

基于DCNN和GLU的武器领域实体关系抽取方法

李晗, 侯守璐, 佟强, 谌彤童, 杨启民, 刘秀磊

引用本文

李晗, 侯守璐, 佟强, 谌彤童, 杨启民, 刘秀磊. 基于DCNN和GLU的武器领域实体关系抽取方法[J]. 计算机科学, 2023, 50(6A): 220200112-7.

LI Han, HOU Shoulu, TONG Qiang, CHEN Tongtong, YANG Qimin, LIU Xiulei. [Entity Relation Extraction Method in Weapon Field Based on DCNN and GLU](#) [J]. Computer Science, 2023, 50(6A): 220200112-7.

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)

[基于句间信息的图注意力卷积网络的文档级关系抽取](#)

Document-level Relation Extraction of Graph Attention Convolutional Network Based on Inter-sentence Information

计算机科学, 2023, 50(6A): 220800189-6. <https://doi.org/10.11896/jsjcx.220800189>

[文档级关系抽取技术研究综述](#)

Review of Document-level Relation Extraction Techniques

计算机科学, 2023, 50(5): 189-200. <https://doi.org/10.11896/jsjcx.220400252>

[PosNet:基于位置的因果关系抽取网络](#)

PosNet: Position-based Causal Relation Extraction Network

计算机科学, 2022, 49(12): 305-311. <https://doi.org/10.11896/jsjcx.211100264>

[边云协同计算中成本感知的物联网数据处理方法](#)

Cost-aware IoT Data Processing in Edge-Cloud Collaborative Computing

计算机科学, 2022, 49(11A): 211000101-7. <https://doi.org/10.11896/jsjcx.211000101>

[基于深度学习与文本计量的技术趋势分析](#)

Analysis of Technology Trends Based on Deep Learning and Text Measurement

计算机科学, 2022, 49(11A): 211100119-6. <https://doi.org/10.11896/jsjcx.211100119>

基于 DCNN 和 GLU 的武器领域实体关系抽取方法

李 晗¹ 侯守璐¹ 佟 强^{1,2} 谌彤童³ 杨启民¹ 刘秀磊^{1,2}

1 北京信息科技大学数据与科学情报分析实验室 北京 100101

2 北京材料基因工程高精尖创新中心(北京信息科技大学) 北京 100101

3 北京跟踪与通信技术研究所 北京 100094

(yhlihan@126.com)

摘 要 武器领域的非结构化文本数据通常十分复杂,单句内可能存在“一武器与多个武器相关联”或“两武器之间存在多种关系”等情况,为此提出基于膨胀卷积神经网络和门控线性单元的实体关系抽取方法以处理该类型数据中存在的关系重叠问题。该方法将拼接了词向量和位置向量的句子编码向量传入带有门控机制的膨胀卷积神经网络模型,引入可以快速抽取句内命名实体特征信息的自注意力机制,通过分层次的序列标注方式识别出句中全部实体以及每个主实体对应的所有关系和客实体,进而生成武器领域实体关系三元组。实验结果显示,该方法在自行标注的武器领域数据集上的 F1 值达 81.1%,具备一定的实体关系抽取能力,在不同重叠类型下的 F1 值均高于 78%,能够解决非结构化数据的关系重叠问题,同时在公开数据集 NYT 上也有良好的表现。

关键词: 关系抽取;关系重叠;膨胀卷积神经网络;门控线性单元

中图法分类号 TP391

Entity Relation Extraction Method in Weapon Field Based on DCNN and GLU

LI Han¹, HOU Shoulu¹, TONG Qiang^{1,2}, CHEN Tongtong³, YANG Qimin¹ and LIU Xiulei^{1,2}

1 Laboratory of Data Science and Information Studies, Beijing Information Science and Technology University, Beijing 100101, China

2 Beijing Advanced Innovation Center for Materials Genome Engineering, Beijing Information Science and Technology University, Beijing 100101, China

3 Beijing Institute of Tracking and Telecommunications Technology, Beijing 100094, China

Abstract Unstructured text data in the field of weapons is usually very complex. In a single sentence, one weapon may be associated with multiple weapons or there may be multiple relations between two weapons. An entity relation extraction method based on dilated convolutional neural network and gated linear unit is proposed to solve the problem of overlapping relation in this type of data. This method introduces the sentence coding vector into the dilated convolutional neural network model with gated linear unit, which combines word vector and position vector. And it introduces the self-attention mechanism to extract the feature information of entities in sentences quickly. Through hierarchical sequence annotation, this model identifies all entities in the sentence and all relations and object entities corresponding to each subject entity, and generates the entity relation triplet in the field of weapons. The F1 value of this method on the self-labeled weapon field data set is 81.1%, and it has a certain entity relation extraction ability, according to the experimental results. The F1 value for various overlap types is greater than 78%, which solves the problem of unstructured data relation overlap. At the same time, it performs admirably on the NYT public data set.

Keywords Relation extraction, Overlapping relation, Dilated convolutional neural network, Gated linear unit

1 引言

伴随着信息科技的迅速发展,自然语言处理、机器学习、计算机视觉等人工智能技术逐渐深入军事领域并推动着新一轮军事变革。从结构化数据库、半结构化页面、非结构化文本等多类型数据中快速获取有价值信息,从而形成相关军事

武器知识库,能够有效提高情报资源利用率,更好地辅助研究人员开展军事知识的应用研究。

在武器领域的实体关系抽取实际应用场景中,存在着关键数据获取成本高、各国武器命名规则不统一、隐含信息抽取难度大、实体关系重叠、信息孤岛等问题,致使抽取效果不尽如人意。由于问题过于繁杂,故本文仅探讨武器领域非结构

基金项目:国家重点研发计划(2021YFB2600600);促进高校分类发展重点研究培育项目(2121YJPY225,2121YJPY226);北京信息科技大学校研基金项目;科研机构创新能力建设-数据科学与情报分析研究所

This work was supported by the National Key R&D Program of China(2021YFB2600600), Promoting the Classified Development of Universities-Key Research and Cultivation Projects(2121YJPY225,2121YJPY226), Natural Science Foundation of Beijing Information Science & Technology University and Innovation Capacity Building of Scientific Research Institutions-Institute of Data Science and Information Analysis.

通信作者:刘秀磊(liuxiulei@bistu.edu.cn)

化文本数据中的关系重叠问题。

通过观察我们可以发现,武器领域的文本数据往往十分复杂,从单句话中可能抽取不同重叠类型的三元组,除三元组内无相同实体(Normal)的情况外,还可能在不同三元组包含相同实体对(Entity Pair Overlap,EPO)和不同三元组包含同一实体(Single Entity Overlap,SEO)等诸多情况。例如,“ZTZ-99 主战坦克由中国研制并率先投入使用”中存在实例“ZTZ-99 主战坦克”和“中国”,它们分别属于“坦克装甲车辆”和“国家”这两类实体,根据预定义的本体结构,上述实体间存在“制造国”和“使用国”的关系,因而需要从句中抽取(ZTZ-99 主战坦克,制造国,中国)和(ZTZ-99 主战坦克,使用国,中国)两个实体关系三元组。

但实际上目前大部分方法并不能准确抽出全部三元组,因为它们将实体关系抽取视为“一对一”的抽取任务,即默认单句话中只存在一组实体对及其之间的关系,忽略了三元组共享同一实体对或仅共享一个实体的情况,从而导致模型处理该类型数据时 F1 值较低,抽取效果较差。

近些年,许多学者对实体关系抽取任务中存在的关系重叠问题进行了研究。2018 年 Zeng^[1]提出一种基于复制机制的端到端实体关系抽取模型,该模型通过机器翻译的编码方式,用指针依次标注关系、主语和谓语,虽然能够从不同重叠程度的句子中联合提取关系事实,但忽略了关系与命名实体之间的内部联系。随后,Wei 等^[2]提出一种新型级联二元标记框架 HBT(又称 CasRel),其弥补了流水线式抽取带来的错误传播问题,在 NYT 和 WebNLG 上均表现优异,但该方法并不能识别嵌套实体。对此,Wang 等^[3]提出单阶段联合抽取的 TPLinker 模型,该模型在解决关系重叠和实体嵌套问题的同时也有效避免了暴露偏差,实验结果显示其在 NYT 数据集上的 F1 值较 HBT 提升了 2.3%。

尽管这些方法均取得了一定成效,但针对武器领域非结构化文本数据的关系重叠问题仍存在较大空白。为填补上述空缺,本文提出了应用于武器领域的基于膨胀卷积神经网络(Dilated Convolutional Neural Network,DCNN)和门控线性单元(Gated Linear Units,GLU)的实体关系抽取方法。该方法将训练好的词向量与位置向量相拼接并传入带有 GLU 机制的 DCNN 模型,使用自注意力机制捕获句子编码向量的语义信息,最后运用指针标注法识别句中全部命名实体。

此方法将主实体与客实体通过关系来做相互映射,通过分层次的序列标注方式抽取每个主实体具有的所有武器领域关系以及与之相对应的全部客实体,以解决关系重叠问题。除此之外,为了在模型中融合主实体与客实体的信息,增强两者之间的依赖性,当抽取某一主实体的关系和客实体时,会将该主实体的编码向量作为先验特征传入双向长短期记忆网络(Bi-directional Long Short-Term Memory,Bi-LSTM)模型,并把输出结果与含有主实体相对位置信息的向量进行拼接,在得到与句子编码向量等长的新向量后,完成关系和客实体的识别、标注工作。

2 相关工作

实体关系抽取需要完成命名实体识别(Named Entity Recognition,NER)和关系抽取(Relation Extraction,RE)两项任务^[4],前者主要负责抽取具有特定实体类型的命名实体,而

后者则负责抽取命名实体间的预定义关系并生成形如(主实体,关系,客实体)的实体关系三元组。

2.1 命名实体识别

早期由于技术的限制,一般使用基于字典和规则的方法来完成命名实体识别工作。该方法能够根据应用领域的特点制定相应的规则,因而拥有优异的抽取效果,但过于依靠人工也使得构建成本极高且难以处理大量数据。为此,人们采用机器学习方法以分类或序列标注的方式完成 NER 任务,降低了模型对人工定义特征的依赖^[5],例如隐马尔可夫模型(Hidden Markov Model,HMM)^[6]、条件随机场(Conditional Random Fields,CRF)^[7]、支持向量机(Support Vector Machine,SVM)等模型。

但是机器学习必须依靠庞大的标注数据才能完成训练工作,数据准备环节成本极高。因此,人们将深度学习思想引入 NER 任务,利用卷积神经网络、循环神经网络(Recurrent Neural Network,RNN)、注意力机制(Attention Mechanism,AM)等神经网络模型训练词向量后,由 CRF 输出命名实体的标注结果。

近些年,出现了许多表现优异的 NER 模型。Li 等^[8]提出了 Dice Loss 模型,该模型采用了基于自适应损耗函数的方法来处理数据不均衡的问题,在 CoNLL-2003 和 MSRA 上的 F1 值分别为 93.33% 和 96.72%。Zhong^[9]提出的基于跨度结构的 Ours: cross-sentence ALB 模型,改变了传统标注方式,在 ACE 2005 公开数据集上达到了 SOTA,较先前最优模型 BERT-MRC^[10]的 F1 值增长 4.02%。除此之外,ACE+document-context 模型^[11]、Co-regularized LUKE 模型^[12]、FLERT XLM-R 模型^[13]和 PL-Marker 模型^[14]也表现极佳,在 CoNLL-2003 上的 F1 值均超过 94%。

2.2 关系抽取

关系抽取按技术发展过程可分为规则模板、机器学习、深度学习和远程监督学习四大类,其中深度学习方法应用最为广泛,该方法包括流水线和联合学习两种抽取方式^[15]。

2.2.1 流水线抽取

在流水线抽取方法中,NER 任务和 RE 任务相互独立,即先识别命名实体再进行关系分类。例如 Santos^[16]提出了排序卷积神经网络 CR-CNN,该模型通过排序的方式来处理关系分类任务,使用的成对排名损失函数能够有效减少人工分类对模型的影响。Wang 等^[17]将依赖两个层次的 Attention 机制引入 Att-Pooling-CNN 模型,分别计算句中词语与实体对的注意力以及词语与各实体概念的注意力,这种多层次注意力机制能够较好地适应异构上下文模式。Wang 等^[18]提出的 Entity-Aware BERT 模型仅需对输入文本进行一次编码,即可抽取出句内实体的多个关系,大幅降低了数据传递过程中的计算成本,有效提升了模型处理长文本语料时的运行效率。Cohen^[19]把有监督的关系分类任务转化为类似问答(Question Answering,QA)的跨度预测(Span-Prediction,SP)问题,在数据集 SemEval-2010 Task 8 上达到 SOTA。

流水线这种分阶段的处理方式尽管操作简单、自由灵活,但却忽略了两个任务之间的联系,因此极易导致误差信息不断传播,影响三元组抽取效果。

2.2.2 实体关系联合抽取

实体关系联合抽取方法对流水线式进行了改进,它利用

统一的模型,通过共享参数或标注序列的手段来完成 NER 和 RE 任务。

Miwa^[20]于2016年提出的一种经典端到端(End-to-End)模型 SPTree,利用词序和依存树结构信息完成实体关系抽取工作。次年,Zheng 等^[21]提出了新型标注方案 NovelTagging,其以就近原则的序列标注方式抽取三元组,取得了不错效果。但该方法未考虑关系重叠问题,故 Bekoulis 等^[22]进行了改进,他将 RE 看作每个关系独立的多分类任务,利用 multi-head 抽取句中复杂的实体关系。

2018 年文献[1]率先对实体关系重叠问题进行研究,提出一种引入复制机制的 Seq2Seq 模型,使用编码器将文本转化为一定长度的语义向量,使用 One Decoder 和 Multi Decoder 两种解码器抽取并生成多种类型的实体关系三元组。为了解决传统 Seq2Seq 方法抽取关系时模型需要关注三元组顺序而易造成大量资源浪费的问题,Sui^[23]提出了 SPN 模型,利用集合预测思想摆脱了预测多个三元组顺序的负担,通过 transformer 内部的多头自注意力机制预测输入句子中全部可能的三元组关系,结果显示在 NYT 和 WebNLG 上该模型拥有 92.5%和 93.4%的极高 F1 值。联合学习方法虽然提升了模型运行性能,但也带来了曝光偏差问题,对此 TPLinker 采用单阶段联合抽取模式,通过单实体内头尾握手、实体对间头握手、实体对间尾握手这 3 种握手方式标记出每种关系类型的实体边界,从而确保训练阶段与推理阶段一致,有效解决了实体嵌套和多关系重叠问题。目前 REBEL^[24]是实体关系联合抽取的 SOTA 模型,它依靠生成式方法处理关系提取和关系

分类任务,使用自回归转换器模型将输入文本中可能存在的所有三元组抽取出来。在公开数据集 NYT 上该模型的 F1 值相较于 TPLinker 模型提高了 1.5%,相较于 SPN 模型提高了 0.9%。

3 基于 DCNN 和 GLU 的武器领域实体关系抽取模型

为了解决三元组重叠问题,本文将关系分类任务转化为类似于 Seq2Seq 模型的序列分层标注任务:首先抽取句子 x 中的全部主实体($s_1, s_2, \dots, s_n$),然后将其中一个主实体 s_i 作为条件,针对预定义的每一种关系($r_1, r_2, \dots, r_m$),依次标注出与该主实体对应的客实体($o_{i1}, o_{i2}, \dots, o_{im}$)并生成 (s, r, o) 形式的三元组,例如 (s_i, r_1, o_{i1}) 。重复上述步骤,直到抽取句中每一个 s 所对应的全部 r 和 o ,如式(1)所示。

$$P(s, r, o | x) = P(s | x)P(r, o | s, x) \tag{1}$$

这种序列分层标注方法将主实体与客实体通过关系来做相互映射,一定程度上保证了三元组的完整性,能够解决三元组重叠问题。

图 1 是本文提出的基于 DCNN 和 GLU 的实体关系抽取方法的模型结构图。以抽取出的第一个主实体“Pz68 主战坦克”为例,针对预定义的关系“产国”,模型抽取出了对应的客实体“瑞士”;针对预定义的关系“生产商”,模型抽取出了对应的客实体“联邦制造厂”;针对预定义的“前型”等其他关系,模型没有抽取出相关客实体。因此模型可以抽取出两个以“Pz68 主战坦克”作为主实体的关系三元组,分别是(Pz68 主战坦克,制造国,瑞士)和(Pz68 主战坦克,生产商,联邦制造厂)。

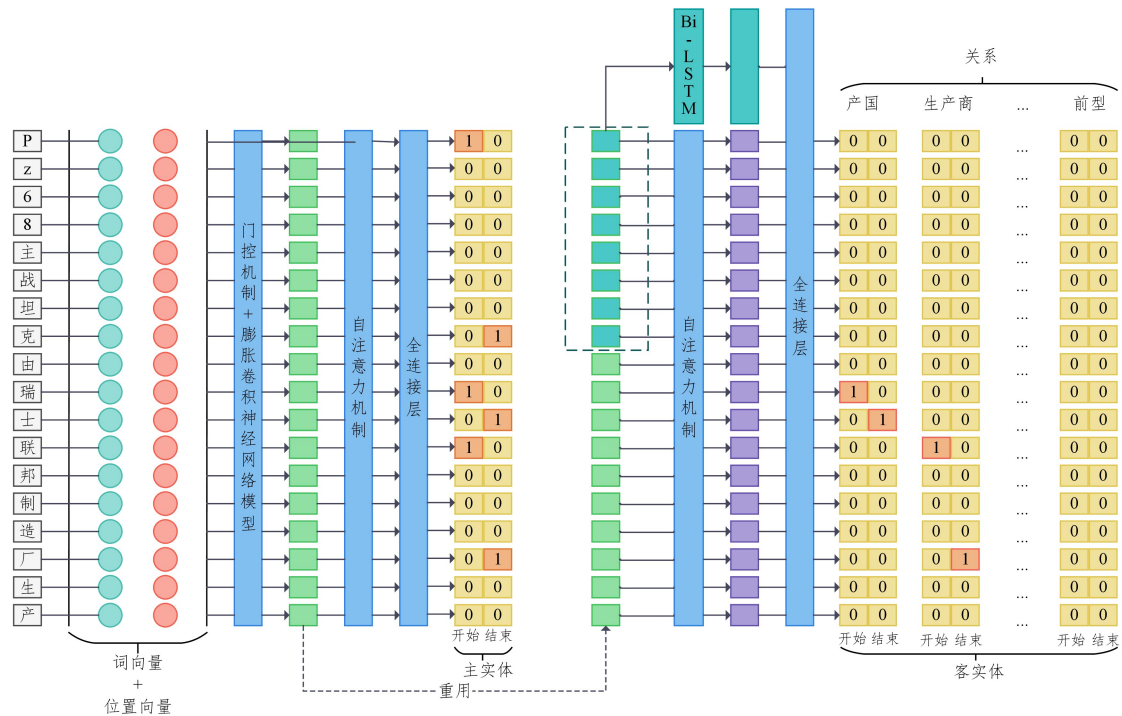


图 1 基于 DCNN 和 GLU 的实体关系抽取模型结构

Fig. 1 Structure of entity relation extraction model based on DCNN and GLU

3.1 词嵌入

本文模型首先使用 jieba 中文分词工具完成武器领域文本数据预处理工作,然后利用 Mikolov 等^[25]提出的基于深度学习思想的 Word2Vec 工具训练编码向量,最终得到词向量 $W^{w2v} = \{w_1^{w2v}, w_2^{w2v}, \dots, w_n^{w2v}\}$ 。

由于抽取命名实体时,模型需要记录候选实例名在句中的开始和结束位置,并且当主实体作为先验特征传入关系预测层时,模型也需要计算该实体在句中的相对位置,因此本文引用 Vaswani 等^[26]提出的式(2)来计算位置向量。

$$\begin{cases} PE_{2i}(p) = \sin \frac{p}{1000^{\frac{2i}{d_{pos}}}} \\ PE_{2i+1}(p) = \cos \frac{p}{1000^{\frac{2i}{d_{pos}}}} \end{cases} \quad (2)$$

该公式将句子的第 p 位与 d_{pos} 维的位置向量相互映射, 向量偶数位元素 $2i$ 的大小为 $PE_{2i}(p)$, 奇数位元素 $2i+1$ 的大小为 $PE_{2i+1}(p)$ 。其中, d_{pos} 等于词向量的最大维度。利用此方法可得到字符的位置向量 $W^{pos} = \{\omega_1^{pos}, \omega_2^{pos}, \dots, \omega_n^{pos}\}$ 。

把相同维度的词向量 W^{w2v} 和位置向量 W^{pos} 相加, 即可得到最终的嵌入向量 W , 如式(3)所示。

$$W = W^{w2v} + W^{pos} \quad (3)$$

3.2 编码器

3.2.1 膨胀卷积神经网络

为提升模型捕获词语特征的能力, 本文所提模型使用膨胀卷积结构替代普通卷积进行训练, 完成文本向量编码工作。与普通卷积相比, 当两者参数相同时, 前者拥有更大的感受野 (Receptive Field, RF), 能够更好地获取全局信息。当网络层数增加时, 普通卷积的等效卷积核始终不变, 感受野与层数呈线性相关的关系; 在相同条件下, 随着每层膨胀率指数的增长, 膨胀卷积的等效卷积核和感受野也会呈指数增长。

膨胀卷积的感受野 RF 由卷积核 k 、膨胀率 d 、步长 S 等参数共同确定, 当 $d=1$ 时表示普通卷积。输入层的感受野值为 1, 第 $i(i \geq 1)$ 层等效卷积核 K_i 和感受野 RF_i 的计算分别如式(4)和式(5)所示。

$$K_i = k_i + (k_i - 1)(d_i - 1) \quad (4)$$

其中, k_i 是第 i 层卷积核的值, d_i 是第 i 层的膨胀率。

$$RF_i = RF_{i-1} + (K_i - 1) \times \prod_{j=1}^{i-1} S_j \quad (5)$$

其中, RF_{i-1} 表示第 $i-1$ 层膨胀卷积的感受野大小, $\prod_{j=1}^{i-1} S_j$ 表示第 $i-1$ 层及之前所有层的步长之积。根据式(4)和式(5)可推出式(6)。

$$RF_i = RF_{i-1} + (k_i - 1) \times \prod_{j=1}^{i-1} S_j \times d_i \quad (6)$$

当 $i=0$ 时, $RF_0=1$ 表示输入层。

3.2.2 门控机制

为了解决网络层数加深带来的过拟合问题, 本文在普通一维卷积中添加门控机制, 利用门卷积替代递归网络中常用的递归连接, 同时引入残差结构增强词语间的关联程度, 从而防止出现梯度消失现象。

假设待处理序列为 $X=[x_1, x_2, \dots, x_n]$, 那么添加了门控机制的一维卷积的输出结果可用式(7)表示。

$$Y = \text{Conv1D}_1(X) \otimes \sigma(\text{Conv1D}_2(X)) \quad (7)$$

其中, Conv1D_1 和 Conv1D_2 的卷积核数、窗口大小均相同, 唯一不同的是后者加入了一个如式(8)所示的 sigmoid 激活函数。将 Conv1D_1 和被 sigmoid 函数激活后的 Conv1D_2 逐位相乘, 则可得输出结果。

$$\sigma(x) = \frac{1}{1 + e^{-x}} \quad (8)$$

接下来利用式(9)的残差结构进一步降低梯度消失的可能, 同时增加信息传输通道。

$$Y = X + \text{Conv1D}_1(X) \otimes \sigma(\text{Conv1D}_2(X)) \quad (9)$$

由于 Conv1D_1 没有添加 sigmoid, 因此它只是一个线性变换, Conv1D_1 与 $\text{Conv1D}_1(X) - X$ 等效。令 $\alpha = \sigma(\text{Conv1D}_2(X))$, 那么可将式(9)改写为式(10)。

$$Y = X \otimes (1 - \alpha) + \text{Conv1D}_1(X) \otimes \alpha \quad (10)$$

由上式可知, 在待处理序列 X 中, 有 $1-\alpha$ 概率的信息可直接通过控制门, 另外 α 概率的信息需激活函数处理后才可通过。

3.3 解码器

本文所提模型将自注意力机制引入解码环节, 以降低模型对外部信息的依赖, 利用词语间的相似性来挖掘序列内部的关联。

该机制主要包括 7 个步骤: 1) 预处理数据得到嵌入向量; 2) 初始化权重矩阵 W_Q, W_K, W_V ; 3) 利用输入向量和权重矩阵分别计算 Q, K, V 3 个向量矩阵; 4) 计算每个向量的注意力得分 QK^T ; 5) 数据归一化; 6) 给归一化后的注意力得分添加 softmax 激活函数; 7) 计算加权的输入向量评分; 8) 计算评分总和, 得到输出结果。计算如式(11)所示。

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (11)$$

其中, d_k 为权重矩阵 W_K 的维度, 在归一化时该参数可以调节注意力得分的大小。

4 实验

本文提出的模型使用 Keras2.2.4 和 Tensorflow1.8 的深度学习框架, 服务器内存 8GB, 搭载 Linux 操作系统, 采用 Python2.7 编程语言。

本次实验中, 词向量和位置向量的维度为 200, batch 为 64, 模型最大迭代次数为 300, 使用膨胀率依次为 1, 2, 5 的 12 层神经网络, dropout 为 0.2。利用 Adam 算法优化参数, 初始学习率为 0.001。

4.1 验数据

本文在自行标注的武器领域数据集和公开数据集上分别进行实验, 以验证模型效果。

4.1.1 武器领域数据集

本文选取武器领域中内容较丰富的百科类网页, 通过地理范围、实体类型和数据语言 3 个方面去除噪音及不相关数据, 筛选数据后通过标注、审核生成武器领域数据集, 共计 4000 条句子。该数据集涵盖 8 类实体以及 13 种关系类型, 如实体类型“部队”“国家或地区”“制造厂”“轻装武器”“炮弹与导弹”“坦克装甲车辆”等, 关系“产国”“生产商”“前型”“装备”等。本实验以 8:1:1 的比例划分数据集, 即 3200 条训练数据用于训练模型, 400 条验证数据用于评判模型参数, 400 条测试数据用于评估模型泛化能力。

根据数据集中单句内三元组信息的重叠情况, 将数据分为 Normal, EPO 和 SEO 3 种重叠类型。在训练集和测试集中, 句内无实体重叠的句子共有 822 条, 句内共享一个实体对的句子共有 44 条, 句内仅共享一个实体的句子共有 2759 条。具体数量如表 1 所列。

表 1 武器领域数据集中关系重叠的句子分布情况

Table 1 Distribution of sentences with overlapping relation on weapon field dataset

Overlapping types	Number of sentences	
	train	test
Normal	737	85
EPO	38	6
SEO	2447	312
Total sentences	3200	400

其中少数句子既含有主客实体均相同的三元组对,又含有主实体或客实体其一相同的三元组对,如表 2 所列。对于这种句子,在统计时既把它算作 EPO 类型,又把它算作 SEO 类型。

表 2 三元组重叠特殊样例
Table 2 Special example of triple overlap

Text data	Overlapping types	Entity relation triplet
Pz68 主战坦克是瑞士把 Pz61 主战坦克进一步发展的改良型,由联邦制造厂生产	EPO	Pz68 主战坦克,产国,瑞士
		Pz68 主战坦克,使用国,瑞士
	SEO	Pz68 主战坦克,使用国,瑞士
		Pz68 主战坦克,前型,Pz61 主战坦克
		Pz68 主战坦克,生产商,联邦制造厂
		Pz68 主战坦克,产国,瑞士

4.1.2 公开数据集

本文选取由远程监督方式生成的 NYT(NewYork Times)^[27]公开数据集,采用 Zeng^[1]发布的预处理数据,共包含 24 种预定义关系、56 195 条训练数据、5 000 条验证数据以及 5 000 条测试数据。在训练集和测试集中,句内无实体重叠(Normal)的句子共有 40 279 条,句内共享一个实体对(EPO)的句子共有 10 760 条,句内仅共享一个实体(SEO)的句子共有 16 032 条。具体数量如表 3 所列。

表 3 NYT 数据集中关系重叠的句子分布情况
Table 3 Distribution of sentences with overlapping relation on NYT dataset

Overlapping types	Number of sentences	
	train	test
Normal	37 013	3 266
EPO	9 782	978
SEO	14 735	1 297
Total sentences	56 195	5 000

4.2 评价指标

本文采用目前广泛使用的指标来评价模型性能,分别是 Precision(精确率)、Recall(召回率)和 F1 值,如式(12) — 式(14)所示。

$$Precision = \frac{Relation_{TP}}{Relation_{TP} + Relation_{FP}} \times 100\% \tag{12}$$

$$Recall = \frac{Relation_{TP}}{Relation_{TP} + Relation_{FN}} \times 100\% \tag{13}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \times 100\% \tag{14}$$

其中,Relation_{TP}表示模型正确抽取的关系三元组数量,即模型预测的三元组与样本中的真实三元组相同;Relation_{FP}则相反,表示模型错误抽取的关系三元组数量;Relation_{TP} + Relation_{FN}表示样本的关系三元组数量。

4.3 实验分析

4.3.1 武器领域数据集

为验证基于 DCNN 和 GLU 实体关系抽取模型(DCNN_GLU_Joint)在武器领域的可行性,将其与 3 个早期经典实体关系联合抽取模型进行比较,分别是 Zeng^[1]提出的 CopyRE 模型、Fu^[28]提出的 GraphRel 模型、Wei 等^[2]提出的 CasRel 模型。在武器领域数据集的对比实验结果如表 4 所列。

表 4 不同模型在武器领域数据集上的实验结果

Table 4 Experimental results of different models on weapon field data sets

(单位:%)

Model	Precision	Recall	F1
CopyRE(MultiDecoder)	40.2	23.4	29.6
GraphRel(2P)	43.5	32.1	36.9
CASREL	81.9	76.2	78.9
DCNN_GLU_Joint	82.7	79.5	81.1

从表 4 可以看出,在武器领域数据集上,与早期经典实体关系联合抽取模型 CopyRE(MultiDecoder)和 GraphRel(2P)、CASREL 相比,本文提出的 DCNN_GLU_Joint 模型在 F1 值上均有一定提升,达到了 81.1%。这是因为武器领域数据十分特殊,包含了大量 SEO 类型的重叠关系,每个模型处理该类型数据的能力各不相同。CopyRE(MultiDecoder)模型虽然考虑到了句子中的实体重叠现象,但是在使用多解码器生成实体关系三元组时未考虑实体与关系间的关联,因此丢失了大量重叠关系信息。GraphRel 尽管考虑了该情况,利用基于图卷积网络的结构将文本数据转化为关系图,增强了命名实体识别与关系抽取任务之间的联系,但由于把关系看作离散的标签,使得训练过程中产生了较多负面示例,导致抽取出的实体关系三元组并不完整。本文提出的 DCNN_GLU_Joint 模型则采用序列分层标注的方法把主实体与客实体进行相互映射,帮助模型识别出每个实体的全部关系,增强了主实体与客实体之间的依赖程度,因此召回率较上述两个基线模型有了大幅提升。与 CASREL 相比,本文模型引入了自注意力机制,该机制能够降低外部信息对模型的影响,使其可以提取丰富的数据内部特征,更加适合小样本数据集的实体关系抽取任务,故 F1 略有提升。

为了验证 DCNN_GLU_Joint 模型具有解决武器领域关系重叠问题的能力,本文统计了该模型在不同重叠类型的武器领域数据上的实验结果,如表 5 所列。

表 5 模型在不同重叠类型的武器领域数据上的实验结果
Table 5 Experimental results of model in different overlapping types of weapon field dataset

(单位:%)

Overlapping types	Precision	Recall	F1
Normal	77.4	80.2	78.8
EPO	85.8	77.9	81.7
SEO	82.1	79.8	80.9

由表 5 可见,本文模型在无重叠、实体对重叠、单实体重叠这 3 种重叠类型下的 F1 值均高于 78%。其中,EPO 重叠类型的实体关系三元组抽取效果最好,F1 值达 81.7%;Normal 重叠类型的实体关系三元组抽取效果最差,F1 值为 78.8%,但整体上 3 种重叠类型的 F1 值相差不大。由此可以推断,本文模型能够处理武器领域关系数据集中不同重叠类型的句子,具有解决非结构化文本数据内关系重叠问题的能力。

4.3.2 公开数据集

为验证 DCNN_GLU_Joint 模型的泛化能力,本文在公开数据集 NYT 上进行实验并将其与 3 个基线模型进行对比,实验结果如表 6 所列。

表 6 不同模型在 NYT 数据集上的对比实验结果
Table 6 Experimental results of different models on NYT dataset
(单位: %)

Model	Precision	Recall	F1
CopyRE(MultiDecoder) ^[1]	61.0	56.6	58.7
GraphRel(2P) ^[28]	63.9	60.0	61.9
CASREL ^[2]	89.7	89.5	89.6
DCNN_GLU_Joint	86.2	85.7	85.9

根据表 4 和表 6 可知,与武器领域数据集相比,各模型在 NYT 数据集上的 F1 值均有所提升,其中 CopyRE(MultiDecoder)和 GraphRel(2P)模型的提升效果显著;CASREL 和本文模型也有提升,但增长幅度并不明显。这是因为在 NYT 训练数据集中,具有 Normal 重叠关系的句子占比最高,达 65.87%;在武器领域训练数据集中,具有 SEO 重叠关系的句子占比最高,达 76.47%。由于模型结构特殊性,CopyRE(MultiDecoder)和 GraphRel(2P)更加适用于处理普通“一对一”关系的句子,故在 NYT 上的表现更好。虽然 CASREL 和本文模型的 F1 值也有所提升,但其本身更加适用于处理“一对多”或“多对多”关系的句子,因此在 SEO 重叠关系占比高的武器领域关系数据集上依然有较好的抽取效果,在 NYT 上的提升效果并不明显。观察表 6 可知,在 NYT 数据集上,CASREL 模型的抽取效果优于本文模型。这是因为两者使用的预训练语言模型不同,CASREL 采用 BERT 结构,而本文模型使用 Word2Vec 工具。在大样本数据集中,前者可以动态地表示词向量,获取更为丰富的表达特征;而后者发现新词的能力较差,故不能很好地适应数据量庞大的 NYT 公开数据集。

结束语 本文基于膨胀卷积神经网络和门控线性单元提出了一种实体关系抽取模型,以分层次的序列标注方式识别命名实体及其之间的关系。实验结果显示该模型在武器领域具有良好的抽取效果,一定程度上能够解决该领域数据的关系重叠问题。

虽然模型初见成效,但在实验过程中也发现了若干可改进之处,在接下来的工作中会以此为研究重点继续努力。由于 Word2Vec 只能静态地训练词向量,不能充分地学习特征信息,因此后续可将其替换为动态训练词向量的 BERT 模型;由于模型需要抽取预定义的所有关系且一次只能处理一个主实体,训练时间较长,因此后续可引入军事领域三元组知识库,实现快速的远程监督式搜索。

参 考 文 献

[1] ZENG X,ZENG D,HE S,et al. Extracting Relational Facts by an End-to-End Neural Model with Copy Mechanism[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Melbourne, Australia: Association for Computational Linguistics, 2018:506-514.

[2] WEI Z,SU J,WANG Y,et al. A Novel Cascade Binary Tagging Framework for Relational Triple Extraction[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2020: 1476-1488.

[3] WANG Y,YU B,ZHANG Y,et al. TPLinker:Single-stage Joint Extraction of Entities and Relations Through Token Pair Link-

ing[J]. arXiv:2010.13415,2020.

[4] XU J,ZHANG Z X,WU Z X. Review on Techniques of Entity Relation Extraction[J]. New Technology of Library and Information Service,2008,24(8):18-23.

[5] CHEN H W,ZHANG X Y,WANG T. Research on Named Entity Recognition[J]. Computer Science,2005,32(4):44-48.

[6] TRAN K M,BISK Y,VASWANI A,et al. Unsupervised Neural Hidden Markov Models[C]//Proceedings of the Workshop on Structured Prediction for NLP. 2016:63-71.

[7] MA X,HOFVY E. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Berlin, Association for Computational Linguistics, 2016:1064-1074.

[8] LI X,SUN X,MENG Y,et al. Dice Loss for Data-imbalanced NLP Tasks[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics,2020:465-476.

[9] ZHONG Z,CHEN D. A Frustratingly Easy Approach for Entity and Relation Extraction[C]//Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics;Human Language Technologies. Association for Computational Linguistics,2021:50-61.

[10] LI X,FENG J,MENG Y,et al. A Unified MRC Framework for Named Entity Recognition[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics,2020:5849-5859.

[11] WANG X,JIANG Y,BACH N,et al. Automated Concatenation of Embeddings for Structured Prediction[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing(Long Papers). Association for Computational Linguistics,2021:2643-2660.

[12] ZHOU W,CHEN M. Learning from Noisy Labels for Entity-Centric Information Extraction[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021: 5381-5392.

[13] SCHWETER S,AKBIK A. FLERT: Document-Level Features for Named Entity Recognition[J]. arXiv:2011.06993,2020.

[14] YE D,LIN Y,SUN M. Pack Together:Entity and Relation Extraction with Levitated Marker[J]. arXiv:2109.06067,2021.

[15] E H H,ZHANG W J,XIAO S Q,et al. Survey of Entity Relationship Extraction Based on Deep Learning[J]. Journal of Software,2019,30(6):1793-1818.

[16] SANTOS C N,XIANG B,ZHOU B. Classifying Relations by Ranking with Convolutional Neural Networks[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Long Papers). Beijing: Association for Computational Linguistics,2015:626-634.

[17] WANG L,CAO Z,MELO G,et al. Relation Classification via Multi-Level Attention CNNs[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Berlin: Association for Computational Linguistics,2016:1298-1307.

[18] WANG H,TAN M,YU M,et al. Extracting Multiple-Relations in One-Pass with Pre-Trained Transformers[C] // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence; Association for Computational Linguistics,2019;1371-1377.

[19] COHEN A D,ROSENMAN S,GOLDBERG Y. Relation Classification as Two-way Span-Prediction[J]. arXiv: 2010. 04829, 2020.

[20] MIWA M,BANSAL M. End-to-End Relation Extraction using LSTMs on Sequences and Tree Structures[C] // Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Berlin; Association for Computational Linguistics,2016;1105-1116.

[21] ZHENG S,WANG F,BAO H,et al. Joint Extraction of Entities and Relations Based on a Novel Tagging Scheme[C] // Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vancouver; Association for Computational Linguistics,2017;1227-1236.

[22] BEKOULIS G,DELEU J,DEMEESTER T,et al. Joint entity recognition and relation extraction as a multi-head selection problem[J]. Expert Systems with Applications, 2018, 114: 34-45.

[23] SUI D,CHEN Y,LIU K,et al. Joint Entity and Relation Extraction with Set Prediction Networks[J]. arXiv:2011. 01675,2020.

[24] HUGUET CABOT P L,NAVIGLI R,REBEL;Relation Extraction By End-to-end Language generation[C] // Findings of the Association for Computational Linguistics;EMNLP 2021. Punta Cana; Association for Computational Linguistics, 2021: 2370-2381.

[25] MIKOLOV T,CHEN K,CORRADO G,et al. Efficient Estima-

tion of Word Representations in Vector Space[J]. arXiv:1301. 3781,2013.

[26] VASWANI A,SHAZEER N,PARMAR N,et al. Attention is All You Need[C] // Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017; 6000-6010.

[27] RIEDEL S, YAO L, MCCALLUM A. Modeling Relations and Their Mentions without Labeled Text[C] // Machine Learning and Knowledge Discovery in Databases. Berlin; Springer, 2010: 148-163.

[28] FU T J,LI P H,MA W Y. GraphRel: Modeling Text as Relational Graphs for Joint Entity and Relation Extraction[C] // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence; Association for Computational Linguistics,2019;1409-1418.

LI Han, born in 1997, postgraduate. Her main research interests include knowledge graph and relation extraction, and so on.

LIU Xiulei, born in 1981, Ph.D, professor, is a member of China Computer Federation. His main research interests include ontology matching, semantic sensor, knowledge graph, semantic Web and semantic search, and so on.

220200112-7