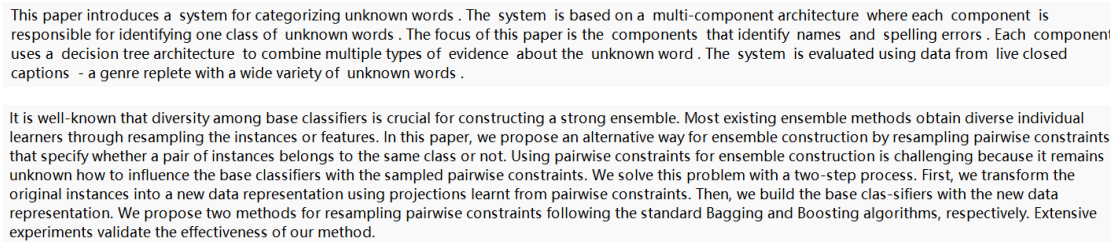
基于所构建的概念模型对语料进行关系抽取并构建知识图谱

张轩 董潇 林兴波

1、选用的数据集

SCIERC 数据集来源于 AI 领域的 12 个会议和研讨会的论文摘要，具体的数据实例是不同 ai 领域的科学文献的摘要信息，语言为英文，主要提供了关于文献的研究内容与对象、所使用的方法、得到的研究成果等方面的信息。



图表 1 数据实例

2、研究问题与概念模型

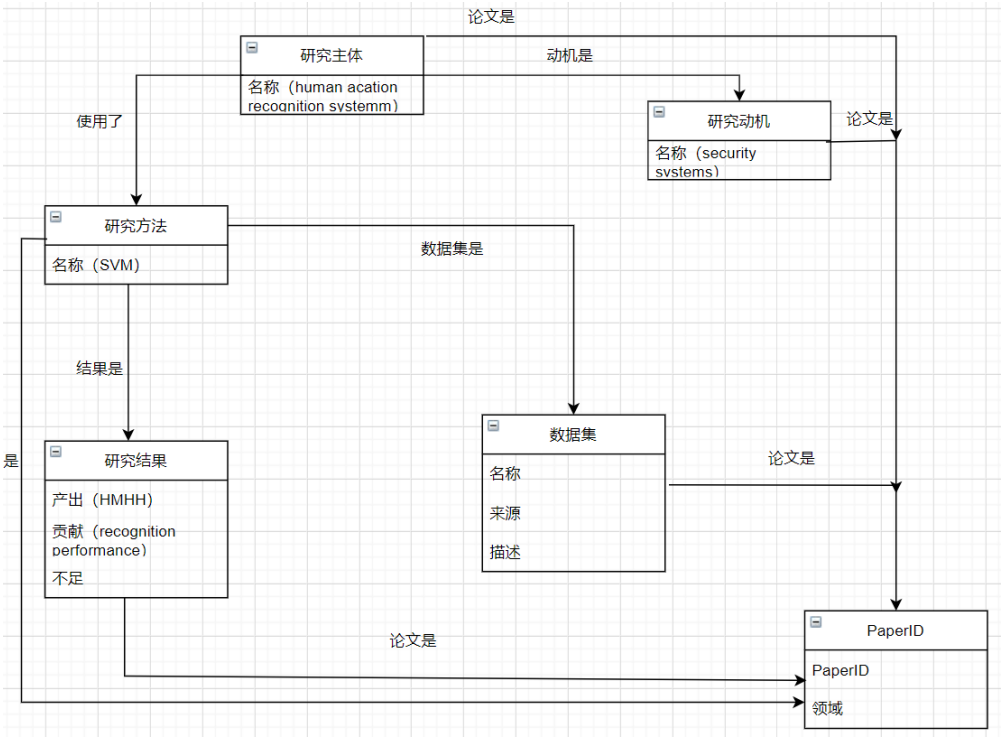
2.1、研究问题

研究问题是：在 AI 领域，如何使用论文摘要快速了解论文？

选择的研究问题具有一定的现实需求和研究意义：即随着数字化学术文献信息迅速增长，系统用户无法以很方便的方式对大量学术文献中的信息进行管理、利用。导致大量学术文献中的知识难以得到捕获、共享和使用。科技文献的摘要信息能够对一篇研究文献所涉及的内容对象、主要工作任务、研究方法、数据使用和主要研究成果进行选取与总结。因此通过对论文摘要信息总结可以快速了解一篇科技论文。

2.2、概念模型

我们根据作业二的研究问题构建如图所示的概念模型。

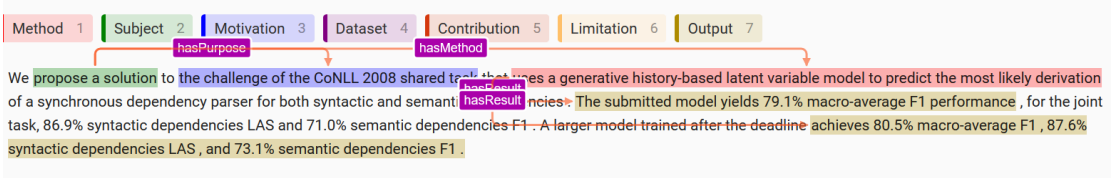


图表 2 基于研究问题的概念模型

3、数据标注过程说明

3.1、数据说明

本次标注数据为 SCIERC 数据库中的摘要数据，共计 500 条，在问题导向的概念建模统一标准下，我们对 7 种实体类型（Method、Subject、Motivation、Dataset、Contribution、Limitation、Output）和 4 种关系类型（hasMethod、hasPurpose、hasResult、hasDataset）进行标注。500 条数据的标注主要由两名同学完成，董潇标注 355 条数据，林兴波标注 145 条数据，平均一条数据的标注时间在 2 分钟左右。



图表 3 数据标注示例

3.2、标注过程说明

我们使用线上平台 label-studio 进行数据标注，最后标注好的数据导出为 JSON 文件。

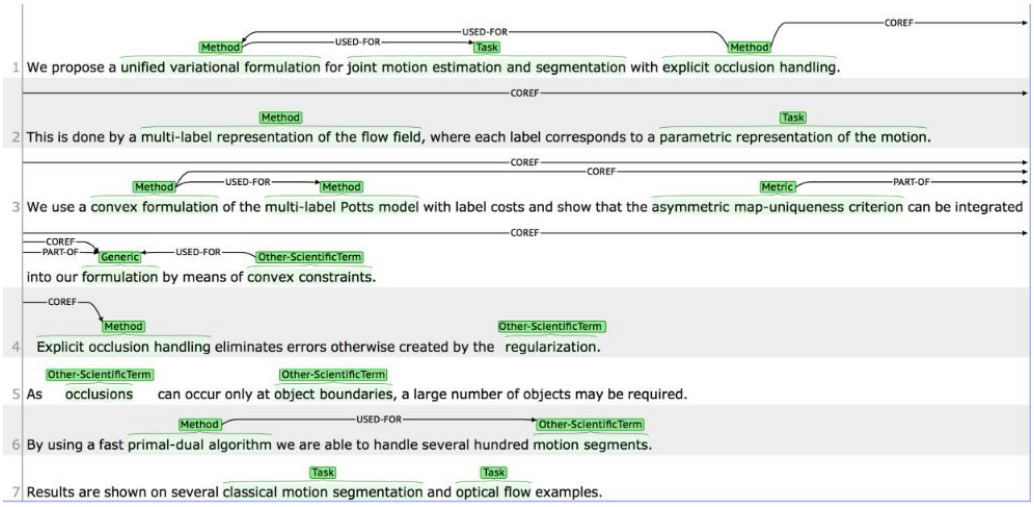
根据作业二的准备，我们确定了基本标注内容，即 7 种实体与 4 种关系。开始标注前，我们讨论确定相应标准与规则：（1）了解英文表述习惯，明确不同类型实体的具体案例，确定哪种类型的信息和语句适合被标注为何种实体；（2）尽可能标注较少的词句还原更全面、完整的信息，避免出现长句。

在标注的过程中，由于数据的特殊性以及标注者自身对文本的理解差异，我们也遭遇了“标注迷雾”，具体表现为：（1）数据对象是摘要文本，研究对象、研究方法、研究成果等信息主要以较复杂的句子的形式呈现，选取精准的单词短语进行实体描述比较困难。摘要信息对于研究对象、研究方法和成果的论述也会横跨多个短语或者句子，因此需要判断取舍。（2）AI 领域研究论文摘要是英文写作，涉及专业表达的专业性较强，对标注人员的理解能力要求高。出于英文本身的特点，存在代词指代判定、特殊语法（定语从句、不定式表达等）理解问题。同一实体（如方法）可能存在不同的表达，存在逻辑顺序，此时容易对标注造成困扰。

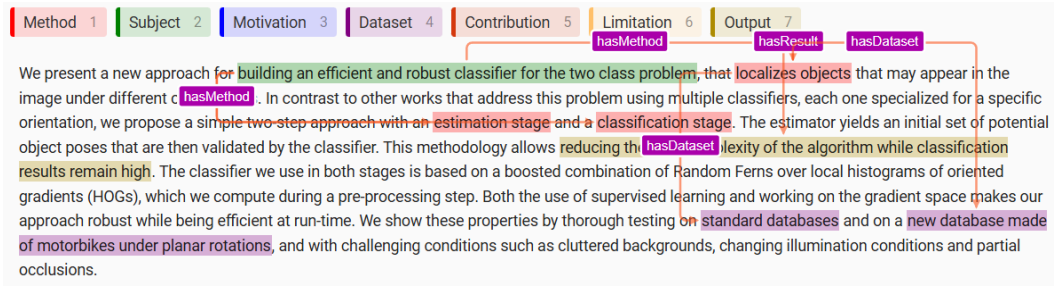
我们预标注一部分数据，发现存在概念不明、标注词句过长、标注质量较低等问题。与老师进行交流并在组内讨论，我们重新确定数据标注流程：（1）两位同学主要负责数据标注，另一位同学作为独立第三方按照一定比例随机抽取两位同学标注的数据，重新标注并与已有标注进行对比，检验对比质量；（2）严格按照实体类型的标准标注，宁缺毋滥，对于不确定的数据可以保留，组内讨论确定；（3）适当放宽对词句长度限制，精准体现数据类型的同时要具有数据差异。（4）标注一部分数据后，放入关系抽取模型检验标注质量。

3.3、与作业 1 数据的区别

作业 1 中的模型主要是运用原论文已经预先标注好的实体和关系数据。本组重新选定标注的实体和关系，因而实体与实体间的关系均有所改变，并无复用原标注数据。原标注数据是单句内的实体间的关系，而且大多数实体是较短的单词或短语。我们标注的实体数据大多较长，且标注的实体分属不同的句子，属于跨句关系，因此在训练处理时有所不同。



图表 4 原数据标注示例



图表 5 本组数据标注示例

4、模型说明

4.1、模型结构

选用模型来自普林斯顿大学的 PURE 模型[1]。该模型分为实体模型和关系模

[1] Zexuan Zhong and Danqi Chen. 2021. A Frustratingly Easy Approach for Entity and Relation Extraction. In Proceedings of the 2021 Conference of the North American Chapter of the

型，实体模型用于识别文本中的命名实体，并为每个候选跨度预测实体类型。实体模型使用预训练的语言模型（如 BERT）来获取每个 token 的上下文表示，对于每个可能的跨度，定义一个跨度表示，通常包括跨度的开始和结束 token 的表示以及一些学习到的跨度宽度特征，将跨度表示输入到一个前馈网络中，预测实体类型的概率分布。关系模型根据实体模型的输出，对每对实体进行关系分类。接收实体模型预测的实体对作为输入，在输入句子中插入标记 token 来区分实体对的主体和客体，并标明它们的类型。使用第二个预训练的语言模型来获取包含标记的新句子的表示，定义一个跨度对表示，通常包括主体和客体标记的开始位置的表示，将跨度对表示输入到前馈网络中，预测关系类型的概率分布。首先运行实体模型来识别文本中的所有实体，然后，对于实体模型识别的每一对实体，使用关系模型来预测它们之间的关系。使用标准的评估协议，如微平均 F1 分数，来评估实体识别和关系提取的性能。这个模型的创新之处在于其简单性以及在不牺牲准确性的情况下提高了推理速度。通过独立的实体和关系模型，以及在关系模型中早期融合实体信息，该方法在标准基准测试中取得了优异的性能。

在本次模型构建中，我们基于 PURE 模型，微调了实体类型和关系类型，并没有使用实体模型来识别文本中的所有实体，而是直接使用人工标注的实体进行关系预测。

4.2、训练框架和超参数

训练框架	Pytorch 1.4.0	Transformers 库 3.0.2	Python 版本 3.7
计算平台	Kaggle	1 张 GPU-P100	
超参数	学习率 2e-5	Batch size 32	10 个 epoch
训练时长	SciBert 训练 887s	BertBase 训练 882s	

图表 6 训练基本要素

本次作业中，我们使用 SciBert 与 BaseBert 两种预训练的语言模型进行关系模型的训练，我们没有进行实体抽取，将标注好的实体直接用于关系抽取。

Association for Computational Linguistics: Human Language Technologies, pages 50–61, Online.
Association for Computational Linguistics.

通过比较这两种模型的性能，我们选择其中较优的那一个作为最终模型。

较原标注数据基础上的模型训练而言，SciBert 的训练时间快了近 2200s，BaseBert 的训练时间快了近 2500s。

4.3、模型训练结果

将关系模型的抽取抽取结果看做一个多分类任务，以下是模型训练结果：

Class	Precision	Recall	F1-score
hasDataset	0.6667	0.3333	0.4444
hasMethod	0.9519	0.9706	0.9612
hasPurpose	0.8000	0.5333	0.6400
hasResult	0.8679	0.9787	0.9200

图表 7　SciBert 模型的各关系分类 P 值、R 值、F1 值

```
Macro-Precision: 0.8560
Macro-Recall: 0.8437
Macro-F1: 0.8496

Micro-Precision: 0.9644
Micro-Recall: 0.9644
Micro-F1: 0.9644
```

图表 8 SciBert 的 Macro 与 Micro 指标

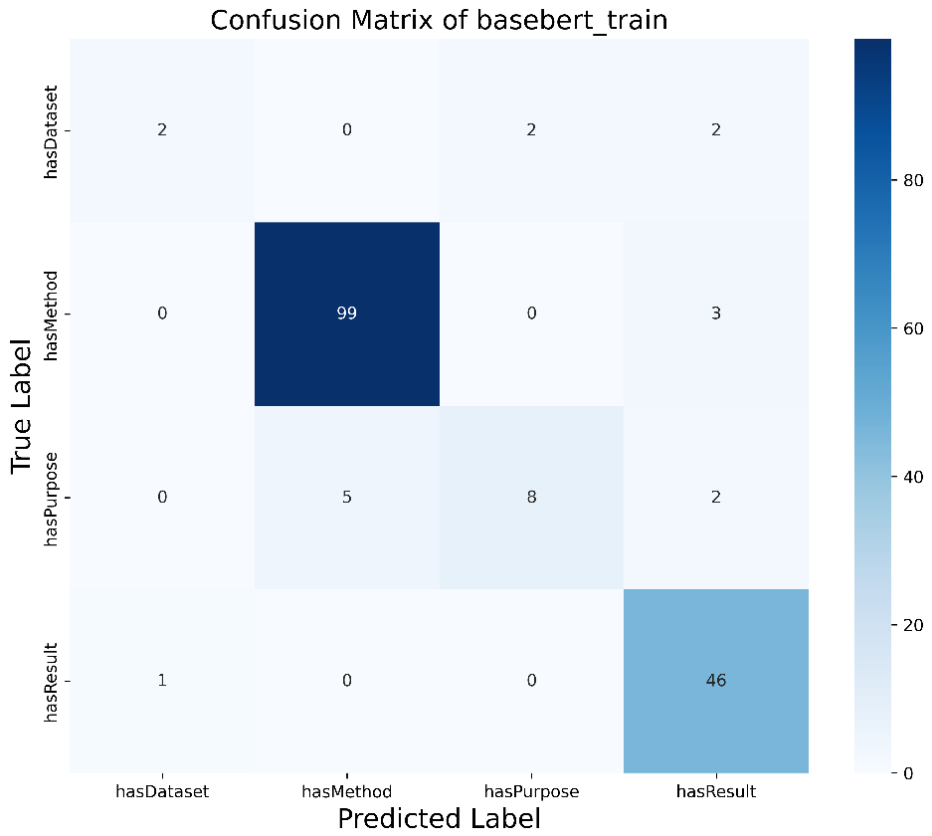
Class	Precision	Recall	F1-score
hasDataset	0.5714	0.5714	0.5714
hasMethod	0.9852	1.0000	0.9925
hasPurpose	0.8824	0.8333	0.8571
hasResult	0.9848	0.9701	0.9774

图表 9　BaseBert 模型的各关系分类 P 值、R 值、F1 值

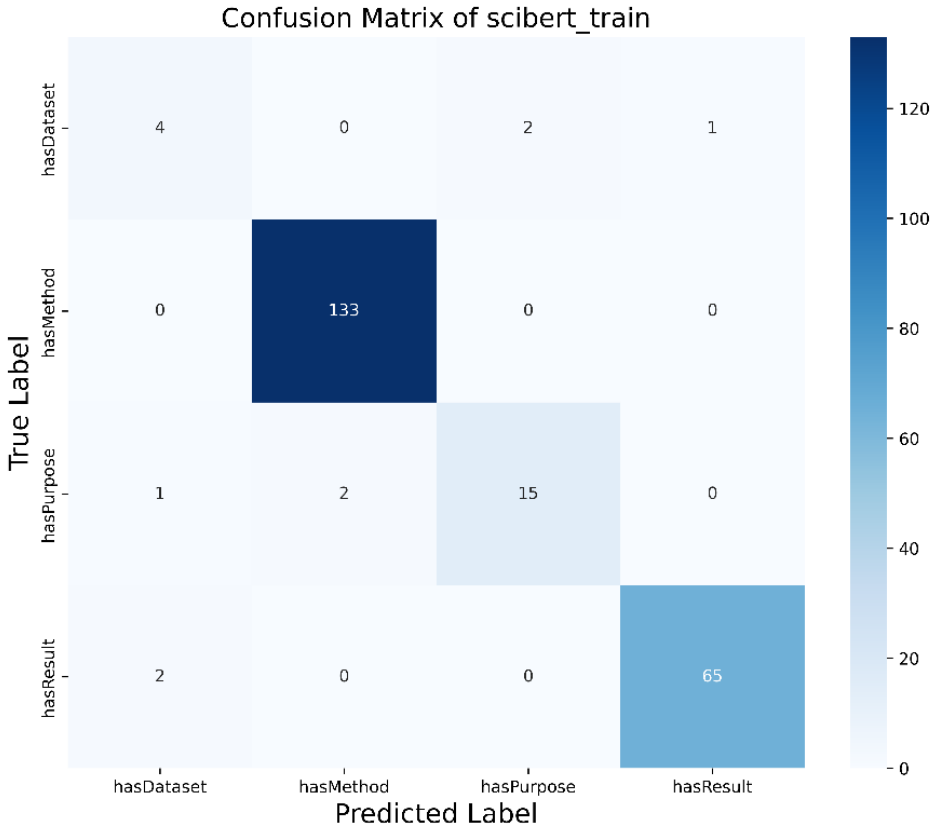
```
Macro-Precision: 0.8216
Macro-Recall: 0.7040
Macro-F1: 0.7414

Micro-Precision: 0.9118
Micro-Recall: 0.9118
Micro-F1: 0.9118
```

图表 10 BaseBert 模型的 Macro 与 Micro 指标



图表 11 SciBert 模型的混淆矩阵



图表 12 BaseBert 模型的混淆矩阵

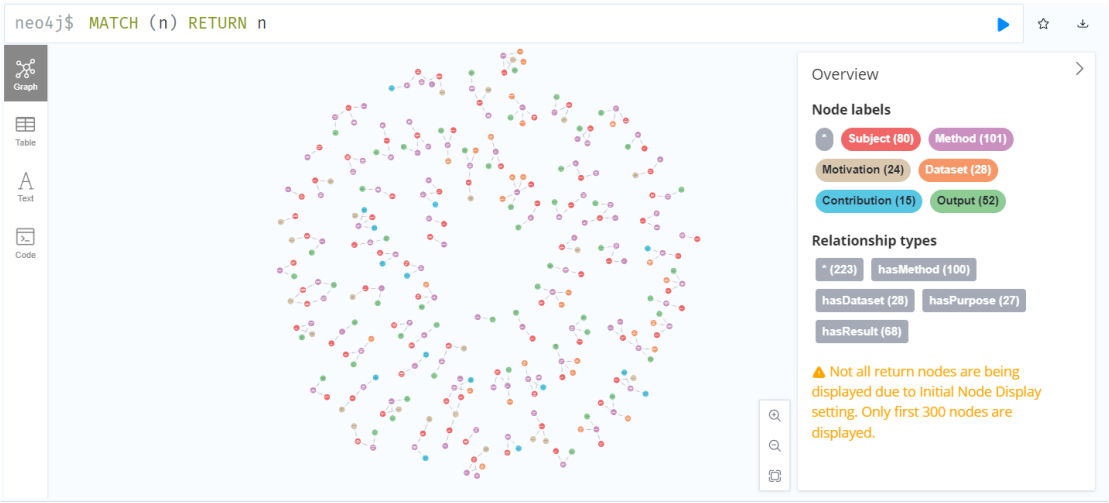
4.3、结果分析

通过一些指标对比可知：SciBert 模型的结果更好。SciBert 的 Macro-Precision 值、Macro-Recall 值和 Macro-F1 值均在 0.85 左右，Micro-Precision 值、Micro-Recall 值和 Micro-F1 值均在 0.96 以上，从指标上来看，模型训练结果比较好。相对而言，BaseBert 模型的 Macro-Precision 值、Macro-Recall 值和 Macro-F1 值分别为 0.82、0.7、0.74，Micro-Precision、Micro-Recall 和 Micro-F1 值均在 0.91 左右。

由于关系数据分布不均衡，hasDataset 这一关系比较少，所以其准确率与召回率均比较低。

5、知识图谱构建

使用 py2neo 库将本地数据批量存储到 neo4j 数据库上。节点（实体）有 1927 个，边（关系）有 1409 条。节点的类型有"Method"，"Subject"，"Motivation"，"Dataset"，"Contribution"，"Limitation"，"Output"，边的类型有"hasMethod"，"hasPurpose"，"hasDataset"，"hasResult"。节点属性值有 name 和 paper，name 代表类的一个实例，paper 代表所属论文的编号。



图表 13 知识图谱整体构建情况示意图

由图可见，构建的知识图谱以单篇论文摘要为单位呈现离散的状态，每个小的集群表示的则是一篇论文摘要的论文 id、研究对象、研究方法、数据集和研究结果。

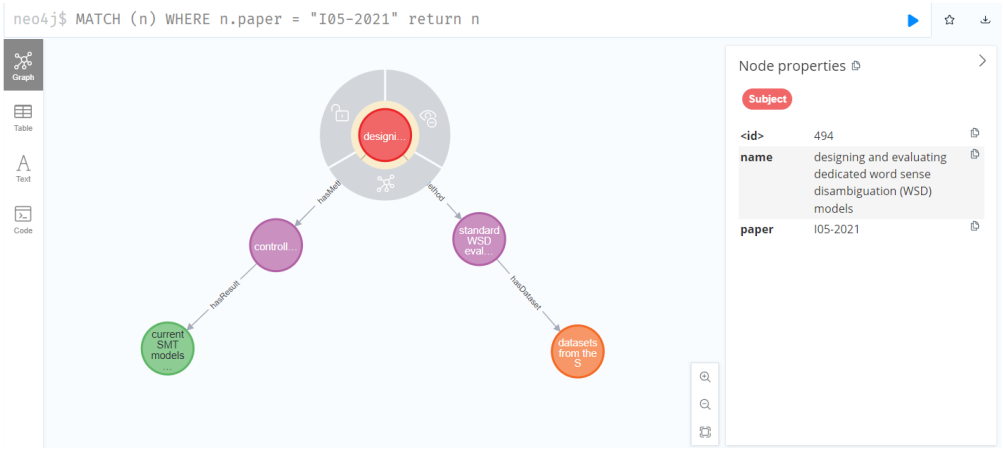
6、通过知识图谱回答研究问题

我们的研究问题是：在 AI 领域，如何使用论文摘要快速了解论文？

我们采用两个案例来说明现有知识图谱的一些功能和发现：

（1）AI 论文摘要一般包含什么内容？

Cypher 语言为：MATCH(n) WHERE n.paper ="I05-2021"return n

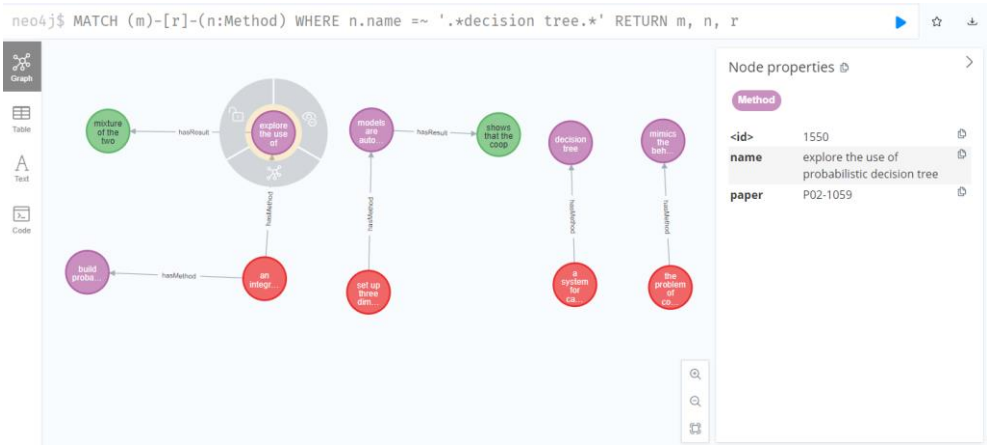


图表 14 基于论文摘要 id 的查询

通过查询论文 I05-2021，可以知道这篇论文的主要研究任务和研究对象是设计一个模型，通过控制和标准化 WSD 评估的研究方法并选用来自 S 的数据集，得到 SMT 模型建立的研究成果。这呈现了基于论文摘要的基本架构。

（2）使用了同一种方法的 AI 论文摘要有哪些，其研究对象是什么，有何结果？

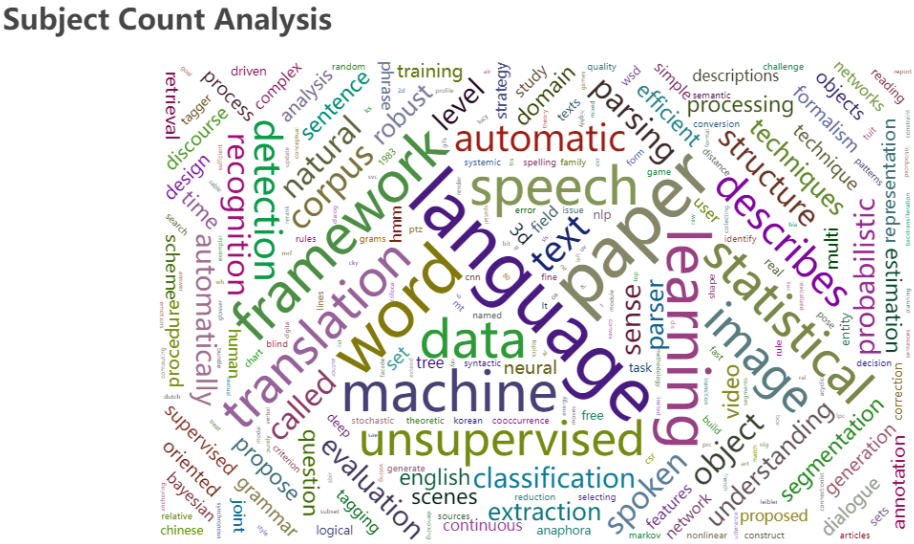
Cypher 语言为：MATCH (m)-[r]-(n:Method) WHERE n.name =~ '.*decision tree.*' RETURN m, n, r



图表 15 基于共有研究方法的查询

可以检索查看研究方法包含“decision tree”的论文及其研究对象，使用了决策树方法的论文研究对象包括自然语言生成概率模型、文本输入生成语音基频轮廓、人工创建摘要、自动识别谓词自变量结构和口语对话中的代词解析等。综合来看，说明了决策树的方法在自然语言研究领域的广泛应用。

我们对数据中的研究对象和研究方法进行可视化，绘制词云图，如下所示。可知论文的研究对象多集中在自然语言处理、图像自动化处理、文本探测与识别等领域。研究方法多集中在分类、迭代、特征提取等。但是由于数据标注质量以及标注数据较长，因此在制作词云图展示结果时效果一般。



图表 16 研究对象词云图

图表 17 研究方法词云图

7、知识图谱改进方向

7.1、成果

本次作业，我们基于所构建的概念模型对语料进行关系抽取并构建相应的知识图谱，并利用这一知识图谱进行探索性研究，基本完成了预定工作和目标。通过我们所建立的知识图谱可以比较直观地了解一篇 AI 领域的论文摘要基本构成和基本内容，了解内容之间的关联和应用方向，帮助快速获取论文的基本思想。同时，使用共现的方式可以查看同一技术手段在不同实验（文章）中的应用以及结果，这可以为技术的发展进步提供一些思考。

7.2、不足与改进

我们以文本摘要作为标注对象，研究方法、研究成果等实体的标注无可避免地需要标注较长的词或者句子，这影响数据标注的质量。而且在标注过程中，数据里的实体类别数量并不均衡，Dataset 较少。由于我们以一篇论文摘要为单独标注对象，关注其内部内容的关系，所以知识图谱呈现离散的“孤岛”，因此利用构建的知识图谱进行问题研究无法挖掘更深层次的价值和联系。

在现有问题的基础上，我们希望对作业有如下改进：（1）优化标注标准和流程，继续提高数据标注质量；（2）优化模型，优化模型超参数或者尝试其他模型；（3）在高质量的数据基础上，发现不同文章、不同节点之间更广泛的、更深层次的联系。（4）利用高质量的标注数据构建知识图谱，希望发现更普遍的联系和规律。

8、成员分工

张轩	讨论、模型训练、知识图谱构建、小组展示
董潇	讨论、数据标注、PPT 制作
林兴波	讨论、数据标注、撰写报告、小组展示