Minimize Exposure Bias of Seq2Seq Models in Joint Entity and Relation Extraction

Ranran Haoran Zhang*[1], Qianying Liu*[2], Aysa Xuemo Fan[1], Heng Ji[1], Daojian Zeng[4], Fei Cheng[2], Daisuke Kawahara[3] and Sadao Kurohashi[2]

[1] University of Illinois at Urbana-Champaign
[2] Graduate School of Informatics, Kyoto University
[3] School of Fundamental Science and Engineering, Waseda University
[4] Hunan Normal University
{haoranz6,xuemof2,hengji}@illinois.edu; ying@nlp.ist.i.kyoto-u.ac.jp
{feicheng,kuro}@i.kyoto-u.ac.jp; dkw@waseda.jp; zengdj916@163.com

EMNLP 2020 finding

本篇文章作者主要想解决的问题是因为训练过程的解码与测试过程的解码方式不一样，训练时有teacher forcing带有预先设定的顺序，但是测试时没有gold-standard labels，所以会造成Exposure bias的问题，降低了模型的表现。

编码器:

1、 使用word embedding和Bi-LSTM对句子进行编码，得到句子表示

$$[\boldsymbol{s}_0^E, \boldsymbol{s}_1^E, \ldots, \boldsymbol{s}_n^E] = \text{Encoder}([x_0, x_1, \ldots, x_n]) \quad (1)$$

2、 对上述结果经过一个卷积层，得到句子的一个辅助表示，用于后续进行attention计算

$$\boldsymbol{o}_0 = \text{Conv}_{en}([\boldsymbol{s}_0^E, \boldsymbol{s}_1^E, \ldots, \boldsymbol{s}_n^E]) \quad (2)$$

解码器: （LSTM）

使用w_{sos}作为解码开始标识符

使用**W**r表示关系特征

实体的表示如下，e1代表实体的第一个token，e2代表实体的第二个token

(c) entity embedding: $\boldsymbol{w}_t^e = \boldsymbol{o}_{t-1}^{e1} + \boldsymbol{o}_{t-1}^{e2} \in \mathbb{R}^h$,

1、 使用上一步解码出的token的特征与上一步解码器的hidden特征拼接在一起，进行当前步骤的解码

$$\boldsymbol{s}_t^D = \text{Decoder}(\boldsymbol{w}_t, \boldsymbol{s}_{t-1}^D)$$

2、 对当前解码出的token特征与经过卷积层的句子表示拼接在一起，计算attention值

$$\boldsymbol{a}_t = \text{Attention}(\boldsymbol{o}_{t-1}, \boldsymbol{s}_t^D) \quad (4)$$

3、 将计算出的at与Ot-1进行拼接后，进行卷积操作，更新O（需要将at重复n次，因为at是1Xh，O是n X h）

$$\boldsymbol{o}_t = \text{Conv}_{de}([\boldsymbol{a}_t; \boldsymbol{o}_{t-1}^{0:n}]) \quad (5)$$

4、 根据上述得到的Ot进行关系分类，Max表示max polling

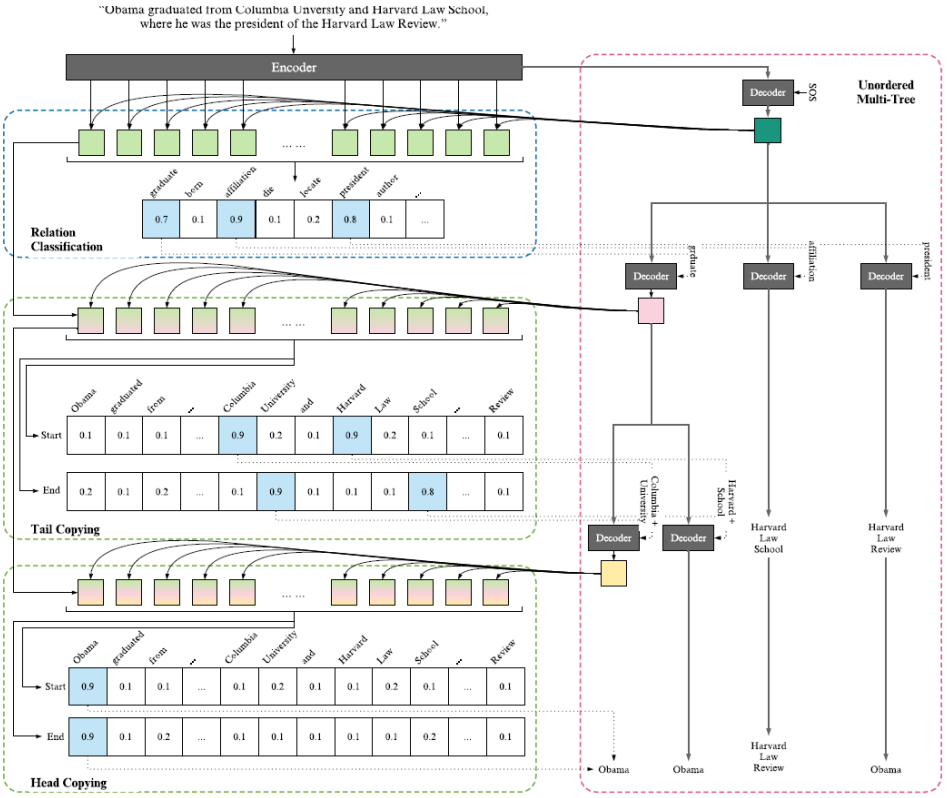
$$\boldsymbol{prob}_r = \sigma(\mathrm{Max}(\boldsymbol{o}_t \boldsymbol{W}_r + \boldsymbol{b}_r)) \qquad (6)$$

5、 根据上述的Ot进行实体预测，激活函数使用sigmoid，一次可预测多个实体

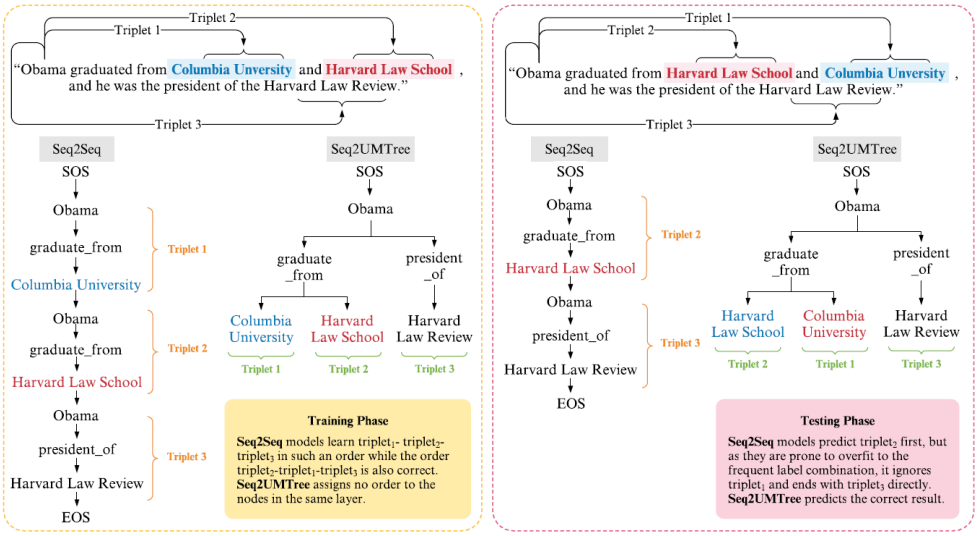
$$\begin{aligned} \boldsymbol{prob}_{e_b} &= \sigma(\boldsymbol{W}_{e_b}^T \boldsymbol{o}_t + b_{e_b}) \\ \boldsymbol{prob}_{e_e} &= \sigma(\boldsymbol{W}_{e_e}^T \boldsymbol{o}_t + b_{e_e}) \end{aligned} \qquad (7)$$

模型在训练时仍然使用teacher forcing的方式，decoder部分使用gold-standard labels的embedding，在论文中作者未说明怎么编码gold-standard labels和relation的特征，因此我自己认为实体的表示仍然使用第一个token和最后一个token特征相加表示（上面已说明），而关系编码并未说明

模型图:



模型示例：



实验结果：

	NYT				DuIE			
	test#	Prec	Rec	F1	test#	Prec	Rec	F1
CopyMTL	.978	.685	.648	.666	.962	.496	.394	439
WDec	.988	**.843**	**.764**	**.802**	.919	.641	.542	.587
MHS	.995	.798	.739	.768	.984	**.772**	.623	.690
Seq2UMTree	1.00	.791	.751	.771	1.00	.756	**.730**	**.743**

Table 1: Main Results on NYT and DuIE. #test is the valid sentence percentage of the test set to the models.

从上述来看模型在NYT上表现并不好，因此作者对NYT数据集进行了分析，发现90%的测试集与训练集重叠，而在DuIE上表现不错（30%重叠）
作者又进行了关于解码三元组顺序对模型表现得影响：

	Order	Prec	Rec	F1
NYT	t, r, h	.788	.694	.738
	r, t, h	.791	**.751**	**.771**
	t, h, r	.765	.495	.601
	h, t, r	.756	.548	.635
	r, h, t	.789	.737	.762
	h, r, t	**.796**	.685	.737
DuIE	t, r, h	.766	.663	.711
	r, t, h	.756	**.730**	**.743**
	t, h, r	**.802**	.330	.467
	h, t, r	.794	.120	.208
	r, h, t	.760	.712	.735
	h, r, t	.731	.728	.729

Table 4: Different orders of Seq2UMTree.

由此发现不同顺序会有不同结果，作者也是固定顺序取最好得结果

总体而言，我认为作者并未体现出在decoder的无序解码，仍然固定了顺序，取得结果也是（r-t-h）顺序得到的结果，但是也能说明不同的解码顺序对于模型的表现会有不同影响，而（r-t-h）顺序表现最好，作者给出的解释是认为先识别出关系对于模型识别实体会提供限制的信息帮助识别实体。

个人认为本篇文章并未能解决所提出Exposure bias的问题，没有实现无序解码，只是说明了解码的顺序对模型的影响，因此此问题仍然有待解决，再者就是NYT数据集的不规范性，训练集与测试集高度重叠，不能真实反应模型的表现