CBLUE A Chinese Biomedical Language Understanding Evaluation-2106.08087v6

全文摘要

本文介绍了一个名为 CBUE（Chinese Biomedical Language Understanding Evaluation）的生物医学自然语言理解评估基准。该基准收集了真实世界中的生物医学数据，并包括命名实体识别、信息提取、临床诊断标准化等任务。为了在这些任务上进行评估，作者使用了当前可用的 11 个预训练中文模型，并报告了实验结果。实验结果显示，最先进的神经网络模型的表现远不如人类水平。该基准旨在促进跨语言自然语言理解的研究和应用。

论文速读

论文方法

方法描述

本文提出的 CBLUE 数据集包含 8 个不同的任务，包括命名实体识别、关系提取以及单句/句子对分类等。该数据集的特点是多样性，涵盖临床试验、电子病历、医学论坛、教科书和搜索引擎日志等多种来源，并且具有真实世界分布和长尾分布等特点。此外，该数据集还提供了用于特定转移学习场景的 CHIP-STS 数据集，其中测试集与训练集具有不同的分布。为了解决这些问题，作者采用了多种策略，如使用多个数据源、维护高质量的数据标注、使用匿名化技术保护隐私等。

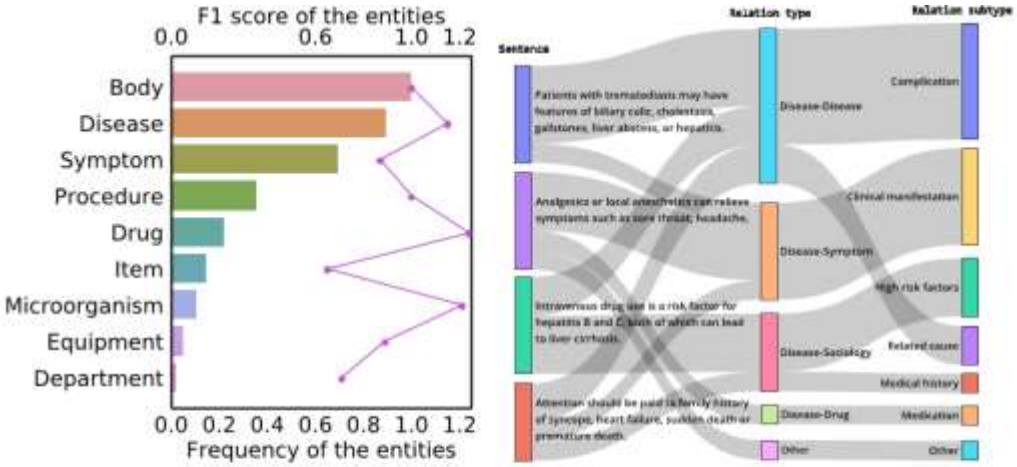
方法改进

在数据收集阶段，作者通过过滤无意义文本、随机抽样等方式保证了数据的质量和多样性。同时，在数据标注过程中，作者采用了严格的控制问题来防止不

诚实的行为，并通过多轮审核确保数据质量。此外，作者还提供了工具包以支持主流预训练模型和各种目标任务。

解决的问题

CBLUE 数据集的目的是提供一个多样化、真实世界分布的基准数据集，以便更好地评估自然语言处理模型的性能。该数据集的多样性和真实世界分布特点使得它能够更准确地反映实际应用场景中的挑战。此外，作者还考虑到了隐私保护等问题，从而进一步提高了数据集的质量和可靠性。



(a) CMeEE (b) CMeIE

Figure 1: Analysis of the named entity recognition and information extraction datasets. (a) illustrates the entity (coarse-grained) distribution in CMeEE and the impact of data distribution on the model's performance. We set entity type Body with the maximum number of entities to 1.0, and others to the ratio of number or F1 score to Body. (b) shows the relation hierarchy in CMeIE.

论文实验

本文介绍了对中文医学信息抽取（Medical Information Extraction）任务的基准测试——CBLUE，并进行了多种预训练模型的比较实验。实验结果表明，不同的预训练模型在不同任务上的表现存在差异，且人类的表现要优于机器。此外，文中还分析了错误案例和数据特点，指出了中文语言的独特之处，需要更强大的模型来解决这些挑战。最后，文章列出了所有实验的超参数和评估指标。

		CMeEE	CMeIE	CDN	CTC	STS	QIC	QTR	QQR
Trained annotation	annotator 1	69.0	62.0	60.0	73.0	94.0	87.0	75.0	80.0
	annotator 2	62.0	65.0	69.0	75.0	93.0	91.0	62.0	88.0
	annotator 3	69.0	67.0	62.0	80.0	88.0	83.0	71.0	90.0
	avg	66.7	64.7	63.7	76.0	91.7	87.0	69.3	86.0
	majority	**67.0**	**66.0**	**65.0**	**78.0**	**93.0**	**88.0**	**71.0**	**89.0**
	best model	62.4	55.9	59.3	70.9	85.6	85.5	62.9	84.7

Table 4: Human performance of two-stage evaluation scores with the best-performed model. "avg" refers to the mean score from the three annotators. "majority" indicates the performance taken from the majority vote of amateur humans. Bold text denotes the best result among human and model prediction.

Query	Model			Gold
	BERT	BERT-ext	MedBERT	
请问淋巴细胞比率偏高、中性细胞比率偏低有事吗? Does it matter if the ratio of lymphocytes is high and the ratio of neutrophils is low?	病情诊断 Diagnosis	病情诊断 Diagnosis	指标解读 Test results analysis	指标解读 Test results analysis
咨询：请问小孩一般什么时候出水痘? Consultation: When do children usually get chickenpox?	其他 Other	其他 Other	其他 Other	疾病表述 Disease description
老人收缩压160，舒张压只有40多，是什么原因？怎么治疗? The systolic blood pressure of the elderly is 160, and the diastolic blood pressure is only more than 40. What is the reason? How to treat?	病情诊断 Diagnosis	病情诊断 Diagnosis	病情诊断 Diagnosis	治疗方案 Treatment

Table 6: Case studies in KUAKE-QIC. We evaluate the performance of baselines with 3 sampled instances. The correlation between Query and Title is divided into 3 levels (0-2), which means '*poorly related or unrelated*', '*related*' and '*strongly related*'. BERT = BERT-base, BERT-ext = BERT-wwm-ext-base, MedBERT = PCL-MedBERT.

Method	Value
warmup_proportion	0.1
weight_decay	0.01
adam_epsilon	1e-8
max_grad_norm	1.0

Table 15: Common hyper-parameters for all CBLUE tasks

Model	epoch	batch_size	max_length	learning_rate
bert-base	5	32	128	4e-5
bert-wwm-ext	5	32	128	4e-5
roberta-wwm-ext	5	32	128	4e-5
roberta-wwm-ext-large	5	12	65	2e-5
roberta-large	5	12	65	2e-5
albert-tiny	10	32	128	5e-5
albert-xxlarge	5	12	65	1e-5
zen	5	20	128	4e-5
macbert-base	5	32	128	4e-5
macbert-large	5	12	80	2e-5
PCL-MedBERT	5	32	128	4e-5

Table 16: Hyper-parameters for the training of pre-trained models with a token classification head on top for named entity recognition of the CMeEE task.

Model	epoch	batch_size	max_length	learning_rate
bert-base	8	32	128	5e-5
bert-wwm-ext	8	32	128	5e-5
roberta-wwm-ext	8	32	128	4e-5
roberta-wwm-ext-large	8	16	80	4e-5
roberta-large	8	16	80	2e-5
albert-tiny	10	32	128	4e-5
albert-xxlarge	8	16	80	1e-5
zen	8	20	128	4e-5
macbert-base	8	32	128	4e-5
macbert-large	8	20	80	2e-5
PCL-MedBERT	8	32	128	4e-5

Table 18: Hyper-parameters for the training of pre-trained models with a classifier for the entity pairs relation prediction of the CMeIE task.

论文总结

文章优点

该论文提出了一种新的中文生物医学自然语言处理基准测试集——CBLE，该测试集包括了八个任务，涵盖了临床诊断标准化、信息提取、问答系统等多个领域。此外，该论文还提供了详细的实验结果和分析，对于研究者来说具有很高的参考价值。

方法创新点

该论文的方法创新点在于提出了一个全面的中文生物医学自然语言处理基准测试集，并且在该测试集中包含了多个不同的任务类型。此外，该论文还使用了多个人工智能模型来评估这些任务的表现，从而更加客观地评价了这些模型的效果。

未来展望

该论文为中文生物医学自然语言处理的研究提供了一个重要的基础，未来可以进一步扩展该测试集，增加更多的任务类型，以更好地反映中文生物医学自然语言处理的实际需求。同时，也可以尝试使用更先进的深度学习技术来改进这些任务的表现，提高其应用价值。