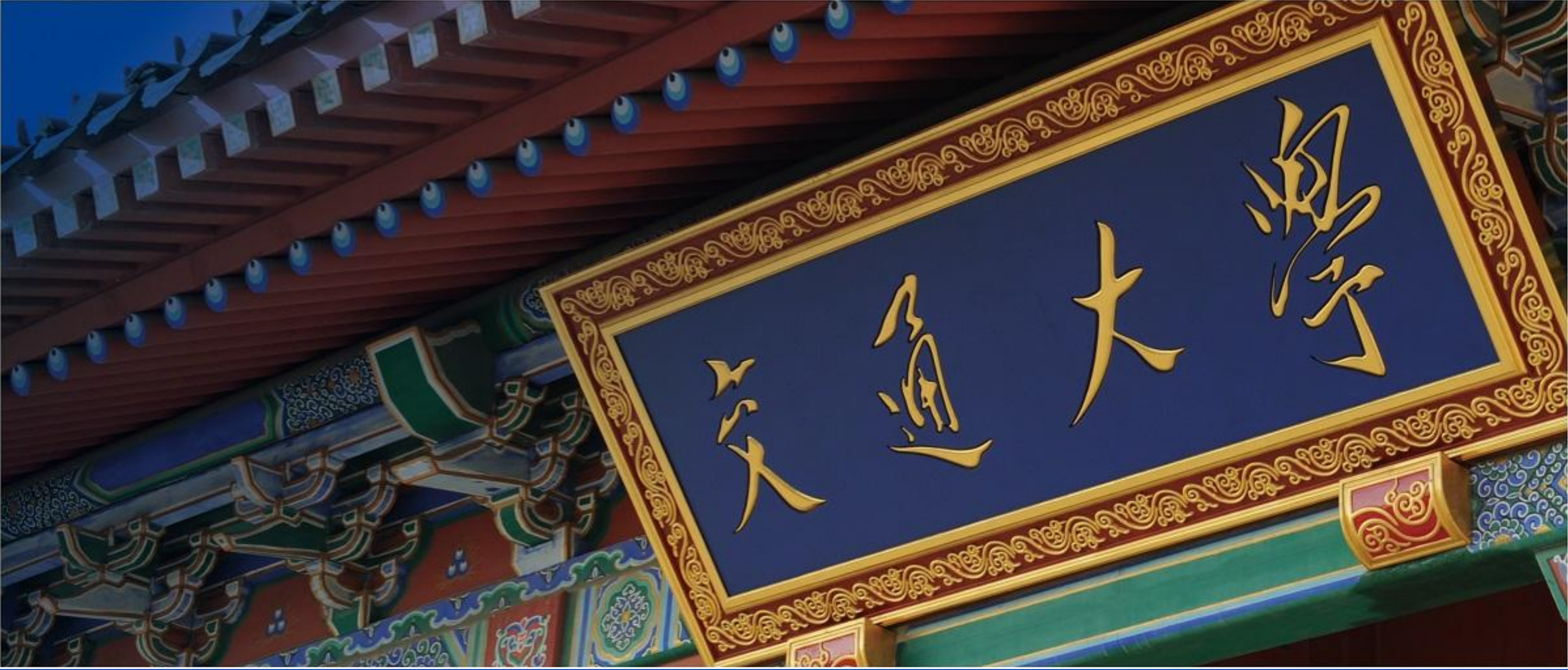

Rumor Detecting with RNN

房宇辰

2018年6月

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

目录 Contents

目录 Contents

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

Introduction

- Nowadays social platforms are becoming perfect places for rumors to spread.
- People post cues about one event over time, generating a long-distance dependencies of evidence.
- We want to find an approach to modeling the sequences of posts and judging whether an event is rumor.

目录 Contents

上海交通大學
Shanghai Jiao Tong University

Data Process
Dataset

- The experiment is based on a crawled Weibo dataset, which contains 3,805,656 posts and 4,664 events.

- There are 2,313 rumors and 2,351 non-rumors.

Data Process
Paragraphing

- The number of posts about an event can be huge and varied.
- An event may only be discussed heatedly for a short range of time.
- We need to find the time ranges where these events are heatedly discussed and paragraph them into sequences with similar length.

Data Process
Paragraphing

Input : Relevant posts of $E_i = \{(m_{i,j}, t_{i,j})\}_{j=1}^{n_i}$,
Reference length of RNN N
Output: Time intervals $I = \{I_1, I_2, \ldots\}$

/* Initialization */
$L(i) = t_{i,n_i} - t_{i,1}$; $\ell = \frac{L(i)}{N}$; $k = 0$;
while *true* **do**
  k ++;
  $U_k \leftarrow Equipartition(L(i), \ell)$;
  $U_0 \leftarrow$ {empty intervals} $\subseteq U_k$;
  $U'_k \leftarrow U_k - U_0$;
  Find $\bar{U}_k \subseteq U'_k$ such that $\bar{U}_k$ contains continuous intervals that cover the longest time span;
  if $|\bar{U}_k| < N$ && $|\bar{U}_k| > |\bar{U}_{k-1}|$ **then**
    /* Shorten the intervals */
    $\ell = 0.5 \cdot \ell$;
  else
    /* Generate output */
    $I = \{I_o \in \bar{U}_k | I_1, \ldots, I_{|\bar{U}_k|}\}$;
    return I;
  end
end
return I;

Data Process
Vectorization

- Use tf-idf value to vectorize the paragraphs

$$\mathrm{tf_{i,j}} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

$$\mathrm{idf_{i}} = \log \frac{|D|}{|\{j : t_i \in d_j\}|}$$

$$\mathrm{tfidf_{i,j}} = \mathrm{tf_{i,j}} \times \mathrm{idf_{i}}$$

- We take top-6000 words according to their tf-idf value as the dictionary.

目录 Contents

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

Modeling

- Simple LSTM
- Multi-layer RNN
- RNN with attention mechanism

Modeling

- We found that when the length of RNN(N) is less than 50, the performance is better when N is larger. And the performance began to drop when N is larger than 50.
- Use a simple attention to make the model able to deal with longer sequences.
- $a_i = f(H_i, x_{last})$

目录 Contents

Experiment

Model	Accuracy	F1 Score	AUC
Simple LSTM	0.9048	0.9024	0.9640
Multi-layer RNN	0.9119	0.9084	0.9668
RNN with Attention	0.9238	0.9223	0.9728

目录 Contents

Reference

[1] Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural Machine Translation by Jointly Learning to Align and Translate, 1–15. https://doi.org/10.1146/annurev.neuro.26.041002.131047

[2] Ma, J., Gao, W., Mitra, P., Kwon, S., & Jansen, B. J. (2016). Detecting Rumors from Microblogs with Recurrent Neural Networks Detecting Rumors from Microblogs with Recurrent Neural Networks, (July), 3818–3824.

[3] Kim, J., Tabibian, B., Oh, A., Schoelkopf, B., & Gomez-Rodriguez, M. (2017). Leveraging the Crowd to Detect and Reduce the Spread of Fake News and Misinformation. https://doi.org/10.1145/3159652.3159734

[4] Fan Yang, Yang Liu, Xiaohui Yu, and Min Yang. Automatic detection of rumor on sina weibo. In Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics, 2012.

Thanks!