小样本命名实体识别方法研究

开题报告

计算机科学与技术学院 学硕2102 2212523 胡怡淳

指导教师 卢婷

1. 研究背景及意义

命名实体识别旨在从非结构化的文本中识别出有意义的实体，并将其分类为一组预定义的实体类型，如人名、机构名、地名、医疗代码、日期、数量等。命名实体通常含有丰富的语义信息，并与语料中的关键信息有着紧密的联系，因此命名实体识别是许多下游任务的重要组成部分，如关系抽取、事件抽取、问答、知识图谱构建等。命名实体识别的结果会对后续任务的性能产生重大影响，可能会导致误差传递等问题，因此命名实体识别一直以来都是热门的研究课题。

尽管如今基于深度学习的命名实体识别方法已经取得了非常优异的性能，但是这些方法通常都需要大量的人工标注语料数据才能达到足够好的效果。对于那些标注数据就比较少的领域，比如金融、餐饮、影视等，从头开始构建一个高质量的标注数据集需要花费巨大的代价，需要丰富的领域内知识以及大量的人力和时间成本。因此，现在有越来越多的研究聚焦于更具挑战性的小样本命名实体识别任务。在小样本命名实体识别中，模型只通过非常少的标注数据来学习如何识别和分类命名实体，这样大幅降低了数据的标注成本，也更加符合现实当中的实际情况。

元学习作为小样本学习任务常用的解决方案，在小样本命名实体识别任务中也取得了不错的效果。其根本思想是通过原型网络将每个类别的数据映射到一个相同的空间中，并取这些向量的均值作为该类型的原型来代表这个类型，通过测试数据与各原型之间的距离来决定最终的类别。

迁移学习方法通常是先在一个拥有较多标注数据的高资源领域中进行预训练，学习一般的表征，然后再将其迁移到标注数据较少的低资源领域，在少量数据上进行微调，使模型适应新的任务。由于预训练时的源领域和微调时的目标领域各自包含的实体类型通常并不一致，需要在微调时舍弃原来的分类层，并添加新的分类层重新训练，导致分类层的训练并不充分，最终模型的性能不够理想。针对这一问题，许多研究提出对类别信息进行统一的建模，使得模型能够适应不同类别的情况。

随着大规模预训练语言模型的发展，许多任务在预训练模型上加以微调就能达到不错的效果。但一般的微调方法会引入大量新的参数。为了不引入新的参数，同时充分利用预训练语言模型的先验知识，提出了基于提示学习的方法，借鉴了遮蔽语言模型的方式，将下游任务转化为完形填空式的问题。具体来说就是为任务构建提示模板，然后通过语言模型对模板进行补全，从而得到预测结果。

1. 研究内容

2.1 类别描述信息的自动生成

为了使模型能适应不同的实体类别，提出了对类别信息进行统一建模的方法，而其中的类别描述信息大多数使用的是类名，类名作为最基本的类别描述，获取容易但包含的信息量也较少。也有一些研究使用了较为详细的描述信息，包含精确的定义以及一些示例，但这种描述通常来自于数据集本身提供的注释指南，或者是一些专业可靠的百科信息，并不适用于所有情况。我们提出一种能从小样本数据中自动生成类别描述的方法，既可以为模型提供更多的先验知识，提升模型的性能，又能适用于大多数的情况。

2.2 实体标签之间依赖关系的概括与推广

条件随机场已被证明对于命名实体识别任务有非常大的提升，但在小样本上训练它是十分困难的，所以有研究提出了利用抽象转移矩阵来概括一些抽象标签之间的依赖关系，再将其推广到任意目标领域，以此来替代原先耗时的条件随机场的训练。而源领域和目标领域之间实体分布的存在差异性，以及其中的概括与推广的方法的不同可能会对性能产生影响。我们对现有方法进行了改进。

2.3 小样本命名实体识别的数据增强

数据增强一直以来都是解决资源稀缺问题最主要的手段，现有研究针对命名实体识别问题的数据增强方法主要还是针对几百几千条数据的低资源情况，对于小样本这种每个类只有几个样本的极低资源的情况并没有非常合适的解决方案。我们提出一种跨领域数据增强方法，利用数据丰富领域的数据为小样本数据进行增强。

1. 研究方法（技术路线）

3.1 类别描述信息的自动生成

现有方法中的类别描述信息基本上来自于外部数据或者是数据集本身提供的，我们希望能仅从数据本身提取可靠的类别描述。由于人在学习一种新的事物时通常是从例子开始，我们也考虑根据一些例子来构造类别的描述。具体来说，我们先从已有的数据中，将一个实体类别概括成几个例子，比如将地点类型的实体概括为城市、省份和国家或者是上海、北京和中国等等，然后根据这些例子构建出一个类别描述。

3.2 实体标签之间依赖关系的概括与推广

首先根据现有数据，概括得到一个体现抽象标签之间依赖关系的抽象转移矩阵，再将其推广到目标领域。如图3-1所示，图中的方法将源领域的标签概括为了B、I和O三种，并统计了三种标签与相同类型的B标签和I标签（sB和sI）、不相同类型的B标签和I标签（dB和dI）以及O标签之间转移的概率，得到抽象转移矩阵。而在将抽象转移矩阵推广到目标领域时，该方法采用的是直接将相应的抽象转移概率填入对应位置，比如B-time和I-time之间的转移对应B和sI之间的转移，概率为0.5。

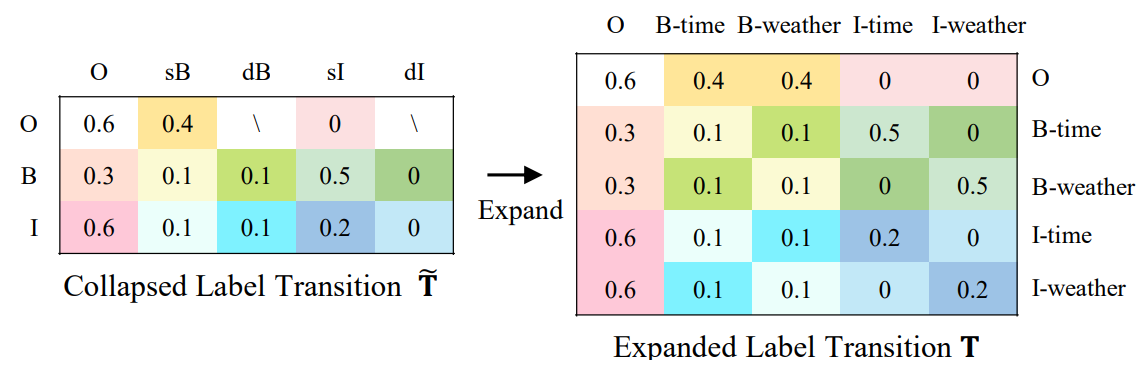


图3-1 实体标签之间依赖关系的概括与推广

3.3 命名实体识别的数据增强

命名实体识别不同于其他自然语言处理任务，它是一个字符级的分类任务，对局部更加敏感，因此有些数据增强的方法并不适用。

现有的命名实体识别数据增强方法主要是传统的自然语言处理的增强方法以及基于模型的方法。传统方法主要包括：同标签词替换；同义词替换；根据外部实体词典或训练集中的相同类型实体进行实体替换；对于句子中不重要的部分进行错拼、删除、替换；根据语义对句子进行划分，然后对切分后的片段进行重新排序等。基于模型的方法首先会将标签合并入输入句子中，实现标签序列化，如图3-2所示，然后遮蔽一部分实体，利用模型预测遮蔽部分词语的能力来进行数据增强。另外还有一些基于MixUp的方法，跨领域增强的方法，多种增强方式相结合的方法等。

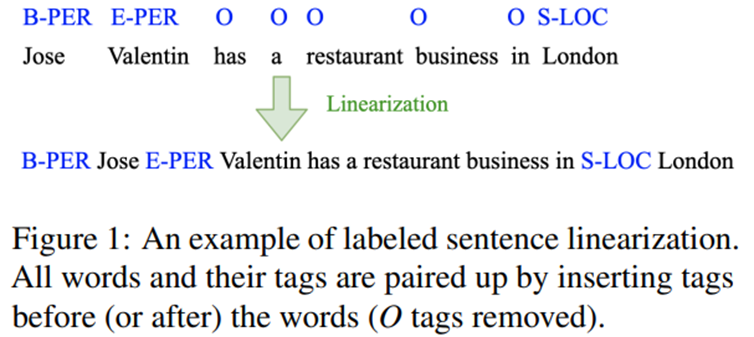


图3-2 标签序列化

1. 拟采取的研究方案

本课题的研究内容主要分为两个部分：基于迁移学习的小样本命名实体识别，以及小样本命名实体识别的数据增强。

基于迁移学习的小样本命名实体识别拟采取的研究方案如下：

（1）设计一种从小样本中自动生成类别描述的方法，并利用该方法获取类别的描述信息；

（2）使用类别描述和输入的表征之间的点积来度量其相似度，根据相似度预测单词所属的实体类别；

（3）改进实体标签之间依赖关系的概括与推广的方法，从训练集中概括抽象转移矩阵并将其推广到目标领域；

（4）对比尝试多种构造类别描述的方法，最后选择采用效果最好的方法；

（5）选取与参考论文相同的数据集，对比在相同实验设置的情况下，不同方法表现的差异。

小样本命名实体识别的数据增强拟采取的研究方案如下：

（1）设计一种跨领域数据增强的方法，利用源领域数据来增强目标领域的小样本数据；

（2）迁移学习与数据增强相结合，先在源领域进行预训练，再在目标领域的增强数据上进行微调；

（3）选取与参考论文相同的数据集，对比在相同实验设置的情况下，不同方法表现的差异。

1. 预期目标

（1）设计并实现一种能从小样本数据中自动生成类别描述的方法，其生成的类别描述能为模型带来更多有用的信息和先验知识，最终能对模型的性能有所提升。

（2）改进现有的实体标签之间依赖关系的概括与推广的方法，为模型的性能带来更进一步的提升。

（3）设计并实现一种跨领域数据增强的方法，利用资源丰富领域的数据来增强目标领域的小样本数据，提升小样本命名实体识别的效果。

1. 难点及存在的问题

（1）从小样本数据中自动生成类别描述。小样本本身包含的信息量就少，如何根据少量的样本，在不借助外部百科、字典等的情况下，自动生成出可靠的类别描述，是本研究的一个难点。

（2）小样本数据集的构建。由于命名实体识别任务不同于其他句子级别的自然语言处理任务，在一句句子中可能会出现多个以及多类实体，因此最终采样得到的小样本数据集中每个实体类别的样本的数量可能会和预期有比较大的出入。如何设计采样方法，使得最终得到的数据集能够更加贴合我们所期望的小样本设置也是一个问题。

（3）跨领域数据增强。跨领域数据增强的现有研究较少，如何能够充分利用数据丰富的高资源领域的数据为低资源领域生成有用的增强数据，并且该方法能适用于小样本命名实体识别，这也是本课题的一个难点。

1. 工作计划安排

|  |  |
| --- | --- |
| **时间阶段** | **具体工作** |
| 2023.05~2023.06 | 查阅资料，阅读文献，复现代码。 |
| 2023.07~2023.09 | 实现所提出的方法，并选取部分数据集进行实验，验证方法的性能。 |
| 2023.10~2023.12 | 根据实验结果调整方法的细节，并进行大规模数据集的实验。 |
| 2024.01~2024.02 | 与现有方法进行对比，进一步优化算法。 |
| 2024.03~2024.05 | 撰写毕业论文，准备毕业答辩。 |