使用多种语言模型分类长篇语料句子

李　新[1]，孙　润[2]

（1. 广东海洋大学外国语学院语言实训中心，广东湛江　524088；

2. 广东海洋大学外国语学院，广东湛江　524088）

[摘　要]为了进行生态翻译学研究，我们实施了一系列数据提取实验，利用多种语言模型对长篇语料中的句子进行分类研究。该研究旨在探索语言模型在文学作品研究环境下的应用，以进一步提高文学作品研究的质量和效率。通过收集研究语料和训练语料，并运用包括 Bert、ChatGPT、Claude、ChatGLM 等多种语言模型，我们对大规模文本进行了分类和筛选。通过这项实验，我们希望能够识别粗语以及谚语俗语等语义特征，为跨语言翻译提供更深入的理解和支持。在实验结果中，我们观察到利用多种语言模型能够对长篇语料中的句子进行有限的分类，在一定程度上减轻了人工分类的工作量，并且本实验数据能为相关翻译研究提供支持。

[关键词]《红高粱家族》；平行语料；Bert；ChatGPT；ChatGLM；Ernie

[作者简介]李新（1987—），男，广东湛江人，工程师，硕士，研究方向：计算机技术与语言处理。孙润（1989—），男，湖北黄冈人，广东海洋大学外国语学院讲师，博士在读，研究方向：普通语言学及文学翻译。

[基金项目]本文系广东海洋大学人文社科项目“魔幻现实主义风格下的莫言作品英译研究——虚幻与现实融合下的中国底层人物形象”（项目编号：C22862）。

[DOI]https://doi.org/10.62662/kxwxz0109007

[本刊网址]www.oacj.net

一、前期工作回顾

本项目的研究对象为《红高粱家族》的中英文语料，中文为莫言的《红高粱家族》，英文为葛浩文翻译的“Red Sorghum”，相关语料从网络获取。得到语料后，对中英文语料进行清洗、分句和词性标注，使用到的工具包括 nltk、ltp、jieba 等。其中《红高粱家族》句子总数为 7379，“Red Sorghum”句子总数为 8619。

此外，为了研究中英文对应句子的翻译特点，我们还需要对中英文语料进行锚点对齐，将中英文双边连续 3 句高相似度句子作为锚点，经过检查这种策略得到的锚点的准确度较高，并且锚点较多，总共锚点簇数 420、锚点数 768。这意味着可以根据锚点按块对语料进行分块细化对齐。

二、利用 ChatGPT 进行长篇双语语料对齐

鉴于 ChatGPT 3.5 近年来火爆全球，且其上下文理解、跨语言分析、自动化处理等通用智能能力让人感觉到其“无所不能”的潜力，且尝试过的一些所谓智能对齐软件，例如 LanguageX，对齐效果也一般，故有必要尝试一下新的手段。当然使用 ChatGPT 进行语料对齐也有“杀鸡用牛刀”的意味。

我们选择一些使用成本较低的、国内能访问的一些 ChatGPT 接入站或者套壳网站来使用，且网站支持 ChatGPT 3.5-16K，它可以支持长文本分析，非常适合用来进行长文本对齐。

每次我们选取 20 句中英文句对进行对齐，并要求 ChatGPT 以句子为单位用 Turple 元组形式输出对齐结果，然后人工大致分析对齐结果。如果输出不理想，例如出现删除句子、改变句子或者大量调换位置、大量空对齐的现象时，需要重新要求 ChatGPT 输出，或修改提示语或者减少句子数量。我们还发现当中英文结尾句子恰好对应时，ChatGPT 的对齐效果可以很好。另外我们还发现对齐效果可能受时间段影响，某段时间，特别是晚上 9 点，使用人数过多可能造成服务器负担，影响算力的分配。这时我们选择 Claude 模型进行输出也能得到不错的效果。

由于 ChatGPT 存在输出格式错误的情况，例如没有以 Turple 元组格式输出、转译字符错误，我们还进行了编程，对收集输出后的结果重新划分中英

文句子。尽管重新划分后的句子存在某些多余的字符,但是对后面模型推理的结果估计影响较小。

通过使用 ChatGPT 结合人工进行对齐,最后得到对齐句对为 7077 对,其中 39 句没有对应的中文、231 句没有对应的英文。39 句没有对应的中文,是因为内容色情敏感被删除了,而 231 句没有对应的英文,则是因为译者没有翻译,而其他译者没有翻译的中文内容已经和其他句子合并到一起。

三、使用 Bert 对粗语进行分类

Bert 是一种基于 Transformer 架构的预训练语言模型,它在大量的语言分类任务里表现出优异的性能。由于粗语与网络冒犯性用语类似,可以借鉴其方法进行分类。

(一)获取数据集

训练模型使用的正面数据集为 THUCNews 的中文新闻数据集,由于是新闻网站,可以认为是不包含粗语的,而负面数据集则从百度搜索得到的语录网站中收集。接着处理收集到的数据,例如补全粗语屏蔽掉的部分、删除重复句子、切分为短句子、去除含英文的句子等等。

最后随机选取 4 万句正面句子和 619 句负面句子混合后分为 8:1:1 作为训练集、验证集和测试集。

(二)模型结构

使用的模型结构较为通用,如图 1 所示,首先是 Bert 模型,接着定义了一个 Dropout 层随机丢弃神经元,以减少过拟合,最后是 Linear 线性层和 ReLU 激活函数。使用过的预训练模型包括 Chinese-BERT-wwm 和 WoBERT,这两种模型都是按词进行掩码训练得到的,而使用 WoBERT 推断得到的负面句子较多。

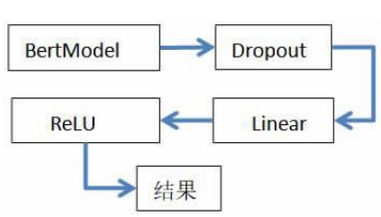


图 1　模型结构

(三)训练测试结果

尽管经过很少的训练轮数后,测试集中的准确率很高,能达到 98% 以上,但是模型应用到实际数据推断时,精确率却很低。其中使用 WoBERT 推理出 486 句,而人工筛选只有 62 句符合要求,而使用 Chinese-BERT-wwm 推理出 128 句,而人工筛选只有 30 句符合要求。虽然精确率较低,但这可能与筛选人员对粗语的理解有关,并且其他“误判的”句子确实也存在语言上负面情绪的特点,例如图 2 的 4810 行的“你别张狂”、5027 行的“小瞧我们”、6456 行的“秃头鬼子”等等。

图 2　WoBERT(上)和 Chinese-BERT-wwm(下)推理结果

四、使用 ChatGLM、Ernie 对谚语俗语进行分类

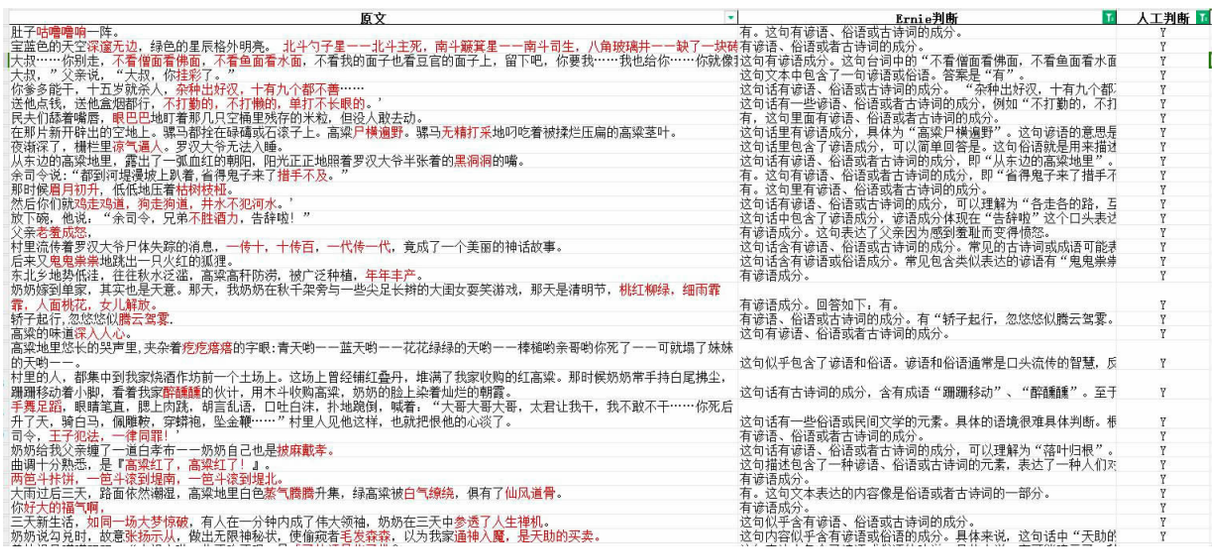
当我们尝试利用 Bert 模型用相同的方法进行谚语俗语分类时,发现得到的结果却极差,数据少且错误率极高。后面考虑到谚语俗语会不会存在结构化的句式特点,试图使用词性进行训练和推理时,发现结果依然不理想。猜测可能是谚语俗语的词汇和句式多样化,从训练数据中无法找到公共的特征。

接着只能求助于大模型,尝试使用 API 的方式调用大模型进行分类。由于 ChatGLM3-6B 是免费的,而其在线对话网站尝试用提示语分类谚语俗语句子的效果也较好,于是决定使用 ChatGLM3-6B 进行谚语俗语分类。

我们下载了 20G 的模型,以本地方式进行实

验。由于测试时有些句子会被识别为民谣和流行语，所以输入提示语时包含了民谣和流行语作为关键字。程序经过 2 天的运行得到了 2300 句谚语俗语。由于分类结果较多，且错误率也较高，于是需要进一步进行筛选。我们选择 Ernie 对分类结果进行二次分类。

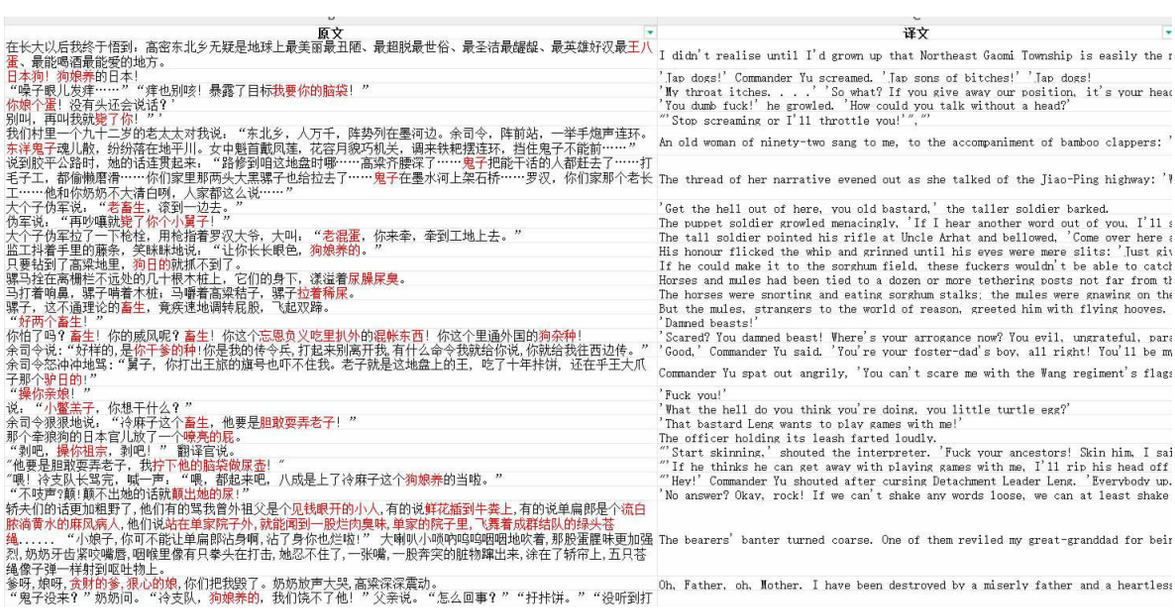
使用 Ernie 进行二次分类时需要联网，免费 Token 有 100 万个。经过大半天的运行，我们得到了 645 句分类为谚语俗语的句子，经过人工筛选后得到 164 句，如图 3 所示。



原文	Ernie判断	人工判断
肚子咕噜噜响一阵。	有。这句有谚语、俗语或古诗词的成分。	Y
宝蓝色的天空深窿无边，绿色的星辰格外明亮。 北斗勺子星——北斗主死，南斗簸箕星——南斗司生，八角玻璃井——缺了一块砖	有谚语、俗语或者古诗词的成分。	Y
大叔……你别走，不看僧面看佛面，不看鱼面看水面，不看我的面子也看豆官的面子上，留下吧，你要我……我也给你……你就像	这句有谚语成分。这句台词中的“不看僧面看佛面，不看鱼面看水面	Y
大叔，”父亲说，“大叔，你挂彩了。”	这句文本中包含了一句谚语或俗语。答案是“有”。	Y
你爹多能干，十五岁就杀人，杂种出好汉，十有九个都不善……	这句话有谚语、俗语或古诗词的成分。 “杂种出好汉，十有九个都	Y
送他点钱，送他盒烟都行，不打勤的，不打懒的，单打不长眼的。’	这句话有一些谚语、俗语或者古诗词的成分，例如“不打勤的，不打	Y
民夫们舔着嘴唇，眼巴巴地盯着那几只空桶里残存的米粒，但没人敢去动。	有，这句里面有谚语、俗语或者古诗词的成分。	Y
在那片新开辟出的空地上。骡马都拴在绿疃或石磙子上。高粱尸横遍野。骡马无精打采地叼吃着被揉烂压扁的高粱茎叶。	这句话里有谚语成分，具体为“高粱尸横遍野”。这句谚语的意思是	Y
夜谐深了，棚栏里痘气逼人。罗汉大爷无法入睡。	这句话里包含了谚语成分，可以简单回答是。这句俗语就是用来描述	Y
从东边的高粱地里，露出了一弧血红的朝阳，阳光正正地照着罗汉大爷半张着的黑洞洞的嘴。	这句话有谚语、俗语或者古诗词的成分，即“从东边的高粱地里”。	Y
余司令说：“都到河堤漫坡上趴着，省得鬼子来了措手不及。”	有。这句有谚语、俗语或者古诗词的成分，即“省得鬼子来了措手不	Y
那时候眉月初升，低低地压着枯树枝桠。	有。这句里有谚语、俗语或者古诗词的成分。	Y
然后你们就鸡走鸡道，狗走狗道，井水不犯河水。’	这句话有谚语、俗语或古诗词的成分，可以理解为“各走各的路，互	Y
放下碗，他说：“余司令，兄弟不胜酒力，告辞啦！”	这句话中包含了谚语成分，谚语成分体现在“告辞啦”这个口头表达	Y
父亲老羞成怒。	有谚语成分。这句表达了父亲因为感到羞耻而变得愤怒。	Y
村里流传着罗汉大爷尸体失踪的消息，一传十，十传百，一代传一代，竟成了一个美丽的神话故事。	这句话含有谚语、俗语或古诗词的成分。常见的古诗词或成语可能表	Y
后来又鬼鬼祟祟地跳出一只火红的狐狸。	这句话含有谚语或俗语成分。常见包含类似表达的谚语有“鬼鬼祟祟	Y
东北乡地势低洼，往往秋水泛滥，高粱高秆防涝，被广泛种植，年年丰产。	有谚语成分。	Y
奶奶嫁到单家，其实也是天意。那天，我奶奶在秋千架旁与一些尖足长辫的大闺女耍笑游戏，那天是清明节，桃红柳绿，细雨霏霏，人面桃花，女儿解放。		
轿子起行，忽悠悠似腾云驾雾。	有谚语成分。回答如下：有。	Y
高粱的味道深入人心。	有谚语、俗语或者古诗词的成分。有“轿子起行，忽悠悠似腾云驾雾。	Y
高粱地里悠长的哭声里，夹杂着疙疙瘩瘩的字眼：青天哟——蓝天哟——花花绿绿的天哟——棒槌哟亲哥哟你死了——可就塌了妹妹的天哟——。	这句有谚语、俗语或者古诗词的成分。	Y
村里的人，都集中到我家烧酒作坊前一个土场上。这场上曾经铺红叠丹，堆满了我家收购的红高粱。那时候奶奶常手持白尾拂尘，蹦蹦移动着小脚，看着我家酿醺醺的伙计，用木斗收购高粱，奶奶的脸上染着灿烂的酮霞。	这句似乎包含了谚语和俗语。谚语和俗语通常是口头流传的智慧，反	Y
手舞足蹈，眼睛笔直，腮上肉跳，胡言乱语，口吐白沫，扑地跪倒，喊着：“大哥大哥大哥，太君让我干，我不敢不干……你死后升了天，骑白马，佩雕鞍，穿鳝袍，坠金鞭……”村里人见他这样，也就把恨他的心谅了。	这句话有谚语成分，含有成语“蹦蹦移动”、“酮醺醺”。至于	Y
司令，王子犯法，一律同罪！’	这句话有一些俗语或民间文学的元素。具体的语境很难具体判断。根	Y
奶奶给我父亲缠了一道白孝布——奶奶自己也是披麻戴孝。	有谚语、俗语或者古诗词的成分。	Y
曲调十分熟悉，是『高粱红了，高粱红了！』。	这句话有谚语、俗语或者古诗词的成分，可以理解为“落叶归根”。	Y
两爸斗井饼，一爸斗滚到堤南，一爸斗滚到堤北。	这句描述包含了一种谚语、俗语或古诗词的元素，表达了一种人们对	Y
大雨过后三天，路面依然潮湿，高粱地里白色蒸气腾腾升集，绿高粱被白气缭绕，俱有了仙风道骨。	有谚语成分。	Y
你好大的福气啊，	有。这句文本表达的内容像是俗语或者古诗词的一部分。	Y
三天新生活，如同一场大梦惊破，有人在一分钟内成了伟大领袖，奶奶在三天中参透了人生禅机。	有谚语成分。	Y
奶奶说勾兑时，故意张扬示从，做出无限神秘状，使偷窥者毛发森森，以为我家通神人魔，是天助的买卖。	这句似乎含有谚语、俗语或古诗词的成分。	Y
	这句内容似乎含有谚语或俗语的成分。具体来说，这句话中“天助的	Y

图 3　机器分类和人工筛查的结果

最后我们再次使用 ChatGLM3-6B 对粗语再次分类和人工筛选，并和 Bert 模型的结果合并去重后得到 155 句粗语，如图 4 所示。



原文	译文
在长大以后我终于悟到：高密东北乡无疑是地球上最美丽最丑陋、最超脱最世俗、最圣洁最龌龊、最英雄好汉最王八蛋、最能喝酒最能爱的地方。	I didn't realise until I'd grown up that Northeast Gaomi Township is easily the r
日本狗！狗娘养的日本！	'Jap dogs!' Commander Yu screamed. 'Jap sons of bitches!' 'Jap dogs!
“嗓子眼儿发痒……”“痒也别咳！暴露了目标我要你的脑袋！”	'My throat itches. . . .' 'So what? If you give away our position, it's your head
你娘个蛋！没有头还会说话？’	'You dumb fuck!' he growled. 'How could you talk without a head?'
别叫，再叫我就毙了你！”’	"'Stop screaming or I'll throttle you!'""
我们村里一个九十二岁的老太太对我说：“东北乡，人万千，阵势列在墨河边。余司令，阵前站，一举手炮声连环。东洋鬼子魂儿散，纷纷落在地平川。女中魁首戴凤莲，花容月貌巧机关，调来铁耙摆连环，挡住鬼子不能前……”说到胶平公路时，她的话连贯起来：“路修到咱这地盘时哪……高粱齐腰深了……鬼子把能干活的人都赶去了……打毛子工，都偷懒磨滑……你们家里那两头大黑骡子也给拉去了……鬼子在墨水河上架石桥……罗汉，你们家那个老长工……他和你奶奶不大清白咧，人家都这么说……”	An old woman of ninety-two sang to me, to the accompaniment of bamboo clappers: '
大个子伪军说：“老畜生，滚到一边去。”	The thread of her narrative evened out as she talked of the Jiao-Ping highway: 'W
伪军说，“再吵嚷就毙了你个小舅子！”	'Get the hell out of here, you old bastard,' the taller soldier barked.
大个子伪军拉了一下枪栓，用枪指着罗汉大爷，大叫，“老混蛋，你来牵，牵到工地上去。”	The puppet soldier growled menacingly, 'If I hear another word out of you, I'll s
监工抖着手里的鞭条，笑眯眯地说，“让你长长眼色，狗娘养的。”	The tall soldier pointed his rifle at Uncle Arhat and bellowed, 'Come over here s
只要钻到了高粱地里，狗日的就抓不到了。	His honour flicked the whip and grinned until his eyes were mere slits: 'Just giv
骡马拴在高栅栏不远处的几十根木桩上，它们的身下，漾溢着尿臊屎臭。	If he could make it to the sorghum field, these fuckers wouldn't be able to catch
马打着喷鼻，骡子啃着木桩；马嚼着高粱秸子，骡子拉着稀屎。	Horses and mules had been tied to a dozen or more tethering posts not far from th
骡子，这不通理论的畜生，竟疾速地调转屁股，飞起双蹄。	The horses were snorting and eating sorghum stalks; the mules were gnawing on the
“好两个畜生！”	But the mules, strangers to the world of reason, greeted him with flying hooves.
你怕了吗？畜生！你的威风呢？畜生！你这个忘恩负义吃里扒外的混帐东西！你这个里通外国的狗杂种！	'Damned beasts!'
余司令说：“好样的，是你干爹的种！你是我的传令兵，打起来别离开我，有什么命令我就给你说，你就给我往西边传。”	'Scared? You damned beast! Where's your arrogance now? You evil, ungrateful, para
余司令怒冲冲地骂：“舅子，你打出王旅的旗号也吓不住我。老子就是这地盘上的王，吃了十年拌饼，还在乎王大爪子那个驴日的！”	'Good,' Commander Yu said. 'You're your foster-dad's boy, all right! You'll be m
“操你亲娘！”	Commander Yu spat out angrily, 'You can't scare me with the Wang regiment's flags
说，“小鳖羔子，你想干什么？”	'Fuck you!'
余司令狠狠地说，“冷麻子这个畜生，他要是胆敢耍弄老子！”	'What the hell do you think you're doing, you little turtle egg?'
那个牵狼狗的日本官儿放了一个嘹亮的屁。	'That bastard Leng wants to play games with me!'
“剥吧，操你祖宗，剥吧！”翻译官说。	The officer holding its leash farted loudly.
“他要是胆敢耍弄老子，我拧下他的脑袋做尿壶！“	"'Start skinning,' shouted the interpreter. 'Fuck your ancestors! Skin him. I sai
“喂！冷支队长骂完，喊一声，“喂，都起来吧，八成是上了冷麻子这个狗娘养的当啦。”	"'If he thinks he can get away with playing games with me, I'll rip his head off
“不吱声？颠！颠不出她的话就颠出她的屎！”	"'Hey!' Commander Yu shouted after cursing Detachment Leader Leng. 'Everybody up.
轿夫们的话更加粗野了，他们有的骂我曾外祖父是个见钱眼开的小人，有的说鲜花插到牛粪上，有的说单扁郎是个流白脓滴黄水的麻风病人，他们说站在单家院子外，就能闻到一股烂肉臭味，单家的院子里，飞舞着成群结队的绿头苍蝇…… “小娘子，你可不能让单扁郎沾身啊，沾了身你也烂啦！” 大喇叭小唢呐呜呜咽咽地吹着，那股蛋腥味更加强烈，奶奶牙齿紧咬嘴唇，咽喉里像有只拳头在打击，她忍不住了，一张嘴，一股奔突的脏物躁出来，徐在了轿帘上，五只苍蝇像子弹一样射到呕吐物上。	'No answer? Okay, rock! If we can't shake any words loose, we can at least shake
爹呀，娘呀，贪财的爹，狠心的娘，你们把我毁了。奶奶放声大哭，高粱深深震动。	The bearers' banter turned coarse. One of them reviled my great-granddad for beir
“鬼子没来？”奶奶问。“冷支队，狗娘养的，我们饶不了他！”父亲说。“怎么回事？”“扦拌饼。”“没听到打	Oh, Father, oh, Mother. I have been destroyed by a miserly father and a heartless

图 4　粗语进行机器分类和人工筛查的结果

五、总结

我们通过利用多种语言模型对《红高粱家族》及其译文按句子进行对齐，以及对粗语和谚语俗语进行分类，最后得到 155 句粗语的平行语料以及 164 句谚语俗语的平行语料。该实验为下一步生态翻译学研究奠定了基础，为相关研究提供借鉴作用。

通过此项研究，我们感觉到语言模型对于特殊领域的语义特征的识别在性能上依然不足，同时缺少能直接使用的面向专业领域的模型，期待未来的语言工具能在某一垂直领域做得更智能化。

另外我们总结了此次研究工作中的不足和未来可以改进的地方:1. 我们使用 ChatGPT 进行语料对齐时,由于人工检查的不到位,语料有某些部分存在错误,例如句子重复、中英文位置错误等;2. 训练的文本数据量较少,可以尝试进行数据增强,利用已有的大模型进行句子改写和生成;3. 我们从中文的角度出发分类粗语和谚语俗语,或许从英文的角度出发进行分析可能得到不一样的数据;4. 尝试多种提示语,或者多种提示语结合,或者将句子分割进行多段分析综合判断。

参考文献：

[1] J. Devlin, M. W. Chang, K. Lee, K. Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[A]. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies [C]. Kerrville: Association for Computational Linguistics, 2019.

[2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin. Attention Is All You Need[A]. NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems[C]. New York: Curran Associates Inc., 2017.

[3] Jian Zhu, Zuoyu Tian, Sandra Kübler. UM-IU@LING at SemEval-2019 Task 6: Identifying Offensive Tweets Using BERT and SVMs [EB/OL]. https://arxiv.org/abs/1904.03450, 2019-4-6.

[4] Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, Hua Wu. ERNIE: Enhanced Representation through Knowledge Integration [EB/OL]. https://arxiv.org/abs/1904.09223, 2019-4-19.

Classifying Sentences of Long Textual Corpus by Using Various Language Models

LI Xin[1], SUN Run[2]

(1. Language Training Center, College of Foreign Languages, Guangdong Ocean University, Zhanjiang Guangdong 524088;
2. College of Foreign Languages, Guangdong Ocean University, Zhanjiang Guangdong 524088, China)

Abstract: In order to conduct research in eco-translation studies, we carried out a series of data extraction experiments that involved using multiple language models to classify sentences in lengthy corpora. This study aims to explore the application of language models in the context of literary works research to enhance the quality and efficiency of literary studies. By collecting research and training corpora, and employing various language models including Bert, ChatGPT, Claude, and ChatGLM, we performed classification and filtering on a large-scale text corpus. Through this experiment, we hoped to identify semantic features such as profanity and idioms, and provide a deeper understanding and support for cross-language translation. In the experimental results, we observed that utilizing multiple language models allowed for limited classification of sentences in lengthy corpora, thus alleviating the workload of manual classification to some extent. Additionally, this experimental data can provide support for related translation research.

Key words: "Red Sorghum"; parallel corpus; Bert; ChatGPT; ChatGLM; Ernie