CCL24-Eval任务6系统报告：基于深度学习模型的中小学作文修辞识别与理解评测

李晨阳[1,2]，张龙[1,2]，郑秋生[1,2]
[1]中原工学院 前沿信息技术研究院，河南 郑州 450007
[2]河南省网络舆情监测与智能分析重点实验室，河南 郑州 450007
2312826399@qq.com

摘要

在中小学生的学习进程中，修辞手法是阅读和写作技巧的核心，也是优秀文学作品的关键元素。然而，识别与理解学生文章中的修辞使用需要大量的人工，为教师的作文评估和教学提出了挑战。最近的研究开始使用计算机技术来自动评审作文，其中修辞的使用是评估的重要部分。本文介绍了我们在第二十三届中文计算语言大会中中小学作文修辞识别与理解评测中的所用的参赛方法。在本次评测中，我们针对不同任务，分别使用了传统模型分类模型和大模型，再利用伪标签、数据增强等方法提升模型性能。实验结果表明，我们的方法取得了较为先进的效果。

关键词： 预训练模型 ； 数据增强 ； 半监督学习

System Report for CCL24-Eval Task 6: Assessment of Rhetoric Recognition and Understanding in Primary and Secondary School Essays Based on Deep Learning Models

Chenyang Li[1,2], Long Zhang[1,2],Qiusheng Zheng[1,2]
[1]Frontier Information Technology Research Institute,
Zhongyuan University of Technology, Zhengzhou 450007 China
[2]Henan Key Laboratory on Public Opinion Intelligent Analysis, Zhengzhou China
2312826399@qq.com

Abstract

In the learning process of primary and secondary school students, rhetorical devices are at the core of reading and writing skills, and are key elements of excellent literary works. However, identifying and understanding the use of rhetoric in students' essays requires a lot of manual effort, posing a challenge for teachers in essay evaluation and instruction. Recent research has begun to use computer technology to automatically evaluate essays, where the use of rhetoric is an important part of the assessment. This paper introduces our methods used in the evaluation of rhetorical recognition and understanding in primary and secondary school essays at the 23rd Chinese Computational Linguistics Conference. In this evaluation, we employed both traditional classification models and large models for different tasks, and further enhanced model performance through pseudo-labeling and data augmentation. Experimental results indicate that our methods achieved relatively advanced performance.

Keywords: pretrained models , data augmentation , semi-supervised learning

1 引言

修辞手法在中小学生的学习过程中占据了重要的地位，它是阅读理解和写作技巧的核心，也是塑造优秀文学作品的重要元素。然而，识别和理解修辞的使用需要大量人工成本，对教师的作文评估和教学提出了挑战。目前，随着教育的发展和网络的普及，许多研究者和机构开始探索利用计算机技术来实现作文的自动评改，其中修辞手法的应用是重要考量因素。作文的修辞使用被视为反映学生文采和语言表达能力的重要指标，对于评估作文质量，指导学生提升表达能力有重要意义。一些现有的研究采用对齐策略等规则，从句子结构，语义信息等语言学特征角度进行排比和比喻等修辞手法的识别；而另一些工作则将修辞理解视为构成抽取任务。然而，这些研究通常存在以下问题：①针对不同的修辞类别进行独立识别，缺乏泛用性；②识别粒度粗，缺乏多层次，细粒度的修辞类型定义；③缺乏对不同修辞类型的修辞对象和内容的定义，无法为学生作文提供全面的指导意见。在这样的背景下，第二十三届中国计算语言学大会发布了中小学作文修辞识别与理解评测。该评测的数据集源自以汉语为母语的中小学生考试作文，包括记叙文和议论文等多种文体。任务目标是系统地定义中小学作文中出现的细粒度修辞类型，从修辞形式和内容两个方面进行识别，并对修辞的使用评分。对于本评测的任务1和任务2，我们采用传统模型进行分类，并使用伪标签、数据增强等方法来提升性能，对于任务3，我们把该信息抽取任务视为生成任务，并引入Qwen大模型(Bai et al., 2023)进行生成。

2 相关工作

文本分类在自然语言处理和文本挖掘中具有重要的作用，通过不断学习文本特征进行预测分类，在各个方面的研究中都具有十分重要的意义和研究价值 (Minaee et al., 2021)。传统的文本分类是基于机器学习方法 (Cheng, 2020)，包括支持向量机、决策树、朴素贝叶斯等，但这些方法都只解决了词汇层面的问题，无法有效学习和反映语句之间的语义相关性和深层语义特征。近年来，深度学习技术在计算机视觉和自然语言处理领域都取得了显著的进展。在自然语言处理任务中，基于深度学习的文本分类模型备受关注和研究，如CNN (Wan et al., 15) (Wang et al., 2017)、RNN (Le et al., 2017)、GNN (Yao et al., 2018)、Attention (Kim et al., 2018)和预训练模型。它们在文本分类等自然语言处理任务中都表现出了优秀的效果。特别是预训练模型，在预训练时就已经接触了大量的文本数据，因此能学习到更加丰富的语义信息，其在文本分类等任务中也具有更高的准确性和泛化能力。

半监督学习是近年来新兴的一种智能学习范式，其利用未标记的数据提高模型性能 (Rizve et al., 2021)，传统的监督学习方法需要使用有标签的数据来建模。然而，在现实世界中给训练数据打标签可能需要昂贵的代价，或者耗费大量的时间。从领域专家那里获得这些标签数据有隐含的成本，例如有限的时间和财务资源。对于涉及使用大量类标签进行学习且有时具有相似性的应用程序尤其如此。半监督学习（SSL）模型能够允许模型在其监督学习中集成部分或者全部的未标签数据来解决这一固有的瓶颈。其目标是通过这些新标记的数据最大化模型的学习性能，同时最小化标注数据的成本。

模型融合是一种训练多个模型并进行融合的方法，旨在通过融合模型结果来超越单个模型的表现。常用的模型的融合方法共有三种，第一种是投票法，适用于分类任务，即对多个学习模型的预测结果过进行投票，以少数服从多数的方式来确定最后的结果，还可以根据人工设置或者根据模型评估分数来设置权重。第二种是平均法，适用于回归和分类任务，即对于学习模型的预测概率进行平均。第三种是交叉融合法，主要思路就是把原始的训练集先分成两部分，例如按9：1划分训练集和测试集，在第一轮训练时，使用训练集训练多个模型，然后对测试集进行预测，在第二轮训练时，直接用第一轮训练的模型在测试集上的预测结果作为新特征继续训练。

3 实现方法

如图1所示，我们先对数据进行预处理，然后对当前主流深度学习模型进行评估，选出表现较好的模型作为我们的基线模型，最后采用伪标签、数据增强等方法来提升分数。

3.1 任务1、2实现方法

任务一与任务二皆为嵌套多标签分类任务，由于这两个任务使用方法一样，我们以任务一为例进行介绍。我们分别对其训练粗粒度和细粒度两个模型，最后在进行粗细结果的融合。

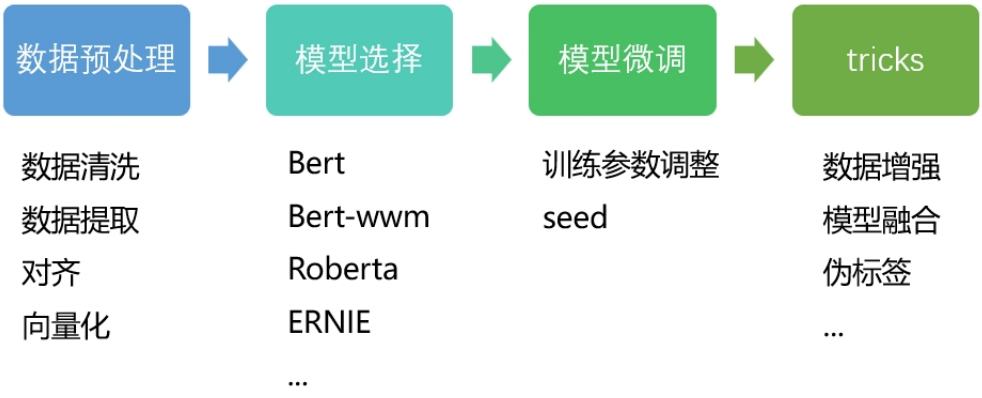

图 1: 模型的数据预处理、预测以及后处理过程

3.1.1 模型选择

如表3.1.1所示，我们选取了当前主流的预训练模型，包括Bert，Bert的变体模型和Ernie，再次基础上对每个模型进行微调。从直观来讲，网络模型越大，层数越深，学习能力越强大，因此我们的Ernie模型 (Sun et al., 2019)选择了20层网络结果的Ernie3.0-xbase进行测试，同时我们后面提出的Ernie均指Ernie3.0-xbase。经过验证，我们发现ernie的效果最高，于是我们把ernie作为我们的基线模型。

模型	Bert_base[0]	Bert_wwm[1]	Bert_wwm_ext[2]	Ernie3.0_xbase[3]	Rebort_wwm_ext[4]
验证集F1	31.1	32.8	31.8	33.4	32.1

表 1: 各模型验证集F1值

3.1.2 数据增强

为了提升F1分数，我们对每个类别的总数和每个类别对应的F1值进行了比较，如图2所示，我们发现了数据分布不平均，部分标签对应的数据量过少的问题，例如属于明喻这一类别的数量有147条，而属于副词这一类别的数量只有4条。这也导致副词这一类别的f1值为0。为了提升该类别的f1分数，我们采用数据增强的方式来使得模型学习到更多的相关特征。采用同义词替换、随机词插入和相似句生成等方法对这些标签的数据进行了数据增强，增强例子如图3所示。在对数据总量增加了三倍之后，我们使用增强后的数据对基线模型进行了重新训练。结果表明，无论是在验证集还是测试集上，模型的性能都有了显著的提升。

3.1.3 基于半监督学习的伪标签生成

伪标签方法来自于半监督学习，其核心思想是借助无标签的数据来提升有监督过程中的模型性能。由半监督学习生成伪标签的过程如图4所示，伪标签方法主要是将模型对无标签的测试数据的预测结果重新加入到训练集中，从而增大数据量以提升模型效果(Li et al., 2023)。这种方法适用于模型精度较高的情况。我们在预测过程中设置了一个阈值0.98。在预测测试集过程中，模型输出概率值大于该阈值的数据都被保存作为伪标签。这些伪标签在我们看来都是接近完全正确的数据。每次的伪标签都接近1000条。这表明我们的数据集在原本数量的基础上又增加了1000条数据。经过多轮的伪标签训练后，筛选出的伪标签内容会越来越接近，最终模型达到了拟合的状态，此时在进行后续的伪标签方法已经不能够再提升测试集的f1值。于是我们采用模型融合的方法来进一步提升f1分数。

[0]https://huggingface.co/google-bert/bert-base-chinese
[1]https://huggingface.co/hfl/chinese-bert-wwm
[2]https://huggingface.co/hfl/chinese-bert-wwm-ext
[3]https://huggingface.co/nghuyong/ernie-3.0-xbase-zh
[4]https://huggingface.co/hfl/chinese-roberta-wwm-ext-large

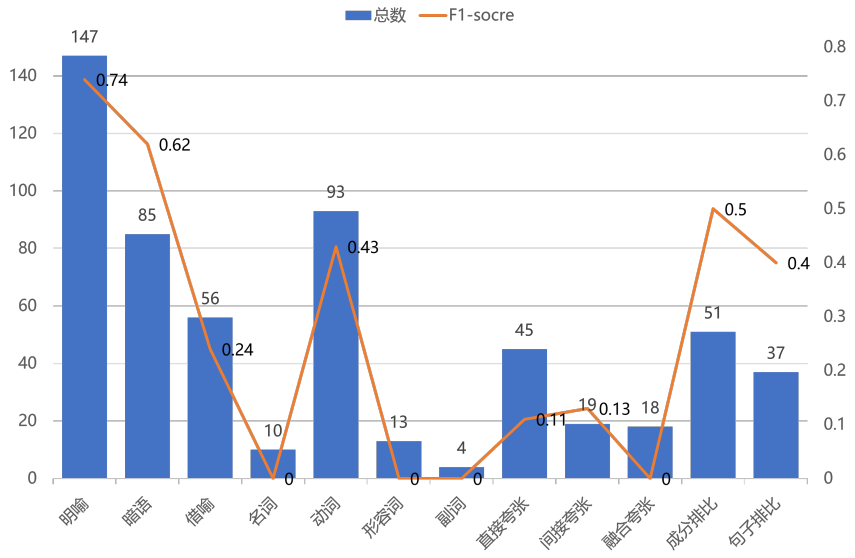

图 2: 各类别对应的训练集数量

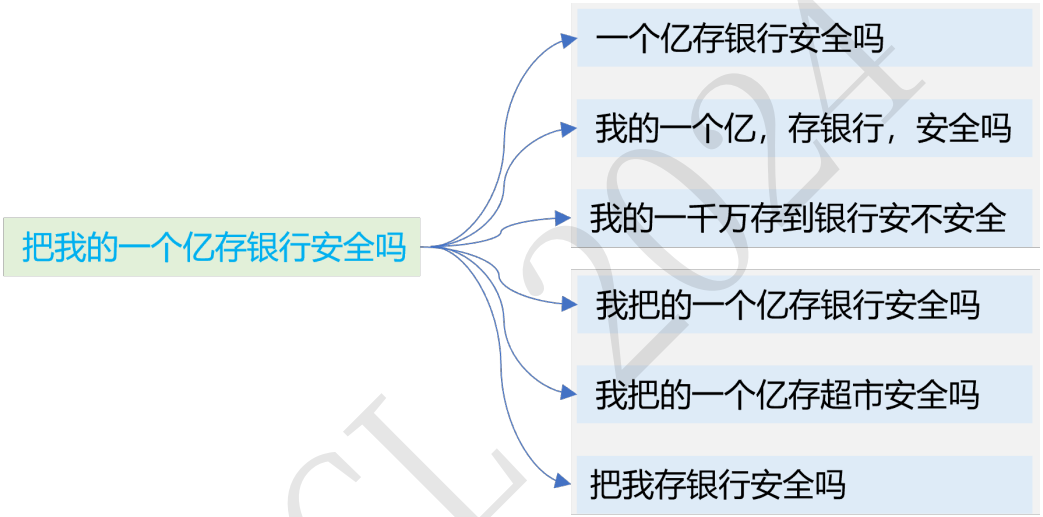

图 3: 数据增强样例

3.1.4 模型融合

关于模型融合，周志华教授(2021)的《机器学习》一书中提到：模型融合要好而不同，即模型差异性越大，融合效果越好。我们从两方面来增加差异化，一是使用不同的两个模型，bert和ernie。二是重新划分训练集和验证集来改变模型输入。再使用这两个模型对测试集进行预测时，我们没有直接输出预测的类别，而是输出了每个类比的概率，然后使两个模型预测的类别概率进行等权相加，得出模型融合后预测的新概率，最后选取具有最大概率的那一个类别作为预测结果。

3.2 任务3实现方法

任务3为信息抽取任务，我们最初尝试使用传统在的信息抽取类模型来完成，但效果并不理想，继而转向了语言大模型，把其作为文本生成任务来实现。

3.2.1 数据格式化及模型选择

我们首先尝试了使用prompt模板的ChatGLM4，发现生成的答案除了格式不规范外还会出现很多原句中没有出现过的词，这显然也是不符合结果要求的。于是我们采用微调大模型的方式来实现。首先是对数据进行格式化处理。如图所示，我们把该任务组为序列生成任务，

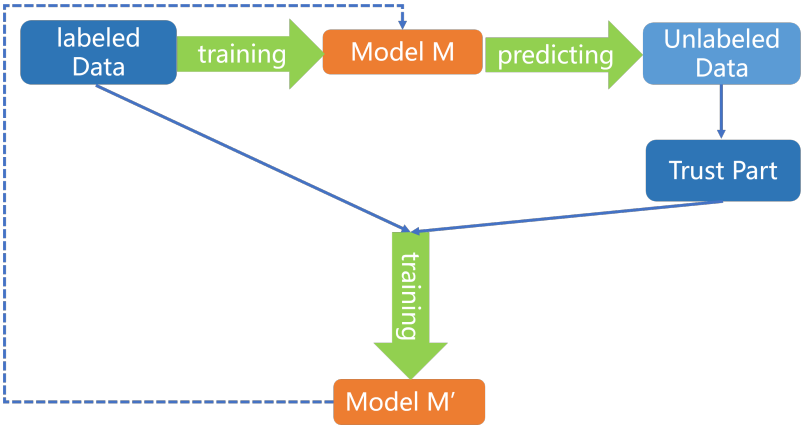

图 4: 基于半监督学习的伪标签生成

而不是直接生成符合本任务的数字答案。如图5所示，我们把原句作为输入，把连接词、描写对象、描写内容3个属性作为输出，其间用“#”进行分隔,如果某个属性为空，则用”null”表示，如果3个属性都为空，则用”null#null#nul”表示。最后针对输出的结果在输入句子的基础上进行位置对比而输出符合任务格式的数据。我们尝试了ChatGLM6b和Qwen7b等开源大模型，采用qlora(Dettmers et al., 2024)和p-tuning(Liu et al., 2022)等方式针对本任务数据进行微调。我们发现Qwen7b效果较好，且当使用qlora进行微调时，把秩的大小设置为64要明显好于其他参数。于是把其当作基线模型进行后续的测试。


```
{
  "paragraphId": 2,
  "sentenceId": 13,
  "sentence": "他变的行尸走肉,成天郁郁寡欢。",
  "componentList": [
    {
      "conjunction": null,
      "conjunctionBeginIdx": null,
      "conjunctionEndIdx": null,
      "tenor": "他",
      "tenorBeginIdx": 0.0,
      "tenorEndIdx": 0.0,
      "vehicle": "变的行尸走肉",
      "vehicleBeginIdx": 1.0,
      "vehicleEndIdx": 6.0
    }
  ]
```

```
{
  "conversations": [
    {
      "from": "user",
      "value": "他变的行尸走肉,成天郁郁寡欢。"
    },
    {
      "from": "assistant",
      "value": "null#他#变的行尸走肉"
    }
  ]
},
```


图 5: 左为原始数据格式，右为格式化后格式

3.2.2 伪标签

在本任务中，我们同样对数据进行分析，并发现数据总体数量是比较少的，属于小样本任务。为了提升后续的分数，我们同样采用任务一中的数据增强方法，以此增加数据集的数量。在增加了1倍数据后，经过训练后发现在其结果在验证集上有了很大的提升，但在测试集上却有所下降，出现了过拟合的状态，可能是由于数据增强的方式破坏了原本的数据分布导致，

于是我们采取了同样能增加训练集数量的伪标签方法。伪标签方法适用于模型精度比较高的情况，此时我们的模型准确率已经达到69的分数，故可以采用该技巧。由于大模型无法想传统分类模型那样输出预测的概率，故不能想任务一那样设置阈值，于是我们结合模型融合的经验，我们利用微调后的ChatGLM和Qwen进行预测，并去预测结果相同的部分加入到训练集中重新训练。每次的伪标签数量大概在1500条，这表明我们的数据集在原本数量的基础上又增加了1500条数据。结果表明，在使用伪标签后，评价指标有所提升。

4 实验结果

如下表所示，分别列出了3个任务的实验结果。任务1和任务3展示了我们在基线模型的基础上，使用了数据增强和伪标签等方法后的线上F1值，相对任务的基线模型分数45.66，我们有了很大的提升。针对于任务3，在使用大模型之前，我们也尝试使用了类似Bert的传统分类模型，然而，传统模型在测试集上的表现要远低于基线模型的结果。这一结果表明基座模型对于推理效果有着显著的影响，而大模型在预训练获得的世界知识和涌现能力对空间语义理解能力任务有着重要帮助。我们在对大模型微调时也面临着一个普遍问题，即幻觉现象(Huang et al., 2023)。当模型生成的文本不遵循原文或者不符合事实时，我们就认为模型出现了幻觉，尽管我们在训练集中标注的答案都是在输入文本中出现过的，但模型在结果预测时仍会出现除输入文本以外的词，为此，我们暂时只采用正则表达式来过滤掉这些无效答案。

模型	线上F1值
Baseline	45.66
Ernie3	44.72
Ernie3+数据增强	46.16
Ernie3+数据增强+伪标签	49.83
Ernie3+数据增强+伪标签+bert_wwm(模型融合)	51.48

表 2: 任务1结果

模型	线上F1值
Baseline	56.89
Ernie3	49.78
Ernie3+数据增强	53.56
Ernie3+数据增强+伪标签	54.25
Ernie3+数据增强+伪标签+bert_wwm(模型融合)	55.11

表 3: 任务2结果

表 4: 任务3结果

模型	线上F1值
Baseline	20.85
ChatGLM4+promptChatGLM-4[5]	24.68
ChatGLM6b[6]	57.00
Qwen1.5-7bQwen1.5-7b[7]	63.30
Qwen7b-1.5+伪标签	69.51

表 5: 任务2结果

[5]https://chatglm.cn
[6]https://github.com/THUDM/ChatGLM3
[7]https://github.com/QwenLM/Qwen1.5

5 总结

通过大量实验发现，对于本任务数据集，大多数模型预测的结果分数相近，并且由于本任务数据集数规模比较小，因此数据增强方法伪标签方法都可以增加数据集的规模，提升测试集的F1分数，并增加模型的泛化性。使用多轮伪标签方法后，后续筛选得出的伪标签几乎不会有变化，导致模型的性能不再有提升。这时可以采用模型融合技术，取差异较大的多个模型，分别学习不同的输入，使得多个模型之间学到的知识尽量不同，这样使得多个模型可以更好的融合，提高性能。进一步优化方面，针对训练集存在的过拟合问题，可以考虑在划分训练集和验证集时进行数据均衡。使用模型融合时，可以先采用五折交叉验证法来训练多个模型，然后再对多个预测结果取平均。

参考文献

Bao Guo, Chunxia Zhang, Junmin Liu, and Xiaoyi ma. 2019. Improving text classification with weighted word embeddings via a multi-channel textcnn model. *Neurocomputing*,363:366 – 374.

Chenyang Li, Long Zhang, Qiusheng Zheng, Zhongjie Zhao, and Ziwei Chen. User preference prediction for online dialogue systems based on pre-trained large model. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 349–357. Springer, 2023.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Jiang Cheng. 2020. Research and implementation of Chinese long text classification algorithm based on deep learning. *University of the Chinese Academy of Sciences (Institute of artificial intelligence, Chinese Academy of Sciences)*.

Le, H. T., Cerisara, C., and Denis, A. 2017. Do convolutional networks need to be deep for text classification? *arXiv preprint arXiv:1707.04108*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*, 2023

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 61–68, 2022.

Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., and Gao, J. 2021. Deep learning based text classification: a comprehensive review. *ACM computing surveys (CSUR)*,54(3), 1-40.

Rizve, M. N., Duarte, K., Rawat, Y. S., and Shah, M. 2021. In defense of pseudo-labeling:an uncertainty-aware pseudo-label selection framework for semi-supervised learning. *arXiv preprint arXiv:2101.06329*.

Seonhoon Kim, Jin Hyun Hong, inho Kang, and nojun kwak. 2019. Semantic sense matching with densely connected recurrent and co-attentive information. *Proceedings of the AAAI Conference on Artificial Intelligence*,33(01), 6586-6593.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36, 2024.

Wan, S., Lan, Y., Guo, J., Xu, J., Pang, L., and Cheng, X. 2016. A deep architecture for semantic matching with multiple positive sense representations. *Proceedings of the AAAI Conference on Artificial Intelligence*,30(1). https://doi.org/10.1609/aaai.v30i1.10342.

Wang, Z., Hamza, W., and Florian, R. 2017. Bilateral multi-perspective matching for natural language sentences. *In procedures of the twenty Sixth International Joint Conference on artistic intelligence*, ijcai-17, pages 4144 – 4150.

Yao, L., Mao, C., and Luo, Y. 2019. Graph revolutionary networks for text classification. *In Proceedings of the AAAI conference on artificial intelligence*,Vol. 33, No. 01, pp. 7370-7377

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, danxiang Zhu, Hao Tian, and Hua wuErnie. 2019. Enhanced representation through knowledge. *arXiv preprint arXiv:1904.09223*.

Zhi-Hua Zhou. *Machine learning*. Springer nature, 2021.