2023年上海交通大学数学建模校内赛

（本赛题的论文写作要求与全国大学生数学建模竞赛一致，请先阅读“全国大学生数学建模竞赛论文格式规范”）

B 题 大语言模型

ChatGPT的成功引起人们对大型语言模型（LLM,Large Language Model）兴趣。LLM已经在自然语言理解类任务（文本分类、句子关系判断、情感倾向判断、命名实体识别）和自然语言生成类任务（机器翻译、文本摘要、应答系统）都获得广泛应用。自2017年以来，transformer已经成为自然处理语言(NLP)的(学术和行业)标准。通过更好理解用户的询问，Google用BERT提高搜索引擎能力。来自OpenAI的GPT系列能产生更适合人类的文本和图像。这些成功都是基于transformer的注意力(attention)机制[1]。它和迁移学习和扩展神经网络成为该行业的主流观点。

Hugging Face Transformers库提供了这方面的一个生态系统，包括模型、数据和应用等。用户往往如下操作：(1)从该网站下载合适的预训练模型和相应数据；(2)根据用户数据对预训练模型微调，对微调后模型进行评价； (3)应用自己模型或者在该网站上发布模型和数据。**请完成问题1到问题5中的3个问题(5选3)，以及第6个问题。**

问题1：机器翻译是通过机器把一种语言翻译为另一种语言。(1)分析机器翻译的原理，特别是注意力机制。请给出一些评价翻译质量的指标。(2)下载相关的预处理模型和数据，用两个中英互相翻译的例子加以验证。

问题2：情感倾向判断是对一段文字判断它属于哪种情感（愤怒、恐惧、高兴、喜爱、悲伤、惊喜等）。(1)分析情感倾向判断的原理。(2)下载相关的预处理模型和数据，详细说明下载的数据。构造6段文字用程序验证它们分别属于上述的6种情感类别。

问题3：文本生成是在用户提示下让机器自动生成后续文本。(1)分析文本生成的原理，并讨论生成文本质量的指标。(2)下载相关的预处理模型和数据，在适当提示下生成一篇2000字的英文稿或者500字的中文稿。

问题4：文本摘要生成是把用户提供的一段文字用机器自动生成一段摘要。(1)分析文本摘要的原理，并讨论文本摘要质量的指标。(2) 下载相关的预处理模型和数据，用两个例子加以验证。请分析哪些因素和文本摘要有很大关系。

问题5：应答系统是指机器回答用户的问题。聊天机器人属于这类系统。(1) 分析应答系统的原

理，并讨论应答系统质量的指标。(2) 下载相关的预处理模型和数据，请用两个例子加以验证，每个例子至少和机器进行10次交互。

问题6：请解决以下两个子问题。(1) 以上述的某一个应用为例，请找相关数据对预训练模型进行训练，产生新模型。(2)和原来的预训练模型想比较，用例子说明新模型的优点，比如预测的准确度得到提高。

参考文献

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin, Attention Is All You Need, https://arxiv.org/abs/1706.03762

[2] Lewis Tunstall, Leandro von Werra, and Thomas Wolf, Natural Language Processing with Transformers Building Language Applications with Hugging Face, Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

[3] Transformer在翻译中的应用的介绍： https://jalammar.github.io/illustrated-transformer/

[4] The Transformer Family：https://lilianweng.github.io/posts/2020-04-07-the-transformer-family/

[5] Xavier Amatriain, TRANSFORMER MODELS: AN INTRODUCTION AND CATALOG, https://arxiv.org/pdf/2302.07730.pdf

[6] 通向AGI之路：大型语言模型（LLM）技术精要。https://zhuanlan.zhihu.com/p/597586623