

融入事件实体知识的汉越跨语言新闻事件检索

薛振宇^{1,2}, 余正涛^{1,2}, 高盛祥^{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)

摘要: 现有汉越跨语言新闻事件检索方法较少使用新闻领域内的事件实体知识, 在候选文档中存在多个事件的情况下, 与查询句无关的事件会干扰查询句与候选文档间的匹配精度, 影响检索性能。提出一种融入事件实体知识的汉越跨语言新闻事件检索模型。通过查询翻译方法将汉语事件查询句翻译为越南语事件查询句, 把跨语言新闻事件检索问题转化为单语新闻事件检索问题。考虑到查询句中只有单个事件, 候选文档中多个事件共存会影响查询句和文档的精准匹配, 利用事件触发词划分候选文档事件范围, 减小文档中与查询无关事件的干扰。在此基础上, 利用知识图谱和事件触发词得到事件实体丰富的知识表示, 通过查询句与文档事件范围间的交互, 提取到事件实体知识表示与词以及事件实体知识表示之间的排序特征。在汉越双语新闻数据集上的实验结果表明, 与BM25、Conv-KNRM、ATER等基线模型相比, 该模型能够取得较好的跨语言新闻事件检索效果, NDCG和MAP指标最高可提升0.712 2和0.587 2。

关键词: 跨语言检索; 事件实体; 事件触发词; 事件范围; 排序学习; 事件检索

开放科学(资源服务)标志码(OSID):



中文引用格式: 薛振宇, 余正涛, 高盛祥. 融入事件实体知识的汉越跨语言新闻事件检索[J]. 计算机工程, 2022, 48(8): 274-282, 291.

英文引用格式: XUE Z Y, YU Z T, GAO S X. Chinese-Vietnamese cross-language news event retrieval incorporating event entity knowledge[J]. Computer Engineering, 2022, 48(8): 274-282, 291.

Chinese-Vietnamese Cross-Language News Event Retrieval Incorporating Event Entity Knowledge

XUE Zhenyu^{1,2}, YU Zhengtao^{1,2}, GAO Shengxiang^{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] The existing Chinese-Vietnamese cross-language news event retrieval methods are not sufficiently integrated into the knowledge of event entities in the news field. Furthermore, when there are multiple events in the candidate document, events unrelated to the query sentence interfere with the matching accuracy between the query sentence and the candidate documents, which affects retrieval performance. This study proposes a Chinese-Vietnamese cross-language news event retrieval model incorporating event entity knowledge. The query translation method is used to translate Chinese event query sentences into Vietnamese event query sentences, and the cross-language news event retrieval problem is transformed into a monolingual news event retrieval problem. Considering that there is only a single event in the query sentence, the coexistence of multiple events in the candidate document affects the exact match between the query sentence and the document. The event trigger word is used to divide the event range of the candidate document and to reduce the interference of events unrelated to the query in the document. On this basis, the knowledge graph and event trigger words are used to obtain the rich knowledge representation of event entities. Through the interaction between the query sentence and the document event scope, the ranking features between the knowledge representation of event entities and the knowledge representation of words and event entities are extracted. The experimental results on the Chinese-Vietnamese bilingual news dataset show that compared with baseline models such as BM25, Conv-KNRM, and ATER, the proposed model achieves better cross-language news event retrieval performance; furthermore, using the proposed model, the NDCG and MAP indicators can be improved by up to 0.712 2 and 0.587 2.

基金项目: 国家自然科学基金(61972186, 61762056, 61472168); 国家重点研发计划(2018YFC0830105, 2018YFC0830101, 2018YFC0830100); 云南省重大科技专项(202002AD080001); 云南省高科技人才项目(201606, 202105AC160018); 云南省基础研究计划(202001AS070014, 2018FB104)。
作者简介: 薛振宇(1996—), 男, 硕士研究生, 主研方向为自然语言处理、跨语言信息检索; 余正涛, 教授、博士; 高盛祥(通信作者), 副教授、博士。

收稿日期: 2021-05-10 **修回日期:** 2021-09-05 **E-mail:** gaoshengxiang.yn@foxmail.com

【Key words】cross-language retrieval; event entity; event trigger; event range; ranking learning; event retrieval

DOI: 10.19678/j.issn.1000-3428.0061596

0 概述

汉越跨语言新闻事件检索任务是指用户将包含事件信息的汉语查询句输入检索系统后,检索系统为用户返回一系列与查询句中事件信息相关的越南语新闻文档。目前主流的跨语言信息检索系统采用查询翻译方法^[1]、文档翻译方法^[2]或中间语言翻译方法^[3]。其中,查询翻译方法首先将查询句翻译成候选文档所使用语言下的查询句,然后使用翻译后的查询句对候选文档进行检索排序。因为该方法只翻译查询句,翻译难度较低且正确率高,所以在跨语言信息检索任务中最常使用。

目前主流的检索模型有基于特征的检索模型^[4]和基于神经网络的检索模型^[5]。

基于特征的检索模型由于依赖于人工寻找特征且找到的特征数量有限,导致模型参数量较少,使得模型泛化能力降低,因此,其相较于基于神经网络的检索模型在完成检索任务时表现较差。但是,有一些基于特征的检索模型在融合实体语义信息后性能会得到较大提升,包括利用实体标注进行文本表示的检索模型^[6]、多排序特征的检索模型^[7]、基于查询句与文档间实体连接的检索模型^[8]以及基于知识图谱实现查询句和文档软匹配的检索模型^[9],这些模型均是通过融合实体语义信息来提高检索性能。

基于神经网络的检索模型又可分为基于表示的检索模型和基于交互的检索模型。基于表示的检索模型^[10]在初始阶段对查询句和文档单独进行处理,然后使用神经网络分别编码,得到各自的文本表征并进行相似度计算,最后将文本表征的相似度作为查询和文档的相似度得分,根据相似度得分对文档进行排序。这种基于表示的检索模型在最后阶段才会利用文本表征计算查询与文档间的相似度,模型的效果过于依赖文本表征的质量,并且会丢失对模型效果有正向作用的句法和词法等基础的文本特征。基于交互的检索模型^[11]在开始阶段就计算查询句与文档之间的词级别的语义相似度作为基础的交互特征,并在交互特征的基础上进一步抽取层次交互特征,得到查询句与文档交互固定维度的表示,最后通过计算相似度对文档进行打分排序。由于基于交互的检索模型尽可能早地将查询句和文档进行了交互,捕获到了查询句与文档之间相对更基础的特征,因此检索效果相较于基于表示的检索模型提升显著。

基于交互的检索模型利用神经网络和词级交互信息学习相对复杂的排序模型,其在开放域的检索

任务上性能优于基于特征的检索模型和基于表示的检索模型。然而,在汉越跨语言新闻事件检索任务中包含事件描述的候选文档中含有大量新闻事件领域内特有的事件实体^[12],如人名、地名、组织机构名、特定政治概念名等事件实体。目前,事件实体的语义信息能否融入基于交互的检索模型中来提高汉越跨语言新闻事件检索模型排序的性能尚不明确,并且在事件检索任务中,一篇候选文档中可能包含多个事件,这会干扰事件查询句和包含该事件信息的候选文档匹配的性能。以越南语候选文档中的事件描述“Tổng Giám đốc Tổ chức Y tế thế giới (WHO) Tedros có bài phát biểu mừng năm mới, sự kiện xảy ra trong năm 2020 sẽ cung cấp bài học và lưu ý khi bước vào năm 2021, điều quan trọng nhất là Chính phủ phải tăng ngân sách cho y tế công cộng, bao gồm cung cấp tài chính cho mọi người có được vắc-xin COVID-19.”为例,该描述中包含3种不同的事件,分别为:“Tedros có bài phát biểu mừng năm mới”,中文解释为“Tedros 发表新年演讲”;“phải tăng ngân sách cho y tế công cộng”,中文解释为“政府必须增加公共卫生预算”;“tài chính cho mọi người có được vắc-xin COVID-19”,中文解释为“资助所有人获得 COVID-19 疫苗”。假定用户对“Tedros 发表新年演讲”这一事件感兴趣,用户输入的查询句可能是“Tedros biểu năm mới”,在查询句与候选文档进行排序匹配时,其余2种事件会增加噪声,扩大匹配的事件范围,降低查询句与候选文档匹配的准确度,影响检索模型的性能。

一篇候选文档中可能包含多个事件,为了能在候选文档中准确地找到与查询句中提及的事件相关的事件范围,每个事件均有相应的事件触发词,在事件查询句中也有事件触发词的情况下,可以将事件触发词作为分类不同事件的依据。例如“Tedros 发表新年演讲”事件中的触发词为“biểu”。本文提出一个融入事件实体知识的基于交互的汉越跨语言新闻事件检索模型。对汉语查询句进行翻译后,利用无监督标注方法 PredPat^[13]识别查询句与候选文档中的事件触发词划分候选文档事件范围,利用事件实体、事件触发词和多语言知识图谱获得查询句与事件范围中事件实体的语义知识表示。在此基础上,使用基于交互的检索排序模型并融入事件实体的语义知识表示,对查询句和文档进行匹配排序,从而提升模型检索性能。

1 本文汉越跨语言新闻事件检索模型

1.1 模型结构

本文构建一个融入事件实体知识的汉越跨语言新闻事件检索模型,模型结构如图1所示。其中,查询句为汉语事件查询句,候选文档为越南语新闻文档。首先将汉语查询句翻译为越南语查询句;然后识别出翻译后的查询句与越南语文档中的事件触发词,并基于文档中的事件触发词对文档划分事件范围;之后使用越南语事件实体识别方法^[12]识别出查

询句与文档事件范围中的事件实体,进而基于多语言知识图谱和事件触发词对事件实体的语义进行扩充;最后使用基于交互的检索模型框架分别提取查询句中的词和文档事件范围中的词、查询句中的词和文档事件范围中扩充后的事件实体、查询句中扩充后的事件实体和文档事件范围中的词以及查询句中扩充后的事件实体和文档事件范围中扩充后的事件实体交互所产生的排序特征,根据排序特征计算查询句与文档最终的排序得分。

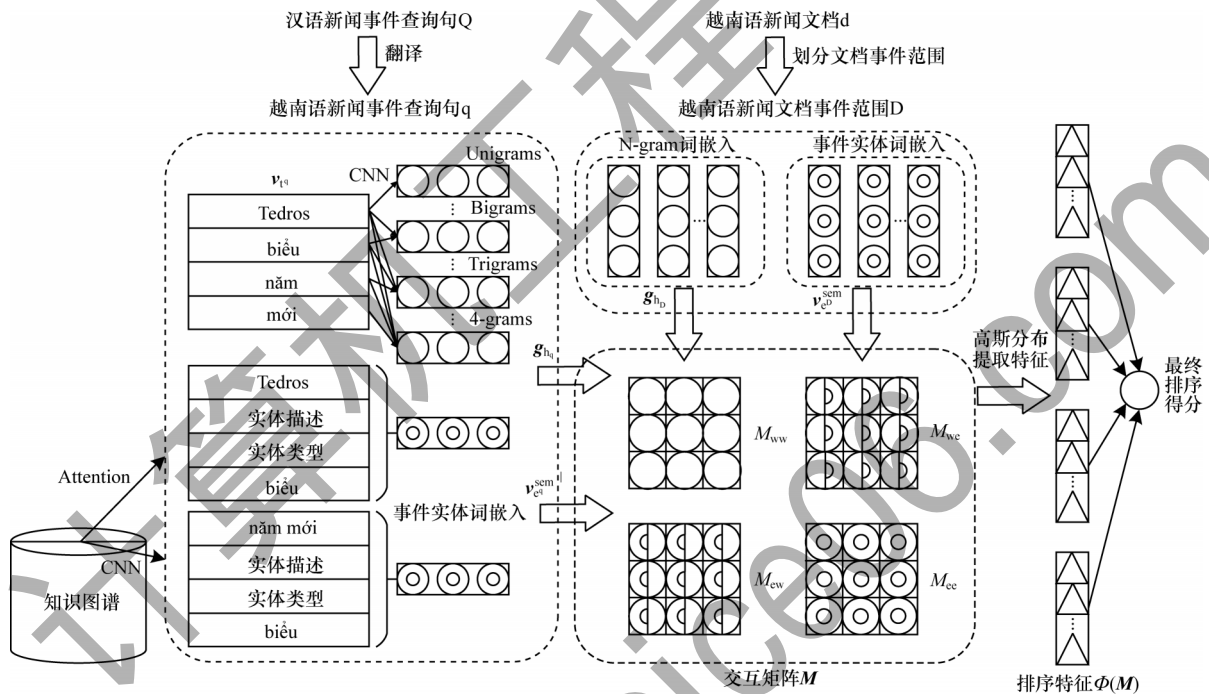


图1 融入事件实体知识的汉越跨语言新闻事件检索模型结构

Fig.1 Structure of Chinese-Vietnamese cross-language news event retrieval model incorporating event entity knowledge

1.2 越南语事件查询句生成

汉越跨语言新闻事件检索首先面临的问题是汉越之间存在的语言隔阂对检索造成的阻碍。目前,跨语言新闻事件检索研究中最常用的方法是查询翻译方法,即先使用现有的翻译工具将事件查询句的语言翻译为候选新闻文档所用语言,再利用翻译后的事件查询句进行事件检索。因此,本文通过现有的翻译工具将汉语事件查询句Q翻译为越南语事件查询句q。

1.3 文档事件范围检测

单个候选新闻文档中包含多个事件,若对整个文档与事件查询句进行匹配,会产生较大的匹配误差。因此,对于每个文档d,本文使用PredPatt方法识别d中所有的事件触发词 $T_{trg} = \{t_{trg}^1, t_{trg}^2, \dots, t_{trg}^f\}$ 。假定其中一个触发词 t_{trg}^i 的位置为 l ,窗口大小为 p ,则取 $l-p$ 至 $l+p$ 范围内的词作为该触发词在文档中的事件范围 D_i 。计算出所有触发词对应的事件范围之后,可以将文档d看作 f 个事件范围的集合,即 $d = \{D_1, D_2, \dots, D_f\}$ 。

1.4 事件实体的语义知识表示

本文利用越南语事件实体识别方法识别出查询句q与文档事件范围D中的事件实体,并在多语言知识图谱ConceptNet^[14]中找到其对应的实体类型和实体描述,融入本文模型。事件触发词位于2个实体之间并连接2个事件实体,可以表征事件实体之间的关系。本文使用PredPatt识别出查询句与文档事件范围中的事件触发词,并融入事件触发词本身的语义信息。最终,事件实体的语义表示包含以下4种大小为 L 维的词嵌入:

1) 实体词嵌入

将实体词e通过词嵌入层 Emb_e 得到大小为 L 维的实体词词向量 v_e^{emb} ,计算公式如式(1)所示:

$$v_e^{emb} = Emb_e(e) \quad (1)$$

2) 实体描述词嵌入

通过ConceptNet找到实体词e对应的包含 m 个词的实体描述。首先将描述中的每一个词w通过词嵌入层 Emb_w 得到大小为 L 维的词向量 v_w ,将 m 个词

向量视作一个整体向量矩阵 V_w 。然后将 V_w 通过卷积操作得到代表长度为 h 的 n -gram 向量 g_e^i , 计算公式如式(2)所示:

$$g_e^i = \text{relu}(W_{\text{CNN}} \cdot V_w^{i:i+h} + b_{\text{CNN}}) \quad (2)$$

其中: W_{CNN} 与 b_{CNN} 是卷积核的2个参数。

将卷积后的结果通过最大池化层得到实体描述词向量 v_e^{des} , 计算公式如式(3)所示:

$$v_e^{\text{des}} = \max(g_e^1, g_e^2, \dots, g_e^j, \dots, g_e^m) \quad (3)$$

3) 实体类型词嵌入

通过 ConceptNet 发现实体词 e 隶属于 n 种实体类型, 表示为 $F_e = \{f_1, f_2, \dots, f_j, \dots, f_n\}$ 。首先将实体词 e 通过实体类型嵌入层 Emb_u 得到 f_j 的向量表示 v_{f_j} :

$$v_{f_j} = \text{Emb}_u(e) \quad (4)$$

然后使用注意力机制将 n 种实体类型融合为一个实体类型词向量 v_e^{type} :

$$v_e^{\text{type}} = \sum_j a_j v_{f_j} \quad (5)$$

其中: a_j 为注意力分数; P_j 是查询或文档事件范围向量表示与 f_j 实体类型向量表示的点积; 利用词袋模型^[15]对查询句或文档事件范围进行编码, W_{bow} 是一个参数矩阵。

$$a_j = \frac{\exp(P_j)}{\sum_i \exp(P_i)}, P_j = \left(\sum_i W_{\text{bow}} v_{t_i} \right) \cdot v_{f_j} \quad (6)$$

4) 触发词嵌入

通过 PredPatt 方法识别出查询句或文档事件范围中的事件触发词 t_{trg} , 并通过词嵌入层 Emb_{trg} 得到大小为 L 维的触发词词向量 $v_{\text{trg}}^{\text{emb}}$, 计算公式如式(7)所示:

$$v_{\text{trg}}^{\text{emb}} = \text{Emb}_{\text{trg}}(t_{\text{trg}}) \quad (7)$$

通过线性层对上述4种词嵌入进行融合, 得到事件实体最终的语义表示, 计算公式如式(8)所示:

$$v_e^{\text{sem}} = v_e^{\text{emb}} + W_e (v_e^{\text{des}} \oplus v_e^{\text{type}} \oplus v_{\text{trg}}^{\text{emb}})^T + b_e \quad (8)$$

其中: W_e 是维度大小为 $L \times 3L$ 的矩阵; b_e 是维度大小为 L 的向量。

1.5 融入事件实体知识的检索排序

本文使用基于交互的检索模型作为融合越南语事件实体知识的模型框架, 对于单个查询句 q 和单个文档事件范围 D , 基于交互的检索模型通过建立两者之间的词级交互矩阵^[16], 使用 q 与 D 之间的词向量相似度来衡量 q 与 D 之间的相似度。

基于交互的排序模型首先将 q 与 D 中的每个词 t 通过词嵌入层 Emb_w 得到大小为 L 维的词向量 v_t :

$$v_t = \text{Emb}_w(t) \quad (9)$$

然后基于得到的查询词向量和文档词向量生成交互矩阵 M , 其中每一个元素 M^{ij} 表示 q 中第 i 个词

向量与 D 中第 j 个词向量之间余弦相似度的值, 计算公式如下:

$$M^{ij} = \cos(v_{q_i}, v_{d_j}) \quad (10)$$

本文借鉴 XIONG 等^[17]提出的基于词与实体交互的匹配模型。该模型首先利用词袋模型处理查询句与文档中的词与标注好的实体, 然后使用已有的不同排序模型(如 BM25^[18]、TF-IDF^[19]等)分别计算查询词与文档词的排序得分、查询词与文档实体的排序得分、查询实体与文档词的排序得分和查询实体与文档实体的排序得分, 最后将这4种排序得分作为特征融入模型, 计算最终的查询句与文档的排序得分。然而, 该模型的特征提取效果过度依赖于已有的检索排序模型且提取过程繁琐。考虑到这一点, 本文在汉越跨语言新闻事件检索这一特定任务中, 使用交互矩阵 $M = \{M_{ww}, M_{we}, M_{ew}, M_{ee}\}$ 来衡量查询词或查询实体与文档词或文档实体之间的相似程度, 其中: M_{ww} 、 M_{we} 、 M_{ew} 、 M_{ee} 分别表征查询句中词与事件范围中词的交互($q^w - D^w$)、查询句中词与事件范围中事件实体的交互($q^w - D^e$)、查询句中事件实体与事件范围中词的交互($q^e - D^w$)和查询句中事件实体与事件范围中事件实体的交互($q^e - D^e$)。

将 q 与 D 中的每个词通过词嵌入层 Emb_w 后分别得到查询词向量 v_{q_i} 和文档事件范围词向量 v_{d_j} 。将查询词向量 v_{q_i} 视作一个整体向量矩阵 V_{T_q} , 将 V_{T_q} 通过卷积操作得到代表长度为 h_q 的 n -gram^[20] 向量 $g_{h_q}^i$:

$$g_{h_q}^i = \text{relu}(W_{\text{CNN}} \cdot V_{T_q}^{i:i+h_q} + b_{\text{CNN}}) \quad (11)$$

其中: W_{CNN} 与 b_{CNN} 是卷积核的2个参数。

同理, 得到代表长度为 h_D 的 n -gram 事件范围向量 $g_{h_D}^j$ 。

因此, M_{ww} 、 M_{we} 、 M_{ew} 、 M_{ee} 中各元素的计算方式如下:

$$M_{ww}^{ij} = \cos(g_{h_q}^i, g_{h_D}^j), M_{ee}^{ij} = \cos(v_{q_i}^{\text{sem}}, v_{d_j}^{\text{sem}}) \quad (12)$$

$$M_{ew}^{ij} = \cos(v_{q_i}^{\text{sem}}, g_{h_D}^j), M_{we}^{ij} = \cos(g_{h_q}^i, v_{d_j}^{\text{sem}})$$

交互矩阵 $M = \{M_{ww}, M_{we}, M_{ew}, M_{ee}\}$ 可以插入到任何基于交互的检索模型中, 本文使用 Conv-KNRM^[21] 这一基于交互的检索模型作为结合 M 的模型框架。以 M_{ww} 为例, Conv-KNRM 使用 K 个高斯分布从 M_{ww} 中提取排序特征 $\phi(M_{ww})$, 将每一个高斯分布 K_k 特征计算的结果为一个 soft-TF 值^[22], 因此, K 个高斯分布对 M_{ww} 处理之后生成了一个 K 维特征向量 $\phi(M_{ww}) = \{K_1(M_{ww}), K_2(M_{ww}), \dots, K_K(M_{ww})\}$, 计算公式如式(13)所示:

$$K_k(M_{ww}) = \sum_j \exp\left(-\frac{M_{ww}^{ij} - \mu_k}{2\delta_k^2}\right) \quad (13)$$

其中: μ_k 和 δ_k 分别表示第 k 个高斯分布的均值和方差。同理, 得到 $\phi(M_{we})$ 、 $\phi(M_{ew})$ 和 $\phi(M_{ee})$ 。

将 $\phi(M_{ww})$ 、 $\phi(M_{wc})$ 、 $\phi(M_{cw})$ 和 $\phi(M_{cc})$ 拼接成最终的排序特征 $\Phi(M)$:

$$\Phi(M) = \phi(M_{1,1}) \oplus \dots \oplus \phi(M_{h_q, h_D}) \oplus \dots \oplus \phi(M_{c,c}) \quad (14)$$

每一个 $\phi(M_{h_q, h_D})$ 包含三部分, 分别是查询句中 h_q -gram 的词与事件范围中 h_D -gram 的词之间的排序特征 $\phi(M_{ww}^{h_q, h_D})$ 、查询句中事件实体与事件范围中 h_D -gram 的词之间的排序特征 $\phi(M_{cw}^{h_q, h_D})$ 、查询句中 h_q -gram 的词与事件范围中事件实体之间的排序特征 $\phi(M_{wc}^{h_q, h_D})$ 。 $\phi(M_{h_q, h_D})$ 的表示形式如下:

$$\phi(M_{h_q, h_D}) = \phi(M_{ww}^{h_q, h_D}) \oplus \phi(M_{cw}^{h_q, h_D}) \oplus \phi(M_{wc}^{h_q, h_D}) \quad (15)$$

本文在得到最终排序特征 $\Phi(M)$ 的基础上, 利用排序学习^[23]得到查询句与事件范围最终的排序得分, 计算公式如式(16)所示:

$$f(q, D) = \tanh(\omega_r^T \Phi(M) + b_r) \quad (16)$$

其中: ω_r 和 b_r 是排序学习的参数; \tanh 为激活函数。

由于一个文档 d 中含有 f 个事件范围, 即 $d = \{D_1, D_2, \dots, D_f\}$, 因此对于查询句 q 和文档 d , 取 q 与 D_1, D_2, \dots, D_f 中最大的排序得分作为 q 与 d 最终的排序得分:

$$f(q, d) = \max_x \{ \tanh(\omega_r^T \Phi(M)_x + b_r) \} \quad (17)$$

其中, $\Phi(M)_x$ 表示查询句 q 与事件范围 D_x 计算所得的排序特征。

最后, 通过优化如下所示的合页损失函数^[24]对模型进行训练:

$$l = \sum_q \sum_{d^+, d^- \in G_q^{+-}} \max(0, 1 - f(q, d^+) + f(q, d^-)) \quad (18)$$

其中: G_q^{+-} 表示越南语新闻文档集中所有的文档; d^+ 表示与查询 q 相关的文档; d^- 表示与查询 q 不相关的文档。

本文通过反向传播优化模型参数, 在此过程中, 对实体词词嵌入、实体描述词嵌入、实体类型词嵌入、触发词词嵌入、词级交互和特征提取进行联合学习。

2 实验结果与分析

2.1 实验数据和实验参数设置

本文实验使用的汉越双语新闻数据集包含汉语数据集和越南语数据集, 其中, 汉语数据集用于构建汉语事件查询句, 越南语数据集用于构建越南语候选文档。为了对比本文模型在越南语单语新闻事件检索任务和汉越跨语言新闻事件检索任务中的性能差异, 利用越南语数据集人工构建与汉语事件查询句数量相同的越南语事件查询句。查询句与文档的相关性标签由人工标注, 1 表示相关, 0 表示不相关。实验中用到的汉语和越南语数据集中查询句子数与候选文档数的详细统计信息如表 1 所示。

表 1 数据集中查询句与候选文档数量

Table 1 Number of query sentences and candidate documents in the data set

数据集	类型	候选文档数量	查询句数量
越语事件数据集	训练集	20 150	3 000
	验证集	6 717	1 000
	测试集	6 717	1 000
汉语事件数据集	训练集	—	3 000
	验证集	—	1 000
	测试集	—	1 000

在本文实验中: 窗口大小 p 的值设定为 5; 实体词嵌入、实体描述词嵌入、实体类型词嵌入和触发词嵌入的维度 L 设定为 300; CNN 中滤波器个数为 128; 使用 Adam 优化器优化模型参数, 初始学习率设置为 0.001, 训练轮次为 100 轮。针对越南语事件查询句和越南语候选新闻文档所使用的事件实体识别模型为融合词典与对抗迁移的越南语事件实体识别模型^[12], 该模型经过训练之后, 在越南语新闻数据集上识别效果较好, 越南语事件实体识别的 F1 值达到 90.05%。实验中使用的翻译工具为目前汉越翻译性能较好的 Google 在线翻译软件。汉语事件查询句、越南语事件查询句和使用翻译工具翻译后的越南语事件查询句均是只包含一种新闻事件的单一事件查询句。本文使用的所有检索模型均在 Nvidia Tesla P100 GPU 上进行训练和测试, 在汉越双语新闻数据集上, 本文提出的融入事件实体知识的汉越跨语言新闻事件检索模型每一轮数据训练时间约为 0.8 h。

2.2 评价指标

在实验中使用 NDCG^[25] (Normalized Discounted Cumulative Gain) 和 MAP^[26] (Mean Average Precision) 作为评价指标。

$$DCG@k = \sum_{i=1}^k \frac{2^{r_i} - 1}{\text{lb}(1 + i)} \quad (19)$$

$$NDCG@k = \frac{DCG@k}{\text{idealDCG}@k} \quad (20)$$

$$AP = \frac{1}{R} \times \sum_{i=1}^R \frac{I}{\text{position}(I)} \quad (21)$$

其中: k 表示 k 个文档的集合; r_i 表示排序列表中第 i 个文档与查询句的相关度。将 $DCG@k$ 按照相关度从大到小对文档进行排序后即得到 $\text{idealDCG}@k$; R 表示与查询句相关的文档总个数; $\text{position}(I)$ 表示在检索结果列表中从前往后第 I 个相关文档在列表中的位置; MAP 是对多个查询语句的 AP 求均值。

2.3 对比实验与结果分析

将本文模型与基线模型进行比较, 验证本文方法的有效性。基线模型分为基于特征和基于神经网络 2 类。基于特征的检索模型包括 RankSVM^[27] 和 Coor-Ascent^[28] 2 种排序学习模型以及基于词的无监督检索模型 BM25; 基于神经网络的检索模型包括 ARC-I^[29]、ARC-II^[29]、CDSSM^[10]、MatchPyramid^[30]、DRMM^[31]、

K-NRM^[32]、Conv-KNRM、BERT-ATT-DBSCAN^[33]、ATER^[34]和 BERT-MaxS^[34],其中,ARC-I、CDSSM、BERT-ATT-DBSCAN和 ATER 是基于表示的检索模型,ARC-II、MatchPyramid、DRMM、K-NRM、Conv-KNRM 和 BERT-MaxS 模型是基于交互的检索模型。

1) RankSVM 模型将文档检索排序问题转化为文档对的分类问题,然后针对此分类问题利用 SVM 模型^[35]进行求解。

2) Coord-Ascent 是一种用于无约束优化问题的常见优化方法。该模型在检索过程中通过一系列的一维搜索来求解最终的多元目标函数。

3) BM25 模型是在融合 TF-IDF 特征的基础上计算查询句与文档相关性的模型,其先计算每个查询词与文档的相关度,再将得到的所有的词与文档的相关度进行加权求和,最后计算出最终的查询句与文档之间的相关度值。

4) ARC-I 模型使用 CNN 来进行文本匹配,其先将查询句和文档表示成 2 个定长的向量,再将 2 个向量拼接成一个向量整体,最后把向量整体输入多层感知器中,多层感知器的输出结果即为查询句与文档的匹配得分。

5) ARC-II 模型是 ARC-I 模型的扩展,其先将查询句与文档表示成向量,利用滑动窗口来选取词向量组,将词向量组作为一个固定单元进行卷积,将卷积后的结果作为查询句与文档相互作用的初步向量表示,再对其进行多次卷积和池化操作,最后将结果送入多层感知器中得到查询句与文档之间的匹配得分。

6) CDSSM 模型先将查询句与文档中的每一个词表征为词向量的形式,对设定的滑动窗口内的词向量进行卷积进而生成一个短语向量表示,再对短语向量表示进行池化操作。因为滑动窗口可以动态选取不同词向量,获取到句子中单词顺序信息,所以该模型能够对查询句与文档间的匹配关系进行更完整的描述。

7) MatchPyramid 模型构建查询句与文档间的匹配矩阵,使用卷积操作提取匹配矩阵中的特征,进而利用这些特征计算查询句与文档间的相似度。

8) DRMM 模型选取查询句中的一个词,将该词与文档中所有的词分别构成词组对,对于每一个词组对,使用余弦距离计算其相似度。该模型利用计算出的不同相似度构建直方图,因而可以有效区分查询句与文档之间相似的程度。

9) K-NRM 模型先将查询句和文档转化为向量形式,利用查询句向量和文档向量构建交互矩阵 M ,再引入 K 个核函数,通过核函数池化的方式计算查询句与文档的相似程度。

10) Conv-KNRM 模型相较于 K-NRM 模型,在进行核函数池化之前,分别利用滑动窗口对查询句向量和文档向量进行卷积操作,得到新的特征向量。在此基础上,对于查询句和文档的新特征向量,两两进行余弦相似度计算形成交互矩阵 M 。最后,使用 K 个核函数池化的方式计算出查询句与文档的相似程度。

11) BERT-ATT-DBSCAN 模型先将查询句和文档分别利用加入注意力机制的 BERT^[36]模型转换为向量形式,再利用 DBSCAN 聚类算法对查询句向量与文档向量进行聚类得到向量簇,通过计算查询句向量簇与文档向量簇的余弦相似度找到与查询句相关联的文档集合。

12) ATER 模型使用 BM25 算法计算出查询句与文档的相关度值,并使用 BERT 模型将查询句和文档分别转换为向量形式,利用编码器-解码器架构计算出查询句与文档之间的相关度值。在此基础上,将 2 种相关度值进行加权求和,得到查询句与文档最终的相关度得分。

13) BERT-MaxS 模型使用 BM25 模型计算查询句与文档的相关度值,并将文档切分为句子集合并分别与查询句进行拼接,使用基于 BERT 的排序模型计算查询句与每个文档句的相关度值。在此基础上,取最高相关度值与 BM25 算法计算得出的相关度值进行加权求和,得到查询句与文档最终的相关度得分。

在查询句为越南语事件查询句的情况下,对越南语候选文档进行检索排序。比较本文模型与基线模型在越南语数据集上检索性能的差异,实验结果如表 2 所示。

表 2 在越南语数据集上的越南语单语新闻事件检索性能

Table 2 Retrieval performance of Vietnamese monolingual news events on Vietnamese data set

模型	NDCG @1	NDCG @3	NDCG @5	NDCG @10	MAP
BM25	0.132 2	0.130 9	0.218 7	0.224 6	0.128 9
RankSVM	0.141 3	0.139 8	0.220 5	0.249 7	0.146 6
Coord-Ascent	0.160 6	0.168 4	0.257 7	0.297 3	0.159 0
ARC-I	0.184 4	0.204 5	0.201 1	0.299 4	0.185 3
ARC-II	0.213 4	0.227 9	0.251 8	0.316 5	0.210 7
CDSSM	0.186 9	0.214 1	0.206 3	0.306 4	0.190 4
MatchPyramid	0.225 4	0.236 7	0.258 2	0.345 2	0.210 3
DRMM	0.197 9	0.230 6	0.255 0	0.328 8	0.201 3
K-NRM	0.269 0	0.290 1	0.349 0	0.412 9	0.247 6
Conv-KNRM	0.338 4	0.394 7	0.449 4	0.538 6	0.332 1
BERT-ATT-DBSCAN	0.370 9	0.402 8	0.413 3	0.579 2	0.368 1
ATER	0.465 3	0.498 7	0.586 0	0.697 1	0.472 8
BERT-MaxS	0.673 1	0.694 1	0.709 7	0.800 3	0.659 8
本文模型	0.672 8	0.696 5	0.708 3	0.812 9	0.660 1

从表 2 的对比结果可以看出,本文模型检索性能优于其他检索模型。其中,基于神经网络的检索模型性能均优于基于特征的检索模型,相较于传统的 BM25 检索模型获得大幅度提升,在 NDCG@1、NDCG@3、NDCG@5、NDCG@10 和 MAP 评价指标上分别提升 0.540 6、0.565 6、0.489 6、0.588 3 和 0.531 2。与 Conv-KNRM 模型相比,本文模型在 NDCG@1、NDCG@3、NDCG@5、NDCG@10 和 MAP 评价指标上分别提升 98.82%、76.46%、57.61%、50.93%、98.77%,原因是本文在将 Conv-KNRM 作为检

索模型框架的基础上,把事件实体的分布式表示作为外部知识融入排序过程中,不仅进行查询句与文档间词与词之间的匹配,而且增加了查询句与文档之间的词与事件实体的匹配、事件实体与事件实体的匹配,同时利用事件触发词划定文档中的事件范围,缩小了查询句与文档匹配的空间,提升了匹配效率。与基线模型中性能最佳的BERT-MaxS相比,虽然BERT-MaxS在NDCG@1和NDCG@5指标上均略高于本文模型,但该检索模型是基于BERT模型构建的,模型训练所需数据量较大,模型参数较多,

完成一次检索过程的时间复杂度较高。

为探究事件实体知识的不同部分对模型检索性能的影响,进行越南语单语新闻事件检索的消融实验,在以下4种情况下对比检索性能:1)检索模型Conv-KNRM;2)在Conv-KNRM基础上分别加入4种词嵌入(实体词嵌入、实体描述词嵌入、实体类型词嵌入和触发词嵌入);3)在Conv-KNRM基础上划分文档事件范围;4)在Conv-KNRM基础上两两加入4种词嵌入。消融实验结果如表3所示。

表3 在越南语数据集上的消融实验结果

Table 3 Ablation experiment result on Vietnamese data set

模型	NDCG@1	NDCG@3	NDCG@5	NDCG@10	MAP
Conv-KNRM	0.338 4	0.394 7	0.449 4	0.538 6	0.332 1
+实体词嵌入	0.409 1	0.405 2	0.451 0	0.549 4	0.408 9
+实体描述词嵌入	0.497 3	0.489 9	0.499 0	0.587 3	0.489 3
+实体类型词嵌入	0.345 1	0.401 4	0.450 3	0.537 2	0.347 7
+触发词嵌入	0.355 8	0.387 1	0.438 5	0.549 0	0.354 3
+文档事件范围划分	0.457 2	0.443 8	0.479 6	0.568 8	0.457 9
+实体词嵌入和实体描述词嵌入	0.512 8	0.539 5	0.557 0	0.623 3	0.512 4
+实体词嵌入和实体类型词嵌入	0.448 3	0.459 2	0.466 1	0.557 5	0.439 6
+实体词嵌入和触发词嵌入	0.416 5	0.430 9	0.458 3	0.507 2	0.417 7
+实体、描述、类型和触发词嵌入	0.540 2	0.557 3	0.569 2	0.698 7	0.539 8
本文模型	0.672 8	0.696 5	0.708 3	0.812 9	0.660 1

从表3中可以看出:

1)在4种词嵌入类型中,实体描述词嵌入对于模型检索性能的提升最大,在融入实体描述词嵌入后,模型相较于Conv-KNRM在MAP评价指标上提升了0.157 2。

2)在只融入实体类型词嵌入的情况下,模型相较于Conv-KNRM在MAP评价指标上只提升0.015 6;但是在同时融入实体词嵌入和实体类型词嵌入的情况下,模型相较于Conv-KNRM在MAP评价指标上提升了0.107 5。由此可见,相较于只融入实体类型词嵌入的情况,只有把实体类型词嵌入和其他词嵌入一同融入时,模型性能才得到较大提升。

3)融入4种词嵌入(实体词嵌入、实体类型词嵌入、实体描述词嵌入和触发词嵌入)后模型的MAP评价指标相较于Conv-KNRM提升0.207 7,充分证明了通过知识图谱和事件触发词找到并融合成的事件实体语义表示可以有效提升查询句与文档的匹配性能。

4)在只对文档划分事件范围后,模型的MAP评价指标相较于Conv-KNRM提升0.125 8。

5)相较于同时融入4种词嵌入后的模型,本文模型在NDCG@1、NDCG@3、NDCG@5、NDCG@10和MAP评价指标上均提升较高。由此可见,对文档划分事件范围后,可以缩小模型匹配的空间,大幅提升模型性能。

在查询句为汉语事件查询句的情况下,对越南

语候选文档进行检索排序。比较本文模型与基线模型在汉越双语新闻数据集上的检索性能,实验结果如表4所示。

表4 在汉越双语新闻数据集上的汉越跨语言新闻事件检索性能

Table 4 Retrieval performance of Chinese-Vietnamese cross-language news events on Chinese-Vietnamese bilingual news data set

模型	NDCG@1	NDCG@3	NDCG@5	NDCG@10	MAP
BM25	0.087 2	0.079 1	0.081 1	0.103 4	0.084 3
RankSVM	0.074 3	0.079 9	0.085 4	0.097 4	0.072 1
Coor-Ascent	0.089 1	0.087 3	0.092 5	0.105 2	0.087 7
ARC-I	0.091 2	0.099 2	0.097 1	0.118 0	0.090 6
ARC-II	0.134 4	0.139 2	0.130 8	0.152 2	0.136 7
CDSSM	0.099 8	0.107 3	0.112 0	0.120 7	0.094 2
MatchPyramid	0.139 0	0.142 2	0.149 0	0.168 7	0.137 5
DRMM	0.122 6	0.128 9	0.149 8	0.157 0	0.122 5
K-NRM	0.198 7	0.209 3	0.213 8	0.253 1	0.197 5
Conv-KNRM	0.202 8	0.219 9	0.247 3	0.309 7	0.201 9
BERT-ATT-DBSCAN	0.294 0	0.341 7	0.398 6	0.427 3	0.285 4
ATER	0.400 1	0.398 2	0.473 5	0.583 1	0.408 5
BERT-MaxS	0.597 3	0.610 8	0.622 2	0.629 7	0.597 7
本文模型	0.662 9	0.689 7	0.693 2	0.809 6	0.659 3

从表 4 中可以看出:各模型检索性能相较于越南语单语新闻事件检索性能均有所降低。这是因为模型性能受所使用的翻译工具影响,翻译工具的翻译质量不高,会导致翻译生成的越南语事件查询句并不完全符合越南语正常的语法和句式表达。对于所有对比基线模型,翻译生成的越南语事件查询句质量低的情况严重影响了模型的检索性能。相较于只在越南语数据集上的 NDCG@1、NDCG@3、NDCG@5、NDCG@10 和 MAP 评价指标:RankSVM 分别降低 0.067 0、0.059 9、0.135 1、0.152 3 和 0.074 5;Conv-KNRM 分别降低 0.135 6、0.174 8、0.202 1、0.228 9 和 0.130 2;而本文模型性能降低相对较少,分别仅降低 0.009 9、0.006 8、0.015 1、0.003 3 和 0.000 8。这是因为本文模型依赖于使用多语言知识图谱和

事件触发词生成事件实体的语义知识表示,进而进行查询句与文档的词与词之间、词与事件实体之间和事件实体与事件实体之间的匹配排序。而事件实体往往可以被翻译工具翻译正确,从而降低了词与事件实体之间和事件实体与事件实体之间的匹配排序误差,因此,本文模型受翻译工具翻译质量的影响较小,能够较好地进行汉越跨语言新闻事件检索排序。

为探索查询句的翻译操作是否会影响事件实体知识的不同部分对模型检索性能所产生的促进作用,在汉越跨语言新闻事件检索时,对本文模型进行消融实验。实验设置与越南语单语新闻事件检索的消融实验设置相同,实验结果如表 5 所示。

表 5 在汉越双语新闻数据集上的消融实验结果

Table 5 Ablation experiment result on Chinese-Vietnamese bilingual news data set					
模型	NDCG@1	NDCG@3	NDCG@5	NDCG@10	MAP
Conv-KNRM	0.202 8	0.219 9	0.247 3	0.309 7	0.201 9
+实体词嵌入	0.408 3	0.404 6	0.450 2	0.547 9	0.407 8
+实体描述词嵌入	0.496 1	0.488 5	0.498 2	0.586 7	0.488 6
+实体类型词嵌入	0.297 6	0.329 8	0.344 1	0.398 5	0.296 2
+触发词嵌入	0.348 0	0.372 6	0.411 3	0.505 9	0.343 5
+文档事件范围划分	0.317 9	0.346 5	0.384 4	0.434 2	0.317 5
+实体词嵌入和实体描述词嵌入	0.510 6	0.537 2	0.556 1	0.619 8	0.510 7
+实体词嵌入和实体类型词嵌入	0.409 2	0.416 7	0.451 5	0.548 1	0.408 5
+实体词嵌入和触发词嵌入	0.415 2	0.429 8	0.453 6	0.500 1	0.413 4
+实体,描述,类型和触发词嵌入	0.538 8	0.549 6	0.560 7	0.698 3	0.538 6
本文模型	0.662 9	0.689 7	0.693 2	0.809 6	0.659 3

从表 5 中可以看出:经过查询句翻译之后,事件实体知识的各部分依然可以对模型的检索性能产生促进作用;在 4 种词嵌入类型中,实体描述词嵌入对于模型检索性能的提升最大,在融入实体描述词嵌入后,本文模型相较于 Conv-KNRM 在 NDCG@1、NDCG@3、NDCG@5、NDCG@10 和 MAP 评价指标上分别提升 0.293 3、0.268 6、0.250 9、0.277 0 和 0.286 7。

3 结束语

本文通过融入事件实体知识,提出一种新的汉越跨语言新闻事件检索模型。将汉语查询句翻译为越南语查询句并识别出候选文档中的事件触发词,基于触发词对文档划分事件范围,同时识别查询句中的事件触发词并使用事件实体识别方法识别出事件范围和查询句中的事件实体,基于知识图谱和触发词得到事件实体的知识表示,将事件实体知识融入基于交互的排序学习算法中对候选文档进行排序。实验结果表明,本文模型在汉越双语新闻数据集上相较于对比的基线模型取得了最佳的跨语言新闻事件检索效果。但是本文模型在汉越双语新闻数

据集上的检索性能相较于其在越南语数据集上的检索性能有所降低,原因在于模型检索的性能受到所使用翻译工具的翻译性能的限制。同时,其在查询句与文档匹配排序的过程中未考虑查询句中的事件触发词与文档中的事件触发词的歧义对查询句和文档匹配过程所造成的影响。后续将通过融入双语词典或引入双语词向量空间来辅助提升翻译工具的翻译效果,并且探索如何在查询句和文档匹配阶段进行事件触发词消歧,从而进一步提升模型匹配的性能。

参考文献

[1] G I A N G L T, H U N G V T, P H A P H C, et al. Building structured query in target language for Vietnamese-English cross language information retrieval systems [J]. International Journal of Engineering Research and Technology, 2015, 4(4): 146-151.

[2] E L A Y E B B, B O U N H A S I. Arabic cross-language information retrieval: a review[J]. ACM Transactions on Asian and Low-Resource Language Information Processing, 2016, 15(3): 1-44.

[3] R A H I M I R, M O N T A Z E R A L G H A E M A, S H A K E R Y A. An axiomatic approach to corpus-based cross-language

- information retrieval[J]. Information Retrieval Journal, 2020, 23(3): 191-215.
- [4] PASSALIS N, TEFAS A. Entropy optimized feature-based bag-of-words representation for information retrieval[J]. IEEE Transactions on Knowledge and Data Engineering, 2016, 28(7): 1664-1677.
 - [5] YAN R, SONG Y, WU H. Learning to respond with deep neural networks for retrieval-based human-computer conversation system [C]//Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA: ACM Press, 2016: 55-64.
 - [6] XIONG C Y, CALLAN J, LIU T Y. Bag-of-entities representation for ranking[C]//Proceedings of 2016 ACM International Conference on the Theory of Information Retrieval. New York, USA: ACM Press, 2016: 181-184.
 - [7] DALTON J, DIETZ L, ALLAN J. Entity query feature expansion using knowledge base links[C]//Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA: ACM Press, 2014: 365-374.
 - [8] LIU X T, FANG H. Latent entity space: a novel retrieval approach for entity-bearing queries [J]. Information Retrieval Journal, 2015, 18(6): 473-503.
 - [9] ENSAN F, BAGHERI E. Document retrieval model through semantic linking[C]//Proceedings of the 10th ACM International Conference on Web Search and Data Mining. New York, USA: ACM Press, 2017: 181-190.
 - [10] SHEN Y L, HE X D, GAO J F, et al. A latent semantic model with convolutional-pooling structure for information retrieval[C]//Proceedings of the 23rd ACM International Conference on Information and Knowledge Management. New York, USA: ACM Press, 2014: 101-110.
 - [11] DAI Z Y, XIONG C Y, CALLAN J, et al. Convolutional neural networks for soft-matching N-grams in ad-hoc search[C]//Proceedings of the 11th ACM International Conference on Web Search and Data Mining. New York, USA: ACM Press, 2018: 126-134.
 - [12] 薛振宇, 线岩团, 余正涛, 等. 融合词典与对抗迁移的越南语事件实体识别[J]. 计算机工程, 2022, 48(3): 107-114, 145.
XUE Z Y, XIAN Y T, YU Z T, et al. Vietnamese event entity recognition combining dictionary and adversarial transfer[J]. Computer Engineering, 2022, 48(3): 107-114, 145. (in Chinese)
 - [13] ZHANG S, RUDINGER R, VAN DURME B. An evaluation of PredPatt and Open IE via stage 1 semantic role labeling[C]//Proceedings of the 12th International Conference on Computational Semantics. Washington D. C., USA: IEEE Press, 2017: 1-5.
 - [14] RODOSTHENOUS C, LYDING V, SANGATI F, et al. Using crowdsourced exercises for vocabulary training to expand conceptnet[C]//Proceedings of the 12th Language Resources and Evaluation Conference. Marseille, France: European Language Resources Association, 2020: 307-316.
 - [15] ZHANG Y, JIN R, ZHOU Z H. Understanding bag-of-words model: a statistical framework [J]. International Journal of Machine Learning and Cybernetics, 2010, 1(1/2/3/4): 43-52.
 - [16] 陈鑫, 李伟康, 洪宇, 等. 面向问句复述识别的多卷积自交互匹配方法研究[J]. 中文信息学报, 2019, 33(10): 99-108, 118.
CHEN X, LI W K, HONG Y, et al. A multi-convolution self-interaction method for question paraphrase identification[J]. Journal of Chinese Information Processing, 2019, 33(10): 99-108, 118. (in Chinese)
 - [17] XIONG C Y, CALLAN J, LIU T Y. Word-entity duet representations for document ranking[C]//Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA: ACM Press, 2017: 763-772.
 - [18] KADHIM A I. Term weighting for feature extraction on Twitter: a comparison between BM25 and TF-IDF [C]//Proceedings of International Conference on Advanced Science and Engineering. Washington D. C., USA: IEEE Press, 2019: 124-128.
 - [19] KIM S W, GIL J M. Research paper classification systems based on TF-IDF and LDA schemes[J]. Human-Centric Computing and Information Sciences, 2019, 9(1): 1-21.
 - [20] CAO S, LU W, ZHOU J, et al. cw2vec: learning Chinese word embeddings with stroke n-gram information [C]//Proceedings of the AAAI Conference on Artificial Intelligence. [S. l.]: AAAI Press, 2018: 5053-5061.
 - [21] HOFSTÄTTER S, REKABSAN N, EICKHOFF C, et al. On the effect of low-frequency terms on neural-IR models[C]//Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA: ACM Press, 2019: 1137-1140.
 - [22] SUN Y F. A comparative evaluation of string similarity metrics for ontology alignment[J]. Journal of Information and Computational Science, 2015, 12(3): 957-964.
 - [23] 杨晋吉, 胡波, 王欣明, 等. 一种知识图谱的排序学习个性化推荐算法[J]. 小型微型计算机系统, 2018, 39(11): 2419-2423.
YANG J J, HU B, WANG X M, et al. Personalized recommendation algorithm for learning to rank by knowledge graph [J]. Journal of Chinese Computer Systems, 2018, 39(11): 2419-2423. (in Chinese)
 - [24] RAZZAK I, KHAN T M. One-class support tensor machines with bounded hinge loss function for anomaly detection[C]//Proceedings of International Joint Conference on Neural Networks. Washinton D. C., USA: IEEE Press, 2020: 1-8.
 - [25] JIN X B, GENG G G, XIE G S, et al. Approximately optimizing NDCG using pair-wise loss [J]. Information Sciences, 2018, 453: 50-65.
 - [26] GROS D, HABERMANN T, KIRSTEIN G, et al. Anaphora resolution: analysing the impact on mean average precision and detecting limitations of automated approaches [J]. International Journal of Information Retrieval Research, 2018, 8(3): 33-45.
 - [27] PENG L, LU G Q. Research on query term expansion based on RankSVM and LDA model [J]. Journal of Physics: Conference Series, 2020, 1684(1): 1-10.
 - [28] SAYED M F, OARD D W. Jointly modeling relevance and sensitivity for search among sensitive content [C]//Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA: ACM Press, 2019: 615-624.

(下转第 291 页)

(上接第 282 页)

- [29] HU B, LU Z, LI H, et al. Convolutional neural network architectures for matching natural language sentences[C]//Proceedings of NIPS' 14. Montreal, Canada: Neural Information Processing Systems Foundation, 2014:2042-2050.
- [30] TANG Z L, LI J. Jointly considering Siamese network and MatchPyramid network for text semantic matching[J]. IOP Conference Series: Materials Science and Engineering, 2019, 490:1-10.
- [31] GUO J F, FAN Y X, AI Q Y, et al. A deep relevance matching model for ad-hoc retrieval[C]//Proceedings of the 25th ACM International Conference on Information and Knowledge Management. New York, USA: ACM Press, 2016:55-64.
- [32] XIONG C Y, DAI Z Y, CALLAN J, et al. End-to-end neural ad-hoc ranking with kernel pooling[C]//Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA: ACM Press, 2017:55-64.
- [33] 曹旭友,周志平,王利,等. 基于 BERT+ATT 和 DBSCAN 的长三角专利匹配算法[J]. 信息技术, 2020, 44(3): 1-5, 12.
- CAO X Y, ZHOU Z P, WANG L, et al. Yangtze River delta patent matching algorithm based on BERT+ATT and DBSCAN[J]. Information Technology, 2020, 44(3): 1-5, 12. (in Chinese)
- [34] 胡瀚. 基于 BERT 的神经排序模型研究[D]. 武汉: 华中师范大学, 2020.
- HU H. Research on BERT-based neural ranking models [D]. Wuhan: Central China Normal University, 2020. (in Chinese)
- [35] 周艳平, 李金鹏, 宋群豹. 一种基于 SVM 及文本密度特征的网页信息提取方法[J]. 计算机应用与软件, 2019, 36(10): 251-255, 261.
- ZHOU Y P, LI J P, SONG Q B. A Web page information extraction method based on SVM and text density features [J]. Computer Applications and Software, 2019, 36(10): 251-255, 261. (in Chinese)
- [36] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[EB/OL]. [2021-06-10]. <https://arxiv.org/abs/1810.04805>.

编辑 金胡考