Leilan Zhang

"The only thing we have to fear is fear itself."

github.com/zll17 · zhangleilan@gmail.com · (+86) 188-1139-3706

EDUCATION

Tsinghua University *Beijing, China*

M.S. in Computer Science *2017 – 2020*

- GPA: 3.41 / 4.0
- Research Instituion: Center of Speech and Language Technologies (CSLT)
- Research Area: Natural Language Processing, Topic Modeling, Machine Learning
- Main Courses: Machine Learning, Convex Optimization, Computational Linguistics, Functional Analysis, Algorithm Analysis

Sichuan University *Chengdu, China*

B.S. in Mathematics *2012 – 2016*

- GPA: 3.6 / 4.0
- Single First class Scholarship (2013 - 2014)
- Provincial Second Prize in the Sixth National College Students Mathematics Competition (2014)
- Relevant Courses: Mathematical Analysis, Complex Analysis, Probability Theory, Statistics, Differential Equation, Differential Manifold, Advanced Programming, C Programming, Data structure and Algorithm

PUBLICATIONS

- **Leilan Zhang**, Qiang Zhou. *Automatically Annotate TV Series Subtitles for Dialogue Corpus Construction*. 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Lanzhou, China, 2019, pp. 1029-1035.
- **Leilan Zhang**, Qiang Zhou. *Topic Segmentation for Dialogue Stream*. 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Lanzhou, China, 2019, pp. 1036-1043.
- **Master Thesis**: Research on Clustering Algorithm for Subtitle Dialogue Text Based on Neural Topic Model.
- Qiang Zhou, **Leilan Zhang**. "A script-based subtitle scene and speaker information automatic labeling method and system". CN Patent Application 201811633842.5, filed July 2019. Patent pending.
- Qiang Zhou, **Leilan Zhang**. "Subtitle dialogue stream theme segmentation method and device". CN Patent Application 201910906359.8, filed Feb 2020. Patent pending.

PROJECTS

- **Neural Topic Models Open-Source Library** [*Github*] [*Doc*] *Mar 2020*
  - Provided an out-of-box library with simple and unified interfaces for the above NTMs, and the evaluation on short Chinese text datasets of the implementation shows superior results to LDA.
  - Implemented a series of neural topic models based on VAE and WAE (including NVDM-GSM, W-LDA, ETM, BATM).
  - Proposed and implemented two improved topic models (GMNTM and WTM-GMM), which adopt Gaussian mixture prior distribution to accommodate short textual corpus. The repository on Github has received 71 stars and 19 forks so far.

- **Dialogue System for Music Recommendation** [*Doc*] [*Demo*] *May 2019*
  - Crawled 7000+ song lyrics from the Internet, adopted LDA to cluster the data, then built a user intent recognition model based on TextCNN (acc:97%).

- Implemented a chatting module and a music recommendation module. The recommendation module supports two modes: (1) multi-turn songs query (2) songs recommendation according to user preferences. The recommendation module is based on node2vec.

- **Topic Segmentation System for Text Stream** [*Code and Data*] *Apr 2019*
  - For purpose of cutting a text stream into small segments according to the topic variation, proposed a model combined with BERT and Temporal Convolutional Network (TCN), in which BERT is used to embed sentences to obtain their semantic representation and TCN is adopted to detect the topic conversion boundary.

- **Automatic Speaker Tag Annotation System for Subtitles** [*Data*] *Nov 2018*
  - The goal is to annotate speaker tags to bilingual TV subtitles semi-automatically.
  - Collected raw scripts and subtitles of 4 TV and adopted **ElasticSearch** to align utterances between scripts and subtitles roughly.
  - Designed models to detect and correct alignment errors for post-processing based on TCN, and finally constructed a multi-turn Chinese dialogue corpus containing nearly 260,000 utterances with annotation accuracy of 94%.

Professional Activities

- **Volunteering work**

  Asia-Pacific Signal and Information Processing Associa Conference *Lanzhou, Nov 2019*
  China National Conference on Computational Linguistics *Changsha, Oct 2018*

- **Talks**

  Doctoral and Master Students Forum in Computer Science, Tsinghua Univ. *Beijing, May 2019*

- **Internship**

  Research intern, Chatbot group, Duoyi AI Research Institute *Guangzhou, May 2016*

Skills

- Proficient in Math (achieved 147/150 in National Graduate Entrance Examination of China)
- Programming Languages: Python, C++, Java, Mathematica, SQL
- Machine Learning Framework: PyTorch, Sklearn, Pandas, Keras, Tensorflow, Gensim
- General Tools: Linux, Git, LaTeX, SVN
- Sports: good at cycling (spent 20 days riding a bicycle from Chengdu to Tibet, 2100 km in total)