

CCKS 2021 技术评测任务书

面向保险领域的低资源文档信息抽取

一、任务描述

随着知识图谱在各个行业的应用，使用信息抽取技术从文档中挖掘大规模图谱知识已经成为了一个研究热点。然而当前信息抽取任务多关注于从无结构化文本中抽取知识，忽略了文档中丰富的格式信息。同时，在真实的业务场景下，文档级信息抽取标注数据的获取成本十分高昂，现有信息抽取方法在少量标注样本下的表现性能不佳，已经成为阻碍行业知识图谱规模化应用的主要瓶颈之一。为此，阿里巴巴和中国人寿联合举办“面向保险领域的低资源文档信息抽取”评测任务，希望通过此次评测任务及开放的中文文档信息抽取数据集，助力中文文档信息抽取技术的发展及其商业化落地应用。

本次评测面向真实业务场景下的商业文档理解需求，提供保险领域的少量标注数据和大量无标注数据，期望参与者能够充分利用文档格式信息，在低资源标注场景下完成“文档标题层级抽取”和“文档开放信息抽取”两个子任务。

输入：文档的 PDF 格式文件及对应的解析后文件

示例：

一、文档的 PDF 格式文件（部分）

中国人寿保险股份有限公司 国寿 e 家吉祥送福综合意外伤害保险（2013 版）利益条款

第一条 保险合同构成

国寿 e 家吉祥送福综合意外伤害保险（2013 版）合同（以下简称本合同）由保险单及所附国寿 e 家吉祥送福综合意外伤害保险（2013 版）利益条款（以下简称本合同利益条款）、短期保险基本条款（以下简称本合同基本条款）、批注、附贴批单、投保单，以及与本合同有关的投保文件、声明和其他书面协议共同构成。

第二条 投保范围

凡出生二十八日以上、七十周岁以下的身体健康者，均可作为被保险人，由具有完全民事行为能力的本人或对其具有保险利益的其他人作为投保人，向本公司投保本保险。

第三条 保险期间和续保

本合同的保险期间最长为一年；除另有约定外，自本合同生效之日起至约定终止日二十四时止。

投保人可于保险期间届满前或在本合同约定的交费宽限期内，经本公司同意后，向本公司交付续保保险费，本合同于保险期间届满的次日起延续有效。本合同可按上述方式续保至被保险人年满七十五周岁后的第一个生效对应日。

本公司保留终止本合同续保的权利，并有权调整保险费收费标准。

第四条 保险责任

本合同的保险责任分为基本责任和可选责任两部分，其中基本责任必须投保，可选责任可以选择投保，由投保人在投保时确定并在保险单上载明，本合同的保险责任一经确定，在保险期间内不得变更。

一、基本责任

1. 意外伤害身故保险责任

在本合同保险期间内，被保险人遭受意外伤害，并自该意外伤害发生之日起一百八十日内因该意外伤害导致身故的，若该被保险人身故时已满十八周岁，本公司按本合同约定的基本保

二、 解析后的文件（部分, 分别为 word, x0, y0, x1, y1, fontsize fontname）：

国	81.12	790.3197	87.600006	793.8383	6.0	DPNNBC+YouYuan
寿	87.600006	790.3197	94.08001	793.8383	6.0	DPNNBC+YouYuan
e	95.70001	790.3197	98.94001	793.8383	6.0	DPNNBC+YouYuan
家	100.56001	790.3197	107.040016	793.8383	6.0	DPNNBC+YouYuan
吉	107.10028	790.3197	113.580284	793.8383	6.0	DPNNBC+YouYuan
祥	113.580284	790.3197	120.06029	793.8383	6.0	DPNNBC+YouYuan
送	120.06029	790.3197	126.54029	793.8383	6.0	DPNNBC+YouYuan
福	126.600555	790.3197	133.08055	793.8383	6.0	DPNNBC+YouYuan
综	133.08055	790.3197	139.56055	793.8383	6.0	DPNNBC+YouYuan
合	139.6208	790.3197	146.1008	793.8383	6.0	DPNNBC+YouYuan
意	146.1008	790.3197	152.5808	793.8383	6.0	DPNNBC+YouYuan
外	152.5808	790.3197	159.06079	793.8383	6.0	DPNNBC+YouYuan
伤	159.06079	790.3197	165.54079	793.8383	6.0	DPNNBC+YouYuan

子任务一：文档标题层级抽取

输出：文档内所有的标题及其目录层级编号

示例：

{“1”，“保险合同构成”}
{“2”，“投保范围”}
{“3”，“保险期间和续保”}
{“4”，“保险责任”}
{“4.1”，“基本责任”}
{“4.1.1”，“意外伤害身故责任”}

注：当文档中标题无数字编号时，参赛者需根据文档生成对应的数字编号。

子任务二：文档开放信息抽取

输出：文档内以当前保险产品为头实体的所有属性-属性值对

示例：

{“保险合同构成”，“国寿 e 家吉祥送福综合意外伤害保险（2013 版）合同（以下简称本合同）由保险单及所附国寿 e 家吉祥送福综合意外伤害保险（2013 版）利益条款（以下简称本合同利益条款）、短期保险基本条款（以下简称本合同基本条款）、批注、附贴批单、投保单，以及与本合同有关的投保文件、声明和其他书面协议共同构成。”}

{“投保范围”，“凡出生二十八日以上、七十周岁以下的身体健康者，均可作为被保险人，由具有完全民事行为能力的本人或对其具有保险利益的其他人作为投保人，向本公司投保本保险。”}

{“保险期间和续保”，“本合同的保险期间最长为一年；除另有约定外，自本合同生效之日起至约定终止日二十四时止。

投保人可于保险期间届满前或在本合同约定的交费宽限期内，经本公司同意后，向本公司交付续保保险费，本合同于保险期间届满的次日起延续有效。本合同可按上述方式续保至被保险人年满七十五周岁后的第一个生效对应日。\\n 本公司保留终止本合同续保的权利，并有权调整保险费收费标准。”}

三、数据描述

本次数据来自真实保险产品文档，由专业人员标注，训练集、验证集及测试集的说明如下：

训练集：共包含 50 篇有标注和 2000 篇无标注的文档，每篇文档都包含 PDF 格式的文档及其对应的解析后文件，用于竞赛模型训练。其中，有标注文档还带有额外的标签文件。

验证集：共包含 50 篇有标注文档，用于竞赛模型训练和参数调试。

测试集：共包含 100 篇有标注文档，为本次评测的最终测试集。另外为了防止针对测试集的调试，数据中将会额外加入混淆数据。该部分数据在评测结束前一周发布。

四、评价指标

本次评测采用调整后的精准率（Precision,P）、调整后的召回率（Recall， R）、F1 值（F1-measure,F1）来评估文档信息抽取的效果：

$$\text{精准率 (precision)} = \frac{\sum_{\text{抽取的元组 } i}^m x_i}{\text{抽取的元组总数量 } m}$$

$$\text{召回率 (recall)} = \frac{\sum_{\text{抽取的元组 } i}^m x_i}{\text{标注的元组总数量 } n}$$

$$F1 = \frac{2 * \text{precision} * \text{recall}}{(\text{precision} + \text{recall})}$$

在子任务一中，如果元组*i*和标注答案中的一个元组严格匹配， x_i 为 1，否则为 0， n 为标注的元组总数量。在子任务二中，属性采用严格匹配的方式，属性值采用 gestalt 匹配^[1]进行模糊计算：

$$x_i = 0.5 * \text{strict_match}(\text{元组 } i \text{ 的属性, 元组 } j \text{ 的属性}) + 0.5 * \text{gestalt_match}(\text{元组 } i \text{ 的属性值, 元组 } j \text{ 的属性值})$$

其中元组*j* ∈ 标准答案，是在保证 $\sum_{\text{抽取的元组 } i}^m x_i$ 最大的条件下，通过求解预测元组与标准答案元组间的线性匹配问题^[2]得到的与元组*i*匹配的标准答案中的元组。

以 Python 为例，两个字符串的 gestalt 匹配得分计算方法如下：

```
import difflib
a="阿里巴巴达摩院"
b="阿里巴巴"
score = difflib.SequenceMatcher(None, a, b).ratio()
```

[1] Ratcliff J W, Metzener D E. Pattern-matching-the gestalt approach[J]. Dr Dobbs Journal, 1988, 13(7): 46-&.

[2] Wikipedia. 2017. Assignment problem— Wikipedia, The Free Encyclopedia. (2017) https://en.wikipedia.org/wiki/Assignment_problem

五、任务提交

提交方式：

本次任务将采取刷榜的方式，在评测阶段，首先分别给出提交结果在两个子任务上的得分，并采用宏平均作为最终计算系统综合得分的方法。

第一阶段：验证集发布后，允许参赛队伍多次向平台提交结果。参赛者将所有预测结果写入一个结果文件中上传给系统（每日最多可上传 3 次），如果新提交结果好于之前提交结果，排行榜中的成绩将自动进行更新覆盖。

第二阶段：测试集发布后，验证集提交通道将会关闭，其余事项与第一阶段相同。

提交格式：

选手将结果保存为 result.txt，以 utf-8 编码格式保存，每行为一个 JSON 字符串，形式如下。

```
{
  "docid": "文档编号",
  "title": [{
    "number": "文档标题的层级编号",
```

```
    "text": "文档标题的文本内容，需要与 number 相对应",
  }, .....],
  "knowledge": [{
    "key": "当前保险产品的属性（一般为文档子标题）",
    "value": "当前保险产品的属性值，需要与 key 相对应",
  }, .....]
}
```

最终提交文件要求：

每一个参赛队需提交的材料如下：

1. 运行结果文件
2. 代码及说明
3. 方法描述文档

代码及其文档需打包成一个文件（tar, zip, gzip, rar 等均可），用“队名-code.xxx”命名，要求提交所有的程序代码及相关的配置说明，确保程序能够正确运行，且所得结果需与最好成绩的 result.txt 相符。

评测平台：

本次评测将依托阿里天池平台（<https://tianchi.aliyun.com/>）展开，请有意向的参赛队伍关注平台上的竞赛列表。

六、时间安排

- 任务征集截止：3 月 15 日
- 任务准备时间：3 月 20 日—4 月 1 日
- 评测任务发布：4 月 1 日
- 报名时间：4 月 1 日—7 月 15 日
- 训练及验证数据发布：4 月 20 日
- 测试数据发布：7 月 15 日
- 提交测试结果：7 月 20 日
- 评测论文提交：8 月 5 日
- CCKS 会议日期(评测报告及颁奖)：8 月 18 日—21 日

七、规则

1. 报名方式：用淘宝或阿里云账号登入天池官网，完成个人信息注册，即可报名参赛。本次比赛的参赛对象仅限全日制在校大学生（本科、硕士、博士均可）和企业员工；
2. 参赛选手需确保注册时提交信息准确有效，所有的比赛资格及奖金支付均以提交信息为准；
3. 参赛选手在管理系统中组队，参赛队伍成员数量不得超过 5 个，报名截止日期之后不允许更改队员名单；
4. 每支队伍需指定一名队长，队伍名称不超过 15 个字符，队伍名的设定不得违反中国法律法规或公序良俗词汇，否则组织者有可能会解散队伍；
5. 每名选手只能参加一支队伍，一旦发现某选手以注册多个账号的方式参加多支队伍，将取消相关队伍的参赛资格；

6. 允许使用开源代码或工具,但不允许使用任何未公开发布或需要授权的代码或工具;
7. 除主办方提供的数据集外,参赛选手允许使用预训练数据(如词向量、字向量等),但是不能用额外的标注数据;

八、任务组织者及联系人

组织者: 阿里巴巴达摩院机器智能技术、中国人寿保险股份有限公司

联系人: 唐呈光 chengguang.tcg@alibaba-inc.com

石薇薇 shiweiwei.sww@alibaba-inc.com

届时会开放钉钉群作为任务讨论平台。

九、奖项安排

一等奖(1名): 15000

二等奖(2名): 10000

三等奖(3名): 5000

创新技术奖(2名): 3000