# [CCKS-千言通用信息抽取竞赛](http://wiki.baidu.com/pages/viewpage.action?pageId=1833484125)

**会议地点：秦皇岛燕山大学**

**会议时间：8月25日—28日**

# 任务描述

信息抽取旨在将非结构化文本中的信息进行结构化，是自然语言处理的基础技术和重要研究领域。传统的信息抽取任务与评测通常针对特定的文本领域和单一的抽取任务，难以评估相关技术与方法在通用场景和任务下的抽取性能。

为此，中国科学院软件研究所、百度公司与千言开源数据集项目联合发起业界首个通用信息抽取评测——“CCKS-千言通用信息抽取竞赛”。本赛事是千言多形态信息抽取的全新升级，不局限于传统的单任务信息抽取的评测范式，而是将多种不同的信息抽取任务用统一的通用框架进行描述，着重考察相关技术方法在面对新的、未知的信息抽取任务与范式时的适应与迁移能力，从而满足当下信息抽取领域快速迭代、快速迁移的实际需求，更贴近实际业务应用。

评测地址：https://aistudio.baidu.com/aistudio/competition/detail/161/0/introduction

## 任务形式：

信息抽取任务旨在根据特定的抽取需求从非结构化文本中自动抽取结构化信息。其中，特定的抽取需求是指抽取任务中的抽取框架，主要由抽取类别（人物名称、公司名称、企业上市事件）及目标结构（实体、关系、事件等）组成。 本任务为中文信息抽取任务，即按照特定的抽取框架S，从给定的一组自由文本X中抽取出所有符合抽取需求的信息结构Y（实体、关系、事件记录等）。对于同一输入文本，不同的抽取框架会抽取不同的信息结构。示例如下：

============================================

**抽取框架1：金融事件抽取**

**输入文本（Text）：**

宁波容百新能源科技股份有限公司(简称“容百科技”，证券代码：688005)在科创板上市。

**抽取需求（Schema）：**

|  |  |
| --- | --- |
| 事件类型 | 上市 |
| 事件论元1 | 上市时间 |
| 事件论元2 | 上市板块 |
| 事件论元3 | 上市企业 |
| 事件论元4 | 融资金额 |

**事件定义：**

企业通过证券交易所首次公开向投资者增发股票，以期募集用于企业发展资金的过程。文本中不存在的槽位可空缺。

<上市企业> 于 <上市时间> 在 <上市板块> 上市，一共融资 <融资金额>。

**论元定义：**

* **上市企业：**是指所发行的股票经过国务院或者国务院授权的证券管理部门批准在证券交易所上市交易的股份有限公司。
* **上市时间：**是指证券管理部门在证券交易所上市交易的时间
* **上市板块：**是指主板、中小板、创业板、其他。
* **融资金额：**是指上市企业通过“上市”这一行为融到的总资本。

--------------------------------------------

**示例输出1**

**信息结构（Structure）:**

|  |  |
| --- | --- |
| 事件类型 | 上市 |
| 上市板块 | 科创板 |
| 上市企业 | 宁波容百新能源科技股份有限公司 |

============================================

**示例输入2 冬奥事件抽取**

**输入文本（Text）：**

2月8日上午北京冬奥会自由式滑雪女子大跳台决赛中中国选手谷爱凌以188.25分获得金牌！

**抽取需求（Schema）：**

|  |  |
| --- | --- |
| **事件类型** | **夺冠** |
| 事件论元1 | 夺冠时间 |
| 事件论元2 | 冠军 |
| 事件论元3 | 夺冠赛事 |

**事件定义：**

参赛选手在体育竞赛中夺得第一名。

<冠军> 于 <夺冠时间> 在 <夺冠赛事>获得冠军。

**论元定义：**

* **冠军：**是指在竞赛中获取第一名的参赛选手。
* **夺冠时间：**是指参赛选手获得冠军的时间。
* **夺冠赛事：**是指参赛选手获得冠军的赛事。

--------------------------------------------

**示例输出2**

**信息结构（Structure）:**

|  |  |
| --- | --- |
| **事件类型** | **夺冠** |
| 夺冠时间 | 2月8日上午 |
| 冠军 | 谷爱凌 |
| 夺冠赛事 | 北京冬奥会自由式滑雪女子大跳台决赛 |

**注：**本例与实例3虽然是相同的文本输入，但是由于抽取需求不同，抽取的目标结构也不同。

============================================

**示例输入3 人物信息**

**输入文本（Text）：**

2月8日上午北京冬奥会自由式滑雪女子大跳台决赛中中国选手谷爱凌以188.25分获得金牌！

**抽取需求（Schema）：**

|  |  |
| --- | --- |
| **关系类型** | **国籍** |
| 关系论元1 | 主体 |
| 关系论元2 | 客体 |

|  |  |
| --- | --- |
| **关系类型** | **地理位于** |
| 关系论元1 | 主体 |
| 关系论元2 | 客体 |

--------------------------------------------

**示例输出3**

**信息结构（Structure）:**

|  |  |
| --- | --- |
| **关系类型** | **国籍** |
| 主体 | 谷爱凌 |
| 客体 | 中国 |

**注：**本例中包含两个抽取需求，但是文本中只包含“国籍”关系描述，故只抽取出包含“国籍”关系的信息结果。

============================================

**示例输入4 对话情感抽取**

**输入文本（Text）：**

A:我昨晚又看了肖申克的救赎，简直百看不厌

B:是的，它是几代人的经典。

A:不过，这部剧的主演，我觉得蒂姆罗宾斯的演技不算那么出众。

**抽取需求（Schema）：**

|  |  |
| --- | --- |
| **情感类型** | **正向情感** |
| 情感论元1 | 目标对象 |
| 情感论元2 | 情感表达 |

|  |  |
| --- | --- |
| **情感类型** | **负向情感** |
| 情感论元1 | 目标对象 |
| 情感论元2 | 情感表达 |

--------------------------------------------

**示例输出4**

**信息结构（Structure）:**

|  |  |
| --- | --- |
| **情感类型** | **正向情感** |
| **目标对象** | 肖申克的救赎 |
| **情感表达** | 百看不厌 |

|  |  |
| --- | --- |
| **情感类型** | **正向情感** |
| **目标对象** | 肖申克的救赎 |
| **情感表达** | 几代人的经典 |

|  |  |
| --- | --- |
| **情感类型** | **负向情感** |
| **目标对象** | 蒂姆罗宾斯 |
| **情感表达** | 不算那么出众 |

# 评测细则

## 数据描述

本评测的数据及抽取框架主要来自于千言数据平台和百度通用信息抽取的应用案例。本次评测构建了多领域多场景下的多种抽取框架，包含医疗、法律、金融等领域和实体抽取、关系抽取、事件抽取等多种抽取任务，以期评测现有技术对通用领域下的信息抽取能力以及对新任务、场景的迁移能力。

参赛者可以通过已有模型以及千言平台、AI Studio平台上公开可获得的数据进行快速数据构建和现有模型迁移。

数据组成主要包含两个部分：

* **6个Seen Schema（已知框架）:** 主要来自千言平台与AI Studio平台上可获取的数据，参赛者可根据平台数据构建模型，该赛道主要评测现有技术基于标记数据构建模型的能力。
* **4个Unseen Schema（未知框架）：**主要来自百度信息抽取的实际案例，评测方仅提供少量的验证数据，用与参赛者进行抽取需求确认和模型验证，该赛道主要评测现有技术面向新的抽取需求的迁移能力。

本次评测数据分为三次发布：

* **4月25日：Seen Schema定义文件、验证数据及A榜测试集数据。** 该部分数据主要来源于千言数据集平台中的各类数据。每一个Schema 包含结构和类型定义，并提供少量验证数据。验证数据用于帮助参赛选手确认标注规范（例如标注边界等）。
* **6月20日：Unseen Schema定义和少量对应的验证数据。**每一个Schema 包含结构和类型定义，并提供少量验证数据。验证数据用于帮助参赛选手确认标注规范（例如标注边界等）。
* **7月25日：B榜测试集数据(最终测试集)。** 冲刺阶段参赛者需要对纯文本数据及对应的抽取需求（同时包含seen和unseen）进行信息抽取，最后提交抽取结果。

## 评价指标

**基本概念：**

* **文本块抽取结果 (span)**：文本字符串
* **实体抽取结果为二元组**：(entity\_type, span)
* **关系抽取及情感抽取结果为三元组**：(relation\_type, span1, span2)
* **事件触发词抽取为二元组**：(event\_type, span)
* **事件论元抽取为三元组**：(event\_type, arg\_type, span)

**评价指标：使用 Span 的字符串匹配进行评价。**详细指标如下：

* **实体抽取（Entity）：**输出结果和人工标注结果中的entity\_type和span字符串都相同时，TP+=1；
* **关系抽取（Relation）**：输出结果和人工标注结果中的Relation Type、Arg1 Span字符串和Arg2 Span字符串都相同时，TP+=1；
* **事件触发词抽取（Event Trigger）**：输出结果和人工标注结果中的Event Type和Trigger Span字符串都相同时，TP+=1。评分仅作参考，不计入最后评分；
* **事件论元抽取（Event Argument）**：输出结果和人工标注结果中的Event Type、Arg Type和Arg Span字符串都相同时，TP+=1；
* **情感抽取（Sentiment）**：输出结果和人工标注结果中的情感类型、目标对象Span字符串和情感表达Span字符串都相同时，TP+=1 ；

P = #TP / #Pred

R = #TP / #Gold

F = 2 \* P \* R / (P + R)

**总体打分：**

* **单任务F1：**所有实例的抽取结果计算微平均Micro-F1。
* **总体F1：**将四个抽取任务\*\*（实体抽取、关系抽取、事件论元抽取、情感抽取）\*\* 的Micro-F1取平均。

赛事基于总体F1进行榜单排名及评奖。

# 时间安排

|  |  |
| --- | --- |
| **赛程** | **时间** |
| 评测任务发布 | 2022/4/11 |
| 报名时间 | 2022/4/11 - 2022/7/25 23:59:59 |
| 验证数据发布 | 2022/4/25 |
| 测试数据发布 | 2022/7/25 |
| 测试提交截止 | 2022/7/31 23:59:59 |
| 评测论文提交 | 2022/8/12 23:59:59 |
| CCKS会议 (评测报告及颁奖) | 2022/8/25-2022/8/28 |

# 任务提交

## 所有队伍任务提交

本次任务将采取排行榜的方式，验证集发布后，允许参赛队伍每天至多向AI Studio平台（https://aistudio.baidu.com） 提交三次结果，排名每十分钟更新一次。

参赛队伍可在测试集B榜（最终测试集）发布之前，随时上传测试集A榜的计算结果，管理系统会及时更新各队伍的最新排名情况，测试集A榜不计入比赛最终成绩，我们会准备部分百度周边礼品奖励周冠军和周最佳进步团队。

**创新奖：**所有队伍都可参加创新奖评选，提交内容见获奖队伍提交内容。 主办方将在Unseen和Seen赛道中各选出一支队伍颁发创新奖。

其中，Unseen赛道的创新奖获奖团队将被进一步推荐参评CCKS评测十佳创新奖，并有机会获得5000元额外奖金。

## 获奖队伍材料提交（供资格审查，符合要求后颁发奖项和奖金）

**提交材料：** 所有获奖团队需额外提交相关材料以供资格审查，要求如下：

1. 测试集结果文件，用result.txt命名（UTF-8格式），格式同线上提交形式；
2. 相关代码及说明；
3. 论文或方法描述文档，优秀评测论文将被推荐至CCKS主会及对应期刊发表。

**提交方式：** 将以上三个文件在任务提交截止日期前发送至邮箱luge\_ai@126.com。

邮件的标题为：“CCKS2022-参赛队名称”，例如“CCKS2022-通用信息抽取队”。
代码及其文档需打包成一个文件（tar，zip，gzip，rar等均可），作为邮件附件传送，要求提交所有的程序代码及相关的配置说明，程序应当可以运行且所得结果与result.txt相符。

如果方法使用了额外公开的数据资源，要求说明并提供资源文件或地址。

# 赛事交流



# 评测规则

1. 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；
2. 参赛选手报名后可在“我的团队”页面组队。每支队伍需指定一名队长，队伍成员总数最多不超过5人；报名截止日期之后不允许更改队伍成员名单，请谨慎选择队员组队；
3. 队伍名的设定不得违反中国法律法规或社会公序良俗，且参赛队伍命名中不可出现“百度官方”，“飞桨官方”，“paddle官方”，“官方baseline”等字样。若命名违规的队伍在收到比赛主办方警告后仍未修改队伍名称，比赛主办方有权解散该队伍 ；
4. 每名参赛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；
5. 参赛队伍可在参赛期间随时上传测试集的预测结果，每天最多评测3次，比赛管理系统会实时更新各队伍的当前最高成绩与当前最新榜单排名情况；

# 评测奖励

本评测总奖励7万元，奖金设置如下：

* 一等奖30000元
* 二等奖20000元
* 三等奖10000元
* 创新奖5000元，共设立两个。