计算机科学
COMPUTER SCIENCE

基于跨层级多视角特征的多语言事件探测

张志远, 张维彦, 宋雨秋, 阮彤

引用本文

张志远, 张维彦, 宋雨秋, 阮彤. 基于跨层级多视角特征的多语言事件探测[J]. 计算机科学, 2024, 51(5): 208-215.

ZHANG Zhiyuan, ZHANG Weiyan, SONG Yuqiu, RUAN Tong. Multilingual Event Detection Based on Cross-level and Multi-view Features Fusion [J]. Computer Science, 2024, 51(5): 208-215.

相似文章推荐（请使用火狐或 IE 浏览器查看文章）

Similar articles recommended (Please use Firefox or IE to view the article)

基于提示学习的生成式医疗对话理解方法

Prompt Learning-based Generative Approach Towards Medical Dialogue Understanding

计算机科学, 2024, 51(5): 258-266. https://doi.org/10.11896/jsjkx.230300007

面向中文电子病历的多粒度医疗实体识别

Multi-granularity Medical Entity Recognition for Chinese Electronic Medical Records

计算机科学, 2021, 48(4): 237-242. https://doi.org/10.11896/jsjkx.200100036

融入语言模型和注意力机制的临床电子病历命名实体识别

Clinical Electronic Medical Record Named Entity Recognition Incorporating Language Model and Attention Mechanism

计算机科学, 2020, 47(3): 211-216. https://doi.org/10.11896/jsjkx.190200259

数据集分类可用性评估的置信区间方法

Confidence Interval Method for Classification Usability Evaluation of Data Sets

计算机科学, 2019, 46(1): 78-85. https://doi.org/10.11896 / j.issn.1002-137X.2019.01.012

三维网格模型增量式聚类检索

3D Mesh Model Retrieval Using Incremental Clustering

计算机科学, 2011, 38(11): 248-251.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230200131

基于跨层级多视角特征的多语言事件探测

张志远　张维彦　宋雨秋　阮　彤
华东理工大学信息工程与科学学院　上海 200237
(1127657064@qq.com)

摘　要　多语言事件探测任务的目标是将多种语言的新闻文档集合组织成不同的关键事件,其中每个事件可以包含不同语言的新闻文档。该任务有助于各种下游任务应用,如多语言知识图谱构建、事件推理、信息检索等。目前,多语言事件探测主要分为先翻译再事件探测与先单语言检测再跨多种语言对齐两种方法,前者依赖翻译的效果,后者需要为每种语言单独训练模型。为此,提出了一种名为基于跨层级多视角特征融合的多语言事件探测方法,端到端地进行多语言事件探测任务。该方法从不同层级利用文档的多视角特征,获得了高可靠性的多语言事件探测结果并提升了低资源语言事件探测的泛化性能。在9种语言混合的新闻数据集上进行的实验表明,所提方法的BCubed F1值提升了4.63%。

关键词:多语言预训练模型;多语言事件探测;新闻文档聚类;加权相似度;增量聚类
中图分类号　TP391

Multilingual Event Detection Based on Cross-level and Multi-view Features Fusion

ZHANG Zhiyuan,ZHANG Weiyan,SONG Yuqiu and RUAN Tong
School of Information Science and Engineering,East China University of Science and Technology,Shanghai 200237,China

Abstract　The goal of the multilingual event detection task is to organize a collection of news documents in multiple languages into different key events,where each event can include news documents in different languages. This task facilitates various downstream task applications,such as multilingual knowledge graph construction,event reasoning,information retrieval,etc. At present,multilingual event detection is mainly divided into two methods:translation first and then event detection,and single language detection first and then alignment across multiple languages. The former relies on the effect of translation while the latter requires a separate training model for each language. To this end,this paper proposes a multilingual event detection method based on cross-level multi-view feature fusion,which performs end-to-end multilingual event detection tasks. This method uses the multi-view features of documents from different levels to obtain high reliability. It improves the generalization performance of low-resource language event detection. Experiments on a news dataset with a mixture of nine languages show that the proposed method improves the BCubed F1 value by 4.63%.

Keywords　Multilingual pre-training model,Multilingual event detection,News documents clustering,Weighted similarity,Incremental clustering

1　引言

多语言事件探测任务(Multilingual Event Detection,MED)旨在对多语言新闻文本进行主题归纳,并将其按主题(如女排夺冠、拜登当选等)进行事件分类。MED是自然语言处理(NLP)的重要任务,有助于各种下游任务应用,如多语言知识图谱构建[1]、事件推理[2]、信息检索[3]等。该任务输入多语言新闻集,输出若干个关键事件,其中每个关键事件是一组关于特定主题的多语言新闻文档集,具体示例如图1所示。

传统的事件探测方法大多是基于词频-逆文本频率(Term Frequency-Inverse Document Frequency,TF-IDF)统计关键词词频的思想。基于本文的调研[4],事件探测的关键词往往是实体。但是,在多语言场景下,同义实体表述不一,极易造成事件探测错误。

为解决上述多语言事件探测面临的问题,研究者们提出了一系列解决方法。这些方法主要分为两类:1)先将所有新闻文档翻译为同一种语言,再进行事件探测[5]。这类基于翻译的方法高度依赖于翻译质量,具体而言,当出现突发新闻时,机器翻译会产生超出词表(Out Of Vocabulary)的困扰,造成翻译失真[6],从而导致事件探测效果不佳。2)首先利用多个单语言模型对新闻文档进行单语言事件探测[7-8],再利用多语言预训练模型对前一步获取的单语言关键事件进行跨多种语言对齐。这类两阶段的管道方法(Pipeline Method)避免了翻译失真带来的噪音,但需要训练多个单语种模型以及一个跨多种语言对齐模型,资源消耗大。此外,还存在泛化性能差的缺陷,在低资源语言上尤为严重。

为此,本文提出了一种基于跨层级多视角特征融合的多语言事件探测方法(Cross-level Multi-view Features Fusion

到稿日期:2023-02-19　返修日期:2023-06-21
通信作者:阮彤(ruantong@ecust.edu.cn)

Method for Multilingual Event Detection，CMFF-MED)，由 3 个模块构成：多语言实体语义匹配模块、多视角特征提取模块和多特征事件探测模块。首先，多语言实体语义匹配模块利用知识引导的跨语言语义匹配模型对不同语言间和同一语言内的同义实体进行对齐，解决了同义实体表述不一的难题。其次，多视角特征提取模块(包括多语言同义实体层级视角、多语言句子层级视角和时间戳层级视角)利用 3 个层级视角模型获取多个不同语言的特征，可以避免使用大量单语言模型来获取特征，缓解了计算资源的过度消耗，同时提升了低资源语言的泛化性能。最后，多特征事件探测模块利用加权相似度特征函数来融合多视角特征，从而获得高可靠性的多语言事件探测结果。

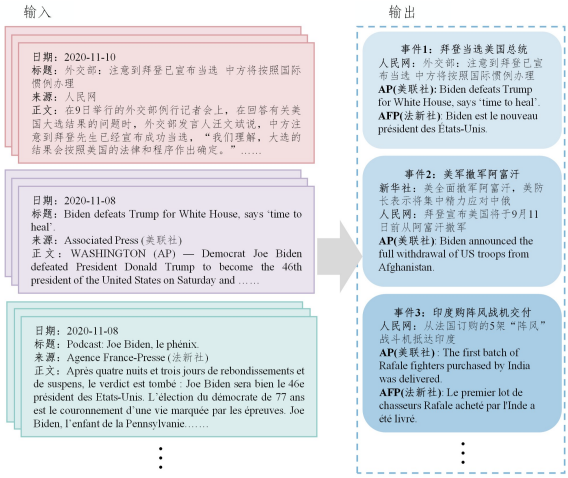


图 1 多语言事件探测示例

Fig. 1 Example of multilingual event detection

总的来说，本文的主要贡献有以下 3 点：

1)提出了一种基于跨层级多视角特征融合的多语言事件探测方法，用于解决多语言事件探测任务，称为 CMFF-MED。该方法利用多视角特征提取技术极大地改善了两阶段管道方法存在模型过多的问题，同时提高了低资源语言的泛化能力。

2)提出了一种基于知识引导的跨语言语义匹配模型。该模型利用外部跨语言知识基于孪生网络来增强实体语义表示，从而获得多语言同义实体集，解决了多个不同语言间和同一语言内的同义实体表述不一的难题。

3)实验结果表明，CMFF-MED 在多语言新闻数据集上显著优于最先进的方法，相较于现有的最先进方法，BCubed F1 提升了 4.63%。

2 相关工作

本文的相关工作包含两个部分，即单语言事件探测方法和多语言事件探测方法。

单语言事件探测方法可分为基于统计的方法和基于神经网络的方法。1)基于统计的方法：Laban 等[9]提出了 NewsLens，其通过 TF-IDF 提取关键词后以关键词表征文章，并通过社区检测算法将这些关键词对齐，从而构建不同的事件；Staykovski 等[10]将 TF-IDF 特征与段落向量(Paragraph Vector)特征相结合，形成了新的聚类方法。2)基于深度神经网络的方法：Li 等[11]以卷积编码器提取新闻特征，提出利用重构损失和聚类损失训练网络来提高聚类效果。然而受限于聚类算法，这些方法难以处理在线新闻文档流。Guo 等[12]利用深度神经网络将文章编码后，采用半监督学习的方法，对于首次探测时不确定的文章，稍后再探测，提升了聚类的效果。Saravanakumar 等[13]为 BERT[14]的输入添加了实体标签隐向量，增强了模型对实体的感知，并结合统计特征进行英语的事件探测。Zhang 等[15]基于 SBERT[16](Sentence-BERT)进行改进，结合提出的峰值短语图网络进行了细粒度的事件探测。

多语言事件探测方法分为基于翻译的方法和先做单语言事件探测再将单语言的结果相链接的管道方法。1)基于翻译的方法：Yang 等[5]提出的模型利用双语词典获得翻译对，通过最小化翻译对的欧氏距离来完成多语言事件探测。然而，低资源的语言可能缺乏双语词典，为此，Liu 等[6]设计了一种基于上下文的翻译方法来构建不同语言之间的词汇映射，利用该词汇映射改善多语言事件探测的效果。但由于多语言通过翻译实现，因此这些方法对于突发事件中的超出词典现象无法起到较好的效果。2)先做单语言事件探测，再将单语言的结果相链接的两阶段方法：Yang 等[17]针对汉语、越南语分别训练神经主题模型，得到通用的表征后进行聚类，从而发现双语关键事件。Miranda 等[7]针对多语言新闻集进行实验，首先根据统计特征将文档集组织成单语言的事件集合，然后通过多语言词向量链接不同的单语事件，得到跨多种语言事件。Linger 等[8]在此基础上使用了更先进的多语言上下文向量，并引入了一种名为重播的策略，在探测后期对不确定的事件重新探测，改进了多语言事件探测的流程。

3 跨层级多视角特征融合的 MED 方法

本章详细论述多语言事件探测任务定义及所提出的 CMFF-MED 方法的细节。

3.1 任务定义

定义 1(多语言同义实体集) 本文中的多语言包含英语(en)、法语(fr)、汉语(zh)、俄语(ru)、西班牙语(es)、德语(de)、意大利语(it)、斯洛文尼亚语(sl)和克罗地亚语(hr)9 种语言。同义实体既存在于不同语言间，也存在于同一语言中，如图 2 所示。多语言同义实体集 E_{sy}^{i} 指多个语言表述的同一真实世界特定实体的实体集，其中 $E_{sy}^{i}=\{e_{l_1}^{1},e_{l_1}^{2},\cdots,e_{l_1}^{m_1},e_{l_2}^{1},e_{l_2}^{2},\cdots,e_{l_2}^{m_2},e_{l_n}^{1},e_{l_n}^{2},\cdots,e_{l_n}^{m_n}\}$ $(2\leqslant n\leqslant 9, l_n\in\{en,fr,zh,ru,es,de,it,sl,hr\}, m_n\geqslant 0)$。

定义 2(多语言实体集) 多语言实体集指多语言同义实体集的集合 $E_{SY}=\{E_{sy}^{1},E_{sy}^{2},\cdots,E_{sy}^{n}\}$。

定义 3(关键事件) 关键事件 c_i 是一组与某真实世界特定主题相关的多语言新闻文档 d 集合 $c_k=\{d_1,d_2,\cdots,d_n\}$。

定义 4(多语言事件探测) 给定一个多语言新闻文档集 $D=\{d_1,d_2,\cdots,d_m\}$，利用模型得到一组关键事件集合 $C=\{c_1,c_2,\cdots,c_k\}$，其中每个关键事件 $c_k\in D$，所有关键事件互不重叠(即对任意 $i\neq j, c_i\cap c_j=\varnothing$)。

定义 5(关键事件池) 本文所提方法进行增量式的多语言事件探测，关键事件池 C_p 用于保存其中间结果。初始化时 $C_p=\{c_1\}$(其中 c_1 包括第一篇新闻文档)；在探测过程中，关键事件池中的关键事件数量将随着文档流的输入而逐渐增加；在探测过程结束后，关键事件池包含输出的关键事件集合(即 $C_p=C=\{c_1,c_2,\cdots,c_k\}$)。



注：任一多语言同义实体集包含两种以上语言，每种语言包含一个以上实体。

图 2 多语言同义实体集

Fig. 2 Multilingual synonymous entity set

3.2 算法框架

本文提出 CMFF-MED 方法来解决多语言事件探测任务，该方法包括 3 个模块：1）多语言实体语义匹配模块，利用知识引导的跨语言语义匹配模型对多个不同语言的同义实体进行语义匹配。2）多视角特征提取模块，从 3 个视角来提取多语言新闻文档的特征，具体包括多语言同义实体层级视角、多语言句子层级视角和时间戳层级视角。3）多特征事件探测模块，首先利用加权相似度特征函数自适应融合多语言新闻文档的多视角特征，检索与其最相似的关键事件；其次利用关键事件探测模型来判断当前文档是否属于所选的最相似关键事件，反之创建新关键事件。

整体框架如图 3 所示。

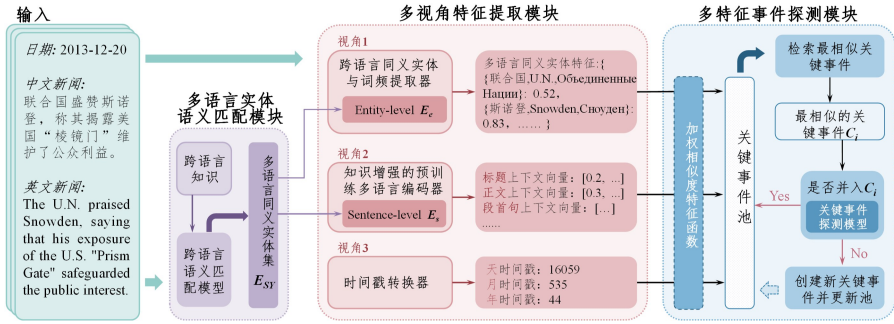


图 3 基于跨层级多视角特征融合的多语言事件探测框架示意图

Fig. 3 Framework of CMFF-MED

3.3 多语言实体语义匹配模块

输入外部多语言文档集 $\boldsymbol{D}_{ex}$[18]，利用多语言实体语义匹配模块，输出多语言实体集。其核心思想是通过模型对不同语言的同义实体进行跨多种语言对齐。具体而言，首先输入 $\boldsymbol{D}_{ex}$，利用多语言实体识别模型，输出多个不同语言的实体集合。其次，利用跨语言知识引导的跨语言语义匹配模型，输出多语言同义实体集。最后，全面利用多语言同义实体集中跨越不同层级的信息，其中包括实体层级信息和句子层级信息，分别供多语言实体对齐特征提取器和知识增强的预训练多语言编码器使用。

首先，输入外部多语言文档集 $\boldsymbol{D}_{ex}$，利用多头自注意力机制的多语言预训练模型（Multilingual Bidirectional Encoder Representations from Transformers，mBERT）配合条件随机场（Conditional Random Field，CRF）作为多语言实体识别模型，输出 $\boldsymbol{D}_{ex}$ 中包含的多种语言的实体集合 $\boldsymbol{E}_{ex}$，该过程表示如式（1）所示：

$$\boldsymbol{E}_{ex}=Model_{en}(\{v_{ex}^1,v_{ex}^2,\cdots,v_{ex}^m\}) \tag{1}$$

其中，$\boldsymbol{v}_{ex}^1$ 为文档 d_{ex}^i 通过分词器获得的文档令牌向量，d_{ex}^i 为一个外部文档，$\boldsymbol{D}_{ex}=\{d_{ex}^1,d_{ex}^2,\cdots,d_{ex}^m\}$，$Model_{en}$ 为多语言实体识别模型。

其次，给定 $\boldsymbol{E}_{ex}$，利用跨语言语义匹配模型，来获取多语言实体集 $\boldsymbol{E}_{sy}$。本文基于 mBERT 搭建孪生网络作为跨语言语义匹配模型，该模型的输入为 $\boldsymbol{E}_{ex}$ 中的两个实体，输出为它们是否指向同一真实世界实体，若是，则合并其所属的多语言同义实体集，如图 3 所示。初始化时，将每个实体单独视作一个多语言同义实体集，将 $\boldsymbol{E}_{ex}$ 中的实体成对输入模型，即可获得 $\boldsymbol{E}_{sy}$。本文利用 MUSE[19] 提供的多语言并行词典对跨语言语义匹配模型进行训练，使其获得额外的实体知识，具体细节见 4.1 节。

3.4 多视角特征提取模块

给定待探测的多语言新闻文档，利用多视角特征提取模块，输出多语言同义实体特征、多语言上下文特征和时间戳特征。其核心思想是针对不同语言的新闻文档提取统一的文档特征，避免使用大量单语言模型来获取特征，缓解了计算资源的过度消耗，同时提高了低资源语言的泛化性能。具体而言，输入一篇待探测的多语言新闻文档，首先利用多语言实体对齐特征提取器，输出多语言实体对齐特征。其次利用实体增强的 mBERT 编码器，输出多语言上下文特征。最后利用时间戳转换器，输出时间戳特征。

1）基于多语言同义实体层级视角，提出多语言实体对齐特征提取器 E_e。给定待探测的多语言新闻文档，利用基于编辑距离的字符匹配模型，输出多语言同义实体特征。其核心思想是利用多语言同义实体集来统一不同语言文档的关键词。具体而言，首先，输入待探测的多语言新闻文档 d_i，利用多语言实体识别模型获取其包含的实体集 E_{d_i}。其次，利用基于编辑距离的字符匹配模型，来辅助查询 E_{sy} 与 E_{d_i} 的交集，同时对抗错误拼写、译名差别等噪声数据。将查询到的 E_{sy}^i 作为该文档多语言实体对齐特征向量 $\boldsymbol{h}_e$ 的键。最后，计算 $\boldsymbol{h}_e$ 的每个键的 TF-IDF 分数，作为键对应的值。该分数为其包含的单语实体的 TF-IDF 分数之和，如式（2）所示：

$$score_{tf}^{E_{sy}^i}=\sum_{j=1}^{m}(score_{tf}^{e_j}) \tag{2}$$

其中，e_j 为 E_{sy}^i 中包含的 m 个实体。记录这些由多语言同义

实体集和 TF-IDF 分数组成的键值对，即为文档d_i的多语言实体对齐特征向量$\boldsymbol{h}_e$。

2）基于多语言句子层级视角，提出知识增强的预训练多语言编码器E_s。给定待探测的多语言新闻文档，利用实体知识增强的 mBERT 模型，输出多语言上下文向量特征。其核心思想是利用多语言预训练模型来统一不同语言文档的上下文向量表示。具体而言，首先，为了将实体知识更好地融入 mBERT 模型，使模型更适合新闻领域的事件探测，本模型利用D_{ex}对 mBERT 模型进行额外的预训练，具体细节见 4.2 节；之后，输入待探测的多语言新闻文档d_i，利用实体知识增强的 mBERT 模型对文档进行编码，即可获得文档对应的上下文向量。针对新闻文档而言，标题能够概括事件主题，首段对事件进行总体描述，每个段首句统领该段，这些要素对于事件探测具有重要意义。为了提高文档上下文特征的质量与多样性，对于每个文档，$\boldsymbol{h}_c$分为 4 种上下文向量：新闻标题的上下文向量、新闻首段的上下文向量、新闻所有段首句的上下文向量，以及新闻正文的上下文向量。如式(3)所示：

$$\boldsymbol{h}_c = mBERT_{en}(\{\boldsymbol{v}_i^{ti}, \boldsymbol{v}_i^{fp}, \boldsymbol{v}_i^{fs}, \boldsymbol{v}_i^{te}\}) \tag{3}$$

其中，$\boldsymbol{v}_i^{ti}$，$\boldsymbol{v}_i^{fp}$，$\boldsymbol{v}_i^{fs}$以及$\boldsymbol{v}_i^{te}$分别为文档d_i的标题、首段、所有段首句以及正文的文档令牌向量。

3）基于时间戳层级视角，提出时间戳转换器E_{ts}。首先，输入待探测的多语言新闻文档d_i，提取其中的日期$date_{d_i}$，将其转化为时间戳ts_{d_i}。对于多语言事件探测任务而言，时、分、秒过于短暂，意义不大，因此，将时间戳转换为以天为单位的值$ts_{d_i}^{\mathrm{d}}$、以月为单位的值$ts_{d_i}^{\mathrm{m}}$和以年为单位的值$ts_{d_i}^{\mathrm{y}}$，从而获得时间戳特征$\boldsymbol{h}_t = \{ts_{d_i}^{\mathrm{d}}, ts_{d_i}^{\mathrm{m}}, ts_{d_i}^{\mathrm{y}}\}$。

将 3 个视角获取的特征向量相连接，即获得多视角文档特征向量$\boldsymbol{h}_d = \{\boldsymbol{h}_e, \boldsymbol{h}_c, \boldsymbol{h}_t\}$。

3.5 多特征事件探测模块

给定文档的多视角文档特征向量$\boldsymbol{h}_d$，利用多特征事件探测模块，输出多语言事件探测结果。其核心思想是加权融合多视角特征，自适应地利用不同重要性的特征，从而增量式地获得高可靠性的多语言事件探测结果。具体而言，1）初始化关键事件池C_p；2）检索最相似事件，输入该文档的多视角文档特征向量$\boldsymbol{h}_d$与关键事件池C_p中的每个关键事件的多视角特征$\boldsymbol{h}_{c_j}$，利用加权相似度特征函数，检索最相似的关键事件c_{top}；3）输入文档的多视角文档特征向量$\boldsymbol{h}_d$以及c_{top}的特征向量$\boldsymbol{h}_{c_{\mathrm{top}}}$，利用关键事件探测模型$m_{\mathrm{de}}$，输出该文档是否合并至$c_{\mathrm{top}}$；4）根据模型输出更新$c_{\mathrm{top}}$的特征，或创建新的关键事件。

1）首先对关键事件池C_p进行初始化。在多语言事件探测任务开始时，第一篇文档用于初始化第一个关键事件c_1。具体来说，c_1的多视角特征$\boldsymbol{h}_{c_1}$继承自第一篇文档的多视角文档特征向量$\boldsymbol{h}_d$，其中$\boldsymbol{c}_e = \boldsymbol{h}_e$，$\boldsymbol{c}_c = \boldsymbol{h}_c$，$c_t = \{\boldsymbol{h}_t, \boldsymbol{h}_t, \boldsymbol{h}_t\}$，之后如果有新的文档合并入$c_1$，则根据式(4)－式(6)对$\boldsymbol{h}_{c_1}$进行更新：

$$\boldsymbol{c}_e^{\mathrm{new}} = \boldsymbol{h}_e \cup \boldsymbol{c}_e^{\mathrm{old}} \tag{4}$$

$$\boldsymbol{c}_c^{\mathrm{new}} = (\boldsymbol{c}_c^{\mathrm{old}} * n + \boldsymbol{h}_c)/(n+1) \tag{5}$$

$$\boldsymbol{c}_t^{\mathrm{new}} = \left\{ \max(\boldsymbol{h}_t, \boldsymbol{c}_{t_1}^{\mathrm{old}}), \min(\boldsymbol{h}_t, \boldsymbol{c}_{t_2}^{\mathrm{old}}), \frac{\boldsymbol{c}_{t_3}^{\mathrm{old}} * n + \boldsymbol{h}_t}{n+1} \right\} \tag{6}$$

其中，n为更新前c_1包含的文档数量，old 和 new 作为上标的变量分别为c_1更新前后对应的特征向量。

2）其次，加权相似度特征函数，检索最相似的关键事件c_{top}，其中不同视角的特征，有着不同的相似度计算方式。具体来说，对于多语言实体对齐特征，通过式(7)计算其相似度，其中$\boldsymbol{h}_e$与$\boldsymbol{c}_e$分别是文档与关键事件的多语言实体对齐特征。

$$sim_e(\boldsymbol{h}_e, \boldsymbol{c}_e) = \boldsymbol{h}_e \cdot \boldsymbol{c}_e \tag{7}$$

对于多语言上下文向量特征，通过式(8)计算其相似度，其中$\boldsymbol{h}_c$与$\boldsymbol{c}_c$分别是文档与关键事件的上下文向量特征。

$$sim_c(\boldsymbol{h}_c, \boldsymbol{c}_c) = \sum_{j=1}^{4} \left(\frac{\boldsymbol{h}_{cj} * \boldsymbol{c}_{cj}}{|\boldsymbol{h}_{cj}| |\boldsymbol{c}_{cj}|} \right) \tag{8}$$

对于时间戳特征，通过式(9)计算其相似度，其中$\boldsymbol{h}_t$和$\boldsymbol{c}_t$分别是文档与关键事件的时间戳特征，u和σ为超参数。

$$sim_t(\boldsymbol{h}_t, \boldsymbol{c}_t) = \sum_{j=1}^{3} \left(\exp\left(-\frac{(\boldsymbol{h}_t - \boldsymbol{c}_{t_j} - u)^2}{2\sigma^2} \right) \right) \tag{9}$$

结合 3 种特征的相似度，以加权相似度特征函数$score_{w(d,c)}$来表示文档$\boldsymbol{d}$和关键事件池C_p中各关键事件$\boldsymbol{c}$之间的相似性，如式(10)所示：

$$score_w(\boldsymbol{h}_d, \boldsymbol{h}_c) = sim_{e(\boldsymbol{h}_e, \boldsymbol{c}_e)} * + sim_{c(\boldsymbol{h}_{cj}, \boldsymbol{c}_{cj})} * \beta + sim_t(\boldsymbol{d}_{ts}, \boldsymbol{c}_{tsk}) * \gamma \tag{10}$$

其中，α，β，γ为可调节的超参，用于对特征进行加权融合，自适应地衡量、利用不同特征的重要性。比较文档与关键事件池C_p中各个关键事件的排名分数，即可找到与文档最相似的关键事件c_{top}，如式(11)所示：

$$top = \mathrm{argmax}(\{score_w(\boldsymbol{h}_d, \boldsymbol{h}_{c_1}), score_w(\boldsymbol{h}_d, \boldsymbol{h}_{c_2}), \cdots, score_w(\boldsymbol{h}_d, \boldsymbol{h}_{c_k})\}) \tag{11}$$

3）再次，输入文档特征向量$\boldsymbol{h}_d$与c_{top}，利用关键事件探测模型m_{de}，输出该文档是否要并入c_{top}。该模型为全连接神经网络(Fully-Connected Neural Network，FNN)分类器，由全连接网络与软最大值(Softmax)层组成，输入文档$\boldsymbol{d}$的多视角文档特征向量$\boldsymbol{h}_d$以及c_{top}的特征向量$\boldsymbol{h}_{c_{\mathrm{top}}}$，输出为真或假，如式(12)所示：

$$output = \mathrm{argmax}(\mathrm{softmax}(DNN(\boldsymbol{h}_d, \boldsymbol{h}_{c_{\mathrm{top}}}))) \tag{12}$$

其中，$output \in \{0, 1\}$，0 代表假，1 代表真。

4）根据 output 执行不同的操作：(1)若模型输出为假，则创建新的关键事件c_j加入关键事件池C_p，其中c_j的多视角特征$\boldsymbol{h}_{c_j}$继承自文档d的多视角文档特征向量$\boldsymbol{h}_d$；(2)若输出为真，则将该文档合并至c_{top}并根据$\boldsymbol{h}_d$更新c_{top}的特征向量$\boldsymbol{h}_{c_{\mathrm{top}}}$，具体如式(4)－式(6)所示。

3.6 算法流程

CMFF-MED 的多语言事件探测流程如算法 1 所示。

算法 1 实体增强的多语言事件探测算法

输入：待探测的多语言新闻文档集 $D=\{d_1, d_2, \cdots, d_m\}$，爬取的多语言公共新闻文档集$D_{ex}$，多语言实体语义匹配模块$M_{se}$，多视角特征提取模块$M_{fe}$，多特征事件探测模块$M_{de}$

输出：关键事件集合 $C=\{c_1, c_2, \cdots, c_k\}$

初始化：第一个新闻文档初始化为第一个关键事件 $C=\{c_1\}$，通过M_{se}获取D_{ex}中包含的多语言同义实体集

1. 循环 $d_i \in D$：

[1] https://github.com/priberam/news-clustering

2. 通过M_{fe}获取d_i的多视角特征向量$\mathbf{h}_{d_i}$
3. Sim=$\{\}$
4. 循环 $c_j \in C$:
5. 获取c_j的特征向量$\mathbf{h}_{c_j}$
6. $sim_j = score_w(\mathbf{h}_{d_i}, \mathbf{h}_{c_j})$
7. 将sim_j加入 Sim
8. 记 k 为 Sim 中最大值对应的下标
9. 最相似关键事件c_{top}为c_k
10. 将$\mathbf{h}_{d_i}$与$\mathbf{h}_{c_k}$输入M_{de}:
11. 如果输出为真:
12. 将d_i加入c_k,更新c_k的特征向量$\mathbf{h}_{c_k}$
13. 否则:
14. 以d_i为基础创建新关键事件,加入 C
15. 输出 C=$\{c_1, c_2, \cdots, c_k\}$

4 训练细节与推理

4.1 跨语言语义匹配模型的训练策略

跨语言语义匹配模型的结构如图 4 所示,本文引入 MUSE[19]提供的多种语言对的平行字典作为跨语言知识,对其进行训练。

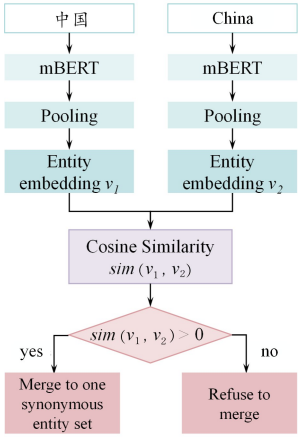


图 4 跨语言语义匹配模型

Fig. 4 Cross-lingual semantic matching model

MUSE[19]提供了 110 种语言对的平行字典,其中有 14 种与本文的 9 种语言相关。英语、法语、德语、意大利语和西班牙语 5 种语言之间任意语言对均有平行字典,共$C_5^2=10$种。汉语、俄语、斯洛文尼亚语和克罗地亚语 4 种语言只与英语有平行字典,共 4 种。根据这些平行字典构建训练语料,作为跨语言知识,用于训练跨语言语义匹配模型。具体来说:正样本为平行字典中每一行的两个不同语言的同义词,负样本为随机挑选的两个不同义的实体。为了避免正负样本不均衡,本文按照正负样本 1∶10 的比例训练跨语言语义匹配模型。另外,考虑到不同语言的样本量不平衡的问题[20],本文依据式(13)确定不同语言的样本被训练的比例。

$$P(L)=|L|^{\delta}/N \tag{13}$$

其中,$P(L)$为语言 L 的样本被训练的比例,δ 为超参数,$|L|$为语言 L 的总样本量,N 为多种语言的样本量。

4.2 实体知识增强的 mBERT 模型的训练策略

实体知识增强的 mBERT 模型的训练策略考虑两个层面:数据层面和预训练任务层面。具体地,在数据层面,本文首先根据D_{ex}以及E_{sy}构建训练样本,具体来说,从D_{ex}中提取不同语言对(共有$C_9^2=36$种)的句子对$stc_{pair}^{i,j}=\{stc_i, stc_j\}$,使得两个句子中有实体属于同一多语言同义实体集$E_{sy}^k$;其次选取句子对时,优先选取出现次数多的句子,句子中实体反复出现,说明有更高的概率句子在描述该实体,因此信息熵更高。在预训练任务层面,mBERT 的预训练任务为预测下一个句子(Next Sentence Predict,NSP)任务,具体来说,将句子成对输入,预测两个句子是否为同一段文本中的连续句子。研究[21]指出,该任务会退化为判断两个句子是否描述的是同一主题,因此,本文将句子对同时输入 mBERT 中进行训练,旨在学习跨语言同主题的句子的特征,鼓励模型将不同语言句子对包含的信息融合。

4.3 关键事件探测模型的推理阶段

多语言事件探测任务的数据集往往包含大量新闻文档,记为 N 篇,这些文档组成 M 个关键事件。若使用常规非增量算法,计算文档间相似度的次数的渐进上界为 $O(N^2)$,而增量式的算法能将该次数减少为 $O(MN)$。以本文实验的数据集为例,平均每个关键事件包含约 30 篇文档,则增量式的算法提高了近 30 倍的效率,避免了计算资源的大量消耗。因此,本文方法与对比实验大多采用增量式的算法。

5 实验及实验结果分析

5.1 实验设置

5.1.1 数据集

本文的实验中使用的新闻语料库[1)]包含括 33 807 篇新闻文档,涵盖了 9 种语言,每种语言的新闻数量如表 1 所列。其中训练集为 2013 年 12 月 18 日到 2014 年 2 月 2 日的新闻文档,而测试集为 2014 年 11 月 2 日到 2015 年 8 月 25 日的新闻文档,从而保证了测试集较新,并且测试集与训练集中的关键事件没有重合。

另外,从表中可以看出,该数据集的训练集中只有英语、西班牙语、德语和汉语 4 种语言,而测试时需要对 9 种语言混合的新闻文档流进行探测。这就需要多语言事件探测模型具备跨语言迁移能力,以确保低资源语言能够获得良好的零样本泛化性能。

表 1 数据集详细数据

Table 1 Details of experiment datasets

	Language	en	es	de	zh	ru	it	fr	sl	hr
Train	Documents	12 233	4 527	4 032	10	—	—	—	—	—
	Key events	593	416	377	1	—	—	—	—	—
Test	Documents	8 726	2 177	2 101	440	231	88	61	37	13
	Key events	222	149	118	9	1	2	2	3	2

5.1.2 对比方法

本文进行多达 9 种语言的事件探测任务，该任务较新，相关研究较少，据我们所知，只有 Miranda 等[7]与 Linger 等[8]利用两阶段管道式方法进行了实验。本文则进行端到端的多语言事件探测，实验的具体对比方法如下：

1）K-Means：该方法采用多语言预训练模型表示文档，用余弦相似度计算文档相似度，基于 K-Means 算法进行多语言事件探测，属于非增量式算法，实验中需要指定正确的事件个数。模型由本文实现。

2）SinglePass：该方法采用多语言预训练模型表示文档，用余弦相似度计算文档相似度，基于 Single-Pass 算法进行多语言事件探测，属于增量式算法，实验中不需要指定正确的事件个数。模型由本文实现。

3）Miranda 等提出的两阶段管道方法[7]：该方法采用多种统计特征使文本与单语关键事件相关联，然后使用多语言词向量将不同的单语关键事件跨多种语言对齐，属于增量式算法。模型由本文复现。

4）Linger 等提出的“重播”方法[8]：该方法基于 Miranda 等[7]的方法，在探测流程中引入了“重播”过程，并使用了质量更高的多语言预训练语言模型上下文向量来将不同的单语关键事件跨多种语言对齐，属于增量式算法。

5）CMFF-MED$_{ti}$：本文提出的方法，其中多视角特征只使用时间戳特征来进行多语言事件探测。

6）CMFF-MED$_{en}$：本文提出的方法，其中多视角特征只使用多语言实体对齐特征来进行多语言事件探测。

7）CMFF-MED$_{en+ti}$：本文提出的方法，其中多视角特征只使用多语言实体对齐特征和时间戳特征来进行多语言事件探测。

8）CMFF-MED$_{co}$：本文提出的方法，其中多视角特征只使用多语言上下文特征来进行多语言事件探测。

9）CMFF-MED$_{co+ti}$：本文提出的方法，其中多视角特征只使用多语言上下文特征和时间戳特征来进行多语言事件探测。

10）CMFF-MED$_{en+co}$：本文提出的方法，其中多视角特征只使用多语言实体对齐特征和多语言上下文特征来进行多语言事件探测。

11）CMFF-MED$_{en+co+ti}$：本文提出的方法，使用全部 3 种多视角特征进行多语言事件探测。

5.1.3 评价指标

本文实验以贝塔立方评估（Beta-Cubed，BCubed）分数衡量实验效果，包含 Bcubed F1 值、Bcubed 准确率和 Bcubed 召回率分数。Bcubed 分数是多语言事件探测任务的常用指标，计算方式如式（14）－式（16）所示：

$$Bcubed_{\text{precision}}=\frac{\sum_{i\in C}\sum_{j\in C}n_{ij}^2}{\sum_{i\in C}(\sum_{j\in C}n_{ij})^2} \tag{14}$$

$$Bcubed_{\text{recall}}=\frac{\sum_{i\in C}\sum_{j\in C}n_{ij}^2}{\sum_{j\in C}(\sum_{i\in C}n_{ij})^2} \tag{15}$$

$$Bcubed_{\text{F1}}=\frac{2*Bcubed_{\text{precision}}*Bcubed_{\text{recall}}}{Bcubed_{\text{precision}}+Bcubed_{\text{recall}}} \tag{16}$$

其中，C 为探测得到的关键事件集合，n_{ij} 表示关键事件 i 中与关键事件 j 中在标注数据里属于同一关键事件的文档数量，$\sum_{j\in C}n_{ij}$ 表示关键事件 i 中的文档总数。

5.1.4 实验环境及超参

本文实验是在 GPU 服务器上进行的，配置如表 2 所列，实验中使用的超参数如表 3 所列。

表 2 实验环境配置

Table 2 Experimental environment configuration

配置	规格
GPU	RTX 3090
编程语言	Python3.8
深度学习框架	Pytorch1.8.1

表 3 超参数设置

Table 3 Hyperparameters setting

超参数	参数值
Mbert 上下文向量维度	768
Mbert 优化器	Adam
Mbert 学习率	2×10^{-5}
Mbert 批处理大小	8
字符匹配模型最大编辑距离	2
时间戳相似度的 u	0
时间戳相似度的 σ	3
加权相似度特征函数的 α	0.4
加权相似度特征函数的 β	0.4
加权相似度特征函数的 γ	0.2
语义匹配模型语料选取的 δ	0.3
m_{de} 网络层数与神经元数	2 * 100

5.1.5 主实验结果与分析

主实验结果如表 4 所列。

表 4 主实验结果

Table 4 Main experiment results

Method	Key Event Count	B-Cubed Metrics		
		P/%	R/%	F1/%
K-Means	394	61.45	58.93	60.06
SinglePass	455	71.38	71.51	71.44
Miranda 等[7]	587	89.28	73.87	80.86
Linger 等[8]	606	80.25	83.97	82.06
CMFF-MED$_{ti}$	385	32.23	41.60	36.32
CMFF-MED$_{en}$	805	89.91	74.81	81.67
CMFF-MED$_{en+ti}$	759	89.94	74.97	81.78
CMFF-MED$_{co}$	763	91.46	77.46	83.88
CMFF-MED$_{co+ti}$	708	91.75	77.54	84.05
CMFF-MED$_{en+co}$	785	84.85	**87.14**	85.98
CMFF-MED$_{en+co+ti}$	731	**94.75**	79.89	**86.69**

由表 4 可以看出，本文利用多视角特征的 CMFF-MED 取得了最佳性能，与已有算法相比，BCubed F1 值提高了 4.63%。从每个视角特征的消融实验可以看出，每个视角的特征对多语言事件探测的效果均有提升。根据表中实验结果可以得出以下结论：

1）单独使用时间戳特征进行实验的探测效果最差，因为时间戳特征中不包括文档正文内容。

2）对比CMFF-MED$_{en}$，CMFF-MED$_{co}$以及 CMFF-MED$_{en+co}$的性能指标可以发现：单独使用多语言实体特征或多语言上下文特征的性能表现一般。当两者结合使用时，可以综合多语言同义实体层级视角和多语言句子层级视角的特征，既利用多语言同义实体库对文档中的跨语言同义词进行统计，又利用实体增强的 mBERT 模型对文档语义编码至高维空间，从而显著提升了性能。

观察该设置下具体的事件探测结果可以发现，与其他对比实验相比，该实验设置的结果产生了较多的碎片关键事件，一些本该属于同一关键事件的新闻文档分裂成了多个小关键

事件。例如：由 378 篇文档组成的凯特王妃顺产关键事件分裂为了 97 个关键事件，其中有 1 个关键事件包含 171 篇文档，3 个关键事件文档数量大于 10 篇，其他 92 个关键事件文档数量小于 5 篇。由于 Bcubed 指标本身的特性，这种情况会降低准确性并提高召回率。碎片关键事件的产生主要有两个原因：(1)不同实验设置下训练的不同关键事件探测模型导致了不同的事件文档合并倾向，该实验设置的关键事件探测模型倾向于更少的合并，从而产生了更多的碎片关键事件；(2)该实验设置未使用时间戳特征，本该合并的时间上相距较短的关键事件未被合并。

3) 对比 CMFF-MED$_{en}$ 与 CMFF-MED$_{en+ti}$，CMFF-MED$_{co}$ 与 CMFF-MED$_{co+ti}$，CMFF-MED$_{en+co}$ 与 CMFF-MED$_{en+co+ti}$ 这 3 组实验设置的性能指标结果可以发现，时间戳特征的使用有助于减少碎片化的关键事件，提升准确率，但召回率有所下降，总体来说是有助于提升实验性能的。

5.2 参数分析实验

5.2.1 权重分析实验

本节针对加权相似度特征函数的超参数 α,β,γ 进行调整后再进行实验，其中 α,β,γ 分别为 3 种视角特征的权重，具体实验结果如图 5、图 6 所示。

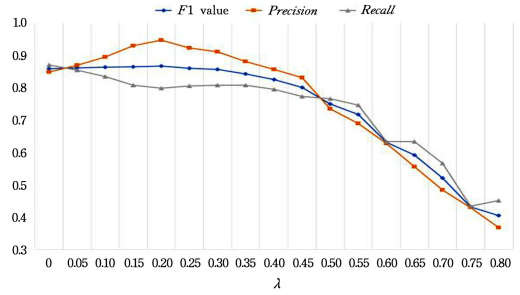


图 5 权重分析实验 1

Fig. 5 Weight analysis experiment 1

图 5 实验中固定 $\alpha=\beta$，不断调整 γ，(即 $\alpha=\beta=\frac{1-\gamma}{2}$)，由实验结果可以发现，当 γ 值过高时，多语言同义实体层级视角特征以及多语言句子层级视角特征被忽略，实验效果急剧下降。当 $\gamma=0.2$ 时，能够获取最佳的实验效果。

图 6 实验中固定 $\gamma=0.2$，不断调整 α 与 β(其中 $\beta=1-\gamma-\alpha$)。由实验结果可以发现，当 $\alpha=0.4$ 时，能够平衡多语言同义实体层级视角特征以及多语言句子层级视角特征，获取最佳的实验效果。在本文的其他实验中，均令 $\alpha=0.4,\beta=0.4,\gamma=0.2$。

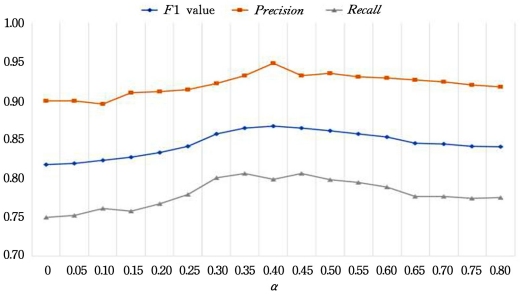


图 6 权重分析实验 2

Fig. 6 Weight analysis experiment 2

5.2.2 泛化性分析实验

本文提出的方法利用多视角特征提取模块，针对多种语言的文档从 3 个层级视角提取特征。通过多语言同义实体集和预训练的多语言编码器，不同语言的文档特征可以映射到同一向量空间，从而提升了对于小语种新闻文档事件探测的泛化性能。本文多语言事件探测的训练集中包括英语、西班牙语、德语、汉语 4 种语言，而测试集中包括 9 种语言。对各个语言的测试集文档进行单语言事件探测，实验结果如表 5 所列。

表 5 泛化性分析实验

Table 5 Generalization analysis experiment

Language	Key Event Count	B-Cubed Metrics		
		P/%	R/%	$F1$/%
English	389	93.93	93.01	93.47
Spanish	284	94.46	88.31	91.28
German	191	96.17	91.96	94.02
Chinese	17	98.77	94.22	96.44
Russian	1	100.00	100.00	100.00
Italian	3	97.81	98.82	98.31
French	3	99.63	98.04	98.83
Slovenian	8	87.36	76.76	82.15
Croatian	4	89.94	71.80	79.85

由表 5 可以看出，即使训练集中没有包含低资源语言的文档，也能有效地进行事件探测，证明了本文方法在低资源语言上的泛化性能。

5.2.3 分类器对比实验

本文比较了决策树(Decision Tree，DT)分类器、随机梯度下降(Stochastic Gradient Descent，SGD)分类器、K 近邻(K Nearest Neighbors)分类器、自适应增强(Adaptive Boosting，AdaBoost)分类器、支持向量机(Support Vector Machine，SVM)分类器以及全连接神经网络分类器作为关键事件探测模型 m_{de} 的多语言事件探测效果，结果如表 6 所列。

表 6 分类器对比实验

Table 6 Classifier comparison experiment

Language	Key Event Count	B-Cubed Metrics		
		P/%	R/%	$F1$/%
DT	568	71.38	84.94	77.57
SGD	656	84.62	76.73	80.48
KNN	684	84.55	79.41	81.90
AdaBoost	781	93.89	76.36	84.22
SVM$_{linear}$	793	91.56	76.94	83.61
SVM$_{rbf}$	822	91.03	74.82	82.13
SVM$_{poly}$	834	78.95	65.96	71.87
FNN	731	94.75	79.89	86.69

由表 5 可以看出，全连接神经网络分类器的实验效果最好。

结束语 基于对跨层级多视角特征的利用，本文提出了一种多语言事件探测的方法，该方法有以下特点：1)利用多语言实体语义匹配模块对多种语言的同义实体进行对齐，解决了同义实体表述不一的难题；2)利用多视角特征提取模块，从 3 个层级视角提取文档特征，提高了低资源语言的泛化性能；3)利用多特征事件探测模块，加权融合多视角特征，从而获得了高可靠性的多语言事件探测结果。本文存在的问题：从实验结果来看，仍有较多碎片化的关键事件，拉低了性能。未来

的工作包括改进探测的流程来减少关键事件的数量，以及在更大更多语言的新闻数据集上进行事件探测等。

参 考 文 献

[1] HUANG Z，LI Z，JIANG H，et al. Multilingual Knowledge Graph Completion with Self-Supervised Adaptive Graph Alignment[C]// Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. 2022：474-485.

[2] AHUJA K，KUMAR S，DANDAPAT S，et al. Multi Task Learning For Zero Shot Performance Prediction of Multilingual Models[C]// Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. 2022：5454-5467.

[3] FUJINUMA Y，JORDAN L，KANN K，et al. Match the Script，Adapt if Multilingual：Analyzing the Effect of Multilingual Pretraining on Cross-lingual Transferability[C]// Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. 2022：1500-1512.

[4] GUZMAN L，LAI V，POURAN A，et al. Event Detection for Suicide Understanding[C]// Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics. 2022：1952-1961.

[5] YANG W，BOYD-GRABER J，PHILIP R. A Multilingual Topic Model for Learning Weighted Topic Links Across Corpora with Low Comparability[C]// Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. 2019：1243-1248.

[6] LIU J，CHEN Y，LIU K，et al. Neural Cross-Lingual Event Detection with Minimal Parallel Resources[C]// Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. 2019：738-748.

[7] MIRANDA S，ZNOTINS A，COHEN S，et al. Multilingual clustering of streaming news[C]// Proceedings of the 2018 Conference onEmpirical Methods in Natural Language Processing. 2018：4535-4544.

[8] LINGER M，HAJAIEJ M. Batch clustering for multilingual news streaming[C]// Proceedings of the Text2Story'20 Workshop. 2020：55-61.

[9] LABAN P，HEARST M. newsLens：building and visualizing long-ranging news stories[C]// Proceedings of the Events and Stories in the News Workshop. 2017：1-9.

[10] STAYKOVSKI T，BARRON-CEDENO A，MARTINO G，et al. Dense vs. Sparse Representations for News Stream Clustering[C]// Proceedings of the Text2Story'19 Workshop. 2019：47-52.

[11] LI Y，YU Z，GAO S，et al. Case-related News Detection Based on Case Element and Deep Clustering Method[J]. Journal of Chinese Information Processing，2021，35(11)：60-69.

[12] GUO H，WANG Z，ZHU Q，et al. Event Clustering Method for Chinese Social Text Based on Semi-supervised Learning [J]. Journal of Chinese Information Processing，2022，36(2)：152-159.

[13] SARAVANAKUMAR K，BALLESTEROS M，CHANDRASEKARAN M，et al. Event-Driven News Stream Clustering using Entity-Aware Contextual Embeddings[C]// Proceedings of The 16th Conference of the European Chapter of the Association for Computational Linguistics. 2021：2330-2340.

[14] DEVLIN J，CHANG M，LEE K，et al. BERT：Pre-training of Deep Bidirectional Transformers for Language Understanding[C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics. 2019：4171-4186.

[15] ZHANG Y，GUO F，SHEN J，et al. Unsupervised Key Event Detection from Massive Text Corpora[C]// Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2022：2535-2544.

[16] REIMERS N，GUREVYCH I. Sentence-BERT：Sentence Embeddings using Siamese BERT-Networks[C]// Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. 2019：3980-3990.

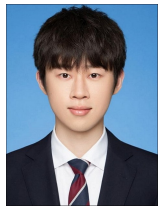
[17] YANG W，YU Z，GAO S，et al. Chinese-Vietnamese news topic discovery methodbased on cross-language neural topic model [J]. Journal of Computer Applications，2021，41(10)：2879-2884.

[18] CONNEAU A，LAMPLE G，DENOYER L，et al. Word Translation Without Parallel Data[C]// Proceedings of the 6th International Conference on Learning Representations. 2018：1-14.

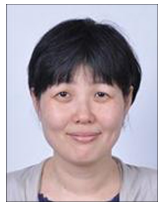
[19] FENG F，YANG Y，CER D，et al. Language-agnostic BERT Sentence Embedding[C]// Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. 2022：878-891.

[20] XUE L，CONSTANT N，ROBERTS A，et al. mT5：A Massively Multilingual Pre-trained Text-to-Text Transformer[C]// Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics. 2021：483-498.

[21] YANG Z，DAI Z，YANG Y，et al. XLNet：Generalized Autoregressive Pretraining for Language Understanding[C]// Advances in Neural Information Processing Systems 32. 2019：5754-5764.



ZHANG Zhiyuan，born in 1998，postgraduate. His main research interests include natural language processing and multilingual pretraining model.



RUAN Tong，born in 1973，professor，Ph.D supervisor. Her main research interests include text extraction knowledge graph and data quality assessment.

（责任编辑：何杨）