无监督广播新闻摘要生成：最大边际相关性（MMR）与潜在语义分析（LSA）的比较研究

Majid Ramezani计算机智能系统实验室 计算机工程系 大不里士大学 伊朗大不里士
m_ramezani@tabrizu.ac.ir

Mohammad-Salar Shahryari计算机工程系 大不里士大学 伊朗大不里士*s.shahryari@tabrizu.ac.ir*

Amir-Reza Feizi-Derakhshi计算机智能系统实验室 计算机工程系 大不里士大学 伊朗大不里士
derakhshi97@ms.tabrizu.ac.ir

Mohammad-Reza Feizi-Derakhshi计算机智能系统实验室 计算机工程系 大不里士大学 伊朗大不里士*mfeizi@tabrizu.ac.ir*

摘要— 自动语音摘要生成方法分为两类：有监督方法和无监督方法。有监督方法基于一组特征，而无监督方法则依据一组规则进行摘要生成。潜在语义分析（LSA）和最大边际相关性（MMR）被认为是自动语音摘要生成中最重要且最广为人知的无监督方法。本研究旨在考察上述两种无监督方法在波斯语广播新闻转录文本摘要生成中的表现。结果表明，在通用摘要生成中，LSA 优于 MMR；而在广播新闻摘要生成的基于查询的摘要任务中，MMR 优于 LSA。

关键词—广播新闻摘要、无监督摘要生成、最大边际相关性（MMR）、潜在语义分析（LSA）

I. 引言

毫无疑问，信息在人类当今生活中的作用不容忽视。鉴于可获取的信息远超实际所需，且信息量呈指数级增长，我们需要一些工具，从各种信息资源中获取关键信息。在这方面，自动摘要系统旨在从一个或多个信息资源中提取最重要且可用的信息。它们被视为通读整篇文档的有效替代方式。此外，它们还能节省大量时间。

自动摘要领域的研究始于1950年，最初聚焦于文本自动摘要。近年来，这项研究受到了广泛关注（[1]-[11]）。自动文本摘要取得的成功促使研究人员转向自动语音摘要，即从语音文档中提取最关键信息，并去除冗余和不相关的信息 [12]。因此，人们得以更便捷地获取所需信息，并在大量语音文档中进行信息检索，例如广播新闻、讲座、演示文稿、多方会议录音、自发对话、语音邮件、电话对话等[13]-[17]。

一般而言，语音摘要方法分为两类：有监督（或基于特征）方法，例如句子位置技术、提示词以及时长和基频等声学/韵律特征；无监督（或基于规则）方法，例如最大边际相关性（MMR）和潜在语义分析（LSA）[18]。

关于广播新闻语音摘要，已有大量研究成果发表。在这些研究中，语音文档的文本转录稿是可用的。为此，研究者采用了基于文本和基于语音的特征。一些研究人员使用有监督方法研究了语音摘要。[19] 采用一组语言特征（如词频、提示词、标题词和位置）以及非语言特征（如韵律信息）对日本广播新闻进行摘要。Banerjee 和 Rudnicky [20] 利用基于文本的特征，开发了一个用于多人对话的抽取式摘要系统。 [21] 利用各种信息对广播新闻进行摘要，包括词汇和结构信息以及声学/韵律信息。其他研究（如[22] ）则考察了不同特征，例如话语、结构和词汇特征，以及声学/韵律特征，以对英语和日本广播新闻进行摘要。[23] 的作者利用支持向量机（SVM）提出了一种对语音文档进行摘要的方法，同时考虑了话语重要性和冗余。在会议摘要方面，Liu 和 Liu 在[24] 中开发了一个基于丰富的说话人敏感特征集的系统。此外，Wang 和 Cardie 在 [25] 中以领域无关的方式阐述了一个摘要生成框架。Zhang 等人在 [26] 中利用修辞结构研究了讲座语音摘要。

此外，已有多项研究采用无监督方法进行语音摘要。ClusterRank[27] 是一种用于抽取式会议摘要的基于图的系统，它采用了 TextRank[28] 算法的扩展版本（该算法用于从文本中抽取句子）。TextRank 使用 Google 的 PageRank 为每个句子计算（相关性）分数。[29]中的作者考虑采用基于图的方法来进行语音

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

摘要生成。他们将韵律特征与无监督语音摘要框架相结合。Chen 等人[30] 和 Lee 等人[31] 在讲座摘要中使用了基于图的方法。其他无监督方法，如语言建模方法（LM）和主题混合模型（TMM），则用于自动摘要，通过 n 元语法预测词项[32]。在无监督方法中，潜在语义分析和最大边际相关性被认为是转录文本摘要中最知名、应用最广泛的方法。抽取式摘要有两个目标：最大化信息覆盖量，并尽量减少冗余[33]。最大边际相关性 [34] 是一种贪心的基于查询的方法，通过逐句选择生成最终摘要，并考虑两个标准：与用户查询的相似度以及选定句子的最低冗余度。同样，口语对话也通过以下步骤进行摘要：[35] 消除语音不流畅现象、句子边界检测、识别并关联问答区域，以及主题分割。随后，通过 [36] 在最大边际相关性方法中识别并抽取关键词组，来确定最佳查询。

在潜在语义分析 [37], 中，词语的隐含语义关系会根据语境用法加以识别和获取。潜在语义分析已被 [38], 采用，并被视为一种方法，无需使用 WordNet 等词汇分类词典，即可识别文本中最重要的主题，用于广播新闻的单文档和多文档摘要生成。 [39] 还将其用于中文广播新闻的故事分段。此外，[40] 还将其用于马来语广播新闻的信息检索（IR），以减少同义词对信息检索造成的负面影响，并识别故事边界。Hofmann 提出了一种名为概率潜在语义分析 [41]的技术。由此可见，该技术源自潜在语义分析方法。与潜在语义分析相比，这种技术具有更坚实的统计基础。

本研究旨在使用最大边际相关性和潜在语义分析方法进行无监督广播新闻摘要生成，并评估其效果。为此，研究采用波斯语广播新闻。第 2 节将介绍最大边际相关性和潜在语义分析，并详细探讨其理论。随后，第 3 节将展示通过在广播新闻中实现自动摘要系统而获得的实验结果，并对这些结果进行分析。最后，第 4 节将给出结论。

II. 方法

本研究将介绍如何使用 MMR 和 LSA 对广播新闻进行摘要生成。通常，摘要生成可采用通用方式或基于查询的方式。通用摘要生成包括提取文档中最重要的信息。通过通用摘要生成得到的摘要被视为原始文档的完整替代品。相比之下，基于查询的摘要生成会根据查询所确定的用户所需信息来呈现摘要。在接下来的章节中，将从通用方法和基于查询的方法两个角度，探讨使用 LSA 或 MMR 方法进行的摘要生成。

A. 最大边际相关性（MMR） 最大边际相关性（MMR）

是信息检索中最广为人知的平衡相关性与多样性的方法之一 [42]。该方法由 Carbonell 和 Goldstein [34] 首次提出。它根据句子与输入查询的相关性，以及句子与此前选定句子的相似度，对句子进行加权组合并据此选取。该方法的核心思想是根据输入查询对句子进行迭代排序。每次迭代中，选取一个此前尚未被选中的句子，该句子与查询的相似度更高，同时与此前选定的句子相似度更低。根据该方法，对输入文档中的每个 S_i 句子，均按公式 (1) 计算最大边际相关性得分。

$$Score^{MMR}(S_i) = argmax_{Si}\{\lambda(Sim_1(S_i, Q)) - (1-\lambda)(Sim_2(S_i, Summ))\} \quad (1)$$

其中，Q 是查询向量，Summ 是被选入最终摘要的那些句子的向量，而 λ 是一个介于 0 和 1 之间的系数，用于控制每次选择中冗余程度以及与输入查询的相似度（相关性）。Sim_1 用于衡量用户查询与当前句子之间的相似度，而 Sim_2 用于评估当前句子与截至目前已选句子之间的相似度。在该方法中，Sim_1 和 Sim_2 使用余弦相似度函数作为相似度度量。MMR 得分最高的句子将被迭代选出并纳入摘要，直至摘要达到预先设定的适当长度。在这方面， λ 系数至关重要，结果的质量取决于该系数的取值。由此可见，查询相似度（相关性）与多样性之间存在权衡。这种关系受 λ 的取值影响。如果 $\lambda=1$，则相关性达到最大值。如果 $\lambda=0$，则多样性达到最大值。

尽管如此，MMR 本质上是一种基于查询的方法，也可以用作通用方法 [43]。在这种情况下，文档向量（D）会在（1）中取代查询向量（Q）。

B. 潜在语义分析（LSA）

潜在语义分析（LSA）是一种用于提取和表示词语语境用法的方法 [44]。通过这种方式，可以识别词语与其他词语集合之间的语义相似度。LSA 是一种向量空间方法，可以提取词语-词语、词语-段落和段落-段落关系。这些关系、人类认知活动与语义表示之间存在高度相关性 [45]。这种方法能够有效提取并识别说话者或作者在写作和言语表达过程中于心中形成的语义关系。此外，LSA 还可以通过恰当地近似人类判断，为机器提供词语之间的语义相似度，并预测不同文本部分之间的语义关系。

潜在语义分析是一种由两个基本步骤组成的全自动数学统计方法。第一步是用一个矩阵表示输入文本，其行对应输入文本中的不重复词语。列则对应句子、段落或其他文本单元（例如段落或文档）。在第一步中，潜在语义分析会创建一个词项-句子矩阵（A），以矩阵形式表示输入文本。矩阵（A_i）的每个列向量都表示第 i^{th} 个句子的加权词频向量。若输入文本中有 t 个不重复词语/词项和 s 个句子，则 A 是一个 t×s 矩阵（t>>s）。矩阵的每个单元格都记录相应句子中词语/词项的频数。当然，确定矩阵单元格数值的方法有多种，例如出现次数、二元

出现次数的表示方式、词频- 逆文档频率（词频-逆文档频率）、对数熵、根类型以及改进的词频-逆文档频率。

LSA 的第二步是对 A 矩阵应用奇异值分解（SVD）。应用 SVD 后，A 矩阵被分解为三个矩阵，即 V、Σ 和 U。

$$A_{t\times s} = U_{t\times c}\ \Sigma_{c\times c}\ V^{T}_{c\times s} \qquad (2)$$

其中，U 是一个 t×c 矩阵，且其列相互正交，这些列称为左奇异值。$\sum=diag(\delta_1, \delta_2, \ldots, \delta_c)$ 是一个对角矩阵，维度为 c×c，其主对角线上的元素是 A 的特征值。特征值已按降序排列在主对角线上。V^T 矩阵是一个正交的 c×s 矩阵，其列称为右奇异值。如果 rank(A)=r，则矩阵 Σ 满足以下关系：

$$\sigma_1 \geq \sigma_2 \geq \sigma_3 \geq \ldots \geq \sigma_r \geq \sigma_{r+1} = \ldots\ldots = \sigma_s = 0$$

对输入矩阵应用奇异值分解算子后得到的矩阵可作如下解释[46]：U 中的每一列（U 是一个 t×c 矩阵）都表示输入文本中的一个独特概念或主题。在这个矩阵中，行对应原始空间中的词语或术语。每个单元格表示相应术语在相应概念或主题中的权重。维度为 c×c 的对角矩阵 Σ 包含原始文本中各个概念或主题的重要性。由于单元格按降序排列，因此可以通过降维忽略重要性较低的概念。此外，V^T 是输入文本中句子的一种新表示。不同之处在于，在这种表示中，句子并非使用原始输入文本中的词语和术语来表示，而是使用 U 中主题的术语来表示。需要注意的是，尽管 LSA 是一种通用方法，但它也可以用作基于查询的方法。下文将详细介绍这两种方法。

1) 通用潜在语义分析自动摘要生成

Gong和Liu首次提出了使用潜在语义分析进行自动摘要生成的方法 [38]。此后，Steinberger和Jezek [47] 以及 Murray 等人 [48] 对其进行了改进。如下面所述，这些方法都执行通用摘要，并包含潜在语义分析的两个主要步骤以及一个句子选择步骤。在本研究中，句子选择是所有基于潜在语义分析的方法的核心重点。

a) Gong和Liu

Gong和Liu [38] 被认为是基于潜在语义分析的自动摘要中最关键的因素。根据他们的建议，在创建 t×s 矩阵并通过奇异值分解对其进行分解后，即可获得最终摘要。接着，将主要依赖 V^T 矩阵的句子选择步骤用于生成摘要。根据他们的句子选择算法，首先选择第一个概念（即最重要的概念，位于 V^T 的第一行）。然后，将与该概念最相关的句子（对应行中数值最高的句子）纳入最终摘要。对其他概念重复此过程，直到提取出预先设定数量的句子。图 1 展示了 V^T 矩阵的示例。根据该算法，首先选择概念0，然后将 S_1（概念0中数值最高的单元格）纳入最终摘要。

	S_0	S_1	S_2	S_3	S_4
概念0	0.557	0.691	0.241	0.110	0.432
概念1	0.345	0.674	0.742	0.212	0.567
概念2	0.732	0.232	0.435	0.157	0.246
概念3	0.628	0.836	0.783	0.265	0.343

图1. V^T 矩阵示例（Gong & Liu）。从 V^T 的每一行中选取得分最高的句子，直到选出预先设定数量的句子。

	概念0	概念1	概念2	概念3	长度
S_0	0.846	0.334	0.231	0.210	0.432
S_1	0.455	0.235	0.432	0.342	0.543
S_2	0.562	0.632	0.735	0.857	0.723
S_3	0.378	0.186	0.248	0.545	0.235

图2。V矩阵示例（Steinberger & Jezek）。对于 V 的每一行，计算使用 n 个概念的句子的长度。n 作为输入参数给出。达到这些长度时，将 Σ 中的值用作重要性。

	S_0	S_1	S_2	S_3	S_4
概念0	0.557	0.691	0.241	0.110	0.432
概念1	0.345	0.674	0.742	0.212	0.567
概念2	0.732	0.232	0.435	0.157	0.246
续 3	0.628	0.836	0.783	0.265	0.343

图3。V^T 矩阵示例（Murray 等人）。从 V^T 的每一行中，选取得分较高的一个或多个句子。所选句子的数量由 Σ 确定。

该算法存在一些缺陷。首先，降维空间的大小应与最终摘要的长度相等。因此，可能会选入一些重要性较低的句子。其次，有些句子与所选概念相关，但对应单元格并非相应行中的最高值。按照该算法，这些句子无法被纳入摘要（尽管它们有助于提高最终摘要的相关性）。第三，与潜在语义分析不同，该算法将所有概念视为同等重要，而其中一些概念在原始文档中并不重要。

b) Steinberger和Jezek

Steinberger和Jezek [47] 提出了一种算法，以改进Gong和Liu提出的方法的缺陷。在该算法中，完成潜在语义分析（LSA）的两个基本步骤后，使用Σ和V矩阵来选择句子。该算法的句子选择从计算V矩阵各行中句子向量的长度开始。计算句子向量的长度时，使用索引小于或等于新空间维度数量的概念。该算法将新空间的维度大小作为输入。Σ是一个对角矩阵，其中包含A矩阵的特征值，这些特征值按降序排列在主对角线上。从概念上讲，该矩阵体现了原始文档中提出的概念转移到降维空间后的重要性。因此，将该矩阵中的数值作为乘法系数时，与文档关联度更高的概念就会具有更高的重要性。如果新空间的维度为n，则根据（3）计算第i^{th} 个句子的长度。

$$Length\ i = \sqrt{\sum_{j=1}^{n} V_{ij} \times \sum_{jj}} \qquad (3)$$

计算完所有句子的长度后，选择长度最长的句子作为摘要。由此可见，该算法可以从每个概念中选择多个句子。与重要概念关联度更高的句子也会被选中。显然，该算法并未将奇异值分解（SVD）提取出的各种概念视为同等重要。图2展示了一个示例V矩阵。如果新空间的维度为3，则通过考察前3个概念来计算句子的长度，最后，S_2 被选为第一个句子。

c) Murray 等人

L算法，该算法由 Murray 等人提出，用于解决 Gong 和 Liu 提出的算法存在的问题。该算法在进行句子选择之前会执行潜在语义分析方法的两个步骤。该算法与 Gong 和 Liu 提出的算法相似，但主要区别在于：这里会选出 n 个最佳句子，而不是从每个概念中选出最佳句子。n 的值通过矩阵 Σ 中的奇异值计算得出。该算法在句子选择中使用 Σ 和 V^T。通过计算 Σ 中相应奇异值占 Σ 中所有奇异值总和的百分比，确定从每个概念中选出的句子数量。因此，每个概念中选出多个句子是可行的。此外，不同提取概念的重要性并非一视同仁。此外，维度缩减不取决于最终摘要的长度。图3展示了一个 V^T 矩阵示例。假设必须从概念0中选出两个句子，并从概念1中选出一个句子。根据该算法，从概念0中选出 S_0 和 S_1，从概念1中选出 S_2。

2) 基于查询的潜在语义分析 在基于查询的

在这种方法中，得到的结果会受到用户查询的影响。通常，可以通过字面匹配用户查询中的词语与文档中的词语，将用户查询用于信息检索。但需要注意的是，字面或词汇匹配及对应关系可能导致错误结果 [37]。通常，同一个概念可以有不同的表述方式。对于这一问题，这类文档无法被识别，也不会被视为相关文档。此外，大多数词语都具有多义性。因此，可能会将用户查询中的词语与无关文档中的词语进行比较。在基于查询的潜在语义分析中，为了解决词汇匹配的问题，文档和用户查询都会被投影到新空间中。在这个新空间里，即使查询与文档没有任何共同词语，两者之间也可能具有较高的余弦相似度。这是因为根据潜在语义分析，查询中的词语和句子在语义上可能相似。可以说，潜在语义分析方法使用奇异值分解算子，在降维空间中表示文档和查询，而不是根据彼此独立的词语向量来表示它们。由于新空间中的因子数或维度数少于分别表示各个词语时的数量，因此词语并非彼此独立。例如，如果两个词语出现在相似的语境中，它们在降维空间中会由同一个向量表示。这样一来，即使句子（文档）与查询没有共同词语，也仍然可以检索到。与通用方法类似，在基于查询的潜在语义分析中，首先将处理后的文本表示为矩阵 A。然后，通过应用奇异值分解（参见（2）），将其投影到潜在语义空间中。

最后，应使用（4）将用户查询 q 转换到新空间中。这也为评估新语义空间中查询向量（$\hat{q}$）与句子向量之间的相似度提供了逻辑和数学依据。

$$\hat{q} = q^T U_{t\times r} \sum^{-1}_{r\times r} \tag{4}$$

其中 q 是查询向量。然后，将查询向量乘以 $U_{t\times r}$，即可将其从原始空间转换到新空间。最后，再将其乘以 $\sum^{-1}_{r\times r}$，为每个维度赋予适当的权重。这样便为在新空间中比较查询向量与句子向量提供了所需的基础。为此，选取输入文本中余弦相似度高于所需阈值的 n 个句子作为摘要。

3) 概率潜在语义分析（PLSA） 概率潜在语义

分析（PLSA）[41] 是一种在潜在语义分析技术基础上，通过强调统计方面而得到的方法。换言之，该技术通过强调最大似然原理，具有更坚实的统计基础。PLSA 使用 Hofmann方面模型。根据此方法，一组未观测变量 $z\epsilon Z=\{z_1, z_2,..., z_k\}$（即文档中出现的主题）与两组观测变量相关联：一组是文档 $d\epsilon D=\{d_1, d_2, ..., d_m\}$（此处视为句子），另一组是这些文档中的词语 $w\epsilon W=\{w_1, w_2, ... w_n\}$。在此方法中，通过观察词语与文档（w, d）的共现，使用公式（5）计算每种共现的概率：

$$P(d, w)= P(d)\, P(w|d)\ ,\ P(w|d)=\sum_{(z\in Z)} P(w|z)P(z|d)$$

$$P(d, w)= P(d) \sum_{z\in Z} P(w|z)P(z|d) \tag{5}$$

其中，P(d) 是选取文档 d 的概率，P(w|z) 是从主题 z 中选取词语 w 的概率，P(z|d) 表示在文档 d 中选取主题 z 的概率，而 P(d, w) 表示每个文档与词语对的似然。下一步，根据最大似然原理，通过最大化对数似然来优化 P(d)、P(w|z) 和 P(z|d) 的值，并确定其最大值。为此，可以使用期望最大化算法。这种算法是统计学中的一种迭代方法，用于在模型依赖一组未观测到的潜变量 [41] 时确定最大似然。该算法包含两个步骤：期望步（E步），计算与潜变量相关的先前概率；以及最大化步（M步），更新参数。Hofmann 提出了用于 E步的（6）：

$$P(z|d, w)=\frac{P(z)P(d|z)P(w|z)}{\sum_{\acute{z}\in Z} P(\acute{z})P(d|\acute{z})P(w|\acute{z})} \tag{6}$$

此外，他还为 M 步提出了 (7)、(8) 和 (9)：

$$P(w|z) = \frac{\sum_d n(d,w)\, P(z|d,w)}{\sum_{d,w'} n(d,w')P(z|d,w')} \tag{7}$$

$$P(d|z) = \frac{\sum_w n(d,w)P(z|d,\ w)}{\sum_{d',w} n(d',w)P(z|d',w)} \tag{8}$$

$$P(z) = \frac{\sum_{d,w} n(d,w)P(z|d.w)}{\sum_{d,w} n(d,w)} \tag{9}$$

其中，n(d, w) 是 w 在 d 中的频数。

E步包括计算该词的概率，*w*出现在文档 *d* 中可以由因子 *z*（主题）来解释。然后，通过交替执行 E 步和 M步，可以找到收敛点。该点是对数似然的局部最大值。

基于PLSA的摘要算法包含以下步骤:

i. 每篇文档均以词项频数矩阵表示。ii. 按照式（7）、（8）和（9）的定义计算P(z)、P(d|z)和P(w|z)的值，直至满足期望最大化算法的收敛准则。P(d|z)表示给定主题z中，文档d（句子）的重要性。P(z)表示文档d中主题z的重要性。iii. 使用式（10）计算每个句子的R分数。在这种情况下，根据句子所包含的主题为所有句子评分。

$$R=\sum_z P(d|z)P(z) = P(d) \quad (10)$$

iv. 最后，选取R分数最高的句子来构成摘要。

III. 实验与结果

A. 数据集。语料

库来自对超过15小时波斯语广播新闻的人工转录，包含58篇新闻文档。四名母语转录员在45天内从三个广播频道和四个电视频道的新闻中收集了该语料库。语料库包含超过115,000个词语和7,000个句子。每篇新闻文档平均包含122.5个句子，语料库中句子的平均长度为16.5。为生成用作评估摘要系统参考的标准摘要，四名人工摘要员对所有文档进行了摘要生成，涵盖通用摘要生成和查询式摘要生成（针对预先指定的查询）。每份标准摘要平均包含36.7个句子。

B. 结果与讨论

为了评估所提出的摘要系统，我们使用了精确率、召回率和F1分数，并采用压缩率为30%的ROUGE-1进行评估。

$$Precision = \frac{TP}{Tp+ FP} \quad (11)$$

$$Recall = \frac{TP}{TP+FN} \quad (12)$$

$$F1\text{-}score = \frac{2 \times Precision \times Recall}{Precision+Recall} \quad (13)$$

$$ROUGE\text{-}n = \frac{\sum_{S\in\{ReferenceSummaries\}}\sum_{gram_n\in S} Count_{match}(gram_n)}{\sum_{S\in\{ReferenceSummaries\}}\sum_{gram_n\in S} Count(gram_n)} \quad (14)$$

其中，在式（11）和（12）中，TP、FP和FN分别表示真正例、假正例和假负例的数量。在式（14）中，n表示n元语法的长度（本研究中n=*1*），$Count_{match}(gram_n)$是候选摘要中共同出现的n元语法的最大数量，而$Count(gram_n)$是参考摘要中的n元语法数量。

1）使用最大边际相关性进行摘要生成 如前所述，在MMR中 d

文档经过摘要处理，以在相关性和冗余之间取得平衡。随后设定 λ （见（1））。λ 的值越高，对相关性的强调就越多。而 λ的值越低，

表 I. 基于查询的MMR评估结果

λ	精确率 (%)	召回率 (%)	F值（%）	ROUGE-1
0.0	33.34	21.87	26.22	0.2437
0.1	35.72	23.54	28.22	0.2494
0.2	35.72	23.54	28.22	0.2267
0.3	65.87	46.25	54.17	0.4310
0.4	71.43	49.28	58.09	0.4487
0.5	61.90	42.05	49.88	0.3982
0.6	76.19	49.26	59.56	0.4847
0.7	72.22	48.79	57.97	0.4839
0.8	69.84	46.94	55.88	0.4681
0.9	69.84	46.94	55.88	0.4681
1.0	**83.34**	**54.66**	**65.69**	**0.5379**

表 II. 通用MMR的评估结果

λ	精确率（%）	召回率 (%)	F值（%）	ROUGE-1
0.0	22.22	20.04	21.06	0.2082
0.1	24.61	22.26	23.36	0.2249
0.2	19.84	18.10	18.92	0.1678
0.3	37.30	33.13	35.08	0.3345
0.4	38.10	33.45	35.61	0.3227
0.5	33.32	29.56	31.32	0.2921
0.6	26.19	23.18	24.58	0.2275
0.7	**40.48**	**35.68**	**37.91**	**0.3807**
0.8	**40.48**	**35.68**	**37.91**	**0.3807**
0.9	**40.48**	**35.68**	**37.91**	**0.3957**
1.0	**40.48**	**35.68**	**37.91**	**0.3957**

多样性就越高。本研究通过为 λ 赋予不同的值 *{0, 0.1, 0.2, ..., 1.0}*，考察了采用这种方法得到的结果。表 I 展示了使用基于查询的MMR得到的四项评估指标的结果。数据集中的每篇新闻文档都预先设定了相应的查询。

可以看出，随着 λ 的增加，所有评估指标均有所提升，因此在 $\lambda=1$ 时取得了最佳结果。因此可以推断，在仅考虑选定句子与查询之间相似度的情况下，采用基于查询的MMR进行广播新闻摘要生成可取得最佳结果（$\lambda=1$）。

表 II 展示了使用通用MMR生成的摘要在四项评估指标上的评估结果。可以看出，与通用MMR类似，随着 λ 的增加，系统性能有所提升，因此在 λ 取最大值（0.7、0.8、0.9 和 1.0）时取得了最佳结果。比较基于查询的MMR与通用MMR可知，基于查询的MMR总体上优于通用MMR。

2）使用潜在语义分析进行摘要生成 表 III 展示了评估

采用通用潜在语义分析和基于查询的潜在语义分析进行广播新闻摘要生成的结果。在潜在语义分析的通用方法中，Murray 等人提出的算法取得了最佳结果。可以看出，总体而言，通用潜在语义分析在所有评估指标上均优于基于查询的潜在语义分析。表 IV 展示了基于概率潜在语义分析的摘要生成评估结果。摘要生成针对 n= *2*、3、4、5、6 的不同取值进行。由于在 n=*1* 时，只有一个主题的句子会被选入最终摘要，因此未考虑 n=*1* 的取值。可以看出，将 n 的值增加至 n=*3* 可取得最佳结果。

3）最大边际相关性与潜在语义分析：哪种方法最适合广播新闻摘要生成?

表 V 对比了通用潜在语义分析与最大边际相关性，表 VI 对比了基于查询的潜在语义分析与基于查询的最大边际相关性。可以看出，就广播新闻

表 III.　通用潜在语义分析和基于查询的潜在语义分析的评估结果

	方法	精确率 (%)	召回率 (%)	F值 (%)	ROUGE-1
Generic	Gong & Liu	55.96	49.39	52.46	0.5737
	Steinberger & Jezek	57.14	50.68	53.70	0.5339
	(n=1)				
	Murray 等人	**60.32**	**53.35**	**56.61**	**0.5930**
基于查询		54.76	42.59	47.08	0.4854

表 IV. 概率潜在语义分析的评估结果

n	精确率（%）	召回率 (%)	F值（%）	ROUGE-1
2	**58.31**	28.96	38.39	0.3167
3	43.32	**38.08**	**40.48**	**0.4012**
4	53.33	23.97	32.57	0.2507
5	50.83	21.47	29.56	0.2258
6	48.32	19.48	38.68	0.2035

表 V. 通用潜在语义分析与通用MMR的比较

方法	精确率 (%)	召回率 (%)	F值 (%)	ROUGE-1
潜在语义分析（Murray 等人）	**60.32**	**53.35**	**56.61**	**0.5930**
最大边际相关性 (λ=1.0)	40.48	35.68	37.91	0.3957

表 VI。比较基于查询的潜在语义分析与基于查询的MMR

方法	精确率 (%)	召回率 (%)	F值 (%)	ROUGE-1
LSA	54.76	42.59	47.08	0.4854
最大边际相关性 (λ=1.0)	**83.34**	**54.66**	**65.69**	**0.5379**

摘要生成而言，通用MMR与基于查询的MMR相比，基于查询的MMR取得了更好的结果。此外，通用潜在语义分析与基于查询的潜在语义分析相比，通用潜在语义分析取得了更好的结果。另一方面，考虑到MMR本质上是一种基于查询的方法，而LSA本质上是一种通用方法，基于查询的方法和通用方法分别取得了最佳结果。

IV. 结论

本研究探讨了无监督广播新闻摘要生成。最大边际相关性（MMR）和潜在语义分析（LSA）是自动语音摘要领域最知名的无监督方法，广泛应用于摘要生成。我们提出了若干系统，用于生成通用和基于查询的自动广播新闻摘要，这些系统基于最大边际相关性（MMR）和潜在语义分析（LSA）。结果表明，在基于查询的摘要生成中，MMR 优于 LSA；而在通用摘要生成中，LSA 优于 MMR。

参考文献

[1] E. Lloret 和 M. Palomar，“文本摘要研究进展：文献综述”，《人工智能评论》，第 37 卷，第 1 期，第 1–41 页，2012 年，doi: 10.1007/s10462-011-9216-z。[2] H. Saggion 和 T. Poibeau，“自动文本摘要：过去、现在与未来”，载于《多源多语言信息抽取与摘要》，Springer，2013 年，第 3–21 页。[3] M. Ramezani 和 M.-R. Feizi-Derakhshi，“自动文本摘要：概述”，《应用人工智能》，第 28 卷，第 2 期，第 178–215 页，2014 年，doi: 10.1080/08839514.2014.862783。[4] Y. K. Meena、P. Dewaliya 和 D. Gopalani，“抽取式自动文本摘要的最优特征集”，载于《高级计算与通信技术（ACCT）》，2015 年第五届国际会议，2015 年，第 35–40 页，doi: 10.1109/ACCT.2015.123。[5] M. Ramezani 和 M.-R. Feizi-Derakhshi，“使用 FarsNet 的基于本体的自动文本摘要”，《高级计算机科学国际期刊》，第 4 卷，第 2 期，第 88–96 页，2015 年。[6] M. Ramezani 和 M.-R. Feizi-Derakhshi，“实现更连贯的自动文本摘要；一种基于本体的方法”，《先进数学与计算机科学杂志》，第 1–15 页，2016 年，doi：10.9734/BJMCS/2016/27549。[7] P. Sethi、S. Sonawane、S. Khanwalker 和 R. B. Keskar，“新闻文章的自动文本摘要”，载于 2017 年大数据、物联网与数据科学国际会议（BID），2017 年，第 23–29 页，doi：10.1109/BID.2017.8336568。[8] M.-Y. Day 和 C.-Y. Chen，“用于自动文本摘要的人工智能”，载于 2018 年 IEEE 信息重用与集成国际会议（IRI），2018 年，第 478–484 页，doi: 10.1109/IRI.2018.00076。[9] J. N. Madhuri 和 R. G. Kumar，“使用句子排序进行抽取式文本摘要”，载于 2019 年数据科学与通信国际会议（IconDSC），2019 年，第 1–3 页，doi: 10.1109/IconDSC.2019.8817040。[10] W. S. El-Kassas、C. R. Salama、A. A. Rafea 和 H. K. Mohamed，“EdgeSumm：用于自动文本摘要的基于图的框架”，《信息处理与管理》，第 57 卷，第 6 期，文章编号 102264，2020 年，doi：10.1016/j.ipm.2020.102264。[11] C. Zong、R. Xia 和 J. Zhang，“自动文本摘要”，载于《文本数据挖掘》，Springer，2021 年，第 285–333 页。[12] G. Yasmin、S. Debnath 和 A. K. Das，“无需语言学知识、基于声学特征进行帧选择的自动语音摘要”，载于《模式识别中的计算智能》，Springer，2022 年，第 463–470 页。[13] S.-H. Liu、K.-Y. Chen、B. Chen、H.-M. Wang 和 W.-L. Hsu，“利用流形学习进行广播新闻抽取式摘要”，载于 2017 年 IEEE 声学、语音与信号处理国际会议（ICASSP），2017 年，第 5805–5809 页，doi: 10.1109/ICASSP.2017.7953269。[14] M. V. M. Cirne 和 H. Pedrini，“VISCOM：一种使用颜色共生矩阵的稳健视频摘要方法”，《多媒体工具与应用》，第 77 卷，第 1 期，第 857–875 页，2018 年，doi: 10.1007/s11042-016-4300-7。[15] M. Elfeki 和 A. Borji，“通过动作显著性排序进行视频摘要”，载于 2019 年 IEEE 冬季计算机视觉应用会议（WACV），2019 年，第 754–763 页，doi: 10.1109/WACV.2019.00085。[16] S.-H. Liu、K.-Y. Chen 和 B. Chen，“结合邻近度与句子相关性信息，改进用于广播新闻抽取式摘要的语言建模”，《ACM 亚洲低资源语言信息处理汇刊》，第 19 卷，第 3 期，第 1–19 页，2020 年，doi：10.1145/3377407。[17] V. Tiwari 和 C. Bhatnagar，“视频摘要近期研究综述：方法与技术”，《多媒体工具与应用》，第 1–35 页，2021 年，doi：10.1007/s11042-021-10977-y。[18] H. H. Kim，“社交语音摘要”，《美国信息科学与技术学会会议录》，第 48 卷，第 1 期，第 1–4 页，2011 年。[19] A. Inoue、T. Mikami 和 Y. Yamashita，“利用韵律信息改进语音摘要”，载于 Speech Prosody 2004 国际会议，2004 年，第 0–3 页。[20] S. Banerjee 和 A. I. Rudnicky，“用于自动检测多人对话中重要话语的抽取式摘要基线”，载于口语技术研讨会，2008 年。SLT 2008。IEEE，2008 年，第 177–180 页，doi：10.1109/SLT.2008.4777869。[21] S. R. Maskey，“广播新闻语音的自动摘要”，2008 年。[22] J. Zhang、R. R. H. Y. Chan、P. Fung 和 L. Cao，“广播新闻与讲座语音摘要的比较研究”，载于 INTERSPEECH，2007 年，第 2781–2784 页，访问日期：2013 年 10 月 5 日。[在线]。可获取于: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.152.5325& rep=rep1&type=pdf。[23] H. Lee、Y. Chou、Y.-B. Wang 和 L.-S. Lee，“通过结构化支持向量机联合考虑话语重要性与冗余的有监督口语文档摘要”，2012 年。[24] F. Liu 和 Y. Liu，“探索会议摘要中的说话人特征”，载于 INTERSPEECH，2010 年，第 2518–2521 页。[25] L. Wang 和 C. Cardie，“面向聚焦式会议摘要的领域无关摘要生成”，载于 ACL (1)，2013 年，第 1395–1405 页。[26] I. Zhang、H. Yuan 和 X. Pan，“ 用于讲座语音摘要的修辞状态支持向量机 ”，《信息技术杂志》，第 13 卷，1450，2014 年，7，1447 10 3923 2014 1447 1450 第 期，第 – 页，doi：. /itj. . . .[27] N. Garg、B. Favre、K. Reidhammer 和 D. Hakkani Tür，“ClusterRank：一种用于会议摘要的基于图的方法”，Idiap，2009 年。[28] R. Mihalcea 和 P. Tarau，“TextRank：为文本带来秩序”，2004 年。[29] S. K. Jauhar、Y.-N. Chen 和 F. Metze，“基于韵律、采用双层互增强随机

游走的无监督语音摘要”，2013 年。[30] Y.-N. Chen、Y. Huang、C.-F. Yeh 和 L.-S. Lee，“通过对自动提取的关键词构建的图进行随机游走，实现口语讲座摘要”，载于 INTERSPEECH，2011 年，第 933–936 页。[31] H. Lee 等，“通过语义结构化和面向个性化学习的原型课程讲座系统实现口语知识组织”，《IEEE/ACM 音频、语音与语言处理汇刊》，第 22 卷，第 5 期，第 883–898 页，2014 年，doi：10.1109/TASLP.2014.2310993。[32] S.-H. Liu、K.-Y. Chen、B. Chen、H.-M. Wang、H.-C. Yen 和 W.-L. Hsu，“结合相关性语言建模与清晰度度量进行抽取式语音摘要”，《IEEE/ACM 音频、语音与语言处理汇刊》，第 23 卷，第 6 期，第 957–969 页，2015 年，doi：10.1109/TASLP.2015.2414820。[33] D. Gillick、K. Riedhammer、B. Favre、D. Hakkani-t 和 D. Hakkani-Tur，“会议摘要的全局优化框架”，载于《声学、语音与信号处理》，2009 年。ICASSP 2009。IEEE 国际会议，2009 年，第 2 期，第 4769–4772 页，doi：10.1109/ICASSP.2009.4960697。[34] J. Carbonell 和 J. Goldstein，“使用最大边际相关性（MMR）进行基于多样性的重排序，以重新排列文档并生成摘要”，《第 21 届 ACM SIGIR 信息检索研究与开发年会论文集——SIGIR ’98》，第 335–336 页，1998 年，doi：10.1145/290941.291025。[35] K. Zechner，“开放领域口语对话的自动摘要”，Google，2001 年。[36] K. Riedhammer、B. Favre、D. Hakkani-t 和 D. Hakkani-Tur，“一种基于关键短语的交互式会议摘要方法”，载于口语技术研讨会，2008 年。SLT 2008。IEEE，2008 年，第 153–156 页，doi：10.1109/SLT.2008.4777863。[37] S. C. Deerwester、G. W. Furnas、T. K. Landauer、R. A. Harshman 和 S. T. Dumais，“通过潜在语义分析进行索引”，《JASIS》，第 41 卷，第 6 期，第 391–407 页，1990 年，doi：10.1002/(SICI)1097- 4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9。[38] Y. Gong 和 X. Liu，“使用相关性度量和潜在语义分析进行通用文本摘要”，载于第 24 届 ACM SIGIR 信息检索研究与开发国际年会论文集，2001 年，第 19–25 页，doi：10.1145/383952.383955.[39] Y. Yang 和 L. Xie，《用于中文广播新闻基于 texttiling 的自动故事分段的子词潜在语义分析》，载于《中文口语处理》，2008. *ISCSLP’08.* 第 *6* 届国际研讨会，2008 年，第 00 卷，第 c 期，第 1–4 页，doi：10.1109/CHINSL.2008.ECP.101.[40] Z. A. K. Aleqili 和 T. P. Tan，《马来语广播新闻故事边界的无监督识别》，《新兴技术与网络智能期刊》，第 5 卷，第 1 期，2013 年。[41] T. Hofmann，《概率潜在语义索引》，载于第 22 届 ACM SIGIR 信息检索研究与开发国际年会论文集，1999 年，第 50–57 页，doi：10.1145/312624.312649。[42] R. M. Alguliev、R. M. Aliguliyev、M. S. Hajirahimova 和 C. a. Mehdiyev，《MCMR：最大覆盖与最小冗余文本摘要模型》，《专家系统与应用》，第 38 卷，第 12 期，第 14514–14522 页，2011 年 11 月，doi：10.1016/j.eswa.2011.05.033。[43] A. Nenkova 和 K. R. McKeown，《自动摘要》，第 5 卷，第 2 期。Now Publishers Inc，2011 年。[44] T. K. Landauer、S. T. Dumais 和 S. T. Dutnais，《柏拉图问题的解答：知识获取、归纳与表征的潜在语义分析理论》，《心理学评论》，第 104 卷，第 2 期，第 211–240 页，1997 年，doi：10.1037/0033-295X.104.2.211。[45] T. K. Landauer、D. S. McNamara、S. Dennis 和 W. Kintsch，《潜在语义分析手册》。Psychology Press，2013 年。[46] A. Nenkova 和 K. McKeown，《文本摘要技术综述》，载于《文本数据挖掘》，C. C. Aggarwal 和 C. Zhai 编。美国马萨诸塞州波士顿：Springer，2012 年，第 43–76 页。[47] J. Steinberger 和 K. Jezek，《在文本摘要与摘要评估中运用潜在语义分析》，载于 *ISIM’04* 论文集，2004 年，第 93–100 页。[48] G. Murray、S. Renals 和 J. Carletta，《会议录音的抽取式摘要》。 2005，[在线]。可访问：http://www.isca- speech.org/archive/interspeech_2005。J. Clerk Maxwell，《电学和磁学论》，第 3 版，第 2 卷。牛津：Clarendon，1892 年，第 68–7 页