# 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" \t "_blank)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。