BERT 4EVER for SMP2020-EWECT

作者：林楠铠, 朱昆睿，刘皓楠，蒋盛益

广东外语外贸大学

摘要

在 SMP2020-EWECT 微博情绪分类任务中，我们利用文本替换、回译、伪标签多种策略对数据进行增强与扩充，并基于多个 SOTA 预训练模型进行对比实验，采用不同方法增强的数据在效果最佳的预训练模型上进行微调，得到多个模型并进行软融合。在最终的评测数据中，我们的模型具有较好的鲁棒性，最终指标得分为 0.7346，排名第四。

关键词：自然语言处理，社会媒体情绪分类，数据增强，预训练模型

1. 引言

社交媒体平台上的文本情绪识别一直是自然语言处理研究重点之一。微博作为重要的中文社交媒体平台，及时分析微博文本的情绪具有重要的社会价值。SMP2020-EWECT 发布了微博情绪分类任务，提供疫情微博以及通用微博两类文本数据用于评测，其中疫情微博训练集共有 8606 条，通用微博训练集共有 27768 条，每条微博被标注为以下六个类别之一：无情绪（neural），积极（happy），愤怒（angry），悲伤（sad），恐惧（fear），惊奇（surprise），数据标签分布情况如表 1 所示。我们利用自然语言处理技术对微博文本进行情绪分类。本报告将介绍我们在本次评测中所采用的方案，下文将介绍：1. 采用的模型；2. 三类数据增强的方法；3. 模型微调方法；4. 模型融合方法。

表 1 数据标签分布情况

数据集	标签数目					
	angry	neural	happy	sad	surprise	fear
通用数据集	8344	5749	5379	4990	2086	1220
疫情数据集	1322	1460	4423	649	197	555

2. 模型及方法介绍

2.1 Roberta

自 BERT[1]刷新自然语言处理多项任务以来，业界又提出了各种不同的预训练语言模型。在训练阶段，这类模型利用在无标签数据上构建特定训练目标的方法学习语言的通用特征，在使用时，模型常被用作语言特征的抽取器，获取句子编码。在比赛过程中，我们最终使用的语言特征抽取器 RoBERTa_WWM_Ext[2]也是 BERT 的衍生模型，该模型由三部分组成：

RoBERTa 代表的是预训练语言模型的名字，这种模型由 24 层的 Transformer[3]组成，每层以 1024 为隐层大小、采用 16 头的自注意力机制，与 BERT 模型结构基本一致。RoBERTa 模型在 BERT 训练方法的基础上做了一系列的修改，其中主要包括：采用动态的词语遮蔽训练 Mask LM 任务，去除 Next Sentence Prediction 任务，采用更大的 mini-batch 量级和更多的训练迭代次数等，最终让预训练模型参数得到更充分的优化。

第二部分的 WWM[4]全称是 Whole Word Masking(全词覆盖)，这是一种对 Mask LM 任务训练方法的优化。BERT 在英文分词处理过程中，单词在经过 Word Piece 分词器后会被切分为多个部分，而在中文方面，模型接收字符级别的分词句子输入。在 WWM 被提出之前，Mask LM 任务让语言模型预测被 Mask 符号随机覆盖掉的分词片段。全词覆盖策略用多个 Mask 符号覆盖完整的单词或中文词语，让模型能够更好地捕捉到分词特征。

最后的 Ext 部分是 Extended Data 的缩写，研究机构通过引入更多的中文语料扩充语言模型训练数据，使预训练模型有更好的信息抽取能力。

2.2 基于文本替换的伪数据生成

在正式进入训练阶段前，我们分别采用了<PAD>符号随机遮蔽与同义词替换两种数据变换策略，生成增强数据。

第一种方式以一定的概率[1]，将输入句子中的词语替换成用于补齐 BERT 预训练模型句子的<PAD>符号（如表 2 所示），这种数据增强方式不仅增加了模型的训练数据，而且减少了模型对某些特定词语的依赖，可以较好地提升模型的泛化能力。另一种基于文本替换的伪数据生成策略是基于同义词表[2]实现（如表 2 所示），可以在不改变句子原意的基础上增加训练数据。处理过程中，我们采用

[1] 最终模型采用的替换概率为 0.3.

[2] https://github.com/Keson96/SynoCN

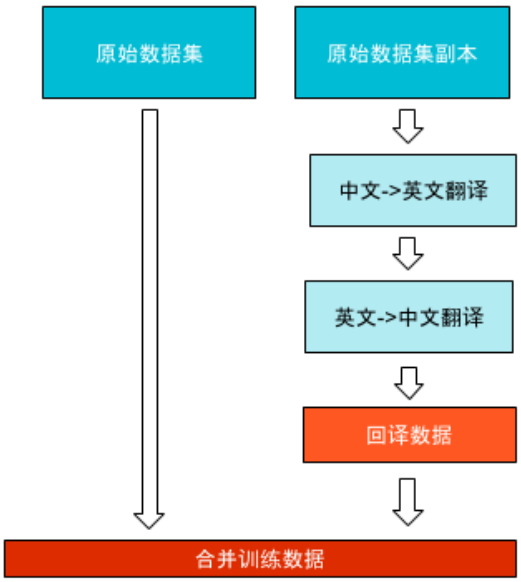
Thulac[5][6]进行中文分词操作。

表 2 文本替换示例

原句	[CLS] 对 着 这个 症状 ， 没 病 的 都 害怕 [SEP]
<PAD>替换	[CLS] 对 着 这个 <PAD> ， 没 病 的 都 害怕 [SEP]
同义词替换	[CLS] 对 着 这个 症状 ， 没 病 的 都 畏惧 [SEP]

2.3 基于回译法的伪数据生成

除了基于文本替换生成伪数据进行数据增强，我们还采用回译法进行数据增强。基于回译法的伪数据生成方法主要利用翻译句子和原始句子之间的差异来丰富训练数据的表达方式。具体的实现过程如下图一所示。



图一 回译法流程图

我们首先将原始数据集利用谷歌中译英翻译接口转换成英文句子，再把得到的英文句子翻译回中文。通过这两个步骤，我们可以得到大量的句子的同义转换。最终，我们合并原始数据集与回译数据集，得到数据量为原来两倍的训练数据集用于模型训练。

2.4 基于伪标签预测的伪数据生成

我们通过采用原始数据集训练一个情绪分类模型，获取疫情期间网民情绪识别情感极性数据集[3]，对 58286 条数据进行伪标签预测，将生成的伪标签作为样

[3] https://www.datafountain.cn/competitions/423/datasets

本的真实标签，加入到原始数据集中重新训练新的模型。由于原始数据集中“无情绪（neural）”与“积极（happy）”样本占比较大，通过原始数据集训练出来的模型将其他标签错误预测成这两类标签的可能性越大，产生的伪标签数据中这两个标签占比较大，因此针对这两类情绪标签进行过滤，分别各选取 10000 条伪标签，最终保留的伪标签数据集分布如表 3 所示。

表 3 伪标签数据标签分布情况

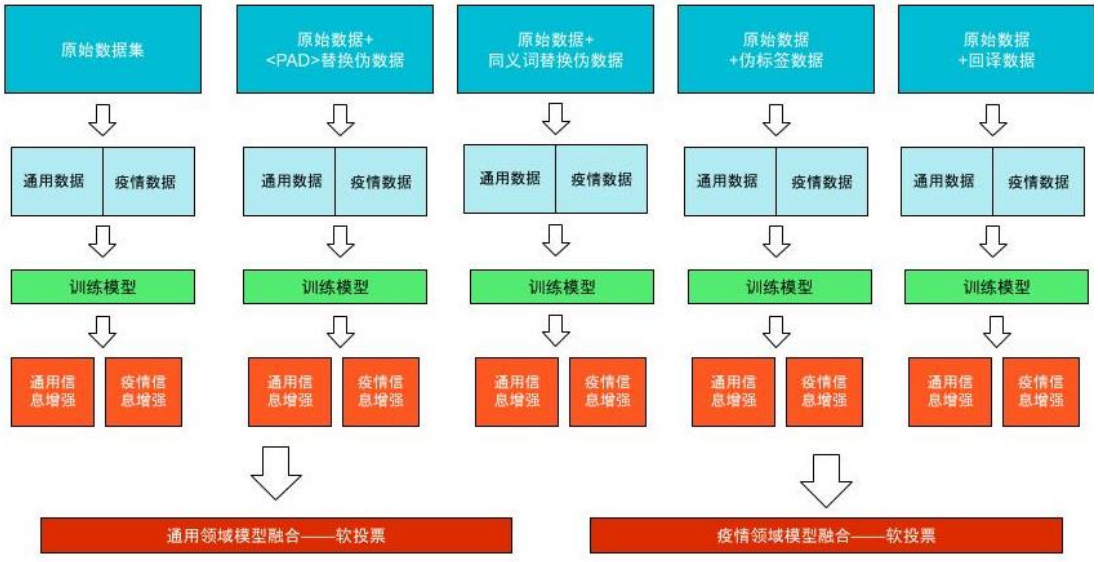
情绪	angry	neural	happy	sad	surprise	fear
标签数目	8156	10000	10000	4951	1564	2515

2.5 领域信息增强的 fine-tune 训练

我们将通用数据和疫情数据两个领域的训练集混合形成一个数据集训练模型，在训练之后生成的模型包含了通用领域的信息和疫情领域的信息，为了让模型在各个领域表现效果更好，我们将模型针对不同领域的数据进行了再次的微调，增强每个领域的信息。

2.6 多模型融合

我们分别训练了五个基于不同数据的 RoBERTa_WWM_Ext 模型，再分别对每个模型进行通用领域信息/疫情领域信息增强后，将增强后的五个通用领域的模型进行软投票融合预测通用领域的数据，增强后的五个疫情领域的模型进行软投票融合预测疫情领域的数据，系统框架如图二所示。



图二 框架流程图

3. 实验设置

本论文基于 PaddlePaddle[4]与 PaddleHub[5]实现社交媒体情绪分类模型，训练阶段采用五折交叉验证的方式，训练过程的详细设置如表 4 所示。

表 4 模型参数设置

参数	参数值
Weight Decay	0.01
Warmup Proportion	0.1
文本最长长度	128
Epoch	3
学习率	4e-6

4. 实验结果及分析

我们将通用数据和疫情数据两个领域的训练集混合形成一个数据集训练模型，对比了 RoBERTa-WWM-Ext、BERT-WWM-Ext、Erine 与 SKEP 四个模型的表现性能，结果如表 5 所示，在通用数据和疫情数据两个领域的数据上，RoBERTa-WWM-Ext 在验证集和测试集上表现效果均达到最佳效果，故采用 RoBERTa-WWM-Ext 模型作为基模型。

表 5 不同模型试验结果

模型	通用数据		疫情数据	
	验证集平均F_1值	测试集F_1值	验证集平均F_1值	测试集F_1值
RoBERTa-WWM-Ext	**0.7664**	**0.7821**	**0.6619**	**0.6685**
BERT-WWM-Ext [1]	0.7553	0.7610	0.6485	0.6533
SKEP [7]	0.7492	0.7764	0.6482	0.6288
Erine [8]	0.7301	0.7623	0.6104	0.6081

我们对比了不同的训练方式：将两个领域的训练数据合并成一个训练集并训练一个模型预测两个领域的数据（以下称“策略一”）与将两个领域的训练数据独立训练两个模型进行预测（以下称“策略二”）两种策略的效果，结果如表 6 所示，结果显示，策略一的效果优于策略二。

表 6 数据合并操作的对比实验

策略	通用数据		疫情数据	
	验证集平均F_1值	测试集F_1值	验证集平均F_1值	测试集F_1值
数据合并	**0.7664**	**0.7821**	**0.6619**	**0.6685**
分领域训练	0.7622	0.7805	0.6079	0.6633

[4] https://github.com/PaddlePaddle/Paddle
[5] https://github.com/PaddlePaddle/PaddleHub

我们针对基于文本替换的伪数据生成策略、基于回译法的伪数据生成策略、基于伪标签预测的伪数据生成策略进行了探究，实验结果如表 7 所示。

表 7 不同策略的对比实验

策略	通用数据		疫情数据	
	验证集平均F_1值	测试集F_1值	验证集平均F_1值	测试集F_1值
无策略	0.7664	0.7821	0.6619	0.6685
<PAD>替换	0.7673	0.7792	0.6698	0.6657
同义词替换	0.7643	0.7812	0.6627	0.6725
回译法	0.7685	0.7745	0.6630	0.6650
伪标签	0.7689	0.7776	0.6662	0.6768

我们进一步探究了领域信息增强的 fine-tune 策略的有效性，结果如表 8 所示，在疫情数据中，该策略在验证集和测试集上效果均有明显的提升，而在通用数据上，无数据生成策略、<PAD>替换策略与同义词替换策略三个模型在测试集均有一定下降。

表 8 领域信息增强的 fine-tune 实验对比

策略		通用数据		疫情数据	
		验证集平均F_1值	测试集F_1值	验证集平均F_1值	测试集F_1值
无策略	无信息增强	0.7664	0.7821	0.6619	0.6685
	信息增强	**0.7682**	0.7801	**0.6753**	**0.6732**
<PAD>替换	无信息增强	0.7673	0.7792	0.6698	0.6657
	信息增强	**0.7695**	0.7775	**0.6709**	**0.6845**
同义词替换	无信息增强	0.7643	0.7812	0.6627	0.6725
	信息增强	**0.7681**	0.7767	**0.6708**	**0.6748**
回译法	无信息增强	0.7685	0.7745	0.6630	0.6650
	信息增强	**0.7686**	**0.7809**	**0.6695**	**0.6802**
伪标签	无信息增强	0.7689	0.7776	0.6662	0.6768
	信息增强	**0.7713**	**0.7819**	**0.6780**	**0.6810**

结合上述实验结果，我们发现当验证集效果提升时测试集效果会出现波动，因此验证集与测试集不属于同一数据分布，因此，我们综合考虑验证集与公榜测试集的结果，结果如表 9 所示，最终我们选择在验证集与测试集平均表现效果最好的模型（“信息增强下五个模型软融合”），平均值为 0.7244，最终私榜指标得分为 0.7346，排名第四。

表 9 模型融合结果

模型	通用数据		疫情数据		平均值
	验证集平均F_1值	测试集F_1值	验证集平均F_1值	测试集F_1值	
无信息增强下五个模型软融合	0.7671	0.7831	0.6647	0.6789	0.7235
信息增强下五个模型软融合	0.7691	0.7799	0.6729	0.6756	**0.7244**

4. 总结

本次测评给我们队伍提供了一个很好的机会验证各种不同的数据增强策略的有效性，最终系统通过模型融合的方式集成各种数据增强策略。此外，结合比赛中包含两个数据集的特点，本项目采用的合并数据集训练与分领域微调的策略也在比赛后阶段为系统带来了一定程度的效果提升。在多种策略的辅助下，我们的模型具有很强的泛化性能，从而在最终阶段表现较优，最终指标得分为 0.7346，排名第四。

相关文献

[1] Jacob D, Ming-Wei C, Kenton L, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[J]. 2019

[2] Yinhan L, Myle O, Naman G, et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach[J]. 2020

[3] Ashish V, Noam S, Niki P, et al. Attention Is All You Need[J]. 2018

[4] Cui Y, Che W, Liu T, et al. Pre-Training with Whole Word Masking for Chinese BERT[J]. 2019.

[5] Li Z , Sun M . Punctuation as Implicit Annotations for Chinese Word Segmentation[J]. Computational Linguistics, 2009.

[6] Maosong Sun, Xinxiong Chen, Kaixu Zhang, Zhipeng Guo, Zhiyuan Liu. THULAC: An Efficient Lexical Analyzer for Chinese. 2016.

[7] Tian H, Gao C, Xiao X, et al. SKEP: Sentiment Knowledge Enhanced Pre-training for Sentiment Analysis[J]. 2020.

[8] Sun Y, Wang S, Li Y, et al. ERNIE: Enhanced Representation through Knowledge Integration[J]. 2019.