Document-Level Relation Extraction with Adaptive Thresholding and Localized Context Pooling

Wenxuan Zhou[1], Kevin Huang[2], Tengyu Ma[3], Jing Huang[2]

University of Southern California[1], JD AI Research[2], Stanford University[3]

USC Viterbi School of Engineering

JDAI Research

Stanford

Document-Level Relation Extraction

John Stanistreet was an Australian politician. He was born in Bendigo to legal manager John Jepson Stanistreet and Maud McIlroy. (…4 sentences…) In 1955 John Stanistreet was elected to the Victorian Legislative Assembly as the Liberal and Country Party member for Bendigo. Stanistreet died in Bendigo in 1971.

Subject: John Stanistreet **Object**: Bendigo

Relation: place of birth; place of death

Goal: identify the relationships between the subject and object entities.

Challenges

John Stanistreet was an Australian politician. He was born in Bendigo to legal manager John Jepson Stanistreet and Maud McIlroy. (…4 sentences…) In 1955 John Stanistreet was elected to the Victorian Legislative Assembly as the Liberal and Country Party member for Bendigo. Stanistreet died in Bendigo in 1971.

Subject: John Stanistreet **Object**: Bendigo

Relation: place of birth; place of death

Document-level RE
(DocRED)

Billy Mays, the bearded, boisterous pitchman who, as the undisputed king of TV yell and sell, became an unlikely pop culture icon, died at his home in Tampa, Fla, on Sunday.

Subject: Billy Mays **Object**: Tampa

Relation: city_of_death

Sentence-level RE
(TACRED)

Sentence-level RE (TACRED, SemEval 2010): mention-level, one entity pair, single-label.

Document-level RE (DocRED, CDR, GDA): entity-level, multiple entity pairs, can be multi-label.

Challenges: Multi-entity

For document-level RE, one document contains multiple entity pairs, and one entity has multiple mentions.

Problems:

1. For a specific entity pair, only some of their mentions/context are relevant.
2. For one entity in different pairs, the relevant mentions/context may be different.

Challenges: Multi-label

One entity pair may be associated with multiple relations. In DocRED, 7% of entity pairs have more than 1 label.

Current approach: reduce the problem to binary classification.

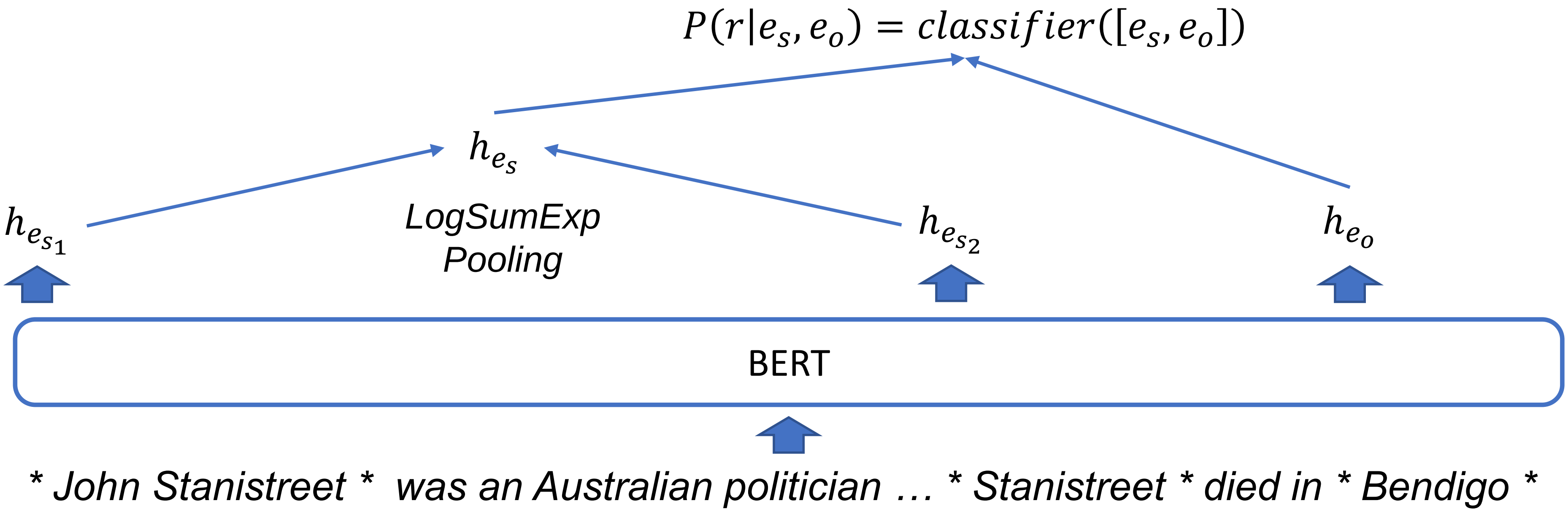
Problems:

1. Binary classification ignores the dependencies among classes.
2. The predicted classes are obtained by applying a heuristic threshold to prediction scores. However, the prediction scores are not calibrated, thus one global threshold does not suffice.

Contributions

1. We propose localized context pooling, which transfers pre-trained attention to identify relevant context that is relevant to entity pairs.
2. We propose adaptive-thresholding loss, which enables the learning of an adaptive threshold that is dependent on entity pairs.
3. Experiments on three public document-level relation extraction datasets demonstrate that our ATLOP model achieves state-of-the-art performance.

Base Model



Mention-level embedding:

- Insert a “*” symbol before and after each entity mention.
- Take the embedding of “*” before the mention as mention-level embedding.

Entity-level embedding: for entities that have multiple mentions, we use logsumexp pooling to aggregate the entity mentions:

$$h_e = \log\left(\Sigma_j \exp h_{e_j}\right)$$

Base Model (cont.)

Classifier: given entity embedding h_{e_s} and h_{e_o}, we fist map them to task-specific representation z:

$$z_s = \tanh(W_s h_{e_s})$$
$$z_o = \tanh(W_o h_{e_o})$$

Then we use grouped bilinear layer to get class probability:

$$[z_s^1, \ldots, z_s^k] = z_s$$
$$[z_o^1, \ldots, z_o^k] = z_o$$
$$P(r|e_s, e_o) = \sigma\left(\sum_{i=1}^{k} z_s^i W_r^i z_o^i + b_r\right)$$

Localized Context Pooling

The relevant mentions/context may be different for different entity pairs.

Intuition: the attention in pre-trained language models (BERT) captures relevant context for each token, we can use the attention to help determine the relevant context for both entities.

For two tokens i, j, a token k is important to both tokens if both $a_{i \rightarrow k}$ and $a_{j \rightarrow k}$ are high, thus we can use $a_i \cdot a_j$ to locate important tokens.

Localized Context Pooling (cont.)

Given an attention matrix A from the pre-trained language model, we use the attention of "*" at the start of mentions as the mention-level attention, and average mention-level attentions of the same entity as the entity-level attention A^E. Then we can obtain the localized context by:

$$A^{(s,o)} = A_s^E \cdot A_o^E$$

$$q^{(s,o)} = \sum_{i=1}^{H} A_i^{(s,o)} \text{ (average attention heads)}$$

$$a^{(s,o)} = q^{(s,o)} / 1^T q^{(s,o)} \text{ (normalize to 1)}$$

$$c^{(s,o)} = H^T a^{(s,o)}$$

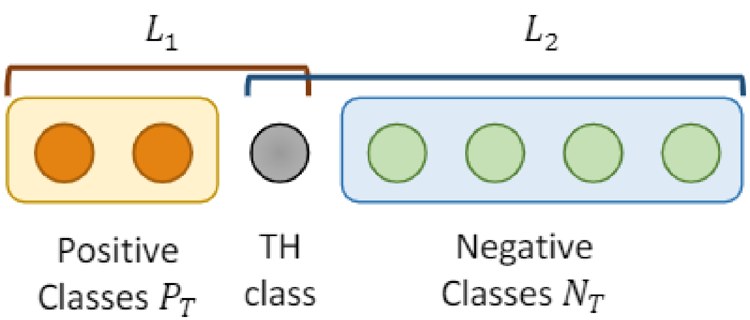
We add the localized context to the entity pair representation by:

$$z_s^{(s,o)} = \tanh(W_s h_{e_s} + W_{c_1} c^{(s,o)})$$

$$z_o^{(s,o)} = \tanh(W_o h_{e_o} + W_{c_2} c^{(s,o)})$$

Adaptive Thresholding

The class probability is not calibrated so the same probability does not mean the same for all pairs, thus we propose to use a learnable adaptive threshold.



P_T: positive classes.
N_T: negative classes.
TH: adaptive threshold.

We should have:

$$P(r \in P_T) > P(TH) > P(r \in N_T)$$

Then in inference, we return classes that have higher probability than TH as positive classes.

Adaptive Thresholding (cont.)

$$\mathcal{L}_1 = -\sum_{r \in \mathcal{P}_T} \log \left(\frac{\exp(\mathrm{logit}_r)}{\sum_{r' \in \mathcal{P}_T \cup \{\mathrm{TH}\}} \exp(\mathrm{logit}_{r'})} \right),$$

$$\mathcal{L}_2 = -\log \left(\frac{\exp(\mathrm{logit}_{\mathrm{TH}})}{\sum_{r' \in \mathcal{N}_T \cup \{\mathrm{TH}\}} \exp(\mathrm{logit}_{r'})} \right),$$

$$\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2.$$

L_1: positive classes have higher logits than TH.
L_2: TH has higher logits than negative classes.

Experiments: Main Results

We test our model on three document-level RE datasets DocRED, CDR and GDA.

Model	**Dev**		**Test**	
	Ign F_1	F_1	Ign F_1	F_1
Sequence-based Models				
CNN (Yao et al. 2019)	41.58	43.45	40.33	42.26
BiLSTM (Yao et al. 2019)	48.87	50.94	48.78	51.06
Graph-based Models				
BiLSTM-AGGCN (Guo, Zhang, and Lu 2019)	46.29	52.47	48.89	51.45
BiLSTM-LSR (Nan et al. 2020)	48.82	55.17	52.15	54.18
BERT-LSR$_{\text{BASE}}$ (Nan et al. 2020)	52.43	59.00	56.97	59.05
Transformer-based Models				
BERT$_{\text{BASE}}$ (Wang et al. 2019a)	-	54.16	-	53.20
BERT-TS$_{\text{BASE}}$ (Wang et al. 2019a)	-	54.42	-	53.92
HIN-BERT$_{\text{BASE}}$ (Tang et al. 2020a)	54.29	56.31	53.70	55.60
CorefBERT$_{\text{BASE}}$ (Ye et al. 2020)	55.32	57.51	54.54	56.96
CorefRoBERTa$_{\text{LARGE}}$ (Ye et al. 2020)	57.35	59.43	57.90	60.25
Our Methods				
BERT$_{\text{BASE}}$ (our implementation)	54.27 ± 0.28	56.39 ± 0.18	-	-
BERT-E$_{\text{BASE}}$	56.51 ± 0.16	58.52 ± 0.19	-	-
BERT-ATLOP$_{\text{BASE}}$	59.22 ± 0.15	61.09 ± 0.16	59.31	61.30
RoBERTa-ATLOP$_{\text{LARGE}}$	$\mathbf{61.32 \pm 0.14}$	$\mathbf{63.18 \pm 0.19}$	**61.39**	**63.40**

DocRED

Model	CDR	GDA
BRAN (Verga, Strubell, and McCallum 2018)	62.1	-
CNN (Nguyen and Verspoor 2018)	62.3	-
EoG (Christopoulou, Miwa, and Ananiadou 2019)	63.6	81.5
LSR (Nan et al. 2020)	64.8	82.2
SciBERT (our implementation)	65.1 ± 0.6	82.5 ± 0.3
SciBERT-E	65.9 ± 0.5	83.3 ± 0.3
SciBERT-ATLOP	$\mathbf{69.4 \pm 1.1}$	$\mathbf{83.9 \pm 0.2}$

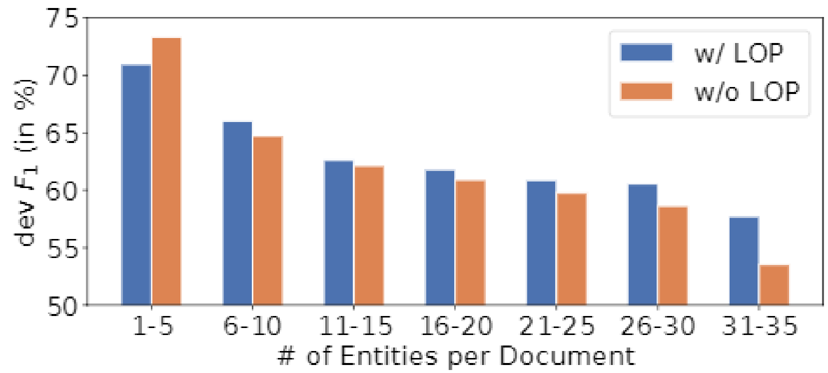
CDR and GDA

Our model achieves SOTA performance on all datasets.

Experiments: Ablation Study

Model	Ign F_1	F_1
BERT-ATLOP$_{\text{BASE}}$	59.22	61.09
– Adaptive Thresholding	58.32	60.20
– Localized Context Pooling	58.19	60.12
– Adaptive-Thresholding Loss	39.52	41.74
BERT-E$_{\text{BASE}}$	56.51	58.52
– Entity Marker	56.22	58.28
– Group Bilinear	55.51	57.54
– Logsumexp Pooling	55.35	57.40

Strategy	Dev F_1	Test F_1
Global Thresholding	60.14	60.62
Per-class Thresholding	**61.73**	60.35
Adaptive Thresholding	61.27	**61.30**



1. Both adaptive thresholding and localized context pooling are effective.
2. Adaptive thresholding performs better than both global thresholding and per-class thresholding.
3. Local context pooling is more effective for documents containing many entities.

Conclusion

- We propose two novel techniques, adaptive thresholding and localized context pooling.
- Our model achieves SOTA performance on three document-level RE datasets.
- Code released at https://github.com/wzhouad/ATLOP