在临床文件中对评估和计划的结构化理解

Doron Stupp*, Ronnie Barequet, I-Ching Lee, Eyal Oren, Amir Feder, Ayelet Benjamini, Avinatan Hassidim, Yossi Matias, Eran Ofek & Alvin Rajkomar

*.通讯作者: doronst@google.com

摘要

医生将他们对诊断和治疗的详细思考过程以非结构化文本的形式记录在名为评估和计划的临床笔记中。这些信息在临床上比为一次会诊分配的结构化收费代码更丰富，但由于临床语言和文档习惯的复杂性，很难可靠地提取。我们描述并发布了一个数据集，其中包含了579份入院和进展记录的注释，这些注释来自于公开的、去识别的MIMIC-III ICU数据集，其中有超过30,000个标签来识别活跃的问题、它们的评估以及相关行动项目的类别（如药物治疗、实验室测试）。我们还提出了基于深度学习的模型，接近人类的表现，其F1得分为0.88。我们发现，通过采用弱监督和特定领域的数据增强，我们可以在不牺牲性能的情况下提高各部门的概括性，并减少人类标记的笔记数量。

简介

看完病人后，医生会写一份正式的临床笔记，记录病人的病史、发现和结论。其中一个关键部分称为评估和计划，通常写成一个结构松散的、以问题为导向的病情清单（如 "类风湿性关节炎"），其评估（如 "新发炎"）和计划（如 "颈部CT并开始治疗"）。文本本身是自由形式的，并以医生--专业--和组织--的方式写成，使得它很难用算法解析成结构化数据。这种结构化的数据是有价值的，例如，推动各种诊断和治疗发生时间的纵向视觉化。依靠固有的结构化数据元素，如编码的ICD-9/10诊断，可能是不可靠的[1]，自由文本是医生评估的更多信息反映。

之前处理非结构化临床笔记数据转换为更多结构化形式的研究主要集中在实体匹配和链接层面（如识别疾病、药物及其关系）。以前的工作已经解决了这个问题的各个方面，识别实体[2–6]和他们的关系[7,8]，识别突出的事件[9]和建立个性化的临床知识图谱[10]。然而，对完整的笔记或章节的结构化，大部分仍然没有探索过。最密切相关的工作是Mullenbach等人，[11]，他们专注于识别和分类出院总结中的出院指示，这是住院后后续门诊护理的重要任务。在这项工作中，我们关注的是住院病人的评估和计划部分及其更丰富的以问题为导向的结构，它不仅包括行动项目（类似于出院指示），还包括它们的相关情况及其在入院和进展记录中的评估。

在本文中，我们描述了一个数据集和深度学习模型，用于将非结构化的评估和计划（A&P）部分解析为表明条件、评估和计划行动项目的片段。我们提出了一个由公开的、去掉身份识别的MIMIC-III ICU数据集中的579份笔记组成的数据集，[12]，其中的A&P部分由临床医生注释为主动问题及其相关的评估描述和计划行动项目（超过3万个注释）。然后我们提出了一系列深度学习和手工设计的模型，并分析了它们的性能。我们表明，在这个任务上训练的模型接近于人类可比较的性能。此外，在临床灵感启发的启发式方法上进行弱监督训练，再加上人类标签以及使用数据增强，可以提供最佳的性能。然后，我们表明这些技术可以减少对人类标记数据的要求，并帮助模型普及到医院其他部门所写的笔记。

结果

在高层次上，我们让医疗专业人员对公共数据集中的评估和计划部分进行注释，并使用包括弱监督和数据增强在内的各种建模技术训练一系列模型来进行相同的注释。然后，我们研究了在训练过程中，有多少手工注释的笔记是必要的，以便在不同医院部门的笔记中获得良好的性能和概括性。

标签和数据收集

我们将A&P部分中与问题导向结构有关的文本跨度分为三类：主动问题的标题（如 "败血症"），它的描述评估（如 "尿毒症、肺炎的dd"），以及计划行动项目（如 "WBC趋势"）。与这些类别无关的文字则不加标记。行动项目被进一步细分为八个不同的类别：药物、观察/化验、成像、咨询、营养、治疗程序、其他诊断程序或其他。补充图2显示了每个类别的频率。补充图1显示了一个注释说明的例子。

我们从MIMIC-III ICU数据集中的医生书写的笔记中随机选择了579个A&P部分，该数据集包含了来自Beth-Israel Deaconess医疗中心ICU病房的5万多个病人的住院记录[12]。所选笔记包括ICU病房的入院和进展笔记。笔记的抽样是为了让每个病人最多只有一张笔记。这些笔记中约有90%用于训练，10%用于测试集。除非另有说明，所有的结果都是基于测试集的。

标签是通过两种程序产生的：人类评分员和临床启发式方法。训练集中的每一个笔记都由一个评分者进行标注，测试集的笔记由六个标注者进行标注，由一个医生作为基础真理（见方法）。

以Jaccard相似度衡量的人类评测者之间的一致性对于跨度类型是0.77（CI 0.75-0.79），对于行动项目类型是0.62（CI 0.6-0.64）（都是跨度级别的微观平均值，见图1）。虽然平均而言，测量者之间的一致性很高，但与地面真实相比，跨度级的微观平均F1得分从0.62到0.93不等，性能上存在差异。与训练集相比，验证集是由具有最高性能（与基本事实相比）的评分者标记的，这可能会导致更高的质量。对于行动项目类型的分类，混淆矩阵（补充图4）显示，大部分分歧出现在频率较低的标签（"其他诊断程序 "和 "其他"）。值得注意的是，分歧在语义上是一致的--例如，将药物标记为治疗程序。从质量上看，这些案例大多涉及边缘案例，如氧气、血制品、液体等，在这些案例中，标签说明要求使用 "药物治疗"，但评分者在许多情况下选择了治疗程序。

临床启发式是用正则表达式来实现的，以捕捉活动问题的圆点列表，然后是圆点列表的行动项目（例子见方法，补充图1）。重要的是，启发式只捕捉到跨度类型（问题标题、描述或行动项目）。由于缺乏一个简单的启发式方法，行动项目类型没有被包括在临床启发式方法中。

模式

我们将识别文本中活动问题的标题、描述和行动项目的跨度的任务建模为序列标记任务。该模型由一个2层双向长短期记忆网络（LSTM）[13]，上面有一个条件随机场（CRF）预测头[14]。对于每个标记，该模型预测其跨度类型和行动项目类型（如果适用）（见方法）。

由于在临床领域获得人类标记的数据很昂贵，我们试图评估在模型中注入领域知识是否能提高数据效率，并有可能提供性能的提升。具体来说，我们专注于弱监督和数据增强。这两种技术都是机器学习中已知的提高模型性能和降低数据要求的技术，并在NLP和其他任务中得到了探索，包括在医学领域[15,16]。
为了训练弱监督模型，我们使用了25,000条用上述临床启发式方法标记的笔记。这些笔记的选择与人类标记的笔记类似，属于不同的病人，与人类评分者已经标记的笔记没有重叠。为了实现数据增强，我们通过算法重组了标记的评估和计划部分（无论是人类还是启发式标记的），以包含同行的行动项目、与描述交错的行动项目和混合bulleting（见补充图5的例子）。

模型的训练分为两个阶段，一个是 "预训练 "阶段，有或没有弱监督，有或没有增强，另一个是在非增强的人类标记数据上的第二阶段（见方法）。具体来说，在第一阶段，4个模型被训练为使用弱监督和/或数据增强的4种可能组合。用弱监督训练的模型是在用临床启发式标记的25,000个随机笔记上预先训练的。为了进行公平的比较，只在人类标记的数据上训练的模型的总步数相同，并采用类似的早期停止策略（见方法）。作为一个基线，我们将我们的模型的性能与临床启发式的性能进行比较。这为这项新任务提供了一个基于规则的基本基线。

所有的模型都达到了明显优于基线模型的性能，并且在跨度类型的性能上接近了中位数的评分者（表1）。补充图6中显示了跨度类型的模型损失模式。这些损失模式在质量上与测评员之间的比较中观察到的相似。

训练集大小对模型性能的影响

然后，我们研究了如图2所示的添加额外标签时的模型性能改进。在不同的训练数据集规模下，用弱监督和增量训练提高了跨度水平的性能，在对数百个音符进行训练时，性能趋于一致。重要的是，一个在25个音符上进行弱监督和增强训练的模型（每个面板上从左边开始的第二个数据点）在跨度层面上取得了与在没有弱监督和没有增强的481个音符的完整训练数据中训练的模型相当的性能。然而，这种效果在标记层面上并不明显。值得注意的是，行动项目类型的标签对于弱监督来说是不可用的，因为它们没有被启发式标记，导致两种情况下的性能相似。

各个部门的业绩归纳

临床模型往往是在一个中心开发的，但部署在其他医院，而这些医院的数据分布是未知的，这使得模型的通用性难以确定。为了估计这一点，我们在一组特定的科室上训练了一系列的模型，并在不同的科室上进行测试。我们按各自的服务将笔记分为两类--内科或外科（或其他，见方法）。不同的服务有不同的跨度分布（补充图3）和不同的典型笔记结构（例如，见补充图1中的笔记）。手术记录倾向于采用面向系统的模板格式，大约30%的手术记录具有完全相同的活动问题集。描述和行动项目通常是交错排列的。来自医疗服务的笔记倾向于以病情为导向，问题标题更多样化。描述大多紧跟在问题标题之后，后面有一个行动项目的圆点清单（类似于基于规则的启发式结构）。因此，我们在单一服务上训练模型，并在两种服务上测试它们（图3）。用增量或弱监督训练的模型在服务中和跨服务中的表现都趋于更好。然而，在手术记录上训练的模型和在医疗记录上评估的模型在跨服水平上的表现明显更好，当也用弱监督训练时。这可能是由于启发式方法在医疗服务笔记上的表现相对较高。

预测探索

评估和计划部分的结构可以发现病人护理的各个方面，这些方面主要或仅在文本中规定。为了解决这种结构的潜在效用，我们对MIMIC-III中的所有141K医生笔记进行了预测，共预测了4.7M的跨度。然后我们探讨了相关的行动项目和一系列常见问题的描述（表2）。从质量上看，检测到的跨度确实与临床相关。例如，"心血管 "与相关药物（阿司匹林、β受体阻滞剂）有关，"呼吸衰竭 "与相关研究（胸部X射线、血气、培养）有关。这些跨度中的许多可能很难与更传统的关系提取方法的问题联系起来，因为它们不对应于明确的实体（如 "避免肾脏毒素"）或模糊不清（"是"-激励性肺活量计，"p.t."-物理治疗和病人）。

讨论

我们展示了一种方法，通过利用临床启发式方法和数据增强技术形式的文档领域知识，可以将非结构化的临床医生评估和计划部分可扩展和准确地解析为结构化的数据元素，而不需要标记大量的笔记。我们还为临床社区发布了一个经过专业标记的评估和计划的数据集，以提高机器学习模型在这项任务上的性能。

评估和计划部分包含关于病人状态的临床思考要点。这包括可能难以从结构化数据中解读的信息，如化验单或用药单（即获取数据的意图）。这项任务本身就很难，因为不同机构、部门和个别医生的记录模式是不同的。临床医生评分者本身对如何标记笔记有许多分歧，需要非琐碎的标记说明（补充文本1）。有趣的是，我们发现该模型的损失模式在质量上与评估员之间的分歧所发现的模式相似。

为了实现这一性能，我们利用了两种方法来补充模型的专业性--来自启发式方法的弱监督和数据增量。这两种方法以前都成功地用于一些领域[15]，包括医学文本[16]。这些方法在时间上与人工标注存在着内在的权衡。补充领域知识的开发成本很高，因为启发式方法本身就是从临床医生那里学来的，而医学专家的标注时间本身就可能被证明是昂贵的。我们表明，用领域知识补充模型对低数据制度和跨部门都是有益的。在为新的医院系统调整模型时，这可能被证明是有用的，可以推广到许多新的地方，最初的实施是一次性的成本，不像需要经常性的标注。

这项工作与最近的研究成果相联系，将各种笔记和笔记部分结构化，如出院说明[11]，影像报告[10]，以促进对医疗笔记的计算理解。在这里，我们主要关注住院病人的进展结构，然而，类似的结构可能也适用于门诊病人的进展记录。

我们的工作有几个局限性。这项工作的一个关键发现是，解析评估和计划部分本身具有主观因素，这一点在评分者之间的分歧中得到了强调。这导致了所发布的数据集的一个局限性，即训练和验证数据只用一个评分者来标注。由于性能只在测试集上测量，可能会出现训练数据受低质量评级的影响较大，导致模型性能不理想的情况。至于模型，虽然本文提出的模型可能会通过利用自然语言处理的深度学习的最新进展来进一步改进（例如，通过使用BERT[17]或T5[18]等转化器），但本文提出的较小的模型是常用的，可以更容易地部署在现实世界的EHR。最后，跨部门的概括可能无法正确模拟跨地点的概括。一些差异

不同笔记的差异预计会来自于科室做法以外的其他来源--病人群体、病情严重程度、护理频率、EHR系统的特定模式和其他不可预见的因素都可能导致不同的记录风格。

未来的工作可以利用解析后的评估和计划来驱动功能，帮助临床医生更快地了解病人的轨迹和历史。例如，可以利用问题及其相关行动项目的时间序列，例如，建立患者尝试各种条件和治疗的轨迹的时间线。同样地，与关系提取一起，问题和相关行动项目的提取可以用来准确地建立整个病人群体的知识图谱。

最后，我们提出了用于构建临床医生笔记中评估和计划部分的模型和相关的临床医生注释的数据集。我们表明，通过将领域的专业知识以弱监督的形式与临床启发的启发式方法和策划的特定领域的数据增量结合起来，可以在有限的人类标记的笔记中实现出色的性能，并在各部门之间保持。

方法

标签和数据收集

笔记从MIMIC-III[12] 笔记事件表中抽出，并根据类别栏进行过滤，只保留医生笔记。作者进一步手工检查笔记，只保留至少有一个非合成的评估和计划部分的笔记，总共放弃了2个笔记。总共抽出579份笔记，分为48份笔记的黄金集、481份笔记的训练集和50份笔记的验证集。6名住院医师和医学生根据结果中描述的标签说明对这些笔记进行了标注（并在补充文本1中全面介绍）。提出了具体的指导意见，以区分描述和行动项目，分割和合并跨度以及区分行动项目的类别。
训练集和验证集的笔记分别由一个评分者标注。黄金集由所有6位评分者进行标注，其中一位评分者（主要作者，评分者ID 1）作为基础真理。评卷人1在审查其他评卷人之前给测试集贴了一次标签，作为评卷人内部比较，之后又贴了一次，作为黄金集。评分者遵循补充文本1中的标签说明。评分者在评分前要先完成一小部分笔记（并接受反馈）或使用说明中的标准化测验进行评估。标签进行了自动规范化处理，以捕捉整个单词的边界，并去除侧面的非字母数字字符。在重复尝试之间，约有2个月的时间来衡量评分者内部的一致性。测评者之间和测评者内部的一致性被计算为测评者之间的平均成对雅卡德指数。

服务归属

每张票据都与写票据时治疗病人的服务有关。具体来说，服务表被使用，并与入院记录表相匹配，在每张入院记录中取最近的服务。服务被大致分为 "内科"、"外科 "和 "其他"。"内科 "包括 "MED"（内科）、"CMED"（心脏科）、"NMED"（神经科）和 "OMED"（骨科）。"外科 "包括 "SURG"（普通外科）、"TRAUM"（创伤）、"CSURG"（心脏外科）、"NSURG"（神经外科）、"TSURG"（胸外科）、"VSURG"（血管外科）、"ORTHO"（矫形外科）、"ENT"（耳鼻喉科）、"GU"（泌尿外科）和 "GYN"（妇科）。

注入领域专业知识

模型中以两种方式捕捉领域的专业知识：（i）对临床启发式的弱监督和（ii）数据增强。启发式是使用正则表达式来捕捉活跃的问题及其相关描述和行动项目。该启发式捕捉包含行动项目的圆点或编号列表的活动问题（完整的python实现可以在代码中找到）。启发式对行动项目的类型是无视的。用这种启发式方法注释的跨度既可以作为基线模型，也可以作为下面描述的用弱监督训练模型的标签。

数据扩充的目的是为了捕捉内联的行动项目、与描述交错的行动项目和混合bulleting。A&P部分被随机选择来进行数据增强。根据标签（启发式的或人工标注的），各部分被分解为有注释的跨度，并根据增强策略进行重建。各个部分可以经历几个独立的增强序列，它们都被用作训练数据。扩增的例子见补充图5。

数据生成

人类标记的笔记和25,000个启发式标记的笔记被处理以产生模型的训练、验证和测试集。25,000份启发式标记的笔记的取样与人类标记的笔记类似。重要的是，这些集合在笔记或病人层面上没有重叠。此外，这两组笔记的取样是这样的：每个病人在样本中只有一个笔记。训练集包括基于启发式的和人类的标签，验证集和测试集只包含额定的验证和地面真理标签的黄金集（即没有启发式的标签）。笔记经历了评估和计划部分的提取，基于白空间的标记化，保持标签和断行，数据增强（如上所述）和转换为TensorFlow实例。由人工生成的文本（如ICD代码）组成的A&P部分被删除。数据增强被随机地应用于评估和计划部分，以产生具有独特特征的部分，如不一致的标题，非标题的行动项目和交错的问题描述和行动项目（见补充图5的例子）。数据增强被应用于人类标记的和启发式标记的笔记。扩增是随机抽样的，平均每个笔记产生2个扩增的视图（泊松分布）。除了原始音符之外，还将增强的视图输入到模型中。

模型

该模型由一个多层双向LSTM组成，上面有一个CRF头。[19]Token id嵌入是由在维基百科上训练的Word2Vec嵌入初始化的（链接：https://tfhub.dev/google/Wiki-words-250-with-normalization/2）。该模型是以类似于命名实体识别模型的序列标记目标来训练的。它预测每个标记是否属于一个跨度（问题标题、描述或行动项目），对于行动项目，预测行动项目类型（药物治疗、观察/实验、成像、咨询、营养、治疗程序、其他诊断程序或其他）。跨度标签被编码为IOB2，行动项目标签被编码为分类。然后用CRF负对数似然损失对两个头（跨度类型和行动项目类型）进行模型训练。该模型使用TensorFlow[20] 2 Keras API和TensorFlow模型园[21]。对于推理，跨度类型取自跨度类型CRF头的Viterbi解码，行动项目类型预测为跨预测跨度的最大似然类型，由logits计算（相当于Viterbi解码与对角线过渡矩阵）。更多信息请参见源代码。

所有的模型都以相同的超参数和相同的训练方案进行训练。简而言之，超参数是嵌入大小为250，LSTM隐藏维度为256（每个方向），学习率为1e-3，线性衰减。该方案由两个阶段组成，在第一阶段，模型在启发式标签或人类标签的弱监督下进行预训练，无论是否有增量，共2000步。在第二阶段，模型在有标签的数据上再训练500步，没有增强，学习率恒定为5e-4。在第二阶段，根据跨度和行动项目类型的宏观平均标记水平准确性进行早期停止。

评价

模型和评级是在跨度、标记和行动项目类型层面上进行评估的。在跨度层面，如果在预测跨度和地面真实跨度之间发现任何重叠，则认为跨度正确。每个跨度只能匹配一个其他跨度，最大的重叠跨度被认为是匹配的。在标记层面，如果与跨度类型相匹配，每个非空间标记被认为是正确的。对于行动项目类型，如果跨度匹配（通过跨度标准），并且预测的行动项目类型与基础事实相同，则认为类型是正确的。然后，计算精度、召回率、Jaccard和F1分数。使用聚类引导法计算指标的95%的百分位数置信区间，在笔记层面进行聚类。

鸣谢

作者要感谢刘云、Aviel Atias、Doron Sharabani和Eyal Marcus在项目期间的有益讨论以及他们对手稿的评论。作者受雇于谷歌公司，并拥有Alphabet公司的股权。

代码可用性

代码是可在在GitHub https://github.com/google-research/google-research/tree/master/assessment_plan_modeling，采用Apache 2.0许可。该资源库包含了从笔记和评级中生成TensorFlow例子所需的代码，并针对地面真相训练模型。

数据可用性

注释可在Zenodo网站上获得：https://doi.org/10.5281/zenodo.6413405。注释以CSV的形式提供，每行包含一个注释的跨度。每一行都包含原始注释行ID（来自MIMIC-III注释事件表）、跨度的字符索引、跨度和行动项目类型以及评估者的唯一ID。注释的分层是指训练、验证和测试集。测试集被进一步划分为地面真实和其他评分者。

参考文献

1. Lindenauer, P. K., Lagu, T., Shieh, M.-S., Pekow, P. S. & Rothberg, M. B. Association of diagnostic coding with trends in hospitalizations and mortality of patients with pneumonia, 2003-2009.*JAMA* **307**, 1405-1413 (2012).
2. Savova, G. K. *et al.* Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications.*J. Am.Med.Inform.Assoc.* **17**, 507-513 (2010).
3. Aronson, A. R. & Lang, F.-M.MetaMap的概述：历史视角和最新进展。*J. Am.Med.Inform.Assoc.* **17**, 229-236 (2010).
4. Mottaghi, A., Sarma, P. K., Amatriain, X., Yeung, S. & Kannan, A. Medical symptom recognition from patient text：*arXiv [cs.CL]* (2020).
5. Feder, A. *et al.* Active deep learning to detect demographic traits in free-form clinical notes. *J.生物医学。Inform.***107**, 103436 (2020).
6. Neumann, M., King, D., Beltagy, I. & Ammar, W. ScispaCy：用于生物医学自然语言处理的快速和稳健模型。 *arXiv [cs.CL]* (2019)。
7. Uzuner, Ö., South, B. R., Shen, S. & DuVall, S. L. 2010 i2b22/VA对临床文本中概念、断言和关系的挑战。*J. Am.Med.Inform.Assoc.* **18**, 552-556 (2011).
8. Wu, S. *et al.* Deep learning in clinical natural language processing: a methodical review.*J. Am.Med.Inform.Assoc.* **27**, 457-470 (2020).
9. Zhao, J., Agrawal, M., Razavi, P. & Sontag, D. Directing Human Attention in Event Localization for Clinical Timeline Creation.**149**, 80-102 (2021).
10. Jain, S.等人. RadGraph：从放射学报告中提取临床实体和关系。*arXiv [cs.CL]* (2021)。
11. Mullenbach, J. *et al.* CLIP: A Dataset for Extracting Action Items for Physicians from Hospital Discharge Notes. *arXiv [cs.CL]* (2021).
12. Johnson, A. E. W. *et al.* MIMIC-III, a free accessible critical care database.*Sci Data* **3**, 160035 (2016).
13. Hochreiter, S. & Schmidhuber, J. 长短期记忆.*Neural Comput.***9**, 1735-1780 (1997).
14. Huang, Z., Xu, W. & Yu, K. Bidirectional LSTM-CRF Models for Sequence Tagging. *arXiv [cs.CL]* (2015).

15. Feng, S. Y. *et al.* A Survey of Data Augmentation Approaches for NLP. *ArXiv [cs.CL]* (2021).
16. Fries, J. A. *et al.* Ontology-driven weak supervision for clinical entity classification in electronic health records.*Nat.Commun.***12**, 2017 (2021).
17. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv [cs.CL]*（2018）。
18. Raffel, C. *et al.* Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *arXiv [cs.LG]*（2019）。
19. Mikolov, T., Chen, K., Corrado, G. & Dean, J. Efficient Estimation of Word Representations in Vector Space. *arXiv [cs.CL]* (2013).
20. Martín Abadi*等人*，TensorFlow：异构系统上的大规模机器学习。(2015).
21. Li, J. TensorFlow Model Garden. https://github.com/tensorflow/models (2020).

表和图

表1

	行动项目类型 (1063)						跨度水平 (2016)						代币水平 (2016)					
	微型-F1			宏观-F1			微型-F1			宏观-F1			微型-F1			宏观-F1		
	平均值	CI (95%)		平均值	CI (95%)		平均值	CI (95%)		平均值	CI (95%)		平均值	CI (95%)		平均值	CI (95%)	
评级	0.744	0.705	0.778	0.807	0.778	0.834	0.847	0.822	0.870	0.841	0.817	0.864	0.853	0.819	0.883	0.860	0.832	0.885
评级+Aug	0.757	0.726	0.787	0.806	0.782	0.830	0.866	0.842	0.888	0.862	0.836	0.885	0.862	0.832	0.888	0.872	0.845	0.895
预培训	0.734	0.698	0.769	0.786	0.755	0.815	0.849	0.827	0.870	0.843	0.823	0.863	0.854	0.825	0.880	0.863	0.840	0.885
预训+扩增	0.771	0.744	0.797	0.814	0.786	0.840	0.883	0.865	0.898	0.878	0.860	0.894	0.874	0.847	0.899	0.881	0.855	0.899
中位数评价者	0.842	0.820	0.861	0.885	0.864	0.902	0.911	0.898	0.924	0.906	0.891	0.919	0.910	0.892	0.927	0.916	0.900	0.932
启发式	-	-	-	-	-	-	0.668	0.615	0.716	0.756	0.717	0.789	0.634	0.573	0.695	0.747	0.710	0.782

表1-模型性能。在不同的训练方案中，模型性能为F1得分。Aug--数据增强。宏观和微观分别表示简单和比例加权平均数。CI(95%)表示95%百分位数的自举置信区间，集中在笔记中。括号内的数字表示测试集中跨度的总数。有关跨度和行动项目类型的详细情况，见补充表2。

表2

问题标题	共同行动项目	
"神经病学"	"神经系统检查q：4小时"（9.7%）。	"Dilaudid prn" (0.5%)
	"percocet prn" (0.4%)	"ICP监控" (0.36)
"呼吸衰竭"	"每日CXR" (0.6%)	"vap bundle" (0.2%)
	"F/U文化" (0.2%)	"继续机械通风"（0.3%）。
"贫血"	"瓜沥大便"（3.9%）。	"趋势HCT" (2.2%)
	"继续PPI" (0.5%)	"活动类型和屏幕"（1.3%）。
"心血管"	"阿司匹林"（11.6%）。	"β-阻断剂"（7.4%）。
	"他汀类药物"（4.5%）。	"全面抗凝"（1.6%）。
"肺部"	"是" (12.8%)	"气管"(2.5%)
	"nebs" (0.5%)	"Cont ett（呼吸机模式：CPAP+PS）"（1.8%）。
"传染病"	"检查文化" (13.0%)	"periop abx" (0.6%)
	"vanco" (0.4%)	"今天的痰液培养"（0.4%）。
"咨询"	"神经外科"（9.6%）。	"P.T" (8.6%)
	"CT手术"（8.2%）。	"神经病学"（5.9%）。
"ppx"	"ppi" (13.1%)	"肠道调理"（12.5%）。
	"pneumoboots" (8.0%)	"肝素平方米"（4.6%）。
"急性肾衰竭"	"肾上药"（4.0%）。	"避免肾脏毒素" (3.0%)
	"趋势cr" (2.0%)	"F/U肾脏复查" (1.4%)
"糖尿病"	"iss" (2.4%)	"嘶嘶声"（1.4%）。
	"糖尿病饮食"（1.2%）。	"qid手指棒" (1.2%)

表2--常见状况的共同行动项目。对于每个问题类型（左栏），都有共同的行动项目（右栏）。括号内的百分比表示与指定问题标题相关的跨度中提到这一确切行动项目的百分比。跨度以 "原样 "呈现，无歧义或缩略语扩展。

图一

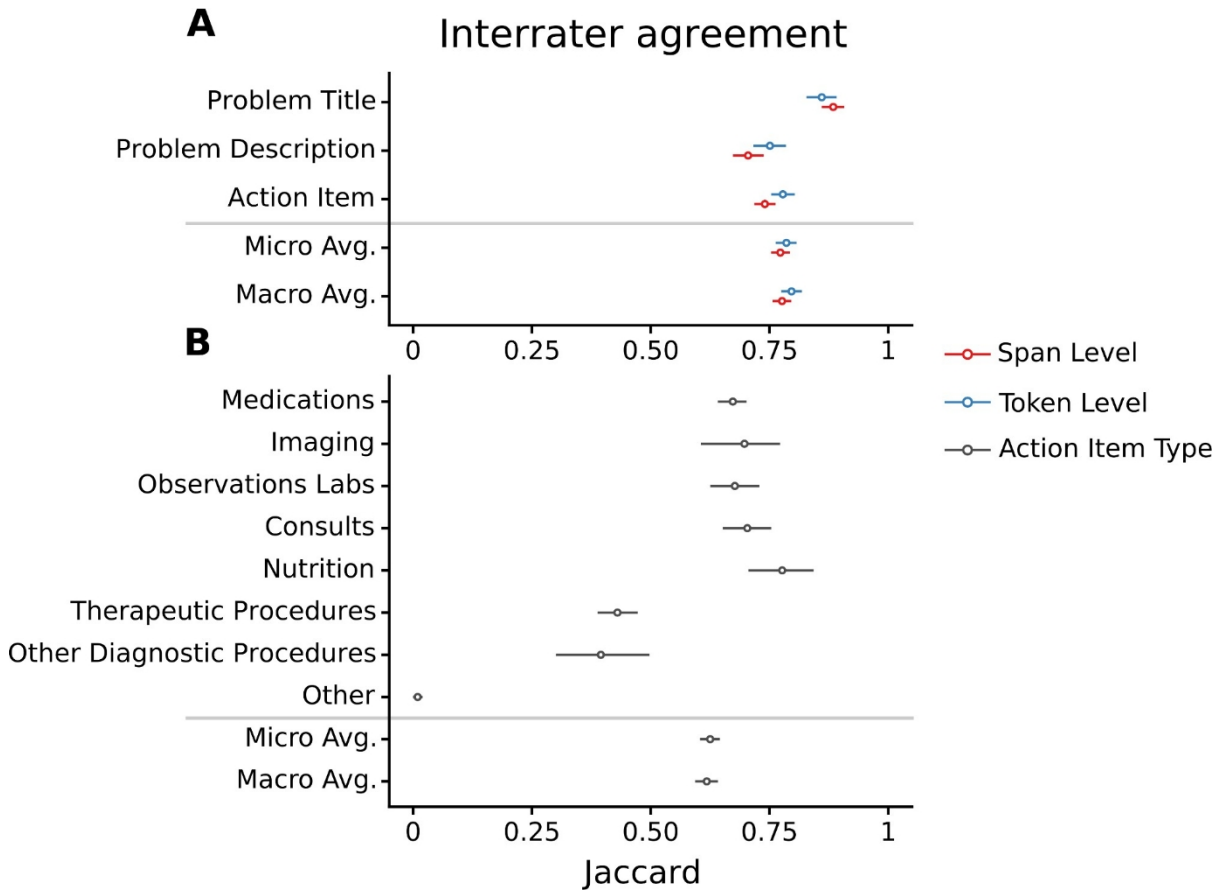

图1--评分者之间的一致性。测评者之间的一致性以测评者之间的平均成对雅卡德指数来衡量。在跨度类型（A）和行动项目类型（B）之间测量一致性。跨度类型的雅卡德指数在跨度（红色）和标记（蓝色）层面上显示。所显示的置信区间是95%的百分位数自举区间，在不同的笔记中聚类。宏观和微观平均数分别表示简单和比例加权的平均数。

图2

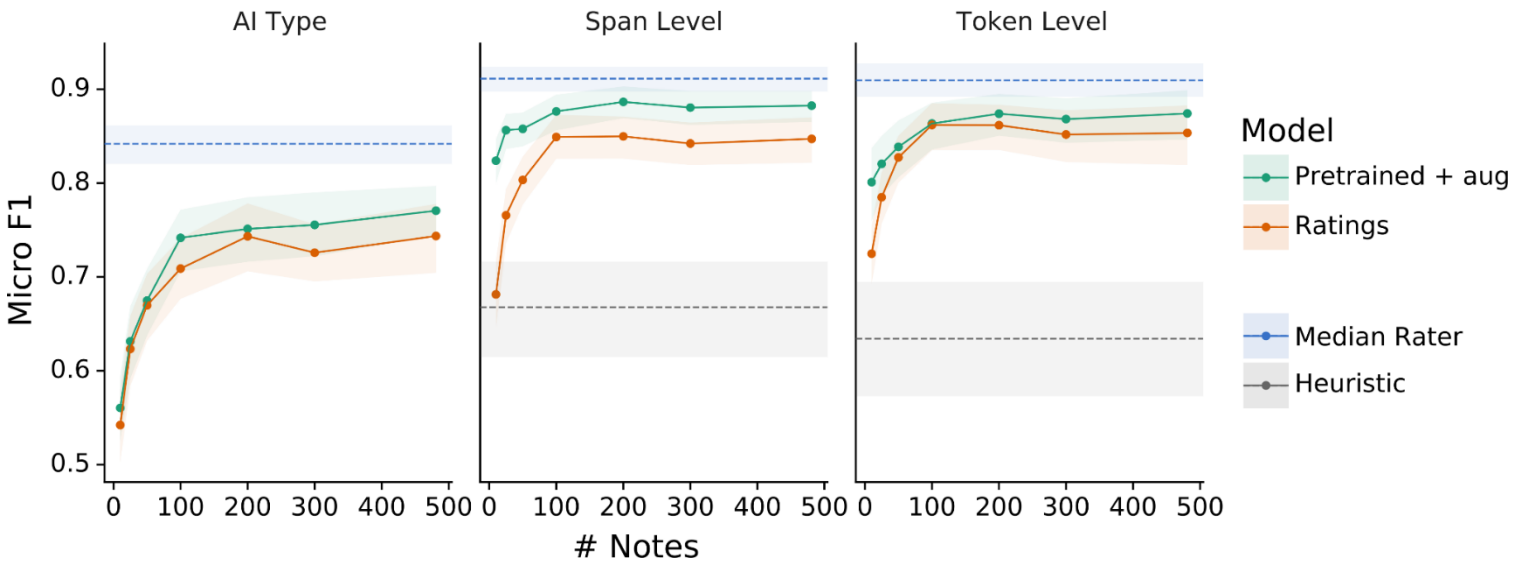

图2 - 有限数据下的表现。模型表现为不同行动项目类型（AI类型，左边）或跨度类型在跨度（中间）或标记（右边）层面的微观平均F1得分。每个点代表了给定模型的微观F1得分（按颜色），在特定的音符数量。所有面板中最右边的数据点代表完整的训练数据大小（481个音符）。来自聚类引导的95%的置信区间表示为与各自线条相同颜色的浅色阴影条。启发式表示临床启发式的表现（对行动项目类型类型不可用，见方法）。

图3

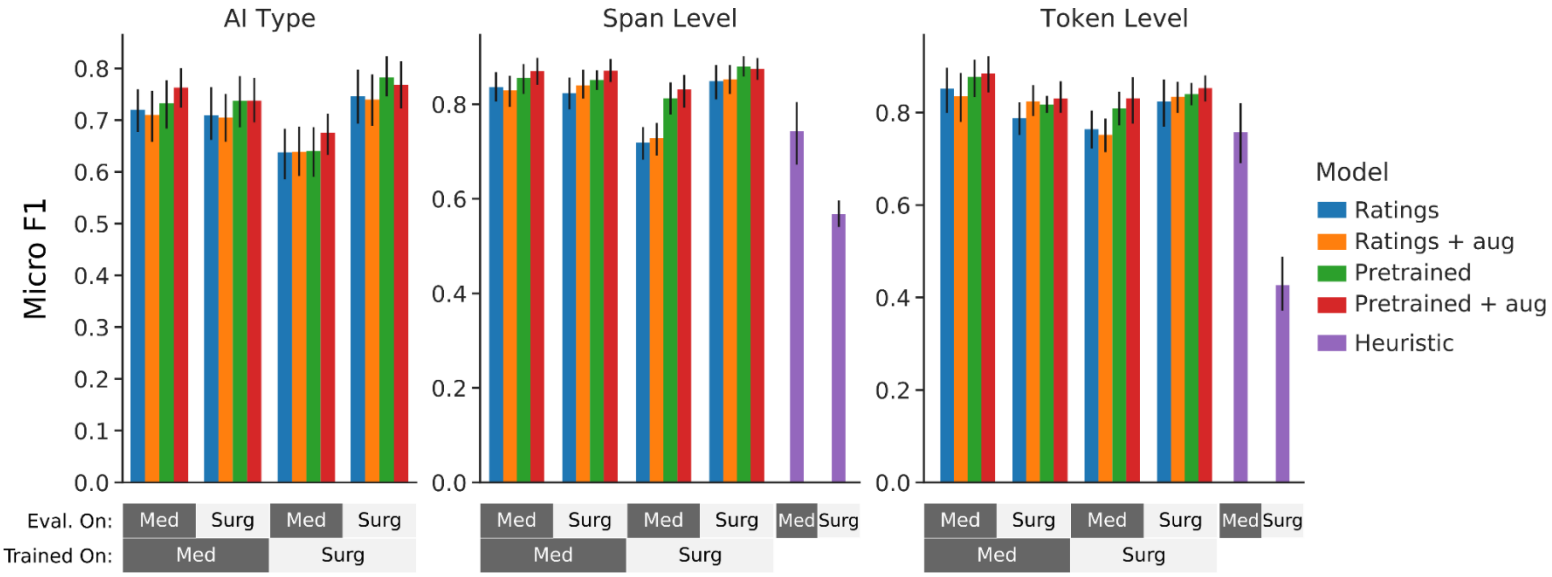

图3 - 部门之间的通用性。在一个部门的笔记上训练的模型性能为微观平均F1分数（底部的条带），在另一个部门的笔记上评估（上面的条带）。在有无弱监督（预训练）和有无增强（缩写为AUG）的情况下，对模型性能进行了比较。临床启发式是用来比较的（紫色）。性能是在标记水平（右）和跨度水平（中）以及行动项目类型（左，人工智能类型）上测量的。误差条表示来自笔记的聚类引导的95%百分位数置信区间。Eval. on - 评价过的，Med - 医疗服务，Surg - 外科服务，AI type - 行动项目类型。