基于阅读理解的汉越跨语言新闻事件要素抽取方法

朱恩昌[1,2]， 余正涛[1,2]， 高盛祥[1,2,*]， 黄于欣[1,2]， 郭军军[1,2]
(1.昆明理工大学信息工程与自动化学院，云南昆明 650500;
2.昆明理工大学云南省人工智能重点实验室，云南昆明 650500)
978105863@qq.com,ztyu@hotmail.com,gaoshengxiang.yn@foxmail.com,
huangyuxin2004@163.com,guojjgb@163.com

摘要

新闻事件要素抽取旨在抽取新闻文本中描述主题事件的事件要素，如时间、地点、人物和组织机构名等。传统的事件要素抽取方法在资源稀缺型语言上性能欠佳，且对长文本语义建模困难。对此，本文提出了基于阅读理解的汉越跨语言新闻事件要素抽取方法。该方法首先利用新闻长文本关键句检索模块过滤含噪声的句子。然后利用跨语言阅读理解模型将富资源语言知识迁移到越南语，提高越南语新闻事件要素抽取的性能。在自建的汉越双语新闻事件要素抽取数据集上的实验证明了本文方法的有效性。

关键词： 新闻事件要素抽取；长文本语义建模；跨语言知识迁移；阅读理解

News Events Element Extraction of Chinese-Vietnamese Cross-language Using Reading Comprehension

**Enchang Zhu[1,2]， Zhengtao Yu[1,2]， Shengxiang Gao[1,2,*]，
Yuxin Huang[1,2]， Junjun Guo[1,2]**
(1.Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, Yunnan 650500, China ;
2. Yunnan Key Laboratory of Artificial Intelligence, Kunmong University of Science and Technology, Kunming, Yunnan 650500, China)
978105863@qq.com,ztyu@hotmail.com,gaoshengxiang.yn@foxmail.com,
huangyuxin2004@163.com,guojjgb@163.com

Abstract

Argument extraction of news text events, the task of extracting extract the event argument that describe the topic events in the news text, such as time, location, people, and organization.Traditional event element extraction methods have poor performance on resource-scarce languages, and it is difficult to model the semantics of long texts.In this paper, we propose News Events Element extraction of Chinese-Vietnamese Cross-language Using Reading Comprehension.Firstly, we use the news long text key sentence retrieval module to filter noisy sentences.Then we use the cross-language reading comprehension model to transfer rich resource language knowledge to Vietnamese to improve the performance of Vietnamese news event element extraction. Experiments on the self-built Chinese-Vietnamese bilingual data set prove the effectiveness of this method.

Keywords: News event element extraction , News event element extraction , Cross-language knowledge transfer , Reading comprehension

项目基金： 国家自然科学基金(61972186， 61762056， 61472168)； 云南省重大科技专项计划项目(202002AD080001)；云南省高新技术产业专项（201606）

1 引言

越南与我国毗邻，在国家“一带一路”战略大环境下，越南与我国交流越来越密切，相关的新闻事件越来越多，而这些报道分布在国内及越南相关网站及媒体上，呈现为中文或者越南文，如何能够及时有效了解国内及越南的新闻事件信息意义重大。新闻事件要素抽取任务旨在抽取新闻文本中描述主题事件的事件要素，如时间、地点、人物和组织机构名等。新闻事件要素抽取是新闻事件抽取的重要子任务之一，是新闻文本相似度计算、新闻事件关联关系分析以及事件检索等下游任务的基础。近年来，(Liao and Grishman, 2010; Hong et al., 2011; Li et al., 2013; Chen et al., 2015)在事件抽取方法已经做了大量的工作并取得了很大的进展，但是事件要素抽取仍然面临着很多的挑战并成为了提高事件抽取模型整体性能的瓶颈。在事件要素抽取方面，基于端到端的神经网络模型取得了很好的效果 (Wang et al., 2019b; Yang et al., 2019)。这些方法在有大规模标注数据的语种上有很好的效果，如英语和中文。但在只有少量或者无标注数据的小语种上的性能还有很大的提升空间，如东南亚国家的非通用语言（越南语、缅甸语、泰语、老挝语以及柬埔寨语等）。跨语言新闻事件要素抽取试图通过跨语言知识迁移的方法将富资源语言知识迁移到资源稀缺型语言来提高目标语言端的新闻事件要素抽取性能。 (Liu et al., 2018a)为了充分利用两种语言中的互补信息，提出了一种跨语言门控注意力机制。 (Lu et al., 2020)提出了跨语言结构迁移，将语言的通用表示(依存树、全连通图)通过图卷积网络编码到一个公共的语义空间中，从而实现跨语言结构迁移，一定程度上解决了上述挑战。上述方法都是利用句子级别的语法或语义信息从单一句子中实现抽取事件要素，但是由于事件构成的复杂性以及语言表述的多样性，在新闻报道中多数情况下需要多个语句才能完整地描述一个事件。此时，需要对多个语句进行分析，才能获取完整的事件要素。篇章级事件要素抽取需要捕获长距离的语义信息，支持跨事件的关联性分析，因此需要更强的语义理解和推断能力。对此， (Du and Cardie, 2020)提出了多粒度编码的模型，动态融合句子级的局部特征和段落级全局特征，在一定程度上解决了篇章语义建模困难的问题。综上所述，目前跨语言新闻事件要素抽取仍然面临着以下两个方面的挑战：1）基于端到端神经网络新闻事件要素抽取的方法，严重依赖大规模且高质量的标注语料，而低资源语言标注语料稀缺，在传统事件要素抽取模型上效果有待提升；2）篇章级长文本语义建模困难：新闻文本通常围绕一个主题事件进行报道，而描述主题事件的事件要素往往分散在多个句子中，需要捕获长距离的语义信息，传统基于RNN端到端的模型无法解决长距离依赖问题，而基于Transformer的BERT一定程度上弥补了RNN的缺陷，但新闻文本在进行预处理后，其文本长度远超过BERT能处理的最大长度。

本质上，新闻事件要素抽取可以看作是一个机器阅读理解（MRC）问题 (Hermann et al., 2015; Chen et al., 2016)，基于机器阅读理解的新闻事件要素方法能够让问题编码一些先验语义知识，使模型具有更强的推理能力，关注更深层次的篇章级上下文语义信息。在其思想的启发下，将阅读理解的思想应用于新闻事件要素抽取任务。而针对越南语新闻要素标注语料稀缺和BERT不能有效处理长文本的问题，我们在模型中分别引入了跨语言知识迁移和 (Ding et al., 2020)提出的CogLTX方法，并提出了一种基于机器阅读理解的汉越跨语言新闻事件要素抽取方法。模型整体流程如图1所示。首先利用新闻长文本关键子句检索模型检索出与问题高度相关的句子，过滤包含噪声的句子，以解决新闻长文本语义建模困难的问题。其次利用新闻事件要素抽取模型提取候选句子中的新闻事件要素。其核心思想是，首先使用中文阅读理解数据集预训练源语言端的阅读理解模型。然后利用mBERT（multilingual BERT）同时建模中文和越南语<文本，问题，答案>之间的关系，并融合两种语言的表示来实现中文到越南语的跨语言知识迁移。在自建的汉越双语新闻事件要素抽取语料上的实验证明，我们提出的基于阅读理解的汉越跨语言新闻事件要素抽取方法，优于传统的基线方法，验证了跨语言阅读理解对资源稀缺型语言（越南语）新闻事件要素抽取任务的有效性。

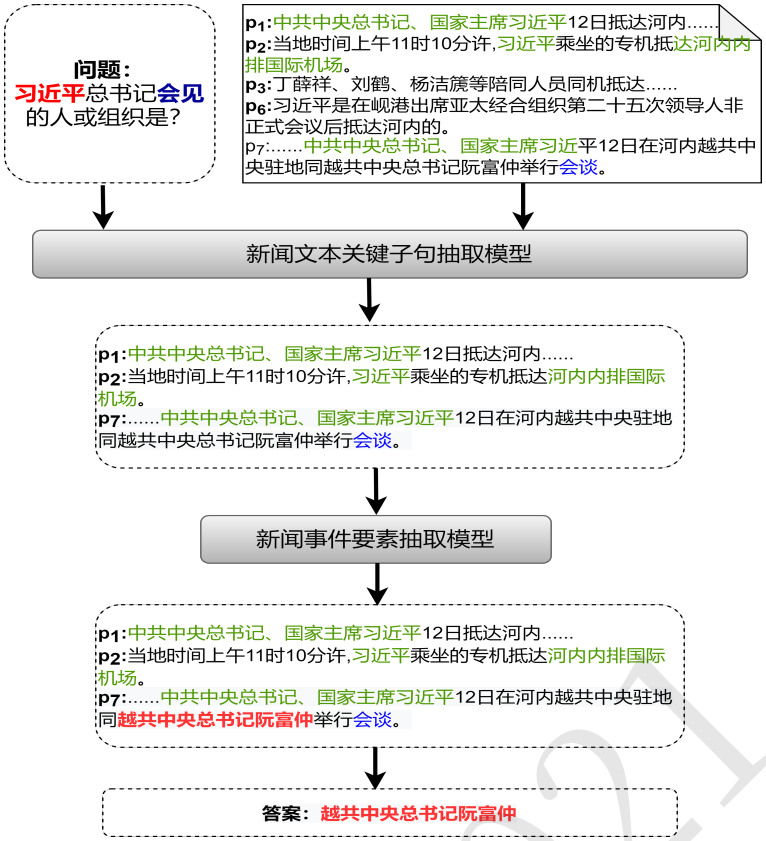

图 1. 模型整体流程图.

2 相关工作

2.1 事件要素抽取

事件抽取是信息抽取的主要研究任务，而事件要素抽取是事件抽取的关键子任务之一，旨在抽取文本中指定事件类型的事件要素，如事件发生的时间、地点以及涉及到的组织机构名等实体信息。现有的事件要素抽取工作主要集中在单语场景下，它可以分为：基于特征的方法和基于神经网络的方法。其中**基于特征的方法**，其主要利用人工设计的特征进行事件要素抽取，这些特征包括词法（lexical）、句法（syntactic）特征，文档级（document-level）特征 (Ji and Grishman, 2008)，实体级（entity-level）特征 (Hong et al., 2011)以及其他更为复杂的特征。基于特征表示的方法具有很好的可解释性，但需要大量的专家知识，领域迁移性较差。随着计算能力的提升以及高质量数据集的出现，**基于神经网络的方法**成为目前的主流方法。(Chen et al., 2015)提出了动态多池化卷积神经网络，根据不同的事件和事件要素特征进行不同的池化处理来更好地捕捉文本特征。 (Liu et al., 2018b)提出了一种新颖的联合多个事件提取框架（JMEE），通过引入句法树来增强信息流动，以解决句子编码的长距离依赖问题，利用基于注意力的图卷积网络来模型化图信息，从而联合抽取多个事件的触发词和事件要素。 (Ma et al., 2020)针对数据稀疏、事件要素重叠等问题，提出了利用训练语料中未标记数据和语法信息的资源增强的事件要素抽取，在ACE2005数据集上取得了新的进展。

目前事件要素抽取的研究主要集中于利用句子级别的语法或语义特征从单一句子中实现事件要素抽取，但由于事件构成的复杂性及语言表述的多样性，在大多数情况下需要多个语句才能完整地描述一个事件。为了获取更多的上下文信息来指导事件要素抽取，近年来也存在一些研究者尝试研究篇章级的事件抽取方法。 (Zhao et al., 2018; Duan et al., 2017)等人则通过使用文档嵌入的方法，获取文档级的全局信息，联合篇章向量、词向量、实体类型向量进行事件抽取。 (Zheng et al., 2019)提出了一种端到端模型从财务公告抽取金融事件的信息，通过将事件结构化任务转换为路径扩展子任务，进而生成一个基于实体的有向无环图实现文档级事件抽取。 (Du and Cardie, 2020)提出了句子级和段落级的多粒度编码器，以动态融合句子级和段落级特征，提高了篇章级的事件要素抽取性能。相比句子级别的事件抽取方法，篇章级事件抽取

需要捕获长距离的语义信息，支持跨事件的关联性分析，因此要求模型具有更强的语义理解和推断能力。并且，基于神经网络的事件要素抽取方法严重依赖于大规模的人工标注数据训练模型，他们不适用于数据标注稀缺的语言。

跨语言事件要素抽取旨在联合多语言训练数据共同训练事件要素抽取模型，以缓解目标语言事件要素标注语料稀缺的问题。目前跨语言事件要素抽取的研究工作相对较少。在现有的工作中，(Zhu et al., 2014)使用基于机器翻译的方法把英文语料翻译成中文来扩充中文数据进行模型训练。(Hsi et al., 2016)探讨了词向量投影与多语言特征相结合的方法，用于英中双语事件识别。(Subburathinam et al., 2019; Lu et al., 2020)基于多语言之间依存关系的通用性，提出了跨语言结构迁移模型，提高了目标语言事件抽取模型性能。尽管这些工作取得了一定的进展，他们往往依赖大规模的平行资源。绝大多数的低资源语言平行资源十分稀缺，这一定程度上限制了这些方法的应用。

2.2 机器阅读理解

机器阅读理解任务(Machine Reading Comprehension, MRC)旨在测试计算机能否理解输入文本。给定一个段落和一个相应的问题，要求计算机对该问题进行回答。在最新的工作中，基于BERT的模型取得了超越人类表现的性能 (Devlin et al., 2018)。通过预训练，它的性能显著优于先前的基于双向注意力机制的阅读理解模型 (Seo et al., 2016)和多层注意力机制的阅读理解模型 (Cui et al., 2016)。现有的工作主要集中在单语的机器阅读理解，在资源稀缺的语言中，其性能有待进一步的提升。基于此，(Asai et al., 2018)提出了基于机器翻译的跨语言阅读理解框架，有效提高了目标语语言的机器阅读理解性能。基于机器翻译的机器阅读理解会给模型引入噪声，针对该问题 (Cui et al., 2019)提出了Dual-BERT跨语言机器阅读理解模型，充分利用源语言大规模的标注数据来提升目标语言的性能。机器阅读理解在信息抽取领域也得到了很好的应用。例如，(Li et al., 2019)使用机器阅读理解来建模关系抽取问题。传统的事件要素抽取问题都将其看作是一个分类任务，基于分类任务的事件要素抽取方法面临着数据稀疏问题并且忽略了事件类型和事件要素之间的关系，为了解决以上问题，(Liu et al., 2020)提出了基于机器阅读理解（MRC）的事件抽取方法，在检测事件的基础上进行事件要素抽取。

3 基于阅读理解的汉越跨语言新闻事件要素抽取

本节中，我们将详细介绍基于阅读理解框架的汉越跨语言新闻事件要素抽取模型。我们的模型旨在从越南语的新闻文本中抽取指定事件类型的事件要素。我们首先利用新闻长文本关键句检索模型从新闻文本中抽取出与问题高度相关的句子，过滤含噪声的句子，然后从候选句子中提取答案，从而实现越南语新闻事件要素抽取。

给定一个问题$\mathrm{Q}=\{\mathrm{q}_1,\mathrm{q}_2,...,\mathrm{q}_n\}$,其中$n$表示问题中序列长度。一个新闻文本$P=\{p_1^{(1)},p_2^{(1)},...,p_{l1}^{(1)},...,p_1^{(k)},p_2^{(k)},...,p_{lk}^{(k)}\}$,其中$p_i^{(k)}$表示新闻文本中第$i$个子句的第$k$个词。我们的模型首先通过关键句检索模型从长文本中检索出与问题高度相关的关键句p_i，在检索出关键句p_i的基础上预测给定问题答案跨度的开始位置和起始位置。模型包括新闻文本关键句检索模块和事件要素抽取模块。

3.1 新闻文本关键句检索模型

基于BERT的语言模型在很多自然语言下游任务上取得了巨大的成功。BERT处理文本的最大长度是512个字符，而新闻文本长度远超512字符。由于新闻文本中包含的事件信息非常离散，采用传统滑动窗口的方法直接对新闻文本进行截取会造成新闻文本中关键信息的丢失。为了解决以上问题，我们训练了一个新闻文本句子打分模型，来动态检索出有可能包含答案的序列，我们称其为KSR。其基本假设是：一个新闻文本其主要围绕一个核心事件进行报到，其包含的冗余信息过多，而对于新闻主题事件要素抽取任务，我们只需要抽取新闻文本中几个关键句子即可。具体步骤包括将新闻文本数据切分成子序列和关键子序列的检索。

切分新闻文本数据：我们通过使用动态规划算法将新闻长文本P切分成$[\mathrm{X}_0,\mathrm{X}_1,...,\mathrm{X}_{\mathrm{T}-1}]$，算法1给出了切块的完整算法。由于BERT处理文本序列的最大长度是512字符，在具体实现中，我们根据不同的语种设置了不同子序列的最大值，在中文新闻文本数据中我们直接根据标点符号进行切分，而在越南语新闻文本数据中将其设为$B_{vi}=86$。在分块的基础上，我们根据标点符号的代价$cost$对其进行合并和排序，和原新闻文本P保持相同的顺序。

算法 1 新闻长文本分割算法

输入: 新闻长文本 P，标点成本 cost,最大句子长度 B.
输出: 分割后的新闻句子$[\mathrm{X}_0, \mathrm{X}_1, ..., \mathrm{X}_{\mathrm{T}-1}]$.

Initialize $f[0]...f[B-1]$ *as* 0
Initialize $from[0]...from[B-1]$ *as* -1
for each $i \in [B, len(P)-1]$ **do**
    $f[i] = +\infty$.
    for each j $\in [i-B, i-1]$ **do**
        if *word is punctuation* **then**
            $v = \cos t[word] + f[i]$
        else
            $v = \cos t[word] + f[i]$
        end if
        if $v < f[i]$ **then**
            $f[i] = v, from[i] = j$
        end if
    end for
end for
$t = len(P) - 1, blocks = []$
while $t \geq 0$ **do**
    prepend $p[from[t]+1...t]$ *to blocks*
    $t = from[t]$
end while
return *blocks*

检索关键句子： 其目标是从新闻长文本P中检索出关键句子Z，算法2给出了检索关键句子的完整算法。在新闻报道中，往往需要多个句子才能完整地描述一个主题事件，则新闻主题事件要素存在多个句子中，这就要求模型具有更强的语义理解能力和推断能力。该模块中我们训练了一个评分模型，对新闻长文本中的句子序列$[\mathrm{X}_0, \mathrm{X}_1, ..., \mathrm{X}_{\mathrm{T}-1}]$进行评分，以实现动态地抽取出和问题高度相关的句子。具体流程如下：我们首先将评分模型的输入初始化为$Z^+ = [[CLS]Q[SEP][X_0, X_1, ..., X_{T-1}]]$。然后训练一个基于mBERT（multilingual BERT）的一个评分模型$judge$，对每个子序列进行评分$judge([Z^+[SEP]X_i])[X_\mathrm{i}]$，将得分最高的子序列加到$Z$中，其中$len(Z) \leq L$。经过多次迭代推理最终得到新闻长文本的关键子序列$Z$。

$$judge(Z^+) = sigmoid(MLP(mBERT(Z^+))) \in (0,1)^{len(Z^+)} \tag{1}$$

每个子序列的分数为序列中每个词的平均分，将其表示为：$judge(Z^+)[X_i]$。

3.2 新闻事件要素抽取模型

新闻事件要素抽取的目的是从候选句Z中提取问题的答案。由于越南语的新闻事件要素抽取语料标注困难且稀缺，而中文具有大规模的机器阅读理解语料且中文新闻事件要素抽取语料标注相对简单。基于此，我们提出了一种新的基于跨语言知识迁移的新闻事件要素抽取模型。该模型可以同时对源语言和目标语言的训练数据进行建模，从而实现源语言到目标语言的知识迁移，提高目标语言的新闻事件要素抽取性能。其模型结构图如图2所示。

算法 2 检索新闻长文本关键句

输入: 新闻长文本 $P = [X_0, X_1, ..., X_{T-1}]$,strides = [strides0,...,strides m - 1]
输出: 关键句序列 Z

Initialize $Z = []$
for each $i \in [0, m-1]$ **do**
  for each $X_i \in [X_0, X_1, ..., X_{T-1}]$ **do**
    $score[X_i] = judge([Z\ X_i]^+)[X_i]$
  end for
  Fill Z up tovlength L with highest scoring blocks.
  $score[Z_0, Z_1, ...] = judge(Z^+)$.
  Retain $\sum_{j=1}^{i} stride_j$ *highest scoring blocks in Z.*
end for
return Z

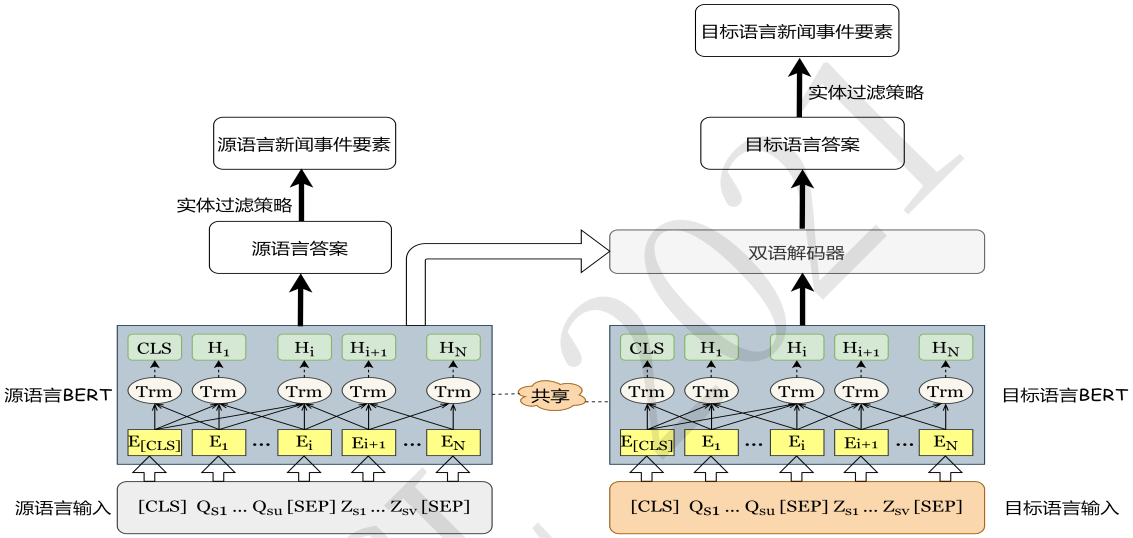

图 2. 基于跨语言知识迁移的新闻事件要素抽取模型图.

本小节主要从以下四个方面来介绍该模型的具体细节：1）问题生成，针对每种事件类型生成一组自然问题；2）双语编码器，利用mBERT（multilingual BERT）对源语言和目标语言进行编码，提取新闻文本特征；3）双语解码器，利用多头注意力机制来提取源语言知识，以提高目标语言答案抽取性能；4）新闻事件要素过滤层：通过启发式规则来过滤非法答案，得到新闻事件要素集。

3.2.1 问题生成

问题生成是连接新闻事件要素抽取任务和机器阅读理解任务的桥梁。问题的语义信息编码了关于新闻事件要素的先验知识。我们的方法使用一种基于模板的问题生成方法。值得注意的是，为了使生成的问题更加的自然，必须针对不同的语义角色选择不同的问题疑问词。比如，针对语义角色时间的问题疑问词应该是“何时”；针对地点的问题疑问词应该是“何地”；针对访问事件中的接见了谁/和谁举行了会谈应该以“谁”作为疑问词。基于此，我们首先将所有的语义角色划分为不同的类别：与时间相关的语义角色、与地点相关的语义角色、与人物相关的语义角色、与组织机构相关的语义角色，然后为每个类别设计不同的问题模板。不同问题疑问词如表1所示。进而根据问题疑问词生成问句表述。为了让问题编码先验知识，我们针对每个事件类型设置了如下的问题模板：｛事件触发词｝ + ｛疑问词｝。例如：触发词是[签署]，则问题可表示为：何时签署了《区域全面经济伙伴关系协定》?

类别	疑问词（中文）	疑问词（越南语）
时间	何时/什么时候	khi nào
地点	何地/在哪里	Ở đâu/ Nó đâu rồi
人物和组织机构	谁/哪个组织或公司	WHO/Tổ chức hoặc công ty nào

表 1: 问题疑问词模板.

3.2.2 双语编码器

BERT在自然语言处理领域具有里程碑的意义。BERT本质上是通过在大量语料的基础上利用自监督学习的方法为每个字或词学习一个好的特征表示。在该方法中，我们使用mBERT（multilingual BERT）来对中文和越南语的问题Q和新闻文本的关键子序列Z进行编码。给定越南语的问题Q_T和关键子序列Z_T，使用特殊字符$[CLS]$和$[SEP]$拼接成P_{T}输入到mBERT。

$$P_T = [CLS]Q_T[SEP]Z_T[SEP] \tag{2}$$

相应地，我们将中文问题Q_S和关键字序列Z_S拼接成P_S输入到mBERT中。值得注意的是，我们的中文和越南语的训练语料在事件级别上是对齐的，即描述的是同一类事件。由于新闻文本描述同一类核心事件的事件要素大体一致，则可通过共享mBERT编码器来实现源语言到目标语言的知识迁移。P_{T}和P_S经过编码后分别得到隐层表示$B_T \in {}^{L_T*h}$，$B_S \in {}^{Ls*h}$，其中L表示输入关键文本的长度，h表示mBERT的隐层大小。

3.2.3 双语解码器

为了提高越南语的新闻事件要素抽取性能，我们使用了一个多头注意力机制层来实现从中文到越南语的跨语言知识迁移，以提取对越南语答案抽取有用的信息。在此过程中，我们分别将目标语言BERT的深度表示B_T和源语言BERT的深度表示B_S看作多头注意力机制中的键和值，将其视为一个解码的过程。传统Transformer的多头注意力机制中的点乘注意力可以表示如下：

$$A_{TS} = B_T.B_S, A_{TS} \in {}^{L_T*L_S} \tag{3}$$

为了将中文的表示融合到越南语表示中，我们改进了Transformer的多头注意力机制。先计算B_T和B_S的自我注意力，如公式（4）和（5），其目的是先利用自我注意力机制来过滤和答案抽取无关的冗余信息。

$$A_T = \mathrm{softmax}(B_T.B_T^\top) \tag{4}$$

$$A_S = \mathrm{softmax}(B_S.B_S^\top) \tag{5}$$

进而我们计算A_T和A_S之间的注意力得到$\tilde{A}_{TS}$，进一步提升注意力计算的精度。在此基础上，我们通过公式（7）计算$\tilde{A}_{TS}$与B_S之间的注意力权重R'。进而我们通过残差连接和层归一化以获得最终的表示H_T。

$$\tilde{A}_{TS} = A_T.A_{TS}.A_T^\top, \tilde{A}_{TS} \in {}^{\mathrm{L}_T*L_S} \tag{6}$$

$$R' = \mathrm{softmax}(\tilde{A}_{TS}).B_S \tag{7}$$

$$R = W_r R' + b_r, W_r \in {}^{h*h} \tag{8}$$

$$H_T = \mathrm{softmax}[B_T, LayerNorm(B_T + R)] \tag{9}$$

最终我们利用H_T来预测目标语言越南语答案的跨度Z_T^s, Z_T^e。其中，答案的开始位置Z_T^s计算如公式（10）所示：

$$Z_T^s = \mathrm{softmax}(W_T^\top H_T + b), W_T \in {}^{2h} \tag{10}$$

进而我们可以计算其交叉熵损失。

$$L_T = -\frac{1}{N}\sum_{i=1}^{N}(y_T^S \log(Z_T^S) + y_T^e \log(Z_T^e)) \tag{11}$$

3.2.4 新闻事件要素过滤

由于同一个新闻文本中，描述其核心事件的事件要素会重复出现，而阅读理解任务并没有考虑这种情况。针对该问题，我们在新闻事件要素过滤层设计了以下几个启发式规则来过滤非法答案：（1）有效答案的开始位置在结束位置之前；（2）有效答案的似然概率Z_T^S应该大于一定的阈值δ,δ是模型的超参数。除此之外，在新闻文本中实体信息已知的情况下，可以进一步过滤答案，只保留和已知实体边界重合的答案。这种策略被称之为标准实体过滤（golden entity refinement）策略。最后模型将事件要素答案生成结果进行汇总，作为最终的新闻事件要素抽取结果。

4 实验与结果分析

4.1 3.1数据集与评价指标

实验数据集：在实验中，本文使用的数据集包含阅读理解中文数据集、中文事件可比数据集（即和越南语描述的相同类型事件的中文新闻文本）和越南语新闻事件要素抽取数据集。其中，阅读理解中文数据集采用CMRC 2018。到目前为止，还没有公开的汉越双语新闻事件要素抽取数据集，因此依据ACE2005数据集标准结合任务构建了汉越双语新闻事件要素抽取数据集。首先在越南网站爬取了708篇越南新闻文本，并根据抽取式阅读理解的形式进行标注。然后根据预先定义的事件类型的关键词爬取并筛选了932篇中文新闻，形成了中文事件可比数据集。

评价指标：为了保证和以往事件抽取工作的可比性，使用精确率（Precision），召回率（Recall）和F1值（F1）作为评价指标。考虑到越南语的数据集的规模偏小，我们进行了显著性检验（significance test），显著水平设为$\rho = 0.05$。

4.2 实验参数设置

本文采用mBERT（multilingual BERT）来构建新闻长文本关键子序列检索模型和阅读理解模型。在新闻长文本关键字序列抽取和新闻事件要素抽取两个阶段，均使用Adam算法对模型调优，可以自动调整训练过程中的学习率，其两个阶段的学习率分别为4×10^{-5}和10^{-4}。训练过程还引入了梯度裁剪策略，它可以加快模型训练的收敛速度。其实验相关参数设置如表3所示。

参数名	参数值
mBERT基础参数	默认值
答案预测阈值δ	0.3
批处理大小	32
训练轮数	4
strides	[3,5]

表 2: 实验参数设置.

4.3 实验结果与分析

4.3.1 基线方法

本文将提出的基于阅读理解的汉越跨语言新闻事件要素抽取方法与下述的事件要素抽取方法进行了对比。

• DMCNN：是由 (Chen et al., 2015)所提出的基于卷积神经网络 (Convolutional Neural Networks, CNNs) 的事件抽取方法；

• JRNN：它是由 (Nguyen et al., 2016)所提出的基于递归神经网络(Recurrent Neural Networks, RNNs) 的事件抽取方法；

• JMEE：它是由 (Liu et al., 2018b)所提出的事件抽取方法，使用图卷积神经网络 (Graph Convolutional Neural Networks, GCNs) 挖掘句法信息来进行事件抽；

• MTL-CRF：它是由 (He and Duan, 2019)提出的基于CRF的方法，设计了一个有效挖掘不同事件之间事件要素关系的多任务学习的序列标注模型，同时抽取了事件触发词和事件要素；

• DMBERT：它是由 (Wang et al., 2019a)提出的的有效利用预先训练语言模型的方法并使用动态多池化方法来聚合特征。为了与本文提出的方法进行比较，本文将其预训练语言模型修改为mBERT；

• RCEE：它是由 (Liu et al., 2020) 提出的基于机器阅读理解（MRC）的事件抽取方法，在检测事件的基础上进行事件要素抽取；

同时，也将本文提出的方法与目前篇章级的事件要素抽取模型进行对比，这些方法包括：

• DCFEE：它是由 (Zheng et al., 2019)提出的一种端到端模型从财务公告抽取金融事件的信息，通过将事件结构化任务转换为路径扩展子任务，进而生成一个基于实体的有向无环图实现文档级事件抽取。

• MGRDEE： (Du and Cardie, 2020)提出了句子级和段落级的多粒度编码器，以动态融合句子级和段落级特征，提高了篇章级的事件要素抽取性能。

方法	精确率	召回率	F1值
DMCNN	51.9	47.6	49.7
JRNN	48.4	50.6	49.5
JMEE	60.3	47.9	53.64
MTL-CRF	63.7	46.8	54.1
DMBERT	58.9	50.2	54.2
RCEE	58.2	59.3	58.7
DCFEE	62.9	60.3	61.6
MGRDEE	**63.8**	59.7	61.7
Ours	63.4	**63.7**	**63.5***

表 3: 汉越跨语言新闻事件要素抽取实验结果（%）.

从表3中的实验结果可以看出，本文提出的基于阅读理解的汉越跨语言新闻事件要素抽取方法优于其他方法。其中 $*$ 代表显著性水平为$\rho = 0.05$。和篇章级事件要素抽取方法MGRDEE方法进行对比发现，我们的方法在召回率（R）和F1值上有明显的提升，召回率和F1值分别提升了4.0%和1.8%。篇章级事件抽取方法MGRDEE动态融合了句子级的局部特征和篇章级的全局特征，有效的提高了事件要素抽取的精确率（P），其可能原因是本文提出的方法首先利用关键句检索出与答案相关度高的句子，再基于候选句抽取出新闻主题事件要素，在此过程中可能会带来一定误差传递，进而影响了模型的精确率，但基于序列标注的事件要素抽取方法会导致数据稀疏，召回率较低。对于篇章级事件抽取模型DCFEE，以金融文档中的一个关键句子为事件中心句进行要素的补充，由于在新闻报到中，描述一个核心事件往往需要多个句子，新闻事件的事件要素分散在不同的句子中，从而导致了该模型在本文所构建的数据集上效果不明显。同样，在多分类任务中，对于部分新闻事件要素标注较少的类别很难识别。

综上可以看出：1）在所有的方法中，在越南语有部分标注语料的前提下，本文提出的方法基本上取得了最好的性能，篇章级新闻事件要素抽取方面显著优于其他方法（F1值提升了1.8%），这直接证明了方法的有效性。2）本文提出的方法取得了较高的召回率，这说明了与其他的方法相比，其可以预测更多的样例。

4.3.2 消融实验

为了进一步的验证本文方法的有效性，本文分别设置了以下两组消融实验：新闻文本关键句检索模型和跨语言知识迁移对模型性能的影响。从表4中的实验结果我们可以看出，新闻文本关键句检索可以有效捕获篇章级局部和全局语义信息，明显提升了越南语新闻事件要素抽取模型性能（F1相差4.4%）。其中 $*$ 代表显著性水平为$\rho = 0.05$。由于新闻长文本中含有大量的冗余信息，让模型直接对新闻文本进行语义建模，对给模型带入大量的噪声数据，从而影响了越南语新闻事件要素抽取模型的性能。

为了验证跨语言知识迁移对越南语新闻事件要素抽取模型性能的影响，本文在设置了如

方法	精确率	召回率	F1值
Ours	**63.4**	**63.7**	**63.5***
w/o KSR	58.7	59.6	59.1(-4.4)

表 4: 新闻文本关键句检索模块消融实验结果（%）.

表5所示的对比实验。其中 $*$ 代表显著性水平为$\rho = 0.05$。从实验结果可以看出，如果不使用中文阅读理解语料和中文可比事件语料进行预训练模型，越南语新闻事件要素抽取模型的性能会显著下降（F1值分别下降了3.8%和1.6%），这也证明了模型通过共享编码器，实现了跨语言知识的迁移，提高了越南语新闻事件要素抽取性能

方法	精确率	召回率	F1值
Ours	**63.4**	**63.7**	**63.5***
w/o Per-Train	59.2	60.3	59.7(-3.8)
w/o Source BERT	61.6	62.2	61.9(-1.6)

表 5: 跨语言知识迁移消融实验结果（%）.

5 结论

本文提出了基于阅读理解的汉越跨语言新闻事件要素抽取方法。该方法包含一个新闻文本关键句检索模块过滤包含噪声的句子和一个跨语言新闻事件要素抽取模块提取候选句子中的新闻事件要素。该方法通过过滤新闻长文本中的噪声数据有效解决篇章语义建模困难的问题，并通过跨语言知识迁移提高资源稀缺型语言的新闻事件要素抽取性能。通过实验证明我们提出的方法在越南语有部分标注语料的条件下优于现有的事件要素抽取模型。在下一步的研究工作中，我们进一步研究在目标语言没有标注语料的前提下，基于半监督或者无监督的新闻事件要素抽取方法。

参考文献

Akari Asai, Akiko Eriguchi, Kazuma Hashimoto, and Yoshimasa Tsuruoka. 2018. Multilingual extractive reading comprehension by runtime machine translation. *arXiv preprint arXiv:1809.03275*.

Yubo Chen, Liheng Xu, Kang Liu, Daojian Zeng, and Jun Zhao. 2015. Event extraction via dynamic multipooling convolutional neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 167–176.

Danqi Chen, Jason Bolton, and Christopher D Manning. 2016. A thorough examination of the cnn/daily mail reading comprehension task. *arXiv preprint arXiv:1606.02858*.

Yiming Cui, Zhipeng Chen, Si Wei, Shijin Wang, Ting Liu, and Guoping Hu. 2016. Attentionover-attention neural networks for reading comprehension. *arXiv preprint arXiv:1607.04423*.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2019. Crosslingual machine reading comprehension. *arXiv preprint arXiv:1909.00361*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pretraining of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ming Ding, Chang Zhou, Hongxia Yang, and Jie Tang. 2020. Cogltx: Applying bert to long texts. *Advances in Neural Information Processing Systems*, 33.

Xinya Du and Claire Cardie. 2020. Documentlevel event role filler extraction using multi-granularity contextualized encoding. *arXiv preprint arXiv:2005.06579*.

Shaoyang Duan, Ruifang He, and Wenli Zhao. 2017. Exploiting document level information to improve event detection via recurrent neural networks. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 352–361.

RF He and SY Duan. 2019. Joint chinese event extraction based multi-task learning. *J. Softw*, 30(4):1015–1030.

Karl Moritz Hermann, Tomáš Kočiskỳ, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *arXiv preprint arXiv:1506.03340*.

Yu Hong, Jianfeng Zhang, Bin Ma, Jianmin Yao, Guodong Zhou, and Qiaoming Zhu. 2011. Using crossentity inference to improve event extraction. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 1127–1136.

Andrew Hsi, Yiming Yang, Jaime G Carbonell, and Ruochen Xu. 2016. Leveraging multilingual training for limited resource event extraction. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1201–1210.

Heng Ji and Ralph Grishman. 2008. Refining event extraction through crossdocument inference. In *Proceedings of ACL-08: Hlt*, pages 254–262.

Qi Li, Heng Ji, and Liang Huang. 2013. Joint event extraction via structured prediction with global features. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 73–82.

Manling Li, Ying Lin, Joseph Hoover, Spencer Whitehead, Clare Voss, Morteza Dehghani, and Heng Ji. 2019. Multilingual entity, relation, event and human value extraction. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 110–115.

Shasha Liao and Ralph Grishman. 2010. Using document level crossevent inference to improve event extraction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 789–797.

Shulin Liu, Kang Liu, Shizhu He, and Jun Zhao. 2016. A probabilistic soft logic based approach to exploiting latent and global information in event classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

Jian Liu, Yubo Chen, Kang Liu, and Jun Zhao. 2018a. Event detection via gated multilingual attention mechanism. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Xiao Liu, Zhunchen Luo, and Heyan Huang. 2018b. Jointly multiple events extraction via attention-based graph information aggregation. *arXiv preprint arXiv:1809.09078*.

Jian Liu, Yubo Chen, Kang Liu, Wei Bi, and Xiaojiang Liu. 2020. Event extraction as machine reading comprehension. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1641–1651.

Di Lu, Ananya Subburathinam, Heng Ji, Jonathan May, Shih-Fu Chang, Avirup Sil, and Clare Voss. 2020. Crosslingual structure transfer for zero-resource event extraction. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 1976–1981.

Jie Ma, Shuai Wang, Rishita Anubhai, Miguel Ballesteros, and Yaser Al-Onaizan. 2020. Resource-enhanced neural model for event argument extraction. *arXiv preprint arXiv:2010.03022*.

Thien Huu Nguyen and Ralph Grishman. 2015. Event detection and domain adaptation with convolutional neural networks. In *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)*, pages 365–371.

Thien Huu Nguyen, Kyunghyun Cho, and Ralph Grishman. 2016. Joint event extraction via recurrent neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 300–309.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2016. Bidirectional attention flow for machine comprehension. *arXiv preprint arXiv:1611.01603*.

Ananya Subburathinam, Di Lu, Heng Ji, Jonathan May, Shih-Fu Chang, Avirup Sil, and Clare Voss. 2019. Crosslingual structure transfer for relation and event extraction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 313–325.

Xiaozhi Wang, Xu Han, Zhiyuan Liu, Maosong Sun, and Peng Li. 2019a. Adversarial training for weakly supervised event detection. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 998–1008.

Xiaozhi Wang, Ziqi Wang, Xu Han, Zhiyuan Liu, Juanzi Li, Peng Li, Maosong Sun, Jie Zhou, and Xiang Ren. 2019b. Hmeae: Hierarchical modular event argument extraction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5781–5787.

Sen Yang, Dawei Feng, Linbo Qiao, Zhigang Kan, and Dongsheng Li. 2019. Exploring pretrained language models for event extraction and generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5284–5294.

Alexander Yeh. 2000. More accurate tests for the statistical significance of result differences. *arXiv e-prints*, pages cs–0008005.

Yue Zhao, Xiaolong Jin, Yuanzhuo Wang, and Xueqi Cheng. 2018. Document embedding enhanced event detection with hierarchical and supervised attention. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 414–419.

Shun Zheng, Wei Cao, Wei Xu, and Jiang Bian. 2019. Doc2edag: An endtoend document-level framework for chinese financial event extraction. *arXiv preprint arXiv:1904.07535*.

Zhu Zhu, Shoushan Li, Guodong Zhou, and Rui Xia. 2014. Bilingual event extraction: a case study on trigger type determination. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 842–847.