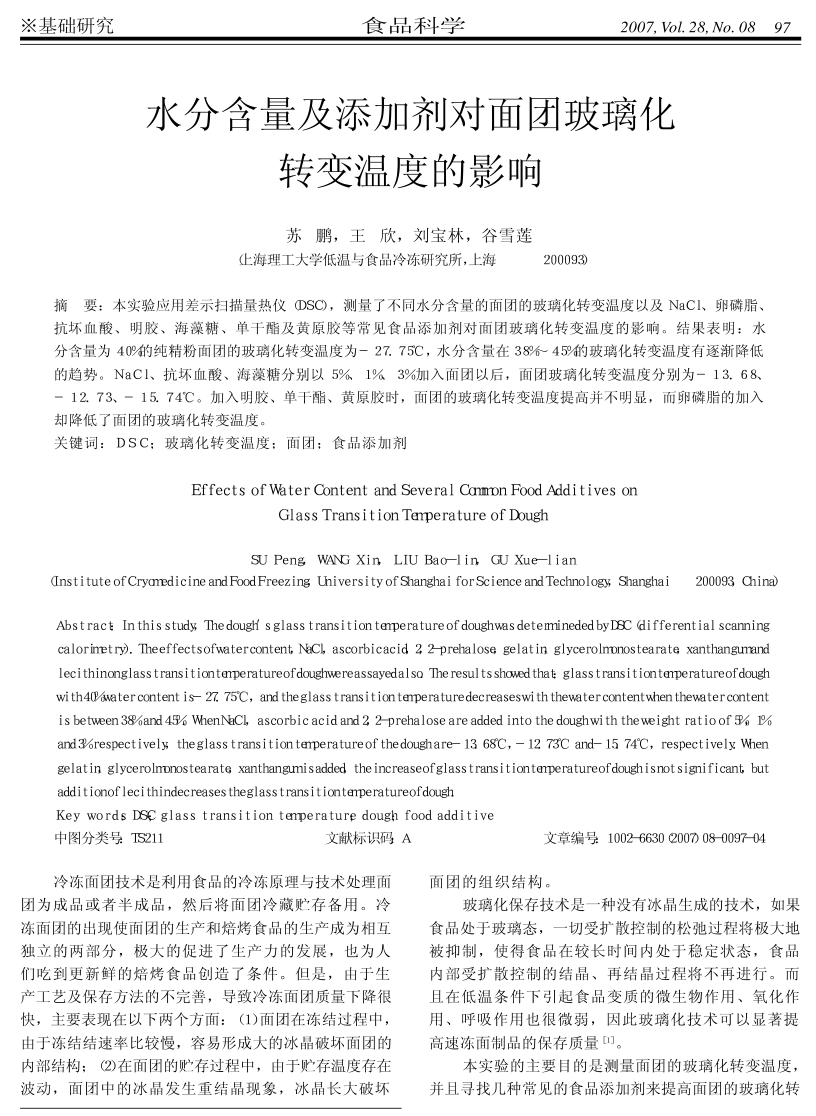
## **一个文献挖掘的案例及流程**

北京迈高材云科技有限公司

## 文献原文件

原始文献pdf文件，此时还不是模型能够理解的数据。



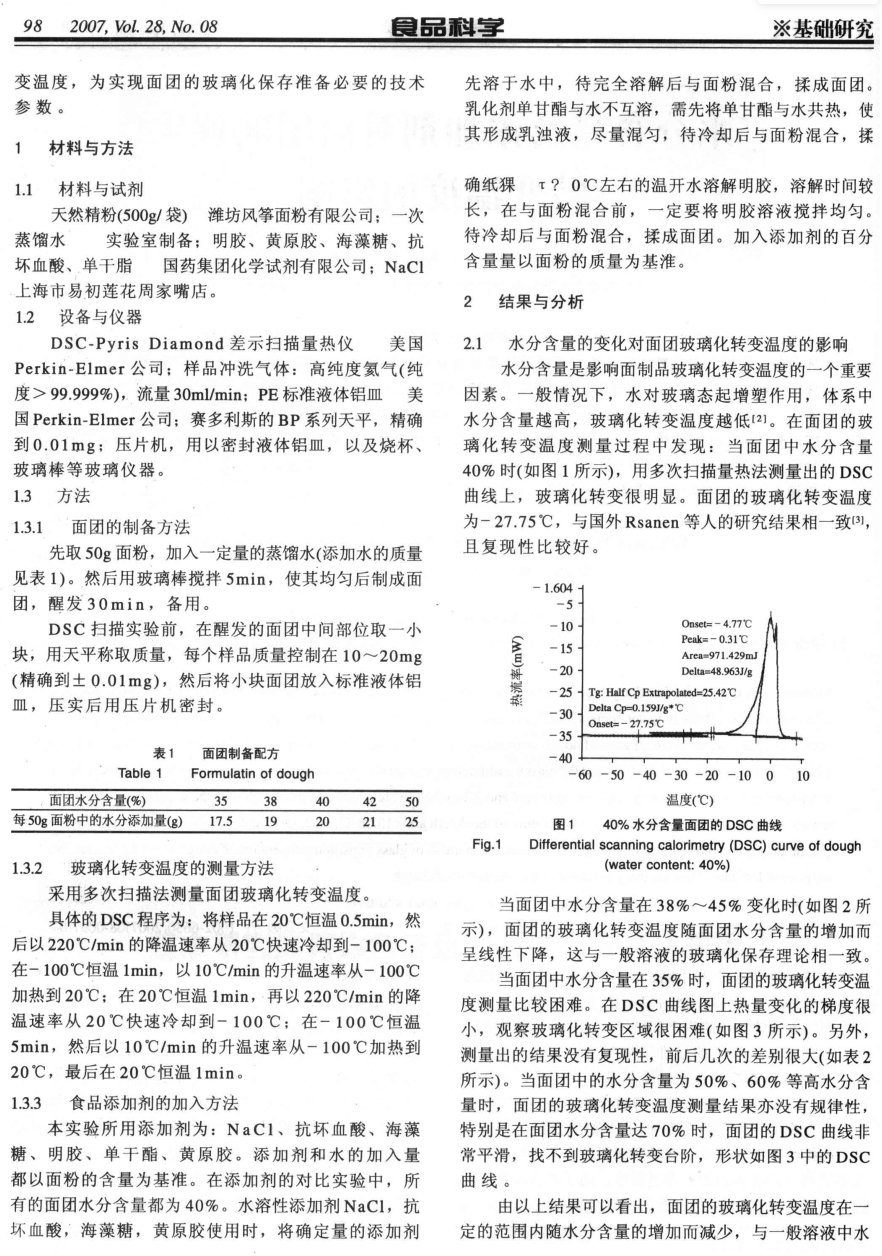


图 1 原始文献

## 文本识别和提取构建语料库

使用图像识别算法，识别出文献中的文本数据，作为后续信息抽取的基础语料库。

首先会进行布局识别， 识别出文本、表格、以及图像。

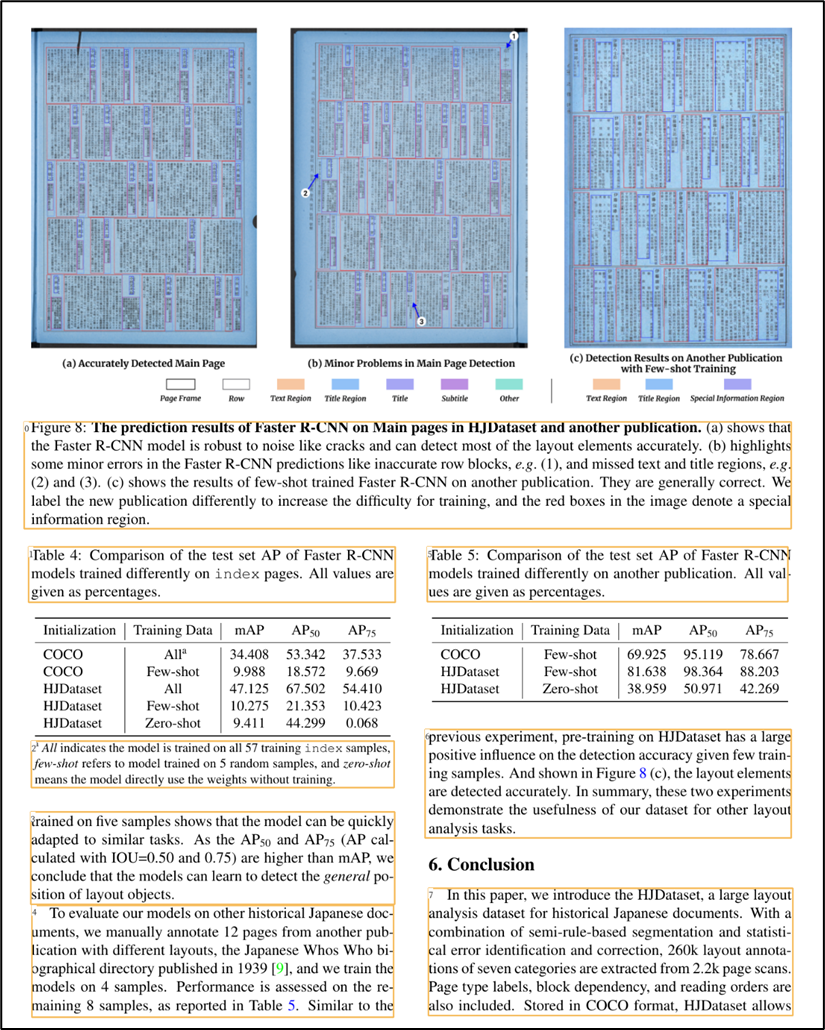


图2 布局检测， 标注出文本，检测出的文字包用黄线标注 （这里以其它文献为例）

经布局检测和文本识别后，形成的文献语料库样例如下：

“*本实验应用差示扫描量热仪 ODSC，测量了不同水分含量的而团的玻璃化转变温度以及 NaCl、卵磷脂、抗坏血酸、明胶、海藻糖、单千酯及黄原胶等常见食品添加剂对面团玻璃化转变温度的影响*。”

“*水分含量是影响面制品玻璃化转变温度的一个重要因素。一般情况下，水对玻璃态起增塑作用，体系中水分含量越高，玻璃化转变温度越低。在面团的玻璃化转变温度测量过程中发现：当面团中水分含量 40%时，用多次扫描量热法测量出的DSC 曲线上，玻璃化转变很明显。面团的玻璃化转变温度为-27.75°C。*”

“*先取50g 面粉， 加入一定程度的蒸馏水，（添加水的质量见表 1），然后用玻璃棒搅拌 5min, 使其均匀后制成面团，醒发 30min, 备用。*” …

## 数据标注

数据标注，可用于信息抽取，可用于文本／语料分类。

假设我们需要提取如“玻璃化转变温度“的值， 以及“面团的制备工艺” 。

一般需要人工标注，找出哪些语料片段含有“玻璃化转变温度“及其数值，”面粉工艺“等信息。一般可通过在线标注平台工具（如label studio、doccano），进行人工标注。可以设定相应的标注任务，降低标注成本。无需对所有的语料进行标注，视语料库的大小，如仅标注１０００－２０００条。

下图为在线平台标注工具，人工进行数据标注的界面。

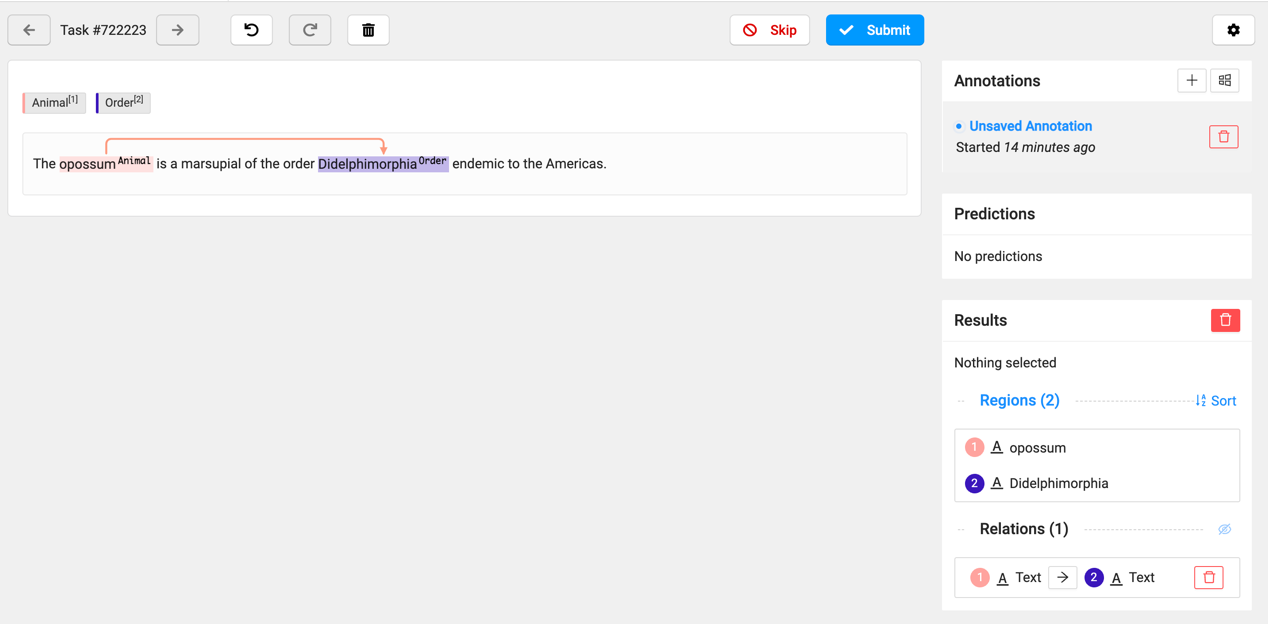


图1 使用标注工具可以方便的从文本中框选所需要的文本数据，后续通过该标注工具可以批量处理为训练数据。图中颜色显示的部分为标注的数据。

标记后的文本如下：

“*水分含量是影响面制品玻璃化转变温度的一个重要因素。一般情况下，水对玻璃态起增塑作用，体系中水分含量越高，玻璃化转变温度越低。在面团的玻璃化转变温度测量过程中发现：当面团中水分含量 40%时，用多次扫描量热法测量出的DSC 曲线上，玻璃化转变很明显。面团的玻璃化转变温度为-27.75°C*”。

“*先取50g 面粉， 加入一定程度的蒸馏水，（添加水的质量见表 1），然后用玻璃棒搅拌 5min, 使其均匀后制成面团，醒发 30min, 备用。*”

标注后的语料段，可用于构造用于语料筛选模型的指令微调数据，和用于文献数据提取模型的指令微调数据（详见下）。

## 用于“文献数据提取模型”的指令性微调数据构造

大模型能够理解自然语言的指令，需要编写指令作为输入，告诉模型需要抽取的信息以及格式。指令需要包含以下几个重点：

1. 模型所扮演的角色或任务，指明模型需要扮演的工具以及需要完成的任务。
2. 模型需要提取信息的具体内容范围，指明模型抽取信息的纲要，即需要抽取哪些信息
3. 模型输出的格式，指明需要模型输出的内容，一般为格式化的信息输出。

指令微调包括定义输入和输出。一般可用JSON格式（模型和人均可识别）。一个完整的JSON指令微调数据示例，见如下：

｛“输入“：”你是一个基础模型，请按如下格式提取文献中的有用数据信息：{“玻璃化转变温度“：“”}，输出即为格式化的信息。下面给出文献：

“水分含量是影响面制品玻璃化转变温度的一个重要因素。一般情况下，水对玻璃态起增塑作用，体系中水分含量越高，玻璃化转变温度越低。在面团的玻璃化转变温度测量过程中发现：当面团中水分含量 40%时，用多次扫描量热法测量出的DSC 曲线上，玻璃化转变很明显。面团的玻璃化转变温度为-27.75°C。“

文献信息输出:，”输出“：“{“玻璃化转变温度“：“-27.75°C”}”｝

所有的标注数据，可用于构造一批指令性微调数据，用于基于LAMMA 或GLM进行文献数据提取模型的微调。

## 用于“语料筛选模型”的指令性微调数据构造

语料筛选模型主要用于把含有所需信息的语料片段，从语料库中找出来。

语料筛选模型， 可基于LAMMA（英文语料）或GLM（中文语料） 微调得到。因此需要构造指令微调数据，用于微调语料筛选模型。

语料筛选模型的指令性微调数据，输入是语料片段，输出是需含有所需信息的语料片段。一个JSON格式的指令性微调数据，示例如下

{“输入“：”你是一个基础模型，请按如下格式筛选语料：{“含玻璃化转变温度的语料片段“”}，输出即为格式化的信息。下面给出文献：

“水分含量是影响面制品玻璃化转变温度的一个重要因素。一般情况下，水对玻璃态起增塑作用，体系中水分含量越高，玻璃化转变温度越低。在面团的玻璃化转变温度测量过程中发现：当面团中水分含量 40%时，用多次扫描量热法测量出的DSC 曲线上，玻璃化转变很明显。面团的玻璃化转变温度为-27.75°C。“

“先取50g 面粉， 加入一定程度的蒸馏水，（添加水的质量见表 1），然后用玻璃棒搅拌 5min, 使其均匀后制成面团，醒发 30min, 备用。”

含所需信息的语料片段:，”输出“：“{“含玻璃化转变温度的语料片段“：水分含量是影响面制品玻璃化转变温度的一个重要因素。一般情况下，水对玻璃态起增塑作用，体系中水分含量越高，玻璃化转变温度越低。在面团的玻璃化转变温度测量过程中发现：当面团中水分含量 40%时，用多次扫描量热法测量出的DSC 曲线上，玻璃化转变很明显。面团的玻璃化转变温度为-27.75°C。”}”｝

## 微调得到“语料筛选模型”

语料筛选模型主要用于把含有所需信息的语料片段，从语料库中找出来。

语料筛选模型，可基于上述构造的，用于语料筛选模型的指令性微调数据，基于LAMMA 或GLM，微调得到。

## 微调得到“文献数据提取模型”

文献数据提取模型，指用于从经语料筛选模型后的、含有所需信息的语料库中，提取数据。

文献数据提取模型，可基于上述构造的，用于文献数据提取模型的指令性微调数据，基于LAMMA 或GLM，微调得到。

## 文献数据提取

将经“语料筛选模型”筛选后的、含有所需信息的语料库，提供给文献数据提取模型， 进行数据提取，最终得到所需的数据。