关系抽取调研报告

姓名：卜梦煜　　学号：1120192419　　班级：07111905

【调研论文】

1. Fu T J， Ma W Y . GraphRel: Modeling Text as Relational Graphs for Joint Entity and Relation Extraction[C]// ACL. 2019.

论文链接：https://aclanthology.org/P19-1136.pdf

2. Wei Z P，Su J L，Wang Y，Tian Y，Chang Y . A Novel Cascade Binary Tagging Framework for Relational Triple Extraction[C]// ACL. 2020.

论文链接：https://arxiv.org/pdf/1909.03227.pdf

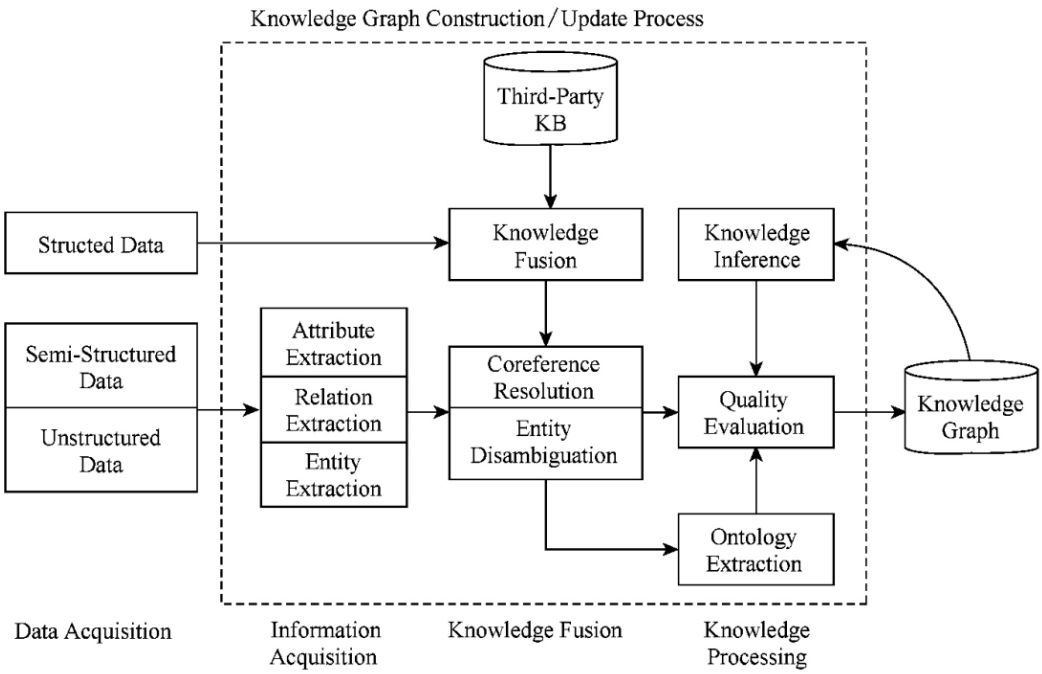
3. Huang Q， Zhu S， Feng Y，et al. Three Sentences Are All You Need: Local Path Enhanced Document Relation Extraction[C]// ACL. 2021.

论文链接：https://arxiv.org/pdf/2106.01793.pdf

第 1 章 文献综述

1.1 任务描述

知识图谱是基于知识的、结构化的语义知识库，本质上是一种大型的语义网络。知识图谱通常以如下形式定义：(E,R,O)，其中，E 为实体集合，R 为关系集合，O 为观察到事实的三元组集合。其中，每个三元组以关系三元组(h,r,t)形式给出，h 为头实体，r 为关系，t 为尾实体，即以实体为点，关系为有向边，以三元组为基本结构的构建网状知识图谱。知识图谱构建的基本流程如下：[1]



关系抽取是知识图谱构建中最基础的问题之一，即实体间关系的提取与表示。关系抽取任务目标是，对于实体抽取后得到的离散实体，从原始文本中提取出实体之间的关联关系，从而建立相互关联的网状知识结构。

关系抽取是知识图谱构建任务中，信息抽取任务的子任务之一。作为知识图谱最基本问题之一，解决好关系抽取任务对后续知识融合、知识加工、知识更新任务有着重要意义。

1.2 研究现状

从发展历史来看，关系抽取的主要方法有几个发展阶段：人工构造语法和语义规则的模式匹配方法、统计机器学习方法、基于特征向量或核函数的有监督学习方法、基于语言模型的半监督和无监督方法、面向开放域自动生成关系类别的方法、面向开放域和封闭领域相结合的方法等。

从抽取范围来看，关系抽取任务主要分为句子级别的关系抽取与文档级别的关系抽取。前者难度低于后者，研究也相对更成熟。

从关系打分函数来看，关系的打分函数构造主要分为基于距离的打分函数和基于语义理解的打分函数。基于距离的方法主要有 Trans 系列方法[2]和变换系列方法，如 TransE、TransH、TransR、TransD、RotatE 等；基于语义理解的打分函数主要有神经网络系列方法，如 Transformer[3]、图神经网络等。

当前关系抽取面临的主要问题是，如何准确挖掘多个实体间关系，以及如何

挖掘实体间的隐含关系。这两个问题的本质是对语料内容的深层次理解与提取。

1.3 调研论文简述

GraphRel: Modeling Text as Relational Graphs for Joint Entity and Relation Extraction.该论文提出了 GraphRel 模型，主要思想是利用加权图卷积网络联合学习命名实体和关系的端到端关系抽取模型，以更好地提取实体间的隐含特征和复杂关系。

A Novel Cascade Binary Tagging Framework for Relational Triple Extraction.该论文提出了 CASREL 框架，主要思路是采用全新的视角代替以往分类的视角，将关系 r 建模为 h 到 t 的映射函数，即尾实体分类器$\mathrm{f_r}(h) \to t$，而不是以往的关系分类器$\mathrm{f(h,t)} \to \mathrm{r}$，以更好地表示实体间复杂关系。

Three Sentences Are All You Need Local Path Enhanced Document Relation Extraction.该论文提出了文档级关系抽取的新方法，抽取少数关键句子抽取实体关系，而不是先提取整个文档语义再抽取实体间关系。

第 2 章 GraphRel Modeling Text as Relational Graphs for Joint Entity and Relation Extraction

2.1 论文方法

调研论文提出了一种用于实体识别和关系提取的神经端到端联合的模型 GraphRel，主要内容分为两部分：第一，利用 Bi-LSTM 自主提取每个单词的序列特征，利用 Bi-GCN 提取每个单词的区域特征，从而能够自主提取单词的隐藏特征，进而得到每个实体对的关系；第二，在第一阶段获得实体对关系的基础上，建立关系加权 GCN，通过图上关系学习，得到实体与关系之间的相互作用和可能存在的重叠关系。

2.1.1 第一阶段

2.1.1.1 Bi-LSTM

首先使用 LSTM 作为双向 RNN 的单元，对于每个词，将词嵌入表示与词性嵌入组合作为初始输入特征：

$$\mathrm{h}_u^0 = Word(u) \oplus POS(u)$$

其中h_u^0表示单词 u 的初始特征，$Word(u)$表示单词 u 的词嵌入表示，由已有的预训练结果得到，$POS(u)$嵌入表示单词 u 的词性，是随机初始化的，需要通过 GraphRel 训练得到。

2.1.1.2 Bi-GCN

输入的原始句子是一个词序列，没有固定的图结构，因此使用依赖解析器（denpendency parse）为输入句子创建依赖树。使用依赖树作为输入句的邻接矩阵，进而使用 GCN 方法提取区域依赖特征。为同时考虑单词的输入、输出特性不同，使用双向 GCN，表达式如下：

$$\vec{h}_u^{l+1} = Relu(\sum_{v\in\vec{N}(u)}(\vec{W}^l h_v^l + \vec{b}^l))$$

$$\overleftarrow{h}_u^{l+1} = Relu(\sum_{v\in\overleftarrow{N}(u)}(\overleftarrow{W}^l h_v^l + \overleftarrow{b}^l))$$

$$h_u^{l+1} = \vec{h}_u^{l+1} \oplus \overleftarrow{h}_u^{l+1}$$

其中h_u^l表示单词 u 在第 l 个隐藏层的隐藏特征，$\vec{N}(u)$表示输出单词 u 的单词，$\overleftarrow{N}(u)$表示输入单词 u 的单词。$\vec{W}$、$\vec{b}$、$\overleftarrow{W}$、$\overleftarrow{b}$都是需要学习的卷积权值，将输入单词 u 的特征$\overleftarrow{h}_u^l$和输出单词 u 的特征$\vec{h}_u^l$连接，作为单词的最终特征h_u^l。

2.1.1.3 实体抽取与关系抽取

对实体预测，利用 Bi-LSTM 与 Bi-GCN 提取出的单词特征，对单词实体预测。只需用一层 LSTM 模型完成，其中损失函数定义为分类损失$eloss_{1p}$。

对关系抽取，利用 Bi-LSTM 与 Bi-GCN 提取出的单词特征，进行实体间关系抽取。需要删掉所有依赖边，从而可以对所有词对做预测。将词w_1、w_2关系趋势定义为$S(w_1, r, w_2)$，对每个关系 r，需要学习权重矩阵W_r^1、W_r^2、W_r^3，S 定义如下：

$$S(w_1, r, w_2) = W_r^3 Relu(W_r^1 h_{w1} \oplus W_r^2 h_{w2})$$

通过对所有单词对(w_1, w_2)计算关系倾向得分，注意，无关系表示为$S(w_1, null, w_2)$。再对计算出的单词对(w_1, w_2)的所有 S 应用 Softmax，得到$P_r(w_1, w_2)$，表示单词对(w_1, w_2)为关系 r 的概率。

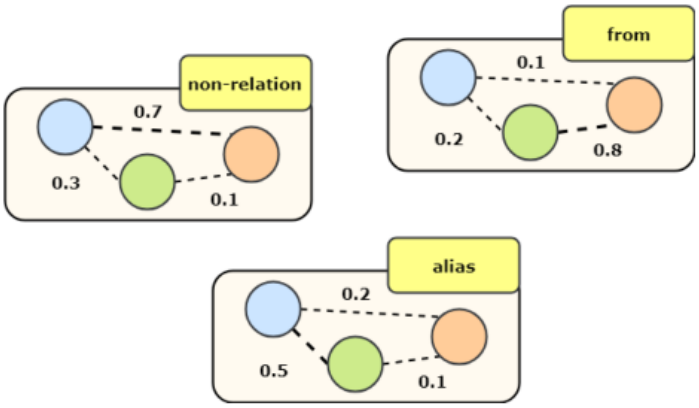
关系抽取时的损失函数定义为关系损失$rloss_{1p}$，用$P_r(w_1, w_2)$计算得到。

2.1.2 第二阶段

第一阶段解决了实体提取和实体间联系的预测问题和文本词对之间隐含特

征的问题，未解决命名实体与关系之间的联系和文本词对之间重叠关系的问题。因此，调研论文在第二阶段提出了全新的关系加权 GCN，以解决这个问题。

经过第一阶段，可得到所有单词对在所有关系上预测的概率$P_r(w_1,w_2)$，以$< w_1,w_2 >$为边，$P_r(w_1,w_2)$为边权，对每个关系 r 构建完整的关系加权图。



对每个关系图，采用 Bi-GCN，考虑不同影响程度和聚合作为综合词特征，公式如下：

$$h_u^{l+1} = Relu\left(\sum_{v\in V}\sum_{r\in R} P_r(u,v)\times(W_r^l h_v^l + b_r^l)\right) + h_u^l$$

其中$P_r(u,v)$表示边权值，W_r和b_r表示关系 r 下的 GCN 权值，需要训练得到。

注意到第二阶段的 Bi-GCN 也考虑了传入、传出的情况，且采用关系加权传播，从而为每个词提取了更加充分的特征。

利用第二阶段新训练的特征h_u^l，再次进行实体预测和关系抽取，损失函数与第一阶段类似，为分类损失$eloss_{2p}$和关系损失$rloss_{2p}$。

2.2 实验结果

GraphRel 中定义了两种损失：实体损失和关系损失，并依次定义了总损失函数：

$$loss_{all} = (eloss_{1p} + rloss_{1p}) + \alpha(eloss_{2p} + rloss_{2p})$$

其中 α 为第一阶段和第二阶段的损失之间的权重。训练时最小化总损失，并以端到端的方式训练整个 GraphRel。

调研论文将提出的 GraphRel 与两个基线 NovelTagging、MultiDecoder 的结果做了对比。NovelTagging 是一种序列标记，可以预测每个句子中单词的实体类和关系类。MultiDecoder 认为关系是一种端对端的问题，并使用动态解码器

提取关系三元组。

评估指标方面，三者采用相同的 F1-Score 评分标准，当且仅当实体对与关系均正确时，预测的是三元组是正确的。

数据集方面，调研论文在 NYT 和 WebNLG 数据集上做评估。

Method	NYT			WebNLG		
	Precision	Recall	F1	Precision	Recall	F1
NovelTagging	62.4%	31.7%	42.0%	**52.5%**	19.3%	28.3%
OneDecoder	59.4%	53.1%	56.0%	32.2%	28.9%	30.5%
MultiDecoder	61.0%	56.6%	58.7%	37.7%	36.4%	37.1%
GraphRel_{1p}	62.9%	57.3%	60.0%	42.3%	39.2%	40.7%
GraphRel_{2p}	**63.9%**	**60.0%**	**61.9%**	44.7%	**41.1%**	**42.9%**

NovelTagging 和 MultiDecoder 都使用顺序结构。由于 NovelTagging 假设一个实体属于单一的关系，所以准确率很高，但查全率很低，且无法解决实体对之间的重叠关系问题。MultiDecoder 使用动态解码器来生成关系三元组，由于 RNN 展开的固有限制，它能产生的三元组数量有限。GraphRel 可以解决这两个问题，因此认为 GraphRel 是最平衡的方法，保持高准确率和召回率，产生更高的 F1 分数。

2.3 方法优点

我认为，本调研论文提出的 GraphRel 模型有以下优点：

（1）考虑了线性结构和依赖结构，从而能够有效地学习单词对之间的隐藏特征。

（2)对实体和关系进行了端对端的联合建模,同时考虑所有单词进行预测，最大限度地保留了单词之间的关系。

（3）仔细考虑了实体与关系之间的相互作用，能够较好地预测实体之间可能存在的重叠关系。

第 3 章 A Novel Cascade Binary Tagging Framework for Relational Triple Extraction

作为知识图谱中知识抽取任务的子任务，关系抽取任务往往不是单独出现的。

目前主流的两种做法，第一种是先进行实体抽取，再进行关系抽取；第二种是将实体抽取、关系抽取联合进行，即实体-关系联合抽取。

根据实体间关系的复杂程度，关系可分为三类：普通关系，即关系之间没有重叠部分；实体对重叠 EPO，即关系两端的实体相同，但两个实体有多种关系；单一实体重叠 SEO，即关系两端有一个实体共享。

第一种流水线方法的缺点是，错误会积累，即一旦实体识别错误，那么关系构建就必然会错误，且离散地抽取实体会导致对句意提取不充分，对 EPO、SEO 的提取效果并不理想。因此，目前的研究往往选择二者联合学习的方式。本调研论文采用的就是第二种方式。

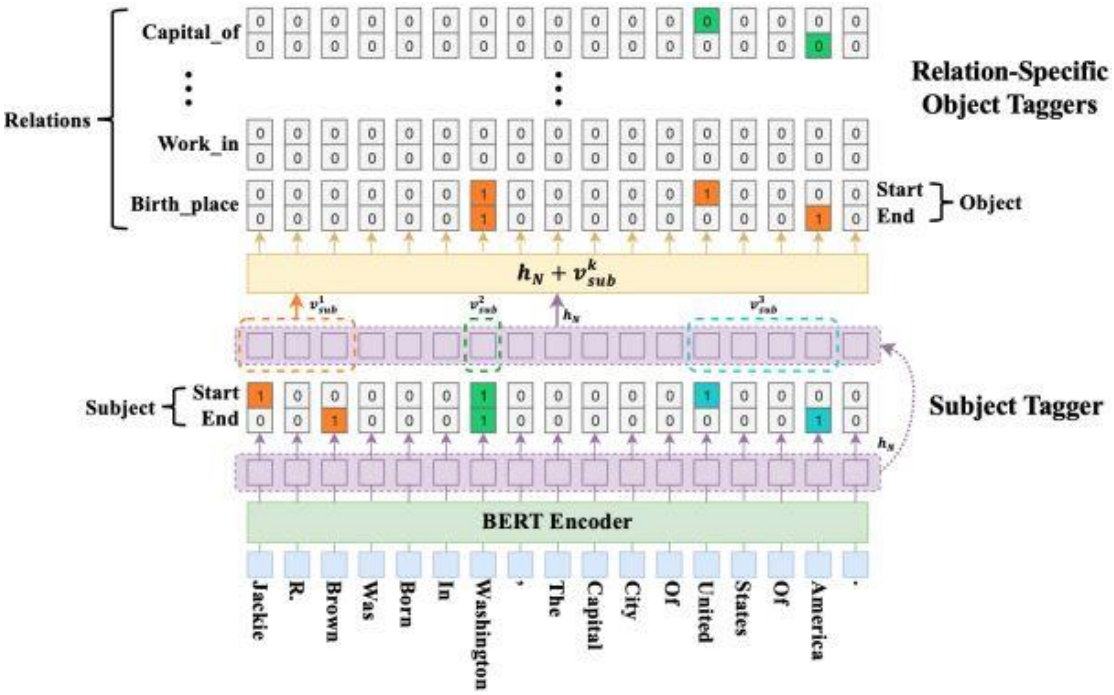
3.1 论文方法

调研论文以解决三元组重叠问题为目标，提出了新的 CASREL 框架，致力于更好地提取复杂关系，如一对多、多对一、多对多的关系，以及实体间多重关系。

该框架核心思想是将关系建模为头实体到尾实体的映射函数，即$\mathrm{f_r}(h) \to t$，而不是对实体对的标签。具体来说，若将关系视为实体对的标签，则学习过程相当于训练关系的分类器，即$\mathrm{f(h,t)} \to \mathrm{r}$，这种方法的坏处是显而易见的，比如实体对确定的条件下，分类出的关系会是唯一的，这显然无法表示 EPO 关系。而对于训练关系特定的尾实体标注器$\mathrm{f_r}(h) \to t$，每个标注器都将在给定关系和尾实体的条件下识别出所有可能的尾实体。

在 CASREL 框架下，关系三元组抽取问题被分解为两个过程：识别语料中所有可能的头实体，然后针对每个头实体，使用每种关系特定的标注器同时识别出所有可能的关系和对应的尾实体。

CASREL 模型分三个模块：BERT 编码层模块、主体标记模块、特定关系下客体标记模块，框架流程图如下：



3.1.1 BERT Encoder 层

BERT 编码层模块称为 Encoder 层，主要完成的是对句子中的词进行编码，以训练出包含词义的词嵌入表示。根据调研论文，Encoder 层可替换为其他的编码结构，如 LSTM 等。编码层原理如下：

$$h_0 = SW_s + W_p$$

$$h_\alpha = Trans(h_{\alpha-1}), \alpha \in [1, N]$$

其中 S 为输入词的独热编码表示，s 为 embedding 表，p 为表示输入词位置的位置向量，Trans 为一个 Transformer 的 Encoder 层。

3.1.2 Cascade Decoder 层

主体标记模块与特定关系下客体标记模块级联形成 Cascade Decoder 层，主要完成的是识别语料中所有可能的头实体，以及针对每个头实体，使用每种关系特定的标注器同时识别出所有可能的关系和对应的尾实体。

主体识别模块对 BERT Encoder 层输入的词嵌入表示解码，构建二分类器预测出实体的头尾索引位置，由此确定实体边界，进而识别出实体。

特定关系下客体标记模块同时识别每个头实体的关系和尾实体。解码时不仅考虑了 BERT 编码向量，还考虑了识别出的头实体特征。解码公式如下：

$$p_i^{start\_o} = \sigma(W_{start}^r(x_i + v_{sub}^k) + b_{start}^r)$$

$$p_i^{end\_o} = \sigma(W_{end}^r(x_i + v_{sub}^k) + b_{end}^r)$$

对每个关系 r 都用$\mathbf{W}^{r}$预测一次，若识别出相应的尾实体 t，则认为存在三元组(h,r,t)。

3.2 实验结果

调研论文在 NYT 和 WebNLG 两个数据集上做验证。

对编码层，分别采用随机初始化参数的 BERT 编码器、LSTM 编码器、预训练 BERT 编码器三种测试。测试结果如下：

Method	NYT			WebNLG		
	Prec.	*Rec.*	*F1*	*Prec.*	*Rec.*	*F1*
NovelTagging (Zheng et al., 2017)	62.4	31.7	42.0	52.5	19.3	28.3
CopyR$_{OneDecoder}$ (Zeng et al., 2018)	59.4	53.1	56.0	32.2	28.9	30.5
CopyR$_{MultiDecoder}$ (Zeng et al., 2018)	61.0	56.6	58.7	37.7	36.4	37.1
GraphRel$_{1p}$ (Fu et al., 2019)	62.9	57.3	60.0	42.3	39.2	40.7
GraphRel$_{2p}$ (Fu et al., 2019)	63.9	60.0	61.9	44.7	41.1	42.9
CopyR$_{RL}$ (Zeng et al., 2019)	77.9	67.2	72.1	63.3	59.9	61.6
CopyR$^{*}_{RL}$	72.8	69.4	71.1	60.9	61.1	61.0
CASREL$_{random}$	81.5	75.7	78.5	84.7	79.5	82.0
CASREL$_{LSTM}$	84.2	83.0	83.6	86.9	80.6	83.7
CASREL	**89.7**	**89.5**	**89.6**	**93.4**	**90.1**	**91.8**

由实验结果可知：

（1）CASREL 框架性能明显好于其他模型。

（2）采用预训练 BERT 做编码器的效果明显好于随机初始化参数 BERT 编码器和 LSTM 编码器。

3.3 方法优点

我认为，本调研论文提出的 CASREL 框架有以下优点：

（1）采用全新的视角看待关系三元组，将关系建模成头实体到尾实体的函数，训练时学习关系特定的尾实体标注器。

（2）能够较好地处理复杂关系，包括实体间一对多、多对一、多对多关系，实体间多种关系等。

第 4 章 Three Sentences Are All You Need Local Path Enhanced Document Relation Extraction

关系抽取任务从抽取范围来看，分为文档级别的关系抽取和句子级别的关系抽取。由于文档级别的关系抽取中，使用不止一个句子描述实体之间的关系，因此文档级别的关系抽取难度要高于句子级别的关系抽取任务。

由于每次对全部文本提取内容不仅浪费存储空间，而且难以抽取出其中两个词的词义和关系，因此需要缩小预料使用范围。本调研论文在 DocRED、CDR 和 GDA 文档级关系抽取语料上统计了实体关系支撑证据的句子条数，结果显示，95%的实例需要不超过 3 句支撑语句，87%的实例需要不超过 2 句支撑语句。因此，筛选出关键的几个句子即可完成文档级别的抽取任务。

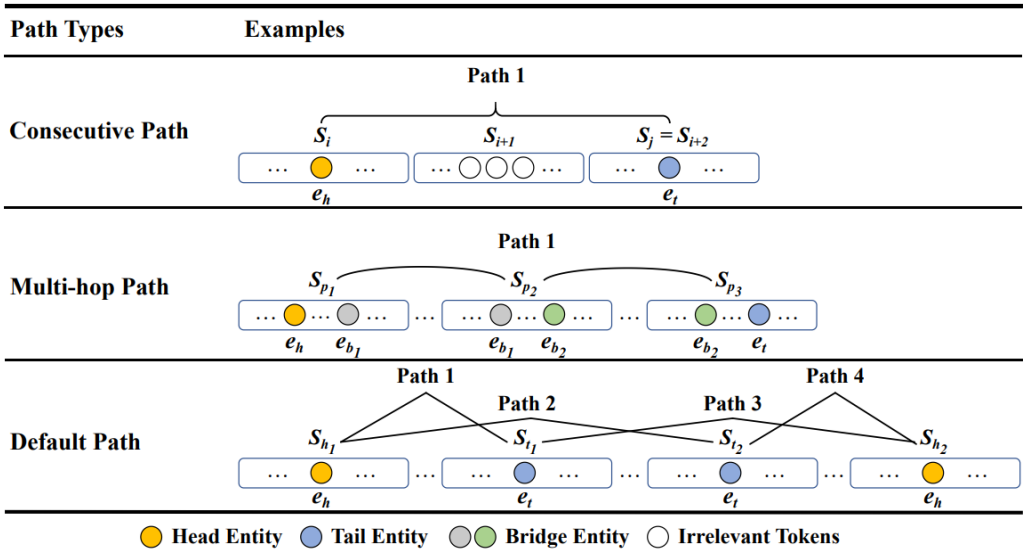
4.1 论文方法

调研论文设计了三条启发式规则寻找三类路径作为支撑证据：

（1）Consecutive Paths：特别的，连续的路径被认为是两个实体在文本中比较相近的情况，所以如果他们有三个连续的句子，那么就把这些句子作为一个路径。

（2）Multi-Hop Paths：当两个实体距离比较远，则需要 Multi-Hop 路径，通过一些和两个实体共同出现的其他实体作为桥梁连接。

（3）Default Paths：如果以上两种情况都不适用，收集所有的句子对，这些句子对中一个包含头实体，一个包含尾实体，来作为默认的路径。



本调研论文结果显示，87.5%的情况能用上述启发式规则给出覆盖。

将利用启发式规则提取出的语句输入 Bi-LSTM 模型进行关系抽取。对抽取结果计算 F1 准确率。

4.2 实验结果

本调研论文在 DocRED 数据集上做验证，准确率结果如下：

Model	Dev			Test
	Intra-F1	Inter-F1	F1	F1
CNN	51.87	37.58	43.45	42.26
BiLSTM	57.05	43.49	50.94	51.06
HIN-Glove	60.83	48.35	52.95	53.30
GAT	58.14	43.94	51.44	49.51
GCNN	57.78	44.11	51.52	51.62
EoG	58.90	44.60	52.15	51.82
AGGCN	58.76	45.45	52.47	51.45
LSR-Glove	60.83	48.35	55.17	54.18
GAIN-Glove	61.67	48.77	55.29	55.08
Paths+BiLSTM	**62.73**	**49.11**	**56.54**	**56.23**

结果显示，使用启发式规则+Bi-LSTM 框架的准确率优于使用复杂图神经网络的模型。

4.3 方法优点

我认为，本调研论文提出的启发式规则+Bi-LSTM 框架有以下优点：

（1）模型简单，节约存储空间。模型仅需 3 条启发式规则提取出关键句子，再输入 Bi-LSTM 框架抽取实体关系。这明显比利用复杂的图神经网络提取整个文档的语义简单。

（2）模型效果好。启发式规则提取出关键句子，对少数关键句子做关系抽取，效果会明显好于对整个文档学习含义的图神经网络，获得文档级别关系抽取任务的 SOTA。

第 5 章 调研论文横向比较

三篇论文内容上各有侧重，采用不同的模型结构完成关系抽取任务。

《GraphRel: Modeling Text as Relational Graphs for Joint Entity and Relation Extraction.》从加权图卷积神经网络的角度对关系做建模，以最大程度地提取文本信息，对文本间复杂关系有较好的表示能力。

《A Novel Cascade Binary Tagging Framework for Relational Triple Extraction.》采用全新的视角看待关系三元组，将关系建模为头实体到尾实体的映射函数，训练每种关系的尾实体标注器，从而从原理上解决了复杂关系表示问题。实际效果优于 GraphRel 模型。

《Three Sentences Are All You Need Local Path Enhanced Document Relation Extraction.》处理的是文档级别的关系抽取任务。不同于以往的模型需要使用复杂的网络结构对整个文档提取语义，本文聚焦于减少训练所用的句子数量，制定启发式规则提取出少数关键句子做关系抽取。

参考文献

[1] 张吉祥,张祥森,武长旭,赵增顺.知识图谱构建技术综述[J/OL].计算机工程:1-16[2021-11-20].https://doi.org/10.19678/j.issn.1000-3428.0061803.

[2] https://blog.csdn.net/weixin_40449300/article/details/88771302

[3] Vaswani A, Shazeer N, Parmar N, et al. Attention Is All You Need[J]. arXiv, 2017.