CopyMTL: Copy Mechanism for Joint Extraction of Entities and Relations with Multi-Task Learning

Daojian Zeng[*§], **Haoran Zhang**[*†] , **Qianying Liu**[‡]
[§]Changsha University of Science & Technology, Changsha, 410114, China
[†]University of Illinois at Urbana-Champaign, Illinois, 61820, USA
[‡]Kyoto University, Kyoto, 606-8501, Japan
zengdj916@163.com, haoranz6@illinois.edu, ying@nlp.ist.i.kyoto-u.ac.jp

Background: CopyRE

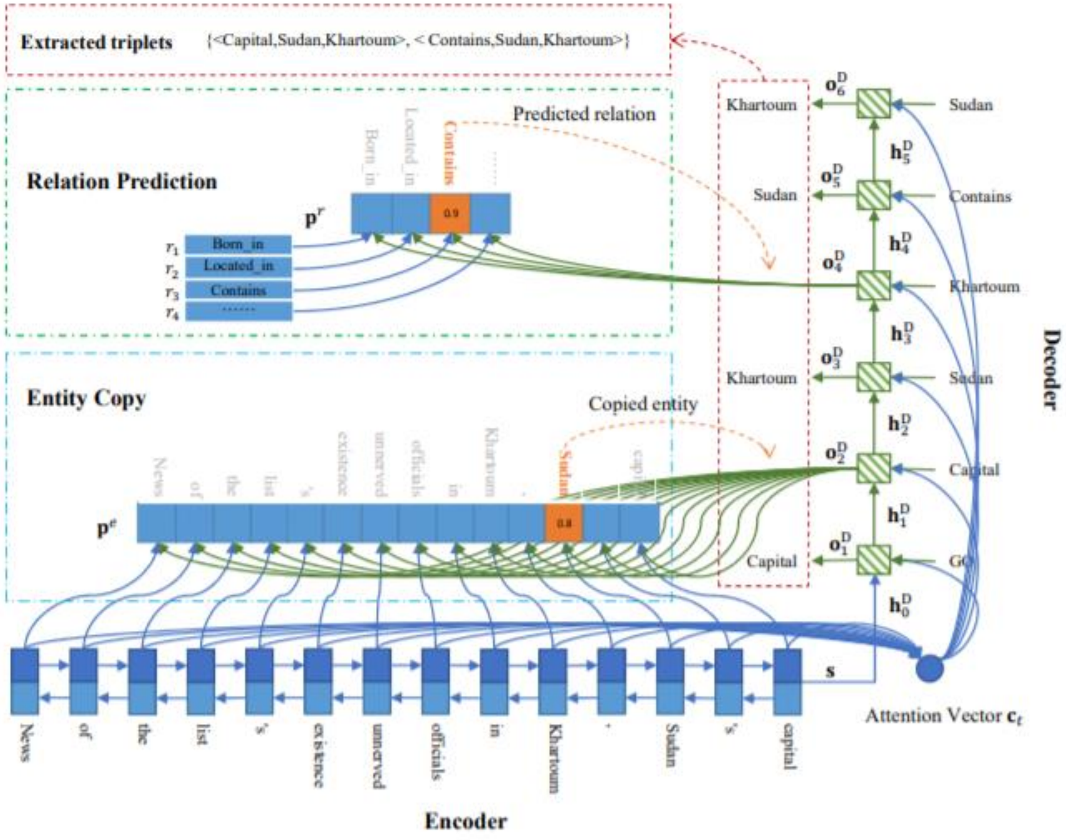
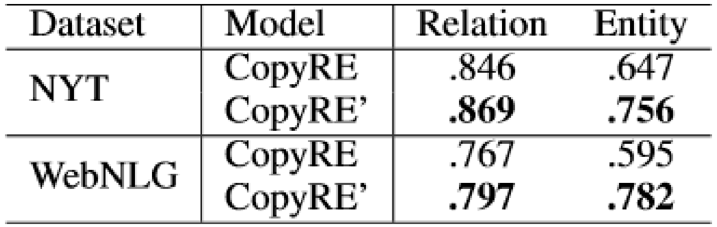


Figure 2: The overall structure of OneDecoder model. A bi-directional RNN is used to encode the source sentence and then a decoder is used to generate triples directly. The relation is predicted and the entity is copied from source sentence.

Problem

- First, the entity copying in CopyRE is unstable and it depends on an unnatural mask to differ the head(h) and tail(t) entities

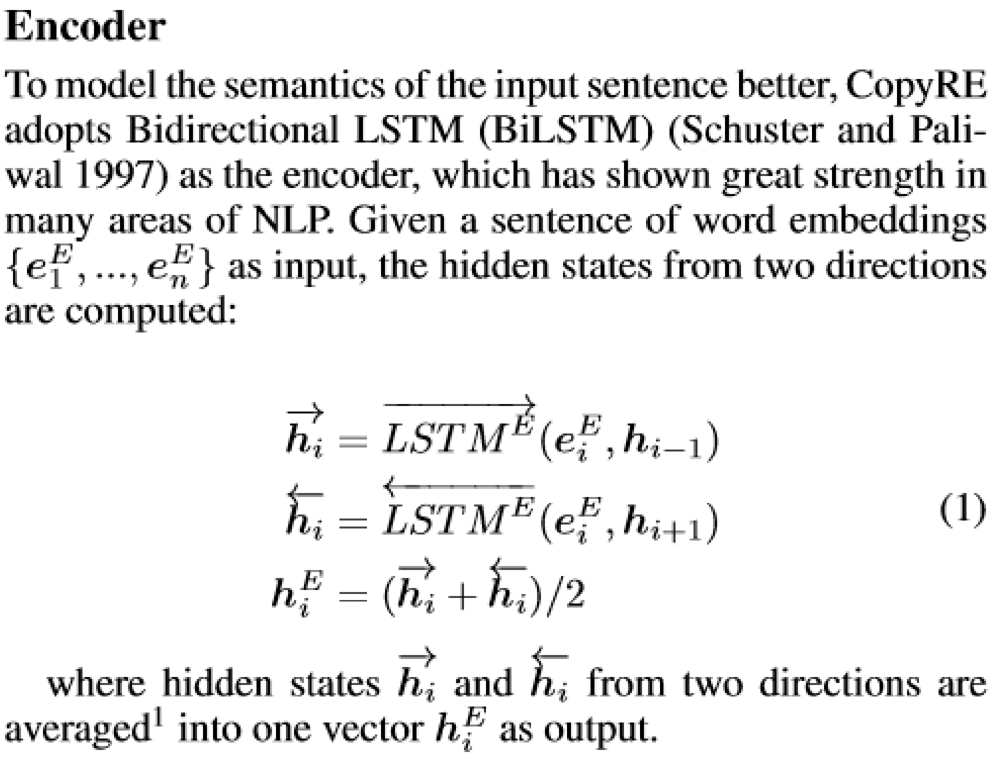


Dataset	Model	Relation	Entity
NYT	CopyRE	.846	.647
	CopyRE’	**.869**	**.756**
WebNLG	CopyRE	.767	.595
	CopyRE’	**.797**	**.782**

Table 4: F1 scores on subtasks.

- Second, CopyRE cannot extract entities that have multiple tokens. The copy-based decoder always points to the last token of any entities, which limits the applicability of the model.

Background

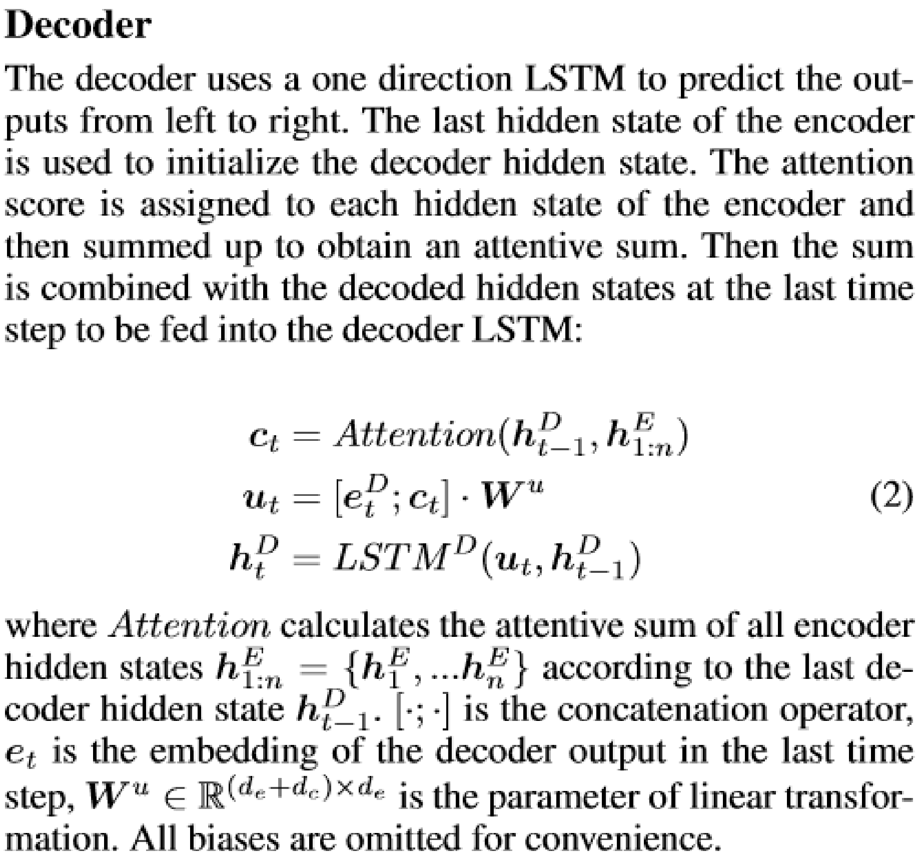


Encoder

To model the semantics of the input sentence better, CopyRE adopts Bidirectional LSTM (BiLSTM) (Schuster and Paliwal 1997) as the encoder, which has shown great strength in many areas of NLP. Given a sentence of word embeddings $\{\boldsymbol{e}_1^E, ..., \boldsymbol{e}_n^E\}$ as input, the hidden states from two directions are computed:

$$\begin{aligned} \overrightarrow{\boldsymbol{h}_i} &= \overrightarrow{LSTM^E}(\boldsymbol{e}_i^E, \boldsymbol{h}_{i-1}) \\ \overleftarrow{\boldsymbol{h}_i} &= \overleftarrow{LSTM^E}(\boldsymbol{e}_i^E, \boldsymbol{h}_{i+1}) \\ \boldsymbol{h}_i^E &= (\overrightarrow{\boldsymbol{h}_i} + \overleftarrow{\boldsymbol{h}_i})/2 \end{aligned} \tag{1}$$

where hidden states $\overrightarrow{\boldsymbol{h}_i}$ and $\overleftarrow{\boldsymbol{h}_i}$ from two directions are averaged[1] into one vector $\boldsymbol{h}_i^E$ as output.



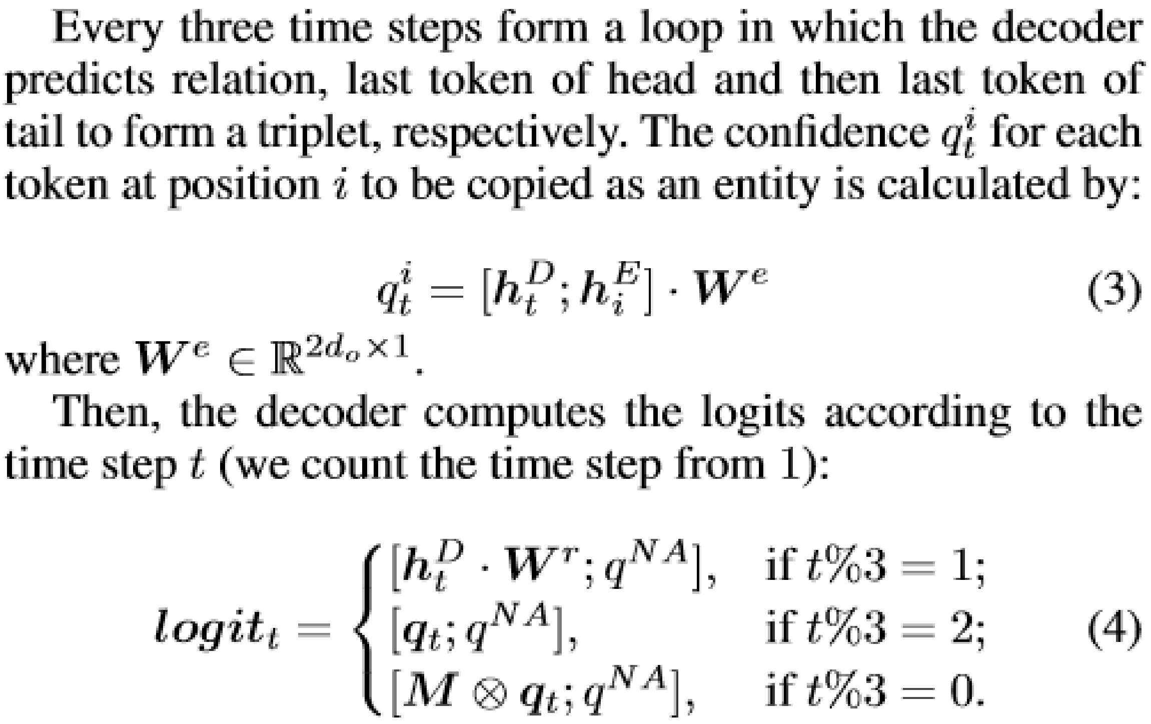
Decoder

The decoder uses a one direction LSTM to predict the outputs from left to right. The last hidden state of the encoder is used to initialize the decoder hidden state. The attention score is assigned to each hidden state of the encoder and then summed up to obtain an attentive sum. Then the sum is combined with the decoded hidden states at the last time step to be fed into the decoder LSTM:

$$\begin{aligned} \boldsymbol{c}_t &= Attention(\boldsymbol{h}_{t-1}^D, \boldsymbol{h}_{1:n}^E) \\ \boldsymbol{u}_t &= [\boldsymbol{e}_t^D; \boldsymbol{c}_t] \cdot \boldsymbol{W}^u \\ \boldsymbol{h}_t^D &= LSTM^D(\boldsymbol{u}_t, \boldsymbol{h}_{t-1}^D) \end{aligned} \tag{2}$$

where $Attention$ calculates the attentive sum of all encoder hidden states $\boldsymbol{h}_{1:n}^E = \{\boldsymbol{h}_1^E, ...\boldsymbol{h}_n^E\}$ according to the last decoder hidden state $\boldsymbol{h}_{t-1}^D$. $[\cdot;\cdot]$ is the concatenation operator, $\boldsymbol{e}_t$ is the embedding of the decoder output in the last time step, $\boldsymbol{W}^u \in \mathbb{R}^{(d_e+d_c)\times d_e}$ is the parameter of linear transformation. All biases are omitted for convenience.

copyRE



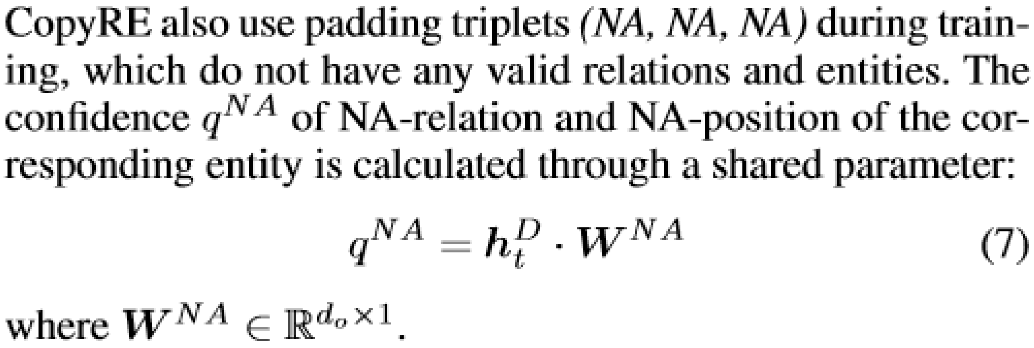
Every three time steps form a loop in which the decoder predicts relation, last token of head and then last token of tail to form a triplet, respectively. The confidence q_t^i for each token at position i to be copied as an entity is calculated by:

$$q_t^i = [\boldsymbol{h}_t^D; \boldsymbol{h}_i^E] \cdot \boldsymbol{W}^e \tag{3}$$

where $\boldsymbol{W}^e \in \mathbb{R}^{2d_o \times 1}$.

Then, the decoder computes the logits according to the time step t (we count the time step from 1):

$$\boldsymbol{logit}_t = \begin{cases} [\boldsymbol{h}_t^D \cdot \boldsymbol{W}^r; q^{NA}], & \text{if } t\%3 = 1; \\ [\boldsymbol{q}_t; q^{NA}], & \text{if } t\%3 = 2; \\ [\boldsymbol{M} \otimes \boldsymbol{q}_t; q^{NA}], & \text{if } t\%3 = 0. \end{cases} \tag{4}$$

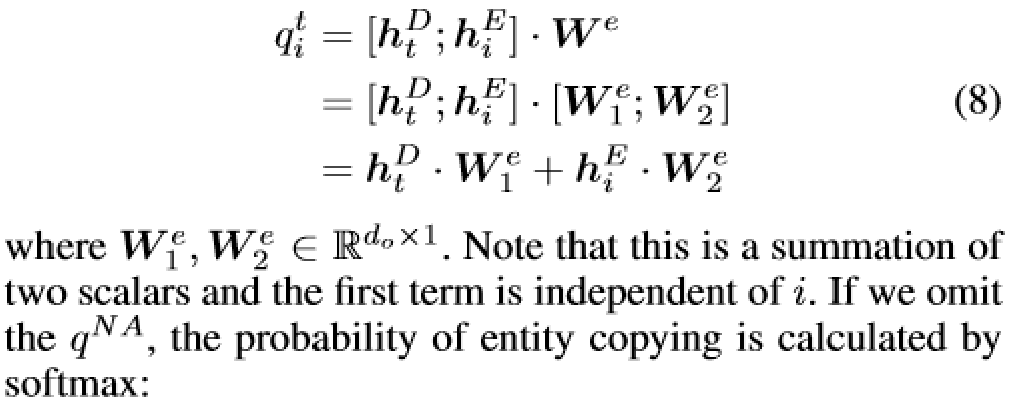


CopyRE also use padding triplets *(NA, NA, NA)* during training, which do not have any valid relations and entities. The confidence q^{NA} of NA-relation and NA-position of the corresponding entity is calculated through a shared parameter:

$$q^{NA} = \boldsymbol{h}_t^D \cdot \boldsymbol{W}^{NA} \tag{7}$$

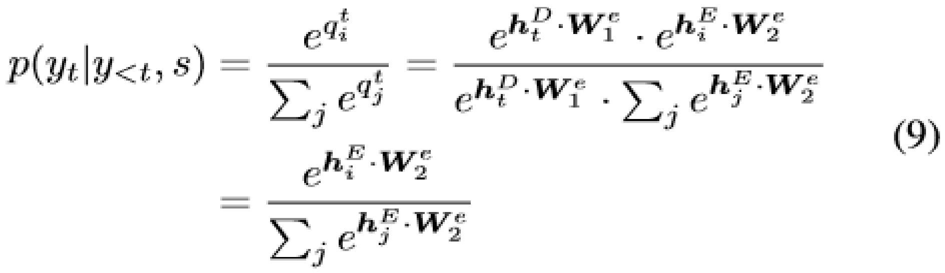
where $\boldsymbol{W}^{NA} \in \mathbb{R}^{d_o \times 1}$.

Unstable and Abnormal dependency to the mask

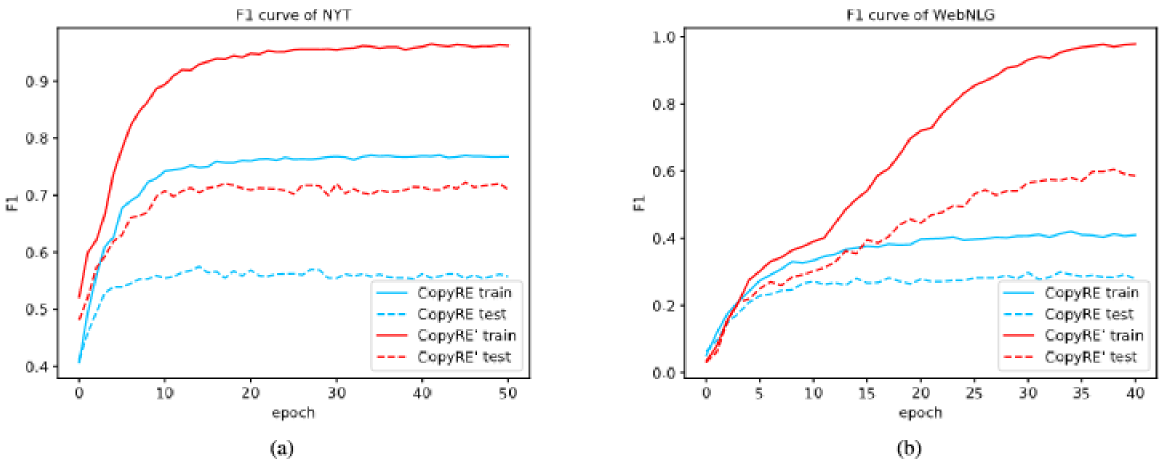


$$\begin{aligned} q_i^t &= [\boldsymbol{h}_t^D; \boldsymbol{h}_i^E] \cdot \boldsymbol{W}^e \\ &= [\boldsymbol{h}_t^D; \boldsymbol{h}_i^E] \cdot [\boldsymbol{W}_1^e; \boldsymbol{W}_2^e] \\ &= \boldsymbol{h}_t^D \cdot \boldsymbol{W}_1^e + \boldsymbol{h}_i^E \cdot \boldsymbol{W}_2^e \end{aligned} \tag{8}$$

where $\boldsymbol{W}_1^e, \boldsymbol{W}_2^e \in \mathbb{R}^{d_o \times 1}$. Note that this is a summation of two scalars and the first term is independent of i. If we omit the q^{NA}, the probability of entity copying is calculated by softmax:



$$\begin{aligned} p(y_t|y_{<t}, s) &= \frac{e^{q_i^t}}{\sum_j e^{q_j^t}} = \frac{e^{\boldsymbol{h}_t^D \cdot \boldsymbol{W}_1^e} \cdot e^{\boldsymbol{h}_i^E \cdot \boldsymbol{W}_2^e}}{e^{\boldsymbol{h}_t^D \cdot \boldsymbol{W}_1^e} \cdot \sum_j e^{\boldsymbol{h}_j^E \cdot \boldsymbol{W}_2^e}} \\ &= \frac{e^{\boldsymbol{h}_i^E \cdot \boldsymbol{W}_2^e}}{\sum_j e^{\boldsymbol{h}_j^E \cdot \boldsymbol{W}_2^e}} \end{aligned} \tag{9}$$

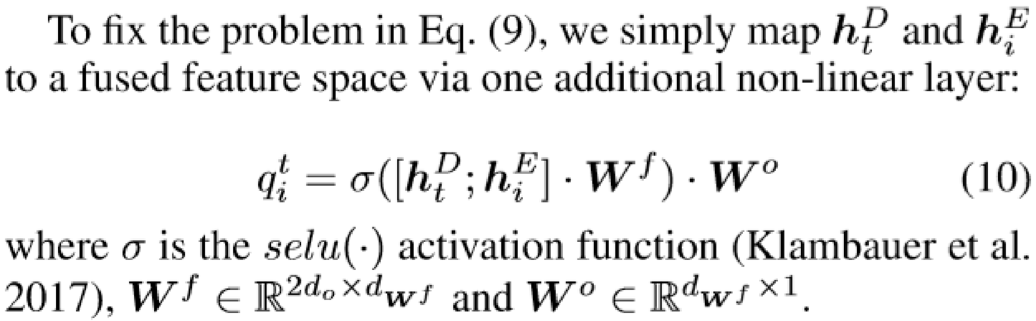


(a)

(b)

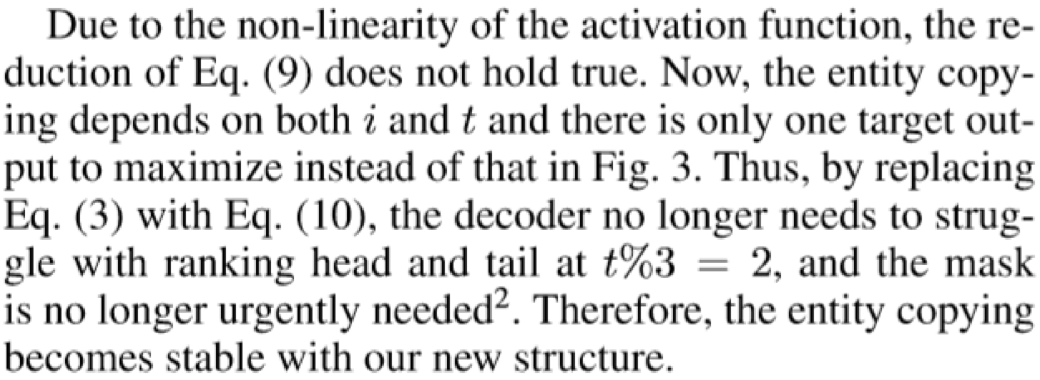
Figure 4: The training curves of CopyRE and CopyRE' on NYT and WebNLG.

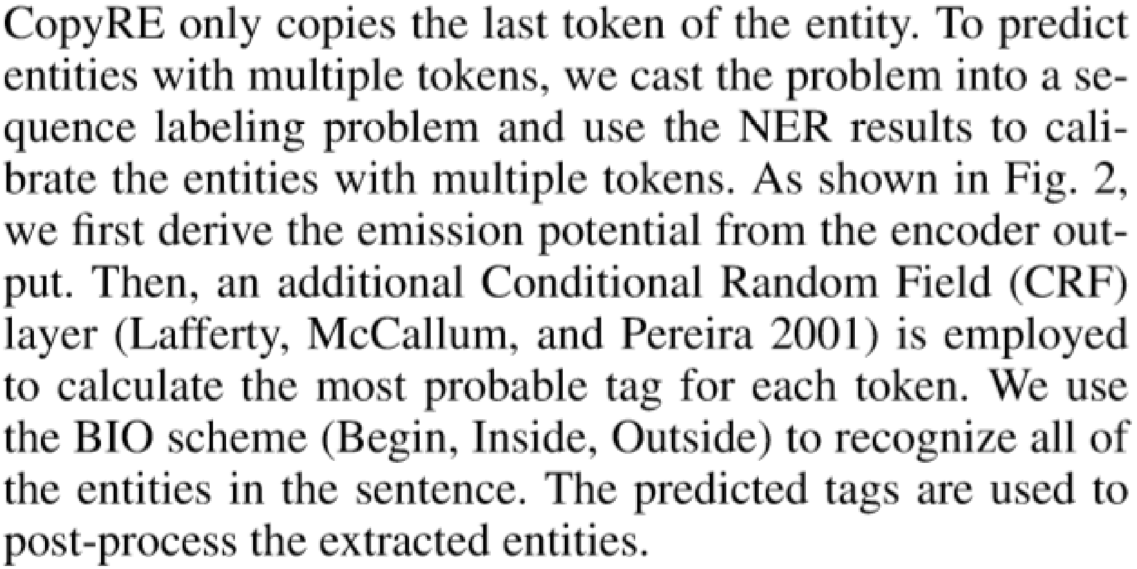
Our Change


To fix the problem in Eq. (9), we simply map $\boldsymbol{h}_t^D$ and $\boldsymbol{h}_i^E$ to a fused feature space via one additional non-linear layer:

$$q_i^t = \sigma([\boldsymbol{h}_t^D; \boldsymbol{h}_i^E] \cdot \boldsymbol{W}^f) \cdot \boldsymbol{W}^o \qquad (10)$$

where σ is the $selu(\cdot)$ activation function (Klambauer et al. 2017), $\boldsymbol{W}^f \in \mathbb{R}^{2d_o \times d_{\boldsymbol{W}^f}}$ and $\boldsymbol{W}^o \in \mathbb{R}^{d_{\boldsymbol{W}^f} \times 1}$.


Due to the non-linearity of the activation function, the reduction of Eq. (9) does not hold true. Now, the entity copying depends on both i and t and there is only one target output to maximize instead of that in Fig. 3. Thus, by replacing Eq. (3) with Eq. (10), the decoder no longer needs to struggle with ranking head and tail at $t\%3 = 2$, and the mask is no longer urgently needed[2]. Therefore, the entity copying becomes stable with our new structure.


CopyRE only copies the last token of the entity. To predict entities with multiple tokens, we cast the problem into a sequence labeling problem and use the NER results to calibrate the entities with multiple tokens. As shown in Fig. 2, we first derive the emission potential from the encoder output. Then, an additional Conditional Random Field (CRF) layer (Lafferty, McCallum, and Pereira 2001) is employed to calculate the most probable tag for each token. We use the BIO scheme (Begin, Inside, Outside) to recognize all of the entities in the sentence. The predicted tags are used to post-process the extracted entities.

Model

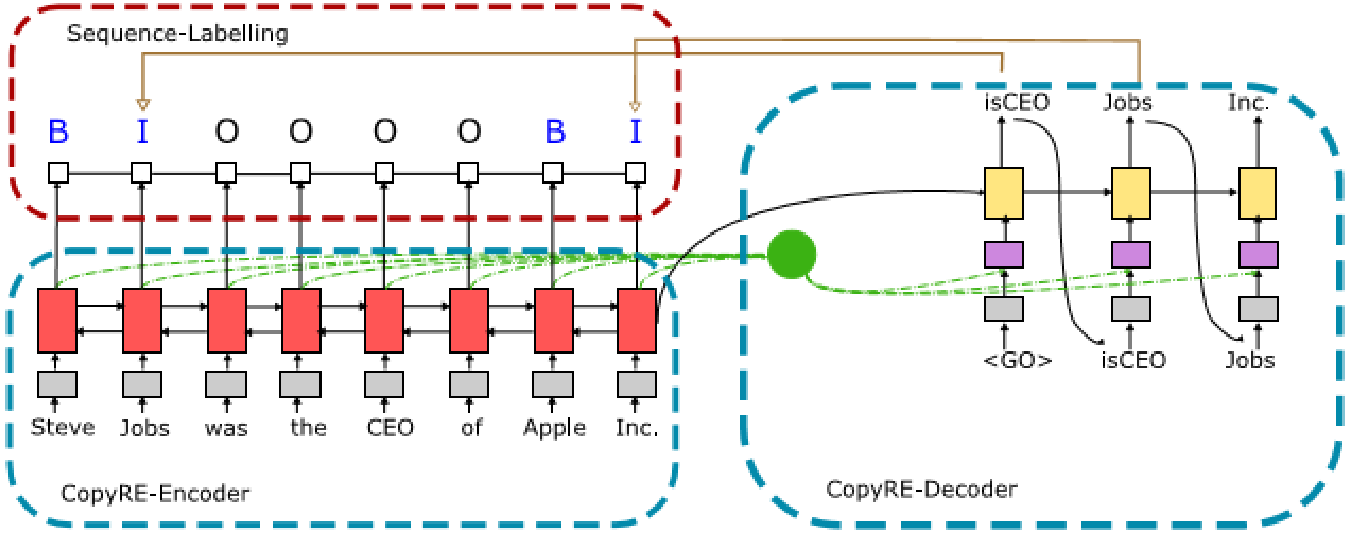
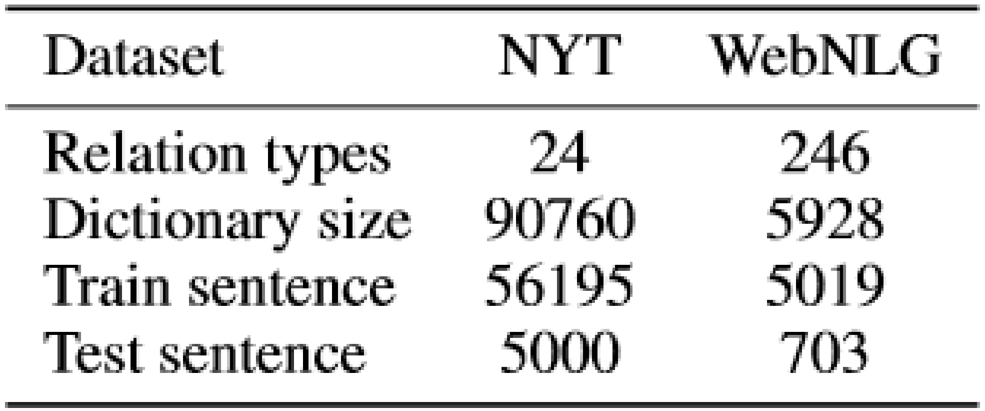


Figure 2: The overview of CopyMTL model for joint extraction of relation and entity. The CopyRE model does not contain the CopyMTL-Tagging part, i.e., the sequence-labeling part in the figure.

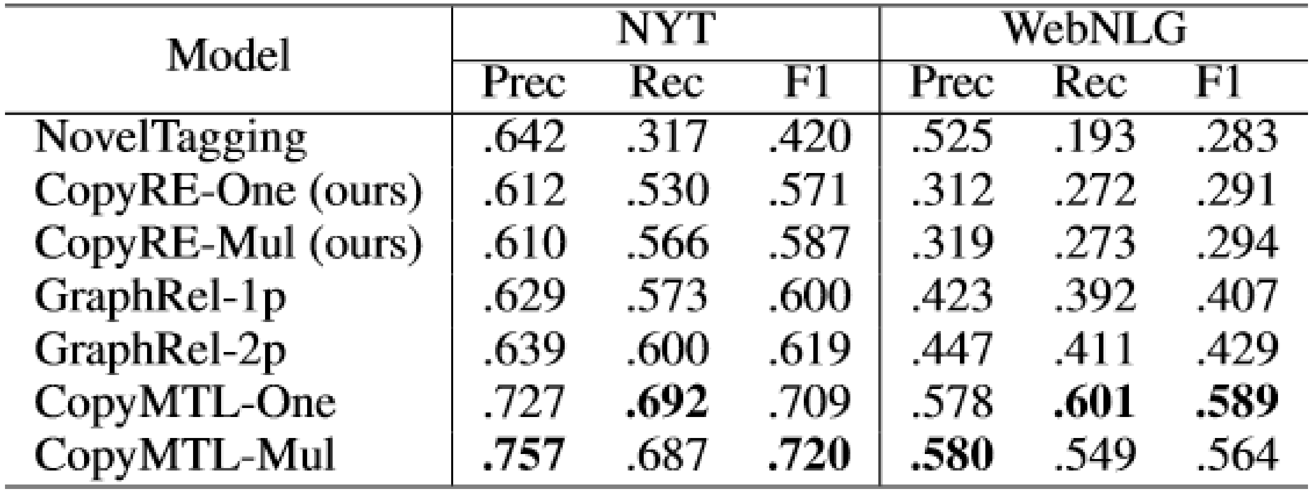
Dataset



Dataset	NYT	WebNLG
Relation types	24	246
Dictionary size	90760	5928
Train sentence	56195	5019
Test sentence	5000	703

Table 2: Statistics of the datasets.

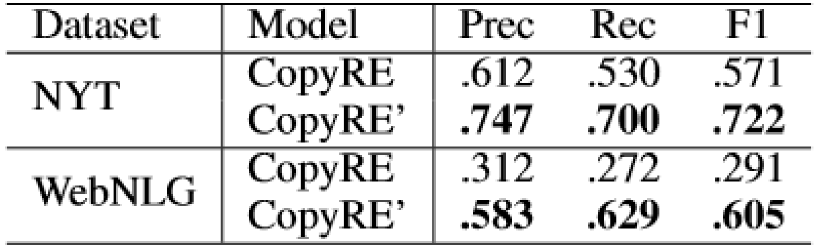
Experiment



Model	NYT			WebNLG		
	Prec	Rec	F1	Prec	Rec	F1
NovelTagging	.642	.317	.420	.525	.193	.283
CopyRE-One (ours)	.612	.530	.571	.312	.272	.291
CopyRE-Mul (ours)	.610	.566	.587	.319	.273	.294
GraphRel-1p	.629	.573	.600	.423	.392	.407
GraphRel-2p	.639	.600	.619	.447	.411	.429
CopyMTL-One	.727	**.692**	.709	.578	**.601**	**.589**
CopyMTL-Mul	**.757**	.687	**.720**	**.580**	.549	.564

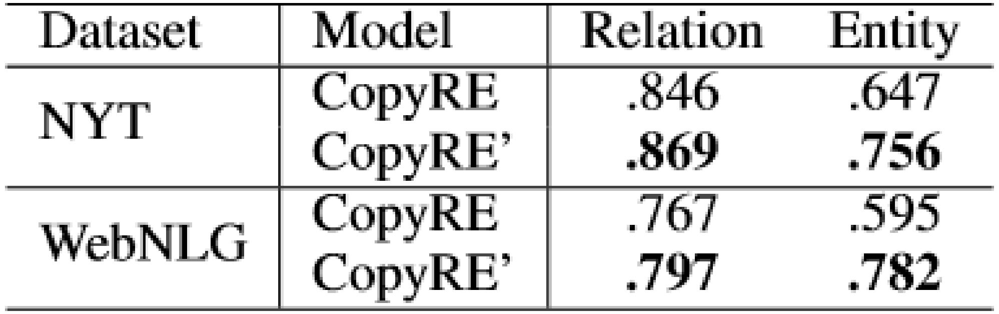
Table 1: Results of the compared models on NYT and WebNLG, in which CopyRE uses less strict evaluation.

Experiment



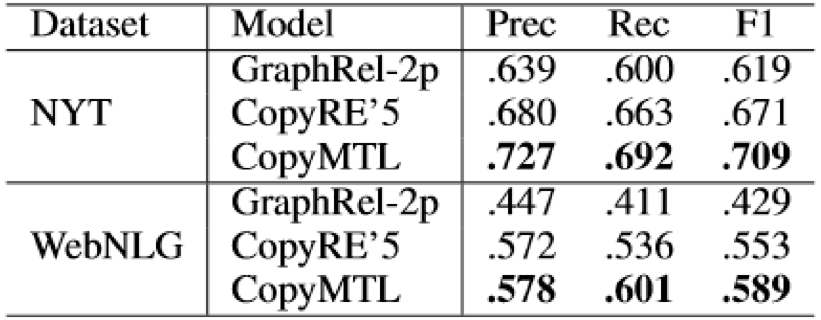
Dataset	Model	Prec	Rec	F1
NYT	CopyRE	.612	.530	.571
	CopyRE’	**.747**	**.700**	**.722**
WebNLG	CopyRE	.312	.272	.291
	CopyRE’	**.583**	**.629**	**.605**

Table 3: Results of CopyRE and CopyRE’ on NYT and WebNLG. These models do not consider entities with multiple tokens and use less strict evaluation that ignores entity with multiple tokens.



Dataset	Model	Relation	Entity
NYT	CopyRE	.846	.647
	CopyRE’	**.869**	**.756**
WebNLG	CopyRE	.767	.595
	CopyRE’	**.797**	**.782**

Table 4: F1 scores on subtasks.



Dataset	Model	Prec	Rec	F1
NYT	GraphRel-2p	.639	.600	.619
	CopyRE’5	.680	.663	.671
	CopyMTL	**.727**	**.692**	**.709**
WebNLG	GraphRel-2p	.447	.411	.429
	CopyRE’5	.572	.536	.553
	CopyMTL	**.578**	**.601**	**.589**

Table 5: Results of different multi-token models on NYT and WebNLG