第九周

《威胁情报知识图谱构建与应用关键技术研究》中所使用的实体抽取和关系抽取方法，是作者根据威胁情报领域中的非结构化数据，采用将词特征、字符特征、实体边界特征以及实体上下文特征融合起来作为实体抽取时神经网络所使用的特征，采用实体间关联的全局语义特征和局部语义特征融合起来以及表征威胁情报实体关系的语义特征作为分类器特征。

本质上是将实体抽取和关系抽取分为两部分进行，先进行实体的抽取后进行关系的抽取。这种分步的抽取方法叫做流水线方法。流水线方法在命名实体识别的基础上进行关系抽取，会导致实体识别中产生的错误影响关系预测结果，造成错误传播。

基于流水线的方法进行关系抽取的主要流程可以描述为：针对已经标注好的目标实体对的句子进行关系抽取，最后把存在实体关系的三元组作为预测结果输出。主流的方法主要有基于 RNN 的实体关系抽取方法、基于 CNN 的实体关系抽取方法、基于 LSTM 的实体关系抽取方法。

流水线方法中存在的共同的问题，1、错误传播，实体识别模块的错误会影响到接下来关系分类性能。2、忽视了两个子任务之间存在的关系，导致信息丢失。3、产生冗余信息，由于对识别出来的实体进行两两配队，然后再进行关系分类，那些没有关系的实体对就会带来多余信息，提升错误率。

在《威胁情报知识图谱构建与应用关键技术研究》中所使用的流水线方法也同样不能避免这样的问题。要解决这些问题就要使用联合学习方法。

联合学习方法能够利用实体和关系间紧密的交互信息，同时抽取实体并分类

实体对的关系，很好的解决了流水线方法所存在的问题。

联合学习方法通过实体识别和关系分类联合模型，直接得到存在关系的实体三元组。在联合学习方法中建模的对象不同，联合学习方法又可以分为参数共享方法和序列标注方法。参数共享方法分别对实体和关系进行建模，而序列标注方法则是直接对实体-关系三元组进行建模。

基于参数共享的实体关系抽取方法。

针对的是流水线方法中存在的错误累计传递为题和忽视两个子任务间关系依赖的问题。在此方法中，实体识别子任务和关系抽取子任务通过共享联合模型的编码层来进行联合学习，通过共享编码层，在训练时，两个子任务都会通过后向传播算法更新编码层的共享参数，以此来实现两个子任务之间的相互依赖，最终找到全局任务的最佳参数。在联合学习模型中，输入的句子在通过共享的编码层后，在解码层会首先进行实体识别子任务，再利用实体识别的结果，并对存在关系的实体对进行关系分类，最终输出实体-关系三元组。

Miwa 等人在 2016 年首次将神经网络的方法用于联合表示实体和关系,其模型图如图 5 所示.在该模型 中,实体识别子任务和关系分类子任务共享编码层的 LSTM 单元序列表示(编码层包括 LSTM 单元和隐藏层). 该方法将实体识别任务当作序列标注任务,使用双向序列 LSTM 输出具有依赖关系的实体标签;之后,通过在双 向序列 LSTM 单元上堆叠双向树结构 LSTM 的方法,使关系分类子任务和实体识别子任务共享编码层的 LSTM 单元序列表示,同时,在关系分类子任务中捕获词性标签等依赖特征和实体识别子任务中输出的实体序列,形成 1800 Journal of Software 软件学报 Vol.30, No.6, June 2019 依存树,最终根据依存树中目标实体间的最短路径对文本进行关系抽取.但该模型中的关系分类子任务和

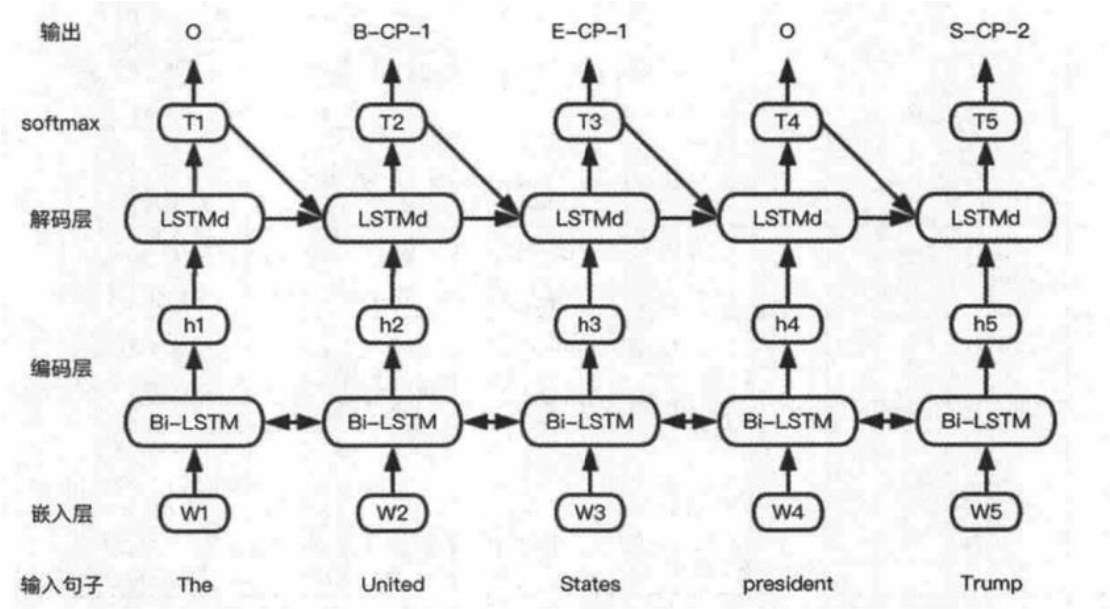
实体 识别子任务仅共享了编码层的双向序列 LSTM 表示,从严格意义上来说不是真正的联合模型.但是该模型的提 出,为之后真正意义上联合学习模型的提出奠定了基础,是基于深度学习方法做联合学习模型的启发者。

基于序列标注的实体关系抽取方法。

基于参数共享的实体关系抽取方法，改善了传统流水线方法中存在的错误累计传递问题和忽视两个子任务间关系依赖的问题，但因其在训练时还是需要先进行命名实体识别子任务，再根据实体预测信息对实体进行两两匹配，最后进行关系分类子任务，所以在模型实现过程中仍然会产生没有关系的实体这种冗余信息，为解决这种问题，基于新序列标注方法的实体，关系联合抽取方法被提出。

Zheng 等人在 2017 年提出了基于新的标注策略的实体关系抽取方法,把原来涉及到命名实体识别和关 系分类两个子任务的联合学习模型完全变成了一个序列标注问题.在该方法中,共包含 3 种标注信息: (1) 实体中词的位置信息{B,I,E,S,O},分别表示{实体开始,实体内部,实体结束,单个实体,无关词};(2) 实体关系类型信息,需根据实际需要自定义关系类型并编码,如{CF,CP,…};(3) 实体角色信息{1,2},分别表示{实体 1,实 体 2}.该方法能使用序列标注的方法同时识别出实体和关系,避免了复杂的特征工程,通过一个端到端的神经网 络模型直接得到实体-关系三元组,解决了基于参数共享的实体关系抽取方法可能会带来的实体冗余的问题.新 序列标注方法的模型图如图所示.在该端到端的神经网络模型中,对输入的句子,首先,编码层使用 Bi-LSTM 来 进行编码;之后,解码层再使用 LSTM 进行解码;最终,输出模型标注好的实体-关系三元组.另外,Zheng 等人在这篇论文中还对该端到端模型增加了偏置损失函数,该函数增强了相关实体对之间的联系,削弱了无效实体标 签的影响力,提高了关系分类的准确率;并基于这种新的标注方

法,该论文中还学习用不同的端到端模型来解决 关系抽取问题。



目前联合学习方法中虽然有基于参数共享的实体关系抽取方法和基于序列标注的实体关系抽取方法，这两种方法能够很好的解决流水线方法中存在的问题，但仍存在对于有监督领域中重叠实体识别的问题。这也是今后需要考虑研究的地方。