Project Proposal

云南大学 信息学院

Joseph Tomasulo
jtoma5@protonmail.com

2020 年 3 月 18 日

A The Problem

Many students and professors at Yunnan University publish papers in English. Since English is not a common native language for this group, this can cause some difficulty. Some methods to overcome this problem are to improve one's English, ask for help from more linguistically-proficient peers, or even to pay third parties to edit a paper before submission. Still, papers in the field of Natural Language Processing commonly contain numerous grammatical errors and repetitive or clumsy phrasing is almost the norm. This seems to be the case generally, and the issue is not specific to publications from any particular country. A consequence worth avoiding might be that the attention given to research is less proportional to the quality of the science and more proportional to the fluency with which it is presented.

本校很多老师和学生都用英文写论文，但是因为英语不是母语，有的时候会遇到困难，所以有的人很努力的学习英语，有的依靠英语特别棒的朋友，还有另一种选择就是在提交论文之前付钱给第三方来修改。怎么去判断一片论文的英语？Word 的语法检查可靠吗？我们都希望一片论文受不受欢迎是个科学质量和重要性的问题而不是一个语言的问题。

A.1 Proposed Partial Solution: Automatic native language identification

Native language identification (NLI) is an automated procedure that takes text input and outputs the predicted native language of the author. A high performance NLI model could serve as a benchmark for researchers wishing to submit a paper in a given language.

母语识别是个自动过程，输入一个文档，输出该文档的作者的母语。高性能的 NLI 模型可以作为研究人员的语言质量基准。

A.2 Significance:

OpenAI staggered the release of GPT-2 out of fear that the text generation capabilities of the largest model could be used for nefarious purposes. This caused an uproar in the community and it shows that there is a need for additional metrics for classifying text. In her TWIML interview, Nasrin Mostafazadeh of Elemental Cognition suggested that we have no way of evaluating whether these models, that can produce quite convincing text, have the common sense of even a young child. Common sense being too wishful a goal, the following strategy lays out a plan to identify a native's grasp of language.

由于担心最大模型的文本生成功能可用于恶意目的，OpenAI 错开了 GPT-2 的发布。这在社区引起轩然大波，表明需要更多的标准用来分类文本。在她的 TWIML 访谈中，Nasrin Mostafazadeh (Elemental Cognition) 建议我们无法评估这些能产生相当有说服力的文本的模型。她认为这些模型其实没有小孩子的常识。常识是太一厢情愿的一个目标，下面列出了一个计划，以识别人对母语的掌握。

B Research Strategy

B.1 Data Acquisition

A dataset called TOEFL 11 was designed for this task in 2013 by Blanchard, et al. This dataset is commercial and as such, probably unavailable. Still, since in NLP, many papers are publicly available on arxiv.org, it should be possible to procure enough text. So the main strategy for data acquisition will be to scrape or use one of the bulk download tools provided by arxiv.org.

即使已经有为识别母语而设计的数据集 (TOEFL11, ICLE)，它们是商业性的，并且用托福考试之类的英语而不是科学文章，因此最好编写新的数据集。幸运的是 arXiv.org 拥有来自世界各地的免费科学文章。他们也提供一个下载工具。

B.2 Dataset Creation

Since the author's native language is not usually available in metadata, it will have to suffice to rely on other characteristics to make this judgment. None of these will be perfect. One solution is to hold a list of universities and the countries in which they are situated. Because of the relatively low number of exchange students in many countries, this should produce reasonable data. To enhance this criterion, holding a list of most common names by country could be used to filter out potential exchange students.

由于元数据中没有作者的母语，因此有必要使用其他规则，所以数据将不是完美的。由于外国学生的人数通常比本国学生少得多，一种方法是把大学附属关系作为母语指标。为了提高这个标准，可以使用按国家列出的最常见的名字来筛选外国学生的论文。

B.3 Classification

There are many ways this may be done. One begins by encoding the text using pretrained language models such as BERT or GPT-2. Each paper would need to be broken up into chunks according to the size of GPU RAM. Then train a model to classify the native language of the author of encoded text. Finally aggregate the results for each paper and make a prediction.

因为之前的母语识别工作是在大型语言模型出现之前进行的，所以测量像 BERT 或 GPT-2 这样模型的性能应该很有趣。

B.4 Delivery

Although it would be possible to distribute a version that works clientside, since people would need such a service only when getting ready to submit a paper, it is sufficient to host the application on a server and ask users to connect through a web application.

尽管可以发布一个客户端运行的模块，但由于人们只有在准备提交论文时才需要这样的服务，因此在服务器上运行模型并开发一个 web 应用程序就足够了。