论文题目: 中文文本自动校对系统设计与实现

学位类别： 工程硕士

学科专业： 软件工程

年 级： 2014级

研 究 生： 张 涛

指导教师： 李天瑞

贾 真

二零一七年五月

国内图书分类号：TP391.1 密级：公开

国际图书分类号：004

**西 南 交 通 大 学**

**研 究 生 学 位 论 文**

中文文本自动校对系统设计与实现

年 级 二○一四级

姓 名 张 涛

申请学位级别 硕 士

专 业 软件工程

指 导 老 师 李 天 瑞

副指导 老 师 贾 真

二○一七年五月十六日

Classified Index: TP391.1

U.D.C:004

Southwest Jiaotong University

Master Degree Thesis

DESIGN AND IMPLEMENTATION OF CHINESE TEXT AUTOMATIC PROOFREADING SYSTEM

Grade: 2014

Candidate: Zhang Tao

Academic Degree Applied for: Master Degree

Speciality: Software Engineering

Supervisor: Li Tianrui

Co-Supervisor: Jia Zhen

May. 16, 2017

**西南交通大学**

**学位论文版权使用授权书**

本学位论文作者完全了解学校有关保留、使用学位论文的规定，同意学校保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。本人授权西南交通大学可以将本论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复印手段保存和汇编本学位论文。

本学位论文属于

1．保密□，在 年解密后适用本授权书；

2．不保密√，使用本授权书。

（请在以上方框内打“√”）

学位论文作者签名： 指导老师签名：

日期：2016年 月 日 日期：2016年 月 日

西南交通大学硕士学位论文主要工作（贡献）声明

本人在学位论文中所做的主要工作或贡献如下：

1. 针对中文文本自动校对系统展开系统需求分析，详细分析系统的用户需求，业务需求以及功能需求。

2. 用MVC框架对系统框架进行设计，详细的分析系统中每一层所包含的功能，同时分析和设计了审校服务的具体功能。

3. 实现中文文本自动校对系统的后台功能，其中包括采用CRF结合分词的技术识别文本中的实体,并采用实体链接的技术来实现词语审校中的实体名称审校，而词语审校中的常用词语审校即敏感词审校则采用构建字典树的方式来实现；针对标点符号和数字审校主要采用的是构建规则库来实现；而对于拼音审校本文利用pinyin4j工具包来获取正确词的拼音，再与原文中的标注的拼音来进行审校的。

4. 利用Spring MVC框架来实现中文文本自动校对系统的Web框架。其中包括前台Jsp界面的编写，控制层逻辑的实现。

本人郑重声明：所呈交的学位论文，是在导师指导下独立进行研究工作所得的成果。除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的研究成果。对本文的研究做出贡献的个人和集体，均已在文中作了明确说明。本人完全了解违反上述声明所引起的一切法律责任将由本人承担。

学位论文作者签名：

日期：

# 摘 要

互联网信息快速增长给人们生活带来了丰富的信息，但同时也造成了互联网信息质量的下降。新闻出版及广播电视等行业对文本质量有较高的要求，而这些行业的审稿工作目前仍以人工审校为主。以中文文本为例，在中文词语、汉语拼音、数字号码、标点符号等方面均存在或多或少的问题。随着因特网上中文文本的快速增长，相关错误在不断累积，这极大降低了文本的利用价值，同时加重了人工审校的负担。此外，中文载体的多样性也使得人工审校愈发力不从心。然而现有的文本审校软件面对不同形式、不同格式、不同载体的审校问题时显得捉襟见肘。以上现状表明，应对中文文本的自动审校、智能审校给予足够的重视。

基于以上背景，本文针对中文文本常见的错误构建了一个自动校对系统，并从以下几个方面展开了工作。

1. 对中文文本自动校对系统展开系统需求分析，详细分析系统的用户需求，业务需求以及功能需求。

2. 采用MVC框架对审校系统进行设计，详细地分析并设计了每一层所包含的功能，同时分析和设计了审校服务的具体功能，具体包括词语审校、标点符号审校数、数字审校和拼音审校。

3. 深入研究了词语审校、标点符号审校、数字审校以及拼音审校的实现方法，采用CRF结合分词的技术识别文本中的实体，并采用实体链接的技术来实现词语审校中的实体名称审校，而词语审校中的常用词语审校和敏感词审校则采用构建字典树的方式来实现；针对标点符号和数字审校主要采用的是构建规则库来实现；对于拼音审校，本文先利用Pinyin4j工具包获取带有标注拼音词语的正确拼音，然后再与原文中的标注拼音进行比较，若二者不同则对原文中的标注拼音进行校对。

4. 实现B/S模式的中文文本自动校对系统。本文采用的Spring MVC框架来实现该系统的Web框架，包括Spring MVC框架的配置，前台Jsp的编写以及控制器程序的编写等。

关键词：文本审校；命名实体识别；命名实体链接；字典树；Spring MVC

# Abstract

The rapid growth of Internet information has brought people's lives a wealth of information, but at the same time also decreased the quality of the Internet information. But the News publishing and television broadcasting industry have higher requirements for quality of the text, and the review work of those industries is still manually. For example (in Chinese), there are more or less problems in Chinese words, pinyin, numbers, punctuation, etc. And these errors are accumulating with the rapid increase in the number of Chinese network text. This makes artificial revisions more and more difficult, and it also reducing the value of the use of text at the same time. In addition, there are various forms of Chinese carrier greatly increased the burden of manual review. The existing text proofreading software is difficult to deal with various forms, different formats, different carrier text. we must give sufficient attention to the automatic and intelligent proofreading of the Chinese text because of it’s present situation.

Based on the above background, this thesis constructs a Chinese text automatic proofreading system for the common mistakes in Chinese text, and the work is carried out in the following aspects:

(1) The system requirements analysis of Chinese text automatic Proofreading system is carried out. We make a detailed analysis on user requirements, business requirements and functional requirements for this system.

(2) The MVC framework is used to design the system frame, and the functions of each layer in the system are analyzed in detail. The specific functions of the proofreading service are also analyzed and designed in detail, including the word proofreading, the punctuation proofreading, the digital proofreading and the phonetic proofreading.

(3) The method of the word proofreading, the punctuation proofreading, the digital proofreading and the pinyin proofreading is studied in this thesis. The named entity recognition and named entity linking are used to proofreading the name of named entity, and we used CRF and word segmentation technology in named entity recognition. And we build the trie tree to proofread the common word and the sensitive word. For punctuation and digital proofreading is mainly achieved by building the rule base. For phonetic proofreading, we use the Pinyin4j toolkit to get the correct pinyin for the word， and then compare it to the marked pinyin of the word in the text. If it is different from the marked pinyin, then use it to proofread the marked pinyin.

(4) To achieve the chinese text automatic proofreading system based on B/S model. We used the Spring MVC framework to achieve the web framework of this system. Including the configuration of the Spring MVC framework, the programing of the foreground Jsp and the controller.

**Key words:** text proofreading; Named entity recognization; Named entity linking; tire tree; Spring MVC

目 录

[摘 要 I](#_Toc478044826)

[Abstract II](#_Toc478044827)

[第1章 绪 论 1](#_Toc478044828)

[1.1 论文的研究背景及意义 1](#_Toc478044829)

[1.2 国内外研究现状 2](#_Toc478044830)

[1.3 本文的研究内容及结构安排 4](#_Toc478044831)

[1.3.1 本文的研究内容 4](#_Toc478044832)

[1.3.2 论文的结构安排 5](#_Toc478044833)

[第2章 相关技术研究 7](#_Toc478044834)

[2.1 条件随机场简介 7](#_Toc478044835)

[2.2 自然语言相关技术介绍 9](#_Toc478044836)

[2.2.1 分词技术介绍 9](#_Toc478044837)

[2.2.2 命名实体识别技术介绍 11](#_Toc478044838)

[2.3 Spring MVC框架介绍 12](#_Toc478044839)

[2.4 本章小结 14](#_Toc478044840)

[第3章 需求分析及设计 15](#_Toc478044841)

[3.1 系统需求分析 15](#_Toc478044842)

[3.1.1 用户需求 15](#_Toc478044843)

[3.1.2 业务需求 15](#_Toc478044844)

[3.1.3 功能需求 15](#_Toc478044845)

[3.2 系统设计 18](#_Toc478044846)

[3.2.1 系统体系结构设计 18](#_Toc478044847)

[3.2.2 审校服务功能设计 19](#_Toc478044848)

[3.3 本章小结 20](#_Toc478044849)

[第4章 中文文本审校功能 21](#_Toc478044850)

[4.1 词语审校 21](#_Toc478044851)

[4.1.1 实体名称审校 21](#_Toc478044852)

[4.1.2 常用词语审校 29](#_Toc478044853)

[4.1.3 敏感词检测 35](#_Toc478044854)

[4.2 标点符号审校 36](#_Toc478044855)

[4.2.1 标点符号错误类型介绍 36](#_Toc478044856)

[4.2.2 标点符号审校 38](#_Toc478044857)

[4.3 数字审校 39](#_Toc478044858)

[4.3.1 数字审校概述 39](#_Toc478044859)

[4.3.2 数字审校流程 39](#_Toc478044860)

[4.4 拼音审校 41](#_Toc478044861)

[4.4.1 拼音审校概述 41](#_Toc478044862)

[4.4.2 拼音审校流程 41](#_Toc478044863)

[4.5 本章小结 43](#_Toc478044864)

[第5章 系统实现与演示 44](#_Toc478044865)

[5.1 前言 44](#_Toc478044866)

[5.2 系统框架 44](#_Toc478044867)

[5.3 系统功能展示 45](#_Toc478044868)

[5.2.1 注册 45](#_Toc478044869)

[5.2.2 登录 46](#_Toc478044870)

[5.2.3 个人信息维护 47](#_Toc478044871)

[5.2.4 在线审校 47](#_Toc478044872)

[5.2.5 离线审校 48](#_Toc478044873)

[5.3 本章小节 49](#_Toc478044874)

[结 论 50](#_Toc478044875)

[本文工作总结 50](#_Toc478044876)

[未来工作展望 50](#_Toc478044877)

[致 谢 51](#_Toc478044878)

[参考文献 52](#_Toc478044879)

[攻读硕士学位期间发表的论文及科研成果 56](#_Toc478044880)

# 第1章 绪 论

## 1.1 论文的研究背景及意义

语言是[人类](http://baike.baidu.com/view/14713.htm)区别其他动物的本质特性。在所有生物中，只有人类才具有语言能力。人类的多种[智能](http://baike.baidu.com/view/375267.htm)都与语言有着密切的关系。人类的逻辑思维以语言为形式，人类的绝大部分知识也是以语言文字的形式记载和流传下来的。因而，它也是[人工智能](http://baike.baidu.com/view/2949.htm)的一个重要，甚至核心部分。用自然语言与[计算机](http://baike.baidu.com/view/3314.htm)进行通信，这是人们长期以来所追求的。因为它既有明显的实际意义，同时也有重要的理论意义：人们可以用自己最习惯的语言来使用计算机，而无需再花大量的时间和精力去学习不符合自然和习惯的各种计算机语言；人们也可通过它进一步了解人类的语言能力和智能的机制。实现人机间自然语言通信意味着要使计算机既能理解自然语言文本的意义，也能以自然语言文本来表达给定的意图、思想等。前者称为[自然语言理解](http://baike.baidu.com/view/464346.htm)，后者称为[自然语言生成](http://baike.baidu.com/view/3458345.htm)。因此，自然语言处理大体包括了自然语言理解和自然语言生成两个部分。历史上对自然语言理解研究得较多，而对自然语言生成研究得较少。但这种状况已有所改变。无论实现自然语言理解，还是自然语言生成，都远不如人们原来想象的那么简单，而是十分困难的。从现有的理论和技术现状看，通用的、高质量的自然语言处理系统，仍然是较长期的努力目标。自然语言处理有着广泛的应用，如机器翻译，全文信息检索，自动文摘生成等。其中中文文本处理就是其中一个重要的应用。

互联网是历史上发展最快的一种电子通信技术，互联网信息的加速爆炸使得信息时代逐步转变为大数据时代，从面向计算走向面向数据。据报告，到2020年，中国的数据量将突破8ZB，是2012年的23倍；同期，到2020年，全球数据量将突破40ZB，是2012年的40倍[1]。2016年度发布的《中国语言生活状况报告》也越来越关注网络文本及外来语。但不可否认的是，互联网信息泛滥、价值密度低大大制约了互联网信息的使用。新闻出版及广播电视等行业对文本审校有较高的要求，这也是语言文字工作的重点，但此方面的工作目前仍以人工审校为主。以中文文本为例，在中文文字、汉语拼音、中文词语、数字号码、标点符号等方面均存在或多或少的问题，使得人工审校日益显得力不从心，同时大大降低了文本的利用价值。此外，多种形式的中文载体的存在大大加重了人工审校的负担。现有的文本审校软件面对各种不同形式、不同格式、不同载体的文本审校时显得捉襟见肘，导致互联网信息的质量江河日下，出现类似于“劣币驱逐良币”的“劣文驱逐良文”现象。在专门的中文编辑审校网上对中文字词、句章等错误有较为全面的阐述。在部分网络文本中甚至直接提供有“纠错”的链接，将文本审校留给读者。中文网络文本的现状使得现今不得不对中文网络文本的自动审校、智能审校给予足够的重视。

文本校对在很多方面都有着重要的应用，但在中文领域，目前还没有开放的中文文本校对系统。如何构建一个能够针对文本中出现词语错误进行自动审校的系统是一个很有意义的课题。

## 1.2 国内外研究现状

审校是出版编辑过程里的一个必须工序，在很大程度上是一项智力活动。尽管已跨入后信息时代，但现阶段的审校工作仍主要以人工审校为主，导致宝贵的人力资源耗费在低水平的文字审阅上。计算机的广泛应用催生了文本审校系统，现有的中文文本审校系统主要有校对助手、校对能手、黑马校对等。由于Microsoft Office办公软件在办公领域的支配地位，Office 2013提供有中文文字校对工具Office Proofing Tools 2013。

从现有的文本审校系统来看，基本上所有的审校软件均能解决文字错误，属于较为简单的审校，对于词语审校则较为困难，实际上，大部分审校软件均未能很好的解决词语审校，这就解释了在互联网信息泛滥的时代，网络文本中词语由于音近、形近、义近等因素而导致的词语误用远高于文字误用。更高层次的语义审校更是漏洞百出。同时，现有的审校软件均基于单机模式或多机模式，尚未有基于B/S模式的审校软件，而且对不同形式文本的支持也不够，基于B/S模式且支持不同形式文本的文本审校软件是文本审校软件的发展趋势。

此外，现有的文本审校系统对文字给予了足够的关注，但却忽视了同样重要的标点符号。标点符号的混用、误用、冗余、遗漏等因素导致全角标点与半角标点混用、中文符号与外文符号混用、顿号与逗号误用、句号冗余、句号遗漏等错误频出，文本审校软件必须对标点符号审校基于更多的关注。在设计开发新的文本审校软件之间，先简要分析几款应用较广的文本审校软件以明确设计开发的重点与要点。

本节从已有的审校系统和针对文本审校方法研究两个方面进行国内外现状分析。

(1) 文本审校方法研究：

国外对英文自动校对技术的研究起步比较早，早在20世纪60年代，国外的研究机构就开始了基于英文文本的自动校对技术研究。经过多年的努力，已经取得了一些研究成果。譬如，IBM Thomas J. Watson 研究中心首先于1960年在IBM/360和IBM/370上用UNIX实现了一个TYPO英文拼写检查器；1971年，斯坦福大学的Ralph Gorin 在DEC-10机上实现了一个英文拼写检查程序spell[2]。Karen Kukich将英文文本检错总结归纳为以下三个问题[3]：

（1）非词错误检测，因为在英文文本中存在单词拼写错误的情况，所以针对英文文本纠错其中一个问题就是检测英文文本中存在的单词错误；

（2）独立词错误的校对，检测到错误单词之后，如何选择正确的单词对文本中错误的单词进行校对也是一个困扰校对研究者的问题；

（3）上下文词语纠错，在英文文本中除了单个词的错误情况，还会存在上下文单词搭配错误的情况，这就是语法纠错级别的问题，也是文本校对最困难的问题所在。

近年来，特别是进入21世纪后，随着新技术如春笋般不断涌现，如语音识别技术、OCR识别技术等。不过，这些效率很高的新技术，也并不能保证其识别结果100%正确，因此对识别后文档中可能存在的错误进行校对就显得更加迫切和必要，各方面对此开展的研究工作也在不断进行，并且取得了一些可喜的成果[4-10]。

国内关于文本自动校对的研究开始得比较晚，开始于20世纪90年代。研究的重点也大多数都是针对中文文本的校对技术。1998年于勐和姚天顺两人以模式匹配的方法和3元文法分析的方法为基础，结合语法属性标注和分析的方法，提出了一种混合的中文文本校对方法，其结果优于任何一种方法的单独应用。该方法建立于错误模式匹配和3元文法分析的基础之上，又加入了基于语法属性标注的语法分析方法[11]。2001年张磊等人在《中文文本自动校对》一文中详细介绍了中文文本自动校对的研究现状,包括文本中错误的分类，中文自动校对中常用的语言模型[12]。2003年龚小谨等人在《中文文本自动校对中的语法错误检查》[14]一文中提出了检查文本自动校对中出现的语法错误的方法。他们将中文文本的语法错误分为搭配错误和与句型成分相关的错误两大类。分别采用模式匹配的方法和基于句型成分分析的方法进行检查，这两种方法的结合，可以同时考虑局部和全局的语法限制信息，并且降低了语法检查的复杂度。同年他们又提出了检查文本自动校对过程中出现的语义错误的方法。他们讨论了中文文本自动校对中的语义错误校对技术，综合使用了基于实例、基于统计和基于规则的搭配关系进行检查，提出统计和规则相结合的校对策略，既能检查局部语义限制，也能检查长距离的语义搭配，收到了较好的效果，也为中文文本自动校对的发展提供了新的思路[15]。2004年骆卫华等人提出了词法、句法、语义多层次结合校对策略，从而能够检查以往无法查出的错误。描述了自动校对系统的整体框架，并具体给出可操作的实现方法。这些方法针对非受限领域的文本，为中文自动校对技术的发展提供了新的思路[16]；2014年南京大学的王思雨等人结合CSSCI数据库的实际情况，构建了一个基于混淆集和上下文特征分析的文本自动校对系统模型[17]。

(2) 已有的中文文本审校系统：

校对助手是一款文本文件的辅助审校工具软件，它界面简洁，功能集中，审校智能。用户可自行定义审校词库，并支持词库更新。审校词汇基本库容纳了一千余条常见易错词组词条，特别适用于文档的初级审校，审校速度快；高级词库包含了一些相对罕见或使用率较低的易错词汇，审校更为全面彻底；专业词库则面向不同专业领域文稿，审校结果更为精准。审校发现的可能错误均列表显示，用户可根据具体语境自主决定是否进行纠正。该软件可以用于文本文件(\*.txt)、方正书版文件(\*.fbd)及网页文件(\*.htm,\*.html)内容的直接审校。

校对能手又名编辑助手，是一款中文文稿审校软件。该软件针对文稿的起草、修改、抄写、打字、扫描、排版等过程中出现的各种差错，检查出可能存在的差错，可以快速对中文文档资料进行审校，列出勘误表，供改稿时参考。此外，该软件还有姓名筛查功能、人名地名索引自动编排功能、字典功能、中文排序功能等，特别适用于快速审校文稿和辅助大型图书编辑工作。

黑马校对是北京黑马公司开发的新一代审校系统，其内含多版本，包括S2、PS、Word、WPS、小样版、飞腾插件版和PDF插件版九个校对界面。黑马校对采用超大规模词库和重点词监控等先进的审校计算技术，在审校质量、审校功能和易用性等方面都有了飞跃性的提高，达到了当前电脑审校软件发展的最高水平，智能技术居于国际领先水平。但是黑马校对目前收费较高，并没有免费对外开放。

不可否认的是，上述三款软件确实对文本审校起到较大的助益，在一定程度上大大降低了人工审校的工作量，但仍存在一些不足。实际上，上述三款审校软件均不能很好的解决中文编辑审校网上列举的各类差错。包括上述三款文本审校软件在内的大部分中文文本审校软件主要存在下述问题：

1、基于单机模式开发，制约了其使用范围。几乎所有的审校软件均是基于单机模式或多机模式开发，必须先安装才能使用，而基于浏览器的B/S 模式是软件应用的一大趋势，应考虑基于B/S 架构的文本审校软件的开发与使用；

2、以文本为审校的基本单位，不支持大批量的文本审校。包括黑马校对在内的审校软件在很大程度上仍以单个的文本为审校的基本单位，对文件的支持不够，在很多方面带有人工审校的残余，应采取措施解决大批量文本的审校问题；

3、更新滞后，大部分不更新或只提供很少的更新。黑马校对已升级到V18，校对能手已升级到V13，但为了与时俱进，二者对文件更新都格外关注，将维护工作转移给用户，相比之下，Web Service 式的文本审校服务更值得重视。通过对上述三款应用较广的文本审校软件进行分析发现，基于B/S 架构模式设计开发、支持不同形式、不同格式、不同载体的大批量文本审校及以Web Service 提供审校服务形式的文本审校软件在文本审校中有较为广阔的应用前景，新开发的文本审校软件应从B/S 架构、大批量、Web Service 三方面入手。

## 1.3 本文的研究内容及结构安排

### 1.3.1 本文的研究内容

本学位论文主要研究内容如下：

（1）对中文文本自动校对系统展开系统需求分析，详细分析系统的用户需求，业务需求以及功能需求。

（）用MVC框架对系统框架进行设计，详细地分析系统中每一层所包含的功能，同时分析并设计了审校服务的具体功能，其中包括词语审校、标点符号审校、数字审校和拼音审校。

（3）深入研究了词语审校、标点符号审校、数字审校以及拼音审校的实现方法，采用CRF加分词的技术识别文本中的实体并采用实体链接的技术来实现词语审校中的实体名称审校，而词语审校中的常用词语审校和敏感词审校则采用构建字典树的方式来实现；针对标点符号和数字审校主要采用的是构建规则库来实现；而对于拼音审校本文利用Pinyin4j工具包被标注拼音的词语的正确拼音，再与原文中的标注的拼音来进行比较，若二者不同再对原文中的标注拼音进行校对。

（4）实现B/S模式的中文文本自动校对系统。本文采用的Spring MVC框架来实现该系统的web框架，包括Spring MVC框架的配置，前台Jsp的编写以及控制器程序的编写等。

### 1.3.2 论文的结构安排

本学位论文主要研究中文文本自动校对系统的问题，共分为五章，内容主要包括：

第1章：介绍中文文本自动校对系统的研究背景与研究意义，分析了文文本自动校对的相关技术与研究现状。

第2章：介绍了本文所用的相关技术。主要介绍了条件随机场（CRF）、分词、实体识别、实体链接以及Spring MVC。

第3章：从用户需求、业务需求和功能需求三个角度分析了该系统的需求，紧接着从MVC三层架构分析了该系统的体系结构设计，最后对审校服务的内部功能进行了设计。

第4章：重点介绍了审校系统中的词语审校功能的实现算法和流程，从实体名称审校、常用词语审校以及敏感词检测三个方面介绍了词语审校，其中实体名称审校采用的是实体链接的方法来实现的，而常用词语审校和敏感词检测则是采用构建字典树来实现的，二者区别在于常用词语审校则是采用模糊匹配的方法来查找词库，而敏感词则是采用精确查找的方式来查找词库，而且二者所加载的词库也不相同。本章还介绍了标点符号、数字审校和拼音审校的方法，对于这三类错误，本文采取的是构建正则表达式的方法来实现的。

第5章：主要介绍了如何使用Spring MVC搭建中文文本自动校对系统的Web框架，并详细地介绍了注册、登录、个人信息维护、在线审校以及离线审校的操作流程。重点介绍了每个功能的操作流程。

结论部分论文总结了主要研究成果，提出了下一步的研究方向。

# 第2章 相关技术研究

## 2.1 条件随机场简介

条件随机场（Conditional Random Fields，CRFs）是J.Lafferty[19]等人在2001年提出来的，是一种常用于模式识别和机器学习的统计建模的方法，主要用于标注和划分序列结构数据。普通的分类器在预测单个样本的标签时不会考虑相邻的样本，而CRF则会考虑上下文。CRF是一种无向图模型，它结合了最大熵模型和隐马尔可夫模型的特点，近几年来在自然语言处理和图像处理等领域得到了广泛的应用[20-25]。CRF在自然语言处理领域主要用于分词、词性标注和命名实体识别等序列标注性的任务，并且已经取得了很好的效果。

CRF是一种用来标注和划分序列结构数据的概率结构模型，假设*X*是一组随机的观测序列，并且*Y*是一组相对应的标志序列，则对任意都有，其中是一组特定的标志序列集。例如,若X为一组自然语言句子的词序列，则Y只能是所有词性集中的一部分。条件随机场是通过条件概率来描述其概率模型，而非用联合概率分布来描述其概率模型。

CRF的定义如下：假设是一个无向图结构，*V*为所有顶点的集合，*E*为所有的边的集合，则有，即每一个随机变量*Yv*对应于V中的一个顶点，而*Yv*的取值范围可能是标志集合。在这种情况下就是一个条件随机场，当条件为观测序列X时，则对于每一个随机变量*Yv* 都会满足如下的马尔科夫特性：

 (2-1)

其中，表示*w*和*v*是图G中的两个临近的节点。

理论上，若标记序列Y中的所有*Yv*之间有一定的条件独立性，则图G的结构可以是任意的。这样就可以用最简单、最普通的链式结构图来描述标记序列的概率模型，节点对应标记序列Y中的元素，如图2-1所示或图2-2所示。



图2-1 CRF的链式结构图



图2-2 CRF链式结构图的另一种表示

所有观测序列X中的各个元素之间并不存在图结构，而只是将其作为条件，对其中的各个元素并不做任何独立性假设。

J.Lafferty等人定义当观测序列X给定时，则标记序列Y的概率为：

 (2-2)

其中，为转移函数，表示观测序列X对应的标注序列在*i*及*i-*1位置上标记的转移概率；而则是状态函数，表示的是观测序列X的第i个位置上的标记概率；和分别对应与和的权重，这两个权重是需要从样本训练中估计出来的。

在定义特征函数时，可以定义一组观测序列的{0,1}分布特征来表示训练样本中的某些分布特性，如：



转移函数的形式可以定义为如下形式：



为了描述起来方便，可以将状态函数的形式书写为：

 (2-3)

如此，特征函数则可以统一表示为：

 (2-4)

式2-4中的每一个局部特征函数表示状态特征或者转移函数。所以对于一个给定的观测序列对应的标记序列的条件随机场的条件概率如下：

 (2-5)

其中，分母*Z(X)*是归一化因子，其具体形式如式2-6所示：

 (2-6)

以上就是条件随机场的算法的基本介绍和理论公式。

## 2.2 自然语言相关技术介绍

### 2.2.1 分词技术介绍

众所周知，词是能够独立运用的最小的语言单位。但是汉语并不像英文那般词与词之间存在天然的空格。在汉语文本中所有的词与词是紧紧粘合在一起的，并没有任何空格之类的显示标志指示词的边界。因而，如何将汉语文本进行自动分词就成了计算机处理汉语文本的一个首要的基础工作了，同样也是诸多汉语应用系统不可或缺的一个重要的环节。在过去的20多年，国内外有众多学者在这一领域做了大量的研究工作，并且也取得了比较好的成果[26-35]。

1、汉语分词中的基本问题

汉语自动分词其实就是让计算机系统在汉语文本的词与词之间自动添加空格与其他边界之类的标记。但是这样一个看似非常简单的问题，却困扰了几代汉语自然语言研究者。汉语自动分词的难点可以归结为以下几点：

（1）汉语分词规范

刘开瑛[36]在2000年曾指出过，“词”这个概念一直是汉语语言学界纠缠不清但是又没法抛开的问题。关于什么是中文的词，到目前为止仍然没有一个能够被广泛接受的权威性定义，更没有什么权威性的词表。其主要困难来自两个方面：一个是单个词与词素之间划分不清楚；还有一个就是词与短语之间的划界不清楚。此外，对于汉语的“词”的认识，我们普通人与语言学专家的标准又有很大的区别。有关专家的调查结果表明，那些母语为汉语的测试者之间，对汉语文本中出现的词语的认同率只有大约70%，所以从严格意义上来说，自动分词是一个没有明确定义的问题[37]。

（2）歧义切分

歧义是指同样一句话可能会存在两种或者多种切分方法。歧义是汉语文本中普遍存在的问题，因此如何切分汉语文本中的歧义句子也是汉语自动分词的一个研究重点。汉语文本中的歧义大致可以分为两类：

a）交集性歧义：所谓交集性歧义指的是若AB为一个词，且BC为一个词时，当ABC组成一个汉字串时就形成了交集型歧义。例如“结合成”，一种是切分为“结合| 成”，而另一种切分为“结| 合成”。这种类似的词还有很多，可以说是普遍存在于汉语文本中。如“发展中国家”、“从小学起”、“部分居民生活水平”等等。

b）组合型歧义：所谓的组合型歧义是指若一个词AB其内部结构A、B也都能各自为词的结构。例如“起身”，根据不同的语境就有不同的划分方法，如“站| 起| 身| 来”，“起身| 去| 上海”。类似的还有“将来”、“才能”、“现在”等等。

（3）未登录词

未登录词又称生词，一般指的是未收录在分词词表中的词，有时候也解释为在训练语料中未出现过的词。随着互联网的发展，像“当红炸子鸡”、“蓝瘦香菇”、“狗带”等这种类似的网络用语简直是层出不穷。这无形中也增加了自动分词的难度。对于大规模的真实文本来说，未登录词对于分词精度的影响远远超过了歧义切分。

2、汉语分词的方法

自汉语分词问题提出来至今，众多专家学者在这一领域付出了大量的努力和心血，并提出了很多分词的方法。汉语分词方法经历了几个阶段：最早大家提出的方法大多是基于字符串匹配的方法；但后面随着统计机器学习的发展，许多研究者提出了许多基于机器学习的分词方法；这几年随着深度学习成为研究的热门，又有许多研究者提出了一些基于深度学习的自动分词方法。所以分词方法大致可以分为基于字符串匹配的分词方法、基于传统统计方法的分词以及基于深度学习的分词方法。以下将分别介绍这几类分词方法。

（1）基于字符串匹配的分词方法

这种方法也被称作为机械分词方法，它是按照一定的策略将待分词的汉语语句与一个足够大的词库中的词条进行配，若在词库中找到某个字符串则表示匹配成功（即成功识别出一个词）。这种分词有多种匹配策略，根据扫描方向的不同，可以将串匹配分词方法分为正向匹配和逆向匹配；而根据长短优先原则，则又可将串匹配分词方法分为最长匹配原则和最短匹配原则；根据是否与词性标注过程相结合，又可以分为单纯分词方法和分词与标注相结合的一体化方法。常用的几种机械分词方法如下：正向最大匹配法（由左到右的方向）；逆向最大匹配法（由右到左的方向）；最少切分（使每一句中切出的词数最小）。

（2）基于机器学习的分词方法

一般来说，词是字的稳定的组合，因此在上下文中，相邻的字同时出现的次数越多，就越有可能构成一个词。因此字与字相邻共现的频率或概率能够较好的反映成词的可信度。可以对语料中相邻共现的各个字的组合的频度进行统计，计算它们的互现信息。定义两个字的互现信息，计算两个汉字X、Y的相邻共现概率。互现信息体现了汉字之间结合关系的紧密程度。当紧密程度高于某一个阈值时，便可认为此字组可能构成了一个词。这种方法只需对语料中的字组频度进行统计，不需要切分词典，因而又叫做无词典分词法或统计取词方法。但这种方法也有一定的局限性，会经常抽出一些共现频度高、但并不是词的常用字组，例如“这一”、“之一”、“有的”、“我的”、“许多 的”等，并且对常用词的识别精度差，时空开销大。实际应用的统计分词系统都要使用一部基本的分词词典（常用词词典）进行串匹配分词，同时使用统计方法识别一些新的词，即将串频统计和串匹配结合起来，既发挥匹配分词切分速度快、效率高的特点，又利用了无词典分词结合上下文识别生词、自动消除歧义的优点。到底哪种分词算法的准确度更高，目前并无定论。对于任何一个成熟的分词系统来说，不可能单独依靠某一种算法来实现，都需要综合不同的算法。目前这类分词方法有基于隐马尔可夫（HMM）的分词方法和基于条件随机场（CRF）的分词方法等等。

（2）基于深度学习的分词方法

在AlphaGo战胜世界围棋高手后，这两年Deep Learning（深度学习）的热度越来越高。深度学习来源于传统的神经网络模型。传统的神经网络一般由输入层、隐藏层和输出层组成，其中隐藏层的数目按需确定。深度学习可以简单地理解为多层神经网络，但是深度学习却不仅仅是神经网络。深度模型将每一层的输出作为下一层的输入特征，通过将底层的简单特征组合成为高层的更抽象的特征来进行学习。在训练过程中，通常采用贪婪算法，一层层地训练，比如在训练第k层时，固定训练好的前k-1层的参数进行训练，训练好第k层之后，以此类推进行一层层训练。目前已有学者将深度学习模型应用于汉语文本分词中去[38]。

### 2.2.2 命名实体识别技术介绍

命名实体识别也称作为专名识别，是指识别文本中特定意义的实体的过程，主要是人名、地名、机构名等专有名词。命名实体识别是自然语言处理的一项极其关键的技术，在自然语言领域有着广泛的应用。例如信息抽取，问答系统、句法分析、机器翻译等应用都离不开命名实体识别，命名实体识别在自然语言处理技术走向实用化的过程中占有特别重要的地位。

命名实体是命名实体识别的研究主体，一般包括3大类(实体类、时间类和数字类)和7小类(人名、地名、机构名、时间、日期、货币和百分比)命名实体。评判一个命名实体是否被正确识别包括两个方面：实体的边界是否正确；实体的类型是否标注正确。主要错误类型包括文本正确，类型可能错误；反之，文本边界错误,而其包含的主要实体词和词类标记可能正确。

命名实体识别的主要技术方法分为：基于规则和词典的方法、基于统计的方法、二者混合的方法等。

1. 基于规则和词典的方法

基于规则的方法多采用语言学专家手工构造规则模板，选用特征包括统计信息、标点符号、关键字、指示词和方向词、位置词(如尾字)、中心词等方法，以模式和字符串相匹配为主要手段，这类系统大多依赖于知识库和词典的建立。基于规则和词典的方法是命名实体识别中最早使用的方法，一般而言，当提取的规则能比较精确地反映语言现象时，基于规则的方法性能要优于基于统计的方法。但是这些规则往往依赖于具体语言、领域和文本风格，编制过程耗时且难以涵盖所有的语言现象，特别容易产生错误，系统可移植性不好，对于不同的系统需要语言学专家重新书写规则。基于规则的方法的另外一个缺点是代价太大，存在系统建设周期长、移植性差而且需要建立不同领域知识库作为辅助以提高系统识别能力等问题。

2. 基于统计的方法

基于统计[机器学习](http://lib.csdn.net/base/machinelearning)的方法主要包括：隐马尔可夫模型(Hidden Markov Model, HMM)、最大熵(Maxmium Entropy, ME)、支持向量机(Support Vector Machine, SVM)、条件随机场( Conditional Random Fields, CRF)等。

在这4种学习方法中，最大熵模型结构紧凑，具有较好的通用性，主要缺点是训练时间复杂性非常高，有时甚至导致训练代价难以承受，另外由于需要明确的归一化计算，导致开销比较大。而条件随机场为命名实体识别提供了一个特征灵活、全局最优的标注框架，但同时存在收敛速度慢、训练时间长的问题。一般说来，最大熵和支持向量机在正确率上要比HMM高一些，但是HMM在训练和识别时的速度要快一些，主要是由于在利用Viterbi[算法](http://lib.csdn.net/base/datastructure)求解命名实体类别序列的效率较高。隐马尔可夫模型更适用于一些对实时性有要求以及像信息检索这样需要处理大量文本的应用，如短文本命名实体识别。

基于统计的方法对特征选取的要求较高，需要从文本中选择对该项任务有影响的各种特征，并将这些特征加入到特征向量中。依据特定命名实体识别所面临的主要困难和所表现出的特性，考虑选择能有效反映该类实体特性的特征集合。主要做法是通过对训练语料所包含的语言信息进行统计和分析，从训练语料中挖掘出特征。有关特征可以分为具体的单词特征、上下文特征、词典及词性特征、停用词特征、核心词特征以及语义特征等。

基于统计的方法对语料库的依赖也比较大，而可以用来建设和评估命名实体识别系统的大规模通用语料库又比较少。

## 2.3 Spring MVC框架介绍

Spring最早是为了解决企业级应用开发而由Rod Johnson创建的一个开源框架。使用Spring可以让简单的Javabean实现之前只有企业级Javabean才能完成的事情。而且Spring并不仅仅局限于服务器端开发，任何Java引用都能在多个方面从Spring中获益。

Spring MVC则是Spring Frame work的后续产品，是一种用Java实现Web MVC设计模式的轻量级Web框架。Spring MVC框架是一种基于请求驱动的Web框架，并且还用了前端控制器模式。Spring MVC可以通过配置处理映射器、视图解析器等来实现你想要的功能。



图2-3 Spring MVC请求流程

如图2-3所示为Spring MVC的请求流程。首先用户从前台发送请求，此时请求会离开浏览器，并带有用户请求的内容。请求到达第一站即是Spring的前段控制器DispatcherServlet，而DispatcherServlet的任务就是将请求发给用于处理用户请求的控制器（controller）。在一个应用中有可能会存在多个控制器，所以DispatcherServlet需要查询处理映射器（handler mapping）才能知道将用户请求具体发送给那个控制器来处理。处理映射器会根据请求所带的URL信息来进行决策。一旦选择好了控制器，DispatcherServlet则会将请求发送给所选中的控制器。用户请求到了控制器后，请求就会卸下其所负载的信息，并等待控制器处理这些信息。控制器处理完成之后，通常会产生一些信息，这些信息需要返回给用户，并在前台页面上显示。这些信息被称之为模型，控制器将模型数据封装好了之后就会连同用户渲染输出的视图名称一起发回给DispatcherServlet，这样控制器就不会与某个特定的视图相耦合。传递给DispatcherServlet的视图名称并不能直接认为就是某个特定的Jsp，甚至并不能确定用户渲染的视图就是Jsp，控制器所传递的仅仅是一个视图逻辑名称而已。DispatcherServlet需要使用视图解析器（View Resolver）才能确定视图名具体匹配哪一个特定的视图。DispatcherServlet确定了哪个视图用来渲染模型数据后，就会将数据传递给视图，即交付模型数据。最后就是视图对模型数据进行渲染展示了。以上就是整个Spring MVC用户请求的整个流程。

## 2.4 本章小结

本章首先介绍了条件随机场（CRF）的基本知识及其原理，详细地介绍了CRF理论公式的推导过程。然后介绍了自然语言中两种常用的技术：中文自动分词和命名实体识别，分析了两项技术目前常用方法。最后本文还介绍了目前非常流行而且用得很广泛的MVC框架——Spring MVC，并详细地介绍了Spring MVC请求的流程。

# 第3章 需求分析及设计

## 3.1 系统需求分析

### 3.1.1 用户需求

中文文本审校系统目前只设计了一个用户角色，就是系统的会员用户。用户需要先注册成该系统的会员才能使用该系统的功能。该用户可以自己维护自己的个人信息。同时该用户还能够使用在线和离线两种方式使用该系统对文本进行审校。

### 3.1.2 业务需求

该系统主要包含两个业务，即在线审校和离线审校。以下就分别介绍这两个业务。

1 在线审校

用户可以使用在线审校校对小段文本。在线审校的优点是审校速度快，短时间内就能够返回审校结果，其缺点是在线审校只能校对短文本，有一定的字数限制。

2 离线审校

离线审校则与在线审校不同，离线审校可以允许用户上传Word文档，后台再对该文本进行校对，校对完成之后会在用户界面提示，最后用户再将校对好的Word文件下载下来。

### 3.1.3 功能需求

由用户需求和业务需求的分析可以知道该系统存在一个用户角色，即该系统的会员，用户首先需要注册成为该系统的会员。注册成功后用户可以修改自己的个人信息，同时还可以使用离线审校和在线审校的两种系统所提供的服务，离线审校需要用户先上传word文件然后才能进行审校，审校完成后还需要下载审校结果文件，审校结果文件仍然是word文件。根据这些信息可以得到该系统的用例图，如图3-1所示。以下就分别对这几个用例做出描述。

1 注册

（1）说明

若用户想要使用本系统，首先需要进行个人信息注册，只有注册为本系统的会员才能使用该系统。

（2）事件流

（a）用户通过浏览器进入该系统登录界面，选择注册，系统转到注册页面；

（b）用户填写用户名、密码以及确认密码，用户名为长度为6-20的字母、数字或下划线组成；

（c）系统验证用户名是否有效以及是否重复，同时还需要验证两次密码是否一致，若验证不通过则给出相关提示，直至用户名及密码通过检验；

（d）系统自动生成用户ID，并提示注册已成功，跳转至系统登录页面。

（3）前置条件

用户进入系统注册页面，并申请注册。

（4）后置条件

注册成功，系统添加用户个人信息至数据库。



图3-1 系统用例图

2 登录

（1）说明

会员需要提供正确的用户名及密码，才能登录该系统，使用该系统所提供的服务。

（2）事件流

（a）会员输入该系统的网址，进入登入页面；

（b）会员输入正确的用户名及密码；

（c）系统验证用户名及密码是否正确，若不正确则给出提示，若用户名密码都正确则跳转至该系统首页。

（3）前置条件

用户必须已经注册成为该系统的会员。

（4）后置条件

会员成功登录该系统。

3 个人信息维护

（1）说明

系统会员在完成登录后可以根据自己的需要修改个人信息。

（2）事件流

（a）系统会员在完成登录后，选择个人信息修改，进入个人信息维护界面；

（b）会员填写需要修改的信息，然后点击完成；

（c）系统验证会员所提供信息的有效性，若均有效则提示信息修改成功。

（3）前置条件

会员必须登录该系统。

（4）后置条件

会员信息修改成功。

4 在线审校

（1）说明

该系统的会员可以使用本系统所提供的在线审校功能来校对短文本。

（2）事件流

（a）系统会员在完成登录后，选择在线审校，进入在线审校界面；

（b）用户输入带审校文本，点击审校按钮；

（c）系统后台对用户输入的文本进行审校，审校完成后返回审校结果；

（d）在线审校界面完成对审校结果的显示。

（3）前置条件

会员必须登录该系统。

（4）后置条件

无

5 离线审校

（1）说明

该系统的会员可以上传word文件，并使用本系统的离线审校功能来完成对已上传文件的审校。

（2）事件流

（a）系统会员在完成登录后，选择离线审校，进入离线审校界面；

（b）会员从本地将待审校的word文件上传至服务器中；

（c）word文件上传成功之后，会员发出离线审校请求；

（d）服务器接收到离线审校请求，调用服务器端审校程序，对上传到服务器上的文件进行审校，并将审校结果写入一个新的word文件中。

（e）前台给出审校成功的提示，用户选择审校好的文件进行下载。

（3）前置条件

会员必须登录该系统。

（4）后置条件

无。

## 3.2 系统设计

经过3.1的系统需求得出了该系统的主要的功能，本节将详细的介绍该系统的设计。

### 3.2.1 系统体系结构设计

中文文本自动校对系统是基于B/S的网络服务系统，采用的是Spring MVC的三层Web架构。所谓的MVC架构就是将视图层（View）、控制层（Controller）和模型层（Model）分开，这样可以极大的降低程序的耦合程度，便于以后对系统的维护和升级。该系统的整体架构图如图3-2所示。其中该系统的View层为用户与该系统交互的接口，用户所能看到的所有界面都可以称为视图层，该层主要用Jsp实现，而且还会用到一些JavaScript、CSS和JSTL标签等这些前端常用的技术。该系统的前台主要包含在线审校、离线审校和个人信息维护三部分功能。



图3-2 系统架构图

因为本系统采用的是Spring MVC，所以该系统的控制层采用的还是原生的Java编写的。控制层主要负责接收前台的action请求，并确定调用哪部分后台功能来处理前台的action请求。

模型层主要包括业务层和数据持久层（Dao），业务层主要负责处理业务逻辑，而数据持久层主要负责对数据库进行数据的存取操作，这样就进一步降低了系统的耦合性。该系统后台的业务主要包括两个部分功能：审校服务功能和个人信息维护功能。其中审校服务主要完成对中文文本进行校对的工作，而个人信息维护则是完成对个人信息表的查询和修改。

由于在离线审校的时候需要在服务器端使用审校服务校对用户上传的文件，所以本文将审校服务独立成一个API。这样既能在在线审校时调用审校服务，同时又方便了在后台启动离线审校服务。

### 3.2.2 审校服务功能设计

该系统的审校服务主要针对的是现有的中文网络文本常见的中文词语、汉语拼音、数字号码、以及标点符号等常见的错误进行审校。审校服务具体的功能框架图如图3-3所示。其中每块具体所需要审校的内容如下



图3-3 审校服务功能

（1）词语审校

包括由于音近、形近、义近等因素导致的人名、地名、机构名等实体名称及敏感词语的审校等，如将“王歧山”校正为“王岐山”、将“荷泽”校正为“菏泽”、将“大义禀然”校正为“大义凛然”及将“寒喧”校正为“寒暄”等。

（2）拼音审校

包括字符审校与大小写审校等，如将“Màodùn（冒顿）”校正为“Mòdú”及将“Zhāngyáng Wénān（张杨文安）”校正为“Zhāng-Yáng Wénān”等。

（3）数字审校

数字审校，包括日期、年龄、号码、专用数字等数字的审校，主要通过设定正则表达式形式的审校规则实现，如标注车牌号、手机号及专用数字中的错误等。

（4）标点符号审校

标点符号审校，包括标点符号的误用审校、冗余审校及遗漏审校等，如将误用的问号校正为逗号、剔除冗余的引号及添加遗漏的句号等。

## 3.3 本章小结

本章首先从用户需求、业务需求和功能需求三个角度分析了该系统的需求，紧接着从MVC三层架构分析了该系统的体系结构设计，最后对审校服务的内部功能进行了设计。

# 第4章 中文文本审校功能

对文本进行审校是中文文本自动校对系统的主要核心功能。具体包括词语审校，标点符号审校，数字审校和拼音审校。本章将详细阐述这几个功能的具体实现方法。

## 4.1 词语审校

词语审校实际上是指错别字审校，即校正词语中的一个或多个错误用字。在中文文本，尤其是网络文本中，错别字极为常见。在这里，将词语审校划分为人名、地名、机构名等实体名称和常用词的审校。词语审校的主要方式是采用错别字对照表的方式进行。现今的大多数文本审校软件，包括百度词典、有道词典等在线中文辞典及百度中文输入法、Sogou中文输入法等均提供有海量的词库实现对词语的审校，词语审校的主要依据就是其所提供的词汇库。黑马校对V18更是提供有5000万专业词汇，Sogou中文输入法的中文词汇也已超过4000万。海量的词汇是中文文本审校的基石，为了更有效的提供中文审校服务，必须构建足够数量的中文词汇库。在词语审校方面，区别于通常的专业词语审校，可以将词语分为人名、地名、成语、机构名、敏感词语及其它词语等。通过构建词语审校表或审校规则实现不同类别词语的审校。下面分别介绍不同类别词语的审校方案。

### 4.1.1 实体名称审校

实体名称审校其实就是判断文本中的实体名称与知识库中的实体是否一致的过程，若不一致，则用知识库中的实体名称对文本中的实体名称进行纠正。实体名称审校具体分为以下几个部分。

1、知识库的构建

实体（Entity）是自身客观存在并具有唯一性的具体事物。实体有别于本体，本体是现实生活中某一类实体的总称。本文为了审校实体名称首先需要构建一个包含大量实体的知识库。本文知识库的构建过程如下：

（1）爬取网页文本

构建实体知识库需要有大量的实体信息，而要获取这些大量的实体信息则需要从互联网上爬取大量的网页文本。本文采用的数据源为百度百科。采用百度百科作为数据源原因是：百度百科涵盖了大量的实体信息，而且覆盖面极其广泛；百度百科实体信息更新频率高，百度百科允许所有用户自己创建和编辑词条，这就使得更多的用户可以参与到创建和编辑百度百科的词条中来。所以通过爬取百度百科可以获取到更多质量更高的实体信息。

百度百科的大多数信息都是以HTML网页的形式展现给用户的。所以要获取到百度百科中的实体信息，首先要做的事就是将百度百科的HTML网页文本给爬取下来。如图4-1为四大天王之一刘德华的百度百科词条页面及其对应的HTML源代码。所以要抽取该实体的信息，首先需要把其对应的网页HTML文本爬取下来。本文总共从百度百科网页上爬取了1440000个HTML网页页面。其中部分列表如图4-2所示。



图4-1 百度百科词条及对应HTML源码

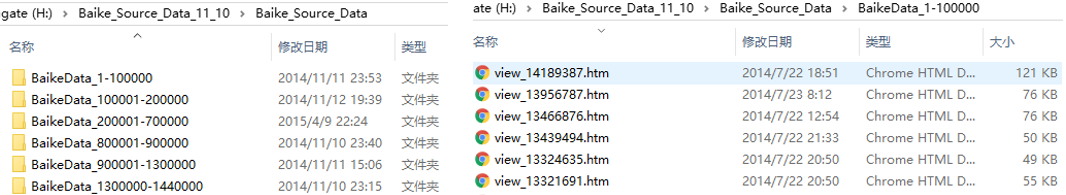


图4-2 百度百科部分网页页面列表

（2）抽取并保存网页文本中的实体信息

在完成网页文本爬取及保存后，接下来要做的工作就是抽取网页HTML文本中的信息。由图4-1可以看出，HTML源码都具有一定的规律性，每一段信息的div都有其唯一的class相对应，就如图中所示第一段的概要的对应的div的class为“lemma-Summary”，而基本信息对应div的class为“basic-info”。在明确每段内容所对应的class之后只需要正则表达式就很容易能从中把要抽取的内容抽取出来。本文目前只需要实体的基本信息，所以在抽取的时候只是抽取了实体的基本信息。首先需要获取文件夹中的网页文本的数目，即遍历每一个HTML文件，再则是将文件读入到内存中，其次则是利用正则表达式Regex=”basic-Info-block” 抽取每个网页中的基本信息。在完成抽取后，则是将实体的基本信息存放到MongoDB数据库中。具体的抽取流程如图4-3所示。



图4-3 百科页面抽取流程图

2、实体识别

本文的实体识别采用的是西南交通大学分词和CRF模型相结合的方法。实体识别的详细流程如图4-4所示。

（1）利用分词的词性进行实体识别

分词是中文自然语言处理的一个重要的课题，大多数的中文自然语言应用都必须以中文分词为基础。本文使用西南交通大学分词系统对中文文本进行分词。因为西南交大分词系统分完词后会带有词性标注，所以本文利用分词后文本中的词性标注来做部分命名实体识别，西南交大分词系统具体的词性和词性说明如表4-1所示。具体步骤分为两步：首先利用分词系统对文本进行分词；其次，查询分词后的文本，若文本中出现词性为nr、ns、nt的词，则将其词性改为entity。因为本文只考虑对人名、地名和机构名的识别，所以本文进行实体识别时，也只取nr、ns、nt三种词性。

表4-1 西南交大分词系统词性和词性说明表

|  |  |  |  |
| --- | --- | --- | --- |
| 词性 | 词性说明 | 词性 | 词性说明 |
| nr | 人名 | nt | 机构团体名 |
| ns | 地名 | nh | 医药疾病等健康相关名词 |
| nb | 生物名 | nf | 食品 |
| g | 学术词汇 | w | 标点符号 |
| u | 助词 | m | 数量词 |



图4-4 实体识别流程

（2）利用CRF进行实体识别

中文文本在经过分词之后有些命名实体会被拆分为多个词语，例如“成都汇康中西医结合医院”分词后的结果为“成都/ns 汇/ng 康/nz 中西医结合医院/nth”，这样就无法准确的通过分词的词性标注来识别出正确的实体名称。而CRF则具有组块的功能，可将被拆开的实体名称组合成一个完整的命名实体。

使用CRF做命名实体识别首先要做的事就是确定特征模板。特征模板一般采用当前词与前后个词及其标记进行组合，即以当前词与前后*n*个词及其标记作为观测窗口。由于当窗口过大时算法的执行效率会过低，而且准确率也不一定会很高，而当窗口过小时所涵盖的信息又太少，不足以完成命名实体识别的任务，因此，在一般情况下*n*的取值范围为2~3。而本文的*n*取值为2，具体的特征模板如表4-2所示。

表4-2 CRF特征模板

|  |  |  |  |
| --- | --- | --- | --- |
| ID | 特征模版 | ID | 特征模版 |
| 1 | U01:%x[-2,0] | 8 | U08:%x[-2,0]/%x[-1,0] |
| 2 | U02:%x[-1,0] | 9 | U09:%x[-1,0]/%x[0,0] |
| 3 | U03:%x[0,0] | 10 | U10:%x[0,0]/%x[1,0] |
| 4 | U04:%x[1,0] | 11 | U11:%x[1,0]/%x[2,0] |
| 5 | U05:%x[2,0] | 12 | U12:%x[-2,0]/%x[-1,0]/%x[0,0] |
| 6 | U06:%x[-1,0]/%x[0,0] | 13 | U13:%x[-1,0]/%x[0,0]/%x[1,0] |
| 7 | U07:%x[0,0]/%x[1,0] | 14 | U14:%x[0,0]/%x[1,0]/%x[2,0] |

基于CRF的命名实体识别是一种有监督的识别方法，因此需要用大量的已标注语料来对CRF模型进行参数训练。本文采用的训练语料为NLPCC2016问答评测任务所提供的训练数据，该数据集总共有14609条句子，本文取了其中的10000条作为训练数据，并对训练数据进行了标注，其中部分标注好的训练数据如图4-5所示，其中用中括号（[]）括起来的就是标注好的实体。

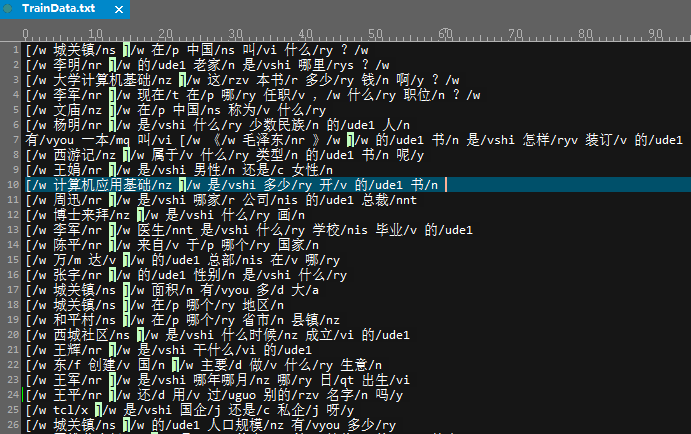


图4-5 CRF命名实体识别部分标注数据

3、实体链接

命名实体链接是指将文本中已经识别出的命名实体链接到知识库中的一个具体真实实体的过程。实体链接是目前自然语言处理领域的一种常用技术。本文的实体链接主要分为以下几个步骤。

（1）同义词表构建

因为本文是文本校对系统，所以在文本中出现的实体名称有可能会是错误的名称，而直接用错误的名称查知识库有可能查不到相应的实体，而且在汉语文本中实体的名称可以是灵活多变的，因为很多名称都存在简称或别名等其他称呼（如科比布莱恩特，别名或简称可为科比、小飞侠、黑曼巴等等），而知识库中存的则是标准的全名，所以当直接用别名或简称来查知识库时也查找不到对应的实体，这就需要构建同义词表。

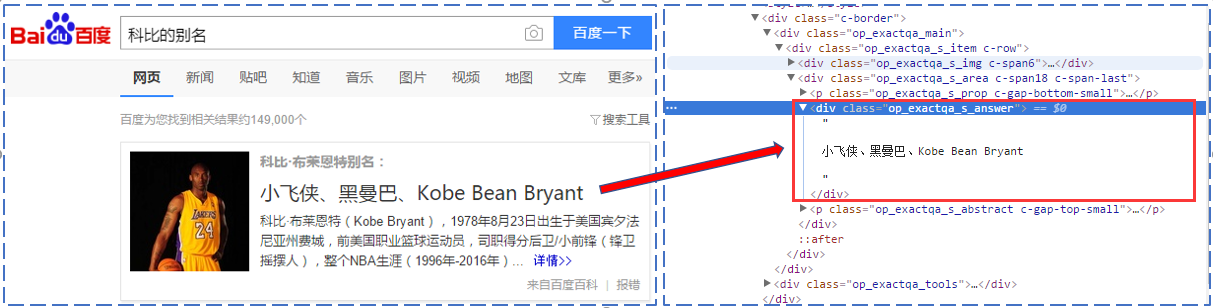


图4-6 百度同义词示例



图4-7 百度百科词条基本信息示例

百度在百度一个人的别名的时候多数情况下会自动给返回其抽取好的别名，如图4-6所示，为百度“科比的别名”后的结果，百度会自动给出其抽取好的别名。本文的同义词表就是借助百度的这个功能来构建的。从图中还可以看出其别名是放在一个class为“op\_exactqa\_s\_answer”的div中。所以在抽取时只需要抽取该div中的内容即可。但也有些情况在百度一个实体名称的别名时候百度并不会直接给出答案，所以这时候需要获取该实体的百度百科词条页面中的基本信息，因为很多实体的百科词条页面中会包含该实体的别名，示例如图4-7所示。这时候只需要抽取该实体百度百科词条页面中的别名即可。具体同义词表构建的算法4-1所示。同义词表达格式为“实体名称 别名1、别名2、… 、别名n”。以歌星陈奕迅为例，其同义词表的格式为：“陈奕迅 医生、吹神、陈小胖、所长”。

|  |
| --- |
| **算法4-1** ：同义词表构建算法 |
| **输入**：实体的名称entityName  **输出**：抽取出的同义词表nickNames  **begin**  1) **DownLoad** content←download(entityName+"的别名") /\*利用百度引擎搜索实体名称的别名并保存其结果\*/  2) **Intialize** XPath Engine htmlCleaner /\*初始化XPath处理引擎\*/  3) **compute** rootNode**←**htmlCleaner(*content*) /\* 将页面源码*content*转化为XPath表达式树\*/  4) **nickNames**←rootNode.evaluateXPath(".//div[@class='op\_exactqa\_s\_answer']") /\*利用XPath表达式抽取其中class为op\_exactqa\_s\_answer中的内容\*/  5) **if** nickNames为空 /\*若不能直接通过百度别名抽取出其别名\*/  6) **entityUrl**←getUrl(entityName) /\*利用百度引擎搜索实体名称并抽取其百度百科的url\*/  7) **if** entityUrl不为空  8) **DownLoad** baikeContent←download(entityUrl) /\*保存实体百科页面\*/  9) **compute** rootNode**←**htmlCleaner(*content*) /\* 将页面源码*content*转化为XPath表达式树\*/  10) **infobox←**rootNode .evaluateXPath(".//div[@class = 'basic-info']"); /\*利用XPath表达式抽取百科页面中的基本信息\*/  10) **nickNames←**getNickName(infobox) /\*抽取infobox中的别名\*/  10）  **end if**  11) **end if**  18) **return** nickNames  **End** |

（2）实体链接

有了同义词表后，接下来的工作就是进行实体链接。本文的实体链接分为两个步骤，首先用实体名称直接查找，若用实体名称无法找到，则通过匹配同义词表查找。直接查找比较简单，就是通过查询语句用实体名称直接查询知识库，这里就不做详细的介绍了。但是对于那些可能有错词或者是别名的实体名称，而知识库中存的都是标准的名称，这样就无法直接查找知识库找出想要的实体，所以就必须通过相应的办法找出完整准确的实体名称，再用正确的实体名称去查询知识库找出相应的实体。本文采用的就是同义词表匹配的方法。

同义词表匹配的方法首先是遍历同义词表，看是否存在与实体名称*entityName*长度一致的词*nickName*，若存在计算*nickName*与*entityName*的拼音相似度*spellSimilarity*。本文用得是最长公共子序列去计算两个词的拼音相似度具体计算方法如下：设entityName的拼音长度为*n*，nickName的拼音长度为*m*，entityName的拼音与nickName的拼音中的最长公共子序列长度为*k*，则*spellSimilarity*的计算公式如下：

 (4-1)

拼音相似度主要解决的是由于用户使用拼音输入法导致的实体名称错误，而大多数人在使用拼音输入法将实体名称打错时，错误的实体名称与正确的实体名称之间的拼音差别一般不会太大，甚至很多情况下是一致的。所以在使用拼音输入法来匹配正确的词的时候相似度的阈值会比较大，本文经过实验调参将拼音相似度的阈值设置为0.95，当则用该*nickName*对应的标准的实体名称去查知识库。

实体名称出现错误一部分原因是因为拼音相似或相近导致的，还有一部分可能是由于字形相似或相近导致的。而字形相似或相近的实体名称，其拼音可能差距会比较大，因而用拼音相似度的算法无法找出正确的实体名称。这就需要计算另外一种相似度。受拼音相似度启发，本文采用了一种汉字的笔画相似度来解决此类由于字形相近而导致的错误。当时，就计算entityName和nickName二者的笔画相似度*strokesSimilarity*。笔画相似度同样采用的是最长公共子序列来计算。设entityName笔画的长度为n，nickName的笔画长度为m，二者笔画的最长公共子序列的长度为k，则*strokesSimilarity*的计算公式如下：

 (4-2)

两个词的笔画相似度的差异性会比拼音相似度要大一些，所以笔画相似度的阈值的取值会比拼音相似度的阈值要小一些。本文经实验验证将笔画相似度的阈值设置为0.86。当时，则用该nickName对应的标准的实体名称去查知识库。

4、实体名称审校

实体名称审校其实就是比较文本中的实体名称与知识库中的实体名称是否一致的过程。在完成实体链接之后如果知识库中存在该名称的实体，若链接成功则会返回一个实体的基本信息，若链接不成功则表示该实体在知识库中不存在，返回空值。若知识库中存在与文本中实体名称相同或者相似的实体，则比较返回实体的实体名称与文本中的实体名称是否一致，若一致，则表示该处实体名称没有错误，不予审校，否则用知识库中的实体名称来纠正文本中的实体名称。若知识库中不存在该实体名称的实体，跳过该处实体名称。

### 4.1.2 常用词语审校

实体名称审校只是整个词语审校中的一部分，文本中有很多的错误还是集中在普通词汇上面。所以常用的普通词汇的审校也是词语审校的一个重要的内容。词语审校主要分为以下几个部分。

1、词库的构建

普通词语审校主要采用的是与词库中正确的词进行比较来实现，所以词库是普通词语审校的基础。本文采用的词库来自搜狗细胞词库。

2、字典树构建

普通词语需要大量的词库，而如何高效地从大量的词库中查找出特定词语就很重要。本文采用了字典树的方式来加载词库，以下就是字典树的介绍及构建过程。

（1）字典树介绍

在计算机科学中，字典树（Trie树）也被称作单词查找树，它是一种哈希树变种的搜索树。字典树主要应用于统计、排序和保存大量的字符串（但并不仅仅限制于字符串）。字典树的键值通常就是字符串，与二叉搜索树不同的是字典树上每个节点的键值都是依赖于其本身的，并不依赖于其他键值。节点的所有后代都具有与该节点相关联的字符串的公共前缀，而且字典树的根节点为空节点。字典树的值不一定与每一个节点相关联，相反字典树的值一般倾向于仅与叶子节点关联。在图4-8的字典树的例子可以看出，每个非叶子节点上只存储key值（单词公共前缀），并不存储value值（单词的词频）。每个单词的词频只在叶子节点存储。



图4-8 字典树示例

（2）字典树构建

字典树构建首先需要构建出字典树的节点类。由上文介绍可知，字典树的非叶子节点与叶子节点是有区别的，有些值只在叶子节点存储，而非叶子节点则不需要存储。所以本文需要构建两个类来存放字典树的节点，一个用来实例化非叶子节点，另一个类用来实例化叶子节点。两个类具体的代码结构如代码4-1和代码4-2所示。非叶子节点只包含一个Map，该Map的key值为一个汉字，表示一个词的第n个字，而其value值则是一个TreeNode的对象。而叶子节点则是继承非叶子节点类，在非叶子节点的基础上增加了词和词频这两个属性。

|  |
| --- |
| **代码4-1** 非叶子节点类 |
| **class** TreeNode {  Map<Character, TreeNode> subNodes; // 子节点  public TreeNode() { //创建新的节点  if (subNodes == null)  subNodes = new HashMap<Character, TreeNode>();  }  public void insert(char k, TreeNode sub) { // 插入节点  if (subNodes == null)  subNodes = new HashMap<Character, TreeNode>();  subNodes.put(k, sub);  }  public TreeNode getSubNode(char k) { // 得到指定位置的值  if (subNodes == null)  return null;  return subNodes.get(k);  }  public boolean isAlsoLeaf() { //判断是否为叶子节点  return false;  }  } |

|  |
| --- |
| 代码4-2：叶子节点类 |
| **class** LeafNode extends TreeNode {  int freq; /\*一个词的词频\*/  String word; /\*一个完整的词\*/  public LeafNode(TreeNode treeNode) {  super.subNodes = treeNode.getnodeMap();  }  public boolean isAlsoLeaf() {//判断是否为叶子节点  return true;  }  } |

构建完两个节点类之后，接下来就是利用这两个类完成字典树的构建过程。构建字典树就是把存储词库的文件读取并加入到内存的过程。具体的构建流程图如图4-9所示。



图4-9 字典树构建流程图

3 常用词语审校

常用词语审校的核心思想就是找出文本中的错误词语，然后再用词库中正确的词对该文本中的词语进行校对。这样如何发现文本中的错误词语就成了该部分的关键问题之一。本文采用的是n元组块后的词与词库中的词计算相似度的方法来查找文本中出现的错误的。词语审校过程的流程图如图4-10所示。以下分别介绍审校过程各个部分的工作：

（1）分词

分词就是利用分词系统对输入的文本进行分词。因为之后的所有处理过程都基于分词的结果，所以第一步首先就是要对输入的文本进行分词。

（2）组块

许多分词系统是按照词库来分词的，所以对于那些错词，很多情况下分词系统是无法将其分正确的。例如“项目名称”，正确的分词结果是“项目/n 名称/n”，若将其写为“项木名称”那分词系统则会将错误的“项木”分开，其结果为“项/q 木/ng 名称/n”，这是因为分词词库里面根本没有“项木”这个词。同样在审校的词库里面也没有“项木”这个词。若想要查找出错词“项木”就必须要进行组块。本文采用的组块策略是最长词优先的原则，先进行长词组块，再逐步缩短。具体步骤如下：

a）用该词与其后的两个词进行组块，然后再判断组合的块是否为常用词并检测是否有错，若有错则保存错词及其校对结果并跳过后面的步骤，并从其后的第三个词重新开始步骤a；

b）若步骤a没找到错误，则用该词与其后的一个词进行组块，然后再判断组合的块是否为常用词并检测是否有错，若有错则保存错词及其校对结果并跳过后面的步骤，并从其后的第二个词重新开始步骤a；

c）若步骤b还是没有找到错误，则直接检测当前词是否有错，若有错则保存错词及审校结果，然后从下一个词开始重新开始步骤a，直至所有词都已处理完。



图4-10 常用词语审校流程图

（3）字典树模糊查询

当待查词为正确的词的时候自然不需要模糊查询，可以直接按字逐层查询字典树即可得到查询结果。但若词语中存在一个或多个字出现错误时，这时候直接查询字典树无法找到与之对应的词。所以本文采用了字典树模糊查询的方法。字典树模糊查询的具体流程如图4-11所示。



图4-11 字典树模糊查询流程图

由图4-11可知，字典树模糊查询主要分为以下几步：

a）在字典树中查询待查词的第一个字，并建立了数据结构Map<trie, List<TreeNode>> map来保存查询结果。查询主要采用两种策略，首先是直接查询，若能在根节点的子节点中直接查询到该字，则将该字对应的节点保存到map中，若不能找到则遍历根节点的所有子节点，利用拼音和笔画相似度（分别如公式4-1和4-2所示）来查找与该字相似的字，并将所有相似的字所对应的节点保存到map中；

b）遍历map中相应层的List<TreeNode>，从每一个TreeNode的下一个节点来查询待查词的下一个字，查询策略还是与a中的查询策略一致，首先是直接查询，若能在该节点的子节点中直接查询到该字，则将该字对应的节点保存到map中，若不能找到则遍历该节点的所有子节点，利用拼音和笔画相似度（分别如公式4-1和4-2所示）来查找与该字相似的字，并将所有相似的字所对应的节点保存到相应层的map中，并从下一个字开始重复步骤b，直至所有字均被查找完成；

c）当待查词的所有字均已查找完毕，则遍历map中的最大key值（即最后一层）的List<TreeNode>，判断每一个节点是否为叶子节点，若为叶子节点则将叶子节点中的词保存到词列表（wordList）中去，遍历完成之后则返回wordList。

（4）审校结果获取

经过了步骤3，若其中的某个词在词库里面存在相同或相似的词的时候都会得到一个由一个或多个词组成的wordList。当wordList中只有一个词的时候比较简单，只需要比较两个词是否一致就行了，若一致忽略该处的词，若不一致，则将wordList中的词作为审校结果返回；但当wordList中有多个词的时候则需要将待审校词与wordList中的词逐一进行相似度计算。最后将相似度最大的词作为审校结果返回。

### 4.1.3 敏感词检测

敏感词一般指的是带有反执政党、暴力、不健康色彩的词或者不文明的用语。在很多网站都很注重敏感词的检测与屏蔽。本系统也设计了针对敏感词检测的方法。

本文的敏感词检测主要分为以下几个步骤：

1 词库构建

本文的敏感词词库是在网上直接下载的，里面包含492个敏感词。其中部分敏感词如表4-3所示。

表4-3 敏感词表示例

|  |  |  |  |
| --- | --- | --- | --- |
| 编号 | 敏感词 | 编号 | 敏感词 |
| 1 | “六四”政治风波 | 5 | 台湾主权独立 |
| 2 | tmd | 6 | 藏独 |
| 3 | 台独 | 7 | 大跃进 |
| 4 | 台湾独立 | 8 | 呆子 |

2 字典树构建

本节的字典树还是采用的4.1.2中的字典树的结构，构建的流程也与上一节中的构建流程是一样的，唯一的区别在于加载的词库不同。

3 敏感词检测

敏感词主要就是检测待审校文本中是否存在敏感词，若存在给出相应的标识。本文采用的是顺序遍历待审校文本，然后对文本中的每一个字查找字典树的方式来检测文本中是否存在敏感词。设待审校文本的长度为*n*，因为字典树中采用的是HashMap，其key值为每个词的每个字，所以在字典树中查找每个字的时间复杂度为*O*(1)。这样就可以得出通过字典树来检测一个长度为n的文本中是否存在敏感词的时间复杂度就是遍历该文本的时间复杂度，即*O*(n)。敏感词具体检测的算法流程图如图4-12所示。



图4-12 敏感词检测流程图

## 4.2 标点符号审校

### 4.2.1 标点符号错误类型介绍

目前市场上的一些校对软件以及学术界很少提及对标点符号的校对问题研究，基本没有有效的方法提出。但这并不意味着标点符号校对问题无从下手，本文认为可以结合标点符号的习惯用法，提取出特定的规则来对标点符号进行审校。标点符号审校主要包括标点符号的误用、冗余等，现分类阐述如下。

1 标点符号误用

标点符号误用审校主要是对误用的标点符号进行审校，如分号误用为句号、句号误用为逗号及书名号的误用等。其中标点符号误用举例如下：

（1）非疑问句误用问号：

我不知道这条路谁能走通？但我一定要坚定不移地走下去。

（2）选择问句中间的停顿误用问号：

今天去呢？还是明天去呢？我实在是拿不定主意。

（3）倒装句中问号与叹号前置错误：

到底怎么办？这件事。

（4）分句之间误用顿号：

他留给我的印象依然如故：天真中透示着纯真、活泼中蕴含着稚气。

（5）联合词组不同层次的词语之间误用顿号：

上海的越剧、沪剧、淮剧、湖北的黄梅戏、河南的豫剧，在这次会演中，都带来了新剧目。

（6）句中未用逗号径直用分号：

任弼时同志一生三怕：一怕工作少；二怕麻烦人；三怕多用钱。

（7）句中已用句号再用分号：

一、学习贵在自觉。要有笨鸟先飞的精神，自我加压；二、学习贵在刻苦。要有锲而不舍的精神，持之以恒；……

2 标点符号冗余

标点符号重复审校主要是对重复的标点符号进行审校，如冒号的重复及“等等”与省略号的重复等。标点符号冗余主要包括以下几种情况：

（1）概数用顿号：

一天才走五、六里。

（2）集合词语用顿号：

如今，科学、技术的发展，必然会推动工、农业的发展。

（3）连词前用顿号：

观众长时间地等待，只为一睹她的风采、或签上一个名。

（4）短暂停顿用冒号：

他为什么总是说：“我疼你，我要保护你”之类的话？

（5）省略号和“等”“等等”同时使用：

一起去的有十七八个人，有我姐姐、表弟、表妹、同学……等。

（6）冒号中再用冒号：

昨天开校会，校长宣布：学校要实行两项改革措施：一是岗位工资制，二是岗位聘任制。

### 4.2.2 标点符号审校

1 正则表达是构建

要使用正则表达式校对文本中出现的标点符号，首先需要构建一个审校标点符号的正则表达式库。本文针对不同类型标点符号错误类型构建了34条正则表达式，其中部分正则表达式及其校对的类型如表4-4所示。其中每条规则的“@@”后面是对应的审校意见。

表4-4 标点符号校对正则表达式示例

|  |  |  |
| --- | --- | --- |
| 编号 | 模式串 | 类型 |
| 1 | [\u4e00-\u9fa5]? (…+)等@@ | 省略号与等不能连用 |
| 2 | [\u4e00-\u9fa5](：)[\u4e00-\u9fa5]+：. @@， | 冒号不能连续使用 |
| 3 | .”[\u4e00-\u9fa5]+[说道讲答问](，)[^“]. @@。 | 引用结束时要用句号 |
| 4 | [\u4e00-\u9fa5]?呢|吗(。) @@？ | 语气词后面一般不用句号 |
| 5 | [\u4e00-\u9fa5]+ (？{4,}) @@？ | 问号连用次数一般不大于三次 |
| 6 | [\u4e00-\u9fa5]+ (、)以?及@@， | 顿号不与“以”、“及”连用 |



图4-13 标点符号审校流程图

2 审校步骤

标点符号校对以段落为单位对标点符号进行扫描，将段落中出现的标点符号存入标点符号队列中，接着再遍历标点符号列表，根据标点符号类型选择对应的规则库对该段落进行校对，如果匹配到了则说明该段落中出现了此标点符号的误用错误，并根据规则返回所对应的解决方案与出错位置。对于括号与书名号等配对标点符号则采用堆栈进行校对，简言之就是如果扫描到了配对标点符号前向符号““”、“《”、“（”、“{”、“[”等则将其压栈，并记录此标点符号在段落中的位置。若为后向标点符号“””、“》”、“）”、“}”、“]”等则检验栈中是否有相应的前向标点，若有则将其从堆栈中弹出；如果不存在对应的前向标点，则将当前标点记为错误。具体标点符号的审校流程如图4-13所示。

## 4.3 数字审校

### 4.3.1 数字审校概述

数字审校主要是对日期、年龄等错误或不规范数字的审校，现分类阐述如下：

1 日期审校

日期审校主要是对错误日期或错误时间的审校，如识别并标注错误形式的年份（如“19999年”）或错误形式的时间（如“2007年2月29日”）等。

2 年龄审校

年龄审校主要是对不现实的年龄数据的审校，如210岁等。

3 号码审校

号码审校主要包括对有固定格式的号码的审校，如身份证号码、手机号码、电话号码、邮政编码、电子邮件、车牌号、IP地址等。

4 专用数字审校

专用数字审校主要是对具有固定书写格式的专用数字的审校，如将“3000里江山”校正为“三千里江山”等。

5 其它审校

其它审校主要包括对大小写数字的混用、全角半角混用及不同格式数字的审校等，如将“康熙18年”校正为“康熙十八年”，将“19７８年”校正为“1978年”等。

### 4.3.2 数字审校流程

数字审校基本采用的是通过正则表达式来检测并校对待审校文本中出现的数字。具体审校过程分为以下两个部分。

1 正则表达式构建

要使用正则表达式校对文本中出现的数字，首先需要构建一个正则表达式库。本文针对不同类型的数字格式构建了26条正则表达式，其中部分正则表达式及其校对的类型如表4-5所示。

表4-5 数字审校正则表达式示例

|  |  |  |
| --- | --- | --- |
| 编号 | 模式串 | 类型 |
| 1 | (?<hzdateymd>\d+(年|\-|\.|/)\d+(月|\-|\.|/)\d+(日|)) | 检验年月日是否有效 |
| 2 | (?<pemail>\w+([-+.]\w+)\*@\w+([-.]\w+)\*\.\w+([-.]\w+)\*) | 检验邮箱是否有效 |
| 3 | (?<ptel>((13[0-9])|(15[^4,\D])|(18[0,5-9]))\d{8}$) | 检验手机号是否有效 |
| 4 | (?<pip>(1\d{2}|2[0-4]\d|25[0-5]|[1-9]\d|[1-9])\.(1\d{2}|2[0-4]\d|25[0-5]|[1-9]\d|\d)\.(1\d{2}|2[0-4]\d|25[0-5]|[1-9]\d|\d)\.(1\d{2}|2[0-4]\d|25[0-5]|[1-9]\d|\d)$) | 检验ip地址是否有效 |
| 5 | (?<agehuman>\d+)(\D\*)(岁|出头|高龄) | 检验年龄 |
| 6 | (?<pbuscard>(冀|豫|云|辽|黑|湘|皖|鲁|新|苏|浙|赣|鄂|桂|甘|晋|蒙|陕|吉|闽|贵|粤|川|青|藏|宁|渝|京|津|沪|琼){1}[A-Z]{1}( |•|-){0,1}[A-Z\_0-9]{5}) | 检验车牌号是否有效 |



图4-14 数字审校流程图

2 数字审校

数字审校是采用正则表达式的规则来进行校对，其具体过程就是遍历正则表达式库，查找待审校文本中是否存在该正则表达式格式所对应的数字，若存在则判断其是否正确，若不正确则给出错误提示。例如“2007年2月29日”，首先利用正则表达式“(?<hzdateymd>\d+(年|\-|\.|/)\d+(月|\-|\.|/)\d+(日|))”将该日期找出来，因为2007年2月只有28天，并没有29日，所以提示该日期为错误日期。由于日期审校、年龄审校、号码审校等没有校对的依据，故只做正确性校对，没有修改建议。具体数字审校的流程如图4-14所示。

## 4.4 拼音审校

### 4.4.1 拼音审校概述

拼音审校包括不符合拼音规范的拼音的审校及大小写错误的审校，现分类阐述如下：

1 字符审校

字符审校是对汉语拼音字符错误的审校，除多音字导致的拼音错误审校外，还包括人名字符审校及地名字符审校等，主要是对专用的人名及专用或通用的地名拼音的审校，还包括其它原因导致的字符错误的审校，如将 “Běljīng（北京）”校正为“Běijīng”等

2 大小写审校

大小写审校是对汉语拼音大小写的审校，除句首大写及专用词语开头大写外，还包括人名大小写审校及地名大小写审校等，主要是对专用的人名及专用或通用的地名拼音中大小写的审校，如将“Lǐhuá（李华）”校正为“LǐHuá（李华）”、将“Běijīngshì（北京市）”校正为“Běijīng Shì（北京市）”等。

### 4.4.2 拼音审校流程

本文拼音审校主要针对的是文本中出现的拼音标注，如“北京（běijīng）”。审校过程主要是获取标注的拼音，然后用pinyin4j工具包来获取被标注拼音的正确拼音，然后在对其比较其拼音是否正确，若不正确给出用pinyin4j工具获取的拼音作为审校结果，拼音审校具体流程图如图4-15所示。



图4-15 拼音审校流程图

图4-15中的检测标注拼音的方法如下：

拼音审校的第一步就是检测待审校文本中是否存在标注拼音，检测方法是通过正则表达式“[(a-z)|(A-Z)]?(ā|á|ă|à|ē|é|ĕ|è|ī|í|ĭ|ì|ō|ó|ŏ|ò|ǖ|ǘ|ǚ|ǜ|ü|ū|ù|ú|ŭ|ù|)?[(a-z)|(A-Z)]?”来检测，若存在则利用规则"[^aoeiuü]?h?[iuü]?(ai|ei|ao|ou|er|ang?|eng?|ong|a|o|e|i|u|ng|n)?"来对检测出来的拼音进行切分，如“BĕijīngShì”,则可以切分为“Bĕi jīng Shì”，然后在根据切分出来的拼音个数在标注拼音标注前后来获取被标注拼音的汉字，用pinyin4j工具包分别获取两组词的拼音，并与转化为小写并切分好的标注拼音进行相似度计算，取相似度最高的词作为被标注拼音的汉字。如“北京市（BĕijīngShì）今天天气很好”，因为拼音被切分为三段，即为三个汉字的拼音，所以取该拼音前后的三个汉字，分别为“北京市”和“今天天”，接着获取这两个词的拼音分别为“bĕi jīng shì”和“jīn tiān tiān”，经计算二者与转换为小写并切分好的标注拼音“bĕi jīng shì”的相似度分别为1.0和0.32。所以最终取“北京市”作为被标注拼音的汉字。这里所用的相似度算法还是公式(4-1)中所提的最长公共子序列计算相似度的方法。

## 4.5 本章小结

本章重点介绍了审校系统中的词语审校功能的实现算法和流程，从实体名称审校、常用词语审校以及敏感词检测三个方面介绍了词语审校，其中实体名称审校采用的是实体链接的方法来实现的，而常用词语审校和敏感词检测则是采用构建字典树来实现的，二者区别在于常用词语审校则是采用模糊匹配的方法来查找词库，而敏感词则是采用精确查找的方式来查找词库，而且二者所加载的词库也不相同。本章还介绍了标点符号、数字审校和拼音审校的方法，对于这三类错误，本文采取的是构建正则表达式的方法来实现的。

# 第5章 系统实现与演示

## 5.1 前言

本章主要介绍中文文本自动系统web框架的实现技术以及每个前台功能的操作流程及展示。

## 5.2 系统框架

本文采用的Spring MVC框架来实现中文文本自动校对系统的Web框架。Spring MVC是基于请求驱动的Web框架，用户提交的请求通过DispatcherServlet来通过查询处理器映射器来确定由哪个控制器来处理。所以搭建Spring MVC框架首先要在web.xml中配置DispatcherServlet配置文件，具体配置如下：

|  |
| --- |
| <servlet>  <servlet-name>proofread</servlet-name>  <servlet-class>org.springframework.web.servlet.DispatcherServlet</servlet-class>  <init-param>  <param-name>contextConfigLocation</param-name>  <param-value>/WEB-INF/proofread-servlet.xml</param-value>  </init-param>  <load-on-startup>1</load-on-startup>  </servlet> |

在web.xml中配置好了DispatcherServlet配置文件后，接着就是在DispatcherServlet配置文件中配置DispatcherServlet。在DispatcherServlet配置文件中需要配置以下几部分内容：

（1）组件扫描基础包，其中包含各个功能的控制器；

（2）资源解析器，负责解析系统所需资源所在位置；

（3）自动扫描组件，开启自动扫描组件后，DispatcherServlet会根据注解自动适配控制器；

（4）视图解析器，负责解析model返回的视图逻辑名称，并将其转换为特定的视图名称；

（5）文件解析器，负责解析文件上传及下载。

其具体配置如下所示：

|  |
| --- |
| <!-- 组件扫描包 -->  <context:component-scan base-package="com.swjtu.controller" />  <!—自动扫描组件 -->  <mvc:annotation-driven />  <!—资源解析器 -->  <mvc:resources mapping="/members/\*\*" location="/WEB-INF/members/" />  <mvc:resources mapping="/assets/\*\*" location="/WEB-INF/assets/" />  <!—视图解析器 -->  <bean class="org.springframework.web.servlet.view.InternalResourceViewResolver">  <property name="prefix">  <value>/WEB-INF/views/</value>  </property>  <property name="suffix">  <value>.jsp</value>  </property>  </bean>  <!—文件解析器-->  <bean id="multipartResolver" class="org.springframework.web.multipart.commons.CommonsMultipartResolver" p:defaultEncoding="UTF-8" p:maxUploadSize="5400000" p:uploadTempDir="files">  </bean> |

## 5.3 系统功能展示

### 5.2.1 注册

用户注册是该系统的一个基本功能，用户首先在图5-1所示的注册页面填写注册所需的基本信息，这些都是必不可少的信息，邮箱需要用于邮箱验证，而密码则是用户登录系统的凭证；填写完邮箱和密码之后点击注册，系统会给注册邮箱发一份验证邮件如图5-2所示，用户点击验证邮件中的链接即可完成注册进入系统，即用户个人信息维护界面，如图5-3所示。



图5-1 用户注册



图5-2 邮箱验证邮件



图5-3 个人信息维护

### 5.2.2 登录

用户已经完成注册后可以随时登录中文文本自动校对系统，用户只需要在如5-4所示的登录界面输入正确的邮箱和密码即可登录到该系统。



图5-4 用户登录

### 5.2.3 个人信息维护

用户刚刚完成注册后进入该系统，用户的基本信息除了邮箱之外其他的都是空的如图5-3所示，用户可以在各项基本信息框内输入自己的基本信息，然后点保存即可，修改后的基本信息如图5-5所示。用户还可以在该界面修改密码，只要验证了原密码即可对密码进行修改。



图5-5 修改个人信息及密码

### 5.2.4 在线审校

在线审校是该系统的主要功能之一，用户进入在线审校界面后可以在左边的输入框内输入要审校的文本，然后点击审校，待后台完成审校后在右侧校对结果框内会输出后台校对的审校的结果，其中红色的代表错误或者出现疑似错误的地方。如图5-6所示，其中第一条“形像代言人”中的“形像”二字为错词，正确的应该是“形象”，系统会给出校对结果以及原文本供用户选择。



图5-6 在线审校

### 5.2.5 离线审校

离线审校主要针对的是word文档的审校，用户进入离线审校界面后可以选择要上传的word文件，然后点击上传，如图5-7所示。用户完成文件上传之后，word文档交由系统后台去审校，待用户审校完成之后会在离线审校界面显示出来审校完成后的文件供用户下载，用户只需要点击文件名后的下载链接即可完成审校完成后的结果的下载，如图5-9所示。

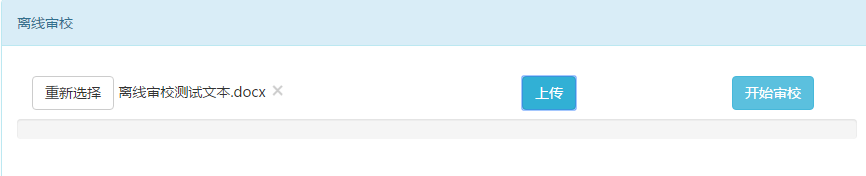


图5-7 文件上传

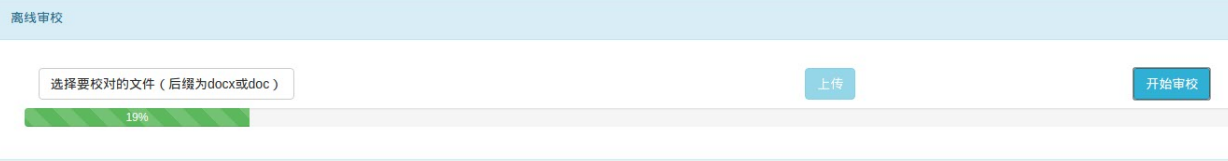
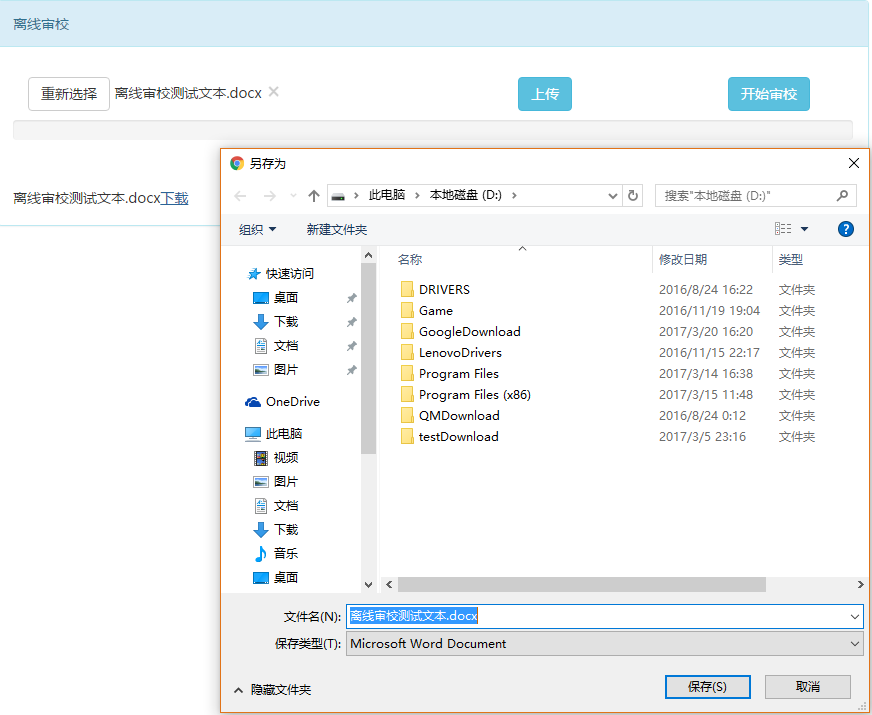


图5-8 离线审校



如5-9 离线审校结果下载

## 5.3 本章小节

本章主要介绍了如何使用Spring MVC搭建中文文本自动校对系统的Web框架，并详细的介绍了注册、登录、个人信息维护、在线审校以及离线审校的操作流程。重点介绍了每个功能的操作流程。

# 结 论

## 本文工作总结

互联网信息快速增长的同时也造成了网络文本的参差不齐，而目前大多数审校工作特别是在新闻出版和电视广播等行业仍然是以人工审校为主。本文针对这一现象构建了一个基于B/S模式的中文文本自动校对系统。主要研究内容包括如下几个方面：

1. 针对中文文本自动校对系统展开系统需求分析，详细分析系统的用户需求，业务需求以及功能需求。

2. 用MVC框架对系统框架进行设计，详细地分析系统中每一层所包含的功能，并详细地分析并设计了审校服务的具体功能。

3. 实现中文文本自动校对系统的后台功能，其中包括采用CRF加分词的技术识别文本中的实体并采用实体链接的技术来实现词语审校中的实体名称审校，而词语审校中的常用词语审校即敏感词审校则采用构建字典树的方式来实现；针对标点符号和数字审校主要采用的是构建规则库来实现；而对于拼音审校本文利用pinyin4j工具包来获取正确词的拼音，再与原文中的标注的拼音比较来进行审校的。

4. 利用Spring MVC框架来实现中文文本自动校对系统的web框架。其中包括前台Jsp界面的编写，控制层逻辑的实现。

## 未来工作展望

本文提出的审校方法虽然可以审校出文本中很多常见错误，也能给出相应的修改意见，但是局限于时间和作者本身的能力，该系统的审校方法还有很大改进空间。接下来的工作可以从以下几个方面来进行：

（1）本文的词语审校针对单个词的错误利用词典的方式来校对的，并没有考虑句子的语义。下阶段可以从语义的角度来考虑词语审校，根据句子语义来选择合适的词对错误文本进行审校。方法则可以尝试一下用深度学习模型。

（2）本文的标点符号审校、数字审校都是采用规则来进行的，下阶段可以尝试寻找合适的机器学习或者深度学习模型来解决这两部分的审校工作。

# 致 谢

光阴荏苒，日月如梭，转眼间三年的硕士生涯就即将结束，但在内心还是有诸多的不舍与留恋。

首先，我要由衷的感谢我的导师李天瑞教授，感谢李老师一直以来对我生活和学习上的无私的指导和帮助。李老师每周会不定期询问我的生活和学习状况。每当我遇到困惑和难题时，李老师总能敏锐的发现问题的关键点，并会很耐心的给予我详细的指导。李老师对待工作和学术严谨认真的态度以及对新知识的明锐洞察力对我影响极大，使我受益终身。同时我还要感谢我的指导老师贾真，我从研一开始就进入了贾老师带的自然语言处理组，这三年贾老师对我学习和生活都给予了极大的帮助。

其次我还要感谢上一届的师兄黄峻福，他和我都在自然语言处理组，他在校期间给予了学习上很大的帮助，每当我有问题找他时他都会无私的帮助我。感谢博士后刘胜久师兄，他在我硕士三年期间对我无私的帮助和细心的指导。感谢我们自然语言处理组的其他同学，和你们一起做比赛和项目使我获益良多。

感谢我的父母，是你们一直以来对我默默的支持和付出才能让我专心于学习。你们的爱和鼓励使我不断前行，无所畏惧。

感谢各位评审和答辩的各位老师，谢谢你们能够在百忙之中抽出时间来对我的论文进行审阅和答辩，感谢你们提出的宝贵的意见。

# 参考文献

1. Adamov A. Distributed file system as a basis of data-intensive computing[C]. 2012 6th International Conference on Application of Information and Communication Technologies (AICT), IEEE, 2012: 1-3.
2. Peterson J L. Computer programs for detecting and correcting spelling errors[J]. Communications of the ACM, 1980, 23(12): 676-687.
3. Kukich, Karen. Techniques for automatically correcting words in text. ACM Computing Surveys (CSUR), 1992, 24(4): 377-439.
4. Pollock J J, Zamora A. Automatic spelling correction in scientific and scholarly text[J]. Communications of the ACM, 1984, 27(4): 358-368.
5. Riseman E M, Hanson A R. A contextual postprocessing system for error correction using binary n-grams[J]. Computers, IEEE Transactions on, 1974, 100(5): 480-493.
6. Riseman E M, Ehrich R W. Contextual word recognition using binary digrams[J]. Computers, IEEE Transactions on, 1971, 100(4): 397-403.
7. Golding A R, Schabes Y. Combining trigram-based and feature-based methods for context-sensitive spelling correction[C]. Proceedings of the 34th annual meeting on Association for Computational Linguistics. Association for Computational Linguistics, 1996: 71-78.
8. Evershed, John, and Kent Fitch. Correcting noisy OCR: Context beats confusion. Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage. ACM, 2014, 45-51.
9. Kumar, Atul, and Gurpreet Singh Lehal. Automatic Text Correction for Devanagari OCR. Indian Journal of Science and Technology, 2016, 9(45).
10. Bulatov, Konstantin, et al. Trigram-based algorithms for OCR result correction. Proceedings in Ninth International Conference on Machine Vision. International Society for Optics and Photonics, 2017, 103410O-103410O.
11. 于勐, 姚天顺. 一种混合的中文文本校对方法[J]. 中文信息学报, 1998, 12(2): 31-36.
12. 李建华, 王晓龙. 多特征的中文文本校对算法的研究[J]. 计算机工程与科学, 2001, 23(3): 93-96.
13. 陈笑蓉, 秦进, 汪维家, 等. 中文文本校对技术的研究与实现[J]. 计算机科学, 2003, 30(11): 53-55.
14. 龚小谨, 罗振声. 中文文本自动校对中的语法错误检查[J]. 计算机工程与应用, 2003, 39(8): 98-100.
15. 骆卫华, 罗振声, 龚小谨. 中文文本自动校对的语义级查错研究[J]. 计算机工程与应用, 2003, 39(12): 115-118.
16. 骆卫华, 罗振声, 龚小瑾. 中文文本自动校对技术的研究[J]. 计算机研究与发展, 2004, 41(1): 244-249.
17. 王斯宇, 邵波. 基于CSSCI的文本自动校对系统的构建与实现[J]. 高校图书馆工作, 2014, 34(6): 50-54.
18. 张仰森, 唐安杰, 张泽伟. 面向政治新闻领域的中文文本校对方法研究[J]. 中文信息学报, 2014, 28(6): 79-84.
19. Lafferty, John, Andrew McCallum, and Fernando Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. Proceedings of the eighteenth international conference on machine learning, ICML. 2001. 1(1). 282-289.
20. Yao, Kaisheng, et al. Recurrent conditional random field for language understanding. Proceedings of Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on. 2014, 4077-4081.
21. Ammar, Waleed, Chris Dyer, and Noah A. Smith. "Conditional random field autoencoders for unsupervised structured prediction." Advances in Neural Information Processing Systems. 2014:3311-3319.
22. 张盛, 李芳. 基于迭代两步 CRF模型的评价对象与极性抽取研究[J]. 中文信息学报, 2015, 29(1): 163-169.
23. 栗伟, 赵大哲, 李博, 等. CRF与规则相结合的医学病历实体识别[J]. 计算机应用研究, 2015, 4, 1082-1086.
24. Uzunbas, Mustafa Gokhan, Chao Chen, and Dimitris Metaxas. "An efficient conditional random field approach for automatic and interactive neuron segmentation." Medical image analysis 2016, 27: 31-44.
25. Vemulapalli, Raviteja, et al. Gaussian conditional random field network for semantic segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, 3224-3233.
26. 吴胜远. 一种汉语分词方法. 计算机研究与发展, 1996, 33(4): 306-311.
27. Howell, Peter, Stevie Sackin, and Kazan Glenn. Development of a Two-Stage Procedure for the Automatic Recognition of Dysfluencies in the Speech of Children Who StutterII. ANN Recognition of Repetitions and Prolongations With Supplied Word Segment Markers. Journal of Speech, Language, and Hearing Research, 1997, 40(5): 1085-1096.
28. 黄昌宁. 中文信息处理中的分词问题. 语言文字应用, 1997, 1: 74-80.
29. Xianghao, Guo, Zhong Yixin, and Yang Li. A Fast Algorithm for Chinese Words Automatic Segment Based on Two letters word family Structure. Journal of the China Society for Scientific and Technical Information, 1998, 05.
30. 孙茂松, 左正平, 黄昌宁. 汉语自动分词词典机制的实验研究. 中文信息学报 2000, 14(1): 1-6.
31. Dong, W. U. Word Segment and Search Techniques for Chinese Information Search Engines [J]. Computer Applications 7, 2004.
32. Bernhard, Delphine. Unsupervised morphological segmentation based on segment predictability and word segments alignment. Proceedings of 2nd Pascal Challenges Workshop. 2006, 19-24.
33. Siwei, Lai, et al. Chinese Word Segment Based on Character Representation Learning. Journal of Chinese information processing, 2013, 27(5): 8-14.
34. 来斯惟, 徐立恒, 陈玉博, 等. 基于表示学习的中文分词算法探索. 中文信息学报, 2013, 27(5): 8-15.
35. 尹存燕, 黄书剑, 戴新宇, 等. 中英命名实体识别及对齐中的中文分词优化. 电子学报, 2015, 43(8), 1481-1487.
36. 刘开瑛.中文文本自动分词和标注. 商务印书馆, 2000.
37. 黄昌宁, 高剑峰, 李沐. 对自动分词的反思. 语言计算与基于内容的文本处理 2003.
38. Chen, Xinchi, et al. Long Short-Term Memory Neural Networks for Chinese Word Segmentation. EMNLP. 2015:197-1206.
39. Derczynski L, Maynard D, Rizzo G, et al. Analysis of named entity recognition and linking for tweets[J]. Information Processing & Management, 2015, 51(2): 32-49.
40. Prokofyev R, Demartini G, Cudré-Mauroux P. Effective named entity recognition for idiosyncratic web collections[C]. Proceedings of the 23rd international conference on World wide web. ACM, 2014: 397-408.
41. Konkol M, Brychcín T, Konopík M. Latent semantics in Named Entity Recognition[J]. Expert Systems with Applications, 2015, 42(7): 3470-3479.
42. Yao T, Ding W, Erbach G. Correcting word segmentation and part-of-speech tagging errors for Chinese named entity recognition[M]. The Internet Challenge: Technology and Applications. Springer Netherlands, 2002: 29-36.
43. Han X, Sun L, Zhao J. Collective entity linking in web text: a graph-based method[C]. Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval. ACM, 2011: 765-774.
44. Hachey B, Radford W, Nothman J, et al. Evaluating entity linking with Wikipedia[J]. Artificial intelligence, 2013, 194: 130-150.
45. Demartini G, Difallah D E, Cudré-Mauroux P. ZenCrowd: leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking[C]. Proceedings of the 21st international conference on World Wide Web. ACM, 2012: 469-478.
46. Lin T, Etzioni O. Entity linking at web scale[C]. Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction. Association for Computational Linguistics, 2012: 84-88.
47. Bunescu R C, Pasca M. Using Encyclopedic Knowledge for Named entity Disambiguation[C]. In EACL. 2006, 6: 9-16.
48. Han X, Zhao J. Named entity disambiguation by leveraging wikipedia semantic knowledge[C]. Proceedings of the 18th ACM conference on Information and knowledge management. ACM, 2009: 215-224.
49. Zheng Z, Si X, Li F, et al. Entity disambiguation with freebase[C]. Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology-Volume 01. IEEE Computer Society, 2012: 82-89.
50. Alhelbawy A, Gaizauskas R. Graph ranking for collective named entity disambiguation[C]. Annual Meeting of the Association for Computational Linguistics. 2014: 75-80.
51. Nguyen D B, Hoffart J, Theobald M, et al. Aida-light: High-throughput named-entity disambiguation[J]. Linked Data on the Web at WWW2014, 2014.

# 攻读硕士学位期间发表的论文及科研成果

**参与的科研项目**

[1] 国际横向项目：语义分析. 起止时间：2014.6-2015.6

[2] 国家自然科学基金项目：基于粒计算的多源异构动态数据挖掘关键技术研究（项目编号：61573292）. 起止时间：2016.1-2019.12.