RICH PUNCTUATIONS PREDICTION USING LARGE-SCALE DEEP LEARNING

Xueyang Wu, Su Zhu, Yue Wu, and Kai Yu

Key Laboratory of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering
SpeechLab, Department of Computer Science and Engineering, Brain Science and Technology Research Center
Shanghai Jiao Tong University, Shanghai, China

Motivation

- **Rich punctuations paly important roles in many NLP tasks.**
- **ASR systems provide plain word streams.**
- **Small-scale models are hard to guarantee performance and generalization ability.**
- **We need data and model with larger scale to adapt to various genres of text.**

Data Comparison

- **Data used in previous research**
  - PTB, CTB
  - Manual speech transcripts

☹ small, out-of-date
☹ limited, expensive

- **Features used in previous research**
  - Syntax information
  - Prosodic cues
- **What we used**
  - 😃 abundant diversity
  - 😃 large scale
  - 😃 formal/informal

Dataset		Contents	Size (Byte)
formal corpus	training	Sina News, Wikipedia	1.4G
	in-domain	Sina News, Wikipedia	7.5M
	out-of-domain	People's Daily, Articles	12M
informal corpus	training	Weibo	1.5G
	in-domain	Weibo	4.5M
	out-of-domain	Speech Transcriptions	1.2M

Task Formulation

- **Punctuation Symbols → Labels**

Original Symbols	Unified	Punctuation	Label
, 、 ; ； —— ： : “ ” ‘ ’	,	Comma	CO
。 . …	。	Period	PE
! ！	！	Exclamation Mark	EX
? ？	？	Question Mark	QU
Space, None		None	EM

middle stop; full stop → sentence boundary denoted by SE

- **Sequence Labeling**

你知道吗？水稻、小麦、玉米，都是中国的主要农作物。

Preprocessing

你	知道	吗	水稻	小麦	玉米	都是	中国	的	主要	农作物
EM	EM	QU	CO	CO	CO	EM	EM	EM	EM	PE

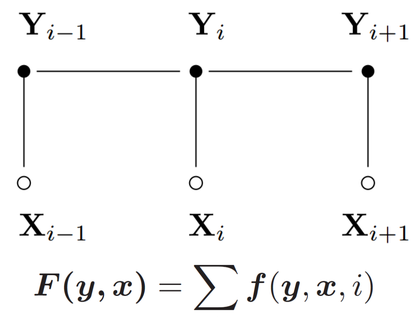
- **Evaluation Metrics**
  - **Precision** (denoted by **P**), **Recall** (denoted by **R**) and **F1-score** (the harmonic mean of precision and recall, denoted by **F**)
  - Only evaluate for **CO, PE, EX, QU,** and **SE**

Methods

- **CRF-based Model**

No.	Feature	Definition
W0	w_{i-4}	The fourth word before the current word.
...	...	...
W8	w_{i+4}	The fourth word after the current word.
P01	$p_{i-4}p_{i-3}$	The fourth and third POS tags before the current POS tags.
...	...	...
P67	$p_{i+3}p_{i+4}$	The fourth and third POS tags after the current POS tags.
P012	$p_{i-4}p_{i-3}p_{i-2}$	The fourth, third and second POS tags before the current POS tags.
...	...	...
P567	$p_{i+2}p_{i+3}p_{i+4}$	The fourth, third and second POS tags after the current POS tags.

CRF Feature-Templates



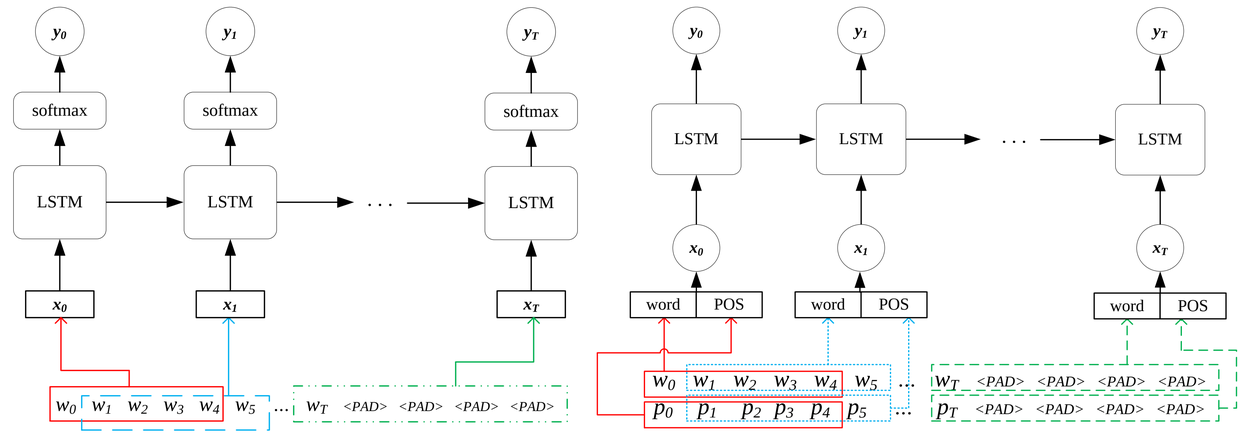
$$\boldsymbol{F}(\boldsymbol{y},\boldsymbol{x}) = \sum_{i} \boldsymbol{f}(\boldsymbol{y},\boldsymbol{x},i)$$

$$p_{\boldsymbol{\lambda}}(\boldsymbol{Y}|\boldsymbol{X}) = \frac{\exp(\boldsymbol{\lambda}\cdot\boldsymbol{F}(\boldsymbol{Y},\boldsymbol{X}))}{Z_{\boldsymbol{\lambda}}(\boldsymbol{X})}$$

$$Z_{\boldsymbol{\lambda}}(\boldsymbol{X}) = \sum_{\boldsymbol{y}} \exp(\boldsymbol{\lambda}\cdot\boldsymbol{F}(\boldsymbol{Y},\boldsymbol{X}))$$

- **Rule Settings**
  - **Conjunction**: insert **comma** before a conjunction.
  - **Parenthesis**: insert **comma** or **period** to the front and rear of parenthesis.
  - **Interrogative sentence**: insert **question mark** after the tone word if interrogative indicator is detected.
  - **Exclamatory sentence**: insert **exclamation mark** after the tone word if exclamatory indicator is detected.

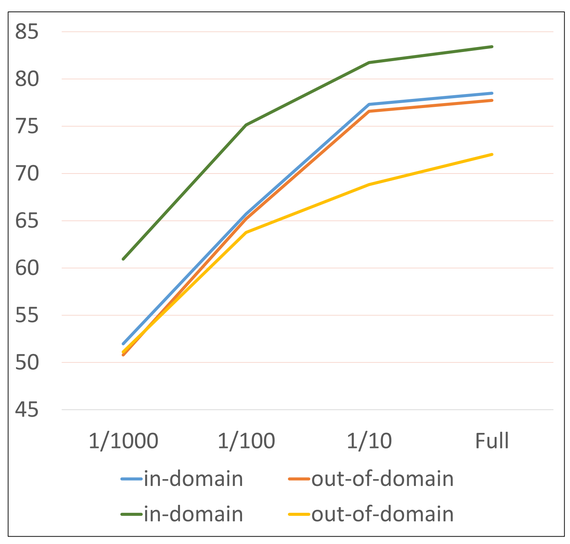
- **LSTM Model and Multiview-LSTM Model**
  - **Words** $x_t = [w_t, w_{t+1}, w_{t+2}, w_{t+3}, w_{t+4}]$
  - **POS-tag / Chunking-tag**
    $v_t = [k_t, k_{t+1}, k_{t+2}, k_{t+3}, k_{t+4}]$ $v_t = [p_t, p_{t+1}, p_{t+2}, p_{t+3}, p_{t+4}]$
  - **Multiview:** concatenate $\boldsymbol{x}_t$ and $\boldsymbol{v}_t/\boldsymbol{v}_t$



Experiments

- **CRF-based Model vs. LSTM Model (left, F1-score of each label)**
  - LSTM significantly outperform CRF-based model.
  - LSTM model knows better about *when to stop*.
  - In Chinese, the selection of punctuations relies more on long-term context. It's hard to **distinguish comma** and **period**.
- **Scale vs. Performance (right, F1-score of sentence boundary)**
  - Increasing data scale helps improves model performance
  - Low model complexity may limits the potential of large data

Training	Testing	Label	CRF-based	LSTM
formal corpus	in-domain	CO	47.43	**66.15**
		PE	55.78	**66.63**
		QU	43.03	**52.98**
		EX	**7.91**	2.79
		SE	55.99	**78.49**
	out-of-domain	CO	47.75	**66.60**
		PE	60.33	**68.05**
		QU	45.51	**52.58**
		EX	4.95	**11.41**
		SE	56.98	**77.74**
informal corpus	in-domain	CO	54.96	**69.57**
		PE	37.76	**58.78**
		QU	53.63	**57.36**
		EX	67.66	**62.97**
		SE	59.58	**83.42**
	out-of-domain	CO	30.52	**48.11**
		PE	37.07	**52.62**
		QU	49.02	**61.66**
		EX	7.09	**32.99**
		SE	41.87	**72.02**



Corpus	testing	1/1000	1/100	1/10	Full
formal corpus	in-domain	51.98	65.71	77.32	78.49
	out-of-domain	50.79	65.2	76.58	77.74
informal corpus	in-domain	60.94	75.14	81.74	83.42
	out-of-domain	51.12	63.76	68.83	72.02

- **Multiview-LSTM**
  - F1-score of SE on different training scales using **words**, **words+POS tags** and **words+chunking**.
  - When the sizes of corpora are small, extra information helps improve performance.
  - When the sizes become larger, the promotion will vanish.

Testing	Formal Corpus	Words	+POS	+chunking
in-domain	1/1000	51.98	**55.51**	54.96
	1/100	65.71	**67.25**	65.81
	1/10	**77.32**	77.06	76.70
out-of-domain	1/1000	51.50	**56.26**	56.16
	1/100	65.20	**67.25**	65.07
	1/10	76.58	**76.75**	75.87

Conclusion

- **Long-term dependency matters!**
- **The more data, the higher performance! (even though there are some noise)**
- **Syntax information can boost the model when using small corpora.**
- **Need more design of both architecture and algorithm.**

{wuxueyang, paul2204, yuewu619, kai.yu}@sjtu.edu.cn