虽然 DualRL2.0 在训练过程中加入了流畅性指导策略，但仍无法兼顾 3 个指标，在流畅性上还有提升的空间。

表 2 人工抽样 60 份病历的评分

Table 2 60 medical records score by artificial sample

Model		Metric		
		Style	Content	Fluency
DualRL[16]	Person1	3.712	4.023	2.313
	Person2	4.567	4.126	2.227
	Person3	4.027	3.899	2.441
	Mean	4.102	4.016	2.327
DualRL2.0	Person1	3.875	4.223	2.568
	Person2	4.532	4.138	2.346
	Person3	4.331	3.684	2.546
	Mean	4.246	4.015	2.487

(2)自动评估

自动评价同样采用随机抽取 60 份转换前后的病历。使用预训练的风格判别模型对风格转换的强度进行评估，评估指标为为 60 个样本中正确转换成目标风格的比率。本文通过困惑度(Perplexity，PPL)对生成的句子进行打分，困惑度越大代表模型的流畅性越差。

如表 3 所列，自动评估可以得出与人工评估方式同样的结论。DualRL2.0 和 DualRL 在内容一致性上几乎保持不变，但风格转换强度和流畅性都有小幅度的提高。

表 3 自动抽样 60 个病历的评分

Table 3 60 medical records score by automatic sample

Model	Score		
	Style	Content	PPL
DualRL	0.850	0.737	570.78
DualRL2.0	0.863	0.735	386.24

如表 4 所列，从抽取的样例中不难发现，经 DualRL 转换后，生成的句子和转换前的句子语义无明显变化，但会出现类似于机器翻译中的漏翻、错翻问题，如“胀痛”等症状，在转换时并未对“胀”进行有效翻译。DualRL2.0 模型可以对源句子中的信息进行有效保留，并且生成目标风格的文本。

表 4 样本病历转换示例

Table 4 Medical record sample transform example

病历源文本	缘于 4 个月前剑突下胀痛，为阵发性，程度不剧，可忍受，不伴肩背放射痛。术后可恢复，黄疸逐渐消退，予办理出院
DualRL	起因于 4 个月前剑突下疼痛，因为阵发性，程度不强烈，可以忍受，不伴随有肩背放射性疼痛。我手术然后恢复可以，黄疸渐渐减退，予办理出院
DualRL2.0	起因于 4 个月前剑突下胀痛，作为阵发性，程度不强，可以忍受，没有伴随有肩背放射性疼痛。我手术后恢复尚可，黄疸逐渐消退，给予办理了出院

对于比较专业的词汇如“剑突下胀痛”，本文模型选择不进行转换。DualRL2.0 在翻译过程中同样会产生一些坏样本，在后续清理转换后的语料时，仅对句子长度发生明显变化的样例进行清理，对其他语料进行保留。因此本文设计的 DualRL2.0 模型在牺牲一定流畅性和语义的情况下，完成了病历文本的风格转换。

1) https://github.com/paddlepaddle/models

2) https://pypi.org/project/pandas/0.25.3/

3) https://developer.nvidia.com/cuda-92-download-archive

4) https://github.com/pytorch/pytorch

4.3 训练和测试

如图 4 所示，本文利用 Pandas 工具[1)]将脱敏数据处理成 DataFrame 格式，从而分析不同种类标签的数量、数据缺失情况等。本文对数据进行去重、去停用词等。由于 ERNIE 不需要做分词处理，本文直接将字符级别作为输入，并依照 7∶2∶1 的比例构造训练集、验证集和测试集。本文使用 ERNIE 提供的词典，建立词汇到索引和索引到词汇的两组映射。

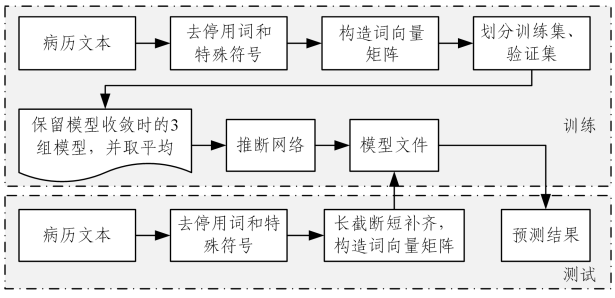

图 4 训练和测试流程图

Fig. 4 Workflow of training and test

本文在胶囊单元间迭代训练，模型根据验证集不断调整参数。当模型在验证集上的效果不再提升时，保留最后 3 组检查点的平均值作为最终的模型调用值，用于后续在部署模型时候被调用。

4.4 参数设置

本文分别使用 $BERT_{base}$，$ALBERT_{tiny}$，$ALBERT_{base}$[17]，$ERNIE_{1.0}$ 词向量进行对比，每种词向量的参数设定如表 5 所列。本文对 4 种常用的判别模型 CNN，RNN，RCNN，DPCNN 进行对比实验。在默认 5 个 Epoch 后，实验在验证集上如果达到 1 000 个 Batch 时效果无提升，则停止模型的训练。对照模型均使用 BERT/ERNIE 模型作为词汇的文本向量化方法。

表 5 词向量及参数

Table 5 Word vectors and parameters

Model	Parameter				
	Layer num	Embedding	Hidden	Vocabulary	Parameter sharing
$BERT_{base}$	12	768	768	21 128	False
$ALBERT_{tiny}$	4	128	312	21 128	True
$ALBERT_{base}$	12	128	768	21 128	True
$ERNIE_{1.0}$	12	768	768	18 000	False

本文的疾病辅诊模型继续采用 Adam 作为优化器，学习率为 $5*10^{-5}$，卷积核尺寸为(2,3,4)，路由次数为 3。为了能够公平且准确地评价不同算法的优劣，本文分别使用准确率(P)、召回率(R)和 F1 作为实验效果的有效性评估。F1 评价指标能较好地衡量算法的分类结果，且 F1 值越高说明算法的分类效果越好。

4.5 实验结果与分析

(1)基于电子病历文本的疾病分类

本文以 BERT 作为词向量的常用文本分类模型，并以此为对照。从表 6 中不难发现，单独的 BERT 即可达到较高的准确率。使用 BERT/ERNIE 作为词向量输入到其他文本分类模型，分类效果并未得到提高。其原因可能是 BERT 和 ERNIE 中含有众多的训练参数，而后续的 CNN/RNN/Capsule 层参数量过少，对结果的提升效果不明显，甚至会干扰训练结果。ERNIE 模型在病历文本数据集上可以获得比 BERT 更好的效果。虽然 ERNIE 作为词向量直接输入到胶囊单元时，效果并未提升，但添加抽取 N 元特征的 CNN 层之后，$F1$ 值有 1.36%～3.68%的提升，说明胶囊网络对高层特征的信息提取能力更强。

表 6 模型测试效果

Table 6 Model testing results

Model	$F1$/%
BERT/ERNIE	85.24/85.53
(BERT/ERNIE)+CNN	84.76/84.85
(BERT/ERNIE)+RNN	84.85/84.92
(BERT/ERNIE)+RCNN	84.62/84.83
(BERT/ERNIE)+DPCNN	84.77/85.02
$ALBERT_{tiny}$	83.21
$ALBERT_{base}$	84.67
ERNIE+Capsule	85.46
ERNIE+CNN+Capsule	86.89

(2)基于病情自述文本的疾病分类

风格转换后生成的文本虽然在风格转换强度和内容一致性上取得了较好的效果，但是在生成语言流畅性上有待提高。由于利用神经网络进行文本分类的过程中，模型会不断抽取对分类有关键影响的强特征，因此本文考虑是否可以不进行风格转换，直接用电子病历文本进行训练，用病情自述文本进行测试。故本文的测试模型分别选用基于原始电子病历和转换后的风格自述文本训练的 ERNIE+CNN+Capsule 组合模型。

本文取 60 份电子病历的症状特征，使用病情自述风格描述后输入到测试模型，每份电子病历的疾病类别作为病情自述风格文本的类别。DualRL2.0 模型可以对源句子中的信息进行有效保留，并且生成目标风格的文本。而对于比较专业的词汇如“剑突下胀痛”，该模型选择不进行转换。实验结果如表 7 所列。由于进行测试的文本和训练文本风格差距较大，模型在验证集上准确率有待提高。但利用经 DualRL2.0 模型风格转换文本训练出的模型，可以在一定程度上减缓这个问题。如何进一步缩小两种数据分布不同带来的影响，将是下一步研究的重点。

表 7 是否使用风格迁移的结果对照

Table 7 Result comparison of using style transfer or not

	Without style transfer	style transfer
$F1$	0.632	0.7

4.6 泛化性评估

本文将在开源中文 THUCNews 数据集[2)]上进行泛化性评估。该数据集源自新浪在 2005－2011 年间的新闻数据，包含 74 万篇新闻文档。本文从该数据集中抽取 20 万条标题样例。为了消除类别不均衡的影响，每个单独类别的数据都固定在 2 万条。实验中数据集、验证集、测试集的比例为9∶0.5∶0.5。对比模型的实现及参数设置与基于 BERT 的中文文本

1) https://pandas.pydata.org/

2) http://thuctc.thunlp.org/

分类[1]一致。实验同样使用 F1 作为评估指标，泛化性实验结果如表 8 所列。

表 8 组合模型在 THUCNews 数据集上的测试效果

Table 8 Results of combination model on THUCNews dataset

Model	F1/%
BERT	94.83
ERNIE	94.61
(BERT/ERNIE)+CNN	94.44/94.45
(BERT/ERNIE)+RNN	94.57/94.54
(BERT/ERNIE)+RCNN	94.51/94.50
(BERT/ERNIE)+DPCNN	94.47/94.60
(BERT/ERNIE)+Capsule	94.52/94.54
ERNIE+CNN+Capsule	94.95

从表 8 中不难发现，在类似 THUCNews 这种覆盖较多领域的中文数据集上，BERT 模型和 ERNIE 模型的表现差距较小，两者可以依靠参数的调节来超越彼此。在此基础上使用 ERNIE 提取字符级别语义特征，使用 CNN 提取 N 元特征，再使用胶囊网络对特征进行聚类的组合模型取得了比单独使用 BERT/ERNIE 更好的效果，由此证明了该组合模型是高效的。

结束语 本文将词向量和文本判别技术应用到电子病历的语义解析任务中，针对 BERT 训练过程中未加入实体和关系信息的局限性，采用预训练的 ERNIE 模型对文本特征进行准确表达。在 CNN 模型提取 N 元特征后，使用胶囊单元对特征进行聚类，以更好地捕获文本的高层语义特征，同时减少对数据量的需求。本文分别针对电子病历和病情自述文本进行疾病辅助诊断模型评估。组合模型在真实的电子病历数据集上具有良好的判别效果。此外，本文训练了电子病历文本到病情自述文本的风格转换模型。针对任务中训练数据不足和传统文本分类模型在疾病数据会出现泛化性差的现象，本文在开源中文数据集上对胶囊网络模型进行了泛化性评估，证明了该模型能很好地应用于疾病文本分类。

在今后的研究中，将尝试融合病灶图像和文本信息，并且加入诸如检验指标等其他特征信息，以完善疾病辅助诊断的量化分析，进一步细化或精化分类，提高辅助诊断的准确率。

参 考 文 献

[1] 国卫办医发〔2017〕8 号. 关于印发电子病历应用管理规范(试行)的通知[OL]. http://www.nhc.gov.cn/mohwsbwstjxxzx/s8553/201702/fb49f9487d884645b7247218b764bba3.shtml.

[2] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[C]// Neural Information Processing Systems. 2012:1106-1114.

[3] ZHANG Y Q, GU D Y. Review of Computer Aided Diagnosis for Parkinson's Tremor and Essential Tremor[J]. Computer Science, 2019, 46(7):22-29.

[4] CHEN D Y, ZHAO H, ZHANG X. Aided Diagnosis Method for Diseases Based on the Domain Semantic Knowledge Base[J]. Journal of Software, 2020, 31(10):3167-3183.

[5] HOCHREITER S, SCHMIDUBER J. Long short-term memory [J]. Neural Computation, 1997, 9(8):1735-1780.

[6] KIM Y. Convolutional neural networks for sentence classification[C]// Conference on Empirical Methods in Natural Language Processing. 2014:1746-1751.

[7] ZHANG Z Y, HAN X, LIU Z Y, et al. ERNIE: Enhanced language representation with informative entities[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019:1441-1451.

[8] SABOUR S, FROSST N, HINTON G E. Dynamic routing between capsules[C]// Neural Information Processing Systems. 2017:3856-3866.

[9] LI F F, PERONA P. A Bayesian hierarchical model for learning natural scene categories[C]// Computer Vision and Pattern Recognition. 2005:524-531.

[10] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[C]// International Conference on Learning Representations (Workshop Poster). Scottsdale, Arizona, USA, 2013.

[11] CHEN T Q, GUESTRIN C. XGBoost: a scalable tree boosting system[C]// Knowledge Discovery and Data Mining. 2016:785-794.

[12] PETERS M E, NEUMANN M, IYYER M, et al. Deep contextualized word representations[C]// The North American Chapter of the Association for Computational Linguistics. 2018:2227-2237.

[13] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pretraining of Deep Bidirectional Transformers for Language Understanding [C]// The North American Chapter of the Association for Computational Linguistics. 2019:4171-4186.

[14] ZHANG H, GOODFELLOW I J, METAXAS D N, et al. Self-Attention Generative Adversarial Networks[C]// International Conference on Machine Learning. 2019:7354-7363.

[15] ZHU J Y, PARK T, ISOLA P, et al. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks[C]// International Conference on Computer Vision. 2017:2242-2251.

[16] LUO F L, LI P. A dual reinforcement learning framework for unsupervised text style transfer[C]// International Joint Conference on Artificial Intelligence. 2019:5116-5122.

[17] LAN Z Z, CHEN M D, GOODMAN S, et al. ALBERT: A Lite BERT for self-supervised learning of language representations [C]// International Conference on Learning Representations. 2020.

FAN Hong-jie, born in 1984, Ph.D, lecturer. His main research interests include data exchange and knowledge graphs.

YE Song-tao, born in 1983, Ph.D, associate professor. His main research interests include truth discovery, data analysis and data mining.

(责任编辑：柯颖)

[1] https://github.com/649453932/