摘要

自然语言具有多义性和模糊性，尽管在编写军事文本时会尽量标准化、规范化，但特定词汇在不同上下文中仍可能具有多种含义。因此，对军事关键实体识别任务来说，需要结合实体消歧帮助理解实体对象的具体含义，以确保后续信息检索的有效性。而军事数据具有高度保密性，相关技术需要在无网环境中进行应用。因此本文针对如何构建本地知识库、如何利用大本地知识库和大语言模型进行实体链接和跨语言实体对齐技术对本地知识库进行补充展开研究。本文主要完成以下研究工作：

（1）在对信息安全和隐私要求较高的场景中，网络连接受到严格限制，相关技术要在无网环境中进行应用。针对上述问题，为实现在离线状态下提供稳定可靠的知识支持，提出通过爬虫的方法对维基数据和维基百科中的军事实体信息进行抽取，并对数据库进行分析选型，构建本地知识库。

（2）在实体链接中，军事数据存在多义性和模糊性的问题，特定词汇在不同上下文中可能具有多种含义。针对以上问题，提出基于本地知识库和实体标准名及别名词典的实体候选集生成方法，高效地从本地知识库中召回候选集。并在链接决策阶段，引入大语言模型进行提及规范化和决策增强。在军事数据集上进行实验，MRR 和 Hits@10 指标分别达到 0.767 和 0.850。

（3）跨语言军事实体对齐旨在创建不同语言直接的军事实体对齐关系，该任务能有效提高本地知识库丰富性，进而提高实体链接质量。针对现有的军事实体中汉英链接匮乏的问题，提出引入自负采样和多个负队列的图神经网络方法，以实现跨语言实体对齐。通过自负采样，有效地捕获实体关系和复杂图形中隐藏的语义，提高中英文军事实体对齐准确性。在军事数据集上对不同参数进行实验，并与 GCN-Align、BootEA 方法进行了对比试验。该方法在 MRR 和 Hits@10 上分别达到 0.775、0873，验证了本方法在中英跨语言军事实体对齐的可行性。

通过上述实体链接方法研究，实现了军事实体本地知识库构建，完成了实体链接及跨语言实体对齐，为军事知识研究打下了基础。

关键词：本地知识库　实体链接　跨语言实体对齐　图神经网络　自负采样

ABSTRACT

Natural language is inherently polysemous and ambiguous. Although military texts are standardized and normalized as much as possible when written, specific words may still have multiple meanings in different contexts. Therefore, for the task of military key entity recognition, it is necessary to incorporate entity disambiguation to help understand the specific meaning of the entity in question, thereby ensuring the effectiveness of subsequent information retrieval. Given that military data is highly confidential, the relevant technology needs to be applied in an offline environment. Thus, this thesis aims to explore how to construct a local knowledge base and how to leverage a large local knowledge base and a large language model for entity linking and cross-language entity alignment to enrich the local knowledge base. The primary research contributions of this thesis include:

(1) In scenarios where information security and privacy are of paramount importance, network connections are strictly limited, and related technologies must be applied in a network-free environment. To address the aforementioned challenges and achieve stable and reliable knowledge support offline, this study proposes to extract military entity information from Wikidata and Wikipedia using web crawling techniques. The extracted data will then be analyzed and selected to construct a local knowledge base.

(2) In entity linking, military data is characterized by polysemy and ambiguity, with specific words often having multiple meanings across different contexts. To address these challenges, we propose an entity candidate set generation method that leverages a local knowledge base and an entity standard name and alias dictionary to efficiently recall candidate sets. Additionally, we introduce a large language model to enhance mention normalization and decision-making in the linking decision stage. Experiments conducted on a military dataset yielded an MRR of 0.767 and Hits@10 of 0.850.

(3) Cross-language military entity alignment aims to establish direct alignment relationships between military entities in different languages, a task that can effectively enhance the richness of the local knowledge base and, consequently, the quality of entity links. To address the existing issue of a lack of Chinese-English links in military entities, we propose a graph neural network approach that incorporates self-negative sampling and multiple negative queues to achieve cross-lingual entity alignment. By

employing self-negative sampling, this approach effectively captures the hidden semantics in entity relationships and complex graphs, thereby improving the accuracy of Chinese-English military entity alignment. Experiments conducted on military datasets with various parameters demonstrate that our method outperforms GCN-Align and BootEA, achieving MRR and Hits@10 scores of 0.775 and 0.873, respectively. These results verify the feasibility of our method for cross-language military entity alignment between Chinese and English.

Through the research on the entity linking method mentioned above, the construction of a local knowledge base for military entities has been realized. Entity linking and cross-language entity alignment have been completed, thereby laying the foundation for military knowledge research.

Keywords:Local Knowledge Base Entity Linking Cross-Language Entity Alignment Graph Neural Networks Ad Hoc Sampling

目录

第一章 绪论

1.1 研究背景及意义

自然语言具有多义性和模糊性，尽管在编写特殊领域文本时会尽量标准化、规范化，但特定词汇在不同上下文中仍可能具有多种含义，例如，“支援”通常指火力支援，即通过炮火、空中打击、导弹发射等手段来掩护己方部队，压制或消灭敌方力量，而在后勤运输场景中，“支援”则指提供补给和物资，包括提供弹药、食物、医疗物资等，确保前线部队的正常运作和战斗能力的维持。因此，对特殊领域关键实体识别任务来说，需结合实体消歧帮助理解从特殊领域文本中识别出的实体对象的具体含义，以确保后续信息检索的有效性。

实体链接是一种常用的实体消歧方法，该方法是将人名、地名、机构名等实体对象在文本中的提及链接至知识库中的标准条目，通过为这些实体赋予唯一的标识符，以消除实体在不同上下文中的模糊性和歧义性。由于知识库能够提供更丰富的语义信息，基于知识库的实体链接过程有助于系统正确准确特殊领域概念，减少误差和信息不一致，进而提升辅助决策的可靠性。

现有研究主要集中在利用大规模知识库（如维基百科）进行实体链接，这些知识库通常依赖于互联网进行实时更新和扩展。在处理特殊领域文本时，往往无法适应封闭环境中的数据安全性和隐私性需求。为保障特殊领域数据的安全性和隐私性，相关技术需要在无网络环境中进行应用，而由于数据源固定、信息更新缓慢，这种封闭的环境可能导致分析结果时效性受限，难以捕捉在动态特殊领域行动/战斗环境中新出现的实体或实体命名方式的变化。因此，为提升知识覆盖度，需结合知识对齐技术，及时扩展和补充外部环境中的知识。

1.2 国内外研究现状

1.2.1 实体链接技术

（1） 国内现状

中文用语灵活，同音字、多义词和省略表达的频率较高，很多词汇在不同上下文中具有不同的含义，并经常省略代词或结构助词，导致信息不完整，更是增

加了歧义性。因此，相较英文实体链接，中文实体链接需要使模型具备更加复杂的语义分析和上下文理解能力。Liu 等人所提出多视角增强蒸馏（MVD）框架[1]通过分割多个视图来避免不相关信息被过度压缩到与提及相关的视图中，并设计交叉对齐和自我对齐机制保留全局视图，以嵌入实体的整体信息，防止统一信息的分散。何展鹏提出使用加入知识库标记的 BERT 模型进行实体识别，并在孪生网络加入知识库标记及实体子图以便进行实体消歧[2]，该方法能够有效地处理因口语化严重、短文本信息少、实体多歧义等问题所导致的中文实体链接挑战。

另一方面，国内研究人员针对不同领域的实体链接也进行了探索。生物医学领域存在显著的疾病命名变异现象，包括同义词（如“高血压”和“原发性高血压”）、缩略语（如“ACS”）、词形变化和语序调整等多种表现形式。Yan 等人[3]利用无标注的大规模中医问诊语料库构建人工标注数据集，然后将实体的聚类信息加入医学知识库中进行实体链接，以缓解知识库稀疏的问题。杨书鸿等人[4]提出了融合外部知识和图卷积神经网络的生物医学信息联合识别模型，他们通过构建包含实体和语义关系的异构图，利用图卷积神经网络迭代地融合本地知识图和外部知识图中的交互信息，进而抽取生物医学实体对之间的关系。统一医学语言系统作为实体消歧的外部知识库，实体链接器根据注意力权重选择对应实体。

随着社交媒体平台迅速发展，越来越多的网络文本以短文本形态呈现。短文本存在上下文信息受限、表达不正式、语法结构不完整等特点，与英文短文本相比，中文短文本由于汉字简繁转换、音译名词识别及标点符号多样性等问题，面临着更突出的上下文信息匮乏、表达非正式化、语法结构残缺等问题。毛二松等人[5]提出一种基于词向量的中文微博实体链接方法，首先从中文维基百科抽取同义词构建同义词表，然后利用词向量解决错别字和外来人名音泽等问题，最后计算实体指称项和候选实体的语义相似度进行实体链接。张晟旗等人[6]提出基于局部注意力机制的中文短文本实体链接方法，通过拼接实体指称上下文和候选实体的描述文本将短文本转换为长文本，并引入局部注意力机制缓解长距离依赖问题，强化局部上下文信息。姜丽婷等人[7]同样将待消歧文本与候选实体拼接为一个句子，以突出实体指称的上下文语义信息，分别使用 CNN 模型和 GCN 模型获取文本的局部特征和语义特征，融合二者语义信息进行消歧，从而解决上下文文本特征提取不充分、语义信息获取较少的问题。詹飞等人[8]基于多任务学习的方法，构建多任务学习模型，以短文本实体链接为主任务，并引人实体分类作为辅助任务，促

使模型学习到更加通用的底层表达，提高模型的泛化能力，缓解短文本实体链接的信息不充分问题。

（2） 国外现状

早期提出的实体链接方法主要依赖于手动构建的知识库和相应的链接规则，处理效率低下且欠缺灵活性，难以适应新的实体和信息，导致其应用范围有限。近年来出现的预训练语言模型（PTM）能更好地理解复杂上下文和语义信息，实现更准确的实体匹配和消除歧义，促使实体链接研究取得了显著进步，如今实体链接的应用领域已拓展至社交媒体分析、知识图谱构建和智能搜索等多个方面。

PTM 通常依赖于大规模标准数据进行训练，然而在多数专业领域中，实体链接所需的标注数据往往十分稀缺甚至完全缺失，如何在低/零资源条件下场景下正确链接实体是当前的研究热点之一。针对少样本实体链接问题，Ledell Wu 等人提出了一种基于 BERT 模型的两阶段零样本实体链接方法[9]，第一阶段使用双编码器在密集空间中检索嵌入提及的上下文和实体描述，第二阶段通过交叉编码器对每个候选实体重新排名，该方法不仅在零样本基准测试中取得了最好的结果，在非零样本评估测试中同样表现出色。Edoardo Barba 等人提出了 ExtEnD 方法[10]，该方法非常适用于资源受限的数据环境，通过将任务重新定义为一个文本提取问题，并利用两个基于 Transformer 架构的模型取得了优秀的链接效率及性能表现。

针对零样本实体链接问题，G P Shrivatsa Bhargav 等人提出了一种名为 NeSLET 的方法[11]。该方法通过结合多任务学习和神经符号技术，利用实体类型层次结构的辅助信息，显著提高了零样本实体链接任务的性能，尤其是在训练数据极少的情况下。NeSLET 通过同时训练实体链接和层次实体类型预测来提升模型性能。证明了在低数据环境下，考虑实体类型的层次结构对于提高实体链接的准确性是有益的。Lajanugen Logeswaran 等人提出利用强大的阅读理解模型，并结合预训练和领域适应性预训练策略，在缺乏领域内标记数据的情况下，将提及链接到之前没有学习过的实体[12]。Tom Ayoola 等人提出了名为 ReFinED 的端到端实体链接方法[13]，该方法使用细粒度实体类型和实体描述执行实体链接，能在单次前向传递中完成文档中所有提及的提及检测、细粒度实体类型划分和实体消歧，达到了相当具有快的执行速度，并能泛化到 Wikidata 等大规模知识库。

1.2.2 实体对齐技术

（1） 国内现状

目前我国关于知识对齐技术的研究呈现出快速发展的趋势，尤其聚焦于多源异构数据和知识图谱的跨领域融合。在实际应用中，注重将知识对齐技术应用于智慧城市、医疗健康和工业互联网等具体领域，并在此基础上发展以规则为驱动的半监督对齐方法和领域适应性较强的 DL 模型。此外，由于中文文本在多义词处理和复杂上下文理解方面的独特需求，研究者更加倾向于针对特定语言特性优化对齐模型，如加入基于先验知识的语义匹配和自适应学习机制。

康世泽等人[14]提出了一种结合知识图谱的内部结构和实体描述信息共同进行跨语言实体对齐的模型。通过训练基于知识图谱结构信息的知识向量找到可能被对齐的实体对，再结合实体描述信息利用改进后的共享参数模型选出最终的对齐实体。于娟等人[15]提出了 Kernel-XGBoost 模型，使用 Kernel 提取术语的模糊对齐特征，采用分类器 XGBoost 学习对齐术语对的特征。能够实现跨语言单词和多词术语的一对多对齐，并提高准确率。聂铁铮等人[16]提出了一种基于上下文的跨语言知识图谱实体对齐方法，通过探索中心实体周围的复杂关联信息来学习实体嵌入。利用 Bi-LSTM 模型和图注意力机制构造相邻的关系信息和结构信息，更好地学习实体的向量表示来实现实体对齐。

余传明等人[17]提出了一种融合双语词嵌入的主题对齐模型（TAM），通过双语词嵌入扩充语义对齐词汇词典，在传统双语主题模型基础上设计辅助分布用于改进不同词分布的语义共享，以此改善跨语言和跨领域情境下的主题对齐效果。贾熹滨等人[18]提出了一种领域对齐对抗的无监督跨领域文本情感分析算法。该算法采用渐进迁移策略，分层减小语义差异，并通过协同优化的自适应方法实现跨领域知识迁移。张文韩等人[19]提出了基于多层结构化语义知识增强的跨领域命名实体识别（NER）模型，通过在多个层级实现对源领域和目标领域文本各自蕴含的结构化表示的对齐来促进实体识别能力跨领域迁移。

王欢等人[20]提出一种基于多模态知识图谱的中文跨模态实体对齐方法，将图像信息引入实体对齐任务，针对领域细粒度图像和中文文本，设计 CCMEA 单双流交互网络模型，基于自监督与对比学习，利用大量无标签数据对模型进行预训练，摒弃了传统监督及半监督方法依赖大量标注数据的弊端。吴含笑等人[21]提出

一种基于度量正则化的红外与可见光跨模态行人重识别算法。在对跨模态任务中 4 种模态匹配行为的距离度量矩阵进行鉴别性优化的同时，约束这 4 种匹配行为之间的差异，提高算法对模态变化的鲁棒性。郭乐铭等人[22]多尺度视觉特征提取及跨模态对齐的连续手语识别方法（MECA）。设计了一种并行多尺度视觉特征提取模型，加强了视觉特征模型对于不同时长手语动作的表征能力。同时，采用动态时间归整 DTW 进行视觉特征及文本特征的跨模态对齐，建立了表示同一手语单词的手语视觉特征之间的关联性。何佳月等人[23]提出了一种基于多级跨模态对齐的 SAR 图像舰船检测算法（MCMA-Net），通过将光学模态中丰富的知识迁移到 SAR 模态来增强 SAR 图像的特征表示。

（2） 国外现状

在军事领域中，由于各国军事术语体系和语言表达存在显著差异，同一军事目标往往存在多种语言表述形式，这给国际军事协作带来了实质性障碍。现代战争形态下，多国联合作战已成为常态，每个环节都需要精确的信息交互，而一个术语的误译或延迟会影响到作战结果。由于军事行动对信息准确性和时效性的严格要求，跨语言实体对齐技术在此领域发挥着关键性作用。

为完成跨语言知识对齐任务，A Sakor 等人[24]提出以维基数据视为基础知识库，使用英语语言模型接收英语的简短自然语言文本，输出一个实体和关系的排名列表，在 Wikidata 中标注适当的候选实体后，采用优化方法进行链接任务。Martin Gerlach 等人提出了一种机器在环的多语言实体链接系统[25]，该系统通过运用维基项目中的数百万个锚文本和数十亿用户阅读会话的数据，构建了一个上下文和语言不可知的实体链接模型。实验结果表明，该链接推荐系统在覆盖 6 种不同语言时，能够实现 80%以上的精确度，同时确保至少 50%的召回率。Wang 等人[26]提出了一种基于预融合知识图谱的多视图 KRL 的新型跨语言实体对齐框架 FuAlign。FuAlign 在统一的嵌入空间中表示实体，从而避免了不同嵌入空间之间容易出错的转换。其次，所提出的多视图表示学习模型可以捕捉知识图谱中的语义、实体上下文和长期实体依赖关系等不同类型的信息。

跨领域知识对齐是指将来自不同领域或知识库的概念、实体或关系建立映射关系的过程。其主要目标是解决不同领域间知识表示的异构性问题，包括概念层次的差异、术语表达的差异以及关系类型的差异等。这个过程通常需要处理以下几个关键挑战：相同概念在不同领域可能有不同表达，不同知识库的组织结构可

能不同，概念的详细程度可能不一致。通过跨领域知识对齐，可以实现不同来源知识的互操作和融合，从而支持跨领域的知识推理和应用。Pan 等人[27]提出了一种光谱特征对齐算法，借助与域无关的词作为桥梁，将不同域中特定于域的词对齐到统一的聚类中。通过将这些领域词聚类对齐，能够有效缩小源域与目标域的语义差异，从而提升目标领域情感分类器的训练精度。Chen[28]等人提出一个名为图最优传输（GOT）的原则型框架，在框架中，跨领域对齐被表述为图形匹配问题，实体则被表示为动态图。Zhu[29]等人针对基于 DL 的多源无监督域适配算法，提出了一个包含两个对齐阶段的新框架，它不仅能分别对齐多特定特征空间中每对源域和目标域的分布，还能利用特定域的决策边界对齐分类器的输出。

跨模态知识对齐指的是在不同模态之间建立一致性的表示。例如，图片和文本描述之间的对齐需要模型理解图片内容并生成相应的语言描述。其核心任务是将不同模态中表达相同或相关语义内容的信息进行匹配和对齐，以实现跨模态知识的融合和迁移。Castrejon 等人[30]引入一个新的跨模态场景数据集，提出了对跨模态卷积神经网络进行正则化的方法，使它们拥有与模态无关的共享表征，以实现跨模态迁移。Messina 等人[31]提出了一种名为变换器编码器推理和对齐网络（TERAN）的新方法。TERAN 强化图像和句子底层组件（即分别为图像区域和单词）之间的细粒度匹配，以保留两种模态的丰富信息，为研究高效的大规模跨模态信息检索方法铺平了道路。Park 等人[32]引入了一种新颖的特征学习框架，利用跨模态人物图像之间的密集对应关系。在像素级解决跨模态差异问题，从而更有效地抑制人物表征中与模态相关的特征。Farooq 等人[33]提出了一种独特的设计来增强基于视觉部分的特征连贯性和局部性信息。它能够在特征学习阶段隐式地学习模态之间的对齐语义。统一的特征学习有效地利用文本数据作为视觉表征学习的超级注释信号，并自动剔除无关信息。Cheng 等人[34]提出一种新型的跨模态图像-文本检索网络，在框架中设计了一个语义对齐模块，以充分探索图像和文本之间的潜在对应关系，其中使用了注意力和门机制来过滤和优化数据特征，从而获得更具区分度的特征表征。

1.3 本文工作

本文对基于在线百科的实体链接技术进行研究，包括构建本地知识库、实体链接和跨语言实体对齐。本文完成的工作如下：

（1）本地知识库构建

在没有实时网络支持的环境中，构建本地知识库是实现知识管理和应用的关键。本文对不同类型数据库进行分析比较，选择 Elasticsearch 作为核心数据库。此外，通过爬虫技术从维基数据中提取结构化数据，利用大语言模型补全部分缺失信息，并爬取分析实体对应的维基百科中丰富的文本及图片内容进行信息扩充。

（2）实体链接

实体链接作为本文研究的核心任务之一，旨在实现文本中识别到的实体指称与本地知识库中标准实体之间的精确映射。为实现这一目标，本文提出了一种结合本地知识库和 LLM 的实体链接方法。在候选实体生成阶段，结合实体标准名和别名词典，利用 ES 数据库的倒排索引机制，快速召回候选实体集。利用 LLM 对提及进行规范化处理后，在链接决策阶段，依据置信度对候选集进行重新排序。

（3）跨语言实体对齐

在跨语言实体对齐任务中，由于军事领域信息的保密性和时效性，维基百科中跨语言链接数量较少，且中英文的数量不匹配。针对军事领域特有的数据稀疏性和语言差异等挑战，本文采用结合自负采样和多个负队列的图神经网络方法来实现跨语言实体对齐，利用 LaBSE 得到不同的语言的实体和关系的词向量，计算中英文实体之间的相似度，并通过自负采样获取更多的实体关系特征向量。通过对模型参数进行训练和更新，成功构建跨语言实体对齐模型。

1.4 本文的组织架构

本文共分为 5 章，主要内容如下概述：

第一章是绪论部分。该部分系统阐述了实体链接与跨语言对齐技术的研究进展。通过分析当前技术的发展现状，明确了本文要解决的问题。同时引出了军事领域实体链接的特殊性与应用价值，说明了本文研究的意义，最后确立了本文要解决的关键问题和研究内容。

第二章是本地知识库的构建。这部分首先对各类数据库进行分析选型并完成安装配置，然后对从维基数据爬取的实体信息进行清洗补全，按照数据特征进行索引补全后，构建本地知识库。

第三章完成了实体链接任务，将文本中识别出的实体提及链接到知识库中。由于自然语言具有多义性和模糊性，特定词汇在不同上下文中可能具有多种含义。因此，需要结合实体消歧帮助理解实体对象的具体含义，而实体链接是一种常用的实体消歧方法。本章基于本地知识库和实体标准名及别名词典，得到输入提及的候选实体集，并通过 LLM 对提及进行规范化，根据置信度对候选集重新进行排序，提高实体链接的有效性。

第四章是跨语言实体对齐。在维基数据中存在很多未链接的中英文实体词条，针对该问题，本章提出了结合自负采样和多个负队列的图神经网络方法来实现跨语言实体对齐。该方法融合维基百科中的中英文军事知识，进行跨语言实体对齐，并与其它方法进行对比，给出了以 MRR、Hit@10 为评价标准的实验分析结果。

第五章是结论与展望。总结论文的主要内容，并对跨语言军事实体链接的未来发展提出设想和展望。

第二章 本地知识库构建

2.1 数据库选型

在没有实时网络支持的环境中，需要注重对本地数据的最大化利用，减少对外部数据的依赖。数据库能够通过存储静态知识，允许系统在不访问在线资源的情况下识别和链接实体，数据库性能和查询效率直接影响实体链接效果，因此需要根据数据量、访问频率、更新需求等选择合适的数据库类型。

将本地数据库作为知识库时，根据查询字段所召回候选实体的质量决定了候选实体生成的质量。传统的检索方法往往依赖关键字进行匹配，难以处理针对同一实体的不同表述查询，而向量检索通过将数据转换成高维向量，可以实现高效的语义相似性搜索，快速、准确地找到相似内容，有效应对多义词、同义词等所带来的问题。向量检索过程如图 2.1 所示，首先通过嵌入模型将文本数据转换为映射至多维空间的数值向量，接着由支持向量存储与检索的数据库，对输入向量的语义相似性进行细致分析，从而显著提高搜索和数据分类的准确性。

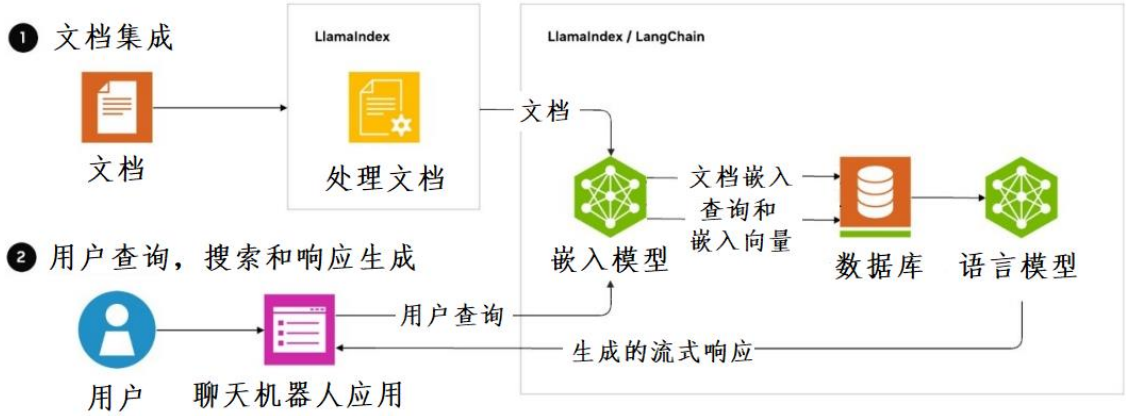

图2.1 向量查询流程

专用向量数据库（如 Milvus、Faiss、Annoy）、ES、Redis 都可以用来存储和查询向量数据，其中：向量数据库是将向量与其他数据项一起进行存储的数据库，能将文本或图像数据块等表示为支持随时创建、读取、更新和删除的数值向量；ES 过去常被用作文档型数据库，自 2022 年 2 月发布的 8.0 版本开始，提供了基于向量的搜索和自然语言处理（NLP）功能，目前已转型为了向量数据库；Redis 属于内存数据库，自 2013 年 12 月发布的 2.4 版本开始，正式加入了向量搜索功能，

并在 2023 年 8 月发布的 7.2 版本中，将向量检索作为其主要功能，以帮助加速 AI 应用的性能。其各自特点如下：

（1）专用向量数据库：专门为高效的向量检索和相似性搜索而设计，通常提供 IVF、HNSW 等复杂索引结构以优化大规模向量的存储和检索，因此能处理非常大的数据集，并支持高效的相似性搜索，如通过计算余弦相似度或欧几里得距离来找到最相似的向量；

（2）ES：为全文文本搜索设计的分布式搜索和分析引擎，支持文本、数值、地理位置等多种数据模型，并能通过向量字段支持向量检索和机器学习功能，供了水平扩展、容错和高可用性的特性，并富含 Kibana、Logstash、Beats 等庞大的生态系统，可用于数据可视化、日志分析和数据提取。

（3）Redis：支持字符串、哈希、列表、集合、有序集合等多种数据结构，能够提供极高的读写速度，适合需要快速响应的场景，可作为缓存层或需要快速读写操作的短期数据存储解决方案。

由于在实体链接过程中，不仅涉及向量检索，还需要对文本内容本身进行查询，向量数据库专注于高效向量相似性搜索，但缺乏对纯文本或结构化数值等非向量数据的灵活支持，实际表现不如 ES，后者在面对向量、文本混合检索的任务需求时非常有用，例如在处理文档或用户查询时，首先进行文本过滤，再进行向量相似性排序。因此，我们采用了 ES 作为用于存储大规模实体知识的本体数据库。

2.2 离线知识库部署

2.2.1 数据收集与预处理

（1） 数据来源

维基数据是一个自由、协作的多语言知识库，旨在为维基百科等提供结构化数据，其数据支持通过 SPARQL 查询语言进行高效检索。数据模型以实体为核心，每个实体都有唯一的标识符，并包含丰富的属性和关系。此外，维基数据的实体数据与维基百科紧密关联，许多实体都链接到对应的维基百科页面。而维基百科作为全球最大的在线百科全书，提供了丰富的文本内容，可以从中提取实体的语义信息，补充维基数据中不够完整的实体描述。如图 2.2 所示，这种结合为本地知识库的构建提供了丰富的数据来源。

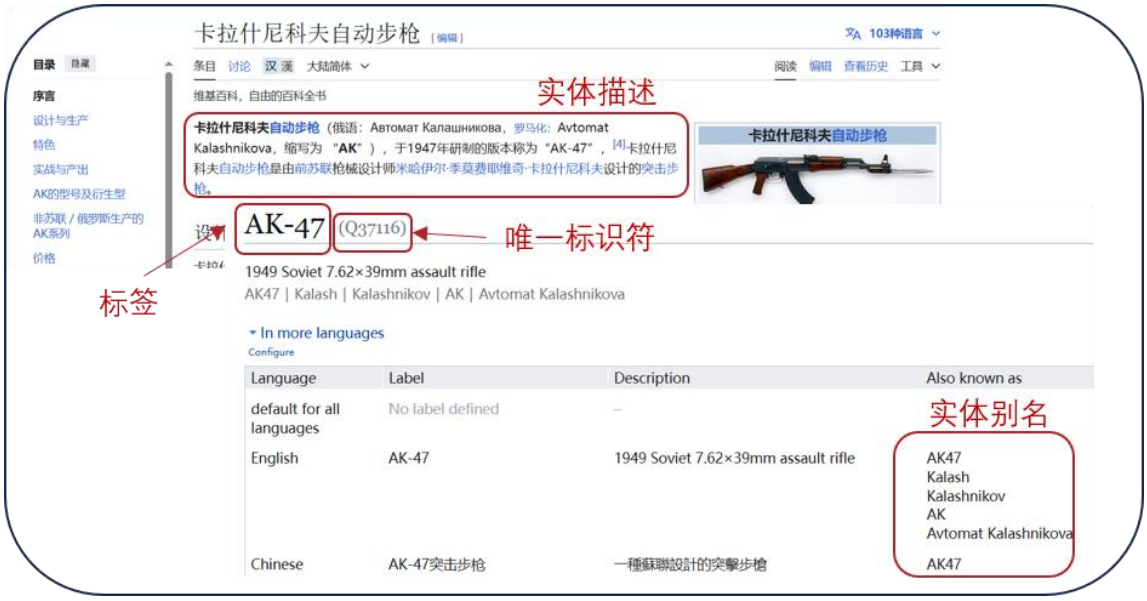

图2.2 维基数据实体示例

为了构建军事领域的本地知识库，本文通过爬虫技术从维基数据中获取相关实体的结构化数据。爬取过程基于维基数据中的实体 ID，通过递归查询其父类和子类关系，逐步获取实体信息。具体而言，通过 SPARQL 查询接口和特定领域中父类在维基数据中的唯一标识符，获取到实体的标签、别名、描述及对应的维基百科页面链接。通过维基百科链接，获取 HTML 内容和图片信息，进一步丰富实体信息。

（2） 数据预处理

如图 2.3 所示，在数据预处理阶段，我们对获取的实体数据进行了系统化的处理，以确保数据的完整性和可用性，为后续的实体链接任务奠定基础。

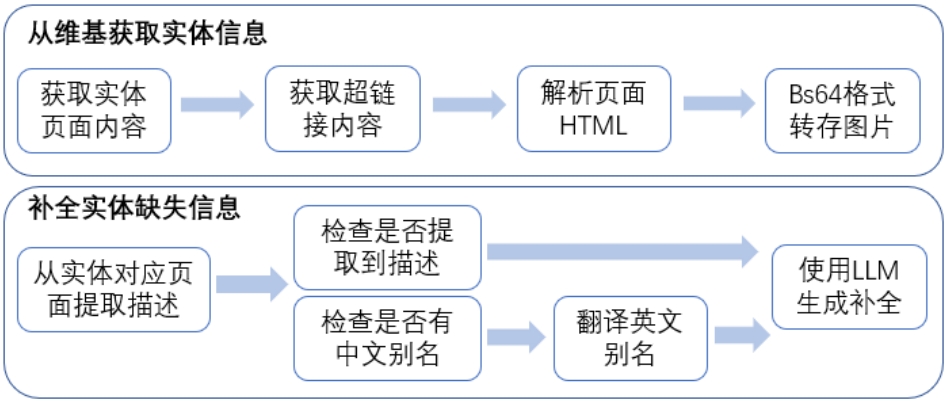

图2.3 数据预处理流程

维基百科为实体提供了丰富的文本内容，通过维基数据提供的实体对应百科页面链接，调用 API 获取页面 HTML。在百科页面中，通常包含多个指向其他页面的超链接。为进一步丰富实体信息，解析页面 HTML，提取超链接对应的页面内容。这些信息的获取能够为实体提供更丰富的背景知识。为了增强实体的可视化信息，从获取的所有页面中提取图片，并以 Base64 格式存储。通过将图片转换为 Base64 格式，能够在离线环境中高效地存储和展示，进一步丰富知识库的内容。

在数据获取过程中，发现维基数据中大部分实体的描述不够具有代表性，无法准确反映实体的核心语义，而对应的百科页面第一段内容往往是对该实体的定义或具体描述。因此，尝试从前面阶段下载得到的页面内容中提取实体描述。针对无法直接提取得到描述的实体，则通过提示大语言模型（LLM）生成。由于维基数据中部分实体缺失中文别名，这在一定程度上不利于后续实体链接候选生成。为了弥补这一不足，尝试将英文别名翻译为中文。如果无法直接通过翻译补全中文别名，则利用 LLM 进行补充。

经过上述预处理步骤，成功获取以下数据：4951 个实体及其别名、描述、对应百科内容，82383 个超链接对应的页面内容，14421 张图片的 Base64 编码内容。分别以 JSONL 格式存储在 data1.jsonl、relink.jsonl 及 image.jsonl 文件中。

2.2.2 索引构建

实现 ES 数据库本地安装后，首先创建了名为“data1”的索引。为了提升索引和查询的性能，需要在创建索引前预先对分析器（Analyzer）和映射（Mapping）进行配置，确保所有索引的数据结构和处理方式一致，以便更好地进行索引管理，避免自动映射带来的误差。

在 ES 中，分析器用于将全文本转换为一系列词符（token），不管是构建索引还是向量搜索，都需要使用分析器。分析器共分为三类，分别是：（1）字符过滤器（Character Filters），是用于对输入文本进行清洗处理的分析器，按照类型可细分为从文本中去除 HTML 元素并用解码值代替“&amp”、“&nsbp”等 HTML 实体的 html strip 字符过滤器，自定义映射以便替换指定字符的 mapping 字符过滤器，以及使用 Java 正则表达匹配应替换特定字符串的 pattern replace 字符过滤器；（2）分词器（Tokenizer）是分析器中必须的组件，ES 自带的标准分词器提供基于语法

的分词方式，适用于大多数语言，但在中文分词上的表现较差，一般需要安装第三方中文分词插件；（3）词元过滤器（Token Filters）对分词后的结果进行过滤处理，常用有大小写转换、停用词处理、同义词处理等。

配置分析器时，必需包含 1 个分词器，而根据任务需要，可以包含 0 个或多个字符过滤器和词元过滤器，多个字符/词元过滤器的执行顺序为其配置顺序。对索引“data1”配置并使用拼音分词器、IK 智能分词器和缩写分词器，同时对映射进行配置，映射用于定义存储于 ES 中的文档及其所包含字段是如何被存储和索引的。为避免不必要的分析处理，选择使用静态映射机制，在写入数据之前显示地指定字段属性。在该索引中，主要用到 6 类字段，分别表示实体标准名、中英别名、中英描述、维基百科内容、维基链接及描述对应的向量特征。同时针对每个字段需要配置数据类型、默认值、所使用的分析器等属性，对数据处理方式及其规则进行限制，如表 2.1 所示。

表2.1 索引字段配置

字段	数据类型	分词搜索	拼音搜索	缩略搜索
label	text	✓	✓	✓
aliases_zh	text	✓	✓	
descriptions_zh	text	✓	✓	
description_ vector	vector			
content	text	✓	✓	✓
link	text			

2.2.3 数据入库

该阶段是将处理得到的实体知识加载到本地 ES 数据库中，形成可查询的离线知识库。由于 ES 只支持 JSON 格式的数据作为输入，在将特征提取后得到的向量数据以及实体名称、描述等将文本数据插入到 ES 倒排索引、完成数据入库之前，首先要经过一定的格式预处理，按照预先定义好的数据库索引结构将其封装为一个 JSON 数据，确保待入库数据格式统一。如图 2.4 所示，入库后在 ElasticSearch-Head 中可以查看到索引“data1”、“data1_relink”和“data1_image”，分别存储实体知识、超链接内容及图片信息。

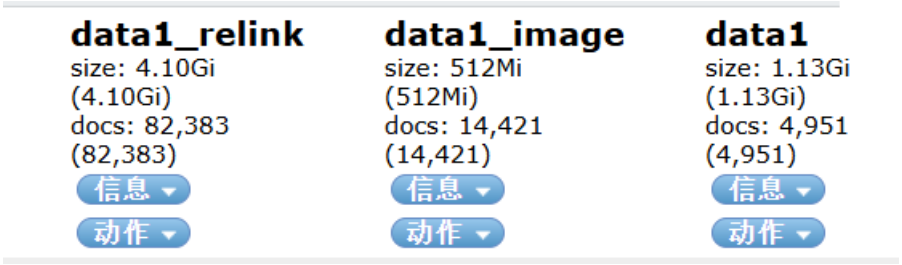

图2.4 索引概览

2.3 本章小结

本章首先对不同类型数据库进行分析，选择最合适的数据库进行安装配置。针对该本地知识库的构建任务，分析维基数据和维基百科中提供的实体信息，提出了结合维基数据结构化信息和维基百科文本及图片信息的数据获取流程。将获取到的数据进行去重、补全等操作后，按预先构建的格式封装为 JSON 文件并入库，完成本地知识库的初步构建。

第三章 实体链接

3.1 候选实体生成

在实体链接任务中，候选实体生成是关键的第一步，其目标是从知识库中快速召回与输入提及相关的候选实体集合。在本项目中，候选实体生成的工作主要依赖于 ES 数据库，通过其强大的全文检索能力和倒排索引机制，实现高效的信息检索。ES 的倒排索引原理是将文档中的内容分解为一系列单词，为每个单词创建一个索引，指向包含该词的所有文档，如图 3.1 所示。每当收到一次查询请求，ES 都将调用分析器对实体提及进行分词处理，将每个拆分后得到的词作为搜索用的关键词，在“倒排索引”中进行全文检索，并根据结果相关性评分对结果集进行排序，该评分能够反映知识库中关于实体的数据记录与查询提及的匹配程度。

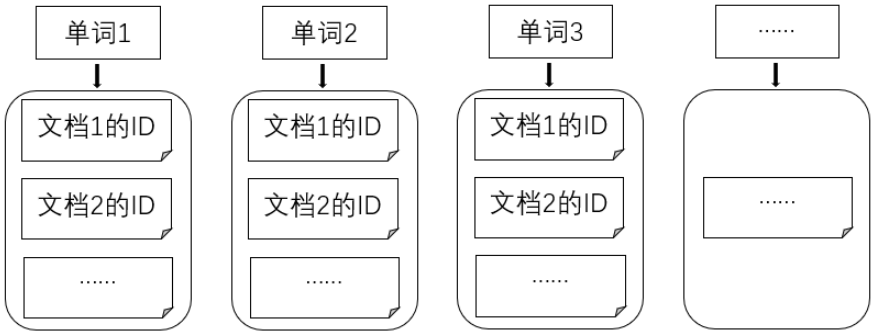

图3.1 ES 数据库的倒排索引

在候选实体生成过程中，主要依赖于输入提及与知识库中实体标准名的精确匹配。通过将输入提及与知识库中的实体标准名进行直接比较，系统能够快速识别出与输入直接匹配的实体。此外，为了进一步扩展召回范围，系统还利用实体的别名词典，如表 3.1 所示。

表3.1 实体的别名词典

标准名	实体别名
AK-47	AK47、AK47 自动步枪、AK-47 突击步枪、卡拉什尼科夫自动步枪
XB-39	波音 XB-39 超级堡垒轰炸机、XB-39 超级堡垒、波音超级堡垒 XB-39
A-36	北美野马攻击机、北美 A-36 阿帕奇、A-36 入侵者、A-36 俯冲轰炸机

别名词典中包含了实体的各种别名信息，通过这些别名信息，系统能够捕捉到与输入提及语义相同但表达形式不同的实体。通过这种方式，系统能够快速识别与输入直接匹配的实体，并借助别名信息为非标准名提及捕捉知识库中语义相同的实体。

3.2 候选实体排序

在军事文本中，常常包含大量专业术语和特定概念，其描述往往较为简略或模糊，且可能缺乏足够的上下文信息。这些特点使得仅依靠 ES 数据库对召回结果进行评分排序时，结果集排序顺序可能并不完全合理。ES 数据库虽然能够快速召回与输入提及相关的候选实体集合，但在面对军事文本中的复杂情况时，其基于倒排索引和相关性评分的机制可能无法充分考虑实体提及的多样性和歧义性。

大语言模型（LLM）在预训练阶段学习到了广泛的百科知识和通用信息，具备丰富的背景知识。在军事领域，许多专业术语和缩写虽然在离线知识库中没有明确关联，但 LLM 可能通过其预训练数据学习过这些信息，并能有效识别出来。例如，“RQ-1”是美军无人侦察机“捕食者”的别称。因此，在本项目中利用 LLM 对提及进行规范化，将非标准的实体提及转化为标准化形式，从而提高实体链接的准确性。此外，利用 LLM 丰富的背景知识，对 ES 搜索返回的结果集重新进行排序。

你现在是军事领域专家，需要参照以下例子给出提及对应的别名和定义。

例子：提及：Steyr HS .50、别名：斯泰尔HS .50狙击步枪、定义：斯泰尔HS .50（Steyr HS.50）是由奥地利斯泰尔-曼利夏公司研制的一款手动枪机式反器材狙击步枪。

输入提及：{mention}

请按照标签：{mention}、中文别名：、英文别名：、定义：的格式直接返回所需内容，不要解释或附加内容。

图3.2 用于规范化提及的提示词模板

在提及规范化过程中，利用 LLM 自动生成涵盖该提及同义词、常用名词的别名，同时生成描述了提及主要属性和背景的简洁定义。为了让 LLM 生成符合要求的提及别名及其定义，需要设计一个清晰的提示模板，以指导 LLM 生成所需内容。在该方案中，构建的提示如图 3.2 所示。

在链接决策阶段，将候选集中实体的标准名、别名、描述组织成多选项，如图 3.3 所示指导 LLM 结合规范化后的提及，对各选项的匹配度进行判断，给出每个候选实体的匹配度，分配一个置信度分数。置信度分数越高，说明 LLM 对该选项的相关性判断越高。依据置信度最高的实体 ID，在 ES 数据库中搜索对应的内容字段，提取并返回该字段内容。

```
现在你是军事领域专家，需要根据输入信息与选项列表的候选的匹配度进行从高到低排序
输入标签名：{input_label}
输入中文别名：{input_aliases_zh}
输入英文别名：{input_aliases_en}
输入定义：{input_definition}
选项列表：{''.join(options)}
请根据输入信息与选项的匹配度，从高到低严格返回所有候选的link值，确保返回的link
值是原始选项列表中的link值的排序，不能有缺失或重复，不要解释或附加内容。
```

图3.3 用于候选集排序的提示词模板

3.3 实验结果分析

为了验证本文提出的基于本地知识库的实体链接方法的有效性，本节通过一系列实验对模型的性能进行了全面评估，如表 3.2 所示。实验的目的主要是分析不同配置下模型在实体链接任务中的表现，并探讨各组件对模型性能的贡献。

其中 label 表示搜索时使用了知识库中的标准名字段，aliases 表示搜索时使用了知识库中的别名词典，*n 表示该模块所赋权重为 n，LLM 表示对候选集按置信度重新进行排序。

分析实验结果，可以发现仅使用知识库中标准名字段时，模型的 MRR 和 Hits@k 指标表现较差，说明单一字段在实体链接任务中存在局限性。在加入别名词典后，模型性能进一步提升。通过调整标准名和别名权重，指标有所变化，说

明权重的分配对模型性能有所影响，其中 label*1+aliases*1 的组合表现最优，能更好地平衡标准名和别名的作用。在结合 LLM 对候选集进行置信度重排后，模型 MRR 和 Hits@1 指标进一步提升，表明 LLM 在实体链接任务的链接决策阶段能够提升模型的准确性。

表3.2 实体链接实验结果

	MRR	**Hits@1**	**Hits@5**	**Hits@10**
label	0.289	0.262	0.323	0.350
aliases	0.738	0.691	0.806	0.837
label*1+aliases*1	0.762	0.717	0.828	0.850
label*1+aliases*2	0.761	0.717	0.828	0.844
label*1+aliases*3	0.758	0.711	0.827	0.844
label*2+aliases*1	0.759	0.714	0.822	0.847
label*3+aliases*1	0.748	0.701	0.814	0.835
label*1+aliases*1+LLM	0.767	0.728	0.813	0.850

3.4 本章小结

本章主要对基于本地知识库的实体链接任务进行了深入研究，旨在提高实体链接的准确性和效率。针对军事文本中实体提及的多样性和歧义性问题，提出了一种基于本地知识库的实体链接方法。该方法首先利用 ES 的倒排索引机制和预先处理得到的实体标准名及别名词典，快速生成候选实体集合，然后通过 LLM 对提及进行规范化处理，并结合 LLM 的背景知识对候选实体进行重新排序。试验结果表明，该方法在处理军事文本中的实体链接任务时表现出色，特别是结合 LLM 重排后，模型准确性有一定提升。此外，通过实验验证了不同权重分配对模型性能的影响，进一步优化了实体链接的流程。

第四章 跨语言实体对齐

4.1 跨语言实体对齐数据集构建

利用英文维基百科补充第二章中所构建的离线知识库，可以扩充知识库信息并提升实体链接性能。本文采用双语军事知识作为实验数据，如表 4.1 所示。主要包含中英文实体对（如“卡拉什尼科夫自動步槍”和“AK-47”）及从维基百科中获取的关联实体三元组。

表4.1 中英文链接实例

中文链接	英文链接
https://zh.wikipedia.org/wiki/卡拉什尼科夫自動步槍	https://en.wikipedia.org/wiki/AK-47

为更高效地验证模型性能，本文对实体集进行编码，如表 4.2 所示，对每个中文实体分别给出对应的实体 ID。

表4.2 跨语言实体编码示例

ID	实体名称
3858	https://zh.wikipedia.org/wiki/724 型气垫登陆艇
3860	https://zh.wikipedia.org/wiki/73 式装甲運兵車
3861	https://zh.wikipedia.org/wiki/73 式輕機槍
3862	https://zh.wikipedia.org/wiki/74 式戰車
3863	https://zh.wikipedia.org/wiki/74 號反坦克手榴彈
3864	https://zh.wikipedia.org/wiki/75/34 式自行火炮
3877	https://zh.wikipedia.org/wiki/815 型電子偵察艦

并给出表 4.3 所示的对齐后的编码集，用实体对应的唯一 ID 替换其标准名，如“6844”和“12475”，其中 6844 为中文实体“卡拉什尼科夫自動步槍”的编码，12475 为英文实体“AK-47”的编码，这个编码对表示编码 6844 和 12475 所对应的实体具有跨语言对齐关系。

表4.3 跨语言实体对齐表

中文实体 ID		英文实体 ID	
7426	奇力_XR-8	16354	Kellett_XR-8
5021	MAZ-537G	17004	MAZ-537
6337	侦察机	18595	Reconnaissance_aircraft
5780	XCG-17 滑翔機	14462	Douglas_XCG-17
8490	柯蒂斯獵鷹	14193	Curtiss_Falcon
4335	EBRC 美洲豹	14491	EBRC_Jaguar

此外，本文还对实体三元组中的关系进行编码，如“生产商”和“Manufacturer ”，其中“生产商”的编码为 292，“Manufacturer ”的编码为 624，通过编码可以构建实体间的关系集合。其结果部分如表 4.4、表 4.5 所示。

表4.4 中文关系表

实体 ID	实体名	关系 ID	关系名	实体 ID	实体名
5533	Sd.Kfz._253 半履带车	185	悬挂	2579	扭杆
4958	M35 卡車	94	原产地	9872	美国
4958	M35 卡車	292	生产商	10616	起亞汽車
7080	哥倫比亞級核潛艇	337	舰种	7833	弹道导弹潜艇

表4.5 英文关系表

实体 ID	实体名	关系 ID	关系名	实体 ID	实体名
19168	Sd.Kfz._253	534	Engine	17256	Maybach
11649	.17_HM2	515	Designer	15702	Hornady
18096	PPD-40	796	Type	19778	Submachine_gun
18096	PPD-40	815	Wars	19615	Spanish_Civil_War

4.2 基于图神经网络的跨语言实体对齐

4.2.1 总体框架

跨语言军事实体对齐的核心任务是在不同语言知识库间建立语义等价关系。具体而言，对于给定的中文军事知识库G_1和英文军事知识库G_2，旨在从G_1中识别

出与G_2中给定实体表达相同意思且语义或语言不同的实体。例如，中文实体“卡拉什尼科夫自动步枪”与英文实体“AK-47”虽表述形式不同，但均指向同一款武器，构成跨语言等价实体对。本文研究重点在于解决维基百科多语言版本间军事实体映射问题，为维基百科上中文军事实体找到等价的英文军事实体。

为有效提高跨语言军事实体对齐的准确性，本文提出了一种多特征融合的实体对齐方法，将各语言实体特征映射到统一的语义空间进行表征学习。该方法的总体框架如图 4.1 所示。对于中英文知识库中不同语言的实体和实体间关系，采用预训练的多语言模型 LaBSE 进行预处理，生成实体及其对应的属性（如生产商、原产地等）的词向量表示。对每个语言的实体进行编码，得到实体的语义向量表示，进行映射，得到一个共同的语义空间。在这一共享空间中，中文和英文的实体语义向量得以统一表示，进一步增强了跨语言对齐的准确性。为了增强跨语言实体之间的对齐效果，并进一步改进模型的自监督优化，引入了自负采样技术和多个负队列技术。为进一步改进实体嵌入，使用邻域聚合器将邻居实体的信息聚合到中心实体，为避免多跳邻居带来的噪声，确保自监督学习在实体对齐任务中的有效性，使用单层的单头图注意力网络进行聚合。对于嵌入统一空间的实体，通过度量实体之间的相对相似性来实现实体对齐。

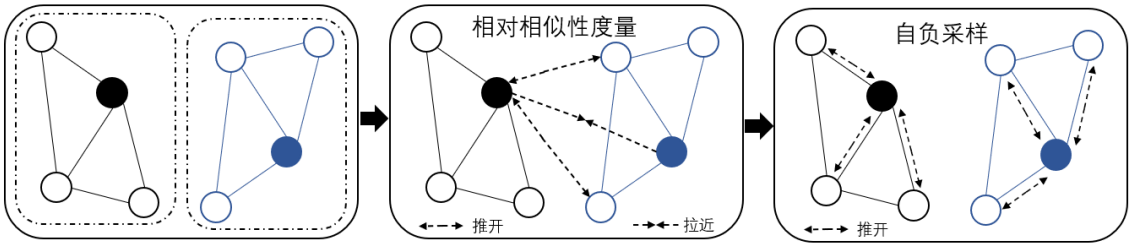

图4.1 总体框架

4.2.2 相对相似性度量

在跨语言实体对齐任务中，我们采用噪声对比估计（NCE）损失函数进行模型优化。给定两个 KGs G_x 和 G_y，分别服从概率分布 p_x 和 p_y，正实体对 $(x,y)\in R^n\times R^n$ 的表示分布为 P_{pos}。对于给定的对齐实体对 $(x,y)\sim p_{pos}$，在温度参数 τ 的控制下，通过满足 $\|f(\bullet)\|=1$ 的编码器 f 获得的表征向量满足特定分布条件，此时可定义监督式噪声对比估计 NCE 损失函数为:

$$\mathcal{L}_{NCE} \triangleq -\log \frac{e^{f(x)^T f(y)/\tau}}{e^{f(x)^T f(y)/\tau} + \sum_i e^{f(x)^T f(y_i^-)/\tau}} \tag{4-1}$$

$$= \underbrace{-\frac{1}{\tau} f(x)^T f(y)}_{alignment} + \underbrace{\log(e^{f(x)^T f(y)/\tau} + \sum_i e^{f(x)^T f(y_i^-)/\tau})}_{uniformity}$$

其中，“alignment”项用于将正对拉近，“uniformity”项用于将负对退远。先前的研究表明，NCE 损失具有以下渐进性质：对于固定的$\tau > 0$，当负样本数量$M -> \infty$时，对比损失 LNCE 收敛到其极限，绝对偏差在$O\left(M^{-2/3}\right)$中衰减。如果存在完全均匀的编码器 f，它将形成均匀项的精确最小化器。在实体对齐任务中，核心挑战在于确保语义等价关的实体获得相似的向量表示，尽管它们在命名和结构上可能存在显著差异。此外，预训练语言模型可以通过将语义相近的实体映射到嵌入空间的邻近区域，从而可以确保等式中“alignment”项，即$f\left(x\right)^T f\left(y\right)$相对较大。因此，优化 NCE 损失的核心在于改进“uniformity”项。

$$\mathcal{L}_{RSM} = -\frac{1}{\tau} + \underset{\{y_i^-\}_{i=1}^M \overset{i.i.d}{\sim} p_y}{E} [\log(e^{1/\tau} + \sum_i e^{f(x)^T f(y_i^-)/\tau}] \tag{4-2}$$

$$\leq \mathcal{L}_{ASM} \leq \mathcal{L}_{RSM} + \frac{1}{\tau}[1 - \min_{(x,y)\sim p_{pos}} (f(x)^T f(y))]$$

通过优化 LRSM，对齐的实体会相对靠近，而未对齐的实体会远离，也就是在没有正标签时，将未对齐的实体推得足够远。因此，在模型中，尝试将负对推远，摆脱对正数据即标签的使用。

4.2.3 自负采样

在负采样过程中，由于缺乏标签信息的监督，可能意外地将实际应对齐的实体对采样为负样本，这种现象称为“对齐碰撞”。当负样本采样规模较小时，这种碰撞概率可以忽略，但模型性能的提升往往依赖于大量负样本，在这种情况下碰撞概论不可忽略。为了缓解这个问题，假设模型从G_x，G_y的单空间中学习，从G_x中对实体$x \in G_x$采负本样x_i^-，可以简单地排除x来避免自负采样。

有噪声的 ASM 表示如下：

$$\mathcal{L}_{ASM|\lambda,x}(f;\tau,M,p_y)=\underset{\substack{(x,y)\sim p_{pos}\\ \{y_i^-\}_{i=1}^{M}\,i.i.d\,p_y}}{E}[-\log\frac{e^{f(x)^T f(y)/\tau}}{\lambda e^{f(x)^T f(y)/\tau}+\sum_i e^{f(x)^T f(y_i^-)/\tau}}] \tag{4-3}$$

增加负样本的数量会导致额外的计算成本，为了解决这个问题，建立负队列将先前编码的批次存储为编码负样本，以有限的成本托管数千个编码负样本。为了适应模型中的自负采样策略，维护两个与输入 KG 相关联的负队列。

模型的训练过程采用延迟更新机制，梯度更新仅在满足特定队列条件时触发，即等待任一队列累积到预设阈值 1+K，其中“1”代表当前批次的样本量，K 代表历史批次缓存数量。假设|E|为 KG 中的实体数量，K 和批次大小 N 受以下约束：

$$(1+K)\times N < MIN(|E_x|,|E_y|) \tag{4-4}$$

当前批次实际使用的负样本数量为（1+K）× N – 1。

负队列带来的主要挑战是过时的编码样本，尤其是在训练早期编码的样本，在此期间模型参数变化很大。因此，仅使用一个经常更新的编码器的端到端训练实际上可能会损害训练。为了缓解这种情况，本文采用双编码器动量训练框架，其中在线编码器通过反向传播实时更新参数 $\theta online$，而目标编码器 $\theta target$ 则进行异步更新，更新方法如下：

$$\theta_{target} \leftarrow m\cdot\theta_{target}+(1-m)\cdot\theta_{online}, m\in[0,1) \tag{4-5}$$

4.3 实验结果分析

（1） 实验数据说明

通过维基数据得到对齐实体，并通过维基百科获取实体与其它实体的关系三元组。本文构建的数据集包含三个核心部分，分别为记录中英文军事实体 ID 对应关系的映射表 ref_ent_ids、中文军事实体关系三元组集合 triples_1 和英文军事实体关系三元组集合 triples_2。实验数据集包含 21211 个中英文实体，其中中文军事实体关联的属性三元组数量为 48859，英文军事实体当中属性三元组数量为 35655，数据覆盖了主要武器装备、军事组织等核心概念。其中，训练集、验证集、测试集划分比为 8：1：1。

（2） 评价指标说明

在信息检索系统的性能评估中，主要采用以下两个核心指标来衡量模型的性能：平均倒数排名（MMR）和命中率 Hits@k。其中，MRR 计算所有查询结果中正确匹配项的倒数值的平均数。具体而言，对于给定查询，若其正确答案在结果列表中的位置为$rank_i$，则该查询得分为$1/rank_i$，系统整体的 MRR 值为所有查询得分的算术平均值。Hits@K 指标衡量的是系统在前 K 个返回结果中包含正确答案的比例。对于单个查询，当正确答案出现在前 K 个结果中时计为 1，否则为 0。系统整体的 Hits@K 值为所有查询得分的平均值。这两个指标的计算方式如下所示：

$$MRR = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{\mathrm{rank_i}} \tag{4-6}$$

$$HITS@\mathrm{n} = \frac{1}{|S|}\sum_{i=1}^{|S|}\mathrm{II}(rank_i \le n) \tag{4-7}$$

在实际研究过程中，我们采用 MRR、Hits@1、Hits@10 评估实体链接模型的有效性。

（3） 模型参数设置

Batch_size 为每次迭代时用于训练的样本数量，设置为 64。dropout 的参数为 0.3，每次训练时随机丢弃 30%的神经元，以减少神经元之间的相互依赖，提高模型的泛化能力。图注意力网络的层数为 1，用于捕获深层次的图结构信息。其余实验参数设置如表 4.6 所示。

表4.6 跨语言知识对齐实验参数

参数	参数值
momentum m	0.9999
温度 τ	2
Lr	10^{-6}
Queue size	64

其中，momentum m 用于更新动量编码器的动量系数，τ 用于对比学习中缩放嵌入向量的距离。lr 用于控制模型在每次迭代中对权重的调整步伐，选择合适的值可以让模型在较快学习速度中得到较为稳定的优化结果。Queue size 为用于负样本对比学习时的负样本队列长度。

（4） 超参数敏感性分析

通过调整模型 negative queue size 参数，验证模型在不同负队列大小时的性能，如表 4.7 所示。可以发现当负样本数量较少时，训练效率高，但模型无法充分学习到负样本的分布，导致模型性能较差。随着负样本队列大小的增加，模型性能在一定范围内逐渐提升，当队列大小超过一定范围时，性能提升逐渐停滞，同时训练时间显著增加。

表4.7 模型在不同 Negative queue size 时的表现

Negative queue size	**测试集准确率**	**训练时间/秒**	**hit@1**	**hit@10**
1	0.760	550.48	0.702	0.866
20	0.771	1748.81	0.718	0.857
40	0.775	3007.70	0.722	0.862
60	0.774	4189.92	0.720	0.863
80	0.772	6495.55	0.717	0.862

较小的 batch size 导致模型参数更新频繁，虽然可以快速适应数据，但优化路径不够稳定，表现出较低的准确率和较长的训练时间。随着 batch size 的增加，模型的准确度逐渐提升，训练时间缩短，如表 4.8 所示。但过大的 batch size 可能会导致模型的泛化能力下降，因此需要在准确率和泛化能力之间取得平衡，选择合适参数。

表4.8 模型在不同 Batch size 时的表现

Batch size	**测试集准确率**	**训练时间/秒**	**hit@1**	**hit@10**
20	0.751	4939.40	0.669	0.843
40	0.765	4598.74	0.711	0.861
60	0.774	4505.09	0.723	0.866
80	0.778	4420.48	0.724	0.870
100	0.780	4366.77	0.723	0.871

动量系数 momentum 的合理设置对模型训练具有双重影响，较低值虽然能加速收敛，但容易引发表征坍塌问题，从而降低最终性能，而过大的 momentum 收敛速度太慢。如表 4.9 所示，适当大的 momentum 对模型性能提升较好。

表4.9 模型在不同 Momentum 时的表现

Momentum	测试集准确率	训练时间/秒	hit@1	hit@10
0.9	0.751	4387.05	0.697	0.837
0.99	0.750	4388.93	0.696	0.837
0.999	0.754	4389.82	0.701	0.845
0.9999	0.780	4366.77	0.723	0.871
0.99999	0.701	4392.44	0.638	0.813

（5） 实验结果分析

为验证本文方法的有效性，进行了邻域聚合器、相对相似性度量及自负采样三个模块的消融实验及与两个模型的对比试验，结果如表 4.10 所示。分别是完整模型、移除邻域聚合器模块的模型、移除相对相似性度量模块的模型、移除自负采样模块的模型、基于 GCN 的知识图对齐和基于 Bootstrap 的实体对齐。

表4.10 跨语言军事实体对齐实验结果

	MRR	Hit@1	Hit@10
本方法	0.775	0.728	0.873
相对相似性度量+自负采样	0.716	0.646	0.863
邻域聚合器+自负采样	0.725	0.658	0.867
相对相似性度量+邻域聚合器	0.753	0.723	0.855
GCN-Align	0.581	0.498	0.755
BootEA	0.489	0.432	0.605

邻域聚合器模块：使用图注意力网络作为邻域聚合器，聚合输入的实体初始嵌入的邻居嵌入信息。实体的邻居往往包含与其语义相关的信息，该模块通过聚合邻域信息，能够有效丰富实体的嵌入信息。在跨语言对齐任务中，实体的语义信息至关重要，该模型通过引入邻域信息，能够帮助模型更好地理解实体的语义，从而更加准确地识别不同知识图谱中语义相似的实体。

相对相似性度量模块：LaBSE 的编码输入将语义相近的实体映射到嵌入空间的邻近区域，因此该模块的核心思想是推动未对齐的负样本远离，避免对正样本对的依赖，从而通过对比学习进一步优化 LaBSE 的编码输出。在负样本数量足够多的情况下，该模块能有效提高实体对齐模型的性能。

自负采样模块：对于源实体，该模块从其所处的知识图谱中采样负样本，而不是从目标知识图谱中采样。通过这种方式，避免将实际对齐的实体对采样为负样本。通过实验结果可知，该模块使模型能够更有效地学习实体的语义表示。

基于 GCN 的知识图对齐（GCN-Align）：GCN-Align 是通过训练将实体表征为低维向量，基于嵌入空间中实体间的距离，结合实体的结构嵌入和属性嵌入信息，发现实体对齐。通过结合实体结构和属性信息进行对齐预测，这种方法在军事知识数据集上，其 MRR 和 Hit@10 成绩较低。这可能是因为该方法中嵌入的是实体的属性信息而非具体属性值，在军事实体中，同类型实体具有的属性重合度较高。例如，同属“战斗机”的“A-6 入侵者式攻击机”和“拉格-3 战斗机”都具有属性“类型”、“制造商”、“主要用户”和“衍生型”。这在一定程度上影响了模型对同类型实体的属性信息的学习，从而无法很好地获得有效信息进行对齐预测。

基于 Bootstrap 的实体对齐（BootEA）：BootEA 是利用已知的一小部分跨语言实体对齐的信息，采用迭代式自适应对齐策略，结合相似性计算与语义约束的双重优化机制，逐步扩充高置信度实体对齐对。该模型在采用迭代式训练策略时，需要连续构建多个子模型，计算资源消耗大且训练周期长。此外，模型性能受到初始种子对齐的质量的影响初始阶段的微小偏差会在后续迭代过程中被逐步放大，最终影响整体对齐精度。BootEA 的模型架构中引入了多组需要人工调校的超参数，这些参数的设置不仅增加了使用复杂度，还导致实验结果存在较大波动，难以保证稳定的性能表现。

（6） 整体测评

为了全面评估跨语言对齐模块对整体任务的贡献，本文进一步进行了整体测评。利用跨语言实体对齐方法对齐中英文实体对，并将对齐后的数据补充到第二章构建的本地知识库中，进一步验证将对齐后的数据补充到本地知识库后，是否能够提升实体链接的整体性能。

表4.11 补充本地知识库后实体链接效果

	MRR	**Hit@1**	**Hit@10**
实体链接	0.767	0.728	0.850
跨语言对齐+实体链接	0.781	0.739	0.867

如表 4.11 所示，补充跨语言对齐数据后，实体链接模型的性能得到了一定提升。跨语言对齐模块能够为本地知识库补充丰富的实体信息，从而提高实体链接的准确性。

4.4 本章小结

在本章中，我们首先基于中英文维基百科的军事相关条目，构建了包含实体-关系三元组的领域数据集，为后续跨语言对齐研究提供支持。针对当前跨语言军事实体对齐面临的主要挑战，提出了采用自适应负采样策略和多个负队列的图神经网络方法实现跨语言的实体对齐。使用 LaBSE 对实体和三元组进行编码，得到实体的词向量。在图神经网络中将对齐实体拉近，非对齐实体推远，能够深入挖掘实体间潜在语义关联和复杂的图结构特征。通过实验发现，本章提出的模型与其它模型相比，在 MRR 和 HITS@10 上展现出显著优势。通过本章的研究，构建并训练了跨语言军事实体对齐模型，用来实现跨语言实体对齐，提升了实体链接质量。

第五章 总结与展望

5.1 论文总结

在知识全球化进程加速的背景下，军事领域的多语言协同与知识共享需求日益凸显，但与此同时也面临着军事数据的高机密性和文本的多义性和模糊性问题。为应对上述挑战，建立一个基于本地知识库的实体链接非常有必要。本文首先提出了一个基本维基数据结构化信息和维基百科文本及图像信息的实体获取和补全算法，构建了本地知识库，接着研究了基于本地知识库和大语言模型的实体链接方法，为实现中英文知识库的有效融合，将图神经网络框架与自负采样和多个负队列结合，显著提升了跨语言军事实体对齐的精度。实验结果表明，该方法在军事领域实体链接任务中表现出显著优势。具体工作如下：

（1）本地知识库构建

针对军事无网场景，本地知识库的构建是本文研究任务开展的基础。首先针对军事知识数据的获取问题，本文通过爬虫技术获取维基数据中的结构化数据和维基百科中的文本及图像信息。对获取的数据进行分析，通过清洗算法对冗余数据进行清洗，并补全部分缺失数据。通过分析各类数据库，选择使用 ES 数据库，并完成安装配置，构建索引信息。将处理后的数据按规定格式封装为 JSON 文件，并加载到数据库中，成功构建本地知识库。

（2）实体链接

为更好实现军事实体链接，为本地知识库中实体构建别名词典。利用 ES 数据库倒排索引机制，结合实体标准名和别名词典进行高效的检索，快速召回候选实体集。对各模块赋予不同权重，进行实验，选择最优配比方案。为进一步提高候选实体集排序，利用大语言模型对提及进行规范化处理，并构建多选项，按置信度对候选集进行排序。在军事实体数据集上进行验证实验，结果表明该方法准确性较高。

（3）跨语言中英文实体对齐

在跨语言军事实体对齐方面，提出了一种引入自负采样和多个负对列的图神经网络方法来实现中英文跨语言军事实体对齐。该方法采用自适应负采样策略构建多个负队列，在特征空间内缩小等价实体的嵌入距离，同时扩大非等价实体的

差异，有效学习实体关系的分布式特征。基于军事领域数据集进行训练，最终构建出具有较好效果的跨语言对齐模型。对比试验表明，该方案有效提升了跨语言军事实体对齐的准确性。

5.2 未来工作展望

针对特殊邻域特有的保密性和时效性要求，尽管本文提出的方法取得了一定成效，但在以下方面仍存在改进空间：

（1）在数据抽取与本地知识库搭建方面，本文设计从维基百科自动爬取军事知识文本并进行处理。该方法存在信源单一问题，知识可靠性完全依赖于维基百科，而维基百科开放编辑特性可能会导致信息可信度较低。因此，如何构建多源军事知识融合框架将称为后续研究的重点方向。此外，针对军事无网场景，如何在离线环境中高效构建和更新本地知识库，确保知识库的时效性和安全性也将成为研究重点。

（2）在实体链接方面，本文提出的方法在候选实体集生成中依赖于构建的别名词典，而别名词典数据主要来自于从维基数据获取的信息，检索时易受词典的信息限制。可以尝试使用大语言模型，提高提及对应实体出现在候选实体集中的概率。

（3）在跨语言军事实体的对齐方面，多语言实体匹配是构建跨语言知识库的核心环节。本文采用图神经网络对实体关系和隐藏特征进行捕获，该方法能显著提升跨语言实体匹配的准确性，但仍存在改进空间，对于领域特有的专业术语和缩略语，可以进一步设计专门的语义消歧模块。

参考文献

[1] Liu Y, Tian Y, Lian J, et al. Towards better entity linking with multi-view enhanced distillation[J]. arXiv preprint arXiv: 2305. 17371, 2023.

[2] 何展鹏. 基于知识库标记预训练孪生神经网络的中文实体链接[J]. Computer Science and Application, 2022, 12: 1202.

[3] Yan C, Zhang Y, Liu K, et al. Enhancing unsupervised medical entity linking with multi-instance learning[J]. BMC medical informatics and decision making, 2021, 21: 1-10.

[4] 杨书鸿, 牛玥, 刘力铭. 融合外部知识和图卷积神经网络的生物医学事件联合识别[J]. 科学技术与工程, 2024, 24(22): 094610.

[5] 毛二松, 王波, 唐永旺等. 基于词向量的中文微博实体链接方法[J]. 计算机应用与软件, 2017, 34(4): 11- 15, 41.

[6] 张晟旗, 王元龙, 李茹等. 基于局部注意力机制的中文短文本实体链接[J]. 计算机工程, 2021, 47(11): 77-83, 92.

[7] 姜丽婷, 古丽拉·阿东别克, 马雅静. 基于混合卷积网络的短文本实体消歧[J]. 中文信息学报, 2021, 35(11): 101-108.

[8] 詹飞, 朱艳辉, 梁文桐等. 基于多任务学习的短文本实体链接方法[J]. 计算机工程, 2022, 48(3): 315-320.

[9] Wu L, Petroni F, Josifoski M, et al. Scalable zero-shot entity linking with dense entity retrieval[J]. arXiv preprint arXiv: 1911.03814, 2019.

[10] Barba E, Procopio L, Navigli R. ExtEnD: Extractive entity disambiguation[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2022: 2478-2488.

[11] Bhargav G P S, Khandelwal D, Dana S, et al. Zero-shot entity linking with less data[C]//Findings of the Association for Computational Linguistics: NAACL 2022. 2022: 1681-1697.

[12] Logeswaran L, Chang M W, Lee K, et al. Zero-shot entity linking by reading entity descriptions[J]. arXiv preprint arXiv: 1906.07348, 2019.

[13] Ayoola T, Tyagi S, Fisher J, et al. Refined: An efficient zero-shot-capable approach to end-to-end entity linking[J]. arXiv preprint arXiv: 2207.04108, 2022.

[14] 康世泽, 吉立新, 刘树新等. 一种基于实体描述和知识向量相似度的跨语言实体对齐模型[J]. 电子学报, 2019, 47(9): 1841-1847.

[15] 于娟, 张晨. 基于 Kernel-XGBoost 的跨语言术语对齐方法[J]. 计算机科学, 2022, 49(z2): 11119.

[16] 聂铁铮, 马新月, 申德荣等. 基于上下文的跨语言知识图谱实体对齐方法[J]. 山西大学学报（自然科学版）, 2021, 44(3): 438-444.

[17] 余传明, 原赛, 胡莎莎等. 基于 DL 的多语言跨领域主题对齐模型[J]. 清华大学学报（自然科学版）, 2020, 60(5): 430-439.

[18] 贾熹滨, 曾檬, 米庆等. 领域对齐对抗的无监督跨领域文本情感分析算法[J]. 计算机研究与发展, 2022, 59(6): 1255-1270.

[19] 张文韩, 刘小明, 杨关等. 多层结构化语义知识增强的跨领域命名实体识别[J]. 计算机研究与发展, 2023, 60(12): 2862876.

[20] 王欢, 宋丽娟, 杜方. 基于多模态知识图谱的中文跨模态实体对齐方法[J]. 计算机工程, 2023, 49(12): 88-95.

[21] 吴含笑, 赵倩倩, 朱建清等. 基于度量正则化的红外与可见光跨模态行人重识别算法[J]. 计算机科学, 2023, 50(z1): 34351.

[22] 郭乐铭, 薛万利, 袁甜甜. 多尺度视觉特征提取及跨模态对齐的连续手语识别[J]. 计算机科学与探索, 2024, 18(10): 2762-2769.

[23] 何佳月, 宿南, 徐从安等. 从光学到 SAR: 基于多级跨模态对齐的 SAR 图像舰船检测算法[J]. 遥感学报, 2024, 28(7): 1789-1801.

[24] Sakor A, Singh K, Patel A, et al. Falcon 2. 0: An entity and relation linking tool over wikidata[C]//Proceedings of the 29th ACM International Conference on Information & Knowledge Management. 2020: 3141-3148.

[25] Gerlach M, Miller M, Ho R, et al. Multilingual entity linking system for wikipedia with a machine-in-the-loop approach[C]//Proceedings of the 30th ACM International Conference on Information & Knowledge Management. 2021: 3818-3827.

[26] Wang C, Huang Z, Wan Y, et al. FuAlign: Cross-lingual entity alignment via multi-view representation learning of fused knowledge graphs[J]. Information Fusion, 2023, 89: 41-52.

[27] Pan S J, Ni X, Sun J T, et al. Cross-domain sentiment classification via spectral feature alignment[C]//Proceedings of the 19th international conference on World wide web. 2010: 751-760.

[28] Chen L, Gan Z, Cheng Y, et al. Graph optimal transport for cross-domain alignment[C]//International Conference on Machine Learning. PMLR, 2020: 1542-1553.

[29] Zhu Y, Zhuang F, Wang D. Aligning domain-specific distribution and classifier for cross-domain classification from multiple sources[C]//Proceedings of the AAAI conference on artificial intelligence. 2019, 33(01): 5989-5996.

[30] Castrejon L, Aytar Y, Vondrick C, et al. Learning aligned cross-modal representations from weakly aligned data[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 2940-2949.

[31] Messina N, Amato G, Esuli A, et al. Fine-grained visual textual alignment for cross-modal retrieval using transformer encoders[J]. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 2021, 17(4): 1-23.

[32] Park H, Lee S, Lee J, et al. Learning by aligning: Visible-infrared person re-identification using cross-modal correspondences[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 12046-12055.

[33] Farooq A, Awais M, Kittler J, et al. Axm-net: Implicit cross-modal feature alignment for person re-identification[C]//Proceedings of the AAAI conference on artificial intelligence. 2022, 36(4): 4477-4485.

[34] Cheng Q, Zhou Y, Fu P, et al. A deep semantic alignment network for the cross-modal image-text retrieval in remote sensing[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2021, 14: 4284297.

致谢

曾无数次幻想写下这篇致谢，却迟迟未能落笔。回望四年，并非全然坦荡，但始终沐光而行。此刻，真诚以文字致以最深的感谢。

感谢我的父母过往 22 年对我每一个决定的支持，感谢他们给予的爱与包容。足够幸运能在这样普通但幸福的家庭中成长，父母的托举让我看到了很多未曾见过的风景，让我大胆地去看这个世界，也让我遇见任何事都能有底气。他们给与我足够的空间，让我拥有直面未来的勇气。

感谢我的毕设指导老师孔婉秋，自确定选题以来，孔老师始终以严谨认真的态度毫无保留地指导我的毕设。从选题的数据收集到实验设计，每一个环节都给与我细致入微的帮助。在我实验遇到问题时，孔老师总是耐心解答，给与指导，为我拨开迷雾，让我重新找到方向。

感谢在我求学路上给予我帮助的学长学姐们，正是因为他们的关怀和支持，让我在校园生活中感到温暖。感谢领航学长在我初入校园时为我指解答疑惑，在我焦虑学业时给与我鼓励，让我重拾信心。

感谢我的朋友们出现在我的生命中，在我人生的不同阶段与我同频共振。无论如今是否仍保持联系，因为有她们的出现，我的青春才格外热烈。特别感谢我的朋友卢林菲，相识六年，我们志趣相投。感谢她在我囫囵青春里画出斑斓，与我共同创建那些值得回忆的瞬间。

感谢一路坚持的自己，求学之路道阻且长，感谢自己始终没有放弃。回想每一次为考试与比赛焦虑痛苦的深夜，每一次没有达到预期时的怀疑和内耗，那些以为天大的事，如今看来都不过是生命旅途中的小小波澜。感谢自己能够坦然面对失败，同时也祝愿自己未来能够星庚顺正，运限光辉，无灾无难，百事吉昌。

在迄今为止的生命里，万分感谢遇见的一切。感谢西泠印社旁的吴山居，伴我走过十余个夏天。感谢塔尔城恣意热烈的鲜花，自此，代码不再冰冷枯燥，因为爱赋予了它血肉与温度。

最后，向这段平凡普通，但又灿烂盛大的本科生涯致谢。