CN 43-1258/TP
ISSN 1007-130X

计算机工程与科学
Computer Engineering & Science

第 39 卷第 5 期 2017 年 5 月
Vol. 39, No. 5, May. 2017

文章编号:1007-130X(2017)05-0978-06

基于双语 LDA 的跨语言文本相似度计算方法研究*

程　蔚[1,2],线岩团[1,2],周兰江[1,2],余正涛[1,2],王红斌[1,2]

(1.昆明理工大学信息工程与自动化学院,云南 昆明 650500;
2.昆明理工大学智能信息处理重点实验室,云南 昆明 650500)

摘　要:基于双语主题模型思想分析双语文本相似性,提出基于双语 LDA 跨语言文本相似度计算方法。先利用双语平行语料集训练双语 LDA 模型,再利用该模型预测新语料集主题分布,将新语料集的双语文档映射到同一个主题向量空间,结合主题分布使用余弦相似度方法计算新语料集双语文档的相似度,使用从类别间和类别内的主题分布离散度的角度改进的主题频率-逆文档频率方法计算特征主题权重。实验表明,改进后的权重计算对于基于双语 LDA 相似度算法的召回率有较大提高,算法对类别不受限且有较好的可靠性。

关键词:双语 LDA;跨语言文本相似度;余弦相似度;主题频率-逆文档频率

中图分类号:TP391　　**文献标志码**:A

doi:10.3969/j.issn.1007-130X.2017.05.024

A cross-lingual document similarity calculation method based on bilingual LDA

CHENG Wei[1,2], XIAN Yan-tuan[1,2], ZHOU Lan-jiang[1,2], YU Zheng-tao[1,2], WANG Hong-bin[1,2]

(1. School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500;
2. Key Laboratory of Intelligent Information Processing,
Kunming University of Science and Technology, Kunming 650500, China)

Abstract: Based on the idea of bilingual topic model, we analyze similarity of bilingual documents and propose a cross-lingual document similarity calculation method based on bilingual LDA. Firstly we use the bilingual parallel documents to train the bilingual LDA model and then use the trained model to predict the topic distribution of the new corpus. The new corpus's bilingual documents are mapped to the vector space of the same topic. We use the cosine similarity method and topic distribution combined to calculate the similarity of the bilingual documents of the new corpus. We improve the topic frequency inverse document frequency method from the aspect of the dispersion of in-category and the between-category topic distribution, and utilize the improved method to calculate feature topic weights. Experimental results show that the improved weight calculation method can enhance the recall rate, enable the LDA similarity calculation algorithm not limited to certain categories, and it is reliable.

Key words: bilingual LDA; cross-lingual document similarity calculation; cosine similarity; topic frequency-inverse document frequency

* **收稿日期**:2015-12-29;**修回日期**:2016-02-23
基金项目:国家自然科学基金(61363044,61462054);云南省科技厅面上项目(2015FB135);云南省教育厅科学研究基金(2014Z021);昆明理工大学省级人培项目(KKSY201403028)。
通信作者:周兰江(915090822@qq.com)
通信地址:650500 云南省昆明市呈贡区昆明理工大学信息工程与自动化学院
Address: School of Information Engineering and Automation, Kunming University of Science and Technology, Chenggong District, Kunming 650500, Yunnan, P. R. China

1　引言

跨语言文本相似度计算是跨语言论文查重、平行语料库构建、跨语言文本推荐、跨语言信息检索等领域的基础工作,有着重要意义。现有的方法主要基于双语词典或双语知识库、平行语料库、统计翻译以及机器翻译等方法。Steinberger[1]使用多语言词典 EUROVOC,将多语言文本用中间语义表征,不再依赖语言,以此来计算文本相似度;何文垒[2]使用 WordNet 计算中英文文本相似度,提出了一种独立于语言的中间语义层,在这之上实现了一种名词语义哈希编码,将文本转化成中间层语义哈希特征序列,从而计算文本相似度;Pottast[3]基于平行语料库提出 CL-ESA(Cross-Language Explicit Semantic Analysis)方法,使用维基百科将要计算相似度的两种语言文本表示成向量,然后通过余弦相似度计算两种语言文本的相似度;王洪俊[4]基于双语词典以及统计翻译模型,计算文档互译词对数,为互译词对赋予权重,使用改进的 Dice 方法进行双语文档相似度计算;Alberto[5]基于统计翻译模型提出 CLiPA(Cross-Lingual Plagiarism Analysis)算法,该算法根据贝叶斯原理估计两种语言文档互译的概率,翻译概率由统计翻译模型计算得到;Uszkoreit[6]提出通过查询基于词典翻译的双语文本的 N-gram 识别双语对齐文本的方法;Hasan[7]提出使用 MI(Mutual Information)和 RIDF(Residual Inverse Document Frequency)两种方法计算双语文档的相似度,通过统计方法实现中日文档对齐;Maike[8]是将源语言文本翻译成目标语言文本,并在目标语言空间计算相似度。上述跨语言文本相似度计算过程中,双语知识是匹配的基础条件。双语词典无法解决未登录词问题,使用词频计算相似度会导致维度过高,机器翻译依赖于翻译的质量。新的方法从主题层面出发尝试解决相似性计算问题,Preiss[9]将源语言主题模型翻译成目标语言主题模型后计算文档的主题相似性;Ivan[10]使用双语主题模型获得平行语料库的词概率分布结合相似度方法发现互译词对;王振振[11]使用 LDA(Latent Dirichlet Allocation)得到中文语料集的主题概率分布结合 KL 散度计算中文文本相似度;Yuan[12]先从维基百科上抽取汉藏实体扩展汉藏词典,利用 LDA 将文本映射到主题的特征空间上,然后根据新闻文本的特点计算跨语言文本相似性。由于主题模型可以挖掘文本潜在语义信息,发现词与主题关系、主题与文档关系且能对文本降维,双语主题模型不需要双语知识,也避免了使用词频逆文档频率计算相似度忽视语义信息,可以有效解决传统方法存在的问题。双语 LDA 是由平行语料训练产生,其主要思想是平行文本共享同一主题分布,同一个主题可以由不同语言来表示。Ni[13]提出 ML-LDA(MultiLingual Topic-Latent Dirichlet Allocation)模型从维基百科平行语料中提取主题,每个主题可以由多种语言表示,使得多语言文档可以用统一的主题表示,Mimno[14]提出的 PLTM(PolyLingual Topic Model)模型与之基本一致。文献[9]的方法依赖翻译质量,文献[12]依然需要双语知识库。所以针对双语主题模型特点,本文提出基于双语 LDA 主题模型,将双语文本映射到同一个主题向量空间,在这个向量空间内结合主题分布使用余弦相似度方法计算双语文本相似度,采用改进的主题频率-逆文档频率 IT-FIDF(Improved Topic Frequency Inverse Document Frequency)方法计算特征主题权重,这样算法无需翻译就能将双语文档放在同一向量空间下进行相似度计算,而且不需要双语知识,最后根据相似度找到与源语言相似的目标语言文本。

本文第 2 节将介绍双语 LDA 的构成,第 3 节将介绍双语文本相似度计算,第 4 节将介绍相关实验和结果分析,第 5 节将对本次研究进行总结。

2　双语 LDA

该模型是 LDA[15]的一种扩展,不同于传统 LDA 的每个文档都有独立的主题分布。平行语料共享主题分布,不同语言描述同一主题,且平行语料是由多语言描述的,词分布可以不同,根据文献[13,14]对于多语言主题模型的描述,在此基础上构建双语 LDA 模型,图 1 表示其概率图。

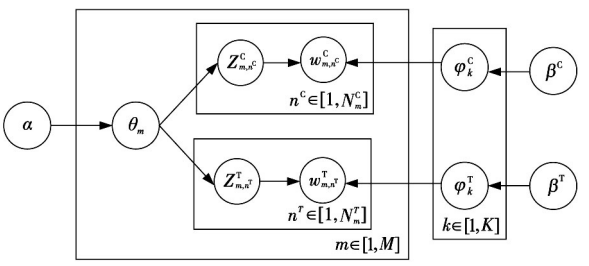


Figure 1　Bilingual LDA probability graph

图 1　双语 LDA 概率图

图1中,C 代表源语言,T 代表目标语言,θ_m 表示第 m 对双语平行文档的主题概率分布,φ_k^C、φ_k^T 分别表示主题 Z_k 在源语言和目标语言的词汇分布概率,Z_{m,n^C}^C、Z_{m,n^T}^T 分别表示第 m 对双语平行文档的源语言和目标语言的第 n 个词项的隐含主题,w_{m,n^C}^C、w_{m,n^T}^T 分别表示第 m 对双语平行文档的源语言和目标语言的第 n 个词项,M 表示文档集总对数,N_m^C、N_m^T 分别表示第 m 对双语平行文档的源语言和目标语言词项总数,θ_m 服从 Dirichlet 分布且 α 是其先验参数并用于产生主题,φ_k^C、φ_k^T 服从 Dirichlet 分布且 β^C、β^T 是其对应的先验参数并用于产生词项。图2是双语 LDA 生成过程。

对主题 k, $k\in[1,K]$

采样源语言词概率分布 $\varphi_k^C \sim Dirichlet(\beta^C)$;

采样目标语言词概率分布 $\varphi_k^T \sim Dirichlet(\beta^T)$;

对语料中的第 m 对源语言、目标语言文档对, $m\in[1,M]$

采样主题概率分布 $\theta_m \sim Dirichlet(\alpha)$;

对源语言文档 d_m^C 的第 n^C 个词, $n^C\in[1,N_m^C]$

选择隐含主题 $Z_{m,n^C}^C \sim Multinomial(\theta_m)$;

生成一个单词 $w_{m,n^C}^C \sim Multinomial(\varphi_k^C, Z_{m,n^C}^C)$;

对目标语言文档 d_m^T 的第 n^T 个词, $n^T\in[1,N_m^T]$

选择隐含主题 $Z_{m,n^T}^T \sim Multinomial(\theta_m)$;

生成一个单词 $w_{m,n^T}^T \sim Multinomial(\varphi_k^T, Z_{m,n^T}^T)$;

Figure 2 Bilingual LDA generation process

图2 双语 LDA 生成过程

Gibbs sampling 公式如下:

$$P(Z_{i,l}=k \mid Z_{\neg i,l}, w^l) \propto \frac{n_{k,\neg i,l}^v+\beta_l^v}{\sum_{v=1}^{V}(n_{k,l}^v+\beta_l^v)-1} \cdot \frac{n_{d^l,\neg i}^k+\alpha_k}{\sum_{k=1}^{K}(n_{d^l}^k+\alpha_k)-1}$$

其中,$l\in(C,T)$,$n_{k,l}^v$ 为主题 Z_k 中出现词 v 的次数,β_l^v 为词项 v 的 Dirichlet 先验,$n_{d^l}^k$ 为文档 d^l 中出现主题 Z_k 的次数,α_k 为主题 Z_k 的 Dirichlet 先验,$\neg i$ 表示排除第 i 项。利用 Gibbs sampling 公式最终得到:

$$\varphi_k^l=\frac{n_{k,l}^v+\beta_l^v}{\sum_{v=1}^{V}(n_{k,l}^v+\beta_l^v)}$$

其中,φ_k^l 为主题 Z_k 中词项概率 v 分布,对于新的语料集,保持上述模型的词分布不变,进而通过 Gibbs sampling 可以计算新语料集的主题分布:

$$P=(Z_k \mid d^l)=\theta_{d^l,k}=\frac{n_{d^l}^k+\alpha_k}{\sum_{k=1}^{K}(n_{d^l}^k+\alpha_k)}$$

其中,$\theta_{d^l,k}$ 表示主题 Z_k 在文档 d^l 出现的概率。

3 双语文本相似度计算

对相似性文档做出如下定义:描述同一话题或主题的文档视为相似性文档。

相似度计算流程:先对语料进行预处理,包括分词、去除停用词等;然后用训练集训练得到双语 LDA 模型和 φ_k^l,保持 φ_k^l 不变结合双语 LDA 预测新语料集的 $\theta_{d^l,k}$,构建出新语料集的主题向量空间,利用该 $\theta_{d^l,k}$ 计算新语料集源语言文本与目标语言文本的相似度,筛选出符合要求的相似性文本。

跨语言文本相似度计算流程如图3所示。

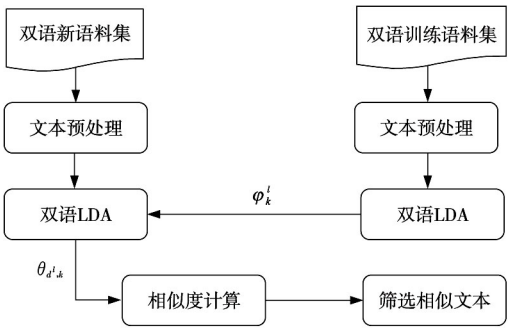


Figure 3 Cross-lingual text similarity calculation process

图3 跨语言文本相似度计算流程

3.1 相似度计算

常见的相似度计算有余弦相似度、Dice 系数、Jaccard 系数等,本文采用第一种方法计算相似度。三种方法计算公式如下:

余弦相似度:

$$sim(\boldsymbol{x},\boldsymbol{y})=\frac{\sum_{i=1}^{n}x_iy_i}{\sqrt{\sum_{i=1}^{n}x_i^2\sum_{i=1}^{n}y_i^2}}$$

Dice 系数:

$$sim(\boldsymbol{x},\boldsymbol{y})=\frac{2\sum_{i=1}^{n}x_iy_i}{\sum_{i=1}^{n}x_i^2+\sum_{i=1}^{n}y_i^2}$$

Jaccard 系数:

$$sim(\boldsymbol{x},\boldsymbol{y})=\frac{\sum_{i=1}^{n}x_iy_i}{\sum_{i=1}^{n}x_i^2+\sum_{i=1}^{n}y_i^2-\sum_{i=1}^{n}x_iy_i}$$

其中,$\boldsymbol{x}$,$\boldsymbol{y}$ 是空间里的两个向量。采用双语 LDA 会将双语语料集映射到同一个向量空间里,余弦相似度方法不会放大数据对重要部分的影响,只在方向上区分差异;而 Dice 系数相比余弦相似度是分

母由两向量长度平方的几何平均改为算数平均;Jaccard 在 Dice 系数基础上分子分母同时减去两向量的内积,取值都在[0,1]。余弦相似度方法在两个向量相差较大时效果要好于 Dice 方法,而 Jaccard 和 Dice 算法性质比较接近。故只选取余弦相似度作为本文的相似度计算方法。

采用余弦相似度计算文本相似度需要构建文本向量,设两个文本向量 $\boldsymbol{d}^{\mathrm{C}}(z_1^c, z_2^c, z_3^c, \cdots, z_K^c)$、$\boldsymbol{d}^{\mathrm{T}}(z_1^t, z_2^t, z_3^t, \cdots, z_K^t)$,分别表示源语言文档和目标语言文档,$z$ 表示主题,K 表示主题个数,我们需要为每个主题添加重要程度,重要程度的计算由权重计算表示。

计算源语言文档 d^{C} 和目标语言文档 d^{T} 的文档相似度表示成如下:

$$sim(\boldsymbol{d}^{\mathrm{C}}, \boldsymbol{d}^{\mathrm{T}}) = \cos(\boldsymbol{d}^{\mathrm{C}}, \boldsymbol{d}^{\mathrm{T}}) = \frac{\sum_{k=1}^{K} w_{d^{\mathrm{C}},k} w_{d^{\mathrm{T}},k}}{\sqrt{\sum_{k=1}^{K} w_{d^{\mathrm{C}},k}^2} \sqrt{\sum_{k=1}^{K} w_{d^{\mathrm{T}},k}^2}}$$

其中,$w_{d^{\mathrm{C}},k}$ 表示源语言文档 $\boldsymbol{d}^{\mathrm{C}}$ 主题 Z_k^{c} 的权重,$w_{d^{\mathrm{T}},k}$ 表示目标语言文档 $\boldsymbol{d}^{\mathrm{T}}$ 主题 Z_k^{T} 的权重,向量夹角越小,相似度值越大,$\boldsymbol{d}^{\mathrm{C}}$、$\boldsymbol{d}^{\mathrm{T}}$ 就越相似。

3.2　特征主题权重计算

特征词权重的常用方法是词频-逆文档频率,该方法很好地反映了一个词对特定文档的贡献程度,是广泛使用的一种有效的权重计算方法。所以,特征主题权重参考特征词权重计算方法,采用主题频率-逆文档频率 $TFIDF$(Topic Frequency Inverse Document Frequency)计算特征主题权重。

3.2.1　主题频率-逆文档频率

$$TFIDF = TF \times IDF = P(Z_k \mid d^l) \times \log \frac{|M|}{1 + |d^l : P(Z_k \mid d^l) > \lambda|} \quad (1)$$

其中,TF 是主题频率(Topic Frequency),即主题 Z_k 在文档 $\boldsymbol{d}^l$ 出现的频率,也就是 $P(Z_k \mid \boldsymbol{d}^l)$,$IDF$ 是逆文档频率,即 $\log \frac{|M|}{1 + |\boldsymbol{d}^l : P(Z_k \mid \boldsymbol{d}^l) > \lambda|}$,$|M|$代表文档集文档总数,$|\boldsymbol{d}^l : P(Z_k \mid \boldsymbol{d}^l) > \lambda|$ 代表含有主题 Z 的文档数,λ 为设定的常数,$P(Z_k \mid \boldsymbol{d}^l) = \theta_{d^l,k}$,该公式体现了主题对特定文档的贡献程度。

根据相似性文档定义,相似性文档一定同属于同一类,所以公式(1)存在一些不足,其类别内和类别间区分能力不强,不能给予代表类别的特征主题高权重,导致同类别的相似性文档匹配较弱。比如类别 C 中含有主题 Z 的文档较多,其他类别含的较少,说明主题 Z 能够代表类别 C,那应该赋予较高权重。而公式(1)并不会增加能够代表类别 C 的主题 Z 的权重,因为当 C 类主题 Z 增加时,$|\boldsymbol{d}^l : P(Z_k \mid \boldsymbol{d}^l) > \lambda|$ 也会增加,而 IDF 的值就会降低,同时,当特征主题在类别 C 中均匀分布也应该给予高权重,所以在此基础上,又改进了 $TFIDF$。

3.2.2　改进的主题频率-逆文档频率

从类别间和类别内的离散度角度出发,改进了 $TFIDF$ 得到 $ITFIDF$。

(1)公式(1)的 IDF 没有考虑特征主题在类间分布情况,首先对 IDF 进行改进,增加在一个类中出现较多的特征主题权重。改进的 IDF 为:

$$IDF = \log \frac{(|C_d^l : P(Z_k \mid \boldsymbol{d}^l) > \lambda| + 1) \times |M|}{1 + |\boldsymbol{d}^l : P(Z_k \mid \boldsymbol{d}^l) > \lambda|} \quad (2)$$

其中,$|C_d^l : P(Z_k \mid \boldsymbol{d}^l) > \lambda|$ 为 C 类中含有主题 Z 的文档数,$|C|$ 为 C 类中所有的文档数,设 $|\boldsymbol{d}^l : P(Z_k \mid \boldsymbol{d}^l) > \lambda| = |C_d^l : P(Z_k \mid \boldsymbol{d}^l) > \lambda| + |O|$,$|O|$ 代表其他类别含有主题 Z 的文档数,设:

$$f(|C_d^l : P(Z_k \mid \boldsymbol{d}^l) > \lambda|) = \frac{|C_d^l : P(Z_k \mid \boldsymbol{d}^l) > \lambda| + 1}{1 + |C^l : P(Z_k \mid \boldsymbol{d}^l) > \lambda| + |O|}$$

则随着 $|C_d^l : P(Z_k \mid \boldsymbol{d}^l) > \lambda|$ 增加,f 值也会增加,也就是提高了 IDF 值。公式(2)就考虑到了主题 Z 在类间的分布情况,增加了对该类别贡献程度,使其更好地代表该类别文档,加 1 是为了防止分子出现 0 的情况。

(2)另外 IDF 也没有考虑到特征主题在类内的分布情况。通过考察特征主题在类内的离散程度 D 分析特征主题在类内的分布情况,离散程度可以反映出特征主题在一个类中的分布情况,用主题概率分布的标准差表示,改进如下:

$$\overline{P'(Z_k \mid \boldsymbol{d}^l)} = \frac{1}{|C|} \sum_{i=1}^{|C|} P(Z_k \mid \boldsymbol{d}^{li}) \quad (3)$$

$$D = \sqrt{\frac{1}{|C|} \sum_{i=1}^{|C|} (P(Z_k \mid \boldsymbol{d}^{li}) - \overline{P'(Z_k \mid \boldsymbol{d}^l)})^2} \quad (4)$$

公式(3)表示主题 Z 在 C 类文档中的概率分布的平均值,公式(4)表示主题 Z 在 C 类文档中的分布的离散程度。公式(4)的值越低就代表主题 Z 在 C 类中分布越均匀,也就是说主题 Z 更能代表 C 类。由于类内离散程度与特征主题的分类能力成反比,因此类内修正公式用(1−D)代替,最终 $ITFIDF$ 为:

$$ITFIDF = P(Z_k \mid \boldsymbol{d}^l)\log \frac{(|C_d^l : P(Z_k \mid \boldsymbol{d}^l) > \lambda| + 1) \times |M|}{1 + |\boldsymbol{d}^l : P(Z_k \mid \boldsymbol{d}^l) > \lambda|}(1 - D) \quad (5)$$

经过改进后,公式(5)就增强了类间和类内的区别能力,使代表某个类的主题权重得到提高。

4 实验与分析

为了验证本文提出的方法有效性,选取了来自中国国际广播电台下的中国百科中泰对齐网页,涉及环境保护、教育、建筑、外交等 24 个类别,共 583 对,该语料集类别丰富,且为篇章对齐,符合本文的研究需要,所以选用该语料集作为本次实验的语料集。

本文所使用的语料集其文档类别分布如表 1 所示。

Table 1 Document category distribution(pair)

表 1 文档类别分布情况

类别	数量	类别	数量	类别	数量
1、国土、资源和人口	14	2、中国历史	30	3、中国戏曲	14
4、中国台湾	9	5、中国古典文学	15	6、政治	11
7、中国新疆与西藏	39	8、民间故事	66	9、网上博物馆	25
10、中国妇女	13	11、中国古代名人	26	12、中国民间美术	66
13、中国传统医学	15	14、中国民俗	18	15、中国世界遗产	14
16、中国乐器	28	17、中国音乐家	41	18、经济	28
19、对外交往	19	20、中国旅游	13	21、中国民族与宗教	14
22、中国建筑	15	23、中国教育与科技	28	24、中国环境保护	22

测试语料设置三组:($Te1$)第一组:8~9(类),($Te2$)第二组:13~17(类),($Te3$)第三组:20~24(类)。训练语料设置三组:($Tr1$)第一组:1~10(类),($Tr2$)第二组:11~19(类),($Tr3$)第三组:1~19(类)。

实验开始首先对文档集进行预处理,包括中泰文分词、去除中泰文停用词等;然后设置 LDA 训练参数:$\alpha = 50/K$,$\beta = 0.01$,$K = 200$,迭代次数为 1 000 次。经过实验 λ 设置为 0.03,阈值设置为 0.25。采用准确率 P、召回率 R、F 值通用评价方式考察性能,如下:

$$P = \frac{\sum P_i}{N}, P_i = \frac{C_p}{C}; R = \frac{\sum R_i}{N}, R_i = \frac{C_p}{Q}; F = \frac{2PR}{P + R}$$

其中,C_p 表示相似度值超过阈值且与源语言文档相似的文档的集合,C 表示所有相似度值超过阈值的文档集合,N 表示所有源语言文档数量,Q 表示所有与源语言文档相似的文档集合。

实验 1 为比较 $TFIDF$ 和 $ITFIDF$ 对相似度性能的影响,使用 $Tr1$、$Tr2$、$Tr3$ 训练集分别训练双语 LDA,$Te1$、$Te2$、$Te3$ 作为测试集语料,测试集用来测试采用不同权重计算方法的本文相似度算法中泰文匹配效果的,作用相当于图 3 中的新语料集,以采用了 $TFIDF$ 计算权重的本文方法为基线系统。表 2 是不同权重计算方法对中泰文匹配效果的影响。

Table 2 Chinese-Thai document match effect by different weight calculation methods

表 2 不同权重计算方式的中泰文匹配效果

	训练集	测试集	$TFIDF$			$ITFIDF$		
			P	R	F	P	R	F
中文匹配泰文	$Tr1$	$Te2$	0.75	0.52	0.61	0.78	0.67	0.72
	$Tr2$	$Te1$	0.72	0.51	0.60	0.76	0.66	0.71
	$Tr3$	$Te3$	0.77	0.53	0.63	0.81	0.69	0.75
泰文匹配中文	$Tr1$	$Te2$	0.76	0.53	0.62	0.79	0.70	0.74
	$Tr2$	$Te1$	0.73	0.52	0.61	0.77	0.68	0.72
	$Tr3$	$Te3$	0.79	0.55	0.65	0.83	0.72	0.77

实验结果表明,采用 $ITFIDF$ 计算权重比 $TFIDF$ 的相似度方法在召回率上有明显提升。原因在于 $ITFIDF$ 考虑到了类别间和类别内的主题分布情况,提高了能代表特定类别的主题权重;同时发现,当训练集文档增加时,准确率和召回率都有提升,说明算法在规模较大的语料集训练下会有更好的效果;训练集和测试集都是不同类别文章,在不同类别下的测试结果相差不大,说明算法对于类别敏感性不高,对文档类别不受限。

实验 2 为比较本文方法与现有方法的性能高低,本文方法使用 $Tr3$ 作为训练集,$Te3$ 作为测试集,权重计算选用 $ITFIDF$。选取有代表性的方法(文献[9]翻译单语主题模型方法)作为基线系统。表 3 是本文方法和文献[9]方法的中文匹配泰文的效果对比。

Table 3 Comparison between the proposed method and the translating monolingual topic model method

表 3 本文方法与翻译单语主题模型方法的效果对比

翻译单语主题模型方法			本文方法		
P	R	F	P	R	F
0.69	0.47	0.56	0.81	0.69	0.75

文献[9]通过翻译源语言单语主题模型至目标语言主题模型,翻译的准确率影响了相似文档匹配

性,而且利用余弦直接计算文档-主题相似度,没有突出不同文档的主题贡献以及不同类别文档的主题贡献,所以准确度不高。双语 LDA 可以将双语文档映射到同一主题特征空间内,无需借助翻译,同时使用了 *ITFIDF* 计算特征主题权重,除了突出主题对不同文档的贡献还突出对不同类别的贡献,克服了已有方法的不足,可靠性较好。

5　结束语

实验表明了本文所提算法的有效性。算法只需源语言和目标语言能分词,有一定的平行文本就可以做跨语言文本相似度计算。它降低了时间复杂度和维度,挖掘了文档潜在语义,无需双语知识和相关语言特征,没有未登录词问题。另外,*ITFIDF* 增强了主题对类别贡献程度,使得针对不同类型文档有更好的匹配性。由于主题模型只考虑文本的语义信息,是一种词袋模型,而文本的句法、语法、词义、篇章结构、上下文信息等均未考虑,使得双语互译文本识别准确率不高。下一步工作将考虑主题模型与文本的篇章结构、上下文信息以及词义等融合,加强双语互译文本的识别。

参考文献:

[1] Steinberger R, Pouliquen B, Hagman J. Cross-lingual document similarity calculation using the multilingual thesaurus EUROVOC[C]// Proc of the 3rd Conference on Computational Linguistics and Intelligent Text Processing, 2002: 415-424.

[2] He Wen-lei. Research on WordNet based Chinese-English cross language similarity measurement[D]. Shanghai: Shanghai Jiao Tong University, 2011. (in Chinese)

[3] Pottast M, Stein B, Anderka M. A wikipedia-based multilingual retrieval model[C]// Proc of European Conference on Advances in Information Retrieval, 2008: 522-530.

[4] Wang Hong-jun, Shi Shui-cai, Yu Shi-wen, et al. Cross-language similar document retrieval[J]. Journal of Chinese Information Processing, 2007, 21(1): 30-37. (in Chinese)

[5] Alberto B C, Paolo R, David P, et al. On cross-lingual plagiarism analysis using a statistical model[C]// Proc of ECAI 2008 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse(PAN08), 2008: 9-13.

[6] Uszkoreit J, Ponte J M, Popat A C, et al. Large scale parallel document mining for machine translation[C]// Proc of the 23rd International Conference on Computational Linguistics, 2010: 1101-1109.

[7] Hasan M M, Matsumoto Y. Multilingual document alignment-a study with Chinese and Japanese[C]// Proc of the 6th Natural Language Processsing Pacific Rim Symposium (NLPRS2001), 2001: 617-623.

[8] Maike E, Andrew F, Kotaro N. Calculating wikipedia article similarity using machine translation evaluation metrics[C]// Proc of Workshops of International Conference on Advanced Information Networking and Applications, 2011: 620-625.

[9] Preiss J. Identifying comparable corpora using LDA[C]// Proc of 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2012: 558-562.

[10] Ivan V, WimDe S, Moens M F. Identifying word translations from comparable corpora using latent topic models[C]// Proc of Annual Meeting of the Association for Computational Linguistics, 2011: 479-484.

[11] Wang Zhen-zhen, He Ming, Du Yong-ping. Text similarity computing based on topic model LDA[J]. Computer Science, 2013, 40(12): 229-232. (in Chinese)

[12] Sun Yuan, Zhao Qian. Tibetan-Chinese cross language text similarity calculation based on LDA topic model[J]. The Open Cybernetics & Systemics Journal, 2015, 9(1): 2911-2919.

[13] Ni X, Sun J T, Hu J, et al. Mining multilingual topics from wikipedia[C]// Proc of the 18th International Conference on World Wide Web, 2009: 1155-1156.

[14] Mimno D, Wallach H, Naradowsky J, et al. Polylingual Topic Models[C]// Proc of the EMNLP, 2009: 880-889.

[15] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet Allocation [J]. Journal of Machine Learning Research, 2003(3): 993-1022.

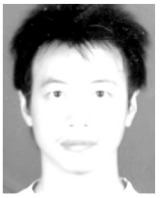
附中文参考文献:

[2] 何文垒. 基于 WordNet 的中英文跨语言文本相似度研究[D]. 上海:上海交通大学, 2011.

[4] 王洪俊,施水才,俞士汶,等. 跨语言相似文档检索[J]. 中文信息学报, 2007, 21(1): 30-37.

[11] 王振振,何明,杜永萍. 基于 LDA 主题模型的文本相似度计算[J]. 计算机科学, 2013, 40(12): 229-232.

作者简介:



程蔚(1989-),男,安徽马鞍山人,硕士生,研究方向为信息检索和自然语言处理。E-mail: chengwei9000@126.com

CHENG Wei, born in 1989, MS candidate, his research interests include information retrieval, and natural language processing.