Global-to-Local Neural Networks for Document-Level Relation Extraction

Difeng Wang[†] **Wei Hu**[†,‡,*] **Ermei Cao**[†] **Weijian Sun**[§]
[†] State Key Laboratory for Novel Software Technology, Nanjing University, China
[‡] National Institute of Healthcare Data Science, Nanjing University, China
[§] Huawei Technologies Co., Ltd.
{dfwang,emcao}.nju@gmail.com, whu@nju.edu.cn, sunweijian@huawei.com

收录会议：EMNLP2020

动机：

（1）一篇文章有很多实体；

（2）一个实体有很多 mention。

所以篇章级模型需要建模多实体之间复杂的交互并综合多 mention 的上下文信息，以更好地对实体进行表示。

方法贡献：

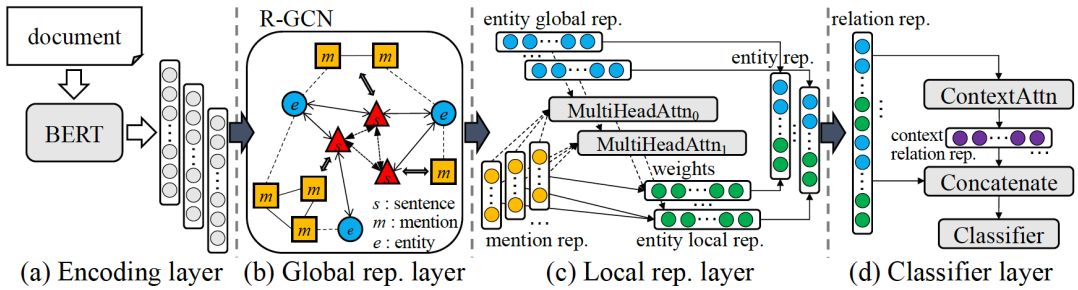

Figure 2: Architecture of the proposed model.

解决了三个挑战：

（1）如何建模一个篇章中的复杂语义？

使用 BERT 捕获语义特征和常识，构建基于启发式规则的异质图，来建模篇章中所有 mention、实体和句子间的复杂交互。

包含了三种节点和五种边。和 EoG 基本类似，与 GAIN 相比缺少了同一个实体 mention 之间的全连接，缺少了 document node，多了一个

sentence node 和 entity node，而且他是将 entity 和 mention 放在一个图中，而 GAIN 是分两个图分别对 mention 和 entity 编码。

- *Mention nodes,* which model different mentions of entities in $\mathcal{D}$. The representation of a mention node m_i is defined by averaging the representations of contained words. To distinguish node types, we concatenate a node type representation $\mathbf{t}_m \in \mathbb{R}^{d_t}$. Thus, the representation of m_i is $\mathbf{n}_{m_i} = [\text{avg}_{w_j \in m_i}(\mathbf{h}_j); \mathbf{t}_m]$, where $[;]$ is the concatenation operator.
- *Entity nodes,* which represent entities in $\mathcal{D}$. The representation of an entity node e_i is defined by averaging the representations of the mention nodes to which they refer, together with a node type representation $\mathbf{t}_e \in \mathbb{R}^{d_t}$. Therefore, the representation of e_i is $\mathbf{n}_{e_i} = [\text{avg}_{m_j \in e_i}(\mathbf{n}_{m_j}); \mathbf{t}_e]$.
- *Sentence nodes,* which encode sentences in $\mathcal{D}$. Similar to mention nodes, the representation of a sentence node s_i is formalized as $\mathbf{n}_{s_i} = [\text{avg}_{w_j \in s_i}(\mathbf{h}_j); \mathbf{t}_s]$, where $\mathbf{t}_s \in \mathbb{R}^{d_t}$.

- *Mention-mention edges.* We add an edge for any two mention nodes in the same sentence.
- *Mention-entity edges.* We add an edge between a mention node and an entity node if the mention refers to the entity.
- *Mention-sentence edges.* We add an edge between a mention node and a sentence node if the mention appears in the sentence.
- *Entity-sentence edges.* We create an edge between an entity node and a sentence node if at least one mention of the entity appears in the sentence.
- *Sentence-sentence edges.* We connect all sentence nodes to model the non-sequential information (i.e., break the sentence order).

（2）如何高效地学习实体的表示？

设计了一个 Global-to-Local 的架构，在异质图上使用 R-GCN 来编码实体粗粒度的信息，即学习实体的全局表示；使用 multi-head attention 来聚合某个实体的多 mention 的细粒度的信息，即学习实体的局部表示。

Global 模块，使用 RGCN 来聚合节点的信息，得到实体的 global 的表示。

Finally, we employ an L-layer stacked R-GCN (Schlichtkrull et al., 2018) to convolute the global heterogeneous graph. Different from GCN, R-GCN considers various types of edges and can better model multi-relational graphs. Specifically, its node forward-pass update for the $(l+1)^{\text{th}}$ layer is defined as follows:

$$\mathbf{n}_i^{l+1} = \sigma\Big(\sum_{x\in\mathcal{X}}\sum_{j\in\mathcal{N}_i^x}\frac{1}{|\mathcal{N}_i^x|}\mathbf{W}_x^l\mathbf{n}_j^l + \mathbf{W}_0^l\mathbf{n}_i^l\Big), \quad (2)$$

在 local 模块，对于一个特定的实体，使用 multi-head attention 来聚合他所有 mention 的信息。

其中，“Local”可以这样理解：

（1）从 encoding 层聚合原始的 mention 信息；

（2）一个实体在不同实体对中有不同的表示。

$$\text{MHead}(\mathcal{Q},\mathcal{K},\mathcal{V}) = [\text{head}_1;\ldots;\text{head}_z]\mathbf{W}^{\text{out}}, \quad (3)$$

$$\text{head}_i = \text{softmax}\Big(\frac{\mathcal{Q}\mathbf{W}_i^{\mathcal{Q}}(\mathcal{K}\mathbf{W}_i^{\mathcal{K}})'}{\sqrt{d_v}}\Big)\mathcal{V}\mathbf{W}_i^{\mathcal{V}}, \quad (4)$$

where $\mathbf{W}^{\text{out}} \in \mathbb{R}^{d_n\times d_n}$ and $\mathbf{W}_i^{\mathcal{Q}}, \mathbf{W}_i^{\mathcal{K}}, \mathbf{W}_i^{\mathcal{V}} \in \mathbb{R}^{d_n\times d_v}$ are trainable parameter matrices. z is the number of heads satisfying that $z\times d_v = d_n$.

$$\begin{aligned}\mathbf{e}_a^{\text{loc}} &= \text{LN}\big(\text{MHead}_0(\mathbf{e}_b^{\text{glo}}, \{\mathbf{n}_{s_i}\}_{s_i\in\mathcal{S}_a}, \{\mathbf{n}_{m_j}\}_{m_j\in\mathcal{M}_a})\big),\\ \mathbf{e}_b^{\text{loc}} &= \text{LN}\big(\text{MHead}_1(\mathbf{e}_a^{\text{glo}}, \{\mathbf{n}_{s_i}\}_{s_i\in\mathcal{S}_b}, \{\mathbf{n}_{m_j}\}_{m_j\in\mathcal{M}_b})\big),\end{aligned} \quad (5)$$

为了保证 K 和 V 的数量是一致的，如果多个 mention 出现在一个句子中，那么一个句子在 K 中就会出现多次？

（3）如何利用其它关系的影响？

（这里刚开始没有读懂，但是实际上讲的是这么回事儿：每篇文章都有它的主题(topic)，当 topic 定下来之后，对应的关系范围也就定了下来，所以本文想针对每个实体对得到一个与主题相关的关系的表示。）

使用 self-attention 来学习每个实体对的上下文关系表示，

topics are not. Thus, we use self-attention (Sorokin and Gurevych, 2017) to capture *context relation representations*, which reveal the topic information of the document:

$$\mathbf{o}_c = \sum_{i=0}^{p} \theta_i \mathbf{o}_i = \sum_{i=0}^{p} \frac{\exp(\mathbf{o}_i \mathbf{W} \mathbf{o}'_r)}{\sum_{j=0}^{p} \exp(\mathbf{o}_j \mathbf{W} \mathbf{o}'_r)} \mathbf{o}_i, \quad (7)$$

where $\mathbf{W} \in \mathbb{R}^{d_r \times d_r}$ is a trainable parameter matrix. d_r is the dimension of target relation representations. $\mathbf{o}_i$ ($\mathbf{o}_j$) is the relation representation of the i^{th} (j^{th}) entity pair. θ_i is the attention weight for $\mathbf{o}_i$. p is the number of entity pairs.

实验结果

问题：

1.这篇指出 EoG（Christopoulouet al. 2019）不加区分的整合全篇中的各种信息，因此不相关的信息也会被聚合进来，变成噪声，会损害模型的预测准确率。但是实际上全篇都没有给出依据，而且本文的构图和 EoG 是类似的，也没有说怎么去除的噪声。

2.这篇文章在 related work 中 Survey 到了 CorefBERT 和 LSR，但没有和他们对比实验结果。可能因为是同期就不用比吧，但是感觉太刻意了，还不如不写 CorefBERT 和 LSR

3. 还是把 document 看做是一个句子来编码，这样会导致在编码的时候无法解决 long-distance 编码的问题。

4. mention node 和 sentence node 过了 R-GCN 之后的表示没有用上

5. 如果两个实体对有 mention 共现在一个句子中，那么他们的 mention 对各自实体的权重就更大？这里只是直观的假设，并没有看到实验结果给出来的验证。

6.实验结果没有 GloVe 的版本，Train+Dev 的 setting 是为了应付数据集小的，比如 CDR 就可以用，但用在 DocRED 上并不合适

8. EoG 的复现和 LSR 的 EoG 的结果差的有点多

9. Wang 工作的缺点分析的太多了

10.用 CDR 的例子做 case study 不如 DocRED 的好

可以借鉴的点：

1. related work 很多非 ACL 系的 paper，可以阅读一下，引入这些也体现了综述的专业性。
2. 提出了 QA-based RE 的一些局限性，感觉可以借鉴他们的观点对 QA-based RE 进行改进，或者移植到 doc-level 上

Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. 2017. Zero-shot relation extraction via reading comprehension. In CoNLL, pages 333–342, Vancouver, Canada. ACL.
Lin Qiu, Hao Zhou, Yanru Qu, Weinan Zhang, Suoheng Li, Shu Rong, Dongyu Ru,

Lihua Qian, Kewei Tu, and Yong Yu. 2018. QA4IE: A question answering based framework for information extraction. In ISWC, pages 198–216, Monterey, CA, USA. Springer.
Shanshan Liu, Xin Zhang, Sheng Zhang, Hui Wang, and Weiming Zhang. 2019. Neural machine reading comprehension: Methods and trends. Applied Sciences, 9(18):3698.

感觉是个可以探究的点。

A. 设计问题的模板需要领域知识；

B. 在没有答案或者多答案的时候会表现的很差？但是没有经过验证啊，DocRED是个比较好的验证的数据集，可以拿来做做实验。

3. Error analysis的分析可以借鉴，

4. Appendix加入Notation的想法可以借鉴

5. 去掉组件在case上的效果，挺有说服力的