Datawhale Paper Show

Document-Level RE with Adaptive Thresholding and Localized Context Pooling -AAAI2021

Reporter : 陈海顺 2021.03.27

Multi-Class vs. Multi-Label

看不见的客人 Contratiempo (2016)

导演: 奥里奥尔·保罗
编剧: 奥里奥尔·保罗
主演: 马里奥·卡萨斯 / 阿娜·瓦格纳 / 何塞·科罗纳多 / 巴巴拉·莱涅 / 弗兰塞斯克·奥雷利亚 / 更多...
类型: 剧情 / 悬疑 / 惊悚 / 犯罪
制片国家/地区: 西班牙
语言: 西班牙语
上映日期: 2017-09-15(中国大陆) / 2016-09-23(奇幻电影节) / 2017-01-06(西班牙)
片长: 106分钟
又名: 死无对证(港) / 布局(台) / The Invisible Guest

Multi-Label

Multi-Class

- **Multi-label classification**

  Multi-label classification is a predictive modeling task that involves predicting zero or more mutually class labels.

- **Multi-class classification**

  Multiclass classification is the problem of classifying instances into one of three or more classes.

Multi-Label Problem in document RE

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; Relation 1

place of death Relation 2

One entity with multiple possible relations

One entity pair occurs multiple times associated with multiple possible relations.

Document with multiple entity in different context

One document contains multiple entity pairs in different should not be the same representation

Multi-Label Problem in document RE

- A common baseline

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

- Encoder

$$d = [x_t]_{t=1}^{l}$$

$$[\boldsymbol{h}_1, \boldsymbol{h}_2, ..., \boldsymbol{h}_l] = \mathrm{BERT}([x_1, x_2, ..., x_l]).$$

$$h_{e_i} = \sum_{j=1}^{N_{e_i}} h_{m_j^i}$$

- Binary classification

$$\boldsymbol{z}_s = \tanh\left(\boldsymbol{W}_s \boldsymbol{h}_{e_s}\right),$$

$$\boldsymbol{z}_o = \tanh\left(\boldsymbol{W}_o \boldsymbol{h}_{e_o}\right),$$

$$P(r \mid e_s, e_o) = \mathrm{sigmoid}\left(z_s^T W_r z_o + b_r\right)$$

Multi-Label Problem in document RE

- An enhanced baseline

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

- Encoder

$$[\boldsymbol{h}_1, \boldsymbol{h}_2, ..., \boldsymbol{h}_l] = \mathrm{BERT}([x_1, x_2, ..., x_l]).$$

$$\boxed{\boldsymbol{h}_{e_i} = \log \sum_{j=1}^{N_{e_i}} \exp\left(\boldsymbol{h}_{m_j^i}\right).}$$

- Binary classification

$$\boldsymbol{z}_s = \tanh\left(\boldsymbol{W}_s \boldsymbol{h}_{e_s}\right),$$
$$\boldsymbol{z}_o = \tanh\left(\boldsymbol{W}_o \boldsymbol{h}_{e_o}\right),$$

$$\boxed{\begin{aligned}\left[\boldsymbol{z}_s^1; ...; \boldsymbol{z}_s^k\right] &= \boldsymbol{z}_s, \\ \left[\boldsymbol{z}_o^1; ...; \boldsymbol{z}_o^k\right] &= \boldsymbol{z}_o,\end{aligned}}$$

$$\mathrm{P}\left(r|e_s, e_o\right) = \sigma\left(\sum_{i=1}^{k} \boldsymbol{z}_s^{i\top} \boldsymbol{W}_r^i \boldsymbol{z}_o^i + b_r\right)$$

Multi-Label Problem in document RE

- Encoder

$$[\boldsymbol{h}_1, \boldsymbol{h}_2, ..., \boldsymbol{h}_l] = \text{BERT}([x_1, x_2, ..., x_l]).$$

$$\boxed{\boldsymbol{h}_{e_i} = \log \sum_{j=1}^{N_{e_i}} \exp\left(\boldsymbol{h}_{m_j^i}\right).}$$

Entity pairs should not based on the **same contextual embedding** !!

- Binary classification

$$\boldsymbol{z}_s = \tanh\left(\boldsymbol{W}_s \boldsymbol{h}_{e_s}\right),$$

$$\boldsymbol{z}_o = \tanh\left(\boldsymbol{W}_o \boldsymbol{h}_{e_o}\right),$$

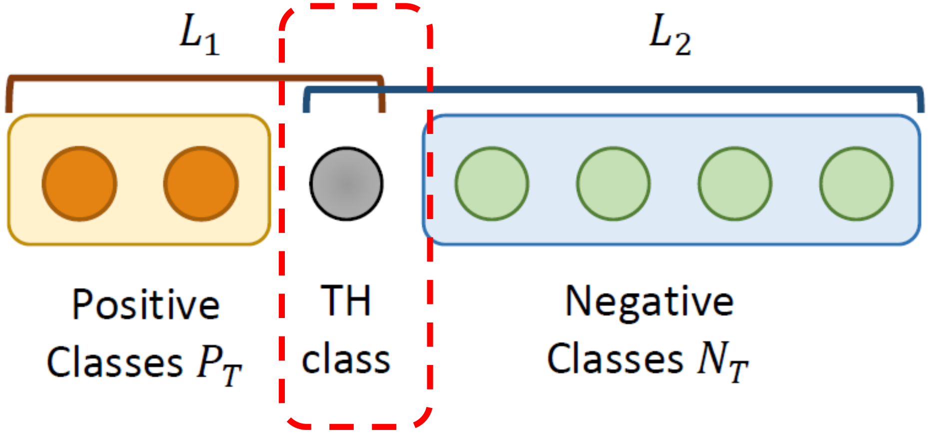
$$\boxed{P(r \mid e_s, e_o) = \text{sigmoid}\left(z_s^T W_r z_o + b_r\right)}$$

The model may have different confidence for different entity pairs or classes and **one global threshold** does not suffice !!

Document-Level RE with (1) Adaptive Thresholding and (2) Localized Context Pooling

Adaptive Thresholding

| Adaptive Thresholding



Learnable and Adaptive

- **A learnable, adaptive thresholding**

The logits of positive labels should be higher than threshold while those negative labels should be lower

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

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

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

- **Reduce decision errors during inference**

Adaptive Thresholding

- A learnable, adaptive thresholding

The logits of positive labels should be higher than threshold while those negative labels should be lower

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

Push the logits of all positive labels to be higher than the TH class.

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

Pull the logits of negative labels to be lower than the TH class.

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

Localized Context Pooling

| Localized Context Pooling

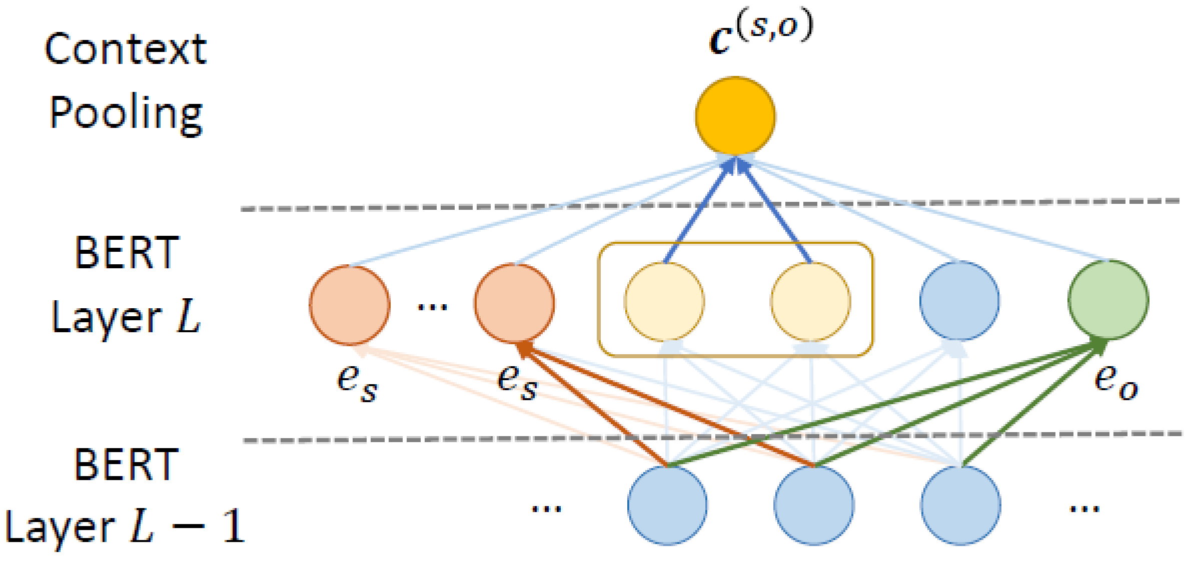


Illustration of localized context pooling

- Entity-level attention

$\boldsymbol{A}_{ijk,1\le i\le H,1\le j,k\le l}$ Token-level

↓

$\boldsymbol{A}_{j=j'}$ Mention-level

↓ Average

$\boldsymbol{A}_i^E \in \mathbb{R}^{H\times l}$ Entity-level

Localized Context Pooling

- Relevant context aggregation

$$
\begin{aligned}
\boldsymbol{A}^{(s,o)} &= \boldsymbol{A}_s^E \cdot \boldsymbol{A}_o^E, \\
\boldsymbol{q}^{(s,o)} &= \sum_{i=1}^{H} \boldsymbol{A}_i^{(s,o)}, \\
\boldsymbol{a}^{(s,o)} &= \boldsymbol{q}^{(s,o)} / \boldsymbol{1}^\intercal \boldsymbol{q}^{(s,o)}, \\
\boldsymbol{c}^{(s,o)} &= \boldsymbol{H}^\intercal \boldsymbol{a}^{(s,o)},
\end{aligned}
$$

Attention heads aggregation

Normalize to 1

Context aggregation

- Add the localized context to entity pair representation

$$
\begin{aligned}
\boldsymbol{z}_s^{(s,o)} &= \tanh\left(\boldsymbol{W}_s \boldsymbol{h}_{e_s} + \boldsymbol{W}_{c_1} \boldsymbol{c}^{(s,o)}\right), \\
\boldsymbol{z}_o^{(s,o)} &= \tanh\left(\boldsymbol{W}_o \boldsymbol{h}_{e_o} + \boldsymbol{W}_{c_2} \boldsymbol{c}^{(s,o)}\right),
\end{aligned}
$$

Experiment

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 et al., 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., 2019b)	-	54.16	-	53.20
BERT-TS$_{\text{BASE}}$ (Wang et al., 2019b)	-	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)	57.84	59.93	57.68	59.91
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**

Experiment

- 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

-20 ?

Experiment

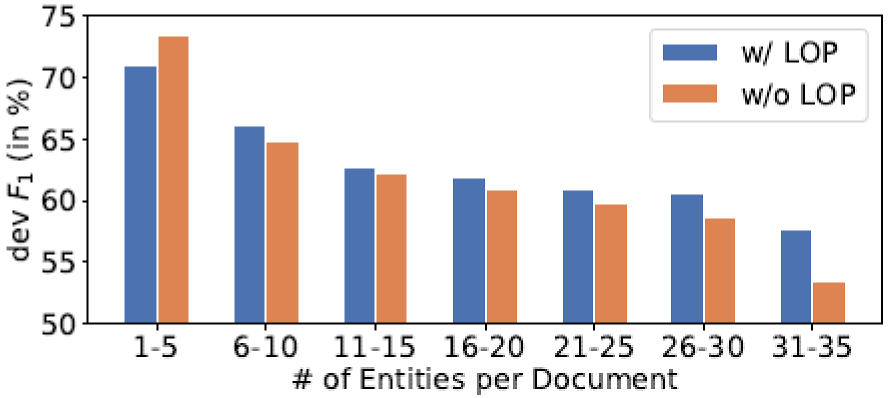


Figure 4: Dev F_1 score of documents with the different number of entities on DocRED. Our localized context pooling achieves better results when the number of entities is larger than 5. The improvement is more significant when the number of entities increases.

John Stanistreet was an Australian politician. He wa s born in Bendigo to legal manager John Jepson Stan istreet and Maud McIlroy. (… 4 sentences …) In 195 5 John Stanistreet was elected to the Victorian Legisl ative Assembly as the Liberal and Country Party mem ber for Bendigo, but he was defeated in 1958. Stanistr eet died in Bendigo in 1971.

Subject: John Stanistreet **Object**: Bendigo

Relation: place of birth; place of death

Figure 5: Context weights of an example from DocRED. We visualize the weight of context tokens $a^{(s,o)}$ in localized context pooling. The model attends to the most relevant context *born* and *died* for entity pair (*John Stanistreet*, *Bendigo*).

| Review

- Graph-based methods

  Graph-structured data may be flexible to model complex interactions.

  GNNs for preforming reasoning and inference

- PLMs-based methods

  A paradigm that has proven to be extremely successful for many NLP tasks

  Informative priors are provided for model in encoding text

Review

- Graph-based methods
  - Reasoning with Latent Structure Refinement for Document-Level Relation Extraction (ACL 2020)
  - Global-to-Local Neural Networks for Document-Level Relation Extraction (EMNLP 2020)
  - Double Graph Based Reasoning for Document-level Relation Extraction (EMNLP 2020)
  - Document-Level Relation Extraction with Reconstruction (AAAI 2021)
- PLMs-based methods
  - A Novel Document-Level Relation Extraction Method Based on BERT and Entity Information (IEEE Access)
  - Entity Structure Within and Throughout: Modeling Mention Dependencies for DocRED (AAAI 2021)
  - Multi-view Inference for Relation Extraction with Uncertain Knowledge(AAAI 2021)
  - Entity and Evidence Guided Relation Extraction for DocRED (Arxiv 2021)

Code & Source

DocRED 官方刷分榜单:

https://competitions.codalab.org/competitions/20717

DocRED 数据下载及Baseline:

https://github.com/thunlp/DocRED

Thanks!