开题报告

一、选题依据与价值

命名实体识别（Named Entity Recognition，NER）是自然语言处理（natural language processing，NLP）的一项基本任务。主要是将非结构化文本中的人名、地名、机构名和具有特定意义的实体抽取出来并加以归类，进而组织成半结构化或结构化的信息，再利用其他技术对文本实现分析和理解目的，这对于文本的结构化起着至关重要的作用。命名实体识别技术在信息抽取、信息检索、问答系统等多种自然语言处理技术领域有着广泛的应用。

在命名实体识别任务中，尽管传统基于规则[2-4]和基于机器学习[5-7]的方法，以及后期基于深度学习[8-9]的命名实体识别方法已经取得了非常高的识别精度，但是训练模型需要大规模标注数据，模型性能与标注数据量成正比，在训练语料匮乏的特定领域上，性能差强人意。例如，在医学领域，有大量的未标注的医疗记录、临床报告和生物医学文献。对这些数据进行标注是相当繁重的，因为它需要扎实的医学知识背景，而这只能由医学专家来完成。这种标记数据不足的情况在许多现实世界的应用中普遍存在。为了使命名实体识别模型有更广泛的用途，减少其对标记数据的依赖性，研究人员竭力研究小样本情形下的命名实体识别任务。

在自然语言处理中，实体链接（Named Entity Linking，EL）指将文本中的实体指称链接到知识库(knowledge base, KB)中相应的实体，实现实体消歧(Entity Disambiguation)，从而帮助人类和计算机理解文本具体含义的任务。例如，在文本“苹果发布了最新产品”中，表述“苹果”在知识库中对应的实体有“苹果(科技产品)”、“苹果(苹果产品公司)”和“苹果(苹果树果实)”等，实体链接就是将指称“苹果”链接到知识库中的“苹果(苹果产品公司)”，消除其他义项导致的歧义的过程。实体链接能够利用知识库丰富的语义信息，在许多领域发挥非常重要的基础性作用，例如问答系统、语义搜索和信息抽取等。实体链接也具有扩充知识库的重要功能，可以用于更新实体和关系，是知识图谱构建中的一个重要环节。实体链接研究中常用的知识库包括维基百科、Freebase 和 YAGO 等。

实体链接由两个关键模块组成：候选实体生成模块和实体消歧模块。候选实体生成的目的是给定文本中的实体指称，从知识库中选择与之相关联的命名实体的子集作为候选实体。随后这组候选实体交给实体消歧模块，该模块的目标是选择候选实体中与实体指称最匹配的实体。

候选实体生成的首要任务是识别出文档中的实体指称，即需要链接到知识库进行消歧的词或短语，这一过程与命名实体识别（named entity recognition，NER）任务较为类似。获得了实体指称后，再从知识库选取相关的候选实体。

当前实体链接方法的一个常见问题是，它们通常需要大量的训练数据来以最佳状态运行。通常情况下，这些方法需要数百万个标记的项目，这导致实体链接的适用范围非常受限。实体链接方法通常需要大规模的标记数据，其中包含了命名实体以及它们在知识库中的正确链接。手动标注这些数据是一项耗时且费力的任务，需要专业知识和专业人员的

参与。因此，获取足够数量和高质量的训练数据对于训练准确的实体链接模型来说是一项挑战。同时，实体链接任务中的实体指称和知识库中的实体是非常多样化的，而且存在数据稀疏性的问题。某些实体可能只在少数文本中指称，而且可能没有相应的知识库链接。这导致在训练数据中很难覆盖所有实体和链接情况，从而影响实体链接方法的泛化能力。

Cao 等人[1]已经表明，学习更好的指称和实体表示是改进实体链接的关键。因此在候选实体生成模块，有必要加强对命名实体识别的研究，从而在选取候选实体时更具有针对性，进而优化实体链接的性能。

本课题旨在研究在小样本情形下，如何通过少量的标注数据缓解命名实体识别的标记数据不足的问题性，以及解决实体链接忽视实体类别丰富信息的问题，并通过联合训练的形式缓解命名实体识别与实体消歧的错误传播问题。。

通常情况下，小样本指的是标注样本数量远远少于需要进行实体识别的目标领域或场景中的实体数目。在这种任务中，由于标注样本的稀缺性，传统的机器学习方法往往难以取得令人满意的结果。因此，针对小样本 NER 任务，需要采用特定的技术和方法，目前使用的方法大致可以分为两类，一类是基于提示学习[10]的方法，一类是基于距离度量的方法，通过这些方法以充分利用有限的样本数据并提高 NER 模型的性能和泛化能力。

尽管已有的命名实体识别和实体链接方法已经取得了不错的效果，但依然存在以下问题与挑战：

(1) 候选实体生成阶段，小样本命名实体识别方法缺少标记数据：由于标注信息的缺乏，常常面临着训练数据不足的挑战。传统的基于监督学习的方法在这种情况下表现不佳。因此，我们需要开发一种能够充分利用有限标注数据的方法来进行命名实体识别，考虑引入实体类别信息及其描述信息，从而补充标记数据不足的缺陷。

(2) 实体消歧阶段，传统硬原型的方法忽视了实体类别的丰富描述信息：在传统的实体消歧方法中，常常使用硬原型来表示实体类别，其中只考虑了均值或者单个类别名，忽视了对于类别的丰富描述信息。这样的表示方法在对实体进行刻画时能力有限。为了克服这一问题，本课题引入了对实体类别的自然语言定义。通过将实体类别和实体描述作为提示信息输入模型，我们可以获得一个具有丰富语义信息的软原型，从而提高实体消歧的准确性和表达能力。

(3) 忽视了命名实体识别和实体链接的联合训练方法：联合训练的优势：在过去，命名实体识别和实体消歧通常被视为一个串联的过程，即先进行命名实体识别，然后再进行实体链接。然而，这种串联方式容易导致错误的传播，从而影响最终的实体链接效果。为了解决这个问题，本课题采用了联合训练的模式，将命名实体识别和实体消歧这两个任务进行了统一的优化。通过联合训练，这两个任务可以相互促进，共享模型参数和特征表示，从而提高整体的性能。这种联合训练方式可以更好地捕捉实体在文本中的上下文和语义信息，从而提升实体链接的准确性和鲁棒性。

二、国内外研究现状

2.1 小样本命名实体识别方法

通常情况下，小样本指的是标注样本数量远远少于需要进行实体识别的目标领域或场景中的实体数目。在这种任务中，由于标注样本的稀缺性，传统的机器学习方法往往难以取得令人满意的结果。因此，针对小样本 NER 任务，需要采用特定的技术和方法，目前使用的方法大致可以分为两类，一类是基于提示学习(Prompt Learning, PL) [10]的方法，一类是基于距离度量[15]的方法，通过这些方法以充分利用有限的样本数据并提高 NER 模型的性能和泛化能力。

基于提示的方法侧重于基于提示学习优化命名实体识别的预训练语言模型[11-12]，这类方法在很大程度上依赖于模板、提示或优秀样本的数量。基于度量的方法主要目的是在源域中学习具有良好可泛化性的特征空间，然后通过最邻近的类别原型[13]或最邻近样本[14]对测试样品进行分类。

2.1.1 基于提示学习的小样本命名实体识别方法

最近，提示学习成为自然语言处理中的一种流行技术，在处理小样本问题方面显示出巨大的潜力[16]。提示学习方法通过构建特定的提示模板来引导模型进行学习。这些模板可以是语言模式或者问题形式，例如“[ENTITY]的实体类型是[ENTITY_TYPE]”。通过这样的提示模板，模型能够关注实体类型并进行识别。研究者们通过设计不同类型的提示模板，使模型能够灵活地适应各种 NER 任务。

受提示学习启发，TemplateNER[17]是一种为小样本命名实体识别引入了手工模板的方法。该方法通过枚举所有可能的潜在实体边界来识别命名实体，每次输入都需要进行多次向前传播，然而这种方法在计算上是非常耗时的。由于模板的设置基于所有可能的实体边界，TemplateNER 需要对输入文本进行多次模型推断，以确定最佳的实体边界和类型。这导致了额外的计算开销，因为每个可能的边界都需要进行前向传播和预测。当处理较长的文本或者大量的潜在实体边界时，模型的计算时间会进一步增加。

ProML[18]是一种创新的小样本命名实体识别（NER）模型，它结合了提示学习和度量学习的思想。为了改善度量学习的性能，ProML 将标签语义融入到模型中，通过设计多个提示模板来增强标签语义，并将多个基于提示的表示进行组合。在 ProML 中，提示模板的设计是关键的一步。这些模板旨在捕捉命名实体的语义特征，并通过提示学习的方式将这些特征融入到模型中。每个模板可以关注不同的语义方面，例如实体类型、上下文信息、实体间关系等。通过设计多个模板，ProML 可以从多个角度获取关于命名实体的丰富语义信息。

PromptNER[19]方法在适应新领域 NER 任务时，除了标准的几个小样本示例外，还引入了一组实体定义，这些实体定义是为了指导模型生成正确的潜在实体列表，并提供相

应的解释，以证明它们与所提供的实体类型定义的兼容性。当给定一个句子时，PromptNER利用大语言模型生成潜在实体的列表。这些潜在实体是通过在模型中引入特定的提示来生成的。这些提示可以是与实体类型相关的词语、短语或句子结构。通过在句子中插入这些提示，模型能够生成与所提供的实体类型定义相匹配的潜在实体。

尽管基于提示的小样本命名实体识别方法具有一定的优势，但这类方法依然存在着一些缺陷：

(1) 过于依赖于模板，模板的质量对模型的性能起着关键作用，高质量的模板应该准确地捕捉命名实体的特征和上下文信息，并且能够引导模型正确地预测命名实体的边界和类型。如果模板存在错误、模糊性或不完整性，模型的性能可能会受到影响。提示模板的设置会大大影响模型的性能；

(2) 过于依赖于标注数据，基于提示的方法通常需要大量的模板和提示样本来进行训练和优化，这意味着模型的性能高度依赖于这些样本中的信息。如果遇到新领域或特定领域中缺乏相关的模板和提示样本，该方法可能无法有效识别命名实体；

(3) 难以适应实体的多样性：基于提示的方法在处理多样性命名实体时可能遇到困难。命名实体可以具有各种形式和语法结构，而提示样本的覆盖范围可能有限。这种情况下，模型可能会出现泛化能力不足的问题，无法准确地捕捉新颖或复杂的命名实体。

2.1.2 基于度量的小样本命名实体识别方法

与基于模板的方法相比，基于距离度量的方法[20]在小样本 NER 任务中更受欢迎且有效。这个范式的关键思想是学习一个相似性度量来测量测试样本和参照物之间的语义相似性，例如原型学习和最近邻居方法。

基于度量学习的方法通过计算待分类样本和已知分类样本之间的距离，找到邻近类别来确定待分类样本的分类结果。基于度量学习方法的通用流程具有两个模块：嵌入模块和度量模块，将样本通过嵌入模块嵌入向量空间，再根据度量模块给出相似度得分。基于度量的学习方法可以学习语义空间中实体的表示，包括原型学习、对比学习的方法。

原型网络[21]提出了一种可以用于小样本学习的原型网络（prototypical networks）。该网络能识别出在训练过程中从未见过的新的类别，并且对于每个类别只需要很少的样例数据。原型网络将每个类别中的样例数据映射到一个空间当中，并且提取他们的“均值”来表示为该类的原型。使用欧几里得距离作为距离度量，训练使得本类别数据到本类原型表示的距离为最近，到其他类原形表示的距离较远。测试时，对测试数据到各个类别的原型数据的距离做归一化处理，来判断测试数据的类别标签。

CONTaiNER[22]使用对比学习的思想，将源域的样本和目标域的区别开来，从而获得一个较好的类别原型表示，同时使用高斯分布表征来优化句子中指称之间的分布距离，代替之前优化标签类特征的方法，能有效防止模型对源数据域过拟合学习的问题，同时使得模型更具有泛化能力。

上述基于度量的学习方法需要在一个模型中同时学习实体边界和实体类型。当源域和目标域差距较大时，由于只有少量领域适应的支持集样例，很难捕捉到如此复杂的结构信息，因此其性能会急剧下降。结果就是导致边界信息学习不充分，同时使得错误地对实

体边界进行分类，不能获得令人满意的性能。同时，上述基于度量的方法效果受到许多非实体类(即“Other”类)的干扰[23]。

为了绕过这些问题，两阶段方法[24]来处理小样本命名实体识别成为一个新趋势，它旨在将命名实体识别任务拆分为两个独立的子任务，即小样本实体边界提取和小样本实体分类，并在每个阶段执行一个任务。

两阶段方法 MAML-ProtoNet[25]中，对于小样本实体边界检测，该方法将其建模为一个序列标记问题，以避免处理嵌套实体。注意，检测模型的目的是定位命名实体，是类别不可知的。该方法只将检测到的实体跨度提供给分类模型进行实体类别的推断，通过这种方法也可以消除噪声“Other”类别原型的问题。在训练边界检测器时，该方法特别使用模型无关的元学习算法 MAML[26]来寻找一个良好的模型参数初始化。在用目标域支持集更新后，该模型的类别无关性使之能够快速适应新实体类，从而将模型能够更好地转移到目标域。对于小样本实体分类，该方法借鉴了标准的原型网络实现了分类模型，并提出了 MAML-ProtoNet 来缩小源域和目标域之间的差距。与仅使用支持集进行推理阶段相似度计算的 ProtoNet 相比，提出的 MAML-Proto 还利用这些样本修改实体边界和原型的共享嵌入空间，将来自同一实体类的边界表示进行聚类，同时将来自不同实体类别的边界表示分散，以获得更准确的精确率。

由于两阶段的分解方法在第一阶段只需要处理一个独立的边界检测任务，因此该模型在目标域可以忽视类别间的差异，从而在目标域可以找到更精确的边界值，并且比端到端方法获得更好的性能。在取得良好进展的同时，这种两阶段原型网络仍然面临两个具有挑战性的问题，即过度检测的虚假边界和相应阶段的不准确和不稳定的原型。

(1) 边界检测容易出现假正例：两阶段的分解方法在目标域测试时存在一个常见问题，即在实体边界提取任务中错误地识别出只存在于源域中的实体边界，从而导致后续的实体分类时，会将这一错误边界分配一个目标域的标签。例如源域的类别为地点、时间，目标域的类别为人名、机构，由于在源域训练时有许多带有地点的样本，因此测试阶段句子中的地点也很容易被跨度检测器识别出来，继而在后续实体分类时，错误地给一个地点实体分配目标域的标签。

(2) 类别原型缺少备语义信息：两阶段的分解方法中的原型网络旨在学习一个类型不可知的相似性度量函数，其目标是通过衡量测试样本与类别原型之间的距离来对实体进行分类。这种方法利用支持集中的少量样本构建原型，在类型无关的特征空间中进行表示。然而，由于原型是基于支持集构造的，并且使用的样本数量较少，所以它们可能存在准确性和稳定性方面的问题。这可能导致测试样本与原型之间的距离无法准确地反映实体的相似性，从而影响到实体分类的准确性。

2.2 实体链接方法

实体链接，也称为命名实体消歧，旨在将非结构化文本中有歧义的实体指称与知识库中的具体实体相关联起来，主要解决实体名的歧义性和多样性问题。实体链接由两个关键技术组成：候选实体生成模块和实体消歧模块。候选实体生成的目的是给定文本中的实体指称，从知识库中选择与之相关联的命名实体的子集作为候选实体。随后这组候选实体交

给实体消歧模块，该模块的目标是选择其中与之最匹配的实体。

2.2.1 候选实体生成方法

Logeswaran 等[27]提出了基于 BM25 的方法得到了候选实体集合。BM25 方法将关键字匹配方法与 TF-IDF 技术相结合，其将实体指称项以单词为单位进行分割，之后根据每个单词与实体指称项文本的相关性计算出每个单词的权重，之后计算出每个单词与实体摘要信息的带权相关性分数。所有单词的相关性分数之和就是当前实体与实体指称项的相似度。之后对实体相似度分数排序，前 K 个实体即为实体指称项的候选实体。BM25 方法凭借其简单的逻辑和功能强大的索引，能够高效地得到实体指称项的候选实体。但是，这种基于关键词（Term）的检索方法可能会由于忽略关键字的排序信息而无法很好地捕获文档语义，进而降低候选实体排序阶段的召回率。

文献[28-30]在候选实体生成阶段采用了 Bi-Encoder 模型对文本中的实体指称项生成候选实体。Ledell 等[28]利用不同的 Bert 编码器对实体指称项以及知识库中的实体进行编码，得到各自的向量表示。对于实体指称项来说，其编码器的输入是实体指称项及其所在文本。在用[SEP]标志位将实体指称项与其所在文本隔开的同时，用[CLS]作为输入起始标志。对于实体，其编码器的输入是实体的名称及其摘要。在用[SEP]标志位将实体的名称与摘要隔开的同时，用[CLS]作为输入起始标志。该模型在利用编码器对文本信息进行捕捉后，选取[CLS]位输入的编码作为整个句子的输入结果，在得到实体指称项以及实体的编码表示后，利用余弦相似度计算后的前 64 位实体作为实体指称项的候选实体。

Partalidou 等[29]在文献[28]的基础上深度探究 Bi-Encoder 在不同向量表示、不同激活函数形式下的召回率。其实体指称项的输入是实体指称项名称及其所在上下文，中间使用[SEP]标志位隔开，句首以[CLS]开始。实体的输入在加入实体名称、摘要的基础上，还加入了实体的类型信息，之后选择将句子中所有特殊标志位的编码信息进行拼接作为句子的输出。在得到候选实体与实体指称项的编码表示后，选择点积结果排序的前 64 位作为实体指称项的候选实体。文献[30]在文献[28]的基础上加入了知识库中实体的额外信息——实体类型、链接先验概率、知识库编码。

2.2.2 实体消歧方法

在实体消歧这一阶段，文献[29]使用 Cross-Encoder 模型用来对候选实体进行精细排序。Zhang 等[28]使用 Bert 编码器对候选实体进行编码。并且在对候选实体进行编码前选择将其与实体指称项的相关信息进行拼接，中间用[SEP]标志位隔开，同时句首使用[CLS]作为起始标志。之后[CLS]位的编码会通过一个线性层，通过这种激活方式得到候选实体与实体指称项的相似度。相似度最高的一位候选实体将作为实体指称项应该链接的对象。Cross-Encoder 模型对需要较高的运算力，但是在一定程度上会提高预测的准确率。

Yao 等[31]提出的模型是实体链接问题在候选实体排序阶段的创新。该方法提出在对文本进行编码时，Bert 需要的位置编码长度与输入的序列长度成正比。因此当实体指称项与实体的文本信息拼接后，需要位置编码更长编码器来捕捉语义信息。当 Bert 输入长度

小于 512 时，所有位置的向量都是存在关联这一前提入手，假设当输入超过 512 的编码值与前 512 位 token 的值是相似的。因此引用了重复编码的方法，通过重复较小位置 token 的编码信息来初始化较大位置 token 的编码。在利用该模型得到候选实体的编码信息后，将句首[CLS]位的编码通过线性层进行池化。之后将所有候选实体的分数送入 Softmax 函数进行归一化处理，结果最大者作为当前实体指称项应该链接的对象。

Tang 等[32]在 Cross-Encoder 的基础上采用双向注意力机制完成对候选实体的排序。其对候选实体的信息以段落为单位分割，然后逐个与实体指称项的文本拼接并送入 Bert 编码器。之后将每个编码器的第一位输出送入注意力交互层进行交互，并经过池化层后得到候选实体的编码表示。同时将实体指称项以段落为单位进行切分，并分别送入 Bert 编码器。之后将每个编码器的第一位输出连同候选实体的编码表示一起送入注意力交互层进行交互，并在池化且过线性层后得到候选实体与实体指称项的相关性分数。在得到所有候选实体的相关性分数后，选择拥有最高分数的候选实体作为本次预测的结果。

2.2.3 命名实体识别和实体消歧联合训练方法

传统实体链接方法分别解决的实体检测和实体消歧两个任务，但而没有利用两个任务之间的的相互依赖性。

GENRE[33]表示实体是实体链接任务中如何表示和聚合知识的核心。例如，维基百科等百科全书是由实体构成的，在给定查询的情况下检索此类实体的能力对于实体链接、开放域问答等任务至关重要。作者提出了 GENRE 模型，这是第一个通过生成实体的唯一名称来检索实体的系统，从左到右，以自回归方式逐个标记。模型通过自回归公式直接捕获上下文和实体名称之间的关系，有效地对两者进行交叉编码；该方式使得内存占用量大大减少，因为其编码器-解码器架构的参数与词汇量而非实体数量成比例；并在不对负数据进行子采样的情况下计算 softmax 损失。模型在实体消歧、端到端实体链接和文档检索任务上对 20 多个数据集进行了实验。

Rychlikowski 等[34]在 SlavNER 共享任务中，利用了大量的非结构化和结构化文档。非结构化信息指的是语言模型的无监督训练和词汇单元嵌入的数据。结构化文本指的是维基百科的结构化数据 Wikidata，以及词形还原规则和现实世界实体的来源信息。借助这些资源，该系统可以识别、规范化和链接实体，同时仅使用少量标记数据进行训练。

Tedeschi[35]研究了如何利用命名实体识别的优势来缩小实体链接系统之间的差距，并在大数据集和小数据集中进行验证。更具体地说，该方法通过维基百科和 WordNet 构造了实体对实体类别的映射数据集，从而为实体引入了实体类别。模型展示了一个实体链接系统如何以及在多大程度上可以从命名实体识别模型中受益，从而增强其实体的上下文表示，并改善候选实体的选择，选择与实体提及更相关的候选实体。

然而现有的联合训练方法忽视了实体类别的丰富描述信息，本课题将在此做一些改进。

2.3 研究现状总结

分析当前研究成果，如表 2-1 所示：

表 2-1 研究现状总结

研究内容		模型	特点	年份
小样本命名实体识别	基于提示学习	TemplateNER[17]	基于模板，枚举所有可能的潜在实体边界来识别命名实体	2021
		ProML[18]	结合了提示学习和度量学习的思想	2022
		PromptNER[19]	引入了一组实体定义，这些实体定义是为了指导模型生成正确的潜在实体列表	2023
	基于度量学习	ProtoNet[21]	提出了一种可以用于小样本学习的原型网络	2017
		CONTaiNER[22]	使用对比学习的思想，将源域和目标域的样本区别开来	2022
		MAML-ProtoNet[25]	两阶段的模型，先进行实体边界检测，再进行实体分类	2022
小样本实体链接方法	候选实体生成方法	BM25[27]	BM25 方法将关键字匹配方法与 TF-IDF 技术相结合	2019
		Ledell 等[28]	利用不同的 Bert 编码器对实体指称以及知识库中的实体进行编码	2020
		Partalidou 等[29]	在[28]的基础上深度探究 Bi-Encoder 在不同向量表示	2021
		Ristoski 等[30]	在[28]的基础上加入了知识库中实体的类型、链接先验概率等额外信息	2021
	实体消歧方法	Ledell 等[28]	使用 Bert 编码器对候选实体进行编码	2020
		Partalidou 等[29]	使用 Cross-Encoder 模型用来对候选实体进行精细排序	2021
		Yao 等[31]	在对文本进行编码时，Bert 需要的位置编码长度与输入的序列长度成正比。	2020
		Tang 等[32]	在 Cross-Encoder 的基础上采用双向注意力机制完成对候选实体的排序	2021
联合训练模型		GENRE [33]	第一个通过生成实体的唯一名称来检索实体的系统	2020
		Rychlikowski 等[34]	利用了大量的非结构化和结构化文档	2021
		Tedeschi 等[35]	研究了如何利用命名实体识别的优势来缩小实体链接系统之间的差距	2021

三、研究目标与研究内容

3.1 研究目标

综合分析小样本命名实体识别和小样本实体链接的研究现状，并针对现有基于分解的两步命名实体识别方法和实体链接的缺陷，本课题的研究目标为提出一种基于对比学习的由小样本命名实体识别方法优化的实体链接方法。

3.2 研究内容

基于上述研究目标，结合下游任务的应用场景和现实需求，本课题的具体研究内容主要包括以下三个部分：

(1) 一种基于对比学习的命名实体识别模型：通过命名实体类别的语义表示构建类别软原型，并将其为实体指称的提示进行对比学习，从而优化命名实体识别模型；
(2) 一种基于对比学习的实体链接模型：为知识库中的实体分配一个实体类别，并将实体类别和实体描述作为该实体的提示，与文本中的实体指称进行对比学习，从而优化实体链接模型；
(3) 一种命名实体识别和实体链接联合训练模型：同时考虑命名实体识别和实体链接两个任务，以共同学习模型在识别命名实体的同时进行实体链接的能力。

3.2.1 一种基于对比学习的小样本命名实体识别模型

小样本命名实体识别任务的目标是基于目标域少量标记样本识别出命名实体。本课题提出了一种具有类别语义信息的两阶段命名实体识别方法，由实体边界检测模块和实体分类模块组成。如图 3-1 所示，模型的输入为自然语言文本，先由边界检测模块识别出实体边界，再由实体分类模块经过对比学习得到实体类别，最终模型的输出为文本中的实体指称及对应实体类别。

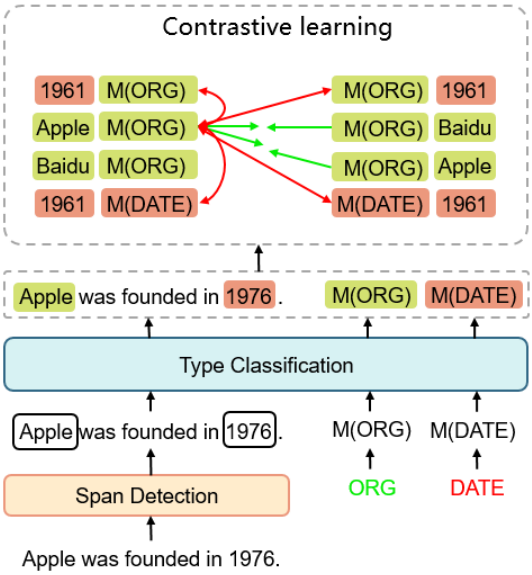

图 3-1 基于对比学习的命名实体识别模型

3.2.1.1 实体边界检测模块

实体边界检测模块旨在定位输入序列中的所有命名实体位置，并且是类型无关的，也就是说，不区分特定的实体类别。因此，模型的参数可以在不同的领域和类别之间共享。

具体而言，对于实体边界检测，将其建模为序列标注问题，以避免处理实体嵌套问题。需要注意的是，实体边界检测模型旨在定位命名实体的位置，而与实体类别无关。本模型忽视实体边界以外的词，仅将检测到的实体边界输入给实体分类模块进行实体类别推断，因此也消除了噪声"Other"的原型难以表示的问题。本模块通过一个带有线性层的预训练语言模型来训练边界检测模型，以促进内部表示中领域不变量的学习，而不是领域特定特征。通过这种方式可以使模型学习到类别无关的实体边界的元知识，从而在一个新的领域也能表现出不错的性能，即模型能够更好地迁移到目标领域。

本模块使用预训练语言模型 BERT 作为编码器，给定一个输入句子 $X = \{x_1, x_2, \ldots, x_N\}$，编码器为每个 token 生成对应的上下文表示，如下所示：

$$\mathbf{H} = [\mathrm{h}_1, \ldots, \mathrm{h}_\mathrm{N}] = f_{\theta 1}([x_1, \ldots, x_N]) \tag{1}$$

然后将上下文表示 $\mathbf{H}$ 输入一个分类层，该分类层由一个 dropout 层和一个线性层组成，再使用 softmax 函数得到词的概率分布：

$$\mathbf{p}(\boldsymbol{x_i}) = softmax(Dropout(\mathbf{W} \cdot \mathbf{h_i} + \mathbf{b})) \tag{2}$$

其中 W 是权重矩阵，b 是偏移量。随后使用概率分布的平均交叉熵和真实标签表示实体边界检测模块的训练损失：

$$L_{span} = \frac{1}{N}\sum_{i=1}^{N} CrossEntropy(y_i, \mathrm{p}(\boldsymbol{x_i})) \tag{3}$$

其中y_i 表示第 i 个词是否为实体词，当第 i 个标记为 O 时，$y_i = 0$ ；否则，$y_i = 1$。在训练过程中，通过该模块的损失函数L_{span}，不断修正模型参数θ_1、W、b，使得损失最小化。

3.2.1.2 实体分类模块

在传统两阶段的命名实体识别方法中，往往会通过原型网络来学习一个类型不可知的相似度度量函数，即原型并不具备类别的语义信息，再通过测试样本中的实体与学习到的原型的距离来进行分类。由于原型是在类型无关特征空间中仅仅使用很少的支持集样本构造的，得到的原型可能是不准确和不稳定的。

因此本课题提出了一个基于对比学习的具有类型语义信息的类别原型构造方法，通过联合利用类型描述和支持集样本作为参考来构造更准确和稳定的原型。通过这种方式，类型描述可以作为原型的指导，从而优化实体分类的性能。如图 3-1 所示，将边界检测模块得到的实体词符和对应的实体类别，分别输入到实体分类模块，得到对应的上下文表示，并经过连接操作使得实体词符具有感知实体类别的能力。

给定一个输入句子$\mathrm{X} = \{x_1, x_2, \ldots, x_N\}$，其中实体边界检测阶段识别出的实体词符的序列为$\mathrm{E} = \{e_1, e_2, \ldots, e_M\}$，E 是 X 的子集，实体词符对应的真实标签为$\mathrm{Y} = \{y_1, y_2, \ldots, y_M\}$。上述实体词符对应的实体类型为$\mathrm{T} = \{t_1, t_2, \ldots, t_M\}$，手动将实体类型转换为相应的类型描述，形式化地用 Map 映射函数来表示，即$\mathrm{T}' = \mathrm{Map(T)} = \{t'_1, t'_2, \ldots, t'_M\}$，例如对于 PER 这一类别，对应的实体描述为 Map(PER) = A person is a human individual。

然后通过对比学习的方式获得实体软原型。为了获得具有类型语义感知的实体词符，并进一步用于计算对比损失，我们将实体词符与其相应的类别描述进行连接，并按两个顺序排序，即“实体词符 - 类型描述”顺序和“类型描述 - 实体词符”顺序。如图 3-1 所示，实体词符和类别描述分别送入实体分类模块，并使用独立于实体边界检测的另一个编码器$f_{\theta 2}$来获得实体词符的上下文表示：

$$\mathrm{h_i^{el}} = f_{\theta 2}(e_i) \oplus f_{\theta 2}(Map(y_i)) \tag{4}$$

$$\mathrm{h_i^{le}} = f_{\theta 2}(Map(y_i)) \oplus f_{\theta 2}(e_i) \tag{5}$$

其中，$\oplus$是连接操作符，$\mathrm{h_i^{el}}$和 $\mathrm{h_i^{le}}$表示实体词符e_i的两种类型感知表示，它们分别按“实体词符 - 类型描述”顺序和“类型描述 - 实体词符”顺序获得。

获得上述表示后，就可以通过对比学习的方法，首先构建正负例。使那些具有相同类别但顺序不同的样本互为正例，具有不同类别的样本互为负例，如图 3-1 对比学习模块所示。对比损失计算如下：

$$L_{type} = -\sum_{i=1}^{M} \log \frac{\frac{1}{\|Z_i\|} \sum_{z \in Z_i} \exp\left(sim\left(\mathrm{h_i^{el}}, \mathrm{h_z^{le}}\right)/\tau\right)}{\sum_{j=1}^{M} \exp\left(sim\left(\mathrm{h_i^{el}}, \mathrm{h_j^{le}}\right)/\tau\right)} \tag{6}$$

$$sim\left(\mathrm{h_i^{el}}, \mathrm{h_z^{le}}\right) = \frac{\mathrm{h_i^{el}}, \mathrm{h_j^{le}}^{T}}{\sum_{k=1}^{M}\left(\mathrm{h_k^{el}}, \mathrm{h_j^{le}}^{T}\right)} \tag{7}$$

其中 M 是实体词符的数量，Z_i 是具有相同类别的正样本集，函数 sim()采用了用带有归

一化因子的点积运算作为相似性度量函数，τ 为一个温度超参数。在训练期间，通过该模块的损失函数L_{type}，不断修正模型参数θ_2，使得损失最小化.

综上，基于小样本的命名实体识别方法的总损失为：

$$L_{NER} = \alpha_1 * L_{span} + \beta_1 * L_{type} \tag{8}$$

其中，α_1和β_1是损失函数的权重系数，用于平衡两个模块的重要性。根据具体的任务需求和数据集特点，可以调整这些权重系数来使得两个模块对模型训练的影响程度相对均衡。

通过将边界检测模块和实体分类模块结合起来，我们能够有效地提取出文本中的实体词符，并通过对比学习改进其上下文和实体类别的表示，从而进行准确的分类。这种两阶段的方法允许我们在命名实体识别的过程中充分利用实体类别的信息，从而缓解了实体识别的缺少标注信息的问题。

3.2.2 一种基于对比学习的实体链接模型

小样本实体链接的目标是基于目标域仅有的少量标记样本，将文本中的实体指称与知识库中真实实体进行链接。为了达到这个目标，本论文提出了一种新的实体链接方法，该方法具有实体类别感知的特性。

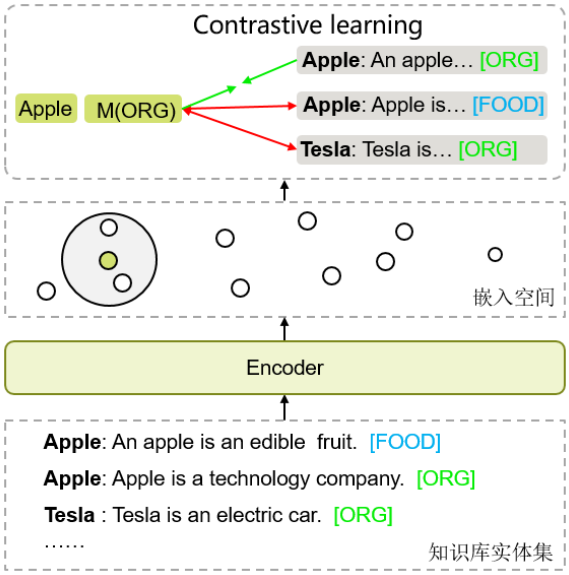

图 3-2 一种基于对比学习的实体链接模型

如图 3-2 所示，本方法为知识库中的每个实体分配了一个实体类别，并将实体类别和实体描述作为该实体的提示信息。其中实体的类别与命名实体识别模块的类别一致，由 Tedeschi 提供 Wikipedia2NER-mapping 数据集支持[36]，获取知识库 Wikipedia 知识库中实体到实体类别的映射。实体的提示信息由 Wikipedia 自身提供。

接下来将命名实体识别模块的实体提及与实体消歧模块的知识库实体映射到同一嵌入空间，从而将文本中的实体指称与带有提示信息的真实实体进行对比学习，通过优化实体链接模型来实现更准确的链接。

具体而言，我们将知识库中的实体、实体对应的描述，以及该实体的类别输入到编码

器中，获得实体的上下文表示。并将命名实体识别模块的实体提及与之映射到同一嵌入空间，然后，我们使用搜索算法在知识库中查找与每个实体提及最相近的几个实体，将它们作为候选实体。这样做是为了缩小实体链接的范围，将注意力集中在与实体提及最相关的候选实体上。最后，我们使用对比学习的方法，将候选实体与实体提及进行比较学习。通过比较它们的上下文表示，我们的模型可以学习如何从相似的候选实体中正确地区分出实体提及所指的真实实体。通过这种联合训练的方式，我们的模型可以在实体链接任务中获得更好的效果。

综上，通过将实体类别纳入考虑，并将其与实体描述结合起来，实体类别感知使得模型能够更好地理解实体指称的上下文和语义信息，并与知识库中的真实实体进行匹配。我们的方法可以解决实体缺少类别信息及丰富描述信息的问题，从而获取更加准确和一致的实体链接结果。

3.2.3 一种命名实体识别和实体链接联合训练模型

根据上述对命名实体识别和实体链接两个任务的研究，本课题进一步提出一种用于实体识别和实体消歧的联合训练方法。联合训练是一种将多个相关任务的训练过程结合起来的方法，以提高模型的整体性能和一致性。在命名实体识别和实体链接的联合训练中，这两个任务被同时考虑和训练，优化命名实体识别性能的同时提升实体链接的能力。如图 3-3 所示，图(a)为基于对比学习的命名实体识别模型，输入为自然语言文本，输出为文本中的实体指称及对应实体类别；图(b)为基于对比学习的实体链接模型，输入为实体指称，输出为该实体指称链接到知识库中的真实实体。

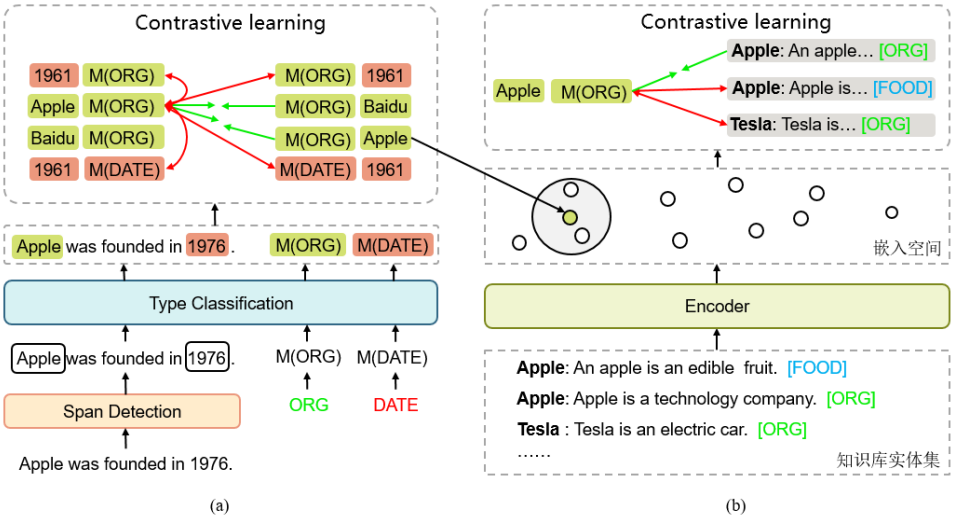

图 3-3 联合训练模型图

在联合训练中，模型的输入是原始文本。模型的输出是预测的命名实体类别和对应的实体链接结果。联合训练的步骤如下：

(1) 对训练数据集合中的文本进行命名实体识别，得到实体指称；

(2) 基于上一步的实体指称，在与知识库中的实体进行映射，得到一个候选实体集合；

(3) 将候选实体与实体指称输入实体消歧模块进行实体预测，得到目标实体，并根据结果调整所述实体消歧模型的参数;

(4) 通过实体识别模型中增设的实体注意力机制层，将实体消歧结果反馈给实体识别模型，并根据实体消歧调整实体识别模型的参数;
(5) 调整实体识别和实体消歧两个模型的参数，并计算两者的损失函数之和，用于实体识别和实体消歧的联合训练。

训练过程中，模型通过最小化命名实体识别和实体链接的损失函数来学习参数，以使得模型能够在两个任务上达到较好的性能。

在命名实体识别和实体链接的联合训练中，本课题设计了一个综合的损失函数来同时考虑这两个任务的训练目标。假设有一个样本的输入文本序列为 X，对应的命名实体识别标签序列为Y_{NER}，实体链接的目标序列为Y_{EL}。模型的输出分别为命名实体识别的预测序列和实体链接预测序列，记为 $\hat{Y}_{NER}$ 和 $\hat{Y}_{EL}$ 。

命名实体识别任务损失函数可以使用交叉熵损失函数来度量 NER 预测结果与真实标签的差异。假设 NER 任务的标签空间为C_{NER}，则损失函数可以定义为：

$$L_{NER} = -\sum_{i=1}^{T}\sum_{c\in C_{NER}} Y_{NER}[i][c] * \log\left(\hat{Y}_{NER}[i][c]\right) \tag{9}$$

实体链接损失函数：实体链接任务通常涉及到实体链接候选集的生成和候选集中每个实体的打分。可以使用最大似然估计损失函数来度量实体链接预测结果与真实目标的差异。假设实体链接的候选集大小为 K，损失函数可以定义为：

$$L_{EL} = -\sum_{i=1}^{T}\sum_{k=1}^{K} Y_{EL}[i][k] * log(\hat{Y}_{EL}[i][k]) \tag{10}$$

最后为了综合命名实体识别和实体链接任务的损失，可以对两个损失函数进行加权求和，得到联合训练的总损失函数：

$$L_{joint} = \alpha * L_{NER} + \beta * L_{EL} \tag{11}$$

其中，α 和 β 是损失函数的权重系数，用于平衡两个任务的重要性。根据具体的任务需求和数据集特点，可以调整这些权重系数来使得两个任务对模型训练的影响程度相对均衡。

通过联合训练命名实体识别和实体链接，模型可以更好地理解命名实体的语义和上下文信息，并将其链接到知识库中的对应实体。这种联合训练模型能够缓解端到端模型的错误传播问题，从而提供更准确和一致的命名实体识别和实体链接结果。

3.3 实验设计

3.3.1 命名实体识别模型的实验设计

本模型的性能评估指标为常用的精确率（Precision）、召回率（Recall）、 F-1 值（F-1 score）。

本模型使用的数据集为命名实体识别领域常用的 Few-NERD[37] 、CoNLL’03[38]、OntoNotes[39]，以及工程项目中自建的组织机构数据集。其中开放域的数据集统计信息见

表 3-1.

表 3-1 数据集统计信息

数据集	领域/来源	类别数	句子数	实体数
Few-NERD	维基百科	66	188, 200	491, 700
CoNLL'03	新闻	4	20, 700	35, 100
OntoNotes	通用领域	18	76, 700	104, 200

本模型的使用的对比基线模型为 ProtoNet[21]、NNShot[40]、StructShot[40]、CONTaiNER[22]、MAML-ProtoNet[25]，其中 ProtoNet、NNShot、StructShot 为基于原型网络的小样本命名实体识别模型，CONTaiNER 为基于对比学习的小样本命名实体识别模型，MAML-ProtoNet 为两阶段的命名实体识别模型。通过与上述模型进行对比，来验证本模型的有效性。

除了对比模型外，本模型还将通过消融实验验证对比学习模块设计的性能和作用，只使用两阶段模型且不含类别描述信息的模型、使用两阶段模型且含有对比学习模块引入类别描述信息的模型。通过上述评价指标和数据集验证对比学习模块的有效性。

为了验证引入类别描述信息的有效性，本课题设置了不同长度的自然语言来描述类别，从而探究提示长度对模型的影响，以及探查最合适的描述长度。

3.3.2 实体链接模型的实验设计

本模型的性能评估指标为常用的精确率（Precision）、召回率（Recall）、 F-1 值（F-1 score）。

本模型使用的知识库为 Wikipedia。

本模型使用的训练数据集为的 AIDA　CoNLL-YAGO[41]，这是实体链接常用的公共数据集。评估数据集为三个较小数据集：MSNBC[42]，AQUAINT[42]， ACE2004[42]，和两个较大数据集：WNED-WIKI[42]和 WNED-CWEB[43]，以及工程项目中自建的组织机构数据集。其中开放域的数据集统计信息见表 3-2.

表 3-2 数据集统计信息

数据集	领域/来源	文章数	可链接实体数
AIDA-YAGO-CoNLL	新闻	388	27, 817
MSNBC	新闻	20	656
AQUAINT	新闻	50	727
ACE2004	新闻	25	257

本模型的使用的对比基线模型为 Partalidou 等[29]、Tang 等[32]、Yao 等[31]，其中 Partalidou 等和 Tang 为基于 Cross-Encoder 的模型，Yao 为基于 BERT 的模型。通过与上述模型进行对比，来验证本模型的有效性。

除了对比模型外，本模型还将通过消融实验验证对比学习模块设计的性能和作用，只使用不含类别描述信息的模型、使用含有对比学习模块引入类别描述信息的模型。通过上述评价指标和数据集验证对比学习模块的有效性。

3.3.3 联合训练模型的实验设计

本联合模型的评估指标和数据集和实体链接模型保持一致。

本联合模型使用的对比基线模型为 GENRE [33]、Rychlikowski [34]的模型作为对比模型。其中 GENRE [33]是第一个通过生成实体的唯一名称来检索实体的系统，同时检测实体和链接实体；Rychlikowski [34] 利用了大量的非结构化和结构化文档联合执行命名实体识别和实体消歧。通过与上述模型进行对比，来验证本模型的有效性。

除了和其他模型进行对比外，本模型还将通过消融实验和自身进行对比，验证联合训练比独立训练更有效。先实体识别再实体消歧的模型，和两任务联合训练的模型进行对比。通过前面的评价指标和数据集验证联合模型的有效性。

四、可行性分析

本课题的研究目的是通过研究一种基于对比学习的小样本命名实体识别方法和一种基于对比学习的小样本实体链接方法，并将两者进行联合训练，从而同时考虑命名实体识别和实体链接两个任务，从而减弱端到端模型引起的错误传播的缺陷。

在研究内容中，本课题构建了一种基于对比学习的小样本命名实体识别模型：通过命名实体类别的语义表示构建类别软原型，并将其为实体指称的提示进行对比学习，从而缓解缺少标记数据的问题；构建了一种基于对比学习的实体链接模型：为知识库中的实体分配一个实体类别，并将实体类别和实体描述作为该实体的提示，与文本中的实体指称进行对比学习，从而解决了模型忽视丰富的类别描述信息的问题；构建了一种命名实体识别和实体链接联合训练模型：同时考虑命名实体识别和实体链接两个任务，以减弱端到端模型引起的错误传播。近几年已经有多篇高质量论文采取了对比学习以及引入语义信息的技术，因此本课题也沿用了这一主流思路，充分利用这些方法在信息抽取邻域的优势。

本课题提出前，本人已经对命名实体识别方法和实体链接方法进行了系统的学习和梳理， 对相关的对比学习和上下文表示算法和研究有了大致的了解，并学习了本课题实施方案中的相关技术基础。

综上所述，本课题的研究能够在计划时间内达到预期目标。

五、预期的成果

在命名实体识别模型中，我们期望缓解小样本命名实体识别的缺少标注数据的问题，

模型通过引入实体类别信息，提供更多的标签信息和上下文约束，从而增强命名实体识别模型的鲁棒性，模型可以学习到更准确的实体边界和实体类别信息，从而在识别命名实体时表现更好。

在实体链接模型中，我们期望解决忽视实体类别信息和丰富描述信息的问题，模型可以更好地理解实体指称和实体描述之间的关系，通过对比学习和搜索算法的辅助，减少错误的实体链接。模型可以从候选实体中选择最相似的实体，并通过上下文匹配和语义相似性判断来区分正确的实体链接。

在命名实体识别和实体链接的联合训练中，我们期望减轻端到端实体链接模型的错误传播问题：联合训练可以利用实体识别和实体链接两个任务之间的相互关联，从而提高实体链接的准确性。通过共享模型参数和特征表示，实体链接可以受益于实体识别的结果，并借助上下文信息更好地将实体指称链接到知识库中的真实实体。

总之，预期的成果是在联合训练中获得更准确和鲁棒的命名实体识别和实体链接模型。这将有助于提高信息抽取和语义理解任务的性能，并在实际应用中更好地应对命名实体链接的挑战。

参考文献

[1] Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni.2021.Autoregressive entity retrieval.

[2] Etzioni et al., “Unsupervised named-entity extraction from the web: An experimental study,” Artif. Intell.,vol.165,no.1,pp.91–134,2005.

[3] K. Humphreys et al., “University of Sheffield: Description of the laSIE-II system as used for MUC-7,” in Proc.7th Message Understanding Conf.,1998, pp.1–20.

[4] G. Krupka and K.IsoQuest, “Description of the nerOWL extractor system as used for MUC-7,” in Proc.7th Message Understanding Conf.,2005, pp.21–28.

[5] S.R.Eddy,“Hidden markov models,” Curr. Opinion Structural Biol., vol. 6, no. 3, pp. 361 – 365, 1996.

[6] M.A. Hearst, S.T. Dumais, E.Osuna, J.Platt, and B.Scholkopf, “Support vector machines,” IEEE Intell.Syst. Their Appl., vol.13, no.4, pp. 18–28, Jul./Aug.1998.

[7] J.D.Lafferty, A.McCallum, and F.C.N.Pereira, “Conditional random fields: Probabilistic models for segmenting and labeling sequence data, ”in Proc.18th Int.Conf.Mach.Learn.,2001, pp.282–289.

[8] Huang, Zhiheng et al. “Bidirectional LSTM-CRF Models for Sequence Tagging.” *ArXiv* abs/1508.01991 (2015): n. pag.

[9] Souza, Fábio, Rodrigo Nogueira, and Roberto Lotufo. "Portuguese named entity recognition using BERT-CRF." arXiv preprint arXiv:1909.10649 (2019).

[10] Cui, Leyang, et al. "Template-based named entity recognition using BART." arXiv preprint

arXiv:2106.01760 (2021).
[11] Leyang Cui,Yu Wu,Jian Liu,Sen Yang,and Yue Zhang.2021.Template-based named entity recog nition using BART.In Findings of the Association for Computational Linguistics:ACL-IJCNLP 2021,pages 1835–1845,Online.Association for Computa tional Linguistics.
[12] Yucheng Huang,Kai He,Yige Wang,Xianli Zhang,Tieliang Gong,Rui Mao,and Chen Li.2022.COP NER:Contrastive learning with prompt guiding for few-shot named entity recognition.In Proceedings of the 29th International Conference on Computa tional Linguistics,pages 2515–2527,Gyeongju,Re public of Korea.International Committee on Com putational Linguistics.
[13] Bin Ji,Shasha Li,Shaoduo Gan,Jie Yu,Jun Ma,Hui jun Liu,and Jing Yang.2022.Few-shot named entity recognition with entity-level prototypical network enhanced by dispersedly distributed proto types.In Proceedings of the 29th International Con ference on Computational Linguistics,pages 1842–1854,Gyeongju,Republic of Korea.International Committee on Computational Linguistics.
[14] Sarkar Snigdha Sarathi Das, Arzoo Katiyar, Rebecca Passonneau, and Rui Zhang. 2022. CONTaiNER: Few-shot named entity recognition via contrastive learning. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 6338–6353, Dublin, Ireland. Association for Computational Linguistics.
[15] Yi Yang and Arzoo Katiyar. 2020. Simple and effective few-shot named entity recognition with structured nearest neighbor learning. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 6365–6375, Online. Association for Computational Linguistics.
[16] Xiao Liu,Yanan Zheng,Zhengxiao Du,Ming Ding,Yujie Qian,Zhilin Yang,and Jie Tang.2021b.Gpt understands,too.arXiv preprint arXiv:2103.10385.
[17] Leyang Cui,Yu Wu,Jian Liu,Sen Yang,and Yue Zhang.2021.Template-based named entity recognition us ing bart.arXiv e-prints,pages arXiv–2106.
[18] Chen Y, Zheng Y, Yang Z. Prompt-Based Metric Learning for Few-Shot NER[J]. arXiv preprint arXiv:2211.04337, 2022.
[19] Ashok D, Lipton Z C. PromptNER: Prompting For Named Entity Recognition[J]. arXiv preprint arXiv:2305.15444, 2023.
[20] Morteza Ziyadi,Yuting Sun,Abhishek Goswami,Jade Huang,and Weizhu Chen.2020.Example-based named entity recognition.CoRR,abs/2008.10570.
[21] Jake Snell,Kevin Swersky,and Richard S.Zemel.2017.Prototypical networks for few-shot learning.In Ad vances in Neural Information Processing Systems 30:Annual Conference on Neural Information Process ing Systems 2017,December 4-9,2017,Long Beach,CA,USA,pages 4077–4087.
[22] Sarkar Snigdha Sarathi Das, Arzoo Katiyar, Rebecca Passonneau, and Rui Zhang. 2022. CONTaiNER: Few-Shot Named Entity Recognition via Contrastive Learning. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 6338–6353,

[23] Peiyi Wang,Runxin Xu,Tianyu Liu,Qingyu Zhou,Yunbo Cao,Baobao Chang,and Zhifang Sui.2021.An enhanced span-based decomposition method for few-shot sequence labeling.In NAACL,pages 5012–5024.

[24] Yongliang Shen,Xinyin Ma,Zeqi Tan,Shuai Zhang,Wen Wang,and Weiming Lu.2021.Locate and label:A two-stage identifier for nested named entity recognition.In ACL,pages 2782–2794.

[25] Ma T, Jiang H, Wu Q, et al. Decomposed Meta-Learning for Few-Shot Named Entity Recognition[J]. arXiv preprint arXiv:2204.05751, 2022.

[26] Chelsea Finn,Pieter Abbeel,and Sergey Levine.2017.Model-agnostic meta-learning for fast adaptation of deep networks.In Proceedings of the 34th International Conference on Machine Learning,ICML 2017,Sydney,NSW,Australia,6-11 August 2017,volume 70 of Proceedings of Machine Learning Research,pages 1126–1135.PMLR.

[27] Logeswaran L, Chang M W, Lee K, et al. Zero-shot Entity Linking by Reading Entity Descriptions[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 3449-3460.

[28] Wu L, Petroni F, Josifoski M, et al. Scalable Zero-shot Entity Linking with Dense Entity Retrieval[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020: 6397-6407.

[29] Partalidou E, Christou D, Tsoumakas G. Improving Zero-Shot Entity Retrieval through Effective Dense Representations[J]. 2021:28-37.

[30] Ristoski P, Lin Z, Zhou Q. KG-ZESHEL: Knowledge Graph-Enhanced Zero-shot Entity Linking[C]//Proceedings of the 11th on Knowledge Capture Conference. 2021: 49-56.

[31] Yao Z, Cao L, Pan H. Zero-shot Entity Linking with Efficient Long Range Sequence Modeling[C]//Findings of the Association for Computational Linguistics: EMNLP. 2020: 2517-2522.

[32] Tang H, Sun X, Jin B, et al. A Bidirectional Multi-paragraph Reading Model for Zeroshot entity linking[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2021, 35(15): 13889-13897.

[33] De Cao, Nicola, et al. "Autoregressive entity retrieval." arXiv preprint arXiv:2010.00904 (2020).

[34] Rychlikowski, Paweł, et al. "Named entity recognition and linking augmented with large-scale structured data." arXiv preprint arXiv:2104.13456 (2021).

[35] Tedeschi, Simone, et al. "Named Entity Recognition for Entity Linking: What works and what's next." Findings of the Association for Computational Linguistics: EMNLP 2021. 2021.

[36] Chelsea Finn,Pieter Abbeel,and Sergey Levine.2017.Model-agnostic meta-learning for fast adaptation of deep networks.In Proceedings of the 34th International Conference on Machine Learning,ICML 2017,Sydney,NSW,Australia,6-11 August 2017,volume 70 of Proceedings of Machine Learning Research,pages 1126–1135.PMLR.

[37] Ning Ding,Guangwei Xu,Yulin Chen,Xiaobin Wang,Xu Han,Pengjun Xie,Haitao Zheng,and

Zhiyuan Liu.2021.Few-NERD:A few-shot named entity recognition dataset.In Proceedings of the 59th An nual Meeting of the Association for Computational Linguistics and the 11th International Joint Confer ence on Natural Language Processing (Volume 1:Long Papers),pages 3198–3213,Online.Association for Computational Linguistics

[38] Erik F.Tjong Kim Sang and Fien De Meulder.2003.Introduction to the CoNLL-2003 shared task:Language-independent named entity recognition.In Proceedings of the Seventh Conference on Natu ral Language Learning at HLT-NAACL 2003,pages 142–147.

[39] Sameer Pradhan,Alessandro Moschitti,Nianwen Xue,Hwee Tou Ng,Anders Björkelund,Olga Uryupina,Yuchen Zhang,and Zhi Zhong.2013.Towards ro bust linguistic analysis using OntoNotes.In Proceed ings of the Seventeenth Conference on Computational Natural Language Learning,pages 143–152,Sofia,Bulgaria.Association for Computational Linguistics.

[40] Yi Yang and Arzoo Katiyar.2020.Simple and effective few-shot named entity recognition with structured nearest neighbor learning.In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP),pages 6365–6375,Online.Association for Computational Linguistics.

[41] Johannes Hoffart,Mohamed Amir Yosef,Ilaria Bor dino,Hagen Fürstenau,Manfred Pinkal,Marc Span iol,Bilyana Taneva,Stefan Thater,and Gerhard Weikum.2011.Robust disambiguation of named en tities in text.In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Process ing,pages 782–792,Edinburgh,Scotland,UK.Asso ciation for Computational Linguistics.

[42] Zhaochen Guo and Denilson Barbosa.2017.Robust named entity disambiguation with random walks.Semantic Web,9:1–21

[43] Evgeniy Gabrilovich,Michael Ringgaard,and Amar nag Subramanya.2013.Facc1:Freebase annotation of clueweb corpora,version 1 (release date 2013-06-26,format version 1,correction level 0).

研究生签名

年　月　日

（注：本页不够可附页）

二、学位论文工作实施进度与安排

起讫 日期	工 作 内 容 和 要 求	备 注
2021.9-2022.8	阅读相关文献，调研现有工作	
2022.9-2022.11	确定研究方向，设计算法框架	
2022.12-2023.2	进行理论分析与验证，设计可行性方案	
2023.03-2023.06	处理数据集，撰写开题报告	
2023.07-2023.10	制定实现方案，搭建模型	
2023.11-2023.01	评估模型性能并进行改进	
2024.02-2024.04	撰写毕业论文初稿,撰写小论文	
2024.05-2024.06	修改毕业论文，小论文投稿，准备毕业论文答辩	

指导教师对开题报告的综合意见	指导教师（签名） 年 月 日

<table>
<tr><td>开题报告审议情况记录</td><td>1、审议小组成员（硕士一般 3-5 人；博士 5－7 人）：
组长：
成员：
2、审议小组意见记录

3、投票表决结果
审议小组出席 _____ 人；通过 _____ 人；不通过 _____ 人。
开题报告质量_____（优、良、中、通过）
4、审议小组组长（签名）
审议小组成员（签名）

年 月 日</td></tr>
<tr><td colspan="2">院（系、所）意见：

院（系、所）负责人签名（或印章）
年 月 日</td></tr>
<tr><td colspan="2">备注：</td></tr>
</table>