SHIJIE YAO

◇ +86 131 2086 8512 ◇ yaoshijie628@gmail.com ◇ shijieyao.github.io ◇ www.linkedin.com/in/shijieyao628

EDUCATION

Master of Science, Speech and Language Processing Sept 2017 - Nov 2018
The University of Edinburgh (UoE), UK

- Degree: First-class Honours (expected) Average score (Level): 75.75 (A3)
- Main courses (Score/Level):

<Natural Language Processing related>			
Accelerated Natural Language Processing	76/A3	Topics in Natural Language Processing	80/A2
Machine Translation	81/A2	Natural Language Understanding	59/C
<Speech related>			
Speech Processing	83/A2	Speech Synthesis	72/A3
Automatic Speech Recognition	70/A3	Introduction to Phonology and Phonetics	75/A3
<Miscellaneous>			
The Human Factor: Working with Users	75/A3	Univariate Statistics and Methodology using R	80/A2
Computer Programming (Python) for Speech and Language Processing			82/A2

- Dissertation: Sequence to Sequence Japanese Neural Lemmatisation: *doweneedsegmentation*?
  (Supervisor: Prof. Sharon Goldwater)
  We applied sequence to sequence models to Japanese morphological analysis of segmentation and lemmatisation by using the framework Nematus. By building models for segmentation, lemmatisation and both, and comparing pipeline models (which first segment text then lemmatise it) with joint end-to-end models (which carry out segmentation and lemmatisation simultaneously), our results and conclusion are: (i) the attempt of applying *seq2seq* models to Japanese segmentation and lemmatisation is a success, as the models all worked reasonably well compared with the baselines (existing Japanese morphological analysers: MeCab, JUMAN and JUMAN++) (ii) regarding performance scores, the pipeline model (F1: 0.9246) outperformed the joint model (F1: 0.9218) by a little bit (iii) the separate optimisation of the pipeline model was time-consuming, but if enough data for training either task in the pipeline could be prepared, the overall performance should be further boosted (iv) the training of the joint model was more efficient especially when the data for training separate tasks were shared.

Bachelor of Arts, Japanese Language and Literature Sept 2013 - Jun 2017
Shanghai Jiao Tong University (SJTU), China

- Major GPA: 3.99/4.0 (92/100) Overall GPA: 3.92/4.0 (91.5/100) Ranking: 1/18
- Dissertation: A Study about *Ateji* in Japanese Popular Song Lyrics: From a Computational Linguistics Perspective
  (Supervisor: Dr. Miyuki Kawasaki)
  We crawled all *ateji* pairs in the lyrics of Japanese popular songs from Utaten, a website of Japanese lyrics collection. We analysed the traits and the tendency in using *ateji*, regarding the rhetorical meanings of the *ateji* word pairs, the genders of the lyricists, the year of issue of the songs, etc.

Exchange study in Faculty of Inter-cultural Studies, Linguistics Apr 2015 - Feb 2016
Kobe University, Japan
GPA: 4.13/4.3

SKILLS AND INTERESTS

Natural Languages	Mandarin & Shanghai-based Wu dialect (native), English & Japanese (proficient)
Artificial Languages	Python, Bash Shell, HTML, CSS, LaTeX
Frameworks/Toolkits	Keras, Nematus, Theano, TensorFlow, Chainer, Kaldi, HTK
Hobbies	Photography, Programming, Piano, Dubbing, Singing

PROFESSIONAL EXPERIENCES

Data Intern in Algorithm Team @ Liulishuo, Shanghai Jun 2017 - Sept 2017

- Wrote Python scripts to pre-process speech and language data for annotation
- Annotated the speech and text data produced by native Chinese learners of English
- Highly praised by colleagues due to efficiency and expertise

PRACTICAL PROJECTS

Natural Language Processing related

- Character-level N-gram Language Modelling (coursework of Accelerated Natural Language Processing): constructed char-level n-gram language models from scratch and computed perplexity for text.
- CKY-algorithm for Grammars and Parsing (ditto): modified CKY-algorithm for constructing parsing trees.
- Distributional Similarity (ditto): implemented various ways of computing word similarities, including cosine similarity, Positive Pointwise Mutual Information and Jensen-Shannon Divergence, on twitter data.
- English-Japanese *seq2seq* Neural Machine Translation (coursework of Machine Translation): modified an MT implementation in Chainer to improve English-Japanese translation performance, by employing dropout and attention.
- Recurrent Neural Network Language Modelling (coursework of Natural Language Understanding): implemented some basic functions of RNN for language modelling in NumPy, to predict subject-predicate agreement in English.

Speech related

- GMM-DNN Hybrid ASR systems (coursework of Automatic Speech Recognition): improved a GMM-DNN hybrid ASR system (based on Kaldi) recognising continuous natural speech by tuning parameters including the number of Gaussian mixture components per state, acoustic features, dynamic features, feature normalisation, gender adaptation, etc.
- Digit Recogniser (coursework of Speech Processing): implemented an ASR system for recognising isolated digits using HTK toolkits.
- Unit Selection Voice Building (coursework of Speech Synthesis): built a voice from self-recorded speech using Festival as the front-end and unit selection methodologies.

EXTRA-CURRICULAR ACTIVITIES

Participant in Hackx SJTU Hackathon, SJTU May 2017

- Carried out a chatbot which helps filtering users' negative wording by applying the tone analyser API provided by IBM Watson

Participant in Rhodes x SJTU Youth Forum, SJTU Mar 2017

- Concluded the group discussion as the representative on the topic of "Be a good teller of China's stories (President Jinping Xi's words) - Promote the international understanding and identification of Chinese culture and voices"

Participant in "Meeting the Prime Minister", SJTU Nov 2013

- Discussed with Mr. David Cameron (the former Prime Minister of the UK) on Sino-UK issues

AWARDS

Outstanding Graduate (Top 1%), SJTU	Jun 2017
Merit Student (Top 3%), SJTU	Nov 2014
Scholarship for Outstanding Students (Top 5%, twice), SJTU	Oct 2014, Oct 2016
Japan "JASSO" Scholarship for International Students	Apr 2015 - Feb 2016