

@高能NLP

nlp中的实体关系抽取方法总结

nlp中的实体关系抽取方法总结



JayLou娄...

知乎专栏《高能NLP》作者

已关注

流光、忆臻、LMdeLiangMi、zenRRan 等 498 人赞同了该文章

本文以QA形式总结了「nlp中的实体关系联合抽取方法」。

2020年5月20日更新：

DeepIE: [github.com/loujie0822/D...](https://github.com/loujie0822/DeepIE)，基于深度学习的信息抽取技术集散地，欢迎大家关注，包含实体、关系、属性、事件、链接&标准化等。

为了更好的阅读体验，建议使用PC端浏览。如需下载本篇文档，可以到我的[github](#)下载。

Question List

- Q1: 与联合抽取对比，Pipeline方法有哪些缺点？
- Q2: NER除了LSTM+CRF，还有哪些解码方式？如何解决嵌套实体问题？
- Q3: Pipeline中的关系分类有哪些常用方法？如何应用弱监督和预训练机制？怎么解决高复杂度问题、进行one-pass关系分类？
- Q4: 什么是关系重叠问题？
- Q5: 联合抽取难点在哪里？联合抽取总体上有何方法？各有哪些缺点？
- Q6: 介绍基于共享参数的联合抽取方法？
- Q7: 介绍基于联合解码的联合抽取方法？
- Q8: 实体关系抽取的前沿技术和挑战有哪些？如何解决低资源和复杂样本下的实体关系抽取？如何应用图神经网络？
- 彩蛋：百度2020关系抽取比赛的baseline可以采取哪些方法？

实体关系抽取（Entity and Relation Extraction，**ERE**）是信息抽取的关键任务之一。ERE是级联任务，分为两个子任务：实体抽取和关系抽取，如何更好处理这种类似的级联任务是NLP的一个热点研究方向。



本文结构

Q1：与联合抽取对比，Pipeline方法有哪些缺点？

Pipeline方法指先抽取实体、再抽取关系。相比于传统的Pipeline方法，联合抽取能获得更好的性能。虽然Pipeline方法易于实现，这两个抽取模型的灵活性高，实体模型和关系模型可以使用独立的数据集，并不需要同时标注实体和关系的数据集。但存在以下缺点：

- 1. 误差积累：实体抽取的错误会影响下一步关系抽取的性能。
- 2. 实体冗余：由于先对抽取的实体进行两两配对，然后再进行关系分类，没有关系的候选实体所带来的冗余信息，会提升错误率、增加计算复杂度。
- 3. 交互缺失：忽略了这两个任务之间的内在联系和依赖关系。

（基于共享参数的联合抽取方法仍然存在训练和推断时的gap，推断时仍然存在误差积累问题，可以说只是缓解了误差积累问题。）

Q2：NER除了LSTM+CRF，还有哪些解码方式？如何解决嵌套实体问题？

虽然NER是一个比较常见的NLP任务，通常采用LSTM+CRF处理一些简单NER任务。NER还存在嵌套实体问题（实体重叠问题），如「《叶圣陶散文选集》」中会出现两个实体「叶圣陶」和「叶圣陶散文选集」分别代表「作者」和「作品」两个实体。而传统做法由于每一个token只能属于一种Tag，无法解决这类问题。笔者尝试通过归纳几种常见并易于理解的 **实体抽取解码方式** 来回答这个问题。

1、序列标注：SoftMax和CRF

本质上是token-level 的多分类问题，通常采用CNNs/RNNs/BERT+CRF处理这类问题。与SoftMax相比，CRF进了标签约束。对这类方法的改进，介绍2篇比较有价值的工作：

- 针对CRF解码慢的问题，LAN^[1]提出了一种逐层改进的基于标签注意力机制的网络，在保证效果的前提下比 CRF 解码速度更快。文中也发现BiLSTM-CRF在复杂类别情况下相比BiLSTM-softmax并没有显著优势。
- 由于分词边界错误会导致实体抽取错误，基于LatticeLSTM^[2]+CRF的方法可引入词汇信息并避免分词错误（词汇边界通常为实体边界，根据大量语料构建词典，若当前字符与之前字符构成词汇，则从这些词汇中提取信息，联合更新记忆状态）。

但由于这种序列标注采取E
如图所示。

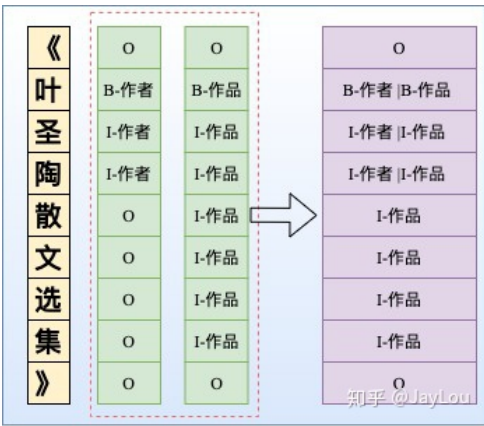


基于BILOU标注框架，笔者尝试给出了2种改进方法去解决实体重叠问题：

- 改进方法1：采取token-level 的多label分类，将SoftMax替换为Sigmoid，如图所示。当然这种方式可能会导致label之间依赖关系的缺失，可采取后处理规则进行约束。



- 改进方法2：依然采用CRF，但设置多个标签层，对于每一个token给出其所有的label，然后将所有标签层合并。显然这可能会增加label数量^[3]，导致label不平衡问题。基于这种方式，文献^[4]也采取先验图的方式去解决重叠实体问题。



2、Span抽取：指针网络

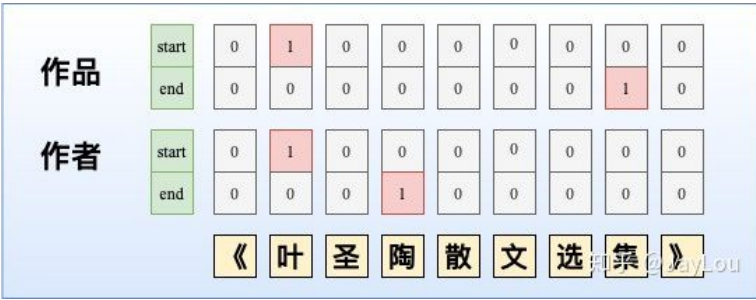
指针网络（PointerNet）最早应用于MRC中，而MRC中通常根据1个question从passage中抽取1个答案片段，转化为2个n元SoftMax分类预测头指针和尾指针。对于NER可能会存在多个实体Span，因此需要转化为n个2元Sigmoid分类预测头指针和尾指针。

将指针网络应用于NER中，可以采取以下两种方式：

第一种：**MRC-QA+单层指针网络**。在ShannonAI的文章中^[5]，构建query问题指代所要抽取的实体类型，同时也引入了先验语义知识。如图所示，由于构建query问题已经指代了实体类型，所以使用单层指针网络即可；除了使用指针网络预测实体开始位置、结束位置外，还基于开始和结束位置对构成的所有实体Span预测实体概率^[6]。此外，这种方法也适合于给定事件类型下的事件主体抽取，可以将事件类型当作query，也可以将单层指针网络替换为CRF。



第二种：**多层label指针网络**。由于只使用单层指针网络时，无法抽取多类型的实体，我们可以构建多层指针网络，每一层都对应一个实体类型。



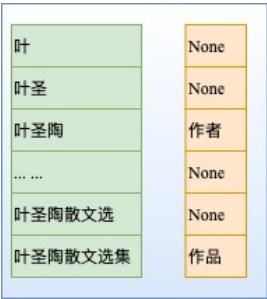
需要注意的是：

- 1) MRC-QA会引入query进行实体类型编码，这会导致需要对愿文本重复编码输入，以构造不同的实体类型query，这会提升计算量。
- 2) 笔者在实践中发现，n个2元Sigmoid分类的指针网络，会导致样本Tag空间稀疏，同时收敛速度会较慢，特别是对于实体span长度较长的情况。

3、片段排列+分类

上述序列标注和Span抽取的方法都是停留在token-level进行NER，间接去提取span-level的特征。而基于片段排列的方式^[7]，显示的提取所有可能的片段排列，由于选择的每一个片段都是独立的，因此可以直接提取span-level的特征去解决重叠实体问题。

对于含T个token的文本，理论上共有 $N = T(T + 1)/2$ 种片段排列。如果文本过长，会产生大量的负样本，在实际中需要限制span长度并合理削减负样本。



需要注意的是：

1. 实体span的编码表示：在span范围内采取注意力机制与基于原始输入的LSTM编码进行交互。然后所有的实体span表示并行的喂入SoftMax进行实体分类。
2. 这种片段排列的方式对于长文本复杂度是较高的。

4、Seq2Seq：

ACL2019的一篇paper中采取Seq2Seq方法^[3]，encoder部分输入的原文tokens，而decoder部分采取hard attention方式one-by-one预测当前token所有可能的tag label，直至输出<eow> (end of word) label，然后转入下一个token再进行解码。



Q3：Pipeline中的关系分类有哪些常用方法？如何应用弱监督和预训练机制？怎么解决高复杂度问题、进行one-pass关系分类？

（注：Pipeline方法中，关系抽取通常转化为一个分类问题，笔者这里称之为「关系分类」）

1、**模板匹配**：是关系分类中最常见的方法，使用一个模板库对输入文本两个给定实体进行上下文匹配，如果满足模板对应关系，则作为实体对之间的关系。常见的模板匹配方法主要包括：

- **人工模板**：主要用于判断实体间是否存在上下位关系。上下位关系的自然语言表达方式相对有限，采用人工模板就可以很好完成关系分类。但对于自然语言表达形式非常多的关系类型而言，这就需要采取统计模板。
- **统计模板**：无须人工构建，主要基于搜索引擎进行统计模板抽取。具体地，将已知实体对作为查询语句，抓取搜索引擎返回的前n个结果文档并保留包含该实体对的句子集合，寻找包含实体对的最长字串作为统计模板，保留置信度较高的模板用于关系分类。

基于模板匹配的关系分类构建简单、适用于小规模特定领域，但召回率低、可移植性差，当遇到另一个领域的关系分类需要重新构建模板。

2、半监督学习

bootstrapping（自举）：利用少量的实例作为初始种子集合，然后在种子集合上学习获得关系抽取的模板，再利用模板抽取更多的实例，加入种子集合中并不断迭代。

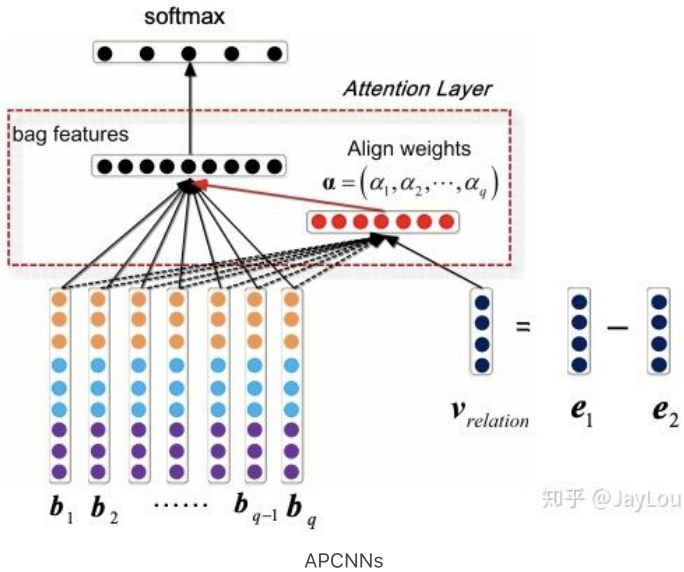
- bootstrapping比较常见的方法有DIPRE和Snowball。和DIPRE相比，Snowball通过对获得的模板pattern进行置信度计算，一定程度上可以保证抽取结果质量。
- bootstrapping的优点构建成本低，适合大规模的关系任务并且具备发现新关系的能力，但也存在对初始种子较为敏感、存在语义漂移、准确率等问题。

远程监督：其主要的基本假设是，如果一个实体对满足某个给定关系，那么同时包含该实体对的所有句子（构成一个Bag）都可能在阐述该关系。可以看出，该假设是一个非常强的假设，实际上很多包含该实体对的句子并不代表此种关系，会引入大量噪声。为了缓解这一问题，主要采取「**多示例学习**」、「**强化学习**」和「**预训练机制**」：

（1）**多示例学习**：主要基于Bag的特征进行关系分类，主要代表文献包括PCNN^[8]、Selective Attention over Instances^[9]、Multi-label CNNs^[10]、APCNNs^[11]，其中Bag的表示主要方式和池化方式为：

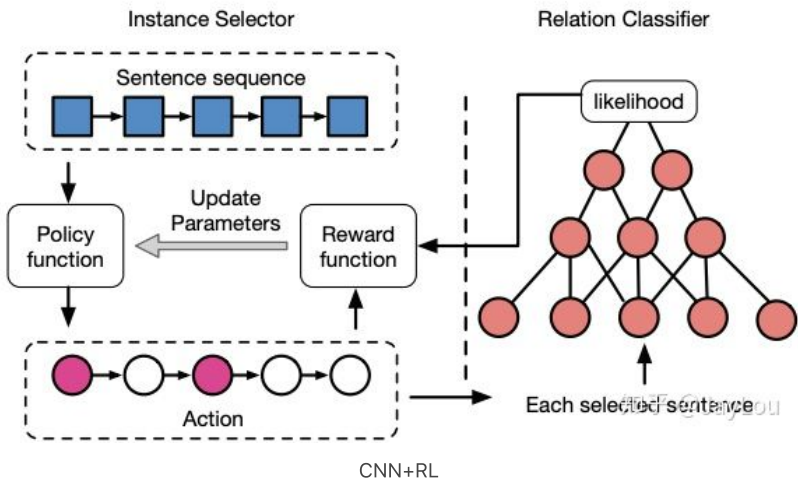
文献	多示例学习中Bag的表示	句子级别池化方式
PCNN	At-Least-One	Piecewise only one sentence
SelectAtt/ APCNNs	Attention weighted sum over bag	Piecewise all sentence
Multi-label CNNs	Max of each feature over bag	Cross all sentence

以APCNNs为例，采取PCNN模型^[8]提取单一句子的特征向量，最后通过attention加权得到Bag级别的特征，关系分类是基于Bag特征进行的，而原始的PCNN模型只选择Bag中使得模型预测得分最高的句子用于模型参数的更新，这会损失很多信息。



(2) 强化学习：在采用多示例学习策略时，可能会出现整个Bag包含大量噪声的情况。基于强化学习的CNN+RL^[12]比句子级别和Bag级别的关系分类模型取得更好效果。

模型主要由样例选择器和关系分类器构成。样例选择器负责从样例中选择高质量的句子，采取强化学习方式在考虑当前句子的选择状态下选择样例；关系分类器向样例选择器反馈，改进选择策略。



(3) 预训练机制：采取“Matching the Blank^[13]”方法，首次在预训练过程中引入关系分类目标，但仍然是自监督的，没有引入知识库和额外的人工标注，将实体mention替换为「BLANK」标识符。

- 该方法认为包含相同实体对的句子对为正样本，而实体对不一样的句子对为负样本。如图， r_A 和 r_B 构成正样本， r_A 和 r_C 构成负样本。
- 不同于传统的远程监督，该方法训练中不使用关系标签，采用二元分类器对句子对进行相似度计算。预训练的损失包含2部分：MLM loss 和 二元交叉熵关系损失。
- 在FewRel数据集上，不进行任何tuning就已经超过了有监督的结果。

r_A	In 1976, e_1 (then of Bell Labs) published e_2 , the first of his books on programming inspired by the Unix operating system.
r_B	The “ e_2 ” series spread the essence of “C/Unix thinking” with makeovers for Fortran and Pascal. e_1 ’s Ratfor was eventually put in the public domain.
r_C	e_1 worked at Bell Labs alongside e_3 creators Ken Thompson and Dennis Ritchie.
Mentions	e_1 = Brian Kernighan, e_2 = Software Tools, e_3 = Unix

3、监督学习：主要分为基于特征、核函数、深度学习三种方法；基于特征的方法需要定义特征集合，核函数不需要定义特征集合、在高维空间进行计算。笔者主要介绍基于深度学习的方法。

过去的几年中，很多基于深度学习的有监督关系分类被提出，大致都采用CNN、RNN、依存句法树、BERT的方法，由于这些方法大都很容易理解，笔者这里不再赘述，只选择介绍3篇比较新颖的文献进行介绍。

3-1 Matching the Blanks: Distributional Similarity for Relation Learning^[13]



这篇文献来自GoogleAI，基于BERT，共采用6种不同结构来进行实体pair的pooling，然后将pooling进行关系分类或关系相似度计算，显示(f)效果最好。

- 1. 标准输入+「CLS」输出；
- 2. 标准输入+mention pooling输出；
- 3. position embedding 输入+mention pooling输出；
- 4. entity markers输入+「CLS」输出；
- 5. entity markers输入+ mention pooling输出；
- 6. entity markers输入+ entity start 输出；

3-2 Extracting Multiple-Relations in One-Pass with Pre-Trained Transformers^[14]

Pipeline方法下的关系分类，同一个句子会有多个不同的实体对，过去的一些方法构造多个（句子，entity1，entity2）进行多次关系分类，本质上是一个multi pass问题，同一个句子会进行重复编码，耗费计算资源。

- 本文将多次关系抽取转化为one pass问题，将句子一次输入进行多个关系分类。在BERT顶层对不同的实体对进行不同的关系预测。
- 本文将还编码词和实体之间的相对距离计算Entity-Aware Self-Attention。如下图所示， $w_{d(i-j)}$ 代表实体 x_i 到token x_j 间相对距离的embedding。

Entity-Aware Self-Attention

3-3 Simultaneously Self-Attending to All Mentions for Full-Abstract Biological Relation Extraction^[15]

已赞同 498

▼

32 条评论

分享

喜欢

收藏

申请转载

...



与上篇文献^[14]类似，这篇文献的依旧采用one-pass对所有实体mention进行关系分类，同时从所有实体mention中定位关系。

不同的地方是从句子级别拓展到文档级别，同时引入NER辅助进行多任务学习，此外，实体信息在进行mention pooling才给定，而不是输入时就给出；进行关系分类时采用Bi-affine方法(sigmoid)，而不是采用Softmax。具体地：

- **Bi-affine Pairwise Scores**：采用Transformer编码，对每个token通过两个独立MLP进行三元组中的head和tail表征，然后Bi-affine通过计算每个三元组的得分：

$$(head - i, relation - l, tail - j) \rightarrow A_{ilj} = (e_i^{head} L) e_j^{tail}$$

- 采用LogSumExp计算得分： $scores(i^{head}, j^{tail}) = \log \sum_{i,j} A_{ij}$

- 计算loss时，给定E个实体对信息再行计算： $\frac{1}{E} \sum_{i=1}^E P(r_i | scores(i^{head}, j^{tail}))$

Simultaneously Self-Attending

Q4：什么是关系重叠&复杂关系问题？

- a：正常关系问题
- b：关系重叠问题，一对多。如“张学友演唱过《吻别》《在你身边》”中，存在2种关系：「张学友-歌手-吻别」和「张学友-歌手-在你身边」
- c：关系重新问题，一对实体存在多种关系。如“周杰伦作曲并演唱《七里香》”中，存在2种关系：「周杰伦-歌手-七里香」和「周杰伦-作曲-七里香」
- d：复杂关系问题，由实

已赞同 498



32 条评论

分享

喜欢

收藏

申请转载





集；

- e: 复杂关系问题，关系交叉导致。如“张学友、周杰伦分别演唱过《吻别》《七里香》”，「张学友-歌手-吻别」和「周杰伦-歌手-七里香」

Q5：联合抽取难点在哪里？联合抽取总体上有哪些方法？各有哪些缺点？

顾名思义，联合模型就是一个模型，将两个子模型统一建模。根据Q1，联合抽取可以进一步利用两个任务之间的潜在信息，以缓解错误传播的缺点（注意△只是缓解，没有从根本上解决）。

联合抽取的难点是如何加强实体模型和关系模型之间的交互，比如实体模型和关系模型的输出之间存在着一定的约束，在建模的时候考虑到此类约束将有助于联合模型的性能。

现有联合抽取模型总体上两大类^[16]：

1、共享参数的联合抽取模型

通过共享参数（共享输入特征或者内部隐层状态）实现联合，此种方法对子模型没有限制，但是由于使用独立的解码算法，导致实体模型和关系模型之间交互不强。

绝大多数文献还是基于参数共享进行联合抽取的，这类的代表文献有：

2、联合解码的联合抽取模型

为了加强实体模型和关系模型的交互，复杂的联合解码算法被提出来，比如整数线性规划等。这种情况下需要对子模型特征的丰富性以及联合解码的精确性之间做权衡^[16]：

- 一方面如果设计精确的联合解码算法，往往需要对特征进行限制，例如用条件随机场建模，使用维特比解码算法可以得到全局最优解，但是往往需要限制特征的阶数。
- 另一方面如果使用近似解码算法，比如集束搜索，在特征方面可以抽取任意阶的特征，但是解码得到的结果是不精确的。

因此，需要一个算法可以在不影响子模型特征丰富性的条件下加强子模型之间的交互。

此外，很多方法再进行实体抽取时并没有直接用到关系的信息，然而这种信息是很重要的。需要一个方法可以同时考虑一个句子中所有实体、实体与关系、关系与关系之间的交互。

Q6：介绍基于共享参数的联合抽取方法？

在联合抽取中的实体和关系抽取的解码方式与Q2中的实体抽取的解码方式基本一致，主要包括：序列标注CRF/SoftMax、指针网络、分类SoftMax、Seq2Seq等。基于共享参数的联合抽取，实体抽取loss会与关系抽取loss相加。

由于很多的相关文献实用性不高，我们只介绍其中具备代表性和易于应用的几篇文献，首先归纳如下：

6-1 依存结构树：End-to-End Relation Extraction using LSTMs on Sequences and Tree Structures^[17]

已赞同 498



32 条评论

分享

喜欢

收藏

申请转载





- **联合抽取顺序**：先抽取实体，再进行关系分类
 - **实体抽取**：采用BIOES标注，SoftMax解码；
 - **关系抽取**：针对实体抽取出的实体对，在当前句子对应的依存句法树中找到能够覆盖该实体对的最小依存句法树，并采用TreeLSTM生成该子树对应的向量表示，最后，根据子树根节点对应的TreeLSTM向量进行SoftMax关系分类。
- **存在问题**：
 - 实体抽取未使用CRF解码，没有解决标签依赖问题。
 - 关系抽取仍然会造成实体冗余，会提升错误率、增加计算复杂度
 - 使用句法依存树，只针对句子级别并且只适用于易于依存解析的语言。
 - 不能解决完整的关系重叠问题，本质上是实体重叠问题没有解决。

6-2 指针网络, Going out on a limb: Joint Extraction of Entity Mentions and Relations without Dependency Trees^[18]

网络结构图和标注框架

- **联合抽取顺序**：识别实体的同时进行关系抽取，不再采取依存树。
 - **实体抽取**：采用BIOES标注，SoftMax解码；解码时利用前一步的label embedding信息。
 - **关系抽取**：采取指针网络解码，指针网络实际上有R层（R为关系总数）。对当前实体查询在其位置前的所有实体（向前查询），并计算注意力得分：

已赞同 498



32 条评论

分享

喜欢

收藏

申请转载





- 存在问题：

- 只向前查询head实体，会存在对tail实体的遗漏；
- 在关系指针网络的gold标签中，对于实体span中每一个token平均分配1/N概率，没有充分利用实体边界信息，这会导致注意力分散。

6-3 Copy机制+seq2seq: Extracting Relational Facts by an End-to-End Neural Model with Copy Mechanism^[19]

- 联合抽取顺序：采用Seq2Seq框架，依次抽取关系、head实体、tail实体。

- Encoder编码: $o_t^E, h_t^E = f(x_t, h_{t-1}^E)$
- Decoder编码: $o_t^D, h_t^D = g(u_t, h_{t-1}^D)$
 - u_t 为decoder部分t时刻的输入, $u_t = [v_t; c_t] \cdot W^u$, 主要有两部分组成:
 - c_t 为attention vector, v_t 为前一步的copy entity 或者 relation embedding;
 - 关系预测: 将 o_t^E 直接喂入SoftMax进行;
 - head实体预测 (Copy the First Entity) :
 - 在当前解码步, 从n个token中选择一个作为实体:
 - $q_i^e = \text{selu}([o_t^D; o_i^E] \cdot w^e)$ 为每一个token的编码, 加入当前解码的输出;
 - 根据 $p^e = \text{softmax}([q^e; q^{NA}])$ 从n个token中选择最大概率的token作为实体;
 - tail实体预测 (Copy the Second Entity)
 - 与head实体预测类似, 只是需要mask上一步预测的head实体 (token)

- 存在问题:

- 只考虑token维度的实体, 丢失了多个token构成的实体, 这是一个明显bug;

6-4 多头选择机制+sigmoid: Joint entity recognition and relation extraction as a multi-head selection problem^[20]





网络结构

本篇文献应用较为广泛，与3-3的文献^[15]十分类似，只是不再提供实体信息、需要对实体进行预测。

- **联合抽取顺序**：先抽取实体，再利用实体边界信息进行关系抽取。
- **实体抽取**：采用BIOES标注，CRF解码；
- **关系抽取**：采用sigmoid进行多头选择，与文献^[15]的做法类似。
 - 对于含n个token的句子，可能构成的关系组合共有 $n \times r \times n$ 个，其中r为关系总数，即当前token会有多个头的关系组合：

- 该方法并没有像文献^[15]分别构建head和tail实体编码，而是直接通过token的编码表示进入sigmoid layer直接构建「多头选择」。
- 引入实体识别后的entity label embedding进行关系抽取，训练时采用gold label，推断时采用predict label。
- 在三元组统一解码时，需要利用实体边界信息组建三元组，因为多头选择机制只能知道token和token之间的关系，但并不知道token隶属的实体类别。
- 存在问题：
 - entity label embedding在训练和推断时存在gap，文献^[21]提出了Soft Label Embedding，并引入了BERT。
 - 鲁棒泛化问题：原作者在文献^[22]引入了对抗训练机制（如今看来，这种对抗训练机制比较简单了）

6-5 SPO问题+指针网络，Joint Extraction of Entities and Relations Based on a Novel Decomposition Strategy ^[23]



- **联合抽取顺序**：是一个spo问题，先抽取实体（主体subject，简称s），再抽取关系（关系predicate及其对应的客体object，简称po）。
- 如上图所示，主体抽取包含「Trump」和「Queens」，然后基于已抽取的主体再进行po抽取。例如对于「Trump」，其对应的关系包含「PO」-「United States」和「BI」-「Queens」；可以看出「Queens」既可以作为subject，也可以是object。

网络结构图

- **主体（s）抽取**：采用指针网络进行解码。
- **关系和客体（po）抽取**：同样采用指针网络进行解码，但事实上采用的是Q2中提到的**多层label指针网络**，即每一层是一个关系label对应的指针网络（用来抽取object）。
 - 在对当前的subject抽取对应的po时，采取多种方式加强了对当前subject的实体感知方式，如 sentence pooling、entity pooling、relative position embedding等；在对object的end pos 解码时也引入start pos的编码信息。
- **存在问题**：
 - 在训练时，subject的选择是随机的，并没有将所有subject统一进行po抽取；没有充分利用信息，可能造成信息损失，因此需要延长epoch训练。

6-6 多轮对话+强化学习：Entity-Relation Extraction as Multi-Turn Question Answering^[24]

多轮对话设计-实体关系抽取

- **联合抽取顺序**：基于人工设计的QA模板，先提取实体，再抽取关系。
 - 文献指出通常的三元组形式存在问题，并不能充分反应文本背后的结构化信息^[25]：如上图的结构化表格，TIME需要依赖Position，Position需要依赖Corp（公司）。进行传统的三元组抽取可能导致依赖关系的间断，因此这种多轮QA方式^[25]：
 - 能够很好地捕捉层级化的依赖关系。
 - 问题能够编码重要的先验关系信息，对实体/关系抽取有所帮助。
 - 问答框架是一种很自然的方法来同时提取实体和关系。
 - 将联合抽取转为一种X
- 这些实体和关系可以以

已赞同 498



32 条评论

分享

喜欢

收藏

申请转载





进行解码（与文献^[5]一脉相承，只是不再使用指针网络，而是CRF）。

- **强化学习：**
 - 笔者在前面已经指出，基于共享参数的联合学习仍然不能完全避免在推断时的误差积累，这篇文献采用强化学习机制进行优化。
 - 在多轮QA中^[25]，Action就是选择一个文本段，Policy就是选择该文本段的概率。对于Reward，使用正确抽取的三元组的数量作为奖励，使用REINFORCE算法寻找最优解。
- **存在问题：**
 - 也许针对三元组形式不能体现文本结构化信息的任务是有一定必要性的，如关系依赖问题。但对于通常的三元组任务，引入question需要对原始文本进行多次编码才能抽取实体和关系，计算复杂度较高。

6-7 片段排列：Span-Level Model for Relation Extraction^[7]

- **联合抽取顺序：**片段排列抽取实体，然后提取实体对进行关系分类；
 - 将片段排列方式生成的候选实体span，进行实体类型SoftMax分类；对于候选实体span不为None的实体span组成实体pair进行关系SoftMax分类；
 - 笔者在前文介绍实体重叠问题时，已经介绍了这种基于片段排列的方式，基于片段排列的方式^[7]，显示的提取所有可能的片段排列，由于选择的每一个片段都是独立的，因此可以直接提取span-level的特征去解决重叠实体问题。
 - 实体span的编码表示：在span范围内采取注意力机制与基于原始输入的LSTM编码进行交互。
- **存在问题：**
 - 对于含T个token的文本，理论上共有 $N = T(T + 1)/2$ 种片段排列，计算复杂度极高。如果文本过长，会产生大量的负样本，在实际中需要限制span长度并合理削减负样本。
 - 进行关系判断时，也会造成实体冗余，提高错误率。

6-8 片段排列：SpERT: Span-based Joint Entity and Relation Extraction with Transformer Pre-training^[26]

SpERT

- **联合抽取顺序：**在输出端进行片段排列进行实体分类，然后进行关系分类。
 - 与6-7^[7]类似，但采取BERT编码表示，在BERT最后输出的hidden层根据候选的实体span进行实体分类，过滤实体类型为None的片段然后进行关系分类。
 - 进行关系分类时，融合多种特征组合：包含实体span的pooling，实体span长度，实体pair之间token的pooling；
- **存在问题：**
 - 虽然缓解了片段排列的高复杂度问题，但关系分类仍有实体冗余问题。

Q7：介绍基于联合解码的联合抽取方法？

在Q6中的基于共享参数的联合抽取的方法中，并没有显式地刻画两个任务之间的交互，同样训练和推断仍然存在gap。

为了加强两个子模型之间的交互，一些联合解码算法被提出^[16]：文献^[27]提出使用整数线性规划（ILP）对实体模型和关系

模实体和关系模型，并通

已赞同 498



32 条评论

分享

喜欢

收藏

申请转载





看为一个结构化预测问题，采用结构化感知机算法，设计了全局特征，并使用集束搜索进行近似联合解码。文献^[30]提出使用全局归一化（Global Normalization）解码算法。文献^[31]针对实体关系抽取设计了一套转移系统（Transition System），从而实现联合实体关系抽取。由于篇幅限制，对上述文献感兴趣的读者可以详细参考原文。

下面笔者介绍3种易于应用的**统一实体和关系标注框架**的联合解码方法。

7-1 Joint extraction of entities and relations based on a novel tagging scheme^[32]

- **总体标注框架：**

- 统一了实体和关系标注框架，直接以关系标签进行BIOES标注。head实体序号为1，tail实体序号为2；
- 存在问题：
 - 不能关系重叠问题，比如一个实体存在于多种关系中的情况。这是一个致命的bug。

7-2 Joint Extraction of Entities and Overlapping Relations Using Position-Attentive Sequence Labeling^[33]

- **总体标注框架：**如上图所示，对于含n个token的句子，共有n个不同标注框架。也就是对于每一个位置的token都进行一次标注，无论实体还是关系都采用BIES标注。
 - 当p=5指向第5个token「Trump」时，其对应的实体为「PER」，此时p=5对应的标签实体有「United States」、「Queens」、「New York City」，分别对应关系「*President of*」、「*Born in*」、「*Born in*」。
 - 本质上将实体和关系融合为一体，共同采用BIES标注，用CRF解码。





- 实体关系提取时，对当前指向位置的实体采用**position attention 机制**进行识别对应的关系实体，该机制融合了 position-aware 和 context-aware 表示：其中 h_p 为当前指示的token位置编码， h_j 为上下文编码， h_t 为当前解码位置的编码。
- **存在问题**：对一个句子进行了n次重复编码，复杂度高， $o(n^2)$

7-3 Joint extraction of entities and relations based on a novel tagging scheme^[34]

- **总体标注框架**：这个方法来自PaddlePaddle/Research，也是百度2020关系抽取的baseline方法，同样也是统一了实体和关系的SPO标注框架。（SPO问题可参考前文的6-5）
 - 使用方法是token level 的多label分类，即每一个token对应多个label。
 - 标注框架十分巧妙，如上图示例中形成的2个spo三元组，「王雪纯-配音-晴雯」和「王雪纯-配音-红楼梦」，存在两个关系「配音-人物」和「配音-作品」，多label标签就以关系标签建立：
 - 假设一共存在R个关系，那label一共为（2*R+2个），如果是subject中的第一个token，则标记为「B-S-关系名称」；如果是object中的第一个token，则标记为「B-O-关系名称」；其余的实体token标记为「I」，不隶属于实体的token标记为「O」；
 - 如对于subject王雪纯中，「王」隶属于两个「B-S-配音-作品」和「B-S-配音-人物」；其余的「雪」「纯」用「I」来标注；
 - 如对于object红楼梦「红」隶属于「B-O-配音-作品」，其余的「楼」「梦」用「I」来标注；

已赞同 498

32 条评论

分享

喜欢

收藏

申请转载





- 如对于object晴雯中「晴」隶属于「B-O-配音-人物」；其余的「雯」用「I」来标注；
- **存在问题：**
 - 上述标注框架还是无法直接解决一些包含实体重叠的关系抽取？
 - 如：《叶圣陶散文选集》中，叶圣陶-作品-叶圣陶散文选集；
 - 上述标注框架也无法直接解决一个句子中的多重同类关系：
 - 如，‘张学友《吻别》周杰伦《菊花台》梁静茹《吻别》’等，需要加入后处理逻辑。

总结：上述统一实体和关系标注框架虽然不能完全解决关系重叠等问题，但在特定场景下，引入一些后处理规则进行约束，这种方式简单明了、易于迭代维护。

Q8：实体关系抽取的前沿技术和挑战有哪些？如何解决低资源和复杂样本下的实体关系抽取？如何应用图神经网络？

在前文中，笔者叙述了pipeline和联合抽取中的一些实体关系抽取方法，其中面临的挑战，笔者初步总结如下并给出一点建议：

1、对于pipeline方法中的NER来说：

虽然很多方法已经很普及，但更需要关注复杂场景下的**实体重叠问题**；此外，对于NER问题其实应用很广，在很多**性能敏感**的场景下，使用深度学习的方法似乎不能满足要求，这时就需要我们采取「词典+规则」的方法，例如：

- 对于医疗场景中的很多实体歧义性并不强，对上下文也不够敏感，这时构建出一个针对目标实体的词表更为有效。
- 对于通用领域中歧义性的实体，是否可以采用多种分词方式和句法分析等融合的方法去寻找实体边界呢？这都值得我们进一步尝试。

此外，应用解决NER的方法是否可以解决一些事件段落切割问题，方便我们将复杂任务进行拆解。

2、对于pipeline方法中的关系分类来说：

首要问题是怎么降低计算复杂度，关系分类时不再对句子重复编码，而是one-pass。

在低资源场景下，采取远程监督的方法确实可以自动进行语料构建，但其中针对样本噪音的降噪方法是否还有提升空间？降噪方法能否做到与模型无关，是否可以借鉴图像分类中很有效的置信学习[35]呢？

此外，预训练语言模型如此火爆，针对关系分类任务，能否在预训练阶段引入更有效的关系分类的目标呢？如前文提到的文献[13]。

3、对于联合抽取任务来说：

难点是如何加强实体模型和关系模型之间的交互，怎么对需要对子模型特征的丰富性以及联合解码的精确性之间做权衡？

此外，很多方法再进行实体抽取时并没有直接用到关系的信息，然而这种信息是很重要的。需要一个方法可以**同时考虑一个句子中所有实体、实体与关系、关系与关系之间的交互**。

引入**图神经网络**是否能够解决关系与关系之间的交互呢？由于篇幅原因，本文不再赘述。感兴趣的读者可以参考ACL2019中的系列文献[36][37][38][39]。

4、对于低资源问题和复杂样本问题来说：

在刘知远老师的《知识图谱从哪里来：实体关系抽取的现状与未来》[40]一文中，详细叙述了这方面的问题：

- 对于**少次关系学习**问题：他们提出了FewRel 2.0[41]，在原版数据集FewRel的基础上增加了以下两大挑战：领域迁移（domain adaptation）和“以上都不是”检测（none-of-the-above detection）。
- 对于**文档级别的关系抽取**

已赞同 498

32 条评论

分享

喜欢

收藏

申请转载

...

关系抽取数据集，文档级关系抽取任务要求模型具有强大的模式识别、逻辑推理、指代推理和常识推理能力^[40]。



此外，如何引入将低资源问题的解决方案引入实体关系抽取中是一个值得探讨的问题，如主动学习、迁移学习（领域自适应、跨语言问题）、元学习、半监督学习等；还有怎么解决不平衡数据下的关系抽取？一些顶会的系列文献^{[43][44][45][46][47][48]}也做了一些尝试，感兴趣的读者可以参考。

笔者注：对于NLP中的低资源问题、复杂样本问题、数据质量问题等，我们将在《高能NLP之路》专栏的下一篇文章中进行详细介绍，希望大家关注。

彩蛋：百度2020关系抽取比赛的baseline可以采取哪些方法？

除了百度官方给出的baseline^[34]，大家可以参考前文提及的^{[20][23]}。

写在最后

由于篇幅有限，并为给读者更好的阅读体验，本文删减了大量对模型内部的解读，更为细节的请阅读原文。

- 如需下载本篇文档，可以到我的github下载。
- 如有错误，请指正。
- 未经允许，不得转载。

参考

1. ^ Hierarchically-Refined Label Attention Network for Sequence Labeling <https://arxiv.org/pdf/1908.08676.pdf>
2. ^ Chinese NER Using Lattice LSTM <https://arxiv.org/pdf/1805.02023.pdf>
3. ^ ^{a b} Neural Architectures for Nested NER through Linearization
4. ^ Nested named entity recognition revisited.
5. ^ ^{a b} A Unified MRC Framework for Named Entity Recognition <https://arxiv.org/pdf/1910.11476.pdf>
6. ^ <https://zhuanlan.zhihu.com/p/89019478>
7. ^ ^{a b c d} Span-Level Model for Relation Extraction <https://www.aclweb.org/anthology/P19-1525.pdf>
8. ^ ^{a b} Distant Supervision for Relation Extraction via Piecewise Convolutional Neural Networks. EMNLP
9. ^ Selective Attention over Instances (Lin 2016)
10. ^ Relation Extraction with Multi-instance Multi-label Convolutional Neural Networks.
11. ^ Distant Supervision for Relation Extraction with Sentence-Level Attention and Entity Descriptions
12. ^ Reinforcement Learning for Relation Classification from Noisy Data
13. ^ ^{a b c} Matching the Blanks: Distributional Similarity for Relation Learning <https://arxiv.org/pdf/1906.03158.pdf>
14. ^ ^{a b} Extracting Multiple-Relations in One-Pass with Pre-Trained Transformers
15. ^ ^{a b c d} Simultaneously Self-Attending to All Mentions for Full-Abstract Biological Relation Extraction <https://www.aclweb.org/anthology/N18-1080.pdf>
16. ^ ^{a b c} 基于深度学习的联合实体关系抽取 <http://www.czsun.site/publications/thesis.pdf>
17. ^ End-to-End Relation Extraction using LSTMs on Sequences and Tree Structures <https://www.aclweb.org/anthology/P16-1105.pdf>
18. ^ Going out on a limb: Joint Extraction of Entity Mentions and Relations without Dependency Trees https://pdfs.semanticscholar.org/bbbd/45338fbd85b0bacf23918bb77107f4cfb69e.pdf?_ga=2.119149259.311990779.1584453795-1756505226.1584453795
19. ^ Extracting Relational Facts by an End-to-End Neural Model with Copy Mechanism
20. ^ ^{a b} Joint entity recognition and relation extraction as a multi-head selection problem
21. ^ BERT-Based Multi-Head Selection for Joint Entity-Relation Extraction
22. ^ Adversarial training for multi-context joint entity and relation extraction
23. ^ ^{a b} Joint Extraction of Entities and Relations Based on a Novel Decomposition Strategy
24. ^ Entity-Relation Extraction as Multi-Turn Question Answering <https://link.zhihu.com/?target=https%3A/arxiv.org/pdf/1905.05529.pdf>
25. ^ ^{a b c d} <https://zhuanlan.zhihu.com/p/65870466>
26. ^ Span-based Joint Entity and Relation Extraction with Transformer Pre-training <https://arxiv.org/pdf/1909.07755.pdf>
27. ^ Joint inference for fine-grained opinion extraction
28. ^ Investigating lstm for joint extraction of opinion entities and relations.
29. ^ Incremental joint extraction of entity mentions and relations.
30. ^ End-to-end neural relation extraction with global optimization
31. ^ Joint extraction of entities ;
32. ^ Joint extraction of entities ;

已赞同 498

32 条评论

分享

喜欢

收藏

申请转载

...

schema. <https://arxiv.org/pdf/1706.05075.pdf>

33. ^ Joint Extraction of Entities and Overlapping Relations Using Position-Attentive Sequence Labeling

34. ^ [a b https://github.com/PaddlePaddle/Research/tree/master/KG/DuIE_Baseline](https://github.com/PaddlePaddle/Research/tree/master/KG/DuIE_Baseline)

35. ^ Confident Learning: Estimating Uncertainty in Dataset Labels

36. ^ Graph Neural Networks with Generated Parameters for Relation

37. ^ GraphRel: Modeling Text as Relational Graphs for Joint Entity and Relation Extraction

38. ^ Attention Guided Graph Convolutional Networks for Relation Extraction

39. ^ Joint Type Inference on Entities and Relations via Graph Convolutional Networks

40. ^ [a b https://www.zhihu.com/search?type=content&q=%E5%85%B3%E7%B3%BB%E6%8A%BD%E5%8F%96](https://www.zhihu.com/search?type=content&q=%E5%85%B3%E7%B3%BB%E6%8A%BD%E5%8F%96)

41. ^ FewRel 2.0: Towards More Challenging Few-Shot Relation Classification

42. ^ DocRED: A Large-Scale Document-Level Relation Extraction Dataset

43. ^ Knowledge-Augmented Language Model and its Application to Unsupervised Named-Entity Recognition

44. ^ Description-Based Zero-shot Fine-Grained Entity Typing

45. ^ Zero-Shot Entity Linking by Reading Entity Descriptions

46. ^ Multi-Level Matching and Aggregation Network for Few-Shot Relation Classification

47. ^ Exploiting Entity BIOES Tag Embeddings and Multi-task Learning for Relation Extraction with Imbalanced Data

48. ^ Massively Multilingual Transfer for NER

编辑于 07-01


「分享学习」

赞赏

还没有人赞赏，快来当第一个赞赏的人吧！

[自然语言处理](#) [深度学习（Deep Learning）](#) [机器学习](#)


文章被以下专栏收录



高能NLP

高能量的NLP分享！不追求数量，质量第一！


已关注



机器学习算法与自然语言处理

公号[机器学习算法与自然语言处理] 微信号yizhennotes

已关注



夕小瑶的卖萌屋

关注同名微信公众号解锁更多NLP、搜索和推荐干货

关注专栏

推荐阅读



知识抽取-实体及关系抽取

徐阿衡



燃烧吧，分类~

LP分类任务的11个关键问

如何解决NLP分类任务的11个关键问题：类别不平衡&低耗时...

JayLo... 发表于高能NLP



SpanBert：对 Bert 预训练的一次深度探索

Andy ... 发表于NLPCA...



从文

李军


32 条评论

已赞同 498 32 条评论 分享 喜欢 收藏 申请转载 ...



写下你的评论...




精选评论 (2)



 **Junebaba**03-24

请教一下，关于6-8 spert，感觉并没有解决了6-7中论文的span复杂度高的问题呀？6-7的论文原先是如果句子长L，那么有 $L(L+1)/2$ 个可能的span，但是实验的时候其实有limit 每个span的长度的，而spert其实也是通过limit span的长度来达到的不是吗？：)


 赞  查看回复

 **JayLou娄杰 (作者)** 回复 Junebaba03-24


更正：实体span的编码表示：在span范围内采取注意力机制与基于原始输入的LSTM编码进行交互。然后所有的实体span表示并行的喂入SoftMax进行实体分类。

 赞  查看回复

评论 (32)

 **壮哉我贾诩文和**03-19

好评！最近做的一个知识图谱的内容正好用到文中的内容，但是没有做比较深的理解，学习一下！

 1

 **Maybewuss**03-20

tql!

 赞

 **Yuflo**03-21

赞

 赞

 **zyc**03-21

赞赞赞

 赞


 **哈哈嘿嘿**03-21

关注了

 赞

 **小黑布莱克**03-23


多层指针网络怎么构建

 赞


 **JayLou娄杰 (作者)** 回复 小黑布莱克03-23


您好！每一个实体label是一个单层指针网络（start和end的），解码时相对复杂一些。具体可参考《Joint Extraction of Entities and Relations Based on a Novel Decomposition Strategy》。

 1


 **小黑布莱克** 回复 JayLou娄杰 (作者)03-24


谢谢，非常棒的一篇分享。值得好好学习！

 1

 **Junebaba**03-24

请教一下，关于6-8 spert，感觉并没有解决了6-7中论文的span复杂度高的问题呀？6-7的论文原先是如果句子长L，那么有 $L(L+1)/2$ 个可能的span，但是实验的时候其实有limit 每个span的长度的，而spert其实也是通过limit span的长度来达到的不是吗？：)

 赞

 **JayLou娄杰 (作者)**

已赞同 498  32 条评论  分享  喜欢  收藏  申请转载 ...

你好！6-7和6-8在实体span排列方式上有区别的，6-7会显式的将所有不同的span排列输入编码器；而6-8会在bert的输出层，对bert的输出隐状态进行span排列。6-7没有端到端处理，复杂度会高一些。6-8没有对文本序列重复编码，直接在输出隐状态进行span排列。6-8也会有 $L(L+1)/2$ 个可能的span。

👍 赞



JayLou姜杰 (作者) 回复 Junebaba

03-24

6-8主要改进6-7[7]中在输入端进行片段排列的高复杂度问题，在输出端排列计算复杂度会低一些。6-8一次编码，而6-7会间所有可能的span排列方式都依次喂入编码器进行分类，这样的复杂度很高。而两者产生的span排列个数是相同的。

👍 赞

查看全部 7 条回复



Jacky Nix

03-24

真的详细，信息量真的大，我都看两天了还没看完



👍 赞



Jacky Nix

03-27

大约花了一周的时间，一篇一篇对照着论文看完了，笔者你真的总结的很好。我觉得这些方法里，Joint entity recognition and relation extraction as a multi-head selection problem和SpERT: Span-based Joint Entity and Relation Extraction with Transformer Pre-training确实相对来说优雅一些。

👍 3



JayLou姜杰 (作者) 回复 Jacky Nix

03-27

感谢你的评价！你的直觉很准确，我只做过3个联合抽取的实验，除了你说的这2篇，还有Joint Extraction of Entities and Relations Based on a Novel Decomposition Strategy。



👍 2



Jacky Nix 回复 Jacky Nix

03-29

请问您spert那篇文章，您有没有在中文语料上实验过？

👍 赞

展开其他 1 条回复



atticus

04-09

很全面，这两天一直在发愁直接用softmax怎么定义合适的损失，车上大概读了一遍，不是很明白，回去再细读！感谢！

👍 赞



扎克斯

04-19

太赞了，这个专栏简直是我的宝藏～

👍 赞



高宇轩

04-22

你好，请问中文实体关系抽取的训练数据有可以分享的吗？一直没找到合适的

👍 赞



建国 回复 高宇轩

04-27

aistudio.baidu.com/aistudio

赞



高宇轩 回复 建国

05-25

多谢

赞

展开其他 1 条回复



孤绿

05-10

您好，请教下“实体关系抽取”的具体定义是？在哪篇论文或者百词词条上有提及到这样的定义？

赞



huntingFor

05-21

这是什么神仙



赞



sjfjfd

08-10

请问multi-head selection那篇论文可以抽取嵌套实体和他们的关系吗？谢谢！！

赞



榕松

09-04

请问百度2020关系抽取比赛关闭了，数据集还能哪里下载？

赞



胡博钦

09-10

必须要赞一下