

# Learning from Context or Names? An Empirical Study on Neural Relation Extraction

Hao Peng<sup>1\*</sup>, Tianyu Gao<sup>2\*</sup>, Xu Han<sup>1</sup>, Yankai Lin<sup>3</sup>, Peng Li<sup>3</sup>, Zhiyuan Liu<sup>1†</sup>,  
Maosong Sun<sup>1</sup>, Jie Zhou<sup>3</sup>

<sup>1</sup>Department of Computer Science and Technology, Tsinghua University, Beijing, China

<sup>2</sup>Princeton University, Princeton, NJ, USA

<sup>3</sup>Pattern Recognition Center, WeChat AI, Tencent Inc, China

{h-peng17, hanxu17}@mails.tsinghua.edu.cn, tianyug@princeton.edu

收录会议：EMNLP 2020

动机：

目前 RE 领域对下面两个问题都没有清晰的理解：

- (1) 哪一类信息影响着现有的 RE 模型去做决策？
- (2) 如何让现有模型的效果得到更进一步的提升？

本文就是为了回答以上两个问题，从经验上去研究了文本中两个可能影响 RE 模型做决策的信息源头：文本上下文（textual context）和实体 mention（名字）

通过经验性的研究发现

- (1) 尽管上下文是支持 RE 模型预测结果的主要源头，RE 模型也严重地依赖实体 mention 的信息，其中主要是实体的类别信息；
- (2) 现有的数据可能通过实体 mention 得到了浅层的启发，因此导致了 RE 基准的高性能

预先试验：

Model	C+M	C+T	OnlyC	OnlyM	OnlyT
CNN	0.547	0.591	0.441	0.434	0.295
BERT	0.683	0.686	0.570	<b>0.466</b>	0.277
MTB	<b>0.691</b>	<b>0.696</b>	<b>0.581</b>	0.433	<b>0.304</b>

Table 1: TACRED results (micro  $F_1$ ) with CNN, BERT and MTB on different settings.

Context+Mention (C+M)：RE 任务最常见的设定，句子加上实体 mention

Context+Type (C+T)：将 C+M 中的 mention 使用它们的 type 代替，类似于 BERT 中的 [MASK] 比如 [person]

Only Context (OnlyC)：把句子中的头尾实体分别用 [SUBJ] 和 [OBJ] 代替，这种情况下，相当于输入的时候把实体的信息去掉了，包括实体的类别信息。

Only Mention (OnlyM)：只使用头尾实体的 mention 去做关系分类。

Only Type (OnlyT)：和 OnlyM 类似，但是只提供实体类别。

分析结论：

- (1) 文本上下文和实体 mention 都提供了支持关系分类的重要信息。

A. 文本上下文提供了支持关系分类的重要信息：C + M v.s. Only M 11-26 个点, C + T v.s. Only T 15-39 个点

B. 实体 mention 提供了支持关系分类的重要信息：C + M v.s. Only C 10-11 个点

实体 mention 中最有用的信息是类别信息：在 BERT 和 MTB 上 C + T v.s. C + M 效果相当，

但是在 CNN 上 C + T 的效果比 C + M 的效果还要好。

(2) Only\*的效果都不好，说明只依赖于单一信息源是不够的，上下文和实体 mention 都对关系分类很重要

(3) 现有的 RE 数据集存在肤浅的提示

Only M 的表现还不错诶。这说明了只用 mention 就能达到一定的效果，说明现有的数据集中 mention 泄露了一些肤浅的提示，这个导致了 RE 模型的高性能。

Case 分析：

C+M
Although her family was from Arkansas, <i>she</i> was born in <i>Washington</i> state, where ... Label: <b>per:state.of.birth</b> Prediction: <b>per:state.of.residence</b>
Dozens of lightly regulated subprime lenders, including New Century Financial Corp., have failed and troubled <i>Countrywide Financial Corp.</i> was acquired by <i>Bank of America Corp.</i> Label: <b>org:parents</b> Prediction: <b>no.relation</b>
C+T
First, <i>Natalie Hagemo</i> says, <i>she</i> fought the Church of Scientology just to give birth to her daughter. Label: <b>no.relation</b> Prediction: <b>per:children</b>
Earlier this week Jakarta hosted the <i>general assembly</i> of the <i>Organisation of Asia-Pacific News Agencies</i> , ... Label: <b>no.relation</b> Prediction: <b>org:members</b>
The boy, identified by the Dutch foreign ministry as <i>Ruben</i> but more fully by Dutch media as <i>Ruben van Assouw</i> , ... Label: <b>per:alternate.names</b> Prediction: <b>no.relation</b>

Table 2: Wrong predictions made only by C+M and only by C+T, where red and blue represent subject and object entities respectively. As the examples suggest, C+M is more easily biased by the entity distribution in the training set and C+T loses some information from mentions that helps to understand the text.

(1) C + M 和 C + T 共享了 95.7%的预测正确的样本，C + M 中 68.1 的错误预测和 C + T 一样，这说明了模型利用实体 mention 主要利用的是类别信息。

(2) 除了 mention 的 type 以外，entity 还提供了其他的信息。

Type	Example
Wrong 42%	..., <i>Jacinto Suarez</i> , Nicaraguan deputy to the <i>Central American Parliament</i> (PARLACEN) said Monday. Label: <b>org:top.members/employees</b> Prediction: <b>no.relation</b>  US life insurance giant MetLife said on Monday it will acquire <i>American International Group</i> unit American Life Insurance company ( <i>ALICO</i> ) in a deal worth 155 billion dollars. Label: <b>org:subsidiaries</b> Prediction: <b>no.relation</b>
No pattern 31%	On Monday, the judge questioned the leader of the <i>Baptist</i> group, <i>Laura Silsby</i> , who ... Label: <b>per:religion</b> Prediction: <b>no.relation</b>
Confusing 27%	About a year later, <i>she</i> was transferred to Camp Hope, <i>Iraq</i> . Label: <b>per:countries.of.residence</b> Prediction: <b>per:stateorprovinces.of.residence</b>

Table 3: Case study on unique wrong predictions made by OnlyC (compared to C+M). We sample 10% of the wrong predictions, filter the wrong-labeled instances and manually annotate the wrong types to get the proportions. We use red and blue to highlight the subject and object entities.

Wrong：存在明显的 pattern，但是模型没有理解到

No pattern：mask 掉 entity pair 后，对于人来说都很难判断的关系

Confusing：mask 掉 entity pair 后，有歧义的关系

这个 case 由于是 Only C 相对于 C + M 特有的，对于 42%的 Wrong，说明 C + M 模型确

实是依赖了 Mention 中的特有模式 (shallow heuristics)，而没有利用上 context 的信息。

这里的 case 分析感觉有点问题，mask 掉 entity pair 之后，只剩 context 了，这个 context 就一定能确保只有一种关系吗？和数据集相关？或许 context 中的实体在 KG 是有关系的，但是 context 没有表达出来这个关系？

本文的方法：

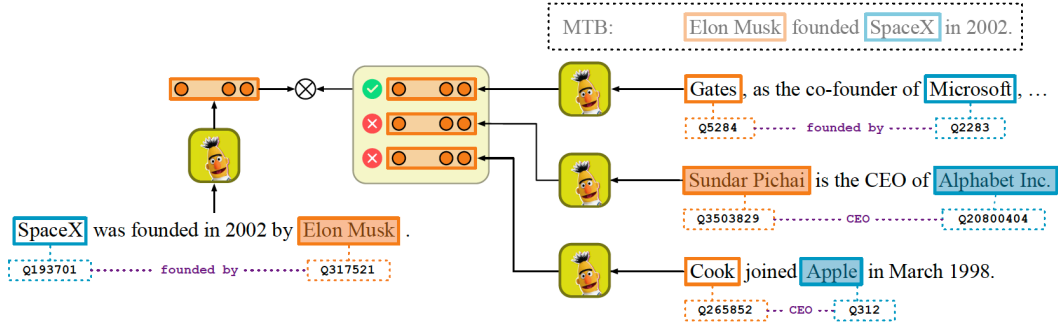


Figure 2: Our contrastive pre-training framework for RE. We assign relations to sentences by linking entity pairs in sentences to Wikidata and checking their relations in the KG. We assume that sentences with the same relation should have similar representations, and those with different relations should be pushed apart. Entity mentions are randomly masked (boxes with colored background) to avoid simple memorization. Compared to MTB (in the dotted box), our method samples data with better diversity, which can not only increase the coverage of entity types and diverse context but also reduce the possibility of memorizing entity names.

(1) 本文采用对比学习的思想(contrastive learning)，这个思想主要是让”neighbor”之间的表示相近，而”non-neighbor”之间的表示更远。这里的”neighbor”就是相近的样本或者实例，对于 RE 任务来说，就是句子表达了相同的关系。

(2) 为了让模型不过拟合实体的特有模式 (shallow heuristic)，随机使用特殊的 token [BLANK]来对实体 mention 进行 mask，随机 mask 的概率为 0.7，

句子的表示就是实体对的表示的拼接

$$\mathbf{x} = \text{ENC}_h(x) \oplus \text{ENC}_t(x), \quad (1)$$

对比预训练的 loss

$$\mathcal{L}_{CP} = -\log \frac{e^{\mathbf{x}_A^T \mathbf{x}_B}}{e^{\mathbf{x}_A^T \mathbf{x}_B} + \sum_{i=1}^{i \leq N} e^{\mathbf{x}_A^T \mathbf{x}_B^i}}. \quad (2)$$

最终的 loss 会加上 MLM 的 loss

Eventually, we have the following training loss:

$$\mathcal{L} = \mathcal{L}_{CP} + \mathcal{L}_{MLM}. \quad (3)$$

这个方法虽然是让模型去学习表达相同 relation 的 context 应该具有相同的表示（这个表示就是句子 entity pair 的拼接），可以一定程度上从多样化的角度学习到关系的不同表达。

Mask 掉 entity 可以避免模型对 entity 的死记硬背，鼓励模型去利用 entity 的 type 信息。生成策略提供了多样的文本上下文来表达相同的关系，可以鼓励模型去学会从各种表达中抽取关系模式。

效果：

Dataset	Model	1%			10%			100%		
		C+M	OnlyC	OnlyM	C+M	OnlyC	OnlyM	C+M	OnlyC	OnlyM
TACRED	BERT	0.211	0.167	0.220	0.579	0.446	0.433	0.683	0.570	<b>0.466</b>
	MTB	0.304	0.231	0.308	0.608	0.496	0.441	0.691	0.581	0.433
	CP	<b>0.485</b>	<b>0.393</b>	<b>0.350</b>	<b>0.633</b>	<b>0.515</b>	<b>0.453</b>	<b>0.695</b>	<b>0.593</b>	0.450
SemEval	BERT	0.367	0.294	0.245	0.772	0.688	0.527	0.871	0.798	0.677
	MTB	0.362	0.330	<b>0.249</b>	0.806	0.744	0.543	0.873	0.807	<b>0.682</b>
	CP	<b>0.482</b>	<b>0.470</b>	0.221	<b>0.822</b>	<b>0.766</b>	<b>0.543</b>	<b>0.876</b>	<b>0.811</b>	0.679
Wiki80	BERT	0.559	0.413	0.463	0.829	0.413	0.655	0.913	0.810	0.781
	MTB	0.585	0.509	0.542	0.859	0.509	0.719	0.916	0.820	0.788
	CP	<b>0.827</b>	<b>0.734</b>	<b>0.653</b>	<b>0.893</b>	<b>0.734</b>	<b>0.745</b>	<b>0.922</b>	<b>0.834</b>	<b>0.799</b>
ChemProt	BERT	0.362	0.362	0.362	0.634	0.584	0.385	0.792	0.777	0.463
	MTB	<b>0.362</b>	0.362	<b>0.362</b>	0.682	0.685	0.403	0.796	0.798	0.463
	CP	0.361	<b>0.362</b>	0.360	<b>0.708</b>	<b>0.697</b>	<b>0.404</b>	<b>0.806</b>	<b>0.803</b>	<b>0.467</b>

Table 5: Results on supervised RE datasets TACRED (micro  $F_1$ ), SemEval (micro  $F_1$ ), Wiki80 (accuracy) and ChemProt (micro  $F_1$ ). 1% / 10% indicate using 1% / 10% supervised training data respectively.

可以借鉴的点：

- (1) 结合 case 来进行分析得出本文的动机
- (2) 和 danqi chen 的 Arxiv 上的 paper 不一样，这里是用预训练的方式，随机 mask 掉 entity，用 type 替换，而 danqi chen 是用 type 的 marker 加在实体两边。
- (3) 附录部分可以借鉴，很多地方之前都没想过可以放在附录。

问题：

- (1) 一个 context 可能不止表达一种关系，也可能不表达关系，这篇文章在这方面的说法可能局限在某个数据集
- (2) 效果其实一般般，设计的模型解释性比较差，对比学习让模型去学习 context 信息的方式比较隐式，是不是用 attention 的方式对每个实体聚合 context 中的信息会更好呢？