中文作家写作风格识别

15307130194

冯梓源

摘要

在文学领域，作家的写作风格是一个相对抽象的概念。在中文领域，怎样让计算机理解并识别作家的写作风格，是一个极具挑战性的命题。本文试图在几篇经典论文的思路上做进一步探索，充分利用课上所学的NLP相关概念与方法，结合以FudanNLP、NLTK为代表的高性能工具包，将中文作家的写作风格以合适的方式编码，借助机器学习甚至深度学习的技术，在一定问题范围内训练分类器，有效解决中文作家写作风格识别的问题。对于新的文本，分类器能够以令人满意的准确率识别可能的作者。

关键词：写作风格、句法依存、循环神经网络、LSTM

问题背景

古今中外，著名作家的优秀作品总是被人们口耳传颂，除了作品本身的主题以外，作家独具一格的遣词造句能力，也是优秀作品的不可或缺的要素。跟其他任何艺术创作相同的是，作家在创作过程中必然会在作品中留下自己的文字风格。如何定义并解读作家的写作风格，无论对文学界还是对语言学界都是极富价值的研究命题。

从经验上看，要区别甚至识别作家的行文风格是很困难的。写作风格实际上是一种个人的行为方式，作家在创作过程中会不知不觉地将其个性和个人社会背景融入或体现于作品中[1]。

问题分析

对于写作风格(writing style)，采用[Sebranek et al. (2006](https://en.wikipedia.org/wiki/Writing_style#CITEREFSebranek_et_al.2006), p. 111)[2]的定义：writing style is the choice of [words](https://en.wikipedia.org/wiki/Word), [sentence](https://en.wikipedia.org/wiki/Sentence_(linguistics)) structure, and [paragraph](https://en.wikipedia.org/wiki/Paragraph) structure, used to convey the meaning effectively.

根据该定义，词语、句子和段落结构的信息能够表达作家的写作风格。接下来会围绕这三方面设计和计算特征。值得注意的是，在这种定义下的写作风格是与主题无关的(subject independent)，即作品的主题以及内容相关信息不应该作为写作风格的衡量。

在一定条件下，识别某位作家的写作风格的问题，就转化为特征选取和多元分类的问题。

问题的难点在于两方面。一、怎样在词、句、段落层面上构建合适的特征。二、怎样选择合适的分类器。

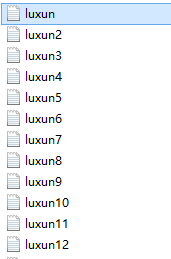
数据收集

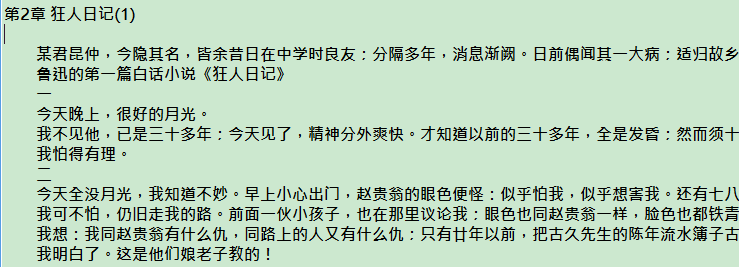
从五六文学网([www.56wen.com](http://www.56wen.com))下载一批中文作家的作品，以txt形式保存。

选取的作家包括：鲁迅、周作人、林语堂、三毛、刘慈欣、王小波、史铁生

每个作家采用收集至少20万字的文本

其中70%作为训练集，10%作为交叉验证集，20%作为测试集





（继续加图！！）

预处理

将文本编码为UTF-8

使用正则表达式匹配，清除文本中的广告。

手动清除非作家本人编写的内容（如编者写的序言、注释）。

划分句子：以现代汉语句子的概念为标准，寻找句号、叹号、问号、冒号、分号、省略号，以此划分每一个句子（特殊情况另外处理），使得每个句子单独占据一行。

分词和词性标注：使用jieba分词的Python包，对每个句子分词，分词的结果附带词性标注，词性标注与“NLPIR汉语分词系统”兼容。

特征选择

一般来说，从文本中提取特征会考虑以下几个方面：

**词汇特征**是基于字符和词语的特征，主要包括词性、词汇丰富度和高频词等 (Abbasi A)。词汇特征在传统英文文学作品作者识别中效果较好；但是由于作品词汇的选择与主题高度相关，词汇特征在跨主题的文本风格分析中效果会受影响。

**语法特征**指的是功能词、标点符号和Ngram等。有研究 (Zhao Y, 2007)表明，在英文语料中功能词能够有效表征作家在写作时的个人语言习惯，对文本识别类型的问题有一定帮助。不同于英语、法语等黏着语，在动词上添加词缀来增强表意能力，汉语属于孤立语（又称分析语），汉语的虚词承担了大部分的语法任务和语义功能 (黄进, 2006)。虚词主要包括：介词、副词、助词、叹词、连词等。

**结构特征**指的是与文本组织和布局相关的特征，如段落数目、段落长度、字体、字号等。显然，这类特征受编辑者的影响很大，同一个作家的作品在不同出版商的手上可能会有不同的字体、字号、缩进等，而且文本的段落设计往往由根据文章的主题需求而定（例如议论文的分段往往带有内容上的逻辑变化，而涉及人物的小说仅仅用分段来表示对话）。所以这类特征在写作风格识别的问题中很难起作用。

**语义特征**是与内容相关的特征，一般来说写作风格识别不应该与内容相关，但是有研究表明，在短文本的文体风格中引入内容相关特征，能让模型表现更好，但是该结论缺乏进一步的验证，在长文本中的表现也未知。

近年的研究

综上所述，选取的文本特征如下：

词汇特征：描述作者使用词汇习惯的指标

词性比例——不同词性的词的数量与总词数之比

词汇丰富度——不同的词数与总词数之比

单现词比例——只出现一次的词的数目与总词数之比

高频虚词（功能词）——介词、连词、叹词、结构助词、语助词、方位词的前若干个高频项，

句法特征：描述作者造句习惯的指标

平均句长——所有句子所含字数（包括标点）的算术平均（服从某个分布？）

短句比例——明显短于平均句长的句子数量与总句数之比

长句比例——明显长于平均句长的句子数量与总句数之比

疑问句比例

感叹句比例

句法依存关系——使用FudanNLP计算每个句子的句法依存关系。

（句法依存关系能够从逻辑上编码一个句子的结构，是最有力的句法特征）

算法及其优化

简单统计学方法

列表分析选取的各种特征在不同作家的文本上的表现：

高频功能词——词云图

句法依存信息统计

相似度衡量：使用两种方法对作家文本风格进行相似度衡量——余弦距离和欧式距离

机器学习方法——多元分类问题

数据增强：将每个作家的全部作品切分为1000份数据，切分的依据是保持段落的完整性，在段间进行切分。在这1000份数据中，使用70%作为训练集，10%作为验证集，20%作为测试集。

聚类

决策树

支持向量机：在对小样本进行分类时，支持向量机具备优势。

深度学习方法

一般来说，深度学习对数据量的要求比较高，训练一个合格的深度模型至少也需要千万级别的数据作为“燃料”。但是在这个问题中，数据是比较缺乏的。一位作家即使再高产，也未必能达到如此巨大的数据量。而且不同作家之间的作品产量相差很大，对深度模型而言这是一个“非平衡样本”的问题。所以，以下对深度模型的使用是一个大胆的尝试，结果未必能够完美重现。

RNN

LSTM

结论

参考文献

[1]胡壮麟.理论文体学[M].北京:外语教学与研究出版社,2000.

*[2]Sebranek, Patrick; Kemper, Dave; Meyer, Verne (2006), Writers Inc.: A Student Handbook for Writing and Learning, Wilmington:*[Houghton Mifflin Company](https://en.wikipedia.org/wiki/Houghton_Mifflin_Company)*,*[ISBN](https://en.wikipedia.org/wiki/International_Standard_Book_Number)[978-0-669-52994-4](https://en.wikipedia.org/wiki/Special:BookSources/978-0-669-52994-4)

[3] 武晓春,黄萱菁,吴立德. 基于语义分析的作者身份识别方法研究[J]. 中文信息学报, 2006, 20(6): 63-70. (武晓春, 2006)

Abbasi AH.Chen. (无日期). Applying authorship analysis to extremist-group web forum messages. IEEE Intelligent Systems, 页 67-75.

Zhao YJZobel. (2007). Searching with style: Authorship attribution in classic literature[C]. In Proceedings of the 30th Australasian Computational Science Conference, 页 59-68.

武晓春黄萱菁,吴立德. (2006). 基于语义分析的作者身份识别方法研究[J]. 中文信息学报, 页 63-70.

# 引用作品

Abbasi AH.Chen. (无日期). Applying authorship analysis to extremist-group web forum messages. IEEE Intelligent Systems, 页 67-75.

Zhao YJZobel. (2007). Searching with style: Authorship attribution in classic literature[C]. In Proceedings of the 30th Australasian Computational Science Conference, 页 59-68.

武晓春黄萱菁,吴立德. (2006). 基于语义分析的作者身份识别方法研究[J]. 中文信息学报, 页 63-70.