一种用于医学命名实体识别与对齐的多任务对抗主动学习框架

一、简介

医学命名实体的自动识别和对齐是构建知识图和构建质量保证系统的基础。我们应用命名实体识别（NER）技术从医学文本中找到关键医学名词（实体）的范围，然后利用命名实体对齐（NEN）技术将从医学文本中提取的实体映射到标准医学标识符（如：MeSH 编码和 OMIM 编码）[1]。我们在预测 NER 任务的标签时，发现其位置信息与 NEN 任务的标签位置有着强关系。鉴于 NER 任务与 NEN 任务之间的相关性，一种用于医学命名实体识别与对齐的多任务模型被提出并取得了较好的结果[1]。

然而，上面的监督学习模型需要一个丰富的实验数据集，但这种标注过程在医学领域中是很昂贵的。当涉及到医学文本时，实验人员对语料库的维护是非常费力耗时的。而主动学习是一种有效的减少标记负荷的半监督算法。科研人员首次将深度神经网络与主动学习相结合应用于命名实体识别，并获得了较好的实验效果[2]。考虑到不同任务的相关性，一些研究者提出了针对语言标注的多任务主动学习模型[3][4]。前面提到的多任务模型采用软共享参数框架并广泛应用于语言模型的训练。这种框架鼓励编码器学习对所有任务都有益的共享特征，并利用解码器提取特定任务的特征来预测任务目标序列。但它不能保证不同任务的私有特征在共享特征空间中不会相互干扰。

为了解决上述问题，我们需要一个框架来独立地学习多任务的共享和私有特征。为了保证共享特征和私有特征能够被独立提取，科研人员提出了一种基于对抗学习的多任务文本分类模型[6]。这也证明了多任务对抗训练的有效性。然而，现在的多任务主动学习模型没有考虑任务私有特征对主动学习查询样本过程的影响。这是现有的多任务主动学习领域所面临的主要挑战。现在的多任务主动学习模型利用任务私有特征进行样本查询。它们根据任务各自的特征学习情况选择当前最不确定的未标注样本进行标注并加入训练集。然而这些多任务主动学习模型中，任务私有特征混合在共享特征空间。因此，不能保证所选择的样本是对各自任务模型是最有利的。

我们提出一种多任务多抗主动学习（**m**ulti-task **a**dversarial **a**ctive **l**earning, MTAAL）框架来弥补当前模型的缺点。在我们的模型中，对抗学习是基于任务和多样性的。基于任务的对抗学习保证了任务特征被限制在多任务的私有特征空间内以避免影响主动学习过程。如此一来，MTAAL 模型选择的样本保证与各任务的特征空间具有较强的相关性，有利于 NER 和 NEN 任务模型性能的提升。基于多样性的对抗学习是一种与多任务特征学习无关的主动学习模型[5]。它的主要原理是选择未标记样本中与标注样本相似性最小的实例，如此一来将多任务特征的影响彻底避免。本文的主要工作和贡献可以总结如下：

1. 通过实验发现了现有多任务主动学习模型的不足之处。现有模型不能保证主动学习选取的样本与各任务的特征空间均有很强的相关性，且对任务性能提升最为有益。

2. 我们提出一个多任务对抗主动学习框架。任务判别器可以避免任务特征混合对查询样本过程的影响。基于多样性的对抗学习是一种有效的多任务主动学习算法。

3. 我们评估了多任务对抗主动学习框架在两种常见的医学 NER 和 NEN 基准上的性能，并获得了比现有的多任务主动学习模型的更优越的结果。

二、数据集

在本次实验中我们使用两个基准医学实体识别与对齐任务数据集。**NCBI** 数据集[7]包含 793 条医学病历文本，均采用 MeSH/OMIM 编码对文本中的疾病等实体标注。**BC5CDR** 数

据集[8]包含 1500 条公开医学病历文本。它采用 MeSH 编码对文本中的疾病和医学实体进行标注。待提取出文本中涉及实体名词的句子后，**NCBI** 数据集占用 2.09MB 大小，**BC5CDR** 数据集占用 4.25MB 大小。

三、研究方法

本文我们设计的多任务主动学习框架（MTAAL）主要包括：共享编码器、任务私有解码器、任务判别器、多样性判别器。共享编码器负责提取输入语句的潜在表示。为了获得不同任务的预测序列标签，我们设计了 NER 和 NEN 的任务私有解码器。通过将任务判别器引入模型，我们实现了基于任务的对抗性学习。基于多样性学习的多样性鉴别器是选择未标记样本进行标记的关键。上面提到的每个模块包含各自特定的序列特征提取器。

为了体现我们框架的优越性，我们需要与前人的方法进行比较，主要包括三种方法：随机查询策略（Random）、基于熵的查询策略（Entropy）[4]、基于最小置信度的查询策略（Least Confidence）[2]。除此之外，我们还将任务判别器应用在上面的基准方法，同样取得了很大的性能提升。最后，我们进行了消融实验探究模型各个组成部分对框架性能的影响。

参考文献

[1] Sendong Zhao, Ting Liu, Sicheng Zhao,and Fei Wang. A neural multi-task learning framework to jointly model medical named entity recognition and normalization. In Proceedings of the 33th AAAI, pages 817–824, 2019.

[2] Yanyao Shen, Hyokun Yun, Zachary C. Lipton, Yakov Kronrod, and Animashree Anandkumar. Deep active learning for named entity recognition. In 6th ICLR, 2018.

[3] Roi Reichart, Katrin Tomanek, Udo Hahn, and Ari Rappoport. Multi-task active learning for linguistic annotations. In Proceedings of the 46th ACL, pages 861–869, 2008.

[4] Fariz Ikhwantri, Samuel Louvan, Kemal Kurniawan, Bagas Abisena, Valdi Rachman, Alfan Farizki Wicaksono, and Rahmad Mahendra. Multi-task active learning for neural semantic role labeling on low resource conversational corpus. In Proceedings of the Workshop on Deep Learning Approaches for Low-Resource NLP, pages 43–50, 2018.

[5] Yue Deng, KaWai Chen, Yilin Shen, and Hongxia Jin. Adversarial active learning for sequences labeling and generation. In Proceedings of the 27th IJCAI, pages 4012–4018, 2018.

[6] Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. Adversarial multi-task learning for text classification. In Proceedings of the 55th ACL, pages 1–10, 2017.

[7] Rezarta Islamaj Dogan, Robert Leaman, and Zhiyong Lu. NCBI disease corpus: A resource for disease name recognition and concept normalization. Journal of Biomedical Informatics, 47:1–10, 2014.

[8] Jiao Li, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Thomas C. Wiegers, and Zhiyong Lu. Biocreative V CDR task corpus: a resource for chemical disease relation extraction. Database, 2016.