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。

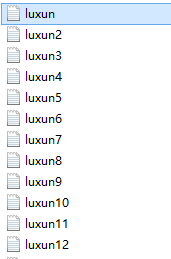
在一定条件下，识别某位作家的写作风格的问题，就转化为特征选取和多元分类的问题。

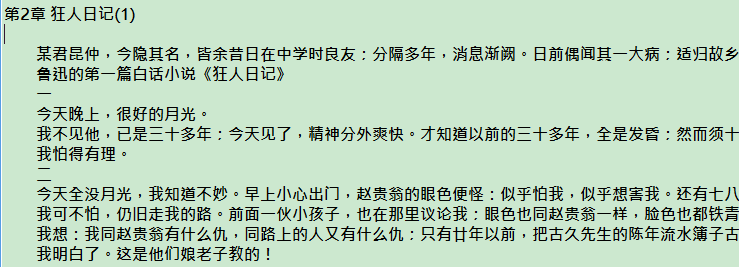
问题的难点在于两方面。一、怎样在词、句、段落层面上构建合适的特征。二、怎样选择合适的分类器。

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





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

借鉴一系列关于英文写作风格的研究，发现词汇特征、语法特征、结构特征和语义特征是常被选择的西大方面。但考虑到中文作家的文本，需要做些调整：

1. 一位作家可能有多种不同主题的作品，而词汇特征在跨主题应用时效果不好；

2. 结构特征（与文本组织和布局相关）易受出版社排版、后期编辑的影响，不宜用来衡量作者的风格；

3. 现有研究表明，中文虚词能有效反映作者的写作习惯，在文体识别上有用。

4. 关于语义特征，有基于HowNet知识库的方法[3]，但由于版权问题无法使用。

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，

（虚词的使用具有主题独立性，能够表示写作风格；而名词、形容词等实词往往跟文章主题密切相关）

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

衡量指标：混淆矩阵

简单统计学方法

相似度衡量

机器学习方法——多元分类问题

聚类

决策树

SVM

深度学习方法

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.