基于双语主题模型和双语词向量的跨语言知识链接

余圆圆[1] 巢文涵[1] 何跃鹰[2] 李舟军[1]
(北京航空航天大学计算机学院 北京 100191)[1] (国家计算机网络应急技术处理协调中心 北京 100029)[2]

摘 要 跨语言知识链接是指在描述相同内容的不同语言的在线百科文章之间建立联系。跨语言知识链接可分为候选集选择和候选集排序两部分。首先,把候选集选择问题转换为跨语言信息检索问题,提出一种将标题与关键词相结合从而生成查询的方法,该方法将候选集选择的召回率大幅提高至93.8%;在候选集排序部分,提出一种融合双语主题模型及双语词向量的排序模型,实现了英文维基百科和中文百度百科之间军事领域的跨语言知识链接。实验结果表明,该模型取得了75%的准确率,显著提高了跨语言知识链接的性能,并且提出的方法不依赖于语言特性和领域特性,因此可以很容易地扩展至其他语言和其他领域的跨语言知识链接。

关键词 跨语言知识链接,跨语言信息检索,双语主题模型,双语词向量
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2019.01.037

Cross-language Knowledge Linking Based on Bilingual Topic Model and Bilingual Embedding

YU Yuan-yuan[1] CHAO Wen-han[1] HE Yue-ying[2] LI Zhou-jun[1]
(School of Computer Science and Engineering, Beihang University, Beijing 100191, China)[1]
(National Computer Network Emergency Response Technical Team/Coordination Center, Beijing 100029, China)[2]

Abstract Cross-language knowledge linking (CLKL) refers to the establishment of links between encyclopedia articles in different languages that describe the same content. CLKL can be divided into two parts: candidate selection and candidate ranking. Firstly, this paper formulated candidate selection as cross-language information retrieval problem, and proposed a method to generate query by combining title with keywords, which greatly improves the recall of candidate selection, reaching 93.8%. In the part of the candidate ranking, this paper trained a ranking model by mixing bilingual topic model and bilingual embedding, implementing military articles linking in English Wikipedia and Chinese Baidu Baike. The evaluation results show that the accuracy of model achieves 75%, which significantly improves the performance of CLKL. The proposed method does not depend on linguistic characteristics and domain characteristics, and it can be easily extended to CLKL in other languages and other domains.

Keywords Cross-language knowledge linking, Cross-language information retrieval, Bilingual topic model, Bilingual embedding

1 引言

在线百科是凝聚了人类知识的丰富资源,对不同语言编写的在线百科中具有相等关系的文章进行链接有利于知识的全球化共享。跨语言知识链接(Cross-Language Knowledge Linking,CLKL)是指在描述相同内容的不同语言的在线百科文章之间建立联系,它在许多研究领域都具有广泛应用,如命名实体翻译、跨语言信息检索和多语言知识库构建。

大多数CLKL以维基百科为研究目标。维基百科是一个内容自由、公开且多语言编辑的在线百科全书,目前大约有71000名活跃撰稿人以299种语言撰写了4600多万篇文章。DBpedia[1]从维基百科的111种不同语言页面中提取结构化信息,构建了一个多语言的多领域知识库,并成为LOD的核心。YAGO,MENTA和BabelNet基于WordNet和维基百科进行构建,成为被广泛使用的大型多语言本体库。

维基百科虽然存在跨语言链接将不同语言编写的文章页面联系起来,但是还是有很多文章缺乏语言间的联系;并且,维基百科中不同语言编写的文章在数量上严重失衡,比如,维

到稿日期:2018-01-24 返修日期:2018-03-12

余圆圆(1993-),女,硕士生,主要研究方向为自然语言处理,E-mail:yuyuanyuan0823@buaa.edu.cn;**巢文涵**(1979-),男,博士,讲师,CCF会员,主要研究方向为自然语言处理、机器翻译,E-mail:chaowenhan@buaa.edu.cn(通信作者);**何跃鹰**(1975-),男,博士,教授级高级工程师,主要研究方向为人工智能、工业物联网安全;**李舟军**(1963-),男,博士,教授,博士生导师,CCF高级会员,主要研究方向为数据挖掘与人工智能、网络与信息安全。

基百科中英文文章条目数有 553 多万，中文文章条目数有 98 万，而韩语文章条目数只有 30 多万；另一方面，在中文和韩文等语言中存在大量单语知识库，例如，百度百科包含了由 640 多万撰稿人撰写的 1500 多万词条页面，互动百科和搜狗百科等作为中文知识库，其包含的中文词条信息是中文维基百科无法企及的。因此，如果能在英文维基百科和其他单语知识库中建立联系，那么就可以弥补各知识库的不足，这对于构建多语言知识库有很大帮助。

此外，CLKL 还面临其他挑战，其中最严重的挑战之一是词汇的相似性，常用的相似性计算方法，如编辑距离，由于语言的差距而不切实可用。鉴于该问题，设计语言无关的特征是实现 CLKL 的基础[2-3]。近年来，词向量的相关研究工作有很多[4-7]，并且在很多研究领域都有较好的表现。Moreno 等[8]和 Blanco 等[9]通过维基语料训练词向量和实体向量，并将其应用于实体链接任务中。Pappu 等[10]通过维基语料和日志信息训练实体向量，实现了多语言的实体链接。受这些研究工作的启发，我们尝试使用词向量信息帮助我们实现 CLKL。

本文提出了一种易扩展的 CLKL 方法，即将 CLKL 分为候选集选择和候选集排序两部分，将候选集选择问题转换为跨语言信息的检索问题，提出使用文章标题和文章关键词生成查询语句进行文档检索的方法，使得候选集选择达到很高的召回率。使用不依赖于语言特性和领域特性的特征进行候选集排序，使得本文方法可较容易扩展到其他语言和其他领域的 CLKL 任务中。

2　相关工作

Wang 等[11]提出了基于链接因子图模型的跨语言知识链接方法，该方法根据本体的出链相似度、入链相似度、开放分类相似度以及作者兴趣相似度进行本体映射，在中文维基百科和英文维基百科中实现了跨语言链接。Pan 等[12]提出了一个面向特定领域的无监督的跨语言知识链接框架，他们首先通过页面的链接信息（如信息框中的链接文本和锚文本）构建单语知识图谱，通过文本相似度获取一些种子跨语言链接实体，然后通过相似度传播算法不断迭代得到更多的跨语言实体的对应关系，从而实现跨语言的知识链接。

Wang 等[13]将 CLKL 称为跨语言文章链接（Cross-language Article Linking，CLAL），并提出一种基于双语主题模型和其他基于翻译的文本特征，将英文维基百科文章与相应的中文百度百科文章联系起来。他们首先根据英文和中文文章的标题相似性，使用跨语言信息检索方法进行候选集选择，然后利用得到的特征学习一个 SVM 模型，对候选集中的所有文章打分，选择可能性最大且高于一定阈值的页面作为最终结果。

自然语言处理中还有一些研究工作与跨语言知识链接密切相关，包括跨语言实体链接和基于维基百科的跨语言链接发现。实体链接（Entity Linking，EL）[14]的目标是确定文本中提到的实体类型，并将实体链接到知识库中的相应条目。维基百科是实体链接任务中使用得最广泛的知识库。跨语言实体链接（Cross-Language Entity Linking，CLEL）是对实体链接任务的扩展，在 CLEL 中待链接的实体和参考知识库的语言不同。Tsai 等[15]通过维基数据同时训练双语词向量和标题向量，然后利用学习到的标题向量信息提取多种特征，进一步训练线性的 ranking SVM 模型，实现了一个支持 12 种语言的跨语言实体链接系统。CLKL 和（CL）EL 之间的主要区别是，在 EL 或 CLEL 中包含待链接实体的文本只是一个字符串，文本中的其他信息并不直接描述该待链接实体，而在 CLKL 中原文章和链接文章都应该包含关于给定实体的详细信息。

维基百科跨语言链接发现的目的是补充不同语言版本的维基百科之间的跨语言链接。Sorg 等[16]基于图和文本特征训练了一个 SVM 分类器，以考量英文和德文维基百科的文章之间是否应该建立链接关系。图特征主要包括链接数目、链接相似性和类别相似性等。文本特征包括标题的编辑距离和文本重叠性。Oh 等[17]在 Sorg 等的基础上进行改进，通过添加一些语言无关的特征来训练分类器，实现了英文和日文维基百科文章之间的跨语言链接。基于维基百科的跨语言链接发现与 CLKL 的不同之处在于，前者在不同语言版本的维基百科之间进行跨语言链接，而后者可以在不同的在线百科中进行跨语言链接。

3　跨语言知识链接的方法

CLKL 主要分为两步：候选集选择和候选集排序。给定一个中文百度百科页面，候选集选择表示从英文维基百科中找出所有可能与这篇中文页面有关系的页面集合，候选集排序通过特征训练打分器对所有候选集页面进行排序。

3.1　候选集选择

由于知识库可能包含数百万篇文章，因此无论从运行时间还是可操作性来说，将两个知识库中的所有文档对进行比较都是不切实际的。为了避免暴力比较，首先选择合适的候选集，候选集选择作为候选集排序的基础，其结果的好坏将直接影响最终效果。因此，设计合理的候选集选择方法对于 CLKL 是至关重要的，候选集选择的目标是：对于有跨语言链接的中文文章，我们应该将其对应的英文文章包含在候选集中，并且候选集的大小应该尽可能小。

将候选集选取问题转换为中英跨语言信息检索（Cross-Language Information Retrieval，CLIR）问题，并且采用两种主要的跨语言信息检索方法：查询翻译（Query Translation）和文档翻译（Document Translation）。查询翻译将中文查询语句翻译为英文，然后用翻译后的英文构造查询语句，检索英文文档。文档翻译将英文文档翻译为中文，然后用原始的中文构造查询语句，对翻译后的中文文档进行检索，将查询翻译得到的前 100 名的结果和文档翻译得到的前 100 名的结果取并集，并将其作为最终的候选集，如图 1 所示。

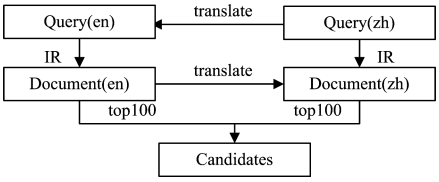

图 1 候选集选择示例图

Fig. 1 Illustration of candidate selection

一般情况下,我们用页面的标题构造查询语句,将页面正文部分作为检索对象。然而,实验发现这种方法的召回率很低,很多查询的候选集不包含正确的文章。考虑到标题内容的有限性和翻译的不准确性可能是导致召回率不高的主要原因,我们提出使用关键词对标题进行补充,将关键词和标题相结合来构造查询语句。关键词提取可以使用 tf-idf 或 textRank 等方法,实验中使用 textRank 从中文页面中提取文章的关键词,并去除 idf 值大于 1000 且词性不属于名词或动词的词。

使用 whoosh 作为全文索引和检索工具,相比关键词,由于标题对于检索匹配度更重要,因此将标题的检索权重设置为关键词的两倍,然后通过使用"OR"操作符连接标题(中文需分词)和关键词构造查询语句,并且选择 BM25 作为排名函数进行文档检索。在文档翻译中,我们使用谷歌翻译对文档进行翻译,考虑到翻译速度,我们只翻译页面标题和第一段文本,使用翻译后的标题和文本建立中文索引。在查询翻译中,使用百度翻译和谷歌翻译对标题进行翻译,对关键词只使用百度翻译,并且用原始的英文标题和全文文本建立英文索引。

举例说明如下:对于百度百科"防弹衣"词条页面,使用 textRank 获取到的页面关键词集合为"人体,能量,断裂,织物,防弹,纱线,弹片,防弹衣,纤维",采用谷歌翻译得到的标题为"Body armor",采用百度翻译得到的标题为"Bulletproof vests",采用百度翻译得到的关键词集合为"human body, energy, fracture, Fabric, bulletproof, Yarn, shrapnel, Body armor, fiber",则查询翻译提交到搜索引擎的查询语句为"(Body)^2 OR (armor)^2 OR (Bulletproof)^2 OR (vests)^2 OR (human body)^1 OR (energy)^1 OR … OR (fiber)^1",文档翻译提交到搜索引擎的查询语句为"(防弹衣)^2 OR (人体)^1 OR (能量)^1 OR … OR (纤维)^1"。

3.2 候选集排序

给定一个中文文档 d 和其英文文档候选集 C_d,对于每一个 $c\in C_d$,通过计算 c 与 d 的相关度得分来对候选集进行排序。相关度得分的计算公式如式(1)所示:

$$s(d,c)=\sum_i w_i \emptyset_i(d,c) \tag{1}$$

其中,$s(d,c)$表示特征的加权和,w_i 表示权重,$\emptyset_i(d,c)$表示特征。下面对使用的特征进行详细介绍。

3.2.1 信息检索特征(Baseline)

将候选集选择步骤中基于查询翻译和文档翻译检索得到的分数进行归一化后作为两个特征,并将其用于训练排序模型。

3.2.2 双语主题模型(Bilingual Topic Model)

主题模型是一个统计生成模型,用于从文档中提取更深层次的潜在主题。特定主题的文档通常会共享某些常见词汇,例如,"导弹"和"将领"常出现在与军事相关的文档中。主题模型允许我们根据文档中的词分布来预测文档的主题及其相应的分布权重。一个广泛使用的主题模型是隐含狄利克雷分布(Latent Dirichlet Allocation,LDA)。LDA 使用概率计算潜在主题分布,每个主题由一系列词的分布来表示,每个词都有一个概率分数来衡量它对这个主题的贡献。

LDA 是一个生成式图模型,用于对离散数据(如文本)的基础主题结构进行建模,如图 2 所示。为了生成一个由 M 个文档和 K 个潜在主题组成的语料,LDA 执行如下操作:对于每个潜在主题 k,其主题-词分布 φ_k 由一个决定潜在主题词分配情况的先验 Dirichlet 分布 dir(α)得到。文档 m 的文档-主题分布 θ_m 由一个决定文档主题分配情况的先验 Dirichlet 分布 dir(α)得到。对于文档 m 中的第 n 个词,首先通过 θ_m 的 multionmial 分布得到其对应的主题 $Z_{m,n}$,最后由 $\varphi_{Z_{m,n}}$ 的 multionmial 分布得到单词 $w_{m,n}$。

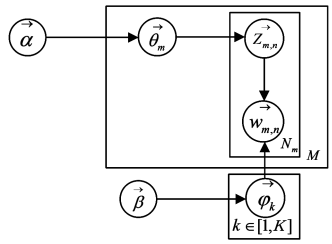

图 2 LDA 图模型

Fig. 2 Graphical model representation of LDA

LDA 一般用于对单语文本进行分析。为了处理双语文本数据,我们使用双语主题模型(Bilingual Topic Model,BLDA)。BLDA 是对 LDA 的扩展,使其能对双语文档对进行建模。每个文档对是一对用不同的语言编写但在语义上相互关联的文档,并且认为一个文档对中的所有文档的主题分布相同,但是对于每个主题 k,它在不同语言上的词分布是不同的。

BLDA 的图模型如图 3 所示。首先,对于每个潜在主题 k,根据先验 Dirichlet 分布 dir(β)得到每种语言下的主题-词分布 $\varphi_{k,l}$;然后对每个文档对 m,通过先验 Dirichlet 分布 dir(α)得到文档-主题分布 θ_m;之后,对于文档对中的每篇文档 m 中的第 n 个词,首先通过 θ_m 的 multionmial 分布得到其对应的主题 $Z_{m,n,l}$,然后由 $\varphi_{Z_{m,n,l}}$ 的 multionmial 分布得到单词 $w_{m,n,l}$。

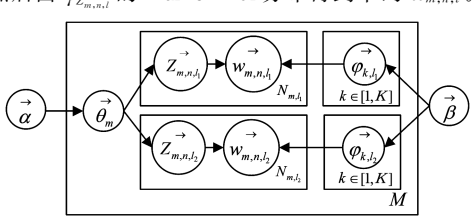

图 3 BLDA 图模型

Fig. 3 Graphical model representation of BLDA

训练好 BLDA 模型后,给定一篇文档,可以通过 BLDA 模

型得到其主题分布向量，从而可以用主题分布的余弦相似度来衡量中英文文章的相似度。将双语主题模型相似度作为一个特征，记为 BTMS(Bilingual Topic Model Similarity)。

3.2.3　双语词向量(Bilingual Embeddings)

为了学习双语词向量，首先训练单语词向量，然后通过一个双语词典学习源语言到目标语言的映射矩阵 W，最后将源语言的词向量与映射矩阵 W 进行运算后投影到目标语言语义空间，从而实现双语词向量的学习。

分别使用百度百科和英文维基百科的军事领域数据学习单语词向量，并且为了学习实体向量，我们对数据进行了相关处理。对于维基百科数据，将文档中出现的锚文本(带链接文本)用其所指向页面的标题替代。如原文本为“A large majority of Finns are members of the Evangelical Lutheran Church，and freedom of religion is guaranteed under the Finnish Constitution.”，替换后的文本为“A large majority of Finns are members of the Evangelical_Lutheran_Church_of_Finland，and freedom of religion is guaranteed under the Constitution _ of _ Finnish.”。其中，“Evangelical Lutheran Church”是锚文本，它指向标题为“Evangelical Lutheran Church of Finland”的页面，我们使用“Evangelical_Lutheran_Church_of_Finland”替换锚文本[18]。对于中文文本，我们将锚文本用其指向的页面标题替换后，将该标题加入到分词词典中，以保证对中文文本进行分词时该标题不被分开。对文本进行这样处理的原因在于：1)我们认为文本中的锚文本非常重要，可以认为它们是一个实体，对于判断中英文文本的相似度有帮助；2)为了学习双语词向量，需要双语词典，通过维基百科的跨语言链接很容易得到中英文标题对应的词典，通过对标题进行以上处理能更容易地获得双语词典。

得到单语词向量之后，我们进一步学习双语词向量。假设 $X\in\mathbb{R}^{n,d_1}$ 和 $Z\in\mathbb{R}^{n,d_2}$ 分别表示双语词典中源语言和目标语言词的词向量矩阵，其中 n 表示词典大小，d_1 表示源语言词向量的维数，d_2 表示目标语言词向量的维数。我们的目标是找到一个线性变换矩阵 $W\in\mathbb{R}^{d_1,d_2}$，使得 $XW\in\mathbb{R}^{n,d_2}$ 逼近 Z，因此学习目标如式(2)所示：

$$\arg\min_W \| XW-Z \|_F^2 \tag{2}$$

我们借鉴 Artetxe 等[19]的方法，通过 SVD 学习线性转换矩阵 W，在得到 W 后，假设源语言中所有词向量矩阵为 $M\in\mathbb{R}^{k,d_1}$，使用 W 矩阵对 M 进行线性转换后得到 $M'=MW\in\mathbb{R}^{k,d_2}$，从而实现源语言到目标语言语义空间的映射。

给定一个百度百科页面和英文维基百科页面，在双语词向量的基础上，使用 3 个特征对这两篇文档的相似度进行比较。

1)Title Embedding Similarity(TES)：中文标题和英文标题的向量余弦相似度。如果没有对应的标题向量信息，那么标题的词向量由组成标题的单词词向量的均值得到。

2)Summary Embedding Similarity(SES)：中文摘要和英文摘要的向量余弦相似度。其中摘要由所有单词词向量的 TF-IDF 权重和表示。

3)Anchor Text Embedding Similarity (ATES)：中文页面中锚文本和英文页面中锚文本的向量余弦相似度。其中用来比较的向量是所有锚文本向量的均值。

3.2.4　共现信息(Co-occurrence Information)

在军事领域中，关于人物、战争和武器装备的文章有很多，我们认为文中出现的英文名称、英文缩写、型号和数字时间等特征是很重要的。比如，在第一句话中出现的英文名称很有可能是该文章对应的英文标题，英文缩写可能是专有名词，数字可能是重要事件的发生时间。因此，我们通过正则表达式对中文文章中的以上信息进行抽取，具体抽取内容和规则如下。

1)U_1：在括号中出现的英文名称，且括号中除了英文名称和空格外不能包含其他字符。

2)U_2：文中所有长度大于 3 的英文大写缩写词。

3)U_3：文中出现的型号，即由数字、字母和一些符号组合的单词。

4)U_4：4 位的数字或者后面的字为“年”的其他数字。

对于中文文章 d 和候选英文文章 c，它们之间的相关分数计算如式(3)所示：

$$score=\log\Big(\sum_{i=1}^{4}\sum_{u\in U_i} w_i I_c(u)+1\Big) \tag{3}$$

其中，w_i 表示权重，U_i 表示使用以上规则抽取到的词集，$I_c(u)$为指示函数。如果 u 在 c 中出现，则值为 1，否则为 0。为了防止得到的分数差距很大，对结果取 log。将该得分作为一个特征用于模型的训练。

3.2.5　空链接处理

在 CLKL 中，不是所有的中文文章都有对应的英文文章，有的中文文章不存在对应的英文跨语言链接文章，这种情况称为空链接(NIL)。对于空链接的处理，目前的做法主要有两种：设定阈值[20]和训练分类器[21]。设定阈值的方法是确定一个置信度分数，当排序最靠前的页面得分低于置信度分数时，认为它是 NIL。训练分类器的方法是使用一些特征训练一个二分类器，判断排序得分最高的页面是否是正确的页面。由于缺少数据用于训练分类器，因此采用设定阈值的方法处理 NIL。

4　实验

4.1　数据准备

实验中使用的数据包括百度百科中的军事领域数据和中英文维基百科数据。对于百度百科中的军事领域文章，首先根据互动百科中的分类树获取军事领域下的所有类别词 C，之后遍历百度百科页面，当页面所属类别与 C 有交集时，认为该页面属于军事领域，并对其进行爬取，最终获取到 68891 个百度百科的军事领域页面。我们从维基转储数据库[1)]上获

1) http://download.wikipedia.org

取了 20170501 版本的中英文维基数据，并且使用开源工具 WikiTailor[22] 从英文维基数据中获取了 307555 个英文军事领域页面。WikiTailor 获取领域数据的方法为：给定一个顶级类别 c^*（如军事），从属于顶层类别的页面中按词频排序获取关键词作为领域相关术语，之后按照广度优先策略不断扩展下一级类别。扩展过程中，通过一些规则来决定此类别是否属于该领域，如该类别是否包括领域相关术语，与该类别同等深度的其他类别属于该领域的比例是否超过一定阈值。同时，设置终止条件，当爬取深度达到一定程度或者当前深度的所有类别属于该领域的比例小于设定阈值时，停止爬取。获得所有领域类别词 C' 后，爬取所有页面类别与 C' 有交集的词条页面。

4.2 模型训练

为了训练 BLDA，通过英文军事领域的 307555 篇文档的中文跨语言链接获取到 36572 个文档对，基于 JGibbLDA[1] 训练了一个领域内的 BLDA 模型。基于 Zhang 等[23] 的研究工作，将训练参数设定为主题数 $K=100$，超参数 α 和 β 分别为 0.5 和 0.1，训练迭代次数为 1000。

单语词向量通过爬取到的中文百度百科中的军事领域语料和英文维基百科中的军事领域语料进行训练，使用 gensim[2] 设置训练维数为 100。之后，以中文维基百科为中介，获取中英文词典来辅助双语词向量的训练。词典的具体构造方法为：对于百度百科中一个锚文本指向的标题 t_1，如果在中文维基百科中存在标题 $t_1=t_2$，并且该标题页面有英文跨语言链接，对应的英文页面标题为 t_3，那么将 (t_1,t_3) 加入到词典中。将中文军事页面中所有锚文本和页面标题都进行上述操作后，就得到最终的双语词典。我们使用的双语词典大小约为 13000。

4.3 实验结果

为了评估本文方法在 CLKL 任务上的表现，我们构造了两个数据集。对于百度百科中的军事领域的文章 d，在中文维基百科中寻找与 d 标题名相同并且存在英文跨语言链接的文章 d_c，对应的英文文章为 d_e。将 (d,d_e) 作为一个文档对加入到数据集 1 中，由此得到的数据集 1 共包含 3135 个文档对。用该数据集来验证本文方法在有跨语言链接文章上的表现。在某些情况下，百度百科的文章可能没有相应的英文维基百科的文章（NIL）。为了验证本文方法在这些空链接文章上的表现，我们手动标注了 500 条没有跨语言链接的中文文章，结合从数据集 1 中随机抽样的 500 条数据作为数据集 2。训练时将数据集按照 4∶1 的比例分为训练集和测试集，由于数据采样的随机性，我们将 10 次实验结果的平均值作为最终结果。

4.3.1 候选集选择结果

为了判断候选集选择是否合理，在数据集 1 上进行了多组对比实验，实验结果如表 1 所列。表 1 中，Query 表示用来构造查询翻译中查询语句的内容；$|C|$ 表示候选集；$|C_{\mathrm{NIL}}|$ 表示候选集为空的查询；$|C_{\mathrm{AVG}}|$ 表示候选集的平均大小；$Recall(C)$ 表示召回率，定义为正确文档包含在候选集中的查询数与所有非空查询数之比。

表 1 候选集的选择结果

Table 1 Selection results of candidate

Query	$\lvert C\rvert$	$\lvert C_{\mathrm{NIL}}\rvert$	$\lvert C_{\mathrm{AVG}}\rvert$	$Recall(C)$
title(google)($q1$)	452228	13	144.25	0.75853
title(baidu)($q2$)	427595	133	136.39	0.81627
keyword(baidu)($q3$)	545512	20	174.01	0.69282
$q1+q2$	452837	2	144.45	0.87687
$q1+q3$	546260	0	174.25	0.89187
$q2+q3$	545663	0	174.06	0.92376
$q1+q2+q3$	546933	0	174.46	0.93844

表 1 中的实验结果由查询翻译和文档翻译两种方法得到。同一实验中两种方法的查询条件相同，由于查询翻译中的查询条件是使用不同的翻译工具获得的，而文档翻译中的查询条件为中文标题和中文关键词，因此表 1 只对查询翻译中的查询条件进行说明。表 1 中，$q1$ 表示使用谷歌翻译得到的标题，$q2$ 表示使用百度翻译得到的标题，$q3$ 表示使用百度翻译得到的关键词。表 1 中的第一行表示使用谷歌翻译的标题对英文索引进行检索，结合使用中文标题对中文索引进行检索得到的实验结果。表 1 中的最后一行表示使用谷歌翻译的标题、百度翻译的标题和百度翻译的关键词对英文索引进行检索，结合使用中文标题和中文关键词对中文索引进行检索得到的结果。

我们比较了百度翻译和谷歌翻译在候选集选择中的表现（表 1 中前两行）。结果表明，百度翻译的表现较好，能达到更高的召回率，但同时使用百度翻译和谷歌翻译时的效果更好（表 1 中第四行）。为了验证关键词对候选集选择的作用，们分别使用标题、关键词、标题和关键词相结合生成查询的方法进行实验。结果表明，使用标题和关键词相结合生成查询的方法时，候选集的平均大小增加，召回率明显提高，最高达到了93.8%。因此，我们认为使用关键词对标题进行补充生成查询，可以提高候选集的选择效果。后续实验都是在采用最后一种策略得到的候选集上进行的。

4.3.2 候选集排序结果

为了验证特征的有效性，使用 linear ranking svm[24] 学习特征权重，并且在数据集 1 上使用序列前向选择方法对不同的特征组合进行测试，从而选择最佳特征组合。实验中采用信息检索特征作为基础特征，在基础特征之上不断增加其他特征，直到所有特征都被选择。以 Mean Reciprocal Ratio (MRR) 和 Accuracy 作为数据集 1 的评价指标，每次选择评价指标最高的特征和前面的特征进行组合，实验结果如表 2 所列。

1) http://jgibblda.sourceforge.net/

2) https://radimrehurek.com/gensim/models/word2vec.html

表 2　不同特征组合在数据集 1 上的结果

Table 2　Results of different feature combination on dataset1

Level	Feature	*MRR*	*Accuracy*
0	**Baseline(BL)**	0.69419	0.61164
1	BL+BTMS	0.75124	0.68070
	BL+SES	0.75178	0.67783
	BL+TES	0.74762	0.68421
	BL+ATES(set1)	0.75963	0.6874
	BL+COI	0.73368	0.66029
2	set1+BTMS	0.77849	0.71276
	set1+SES	0.77784	0.71196
	set1+TES(set2)	0.78101	0.71978
	set1+COI	0.77727	0.71100
3	set2+SES	0.78832	0.72568
	set2+BTMS(set3)	0.79208	0.73541
	set2+COI	0.79184	0.73413
4	set3+SES	0.79570	0.73892
	set3+COI(set4)	0.79889	0.74290
5	**set4+SES(all)**	0.80355	0.75104

注：BTMS：Bilingual Topic Model Similarity；SES：Summary Embedding Similarity；TES：Title Embedding Similarity；ATES：Anchor Text Embedding Similarity；COI：Co-occurrence Information

实验结果显示，本文方法在数据集 1 上的 *MRR* 和 *Accuracy* 分别达到了 0.80355 和 0.75104，表明了其在英文维基百科和百度百科跨语言链接方面的有效性。在基础特征上，增加任意一个特征都能明显改善实验结果，在一个特征组合的基础上，每增加一个特征，实验结果也会提升，所有特征都使用时，实验结果达到最高值，这进一步说明了特征的有效性。且这些特征并不是很依赖于语言特性和领域特性，因此可以很容易地扩展至其他语言和其他领域的跨语言知识链接任务中。

实验中采用 Accuracy 作为评价指标。本文方法在空链接上的表现如表 3 所列，其中 *Accuracy*(*link*)表示方法在有跨语言链接文章中的准确度，*Accuracy*(*nil*)表示方法在空链接文章中的准确度。结果显示，在基础特征上增加我们的特征，效果得到显著提升。

表 3　数据集 2 的结果

Table 3　Results of evaluation on dataset2

Feature	*Accuracy* (link)	*Accuracy* (nil)	*Accuracy* (avg)
Baseline(BL)	0.618	0.661	0.640
BL+BTMS	0.661	0.660	0.661
BL+SES	0.685	0.719	0.702
BL+TES	0.679	0.723	0.701
BL+ATES	0.663	0.723	0.693
BL+COI	0.653	0.686	0.670
all	0.724	0.714	0.719

5　结果分析

5.1　特征的有效性

只使用基础特征时，百度百科中的文章“米歇尔·魏特曼”的跨语言链接被预测为维基页面“Mother Lü”，其正确的维基页面为“Michael Wittmann”。两个页面都是对人物的描写，但是“Mother Lü”是汉朝时期的人物，“Michael Wittmann”则是德国上尉。通过加入 BTMS 特征能正确区分这两篇英文文章，因为它们的双语主题相似度差异很大，分别为 0.10 和 0.89。

百度百科文章“澳大利亚陆军”对应的跨语言链接文章为维基页面“Australian Army”，由于标题相似度极高，基础特征将维基页面“Royal Australian Army Service Corps”预测为最终结果。通过加入双语词向量信息，结果得到修正。

5.2　错误示例

本文方法虽然能有效地实现跨语言链接，但还是存在一些正确的候选页面在候选集排序中没有排名第一。检查结果后，我们可以将错误的原因分为两类。

1）文章相似性太强，如页面“尼南斯·阿农迪亚德战役”的跨语言链接页面为“Nirnaeth Arnoediad”，得分最高的页面为“First Battle of Beleriand”。分析发现正确页面和预测页面分别是第五次和第一次贝尔兰会战，页面中的人名和地名的相似度很高，导致预测错误。

2）文章太过简洁，使得本文方法没有提取到足够的有效信息，导致预测结果错误。如页面“第一帕提亚军团”的对应跨语言链接页面为“Legio I Parthica”，预测页面为“Parthian Empire”。预测页面和中文页面有一定的相关性，并且预测页面的内容丰富，使得它在候选集排序中的得分较高。

结束语　跨语言知识链接是指在不同语言的在线百科全书描述相同内容的文章之间建立联系。本文将跨语言知识链接分为两部分：候选集选择和候选集排序。将候选集选择作为跨语言信息检索任务，使用 textRank 抽取文章关键词，结合文章标题生成查询，极大地提高了候选集选择的召回率。使用双语主题模型、双语词向量和共现信息训练 ranking SVM 模型进行候选集排序。实验结果显示本文方法在自己构建的数据集 1 上的 *MRR* 和 *Accuracy* 分别达到了 0.80355 和 0.75104，表明了本文方法能够有效地生成英文维基百科和百度百科之间的跨语言链接；且本文方法并不是很依赖于语言特性和领域特性，因此可以很容易地扩展到其他语言和其他领域的跨语言知识链接任务中。

将来，我们计划在候选集选择排序部分加入信息框和类别等信息，以训练排序模型，但通过设定阈值过滤空链接的方法并不友好，可以尝试使用其他方法对这种情况进行改进。

参 考 文 献

[1] LEHMANN J，ISELE R，JAKOB M，et al. DBpedia-a large-scale，multilingual knowledge base extracted from Wikipedia [J]. Semantic Web，2015，6(2)：167-195.

[2] WANG Z，LI J，TANG J. Boosting Cross-Lingual Knowledge Linking via Concept Annotation[C]// Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence. Beijing，China：AAAI Press，2013：2733-2739.

[3] WANG Z，PAN L，LI J，et al. Boosting to Build a Large-Scale Cross-Lingual Ontology[C]// China Conference on Knowledge

Graph and Semantic Computing. Singapore: Springer, 2016: 41-53.

[4] RUDER S, VULIC I, SØGAARD A. A survey of cross-lingual embedding models [J/OL]. https://arxiv. org/pdf/1706.04902v2. pdf.

[5] 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.

[6] ARTETXE M, LABAKA G, AGIRRE E. Learning bilingual word embeddings with (almost) no bilingual data[C]// Meeting of the Association for Computational Linguistics. 2017: 451-462.

[7] DUONG L, KANAYAMA H, MA T F, et al. Learning Crosslingual Word Embeddings without Bilingual Corpora[C]// Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. USA: ACL, 2016: 1285-1295.

[8] MORENO J G, BESANÇON R, BEAUMONT R, et al. Combining word and entity embeddings for entity linking[C]// European Semantic Web Conference. Cham: Springer, 2017: 337-352.

[9] BLANCO R, OTTAVIANO G, MEIJ E. Fast and space-efficient entity linking for queries[C]// Proceedings of the Eighth ACM International Conference on Web Search and Data Mining. ACM, 2015: 179-188.

[10] PAPPU A, BLANCO R, MEHDAD Y, et al. Lightweight multilingual entity extraction and linking[C]// Proceedings of the Tenth ACM International Conference on Web Search and Data Mining. ACM, 2017: 365-374.

[11] WANG Z, LI J, WANG Z, et al. Cross-lingual knowledge linking across wiki knowledge bases[C]// International Conference on World Wide Web. ACM, 2012: 459-468.

[12] PAN L, WANG Z, LI J, et al. Domain Specific Cross-Lingual Knowledge Linking Based on Similarity Flooding[C]// International Conference on Knowledge Science, Engineering and Management. Cham: Springer, 2016: 426-438.

[13] WANG Y C, WU C K, TSAI T H. Cross-Language Article Linking with Different Knowledge Bases Using Bilingual Topic Model and Translation Features[J]. Knowledge-Based Systems, 2016, 111(3): 228-236.

[14] SHEN W, WANG J, LUO P, et al. LINDEN: linking named entities with knowledge base via semantic knowledge[C]// Proceedings of the 21st International Conference on World Wide Web. ACM, 2012: 449-458.

[15] TSAI C T, DAN R. Cross-lingual Wikification Using Multilingual Embeddings[C]// Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2016: 589-598.

[16] SORG P, CIMIANO P. Enriching the crosslingual link structure of wikipedia-a classification-based approach[C]// Proceedings of the AAAI 2008 Workshop on Wikipedia and Artifical Intelligence. Chicago, Illinois, 2008: 49-54.

[17] OH J H, KAWAHARA D, UCHIMOTO K, et al. Enriching multilingual language resources by discovering missing cross-language links in wikipedia[C]// Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology-Volume 01. IEEE Computer Society, 2008: 322-328.

[18] SHEARKAT E, MILIOS E E. Vector embedding of wikipedia concepts and entities[C]// International Conference on Applications of Natural Language to Information Systems. Cham: Springer, 2017: 418-428.

[19] ARTETXE M, LABAKA G, AGIRRE E. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance[C]// Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. 2016: 2289-2294.

[20] HOFFART J, ALTUN Y, WEIKUM G. Discovering emerging entities with ambiguous names[C]// Proceedings of the 23rd International Conference on World Wide Web. ACM, 2014: 385-396.

[21] RATINOV L, ROTH D, DOWNEY D, et al. Local and global algorithms for disambiguation to wikipedia[C]// Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Association for Computational Linguistics, 2011: 1375-1384.

[22] BARRÓN-CEDEÑO A, ESPAÑA-BONET C, BOLDOBA J, et al. A factory of comparable corpora from wikipedia[C]// Proceedings of the Eighth Workshop on Building and Using Comparable Corpora. 2015: 3-13.

[23] ZHANG T, LIU K, ZHAO J. Cross Lingual Entity Linking with Bilingual Topic Model[C]// Proceedings of the 23rd International Joint Conference on Artificial Intelligence. Beijing, China: AAAI Press, 2013: 2218-2224.

[24] LEE C P, LIN C J. Large-scale linear ranksvm[J]. Neural Computation, 2014, 26(4): 781-817.