北京大学学报(自然科学版) 第 58 卷 第 3 期 2022 年 5 月
Acta Scientiarum Naturalium Universitatis Pekinensis, Vol. 58, No. 3 (May 2022)
doi: 10.13209/j.0479-8023.2021.118

面向武器装备领域的复杂实体识别

游新冬[1] 葛昊杰[1] 韩君妹[2] 李育贤[1] 吕学强[1,†]

1. 北京信息科技大学网络文化与数字传播北京市重点实验室, 北京 100101; 2. 军事科学院系统工程研究院复杂系统仿真总体重点实验室, 北京 100101; † 通信作者, E-mail: lxq@bistu.edu.cn

摘要 针对武器装备领域复杂实体的特点, 提出一种融合多特征后挂载武器装备领域知识的复杂命名实体识别方法。首先, 使用BERT模型对武器装备领域数据进行预训练, 得到数据向量, 使用Word2Vec模型学习郑码、五笔、拼音和笔画的上下位特征, 获取特征向量。然后, 将数据向量与特征向量融合, 利用Bi-LSTM 模型进行编码, 使用CRF解码得到标签序列。最后, 基于武器装备领域知识, 对标签序列进行复杂实体的触发检测, 完成复杂命名实体识别。使用环球军事网数据作为语料进行实验, 分析不同的特征组合、不同神经网络模型下的识别效果, 并提出适用于评价复杂命名实体识别结果的计算方法。实验结果表明, 提出的挂载领域知识且融合多特征的武器装备复杂命名实体识别方法的F1值达到95.37%, 优于现有方法。

关键词 武器装备; 复杂命名实体识别; 郑码; 领域规则; BERT; 评价方法

Recognition of Complex Entities in Weapons and Equipment Field

YOU Xindong[1], GE Haojie[1], HAN Junmei[2] , LI Yuxian[1], LÜ Xueqiang[1,†]

1. Beijing Key Laboratory of Internet Culture and Digital Dissemination Research, Beijing Information Science and Technology University, Beijing 100101; 2. National Key Laboratory for Complex Systems Simulation, Institute of Systems Engineering, Beijing 100101; † Corresponding author, E-mail: lxq@bistu.edu.cn

Abstract Aiming at the characteristics of complex entities in weapons and equipment field, a complex named entity recognition method is proposed which integrates multi-features and mounts the domain knowledge of weapons and equipment. First, we use the BERT model to pre-train on the weapon equipment field data to obtain the data vector, and use the Word2Vec model to learn context features of Zhengma, Wubi, Pinyin, and strokes to obtain the feature vector. Then the data vector and the feature vector are fused, the Bi-LSTM model is used for encoding, and the CRF decoder is used to obtain the tag sequence. Finally, the detection of complex entities on the label sequence is triggered to complete the recognition of complex named entities. In the experiments, we use the data collected from Global Military Network as the corpus, and analyze the recognition effect of different feature combinations and neural network models. A calculation method suitable for evaluating the recognition results of complex named entities is also proposed. The experimental results show that the F1-value of the proposed method for recognizing complex named entities of weapons and equipment with domain knowledge and fusion of multi-features reaches 95.37%, which outperforms the existing methods.

Key words weapon and equipment; complex named entity recognition; Zhengma; domain rules; BERT; evaluation method

互联网上存储着包含大量武器装备领域知识的文本, 按照数据的组织性质, 可以将其划分为结构化、半结构化和非结构化数据。其中, 占比最高的非结构化文本数据中包含大量的武器装备信息, 可以从中获取大量的武器装备信息, 用于构建武器装备知识图谱。通过命名实体识别任务可以识别其中

北京市自然科学基金(4212020)、国家自然科学基金(62171043)、国防科技重点实验室基金(6412006200404)、北京信息科技大学“勤信人才”培育计划项目(QXTCP B201908)和北京市市教委科研计划(KM202111232001)资助

收稿日期: 2021–09–01; 修回日期: 2021–11–01

包含的实体, 为知识图谱构建提供知识支撑。根据实体的边界特征以及分类特征, 可以划分为普通命名实体和复杂命名实体。复杂命名实体(complex named entity)包含嵌套命名实体(nested named entity)和非连续性命名实体(discontinuous named entity)。嵌套命名实体指在一个实体内部完全包含着另一个相同类别或不同类别的实体, 非连续实体指若干实体边界存在交叉、不完全重叠的实体。实体类别的定义如表 1 所示。

以非结构化文本“纳希莫夫号有 sa-n-6、sa-n-4, sa-n-9 三种导弹和一种 cads-1 弹炮合一系统, 共有四个层次的对空火力。”为例, 嵌套命名实体、非连续命名实体和普通命名实体的标识如图 1~3 所示。

图 1 的例句中包含的普通命名实体为 “纳希莫夫号###舰船”, 嵌套命名实体为“纳希莫夫###人名”, 两者存在完全交叉的边界“夫”, 其中“纳希莫夫”序列片段隶属“person”和“ship” 两个类别。

表 1 实体类别的定义说明

Table 1 Definition table of entity category

实体类别	从边界的角度	从分类的角度
普通命名实体	边界无交叉	标签分类归属唯一
嵌套命名实体	边界完全交叉	标签分类归属不唯一
非连续命名实体	边界不完全交叉	标签分类可能不唯一

图 2 的例句中包含的非连续命名实体为“sa-n-6 导弹###导弹, sa-n-4 导弹###导弹, sa-n-9 导弹###导弹”, 三者的边界并不完全, 中间掺杂“、”和“,”等符号, 却又共享一部分序列片段“导弹”; 标签分类“sa-n-6”的标签是唯一的, 但共享片段“导弹”的标签却不是唯一的, 同时隶属“O”和“missile”。

图 3 的例句中包含普通命名实体“cads-1 弹炮

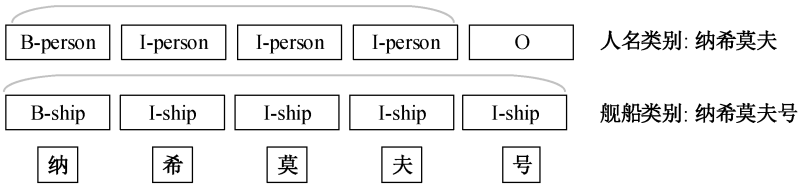


图 1 嵌套命名实体示意图

Fig. 1 Schematic diagram of nested named entities

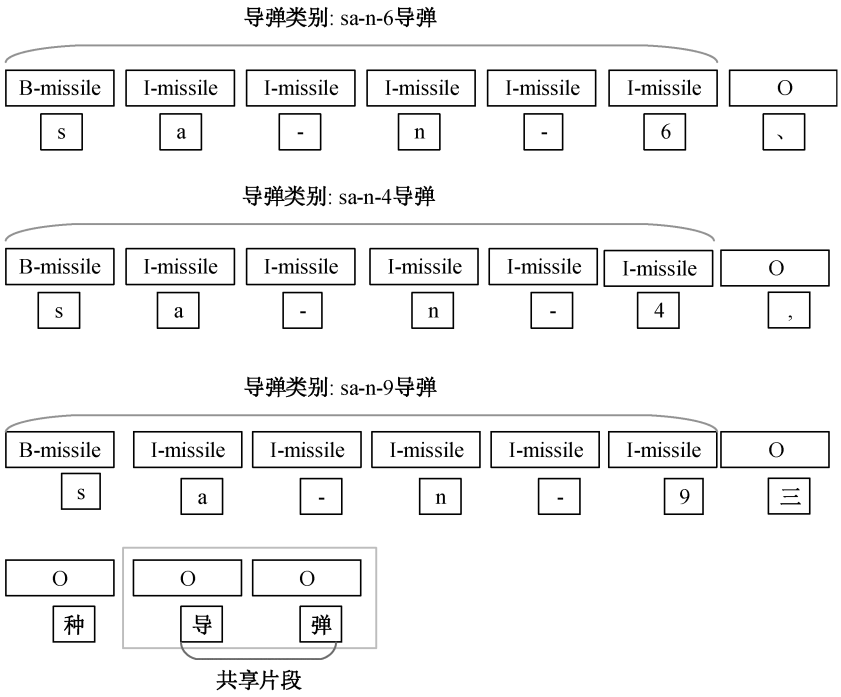


图 2 非连续命名实体示意图

Fig. 2 Schematic diagram of discontinuous named entities

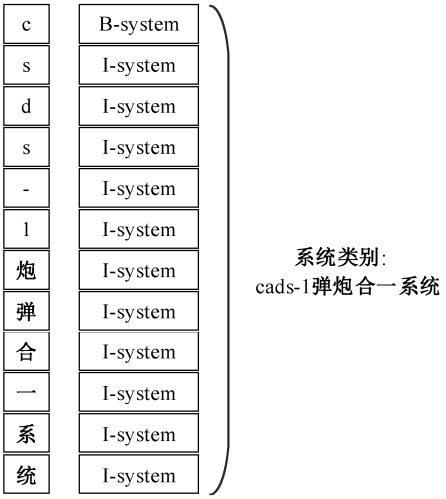


图 3 普通命名实体示意图

Fig. 3　Schematic diagram of flatted named entities

合一系统###系统”, 每个字符的标签是唯一的, 并且实体边界不存在任何交叉的情况。

以上分析说明, 复杂命名实体识别任务可以补充在知识图谱构建过程中普通命名实体识别任务无法抽取的结果, 完成对包含在非结构化文本中的实体细粒度抽取, 从而提高从非结构化文本中抽取出来的知识的质量[1]。因此, 面向武器装备领域的复杂命名实体识别效果的提升可以提高武器装备领域知识图谱的构建效果。

1 相关工作

实体抽取(entity extraction)是命名实体识别与分类的别称[2], 也称为命名实体识别(named entity recognition, NER), 目的是从非结构化文本数据中抽取出由领域限定的命名实体。依据命名实体的类型, 可以将命名实体识别划分为普通命名实体识别(flat named entity recognition)和复杂命名实体识别(complex named entity recognition)。复杂命名实体识别主要解决的是嵌套命名实体(nested NE)和非连续性实体(discontinuous NE)的识别问题。

普通命名实体识别的方法主要经历了 3 个发展阶段。第一个阶段是针对特定领域的命名识别, 关注领域内的固定表达, 即根据数据特点制定领域规则。Collins 等[3]提出 DL-CoTrain 方法, 首先人工构建决策列表(种子规则集), 之后对数据集进行无监督学习, 构建更多的规则集, 最后将规则集进行分类。此方法对人名、地名和机构名的识别和分类精确率超过 91%, 其整体的思想是利用自提升思想, 不断地构建规则集。识别效果和人工构建规则集的质量和数量呈现正相关的关系, 因此此类方法对新数据的变化十分敏感, 可迁移性较差。

随着统计机器学习的发展, 普通命名实体识别进入第二阶段。在该阶段, 命名实体识别被转化为序列标注, 其本质还是分类的问题。在此阶段, 根据识别类型又可以划分为两大派系。一个派系是分阶段识别, 以 Collins 等[3]的 CoBoost 方法为代表, 先识别边界, 后进行实体分类, 其思路是设两个分类器来识别边界, 再设一个分类器对实体类型进行分类。另一个派系以 Petasis 等[4]为代表, 使用隐马尔科夫模型等典型的序列标注方法, 同时进行边界识别和分类。对英文而言, 词是最小的标注单元; 对中文而言, 字是最小的单元。此方法避免了分词带来的错误积累问题。该阶段的方法将重心转移到概率上, 因此拥有一定的迁移能力, 但需要人工选择特征, 对相关人员的特征选择能力也有一定的要求。

随着深度学习技术的发展, 普通命名实体识别进入以深层神经网络(deep neural network, DNN)为基础的第三阶段, 在该阶段, 特征工程和领域知识几乎可以不介入[5]。2019 年, 王子牛等[6]将 BERT 预训练模型加入 Bi-LSTM+CRF 模型, 将位置信息和语义信息加入模型的学习中, 在命名实体识别中达到 94%的 F1 值。本阶段的关注点转移到文本的整体信息上, 无需专家构建规则。

复杂命名实体识别中较为有效的方法是转换任务或堆叠模型。Finkel 等[7]在嵌套实体识别任务中构建解析树, 将每个实体都作为解析树的一部分, 实现针对嵌套命名实体识别的基于 CRF 的解析器。Lu 等[8]将离散数学的超图思想引入嵌套命名实体识别, 允许一条边链接多个顶点, 以此表示嵌套实体。Ju 等[9]通过动态地堆叠普通命名实体识别模块, 从由内而外的角度对嵌套命名实体进行识别, 但是此方法面临错误传播的问题。Xia 等[10]提出 MGNER 架构, 对文本中复杂实体和普通实体都有很好的效果。此方法分为两个阶段, 首先识别实体, 之后再对实体进行分类, 这与序列标注的思想是不同的。Li 等[11]将命名实体识别任务转化为机器阅读理解任务, 从一个全新的视角进行复杂命名实体的抽取, 通过挂载询问语句外部知识库, 对复杂命名实体和普通命名实体达到 95.75%的 F1 值。

普通命名实体识别和复杂命名实体识别的研究

起步较早，并且有一个逐渐深入的过程。然而，面向武器装备领域的命名实体识别因数据具有很强的领域特点、独特的研究背景和不规则的文本表达形式，其研究起步较晚。

针对武器装备领域的普通命名实体识别，也是从规则的方法逐步发展到深度学习的方法。姜文志等[12]利用CRF模型和领域规则相结合的方式，外加非结构化文本的局部特征和外部知识库，再使用规则对识别结果进行优化，完成武器装装备的普通命名实体识别。冯蕴天等[13]使用CRF模型的半监督武器装备命名实体识别方法，并使用词典和规则对识别的结果进行校准。朱佳晖等[14]融合同义词集、字符相似性以及包含相似性等多方面相似性度量特征，完成武器装备实体的识别。尹学振等[15]采用BERT+LSTM+CRF模型，完成武器装备命名实体识别。

针对武器装备领域复杂命名实体识别的研究较少，并聚焦在嵌套实体的识别上。姜文志等[16]采用CRF模型和SVM模型并行抽取之后再合并的方法，完成武器装备的嵌套命名实体识别。单赫源等[17]提出小粒度的命名实体识别方法，按照粒度从小到大的顺序，对武器装备实体进行识别。单义栋等[18]通过对复合的武器装备领域实体做分词处理，再对最小词组做分段标注之后做词位标注完成武器装备的嵌套命名实体识别。

综上所述，在武器装备领域的复杂命名实体识别中，目前面临的问题有分词的错误传播、分词的颗粒度大小不统一以及领域特色挖掘深度不足等。另一方面，学者们往往忽略非连续命名实体，而实体表达形式的非连续性。

本文为提高武器装备领域文本的嵌入表示质量，引入预训练模型，在武器装备领域文本上进行训练；使用领域规则触发器，对边界进行特别处理，以期提高嵌套实体的边界关注度；面对非连续实体，采用武器装备领域规则词性表和领域规则映射表，完成数据规范化和完整性，提升武器装备领域复杂实体识别的精确率和召回率，有效地扩充武器装备的领域知识。本文采用基于BERT的方法获取字级别的特征，解决分词的错误传播问题以及分词颗粒度大小不统一的问题。通过双向长短时记忆神经网络(Bi-LSTM)抽取上下文特征，形成特征矩阵，再经由条件随机场(CRF)解码获得最优标签。最后，根据领域规则触发器对以上阶段的识别结果进行扩充，解决领域特色挖掘深度不足的问题。最终，实现武器装备领域嵌套命名实体和非连续命名实体的识别精确率和召回率的提高。

2 武器装备复杂命名实体识别模型

为避免中文分词的错误传播，目前的中文命名实体识别任务中，最小处理单元(token)普遍是字符。这就导致嵌套命名实体和非连续命名实体的实体内部边界无法得到足够的关注。因此，需要在识别后进行触发检测，提高模型对实体内部边界的关注度。例如，对“纳希莫夫号有sa-n-6、sa-n-4，sa-n-9三种导弹和一种cads-1弹炮合一系统，共有四个层次的对空火力。”使用普通的字符粒度的标注得到的结果如图4所示。

从图4可以发现，标注序列的内部边界和非实体内部边界的token得到同样的标记结果。为了解决字符粒度对实体边界关注不足的问题，本文使用融合多特征的深度学习模型和武器装备领域规则的方法，完成武器装备的复杂命名实体识别，其中武器装备领域规则负责对实体边界进行显式“关注”。本研究的武器装备复杂命名实体识别框架如图5所示。

从图5可以看出，该框架在BERT+Bi-LSTM+CRF多神经网络的协同下，融入郑码、五笔、拼音和笔画特征，在武器装备领域规则触发器的指导下，实现武器装备领域的复杂命名实体识别，即同时解决武器装备领域非结构化文本中包含的嵌套命名实体和非连续命名实体问题。

纳/B-missile 希/I-missile 莫/I-missile 夫/I-missile 号/I-missile 有/O s/I-missile a/I-missile
-/I-missile n/I-missile -6/I-missile 、/O s/B-missile a/I-missile -/I-missile n/I-missile
-/I-missile 4/I-missile ,/O s/B-missile a/I-missile -/I-missile n/I-missile -/I-missile
9/I-missile 三/O 种/O 导/O 弹/O 。/O

图 4 武器装备领域数据集的标注样例

Fig. 4 Annotated sample diagram of data sets in the field of weapons and equipment

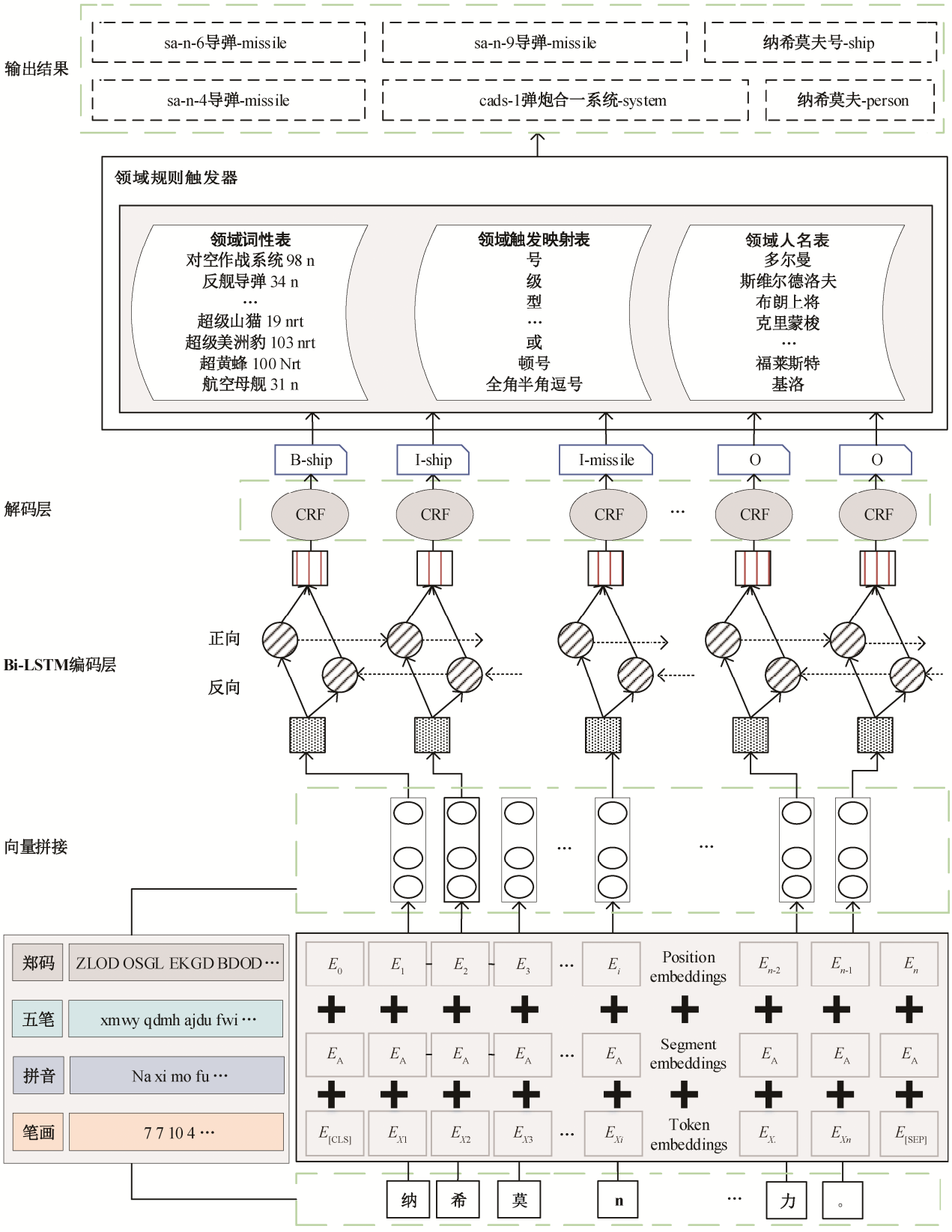


图 5　武器装备复杂命名实体识别模型

Fig. 5　Recognition model diagram of complex named entity of weapons and equipment

本框架分为两个识别阶段。1) 首先在大规模武器装装备领域文本上预训练BERT模型, 得到字符的分布式嵌入表示。之后按照字符与特征(即五笔、郑码、拼音和笔画)映射表, 完成文本的表示映射。利用Word2Vec模型进行上下文的学习, 获取各个特征的分布式嵌入表示。然后在BERT生成特征向量的基础上, 拼接不同的特征组合, 投入Bi-LSTM网络, 利用Bi-LSTM的门控机制, 进一步对现有信息进行编码, 获取特征向量。最后使用条件随机场(CRF), 对特征向量进行解码, 获取第一阶段的识别结果。2) 依据第一阶段的识别结果, 进行武器装备领域的复杂命名实体触发检测。根据领域触发映射表, 选择进入嵌套命名实体或非连续命名实体的领域规则进行筛查。最后, 通过领域规则触发器模块, 获取最终的输出, 同时抽取出非结构化文本中包含的嵌套命名实体和非连续命名实体, 实现面向武器装备领域的嵌套命名实体和非连续命名实体的识别。

2.1 嵌入层模块

2018年, Google开发出基于Transformer的双向编码器(bidirectional encoder representations from transformers, BERT)的预训练语言模型[19]。该模型由12或24个采用自注意力机制的Transformer Block编码器块堆叠而成。BERT在非武器装备领域的命名实体识别任务中使用广泛, 可以挖掘出通用领域非结构化文本的深层次语义信息。因此, 本文将BERT应用在武器装备复杂命名实体识别任务中, 以便提高武器装备领域数据的嵌入表示质量。Google于2013年开源的Word2Vec是文本向量化的经典模型, 该模型已被证明在通用领域上拥有良好的文本向量化效果。本文通过Word2Vec模型, 对不同的特征(郑码、五笔、拼音和笔画)进行上下文学习, 获取稠密向量作为特征, 与文本的嵌入表示进行拼接。

2.1.1 输入序列向量化

在该模块中, 将非结构化的文本视为输入序列 $X=(x_0,\ x_1,\ x_2,\ \ldots,\ x_{n\text{-}1})$, 其中 n 是序列的字符总长度。对输入序列添加用于分类的特殊字符“[CLS]” (classification) 以及用于间隔文本的特殊字符“[SEP]” (separation), 然后使用基于Transformer的双向编码器进行编码。输入序列的向量化过程如下:

$$\boldsymbol{v}_i^{\text{bert}}=\text{BERT}(x_i),\ i\in Z\bigcap i\in[0,n-1], \tag{1}$$

其中, $\boldsymbol{v}_i^{\text{bert}}$ 表示输入序列 X 的第 i 个字符对应的通过BERT模型编码得到的字符向量。BERT表示BERT预训练模型, x_i 表示输入序列的第 i 个字符, i 是属于0到 n–1的整数。

2.1.2 郑码特征

根据中文汉字的发展历史, 可以将其归结为语素文字或表意文字或象形文字[20]。我国著名文学家郑易里创造了郑码(又称字根通用码)。Windows 95/98/NT/2000/XP/Vista/7中文操作系统都选用郑码作为内置编码。郑码有助于挖掘非结构化文本中潜在的语义关系, 可使神经网络通过不同汉字的字型结构组成, 学习到武器装备领域的命名内部和外部实体边界信息。“舰、艇、船”这些语义与结构都相似的字符被划分为“ship”的概率比较大, 并且相较于其他字符, 这些字符充当武器装备领域实体右边界的概率也较大。我们利用郑码与汉字的映射表(http://zmsjit.com), 将武器装备领域的非结构化数据进行转换, 之后使用Word2Vec模型按句进行训练, 获取每个字符的上下文特征, 并提供字符的向量映射。郑码的向量化过程如下:

$$p=f_{\textbf{zhengma}}(X), \tag{2}$$

$$\boldsymbol{v}_i^{\text{zhengma}}=\boldsymbol{e}^{\text{zhengma}}(p_i),\ i\in Z\bigcap i\in[0,n-1], \tag{3}$$

其中, f_{zhengma} 表示将输入的字符序列映射为郑码序列的函数。之后, 按照字符在 X 中序号 i 在郑码向量中查找 x_i 对应的郑码向量。$\boldsymbol{v}_i^{\text{zhengma}}$ 表示与输入序列 x_i 对应的郑码向量。$\boldsymbol{e}$ 表示向量查找表。

2.1.3 五笔特征

作为形码的典型, 五笔根据笔画和字型对汉字进行编码, 使用一个汉字最多四级编码字母的标识。五笔无法避免表达与拼音重复, 例如“亦”的五笔是“you”, 而“you”可以作为汉字的拼音。我们使用五笔特征的目的是与郑码特征互相校正。本文利用与训练郑码特征类似的方法完成五笔特征获取。五笔特征的向量化过程如下:

$$p=f_{\text{wubi}}(X), \tag{4}$$

$$\boldsymbol{v}_i^{\text{wubi}}=\boldsymbol{e}^{\text{wubi}}(p_i),\ i\in Z\bigcap i\in[0,n-1], \tag{5}$$

其中, $\boldsymbol{f}_{\text{wubi}}$ 表示将输入的字符序列映射为五笔序列的函数。之后, 按照字符在 X 中序号 i 在五笔向量中查找 X_i 对应的五笔向量。$\boldsymbol{v}_i^{\text{wubi}}$ 标识与输入序列 x_i 对应的五笔向量。

2.1.4 拼音特征

郑码和五笔都是基于汉字的象形结构，而读音的变化对汉字的语义表达也有不可忽视的作用。本文利用Pinyin工具包构建输入序列与特征序列的映射关系，之后利用Word2vec模型对武器装备领域数据进行训练，完成拼音特征的向量化。拼音向量化过程如下:

$$p = f_{\text{Pinyin}}(X), \tag{6}$$

$$\boldsymbol{v}_i^{\text{pinyin}} = \boldsymbol{e}^{\text{pinyin}}(p_i),\ i \in Z \bigcap i \in [0, n-1], \tag{7}$$

其中，f_{Pinyin}表示将输入的X字符序列映射为拼音特征序列的函数，$\boldsymbol{e}^{\text{pinyin}}$表示拼音和输入序列的映射表，$\boldsymbol{v}_i^{\text{pinyin}}$表示与输入序列$X$中第$i$个输入字符对应的拼音向量。

2.1.5 笔画特征

在武器装备领域数据中，笔画复杂度相似的汉字更有可能归于同一个类别标签。本文参考BERT的位置嵌入思想，增加输入序列字符对应的笔画数，采取Word2Vec模型训练笔画的上下文特征，获取文本的笔画特征。笔画特征向量化过程如下:

$$p = f_{\text{bihua}}(X), \tag{8}$$

$$\boldsymbol{v}_i^{\text{bihua}} = \boldsymbol{e}^{\text{bihua}}(p_i),\ i \in Z \bigcap i \in [0, n-1], \tag{9}$$

其中，f_{bihua}表示将输入的X字符序列映射为笔画特征序列的函数。$\boldsymbol{e}^{\text{bihua}}$表示笔画与输入序列的映射表，$\boldsymbol{v}_i^{\text{bihua}}$表示与输入序列$X$中第$i$个输入字符对应的笔画向量。

将获取的4类特征进行拼接处理，得到嵌入层的最终向量。特征的处理如下:

$$\boldsymbol{v}_i^{w} = \text{Concat}\left(\boldsymbol{v}_i^{\text{bert}}, \boldsymbol{v}_i^{\text{zhengma}}, \boldsymbol{v}_i^{\text{wubi}}, \boldsymbol{v}_i^{\text{pinyin}}, \boldsymbol{v}_i^{\text{bihua}} \cdot k\right), \tag{10}$$

$\boldsymbol{v}_i^{w}$代表第i个字符对应的融合后向量表示。设定k的目的是降低笔画数目分布不均衡对整体向量的影响，经过初期的数据分析，k取值为0.5。

2.2 编码层模块

Hochreiter等[21]提出的长短期记忆网络(long-short term memory, LSTM)解决了传统RNN模型存在的梯度消失问题，并通过门控机制，有选择地遗忘某些信息，比传统的RNN模型拥有更好的信息“筛选”能力。LSTM的门控机制由输入门、遗忘门和输出门分别担负信息的记忆、遗忘和输出的任务。输入门使用的函数如下:

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i), \tag{11}$$

$$\widetilde{C}_i = \tanh(W_C\left[h_{t-1}, x_t\right] + b_C), \tag{12}$$

$$C_t = f_t * C_{t-1} + i_t * \widetilde{C}_t。 \tag{13}$$

遗忘门使用的函数为

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f)。 \tag{14}$$

输出门使用的函数为

$$O_t = \sigma(W_o[h_{t-1}, x_t] + b_o), \tag{15}$$

$$h_t = O_t * \tanh(C_t)。 \tag{16}$$

σ表示激活函数sigmoid，W和b代表模型的训练参数，h_{t-1}和h_t分别表示前一个时间t的前一时刻的隐层状态和当前状态，x_t表示t时刻的输入，$\widetilde{C}_t$和C_t分别表示t时刻的细胞临时状态和细胞当前状态，f_t，i_t和O_t分别表示遗忘门、记忆门和输出门的输出。

从以上公式可以发现，LSTM模型仅注意到单侧的时间序列信息。Graves等[22]为解决LSTM神经网络的单侧注意问题，提出双向长短期记忆网络Bi-LSTM(bidirectional long-short term memory)模型。本文选用Bi-LSTM模型。如图1中编码层所示，前向和后向LSTM分别负责上下文的信息记忆，二者拼接组合，完成对上下文的信息记忆。在Bi-LSTM的输出部分，将双向序列进行拼接，获取包含双向信息的序列。序列处理如下:

$$\boldsymbol{H}_t = [\overrightarrow{\boldsymbol{h}_t}, \overleftarrow{\boldsymbol{h}_t}],\ t \in Z \bigcap t \in [0, n-1], \tag{17}$$

其中，$\boldsymbol{H}_t$代表拼接之后t时刻序列元素的上下文向量，$\overrightarrow{\boldsymbol{h}_t}$和$\overleftarrow{\boldsymbol{h}_t}$分别表示前向和后向序列中$t$时刻序列元素的上下文向量。

2.3 标签解码层

在命名实体识别任务中，经过嵌入和编码之后，需要对编码层输出的向量进行解码，以便获取与输入序列对应的标签序列。我们采用条件随机场(CRF)作为解码器，对编码层的结果进行解码。解码过程如下:

$$Y^* = \arg\max_{Y_r \in Y_x} \text{score}(X, Y_r), \tag{18}$$

其中，Y_r为真实标注数据序列，Y_x表示所有可能的标签序列。通过维特比算法得到最大分数输出，即预测标签序列Y^*。

2.4 领域规则触发器模块

根据武器装备领域的数据特点，归纳出武器装

备领域词性表(表 2)、领域规则触发表、人名和地名规则表。领域规则触发表决定武器装备领域触发器的触发条件, 规则表和词性表作为触发器的触发响应。

武器装备领域词性表用来处理非连续实体, 根据领域词性表, 将量词(如“sa-n-6、sa-n-4, sa-n-9三种导弹”中的“三种”)和动词(“阿斯洛克反潜导弹”中的“反潜”)等非实体内容进行调整、舍弃。在利用HMM模型做词性标注得到的通用领域词性基础上, 依据武器装备领域词语出现的频率设定权重以及在武器装备领域的词性。

例如, “阿斯洛克/nr 反潜/v 导弹/n”。在武器装备领域, “反潜导弹”作为专有名词的固定表达应该被划分为名词, 根据在训练集中的出现频率设定一个词频。根据上述方法构建武器装备领域的领域词性表。

领域规则触发表用来加强对命名实体边界信息的关注度。一方面, 当出现全字母(如“sa-n-9导弹”)或全数字(如“056型护卫舰”)表达的武器装备代号时, 其左侧不可能是人名和地名。基于这个数据特点, 在出现疑似实体时均使用领域规则的人名和地名规则表进行筛查, 抽取实体中包含的人名和地名实体。另一方面, 当出现例如“与、或、和、以及”等词汇时, 两侧有可能是非连续实体的内部边界。之后, 将触发词所在的片段序列进行规范化。依据上阶段的实体类别, 补充非连续的实体内容。例如图2所示的“sa-n-6、sa-n-4, sa-n-9三种导弹”标注序列, 首先对片段进行规范化, 去除“三种”之类的量词, 之后对全角的逗号“，”和半角的逗号“,”规范化为非连续实体间隔。然后依据武器装备补充规律, 将非连续的实体“sa-n-6、sa-n-4, sa-n-9”补充为“sa-n-6导弹###missile、sa-n-4 导弹###missile、sa-n-9导弹###missile”。

武器装备领域的人名和地名词典是结合各国对武器装备命名的惯例而设。例如, 日本几乎没有以人名作为舰船名的传统, 而是用山川(如“金刚号战列舰”)、河流(如“利根号巡洋舰”)和气象(如“雪风号驱逐舰”)等做舰船的名字。美国有很多舰船以政界著名人物(如美国现代海军之父约翰·斯坦尼斯)、军界著名人物(美国太平洋舰队总司令、军事家切斯特·威廉·尼米兹)的姓名来命名, 如“约翰·C·斯坦尼斯号航空母舰”和“尼米兹号航空母舰”。中国的舰船一般以行政区域命名(如“辽宁号航空母舰”)。本文依据各个国家对武器装备的命名习惯, 使用数据采集技术, 从百度百科获取相关国家的人名和地名, 作为武器装备领域的规则表。

表 2 武器装备领域词性表

Table 2 Part-of-speech table of weapons and equipment

词语类别	缩写	词语类别	缩写
副词	d	动词	v
数词	m	时间词	t
数量词	mq	量词	q
名词	n		

3 实验

3.1 实验数据

实验数据源自环球军事网关于舰船和飞机的非结构化文本数据, 最终人工标注数据3000条, 数据标注核对人员对数据进行逐一核对, 确保标注质量。按照7:2:1的比例将数据划分为训练数据、验证数据和测试数据。本实验对9种武器装备实体命名实体进行识别, 类别和标签的对应关系如表3所示。

按照国际惯例, 潜艇类和舰船类划分为并列类别。人名和地名指包含在普通武器装备命名实体之内的字符, 旨在抽取出对武器装备有重大纪念意义的人名和地名实体。

本实验使用的数据采用BIO标记方式, 样例见图4。实验数据的统计信息如表4所示。

3.2 评价指标

为了更好地评价模型的效果, 采用目前主流的基于序列的评价指标以及我们针对复杂实体识别提出的基于实体粒度的评价指标。

3.2.1 基于序列的评价指标

本文选用在命名实体识别领域广泛采用的精确率P、召回率R和F1值作为评价指标:

表 3 实体类别和标签对应表

Table 3 Correspondence table of entity category and label

实体类别	缩写	实体类别	缩写
舰船类	ship	潜艇类	submarine
飞行器类	aircraft	鱼雷类	torpedo
舰载导弹类	missile	发动机类	engine
舰炮类	shells	人名类	person
雷达类	radar	地名类	location
系统类	system	国家/地区类	country

表 4 武器装备领域数据集的统计信息

Table 4 Statistics of data sets in the field of weapons and equipment

武器装备数据集统计项	数值指标
句子条数	3000
句子平均长度	56.41
句中包含命名实体的平均个数	2.4
复杂命名实体所占句子比例	16.1%

$$P=\frac{\mathrm{TP}}{\mathrm{TP+FP}}, \tag{19}$$

$$R=\frac{\mathrm{TP}}{\mathrm{TP+FN}}, \tag{20}$$

$$\mathrm{F1}=\frac{2\times P\times R}{P+R}, \tag{21}$$

其中，TP 表示被正确识别的武器装备领域标签个数，FP 表示被错误识别的武器装备领域标签个数，FN 表示未被识别的武器装备领域标签个数。

2.2.2 基于实体粒度的评价指标

结合武器装备领域复杂实体识别的特点，为了更好地评价算法的性能，我们使用基于实体粒度的无序且可重复的实体计算 F1 数值。主要出于以下两方面的考虑。其一，如果出现嵌套实体，经过领域规则触发器的处理，将会新增实体；如果出现非连续实体，将会扩充出原本序列中未被完全包含的序列标签。这样的计算对复杂命名实体识别任务不公平的。其二，如果序列包含大量的“O”标签或包含“I-xxx”(xxx 指实体类别)为起始的实体序列，若仍然使用基于序列的命名实体评价指标，则会导致评价指标虚高。嵌套实体和非连续实体会存在实体序列重叠或断续的情况，同时标签会存在重复的情况，数据对比如表 5 所示。

基于实体粒度的复杂命名实体的精确率 P^*、召回率 R^*和 F1*值计算如下：

$$P^*=(\text{正确识别出的普通实体数量}+\text{规则触发器筛选出的复杂实体数量})/(\text{模型识别出的普通实体总数量}+\text{复杂实体数量})\times 100\%, \tag{22}$$

$$R^*=(\text{正确识别出的普通实体数量}+\text{规则触发器筛选出的复杂实体数量})/\text{测试数据中实体总数量}\times 100\%, \tag{23}$$

表 5 武器装备领域数据对比表

Table 5 Comparison of data in the field of weapons and equipment

输入序列	纳希莫夫号有 sa-n-6、sa-n-4，sa-n-9 三种导弹和一种 cads-1 弹炮合一系统，共有四个层次的对空火力。
普通实体	纳希莫夫号###ship：sa-n-6###missile、sa-n-4###missile、sa-n-9##missile
嵌套实体	纳希莫夫###person
非连续实体	sa-n-6 导弹###missile、sa-n-4 导弹###missile、sa-n-9 导弹###missile

$$\mathrm{F1}^*=\frac{2\times P^*\times R^*}{P^*+R^*}, \tag{24}$$

其中，正确识别出的实体数量指通过编码器和解码器完成的普通命名实体，复杂实体则是通过在武器装备领域触发器的指导下完成的复杂实体识别。

3.3 实验配置

1) 环境配置。本实验在戴尔服务器上运行，具体的环境配置如表 6 所示。

2) 参数配置。本文 BERT 模型通过调用 HuggingFace 库实现，经过多次实验调参①，最终确定如下参数：隐藏层的维度为 500，批处理大小(Batch_Size)为 16，BERT_embedding 维度为 768，采用的优化器为 Adam，学习率为 0.00001，衰减率为 0.00001，训练轮数为 200，神经网络单元丢弃率为 0.5。

Word2Vec 模型使用 Gensim 工具包实现训练，特征维度统一设定为 128。Bi-LSTM 隐藏层的维度设定为 500。根据对数据分布的分析，将最大长度设定为 100。

3.4 实验结果

以目前命名实体识别模型中效果最好的 BERT+Bi-LSTM+CRF 为基线模型，从两个维度验证明本

表 6　训练环境配置

Table 6 Training environment configuration

配置项	配置情况
操作系统	Linux Ubuntu 16.04
CPU	Intel (R) Xeon (R) Gold 5118 CPU @ 2.30GHz (12-Core)
GPU	8*NVIDIA Tesla V100 (16 GB)
Python	3.6.9
Pytorch	1.14.0
内存	64 G

① https://github.com/Franck-Dernoncourt/NeuroNER

文方法的有效性。纵向维度上，通过 9 个不同的特征组合证明特征融合的有效性；横向维度上，分别将基于实体粒度的未挂载武器装备领域规则触发器的评价指标与挂载武器装备领域规则触发器的评价指标进行对比，证明武器装备领域规则触发器的有效性。对比试验的特征组合如表 7 所示。

为验证各个特征的效果，设置实验 2~4 与基准实验(实验 1)进行对比。为验证不同组合特征的融合效果，设置实验 5~10 与基准实验进行对比。在 Linux 服务器上，经过训练和预测，完成对比试验。基于序列的评价指标对比试验结果如表 8 所示，基于实体粒度的对比试验和挂载武器装备领域规则触发器的对比试验结果如表 9 所示，未挂载武器装备领域规则触发器的对比试验结果如表 10 所示。

从表 8 可以看出，基准实验 BERT+Bi-LSTM+CRF 只用 BERT 提取文本特征，取得的 F1 数值为 92.06%，特征融合效果有一定的提升空间。与基准实验对比，实验 6 提升幅度最大，精确率、召回率和 F1 分别提高 2.25%，4.06%和 3.31%，说明五笔与郑码交互纠正融合的特征在武器装备领域的命名实体识别效果提升中扮演着重要角色。

从表 8 和 9 可以看出，基于实体粒度的评价指标数值明显比基于序列的评价指标低，说明在武器装备领域的复杂命名实体识别中，基于序列的评价方法导致的嵌套实体的交叉片段和非连续实体的共享片段出现无法合理计算的问题，通过本文提出的

表 7 对比试验的特征组合

Table 7 Feature combination table for comparison test

实验编号	特征组合(模板)	说明
1	—	只通过 BERT 模型提取特征
2	五笔	在特征组合 1 基础上，融入五笔特征
3	郑码	在特征组合 1 基础上，融入郑码特征
4	拼音	在特征组合 1 基础上，融入拼音特征
5	笔画	在特征组合 1 基础上，融入笔画特征
6	五笔+郑码	在特征组合 2 基础上，融入郑码特征
7	拼音+笔画	在特征组合 4 基础上，融入笔画特征
8	郑码+五笔+拼音	在特征组合 6 基础上，融入拼音特征
9	郑码+五笔+笔画	在特征组合 6 基础上，融入笔画特征
10	郑码+五笔+笔画+拼音	在特征组合 9 基础上，融入拼音特征

表 8 基于序列的对比实验结果(%)

Table 8 Results of comparative experiments based on tokens (%)

实验编号	特征组合(模板)	精确率 *P*	召回率 *R*	F1
1	—	91.85	93.19	92.06
2	五笔	93.30	94.89	93.38
3	郑码	91.89	93.80	92.44
4	拼音	91.91	93.62	92.32
5	笔画	92.52	93.41	92.52
6	五笔+郑码	94.40 (↑2.25)	97.25 (↑4.06)	95.37 (↑3.31)
7	笔画+拼音	91.70	93.41	92.11
8	郑码+五笔+拼音	91.99	93.86	92.51
9	郑码+五笔+笔画	92.53	93.41	92.52
10	郑码+五笔+笔画+拼音	92.25	93.46	92.36

表 9 基于实体粒度的未挂载武器装备领域规则触发器的对比实验结果(%)

Table 9 Comparison experimental results of rule triggers in the field of unmounted weapons and equipment based on entity granularity (%)

实验编号	特征组合(模板)	精确率 *P**	召回率 *R**	F1*
1	—	83.99	83.71	82.55
2	五笔	83.09	84.97	82.46
3	郑码	84.07	83.59	82.51
4	拼音	84.22	84.21	82.92
5	笔画	84.09	83.77	82.58
6	五笔+郑码	83.67 (↓0.32)	85.37 (↑1.66)	83.09 (↑0.54)
7	笔画+拼音	84.34	83.74	82.74
8	郑码+五笔+拼音	84.88	83.75	83.02
9	郑码+五笔+笔画	84.09	83.77	82.58
10	郑码+五笔+笔画+拼音	84.54	83.82	82.91

表 10 基于实体粒度的挂载武器装备领域规则触发器的对比实验结果(%)

Table 10 Comparison experimental results of rule triggers in the field of mounted weapons and equipment based on entity granularity (%)

实验编号	特征组合(模板)	精确率 *P**	召回率 *R**	F1*
1	—	84.33	87.57	85.00
2	五笔	83.38	88.11	84.66
3	郑码	84.23	87.65	85.01
4	拼音	84.74	88.08	85.47
5	笔画	84.43	87.78	85.09
6	五笔+郑码	83.83	88.71	85.27
7	笔画+拼音	84.78	87.91	85.38
8	郑码+五笔+拼音	85.24 (↑0.91)	87.69 (↑0.12)	85.55 (↑0.55)
9	郑码+五笔+笔画	84.43	87.73	85.09
10	郑码+五笔+笔画+拼音	84.95	87.83	85.45

基于实体粒度的评价方法得到解决。通过分析基准实验和对比试验的结果可以发现，采用基于实体粒度的评价方法并挂载武器装备领域规则触发器后，融合五笔、郑码和拼音的效果最佳，精确率、召回率和F1数值分别提高0.91%，0.12%和0.55%，说明五笔和郑码相互纠错并结合拼音的特征融合，对提升武器装备领域复杂命名实体识别的效果最有效。分别对比实验1和2、实验1和3可以发现，单独融入五笔或郑码特征，F1*值稍有下降，可能是因为单独的一种象形特征存在一些错误的表达，经过二者相互纠错，可以有效地避免这些错误。

对表9和10做横向对比可以发现，在有基准实验和不同特征组合的对比实验中，挂载武器装备领域规则触发器后，评价指标均显著提高。

如表11所示，经过识别样例分析，本文方法可以有效地识别上述武器装备包含的嵌套命名实体信息。

表12所示的文本包含非连续实体，基准模型BERT+Bi-LSTM+CRF只识别出普通命名实体和第一个非连续实体的部分片段“黄铜骑士”，本文方法能够识别出整段的非连续实体“黄铜骑士舰对空导弹、小猎犬舰对空导弹、鞑靼舰对空导弹”。

如表13所示，基准实验只抽取出普通实体(“纳希莫夫号###舰船”、“cads-1弹炮合一系统###系统”)和非连续实体的部分片段(“sa-n-6###导弹、sa-n-4###导弹、sa-n-9###导弹”)，本文方法可以抽取到嵌套实体“纳希莫夫###人名”和完整的非连续实体“sa-n-6导弹###导弹、sa-n-4导弹###导弹、sa-n-9导弹###导弹”。

从纵向角度观察对比试验，结果如图6所示。可以看出，对比实验6序列粒度的F1值最佳，为95.37%，其召回率从基准实验的93.19%显著地提

表11　武器装备嵌套实体识别样例

Table 11　Sample table of weapon and equipment nested entity recognition

样例	1992年1月俄总统叶利钦下令基洛夫级巡洋舰更名：乌沙科夫海军上将号(原基洛夫号)、拉扎耶夫海军上将号(原伏龙芝号)、纳希莫夫海军上将号(原加里宁号)、彼得大帝号(原安德罗波夫号)。
识别参考结果	基洛夫级巡洋舰###舰船、基洛夫###人名、乌沙科夫海军上将号###舰船、乌沙科夫###人名、基洛夫号###舰船、基洛夫###人名、拉扎耶夫海军上将号###舰船、伏龙芝号###舰船、伏龙芝###人名、纳希莫夫海军上将号###舰船、纳希莫夫###人名、加里宁号###舰船、加里宁###人名、彼得大帝号###舰船、彼得大帝###人名、安德罗波夫号###舰船、安德罗波夫###人名
基准实验	基洛夫级巡洋舰###舰船、乌沙科夫海军上将号###舰船、基洛夫号###舰船、拉扎耶夫海军上将号###舰船、伏龙芝号###舰船、纳希莫夫海军上将号###舰船、加里宁号###舰船、彼得大帝号###舰船、安德罗波夫号###舰船
本文方法	基洛夫级巡洋舰###舰船、基洛夫###人名、乌沙科夫海军上将号###舰船、乌沙科夫###人名、基洛夫号###舰船、基洛夫###人名、拉扎耶夫海军上将号###舰船、伏龙芝号###舰船、伏龙芝###人名、纳希莫夫海军上将号###舰船、纳希莫夫###人名、加里宁号###舰船、加里宁###人名、彼得大帝号###舰船、彼得大帝###人名、安德罗波夫号###舰船、安德罗波夫###人名

表12　武器装备非连续实体识别样例

Table 12　Sample table of discontinuous entity recognition of weapons and equipment

样例	“宙斯盾系统”的前身称为先进的水面导弹系统(asms)，asms计划是在1963年11月提出来的，当时的主要目的是用于对付80年代的空中威胁，准备取代黄铜骑士、小猎犬和鞑靼三种舰对空导弹。
识别参考结果	宙斯盾系统###系统、黄铜骑士舰对空导弹###导弹、小猎犬舰对空导弹###导弹、鞑靼舰对空导弹###导弹
基准实验	宙斯盾系统###系统、黄铜骑士######导弹
本文方法	纳希莫夫号###舰船、纳希莫夫###人名、sa-n-6导弹###导弹、sa-n-4导弹###导弹、sa-n-9导弹###导弹、cads-1弹炮合一系统###系统、纳希莫夫###人名

表13　武器装备复杂实体识别样例

Table 13　Sample table of complex entity recognition of weapons and equipment

样例	纳希莫夫号有sa-n-6、sa-n-4 sa-n-9三种导弹和一种cads-1弹炮合一系统，共有四个层次的对空火力。
识别参考结果	纳希莫夫号###舰船、纳希莫夫###人名、sa-n-6导弹###导弹、sa-n-4导弹###导弹、sa-n-9导弹###导弹、cads-1弹炮合一系统###系统、纳希莫夫###人名
基准实验	纳希莫夫号###舰船、纳希莫夫###人名、sa-n-6###导弹、sa-n-4###导弹、sa-n-9###导弹、cads-1弹炮合一系统###系统
本文方法	纳希莫夫号###舰船、纳希莫夫###人名、sa-n-6导弹###导弹、sa-n-4导弹###导弹、sa-n-9导弹###导弹、cads-1弹炮合一系统###系统、纳希莫夫###人名

高到 97.25%。实验 2~5 和实验 8~10 的精确率、召回率和 F1 值均有较显著的提高。实验 7 的精确率(91.70%)相较于基准实验(91.85%)略微下降，原因可能是笔画和拼音隶属不同的维度特征，将其拼接后会制造出噪音(从实验 4 和 5 可以发现拼音和笔画特征对精确率、召回率和 F1 值的提高有一定的促进作用)，即笔画数是用字符的复杂程度描述不同字符之间的特征，而拼音是用字符的读音维度描述不同字符的特征。

将对比实验从横向的角度观察，结果如图 7 所示。挂载武器装备领域规则触发器的情况下，实验 4 中实体粒度指标 F1 值最佳(85.55%)，精确率也最高(85.24%)，其他对比实验均比基准实验的评价值高。不挂载武器装备领域规则触发器的情况下，实验 6 中在实体粒度指标 F1 值最佳(83.09%)，召回率也最佳(85.37%)。对比实验 2 与基准实验，可以发现精确率从 83.99%下降到 83.09%，可能是因为五笔的设计缺陷导致(存在五笔表达与英语单词、拼音特征相同的情况)，会对模型的学习造成一定的干扰。

5 结论

本文提出一种融合多特征挂载领域规则触发器的武器装备领域复杂命名实体识别方法，可以提高武器装备领域复杂实体的识别效果，以最大程度地挖掘所有实体，为武器装备领域知识图谱的构建提供有力的知识支撑。该方法在 BERT，Bi-LSTM 和 CRF 多神经网络协同下，融入郑码、五笔、笔画和拼音特征，其中郑码和五笔可以有效地互相纠正，提升识别效果。在此基础上，利用挂载武器装备领域触发器，不仅解决了嵌套的人名和地名实体识别问题，而且还解决了武器装备的非连续实体的识别问题。本文还提出更适用于复杂命名实体识别的基于实体粒度的评价方法。实验结果表明，与现有的

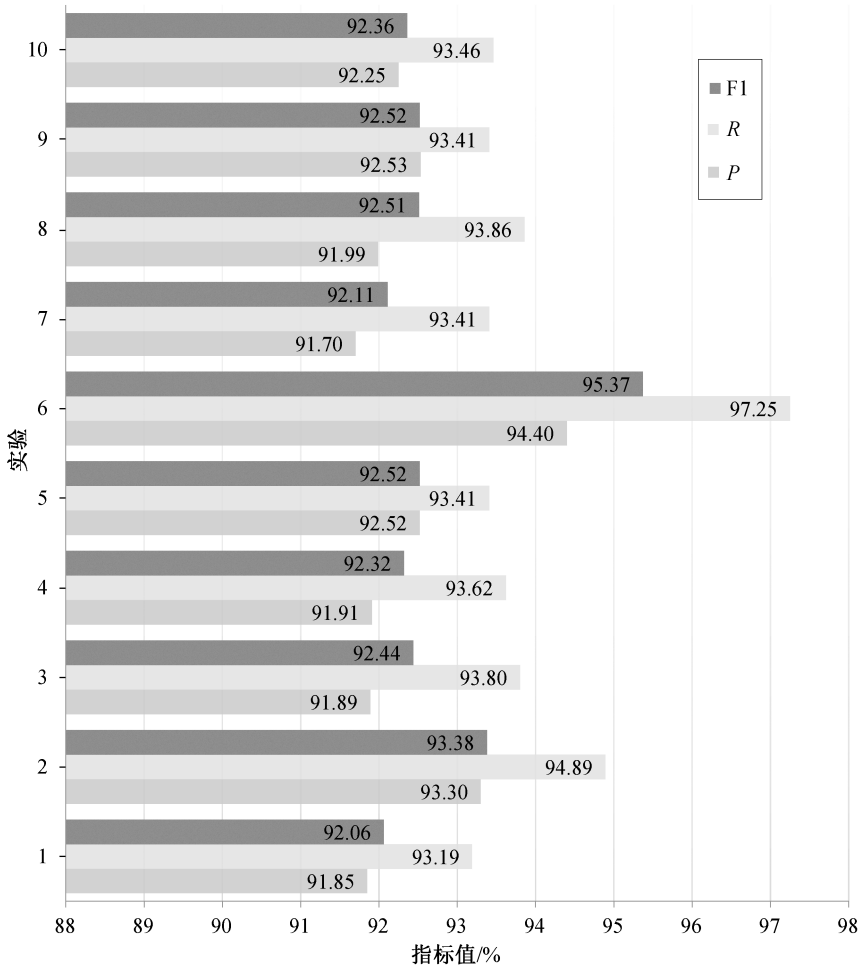
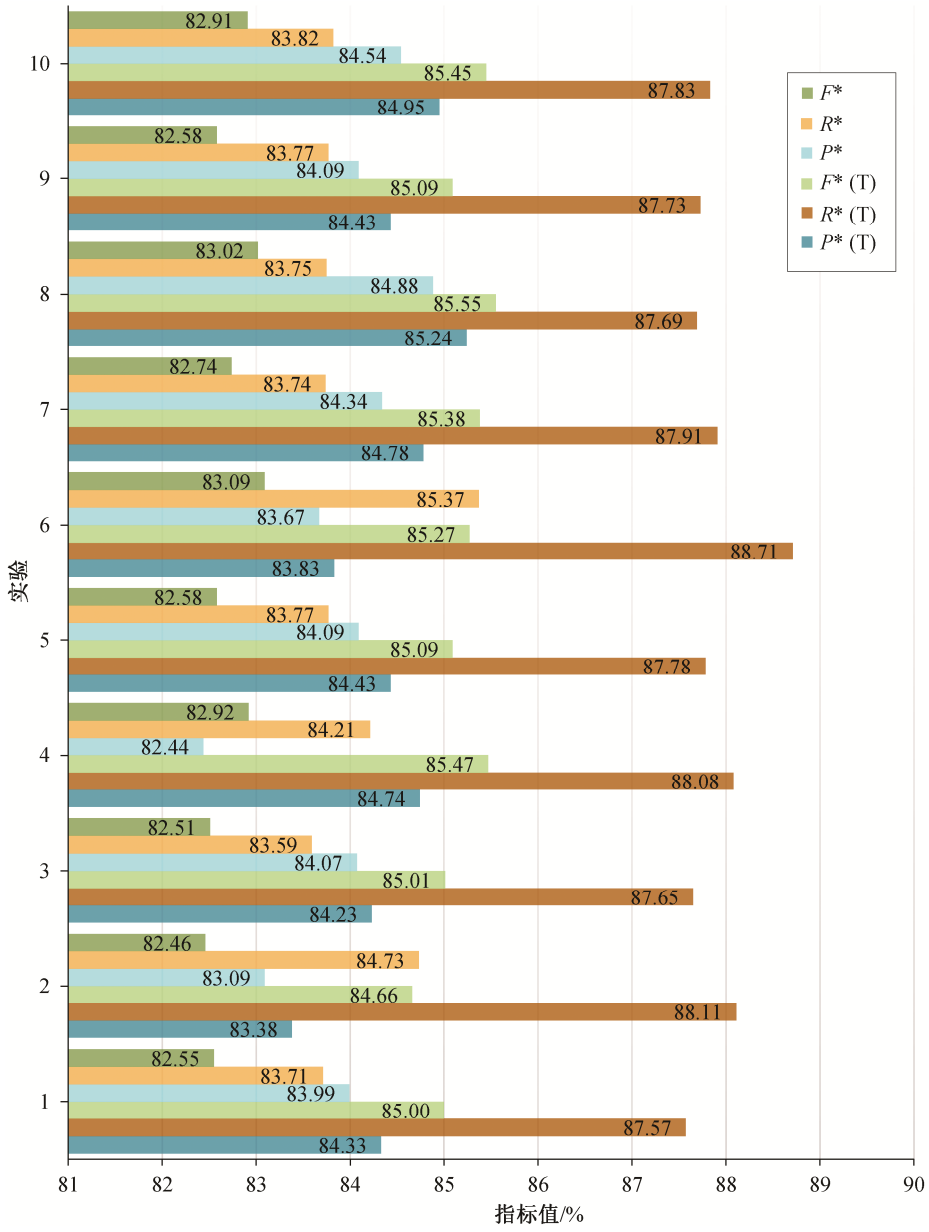


图 6 基于序列的评价指标下的对比实验结果

Fig. 6 Comparison experiment results based on token-based evaluation indicators



T 代表挂载武器装备领域触发器的评价结果

图 7　基于实体粒度的评价指标下的对比实验结果

Fig. 7　Comparative experimental results based on the evaluation index of entity granularity

主流方法相比，本文提出的方法在武器装备领域复杂实体识别任务中效果最佳。

参考文献

[1] 陈曙东，欧阳小叶. 命名实体识别技术综述. 无线电通信技术, 2020, 46(3): 251–260

[2] 刘浏，王东波. 命名实体识别研究综述. 情报学报, 2018, 37(3): 329–340

[3] Collins M, Singer Y. Unsupervised models for named entity classification // Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora. Hong Kong, 1999: 100–110

[4] Petasis G, Cucchiarelli A, Velardi P, et al. Automatic adaptation of proper noun dictionaries through cooperation of machine learning and probabilistic methods // Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Athens, 2000: 128–135

[5] Zhou J T, Zhang H, Jin D, et al. Roseq: robust sequence labeling. IEEE transactions on neural net-

works and learning systems, 2019, 31(7): 2304–2314

[6] 王子牛, 姜猛, 高建瓴, 等. 基于 BERT 的中文命名实体识别方法. 计算机科学, 2019, 46(S2): 138–142

[7] Finkel J R, Manning C D. Nested named entity recognition // Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. Singapore, 2009: 141–150

[8] Lu W, Roth D. Joint mention extraction and classification with mention hypergraphs // Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, 2015: 857–867

[9] Ju M, Miwa M, Ananiadou S. A neural layered model for nested named entity recognition // Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). New Orleans, 2018: 1446–1459

[10] Xia C, Zhang C, Yang T, et al. Multi-grained named entity recognition [EB/OL]. (2019–06–20)[2021–03–05]. https://arxiv.org/abs/1906.08449v1

[11] Li X, Feng J, Meng Y, et al. A unified MRC framework for named entity recognition [EB/OL]. (2020–05–23)[2021–03–05]. https://arxiv.org/abs/1910.11476

[12] 姜文志, 顾佼佼, 丛林虎. CRF 与规则相结合的军事命名实体识别研究. 指挥控制与仿真, 2011, 33(4): 13–15

[13] 冯蕴天, 张宏军, 郝文宁. 面向军事文本的命名实体识别. 计算机科学, 2015, 42(7): 15–18

[14] 朱佳晖, 张文峰, 刘卫平, 等. 基于双向 LSTM 和 CRF 的军事命名实体识别和链接//第六届中国指挥控制大会论文集(上册). 北京, 2018: 470–475

[15] 尹学振, 赵慧, 赵俊保, 等. 多神经网络协作的军事领域命名实体识别. 清华大学学报(自然科学版), 2020, 60(8): 648–655

[16] 姜文志, 顾佼佼, 胡文萱, 等. 基于多模型结合的军事命名实体识别. 兵工自动化, 2011, 30(10): 90–93

[17] 单赫源, 张海粟, 吴照林. 小粒度策略下基于 CRFs 的军事命名实体识别方法. 装甲兵工程学院学报, 2017, 31(1): 84–89

[18] 单义栋, 王衡军, 王娜. 基于多标签的军事领域命名实体识别. 计算机科学, 2019, 46(11A): 9–12

[19] Devlin J, Chang M W, Lee K, et al. BERT: pre-training of deep bidirectional transformers for language understanding [EB/OL]. (2019–05–24)[2021–03–05]. https://arxiv.org/abs/1810.04805

[20] Meng Y, Wu W, Wang F, et al. Glyce: glyph-vectors for chinese character representations [EB/OL]. (2020–05–21)[2021–03–05]. https://arxiv.org/abs/1901.10125

[21] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation, 1997, 9(8): 1735–1780

[22] Graves A, Schmidhuber J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Networks, 2005, 18(5/6): 602–610