

本科毕业设计(论文)

学院(部)	计算机科学与技术学院		
题　　目	面向科技文献领域的 文本纠错系统实现		
年　级	2019 级	专业	计算机科学与技术
班　级	图灵班	学号	1927405039
姓　名	乔子恒		
指导老师	李正华	职称	教授
论文提交日期	2023 年 5 月		

苏州大学
本科毕业设计（论文）独创性声明

本人郑重声明：所提交的本科毕业设计（论文）是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本设计（论文）不含其他个人或集体已经发表或撰写过的研究成果。对本文的研究作出重要贡献的个人和集体，均已在文中以明确方式标明。本人承担本声明的法律责任。

作者签名：乔子恒　　日　期：**2023-5-9**

苏州大学
本科毕业设计（论文）使用授权声明

本人完全了解苏州大学关于收集、保存和使用本科毕业设计（论文）的规定，即：本科毕业设计（论文）的著作权以及文中研究成果的知识产权归属苏州大学。苏州大学有权向国家有关部门或第三方机构送交毕业设计（论文）的复印件和电子文档，允许毕业设计（论文）被查阅和借阅，可以采用影印、缩印或其他复制手段保存和汇编毕业设计（论文），可以将毕业设计（论文）的全部或部分内容编入有关数据库进行检索。

涉密设计（论文）☐

本设计（论文）属　　在_____年___月解密后适用本规定。

非涉密设计（论文）☑

论文作者签名. 乔子恒　　日　期：**2023-5-9**

导 师 签 名：　　日　期：**2023-5-9**

面向科技文献领域的文本纠错系统实现

摘 要

随着信息技术的快速发展，阅读、写作科技文献已经成为大学生、研究生以及学者工作中不可或缺的一部分。然而，由于科技文献通常包含大量的专业术语和复杂的语言结构，因此很容易出现各种文本错误，如拼写错误、标点错误、语法错误等，都会影响到文献的可读性和可信度。本文尝试搭建一个面向科技文献的文本纠错系统，自动检测并纠正科技文献中的错误，辅助人工校对，提高研究人员的工作效率。

我们采用 Flask 后端框架和 Vue 前端框架，使用 SQLite 数据库存储信息，采取前后端分离和分布式系统技术设计并实现了一款面向科技文献领域的文本纠错系统。

在机器学习建模方面，我们的系统基于目前最先进的自然语言处理和深度学习技术，使用 Seq2Seq 和 Seq2Edit 模型对用户提交的论文进行纠错，找出其中的常见文本错误并进行修改。

首先，我们使用爬虫技术获得大量科技文献资料，并提取出文本内容。进而，我们利用数据增强技术得到大规模科技文献领域的文本纠错数据，同时结合真实通用领域文本纠错数据集对模型进行训练，取得了不错的的性能。最后，我们使用模型集成技术，结合多个纠错模型进行投票，使纠错结果的准确率大幅提高，降低了过度纠错的可能性。同时，利用模型集成实现的多种纠错模式也为用户提供了更人性化的纠错体验。

基于本文工作，我们认为面向科技文献领域的文本纠错系统具有很强的现实意义与未来前景，但是目前模型的效果还有很大的提升空间。

关键词：文本纠错；神经机器翻译；数据增强；Flask；Vue

Implementation of a Grammatical Error Correction System for Scientific and Technological Literature

Abstract

With the rapid development of information technology, reading and writing scientific and technological literature has become an indispensable part of the work of college students, graduate students and scholars. However, scientific and technological literature typically includes a large number of specialized terms and complex language structures, which can lead to various text errors, such as spelling errors, punctuation errors, and grammar errors, all of which can affect the readability and credibility of the literature. This article attempts to build a grammatical error correction(GEC) system for scientific and technological literature, which automatically detects and corrects errors in literature, assists manual proofreading, and improves the efficiency of research.

We adopted Flask backend framework and Vue frontend framework, used SQLite database for information storage, and implemented a GEC system for the scientific and technological literature field by employing front-end and back-end separation and distributed system technology in the design.

In the aspect of machine learning modeling, our system is based on the state-of-the-art natural language processing and deep learning technologies. It uses Seq2Seq and Seq2Edit models to correct the papers submitted by users, identifying common textual errors and making corresponding modifications.

Firstly, we use web crawler technology to obtain a large number of scientific and technological literature materials and extract the text content. Then, we employ data augmentation techniques to obtain a large-scale text correction dataset in the scientific and technological literature field, and train the model with a combination of real-world, general-domain GEC datasets, achieving good performance. Finally, we use model ensemble techniques, integrating multiple error correction models for voting, which greatly improves

the accuracy of the error correction results and reduces the likelihood of excessive corrections. At the same time, the various GEC modes implemented through model ensembles also provide users with a more user-friendly GEC experience.

Based on the work presented in this article, we believe that the GEC system for the scientific and technological literature domain has strong practical significance and future prospects. However, there is still much room for improvement in the current model's performance.

Keywords: Grammatical Error Correction(GEC); Neural Machine Translation(NMT); Data Augmentation; Flask; Vue

目 录

前 言 ..... - 1 -

第一章 绪论 ..... - 2 -

1.1 课题背景 ..... - 2 -

1.2 系统开发目的和意义 ..... - 2 -

1.3 课题内容 ..... - 3 -

1.4 论文组织结构 ..... - 3 -

第二章 系统相关技术介绍 ..... **- 5 -**

2.1 系统架构 ..... - 5 -

2.2 WEB 开发工具 ..... - 6 -

2.2.1 Flask 框架 ..... - 6 -

2.2.2 Vue 前端框架 ..... - 7 -

2.2.3 Element UI 组件库 ..... - 7 -

2.3 数据库技术 ..... - 8 -

2.4 本章小结 ..... - 9 -

第三章 纠错相关技术介绍 ..... **- 10 -**

3.1 纠错模型 ..... - 10 -

3.1.1 基于 NMT（神经机器翻译）的方法 ..... - 10 -

3.1.2 Seq2Seq 模型 ..... - 11 -

3.1.3 Seq2Edit 模型 ..... - 12 -

3.2 数据集 ..... - 14 -

3.2.1 HSK 数据集 ..... - 14 -
3.2.2 Lang8 数据集 ..... - 14 -
3.2.3 NaSGEC-thesis 数据集 ..... - 15 -
3.2.4 CCF 各类中文科技期刊数据 ..... - 15 -
3.3 数据增强 ..... - 15 -
3.4 模型优化方法 ..... - 17 -
3.4.1 多轮迭代纠正 ..... - 17 -
3.4.2 模型集成 ..... - 17 -
3.5 模型训练 ..... - 18 -
3.5.1 模型训练流程 ..... - 18 -
3.5.2 结果分析 ..... - 19 -
3.6 本章小结 ..... - 19 -
第四章 系统分析与设计 ..... - 20 -
4.1 系统需求分析 ..... - 20 -
4.1.1 功能性需求 ..... - 20 -
4.1.2 非功能性需求 ..... - 21 -
4.2 系统用例分析 ..... - 21 -
4.2.1 用例图 ..... - 21 -
4.2.2 核心事件流 ..... - 22 -
4.3 系统概要设计 ..... - 25 -
4.3.1 系统总体设计 ..... - 25 -

4.3.2 核心功能模块设计 ..... - 25 -
4.3.3 数据库设计 ..... - 27 -
4.4 系统可行性分析 ..... - 29 -
4.5 本章小结 ..... - 30 -
第五章 系统模块详细设计与实现 ..... - 31 -
5.1 登录注册模块 ..... - 31 -
5.1.1 登录 ..... - 31 -
5.1.2 注册和修改密码 ..... - 32 -
5.2 用户功能模块 ..... - 34 -
5.2.1 管理用户信息 ..... - 34 -
5.2.2 论文上传 ..... - 35 -
5.2.3 纠错测试 ..... - 39 -
5.2.4 查看纠错记录 ..... - 39 -
5.2.5 管理纠错记录 ..... - 42 -
5.3 本章小结 ..... - 43 -
第六章 系统测试 ..... **- 44 -**
第七章 总结与展望 ..... **- 46 -**
7.1 本文总结 ..... - 46 -
7.2 后续工作展望 ..... - 46 -
参考文献 ..... - 48 -

前 言

在文本阅读出现在方方面面的今天，文本错误的问题也随之而来，令人防不胜防。科技文献具有很高的专业性与严谨性，任何一处小的文本错误都会影响论文的质量，甚至造成歧义。但由于科技文献的复杂性，其涵盖大量的专业术语和复杂的语言结构，反而更容易产生拼写错误和语法错误等文本错误。

本文探讨了面向科技文献领域的文本纠错系统的实现，旨在介绍如何应用自然语言处理技术，开发一个能够自动检测和修正科技文献中的拼写错误、语法错误和标点符号错误等问题的文本纠错系统。同时，本论文还涵盖了后端 Flask 和前端 Vue 技术的应用和整合，以及系统的具体实现和测试结果。

本文采用了目前主流的神经机器翻译（NMT）以及 Seq2Edit 模型，利用爬虫技术获取大量科技文献数据，并使用数据增强技术处理无监督数据，对纠错模型做了一系列优化，使模型在科技文献领域的纠错效果更好。

本论文的研究意义在于探讨如何通过文本纠错系统提高科技文献的严谨性和可读性，减少人工校对文献的时间和成本，提高研究人员的工作效率。同时，本论文所涉及的 Flask 和 Vue 技术也具有广泛的应用价值。

第一章 绪论

1.1 课题背景

近年来，随着社会飞速发展，经济与教育水平不断提高，我们接触互联网的机会越来越多。在当今信息时代，中文文本的使用和传播已经成为了人们日常生活和工作中不可或缺的一部分，每天阅读的文本量大大增加。然而，我们经常会在各个地方遇到错误文本，比如错别字、标点使用不当或是有语病等等。可以说，只要有文本写作，就会出现文本错误，进而出现了文本纠错的需求。很多重要的文本在发布前都会经过人工校对审核，尽管如此，我们依然能够在新闻甚至政府文书中见到错别字。

中文文本纠错技术的应用领域十分广泛，可以在中文写作、中文翻译、社交媒体、教育领域以及商业领域等发挥重大作用。由于中文的复杂性和语法规则的多样性，中文文本常常存在着拼写错误、语法错误、标点符号错误、词语搭配错误等问题。这些错误不仅会影响读者的理解和阅读体验，还可能导致误解、误导和误判。中文文本纠错的重要性在于它可以帮助消除这些错误，使文本更加准确和易于理解。通过纠错，可以提高文本的质量和可信度，帮助读者更好地理解和利用文本信息。

1.2 系统开发目的和意义

相比起英文广泛且成熟的文本纠错软件，如 Grammarly、Ginger、ProWritingAid 等，中文在这方面的软件并不多。并且由于中英文在语法规则、词汇量以及环境语境等方面具有明显区别，中文文本纠错（CGEC）的研究人员并不多，相关软件不太成熟。

我们希望能设计并实现一个面向中文科技文献领域的文本纠错系统，专注于科技文献论文的文本纠错。文献论文应具有较高的规范性与严谨度，任何小的文本错误都会影响论文的质量与可信度。往小了说错别字、语法错误等文本错误会降低论文的准确性和可读性，往大了说甚至会引起歧义，造成读者的错误理解。为了降低作者在勘误方面的工作量、提高工作效率、改善用户体验以及促进中文信息技术的

发展，本文设计并实现了面向科技文献领域的文本纠错系统，该系统能够快速、自动地检查和修正中文文本中的错误，避免繁琐耗时的人工校对和编辑。

1.3 课题内容

本课题旨在满足用户需求，基于 Web 技术开发一个面向科技文献领域的文本纠错系统，主要工作内容包括：

1）学习 Web 相关知识，学习并熟练掌握 Flask 后端框架，学会并使用 Vue 前端框架，熟练掌握 SQLite 数据库的使用方法，为设计并实现面向科技文献领域的文本纠错系统做好技术方面的准备。

2）分析面向科技文献领域的文本纠错系统的需求，并在此基础上，深入了解当今在线纠错所面临的问题，最终分析面向科技文献领域的文本纠错系统的功能性需求和非功能性需求，结合互联网模式，设计实现一款合理的文本纠错系统。

3）学习爬虫技术，获取 CCF 的 A、B、C 类期刊的大量科技文献数据，为后续纠错模型的训练提供数据支撑。

4）学习并学会使用当今基于神经机器翻译（NMT）的文本纠错技术，熟练掌握 Seq2Seq 模型以及 Seq2Edit 模型。学习并尝试利用数据增强技术以优化模型并在科技文献领域取得更高的纠错评价指标。

5）详细设计与实现系统，将系统分为登录注册模块和用户模块，使用 Flask 和 Vue 框架，对各功能模块进行实现。

6）设计测试用例，对系统进行各种测试，分析测试结果，验证功能是否正常。

7）验收系统，撰写论文。

1.4 论文组织结构

首先，本文详细介绍了本系统的背景意义和开发目的。接着介绍了本系统所采用的相关技术，包括 Web 方面和纠错模型的技术。然后，本文通过对面向科技文献领域的文本纠错系统的业务流程进行分析，对系统进行了需求分析和概要设计，并详细介绍了系统的实现过程。最后，本文对系统进行测试，并进行总结。全文共分

七个章节：

第一章绪论，介绍面向科技文献领域的文本纠错系统的研发背景和意义，阐述论文的组织结构。

第二章系统相关技术介绍，这一章介绍了系统开发过程中所使用的前后端架构，并介绍其余相关技术，主要包括了后端 Flask 框架、前端 Vue 框架以及数据库 SQLite。

第三章纠错相关技术介绍，这一章介绍了文本纠错方面所采用的技术，包括 Seq2Seq 模型与 Seq2Edit 模型、数据增强方法以及模型集成技术。同时，还简单介绍了科技文献数据的获取以及处理流程。

第四章系统分析与设计，包括系统的功能性和非功能性需求进行分析、系统用例设计、系统总体设计与核心模块设计以及数据库结构。

第五章系统详细设计和实现，本章节将介绍系统的主要功能模块，并描述采用的实现思路。本章节着重介绍纠错系统开发过程中的重点和难点，以便读者了解系统开发的技术细节和实现过程。

第六章系统测试，本文将设计测试用例，验证系统是否能正常工作。测试完成后，若无法达到预期效果，将进一步完善纠错系统，提高系统的健壮性。

第七章总结与展望，对设计开发进行总结，并总结本文所进行的工作和实现的成果。同时，本文将对课题内容进行展望，思考系统的不足之处以及未来的改进方向，以便未来对纠错系统进行进一步的优化。

第二章 系统相关技术介绍

本章介绍纠错系统开发过程中所使用到的技术，包括 Flask 后端框架、Vue 前端框架以及 SQLite 数据库等。

2.1 系统架构

本系统的设计和开发主要对应 MVC 架构模式。MVC 是一种被广泛应用于软件开发的架构模式，它将应用程序分为三个不同的部分：模型（Model）、视图（View）和控制器（Controller）。MVC 框架是一种将应用程序的逻辑层和视图层相分离设计, 并通过控制层连接的开发模式[1]。MVC 的核心思想是将应用程序的不同方面分离，以便更好地组织和维护代码。[2]中详细介绍了 MVC 的架构，如图 2-1 所示。

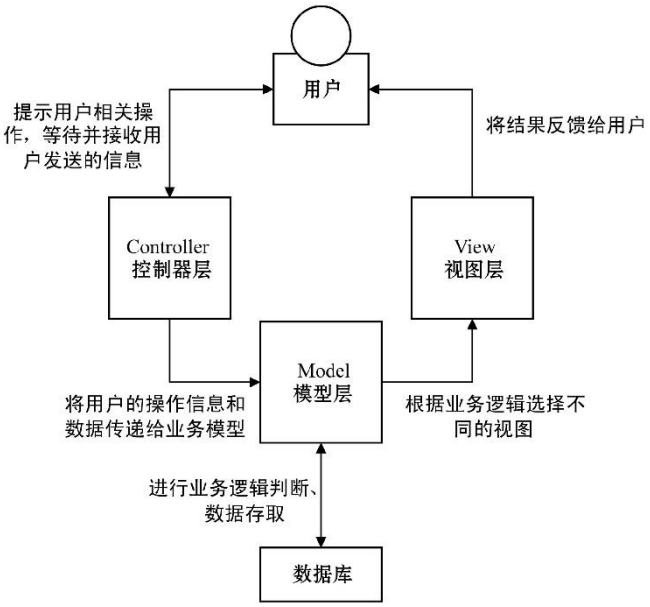


图 2-1 MVC 架构，图片来自文献[2]

（1）模型（Model）

模型是应用程序中负责处理数据和逻辑的部分，它通常包括数据存储、数据处理和业务逻辑等功能。模型通常不知道应用程序的其它部分，它们只是提供数据和行为。模型是 MVC 架构中最核心的部分。

（2）视图（View）

视图是用户界面的部分，负责向用户呈现数据和接收用户输入。视图通常是通过 HTML、CSS 和 JavaScript 等 Web 技术实现的，或者是其他用户界面技术（如 Swing、WinForms 等）。视图的主要任务是呈现数据，并且一般与模型完全分离，它们不包含任何业务逻辑或数据处理代码。

（3）控制器(Controller）

控制器是 MVC 架构的第三个组成部分，它负责协调模型和视图之间的通信。控制器通常响应用户输入，并将用户请求传递给模型来获取或更新数据。一旦模型完成数据的处理，控制器将结果传递给视图进行呈现。控制器通常包含了应用程序的核心业务逻辑和处理代码。

MVC 架构的优点是明确的分离关注点和良好的可扩展性。将应用程序分成三个独立的部分，可以更容易地管理和维护代码，并且能够在不影响其他部分的情况下进行修改和改进。此外，MVC 架构使应用程序的测试更容易，因为模型和视图是完全分离的，可以独立地进行单元测试。

在 MVC 架构中，模型、视图和控制器之间的通信通常是通过一组接口来实现的。这些接口定义了不同部分之间的交互方式，从而使各部分之间的耦合度降到最低。通过使用 MVC 架构，可以创建具有良好可维护性和可扩展性的高质量应用程序。

2.2 WEB 开发工具

2.2.1 Flask 框架

Flask 是一个轻量级的 Python Web 框架，它提供了一种简单、易用的方式来构建 Web 应用程序。相比于其他 Web 框架，Flask 的项目结构简单，核心服务比较轻量，配置比较简单，第三方组件可以自由拆卸，比较适用于小型的项目开发使用[3]。Flask 使用 Python 中的装饰器来定义路由，根据请求的 URL 映射到对应的视图函数进行处理，同时支持多种模板引擎，如 Jinja2、Mako 等，可以将数据渲染到 HTML 页面中，生成最终的响应内容。

Flask 还提供了多种数据库支持，如 SQLAlchemy、MongoDB 等，可以方便地与数据库进行交互，同时还支持中间件和扩展库，可以实现各种额外的功能，如日志、

缓存、邮件发送等。Flask 的优点在于易于学习和上手，同时支持多种插件和扩展库，可以帮助开发者快速实现各种功能。缺点在于对于大型项目或者高并发场景的处理能力相对较弱，需要使用其他的解决方案来处理。总的来说，Flask 是一个非常灵活、简洁的后端框架，可以根据项目需求进行选择和扩展。

2.2.2 Vue 前端框架

Vue.js 框架是一种面向前端界面开发的轻量化框架，其设计模式遵循自下而上的原则，此种开发框架的最本质的特点是具有响应式编程和组态化的特点[4]。它是一个轻量级的框架，主要关注视图层的渲染和数据绑定。Vue.js 的核心是一个响应式的数据绑定系统，可以让开发者通过简单的模板语法将数据和 UI 进行绑定。在数据发生变化时，Vue.js 能够自动更新 UI，从而实现快速的开发和高效的性能。

Vue.js 提供了诸多功能和特性，包括模板语法、组件系统、指令、生命周期钩子函数、计算属性、侦听属性等。其中，组件系统是 Vue.js 的重要特性之一，它使开发者能够将 UI 分解为独立的组件，并进行复用和组合。Vue.js 的组件可以包含模板、样式和行为等多种元素，可以实现灵活、可维护的 UI 构建。

Vue.js 的生态系统非常丰富，包括官方维护的插件和社区维护的插件。其中，Vue Router 和 Vuex 是 Vue.js 的两个官方插件。Vue Router 是一个用于实现前端路由的库，可以方便地管理 URL 和组件之间的映射关系。Vuex 是一个用于管理应用程序状态的库，可以将所有组件的共享状态抽离出来，从而方便地进行统一管理和操作。

Vue.js 是目前国内应用最为普遍的前端开发框架，采用此种开发模式可以大幅度地提升开发效率以及代码的服用效率[5]。Vue.js 也有一些缺点，例如在大型项目中的组件管理和状态管理等方面，需要更加细致的考虑和规划。总的来说，Vue.js 是一个非常流行、易于学习和使用的前端框架，非常适用于本系统的开发。

2.2.3 Element UI 组件库

Element UI 是一套基于 Vue.js 的桌面端组件库，旨在帮助开发人员快速搭建高质量的用户界面。Element UI 提供了丰富的可定制的组件，包括按钮、表格、表单、

弹窗等常见的 UI 元素，以及日历、时间轴、评分、导航等更复杂的组件，可满足各种应用场景和设计需求。

Element UI 的优点之一是其易用性和灵活性。它使用基于 HTML 和 CSS 的语法和设计风格，使其易于学习和使用。开发人员可以通过简单的配置和组合来定制组件样式、主题和行为，以实现所需的外观和功能。

Element UI 还具有优秀的性能和响应性。它采用了现代化的技术，如虚拟滚动、异步加载等，可以有效地减少页面加载时间和内存占用，提高用户体验。

总之，Element UI 是一款功能丰富、易于使用和高性能的 UI 组件库，适用于本系统在内的各种规模的项目和应用。

2.3 数据库技术

SQLite 是一款轻量级的嵌入式关系型数据库管理系统，它是在公共领域下开放源代码的软件，适用于各种平台，包括移动设备和嵌入式系统。SQLite 的设计目标是提供一种零配置、自给自足的数据库解决方案，它不需要专门的服务器进程或配置，也不需要复杂的安装过程。

SQLite 数据库以单个文件形式存储数据，该文件包含完整的数据库结构和数据，因此易于备份和迁移。SQLite 支持多种数据类型，包括整数、浮点数、文本、二进制和 NULL 值，并支持 SQL 标准的多种语法和功能，例如事务、索引、触发器和视图等。它还支持多线程访问和高并发性能，可以处理大量数据并保持数据的完整性和一致性。

SQLite 具有极低的内存占用和高效的性能，可以在资源受限的环境中运行，并且可以与各种编程语言和框架集成。由于 SQLite 数据库的易用性和可移植性，它广泛应用于各种应用程序中，例如移动应用、桌面应用、Web 应用、游戏、数据采集和存储等。总之，SQLite 是一款功能强大、易于使用、高效可靠的数据库管理系统，适用于各种应用场景。

2.4 本章小结

本章介绍了纠错系统的系统架构，本系统使用了轻量级的 Python Web 框架 Flask 和面向前端界面开发的轻量化框架 Vue，利用 Element UI 组件库美化前端页面，并使用具有极低的内存占用和高效性能的 SQLite 数据库存储信息。

第三章 纠错相关技术介绍

本章主要介绍了使用到的纠错模型比如 Seq2Seq 模型和 Seq2Edit 模型，以及使用的数据集、数据增强方法和模型优化方法，最后分析了模型训练结果。

3.1 纠错模型

3.1.1 基于 NMT（神经机器翻译）的方法

神经机器翻译（NMT）是一种利用神经网络模型进行机器翻译的方法。与传统的统计机器翻译不同，神经机器翻译旨在建立一个单个的神经网络，可以共同调整以最大化翻译性能(表现)[6]。它将输入的源语言句子编码为一个向量，然后将其解码为目标语言句子。该过程使用神经网络的编码器-解码器结构来完成,常见的编码器-解码器结构如[7]中图 3-1 所示。

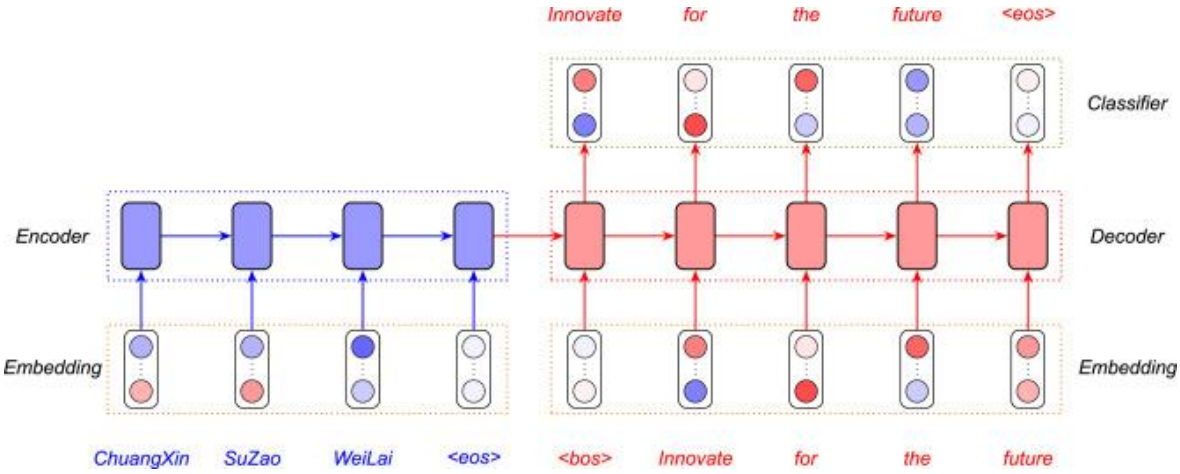


图 3-1 编码器-解码器结构，图片来自文献[7]

在神经机器翻译中，编码器将源语言句子转换成一个连续的向量表示，而解码器将向量表示解码成目标语言句子。这个向量表示被称为“上下文向量”，它包含了输入句子的所有信息，因此可以用来生成输出句子。

在神经机器翻译的训练过程中，需要大量的平行语料库来训练模型。一旦训练好了模型，就可以使用它来翻译新的句子。与传统的统计机器翻译相比，神经机器翻译在翻译质量上更加优秀，因为它可以处理更加复杂的句子结构，而且在处理长

句子时更加有效。

虽然神经机器翻译在翻译质量上有所提高，但仍然存在一些挑战。例如，它仍然难以处理某些语言之间的巨大差异，如英语和汉语之间的语法和词汇差异。此外，神经机器翻译还需要大量的训练数据和计算资源，以获得最佳的性能。

总的来说，神经机器翻译是一种非常强大的翻译方法，已经成为机器翻译领域的主流。由于文本纠错可以看作从错误文本到正确文本的翻译过程，因此同样可以采用神经机器翻译技术来处理。下面介绍目前常用的，也是本系统中使用到的两种模型——Seq2Seq（序列到序列）模型和 SeqEdit（序列到编辑）模型[8]。

3.1.2 Seq2Seq 模型

Seq2Seq（Sequence-to-Sequence）[9]模型是一种基于神经网络的序列生成模型。它由编码器和解码器两个部分组成，可以用于机器翻译、语音识别、对话系统、文本摘要等任务。Seq2Seq 模型接受一个变长的输入序列，经过编码器将其转化为一个固定长度的向量，再通过解码器逐步生成一个与目标语言相对应的输出序列。Seq2Seq 模型的关键是使用了 Attention[10]机制，使得模型能够学习到输入序列中的关键信息，并将其用于生成输出序列，如其论文中图 3-2 所示。

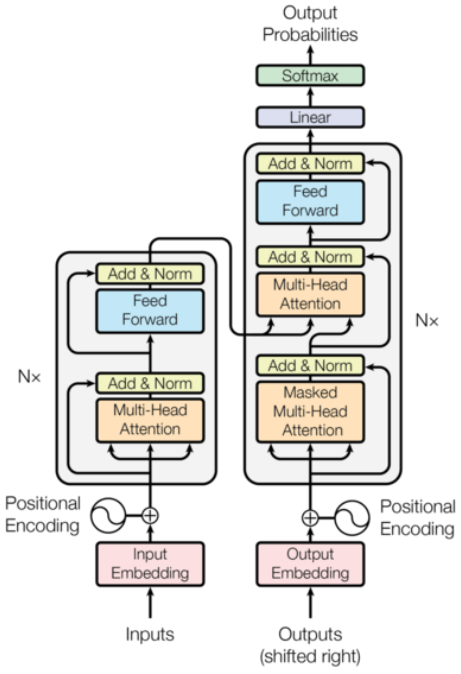


图 3-2 Transformer 架构，图片来自文献[10]

Seq2Seq模型的训练过程通常采用最大似然估计法，通过最小化目标序列和模型生成序列之间的差距来优化模型。Seq2Seq模型在各种序列生成任务中都取得了很好的效果，成为自然语言处理领域的重要技术之一。

在本系统使用的Seq2Seq模型中，我们使用了基于Transformer架构的编码器-解码器结构，专用于文本生成的BART(Bidirectional and Auto-Regressive Transformers)[11]模型。我们采用了面向中文的bart-large-chinese预训练模型，在训练集上进行训练。

3.1.3 Seq2Edit模型

Seq2Seq结构的NMT-based GEC系统仍存在一些问题，比如：1）推理速度较慢；2）对训练数据的需求量大；3）缺乏可解释性，比如没法解释语法错误类型分类[12]。而Seq2Edit[13]可以将序列生成任务简化为序列标注任务来解决这些问题。

Seq2Edit 可以被认为是一种神经机器翻译方法的变体，但它更加专注于对文本的编辑和修正，而不仅仅是传统翻译的转换任务。Seq2Edit 能够将源文本转换为目标文本，并且能够进行多种编辑和修正操作。Seq2Edit 使用编码器-解码器结构，编码器将源文本编码为一个向量表示，解码器则将该向量解码为目标文本，并通过注意力机制对目标文本进行编辑和修改。

Seq2Seq 是把长度为I的源序列x在字符词表V上生成长度为J的目标序列y，概率$P(y|x)$的计算方式如公式(3.1)所示。

$$P(y|x) = \prod_{j=1}^{J} P\left(y_j|y_1^{j-1}, x\right) \tag{3.1}$$

Seq2Edit 不是直接预测目标序列y，而是预测编辑操作的序列，其中每个编辑序列$(t_n, p_n, r_n) \in T \times N_0 \times V$是一个三元组，$r_n$是与可解释标签相关联的字符，用于替换源序列中下标为p_{n-1}到p_n的动作t_n。三元组序列y的计算方式如公式(3.2)所示。

$$P(y|x) = P(t_1^N, p_1^N, r_1^N|x) = \prod_{n=1}^{N} P(t_n, p_n, r_n|t_1^{n-1}, p_1^{n-1}, r_1^{n-1}, x) \tag{3.2}$$

Seq2Edit模型的整体架构如图3-3所示。

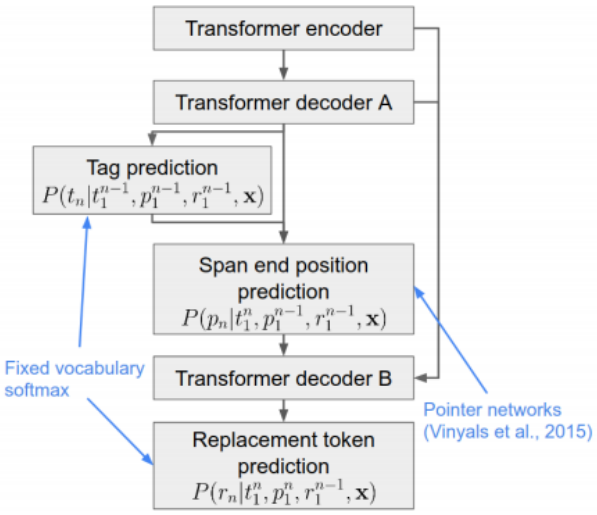


图 3-3 Seq2Edit 模型架构，图片来自文献[13]

与传统的神经机器翻译不同，Seq2Edit 需要对文本的不同操作进行精细预测，例如插入、删除、替换等操作。为此，Seq2Edit 引入了编辑操作序列，使得模型能够在翻译时实时地进行编辑和修正。

我们系统中的 Seq2Edit 模型来自 MuCGEC 开源模型，其通过修改 GECToR(Grammatical Error Correction: Tag, Not Rewrite)[12]模型，以应对中文文本纠错。GECToR 在字符级别上进行文本纠错的处理是通过将输入句子中的字符分解成子字符，将每个子字符作为输入，并将其对应的错误标注为需要修正的错误类型。在 GECToR 的 Transformer 架构中，输入嵌入层用于将输入的字符表示为向量形式，然后送入 Transformer 编码器。编码器根据输入序列中每个字符的嵌入向量，自适应地学习相应的特征表示。最后，根据编码器的输出和解码器标注的标签信息，生成一个修正后的输出句子。

Seq2Edit 模型去除了 GECToR 中的 g-transformations，此转换标签主要面对英文文本，比如更改单复数、合并前后单词、更改字符大小写、拆分单词等等，并不适用于中文文本。模型保留了 GECToR 中的 basic-transformations，比如$KEEP 标签表示当前字符保持不变、$DELETE 标签表示当前字符需要删除、$APPEND_x1 表示当前字符后需要添加字符 x1、$REPLACE_x2 表示当前字符需要替换为 x2。

由于一些字符可能包括多个转换标签，模型规定每个源字符仅保留一个转换，如果有多个转换，将采用第一个不是$KEEP 标记的转换[12]。

3.2 数据集

在模型训练过程中，本文使用了各类数据集，包括 HSK 数据集、中文 Lang8 数据集、CCF 各类中文科技期刊数据以及悟道语料库数据，同时还使用了我们自己标注的科技文献领域 NaSGEC-thesis 数据集作为测试集以评估模型纠错性能，如表 3-1 所示。

表 3-1 数据集

数据集名称	句子	句子（修改）
HSK	156870	96049
Lang8	1220734	1097233
CCF（数据增强）	300000	187415
悟道（数据增强）	300000	182903
NaSGEC-thesis(test)	500	378

3.2.1 HSK 数据集

HSK 数据集是汉语水平考试（HSK）的词汇等级测试中使用的数据集。它包含了大量的汉语单词和短语，涵盖了 HSK1-6 级的词汇量。这个数据集可以被用来开发各种汉语学习应用，比如语言学习软件、词汇测试和语音识别等。同时，HSK 数据集还被广泛用于自然语言处理（NLP）领域的研究，比如文本纠错、文本分类、语言模型和机器翻译等。

3.2.2 Lang8 数据集

Lang8 数据集由来自全球各地的学习者提供的短语和句子组成，旨在帮助学习者改善他们的外语写作技巧。Lang8 数据集包含了大量的带有错误的文本，例如拼写错误、语法错误和用词不当等。这些错误的文本是为了给语言学习者提供一个学习平台，让他们可以通过撰写和修正其他学习者的文本来提高自己的外语写作能力。

3.2.3 NaSGEC-thesis 数据集

这是我们自己标注的面向科技文献领域的中文文本纠错数据集，包含 1500 句高质量纠错数据，主要用来作为测试集评测纠错模型在科技文献领域的纠错指标。

3.2.4 CCF 各类中文科技期刊数据

通过学习爬虫技术，获取了 CCF 推荐的 A、B、C 类中文科技期刊，总计约 8 万篇论文。通过解析 PDF 文档、提取文本、分句、筛选去除噪音数据后得到约 400 万句科技文献领域的句子。这些数据可通过数据增强技术人工生成错误文本，作为训练集训练。

3.3 数据增强

数据增强是指通过对已有数据进行一系列变换和扩充，生成新的训练数据集的技术。在机器学习和深度学习任务中，通常需要大量的训练数据才能取得良好的性能，而现实中的数据往往是有限的。

一种常见的方法是向正确文本中插入错误，人造伪标签数据。由于随机插入生成的错误与真实错误差异较大，训练效果只是差强人意。针对这个问题，有论文提出了先统计各错误的分布，再使用 ERRANT 自动注释工具里的错误类型来辅助生成合成数据，使人造伪数据更符合领域内的真实错误分布的方法[14]。此方法能大大减少人造数据和真实数据的差异，但由于没法完全了解领域内的真实错误分布，因此伪数据和真实数据仍然存在差异，不过效果比随机插入错误生成伪数据要好得多。

此外，自训练[15]是文本纠错领域常用的一种数据增强方法。首先使用真实的有标签数据训练一个模型，然后用该模型来纠正不符合语法的句子，并将纠正结果作为伪数据加入到训练数据中，最后迭代训练模型。虽然将模型预测结果直接用作训练数据会带来一定的噪声，但总体而言，这种方法仍然能够提高模型的性能。

还有一种反向翻译人造伪数据的方式[16]，源于机器翻译领域的研究。通常情况下，GEC 的 Seq2Seq 模型会将错误纠正任务视为机器翻译任务，并通过将有误句子

翻译为正确句子来实现纠错。反向翻译则是指训练一个将正确的句子翻译为错误的句子的模型（即造错模型），并使用该模型生成伪数据。这种方法的优点在于，制造错误模型的训练数据只需通过将原始训练句子对中的顺序进行交换即可轻松获得。

经过研究，本文最终使用第一种数据增强方法。通过观察真实数据集中的错误分布，本文针对性地制定了数据增强策略。首先将句子分为长度为 20 的片段，对每个片段以 0.5 概率随机加入噪声，并按 0.8、0.15、0.05 的概率加入 1、2、3 个噪声。噪声共分为四种类型：替换、添加、删除与词序，其中每个类型又分为字级别和词级别类型（删除分为字级别和标点级别）。

替换类型的噪声针对字词级别分为同音字替换、混淆集替换、同音词替换和近义词替换几个选项。添加类型的噪声分为添加近义词和重复当前 token，并对下标进行随机化处理。删除类型的噪声随机选择 1、2、3 个删除数量。词序类型的噪声则分为交换和先删除再插入两种类型。

以上各选择的概率分布均为观察真实数据集后得到，尽量模拟真实文本错误的分布情况，以让模型达到更好的拟真训练效果。具体的数据增强操作如表 3-2 所示。

表 3-2 数据增强方法

级别	操作	具体类型
词	替换	同音：随机从词级别同音混淆集中采样替换
		近义：随机从大词林的近义词中采样替换
	添加	重复：将当前词复制并插入到附近某个位置
		近义：随机从大词林的近义词中采样插入
		分布：随机从真实添加分布中采样替换
	词序	词序：随机将当前词语与附近某个词语交换位置
字	替换	音近：随机从字级别同音混淆集中采样替换
		形近：随机从字级别形近混淆集中采样替换
	删除	缺失：删除当前字
	添加	重复：将当前字复制并插入到附近某个位置
		分布：随机从真实添加分布中采样替换
	词序	词序：随机将当前字与附近某个字交换位置
标点	删除	缺失：删除当前标点

3.4 模型优化方法

总体而言，目前的研究大多是考虑给模型加入额外信息来辅助纠错，如[17]通过给模型加入上下文信息来帮助模型处理一些篇章级错误。[18]为模型加入句法信息（依存句法，成分句法）来辅助模型纠错。[19]提出了将 GED 模型的结果作为辅助输入参与微调的方法。

此外，还有一些不改动模型的方法，但也能有效提升模型性能，比如下面本文采用的两种方法。

3.4.1 多轮迭代纠正

对于句子中的连续错误，尤其是连续的缺失错误，他们之间通常会存在一定的依赖关系，需要知道前一个 token 才更有利于预测下一个 token，这时模型往往不能一次纠正出所有错误。因此采用多轮迭代纠正[15]，直至模型认为句子没有错误是一种提高模型性能的有效方式。

3.4.2 模型集成

实验表明，Seq2Seq 模型和 Seq2Edit 模型擅长处理的错误类型不同[8]，使用模型集成的方式可以让两种不同架构的模型有效互补，提高纠错性能。

目前的主流模型集成方式有两种，一种是对模型预测结果的概率进行求和平均，另一种则是通过简单的投票来进行集成。投票集成的方式相对而言泛用性更强，不需要考虑模型输出结构，概率求和的方式繁琐一些，但能在更细粒度上进行集成。

举个例子，使用六个模型对同一句话进行纠错，得到六句结果。对其中的每一处编辑进行投票，若某处编辑投票数大于 3 则保留。这样可以增加文本纠错的准确率，同时可以自定义投票阈值以提供不同的纠错效果。

3.5 模型训练

3.5.1 模型训练流程

直接用二语数据集训练出的 Seq2Edit 模型在科技文献领域的纠错效果一般，因此需要对纠错模型做领域移植。从原来的纠错结果上看，模型对一些科技文献领域的专有名词处理得不好，会有漏纠或过纠的情况，因此我们想到利用数据增强技术，使模型学习大量 CCF 科技文献数据，在领域内实现更好的纠错效果。

受资源限制，没法使用大规模文本进行训练，主要采用小规模的数据集。模型采用 Seq2Edit 模型，测试集选用 NaSGEC-thesis，即我们标注的高质量科技论文文本纠错数据集，用于检测模型在科技文献领域的纠错效果。评测时使用 CHERRANT (Chinese ERRANT)工具，此工具对目前英文上通用的可细分类别评估的评测工具 ERRANT 进行了中文适配和修改，支持字、词粒度的评估[8]。

模型训练的过程中使用到的数据集有 HSK+Lang8 数据集、数据增强的 CCF 数据集、数据增强的悟道语料库数据集。模型使用了融入语言结构信息的 BERT[20]预训练语言模型 StructBERT-Large，经一阶段或分阶段训练，训练 epoch 均为 20，训练结果如表 3-3 所示。

表 3-3 训练结果

一阶段训练	二阶段训练	测试集(500)	P	R	$F_{0.5}$
CCF(30w)	×	NaSGEC-thesis	22.4	17.6	21.2
HSK+Lang8(15w)	×	NaSGEC-thesis	8.2	19.9	9.3
CCF(30w)	HSK+Lang8(15w)	NaSGEC-thesis	17.4	21.6	18.1
HSK+Lang8(15w)	CCF(30w)	NaSGEC-thesis	18.7	20.5	19.0
CCF(30w)+HSK+Lang8(15w)	×	NaSGEC-thesis	**23.4**	**21.8**	**23.1**
WuDao(30w)+HSK+Lang8(15w)	×	NaSGEC-thesis	18.9	19.2	18.9
HSK+Lang8(156w)	×	NaSGEC-thesis	**30.5**	12.9	**24.0**
HSK+Lang8(156w)	CCF(30w)	NaSGEC-thesis	18.1	21.3	18.7

3.5.2 结果分析

由于 HSK+Lang8 数据集是二语学习者领域，而测试集是母语科技文献领域数据，两者之间存在较大的领域差异，观察表 3-3 的实验结果可以发现直接使用二语数据集训练出的模型在科技文献领域的纠错效果并不好。

只使用 HSK+Lang8 数据集作为训练集，在数据规模较小时（15 万）$F_{0.5}$ 只有 9.3；数据规模扩大后（156 万）会有明显提升，具有最高的准确率和 $F_{0.5}$ 值，但是召回率较低，只有不到 13%。

数据增强使用的数据来自 CCF 中文期刊，同样是母语科技文献领域数据，只使用数据增强后的 CCF 伪数据集（30 万）进行训练，可以看到性能和使用真实数据训练差距不大。与大规模 HSK+Lang8 数据集相比，准确率低了 8 个点，召回率高了 5 个点。

值得注意的是将 CCF 伪数据集（30 万）和 HSK+Lang8 数据集（15 万）以不同顺序分阶段训练后，$F_{0.5}$ 只有 18~19，比单 CCF 数据集要低不少。

为了探索原因，本文采用了 156 万的大规模 HSK+Lang8 数据集并在 CCF 数据集上进行二阶段训练，数据量扩大后的模型性能没有改善，甚至低于 15 万的 HSK+Lang8 数据集。经分析后我们猜测原因是两个数据集的领域差异过大，且文本错误的差异也过大，在分阶段训练中表现不好，产生了大量的噪音信息。

之后本文尝试将两个数据集拼接起来一起训练，获得了最高的 21.8%召回率与不低的 23.4%准确率，$F_{0.5}$ 与 156 万 HSK+Lang8 数据集很接近。考虑到两者数据量间的巨大差异，我们认为将 CCF 数据集与 HSK+Lang8 数据集相结合是能在科技文献领域实现更好的纠错效果的。

3.6 本章小结

本章介绍了目前主流的神经机器翻译模型 Seq2Seq 以及能够能够提供错误类型的 Seq2Edit 模型，包括对 GECToR 模型的修改和利用，接着介绍了模型训练过程中所使用的数据集，随后阐述了数据增强方法与模型优化方法，最后介绍了模型训练流程以及分析了结果。

第四章 系统分析与设计

本章对纠错系统进行了需求分析、用例分析和概要设计，最后对系统进行可行性分析。

4.1 系统需求分析

对于软件开发而言，需求分析是至关重要的环节，也是首个环节，需求分析质量会对软件开发产生较大的影响，一般来讲，应结合需求分析结果来开展概要设计，根据相关的软件功能，构建程序流程图[22]。

4.1.1 功能性需求

软件需求通常分为功能性需求和非功能性需求（如可靠性、可支持性等），在这些需求中，功能性需求是需求定义的重点[21]。通过对文本纠错技术发展的了解，以及对具有相关纠错功能的软件与系统的调研，本文对面向科技文献领域的文本纠错系统所需实现的功能做了系统分析：

登录注册：用户需要在本网站注册账号，成功注册后通过登录访问网站，获取提交论文和查看记录的权限。

提交论文：用户可以通过不同的纠错模式提交论文，获取不同的纠错结果。

个人信息填写：用户可以完善个人基本信息。

用户隐私：用户有保护自己隐私的权利，提交的论文记录可以从数据库中删除，同时可以选择删除保存的论文原文件。用户可以选择我们团队是否能够使用他们的论文纠错数据用于后续科研。

纠错测试：提交论文获得纠错结果的时间略长，用户可以选择在测试页面提交短句快速查看纠错结果。

纠错反馈：用户可以针对每句话的纠错结果提供不同的反馈，以让我们的纠错模型不断改进，获得更好的用户体验。

4.1.2 非功能性需求

非功能性需求则是评估软件质量的指标，经分析，面向科技文献领域的文本纠错系统非功能性需求如下：

（1）易用性需求

系统的界面和交互设计方面需要符合用户需求，易于用户使用和理解。纠错系统需要设计得简洁高效，使各功能区一目了然，降低用户的学习成本。同时，用户仅需在浏览器输入域名即可浏览网页，不需要下载软件。

（2）隐私和安全性需求

该需求主要涉及到系统在处理用户数据、保护用户隐私方面的要求和保证系统的安全性。用户登录注册存储的密码均采用加密操作，降低密码泄露风险。前端使用 https 加密传输，用户的每次请求都会通过 session 进行身份验证，避免有人恶意绕过登录访问他人信息。系统确保用户能够访问和控制其个人数据，并提供相关的隐私设置。同时，系统提供删除用户个人数据的功能，并确保删除后数据不会被恢复。

（3）可维护性需求

系统开发过程中应注重注释以及保留版本，使得系统更加容易地进行维护和修改，从而减少维护成本和修改风险，提高系统的可靠性和可维护性。

（4）可扩展性需求

系统能很方便地添加其它功能而不影响现有功能的使用，系统可以采用分布式架构，将不同的功能区分布在不同服务器上，以便后续测试与修改。

4.2 系统用例分析

根据功能性需求分析可知，系统只有论文提交者一类用户。按照此类用户的需求，这里给出了用例图，并分析了核心事件流。

4.2.1 用例图

根据需求分析，得到用户用例分析，绘制出用例图，如图 4-1：

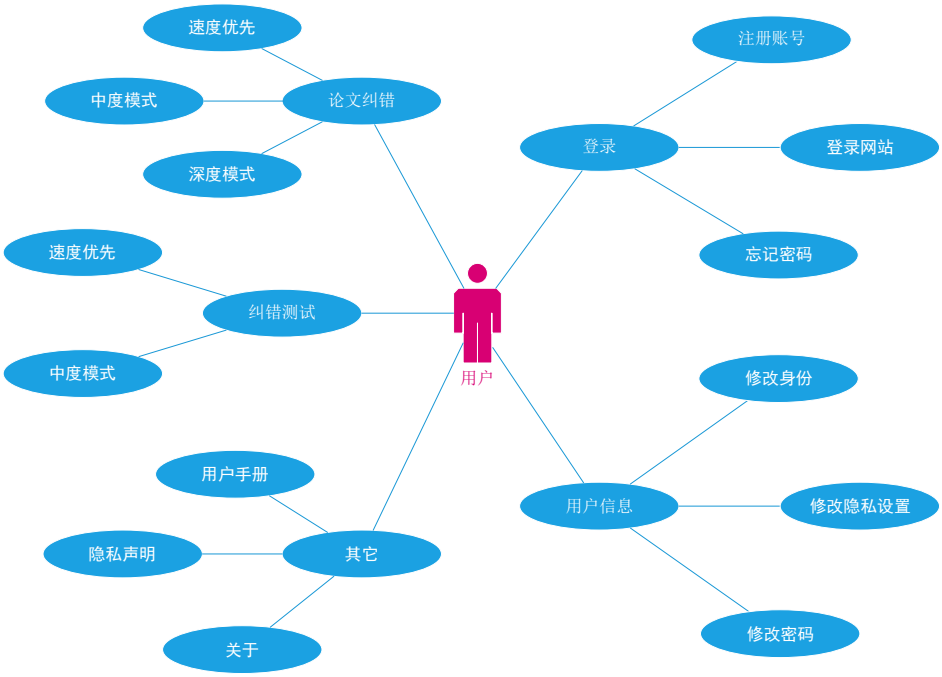


图 4-1 用户用例图

4.2.2 核心事件流

根据需求分析以及用例分析，在本系统中，用户的核心事件流包括登录注册事件流、用户信息事件流，论文提交事件流，纠错测试事件流、纠错记录事件流等。

（1）登录事件

用户输入账号密码登录网页，返回相应记录，如图 4-2 所示。

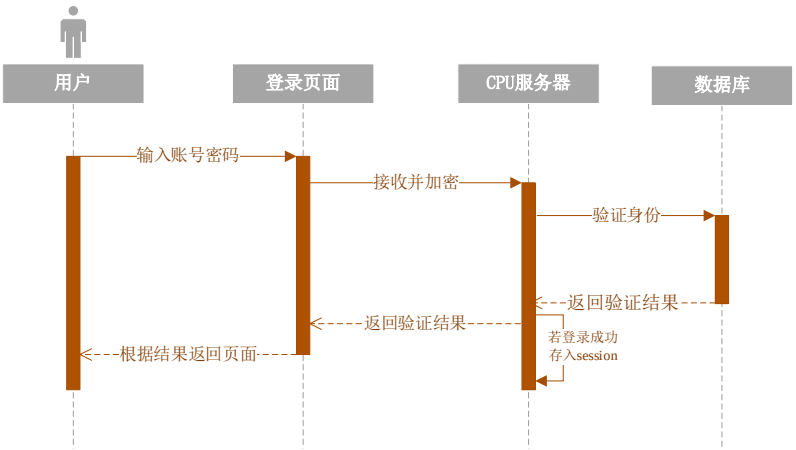


图 4-2 登录时序图

（2）注册事件

用户输入邮箱，接收验证码，满足注册要求的邮箱格式和密码格式后成功注册，如图 4-3 所示。

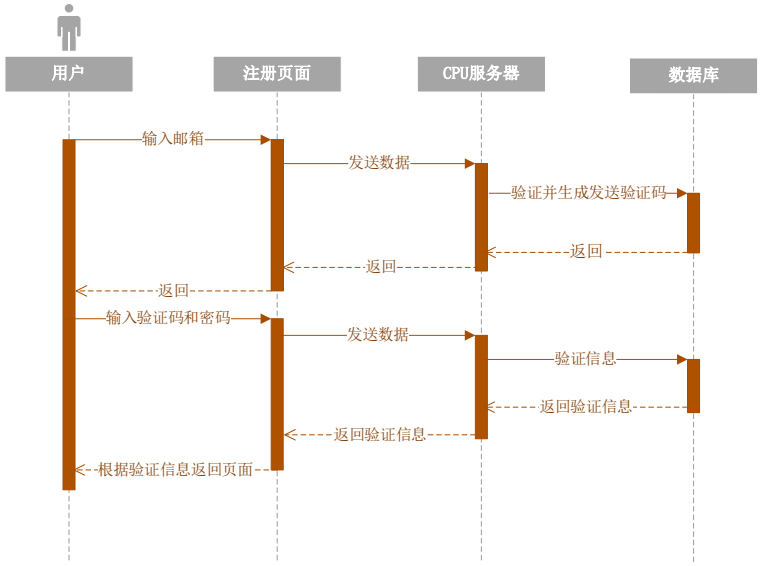


图 4-3 注册时序图

（3）用户信息事件

用户进入用户信息页面，可在此修改个人信息与隐私设置，也可以修改密码，如图 4-4 所示。

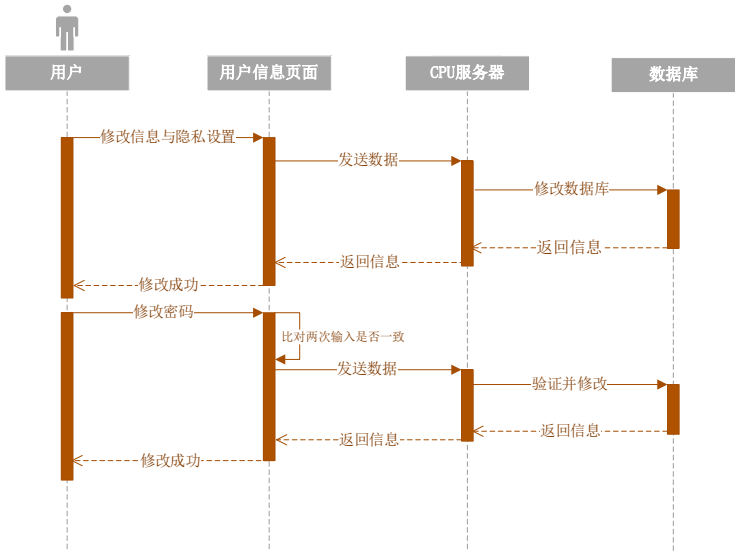


图 4-4 用户信息时序图

（4）论文提交事件

在论文上传页面，用户可以按不同模式提交想要纠错的论文，CPU 服务器将论文处理解析后发送给 GPU 服务器进行纠错，最后把结果返回给前端，如图 4-5 所示。

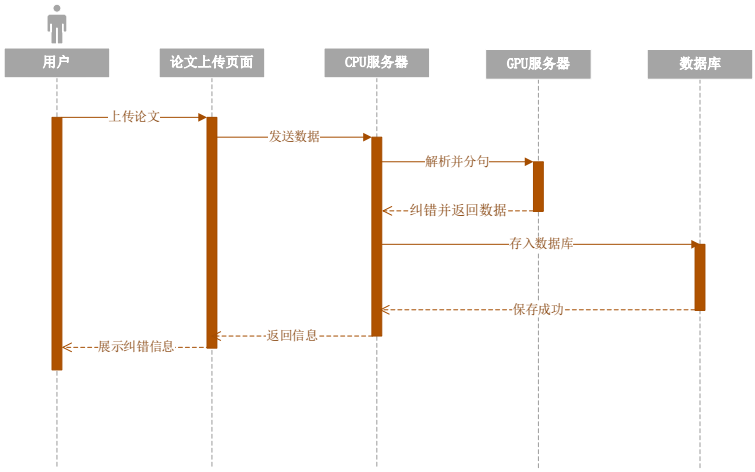


图 4-5 论文提交时序图

（5）纠错测试事件

用户可以在纠错测试页面测试短句纠错结果，而不需要将文件上传，以此快速预览纠错效果，如图 4-6 所示。

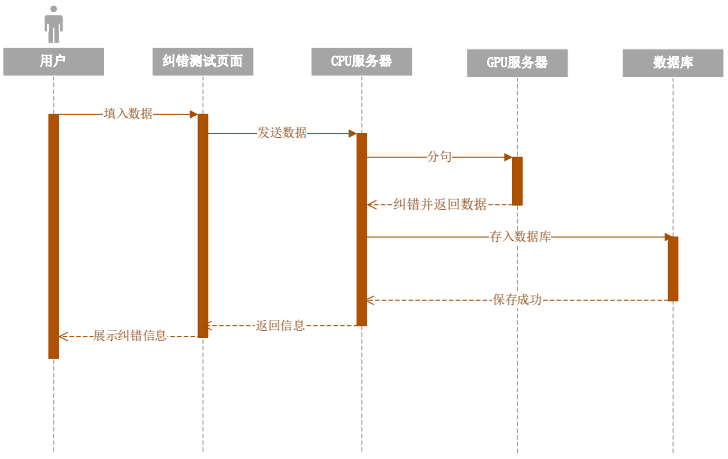


图 4-6 纠错测试时序图

（6）纠错记录事件

在查看纠错记录页面，用户可以选择进行多种操作，比如查看纠错结果、删除记录、下载原文、设置文档隐私等，如图 4-7 所示。

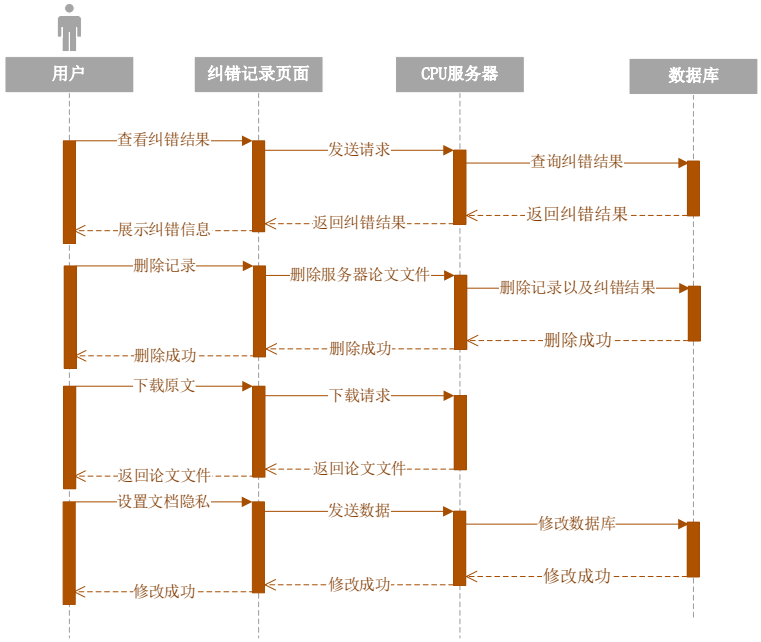


图 4-7 纠错记录时序图

4.3 系统概要设计

4.3.1 系统总体设计

本系统基于 Flask 开发，前端采用 Vue 框架。之所以采用 Flask 框架而非 SpringBoot 框架，是为了方便接入同样由 python 编写的纠错神经网络模型。Flask 是一款轻量级的 Web 应用程序框架，尽管其支持 MVC (Model-View-Controller) 架构模式，但本系统在代码开发过程中并没有刻意分层。

4.3.2 核心功能模块设计

通过对面向科技文献领域的文本纠错系统进行需求分析与用例分析，最终将系统分为登录注册模块和用户模块。图 4-8 给出了系统核心功能模块图。

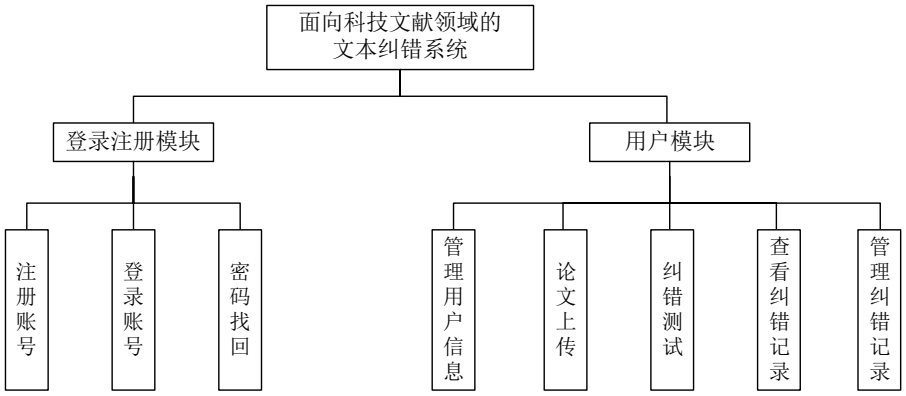


图 4-8 系统核心功能模块图

（1）登录注册模块

注册：为新用户提供注册功能，仅支持邮箱注册，同时暂时限制只能以 edu.cn 结尾的学校邮箱注册。系统会自动生成并发送验证码给对应的邮箱，需要填写正确的验证码才能完成注册。

登录：用户输入账号密码进行登录，后端会自动加密与数据库验证，登录成功后可查看其余页面。

密码找回：重新给用户邮箱发送验证码，填写新密码，验证成功后修改密码。

（2）用户模块

管理用户信息：用户能够修改自己的身份信息与隐私设置，若不想我们团队使用论文和纠错数据，可以设置为不公开。同时用户也能修改密码，需要验证码以及重复输入两次新密码。

论文上传：用户可以上传待纠错的论文，仅支持 pdf、docx 和 doc 后缀的文件。根据不同的需求，用户可以选择不同的纠错模式，其中速度优先模式具有很快的速度和较高的召回率，中度纠错模式在速度、召回率与准确率间达到了平衡点，而深度纠错模式牺牲了部分速度，大幅提高准确率。通过设置不同的模式为用户提供多样化以及人性化的纠错体验。

纠错测试：用户在此提交短句以快速体验纠错结果，而不需要以文件形式上传。为提高响应率以及节省资源，只提供速度优先与中度纠错模式。

查看纠错记录：论文纠错完成后用户可以查看纠错结果，纠错结果是分句后的论文，一行原文本对应一行纠错文本，用户可以通过复制按钮快速定位论文中原文

本位置，进行纠错。同时每句话提供反馈按钮功能，以便后续改进。

管理纠错记录：用户能够删除记录、论文文件以及纠错数据，也能下载论文或是修改论文隐私设置。

4.3.3 数据库设计

（1）概念模型设计

在数据库设计阶段，根据用户需求和业务规则，将业务实体及其属性、关系等概念抽象出来，如图 4-9 所示。

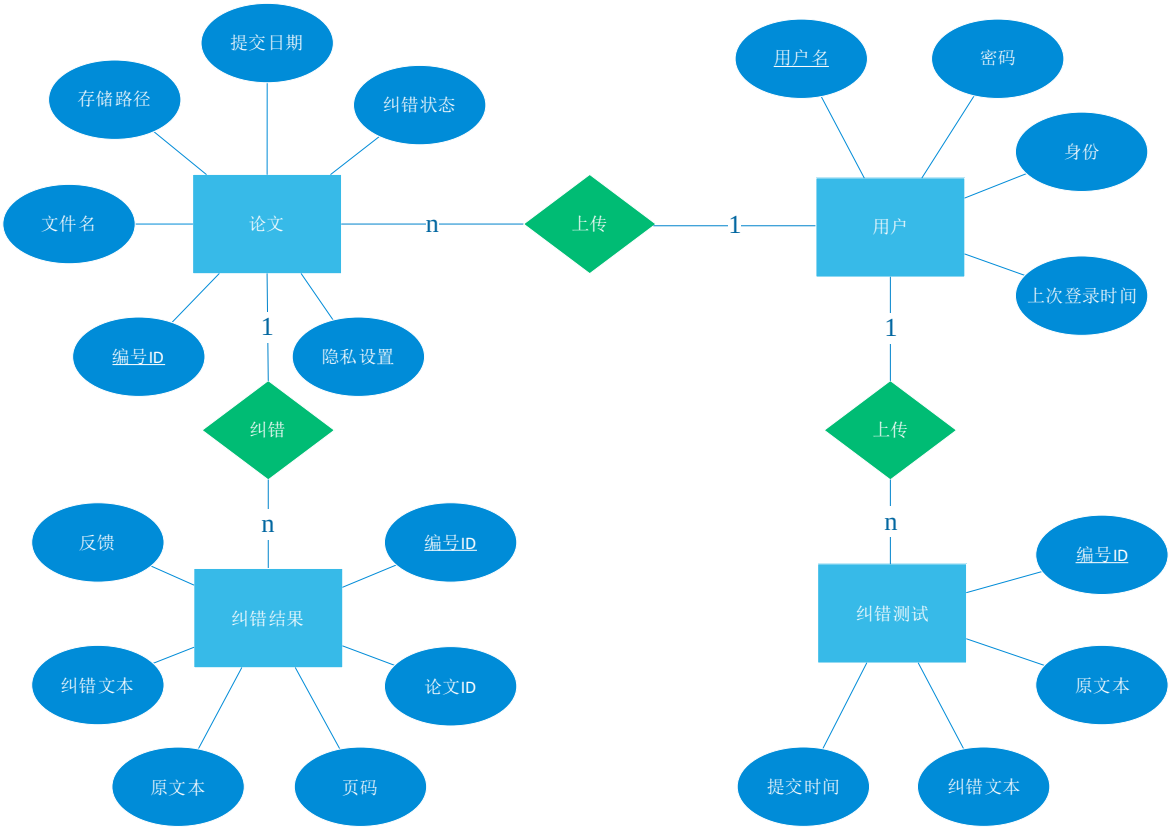


图 4-9 概念模型设计图

（2）逻辑结构设计

用户表 user，表中存储用户的账密、登录时间以及身份信息，如表 4-1 所示。

表 4-1 用户表 user

字段名	数据类型	字段说明	约束
username	varchar	用户名	主键，不为空
password	varchar	密码	不为空
time	date	登录时间	不为空
identity	varchar	用户身份	不为空
privacy	int	隐私设置	不为空
salt	varchar	盐值	不为空

论文提交记录表 commit_record，表中记录了论文编号、所属用户、论文信息等数据，如表 4-2 所示。

表 4-2 论文提交记录表 commit_record

字段名	数据类型	字段说明	约束
file_id	int	文件编号	主键，不为空
username	varchar	所属用户名	外键，不为空
file_name	varchar	论文名	不为空
file_path	varchar	论文存储路径	不为空
json_path	varchar	json 存储路径	不为空
date	date	提交时间	不为空
status	int	论文纠错状态	不为空
is_delete	int	是否删除	不为空
file_privacy	int	文档隐私	不为空
gec_mode	int	纠错模式	不为空

纠错测试表 test，表中存储纠错句子编号、所属用户、纠错信息与时间，如表 4-3 所示。

表 4-3 纠错测试表 test

字段名	数据类型	字段说明	约束
id	int	句子编号	主键，不为空
username	varchar	所属用户名	外键，不为空
source_text	varchar	原文本	不为空
target_text	varchar	纠错文本	不为空
time	date	纠错时间	不为空
judge	int	纠错反馈	不为空

论文纠错结果表 errors，表中存储论文里每句话的纠错结果，如表 4-4 所示。

表 4-4 论文纠错结果表 errors

字段名	数据类型	字段说明	约束
id	int	句子编号	主键，不为空
file_id	int	所属论文编号	外键，不为空
source_text	varchar	原文本	不为空
target_text	varchar	纠错文本	不为空
judge	int	纠错反馈	不为空

验证码表 verify，表中记录用户注册或修改密码时发送的验证码和发送时间，如表 4-5 所示。

表 4-5 验证码表 verify

字段名	数据类型	字段说明	约束
id	int	句子编号	主键，不为空
file_id	int	所属论文编号	外键，不为空
source_text	varchar	原文本	不为空
target_text	varchar	纠错文本	不为空
judge	int	纠错反馈	不为空

4.4 系统可行性分析

系统可行性分析的主要目的是确定计划中的系统是否可以在现有的技术、资源和时间限制下实现，并且是否符合用户需求和预期目标。本系统选择的开发技术有后端 Flask 框架、前端 Vue 框架以及数据库 SQLite 等，均为目前成熟且先进的技术，足够让我们构建一个功能完备的面向科技文献领域的文本纠错系统。

纠错模型可以采用目前十分流行的基于 NMT（神经机器翻译）的文本纠错技术比如 Seq2Seq（序列到序列）模型，以及序列标注 Seq2Edit（序列到编辑）模型。同时，中文纠错领域也有一些公开的数据集，可供模型训练。因此，本系统的开发具备技术上的可行性。

在开发成本部分，系统的开发只需要一台配置了相关环境的电脑，网站的部署与运行可使用组内的服务器。因此，本系统的开发具备经济可行性。

4.5 本章小结

本章介绍了纠错系统的系统分析与设计，先进行了系统需求分析，包括功能性需求和非功能性需求，接着对系统进行用例分析，绘制了用例图和时序图。随后阐述了系统概要设计，简单介绍了系统总体设计，重点说明了系统的功能模块，并对数据库绘制了 E-R 图和逻辑结构表。

第五章 系统模块详细设计与实现

本章详细阐述了系统各功能模块的设计与实现，展示了系统主要功能页面与核心代码。

5.1 登录注册模块

5.1.1 登录

用户只有通过登录才能正常使用纠错系统提供的功能。在登录窗口，用户需要输入正确的邮箱和密码，才能通过登录验证。后端在接收到用户输入的账号密码之后，会先检查账号是否存在，再对密码进行 md5 加密，与数据库中对应账号的密码进行比较。登录界面如图 5-1 所示。



图 5-1 登录界面

用户登录成功后，用户信息会自动存入 session 中。session 可用来保持用户登录状态，也能在用户发送请求时验证身份。session 的存活时间设为了 15 分钟，也就是说每 15 分钟用户需要重新登录。

为了安全性考虑，登录系统后用户每次发送请求都需要经过 session 身份认证，这是为了避免有人通过修改 cookie 绕过别人的登录状态，并修改 url 参数冒充他人进行请求，窃取隐私。然而由于系统采用分布式架构，即前后端分别部署在不同服务器，session 的设置涉及到跨域问题，会发生系统无法定位用户 session 的情况，需要对前后端进行分别的参数设置。session 身份认证的核心代码如下：

```
def valid_username(username):
    session_username = session.get("username")
    if username == session_username:
        print("验证成功")
        return True
    else:
        print("验证失败，非本人操作")
        return False

if not valid_username(username):
    return {
        'code': 707,
        'message': "请先登录"
    }

# 用户登录成功时的操作
session['username'] = username
print("登录用户：", session.get('username'))
cur.execute("UPDATE user SET time = ? WHERE username = ?", (now_time,
username))  # 更新最近登录时间
```

5.1.2 注册和修改密码

用户可以在注册界面创建新账号，输入邮箱后，后端会自动生成随机 6 位验证码，在存入数据库的同时以邮件形式发送给填写的邮箱。接着用户先后需要填写密码、选择身份与填写验证码，最后接受我们的用户隐私保护声明，即可完成注册流程。注册界面如图 5-2 所示。



图 5-2 注册界面

用户注册过程中也有一些约束，比如注册信息不能有空、验证码必须填写正确、邮箱格式检验、密码不能小于 6 位、邮箱不能重复注册等。用户注册成功后，后端会随机生成 20 位字母数字构成的字符串作为盐值，对密码进行 md5 加密，盐值和加密后的密码都会存入数据库 user 表中，待登录时验证使用。为限制用户无限制利用发送验证码功能发送邮件，添加了每 1 分钟只能发送一次，且同一 session 只能发送 15 次验证码的限制（session 只能维持 15 分钟）。

为避免用户忘记密码无法登录的问题，本系统设置了修改密码操作，界面如图 5-3 所示。功能与注册类似，也是需要用户填写邮箱后发送验证码。用户需要连续填写两次新密码，前端负责校验两次密码是否一致，否则不予通过。填写验证码并提交后就能成功修改密码。



图 5-3 修改密码

5.2 用户功能模块

5.2.1 管理用户信息

登录进入系统后就能看到上传论文的界面，如图 5-4 所示。左侧是导航条，可进入不同的界面。用户信息在右上角，如图 5-5 所示。

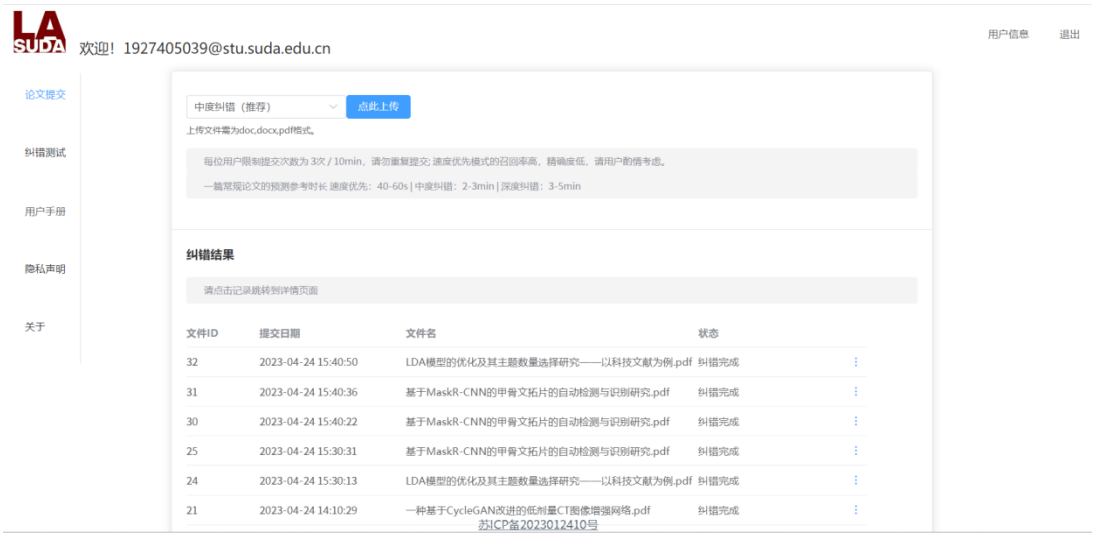


图 5-4 首页



图 5-5 用户信息

目前存储的用户信息并不多，用户可以在这里修改个人信息以及修改密码，修改密码功能与登录注册模块里的忘记密码功能类似，只是不需要输入验证码，只需验证原密码是否一致即可，界面如图 5-6 所示。



图 5-6 修改密码

5.2.2 论文上传

在此界面，用户可以选择上传需要纠错的论文，并且选择不同的纠错模式，不同的模式对应不同的纠错力度、准确率与召回率，如图 5-7 所示。

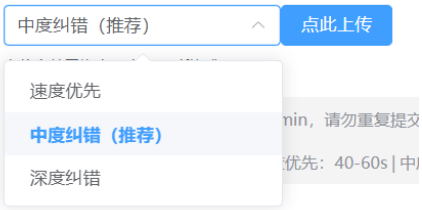


图 5-7 选择纠错模式

其中，速度优先模式只采用了一个 Seq2Edit 模型，因此速度较快，正常篇幅的论文只需 40~60 秒即可纠错完成，具有较高的召回率，缺点是准确率不高，有很多过度纠错的现象。中度纠错模式目前采用了三个 Seq2Edit 模型，这些模型分别采用了 BERT、StructBERT 和 MacBERT[23]预训练模型，同时在训练数据和参数选择上也有区别。采用模型集成技术，对三个模型得到的纠错文本中的编辑进行投票，因此最终得到的纠错结果有了不错的准确率和召回率，在性能和速度上达到了平衡。经测试，一篇正常篇幅的论文在中度纠错模式中需要约 2 分钟的纠错时间。

在深度纠错模式中，采用了多个 Seq2Edit 和 Seq2Seq 共同纠错的方式。与中度纠错模式一样，这些模型在预训练模型、训练集以及参数上有不同的选择，因此具有不同的纠错性能。由于 Seq2Seq 模型的推理时间远大于 Seq2Edit 模型，且因为分

配策略的限制很难并行运行这些模型，对于正常篇幅的论文，深度纠错模式需要约4分钟的纠错时间。速度上的劣势换来的是较高的精准度和较少的过纠现象，给用户带来更好的体验。

在之后的改进方案中，用户可以给深度纠错模式选择不同的阈值，在前端展示不同的纠错结果。阈值的选择影响模型集成中的投票数选择，以提供更多样化的纠错体验。

目前系统仅支持 doc、docx 以及 pdf 论文的上传。后端接收到论文后，会经过解析——提取文本——分句——清理筛选等步骤，提取出用于纠错的句子。对于 pdf 文件，在对比了 PDFMiner、PyPDF2、PyMuPDF 等多种 pdf 解析库之后，本系统最终选择使用效果最好的 PyMuPDF 库用于解析 pdf 并提取文本。得到所有文本后按“。！？；”等标点符号进行分句。最后按照设定的规则进行筛选，比如去除英文超过一半的句子、去除过短句、过滤不合法字符和乱码字符、删除页眉页脚噪音数据等，以上均使用正则表达式进行筛选并清理。对于 docx 文件的处理比较简单，使用 docx 库提取出文本并继续分句和筛选清理，相比起 pdf 文件少了很多噪音数据。而对于 doc 文件，python 没法直接解析，需要先转换成 docx 格式再解析。对于 doc 转 docx 的操作 windows 和 linux 有不同的解决方案，本系统最终是通过 linux 中的 libreoffice 命令执行的。

由于论文纠错需要一定的时间，用户上传论文后不能马上得到后端返回的纠错结果，但是迫切需要得到上传成功的反馈，不能干等后端响应。因此，在查阅了大量资料后，本系统决定使用多线程异步处理来解决此问题。使用 multiprocessing 库，在系统启动时自动创建一个线程池。用户上传论文后，系统会自动将解析筛选与纠错的代码放到多线程中执行，同时立即返回“上传成功”反馈。

在后面的开发过程中，为了减轻 GPU 服务器的 CPU 负担，我们将后端分别部署在一台 CPU 服务器和一台 GPU 服务器上。CPU 服务器负责数据分发以及和前端连通，GPU 服务器只负责运行纠错模型并返回数据给 CPU 服务器。这里截取了关键代码部分如下所示。

```
pool = ThreadPool(processes=processes)  # 建立线程池
def gec_task(file_id, json_path, doc, GEC_mode):
    inputs = preprocess(file_id, json_path, doc)  # 解析论文文件
    data = {
        "inputs": inputs,
        "file_id": file_id,
        "GEC_mode": GEC_mode,
    }
    url = 'http://192.168.134.32:5000/gec_task'  # 请求接口
    requests.post(url, data=json.dumps(data))  # 发送请求
pool.apply_async(gec_task, (file_id, json_path, doc, GEC_mode),
error_callback=Except)  #多线程异步处理
@app.route('/receive', methods=['POST'])  # 接收纠错文本
def receive():
    args = json.loads(request.data)
    ans = args["ans"]
    file_id = args["file_id"]
    conn = sqlite3.connect('./database/mydata.db')
    cur = conn.cursor()
    for i in range(len(ans)):
        cur.execute("INSERT INTO errors VALUES (?, ?, ?, ?, ?, ?)",
                    (None, file_id, 0, ans[i][0], ans[i][1], None))
    conn.commit()
```

```
    # 纠错完成，更新 status
    logger.info("file_id:" + str(file_id) + "保存完成!")
    cur.execute("UPDATE commit_record SET status = 1 WHERE file_id = %s" %
file_id)
    conn.commit()
    conn.close()
    return "200"
```

其中，gec_task 函数为预处理论文文件并得到纠错原文本送入纠错模型的流程，系统将原文本 inputs 和纠错模式 GEC_mode 以 requests 请求的方式发送给 GPU 服务器。GPU 服务器在纠错完成时会将纠错结果从/receive 接口返回，并存入数据库。GPU 服务器接收原文本并纠错的核心代码如下所示。

```
@app.route('/gec_task', methods=['POST'])
def gec_task_interface():
    def Except(e):
        print(e)
        logger.info(e)
        return e

    args = json.loads(request.data)
    inputs = args["inputs"]
    file_id = args["file_id"]
    GEC_mode = args["GEC_mode"]
    if GEC_mode != "深度纠错":
        pool.apply_async(gec_task, (file_id, inputs, GEC_mode),
error_callback=Except)
    else:
        ensemble_pool.apply_async(gec_task, (file_id, inputs, GEC_mode),
error_callback=Except)
    logger.info("新的纠错进程已加入队列,file_id:" + str(file_id))
            return "200"
```

根据纠错模式的不同，放入的线程池也不同，这是因为不同纠错模式的 GPU 资源占用量不同，分配策略也有所不同。

为完善异常处理机制，纠错过程中报错后会记录报错信息并将 fild_id 发送给 CPU 服务器以更新论文纠错状态，核心代码如下所示：

```
try:
    # 纠错过程
except Exception as e:
    logger.info("子进程出错")
    url = 'http://192.168.134.30:5000/error'  # 请求接口
    data = {
        "file_id": file_id
    }
    requests.post(url, data=json.dumps(data))  # 发送请求
    logger.info(e)
return
```

为限制用户大量提交论文，导致资源紧张，系统设置了 10 分钟只能提交 3 篇论文的限制，可以通过数据库中论文提交时间来判断。

5.2.3 纠错测试

考虑到用户可能希望快速体验纠错效果而不想等待上传论文，本系统设计了纠错测试功能模块，界面如图 5-8 所示。



图 5-8 纠错测试

相比起论文上传部分，纠错测试部分的代码简化了很多。首先不再需要解析文件并处理筛选，只需要将句子分句处理即可。其次短句的纠错速度较快，不需要多线程异步处理也能接受，用户仅需略微等待即可得到纠错结果。

5.2.4 查看纠错记录

用户上传论文后立即更新论文提交记录，默认按时间排序。在这里，用户可以看到论文的提交时间、文件名与当前纠错状态，如图 5-9 所示。



图 5-9 查看纠错记录

由于多线程池的机制，在达到了可同时纠错的任务数量上限后，新的纠错任务会进入阻塞状态，也就是所谓的排队。在排队状态时，用户希望看到自己论文的排队数，以估算预计纠错时间。因此，系统需要维护一个纠错任务队列并实时更新。

初步想法是利用现成的线程池，用户每次发送查看记录的请求时自动遍历，由于已运行完成的任务会自动离开线程池，因此当前任务在线程池的下标就是排队数。然而在开发过程中发现有些纠错任务会莫名其妙从线程池中消失不见，但其仍在正常运行。手动在全局变量中维护一个队列仍有这个问题。在查阅了大量资料后才明白，由于本系统使用 uWSGI 技术将 Flask 部署在服务器，而 uWSGI 默认是多进程运行的，不同进程间的数据并不互通，用户请求是随机分配其中进程执行的，因此不一定每次都能读到全局变量。最后我学习了 uWSGI 的缓存技术，手动创建一个队列维护在缓存中，队列里只存论文的 file_id，终于实现了查看任务排队数的功能，如图 5-10 所示。



图 5-10 查看排队数

等到状态显示论文纠错完成时，可点击记录进入纠错结果界面。在该界面中，用户能够看到每句模型认为有错的句子并且给出对应的纠错结果。原文本与纠错文本形成上下对应的形式，并且高亮了删除和添加的部分，方便用户进行观察。纠错结果界面如图 5-11 所示。



图 5-11 查看纠错结果

要实现上述显示效果，实际上是实现找到原文本与纠错文本之间的最长公共子序列（LCS）问题，核心代码如下所示：

```
import difflib
d = difflib.Differ()
for index, item in enumerate(ans):
    source = item[3]  # 原文本
    target = item[4]  # 纠错文本
    l = list(d.compare(source, target))  # 返回纠错文本相比原文本的最小编辑
    tgt = []
    for i in l:
        if i[0] == ' ':
            tgt.append([i[-1], 0])
        elif i[0] == '-':
            tgt.append([i[-1], 1])
        elif i[0] == '+':
            tgt.append([i[-1], 2])
```

简单起见，本系统直接使用 difflib 库实现了此功能，该库可以得到纠错文本相比原文本的最小编辑操作，每个字符前都有一个标签，若为“-”说明需要删除，若为“+”说明需要添加，若为“ ”则保持不变。系统将每个字连同编辑标签一起返回前端，前端根据标签分别渲染每个字，实现了删除和添加字符的高亮效果。

每句纠错记录的右边是用户的反馈按钮，用户可以根据纠错效果选择“赞”或“踩”两种评价，这些信息在未来将用于改善我们的纠错模型。

5.2.5 管理纠错记录

用户可以管理自己账号的每条论文纠错记录，包括下载原文、删除记录、删除原文件以及设置文档隐私等，管理界面如图 5-12 所示。





图 5-12 管理纠错记录

对于下载操作，本系统采用了 Flask 自带的 send_from_directory(directory, filename, as_attachment=True)函数来解决，由于论文上传后数据库保存了论文路径与文件名，只需要将参数填入 send_from_directory 即可。删除功能中默认采用了只删除记录，将数据库对应记录的删除位设为 1，在展示纠错记录时筛去即可。若用户选择了删除数据库数据，则将该论文的纠错数据与存储的文件一并删除。用户论文的隐私设置会影响我们对于该论文及其纠错数据的使用方式，当用户上传隐私数据后可设为“不可使用”。

5.3 本章小结

本章分登录注册模块和用户功能模块，详细介绍了各子功能模块的设计与实现。在给出了系统每一部分的功能说明与设计思路的同时，也针对系统开发过程中的重难点并给出了核心代码。

第六章 系统测试

系统开发完毕后，本文根据系统功能需求进行了系统测试，结果如下：

（1）登录注册模块测试：如表 6-1。

表 6-1 登录注册模块测试表

测试功能	用例	预期结果	实际结果	测试结果
注册	邮箱格式不对	注册失败	提示邮箱格式不对	通过
注册	输入错误验证码	注册失败	提示验证码错误	通过
注册	密码格式不对	注册失败	提示密码格式不对	通过
注册	填写正确信息	注册成功	注册成功	通过
修改密码	重复密码填写错误	修改失败	修改失败	通过
修改密码	输入未注册邮箱	修改失败	修改失败	通过
修改密码	输入正确信息	修改成功	修改成功	通过
登录	输入未注册账号	登录失败	提示邮箱不存在	通过
登录	输入错误密码	登录失败	提示密码错误	通过
登录	输入正确邮箱密码	登录成功	登录成功	通过
登录	退出登录	注销成功	注销成功	通过

（2）用户模块测试，如表 6-2。

表 6-2 用户模块测试表

测试功能	用例	预期结果	实际结果	测试结果
用户信息管理	进入用户信息页并修改	看到用户信息并成功修改	修改成功	通过
修改密码	原密码输入错误	修改失败	提示原密码错误	通过
修改密码	新密码两次输入不一致	修改失败	提示两次输入不一致	通过
修改密码	填写正确信息	修改成功	修改成功	通过
论文上传	选择错误格式文件	上传失败	提示格式错误	通过
论文上传	选择正确格式文件	上传成功	上传成功	通过
论文上传	选择不同的纠错模式	上传成功	上传成功	通过
纠错测试	选择不同纠错模式	测试成功	测试成功	通过
查看纠错记录	选择纠错完成论文	查看纠错数据	查看纠错数据	通过
查看纠错记录	对纠错数据进行反馈	反馈成功	反馈成功	通过
管理纠错记录	下载原文	下载成功	下载成功	通过
管理纠错记录	删除记录	删除成功	删除成功	通过
管理纠错记录	删除纠错数据和论文	删除成功	删除成功	通过
管理纠错记录	设置文档隐私	修改成功	修改成功过	通过

测试结果表明，纠错系统能够正常运行。系统能根据用户的异常输入给出完备的提示，并对用户的所有操作及时给予反馈。本系统能够满足纠错系统开发所需的各类需求。

第七章 总结与展望

7.1 本文总结

本篇论文实现了一款面向科技文献领域的文本纠错系统。该系统基于后端 Flask 和前端 Vue 技术，利用神经机器翻译技术对文本进行纠错，并提供用户界面进行交互操作。本系统可以有效地改善科技文献的语言表达，提高文献的质量，对于科研人员和学术界具有重要的意义。

具体来讲，本文主要做了以下工作：

（1）分析文本纠错技术的课题背景，介绍开发面向科技文献领域的文本纠错系统的目的和意义。接着阐述了系统开发的相关技术以及采用的文本纠错技术。

（2）介绍了采用的数据集，并利用数据增强技术处理爬虫获取的科技文献数据，在小规模数据集上对 Seq2Edit 模型进行优化，使之在科技文献领域上的文本纠错表现更好。

（3）本文通过对用户需求的分析，确定了系统的功能性需求和非功能性需求，并进行了系统用例分析。随后，本文给出了系统的总体设计，设计的数据库结构。最后，本文对系统的可行性进行了分析。

（4）根据系统需求，实现了系统主要功能模块，着重分析了系统详细开发过程中的重难点。

（5）对系统各功能进行测试，确保系统能够正常运行。

7.2 后续工作展望

本系统已经基本完成了文本纠错需求，但还有很多可以不断改进的地方。

（1）可以提供更多样化的纠错模式，提供更多的阈值选择，使用户查看不同等级的纠错结果，提高用户的使用体验。

（2）继续优化模型性能，在保障纠错准确率的前提下加快系统的纠错效率，减少用户的等待时间。

（3）继续做好领域移植工作，使模型在科技文献领域内的纠错表现更好。

（4）引入更多更优秀的纠错模型来改进模型集成，以提高纠错的精度和鲁棒性。

参考文献

[1] 李守振, 张南平, 常国锋.Web 应用分层与开发框架设计研究[J].计算机工程, 2006 (22) :274-276.

[2] 牛作东,李捍东.基于 Python 与 flask 工具搭建可高效开发的实用型 MVC 框架[J].计算机应用与软件,2019,36(07):21-25.

[3] 徐杰.基于 Flask 的分布式 Web 服务架构研究与应用[J].工业控制计算机,2020,33(10):97-98+101.

[4] 旷志光,纪婷婷,吴小丽.基于Vue.js的后台单页应用管理系统的研究与实现[J].现代计算机(专业版),2017,(30):51-55.

[5] 李晓薇.vue.js 前端应用技术分析[J].网络安全技术与应用,2022,(04):44-45.

[6] Bahdanau D , Cho K , Bengio Y . Neural Machine Translation by Jointly Learning to Align and Translate[J]. Computer Science, 2014.

[7] Zhixing Tan, Shuo Wang, Zonghan Yang, Gang Chen, Xuancheng Huang, Maosong Sun, Yang Liu. Neural machine translation: A review of methods, resources, and tools[J]. AI Open, 2020, (1):5-21.

[8] Zhang, Yue, Zhenghua Li, Zuyi Bao, Jiacheng Li, Bo Zhang, Chen Li, Fei Huang and Min Zhang. MuCGEC: a Multi-Reference Multi-Source Evaluation Dataset for Chinese Grammatical Error Correction[J]. North American Chapter of the Association for Computational Linguistics, 2022.

[9] Sutskever, Ilya and Vinyals, Oriol and Le, Quoc V. Sequence to Sequence Learning with Neural Networks[C]. Advances in Neural Information Processing Systems, 2014, (27).

[10] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[11] Lewis, Mike, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov and Luke Zettlemoyer. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension[C]. Annual Meeting of the Association for Computational Linguistics,

2019.

[12] Omelianchuk, K., Atrasevych, V., Chernodub, A.N., & Skurzhanskyi, O.. GECToR – Grammatical Error Correction: Tag, Not Rewrite[C]. Workshop on Innovative Use of NLP for Building Educational Applications, 2020.

[13] Stahlberg, Felix and Shankar Kumar. Seq2Edits: Sequence Transduction Using Span-level Edit Operations[C]. Conference on Empirical Methods in Natural Language Processing, 2020.

[14] Stahlberg F, Kumar S. Synthetic data generation for grammatical error correction with tagged corruption models[J], arXiv preprint arXiv:2105.13318, 2021.

[15] Tarnavskyi M, Chernodub A, Omelianchuk K. Ensembling and Knowledge Distilling of Large Sequence Taggers for Grammatical Error Correction[J]. Annual Meeting of the Association for Computational Linguistics, 2022.

[16] Edunov, S., Ott, M., Auli, M., & Grangier, D.. Understanding Back-Translation at Scale[J]. Conference on Empirical Methods in Natural Language Processing, 2018.

[17] Yuan Z, Bryant C. Document-level grammatical error correction[C]. Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications, 2021: 75-74.

[18] Zhang, Yue, Bo Zhang, Zhenghua Li, Zuyi Bao, Chen Li and Min Zhang. SynGEC: Syntax-Enhanced Grammatical Error Correction with a Tailored GEC-Oriented Parser[J], arXiv preprint arXiv:2210.12484, 2022.

[19] Yuan Z, Taslimipoor S, Davis C. Multi-class grammatical error detection for correction: A tale of two systems[C]. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021: 8722-8736.

[20] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

[21] 邹楠,厉志成.基于统一建模语言的软件需求分析应用研究[J].电脑知识与技术,2022,18(35):22-24+31.

[22] 安永刚.计算机应用软件的需求分析与开发[J].数字技术与应用,2022,40(07):166-

168.

[23] Cui, Yiming, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang and Guoping Hu. Revisiting Pre-Trained Models for Chinese Natural Language Processing[J], ArXiv abs/2004.13922, 2020.

致 谢

在我的毕业设计的开发实现过程中，我首先要感谢我的指导老师，也是我未来三年的研究生导师李正华老师。在这期间，李老师多次为我指导纠错系统的研发方向，也提供了足够的资源让我部署系统与训练模型，在最后的论文撰写时耐心地为我指导论文的细节部分。李老师渊博的知识与端正的科研态度给我留下了无比深刻的印象，在此谨向李老师表达我真诚的谢意。

其次我要感谢蒋浩辰师兄，他在纠错系统的设计与实现过程中帮助我完成了前端的开发。我们无数次遇到问题、讨论问题并解决问题，无数次讨论系统的细节并着手实施。他在系统开发与科研上给予我无数的指导，我才能一次次克服困难，直到纠错系统顺利完成与上线。

最后，我要感谢我的家人朋友，他们一直默默支持着我，给我无言的关心与支撑。四年的学习生活已近尾声，在此我向所有帮助过我的人说声谢谢！

苏州大学本科生毕业设计（论文）任务书

学院（部）：计算机科学与技术学院

<table>
<tr><td colspan="6">设计（论文）选题：面向科技文献领域的文本纠错系统实现</td></tr>
<tr><td>指导教师姓名</td><td>李正华</td><td>职 称</td><td>副教授</td><td>类 别</td><td>毕业论文</td></tr>
<tr><td>学 生 姓 名</td><td>乔子恒</td><td>学 号</td><td>1927405039</td><td>设计（论文）类型</td><td>应用型课题</td></tr>
<tr><td>年 级</td><td>2019</td><td>专 业</td><td>计算机科学与技术</td><td>是否隶属科研项目</td><td>是</td></tr>
<tr><td colspan="6">1、设计（论文）的主要任务及目标
本论文的主要任务是通过自然语言处理和深度学习技术，训练一个面向科技文献领域的纠错模型，自动检测和纠正文本中的语法错误、拼写错误、标点符号错误等等。目标是提高文献的质量和可读性，使得科技文献更加准确、规范、流畅，从而更好地传达作者的意图，帮助读者更好地理解和使用科技文献。
为了保证纠错模型的性能，高质量的训练数据极为重要。本论文的另一目标就是探索获取高质量训练数据的方法。具体方式如下：
1. 通过爬虫等方式获取科技文献领域的原始数据，并用 pdf 解析对其文本进行提取，再通过数据预处理和清洗筛选等步骤得到正常的句子。
2. 尝试一些数据挖掘方式，将其转化为可供模型使用的训练数据。</td></tr>
<tr><td colspan="6">2、设计（论文）的主要内容
数据获取：本论文将尝试通过爬虫来获取原始数据，并在后续使用 pdf 解析器对文本进行提取，再通过数据预处理的方式对数据进行筛选和清洗，以此提高数据质量，将原始数据转化为可供模型使用的训练数据。
数据挖掘：尝试考虑通过一些数据挖掘方式，生成语句对作为训练数据，放入纠错模型训练。
纠错模型训练：文本纠错系统需要接入纠错模型，自动检测文本中的错误，包括拼写错误、语法错误等，并给出预测的纠正答案。
纠错系统后端搭建：使用 flask 框架开发系统后端，方便纠错模型的接入。</td></tr>
<tr><td colspan="6">3、设计（论文）的基本要求
数据获取基本要求：
能通过爬虫抓取不同期刊的论文，获取科技文献领域原始数据。编写数据处理 pipeline，能对爬取到的论文原始数据进行文本提取，并通过数据预处理对其进行筛选和清洗，提高数据质量。
纠错系统基本要求：
支持用户提交 PDF、Docx、Doc 格式论文，能够获得具有参考意义的纠错结果。
页面清晰、简洁、一目了然，方便用户使用。
论文基本要求：</td></tr>
</table>

主题清晰，语言精练、明确，能够清楚地阐释本人的毕业设计工作。
撰写的内容与结果密切相关，分析能够结合理论，文献阅读充分，文笔流畅。

4、主要参考文献

[1] Dahlmeier D, Ng H T, Wu S M. Building a large annotated corpus of learner English: The NUS corpus of learner English[C]//Proceedings of the eighth workshop on innovative use of NLP for building educational applications. 2013: 22-31.
[2] Tajiri T, Komachi M, Matsumoto Y. Tense and aspect error correction for ESL learners using global context[C]//Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). 2012: 198-202.
[3] Yannakoudakis H, Briscoe T, Medlock B. A new dataset and method for automatically grading ESOL texts[C]//Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies. 2011: 180-189.
[4] Yannakoudakis H, Briscoe T, Medlock B. A new dataset and method for automatically grading ESOL texts[C]//Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies. 2011: 180-189.
[5] Yannakoudakis H, Briscoe T, Medlock B. A new dataset and method for automatically grading ESOL texts[C]//Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies. 2011: 180-189.
[6] Baolin Zhang. 2009. Features and functions of the HSK dynamic composition corpus (in Chinese). International Chinese Language Education, 4:71–79.
[7] Zhang Y, Li Z, Bao Z, et al. MuCGEC: a Multi-Reference Multi-Source Evaluation Dataset for Chinese Grammatical Error Correction[J]. arXiv preprint arXiv:2204.10994, 2022.
[8] Zheng Yuan and Ted Briscoe. 2016. Grammatical error correction using neural machine translation. In Proceedings of NAACL-HLT, pages 380–386.
[9] Sascha Rothe, Jonathan Mallinson, Eric Malmi, Sebastian Krause, and Aliaksei Severyn. 2021. A simple recipe for multilingual grammatical error correction.In Proceedings ofACL-IJCNLP, pages 702–707.
[10] Sascha Rothe, Jonathan Mallinson, Eric Malmi, Sebastian Krause, and Aliaksei Severyn. 2021. A simple recipe for multilingual grammatical error correction.In Proceedings ofACL-IJCNLP, pages 702–707.
[11] Zhang Y, Zhang B, Li Z, et al. SynGEC: Syntax-Enhanced Grammatical Error Correction with a Tailored GEC-Oriented Parser[J]. arXiv preprint arXiv:2210.12484, 2022.

5、进度安排

	设计（论文）各阶段任务	起 止 日 期
1	资料收集，明确任务，完成任务书	2023-03-01 至 2023-03-09
2	阅读相关文献资料，完成外文翻译，文献综述	2023-03-10 至 2023-03-19
3	学习前后端技术，完成数据积累与预处理，搭出系统基本框架。	2023-03-20 至 2023-04-05
4	中期检查	2023-04-06 至 2023-04-12
5	优化调整纠错模型，优化系统	2023-04-13 至 2023-04-20
6	撰写毕业论文，准备毕业答辩	2023-04-21 至 2023-05-12
7	完善论文，毕业答辩	2023-05-07 至 2023-05-31

外文文献资料

Stop! Grammar time: University students' perceptions of the automated feedback program Grammarly

Ruth O'Neill
Academic Learning Centre, School of Access Education, CQUniversity, Australia

Alex M. T. Russell
Experimental Gambling Research Laboratory, School of Health, Medical and Applied Sciences, CQUniversity, Australia

Universities are increasingly looking towards online technology to assist their students. Grammarly advertises itself as the world's most accurate online grammar checker, yet little research exists on its performance as a feedback tool in universities. This paper reports on a study investigating students' perceptions of Grammarly when used in conjunction with advice from an academic learning advisor. Using a mixed methods sequential explanatory design, the study compared one group of students' responses to the feedback they received from Grammarly (n = 54) with another group's responses to the traditional non-automated grammar feedback they received from the Academic Learning Centre at CQUniversity (n = 42). Results show that students receiving feedback from Grammarly responded more positively to 9 of the 15 survey items and were significantly more satisfied with the grammar advice that they received compared with non-Grammarly students. No significant differences were registered between cohorts or delivery mode, which suggests that Grammarly can be used effectively by academic learning advisors to provide grammar support for both international and domestic students, online and on campus. It is recommended, however, that the program be used in conjunction with academic learning advisor input as the program is currently not accurate enough for independent use to be justified.

Introduction

Many students in the Australian higher education system struggle with the demands of tertiary education because of language problems (Harris, 2016). This difficulty has been well documented with international students, who tend to exhibit more surface-level errors relating to accurate grammar, punctuation and spelling (Tynan & Johns, 2015). Traditionally, domestic students performed better than international students in relation to grammatical accuracy (Scouller, Bonanno, Smith, & Krass, 2008), but changes in the domestic student population have resulted in members of this cohort also exhibiting difficulties with sentence structure and form (Cocks & Stokes, 2013). Grammar feedback is, therefore, needed for different cohorts to achieve academic success. Responsibility for that feedback is often perceived to lie with learning support units, typically known as the Academic Learning Centre (ALC). This expectation can, however, conflict with the ALC's perception of their role, resulting in students leaving ALC sessions dissatisfied with the amount and quality of the grammar feedback they receive. Automated feedback programs have been advocated as tools to reconcile this tension. Grammarly is one such program and was chosen by CQUniversity as a potential grammar feedback tool based on its ability to reconcile contrasting theories on written corrective feedback. Building on the work of Caveleri and Dianati (2016), the primary purpose of this study was to analyse students' satisfaction with Grammarly as a feedback tool and compare this response with the manual, non-automated feedback traditionally offered at CQUniversity by the ALC. It also aimed to determine whether the program was more suitable for particular student cohorts based on language level, visa status and delivery method.

Literature review

The need for grammar and grammar input

Students need to use standard grammar and adhere to grammatical conventions to succeed in academic writing at universities (Caveleri & Dianati, 2016). Students cannot, however, improve their language skills without input. International students need external input to develop their linguistic skills (Murray, 2011), including the explicit teaching of grammar needed for academic writing (Knoch, Rouhshad, & Storch, 2014). Different groups of domestic students also need practical interventions with regards to punctuation,

Australasian Journal of Educational Technology, 2019, 35(1).

grammar and spelling (McNaught & Shaw, 2016). This includes local students (defined as being born in Australia) who have not received formal grammar training as part of the school curriculum or who have accessed tertiary education through non-traditional means such as enabling programs (Cocks & Stokes, 2013). It also includes domestic non-English speaking background students (DNESB), defined by the Department of Education and Training (n.d.) as being domestic students who arrived in Australia less than 10 years before starting at university and who speak a language other than English at home.

Academic learning advisors' and students' attitudes to grammar intervention

The responsibility for grammar teaching in university settings is traditionally assigned to ALCs (Channock, D'Cruz, & Bisset, 2009) and the academic learning advisors (ALAs) who work there. ALAs at CQUniversity provide assignment feedback in online and face-to-face consultations. These consultations are 60–90 minutes, but ALAs accord a fraction of that time to grammar feedback because it is time-consuming (Dikli & Bleyle, 2014) and deemed less important than the long-term augmentation of higher order skills such as developing an argument (Winder, Kathpalia, & Koo, 2016). Students on the other hand value grammar feedback and identify immediate grammar correction as their number one concern (Huang, 2011). It was noted anecdotally that this conflict in priorities and the time constraints imposed on grammar feedback were distressing some students, because they felt they were not receiving the amount of grammar feedback they needed both for the assignment and their long-term development. They felt that if the ALC did not assist with developing their grammar, they had no other recourse, as lecturers focus on content, not the development of writing skills (Bacon & Anderson, 2004).

Automated feedback programs

Automated feedback programs could reconcile the differences between ALAs' and students' attitudes to grammar feedback, by providing students with the amount of grammar feedback they want and need without compromising ALAs' prioritisation of macro skills development. Research into online grammar checkers is limited (Caveleri & Dianati, 2016), particularly regarding students' responses to them, but studies have been undertaken on automated essay scoring systems which provide grammar feedback as part of their functionality (Ramineni & Williamson, 2012). Students' responses to the grammar feedback that these systems provide are mixed. Students using CorrectEnglish cited improvement in grammar and written accuracy as a positive facet of the program (Wang, Shang, & Briody, 2013), and those using My Editor similarly believed that it helped them to improve their grammar and spelling (Hoon, 2006). Participants in a study by Chen and Cheng (2006), however, were largely unimpressed with the grammar feedback they received. Student dissatisfaction in this study highlights the importance of the facilitator's role in the feedback process as the students' negative responses were attributed to the teachers' attitudes towards the program. One teacher was not conversant with the tool's editing functions and the other was not impressed by them, which meant that neither teacher communicated the best way for the students to use the tool.

Conflicting findings are also reported regarding the student language level that most benefits from automated feedback programs. Advanced students in Chen and Cheng's (2008) study found the focus on accuracy constraining and suggested that the automated feedback would be more useful for beginner and intermediate learners. Dikli (2010) contradicted this assertion by suggesting that automated feedback can overwhelm students with low English proficiency, and Liao (2016) agreed that automated feedback is preferable for higher performing writers because they possess the metacognitive strategies needed to develop grammatical accuracy long term. In their study on Grammarly, Caveleri and Dianati (2016) also stated that the program may benefit more-able writers because students need to have sufficient grammatical understanding to be able to filter suggestions that are incorrect, or that they deem unnecessary and not useful.

The argument as to whether automated feedback programs are more suitable for domestic or non-domestic students is similarly contentious. Hoang and Kunnan (2016) make the point that the feedback systems were not traditionally designed for non-native student writers of English and that international students do not have enough understanding of the target language to interpret and apply the suggestions the programs make. They also suggest that non-native students might lack the metalanguage to understand the suggestions provided by the programs. This argument, however, could just as easily be applied to local students who have acquired language with little to no instruction in metalanguage and rule-based principles.

Australasian Journal of Educational Technology, 2019, 35(1).

A further disadvantage of automated feedback from the student perspective is the lack of human interaction (Dikli, 2010). According to Wang et al. (2013), this is a particular problem for international students and less experienced writers who need instruction, modelling and practice to develop their written English. A lack of dialogue is additionally detrimental to writing proficiency as it lowers student motivation and makes their writing more mechanical and less authentic (Wang et al., 2013). To counteract these issues and render automated feedback useful, Hoang and Kunnan (2016) suggest a combined pedagogy in which the writing instructor acts as an intermediary between the students' work and the feedback. Research suggests that feedback given in this way is most effective (Choi & Lee, 2010) because teachers can provide instruction on how best to use the program (Burston, 2008), provide clarification of errors (Hoang & Kunnan, 2016) and eliminate errors which the program has made.

Grammarly

In 2016, the ALC at CQUniversity piloted the use of Grammarly Premium, in conjunction with ALA advice. Grammarly Premium is an automated proofreading system which can identify errors related to 250 grammar rules (Grammarly Inc., 2017). Users upload their assignment and receive two scores. The first score is based on percentage accuracy, and the second is the total number of errors that the program has identified. Grammarly organises the errors according to six categories: contextual spelling, grammar, punctuation, sentence structure, style and vocabulary enhancement. The number of errors within these specific categories are counted and presented (see Figure 1).

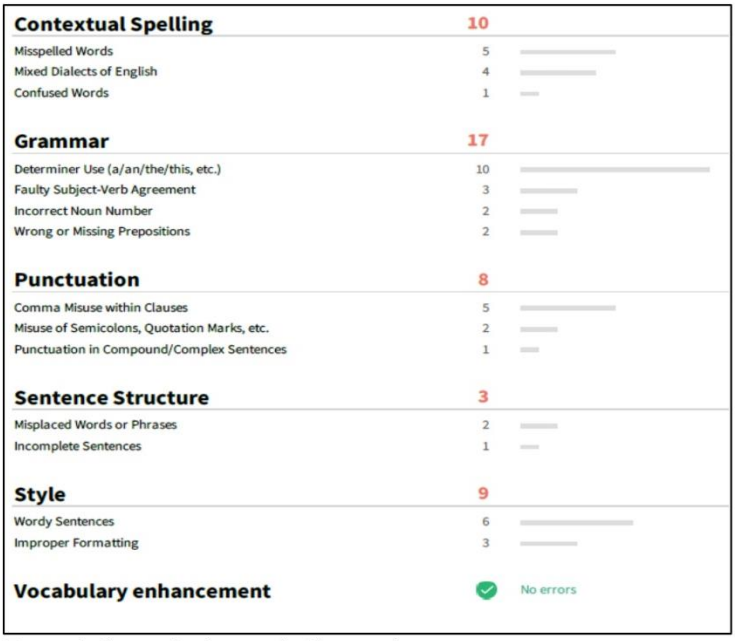


Figure 1. Categorised errors in Grammarly report

Feedback problems and Grammarly as the potential solution

Grammarly was chosen because of its potential to reconcile contrasting theories relating to written corrective feedback. Much has been written about the relative merits of direct and indirect feedback. Proponents of indirect feedback (Jamalinesari, Rahimi, Gowhary, & Azizifar, 2015) state that the error should only be highlighted for the student to independently review and correct, whereas advocates of direct feedback (Bitchener, 2008) maintain that explicit corrections need to be provided by the teacher. Grammarly reconciles these different theoretical approaches by providing teachers with the opportunity to use both. On the left-hand side of the screen, indirect feedback can be given using the underlined error. On the right side of the screen, direct feedback can be provided using the correction and explanation card (see Figure 2).

Australasian Journal of Educational Technology, 2019, 35(1).

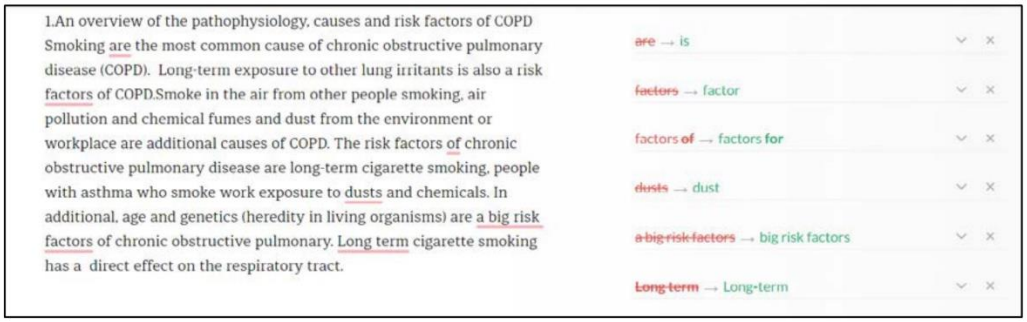


Figure 2: Direct and indirect feedback in Grammarly

Further debate exists over the amount of feedback which should be given. Students value a correct-all approach (Salehfard & Bagheri, 2014) and can become frustrated when their goals for line-by-line editing are not addressed (Linville, 2009). Research, on the other hand, has shown a need for more targeted error correction (Jamalinesari et al., 2015) as focused feedback on high-frequency errors improves students' writing (Linville, 2009). Grammarly again offers a middle ground solution, as its ability to identify 250 breaches in grammatical rules offers an all-inclusive approach, whereas the categorisation system provides the students with cues to address high-frequency errors.

An additional reason for choosing Grammarly was that little research exists on its performance within university settings. Japos (2013) conducted a study of undergraduate research students, finding that Grammarly improved their written accuracy. Qassemzadeh and Soleimani (2016) concluded that Grammarly performed better than traditional teacher intervention with regards to long-term retention of passive voice rules. Students in a study by Reis and Huijser (2016) preferred Grammarly to the alternative feedback system Marking Mate. Results from Caveleri and Dianati's (2016) investigation of students' perceptions of Grammarly within an Australian higher education setting were also largely positive, with students stating that Grammarly was useful and easy to use. They also maintained that it improved their written work and helped develop their understanding of grammatical rules.

This study builds on Caveleri and Dianati (2016) but offers essential points of difference. Firstly, students did not use Grammarly as an independent learning tool but in conjunction with an ALA. Secondly, this study compared Grammarly with the traditional ALC approach of using "Track Changes" in Microsoft Word. This study also had a larger sample size and investigated whether the software better benefits a particular student cohort based on language level, visa status and delivery method.

Research questions

(1) What are students' perceptions of Grammarly compared to perceptions of the traditional ALC grammar feedback approach?
(2) Does language level affect students' response to Grammarly?
(3) Does cohort (international, DNESB, local) affect students' response to Grammarly?
(4) Does delivery mode (face to face, online) affect students' response to Grammarly?

Methodology

Ethical clearance

Ethical clearance for this project was given by the Human Research Ethics Committee at CQUniversity on 12 February 2016.

Setting

Data was collected over one university term between February 2016 and July 2016. Six researchers collected data: five on CQUniversity's Sydney campus and one on CQUniversity's Mackay campus.

Australasian Journal of Educational Technology, 2019, 35(1).

Participants

Researchers agree that a control group is integral for studies aiming to determine the efficacy of corrective feedback (Bitchener, 2008). Studies considering second language acquisition also need to be conducted under controlled experimental conditions using a control and treatment group (Ferris, 2010). As this study is driven by a consideration of best practice in corrective feedback and has a large second language cohort, an experimental group which received feedback from an ALA using Grammarly and a control group which did not use Grammarly were created.

Ninety-six students were involved in this study: 54 students were in the experimental group, and 42 were in the control group. Students in the experimental group were either enrolled in a degree at CQUniversity ($n = 48$) or preparing for university with the CQUniversity English Language Centre (ELC) ($n = 6$). Students in the control group were also either enrolled in a degree ($n = 33$) or in the ELC ($n = 9$).

Instruments

A mixed methods sequential explanatory design with follow-up variant was used (Creswell & Plano Clark, 2011). First, quantitative data was collected and analysed to reveal patterns in the students' responses to the program and identify deviations from those patterns for specific statements and cohorts. Qualitative recordings were designed to explain these patterns; to reveal why students responded in a particular way to the program. It was hoped that using both sets of data in this complementary way would provide greater insight into the program' performance and that this greater insight could be used to inform future use of the program as a feedback tool.

A student survey was, therefore, created and presented in 3 phases: personal details, quantitative responses and qualitative reactions. The survey was designed using the findings from the literature review and by adapting the research instrument employed by Caveleri and Dianati (2016). Three CQUniversity lecturers and Caveleri, co-author of the 2016 investigation into students' perceptions of Grammarly, reviewed the survey for relevance to the research questions and readability.

Personal details

The measures from the personal details phase were:

- Visa status: Students were asked to identify as local, DNESB or international. International students were subdivided into two groups: undergraduate and postgraduate students or those undertaking university preparatory courses in the ELC.
- Delivery method: Whether the students received feedback online or face-to face was recorded.
- Completion of language tests: Scores were recorded for respondents who had completed a language test such as the International English Language Test (IELTS), Pearson Test of English or Occupational English Test. Most had completed IELTS, and the scores for those who had not were converted into IELTS scores using an online test converter (Pearson Education Ltd, 2012).
- Previous assignments at the university: Students who had previously submitted an assignment were asked for the grade they had received on that assignment.

Quantitative phase

A survey using a 5-point Likert-scale measuring tool was completed by the students. Fifteen items were devised to test students' responses to the grammar feedback with scores for each statement ranging from a possible 1 (*strongly disagree*) to 5 (*strongly agree*). Items 1–4 were drawn from the literature review relating to students' beliefs about the importance of grammar feedback and their satisfaction with the grammar feedback received, both overall and in relation to the time accorded, amount provided and focus. All other statements were adapted from the survey designed by Caveleri and Dianati (2016). These aimed to determine the perceived usefulness of the grammar feedback by evaluating the ways in which the students believed the feedback was helpful to their writing and confidence in the short and long term. The ease with which the feedback could be assimilated into the writing was also considered, as was the propensity for this to happen.

Qualitative phase

Students answered three open questions:

- Q1: What did you like about the grammar feedback you received?
- Q2: What did you dislike about the grammar feedback you received?
- Q3: Any additional comments on the grammar feedback you received?

Thematic analysis was used to interpret the qualitative data. Initial themes were identified from the quantitative findings. These themes were coded and then categorised depending on whether they corroborated or refuted the quantitative findings. Additional themes, which did not arise from the statistical analyses, were then categorised from the qualitative comments using inductive analysis.

Procedure

All students had a Grammarly Premium account for the duration of the trial. Students in both the control and experimental groups received a maximum of 15 minutes grammar feedback.

Control group: face-to-face and online

ALAs identified grammar issues in the students' assignments. They discussed their findings and recorded their suggestions in the students' assignments using "Track Changes."

Experimental group: face-to-face

The ALA uploaded each student's assignment into Grammarly Premium. They then downloaded the detailed Grammarly report with feedback across the six categories: contextual spelling, grammar, punctuation, sentence structure, style and vocabulary enhancement. To make the error correction more manageable, the five most frequently occurring errors from the report were dealt with in descending order. Indirect error correction feedback was employed first with the ALA locating an underlined example of the error in the Grammarly report. To ascertain the student's current understanding of the error and encourage self- correction, the student was asked if the underlined section was accurate or not. If the student deemed the section inaccurate, they were asked to write it a different way. If an appropriate correction was made, an extra example of the error was reviewed to confirm competency with the language item. If the ALA felt that further confirmation of understanding was required, then additional error correction checks were employed using the most appropriate technique for the language point. These techniques vary; for example, concept checking questions are used to check meaning and function, visuals such as timelines are used to check verb tenses and highlighting can clarify errors in form. All techniques, however, had the same objective: to encourage the students to think about their choices, articulate their choices and not just passively accept the program's suggestions.

If the student was unsure or unable to articulate their answer, then direct feedback was employed. The student was shown the suggested correction in Grammarly and asked whether this suggestion was appropriate. If the student answered correctly, then another example of this error in the student's work was reviewed. However, if the student was unsure whether the correction was appropriate or not, they were asked to read the Grammarly information card and examples. The ALA used the explanation and examples on the card to unpack the grammar rule and illustrate its use. If the Grammarly information was not clear or detailed enough, the ALA provided additional instruction by modelling and providing examples themselves. In this way gaps in the Grammarly feedback were filled by the ALA, so the students were provided with the instruction and clarification they needed. In addition to clarifying language issues, identifying erroneous and unnecessary suggestions, this procedure also encouraged students to ask questions about the Grammarly feedback process and provided a model of use for the students to replicate.

Experimental group: online

A Grammarly Premium account was created for the student and the student's assignment uploaded. The five most frequently committed errors were identified and examples in the assignment located. Any errors that the program had made in relation to these suggestions were deleted. The assignment was uploaded into

Australasian Journal of Educational Technology, 2019, 35(1).

PDF Annotator, and any errors that Grammarly had missed were highlighted with feedback from the ALA. An email was sent to the student with details of their Grammarly account, a PDF copy of their annotated assignment and advice about which errors to target.

To encourage survey completion, on campus students recorded their responses in a hardcopy survey and the data was manually input into SurveyMonkey. Online students completed the same survey in an online form in SurveyMonkey.

Analyses

The dependent variables were the level of agreement with the 15 statements about the grammar feedback. Independent variables were whether the student received feedback via Grammarly or not; their English proficiency (IELTS score and previous grades; cohort (local students, DNESB, international students, and international students from the ELC) and delivery mode (face-to-face or online). The latter variables were included to determine which specific aspects of Grammarly (or non-Grammarly) appealed to which types of students. Comparisons between those who received feedback via Grammarly and those who did not were conducted using inferential statistics, which are independent samples t-tests. For some statements, variances between the groups were significantly different, violating an assumption of independent samples t-tests. In those cases the more robust Welch t-test was used instead, indicated by degrees of freedom with decimal places. Effect sizes are reported for significant results using Cohen's *d*.

Relationships between reported agreement with statements and English proficiency, as well as previous grades, were conducted using non-parametric (Spearman's) correlations as the agreement with statements variables were skewed, and because the previous grades variable was ordinal in nature. Comparisons between students from different cohorts were conducted using ANOVA with post-hoc Tukey pairwise comparisons. Alpha was set at .05 throughout.

Results

Students' perceptions of Grammarly feedback as compared to perceptions of non-Grammarly feedback

Those who received ALA advice with Grammarly feedback were generally very positive about the experience, as indicated by the high mean in Table 1. In general, the respondents understood the importance of getting grammar feedback on assignments, believed that they had received enough assistance from Grammarly and found the suggestions useful and easy to understand. They believed that the feedback improved their assignment and grade. It also improved their confidence in the assignment and developed their confidence and language skills beyond the assignment. The only mean below 4.4 was for the statement "I made additional grammar corrections not recommended by the ALC") (M = 3.72, SD = 1.47). This suggests that although the students still agreed with the statement, there was some hesitance to do more than that recommended by the ALA. Qualitative responses largely substantiate the positive findings illustrated in Table 1, with six students stating that they would like the university to provide the software on a permanent basis, 11 students claiming that the program helped them, and 14 students describing the feedback using positive verbs and adjectives; for example, "I liked it. Really good. Often use it. If I could give 6/5, I would. Really happy."

Australasian Journal of Educational Technology, 2019, 35(1).

Table 1
Mean and SD ratings for students' statements about grammar feedback

Statement	Grammarly (*n* = 54)		Non-Grammarly (*n* = 42)		Inferential statistics			
	M	*SD*	*M*	*SD*	*t*	*df*	*p*	*d*
It was important for me to get grammar feedback on my assignment	4.93	.26	4.93	.26	-.05	94	.961	-
The advisor spent enough time on grammar feedback to help me improve my assignment	4.74	.62	4.38	.96	2.11	66.43	.039	.45
I got enough grammar feedback on my assignment	4.57	.69	4.33	.87	1.51	94	.135	-
The grammar feedback really focused on my main errors – i.e. those that were made a lot	4.54	.77	4.14	.93	2.23	79.24	.029	.47
It was easy to understand the errors because the explanations given were clear	4.65	.76	4.36	1.01	1.56	73.83	.123	-
I got a lot of useful suggestions about how to improve my grammar in the assignment	4.67	.70	4.29	.97	2.15	71.86	.035	.45
It was easy to make grammatical changes to my work using the feedback	4.65	.68	4.20	.98	2.54	67.55	.014	.53
I made the grammar corrections the ALC advisor recommended during the consultation	4.54	.86	4.38	.94	.85	94	.399	-
I made additional grammar corrections not recommended by the ALC	3.72	1.47	3.12	1.33	2.08	94	.041	.43
The grammar feedback improved my assignment	4.80	.49	4.55	.92	1.59	59.06	.117	-
The grammar feedback made me feel more confident about handing the assignment in	4.76	.58	4.38	.99	2.21	62.56	.031	.47
The grammar feedback improved my grade	4.45	.85	4.17	1.01	1.50	93	.136	-
The grammar feedback developed my language long term (not just for this assignment) as I could understand the grammatical rules more	4.48	.72	4.07	1.07	2.14	68.48	.036	.45
The grammar feedback developed my confidence in my language use long term (not just for this assignment) as I could understand the grammatical rules more	4.46	.75	3.98	1.09	2.47	69.02	.016	.51
I was satisfied with the grammar advice I received	4.75	.55	4.43	.83	2.19	68.09	.032	.45

Note: df = degrees of freedom, with decimal places indicating Welch t-tests was used. *d* refers to Cohen's *d* (effect size).

However, whilst students were largely positive about the program, they did identify six areas of concern. The main area of concern was the accuracy of the feedback, with nine students citing it as a problem: "Not all grammar errors were valid or correct." Students also stated that the pre-set to American English caused the program to notice invalid spelling errors. They identified issues with passive voice, complex sentences and vocabulary choices, which Grammarly suggested students changed based on style and not accuracy.

Students also noted that the program missed errors and that the software suggestions were at times difficult to understand. Further concerns about the software related to technical problems and included issues with uploading documents and compatibility. For two students, the program's faults clearly outweighed the positives and they explicitly expressed a preference for the traditional ALC feedback method, with one of the students writing, "I liked getting feedback via the Word document. Rather than Grammarly. I hated that."

Table 1 also provides the mean and standard deviations for the 15 statements for students who received feedback via traditional, non-Grammarly feedback mechanisms. In general, the pattern of results was similar, in that the non-Grammarly students indicated that they found the feedback useful, that they received enough, and that it assisted their grammar, grades and confidence. However, the means for Grammarly students were significantly higher for some of these statements. Grammarly students were significantly more likely to state that an ALA spent enough time on their grammar feedback to improve their assignment compared to non-Grammarly students, and that the feedback via Grammarly focused on errors that they make a lot. Grammarly students were significantly more likely to agree that they received a lot of useful feedback via Grammarly, and that it was significantly easier to make grammatical changes based on this feedback. Grammarly students reported being significantly more likely to make additional corrections that were not recommended by Grammarly, and were significantly more confident about handing the assignment in. Grammarly students agreed significantly more strongly that the feedback had developed their language, as well as their confidence in their language, outside of the assignment, compared to non-Grammarly students. In general, Grammarly students were significantly more satisfied with the grammar advice that they received compared to non-Grammarly students.

Students' perceptions of Grammarly based on language level

IELTS (and IELTS indexed) scores were captured for 35 Grammarly students and 25 non-Grammarly students, while previous grades were captured for 36 Grammarly students and 25 non-Grammarly students. For Grammarly students, those with lower IELTS scores tended to be more satisfied with the grammar advice they had received (Spearman's rho = -.41, $p < .05$), and those with higher previous grades were significantly more likely to agree that it was easy to make grammatical changes (Spearman's rho = .38, $p < .05$). These results were not statistically significant for non-Grammarly students, and no significant difference was observed between the results. For non-Grammarly students, those with lower IELTS scores were significantly more likely to agree that the grammar feedback improved their grade (Spearman's rho = -.40, $p < .05$). No other correlations were statistically significant, although it is important to acknowledge that due to missing data, the small sample and the conservative nature of Spearman's correlations, these results are most likely underpowered, and thus conclusions from these results are tentative.

Qualitative data supports the finding that graduate students with a lower IELTS score (6 or less) were satisfied with the Grammarly feedback, because students with a score of 6 or less were the group least likely to post negative comments about the program. Students from the ELC, however, who had the lowest language proficiency of all groups, were the most vocal about Grammarly's limitations, with all six students registering concerns. Four of these related to the program's inaccuracies, whilst the other two stated that it "doesn't really make sense" and that it "gives wrong suggestions." These findings suggest that Grammarly can be used with students of different language levels but a different approach may need to be taken when using it with ELC students.

Students' perceptions based on visa status

As the Grammarly and non-Grammarly students were found to differ in their agreement with some statements, differences between cohorts were conducted as a factorial ANOVA, with Grammarly (2 levels: yes vs no), Cohort (4 levels: local, DNESB, international and international ELC) as the main effects, and an interaction effect for each statement. These analyses should be regarded with caution for two reasons. Firstly, some cells were constant for some statements (i.e., all respondents selected the same answer); and secondly, cell sizes were relatively small for this type of analysis. Where significant effects were found, post-hoc Tukey pairwise comparisons were conducted.

Differences between the cohorts were observed for three statements. International students were generally significantly less satisfied with the grammar feedback they received and significantly less likely than

Australasian Journal of Educational Technology, 2019, 35(1).

DNESB students to agree that the grammar feedback developed their confidence in their language long term. Local students were significantly less likely to state that the grammar feedback had developed their language long term compared to DNESB and international ELC students. These results, however, are main effects for cohort, which means that they are averaged across whether or not they received Grammarly feedback and there was no significant interaction for any of these statements, indicating that the results did not differ significantly depending on whether or not they received Grammarly feedback.

Qualitative data supports the finding that the DNESB cohort responded positively to Grammarly. Question 2 asked students to identify what they disliked about the program, and the DNESB cohort posted the fewest negative responses to this question (21% reply rate compared to 66% by both local and international students). This group was also the only one to explicitly identify that Grammarly was useful for them, with one student stating, "I think it is a great tool to assist non-English speaking background students." The high response rate to question 2 from international students, coupled with the statistical finding that they were less satisfied with the grammar feedback, suggests that, as with the ELC students, a more tailored approach may need to be taken when using Grammarly with them.

Student responses by delivery mode

Comparisons were also made between those who had received feedback online compared to on campus. Twenty-seven of the 54 Grammarly students and 13 of the 42 non-Grammarly students received feedback online, while the remainder received feedback on campus. No significant differences were found between those who had received feedback online compared to on campus for any of the statements (largest $F(1,90) = 2.58$, $p = .112$). Interaction effects were also calculated comparing differences between online vs on campus feedback for Grammarly and non-Grammarly students. No significant interaction effects were found (largest $F(1,90) = 3.32$, $p = .072$.

Discussion

The amount of time and detail accorded to grammar feedback was identified in the literature review as a potential cause of dissatisfaction with students visiting ALCs. Quantitative data showed that Grammarly students were significantly more likely to believe that enough time had been spent on grammar feedback. Qualitative data supported these findings as the issue of time was not mentioned by any of the students who received feedback using Grammarly, whereas four students who received feedback in the traditional manner bemoaned the lack of time accorded to grammar feedback. An example of this is from a student who wrote that they "would like more time spent on grammar and language."

Students who received Grammarly advice were also significantly more likely to state that they had received a lot of useful feedback. There were no responses from students in the Grammarly group complaining that they had not received enough grammar feedback. Six students in the non-Grammarly group did, however, express dissatisfaction.

> I haven't received enough grammar feedback. I want to fix my grammar but the advisor didn't pay attention to my grammar mistakes.

The tendency amongst ALAs to prioritise macro skills over grammar was flagged as a potential issue in the literature review. This issue was not referred to by any of the students in the Grammarly group, but students in the control group did believe that ALAs sacrificed grammar feedback to focus on other issues.

> We could not correct the whole assignment as to grammar issues since we should consider and correct other factors such as structure.

The shortfall of grammar feedback was identified in the literature review as a potential hindrance to language development, and some students in the control group did believe that their long-term development was inhibited.

> They just fix the grammar but I don't understand why they did it that way. It's just for the assignment not for overall English development.

Australasian Journal of Educational Technology, 2019, 35(1).

Students' responses to Grammarly also identified features of effective written corrective feedback outlined in the literature review and support the proposition that the program can reconcile different theoretical approaches. Firstly, Grammarly does not just highlight one type of error as students identified a range: "repeated", "serious", "basic"; "errors that I did not know I had made" and mistakes that "I had overlooked." The issue of how much to correct was addressed, with some students expressing satisfaction that the feedback was "detailed", "thorough" and "line by line", whereas other students commented on the targeted nature of the feedback and the fact that it identified their "areas of weakness" and highlighted specific issues they had; for example, article use, sentence structure, commas, spelling, prepositions and word choices.

In addition to identifying positive feedback practices from the literature review, students' responses revealed that Grammarly adheres to other best practice elements in written corrective feedback. Students liked the fact that the feedback targeted their errors. This personalisation is a key factor in feedback assimilation, with students more likely to value feedback that takes into account their individual needs (Hyland & Hyland, 2006). Several students commented on the quickness of the program stating that the feedback was "prompt." Such timely feedback is identified by McGregor, Merchant, and Butler (2008) as crucial to its usefulness as advice is more likely to be acted on whilst fresh in the student's mind. Channock et al. (2009) further qualify the link between time and feedback by stating that students reap the most benefits if they make the changes to the assignment after the review of the first draft, which is what the students in this study did. Several students also made the link between the Grammarly feedback and their grade, which according to Bacon and Anderson (2004) is necessary to stimulate improvements in grammar and punctuation.

Acting on the literature review findings, ALAs were positioned as an intermediary between the Grammarly feedback and the student. This approach reduced negative responses about the program. Firstly, the ALAs managed students' expectations about the feedback by making them explicitly aware that it was not infallible. This understanding of the program's limitations is reflected in the following student's comment: "Some places it could not point out so it is not 100% [sic]. Ok I know it is hard and it's not perfect." Secondly, ALAs helped make the feedback more reliable by identifying and deleting erroneous suggestions and misrepresentations. Inaccuracies are acknowledged to be an issue with AFPs (Hoang & Kunnan, 2013; Liu & Kunnan, 2016) but ALAs can make them less of an issue by reducing their number. They can also use Grammarly's errors as a learning tool so that the student understands why they are inaccurate: "The Grammarly suggestion was wrong but the instructor did a pretty awesome job to clear why [sic]." Finally, and most importantly, ALAs can provide additional instruction and explanation about the students' errors. This is achieved by explaining the metalanguage used by the program, by grading explanations for lower level students, or by providing alternative examples or models. Irrespective of the technique, the outcome remains the same: a clearer explanation of the problem so that students have a better understanding of how they can improve their writing. None of the Grammarly students posted negatively about the explanations they received, whereas four non-Grammarly students stated that they needed "clearer explanations." Similarly, no Grammarly students asked for additional support, whereas non-Grammarly students mentioned a need for further clarification through additional grammar classes, sessions, materials and websites. It is therefore recommended that Grammarly continue to be used in conjunction with an ALA until comparable results are collated, proving its effectiveness as an independent tool.

Conclusion

This study shows that students using the ALC want assistance with their grammar to improve their assignments and develop their long-term writing performance. Grammarly was chosen as a means of providing the grammar feedback students want without compromising the development of their higher order skills. A systematic approach to using Grammarly was devised with supplementary ALA advice provided if students did not understand the program's response to the language item. Results between the students who received Grammarly and non-Grammarly advice were compared. Students who received grammar advice from ALAs in the traditional manner were largely satisfied with the feedback they received, but students who received feedback from Grammarly in conjunction with an ALA were more satisfied in relation to the amount of time spent, the amount of feedback received, the short-term benefits to their grade, the long-term benefits to their writing, their confidence in their writing and their ability to proofread independently. Students also identified several features of best practice in written corrective feedback in the Grammarly approach. This included the speed of response, range of errors addressed and

the targeted, personalised nature of the feedback. Concerns were, however, expressed with regards to the program's accuracy, its propensity to miss errors and correct those that were not wrong and technological glitches. These concerns would have been more prevalent had an ALA not engaged with the feedback first, as the ALA was able to manage students' expectations, reduce erroneous feedback and use the inaccuracies as a learning tool. No significant differences were identified between delivery mode, suggesting that the program can be used with online and on-campus students. Similarly, the results showed that Grammarly can be used with students of different language levels and visa status, but greater discretion or a revised approach may be needed with ELC and international students. Based on these findings, it is therefore recommended that Grammarly be added to the ALC's toolbox, but it should continue to be used in conjunction with an ALA.

Limitations to the study

Statistical power issues exist within this study because the overall sample size is small. A larger sample might find significant differences, so any future research into student perceptions' on Grammarly should use a larger sample. A number of student variables, which could have impacted on responses, were also not considered; for example, the length of time the students had spent learning English and the length of the assignment. Data connected to relational variables was also not collected. Such variables as the student-ALA relationship and the student's preferences in an ALA (e.g., gender) could have affected the students' responses. Data relating to language level used IELTS or IELTS equivalent writing scores, but these are not always a reliable indication of language performance. Asking students to self-report on these scores and university grades is also problematic as students may not report accurately and may enhance performance levels.

Future research

This study focused on students' perceptions of Grammarly. To build a more comprehensive picture of Grammarly as a feedback tool additional research is needed. Firstly, there is a need to consider ALA perceptions of the program and compare these findings with ALA attitudes to the traditional feedback approach. Secondly, as inaccuracy was cited as an issue with the program, further investigations are needed to identify the errors which Grammarly most frequently misses or misidentifies. This will help inform future feedback sessions and render them more useful.

It is also necessary to determine the degree to which Grammarly enhances students' grammatical accuracy. Short-term improvements can be identified by comparing students' grammatical accuracy in the first draft of an assignment (pre-Grammarly feedback) with their grammatical accuracy in their final draft (post-Grammarly feedback). Similar analysis should also be undertaken on the assignments of students who received feedback in the traditional manner so that it can be seen whether accuracy in the short term is more improved in students using Grammarly than in those receiving conventional ALC advice. Long-term improvement in students' grammatical accuracy could be determined through longitudinal studies in which students submit an assignment (post-Grammarly feedback) each term during their university studies. The percentage accuracy of these assignments could be compared with their first submission to determine whether improvements have been made in relation to high-frequency errors and overall written accuracy. These students could also be asked to take a grammar test each term in relation to their high-frequency errors to determine whether their accuracy using the item has improved when the program is not being used.

References

Bacon, D. R., & Anderson, E. S. (2004). Assessing and enhancing the basic writing skills of marketing students. *Business Communication Quarterly, 67*(4), 443–454. https://doi.org/10.1177/1080569904271083

Bitchener, J. (2008). Evidence in support of written corrective feedback. *Journal of Second Language Writing, 17*(2), 102–118. http://dx.doi.org/10.1016/j.jslw.2007.11.004

Burston, J. (2008). BonPatron: An online spelling, grammar, and expression checker. *CALICO Journal, 25*(2), 337–347. Retrieved from http://journals.equinoxpub.com/index.php/CALICO/article/view/23116

Caveleri, M., & Dianati, S. (2016). You want me to check your grammar again? The usefulness of an online grammar checker as perceived by students, *Journal of Academic Language and Learning, 10*(1), 223–236. Retrieved from http://www.journal.aall.org.au/index.php/jall/article/view/393/246

Channock, K., D'Cruz, C., & Bisset, D. (2009). Would you like grammar with that? *Journal of Academic Language and Learning, 3*(2), 1–12. Retrieved from http://www.journal.aall.org.au/index.php/jall/article/view/70/63

Chen, C. E., & Cheng, W. (2006, May). *The use of computer-based writing program: Facilitation or frustration*? Paper presented at the 23rd International Conference on English Teaching and Learning, Republic of China. Retrieved from ERIC database. (ED492078)

Chen, C. E., & Cheng, W. (2008). Beyond the design of automated writing evaluation: Pedagogical practices and perceived learning effectiveness in EFL writing classes. *Language, Learning & Technology*, *12*(2), 94–112. Retrieved from http://llt.msu.edu/vol12num2/chencheng.pdf

Choi, J., & Lee, Y. (2010). The use of feedback in the ESL writing class integrating automated essay scoring (AES). In D. Gibson & B. Dodge (Eds.), *Proceedings of Society for Information Technology & Teacher Education International Conference* (pp. 3008–3012). Chesapeake, VA: Association for the Advancement of Computing in Education. Retrieved from https://www.learntechlib.org/p/33826/

Cocks, T., & Stokes, J. (2013). Policy into practice: A case study of widening participation in Australian higher education. *Widening Participation and Lifelong Learning, 15*(1), 22–38. http://dx.doi.org/10.5456/WPLL.15.1.22

Creswell, J. W., & Plano Clark, V. L. (2011). *Designing and conducting mixed methods research*. Thousand Oaks, CA: Sage.

Department of Education and Training. (n.d,). *Non-English speaking background*. Canberra: Australian Government. Retrieved from http://heimshelp.education.gov.au/sites/heimshelp/resources/glossary/pages/glossaryterm?title=Non-English%20speaking%20background

Dikli, S. (2010). The nature of automated essay feedback. *CALICO Journal, 28*(1) 99–134. Retrieved from https://calico.org/page.php?id=5

Dikli, S., & Bleyle, S. (2014). Automated essay scoring feedback for second language writers: How does it compare to instructor feedback? *Assessing Writing, 22*, 1–17. http://dx.doi.org/10.1016/j.asw.2014.03.006

Ferris, D. R. (2010). Second language writing research and written corrective feedback in SLA: Intersections and practical applications. *Studies in Second Language Acquisition, 32*(2), 181–201. https://doi.org/10.1017/S0272263109990490

Grammarly Inc. (2017). *Grammarly Premium: Keep it clean and write your heart out*. Retrieved from https://www.grammarly.com/premium

Harris, A. (2016). Integrating written communication skills: Working towards a whole of course approach. *Teaching in Higher Education, 21*(3), 287–300. http://dx.doi.org/10.1080/13562517.2016.1138456

Hoang, G. T. L., & Kunnan, A. J. (2016). Automated essay evaluation for English language learners: A case study of MY Access. *Language Assessment Quarterly, 13*(4), 359–376. http://dx.doi.org/10.1080/15434303.2016.1230121

Hoon, T. B. (2006). Online automated essay assessment: Potentials for writing development. In A. Treloar & A. Ellis (Eds.), *Making a difference with Web technologies. Proceedings of AusWeb 2006, The Twelfth Australasian World Wide Web Conference* (pp. 230–234). Lismore: Southern Cross University. Retrieved November 1, 2016, from http://ausweb.scu.edu.au/aw06/papers/refereed/tan3/paper.html

Huang, L. S. (2011). Are we having the effect we want? Implementing outcomes assessment in an academic English language-support unit using a multi-component approach. *Writing Program Administration, 35*(1), 1–11. Retrieved from http://wpacouncil.org/archives/35n1/35n1huang.pdf

Hyland, K., & Hyland, F. (2006). Interpersonal aspects of response: constructing and interpreting teacher written feedback. In K. Hyland & F. Hyland (Eds.), *Feedback in second language writing: Contexts and issues* (pp. 206–224). New York, NY: Cambridge University Press. https://doi.org/10.1017/CBO9781139524742.013

Jamalinesari, A., Rahimi, F., Gowhary, H., & Azizifar, A. (2015). The effects of teacher-written direct vs. indirect feedback on students' writing. *Procedia - Social and Behavioral Sciences, 192*, 116–123. http://dx.doi.org/10.1016/j.sbspro.2015.06.018

Japos, V. (2013). Effectiveness of coaching interventions using Grammarly software and detection software in reducing grammatical errors and plagiarism of undergraduate researchers. *JPAIR Institutional Research, 1*(1), 97–109. http://dx.doi.org/10.7719/irj.v1i1.207

Knoch, U., Rouhshad, A., & Storch, N. (2014). Does the writing of undergraduate ESL students develop after one year of study in an English-medium university? *Assessing Writing, 21*, 1–17. http://dx.doi.org/10.1016/j.asw.2014.01.001

Liao, H. (2016). Enhancing the grammatical accuracy of EFL writing by using an AWE-assisted process approach. *System, 62*, 77–92. http://dx.doi.org/10.1016/j.system.2016.02.007

Linville, C. (2009). Editing line by line. In S. Bruce & B. Rafoth (Eds.), *ESL writers: A guide for writing center tutors* (2nd ed., pp. 116–131). Portsmouth, NH: Heinemann.

Liu, S., & Kunnan, A. J. (2016). Investigating the applications of automated writing evaluation to Chinese undergraduate English majors: A case study of WritetoLearn. *CALICO Journal*, *33(1)*, 71–91. http://dx.doi.org/10.1558/cj.v33i1.26380

McGregor, K. M., Merchant, A. R., & Butler, M. (2008). Making feedback more immediate. *CAL-laborate International, 16*(1) 8–11. Retrieved from https://openjournals.library.sydney.edu.au/index.php/CAL/article/view/6020/6671

McNaught, K., & Shaw, G. (2016). Preparing undergraduate students to be successful writers: Exploring the spelling, grammar and punctuation skills of students identified as potentially 'at risk'. *Journal of Academic Language and Learning, 10*(2), 11–19. Retrieved from http://www.journal.aall.org.au/index.php/jall/article/view/406/253

Murray, N. (2011). Ten 'Good Practice Principles' … ten key questions: Considerations in addressing the English language needs of higher education students. *Higher Education Research & Development, 31*(2), 233–246. http://dx.doi.org/10.1080/07294360.2011.555389

Pearson Education Ltd. (2012). *PTE academic score guide, version 4*. Retrieved from http://pearsonpte.com/wp-content/uploads/2014/07/PTEA_Score_Guide.pdf

Qassemzadeh, A., & Soleimani, H. (2016). The impact of feedback provision by Grammarly software and teachers on learning passive structures by Iranian EFL Learners. *Theory and Practice in Language Studies, 6*(9), 1884–1894. http://dx.doi.org/10.17507/tpls.0609.23

Ramineni, C., & Williamson, D. M. (2012). Automated essay scoring: Psychometric guidelines and practices. *Assessing Writing*, *18*(1), 25–39. http://dx.doi.org/10.1016/j.asw.2012.10.004

Reis, C., & Huijser, H. (2016). Correcting tool or learning tool? Student perceptions of an online essay writing support tool at Xi'an Jiaotong-Liverpool University. In S. Barker, S. Dawson, A. Pardo, & C. Colvin (Eds.), *Show Me The Learning. Proceedings of ASCILITE 2016, 33rd International Conference of Innovation, Practice and Research in the Use of Educational Technologies in Tertiary Education* (pp. 529–533). Adelaide: University of South Australia). Retrieved from http://2016conference.ascilite.org/wp-content/uploads/ascilite2016_reis_concise.pdf

Salehfard, E., & Bagheri, M. S. (2014). Preferences of EFL teachers and students towards corrective feedback in writing. *Modern Journal of Language Teaching Methods, 4(*4), 308–319. Retrieved from http://mjltm.org/en/archive.php?rid=18

Scouller, K., Bonanno, H., Smith, L., & Krass, I. (2008). Student experience and tertiary expectations: Factors predicting academic literacy amongst first year pharmacy students. *Studies in Higher Education, 33*(2),167–178. http://dx.doi.org/10.1080/03075070801916047

Tynan, L., & Johns, K. (2015). Piloting the post-entry language assessment: Outcomes from a new system for supporting research candidates with English as an additional language. *Quality in Higher Education, 21*(1), 66–78. http://dx.doi.org/10.1080/13538322.2015.1049442

Wang, Y. J., Shang H. F., & Briody, P. (2013). Exploring the impact of using automated writing evaluation in English as a foreign language university students' writing. *Computer Assisted Language Learning, 26*(3), 234–257. http://dx.doi.org/10.1080/09588221.2012.655300

Winder, R., Kathpalia, S., & Koo, S. (2016). Writing centre tutoring sessions: Addressing students' concerns. *Educational Studies, 42*(4), 323–339. http://dx.doi.org/10.1080/03055698.2016.1193476

Australasian Journal of Educational Technology, 2019, 35(1).

Corresponding author: Ruth O'Neill, r.oneill@cqu.edu.au

Australasian Journal of Educational Technology © 2019.

Please cite as: O'Neill, R., & Russell, A. (2019). Stop! Grammar time: University students' perceptions of the automated feedback program Grammarly. *Australasian Journal of Educational Technology, 35*(1), 42-56. https://doi.org/10.14742/ajet.3795

中文译稿

停!语法时间:大学生对自动反馈程序 Grammarly 的看法

摘 要

越来越多的大学利用网络技术来帮助学生。Grammarly 标榜自己是世界上最准确的在线语法检查器，但很少有人研究它在大学作为反馈工具的表现。本文报道了一项研究，调查了学生对结合使用 Grammarly 和学术学习顾问的建议的看法。本研究采用混合法研究顺序解释性设计，比较了一组学生对 Grammarly 反馈的反应（n = 54）和另一组学生对中央昆士兰大学学术学习中心的传统非自动语法反馈的反应（n = 42）。结果显示，与没有使用 Grammarly 的学生相比，收到 Grammarly 反馈的学生对 15 个调查项目中的 9 个做出了更积极的回应，并且对他们收到的语法建议明显更满意。两者的传递模式之间没有显著差异，这表明学术学习顾问可以有效地使用 Grammarly 为国内外学生提供语法支持。然而，本文推荐该程序与学术学习顾问的建议结合使用，因为该程序目前还不够准确，不能单独使用。

1 引言

澳大利亚高等教育系统中的许多学生由于语言问题难以满足学术需求(Harris，2016)。尤其是国际学生往往存在更多与准确语法、标点符号和拼写相关的低级错误（Tynan&Johns, 2015）。传统上，国内学生在语法准确性方面表现优于国外学生（Scouller, Bonanno, Smith,& Krass, 2008）。然而，由于国内学生人数的变化，这一群体的成员在句子结构和形态方面也存在困难（Cocks & Stokes, 2013）。因此，不同群体的学生需要语法反馈才能取得学术成功。一般认为，学习支持单位，即学术学习中心（ALC）承担提供这种反馈的责任。但是，ALC 对其角色的认知与学生期望可能存在冲突，导致学生在完成 ALC 课程后对其接收到的语法反馈数量和质量不满意。自动反馈程序被认为是调和这种矛盾的工具。Grammarly 就是这样一个程序，它被中央昆士兰大学挑选为潜在语法反馈工具，因为它能够调和关于书面纠正性反

馈的不同理论。在 Caveleri 和 Dianati（2016）的工作基础上，本研究的主要目的是分析学生对 Grammarly 作为反馈工具的满意度，并与传统上由 ALC 在中央昆士兰大学提供的非自动反馈进行比较。它还旨在确定该方案是否更适合基于语言水平、签证状态和授课方式的特定学生群体。

2 文献综述

2.1 语法和语法输入的需求

学生需要使用标准语法并遵守语法惯例，才能在大学的学术写作中取得成功（Caveleri & Dianati，2016）。然而，如果没有输入，学生就无法提高他们的语言技能。国际学生需要外部输入来发展他们的语言技能（Murray，2011），包括学术写作所需的语法教学（Knoch, Rouhshad, & Storch, 2014）。不同的国内学生群体也需要在标点符号、语法和拼写方面进行实际干预（McNaught & Shaw, 2016）。这包括本地学生（在澳大利亚出生），他们没有接受过作为学校课程一部分的正式语法培训，或通过非传统方式（如授权项目）进入高等教育（Cocks& Stokes, 2013）。它还包括国内非英语背景的学生（DNESB），由教育和培训部（n.d.）定义为国内学生，他们在开始上大学之前到达澳大利亚不到 10 年，平时说英语以外的语言。

2.2 学术学习顾问和学生对语法干预的态度

在大学环境中，语法教学传统上是由 ALCs（Channock, D'Cruz, & Bisset, 2009）和在那里工作的学术学习顾问（ALA）负责的。中央昆士兰大学的 ALA 通过在线和面对面的咨询提供作业反馈。这些咨询是 60-90 分钟，但 ALA 将其中的一小部分时间用于语法反馈，因为它很耗时（Dikli 和 Bleyle，2014）。另一方面，学生们重视语法反馈，并认为能否立即纠正语法是他们最关心的问题（Huang，2011）。有人指出，这种优先事项的冲突和对语法反馈的时间限制使一些学生感到不安，因为他们觉得他们没有得到他们在作业和长期发展方面所需要的语法反馈量。他们认为，如果 ALC 不帮助他们发展语法，他们就没有其他办法，因为讲师只关注内容，而不

是写作技巧的发展（Bacon & Anderson, 2004）。

2.3 自动反馈程序

自动反馈程序可以缓解 ALA 和学生在语法反馈方面的态度差异。这些程序能够提供学生需要和想要的语法反馈，同时不影响 ALA 对宏观技能发展的优先考虑。虽然在线语法检查器的研究较为有限（Caveleri & Dianati, 2016），但对自动论文评分系统的研究已经展开，这些系统提供语法反馈作为其功能的一部分（Ramineni & Williamson, 2012）。学生对这些系统所提供的语法反馈的反应不一。使用 CorrectEnglish 的学生认为改善语法和书写的准确性是该程序的一个积极方面（Wang, Shang, & Briody, 2013），而使用 My Editor 的学生也同样认为该程序帮助他们改善了语法和拼写方面的问题（Hoon, 2006）。然而，在 Chen 和 Cheng（2006）的研究中，参与者对他们收到的语法反馈基本没有印象。在这项研究中，学生的不满突出了教师在反馈过程中的作用的重要性，因为学生的消极反应是由教师对该计划的态度造成的。一位老师不熟悉该工具的编辑功能，另一位老师对其印象不深，这意味着两位老师都没有为学生传达使用该工具的最佳方式。

关于从自动反馈程序中获益最多的学生语言水平，也有矛盾的发现。Chen 和 Cheng（2008）的研究发现，高级学生在使用自动反馈时对准确性的关注存在局限性，并建议将自动反馈应用于初级和中级学习者。然而，Dikli（2010）对此提出反驳，认为自动反馈会让英语水平较低的学生感到不知所措。Liao（2016）也认为，自动反馈更适合表现较好的写作者，因为他们拥有长期发展语法准确性所需的认知策略。在对 Grammarly 的研究中，Caveleri 和 Dianati（2016）也表示，该程序可能对能力较强的写作者有利，因为学生需要有足够的语法理解能力，才能过滤那些不正确的建议，或者他们认为不必要的、没有用的建议。

关于自动反馈程序是更适合国内还是国外学生也同样有争议。Hoang 和 Kunnan（2016）提出，反馈系统在传统上并不是为非英语母语的学生写作者设计的，国际学生对目标语言的理解不够，无法解释和应用程序提出的建议。

从学生的角度来看，自动反馈的另一个缺点是缺乏人际互动（Dikli，2010）。根据 Wang 等人（2013 年）的说法，这对国际学生和经验不足的写作者来说是一个

突出的问题，他们需要指导、示范和练习来发展他们的书面英语。缺乏对话对写作能力也是不利的，因为这会降低学生的积极性，使他们的写作更加机械和不真实(Wang et al., 2013)。为了解决这些问题并使自动反馈变得更加有用，Hoang 和 Kunnan(2016)建议采用组合式教学法，即写作教师作为学生的工作和反馈之间的中介，促进对话和互动。研究表明，以这种方式提供的反馈是最有效的（Choi & Lee, 2010），因为教师可以提供关于如何最好使用程序的指导（Burston, 2008），提供错误的纠正（Hoang & Kunnan, 2016），并过滤程序所犯的错误。

2.4 Grammarly

2016 年，中央昆士兰大学的 ALC 在 ALA 的建议下，试行使用 Grammarly Premium。Grammarly Premium 是一个自动校对系统，可以识别与 250 条语法规则有关的错误（Grammarly Inc., 2017）。用户上传他们的作业并收到两个分数，第一个分数是准确率，第二个分数是程序识别的错误总数。Grammarly 根据六个类别来组织错误：上下文拼写、语法、标点符号、句子结构、风格和词汇增强。这些具体类别中的错误数量被计算并呈现出来（见图 1）。

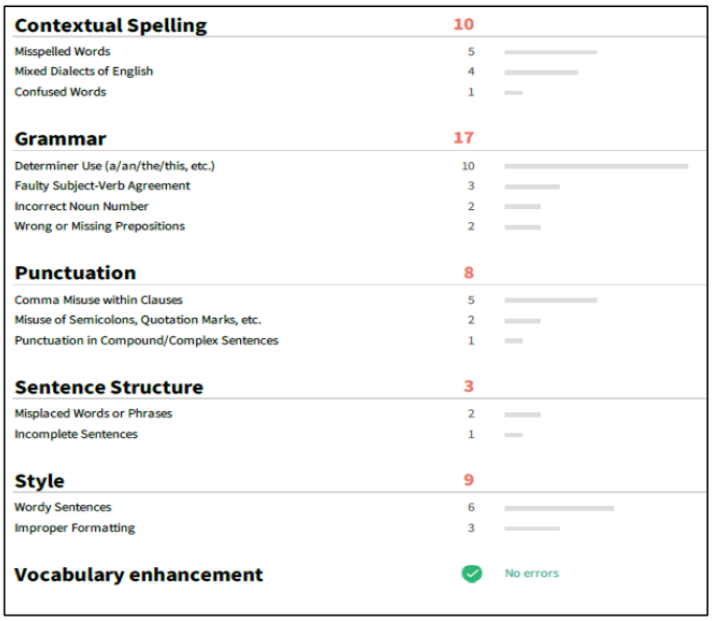


图 1：Grammarly 报告中的错误分类

2.5 反馈问题以及将 Grammarly 作为潜在的解决方案

选择 Grammarly 是因为它有可能调和与书面纠正性反馈有关的对比性理论。关于直接反馈和间接反馈的相对优势，已经有很多文章进行阐述。

支持间接反馈的人（Jamalinesari, Rahimi, Gowhary, & Azizifar, 2015）指出，错误应该只被强调，让学生独立审查和纠正，而直接反馈的支持者（Bitchener, 2008）认为，明确的纠正需要由教师提供。

Grammarly 调和了这些不同的理论方法，为教师提供了使用这两种方法的机会。在屏幕的左侧，可以使用下划线的错误来提供间接反馈。在屏幕的右侧，可以使用更正和解释卡来提供直接反馈（见图 2）。

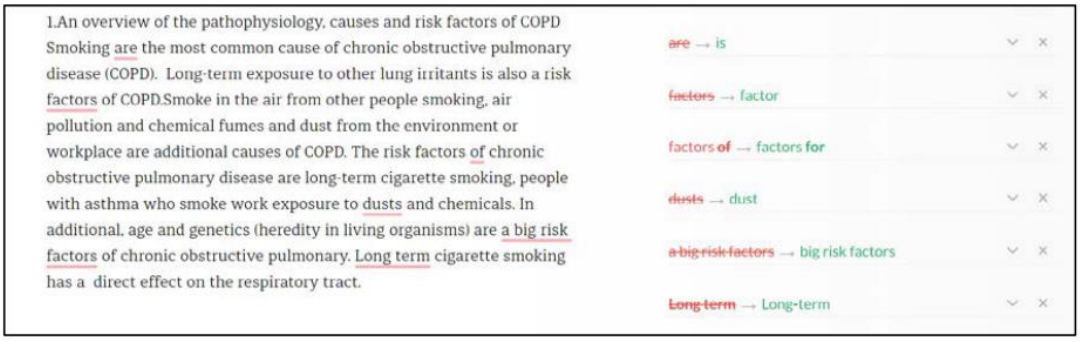


图 2：语法中的直接和间接反馈

关于应该提供多少反馈的问题，目前仍存在争议。研究发现，学生更看重所有错误都被指出的反馈方式（Salehfard 和 Bagheri，2014）。如果学生的目标是逐行编辑，但未达到该目标时，他们可能会感到沮丧（Linville，2009）。然而，其他研究表明，更具针对性的纠错反馈可能更为有效（Jamalinesari 等人，2015），因为集中反馈最常见错误可以提高学生的写作水平（Linville，2009）。Grammarly 则提供了一个中间解决方案，它可以识别违反语法规则的 250 个错误，并提供全面的纠错方法，分类系统则为学生提供了解决高频错误的线索。

选择 Grammarly 的另一个原因是，尽管在大学环境中缺乏关于其表现的研究，但已有一些研究得出了积极的结论。Japos（2013）对本科生进行了一项研究，结果发现 Grammarly 显著提高了他们的写作准确性。另一项研究由 Qassemzadeh 和 Soleimani（2016）进行，他们认为在长期保留被动语态规则方面，Grammarly 的表

现优于传统的教师干预。Reis 和 Huijser（2016）的研究表明，相对于其他反馈系统如 Marking Mate，学生们更喜欢使用 Grammarly。在 Caveleri 和 Dianati（2016）的调查中，他们调查了在澳大利亚高等教育环境中使用 Grammarly 的学生们对其看法，结果基本上都是正面的。学生们表示 Grammarly 很有用，易于使用，并且能够帮助他们提高书面表达和理解语法规则的能力。

这项研究建立在 Caveleri 和 Dianati（2016）的基础上，但提供了基本的区别点。首先，学生没有把 Grammarly 作为独立的学习工具，而是与 ALA 一起使用。其次，本研究将 Grammally 与 Microsoft Word 中使用“跟踪变化”的传统 ALC 方法进行了比较。这项研究还拥有更大的样本量，并根据语言水平、签证状态和授课方式调查了该软件是否对特定的学生群体更有利。

3 方法

3.1 伦理审查

2016 年 2 月 12 日，中央昆士兰大学人类研究伦理委员会为该项目提供了伦理许可。

3.2 参与者

研究人员一致认为，对于旨在确定纠正性反馈效果的研究来说，控制组是不可或缺的（Bitchener, 2008）。考虑到第二语言习得的研究也需要在受控的实验条件下使用控制组和治疗组进行（Ferris, 2010）。由于本研究是出于对纠正性反馈的最佳实践的考虑，并且有大量的第二语言群体，因此我们设立了一个实验组，该实验组使用 Grammarly 接受 ALA 的反馈，而对照组则不使用 Grammarly。

有 96 名学生参与了这项研究。54 名学生为实验组，42 名学生为对照组。实验组的学生要么在中央昆士兰大学攻读学位（n = 48），要么在中央昆士兰大学英语语言中心（ELC）准备上大学（n = 6）。对照组的学生也是在读学位（n = 33）或在 ELC（n = 9）。

3.3 数据采集手段

我们采用了 Creswell 和 Plano-Clark（2011）提出的混合法研究顺序解释性设计，其中涵盖了后续变量。具体而言，我们首先收集并分析了定量数据，以揭示学生对该项目的反应模式，并确定特定声明和群组与这些模式的偏差。随后，我们使用定性记录来解释这些模式，揭示学生为何会以特定的方式对该项目做出回应。我们希望通过这种互补的方式使用这两组数据，以提供更深入的了解该项目的表现，这种更深入的了解可以为未来将该项目作为反馈工具提供参考。

因此，我们创建了一个学生调查，并分三个阶段呈现：个人详细资料、定量反应和定性反应。该调查是利用文献回顾的结果，并通过改编 Caveleri 和 Dianati（2016）采用的研究工具而设计的。

3.4 个人详细资料

- 签证状态：学生被要求确认是本地人、DNESB 或国际学生。国际学生被细分为两组：本科生和研究生或在 ELC 学习大学预科课程的学生。

- 反馈方式：记录了学生是在网上还是面对面地接受反馈。

- 语言水平：记录了已完成语言测试的受访者的分数，如国际英语语言测试（IELTS）、培生英语测试或职业英语测试。大多数人已经完成了雅思考试，没有完成的人的分数通过在线测试转换器转换为雅思分数。

- 作业成绩：曾经提交过作业的学生被要求提供他们在该作业中获得的分数。

3.5 定量阶段

为了测试学生对语法反馈的反应，我们设计了 15 个项目。每个项目要求学生根据一个陈述，从 1（强烈不同意）到 5（强烈同意）的分数来表达他们的看法。这些项目中，项目 1-4 取自文献综述，涉及学生对语法反馈重要性的看法，以及他们对所收到的语法反馈的满意度。其中，我们对学生对语法反馈的总体满意度进行了测量，并分别测试了在给予的时间、提供的数量和重点方面对学生的影响。

3.6 定性阶段

学生回答三个开放性问题。

- Q1：你喜欢你收到的语法反馈中的什么？

- Q2: 你不喜欢你收到的语法反馈中的什么？

- Q3: 对你收到的语法反馈有什么补充意见？

主题分析被用来解释定性数据。我们对这些主题进行编码，然后根据它们是否证实或违背了定量结果进行分类。在统计分析中没有出现的其他主题，则使用归纳分析法从定性评论中进行分类。

3.7 步骤

在试验期间，所有学生都有一个 Grammarly 高级账户。控制组和实验组的学生都接受了最多 15 分钟的语法反馈。

3.8 对照组：面对面与线上

ALA 在学生的作业中发现了语法问题。他们讨论了这些发现，并使用“跟踪修改”在学生的作业中记录他们的建议。

3.9 实验组：面对面

ALA 将每个学生的作业上传至 Grammarly Premium。然后，他们下载了详细的 Grammarly 报告，报告中的反馈涉及六个类别：上下文拼写、语法、标点符号、句子结构、风格和词汇增强。为了使纠错工作更易于管理，报告中最常出现的五个错误按降序处理。

间接纠错反馈首先被采用，ALA 在 Grammarly 报告中找到一个划线的错误例子。为了确定学生目前对错误的理解，并鼓励他们自我纠正，学生会被问到划线部分是否准确。如果学生认为该部分不准确，他们会被要求以不同的方式书写。如果学生

进行了适当的改正，则对该错误的另外一个例子进行审查，以确认其对该语言项目的能力。如果 ALA 认为需要进一步确认理解，则会使用最适合该语言点的技术进行额外的纠错检查。然而，所有的技巧都有相同的目标：鼓励学生思考并阐明他们的选择，而不是仅仅被动地接受程序的建议。

如果学生不确定或无法表达他们的答案，我们会使用直接反馈。我们向学生展示 Grammarly 中的修正建议，并询问该建议是否合适。如果学生回答正确，我们就会查看该学生作品中的另一个错误例子。如果学生不确定该更正是否合适，我们就会要求他们阅读 Grammarly 的信息卡和例子。ALA 使用这些卡片上的解释和例子来解读语法规则并说明其用途。如果 Grammarly 的信息不够清晰或详细，ALA 就会通过示范和提供额外的例子来提供指导。通过这种方式，ALA 填补了 Grammarly 反馈中的空白，因此学生们得到了他们需要的指导和澄清。除了澄清语言问题、找出错误和不必要的建议外，这个程序还鼓励学生提出问题，以便更好地了解 Grammarly 反馈的过程，并为学生提供了一个可复制的使用模式。

3.10 实验组：线上

我们为该学生创建了一个 Grammarly 高级账户，并上传了该学生的作业。我们确定了作业中最常见的五个错误，并提供了作业中这些错误的示例。我们将作业上传到 PDF 注释器，突出显示任何 Grammarly 漏检的错误，并对其进行了 ALA 反馈。

我们向学生发送了一封电子邮件，其中包括他们的 Grammarly 账户的详细信息、他们注释的作业 PDF 副本以及关于哪些错误需要解决的建议。

为了鼓励学生完成调查，校内学生在一份硬拷贝调查中记录了他们的回答，并将数据手动输入 SurveyMonkey。在线学生在 SurveyMonkey 的在线表格中完成同样的调查。

3.11 分析

本研究的因变量为学生对语法反馈的 15 条陈述的同意程度，自变量包括学生是否使用 Grammarly 获得反馈、学生的英语水平（雅思成绩和之前的成绩）、学生所

属的组别（本地学生、DNESB、国际学生和来自 ELC 的国际学生）以及授课模式（面对面或在线）。这些自变量的目的是确定 Grammarly（或非 Grammarly）在哪些具体方面吸引了哪些类型的学生。我们采用了推理统计方法——独立样本 t 检验来比较使用 Grammarly 获得反馈的学生和没有获得反馈的学生之间的差异。然而，我们发现在某些陈述中，各组之间的方差存在明显差异，违反了独立样本 t 检验的假设。因此，在这些情况下，我们采用更为稳健的韦尔奇 t 检验，并用带小数点的自由度来表示结果。最后，我们使用 Cohen's d 来报告显著结果的效应大小。

由于声明同意程度变量呈现偏斜分布，而以前的成绩变量为序数性质，因此我们使用了非参数的（Spearman's）关联分析来研究这两者之间的关系。而对于不同批次的学生间的比较，则采用了方差分析和事后 Tukey 配对比较的方式进行分析。整个过程中，我们设置了显著性水平阈值为 0.05。

4 结果

4.1 学生对 Grammarly 和非 Grammarly 反馈的看法和比较

那些接受了 ALA 建议和 Grammarly 反馈的人普遍对这次经历非常积极，这一点在表 1 中的高平均值得以体现。总的来说，受访者认识到在作业中获得语法反馈的重要性，认为他们从 Grammarly 获得了足够的帮助，并且认

为这些建议很有用且易于理解。他们认为，这些反馈改善了他们的作业和成绩，并提高了他们对作业的信心，同时也在作业之外发展了他们的信心和语言技能。唯一低于 4.4 的平均值是“我做了 ALC 不建议的额外语法修正”（M = 3.72，SD = 1.47）。这表明，尽管学生们仍然同意这一说法，但他们对是否要做比 ALA 建议的更多的事情有些犹豫。定性回答在很大程度上证实了表 1 中的积极结论。有 6 名学生表示希望学校能长期提供该软件，11 名学生声称该项目对他们有帮助，14 名学生用积极的动词和形容词来描述反馈。

表 1 学生对语法反馈陈述的平均和 SD 评分

Statement	Grammarly (n = 54)		Non-Grammarly (n = 42)		Inferential statistics			
	M	SD	M	SD	t	df	p	d
对我来说，获得作业的语法反馈很重要	4.93	.26	4.93	.26	-.05	94	.961	-
顾问花了足够的时间在语法反馈上，帮助我改进作业	4.74	.62	4.38	.96	2.11	66.43	.039	.45
我的作业得到了足够的语法反馈	4.57	.69	4.33	.87	1.51	94	.135	-
语法反馈确实集中那些在写作中出现的错误	4.54	.77	4.14	.93	2.23	79.24	.029	.47
很容易理解这些错误，因为给出的解释很清楚。	4.65	.76	4.36	1.01	1.56	73.83	.123	-
我得到了很多关于如何在作业中改善我的语法的有用建议	4.67	.70	4.29	.97	2.15	71.86	.035	.45
利用反馈意见对我的作品进行语法修改很容易	4.65	.68	4.20	.98	2.54	67.55	.014	.53
我做了 ALC 顾问在咨询时建议的语法修正	4.54	.86	4.38	.94	.85	94	.399	-
我还做了一些 ALC 不建议的语法修正	3.72	1.47	3.12	1.33	2.08	94	.041	.43
语法反馈改善了我的作业	4.80	.49	4.55	.92	1.59	59.06	.117	-
语法反馈使我对提交作业更有信心	4.76	.58	4.38	.99	2.21	62.56	.031	.47
语法反馈提高了我的成绩	4.45	.85	4.17	1.01	1.50	93	.136	-
语法反馈长期地发展了我的语言，因为我可以理解语法规则	4.48	.72	4.07	1.07	2.14	68.48	.036	.45
语法反馈培养了我对长期使用语言的信心，因为我可以更加理解语法规则	4.46	.75	3.98	1.09	2.47	69.02	.016	.51
我对收到的语法建议感到满意	4.75	.55	4.43	.83	2.19	68.09	.032	345

注：df=自由度，小数点位表示使用威尔斯特拉斯函数。d 指 Cohen's d（效应量）。

然而，虽然学生们对这个项目基本持肯定态度，但他们确实发现了六个方面的问题。

关注的主要方面是反馈的准确性，有 9 名学生认为这是一个问题："并非所有的语法错误都是有效的或正确的"。学生们还表示，预设为美式英语导致该程序注意到无效的拼写错误。他们发现了被动语态、复杂句子和词汇选择方面的问题，Grammarly 建议学生根据风格而不是准确性来改变这些问题。

学生们还指出，该程序可能会漏掉一些错误，而且软件提供的建议有时难以理解。此外，该软件在文件上传和兼容性方面也存在一些技术问题，需要进一步关注和改进。对于这两个学生来说，这个软件的缺点显然超过了优点，他们明确表示更喜欢传统的 ALC 反馈方法。

表 1 还提供了通过传统的、非 Grammarly 反馈机制接受反馈的学生的 15 条陈述的平均值和标准差。总的来说，结果是相似的，即非 Grammarly 的学生表示他们认为反馈是有用的，他们得到了足够的反馈，并且反馈有助于提高他们的语法、成绩和信心。然而，Grammarly 学生的平均值在某些方面明显较高。与非 Grammarly 学生相比，Grammarly 学生更有可能说 ALA 在语法反馈上花了足够的时间来改善他们的作业，而且通过 Grammarly 的反馈集中在他们经常犯的错误上。擅长语法的学生明显更有可能同意他们通过 Grammarly 收到了很多有用的反馈，而且根据这些反馈进行语法修改也明显容易。使用 Grammarly 的学生报告说，他们更有可能对 Grammarly 没有推荐的内容进行额外的修改，并且对提交作业的信心明显增强。与不使用 Grammarly 的学生相比，使用 Grammarly 的学生对反馈意见在完成作业之外同时发展了他们的语言，对自己语言的信心明显更高。总的来说，与非 Grammarly 学生相比，Grammarly 学生对他们收到的语法建议的满意度明显更高。

4.2 不同语言水平的学生对 Grammarly 的看法

35 名 Grammarly 学生和 25 名非 Grammarly 学生的雅思（和雅思指数）分数被采集，而 36 名 Grammarly 学生和 25 名非 Grammarly 学生的以往成绩被采集。

对于 Grammarly 学生来说，那些雅思分数较低的学生往往对他们得到的语法建议更加满意（Spearman's rho = -.41, p < .05），而那些以前成绩较好的学生则更有可能认为进行语法修改很容易（Spearman's rho = .38, p < .05）。这些结果对于非 Grammarly 的学生来说没有统计学意义，而且没有观察到结果之间的显著差异。对于非 Grammarly 学生来说，那些雅思成绩较低的学生明显更倾向于同意语法反馈能提高他们的成绩（Spearman's rho = -.40, p < .05）。没有其他相关性在统计学上是显著的，但必须承认，由于数据缺失、样本量小以及 Spearman 相关的保守性质，这些结果很可能是力量不足的，因此从这些结果得出的结论是暂时的。

定性数据支持了雅思成绩较低（6 分或以下）的研究生对 Grammarly 的反馈感到满意的结论，因为 6 分或以下的学生是最不可能对该程序发表负面评论的群体。然而，来自 ELC 的学生是所有小组中语言能力最低的，他们对 Grammarly 的局限性最有发言权，所有 6 名学生都表示了担忧。其中四名学生对该程序的不准确性表示担忧，另外两名学生则表示该程序 “并没有什么意义”，而且“给出了错误的建议”。这些发现表明，Grammarly 可以用于不同语言水平的学生，但在对 ELC 学生使用时可能需要采取不同的方法。

4.3 不同签证身份的学生的看法

由于 Grammarly 和非 Grammarly 的学生在对一些陈述的认同上存在差异，因此对不同组别的差异进行了因子方差分析，以 Grammarly（2 个水平：是与否）、组别（4 个水平：本地、DNESB、国际和国际 ELC）作为主效应，并对每个陈述进行交互效应。由于两个原因，应谨慎看待这些分析。

首先，一些陈述的单元是恒定的（即所有受访者都选择相同的答案）；其次，对于这种类型的分析，单元大小相对较小。在发现重大影响的地方，进行了事后的 Tukey 配对比较。

在三条陈述中观察到了各组学生之间的差异。国际学生对他们收到的语法反馈的满意度普遍较低，而且同意语法反馈长期培养了他们对语言的信心的可能性也明显低于 DNESB 学生。与 DNESB 和国际 ELC 学生相比，本地学生表示语法反馈长期发展了他们的语言的可能性明显较小。然而，这些结果是群组的主效应，这意味着

它们是在是否接受了 Grammarly 反馈的情况下的平均数，而且这些陈述都没有明显的交互作用，这表明结果并不因他们是否接受 Grammarly 反馈而有明显不同。

定性数据支持 DNESB 学生对 Grammarly 的积极反应。在问题 2 中，学生被要求指出不喜欢 Grammarly 的原因，DNESB 学生的负面回答最少（回答率仅为 21%），而本地和国际学生的回答率为 66%。此外，DNESB 学生是唯一一组明确表示 Grammarly 对他们有用的学生群体。有一位学生认为："我认为这是一个很好的工具，可以帮助非英语背景的学生。" 国际学生对问题 2 的回答率非常高，而统计结果显示他们对语法反馈的满意度较低，这表明需要采取更有针对性的方法，以满足这些学生的需求。

4.4 按授课模式划分的学生回答

我们还对那些在网上和在校园里接受反馈的学生进行了比较。54 名 Grammarly 学生中的 27 人和 42 名非 Grammarly 学生中的 13 人在网上得到了反馈，而其余的人则在校园里得到了反馈。在任何陈述中，那些在网上和校园里接受反馈的学生之间没有发现明显的差异（最大的 $F(1,90) = 2.58, p = .112$）。我们还计算了交互效应，比较了 Grammarly 和非 Grammarly 学生的在线与校园反馈之间的差异。没有发现明显的交互效应（最大的 F（1,90）=3.32，p=.072）。

5 讨论

文献综述中指出，给予语法反馈的时间和细节是导致学生对 ALC 不满意的一个潜在原因。定量数据显示，Grammarly 的学生明显更倾向于认为在语法反馈上花了足够的时间。

定性数据支持了这些发现。在使用 Grammarly 接受反馈的学生中，没有人提到时间问题，而以传统方式接受反馈的四名学生则抱怨没有足够的时间进行语法反馈。例如，有一个学生写道，他们"希望能够有更多的时间来关注语法和语言方面"。

接受 Grammarly 建议的学生也明显更有可能表示他们收到了很多有用的反馈。在 Grammarly 组中，没有学生抱怨他们没有收到足够的语法反馈。然而，非

Grammarly 组中有 6 名学生表示不满意。

文献综述中指出，ALA 倾向于优先考虑宏观技能而不是语法，这是一个潜在的问题。Grammarly 组的学生没有提到这个问题，但是对照组的学生确实认为 ALA 牺牲了语法反馈来专注于其他问题。

文献综述中指出，语法反馈的不足是语言发展的一个潜在阻碍，对照组的一些学生确实认为他们的长期发展受到了抑制。

根据文献审查结果，ALA 被定位为 Grammarly 反馈和学生之间的中介，这种方法有助于减少学生对该程序的负面反应。具体而言，ALA 采取了以下措施：首先，管理学生对反馈的期望，让他们明确意识到反馈并不是完美无缺的；其次，通过识别和删除错误的建议和表述，帮助使反馈更加可靠。虽然不准确性是 AFP（自动反馈程序）的一个问题（Hoang & Kunnan，2013；Liu & Kunnan，2016），但是 ALA 可以通过减少错误数量来缓解此问题。同时，ALA 还可以把 Grammarly 的错误视为学习工具，帮助学生了解他们为什么出现错误。最后，也是最重要的，ALA 可以提供额外的指导和解释，以帮助学生更好地理解他们所犯的错误。

没有一个 Grammarly 的学生对他们得到的解释发表负面意见，而四个非 Grammarly 的学生则表示他们需要 "更清晰的解释"。同样，没有一个 Grammarly 学生要求额外的支持，而非 Grammarly 学生则提到需要通过额外的语法课程、讲座、材料和网站来进一步澄清。因此，我们建议 Grammarly 继续与 ALA 一起使用，直到整理出可比较的结果，证明其作为一个独立工具的有效性。

6 结论

这项研究表明，使用 ALC 的学生希望在语法方面得到帮助，以提高他们的作业和发展他们的长期写作表现。Grammarly 被选为一种提供学生想要的语法反馈的方法，而不影响他们更高阶技能的发展。如果学生不理解程序对语言项的响应，还设计了一个系统的 Grammarly 使用方法，并提供 ALA 补充建议。比较了接受语法建议和非语法建议的学生之间的结果。以传统方式接受语法指导的学生对所收到的反馈基本满意，但从 Grammarly 和 ALA 联合获得反馈的学生对所花费的时间、收到的反馈数量、对成绩的短期收益、对写作的长期收益、对写作的信心和独立校对能力方

面更满意。学生们还在 Grammarly 方法中确定了书面纠正反馈的最佳实践的几个特征。这包括响应速度、处理错误的范围以及反馈的针对性和个性化。然而，人们对该程序的准确性、遗漏错误和纠正正确错误的倾向以及技术故障表示了担忧。如果ALA没有首先参与反馈，这些担忧会更加普遍，因为ALA能够管理学生的期望，减少错误的反馈，并将不准确的反馈作为学习工具。在授课模式之间没有发现显著差异，这表明该计划可用于在线和在校学生。同样，结果表明，Grammarly 可以用于不同语言水平和签证状态的学生，但对于 ELC 和国际学生可能需要更大的自由裁量权或修改方法。基于这些发现，因此建议将 Grammarly 添加到 ALC 的工具箱中，但它应该继续与 ALA 一起使用。

6.1 研究的局限性

本研究存在统计能力问题，因为总体样本量较小。更大的样本可能会发现显著的差异，因此未来任何关于学生对 Grammarly 的看法的研究都应该使用更大的样本。另外，一些学生变量也没有被考虑，这些变量可能会影响回答。例如，学生花在学习英语上的时间长度和作业的长度，以及与关系变量相关的数据也没有被收集。学生的 ALA 关系和在 ALA 中的偏好（例如性别）等变量可能会影响学生的反应。此外，与语言水平相关的数据使用雅思或相当于雅思的写作成绩，但这些并不总是语言表现的可靠指标。要求学生自我报告这些分数和大学成绩也是有问题的，因为学生可能不会准确地报告，可能会夸大表现水平。

6.2 未来的研究

本研究主要关注学生对语法的认知。为了更全面地了解 Grammarly 作为反馈工具的作用，还需要进一步的研究。首先，需要考虑 ALA 对该程序的看法，并将这些发现与 ALA 对传统反馈方法的态度进行比较。其次，由于不准确被认为是该程序的一个问题，需要进一步调查以确定 Grammarly 最经常遗漏或错误识别的错误。这将有助于为未来的反馈研究提供信息，并使其更有用。

为了确定 Grammarly 提高学生语法准确性的程度，我们需要比较学生在使用

Grammarly 前后的语法准确性。我们可以通过比较学生在作业初稿（未使用 Grammarly）和终稿（使用 Grammarly 后）中的语法准确性来发现短期的改进。我们还需要对那些以传统方式获得反馈的学生的作业进行分析，以比较使用 Grammarly 的学生和接受传统 ALC 建议的学生在短期内的准确性。要确定学生语法准确性的长期提高，我们可以进行纵向研究。在这些研究中，学生在大学学习期间每学期都需提交一份作业，并得到语法反馈。我们可以比较每个学生的作业初次提交和最终提交的准确率百分比，以确定是否在高频错误和整体写作精度方面取得了改进。此外，我们还可以要求这些学生每学期进行一次语法测试，以评估在不使用 Grammarly 时，他们的准确性是否得到提高。

7 参考文献

[1] Bacon, D. R., & Anderson, E. S. (2004). Assessing and enhancing the basic writing skills of marketing students. Business Communication Quarterly, 67(4), 443–454. https://doi.org/10.1177/1080569904271083

[2] Bitchener, J. (2008). Evidence in support of written corrective feedback. Journal of Second Language Writing, 17(2), 102–118. http://dx.doi.org/10.1016/j.jslw.2007.11.004

[3] Burston, J. (2008). BonPatron: An online spelling, grammar, and expression checker. CALICO Journal, 25(2), 337–347. Retrieved from http://journals.equinoxpub.com/index.php/CALICO/article/view/23116

[4] Caveleri, M., & Dianati, S. (2016). You want me to check your grammar again? The usefulness of an online grammar checker as perceived by students, Journal of Academic Language and Learning, 10(1), 223–236. Retrieved from http://www.journal.aall.org.au/index.php/jall/article/view/393/246

[5] Channock, K., D'Cruz, C., & Bisset, D. (2009). Would you like grammar with that? Journal of Academic Language and Learning, 3(2), 1–12. http://www.journal.aall.org.au/index.php/jall/article/view/70/63

[6] Chen, C. E., & Cheng, W. (2006, May). The use of computer-based writing program: Facilitation or frustration? Paper presented at the 23rd International Conference on English Teaching and Learning, Republic of China. Retrieved from ERIC database. (ED492078)

[7] Chen, C. E., & Cheng, W. (2008). Beyond the design of automated writing evaluation: Pedagogical practices and perceived learning effectiveness in EFL writing classes. Language, Learning & Technology, 12(2), 94–112. Retrieved from http://llt.msu.edu/vol12num2/chencheng.pdf

[8] Choi, J., & Lee, Y. (2010). The use of feedback in the ESL writing class integrating automated essay scoring (AES). In D. Gibson & B. Dodge (Eds.), Proceedings of Society for Information

Technology & Teacher Education International Conference (pp. 3008–3012). Chesapeake, VA: Association for the Advancement of Computing in Education. Retrieved from https://www.learntechlib.org/p/33826/

[9] Cocks, T., & Stokes, J. (2013). Policy into practice: A case study of widening participation in Australian higher education. Widening Participation and Lifelong Learning, 15(1), 22–38. http://dx.doi.org/10.5456/WPLL.15.1.22

[10] Creswell, J. W., & Plano Clark, V. L. (2011). Designing and conducting mixed methods research.Thousand Oaks, CA: Sage.

[11] Department of Education and Training. (n.d,). Non-English speaking background. Canberra: Australian Government.

[12] Dikli, S. (2010). The nature of automated essay feedback. CALICO Journal, 28(1) 99–134. Retrieved from https://calico.org/page.php?id=5

[13] Dikli, S., & Bleyle, S. (2014). Automated essay scoring feedback for second language writers: How does it compare to instructor feedback? Assessing Writing, 22, 1–17.http://dx.doi.org/10.1016/j.asw.2014.03.006

[14] Ferris, D. R. (2010). Second language writing research and written corrective feedback in SLA: Intersections and practical applications. Studies in Second Language Acquisition, 32(2), 181–201. https://doi.org/10.1017/S0272263109990490

[15] Grammarly Inc. (2017). Grammarly Premium: Keep it clean and write your heart out. Retrieved from https://www.grammarly.com/premium

[16] Harris, A. (2016). Integrating written communication skills: Working towards a whole of course approach. Teaching in Higher Education, 21(3), 287–300.http://dx.doi.org/10.1080/13562517.2016.1138456

[17] Hoang, G. T. L., & Kunnan, A. J. (2016). Automated essay evaluation for English language learners: A case study of MY Access. Language Assessment Quarterly, 13(4), 359–376. http://dx.doi.org/10.1080/15434303.2016.1230121

[18] Hoon, T. B. (2006). Online automated essay assessment: Potentials for writing development. In A.Treloar & A. Ellis (Eds.), Making a difference with Web technologies. Proceedings of AusWeb 2006, The Twelfth Australasian World Wide Web Conference (pp. 230–234). Lismore: Southern Cross University. Retrieved November 1, 2016, from http://ausweb.scu.edu.au/aw06/papers/refereed/tan3/paper.html

[19] Huang, L. S. (2011). Are we having the effect we want? Implementing outcomes assessment in an academic English language-support unit using a multi-component approach. Writing Program Administration, 35(1), 1–11. Retrieved from http://wpacouncil.org/archives/35n1/35n1huang.pdf

[20] Hyland, K., & Hyland, F. (2006). Interpersonal aspects of response: constructing and interpreting teacher written feedback. In K. Hyland & F. Hyland (Eds.), Feedback in second language writing: Contexts and issues (pp. 206–224). New York, NY: Cambridge University Press.

https://doi.org/10.1017/CBO9781139524742.013

[21] Jamalinesari, A., Rahimi, F., Gowhary, H., & Azizifar, A. (2015). The effects of teacher-written direct vs.indirect feedback on students' writing. Procedia - Social and Behavioral Sciences, 192, 116–123. http://dx.doi.org/10.1016/j.sbspro.2015.06.018

[22] Japos, V. (2013). Effectiveness of coaching interventions using Grammarly software and detection software in reducing grammatical errors and plagiarism of undergraduate researchers. JPAIR Institutional Research, 1(1), 97–109. http://dx.doi.org/10.7719/irj.v1i1.207

[23] Knoch, U., Rouhshad, A., & Storch, N. (2014). Does the writing of undergraduate ESL students develop after one year of study in an English-medium university? Assessing Writing, 21, 1–17. http://dx.doi.org/10.1016/j.asw.2014.01.001

[24] Liao, H. (2016). Enhancing the grammatical accuracy of EFL writing by using an AWE-assisted process approach. System, 62, 77–92. http://dx.doi.org/10.1016/j.system.2016.02.007

[25] Linville, C. (2009). Editing line by line. In S. Bruce & B. Rafoth (Eds.), ESL writers: A guide for writing center tutors (2nd ed., pp. 116–131). Portsmouth, NH: Heinemann.

[26] Liu, S., & Kunnan, A. J. (2016). Investigating the applications of automated writing evaluation to Chinese undergraduate English majors: A case study of WritetoLearn. CALICO Journal, 33(1), 71–91. http://dx.doi.org/10.1558/cj.v33i1.26380

[27] McGregor, K. M., Merchant, A. R., & Butler, M. (2008). Making feedback more immediate. CALlaborate International, 16(1) 8–11. https://openjournals.library.sydney.edu.au/index.php/CAL/article/view/6020/6671

[28] McNaught, K., & Shaw, G. (2016). Preparing undergraduate students to be successful writers: Exploring the spelling, grammar and punctuation skills of students identified as potentially 'at risk'. Journal of Academic Language and Learning, 10(2), 11–19. http://www.journal.aall.org.au/index.php/jall/article/view/406/253

[29] Murray, N. (2011). Ten 'Good Practice Principles' … ten key questions: Considerations in addressing the English language needs of higher education students. Higher Education Research & Development, 31(2), 233–246. http://dx.doi.org/10.1080/07294360.2011.555389

[30] Pearson Education Ltd. (2012). PTE academic score guide, version 4. Retrieved from http://pearsonpte.com/wp-content/uploads/2014/07/PTEA_Score_Guide.pdf

[31] Qassemzadeh, A., & Soleimani, H. (2016). The impact of feedback provision by Grammarly software and teachers on learning passive structures by Iranian EFL Learners. Theory and Practice in Language Studies, 6(9), 1884–1894. http://dx.doi.org/10.17507/tpls.0609.23

[32] Ramineni, C., & Williamson, D. M. (2012). Automated essay scoring: Psychometric guidelines and practices. Assessing Writing, 18(1), 25–39. http://dx.doi.org/10.1016/j.asw.2012.10.004

[33] Reis, C., & Huijser, H. (2016). Correcting tool or learning tool? Student perceptions of an online essay writing support tool at Xi'an Jiaotong-Liverpool University. In S. Barker , A. Pardo, &

C.Colvin (Eds.), Show Me The Learning. Proceedings of ASCILITE 2016, 33rd International Conference of Innovation, Practice and Research in the Use of Educational Technologies in Tertiary Education (pp. 529–533). Adelaide: University of South Australia). Retrieved from http://2016conference.ascilite.org/wp-content/uploads/ascilite2016_reis_concise.pdf

[34] Salehfard, E., & Bagheri, M. S. (2014). Preferences of EFL teachers and students towards corrective feedback in writing. Modern Journal of Language Teaching Methods, 4(4), 308–319. Retrieved from http://mjltm.org/en/archive.php?rid=18

[35] Scouller, K., Bonanno, H., Smith, L., & Krass, I. (2008). Student experience and tertiary expectations: Factors predicting academic literacy amongst first year pharmacy students. Studies in Higher Education, 33(2),167–178. http://dx.doi.org/10.1080/03075070801916047

[36] Tynan, L., & Johns, K. (2015). Piloting the post-entry language assessment: Outcomes from a new system for supporting research candidates with English as an additional language. Quality in Higher Education, 21(1), 66–78. http://dx.doi.org/10.1080/13538322.2015.1049442

[37] Wang, Y. J., Shang H. F., & Briody, P. (2013). Exploring the impact of using automated writing evaluation in English as a foreign language university students' writing. Computer Assisted Language Learning, 26(3), 234–257. http://dx.doi.org/10.1080/09588221.2012.655300

[38] Winder, R., Kathpalia, S., & Koo, S. (2016). Writing centre tutoring sessions: Addressing students' concerns. Educational Studies, 42(4), 323–339. http://dx.doi.org/10.1080/03055698.2016.1193476

文献综述

文本纠错综述报告

摘　要：文本纠错是自然语言处理的一项重要任务，主要包含查错（GED）和纠错（GEC）两部分内容。随着机器学习、深度学习技术的发展，文本纠错模型的性能有了极大的提升，已被广泛运用到日常写作辅助，公文检查，二语学习等多种领域。从早期基于规则/知识工程的方法，到基于统计机器翻译的方法以及基于神经机器翻译的方法，再到今天我们常用的基于预训练语言模型的方法，针对文本纠错任务的研究已经十分深入。本文将简单介绍近几年文本纠错领域的发展过程和研究成果，同时分析探讨近期热门的大模型对纠错任务的影响。

关键词：文本纠错；统计机器翻译；神经机器翻译；预训练语言模型

A Comprehensive Review of Grammar Error Correction

Abstract: Grammar Error Correction(GEC) is an important task of natural language processing, which mainly includes Grammar Error Detection (GED) and Grammar Error Correction (GEC). With the development of machine learning and deep learning technology, the performance of Grammar Error Correction model has been greatly improved, and has been widely used in daily writing assistance, official document inspection, second language learning and so on. From the early rule-based/knowledge engineering methods, to statistical machine translation based methods and neural machine translation based methods, to today's commonly used methods based on pre-training language models, the research on Grammar Error Correction tasks has been very in-depth. This paper will briefly introduce the development and achievements in the field of text error correction in recent years, as well as the current research trends. Also, the impact of recent popular Foundation Models on error correction tasks will be analyzed.

KeyWords: GEC; SMT; NMT; Pre-training Language Models

1 引言

文本纠错任务，旨在通过模型自动识别并纠正文本中的各式各样的错误，从简单的拼写错误，到复杂的主语缺失，歧义等语法错误，都是文本纠错任务的处理对象。目前，文本纠错任务已被广泛运用到日常写作辅助，公文检查，二语学习等方面，对文本纠错任务的研究具有十分重大的现实意义。

早期人们解决文本纠错任务的方法是由专家撰写知识库和相关规则，模型性能主要取决于规则的完善程度，成本极高。后来随着机器学习的发展，人们提出了将文本纠错视为机器翻译任务，并在此基础上提出了基于统计机器翻译的方法，以及基于神经机器翻译的方法。而在预训练语言模型被提出后，预训练+微调的方式逐渐流行起来，直至今日，基于预训练语言模型的方式已经成为最主流的文本纠错方法。

目前文本纠错领域主要的难点在于数据匮乏，而数据标注的成本又极高。此外，语言差异，领域差异导致的文本错误类型分布、语言风格等不同，使得一种语言、一个领域的数据难以迁移到其它语言、其它领域，这也是导致数据匮乏的重要因素。

2 研究现状

目前最广泛使用的免费公开数据集主要包括 NUCLE[4],Lang-8[2],FCE[3],JFLEG[4],Write&Improve+LOCNESS[5]这几个英文数据集，以及hsk[6],MuCGEC[7]这两个中文数据集。当前对文本纠错领域的研究已经十分深入，我们会在这一段介绍目前基本的文本纠错方法，一些好用的性能提升技术，以及一些简单有效的数据增强方式。

2.1 方法

本段我们将会介绍用于解决 GEC 问题的基本数据驱动方法，包括基于 SMT（统计机器翻译）的方法、基于 NMT（神经机器翻译）[8]的方法、基于分类的方法和基于 LM 的方法。

2.1.1 基于SMT的方法

在 SMT 方法之前,语法错误修正主要是通过基于规则或分类的方法,它们都具有明显的局限性[9]。一方面,设计规则之前通常需要大量的语言信息和专业知识，同时，由于自然语言的灵活性，人们很难找到一个通用的规则，这使得基于规则或分类的方法成本极高。

在此基础上，人们提出了基于统计的机器翻译模型，即通过将 GEC 任务看作是从病句到正确句子的翻译过程，通过在大量平行语料上的训练，让模型学习到语法错误和正确形式之间的关系，从而实现纠错。

2.1.2 基于NMT的方法

虽然基于 SMT 的方法可以吸收大量的平行数据和单语数据以及辅助神经网络组件，但它仍然存在上下文信息不足和泛化能力有限的问题。作为一种解决方案，许多学者开始研究基于 NMT 的 GEC 方法。

随着神经编码器-解码器模型在机器翻译中获得的性能的提高，基于神经编码器-解码器的模型被采用和修改[10]。与基于 SMT 的 GEC 系统相比，基于 NMT 的模型具有两个优点。首先，神经编码器-解码器模型直接从训练并行数据中学习从源到目标的映射，而不是通过 SMT 中所需的特征来捕捉映射规律。其次，基于 NMT 的系统能够比基于 SMT 的方法更有效地纠正看不见的不符合语法的短语和句子，从而提高泛化能力。目前主流的两种方案分别是基于序列到序列的模型（Seq2Seq）和基于序列到编辑的模型（Seq2Edit）。

2.1.2.1 Seq2Seq模型

早些年间，Seq2Seq 模型主要采用 RNN、CNN 等架构，期间为了解决语义向量无法完全表达整个输入序列的信息等问题，Attention 机制被引入。随着 Transformer 在机器翻译领域的成功，主流的语法纠错也都使用了 Transformer 来作为 Seq2Seq 模型的 Encoder 和 Decoder。Transformer 模型完全抛弃了传统 RNN 在水平方向的传播，通过采用 Self-Attention 自注意力机制，只在垂直方向进行传播。我们只需要不断叠加 Self-Attention 层，就能构建一个完整的 Transformer 模型。

2.1.2.2 Seq2Edit模型

Seq2Seq模型依然存在一些问题，比如无法并行计算、训练集需求量大、只生成结果不生成错误类型等[11]。针对这些问题，诞生了 Seq2Edit 模型。Seq2Seq 模型采用 Encoder-Decoder 结构，将编码后的序列信息传给解码端生成序列。相比之下Seq2Edit 模型只有 Encoder 层，在每个 Time-Step 预测生成一个编辑动作，通过这些编辑动作与源文本结合生成目标文本。Seq2Edit 模型的编辑动作一般包括保持不变、删除、添加和替换等，编辑空间远小于 Seq2Seq 模型的词汇空间。总的来说，Seq2Edit 的推理速度相比 Seq2Seq 模型有了很大提升，且精度没有下降。

2.1.3 基于分类的方法

在基于分类的方法中,训练多类分类器以基于来自人工特征工程或深度学习模型的词的上下文的特征来预测混淆集中的正确词。给定要校正的文本，对于文本中出现在混淆集中的每个标记，分类器预测最可能的候选词。分类器是在母语或非母语说话者数据上训练的，因此降低了对注释数据的需求。常用的排序算法有最大熵、朴素贝叶斯、决策树和平均感知器。虽然分类方法曾经很流行，但现在已经不被广泛采用。

2.1.4 基于LM的方法

基于 SMT 的方法和基于 NMT 的方法都是有监督的，因此需要大量的并行训练数据。然而，基于 LM 的无监督方法被应用到 GEC 中，并且获得了与监督方法相当的性能。基于 LM 的方法的优势在于，考虑到大量并行数据在 GEC 中不可用的事实，它们不需要大量的并行注释数据。因此，基于 LM 的 GEC 系统往往是在资源不足的情况下开发的。大多数基于 LM 的 GEC 方法使用 LM 来评估待纠正的假设，其中假设是通过根据设计的规则对源不合语法的句子进行修改或者通过用从混淆集中选择的词替换源句子中的标记来产生的。

2.2 性能提升技术

为了提高模型性能，改进模型结构是一种十分流行的思路，总体而言，目前的研究大多是考虑给模型加入额外信息来辅助纠错，如[12]通过给模型加入上下文信息来帮助模型处理一些篇章级错误。[13]为模型加入句法信息（依存句法，成分句

法）来辅助模型纠错。[14]提出了将 GED 模型的结果作为辅助输入参与微调的方法。

此外，还有一些不改动模型的方法，但也能有效提升模型性能，如模型集成，多轮迭代纠正[15]，重排[14]等。

2.2.1 模型集成

目前的主流模型集成方式有两种，一种是对模型预测结果的概率进行求和平均，另一种则是通过简单的投票来进行集成。投票集成的方式相对而言泛用性更强，不需要考虑模型输出结构，概率求和的方式繁琐一些，但能在更细粒度上进行集成。

2.2.2 多轮迭代纠正

对于句子中的连续错误，尤其是连续的缺失错误，他们之间通常会存在一定的依赖关系，需要知道前一个 token 才更有利于预测下一个 token，这时模型往往不能一次纠正出所有错误，因此采用多轮迭代纠正，直至模型认为句子没有错误是一种提高模型性能的有效方式。

2.2.3 重排

重排，指的是模型预测出 top-k 个可能结果后，对这 k 个候选结果进行重新排序，选出最有可能的作为最终结果。[14]用 GED 模型生成一个错误检测的标签序列，根据其与假设的海明距离，为假设重排顺序，提升模型性能。

2.3 数据增强

由于文本纠错领域的高质量有有标签数据较少，因此数据增强是一种常用的提升模型性能的方式。数据增强，指通过一定的规则生成人造的伪标签数据，将其和原始训练数据混合一起训练模型的方式。

一种常见的数据增强方法是向正确文本中插入错误，人造伪标签数据。值得注意的是，随机插入生成的错误与真实错误差异较大，因此训练效果只是差强人意。针对这个问题，[16]提出了先统计各错误的分布，再使用 ERRANT 自动注释工具里的错误类型来辅助生成合成数据，使人造伪数据更符合领域内的真实错误分布的方法。此方法能大大减少人造数据和真实数据的差异，但由于没法完全了解领域内的

真实错误分布，只能通过开发集大致估计，因此伪数据和真实数据仍然存在差异，不过效果比随机插入错误生成伪数据要好得多。

此外自训练[15]也是文本纠错领域常用的一种数据增强方法，自训练指的是通过真实的有标签数据训练一个模型，然后用其纠正不合语法的句子，将预测结果作为伪数据加入到训练数据中，迭代训练模型。经过实验验证，虽然模型预测的结果直接作为训练数据会带来一定的噪声，但总体而言模型的性能还是有提升的。

还有一种反向翻译人造伪数据的方式[17]，灵感来自机器翻译领域的研究。常规GEC的seq2seq模型将纠错任务视为机器翻译任务，通过将病句翻译为正确句来实现纠错，反向翻译指训练一个将正确句子翻译为错误句子的模型（造错模型），用其生成伪数据，优点是造错模型的训练数据只需将原始训练句对置换顺序，很容易获取。

[18]提出了生成流畅度提升伪实例作为额外训练数据的方法，即在训练期间将预测的不太流畅的句子与其参考句子配对来建立新的错误参考句子对作为训练数据，[19]提出了从每个训练实例中，另外构造多个训练实例，每个训练实例都涉及对特定类型错误的纠正，使用这些额外构建的训练实例和原始实例来依次训练模型的方法。这样做，模型不仅被训练为逐步纠正错误，而且还利用不同类型错误之间的相互依赖性来获得更好的性能。

3 总结与展望

本文介绍了文本纠错任务的基本方法，有早期的基于规则的方法，基于SMT的方法，基于NMT的方法，基于分类的方法以及基于LM的方法，目前最常用的是基于NMT的方法。传统的基于规则、基于统计、基于分类的方法已经逐渐少有人使用。LM方法的特点是无监督学习，可以不依赖平行语料。

目前文本纠错领域的研究已经十分深入，单领域上文本纠错模型的性能已经十分可靠。最近兴起的大模型对于篇章级纠错以及领域差异这些问题有着很大的潜力，不过目前来说，大模型在纠错的同时，还会进行大量改写和润色操作，有不可控的问题。接下来，如何利用大模型做数据增强，或是将大模型和强化学习相结合是很有前景的研究方向。

4 参考文献

[1] Dahlmeier D, Ng H T, Wu S M. Building a large annotated corpus of learner English: The NUS corpus of learner English[C]//Proceedings of the eighth workshop on innovative use of NLP for building educational applications. 2013: 22-31.

[2] Tajiri T, Komachi M, Matsumoto Y. Tense and aspect error correction for ESL learners using global context[C]//Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). 2012: 198-202.

[3] Yannakoudakis H, Briscoe T, Medlock B. A new dataset and method for automatically grading ESOL texts[C]//Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies. 2011: 180-189.

[4] Napoles C, Sakaguchi K, Tetreault J. JFLEG: A fluency corpus and benchmark for grammatical error correction[J]. arXiv preprint arXiv:1702.04066, 2017.

[5] Bryant C, Felice M, Andersen Ø E, et al. The BEA-2019 shared task on grammatical error correction[C]//Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications. 2019: 52-75.

[6] Baolin Zhang. 2009. Features and functions of the HSK dynamic composition corpus (in Chinese). International Chinese Language Education, 4:71–79.

[7] Zhang Y, Li Z, Bao Z, et al. MuCGEC: a Multi-Reference Multi-Source Evaluation Dataset for Chinese Grammatical Error Correction[J]. arXiv preprint arXiv:2204.10994, 2022.

[8] Zheng Yuan and Ted Briscoe. 2016. Grammatical error correction using neural machine translation. In Proceedings of NAACL-HLT, pages 380–386.

[9] Sascha Rothe, Jonathan Mallinson, Eric Malmi, Sebastian Krause, and Aliaksei Severyn. 2021. A simple recipe for multilingual grammatical error correction.In Proceedings ofACL-IJCNLP, pages 702–707.

[10] Satoru Katsumata and Mamoru Komachi. 2020.Stronger baselines for grammatical error correction using a pretrained encoder-decoder model. In Proceedings ofAACL, pages 827–832.

[11] Kostiantyn Omelianchuk, Vitaliy Atrasevych, ArtemChernodub, and Oleksandr Skurzhanskyi. 2020.Gector–grammatical error correction: tag, not rewrite.In Proceedings ofBEA@ACL, pages 163–170.

[12] Yuan Z, Bryant C. Document-level grammatical error correction[C]//Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications. 2021: 75-84.

[13] Zhang Y, Zhang B, Li Z, et al. SynGEC: Syntax-Enhanced Grammatical Error Correction with a Tailored GEC-Oriented Parser[J]. arXiv preprint arXiv:2210.12484, 2022.

[14] Yuan Z, Taslimipoor S, Davis C, et al. Multi-class grammatical error detection for correction: A tale of two systems[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language

Processing. 2021: 8722-8736.

[15] Tarnavskyi M, Chernodub A, Omelianchuk K. Ensembling and Knowledge Distilling of Large Sequence Taggers for Grammatical Error Correction[J]. arXiv preprint arXiv:2203.13064, 2022.

[16] Stahlberg F, Kumar S. Synthetic data generation for grammatical error correction with tagged corruption models[J]. arXiv preprint arXiv:2105.13318, 2021.

[17] Edunov S, Ott M, Auli M, et al. Understanding back-translation at scale[J]. arXiv preprint arXiv:1808.09381, 2018.

[18] Tao Ge, Furu Wei, and M. Zhou. 2018. Fluency boost learning and inference for neural grammatical error correction. In ACL, pages 1055–1065.

[19] Lai S, Zhou Q, Zeng J, et al. Type-Driven Multi-Turn Corrections for Grammatical Error Correction[J]. arXiv preprint arXiv:2203.09136, 2022.

苏州大学本科生毕业设计（论文）中期进展情况检查表

学院（部）：计算机科学与技术学院

<table>
<tr><td colspan="2">学生姓名</td><td>乔子恒</td><td>年级</td><td>2019</td><td>专业</td><td>计算机科学与技术</td><td>填表日期</td><td>2023/4/12</td></tr>
<tr><td colspan="2">设计（论文）选题</td><td colspan="7">面向科技文献领域的文本纠错系统实现</td></tr>
<tr><td rowspan="2">已完成的任务</td><td colspan="8">学习前后端 flask+vue 技术，搭出系统框架。
阅读相关文献，完成外文翻译与文献综述。
通过爬虫技术获取大量科技文献论文，通过数据预处理手段清洗筛选出需要的数据。
学习纠错模型技术，学会使用 seq2edit 模型和 seq2seq 模型。
完善系统功能，实现用户注册登录、上传论文、查看纠错结果、更新个人信息等功能。</td></tr>
<tr><td colspan="2">是否符合任务书要求进度</td><td colspan="6">是</td></tr>
<tr><td rowspan="2">尚须完成的任务</td><td colspan="8">优化调整纠错模型，使之在科技文献数据上表现更好。
完善系统，添加更多功能。
撰写毕业论文。</td></tr>
<tr><td colspan="2">能否按期完成任务</td><td colspan="6">是</td></tr>
<tr><td rowspan="2">存在的问题和解决办法</td><td>存在的问题</td><td colspan="7">1、如何提高模型性能
2、纠错力度无法调整</td></tr>
<tr><td>拟采取的办法</td><td colspan="7">1、通过数据增强等手段，处理科技文献领域数据，训练纠错模型。
2、通过模型集成手段，牺牲纠错速度，提高纠错精度，同时提供纠错阈值的自定义调整。</td></tr>
<tr><td colspan="2">指导教师意见</td><td colspan="7">理清思路
签名：</td></tr>
<tr><td colspan="2">中期检查专家组意见</td><td colspan="7">同意
组长签名：</td></tr>
<tr><td colspan="2">学院（部）意见</td><td colspan="7">同意
教学院长签名：</td></tr>
</table>

检查日期：　2023 年 4 月 12 日

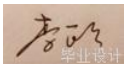
苏州大学本科生毕业设计（论文）答辩记录表

学院（部）：计算机科学与技术学院

学生姓名	乔子恒	年 级	2019	专 业	计算机科学与技术
设计（论文）题目	面向科技文献领域的文本纠错系统实现				
答辩时间	2023 年 5 月 17 日 13：30 至 17：00		答辩地点	理工楼 317	
答辩小组成员： 王宜怀，贾俊铖，周经亚，屈蕴茜，陈茜					
答辩中提出的主要问题及学生回答问题的简要情况： 1. 你是从什么时候开始做的这个工作？ 答：今年 2 月份开始的。 2. 论文的工作专门面对科技文献，在这方面做了什么特殊的措施吗？ 答：是的。模型的训练使用了科技文献领域内数据，该数据是从 CCF 推荐的中文 ABC 类期刊爬取获得的。我将 CCF 数据与二语数据集结合训练，让模型在领域内的纠错表现更好。 3. 你的 PPT 中展示了 Seq2Seq 和 Seq2Edit 模型，主要是基于这两个模型做的吗？ 答：是的，采用了这两类模型，并使用了模型集成技术。其中我主要采用了 Seq2Edit 模型，领域移植实验也是使用 Seq2Edit 来做的。。 答辩小组组长签名： 答辩小组成员签名：，，， 记录人签名： 2023 年 5 月 17 日					

苏州大学本科生毕业设计（论文）成绩评定表

学院（部）：计算机科学与技术学院

<table>
<tr><td colspan="5">设计（论文）题目：面向科技文献领域的文本纠错系统实现</td></tr>
<tr><td>姓 名</td><td>乔子恒</td><td>学 号</td><td colspan="2">1927405039</td></tr>
<tr><td>年 级</td><td>2019</td><td>专 业</td><td colspan="2">计算机科学与技术</td></tr>
<tr><td rowspan="6">指导教师评语</td><td colspan="4">乔子恒同学的毕业内容是实现一个科技文献领域的文本纠错系统，包括 CCF 推荐期刊论文爬取、文本内容提取、数据增强、模型训练、模型集成、可视化系统构建等模块。乔子恒同学学习态度很好、做事情有责任心，努力完成每一项任务，比较完满地完成了预定目标。论文书写规范、认真。</td></tr>
<tr><td colspan="3">评 价 内 容</td><td>得 分</td></tr>
<tr><td colspan="3">设计（论文）方案设计、文献检索、阅读及综述能力、进度等情况评价分（计 25 分）</td><td>22</td></tr>
<tr><td colspan="3">毕业设计（论文）质量和工作量评价分（计 50 分）</td><td>47</td></tr>
<tr><td colspan="3">科学素养、学习态度、纪律表现等情况评价分（计 25 分）</td><td>24</td></tr>
<tr><td colspan="4">成绩（满分 100 分）：93 签名： 2023 年 5 月 15 日</td></tr>
</table>

<table>
<tr><td>评阅教师评语</td><td>该文实现了科技文献领域的文本纠错系统，分析了针对当前科技文献相关文本纠错系统的用户需求，爬取科技文献数据，实现文本纠错可视化操作。该文选题合理，内容完整，层次结构安排科学，主要观点突出，逻辑关系清晰明了，格式规范，系统实现充分完备，工作量充足充分。该文对纠错系统的演示完备，纠错操作流程明了，符合毕业论文要求。

<table>
<tr><td>评 价 内 容</td><td>得 分</td></tr>
<tr><td>毕业设计（论文）文字书写评价分（计 20 分）</td><td>17</td></tr>
<tr><td>毕业设计（论文）质量评价分（计 40 分）</td><td>36</td></tr>
<tr><td>工作量情况评价分（计 20 分）</td><td>19</td></tr>
<tr><td>毕业设计（论文）创新及分析问题、解决问题能力评价分（计 20 分）</td><td>17</td></tr>
</table>
成绩（满分 100 分）：89 签名： 2023 年 5 月 16 日</td></tr>
<tr><td>答辩小组评语</td><td>该生毕业设计完成了一个面向科技文献领域的文本纠错系统实现，较好地完成了毕业设计任务书规定的内容和要求，反映该生能较好地运用所学的理论知识解决设计中的问题，具有一定的独立工作能力。该生能在规定的时间内比较流利和清晰地阐述论文的主要内容，能恰当回答与论文有关的问题。答辩小组经过充分讨论并根据该生论文质量和答辩中的表现，一致认为该生毕业论文成绩达到良好。

<table>
<tr><td>评 价 内 容</td><td>得 分</td></tr>
<tr><td>毕业设计（论文）介绍表达情况评价分（计 20 分）</td><td>15</td></tr>
<tr><td>回答问题表现评价分（计 40 分）</td><td>34</td></tr>
<tr><td>毕业设计（论文）水平和工作量评价分（计 40 分）</td><td>35</td></tr>
</table>
成绩（总分 100 分）：84 组长签名： 2023 年 5 月 17 日</td></tr>
<tr><td colspan="2">按权重折算总成绩 88</td></tr>
<tr><td colspan="2">审定成绩：88 答辩委员会主任签名：
日 期：2023 年 5 月 17 日</td></tr>
</table>

文本复制检测报告单

№:ADBD2023R_20181127141145202305261618518143620409O5

检测文献： 面向科技文献领域的文本纠错系统实现

作者： 乔子恒

检测范围：
中国学术期刊网络出版总库
中国博士学位论文全文数据库/中国优秀硕士学位论文全文数据库
中国重要会议论文全文数据库
中国重要报纸全文数据库
中国专利全文数据库
图书资源
优先出版文献库
大学生论文联合比对库
互联网资源(包含贴吧等论坛资源)
英文数据库(涵盖期刊、博硕、会议的英文数据以及德国Springer、英国Taylor&Francis 期刊数据库等)
港澳台学术文献库
互联网文档资源
CNKI大成编客-原创作品库

时间范围： 1900-01-01至2023-05-26

检测时间： 2023-05-26 16:18:51

总文字复制比：5%

去除引用：3.6% **去除本人：5%** **重合字数：1470** **文献总字数：29672**

单篇最大文字复制比：0.9% **单篇最大重复字数：264**

总段落数： [8] 疑似段落数： [7] 疑似段落最大重合字数： [432]

前部重合字数： [301] 后部重合字数： [1169] 疑似段落最小重合字数： [32]

6.5%	中英文摘要等(总4039字)
1.8%	第一章绪论(总2054字)
14%	第二章系统相关技术介绍(总3094字)
3.6%	第三章纠错相关技术介绍(总7126字)
6.5%	第四章系统分析与设计(总4461字)
2.3%	第五章系统模块详细设计与实现(总6993字)
3.8%	第六章系统测试(总842字)
0%	第七章总结与展望(总1063字)

中国知网文献检测报告单（机构版）