

# Shaobo Liu

Institute of Computing Technology, Chinese Academy of Sciences, No.6 Kexueyuan South Road, Haidian, Beijing, China

☎ (+86) 15652332566 | ✉ averyliu2013@gmail.com | 🏠 liushaobo.me

## Education

---

### Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China

Sep. 2016 - PRESENT

- M.S.E. in Computer Science, at CAS Key Laboratory of Network Data Science and Technology  
Major GPA: 3.64/4.0, Overall GPA: 3.66/4.0

### Harbin Institute of Technology, Harbin, China

Aug. 2012 - Jul. 2016

- B.E. in Software Engineering, at School of Software Engineering  
Major GPA: 3.89/4.0, Overall GPA: 3.85/4.0 (Ranking: 6/116, **Top 6%**)

### Nanyang Technological University, Singapore

Aug. 2015 - Dec. 2015

- Exchange Student at School of Computer Science and Engineering

## Publication

---

- EMNLP 18** *Exploiting Contextual Information via Dynamic Memory Network for Event Detection.* Shaobo Liu, Rui Cheng, Xiaoming Yu, Xueqi Cheng. 2018 Conference on Empirical Methods in Natural Language Processing (Short).

## Internship

---

**NLP Research Intern** Natural Language Processing Group, Alibaba DAMO Academy

Dec. 2018 - PRESENT

## Projects

---

**Research Interest:** Information Extraction, Natural Language Processing, Data Mining, Machine Learning

### INFORMATION EXTRACTION

#### Heterogeneously Supervised Event Detection (In Preparation)

Nov. 2018 - PRESENT

- The scarcity of human-labeled data is a commonly shared problem among natural language processing tasks. The expensive labeling cost results in size-restricted human annotated datasets, thus impeding the further advancement of the field.
- To address this issue, **distant supervision** was proposed using knowledge bases as sources of weak supervision. However, knowledge bases are often incomplete and are inapplicable in domain-specific tasks.
- **Heterogeneous supervision** was proposed to complement the supervision from knowledge bases with human written heuristic labeling rules. However, heuristic labeling rules, unlike humans who make random errors, tend to make patterned errors. The assumption can be made that each labeling rule has its proficient subset within which it can confidently label samples. We intend to utilize reinforcement learning to determine the proficient subset for each labeling rule. Therefore, we can improve overall labeling accuracy by distributing a new, unlabelled sample to the most promising labeling rule.

#### Trigger Detection Dynamic Memory Network

May. 2018 - Aug. 2018

- Proposed the TD-DMN model for the event trigger detection task. The task is cast as a question answering problem, enabling the use of **dynamic memory network** (DMN). Furthermore, each sentence in the document to be classified is deemed as an implicit question by the question module of the TD-DMN, compared to the use of empty questions in the original DMN. For any given sentence of interest, the TD-DMN model exploits its across sentences and within document context through the multi-hop mechanism which iteratively processes the context. Experimental results show that the use of implicit questions and multi-hop mechanism improves the model performance.
- Performed 5-fold cross-validation on the ACE 2005 dataset. The previous works on the ACE-2005 dataset usually used a train, validation and test segmentation which consists of 529, 30, 40 documents respectively. However, the relatively small number of test documents may result in a biased evaluation, and a more reasonable way is to evaluate using 5-fold cross validation by randomly separating documents into five parts.
- Implemented the DEEB-RNN event detection model for comparison on the new data split.
- Wrote a short paper which was accepted by **EMNLP 2018 (1st author)**.

### SOCIAL MEDIA AND NEWS CRAWLER

#### News Crawler for Obtaining Unlabelled News Text

Jul. 2018 - Jul. 2018

- Wrote a news crawler to obtain news text. The crawler first obtains real-time news seeds through RSS subscription, and then it queries the seeds in Bing news search engine to obtain related news.
- Crawled 2.55 million news articles in total from July to October, approximately 30 thousand news a day using a single machine.
- The stored unlabelled news text can be exploited using semi-supervised or weakly-supervised techniques.

## Distributed Social Media Crawling Platform

Dec. 2017 - Apr. 2018

- Wrote a distributed social media crawling platform.
- Decoupled different components using redis database as the message broker.
- Created docker image and identified CAPTCHA using convolution neural networks.

## DATA COMPETITION

### Indoor Positioning by Mining Wifi Signals from Mobile Devices - hosted by Alibaba

Oct. 2017 - Dec. 2017

- The task is to predict which store the user is currently at in a shopping mall using the Wifi names and their signal strength detected by the mobile device.
- The desensitized data were given. The task can be viewed as a binary or a multiple classification problem. One challenge for binary classification is that only positive labels were given. Therefore the negative labels should be carefully designed to achieve greater precision. The multiple classification model was applied to select highly similar but incorrect samples as candidate negative samples using the output prediction probability.
- Used XGBoost, LightGBM and their ensembles to model the task as a multi-classification problem in the preliminary round.
- Manipulated data tables and did feature extraction using SQL in the final round on the cloud computing platform provided.
- Placed the **7th place (top 0.3%)** out of 2845 teams in the final.

### Offline Store Coupon Usage Prediction - hosted by Alibaba

Oct. 2016 - Dec. 2016

- The task is to predict whether a user would use a coupon from a specific merchandise given the information of the coupon, the user and the merchandise.
- Placed the **62th place (top 5%)** out of 1501 teams in the final.

## APPLICATION DEVELOPMENT

### Final Year Project

Aug. 2015 - May. 2016

- Designed a social application called *Kamplus*. Kamplus is a Facebook-like application customized for campus students. Other than messaging, users of Kamplus can post tasks asking for help as well as view tasks posted by others.
- Developed the Kamplus iOS application front end.
- Developed the application server end using Django framework.

### WI Input Method (on Chinese App Store)

May. 2014 - Jun. 2015

- iOS development team leader and core developer.
- Led two group members and developed the Chinese input method application.
- Submitted the application to the App Store and maintained it till June 2015.

## Honors & Awards

---

2018	<b>National Scholarship</b> , dollar amount: \$3000	Beijing, China
2017	<b>7th place</b> , Data competition: Indoor positioning by mining wifi signals from mobile devices	Beijing, China
2016	<b>62th place</b> , Data competition: Offline store coupon usage prediction	Beijing, China
2015	<b>Scholarship</b> , Offered by China Scholarship Council, dollar amount: \$5000	Harbin, China
2015	<b>Finalist</b> , Intel invent 50 competition	Singapore
2014	<b>First class scholarship</b> , Offered by School of Software Engineering	Harbin, China
2013	<b>Second class scholarship</b> , Offered by School of Software Engineering	Harbin, China

## Presentation

---

### 2018 Conference on Empirical Methods in Natural Language Processing

Nov. 2018

- Presented *Exploiting Contextual Information via Dynamic Memory Network for Event Detection* (3E: Short Posters II).

## Skills

---

**Languages** Python, C++, C, SQL, JAVA, Objective-C, Shell

**Courses** Natural Language Processing, Web Data Mining, Modern Information Retrieval, Advanced Artificial Intelligence, etc.

**English** GRE: 324 (V:158 Q:166) + 3.5, TOFEL: 106 (R:29 L:28 S:23 W:26)

## Extracurricular Activity

---

### Volunteered in Social Media Processing Conference

Sep. 2017 - Sep. 2017

- Volunteered to help conference registration, venue arrangement etc.

### Volunteered in "One Family" Korean volunteer club

Aug. 2013 - Aug. 2014

- Participated in environmental protection activities.