技术文档

A27：基于本体的军事知识图谱自动化构建技术及应用

指导老师：徐天阳，李辉

队伍成员：丁志坚，宋佳颖，张文杰，华阳，周芮佳

（一） 研究背景

知识图谱（Knowledge Graph，KG）其本质上源于早期的语义网，能够以更直观的方式对真实世界中的实体及其关系进行清晰的描述，揭示实体之间的关系，使人类更便于认知。同时知识图谱还具有更好更强的组织管理网络大数据的能力，因此知识图谱在新一代的数据应用中也备受青睐，越来越多的基于知识图谱的智能平台先后问世。随着大数据技术的不断革新，知识图谱在智能化的道路上扮演着越来越重要的角色，推动智能技术走向新的未来。

现代的军事化建设在逐步向信息化转型，相关领域人员在面对数量众多的军事领域知识时，如何从大量未整合的数据中快速、准确地查询到自己需要的信息是亟待解决的问题。传统的检索方式主要是基于关键词匹配实现的，得到的结果往往忽略了对关键词的语义理解，无法满足用户的真正需求。自 2012 年谷歌搜索引擎首次融入知识图谱技术后，知识图谱已被广泛应用于智能搜索领域，在医疗、工业生产、金融等特定领域都有很多成功的案例。与其他领域不同，军事领域的数据获取难度大，实体间的关系也较为复杂，这些都为军事知识图谱的构建带来了困难。

军事装备数据的有效组织与存储，是构建智能化军事装备知识系统的重要基石。在军事装备领域，存在大量装备属性、装备间关系等数据，这些数据具有重要的研究与应用价值。然而，由于缺少有效的数据组织与存储结构，在面对海量、分散的军事装备数据时，相关人员难以快速准确地获取装

备信息。而知识图谱是一种以三元组为单位的数据存储结构，其本质是一张语义网络，能有效地描述实体间的关系以及实体的属性信息。知识图谱衍生于大数据时代，不仅具备强大的复杂数据管理能力和语义处理能力，而且其数据可视化能力非常强大。本项目研究基于本体的军事知识图谱自动化构建技术，借用知识图谱强大的数据处理能力和可视化优势，从而为新时代智能化的军事行动的辅助决策等应用提供数据支撑，对军事领域的现代化智能化建设具有重要的作用和意义。

（二） 研究内容及待解决的问题

1. 待解决问题

1）如何使用 OCR、数据清洗、数据纠错和信息抽取等技术，从不可编辑的 PDF 数据文件中提取可训练语料数据是一个亟待解决的问题；

2）如何使用少量示例样本，高效地微调大规模语言模型，以使其在下游任务中表现更好，是一个重要的研究方向；

3）如何利用各种图谱补全策略，解决远程监督的信息不足和知识图谱的实体或关系缺失是一个重要的研究问题；

4）如何结合外部知识，优化输入数据的选择和构造，进一步提高基于知识图谱语言模型的性能和效率是关键；

5）如何借助大型语言预训练模型，增强对话模型的常识信息、提高对话流畅度和融合对话上下文信息是一个重要的研究问题。

2. 研究内容与创新

1）面向非结构化数据的数据治理和信息提取

针对 PDF 文本无法直接编辑的问题，本项目采用了多种技术手段，包括光学字符识别（OCR）图像去噪、增强、二值化、倾斜矫正、分割等操作，提高了图像质量和文字可读性，并将图像中的文字转换为可编辑的文本。接

着，本项目对原始数据进行了检查、修正或删除，以消除数据中的错误、缺失或不一致等问题。在此之后，本项目采用了信息提取（Information Extraction, IE）方法，包括基于深度学习的命名实体识别（Named Entity Recognition, NER）和关系抽取（Relation Extraction, RE），分别从原文件中提取实体信息，以及实体之间的具体关系，从而形成了结构化信息，如三元组等。最后，本项目对提取的三元组信息做进一步的清洗与纠错，以提高数据质量和准确性。这些技术手段的应用，可以从不可编辑的 PDF 数据文件中提取可训练语料数据，从而更好地应用于各种自然语言处理领域的研究和应用。

2）少量标注数据的大模型微调策略

为了使预训练的大规模语言模型更适用于当前任务，提高其在下游任务上的性能，同时减少所需的标注数据量和计算资源，本项目采用了微调（Fine-tuning）方法。针对提取的三元组语料数据，本项目使用飞桨深度学习框架对大模型进行微调，使其能够学习到特定任务的相关知识和模式。具体而言，本项目引导模型学习前缀指令文本的向量表示，从而将其用于指导模型完成特定任务。在微调过程中，模型只学习“指令文本”的输入层嵌入表示，不使用任何额外的层或结构。因此，本项目所采用的微调方法可以在零参数增加的情况下，提高模型微调的效率和泛化能力。

3）针对远程监督的信息不足和知识图谱的实体或关系缺失的图谱补全

依赖模型直接生成基于三元信息的知识图谱往往存在监督信息不全和实体或关系缺失的问题，针对于此，本项目提出了一种综合图谱补全方案。具体方法包括：首先利用统一信息抽取框架抽取三元组得到初始知识图谱，接着使用基于文心语言模型（ERNIE 3.0）的预训练权重来做实体识别，其识别结果与抽取三元组做对比，将低于对比阈值的作为图谱缺陷数据直接

丢弃，将低于阈值且高于临界值的数据作进一步的人工筛选，符合要求的数据同高于阈值的数据一同放入知识图谱之中，完成图谱补全工作。

4）结合外部知识提高对话模型的性能和效率

为了增强对话模型的准确性，本项目结合外部知识来提升对话模型实际性能。首先根据实体检测模型从输入对话中检测出实体信息，并根据这些实体信息完成三部分的知识检索，分别为图谱检索、外部知识检索和阿里云OSS图文数据库检索。图谱和外部知识检索的内容将用于构建prompt信息。语言模型则根据这些 prompt 信息来生成回应信息。此外，web 端同时显示了阿里云检索结果，与生成信息一同反馈与用户。

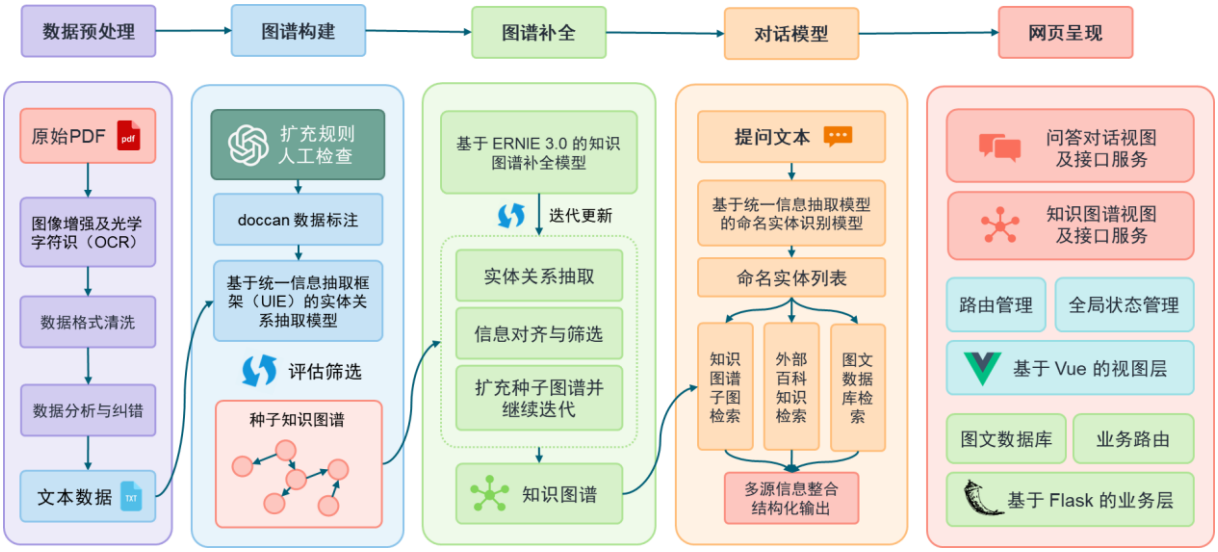
5）结合大型语言预训练模型，增强对话模型的常识信息和对话流畅度

问答系统的核心是问答模型，它能将用户的问题转换为计算机可以理解的形式，并生成相应的答案。基于检索的问答模型使用预定义的知识库来回答问题，而基于生成的问答模型则使用机器学习技术从大量文本数据中学习如何回答问题。基于检索的问答模型需要一个大而全面的知识库来支持它们的运作，而基于生成的问答模型通常需要大量标记数据来训练。最近的研究表明，结合这两种模型的优点可以提高问答系统的效率和准确性。为了实现这个目标，本项目使用基于检索的生成式语言模型，通过外部大型语言预训练模型来提升对话模型的常识信息和流畅度。这个系统综合使用检索式和生成式问答策略来回答问题，在生成最终的答案之前，首先通过检索的方法从知识图谱、外部数据库等信息源检索相关信息，然后基于提示工程的方法引导大规模语言模型输出正确的结果。

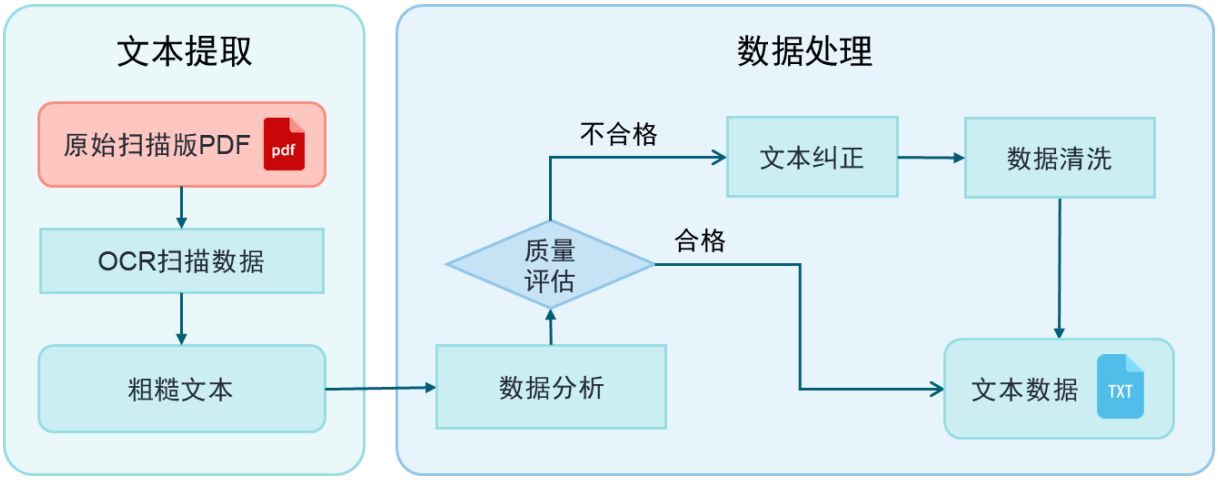
（三） 研究方法与技术路线

赛题的设置分为知识图谱的构建与问答模型的搭建；在此项目中，如下图所示，项目的研发流程可以分为五个部分，分别是（1）数据预处理、（2）

知识图谱构建、（3）知识图谱补全、（4）对话模型的构建以及（5）网页视觉呈现。



1. 数据预处理



在数据预处理阶段，本项目要将非结构化数据转换为结构化数据。由于本项目首先对原始数据进行清洗和整理，以便为后续的知识图谱构建和补全工作打下坚实基础。这一阶段的主要工作包括：

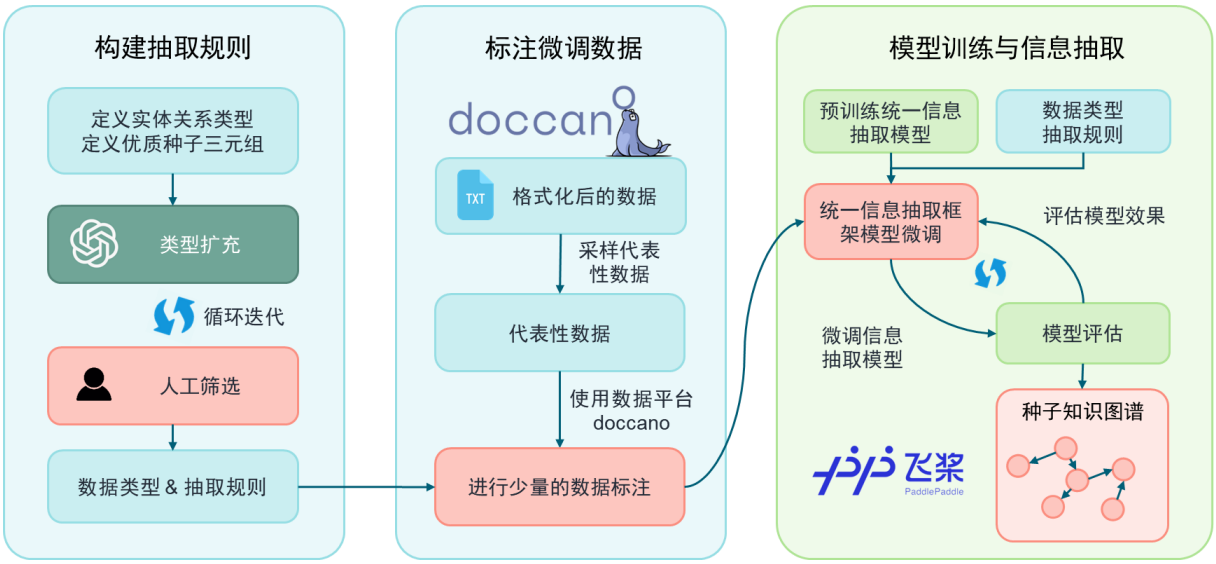
OCR 扫描文本：由于军事资料为 PDF 扫描文件，使用准确度较高的光学字符识别（OCR）技术转换将其转换为可读的文本。在处理扫描文件中的噪声、失真以及版式排列问题的同时要保证文本的质量。

数据评估与分析：在使用 OCR 技术扫描文本后，需要对数据质量进行

评估和分析。通过分析不合理数据的分布情况，针对那些出现多次的错误，编写自动化脚本来分析其对于图谱构建中实体关系抽取的影响，并进行批量处理。

数据清洗：在获取到的文本数据中，存在一些不规范、冗余或无关的内容。需要分析文本识别得到的数据的合理性，并对这些内容进行处理，例如去除特殊字符、修正拼写错误、去除多余空格和换行等。这样可以提高数据的质量，便于后续处理。

2. 领域知识图谱构建



在领域知识图谱构建阶段，本项目利用上一步预处理过的数据通过统一信息抽取模型（Universal Information Extraction，UIE）构建一个初始的军事领域知识图谱。这一阶段的主要工作包括：

构建规则：首先要分析领域特点深入理解军事领域的特点、实体、关系和属性等，明确知识图谱中需要表示的内容。然后设计信息抽取规则，基于领域特点，设计一套优质的规则来自动识别和抽取文本中存在的三元组信息（头实体，尾实体，关系）。

扩充规则：本项目使用 ChatGPT 基于现有规则，辅助生成批量可能的规则，以扩展现有的规则集；之后，对生成的候选规则进行人工筛选，挑选

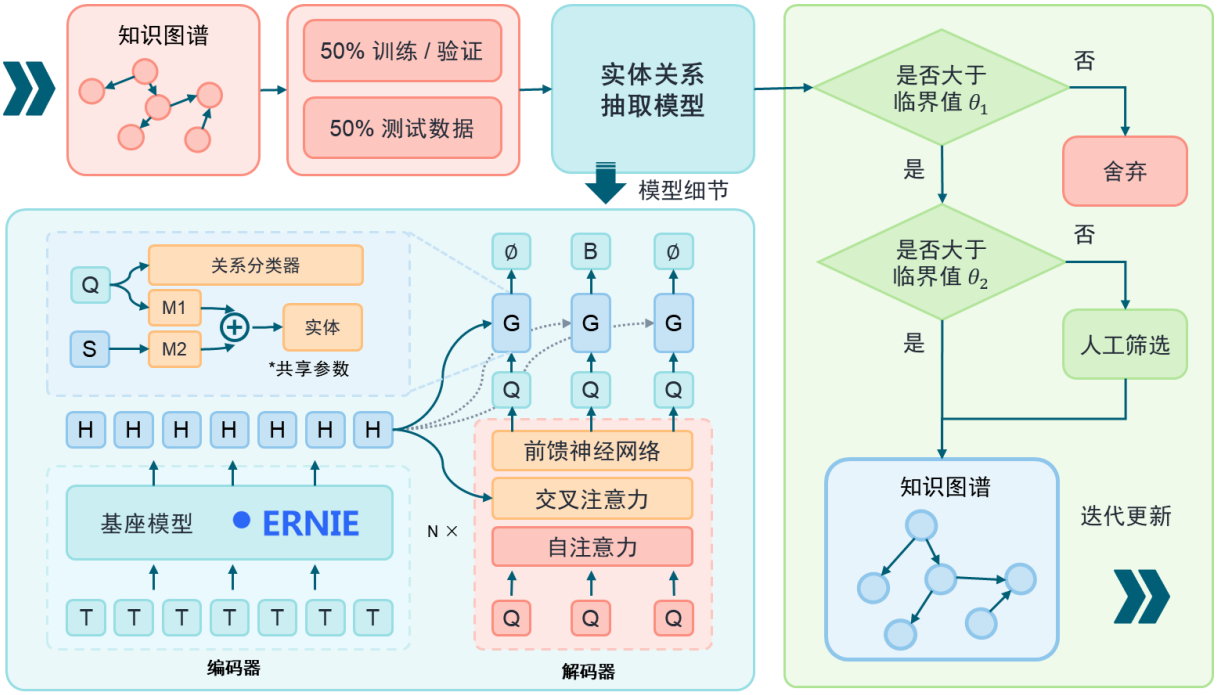
出准确性高、适用范围广的规则，以提高知识图谱构建的效果。

数据标注：先采样数据，从预处理过的数据中，按照一定的策略采样出具有代表性的样本数据。然后使用 doccano 工具对采样出的数据进行标注，添加实体、关系和属性值的标签；利用数据增强方法，能有效扩充训练数据集，有助于提高模型的性能。

训练模型：利用已标注的少量高精度数据来微调信息抽取模型，使其在抽取三元组的时候具有更高的准确性。根据训练结果，反复微调模型的参数，以提高模型在种子知识图谱构建任务上的性能。

构建领域知识图谱：将微调后的信息抽取模型应用于预处理过的数据，自动抽取实体、关系和属性值，构建初始的领域知识图谱。

3. 知识图谱补全



在知识图谱补全阶段，本项目对初始知识图谱进行优化和完善，进一步提高知识图谱的质量。这一阶段的主要工作包括：

模型选择：将图谱的数据来源与常见的实体关系抽取的数据集进行对

比，选出合适的模型。鉴于军事领域数据比较小众，本项目选择了基于集合预测策略的实体关系抽取模型作为本项目的基线模型，集合预测模型具有非自回归的特点，能够很好的处理多关系以及关系重叠的场景，对于噪声数据较高的场景也有很好的泛化能力。在基座模型的选择上，我们选择对中文预料支持更好的文心预训练语言模型（ERNIE 3.0）。

数据划分：由于数据量较少，经多次实验论证，实验中将种子知识图谱的数据的 50% 用于模型的训练以及评估，剩余的 50% 的数据集用作测试集，用于扩充现有的图谱数据。模型的抽取需要经过多次迭代，每次迭代都会将数据重新划分，以确保不同的数据都会经过有效的扩充。

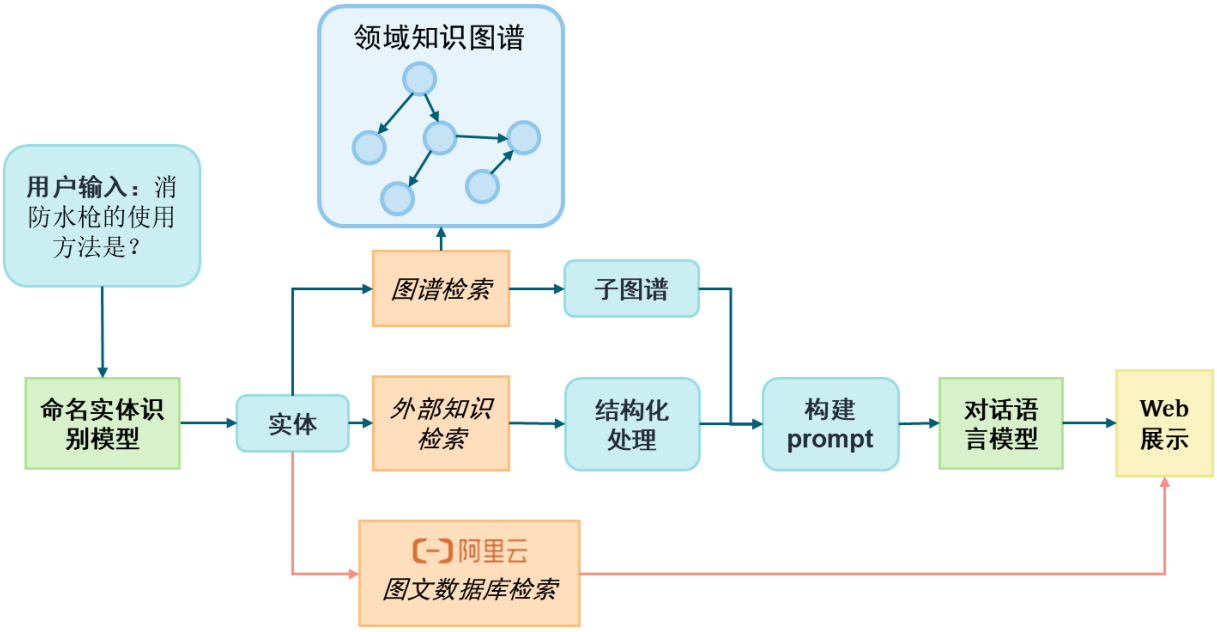
实体关系抽取：利用基于预训练权重的补全模型来抽取测试数据中的三元组，将预测结果与种子知识图谱对齐后，补充知识图谱中缺失的实体以及关系。

图谱补全：将从补全模型中得到的实体关系对依据其合理性分为三类，高于合理性阈值的直接加入图谱中，低于最低临界值的直接丢弃，合理性在二者中间的数据则需人利用自动化筛选脚本经过高效的人为筛选后再将合理的师徒关系对放入图谱中。

4. 对话语言模型

考虑到构建用户交互的问答模型需要具备流畅的对话能力、基于上下文语境的多轮对话能力、外部信息获取能力、多模态信息展示能力；

模型选择：本项目使用基于 ChatGLM 的 62 亿参数的大规模预训练语言模型作为生成模型，并结合了军事领域知识图谱检索能力、图像检索能力、外部知识检索能力。利用提示工程的经验，对上述多维度、多来源信息整合利用，生成最终的回答，并在基于 Vue 和 Flask 的网站上展示结果。



信息检索：针对用户输入的问题，使用统一信息抽取模型 UIE 提取其中的物品类实体，例如 "消防水枪"，然后根据领域知识图谱中的信息，检索与该实体相关的子图谱，以获取与该实体相关的所有信息，例如：该实体的定义、功能、使用方法、维护等信息。在获取领域知识图谱中的信息后，还通过外部知识检索，例如 Wikipedia 等，检索与该实体相关的更广泛的信息。此外，通过阿里云 OSS 图文数据库检索获取到实体相关的图片资源。

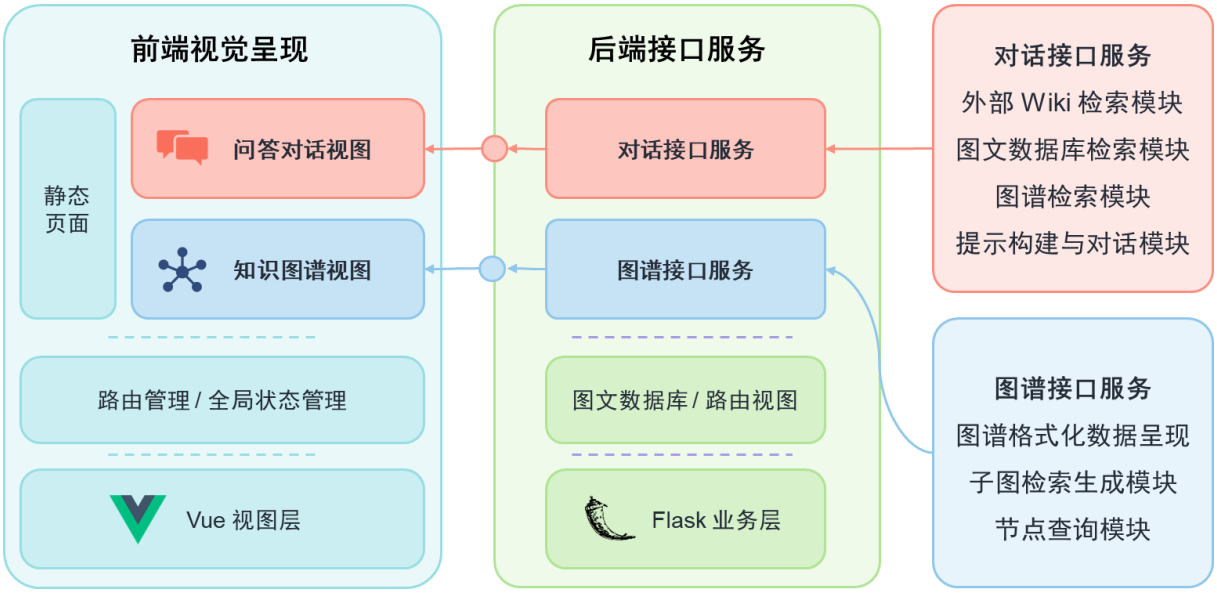
问答模型：接下来，对于所有获取到的信息进行结构化处理，以方便后续处理和展示。然后，使用 prompt 工程构建 prompt 输入，将上述结构化的数据形式作为输入，然后输入到对话模型中，得到最终的响应。使用对话语言模型，可以对输入问题进行自然语言理解和生成，从而产生具有自然流畅性的对话回复。

网页展示：最后，将获取到的答案展示给用户，可以通过网站展示呈现给用户，实现交互式问答页面的展示。在展示过程中，可以根据问题的不同，自动选择最适合的展示方式。

5. 网页构建

网站开发部分采用前后端分离的方式搭建，前端负责视觉效果呈现以

及用户交互，后端负责对话模型以及数据检索服务。



前端视觉呈现：前端部分使用基于 Vue3 框架以及相关开源组件开发搭建，使用 Vue Router 进行路由管理和 Pinia 进行全局状态管理。除了用于展示的静态页面之外，网站的核心功能包括问答对话视图和知识图谱视图。在问答对话视图中，用户可以输入军事领域相关的问题，前端通过接口请求后端进行答案生成并将答案展示给用户。在展示的相关信息中，展示了包括对话响应、维基百科内容、相关图片、相关图谱、相关数据支撑等信息，在知识图谱视图中，使用 Echarts 进行数据可视化展示，后端接口支持图谱子图检索和节点查询功能。

后端接口服务：网站的后端部分采用 Flask 框架进行后端开发，设计对话接口服务、图谱接口服务和图文数据库以及路由视图。在对话接口服务中，接收前端发送的问题，通过外部 Wiki 检索模块和图文数据库检索模块，获取相关的知识点，然后将问题和知识点传给图谱检索模块进行图谱生成，最后将答案返回给前端。在图谱接口服务中，接收前端请求的图谱数据，并通过图谱格式化数据呈现、子图检索生成模块和节点查询模块，获取图谱相关的数据，最后将数据返回给前端。

（五） 效果展示

1. 知识问答

检索能力

2. 图谱可视化

啊实打实

（六） 总结与展望

本次研究主要涉及到面向非结构化数据的数据治理和信息提取、少量标注数据的大模型微调策略、针对远程监督的信息不足和知识图谱的实体或关系缺失的图谱补全、结合外部知识提高对话模型的性能和效率、以及结合大型语言预训练模型，增强对话模型的常识信息和对话流畅度等方面的内容。通过采用多种技术手段，包括光学字符识别（OCR）图像去噪、增强、

二值化、倾斜矫正、分割等操作、以及基于深度学习的命名实体识别和关系抽取等方法，从非结构化数据中提取出结构化信息，形成了三元组等。同时，本项目采用了微调方法来提高大规模语言模型在下游任务上的性能，并使用可微调的预训练语言模型抽取三元组来进行图谱补全工作。在对话模型方面，本项目结合外部知识来提升其准确性和性能，并结合大型语言预训练模型增强对话模型的常识信息和对话流畅度。

尽管本项目取得了一定成果，仍有待改进之处：

1）由于数据量有限，无法支撑构建一个大规模的知识图谱，因此在图谱检索方面还有很长的路要走。

2）在大规模语言模型生成方面，可以利用更多领域预料进行特定场景的微调，以增强其效果。但是由于数据来源受限，无法在此次比赛中实施。

3）尽管预训练语言模型已经极大地提高了远程监督下生成的种子知识图谱的质量，但是要进一步提高图谱的性能上限，高质量的标注数据是必不可少的。由于项目资金和人力资源的限制，图谱补全部分也没有得到更深入的探索。希望未来能够有更多深入的研究。