# 问题驱动的本体构建与知识图谱概念层设计

**张轩 董潇 林兴波**

## 小组分工：

|  |  |
| --- | --- |
| **张轩** | 讨论问题导向的本体构建、构建概念模型、使用protégé构建本体 |
| **董潇** | 讨论问题导向的本体构建、编写本体结构、使用protégé构建本体 |
| **林兴波** | 讨论问题导向的本体构建、构建概念模型、制作PPT |

## 数据来源

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，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。

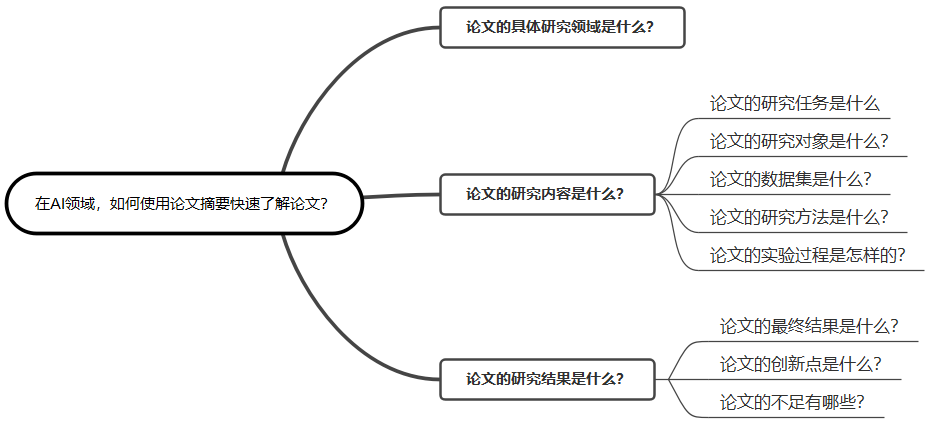
## 研究问题：

**在AI领域，如何使用论文摘要快速了解论文？**

## 现实需求与研究意义：

数字化的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知识进行管理、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[[1]](#footnote-1)。科技文献的摘要信息能够有效对一篇研究文献所涉及的内容对象、主要工作任务、研究方法与数据，和主要研究成果进行选取与总结[[2]](#footnote-2)。

## 问题拆解：



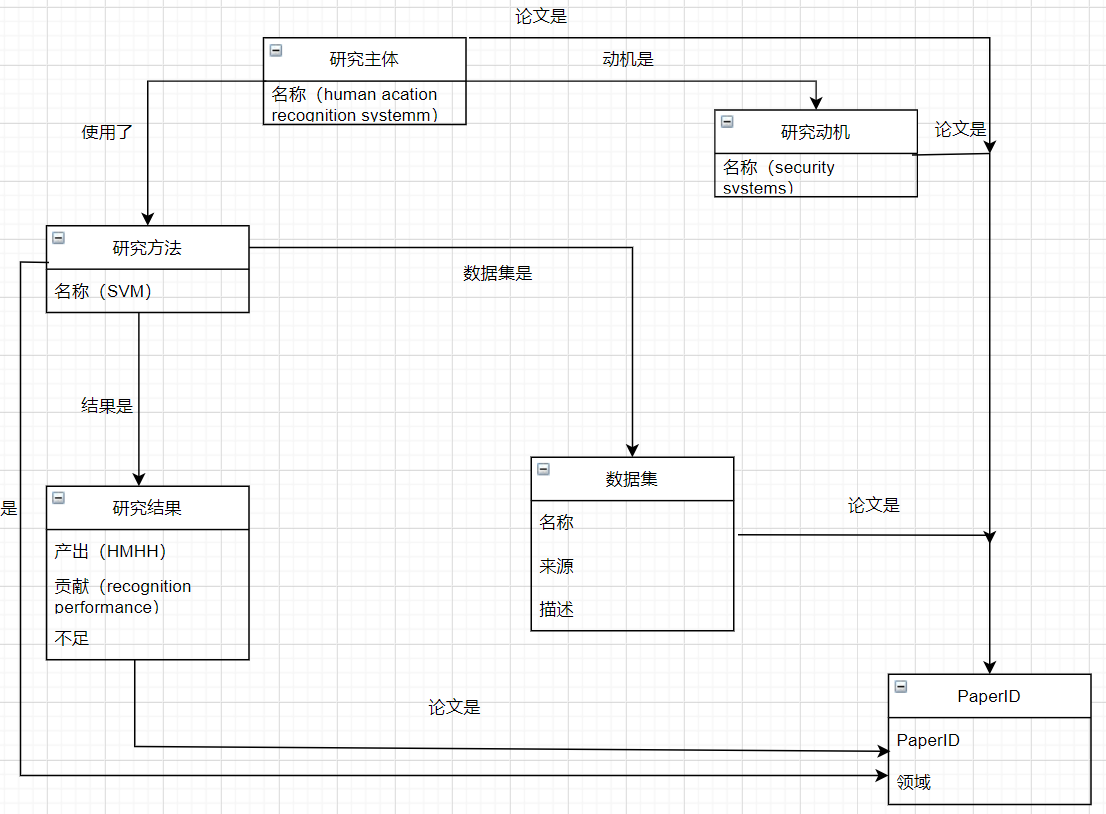
图表 1 研究问题子级体系拆解

## 形式化分析：

|  |  |  |  |
| --- | --- | --- | --- |
| 序号 | 领域问题 | 类 | 关系 |
| 1 | 论文的具体研究领域是什么？ | 研究领域 | 属于 |
| 2 | 论文的研究内容是什么？ | 研究内容 |  |
| 3 | 论文的研究任务是什么 | 研究动机 | 研究动机是 |
| 4 | 论文的研究主体是什么？ | 研究主体 | 研究主体是 |
| 5 | 论文的数据集是什么？ | 数据集 | 数据集是 |
| 6 | 论文的研究方法是什么？ | 研究方法 | 使用的方法是 |
| 7 | 论文的研究结果是怎样的？ | 研究结果 |  |
| 8 | 论文的最终结果是什么？ | 研究结论 | 得出 |
| 9 | 论文的创新点是什么？ | 贡献 | 创新是 |
| 10 | 论文的不足有哪些？ | 不足 | 不足是 |

图表 2 具体研究子问题形式化分析

## 概念建模：



图表 3 概念建模图

## 本体构建的相关参考对象与理论基础：

首先，本组所选用的SciERC数据集的对应科学知识图谱构建研究论文[[3]](#footnote-3)，对统一数据集所定义的实体类型，主要有Task, Method, Metric, Material, Other-ScientiﬁcTerm, Generic六种实体，对ai领域的摘要数据中的信息类型进行区分。

同时功能单元理论定义了读者阅读科学论文的五大常用信息使用任务，包括学习背景知识、学习方法、参考事实、参考论证、跟进研究前沿。研究者王晓梦（2018）[[4]](#footnote-4)进行科学论文功能单元本体设计时，定义了背景、主题、缘起、已有研究、假设、方法、实验、数据、结果、结论、讨论、未来工作等12个一级类，对科学论文中的知识单元进行定位。

结合数据源作为ai领域的科学论文摘要的特点，并参考已有文献对科学论文内容组件的定义以及本体类的构建。论文（摘要）（PaperID）、内容主体（Subject）、方法（Method）、数据集（DataSet）、研究动机（Motivation）、结论（Result）六个一级类别。

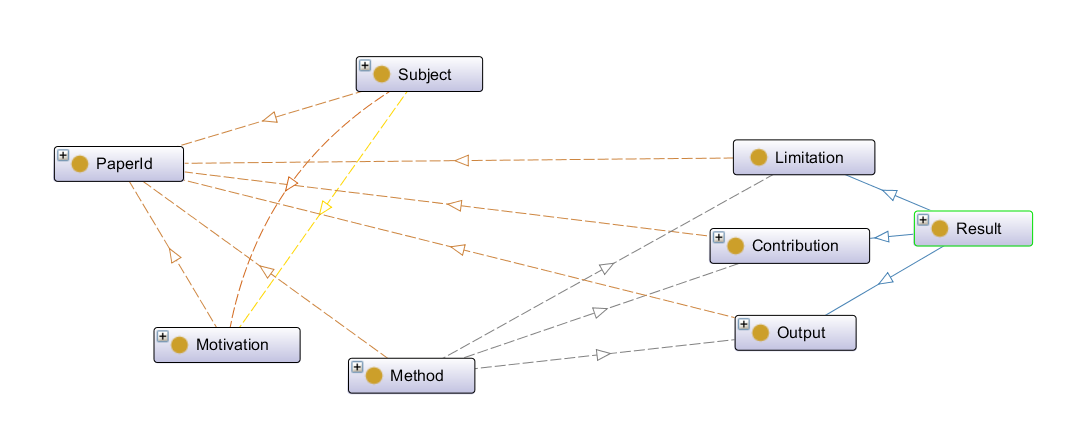
|  |  |  |
| --- | --- | --- |
| 一级类 | 二级类 | 具体解释 |
| 论文（摘要）（PaperID） |  | 论文摘要数据 |
| 研究主体（Subject） | 算法、系统、模型 | 研究对象的形式和内容 |
| 研究方法（Method） |  | 研究所涉及的具体方法，如SVM |
| 数据集（DataSet） |  | 研究所使用的数据集 |
| 研究动机（Motivation） |  | 研究目的与主要工作 |
| 结论（Result） | 产出、贡献、局限 | 指研究最终的成果 |

图表 4 本体类的构建与解释

结合不同实体在科学论文中的地位与作用，建立相应的对象属性，并构建相应关系。

|  |  |  |
| --- | --- | --- |
| 属性 | 定义域、值域 | 具体解释 |
| 出现在（isAppeardIn） | 定义域：  output、motivation、  limitation、dataset、  contribution、subject、method  值域：paperID | 主要指某实体，出现在某篇文献中。如“自动摘要系统构建”出现在“a论文” 中 |
| 使用数据集（hasDataset） | 定义域：method  值域：dataset | 指某一研究方法使用了某项数据库。如“SMV”基于“某数据集” |
| 使用方法（hasMethod） | 定义域：subject  值域：method | 指某一研究主体使用了某研究方法。如“某系统构建”使用“某方法” |
| 目的是（hasPurpose） | 定义域：subject  值域：motivation | 指某一研究主体的研究目的是什么，如“比较两个模型”是为了 “某动机” |
| 结果是（hasResult） | 定义域：method  值域：limitation、output、contribution | 指使用某一研究方法后，得到的相应研究结果 |

图表 5 关系属性解释



图表 6 实体关系属性图

## 创新与不足

创新方面，基于问题导向，探究了在论文摘要中关键信息的关系，能够帮助了解论文的主要内容和结构。

在具体研究成果方面，存在以下两方面的不足：（1）概念的抽象与识别不够精准，类与类的范围、一级类与二级类的包含关系有待进一步商榷；（2）由于概念抽象以及论文摘要数据自身复杂，给后续数据标注带来一定困难。

希望在最终知识图谱建构中进一步完善。

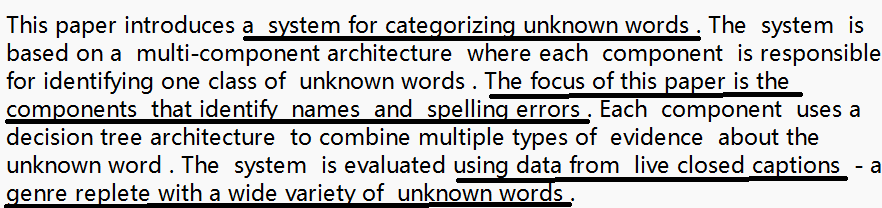
## 数据标注实例：

**PaperID：**A00-1024

**Subject：**a system for categorizing unknown words

**Motivation：**The focus of this paper is the components that identify names and spelling errors.

**Dataset：**using data from live closed captions - a genre replete with a wide variety of unknown words



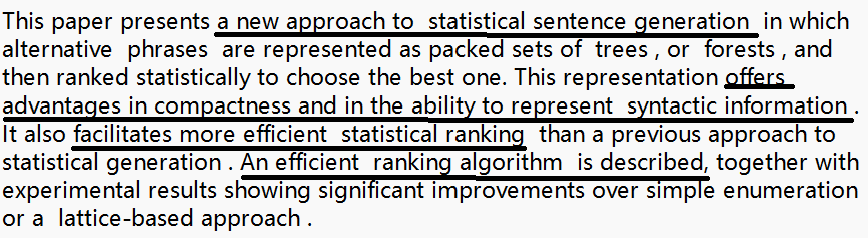
图表 7 数据标注实例1

**PaperID：**A00-2023

**Subject：**a new approach to statistical sentence generation

**Result\_contribution：**offers advantages in compactness and in the ability to represent syntactic information；facilitates more efficient statistical ranking

**Result\_output：**An efficient ranking algorithm is described



图表 8 数据标注实例2

1. 姜彩红,乔晓东,朱礼军.基于本体的专利摘要知识抽取[J].现代图书情报技术,2009(02):23-28. [↑](#footnote-ref-1)
2. 曹盼. 基于领域本体的科技文献自动摘要方法研究[D].重庆大学,2022.DOI:10.27670/d.cnki.gcqdu.2020.001560. [↑](#footnote-ref-2)
3. Multi-Task Identification of Entities, Relations, and Coreference for Scientific Knowledge Graph Construction [↑](#footnote-ref-3)
4. 王晓光,李梦琳,宋宁远.科学论文功能单元本体设计与标引应用实验[J].中国图书馆学报,2018,44(04):73-88.DOI:10.13530/j.cnki.jlis.184005. [↑](#footnote-ref-4)