2021 年 1 月 第 1 期
Vol. 42 No. 1 2021

融合主题及上下文特征的汉缅双语词汇抽取方法

李 越[1],毛存礼[1,2],余正涛[1,2],高盛祥[1,2],王振晗[1,2],张亚飞[1,2]

[1](昆明理工大学 信息工程与自动化学院,昆明 650500)
[2](昆明理工大学 云南省人工智能重点实验室,昆明 650500)
E-mail: maocunli@163.com

摘 要: 缅甸语属于低资源语言,网络中获取大规模的汉-缅双语词汇一定程度上可以缓解汉-缅机器翻译中面临句子级对齐语料匮乏的问题. 为此,本文提出了一种融合主题及上下文特征的汉缅双语词汇抽取方法. 首先利用 LDA 主题模型获取汉缅文档主题分布,并通过双语词向量表征将跨语言主题向量映射到共享的语义空间后抽取同一主题下相似度较高的词作为汉-缅双语候选词汇,然后基于 BERT 获取候选双语词汇相关上下文的词汇语义表征构建上下文向量,最后通过计算候选词的上下文向量的相似度对候选双语词汇进行加权得到质量更高的汉缅互译词汇. 实验结果表明,相对于基于双语词典的方法和基于双语 LDA + CBW 的方法,本文提出的方法准确率上分别提升了 11.07% 和 3.82%.

关 键 词: 汉缅双语词汇; 主题特征; 上下文特征; BERT; 双语词向量
中图分类号: TP311 **文献标识码:** A **文 章 编 号:** 1000-1220(2021)01-0091-05

Method of Chinese Burmese Bilingual Vocabulary Extraction Based on Subject and Context Features

LI Yue[1,2], MAO Cun-li[1,2], YU Zheng-tao[1,2], GAO Sheng-xiang[1,2], WANG Zhen-han[1,2], ZHANG Ya-fei[1,2]
[1](Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, China)
[2](Yunnan Key Laboratory of Artificial Intelligence, Kunming University of Science and Technology, Kunming 650500, China)

Abstract: Burmese is a low-resource language. Obtaining large-scale Chinese-Burmese bilingual vocabularies on the Internet, which can be mitigated due to the lacking of sentence-level alignment corpora in Chinese-Burmese machine translation. Consequently, this article proposes a method of Chinese-Burmese bilingual vocabulary extraction based on subject and context features. Firstly, the topic distribution of the Chinese-Burmese document is obtained by the LDA topic model, what's more the cross-language topic vector is mapped to the shared semantic space through bilingual word vector representation. The words with higher similarity under the same topic are extracted as Chinese-Burmese bilingual candidate vocabulary, we obtain the linguistic semantic representation of the context of the candidate bilingual vocabulary to construct a context vector by BERT. Finally we weight the candidate bilingual vocabulary by calculating the similarity of the context vector of the candidate word to obtain the higher quality Chinese-Myanmar translation Vocabulary. Experimental results show that compared with the method based on bilingual dictionary and the method based on bilingual LDA + CBW, the accuracy of the proposed method is improved by 11.07% and 3.82% respectively.

Key words: Chinese-Myanmar vocabulary; thematic features; contextual features; BERT; bilingual word vector

1 引 言

缅甸语属于一种资源稀缺型语言,汉-缅双语平行资源相对稀缺,但互联网中有一定规模的汉语-缅语双语资源,这些双语资源大多是主题相关,内容相似的可比文档. 汉-缅双语可比文档语料中存在一些具有互译关系的双语词汇数据,这些互译词语一般出现在语义相近但语言不同的上下文环境中. 抽取这些数据能有效改善汉、缅双语平行资源稀缺问题,进一步为开展跨语言检索研究[1,2]及机器翻译[3,4]提供资源支撑.

在先前的工作中,有研究者利用双语 LDA 和上下文向量组合的方法从可比语料中抽取双语词汇,取得不错的效果. 但对于资源稀缺的缅甸语来说,构建汉缅双语 LDA 需要大量标记好的双语平行语料,同时词袋模型表征的上下文向量没有考虑上下文语义和词语位置的影响,且维度较高.

在前人的基础上,为了获取具有上下文语义特征的上下文向量,克服汉缅双语 LDA 难以构建的问题. 本文提出了一种融合主题及上下文特征的汉缅双语词汇抽取方法: 本文首先利用单语 LDA 结合种子词典的方法抽取到具有主题特征的主题双语词汇,然后用多语言 BERT 对主题候选词的上下文语义进行

收稿日期: 2020-01-21 收修改稿日期: 2020-02-21 基金项目: 国家自然科学基金重点项目(61732005)资助; 国家自然科学基金项目(61662041,61761026, 6186019,61972186)资助; 云南省中青年学术和技术带头人后备人才项目(2019HB006)资助; 云南省自然科学基金重点项目(2019FA023)资助. 作者简介: 李 越,男,1995 年生,硕士研究生,研究方向为机器翻译; 毛存礼(通讯作者),男,1977 年生,博士,副教授,CCF 会员,研究方向为自然语言处理、机器翻译; 余正涛,男,1970 年生,博士,教授,博士生导师,CCF 高级会员,研究方向为自然语言处理、机器翻译; 高盛祥,女,1977 年生,博士,副教授,CCF 会员,研究方向为自然语言处理; 王振晗,男,1993 年生,博士研究生,CCF 会员,研究方向为机器翻译; 张亚飞,女,1981 年生,博士,讲师,CCF 会员,研究方向为机器翻译.

向量化表示,得到具有上下文语义特征的表示向量,再计算上下文的相似度得到具有上下文语义特征的双语词汇,最后与主题双语词汇加权组合得到更高质量的双语词汇.

2 相关工作

目前,针对从可比语料抽取双语词汇问题,主要有以下四类方法:

1) 基于双语词典的方法,其主要思想是通过一个种子词典学习到一个映射矩阵,将两种语言的词向量表示在同一语义空间中计算双语词向量的相似度抽取双语词汇,如,Artetxe[5,6]等人提出基于种子词典来抽取双语词汇,在大量的单语语料中训练表征成单语词向量,再通过种子词典学习到双语映射关系,将两种单语词向量映射到同一个语义空间,计算两种语言的词向量的相似度来抽取双语词汇. 但此类方法依赖于大规模且高质量的双语词典.

2) 基于枢轴语言的方法,其主要思想是将源语言和目标语言翻译成一种通用语言,在通用语言的语义空间中计算相似度抽取双语汇. 如,Kim 等人[7,8]提出一种基于枢轴语言抽取双语词汇的方法,首先将源语言转换为英语,再将目标语言转换到英语最后在同一语义空间下计算相似度完成双语词汇的抽取. 然而此类方法需要建立大规模对齐语料库,并且依赖于机器翻译的翻译效果.

3) 基于主题的方法,此类方法的思想是相似的词在跨语言文本当中具有相似的主题分布,如,Vulić等人[9,10]利用双语 LDA[11]将源语言跟目标语言的主题词归纳到同一主题空间下,通过计算源语言与目标语言的词-主题概率分布的相似性来抽取双语词汇. 此类方法需要构建跨语言双语 LDA,需要大量标记好的双语平行句对.

4) 基于上下文的方法,其主要思想是具有相似含义的词很可能出现在跨语言的相似上下文中. 如,从 Rapp 等人[12]开始,他们利用 Harris(1954)[13]提出的分布假设,提出了一种基于上下文的方法(CBM)抽取双语词汇,将跨语言词汇相似度计算问题转化为计算源语言和目标语言词汇对应的上下文向量的相似性来抽取双语词汇. 此类方法的缺点是忽略了词序关系对上下文向量的影响且容易出现高维问题.

3 融合主题及上下文特征的汉缅双语词汇抽取

我们提出的双语词汇抽取方法如图 1 所示,基本思路如下:

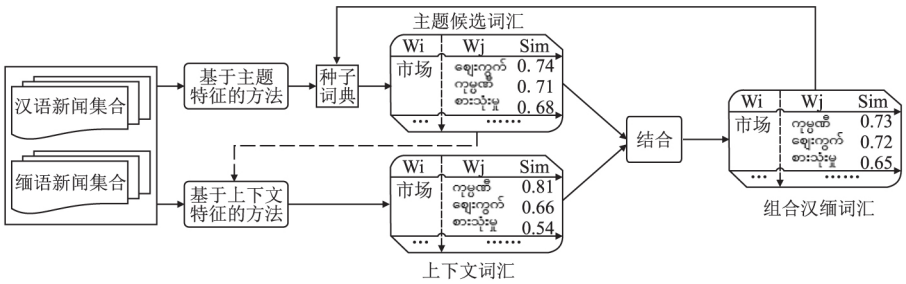

图 1 融合主题特征及上下文特征的汉缅双语词汇抽取架构

Fig. 1 Chinese-Burmese bilingual vocabulary extraction architecture integrating topic features and context features

1) 首先利用单语 LDA 模型抽取汉语、缅甸语文章的主题,然后将抽取到的主题词进行表征,得到汉缅单语主题词向量,再利用种子词典训练双语词向量将汉、缅主题词向量映射到同一语义空间,最后通过计算汉缅主题词向量的相似度抽取汉缅双语互译词汇. 我们称之为主题双语词汇. 主题双语词汇包含一个主题下的汉语词 w_i^S 的翻译候选缅甸词 w_j^T,其中列表中的缅甸词具有主题相似度得分 $Sim_{Topic}(w_i^S, w_j^T)$.

2) 然后基于多语言 BERT 表征主题双语词汇在文档中对应的上下文词汇的语义特征后生成主题词上下文特征向量,进而通过计算汉缅主题双语词汇的上下文特征向量的相似性对候选词汇进行加权,候选词汇上下文相似度评分记为 $Sim_{Context}(w_i^S, w_j^T)$.

3) 最后我们把主题双语词汇跟上下文双语词汇结合起来,得到对应的组合双语词汇 $Sim_{Comb}(w_i^S, w_j^T)$. 这样计算出来的双语词汇质量更高.

3.1 基于主题特征的汉缅双语候选词汇抽取

LDA(Latent Dirichlet Allocation)[14]是用来在一系列文档中发现抽象主题的一种统计模型. 换句话说就是在一篇文章中有一个中心思想,那么一定存在一些出现频率比较高的词. LDA 也是一种生成模型,一篇文章中每个词都是通过“以一定概率选择某个主题,并从这个主题中以一定概率选择某个词语”这样一个过程得到的. 每个主题下的主题词都服从一个多项式分布(Multinomial distribution). LDA 的概率图模型如图 2 所示.

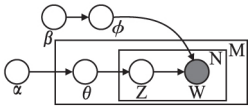

图 2 LDA 概率模型图

Fig. 2 LDA probability model

在图 2 中,α,β 分别是文本-主题和主题-词汇分布的先验参数. d 代表一篇文本. θ 为文本中主题-文档的概率分布参数,ϕ 则为每一个主题分布下词语的分布参数. Z 表示为其中一个主题,W 表示为一个词汇,M 表示文档总数,N 表示所有文档中的词的总数. 由于吉普斯采样具有简单、快速的特点,所以本文采用吉布斯采样[15]方法来训练. 假设我们从一组汉语词汇表 W^S 抽取出一个词 w_i,给定一个汉语主题分布 z_k. 则其词-主题概率分布为:

$$P(w_i|z_k) = \phi_{k,i} = \frac{n+\beta}{\sum_{j=1}^{|WS|} n + W^S\beta} \quad (1)$$

其中,n 代表主题分配到词汇表中的单词次数. $|W^S|$ 代表词汇表中不同单词的总数. $\sum_{j=1}^{|WS|} n$ 分配给主题的单词总数. 首先利用 LDA 主题模型从汉缅新闻篇章抽取到汉缅主题集合,得到每个主题下的词-概率分布,其次通过汉语篇章文本和缅甸语篇章文本抽取到的各自主题来训练汉语和缅甸语的主题词向量,分别令 x_i 表示汉语主题词的连续向量表示,z_j 表示缅语主题词的连续向量表示. 利用种子词典学习到映射矩阵 W,通过映射 Wx_i 到缅语语言空间,计算 Wx_i 与 z_j 向量之间的余弦相似度,如果汉缅双语词向量之间的相似性越高,那它们之间是互译词汇的准确率也越高. 本文采用余弦相似度计算汉缅双语词向量之间的相似度,计算公式如下所示:

$$Sim_{Topic}(wx_i, z_j) = \frac{\sum_{i=1}^{n} wx_i z_j}{\sqrt{\sum_{i=1}^{n}(wx_i)^2 \sum_{j=1}^{n}(z_j)^2}} \tag{2}$$

然后对上述相似度进行排序,选取前 N 个缅语作为汉语单词的候选翻译列表.

3.2　基于 BERT 的候选词汇的上下文表征

汉-缅双语可比文档中具有互译关系的词汇片段一般出现在具有双语词共现及主题相似的句子中,这些对齐词汇具有相似的上下文特征,并且互译词汇及上下文体现了相应的位置对应关系,为此,本文通过计算汉-缅双语词汇上下文特征向量的相似性实现汉-缅双语可比文档中具有互译关系的双语词汇抽取. 首先将汉语中的每个词 w_i^S(S = 汉语) 及缅语中的每个词汇 w_j^T(T = 缅语) 表征为窗口大小为 N 的词向量 v_i^S 和 v_j^T,然后利用汉-缅双语词典作为种子词典将汉语词汇上下文向量 v_i^S 中的每个词语翻译成缅语后得到用缅语表征的汉语词汇上下文向量 $v_i^{'S}$,并通过计算 $v_i^{'S}$ 和缅语中所有 v_j^T 之间的相似性实现汉-缅双语词汇抽取. 但由于利用汉-缅双语词典进行词汇翻译过程中会出现一词多义的歧义现象,而多语言的预训练模型 BERT[16] 是一种考虑词语位置关系以及上下文语义的一种模型,避免了一词多义的问题,即对于同一个词无论上下文的单词是否相同,训练出来的词向量都是一样的. 在此之前词向量本质是个静态方式,静态的意思就是说单词训练好之后就固定了,在以后使用时,单词不会跟着上下文场景变化而变化. 比如乔布斯是苹果的 CEO,他吃了一个苹果,用 word2vec[17,18] 表示“苹果”是一样的向量,而 BERT 模型可以解决此类问题.

基于此,本文采用 Google 开源的 BERT 模型来构造候选词汇的上下文特征表示,可以从候选词汇的前后单词中学习其上下文关系. BERT 的设计基于 Transformer[19] 网络结构. Transformer 对当前的输入,分别计算 Key,Query,Value 向量,并基于上述向量对每个输入使用注意力机制,以获得当前输入与上下文语义的关系和自身所包含的信息. 通过多层累加和多头注意力机制,不断获取当前输入更为合适的向量表示. 所以利用多语言 BERT 模型训练主题双语词汇能得到更好的上下文特征表示,设 S^i 为汉语主题词的上下文特征表示,T^j 为缅语主题词的上下文特征表示,则余弦相似度为:

$$\mathrm{Sim}_{Context}(w_i^S, w_j^T) = \frac{\sum_{i=1}^{n} S_k^i \times T_k^j}{\sqrt{\sum_{i=1}^{n}(S_k^i)^2 \sum_{j=1}^{n}(T_k^j)^2}} \tag{3}$$

一旦提取上下文双语词汇,我们将它们与主题双语词汇相结合. 结合后,词汇的质量将得到提高. 因此,我们进一步使用组合词汇作为新的种子词典,继而抽取到更好的汉缅双语词汇. 通过重复这些步骤,上下文双语词汇和组合双语词汇质量将被反复改进,直至模型收敛.

3.3　基于联合的方法抽取汉缅双语词汇

主题特征可以计算两个词在跨语言主题上的分布相似性,而上下文特征可以计算两个词在跨语言上下文上的分布相似性. 将这两种方式结合起来,可以利用全局知识和上下文知识来计算相似度,使双语词汇的抽取更加可靠和准确. 因此我们将 3.1 节中基于主题特征得到双语主题词汇 w_i^S, w_j^T 的主题相似度得分 $Sim_{Topic}(w_i^S, w_j^T)$ 与 3.2 节中利用 BERT 表征汉缅主题双语词汇的上下文特征向量的相似性对候选词汇进行加权得到双语词汇的上下文相似度评分 $Sim_{\mathrm{Context}}(w_i^S, w_j^T)$ 结合起来. 该组合是计算主题双语词汇与上下文双语词汇的综合相似性得分,定义如下:

$$Sim_{Comb}(w_i^S, w_j^T) = \lambda Sim_{Context}(w_i^S, w_j^T) + (1-\lambda) Sim_{Topic}(w_i^S, w_j^T) \tag{4}$$

其中 λ 是两种方法线性结合过程中的超参数. 我们首先使用主题特征的方法为汉语单词生成一个前 N 个候选列表(缅甸语候选词). 然后通过上下文特征向量计算候选列表词中的相似度. 最后,我们进行组合. 因此,组合过程是一次对基于主题特征抽取的候选词的重新排序实现汉缅双语词汇抽取.

4　实验结果与分析

4.1　实验数据跟参数设置

为了避免数据的单一性,我们分别从汉-缅双语网站、缅甸官方新闻网站、中文新闻网站、微信公众号等网络平台获取 778 篇汉-缅双语可比文档,覆盖了政治,军事,娱乐等多个方面,这些语料包括政治领域 271 篇,军事领域 296 篇,娱乐领域 211 篇,合计 778 篇. 其中汉语的平均句子长度为 23,缅语的平均句子长度为 18,如表 1 所示.

表 1　汉-缅双语可比文档数据集

Table 1　Chinese-Burmese comparable document data set

领域	政治	军事	娱乐	总计
汉语新闻	271	296	211	778
缅语新闻	271	296	211	778

接着我们对搜集到的语料进行预处理,利用昆明理工大学智能信息处理重点实验室研发的缅甸语分词工具对缅甸语进行分词,利用 jieba 分词工具对汉语进行分词,去除停用词等处理. 此外,通过人工方式构建了一个小规模的汉-缅双语种子词典,如表 2 所示.

表 2　训练汉-缅双语词向量的种子词典规模

Table 2　Seed dictionary scale for training Chinese-Burmese bilingual word vectors

训练词典规模	词典大小/KB
汉缅双语词典	327
汉语停用词表	15
缅语停用词表	8

LDA 模型中设置训练的超参数 $\alpha = 0.1$,$\beta = 0.1$,迭代次

数为 500 次,每篇文章的主题数为 5;词向量维度设置 300 维;对于我们提出的方法,我们根据经验设置线性组合参数 $\lambda = 0.8$.

4.2 实验方法和评价指标

为了验证本文方法在汉缅双语词汇抽取的效果,设计了 3 组对比实验.

对比实验 1. 本文与当前其他方法的对比实验

对比实验 2. 不同种子词典规模对词汇抽取的影响

对比实验 3. 在不同 P@N 值下词汇抽取的准确率

本文将准确率 P@N(前 N 个候选翻译的准确率)作为评价指标,定义如下:

$$\text{P@N} = \frac{\sum_{i=1}^{S} |T(w_i)|}{S} \tag{5}$$

其中,S 代表实验中对应的是测试词典中词的总数;w_i 代表测试词典中的源词,$|T(w_i)|$ 代表在测试词典中源词对应的目标词汇.

4.3 实验分析

实验 1. 当前的双语词汇抽取方法与本文方法实验结果比较.

表 3 本文方法与其他方法抽取双语词汇的准确率

Table 3 Accuracy of bilingual vocabulary extraction with this method and other methods

方法	P@1
基于双语词典的方法	38.65
基于枢轴语言的方法	36.45
基于双语 LDA 的方法	37.50
基于 CBW 的方法	33.75
基于双语 LDA + CBW 的方法	45.90
本文方法	**49.72**

由表 3 可知,我们提出的方法可以显著提高汉缅双语词汇的准确率. 实验结果也表明明显优于其他几种方法,同基于双语 LDA + CBW 的方法相比,本文方法准确率提升了 3.82%,主要原因在于 BERT 不仅仅是只关注一个词前文或后文的信息,而是整个模型的所有层都去关注其整个上下文的语境信息,得到更好的上下文特征表示向量. 同基于双语词典的方法和基于枢轴语言的方法相比,本文方法准确率分别提升了 11.07% 和 13.27%. 主要原因在于基于双语词典的方法未考虑到双语可比文档的主题特征对候选翻译的有效约束和基于枢轴语言的方法容易出现一词多译,错译等问题.

实验 2. 种子词典规模对抽取词汇效果的影响.

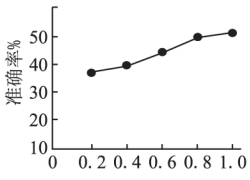

图 3 不同种子词典规模下的准确率

Fig. 3 Accuracy at different seed dictionary sizes

其次,种子词典是汉缅两种语义空间的中间桥梁,其规模大小对抽取的准确率也有着非常重要的影响. 我们将种子词典分成不同比例的规模大小,然后进行对比实验. 实验结果如图 3 所示. 从图 3 中可以看出,伴随着种子词典规模的扩大,抽取到的汉缅双语词汇准确率一直在逐渐上升. 当词典规模比例从 0.8 增加到 1 的时候,准确率上升的比较慢,主要原因是汉缅可比文档中,常见词已经得到补充,而生僻词的出现导致模型达到饱和.

实验 3. 为验证方法的准确率与抽取的候选词个数之间的关系,实验还比较了 P@1、P@5 和 P@10 的准确率. 具体实验结果见表 4.

表 4 本文方法在不同 P@N 值下的准确率

Table 4 Accuracy of this method under different P@N values

方法	P@1	P@5	P@10
本文方法	49.72	69.97	74.58

分析表 4 可知,本文方法的准确率随候选词的增多而逐渐上升,当候选词数量为 1 时便可获得较高的准确率,而当候选词为 10 时,准确率可以达到 74.58%,这同时说明了不同语言在向量空间上具有同构性.

5 总 结

为了抽取汉缅双语词汇,本文提出了一种融合主题及上下文特征的汉缅双语词汇抽取方法. 有效利用了汉缅双语主题的特征信息和上下文信息,进而抽取到质量更高的双语词汇. 实验结果表明,本文方法相比其他仅使用主题特征和上下文特征的方法相比,准确率有明显提升. 同基于双语 LDA + CBW 的方法相比,本文克服了汉缅双语 LDA 难以构建的问题,同时利用 BERT 训练得到具有上下文语义特征的上下文表示向量,进一步提升了汉缅双语词汇的准确率. 在未来的研究当中,我们可以将该方法用于其他稀缺语言中,如汉语-老挝语、柬埔寨等东南亚语言双语词汇抽取,为开展面向汉语-东南亚语跨语言检索及机器翻译研究提供数据支撑.

References:

[1] Grossman D A, Frieder O. Information retrieval: algorithms and heuristics [M]. Berlin: Springer Science & Business Media, 2012.

[2] Sadat F. Using comparable corpora to improve the effectiveness of cross-language information retrieval [C]//International Conference on Natural Language Processing, Springer, Berlin, Heidelberg, 2010: 320-331.

[3] Artetxe M, Labaka G, Agirre E. Bilingual lexicon induction through unsupervised machine translation [C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 2019: 5002-5007.

[4] Cao H, Zhao T, Zhang S, et al. A distribution-based model to learn bilingual word embeddings [C]//Proceedings of the 26th International Conference on Computational Linguistics, Osaka, Japan, 2016: 1818-1827.

[5] Artetxe M, Labaka G, Agirre E. Learning principled bi lingual mappings of word embeddings while presering monolingual invariance [C]//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, Texas, USA, 2016: 2289-2294.

[6] Artetxe M, Labaka G, Agirre E. Learning bilingual word embeddings with (almost) no bilingual data [C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguis-

tics, Vancouver, Canada, 2017: 451-462.

[7] Kim J H, Seo H W, Kwon H S. Bilingual lexicon induction through a pivot language [J]. Journal of the Korean Society of Marine Engineering, 2013, 37(3): 300-306.

[8] Jae-Hoon K, Hong-Seok K, Hyeong-Won S. Evaluating a pivot-based approach for bilingual lexicon extraction [J]. Computational Intelligence and Neuroscience, 2015, 12(1): 1-13.

[9] Vulić I, De Smet W, Moens M F. Identifying word translations from comparable corpora using latent topic models [C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers-Volume 2, Association for Computational Linguistics, Portland, Oregon, USA, 2011: 479-484.

[10] Vulić I, Moens M F. Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings [C]//Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, Queensland, Australia, 2015: 363-372.

[11] Mimno D, Wallach H M, Naradowsky J, et al. Polylingual topic models [C]//Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2, Association for Computational Linguistics, Suntec, Singapore, 2009: 880-889.

[12] Rapp R. Automatic identification of word translations from unrelated English and German corpora [C]//Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics, Association for Computational Linguistics, Maryland, USA, 1999: 519-526.

[13] Harris Z S. Distributional structure [J]. Word, 1954, 10(2-3): 146-162.

[14] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation [J]. Journal of Machine Learning Research, 2003, 3(1): 993-1022.

[15] Heinrich G. Parameter estimation for text analysis [R]. Technical Report, Austin, Texas, USA, 2005.

[16] Devlin J, Chang M W, Lee K, et al. Bert: pre-training of deep bidirectional transformers for language understanding [C]//Proceedings of Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minnesota: Association for Computational Linguistics, Florence, Italy, 2019: 4171-4186.

[17] Mikolov T, Chen K, Corrado G, et al. Efficient estimtion of word representations in vector space [C]//Proceedings of the ICLR Workshop, Scottsdale, Arizona, 2013: 154-162.

[18] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality [C]//Advances in Neural Information Processing Systems, Cambridge, Massachusetts, USA, 2013: 3111-3119.

[19] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [C]//Advances in Neural Information Processing Systems, Cambridge, Massachusetts, USA, 2017: 5998-6008.