NER 论文研究方向调研

Author: Jiaojiao Ye

Date: 17.03.2022

介绍

命名实体识别(Named Entity Recognition，NER)一直是NLP中最主流，也是最基础最重要的任务之一。 本文重点关注最新研究方向，梳理了2021年ACL会议上的19篇定位命名实体识别的论文对其相关问题进行分类，介绍各篇论文的大致思路，对现阶段NER领域进行总结。

整理分类

DIRECTION	PAPER	INSTITUTION	LINK
数据问题	MECT: Multi-Metadata Embedding based Cross-Transformer for Chinese Named Entity Recognition	Jiangnan University,University of Surrey	paper
	Named Entity Recognition with Small Strongly Labeled and Large Weakly Labeled Data	Georgia Institute of Technology, Amazon.com Inc	paper
	FEW-NERD: A Few-shot Named Entity Recognition Dataset	Tsinghua University, Alibaba Group	paper
	Leveraging Type Descriptions for Zero-shot Named Entity Recognition and Classification	University of Cambridge, ASAPP	paper
	Learning from Miscellaneous Other-Class Words for Few-shot Named Entity Recognition	Tsinghua University, JOYY Inc	paper
	Subsequence Based Deep Active Learning for Named Entity Recognition	University College London, University of Cambridge	paper
弱监督学习	Weakly Supervised Named Entity Tagging with Learnable Logical Rules	University of California, San Diego, Bosch Research North America	paper
	BERTifying the Hidden Markov Model for Multi-Source Weakly Supervised Named Entity Recognition	Georgia Institute of Technology, Mohamed bin Zayed University of Artificial Intelligence	paper

DIRECTION	PAPER	INSTITUTION	LINK
嵌套实体识别任务	Locate and Label: A Two-stage Identifier for Nested Named Entity Recognition	Zhejiang University, University of Science and Technology of China	paper
	Nested Named Entity Recognition via Explicitly Excluding the Influence of the Best Path	National Institute of Information and Communications Technology (NICT), Nara Institute of Science and Technology (NAIST)	paper
不连续的实体识别	Discontinuous Named Entity Recognition as Maximal Clique Discovery	Chinese Academy of Sciences, University of Chinese Academy of Sciences	paper
	A Span-Based Model for Joint Overlapped and Discontinuous Named Entity Recognition	Wuhan University, Tianjin University	paper
医疗领域-疾病实体的识别与规范化	A Neural Transition-based Joint Model for Disease Named Entity Recognition and Normalization	PAII Inc., Ping An Technology	paper
	An End-to-End Progressive Multi-Task Learning Framework for Medical Named Entity Recognition and Normalization	Nankai University, Tianjin Key Laboratory of Network and Data Security Technology	paper
远程监督	De-biasing Distantly Supervised Named Entity Recognition via Causal	Chinese Information Processing Laboratory, Chinese Academy of Sciences, University of Chinese Academy of Sciences	paper
范围预测	SPANNER: Named Entity Re-/Recognition as Span Prediction	Fudan University, Carnegie Mellon University	paper
众包学习	Crowdsourcing Learning as Domain Adaptation: A Case Study on Named Entity Recognition	Tianjin University	paper
文档级别NER	Improving Named Entity Recognition by External Context Retrieving and Cooperative Learning	ShanghaiTech University, Shanghai Engineering Research Center of Intelligent Vision and Imaging	paper
辅助学习任务	Modularized Interaction Network for Named Entity Recognition	Beijing Institute of Technology, Nanyang Technological University	paper

主要思路

1. **MECT: Multi-Metadata Embedding based Cross-Transformer for Chinese Named Entity Recognition**

之前在提升中文实体识别上，主要从减少分词错误，增加汉语词的语义和边界信息等方向入手，但这些方法都没有考虑汉字的结构信息，比如"鸟类"都会带“鸟”结构字，如下表。针对上述问题，文本主要提出利用Cross-Transformer的结构，学习更多的元信息(Metadata)来提升中文实体识别效果。这里的元信息，论文是从中文中的象形结构信息来定义的。简单来说，作者就是将汉字从象形的角度拆分更多的元信息出来，嵌入实体识别任务中。

2. **Named Entity Recognition with Small Strongly Labeled and Large Weakly Labeled Data**

本篇也是解决少量样本学习问题的，以往方法主要采用先用少量强标注(人工标注)的数据生成大量弱标注的数据，然后采用弱监督的方式进行实体识别，但这类方法会产生的很大的噪声数据，对任务没多少提升，甚至损害模型。对此，作者提出一个多阶段学习框架：第一步进行弱标签补充；第二步降低噪声，生成强标注数据；第三步在强标注数据进行微调(fine-tuning)识别。

3. **FEW-NERD: A Few-shot Named Entity Recognition Dataset**

本篇是作者发布一个面向少样本学习(few-shot)的实体识别数据集：FEW-NERD，该数据集由8个粗粒度实体类型和66个细粒度实体类型组成的层次结构。数据集包括18万+的句子，49万+的实体。

4. **Leveraging Type Descriptions for Zero-shot Named Entity Recognition and Classification**

从论文题目上可以看出，本篇要解决的问题是：利用类型描述来解决少样本实体识别与实体分类问题。这里的实体识别与分类是一个连续的任务：给定一个文本和实体类型集合，即要识别文本中的实体并判断这个实体属于哪类(如地点，人物等)，这里的实体类型集合可能是变动。本文是要解决少样本来做此任务，具体解决方案是利用实体类型的描述信息来表征成实体类型向量，然后与实体向量进行匹配学习。

5. **Learning from Miscellaneous Other-Class Words for Few-shot Named Entity Recognition**

同样是NER中Few-shot问题，认为之前的方法对标签“O”类的信息学习不够充分，导致更容易让模型过拟合。针对该问题，作者提出了一个新的模型MUCO(Mining Undefined Classes from Other-class)，可以自动从其他类(Other-class)中学习到不同的未定义类，以提高NER少样本学习能力。

6. Subsequence Based Deep Active Learning for Named Entity Recognition

本篇是利用主动学习(Active Learning)来做NER任务，不同的是，作者在主动学习框架上采用子序列(Subsequence)的学习方式，来解决之前方法存在的问题：没有利用语言的连续性和每个实例异构的不确定性，需要对整个句子进行标记。作者提出的方法，利用人工智能算法查询句子中的子序列，并将它们的标签传播到其他句子，解决上述存在的问题。

7. **Weakly Supervised Named Entity Tagging with Learnable Logical Rules**

本篇也是来解决NER中数据标注问题，采用弱监督学习的方式(Weakly Supervised Learning)，来学习NER中一些逻辑规则，通过迭代(bootstrap)的方式生成高质量的逻辑规则，以完全自动化的方式训练神经标注模型。此外，文中进一步设计了一种动态标签选择策略，以确保伪标签的质量，从而避免神经标签的过拟合，提出数据生成的质量。

8. **BERTifying the Hidden Markov Model for Multi-Source Weakly Supervised Named Entity Recognition**

在使用多个数据源的弱监督方式识别NER任务中，存在标签往往不完整、不准确和矛盾的问题。为此，作者提出了一种条件隐马尔可夫模型(CHMM)，该模型能够以无监督的方式从多源噪声标签中有效地推断出真标签。CHMM模型利用预训练语言模型的上下文表示能力，增强了经典的隐马尔可夫模型，同时也提高多源弱监督的学习能力。

9. **Locate and Label: A Two-stage Identifier for Nested Named Entity Recognition**

本篇是解决嵌套命名实体识别(Nested Named Entity Recognition)任务的，其方法思路是借鉴图像领域的目标检测任务，将嵌套任务转化成span的预测。具体是采用两阶段：第一步Locate，即定位实体的边界；第二步Label，即对识别span进行实体类型判断。

10. **Nested Named Entity Recognition via Explicitly Excluding the Influence of the Best Path**

这篇也是做嵌套命名实体识别的(Nested Named Entity Recognition)，作者同样是采用分层识别的方法，不同的是，作者提出的方法可显式地排除最优路径的影响，扩展了现有的次最优路径识别。并且在每个时间步骤保持一组隐藏状态，并有选择地利用它们来构建不同的潜在函数，以便在每个级别进行识别。同时，作者也发现：先识别最内层的实体比传统的先识别最外层的实体的方案具有更好的性能。

11. **Discontinuous Named Entity Recognition as Maximal Clique Discovery**

本篇论文是解决NER中细分任务：识别不连续的实体(Discontinuous Named Entity)。先前的方法主要采用分阶段的方式进行识别，但存在误差传递的问题。本文中，作者提出segment graph(片段图)的方法，图中每个节点(node)代表一个片段(segment)——这个片段可能独立组成一个实体，也可能是不连续实体的一部分，图中的边(edge)连接两个片段，代表属于同一个实体。这样就将不连续的实体识别问题转化成图中最大团发现的问题。

12. **A Span-Based Model for Joint Overlapped and Discontinuous Named Entity Recognition**

本篇跟前面做同样的任务：不连续的实体识别(Discontinuous Named Entity Recognition)。不同的是，本篇还聚焦重叠识别的学习，意思同时解决嵌套与不连续的实体识别(Overlapped and Discontinuous Named Entity Recognition)。为此，提出一种基于范围识别的模型，包括两个步骤：首先，通过遍历所有可能的文本范围来识别实体片段，从而识别重叠的实体。其次，进行关系分类，以判断给定的一对实体片段是重叠的还是连续的。

13. **A Neural Transition-based Joint Model for Disease Named Entity Recognition and Normalization**

在医疗领域，关于疾病实体的识别与规范化(Disease Named Entity Recognition and Normalization)任务，有两种方式建模，一种采用分阶段的pipeline方式，一种是联合学习变成多任务学习(multi-task learning)方式。实验表明后者学习模式表现更好些，但仍存在采用不同decoder而导致边界不一致性的问题，此外在规范化中也没有考虑候选词丰富的文本信息。针对以上两点问题，作者提出基于状态转移的模型，将端到端疾病识别和规范化任务转化为动作序列预测任务，该任务不仅具有共享输入表示的模型，而且通过状态转换在同一搜索空间进行搜索输出。

14. **An End-to-End Progressive Multi-Task Learning Framework for Medical Named Entity Recognition and Normalization**

本文与第8篇一样，是做医疗领域实体识别与规范化的。解决的也是pipeline方式存在误差传递的问题，提出端到端渐进多任务学习框架(End-to-End Progressive Multi-Task Learning)。基于此框架，利用三个渐进任务来减少错误的传播。

15. **De-biasing Distantly Supervised Named Entity Recognition via Causal**

本篇是对NER中远程监督学习(Distantly Supervised Learning)方式的一种改进。原先的方法存在严重的字典偏差，具体表现为虚假的相关性，削弱了学习模型的有效性和稳健性。因此，文中提出结构因果模型(SCM)，从根本上解释了词典偏差，将词典偏差分为词典内偏差和词典间偏差，并分析了其产生的原因。基于SCM，通过因果干预学习对字典偏差进行去偏(De-biasing)。

16. **SPANNER: Named Entity Re-/Recognition as Span Prediction**

近年来，命名实体识别(NER)的研究范式从序列标注转向范围预测(spanNER)。尽管显示spanNER初步有效性，但模型的架构偏差还没有被完全理解。本文首先探讨了范围预测模型与序列标记框架在NER任务上的优缺点，以及如何进一步改进该模型，从而促使基于不同范式的系统实现优势互补。

17. **Crowdsourcing Learning as Domain Adaptation: A Case Study on Named Entity Recognition**

本篇是以NER任务来研究众包学习的(Crowdsourcing Learning)，文中提出了不同的观点：认为所有众包标注都是针对单个标注的黄金标准，并发现众包与领域自适应(Domain Adaptation)高度相似，而这类最新的方法都可以应用到众包中。因此，提出了一种基于领域自适应方法的标注感知学习模型，该模型能捕获有效的领域感知特征。

18. **Improving Named Entity Recognition by External Context Retrieving and Cooperative Learning**

有研究显示，利用文档级的文本信息可以提升实体识别效果，但在很多应用场景下并没有可用文档的信息。针对此问题，作者提出利用搜索引擎来获取原始识别文本语义相关的文档信息，接着使用Re-ranking模块对检索到的文本进行排序，筛选出topk个文本作为额外的文本，最后连同原始文本一起输入Encoder模型进行多任务学习，包括NER学习和样本分布学习，从而提高NER识别效果。

19. **Modularized Interaction Network for Named Entity Recognition**

论文是在NER任务中新增加两个辅助学习任务：Boundary Module 和 Type Module，该两个模块融合形成一个交互网络(Interaction Network)，由交互网络学习NER任务中一些特征任务，包括边界、类型、片段(boundary, type and segment information)，将学到这些特征信息加到NER Module来提高识别效果。虽然该整体框架并不新奇，但笔者看来，论文的创新之处是在于Boundary Module 与Interaction Network进行了精心设计。在Boundary Module上，采用BiLSTM 作为encoder进行边界信息的学习，使用LSTM 作为decoder生成解码信息。

总结

通过收集的19篇有关NER任务来看出， 目前研究学者主要做的都是NER中细分方向的任务，聚焦的点有少样本学习(Few-shot)，弱监督学习(Weakly Supervised Learning)，以及NER中的嵌套，不连续问题。已很少做纯粹的NER任务，这也印证了正常的NER任务也被研究透了，没多少可研究的空间。这些论文基本都涵盖实体识别中存在的各种问题，其中数据问题仍是最大问题。数据问题着眼的点包括few-shot，other tag学习效果，细粒度新数据集，多特征等。

Reference

- https://zhuanlan.zhihu.com/p/409594252