测评题目：面向“基因-疾病”的关联语义挖掘任务（Text mining task for "Gene-Disease" association semantics，GDAS track，CHIP 2022）

1、任务简介/任务详情

- **任务简介：**

在海量科学文献中，“基因-疾病”的关联机理通过突变、基因等系列分子对象及其触发词获得描述，自然语言处理为自动挖掘这一隐性知识条目提供了可能，亦为健康医学信息的自动化处理提供解决方案。本任务包括三个子任务：1、触发词实体识别；2、语义角色标注，3、“基因，调控类型，疾病”三元组抽取。所有数据取自 AGAC 语料库。

- **任务详情：**

➢ 任务一：触发词实体识别

任务一是传统意义下的命名实体识别（Named Entity Recognition），用以识别十二类与“基因-疾病”有关的分子对象及其触发词实体，包括 Var, MPA, Interaction, Pathway, CPA, Reg, PosReg, NegReg, Disease, Gene, Protein 和 Enzyme。

➢ 任务二：语义角色标注

任务二是一个语义角色标注任务（Semantic Role Labeling，or Shallow Semantic Parsing），语义角色包括 ThemeOf 和 CauseOf。该任务捕捉实体之间的语义依赖关系，用以构建“基因-疾病”关联。

➢ 任务三：“基因，调控类型，疾病”三元组抽取

任务三是一个三元组抽取任务（Triplet Extraction），该任务针对“基因-疾病”的关联机理的调控类型进行相关语义的抽取，可利用任务一和任务二所获得的触发词及其语义角色，挖掘其背后的深层语义。此处，调控类型包含 4 种对突变基因的语义描述：功能丧失(LOF)、功能获得(GOF)、功能调节(REG)、和功能的复合变化(COM)。该任务提供了 250 篇训练文本的“基因，调控类型，疾病”三元组结果。

➢ 子任务之间的逻辑关系

参与者可以选择参加任何一个子任务，但是任务一是基于任务二的，任务三可以独立执行或者基于任务一或者任务二的结果。

2、数据说明

- **数据样例：**

任务一和任务二数据格式为 JSON，包括以下内容：

{**"target"**: "http://pubannotation.org/docs/sourcedb/PubMed/sourceid/25805808", **"sourcedb"**: "PubMed", **"sourceid"**: "25805808", **"text"**: "Loss-of-function de novo mutations play an important role in severe human neural tube defects.\nBACKGROUND: Neural tube defects (NTDs) are very common and severe birth defects that are caused by failure of neural tube closure and that have a complex aetiology. Anencephaly and spina bifida are severe NTDs that affect reproductive fitness and suggest a role for de novo mutations (DNMs) in their aetiology.\nMETHODS: We used whole-exome sequencing in 43 sporadic cases affected with myelomeningocele or anencephaly and their unaffected parents to identify DNMs in their exomes.\nRESULTS: We identified 42 coding DNMs in 25 cases, of which 6 were loss of function (LoF) showing a higher rate of LoF DNM in our cohort compared with control cohorts. Notably, we identified two protein-truncating DNMs in two independent cases in SHROOM3, previously associated with NTDs only in animal models. We have demonstrated a significant enrichment of LoF DNMs in this gene in NTDs compared with the gene specific DNM rate and to the DNM rate estimated from control cohorts. We also identified one nonsense DNM in PAX3 and two potentially causative missense DNMs in GRHL3 and PTPRS.\nCONCLUSIONS: Our study demonstrates an important role of LoF DNMs in the development of NTDs and strongly implicates SHROOM3 in its aetiology.", "project": "AGAC2_PubMed_2",**"denotations"**: [{ "id": "T8", "span": { "begin": 771, "end": 778 }, "obj": "Protein" }, { "id": "T7", "span": { "begin": 779, "end": 789 }, "obj": "NegReg" }, { "id": "T6", "span": { "begin": 790, "end": 794 }, "obj": "Var" }, { "id": "T9", "span": { "begin": 823, "end": 830 }, "obj": "Gene" }, { "id": "T10", "span": { "begin": 936, "end": 939 }, "obj": "NegReg" }, { "id": "T11", "span": { "begin": 940, "end": 944 }, "obj": "Var" }, { "id": "T12", "span": { "begin": 961, "end": 965 }, "obj": "Disease" }, { "id": "T3", "span": { "begin": 1224, "end": 1227 }, "obj": "NegReg" }, { "id": "T1", "span": { "begin": 1228, "end": 1232 }, "obj": "Var" }, { "id": "T2", "span": { "begin": 1255, "end": 1259 }, "obj": "Disease" }, { "id": "T5", "span": { "begin": 1284, "end": 1291 }, "obj": "Gene" }], **"relations"**: [{ "id": "R1", "pred": "CauseOf", "subj": "T1", "obj": "T3" }, { "id": "R10", "pred": "ThemeOf", "subj": "T12", "obj": "T10" }, { "id": "R11", "pred": "ThemeOf", "subj": "T5", "obj": "T1" }, { "id": "R2", "pred": "ThemeOf", "subj": "T2", "obj": "T3" }, { "id": "R5", "pred": "CauseOf", "subj": "T6", "obj": "T7" }, { "id": "R6", "pred": "ThemeOf", "subj": "T8", "obj": "T7" }, { "id": "R7", "pred": "ThemeOf", "subj": "T9", "obj": "T6" }, { "id": "R8", "pred": "ThemeOf", "subj": "T9", "obj": "T11" }, { "id": "R9", "pred":

"CauseOf", "subj": "T11", "obj": "T10" }]}

- 数据格式说明:

数据格式为 JSON，包括以下内容:

“target”：注释文本的地址

“sourcedb”：文本来源，AGAC 中的所有文本都来自 PubMed

“sourceid”：文本的 PMID

“text”：文本原始摘要

“denotations”：对应任务一的触发词注释，包括“id”；“span”：实体在文本中的位置信息；“obj”：实体被标注的标签

“relations”：对应任务二触发词之间的语义角色，包括“id”；“pred”：语义角色；“subj”和“obj”：任务一中触发词的“id”，关联方向从“subj”到“obj”

任务三“基因，调控类型，疾病”三元组数据说明：

PMID	Gene	Regulation Type	Disease
19338054	MC1R	LOF	melanoma
18594199			
20399956	Shp2	GOF	juvenile myelomonocytic leukemia
	Shp2	LOF	LEOPARD syndrome

- 数据格式说明:

任务三数据包含四列数据，分别是 PMID，Gene，Regulation Type 和 Disease，每一篇文本中可能不会抽取到三元组关系，也可能会抽取到多条三元组关系。

- **训练及测试数据规模**

- 任务一:

训练数据：250 篇 PubMed 文献

测试数据：2000 篇 PubMed 文献

- 任务二:

训练数据：250 篇 PubMed 文献

测试数据：2000 篇 PubMed 文献

- 任务三：

训练数据：250 篇 PubMed 文献及其抽取出的三元组关系

测试数据：2000 篇 PubMed 文献

- **任务对象**

任务对象定义来源于 AGAC（Active Gene Annotation Corpus）活跃基因注释语料库，该语料库主要用以挖掘突变引起的“基因-疾病”关联机理。AGAC 语料库包括四类分子对象、八类触发词实体，两个语义角色，以及四种用以描述“基因-疾病”关联机理的功能变化。说明：AGAC 中只有同时涉及特定突变和生物学功能或疾病的句子才会被注释。

任务一识别对象：

- 分子实体
  - **Disease(疾病)**
  - **Gene(基因)**
  - **Protein(蛋白)**
  - **Enzyme(酶)**

- 触发词实体
  - **Variation(Var，突变)**：包括 DNA、RNA、蛋白质的突变和分子结构的变化，标注实体如：“*mutations on the Arg248 and Arg282*”，“*mutant R282W*”，“*missense mutations*”
  - **Molecular Physiological Activity (MPA， 分子活性)**：分子水平的活性包括分子活性、基因表达和分子生理活性，标注实体如：“*phosphorylation*”，“*transcription*”，“*histone methylation*”，“*bioactivation of cyclophosphamide*”
  - **Interaction(互作)**：分子和分子或者分子和细胞之间的联系，标注实体如：“*bind*”，“*interaction*”

- ✧ **Pathway(通路)**：包括各种通路，如“*Bmp pathway*”，“*PI3K pathway*”
- ✧ **Cell Physiological Activity (CPA，细胞活性)**：在细胞水平或以上的活动，包括细胞反应性和细胞或器官的发育和生长，标注实体如：“*T helper cell responses*”，“*renal development*”
- ✧ **Regulation (Reg，调控)**：中性的提示词或词组，意思是没有损失或收获，如：“***resulted in***”，“*regulated*”
- ✧ **Positive Regulation (PosReg，正调控)**：表示获得功能的线索词或短语，如：“*facilitates*”，“*enhanced*”，“*increased*”
- ✧ **Negative Regulation (NegReg，负调控)**：表示失去功能的线索词或短语，如：“*suppressed*”，“*decreased*”，“*inhibited*”

任务二标注对象：

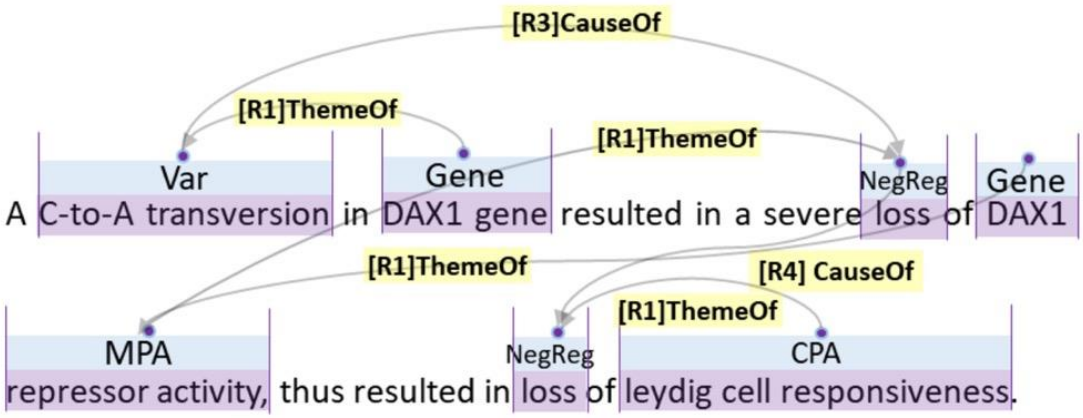
- ➢ 语义角色
  - ✧ **ThemeOf**：从主事（Theme）实体指向当前实体
  - ✧ **CauseOf**：从当前实体指向致事（Cause）实体

任务三抽取对象：

- ➢ 调控类型

LOF 和 **GOF** 表示功能的丧失或获得，**REG** 表示一般性调控关系，**COM** 表示基因与疾病之间的功能改变更为复杂，很难判断是 LOF 还是 GOF。在预测时从文本中提取“gene; regulation type; disease”三元组，如：“SHROOM3; LOF; Neural tube defects”，或者添加 pmid 信息：“25805808; SHROOM3; LOF; Neural tube defects”

- **任务样例**



例句“**A C-to-A transversion in DAX1 gene resulted in a severe loss of DAX1 repressor activity, thus resulted in loss of leydig cell responsiveness**”中：

任务一：实体标注一共有七处

- “C-to-A transversion”标注为“Var”；
- “DAX1 gene”标注为“Gene”；
- “loss”标注为“NegReg”；
- “DAX1”标注为“Gene”；
- “repressor activity”标注为 MPA；
- “loss”标注为“NegReg”；
- “leydig cell responsiveness”标注为 CPA；

任务二：语义角色的识别有六个

- 由“DAX1 gene”指向“C-to-A transversion”的语义角色关系是“ThemeOf”，因为发生该突变的主体是“DAX1 gene”
- 由“C-to-A transversion”指向第一个“loss”的语义角色关系是“CauseOf”，因为突变发生后导致了 DAX1 基因抑制活动的丧失
- 由“DAX1”指向“repressor activity”的语义角色关系是“ThemeOf”，因为发生该抑制活性的主体是“DAX1”基因
- 由“repressor activity”指向第一个“loss”的语义角色关系是“ThemeOf”，因为 DAX1 基因抑制活动的丧失事件的主体是抑制活动“repressor activity”
- 由“leydig cell responsiveness”指向第二个“loss”的语义角色关系是“ThemeOf”，因为间质细胞反应性的丧失事件的主体是“leydig cell responsiveness”

- ✧ 由第一个“loss”指向第二个“loss”的语义角色关系是“CauseOf”，因为 DAX1 基因抑制活动的丧失导致了间质细胞反应性的丧失

任务三：根据给定文本抽取出三元组

例句 1：**Mutations in SHP-2 phosphatase that cause hyperactivation of its catalytic activity have been identified in human leukemias, particularly juvenile myelomonocytic leukemia.**

从生物学的观点来看，催化活性的过度激活（hyperactivation of its catalytic activity）显然是功能获得的一种描述。因此，这句话承载着明确的语义信息，即基因“SHP-2”在突变后发挥着与“少年粒细胞白血病”（ juvenile myelomonocytic leukemia）相关的“GOF”功能。因此任务三从这个句子中提取出的三元组就是“*SHP-2; GOF; juvenile myelomonocytic leukemia*”

例句 2：**Lynch syndrome (LS) caused by mutations in DNA mismatch repair genes MLH1.**

这句话描述了疾病“Lynch syndrome”与基因“MLH1”之间的关联，但短语“caused by”意味着没有损失或获得，因此这句话中的三元组应该是“*MLH1; REG; Lynch syndrome*”。

例句 3：**Here, we describe a fourth case of a human with a de novo KCNJ6 (GIRK2) mutation, who presented with clinical findings of severe hyperkinetic movement disorder and developmental delay. Heterologous expression of the mutant GIRK2 channel alone produced an aberrant basal inward current that lacked G protein activation, lost K+ selectivity and gained Ca2+ permeability.**

“失去 K+选择性，获得 Ca2+渗透性”（ lost K+ selectivity and gained Ca2+ permeability）的描述同时显示了 LOF 和 GOF，因此功能变化不能标记为 LOF 或 GOF，而是标记为 COM，因此这句话中的三元组应该是“*GIRK2; COM; hyperkinetic movement disorder*”

3、评估方法

- **评估指标：**

  Precision，recall，F-score

- **结果提交邮箱：**

  ouyangsizhuo@foxmail.com

- **结果提交格式：**

任务一：参赛者请提交包含触发词实体识别结果的 JSON 文件，格式与数据样例保持一致

任务二：参赛者请提交包含语义角色标注结果的 JSON 文件，格式与数据样例保持一致

任务三：参赛者请提交包含“基因，调控类型，疾病”三元组抽取结果的 TXT 文件，每行报告一个三元组结果，格式如：“25805808; SHROOM3; LOF; Neural tube defects”，其中“25805808”为该三元组来源文献的 PMID

4、报名方法以及测评任务沟通途径（电子邮件、微信群或其他）

报名方法：以 CHIP 评测主办方官方报名渠道为准
联系邮箱：ouyangsizhuo@foxmail.com

5、参赛规则

注意，以下通用规则适用于本评测任务。在此基础上，参赛选手还需遵循具体大赛的特有规则。如有冲突，以后者为准。

1. 参赛选手需要提交“参赛队名，队长信息（姓名，邮箱，联系电话），参赛单位名称”等信息，报名方式见下文。
2. 报名截止到测试数据集发布，在测试数据集发布之后，未报名的选手/队伍不能再报名或提交。
3. 每支队伍需指定一名队长，队伍名称不超过 15 个字符，队伍成员不超过 4 人。
4. 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，将取消所有相关队伍的参赛资格。
5. 允许使用公开和选手个人/组织内部的代码、工具、数据，但需要保证参赛结果可以复现。
6. 针对测试集，选手不允许执行任何人工标注。
7. 参赛选手最终需要提交可运行的代码和方法描述文档，若在排行榜上的结果无法复

现，将取消参赛资格。

8. 欢迎国内外在校生及社会在职人士参加。比赛组织方成员不可参赛。

6、任务组织者

评测任务组织者：

夏静波： xiajingbo.math@gmail.com

评测任务联系人：

欧阳思卓： ouyangsizhuo@foxmail.com

评测任务网页链接：

http://lit-evi.hzau.edu.cn/AGAC-CHIP2022/

7、测评时间安排

- **报名时间：2022 年 6 月 1 日-10 月 15 日**
- **训练及验证数据发布：2022 年 7 月 1 号（250 条）**
- **测试数据发布：2022 年 8 月 25 号（2000 条）**
- **提交测试结果：2022 年 8 月 26-28 号（每天可提交一次，以零点后第一次提交的结果为准，取三天内最高成绩，数据格式为数据样例一样的 JSON 格式）**
- **评测论文提交时间：2022 年 10 月（CHIP 会议前 1 个月）**
- **评测报告及颁奖：2022 年 11 月 24 日**
- **评测学术委员会评测论文审阅：2022 年 12 月**
- **评测论文修回：2022 年 12 月（2 周修改周期）**
- **评测论文集中投稿：2023 年 1 月-**