计算机科学
COMPUTER SCIENCE

融合语义解释和DeBERTa的极短文本层次分类

陈昊飏, 张雷

引用本文

陈昊飏, 张雷. 融合语义解释和DeBERTa的极短文本层次分类[J]. 计算机科学, 2024, 51(5): 250-257.

CHEN Haoyang, ZHANG Lei. Very Short Texts Hierarchical Classification Combining Semantic Interpretation and DeBERTa [J]. Computer Science, 2024, 51(5): 250-257.

相似文章推荐（请使用火狐或 IE 浏览器查看文章）

Similar articles recommended (Please use Firefox or IE to view the article)

多线路信息融合的公交车行程时间预测算法

Bus Travel Time Prediction Algorithm Based on Multi-line Information Fusion

计算机科学, 2019, 46(11): 222-227. https://doi.org/10.11896/jsjkx.180901764

图像超分辨率全局残差递归网络

Global Residual Recursive Network for Image Super-resolution

计算机科学, 2019, 46(6A): 230-233.

微波探火及其在低空平台中的系统设计

Microwave Fire Detection and System Design Based on Low-altitude Aircraft Platform

计算机科学, 2010, 37(3): 141-143.

非负矩阵分解在标签语义分析中的应用

Application of Non-negative Matrix Factorization in Tag Semantics Analysis

计算机科学, 2010, 37(4): 171.

分层特征计算和错误控制的层次分类方法

Hierarchical Classification Approach of Hierarchical Feature Selection and Error Control

计算机科学, 2010, 37(10): 165-168.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.231100134

融合语义解释和 DeBERTa 的极短文本层次分类

陈昊飏　张　雷
南京大学计算机软件新技术全国重点实验室　南京 210023
(chen-haoyang@qq.com)

摘　要　文本层次分类在社交评论主题分类、搜索词分类等场景中有重要应用,这些场景的数据往往具有极短文本特征,体现在信息的稀疏性、敏感性等中,这对模型特征表示和分类性能带来了很大挑战,而层次标签空间的复杂性和关联性使得难度进一步加剧。基于此,提出了一种融合语义解释和 DeBERTa 模型的方法,该方法的核心思想在于:引入具体语境下各个字词或词组的语义解释,补充优化模型获取的内容信息;结合 DeBERTa 模型的注意力解耦机制与增强掩码解码器,以更好地把握位置信息、提高特征提取能力。所提方法首先对训练文本进行语法分词、词性标注,再构造 GlossDeBERTa 模型进行高准确率的语义消歧,获得语义解释序列;然后利用 SimCSE 框架使解释序列向量化,以更好地表征解释序列中的句子信息;最后训练文本经过 DeBERTa 模型神经网络后,得到原始文本的特征向量表示,再与解释序列中的对应特征向量相加,传入多分类器。实验遴选短文本层次分类数据集 TREC 中的极短文本部分,并进行数据扩充,最终得到的数据集平均长度为 12 词。多组对比实验表明,所提出的融合语义解释的 DeBERTa 模型性能最为优秀,在验证集和测试集上的 Accuracy 值、F1-micro 值、F1-macro 值相比其他算法模型有较大的提升,能够很好地应对极短文本层次分类任务。

关键词:极短文本;层次分类;语义解释;DeBERTa;GlossDeBERTa;SimCSE
中图分类号　TP391.1

Very Short Texts Hierarchical Classification Combining Semantic Interpretation and DeBERTa

CHEN Haoyang and ZHANG Lei
State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China

Abstract　Text hierarchy classification has important applications in scenarios such as social comment topic classification and search term classification. The data in these scenarios often exhibits short text features, which is reflected in the sparsity and sensitivity of information. It poses great challenges for model feature representation and classification performance. The complexity and associativity of the hierarchical label space further exacerbate the difficulties. In view of this, a method fusing semantic interpretation and DeBERTa model is proposed, and the core idea of the method is as follows: introducing the semantic interpretation of individual words or phrases in specific contexts to supplement and optimize the content information acquired by the model; combining the disentangled attention and enhanced mask decoder of the DeBERTa model to better grasp the location information and improve the feature extraction ability. The method firstly performs grammatical disambiguation and lexical annotation on the training text, and then constructs the GlossDeBERTa model to perform semantic disambiguation with high accuracy to obtain the semantic interpreted sequence. Then the SimCSE framework is used to make the interpreted sequence vectorized to better characterize the sentence information in the interpreted sequence. Finally, the training text passes through the DeBERTa model neural network to get the feature vector representations of the original text, which is then summed up with the corresponding feature vector in the interpreted sequence, and passed into the multi-class classifier. The experiments select the very short text portion of the short text hierarchical categorization dataset TREC and expand the data, resulting in a dataset with an average length of 12 words. Multiple sets of comparison experiments show that the DeBERTa model proposed in this paper with fused semantic interpretation has the best performance, and the Accuracy, F1-micro, and F1-macro values on the validation and test sets are much better than other algorithmic models, which can well cope with the task of hierarchical categorization of very short texts.

到稿日期:2023-11-20　返修日期:2024-02-22
基金项目:国家自然科学基金(62192783,62376117);南京大学软件新技术与产业化协同创新中心
This work was supported by the National Natural Science Foundation of China(62192783,62376117) and Collaborative Innovation Center of Novel Software Technology and Industrialization at Nanjing University.
通信作者:张雷(ZhangL@nju.edu.cn)

Keywords Very short text, Hierarchical classification, Semantic interpretation, DeBERTa, GlossDeBERTa, SimCSE

1 引言

互联网与社交媒体的迅速发展带来了很多层次分类的应用场景,如平台和研究人员需要对社交媒体数据进行精细的层次主题分类和情感分析,搜索引擎也需要准确理解用户的简短查询词并进行层次分类。但在这些场景中,人们越来越多地使用极短文本,如微博、朋友圈、搜索词等,这给层次分类任务带来了新的挑战。

极短文本层次分类的核心难点在于:极短文本具有比短文本更强的文本长度限制,平均长度通常只有十几个字,蕴含的信息量十分有限,很难获得足够的上下文来准确理解文本含义;层次标签空间十分复杂,存在歧义、多样性和关联性。此外,极短文本还具有高度敏感性,每个单词或字符都可能承载重要信息,一些微小的变化或错误会对分类结果产生显著影响。

就现有的层次分类方法而言,Siddhartha 等[1]提出了一种迁移学习策略(HTrans),通过上层分类器的参数初始化下层分类器,以提升分类性能;Zhou 等[2]提出了一种层次感知的全局模型(HiAGM),层次结构公式化为有向图,引入层次结构编码器来建模标签相关性,进而优化层次分类表现;Chen 等[3]提出了层次感知的标签语义匹配网络(HiMatch),通过建模文本和标签的语义关系、引入联合嵌入损失和匹配学习损失,捕捉不同层次标签之间的语义匹配关系。但是这些层次分类方法在同时处理极短文本和层次分类时往往具有一定的局限性,会受到极短文本信息稀缺、高敏感性、标签空间复杂性等的影响。由于信息有限且标签复杂多样,一些层次分类方法在建模极短文本与标签空间之间的关系时会遇到困难,并且此时引入标签信息所产生的噪声很可能会被放大,对分类结果产生不利影响。

就现有的短文本分类方法而言,首先传统的机器学习方法往往基于统计与概率,如 Huang 等[4]使用词袋模型[5](Bag-Of-Word,BOW)从词向量中提取词频矩阵,结合 TF-IDF 算法提取文本特征,进行短文本分类。Chen 等[6]提出了一种改进的基于潜在狄利克雷分配主题模型和 K-最近邻算法的短文本分类方法,生成的概率主题可以使文本更加关注语义。虽然传统机器学习的方法取得了一定成效,但是仍面临着诸多限制,对于极短文本层次分类,它们难以建模复杂的标签空间,也无法很好地应对极短文本的信息稀缺性、高敏感性、歧义性等。

深度学习则是当前短文本分类的主流方法,深度学习方法基于各种神经网络模型结构,如循环神经网络(Recurrent Neural Networks,RNN)、BERT[7]模型、RoBERTa[8]模型等,有更强大的上下文理解能力和学习特征表示的能力,并具有一定的泛化能力和适应性。Chen 等[9]通过对抗训练提升模型的健壮性,并利用多层双向长短期记忆网络提取语义信息,有效地提高了短文本分类的准确性;Hu 等[10]结合了 CNN(Convolutional Neural Network)与支持向量机,提高了短文本分类的泛化能力;Lyu 等[11]使用 CNN 学习单词的重要性权重矩阵,结合双向 RNN 获取单词表示,在多种文本分类数据集上取得了不错的效果;Yang 等[12]融合了 BERT 模型和 TextCNN,实现了临床试验筛选标准的短文本分类。这些方法在处理普通短文本分类任务时具有不错的表现,但极短文本的高敏感性、歧义性、信息稀缺性以及复杂标签空间依旧会对它们的分类性能造成较大影响。

在整个文本分类领域,还有一些针对拓展文本信息的研究,它们往往结合知识图谱或相关知识库,试图提供额外的知识链接和补充。Liu 等[13]在文本表示和层次标签学习过程中整合了知识图谱,提出了知识增强的分类模型 K-HTC,在 WOS 这类中长文本层次分类任务中取得了不错的效果。Li 等[14]提出了 KAe-RCNN 短文本分类模型,该模型结合了知识感知和双重注意力机制,能够挖掘短文本的隐含语义,具有良好的分类效果。Hoppe[15]在零样本常规文本分类中引入知识图谱作为额外的模态,把未见类别的外部知识与文本连接,并利用知识图谱的显式知识进行分类,优化性能。Zheng 等[16]提出了 SimpleSTC,从外部大型语料库中构建单词关联图来弥补信息的缺乏,并学习文本图来处理标记数据的缺乏,提高了短文本分类推理的速度和效果。这些方法通过补充额外知识来提高文本分类的能力,为我们提供了一定借鉴。但是面对极短文本时,由于知识图谱等涉及实体关系等复杂网状结构,引入的知识有一定的波动性、间接性,而极短文本的特点又很可能导致补充知识中的噪声和波动更加突出。同时,引入的知识可能不够直接、准确或稳定,以致于会干扰模型的特征提取和判断过程。

因此,针对极短文本层次分类的困境,本文提出了一种融合语义解释和 DeBERTa 模型的极短文本层次分类方法。该方法的核心理念是补充单词语境信息和充分利用位置信息。一方面,引入各个字词或词组在具体语境下的直接语义解释,在保证高效率性和高正确性的前提下,较为稳定、直接和准确地实现对极短文本的信息补充,改善信息特征不足和歧义性的问题。另一方面,利用 DeBERTa-V1[17]模型更好地把握文本位置信息,提升特征信息提取能力,其中注意力解耦机制可以同时考虑上下文字词的内容和相对位置,增强的掩码解码器可以补充绝对位置信息。实验遴选短文本层次分类数据集 TREC 中的极短文本部分,并进行大量人工扩充。最终,多组对比实验和消融实验表明,补充单词语境信息和利用位置信息都能提高极短文本层次分类的任务表现,同时,本文提出的融合语义解释的 DeBERTa 模型的综合表现最为优秀,验证集和测试集的 Accuracy 值、F1-micro、F1-macro 都高达 93% 左右,能够很好地处理极短文本层次分类任务。

2 模型结构

本文提出的融合语义解释与 DeBERTa 的模型结构如图 1 所示,主要包括解释序列生成、极短文本训练、解释

序列嵌入、层次分类这四大模块。本模型的核心流程如下：

1)针对原始输入的极短文本，依次进行语法分词、词性标注、词义消歧，得到解释序列文本。

2)极短文本输入经过 Tokenizer 分词后，进入 DeBERTa 神经网络，得到隐藏层状态向量。

3)将解释序列文本向量化，极短文本的隐藏层状态经过全连接层和激活函数后，与解释序列向量相加，传入多分类器。

4)利用多分类器扁平化层次标签空间，进行层次分类。

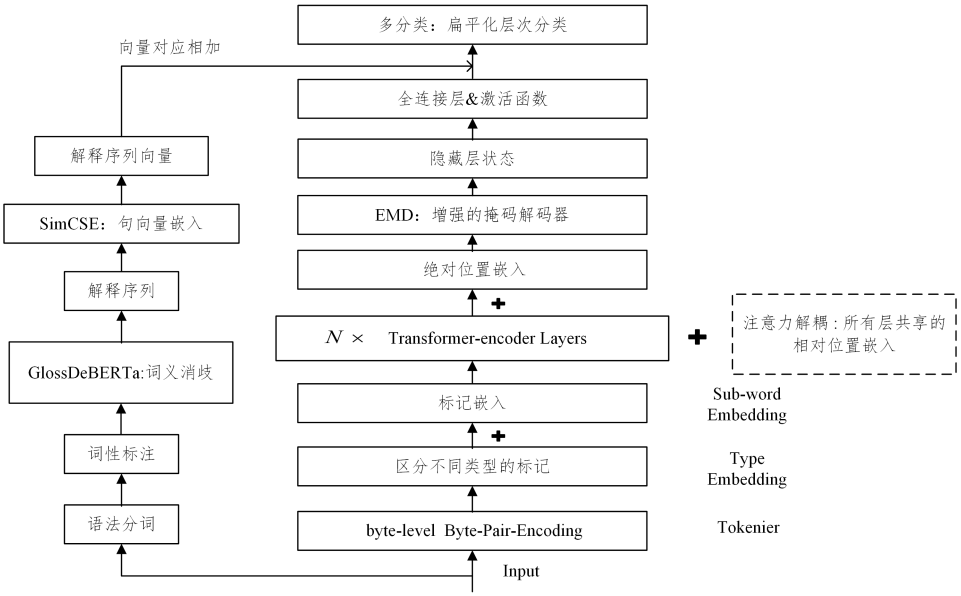


图 1 融合语义解释与 DeBERTa 模型的结构图

Fig. 1 Model structure fusing semantic interpretation with DeBERTa

2.1 解释序列生成

目前，语法分词技术以及词性标注技术发展较为成熟，有着极高的准确率和便捷的调用方法，本模型调用 expert.ai 的 API 进行语法分词与词性标注，举例如图 2 所示。实验中，人工校对了 50 条随机数据的处理结果，正确率达 100%。

ADV	AUX	DET	NOUN
Why	does	the	tide
VERB	**CCONJ**	**VERB**	**PUNCT**
rise	and	fall	?

图 2 语法分词与词性标注

Fig. 2 Syntactic parsing and part-of-speech tagging

词义消歧(Word Sense Disambiguation，WSD)是一项具有挑战性的任务。传统方法的词义消歧准确率往往不超过 60%，如经典的 Lesk 方法[18]、Mona 等[19]提出的利用平行语料库标记意义的方法。但是，目前的 SOTA 模型的准确率已经接近 90%。如 EXTEND 模型引入了两个 Transformer-based 架构，把实体消歧当作文本抽取任务去实现[20]；Huang 等[21]提出了 GlossBERT 模型，核心思路是在有监督神经词义消歧系统中更好地利用义项知识(Gloss Knowledge)，构建了如表 1 所列的上下文-义项对，并对 BERT 训练微调。本模型便基于 GlossBERT 的核心思想，结合 Huang 等发布的义项数据集，对更强大的 DeBERTa 模型训练微调、改造线性层，构造了 GlossDeBERTa 模型，从而获得更高的词义消歧准确率。经过对 100 条数据的人工测试，消歧准确率高达 91%。然后利用 GlossDeBERTa 模型对经过语法分词和词性标注的极短文本进行词义消歧，获得具有弱监督形式的解释序列，举例如图 3 所示。

表 1 上下文-义项对

Table 1 Context-Gloss Pairs

Context-Gloss Pairs	Label
[CLS]How … since you "reviewed" the objectives … program? [SEP] reviewed: a variety show with topical … [SEP]	False
[CLS]How … since you "reviewed" the objectives … program? [SEP] reviewed: look at again; examine again [SEP]	True

"Why : the cause or intention underlying an action or situation, especially in the phrase 'the whys and wherefores' # does : engage in # the : the # tide : the periodic rise and fall of the sea level under the gravitational pull of the moon # rise : move upward # and : and # fall : move downward and lower, but not necessarily all the way # ? : ?"

图 3 解释序列

Fig. 3 Semantic interpretation

2.2 极短文本训练

极短文本首先经过 Tokenizer 分词层，利用字节级别的字节对编码(byte-level Byte-Pair-Encoding，BBPE)将文本分割成有意义的子词单元，BBPE 训练和推断的速度较快，不需要引入超出词汇表的标记；然后经过 Type Embedding 层，该嵌入层能够区分不同的标记，帮助模型理解句子之间的关系和上下文；接着进入 Sub-word Embedding 层，把字词转换为 768 维度的向量表示，它们可以用于捕捉单词的语义信息和上下文关系。之后，将利用 DeBERTa 的核心机制，即注意力解耦机制、增强的掩码解码器。

2.2.1 注意力解耦机制

在 DeBERTa 的注意力解耦机制中，每个单词都使用两个分离的向量来表示，分别编码单词的内容和位置，这样在计算单词对的注意力权重时，能同时考虑单词本义，以及单词在

语境中的相对位置关系。因此，单词对(i,j)的注意力权重可以通过内容和位置的解耦矩阵进行计算，如式(1)所示：

$$A_{i,j}=\{\boldsymbol{H}_i,\boldsymbol{P}_{i|j}\}\times\{\boldsymbol{H}_i,\boldsymbol{P}_{j|i}\}^{\mathrm{T}}=\boldsymbol{H}_i\boldsymbol{H}_j^{\mathrm{T}}+\boldsymbol{H}_i\boldsymbol{P}_{i|j}^{\mathrm{T}}+\boldsymbol{P}_{j|i}\boldsymbol{H}_j^{\mathrm{T}}+\boldsymbol{P}_{j|i}\boldsymbol{P}_{j|i}^{\mathrm{T}} \tag{1}$$

其中，$\boldsymbol{H}$ 是内容向量，$\boldsymbol{P}$ 是 i 相对于 j 的相对位置向量。式(1)说明了注意力权重最终由内容到内容、内容到位置、位置到内容、位置到位置 4 个部分的关系映射组成。但由于使用了相对位置嵌入向量，因此位置到位置的信息较少，便将其去除。

通过 Sub-word Embedding 层后的极短文本向量表示会继续进入堆叠的 Transformer -encoder[22] 层中，Transformer-encoder 结构的核心是多头注意力机制和前馈神经网络。而相对位置的嵌入向量会在每一层的 Transformer-encoder 中被作为输入，以便在自注意力机制中使用，这样利于在模型的不同层次上既能有效地利用字词相对位置信息，又能学习字词上下文含义的信息特征表示，进而提升应对极短文本层次分类的能力。

2.2.2 增强的掩码解码器

相对位置和绝对位置在模型的训练和预测过程中都扮演着重要的角色，在注意力解耦机制考虑相对位置和上下文含义的基础上，还需要引入绝对位置信息，以提供更完善的上下文信息，帮助模型理解空间模态，进行掩码语言预训练任务(Masked Language Modeling，MLM)。另一方面，原始 BERT 在预训练任务和下游微调阶段的不一致性，可能会影响下游任务表现。因此，为了解决上述问题，DeBERTa 使用了增强的掩码解码器(Enhanced Masked Decoder，EMD)，其结构如图 4 所示。

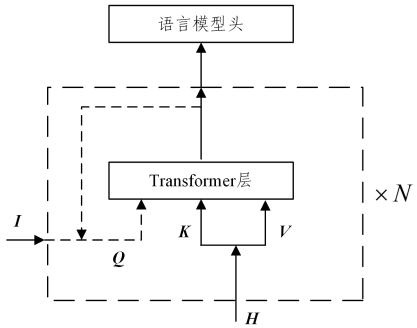


图 4 增强的掩码解码器

Fig. 4 Enhanced masked decoder

图 4 中，$\boldsymbol{H}$ 表示来自前一层 Transformer Layer 的输出；$\boldsymbol{I}$ 表示为解码所准备的任何信息，第一个 EMD 的 $\boldsymbol{I}$ 代表绝对位置信息，每个 EMD 层的输出是下一个 EMD 层的输入 $\boldsymbol{I}$；N 表示 EMD 的层数；$\boldsymbol{Q}$，$\boldsymbol{K}$，$\boldsymbol{V}$ 是注意力机制中的权重矩阵。通过这种方式，DeBERTa 能够在所有 Transformer 层中捕捉相对位置和字词内容信息，然后仅在 EMD 中将绝对位置作为补充信息，而不在前面去过度干涉模型学习，最终能够在模型训练和学习表征的过程中，充分利用极短文本的位置信息，优化性能。

2.3 解释序列嵌入

句向量表征指利用向量来表示句子特征信息的方法。传统的深度方法中，比较流行的是利用 word2vec 训练词向量、结合池化策略来表示句子向量，而在 Transformer 架构出现之后，各类模型凭借更强大的特征抽取能力占据了主流。为了更好地表征解释序列中的句子信息，本模型选择了目前的 SOTA 方法之一，即有监督的 SimCSE[23](Simple Contrastive Learning of Sentence Embeddings)框架。

有监督的 SimCSE 是一种简单的对比学习框架，对比学习的目标是最大化相似样本对的相似度，最小化不相似样本对的相似度，学习到更好的向量表示。有监督的 SimCSE 的核心算法是利用一些 NLI 自然语言推理数据集，构造形如(X_i,X_i^+,X_i^-)的前提假设文本、继承文本和对立文本对，然后利用 BERT 模型分别对其进行编码，得到特征结果 $h_i=f_\theta(x_i)$，$h_i^+=f_\theta(x_i^+)$，$h_i^-=f_\theta(x_i^-)$，训练目标的损失函数如式(2)所示：

$$l_i=-\log\frac{\mathrm{e}^{\mathrm{sim}(h_i,h_i^+)/\tau}}{\sum_{j=1}^{N}(\mathrm{e}^{\mathrm{sim}(h_i,h_j^+)/\tau}+\mathrm{e}^{\mathrm{sim}(h_i,h_j^-)/\tau})} \tag{2}$$

其中，τ 是温度超参数，sim 函数用于计算余弦相似度。

本文模型利用有监督的 SimCSE，对生成的解释序列进行句向量表征，提升了句子嵌入的特征提取效果，并且通过对比学习的优化目标缓解了各向异性问题，使得句向量能够更好地表征句子信息。然后，将解释序列向量与经过全连接层和 Tanh 激活函数的极短文本隐藏层状态向量相加，传入多分类层。在进行向量相加的过程中必须保证向量对齐，否则将会产生重大的负面影响。具体而言，先确定模型使用的数据集划分，然后经过上文操作，获得解释序列向量；在训练前将整体向量设置为模型类的一个属性，并且分别维护训练集和验证集对应的解释序列的下标属性；进入训练状态时，根据 batch size 的大小，在每个 batch 的训练中将 batch size 个解释序列向量与 batch size 个隐藏层状态向量对应相加，确保一一对应；当每轮训练后进入验证状态时，也是同理，同时还需要把解释序列下标属性归零，以便下一轮继续使用。

2.4 层次分类

对于层次标签空间，本文模型利用一个多分类器扁平化标签空间的层次结构，实现标签预测分类。其原因在于，极短文本信息量很可能少于复杂的标签空间信息，建模复杂标签空间或者补充标签知识容易引入噪声，并且极短文本的歧义性、敏感性会进一步放大干扰。经过实验发现，直接使用简单便捷的多分类器更适合极短文本层次分类。多分类器包括全连接层、softmax 函数和多元交叉熵损失函数(categorical cross-entropy)，softmax 函数、多元交叉熵损失函数如式(3)、式(4)所示：

$$\mathrm{Softmax}(x_i)=\frac{\mathrm{e}^{x_i}}{\sum_{c=1}^{C}\mathrm{e}^{x_c}} \tag{3}$$

其中，x_i 是第 i 个节点的输出，C 为分类的类别数量。

$$loss=-\frac{1}{M}\sum_{i=1}^{M}\sum_{j=1}^{K}\boldsymbol{y}_{i,j}\log\boldsymbol{p}_{i,j} \tag{4}$$

其中，M 为样本数，K 为类别数量，$\boldsymbol{y}$ 是观察的概率向量，$\boldsymbol{p}$ 是模型输出的概率向量。

3 实验部分

3.1 实验数据集

为实际验证模型的性能和表现，实验选取 TREC(The

Text REtrieval Conference Question Classification dataset)[24] 短问题层次分类通用数据集，该数据集由短问句组成，包括50个层次分类标签，如表2所列。但是由于其数据量有限、平均长度仍不够短，并且标签数据极不平衡，实验遴选了其中的极短文本，并进行数据扩充以提高数据质量。扩充方法以使用各类大语言模型进行语句生成为主，人工构造为辅，对模型生成内容加以严格的人工审核。实验剔除了有歧义、有语病、直接出现分类标志词、长度超过25个字词、不符合分类定义的大量生成数据，并及时删减了结构相似的语句，尽可能保留和生成各式各样的极短文本，以充分确保数据的有效性。扩充后的数据集数量达8425条，平均长度为12.14，数据也较为平衡，部分标签分布如表3所列。最终，数据集按5∶1∶1的比例随机形成训练集、验证集、测试集。

表2 数据样例

Table 2 Sample data

文本	标签
When did the neanderthal man live ?	39(NUM:date)
How can I enforce new rules to a group of youngsters who have been allowed to do as they please.	26(DESC:manner)

表3 部分标签分布

Table 3 Partial label distribution

标签	数量
2(ENTY:animal)	178
4(ENTY:color)	150
9(ENTY:food)	157
20(ENTY:techmeth)	149
25(DESC:desc)	407
31(HUM:desc)	159
33(LOC:country)	208
48(NUM:volumesize)	163

3.2 评价指标与实验环境

实验首先选择了文本分类评价中常用的Accuracy(准确率)、F1值作为评价指标。Accuracy是一个直观且易于理解的指标，它代表了被正确分类的样本占总样本数的比例；而考虑到扩充的数据集仍具有轻微的不平衡性。实验同时选取了更加关注模型整体性能的F1-micro，以及关注类别性能平衡性的F1-macro作为评价指标，从而能够更全面地评价模型的表现。F1-micro，F1-macro的计算式均如式(5)所示：

$$F1\text{-}mi/ma=\frac{2\times(p_{mi/ma}\times r_{mi/ma})}{(p_{mi/ma}+r_{mi/ma})} \tag{5}$$

其中，p指precision(精确率)，r指recall(召回率)，具体的计算式分别如式(6)所示。

$$\begin{aligned}
precision_{\text{macro}}&=\frac{\sum_{i=1}^{N}\frac{TP_i}{TP_i+FP_i}}{N}\\
recall_{macro}&=\frac{\sum_{i=1}^{N}\frac{TP_i}{TP_i+FN_i}}{N}\\
precision_{\text{micro}}&=\frac{\sum_{i=1}^{N}TP_i}{\sum_{i=1}^{N}TP_i+\sum_{i=1}^{N}FP_i}\\
recall_{\text{micro}}&=\frac{\sum_{i=1}^{N}TP_i}{\sum_{i=1}^{N}TP_i+\sum_{i=1}^{N}FN_i}
\end{aligned} \tag{6}$$

其中，TP表示被模型正确预测为正例的样本数量，FP表示被错误预测为正例的样本数量，FN则是被错误预测为反例的样本数量。而需要注意的是，在多分类任务中，测试数据只属于一个标签，一个标签的每个FN都是另一个类别的FP，因此在最后的计算数值上，F1-micro等于Accuracy。

然后，实验还选取了适合轻微不平衡数据的ROC_AUC_{micro}值作为评价指标。ROC(Receiver Operating Characteristic)曲线是反映模型分类性能的一种图像方式，AUC则是ROC曲线下与坐标轴围成的面积，AUC越接近于1，那么模型的性能就更优秀。在计算ROC_AUC_{micro}值时，ROC曲线的纵坐标TPR、横坐标FPR的计算式如式(7)所示：

$$\begin{aligned}
TPR&=\frac{\sum_{i=1}^{C}TP_i}{\sum_{i=1}^{C}(TP_i+FN_i)}\\
FPR&=\frac{\sum_{i=1}^{C}FP_i}{\sum_{i=1}^{C}(TN_i+FP_i)}
\end{aligned} \tag{7}$$

实验环境的参数设置如表4所列。

表4 环境参数

Table 4 Environmental parameters

详细参数
PyTorch 1.10.0
Python 3.8(ubuntu20.04)
Cuda 11.3
GPU:V100-32GB(32GB)
CPU:10 vCPU Intel Xeon Processor(Skylake,IBRS)

3.3 实验结果与分析

为了充分验证融合语义解释以及DeBERTa模型对增强极短文本层次分类表现的有效性，实验选取了一些热门文本分类模型，共构建了11组实验：1)仅Bi-LSTM[25]，一种双向循环神经网络模型；2)仅BERT，基于Transformer-encoder的神经网络模型；3)融合语义解释的BERT；4)仅TextCNN[26]，一种基于卷积神经网络的文本分类模型；5)仅DistilBERT[27]，对BERT进行知识蒸馏后的轻量级模型；6)融合语义解释的DistilBERT；7)仅RoBERTa，对BERT的优化预训练模型；8)融合语义解释的RoBERTa；9)仅DeBERTa；10)融合语义解释的DeBERTa；11)HiAGM模型，一种基于层次编码器的端到端层次感知全局模型。实验结果如表5所列。其中，融合语义解释的模块简称SI(Semantic Interpretation)；Bi-LSTM和TextCNN等模型由于主体网络结构不同于BERT系列模型，因此不引入语义解释。表5中呈现的数据都是基于数次调优后所选取的最佳结果。从表5中可以看出：

1)横向对比中，融合语义解释的DeBERTa模型表现最为优异，其测试集的Accuracy值、F1-micro值、F1-macro值、AUC值高达93.82，93.82，92.62，96.85，验证集的Accuracy值、F1-micro值、F1-macro值、AUC值高达93.14，93.14，91.78，96.98。这证明了当直接稳定地引入语义解释并结合DeBERTa模型时，模型能够充分利用上下文语义内容信息以及位置信息，从而很好地应对极短文本层次分类的挑战，实现

良好的性能；而传统的一些神经网络模型，如 Bi-LSTM、TextCNN 的表现不如预训练模型。表 6 列出了融合语义解释的 DeBERTa 模型具体标签的准确率情况，从中可以看出融合语义解释的 DeBERTa 模型还具有较好的类别平衡性，能够对每一个标签都作出有效的分类，不会出现个别标签准确率极低的情况。

2）从消融实验的角度，BERT，DistilBERT，RoBERTa，DeBERTa 这 4 个模型在使用融合语义模块后，性能都有一定的提升，BERT，DistilBERT 提升的幅度较大，验证集和测试集的各指标可提升 4～10 个百分点，RoBERTa 融合语义解释后测试集上的准确率、F1-macro 分别提升了 2.37，4.61 个百分点，DeBERTa 验证集和测试集上的指标提升了 1～4 个百分点。同时，还观察到不使用融合语义解释模块时，这 4 个模型的测试集表现均差于验证集表现，而使用之后，4 个模型的测试集表现都好于验证集表现，且提升得更多，这说明了融合语义解释能够很好地提升模型对于极短文本层次分类的泛化能力；具体到 F1-micro 和 F1-macro 时，使用融合语义解释模块后，F1-micro 和 F1-macro 之间的差距进一步缩小了，这侧面体现了模型类别平衡性的增强；对比仅使用这 4 个模型的情况也发现，能够更充分利用位置信息的 DeBERTa 模型的性能表现最好，仅 DeBERTa 比仅 BERT 的表现综合提升了约 5 个百分点。

3）从层次分类研究对比的角度，实验利用近期比较流行的开源 HiAGM 层次分类模型进行极短文本层次分类，取得了一定成效，其指标表现超过了 TextCNN，DistilBERT 等模型，但是仍然低于 BERT、融合语义解释的 DeBERTa 等模型。这侧面体现了一些层次分类研究主要聚焦于利用层次标签信息和常规文本分类，如 HiAGM 利用标签结构特征增强文本信息，但面对敏感的极短文本时，具有标签信息大于文本内容信息或放大标签噪声的风险，从而可能限制了其分类表现；在面对极短文本层次分类时，使用融合语义解释的 DeBERTa 模型是更优的选择。

表 5　实验结果

Table 5　Experimental results

Model	Validation				Test			
	Accuracy	*F1-micro*	*F1-macro*	AUC_{micro}	*Accuracy*	*F1-micro*	*F1-macro*	AUC_{micro}
Bi-LSTM	55.72	55.72	32.44	77.40	58.76	58.76	39.19	78.96
BERT	88.84	88.84	82.03	94.30	86.28	86.28	85.65	93.00
BERT＋SI	90.04	90.04	83.05	94.92	90.34	90.34	85.98	95.07
TextCNN	69.32	69.32	55.89	94.34	67.26	67.26	54.89	83.29
DistilBERT	70.11	70.11	59.79	84.75	67.56	67.56	57.18	83.45
DistilBERT＋SI	74.90	74.90	71.08	87.19	77.11	77.11	72.31	88.32
RoBERTa	91.63	91.63	90.31	95.73	90.85	90.85	87.91	95.33
RoBERTa＋SI	92.34	92.34	90.78	96.09	93.22	93.22	92.52	96.54
DeBERTa	91.33	91.33	89.43	95.58	90.95	90.95	88.25	95.38
DeBERTa＋SI	**93.14**	**93.14**	**91.78**	**96.98**	**93.82**	**93.82**	**92.62**	**96.85**
HiAGM	—	79.65	74.60	—	—	78.09	72.22	—

表 6　最优模型详细 Accuracy 分布

Table 6　Detailed Accuracy distribution of the optimal model

Label：Accuracy	Label：Accuracy	Label：Accuracy	Label：Accuracy	Label：Accuracy
‘ABBR-abb’：1.0	‘ENTY-instru’：1.0	‘ENTY-techno’：0.79	‘HUM-title’：1.0	‘NUM-dist’：1.0
‘ABBR-exp’：1.0	‘ENTY-lang’：1.0	‘ENTY-termeq’：0.75	‘HUM-desc’：1.0	‘NUM-money’：1.0
‘ENTY-animal’：0.94	‘ENTY-letter’：1.0	‘ENTY-veh’：1.0	‘LOC-city’：0.88	‘NUM-ord’：0.73
‘ENTY-body’：1.0	‘ENTY-other’：0.80	‘ENTY-word’：1.0	‘LOC-country’：1.0	‘NUM-other’：0.86
‘ENTY-color’：1.0	‘ENTY-plant’：1.0	DESC-def’：0.84	‘LOC-mount’：0.83	‘NUM-period’：1.0
‘ENTY-cremat’：0.89	‘ENTY-product’：1.0	‘DESC-desc’：0.89	‘LOC-other’：0.95	‘NUM-perc’：1.0
‘ENTY-currency’：1.0	‘ENTY-product’：0.83	‘DESC-manner’：1.0	‘LOC-state’：1.0	‘NUM-speed’：1.0
‘ENTY-dismed’：0.78	‘ENTY-sport’：0.94	‘DESC-reason’：1.0	‘NUM-code’：0.85	‘NUM-temp’：1.0
‘ENTY-event’：1.0	‘ENTY-substan’：0.93	‘HUM-gr’：0.94	‘NUM-count’：0.92	‘NUM-volsize’：0.94
‘ENTY-food’：0.94	‘ENTY-symbol’：1.0	‘HUM-ind’：0.99	‘NUM-date’：1.0	‘NUM-weight’：1.0

考虑到随机划分具有随机性，实验又进行了两次划分，因此实验一共进行了 3 次划分，以证明各模型以及本文方法的表现具有稳定性和鲁棒性。表 7 列出了融合语义解释的 DeBERTa 模型在另外两次划分中的详细结果。

表 7　融合语义解释的 DeBERTa 模型

Table 7　Performance of DeBERTa＋SI

另外两次划分	*Accuracy*	*F1-micro*	*F1-macro*	*AUC*
测试集 1	92.64	92.64	91.13	96.24
验证集 1	92.44	92.44	91.10	96.08
测试集 2	93.33	93.33	92.45	96.60
验证集 2	93.03	93.03	91.54	96.45

表 8 列出了部分模型在这 3 次划分中测试集上的 F1-micro，F1-macro 的均值与方差。从表 7 和表 8 中可以看出，在不同的数据集划分下，融合语义解释的 DeBERTa 模型和其他算法模型具有稳定的性能表现，之前的分析依旧成立。

需要指出的是，融合语义解释的 DeBERTa 模型没有明显的模型偏见性。一方面，从数据集组成来看，模型没有明显的位置偏见和分类标志词偏见，因为极短文本的特性，每个位置上的每个词都对语义理解有重大影响，模型必须关注和理解极短文本中的每个词，才能实现高性能的极短文本层次分类。同时，在数据集处理时，已经删除直接出现分类标志词的

语句，如 technology 和 product，避免了模型仅仅依靠分类标志词去判断的情况。另一方面，从实验结果来看，表 5 说明了模型的类别偏见并不明显，正确率分布十分均衡，表 8 也说明了模型在本数据集上的表现具有较好的鲁棒性。而为了进一步验证模型无明显偏见性和具有良好的泛化能力，又遴选 Twitter US Airline Sentiment 数据集[28]中长度小于 90 的数据进行了补充实验，数据集形式如表 9 所列，实验结果如表 10 所列。从表 10 中可以看出，融合语义解释和 DeBERTa 的模型依旧具有最优的性能，这从侧面说明了本文方法无明显的偏见性，并具有良好的泛化能力。

表 8 部分模型测试集表现的均值和方差

Table 8 Mean and variance of performance of partial models on test set

模型	均值		方差	
	F1-micro	*F1-macro*	*F1-micro*	*F1-macro*
DeBERTa+SI	93.26	92.07	0.23	0.44
DeBERTa	90.49	88.12	0.15	0.12
RoBERTa+SI	92.62	91.74	0.26	0.37
RoBERTa	90.11	87.74	0.46	0.81
BERT+SI	89.64	85.79	0.32	0.49
BERT	86.37	85.58	0.20	0.11
TextCNN	67.20	53.79	4.91	4.69

表 9 数据集形式

Table 9 Dataset formats

平均长度	训练集	验证集	测试集
64	2761	578	579

表 10 补充实验的结果

Table 10 Additional experimental results

模型	验证集		测试集	
	F1-micro	*F1-macro*	*F1-micro*	*F1-macro*
DeBERTa+SI	**94.65**	**94.22**	**94.30**	**94.02**
DeBERTa	93.09	92.48	92.90	92.51
RoBERTa+SI	94.47	93.95	93.42	93.06
RoBERTa	92.97	92.38	92.56	92.13
BERT+SI	93.43	92.81	93.09	92.52
BERT	91.32	90.61	89.08	88.32
TextCNN	82.74	—	78.36	—

结束语 针对极短文本层次分类的挑战，本文提出了一种融合语义解释和 DeBERTa 的模型。该方法先构建 Gloss-DeBERTa 模型获取高准确率的解释序列，并使用 SimCSE 框架获得表征能力更强的解释序列向量，从而直接稳定地将字词在语义中的含义信息引入，以补充模型获取的内容信息；然后，基于 DeBERTa 模型对极短文本进行特征提取，并且充分地利用相对位置与绝对位置信息，最终极短文本向量与解释序列向量相加后传入多分类器。实验结果表明，融合语义解释模块可以提升各模型的表现，并且融合语义解释的 DeBERTa 具有最优异的极短文本层次分类性能，优于其他各类模型。

下一步工作将聚焦于包括特定领域的网络热词的极短文本层次分类任务。不同于本文中使用的通用性极短文本数据集，当前网络上有很多的短文本，如网民评论等，会使用特定领域中的网络热词，这些网络热词的隐含意义超出了传统语义分析理解的范畴，并且往往与句式结构、句子语气等有较大关系。未来将探索包括网络热词的短文本层次分类方法。

参 考 文 献

[1] SIDDHARTHA B, CEM A, FRANCISCO P S, et al. Hierarchical Transfer Learning for Multi-label Text Classification [C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2019: 6295-6300.

[2] ZHOU J, MA C P, LONG D K, et al. Hierarchy-Aware Global Model for Hierarchical Text Classification [C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2020: 1106-1117.

[3] CHEN H B, MA Q L, LIN Z X, et al. Hierarchy-aware label semantics matching network for hierarchical text classification [C]// Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. Stroudsburg, PA: Association for Computational Linguistics, 2021: 4370-4379.

[4] HUANG C M, WANG S L. Research on Short Text Classification Based on Bag of Words and TF-IDF[J]. Software Engineering, 2020, 23(3): 1-3.

[5] WALLACH H M. Topic Modeling: Beyond Bag-of-Words[C]// Proceedings of the 23rd International Conference on Machine Learning. New York: ACM, 2006: 977-984.

[6] CHEN Q, YAO L, YANG J. Short text classification based on LDA topic model[C]// Proceedings of the 2016 International Conference on Audio, Language and Image Processing (ICALIP). Piscataway: IEEE, 2016: 749-753.

[7] DEVLIN J, CHANG M, LEE K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis: ACL, 2019: 4171-4186.

[8] LIU Y, OTT M, GOYAL N, et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach [EB/OL]. https://arxiv.org/abs/1907.11692.

[9] CHEN L C, QIN J, LU W D, et al. Short text classification method based on self-attention mechanism [J]. Computer Engineering and Design, 2022, 43(3): 728-734.

[10] HU Y, LI Y, YANG T, et al. Short Text Classification with A Convolutional Neural Networks Based Method [C]// Proceedings of the 2018 15th International Conference on Control, Automation, Robotics and Vision (ICARCV). Piscataway: IEEE 2016, 2018: 1432-1435.

[11] LY U S, LIU J. Combine Convolution with Recurrent Networks for Text Classification[EB/OL]. https://arxiv.org/abs/2006.15795.

[12] YANG F H, WANG X W, LI J. BERT-TextCNN-based classification of short texts from clinical trials [J]. Chinese Journal of Medical Library and Information Science, 2021, 30(1): 54-59.

[13] LIU Y, ZHANG K, HUANG Z, et al. Enhancing Hierarchical

Text Classification through Knowledge Graph Integration[C]// Findings of the Association for Computational Linguistics: ACL. Stroudsburg. PA: Association for Computational Linguistics, 2023: 5797-5810.

[14] LI B H, XIANG Y X, FENG D, et al. Short Text Classification Model Combining Knowledge Aware and Dual Attention[J]. Journal of Software, 2022, 33(10): 3565-3581.

[15] HOPPE F. Improving Zero-Shot Text Classification with Graph-based Knowledge Representations[C]// Proceedings of the Doctoral Consortium at ISWC 2022. FIZ Karlsruhe, 2022: 3165: 4.

[16] ZHENG K X, WANG Y Q, YAO Q M, et al. Simplified Graph Learning for Inductive Short Text Classification[C]// Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing Stroudsburg. PA: Association for Computational Linguistics, 2020: 10717-10724.

[17] HE P, LIU X, GAO J, et al. DeBERTa: Decoding-enhanced BERT with Disentangled Attention[EB/OL]. https://arxiv.org/abs/2006.03654.

[18] Lesk M. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone [C]// Proceedings of the 5th Annual International Conference on Systems Documentation. New York: ACM, 1986: 24-26.

[19] MONA D, PHILIP R. An unsupervised method for word sense tagging using parallel corpora [C]// Proceedings of 40th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2002: 255-262.

[20] BARBA E, PROCOPIO L, NAVIGLI R. ExtEnD: Extractive entity disambiguation [C]// Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2022: 2478-2488.

[21] HUANG L, SUN C, QIU X, et al. GlossBERT: BERT for Word Sense Disambiguation with Gloss Knowledge[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: Association for Computational Linguistics, 2019: 3509-3514.

[22] VASWANIA, SHAZEER N, PARMARN, et al. Attention is all you need[C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: ACM, 2017: 6000-6010.

[23] GAO T, YAO X, CHEN D. Simcse: Simple contrastive learning of sentence embeddings[EB/OL]. https://arxiv.org/abs/2104.08821.

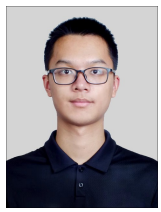
[24] HOVY E, GERBER L, HERMJAKOB U, et al. Toward semantics-based answer pinpointing[C]// Proceedings of the First International Conference on Human Language Technology Research. New York: ACM, 2021: 1-7.

[25] HUANG Z, XU W, YU K. Bidirectional LSTM-CRF models for sequence tagging[EB/OL]. https://arxiv.org/abs/1508.01991.

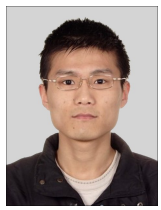
[26] KIM Y. Convolutional Neural Networks for Sentence Classification[C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg, PA: Association for Computational Linguistics, 2014: 1746-1751.

[27] SANH V, DEBUT L, CHAUMONDJ, et al. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. [EB/OL]. https://arxiv.org/abs/1910.01108.

[28] WAN Y, GAO Q. An ensemble sentiment classification system ofTwitter data for airline services analysis[C]// Proceedings of the 2015 IEEE International Conference on Data Mining Workshop (ICDMW). Piscataway: IEEE, 2015: 1318-1325.



CHEN Haoyang, born in 2003, undergraduate. His main research interests include NLP text classification and question answering.



ZHANG Lei, born in 1987, assistant researcher. His main research interests include artificial intelligence, intelligent agents, and multi-agent systems.

(责任编辑:喻藜)