利用半监督自动语音识别和 N-best 转录文本提升基于 BERT 的排序模型在口语内容检索中的抗噪鲁棒性

Yasufumi Moriya，Gareth. J. F. Jones

ADAPT Centre，都柏林城市大学计算机学院，爱尔兰都柏林 9

arXiv:2301.06056v1 [cs.IR] 15 Jan 2023

摘要

研究表明，基于 BERT 的重排序和稠密检索（DR）系统能够提升口语内容检索（SCR）的搜索效果。然而，与准确的人工转录文本相比，使用自动语音识别转录文本时，这两种方法的效果仍可能下降。我们发现，在 How2口语指令视频数据集上进行的已知项目搜索任务中，平均倒数排名（MRR）得分下降了 10%–14%。为缩小这一差距，我们研究了使用半监督自动语音识别转录文本和 N-best 自动语音识别转录文本来减轻语音搜索中的自动语音识别错误，并采用基于 BERT 的排序方法。与标准自动语音识别转录文本相比，半监督自动语音识别转录文本使 MRR 提高了 2%–5.5%；我们针对 BERT DR 系统采用的 N-best早期融合方法则使 MRR 提高了 3%–4%。将半监督转录文本与用于 BERT DR 的 N-best早期融合相结合，使人工转录文本与自动语音识别转录文本之间的搜索效果 MRR 差距缩小了 50% 以上，从 14.32% 降至 6.58%。

索引词— 口语内容检索、基于 BERT 的重排序、BERT 稠密检索、基于 N-best BERT 的检索

1. 引言

视频流媒体服务和播客等领域产生的口语多媒体内容档案不断增长，因而对有效的口语内容检索（SCR）系统的需求也日益增加。与文本文档不同，口语文档无法直接根据其内容进行搜索。因此，口语多媒体内容通常只能依据用户创建的摘要和元数据进行搜索。这种方法的有效性取决于文档是否具备有意义的摘要和元数据。此外，由于许多口语内容较长，SCR 系统理想情况下应能根据内容本身，为用户提供更精细、更具层次的搜索功能。只有在能够获取文档中口语内容的转录文本时，才能提供此类功能。

人工转录文本的制作成本高昂，且内容量巨大，因此通常会使用自动语音识别（自动语音识别）生成转录文本。然而，自动语音识别产生的转录错误会影响搜索效果。尽管已有研究表明，信息检索（IR）通常能够适应中等程度的词错误率（WER），但 Larson 和 Jones 报告称，词错误率超过约 30% 时，可能会显著影响搜索可靠性 [1]。自动语音识别系统的准确度不断提高，某些任务的词错误率已低至极低水平 [2, 3]。然而，当自动语音识别系统的训练与待转录数据之间存在领域不匹配 [4], 、语音较为随意，或音频数据中存在声学噪声时，词错误率仍可能很高 [2]。在许多场景中，数据的领域或音频质量无法控制，例如视频流媒体服务或播客上的用户生成数据；这与实验室环境中使用的音频语料库不同。因此，即使是最先进的自动语音识别系统，转录词错误率仍可能很高。这促使我们探索开发提升 SCR 系统抗噪鲁棒性的方法。

基于 Transformer 的信息检索模型近来已成为传统模型（如 BM25 [5]）的一种高效替代方案。这些新模型大致可分为两类：重排序和稠密检索（DR）。基于 BERT 的重排序模型 MonoBERT [6]直接以查询-文档对为输入，计算其相关性得分。MonoBERT 及其后续模型的缺点在于，基于 Transformer 的推理速度很慢，而且硬件要求可能使其在实际应用中难以实施 [7]。在 DR 中，基于 Transformer 的模型分别对文档和查询进行编码，将文档转换为稠密文档向量。搜索时，只需将查询编码为向量表示。随后，可使用最近邻（NN）等任何基于向量的相似度度量来计算查询-文档相关性得分。这种方法的运行速度远快于重排序 [7]。

本文的贡献如下。(i) 我们研究了基于 BERT 的重排序和稠密检索在口语指令视频已知项目搜索任务中的表现。我们发现，尽管采用了基于 BERT 的检索系统，人工转录文本与自动语音识别转录文本之间的检索效果差距仍约为 10%–14%。(ii) 我们通过半监督训练对自动语音识别系统进行增强，以实现领域自适应，并考察改进后的自动语音识别转录文本对基于 BERT 的排序

978-1-6654-7189-3/22/$31.00 ©2023 欧盟

系统的影响。使用半监督自动语音识别转录文本使平均倒数排名得分提高了 2%–5.5%。(iii) 我们提出在 BERT DR 系统中使用 N-best 自动语音识别转录文本，并发现所提出的早期融合方法使平均倒数排名得分提高了约 3%–4%。结合这两种方法后，平均倒数排名平均提高了 6.5%。

本文其余部分安排如下。下一节将简要回顾基于 BERT 的信息检索和 SCR，随后介绍本研究所用模型的详细信息。接着，我们提出适用于基于 BERT 的稠密检索系统的扩展方法，利用自动语音识别 N-best 转录文本进行早期融合和后期融合。然后，我们将展示基于 BERT 的排序模型的实验结果并进行分析，最后总结全文并探讨未来工作。

2. 与先前工作的关联

口语文档检索（SDR）的主流研究始于文本检索会议（TREC） [8]。此后，研究者持续探索 SCR，例如参与 MediaEval [9]的相关活动，以及 TREC 2020and 2021 [10, 11]的播客检索赛道。近期另一项关于 SCR 的研究见 [12, 13] , , 该研究探讨了在普通话新闻报道检索中使用 BERT 重排序。

多年来，人们提出了许多检索模型，其中效果最稳定的模型之一是 BM25 概率模型 [5], ，它已成功应用于许多 SCR 任务，包括 [9, 10]。近年来，神经信息检索模型陆续问世，包括 [14]。不过，最近又出现了一类使用预训练 Transformer 架构（BERT）的新型检索模型 [15] [16]。研究表明，这些模型优于经典的 BM25 模型 [5] 以及非 BERT 神经模型，例如 Conv-KNRM [14]。这些基于 BERT 的检索模型可分为两类：基于 BERT 的重排序 [6] 和稠密检索（DR） [7, 17]。基于 BERT 的重排序以查询-文档对为输入，直接生成相关性得分；而 DR 则分别将文档和查询编码为稠密向量，并对这些向量执行近似最近邻（ANN）搜索。

N-best 自动语音识别假设已被用作 BERT 模型的输入，用于口语理解，其形式包括格结构 [18], 混淆网络 [19] 和纯文本 [20]。Ganesan 等人在 [20] 中得出结论：与其他格式相比，文本格式的输入更适合 BERT 模型，因为 BERT 模型是在纯文本上预训练的。N-best 假设也已被用于自动语音识别错误纠正 [21]。

我们的论文与现有基于 BERT 的 SCR 任务排序系统研究的不同之处如下。(i) 我们在人工转录文本和自动语音识别转录文本上，分别对 BERT 重排序和 BERT DR 系统进行了实验。TREC 播客赛道 [10, 11]已采用 BERT 重排序和 DR 模型，但系统评估时没有提供人工转录文本。[12, 13], 研究了 BERT 重排序，但未将 BERT DR 模型

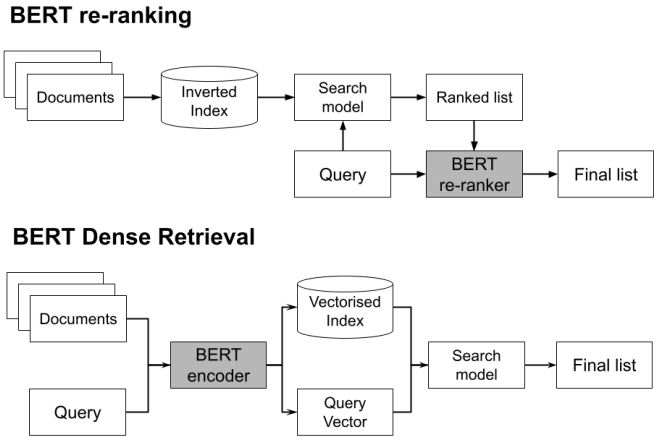


图 1。基于 BERT 的重排序系统和基于 BERT 的稠密检索（DR）系统示意图。

应用于其口语广播检索。(ii) 我们将半监督自动语音识别转录文本用于 SCR 场景，此时无法使用领域自适应自动语音识别系统。现有文献 [1, 8] 指出，转录词错误率与搜索得分之间近乎呈线性关系。我们对半监督自动语音识别转录文本的研究表明，使用半监督 ASR 系统改进转录文本后，BERT 重排序和 DR 模型均能从中受益。(iii) 我们提出在 BERT DR 系统中使用自动语音识别 N-best 转录文本。据我们所知，以往尚未研究过在 SCR 中使用 ASR N-best 的 BERT 排序模型。我们发现，在 SCR 任务中，BERT DR 的早期融合和后期融合方法都能减轻自动语音识别转录文本中的错误。

3. 基于 BERT 的检索模型

本节将详细介绍基于 BERT 的重排序和稠密检索。我们首先介绍这些模型，然后说明它们的训练方式。

3.1. 用于检索系统的 BERT

图 1 展示了 BERT 重排序系统和 BERT DR 系统的结构。BERT 重排序器接收初始检索模型生成的文档排序列表，并使用 BERT 模型对其重新排序；而稠密检索系统使用 BERT 将文档和查询转换为稠密向量，再通过另一种检索模型获取最终的文档排序列表。

BERT 重排序器 BERT 模型会对使用 BM25 等轻量级检索模型从文档库中检索出的排序文档列表进行重排序。最早用于重排序的 BERT 模型是 MonoBERT [6]。假设查询由 N 个标记组成，文档由 M 个标记组成，则 BERT 模型的输入可表示为 $[CLS]q_1$、.....、$q_N[SEP]$

d_1、.....、$d_M[SEP]$，其中 $[CLS]$和 $[SEP]$是 BERT 的特殊标记。将对应于$[CLS]$标记的输出向量输入全连接层，即可计算查询-文档对的相关性得分。

*BERT*稠密检索 使用 BERT 模型将查询和文档编码为向量表示。Karpukhin等人 [17] 使用两个 BERT 模型分别对查询和文档进行编码，而 Xiong 等人 [7] 则使用单个 BERT 模型对查询和文档进行编码（双编码器）。在本研究中，我们的稠密检索系统采用了双编码器框架。查询和文档的表示分别是一个向量，对应于查询标记或文档标记之前的 $[CLS]$ 标记。与重排序器 BERT 输出的相关性得分不同，稠密向量是在固定维度空间中对文档或查询的表示，最终的相关性得分通过计算查询向量和文档向量的相似度分数得到。

相似度分数可以通过点积或余弦相似度计算。这样便可使用近似最近邻等高效检索算法，根据查询与文档的相似度分数生成最终的文档列表 [22]。本研究遵循 [7]，使用点积计算相似度分数。

BERT 重排序与 *BERT* 稠密检索BERT 重排序通常被发现比 BERT DR 能取得更好的检索指标 [16]。然而，Xiong 等人 [7] 证明，在搜索时使用 BERT 重排序的速度可能比 DR 慢 100 倍。由于推理速度的限制，BERT 重排序通常只应用于初始搜索返回的 k 个顶级文档；而 BERT DR 在搜索时只需对查询进行编码，因此可以应用于整个文档集合。

3.2. 用于检索的 BERT 训练

BERT 重排序BERT 重排序模型使用二元交叉熵损失 [6] 进行训练。具体而言，给定一个查询与相关文档对，训练 BERT 模型使其输出 1；而对于查询与不相关文档，训练模型使其输出 0。该过程可表示为

$$L = -\log(f(q_i, d_i^+; \theta)) - \sum_{j=1}^{J} \log(1 - f(q_i, d_{ij}^-; \theta)) \quad (1)$$

其中，$f(q_i, d_i^+; \theta)$ 是 BERT 模型针对第 i 个查询-文档对给出的分数，而 $f(q_i, d_{ij}^-; \theta)$ 是 BERT 模型针对与第 i 个训练样例对应的第 j 个负查询-文档对给出的分数。

*BERT*稠密检索 BERT DR 模型使用负对数似然 [7, 17]进行训练。具体而言，模型经过训练后，会最大化查询与相关文档对的相似度分数，并最小化查询与非相关文档对的相似度分数。形式化表示如下：

$$s(q, d) = sim(g(q; \theta), g(d; \theta)) \quad (2)$$

$$L = -\log \frac{\exp(s(q_i, d_i^+))}{\exp(s(q_i, d_i^+)) + \sum_{j=1}^{J} \exp(s(q_i, d_{ij}^-))} \quad (3)$$

其中，$g(q; \theta)$ 是一个函数，用于获取与 BERT 输出向量对应的查询稠密向量，该输出向量来自 [CLS] 标记；$g(d; \theta)$ 是其文档对应函数。$sim()$ 是一个用于计算这两个向量相似度分数的函数。与 BERT 重排序器使用的交叉熵损失类似，对于第 i 个查询-文档对及其对应的 j 负查询-文档对，可以得到该损失。

为了为每个真实查询-文档对获取负例，通常使用 BM25 模型来确定初始负例。Xiong 等人 [7] 提出了一种改进算法来选择负例，称为近似最近邻负对比学习（ANCE）。该算法首先使用从 BM25 文档排序结果中选出的负例进行 BERT 训练，并在每个指定的检查点使用正在训练的 BERT 模型更新负例。Xiong 等人 [7]为 ANCE 提供了理论依据和实证结果，表明 ANCE 训练能够带来更好的搜索结果，并加快模型收敛。尽管他们的原始论文提出了异步更新负例，即 ANCE 推理与模型训练并行运行，但我们在每个训练轮次结束时暂停模型训练，运行 ANCE 更新，然后使用更新后的负例恢复模型训练。

4. N- BEST 融合，用于 BERT 稠密检索 VAL

如第 2 节所述，使用自动语音识别 N-best 转录文本已在口语语言理解和自动语音识别错误纠正中取得成效 [20, 21]。本文提出了两种用于 BERT DR 的 N-best 融合方法。第一种方法在将文本输入 BERT 之前融合 N-best 转录文本（早期融合）。第二种方法为 N-best 转录文本创建 N 个索引，并合并各索引的得分，以计算每个文档的最终得分（晚期融合）。

4.1. N-best早期融合

BERT 模型的输入序列长度通常限制为 512 个标记，超过此最大长度的标记会被截断 [15]。对于口语语言理解任务，用户话语通常很短，因此可以使用 [SEP]标记 [20]拼接 N-best 转录文本。然而，对于 SCR，文档长度通常要长得多，因此不能简单地使用 [SEP]标记拼接 N-best 文档表示。可以通过 [21], 中介绍的对齐方法解决此问题，并对 N 词

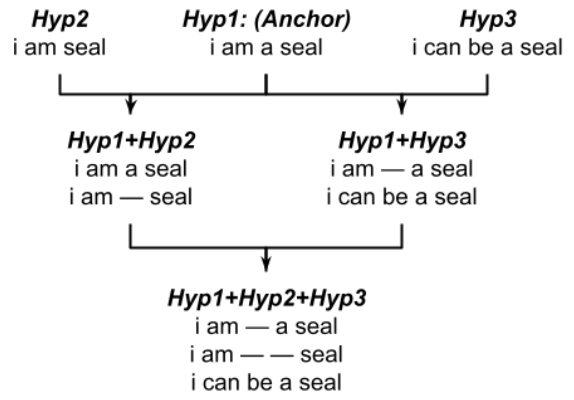


图 2. 用于生成 BERT 输入文本的 N-best 对齐示例。

嵌入取平均值，将 ith 个标记的 N 文档压缩为单个文档表示。

图 2 展示了对 3-best 转录文本进行对齐的示例。1-best 转录文本被用作锚定转录文本，并与另外两份转录文本对齐。随后进一步组合对齐后的转录文本，使 N-best 转录文本的长度一致。如 [21], 所述，使用“—”作为特殊标记，以缓解转录文本之间的长度不匹配。该标记会被添加到其中一个词嵌入向量中。BERT 模型的输入通常是子词标记 [15]。必须在文本完成分词后再 对转录文本进行对齐。如果颠倒这一顺序，对齐后的转录文本在分词后可能会出现长度不匹配。N-best 转录文本对齐后，将标记转换为词嵌入，并对齐后标记中处于 i 位置的内容取平均值。例如，在图 2 中，三份转录文本的首个位置均为“i”，因此得到“i”的平均嵌入。还可以通过拼接嵌入并将拼接后的嵌入输入线性层来扩展此方法，如 [21], 所示；但拼接后的向量可能维度很大，因此本文未采用此方法。

4.2. N-best 后期融合

我们的 BERT DR N-best 后期融合方法为每份 N-best 转录文本创建一个 N 索引。BERT 模型将查询编码为向量，并针对所有 N 索引进行搜索。最终结果通过累加每个索引中各文档的分数计算得出，并按累加分数降序排列。尽管这种方法的缺点是需要创建 N 索引，但在计算出查询向量后，可以使用 N 中央处理器并行对 N 索引执行近似最近邻搜索，以保持运行速度。

5. 实验研究

本节介绍我们的实验研究。我们首先说明实验设置，包括相关细节涵盖实验所用数据、BERT 重排序和深度检索模型的训练与超参数细节，以及标准自动语音识别转录文本和半监督自动语音识别转录文本的制作。随后，我们展示使用口语指令视频语料库进行已知项目搜索任务的实验结果。结果包括将 BERT排序系统应用于不同类型转录文本时的搜索得分、BERT 重排序和 DR 系统的推理时间，以及 N-best 扩展在 BERT DR 系统中的应用。

5.1. 实验设置

测试集与评估 我们的实验研究使用了 How2数据集 [23]。How2 数据集包含 19,770 个视频及其人工转录文本和元数据（包括视频标题）。在实验中，我们使用 How2 视频标题作为搜索查询，以检索相应的视频，从而构成已知项目搜索任务，其中每个查询都对应一个相关目标文档。我们从 19,770 个视频标题中随机选取了 500 个作为评估查询。实验采用标准的已知项目搜索指标——平均倒数排名（MRR），其定义如下：

$$MRR = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{rank_i} \tag{4}$$

其中，N 是用户查询的数量，$rank_i$ 是检索列表中与第 i 个查询相关的文档的排名。在 How2 语料库中，所用集合中只有约 2.5% 超出了 BERT 的长度限制。因此，我们的结果不太可能受到 BERT 长度限制的影响。对这一问题可能造成的影响进行考察，并探讨应对策略，可作为未来的研究方向。

BERT 检索模型数据设置 对于 BERT 重排序模型，我们另外随机选取了 500 个验证查询。对于每个验证查询，从 BM25 排序列表的顶部选取一个正例和一个负例（如果排名第一的文档是正例，则选取排名第二的文档），并使用交叉熵计算验证损失。对于 BERT DR 模型，我们随机选取了 500 个验证查询，并针对每个查询，从 BM25 模型中选取 1 个相关文档和 99 个不相关文档，由 BERT 编码器对其进行排序，以计算相关文档的平均倒数排名。

自动语音识别转录文本为了获取 How2 音频数据的自动语音识别转录文本，我们使用 LibriSpeech 语料库 [24, 25] 开发了一个 Kaldi“chain”系统。这是一种将声学模型与隐马尔可夫模型相结合的传统自动语音识别系统。该系统称为“标准自动语音识别”，使用 LibriSpeech 进行训练，因此在识别 How2 音频时存在领域不匹配，整体词错误率为 31.1%。我们使用“标准自动语音识别”系统生成 How2 训练数据的解码格和一最佳转录文本。将解码格与 LibriSpeech 数据结合后，用于训练一个

领域自适应半监督声学模型；一最佳转录文本则用于训练半监督 N 元语法，以进行解码，并训练用于重打分的循环神经网络语言模型 [4, 26, 27]。该系统称为“半监督自动语音识别”，整体词错误率为 23.95%。此外，我们使用“标准自动语音识别”系统生成格预言机转录文本（即格结构中相对于人工转录文本最正确的假设），其整体词错误率为 11.18%。在 N-best 实验中，我们使用“标准自动语音识别”和“半监督自动语音识别”系统生成 {2, 5, 10, 20}-best 转录文本，以开发 BERT DR 系统的 N-best 版本。

检索模型所用的 BM25 模型是 Lucene 的 Python 版本，版本为 $k1$ 1.2 和 b 0.75[1]。重排序系统以每个查询返回的排名前 1,000 的文档作为输入。所用的预训练 BERT 模型为“bert-base-uncased”，并使用 HuggingFace [28] 提供的对应分词器。对于稠密检索系统，在 BERT 模型对整个文档集合完成编码后，使用 Faiss 软件 [22] 对查询向量和文档向量索引执行相似度搜索。

BERT 模型训练的超参数设置如下。重排序器 BERT 训练了 20 个训练轮次，并采用验证分数最高时的模型检查点。模型通常在第 15 个训练轮次时收敛。通过经验选择的最佳初始学习率为 3×10^{-7}。据 [7] 报告，稠密检索 BERT 在 10 个训练轮次后收敛。找到的最佳初始学习率为 1×10^{-5}。重排序器 BERT 和稠密检索 BERT 均使用了 20 个负例。我们的 BERT 模型使用 Quadro RTX 6000 进行训练。增加负例数量或许能够提升 BERT 模型的性能，但 20 个负例已达到其内存限制。利用内存队列增加负例数量可作为未来的研究方向。训练重排序器和稠密检索 BERT 时，使用 AdamW 优化器，并将权重衰减设为 5×10^{-5}，以避免过拟合。

5.2. 使用格网预言、人工、标准和半监督自动语音识别转录文本的 MRR 分数

表 1 展示了 BM25、BERT 重排序和 BERTDR 系统使用人工（man）、格网预言（oracle）、半监督自动语音识别（semi ASR）和标准自动语音识别（std ASR）转录文本时的 MRR 分数。我们使用配对双尾 t 检验，比较 BM25 与其他基于 BERT 的模型所得 MRR 分数的分布。

*BERT*排序系统比较使用 ANCE 训练的 BERT DR 系统带来了最大的 MRR 提升。与人工转录文本上的 BM25 模型相比，其 MRR 分数提高了约 26%；与标准自动语音识别转录文本相比，提高了 23%，这表明转录错误影响了分数。BERT 重排序系统则提高了

[1]https://lucene.apache.org/

表 1。使用不同转录文本进行 How2 已知项目搜索评估的 MRR 结果。所有 BERT 结果均在统计学上显著优于 BM25，以 $p < 0.05$ ** 标示。

模型	MRR 评估数据			
	man	预言机	半自动语音识别	标准自动语音识别
BM25	40.27	36.02	30.85	29.37
重排序	57.32**	52.46**	48.69**	46.37**
DR	50.61**	45.96**	40.38**	34.81**
DR-ANCE	**66.73****	**63.88****	**56.22****	**52.41****

所有转录类型上的 MRR 分数均比 BM25 模型提高了约 16%–18%。

不同转录文本上的检索比较尽管 BERT 重排序和稠密检索系统提升了平均倒数排名，但人工转录文本与自动语音识别转录文本的得分差距仍然存在。使用标准自动语音识别转录文本进行检索时，BM25 和 BERT 重排序的 MRR 分数比使用人工转录文本低约 11%。转录错误对 BERT DR 系统的影响更大：使用标准自动语音识别转录文本时，MRR 分数低约 15%。尽管半监督转录文本对 BM25 模型提升 MRR 分数的作用不大，BERT 重排序系统的 MRR 分数提高了 2%，而 BERT DR 系统在未经 ANCE 训练时分数提高了 5.5%，经过 ANCE 训练时则提高了约 4%。然而，预言机实验表明，人工转录文本与自动语音识别转录文本之间的 MRR 分数差距仍有进一步缩小的可能。与使用人工转录文本进行检索相比，使用格网预言转录文本进行检索时，包括 BM25 在内的所有模型的 MRR 分数仅低 3% 至 4%。

搜索运行时间的比较与 [7], 类似，我们比较了每个查询检索 100 篇文档的推理时间。BM25 的搜索操作使用 CPU 测量，而 BERT 重排序的推理则使用 GPU 执行。BERT DR 的搜索运行时间是指使用 GPU 对查询进行编码以及使用 CPU 执行近似最近邻操作所需的总时间。为确保比较公平，所有搜索运行均在我们高性能计算系统的同一计算节点上执行。我们发现，BM25 的平均运行时间为 40 毫秒，BERT 重排序为 1.84 秒，BERT DR 则为 18 毫秒。该结果与 [7], 一致：在搜索时，BERT 重排序的速度约比 BERT DR 慢 100 倍，也表明 BERT DR 系统具有实际优势。

5.3. 使用 ASR N-best 的 BERT DR 平均倒数排名分数

接下来，我们展示所提出的 BERT DR 自动语音识别 N-best 扩展的平均倒数排名分数。表 2 总结了 BERT DR 使用来自标准自动语音识别系统或半监督 ASR 系统的 {2,5,10,20}-best 转录文本时的平均倒数排名分数。N 值为 1 时的 MRR 分数对应于平均倒数排名

表 2。采用标准自动语音识别和半监督自动语音识别转录文本的 N-best 扩展时，经过和未经 ANCE 训练的 BERT-DR 模型的平均倒数排名结果。% 变化表示相较于对应的 1-best 模型，平均倒数排名的相对提升/退步。星号 * 和 ** 分别表示相对于基线 1-best 系统，具有 $p < 0.05$ 和 $p < 0.01$ 的统计显著性。

模型	MRR N 值					% 变化
	1	2	5	10	20	
标准自动语音识别						
DR-early	34.81	**35.42**	32.94	33.71	34.24	+1.75 (2-best)
DR-late		35.49	36.94**	37.55**	**38.20****	+9.74 (20-best)
DR-ANCE-早期	52.41	51.77	52.14	52.75	**54.45**	+3.49 (20-best)
DR-ANCE-后期		53.29	54.88**	54.86**	**55.40****	+5.71(20-best)
半监督自动语音识别						
DR-early	40.38	39.40	38.21	37.91	38.72	-2.43 (2-best)
DR-late		40.05	40.21	**41.20**	40.93	+2.03 (10-best)
DR-ANCE-早期	56.22	57.26	57.27	57.92	**60.15****	+6.99 (20-best)
DR-ANCE-后期		57.79*	58.19**	58.79**	**59.20****	+5.30 (20-best)

分数，即表 1 中第 3 行和第 4 行的半自动语音识别列与标准自动语音识别列中的分数。粗体表示各 BERT DR 系统的平均倒数排名分数高于基线 1-best 结果，且在其他 N 值中最高。我们再次进行了配对双尾 t 检验，以比较 1-best BERT 模型与其他 N-best 模型的 MRR 分布。

使用 ANCE BERT 对半监督转录文本进行 20-best 早期融合和后期融合，分别得到 60.15% 和 59.20% 的平均倒数排名（表 2 中半监督自动语音识别部分的第 3 行和第 4 行）。考虑到对标准自动语音识别转录文本使用 1-best BERT DR 得到的平均倒数排名为 52.41%，且标准自动语音识别与人工转录文本之间的平均倒数排名差距为 14.32%（$66.73 - 52.41$），采用 20-best 早期融合方法后，平均倒数排名差距缩小至 6.58%（$66.73 - 60.15$）。这一缩减意味着相对提升超过 50%。

早期融合与后期融合 早期融合方法在不同 N 值下带来的平均倒数排名提升并不一致，而后期融合方法则始终受益于 N 值的增加。尽管 BERT DR 结合 ANCE 训练，对 20-best 转录文本进行早期融合后取得了最高的平均倒数排名分数 60.15%，但将早期融合应用于未结合 ANCE 训练的 BERT DR 时，平均倒数排名分数并未随 N 值增加而提升。在半监督自动语音识别转录文本上，观察到相较于 1-best 系统具有统计显著性（半监督自动语音识别部分第 3 行），而标准自动语音识别转录文本上的 20-best 平均倒数排名分数则未达到统计显著性（标准自动语音识别部分第 3 行）。相比之下，后期融合方法相较于 1-best 系统持续带来提升。使用 20-best 半监督自动语音识别转录文本对结合 ANCE 训练的 BERT DR 进行处理时，观察到最高的平均倒数排名分数 59.20（半监督自动语音识别部分第 4 行）。总体而言，20-best 转录文本对后期融合方法的益处最大，但对半监督自动语音识别转录文本应用未结合 ANCE 训练的 BERT DR 时除外（半监督自动语音识别部分第 2 行）。

在 fu 进一步分析中，我们考察了平均倒数排名（MRR） im-改进和退化情况，逐项查询分析了将 20-best 扩展应用于 ANCE BERT DR 系统的效果。与 1-best 系统相比，早期融合方法使 500 个评估查询中的 147 个平均 MRR 提高了 0.32，而其中 123 个查询的 MRR 分数有所降低，平均降幅为 0.22。其余 240 个查询的 MRR 分数保持不变。采用 20-best ANCE BERT 的后期融合方法使 135 个查询的 MRR 分数平均提高了 0.17，同时使 86 个查询的分数平均降低了 0.10。后期融合方法未改变其余 279 个查询的 MRR 分数。这些结果表明，在基于 1-best 系统计算文档表示时，早期融合方法带来的变化比后期融合更为显著。

6. 结论与未来工作

我们针对已知项目的 SCR 任务，探索了 BERT 重排序和 DR 系统。尽管使用 BERT 排序模型提升了检索效果，但与使用人工转录文本相比，使用自动语音识别转录文本时，BERT 重排序和 DR 系统的 MRR 分数低了 10%–14%。我们研究了半监督自动语音识别转录文本以及用于 BERT DR 的 N-best 扩展。使用半监督转录文本使 MRR 提升了 2%–5.5%，而使用 20-best 转录文本则带来了 3%–4%的增益。将两者结合后，平均提升了 6.5%。

未来的工作包括将解码格用于 BERT 检索。如表 1 所示，将 BERT 排序模型应用于格预言机转录文本后，与人工转录文本之间的 MRR 差距进一步缩小了 3%–5%。这推动了 BERT 排序模型的开发，使其能够有效地将解码格作为输入，从而克服自动语音识别转录错误。

致谢：本研究由爱尔兰科学基金会资助，是都柏林城市大学 ADAPT Centre（资助编号 13/RC/2106）项目的一部分。

7. 参考文献

[1] M. Larson 和 G. J. F. Jones，“口语内容检索：技术与方法综述”，信息检索基础与趋势，第 4 卷，第 4–5 期，第 235–422 页，2012 年。

[2] Y. Wang、A. Mohamed、D. Le、C. Liu、A. Xiao、J. Ma-hadeokar、H. Huang、A. Tjandra、X. Zhang、F. Zhang、C. Fuegen、G. Zweig 和 M. L. Seltzer，《用于混合语音识别的 Transformer 声学建模》，载于*ICASSP*，2020年，第 6874–6878 页。

[3]W-N. Hsu、B. Bolte、Y-H. H. Tsai、K. Lakhotia、R. Salakhutdinov 和 A. Mohamed，《HuBERT：通过隐藏单元掩码预测进行自监督语音表征学习》，音频、语音与语言处理汇刊，第 29 卷，第 3451–3460 页，2021 年。

[4] Y. Moriya 和 G. J. F. Jones，《通过半监督训练和声学模型适配增强用户生成视频的自动语音识别，以实现口语内容检索》，载于统计语言与语音处理，2021 年，第 73–84 页。

[5] S. Robertson 和 H. Zaragoza，“概率相关性框架：BM25 及其扩展”，信息检索基础与趋势，第 3 卷，第 4 期，第 333–389 页，2009 年。

[6] R. Nogueira 和 K. Cho，“使用 BERT 进行段落重排序”，*arXiv* 预印本，2019 年。

[7] L. Xiong、C Xiong、Y Li、K-F. Tang、J. Liu、P. N. Bennett、J Ahmed 和 A. Overwijk，“用于密集文本检索的近似最近邻负对比学习”，载于*ICLR*，2021 年。

[8] J. S. Garofolo、C. G. P. Auzanne 和 E. M. Voorhees，“TREC 口语文档检索专题：一个成功案例”，载于基于内容的多媒体信息获取——第 *1* 卷，2000 年，第 1–20 页。

[9] M. Eskevich、R. Aly、D. Racca、R. Ordelman、S. Chen 和 G.J.F. Jones，《MediaEval 2014 搜索与超链接任务》，载于 *MediaEval* 论文集，2014年。

[10] R. Jones、B. Cartere、A. Clifton、M. Eskevich、G. J. F. Jones、J. Karlgren、A. Papu、SReddy 和 Y.Yu，《TREC 2020 播客赛道概述》，载于 文本检索会议 *(TREC)*，2020年。

[11] R. Jones、B. Cartere、A. Clifton、M Eskevich、G. J. F. Jones、J. Karlgren、A. Pappu、S. Reddy 和 Y. Yu，《TREC 2021 播客赛道概述》，载于文本检索会议（*TREC*），2021年。

[12] H-Y. Lin、T-H. Lo 和 B. Chen，《用于口语文档检索的增强型 BERT 排序模型》，载于*ASRU*，2019年，第 601–606 页。[13] S-W. Fan-Jiang、T-H. Lo 和 B. Chen，《利用基于 BERT 的建模与查询改写进行口语文档检索》，载于*ICASSP*，2020年，第 8144–8148 页。

[14] Z. Dai、C. Xiong、J. Callan 和 Z. Liu，《用于即席搜索中软匹配 n 元语法的卷积神经网络》，载于国际网络搜索与数据挖掘会议，2018年，第 126–134 页。

[15] J. Devlin、M-W. Chang、K. Lee 和 K. Toutanova，《BERT：用于语言理解的深度双向 Transformer 预训练》，载于 *NAACL*，2019年，第 4171–4186 页。

[16] J. Lin、R. Nogueira 和 A. Yates，《用于文本排序的预训练 Transformer：BERT 及其后续模型》，*arXiv* 预印本，2020年。

[17] V. Karpukhin、B. Oguz、S. Min、P. Lewis、L.l Wu、S. Edunov、D. Chen 和 W-T. Yih，“用于开放域问答的稠密段落检索”，载于*EMNLP*，2020年，第 6769–6781 页。

[18] C-W. Huang 和 Y-N. Chen，“将预训练 Transformer 适配到格结构，以用于口语理解”，载于*ASRU*，2019 年，第 845–852 页。

[19] C. Liu、S. Zhu、Z. Zhao、R. Cao、L. Chen 和 K. Yu，“使用 BERT 联合编码词混淆网络和对话上下文以实现口语理解”，载于*Interspeech*，2020年，第 871–875 页。

[20] K. Ganesan、P. Bamdev、J. B、V. Amresh 和 A. Tushar，“N-best 自动语音识别 Transformer：利用多个 ASR 假设提升口语语言理解性能”，载于 *ACL*，2021 年，第 93–98 页。

[21] L. Zhu、W. Liu、L. Liu 和 E. Lin，《利用 N-best 假设改进自动语音识别错误纠正》，载于 *ASRU*，2021年，第 83–89 页。

[22] J. Johnson、M. Douze 和 H. J´egou，《使用 GPU 进行十亿规模的相似性搜索》，*IEEE*《大数据汇刊》，第 7 卷，第 3 期，第 535–547 页，2019年。

[23] R. Sanabria、O. Caglayan、S. Palaskar、D. Elliott、L. Barrault、L. Specia 和 F. Metze，《How2：用于多模态语言理解的大规模数据集》，载于 视觉基础交互与语言研讨会论文集（*ViGIL*）。NeurIPS，2018年。

[24] V. Panayotov、G. Chen、D. Povey 和 S. Khudanpur，“LibriSpeech：一个基于公有领域有声书的自动语音识别语料库”，载于 *ICASSP*，2015年，第 5206–5210 页。

[25] H. Hadian、H. Sameti、D Povey 和 S Khudanpur，“使用无格点 MMI 的端到端语音识别”，载于 *Interspeech*，2018年，第 12–16 页。[26] V. Manohar、H. Hadian、D. Povey 和 S. Khudan- pur，“使用无格点 MMI 进行声学模型的半监督训练”，载于*ICASSP*，2018年，第 4844–4848 页。[27] H. Xu、T. Chen、D. Gao、Y. Wang、K. Li、N. Goel、Y. Carmiel、D. Povey 和 S. Khudanpur，“一种用于自动语音识别的剪枝 RNNLM 格点重评分算法”，载于*ICASSP*，2018年，第 5929–5933 页。[28] T. Wolf、L. Debut、V. Sanh、J. Chaumond、C. Delangue、A. Moi、P. Cistac、T. Rault、R. Louf、M. Funtowicz、J. Davison、S. Shleifer、P. von Platen、C. Ma、Y. Jer- nite、J. Plu、C. Xu、T. Le Scao、S. Gugger、M. Drame、Q. Lhoest 和 A. M. Rush，“Transformer：自然语言处理的最新进展”，载于 *EMNLP*，2020 年，第 38–45 页。