A Frustratingly Easy Approach for Joint Entity and Relation Extraction

Zexuan Zhong **Danqi Chen**
Department of Computer Science
Princeton University
{zzhong, danqic}@cs.princeton.edu

ArXiv 2020

动机：现有的端到端的 RE 模型（即同时抽取实体和关系），通常使用 joint 的模型，要么将他们使用统一的预测架构进行预测（比如基于标注的模型、Table Filling、Seq2Seq 模型等），要么使用共享的表示进行多任务学习（比如 SciIE、DyIE、DyIE++）。这篇文章提出一种 pipeline 的形式，相比前人更加简单，效果更好。

（一般大家都会认为 pipeline 的模型要比 joint 的模型差，由于 pipeline 会有错误累积的问题，但这篇给出了不同的观点，通过实验证明 joint 的方法共享表示反而比 pipeline 串联特征的模型效果更差。本文实验效果还挺不错的，也很有启发意义，详见方法和实验部分）

结果：

在 ACE04、ACE05 和 SciERC 上表现 SOTA

方法：

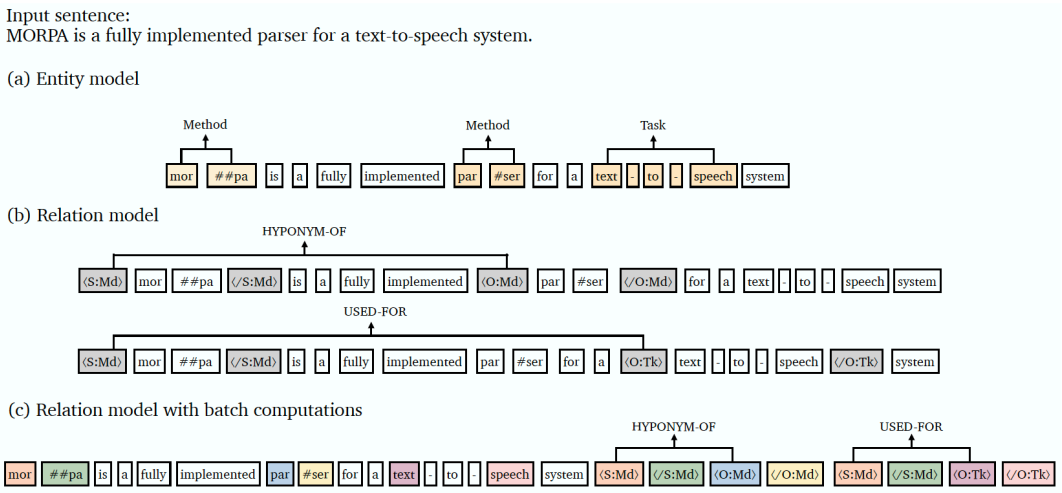

Figure 1: An example from the SciERC dataset (Luan et al., 2018), where a system is expected to identify that MORPA and PARSER are entities of type METHOD, TEXT-TO-SPEECH is a TASK, as well as MORPA is a *hyponym of* PARSER and MORPA is *used for* TEXT-TO-SPEECH. Our entity model (a) predicts all the entities at once and our relation model (b) considers every pair of entities independently by inserting typed entity markers (e.g., [S:MD] = the subject is a METHOD, [O:TK] = the object is a TASK). We also proposed an approximation relation model (c) which supports batch computations. The tokens of the same color in (c) share the positional embeddings. See text for more details.

这篇文章的方法特别简单，模型基于 span 级别的表示，也就是说数据输入是枚举所有 span，定义两个独立的预训练好的编码器（参数不共享，不一起训练，span 表示不共享），即 entity model 和 relation model，entity model 对每个 span 进行实体分类和实体类别分类，relation

model 对每个 span pair 进行关系分类。在 inference 的时候将 entity model 的输出作为 relation model 的输入特征。

Entity model:

Entity model Our entity model is a standard span-based model following prior work (Lee et al., 2017; Luan et al., 2018, 2019; Wadden et al., 2019). We first use a pre-trained language model (e.g., BERT) to obtain contextualized representations $\mathbf{x}_t$ for each input token x_t. Given a span $s_i \in S$, the span representation $\mathbf{h}_e(s_i)$ is defined as:

$$\mathbf{h}_e(s_i) = [\mathbf{x}_{\text{START}(i)}; \mathbf{x}_{\text{END}(i)}; \phi(s_i)],$$

where $\phi(s_i) \in \mathbb{R}^{d_W}$ represents the learned embeddings of span width features. The span representation $\mathbf{h}_e(s_i)$ is then used to predict entity types $e \in \mathcal{E} \cup \{\epsilon\}$:

$$P_e(e \mid s_i) = \text{softmax}(\mathbf{W}_e\text{FFNN}(\mathbf{h}_e(s_i)).$$

We follow Wadden et al. (2019) and use a 2-layer feedforward neural network with ReLU activations.

Relation model：

前人会共享实体识别和关系抽取的 span 表示，这个 span 的表示仅仅捕获了其所处位置的上下文信息。对于一个实体来说，他在不同实体对中的表示是一样的，也就是说不同实体对用的可能是同样的上下文，这明显是不合理的。作者认为前人的这种做法是一个局部最优解，很容易让模型学到的是基于相同上下文的特征去分类，而忽略了每个实体所需要的是不同的上下文。所以他认为需要在 relation model 中学习不同于 entity model 的 span 表示，并对于每个实体对，使用不同的上下文表示。

Our relation model instead processes each pair of spans independently and inserts typed markers at the input layer to highlight the subject and object and their types. Specifically, given an input sentence X and a pair of spans s_i, s_j, where s_i, s_j have a type of $e_i, e_j \in \mathcal{E} \cup \{\epsilon\}$ respectively. We define text markers as $\langle \text{S}{:}e_i \rangle$, $\langle /\text{S}{:}e_i \rangle$, $\langle \text{O}{:}e_j \rangle$, and $\langle /\text{O}{:}e_j \rangle$, and insert them into the input sentence before and after the subject and object spans (Figure 1 (b)).[3] Let $\widehat{X}$ denote this modified sequence with text markers inserted:

$$\widehat{X} = \ldots \langle \text{S}{:}e_i \rangle, x_{\text{START}(i)}, \ldots, x_{\text{END}(i)}, \langle /\text{S}{:}e_i \rangle, \ldots \langle \text{O}{:}e_j \rangle, x_{\text{START}(j)}, \ldots, x_{\text{END}(j)}, \langle /\text{O}{:}e_j \rangle, \ldots.$$

We then apply another pre-trained encoder on $\widehat{X}$ and denote the output representations by $\widehat{\mathbf{x}}_t$. We concatenate the output representations of two start positions and obtain the span pair representation:

$$\mathbf{h}_r(s_i, s_j) = [\widehat{\mathbf{x}}_{\widehat{\text{START}}(i)}; \widehat{\mathbf{x}}_{\widehat{\text{START}}(j)}],$$

where $\widehat{\text{START}}(i)$ and $\widehat{\text{START}}(j)$ are the indices of $\langle \text{S}{:}e_i \rangle$ and $\langle \text{O}{:}e_j \rangle$ in $\widehat{X}$. Finally, the representation $\mathbf{h}_r(s_i, s_j)$ will be used to predict the relation type $r \in \mathcal{R} \cup \{\epsilon\}$:

$$P_r(r|s_i, s_j) = \text{softmax}(\mathbf{W}_r\mathbf{h}_r(s_i, s_j)),$$

这里的 typed marker 也是一个创新，前人只用了头尾 marker，但是没有加入信息，可能是因为前人对于 entity 和 relation 不是两个独立的模型吧

实验结果：

Model	Encoder	ACE05			ACE04			SciERC		
		Ent	Rel	Rel+	Ent	Rel	Rel+	Ent	Rel	Rel+
(Li and Ji, 2014)	-	80.8	52.1	49.5	79.7	48.3	45.3	-	-	-
(Miwa and Bansal, 2016)	L	83.4	-	55.6	81.8	-	48.4	-	-	-
(Katiyar and Cardie, 2017)	L	82.6	55.9	53.6	79.6	49.3	45.7	-	-	-
(Zhang et al., 2017a)	L	83.6	-	57.5	-	-	-	-	-	-
(Luan et al., 2018)♣†	L+E	-	-	-	-	-	-	64.2	39.3	-
(Luan et al., 2019)♣†	L+E	88.4	63.2	-	87.4	59.7	-	65.2	41.6	-
(Li et al., 2019)	Bl	84.8	-	60.2	83.6	-	49.4	-	-	-
(Dixit and Al-Onaizan, 2019)	L+E	86.0	-	62.8	-	-	-	-	-	-
(Wadden et al., 2019)♣†	Bb	88.6	63.4	-	-	-	-	-	-	
(Wadden et al., 2019)♣†	SciB	-	-	-	-	-	-	67.5	48.4	
(Lin et al., 2020)	Bl	88.8	67.5	-	-	-	-	-	-	-
(Wang and Lu, 2020)	ALB	89.5	67.6	64.3	88.6	63.3	59.6	-	-	-
Ours: single-sentence	Bb	88.7	67.0	62.2	88.1	62.8	58.3	-	-	-
	SciB	-	-	-	-	-	-	66.6	48.4	35.1
	ALB	89.7	69.0	65.6	-	-	-	-	-	-
Ours: cross-sentence♣	Bb	90.2	67.7	64.6	89.2	63.9	60.1	-	-	-
	SciB	-	-	-	-	-	-	**68.2**	**50.1**	**36.7**
	ALB	**90.9**	**70.2**	**67.8**	**90.3**	**66.1**	**62.2**	-	-	-

结论：本文验证了对 entity 和 relation 学习不同表示、将实体信息融合在 relation 模型输入和使用全局上下文的重要性。

问题：

没有公布源码和超参，时间复杂度尽管通过 batch 化计算过程，但是因为需要遍历所有的 span 还是有点大。

可以借鉴的点：

1. 他的动机和 doc-level 的挑战有些重叠，并且在篇章级数据集 SciERC 上表现 SOTA，可以试着复现到 DocRED 上。
2. Pipeline 也能做这么好，原因来自于 distinct representation 的比较多，可以借鉴，然后模型参数只对一个任务敏感。我们也需要重视一下 pipeline 了
3. 使用 Rel+来评价，很严格
4. Paper 中还提到加入上下文信息，window=3 个句子，感觉对 doc-level 的数据集挺适用的
5. 关于 joint 模型的综述可以学习下，挺全的。