Zero-shot Sequence Labeling: Transferring Knowledge from Sentences to Tokens

NAACL 2018

Jie Zhou
2020/01/07

Task

1. Sentiment Detection

They may *have* a *SuperBowl* in Dallas,
but Dallas **ain't winning** a SuperBowl.
Not with that quarterback and owner.
@S4NYC @RasmussenPoll

Sentence Level: Negative

Token Level: None

Task

1. Sentiment Detection

They may *have* a *SuperBowl* in Dallas,
but Dallas **ain't winning** a SuperBowl.
Not with that quarterback and owner.
@S4NYC @RasmussenPoll

Sentence Level: Negative

Token Level: None

2. FCE Error Detection

3. Uncertainty Detection ("either … or …", "whether")

Motivation

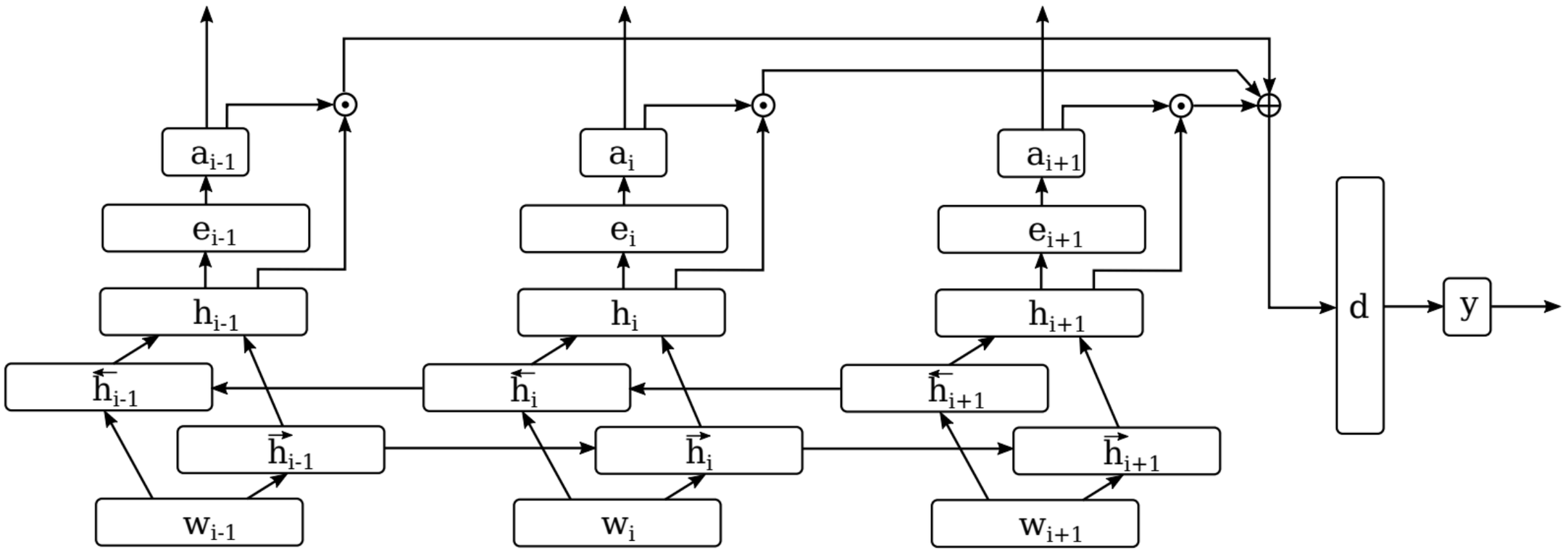
- **Visualization techniques**
  - **Attention**
  - **Gradient**

How to infer **token-level** labels for **binary** sequence tagging problems, using networks trained **only** on **sentence-level** labels?

Observation

- **Attention**-based methods **better** than **gradient**-based methods
- sometimes even **rivaling** the **supervised** oracle network

Model


a_{i-1}
a_i
a_{i+1}
e_{i-1}
e_i
e_{i+1}
h_{i-1}
h_i
h_{i+1}
$\overleftarrow{h}_{i-1}$
$\overleftarrow{h}_i$
$\overleftarrow{h}_{i+1}$
$\overrightarrow{h}_{i-1}$
$\overrightarrow{h}_i$
$\overrightarrow{h}_{i+1}$
w_{i-1}
w_i
w_{i+1}
d
y

Attention

- Word Representation

$$\overrightarrow{h_i} = LSTM(w_i, \overrightarrow{h_{i-1}})$$

$$\overleftarrow{h_i} = LSTM(w_i, \overleftarrow{h_{i+1}})$$

$$\widetilde{h_i} = [\overrightarrow{h_i}; \overleftarrow{h_i}] \qquad h_i = tanh(W_h \widetilde{h_i} + b_h)$$

Attention

- Weight

$$e_i = tanh(W_e h_i + b_e)$$

$$\widetilde{e_i} = W_{\widetilde{e}} e_i + b_{\widetilde{e}}$$

Basic

$$a_i = \frac{exp(\widetilde{e_i})}{\sum_{k=1}^{N} exp(\widetilde{e_k})}$$

Ours

Logistic

$$\widetilde{a_i} = \sigma(\widetilde{e_i}) \qquad a_i = \frac{\widetilde{a_i}}{\sum_{k=1}^{N} \widetilde{a_k}}$$

Attention

- Classification

$$c = \sum_{i=1}^{N} a_i h_i$$

$$d = tanh(W_d c + b_d)$$

$$y = \sigma(W_y d + b_y)$$

$$L_1 = \sum_j (y^{(j)} - \widetilde{y}^{(j)})^2$$

Attention

- Construct loss functions

$$L_2 = \sum_j (min_i(\widetilde{a}_i) - 0)^2$$

$$L_3 = \sum_j (max_i(\widetilde{a}_i) - \widetilde{y}^{(j)})^2$$

min_i(a_i): the min value of a_i
max_i(a_i): the max value of a_i

Alternative Methods

- Gradient

$$g_i = \frac{\partial L_1}{\partial w_i}\Big|_{(y^*,y)}$$

- Relative Frequency Baseline
- Supervised Sequence Labeling

Experiment

	CoNLL 2010					FCE				
	Sent F_1	MAP	P	R	F_1	Sent F_1	MAP	P	R	F_1
Supervised	-	96.54	78.92	79.41	79.08	-	59.13	49.15	26.96	34.76
Relative freq	-	81.78	15.94	**79.98**	26.59	-	37.75	14.37	**86.36**	24.63
LSTM-LAST-BP	84.42	77.90	7.16	66.64	12.92	85.10	46.12	**29.49**	16.07	20.80
LSTM-ATTN-BP	**84.94**	80.38	9.13	71.42	16.18	**85.14**	44.52	27.62	17.81	21.65
LSTM-ATTN-SW	**84.94**	**87.86**	**77.48**	69.54	**73.26**	**85.14**	**47.79**	28.04	29.91	**28.27**

Experiment

	SemEval Negative					SemEval Positive				
	Sent F_1	MAP	P	R	F_1	Sent F_1	MAP	P	R	F_1
Supervised	-	67.70	31.79	44.66	37.02	-	67.41	36.27	50.71	42.24
Relative freq	-	44.15	17.39	15.67	16.48	-	47.64	13.39	**54.69**	21.51
LSTM-LAST-BP	53.65	43.02	8.33	28.41	12.88	70.83	49.06	17.66	35.06	23.48
LSTM-ATTN-BP	**55.83**	50.96	11.55	**31.54**	16.90	**71.26**	53.89	23.45	34.53	27.92
LSTM-ATTN-SW	**55.83**	**54.37**	**29.41**	14.40	**19.23**	**71.26**	**56.45**	**37.19**	25.96	**30.45**

Thanks!