Chang Shu

cs2175@cam.ac.uk +44-7422-419-024

EDUCATIONS

University of Cambridge **Cambridge, United Kingdom**
Ph.D, Computation, Cognition and Language *2022 - 2026*
Research Topic: Human-centred Generative AIs
Supervisor: Prof. Nigel Collier and Dr. Ehsan Shareghi

University of Edinburgh **Edinburgh, United Kingdom**
Master by Research, Linguistics (Distinction) *2022*
Thesis: Pretrained Language Models as the Automatic Re-Annotator
Supervisor: Prof. Bonnie Webber and Dr. Beatrice Alex

University of Edinburgh **Edinburgh, United Kingdom**
Bachelor of Science, Computer Science (First class with honours) *2019*
Thesis: An Investigation on Adaptation for Speech Recognition on Edinburgh Recording Data
Supervisor: Prof. Steve Renals

EXPERIENCES

ILCC, University of Edinburgh **Edinburgh, United Kingdom**
M.Res Degree Project , supervised by Prof. Bonnie Webber and Dr. Beatrice Alex. *Feb. 2021 - Present*

Annotation inconsistency detection and correction based on pre-trained models:

- Explored approaches to automatically detecting annotation inconsistencies and errors and suggesting alternative labels based on the intrinsic knowledge of PLMs and the overall concordance of the corpus.
- Composed three new sentence-level and document-level datasets for evaluating the performance of the proposed re-annotator in the relation extraction task by manually revising DocRED and TACRED.
- Based on the intuition that the examples sharing the same label should obtain similar contextual embeddings from PLMs to be mutually adjacent, we investigated K-Nearest Neighbours (KNN), Kernel Density Estimation (KDE), Transformer and Contrastive Learning to utilize this neighborhood agreement fully.
- The framework achieves competitive results with the macro F1 up to 66.2% and 61.2% while regarding human revisions on TACRED and DocRED dataset as ground truth.

Penn State University and Yale University **New Haven, United States**
Research Intern in NLP, supervised by Dr. Rui Zhang and Prof. Dragomir Radev. *Jul. 2020 - Jan. 2021*

Adversarial augmentation and evaluation for logic-preserved text generation:

- Proposed a cyclic augmentation training framework that strengthens the logic faithfulness of generated text by covering diverse unseen logic variations and a new slot-filling logic evaluation metric that accurately measures the logical consistency with a refined keyword matching mechanism.
- Experiments demonstrate that the framework at most increases the logic fidelity by 10.1% on SQL-to-Text and 1.2% on Logic-to-Text tasks compared to the baseline, and the proposed metrics achieves a +0.66 Pearson correlation coefficient compared with human labels on judging logic consistency, which is better than traditional BLEU and ROUGE metrics.

ET Industrial Brain Group, Alibaba Cloud **Beijing, China**
NLP Engineer Intern *Sep. 2020 - Dec. 2020*

- Involved in building the ET industrial dialogue system for State Grid, a natural language interface for factory staff to access the expediently existing professional knowledge, including repair guidelines, sensor data, and industrial-strength database and knowledge graphs.
- Improved the document retrieve module by developing a dynamically weighting model for the search terms based on the nested NER system with an active learning mechanism.

THUNLP, Tsinghua University **Beijing, China**
Research Assistant in NLP, supervised by Prof. Zhiyuan Liu *Sep. 2019-Jul. 2020*

Knowledge graph embedding with contrastive learning:

- Replaced the random embedder of the entity and relation of ConvE and TuckER with the CompGCN trained in the unsupervised contrastive learning manner. It significantly shortens the converging time of the knowledge

graph embedding models.

AST grounded semantic parser for text-to-SQL:

- Built a graph-based semantic parser for text-to-SQL in context on the SParC dataset, which can explicitly reason on conversation history based on entity cor-reference, and decode the SQL query in each turn by first encoding the historical data Abstract Syntax Tree(AST) predictions with Graph Attention Networks (GATs).

ICSA, University of Edinburgh **Edinburgh, United Kingdom**
Summer Research Engineer in ML platform, supervised by Dr. Boris Grot *Jun. 2019-Aug. 2019*

- Developed a scalable HPC environment on Google Cloud Platform(GCP) with Slurm job scheduler and Lustre file system configured, that has the ability of automatically adjusting the size and configuration of Slurm cluster on demand.
- Configured the experiment environment and docker image for ML performance benchmark MLPerf on aforementioned cloud-based ML cluster, including setting up NCCL, openMPI, and horovod across docker containers hosted on multi-nodes with multiple GPU attached.

CSTR, University of Edinburgh **Edinburgh, United Kingdom**
B. Sc Degree Project, supervised by Prof. Steve Renals *May. 2018-Apr. 2019*

- Collected and built the Edinburgh recording corpus starting from fetching the raw audio data from YouTube. It involves pre-processing the audio data and hand-written transcript, preparing the Kaldi data directory, segmenting the long utterance, and data-splitting.
- Empirically compared the TDNN and LSTM baseline acoustic models and studied the influence of using different training data and language models.
- Investigated the speaker and domain adaptation for acoustic models including LHUC, fine-tuning, i-vector, and fMLLR.
- Researched the acoustic-data driven adaptation for lexicon and evaluated the results with the Edinburgh recording corpus.
- Implemented and analyzed a hybrid ASR system that used both the LHUC based domain adapted acoustic model and acoustic-data driven extended lexicon, which reduced the WER by 9.34% compared to the baseline.

University of Edinburgh **Edinburgh, United Kingdom**
Machine Learning Course Project, advised by Dr. Ivan Titov *Dec. 2018-Apr. 2019*

- Developed a syntax-augmented semantic parser for text-to-SQL with Spider dataset. It firstly converted natural language questions to dependency tree by off-shelf syntax parser and then utilized GCN to communicate structural information between word embeddings on the acquired dependency tree.
- Replaced the original word embedding layer with BERT, and finally acquired 6.7% improvement with the aforementioned two modifications in prediction accuracy over the baseline SyntaxSQL.

PUBLICATIONS

- **Chang Shu**, Yusen Zhang, Xiangyu Dong, Peng Shi, Tao Yu, and Rui Zhang. Logic-consistency text generation from semantic parses. In Findings of the Association for Computational Linguistics: ACL/IJCNLP2021, pages 4414–4426, 2021.
- Zhuohui Wei, **Chang Shu**, Changsheng Zhang, Jingying Huang, and Hongmin Cai. A short review of variants calling for single-cell-sequencing data with applications.The International Journal of Biochemistry CellBiology, 92:218–226, 2017.

AWARDS & ACHIEVEMENTS

- CRoSS Incubator Bursary Award of £600 from Cambridge Enterpreise *2023*
- UKRI EPSRC DTP-Toshiba Studentship of £ 50,000/ year *2022-2026*
- "2+2 double degree programme" scholarship of £ 5,000/ year *2017-2019*

TEACHING

- Tutor and Lab Demonstrator for LI18: Computational Linguistics, University of Cambridge *2022-2023*

SKILLS & LANGUAGES

- **Programming**: Skilled in Python, Shell, C++, Java; Experienced with Pytorch, and related NLP toolkits, AllenNLP, HuggingFace, DGL, PyG; Experienced with large-scale pre-trained language models.
- **Skills**: Knowledgeable in Machine Learning and Pattern Recognition; Expertise in Natural Language Processing (NLG, KG and Semantics Parsing especially); Experienced with Speech Recognition and Distributed System.
- **Languages**: Chinese(Native); English (Proficient); Japanese(Intermediate); Spanish(Beginner)