第一部分

译 前 处 理

文件格式与转换

第1章
Di 1 Zhang

1.1 本章导读

在工作中，我们常常会碰到各类不同的文件格式，了解常用的文件格式并且知道如何通过工具转换文件格式，能够破除文件格式的壁垒，提高工作效率。

无论我们所要处理的文件是哪种格式，其核心要素都是信息与形式。我们在此所说的形式，是指信息呈现与承载的实现形式。例如，我们在读一本电子书时，书中的文字为我们提供了信息，同时，电子书又通过加粗、下划线、分段、编号等各类形式将文字呈现出来。除了信息的呈现形式，我们还要关注信息的承载形式。软件或硬件开发者为了实现特定需求，或是受限于不同时空的技术水平，可能采取不同的技术路径，采用不同的承载形式，推出不同类型的文件格式及版本。诸如文本、音频、视频等不同类型的呈现形式，可能因为厂商不同、代际差异等原因而存在各类不同的文件格式，需要我们借助技术手段加以转换，从而在具体场景中实现信息转换。

本章将介绍的实用技术包括以下内容。

(1) 光学字符识别(OCR)。利用光学字符识别技术，可以将图片等媒介形式中的文字识别出来，变成计算机易于处理的文本。

(2) 语音转文本(STT)。利用语音转文本技术，可以将语音转换为文本，不仅便于快速浏览，而且便于计算机做进一步处理。

(3) 文本转语音(TTS)。利用文本转语音技术，将文字转换成语音，可用于不便阅读的情况，也可用于配音。

这些技术帮助我们突破媒介载体的限制，实现信息呈现形式的转换，不仅是提高信息处理效率的有效方式，还是进一步处理信息的基础。例如，客户提供了一份英语版 MP3 格式文件，希望我们能够提供对应的汉语版 MP3 格式文件，有了语音转文本技术，我们便可以快速听写原文件内容，通过机器翻译快速译为中文，在译后编辑的基础上，利用文本转语音技术为客户提供对应的汉语版 MP3 格式文件。

纯文本(plain text)是我们进行各类格式转换的一种重要中介。纯文本是人与计算机交互的一种基础形式,我们可以认为纯文本只承载信息,而不关注信息的具体实现形式。不过,我们可能也需要在文本格式间做转换。例如,我们有可能需要将 PDF 格式转为 DOC 格式,将图片格式转换为表格,将 DOCX 格式批量转换为 TXT 格式等。不同文本呈现形式的转换也是我们要学习的重点。本章中,我们不仅要学习不同文件格式的转换,还需要学习文件格式转换过程中出现的问题,为种类繁多的文件格式转换奠定自行处理的基础。

在计算机处理信息的过程中,计算机要将二进制的代码转换为人类能够直观读懂的文字,也要将人类所认识的文字转换为二进制的代码,做出存储与处理等一系列操作,这便是计算机编码与解码的过程。在计算机的发展过程中,文字编码不断进化发展,但也依旧存在不少遗留问题,使我们在使用过程中常常碰到乱码问题。因此,除了学习不同格式的转换,我们还有必要了解字符与编码,使我们在碰到乱码问题时,能够有解决思路。

本章将主要实现以下两个目标。

(1) 学习字符与编码,了解出现乱码的原因,学习乱码的常见处理方式。

(2) 学习格式转换,为各类格式的文本处理奠定基础。

1.2 基础知识

1.2.1 字符与编码

我们在计算机上所见到的文字,无论是汉字、字母、数字还是特殊符号,其实都是字符。计算机以二进制的方式实现信息的存储,要在屏幕上呈现具体的文字,就要通过具体的字符编码实现。总体而言,字符编码发展经历了原生、本地化、国际化三个阶段,系统内码也相应经历了 ASCII、ANSI、Unicode 三个阶段,如表 1-1 所示。了解了这三个阶段,我们才能知道乱码产生的原因和解决方法,并且在具体软件的应用过程中,将文件保存为该软件适用的编码格式。

表 1-1 字符编码发展阶段

阶 段	系统内码	说 明
原生	ASCII	ASCII(American Standard Code for Information Interchange,美国信息交换标准代码)是计算机发展早期使用的编码,最早只用于显示英语和一些拓展字符,后拓展至其他西欧语言
本地化	ANSI	ANSI(American National Standards Institute,美国国家标准学会)编码支持的语言范围拓展至汉语等象形文字,由各国家/地区与机构依据这一标准分别制定相应编码,如汉语的 GB 2312、BIG5、GBK 等

续表

阶　段	系统内码	说　　明
国际化	Unicode	Unicode(统一码、万国码、单一码)是国际组织制订的字符编码方案。该编码为各种语言中的每一个字符设定了统一并且唯一的数字编号,理论上可以容纳所有文字和符号,可满足跨语言、跨平台进行文本转换、处理的要求

在原生阶段,计算机刚刚在美国发明,容量非常有限,用于显示的字符也只要能支持英语中的常见字符即可,因此,能够提供 256 个码位的 ASCII 编码足以满足需求。随着计算机技术进一步发展,计算机能够给字符提供的码位进一步增加。同时,计算机开始进入其他国家和地区,需要支持的字符也就更多,编码开始进入本地化阶段,各个地区根据 ANSI 标准编制支持本地字符的编码。在这一阶段,虽然计算机的容量有所提升,但能够提供给字符的码位也依旧有限,因此,同样的码位在不同的地区所代表的字符就可能有所不同。此时,一封从日本发出的邮件,在中国的计算机上打开时,见到的可能是汉字,但合在一起并不能提供有意义的信息,也就出现了我们常说的乱码现象。此后,计算机容量进一步提升,完全能够为所有字符都分配一个独有的码位,字符编码的国际化阶段也随之而来,Unicode 系列编码(包括 UTF-8、UTF-16、UTF-32 等)理论上能够为所有字符提供唯一码位,乱码问题也就随之减少。

然而,我们常常会遇到文件、网页、软件中存在乱码的问题,其主要原因仍然是编码和解码时遵循的编码方式不统一。同时,一些软件由于编写时采用的编码格式并非 Unicode,在使用时可能也无法为 Unicode 提供支持。遇到乱码,我们通常可以找到正确的解码方式,将乱码变成我们可以直接解读的文字或符号。

1.2.2　常见媒介类型

常见媒介类型是指我们在计算机上处理语言服务相关信息所经常涉及的媒介类型,主要包含文本、语音、图片三种类型。这三种类型是我们处理相关信息的基础形式,掌握其处理方式就有了进一步处理视频等其他格式的基础。

文本是计算机最易于处理的媒介形式,大量的数据信息处理都以文本为中介完成,计算机要完成各类格式的转换,通常都离不开文本。无论是 OCR、STT 还是 TTS,本身都离不开文本:OCR 和 STT 生成的是可编辑的文本,TTS 则利用文本生成语音。

语音的优势是直观,语音输入与语音转文本技术解放了我们的双手与双耳,极大地提高了信息传输的效率。我们有时需要将语音转换成文本,或者将一种语言的语音转换成另一种语言的语音,还有时候需要将文本转换成语音。这些工作通常通过人工听录、人工配音等方式完成,但若时间紧迫,或预算不足,这些工作也可以借助相关技术完成,虽然质量不如人工,但可以满足基本需求。

以图片形式呈现的文字是计算机无法理解的,因而通常要将图片上的文字转换成可编辑的字符,才能依靠计算机处理相关信息。例如,客户提供了一张照片,要求我们将上面的英文快速翻译为中文以供参考,此时首先要将图片转换为可编辑的文字,再通过机器翻译完成基本信息的转换。

1.3 案例解析

了解编码格式后，我们就有了处理乱码所需的基础知识。在本节中，我们将学习部分乱码情况的处理方式。此外，我们还将讲解语音、文本、图片等不同媒介类型之间的相互转换。

1.3.1 乱码处理

如前所述，乱码主要是由于编码与解码不一致导致的，通过更改编码方式或解码方式使二者一致，便可解决乱码问题。在翻译软件中，我们通常难以更改解码方式，此时我们需要通过更改编码的方式来解决乱码问题；在网页中，我们通常难以更改编码方式，此时我们就需要通过更改解码方式来解决乱码问题。

1. TXT 编码更改

TXT 格式是纯文本的保存格式，常见的文本编辑器，如 Word、Windows 系统自带的记事本等都可以直接打开。用记事本打开 TXT 文档时，右下角会显示该文档的编码格式。当我们需要修改文件编码格式时，我们可以单击“文件”，选择“另存为”，并在对话框中将编码格式设置为所需的编码。

这种方式不仅用于解决 TXT 文件的乱码问题，也用于更改 TXT 格式的文档编码，使其符合处理工具的要求。例如，北京外国语大学的 ParaConc 语料处理软件仅支持 ANSI 编码，需要将 Unicode 格式的文档照这一方法转换成 ANSI 格式，如图 1-1 所示。

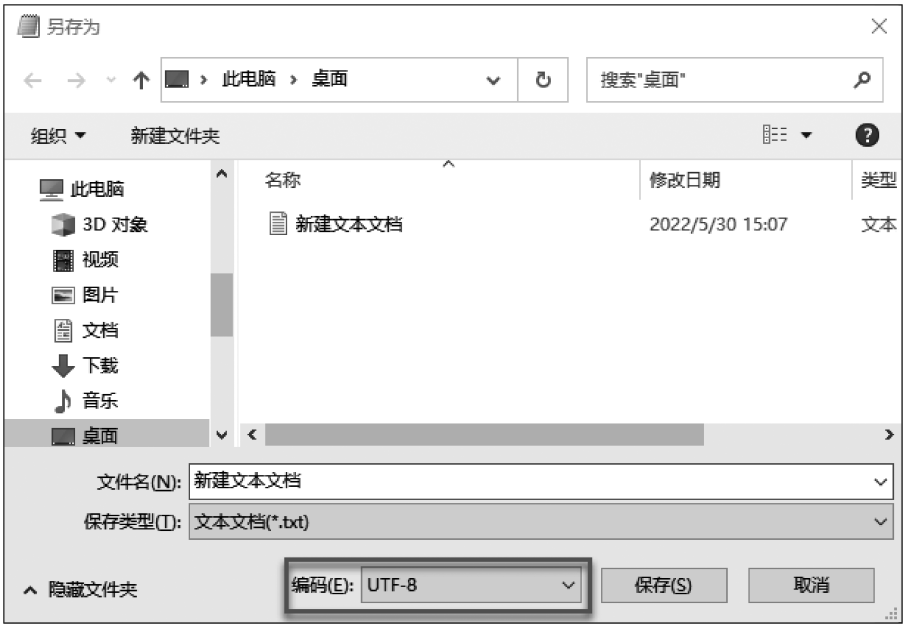

图 1-1 另存为 TXT 格式以更改编码

2. CSV 乱码纠正

如图 1-2 所示，一份从数据库中导出的 CSV 文件出现了乱码，基本可以判断是编码不正确导致的，将该文件保存为正确的编码格式即可。

Filename	Run Count	Last Run Date
D:\software\tianruoV4.47\澶十婁OCR鉱囧瓩璇嗗埆.exe	6	1.329E+17
D:\software\ofiice2016\office2016娉∠晰鏈篭KMSAuto Net.exe	5	1.32902E+17
\\BON-LION\Brian\Project\瀹炵敤缈昏瘧鎶€鏈　晆 绋?,4,132902530154855306 \\BON-LION\usbshare1-1\骞垮窨鏂匽柃绉匽妧鏈夐檺璐ｄ换鍏绽徃\骞垮　鍩哄湴鍏ラ┒\2021 鍩哄湴鍏ラ┖"	4	1.32865E+17
D:\software\EmEditor Professional v13.0.5\EmEditor\EmEditor.exe	4	1.32859E+17
C:\$RECYCLE.BIN\S-1-5-21-3673421196-2154311799-2778244210-500\	3	1.32902E+17
C:\Users\Administrator\Desktop\鏂板缓鏂囦欢澶?(2)\20211124 绗絺簱绔?	3	1.32902E+17
C:\Users\Administrator\Desktop\~$211124 绗絺簱绔?鏈哄櫒缈昏瘧搴旂敤	3	1.32902E+17
E:\鏂匽柃绉匽妧logo	3	1.32846E+17
E:\閲戝０璇跌■璧勮骇绠＄悊骞冲彴\璇跌■浠ｇ爣琛?xlsx	2	1.32903E+17

图 1-2　CSV 乱码示例

出现这种情况是因为在使用 Excel 打开 CSV 文件时，需要利用文件头的 BOM 来识别编码，如果保存的 CSV 没有 BOM 文件头编码，则 Excel 会按照默认编码读取文件，如果文件编码与 Excel 的默认编码不一致，则会出现乱码。在这一案例中，我们只需使用记事本将该 CSV 文件打开，另存为"带有 BOM 的 UTF-8"编码，再使用 Excel 打开时，即可正确读取文件。

3. 网页乱码纠正

有时，一些网站没有使用最新的 UTF 编码，很有可能出现乱码问题。此时，我们只需要将浏览器的解码方式更改为相应的编码，就可以解决乱码问题。IE、360 等浏览器可以通过右击网页直接更改编码，将浏览器所使用的编码方式调整成与网页一致的编码即可解决乱码问题。Chrome、Edge 等浏览器不支持直接右击更改编码，可考虑暂用其他浏览器或使用高级设置更改编码。

1.3.2　光学字符识别

光学字符识别(optical character recognition，OCR)技术是将图片中的文字转换为可编辑文字的一种技术手段。我们在翻译或日常办公中常常要用到光学字符识别技术。例如，客户提供了一本纸质书需要翻译，我们通常不会直接对照书本翻译。对照书本的内容进行翻译，不仅过程烦琐，无法利用划词等方式快捷查词，而且不利于机器翻译、计算机辅助翻译、机器检查等翻译技术的应用。此时，我们通常先将纸质书扫描成电子版，再利用光学识别技术将字符转换为可编辑的文本文件，然后在此基础上选用计算机辅助翻译或机器翻译等形式完成翻译。

1. PDF 转 Word

PDF 是一种常见的文档格式，因其在不同计算机中打开皆可保持原有样式的特性而被广泛使用。在翻译活动中，我们常常会收到客户发来的 PDF 文件，要求我们提供报价。即使是我们自己需要翻译内容，也通常需要在转换出准确文本的基础上进行。对于可编

辑的 PDF 文档，我们有多种方式可以转换为 Word 文档，以便统计字数，完成翻译。以下介绍利用 ILovePDF 网站转换可编辑 PDF 文档的方法。

ILovePDF(ilovepdf. com)是集成多种 PDF 文档处理功能于一体的 SaaS 网站，可以实现 PDF 压缩、拆分、合并、转换等多种功能，如图 1-3 所示。其中，PDF 转 Word 功能每次可免费处理一份 PDF 文档并免费下载。需要注意的是，该功能虽然支持 OCR，但是不支持中文字符的识别，因此应在转换可编辑的 PDF 文档时使用。对于不可编辑的扫描文档，可以使用 ABBYY FineReader 等专业工具做识别转换。

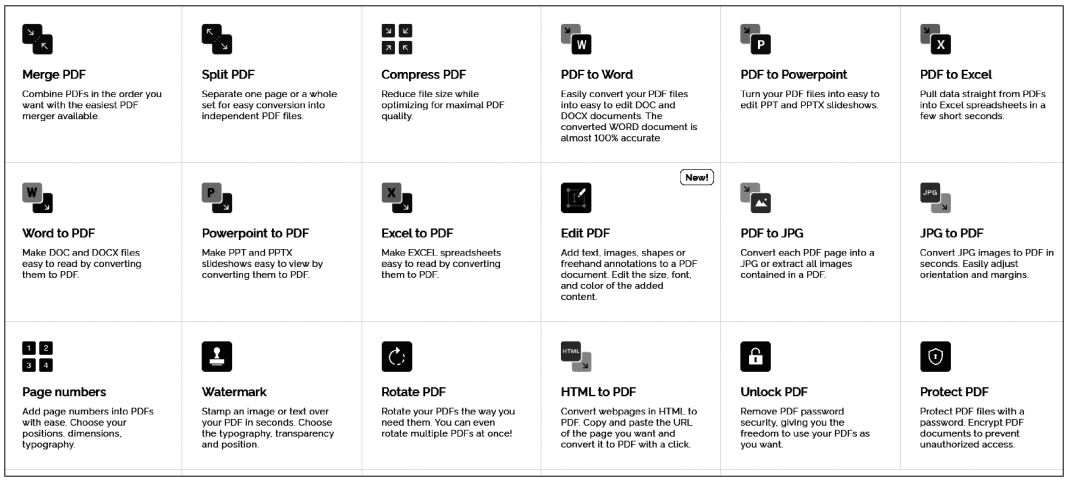

图 1-3　ILovePDF 网站首页展示的部分功能

2. 搜狗 OCR

我们知道，QQ 截图、微信图片的“提取文字”等方式都可以将图片中的文字提取出来，变为可编辑内容。这些方式非常适合用来识别临时性的少量内容，在特定的场景中也非常方便。不过，有时我们可能不想打开专门的软件来做识别，此时利用已经打开的搜狗输入法来截图识别就十分方便。

在搜狗输入法激活状态下按下 Ctrl＋Shift＋M 快捷键，即可调出搜狗输入法菜单。调出菜单后，再按下 Ctrl＋Shift＋O 快捷键，即可调出搜狗输入法的智能输入助手，如图 1-4 所示。利用搜狗智能输入助手，我们可以利用“图片转文本”或“截屏”完成识别转换过程，获得可编辑的文字。若识别内容为完整文字，但有一定数量的错误识别内容，我们还可以复制文字内容，将其粘贴到“智能写作”中，由搜狗提供智能纠错建议，以较快的方式识别常见错误。如果有翻译需求，我们也可以直接在“在线翻译”中粘贴文字，完成多语言的翻译。

利用搜狗智能输入助手的这一过程，实际上与我们平时完成翻译的完整过程类似。首先利用技术完成待译文本的处理，使其易于处理；其次利用机器完成“入口检查”，确保输入的内容没有错误；最后在完成处理的基础上完成翻译。虽然这些技术的处理准确度还有待提升，但是利用这些技术，可以极大减轻人工处理的负担，从而提高处理效率。

图 1-4　智能输入助手

3. 天若 OCR

天若 OCR 是一款接入了多种光学字符识别应用程序接口(API)的集成 OCR 工具，可实现手写字体识别、表格识别等多种功能，分为免费版和专业版。除了官方推出的版本，另有基于早期免费版推出的开源版，此处讲解开源版。该软件与搜狗 OCR 类似，适合用来处理小批量的临时性需求，如图 1-5 所示。

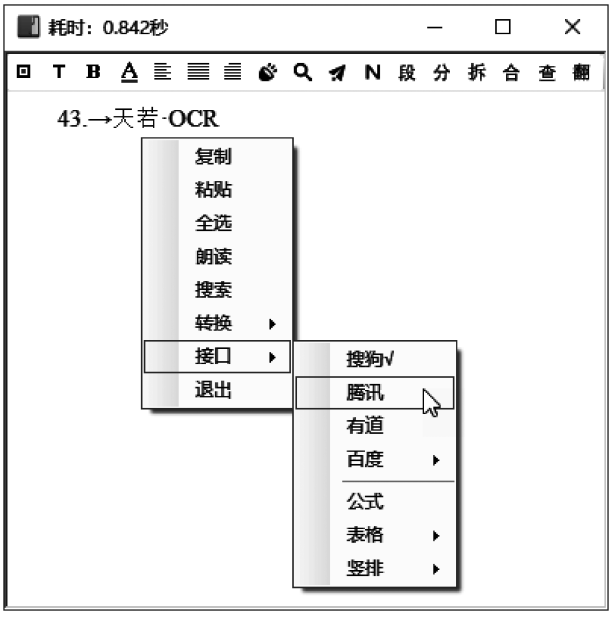

图 1-5　天若 OCR 软件

下载软件解压后，无须安装，只需双击 exe 图标，即可启动软件。该软件启动后会自动最小化至任务栏右侧，需要使用时，可按 F4 快捷键或双击任务栏上的软件图标，即可开始截图。按下鼠标左键，拖动截图区域，即可完成截图。完成截图后，将会跳出识别结果窗口。由于开源版调用的是免费的 OCR 识别接口，因此可能出现接口无法使用的情况。此时可以在截图结果窗口右击，在接口中选择“搜狗”或“有道”，再次尝试即可。在这一菜单中，还有转换功能，可实现中英文标点符号、简繁体、英文大小写转换，而且可将汉字转为拼音。

由于开源版缺少维护，目前有不少附加功能已经无法使用，不过其核心功能依旧可以使用，如识别文字后的拆分、合并等功能，可以分别点击识别结果中的“拆”“合”图标，从而获得更好的段落拆分格式。

1.3.3 语音转文本

语音转文本（speech to text，STT）又称自动语音识别（automatic speech recognition，ASR），可以利用语音识别技术将语音转换为文字，在语音控制、语音听写、语音转写等领域被广泛应用。利用语音听写，我们可以将语音内容转录为文字，并在此基础上进行编辑，这一技术在语言服务领域已经得到较为广泛的应用。例如，要在视频会议中提供双语实时字幕，首先要做的就是将语音内容听写为文字，然后在文字基础上提供机器翻译译文。语音转写则是将已有的录音转换为文字，与语音听写没有本质区别，可以用于会议记录生成、字幕生成等场景。本节主要以 Windows 系统自带的语音输入功能、搜狗语音输入为例讲解语音听写，以搜狗听写讲解语音转写，但不介绍利用智能语音鼠标、录音笔等硬件提供的语音输入与语音转写。

1. Windows 系统语音输入

Windows 系统自带的语音转文本功能可以通过按下 Win＋H 快捷键调用。按下该快捷键后，桌面顶端将出现语音识别工具栏，左击需要输入文字的位置，即可通过语音方式输入内容。系统自带的语音输入方式是一种比较适合用来叙述思路的快捷输入方式。不过这种输入方式的准确度可能不高，如果要获得准确文本，需要在后期做较多的编辑工作。与其他语音输入方式相比，这种语音输入方式也有它的优势。其最主要的优势在于系统自带的语音输入方式实际上结合了语音命令功能，能够通过语音命令控制我们的一些操作。例如，在开启语音输入的情况下，我们可以通过语音输入功能在 Word 当中选中某一个段落的内容或者选中全文。

2. 搜狗语音输入

搜狗语音输入是搜狗语音输入法自带的一种语音输入方式。利用搜狗语音输入，我们可以快速将语音转为文字。如果搜狗输入法的状态栏未出现在屏幕上，我们首先要在任务栏右侧单击搜狗输入法的图标，调出输入法状态栏，再单击输入法状态栏中的麦克风图标调出语音输入面板，然后通过单击语音按钮或按下 F2 快捷键即可开启语音输入，如图 1-6 所示。

如有需要，可通过输入法状态栏左边的搜狗输入法图标进行语音输入快捷键设置，以