**知识库构建与融合**

## 技术方法现状

### 实体知识图谱构建

知识库构建的主要任务是根据相关原始数据进行知识抽取。知识抽取通过自动化或半自动化的知识抽取技术获得实体、关系和属性等可用知识单元。知识抽取的主要三类任务为实体识别、关系抽取和属性抽取，对于每种不同的任务，有其对应的技术方法。

实体识别的目的是从海量的原始数据中提取知识图谱所需的特定实体信息。现阶段，实体识别方法主要为基于规则、基于统计模型及基于神经网络三类[1]。

基于规则的方法通过手工构建规则集，将纯文本等非结构化数据与规则集匹配来实现实体识别。该方法在处理小规模的知识图谱时精度较高，但是随着知识图谱规模的增大，规则构建困难，难以进行大规模扩展并应用于不同领域的知识图谱。

基于统计模型的方法将实体识别作为序列标注问题，以完全或部分标注的语料进行模型训练。常见的统计模型有条件马尔科夫模型、隐马尔科夫、条件随机场和最大熵等。基于统计模型的方法通用性强。但是状态搜索空间庞大、训练时间长、高度依赖特征选取和语料库；由于统计模型的状态空间爆炸问题，难以处理海量数据。

深度学习能够自动地从数据中学习复杂的隐藏特征，所需的领域专业知识和经验知识较少，基于神经网络的 NER 已成为目前 主流方法。主要模型有卷积神经网络CNN、循环神经网络RNN、长短期记忆网络、双向长短期记忆网络等。基于深度学习的实体识别算法有更好的通用性，一句对海量数据处理的能力。

关系抽取通过获取实体之间的某种语义关系或关系的类别，自动识实体对及联系这一对实体的关系所构成的三元组。早期的关系抽取研究方法主要是通过人工构造语法和语义规则进而采用模式匹配的方法来识别实体间的关系，此种方式要求指定规则的人对特定领域有深入的理解和认知且规则制定工作量大。近几年来，关系抽取研究大多是基于神经网络，主要包括卷积神经网络、循环神经网络、注意力机制、图卷积网络、对抗训练和强化学习等技术[2]。

属性抽取通过从不同信息源的原始数据中抽取实体的属性名和属性值，构建实体的属性列表，形成完整的实体概念。目前，属性抽取方法可分为传统的监督、无监督和半监督属性抽取以及基于神经网络的属性抽取和元模式的属性抽取。大部分属性抽取是从文本数据中抽取信息，与实体识别类似，传统方法中监督学习的属性抽取使用统计模型抽取属性。基于神经网络的属性抽取使用序列标注模型抽取文本中的属性值。基于元模式的属性抽取将类型化的文本模式结构命名为元结构，从而在海量语料库中发现元模式。

### 事件知识图谱构建

事件知识图谱多采用“自顶向下”的构建方法，先定义数据模式，再添加实例内容向下细化，构建技术包括事件知识表示、事件知识抽取、事件关系抽取等。事件知识图谱从自然语言文本中抽取事件和实体、属性、关系等并进行知识融合，然后通过本体构建体系框架，以结构化的三元组形式存储。

事件知识表示直接关系到知识推理、知识计算的应用，是事件知识图谱的核心部分。传统的事件表示模型主要基于5W ( Who，When，Where，What，Why) 构建，没有建立事件之间的关系连接。Silver等在2011年提出了一种离散事件表示本体模型，提供了事件之间的关系表示方法。Van Hage等2011年构建了简单事件模型( Simple Event Model，SEM) ，对不同领域中的事件进行建模，建立了事件与时间、地点和实体之间的链接，但无法表示事件之间的关系。Rospocher 等2016年和 Gottschalk 等2018 年在简单事件模型(SEM)基础上建立了事件之间的关联，构建事件知识图谱。

事件关系抽取以事件为基本的语义单元，自动抽取事件之间的逻辑关系，包括事件的共指关系、因果 关系、时序关系等。事件关系的抽取和构建可以揭示事件发展规律，厘清事件关联并全面了解事件，进而构建事件知识图谱。

因果关系指事件之间的作用关系，即某个事件是另一事件的结果。在事件知识图谱的推理应用中具有重大意义。杨竣辉等2016年提出了通过构建事件和事件元素的语义关联，进行事件因果关系的识别方法。付剑峰等2011年提出了一种基于层叠条件随机场的事件因果 关系抽取模型，将事件因果关系建模为序列标注问题。Sorgent 等2013 年通过制定规则进行事件因果关系抽取，并利用贝叶斯推理优化结果。

[1] Goyal, Archana, Vishal Gupta, and Manish Kumar. "Recent named entity recognition and classification techniques: a systematic review." Computer Science Review 29 (2018): 21-43.

[2] LiuQiao, LiYang, and LiuYao DuanHong. "Knowledge graph construction techniques." Journal of computer research and development 53.3 (2016): 582.

## 2、设计思路

传统的实体识别任务主要分为实体类识别、时间类识别和数字类识别三个大类和人名、机构名、地点、时间、日期、货币和百分比识别七个小类。但是本文指控领域涉及的实体类别远远不止这些，因此本方法将带有领域特色的实体识别作为研究重点。

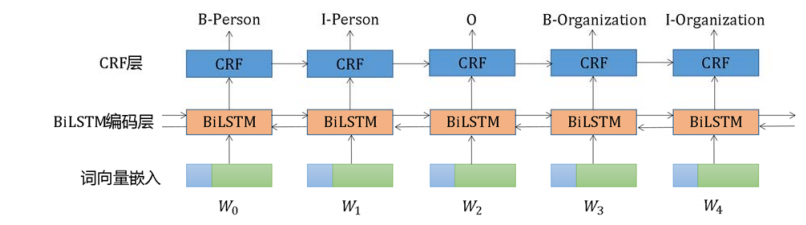
相较于传统的三大类和七小类实体识别，指控领域实体识别的特点是，领域中实体类别的粒度更细且普遍带有军事特征。军事实体识别的任务可以描述为，将指控信息文本语料进行预处理后，从中提取句子特征，将特征输入到实体识别模型中，从而识别出具有特定意义的军事实体。

军事领域关系抽取的任务可以描述为，将军事文本语料进行预处理后，依靠前文已经提取到的军事相关实体，利用神经网络模型对文本数据中实体之间的关系进行抽取，然后结合实体和抽取到的关系组成（实体 1，关系，实体 2）三元组数据，从而完善军事知识图谱的构建。

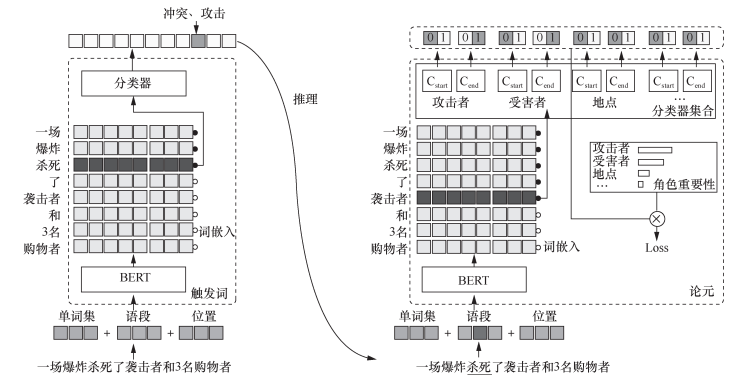
### 事件知识图谱构建

事件是文本中包含的一种特殊信息，事件抽取就是从非结构化的文本数据中抽取与事件有关的各种角色，将信息用结构化数据表示。按照确定事件类别的方法，事件抽取可以被分为限定域事件抽取和开放域事件抽取。事件知识图谱构建与实体知识图谱类似，通常采用基于本体的构建方法进行知识建模，先构建事件知识图谱顶层表示模式，再向下细化补充实例。

军事领域实体标注的任务可以描述为，将军事文本语料进行预处理后，依靠前文已经提取到的军事相关实体，利用神经网络模型对文本数据中的实体进行标注，然后结合实体和抽取到的关系组成三元组数据，完善军事知识图谱的构建。

目前应用较为广泛的实体识别方法是BiLSTM-CRF，该方法将句子中的词向量作为双向长短期记忆网络的输入。通过提取句子特征，结合条件随机场（CRF）完成实体的标注。

BiLSTM-CRF模型结构

预训练模型可以从大规模的语料中提取隐含的语义信息，学习到更好的通用语义表示向量，从而提高下游任务的表现。Yang S等人[2]针对现有远程监督事件抽取方法中存在的问题，将目光转向预训练的语言模型，希望利用从大规模语料库中学习到的知识表示向量来提高模型的性能。其设计了一种基于预训练语言模型的事件抽取（pre-trained language model based event extractor，PLMEE）模型。该模型的结构如图所示[2]。他们将事件抽取看作由两个子任务组成，两个子任务分别是触发词抽取和论元抽取，并提出了以预训练语言模型为基础的触发词抽取器和论元抽取器。基于预训练模型的方法会使事件的语义表示更加精确。

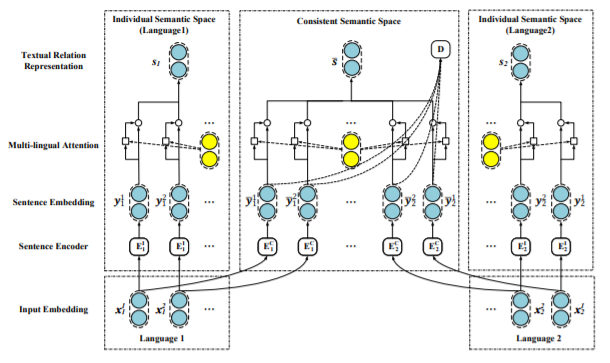
PLMEE模型结构

为了在神经网络中引入句法结构特征，Nguyen T等人[3]提出了一个基于句法依存树的GCN模型用于事件抽取。在GCN中，每个节点的卷积向量是由相邻节点的表示向量计算出来的，可以作为该节点的唯一特征进行分类。另外，模型中通过对当前单词的卷积向量以及句子中提到的实体进行池化操作，克服实体指称无法捕捉的问题。池化操作聚合了卷积向量，从而为事件类型预测生成了单个向量表示。该方法在事件抽取中引入了GCN模型，将句法依存树上的信息进行聚合，首次利用了句子中的结构信息。除此之外，Liu X等人[4]设计了一个新的联合多事件抽取（jointly multiple events extraction，JMEE）框架。该框架利用基于注意力机制的图卷积神经网络进行建模，并通过引入句法依存树中的句法捷径弧来增强信息流，以此来提升在一个句子中抽取多个事件的效果。使用句法捷径弧可以减少将信息流从一个节点转换。

### 模型知识图谱构建

模型类知识图谱的作用主要在于以实体知识图谱以及事件知识图谱为基础，并以既定的时间应对策略为知识图谱构建数据。对指控系统所发生的事件进行及时的决策响应。

类似于实体关系抽取以及时间关系抽取，使用与之相同的方法识别实体。

由于指控系统的专业性以及可靠性要求，以及指控系统的复杂性，常用的基于神经网络的模型通常忽略了输入数据中的噪声，导致模型效果不佳。将对抗训练AT引入关系抽取，对训练数据产生对抗噪声来优化分类算法，能够有效提升模型鲁棒性。Wu等［5］首次提出在关系抽取中应用AT策略，Wang等［6］在此基础上，提出了 AMNRE（Adversarial Multi-lingual Neural Relation Extraction），以 AT策略保证对多种语言的句子表征，提取出语义一致的关系。

[1]薛坤. 面向军事领域的知识图谱构建与应用研究[D].大连理工大学,2020.

[2]Yang S, Feng D, Qiao L, et al. Exploring pre-trained language models for event extraction and generation[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 5284-5294.

[3]Nguyen T H, Grishman R. Graph convolutional networks with argument-aware pooling for event detection[C]//Thirty-second AAAI conference on artificial intelligence. 2018.

[4]Liu X, Luo Z, Huang H. Jointly multiple events extraction via attention-based graph information aggregation[J]. arXiv preprint arXiv:1809.09078, 2018.

[5]Wu Y, Bamman D, Russell S. Adversarial training for relation extraction[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017: 1778-1783.

[6]Wang X, Han X, Lin Y, et al. Adversarial multi-lingual neural relation extraction[C]//Proceedings of the 27th International Conference on Computational Linguistics. 2018: 1156-1166.