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-**T**ask **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、模型

开始介绍模型前，我们需要定义在多任务主动学习场景中的一些符号，方便我们后面的细节描述。我们定义三元组 $(x^L, y^L_{\text{NER}}, y^L_{\text{NEN}}) \sim (X^L, Y^L_{\text{NER}}, Y^L_{\text{NEN}})$ 为一个标注样本。其中 x^L

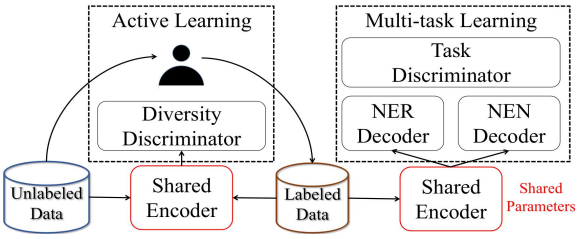

图 1 MTAAL 模型框架

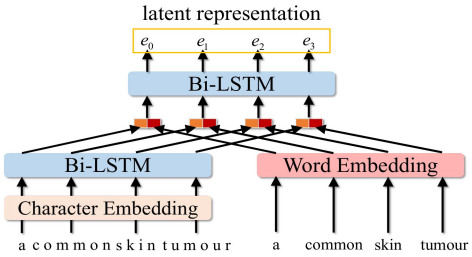

图 2 Shared Encoder 内部结构

代表医学文本，而 y^L_{NER}、y^L_{NEN} 分别代表实体识别任务的标签和实体对齐任务的标签。在主动学习场景中，$x^U \sim X^U$ 代表未标注的样本。在每个查询过程中，我们根据主动学习算法从未标注样本池 X^U 中选取一定数量样本 X^S。对选择出的样本进行标注后加入标注集合。

模型的整体框架如图 1 所示，主要包括三部分：共享编码器（Shared Encoder）、主动学习模块（Multi-task Learning）和多任务学习模块（Active Learning）。下面我们针对每个模块进行详细介绍。

共享编码器（Shared Encoder）E 用于将输入的医疗文本 x 映射为隐表征：$e = E(x)$。其内部结构如图 2 所示。主动学习模块与多任务学习模块利用该表征分别进行多样性样本的查询和 NER、NEN 任务模型的训练。我们的编码器结构考虑对输入文本的字符级和单词级特征建模。对于单词中的每个字符，我们使用预先定义的字符嵌入矩阵将它映射为一个稠密向量。然后利用双向 LSTM（Bi-LSTM）层得到字符级别的上下文表征。对于句中的每个单词，我们使用预训练词向量将它映射为一个稠密向量。为了得到多级别的混合表征，我们将字符表征和单词向量拼接起来，一同输入双向 LSTM 层。最终，我们可以得到输入句子的隐表征 e。

多任务学习模块主要包括了：NER、NEN 的任务私有解码器以及任务判别器。其内部结构如图 3 所示。前面提到共享编码器 E 负责提取输入句子的隐表征。而为了进行 NER、NEN 任务的结果预测，隐表征需要被转换为任务私有特征。因此，我们设计了 NER、NEN 私有解码器（NER、NEN Private Decoder）。这些解码器均由自注意力层（Self-Attention）和双向 LSTM 层组成。

由于任务 $k \in \{\text{NER}, \text{NEN}\}$ 的标签空间不同，它们关注于同一句子的不同内部结构信息。为了明确地学习句子中两个单词之间的关系，我们将自注意力层作用于隐表征。如此一来，各自解码器的自注意力层输出任务私有特征 p_k。最后，私有特征输入双向 LSTM 层并经过 softmax 函数后，我们得到任务的预测概率 $\hat{y}_k$。在训练网络时，我们采用交叉熵损失作为目标函数。现在输入训练样本 $(x^L, y^L_{\text{NER}}, y^L_{\text{NEN}})$，单任务的损失函数可以被定义为：$L_k = -\sum y^L_k \log \hat{y}^L_k$。为了训练多任务模型，多任务损失可被定义为：$L_{\text{Task}} = \sum_{k \in \{\text{NER}, \text{NEN}\}} L_k$。

在传统多任务主动学习模型中，NER、NEN 任务模型的性能受到任务特有特征的影响而变糟。因为共享特征存在于任务私有特征空间，而任务特有特征潜入到共享空间[6]。不规则的特征空间影响了两个任务的性能，继而影响基于不确定估计的主动学习方法。

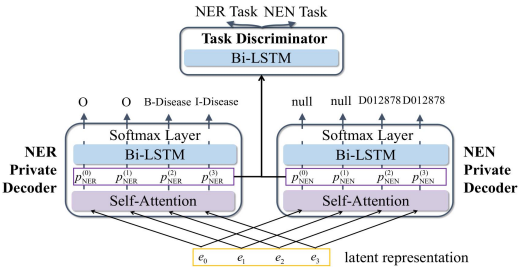

图 3 多任务学习模块

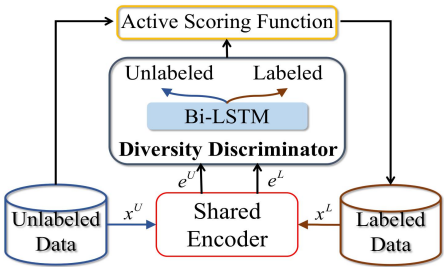

图 4 主动学习模块

为了解决上述问题，我们将任务判别器（Task Discriminator,）TD 引入多任务模块。我们利用共享编码器与任务私有解码器之间的极大极小博弈来促使模型学习到规则的特征空间。TD 将任务私有特征映射为一种概率分布。任务判别器由双向 LSTM 层与 softmax 层组成，被用来判断任务私有特征 p_k 来自于哪一任务。为了训练网络，基于任务的对抗损失可以被定义为：$L_{\text{Task}}^{Adv} = \min_{\theta_E} \max_{\theta_k, \theta_{TD}} \sum_k d_k \log(TD(p_k))$。$d_k$ 表示任务私有特征所属任务的真实标签。

前人提出将多样性作为模型训练目标的主动学习方法取得了很好的效果[5]。而传统的多任务主动学习没有将多样性作为模型训练的显式目标，故训练样本多样性无法得到保证。为了解决上面的问题，我们设计了主动学习模块。**主动学习模块**主要包括：多样性判别器（**D**iversity **D**iscriminator）DD，其内部结构如图 4 所示。多样性判别器同样由双向 LSTM 层以及 softmax 层组成。它负责将来自共享编码器的隐表征 e 映射为一种概率分布，用来估计输入模型的句子来自标注集合还是未标注集合。现给定输入句子 x^L 和 x^U，我们通过共享编码器获得它们的隐表征 $e^L = E(x^L)$ 和 $e^U = E(x^U)$。多样性判别器负责将 e^L 分类为已标记样本（$DD(e^L) = 1$），同时将 e^U 分类为未标记样本（$DD(e^U) = 0$）。因此，优化 DD 的目标函数为：$L_{\text{DD}}^{Adv} = -\log(DD(E(x^L))) - \log(1 - DD(E(x^U)))$。

为了使模型学习到标注样本与未标注样本之间的细粒度表征，共享编码器被训练来迷惑多样性判别器，使它将隐表征 e^L 和 e^U 均分类为标注样本。我们定义相关的损失函数为：$L_{\text{E}}^{Adv} = -\log(DD(E(x^L))) - \log(DD(E(x^U)))$。经过训练，多样性判别器可以从未标注集合中选择出与标注集合最不相同的样本。然后，我们定义主动分数函数来选择未标注样本：$\psi_{\text{Diversity}}(x^U) = 1 - DD(E(x^U)) \in (0,1)$，$x^S = \max_{x^U \sim X^U} \psi_{\text{Diversity}}(x^U)$。

具体模型的训练流程参见附录-Multi-Task Adversarial Active Learning 算法流程。

三、数据集与实验设定

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。我们按照句子结束符将病例文本划分为平均包含 40 个单词的句子。

模型中双向 LSTM 的隐含层神经元数量均被设置为：64，且学习率被设定为 0.001。我们采用预训练词向量 Glove 和 Word2Vec 作为单词级别的输入特征，并在实验中进行了讨论。

实验中，我们一共进行 70 次查询操作，且每次查询后模型被微调训练一个 epoch。针对 **NCBI** 数据集，我们每次查询 64 个样本加入标注集合；而 **BC5CDR** 数据集则是 128 个样本。为了评价 NER、NEN 模型的性能，我们采用 F1 值作为评价指标。

我们的模型要和现有多任务主动学习方法进行比较，具体包括如下几种：

1. **Random**：这是一种最为直接的策略。它忽略样本的属性而以相同的概率选择样本。

2. **Entropy**：NER、NEN 模型经过训练可以计算输入句子的预测概率，由此可以得到基于熵的主动得分函数：$\psi_{\text{Entropy}}(x^U) = -\sum_k \sum_{N_w} \hat{y}_k \log(\hat{y}_k)$。我们对未标注样本按所得分数递增顺序排序，并选择最高得分样本加入标注集合。

3. Least Confidence（**LC**）：该方法同样利用 NER、NEN 模型对输入句子的预测概率。基于最小置信度的主动得分函数为：$\psi_{\text{LC}}(x^U) = \sum_k 1 - \max_{\hat{y}_k} P(\hat{y}_k \mid x^U)$。

四、实验结果分析

Method	ST		MT	
	NER	NEN	NER	NEN
LC+w	0.8408	0.8862	0.8107	0.8869
LC+g	0.8402	0.9066	0.8338	0.9040
Entropy+w	0.8334	0.8864	0.7868	0.8931
Entropy+g	0.8381	0.9008	0.8292	0.9015
Random+w	0.8137	0.8856	0.7651	0.8836
Random+g	0.8188	0.8936	0.8005	0.8923
MTAAL+w	0.8411	0.8873	0.8462	0.9091
MTAAL+g	0.8492	0.9103	**0.8600**	**0.9152**

表 1 BC5CDR 数据集结果

Method	ST		MT	
	NER	NEN	NER	NEN
LC+w	0.7608	0.9158	0.6736	0.9151
LC+g	0.7452	0.9200	0.6752	0.9151
Entropy+w	0.7394	0.9154	0.7257	0.9139
Entropy+g	0.7462	0.9201	0.7291	0.9137
Random+w	0.7099	0.9151	0.6763	0.9138
Random+g	0.7185	0.9194	0.6749	0.9137
MTAAL+w	0.7688	0.9284	0.7682	0.9267
MTAAL+g	0.7542	0.9267	**0.7744**	**0.9287**

表 2 NCBI 数据集结果

上表中，“+g”表示模型使用 Glove 词向量，而“+w”表示 Word2Vec 词向量。表 1 结果为模型查询 55 次后的结果；表 2 结果为模型查询 25 次后的结果。

我们可以看到：处于多任务（MT）场景下，LC、Entropy、Random 方法在 NER、NEN 任务上的结果会低于单任务（ST）场景中的结果。这一点表明：传统多任务主动学习方法确实会受到不规则特征空间的影响，继而影响 NER、NEN 任务模型的性能。而 MTAAL 通过引入任务判别器和多样性判别器，同时保证了学习规则的特征空间以及主动学习过程的样本多样性。因此，本文方法在多任务（MT）场景中取得了最好的结果。预训练词向量 Glove 和 Word2Vec 对 NER、NEN 结果的影响也是不同的。在本次实验中，Glove 词向量的结果要更好一些。该部分实验结果证明了本文方法的有效性。

为了更好地说明任务判别器（TD）的作用，我们将它应用在前人的方法 LC、Entropy、Random 上。具体结果可见图 5。图中虚线为未使用 TD 时的结果，而实线为使用 TD 时的结果。我们可以看到：TD 对于前人方法在 NER、NEN 任务上均有所提升。这也从侧面说明，传统的多任务主动学习方法受到不规则的特征空间影响，NER、NEN 的性能都会降低。通过 TD 的作用，NER、NEN 任务模型学习到规则的特征空间，保证了主动学习过程查询样本的高效性。如此一来，NER、NEN 任务上的模型性能得到了保证。

为了更直观的比较我们与前人方法的差异，我们在图 6 中展示了查询次数对模型性能的影响。随着查询次数的增加，不同方法在 NER、NEN 任务取得的 F1 值都在上升。不过，NER 任务上的 F1 值逐渐平缓，说明模型的性能达到了上限。而在查询过程中，MTAAL 在 NER、NEN 上的得分往往高于其他方法。而且 MTAAL 往往可以查询最少的次数即可使模型达到较优性能。这说明本文将多样性作为多任务主动学习模型的训练目标对于提升 NER、

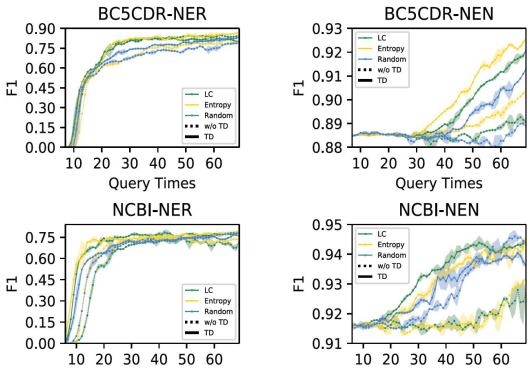

图 5 任务判别器的影响

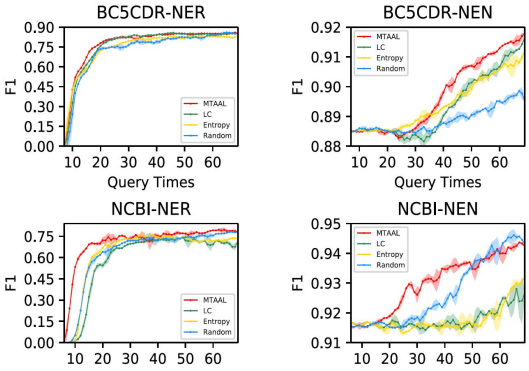

图 6 查询次数的影响

NEN 模型性能是有帮助的。

通过上面的实验，我们证明了本文提出的 MTAAL 模型的有效性。同时，我们更细致地讨论了任务判别器对前人方法的有效性。结合了任务与多样性判别器的 MTAAL 在两个医学数据集上取得了最优结果。

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.

附录

Multi-Task Adversarial Active Learning 算法流程

输入：标注样本集合 $(X^L, Y_{\text{NER}}^L, Y_{\text{NEN}}^L)$、未标注样本集合 X^U；初始化的模型参数：共享编码器 θ_{E}、任务判别器 θ_{TD}、多样性判别器 θ_{DD}、任务私有解码器 $\{\theta_k \mid k \in \{\text{NER}, \text{NEN}\}\}$

1. **for** e=1 to 查询次数 Qn **do**
2. 获取批量样本 $x^L \sim X^L$、$x^U \sim X^U$
3. 最小化损失函数 L_{DD}^{Adv} 以更新参数 θ_{DD}
4. 最小化损失函数 L_{E}^{Adv} 以更新参数 θ_{E}
5. 根据主动得分函数 $\psi_{\text{Diversity}}$ 查询样本 X^S
6. 进行标注 $(Y_{\text{NER}}^S, Y_{\text{NEN}}^S) \leftarrow ORACLE(X^S)$
7. 将查询样本加入标注集合 $(X^L, Y_{\text{NER}}^L, Y_{\text{NEN}}^L) \leftarrow (X^L, Y_{\text{NER}}^L, Y_{\text{NEN}}^L) \bigcup (X^S, Y_{\text{NER}}^S, Y_{\text{NEN}}^S)$
8. 将查询样本从未标注集合移去 $X^U \leftarrow X^U - X^S$
9. **for** 采样小批量样本 $x^L \sim X^L$ 和 $x^U \sim X^U$ **do**
10. 最小化损失函数 L_{Task}^{Adv} 以更新参数 θ_{E}
11. 最大化损失函数 L_{Task}^{Adv} 以更新参数 θ_{TD} 和 $\{\theta_k \mid k \in \{\text{NER}, \text{NEN}\}\}$
12. **end for**
13. **for** 采样小批量样本于 $(X^L, Y_{\text{NER}}^L, Y_{\text{NEN}}^L)$ **do**
14. 最小化损失函数 L_{Task} 以更新参数 θ_{E} 和 $\{\theta_k \mid k \in \{\text{NER}, \text{NEN}\}\}$
15. **end for**
16. **end for**

输出：训练后的模型参数 θ_{E}、θ_{TD}、θ_{DD}、$\{\theta_k \mid k \in \{\text{NER}, \text{NEN}\}\}$