基于跨语言词向量模型的蒙汉查询词扩展方法研究

马路佳　赵小兵　赖文

（中央民族大学 国家语言资源监测与研究少数民族语言中心，北京 100081）

摘要：跨语言信息检索指用户以一种语言提问，检索出另一种或几种语言描述的信息资源的检索技术，是信息检索领域重要的研究方向之一。近年来，跨语言词向量为跨语言信息处理提供了良好的表示形式，受到很多学者的关注。该文利用跨语言词向量实现从汉文查询词到蒙古文查询词扩展和映射，并利用该文提出的串联式查询扩展、串联式查询扩展过滤、交叉验证过滤三种查询扩展方法在进行词向量映射时对候选的蒙古文查询词进行筛选和排序，选择符合上下文的蒙古文词语。实验结果表明：在蒙汉跨语言信息检索任务中引入交叉验证方法对检索结果有很大的提升。

关键词：跨语言信息检索；跨语言词向量；查询扩展

Mongolian-Chinese Cross-Language Query Expansion Based on Cross-Language Word Vectors

Abstract: Cross-Language information retrieval is supposed to make information retrieval of one language according to the queries of other languages, and it is an important branch of information retrieval. Recently, cross-language word vectors perform effectively on cross-language information retrieval, and more and more researchers get interested in this field. This paper takes use of cross-language word vectors to map Chinese query words to Mongolian query words, which will be used to perform information retrieval. Three methods were proposed in this paper to perform mapping, namely Series, Series_opt and Cross_valid. These methods are used to map the Chinses queries as well as to select and sort the mapped words. Experimental results conducted in the real environments show that our proposed algorithm can obtain improvement on Mongolian-Chinses cross-language information retrieval.

Key words: Mongolian-Chinses cross-language information retrieval; Cross-Language word vectors; Query Expansion

---

收稿收稿日期：　　　　**定稿日期：**

基金项目：国家自然科学基金重点项目（61331013）

1　引言

随着网络技术的发展，信息检索已经成为人们充分利用各种信息资源不可或缺的工具。从最初的基于关键字匹配到现在的基于语义的分析、基于上下文的分析、以及应用各种统计方法进行分析等等，已经逐渐形成了一套比较完善的检索算法，并被学术界和工业界广泛应用。然而，随着网络的进一步发展及用户对查询的需求不断提高，单语言信息检索技术所表现出来的局限性越来越明显，人们已经不能满足于仅仅在同一种语言中进行检索，用户逐渐将需求转变为多语言的信息检索。

跨语言信息检索（Cross-Language Information Retrieval,CLIR）是信息检索领域一个重要的分支，由美国康奈尔大学 G．Salton 教授在 1973 年提出[1]，其主要作用是从目标语言表示的文档集中检索到跟源语言相关的文档[2]。跨语言信息检索在一般的信息检索基础上，允许用户使用一种语言来查询其他语言形式的信息，与传统信息检索相比较，目前所说的跨语言信息检索面临的一个关键问题就是检索时查询词使用的源语言与目标文档所使用目标语言需要统一为一种语言形式。目前的跨语言信息检索的方法主要有四种：查询词翻译的方法、文档翻译的方法、中间语言翻译方法和非翻译方法[3]。

查询翻译方法[4]是在信息检索之前，将查询的目标语种转化为翻译所要检索信息的语种，这种转化方式是目前实现跨语言信息检索的主流思想。这种方法的优点是：它可以和传统单语种信息检索技术紧密结合；仅对用户提交的查询进行语言翻译，工作量较小。这种方式的缺点是：检索返回的结果是用目标语言描述的，这对于不懂目标语言的用户来说，将会增加用户利用信息的难度。

文档翻译方法是在信息检索之前，将文档的信息语种转化为查询的目标语种。目前，实现文档翻译方法的技术主要有基于字典翻译文档索引词的方法和机器翻译系统（MTS）。文档翻译方法相比于查询翻译方法，其语境更加宽泛，能够充分利用上下文消除翻译的歧义性。但是文档翻译方法也存在着一些缺陷：它要求所有被检索的信息改变语种自有的符号；同时，现有的大多数机器翻译技术的翻译正确率还无法达到令人满意的程度；而且将数据库中全部文档从目标语种翻译到源语言语种代价昂贵，工作量巨大。目前这种方法在研究和实用上都远不如查询翻译。

中间语种翻译方法是将源语言和目标语言都转换为一种中间语言以实现跨语言信息检索的一种方法。在这种方法中，选择的中间语言应该是计算机容易自动处理和翻译准确率相对更高的语种，如英语等。在跨语言信息检索中会遇到源语种和目标语种之间无法进行直接翻译的问题，这时只能借助于中间语种将源语种和目标语种均翻译成中间语种。在这种情况下，使用中间语言翻译方法实现跨语言信息检索将会是一种不错的选择。

非翻译方法是不利用任何机器翻译方法即可实现跨语言信息检索的一种方法。这种方法目前主要是通过 Deerwester 等人 1990 年提出的浅层语义分析检索方法（LSI）来实现[5]。这种方法的优点是不会出现翻译系统中出现的未登录词无法翻译和歧义问题，但是这种方法对中间语义的提取比较困难。

Mikolov 等[6]首次提出不同的语言之间的词向量空间具有一定的相似性，通过映射源语言词向量到目标语言词向量可以实现“词翻译”，例如，英文词向量空间模型中“movie”对应的词向量和汉文词向量空间中“电影”对应的词向量的余弦距离是最接近的。跨语言词向量训练方式一般分为有监督（在训练过程中使用双语词典等）方式和无监督（在训练过程中不需要双语词典等）方式。最近，Facebook 提出的 MUSE 跨语言词向量训练方法[7]可以不依赖任何平行语料等先验知识，利用对抗学习（General Adversarial networks,GANs）和跨领域相似度局部缩放（cross-domain similarity local scaling, CSLS）等方法来获得跨语言词向量模型，该方法在词翻译等任务上实验的效果对比其他方法要好，很多情况下甚至比监督的方法效果还好。

由于现有的高质量蒙汉平行语料数量较少，训练出优良的蒙汉机器翻译模型尚存在一定的困难，并且利用机器翻译方法来实现跨语言信息检索时要考虑到存储空间和系统的可扩展性要求；同时，未登录词、消岐等方面问题也会在很大程度上影响信息检索的

效率。

本文采用非翻译的方法实现跨语言信息检索，即基于跨语言词向量模型实现语言统一和查询扩展目标，利用 MUSE 跨语言词向量方法以监督方法训练蒙汉跨语言词向量并实现蒙汉跨语言信息检索。该方法在查询前将用户的汉文查询词映射为蒙古文，然后利用映射后的蒙古文进行信息检索，实验表明这种方法切实有效。与文档翻译的方式形成鲜明的对比，本文使用 MUSE 的监督方式来训练蒙汉跨语言词向量，使用了蒙汉词典，但是不需要平行句对，存储空间小，适用性更强；与基于双语词典的方式比较，该方法在一定的程度上改善了未登录词的影响，提升了跨语言信息检索的召回率。

2 相关研究

跨语言词向量(corss-lingual word embeddings)是一种对单语言环境下的模型进行多语言扩展的有效手段。通过平行语料得到不同语种之间词向量的关联，使用这种关联关系实现了跨语言信息扩展的任务。近年来，越来越多的学者将目光转移到跨语言词向量的相关研究上来，大致原因由两个。首先，跨语言词向量可以在多语言环境中推断词语的语义；其次，跨语言词向量可以实现不同语言之间的知识迁移，可以计算多任务语言之间的相关性。

跨语言词向量由 Klementiev 等人[8]在 2012 年首次提出，首先借助神经网络语言模型构建初始词向量，然后利用对齐语料的词共现特征构建跨语言词向量。此后，越来越多的学者将目光转向跨语言词向量研究，并提出不同的学习模型。Faruqui[9]等基于词汇语义内容在语言之间的不变性特征，提出一种基于典型相关性分析的简单技术，并将多语言的特征并入单语言的生成向量中，该方法相比于单语言技术也表现出更好的语义表示性能。但是由于采用串行级联形式，该方法难以同时学习单语言和跨语言的嵌入表示。Chandar A P[10]等使用基于自编码器的方法来实现跨语言的两种语言之间的相关性词向量表示，通过简单的学习在不同语言之间去重建句子级别的词袋表示，可以得到更高的性能，并且不需要词语对齐。这种方式对于句子级别的表达具有较高的性能，但缺乏对词之间的语义表达。此后，很多学者设计不同的目标函数来提升跨语言词向量的性能[11-13]。2015 年后，越来越多的学者们分别设计不同的算法对语料进行随机词混合[14-16]，将得到的混合语料作为训练数据，并将跨语言词嵌入转化为单一语言词嵌入，也得到了较好的效果。

蒙古语信息检索相关研究起步较晚，巩文靖[17]提出采用词相关性扩展、加入距离模型的扩展以及关联词与词对共现距离相结合的扩展方法进行汉蒙信息检索的查询扩展。

目前，跨语言词嵌入仍然是表示学习的一个研究热点问题，并开始逐渐向多语言、多粒度、多功能的方向发展，在跨语言文档分类、跨语言情感分类、跨语言相似度计算、机器翻译、跨语言句法分析等领域得到了广泛的应用。

3 本文提出的方法

蒙汉跨语言信息检索可以采用跨语言词向量扩展和文档翻译两种方法实现，本文提出的跨语言查询扩展模式除了在查询效果方面表现优异之外，在模型的可扩展性、数据冗余性，以及存储空间的消耗方面都表现出较为明显的优势：

① 使用翻译技术，每个文档都需要存储两份，并且随着时间的推移，后续文档需要不断的进行翻译，这就会积累越来越多的冗余数据，而跨语言词向量文件只需在每次训练完成后提供跨语言检索使用，占用的资源少。

② 基于跨语言词向量查询扩展时，由于它依赖的资源比较少(只依赖于两个词向量文件)，那么最终的信息检索系统结构规模小、冗余数据少，且易于扩展(只需要替换两种词向量文件即可)，这很容易实现从一种跨语言对的查询(例如，汉文到蒙古文)扩展移

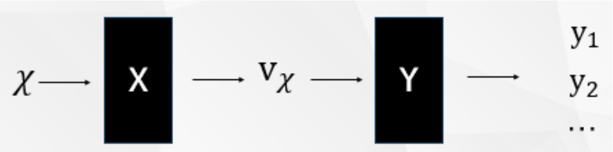
植到其他语言对的检索(例如，汉文到维吾尔文)。

本文提出使用串联式查询扩展(记为 Series)、串联式查询扩展过滤(记为 Series_opt)、交叉验证筛选扩展（记为 Cross_valid）三种策略进行蒙汉跨语言的词向量查询扩展。

3. 1 串联式查询扩展

该策略中对汉文查询式中每一个词进行跨语言词向量扩展，获得蒙古文查询词，然后将所有扩展所得的蒙古文查询词按照汉文查询式中的查询词先后关系串联起来。本文中使用的跨语言词向量映射方式如下图 1 所示：

图 1 跨语言词向量映射方式



上图中 X 和 Y 分别代表源语言（汉文）词向量空间和目标语言（蒙古文）词向量空间，χ 是一个汉文查询词，v_χ 为χ 在 X 中的词向量表示，通过在词向量空间 Y 中找到与 v_χ 余弦距离最近的 k 个词作为其候选扩展词，即上图中的y_1，y_2等词会作为查询词 χ 的候选扩展词。

在串联式查询扩展方法中，利用蒙汉跨语言词向量直接进行映射，因为一个汉文词语翻译为蒙古文时通常会对应多个蒙古文词语，所以本文中选择了为每个汉文词语映射 2 个蒙古文词语。对于汉文查询语句“网络安全”来说，它经过分词之后成为“网络”和“安全”两个词语。假设每个汉文词语扩展两个蒙古文词语， 例如“网络”对应的蒙古文词语为“[illegible]”“ [illegible]”，“安全”对应的蒙古文查询词为“[illegible] ” “[illegible]”，最终的蒙古文查询式为这四个蒙古文词语的串联，即“[illegible] [illegible] [illegible] [illegible]”。

3. 2 串联式查询扩展过滤

对于已分词的汉文查询式，如“传统-文化”(其蒙古文为“[illegible] [illegible]”)，按照 k=2 进行扩展得到的蒙古文查询式“[illegible] [illegible] [illegible] [illegible]”，除了扩展出“[illegible]”和“[illegible]”两个词之外还有其他的词。这种冗余的情况会导致查询结果质量下降。所以本策略中在其查询扩展时将按照跨语言词向量相似度值去进行过滤操作。上例中，串联式扩展中，将对所有的汉文词语进行 k(k=2)个蒙古文词语，而如果一个汉文词向量与一个蒙古文词向量的相似度（本文使用 cosine 距离）超过一定阈值（0.51），那么该蒙古文极有可能是该汉文词对应的翻译词汇，那么另一个就可能是冗余的。

3. 3 交叉验证筛选

该策略考虑到由于一个汉文词语在经过跨语言词向量映射的时候，对应的蒙古文扩展词有多个，那么如何对扩展词进行筛选和排序就成为一个值得考虑的问题，所以我们在对一个汉文查询词进行扩展的时候，除了考虑该汉文词与它的蒙古文候选词之间的相似度，还要考虑到候选扩展词跟其他汉文查询词之间的相似度，以该汉文周围的词作为它的上下文对这些蒙古文短语进行筛选和排序，从而可以从这些蒙古文短语中选择符合当前汉文查询短语上下文的一个，这可以看做一种消岐的方法：

给定查询式 $\mathrm{Q}=\{q_1, q_2, \ldots, q_n\}$，$\vec{q}_i$表示某个查询词 q_i 对应的词向量，$C_i=\{q_{i1}, q_{i2}, \ldots, q_{im}\}$ 为查询词 q_i 对应的 m 个目标语言（蒙古文）候选查询词集合，交叉筛选扩展策略就是通过计算集合中扩展词对于其他查询词的的整体相似度（这里使用 score 表示）

从而从该集合中选出 n 个（本文 n=1）个 score 最高的词作为最终 q_i 的扩展词。本文中当汉文查询式中只有一个词的时候，则选择相似度大于阈值时选择 1 个，否则选择 2 个蒙古文词语作为其最终扩展词。

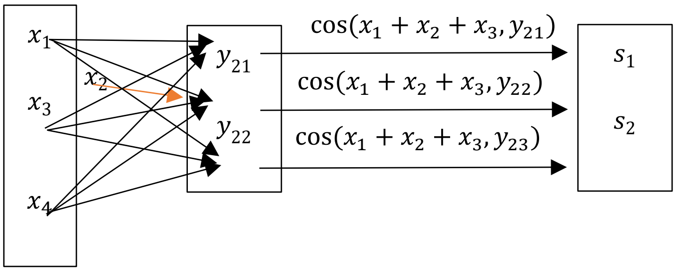
对于q_{il}来说 $\overrightarrow{q_{il}}$ 是它的词向量，那么：

$$score_{il} = \sum_{o=1,o\neq i}^{m} \cos(\overrightarrow{q_{il}}, \overrightarrow{q_o})$$

$$= \cos(\overrightarrow{q_{il}}, \sum_{o=1,o\neq i}^{m} \overrightarrow{q_o})$$

该方法的跨语言词向量映射如图 2 所示：

图 2 交叉验证筛选映射方式



上图中x_1，x_2，x_3，x_4 为汉文查询词，现在考虑从 x_2 的候选扩展词 y_{21}，y_{22}，y_{23} 中选择其最终的扩展词。该方法通过计算每个候选词和x_1，x_2，x_3，x_4 整体的相似度来对这些候选词进行排序，例如 y_{21} 对应与查询式的相似度值为$cos(x_1+x_2+x_3,\ y_{21})$，$y_{21}$，$y_{22}$，$y_{23}$ 对应的结果分别为S_1，S_2，S_3，最后选择最大相似度值对应的候选扩展词作为 x_2 最终的扩展词。

4 实验

4.1 实验数据与准备

本文研究文档库中一共有 28166 篇蒙古文文档，所有蒙古文文档均采用 Unicode 编码。实验使用 21 个蒙古文查询短语，这些查询式来自于训练跨语言词向量时所用训练集词对的高频词，这些扩展式不包含在训练跨语言词向量时所使用的词典中，平均每个查询短语对应 300 个候选文档进行实验验证。

实验中使用的查询式与巩文靖等人[17]相同，如表 1 所示。

表 1 汉文查询式和跨语言词查询扩展后获得的对应的蒙古文查询式

汉文查询式	蒙汉文查询式	汉文查询式	蒙汉文查询式
新闻	ᠰᠣᠨᠢᠨ ᠮᠡᠳᠡᠭᠡ	事业单位	[illegible]
资金	[illegible]	民族文化	ᠦᠨᠳᠦᠰᠦᠲᠡᠨ ᠦ ᠰᠣᠶᠣᠯ
皇帝	[illegible]	制度改革	[illegible]
法律	[illegible]	农业生产	[illegible]
艺术	[illegible]	食品安全	[illegible]
纪律检查	[illegible]	民族教育	ᠦᠨᠳᠦᠰᠦᠲᠡᠨ ᠦ ᠰᠤᠷᠭᠠᠨ ᠬᠦᠮᠦᠵᠢᠯ
传统文化	[illegible]	服务机构	[illegible]
新闻媒体	ᠰᠣᠨᠢᠨ ᠮᠡᠳᠡᠭᠡ [illegible]	养老保险	[illegible]
机构改革	[illegible]	小康社会	[illegible]
社会发展	[illegible]	体育	[illegible]
宗教	[illegible]		

在训练跨语言词向量时需要先训练得到两种语言各自的单语言词向量，本文使用 fastText 来训练蒙古文和汉文的单语言词向量，其中汉文语料来源于中文维基百科，大小为 1.1G，蒙古文语料来自爬取的蒙古文网页数据及 CCWMT2017 中的蒙古文数据，大小为 329MB。相关训练词向量参数为：

--minn：最短子串长度，2 种语言都指定为 2；--maxn：最长子串长度，蒙古文指定为 15，中文指定为 4；--dim：词向量维度 300 维。

训练蒙汉跨语言词向量时训练集词典大小为 3224，测试集词典大小 7732，汉文词不唯一。

训练跨语言词向量时需要使用前面训练得到的蒙古文单语言词向量和汉文单语言词向量，跨语言词向量训练模型使用的为 MUSE[7]，其大致工作方式为：假设 X 和 Y 表示两种语言的词向量空间，W 为转移矩阵（rotation matrix），MUSE 的最终目的是系统通过对抗学习（adversarial learning）和 CSLS（CROSS-DOMAIN SIMILARITY LOCAL SCALING）等一系列方式学习得 W，并最终将词向量空间 WX 与 Y 对齐，主要步骤如下：

1） 利用对抗学习（adversarial learning ）得到转移矩阵 W，达到初步对齐的效果。选择一部分词来测试鉴别器（Discriminator），检测它是否能够区分出来自两个词向量空间的两个词向量。

2） 转移矩阵 W 进行优化。频率高的词会被选作锚点（anchor point），通过缩小这些锚点之间的距离来优化 WX 和 Y 两个向量空间的对齐效果。

3） 使用 CSLS 来提高跨语言词向量映射效果，并结合 CSLS 和 W 来对 X 进行映射。

最终生成的跨语言词向量模型中蒙古文和汉文词向量模型大小分别为：167MB 和 698MB。

4.2 实验结果与对比分析

本文使用平均精度均值(Mean Average Precession，MAP)作为信息检索的评价指标。设 p_i为第 i 个查询式的平均计算精度，一共有 n 个查询式，那么最终的 MAP 值为：

$$MAP = \frac{1}{n}\sum_{i=1}^{n} P_i$$

4.2.1 串联式查询扩展

一般情况下汉文对应的蒙古文词语的数量都会是大于 1 个，所以使用串联式查询扩展策略时，选择对每个汉文词扩展 2 个蒙古文词语，其查询扩展结果如下表 2 所示：

表 2 串联式查询扩展式

汉文查询式	扩展查询式	汉文查询式	扩展查询式
新闻	[illegible] / [illegible]	事业+单位	[illegible] / [illegible]+[illegible] / [illegible]
资金	[illegible] / [illegible]	民族+文化	[illegible] / [illegible] + [illegible] / [illegible]
皇帝	[illegible] / [illegible]	制度+改革	[illegible] / [illegible] + [illegible] / [illegible]
法律	[illegible] / [illegible]	农业+生产	[illegible] / [illegible] + [illegible] / [illegible]
艺术	[illegible] / [illegible]	食品+安全	[illegible] / [illegible] + [illegible] / [illegible]
纪律+检查	[illegible] / [illegible] + [illegible] / [illegible]	民族+教育	[illegible] / [illegible] + [illegible] / [illegible]
传统+文化	[illegible] / [illegible] + [illegible] / [illegible]	服务+机构	[illegible] / [illegible] + [illegible] / [illegible]
新闻+媒体	[illegible] / [illegible] + [illegible] / [illegible]	养老+保险	[illegible] / [illegible] + [illegible] / [illegible]
机构+改革	[illegible] / [illegible] + [illegible] / [illegible]	小康+社会	[illegible] / [illegible] + [illegible] / [illegible]

社会+发展	[illegible] / [illegible] + [illegible] / [illegible]	体育	[illegible] / [illegible]
宗教	[illegible] / [illegible]		

最终结果（MAP 值）： MAP_{series}= 0.4405，准确率：0.7437， 查全率：0.6927。

从上面的蒙古文查询扩展式跟原蒙汉文查询式对应比较可以看出，这种方式没有考虑到扩展时相似度很高的情况，这个时候接近于词翻译效果，所以就会出现扩展冗余的情况。例如在对“民族”和“文化”进行扩展的时候，上面的结果已经显示出扩展结果除了包含“[illegible]”和“[illegible]”之外，还包含其它的词，这些词导致了检索性能的下降，所以下面将根据跨语言词向量相似度对这些可能冗余的词进行过滤。

4.2.2 串联式查询扩展过滤

用于过滤查询扩展词的跨语言词向量相似度阈值实际上是一个观测值或者是经验值，这跟词向量的训练结果很相关。当阈值为 0.51 时，所得的蒙古文查询式的结果如下表 3 所示：

表 3 串联式查询扩展过滤后的查询式

汉文查询式	扩展查询式	汉文查询式	扩展查询式
新闻	[illegible]	事业+单位	[illegible] + [illegible]
资金	[illegible]	民族+文化	[illegible] + [illegible]
皇帝	[illegible]	制度+改革	[illegible] + [illegible] / [illegible]
法律	[illegible]	农业+生产	[illegible] + [illegible]
艺术	[illegible]	食品+安全	[illegible] + [illegible]
纪律+检查	[illegible] + [illegible] / [illegible]	民族+教育	[illegible] + [illegible]
传统+文化	[illegible] + [illegible]	服务+机构	[illegible] + [illegible]
新闻+媒体	[illegible] + [illegible]	养老+保险	[illegible] / [illegible] + [illegible] / [illegible]
机构+改革	[illegible] + [illegible] / [illegible]	小康+社会	[illegible] / [illegible] + [illegible]
社会+发展	[illegible] + [illegible]	体育	[illegible]
宗教	[illegible]		

最终结果（MAP 值）： MAP_{series_opt}= 0.6262，准确率：0. 8097， 查全率：0.7819。

上面的结果表明，通过去除冗余的查询扩展词可以明显地提升系统的检索效果。通过与表 4 对比可以发现，表 5 中有些汉文查询式的扩展查询式得到了明显“裁剪”，以此消除冗余的查询词，例如“传统文化”的扩展词由“[illegible] [illegible] [illegible] [illegible]”变为“[illegible] [illegible]”。对于某些中文查询词来说，对于当前上下文中其蒙古文扩展词的默认排序不太准确，例如“改革”的查询扩展词为“[illegible] [illegible]”，但是在“机构改革”中“改革”的扩展词顺序为“[illegible] [illegible]”比较合理。所以，考虑到上下文语境信息使用“交叉验证筛选扩展”的方法再次对扩展查询式进行优化。

4.2.3 交叉验证筛选

交叉验证筛选扩展主要通过一个汉文查询的词的上下文来对其蒙古文查询式进行排序和筛选合适的扩展查询词，所得的蒙古文查询式的结果如下表 4 所示：

表 4 交叉验证筛选扩展查询式

汉文查询式	扩展查询式	汉文查询式	扩展查询式
新闻	[illegible]	事业+单位	[illegible] + [illegible]
资金	[illegible]	民族+文化	[illegible] + [illegible]
皇帝	[illegible]	制度+改革	[illegible] + [illegible]
法律	[illegible]	农业+生产	[illegible] + [illegible]
艺术	[illegible]	食品+安全	[illegible] + [illegible]

纪律+检查	[illegible] + [illegible]	民族+教育	[illegible] + [illegible]
传统+文化	[illegible] + [illegible]	服务+机构	[illegible] + [illegible]
新闻+媒体	[illegible] + [illegible]	养老+保险	[illegible] + [illegible]
机构+改革	[illegible] + [illegible]	小康+社会	[illegible] + [illegible]
社会+发展	[illegible] + [illegible]	体育	[illegible]
宗教	[illegible]		

最终结果（MAP 值）：MAP_{cross_valid} =0.7068，准确率：0.8519， 查全率：0.8187。

通过与表 5 比较，表 6 中一些汉文查询词在进行跨语言词向量映射时通过该词的上下文对候选词进行调序和筛选，这些汉文查询式对应的蒙古文查询式发生了变化，例如“制度改革”中的“改革”的前 3 个扩展词为“[illegible] [illegible] [illegible]”，而经过调序之后“[illegible]”被调到了第一位；“小康”的前 3 个扩展词最初为“[illegible] [illegible] [illegible]”，而调序之后“[illegible]”调序到了第一位。这几个词的调序属于对于低于相似度阈值的词进行调序，而“教育”与其扩展词的相似度是在阈值之上的。“教育”的前 3 个扩展词及其相似度分别为“[illegible] 0.6912”、“[illegible] 0.6579”、“[illegible] 0.6172”，通过公式 3-1 重新打分之后三者的分值分别为：0.2493、0.2692、0.2208，“[illegible]”得分最高。

3.2.4 3 种查询扩展方式对比

本文以基于机器翻译的方法为基准，分别用本文所提出的三种跨语言信息检索方法进行对比，结果如表 5 所示：

表 5 四种跨语言方法结果比较

跨语言方法	MAP 值	查准率	查全率
文档翻译	0.641	0.837	0.717
串联式查询扩展	0.440	0.744	0.693
串联式查询扩展过滤	0.626	0.810	0.782
交叉验证筛选	0.707*	0.852*	0.819*

对于串联式查询扩展过滤方法，通过去除冗余的查询扩展词可以明显地提升系统的检索效果。对于交叉验证筛选扩展方法，主要通过一个汉文查询的词的上下文来对其蒙古文查询式进行排序和筛选符合当前查询式上下文的扩展查询词，效果再次得到了明显的提升，其 MAP 值超过了基准测试的 MAP 值 0.6012，说明了当前方法的有效性。

另外，对于“十九大”以及“一带一路”，“四风”这样的词，他们本身整体表示一个含义，而在分词的时候对它们进行切分后的词（例如，“四风”被分词工具分为“四”和“风”），这个时候分词后与分此前的含义完全无关。在面对这些词时，文中的方法不在适用，结果不理想，查询结果如表 6 所示。

表 6 非叠加词结果

汉文查询式	蒙古文查询式	查询扩展式
一带一路	[illegible] [illegible] [illegible] [illegible]	[illegible] [illegible]
四风	[illegible] [illegible]	[illegible] [illegible]
十九大	[illegible] [illegible] [illegible] [illegible]	[illegible] [illegible]

对于上面 3 个查询式，结果为 MAP：0.042，准确率：0.215，查全率：0.211.

4 结论

随着跨语言词向量技术的发展，它成为除了双语词典、机器翻译技术之外又一项有力工具。与机器翻译的技术手段相比，MUSE 不需要平行句对，对于监督的跨语言词向量训练方式，只需要一定数量的词典就可达到可用的效果，并且利用这个方式实现的跨语言信息检索系统容易扩展到其它语言对中；基于词典翻译的方式则受限于词典的规模

来说。本文中提出了串联式扩展、串联式查询扩展过滤以及交叉验证筛选查询扩展 3 种蒙汉跨语言词向量映射方案，使得在根据汉文查询词进行选择蒙古文查询词时能够选出较符合上下文的蒙古文词语，通过实验证明了这些方法的有效性。

参考文献

[1] Oard D W. Alternative approaches for cross-language text retrieval[C]//AAAI Symposium on Cross-Language Text and Speech Retrieval. American Association for Artificial Intelligence. 1997, 16.

[2] 麦淑平. 跨语言信息检索技术探析[J]. 中华医学图书情报杂志. 2008.

[3] 黄国斌，王明文，叶浩．一种新的基于中间语义的跨语言信息检索模型[J]．中文信息学报，2009, 23(2): 77r82.

[4] Gao J, Nie J Y, Xun E, et al. Improving query translation for cross-language information retrieval using statistical models[C]//Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2001: 96-104.

[5] 金千里，赵军，徐波．弱指导的统计隐含语义分析及其在跨语言信息检索中的应用[C]//见：语言计算与基于内容的文本处理——全国第七届计算语言学联合学术会议论文集．北京：清华大学. 2003: 527-533.

[6] Mikolov T, Le Q V, Sutskever I. Exploiting Similarities among Languages for Machine Translation[J]. Computer Science, 2013.

[7] Conneau A, Lample G, Ranzato M, et al. Word Translation Without Parallel Data[J]. 2017.

[8] Bhattarai B, Klementiev A, Titov I, et al. Inducing Crosslingual Distributed Representations of Words[C]//COLING. 2012.

[9] Faruqui M, Dyer C. Improving vector space word representations using multilingual correlation[C]//Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics. 2014: 462-471.

[10] AP S C, Lauly S, Larochelle H, et al. An autoencoder approach to learning bilingual word representations[C]//Advances in Neural Information Processing Systems. 2014: 1853-1861.

[11] Gouws S, Bengio Y, Corrado G. Bilbowa: Fast bilingual distributed representations without word aligments[C]//International Conference on Machine Learning. 2015: 748-756.

[12] Soyer H, Stenetorp P, Aizawa A. Leveraging monolingual data for crosslingual compositional word representations[J]. arXiv preprint arXiv:1412.6334, 2014.

[13] Shi T, Liu Z, Liu Y, et al. Learning cross-lingual word embeddings via matrix co-factorization[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). 2015, 2: 567-572.

[14] Vulić I, Moens M F. Bilingual word embeddings from non-parallel document-aligned data applied to bilingual lexicon induction[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). 2015, 2: 719-725.

[15] Gouws S, Søgaard A. Simple task-specific bilingual word embeddings[C]//Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2015: 1386-1390.

[16] Coulmance J, Marty J M, Wenzek G, et al. Trans-gram, fast cross-lingual word-embeddings[J]. arXiv preprint arXiv:1601.02502, 2016.

[17] 巩文婧．基于语言模型的跨汉蒙信息检索技术研究[D]．内蒙古大学, 2012.

作者联系方式：

1. 马路佳，中央民族大学国家语言资源监测与研究少数民族语言中心，yudianer1991@hotmail.com

2. 赵小兵，中央民族大学国家语言资源监测与研究少数民族语言中心，nmzxb_cn@163.com

3. 赖文，中央民族大学国家语言资源监测与研究少数民族语言中心，Lavine_Lai@126.com