

理解低资源主题建模中的跨域适应

Pritom Saha Akash Kevin Chen-Chuan Chang
University of Illinois at Urbana-Champaign, USA
{ pakash2, kcchang } @illinois.edu

Abstract

主题建模在揭示文本语料库中的隐藏语义结构方面起着至关重要的作用，但现有模型在低资源环境中表现不佳，因为目标领域数据有限会导致主题推断不稳定和不连贯。我们通过正式引入低资源主题建模的领域适应来解决这一挑战，在这种情况下，高资源领域为低资源目标领域提供信息，但不会被不相关内容淹没。我们建立了一个有限样本泛化界限，显示有效的知识转移依赖于两个领域中的稳健性能，最小化潜在空间差异，并防止过度拟合数据。基于这些见解，我们提出了 DALTA (Domain-Aligned Latent Topic Adaptation)，一个新的框架，采用共享编码器提取领域不变特征，专用解码器处理领域特定的细微差别，并通过对抗对齐选择性地转移相关信息。在不同的低资源数据集上进行的实验表明，DALTA 在主题连贯性、稳定性和可转移性方面始终优于最先进的方法。

1 介绍

在当今的数字时代，各个领域都产生了大量的非结构化文本语料，使得提取有意义的见解极为重要。主题建模帮助揭示文本中的隐藏模式，从而实现文档分类、文本摘要、内容推荐和趋势分析等应用。虽然概率主题模型 (Blei et al., 2003; Blei and Lafferty, 2006a,b; Mcauliffe and Blei, 2007) 仍然被广泛使用，但深度学习推动了更高级变体的出现。例如，神经主题模型 (NTMs) (Miao et al., 2016; Srivastava and Sutton, 2017; Nguyen and Luu, 2021; Dieng et al., 2020) 利用深度学习来改进主题表示，通常采用变分自编码器 (VAEs) 来建模潜在空间。进一步的进展，例如上下文主题模型 (CTMs) (Bianchi et al., 2020a,b; Grootendorst, 2022)，结合预训练语言模型，通过捕捉上下文依赖性来增强主题连贯性。

尽管取得了成功，主题模型通常假设拥有足够的数据来学习有意义和连贯的主题。然而，在许多现实场景中，尤其是在新兴或小众领域，数据收集受到资源限制、访问受限或知识

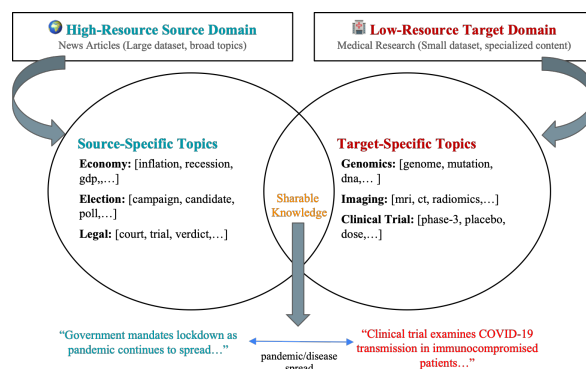


Figure 1: 一个高资源领域（新闻）和一个低资源领域（医学研究）如何共享某些主题（例如，“大流行/疾病传播”）的示意图，同时每个领域保留特定的内容（例如，“经济”与“基因组学”）。目标是只转移相关知识，而不引入无关信息。

快速发展的限制。例如，像量子机器学习这样的领域可能公开的文件不到 1000 份，而法律或医学文本等专业领域受到严格的隐私法规限制。在这种情况下，传统的主题模型难以从资源匮乏的语料库中提取出稳定和连贯的主题。

尽管已经有若干尝试来解决低资源话题建模，但它们存在一些局限性。例如，较早的方法 (Sia and Duh, 2021) 通过自适应平衡离散标记统计与预训练词嵌入来优化 LDA，允许经常出现的词依赖于计数，而不常出现的词则借助外部表示。同样地，基于嵌入的 NTMs (Duan et al., 2022, 2021) 利用预训练词嵌入作为可转移知识，通过自适应学习话题嵌入实现有效的泛化。然而，由于词语语义会随着上下文的变化而变化，静态嵌入可能无法适应未见过的任务，从而限制了它们在特定领域的低资源环境下的有效性。为了解决这个问题，Context-Guided Embedding Adaptation (Xu et al., 2024) 通过根据目标语料库中的句法和语义依赖来调整词嵌入，动态优化话题-词关系。这改善了低资源环境下的话题一致性，但它仍然仅依赖于目标领域的数据，当极端的数据不足限制了有效的调整时，其效果有限。

从上述讨论来看，低资源主题模型的关键挑

战在于如何有效利用外部知识，同时保留目标领域的特定细微差别。如图 1 所示，考虑一个高资源源领域，比如新闻文章，以及一个低资源目标领域，比如医学研究。源领域提供了广泛的主题——其中一些，如疫情和心理健康，可以提供可共享的知识（这里显示为疫情/疾病传播），而其他主题，如选举或经济，则保持为源领域特有且与医学研究无关。同样，目标领域包含特定的概念（例如基因组学、临床试验），这些概念在源领域中并不存在。通过专注于疫情/疾病传播，新闻文章（例如，“政府命令封锁……”）可以为医学研究提供信息（例如，“临床试验研究 COVID-19 传播……”），使模型能够转移相关的健康相关内容。然而，源领域中的选举或经济等主题与目标领域中的基因组学或成像无关。因此，关键挑战是如何在尽量共享可共享主题的同时，确保领域特定的信息保持完整，从而在数据稀少的情况下增强主题发现。

基于这些见解，领域适应 (Ben-David et al., 2010) 提供了一个有前景的框架，可以在主题建模中弥合高资源源数据与低资源目标需求之间的差距。尽管领域适应在监督任务中广泛用于对齐跨领域特征 (Zhao et al., 2019; Li et al., 2021)，其在无监督主题建模中的潜力仍然很大程度上未被探索。为了解决这个问题，我们通过建立一个泛化界限形式化主题建模的领域适应，这表明成功的迁移取决于：(i) 在源数据和目标数据上都取得良好的性能，(ii) 最小化源和目标潜在表示之间的差异，以便仅转移相关知识，以及 (iii) 应用正则化以防止过拟合到源领域。鉴于此，我们推出了新颖的界限最小化算法 DALTA（领域对齐潜在主题适应），该算法利用这些原则选择性地从高资源领域转移有用的主题结构，同时保留目标领域的独特语义特征。

总之，我们的工作提供了以下贡献：

- 我们推导了用于低资源主题建模中的领域适应的有限样本泛化界限，证明了有效转移依赖于源域和目标域的稳健表现，潜在表示的对齐，以及适当的正则化以防止过拟合。
- 基于这些理论见解，我们提出了一个新颖的框架，DALTA（Domain-Aligned Latent Topic Adaptation，域对齐潜在主题自适应），该框架利用共享编码器提取域不变的主题表示，并采用专用解码器捕捉目标特定的语义细微差别。据我们所知，DALTA 是第一个在严格的理论基础上联合优化潜在对齐和主题建模中的域特定学习的方法。
- 我们在不同的低资源数据集上进行了广泛的实验，结果表明，DALTA 在话题一致性、多样性和可迁移性方面持续优于最新的方法。

2 相关工作

2.1 神经主题模型

传统的概率主题模型，例如隐含狄利克雷分布 (LDA) (Blei et al., 2003) 及其扩展 (Blei and Lafferty, 2006a,b; Mcauliffe and Blei, 2007)，已广泛用于发现文本中的潜在语义结构。然而，它们依赖于忽略词序和上下文意义的词袋假设，限制了它们捕捉细微语义的能力。为了克服这些限制，神经主题模型 (NTMs) 利用深度学习，特别是变分自编码器 (VAEs) (Kingma, 2013)，来学习更丰富和灵活的主题表示 (Miao et al., 2016; Srivastava and Sutton, 2017)。上下文文化主题模型 (CTMs) (Bianchi et al., 2020a,b; Grootendorst, 2022) 通过结合预训练语言模型 (PLMs) 进一步改善主题一致性，从而能够捕捉上下文依赖关系。更近期的研究，例如 UTopic (Han et al., 2023)，通过引入对比学习和术语加权来增强主题的一致性和多样性，有效地优化了主题表示。同样，NeuroMax (Pham et al., 2024) 通过最大化主题之间的互信息并强制结构化主题正则化来提高一致性以增强 NTMs。

尽管这些进展存在，NTM 在跨领域或低资源环境中的应用仍面临挑战。大多数模型假设拥有充足的训练数据，并在领域适应性方面表现不佳，因为词汇变化和分布变化会导致主题表示的错位。虽然最近的工作，例如基于提示的 NTM (Pham et al., 2023) 和 LLM 驱动的上下文扩展主题模型 (Akash and Chang, 2024)，尝试利用外部知识源来实现更稳健的主题发现，但它们并没有明确解决如何在不同数据分布的领域之间转移主题知识。因此，目前的 NTM 通常无法有效地超越其训练领域进行泛化，这需要新的方法来平衡领域不变的知识转移与领域特定的适应能力。

2.2 低资源主题建模

小样本主题建模已经通过元学习和基于嵌入的适应进行了探索，但每种方法都有其固有的局限性。小样本方法 (Iwata, 2021) 尝试从有限样本中学习特定于任务的先验以进行泛化。然而，它们的刚性概率假设使其无法有效捕捉新领域中出现的上下文变化和微妙的主题差异。基于嵌入的神经主题模型 (NTMs) (Duan et al., 2022, 2021) 通过利用预训练的词嵌入来改善泛化，从而允许主题发现超越简单的词共现模式。然而，它们对静态表示的依赖性限制了其在词语语义在不同上下文中显著变化的领域中的适应性。

为进一步解决数据匮乏的问题，Meta-CETM (Xu et al., 2024) 适应预训练的上下文嵌入，以改善低资源领域的主题建模。但是，它并没有

明确对齐源域和目标域之间的主题分布，而是依靠通过嵌入优化进行的隐式适应。这种方法假设目标域环境提供了足够的信息来进行有效的适应，但在术语稀少或高度专业化的领域，模型可能会过拟合有限的上下文信号，导致主题表示不稳定和泛化性差。另一方面，FASTopic (Wu et al., 2024) 采用完全预训练的基于变换器的主题模型，避免了对目标域数据进行微调的需要。虽然这种方法提高了效率并避免了对小型目标数据集的过拟合，但它假设源域知识普遍适用。

3 提出的方法

在本节中，我们正式定义了用于低资源主题建模的领域适应问题，并为我们的方法建立了理论基础。我们首先推导出泛化界限，以量化从高资源源领域向低资源目标领域有效知识转移的条件。利用这些见解，我们引入了 DALTA (领域对齐潜在主题适应)，其在保持领域特定结构的同时对齐潜在表示。

Problem 1 (Domain Adaptation for Low-resource Topic Modeling). 让 $\mathcal{X} \subseteq \mathbb{R}^d$ 表示一个 d 维的文档数据空间，其中 $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$ 是文档表示的集合（例如，词袋和/或嵌入）具有边际分布 $p(\mathcal{X})$ 。源域定义为 $(\mathcal{X}_S, p(\mathcal{X}_S))$ ，目标域定义为 $(\mathcal{X}_T, p(\mathcal{X}_T))$ ，其中 $\mathcal{X}_S \neq \mathcal{X}_T$ （例如，不同的词汇或结构）和 $p(\mathcal{X}_S) \neq p(\mathcal{X}_T)$ （例如，分布转移）。主题空间 α_S 和 α_T 分别表示源域和目标域中文档的特定于域的潜在主题比例，而主题词分布 β_S 和 β_T 捕捉每个域中主题和词之间的关系。领域适应通过利用来自源域 $\mathcal{X}_S$ 的知识来推断低资源目标域 $\mathcal{X}_T$ 的有意义的主题，同时解决域之间的词汇不匹配、主题变化和分布差异等挑战。

3.1 泛化界限

为了更好地理解低资源主题建模中领域适应的挑战，考虑我们想要分析来自两个不同领域的文档的情形：计算机科学（源领域）和医学（目标领域）。虽然源领域提供了丰富的文档，但目标领域却面临数据稀缺的问题。我们的目标是开发一个在目标领域中能够很好地泛化的主题模型，即使数据的可用性有限。这一情形引发了基本的问题：我们能否有效利用丰富的源领域数据来改善目标领域的主题建模，以及我们如何在理论上保证这种迁移的有效性？

为了解决这些问题，我们推导了一个关于主题建模中领域适应的有限样本泛化界限，提供了对影响领域间知识转移因素的见解。我们的理论框架基于泛化误差的概念，衡量训练模型在有限样本上对未见数据的表现能力。具体来

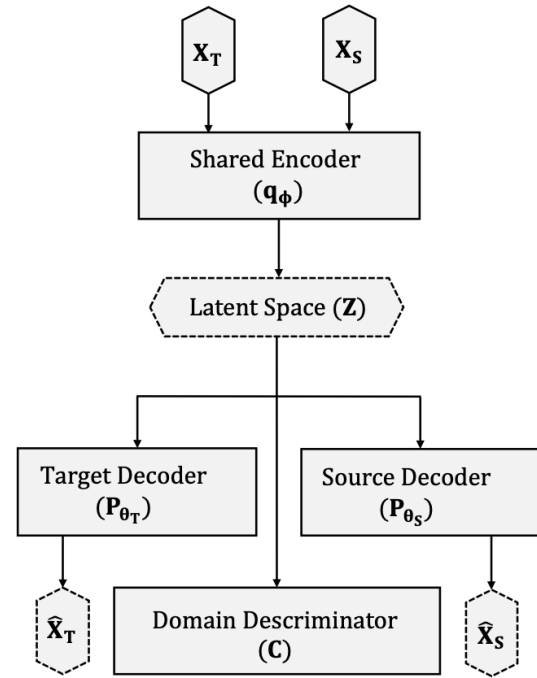


Figure 2: DALTA 框架

说，我们旨在用可观察的源领域和目标领域的量来界定目标领域误差 $\epsilon_T(h)$ 的界限。

Theorem 1 (Generalization bound). 令 $h \in \mathcal{H}$ 是来自假设类 $\mathcal{H}$ 的一个假设，其中 $h: \mathcal{Z} \rightarrow [0, 1]^V$ 从一个潜在语义空间 $\mathcal{Z}$ 映射到词汇表 V 上的概率分布。令 f_S 和 f_T 分别为将潜在表示映射到源域和目标域重构输出的最优函数。定义 $p_S = \frac{n_S}{n_S + n_T}$ 为源样本的比例， $p_T = \frac{n_T}{n_S + n_T}$ 为目标样本的比例。那么，对于每个 $h \in \mathcal{H}$ 和任何 $\delta > 0$ ，源样本和目标样本大小分别为 n_S 和 n_T 的选择概率至少为 $1 - \delta$ 时，目标域误差受以下约束：

$$\epsilon_T(h) \leq \frac{1}{\lambda} \left(\text{KL}(q||p) + p_S \cdot (d_{\mathcal{H}}(\mathcal{D}_S(\mathcal{Z}), \mathcal{D}_T(\mathcal{Z})) + \frac{1}{\lambda} \right)$$

定理 1 的泛化界限强调了主题建模中影响领域适应的关键因素。前两项捕捉了源域和目标域的经验重构误差，强调了需要潜在主题表示来有效地重构跨这两个领域的文档。第三项，KL 散度 $\text{KL}(q||p)$ ，通过约束学习到的潜在表示接近先验来正则化模型，从而减少过拟合。第四项衡量源域和目标域潜在表示之间的差异，较低的值表明更好的领域对齐和改进的学习主题的可迁移性。第五项量化了最优重构函数之间的差异，反映了主题词分布在各领域之间的对齐程度。最后，最后一项涉及高阶复杂性度量，解释了模型的容量以及由于有限样本量引起的统计波动。

3.2 提出的模型：DALTA

受定理 1 中推广误差界的启发，我们引入了 DALTA (Domain-Aligned Latent Topic Adaptation)，一种界限最小化框架（如图 2 所示），旨

在改善低资源主题建模中的领域适应。DALTA 背后的关键思想是专注于优化那些直接影响目标领域模型性能的界限组成部分。由于界限的最后一项反映了模型复杂性和样本方差——这些因素较难掌控，我们优先考虑最小化前五项，这五项涵盖了重建质量、表示对齐和正则化。

为了实现这一目标，DALTA 使用源数据和目标数据最小化经验重建误差，以促进准确的文档重建。它鼓励源和目标潜在空间之间的对齐，以减少领域差异，并促进重建函数的一致性，以确保跨领域的类似性能。此外，KL 散度项被用作正则化器，以保持稳定的潜在空间结构，支持泛化同时减轻过拟合。这一有针对性的优化策略使得 DALTA 能够有效地适应不同领域的主题模型。

为了学习领域不变的信息，我们使用一个共享编码器 $q_\phi: \mathcal{X} \rightarrow \mathcal{Z}$ ，将来自源域和目标域的文档 X 映射到一个共同的潜在空间 Z 。为了实现领域不变的表示，编码器旨在最小化源域和目标域潜在分布之间的差异。我们采用一种对抗训练方法 (Ganin et al., 2016)，引入一个领域判别器 C ，以区分源域和目标域的表示，而编码器 q_ϕ 则被优化以欺骗 C 。对抗目标被表述为：

$$\min_{q_\phi} \max_C \mathcal{L}_{adv} = \mathbb{E}_{\mathcal{X}_S} [\log C(q_\phi(X_S))] + \mathbb{E}_{\mathcal{X}_T} [\log(1 - C(q_\phi(X_T)))] \quad (1)$$

这种极小化-极大化优化通过降低判别器区分两个域的能力，确保编码器学习域不变特征。为了正式描述这一关系，我们将域分类器的性能与源域和目标域之间的差异的联系表达如下：

Proposition 1. 设 $q_\phi: \mathcal{X} \rightarrow \mathcal{Z}$ 为一个共享的编码器，将来自源域 $\mathcal{X}_S$ 和目标域 $\mathcal{X}_T$ 的文档映射到一个共同的潜在空间 $\mathcal{Z}$ 。源和目标潜在分布之间的 $\mathcal{H}$ -散度由以下公式给出：

$$d_{\mathcal{H}}(q_\phi(\mathcal{X}_S), q_\phi(\mathcal{X}_T)) = 2(1 - 2\epsilon_C^*),$$

，其中 ϵ_C^* 是最优域分类器 C^* 的分类误差。

作为 $\epsilon_C \rightarrow 0.5$ （由 C^* 进行的随机猜测）， $d_{\mathcal{H}}(q_\phi(\mathcal{X}_S), q_\phi(\mathcal{X}_T)) \rightarrow 0$ ，表示 $q_\phi(\mathcal{X}_S)$ 和 $q_\phi(\mathcal{X}_T)$ 之间的完美对齐。这个过程直接有助于最小化泛化界中的第四项。

为了捕捉领域特定的特征，DALTA 融入了解码器 $p_{\theta_S}(X_S|Z)$ 和 $p_{\theta_T}(X_T|Z)$ 。每个解码器通过映射到主题比例来推断文档主题分布—— α_S 和 α_T 分别针对源领域和目标领域——然后用于重构文档。这个中间步骤确保了潜在表

示能够捕捉与领域相关的语义。主题数量在不同领域可以变化，并且与潜在空间的大小无关，这提供了处理不同主题细粒度的灵活性。重构目标定义为：

$$\min_{q_\phi, p_{\theta_S}, p_{\theta_T}} \mathcal{L}_{rec} = -\mathbb{E}_{q_\phi(Z|X_S)} [\log p_{\theta_S}(X_S|Z)] - \mathbb{E}_{q_\phi(Z|X_T)} [\log p_{\theta_T}(X_T|Z)], \quad (2)$$

，对应于在泛化界限中最小化前两个项。

为了增强在各个领域间学习表示的一致性，DALTA 引入了一种一致性损失，用于减少源解码器和目标解码器的最优重构函数之间的差异。这是通过在处理对齐的潜在表示时，强制域特定解码器的输出之间的相似性来实现的： $\mathbb{E}_{q_\phi(Z|X_S)} \|p_{\theta_S}(X_S|Z) - p_{\theta_T}(X_T|Z)\|^2$

，这直接通过促进源域和目标域之间的对齐，来贡献于最小化泛化界中的第五项。 (3)

为了防止过拟合并保持平滑的潜在空间结构，DALTA 采用了一个 KL 散度正则项：

$$\min_{q_\phi} \mathcal{L}_{KL} = D_{KL}(q_\phi(Z|X) \| p(Z)), \quad (4)$$

其中 $p(Z)$ 通常被选择为标准高斯先验。这种正则化有助于控制模型的复杂性，解决泛化界的第三项，确保鲁棒的泛化。

将这些组件结合在一起，DALTA 的整体目标被表述为：

$$\min_{q_\phi, p_{\theta_S}, p_{\theta_T}} \max_C \mathcal{L}_{DALTA} = \mathcal{L}_{rec} + \omega_{adv} \mathcal{L}_{adv} + \omega_{cons} \mathcal{L}_{cons} + \omega_{KL} \mathcal{L}_{KL}, \quad (5)$$

，其中 ω_{adv} 、 ω_{cons} 和 ω_{KL} 是平衡对抗性、一致性和正则化损失贡献的超参数。这个综合框架使 DALTA 能够有效地平衡域不变和域特定的学习，促进稳健和适应性强的跨域主题建模。DALTA 的详细训练过程在算法 1 中概述。

Algorithm 1 学习 DALTA

Require: Source domain data $\mathcal{X}_S$, target domain data $\mathcal{X}_T$, learning rates, domain-weight parameter μ , trade-off parameters $\omega_{adv}, \omega_{cons}, \omega_{KL}$

- 1: Initialize encoder parameters ϕ , decoder parameters θ_S, θ_T , and domain discriminator C
- 2: **while** not converged **do**
- 3: Sample mini-batch from $\mathcal{X}_S$ and $\mathcal{X}_T$
- 4: Encode documents: $Z_S = q_\phi(X_S), Z_T = q_\phi(X_T)$
- 5: Compute losses: $\mathcal{L}_{rec}$ (2), $\mathcal{L}_{adv}$ (1), $\mathcal{L}_{cons}$ (3) and $\mathcal{L}_{KL}$ (4)
- 6: Compute total loss: $\mathcal{L}_{DALTA}$ (5)
- 7: Update ϕ, θ_S, θ_T to minimize $\mathcal{L}_{DALTA}$
- 8: Update discriminator C to maximize $\mathcal{L}_{adv}$
- 9: **end while** **return** Optimized parameters $\phi, \theta_S, \theta_T, C$

4 实验

在本节中，我们在资源匮乏的环境下进行了一套全面的实验，评估主题质量、分类准确性、文档聚类性能以及个别损失组件的影响。我们还包括了主题可解释性的定性分析和一个关于源领域选择的案例研究，附录中提供了扩展结果。

4.1 实验设置

我们评估四个不同目标数据集在低资源主题建模中的跨领域适应性，每个数据集每个领域随机抽取 1000 个实例：(1) 20 Newsgroups¹，我们使用科学和宗教来评估技术性和基于信仰的主题之间的适应性；(2) 药物评论²，由患者对名为快诺酮和炔雌醇的两种药物的评论组成，用于评估医学文本中的适应性；(3) Yelp 评论³，代表了非正式、情感丰富的商业评论；以及 (4) SMS 垃圾邮件集合⁴，包含标记为垃圾邮件和非垃圾邮件的短信，用于测试高词汇变异性的短文本数据中的适应性。

对于源数据集，我们使用 AG News 语料库⁵，这是一个大规模新闻数据集，涵盖世界、体育、商业和科学/技术主题。作为一个高资源领

¹<https://www.kaggle.com/datasets/crawford/20-newsgroups>

²<https://www.kaggle.com/datasets/jessicali9530/kuc-hackathon-winter-2018>

³<https://www.kaggle.com/datasets/omkarsabnis/yelp-reviews-dataset>

⁴<https://archive.ics.uci.edu/dataset/228/sms+spam+collection>

⁵<https://www.kaggle.com/datasets/amananandrai/ag-news-classification-dataset>

域，它提供了广泛的主题覆盖，允许模型学习可迁移的主题表示，以适应低资源目标领域。

我们将我们的模型与几个已建立的基线进行比较：(1) LDA (Blei et al., 2003) 将文档建模为主题的混合，每个主题由单词的分布表示。(2) ProdLDA (Srivastava and Sutton, 2017) 使用变分自编码器推断文档主题分布。(3) ETM (Dieng et al., 2020) 融入词嵌入以增强主题一致性。(4) CTM (Bianchi et al., 2020a) 将上下文化的文档嵌入与词袋表示相结合。(5) ECRTM (Wu et al., 2023) 为每个主题强制执行不同的词嵌入聚类以防止主题崩溃。(6) DeTiME (Xu et al., 2023) 利用基于编码器-解码器的大型语言模型生成语义一致的主题嵌入。(7) Meta-CETM (Xu et al., 2024) 在低资源环境中使用目标领域上下文调整词嵌入。(8) FASTopic (Wu et al., 2024) 通过双重语义关系重构范式建模文档、主题和单词之间的语义关系。

有关实现细节和计算基础设施，请参阅附录 B 和 ??。

4.2 主题质量评估

评估指标。为了评估每个模型返回的话题的质量，我们使用以下两个不同的指标——(1) C_V (Wu et al., 2020)：我们使用一种广泛使用的主题建模的一致性评分，名为 C_V 。这是衡量话题可解释性的标准指标。(2) TD (Nan et al., 2019)：主题多样性 (TD)，定义为所有话题中前 10 个词中独特词汇的百分比。

结果和讨论。表 1 展示了多个数据集和主题数量下各种模型的主题连贯性 (C_V) 和主题多样性 (TD) 得分。较高的连贯性得分表明主题内部的语义一致性更好，而较高的多样性得分则反映了在主题间更广泛的独特词汇覆盖率。我们提出的 DALTA 模型在几乎所有设置中始终实现最高的连贯性和多样性得分，这表明其能够生成语义上有意义且多样化的主题。值得注意的是，DALTA 在每个数据集和设置中都在连贯性方面优于所有基线，尤其是在药物评论和垃圾邮件集合等特定领域具有显著的提升。这表明，DALTA 有效地平衡了领域无关的知识转移，同时保留了领域特定的主题结构。

在比较其他模型时，ETM 和 CTM 通常通过利用词向量和上下文化表示来改善连贯性而优于 LDA 和 ProdLDA。然而，这通常以主题多样性的代价换取，导致较少全面的主题覆盖。Meta-CETM 和 FASTopic 在像 Yelp 这样的更一般的数据集中表现良好，但在小众环境中则表现困难，而 DALTA 被证明更加稳定和健壮。一个有趣的趋势是，增加主题的数量（从

Models	Newsgroup								Drug Review								Yelp				SMS Spam Collection			
	Science				Religion				Norethindrone				Norgestimate				k=10		k=20		k=10		k=20	
	k=10		k=20		k=10		k=20		k=10		k=20		k=10		k=20		C _V		TD		C _V		TD	
LDA	0.425	0.696	0.429	0.564	0.424	0.588	0.381	0.508	0.439	0.420	0.444	0.372	0.461	0.472	0.457	0.318	0.394	0.420	0.398	0.358	0.351	0.680	0.391	0.662
ProdLDA	0.410	0.816	0.417	0.834	0.422	0.900	0.390	0.878	0.437	0.796	0.473	0.616	0.472	0.720	0.403	0.662	0.437	0.772	0.453	0.834	0.405	0.828	0.421	0.708
ETM	0.469	0.808	0.408	0.498	0.422	0.784	0.406	0.560	0.439	0.516	0.426	0.304	0.445	0.492	0.450	0.314	0.359	0.688	0.412	0.518	0.434	0.632	0.407	0.336
CTM	0.476	0.804	0.431	0.832	0.407	0.852	0.422	0.830	0.466	0.792	0.470	0.694	0.422	0.724	0.480	0.594	0.398	0.768	0.441	0.746	0.471	0.848	0.476	0.732
ECRTM	0.391	0.636	0.427	0.556	0.410	0.628	0.420	0.524	0.459	0.632	0.410	0.860	0.411	0.596	0.457	0.798	0.392	0.728	0.473	0.472	0.499	0.829	0.493	0.821
DeTime	0.417	0.808	0.411	0.844	0.402	0.900	0.396	0.874	0.355	0.672	0.341	0.652	0.380	0.714	0.345	0.648	0.371	0.716	0.374	0.784	0.378	0.628	0.382	0.562
Meta-CETM	0.396	0.831	0.391	0.891	0.409	0.873	0.403	0.899	0.493	0.845	0.530	0.748	0.426	0.679	0.417	0.872	0.406	0.791	0.437	0.761	0.452	0.879	0.423	0.792
Fastopic	0.406	0.829	0.424	0.905	0.389	0.881	0.418	0.900	0.517	0.811	0.490	0.948	0.413	0.709	0.414	0.900	0.440	0.811	0.454	0.778	0.464	0.814	0.485	0.692
DALTA	0.493	0.836	0.451	0.924	0.431	0.908	0.451	0.918	0.582	0.892	0.571	0.800	0.484	0.732	0.483	0.932	0.448	0.852	0.516	0.808	0.503	0.900	0.505	0.800

Table 1: 主题词的一致性 (C_V) 和多样性 (TD) 分数。 k 表示主题的数量。 每种情况下的最佳结果以粗体显示。

$k = 10$ 到 $k = 20$) 往往能够提高多样性, 但并不总是能够增强连贯性。例如, 在像 Newsgroup 和 Yelp 这样的数据集中, 提高主题数量并不一定会导致更连贯的主题。尽管如此, DALTA 在连贯性和多样性之间保持最佳平衡, 使其特别适合资源有限的主题建模场景。

4.3 文本分类评估

虽然主题模型并非主要为文本分类而设计, 但它们生成的文档主题分布可以作为分类任务中的有用特征。为了评估这些表示是否能够很好地捕捉有意义的文档特征, 我们将它们用作支持向量分类 (SVC) (Cortes and Vapnik, 1995) 和逻辑回归 (LR) (Wright, 1995) 的输入特征。我们使用 5 折交叉验证来评估分类性能, 从而确保基于不同主题模型生成的信息丰富且具有辨别力的文档表示进行稳健的比较。

结果与讨论。表格 2 显示了文本分类的表现。与主题质量相似, 我们提出的 DALTA 模型在大多数情况下实现了最高的分类准确性, 特别是在利基数据集上, 精确的主题分离至关重要。在垃圾邮件合集 (Spam Collection) 中, DALTA 达到了 0.975 (SVC) 和 0.978 (LR), 优于 Meta-CETM 和 FASTopic。同样地, 在药品评论 (Drug Review) 中, DALTA 在两种药物类别, Norethindrone 和 Norgestimate 上都实现了最佳分类准确性, 表明它能够有效捕捉领域特定的词汇以改进分类。

在基线模型中, FASTopic 和 Meta-CETM 在一般数据集上表现良好, 但在小众领域表现不一致, 这可能是因为它们依赖于基于嵌入的适应而非直接的领域对齐。CTM 和 ECRTM 通过增加主题数量而受益, 特别是在新闻组 (例如, 科学、宗教) 中, 其中更细的主题粒度提高了分类性能。LDA 和 ProdLDA 在广泛的领域中表现出竞争力的准确性, 但在小众环境中遇到困难, 这里需要更专业的主题表示。

4.4 聚类性能评估

在我们的分类研究基础上, 我们现在评估 DALTA 的文档-主题分布是否在不使用任何标签的情况下自然形成连贯的聚类。我们使用两种聚类指标进行评估: 纯度 (Purity) 和标准化互信息 (NMI), 参考了赵等人的方法。纯度衡量的是在每个推断的主题聚类中被正确分配的文档比例, 而 NMI 量化推断主题分配与真实主题分配之间的相互依赖性, 从而提供主题连贯性和分离性的见解。这两种指标的较高值表明发现的主题结构与真实标签之间的更好对齐。

与分类结果相似, DALTA 在聚类中始终优于基线模型, 在大多数数据集上实现了最高的纯度和 NMI 分数。值得注意的是, DALTA 在 SMS 垃圾短信集合 (纯度 0.978, NMI 0.287) 和药物评论 (Norgestimate: 纯度 0.604, NMI 0.071) 上表现极佳, 突显了其在低资源和专业领域中增强集群质量的能力。在基线模型中, CTM 和 Meta-CETM 在像 Newsgroup 和 Yelp 这样的结构化数据集上显示了竞争力, 其中 CTM 受益于更高的主题颗粒度。FASTopic 在 Yelp 上表现良好, 利用基于嵌入的适应进行聚类。然而, ETM 和 DeTime 的 NMI 分数较低, 表明在形成分离良好的主题集群上存在困难。这些结果证实, DALTA 的领域感知建模通过学习更连贯且可迁移的主题表示来提高分类和聚类效果。

4.5 消融实验

表 3 使用 Newsgroup Science 数据集评估不同损失项对主题质量和分类性能的影响。虽然此分析仅限于单一数据集, 但它提供了有关每个损失函数如何促进跨域主题适应的重要见解。完整的 DALTA 模型, 结合所有损失项, 实现了连贯性、多样性和分类准确性之间的最佳平衡, 证实每个组件在优化主题建模性能中起着至关重要的作用。

去除 $\mathcal{L}_{adv}$ 会降低分类准确性, 强调了其在领域对齐中的作用。排除 $\mathcal{L}_{consist}$ 则会降低主

Models	Newsgroup								Drug Review								Yelp				SMS Spam Collection			
	Science				Religion				Norethindrone				Norgestimate											
	k=10		k=20		k=10		k=20		k=10		k=20		k=10		k=20		k=10		k=20		k=10		k=20	
	SVC	LR	SVC	LR	SVC	LR	SVC	LR	SVC	LR	SVC	LR	SVC	LR	SVC	LR	SVC	LR	SVC	LR	SVC	LR	SVC	LR
LDA	0.575	0.588	0.540	0.547	0.505	0.513	0.529	0.522	0.564	0.564	0.562	0.578	0.571	0.557	0.575	0.600	0.686	0.686	0.686	0.686	0.864	0.872	0.890	0.897
ProdLDA	0.569	0.635	0.650	0.716	0.496	0.549	0.473	0.505	0.587	0.569	0.570	0.540	0.599	0.620	0.615	0.628	0.667	0.656	0.686	0.686	0.837	0.960	0.864	0.864
ETM	0.297	0.272	0.246	0.263	0.404	0.394	0.404	0.378	0.451	0.452	0.433	0.442	0.493	0.496	0.495	0.478	0.686	0.686	0.686	0.686	0.864	0.864	0.864	0.864
CTM	0.674	0.651	0.665	0.702	0.544	0.557	0.498	0.516	0.566	0.552	0.586	0.576	0.562	0.563	0.627	0.628	0.686	0.686	0.686	0.686	0.886	0.987	0.864	0.864
ECRTM	0.534	0.566	0.591	0.625	0.521	0.541	0.516	0.507	0.584	0.596	0.592	0.552	0.542	0.542	0.630	0.613	0.686	0.685	0.686	0.686	0.881	0.883	0.870	0.891
DeTime	0.254	0.254	0.254	0.254	0.411	0.411	0.410	0.410	0.464	0.464	0.464	0.464	0.503	0.503	0.503	0.503	0.686	0.686	0.686	0.686	0.864	0.864	0.864	0.864
Meta-CETM	0.681	0.729	0.641	0.682	0.506	0.492	0.489	0.515	0.558	0.519	0.543	0.561	0.601	0.591	0.635	0.636	0.649	0.686	0.686	0.686	0.871	0.872	0.853	0.881
Fastopic	0.678	0.702	0.667	0.860	0.510	0.522	0.519	0.530	0.564	0.582	0.572	0.584	0.616	0.601	0.625	0.636	0.685	0.686	0.717	0.717	0.869	0.870	0.869	0.868
DALTA	0.698	0.758	0.685	0.707	0.529	0.549	0.528	0.549	0.598	0.600	0.598	0.600	0.646	0.641	0.646	0.641	0.686	0.686	0.686	0.684	0.975	0.978	0.975	0.978

Table 2: 文本分类准确率经过 5 折交叉验证。每种情况下的最佳结果以粗体显示。

					NG-Science							
$\mathcal{L}_{rect}$	$\mathcal{L}_{recs}$	$\mathcal{L}_{adv}$	$\mathcal{L}_{cons}$	$\mathcal{L}_{KL}$	Topic Quality				Classification			
					k=10		k=20		k=10		k=20	
					C_V	TD	C_V	TD	SVC	LR	SVC	LR
✓	✗	✗	✗	✗	0.458	0.848	0.391	0.868	0.614	0.704	0.671	0.684
✓	✓	✗	✗	✗	0.365	0.808	0.305	0.824	0.658	0.725	0.685	0.707
✓	✗	✓	✗	✗	0.427	0.860	0.394	0.864	0.673	0.693	0.656	0.702
✓	✗	✗	✓	✗	0.350	0.844	0.340	0.844	0.676	0.709	0.635	0.713
✓	✗	✗	✗	✓	0.385	0.856	0.393	0.864	0.683	0.705	0.663	0.675
✓	✗	✗	✓	✓	0.354	0.844	0.406	0.836	0.671	0.723	0.680	0.689
✓	✗	✓	✓	✓	0.383	0.868	0.373	0.836	0.669	0.718	0.649	0.667
✓	✓	✓	✓	✓	0.493	0.836	0.451	0.924	0.698	0.758	0.685	0.707

Table 3: 消融研究。“✓”表示我们使用对应的损失项，“✗”则表示不使用。

题连贯性，这表明一致性约束有助于保持稳定且有意义的主题结构。当省略 $\mathcal{L}_{KL}$ 时，连贯性和分类准确性出现最显著的下降，这表明潜在空间正则化可以防止主题崩溃。有趣的是，没有 $\mathcal{L}_{KL}$ 时主题多样性略有增加，这突显了结构稳定性和多样性之间的权衡。

4.6 定性主题可解释性

为了进一步评估主题一致性和语义清晰度，我们使用 Yelp Reviews 数据集对模型进行定性比较。在此分析中，我们手动检查每个模型生成的两个代表性主题，并使用前五个主题词对其进行总结。此评估不仅使我们能够评估主题的语义独特性，还能评估其实际可解释性。

如表 4 所示，DALTA 产生的主题比其他方法显著更加连贯和主题一致。传统模型如 LDA 和 ProdLDA 往往会产生模糊或通用的主题（例如，好，食物，来），而更先进的模型如 CTM、Meta-CETM 和 Fastopic 则出现主题混淆或语义漂移的迹象。相比之下，DALTA 生成的主题更加明确且更易于解释。例如，它识别出一个专注于食物的主题（塔可，小麦饼，萨尔萨酱，墨西哥，午餐）和一个与社交环境相关的独特上下文主题（厨房，聚会，房间，夜晚，饮料）。这些发现强调了 DALTA 在生成简洁和集中的主题方面的优势，这些主题更好地反映了评论内容结构，同时在低资源环境中结合了其强大的定量表现和实际可解释性。

在本文中，我们通过引入一个理论上的泛化

界限来量化有效知识转移的条件，从而解决了低资源主题建模中的跨领域适应挑战。基于这一见解，我们提出了 DALTA，这是一种将领域不变与领域特定组件分开的模型，以改善主题适应。在给定高资源源语料库和低资源目标语料库的情况下，DALTA 学会对齐主题，同时保留领域特定的信息。实证结果表明，DALTA 始终优于最先进的方法，突显了我们框架在低资源环境中的有效性。

5

局限性 尽管 DALTA 为跨域适应提供了强大的理论基础，但它并未提供一种实用的方法来选择最合适的源域。然而，我们发现其内部训练信号可以被重新用于这个目的。在附录 C 中，我们展示了案例研究，证明了一种简单的对齐分数可以有效识别最佳源域，而无需完全模型收敛。虽然我们的理论界限有助于确定何时知识转移是有益的，但我们并没有实证研究不同的源域如何影响适应性能。这引发了一个未解的问题，即如何系统地给定的低资源目标域识别最佳的源域。此外，DALTA 的有效性取决于域之间主题结构对齐的程度，而这在事先可能并不总是已知的。如果源域和目标域在主题分布上存在显著差异，适应可能导致不对齐或模型性能下降。解决这些挑战需要进一步研究自动化的源域选择策略，以优化跨不同真实世界环境的适应。

Model	Topic 1	Topic 2
LDA	good, place, great, food, come	like, time, place, come, want
ProdLDA	good, food, like, menu, order	schnitzel, alligator, shell, fry, foodie
ETM	burger, pizza, sandwich, salad, meal	like, know, come, say, time
CTM	chicken, soup, beet, salad, mandarin	good, order, come, say, like
ECRTM	order, food, table, burger, salad	taco, pho, boba, chop, gelato
DeTime	flavor, waffle, soup, decor, slaw	great, love, order, time, try
Meta-CETM	burger, fries, shake, cheese, ketchup	spa, massage, facial, candle, relaxing
Fastopic	burger, pork, rice, sushi, sandwich	server, minutes, wait, thanks, manager
DALTA	taco, burrito, salsa, mexican, lunch	kitchen, party, room, night, drink

Table 4: 在 Yelp 评论数据集上由每个模型生成的具有代表性的 5 个词的主题。

6

致谢 这些材料基于以下支持的工作：国家自然科学基金会 IIS 16-19302 和 IIS 16-33755 的支持，浙江大学 ZJU 研究 083650，IBM-伊利诺伊认知计算系统研究中心（C3SR）和 IBM-伊利诺伊探索加速器研究所（IIDAI），来自 eBay 和 Microsoft Azure 的资助，UIUC OVCR CCIL 规划资助 434S34，UIUC CSBS 小额资助 434C8U，以及 UIUC 新前沿计划。本文中表达的任何观点、发现、结论或建议均为作者的个人意见，并不一定反映资助机构的观点。

References

- Pritom Saha Akash and Kevin Chang. 2024. Enhancing short-text topic modeling with llm-driven context expansion and prefix-tuned vaes. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15635–15646.
- Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. 2010. A theory of learning from different domains. *Machine learning*, 79:151–175.
- Federico Bianchi, Silvia Terragni, and Dirk Hovy. 2020a. Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. *arXiv preprint arXiv:2004.03974*.
- Federico Bianchi, Silvia Terragni, Dirk Hovy, Debora Nozza, and Elisabetta Fersini. 2020b. Cross-lingual contextualized topic models with zero-shot learning. *arXiv preprint arXiv:2004.07737*.
- David Blei and John Lafferty. 2006a. Correlated topic models. *Advances in neural information processing systems*, 18:147.
- David M Blei and John D Lafferty. 2006b. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120.
- David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.
- Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.
- Adji B Dieng, Francisco JR Ruiz, and David M Blei. 2020. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453.
- Zhibin Duan, Dongsheng Wang, Bo Chen, Chaojie Wang, Wenchao Chen, Yewen Li, Jie Ren, and Mingyuan Zhou. 2021. Sawtooth factorial topic embeddings guided gamma belief network. In *International Conference on Machine Learning*, pages 2903–2913. PMLR.
- Zhibin Duan, Yishi Xu, Jianqiao Sun, Bo Chen, Wenchao Chen, Chaojie Wang, and Mingyuan Zhou. 2022. Bayesian deep embedding topic meta-learner. In *International Conference on Machine Learning*, pages 5659–5670. PMLR.
- Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35.
- Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.
- Sungwon Han, Mingi Shin, Sungkyu Park, Changwook Jung, and Meeyoung Cha. 2023. Unified neural topic model via contrastive learning and term weighting. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1802–1817.
- Tomoharu Iwata. 2021. Few-shot learning for topic modeling. *arXiv preprint arXiv:2104.09011*.
- Diederik P Kingma. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.
- Bo Li, Yezhen Wang, Shanghang Zhang, Dongsheng Li, Kurt Keutzer, Trevor Darrell, and Han Zhao.

2021. Learning invariant representations and risks for semi-supervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1104–1113.
- Sokhna Diarra Mbacke, Florence Clerc, and Pascal Germain. 2024. Statistical guarantees for variational autoencoders using pac-bayesian theory. *Advances in Neural Information Processing Systems*, 36.
- Jon Mcauliffe and David Blei. 2007. Supervised topic models. *Advances in neural information processing systems*, 20.
- Yishu Miao, Lei Yu, and Phil Blunsom. 2016. Neural variational inference for text processing. In *International conference on machine learning*, pages 1727–1736. PMLR.
- Feng Nan, Ran Ding, Ramesh Nallapati, and Bing Xiang. 2019. Topic modeling with wasserstein autoencoders. *arXiv preprint arXiv:1907.12374*.
- Thong Nguyen and Anh Tuan Luu. 2021. Contrastive learning for neural topic model. *Advances in neural information processing systems*, 34:11974–11986.
- Chau Minh Pham, Alexander Hoyle, Simeng Sun, Philip Resnik, and Mohit Iyyer. 2023. Topicgpt: A prompt-based topic modeling framework. *arXiv preprint arXiv:2311.01449*.
- Duy-Tung Pham, Thien Trang Nguyen Vu, Tung Nguyen, Linh Ngo, Duc Nguyen, and Thien Nguyen. 2024. Neuromax: Enhancing neural topic modeling via maximizing mutual information and group topic regularization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7758–7772.
- Suzanna Sia and Kevin Duh. 2021. Adaptive mixed component lda for low resource topic modeling. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2451–2469.
- Akash Srivastava and Charles Sutton. 2017. Autoencoding variational inference for topic models. *arXiv preprint arXiv:1703.01488*.
- Raymond E Wright. 1995. Logistic regression.
- Xiaobao Wu, Xinshuai Dong, Thong Thanh Nguyen, and Anh Tuan Luu. 2023. Effective neural topic modeling with embedding clustering regularization. In *International Conference on Machine Learning*, pages 37335–37357. PMLR.
- Xiaobao Wu, Chunping Li, Yan Zhu, and Yishu Miao. 2020. Short text topic modeling with topic distribution quantization and negative sampling decoder. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1772–1782.
- Xiaobao Wu, Thong Thanh Nguyen, Delvin Ce Zhang, William Yang Wang, and Anh Tuan Luu. 2024. Fastopic: Pretrained transformer is a fast, adaptive, stable, and transferable topic model. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- Weijie Xu, Wenxiang Hu, Fanyou Wu, and Srinivasan Sengamedu. 2023. DeTiME: Diffusion-enhanced topic modeling using encoder-decoder based LLM. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9040–9057, Singapore. Association for Computational Linguistics.
- Yishi Xu, Jianqiao Sun, Yudi Su, Xinyang Liu, Zhibin Duan, Bo Chen, and Mingyuan Zhou. 2024. Context-guided embedding adaptation for effective topic modeling in low-resource regimes. *Advances in Neural Information Processing Systems*, 36.
- Han Zhao, Remi Tachet Des Combes, Kun Zhang, and Geoffrey Gordon. 2019. On learning invariant representations for domain adaptation. In *International conference on machine learning*, pages 7523–7532. PMLR.

A 泛化界限的证明

Lemma 1. (*Ben-David et al., 2010*) 令 $\mathcal{X} \subseteq \mathbb{R}^d$ 为实例空间, 令 $\mathcal{D}_S$ 和 $\mathcal{D}_T$ 代表 $\mathcal{X}$ 上的源数据分布和目标数据分布。 f_S 和 f_T 分别是源域和目标域的最优标注函数。对于假设 $h \in \mathcal{H}$, 目标域误差 $\epsilon_T(h)$ 被界定为:

$$\epsilon_T(h) \leq \epsilon_S(h) + d_{\mathcal{H}}(\mathcal{D}_S(X), \mathcal{D}_T(X)) + \min \{ \mathbb{E}_S[|f_S - f_T|], \mathbb{E}_T[|f_S - f_T|] \}.$$

Lemma 2 (Reconstruction Guarantee for Bounded Instance Spaces (*Mbacke et al., 2024*)). 设 $\mathcal{X}$ 为一个直径为 $\Delta < \infty$ 的实例空间, 并设 $\mu \in \mathcal{M}_+^1(\mathcal{X})$ 表示数据生成分布。考虑 $\mathcal{Z}$ 为一个带有先验分布 $p(z) \in \mathcal{M}_+^1(\mathcal{Z})$ 的潜在空间, 并设 θ 表示重构函数的参数。对于任何后验分布 $q_\phi(z|\mathbf{x})$ 、正则化参数 $\lambda > 0$ 和置信水平 $\delta \in (0, 1)$, 以下不等式在概率至少为 $1 - \delta$ 的情况下对一个随机样本 $S \sim \mu^{\otimes n}$ 成立:

$$\epsilon(h) \leq \hat{\epsilon}(h) + \frac{1}{\lambda} KL(q||p) + K_\phi K_\theta \Delta + \frac{1}{\lambda} \log \frac{1}{\delta} + \frac{\lambda \Delta^2}{8n},$$

其中

$\epsilon(h) : \mathbb{E}_{\mathbf{x} \sim \mu} \mathbb{E}_{z \sim q_\phi(z|\mathbf{x})} \ell_{rec}^\theta(z, \mathbf{x})$, 为真实数据分布上的期望重构损失,

$\hat{\epsilon}(h) : \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{z \sim q_\phi(z|\mathbf{x}_i)} \ell_{rec}^\theta(z, \mathbf{x}_i)$, 为可用数据上的经验重构损失,

$KL(q||p) : \sum_{i=1}^n KL(q_\phi(z|\mathbf{x}_i)||p(z))$, 为后验分布与先验分布之间的 KL 散度。

Lemma 3. 设 $h \in \mathcal{H}$ 是来自假设类 $\mathcal{H}$ 的一个假设, 其中 $h: \mathcal{Z} \rightarrow [0, 1]^{|V|}$ 从潜在语义空间 $\mathcal{Z}$ 映射到词汇表 V 上的概率分布。 f_S 和 f_T 分别是将潜在表示映射到源域和目标域的重构输出的最优函数。 然后, 对于每个 $h \in \mathcal{H}$ 和任何 $\delta > 0$, 在源和目标样本的选择上, 大小为 n_S 和 n_T , 具有至少 $1 - \delta$ 的概率, 满足以下不等式: $\epsilon_T(h) \leq \hat{\epsilon}_S(h) + d_{\mathcal{H}}(\mathcal{D}_S(Z), \mathcal{D}_T(Z)) +$

$\frac{1}{\lambda} \min \{ \mathbb{E}_S[|f_S - f_T|], \mathbb{E}_T[|f_S - f_T|] \} + \frac{1}{\lambda} \log \frac{1}{\delta} + \frac{\lambda \Delta^2}{8n_S}$
Proof. 根据引理 1, 对于假设 h , 目标域重建误差有如下界限:

$$\epsilon_T(h) \leq \epsilon_S(h) + d_{\mathcal{H}}(\mathcal{D}_S(Z), \mathcal{D}_T(Z)) + \min \{ \mathbb{E}_S[|f_S - f_T|], \mathbb{E}_T[|f_S - f_T|] \},$$

现在, 应用引理 2 中的上界, 我们得到 $\epsilon_T(h) \leq \epsilon_S(h) + d_{\mathcal{H}}(\mathcal{D}_S(Z), \mathcal{D}_T(Z)) +$

$$\min \{ \mathbb{E}_S[|f_S - f_T|], \mathbb{E}_T[|f_S - f_T|] \} + \frac{1}{\lambda} \text{KL}(q_S \| p) + K_{\phi} K_{\theta_S} \Delta + \frac{1}{\lambda} \log \frac{1}{\delta} + \frac{\lambda \Delta^2}{8n_S}$$

Theorem 1 (Generalization bound). 设 $h \in \mathcal{H}$ 为一个假设类 $\mathcal{H}$ 中的假设, 其中 $h: \mathcal{Z} \rightarrow [0, 1]^{|V|}$ 从潜在语义空间 $\mathcal{Z}$ 映射到词汇 V 上的概率分布。 设 f_S 和 f_T 分别为源域和目标域中将潜在表示映射到重构输出的最优函数。 定义 $p_S = \frac{n_S}{n_S + n_T}$ 为源样本的比例, $p_T = \frac{n_T}{n_S + n_T}$ 为目标样本的比例。 那么, 对于每个 $h \in \mathcal{H}$ 和任何 $\delta > 0$, 以至少为 $1 - \delta$ 的概率, 在大小为 n_S 和 n_T 的源和目标样本的选择下, 目标域误差被限制在以下范围内:

$$\epsilon_T(h) \leq \epsilon_S(h) + d_{\mathcal{H}}(\mathcal{D}_S(Z), \mathcal{D}_T(Z)) + \min \{ \mathbb{E}_S[|f_S - f_T|], \mathbb{E}_T[|f_S - f_T|] \} + \frac{1}{\lambda} \text{KL}(q \| p) + \frac{n_S}{n_S + n_T} \left(d_{\mathcal{H}}(\mathcal{D}_S(Z), \mathcal{D}_T(Z)) + \min \{ \mathbb{E}_S[|f_S - f_T|], \mathbb{E}_T[|f_S - f_T|] \} \right) + \frac{1}{\lambda} \log \frac{1}{\delta} + \frac{\lambda \Delta^2}{8n_S}$$

Proposition 1. 设 $q_{\phi}(\mathcal{X}_S, \mathcal{X}_T)$ 为公共编码器, 将来自源域 $\mathcal{X}_S$ 和目标域 $\mathcal{X}_T$ 的文档映射到一个公共潜在空间 $\mathcal{Z}$ 。 源和目标潜在分布之间的 $\mathcal{H}$ 距离由以下式子给出:

$$d_{\mathcal{H}}(q_{\phi}(\mathcal{X}_S), q_{\phi}(\mathcal{X}_T)) = \frac{2(1 - 2\epsilon_{C^*})}{\lambda \text{KL}(q \| p) + p_S \cdot d_{\mathcal{H}}(\mathcal{D}_S(Z), \mathcal{D}_T(Z))},$$

其中 ϵ_{C^*} 是最佳域外分类器 C^* 的分类误差。 **Proof.** 考虑最佳域外分类器 C^* , 其用于区分源和目标的潜在表示。 最佳决策函数由以下表示:

$$C^*(z) = \frac{p_S(z)}{p_S(z) + p_T(z)},$$

其中 $p_S(z)$ 和 $p_T(z)$ 分别表示源和目标潜在表示的概率密度函数。 C^* 的分类误差可以表示为:

$$\epsilon_{C^*} = \frac{1}{2} \int \min(p_S(z), p_T(z)) dz.$$

使用恒等式 $\min(a, b) = \frac{a+b-|a-b|}{2}$, 我们得到:

$$\int \min(p_S(z), p_T(z)) dz = 1 - \frac{1}{2} \int |p_S(z) - p_T(z)| dz.$$

因此, 分类误差简化为:

$$\epsilon_{C^*} = \frac{1}{2} - \frac{1}{4} \int |p_S(z) - p_T(z)| dz.$$

源和目标潜在分布之间的总变差距离定义为:

$$\text{TV}(q_{\phi}(\mathcal{X}_S), q_{\phi}(\mathcal{X}_T)) = \frac{1}{2} \int |p_S(z) - p_T(z)| dz.$$

代入 ϵ_{C^*} 的公式中, 我们得到:

$$\text{TV}(q_{\phi}(\mathcal{X}_S), q_{\phi}(\mathcal{X}_T)) = 1 - 2\epsilon_{C^*}.$$

$\mathcal{H}$ -散度与总变差距离的关系为:

$$d_{\mathcal{H}}(q_{\phi}(\mathcal{X}_S), q_{\phi}(\mathcal{X}_T)) = 2 \text{TV}(q_{\phi}(\mathcal{X}_S), q_{\phi}(\mathcal{X}_T)).$$

因此, 代入总变差距离:

$$d_{\mathcal{H}}(q_{\phi}(\mathcal{X}_S), q_{\phi}(\mathcal{X}_T)) = 2(1 - 2\epsilon_{C^*}).$$

B 实现细节

我们为提出的架构和基线模型设置了特定参数, 以确保公平比较。 所有基线主题模型的迭代次数固定为 100, 而 DALTA 在 20 个周期内收敛, 以实现第 4 节中报告的性能。 对于上下文文化的文档表示, 我们使用句子转换器 all-MiniLM-L6-v2⁶。 评估指标是使用相同的参数设置计算的; 例如, 用于计算 C_V 和 TD 的每个主题的顶部词数固定为 10。 在文本分类实验中, 我们使用 scikit-learn⁷ 的 SVC 和 LR 的默认参数。

我们提出的模型 DALTA 是基于变分自编码器 (VAE) 的结构, 具有 50 维的潜在空间。 主题数在源域中设置为 50, 而在目标域中则根据实验而有所不同。 为了平衡源域和目标域的贡献, DALTA 采用了一种领域加权采样策略, 通过一个概率参数 μ 进行控制。 初始阶段,

⁶<https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2>

⁷<https://scikit-learn.org>

$\mu = 0.7$ ，优先选择源域样本以实现稳定的表示学习。随着训练的进行， μ 逐渐减少到 0.3，将重点转向目标域以进行适应。这确保了模型在捕捉领域不变结构的同时，通过单独的解码器保留目标特定的信息。

所有实验均在一台服务器上进行，该服务器配备了两块 AMD EPYC 7302 3GHz CPU、三块 NVIDIA Ampere A40 GPU (每块 48GB VRAM, 300W)，以及 256GB RAM。

C 案例研究：通过内部信号选择源域

尽管 DALTA 不是为了执行源域选择而设计的，我们探讨其内部训练信号是否可以用作实用的启发式方法，以指导在真实世界的低资源场景中进行选择。这在存在多个候选源域的情况下尤为相关，但为给定目标域选择最兼容的源域仍然是一个未解决的挑战。

我们定义了一个简单的对齐评分，该评分使用两个在 DALTA 训练早期阶段容易获得的量来定义：领域对齐损失 $\mathcal{L}_{adv}$ ，用于衡量源域和目标域的潜表示对齐得有多好，以及目标重建损失 $\mathcal{L}_{rec}^{(T)}$ ，用于衡量模型在目标领域中重建文档的准确程度。对齐评分定义为：

$$\text{Alignment Score} = \mathcal{L}_{adv} - \lambda \cdot \mathcal{L}_{rec}^{(T)},$$

，其中我们设置 $\lambda = 0.001$ 以平衡两个组件。我们在训练仅进行 5 次迭代后计算这个分数——所以它速度很快并且不需要完全收敛。

为了评估这个启发式方法的有效性，我们使用 Newsgroup 数据集进行了两个案例研究。在第一个案例中，我们将 NG SCIENCE 子集视为目标领域，并考虑四个源领域：(i) 20 Newsgroup 语料库中除去 NG SCIENCE 之外的其余部分，(ii) AG News, (iii) Arxiv-CS 摘要，以及 (iv) 药物评论数据。我们观察到 NG（不包括 SCIENCE）产生了最高的对齐分数，并且在主题一致性和多样性方面也表现最好。这表明，即使是部分重叠的源领域，当其潜在空间与目标很好对齐时，也可以为适应提供有价值的归纳偏差。

在第二种情况下，我们使用 NG RELIGION 作为目标领域并评估相同的来源集。AG News 获得了最高的对齐分数，也产生了最佳的主题质量。虽然 NG（不包括 RELIGION）包含更多的主题重叠，其与目标的潜在对齐似乎较弱，可能是由于词汇变化或语义粒度不匹配所致。相比之下，AG News 包含了通用新闻内容——包括政治和社会——这与宗教话语隐含地重叠，导致更好的适应。

这些案例研究表明，DALTA 的内部训练动态可以用于在学习过程的早期估计源效用。尽

管这种对齐评分是启发式的且任务特定的，但它在低资源主题建模中开发轻量级、数据驱动的源域选择策略提供了一个有希望的起点。

Target	Source	DL	RL _T	Score
NG Science	NG (w/o Science)	0.369	564.68	-0.196
	AG News	0.116	522.34	-0.406
	Arxiv-CS	0.129	557.81	-0.428
	Drug Review	0.095	524.99	-0.430
NG Religion	AG News	0.255	726.58	-0.472
	NG (w/o Religion)	0.289	771.92	-0.483
	Drug Review	0.208	732.50	-0.525
	Arxiv-CS	0.110	765.89	-0.656

Table 5: 使用 DALTA 的早期训练信号在源域之间的对齐分数。更高的分数表明与目标域的对齐和建模适合度更好。