组会

林佳兴

2020-10-31

EMNLP 2020

Global-to-Local Neural Networks for Document-Level Relation Extraction

Difeng Wang[†] **Wei Hu**[†, ‡, *] **Ermei Cao**[†] **Weijian Sun**[§]
[†] State Key Laboratory for Novel Software Technology, Nanjing University, China
[‡] National Institute of Healthcare Data Science, Nanjing University, China
[§] Huawei Technologies Co., Ltd.
{dfwang,emcao}.nju@gmail.com, whu@nju.edu.cn, sunweijian@huawei.com

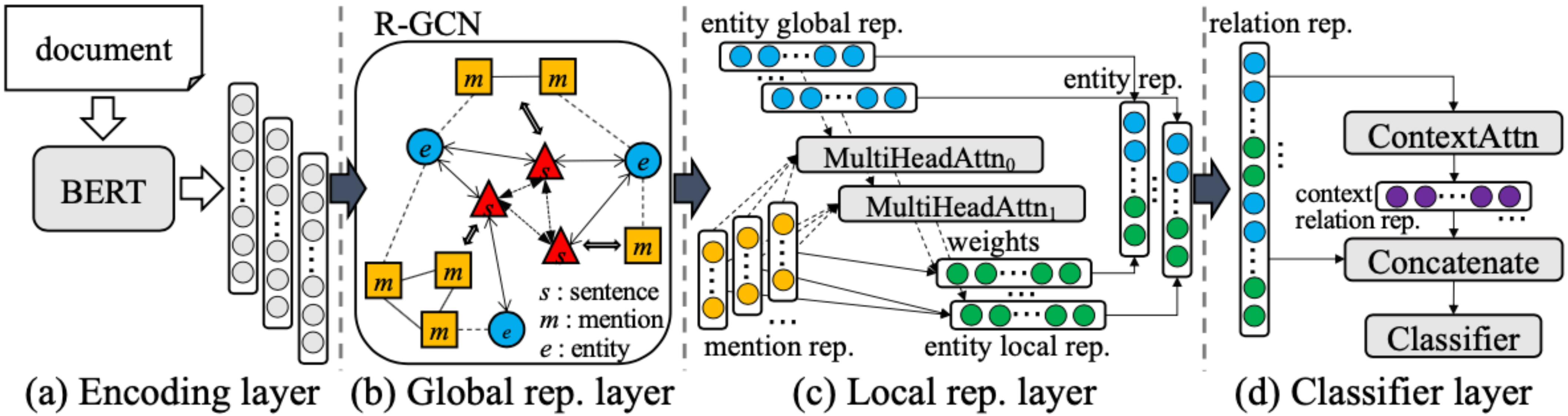
Motivation

- How to model the complex semantics of a document?
  - BERT   heterogeneous graph
- How to learn entity representations effectively?
  - Entity global representations by employing R-GCN
  - Entity local representations by aggregating multiple mentions of specific entities with multi-head attention
- How to leverage the influence from other relations?
  - Context relation representations with self-attention to make final relation prediction.

Contribution

- Synthesize entity global representations, entity local representations and context relation representations
- Demonstrate the superiority of GLRE compared with many state-of-the-art competitors.

Architecture



Global Representation

$$\mathbf{n}_i^{l+1} = \sigma\Big(\sum_{x \in \mathcal{X}} \sum_{j \in \mathcal{N}_i^x} \frac{1}{|\mathcal{N}_i^x|} \mathbf{W}_x^l \mathbf{n}_j^l + \mathbf{W}_0^l \mathbf{n}_i^l\Big),$$

Local Representation

$$
\begin{aligned}
\mathbf{e}_a^{\text{loc}} &= \text{LN}\big(\text{MHead}_0(\mathbf{e}_b^{\text{glo}}, \{\mathbf{n}_{s_i}\}_{s_i \in \mathcal{S}_a}, \{\mathbf{n}_{m_j}\}_{m_j \in \mathcal{M}_a})\big), \\
\mathbf{e}_b^{\text{loc}} &= \text{LN}\big(\text{MHead}_1(\mathbf{e}_a^{\text{glo}}, \{\mathbf{n}_{s_i}\}_{s_i \in \mathcal{S}_b}, \{\mathbf{n}_{m_j}\}_{m_j \in \mathcal{M}_b})\big),
\end{aligned}
$$

$$
\begin{aligned}
\text{MHead}(\mathcal{Q}, \mathcal{K}, \mathcal{V}) &= [\text{head}_1; \ldots; \text{head}_z]\mathbf{W}^{\text{out}}, \\
\text{head}_i &= \text{softmax}\Big(\frac{\mathcal{Q}\mathbf{W}_i^{\mathcal{Q}}(\mathcal{K}\mathbf{W}_i^{\mathcal{K}})'}{\sqrt{d_v}}\Big)\mathcal{V}\mathbf{W}_i^{\mathcal{V}},
\end{aligned}
$$

Context Relation Representation

$$\hat{\mathbf{e}}_a = [\mathbf{e}_a^{\text{glo}}; \mathbf{e}_a^{\text{loc}}; \mathbf{\Delta}(\delta_{ab})],$$
$$\hat{\mathbf{e}}_b = [\mathbf{e}_b^{\text{glo}}; \mathbf{e}_b^{\text{loc}}; \mathbf{\Delta}(\delta_{ba})],$$

$$\mathbf{o}_r = [\hat{\mathbf{e}}_a; \hat{\mathbf{e}}_b].$$

$$\mathbf{o}_c = \sum_{i=0}^{p} \theta_i \mathbf{o}_i = \sum_{i=0}^{p} \frac{\exp(\mathbf{o}_i \mathbf{W} \mathbf{o}_r')}{\sum_{j=0}^{p} \exp(\mathbf{o}_j \mathbf{W} \mathbf{o}_r')} \mathbf{o}_i,$$

Classifier

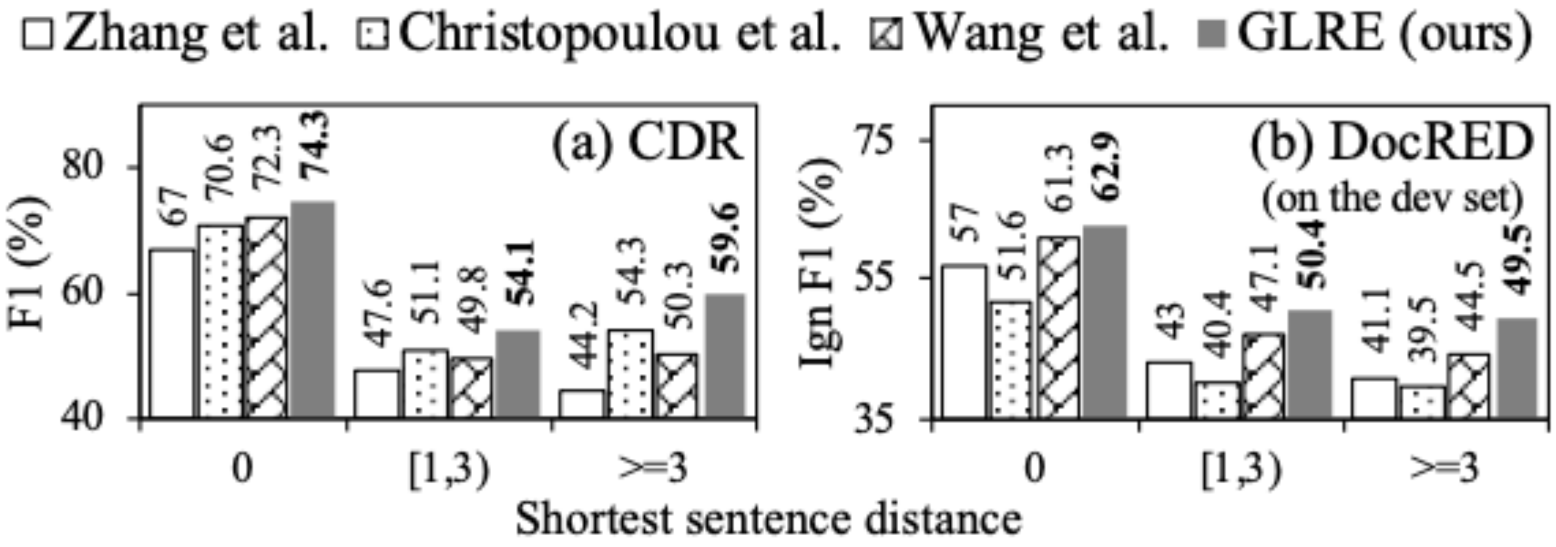
$$\mathbf{y}_r = \mathrm{sigmoid}(\mathrm{FFNN}([\mathbf{o}_r; \mathbf{o}_c])),$$

$$\mathcal{L} = -\sum_{r \in \mathcal{R}} \Big(y_r^* \log(y_r) + (1 - y_r^*) \log(1 - y_r) \Big),$$

Experiment

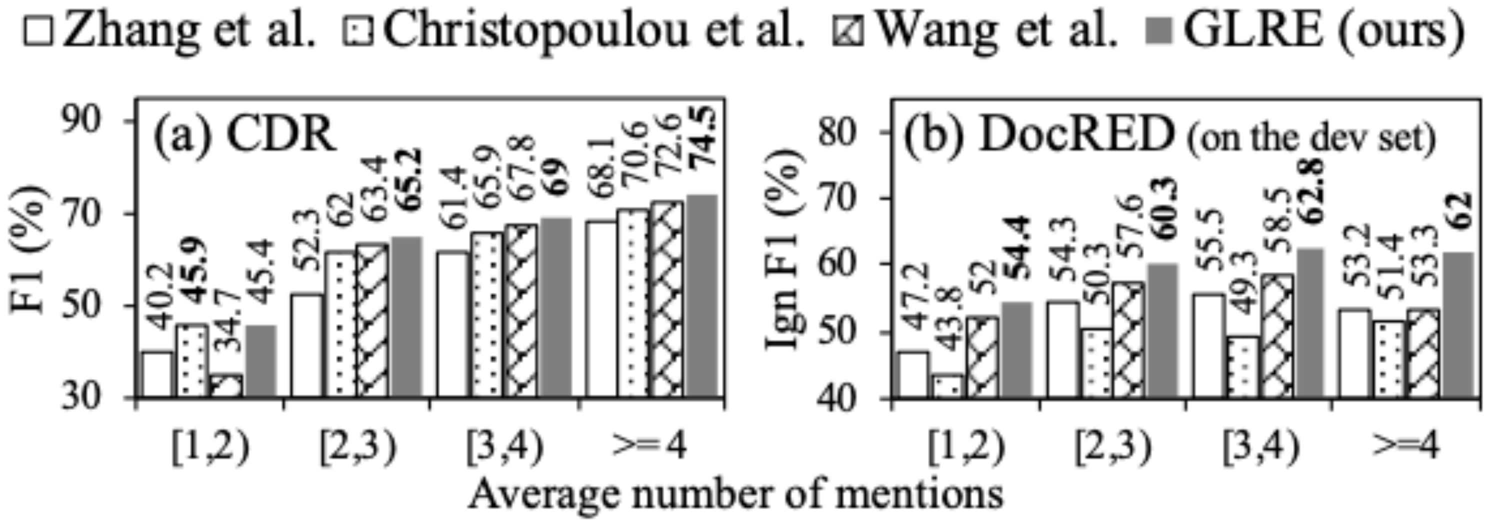
Models	Train		Train + Dev	
	Ign F1	F1	Ign F1	F1
Zhang et al.¶	49.9	52.1	52.5	54.6
Yao et al. (CNN)	40.3	42.3	-	-
Yao et al. (LSTM)	47.7	50.1	-	-
Yao et al. (BiLSTM)	48.8	51.1	-	-
Yao et al. (Context-aware)	48.4	50.7	-	-
Christopoulou et al.¶	49.1	50.9	48.3	50.4
Wang et al.¶	53.1	55.4	<u>54.5</u>	<u>56.5</u>
Tang et al.	<u>53.7</u>	<u>55.6</u>	-	-
GLRE (ours)	**55.4**	**57.4**	**56.7**	**58.9**

Experiment



Zhang et al.
Christopoulou et al.
Wang et al.
GLRE (ours)
(a) CDR
F1 (%)
80
60
40
67
70.6
72.3
74.3
47.6
51.1
49.8
54.1
44.2
54.3
50.3
59.6
0
[1,3)
>=3
(b) DocRED
(on the dev set)
Ign F1 (%)
75
55
35
57
51.6
61.3
62.9
43
40.4
47.1
50.4
41.1
39.5
44.5
49.5
Shortest sentence distance

Experiment



Zhang et al.
Christopoulou et al.
Wang et al.
GLRE (ours)
(a) CDR
F1 (%)
90
70
50
30
40.2
45.9
34.7
45.4
52.3
62
63.4
65.2
61.4
65.9
67.8
69
68.1
70.6
72.6
74.5
[1,2)
[2,3)
[3,4)
>=4
(b) DocRED (on the dev set)
Ign F1 (%)
80
70
60
50
40
47.2
43.8
52
54.4
54.3
50.3
57.6
60.3
55.5
49.3
58.5
62.8
53.2
51.4
53.3
62
[1,2)
[2,3)
[3,4)
>=4
Average number of mentions

Experiment

Models	**CDR**			**DocRED**	
	P	R	F1	Ign F1	F1
GLRE	65.1	72.2	**68.5**	**55.4**	**57.4**
w/o BERT	**69.6**	66.5	68.0	51.6	53.6
w/o Entity global rep.	67.0	65.4	66.2	54.7	56.6
w/o Entity local rep.	60.9	68.5	64.5	54.6	56.4
w/o Context rel. rep.	60.5	**75.1**	67.1	54.6	56.8

Experiment

GLRE	CDR			DocRED	
	P	R	F1	Ign F1	F1
BERT-Base	65.1	72.2	68.5	55.4	57.4
BERT-Large	65.3	72.3	**68.6**	**56.8**	58.9
XLNet-Large	**66.1**	70.5	68.2	**56.8**	**59.0**
ALBERT-xxLarge	57.5	**80.6**	67.1	56.3	58.3

Experiment

[S1] Conrad Oberon Johnson (November 15, 1915–February 3, 2008) was an American music educator, long associated with the city of Houston, who was inducted into the Texas Bandmasters Hall of Fame in 2000. [S2] Born in Victoria, Texas, Conrad Johnson was nine when his family moved to Houston. ...

Case 1 Label: country GLRE: country Wang et al.: N/A

[S1] Michael Imperioli (born March 26, 1966) is an American actor, writer and director best known for ... [S4] He was starring as Detective Louis Fitch in the ABC police drama Detroit 1-8-7 ... [S5] He wrote and directed his first feature film, The Hungry Ghosts, in 2008. ...

Case 2 Label: director GLRE: director Wang et al.: cast

[S1] The Pleistocene coyote (Canis latrans orcutti), also known as the Ice Age coyote, is an extinct subspecies of coyote that lived in western North America during the Late Pleistocene era. [S2] Most remains of the subspecies were found in southern California, though at least one was discovered in Idaho. ...

Case 3 Label: continent GLRE: continent Wang et al.: country

[S1] Operation Unified Resolve is an air and ground operation to flush out and trap al - Qaeda fighters hiding in the eastern Afghanistan provinces. [S2] Launched on 23 June 2003, Operation Unified Resolve is a joint operation between Pakistan, United States, and Afghanistan. [S3] Over 500 troops, mostly from the U.S. 82nd Airborne Division, began hunting the Taliban and al - Qaeda fighters in the provinces of Nangarhar and Kunar on Afghanistans eastern border. ...

Case 4 Label: country GLRE: country w/o local rep.: N/A

[S1] Eclipse is the third novel in the Twilight Saga by Stephenie Meyer ... [S3] Eclipse is preceded by New Moon and followed by Breaking Dawn. ... [S6] Eclipse was the fourth bestselling book of 2008, only behind Twilight, New Moon, and Breaking Dawn. ...

Case 5 Label: author GLRE: author w/o context rel.: creator

Arxiv 2020

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

Wenxuan Zhou[1], Kevin Huang[2], Tengyu Ma[3], Jing Huang[2]
[1]University of Southern California [2]JD AI Research [3]Stanford University
[1]zhouwenx@usc.edu [2]{kevin.huang, jing.huang}@jd.com [3]tengyuma@stanford.edu

Motivation

- Graph-based model和transformer-based model都有提取长距离信息的作用
- 简单使用transformer-based model会导致不同的实体对中使用的entity embedding是相同的，故需改进
- Enhance the entity embedding with additional context

Motivation

- 之前的方法是将多分类转化为多个二分类问题，test时使用global threshold
- Replace the global threshold with a learnable threshold class

Contribution

- Propose adaptive-thresholding loss
- Propose localized context pooling
- Achieve the new state-of-the-art performance on the three benchmark datasets

Encoder

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

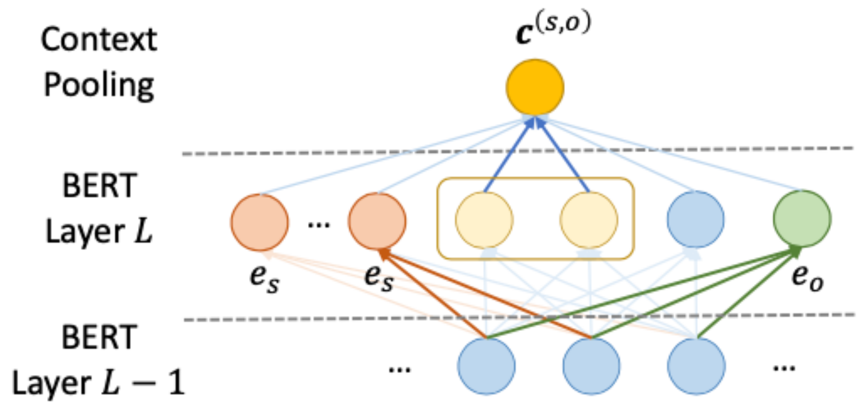
$$\boldsymbol{h}_{e_i} = \log \sum_{j=1}^{N_{e_i}} \exp\left(\boldsymbol{h}_{m_j^i}\right).$$

Classifier

$$
\begin{aligned}
\boldsymbol{z}_s &= \tanh\left(\boldsymbol{W}_s \boldsymbol{h}_{e_s}\right), \\
\boldsymbol{z}_o &= \tanh\left(\boldsymbol{W}_o \boldsymbol{h}_{e_o}\right), \\
\mathrm{P}\left(r \mid e_s, e_o\right) &= \sigma\left(\boldsymbol{z}_s^{\top} \boldsymbol{W}_r \boldsymbol{z}_o + b_r\right),
\end{aligned}
$$

$$
\begin{aligned}
\left[\boldsymbol{z}_s^1; \ldots; \boldsymbol{z}_s^k\right] &= \boldsymbol{z}_s, \\
\left[\boldsymbol{z}_o^1; \ldots; \boldsymbol{z}_o^k\right] &= \boldsymbol{z}_o, \\
\mathrm{P}\left(r \mid e_s, e_o\right) &= \sigma\left(\sum_{i=1}^{k} \boldsymbol{z}_s^{i\top} \boldsymbol{W}_r^i \boldsymbol{z}_o^i + b_r\right),
\end{aligned}
$$

Improved



$$\boldsymbol{A}^{(s,o)} = \boldsymbol{A}_s^E \cdot \boldsymbol{A}_o^E,$$

$$\boldsymbol{q}^{(s,o)} = \sum_{i=1}^{H} \boldsymbol{A}_i^{(s,o)},$$

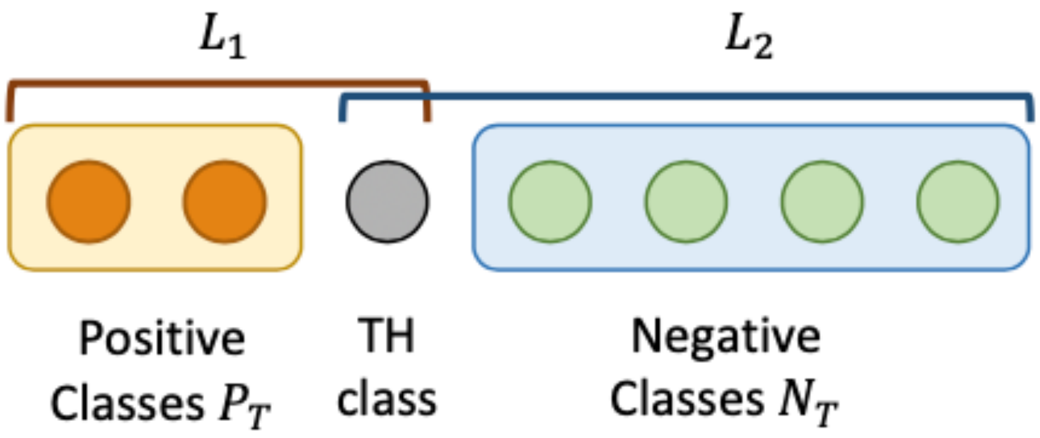
$$\boldsymbol{a}^{(s,o)} = \boldsymbol{q}^{(s,o)} / \mathbf{1}^\intercal \boldsymbol{q}^{(s,o)},$$

$$\boldsymbol{c}^{(s,o)} = \boldsymbol{H}^\intercal \boldsymbol{a}^{(s,o)},$$

$$\boldsymbol{z}_s^{(s,o)} = \tanh\left(\boldsymbol{W}_s \boldsymbol{h}_{e_s} + \boldsymbol{W}_{c_1} \boldsymbol{c}^{(s,o)}\right),$$

$$\boldsymbol{z}_o^{(s,o)} = \tanh\left(\boldsymbol{W}_o \boldsymbol{h}_{e_o} + \boldsymbol{W}_{c_2} \boldsymbol{c}^{(s,o)}\right),$$

Adaptive threshold



Adaptive threshold

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

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

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

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

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

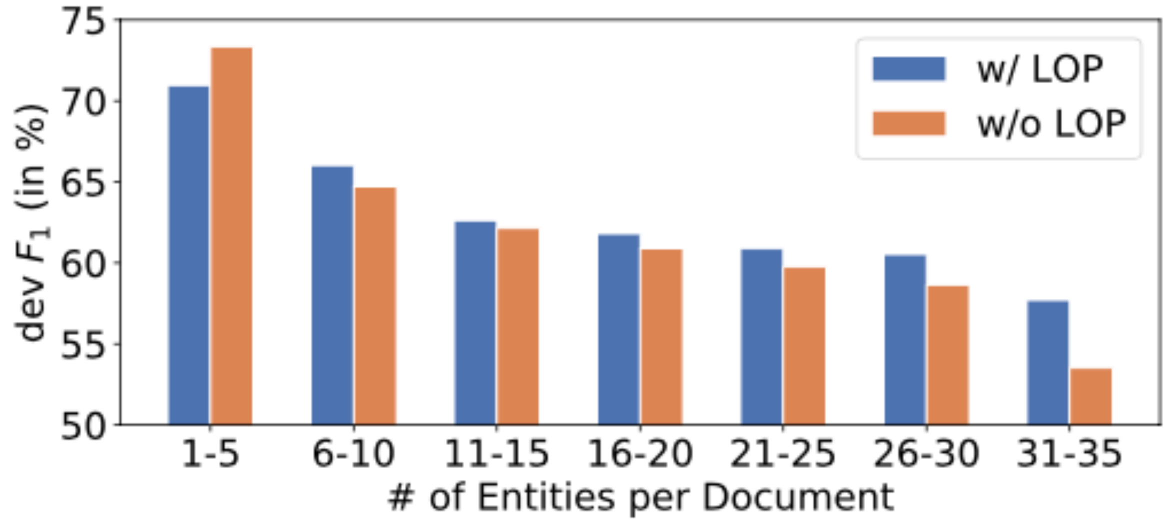
Experiment

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

Experiment

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**

Experiment



75
70
65
60
55
50
dev F_1 (in %)
w/ LOP
w/o LOP
1-5
6-10
11-15
16-20
21-25
26-30
31-35
# of Entities per Document

Experiment

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

Thank you