arXiv:2301.09350v2 [cs.CL] 5 Oct 2023

用于生物医学文献细粒度语义索引的弱监督深度学习大规模研究

Anastasios Nentidis[a, b, *], Thomas Chatzopoulos[a, c], Anastasia Krithara[a], Grigorios Tsoumakas[b], Georgios Paliouras[a]

[a] *Institute of Informatics and Telecommunications, NCSR Demokritos, Athens, Greece*
[b] *School of Informatics, Aristotle University of Thessaloniki, Thessaloniki, Greece*
[c] *Department of Computer Engineering and Informatics, University of Patras, Patras,Greece*

摘要

目的：生物医学文献的语义索引通常以 MeSH 描述符为单位进行，多个彼此相关但又各不相同的生物医学概念往往被归为一组，并作为单一主题处理。本研究提出一种新方法，用于在 MeSH 概念层面对主题标注进行自动细化。方法：由于缺乏标注数据，我们依靠基于文章摘要中概念出现情况的弱监督，并辅以基于词典的启发式方法。此外，我们还研究了深度学习方法，并针对这一任务的特定挑战作出设计选择。我们在大规模回顾性场景中评估了新方法，所用概念均已晋升为描述符。结果：在实验中，概念出现是效果最强的启发式方法，在多个标签上的宏平均F1分数约为 0.63。所提出的方法在此基础上又提升了 4 个百分点以上。结论：结果表明，概念出现是细化 MeSH 概念层面粗粒度标签的一种强有力的启发式方法，而所提出的方法还能进一步提升效果。

*通讯作者*Email address:* nentidis@csd.auth.gr, tasosnent@iit.demokritos.gr (Anastasios Nentidis)[1]本文为发表于《生物医学信息学杂志》的一篇文章的预印本。最终认证版本可在线获取：https://doi.org/10.1016/j.jbi.2023.104499

关键词：语义索引、医学主题词表（MeSH）、生物医学文献、弱监督、深度学习*2010*年数学学科分类：00-01，99-00

---

1. 引言

生物医学文献是不断扩展的生物医学知识宝库。从如此丰富的文献中，为特定主题找出相关文章是一项真正的挑战。为此，美国国家医学图书馆（NLM）使用医学主题词表（MeSH）叙词表中的主题描述符为 MEDLINE数据库中的引文编制索引[2]。这一语义索引过程使 MEDLINE/PubMed[3] 能够提供先进的语义搜索策略，检索与感兴趣主题相关的引文，并解决生物医学概念的同义关系和多义性等问题。

医学主题词表包含三万多个描述符，例如疾病和化学物质，并按层级组织，因此大多数描述符至少是一个更宽泛描述符的子类别。例如，如图1所示，“反射性癫痫”是更宽泛描述符“癫痫”的子类别。其中一些描述符涵盖若干相关但彼此不同的下位*MeSH* 概念[4]，，这些概念被定义为同义词集合。通常，这些概念的范围比描述符的主要主题更窄[5] 。例如，“反射性癫痫”包含一个首选概念，即描述符名称及其同义词“反射性癫痫”，同时还包括更窄概念的其他术语，例如“听源性反射性癫痫”和“触觉反射性癫痫”；这些术语本身并非医学主题词表描述符。

语义索引过程是在描述符层级进行的，因此无法区分这类更具体的概念。因此，语义搜索

[2] https://meshb.nlm.nih.gov/ [3] https://pubmed.ncbi.nlm.nih.gov/

[4] https://www.nlm.nih.gov/mesh/concept_structure.html [5] 在某些情况下，下位概念的范围可能比主题更广。

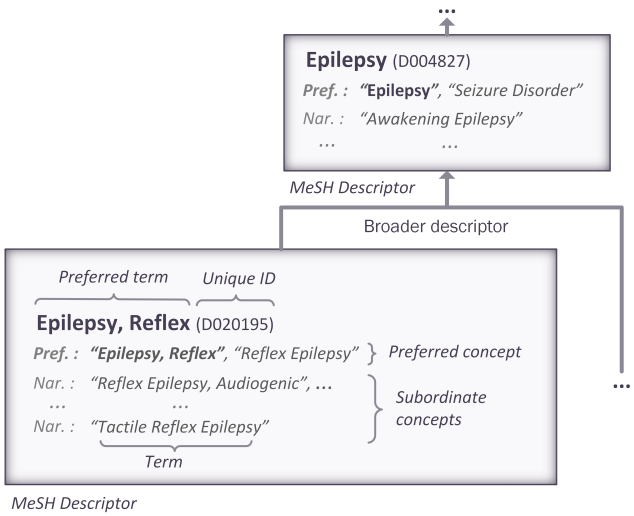

图1：“反射性癫痫”医学主题词表描述符的结构及其层级位置。

无法用于如此细致的层级。本研究旨在实现这种细粒度语义搜索，为生物医学文献提供一种细粒度语义索引（$FGSI$）方法，即将 MeSH 标签从描述符层级自动细化到每个更具体的概念层级。例如，我们希望区分已标注“反射性癫痫”的文章中，哪些属于“听源性癫痫”，哪些属于“触觉反射性癫痫”或“反射性癫痫”的其他子类别。

在 MEDLINE/PubMed 中，按医学主题词表概念层级进行索引超出了当前实践；若强制采用这种方式，索引所用的不同主题数量几乎会翻倍。鉴于生物医学文献的数量不断增长，手动完成这项额外工作并不现实。此外，最先进的自动主题建议系统侧重于已知的医学主题词表主题，并且通常要求每个主题都有多个标注训练示例 [1]。因此，如何为概念层级的新细粒度主题提供可靠的主题建议，仍是一个尚待解决的研究问题。缺乏用于方法开发和评估的标注数据，是这项 $FGSI$ 任务面临的一大挑战。

在这项工作中，我们提出了$Deep\ Beyond\ MeSH$ (DBM)，这是一种用于

基于深度预训练语言模型（即$PubMedBERT$ [2]）细化粗粒度医学主题词表注释的新方法。为克服监督训练中标注数据不足的问题， DBM 采用弱监督，探究此类深度学习（DL）模型能否适应启发式标签。该弱监督方法基于概念出现 (CO)[6], ，其定义为一项信息抽取任务：识别文本中的生物医学命名实体，并将其映射到规范化语义系统中的特定概念，例如统一医学语言系统[7]。

先前的实验采用小型真实数据集来完善两个特定的医学主题词表描述符，以此验证这一思路 [4], ，结果表明， CO 可为 $FGSI$ 任务奠定坚实基础。然而，这一启发式方法仍需在涵盖多个医学主题词表描述符的大型真实标注数据集上进行测试。为此，我们基于 MEDLINE 中医学主题词表索引的发展历程设计了一种回顾性场景，并在此介绍$Retrospective\ Beyond\ MeSH$（$RetroBM$）。这种新方法为多个独立概念构建大规模数据集；这些概念随着时间推移被提升为新的细粒度医学主题词表描述符，也就是说，这些描述符各自涵盖一个 MeSH 概念。

简而言之，推动本研究的主要问题如下：

- CO 是否是大规模概念级 $FGSI$ 的一种有效启发式方法？将CO 与其他启发式方法结合，能否进一步改进？

- 是否可以不依赖真实标注训练数据，而将上述启发式方法用作弱监督来训练用于 $FGSI$ 的深度学习模型？这些模型的预测性能能否超过启发式方法本身？

- 能否利用医学主题词表向细粒度标签的自然扩展，对自动化 $FGSI$ 方法进行大规模回顾性评估？

本文其余部分安排如下。第2节中，我们将介绍一些

[6]这基于 $MetaMap$ 工具 [3]对概念的识别。
[7]`https://www.nlm.nih.gov/research/umls/index.html`

背景以及生物医学文献 $FGSI$ 方面的先前研究。第3节中，我们介绍两种新方法： $RetroBM$ 用于构建评测数据集， DBM 用于在没有真实标注数据的情况下为文档编制细粒度标签索引。第4节中，我们将介绍本文的实验与结果。最后，第5节总结本文，并讨论本文对解决研究动机中提出的问题所作的贡献。

2. 背景与相关工作

与层次文本分类这一更广泛的问题不同， $FGSI$ 任务中已有粗粒度类别标签，该任务在本文中被定义为对现有粗粒度标签进行细化。近期，Mekala等人 [5] 针对这一特殊情况提出了$Coarse2Fine$ ($C2F$)方法。$C2F$ 该方法将细粒度标签在文本中的字面出现作为初始弱细粒度标签，并将其与真实的粗粒度标签结合起来，训练一个由标签条件化的 GPT2 [6] 模型，为任意给定标签生成相关文档。该模型用于为每个细粒度标签生成新文档，并以此训练$BERT$ 分类器 [7]。

数十年来，美国国家医学图书馆的人工专家索引员一直负责对 MEDLINE/PubMed 中的生物医学文献进行语义索引。自 2002 年起，医学文本索引器 (*MTI*) 工具[8], 开始为人工专家提供一些自动生成的主题建议；自 2012 年起，BioASQ 挑战赛 [8]也围绕这一任务组织了专门的共享任务。尽管这项任务面临诸多挑战，例如类别数量庞大，但数百万篇人工索引文章的可用性以及先进的深度学习架构，推动了机器学习模型逐步发展并取得了令人满意的结果 [1, 9, 10]。因此，美国国家医学图书馆逐步转向对所有 MEDLINE 文章进行全自动索引[9]。然而，受限于缺乏充足的数据，细化到医学主题词表概念层级的索引仍未能实现。

[8] https://lhncbc.nlm.nih.gov/ii/tools/MTI.html

[9] https://www.nlm.nih.gov/pubs/techbull/nd21/nd21_medline_2022.html

早期针对生物医学文献 $FGSI$ 的研究致力于通过添加医学主题词表限定词和医学主题词表补充概念记录（SCR）[10] [3], 来丰富基于描述符的标签，但这些做法并未解决描述符过于粗略、将不同概念归为一类的问题。[11]首次指出了在医学主题词表概念层级对生物医学文献进行索引的重要性。近期，我们提出了 $Beyond\ MeSH$ 方法 [12, 4] ，用于在医学主题词表概念层级自动细化医学主题词表主题标注。在该研究中，我们引入了概念出现（CO）启发式方法，为 $FGSI$分配弱标签。该启发式方法在更简单的基于词典的方法基础上进行了扩展，使用$MetaMap$ 工具来处理信息抽取方面的挑战，例如词形变化、多词术语和概念消歧 [13]。在小型数据集上使用常见机器学习算法，对两个特定描述符进行细化的实验表明， CO 启发式方法为这两个用例提供了强有力的基线，只有部分经过训练的分类器能够超越它。

在这项早期工作的基础上，本文在大规模场景中对若干医学主题词表描述符的 promising CO 启发式方法进行了全面评估，同时首次评估了深度学习方法在此任务中的表现，揭示了 $FGSI$ 任务面临的特殊挑战，并提出了相应的解决方案。为了构建大规模实验场景，我们回顾医学主题词表描述符的演变历程，重点关注那些随着时间推移逐渐成为描述符的医学主题词表概念。

3. 材料与方法

本节将介绍：a) $Retrospective\ Beyond\ MeSH$ ($RetroBM$) 方法，用于构建真实标注数据集，以便大规模评估细粒度语义索引（$FGSI$）；b) $Deep\ Beyond\ MeSH$(DBM) 方法，用于将粗粒度医学主题词表标签细化为细粒度标签，而这类标签目前没有可用的真实标注数据。

[10] https://www.nlm.nih.gov/mesh/intro_record_types.html

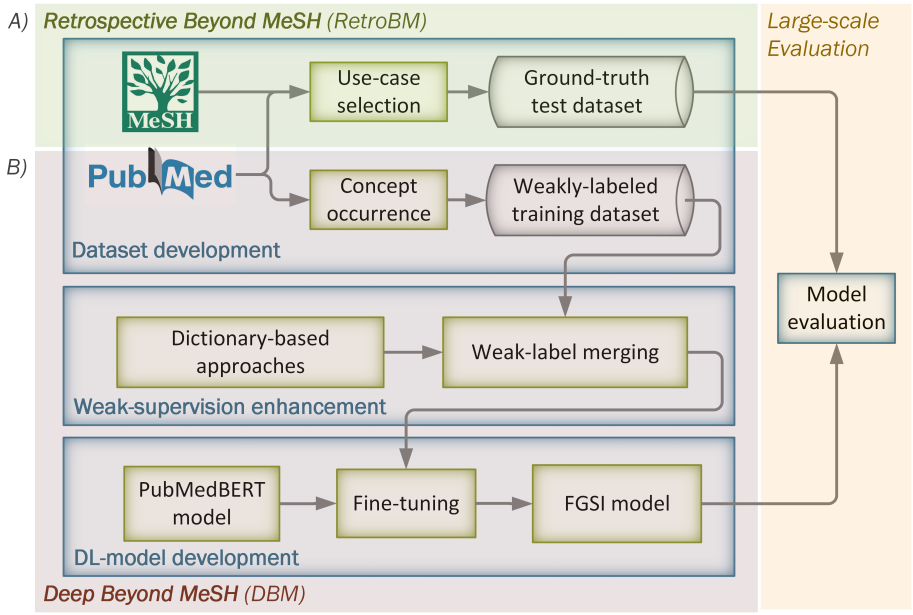

图 2：A）用于构建大规模真实标签 $FGSI$数据集的 $RetroBM$ 方法；B）用于弱监督 $FGSI$ 模型开发的 DBM 方法。$RetroBM$ 数据集用于大规模评估 DBM 模型。

如图 2A 所示， $RetroBM$ 首先通过回顾医学主题词表的演变，选取一些适合作为描述符的 MeSH 概念用例。随后，它基于 MEDLINE/PubMed 中已有的人工标注，为细化这些粗粒度描述符这一任务构建真实标注数据集 $FGSI$ 。

另一方面，DBM（图 2，B）无需真实标签，即可为任何下位医学主题词概念提供 $FGSI$，无论其是否被提升为描述符。 DBM 分为三个部分：a）基于概念出现 （CO）构建弱标注训练数据集；b）采用基于词典的方法增强弱监督；以及 c）通过微调预训练的$PubMedBERT$模型来开发 $FGSI$ 模型。

最后，真实标注的 $RetroBM$ 数据集既用于验证 DBM 方法，也用于评估弱监督 DBM 模型，以应用于

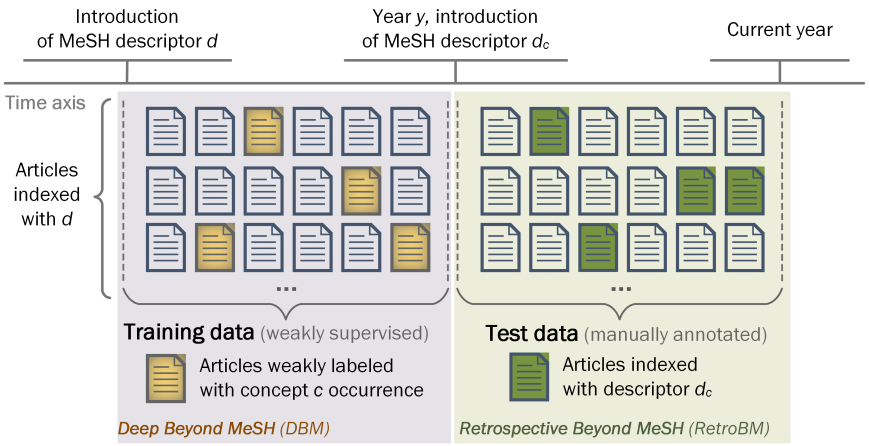

图3：本研究采用的回顾性评估场景基于将 MeSH 概念 c 提升为描述符 d_c，以开发$FGSI$ 数据集。

所选用例。这些方法已在 GitHub[11] 上公开提供。

3.1. 数据集开发

尽管所提出的 DBM 方法通过开发弱标注数据集，避免了在模型开发中使用真实标签的需要，但验证和评估此类 DBM模型仍需要一些真实数据。因此，我们还介绍了 $RetroBM$，这是一种基于 MEDLINE/PubMed 中既有 MeSH 索引演变的方法，用于开发一些带有真实标签的 $FGSI$ 数据集。

这两种方法在一种回顾性评估场景中互为补充：该场景基于将 MeSH 概念 c 提升为专用描述符d_c （见图3），发生于某一年y 。在此场景中， $RetroBM$ 使用在y 年之后以d 编目的文章构建真实测试集数据集，而 DBM 使用在y年之前以 d编目的文章构建相应的弱标注训练数据集[12]。这样做是为了简化流程，并与前瞻性

[11] https://github.com/tasosnent/DBM [12]对于使用 d 编入索引的文章，我们同时考虑直接标注了描述符d 的文章，以及在医学主题词表层级中标注了比 d 更窄的任一描述符的文章。

使用场景保持一致：我们基于现有的、以 d编目的文章开发模型，并将其用于新文章。不过，新文章也可以进行弱标注，并由 DBM 用于模型训练。

3.1.1. Retrospective Beyond MeSH（$RetroBM$）

MEDLINE/PubMed 为若干 MeSH 描述符的父子配对提供了标注数据。然而，并非所有此类配对都适合作为评估 $FGSI$ 任务的实际用例；该任务需要将现有的粗粒度标注细化为细粒度标注。例如，有些情况下，子描述符的年代早于其父描述符，因此子描述符的标注早已可用，实际上无需细化父描述符的标注。为此， $RetroBM$ 依托 MeSH 的演变，识别适用于 $FGSI$ 评估的概念细化案例，使其既合适又符合实际情况。

在y年对 MeSH 叙词表进行年度更新时，如果粗粒度描述符d 的下位概念 c 在文献中的重要性日益提升，就可能被升格为新的 MeSH 描述符d_c。以图4为例，“A型尼曼-匹克病”在2007年y之前一直是“尼曼-匹克病”描述符d的一个较窄下位概念（c），之后则被升格为一个独立的描述符d_c。因此，自2007年以来，与“A型尼曼-匹克病”相关的文章都使用新的专用描述符d_c进行索引。在此类情况下，如果c 是新描述符d_c所包含的唯一概念，那么使用d_c 进行索引就等同于使用c进行索引。因此，c 升格为d_c 之后人工索引的文章，其 $FGSI$ 真实标签均包含c ， $RetroBM$ 可据此构建测试集，如图3所示。

具体而言， $FGSI$ 通过c 开展的任务是：针对每篇已使用d （例如“尼曼-匹克病”）编目过的文章，判断其是否与c （例如“A型尼曼-匹克病”）有关。因此，我们只考虑使用d 编目的文章，并将这些文章称为 $FGSI$ 通过c开展任务的有效 文章。由 $RetroBM$ 生成的真实数据集包含所有在y之后标注的这些有效文章，其中使用d_c 编目的文章为正例，其余文章则是 $FGSI$ 通过c开展任务的负样本。

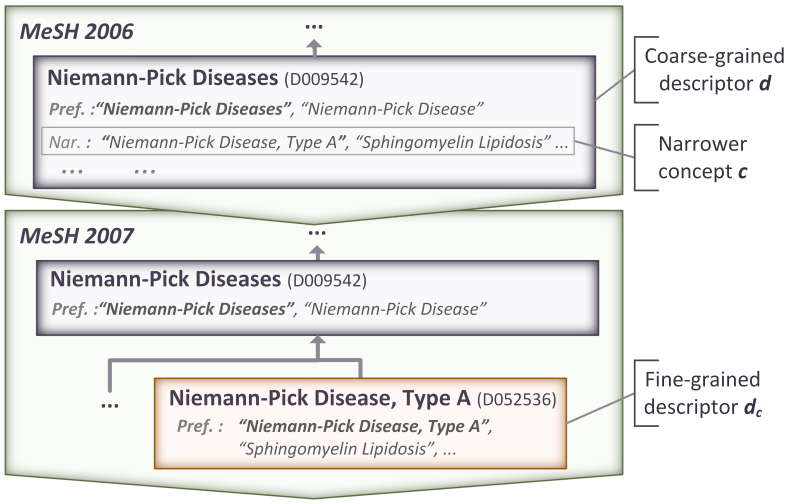

图 4：“A型尼曼-匹克病”从描述符“尼曼-匹克病”的下位概念晋升为 2007年新增的专用描述符。

当然，使用d编目并不足以满足使用d_c编目的条件，但却是必要条件，因为d_c是d的子类别。另一方面，未使用d 编目的文章（对于c而言是无效 的）不属于$FGSI$ 任务的研究范围，因为它们也不可能与其子类别d_c 相关。

基于上述想法， $RetroBM$ 回顾性地寻找此类概念晋升事件，并构建相应的数据集，提出了两项标准。首先， $RetroBM$ 依赖于我们近期提出的一种关于新医学主题词表描述符溯源的类型体系 [14]。在这项工作介绍的类型中，名为细分[13] 的一类表示下位概念c被提升为粗粒度描述符d 的子类别d_c ； 此前，这些下位概念由该描述符收录。这一类型符合图 4 所示的情形， $RetroBM$据此识别出合适的案例（标准 *1*）。对于d_c 而言，存在其他上位概念并不影响这一情形，因为c 始终只属于单个描述符d_c。不过，通过考虑其他类型的演变，RetroBM 可以扩展至类似问题，例如使用新兴概念、具有

[13]我们将溯源码为 1.2 的案例称为 Nentidis *et al.* [14] 中介绍的案例。

多个上位概念的概念进行索引，甚至使用概念组进行粗粒度索引。

除了溯源之外，每个新的描述符 d_c 在 $FGSI$ 中也需要达到细粒度（标准 *2*），具体包括：a) d_c 应专门对应相应的提升概念 c，而不同时涵盖其他下位概念，从而确保所有标注了d_c 的文章实际上都与 c 相关（标准 *2.1*）。b) d_c 应是 MeSH 层级中的叶节点。存在表示子类别的更窄描述符，说明该描述符涵盖了多个概念，因此不具备细粒度（标准 *2.2*）。

3.1.2. 弱标注训练数据集

DBM 方法的第一步使用标注了任意粗粒度 d 的文章，为 $FGSI$ 生成弱标注训练数据，其中包含任意下位概念 c，无需人工进行 d_c 标注。为了生成弱标签， DBM 采用了 $Beyond\ MeSH$ [12, 4],中提出的启发式方法：根据 $MetaMap$ 工具 [13]识别出的概念 c 是否出现在文章的标题或摘要中来进行判断。 DBM 适用于任何至少包含一个更窄下位概念 $c^{'}$的粗粒度描述符 $d^{'}$，以预测带有 $c^{'}$的 $FGSI$ 的标签，无论$c^{'}$ 是否已提升为描述符。例如，在 MeSH 2020 中，约有 7,000 个此类更窄的 MeSH 概念具备充足的弱监督信息，可用于应用 DBM[14]。然而，要验证这一标注过程的准确性，还需要$c^{'}$的真实标签。因此，在本研究中，我们针对由 $RetroBM$ 选定的用例采用 DBM，如图3的评估场景所示。

3.1.3. 多标签设置

$FGSI$ 任务本质上是多标签的，因为同一篇文章可能与多个不同的粗粒度和细粒度标签相关。此外，在 $FGSI$中，我们希望任何方法都能将单个描述符（例如“癫痫”）细化为其任一同级下位概念（例如“听源性癫痫”、“触觉反射

[14]在此估算中，我们考虑描述符 $d^{'}$ 下任何较窄的概念 $c^{'}$，该概念至少出现在 10 篇文章中，且满足：a) 带有 $d^{'}$ 注释；b) 有 $c^{'}$ 出现。

性癫痫”等），而这些概念通常并非互斥。在本研究中，为与回顾性评估场景保持一致，我们根据推广年份y整理了由 $RetroBM$ 和 DBM 开发的 $FGSI$ 数据集，将给定年份中对应任一用例的所有有效文章汇集在一起，并据此分配多标签标注。

3.2. 弱监督增强

DBM 方法的第二部分是增强 CO提供的原始弱监督。本节中，我们研究了一系列基于词典的变体，以及将它们组合起来是否能够提升弱标签的质量。在少数验证案例中开展的本研究结果见第 4.2 小节，可据此选择在 DBM 方法的第二步中采用的策略，以增强弱监督。具体而言，借用数据编程 [15],中引入的术语，我们将这些基于词典的变体以及 CO 称为用于 $FGSI$ 任务的标注函数（LF）。

如果与概念c相关联的任何词典元素在文章的标题或摘要中逐字出现c ，每个基于词典的标注函数（LF）就会为该文章中的概念c分配一个标签[15]。在 名称精确匹配 LF 中，仅匹配c 的名称（例如，“A型尼曼-匹克病”）；而在 同义词精确匹配 中，则匹配其同义词（例如，“经典型尼曼-匹克病”等）。随后，为了提高召回率（见表 1），引入了这两个 LF 的六种变体[16], 。具体而言，首先将词典元素转换为小写（名称小写化、同义词小写化），然后去除标点符号（名称去标点、无标点同义词），最后将短语拆分为单词元元素（名称词元、同义词词元）。

这些基于词典的标注函数依赖的信息与 CO 使用的相同，基本上就是c 中的术语和文章文本。不过，它们可能会产生

[15]含有ndictionary个元素（e_i）的标注函数通过以下查询实现： e_1 或 e_2 .....或 e_n。 [16]应用后六个标注函数时，还会将文章的标题和摘要转换为小写。

表1：“A型尼曼-匹克病”标签的各个标注函数（LF）的词典元素。

标注函数（LF）	缩写	词典元素
name exact	NE	“A型尼曼-匹克病”
synonyms exact	SE	“经典型尼曼-匹克病”，.....
name lowercase	NL	“A型尼曼-匹克病”
synonyms lowercase	SL	“经典型尼曼-匹克病”，...
name no punct	NNP	“A型尼曼-匹克病”
synonyms no punct	SNP	“经典型尼曼-匹克病”，...
name tokens	NT	“尼曼”、“匹克”、“疾病”、“类型”、“A”
synonyms tokens	ST	“经典型”、“尼曼”、“匹克”、“疾病”...

不同的弱标签。这是因为 CO由$MetaMap$计算得出，比精确的术语匹配更为复杂，会考虑词性标注和基于上下文的词义消歧等额外信息 [13]。因此，将这些标注函数与 CO 结合起来可能会带来改进。另一方面，这些标注函数之间的相关性可能会降低集成的准确率。因此，在验证实验中（第4.2小节），通过考察这些标注函数中两个或更多函数的所有组合，来确定适用于DBM的最佳标注函数子集。

为了将标注函数组合成一个集成，我们考察了两种投票分类器：一种基于多数投票 (MV)，另一种基于至少一个 投票 (ALO) [16], ；此外，还考察了一种先进的概率标签建模 (LM)方法，用于组合含噪启发式规则，该方法是在 Snorkel [17]的背景下提出的。 MV 方案规定，只有当大多数标注函数（投票者）都为某篇文章赋予某个标签时，才将该标签赋予这篇文章；这种做法倾向于提高集成的精确率，因为只有少数标注函数预测出的标签会被视为不可靠而忽略。另一方面，在 ALO 方案中，只要任何一个标注函数为某篇文章赋予某个标签，就会将该标签赋予这篇文章。这是 Kambhatla [16],提出的至少 N 个方案的一种特殊情况，其中 $N = 1$。采用这种方法时，即使只有一个标注函数预测出某个标签，该标签也会纳入集成的预测结果，从而倾向于提高召回率。

最后，Snorkel LM 方法将文章的真实标签视为概率生成模型中的一个潜变量，并根据标注函数的加权组合对其进行估计 [15, 17]。最终，如果估计出的标签概率大于 0.5，就会将该标签赋予这篇文章。

3.3. 模型开发

DBM 该方法的最后一部分是为 $FGSI$ 任务开发弱监督深度学习（DL）模型。鉴于其他任务的实验表明，深度预训练模型可能天生具备抵御标签噪声的能力 [18, 19], ，我们据此研究了文本分类领域的一种先进方法是否适用，即对基于$BERT$的预训练模型进行微调；在本研究中，我们使用的是 $PubMedBERT$ 模型，并添加了一个特定于任务的输出层 [2], [7]。具体而言，我们针对若干验证案例进行了实验，以确定 $FGSI$ 任务面临的具体挑战，并采取相应措施加以应对。这些实验既包括本节讨论的基本设计选择的初步实验，也包括第 4.2 小节介绍的验证实验。

在 DBM 中，我们选择多标签场景，即为每个年份 y单独开发一个多标签模型，如图 5[17] 所示。在这种架构中，将每篇文章的标题和摘要拼接并进行分词，然后输入 $PubMedBERT$ 模型。随后，将分类标记 CLS 的嵌入作为输入传入全连接输出层，该层会生成一个大小等于标签数量的向量，如 [20] 所示。这样，不同标签的信息便能整合到同一年份所有用例共享的深层表示中。这尤其适用于同一粗粒度描述符下的兄弟概念，例如图1中的“听源性癫痫”和“触觉反射性癫痫”。初步实验确认，多标签模型至少与采用

[17]可以采用一个适用于多年的通用模型，但这会带来实验上的复杂性，因为同一篇文章可能在一个用例中被划入测试集，而在另一个用例中被划入训练集。

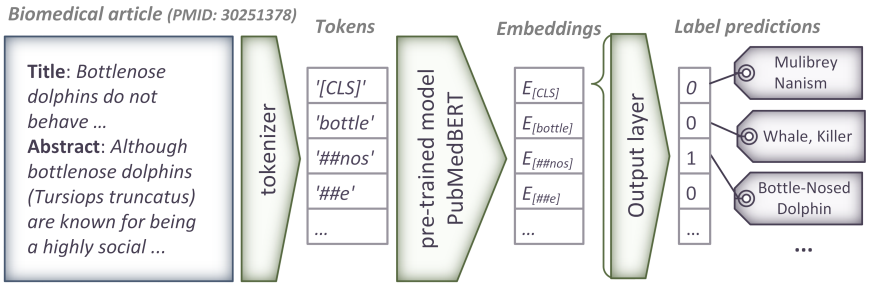

图5：通过 DBM 为单一年份的所有标签开发多标签模型。

一对其余方法的单标签模型效果相当。因此，开发和存储多个单标签模型所需的额外成本并不值得。

关于要微调的具体$BERT$模型，我们选择了$Pub\text{-}MedBERT$ [2], ，这是一种$BERT$ 领域特定变体，采用领域特定词汇表，以便更好地捕捉文本的语义。尽管初步结果中的差异很小，但结果证实$PubMedBERT$ 性能良好，因此我们选择它作为所有 DBM 模型的基础。初步实验还表明，训练轮次的数量会影响模型性能。为此，我们将弱标注训练数据划分为训练部分 和验证部分 （90-10%），仅使用前者进行训练，并在未见过的验证部分中检查损失，以确定损失最小的轮次（最佳轮次）。

为了提高泛化能力，我们还采用了早停策略：一旦验证部分 上的损失开始增加 [21]，便结束微调过程。在这种情况下， $best\ ep$ 模型始终取自早停微调中倒数第二个轮次。此外，我们还考虑了$prev\ ep$ 和 $next\ ep$ 模型，它们分别对应于$best\ ep$之前或之后的轮次。其动机是，性能最佳轮次附近的轮次表现“不完美”，可能是因为它们具有更高的泛化能力。

训练 $FGSI$ 模型面临的一大挑战是标签在两个方面存在不平衡：a) 每个标签的正例数量因用例而异，这使得更受欢迎的细粒度标签占据优势；b) 负样本数量众多，

远多于正例。在极端情况下，大量负样本会使模型无法学会识别正例，因为它们对损失的整体影响微乎其微。

针对标签之间的不平衡，我们考虑了成熟的二元交叉熵（BCE）损失的扩展方法 [22] ，这些方法能够处理多标签文本分类中的不平衡训练数据 [23]。具体而言，重平衡焦点损失（R-BCE-FL）是 BCE 损失的一种扩展，最初用于处理多标签图像分类任务中的标签不平衡 [24] ，最近又扩展到多标签文本分类任务 [23]。R-BCE-FL 通过两种方式修改 BCE。首先，根据相应标签的总体频率对样本进行过采样，以实现重平衡。其次，聚焦参数根据样本的预测概率，提高较难分类样本的权重 [23]。

针对正样本与负样本之间的不平衡，先前研究表明，欠采样有助于解决这一问题 [4]。因此，我们尝试了一种欠采样流程，通过移除负样本，为每个标签达到理想的负样本与正样本比例（$balance\ n$）。这一理想比例无法预先得知，但可以通过验证得出合理估计。若为严格达到某个标签的理想负样本与正样本比例而随机移除负样本，可能会耗尽其他标签的正样本或有效负样本。因此，DBM 的欠采样过程不会移除任何标签的正样本，也不会移除负样本与正样本比例较低的标签的有效 样本。

模型开发过程的多个步骤中都引入了随机性 [25], ，尤其是在划分训练部分和验证部分、对负样本进行欠采样，以及进行随机优化时 [26]。初步实验表明，采用不同的随机种子会产生不同的结果。因此，每项实验均使用六个不同的随机种子重复进行，并采用多数投票法汇总预测结果，从而对模型性能作出更稳定的估计。

3.4. 其他机器学习方法

我们根据近期相关研究（第 2 节），在实验中纳入了三种机器学习方法进行比较。首先，我们考虑了超越医学主题词表 [4]，这是一项先前针对 $FGSI$ 任务开展的研究，旨在细化 MeSH 概念层面的 MeSH 主题标注。沿着这一研究方向，我们针对每个细粒度标签分别开发了逻辑回归（LR）模型，同时考虑词汇特征和语义特征。为与 DBM进行公平比较，我们对负样本采用相同的欠采样方法，并使用六个随机种子重复操作，再以多数投票法汇总结果。我们在弱标注数据的验证部分 （10%）上通过网格搜索选择正则化程度，考察的取值范围从 0.01 到十亿[18]。根据第 4.2 节所述的验证实验，选定了合理的特征数量。

其次，我们还考虑了$Coarse2Fine$ ($C2F$) [5]，这是一项先进研究，旨在将现有的粗粒度标签细化为细粒度类别。尽管 $C2F$ 最初依赖标签名称的字面出现情况，但为公平地与 DBM进行比较，我们为 $C2F$ 提供了相同的弱监督。此外， $C2F$ 处理的是多分类问题，其中每篇文档恰好属于某个粗粒度标签d下的一个细粒度子类别。为处理$FGSI$中“未标注”的样本（这些样本属于d，但不属于其任何子类别），我们为每个d引入一个额外的子类别（$d\ rest$）。不过，在评估 $C2F$ 模型的性能时，会忽略$d\ rest$的预测；与所有方法一样，评估依据的是实际子类别。由于 $C2F$ 通过添加生成样本来处理类别不平衡，我们没有对数据进行欠采样。此外，我们没有尝试不同的随机种子或参数，而是沿用了原始实现[19]。

最后，为了研究按照

[18]具体考虑了以下取值：0.01、0.1、1、10、100、1K、10K、100K、1M、1B。[19] https://github.com/dheeraj7596/C2F

$C2F$生成实例的影响，我们还使用 DBM的实现，基于 $C2F$ 生成的文档作为弱监督，开发了 DBM^{C2F}模型。

3.5. 评估指标

由于精确率（P）和召回率（R）对于 $FGSI$都很重要，因此我们采用F1度量（$F1$）作为总体性能指标，同时考虑不同用例 [27]上的宏平均（$maF1$、 maP、 maR）和微平均（$miF1$、 miP、 miR）。在多标签 $FGSI$ 数据集中，对于特定细粒度标签c 而言$valid$ 的文章并非关注对象，因此在评估该标签时会将其忽略。我们将这一过程称为有效性过滤，并在评估所有方法时采用该过程。

4. 实验与结果

$RetroBM$ 应用于 MeSH 在连续十四年（2006—2019年）中的演变过程，以构建相应的真实标注数据集（测试y）。随后， DBM构建了相应的弱标注训练数据集（弱监督开发y），用于 $FGSI$模型开发（第4.1小节）。在这些数据集中，我们选取了针对 2006年引入的标签的数据集（*2006*年数据集）进行“验证实验”，以确定该方法的架构和配置（第4.2小节）。最后，使用后续年份引入的未见细粒度标签对应的数据集（*2007—2019*年数据集）进行“评估实验”（第4.3小节），旨在确认基于*2006*年数据集所作的选择。

4.1. 回顾性数据集

总计发现 88 个新的细粒度描述符符合 $RetroBM$ 的标准，并据此对2006年至 2019[20]年的 MEDLINE 数据进行了回顾性处理。具体而言，使用了2020 版 MEDLINE/PubMed

[20]每年的完整描述符列表见补充文件 S1。

年度基线存储库来检索文档和医学主题词表标注。此外， DBM 使用 SemMedDBv4.0 进行 CO ，并将数据不足以用于训练和评估的用例排除在实验之外[21]。

具体而言， $RetroBM$ 确定了2006年引入的18个描述符，用于开发$test_{2006}$ 真实数据集，并 DBM 开发了相应的弱标注数据集$WSdev_{2006}$。这两个数据集（*2006*年数据集）各自包含超过40万篇文章，并用于验证实验[22]。其余70个描述符是在2006年之后引入的，用于评估实验（*2007—2019*年数据集）。13个弱标注数据集（$WS\,dev_{2007\text{-}2019}$）及其相应的真实数据集（$test_{2007\text{-}2019}$）的总规模分别超过200万篇和100万篇文章[23]。

-

在这些数据集中，正例极为稀少，凸显了 DBM所采用的平衡方法的重要性。具体而言，在$test_{2006}$, 中，18个标签中任一标签的正例仅约占$test_{2006}$全部文章的6%。此外， $WS\,dev_{2006}$ 和$test_{2006}$中的文章分别约有3%和5%依据弱标签被判定为正例。考虑到弱标签和真实标签，*2007—2019*年数据集 中70个标签的这些比例也相近[24]。这表明，对较早标签的观察结果可能有助于后续引入的标签，从而证实了回顾性场景的合理性。

*4.2. 2006*年数据集的验证实验

验证实验的首要目的是确认概念出现（CO）作为 DBM的弱监督来源的有效性。表 2 展示了 $FGSI$ 任务在测试$_{2006}$上的标注函数性能。 CO 显然是表现最强的标注函数，这证实了此前小规模实验中的观察结果 [4]。在

[21]我们保留了至少包含 10 篇文章的用例，这些文章满足以下条件：a） c 在 $WS\ dev_y$中出现，且 b） d_c在 $test_y$中有注释。出于实际考虑，我们还排除了一个过于宽泛的用例，该用例中单个 d 对应的文章超过 150 万篇。[22] $WS\ dev_{2006}$ 446,377 篇文章， $test_{2006}$ 468,271 篇文章。
[23] $WS\ dev_{2007\text{-}2019}$ 2,068,965 篇文章， $test_{2007\text{-}2019}$ 1,135,657 篇文章。[24]各数据集的详细统计数据见补充文件 S1。

表 2： $FGSI$ 各标注函数在 $test_{2006}$ 真实标注数据上的结果。 var 表示 $maF1$ 在18个标签间的方差。

标注函数（LF）	maP	maR	$maF1$ $\pm var$	miP	miR	$miF1$
concept occurrence (*CO*)	0.715	0.618	**0.634** ±0.06	0.737	0.531	**0.618**
name exact (*NE*)	0.707	0.082	0.112 ±0.04	0.901	0.027	0.052
synonyms exact (*SE*)	0.656	0.212	0.251 ±0.12	0.889	0.046	0.088
name lowercase (*NL*)	0.649	0.441	0.446 ±0.12	0.589	0.203	0.302
synonyms lowercase (*SL*)	0.620	0.365	0.408 ±0.11	0.709	0.251	0.371
name no punct (*NNP*)	0.574	0.440	0.442 ±0.13	0.583	0.196	0.293
synonyms no punct (*SNP*)	0.597	0.349	0.385 ±0.11	0.690	0.206	0.318
name tokens (*NT*)	0.322	0.838	0.377 ±0.07	0.157	0.807	0.263
synonyms tokens (*ST*)	0.274	0.874	0.338 ±0.07	0.081	0.883	0.149

此外，我们观察到，大多数标注函数的精确率都高于召回率，基于词元化的标注函数（NT、ST）[25]除外。CO 与其他每个标注函数之间的平均相关性为 0.48，其中 SL、 NNP和 SNP的相关性最高（0.68），而 ST 与 CO的相关性最低（0.21）。同样，各基于词典的标注函数之间的平均相关性为 0.35，其中$[SL, SNP]$ 是相关性最高的一对（0.95），而 $[SL, ST]$ 是相关性最低的一对（0.06）[26]。

这些实验的第二个目标是确认 CO 能否通过与其他标注函数集成而得到提升，并选择 a) 应采用哪个标注函数子集，以及 b) 在 DBM 中应采用哪种集成方法来增强弱监督。为此，针对每个标注函数子集，评估了三种集成方法在 $test_{2006}$上的 $FGSI$任务中的表现，即多数投票 (*MV*)、至少一个 (*ALO*)和 ***Snorkel*** 标签模型 (*LM*)。结果证实，与单独使用 CO [27]相比，若干标注函数子集提升了性能。

[25]这些标注函数中的某些词元可能过于宽泛，例如表 1 中的“疾病”、“类型”和“a”。[26]所有标注函数的完整相关性矩阵见补充文件 S2。[27]所有标注函数子集的完整结果见补充文件 S3。

总体而言，在 ALO 集成或 LM 集成下，两个标注函数子集[28]分别取得了最高的 $maF1$ 分数（0.689）和 $miF1$ 分数（0.646），比 CO 分别高出5.5个百分点和2.8个百分点。此外，另有六个标注函数子集仅在 ALO 方法下取得了这些最高分。值得注意的是，无论采用哪种集成方法，所有表现最佳的标注函数子集中都包含 CO、 NL和 SL ；仅由这三个标注函数组成的组合，使用 ALO和 LM时也足以取得最佳表现。因此，最终选用子集 $[CO, NL, SL]$，用于增强 DBM中的弱监督。

鉴于 ALO 在比 LM[29]更多的 LF 子集上取得了最佳结果，这表明在当前设置下， ALO 对 LF 子集的选择更具鲁棒性。这可能是因为大多数标注函数的精确率较高，而且表现最佳的子集中都不包含精确率较低的标注函数（NT、ST）；众所周知， ALO能够通过提高总体召回率 [16]，有效地组合精确的投票者。具体而言，只要至少一个标注函数为某篇文章赋予标签， ALO 就会为其赋予标签，不受较谨慎的标注函数可能给出的负向投票影响。因此，在 DBM中，采用了 $[CO, NL, SL]$ （ALO_3）的 ALO 集成作为增强型弱监督方案。

2006年数据集上的验证实验的另一个目标，是为 DBM的模型开发部分做出最终的具体选择。具体而言，我们需要 a) 估算一个合理的负样本与正样本比例（$balance\ n$），以及 b) 选择用于 DBM 预测的模型所对应的训练轮次，也就是说，使用在 验证部分 的 $WSdev$上损失最低的轮次（ $best\ ep$ ）对应的模型，还是使用其前一轮（$prev\ ep$）或后一轮（$next\ ep$）对应的模型。

围绕这一方向，基于 $WS\ dev_{2006}$开发了不同的 DBM 模型，并评估了它们在 $test_{2006}$上的 $FGSI$ 性能。具体而言，通过实验测试了从1到100的不同 $balance\ n$ 值，最终认为10比1的负样本与正样本比例较为合理，并将其用于所有

[28] $[CO, NL, SL]$ 和 $[CO, NL, SL, NNP]$。298个表现最佳的子集使用 ALO ，而使用 LM的有2个。

表3： $FGSI$ $test_{2006}$ 数据上的验证结果。WS表示训练模型所用弱监督的来源。 LR 中的下标表示所选特征的数量。 DBM 中的下标表示训练轮次，其中 p、 b和 n 分别对应 $prev\ ep$、$best\ ep$和 $next\ ep$ 。 var 表示18个标签中 $maF1$ 的方差。

LF/模型 WS		maP	maR	$maF1$ $\pm var$	miP	miR	$miF1$
CO	-	0.715	0.618	0.634 ±0.06	0.737	0.531	0.618
LR_5	CO	0.427	0.822	0.474 ±0.10	0.295	0.655	0.406
LR_{10}	CO	0.358	0.807	0.396 ±0.09	0.102	0.629	0.176
LR_{100}	CO	0.275	0.809	0.318 ±0.06	0.085	0.620	0.149
LR_{1000}	CO	0.279	0.798	0.310 ±0.07	0.134	0.644	0.221
ALO_3	-	0.717	0.698	0.689 ±0.03	0.713	0.591	0.646
DBM_p	ALO_3	0.681	0.744	0.700 ±0.03	0.689	0.624	0.655
DBM_b	ALO_3	0.688	0.738	**0.702** ±0.03	0.690	0.626	0.656
DBM_n	ALO_3	0.689	0.741	**0.702** ±0.03	0.695	0.623	**0.657**

剩余实验。此外，对三种训练轮次方案进行 DBM 实验所得的 $FGSI$ 结果（表3）表明，对应于$next\ ep$ （DBM_n）的模型在两项指标上都取得了最佳结果。因此，在 DBM中采用了 $next\ ep$ 模型作为最终模型。

最后，在相同的 $FGSI$ 任务（*2006*年数据集）上使用逻辑回归（LR）模型进行的验证实验（结果也见表3），使我们能够确定评估实验中应考虑的特征数量。具体而言， LR 模型在 WS dev_{2006} 上进行训练，并使用F-ANOVA统计量进行特征选择，分别考虑5、10、100和1000个特征。结果表明，没有任何 LR 模型的整体表现优于CO 。具体而言，表现最佳的 LR 模型，即使用五个特征的 LR_5 ，在 $F1$ 上约40%的18个标签中，表现优于 CO $test_{2006}$[30]。

[30]补充文件S4中提供了各描述符的详细验证结果。

表4： $FGSI$ 在 $test_{2007-2019}$ 数据上的评估结果。WS表示训练模型所用的弱监督来源。 LR 模型中的下标表示所选特征的数量。 DBM 模型中的下标表示所采用的训练轮次（ $next\ ep$的n ）。 var 表示 $maF1$ 在70个标签上的方差。

LF/模型	WS	maP	maR	$maF1$ $\pm var$	miP	miR	$miF1$
CO	-	0.766	0.590	0.626 $\pm$0.06	0.848	0.554	0.670
NL	-	0.647	0.369	0.408 $\pm$0.12	0.880	0.217	0.348
SL	-	0.567	0.253	0.287 $\pm$0.09	0.809	0.125	0.217
LR_5	CO	0.420	0.807	0.458 $\pm$0.09	0.204	0.783	0.324
$C2F$	CO	0.206	0.946	0.276 $\pm$0.09	0.074	0.981	0.138
$coDBM_n$	CO	0.718	0.647	0.641 $\pm$0.06	0.826	0.573	0.677
$coDBM_n^{C2F}$	CO	0.712	0.626	0.625 $\pm$0.06	0.831	0.568	0.675
MV_3	-	0.782	0.481	0.542 $\pm$0.08	0.873	0.281	0.462
LM_3	-	0.756	0.652	0.661 $\pm$0.06	0.842	0.569	0.679
ALO_3	-	0.764	0.668	**0.677** $\pm$0.05	0.839	0.583	0.688
LR_5	ALO_3	0.403	0.825	0.454 $\pm$0.09	0.196	0.797	0.315
$C2F$	ALO_3	0.203	0.959	0.272 $\pm$0.09	0.076	0.983	0.141
DBM_n	ALO_3	0.725	0.695	0.672 $\pm$0.04	0.839	0.621	**0.714**
DBM_n^{C2F}	ALO_3	0.726	0.679	0.656 $\pm$0.05	0.803	0.588	0.679

4.3. 对*2006*年后引入的新标签进行评估

在最后一组实验中，我们使用*2007—2019*年数据集中的70个未见标签，对 $FGSI$ 上的 DBM 及竞争方法进行了评估。将 CO 监督增强为 ALO_3 是 DBM的重要组成部分。然而，为了探究 ALO_3 的重要性，我们还仅使用 CO 直接训练了 DBM 模型（$coDBM$）。出于同样的原因，我们使用原始（CO）和增强后的（ALO_3）监督分别训练了竞争模型。表 4[31]展示了这些实验在 $FGSI$ 任务上的结果。

这些结果证实， CO 是用于 $FGSI$的强大启发式方法，表现再次优于

[31]各描述符的详细评估结果见补充文件 S5。

基于词典的启发式方法。此外，弱监督增强方案（ALO_3）的总体 $maF1$ 比单独使用CO 高出约 5 个百分点，这一差异具有统计学意义[32]。对于相同的 LF 子集（LM_3），*Snorkel* 标签模型的性能并未显著低于 ALO_3，因为在 70 个标签中的大多数标签上，这两种方法取得的 $F1$ 得分实际上相同。不过，对于全部 70 个标签， ALO_3 的 $F1$ 性能均与 LM_3 相当或更优。

关于训练后的模型，结果显示，逻辑回归（LR）模型整体上未能超越其所获得的监督信号，无论是 CO 还是 ALO_3。 $maF1$ 模型的整体性能也较低，召回率最高，但精确率较低。具体而言，在 70 个标签中，只有 13 个标签上，基于 CO 训练的 $C2F$ 表现优于 CO 本身。像 $C2F$ 这样的最先进方法表现不佳，说明$FGSI$ 任务和数据集存在一些特殊挑战，例如大量“不带标签”的文档不属于任何目标标签。

另一方面，仅使用 CO 的弱监督开发的 $coDBM_n$ 模型，在 $maF1$ 方面相较于 CO 本身，整体性能仅提升约 1.5 个百分点。采用增强监督 （ALO_3）的 DBM_n 模型是训练效果最佳的模型，在 $maF1$ 上较 CO 提升约 4.6 个百分点，且这一提升具有统计显著性[33]。然而，即使是 DBM_n ，其性能也仅与 ALO_3 本身相当。事实上，单独使用 ALO_3 的 $maF1$ 略高，但差异不具有统计显著性[34]。总体而言， DBM 方法以 CO 这一强启发式方法为起点，进一步提升了其效果，并在无法获取真实标签的情况下，将若干粗粒度标签细化为 70 个未见过的细粒度概念。这些结果表明，采用 DBM 设计选择的一些深度学习模型，确实能够改进其所依赖的弱监督。

[32]Wilcoxon 符号秩检验， H_1： $F1\,CO < F1\,ALO_3$， p 值 $= 0.000001$。[33]
Wilcoxon 符号秩检验， H_1： $F1\,CO < F1\,DBM_n$， p 值 $= 0.0065$。[34]
Wilcoxon 符号秩检验， H_1： $F1\,ALO_3 \neq F1\,DBM_n$， p 值 $= 0.088$。

5. 讨论与结论

在这项工作中，我们研究了在缺乏真实标注训练数据的情况下进行细粒度语义索引（$FGSI$）的任务，以及为此开发弱监督深度学习（DL）模型。此外，我们提出了一种自动化方法，基于 MEDLINE 文章为 $FGSI$ 生成大规模真实标注数据集。因此，本文的主要贡献如下：

- 确认概念出现 (CO)是大规模$FGSI$ 的强有力启发式方法。

- 一种用于$FGSI$ 任务的新型弱监督方法：$Deep\ Beyond\ MeSH$ (DBM)。该方法提出了一种新颖的方案，以增强原有的弱监督（ALO_3），并通过特定的设计选择，开发出能够在弱监督本身基础上进一步提升效果的深度学习模型。

- 一种用于自动构建真实标注数据集的新方法：$Retrospective\ Beyond\ MeSH$ ($RetroBM$)，以支持对大规模 $FGSI$ 预测的评估。

$RetroBM$利用医学主题词表向更细粒度标签扩展这一特点，为88个不同的细粒度标签构建了 $FGSI$ 真实标注数据集。据我们所知，这是首个用于 $FGSI$ 的大规模真实标注数据集，总计包含超过一百万篇文章。这些数据使我们得以确认 CO 作为生物医学文献$FGSI$ 启发式方法的有效性，并进一步开展本研究的验证与评估实验。

DBM首先使用 CO 启发式方法来开发弱标注 $FGSI$ 数据集。然后，它将 CO 与基于词典的启发式方法 名称小写化 (NL) 和 同义词小写化 (SL) 相结合，采用至少一个 集成方法 (ALO_3)，显著增强这些弱标签。这两种启发式方法会检查概念名称及其同义词是否在文章中逐字出现。考虑到 NL 和 SL 依赖于，这种增强尤其值得关注。

相同的概念术语已由 CO 本身加以利用，但以不同方式使用，从而检测出仅靠 CO 会遗漏的相关文章。

最后， DBM 利用增强的 ALO_3 弱监督，并基于预训练模型微调这一最先进的方法，在无法获取真实标签的情况下构建用于 $FGSI$ 任务的深度学习模型。对现有方法（即逻辑回归 (LR)和 $Coarse2Fine$ ($C2F$)）进行的实验表明， $FGSI$ 任务颇具挑战性，且这些方法无法取得优于弱标签本身的结果。另一方面， DBM 采用了针对 $FGSI$ 任务挑战的设计选择。具体而言，为处理 $FGSI$ 数据的不平衡问题， DBM 采用 R-BCE-FL 损失和负样本欠采样。为避免在弱监督上过拟合， DBM 采用早停策略，选择损失达到最小值的那个轮次之后的一个轮次（$next\ ep$）。最后，为限制随机性的影响，该方法对使用不同随机种子构建的模型进行多数投票。

-

评估结果显示，没有任何训练后的模型能够超越 ALO_3 本身的整体宏平均 F1（$maF1$）表现， DBM 模型的表现也仅与之相当。尽管如此， DBM 模型整体上取得了有竞争力的 $FGSI$ 表现；在评估实验的新未见标签上，其微平均 F1 （ $miF1$）水平甚至超过了表现极具竞争力的增强弱标注（ ALO_3）本身。这些结果表明，基于深度学习的弱监督 $FGSI$ 结合 DBM ，是支持对生物医学文献进行细粒度访问的一个有前景的方向。

关于 $RetroBM$，一个有趣的拓展方向是调整用例选择标准，构建用于回顾性研究其他索引任务的数据集，例如为一组概念使用描述符进行索引。关于改进这一前景良好的 DBM 方法，以实现基于深度学习的弱监督 $FGSI$，我们计划利用额外的标签信息，例如层级关系、基于对比学习方法的更细粒度文本表示，以及概率弱标签。

6. 致谢

本研究得到希腊研究与创新基金会（HFRI）HFRI博士奖学金项目的资助（奖学金编号：697）。本研究使用的数据资源，包括医学主题词表、MEDLINE/PubMed和SemMedDB，均由美国国家医学图书馆提供。我们感谢Georgios Katsimpras就深度学习开展的有益讨论。

参考文献

[1]A. Nentidis、G. Katsimpras、E. Vandorou、A. Krithara、A. Miranda- Escalada、L. Gasco、M. Krallinger、G. Paliouras，《BioASQ 2022 概述：第十届大型生物医学语义索引与问答 BioASQ 挑战赛》，载于：《计算机科学讲义》（包括《人工智能讲义》和《生物信息学讲义》子系列），第 13390 卷 LNCS，2022 年，第 337–361 页。arXiv:2210.06852，doi:10.1007/978-3-031-13643-6_22。

[2] Y. Gu、R. Tinn、H. Cheng、M. Lucas、N. Usuyama、X. Liu、T. Naumann、J. Gao、H. Poon，《面向生物医学自然语言处理的领域专用语言模型预训练》，《ACM医疗保健计算汇刊》3 (1) (2022) 1–23。doi:10.1145/3458754。

[3] A.R.Aronson, J.G.Mork, C. W. Gay, S. M. Humphrey, W. J. Rogers,《美国国家医学图书馆NLM索引计划的医学文本索引器》，《健康技术与信息学研究》107（第1部分）（2004）268–72。doi:10.3233/978-1-60750-949-3-268.

[4] A. Nentidis、A. Krithara、G. Tsoumakas、G. Paliouras，《Beyond MeSH：基于弱监督的生物医学文献细粒度语义索引》，《信息处理与管理》57（5）（2020）102282。doi:10.1016/j.ipm.2020.102282。

[5] D. Mekala、V. Gangal、J. Shang，《Coarse2Fine：基于粗粒度标注数据的细粒度文本分类》，载于：2021年自然语言处理实证方法会议论文集，计算语言学协会，美国宾夕法尼亚州斯特劳兹堡，2021年，第583–594页。arXiv:2109.10856，doi:10.18653/v1/2021.emnlp-main.46。

[6] I. S. Alec Radford、JeffreyWu、Rewon Child、David Luan、Dario Amodei，《语言模型是无监督的多任务学习者》，技术报告，5月（2020）。网址：https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

[7] J. Devlin、M.-W. Chang、K. Lee、K. Toutanova，《BERT：用于语言理解的深度双向变换器预训练》，载于：2019年计算语言学协会北美分会会议：人类语言技术论文集，第1卷（长篇与短篇论文），计算语言学协会，明尼阿波利斯，明尼苏达州，2019年，第4171–4186页。doi:10.18653/v1/N19-1423。

[8] G. Tsatsaronis、G. Balikas、P. Malakasiotis、I. Partalas、M. Zschunke、M. R. Alvers、D. Weissenborn、A. Krithara、S. Petridis、D. Polychronopoulos等，《BIOASQ大规模生物医学语义索引与问答竞赛概述》，BMC生物信息学 16（1）（2015）138。doi:10.1186/s12859-015-0564-6。

[9]R. You、Y. Liu、H. Mamitsuka、S. Zhu，《BERTMeSH：利用全文进行大规模、高性能 MeSH 索引的深度上下文表征学习》，《生物信息学》37 (5) (2021) 684–692。doi:10.1093/ bioinformatics/btaa837。

[10] A. R. Rae、J. G. Mork、D. Demner-Fushman，《一种用于自动医学主题词表索引的神经文本排序方法》，载于：CEUR研讨会论文集，第 2936 卷，2021 年，第 302–312 页。网址：http://ceur-ws.org/Vol-2936/paper-22.pdf

[11]S. J. Darmoni、L. F. Soualmia、C. Letord、M.-C. Jaulent、N. Griffon、B. Thirion、A. N´ev´eol，《利用医学主题词表概念改进信息检索：罕见病和慢性病案例测试》，《医学图书馆协会杂志》：JMLA 100 (3) (2012) 176–183。doi:10.3163/1536-5050.100.3.007。

[12] A. Nentidis, A. Krithara, G. Tsoumakas, G. Paliouras，《超越医学主题词表：基于弱监督的生物医学文献细粒度语义索引》，载于：2019 年 IEEE 第 32 届国际计算机医学系统研讨会（ CBMS ），IEEE，2019 年，第 180–185 页。doi:10.1109/ CBMS.2019.00045。

[13] A. R. Aronson, F.-M. Lang，《MetaMap 概述：历史视角与最新进展》，美国医学信息学会杂志 17 (3) (2010) 229–236。doi:10.1136/jamia.2009.002733。

[14] A. Nentidis, A. Krithara, G. Tsoumakas, G. Paliouras，《这些新的医学主题词表究竟是什么？》，《国际数字图书馆杂志》22 (4) (2021) 319–337。arXiv:2101.08293，doi:10.1007/s00799-021-00304-z。

[15] A. Ratner, C. De Sa, S. Wu, D. Selsam, C. R´e，《数据编程：快速创建大型训练集》（NIPS）（2016）1–9。arXiv:1605.07723。

[16] N. Kambhatla，《少数票：至少 N 票投票可提高关系抽取的召回率》，载于：COLING/ACL 2006——第21届国际计算语言学会议暨计算语言学协会第44届年会论文集，主会场海报环节，2006年7月，页码 460–466。网址：https://aclanthology.org/P06-2060

[17] A. Ratner、S. H. Bach、H. Ehrenberg、J. Fries、S. Wu、C. R´e，《Snorkel：利用弱监督快速创建训练数据》，《VLDB基金会会议论文集》11（3）（2017）269–282。arXiv:1711.10160v1，doi:10.14778/3157794.3157797。

[18] M. T¨anzer、S. Ruder、M. Rei，《预训练语言模型中的记忆与泛化》03。arXiv:2105.00828。

[19] D. Hendrycks、K. Lee、M. Mazeika，《使用预训练可以提升模型鲁棒性和不确定性》，第36届国际机器学习大会（ICML 2019），2019年6月（2018）（2019），4815–4826。arXiv:1901. 09960。

[20] A. Garcia-Silva、J. M. Gomez-Perez，《使用 BERT 对科学出版物进行分类——自注意力是一种特征选择方法吗？》，载于：D. Hiemstra、M. Moens、J. Mothe、R. Perego、M. Potthast、F. Sebastiani（编），《信息检索进展——第43届欧洲信息检索研究大会（ECIR 2021）》，线上会议，2021年3月28日至4月1日，论文集，第一部分，《计算机科学讲义》第12656卷，施普林格，2021年，第161–175页。doi:10.1007/978-3-030-72113-8_11。

[21] L. Prechelt，《早停——但何时？》，Springer Berlin Heidelberg，柏林、海德堡，2012年，第53–67页。doi:10.1007/978-3-642-35289-8_5。

[22] Y. Bengio、A. Courville、P. Vincent，《表示学习：综述与新视角》，IEEE《模式分析与机器智能汇刊》35 (8) (2013) 1798–1828。doi:10.1109/TPAMI.2013.50。

.[23] Y. Huang、B. Giledereli、A. K¨oksal、A. Ozg¨ur、E. Ozkirimli，《长尾类别分布下多标签文本分类的平衡方法》，《2021年自然语言处理实证方法会议论文集》arXiv:2109.04712。

[24] T. Wu, Q. Huang, Z. Liu, Y. Wang, D. Lin，《长尾数据集上的多标签分类分布均衡损失》，载于：《计算机科学讲义》（包括《人工智能讲义》和《生物信息学讲义》子系列），第12349卷 LNCS，2020年，第162–178页。arXiv:2007.09654，doi:10.1007/978-3-030-58548-8_10。

[25] J. Dodge, G. Ilharco, R. Schwartz, A. Farhadi, H. Hajishirzi, N. S mith, 《预训练语言模型微调：权重初始化、数据顺序和早停》 arXiv:2002.06305。

[26] D. P. Kingma, J. Ba, 《Adam：一种随机优化方法》，第三届国际学习表征会议（ICLR 2015）会议论文集（2014），第 1–15 页，arXiv:1412.6980。

[27] G. Tsoumakas、I. Katakis、I. Vlahavas, 《多标签数据挖掘》，载于《数据挖掘与知识发现手册》，Springer US，波士顿，马萨诸塞州，2009 年，第 667–685 页。doi:10.1007/978-0-387-09823-4_34。