Why Attention? Analyze BiLSTM Deficiency and Its Remedies in the Case of NER

Peng-Hsuan Li jacobvsdanniel@iis.sinica.edu.tw

Tsu-Jui Fu tsu-juifu@ucsb.edu

Wei-Yun Ma ma@iis.sinica.edu.tw

XOR Limitation for BiLSTM

1. Key and Peele (*work-of-art*)
2. You and I (*work-of-art*)
3. Key and I
4. You and Peele

$$s_t = H_t W_p + b$$
$$= \overrightarrow{H}_t \overrightarrow{W}_p + \overleftarrow{H}_t \overleftarrow{W}_p + b$$
$$= \overrightarrow{s}_t + \overleftarrow{s}_t + b$$

i: *inside-work-of-art*
j: *outside*

$$\overrightarrow{s}^1_{2i} + \overleftarrow{s}^1_{2i} + b_i > \overrightarrow{s}^1_{2j} + \overleftarrow{s}^1_{2j} + b_j$$
$$\overrightarrow{s}^2_{2i} + \overleftarrow{s}^2_{2i} + b_i > \overrightarrow{s}^2_{2j} + \overleftarrow{s}^2_{2j} + b_j$$
$$\overrightarrow{s}^3_{2i} + \overleftarrow{s}^3_{2i} + b_i < \overrightarrow{s}^3_{2j} + \overleftarrow{s}^3_{2j} + b_j$$
$$\overrightarrow{s}^4_{2i} + \overleftarrow{s}^4_{2i} + b_i < \overrightarrow{s}^4_{2j} + \overleftarrow{s}^4_{2j} + b_j$$

Left	Right	Tag of "and"
0	0	0
1	1	0
0	1	1
1	0	1

$$\overrightarrow{s}^1_{2i} + \overleftarrow{s}^1_{2i} + b_i > \overrightarrow{s}^1_{2j} + \overleftarrow{s}^1_{2j} + b_j$$
$$\overrightarrow{s}^2_{2i} + \overleftarrow{s}^2_{2i} + b_i > \overrightarrow{s}^2_{2j} + \overleftarrow{s}^2_{2j} + b_j$$
$$\overrightarrow{s}^1_{2i} + \overleftarrow{s}^2_{2i} + b_i < \overrightarrow{s}^1_{2j} + \overleftarrow{s}^2_{2j} + b_j$$
$$\overrightarrow{s}^2_{2i} + \overleftarrow{s}^1_{2i} + b_i < \overrightarrow{s}^2_{2j} + \overleftarrow{s}^1_{2j} + b_j$$

XOR Limitation for CRF-BiLSTM

1. $a\_o\ m\_s\ c\_o$
2. $b\_o\ m\_s\ d\_o$
3. $a\_o\ m\_o\ d\_o$
4. $b\_o\ m\_o\ c\_o$

Viterbi with softmax (*p*) and transition (*q*)

$$lp(\_s|am)+lp(s\_|mc)+lq(oso) > lp(\_o|am)+lp(o\_|mc)+lq(ooo)$$
$$lp(\_s|bm)+lp(s\_|md)+lq(oso) > lp(\_o|bm)+lp(o\_|md)+lq(ooo)$$
$$lp(\_s|am)+lp(s\_|md)+lq(oso) < lp(\_o|am)+lp(o\_|md)+lq(ooo)$$
$$lp(\_s|bm)+lp(s\_|mc)+lq(oso) < lp(\_o|bm)+lp(o\_|mc)+lq(ooo)$$

Viterbi, suppose no transition

$$p(oso|amc) > p(ooo|amc)$$
$$p(oso|bmd) > p(ooo|bmd)$$
$$p(oso|amd) < p(ooo|amd)$$
$$p(oso|bmc) < p(ooo|bmc)$$

$$lp(\_s\_|amc) > lp(\_o\_|amc)$$
$$lp(\_s\_|bmd) > lp(\_o\_|bmd)$$
$$lp(\_s\_|amd) < lp(\_o\_|amd)$$
$$lp(\_s\_|bmc) < lp(\_o\_|bmc)$$

$$lp(\_s|am) + lp(s\_|mc) > lp(\_o|am) + lp(o\_|mc)$$
$$lp(\_s|bm) + lp(s\_|md) > lp(\_o|bm) + lp(o\_|md)$$
$$lp(\_s|am) + lp(s\_|md) < lp(\_o|am) + lp(o\_|md)$$
$$lp(\_s|bm) + lp(s\_|mc) < lp(\_o|bm) + lp(o\_|mc)$$

BiLSTM

$$\overrightarrow{H} = \overrightarrow{LSTM}_2(\overrightarrow{LSTM}_1(X))$$
$$\overleftarrow{H} = \overleftarrow{LSTM}_4(\overleftarrow{LSTM}_3(X))$$
$$H = \overrightarrow{H} \,||\, \overleftarrow{H},$$
$$s_t = H_t W_p + b$$

-> Cross

$$\overrightarrow{H}^1 = \overrightarrow{LSTM}_1(X)$$
$$\overleftarrow{H}^3 = \overleftarrow{LSTM}_3(X)$$
$$\overrightarrow{H}^2 = \overrightarrow{LSTM}_2(\overrightarrow{H}^1 || \overleftarrow{H}^3)$$
$$\overleftarrow{H}^4 = \overleftarrow{LSTM}_4(\overrightarrow{H}^1 || \overleftarrow{H}^3)$$
$$H = \overrightarrow{H}^2 \,||\, \overleftarrow{H}^4$$

-> Att

$$\alpha^i = \sigma\left(\frac{HW^{qi}(HW^{ki})^T}{\sqrt{d_c}}\right)$$
$$C^i = \alpha^i HW^{vi},$$
$$s^c_t = (H_t||C^1_t||C^2_t||...||C^m_t)W_c + b$$

Results

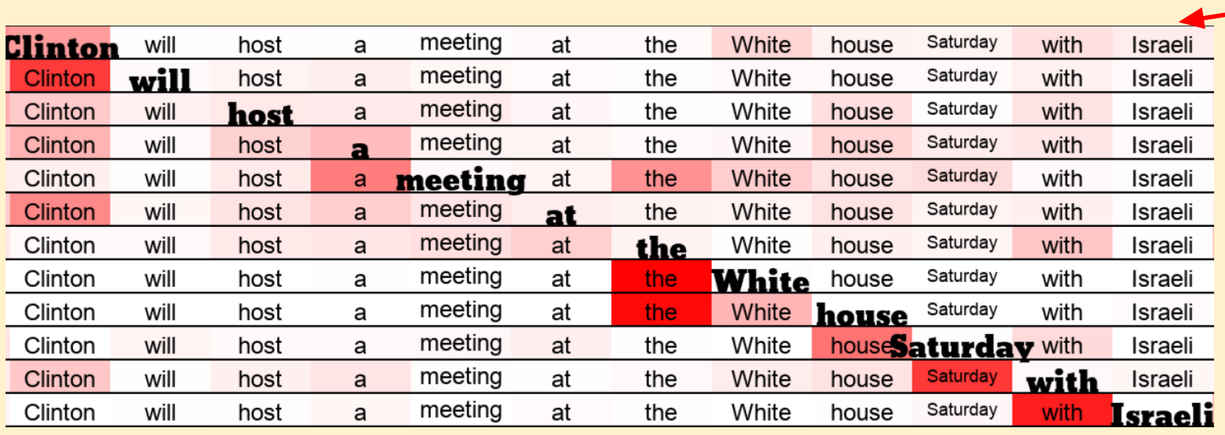
	OntoNotes 5.0			WNUT 2017		
	Prec.	Rec.	F1	Prec.	Rec.	F1
BiLSTM-CNN	86.04	86.53	86.28±0.26	-	-	-
CRF-IDCNN	-	-	86.84±0.19	-	-	-
CRF-BiLSTM(-BiLSTM*)	-	-	86.99±0.22	-	-	38.24
Baseline-BiLSTM-CNN	88.37	87.14	87.75±0.14	53.24	32.93	40.68±1.78
Cross-BiLSTM-CNN	88.37	**88.17**	88.27±0.17	**58.28**	33.92	**42.85**±0.99
Att-BiLSTM-CNN	**88.71**	88.11	**88.40**±0.18	55.82	**34.08**	42.26±0.82

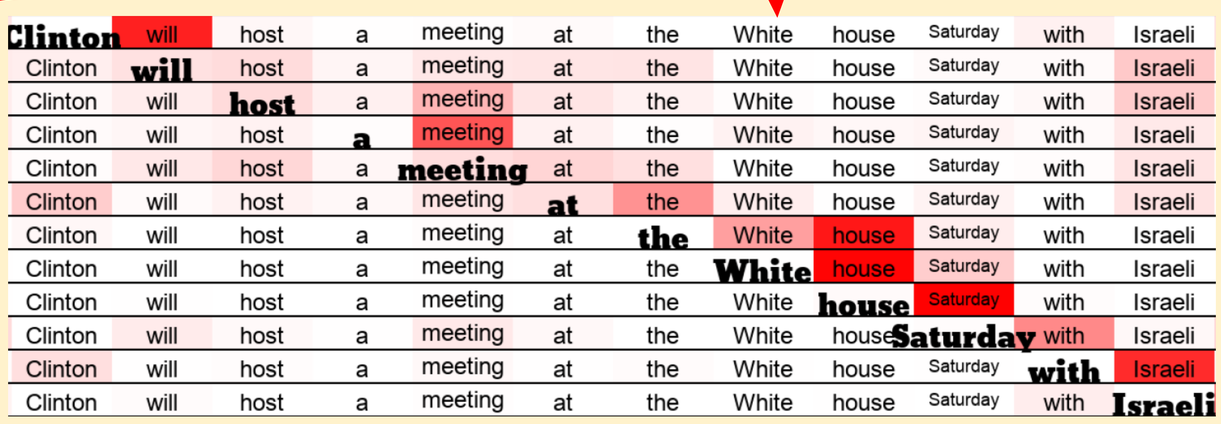
-> By Length

	OntoNotes 5.0				WNUT 2017			
	1	2	3	3+	1	2	3	3+
Cross	+0.3%	+0.6%	+1.8%	+1.3%	+1.7%	+2.9%	+8.7%	+5.4%
Att	+0.1%	+1.1%	+2.3%	+1.8%	+1.5%	+2.0%	+2.6%	+0.9%

-> By Component

	Att-BiLSTM-CNN								Baseline-...
	HC^{all}	H	C^{all}	C^1	C^2	C^3	C^4	C^5	H
O	99.05	-1.68	0.75	0.95	-1.67	-45.57	-0.81	-35.46	-0.03
S	93.74	2.69	-91.02	-90.56	-90.88	-25.61	-86.25	-84.32	0.13
B	90.99	1.21	-52.26	-90.78	-88.08	-90.88	-12.21	-87.45	-0.63
I	90.09	-28.18	-3.80	-87.93	-60.56	-50.19	-57.19	-79.63	-0.41
E	93.23	2.00	-71.50	-93.12	-36.45	-39.19	-91.90	-90.83	-0.38





President ***Clinton*** will host a meeting at ***the White house*** ***Saturday*** with ***Israeli*** and ***Palestinian*** negotiators …