Project Background

云南大学 信息学院

Joseph Tomasulo
jtoma5@protonmail.com

2020 年 5 月 10 日

A Introduction

In computational linguistics, Native Language Identification (NLI) is the task of determining the native language of the author of a text. It may prove to be directly useful for Second Language Acquisition (SLA) researchers. For example, teachers could prepare certain teaching materials that target common errors for Chinese learners of English and others targeting errors that Indian learners of English are likely to make. For example, there are already clear differences between the content and vocabulary of English textbooks for Chinese students and those for French students. But the grammar lessons that they emphasize are largely the same, and any differences are not attributed in a systematic way to errors that one group is likely to make. There is an opportunity for improvement here and NLI migt be able to help. 在计算语言学中，母语识别是确定文本作者的母语的任务。这对第二语言习得研究者来说可能是直接有用的。如果问题的到解决，那么准备针对特定群体的语言学习者最容易犯错误的教材可能是有益的。例如，中国学生和法国学生的英语文本在内容和词汇上已经有明显的差异。但是他们强调的语法基本上是一样的，任何差异不是因为最有可能犯的错误的区别。这里有一个改进的机会，NLI 可能会提供帮助。

Another field of study that would be improved by an improvement in NLI is 的 forensic linguistics. NLI is considered a subtask in both authorship identification and plagiarism detection. 此外，随着 NLI 的提高，法医语言学的研究领域将会的到改进。NLI 被认为是作者识别和剽窃检测中的一个子任务。

B Background

B.1 International Corpus of Learner English

The international Corpus of Learner English is a dataset that six thousand contains essays written in English by intermediate and advanced students from 16 different native language backgrounds. There were multiplie topics and the number of essays written about each topic was not constant across language backgrounds. A system could guess the native language of the author just by looking at the topic. 国际英文学习者语料库是一个数据集，包含 16 个不同母语北京的中高级学生用英语写的六千篇论文。有了很多不同的主题，而且在不同的语言北京下，关于每个主题的论文数量并不一样。系统可以通过主题来猜测作者的母语。所以此数据集现在不再用于做母语识别。

B.2 TOEFL11

TOEFL11 was created specifically for NLI. It contains the written responses for the TOEFL exam of ten tousand students fro m11 different language backgrounds. There are eight different prompts and there are approximately equal numbers of students writing abotu each topic from each language background. 托福 11 是专门为母语识别而创建的数据集。它包含来自 11 不同语言北京的一万名学生的托福考试写作。学生们从八个不同主题选出。大约有同等数量的学生从每一种语言北京写每一个主题。

B.3 NLI Shared Task 2013

There was a competition organized around this dataset in 2013. The competition had three separate phases conducted on TOEFL11 data, any other data, and both TOEFL11 and any other data. 29 teams participated and the most successful teams all used Support Vector Machines (SVM). They differed in the features they chose as input. The choices included character n-grams, word n-grams, POS n-grams. Most teams also used a form a weighting such as log-entroy or TF-IDF. For teams that competed in all three phases there was a consistent result. The accuracy using TOEFL11 was much higher than using any other dataset, but adding other data did slightly improve accuracy. 2013 年有一场关于这个数据集的竞赛。竞赛分为三个阶段而进行：托福 11 数据，任何其他数据，托福和任何其他数据在一起。29 个团队参加，最成功的团队都使用支持向量机。他们选择不同的特征作为输入。选项包括字符 n-grams，单词 n-grams，和 POS n-grams。大部分团队还是用了一种加权形式，如对数熵或 TF-IDF。参加每个阶段的团队都表现相似的结果，使用托福 11 数据精确度比使用任何其他数据集高的多，但是添加其他数据稍微提高了精确度。

B.4 NLI Shared Task 2017

The competition was held again in 2017 with a modified format. In addition to the written essay task, there was also a speech-based task. The speech-based component is outside the scope of this work. Because the text-based component was the same as that of 2013, direct comparison is possible. Owing to the use of various ensemble methods, the best accuracy improved by around 5%. Importantly, teams that used neural networks performed worse than those that used SVM's. 比赛于 2017 年再次举行，形式有所改变，除了写作任务之外，还有口语的任务。口语那个任务不在这个项目的范围内。由于写作部分与上次比赛相同，因此可以直接进行比较。这次最厉害的团队使用了集成技术，精确度提高了 5%。重要的是，有一些团队尝试使用神经网络但是结果没有支持向量机那么好。

B.5 Modified Plan

Since 2013, there has been a trend towards using distributed representations to model language. This advance was enabled by computationaly efficient methods of using neural networks to compute the distributed representations. People said that in any NLP problem, you could add a distributed representation and improve the SOA. When I chose this project, I had only heard about the 2013 competition. I thought the participants hadn't heard of Word2Vec, a popular neural network-based way to make distributed representations. But I can't believe the participats of the 2017 competition did not know about it. Because the chose not to use it, I am questioning whether it will work. I still intend to try it, but I am also preparing to use lexical n-grams with an SVM. 自 2013 年以来，有一种趋势是使用分布式表示来建模语言。这一进展是通过使用神经网络计算分布式表示的高校计算方法实现的。人们说，在任何 NLP 问题中，都可以添加一个分布式表示，这将提高最新技术。当我选择这个项目时，我只听说了 2013 的竞赛。我认为作者不熟悉 Word2Vec，一种流行的基于神经网络的创建分布式表示的方法。但我不相信 2017 年比赛参赛者对此一无所知。因为他们没有使用 word2vec，所以我怀疑它是否有效。我仍然打算尝试，但我也准备使用他们的方法，就是词汇 n-grams 和一个支持向量机。

One thing to note is that since TOEFL11 is a dataset of essays from an English exam, there are bound to be many spelling errors. But with the arXiv dataset that I am making, there should be very few spelling errors as they have all been proofread. As a consequence, the character n-grams that were useful in the two competitions are unlikely to be as useful in my case. 值得一提，由于托福 11 是一个来自英语考试的数据集，肯定会有很多拼写错误。但我在准备的 arXiv 数据集应该很少有拼写错误，因为 arXiv 的论文都已经经过校对。因此，字符 n-grams，一个比赛中很强的特真，对我可能没有用处。