基于深度学习的主题对齐模型研究

余传明[1]，原　赛[2]，胡莎莎[1]，安　璐[3]

（1.中南财经政法大学 信息与安全工程学院,湖北 武汉 430073;
2.中南财经政法大学 统计与数学学院,湖北 武汉 430073;
3.武汉大学 信息管理学院,湖北 武汉 430072）

摘要：在主题深度表示学习的基础上，本文提出了一种融合双语词嵌入的主题对齐模型（Topic Alignment Model, TAM），通过双语词嵌入扩充语义对齐词汇词典，在传统双语主题模型基础上设计辅助分布用于改进不同词分布的语义共享，以此改善跨语言和跨领域情境下的主题对齐效果；提出了两种新的指标，即双语主题相似度（Bilingual Topic Similarity，BTS）和双语对齐相似度（Bilingual Alignment Similarity，BAS），用于评价辅助分布对齐的效果。在跨语言主题对齐任务中，相比于传统的对齐模型，双语对齐相似度提升了约 1.5%；在跨领域主题对齐任务中，F1 值提升了约 10%。TAM 模型在跨语言和跨领域主题对齐上的效果优于传统方法，研究结果对于改进跨语言和跨领域信息处理具有重要意义。

关键词：跨语言主题对齐；跨领域主题对齐；深度学习；双语词嵌入；知识对齐

中图分类号：TP391　　**文献标识码**：A

Research on Topic Alignment Model based on Deep Learning

Yu Chuanming[1]，Yuan Sai[2]，Hu Shasha[1]，An Lu[3]

(1. School of Information and Safety Engineering, Zhongnan University of Economics and Law, Wuhan,430073;
2. School of Statistics and Mathematics, Zhongnan University of Economics and Law, Wuhan,430073;
3. School of Information Management, Wuhan University, Wuhan,430072)

Abstract: Based on the deep representation learning of domain topics, the paper proposed a Topic Alignment Model (TAM) which integrates bilingual word embedding. The semantic alignment lexicon was extended by bilingual word embedding. Based on traditional bilingual topic model, an auxiliary distribution was designed to improve different word distributions semantic sharing, so as to improve the effect of topic alignment in cross-lingual and cross-domain contexts. Two new indicators, i.e. Bilingual Topic Similarity (BTS) and Bilingual Alignment Similarity (BAS), were proposed for supplementary alignment evaluation. In the cross-language topic matching task, the bilingual alignment similarity was improved by about 1.5% compared to the traditional multi-language common cultural theme analysis; in the cross-domain topic alignment task, the F1 value was improved by about 10%. The experimental results show that TAM is better than traditional methods on the tasks of cross-lingual and cross-domain information processing.

Key words: Cross-lingual Topic Alignment; Cross-domain Topic Alignment; Deep Learning; Bilingual Word Embedding; Knowledge Alignment

基金项目：本文系国家自然科学基金面上项目“大数据环境下基于领域知识获取与对齐的观点检索研究”（项目编号 71373286）与教育部哲学社会科学研究重大课题攻关项目“提高反恐怖主义情报信息工作能力对策研究”(项目编号 17JZD034)成果之一。

0 引言

主题模型通常是指以无监督学习的方式，对某个领域的文档集合的潜在语义结构进行自动化分析的数理统计模型[1]。随着大数据技术的兴起，不同语言、不同领域的知识共享与联系日益紧密，来自不同文化的人们可能对同一事物有着相同的主题倾向[2]。例如，同一新闻事件的不同语言评论往往包含相似的兴趣话题[3]，不同语言用户可能存在相同的检索习惯[4]，不同种类的电商评论可能蕴含相同的消费喜好[5]。如何有效的提取和分析不同领域知识的主题信息已成为情报分析、舆情监测、自然语言处理等领域的研究热点[6]。传统的主题模型在大数据环境下，会面临知识的领域跨度和语言跨度问题[7]。领域跨度是由于训练数据和测试数据的总体分布或领域特征不一致，造成不同领域知识所挖掘出来的主题可解释性较差，这是由词汇存在领域差异引起的，如“bank”一词在金融领域中代表“银行”，而在地理领域中代表“河堤”；语言跨度是指由语言多样性造成的同一主题下的词汇在不同语言体系中具有不同的表达方式，如京东和亚马逊中同类商品可以有不同的词汇表示。

为解决传统主题模型在大数据环境下所面临的语言跨度和领域跨度等问题，本文尝试提出一种融合双语词嵌入的领域主题对齐模型，并将其应用到跨语言主题对齐与跨领域主题对齐的实证研究中，通过检验模型的效果，以期为相关领域的主题对齐研究提供借鉴。

1 相关研究

在大数据情境下，主题对齐通常从潜在语义分析出发，旨在发现不同语言、不同领域之间的主题共同特征与关联。具体而言，包括主题的深度表示和主题深度对齐两部分。

1.1 主题的深度表示研究

从技术层面上来看，主题的表示方法可分为统计语义模型和嵌入向量模型。统计语义模型从文档、主题分布、词频入手，计算文档与词汇共生矩阵之间的相似性。潜在语义分析（Latent Semantic Analysis, LSA）和概率潜在语义分析（Probabilistic Latent Semantic Analysis，PLSA）是较早提出的统计语义模型，常用于计算共现词汇文档间的相似性，但它们忽略了词汇之间的顺序关系。潜在狄利克雷分布（Latent Dirichlet Allocation, LDA）通过假定主题分布生成文档、且每个主题是词汇的概率分布，可以捕获词汇背后的语义[8]。尽管 LDA 及其扩展模型考虑了词汇间的语义，但它们忽略了文档中的词汇。为了将语义相关性嵌入到主题模型中，Li[9]等人提出主题向量(TopicVec)模型，通过嵌入链接函数模拟主题中的单词分布以代替 LDA 中的 Dirichlet 分布，其优点在于语义相关性被编码为嵌入空间中的余弦距离，利用变分推理算法将主题嵌入映射到单词的相同空间中。

嵌入向量模型从上下文向量、主题向量、词向量入手，使用邻域信息表示目标词汇的含义。它的主要过程是将主题向量和词向量嵌入到同一语义空间学习分布式表示，根据语义依赖的假定不同，可分为以主题词嵌入模型和 Lda2vec 模型为代表的假定上下文语义依赖单词和主题语义，以及以主题增强词向量模型为代表的假定当前单词语义取决于上下文和主题语义两种。Liu[10]等人提出主题词嵌入模型（Topical Word Embeddings, TWE）将主题集成到词嵌入表示中，并允许主题词嵌入学习不同语境下单词的不同含义，通过计算给定单词和主题词来预测上下文单词嵌入，以最大化给定词和主题下的上下文概率。Zhang[11]等人在 TWE 模型基础上，提出将每个主题视为一个新单词并将主题插入到语料库中，与 TWE 相比，将主题视为新单词的方法考虑了单词与其指定主题之间的内在交互学习，每个单词-主题对共享每个单词和指定主题的参数，减轻了稀疏性问题，并且通过将单词嵌入和指定的主题连接学习可以使同一主题中的单词更具辨别力。Moody[12]提出 Lda2vec 模型用于对初始化文档向量和词向量采用联合训练的方式完成中枢词的上下文向量预测任务，以获得含有主题信息的稀疏文档表示，其训练目标是最小化预测过程中的负采样损失以及文档比例稀疏化过程中 Dirichlet 似然项总和。Li[13]等人提出主题增强词向量模型（Topic Enhanced Word Vectors, TEWV），它基于 CBOW 模型将主题信息和上下文信息集成到词向量中，单词的语义取决于上下文的语义和主题，在 TEWV 模型训练收敛后，可以获得更好单词和主题的低维矢量表示。

1.2 主题的深度对齐研究

主题的深度对齐是将源领域中所学习到的主题特征空间应用到目标领域中，由领域相关知识来度量不同领域间的相似性，大量文献表明，基于领域间高层概念（如特征簇、主题）构成的空间有利于解决跨领域相关任务中知识对齐效果[14]。依照应用层面的不同，可以将领域主题对齐分为跨语言主题对齐与跨领

域主题对齐等。

跨语言主题对齐通常是指将源语言中学习到的主题特征空间对应到目标语言的特征空间中。依据所选用数据的不同，可将其分为词汇、段落、文档、语料库、多模态数据等方面的主题对齐。基于类别相关的双语话题模型 CC-BiLDA 和 CC-BiBTM[15]，通过考虑文本中的同现词语间及关系类别间的相关性在跨语言类别对齐中得到广泛应用。BiSTM 模型利用大多数文档具有段落分层结构这一性质以提升主题对齐模型的精度[16]。文档级别的跨语言主题模型主要有三种，分别为基于双语词典定义的软约束的概率跨语言潜在语义分析[17]、基于潜在狄利克雷分布的双语长文本主题模型[18]以及基于词共现模式的双语短文本主题模型[19]。语料库级别的跨语言主题模型则是对非平行语料库建模，以识别可比语料库或未对齐语料库中的共同主题，如 C-BiLDA[20]。多模态数据的跨语言主题对齐则是从文档、图像等多模态数据中探索模态本身及模态之间的对齐关系，如条件独立的 gRTM（CI-gRTM）模型[21]。

跨领域主题对齐通常是指通过建立潜在共享主题空间，引导源领域与目标领域的主题映射。为减少跨领域文本分类任务中仅使用共享潜在主题作为域桥接的限制，Li[22]等人提出通过推断共享主题和特定域主题间的相关性，可以引导来自不同域的特定域主题之间的映射。Yang[23]等人认为不同领域的文档可以从内容信息和链接结构的角度共享一些共同的主题，以增强相关但不同的分类知识域间的联系，利用辅助链接网络来发现文档之间的直接或间接共引用关系以弥合不同领域之间的差距。当领域间的连接是多角度时，单一共享主题的映射会存在语义表示不完备和偏差性等问题，对此杨奇奇[24]等人提出通过提取多重共享主题和领域独有主题，并以多重共享主题为桥梁来建立领域独有主题之间的多重映射关系。

值得说明的是，目前大多数主题对齐工作局限于单语主题研究范畴，针对多语言情境下领域主题对齐，尚缺乏系统性的研究。鉴于此，本文尝试基于深度学习的跨语言词嵌入技术构建双语语义对齐字典，将不同语言词汇的语义信息和领域信息嵌入到同一语义空间中，同时借助双语主题模型实现主题层次的领域知识对齐。

2 研究问题与研究方法

2.1 研究问题

假定源语言（源领域）和目标语言（目标领域）下的主题集合分别为 TopicA（简称为 A）和 TopicB（简称为 B），主题对齐可通过五元组<id,t1,t2,r,p >（或称为对齐主题）来定义。其中，id 为对齐主题的编号，t1 和 t2 分别代表 A 和 B 中的主题，r 为 t1 和 t2 间的关系（如对等关系、包含关系等），p 为该关系的置信度（通常在 0 到 1 之间）。鉴于本文以“主题对齐模型”为中心议题，因此将实证研究界定在以探究不同语言、不同领域下的主题对等关系为主要研究目标。

2.2 主题对齐模型

为解决跨语言与跨领域情境下的主题对齐问题，本文在主题深度表示学习的基础上提出一种融合双语词嵌入的主题对齐模型（Topic Alignment Model，TAM）。该模型由双语词嵌入模块和主题对齐模块两个部分构成，如图 1 所示。其中，双语词嵌入模块用于将不同语言的词嵌入映射到同一向量空间中，以此丰富双语词汇对齐词典，同时将获得的词汇相似性应用到双语主题对齐中（参见 2.2.1 节）；主题对齐模块则是在前者的基础上对不同语言或领域的主题进行对齐（参见 2.2.2 节）。

模型采用如下的训练过程。首先，使用 Word2vec 工具生成中英文词向量 X 和 Z，对训练语料进行双语互译得到具有翻译关系的训练词典，在双语词嵌入模块中根据训练词典学习单语词嵌入的结构相似性并将其映射到同一空间中，得到映射后的词向量X^*和Z^*，计算X^*和Z^*中词向量的相似度，选择相似度大于预设阈值的前五个中英文词汇构建双语词典D^*。其次，对每个主题$k \in$ K，从超参数为β的$Dirichlet$先验分布中生成辅助分布$\eta^e \sim Dir(\beta \cdot \mathbf{1}_{|\Lambda^e|})$、$\eta^c \sim Dir(\beta \cdot \mathbf{1}_{|\Lambda^c|})$，其中$\mathbf{1}_D$表示分量为 1 的 D 维向量，然后分别生成词分布φ_k^e和φ_k^c。再次，对每篇英文文档$d^e \in \varepsilon$，从超参数为 α 的 $Dirichlet$ 先验分布中生成英文主题向量$\theta_d^e \sim Dir(\alpha \cdot \mathbf{1}_K)$，对于每篇文档$d^e$中每个位置$n$，从多项式分布中抽取词分布和词的主题分布，即$z_n^e \sim Multi(\theta_d^e)$和$w_n^e \sim Multi(\varphi_k^e)$。最后，对每篇中文文档$d^c \in$ C，生成中文主题向量$\theta_d^c \sim Dir(\alpha \cdot \mathbf{1}_K)$，对于每篇文档$d^c$中每个位置$n$，抽取词分布和词的主题分布，即$z_n^c \sim Multi(\theta_d^c)$和$w_n^c \sim Multi(\varphi_k^c)$。

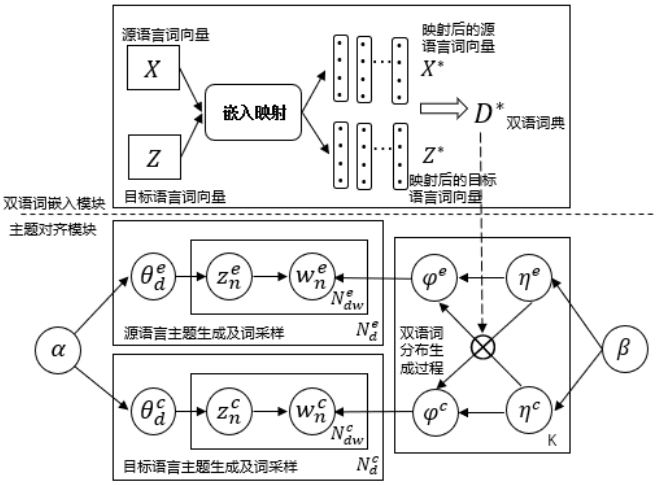

图 1 基于深度学习的主题对齐框架

2.2.1 双语词嵌入模块

双语词嵌入模块将源语言和目标语言的词嵌入映射到同一向量空间，本文借鉴了 Artetx[25]等人提出的双语词嵌入模型，将其分为嵌入映射和词典归纳两部分，其训练过程如图 1 上半部分所示。

在嵌入映射部分，设X和Z分别表示源语言和目标语言的词嵌入矩阵，X_{i*}对应于源语言中第i个词的嵌入表示，Z_{j*}对应于目标语言中第j个词的嵌入表示，同时用二进制矩阵 D 表示单词的对齐关系，即如果源语言的第i个单词与目标语言的第j个单词在训练词典中是对齐的，则$\mathrm{D}_{ij}=1$，否则为 0。那么嵌入映射的目标是找到最优的映射矩阵W^*，使得对于每个词典条目D_{ij}，映射后的源语言嵌入表示$\mathrm{X}_{i*}W$和目标语言嵌入表示Z_{j*}之间的欧几里德距离平方和最小，如公式（1）所示。

$$W^* = \arg\min_{W} \sum_{i} \sum_{j} \mathrm{D}_{ij} ||\mathrm{X}_{i*}W - Z_{j*}||^2 \quad （1）$$

在词典归纳部分，与传统的最近邻检索把源语言单词分配给目标语言中最接近的单词这种方式不同，本文使用映射后的源语言嵌入表示和目标语言嵌入表示的点积作为相似性度量，使用向量化矩阵乘法迭代计算相似度矩阵XWZ^T的每个子矩阵，得到所有单词对之间的相似性度量，每次找到它们相应的最大值，然后组合结果。在此基础上，就可以获得同一语义空间的双语对齐词典。

2.2.2 主题对齐模块

主题对齐模块是在双语词嵌入模块的基础上对不同语言或领域的主题进行对齐，本文在 Shi[26]等人提出的主题对齐模型的基础上，通过融合双语词嵌入获得的对齐词典对其进行改进。

图 1 下半部分代表双语共同主题检测机制，对于英文文档集$\mathcal{E}$和中文文档集C，每种双语共同主题k由英文词汇表Λ^e中的英文词分布φ_k^e和中文词汇表Λ^c中的中文词分布φ_k^c共同表示，同时使用双语词典，该词典由英汉多对多语义对齐的单词对组成。为了捕获多语言文档的常见语义，这里设计了两个辅助分布，分别是维度为Λ^e的η_e和维度为Λ^c的η_c，以生成词分布φ_k^e和φ_k^c。准确地说，分别从$Dirichlet$先验分布$Dir(\beta\cdot\mathbf{1}_{|\Lambda^e|})$和$Dir(\beta\cdot\mathbf{1}_{|\Lambda^c|})$中生成辅助分布$\eta_e$和$\eta_c$，其中$\mathbf{1}_D$表示分量为 1 的 D 维向量。然后从辅助分布$\eta_k^e$和$\eta_k^c$混合采样出词分布$\varphi_k^e$，其表述如公式(2)所示。

$$\varphi_k^e \propto \lambda(\eta_k^c)^T M^{c\to e} + (1-\lambda)\eta_k^e \quad （2）$$

这里，辅助分布η^e、η^c服从$Dir(\beta)$，平衡参数$\lambda\in(0,1)$用于平衡原始主题的特性和传递其他语言信息，$M^{c\to e}$代表Λ^c到Λ^e的映射矩阵$|\Lambda^c|\times|\Lambda^e|$，在同一文档集中$M_{i,j}^{c\to e}$用于映射给定中文词汇$w_i^c$后英文词汇$w_j^e$出现的概率，其概率计算如公式（3）所示。

$$M_{i,j}^{c\to e} = \frac{C(w_j^e)+1}{|T(w_j^c)| + \sum_{w^e\in T(w_j^c)} C(w^e)} \quad （3）$$

这里，$C(w_j^e)$是所有文档中w_j^e的计数，$T(w_j^c)$是在双语词典中找到w_j^c的英文翻译集。同理推导出中文词分布φ_k^c如公式（4）所示。

$$\varphi_k^c \propto \lambda(\eta_k^e)^T M^{e\to c} + (1-\lambda)\eta_k^c \quad （4）$$

由此可见，在主题级别上引入η_k^e和η_k^c可以促进词分布φ_k^e和φ_k^c共享不同语言文档中的共同语义成分。

图 1 左下侧部分分别表示英文文档和中文文档的主题生成过程。N_d^e表示英文文档数量，N_{dw}^e表示英文文档d^e中的单词数，每个英文文档d^e由 K 维主题向量θ_d^e表示，θ_d^e假设由先验分布$Dir(\alpha\cdot\mathbf{1}_K)$生成的。对于英文文档$d^e$中的每个词$w_n^e$，从$\theta_d^e$中生成其主题$z_n^e$，并从相应的英文词分布$\varphi_k^e$中生成单词$w_n^e$。中文文档的主题生成过程与英文类似，即从先验分布$Dir(\alpha\cdot\mathbf{1}_K)$中生成$\theta_d^c$，中文文档$d^c$中每个词$w_n^c$的主题$z_n^c$由$\theta_d^c$生成。并从相应的分布$\varphi_k^c$中生成单词$w_n^c$。

3 实验与讨论

本文的实验步骤是：首先通过双语词嵌入将不同语言下的单语词向量通过空间旋转映射到同一语义空间中，获得语义对齐的双语词典和同一空间的双语词向量，以解决基于翻译方法带来的上下文语义信息缺失和领域信息缺失问题；其次通过跨语言主题对齐实验研究跨语言情境下的主题词对齐效果，并利用谷歌翻译、双语词向量等手段辅助完成对齐相似度的测量；

最后通过跨领域主题文档分类实验研究跨领域下的主题词表示效果，将双语文档采用同一主题空间中的文档-主题向量进行表示，并利用机器学习中的 SVM 算法对文档分类效果进行比较分析。

3.1 数据集

本文以“英国脱欧”、“朝核问题”、“一带一路”为检索词，分别从新浪新闻和必应搜索爬取了共计 60000 条中英文新闻，内容包括新闻标题、摘要、正文共三个维度信息。文本预处理阶段，合并新闻标题、摘要和正文作为基本语料库，中文语料库使用结巴分词及停用词表操作、英文语料库同样进行分词和去停用词操作。预处理后的文本数据集描述如表 1 所示。

表 1 实验数据集描述

事件编号	新闻事件	中文语料		英文语料	
		总条数	平均词数	总条数	平均词数
1	英国脱欧	10000	90	10000	168
2	朝核问题	10000	80	10000	169
3	一带一路	10000	92	10000	166

3.2 基线方法与参数设置

鉴于在双语领域主题对齐方面，目前尚无系统性的研究，选择与本文研究最为接近的 MCTA（Multilingual Common Topic Alignment）模型[26]作为基线方法。

对于 TAM 模型，本文采用监督模式的双语词嵌入方法获得语义对齐的双语词典，其主要参数为词向量维度 200、批尺寸 5000。此外，使用支持向量机（Support Vector Machine, SVM）进行主题分类任务，使用谷歌翻译 API 进行翻译，其余参数设置如表 2 所示。

表 2 模型参数设置

参数类别	MCTA 模型	TAM
主题数 k	10-100	10-100
lambda	0.3、0.5、0.7	0.3、0.5、0.7
alpha	1.0/k	1.0/k
beta	0.005、0.01、0.05	0.005、0.01、0.05
迭代次数（训练）	100	100
迭代次数（测试）	50	50
词向量维度	--	200
批尺寸	--	5000

3.3 评价指标

本文采用的评价指标分为两部分组成，跨语言主题对齐评价指标和跨领域主题对齐评价指标。

在跨语言主题对齐方面，使用交叉集困惑度（Cross-collection Perplexity，CCP）[27]指标来衡量中英双语共享主题的质量，该指标值越低表明模型划分的主题效果越好。具体地，该指标值计算过程为：①对于每个主题$k \in \mathcal{K}$，从双语词典中查询英文单词φ_k^e和中文词φ_k^c；②将φ_k^e翻译成中文得到词分布$T(\varphi_k^e)$，将φ_k^c翻译成英文得到词分布$T(\varphi_k^c)$；③使用$T(\varphi_k^e)$拟合中文词汇$\mathcal{C}$，使用$T(\varphi_k^c)$拟合英文词汇$\mathcal{E}$。其计算如公式（5）所示。

$$\begin{aligned}\mathrm{CCP} &= \frac{1}{2}\exp\left\{-\frac{\sum_{d\in\mathcal{E},w\in d,k\in\mathcal{K}}\log p(k|\theta_d)p\left(w|T(\varphi_k^c)\right)}{\sum_{d\in\mathcal{E}}N_d^e}\right\} \\ &+ \frac{1}{2}\exp\left\{-\frac{\sum_{d\in\mathcal{C},w\in d,k\in\mathcal{K}}\sum_{k\in\mathcal{K}}log\,p(k|\theta_d)p\left(w|T(\varphi_k^e)\right)}{\sum_{d\in\mathcal{C}}N_d^c}\right\}\end{aligned} \quad (5)$$

此外，为更好地对跨语言主题对齐进行评价，本文提出了两种新的指标，即双语主题相似度（Bilingual Topic Similarity，BTS）和双语对齐相似度（Bilingual Alignment Similarity，BAS），用于评价辅助对齐的效果。BTS 是将中英对齐主题词通过谷歌翻译完成余弦相似度测量，如式（6）所示；BAS 是以对齐后的双语词向量为基础，通过查询中英词向量进行余弦相似度测量，如式（7）所示。BAS 与 BTS 指标越高，表明主题对齐效果越好。

$$\mathrm{BTS} = \mathrm{cosine}(\mathrm{vec_s}, \mathrm{vec_t}) = \frac{v_1^s v_1^t + \cdots + v_n^s v_n^t}{\sqrt{{v_1^s}^2 + \cdots + {v_n^s}^2}\sqrt{{v_1^t}^2 + \cdots + {v_n^t}^2}} \quad (6)$$

在公式（6）中，$\mathrm{vec_s} = (v_1^s, v_2^s, \ldots, v_n^s)$表示源语言（中文、英文）主题词词向量，使用谷歌翻译得到源语言主题词对应的目标语言（英文、中文）主题词，词向量为$\mathrm{vec_t} = (v_1^t, v_2^t, \ldots, v_n^t)$，计算词向量的余弦相似度。

$$\mathrm{BAS} = \mathrm{cosine}\left(V_{align}^s, V_{align}^t\right) = \frac{v_1^s v_1^t + \cdots + v_n^s v_n^t}{\sqrt{{v_1^s}^2 + \cdots + {v_n^s}^2}\sqrt{{v_1^t}^2 + \cdots + {v_n^t}^2}} \quad (7)$$

在公式（7）中，V_{align}^s和V_{align}^t表示对齐的源语言（中文、英文）和目标语言（英文、中文）主题词词向量。

在跨领域主题对齐方面，由于该任务更贴近于主题分类，因此采用通用的分类指标进行评价。具体而言，包括准确率(Precision)、召回率(Recall)和综合指标 F1 值。

3.4 跨语言主题对齐结果

本节实验用于对比本文提出的 TAM 与基线 MCTA 模型在跨语言主题层次对齐效果，使用 CCP（交叉集困惑度）、BTS（双语主题相似度）和 BAS（双语对齐相似度）三个指标进行评估，实验过程如图 2 所示。

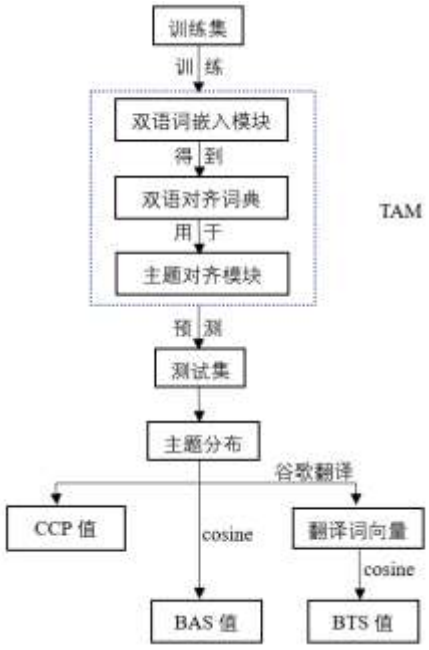

图 2 跨语言主题对齐实验过程

首先，将三个领域的中英双语新闻集按 8:2 划分训练集和测试集，使用训练集学习 TAM，将训练好的 TAM 用于测试集得到其在中英语言下的主题-词汇表示。根据主题分布计算 CCP 指标值，从而评估主题-词汇的好坏程度。对于主题下的双语词分布，如(W_{ch}, W_{en})，使用谷歌翻译得到英文词W_{en}的中文译词W_{en2zn}以及中文词W_{ch}的英文译词W_{ch2en}。然后使用中文搜狗词向量模型获得W_{ch}和W_{en2zn}的向量表示，使用英文维基词向量模型获得W_{en}和W_{ch2en}的向量表示，分别计算两组词向量的余弦相似度的平均值得到 BTS 翻译相似度指标。最后，以具有共同语义空间的双语词向量为基础，计算测试集中不同主题下双语词汇的余弦相似度平均值完成 BAS 对齐相似度指标测量。为保证实验具有可对比性，在其他实验参数相同的情况下，对每个事件选取十组不同的主题个数并对结果取均值。

表 3 显示了双语主题对齐的实验结果。对比 MCTA 和 TAM 在双语主题对齐效果比较可以发现，尽管在不同领域中 CCP、BTS 和 BAS 三个指标值均有所差异，总体上 TAM 的主题对齐效果优于基线 MCTA 模型。具体来看，第一个指标 CCP 用于衡量模型对主题词的拟合程度，其指标值越低越好。通过表 3 可知 TAM 在“英国脱欧”和“一带一路”事件中要优于基线，但在“朝核问题”事件中拟合效果较差，这可能是因为 TAM 使用语义词典作为双语对齐依据，忽视了“朝核问题”事件特有的专用词汇。第二个指标 BTS 通过余弦相似度计算中英互译主题词的相似程度，其相似度越高说明语义信息越接近。由表 3 可以看出，针对 BTS 指标，TAM 在“朝核问题”和“一带一路”事件中要优于基线 MCTA 模型，但在“英国脱欧”事件中相似性较差，这可能是因为翻译过程中一词多义现象造成的。针对第三个指标 BAS，采用余弦相似度计算中英主题词向量在同一空间中的相似程度，可以看出除“英国脱欧”事件相似性略低外，“朝核问题”和“一带一路”事件的 TAM 相似效果要优于基线，这表明所提模型在双语主题对齐方面具有很好的效果。

表 3 MCTA 和 TAM 双语主题对齐效果比较

领域	CCP		BTS		BAS	
	MCTA	TAM	MCTA	TAM	MCTA	TAM
英国脱欧	2404.39	2342.67	0.8975	0.8817	0.8122	0.8115
朝核问题	2269.28	2342.15	0.8810	0.8846	0.8158	0.8172
一带一路	2587.11	2470.72	0.8872	0.9002	0.7434	0.7866
平均	2420.26	**2385.18**	0.8886	**0.8888**	0.7905	**0.8051**

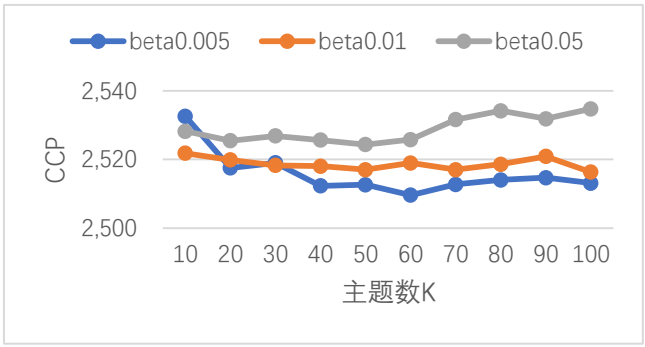

图 3 不同主题下超参数 beta 对 CCP 值影响

为探寻不同实验参数对主题对齐效果的影响，以“一带一路”事件为例，对 TAM 展开扩展实验，并对最优的对齐实验结果进行具体分析。图 3 展示了超参数 beta 和主题数 K 对领域主题对齐任务在 CCP 指标上的影响情况。CCP 可以衡量对齐模型对文档所属主题的不确定性程度，可以用于主题对齐效果评估。超参数 beta 是每个主题下词的多项分布的 Dirichlet 先验参数，beta 值变小，则一个词尽可能属于同一个主题。在 beta 取值为 0.005、0.01、0.05 三种情形时，CCP 值随主题数变动趋势差异明显。具体来看，当 beta 为 0.005 时，CCP 值随主题数增加而迅速减小后趋于平缓，在主题数大于 10 时主题对齐效果要明显好于 beta 为 0.01 和 0.05 的情况；当 beta 为 0.01 时，在所有主题数上 CCP 表现都很平稳，主题数的改变对 CCP 的影响较小；当 beta 为 0.05 时，CCP 值会随主题数的增加而上升，同时可以看出较大的 beta 值在不同的主题数下的表现均较差。总体来看，当 beta 为 0.005 时可以取得更好的对齐效果。

为进一步探究主题对齐的实际效果，本文从 TAM 对齐的主题文档中随机抽取了一部分领域主题，进行定性分析，作为对定量指标评价的补充。表 4 列举了 TAM 在“一带一路”领域中主题数为 40、超参数 beta 为 0.005、平衡参数 lambda 为 0.5 条件下的 5 个对齐的领域主题及相应的双语对齐词汇情况。由表 4 可以看出，编号为 8 的主题可以理解为一带一路下推动境外投资的主题词，如推动（advocates）、香港（hk）；编号为 19 的主题可能代表与欧盟贸易往来，如中国（China）、中欧（europe）等；编号 28 的主题可能代表我国区域发展，如发展（developing）、经济（economy）；编号 34 的主题可能代表促进世界文化交流，如连通（bridge）、世界各地（world）；编号 37 的主题可能代表国内投资市场，如投资（invest）、政策（policy）等。通过对比可知，中英双语文档集合在领域主题上具有较高的一致性，这从定性方面验证了 TAM 在双语领域主题对齐上的有效性。

表 4 主题数为 40 时的部分领域主题与词汇

编号	双语主题词汇
8	推动 战略 国家 国际 发展 香港 沿线 人民币 企业 经济 服务 跨境 贸易 投资 advocates strategic status nation country nations hk closer yuan powers singapore
19	中国 国际 国家 班列 服务 品牌 战略 研究 中欧 沿线 铁路 东盟 检验 口岸 俄罗斯 人才 china status international nation europe countries complex expansion connections
28	战略 经济 发展 改革 加快 推进 创新 实施 建设 供给 区域 京津冀 长江 国家 政策 企业 strategic strategy risks moves growing challenges influence developing expanding economy
34	贯通 新丝路 丝路 大通道 丝绸之路 连通 经济带 助力 世界各地 世博会 民族团结 动能 road belt world york national big bridge island south top located park solar largest energy
37	投资 投资者 板块 主题 市场 政策 互联网 机会 行业 改革 证券 基金 行情 有望 军工 investing invest plate investments calendar policies market shift policy credit presents

3.5 跨领域主题对齐结果

针对跨领域情境下的主题对齐，目前尚缺乏系统性的评价语料和评价指标。为了更准确地量化领域主题的对齐效果，本文将“英国脱欧”、“朝核问题”和“一带一路”三个领域的文档融合在一起，依照源语言和目标语言的不同将文档集合划分为训练集和测试集，将原问题转换为跨语言情境下的领域主题分类问题（即在跨语言情境下，给定一篇文档，识别其所对应的主题）。本文使用 TAM 将中英文文档投影到同一主题空间中，使用机器学习方法训练源语言主题分类模型，并将其用于预测目标语言文档的主题类型。具体而言，实验包括两个子任务，一是使用中文主题向量训练的主题分类模型预测英文测试集中文档的主题；二是使用英文主题向量训练的主题分类模型预测中文测试集中文档的主题。训练集和测试集均包含三个领域主题标签，且数据量以 10:1 划分。由于实验目的是为了比较两种模型（MCTA 和 TAM）的主题分类差异，这里统一采用传统的基分类器，即 L1 正则化的 SVM 算法。

表 5 显示了 MCTA 和 TAM 在领域主题分类任务

上的效果对比。由表 5 可以看出，在中文预测英文（cn-en）和英文预测中文（en-cn）这两类主题分类任务中，TAM 在准确率、召回率、F1 值方面均优于基线 MCTA 模型，且在中文预测英文任务中提升更为明显。具体而言，在中文预测英文主题分类任务中，相对于 MCTA 方法，除“朝核问题”领域中召回率有所降低外，在其他领域中三项指标均有显著提升。在宏平均上，相比于 MCTA 模型（0.722），TAM 的 F1 值提升 10%左右（至 0.822）。这表明使用中文文档的主题向量训练的主题分类模型，可以更好地预测英文文档主题。在英文预测中文主题分类任务中，TAM 模型虽然在“朝核问题”事件表现略差，但总体来看，TAM 与 MCTA 模型在使用英文文档预测中文文档分类任务中取得了与之相近的分类效果。通过对实验过程进行分析，其可能的原因在于英文文档在训练阶段的主题特征提取存在过拟合现象，导致在预测中文文档时并没有获得更好的主题分类效果。

表 5 MCTA 和 TAM 在跨语言跨领域主题对齐上的效果对比

语言	领域	Precision		Recall		F1	
		MCTA	TAM	MCTA	TAM	MCTA	TAM
cn-en	英国脱欧	0.761	0.842	0.696	0.776	0.727	0.808
	朝核问题	0.687	0.819	0.926	0.893	0.789	0.855
	一带一路	0.767	0.808	0.562	0.802	0.649	0.805
	宏平均	0.738	**0.823**	0.728	**0.824**	0.722	**0.822**
en-cn	英国脱欧	0.975	0.976	0.956	0.960	0.965	0.968
	朝核问题	0.981	0.978	0.987	0.984	0.984	0.981
	一带一路	0.967	0.973	0.979	0.983	0.973	0.978
	宏平均	0.974	**0.976**	0.974	**0.975**	0.974	**0.976**

图 4 展示了 TAM 中平衡参数 lambda（在抽取双语对齐词汇时进行概率加权）和主题数 K 对主题分类 F1 值的影响情况。这里以使用中文文档的主题特征来预测英文文档主题类别为例。具体来看，当 lambda 为 0.3 时，随着主题数的增加，F1 值急剧下降，在主题数达到 50 以后，其分类效果趋于平缓，可能的原因是 lambda 取值过小导致主题模型在选择双语主题词汇时，更倾向于同语言词汇，忽视了部分双语主题词间的映射关系。当 lambda 为 0.5 时，表示主题模型在生成主题词时会同等考虑双语对齐词汇，因而当主题数小于 50 时，其 F1 值的变动曲线基本位于 0.3 和 0.7 曲线之间，在主题数 50 以后变动趋于平缓。当 lambda 为 0.7 时，由于主题模型在生成主题词时会优先考虑另一种语言的词汇，虽然这可以捕获更多的另一种语言的主题特征，但忽视了语言内的部分信息，因而其 F1 效果虽然优于 lambda 为 0.3 的情形，但没有 lambda 为 0.5 的效果好。总体来看，当 lambda 为 0.5 时，综合考虑了主题模型在对齐过程中平衡选择双语主题特征的需求，所以在选择参数 lambda 为 0.5、主题数为 20 时，可以达到最优 F1 值为 0.8385。这表明，通过选择合理的参数，TAM 能够较好地解决跨语言情境下的主题分类问题。

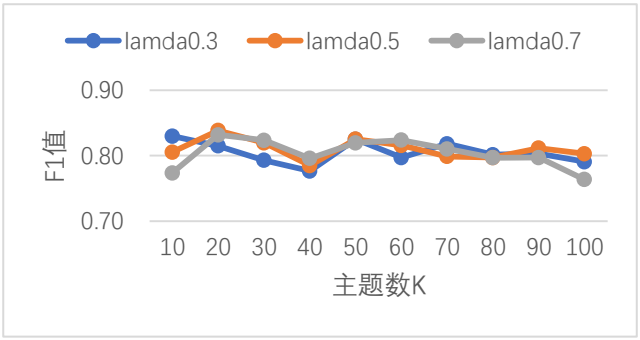

图 4 不同 lambda 下主题数目对 F1 值影响

3.6 讨论

以上实验结果证实了 TAM 在跨语言主题对齐和跨领域主题对齐中均优于基线 MCTA 模型，这证明了 TAM 在解决跨语言跨领域主题对齐问题上的有效性。

在跨语言主题对齐实验中，通过主题对齐任务验证了双语主题的跨语言对齐问题。相比于传统双语主题模型仅能实现句子与句子、篇章与篇章等粗粒度对齐，所提方法通过设计辅助分布促进不同语言词分布间的语义共享，以此实现了跨语言研究中主题级别的细粒度对齐，这为解决跨语言环境下不同领域知识的对齐提供了基础。此外，在探寻不同实验参数对主题对齐效果影响中发现，在固定其他参数条件下，较小的 beta 参数有利于主题词的对齐，通过观察 40 个主题的对应词汇可知，中英双语文档集合在领域主题上

具有较高的一致性。

在跨领域主题对齐实验中，与基线相比，通过双语词嵌入改进后的 TAM 在使用中文文档主题向量预测英文文档主题类别任务中表现更为出色。这是因为双语词嵌入改进了基于翻译的双语词典，对于每个源语言词汇，采用余弦相似度选出了最接近的前五个候选词进行语义对齐，语义对齐能够有效提升主题对齐的效果。首先，语义对齐有效扩充了双语语义映射关系。相比于基于翻译的一对一词汇对，经过双语词嵌入扩充后可实现一对多语义词汇对，极大地丰富了不同领域词汇间的对应关系。例如，在“朝核问题”新闻中，对于英文词汇“North”来说，采用翻译的中文对应词汇应为“北”、“北方的”，而采用双语词嵌入则会根据“North”的上下文语义向量计算出最接近的中文词汇应为“朝鲜”“核武”等相关词汇，这不仅解决了中英文词汇间的语言跨度，还解决了不同领域间的语义联系，实现不同语言下领域知识间的语义对齐。其次，经过词嵌入产生的语义词典不再是严格意义上的翻译关系，而是源语言词汇与目标语言词汇所存在的语义关系。不同语言的单语词向量作为双语词嵌入的输入，可以间接地将单语词汇的上下文信息通过嵌入映射和词典归纳两部操作实现双语语义对齐，这种可以利用捕获到的上下文信息完成对齐过程是基于统计机器翻译所不能比拟的，同时该方法在某些稀疏领域或稀缺语料中亦可以取得很好的效果。

从模型的推广性来看，TAM 除了用于不同语言不同领域的主题检测和对齐任务外，还可推广到更多的应用场景。从知识利用角度来看，使用种子词典就可以完成双语词汇的语义对齐，解决了基于翻译方法带来的语义缺失和领域缺失问题，该方法可推广到语料稀缺情境下的信息处理问题。从语料使用来看，TAM 不局限于传统的平行或可比语料库，可以推广到非对齐语料的相关应用中，能够减轻研究人员建设对齐语料库的工作。从研究任务来看，TAM 可推广到主题级别的词汇对齐任务中，可以获得比传统跨语言主题模型仅实现句子或篇章等粗粒度对齐更多的知识信息。从主题对齐效果来看，所提方法均优于采用翻译词典的基线方法，且主题的可解释性达到实验预期，因此可应用到不同语言与不同领域的舆情挖掘与趋势分析工作中。从创新性来看，利用双语词嵌入改进主题模型中翻译词典缺乏领域信息这一思路，可推广到机器翻译、跨语言信息检索和跨语言自动摘要等情景。

从模型的理论价值来看，模型对于解决跨语言与跨领域情境下的主题一致性和可解释性问题具有重要意义。对于一致性问题，经典的 BiLDA 模型采用单词一对一映射，但该方法所识别的共同主题对应关系局限于词汇对水平，且共享的主题关系会因为双语词频变化失衡，对此本文在 TAM 模型中采用两个辅助分布用于生成双语词分布，以捕获双语知识的常见语义。一种语言的词分布会以一定的平衡概率从两种语言构成的辅助分布中同时获得，这种概率用于平衡原始主题的特性及传递其他语言信息；同时一种语言的辅助分布在生成另一种语言的词分布时，还会以对齐映射矩阵进行加权，由于双语词典可以将双语词汇表示为多对多映射，所以该矩阵中的元素可以根据双语映射词汇间的出现频次进行表示，例如给定中文词汇后英文词汇出现的概率。对于可解释性问题，已有的方法是采用双语词典或知识库弥合语言差异，由于基于翻译的词典并未考虑到词汇的上下文信息和领域信息，造成不同主题下的词汇同质化现象，这对于主题划分、解释和分析带来困难。对此本文采用双语词嵌入方法获得语义对齐的双语词典，在此基础上拓展了语义对齐的词汇。

4 结语

本文以跨语言主题对齐为切入点，提出一种融合双语词嵌入的主题对齐模型，在中英双语新闻数据集中的实验结果表明，该方法在双语主题对齐任务和跨语言主题分类两项任务中均获得了比基线 MCTA 模型更好的效果。TAM 与传统的跨语言主题模型不同之处在于：1）通过双语词嵌入方法将不同单语空间中的词向量对齐到同一双语空间中，构建包含语义信息和领域信息的双语语义对齐词典；2）在主题级别上引入辅助分布促进不同语言词分布的语义共享，以此改善跨语言研究中主题-词汇对齐效果；3）采用部分折叠的 Gibbs 采样器将语义对齐词典中不同语言的领域知识关系结合到共同主题检测中，以便有效地用于模型参数学习。

本文的不足之处在于实验中只选取了中文和英文两种语言，未探讨模型在其它语言上的对齐和分类效果，在后续研究中，考虑选择更多语言以检验模型的有效性。

参考文献

[1] Papadimitriou C.H., Raghavan P., Tamaki H. and Vempala S. Latent semantic indexing: A probabilistic analysis [J]. Journal of Computer and System Sciences,2000, 61(2):217-235.

[2] 夏青, 严馨, 余正涛, 等. 融合要素及主题的汉越双语新闻话题分析[J]. 计算机工程, 2016, 42(9):186-191.

[3] 唐莫鸣, 朱明玮, 余正涛, 等. 基于双语主题和因子图模型的汉语-越南语双语事件关联分析[J]. 中文信息学报, 2017, 31(6):125-131.

[4] 司莉, 陈雨雪, 曾粤亮. 基于多语言本体的中英跨语言信息检索模型及实现[J]. 图书情报工作, 2017(1):100-108.

[5] 余传明, 冯博琳, 田鑫, 等. 基于深度表示学习的多语言文本情感分析[J]. 山东大学学报：理学版, 2018，53（3）：13-23.

[6] 许海云, 董坤, 刘春江, 等. 文本主题识别关键技术研究综述[J]. 情报科学, 2017(01):155-162.
the Association for Computational Linguistics. Association for Computational Linguistics, Berlin,2016: 666–675.

[10] Liu Yang, Liu Zhiyuan, Chua T S, et al. Topical word embeddings[EB/OL].[2018-02-19]. https://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/viewPaper/9314.

[11] Zhang Heng, Zhong Guoqiang. Improving short text classification by learning vector representations of both words and hidden topics[J]. Knowledge-Based Systems, 2016,102(15):76-86.

[12] Moody C E . Mixing Dirichlet Topic Models and Word Embeddings to Make lda2vec [EB/OL].[2018-05-06]. https://arxiv.org/abs/1605.02019.

[13] Li D , Li Y , Wang S . Topic Enhanced Word Vectors for Documents Representation[M]. Singapore:Springer,2017,166-177.

[14] 杨奇奇. 基于多主题空间的跨领域文本分类方法研究[D]. 合肥：合肥工业大学，2017：7-9.

[15] Wu T, Lei Z, Qi G, et al. Encoding Category Correlations into Bilingual Topic Modeling for Cross-Lingual Taxonomy Alignment[M]. Cham: Springer,2017,728-744.

[16] Tamura, A, and Eiichiro S. Bilingual Segmented Topic Model[C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Berlin,2016:1266-1276.

[17] Duo Zhang, Mei Qiaozhu, and Zhai ChengXiang. Cross-lingual latent topic extraction[C]// Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Uppsala,2010:1128-1137.

[18] Tao Zhang,Liu Kang, Zhao Jun. Cross Lingual Entity Linking with Bilingual Topic Model[EB/OL].[2013-06-30].https://www.aaai.org/ocs/index.php/IJCAI/IJCAI13/paper/viewPaper/6268.

[19] Wu Tianxing, Qi Guilin, Wang Haofen,et al. Cross-Lingual Taxonomy Alignment with Bilingual Biterm Topic

[7] 余传明, 安璐. 从小数据到大数据——观点检索面临的三个挑战[J]. 情报理论与实践, 2016, 39(2)：13-19.

[8] Wei Xing. LDA-Based Document Models for Ad-Hoc Retrieval[C]// Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, Seattle,2006: 178-185.

[9] Li S, Chua T S , Zhu J , et al. Generative Topic Embedding: a Continuous Representation of Documents (Extended Version with Proofs)[C]//Proceedings of the 54th Annual Meeting of
Model[EB/OL].[2018-06-21]. https://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/viewPaper/12011.

[20] Heyman G, Ivan V, and Marie-Francine M. C-BiLDA extracting cross-lingual topics from non-parallel texts by distinguishing shared from unshared content[J].Data Mining and Knowledge Discovery,2016, 30(5): 1299-1323.

[21] Rus V, Nobal N, and Rajendra B. Similarity measures based on latent dirichlet allocation[M]. Berlin:Springer,2013,459-470.

[22] Li Lianghao, Jin Xiaoming, Long Mingsheng. Topic Correlation Analysis for Cross-Domain Text Classification [EB/OL].[2018-07-14]. https://www.aaai.org/ocs/index.php/AAAI/AAAI12/paper/viewPaper/5077.

[23] Yang Pei, Gao Wei , Tan Qi, et al. A link-bridged topic model for cross-domain document classification[J]. Information Processing & Management, 2013, 49(6):1181-1193.

[24] 杨奇奇, 张玉红, 胡学钢. 一种基于多桥映射的跨领域文本分类方法[J]. 计算机应用研究, 2018, 35(4): 996-1000.

[25] Artetxe M, Gorka L, and Eneko A. Learning bilingual word embeddings with (almost) no bilingual data[C]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver:ACL,2017:451-462.

[26] Shi Bei, Wai L, and Bing Lidong, et al. Detecting Common Discussion Topics Across Culture from News Reader Comments [C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin:ACL,2016:676-685.

[27] Zhang D, Mei Q, Zhai C. Cross-Lingual Latent Topic Extraction[C]// Meeting of the Association for Computational Linguistics, Uppsala：ACL, 2010: 1128-1137.

余传明（1978-），博士，教授，主要研究领域为数据挖掘、信息检索与商务智能。

原赛（1996-），硕士研究生，主要研究方向为数据挖掘与大数据分析。

胡莎莎（1996-），硕士研究生，主要研究方向为数据挖掘与大数据分析。

安璐（1979-），通信作者，博士，教授，博士生导师，主要研究领域为可视化知识发现、网络数据分析。E-mail：anlu97@163.com。

修改情况说明

REVIEWER #1：

稿件的主要贡献是在主题深度表示学习的基础上，提出一种融合双语词嵌入的主题对齐模型，以改善跨语言和跨领域情境下的主题对齐效果。同时作者还提出了两种新的指标，即 BTS 和 BAS，用于评价辅助分布对齐的效果。作者进行了较为丰富的在跨语言和跨领域主题对齐上的实验及详尽的分析和讨论，实验结果相对于基线方法有较大进步。

建议：1）摘要中，双语翻译相似度 翻译为 bilingual topic similarty 似乎不妥。建议做适当调整

作者答复：

感谢评审专家的宝贵意见。

对于建议 1），为了更准确地表达“bilingual topic similarity”，并使得中英文表达一致，我们将其中文翻译改为“双语主题相似度”。

建议：2）图 4 的题目是否应该改为不同的 lambda 情况下，主题数目对 f1 值的影响？

作者答复：

感谢评审专家的宝贵意见。

对于建议 2），图 4 展示了本文模型中平衡参数 lambda 和主题数 K 对主题分类 F1 值的影响情况。对图 4 的文字描述部分分析了 lambda 取值固定时，主题分类 F1 值随着主题数 K 的变化情况，鉴于此，我们将图 4 的题目修改为“不同 lambda 下主题数目对 F1 值影响”。

REVIEWER #2：

本文提出了一种融合双语词嵌入的主题对齐模型（Topic Alignment Model, TAM）。该模型通过双语词嵌入来扩充语义对齐词汇词典，并通过设计辅助分布用于改进不同词分布的语义共享，旨在改善跨语言和跨领域情境下的主题对齐效果。相比于传统的对齐模型，该模型在跨语言主题对齐任务中的双语对齐相似度提升了约 1.5%，在跨领域主题对齐任务中的 F1 值提升了约 10%。

论文条理清晰，观点新颖，实验详尽，有理有据。

作者答复：

感谢评审专家的宝贵意见。

REVIEWER #3：

针对双语主题模型的对齐问题，作者提出了一种融合双语词嵌入的主题对齐模型（Topic Alignment Model, TAM），通过双语词嵌入扩充语义对齐词汇词典，改善跨语言和跨领域情境下的主题对齐效果。同时提出了两种新的评价指标 BTS 和 BAS。实验较充实，实验结果优于传统的基线模型。

该工作具有较好的研究动机，实验较为充分。但是存在以下问题：（1）排版不美观，例如表 2 的参数设置，可以通过语言描述，如果使用表格表示，需要在同一页面展示且不要占用过多的篇幅；且第五页具有较多的空白，需要进一步调整。（2）实验对比的基线模

型可以再多做几个，仅有一个对比实验显得过于单薄。（3）建议将 BAT 和 BTS 的数学表达展示出来，可能会更好。

作者答复：

（1）为了使排版更美观,我们删除了第 5 页多余的空白，并将 3.2 节最后一段以及表 2 修改为如下：

“对于 TAM 模型，本文采用监督模式的双语词嵌入方法获得语义对齐的双语词典，其主要参数为词向量维度 200、批尺寸 5000。此外，使用支持向量机（Support Vector Machine, SVM）进行主题分类任务，使用谷歌翻译 API 进行翻译，其余参数设置如表 2 所示。

表 2 模型参数设置

参数类别	MCTA 模型	TAM
主题数 k	10-100	10-100
lamda	0.3、0.5、0.7	0.3、0.5、0.7
alpha	1.0/k	1.0/k
beta	0.005、0.01、0.05	0.005、0.01、0.05
迭代次数（训练）	100	100
迭代次数（测试）	50	50
词向量维度	--	200
批尺寸	--	5000

”

（2）感谢评审专家的宝贵意见。针对双语领域主题对齐研究，我们采用关键词“cross-lingual topic alignment”及“multi-lingual topic alignment”等对国内外文献进行了广泛搜索。目前来看，相关的研究并不多见。就搜索结果而言，MCTA（Multilingual Common Topic Alignment）模型与本文探讨的研究问题最为接近。鉴于此，我们将其与本文的实验进行了对比研究。此外，我们正在持续检索和关注双语领域主题对齐的最新研究进展，若发现有新的相关研究出现，我们将尝试与本文模型进行对比。

（3）为了更清晰地表达 BTS 和 BAS 评价指标，在 3.3 节第三段末尾增加如下内容：

“

$$\mathrm{BTS}=\mathrm{cosine}(\mathrm{vec_s},\mathrm{vec_t})=\frac{v_1^s v_1^t+\cdots+v_n^s v_n^t}{\sqrt{{v_1^s}^2+\cdots+{v_n^s}^2}\sqrt{{v_1^t}^2+\cdots+{v_n^t}^2}} \quad (6)$$

在公式（6）中，$\mathrm{vec_s}=(v_1^s,v_2^s,\ldots,v_n^s)$表示源语言（中文、英文）主题词词向量，使用谷歌翻译得到源语言主题词对应的目标语言（英文、中文）主题词，词向量为$\mathrm{vec_t}=(v_1^t,v_2^t,\ldots,v_n^t)$，计算词向量的余弦相似度。

$$\mathrm{BAS}=\mathrm{cosine}\left(V_{align}^s,V_{align}^t\right)=\frac{v_1^s v_1^t+\cdots+v_n^s v_n^t}{\sqrt{{v_1^s}^2+\cdots+{v_n^s}^2}\sqrt{{v_1^t}^2+\cdots+{v_n^t}^2}} \quad (7)$$

在公式（7）中，V_{align}^s和V_{align}^t表示对齐的源语言（中文、英文）和目标语言（英文、中文）主题词词向量。”

再次诚挚地感谢您的宝贵意见！