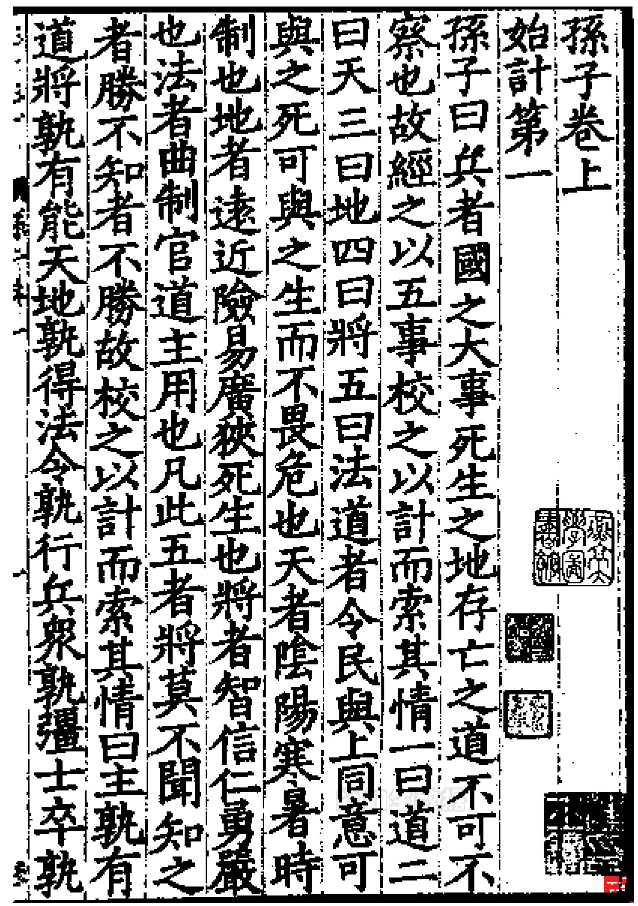
Sentence Segmentation for Ancient Chinese Texts Based on RNN

一种基于循环神经网络的古文断句方法

Boli Wang, Xiaodong Shi, Jinsong Su
Xiamen University

Introduction

- Ancient Chinese texts
  - No punctuations



孫子卷上
始計第一
孫子曰兵者國之大事死生之地存亡之道不可不
察也故經之以五事校之以計而索其情一曰道二
曰天三曰地四曰將五曰法道者令民與上同意可
與之死可與之生而不畏危也天者陰陽寒暑時
制也地者遠近險易廣狹死生也將者智信仁勇嚴
也法者曲制官道主用也凡此五者將莫不聞知之
者勝不知者不勝故校之以計而索其情曰主孰有
道將孰有能天地孰得法令孰行兵衆孰彊士卒孰

Introduction

- Ancient Chinese texts
  - No punctuations
- Ju Dou (句读)
  - Rely on language expertise
  - Time-consuming

汗 汗 吐 吐 下

而下者有醫用藥汗之下之者中間節目頗多汗藥宜早
下藥宜遲此亦大綱之論耳
張仲景曰凡發汗服三劑若汗不出者死病也
千金曰熱病脉躁盛而不得汗者此陽極之極脉也死
王三陽曰亦有黄耆建中補之而後汗出者若服三劑發
汗藥汗復不出其死必矣宜仔細候脉消息之毋妄治也
華佗曰傷寒三日四日邪在胸中者宜吐之凡吐用瓜蒂
散或温茶湯乘熱與之以指探喉中即吐也
張仲張景曰大法春宜吐中病即止不欲過也誤吐則損
人上焦元氣

金匱玉函曰虛者十補勿一瀉之虛家下之為重虛內竭
津液故令大渴求水者陽氣未竭而猶可愈惡水者陽氣
已竭則難可制
難經曰實實虛虛損不足益有餘此中工所害也
凡傷寒之病多從風寒得之凡中風與傷寒自古通謂之傷寒始表中風寒入裏
則難消矣不在證治擬欲攻之則尤當先解表乃可下之
若表已解而內不消非大滿尤生寒熱則病不除言表雖罷非大堅滿者
亦未可下之若表已解而內不消大滿大實堅滿燥屎自可除下
之若不宜不下而便攻之內虛熱入協熱遂利煩燥諸變
不可勝數輕者困篤重者必死矣

Introduction

- Ancient Chinese texts
  - No punctuations

- Ju Dou (句读)
  - Rely on language expertise
  - Time-consuming

- Digitalized ancient Chinese texts
  - Segmented: only about 4000

- Automatic sentence segmentation

Related work

- Rule-based Methods
  - RegExp (Huang, 2008)
- Statistical Methods
  - N-gram (Chen, 2007)
  - CRF (H. Zhang, 2009) (K. Zhang, 2009) (Huang, 2010)
    - Handcrafted local context features
    - State-of-the-art
- NN-based Methods
  - Sequence generation: RNNLM (Wang, 2016)
    - Underperform CRF
    - High time complexity

- **This paper**: seq-to-seq

Our approach

– Sequence labeling: $o_t^{[i]} = \mathrm{P}(y_t = i | x_1 x_2 \dots x_N)$

– Punctuations: 。？！，；：

曰：“朕此衣已三浣矣。”

曰/S 朕/B 此/M 衣/M
已/M 三/E3 浣/E2 矣/E

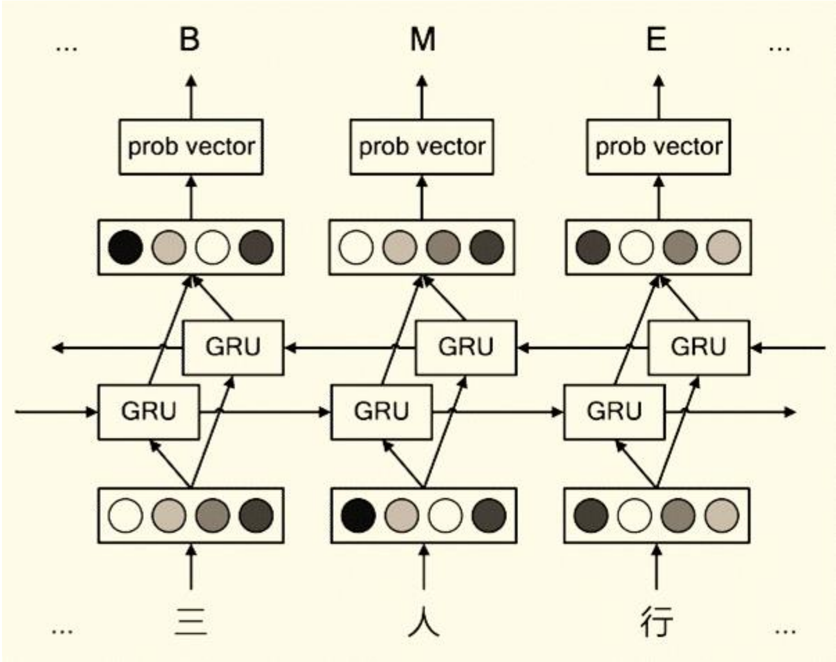
– 6-tag set (H. Zhang, 2009)

$$T = \{B, M, E3, E2, E, S\}$$

B – the first character of a clause
M – the middle character of a clause
E3 – the last 3rd character of a clause
E2 – the last 2nd character of a clause
E – the last character of a clause
S – a single character forming a clause

Our approach

- Model: RNN
  - Input layer
    Character embedding: $x_t \rightarrow e_t$
  - Hidden layer:
    Bi-directional GRU
    $\overrightarrow{h}_t = f(e_t, \overrightarrow{h}_{t-1}) \quad \overleftarrow{h}_t = f(e_t, \overleftarrow{h}_{t+1})$
  - Output layer: softmax

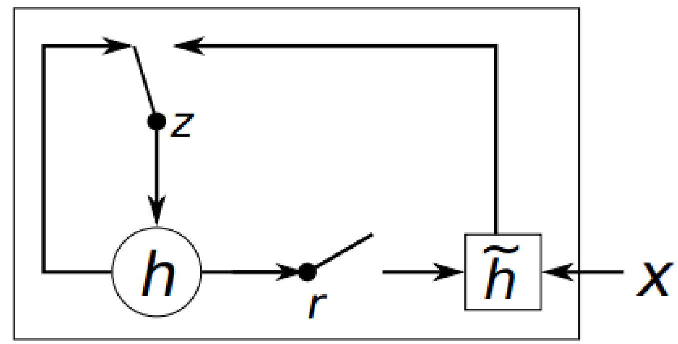


$$o_t^{[i]} = \frac{exp(W^{[i]}_{\overrightarrow{h}o} \overrightarrow{h}_t + W^{[i]}_{\overleftarrow{h}o} \overleftarrow{h}_t + b_o)}{\sum_{j \in T} exp(W^{[j]}_{\overrightarrow{h}o} \overrightarrow{h}_t + W^{[j]}_{\overleftarrow{h}o} \overleftarrow{h}_t + b_o)}$$

Our approach

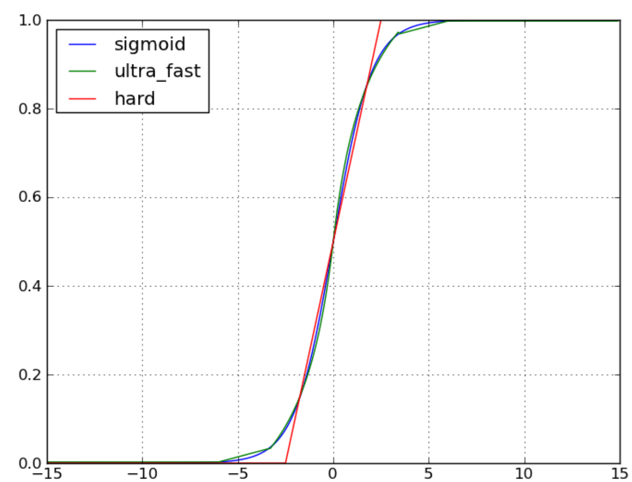
- Gated Recurrent Unit (Cho, 2014)

$$r_t = sigmoid(W_r x_t + U_r h_{t-1} + b_r),$$
$$\tilde{h}_t = tanh(W_h x_t + U_h(r_t \odot h_{t-1}) + b_h),$$
$$z_t = sigmoid(W_z x_t + U_z h_{t-1} + b_z),$$
$$h_t = z_t \odot h_{t-1} + (1 - z_t) \odot \tilde{h}_t,$$



- Our implementation
  - using **hard_sigmoid**

$$\text{hard\_sigmoid}(x) = \begin{cases} 0, & x < -2.5 \\ 0.2x + 0.5, & -2.5 \le x < 2.5 \\ 1, & x \ge 2.5 \end{cases}$$



Our approach

- Training
  - Maximum-likelihood estimation

$$Loss(\theta) = -\sum_i log(P(y_i|x_i, \theta))$$

  - Mini-batch gradient descent
    - Back-Propagation Through Time (Werbos, 1990)
    - RMSProp (Tieleman, 2012)
    - Gradient Re-normalization (Salimans, 2016)
    - Early Stopping (Yao, 2007)

- Pre-training: using a GRU-RNNLM

Our approach

- Decoding
  - Given a char sequence $x_{1:N}$
  - Goal: find $y_{1:N}$ to maximize $s(x_{1:N}, y_{1:N})$

$$s(x_{1:N}, y_{1:N}) = \sum_{i=1}^{N} (o_i^{[y_i]} + \alpha \cdot \underline{A_{y_{i-1} y_i}} + \beta \cdot \underline{C(y_{i:N})})$$

state transition probability

length penalty (# of sents)

  - Tag inference: beam search

Experiments

- Datasets

Dataset	Number of characters	Charset size	Source
Training Set	237M	23905	STCC
Development Set	0.01M	1890	HDST
Test Set 1	0.32M	6188	4HN
Test Set 2	0.36M	5755	HDST&4HN

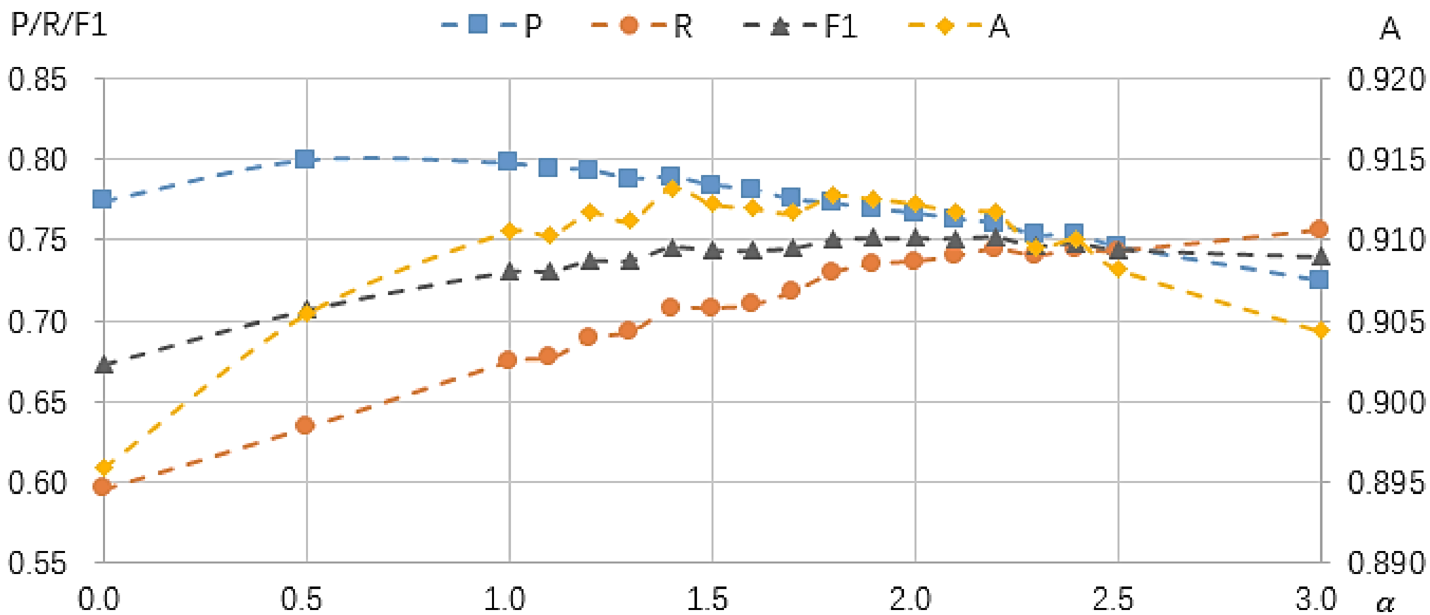
  - Superfection Traditional Chinese Corpus
    http://cloudtranslation.cc/corpus_tc.html

- Baselines
  - RNNLM: CLM1 / CLM6 (Wang, 2016)
  - CRF: reproduce (H. Zhang, 2009) with CRF++

Experiments

- Parameter Selection
  - Weight of transition score $\boldsymbol{\alpha}$

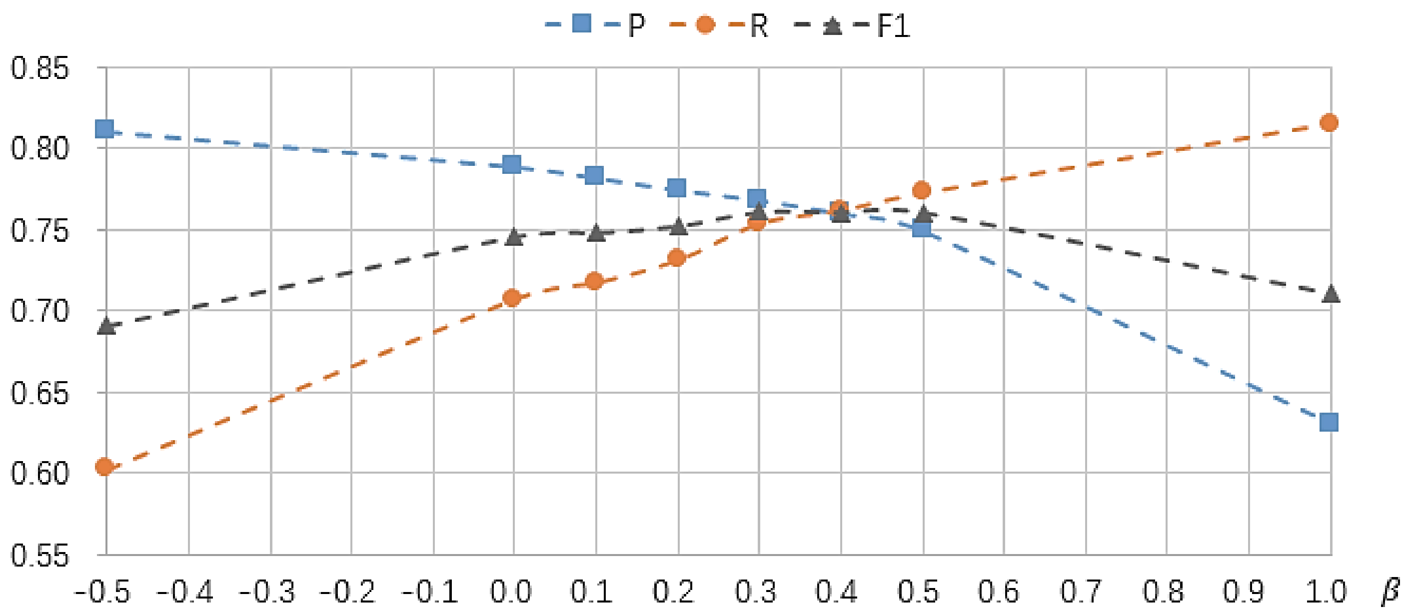
    Fix $\boldsymbol{beam = 50}$, $\boldsymbol{\beta = 0}$ => highest accuracy when $\boldsymbol{\alpha = 1.4}$



Experiments

- Parameter Selection
  - Weight of length penalty $\boldsymbol{\beta}$

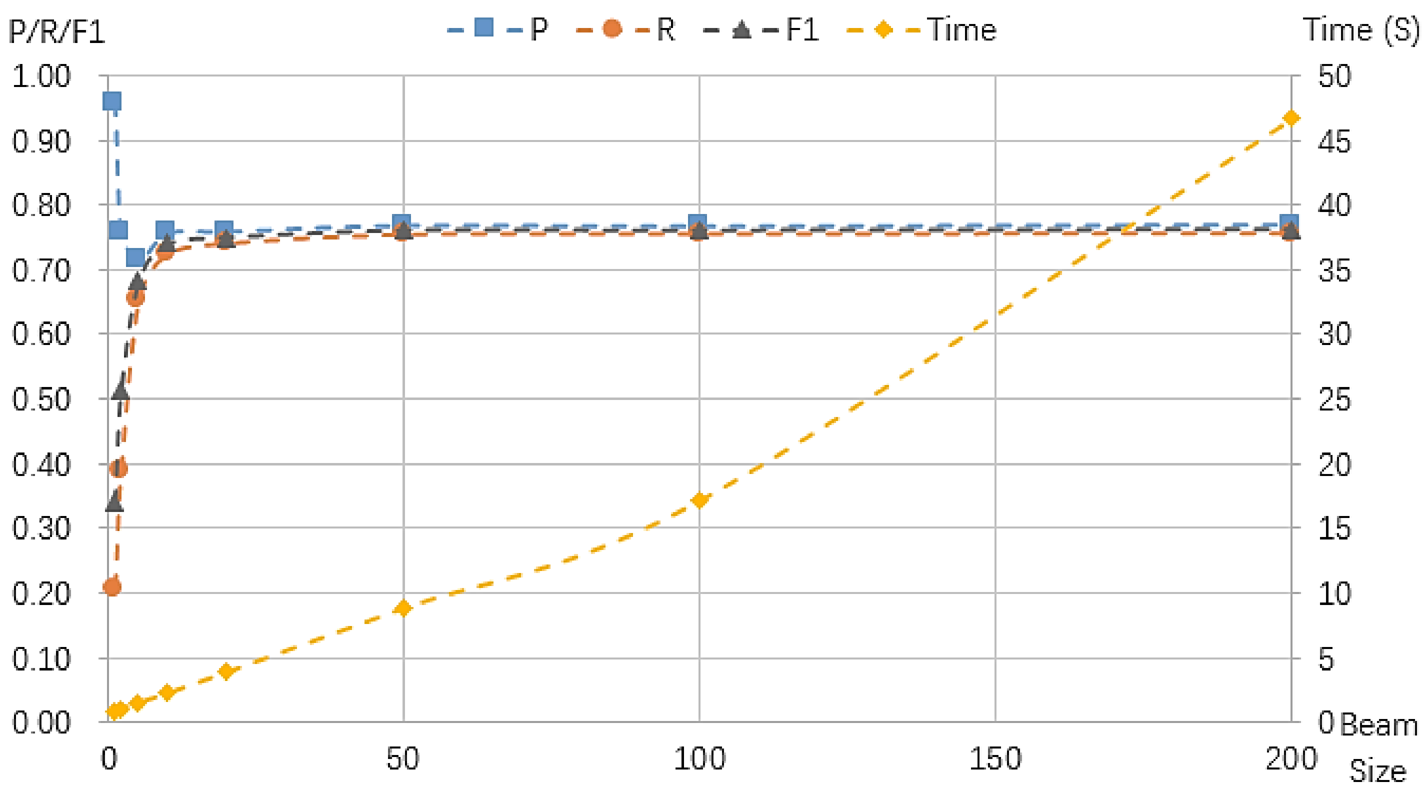
    Fix $\boldsymbol{beam = 50}$, $\boldsymbol{\alpha = 1.4}$ => highest F1 when $\boldsymbol{\beta = 0.3}$



Experiments

- Parameter Selection
  - Size of beam

    Fix $\alpha = 1.4$, $\beta = 0.3$



Experiments

- Results

Models	Test Set 1			Test Set 2		
	P	**R**	**F1**	**P**	**R**	**F1**
CLM6	0.782	0.577	0.664	0.728	0.636	0.679
CLM6+0.65LP	0.749	0.673	0.709	0.686	0.721	0.703
CLM1+0.65LP	0.721	0.733	0.727	0.639	**0.769**	0.698
CRF	**0.816**	0.662	0.731	**0.786**	0.710	0.746
GRU-RNN	0.755	0.677	0.714	0.734	0.633	0.680
GRU-RNN+1.4ST[50]	0.785	0.720	0.751	0.760	0.719	0.739
GRU-RNN+1.4ST+0.3LP[50]	0.761	0.746	0.753	0.735	0.753	0.744
GRU-RNN+1.4ST+0.3LP[200]	0.763	**0.748**	**0.755**	0.738	0.756	**0.747**

Conclusion

- Bi-directional GRU-RNN outperform CRF and RNNLM methods.
- Integrations of transition score and length penalty are effective.

- Future works
  - Classify texts by ages and train sub-models
  - Segmentation → punctuation

Thank you!

by Boli Wang
me@bo-li.wang